

Mammalian Comparative Genomics and Epigenomics

by

Tarjei Sigurd Mikkelsen

S.B. Mathematics with Computer Science
Massachusetts Institute of Technology, 2001

M. Eng. Biomedical Engineering,
Massachusetts Institute of Technology, 2002

Submitted to the Harvard-MIT Division of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2009

© Massachusetts Institute of Technology 2009. All rights reserved.

Signature of the Author
Division of Health Sciences and Technology
April 17, 2009

Certified by
Eric S. Lander, PhD
Professor of Biology
Thesis Supervisor

Accepted by
Ram Sasisekharan, PhD
Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering
Director, Harvard-MIT Division of Health Sciences and Technology

[This page is intentionally left blank]

Mammalian Comparative Genomics and Epigenomics

Tarjei Sigurd Mikkelsen

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on April 17, 2009 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

The human genome sequence can be thought of as an instruction manual for our species, written and rewritten over more than a billion of years of evolution. Taking a complete inventory of our genome, dissecting its genes and their functional components, and elucidating how these genes are selectively used to establish and maintain cell types with markedly different behaviors, are key challenges of modern biology. In this thesis we present contributions to our understanding of the structure, function and evolution of the human genome. We rely on two complementary approaches.

First, we study signatures of evolutionary processes that have acted on the genome using comparative sequence analysis. We generate high quality draft genome sequences of the chimpanzee, the dog and the opossum. These species share a last common ancestor with humans approximately 6 million, 80 million and 140 million years ago, respectively, and therefore provide distinct perspectives on our evolutionary history. We apply computational methods to explore the functional organization of the genome and to identify genes that contribute to shared and species-specific traits.

Second, we study how the genome is bound by proteins and packaged into chromatin in distinct cell types. We develop new methods to map protein-DNA interactions and DNA methylation using single-molecule based sequencing technology. We apply these methods to identify new functional sequence elements based on characteristic chromatin signatures, and to explore the relationship between DNA sequence, chromatin and cellular state.

Thesis Supervisor: Eric S. Lander
Title: Professor of Biology

[This page is intentionally left blank]

Acknowledgements

I am indebted to Eric Lander for his constant support and mentorship during my graduate career. I am also grateful to Jill Mesirov, James Galagan, Simon Kasif, David Gifford and Nicholas Patrikalakis for inviting me to work in their laboratories during my undergraduate years.

I would like to acknowledge the whole staff of the Broad Institute Genome Biology Program, the Genome Sequencing and Analysis Platform, the Genetic Analysis Platform, and the former Whitehead/MIT Center for Genome Research for generating the vast majority of the sequence and expression data analyzed throughout this thesis; Kerstin Lindblad-Toh, Mike Zody, Xiaohui Xie, Chad Nusbaum and Michele Clamp for many helpful discussions on sequencing, comparative genomics and related topics; Brad Bernstein and Alex Meissner for close collaborations on chromatin biology and epigenetics; Mike Kamal, David Jaffe, Manuel Garber, Evan Mauceli, Manfred Grabherr, Jean Chang and Andrey Sivachenko for providing significant computational support; Andreas Gnirke for patiently teaching me DNA affinity capture and a variety of other laboratory techniques; and Xiaolan Zhang, Li Wang, Hongcang Gu, Jake Jaffe, Veronica Saenz-Vash, Manching Ku and Dana Huebert for providing laboratory support and performing critical experiments.

I owe my gratitude to my parents Helga and Bjørn Ove Mikkelsen for their unconditional love and support.

[This page is intentionally left blank]

Table of contents

Title page	1
Abstract	3
Acknowledgements	5
Chapter 1: Introduction	9
Chapter 2: The chimpanzee genome	47
Chapter 3: The dog genome	147
Chapter 4: The opossum genome	197
Chapter 5: Proteomic analysis of conserved non-coding sequences	249
Chapter 6: Maps of histone methylation at key developmental loci	269
Chapter 7: Genome-wide maps of histone methylation	297
Chapter 8: Genome-scale maps of DNA methylation	333
Chapter 9: Integrative analysis of cellular reprogramming	369
Chapter 10: Future directions	399
Appendix 1: Mikkelsen, T. S. <i>et al.</i> Nature (2005) 437, 69-97	409
Appendix 2: Lindblad-Toh, K., Wade, C., Mikkelsen, T. S. <i>et al.</i> Nature (2005) 438, 803-819	429
Appendix 3: Mikkelsen, T. S. <i>et al.</i> Nature (2007) 447, 167-177	447
Appendix 4: Xie, X., Mikkelsen T. S. <i>et al.</i> PNAS (2007) 104, 7145-7150	459
Appendix 5: Bernstein, B. E., Mikkelsen, T. S. <i>et al.</i> Cell (2006) 125, 315-326	465
Appendix 6: Mikkelsen, T. S. <i>et al.</i> Nature (2007) 448, 553-560	477
Appendix 7: Meissner, A.*, Mikkelsen T. S.* <i>et al.</i> Nature (2008) 454, 766-770	487
Appendix 8: Mikkelsen, T. S. <i>et al.</i> Nature (2008) 454, 49-55	491

[This page is intentionally left blank]

Chapter 1: Introduction

[This page is intentionally left blank]

The completion of the Human Genome Project ^{1,2} at the beginning of this decade represents a major milestone in the human scientific endeavor. The DNA sequence generated by the project can be thought of as an instruction manual for our species, written and rewritten over more than a billion years of evolution. Genes encoded by the sequence specify the synthesis of the molecular building blocks necessary to create, sustain and propagate human life. Interactions of these genes with each other and their environment control the process of development, in which a single fertilized egg gives rise to daughter cells that progressively divide, differentiate and organize into the assembly of trillions of specialized cells that make up the human body. Taking a complete inventory of our genome, dissecting its genes and their functional components, and elucidating how these genes are selectively used to establish and maintain cell types with markedly different behaviors, are key challenges of modern biology.

In this thesis we present contributions to our understanding of the structure, function and evolution of the human genome. We rely on two complementary approaches. First, we study signatures of evolutionary processes that have acted on the genome using comparative sequence analysis. We generate high quality draft genome sequences of the chimpanzee, the dog and the opossum. These species share a last common ancestor with humans approximately 6 million, 80 million and 140 million years ago, respectively, and therefore provide distinct perspectives on our evolutionary history. We apply computational methods to explore the functional organization of the genome and to identify genes that contribute to shared and species-specific traits. Second, we study how the genome is bound by proteins and packaged into chromatin in distinct cell types. We develop new methods to map protein-DNA interactions and DNA methylation using single-molecule based sequencing technology. We apply these methods to identify new functional sequence elements based on characteristic chromatin signatures, and to explore the relationship between DNA sequence, chromatin and cellular state.

To provide context, we begin by reviewing relevant literature on genome and chromatin biology pre-dating the inception of this thesis (fall 2004). Readers with limited background in biological sciences might also find it useful to refer to general textbooks on molecular and cellular biology ^{3,4}, development ⁵, epigenetics ⁶, evolutionary theory ⁷⁻⁹ and genomics ¹⁰. Next, we briefly review advances in DNA sequencing, which is an enabling technology for this thesis. Finally, we summarize our specific contributions.

Genome biology

Genome biology is the study of the information content and global properties of the genetic material of living organisms. The basic unit of heredity is a gene. The physical manifestation of a gene is a set of nucleotide sequences (DNA in all cellular organisms) that specify (i) the synthesis of one or more gene products (RNA or proteins) and (ii) under what conditions these products are to be produced. A chromosome is a linear or circular DNA polymer that contains physically linked genes, potentially interspersed among nucleotide sequences that are inherited but are functionally inert or do not contribute to organismal traits. A genome is the complete set of chromosomes in a living cell (because eukaryotic cells carry two copies of each chromosome, we usually refer to their genomes as the complete haploid set of chromosomes). Major questions in the field follow three related themes:

Content: A key step in understanding a genome is to comprehensively identify the biological information encoded in its nucleotide sequence. How much of the total sequence encodes genes? Where is each gene located? What are their RNA or protein products? How are the coding and regulatory components of each gene organized? Are there other classes of biologically active or inert sequence features? Do chromosomes, or regions of chromosomes, differ significantly in their gene content or physical properties?

Regulation: Only a subset of the genes in a genome is expressed in any given cell type or state. How is regulatory information encoded in the genome? Are genes regulated independently, or does the genome structure facilitate coordinated regulation of multiple genes? How are differential gene expression patterns maintained through cell division and development?

Evolution: Understanding *why* a genome is organized and regulated the way it is requires understanding its evolutionary history. What is the nature of the evolutionary processes acting on a genome? How do these processes constrain or enable molecular and organismal change? What is the relative importance of genetic drift and natural selection? How rapidly do ancestral genes change or disappear? How do new gene products and regulatory elements emerge?

Answers to these questions would greatly inform future genetic and functional analyses of human development, physiology, disease and evolutionary history. We note that genome content, regulation and evolution can vary dramatically between species from different taxonomic kingdoms, phyla and classes. Here, we focus on the genomes of mammals, the class of vertebrate animals to which the human species belongs.

Sanger (dideoxy) sequencing

Efficient methods for the determination of the nucleotide sequence of a DNA polymer was first demonstrated in the mid-1970s when Maxam, Gilbert and Sanger independently published descriptions of sequencing methods that relied on gel electrophoresis to resolve DNA fragments that encoded sequence information at base pair resolution¹¹⁻¹³. While Maxam-Gilbert sequencing initially became the most widely used methodology, the Sanger sequencing method based on dideoxy chain termination eventually proved to be more practical and has been used in the vast majority of genome sequencing projects.

Initially a labor intensive process that yielded at most a few hundred bases of sequence information per experiment, dideoxy sequencing has been successfully scaled to a level where it has become feasible to read all nucleotides in the human genome many times over. Getting to this point involved extensive modification of the original chemistry to make it more amenable to automation. In particular, introduction of fluorescently labeled dideoxy terminators¹⁴ and engineered DNA polymerases¹⁵ proved to be key innovations. Replacing traditional slab gel electrophoresis with capillary electrophoresis¹⁶ greatly reduced reagent and labor costs. Large investments have also been made in parallelized and robotic sample preparation^{17,18}.

Because the dideoxy sequencing process is in practice limited to reading less than 1,000 bp from any one template, indirect strategies are required to infer the contiguous sequence of larger naturally occurring DNA polymers, such as each chromosome in a genome. The fundamental strategy for genome sequencing over the last three decades have been shotgun sequencing¹⁹⁻²². In this approach, a long DNA template is fragmented by mechanical shearing or enzyme digestion. Resulting DNA fragments are selected at random, ligated into a common sequencing vector, amplified by bacterial cloning, and sequenced using universal primers. The initial template is reconstructed computationally by searching for overlap between the sequenced fragments. Simple mathematical analysis can predict the expected number of gaps in the resulting sequence as a function of fragment coverage²³. Such gaps can be filled in by targeted sequencing as required.

The first DNA genome to be fully sequenced was the 5.4 kb genome of the bacteriophage phi-X174²⁴. Gradual improvements to sequencing and computational methods eventually led to sequencing of the first genome of a free-living organism, the bacterium *Haemophilus influenzae* (1.8 Mb), in 1995²⁵. The genome sequences of a variety of model organisms quickly followed, including the yeast *Saccharomyces cerevisiae* (12 Mb)²⁶, the roundworm *Caenorhabditis elegans* (97 Mb)²⁷ and the fruit fly *Drosophila melanogaster* (120 Mb)²⁸.

To obtain complete coverage of the 3,000 Mb human genome, the Human Genome Project employed a common variation of the shotgun strategy known as hierarchical shotgun sequencing. This two-step process begins with cloning large 100-200 kb fragments of the human genome into bacterial or P1-derived artificial chromosome (BAC/PAC) vectors. Overlapping BAC/PAC clones are shotgun sequenced individually and then combined into a largely contiguous whole genome sequence.

In a parallel commercial effort, Celera attempted to generate a human genome sequence using whole genome shotgun (WGS) sequencing²⁹, which skips the BAC/PAC subcloning step³⁰. At present, WGS is the dominant strategy for genome sequencing. Generating 20-40 million shotgun sequence reads, ideally from both ends of fragments of several known sizes (paired end-sequencing), is sufficient to infer the contiguous sequence of a mammalian genome using whole genome assembly software such as ARACHNE^{31,32} and PCAP³³.

The human genome sequence

Gradual advances in DNA sequencing technology allowed targeted sequencing of genomic regions of increasing size, from the 48kb growth hormone locus³⁴ and the 106 kb FosB/ERCC1 locus³⁵, to the 685 kb beta T cell receptor locus³⁶, and eventually the two smallest autosomes, chromosome 21 (25 Mb³⁷) and chromosome 22 (33.4 Mb³⁸). These early studies provided anecdotal insights into sequence composition, gene structure and evolution. The initial draft sequence of the whole human genome provided the first opportunity to assess the generality of these insights and to generate a comprehensive map. Here, we review some of its large-scale features.

Nucleotide composition. 41% of human DNA consists of GC base pairs, while AT base pairs make up the remainder. In the 1970s, density gradient separation studies revealed significant variation in GC content between large genomic fragments^{39,40}. Analysis of the complete genome sequence confirmed these studies and showed that the GC content of 20 kb intervals varies from ~30%-60%, a 15-fold higher spread than expected from uniform random fluctuations alone. Notably, several genomic properties, such as ‘dark’ and ‘light’ cytogenetic bands, repetitive element composition, gene density and recombination rate are strongly correlated with GC content, suggesting that GC-rich and GC-poor regions of the genome may have different functions⁴¹⁻⁴³.

A key question is therefore whether the large-scale variation in GC content is the direct product of natural selection, or simply a consequence of variation in neutral evolutionary processes acting across the genome⁴⁴. Bernardi and colleagues have proposed that variation in GC content reflects natural selection on thermal stability in species with high body temperatures, because GC-

rich DNA tends to be more stable than AT-rich DNA⁴⁵. Alternative mechanisms that do not invoke selection include variation in base misincorporation patterns during DNA replication due to changes in the free nucleotide pool during the cell cycle⁴⁶; variable cytosine deamination rates (see below) in GC-rich and AT-rich DNA⁴⁷; and ‘biased gene conversion’, where heterozygous sites are preferentially repaired to GC during homologous recombination⁴⁸ and would therefore lead to variation in GC content as a consequence of variation in recombination rates^{49,50}. A better understanding of mammalian genome evolution is needed to differentiate between these hypotheses.

CpG islands. So-called ‘CpG islands’ are smaller sequence features related to GC content⁵¹. CpG dinucleotides in the human genome are generally methylated on the cytosine base. Spontaneous deamination of C residues gives rise to uracil residues that are recognized and repaired by the cell, whereas deamination of methyl-C residues gives rise to T residues. CpG dinucleotides therefore steadily mutate to TpG nucleotides, leading to a scarcity of CpGs across the genome. However, small ‘islands’ of relatively high CpG density, usually spanning less than a thousand base pairs, can be found throughout the genome. CpGs in these islands are generally unmethylated. Many, but not all, known genes have CpG islands at their 5’ ends, suggesting that CpG islands are involved in regulation of gene expression and might be useful markers for the identification of novel genes⁵¹⁻⁵⁴. Approximately 29,000 CpG islands have been identified across the non-repetitive fraction of the human genome. They are unevenly distributed across each chromosome, in a pattern that correlates with gene density. The human genome also harbors tens of thousands of copies of transposons that meet the classic definition of a CpG island⁵³. These transposon associated CpG islands tend to be shorter than gene-associated islands and are generally thought to not have any biological role⁵⁵, although there is little direct evidence for this distinction.

Repetitive elements. Approximately half of the human genome sequence can be recognized as copies, or relics, of transposable elements, parasitic sequence elements that have copied and inserted themselves throughout the genome^{56,57}. The vast majority of transposable elements can be assigned to one of four major categories, based on their sequence composition and replication strategies: Long interspersed elements (LINEs), short interspersed elements (SINEs), long terminal repeat (LTR) retrotransposons and DNA transposons.

LINEs are 6-8 kb autonomous sequence elements that contain an internal promoter and two ORFs. LINE RNA transcripts associate with their encoded proteins and are translocated back to the nucleus, where they are reverse transcribed directly into the genome. Reverse transcription is inefficient, leading to large numbers of truncated, nonfunctional insertions. Approximately 21% of the human genome is clearly derived from insertions of each of three distantly related LINE

families: LINE1, LINE2 and LINE3. Only LINE1 is thought to remain active. LINEs tend to be clustered in AT-rich, gene-poor regions of the genome.

SINEs are 100-400 bp non-autonomous sequence elements that contain an internal promoter but no ORFs. They are thought to use their 3' ends to 'hijack' the LINE machinery for transposition^{58,59}. Approximately 13% of the human genome can be recognized as being derived from SINE insertions. The Alu element is the most abundant, with one million recognizable copies, and remains active in modern humans. SINEs tend to be clustered in GC-rich, gene-rich regions of the genome.

LTR retroposons are flanked by direct repeats that contain promoter activity. Autonomous LTR elements (retrotransposons) contain *gag* and *pol* genes that enable reverse transcription in a cytoplasmic virus-like particle and subsequent integration into the genome. LTR insertions have contributed at least 8% of the human genome sequence. The majority of these insertions have lost their internal coding sequences due to homologous recombination between the flanking repeats. LTR activity is thought to be on the brink of extinction in the human genome.

DNA transposons contain one ORF flanked by inverted repeats. It encodes a protein that binds near the flanking repeats and moves the transposon to a new location through a 'cut-and-paste' mechanism. DNA transposons in the human genome can be grouped into at least seven major families with apparently independent origins. They have contributed at least 3% of the genome sequence, but appear to have become completely inactive over the last ~50 million years.

Copies of transposable element insertions gradually diverge as they age, until they degenerate into sequences that bear no resemblance to their parental elements. Simulations suggest that the mutation rate of the human genome (see below) is sufficient to obscure the origin of any insertion older than ~150-200 million years⁶⁰. Because transposable elements similar to those found in the human genome are known to be substantially older than this, it is likely that much of the remaining 'unique' sequence of the human genome are also derived from them⁵⁶.

A key question surrounding transposable elements is how these 'selfish' pieces of DNA affect the fitness and evolution of their hosts. Transposons were first described by McClintock in the 1950s as mobile 'controlling elements' that could both induce chromosome breaks and affect the expression of nearby genes in maize^{61,62}. In the 1960s, Britten and colleagues showed that about half of the genomes of mammals had to consist of families of repetitive sequences, based on the reassociation kinetics of DNA fragments obtained from them^{63,64}. Since then, a number of researchers have proposed key roles for transposable elements in the evolution of higher organisms, for example as templates for organization of chromatin, as sources of genetic variation due to

insertion mutagenesis and homologous recombination, as vehicles for copying existing functional elements to new locations, and as sources of proteins and regulatory elements that can be exapted and modified for the benefit of their host⁶⁵⁻⁷². Individual examples of most of these models have been identified. Homologous recombination between Alu elements has been shown to cause chromosomal rearrangements in the human genome⁷³⁻⁷⁶. A small number of functional ‘retrogenes’ are thought to have been generated by LINE-mediated reverse transcription of non-LINE RNA transcripts^{77,78}. About fifty human protein-coding sequences^{56,79} and a few dozen known regulatory elements^{77,80} clearly originate from exaptation of transposon sequences. It remains to be determined whether these examples were isolated incidents or represent major sources of innovation in our evolutionary history.

Segmental duplications. The human genome contains numerous instances of two or more ~1-200 kb intervals with very high nucleotide sequence similarity (>90%) that are not copies of transposable elements⁸¹⁻⁸³. The initial draft genome sequence showed that such ‘segmental duplications’ cover at least ~3.6% of the genome sequence, although recent duplications can be difficult to assemble correctly from shotgun sequencing data. The estimate was later revised up to at least 5.3% based on a more complete version of the sequence². Duplications can occur both within a chromosome and between non-homologous chromosomes, and their distribution varies significantly between different chromosomes and chromosomal regions. Interchromosomal duplications are particularly common in pericentromeric and subtelomeric regions, apparently due to a steady bombardment of insertional translocations⁸⁴. Interchromosomal duplications often occur in clusters and are predisposed to recurrent deletions or rearrangements due to paralogous recombination. Such events have been associated with a variety of genetic diseases, including Prader-Willi/Angelman, DiGeorge and Williams’ syndromes⁸⁵⁻⁸⁸. Segmental duplications have also been shown to harbor novel gene families evolving under strong positive selection⁸⁹, suggesting that they may provide a useful substrate for evolutionary innovation.

Gene content. Genes can generally be divided into protein-coding and non-coding RNA genes based on the functional form of the products they specify. Protein-coding sequences can easily be identified in the genomes of simple organisms, such as bacteria and yeast, based on long open reading frames (ORFs). The task is much more difficult in the genomes of complex organisms such as humans, because their protein-coding sequences make up only a small fraction of the total genomic sequence, and because most functional ORFs are split into multiple short exons separated by longer, non-coding introns that are spliced out before translation. This creates a signal-to-noise problem, because the number of potential ORFs in the genomic sequence is much higher than the

number of ORFs that are ever transcribed, spliced and translated into proteins in the organism. Efforts to create a complete catalog of human genes have therefore relied on the indirect strategy of first isolating and identifying gene products through cDNA or amino acid sequencing, and then aligning these products to the genome to identify the genes which specified them or novel genes with similar structures or protein domains.

Various attempts were made at estimating the total number of protein-coding genes prior to completion of the human genome sequence. Extrapolation from early work on messenger RNA reassociation kinetics suggested about 40,000 distinct genes⁹⁰. Extrapolation from the number of CpG islands and the frequency with which they are associated with genes suggested about 70,000-80,000⁹¹. Different analyses of expressed sequence tags (ESTs⁹²) led to predictions ranging from 35,000 to 120,000^{93,94}.

With the availability of the genome sequence, computational methods were used to identify any potential coding sequences in the genome based on sequence similarity to any known EST, messenger RNA, protein sequence or domain from any species. This effort led to a set of ~32,000 predicted genes. These predictions were then merged with ~15,000 well-characterized human transcripts and proteins known at the time. Various considerations on the accuracy of the prediction algorithms and the completeness of the initial draft sequence led to the estimate that ~24,000 of the predicted genes would turn out to be real, and that the genome contains an additional 5,000-10,000 protein-coding genes, for a total of 30,000-35,000. Subsequent corrections to, and increased coverage of, the genome sequence, improved computational methods and expanded cDNA collections have led to downward adjustments to this estimate, down to 19,600-25,000 at present². The lower bound is the number of currently known genes. The upper bound is likely to be conservative, based on the low number of novel genes having been discovered despite comprehensive cDNA sequencing efforts. In total, the known and predicted protein-coding sequences make up ~1.5% of the human genome.

The genome also contains a variety of non-coding genes that specify RNA molecules that contribute directly to cellular processes, rather than being translated into proteins⁹⁵⁻⁹⁷. So far 1,000-2,000 non-coding genes with known or putative functions have been identified in the genome, including transfer RNAs, ribosomal RNAs, small nucleolar RNAs, spliceosomal RNAs, regulatory micro RNAs, and a small number of idiosyncratic long RNAs associated with imprinting or X inactivation, such as XIST⁹⁸. Unlike protein-coding genes, which are transcribed into messenger RNAs that all share a common primary structure and grammar (the genetic code), non-coding RNAs exist in a wide variety of sizes and structures. It is therefore difficult to estimate the

completeness of the current non-coding gene catalog. Chromosome-wide surveys of transcriptional activity suggest that the genome gives rise to a large number of currently unclassified RNA transcripts⁹⁹, but to what extent these represent functional gene products or reproducible transcriptional noise remains unclear.

Locating the regulatory sequence elements that specify when and where each gene product is to be synthesized is considerably more challenging than locating the sequences that specify the product itself. In contrast to the genomes of many simple organisms where regulatory elements are generally located near transcription start sites, human regulatory elements can be located far from the sequences which transcription they control. For example, the *Shh* gene is controlled in part by a regulatory element situated over a million nucleotides away from its protein-coding sequence, within an intron of a different protein-coding gene¹⁰⁰. Moreover, analysis of known regulatory elements have so far failed to uncover characteristic patterns that make them clearly stand out against the vast background of non-functional sequences. Alternative approaches are required to construct a parts list of the human genome that include comprehensive coverage of both non-coding genes and regulatory elements.

Comparative analysis of genome sequences

Comparative analysis has emerged as a promising tool for studying genome evolution and for locating functional information encoded in DNA sequences. This approach relies on evolutionary theory and the principle of common descent. As genetic information is passed from generation to generation, individuals acquire mutations through errors in DNA replication or repair. Each mutation has a small chance of spreading throughout a population and becoming a fixed part of a species' genome. Mutations that happen to change functional sequence information might have positive or negative effects on the fitness of the individuals that carry them. Natural selection will act on such mutations to increase or decrease their chance of fixation, respectively. As evolution ceaselessly tinkers with gene pools over millions of years, new species form and diverge from common ancestors. Comparing the genomes of related species to identify sequences that have changed less than would be expected in the absence of selection should therefore help pinpoint shared genes. Sequences that have changed more than expected might also help pinpoint genes responsible for the emergence of novel traits.

Comparative analysis of genomic sequences dates to at least 1975, when Pribnow and Schaller identified the -10 RNA polymerase recognition site, an essential part of prokaryotic promoters, in part by comparing promoter sequences from bacteriophages and *E. coli*^{101,102}. In one

of the first applications to mammalian biology, Tagle and colleagues used ‘phylogenetic footprinting’ of the galago, rabbit and mouse ϵ and γ globulin loci to identify conserved regulatory elements likely to control their expression in primates¹⁰³. They also noted that the transition from embryonic to fetal γ globulin expression in primates correlates with an elevated non-synonymous substitution rate.

The first comparative analysis of extended genomic regions between human and other mammals involved the rat γ -crystallin locus¹⁰⁴ and the mouse β -globin cluster¹⁰⁵. Both analyses found low overall sequence similarity outside of the orthologous ORFs, in part due to insertions of large numbers of lineage-specific transposable elements, but noted the presence of small conserved non-coding sequences of unknown function. In contrast, comparative analysis of the 100 kb human and mouse α/δ T cell receptor loci showed extensive organizational and non-coding sequence conservation, suggesting that this region either contains a large number of regulatory sequence elements or has an unusually low mutation rate¹⁰⁶. In 1997, Oeltjen and colleagues demonstrated that conserved non-coding sequences identified by comparison of the human and mouse Bruton’s tyrosine kinase (BTK) loci possessed regulatory activities in transient transfection experiments¹⁰⁷. Similar findings were later made in, for example, orthologous interleukin¹⁰⁸ and Hox¹⁰⁹ gene clusters. These results led to a growing appreciation of the value of sequencing additional mammalian and vertebrate genomes to facilitate detection of novel coding and regulatory sequences in the human genome^{110,111}.

The utility of whole-genome comparative analysis for identification of protein-coding sequences, non-coding RNAs and regulatory elements, as well as for gaining insight into genome evolution, has been conclusively demonstrated for model systems such as yeast¹¹²⁻¹¹⁵ and worm¹¹⁶. Here, we review lessons learned from comparative genome analysis of mammals and more distantly-related vertebrates.

Comparison of the human and murid genomes

Sequencing of the mouse genome⁶⁰, and later the rat genome¹¹⁷, launched the era of mammalian comparative genomics. The mammalian lineages leading to modern humans and murids diverged approximately 75 million years ago, while the lineages leading to mice and rats diverged 12-24 million years ago^{118,119}. Thus, human-murid and mouse-rat comparisons provided two different perspectives on mammalian evolution. Here, we review key lessons from the whole-genome analyses.

Synten. The human and murid genomes have each been repeatedly shuffled by chromosomal rearrangements after their lineages separated. The rate of these rearrangements have been low enough, however, that local gene order generally remains the same. Early evidence of conserved gene order in mammals came from the observation that the albino and pink-eye dilution mutants were closely linked in both mouse and rat^{120,121}. Consistent with this, comparison of the mouse and rat genome sequences show that they can be divided into ~100 orthologous segments with largely intact gene order. Extrapolating from linkage and cytogenetic data for 83 loci, Nadeau and Taylor estimated that the human and mouse genomes could be parsed into ~180 syntenic segments¹²². Analysis of the complete genome sequences identified ~300 syntenic segments varying in length from 1 to 65 Mb, which covers >90% of each genome. A cytogenetic analysis that included outgroups suggested that the rate of rearrangement was significantly higher in the lineage leading to the murids¹²³.

A question arising from the analysis of mammalian synteny is whether the chromosomal breaks involved in large-scale rearrangements occur randomly across the genome. In 1973, Ohno postulated the random breakage model of genome evolution¹²⁴, which implies that the number of breakpoints can be used as a measure of genetic distance and that significant deviation from random breakage can be used as a test for selection on gene order. Genomic synteny maps of human and mouse genomes were at first deemed to be consistent with this model⁶⁰. However, based on analysis of human, mouse and rat, Pevzner and colleagues have argued that breakage might be biased towards structurally unstable ‘hotspots’^{125,126}. High resolution synteny analysis involving additional mammals should help resolve this issue.

Shared and lineage-specific genes. Catalogs of known and predicted genes in the mouse and rat genomes have been generated by computational analysis of cDNA and protein evidence (see above). Both genomes are predicted to contain 20,000-25,000 protein-coding genes and ~1,000 known non-coding RNA genes. Similar to human, substantial evidence of additional non-coding transcription of largely unknown function has also been found in mouse¹²⁷.

The vast majority (>99%) of known and predicted mouse and rat protein-coding genes have clear homologs in the human genome, which is consistent with the notion that duplication of ancestral genes, rather than *de novo* evolution from non-coding sequences, is the most common mechanism for generating new genes¹²⁸. The fraction of human protein-coding genes that have single, unambiguous (1:1) orthologs in each of the murid genomes is estimated to be 80-90% (after accounting for shortcomings in the genome assembly). This is similar to the fraction of 1:1 orthologs between mouse and rat (86-94%), despite the difference in divergence times.

Protein-coding genes that are not part of 1:1 ortholog pairs are frequently found in clusters of lineage-specific duplications in one or more of the species, usually within syntenic segments. The largest such clusters involve olfactory receptor genes and the cytochrome P450 gene family, which is involved in xenobiotic metabolism. Both of these gene families have undergone parallel duplication, expansion and gene loss in both the human and murid lineages. Interestingly, the majority of lineage-specific duplications in these and other families appear to be of very recent origin, based on their high synonymous sequence similarities. One possible interpretation of this pattern is that many lineage-specific genes represent transient duplication events that are destined for deletion due to lack of functional benefit. The functional proteome of humans and murids might therefore be even more similar than ~90% 1:1 orthology suggests.

An emerging theme from the field of evolutionary developmental biology is that changes in gene regulation rather than gene products might be the most important driver of evolution of morphological diversity, such as that seen across the class of mammalian species^{129,130}. As early as 1975, Wilson and colleagues argued that the substitution rate between known human and chimpanzee proteins were too low and symmetrical to account for the accelerated anatomical evolution evident in modern humans^{131,132}. Comparing the rates of evolutionary innovation in coding and regulatory sequences is not feasible, however, without a much better understanding of the identity and evolution of the latter.

Nucleotide divergence. Approximately 40% of the human genome can be aligned to the murid genomes at the nucleotide level. The remaining 60% is thought to be accounted for by insertion of lineage-specific transposable elements and turnover of ancestral sequence. Using orthologous transposable element relics as a proxy for neutrally evolving sequence, the substitution rate between human and mouse genomic DNA has on average been ~0.5 per site, with the absolute rate being approximately twice as high on the murid lineage compared to human. The rates of small insertion and deletion events (not related to transposable elements) have been ~9 and ~21 per kb, respectively. These rates all vary significantly across the genome, in a pattern that shows complex correlations with sequence properties such as GC and transposable element content. The source of this variation is not fully understood, but may include reduction in genetic diversity due to natural selection¹³³, biased gene conversion¹³⁴ or transcription associated mutagenesis¹³⁵. Gaining a better understanding of substitution rate variation is critical to any genomic analysis that relies on accurate detection of signals of positive or negative selection, such as phylogenetic footprinting.

Proportion of the genome under negative selection. Sequencing of the mouse genome provided the first opportunity to estimate how much of the human genome has been evolving under

negative selection since our last common ancestor, and therefore presumably contains functional sequence elements. Waterston and colleagues estimated this fraction by measuring human-mouse sequence similarity across short (50 or 100 bp) intervals of aligned, orthologous sequence⁶⁰. They then compared the resulting distribution of sequence similarities to the distribution expected under neutral evolution, as estimated from relics of ancestral transposable elements. Variations on this method consistently showed a ~5% excess of intervals with high sequence similarity.

The 5% estimate came as something of a surprise because it is three times larger than what can be accounted for by protein-coding sequences alone, suggesting that the majority of functional information in mammalian genomes have other roles. Some conserved non-coding sequences overlap promoters, enhancers and untranslated regions of protein-coding genes, but many are not associated with any known gene. In fact, there appears to be no simple spatial correlation between conserved coding and non-coding sequences, and conserved non-coding sequences are frequently found in long ‘gene deserts’¹³⁶. The strength of negative selection also appears to vary significantly between different sequence elements. At one extreme, Bejarano and colleagues have identified several hundred ‘ultraconserved elements’ associated with developmental genes that show perfect nucleotide identity between human, mouse and rat over at least 200 bp¹³⁷. Such high levels of constraint have been difficult to explain, because degeneracy in the genetic code and in transcription factor binding preferences suggests that at least some sites is both protein-coding and regulatory elements should have little functional impact.

How accurate the 5% estimate is remains an open question. It critically depends on assumptions made about how mammalian genomes change in the absence of selection. It is difficult to assess how general these assumptions are from analysis of only three complete genome sequences. For example, hypothetical mutational ‘cold spots’ could yield a high rate of false positives. Moreover, in contrast to early optimism¹¹⁰, comparison of the human and murid genomes does not provide sufficient statistical power to accurately identify most individual sequence elements under selection, which limits opportunities for follow-up studies. Comparing the human genome to that of more distantly related species, such as chicken¹³⁸ and fish¹³⁹, have yielded higher specificity (but lower sensitivity) and helped to accurately predict new regulatory elements^{140,141}. While most human protein-coding sequences have conserved orthologs in these species, significantly less conserved non-coding elements can be found (2-3 fold less with chicken, >100 fold less with fish). Due to the high background substitution rate (>1.7 per site for human-chicken), it is difficult to say whether this stems from failure to align orthologous elements, lack of specificity in the human-murid comparisons, or evolutionary innovation in mammals. Sequencing of additional

mammalian genomes should help inform these issues and unlock the full power of comparative analysis for understanding human biology.

How close the 5% estimate is to the fraction of the human genome that is functional at present is also unclear. If lineage-specific loss or gain of regulatory or other non-coding elements is common, as a comparison of limited sequences from eight mammals has suggested¹⁴², many would have been missed altogether by the human-murid analysis. Additional mammals might help identify ancestral elements that have been lost in the murids. Identification of new elements specific to the human lineage is difficult, however, because comparative analysis of closely related species has less statistical power. Complementary approaches that do not rely on sequence conservation might therefore be required to compile a comprehensive catalog of functional elements in the human genome.

Chromatin

In higher organisms, the genome exists in the cell nucleus as part of a complex combination of DNA, RNA and proteins referred to as chromatin³. X-ray diffraction and electron-microscopy studies in the early 1970s suggested that chromatin might be organized into a repeating, “spheroid” structure^{143,144}. In 1974, Kornberg and Thomas demonstrated that the fundamental repeating unit of chromatin is the nucleosome, which consists of ~147 base pairs of DNA wrapped around an octamer of different histone proteins (H2A, H2B, H3, H4). Nucleosomes are connected by 20-100 bp of free linker DNA. Each nucleosome is stabilized by several hundred protein-DNA interactions¹⁴⁵.

Cellular processes that utilize genetic information, such as gene activation or replication, require temporary disruption of nucleosome interactions to gain access to the DNA sequence. Early studies of eukaryotic gene regulation showed that active genes and functional elements tend to be associated with less compact chromatin than silent genes, as measured by nuclease accessibility^{146,147}. “Open” chromatin can therefore be thought of as a signature of active genes or regulatory elements in a given cell type or state. Indeed, mapping sequences that are hypersensitive to DNase I digestion has been shown to be an effective and specific method for identification of active promoters and other regulatory elements¹⁴⁸. For example, the locus control region (LCR) which regulates β -globin expression during erythroid development was identified and dissected using nuclease hypersensitivity mapping^{149,150}.

The open chromatin structure observed at active genes is not simply an indirect consequence of non-histone proteins interacting with the DNA sequence. In the late 1980s,

pioneering experiments by Han and Grunstein demonstrated that histone depletion lead to widespread transcriptional initiation *in vivo*¹⁵¹. In the early 1990s, genetic and biochemical experiments showed that sequence-specific transcription factors mediated gene activation in part through recruitment of ATP-dependent chromatin remodeling complexes¹⁵²⁻¹⁵⁴. These observations imply that the nucleosome structure contributes directly to gene repression. Controlling chromatin compaction might provide a mechanism for controlling which parts of the information encoded in the genome is accessible at any given time and therefore help stabilize lineage commitment and cell state¹⁵⁵.

Modulation of DNA accessibility is mediated by chromatin remodeling complexes, but often involves covalent modifications of histones or the DNA itself. Various evidence from model organisms suggest that these modifications might provide highly informative signatures of different processes involved in genome regulation.

Histone modifications. Histones are globular proteins with flexible N-terminal tails. Biochemical studies have revealed that specific residues on these tails can be subject to a variety of post-translational covalent modifications, including acetylation, methylation, phosphorylation, ubiquitination and ADP-ribosylation¹⁵⁶. Pioneering experiments in yeast showed that the N-terminal tail of histone H4 is required for stable repression of the silent mating type loci¹⁵⁷. The same group later showed that acetylation of specific residues in the H4 tail are required for transcriptional activation¹⁵⁸.

Histone acetylation was at first thought to mediate its activating function primarily by removing positive charge from lysine residues and therefore decrease the strength of electrostatic interactions with the negatively charged DNA backbone. Eventually, it became clear that acetylated residues could also directly block interactions with repressive chromatin remodeling proteins SIR3/4¹⁵⁹ and serve as recognition sites for a protein domain found in a variety of transcriptional co-activators¹⁶⁰. Searches for additional proteins that interact with histone modifications soon revealed that methylation of H3 lysine 9 (H3K9me) serves as a recognition mark for a different class of repressive proteins (HP1/Swi6)^{161,162}.

It quickly became clear that multiple pathways converge on the histones and use covalent tail modifications as anchoring points for regulatory enzymes. A growing number of histone tail modifications have been associated activating and repressive protein complexes, primarily from studies in model systems such as yeast and fly. For example: H3 lysine 4 methylation (H3K4me) has been associated with transcriptional initiation, potentially due to interactions between the initiating form of RNA polymerase and H3K4 methyl-transferases¹⁶³⁻¹⁶⁵; H3K27me has been

shown to be catalyzed and recognized by repressive Polycomb group (PcG) proteins¹⁶⁶⁻¹⁶⁸; H3 lysine 36 methylation (H3K36me) has also been shown to mediate transcriptional repression, but appears to be associated with the elongating form of RNA polymerase *in vivo*, potentially to prevent aberrant initiation in active gene bodies¹⁶⁹. H4 lysine 20 methylation (H4K20me) has been associated with heterochromatin formation¹⁷⁰.

The potential for combinatorial modification patterns led to the formulation of the 'histone code' hypothesis, which states that different histone modifications can influence each other and have synergistic or antagonistic effects on gene activity or related processes¹⁷¹. A better understanding of the genomic distribution and functional role of histone modifications will be required to test this hypothesis.

DNA methylation. As alluded to above, cytosines in CpG dinucleotides are frequently methylated in vertebrate and mammalian genomes. In 1975, Riggs and Holliday suggested that this covalent chromatin modification might have a role in regulating gene expression^{172,173}. The discovery that promoter-associated CpG islands were strongly represented in the unmethylated fraction of the mouse genome provided anecdotal support for this hypothesis^{51,52}. Direct evidence for a causal role came from studies showing that DNA methylation can directly interfere with transcription factor binding¹⁷⁴⁻¹⁷⁶, and from the discovery of methyl-CpG binding domain proteins that target repressive nucleosome remodeling complexes to methylated sequences¹⁷⁷⁻¹⁸¹.

Genetic evidence suggests that a single enzyme, DNMT1¹⁸², is responsible for maintaining DNA methylation patterns in mammalian cells^{183,184}. Two additional enzymes, DNMT3A and DNMT3B, possess *de novo* methyltransferase activities, are highly expressed in developing embryos and are responsible for establishing global DNA methylation patterns following implantation^{184,185}. DNMT1 shows strong preference for hemi-methylated over unmethylated CG/GC base pairs. This preference provides a clear mechanism for faithful transmission of established DNA methylation patterns through mitosis. After DNA replication, unmethylated CG/GC pairs give rise to two unmethylated copies while symmetrically methylated CG/GC pairs give rise to two hemi-methylated copies. DNMT1 then 'completes' the methylation of the hemi-methylated copies. Interestingly, experiments by Jones and Taylor have demonstrated that transient inhibition of DNMT1 methylation in differentiated cells using small molecules appear to revert cultured fibroblasts to a more pluripotent state in a cell-cycle dependent manner^{186,187}. Thus, DNA methylation appears to play a key role in maintaining differentiated cell states. DNA methylation has also been implicated in repression of transposable elements, genomic imprinting and X inactivation¹⁸⁸.

Chromatin state and cell state. Lineage-commitment and cellular differentiation lie at the core of animal development. As embryonic cells grow and divide, they organize into anatomical structures and take on specialized states. These states are associated with differential gene expression patterns that can be remarkably stable, even after the signals that established them are no longer present. Seminal experiments by Hadorn and colleagues demonstrated the concept of ‘cellular memory’¹⁸⁹: clusters of cells from fly embryos that were destined to give rise to particular adult structures were isolated and made to proliferate substantially longer than they would during normal development. When reintroduced into embryos, these cells still differentiated into the anatomical structures and appendages they were initially programmed for. Understanding the molecular basis of this memory is a key challenge in developmental biology.

It has long been speculated that differentiated cells retain the genetic material necessary to control all of development¹⁹⁰. Cloning and nuclear reprogramming experiments in amphibians and mammals¹⁹¹⁻¹⁹⁵ eventually proved that this was the case. All nucleated cells contain the same genetic information as the totipotent zygote (with a few exceptions, such as mature B and T cells due to recombination of the immunoglobulin loci). Thus, lineage-commitment and cell state must be maintained through cell division by ‘epigenetic’* mechanisms. There are two known molecular systems that appear contribute to epigenetic inheritance: DNA methylation (see above) and the Polycomb/Trithorax system^{188,196}. Both systems are closely related to chromatin state.

The Polycomb/Trithorax system was first described in *Drosophila melanogaster*¹⁹⁷. The Polycomb group (PcG) proteins were named after mutations in 18 different genes that cause ectopic sex combs on the legs of male flies. The phenotype is caused by failure to repress expression of Hox genes, which are transcriptional regulators that specify anatomical patterns in the embryo. The Trithorax group (TrxG) proteins are named after mutations in 17 different genes that act as suppressors of the ectopic sex comb phenotype. Biochemical purification and analysis of the PcG and TrxG proteins show that they are components of large complexes that interact with and modify histone tails. As noted above, the recruitment and repressive activity of PcG proteins have been closely linked to tri-methylation of H3 lysine 27 (H3K27me3). The activity of TrxG proteins is less understood, but appears to be associated with methylation of H3 lysine 4 (H3K4me)¹⁹⁸. PcG and TrxG activities are thought to initially be recruited to target loci through sequence-specific binding to Polycomb Response Elements (PREs) scattered throughout the fly genome¹⁹⁹.

* The term ‘epigenetic’ has a variety of proposed definitions. Here, we adapt the recent definition of Bird²²¹ as “the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states”.

Human and other mammalian genomes contain highly conserved homologs of most of the PcG and TrxG proteins identified in fly¹⁹⁶. Recent experiments in mice have shown that PcG proteins are involved in maintaining the pluripotent state of embryonic stem cells, are required for successful embryonic development and contribute to X chromosome inactivation in differentiated female cells²⁰⁰⁻²⁰². Few mammalian PcG/TrxG targets are known, however, and no homologs of fly PREs have so far been identified. Unlike DNA methylation, the molecular basis for mitotic inheritance of PcG-mediated repression is also not yet understood¹⁹⁶.

Mapping chromatin state

Mapping the location of histone tail modifications and DNA methylation at high resolution might be a highly informative approach to identify functional elements, particularly activating and repressive regulatory elements. Gaining a better understanding of the distribution of epigenetic marks should also be helpful for exploring the relationship between DNA sequence, chromatin state and differentiation. To date, virtually nothing is known about the distribution of histone modifications and DNA methylation across mammalian genomes.

Accurate assays have been developed for mapping protein-DNA interactions, including histone tail modifications at single sequence elements or genes. The most practical assay is chromatin immunoprecipitation (ChIP), in which an antibody is used to enrich DNA from genomic regions carrying a specific epitope.²⁰³⁻²⁰⁵ The level of enrichment relative to the expected background can be measured by quantitative PCR. The ‘gold standard’ assay for DNA methylation is bisulfite sequencing, which involves chemical conversion of unmethylated cytosines to uracils²⁰⁶. Bisulfite treated fragments can be amplified by PCR and sequenced to reveal DNA methylation patterns at nucleotide resolution. The major challenge to generating genome-wide chromatin state maps lies in scaling the existing assays to allow profiling of multiple chromatin modifications across large mammalian genomes in multiple cell types without incurring prohibitive labor or reagent costs.

Proof-of-principle experiments have recently demonstrated the feasibility of mapping chromatin state across small genomes using microarray and sequencing technologies. Several groups have used microarrays containing amplified intergenic regions and ORFs to assay the distribution of H3K4 methylation, H3/H4 acetylation and related enzymes across the yeast genome^{164,207,208}. A variation of the serial analysis of gene expression (SAGE) method²⁰⁹ has also been used to directly sequence ChIP fragments from the yeast genome to identify acetylated regions²¹⁰. Scaling these methods to mammals, which genomes are two orders of magnitude larger than yeast,

and which utilize a significantly larger array of cell types and states, would be cost-prohibitive using either microarrays or Sanger sequencing, however. New technologies are therefore needed to meet this challenge.

Single molecule-based sequencing

Initially motivated by the Human Genome Project, tremendous effort has been expended into developing cost-efficient, high-throughput instruments capable of decoding any collection of DNA sequences. Similar to the production of semiconductors, the cost of dideoxy sequencing has decreased exponentially over the last two decades²¹¹. However, with the stated post-Human Genome Project goal of sequencing a complete human genome for USD \$1,000, the consensus in the DNA sequencing community has been that an entirely new generation of sequencing technology has to be developed.

Various non-Sanger sequencing instruments have recently been developed and commercialized. These instruments are based on the unifying principle of cyclic array sequencing. They achieve increased throughput and decreased costs by using a single reagent volume to simultaneously infer the sequence of millions (potentially billions) of DNA features immobilized on a surface. Each DNA feature may be a single molecule or an ensemble of identical molecules in close spatial proximity. Sequencing takes place in progressive cycles where in each cycle an enzymatic process is used to interrogate one or more nucleotide position from all of the DNA features in parallel. The outcome of each cycle is reported by light or fluorescence signals and captured using CCD imaging of the array. After multiple sequencing cycles, the location and composition of each DNA feature are inferred from the resulting series of images. The details of DNA feature generation, deposition and interrogation differ significantly between instruments designs.

Pyrosequencing of emulsion PCR features. In this approach, a collection of DNA templates are fitted with common adapters. Emulsion PCR²¹² is used to generate DNA features consisting of a large number of identical copies of each template immobilized on the surface of micrometer-scale paramagnetic beads. These beads are deposited across millions of picoliter-scale wells etched into the surface of a fiber optic bundle. The concentration of beads is titrated to maximize throughput while minimizing the number of wells expected to contain multiple beads. Sequence interrogation is done using the pyrosequencing method²¹³. In each cycle, one of the four DNA nucleotides is introduced into the common reaction volume and polymerase-mediated incorporation events are detected by monitoring luciferase-based light generation upon

pyrophosphate release. Parallel incorporation across identical templates on a single bead amplifies the signal for robust detection. Unincorporated nucleotides are removed and the process is repeated. A commercial implementation of this method by 454 Life Sciences can generate >100 bp sequence reads at a cost per read that is roughly an order of magnitude lower than for capillary Sanger sequencing.

Sequencing of emulsion PCR features by ligation. In this approach, DNA features are also prepared by bead-based emulsion PCR. The amplified beads are then randomly distributed on a glass slide and immobilized by a thin layer of polyacrylamide gel or by direct covalent attachment to the surface. Sequencing is achieved by sequence-specific ligation rather than polymerase-based extension [214]. In each cycle, an anchor primer is first hybridized to a universal adapter on each DNA template. Next, the slide is exposed to a population of fluorescently labeled degenerate nonamers (single-stranded 9 bp DNA sequences). The nonamer population is designed such that the attached fluorophore identifies the base at one particular position within it. The ligase discriminates for sequence complementarity up to some distance from the ligation site, ensuring that nonamers with one of the fluorophores are preferentially ligated to each DNA feature. After ligation, the array is imaged in four colors. Then the ligation products of the anchor primers and 9-mers are stripped from the beads, and the process is repeated. A commercial implementation of this method by Applied Biosystems can generate short (25-35 bp) reads at a cost per read that is roughly two orders of magnitude lower than for capillary Sanger sequencing.

Sequencing of bridge PCR features by synthesis. In this approach, DNA features are generated directly on a glass slide by bridge PCR ²¹⁵. Two universal primers are first immobilized to the glass surface. The primers are complementary to adaptors ligated onto each DNA template and serve to capture the fragments on the surface. Upon thermal cycling with all non-DNA reagents moving freely in the aqueous phase, DNA features corresponding to a cluster of ~1,000 identical DNA templates are “grown” on the surface. After amplification, one of the two primers is released from the slide, resulting in only one of the two amplicon strands remaining in each cluster. Sequencing is achieved by cyclic polymerase-based incorporation of fluorescently labeled nucleotides, starting from a universal sequencing primer. Reversible terminators ensure that only one nucleotide is incorporated in each cycle. After removal of unincorporated nucleotides, the array is imaged in four colors, allowing identification of the identity of one base in each cluster. The reversible terminator group is subsequently cleaved off the clusters, and the process is repeated by extending the previous synthesis products, allowing interrogation of the next base. A commercial

implementation of this method by Solexa can generate short (25-50 bp) sequence reads, at a cost per read that is also roughly two orders of magnitude lower than for capillary Sanger sequencing.

Sequencing of single molecule features by synthesis. The ultimate realization of high-density cyclic array sequencing is interrogation of single molecule DNA features. In one version of this approach, adapter flanked DNA templates are first immobilized on a quartz slide. Fluorescently labeled universal primers are then hybridized to the templates and imaged to identify the location of each DNA feature. In the subsequent cyclic steps, fluorescently labeled nucleotides are incorporated by a polymerase, imaged, and then inactivated by photobleaching. Observations of single molecule fluorescence are made with a conventional microscope equipped with total internal reflection illumination, which reduces background fluorescence. In addition, single-pair fluorescence resonance energy transfer (spFRET) is used to minimize noise. The first incorporated nucleotide is labeled with a donor fluorophore (Cy3) and the subsequent nucleotides by an acceptor (Cy5). Excitation of the donor leads to fluorescence from acceptors within a limited spatial range that avoids unincorporated nucleotides on the slide. Photobleaching of the incorporated acceptors does not affect the donor fluorophore, rendering it active for the next cycle. An implementation of this method by Helicos²¹⁶ is currently being adapted for commercialization.

The DNA sequencer as a general-purpose laboratory tool. The launch and continued improvement of commercial and academic next-generation sequencing instruments have generated tremendous excitement in the DNA sequencing field. New applications that were effectively out of reach with traditional dideoxy sequencers are continually being developed and explored. Importantly, the promise of next-generation sequencing technologies extends far beyond deeper and more comprehensive cataloguing of genomes and genetic variation. Automated sequencing can in principle be used to capture the result of any assay for which the end product is a collection of DNA molecules that encode its outcome. It does not matter whether the DNA is naturally occurring or the product of a designed enzymatic process. As long as sequence reads are long enough to capture the information encoded in each DNA molecule, and throughput is sufficient to sample the collection to the required depth, a sequencing instrument can provide a high resolution, digital measurement of the assay result.

Contributions of this thesis

We begin by describing a comparative analysis of the chimpanzee and human genomes (Chapter 2). The chimpanzee is the first non-human primate to be sequenced. We generate a nearly complete catalog of genetic changes that have occurred since our last common ancestor, and use this catalog to explore the magnitude and variation in mutational and selective forces acting on our genomes. We show that patterns of evolution in human and chimpanzee protein-coding genes are highly correlated, and dominated by the fixation of neutral and slightly deleterious alleles, as predicted by the nearly neutral theory of evolution²¹⁷. We find evidence of human-specific positive selection in genes encoding transcription factors and in regions devoid of protein-coding genes, which is consistent with the hypothesis that the accelerated anatomical evolution in the human lineage was driven by changes in gene regulation¹³².

We next describe a comparative analysis of the dog genome with those of human and mouse (Chapter 3). We characterize patterns of gene and genome evolution in eutherian (placental) mammals, and provide strong support for the previous estimate⁶⁰ that 5% of the human genome has evolved under purifying selection since the last common ancestor of eutherian mammals, and that the majority of these sequences do not encode proteins. We show that the majority of highly conserved non-coding elements are clustered in large gene deserts surrounding a few hundred genes encoding developmental transcription factors, signaling molecules and axon guidance receptors. This suggests a model of mammalian genome organization where a large fraction of the highly conserved regulatory elements control a small set of key regulatory genes.

We next describe a comparative analysis of the opossum genome with those of human, mouse and dog (Chapter 4). As the first metatherian (marsupial) species to be sequenced, the opossum provides a much closer outgroup to the eutherian mammals than previously sequenced vertebrates. Our analysis reveals a sharp difference in evolutionary innovation between protein-coding and non-coding elements. Lineage-specific differences in protein-coding gene content are rare and appear to be largely due to diversification and rapid turnover in gene families involved in environmental interactions. By contrast, one-fifth of the non-coding elements conserved in eutherian mammals are recent inventions that postdate the split from metatherians. A substantial proportion of these recent elements arose from sequence inserted by transposable elements, including tens of thousands of elements surrounding key regulatory genes, pointing to transposons as a major creative force in the evolution of mammalian gene regulation.

In order to uncover which bits of genomic information are utilized in a particular cell state, and to gain clues about their functions, we turn to biochemical analysis of proteins that interact with

and package DNA in the cell nucleus. We first develop a method for identifying proteins that recognize specific DNA sequences *in vitro* using affinity capture and mass spectrometry. Using this method, we discover sequence- and context-specific recruitment of proteins involved in chromatin remodeling to highly conserved non-coding elements (Chapter 5). We next characterize the chromatin state of selected loci enriched for highly conserved non-coding elements and key regulatory genes in pluripotent and differentiated cells using chromatin immunoprecipitation and microarray (ChIP-Chip) technology (Chapter 6). We describe a novel chromatin pattern, termed “bivalent domains”, which is associated with regulatory genes expressed at low levels and might contribute to maintaining embryonic stem (ES) cells in a pluripotent state. Moreover, we find a strong correlation between genome sequence and chromatin state in undifferentiated cells that become notably weaker upon differentiation. These results highlight the importance of DNA sequence in establishing the epigenomic landscape during development.

To facilitate more comprehensive analyses of chromatin state, we next develop methods for mapping protein-DNA interactions using chromatin immunoprecipitation and single-molecule based sequencing technology (ChIP-Seq; Chapter 7). We apply these methods to generate the first genome-wide maps of histone methylation patterns in undifferentiated and differentiated mouse cells. We identify three broad categories of promoters in the mouse genome based on their sequence composition and histone methylation patterns in ES cells, and show that lineage commitment is accompanied by characteristic chromatin changes that parallel changes in gene expression and transcriptional competence. We also demonstrate the potential for using characteristic histone methylation patterns to identify active genomic elements such as promoters, transcribed protein-coding and non-coding genes, transposons, imprinting control regions and other distal regulatory elements.

To complement our studies of protein-DNA interactions, we next develop efficient methods for mapping DNA methylation across mammalian genomes using high-throughput reduced representation bisulfite sequencing (RRBS; Chapter 8). We show that DNA methylation is a dynamic chromatin modification that is closely correlated with histone methylation patterns and that undergoes extensive change during cellular differentiation, particularly in regulatory elements outside of core promoters. We also show that the majority of differences in DNA and histone methylation between cell types are found outside of promoters and protein coding-sequences, which provides additional support for the notion that the majority of functional elements in mammalian genomes do not encode proteins⁶⁰. Finally, we show that while developmentally regulated methylation of promoters is largely limited to imprinted and germ line-specific genes, extended culture of cells *in*

vitro can induce aberrant methylation of specific regulatory genes in a pattern similar to that observed in some primary tumors.

Finally, we explore direct reprogramming of somatic cells to an undifferentiated state using ectopic expression of transcription factors²¹⁸⁻²²⁰. We generate histone methylation, DNA methylation and gene expression profiles of mouse cells at different stages of this reprogramming process (Chapter 9). Our data suggest that cells may become trapped in partially reprogrammed states owing to incomplete repression of key regulatory genes, and that DNA de-methylation is an inefficient step in the transition to pluripotency. We show that inhibition of DNA methyltransferase can improve the efficiency of the reprogramming process, which supports the notion that chromatin modifications contribute to the stabilization of differentiated cell states.

Our results provide strong experimental support for some previous theories about genome evolution and function, and generate several novel hypotheses. In addition, our methods provide a technological framework for characterization of chromatin state across diverse mammalian cell populations. Insights from our studies are already guiding multiple large-scale efforts to comprehensively annotate the human genome, as well as the genomes of model organisms, and to further explore the relationship between chromatin and cellular state in development and disease.

References

1. Lander, E.S. *et al.*, Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860-921 (2001).
2. Consortium, I.H.G.S., Finishing the euchromatic sequence of the human genome. *Nature* 431 (7011), 931-945 (2004).
3. Alberts, B. *et al.*, *Molecular Biology of the Cell*, 4th ed. (Garland Science, New York, 2002).
4. Lodish, H. *et al.*, *Molecular Cell Biology*, 5th ed. (W. H. Freeman and Co., New York, 2003).
5. Gilbert, S.F., *Developmental Biology*, 8th ed. (Sinauer, Sunderland, MA, 2006).
6. Allis, C.D., Jenuwein, T., & Reinberg, D., *Epigenetics*, 1st ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2007).
7. Barton, N.H., Briggs, D.E.G., Eisen, J.A., Goldstein, D.B., & Patel, N.H., *Evolution*, 1st ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2007).
8. Li, W.-H., *Molecular Evolution*, 1st ed. (Sinauer, Sunderland, MA, 1997).
9. Rice, S.H., *Evolutionary Theory*, 1st ed. (Sinauer, Sunderland, Ma, 2004).
10. Gibson, G. & Muse, S.V., *A Primer of Genome Science*, 2nd ed. (Sinauer, Sunderland, MA, 2004).
11. Maxam, A.M. & Gilbert, W., A new method for sequencing DNA. *Proc Natl Acad Sci U S A* 74 (2), 560-564 (1977).
12. Sanger, F. & Coulson, A.R., A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94 (3), 441-448 (1975).
13. Sanger, F., Nicklen, S., & Coulson, A.R., DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74 (12), 5463-5467 (1977).
14. Prober, J.M. *et al.*, A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238 (4825), 336-341 (1987).
15. Tabor, S. & Richardson, C.C., A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc Natl Acad Sci U S A* 92 (14), 6339-6343 (1995).
16. Dovichi, N.J., DNA sequencing by capillary electrophoresis. *Electrophoresis* 18 (12-13), 2393-2399 (1997).
17. Meldrum, D., Automation for genomics, part two: sequencers, microarrays, and future trends. *Genome Res* 10 (9), 1288-1303 (2000).
18. Meldrum, D., Automation for genomics, part one: preparation for sequencing. *Genome Res* 10 (8), 1081-1092 (2000).
19. Anderson, S., Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res* 9 (13), 3015-3027 (1981).
20. Deininger, P.L., Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Anal Biochem* 129 (1), 216-223 (1983).
21. Gardner, R.C. *et al.*, The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res* 9 (12), 2871-2888 (1981).
22. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., & Petersen, G.B., Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* 162 (4), 729-773 (1982).

23. Lander, E.S. & Waterman, M.S., Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2 (3), 231-239 (1988).
24. Sanger, F. *et al.*, Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265 (5596), 687-695 (1977).
25. Fleischmann, R.D. *et al.*, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269 (5223), 496-512 (1995).
26. Goffeau, A. *et al.*, Life with 6000 genes. *Science* 274 (5287), 546, 563-547 (1996).
27. *C. elegans* Sequencing Consortium, Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282 (5396), 2012-2018 (1998).
28. Adams, M.D. *et al.*, The genome sequence of *Drosophila melanogaster*. *Science* 287 (5461), 2185-2195 (2000).
29. Myers, G., Whole-genome DNA sequencing. *Comput Sci Eng* 1, 33-43 (1999).
30. Venter, J.C. *et al.*, The Sequence of the Human Genome. *Science* 291 (5507), 1304-1351 (2001).
31. Jaffe, D.B. *et al.*, Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13 (1), 91-96 (2003).
32. Batzoglou, S. *et al.*, ARACHNE: a whole-genome shotgun assembler. *Genome Res* 12 (1), 177-189 (2002).
33. Huang, X., Wang, J., Aluru, S., Yang, S.P., & Hillier, L., PCAP: a whole-genome assembly program. *Genome Res* 13 (9), 2164-2170 (2003).
34. Hirt, H. *et al.*, The human growth hormone gene locus: structure, evolution, and allelic variations. *DNA* 6 (1), 59-70 (1987).
35. Martin-Gallardo, A. *et al.*, Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. *Nat Genet* 1 (1), 34-39 (1992).
36. Rowen, L., Koop, B.F., & Hood, L., The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* 272 (5269), 1755-1762 (1996).
37. Hattori, M. *et al.*, The DNA sequence of human chromosome 21. *Nature* 405 (6784), 311-319 (2000).
38. Dunham, I. *et al.*, The DNA sequence of human chromosome 22. *Nature* 402 (6761), 489-495 (1999).
39. Thiery, J.P., Macaya, G., & Bernardi, G., An analysis of eukaryotic genomes by density gradient centrifugation. *J Mol Biol* 108 (1), 219-235 (1976).
40. Filipinski, J., Thiery, J.P., & Bernardi, G., An analysis of the bovine genome by Cs₂SO₄-Ag density gradient centrifugation. *J Mol Biol* 80 (1), 177-197 (1973).
41. Hurst, L.D. & Eyre-Walker, A., Evolutionary genomics: reading the bands. *Bioessays* 22 (2), 105-107 (2000).
42. Zoubak, S., Clay, O., & Bernardi, G., The gene distribution of the human genome. *Gene* 174 (1), 95-102 (1996).
43. Gardiner, K., Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends Genet* 12 (12), 519-524 (1996).
44. Eyre-Walker, A. & Hurst, L.D., The evolution of isochores. *Nat Rev Genet* 2 (7), 549-555 (2001).
45. Bernardi, G., Isochores and the evolutionary genomics of vertebrates. *Gene* 241 (1), 3-17 (2000).
46. Wolfe, K.H., Sharp, P.M., & Li, W.H., Mutation rates differ among regions of the mammalian genome. *Nature* 337 (6204), 283-285 (1989).

47. Fryxell, K.J. & Zuckerkandl, E., Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol* 17 (9), 1371-1383 (2000).
48. Brown, T.C. & Jiricny, J., Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* 54 (5), 705-711 (1988).
49. Fullerton, S.M., Bernardo Carvalho, A., & Clark, A.G., Local rates of recombination are positively correlated with GC content in the human genome. *Mol Biol Evol* 18 (6), 1139-1142 (2001).
50. Eyre-Walker, A., Recombination and mammalian genome evolution. *Proc Biol Sci* 252 (1335), 237-243 (1993).
51. Bird, A., Taggart, M., Frommer, M., Miller, O.J., & Macleod, D., A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* 40 (1), 91-99 (1985).
52. Bird, A.P., CpG-rich islands and the function of DNA methylation. *Nature* 321 (6067), 209-213 (1986).
53. Gardiner-Garden, M. & Frommer, M., CpG islands in vertebrate genomes. *J Mol Biol* 196 (2), 261-282 (1987).
54. Larsen, F., Gundersen, G., Lopez, R., & Prydz, H., CpG islands as gene markers in the human genome. *Genomics* 13 (4), 1095-1107 (1992).
55. Takai, D. & Jones, P.A., Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* 99 (6), 3740-3745 (2002).
56. Smit, A.F., Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 9 (6), 657-663 (1999).
57. Prak, E.T. & Kazazian, H.H., Jr., Mobile elements and the human genome. *Nat Rev Genet* 1 (2), 134-144 (2000).
58. Okada, N., Hamada, M., Ogiwara, I., & Ohshima, K., SINEs and LINEs share common 3' sequences: a review. *Gene* 205 (1-2), 229-243 (1997).
59. Dewannieux, M., Esnault, C., & Heidmann, T., LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* 35 (1), 41-48 (2003).
60. Waterston, R.H. *et al.*, Initial sequencing and comparative analysis of the mouse genome. *Nature* 420 (6915), 520-562 (2002).
61. McClintock, B., The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* 36 (6), 344-355 (1950).
62. McClintock, B., Induction of Instability at Selected Loci in Maize. *Genetics* 38 (6), 579-599 (1953).
63. Waring, M. & Britten, R.J., Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. *Science* 154 (750), 791-794 (1966).
64. Britten, R.J. & Kohne, D.E., Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* 161 (841), 529-540 (1968).
65. Britten, R.J., Mobile elements inserted in the distant past have taken on important functions. *Gene* 205 (1-2), 177-182 (1997).
66. Robins, D.M. & Samuelson, L.C., Retrotransposons and the evolution of mammalian gene expression. *Genetica* 86 (1-3), 191-201 (1992).
67. Kazazian, H.H., Jr., Mobile elements: drivers of genome evolution. *Science* 303 (5664), 1626-1632 (2004).
68. Bowen, N.J. & Jordan, I.K., Transposable elements and the evolution of eukaryotic complexity. *Curr Issues Mol Biol* 4 (3), 65-76 (2002).

69. Deininger, P.L., Moran, J.V., Batzer, M.A., & Kazazian, H.H., Jr., Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* 13 (6), 651-658 (2003).
70. Jaenisch, R., Endogenous retroviruses. *Cell* 32 (1), 5-6 (1983).
71. Syvanen, M., The evolutionary implications of mobile genetic elements. *Annu Rev Genet* 18, 271-293 (1984).
72. Finnegan, D.J., Eukaryotic transposable elements and genome evolution. *Trends Genet* 5 (4), 103-107 (1989).
73. Myerowitz, R. & Hogikyan, N.D., A deletion involving Alu sequences in the beta-hexosaminidase alpha-chain gene of French Canadians with Tay-Sachs disease. *J Biol Chem* 262 (32), 15396-15399 (1987).
74. Lehrman, M.A., Russell, D.W., Goldstein, J.L., & Brown, M.S., Exon-Alu recombination deletes 5 kilobases from the low density lipoprotein receptor gene, producing a null phenotype in familial hypercholesterolemia. *Proc Natl Acad Sci U S A* 83 (11), 3679-3683 (1986).
75. Calabretta, B., Robberson, D.L., Barrera-Saldana, H.A., Lambrou, T.P., & Saunders, G.F., Genome instability in a region of human DNA enriched in Alu repeat sequences. *Nature* 296 (5854), 219-225 (1982).
76. Rouyer, F., Simmler, M.C., Page, D.C., & Weissenbach, J., A sex chromosome rearrangement in a human XX male caused by Alu-Alu recombination. *Cell* 51 (3), 417-425 (1987).
77. Brosius, J., Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* 107 (1-3), 209-238 (1999).
78. Wang, P.J., X chromosomes, retrogenes and their role in male reproduction. *Trends Endocrinol Metab* 15 (2), 79-83 (2004).
79. Britten, R.J., Coding sequences of functioning human genes derived entirely from mobile element sequences. *Proc Natl Acad Sci U S A* 101 (48), 16825-16830 (2004).
80. Britten, R.J., Cases of ancient mobile element DNA insertions that now affect gene regulation. *Mol Phylogenet Evol* 5 (1), 13-17 (1996).
81. Ji, Y., Eichler, E.E., Schwartz, S., & Nicholls, R.D., Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res* 10 (5), 597-610 (2000).
82. Eichler, E.E., Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res* 8 (8), 758-762 (1998).
83. Mazzarella, R. & Schlessinger, D., Pathological consequences of sequence duplications in the human genome. *Genome Res* 8 (10), 1007-1021 (1998).
84. Horvath, J.E., Bailey, J.A., Locke, D.P., & Eichler, E.E., Lessons from the human genome: transitions between euchromatin and heterochromatin. *Hum Mol Genet* 10 (20), 2215-2223 (2001).
85. Amos-Landgraf, J.M. *et al.*, Chromosome breakage in the Prader-Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. *Am J Hum Genet* 65 (2), 370-386 (1999).
86. Christian, S.L., Fantes, J.A., Mewborn, S.K., Huang, B., & Ledbetter, D.H., Large genomic duplicons map to sites of instability in the Prader-Willi/Angelman syndrome chromosome region (15q11-q13). *Hum Mol Genet* 8 (6), 1025-1037 (1999).
87. Edelmann, L., Pandita, R.K., & Morrow, B.E., Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome. *Am J Hum Genet* 64 (4), 1076-1086 (1999).
88. Shaikh, T.H. *et al.*, Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* 9 (4), 489-501 (2000).

89. Johnson, M.E. *et al.*, Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413 (6855), 514-519 (2001).
90. Lewin, B., *Gene Expression*. (wiley, New York, 1980).
91. Antequera, F. & Bird, A., Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* 90 (24), 11995-11999 (1993).
92. Fields, C., Adams, M.D., White, O., & Venter, J.C., How many genes in the human genome? *Nat Genet* 7 (3), 345-346 (1994).
93. Liang, F. *et al.*, Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet* 25 (2), 239-240 (2000).
94. Ewing, B. & Green, P., Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* 25 (2), 232-234 (2000).
95. Storz, G., An expanding universe of noncoding RNAs. *Science* 296 (5571), 1260-1263 (2002).
96. Szymanski, M., Erdmann, V.A., & Barciszewski, J., Noncoding regulatory RNAs database. *Nucleic Acids Res* 31 (1), 429-431 (2003).
97. Bartel, D.P., MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116 (2), 281-297 (2004).
98. Brown, C.J. *et al.*, A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* 349 (6304), 38-44 (1991).
99. Kapranov, P. *et al.*, Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296 (5569), 916-919 (2002).
100. Lettice, L.A. *et al.*, A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 12 (14), 1725-1735 (2003).
101. Pribnow, D., Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc Natl Acad Sci U S A* 72 (3), 784-788 (1975).
102. Schaller, H., Gray, C., & Herrmann, K., Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd. *Proc Natl Acad Sci U S A* 72 (2), 737-741 (1975).
103. Tagle, D.A. *et al.*, Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* 203 (2), 439-455 (1988).
104. den Dunnen, J.T., van Neck, J.W., Cremers, F.P., Lubsen, N.H., & Schoenmakers, J.G., Nucleotide sequence of the rat gamma-crystallin gene region and comparison with an orthologous human region. *Gene* 78 (2), 201-213 (1989).
105. Shehee, W.R. *et al.*, Nucleotide sequence of the BALB/c mouse beta-globin complex. *J Mol Biol* 205 (1), 41-62 (1989).
106. Koop, B.F. & Hood, L., Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat Genet* 7 (1), 48-53 (1994).
107. Oeltjen, J.C. *et al.*, Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res* 7 (4), 315-329 (1997).
108. Loots, G.G. *et al.*, Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288 (5463), 136-140 (2000).
109. Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., & Lander, E.S., Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res* 10 (7), 950-958 (2000).
110. Hardison, R.C., Oeltjen, J., & Miller, W., Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* 7 (10), 959-966 (1997).

111. Pennacchio, L.A. & Rubin, E.M., Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* 2 (2), 100-109 (2001).
112. Cliften, P. *et al.*, Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301 (5629), 71-76 (2003).
113. Cliften, P.F. *et al.*, Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res* 11 (7), 1175-1186 (2001).
114. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., & Lander, E.S., Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423 (6937), 241-254 (2003).
115. Kellis, M., Birren, B.W., & Lander, E.S., Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428 (6983), 617-624 (2004).
116. Stein, L.D. *et al.*, The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol* 1 (2), E45 (2003).
117. Gibbs, R.A. *et al.*, Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428 (6982), 493-521 (2004).
118. Adkins, R.M., Gelke, E.L., Rowe, D., & Honeycutt, R.L., Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. *Mol Biol Evol* 18 (5), 777-791 (2001).
119. Springer, M.S., Murphy, W.J., Eizirik, E., & O'Brien, S.J., Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc Natl Acad Sci U S A* 100 (3), 1056-1061 (2003).
120. Clark, F.H., The Inheritance and Linkage Relations of a New Recessive Spotting in the House Mouse. *Genetics* 19 (5), 365-393 (1934).
121. Castle, W.W., Observations of the occurrence of linkage in rats and mice. *Car Inst Wash Pub* 288, 29-36 (1919).
122. Nadeau, J.H. & Taylor, B.A., Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci U S A* 81 (3), 814-818 (1984).
123. Stanyon, R., Stone, G., Garcia, M., & Froenicke, L., Reciprocal chromosome painting shows that squirrels, unlike murid rodents, have a highly conserved genome organization. *Genomics* 82 (2), 245-249 (2003).
124. Ohno, S., Ancient linkage groups and frozen accidents. *Nature* 244 (5414), 259-262 (1973).
125. Bourque, G., Pevzner, P.A., & Tesler, G., Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res* 14 (4), 507-516 (2004).
126. Pevzner, P. & Tesler, G., Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A* 100 (13), 7672-7677 (2003).
127. Okazaki, Y. *et al.*, Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420 (6915), 563-573 (2002).
128. Ohno, S., *Evolution by Gene Duplication*. (Springer, Berlin, 1970).
129. Carroll, S.B., Endless forms: the evolution of gene regulation and morphological diversity. *Cell* 101 (6), 577-580 (2000).
130. Davidson, E.H., *Genomic Regulatory Systems: Development and Evolution*. (Academic Press, 2001).
131. Cherty, L.M., Case, S.M., & Wilson, A.C., Frog perspective on the morphological difference between humans and chimpanzees. *Science* 200 (4338), 209-211 (1978).
132. King, M.C. & Wilson, A.C., Evolution at two levels in humans and chimpanzees. *Science* 188 (4184), 107-116 (1975).

133. Charlesworth, B., The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res* 63 (3), 213-227 (1994).
134. Nachman, M.W., Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet* 17 (9), 481-485 (2001).
135. Francino, M.P. & Ochman, H., Strand asymmetries in DNA evolution. *Trends Genet* 13 (6), 240-245 (1997).
136. Dermitzakis, E.T. *et al.*, Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res* 14 (5), 852-859 (2004).
137. Bejerano, G. *et al.*, Ultraconserved elements in the human genome. *Science* 304 (5675), 1321-1325 (2004).
138. International Chicken Genome Sequencing Consortium, Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432 (7018), 695-716 (2004).
139. Jaillon, O. *et al.*, Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431 (7011), 946-957 (2004).
140. Boffelli, D., Nobrega, M.A., & Rubin, E.M., Comparative genomics at the vertebrate extremes. *Nat Rev Genet* 5 (6), 456-465 (2004).
141. Nobrega, M.A., Ovcharenko, I., Afzal, V., & Rubin, E.M., Scanning human gene deserts for long-range enhancers. *Science* 302 (5644), 413 (2003).
142. Smith, N.G., Brandstrom, M., & Ellegren, H., Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* 84 (5), 806-813 (2004).
143. Hewish, D.R. & Burgoyne, L.A., Chromatin sub-structure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochem Biophys Res Commun* 52 (2), 504-510 (1973).
144. Olins, A.L. & Olins, D.E., Spheroid chromatin units (v bodies). *Science* 183 (122), 330-332 (1974).
145. Davey, C.A., Sargent, D.F., Luger, K., Maeder, A.W., & Richmond, T.J., Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J Mol Biol* 319 (5), 1097-1113 (2002).
146. Wu, C., The 5' ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNase I. *Nature* 286 (5776), 854-860 (1980).
147. Elgin, S.C., Anatomy of hypersensitive sites. *Nature* 309 (5965), 213-214 (1984).
148. Gross, D.S. & Garrard, W.T., Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 57, 159-197 (1988).
149. Tuan, D., Solomon, W., Li, Q., & London, I.M., The "beta-like-globin" gene domain in human erythroid cells. *Proc Natl Acad Sci U S A* 82 (19), 6384-6388 (1985).
150. Forrester, W.C., Thompson, C., Elder, J.T., & Groudine, M., A developmentally stable chromatin structure in the human beta-globin gene cluster. *Proc Natl Acad Sci U S A* 83 (5), 1359-1363 (1986).
151. Han, M. & Grunstein, M., Nucleosome loss activates yeast downstream promoters in vivo. *Cell* 55 (6), 1137-1145 (1988).
152. Hirschhorn, J.N., Brown, S.A., Clark, C.D., & Winston, F., Evidence that SNF2/SWI2 and SNF5 activate transcription in yeast by altering chromatin structure. *Genes Dev* 6 (12A), 2288-2298 (1992).
153. Tsukiyama, T., Becker, P.B., & Wu, C., ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. *Nature* 367 (6463), 525-532 (1994).
154. Kwon, H., Imbalzano, A.N., Khavari, P.A., Kingston, R.E., & Green, M.R., Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex. *Nature* 370 (6489), 477-481 (1994).

155. Felsenfeld, G., Chromatin unfolds. *Cell* 86 (1), 13-19 (1996).
156. van Holde, K.E., in *Chromatin*, edited by A. Rich (Springer, New York, 1988), pp. 111-148.
157. Kayne, P.S. *et al.*, Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. *Cell* 55 (1), 27-39 (1988).
158. Durrin, L.K., Mann, R.K., Kayne, P.S., & Grunstein, M., Yeast histone H4 N-terminal sequence is required for promoter activation in vivo. *Cell* 65 (6), 1023-1031 (1991).
159. Hecht, A., Laroche, T., Strahl-Bolsinger, S., Gasser, S.M., & Grunstein, M., Histone H3 and H4 N-termini interact with SIR3 and SIR4 proteins: a molecular model for the formation of heterochromatin in yeast. *Cell* 80 (4), 583-592 (1995).
160. Dhalluin, C. *et al.*, Structure and ligand of a histone acetyltransferase bromodomain. *Nature* 399 (6735), 491-496 (1999).
161. Bannister, A.J. *et al.*, Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* 410 (6824), 120-124 (2001).
162. Nakayama, J., Rice, J.C., Strahl, B.D., Allis, C.D., & Grewal, S.I., Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* 292 (5514), 110-113 (2001).
163. Ng, H.H., Robert, F., Young, R.A., & Struhl, K., Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell* 11 (3), 709-719 (2003).
164. Bernstein, B.E. *et al.*, Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci U S A* 99 (13), 8695-8700 (2002).
165. Krogan, N.J. *et al.*, The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. *Mol Cell* 11 (3), 721-729 (2003).
166. Cao, R. *et al.*, Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* 298 (5595), 1039-1043 (2002).
167. Czermin, B. *et al.*, Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell* 111 (2), 185-196 (2002).
168. Lavigne, M., Francis, N.J., King, I.F., & Kingston, R.E., Propagation of silencing; recruitment and repression of naive chromatin in trans by polycomb repressed chromatin. *Mol Cell* 13 (3), 415-425 (2004).
169. Li, B., Howe, L., Anderson, S., Yates, J.R., 3rd, & Workman, J.L., The Set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II. *J Biol Chem* 278 (11), 8897-8903 (2003).
170. Nishioka, K. *et al.*, PR-Set7 is a nucleosome-specific methyltransferase that modifies lysine 20 of histone H4 and is associated with silent chromatin. *Mol Cell* 9 (6), 1201-1213 (2002).
171. Jenuwein, T. & Allis, C.D., Translating the histone code. *Science* 293 (5532), 1074-1080 (2001).
172. Riggs, A.D., X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet* 14 (1), 9-25 (1975).
173. Holliday, R. & Pugh, J.E., DNA modification mechanisms and gene activity during development. *Science* 187 (4173), 226-232 (1975).
174. Watt, F. & Molloy, P.L., Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev* 2 (9), 1136-1143 (1988).
175. Hark, A.T. *et al.*, CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* 405 (6785), 486-489 (2000).

176. Tate, P.H. & Bird, A.P., Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr Opin Genet Dev* 3 (2), 226-231 (1993).
177. Lewis, J.D. *et al.*, Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* 69 (6), 905-914 (1992).
178. Hendrich, B. & Bird, A., Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol* 18 (11), 6538-6547 (1998).
179. Prokhortchouk, A. *et al.*, The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev* 15 (13), 1613-1618 (2001).
180. Boyes, J. & Bird, A., DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell* 64 (6), 1123-1134 (1991).
181. Jones, P.L. *et al.*, Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* 19 (2), 187-191 (1998).
182. Bestor, T., Laudano, A., Mattaliano, R., & Ingram, V., Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J Mol Biol* 203 (4), 971-983 (1988).
183. Li, E., Bestor, T.H., & Jaenisch, R., Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 69 (6), 915-926 (1992).
184. Lei, H. *et al.*, De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* 122 (10), 3195-3205 (1996).
185. Okano, M., Xie, S., & Li, E., Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* 19 (3), 219-220 (1998).
186. Taylor, S.M. & Jones, P.A., Multiple new phenotypes induced in 10T1/2 and 3T3 cells treated with 5-azacytidine. *Cell* 17 (4), 771-779 (1979).
187. Jones, P.A. & Taylor, S.M., Cellular differentiation, cytidine analogs and DNA methylation. *Cell* 20 (1), 85-93 (1980).
188. Jaenisch, R. & Bird, A., Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 33 Suppl, 245-254 (2003).
189. Hadorn, E., Transdetermination in cells. *Sci Am* 219 (5), 110-114 passim (1968).
190. Delage, Y., *La structure du protoplasma et les théories sur l'hérédité et les grands problèmes de la biologie générale*. (C. Reinwald, Paris, 1895).
191. Briggs, R. & King, T.J., Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs. *Proc Natl Acad Sci U S A* 38 (5), 455-463 (1952).
192. Gurdon, J.B. & Byrne, J.A., The first half-century of nuclear transplantation. *Proc Natl Acad Sci U S A* 100 (14), 8048-8052 (2003).
193. Wakayama, T., Perry, A.C., Zuccotti, M., Johnson, K.R., & Yanagimachi, R., Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei. *Nature* 394 (6691), 369-374 (1998).
194. Wilmut, I., Schnieke, A.E., McWhir, J., Kind, A.J., & Campbell, K.H., Viable offspring derived from fetal and adult mammalian cells. *Nature* 385 (6619), 810-813 (1997).
195. Hochedlinger, K. & Jaenisch, R., Monoclonal mice generated by nuclear transfer from mature B and T donor cells. *Nature* 415 (6875), 1035-1038 (2002).
196. Ringrose, L. & Paro, R., Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu Rev Genet* 38, 413-443 (2004).

197. Kennison, J.A., The Polycomb and trithorax group proteins of *Drosophila*: trans-regulators of homeotic gene function. *Annu Rev Genet* 29, 289-303 (1995).
198. Beisel, C., Imhof, A., Greene, J., Kremmer, E., & Sauer, F., Histone methylation by the *Drosophila* epigenetic transcriptional regulator Ash1. *Nature* 419 (6909), 857-862 (2002).
199. Horard, B., Tatout, C., Poux, S., & Pirrotta, V., Structure of a polycomb response element and in vitro binding of polycomb group complexes containing GAGA factor. *Mol Cell Biol* 20 (9), 3187-3197 (2000).
200. O'Carroll, D. *et al.*, The polycomb-group gene *Ezh2* is required for early mouse development. *Mol Cell Biol* 21 (13), 4330-4336 (2001).
201. Plath, K. *et al.*, Role of histone H3 lysine 27 methylation in X inactivation. *Science* 300 (5616), 131-135 (2003).
202. Valk-Lingbeek, M.E., Bruggeman, S.W., & van Lohuizen, M., Stem cells and cancer; the polycomb connection. *Cell* 118 (4), 409-418 (2004).
203. Braunstein, M., Rose, A.B., Holmes, S.G., Allis, C.D., & Broach, J.R., Transcriptional silencing in yeast is associated with reduced nucleosome acetylation. *Genes Dev* 7 (4), 592-604 (1993).
204. Solomon, M.J., Larsen, P.L., & Varshavsky, A., Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 53 (6), 937-947 (1988).
205. Hecht, A. & Grunstein, M., Mapping DNA interaction sites of chromosomal proteins using immunoprecipitation and polymerase chain reaction. *Methods Enzymol* 304, 399-414 (1999).
206. Frommer, M. *et al.*, A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* 89 (5), 1827-1831 (1992).
207. Kurdistani, S.K., Tavazoie, S., & Grunstein, M., Mapping global histone acetylation patterns to gene expression. *Cell* 117 (6), 721-733 (2004).
208. Robyr, D. *et al.*, Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell* 109 (4), 437-446 (2002).
209. Velculescu, V.E., Zhang, L., Vogelstein, B., & Kinzler, K.W., Serial analysis of gene expression. *Science* 270 (5235), 484-487 (1995).
210. Roh, T.Y., Ngau, W.C., Cui, K., Landsman, D., & Zhao, K., High-resolution genome-wide mapping of histone modifications. *Nat Biotechnol* 22 (8), 1013-1016 (2004).
211. Collins, F.S., Genome research: the next generation. *Cold Spring Harb Symp Quant Biol* 68, 49-54 (2003).
212. Tawfik, D.S. & Griffiths, A.D., Man-made cell-like compartments for molecular evolution. *Nat Biotechnol* 16 (7), 652-656 (1998).
213. Ronaghi, M., Uhlen, M., & Nyren, P., A sequencing method based on real-time pyrophosphate. *Science* 281 (5375), 363, 365 (1998).
214. Shendure, J. *et al.*, Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309 (5741), 1728-1732 (2005).
215. Fedurco, M., Romieu, A., Williams, S., Lawrence, I., & Turcatti, G., BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* 34 (3), e22 (2006).
216. Braslavsky, I., Hebert, B., Kartalov, E., & Quake, S.R., Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A* 100 (7), 3960-3964 (2003).
217. Ohta, T., Slightly deleterious mutant substitutions in evolution. *Nature* 246 (5428), 96-98 (1973).

218. Takahashi, K. & Yamanaka, S., Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126 (4), 663-676 (2006).
219. Okita, K., Ichisaka, T., & Yamanaka, S., Generation of germline-competent induced pluripotent stem cells. *Nature* 448 (7151), 313-317 (2007).
220. Wernig, M. *et al.*, In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* 448 (7151), 318-324 (2007).
221. Bird, A. Perceptions of epigenetics. *Nature* 447, 396-398 (2007).

[This page is intentionally left blank]

Chapter 2: The chimpanzee genome

In this chapter, we describe the first comprehensive comparative analysis of the human and chimpanzee genome sequences.

This work was first published as

The Chimpanzee Sequencing and Analysis Consortium (Mikkelsen, T. S. *et al.*). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).

This publication is attached as Appendix 1. Supplementary notes can be found at the end of the chapter. Supplementary data is available online from <http://www.nature.com/nature>

The text was co-authored with analysis section leaders LaDeana W. Hillier (Genome Sequencing and Assembly), Evan E. Eichler (Insertions and Deletions, Transposable Element Insertions, Large-scale rearrangements), Michael C. Zody (Human Population Genetics), with significant input from many members of the analysis consortium.

[This page is intentionally left blank]

We present a draft genome sequence of the common chimpanzee (*Pan troglodytes*). Through comparison with the human genome, we generate a largely complete catalog of the genetic differences that have accumulated since the human and chimpanzee species diverged from our common ancestor, constituting approximately 35 million single-nucleotide changes, 5 million insertions and deletions, and various chromosomal rearrangements. We use this catalog to explore the magnitude and regional variation of mutational forces shaping these two genomes, and the strength of positive and negative selection acting on their genes. In particular, we find that the patterns of evolution in human and chimpanzee protein coding genes are highly correlated, and dominated by the fixation of neutral and slightly deleterious alleles. We also use the chimpanzee genome as an outgroup to investigate human population genetics and to identify signatures of selective sweeps in recent human evolution.

More than a century ago Darwin¹ and Huxley² posited that humans share recent common ancestors with the African great apes. Modern molecular studies have spectacularly confirmed this prediction and have refined the relationships, showing that the common chimpanzee (*Pan troglodytes*) and bonobo (*Pan paniscus* or so-called pygmy chimpanzee) are our closest living evolutionary relatives³. Chimpanzees are thus especially suited to teach us about ourselves, both in terms of their similarities and differences with human. For example, Goodall's pioneering studies on the common chimpanzee revealed startling behavioral similarities such as tool use and group aggression^{4,5}. By contrast, other features are obviously specific to human including habitual bipedality, a greatly enlarged brain and complex language⁵. Important similarities and differences have also been noted for the incidence and severity of several major human diseases⁶.

Genome comparisons of human and chimpanzee can help to reveal the molecular basis for these traits as well as the evolutionary forces that have molded our species, including underlying mutational processes and selective constraints. Early studies sought to draw inferences from sets of a few dozen genes⁷⁻⁹, while recent studies have examined larger datasets such as protein-coding exons¹⁰, random genomic sequences^{11,12}, and an entire chimpanzee chromosome¹³.

Here, we report a draft sequence of the genome of the common chimpanzee, and undertake comparative analyses with the human genome. This comparison differs fundamentally from recent comparative genomic studies of mouse, rat, chicken and fish¹⁴⁻¹⁷. Because the latter species have diverged substantially from the human lineage, the focus in such studies is on accurate alignment of the genomes and recognition of regions of unusually high evolutionary conservation to pinpoint functional elements. Because the chimpanzee lies at such a short evolutionary distance, nearly all of the bases are identical by descent and sequences can be readily aligned except in recently derived,

large repetitive regions. The focus thus turns to differences rather than similarities. An observed difference at a site nearly always represents a single event, not multiple independent changes over time. Most of the differences reflect random genetic drift, and thus they hold extensive information about mutational processes and negative selection that can be readily mined with current analytical techniques. Hidden among the differences is a minority of functionally important changes that underlie the phenotypic differences between the two species. Our ability to distinguish such sites is currently quite limited, but the catalog of human-chimpanzee differences opens this issue to systematic investigation for the first time. We would also hope that, in elaborating the few differences that separate the two species, we will increase pressure to save chimpanzees and other great apes in the wild.

Our results confirm many earlier observations, but notably challenge some previous claims based on more limited data. The genome-wide data also allow some questions to be addressed for the first time. (Here and throughout, we refer to chimpanzee-human comparison as representing hominids and mouse-rat comparison as representing murids. Of course, each pair covers only a subset of the clade). The key findings include:

- Single-nucleotide substitutions occur at a mean rate of 1.23% between copies of the human and chimpanzee genome, with 1.06% or less corresponding to fixed divergence between the species.

- Regional variation in nucleotide substitution rates is conserved between the hominid and murid genomes, but rates in subtelomeric regions are disproportionately elevated in the hominids.

- Substitutions at CpG dinucleotides, which constitute a quarter of all observed substitutions, occur at more similar rates in the male and female germ-lines than non-CpG substitutions.

- Insertion and deletion (indel) events are fewer in number than single-nucleotide substitutions, but result in ~1.5% of the euchromatic sequence in each species being lineage-specific.

- There are notable differences in the rate of transposable element insertions: SINEs have been three-fold more active in humans, while chimpanzees have acquired two new families of retroviral elements.

- Orthologous proteins in human and chimpanzee are extremely similar, with ~29% being identical and the typical ortholog differing by only two amino acids, one per lineage.

- The normalized rates of amino acid altering substitutions in the hominid lineages are elevated relative to the murid lineages, but close to that seen for common human polymorphisms,

implying that positive selection during hominid evolution accounts for a smaller fraction of protein divergence than suggested in some previous reports.

- The substitution rate at silent sites in exons is lower than the rate at nearby intronic sites, consistent with weak purifying selection on silent sites in mammals.

- Analysis of the pattern of human diversity relative to hominid divergence identifies several loci as potential candidates for strong selective sweeps in recent human history.

In this paper, we begin with information about the generation, assembly and evaluation of the draft genome sequence. We then explore overall genome evolution, with the aim of understanding mutational processes at work in the human genome. We next focus on the evolution of protein-coding genes, with the aim of characterizing the nature of selection. Finally, we briefly discuss initial insights into human population genetics.

In recognition of its strong community support, we will refer to chimpanzee chromosomes using the orthologous numbering nomenclature proposed by McConkey¹⁸, which renumber the chromosomes of the great apes from the ICSN (1978) standard to directly correspond to their human orthologs, using the terms 2A and 2B for the two ape chromosomes corresponding to human chromosome 2.

Genome Sequencing and Assembly

We sequenced the genome of a single male chimpanzee (“Clint”; Yerkes pedigree number C0471), a captive-born descendant of chimpanzees from the West Africa sub-species (*Pan troglodytes verus*), using a whole-genome shotgun (WGS) approach^{19,20}. The data were assembled using both the PCAP and ARACHNE programs^{21,22} (see Supplementary Information). The former was a *de novo* assembly, while the latter made limited use of human genome sequence (NCBI build 34)^{23,24} to facilitate and confirm contig linking. The ARACHNE assembly has slightly greater continuity (Table 1) and was used for analysis in this paper. The draft genome assembly, generated from ~3.6-fold sequence redundancy of the autosomes, and ~1.8-fold redundancy of both sex chromosomes, covers ~94% of the chimpanzee genome with >98% of the sequence in high-quality bases. 50% of the sequence (N50) is contained in contigs of length greater than 15.7 kb and supercontigs of length greater than 8.6 Mb. The assembly represents a consensus of two haplotypes, with one allele from each heterozygous position arbitrarily represented in the sequence.

Table 1: Chimpanzee Assembly Statistics

Assembler	PCAP	ARACHNE
Major contigs ¹	400,289	361,782
Contig length (N50) ²	13.3 Kb	15.7 Kb
Supercontigs	67,734	37,846
Supercontig length (N50)	2.3 Mb	8.6 Mb
Sequence redundancy: all bases (Q20)	5.0x (3.6x)	4.3x (3.6x)
Physical redundancy	20.7	19.8
Consensus bases	2.7 Gb	2.7 Gb

¹Contigs > 1 kb. ²N50 length is the size x such that 50% of the assembly is in units of length at least x.

Assessment of Quality and Coverage

The chimpanzee genome assembly was subjected to rigorous quality assessment, based on comparison to finished chimpanzee BACs and to the human genome (see Supplementary Notes).

Nucleotide-level accuracy is high by several measures. About 98% of the chimpanzee genome sequence has quality scores²⁵ of at least 40 (Q40), corresponding to an error rate of $\leq 10^{-4}$. Comparison of the WGS sequence to 1.3 Mb of finished BACs from the sequenced individual is consistent with this estimate, giving a high-quality discrepancy rate of 3×10^{-4} substitutions and 2×10^{-4} indels, which is no more than expected given the heterozygosity rate (see below) since 50% of the polymorphic alleles in the WGS sequence will differ from the single-haplotype BACs. Comparison of protein coding regions aligned between the WGS sequence, the recently published sequence of chromosome 21¹³ (formerly chromosome 22¹⁸) and the human genome, also revealed no excess of substitutions in the WGS sequence (see Supplementary Notes). Thus, by restricting our analysis to high-quality bases, the nucleotide-level accuracy of the WGS assembly is essentially equal to that of 'finished' sequence.

Structural accuracy is also high based on comparison with finished BACs from the primary donor and other chimpanzees, although the relatively low level of sequence redundancy limits local contiguity. Based on comparisons with the primary donor, some small supercontigs (most < 5kb) have not been positioned within large supercontigs (~1 event per 100 kb); these are not strictly errors but nonetheless affect the utility of the assembly. There are also small, undetected overlaps (all < 1kb) between consecutive contigs (~1.2 events per 100 kb) and occasional local misordering of small contigs (~0.2 events per 100 kb). No misoriented contigs were found. Comparison with the finished chromosome 21 sequence yielded similar discrepancy rates (see Supplementary Information).

The most problematic regions are those containing recent segmental duplications. Analysis of BAC clones from duplicated (n=75) and unique (n=28) regions showed that the former tend to be fragmented into more contigs (1.6-fold) and more supercontigs (3.2-fold). Discrepancies in contig order are also more frequent in duplicated than unique regions (~0.4 vs. ~0.1 events per 100 kb). The rate is two-fold higher in duplicated regions with the highest sequence identity (>98%). If we restrict the analysis to older duplications ($\leq 98\%$ identity) we find fewer assembly problems: 72% of those that can be mapped to the human genome are shared as duplications in both species. These results are consistent with the described limitations of current WGS assembly for regions of segmental duplication²⁶.

Chimpanzee Polymorphisms

The draft sequence of the chimpanzee genome also facilitates genome-wide studies of genetic diversity among chimpanzees, extending recent work²⁷⁻³⁰. We analyzed sequence reads from the primary donor, four other western and three central chimpanzees (*Pan troglodytes troglodytes*) to discover polymorphic positions within and between these individuals (see Supplementary Information).

A total of 1.66 million (M) high-quality single nucleotide polymorphisms (SNPs) were identified, of which 1.01 M are heterozygous within the primary donor, Clint. Heterozygosity rates were estimated to be 9.5×10^{-4} for Clint, 8.0×10^{-4} among western chimpanzees and 17.6×10^{-4} among central chimpanzees, with the variation between western and central chimpanzees being 19.0×10^{-4} . The diversity in western chimpanzees is similar to that seen for human populations³¹, while the level for central chimpanzees is roughly twice as high.

The observed heterozygosity in Clint is broadly consistent with western origin, although there are a small number of regions of distinctly higher heterozygosity. These may reflect a small amount of central ancestry, but more likely reflect undetected regions of segmental duplications present only in chimpanzees.

Genome Evolution

We set out to study the mutational events that have shaped the human and chimpanzee genomes since their last common ancestor. We explored changes at the level of single nucleotides, small insertions and deletions, interspersed repeats and chromosomal rearrangements. The analysis is nearly definitive for the smallest changes, but is more limited for larger changes, particularly lineage-specific segmental duplications, owing to the draft nature of the genome sequence.

Nucleotide Divergence

Best reciprocal nucleotide-level alignments of the chimpanzee and human genomes cover ~2.4 Gb of high quality sequence, including 89 Mb from chromosome X and 7.5 Mb from chromosome Y.

Genome-wide rates. We calculate the genome-wide nucleotide divergence between human and chimpanzee to be 1.23%, confirming recent results from more limited studies^{12,32,33}. The differences between one copy of the human genome and one copy of the chimpanzee genome include both the sites of fixed divergence between the species and some polymorphic sites within each species. By correcting for the estimated coalescence times in the human and chimpanzee

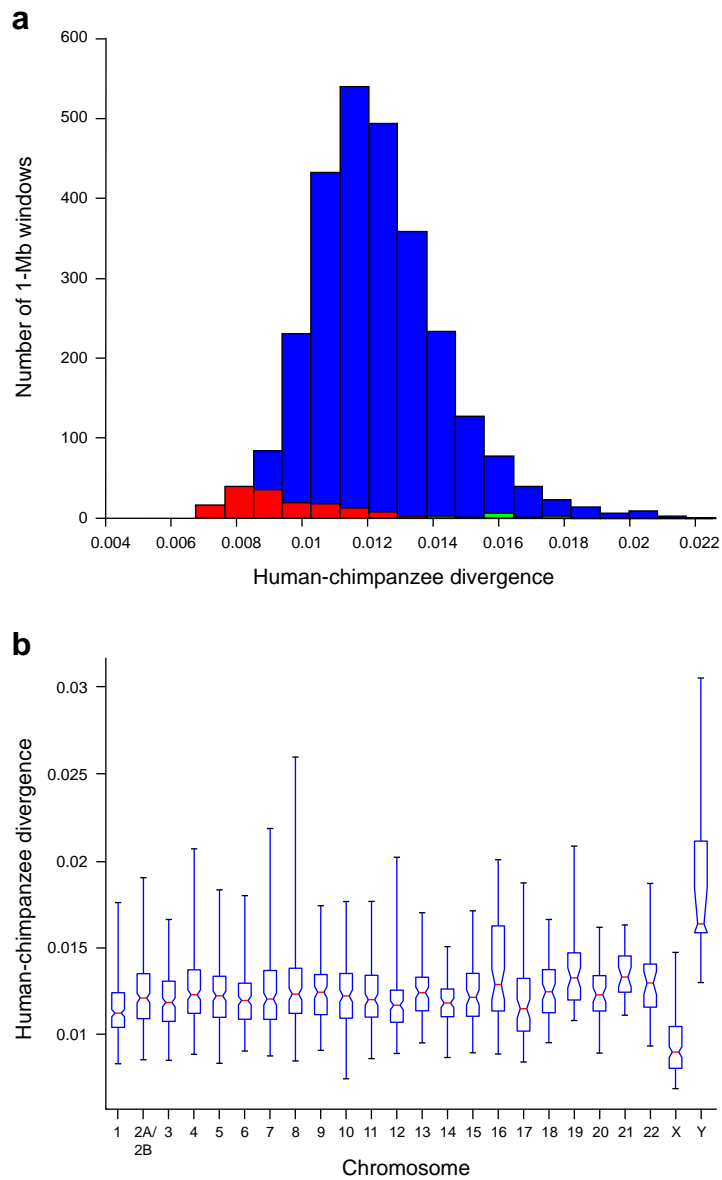


Figure 1. Human-chimpanzee divergence in 1 Mb segments across the genome. (a) Distribution of divergence of the autosomes (blue), the X chromosome (red) and the Y chromosome (green). (b) Distribution of variation by chromosome, shown as a box plot; the edges of the box correspond to quartiles; the notches to the standard error of the median; and the whiskers to the range. The X and Y chromosomes are clear outliers, but there is also high local variation within each of the autosomes.

populations (see Supplementary Information), we estimate that polymorphism accounts for 14-22% of the observed divergence rate and thus that the fixed divergence is ~1.06% or less.

Nucleotide divergence rates are not constant across the genome, as has been seen in comparisons of the human and murid genomes^{16,17,24,34,35}. The average divergence in 1 Mb segments fluctuates with a standard deviation of 0.25% (coefficient of variation, CV = 0.20), which is much greater than the 0.02% expected assuming a uniform divergence rate (Figure 1a; Figure 2).

Regional variation in divergence could reflect local variation in either mutation rate or other evolutionary forces. Among the latter, one important force is genetic drift, which can cause substantial differences in divergence time across loci when comparing closely related species, since the divergence time for orthologs is the sum of two terms: t_1 , the time since speciation, and t_2 , the coalescence time for orthologs within the common ancestral population³⁶. While t_1 is constant across loci (~6-7 Mya³⁷), t_2 is a random variable that fluctuates across loci (with a mean that depends on population size and here may be on the order of 1-2 million years³⁸). However, because of historical recombination, the characteristic scale of such fluctuations will be on the order of tens of kb, which is too small to account for the variation observed for 1 Mb regions³⁹ (see Supplementary Notes). Other potential evolutionary forces are positive or negative selection. While it is more difficult to quantify the expected contributions of selection in the ancestral population⁴⁰⁻⁴², it is clear that the effects would have to be very strong to explain the large-scale variation observed across mammalian genomes^{16,43}. There is tentative evidence from in-depth analysis of divergence and diversity that natural selection is not the major contributor to the large-scale patterns of genetic variability in humans⁴⁴⁻⁴⁶. For these reasons, we suggest that the large-scale variation in the human-chimpanzee divergence rate primarily reflect regional variation in mutation rate.

Chromosomal variation in divergence rate. Variation in divergence rate is evident even at the level of whole chromosomes (Figure 1b). The most striking outliers are the sex chromosomes, with a mean divergence of 1.9% for chromosome Y and 0.94% for chromosome X. The likely explanation is a higher mutation rate in the male vs. female germline⁴⁷. Indeed, the ratio of the male-female mutation rates (denoted α) can be estimated by comparing the divergence rates among the sex chromosomes and the autosomes and correcting for ancestral polymorphism as a function of population size of the most recent common ancestor (MRCA; see Supplementary Notes. Estimates for α range from 3-6, depending on the chromosomes compared and the assumed ancestral population size. This is significantly higher than recent estimates of α in the murids (~1.9)¹⁷ and resolves a recent controversy based on smaller datasets^{12,24,48,49}.

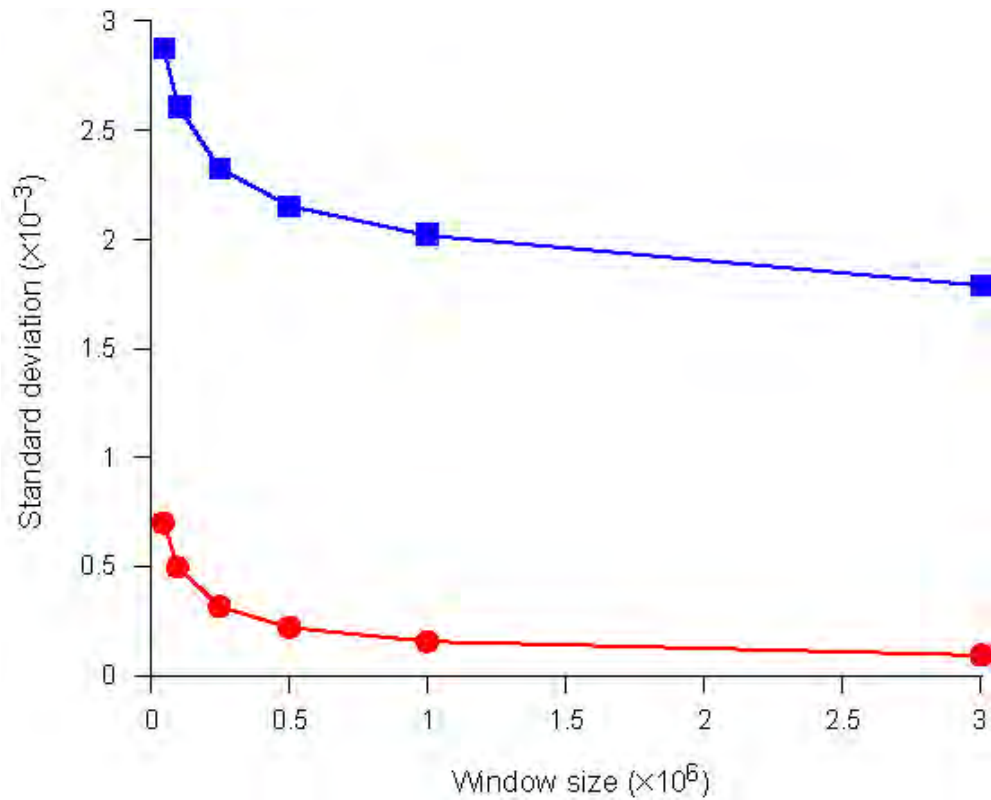


Figure 2. Distribution of observed (blue) and Poisson expectation (red) standard deviations of human-chimpanzee divergence over different window sizes. The observed variation is consistently larger than expected, but sample variance starts to increase rapidly in windows less than ~250 kb.

The higher mutation rate in the male germline is generally attributed to the 5-6-fold higher number of cell divisions undergone by male germ cells⁴⁷. We reasoned that this would affect mutations resulting from DNA replication errors (the rate should scale with the number of cell divisions) but not mutations resulting from DNA damage such as deamination of methyl CpG to TpG (the rate should scale with time). Accordingly, we calculated α separately for CpG sites, obtaining a value of ~ 2 from the comparison of rates between autosomes and chromosome X. This intermediate value is a composite of the rates of CpG loss and gain, and is consistent with roughly equal rates of CpG to TpG transitions in the male and female germ-line^{50,170}.

Significant variation in divergence rates is also seen among autosomes (Figure 1b; $p < 3 \times 10^{-15}$, Kruskal-Wallis test over 1 Mb windows), confirming earlier observations based on low-coverage WGS sampling¹². Additional factors thus influence the rate of divergence between chimpanzee and human chromosomes. These factors are likely to act at length scales significantly shorter than a chromosome, because the standard deviation across autosomes (0.21%) is comparable to the standard deviation seen in 1 Mb windows across the genome (0.13-0.35%). We therefore sought to understand local factors that contribute to variation in divergence rate.

Contribution of CpG dinucleotides. Sites containing CpG dinucleotides in either species show a substantially elevated divergence rate of 15.2% per base; they account for 25.2% of all substitutions while constituting only 2.1% of all aligned bases. The divergence at CpG sites represents both the loss of ancestral CpGs and the creation of new CpGs. The former process is known to occur at a rapid rate per base due to frequent methylation of cytosines in a CpG context and their frequent deamination^{51,52}, whereas the latter process likely proceeds at a rate more typical of other nucleotide substitutions. Assuming that loss and creation of CpG sites are close to equilibrium, the mutation rate for bases in a CpG dinucleotide must be 10-12-fold higher than for other bases (see Supplementary Notes and ref. 50).

Because of the high rate of CpG substitutions, regional divergence rates would be expected to correlate with regional CpG density. CpG density indeed varies across 1 Mb windows (mean = 2.1%, CV = 0.44 vs. 0.0093 expected under a Poisson distribution), but only explains 4% of the divergence rate variance. In fact, regional CpG and non-CpG divergence is highly correlated ($r = 0.88$; Figure 3), suggesting that higher-order effects modulate the rates of two very different mutation processes (see also ref. 46).

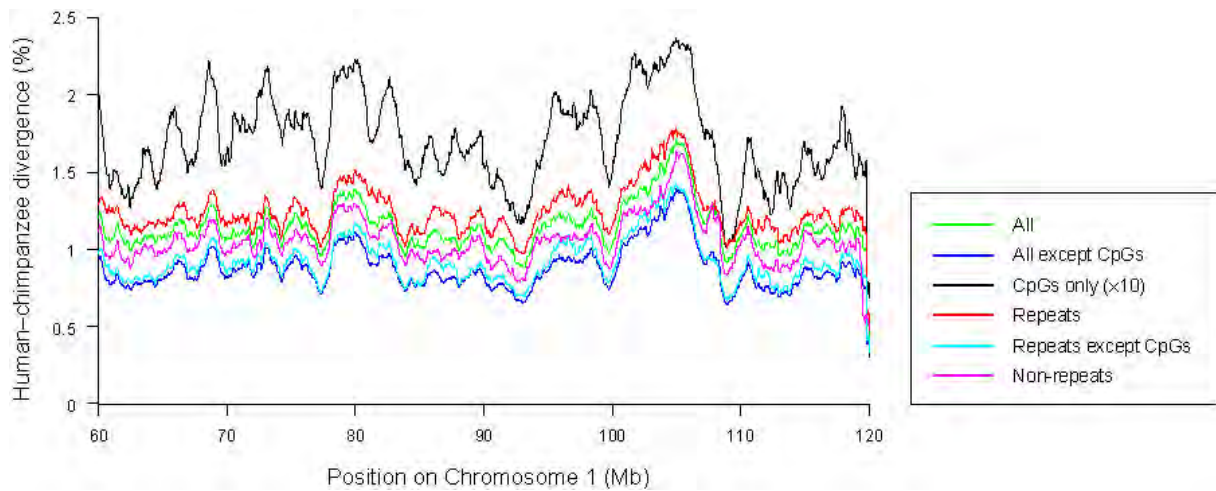


Figure 3. Co-variation of the divergence rate of different sequence classes in sliding 1 Mb windows. CpG and non-CpG divergence is highly correlated. As is repetitive and non-repetitive sequence divergence.

Increased divergence in distal regions. The most striking regional pattern is a consistent increase in divergence towards the ends of most chromosomes (Figure 4). The terminal 10 Mb of chromosomes (including distal regions and proximal regions of acrocentric chromosomes) averages 15% higher divergence than the rest of the genome ($p_{MW} < 10^{-30}$), with a sharp increase towards the telomeres. The phenomenon correlates better with physical distance than relative position along the chromosomes and may partially explain why smaller chromosomes tend to have higher divergence (Figure 5; see also ref. 15). These observations suggest that large-scale chromosomal structure, directly or indirectly, influences regional divergence patterns. The cause of this effect is unclear, but these regions (~15% of the genome) are notable in having high local recombination rate, high gene density and high GC content.

Correlation with chromosome banding. Another interesting pattern is that divergence increases with the intensity of Giemsa staining in cytogenetically-defined chromosome bands, with the regions corresponding to Giemsa dark bands (G-bands) showing 10% higher divergence than the genome-wide average ($p_{MW} < 10^{-14}$) (e.g. Figure 4). In contrast to terminal regions, these regions (17% of the genome) tend to be gene-poor, GC-poor and low in recombination^{53,54}. The elevated divergence seen in two such different types of regions suggests that multiple mechanisms are at work, and that no single known factor, such as GC content or recombination rate, is an adequate predictor of regional variation in the mammalian genome by itself (Figure 6). Elucidation of the relative contributions of these and other mechanisms will be important in formulating accurate models for population genetics, natural selection, divergence times and the evolution of genome-wide sequence composition⁵⁵.

Correlation with regional variation in the murid genome. Given that sequence divergence shows regional variation in both hominids (human-chimpanzee) and murids (mouse-rat), we asked whether the regional rates are positively correlated between orthologous regions. Such a correlation would suggest that the divergence rate is driven, in part, by factors that have been conserved over the ~75 Myr since rodents, humans and apes shared a common ancestor. Comparative analysis of the human and murid genomes has suggested such a correlation⁵⁶⁻⁵⁸, but the chimpanzee sequence provides the first direct opportunity to compare independent evolutionary processes between two mammalian clades.

We compared the local divergence rates in hominids and murids across major orthologous segments in the respective genomes (Figure 7). For orthologous segments that are non-distal in both hominids and murids, there is a strong correlation between the divergence rates ($r = 0.5$, $p < 10^{-11}$). In contrast, orthologous segments that are centered within 10 Mb of a hominid telomere have

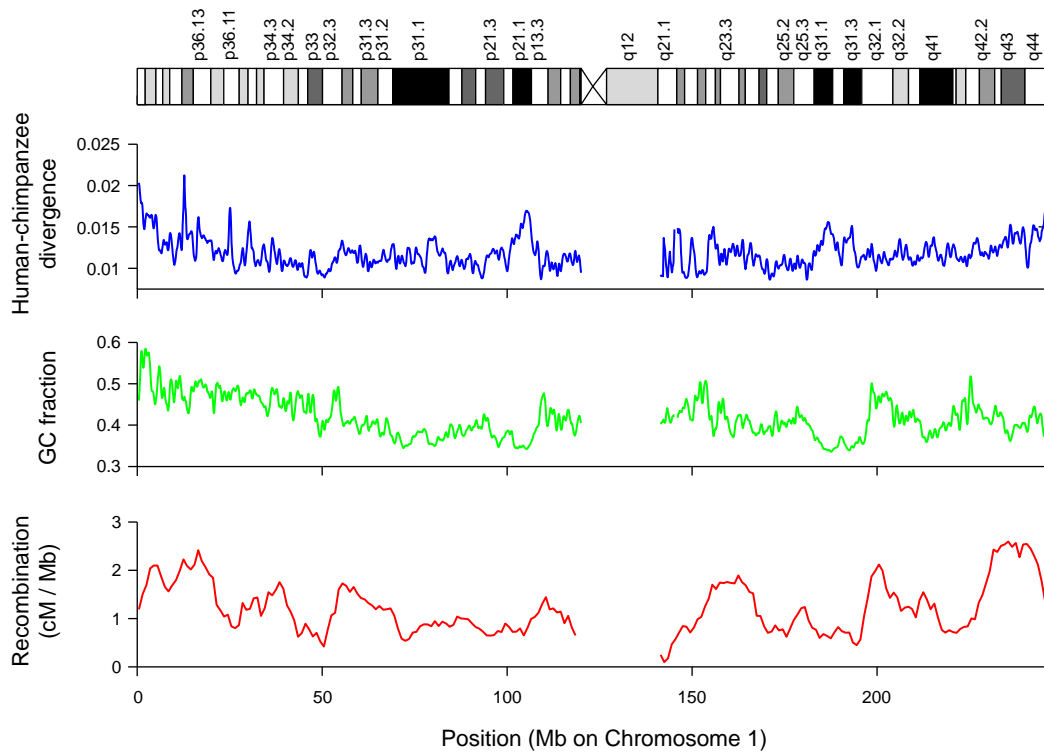


Figure 4. Regional variation in divergence rates. Human-chimpanzee divergence (blue), GC content (green) and human recombination rates [169] (red) in sliding 1 Mb windows for human and chimpanzee chromosome 1. Divergence and GC content is noticeably elevated near the 1p telomere, a trend that holds for most subtelomeric regions (see text). Internally on the chromosome, regions of low GC content and high divergence often correspond to the dark G-bands.

disproportionately high divergence rates and GC content relative to the murids ($p_{MW} < 10^{-11}$ and $p_{MW} < 10^{-4}$), implying that the elevation in these regions is, at least partially, lineage-specific. The same general effect is observed (albeit less pronounced) if CpG dinucleotides are excluded (Figure 8). Increased divergence and GC content might be explained by “biased gene conversion”⁵⁹ due to the high hominid recombination rates in these distal regions. Segments that are distal in murids do not show elevated divergence rates, which is consistent with this model, because the recombination rates of distal regions are not as elevated in mouse and rat⁶⁰.

Taken together, these observations suggest that sequence divergence rate is influenced by both conserved factors (stable across mammalian evolution) and lineage-specific factors (such as proximity to the telomere or recombination rate, which may change with chromosomal rearrangements).

Insertions and Deletions

We next studied the indel events that have occurred in the human and chimpanzee lineage by aligning the genome sequences to identify length differences. We will refer below to all events as insertions relative to the other genome, although they may represent insertions or deletions relative to the genome of the common ancestor.

The observable insertions fall into two classes: (i) ‘completely covered’ insertions, occurring within continuous sequence in both species, and (ii) ‘incompletely covered’ insertions, occurring within sequence containing one or more gaps in the chimpanzee, but revealed by a clear discrepancy between the species in sequence length. Somewhat different methods are needed for reliable identification of modest-size insertions (1 base to 15 kb) and large insertions (>15 kb), with the latter only being reliably identifiable in the human genome (see Supplementary Information).

The analysis of modest-sized insertions reveals ~32 Mb of human-specific sequence and ~35 Mb of chimpanzee-specific sequence, contained in ~5 M events in each species (Supplementary Notes and Figure 9). Nearly all of the human insertions are completely covered, whereas only half of the chimpanzee insertions are completely covered. Analysis of the completely covered insertions shows that the vast majority are small (45% of events cover only 1 bp, 96% are < 20 bp, and 98.6% are <80 bp), but that the largest few contain the majority of the sequence (with the ~70,000 indels larger than 80 bp comprising 73% of the affected bp) (Figure 10). The latter indels >80 bp fall into three categories: about one-quarter are newly inserted transposable elements; more than a third are due to microsatellite and satellite sequences; and the remainder are assumed to be mostly deletions in the other genome.

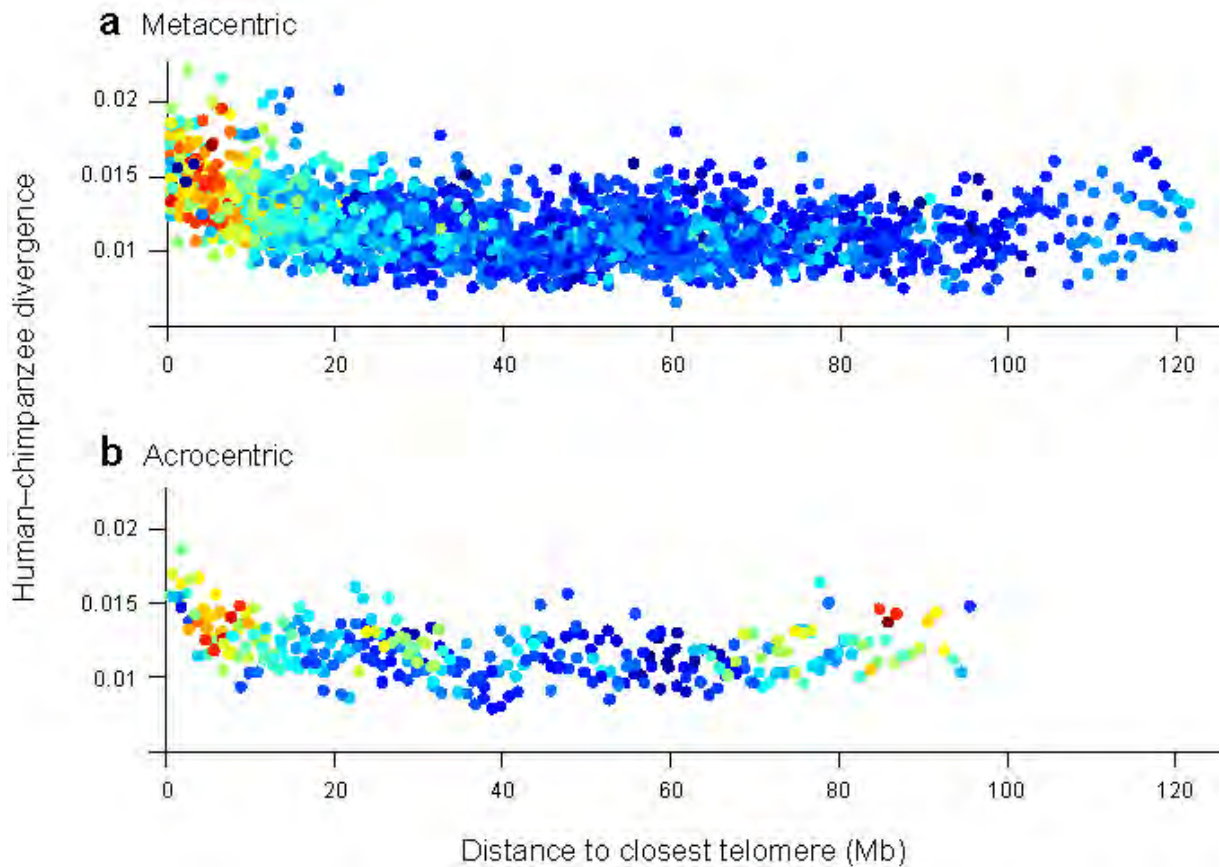


Figure 5. Correlation between human-chimpanzee divergence and distance to the closest telomere for 1 Mb windows on metacentric (a) or acrocentric (b) chromosomes. Each dot corresponds to a unique 1 Mb window. The colors of the dots represent their mean recombination rate (red = highest, dark blue = lowest).

The analysis of larger insertions (>15 kb) identified 163 human regions, containing 8.3 Mb of human-specific sequence (Figure 11). These cases include 34 regions that involve exons from known genes, which are discussed in a subsequent section. Although we have no direct measure of large insertions in the chimpanzee genome, it appears likely that the situation is similar.

Based on this analysis, we estimate that the human and chimpanzee genomes each contain 40-45 Mb of species-specific euchromatic sequence, and the indel differences between the genomes thus total ~90 Mb. This difference corresponds to ~3% of both genomes and dwarfs the 1.23% differences resulting from nucleotide substitution; this confirms and extends several recent studies^{61-64, 65}. Of course, the number of indel events is far fewer than the number of substitution events (~5 million vs. ~35 million).

Transposable Element Insertions

We next used the catalog of lineage-specific transposable element copies to compare the activity of transposons in the human and chimpanzee lineages (Table 2).

Endogenous Retroviruses. Endogenous retroviruses (ERVs) have become all but extinct in the human lineage, with only a single retrovirus (HERV-K) still active²⁴. HERV-K was found to be active in both lineages, with at least 73 human-specific insertions (7 full length and 66 solo LTRs) and at least 44 chimpanzee-specific insertions (1 full length and 44 solo LTRs). A few other ERV classes persisted in the human genome beyond the human-chimpanzee split, leaving ~9 human-specific insertions (all solo LTRs, including five HERV9) before dying out.

Against this background, it was surprising to find that the chimpanzee genome has two active retroviral elements (PtERV1, PtERV2) that are unlike any older elements in either genome; these must have been introduced by infection of the chimpanzee germline. The smaller family (PtERV2) has only a few dozen copies, which nonetheless represent multiple (~5-8) invasions, because the sequence differences among reconstructed subfamilies is too great (~8%) to have arisen by mutation since divergence from human. It is closely related to a baboon endogenous retrovirus (BaEV, 88% ORF2 product identity) and a feline endogenous virus (ECE-1, 86% ORF2 product identity). The larger family (PtERV1) is more homogeneous and has over 200 copies. While older ERVs, like HERV-K above, are primarily represented by solo LTRs, resulting from LTR-LTR recombination, more than half of the PtERV1 copies are still full-length, likely reflecting the young age of the elements. PtERV1-like elements are present in the rhesus monkey, olive baboon and African great apes but not in human, orangutan or gibbon, suggesting separate germline invasions in these species⁶⁶.

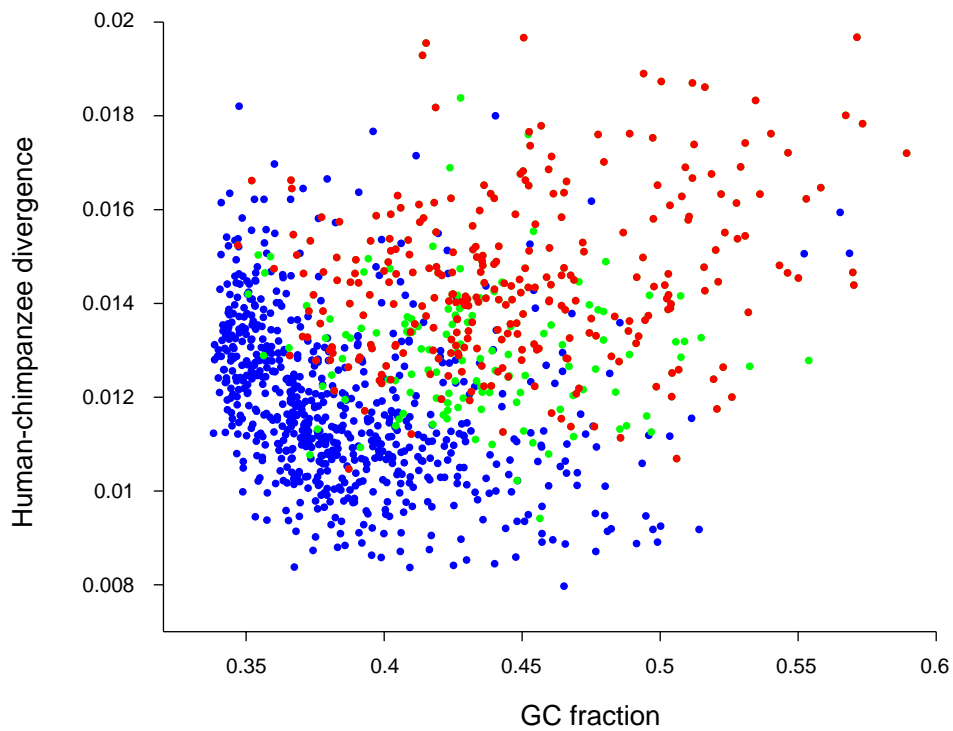


Figure 6. Divergence rates vs. GC content for 1 Mb segments across the autosomes. Conditional on recombination rate, the relationship between divergence and GC content varies. In regions with recombination rates less than 0.8 cM/Mb (blue), there is an inverse relationship, where high divergence regions tend to be GC-poor and low divergence regions tend to be GC-rich. In regions with recombination rates greater than 2.0 cM/Mb, whether within 10 Mb (red) or proximal (green) of chromosome ends, both divergence and GC-content are uniformly high.

Table 2: Transposable element activity in the human and chimpanzee lineages

Element	Chimpanzee ¹	Human ¹
Alu	2340 (0.7 Mb)	7082 (2.1 Mb)
LINE1	1979 (>5 Mb)	1814 (5.0 Mb)
SVA	757 (>1 Mb)	970 (1.3 Mb)
ERV class 1	234 (>1 Mb) ²	5 (8 kb) ³
ERV class 2	45 (55 kb) ⁴	77 (130 kb) ⁴
(Micro)satellite	7054 (4.1 Mb)	11101 (5.1 Mb)

¹ Number of lineage-specific insertions (total size of inserted sequences) in the aligned parts of the genomes. ² PtERV1 and PtERV2. ³ HERV9. ⁴ Mostly HERV-K.

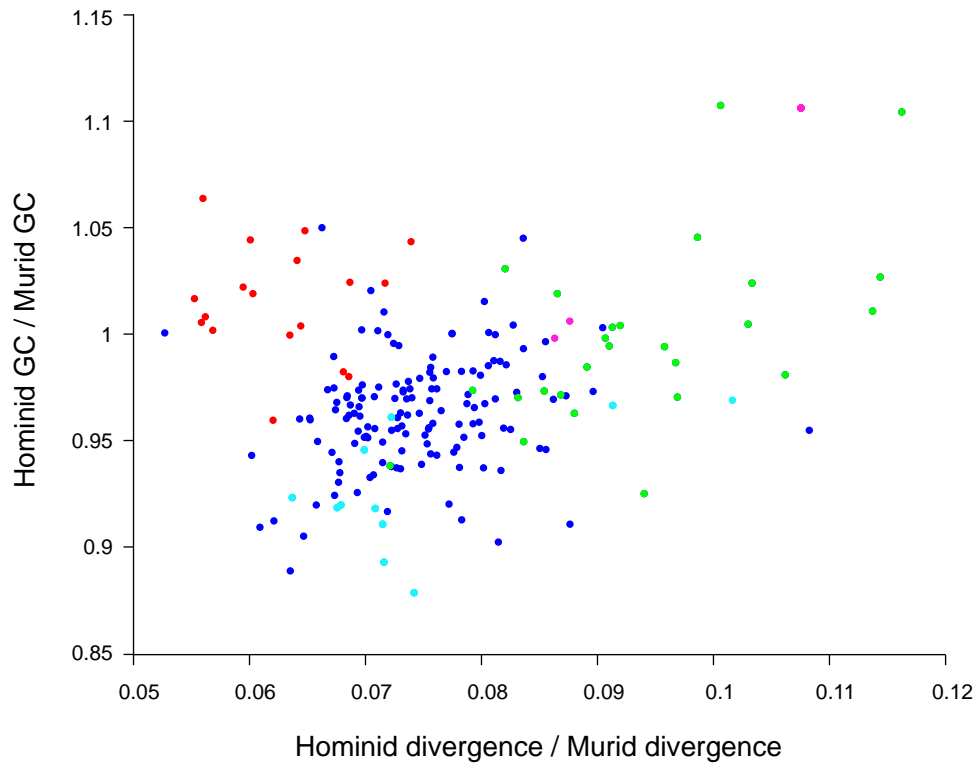


Figure 7. Disproportionately elevated divergence and GC content near hominid telomeres. Scatter plot of the ratio of human-chimpanzee divergence over mouse-rat divergence vs. the ratio of human GC-content over mouse GC-content across 199 syntenic blocks for which more than 1 Mb of sequence could be aligned between all four species. Blocks whose center is within 10 Mb of a telomere in hominids only (green) or in hominids and murids (magenta), but not in murids only (light blue), show a significant trend towards higher ratios than internal blocks (blue). Blocks on the X chromosome (red) tend to show a lower divergence ratio than autosomal blocks, consistent with a smaller difference between autosomal and X divergence in murids than in hominids (lower α).

Higher Alu activity in humans. SINE (Alu) elements have been three-fold more active in humans than chimpanzee (~7000 vs. ~2300 lineage-specific copies in the aligned portion), refining the rather broad range (2-7-fold) estimated in smaller studies^{13,65,67}. Most chimpanzee-specific elements belong to a subfamily (AluYc1) that is very similar to the source gene in the common ancestor. By contrast, most human-specific Alus belong to two new subfamilies (AluYa5 and AluYb8) that have evolved since the chimpanzee-human divergence and differ substantially from the ancestral source gene⁶⁷. It seems likely that the resurgence of Alu elements in humans is due to these potent new source genes. However, based on an examination of finished sequence, the baboon shows a 1.6 fold higher Alu activity relative to human new insertions, suggesting that there may also have been a general decline in activity in the chimpanzee⁶⁵.

Some of the human-specific Alus are highly diverged (92 with > 5% divergence), which would seem to suggest that they are much older than the human-chimpanzee split. Possible explanations include: gene conversion by nearby older elements; processed pseudogenes arising from a spurious transcription of an older element; precise excision from the chimpanzee genome; or high local mutation rate. In any case, the presence of such anomalies suggests that caution is warranted in the use of single repeat elements as homoplasy-free phylogenetic markers.

New Alus target AT-rich DNA in human and chimpanzee genomes. Older SINE elements are preferentially found in gene-rich, GC-rich regions whereas young SINE elements are found in gene-poor AT-rich regions where LINE-1 (L1) copies also accumulate^{24,68}. The latter distribution is consistent with the fact that Alu retrotransposition depends on L1 proteins⁶⁹. Murid genomes revealed no change in SINE distribution with age¹⁷.

The human pattern could reflect either preferential retention of SINEs in GC-rich regions, due to selection or mutation bias, or a recent change in Alu insertion preferences. With the availability of the chimpanzee genome, it is possible to classify the youngest Alu copies more accurately and thus to begin to distinguish these possibilities.

Analysis shows that lineage-specific SINEs in both human and chimpanzee are biased toward AT-rich regions, as opposed to even the most recent copies in the MRCA (Figure 12). This strongly favors that SINEs are indeed preferentially retained in GC-rich DNA, but comparison with a more distant primate is required to formally rule out that the insertion bias of SINEs did not change just prior to speciation.

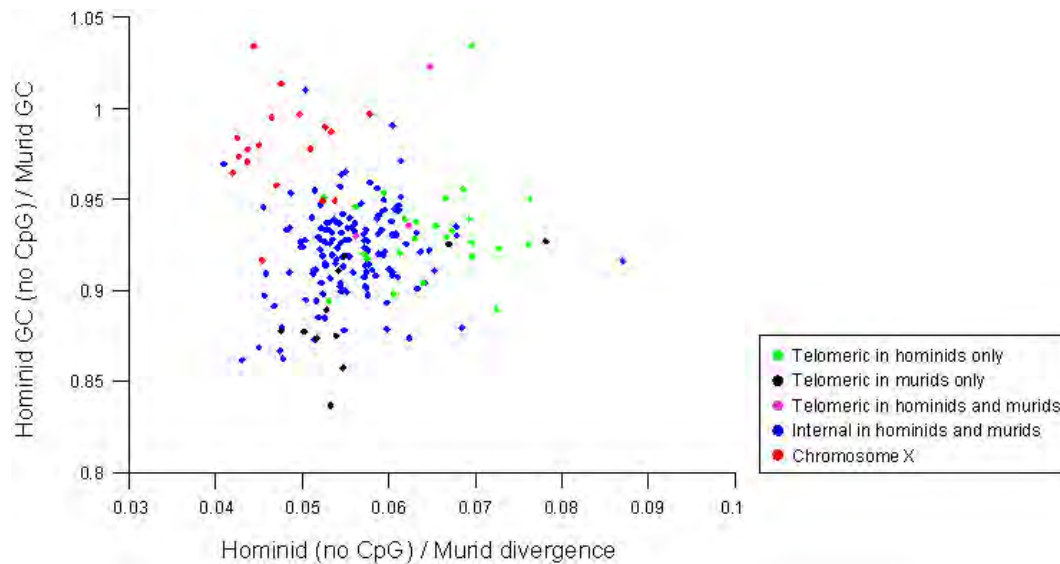


Figure 8. The ratio of human-chimpanzee non-CpG divergence over mouse-rat divergence vs. the ratio of human GC-content over mouse GC-content across 199 syntenic blocks greater than 1 Mb. Hominid-specific acceleration in subtelomeric blocks is evident even when ignoring CpG sites.

Equal activity of L1 in both species. The human and chimpanzee genomes both show ~2000 lineage-specific L1 elements, contrary to previous estimates based on small samples that L1 activity is 2-3-fold higher in chimpanzee ⁷⁰.

Transcription from L1 source genes can sometimes continue into 3' flanking regions, which can then be co-transposed ^{71,72}. Human-chimpanzee comparison revealed that ~15% of the species-specific insertions appear to have carried with them at least 50 bp of flanking sequence (followed by a polyA tail and a target site duplication). In principle, incomplete reverse transcription could result in insertions of the flanking sequence only (without any L1 sequence) mobilizing gene elements such as exons, but we found no evidence of this.

Retrotransposed gene copies. The L1 machinery also mediates retrotransposition of host mRNAs, resulting in many intron-less (processed) pseudogenes in the human genome ⁷³⁻⁷⁵. We identified 163 lineage-specific retrotransposed gene copies in human and 246 in chimpanzee. Correcting for incomplete sequence coverage of the chimpanzee genome, we estimate that there are ~200 and ~300 processed gene copies in human and chimpanzee, respectively. Processed genes thus appear to have arisen at a rate of ~50 per Myr since the divergence of human and chimpanzee; this is lower than the estimated rate for early primate evolution ⁷³, perhaps reflecting the overall decrease in L1 activity. As expected ⁷⁶, ribosomal protein genes constitute the largest class in both species. The second largest class in chimpanzee corresponds to zinc finger C2H2 genes, which are not a major class in the human genome.

The retroposon SVA and distribution of CpG islands by transposable elements. The third most active element since speciation has been SVA, which created about 1000 copies in each lineage. SVA is a composite element (~1.5-2.5 kb) consisting of two Alu fragments, a tandem repeat and a region apparently derived from the 3' end of a HERV-K transcript; it is likely mobilized by L1 ^{77,78}. This element is of particular interest because each copy carries a sequence that satisfies the definition of a CpG island ⁷⁹ and contains potential transcription factor binding sites; the dispersion of 1000 SVA copies could therefore be a source of regulatory differences between chimpanzee and human. At least three human genes contain SVA insertions near their promoters, one of which have been found to be differentially expressed between the two species ^{80 81}, but additional investigations will be required to determine if the SVA insertion directly caused this difference.

Homologous recombination between interspersed repeats. Human-chimpanzee comparison also makes it possible to study homologous recombination between nearby repeat elements as a source of genomic deletions. We found 612 deletions (totaling 2Mb) in the human

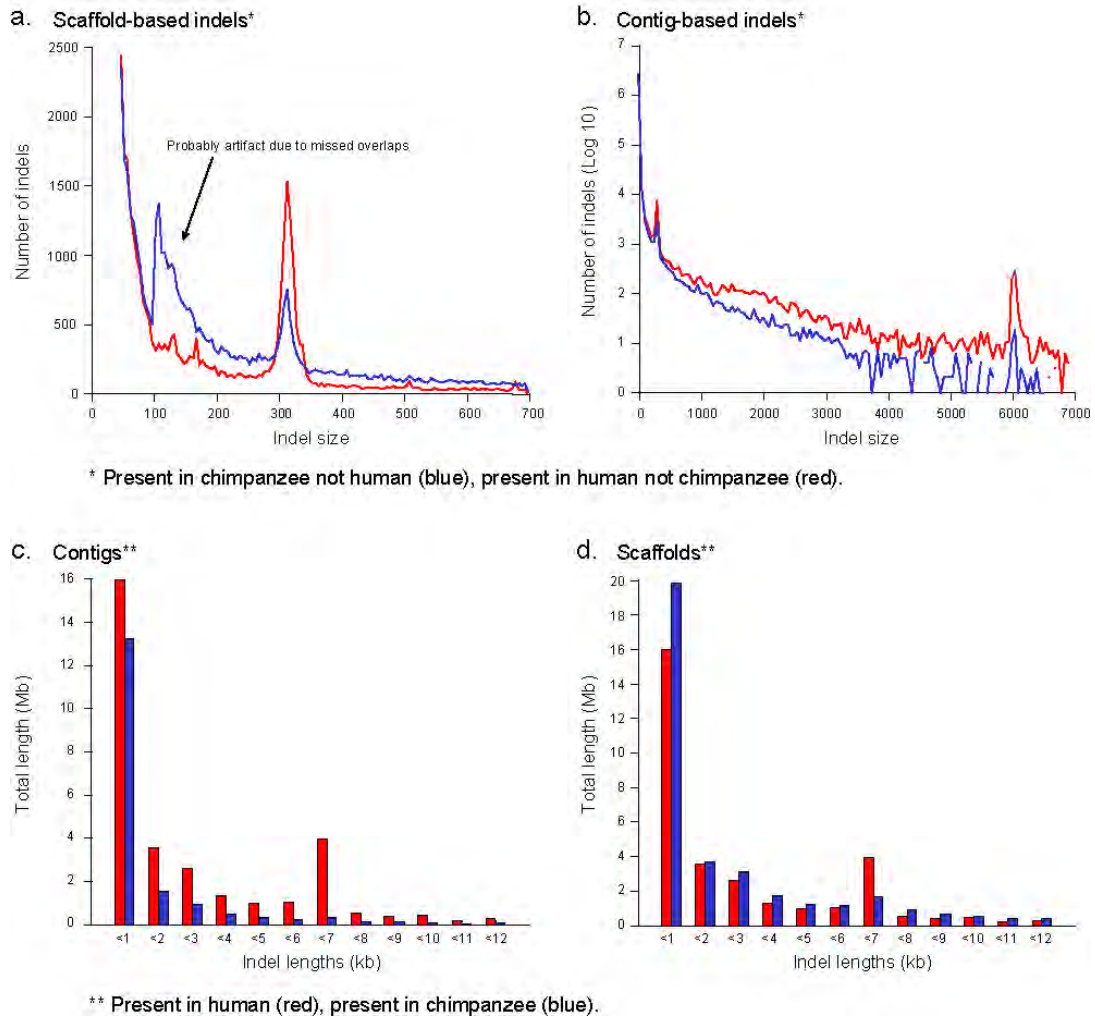


Figure 9. Length distribution of small indels (< 15 kb) detected within scaffolds (a and d) or contigs only (b and c). For chimpanzee “insertions”, the former is an over-estimate of the number of actual indels due to assembly artifacts, whereas the latter is an under-estimate, due to the small contig size.

genome that appear to have resulted from recombination between two nearby Alus present in the common ancestor and 914 such events in the chimpanzee genome. (The events are not biased to AT-rich DNA and thus would not explain the preferential loss of Alus in such regions discussed above). Similarly, we found 26 and 48 instances involving adjacent L1 copies and 8 and 22 instances involving retroviral LTRs in human and chimpanzee, respectively. None of the repeat mediated deletions removed an orthologous exon of a known human gene in chimpanzee.

The genome comparison allows one to estimate the dependency of homologous recombination on divergence and distance. Homologous recombination appears to occur between quite (>25%) diverged copies (Figure 13), while the number of recombination events (n) varies inversely with the distance (d , in bases) between the copies (as $n \approx 6 \times 10^6 d^{-1.7}$; $r^2 = 0.9$).

Large Scale Rearrangements

Finally, we examined the chimpanzee genome sequence for information about large-scale genomic alterations. Cytogenetic studies have shown that human and chimpanzee chromosomes differ by one chromosomal fusion, at least nine pericentric inversions, and in the content of constitutive heterochromatin⁸². Human chromosome 2 resulted from a fusion of two ancestral chromosomes that remained separate in the chimpanzee lineage (chromosomes 2A and 2B in the revised nomenclature¹⁸, formerly chimpanzee chromosomes 12 and 13); the precise fusion point has been mapped and its duplication structure described in detail^{83,84}. In accord with this, alignment of the human and chimpanzee genome sequences shows a break in continuity at this point.

We searched the chimpanzee genome sequence for the precise locations of the 18 breakpoints corresponding to the 9 pericentric inversions. By mapping paired-end sequences from chimpanzee large insert clones to the human genome, we were able to identify 13 of the breakpoints within the assembly from discordant end alignments. The positions of five breakpoints (on chromosomes 4, 5, and 12) were tested by FISH analysis and all were confirmed. Also, the positions of three previously mapped inversion breakpoints (on chromosomes 15 and 18) matched closely those found in the assembly^{85,86}. The paired-end analysis works well in regions of unique sequence, which constitute the bulk of the genome, but is less effective in regions of recent duplication due to ambiguities in mapping of the paired-end sequences. Beyond the known inversions, we also found suggestive evidence of many additional smaller inversions, as well as older segmental duplications (< 98% identity). However, both smaller inversions and more recent segmental duplications will require further investigations.

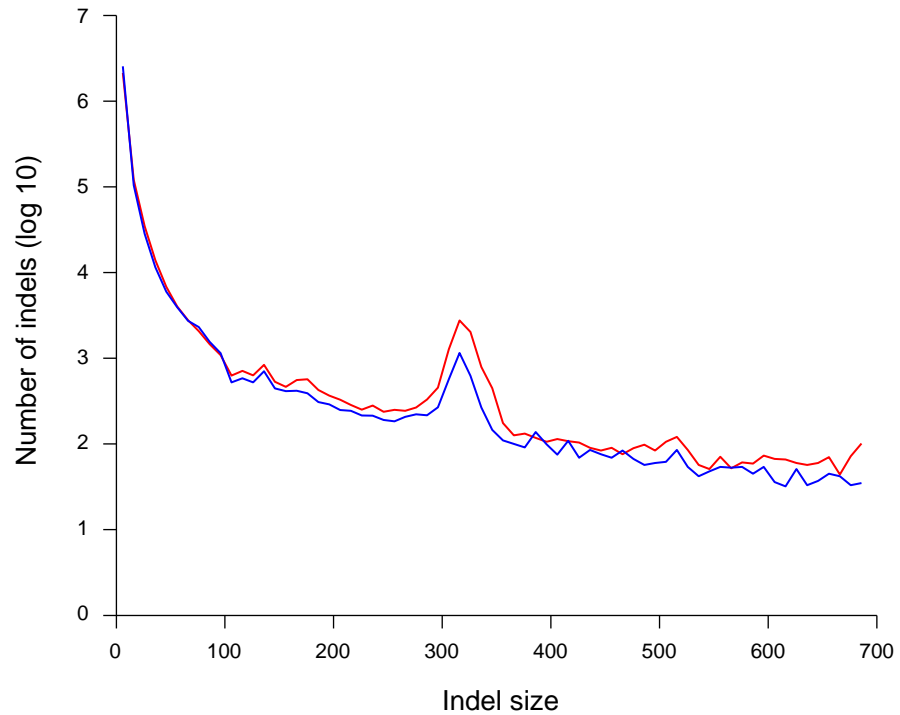


Figure 10. Length distribution of small insertion-deletion events (<15 kb), as determined using bounded sequence gaps. Sequences present in chimpanzee but not in human (blue) or present in human but not in chimpanzee (red) are shown. The prominent spike around 300 nucleotides corresponds to SINE insertion events. The vast majority of indels are smaller than 20 bp, but larger indels account for the bulk of lineage-specific sequence in the two genomes.

Gene Evolution

We next sought to use the chimpanzee sequence to study the role of natural selection in the evolution of human protein coding genes. Genome-wide comparisons can shed light on many central issues, including: the magnitude of positive and negative selection; the variation in selection across different lineages, chromosomes, gene families and individual genes; and the complete loss of genes within a lineage.

We began by identifying a set of 13,454 pairs of human and chimpanzee genes with unambiguous 1:1 orthology for which it was possible to generate high-quality sequence alignments covering virtually the entire coding region (see Supplementary Notes). The list contains a large fraction of the entire complement of human genes, although it underrepresents gene families that have undergone recent local expansion (such as olfactory receptors and immunoglobins). To facilitate comparison with the murid lineage, we also compiled a set of 7,043 human, chimpanzee, mouse and rat genes with unambiguous 1:1:1:1 orthology and high-quality sequence alignments.

Average Rates of Evolution

To assess the rate of evolution for each gene, we estimated K_a , the number of coding base substitutions that result in amino acid change as a fraction of all such possible sites (the non-synonymous substitution rate). Because the background mutation rate varies across the genome, it is crucial to normalize K_a for comparisons between genes. A striking illustration of this variation is the fact that the mean K_a is 37% higher in the rapidly diverging distal 10 Mb of chromosomes than in the more proximal regions. Classically, the background rate is estimated by K_s , the synonymous substitution rate (coding base substitutions that because of codon redundancy do not result in amino acid change). Because a typical gene has only a few synonymous changes between humans and chimpanzees, and not infrequently is zero, we exploited the genome sequence to estimate the local intergenic/intronic substitution rate, K_i , where appropriate. K_a and K_s were also estimated for each lineage separately using mouse and rat as outgroups (Figure 14).

The K_a/K_s ratio is a classical measure of the overall evolutionary constraint on a gene, where $K_a/K_s \ll 1$ indicates that a substantial proportion of amino acid changes must have been eliminated by purifying selection. Under the assumption that synonymous substitutions are neutral, $K_a/K_s > 1$ implies, but is not a necessary condition for, adaptive or positive selection. The K_a/K_i ratio has the same interpretation. The ratios will sometimes be denoted below by ω with an appropriate subscript (for example, ω_{human}) to indicate the branch of the evolutionary tree under study.

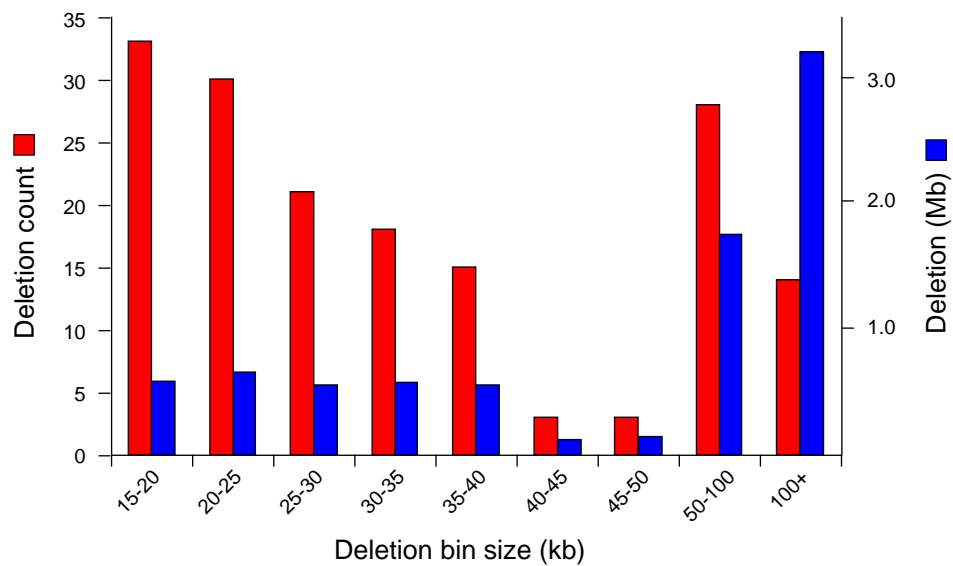


Figure 11. Length distribution of large insertion-deletion events (>15 kb), as determined using paired-end sequences from chimpanzee mapped against the human genome. Both the total number of candidate human insertions/chimpanzee deletions (blue) and the number of bases altered (red) are depicted.

Table 3: Comparison of K_a/K_s for divergence and human diversity

	ΔA	ΔS	K_a/K_s	% excess ¹	CI ²
Human-chimpanzee divergence	38773	61737	0.23	-	-
HapMap (European ancestry) ³					
Rare derived alleles (< 15%)	1614	1540	0.39	67	[59, 75]
Common alleles	1199	1907	0.23	0	[-5, 6]
Frequent derived alleles (> 85%)	209	356	0.22	-7	[-19, 7]
HapMap (African ancestry) ³					
Rare derived alleles (< 15%)	849	842	0.36	61	[50, 72]
Common alleles	495	803	0.22	-2	[-10, 7]
Frequent derived alleles (> 85%)	59	82	0.26	15	[-11, 48]
Affymetrix 120K (Multi-ethnic) ⁴					
Rare derived alleles (< 15%)	74	82	0.33	44	[14, 80]
Common alleles	77	137	0.21	-11	[-28, 12]
Frequent derived alleles (> 85%)	10	15	0.25	6	[-42, 95]

ΔA = Number of observed non-synonymous substitutions. ΔS = Number of observed synonymous substitutions. ¹ A negative fraction indicates excess of non-synonymous divergence over polymorphism. ² 95% confidence intervals assuming non-synonymous substitutions are Poisson distributed. ³ Source: <http://www.hapmap.org> (Public Release #13). ⁴ Source: <http://www.affymetrix.com>

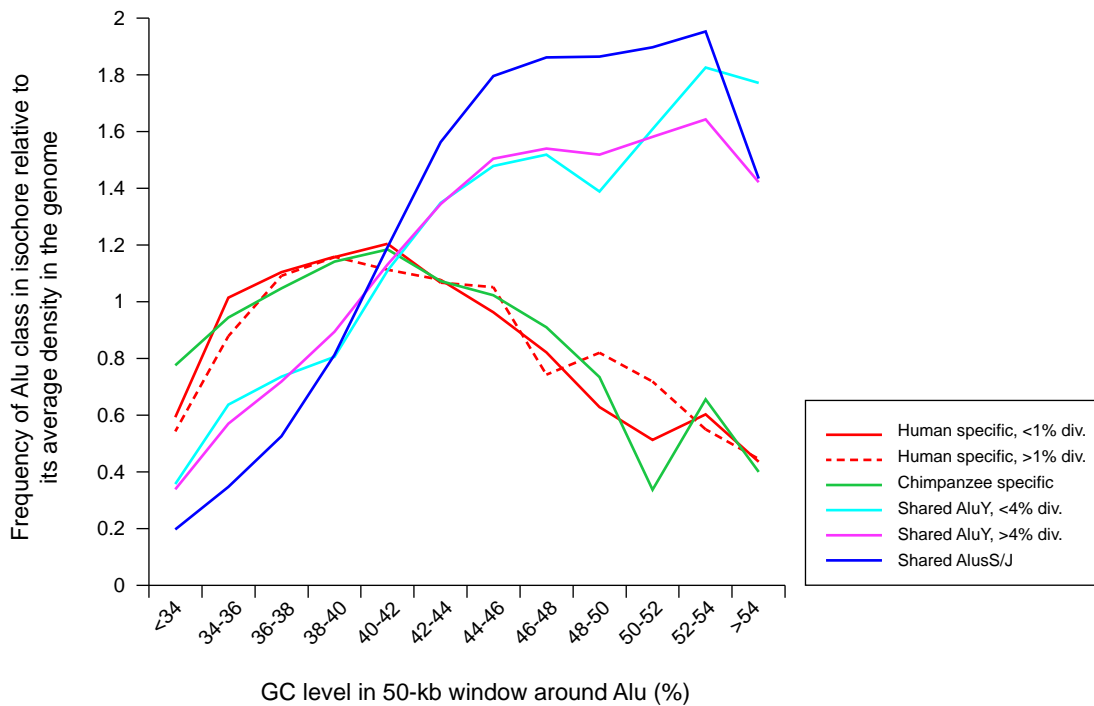


Figure 12. Correlation of Alu age and distribution by GC content. Alu elements that inserted after human-chimpanzee divergence are densest in the GC-poor regions of the genome (peaking at 36-40% GC), whereas older copies, common to both genomes, crowd GC-rich regions. The figure is similar to Figure 23 of ref. 24, but the use of chimpanzee allows improved separation of young and old elements, leading to a sharper transition in the pattern.

Evolutionary constraint on amino acid sites within hominid lineage. Overall, human and chimpanzee genes are extremely similar, with the encoded proteins identical in the two species in 29% of cases. The median number of non-synonymous and synonymous substitutions per gene are 2 and 3, respectively. About 5% of the proteins show in-frame indels, but these tend to be small (median = 1 codon) and to occur in regions of repeated sequence. The close similarity of human and chimpanzee genes necessarily limits the ability to make strong inferences about individual genes, but there is abundant data to study important sets of genes.

The K_a/K_s ratio for the human-chimpanzee lineage (ω_{hominid}) is 0.23. The value is much lower than some recent estimates based on limited sequence data (ranging as high as 0.63⁷), but is consistent with an estimate (0.22) from random EST sequencing⁴⁴. Similarly, K_a/K_i was also estimated as 0.23.

Under the assumption that synonymous mutations are selectively neutral, the results imply that 77% of amino acid alterations in hominid genes are sufficiently deleterious as to be eliminated by natural selection. Because synonymous mutations are not entirely neutral (see below), the actual proportion of amino acid alterations with deleterious consequences may be higher. Consistent with previous studies⁸, we find that K_a/K_s of human polymorphisms with frequencies up to 15% is significantly higher than that of human-chimpanzee differences and more common polymorphisms (Table 3), implying that at least 25% of the deleterious amino acid alterations may often attain readily detectable frequencies and thus contribute significantly to the human genetic load.

Evolutionary constraint on synonymous sites within hominid lineage. We next explored the evolutionary constraints on synonymous sites, specifically four-fold degenerate sites. Because such sites have no effect on the encoded protein, they are often considered to be selectively neutral in mammals.

We re-examined this assumption by comparing the divergence at four-fold degenerate sites with the divergence at nearby intronic sites. Although overall divergence rates are very similar at four-fold degenerate and intronic sites, direct comparison is misleading because the former have a higher frequency of the highly mutable CpG dinucleotides (9% vs. 2%). When CpG and non-CpG sites are considered separately, we find that both CpG sites and non-CpG sites show dramatically lower divergence in exonic synonymous sites than in introns (~50% and ~30% lower, respectively). This result resolves recent conflicting reports based on limited datasets^{44,87} by showing that such sites are indeed under constraint.

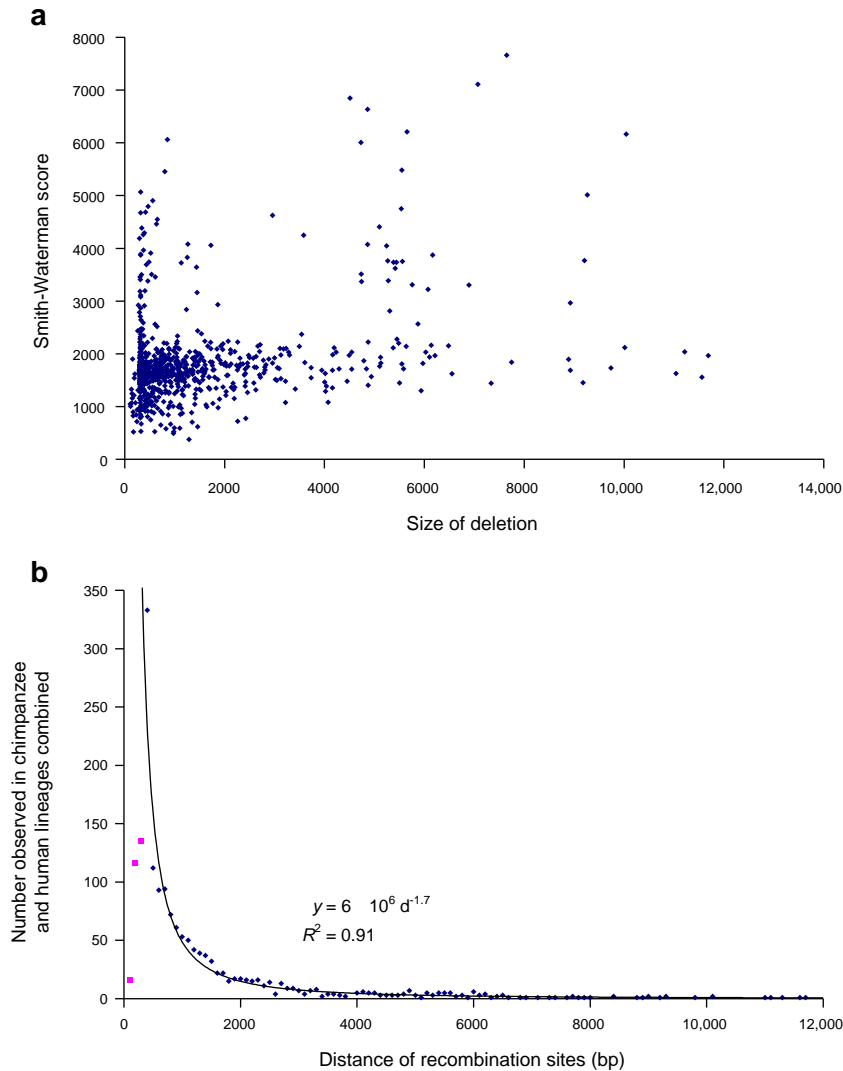


Figure 13. Dependency of homologous recombination between Alu elements on divergence and distance. (a) While homologous recombination occurs between quite divergent (Smith-Waterman score < 1000), closely spaced copies, more distant recombination seems to favor a better match between the recombining repeats. (b) The frequency of Alu-Alu mediated recombination falls dramatically as a function of distance between the recombining copies. The first three points (magenta) involve recombination between left or right arms of one Alu inserted into another. The high number of occurrences at a distance of 300-400 nucleotides is due to the preference of integration in the A-rich tail; exclusion of this point does not change the parameters of the equation.

The constraint does not appear to result from selection on the usage of preferred codons, which has been detected in lower organisms,⁸⁸ such as bacteria⁸⁹, yeast⁹⁰ and flies⁹¹. In fact, divergence at four-fold degenerate sites increases slightly with codon usage bias (Kendall's $\tau = 0.097$, $p < 10^{-14}$). Alternatively, the observed constraint at synonymous sites might reflect 'background selection' – that is, the indirect effect of purifying selection at amino acid sites causing reduced diversity and thereby reduced divergence at closely linked sites⁴¹. Given the low rate of recombination in hominid genomes (a 1 kb region experiences only ~1 crossover per 100,000 generations or 2 M years), such background selection should extend beyond exons to include nearby intronic sites⁹². However, when the divergence rate is plotted relative to exon-intron boundaries, we find that the rate jumps sharply within a short region of ~7 bp at the boundary (Figure 15). This pattern strongly suggests that the action of purifying selection at synonymous sites is direct rather than indirect, suggesting other signals, e.g. those involved in splice site selection, may be embedded in the coding sequence and therefore constrain synonymous sites.

Comparison with murids. An accurate estimate of K_a/K_s makes it possible to study how evolutionary constraint varies across clades. Ohta⁹³ predicted more than thirty years ago that selection against deleterious mutations would depend on population size, with mutations being strongly selected only if they reduce fitness by $s \gg 1/4N$ (where N is effective population size). This would predict that genes would be under stronger purifying selection in murids than hominids, owing to their presumed larger population size. Initial analyses (involving fewer than 50 genes⁹⁴) suggested a strong effect, but the wide variation in estimates of K_a/K_s in hominids^{7,8,95} and murids⁹⁶ has complicated this analysis⁴⁴.

Using the large collection of 7,043 orthologous quartets, we calculated mean K_a/K_s values for the various branches of the four-species evolutionary tree (Figure 14). The K_a/K_s ratio for hominids is 0.20. (This is slightly lower than the value of 0.23 obtained with all human-chimpanzee orthologs, probably reflecting slightly greater constraint on the class of proteins with clear orthologs across hominids and murids).

The K_a/K_s ratio is strikingly lower for murids than for hominids ($\omega_{\text{murid}} \sim 0.13$ vs. $\omega_{\text{hominid}} \sim 0.20$) (Figure 14). This implies that there is a ~35% excess of the amino acid changing mutations in the two hominids, relative to the two murids. Excess amino acid divergence may be explained by either increased adaptive evolution or relaxation of evolutionary constraints. As shown in the next section, the latter appears to be the principal explanation.

Relaxed constraints in human evolution. The K_a/K_s ratio can be used to make inferences about the role of positive selection in human evolution^{97,98}. Because alleles under positive selection

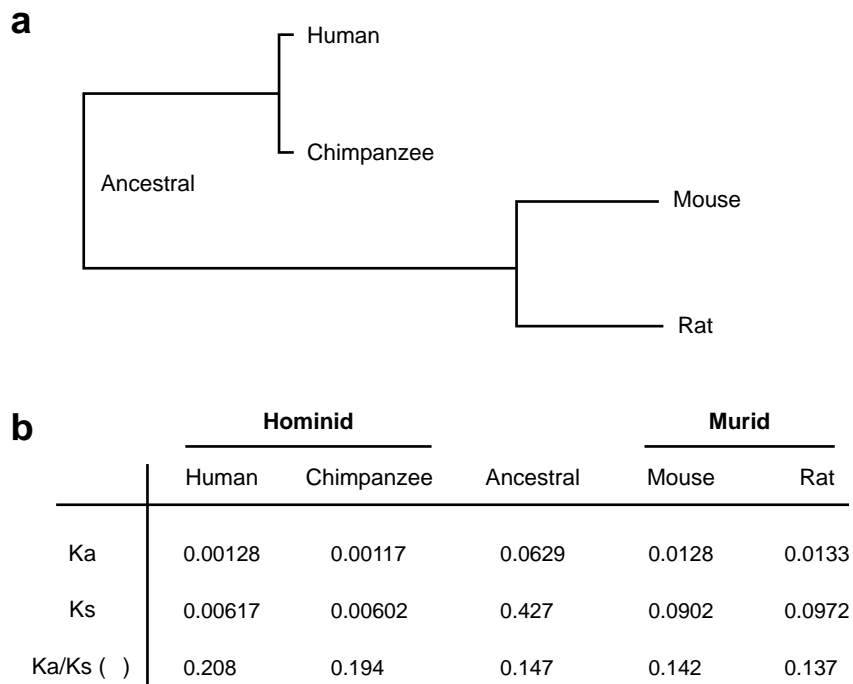


Figure 14. Human-chimpanzee-mouse-rat tree with branch specific Ka/Ks (ω) values. (a) Evolutionary tree. The branch lengths are proportional to the absolute rates of amino acid divergence. (b) Maximum-likelihood estimates of the rates of evolution in protein-coding genes for humans, chimpanzees, mice and rats. In the text, ω [hominid] is the Ka/Ks of the combined human and chimpanzee branches and ω [murid] of the combined mouse and rat branches. The slight difference between ω [human] and ω [chimp] is not statistically significant; masking of some heterozygous bases in the chimpanzee sequence may contribute to the observed difference (see Supplementary Notes).

spread rapidly through a population, they will be found less frequently as common human polymorphisms than as human-chimpanzee differences⁸. Positive selection can thus be detected by comparing the K_a/K_s ratio for common human polymorphisms with the K_a/K_s ratio for hominid divergence. Fay and colleagues⁸ estimated these ratios as $\omega_{\text{polymorphism}} \sim 0.20$ based on an initial collection of common SNPs in human genes and $\omega_{\text{divergence}} \sim 0.34$ based on comparison of human and Old World monkey genes. They thus inferred the proportion of amino acid changes attributable to positive selection to be $\sim 35\%$. This would imply a huge quantitative role for positive selection in human evolution.

With the availability of extensive data for both human polymorphism and human-chimpanzee divergence, we repeated this analysis (using the same set of genes for both estimates). We find that $\omega_{\text{polymorphism}} \sim 0.21-0.23$ and $\omega_{\text{divergence}} \sim 0.23$ are statistically indistinguishable (Table 3). Although some of the amino acid substitutions in human and chimpanzee evolution must surely reflect positive selection, the results indicate that the proportion of changes fixed by positive selection appears to be much lower than the previous estimate⁸. (Because the previous results involved comparison to Old World monkeys, it is possible that they reflect strong positive selection earlier in primate evolution. However, we suspect that they reflect the fact that relatively few genes were studied and that different genes were used to study polymorphism and divergence).

Relaxed negative selection pressures thus primarily explain the excess amino acid divergence in hominid genes relative to murids. Moreover, since both ω_{human} and $\omega_{\text{chimpanzee}}$ are similarly elevated this explanation applies equally to both lineages.

We next sought to study variation in the evolutionary rate of genes within the hominid lineage, by searching for unusually high or low levels of constraint for genes and sets of genes.

Rapid evolution in individual genes.

We searched for individual genes that have accumulated amino acid substitutions faster than expected given the neutral substitution rate, as potentially being under strong positive selection. A total of 585 of the 13,454 human-chimpanzee orthologs (4.4%) have observed $K_a/K_i > 1$ (see Supplementary Information). However, given the low divergence, the K_a/K_i statistic has large variance. Simulations show that estimates of $K_a/K_i > 1$ would be expected to occur simply by chance in at least 263 cases, if purifying selection is allowed to act non-uniformly across genes (Figure 16).

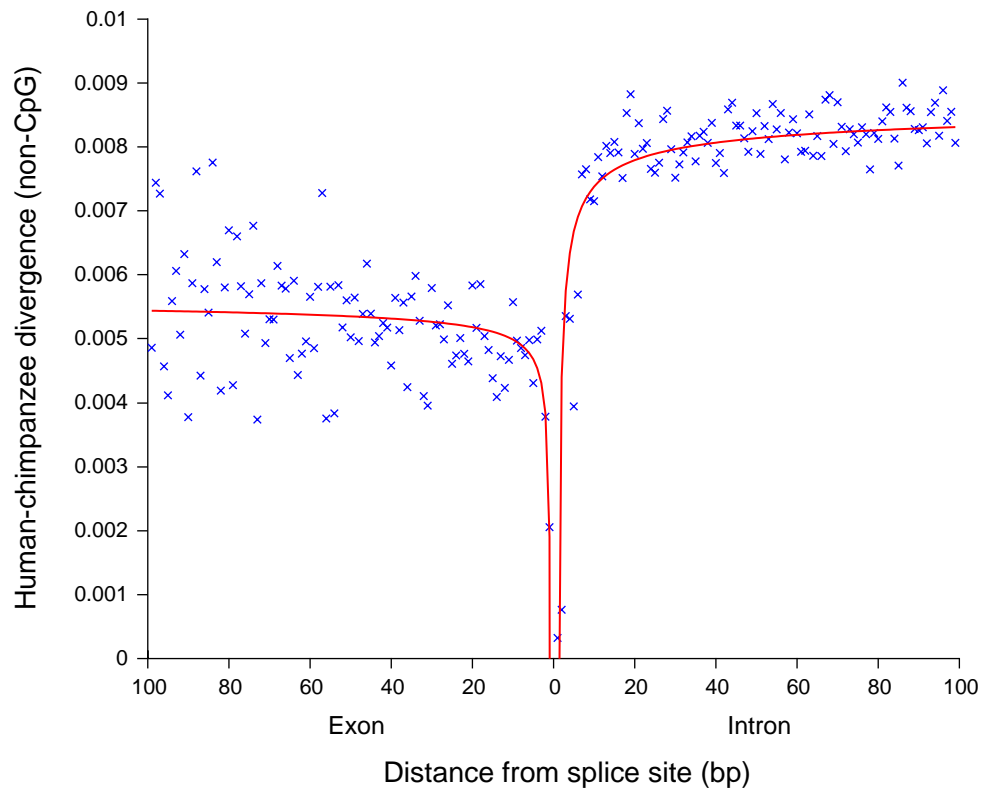


Figure 15. Purifying selection on synonymous sites. Mean divergence around exon boundaries at non-CpG exonic four-fold degenerate sites and intronic sites, relative to the closest mRNA splice junction. The divergence rate at exonic four-fold degenerate sites is significantly lower than at nearby intronic sites ($p[\text{MW}] < 10^{-27}$), suggesting that purifying selection limits the rate of synonymous codon substitutions.

Nonetheless, this set of 585 may be enriched for genes that are under positive selection. The most extreme outliers include glycoporphin C, which mediates one of the *P. falciparum* invasion pathways in human erythrocytes⁹⁹; granulysin, which mediates antimicrobial activity against intracellular pathogens such as *M. tuberculosis*¹⁰⁰; as well as genes that have previously been shown to be undergoing adaptive evolution, such as the protamines and semenogelins involved in reproduction¹⁰¹ and the Mas-related gene family involved in nociception¹⁰². With similar follow up studies on candidates from this list, one may be able to draw conclusions about positive selection on other individual genes. In subsequent sections, we examine the rate of divergence for sets of related genes with the aim of detecting subtler signals of accelerated evolution.

Variation in Evolutionary Rate Across Physically Linked Genes

We explored how the rate of evolution varies regionally across the genome. Several studies of mammalian gene evolution have noted that the rate of amino acid substitution shows local clustering, with proteins encoded by nearby genes evolving at correlated rates^{16,103-105}.

Variation across chromosomes. Navarro and Barton¹⁰⁶ recently reported, based on the analysis of ~100 genes, that the normalized rate of protein evolution is greater on the nine chromosomes that underwent major structural rearrangement during human evolution (chromosomes 1, 2, 5, 9, 12, 15, 16, 17, 18); they suggested that such rearrangements led to reduced gene flow and accelerated adaptive evolution. A subsequent study of a collection of chimpanzee ESTs gave contradictory results^{107,108}. With our larger data set, we re-examined this issue and found no evidence of accelerated evolution on chromosomes with major rearrangements, even if we considered each rearrangement separately.

Among all chromosomes, the most extreme outlier is chromosome X with a mean K_a/K_i of 0.32. The higher mean appears to reflect a skewed distribution at both high and low values, with the median value (0.17) being more in line with other chromosomes (0.15). The excess of low values may reflect greater purifying selection at some genes, owing to the hemizyosity of chromosome X in males. The excess of high values may reflect increased adaptive selection also resulting from hemizyosity, if a considerable proportion of advantageous alleles are recessive¹⁰⁹. Interestingly, the higher K_a/K_i on X versus autosomes is largely restricted to genes expressed in testis⁸¹.

Variation in local gene clusters. We next searched for genomic neighborhoods with an unusually high density of rapidly evolving genes. Specifically, we calculated the median K_a/K_i for sliding windows of 10 orthologs and identified extreme outliers ($p < 0.001$ compared to random ordering of genes; see Supplementary Information). A total of 16 such neighborhoods were found,

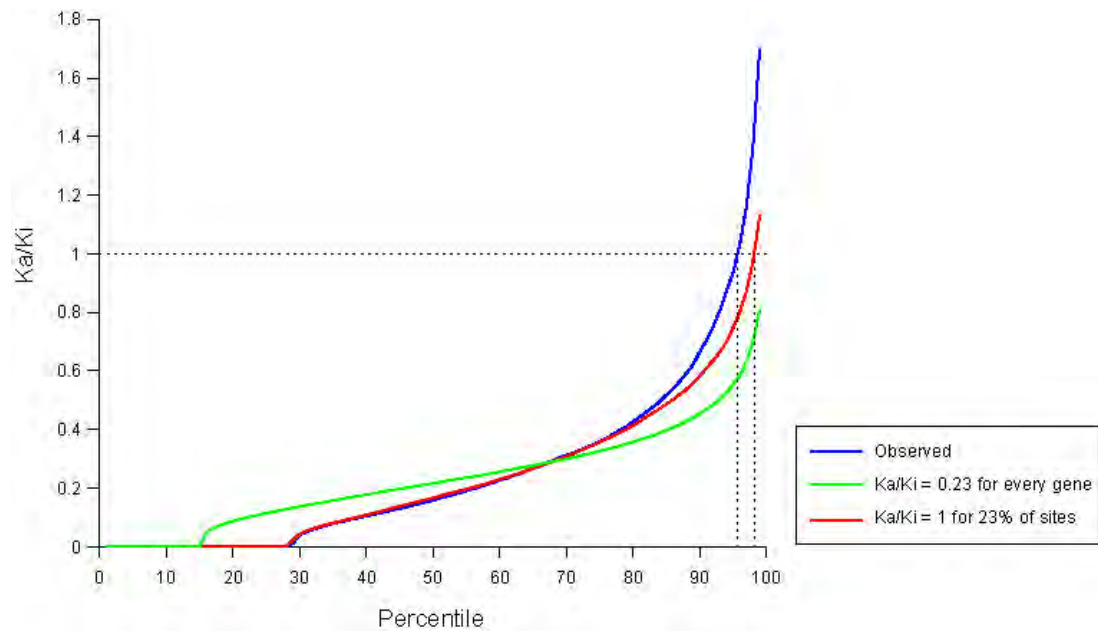


Figure 16. Cumulative distribution of Ka/Ki values for 13,454 orthologs as observed (blue), as expected if all orthologs evolved at $Ka/Ki = 0.23$ (green) and as expected if 23% of codons evolved at $Ka/Ki = 1$ and the rest at $Ka/Ki = 0$ (red). There is a small excess of orthologs with $Ka/Ki > 1$ in the observed distribution, possibly indicating an enrichment of genes under strong positive selection.

Table 4 Rapidly diverging gene clusters in human and chimpanzee

Location		Median
(HSA)	Cluster	K_d/K_i ¹
1q21	Epidermal differentiation complex	1.46
6p22	Olfactory receptors and HLA-A	0.96
20p11	Cystatins	0.94
19q13	Pregnancy-specific glycoproteins	0.94
17q21	Hair keratins and keratin associated proteins	0.93
19q13	CD33-related Siglecs	0.90
20q13	WAP domain protease inhibitors	0.90
22q11	Immunoglobulin lambda/Breakpoint Critical Region	0.85
12p13	Taste receptors, type 2	0.81
17q12	Chemokine (C-C motif) ligands	0.81
19q13	Leukocyte-associated Ig-like receptors	0.80
5q31	Protocadherin-beta	0.77
1q32	Complement component 4-binding proteins	0.76
21q22	Keratin associated proteins and uncharacterized ORFs	0.76
1q23	CD1 antigens	0.72
4q13	Chemokine (C-X-C motif) ligands	0.70

¹ Maximum median K_d/K_i if the cluster stretched over more than one window of 10 genes.

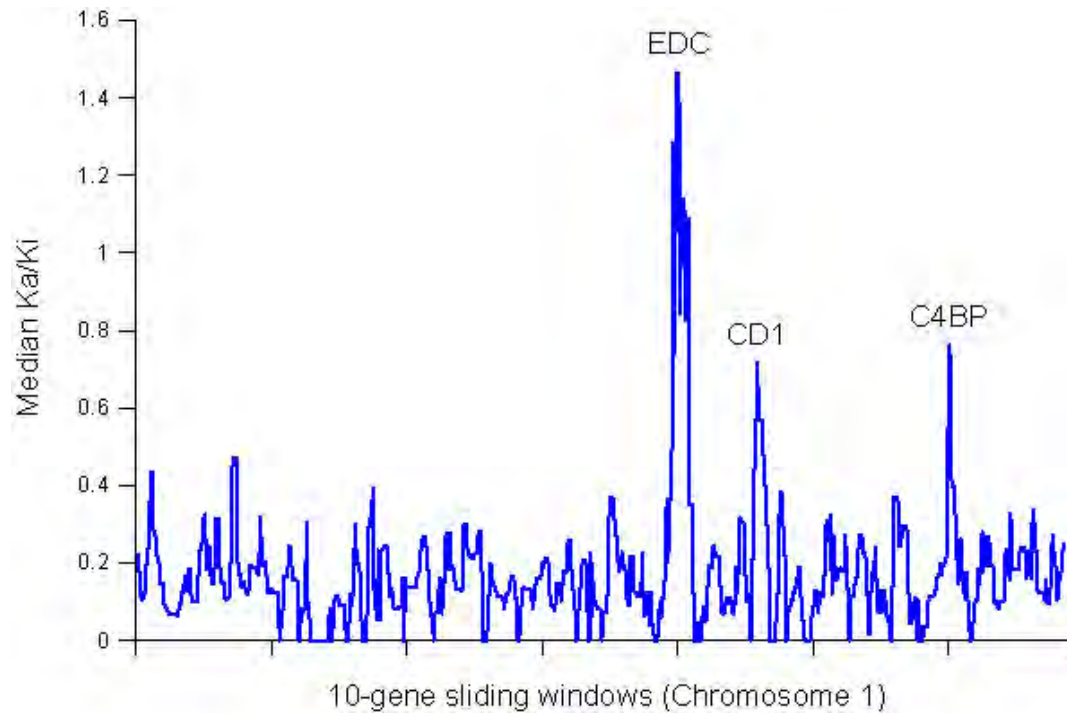


Figure 17. Median Ka/Ki over sliding 10 gene windows across chromosome 1. Three peaks, corresponding to rapidly evolving gene clusters, are visible.

which greatly exceeds random expectation (Table 4). Repeating the analysis with larger windows (25, 50 and 100 orthologs) did not identify additional rapidly diverging regions.

In nearly all cases, the regions contain local clusters of phylogenetically and functionally related genes. The rapid diversification of gene families, postulated by Ohno (1970), can thus be readily discerned even at the relatively close distance of human-chimpanzee divergence. Most of the clusters are associated with functional categories such as host defense and chemosensation (see below). Examples include the epidermal differentiation complex encoding proteins that help form the cornified layer of the skin barrier (Figure 17), the WAP-domain cluster encoding secreted protease inhibitors with antibacterial activity and the Siglec cluster encoding *CD33*-related genes. Rapid evolution in these clusters does not appear to be unique to either human or chimpanzee^{110,111}.

Variation in Evolutionary Rate Across Functionally Related Genes

We next studied variation in the evolutionary rate of functional categories of genes, based on the Gene Ontology (GO) classification¹¹².

Rapidly and slowly evolving categories within the hominid lineage. We started by searching for sets of functionally related genes with exceptionally high or low constraint in humans and chimpanzees. For each of the 809 categories with at least 20 genes, K_a/K_s was calculated by concatenating the gene sequences. The category-specific ratios were compared to the average across all orthologs to identify extreme outliers using a metric based on the binomial test (see Supplementary Information). The numbers of observed outliers below a specific threshold (test statistic < 0.001) were then compared to the expected distribution of outliers given randomly permuted annotations.

A total of 98 categories showed elevated K_a/K_s ratios at the specified threshold (Table 5). Only 30 would be expected by chance, indicating that most (but not all) of these categories undergo significantly accelerated evolution relative to the genome-wide average ($p < 10^{-4}$). The rapidly evolving categories within the hominid lineage are primarily related to immunity and host defense, reproduction and olfaction, which are the same categories known to be undergoing rapid evolution within the broader mammalian lineage, as well as more distantly related species^{15,16,113}. Hominids thus appear to be typical of mammals in this respect (but see below).

A total of 251 categories showed significantly low K_a/K_s ratios (versus ~32 expected by chance; $p < 10^{-4}$). These include a wide range of processes including intracellular signaling, metabolism, neurogenesis and synaptic transmission, which are evidently under stronger-than-

average purifying selection. More generally, genes expressed in the brain show significantly stronger average constraint than genes expressed in other tissues⁸¹.

Differences between hominid and murid lineages. Having found gene categories that show substantial variation in *absolute* evolutionary rate within hominids, we next examined variation in *relative* rates between murids and hominids. The K_a/K_s of each of the GO categories are highly correlated between the hominid and murid ortholog pairs, suggesting that the selective pressures acting on particular functional categories have been largely proportional in recent hominid and recent murid evolution (Figure 18). However, there are several categories with significantly accelerated non-synonymous divergence on each of the lineages, which could represent functions that have undergone lineage-specific positive selection or a lineage-specific relaxation of constraint (Supplementary Notes).

A total of 59 categories (versus 11 expected at random, $p < 0.0003$) show evidence of accelerated non-synonymous divergence in the murid lineage. These are dominated by functions and processes related to host defense, such as immune response and lymphocyte activation. Examples include genes encoding interleukins and various T-cell surface antigens (*CD4*, *CD8*, *CD80*). Combined with the recent observation that genes involved in host defense have undergone gene family expansion in murids^{16,17}, this suggests that the immune system has undergone extensive lineage-specific innovation in murids. Additional categories that also show relative acceleration in murids include chromatin-associated proteins and proteins involved in DNA repair. These categories may have similarly undergone stronger adaptive evolution in murids or, alternatively, they may contain fewer sites for mutations with slightly deleterious effects (with the result that the K_a/K_s ratios are less affected by the differences in population size^{94,114}).

Another 58 categories (versus 14 expected at random, $p < 0.0005$) show evidence of accelerated evolution in hominids, with the set dominated by genes encoding proteins involved in transport (e.g. ion transport), synaptic transmission, spermatogenesis and perception of sound (Table 6). Notably, some outliers include genes with brain related functions, compatible with a recent finding¹¹⁵. However, as above, it is possible that these categories could have more sites for slightly deleterious mutations and thus be more affected by population size differences. Sequence information from more species and from individuals within species will be necessary to distinguish between the possible explanations.

Table 5 GO categories with the highest divergence rates in hominids

	<i>N</i>	<i>AA</i>	<i>K_a/K_s</i>
GO:0007606 Sensory perception of chemical stimulus	59	0.018	0.590
GO:0007608 Perception of smell	41	0.018	0.521
GO:0006805 Xenobiotic metabolism	40	0.013	0.432
GO:0006956 Complement activation	22	0.013	0.428
GO:0042035 Regulation of cytokine biosynthesis	20	0.011	0.402
GO:0007565 Pregnancy	34	0.014	0.384
GO:0007338 Fertilization	24	0.010	0.371
GO:0008632 Apoptotic program	36	0.010	0.358
GO:0007283 Spermatogenesis	80	0.008	0.354
GO:0000075 Cell cycle checkpoint	27	0.006	0.354

AA = Amino acid divergence. N = number of orthologs. Listed are the 10 categories in the taxonomy 'biological process' with the highest K_a/K_s ratios, which are not significant solely due to significant subcategories.

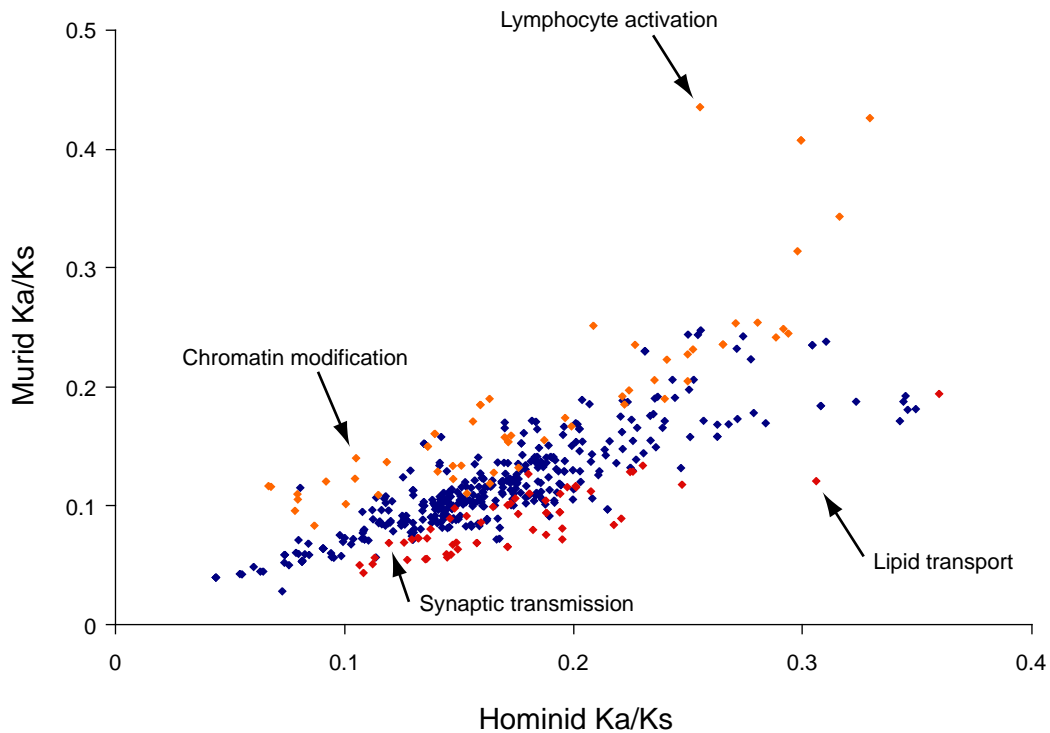


Figure 18. Hominid and murid Ka/Ks (ω) in GO categories with more than 20 analyzed genes. GO categories with putatively accelerated (test statistic < 0.001 ; see Methods) non-synonymous divergence on the hominid lineages (red) and on the murid lineages (orange) are highlighted. Due to the hierarchical nature of GO, the categories do not all represent independent data points. A non-redundant list of significant categories is provided in Table 8.

Table 6: GO categories with accelerated divergence rates in hominids relative to murids

	<i>N</i>	<i>AA</i> hominids	<i>AA</i> murids	<i>K_a/K_s</i> hominids	<i>K_a/K_s</i> murids
GO:0007283 Spermatogenesis	43	0.0075	0.054	0.323	0.188
GO:0006869 Lipid transport	22	0.0081	0.051	0.306	0.120
GO:0006865 Amino acid transport	24	0.0058	0.033	0.218	0.084
GO:0015698 Inorganic anion transport	29	0.0061	0.027	0.195	0.072
GO:0006486 Protein amino acid glycosylation	50	0.0056	0.040	0.166	0.100
GO:0019932 Second-messenger-mediated signaling	58	0.0049	0.036	0.159	0.083
GO:0007605 Perception of sound	28	0.0052	0.033	0.158	0.085
GO:0016051 Carbohydrate biosynthesis	27	0.0047	0.028	0.147	0.067
GO:0007268 Synaptic transmission	93	0.0040	0.025	0.126	0.069
GO:0006813 Potassium ion transport	65	0.0035	0.022	0.113	0.056

AA = Amino acid divergence. N = number of orthologs. Listed are the 10 categories in the taxonomy biological process with the strongest evidence for accelerated evolution in hominids relative to murids, which are not significant solely due to significant subcategories.

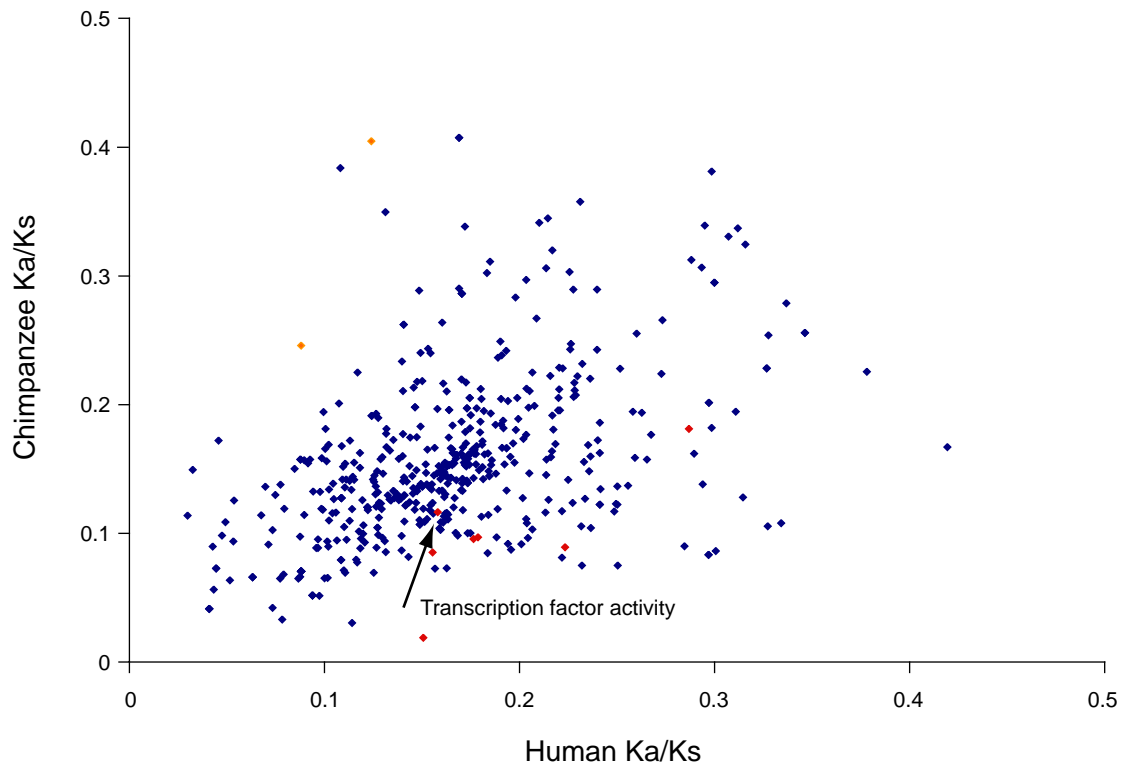


Figure 19. Human and chimpanzee Ka/Ks (ω) in GO categories with more than 20 analyzed genes. GO categories with putatively accelerated (test statistic < 0.001 ; see Methods) non-synonymous divergence on the human lineage (red) and on the chimpanzee lineage (orange) are highlighted. The variance of these estimates are larger than that seen in the hominid-murid comparison due to the small number of lineage-specific substitutions. Due to the hierarchical nature of the GO ontology, the categories do not all represent independent data points.

Differences between human and chimpanzee lineage. One of the most interesting questions is perhaps whether certain categories have undergone accelerated evolution in humans relative to chimpanzees, since such genes might underlie unique aspects of human evolution.

As was done for hominids and murids above, we compared non-synonymous divergence for each category to search for relative acceleration in either lineage (Figure 19). Seven categories show signs of accelerated evolution on the human lineage relative to chimpanzee, but this is only slightly more than the 4 expected at random ($p < 0.22$). Intriguingly, the single strongest outlier is ‘transcription factor activity’, with the 348 human genes studied having accumulated 47% more amino acid changes than their chimpanzee orthologs. Genes with accelerated divergence in human include homeotic, forkhead and other transcription factors that play key roles in early development. However, given the small number of changes involved, additional data will be required to confirm this trend. There was no excess of accelerated categories on the chimpanzee lineage.

We also compared human genes with and without disease associations, including mental retardation, for differences in mutation rate when compared to chimpanzee. Briefly, no significant differences were observed in either the background mutation rate or in the ratio of human-specific changes to chimpanzee specific amino acid changes (see Supplemental Notes).

We thus find minimal evidence of acceleration unique to either the human or chimpanzee lineage across broad functional categories. This is not entirely due to general lack of power resulting from the small number of changes since the divergence of human and chimpanzee, because one can detect acceleration of categories in either hominid relative to either murid. For example, 29 accelerated categories versus 9 expected at random ($p < 0.02$) can be detected on the human lineage, and 40 categories versus 11 expected at random ($p < 0.007$) on the chimpanzee lineage, relative to mouse. But the outliers are largely the same for both human and chimpanzee, indicating that the fraction of amino acid mutations that have contributed to human- and chimpanzee-specific patterns of evolution must be small relative to the fraction that have contributed to a common hominid and, to a large extent, mammalian pattern of evolution.

Clark and colleagues¹⁰ recently reported that numerous gene categories (including signal transduction, ion transport and hearing) are under accelerated positive selection in the human lineage relative to the chimpanzee lineage. Their analysis used a likelihood approach to detect codons under positive selection in human or chimpanzee coding sequences, using mouse as an outgroup. However, this approach may be highly sensitive to false positives in the presence of relaxed constraints in the hominid lineages¹¹⁶, or potential alignment artifacts. As shown above, with the potential exception of some developmental regulators, the categories that Clark *et al.*

reported as showing the strongest evidence of human-specific positive selection are among those that show the highest accelerated divergence in both human and chimpanzee. This suggests that their results may be enriched for false positives due to relaxed constraints, rather than human-specific positive selection. Additional data may enhance the statistical power of such analysis, but, at present, strong evidence of positive selection unique to the human lineage is limited to a handful of genes¹¹⁷.

Our analysis above largely omitted genes belonging to large gene families, because gene family expansion makes it difficult to define 1:1:1:1 orthologs across hominids and murids. One of the largest such family, the olfactory receptors, is known to be undergoing rapid divergence in primates. Directed study of these genes in the draft assembly has suggested that more than a hundred functional human olfactory receptors are likely to be under no evolutionary constraint¹¹⁸. Our analysis also omitted the majority of very recently duplicated genes due to their lower coverage in the current chimpanzee assembly. However, recent human-specific duplications can be readily identified from the finished human genome sequence, and have previously been shown to be highly enriched for the same categories found to have high absolute rates of evolution in 1:1 orthologs here, i.e. olfaction, immunity and reproduction²³.

Gene disruptions in Human and Chimpanzee

Whereas most genes have undergone only subtle substitutions in their amino acid sequence, a few dozen have suffered more drastic changes. We found a total of 53 known or predicted human genes that are either deleted entirely (36) or partially (17) in chimpanzee. We have so far tested and confirmed 15 of these cases by PCR or Southern blotting. An additional 8 genes have sustained large deletions (>15 kb) entirely within an intron. Some genes may have been missed in this count, owing to limitations of the draft genome sequence. In addition, some genes may have suffered chain termination mutations or altered reading frames in chimpanzee, but accurate identification of these will require higher-quality sequence. The sensitivity of the reciprocal analysis of genes disrupted in human is currently limited by the small number of independently predicted gene models for the chimpanzee¹¹⁹. Some of the gene disruptions may be related to interesting biological differences between the species, as discussed below.

Genetic Basis for Human- and Chimpanzee-specific Biology

Given the substantial number of neutral mutations, only a small subset of the observed gene differences is likely to be responsible for the key phenotypic changes in morphology, physiology

and behavioral complexity between humans and chimpanzees. Determining which differences are in this evolutionarily important subset and inferring their functional consequences will require additional types of evidence, including information from clinical observations and model systems¹²⁰. We describe some novel examples of genetic changes for which plausible functional or physiological consequences can be suggested.

Apoptosis. Mouse and human are known to differ with respect to an important mediator of apoptosis, *caspase-12*^{121,122 123}. The protein triggers apoptosis in response to perturbed calcium homeostasis in mice, but humans appear to lack this activity owing to several mutations in the orthologous gene that together affect the protein produced by all known splice forms; the mutations include a premature stop codon and a disruption of the SHG box required for enzymatic activity of caspases. By contrast, the chimpanzee gene encodes an intact open-reading frame and SHG box, indicating that the functional loss occurred in the human lineage. Intriguingly, loss-of-function mutations in mice confer increased resistance to amyloid-induced neuronal apoptosis without causing obvious developmental or behavioral defects¹²⁴. The loss-of-function in humans may contribute to the human-specific pathology of Alzheimer's disease, which involves amyloid-induced neurotoxicity and deranged calcium homeostasis.

Inflammatory Response. Human and chimpanzee show a notable difference with respect to important mediators of immune and inflammatory response. Three genes (*IL-1F7*, *IL-1F8* and *ICEBERG*) that act in a common pathway involving *caspase-1* all appear to be deleted in chimpanzee. *ICEBERG* is thought to repress *caspase-1* mediated generation of pro-inflammatory *IL-1* cytokines, and its absence in chimpanzee may point to species-specific modulation of the interferon-gamma and lipopolysaccharide-induced inflammatory response¹²⁵.

Parasite resistance. Similarly, we found that two members of the primate-specific APOL gene cluster (*APOLI* and *APOLA*) have been deleted from the chimpanzee genome. The Apol1 protein is associated with the high-density lipoprotein fraction in serum and has recently been proposed to be the lytic factor responsible for resistance to certain subspecies of *Trypanosome brucei*, the parasite that causes human sleeping sickness and the veterinary disease *nagana*¹²⁶. The loss of the *APOLI* gene in chimpanzees could thus explain the observation that human, gorilla and baboon possess the trypanosome lytic factor, whereas the chimpanzee does not¹²⁷.

Sialic acid biology-related proteins. Sialic acids are cell surface sugars that mediate many biological functions¹²⁸. Of 54 genes involved in sialic acid-biology, 47 were suitable for analysis. We confirmed and extended findings on several that have undergone human-specific changes, including disruptions, deletions and domain-specific functional changes^{110,119,129}. Human- and

chimpanzee-specific changes were also found in otherwise evolutionarily conserved sialyl motifs in four sialyl transferases (*ST6GAL1*, *ST6GALNAC3*, *ST6GALNAC4*, and *ST8SIA2*), suggesting changes in donor and/or acceptor binding¹²⁹. Lineage-specific changes were found in a complement factor H (*HF1*) sialic acid domain binding associated with human disease¹³⁰. Human *SIGLEC11* has undergone gene conversion with a nearby pseudogene, which correlates with the acquisition of human-specific expression and binding properties¹⁷¹.

Human disease alleles

We next sought to identify putative functional differences between the species by searching for instances in which a human disease-causing allele appears to be the wild-type allele in the chimpanzee. Starting from 12,164 catalogued disease variants in 1,384 human genes, we identified 16 cases in which the altered sequence in a disease allele matched the chimpanzee sequence, and had plausible support in the literature (Table 7). Upon resequencing in seven chimpanzees, 15 cases were confirmed homozygous in all individuals, whereas one (*PONI* I102V) appears to be a shared polymorphism.

Six cases represent *de novo* human mutations associated with simple Mendelian disorders. Similar cases have also been found in comparisons of more distantly related mammals¹³¹, as well as between insects¹³², and have been interpreted as a consequence of a relatively high rate of compensatory mutations. If compensatory mutations are more likely to be fixed by positive selection than by neutral drift¹³², then the variants identified here might point towards adaptive differences between humans and chimpanzees. For example, the ancestral Thr29 allele of cationic trypsinogen (*PRSSI*) causes autosomal dominant pancreatitis in humans¹³³, suggesting that the human-specific Asn29 allele may represent a digestion-related molecular adaptation¹³⁴.

The remaining 10 cases represent common human polymorphisms that have been reported as associated with complex traits, including coronary artery disease and diabetes mellitus. In all of these cases, we confirmed that the disease-associated allele in humans is indeed the ancestral allele by showing that it is carried not only by chimpanzee but also by outgroups such as the macaque. These ancestral alleles may thus have become human-specific risk factors due to changes in human physiology or environment, and the polymorphisms may represent ongoing adaptations. For example, *PPARG* Pro12 is the wild-type allele in chimpanzee but has been clearly associated with increased risk of type 2 diabetes in human¹³⁵. It is tempting to speculate that this allele may represent an ancestral ‘thrifty’ genotype¹³⁶.

Table 7 Candidate human disease variants found in chimpanzee

Gene	Variant ¹	Disease association	Ancestral ²	Freq. ³
<i>AIRE</i>	P252L ¹⁵⁵	Autoimmune syndrome	Unresolved	0
<i>MKKS</i>	R518H ¹⁵⁶	Bardet-Biedl syndrome	Wildtype	0
<i>MLH1</i>	A441T ¹⁵⁷	Colorectal cancer	Wildtype	0
<i>MYOC</i>	Q48H ¹⁵⁸	Glaucoma	Wildtype	0
<i>OTC</i>	T125M ¹⁵⁹	Hyperammonemia	Wildtype	0
<i>PRSS1</i>	N29T ¹³³	Pancreatitis	Disease	0
<i>ABCA1</i>	I883M ¹⁶⁰	Coronary artery disease	Unresolved	0.136
<i>APOE</i>	C130R ¹⁶¹	Coronary artery disease and Alzheimer's disease	Disease	0.15
<i>DIO2</i>	T92A ¹⁶²	Insulin resistance	Disease	0.35
<i>ENPP1</i>	K121Q ¹⁶³	Insulin resistance	Disease	0.17
<i>GSTP1</i>	I105V ¹⁶⁴	Oral cancer	Disease	0.348
<i>PON1</i> ⁴	I102V ¹⁶⁵	Prostate cancer	Wildtype	0.016
<i>PON1</i>	Q192R ¹⁶⁶	Coronary artery disease	Disease	0.3
<i>PPARG</i>	A12P ¹³⁵	Type 2 Diabetes	Disease	0.85
<i>SLC2A2</i>	T110I ¹⁶⁷	Type 2 Diabetes	Disease	0.12
<i>UCP1</i>	A64T ¹⁶⁸	Waist-to-hip ratio	Disease	0.12

¹ Benign variant, codon, disease/chimpanzee variant. ² Ancestral variant as inferred from closest available primate outgroups (Supplementary Information). ³ Frequency of the disease allele in human study population. ⁴ Polymorphic in chimpanzee.

The current results must be interpreted with caution, because few complex disease associations have been firmly established. The fact that the human disease allele is the wild-type allele in chimpanzee may actually indicate that some of the putative associations are spurious and not causal. However, this approach can be expected to become increasingly fruitful as the quality and completeness of the disease mutation databases improve.

Human Population Genetics

The chimpanzee plays a special role in informing studies of human population genetics, a field that is undergoing rapid expansion and acquiring new relevance to human medical genetics¹³⁷. The chimpanzee sequence allows recognition of those human alleles that represent the ancestral state and the derived state. It also allows estimates of local mutation rates, which serve as an important baseline in searching for signs of natural selection.

Ancestral and derived alleles. Of ~7.2 M SNPs mapped to the human genome in the current public database, we could assign the alleles as ancestral or derived in 80% of the cases according to which allele agrees with the chimpanzee genome sequence¹³⁸ (see Supplementary Notes). In remaining cases, no assignment could be made because the orthologous chimpanzee base differed from both human alleles (1.2%); was polymorphic in the chimpanzee sequences obtained (0.4%); or could not be reliably identified with the current draft sequence of the chimpanzee (18.8%), with many of these occurring in repeated or segmentally duplicated sequence. The first two cases arise presumably because a second mutation occurred in the chimpanzee lineage. It should be possible to resolve most of these cases by examining a close outgroup such as gorilla or orangutan.

Mutations in the chimpanzee may also lead to the erroneous assignment of human alleles as derived alleles. This error rate can be estimated as the probability of a second mutation resulting in the chimpanzee sequence matching the derived allele (see Supplementary Notes). The estimated error rate for typical SNPs is 0.5%, owing to the low nucleotide substitution rate. The exceptions are those SNPs for which the human alleles are CpG and TpG and the chimpanzee sequence is TpG. For these, a non-negligible fraction may have arisen by two independent deamination events within an ancestral CpG dinucleotide, which are well known mutational hotspots⁵⁰ (also see above). Human SNPs in a CpG-context for which the orthologous chimpanzee sequence is TpG account for 12% of the total, and have an estimated error rate of 9.8%. Across all SNPs, the average error rate, ϵ , is thus estimated to be ~1.6%.

We compared the distribution of allele frequencies for ancestral and derived alleles, using a database of allele frequencies for ~120,000 SNPs (see Supplementary Notes). As expected,

ancestral alleles tend to have much higher frequencies than derived alleles. Nonetheless, a significant proportion of derived alleles have high frequencies: 9.1% of derived alleles have frequency $\geq 80\%$.

An elegant result in population genetics states that, for a randomly interbreeding population of constant size, the probability that an allele is ancestral is equal to its frequency¹³⁹. We explored the extent to which this simple theoretical expectation fits the human population. We tabulated the proportion $p_a(x)$ of ancestral alleles for various frequencies of x and compared this with the prediction $p_a(x) = x$ (Figure 20).

The data lie near the predicted line, but the observed slope (0.83) is substantially less than 1. One explanation for this deviation is that some ancestral alleles are incorrectly assigned (an error rate of ε would artificially decrease the slope by a factor of $1-2\varepsilon$). However, with ε estimated to be only 1.6%, errors can only explain a small part of the deviation. The most likely explanation is the presence of bottlenecks during human history, which tend to flatten the distribution of allele frequencies. Theoretical calculations indicate that a recent bottleneck would decrease the slope by a factor of $(1-b)$, where b is the inbreeding coefficient induced by the bottleneck. This suggests that measurements of the slope in different human groups may shed light on population-specific bottlenecks. Consistent with this, preliminary analyses of allele frequencies in several regions for SNPs obtained by systematic uniform sampling indicate that the slope is significantly lower than 1 in European and Asian samples and close to 1 in an African sample (see Supplementary Notes).

Signatures of strong selective sweeps in recent human history. The pattern of human genetic variation holds substantial information about selection events that have shaped our species. Strong positive selection creates the distinctive signature of a ‘selective sweep’, whereby a rare allele rapidly rises to fixation and carries the haplotype on which it occurs to high frequency (the ‘hitchhiking’ effect). The surrounding region should show two distinctive signatures: (i) a significant reduction of overall diversity and (ii) an excess of derived alleles with high frequency in the population, owing to hitchhiking of derived alleles on the selected haplotype (see Supplementary Information). The pattern might be detectable up for 250,000 years after a selective sweep has ended¹⁴⁰. Notably, the chimpanzee genome provides crucial baseline information required for accurate assessment of both signatures.

The size of the interval affected by a selective sweep is expected to scale roughly with s , the selective advantage due to the mutation. Simulations can be used to study the distribution of the interval size (see Supplementary Notes). With $s = 1\%$, the interval over which heterozygosity falls by 50% has modal size of 600kb and a probability of greater than 10% of exceeding 1Mb.

We undertook an initial scan for large regions (> 1 Mb) with the two signatures suggestive of strong selective sweeps in recent human history. We began by identifying regions in which the observed human diversity rate was much lower than the expectation based on the observed divergence rate with chimpanzee. The human diversity rate was measured as the number of occurrences from a database of 1.92 M SNPs identified by shotgun sequencing in a panel of African-American individuals (see Supplementary Information). The comparison with the chimpanzee eliminates regions in which low diversity simply reflects a low mutation rate in the region. Regions were identified based on a simple statistical procedure (see Supplementary Notes). Six genomic regions stood out as clear outliers that show significantly reduced diversity relative to divergence (Table 8).

We next tested whether these six regions show a high proportion of SNPs with high-frequency derived alleles (defined here as alleles with frequency $\geq 80\%$). Within each region, we focused on the 1 Mb interval with the greatest discrepancy between diversity and divergence and compared it to 1 Mb regions throughout the genome. For the database of 120,000 SNPs with allele frequencies discussed above, the typical 1 Mb region in the human genome contains ~ 40 SNPs and the proportion p_h of SNPs with high-frequency derived alleles is $\sim 9.1\%$. All six regions identified by our scan for reduced diversity have a higher than average fraction of high frequency derived alleles; all six fall within the top 10% genome-wide and three fall within the top 1%. Although this is not definitive evidence for any particular region, the joint probability of all six regions randomly scoring in the top 10% is 10^{-6} . The results suggest the six regions as candidates for strong selective sweeps during the past quarter-million years¹⁴⁰. The regions differ notably with respect to gene content, ranging from one containing 57 annotated genes (chromosome 22) to another with no annotated genes whatsoever (chromosome 4). We have no evidence to implicate any individual functional element as a target of recent selection at this point, but the regions contain a number of interesting candidates for follow-up studies. Intriguingly, the chromosome 4 gene desert, which is conserved across vertebrates¹⁵, has been implicated in two independent studies as being associated with obesity^{141,142}.

In addition to the six regions, one further genomic region deserves mention: an interval of 7.6 Mb on chromosome 7q (see Supplementary Notes). The interval contains several regions with high scores in the diversity-divergence analysis (including the seventh highest score overall) as well as in the proportion of high-frequency derived alleles. The region contains the *FOXP2* and *CFTR* genes. The former has been the subject of much interest as a possible target for selection during human evolution¹⁴³ and the latter as a target of selection in European populations¹⁴⁴.

Convincing proof of past selection will require careful analysis of the precise pattern of genetic variation in the region and the identification of a likely target of selection. Nonetheless, our findings suggest that the approach outlined here may help to unlock some of the secrets of recent human evolution through a combination of within-species and cross-species comparison.

Table 8 Human regions with strongest signal of selection based on diversity relative to divergence

Chr	Start(Mb)	End(Mb)	Score	Skew p-value	Genes
1	48.58	52.58	103.3	0.071	14 known genes from ELAVL4 to GPX7
2	144.35	148.47	84.8	0.074	ARHGAP15 (partial), GTDC1, and ZFHX1B
22	36.15	40.22	81.8	0.00022	57 known genes from CARD10 to PMM1
12	84.69	89.01	80.9	0.031	10 known genes from PAMCI to ATP2B1
8	34.91	37.54	76.9	0.00032	UNC5D and FKSG2
4	32.42	35.62	55.9	0.00067	No known genes or Ensembl predictions

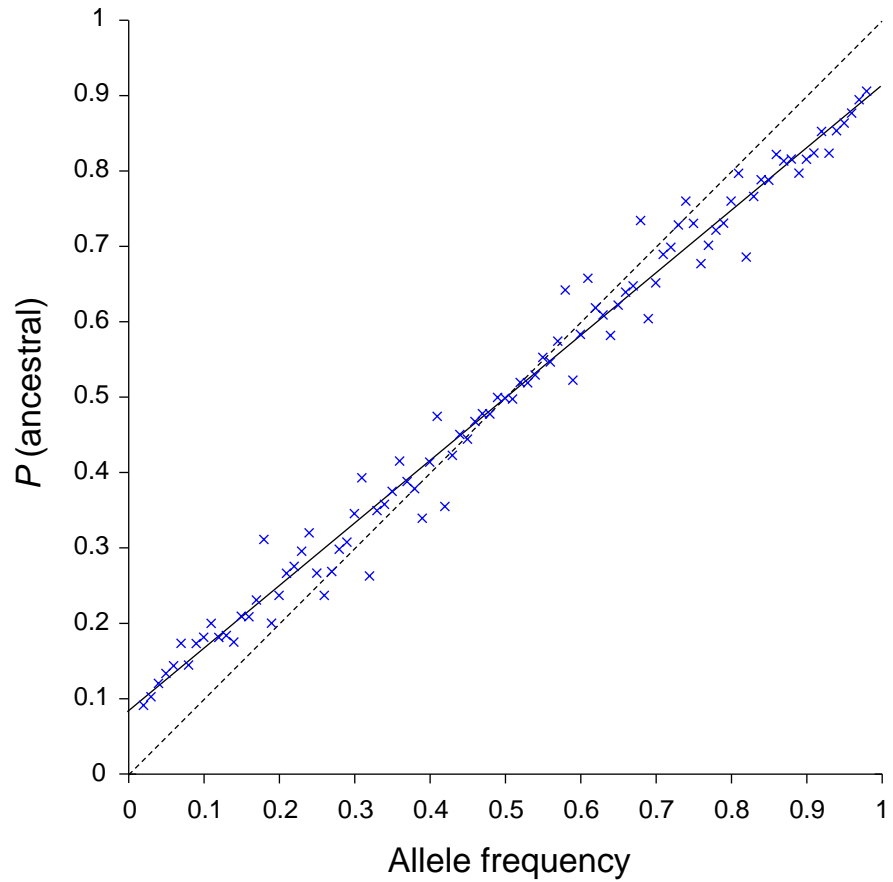


Figure 20. The observed fraction of ancestral alleles in 1% bins of observed frequency. The solid line shows the regression ($b = 0.83$). The dotted line shows the theoretical relationship $p_a(x) = x$. Note that because each variant yields a derived and an ancestral allele, the data are necessarily symmetric about 0.5.

Discussion

Our knowledge of the human genome is greatly advanced by the availability of a second hominid genome. Some questions can be directly answered by comparing the human and chimpanzee sequences, including estimates of regional mutation rates and average selective constraints on gene classes. Other questions can be addressed in conjunction with other large datasets, such as issues in human population genetics for which the chimpanzee genome provides crucial controls. For still other questions, the chimpanzee genome simply provides a starting point for further investigation.

The hardest such question is: what makes us human? The challenge lies in the fact that most evolutionary change is due to neutral drift. Adaptive changes comprise only a small minority of the total genetic variation between two species. As a result, the extent of phenotypic variation between organisms is not strictly related to the degree of sequence variation. For example, gross phenotypic variation between human and chimpanzee is much greater than between the mouse species *M. musculus* and *M. spretus*, although the sequence difference in the two cases is similar. On the other hand, dogs show considerable phenotypic variation despite having little overall sequence variation (~0.15%). Genomic comparison dramatically narrows the search for the functionally important differences between species, but specific biological insights will be needed to sift the still-large list of candidates to separate adaptive changes from neutral background.

Our comparative analysis suggests that the patterns of molecular evolution in the hominids are typical of a broader class of mammals in many ways, but distinctive in certain respects. As with the murids, the most rapidly evolving gene families are those involved in reproduction and host defense. In contrast to the murids, however, hominids appear to experience substantially weaker negative selection; this likely reflects their smaller population size. Consequently, hominids accumulate deleterious mutations that would be eliminated by purifying selection in murids. This may be both an advantage and a disadvantage. Although decreased purifying selection may tend to erode overall fitness, it may also allow hominids to ‘explore’ larger regions of the fitness landscape and thereby achieve evolutionary adaptations that can only be reached by passing through intermediate states of inferior fitness^{146,147}.

Although the analyses presented here focus on protein-coding sequences, the chimpanzee genome sequence also allows systematic analysis of the recent evolution of gene regulatory elements for the first time. Initial analysis of both gene expression patterns and promoter regions suggest that their overall patterns of evolution closely mirror that of protein-coding regions. In a companion paper⁸¹, we show that the rates of change in gene expression among different tissues in human and chimpanzee correlate with the nucleotide divergence in the putative proximal promoters

and even more interestingly with the average level of constraint on proteins in the same tissues. Keightley and colleagues¹⁴⁵ have similarly used the chimpanzee sequence described here to show that gene promoter regions are also evolving under dramatically less constraint in hominids than in murids.

The draft chimpanzee sequence here is sufficient for initial analyses, but it is still imperfect and incomplete. Definitive studies of gene and genome evolution – including pseudogene formation, gene family expansion and segmental duplication – will require high-quality finished sequence. In this regard, we note that efforts are already underway to construct a BAC-based physical map and to increase the shotgun sequence coverage to ~6-fold redundancy. The added coverage alone will not impact the analysis greatly, but plans are in place to produce finished sequence for difficult and important segments of the genome.

Finally, our close biological relatedness to chimpanzees not only allows unique insights into human biology, it also creates ethical obligations. Although the genome sequence was acquired without harm to chimpanzees, the availability of the sequence may increase pressure to use chimpanzees in experimentation. We strongly oppose reducing the protection of chimpanzees and instead advocate the policy positions suggested by Gagneux and colleagues elsewhere in this issue¹⁴⁸. Furthermore, the existence of chimpanzees and other great apes in their native habitats is increasingly threatened by human civilization. More effective policies are urgently needed to protect them in the wild. We hope that elaborating how few differences separate our species will broaden recognition of our duty to these extraordinary primates who stand as our siblings in the family of life.

Methods

Sequencing and assembly. Approximately 22.5 million sequence reads were derived from both ends of inserts (paired end reads) from 4, 10, 40 and 180 kb clones, all prepared from primary blood lymphocyte DNA. Genomic resources available from the source animal include a lymphoid cell line (S006006) and genomic DNA (NS06006) at Coriell Cell Repositories (<http://locus.umdnj.edu/ccr/>), as well as a BAC library (CHORI-251)¹⁴⁹ (see also Supplementary Notes).

Genome alignment. BLASTZ¹⁵⁰ was used to align non-repetitive chimpanzee regions against repeat-masked human sequence. BLAT¹⁵¹ was subsequently used to align the more repetitive regions. The combined alignments were chained¹⁵² and only best reciprocal alignments were retained further analysis.

Insertions and deletions. Small insertion/deletion events (<15kb) were parsed directly from the BLASTZ genome alignment by counting the number and size of alignment gaps between bases within the same contig. Sites of large-scale insertion/deletion (indels >15kb) were detected from discordant placements of paired sequence reads against the human assembly. Size thresholds were obtained from both human fosmid alignments on human sequence (40 +/- 2.58 kb), and chimpanzee plasmid alignments against human chromosome 21 (4.5 +/-1.84 kb). Indels were inferred by two or more pairs surpassing these thresholds by more than two standard deviations and the absence of sequence data within the discordancy.

Gene annotation. A total of 19,277 human RefSeq transcripts¹⁵³, representing 16,045 distinct genes, were indirectly aligned to the chimpanzee sequence via the genome alignment. After removing low quality sequences and likely alignment artifacts, an initial catalog containing 13,454 distinct 1:1 human-chimpanzee orthologs was created for the analyses described here. A subset of 7,043 of these genes with unambiguous mouse and rat orthologs were realigned using Clustal W¹⁵⁴ for the lineage specific analyses. Updated gene catalogs can be obtained from www.ensembl.org.

Rates of divergence. Nucleotide divergence rates were estimated using baseml with the REV model. Non-CpG rates were estimated from all sites that did not overlap a CG dinucleotide in either human or chimpanzee. K_a and K_s were estimated jointly for each ortholog using codeml with the F3x4 codon frequency model and no additional constraints, except for the comparison of divergent and polymorphic substitutions where K_a/K_s for both was estimated as $(\Delta A/N_a)/(\Delta S/N_s)$ with N_s/N_a , the ratio of synonymous to non-synonymous sites, estimated as 0.36 from the ortholog alignments. Unless otherwise specified, K_a/K_s for a set of genes was calculated by summing the number of substitutions and the number of sites to obtain K_a and K_s for the concatenated set before taking the

ratio. Hominid and murid pair-wise rates were estimated independently from codons aligned across all four species. Human and chimpanzee lineage specific K_a and K_s were estimated on an unrooted tree with both mouse and rat included. Lineage-specific rates were also estimated by parsimony, with essentially identical results (see Supplementary Notes). K_i was estimated from all interspersed repeats within 250kb of the midpoint of each gene.

Accelerated evolution in GO categories. The binomial probability of observing X or more non-synonymous substitutions, given a total of $X + Y$ substitutions and the expected proportion x from all orthologs, was calculated by summing substitutions across the orthologs in each GO category. For the absolute rate test, Y = the number of synonymous substitutions in orthologs in the same category. For the relative rate tests, Y = the number of non-synonymous substitutions on the opposite lineage. Note that this binomial probability is simply a metric designed to identify potentially accelerated categories, it is not a p-value that can be used to directly reject the null hypothesis of no acceleration in that particular category. For each test, the observed number of categories with a binomial probability less than 0.001 was compared to the expected distribution of such outliers by repeating the procedure 10,000 times on randomly permuted GO annotations. The significance of the number of observed outliers n was estimated as the proportion of random trials yielding n or more outliers.

Detection of selective sweeps. The observed number of human SNPS, u_i , human bases, m_i , human-chimpanzee substitutions, v_i , and chimpanzee bases, n_i , within each set of non-overlapping 1 Mb windows along the human genome were used to generate two random numbers, x_i (adjusted human diversity) and y_i (adjusted human-chimpanzee divergence), from the distributions:

$$x_i \sim \text{Beta}(u_i + a, m_i - u_i + b)$$

$$y_i \sim \text{Beta}(v_i + c, n_i - v_i + d)$$

where $a = 1$, $b = 1000$, $c = 1$, and $d = 100$. These numbers were then fit to a linear regression:

$$x|y \sim \text{N}(\alpha_0 + \alpha_1 y, \beta^2)$$

A p-value for each window was calculated for each window based on (x_i, y_i) and the regression line. This was repeated 100 times and the average of the p-values taken as the p-value for diversity given divergence in each window. Overlapping windows with $p < 0.1$ containing at least one window of $p < 0.05$ were coalesced and scored as the sum of their $-\log(p)$ scores.

References

1. Darwin, C. *The decent of man, and selection in relation to sex*. (D Appleton and Company, New York, 1871).
2. Huxley, T. H. *Evidence as to Man's place in Nature* (Williams and Norgate, London, 1863).
3. Goodman, M. The genomic record of Humankind's evolutionary roots. *Am J Hum Genet* 64, 31-9 (1999).
4. Goodall, J. Tool-Using and Aimed Throwing in a Community of Free-Living Chimpanzees. *Nature* 201, 1264-6 (1964).
5. Whiten, A. et al. Cultures in chimpanzees. *Nature* 399, 682-5 (1999).
6. Olson, M. V. & Varki, A. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat Rev Genet* 4, 20-8 (2003).
7. Eyre-Walker, A. & Keightley, P. D. High genomic deleterious mutation rates in hominids. *Nature* 397, 344-7 (1999).
8. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Positive and negative selection on the human genome. *Genetics* 158, 1227-34 (2001).
9. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* 188, 107-16 (1975).
10. Clark, A. G. et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302, 1960-3 (2003).
11. Hellmann, I. et al. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res* 13, 831-7 (2003).
12. Ebersberger, I., Metzler, D., Schwarz, C. & Paabo, S. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* 70, 1490-7 (2002).
13. Watanabe, H. et al. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* 429, 382-8 (2004).
14. Jaillon, O. et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946-57 (2004).
15. Hillier, L. W. et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695-716 (2004).
16. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-562 (2002).
17. Rat Genome Sequencing Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428, 493-521 (2004).
18. McConkey, E. H. Orthologous numbering of great ape and human chromosomes is essential for comparative genomics. *Cytogenet Genome Res* 105, 157-8 (2004).
19. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* 162, 729-73 (1982).
20. Myers, E. W. in *Computing in Science and Engineering* 33-43 (1999).
21. Huang, X., Wang, J., Aluru, S., Yang, S. P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res* 13, 2164-70 (2003).

22. Jaffe, D. B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13, 91-6 (2003).
23. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945 (2004).
24. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 409, 860-920 (2001).
25. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8, 175-85 (1998).
26. She, X. et al. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* 431, 927-30 (2004).
27. Fischer, A., Wiebe, V., Paabo, S. & Przeworski, M. Evidence for a complex demographic history of chimpanzees. *Mol Biol Evol* 21, 799-808 (2004).
28. Yu, N. et al. Low nucleotide diversity in chimpanzees and bonobos. *Genetics* 164, 1511-8 (2003).
29. Kaessmann, H., Wiebe, V., Weiss, G. & Paabo, S. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat Genet* 27, 155-6 (2001).
30. Kitano, T., Schwarz, C., Nickel, B. & Paabo, S. Gene diversity patterns at 10 X-chromosomal loci in humans and chimpanzees. *Mol Biol Evol* 20, 1281-9 (2003).
31. International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928-933 (2001).
32. Chen, F. C. & Li, W. H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* 68, 444-56 (2001).
33. Fujiyama, A. et al. Construction and analysis of a human-chimpanzee comparative clone map. *Science* 295, 131-4 (2002).
34. Hardison, R. C. et al. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* 13, 13-26 (2003).
35. Webster, M. T., Smith, N. G., Lercher, M. J. & Ellegren, H. Gene expression, synteny, and local similarity in human noncoding mutation rates. *Mol Biol Evol* 21, 1820-30 (2004).
36. Rosenberg, H. F. & Feldmann, M. W. *The Relationship Between Coalescence Times and Population Divergence Times* (Oxford Univ. Press, Oxford, 2002).
37. Vignaud, P. et al. Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* 418, 152-5 (2002).
38. Wall, J. D. Estimating ancestral population sizes and divergence times. *Genetics* 163, 395-404 (2003).
39. Reich, D. E. et al. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 32, 135-42 (2002).
40. Maynard Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet Res* 23, 23-35 (1974).
41. Hudson, R. R. & Kaplan, N. L. Deleterious background selection with recombination. *Genetics* 141, 1605-17 (1995).
42. Charlesworth, B. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res* 63, 213-27 (1994).
43. Birky, C. W., Jr. & Walsh, J. B. Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci U S A* 85, 6414-8 (1988).

44. Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S. & Przeworski, M. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* 72, 1527-35 (2003).
45. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* 18, 337-40 (2002).
46. Hellmann, I. et al. Why do human diversity levels vary at a megabase scale? *Genome Res* (in the press).
47. Li, W. H., Yi, S. & Makova, K. Male-driven evolution. *Curr Opin Genet Dev* 12, 650-6 (2002).
48. Bohossian, H. B., Skaletsky, H. & Page, D. C. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* 406, 622-5 (2000).
49. Makova, K. D. & Li, W. H. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* 416, 624-6 (2002).
50. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A* 101, 13994-4001 (2004).
51. Bulmer, M., Wolfe, K. H. & Sharp, P. M. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. *Proc Natl Acad Sci U S A* 88, 5974-8 (1991).
52. Ehrlich, M., Zhang, X. Y. & Inamdar, N. M. Spontaneous deamination of cytosine and 5-methylcytosine residues in DNA and replacement of 5-methylcytosine residues with cytosine residues. *Mutat Res* 238, 277-86 (1990).
53. Craig, J. M. & Bickmore, W. A. Chromosome bands--flavours to savour. *Bioessays* 15, 349-54 (1993).
54. Holmquist, G. P. Chromosome bands, their chromatin flavors, and their functional features. *Am J Hum Genet* 51, 17-37 (1992).
55. Ellegren, H., Smith, N. G. & Webster, M. T. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev* 13, 562-8 (2003).
56. Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S. & Sidow, A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res* 13, 813-20 (2003).
57. Cooper, G. M. et al. Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res* 14, 539-48 (2004).
58. Yang, S. et al. Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. *Genome Res* 14, 517-27 (2004).
59. Birdsell, J. A. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.* 19, 1181-1197 (2002).
60. Jensen-Seaman, M. I. et al. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* 14, 528-38 (2004).
61. Fortna, A. et al. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* 2, E207 (2004).
62. Britten, R. J. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc Natl Acad Sci U S A* 99, 13633-5 (2002).
63. Frazer, K. A. et al. Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res* 13, 341-6 (2003).
64. Locke, D. P. et al. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res* 13, 347-57 (2003).
65. Liu, G. et al. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* 13, 358-68 (2003).

66. Yohn, C. T. et al. Lineage-Specific Expansions of Retroviral Insertions within the Genomes of African Great Apes but Not Humans and Orangutans. *PLoS Biol* 3, 1-11 (2005).
67. Hedges, D. J. et al. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res* 14, 1068-75 (2004).
68. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 9, 657-63 (1999).
69. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* 35, 41-8 (2003).
70. Mathews, L. M., Chi, S. Y., Greenberg, N., Ovchinnikov, I. & Swergold, G. D. Large differences between LINE-1 amplification rates in the human and chimpanzee lineages. *Am J Hum Genet* 72, 739-48 (2003).
71. Pickeral, O. K., Makalowski, W., Boguski, M. S. & Boeke, J. D. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res* 10, 411-5 (2000).
72. Goodier, J. L., Ostertag, E. M. & Kazazian, H. H., Jr. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum Mol Genet* 9, 653-7 (2000).
73. Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* 13, 2541-58 (2003).
74. Torrents, D., Suyama, M., Zdobnov, E. & Bork, P. A genome-wide survey of human pseudogenes. *Genome Res* 13, 2559-67 (2003).
75. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* 24, 363-7 (2000).
76. Zhang, Z., Harrison, P. & Gerstein, M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res* 12, 1466-82 (2002).
77. Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H., Jr. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* 73, 1444-51 (2003).
78. Shen, L. et al. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *J Biol Chem* 269, 8466-76 (1994).
79. Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* 99, 3740-5 (2002).
80. Enard, W. et al. Intra- and interspecific variation in primate gene expression patterns. *Science* 296, 340-3 (2002).
81. Khaitovich, P. et al. Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees. *Science* (in the press).
82. Yunis, J. J., Sawyer, J. R. & Dunham, K. The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. *Science* 208, 1145-8 (1980).
83. Fan, Y., Linardopoulou, E., Friedman, C., Williams, E. & Trask, B. J. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13-2q14.1 and paralogous regions on other human chromosomes. *Genome Res* 12, 1651-62 (2002).
84. Fan, Y., Newman, T., Linardopoulou, E. & Trask, B. J. Gene content and function of the ancestral chromosome fusion site in human chromosome 2q13-2q14.1 and paralogous regions. *Genome Res* 12, 1663-72 (2002).
85. Locke, D. P. et al. Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol* 4, R50 (2003).

86. Dennehey, B. K., Gutches, D. G., McConkey, E. H. & Krauter, K. S. Inversion, duplication, and changes in gene context are associated with human chromosome 18 evolution. *Genomics* 83, 493-501 (2004).
87. Subramanian, S. & Kumar, S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res* 13, 838-44 (2003).
88. Duret, L. Detecting genomic features under weak selective pressure: the example of codon usage in animals and plants. *Bioinformatics* 18 Suppl 2, S91 (2002).
89. Sharp, P. M. & Li, W. H. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res* 14, 7737-49 (1986).
90. Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G. & Peden, J. F. DNA sequence evolution: the sounds of silence. *Philos Trans R Soc Lond B Biol Sci* 349, 241-7 (1995).
91. Moriyama, E. N. & Powell, J. R. Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. *J Mol Evol* 45, 378-91 (1997).
92. McVean, G. A. et al. The fine-scale structure of recombination rate variation in the human genome. *Science* 304, 581-4 (2004).
93. Ohta, T. Slightly deleterious mutant substitutions during evolution. *Nature* 246, 96-98 (1973).
94. Ohta, T. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J Mol Evol* 40, 56-63 (1995).
95. Eyre-Walker, A., Keightley, P. D., Smith, N. G. & Gaffney, D. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol Biol Evol* 19, 2142-9 (2002).
96. Makalowski, W. & Boguski, M. S. Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. *J Mol Evol* 47, 119-21 (1998).
97. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351, 652-4 (1991).
98. Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* 132, 1161-76 (1992).
99. Maier, A. G. et al. *Plasmodium falciparum* erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. *Nat Med* 9, 87-92 (2003).
100. Stenger, S. et al. An antimicrobial activity of cytolytic T cells mediated by granulysin. *Science* 282, 121-5 (1998).
101. Swanson, W. J. & Vacquier, V. D. The rapid evolution of reproductive proteins. *Nat Rev Genet* 3, 137-44 (2002).
102. Choi, S. S. & Lahn, B. T. Adaptive evolution of MRG, a neuron-specific gene family implicated in nociception. *Genome Res* 13, 2252-9 (2003).
103. Hardison, R. C. et al. Global predictions and tests of erythroid regulatory regions. *Cold Spring Harb Symp Quant Biol* 68, 335-44 (2003).
104. Lercher, M. J., Chamary, J. V. & Hurst, L. D. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res* 14, 1002-13 (2004).
105. Williams, E. J. & Hurst, L. D. The proteins of linked genes evolve at similar rates. *Nature* 407, 900-3 (2000).
106. Navarro, A. & Barton, N. H. Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. *Science* 300, 321-4 (2003).
107. Zhang, J., Wang, X. & Podlaha, O. Testing the chromosomal speciation hypothesis for humans and chimpanzees. *Genome Res* 14, 845-51 (2004).

108. Lu, J., Li, W. H. & Wu, C. I. Comment on "Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes". *Science* 302, 988; author reply 988 (2003).
109. Charlesworth, B., Coyne, J. A. & Orr, H. A. Meiotic drive and unisexual hybrid sterility: a comment. *Genetics* 133, 421-32 (1993).
110. Angata, T., Margulies, E. H., Green, E. D. & Varki, A. Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. *Proc Natl Acad Sci U S A* 101, 13251-6 (2004).
111. Teumer, J. & Green, H. Divergent evolution of part of the involucrin gene in the hominoids: unique intragenic duplications in the gorilla and human. *Proc Natl Acad Sci U S A* 86, 1283-6 (1989).
112. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-9 (2000).
113. Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* 15, 496-503 (2000).
114. Weinreich, D. M. The rates of molecular evolution in rodent and primate mitochondrial DNA. *J Mol Evol* 52, 40-50 (2001).
115. Dorus, S. et al. Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell* 119, 1027-40 (2004).
116. Zhang, J. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol* 21, 1332-9 (2004).
117. Vallender, E. J. & Lahn, B. T. Positive selection on the human genome. *Hum Mol Genet* 13 Spec No 2, R245-54 (2004).
118. Gilad, Y., Man, O. & Glusman, G. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* 15, 224-30 (2005).
119. Varki, A. How to make an ape brain. *Nat Genet* 36, 1034-6 (2004).
120. Enard, W. & Paabo, S. Comparative primate genomics. *Annu Rev Genomics Hum Genet* 5, 351-78 (2004).
121. Saleh, M. et al. Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature* 429, 75-9 (2004).
122. Fischer, H., Koenig, U., Eckhart, L. & Tschachler, E. Human caspase 12 has acquired deleterious mutations. *Biochem Biophys Res Commun* 293, 722-6 (2002).
123. Puente, X. S., Sanchez, L. M., Overall, C. M. & Lopez-Otin, C. Human and mouse proteases: a comparative genomic approach. *Nat Rev Genet* 4, 544-58 (2003).
124. Nakagawa, T. et al. Caspase-12 mediates endoplasmic-reticulum-specific apoptosis and cytotoxicity by amyloid-beta. *Nature* 403, 98-103 (2000).
125. Humke, E. W., Shriver, S. K., Starovasnik, M. A., Fairbrother, W. J. & Dixit, V. M. ICEBERG: a novel inhibitor of interleukin-1beta generation. *Cell* 103, 99-111 (2000).
126. Vanhamme, L. et al. Apolipoprotein L-I is the trypanosome lytic factor of human serum. *Nature* 422, 83-7 (2003).
127. Seed, J. R., Sechelski, J. B. & Loomis, M. R. A survey for a trypanocidal factor in primate sera. *J Protozool* 37, 393-400 (1990).
128. Angata, T. & Varki, A. Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective. *Chem Rev* 102, 439-69 (2002).

129. Sonnenburg, J. L., Altheide, T. K. & Varki, A. A uniquely human consequence of domain-specific functional adaptation in a sialic acid-binding receptor. *Glycobiology* 14, 339-46 (2004).
130. Pangburn, M. K. Host recognition and target differentiation by factor H, a regulator of the alternative pathway of complement. *Immunopharmacology* 49, 149-57 (2000).
131. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A* 99, 14878-83 (2002).
132. Kulathinal, R. J., Bettencourt, B. R. & Hartl, D. L. Compensated deleterious mutations in insect genomes. *Science* 306, 1553-4 (2004).
133. Pfutzer, R. et al. Novel cationic trypsinogen (PRSS1) N29T and R122C mutations cause autosomal dominant hereditary pancreatitis. *Gut* 50, 271-2 (2002).
134. Chen, J. M., Montier, T. & Ferec, C. Molecular pathology and evolutionary and physiological implications of pancreatitis-associated cationic trypsinogen mutations. *Hum Genet* 109, 245-52 (2001).
135. Altshuler, D. et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 26, 76-80 (2000).
136. Neel, J. V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am J Hum Genet* 14, 353-62 (1962).
137. International HapMap Consortium. The International HapMap Project. *Nature* 426, 789-96 (2003).
138. Hacia, J. G. et al. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet* 22, 164-7 (1999).
139. Watterson, G. A. & Guess, H. A. Is the most frequent allele the oldest? *Theor Popul Biol* 11, 141-60 (1977).
140. Przeworski, M. The signature of positive selection at randomly chosen loci. *Genetics* 160, 1179-89 (2002).
141. Stone, S. et al. A major predisposition locus for severe obesity, at 4p15-p14. *Am J Hum Genet* 70, 1459-68 (2002).
142. Arya, R. et al. Evidence of a novel quantitative-trait locus for obesity on chromosome 4p in Mexican Americans. *Am J Hum Genet* 74, 272-82 (2004).
143. Enard, W. et al. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 418, 869-72 (2002).
144. Schroeder, S. A., Gaughan, D. M. & Swift, M. Protection against bronchial asthma by CFTR delta F508 mutation: a heterozygote advantage in cystic fibrosis. *Nat Med* 1, 703-5 (1995).
145. Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. Evidence for Widespread Degradation of Gene Control Regions in Hominid Genomes. *PLoS Biol* 3, e42 (2005).
146. Ohta, T. Evolution by nearly-neutral mutations. *Genetica* 102-103, 83-90 (1998).
147. Hayakawa, T., Altheide, T. K. & Varki, A. Genetic basis of human brain evolution: accelerating along the primate speedway. *Dev Cell* 8, 2-4 (2005).
148. Gagneux, G., Moore, J. & Varki, A. Great Apes in Captivity: Ethical and Scientific Challenges in the Genome Era. *Nature* This issue (2005).
149. Osoegawa, K. et al. An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* 52, 1-8 (1998).
150. Schwartz, S. et al. Human-mouse alignments with BLASTZ. *Genome Res* 13, 103-7 (2003).
151. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* 12, 656-64 (2002).

152. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* 100, 11484-9 (2003).
153. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res* 31, 34-7 (2003).
154. Higgins, D. G., Thompson, J. D. & Gibson, T. J. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* 266, 383-402 (1996).
155. Meloni, A. et al. Delineation of the molecular defects in the AIRE gene in autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy patients from Southern Italy. *J Clin Endocrinol Metab* 87, 841-6 (2002).
156. Beales, P. L. et al. Genetic and mutational analyses of a large multiethnic Bardet-Biedl cohort reveal a minor involvement of BBS6 and delineate the critical intervals of other loci. *Am J Hum Genet* 68, 606-16 (2001).
157. Cunningham, J. M. et al. The frequency of hereditary defective mismatch repair in a prospective series of unselected colorectal carcinomas. *Am J Hum Genet* 69, 780-90 (2001).
158. Mukhopadhyay, A. et al. Mutations in MYOC gene of Indian primary open angle glaucoma patients. *Mol Vis* 8, 442-8 (2002).
159. Tuchman, M., Jaleel, N., Morizono, H., Sheehy, L. & Lynch, M. G. Mutations and polymorphisms in the human ornithine transcarbamylase gene. *Hum Mutat* 19, 93-107 (2002).
160. Clee, S. M. et al. Common genetic variation in ABCA1 is associated with altered lipoprotein levels and a modified risk for coronary artery disease. *Circulation* 103, 1198-205 (2001).
161. Fullerton, S. M. et al. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am J Hum Genet* 67, 881-900 (2000).
162. Mentuccia, D. et al. Association between a novel variant of the human type 2 deiodinase gene Thr92Ala and insulin resistance: evidence of interaction with the Trp64Arg variant of the beta-3-adrenergic receptor. *Diabetes* 51, 880-3 (2002).
163. Pizzuti, A. et al. A polymorphism (K121Q) of the human glycoprotein PC-1 gene coding region is strongly associated with insulin resistance. *Diabetes* 48, 1881-4 (1999).
164. Katoh, T. et al. Human glutathione S-transferase P1 polymorphism and susceptibility to smoking related epithelial cancer; oral, lung, gastric, colorectal and urothelial cancer. *Pharmacogenetics* 9, 165-9 (1999).
165. Marchesani, M. et al. New paraoxonase 1 polymorphism I102V and the risk of prostate cancer in Finnish men. *J Natl Cancer Inst* 95, 812-8 (2003).
166. Humbert, R. et al. The molecular basis of the human serum paraoxonase activity polymorphism. *Nat Genet* 3, 73-6 (1993).
167. Barroso, I. et al. Candidate gene association study in type 2 diabetes indicates a role for genes involved in beta-cell function as well as insulin action. *PLoS Biol* 1, E20 (2003).
168. Herrmann, S. M. et al. Uncoupling protein 1 and 3 polymorphisms are associated with waist-to-hip ratio. *J Mol Med* 81, 327-32 (2003).
169. Kong, A. et al. A high-resolution recombination map of the human genome. *Nat Genet* 31, 241-7 (2002).
170. Taylor, J., Tyekucheva, S., Zody, M., Chiaromonte, F. & Makova K. D. Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol Biol Evol* 23, 565-573 (2006).
171. Hayakawa, T. et al. A human-specific gene in microglia. *Science* 309, 1693 (2005).

Supplementary Notes: Genome Sequencing and Assembly

The chimpanzee genome sequence was generated from native DNA derived from Clint, a captive-born male chimpanzee from the Yerkes Primate Research Center (Atlanta, USA). Whole genome shotgun libraries were constructed in plasmid pOT4. Protocols are available from the Washington University website (genome.wustl.edu/tools/protocols).

ARACHNE Assembly (Jaffe 2003). The ARACHNE assembly was created using reads from the primary donor only, except as noted below. Human genome sequence (NCBI build 34) was also used in a limited fashion. We created two distinct assemblies, one called the modified de novo assembly (MDN) and the other called the validated chimpanzee-on-human assembly (VCH). We then performed a partial merger to obtain the final assembly described by the Consortium.

The MDN assembly started with a partially formed ARACHNE assembly, constructed without reference to human genome data. This assembly was iteratively modified using several procedures, some of which exploited human data in the following capacities:

1. We normally require two read pairs to join supercontigs, but if we had a single read pair which landed consistently on human sequence (implying that the pair was not chimeric), we allowed a join based on it.
2. Joins were broken in contigs having the following characteristics:
 - there was a weak sequence link holding the contig together, and
 - there was no link across the join (within the contig), and
 - there was no link from both sides to a common contig, and
 - the reads abutting the juncture did not align consistently to the human sequence through the region.
3. The calculated gap size between two consecutive contigs in a supercontig was replaced with the corresponding gap size in human sequence if:
 - these contigs aligned to human sequence, and
 - if the gap size on human was within two standard deviations of the assembly gap size, as predicted from insert characteristics.

The VCH assembly started with alignments of chimpanzee reads to the human build. We utilized only those reads whose placement was unique, or for which unique placement could be inferred from pairing. These were formed into overlapping “piles”, based on relative positions inferred from human sequence and confirmed by chimpanzee-chimpanzee overlap. Piles were

assembled into contigs, which were then “validated” by requiring that SNPs between reads within a given contig were consistent with the presence of no more than two haplotypes. These validated contigs were formed into the VCH assembly.

The MDN and VCH assemblies were then aligned using shared reads to seed the alignments. A merged assembly was formed by importing sequence from VCH into MDN, but only in cases where the two assemblies were consistent. This merged assembly was then iteratively modified, as the MDN had been.

At this stage we temporarily inserted non-Clint reads into the assembly, without changing the sequence in the contigs. Supercontigs were then merged, where possible, after which the non-Clint reads were removed. Those supercontigs that were thereby disconnected were allowed to fall apart. As a final step, prior to release, we manually identified and removed two global misassemblies.

PCAP Assembly As described in (Huang 2003).

Assembly Statistics. Sequence coverage is calculated on the basis of assuming a 3 Gb euchromatic genome size. An additional 2.1 million chimpanzee SNP reads were used as input and later excised from the ARACHNE assembly, as described above.

Genome Alignment. We evaluated three independent methods for aligning chimpanzee contigs to human genome sequence. The first utilized BLASTZ (Schwartz 2003) to align and score non-repetitive chimpanzee regions against repeat-masked human sequence and BLAT (Kent 2002) to process the more repetitive regions. Alignment chains differentiated between orthologous and paralogous alignments (Kent 2003) and only “reciprocal best” alignments were retained in the alignment set.

In the second method, each contig was aligned to human sequence with a hashing procedure based on affine Smith-Waterman calculations for locally refining alignments without repeat masking (D. Jaffe and T.S. Mikkelsen, unpublished). Each alignment was assigned a confidence value based on alignment strength and frequency.

The last method split chimp contigs into 1 kb segments and aligned them to human sequence with BLAT. This step was followed by re-alignment using `cross_match` (P. Green, personal communication) to exploit base quality values and tagging of unique alignments.

Results of the three alignment methods were compared using detailed analyses of local regions using finished chimpanzee sequence, and quality of human RefSeq alignments to the chimpanzee genome when applying the chimp-human alignments. Coverage and consistency with respect to human sequence was also considered. For example, in comparing to a 1.3 Mb region of

finished chimp sequence, less than 0.25% of the aligned bases came from alignments that were clearly spurious. There were 3.068 Gb of scaffold length where all three methods agreed upon chromosomal placement and 3.087 Gb where at least two of the methods concurred. Only 9.9 Mb showed disagreement between all three methods, but in 98% of these cases less than 2 kb of the scaffold could be uniquely placed. There was disagreement as to where 3.8 Mb belonged on specific chromosomes. Of that, only 293 kb had more than 1 kb uniquely mapped.

Creation of Chimp AGP Files. The 37,931 chimpanzee scaffolds comprise 2.73 Gb of sequence and span 3.109 Gb of the genome. Of these, a total of 33,180 (2.70 Gb of sequence spanning 3.077 Gb) scaffolds had significant alignments to the human genome. The process of constructing a path of scaffolds designed to represent the chimpanzee genome was as follows.

All alignments completely contained within other alignments were first removed from the alignment set. Next, alignments where more than 90% of the scaffold length aligned to the human genome in a single chain were examined. When a subregion of such an alignment also aligned to another region of the genome, the alignment was removed. The next step involved examining the subset of scaffolds for which more than 5% of the scaffold content aligned elsewhere in the genome. If the alternate alignments were within 3 kb of the main (90%) alignment, the alignments were merged. If the alternate alignments were each less than 100 bp in length, they were also removed. If the 5% was aligned on the same chromosome but in the random portion, then only the main 90% portion was retained. Only those alignments determined to be chimeric were removed, all others were retained.

All alignments were retained where at least 60% of the scaffold length was aligned to one "region" of the genome. For remaining alignments, when all alignments were to the same chromosome they were assigned to their respective chromosomes, but to the random drawer. When alignments were to various chromosomes, they were assigned to chrUn_random.

All alignments in the >90% and >60% categories were examined if the chimp scaffold spanned more than 1.5 times its counterpart in the human genome. When the offending supercontig only subsumed other supercontigs less than 1.5 kb, they were retained. For those that would completely overlap large contigs, the alignments were manually reviewed to determine if the alignment should be broken.

At this point scaffolds spanning a total of 2.85 Gb were anchored to the human genome sequence (excluding those in the _random bins). All scaffolds that were completely overlapped by another scaffold based on the human position were then removed. Also removed were the smaller of two neighboring contigs when there was an overlap of 60% (based on human) between neighboring

scaffolds. The total anchored sequence after these steps dropped to 2.74 Gb (2.41 Gb of actual contig length), or 88% of the total chimpanzee sequence. An additional span of 280 Mb (240 Mb of actual sequence) was assigned to a chromosome, but in the _random portion, and the final 91 Mb was unassigned. For gaps between supercontigs, sizes were estimated using the alignments to the human genome.

Centromeres were introduced into the chimp at the positions of the centromeres in the human chromosomes. An additional centromere was introduced in Chromosome 2B (formerly PTR13) at the site of the 30 kb of alpha-satellite. Finally, nine documented/known human inversions (Yunis 1982) were introduced into the ordering as was the fusion of human chromosome 2 from chimpanzee chromosomes 2A and 2B (formerly PTR12 and PTR13).

Detailed Assessment of Assemblies. We sought to evaluate the consistency of the assemblies by examining pairing rates and read depths and by comparing assemblies to various data sets. In terms of pairing rates, 90% of the read pairs in the PCAP assembly fell within the distance of their nominal insert sizes. We also examined regions of the assembly that appeared to be collapsed repetitive data by examining local depth of coverage. For example, only 910 kb of the PCAP assembly fell in areas where the depth of coverage exceeded 20 reads. Using a Poisson model for independently and identically distributed sequence with a rate of 4, i.e. for 4x nominal coverage, such regions would have a probability on the order of 10^{-8} of occurring by mere chance. Therefore, they are very likely collapsed repeat regions. Examination of the content of these regions revealed four times fewer simple repeats and low complexity masked regions as compared to the genome as a whole, but 26 times more satellite sequences. About 34% of these bases were masked as interspersed repeats as compared to 42% for the genome as a whole.

As a measure of accuracy of the assembly, we examined the alignments between the chimpanzee assembly and the finished chimpanzee clones for interweaving of supercontigs and mis-ordering of contigs within a supercontig. Both assemblies had some issues arise as a result of the moderate level of coverage, although the ARACHNE assembly had fewer overall. There were cases in both assemblies where overlaps were suggested by human sequence, but which were not recognized by the assemblers. The lengths of these provisional overlaps were always less than 1000 bp.

We also assessed the assemblies by more direct comparisons with the human genome. For example, none of the final assemblies contained any global misjoins. These were defined as regions of at least 50 kb having cross-over or consecutive alignment to more than one human chromosome. Human mRNAs and oriented ESTs were also used to assess coverage of the assemblies. Both

assembly methods performed similarly with respect to percentages of human mRNAs and ESTs aligned.

Absolute values are not of interest for certain comparisons between the chimpanzee and human genome, since many of these cases are true chimpanzee/human differences. However, relative numbers of inter- and intrachromosomal differences were used to compare performance of each assembler. For example, we considered

1. numbers of supercontigs with orientation issues with respect to the human genome
2. numbers of supercontigs where successive pieces of the alignments jumped more than 300 kb in human coordinates
3. numbers of supercontigs where individual contigs within them seemed to be out of order with respect to the human genome.

We specifically examine only supercontigs larger than 150 kb. PCAP had fewer orientation and "jumps" along the human genome, although both assemblies had a similar number of ordering anomalies. In particular, about 73% of those supercontigs exhibited at least one of the above issues, confirming the complex relationship between the chimp and human genomes.

We also examined the assemblies in terms of several additional quantities for which finished BAC sequences were available. These BACs were determined to be either unique or duplicated, based on segmental duplication analysis of the human genome (NCBI build 34). Overall, the mean number of aligned bases is reduced by 30% in duplicated regions (determined by merging Blast (Altschul 1990) alignments where neighboring HSPs were within 80 bp and sequence contigs were in the same orientation). The number of supercontigs and the number of discordances, or disruptions in the linear relationship of adjacent contigs within a supercontig, increases by 3 to 4 times.

We estimate the genome coverage to be about 94%, based on comparison to 12 finished CHORI-251 BAC clones. These clones collectively comprise a total of 1,265,617 bases of sequence. ARACHNE covers 1,186,774 bases, or 93.8% of the clones, while PCAP covers 1,189,836 bases, or 94.0% of the clones.

Comparison with the finished chromosome 21 sequence. We also sought to evaluate the contiguity of the WGS assembly by comparing it to the BAC-sequenced chromosome 21 (Watanabe, 2004). Using the same alignment procedure as described above, a total of 3,462 contigs could be unambiguously aligned to cover 95% of the finished sequence. 57% of this sequence is contained in a single 17.8 Mb supercontig, 90% is contained in the 10 largest supercontigs, and the remainder is contained in 201 smaller supercontigs, often interwoven within the larger ones (~0.6

events per 100 kb). There are 371 undetected contig overlaps with a median length of 101 bp (~1.2 events per 100 kb). With the exception of 88 linked contig pairs that could not be unambiguously ordered in the assembly (~0.3 events per 100 kb), there were no ordering or orientation conflicts between the WGS supercontigs and the finished sequence. These discrepancy rates are nearly identical to the same source BAC comparison. Because chromosome 21 is particularly duplication poor (only ~ 650 kb of segmental duplications) and the pericentromeric region appears to be underrepresented in the finished sequence, these rates may still underestimate the number of discrepancies in highly duplicated/repetitive regions.

Because the error rate at high quality bases in the WGS assembly is significantly lower than the polymorphism rate in the chimpanzee population, the substitution rate between the WGS sequence and the finished chromosome 21 sequence is not informative for accurately determining this error rate (the Q20 substitution rate is 1.2×10^{-3} , which is within 10^{-4} of the expected heterozygosity given that the mutation rate of chromosome 21 is approximately 20% higher than the genome mean). However, by comparing both to the orthologous human sequence, we can determine whether there is any bias in the WGS analysis. We therefore constructed a gene catalog for each of the two chromosome sequences using the same approach as used globally in the manuscript. Using the UCSC Browser knownGenes track for human chromosome 21, orthologous chimpanzee coding regions were extracted for 493 transcripts using the uniquely placed contig alignments, and then filtered for alignment/sequence artifacts. The finished chromosome 21 sequence yielded a mean base coverage of the human coding regions of 95.6%, and had 68 transcripts with frame-shifts or premature stop codons. Manual inspection suggested that the majority of these artifacts are likely due to problems with the human gene predictions (errors in the cDNA evidence, see below). The WGS chromosome 21 sequence yielded a mean base coverage of the human coding regions of 89.3%, and had 95 transcripts with frame-shifts or premature stop codons. Manual inspection suggested that the majority of the 27 artifacts that were not also present in the finished sequence were caused by 1-2 bp indels in the chimpanzee WGS sequence. After removing all but the longest transcript from each gene, a total of 219 unique, artifact free coding region alignments could be extracted from the finished sequence, and 196 from the WGS sequence.

We counted the number of substitutions between the two chimpanzee sequences and the human sequence in the aligned coding regions at all unmasked sites in the WGS assembly. The number of differences between the WGS and finished sequences is 3.88×10^{-4} , well within the range expected from heterozygosity. Both chimpanzee sequences were also roughly equidistant from the human sequence, indicating no detectable excess of substitutions in the WGS sequence.

We also counted the number of substitutions between the finished chimpanzee sequence and the human sequence after including those sites which were masked as low quality in the WGS sequence, shown in the last column on the table below. Including these sites increase the observed substitution rate for the finished sequence by 2.6×10^{-4} , 50% of which is contributed by only 2.5% of the compared genes. This may indicate a slight bias against inclusion of highly diverged regions in the WGS sequence (most likely by mistaking true SNPs for sequencing errors). As described elsewhere, a bias may contribute to the slightly higher number of substitutions observed on the human lineage compared to the chimpanzee lineage, but this would not affect any conclusions in the manuscript. Alternatively, low quality regions may not be independent in the WGS and finished sequences, and including sites corresponding to masked bases in the WGS assembly may slightly increase the error rate of the finished sequence.

Human-chimpanzee ortholog alignments. In order to build a first-generation chimpanzee gene catalog we aligned human RefSeq cDNA sequences to the human genome (NCBI build 34) and in turn transferred those alignments to the chimpanzee sequence. Discrepancies between the cDNA and human genome sequence were carefully flagged (see below). Each gene alignment was transferred to the chimpanzee genome sequence by identifying the orthologous chimpanzee bases of the aligned cDNA via the whole-genome BLASTZ alignments (see above). Discrepancies between the human and chimpanzee sequences were also noted.

Proofreading the Human Genome. As mentioned above, not all apparent discrepancies between human gene predictions and the chimpanzee sequence are due to errors in the latter. A significant source of such problems may in fact be discrepancies between the human genome sequence and the transcript evidence on which the gene prediction was based.

We used the chimpanzee alignment to examine such discrepancies among existing human genomic data. As described above, we first identified the positions of each alignment gap between the cDNAs and the human genome by searching for indels in the alignment descriptions, and then determining how far we could slide them in either direction without introducing a mismatch or merging the gap with another gap. This allowed us to find all equivalent instances of the alignment that have the same score under an affine gap scoring scheme. So for example, for the BLAT alignment

```

                01234  56 7
refseq    CGTAT--AT-C
genome    CGTATATATGC
                01234567891

```

we would report each gap in the format “<rmin> <rmax> <rinsert> <gmin> <gmax> <ginsert>” and the leftmost gap would be described as “1 6 0 1 8 2”, meaning that there is a 2 bp insertion in between positions 1 and 8 in the genome, and a corresponding 2 bp gap between positions 1 and 6 in the RefSeq sequence.

Next, we looked at the alignment of the human genome to the chimpanzee genome. So for example, if the BLASTZ genome alignment showed

```

01234567891
human   CGTATATATGC
chimp   CG--TATATGC
01  2345678

```

we would report that the chimpanzee genome supported the RefSeq at this 2 bp discrepancy, and that it supported the human genome at the other 1 bp discrepancy.

Limiting our analysis to gaps < 10 bp in size, we identified 11,986 unique UTR gaps and 2,582 unique CDS gaps in 6,216 gapped cDNAs alignments. The gaps occurring in the cDNAs are rarely disruptive to the annotated RefSeq CDS. They tend to be close to the end of the CDS, and often have compensating frameshifts nearby, for example:

```

human_cdna   GTTGGCCGCGG-CTGCGAGGACGGGTGCCC
human        GT-GGCCGCGGCCTGCGAGGACGGGTGCCC
chimp        GT-GGCCGCGGCCTGCGAGGACGAGTGCCC
qual         44 555555555555544433333444455

```

The gaps occurring in the genome sequence tend to be more randomly distributed, and are rarely compensated for, e.g.:

```

human_cdna   GATGGGCTCGTCCGCGGAGGACGCGTTGAC
human        GATGGGCTCG-CCGCGGAGGACGCGTTGAC
chimp        GATGGGCTCGTCCGCGGAGGACGCGTTGAC
qual         44444444444444444444444444444444

```

The nature of these inconsistencies (over-representation in UTRs and small effects on the RefSeq annotated CDSs) suggests that they are due mainly to sequencing errors in cDNAs. Disruptive indels are likely to have been filtered out of RefSeq, but probably would not have been discovered in the genomic sequencing/assembly process. Such errors can now be corrected appropriately. On the other hand, it is possible that at least some of these inconsistencies are actual polymorphisms in human sequence. We genotyped 90 CDS indels in the 24-individual NIH diversity panel (Collins 1998). Of 87 successful assays, the chimpanzee sequence correctly predicted the human sequence in

all but three cases. The predicted sequence was found to be monomorphic in the sample in 90% of cases and polymorphic in the remainder. This strongly suggests that ~90% of observed discrepancies between current human cDNAs and the human genome sequence represent errors in the cDNAs, or deleterious mutations acquired by the source cell lines, which is consistent with the conclusions of a similar EST-based analysis (Furey 2004).

Chimpanzee Polymorphisms. Sequences for SNP discovery were generated from three western and three central chimpanzees, *Pan troglodytes verus* and *Pan troglodytes troglodytes*, respectively. For the chimpanzees other than Clint, 4 kb libraries were constructed and ~0.5x whole-genome shotgun reads were generated. The western chimpanzees were captive-born descendants of chimpanzees shipped from Sierra Leone and the central chimpanzees were wild born, confiscated by customs and brought to a primate center in Gabon. BAC clones from two other (putative) western chimpanzees, "Donald" and "Gon" provided an additional end sequences. For more discussion on chimpanzee nucleotide diversity, see Kaessmann (1999), Deinard (2000), and Yu (2003).

Supplementary Notes: Genome Evolution

Divergence rate estimates. Regional divergence rates were estimated over all bases in a chosen segment/window that passed the relaxed NQS(30,25) quality filter (quality score 30 at the compared base, 25 at the five flanking bases on each side, and any number of flanking substitutions allowed), using the baseml program of PAML (Yang 1997) with the REV substitution model. Due to the low level of divergence, the REV model estimate and the observed divergence rate (diverged bases/total bases) was always highly similar.

CpG and non-CpG divergence rates. We observe a rate of divergence at sites in CpG context of 15.2%, compared to a rate of 0.92% in other contexts. The simple assumption would be that the CpG to non-CpG mutation ratio is 16.5. However, some number of these mutations are into, rather than out of, the CpG context, and are in fact normal, not CpG mutations. We can, however, calculate the real ratio, which we define as X , the ratio of CpG mutations to non-CpG mutations (note that X is not the rate of deamination events, so CpG mutations consist of $(X-1)/X$ caused by deamination and $1/X$ resulting from normal replication error). Separating observed mutations into loss and gain requires the assumption that the total fraction of CpG in the genome is roughly in equilibrium (Sved 1990), which seems valid given the high rate of CpG loss, the long history of primate evolution, and the fact that both humans and chimpanzees have almost identical counts of CpGs. As we note, some CpGs may be gained through non-mutational methods such as mobile elements rich in CpG, like Alu or SVA, but the total number of CpGs added to either genome by this method is no more than 500 k, at most 2% of the approximately 25 M CpGs in the aligned portion of either genome.

To calculate the true ratio, we reassign some of the observed CpG mutations to be non-CpG and recalculate the denominators. If we observe a fraction of bases currently in CpG context in one or both genomes, we can break this down into the number of ancestral CpGs plus those gained by mutation into CpG less those lost in both genomes and now classified (with equal probability) as two non-CpG mutations or a non-CpG apparently unchanged base (if the same base mutated in both copies). Assuming that the number of CpGs created equals the number destroyed, this yields a quadratic equation which can be solved to get a true rate of CpG mutation of 4.7% (per genome), with an ancestral fraction of CpGs of 1.78%. The rate of non-CpG mutation increases (because we reassigned more than half of the CpG mutations as non-CpG while only slightly increasing the number of such sites) to 0.535%, (per genome, or ~1.07% divergence) for a ratio of $X = 8.8$.

This is the observed, rather than the instantaneous, rate, although at this divergence they are approximately equal. However, several other factors could influence the estimate. First, the chimpanzee genome is a heterozygous draft, and the CpG context bases are ~10 fold more likely to be polymorphic, which would result in lowered quality scores, and they might be excluded from analysis, which might underestimate the fraction of CpG positions (in fact, human build 34 has 1.98% of its bases in CpG context compared to our estimate of 1.78% estimated in the chimpanzee-human ancestor). Also, this rate assumes that all CpGs are equally susceptible to deamination events. However, approximately 7% of all CpGs are in CpG islands, and presumably protected from methylation (in fact, their mutation rate is only ~0.8%), which would imply a larger mutation ratio at the remaining sites (although the globally observed rate per site remains constant regardless of fraction methylated). In the end, this question could be illuminated further by the sequence of a close outgroup, such as orangutan, baboon, or macaque, which could determine the ancestral human-chimp base at high confidence.

Proportion of fixed differences. Assuming constant mutation rates and no selection, the proportion of observed divergent sites that are non-polymorphic in both the human and chimpanzee populations is $1 - (TH+TC)/(2*THC)$ where TH is the mean time to the most recent common ancestor (TMRCA) of a chromosomal segment in the human population, TC is the TMRCA in the chimpanzee population and THC is the TMRCA of humans and chimpanzees. From coalescence theory (Rosenberg 2002), the expected TMRCA in the human and chimpanzee populations is $4*N_e*g$, where N_e is the effective population size (10,000 for humans, 10,000-20,000 for chimpanzees) and g is the generation time (assumed to be 25 years), giving TH = 1 Myr and TC = 1-2 Myr (see also Excoffier 2002). Assuming THC = 7 Myr, we get a fixed proportion of 0.78-0.86.

Expected variation in divergence due to variation in Time to the Most Recent Common Ancestor. In order to estimate a conservative upper bound on divergence variation due to TMRCA variation, we assumed that 2/3rds of the observed divergence has accumulated since the human-chimpanzee split; that TMRCA is exponentially distributed with a mean of 1/3 times the observed divergence (a conservative upper bound) and that blocks of constant TMRCA are on average 10 kb long (approximately the length of linkage disequilibrium blocks in African populations (Reich 2001) and a likely overestimate for the larger human-chimpanzee ancestral population) and randomly distributed across the autosomes. We estimated the expected standard deviation as 0.07% (roughly one-quarter of the observed standard deviation) from a simulated ensemble of 2,000 random windows of 1 Mb length, assuming constant mutation rates and no

sample variance, repeated 1,000 times. Given that this should be a conservative upper bound, the majority of the variation observed at the megabase scale is unlikely to be due to drift.

Male mutation rate bias for CpGs and non-CpGs. After masking ampliconic regions in Y, pseudoautosomal regions in X and Y, and segmental duplications in all chromosomes, we estimated alpha from the three possible comparisons between X and Y, X and autosomes and Y and autosomes (Taylor 2006). Since the divergence between human and chimpanzee is low and the effective population size is different for X, Y and the autosomes, alpha estimates should be corrected for the effects of pre-existing polymorphism in the ancestral human-chimpanzee population (Makova 2002). Assuming a similar effective population size as contemporary chimpanzees, which is twice as high as contemporary humans (Fischer 2004; Yu 2003), corrected alpha estimates range between 3-6, depending on the pair-wise comparison. The X/A comparisons are likely to be the most accurate because the Y chromosome data is scarce (particularly for diversity on Y - and we need this for correction), and challenging to align correctly. If the relative time spent in the male and female germlines is the dominant factor leading to differences among rates on X, Y, and autosomes, then alpha estimated from the three comparisons should be similar. This can be achieved if we assume a three times higher population size for the human-chimpanzee ancestral population compared with that in contemporary humans and leads to alpha ~5.

However, alpha seems to be not the same for all types of mutations: Intriguingly, alpha estimated at CpG dinucleotides is lower than at all sites. This is consistent with the expectation that CpG to TpG transitions caused by spontaneous deamination of methylated cytosines are time-dependent, rather than dependent on the number of germline cell divisions (Nachman 2000). A close outgroup will be required to separate mutations that create CpGs (and are expected to be replication-dependent) from those that eliminate CpGs (and are expected to be time-dependent).

Detection of Deletions within Bounded Alignments: Small insertion/deletion events (<15kb) were parsed directly from the BLASTZ genome alignment by counting the number and size of alignment gaps between bases within the same scaffold (“scaffold-based indels”) or contig (“contig-based indels”). The size distributions of the bounded indels are given in Figure 9. Missed contig overlaps in the draft assembly create artificial chimpanzee “insertions”, leading to a slight overestimate of the number of unalignable chimpanzee bases from the scaffold-based indels (35.18 Mb total sequence). On the other hand, the relatively small contig sizes lead to an underestimate of small indels in the contig-based set (17.45 Mb total sequence). Together, these sets provide conservative upper and lower bounds on the number of chimpanzee “insertions” in the aligned draft sequence.

Detection of Deletions by Paired-end Placement: Sites of large-scale insertion/deletion (indels >15kb) were detected by optimal placement of paired sequence reads (8.94M fosmid pairs [1788427 reads]), 6.88M plasmid pairs [WASHU: 7339999 reads, MIT: 6420133 reads] and 0.084M BAC pairs [RIKEN: 45828, WASHU: 122614 reads]) against the human assembly (April 2003, build33). Our detection methodology utilizes only high quality read pair alignments to the finished human genome, thereby circumventing false-positives which may be detected by using the draft chimpanzee assembly alone (method in preparation). The distance between the reads of a single clone should reflect the size of the cloned insert (concordant read pair). If the pairs do not place in the correct orientation, or define a region smaller/larger than the expected it is considered discordant. Indels (>15kb) were identified by two or more discordant placement from the same vector, with support from at least one plasmid; macro events (>100kb) are defined by BAC discordant placements. Size thresholds were obtained from read pair distribution of both human fosmids alignments on human sequence ($X=40\text{kb}$; $SD: \pm 2.58\text{kb}$), and chimpanzee plasmid alignments against human chromosome 21 ($X=4.5\text{kb}$; $SD: \pm 1.84\text{kb}$). Size discrepancies were determined to fall within two standard deviations from mean distribution. By identifying read pairs which surpass our thresholds we are able to detect both chimpanzee deletion and potential insertion events in respect to the human genome. Three events were required before considering an indel: each indel must be defined by two or more discordant pairs and the absence of sequence data within the discordancy. This eliminates potential cloning artifacts from further consideration. Other confounding sequence properties (recent duplications, retroelements, etc.) were also considered during this analysis.

Unmapped Chimpanzee Sequence: Roughly 90% of the scaffolds that did not align to the human genome at all contained previously characterized repeats (~5.9 Mb of total sequence). As expected, satellite repeats were largely represented in set. Subterminal satellite repeats, found in many chromosome arms in the chimpanzee and gorilla, represents the largest percentage of identified repeats (62% of all masked bases in unplaced scaffolds) (Royle 1994). Centromeric satellite repeats were also detected, with over 2.1Mb (19% of all masked bases) of sequence consisting of alpha satellite repeat (ALR) and 1.2 Mb of sequence (10.8% of all masked basepairs) of beta satellite repeat (BSR). Only ~1% of all masked bases were due to complex and simple repeats (0.132 Mb; 1.2% all masked bases). HERV and LTR sequences comprise 2.5% of all masked bases (271274 bp).

Estimate of indel basepairs: The total number of insertion/deletion bases between chimpanzee and human was estimated as follows. The number of unaligned bases (“insertions” < 15

kb) within sequence scaffolds was 31.78 Mb (2347812 events) and 35.18 Mb (2741577 events) for human and chimpanzee respectively. The number of chimpanzee deletions >15 kb was estimated by paired-end sequence to be 8.2 Mb (163 events). 5.9 Mb of chimpanzee sequence could not be mapped back to human using low sequence threshold cutoffs. We estimate a similar amount of such sequence for human. In total, we estimate 95.2 Mb ($31.78+35.18+5.9 *2 + 8.2*2$) or 3.2% difference between chimp and human.

Processed Pseudogene Analysis. Based on a divergence time of 6 million years and the number of processed pseudogenes reported in the paper, we estimate a minimum rate of retrotransposition as 40 and 60 events per million years. This is significantly reduced when compared to a constant rate of retrotransposition after the human-mouse split. We estimate 17,000 human processed pseudogenes have emerged since the human-mouse divergence with a concomitant rate of 170 events (17,000/100) per My (Torrents 2003).

New repeat-derived CpG islands in humans: Some interspersed repeat elements contain CpG-rich regions that could theoretically become functional CpG-islands if inserted in the promoter region of a host gene.. At least 3 of ~1000 human-specific CpG islands have been inserted in the promoter region of known genes, but additional data will be required to determine whether these insertions have led to changes in gene expression patterns.

Repeat-mediated homologous recombination: We curated the results by hand to eliminate assembly artifacts in chimpanzee and cases of expansion or contractions of tandem duplications. We limited the analysis to those indels with breakpoints well within the repeat elements. This removes duplications that simply have an element on one site, but also results in an underestimate by missing those recombination events with breakpoints on the edge of repeats.

Detection of large-scale Inversions: Optimal BAC read pair (32,826) placements against the human genome (build33) were evaluated for large discordant placement (greater than 2Mb) to identify sites of inversion. Reads pairs which are incorrectly orientated and lack concordant read pair placement within breakpoints may identify sites of rearrangement. Large-scale inversions (>2Mb) were initially determined by 2 or more discordant BACs spanning the same region. Plasmid and fosmid discordant placements were then used to refine the breakpoints and increase confidence in a potential rearrangement. Breakpoints supported by 2 or more fosmids, plasmids, and BAC discordant read pair placement were selected for experimental validation. We utilized a previously described method (Nickerson 1998) to validate breakpoints. Two or more probes were selected in the human genome that mapped on either side of the breakpoint region identified by our study. Two-color FISH experiments (or single FISH experiment if the region was unique) were then

performed with each pair of BAC probes. True inversions in the chimpanzee lineage will appear as separate FISH foci/ split signal within a chimpanzee metaphase chromosomes as opposed to a merged/single signals within the human genome (Nickerson 1998)

Segmental Duplication Analysis. Segmental duplication is very difficult to analyze on the basis of draft genome sequence, because sequence from duplicated regions may be collapsed together and sequence from a single region may fail to be assembled together (resulting, respectively, in under- and over-estimates of the extent of duplication). With near-complete sequence, it is now possible to estimate that ~5.3% (150.8 Mb) of the human genome resides in regions of segmental duplications (defined as stretches of >1 kb in length matching other regions with >90% identity (IHGSC 2004, She 2004). The chimpanzee genome assembly shows a lower amount of segmental duplication (136.7 Mb), with greater fragmentation and more regions with >99% identity. However, these apparent differences from the human are likely predominantly to reflect limitations of the draft genome assembly (She 2004). Like the human genome, the chimpanzee assembly shows both extensive interchromosomal and intrachromosomal duplication; in contrast, the mouse and rat genomes have predominantly intrachromosomal duplications (Bailey 2004, Tuzun 2004, Cheung 2003).

Supplementary Notes: Gene evolution

Ka, Ks, Ki. The Ka and Ks rates are estimated jointly by PAML (Yang 1997) from all aligned bases with quality score > 20 in orthologous coding regions, using the F3x4 codon frequency model and the REV substitution matrix. Lineage specific rates were estimated using an unrooted tree including human, chimpanzee, mouse and rat. Ki is estimated by PAML as substitutions per nucleotide for all aligned, orthologous nucleotides that passed the relaxed NQS(30,25) filter (any number of flanking substitutions allowed) in non-coding, interspersed repeats (not low-complexity) within 250kb, centered on each gene.

We note that CpG dinucleotides are not explicitly modeled, which leads to bias in the rate estimates due to differential sequence content. On average, Ks appears to be more sensitive to CpG content than Ka, which again is more sensitive than Ki. For example, genes in GC-rich regions appear to have lower mean Ka/Ks values and higher Ka/Ki values than genes in GC-poor regions. To the best of our knowledge, none of the results presented in this work are directly affected by this apparent bias, but we caution that for applications where absolute rate estimates are crucial, more sophisticated evolutionary models should be used. Models that explicitly incorporate context-dependent mutation rates are currently under development (e.g. Hwang 2004, Siepel 2004).

Bias due to quality masked chimpanzee SNPs. The apparent slight excess of amino acid and synonymous substitutions on the human lineage in the hominid-murid comparison may be partly explained by a data artifact. Because the human assembly is a tiling of single haplotypes, whereas the chimpanzee assembly is a consensus of two haplotypes, some heterozygous positions in the chimpanzee WGS reads are effectively masked as low quality positions. This may lead to a slightly disproportionate contribution of human diversity to the human divergence rate. Additional high quality sequence will be required to robustly test for small differences in the human and chimpanzee substitution rates.

Ka/Ks of polymorphism. Polymorphisms from the HapMap or Affymetrix datasets that overlapped one of the aligned 13,454 orthologs were classified as synonymous or nonsynonymous according to whether they changed the overlapped amino acid when the flanking bases were fixed as the human and chimpanzee reference sequences. The Ka/Ks of the polymorphisms is calculated as $(\Delta A / N_a) / (\Delta S / N_s)$ where N_a and N_s are the number of nonsynonymous and synonymous sites sampled from, and ΔA and ΔS are the number of observed nonsynonymous and synonymous differences. N_s / N_a was estimated as ~0.36 from the aligned orthologs. The % excess of nonsynonymous substitutions were calculated as $1 - (\Delta A_{\text{diversity}} / \Delta S_{\text{diversity}}) / (\Delta A_{\text{divergence}} / \Delta S_{\text{divergence}})$.

The confidence interval for this fraction was estimated by assuming that both ΔA counts were Poisson distributed and both ΔS fixed, and calculating a likelihood-based 95% confidence interval for a ratio of two rates (Graham 2003).

Simulation of expected Ka/Ki distributions. The expected distribution of observed Ka/Ki values over 13,454 orthologs under the null hypothesis of no positive selection were simulated by randomly redistributing the observed number of non-synonymous substitutions between the orthologs under the constraints that (i) the actual (not observed) Ka/Ki = 0.23 for all genes, or (ii) 23% of the non-synonymous sites, distributed between genes according to a beta distribution, evolved at Ka/Ki = 1, and the remaining at Ka/Ki = 0, and that the number of orthologs with observed Ka/Ki > 0 were equal to the observed fraction (~29%). The mean numbers of orthologs with observed Ka/Ki > 1 over 100 trials were (i) 0 and (ii) 263. Note that this does not guarantee that the apparent excess of genes with observed Ka/Ki > 1 is due to positive selection, 263 is simply a *lower bound* on the number of cases that can be explained by stochastic fluctuations.

Identification of rapidly diverging gene clusters. We ordered all aligned human-chimpanzee orthologs by their genomic positions and calculated the median Ka/Ki ratio for sliding windows over 10 genes, with a step size of 2, across the autosomes (because X chromosome Ka/Ki are not directly comparable for this purpose). We estimated the distribution of Ka/Ki under the null hypothesis of no clustering by repeating the same procedure 1,000 times on the same orthologs in random orderings. Empirical P-values were calculated by comparing the observed Ka/Ki to this null distribution, and windows with a P-value of less than 0.001 were reported. Approximately 13,000 orthologs were scanned, implying that approximately 1,300 independent tests were performed, we therefore calculated the significance of seeing 16 clusters with P-value less than 0.001 from a binomial distribution with $n=1,300$ and $p=0.001$. When multiple, overlapping windows corresponded to the same gene cluster, only the most significant window was kept. The list of top clusters also remained highly similar when the procedure was repeated using Ka only. Repeating the analysis with larger window sized did not identify any regions that could not be explained by the inclusion of one or more of these original clusters.

No accelerated evolution in rearranged chromosomes. Navarro et al. (2003) proposed that orthologs within one or more of the pericentric inversions between human and chimp might show higher Ka/Ks than the genome-wide average. We re-evaluated this hypothesis on our considerably larger dataset, using a binomial test for increased Ka/Ks relative to genes on the collinear chromosomes. All P-values are corrected for multiple hypothesis testing. A few of the regions contained few or no aligned orthologs.

Only the HSA4 and HSA5 inversions show a significant increase in Ka/Ks, but it is significantly smaller than the 2.2-fold increase observed on the smaller dataset in (Navarro 2003), and it is primarily due to low Ks, rather than the overall accelerated divergence that is predicted by the model of Navarro et al. Furthermore, there is no increase in Ka/Ki. However, we note that because of the numerous confounding factors, such as differences in gene content and sequence composition, it is difficult to directly compare the rates of protein evolution between relatively small genomic regions. Thus, our negative findings do not necessarily disprove the chromosomal speciation model described by Navarro et al. *per se*, just the huge quantitative signal implied by their initial report.

Identification of rapidly and slowly evolving categories. Unique LocusLink identifiers (June 2004) of 13438 human-chimpanzee mRNA alignments GO categories (May 2004) were assigned to 9205 orthologs via GenMapper (Do 2004). We used the following approach to identify GO categories that have a Ka/Ks ratio significantly above or below average.

First, the ‘concatenated’ k_a and k_s for all genes in a GO taxonomy T were calculated as

$$k_a = \frac{\sum_{i \in T} a_i}{\sum_{i \in T} A_i} \quad k_s = \frac{\sum_{i \in T} s_i}{\sum_{i \in T} S_i}$$

where n_i and N_i are the numbers of non-synonymous substitutions and sites, and s_i and S_i are the numbers of synonymous substitutions and sites in gene i , as estimated by PAML, respectively.

The expected proportion of non-synonymous substitutions p_A in a GO category C was then estimated as:

$$p_A = \frac{k_a \sum_{i \in C} A_i}{k_a \sum_{i \in C} A_i + k_s \sum_{i \in C} S_i}$$

Finally, for a given category, the probability p_c of observing an equal or higher number of non-synonymous substitutions, conditional on the total number of observed substitutions, was calculated assuming a binominal distribution:

$$p_C = \sum_{j=a_C}^{a_C+s_C} \binom{a_C+s_C}{j} p_A^j (1-p_A)^{a_C+s_C-j}$$

where a_C and s_C are the total number of non-synonymous and synonymous substitutions in GO category C , respectively.

Defined in this manner, p_C , is a statistic whose expected value has two relevant properties: (1) given a fixed number of substitutions, its value decreases monotonically as the category Ka/Ks increases relative to the GO taxonomy average and (2) given a fixed Ka/Ks ratio, its value decreases monotonically as the total number of substitutions in a category grows. A GO category with a low p_C value is therefore more likely to be rapidly evolving relative to all other genes than a GO category with a high p_C value. However, because of the large number of categories tested and the unknown variance of p_C , it is not a conservative p-value for category specific acceleration.

Slowly evolving categories was detected correspondingly by calculating the probability that a category contains equal or less non-synonymous substitutions, conditional on the total number of observed substitutions.

To determine whether a subset of the categories are evolving under significantly high (low) constraints we first computed the number of GO categories with at least 20 orthologs and p_C less than a threshold value (0.05, 0.01 or 0.001). We then repeated this procedure 10,000 times on the same dataset after randomly permuting the GO annotations (all GO categories assigned to a specific gene are kept together in order to preserve the hierarchical structure of the GO categories). Finally, we tested the null hypothesis that the number of biologically meaningful categories with p_C below the chosen threshold is no more than expected from randomly composed categories by counting how many of the latter have more low p_C values. A rejection of this null hypothesis implies that the level of constraint is significantly higher (lower) than average in some biologically meaningful categories. The average number of categories with p_C below the threshold identified in the randomized datasets is the expected number of false positives among the putatively rapidly (slowly) evolving categories.

Since categories are overlapping, the list of significant categories can contain redundant information. Therefore we used a 'refinement' algorithm to generate the non-redundant Table 5. This algorithm removes parent categories that do not have p_C values below the chosen threshold after the genes in their child categories with low p_C have been removed.

As an alternative to the binomial statistics described above, we also used a non-parametric approach to test for categories with a significantly high or low median Ka/Ki ratio. For this purpose, we first used a Wilcoxon two-sample test with a correction for ties as implemented in the R package (<http://www.r-project.org>) to calculate the probability of equal medians, and then repeated the permutation test as described above to estimate the significance of this statistic. The results are largely similar to the binomial approach, but with slightly more outliers and a lower expected false positive rate, potentially due to the lower variance of Ki compared to Ks. However, the lack of

lineage-specific K_i estimates, combined with the large number orthologs with observed K_a or $K_s = 0$ between human and chimpanzee make rank sum statistics ill-suited for the relative rate tests described below.

Identification of lineage-specific acceleration in GO categories. Unique LocusLink identifiers (June 2004) of the 7043 human, chimpanzee, mouse, rat mRNA alignments were used to assign GO categories (May 2004) to 4805 genes via GenMapper (Do 2004).

We used a similar approach to the binomial test described above to identify GO categories that have an excess of non-synonymous changes on one lineage (e.g. between mouse and rat or on the human lineage) than on another lineage (e.g. between human and chimpanzee or on the chimpanzee lineage). Instead of calculating the expected proportion of non-synonymous to total substitutions, we calculate the genome-wide proportion of non-synonymous changes on one lineage over the total number of non-synonymous changes between the two compared lineages:

$$p_x = \frac{x}{x+y} \quad \text{for } x = \sum_i x_i \quad \text{and} \quad y = \sum_i y_i$$

where x_i and y_i are the numbers of non-synonymous changes on lineages x and y for gene i , respectively. For the hominid vs. murid tests, the numbers of substitutions were estimated independently from pair-wise comparisons of human-chimpanzee or mouse-rat alignments across the same codons. For the human vs. chimpanzee and human vs. mouse tests, the numbers of lineage-specific substitutions were estimated jointly with mouse and rat as outgroups using PAML as described above.

$$P_C = \sum_{j=x_C}^{x_C+y_C} \binom{x_C+y_C}{j} p_x^j (1-p_x)^{x_C+y_C-j}$$

As described for the absolute rate tests, we then computed this statistic for every GO category with more than 20 orthologs, as well as for every category in 10,000 randomly permuted data sets. The top 10 non-redundant categories from the hominid vs. murid comparison were reported in Table 6.

In order to rule out any estimation bias due to using PAML to estimate lineage-specific rates from relatively little sequence information, we repeated the relative rate tests using the number of observed amino acid changes only. For the hominid vs. murid tests, the observed numbers of amino acids between each of the two species pairs were counted. For the human vs. chimpanzee and human/chimpanzee vs. mouse tests, only the amino acid changes that could be unambiguously assigned to one of the three lineages using mouse as the outgroup were counted. The results are essentially identical to those from the non-synonymous rate estimates.

As shown in the paper, there is a significant correlation between synonymous and non-synonymous substitution rates across orthologs. In order to test whether any category-dependent non-synonymous acceleration is due to systematic changes in mutation rate we repeated the relative rate analysis for synonymous substitutions. Although there is a trend towards an excess of synonymous substitutions in some categories between hominids and murids, the variation is significantly less than for non-synonymous substitutions. The strongest outliers are cell surface receptors and adhesion related genes, which have previously been noted to be biased towards mutational hotspots in the human genome (Chuang 2004).

Human Disease Genes: Overall substitution rate. We downloaded all available 1116 OMIM disease genes using EnsMart (www.ensembl.org) and could find a human-chimpanzee cDNA alignment for 882 of them. We compared K_a , K_s , K_i , K_a/K_s and K_a/K_i between disease genes and non-disease genes and used a Mann-Whitney test to estimate the significance of the observed differences.

Recent studies (Huang 2004, RGSC 2004) have reported that human disease genes show an elevated synonymous substitution rate (K_s) in human-rat comparisons. The authors suggest that disease genes may tend to be located in genomic regions with elevated mutation rates. We re-examined this question using the chimpanzee sequence, which allows us to study both K_s and K_i . Disease genes indeed show higher K_s (0.0138 vs. 0.0126, $p_{MW} < 0.001$), but no elevation in K_i . This indicates that the effect is not due to known disease genes being located in regions with higher overall mutation rates. In fact, the elevated K_s in disease genes appears to be explained by a higher proportion or rate of substitution of CpGs at silent sites and because these CpGs have a higher substitution rate, the effect disappears when only non-CpG sites are considered.

We also tested whether there were more amino acid changes on the human lineage than on the chimpanzee lineage in disease genes and in mental retardation genes (Inlow 2004) using the binominal test described above. We do not find a significant difference in the ratio of human-specific changes to chimpanzee-specific amino acid changes between these two classes and genes that have no disease association. This is also true for genes on the X-chromosome, which does not support the hypothesis that mental retardation genes on the X-chromosome would have contributed particularly to the evolution of human cognition as proposed by Zechner et al (2001).

Human Disease Genes: Specific disease alleles. We downloaded all amino acid variants annotated as “disease” variants in SWISS-PROT (www.expasy.org) and all amino acid variants annotated in HGMD (www.celeradiscoverysystem.com), mapped them to their corresponding position in human-chimpanzee mRNA alignments and filtered out positions of low sequence quality

($Q < 20$ in at least one codon position). This resulted in a combined list of 12164 disease variants at 10886 amino acid positions in 1384 different genes. At 46 positions in 41 different genes the chimpanzee variant was identical to the human disease variant. We manually screened the literature for these variants to confirm a plausible causative association of the variant with the disease and excluded 30 variants. We also checked the chimpanzee trace archive to confirm the chimpanzee sequence and excluded one variant. For the remaining 15 variants in 14 different genes, we confirmed the ancestral state by genotyping them across a panel of primate samples. One additional variant (PPARG) was manually added due to the wrong allele being assigned as the disease allele in HGMD.

Supplementary Notes: Human Population Genetics

Alignment of Reads and Variant Calling. Human reads from the Baylor African-American diversity panel (International HapMap Consortium, 2003) consisting of pooled DNA from 4 male and 4 female African-American donors generated at the Whitehead Institute/MIT Center for Genome Research and the Baylor College of Medicine Human Genome Sequencing Center were downloaded from the NCBI trace repository (<http://www.ncbi.nlm.nih.gov/Traces>). A total of 5,316,404 reads were downloaded and quality screened to eliminate reads which did not have: length ≥ 500 bp, $\geq 60\%$ of length at Phred score ≥ 20 , and at least 100 (for WIBR) and 50 (for BCM) passing reads on their sequencing plate. The reads were aligned to Build 34 of the human genome using the alignment portion of the Arachne assembler (Jaffe 2003). Reads were discarded at this phase if they were not uniquely placed or were not placed consistently with their annotated paired end. SNPs were called using the neighborhood quality standard (Altshuler 2000) with a window size of 11, minimum score for the variant base of 30, minimum flank score of 25, maximum mismatches within the window 2, and maximum indels in the window 0. We discarded any alignments, which yielded fewer than 200 NQS bases at that threshold or a SNP rate greater than 1%. This resulted in 3,497,810 read alignments covering 1.945 billion NQS bases and yielding 1,924,196 discrepancies vs the reference sequence (average heterozygosity of 9.9×10^{-4}).

For human-chimpanzee divergence, we started with 23,021,928 chimpanzee reads sequenced at the Whitehead Institute/MIT Center for Genome Research, Washington University Genome Sequencing Center, The Institute for Genomic Research, and RIKEN (Fujiyama 2002). We applied similar criteria to the reads, modified as follows: only 50% of raw bases \geq Phred 20 were required, and the reads had to have 30% of bases matching only the quality portion of NQS prior to alignment and at least 200 such bases (because these trimmed reads were also being used for chimpanzee-chimpanzee comparisons at the read level, we needed screens which were independent of the human reference genome). These reads were then aligned to the human genome and variants called as above, with the exception that we placed no upper bound on divergence rate. We estimate the genome-wide average divergence to be 1.23%. An alternative analysis not applying the mismatch/indel restriction on the NQS windows raises the estimate slightly to 1.27%.

Assignment of Ancestral Alleles. We started with the most recent build of NCBI's dbSNP, which had been mapped to build 34 of the human genome (from <http://genome.ucsc.edu>). In order to assign ancestral alleles, we used the same chimpanzee read alignments to human that were used to call chimpanzee SNPs. Repetitive or segmentally duplicated regions of the human genome were

not covered by this method. This yields calls for 79.6% of human SNPs, with 1.2% having a chimp base that agrees with neither human allele and 0.4% being polymorphic in chimpanzee. If we then use the draft assembly alignment to human to augment the coverage, we can cover an additional 6-10% of SNPs (depending on quality threshold on the chimp assembly base) with a rate of chimp bases matching neither human allele which is slightly higher than the uninformative calls for the original alignment.

Estimating Error Rate of Ancestral Allele Assignments. We estimated the rate of error in the assignment of ancestral bases as the probability that the chimp base matches one of the two human alleles, but not the one that is ancestral in the human population. There are two simple cases in which this could happen, first where the chimp base has mutated to the same base as the derived human allele and second where the human base has experienced a fixed mutation at some point in the past and then experienced a reversion mutation that is still segregating. Cases involving more mutations are possible but are at least two orders of magnitude less likely than these. Since most segregating variants are <1 Mya, we take the probability of a prior fixed change in human to be about equal to the probability of a change in chimp, so in the general case, both of these events have equal probability:

$$P_{\text{error}} = (P_{\text{change}})(1 - P_{\text{change}})(P_{\text{same}}) + (1 - P_{\text{change}})(P_{\text{change}})(P_{\text{same}})$$

P_{change} is ~ half the observed divergence, or 0.00615. P_{same} is the probability that both mutations are identical, which is 0.5, given a 2:1 transition:transversion ratio. This is all contingent on the human base currently being polymorphic, so we take $P_{\text{hs-poly}} = 1$ and drop it. This makes P_{error} 0.6%.

Breaking these sites down into CpG and non-CpG reveals a more complex story. Polymorphic sites in the human genome that are not in a CpG context for either allele will essentially follow the equation above, with P_{change} reduced to our estimated non-CpG mutation rate, 0.00535, for P_{error} of 0.5%. A small number of these sites may have ancestrally been CpG, which would create a higher error rate (see below), but we estimate only 0.16% of the genome was ancestrally CpG and has mutated out, so the effect will be negligible.

However, for sites which are in a CpG context for one of their alleles in human, we need to consider more seriously the possibility of multiple mutations. Because of the high frequency mutation of CpG to TpG, CpG context alleles in human whose chimp alignment is to ApG or GpG will have the CpG as the derived variant in human 85-95% of the time and thus be only slightly more likely to be erroneously estimated than in the non-CpG case. Similarly, mutants in the human which are [C/X]pG and $X \neq T$ will rarely (<20%) be ancestrally CpG. Thus, we limit our analysis to those mutations where the human is [C/T]pG and the chimp is CpG or TpG. (The former case,

chimp CpG, will stand as illustrative for all the cases not explicitly calculated that the CpG effect on error is small given that observation.)

For each observed state, human = [C/T]pG and chimp = CpG or TpG, we define prior probabilities of observing each of the four likely combinations of ancestral states of the human chimp ancestor, HCA = C or T, and the most recent common human ancestor (the ancestral human allele), MRCA = C or T as follows:

$$P(\text{HCA} = \text{X}, \text{MRCA} = \text{X} \mid \text{Pt} = \text{X}) = P_{\text{XpG}} \cdot (1 - P_{\text{XpG-ch}})^2 \cdot P_{\text{XpG-p}}$$

$$P(\text{HCA} = \text{X}, \text{MRCA} = \text{Y} \mid \text{Pt} = \text{X}) = P_{\text{XpG}} \cdot (1 - P_{\text{XpG-ch}}) \cdot P_{\text{XpG>YpG}} \cdot P_{\text{YpG-p}}$$

$$P(\text{HCA} = \text{Y}, \text{MRCA} = \text{X} \mid \text{Pt} = \text{X}) = P_{\text{YpG}} \cdot P_{\text{YpG>XpG}}^2 \cdot P_{\text{XpG-p}}$$

$$P(\text{HCA} = \text{Y}, \text{MRCA} = \text{Y} \mid \text{Pt} = \text{X}) = P_{\text{YpG}} \cdot P_{\text{YpG>XpG}} \cdot (1 - P_{\text{YpG-ch}}) \cdot P_{\text{YpG-p}}$$

Where:

$P_{[\text{X/Y}]pG}$ = probability that the ancestral sequence was [X/Y]pG at any base

$$P_{\text{CpG}} = 0.0178, P_{\text{TpG}} = 0.145$$

$P_{[\text{X/Y}]pG-ch}$ = probability that an ancestral [X/Y]pG has changed at all

$$P_{\text{CpG-ch}} = 0.047, P_{\text{TpG-ch}} = 0.00535$$

$P_{\text{XpG>YpG}}$ = probability that an ancestral XpG will mutate to YpG

$$P_{\text{CpG>TpG}} = (0.047)(8.45/8.8) = 0.045, P_{\text{TpG>CpG}} = (0.00535)(.65) = 0.003$$

$P_{[\text{X/Y}]pG-p}$ = relative probability that a given human site will become a C/T variant

The base rate of polymorphism will divide out, so we take this as 0.65 for TpG and 8.45 for CpG

Cases 1 and 3, where MRCA = Pt, will yield correct inferences of the human ancestral allele, while 2 and 4, MRCA \neq Pt, will yield errors. The ratio of the sum of the latter two to the total will give the error rate. For Pt = C, this gives us $P_{\text{error}} = 0.6\%$, as suggested, only slightly different than the non-CpG case. However, when Pt = T, we get $P_{\text{error}} = 9.8\%$, thus these bases will be a significant source of error.

Effect of Bottlenecks on Ancestral Allele Probabilities. We estimated the effect of a bottleneck on ancestral allele frequencies as follows. Using a diffusion approximation for how the frequency of an allele changes with time, one can show that, under the simplest demographic assumptions, the probability density that a derived allele has frequency f (for $0 < f < 1$) is

$$\lim_{\varepsilon \rightarrow \infty} \int_{s=0}^{\infty} K(\varepsilon, f; s) \partial s / \varepsilon = \frac{2}{f}$$

where $K(x,y;t)$ is the transition probability that an allele initially at frequency x is at frequency y after time t (Patterson 2005). From a diffusion perspective, a genetic bottleneck is a time interval in which the allele frequencies diffuse, but no new mutations occur. In essence, ‘genetic time’ is stretched. For a bottleneck with inbreeding coefficient b , the corresponding time interval has length

$$t(b) = -\log(1-b)$$

After a bottleneck of inbreeding coefficient b , the frequency distribution of derived alleles will be given by

$$\lim_{\varepsilon \rightarrow \infty} \int_{s=t(b)}^{\infty} K(\varepsilon, f; s) \partial s / \varepsilon$$

where the range of integration starts before the bottleneck. As no new mutations are introduced during the bottleneck, the low frequency alleles after the bottleneck will be overrepresented in the population by alleles of previously higher frequency (and larger probability of being ancestral) that drifted downward in frequency during the bottleneck. The above equation can be evaluated numerically. The slope following a bottleneck decreases from 1 to roughly $(1-b)$.

Allele Frequency Dataset. The genome-wide dataset that we analyze here (from Affymetrix: (www.affymetrix.com/support/technical/sample_data/genotyping_data.affx) is composed of a collection of individuals from multiple populations (6 Venezuelan, 6 Chinese, 6 African-American, 12 Caucasian, and 24 of unknown origin); accordingly the effect of recent bottlenecks on the distribution of ancestral probabilities is a mixture reflecting the various subpopulations. We have also analyzed a much smaller set of data, generated across several of the ENCODE regions, for which we have separate results for European, Asian, and West African HapMap samples. These show that the European and Asian slopes are well below 1, consistent with the effects of an out of Africa bottleneck, while the West African population has a slope close to 1.

Excess of Derived Alleles After Selective Sweep. Within the immediate region (i.e., in the absence of recombination during the sweep) of an advantageous allele under selection, ancestral variation will be completely removed. Distant from the selected allele (i.e., where recombination has removed association), there will be no effect. Within the region where recombination occurs, but rarely, during the sweep, some alleles will be swept to high frequency while others will be driven to low frequency. The probability that an allele exists on the selected background is given by its frequency f , while the number of derived alleles at frequency f is proportional to $1/f$. As a result, the distribution of pre-sweep allele frequencies after the sweep is uniform across pre-sweep allele

frequencies, meaning that high frequency alleles are equally likely to be derived or ancestral, creating a large excess of high frequency derived alleles (Fay 2000). This signal is highly specific for selection, but not especially sensitive, especially at long times past the end of the sweep, as high frequency derived alleles created by the sweep move to fixation and all new low frequency alleles introduced by mutation during and after the sweep are at low frequency, rapidly restoring the balance of high frequency ancestral alleles (Przeworski 2002).

Expected Width of Reduction of Diversity in Selected Regions. We performed simulations with the program *cosi* (Schaffner 2005) using median values for human recombination (1 cM/Mb) and selective sweeps of $s = 0.005, 0.01, \text{ and } 0.02$ ending 5000 generations ($\sim 125,000$ yrs) ago, we found that the probability that region over which heterozygosity is reduced to below 50% of the average value exceeds 1 Mb is 6.3%, 13.9%, and 40.6%, respectively.

Scoring of Low Diversity Relative to Divergence Regions. We identified regions in which the observed human diversity rate was much lower than the expectation based on the observed divergence rate with chimpanzee.

We compared the human diversity to the chimpanzee divergence to eliminate regions in which low diversity simply reflects a low mutation rate in the region. In order to capture the uncertainty in diversity and divergence estimates within each window, we looked at each set of non-overlapping windows (since the window step is 1/100 the size, there are 100 such sets). Within each window, we took the observed number of human SNPS, u_i , human NQS bases, m_i , human-chimpanzee substitutions, v_i , and chimpanzee NQS bases, n_i , and generated two random numbers from the distributions:

$$\begin{aligned} x_i &\sim \text{Beta}(u_i + a, m_i - u_i + b) \\ y_i &\sim \text{Beta}(v_i + c, n_i - v_i + d) \end{aligned}$$

where $a = 1$, $b = 1000$, $c = 1$, and $d = 100$. We then took x_i as the human diversity and y_i as the human-chimp divergence for each window i and fit a linear regression

$$x | y \sim N(\alpha_0 + \alpha_1 y, \beta^2)$$

A p-value for each window was then calculated for each window based on (x_i, y_i) and the regression line. This was repeated 100 times and the average of the p-values taken as the p-value for diversity given divergence each window. The window was assigned a score proportional to $-\log(\text{p-value})$.

Because we were looking for a signal where diversity was low relative to divergence, we were concerned that regions where divergence might be artificially high would preferentially appear

in our analysis. In order to avoid finding such regions, which might be true but were deemed likely to be enriched in artifacts, we aggressively screened the windows. The $-\log(\text{p-value})$ score was set to 0 for any window matching any of the following: low human or chimpanzee NQS coverage (NQS bases ≤ 0.5 max NQS coverage), in the highest quartile of human chimpanzee divergence, within 3 Mbp of a human centromere or telomere, or within 1 Mbp of a large gap in the human genome.

After filtering, we coalesced regions as the maximal overlapping windows with $p < 0.1$ containing at least one window of $p < 0.05$ and scored them as the sum of their $-\log(\text{p-value})$ scores, thus weighting for both length and strength.

FOXP2 – CFTR Region. The genomic region on 7q containing both FOXP2 and CFTR stands out as unusual, although no specific part of it scores exceptionally high in the diversity-divergence test. The region is 7.58 Mb long and is covered by 3 separate regions, running from 112.88 to 114.41, 114.83 to 117.15, and 117.77 to 120.46 Mb, and covering 6.55 Mb of the extended region. Were these regions merged into a single region, their combined div-div score would be 94.4, ranking it as the second highest scoring region. Two of the three regions show large windows of severe derived allele frequency skew, but only in the central region does it come close to overlapping the highest diversity-divergence score. Intriguingly, well outside this region and flanking it, at 106 to 108 and 121 to 123 Mb are two other large regions of severe derived allele frequency skew. It is tempting to speculate that since the hitchhiking model posits the derived allele skew in the flanks of the region, this may be the relic of a very powerful sweep affecting the entire extended region, although such an observation would require a selective coefficient on the order of 0.1-0.2. Alternatively, an undetected inversion of the region combined with positive selection could also have led to these results. Although our data fail to strongly confirm prior evidence of positive selection in recent human history, the region clearly bears more detailed examination.

References for the Supplementary Notes

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403-410 (1990).
- Altshuler, D. et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407, 513-516 (2000).
- Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res* 14, 789-801 (2004).
- Cheung, J. et al. Recent segmental and gene duplications in the mouse genome. *Genome Biol* 4, R47 (2003).
- Chuang, J.H. & Li, H. Functional bias and spatial organization of genes in mutational hot and cold regions of the human genome. *PLoS Biol* 2, E29 (2004).
- Collins, F. S., Brooks, L. D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* 8, 1229-1231 (1998).
- Deinard, A. S. and Kidd, K. Identifying conservation units within captive chimpanzee populations. *American Journal of Physical Anthropology* 111: 25-44 (2000).
- Do, H. H. & Rahm, E. in *Proceedings 9. International Conference on Extending Database Technology* (Springer LNCS, 2004).
- Excoffier, L. Human demographic history: refining the recent African origin model. *Curr Opin Genet Dev* 12, 675-685 (2002).
- Fay, J.C. and Wu, C.-I. Hitchhiking Under Positive Darwinian Selection. *Genetics* 155: 1405-1413 (2000).
- Fischer, A., Wiebe, V., Paabo, S. and Przeworski, M. Evidence for a complex demographic history of chimpanzees. *Mol Biol Evol* 21, 799-808 (2004).
- Fujiyama, A. et al. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* 15, 224-230 (2005).
- Furey, T. S. et al. Analysis of Human mRNAs With the Reference Genome Sequence Reveals Potential Errors, Polymorphisms, and RNA Editing. *Genome Res* 14, 2034-2040 (2004).
- Graham, P. L. et al. Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method. *Statistics in Medicine* 22, 2071-2083 (2003).
- Huang, H. et al. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol* 5, R47. (2004)
- Huang, X., Wang, J., Aluru, S., Yang, S. P., and Hillier, L. PCAP: A whole-genome assembly program. *Genome Research* 13: 2164-2170 (2003).
- Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A* 101, 13994-4001 (2004).
- Inlow, J. K. & Restifo, L. L. Molecular and comparative genetics of mental retardation. *Genetics* 166, 835-81 (2004).

- International HapMap Consortium. The International HapMap Project. *Nature* 426, 789-96 (2003).
- International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945 (2004).
- Jaffe, D. B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Research* 13: 91-96 (2003).
- Kaessmann, H., Wiebe, V., and Pääbo, S. Extensive nuclear DNA sequence diversity among chimpanzees. *Science* 286: 1159-1162 (1999).
- Kent, W. J. BLAT -- the BLAST-like alignment tool. *Genome Research* 12: 656-664 (2002).
- Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences* 100: 11484-11489 (2003).
- Makova, K. D. and Li, W.-H. Strong male-driven evolution of DNA-sequences in humans and apes. *Nature* 416, 624-626 (2002).
- Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297-304 (2000).
- Navarro, A. & Barton, N. H. Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes. *Science* 300, 321-4 (2003).
- Nickerson, E. and Nelson, D. L. Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. *Genomics* 50, 368-372 (1998).
- Patterson, N.J. How old is the most recent ancestor of the two copies of an allele? *Genetics* 169(2): 1093-1104 (2005). Pmid: 15520271
- Przeworski, M. The Signature of Positive Selection at Randomly Chosen Loci. *Genetics* 160(3): 1179-89 (2002). Pmid: 11901132
- Rat Genome Sequencing Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428, 493-521 (2004).
- Reich, D. E. et al. Linkage disequilibrium in the human genome. *Nature* 411, 199-204 (2001).
- Rosenberg, H. F. and Feldmann, M. W. The relationship between coalescence times and population divergence times. Oxford University Press, Oxford (2002).
- Schwartz, S. et al. Human-mouse alignments with BLASTZ. *Genome Research* 13: 103-107 (2003).
- Schaffer, S. F., Foo, C., Gabriel, S. Reich, D., Daly, M. J., Altshuler, D. Calibrating a coalescent simulation oh human genome sequence variation. *Genome Res* 15, 1576-1583 (2005).
- She, X. et al. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927-30 (2004).
- Siepel, A. & Haussler D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* 21, 468-88 (2004).

- Sved J. & Bird A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci U S A* 87, 4692-6 (1990).
- Taylor, J., Tyekucheva, S., Zody, M. Chiaromonte, F., Makova K. D. Strong and weak male mutation bias at different sites in the primate genomes: insights fomr the human-chimpanzee comparison. *Mol Biol Evol* 23, 565-573 (2006).
- Torrents, D., Suyama, M., Zdobnov, E. and Bork, P. A genome-wide survey of human pseudogenes. *Genome Research* 13, 2559-2567 (2003).
- Tuzun, E., Bailey, J. A. & Eichler, E. E. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res* 14, 493-506 (2004).
- Watanabe, H. et al. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* 429, 382-8 (2004).
- Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13, 555-6 (1997). [<http://abacus.gene.ucl.ac.uk/software/paml.html>]
- Yu, N. et al. Low nucleotide diversity in chimpanzees and bonobos. *Genetics* 164: 1511-1518 (2003).
- Yunis, J. J. and Prakash, O. The origin of man: A chromosomal pictorial legacy. *Science* 215: 1525-1530 (1982).
- Zechner, U. et al. A high density of X-linked genes for general cognitive ability: a run-away process shaping human evolution? *Trends Genet* 17, 697-701 (2001).

Chapter 3: The dog genome

In this chapter, we describe the first comprehensive comparative analysis of the dog, human and mouse genome sequences.

This work was first published as part of

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).

The full publication is attached as Appendix 2. Supplementary notes can be found at the end of the chapter. Supplementary data is available online from <http://www.nature.com/nature>

The text in this chapter was written with significant contributions from members of the analysis consortium.

[This page is intentionally left blank]

We report a high-quality draft sequence of the domestic dog. The dog is of particular interest to the field of comparative genomics because the species provides key evolutionary information and because existing breeds display extraordinary phenotypic diversity for morphological, physiological and behavioral traits. Sequence comparison with the primate and rodent lineages sheds light on the structure and evolution of genomes and genes. Notably, the majority of the most highly conserved non-coding sequences in mammalian genomes are clustered near a small subset of genes with key roles in development.

Man's best friend, *Canis familiaris*, occupies a special niche in genomics. The unique breeding history of the domestic dog provides an unparalleled opportunity to explore the genetic basis of disease susceptibility, morphological variation and behavioral traits. The dog's position within the mammalian evolutionary tree also makes it an important guide for comparative analysis of the human genome.

The history of domestic dogs traces back at least 15,000 and possibly over 100,000 years to their original domestication from the gray wolf in East Asia¹⁻⁴. Dogs evolved in a mutually beneficial relationship with humans, sharing living space and food sources. In recent centuries humans have selectively bred dogs that excel at herding, hunting, and obedience, and, in the process, created breeds rich in behaviors that both mimic our own and support our needs. Similarly, dogs were bred for desired physical characteristics such as size, skull shape, coat color and texture⁵, producing breeds with closely delineated morphologies. This evolutionary experiment has produced the most diverse domestic species, harboring more morphological diversity than exists within the remainder of the family Canidae⁶.

As a consequence of these stringent breeding programs and periodic population bottlenecks during, for example, the world wars, many of the ~400 modern dog breeds also exhibit high prevalence of specific diseases, including cancers, blindness, heart disease, cataracts, epilepsy, hip dysplasia and deafness^{7,8}. The majority of these diseases are also commonly seen in the human population, and clinical manifestations in the two species are often similar⁹. The high prevalence of specific diseases within certain breeds, suggest that a limited number of loci underlie each disease, making their genetic dissection potentially more tractable in dogs than in humans¹⁰.

Genetic analysis of traits in dogs is enhanced by the close relationship between humans and canines in modern society. Through the efforts of the American Kennel Club (AKC), and similar organizations worldwide, extensive genealogies are easily accessible for most purebred dogs. With the exception of human, dog is the most intensely studied animal in medical practice, with detailed family history and pathology data often available⁸. Using genetic resources

developed over the last 15 years¹¹⁻¹⁶, researchers have already identified mutations in genes underlying ~25 Mendelian diseases^{17,18}. There are also growing efforts to understand the genetic basis of phenotypic variation, such as skeletal morphology^{10,19}.

The dog is similarly important for comparative analysis of mammalian genome biology and evolution. The four mammalian genomes that have been intensely analyzed to date (human^{20,21,22}, chimpanzee²³, mouse²⁴ and rat²⁵) represent only one of the four clades of placental mammals (*Euarchontoglires*). The dog represents the neighboring clade, *Laurasiatheria*²⁶. It thus serves as an outgroup to the *Euarchontoglires* and increases the total branch length of the current tree of fully sequenced mammals, thereby providing additional statistical power to search for conserved functional elements in the human genome^{24,27-33}; it helps us draw inferences about the common ancestor of the two clades, called the boreoeutherian ancestor; and it provides a bridge to the two remaining clades (*Afrotheria* and *Xenarthra*) that will be helpful for anchoring low-coverage sequence currently being produced from such species as elephant and armadillo²⁸.

Here, we report a high-quality draft sequence of the dog genome covering ~99% of the euchromatic genome. The completeness, nucleotide accuracy, sequence continuity and long-range connectivity are extremely high; these quality measures exceed those for the recent draft sequence of the mouse genome²⁴, reflecting improved algorithms, higher quality data, deeper coverage and intrinsic genome properties.

We have analyzed these data to study genome structure, gene evolution, haplotype structure and phylogenetics of the dog. The key findings include:

- The evolutionary forces molding the mammalian genome differ among lineages, with the average transposon insertion rate being lowest in dog, the deletion rate being highest in mouse and the nucleotide substitution rate being lowest in human.

- Human-dog comparison shows that ~5.3% of the human genome contains functional elements that have been under purifying selection in both lineages. Nearly all of these elements are confined to regions that have been retained in mouse, indicating that they represent a common set of functional elements across mammals.

- Fully 50% of the most highly conserved non-coding sequence in the genome shows striking clustering in ~200 gene-poor regions, most of which contain genes with key roles in establishing or maintaining cellular identity, such as transcription factors or axon guidance receptors.

- Sets of functionally related genes show highly similar patterns of evolution in the human and dog lineages. This suggests caution about interpreting accelerated evolution in human

relative to mouse as representing human-specific innovations (such as in genes involved in brain development), because comparable acceleration is often seen in the dog lineage.

Generating a draft genome sequence

We sequenced the genome of a female Boxer using the whole-genome shotgun (WGS) approach^{22,24} (see Methods). A total of 31.5 million sequence reads, providing ~7.5-fold sequence redundancy, were assembled with an improved version of the ARACHNE program³⁴, resulting in an initial assembly (CanFam1.0) used for much of the analysis below and an updated assembly (CanFam2.0), with minor improvements (Table 1).

Genome assembly. The recent genome assembly spans a total distance of 2.41 Gb, consisting of 2.38 Gb of nucleotide sequence with the remaining 1% in captured gaps. The assembly has extremely high continuity. The N50 contig size is 180 kb (that is, half of all bases reside in a contiguous sequence of 180 kb or more) and the N50 supercontig size is 45.0 Mb (Table 1). In particular, this means that the majority of genes should contain no sequence gaps and that most canine chromosomes (mean size 61 Mb) have nearly all of their sequence ordered and oriented within one or two supercontigs. Notably, the sequence contigs are ~50-fold larger than the earlier survey sequence of the Standard Poodle¹⁶.

The assembly was anchored to the canine chromosomes using data from both RH and cytogenetic maps^{11,13,14}. Roughly 97% of the assembled sequence was ordered and oriented on the chromosomes, showing an excellent agreement with the two maps. There were only three discrepancies, which were resolved by obtaining additional FISH data from the sequenced Boxer. The 3% of the assembly that could not be anchored consists largely of highly repetitive sequence, including eight supercontigs of 0.5-1.0 Mb composed almost entirely of satellite sequence.

The nucleotide accuracy and genome coverage of the assembly is high. Of the bases in the assembly, 98% have quality scores exceeding 40, corresponding to an error rate of less than 10^{-4} and comparable to the standard for finished human sequence³⁵. When we directly compared the assembly to 760 kb of finished sequence (in regions where the Boxer is homozygous to eliminate differences due to polymorphisms; see below), we found that the draft genome sequence covers 99.8% of the finished sequence and that bases with quality scores exceeding 40 have an empirical error rate of 2×10^{-5} .

Explaining the high sequence continuity. The dog genome assembly has superior sequence continuity (180 kb) than the WGS assembly of the mouse genome (25 kb) obtained several years ago²⁴. At least three factors contribute to the higher connectivity of the dog assembly (see Supplementary Notes). First, a new version of ARACHNE with improved

Table 1: Assembly statistics for CanFam1.0 and 2.0

	CanFam1.0	CanFam2.0
N50 contig size	123 kb	180 kb
N50 supercontig size	41.2 Mb	45.0 Mb
Assembly size, total bases	2.360 Gb	2.385 Gb
Number of anchored supercontigs	86	87
Portion of genome in anchored supercontigs	96%	97%
Sequence in anchored bases	2.290 Gb	2.309 Gb
Portion of assembly in gaps	0.9%	0.8 %
Estimated genome size: anchored bases, spanned gaps (21Mb and 18Mb respectively) and centromeric sequence (3Mb each)	2.411 Gb	2.445 Gb
Portion of assembly in 'certified regions', without assembly inconsistency	99.3%	99.6%

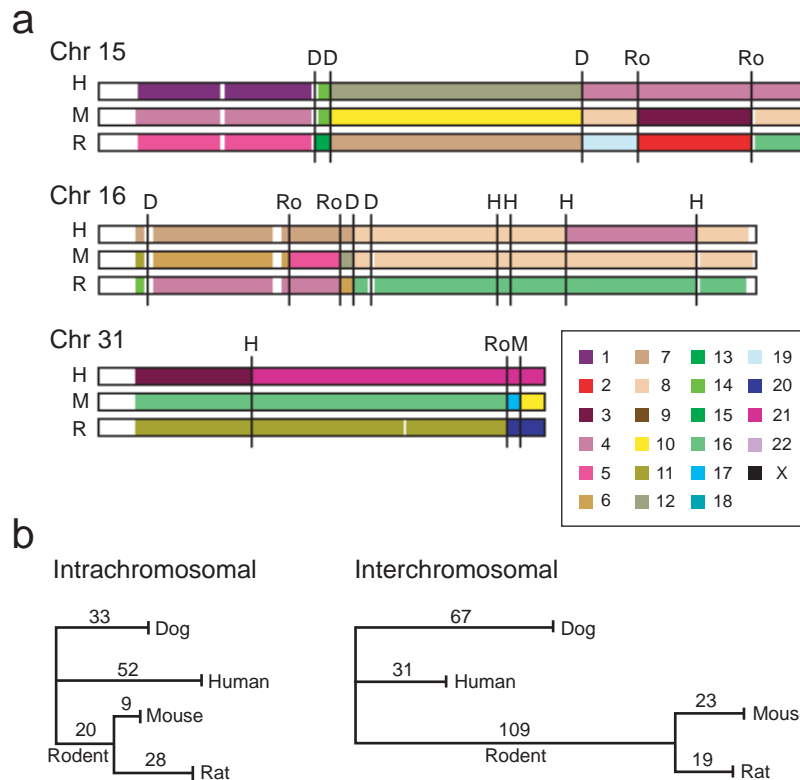


Figure 1. Conserved synteny between the human, dog, mouse and rat genomes. (a) Diagram of syntenic blocks (>500 kb) along dog chromosomes 15, 16 and 3, with colors indicating the chromosome containing the syntenic region in other species. Synteny breakpoints were assigned to one of five lineages: dog (D), human (H), mouse (M), rat (R) or the common rodent ancestor (Ro). (b) Lineage-specific intrachromosomal and interchromosomal breaks displayed on phylogenetic trees. Intrachromosomal breaks are seen more frequently in the human lineage than in mouse and rat, while interchromosomal breaks are somewhat more common in dog and considerably more common in rodents.

algorithms was used. Assembling the dog with the previous software decreases N50 contig size from 180 kb to 61 kb, while assembling the mouse with the new version increased it from 25 kb to 35 kb. Second, the amount of recently duplicated sequence is roughly two-fold lower in dog than mouse; this improves contiguity because sequence gaps in both organisms tend to occur in recently duplicated sequence. Third, the dog sequence data has both higher redundancy (7.5 vs. 6.5-fold) and higher quality (in terms of read length, pairing rate and tight distribution of insert sizes). The contig size for dog drops by about 32% when the data redundancy is decreased from 7.5- to 6.5-fold. A countervailing influence is that the dog genome contains polymorphism, while the laboratory mouse is fully inbred .

Genome landscape and evolution

Our understanding of the evolutionary processes that shape mammalian genomes has greatly benefited from the comparative analysis of primate^{21, 23}, and rodent^{24, 25} genomes sequenced to date. However, the rodent genome is highly derived relative to that of the common ancestor of the eutherian mammals. The dog genome, as the first extensive sequence from an outgroup to the clade including primates and rodents, therefore offers a fresh perspective on mammalian genome evolution. Accordingly, we examined the rates and correlations of large-scale rearrangement, transposon insertion, deletion, and nucleotide divergence across three major mammalian orders (primates, rodents and carnivores).

Conserved synteny and large-scale rearrangements. We created multi-species synteny maps from anchors of unique, unambiguously aligned sequences (see Supplementary Notes), showing regions of conserved synteny among dog, human mouse and rat. Roughly 94% of the dog genome lies in such regions of conserved synteny with the three other species.

Given a pair of genomes, we refer to a ‘syntenic segment’ as a region that runs continuously without alterations of order and orientation and a ‘syntenic block’ as a region that is contiguous in two genomes but may have undergone internal rearrangements. The syntenic breakpoints between blocks thus primarily reflect interchromosomal exchanges, while breakpoints between syntenic segments reflect intrachromosomal rearrangements. In the analysis below, we focus on syntenic segments of at least 500 kb.

We identified a total of 391 syntenic breakpoints across dog, human, mouse and rat (Figure 1). With data for multiple species, it is possible to assign events to specific lineages. We counted the total number of breakpoints along the human, dog, mouse and rat lineages, with the values for each rodent lineage reflecting all breakpoints since the common ancestor with human. The total number of breakpoints in the human lineage is substantially smaller than in dog, mouse

or rat lineages (83 vs. 100, 161 and 176). However, there are more intrachromosomal breakpoints in the human lineage than in dog (52 vs 33).

Although the overall level of genomic rearrangement has been much higher in rodent than in human, comparison with dog shows that there are regions where the opposite is true. In particular, of the many intrachromosomal rearrangements previously observed between human chromosome 17 and the orthologous mouse sequence²⁴, the majority has occurred in the human lineage (see Supplementary Notes). Human chromosome 17 is rich in segmental duplications and gene families²¹, which may contribute to its genomic fragility^{37,38}.

Genomic insertion and deletion. The dog has a smaller euchromatic genome than mouse and human by ~150 Mb and ~500 Mb, respectively. The smaller total size is reflected at the local level, with 100 kb blocks of conserved synteny in dog corresponding to regions whose median size is ~3% larger in mouse and ~15% larger in human.

To understand the balance of forces that determine genome size, we studied the alignments of the human, mouse and dog genomes (Figure 2). In particular, we identified the lineage-specific interspersed repeats within each genome, consisting of particular families of SINE, LINE and other transposable elements that are readily recognized by sequence analysis. The remaining sequence was annotated as ‘ancestral’, consisting of both ancestral unique sequence and ancestral repeat sequence; these two categories were combined because the power to recognize ancient transposon-derived sequences degrades with repeat age, particularly in the rapidly diverging mouse lineage²⁴.

This comparative analysis indicates that different forces account for the smaller genome sizes in dog and mouse relative to human. The smaller size of the dog genome is primarily due to the presence of substantially less lineage-specific repeat sequence in dog (334 Mb) than in human (609 Mb) or mouse (954 Mb). This reflects a lower activity of endogenous retrovirus and DNA transposons (~26,000 extant copies in dog vs. ~183,000 in human), as well as the fact that the SINE element in dog is smaller than in human (although similar to that in mouse). As a consequence, the total proportion of repetitive elements (both lineage-specific and ancestral) recognizable in the genome is lower for dog (34%) than for mouse (40%) and human (46%). By contrast, the smaller size of the mouse genome is primarily due to a higher deletion rate.

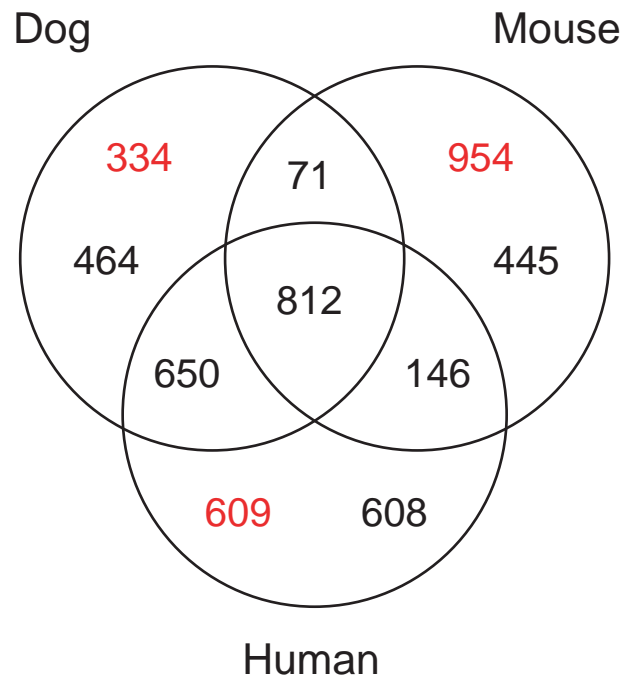


Figure 2. Venn diagram showing the total length (Mb) of alignable and unique sequences in the euchromatic portions of the dog, human and mouse genomes, as inferred from genome-wide BLASTZ alignments (see Methods and Supplementary Notes). Overlapping partitions represent orthologous ancestral sequences. Each lineage-specific partition is further split into the total length of sequence classified as either lineage-specific interspersed repeats (red) or non-repetitive sequence (black). The latter is assumed to primarily represent ancestral sequences deleted in the two other species.

Specifically, the amount of extant ‘ancestral sequence’ is much lower in mouse (1474 Mb) than in human (2216 Mb) or dog (1997 Mb). Assuming an ancestral genome size of 2.8 Gb²⁴ and that deletions occur continuously, this suggests that the rate of genomic deletion in the rodent lineage has been approximately two and a half times higher than in the dog and human lineages (see Supplementary Notes). As a consequence, the human genome shares ~650 Mb more ancestral sequence with dog than with mouse, despite our more recent common ancestor with the latter.

Active SINE family. Despite its relatively low proportion of transposable element-derived sequence, the dog genome contains a highly active carnivore-specific SINE family (defined as SINEC_Cf; RepBase release 7.11)¹⁶. The element is so active that many insertion sites are still segregating polymorphisms that have not yet reached fixation. Of ~87,000 young SINEC-Cf elements (defined by low divergence from the consensus sequence), nearly 8% are heterozygous within the draft genome sequence of the Boxer. Moreover, comparison of genome sequence between Boxer and Standard Poodle reveals more than 10,000 insertion sites that are bimorphic, with thousands more certain to be segregating in the dog population^{16,39}. By contrast, the number of polymorphic SINE insertions in the human genome is estimated to be fewer than a thousand⁴⁰.

The biological impact of these segregating SINE insertions is unknown. SINE insertions can be mutagenic through direct disruption of coding regions, as well as indirect effects on regulation and processing of mRNAs³⁹. Such SINE insertions have already been shown to be responsible for two diseases; narcolepsy and centronuclear myopathy in dog^{41,42}. It is conceivable that the genetic variation resulting from these segregating SINE elements has provided important raw material for the selective breeding programs that produced the wide phenotypic variations among modern dog breeds^{16,43}.

Sequence composition. The human and mouse genomes differ significantly in sequence composition, with the human having slightly lower average G+C content (41% vs. 42%) but much greater variation across the genome. The dog closely resembles the human in its distribution of G+C content (Figure 3a; Spearman’s $\rho = 0.85$ for dog-human and 0.76 for dog-mouse), even if we consider only nucleotides that can be aligned across all three species. The wider distribution of G+C content in human and dog is thus likely to reflect the boreoeutherian ancestor^{44,45}, with the more homogeneous composition in rodents having arisen primarily by lineage-specific changes in substitution patterns^{46,47} rather than deletion of sequences with extreme G+C content.

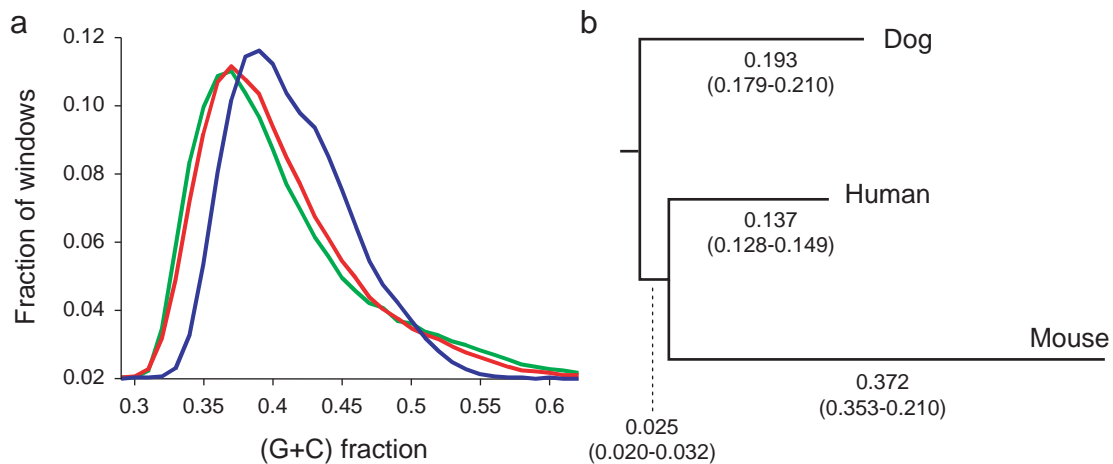


Figure 3. Sequence composition and divergence rates. a, Distribution of G+C content in 10 kb windows across the genome in dog (green), human (red) and mouse (blue). b, Median lineage-specific substitution rates based on analysis of ancestral repeats aligning across all three genomes. Analysis was performed in non-overlapping 1 Mb-windows across the dog genome containing at least 2 kb of aligned ancestral repeat sequence (median 8.8 kb). The tree was rooted with the consensus sequences from the ancestral repeats. Numbers in parentheses give the 20th and 80th percentiles across the windows studied.

Table 2: Substitution rates and evolutionary constraint (Ka/Ks) for 1:1:1 orthologs among dog, mouse and human

	Median and 20/80 percentiles			Spearman's <i>rho</i>		
	Dog*	Mouse	Human	Dog Human	Dog Mouse	Human Mouse
Ks	0.210 (0.138-0.322)	0.416 (0.310-0.558)	0.139 (0.0928-0.214)	0.47	0.50	0.52
Ka	0.021 (0.006-0.051)	0.038 (0.013-0.087)	0.017 (0.005-0.040)	0.87	0.87	0.86
Ka/Ks	0.095 (0.030-0.221)	0.088 (0.031-0.197)	0.112 (0.034-0.272)	0.80	0.85	0.82

* Estimates are based on unrooted tree. Dog branch thus includes the branch from the boreoeutherian ancestor to the primate-rodent split.

Rate of nucleotide divergence. We estimated the mean nucleotide divergence rates in 1 Mb windows along the dog, human and mouse lineages based on alignments of all ancestral repeats, using the consensus sequence for the repeats as a surrogate outgroup (Figure 3b; Supplementary Notes).

The dog lineage has diverged more rapidly than the human lineage (median relative rate of 1.18, longer branch length in 95% of windows), but at only half the rate of the mouse lineage (median relative rate of 0.48, longer branch length in 100% of windows). The absolute divergence rates are somewhat sensitive to the evolutionary model used and the filtering of alignment artifacts, but the relative rates appear to be robust and are consistent with estimates from smaller sequence samples with multiple outgroups^{28, 48, 49}. The lineage-specific divergence rates (human < dog < mouse) are likely explained by differences in metabolic rates^{50, 51} or generation times^{52, 53}, although the relative contributions of these factors remain unclear⁴⁹.

Correlation in nucleotide divergence. As seen in other mammalian genomes²³⁻²⁵, the average nucleotide divergence rate across 1 Mb windows varies significantly across the dog genome (coefficient of variation = 0.11, versus 0.024 expected under a uniform distribution). This regional variation shows significant correlation in orthologous windows across the dog, human and mouse genomes, but the strength of the correlation appears to decrease with total branch length (pair-wise correlation for orthologous 1 Mb windows: Spearman's $\rho = 0.49$ for dog-human and 0.24 for dog-mouse). Lineage-specific variation in the regional divergence rates may be coupled with changes in factors such as sequence composition or chromosomal position^{23, 54}. Consistent with this, the ratios of lineage-specific divergence rates in orthologous windows are positively correlated with the ratios of current G+C content in the same windows (dog-human: Spearman's $\rho = 0.16$; dog-mouse: $\rho = 0.24$).

Male mutation bias. Comparison of autosomal and X chromosome substitution rates can be used to estimate the relative mutation rates in the male and female germ lines (α), because the X chromosome resides in females twice as often as in males. Using the lineage-specific rates from ancestral repeats, we estimate α as 4.8 for the lineage leading to human, and 2.8 for the lineages leading to both mouse and dog. These values are intermediate between recent estimates from murids^{24, 25} and from hominids²³, and suggests that male mutation bias may have increased in the lineage leading to humans.

Mutational hotspots and chromosomal fission. Genome comparison of human with both chicken⁵⁵ and chimpanzee²³ have previously revealed that sequences close to a telomere tend to have elevated divergence rates and G+C content relative to interstitial sequences. It has been unclear whether this subtelomeric elevation is an inherent characteristic of the sequence

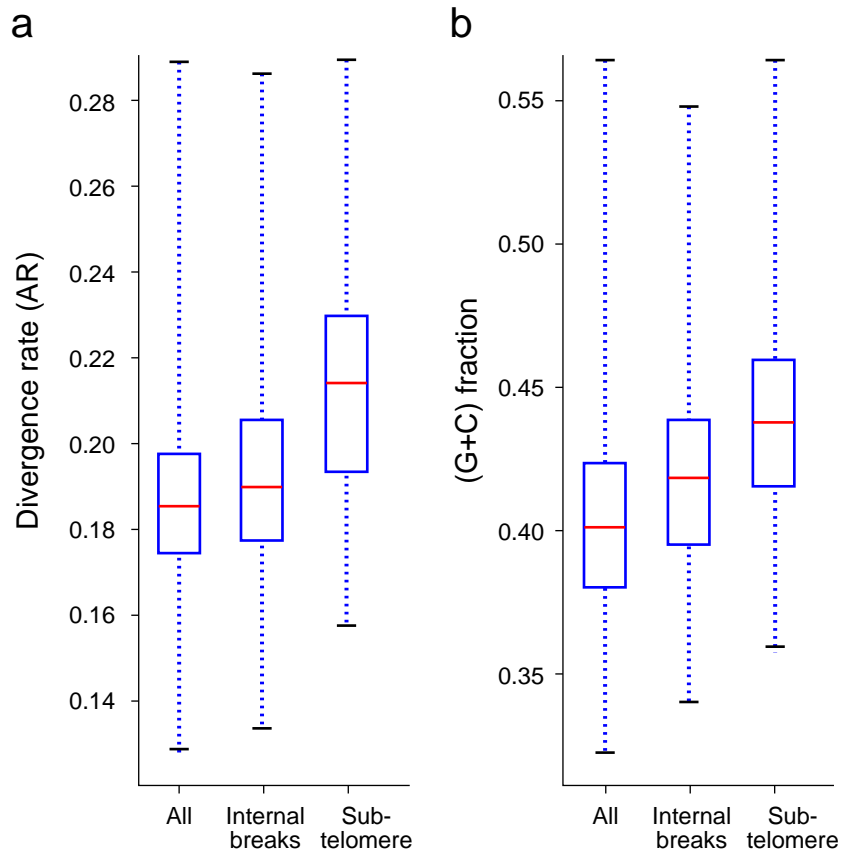


Figure 4. Dependence between divergence, (G+C) content and human-dog synteny breakpoints. a, distribution of divergence on the dog-lineage (as estimated from ancestral repeats) for all 1 Mb non-overlapping windows, windows centered on a non-telomeric synteny break, and windows within 1 Mb of a telomere. (Red line = median, box = upper and lower quartiles, whiskers = range of data) b, distribution of (G+C) content in the dog genome for all 1 Mb non-overlapping windows, windows within 1 Mb of a non-telomeric synteny break, and windows within 1 Mb of a telomere.

itself or a derived characteristic causally connected with its chromosomal position. We find a similar elevation in both divergence (median elevation = 15%, $p_{\text{Mann-Whitney}} < 10^{-5}$) and G+C content (median elevation = 9%, $p_{\text{Mann-Whitney}} < 10^{-9}$) for subtelomeric regions along the dog lineage, with a sharp increase towards the telomeres (Figure 4).

This phenomenon is manifested at other synteny breaks, not only those at telomeres. We also observed a significant elevation in divergence and G+C content in interstitial regions that are sites of syntenic breakpoints^{54,56}. This indicates that these properties are correlated with susceptibility of regions to chromosomal breakage.

Proportion of genome under purifying selection

One of the striking discoveries that emerged from the comparison of the human and mouse genomes^{21,24} was the inference that ~5.2% of the human genome shows greater-than-expected evolutionary conservation (compared to the background rate seen in ancestral repeat elements, which are presumed to be non-functional). This proportion greatly exceeds the 1-2% that can be explained by protein coding regions alone. The extent and function of the large fraction of non-coding conserved sequence remain unclear⁵⁷, but it is likely to include regulatory elements, structural elements and RNA genes.

Low turnover of conserved elements. We repeated the analysis of conserved elements based on the human and dog genomes. Briefly, the analysis involves calculating a conservation score S_{HD} , normalized by the regional divergence rate, for every 50 bp-window in the human genome that can be aligned to dog. The distribution of conservation scores for all genomic sequences is compared to the distribution in ancestral repeat sequences (which are presumed to be diverging at the local neutral rate), showing a clear excess of sequences with high conservation scores. By subtracting a scaled neutral distribution from the total distribution, one can estimate the distribution of conservation scores for sequences under purifying selection. Moreover, for a given sequence with conservation score S_{HD} , one can also assign a probability, $p_{\text{selection}}(S_{\text{HD}})$, that the sequence is under purifying selection (see²⁴ and Supplementary Notes).

The human-dog comparison indicates that ~5.3% of the human genome is under purifying selection (Figure 5a), which is equivalent to the proportion estimated from human-rodent analysis. The obvious question is whether the bases conserved between human and dog coincide with the bases conserved between humans and rodents^{25,58}. Because the conservation scores do not unambiguously assign sequences as either selected or neutral (but rather only assign probability scores for selection), we cannot directly compare the conserved bases. We therefore devised the following alternative approach.

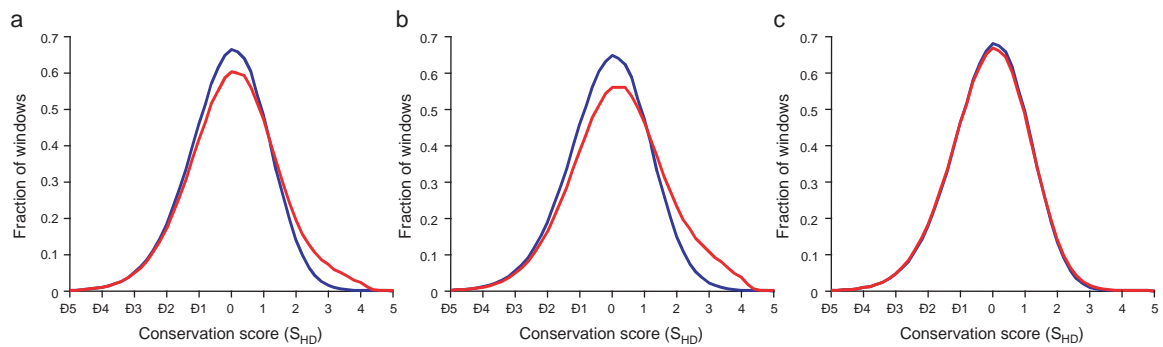


Figure 5. Conservation of orthologous sequence between human and dog.

a, Histogram of conservation scores, S , for all windows of 50-bp across the human genome with at least 20 bases of orthologous sequence aligning to the dog genome, for all aligning sequences (red) and for ancestral repeat sequence only (blue).

b, Conservation scores for the subset of window that also have at least 20 bases of orthologous sequence aligning to the mouse genome. c, Conservation scores of the complementary subset of windows lacking such orthologous sequence in mouse.

We repeated the human-dog analysis, dividing the 1462 Mb of orthologous sequence between human and dog into those regions with (812 Mb) or without (650 Mb) orthologous sequence in mouse (Figure 2). The first set shows a clear excess of conservation relative to background, corresponding to ~5.2% of the human genome (Figure 5b). In contrast, the second set shows little or no excess conservation, corresponding to at most 0.1% of the human genome (Fig 6c). This implies that hardly any of the functional elements conserved between human and dogs have been deleted in the mouse lineage (see Supplementary Notes).

The results strongly suggest that there is a common set of functional elements across all three mammalian species, corresponding to ~5% of the human genome (~150 Mb). These functional elements reside largely within the 812 Mb of ancestral sequence common to human, mouse and dog. If we eliminate ancestral repeat elements within this shared sequence as largely non-functional, the majority of functional elements can be localized to 634 Mb and constitute, perhaps, 24% of this sequence.

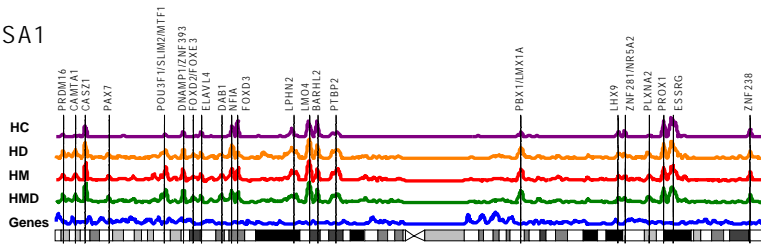
It should be noted that the estimate of ~5% pertains to conserved elements across distantly related mammals. It is possible that there are additional weakly constrained or recently evolved elements within narrow clades (for example, primates) that can only be detected by sequencing more closely related species ²⁹.

Clustering of highly conserved non-coding elements. We next explored the distribution of conserved non-coding elements (CNEs) across the mammalian genomes. For this purpose, we calculated a conservation score S_{HMD} based on simultaneous conservation across all three species. We defined highly conserved non-coding elements (HCNEs) to be 50-bp windows that do not overlap coding regions and for which $p_{\text{selection}}(S_{\text{HMD}})$, the probability of being under purifying selection given the conservation score, is at least 95%. We identified ~140,000 such windows (6.5 Mb total sequence), comprising ~0.2% of the human genome and representing the most conserved ~5% of all mammalian CNEs.

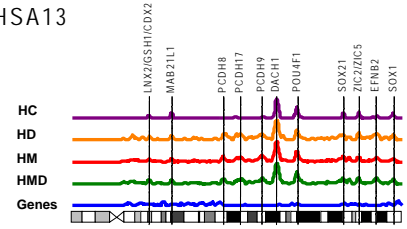
The density of HCNEs show striking peaks, when plotted in 1 Mb windows across the genome (Figure 6), with fully 50% lying in 204 regions that span less than 14% of the human genome. These regions are generally gene-poor, together containing only ~6% of all protein coding sequence.

The genes contained within these gene-poor regions are of particular interest. At least 182 of the 204 regions contain genes with key roles in establishing or maintaining cellular 'state'. At least 156 of the regions contain one or, in a few cases, several transcription factors involved in differentiation and development ⁵⁹. Another 26 regions contain a gene important to neuronal specialization and growth, including several axon guidance receptors. The proportion of

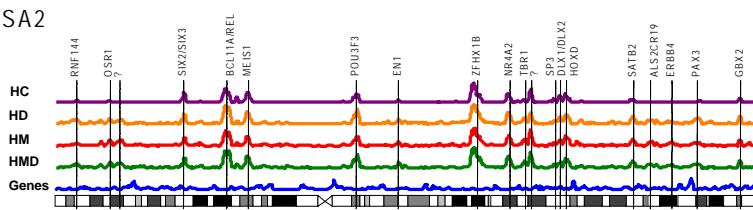
HSA1



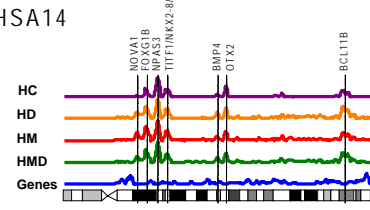
HSA13



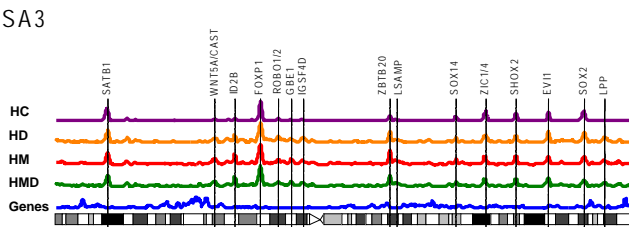
HSA2



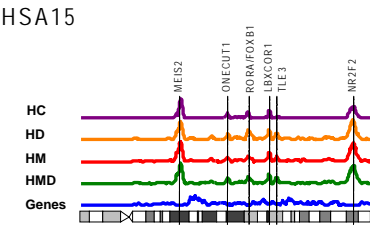
HSA14



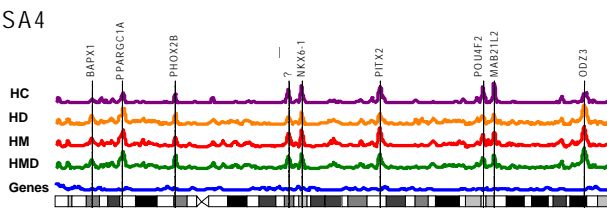
HSA3



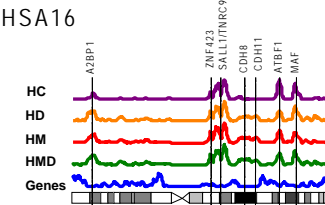
HSA15



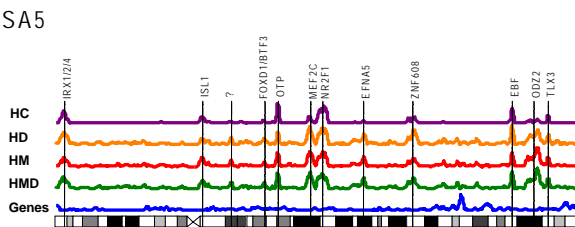
HSA4



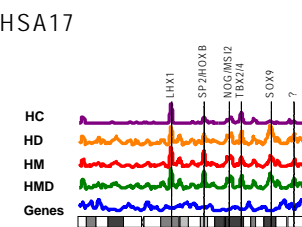
HSA16



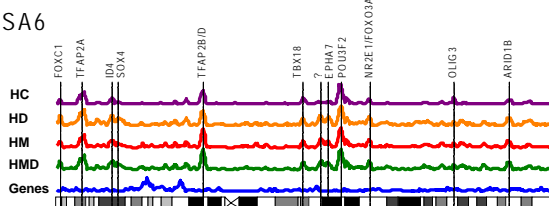
HSA5



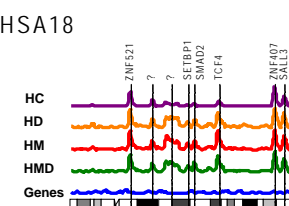
HSA17



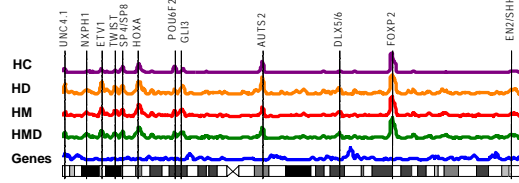
HSA6



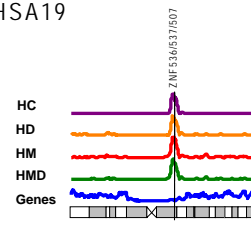
HSA18



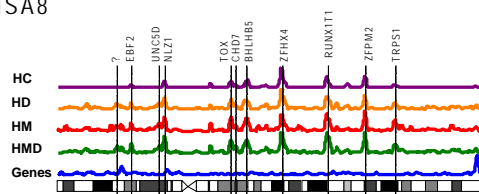
HSA7



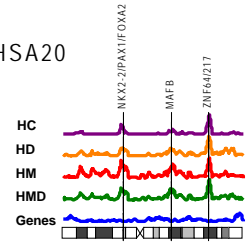
HSA19



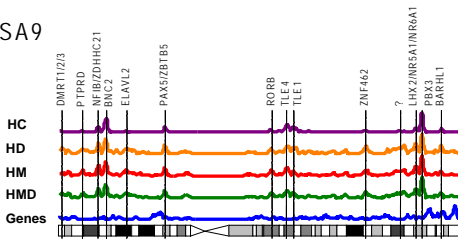
HSA8



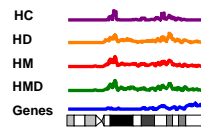
HSA20



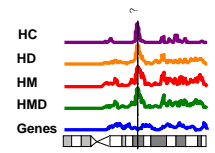
HSA9



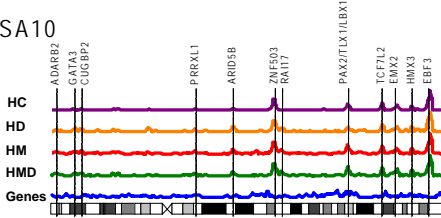
HSA21



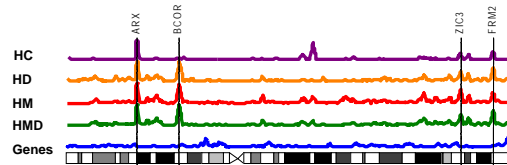
HSA22



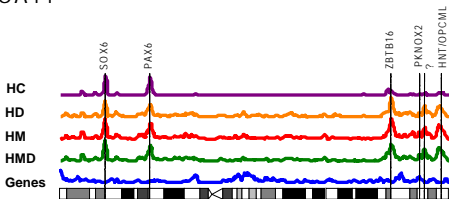
HSA10



HSAX



HSA11



HSA12

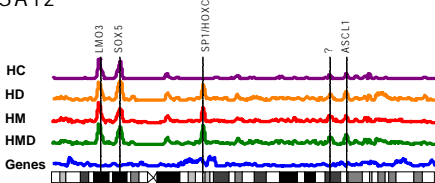


Figure 6. Density of the top 5% most conserved 50 bp windows aligned between human and chicken (HC), where the conservation score SHC is calculated as described in the supplementary notes, but without the AR normalization term. Density of the top 1% most conserved 50 bp windows out of all aligned windows between human and dog (HD) or mouse (HM), or between human, mouse and dog (HMD), and density of known genes, both in 1 Mb sliding windows across human chromosomes (blue).

developmental regulators is far greater than expected by chance ($P < 10^{-31}$; see Supplementary Notes).

We tested whether the HCNEs within these regions tend to cluster around the genes encoding development regulators. Analysis of the density of HCNEs in the intronic and intergenic sequences flanking every gene in the 204 regions revealed that the 197 genes encoding developmental regulators show an average of ~10-fold enrichment for HCNEs relative to the full set of 1285 genes in the regions (Figure 7). The enrichment sometimes extends into the immediately flanking genes.

We note that the 204 regions include nearly all of the recently identified clusters of conserved elements between distantly-related vertebrates, such as chicken and pufferfish^{55 59-62}. For example, they overlap 56 of the 57 large intervals containing conserved non-coding sequence identified between human and chicken⁵⁵. The mammalian analysis, however, detects vastly more CNEs (>100-fold more sequence than with pufferfish⁵⁹ and 2-3-fold more than with chicken) and identifies many more clusters. The limited sensitivity of these more distant vertebrate comparisons may reflect the difficulty of aligning short orthologous elements across such large evolutionary distances or the emergence of mammal-specific regulatory elements. In any case, mammalian comparative analysis may be a more powerful tool for elucidating the regulatory controls across these important regions.

Although the function of conserved non-coding elements is unknown, it seems likely based on recent studies^{59, 63-66} that many are likely to regulate gene expression. If so, the results above suggest that ~50% of all mammalian HCNEs may be devoted to regulating ~1% of all genes. In fact, the distribution may be even more skewed, as there are additional genomic regions with only slightly lower HCNE density than the 204 studied above. All of these regions clearly merit intensive investigation to assess indicators of regulatory function. We speculate that these regions may also harbor characteristic chromatin structure and modifications, potentially involved in establishment or maintenance of cellular state.

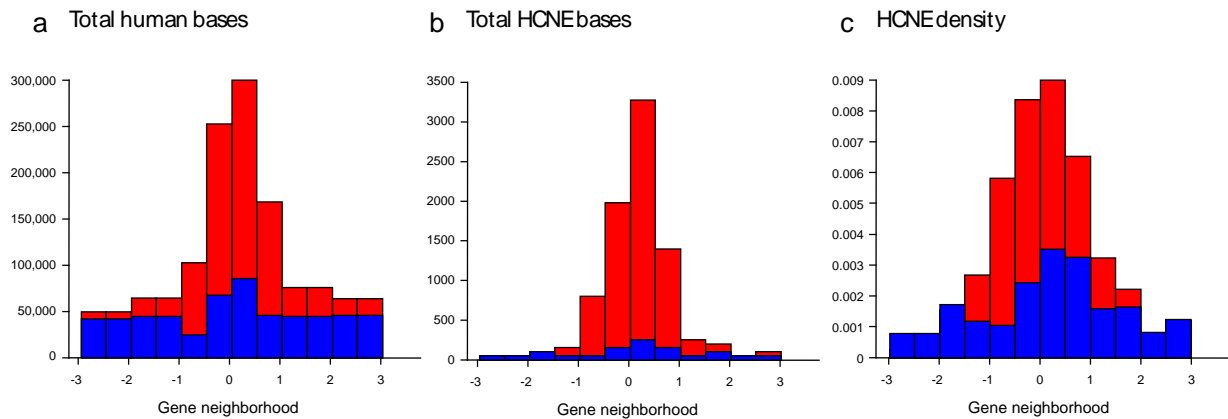


Figure 7. Enrichment of HCNEs in the immediate neighborhood of genes encoding developmental regulators in the 204 highly conserved regions (see text). Histograms show the median number of (a) total bases, (b) HCNE bases and (c) HCNE density in the intronic and surrounding intergenic sequence, for the 197 known or putative development regulators (indicated by top of red bar) and for all of the 1,285 genes (blue bar). The histogram is centered at the 5'-end of the gene and each bin corresponds to half of the normalized distance to the flanking upstream or downstream gene, as indicated. The sequences surrounding the developmental genes are typically longer, have more HCNE sequence and have a higher density of HCNE sequence than other genes in the regions. This suggests that they represent the primary targets of regulation in the regions.

Genes

Accurate identification of the protein-coding genes in mammalian genomes is essential for understanding the human genome, including cellular components, regulatory controls and evolutionary constraints. The number of human protein-coding genes has been a topic of considerable debate, with estimates steadily falling from ~100,000 to 20-25,000 over the past decade^{21, 22, 67-70}. We analyzed the dog genome to elucidate the human gene complement and to assess the evolutionary forces shaping mammals. (Below, 'gene' refers only to a protein-coding gene.)

Gene predictions in dog and human. We generated gene predictions for the dog genome using an evidence-based method (see Supplementary Notes). The resulting collection contains 19,300 dog gene predictions, with nearly all being clear homologs of known human genes.

The dog gene count is substantially lower than in the ~22,000 gene models in the current human gene catalog (Ensembl build 26). For many predicted human genes, we find no convincing evidence of a corresponding dog gene. Much of the excess in the human gene count is due to spurious gene predictions in the human genome¹²¹.

Gene duplications. Gene duplication is thought to contribute substantially to functional innovation^{69, 71}. We identified 216 gene duplications that are specific to the dog lineage and 574 that are specific to the human lineage, using the synonymous substitution rate K_S as a distance metric while taking care to discard likely pseudogenes. (The CanFam 2.0 assembly contains roughly two dozen additional gene duplications, mostly olfactory receptors.) Human genes are thus 2.7-fold more likely to have undergone duplication than are dog genes over the same time period. This may reflect increased repeat-mediated segmental duplication in the human lineage⁷².

Although gene duplication has been less frequent in dog than human, the affected gene classes are very similar. Prominent among the lineage-specific duplication are genes that function in adaptive immunity, innate immunity chemosensation and reproduction, as has been seen for other mammalian genomes^{24, 25, 69, 71}. Reproductive competition within the species and competition against parasites have thus been major driving forces in gene family expansion.

The two families with the largest numbers of dog-specific genes are the histone H2Bs and the α -interferons, which cluster in monophyletic clades when compared to their human homologs. This is particularly notable for the α -interferons, where the gene families within the 6 species (human, mouse, rat, dog, cat and horse) are apparently monophyletic. This may be due either to

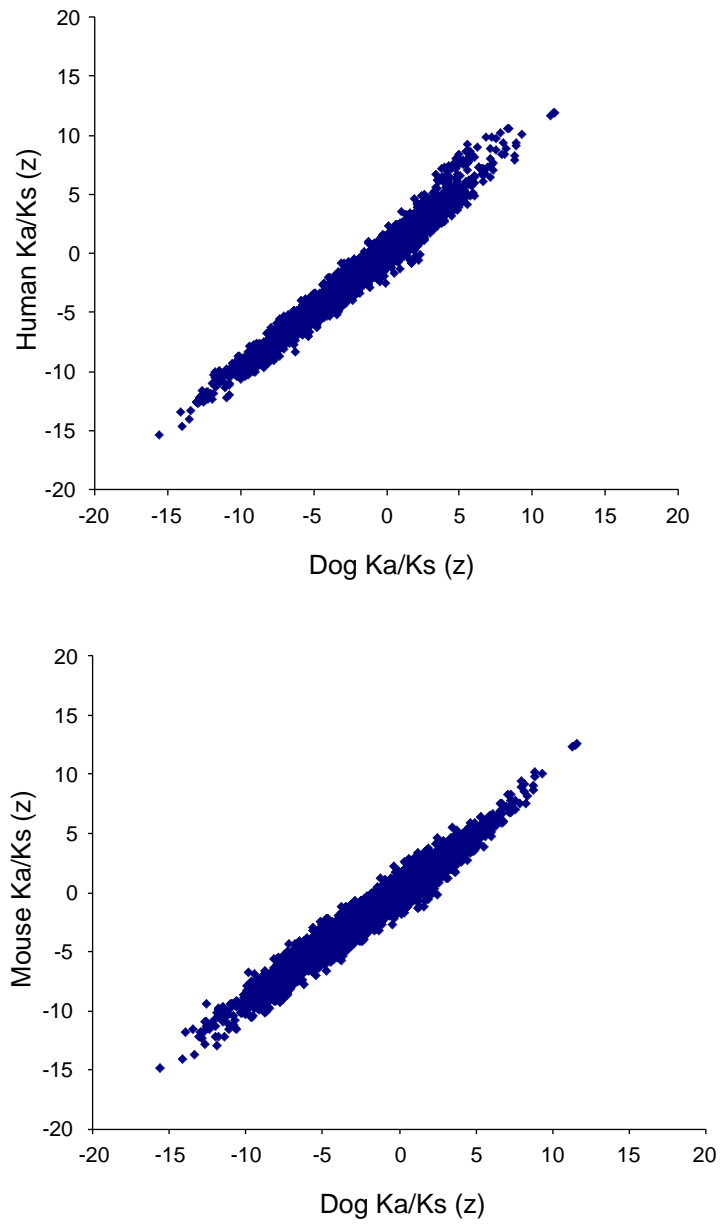


Figure 8. Correlation of Ka/Ks rank-sum z-scores for 4,950 gene sets between human and dog (top) and between mouse and dog (bottom).

coincidental independent gene duplication in each of the 6 lineages or ongoing gene conversion events that have homogenized ancestral gene duplicates⁷³.

Evolution of orthologous genes across three species. The dog genome sequence allows us for the first time to characterize the large-scale patterns of evolution in protein-coding genes across three major mammalian orders. We focused on a subset of 13,816 human-mouse-dog 1:1:1 orthologs. For each, we inferred the number of lineage-specific synonymous (Ks) and non-synonymous (Ka) substitutions along each lineage and calculated the Ka/Ks ratio (Table 2), a traditional measure of the strength of selection (both purifying and directional) on proteins⁷⁴.

The median Ka/Ks differs sharply across the three lineages ($p_{MW} < 10^{-44}$), with the dog lineage being intermediate between mouse and human. Population genetic theory predicts⁷⁵ that the strength of purifying selection should increase with effective population size (N_e). The observed relationship (mouse < dog < human) is thus consistent with the evolutionary prediction, given the expectation that smaller mammals tend to have larger effective population sizes⁷⁶.

We next searched for particular classes of genes showing deviations from the expected rate of evolution for a species. Such variation in rate (heterotachy) may point to lineage-specific positive selection or relaxation of evolutionary constraints⁷⁷. We developed a statistical method, similar to the recently-described Gene Set Enrichment Analysis (GSEA)⁷⁸⁻⁸⁰ to detect evidence of heterotachy for sets of functionally related genes (see Supplementary Notes). Briefly, the approach involves ranking all genes by Ka/Ks ratio, testing whether the set is randomly distributed along the list and assessing the significance of the observed deviations by comparison with randomly permuted gene sets. In contrast to previous studies focused on small numbers of genes with prior hypotheses of selection, this approach detects signals of lineage-specific evolution in a relatively unbiased manner and can provide context to the results of more limited studies.

A total of 4,950 overlapping gene sets were studied, defined by such criteria as biological function, cellular location or co-expression. Overall, the deviations between the three lineages are small, and median Ka/Ks ratios for particular gene sets are highly correlated for each pair of species (Figure 8). However, there is notably greater relative variation in human-mouse and dog-mouse comparisons than in human-dog comparisons (Figure 9).

This suggests that observed heterotachy between human and mouse must be interpreted with caution. For example, there is a great interest in the identification of genetic changes underlying the unique evolution of the human brain. A recent study⁸¹ highlighted 24 genes involved in brain development and physiology that show signs of accelerated evolution in the lineage leading from ancestral primates to humans when compared to their rodent orthologs. We

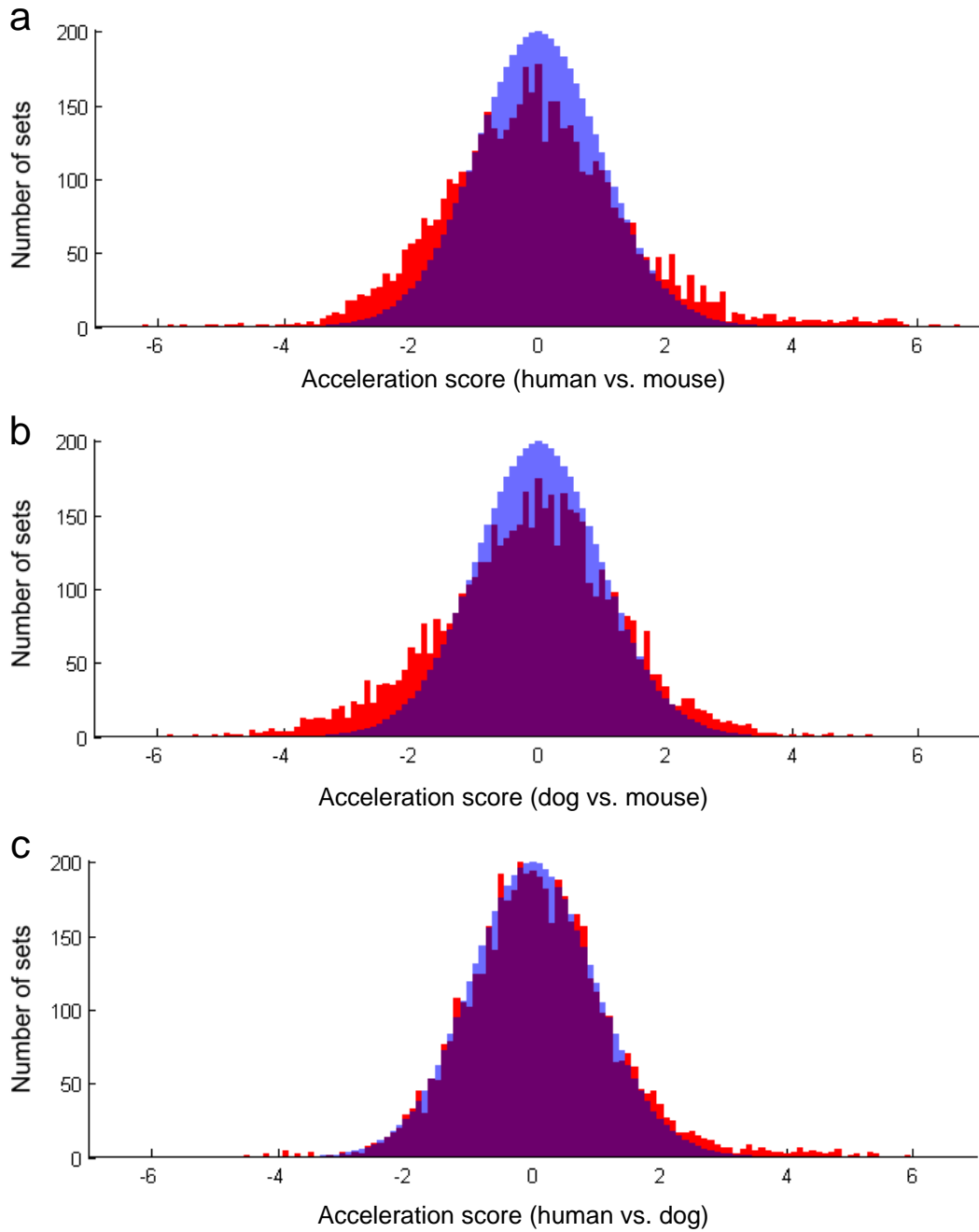


Figure 9. Histogram of observed z-scores for accelerated evolution in 4,950 gene sets (red), and the expected distribution based on 10,000 randomized trials (blue) for human relative to mouse (a), dog relative to mouse (b) and human relative to dog (c). Note that the comparison between human and dog fits the expected distribution significantly better than either comparison involving mouse, possibly due to murid-specific changes in sequence composition or evolutionary constraints.

observe the same trend for the 18 human genes that overlap with the genes studied here, but find at least as many genes with higher relative acceleration in the dog lineage (see Supplementary Notes). Heterotachy relative to mouse therefore does not appear to be a distinctive feature of the human lineage. It may reflect decelerated evolution in the rodent lineage or, perhaps, independent adaptive evolution in the human and dog lineages⁸².

A small number of gene sets show evidence of significantly accelerated evolution on the human lineage, relative to both mouse and dog (32 sets at $z \geq 5.0$ versus 0 sets expected by chance, $p < 10^{-4}$; Figure 10a). These sets fall into two categories: genes expressed exclusively in testis and (nuclear) genes encoding subunits of the mitochondrial electron transport chain (ETC) complexes. The former are believed to undergo rapid evolution as a consequence of sperm competition across a wide range of species⁸³⁻⁸⁵ and lineage-specific acceleration suggests that sexual selection may have been a particularly strong force in primate evolution. The selective forces acting on the latter set are less obvious. Because of the importance of mitochondrial ATP generation for sperm motility⁸⁶ and the potentially antagonistic co-evolution of these genes with maternally inherited mtDNA-encoded subunits⁸⁷, we propose that sexual selection may also be the primary force behind the rapid evolution of the primate ETC genes. Given the ubiquitous role of mitochondrial function, however, such sexual selection may have led to profound secondary effects on physiology⁸⁸.

We found no sets with comparably strong evidence for dog-specific accelerated evolution. There is, however, a small excess of sets with moderately high acceleration scores (19 sets at $z \geq 3.0$ versus 5 sets expected by chance, $p < 0.02$; Figure 10b). These sets, which are primarily related to metabolism, may contain promising candidates for follow-up studies of molecular adaptation in carnivores.

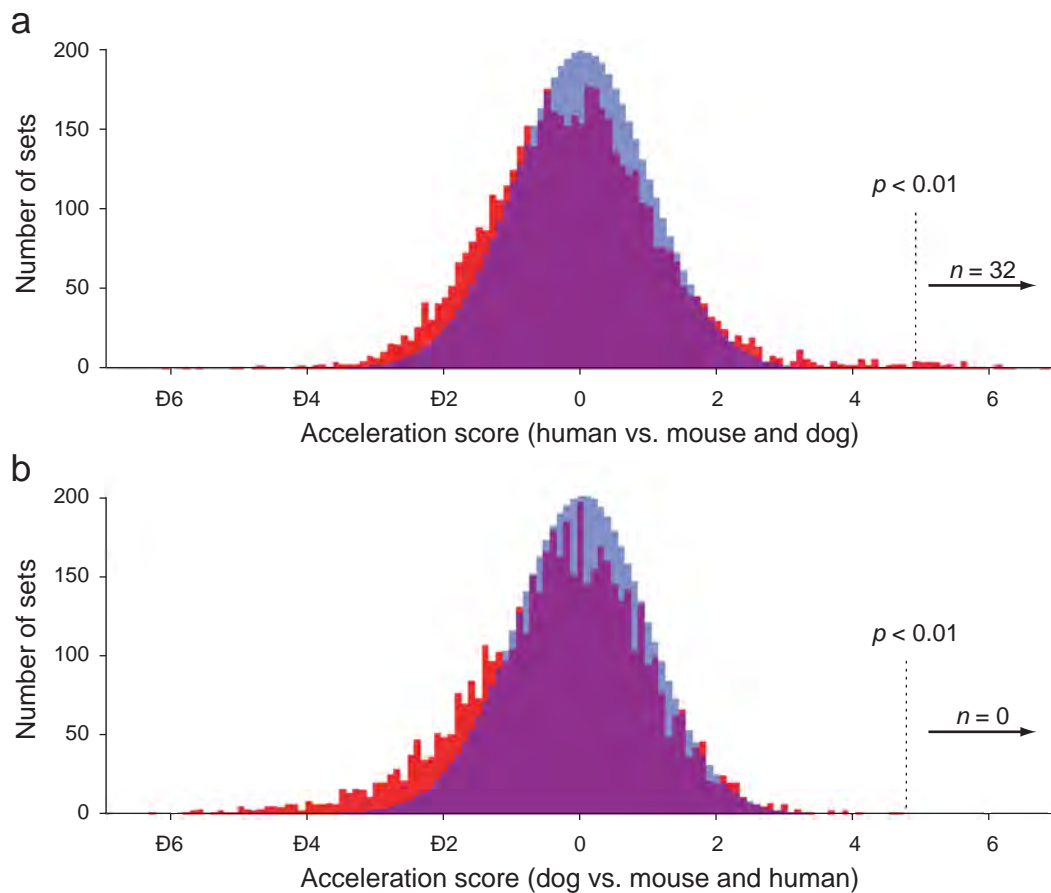


Figure 10. Genes sets showing accelerated evolution along the human and dog lineages. a, Distribution of acceleration scores along the human lineage relative to both mouse and dog, observed for 4,950 gene sets (red) and expected based on 10,000 randomized trials (blue). The dotted line shows the acceleration score for which the probability of observing even a single set by random chance (out of the 4,950 sets tested) is less than 1%. In fact, 32 sets show acceleration scores on the human lineage exceeding this threshold. b, The observed (red) and expected (blue) distribution of acceleration scores for dog, relative to both human and mouse.

Conclusion

Genome comparison is a powerful tool for discovery. It can reveal unknown – and even unsuspected – biological functions, by sifting the records of evolutionary experiments that occurred over 100 million years or over only 100 years. The dog genome sequence illustrates the range of information that can be gleaned from such studies.

Mammalian genome analysis is providing the first global picture of gene regulation in the human genome. Initial comparison with rodents revealed that ~5% of the human genome is under purifying selection and that the majority of this sequence is not protein-coding. The dog genome is now further clarifying this picture, as our data suggest that this ~5% represents functional elements common to all mammals. The distribution of these elements relative to genes is highly heterogeneous, with roughly half of the most highly conserved non-coding elements apparently devoted to regulating ~1% of human genes; these genes play key roles in development, and understanding the extraordinary regulatory clusters that surround them may reveal how cellular states are established and maintained. In recent papers^{32, 108}, the dog genome sequence has been used to greatly expand the catalog of mammalian regulatory motifs in promoters and 3'-untranslated regions. The dog genome sequence is also being used to substantially revise the human gene catalog. Despite these advances, it is clear that mammalian comparative genomics is still in its early stages. Progress will be dramatically accelerated by the availability of many additional mammalian genome sequences, initially with light coverage²⁸ but eventually with near-complete coverage.

Beyond its role in studies of mammalian evolution, the dog has a special role in genomic studies because of the unparalleled phenotypic diversity among closely related breeds. The dog is a testament to the power of breeding programs to select naturally occurring genetic variants with the ability to shape morphology, physiology and behavior. Genome comparison within and across breeds can reveal the genes that underlie such traits, informing basic research on development and neurobiology. It can also identify disease genes that were carried along in breeding programs; the potential benefits include insights into disease mechanism, as well as the possibility of clinical trials in affected dogs to accelerate new therapeutics to improve health in both species. The SNP map of the dog genome generated in a related analysis confirms that dog breeds show the long-range haplotype structure expected from intensive breeding. Moreover, analysis shows that the current collection of ~2.5 million SNPs should be sufficient to allow association studies of nearly any trait in any breed. Realizing the full power of dog genetics now only awaits the development of appropriate genotyping tools, such as multiplex 'SNP chips'¹⁰⁹; this is already underway.

For millennia, dogs have accompanied humans on their travels. It is only fitting that the dog should be a valued companion on our journeys of scientific discovery as well.

Methods

WGS sequencing and assembly. Approximately 31.5 million sequence reads were derived from both ends of inserts (paired end reads) from 4-, 10-, 40- and 200-kb clones, all prepared from primary blood lymphocyte DNA from a single female Boxer. The assembly was carried out using an interim version of ARACHNE2+ (www.broad.mit.edu/wga/). The particular individual was chosen for sequencing because it has the lowest heterozygosity rate among ~120 dogs tested at a limited set of loci; subsequent analysis showed that the genome-wide heterozygosity rate is not substantially different than other breeds⁹¹.

Genome alignment and comparison. Synteny maps were generated using standard methods²⁴ from pair-wise alignments of repeat masked assemblies using PatternHunter¹¹⁰ on CanFam2.0. All other comparative analyses were performed on BLASTZ/MULTIZ^{111, 112} genome-wide alignments obtained from the UCSC genome browser (genome.ucsc.edu) based on CanFam1.0. Known interspersed repeats were identified and dated using RepeatMasker and DateRepeats¹¹³. The numbers of orthologous nucleotides were counted directly from the alignments using human (hg17) as the reference sequence for all overlaps except the dog-mouse overlap, where pair-wise (CanFam1.0, mm5) alignments were used.

Divergence rate estimates. Orthologous ancestral repeats were excised from the genome alignment and realigned with the corresponding RepBase consensus using ClustalW. Nucleotide divergence rates were estimated from concatenated repeat alignments using baseml with the REV substitution model¹¹⁴. Orthologous coding regions were excised from the genome alignments using the annotated human CDS from Ensembl and the UCSC browser Known Genes track (Oct. 2004) as reference. K_A and K_S were estimated for each ortholog triplet using codeml with the F3x4 codon frequency model and no additional constraints.

Detection and clustering of sequence conservation. Pair-wise conservation scores and the fraction of orthologous sequences under purifying selection were estimated as in²⁴. The three-way conservation score S_{HMD} was defined as

$$S_{\text{HMD}} = (p - u) / \sqrt{(u(1 - u) / n)}$$

where n is the number of nucleotides aligned across all three genomes for each non-overlapping 50-bp window with more than 20 aligned bases, p is the fraction of nucleotides identical across all three genomes, and u is the mean identity of ancestral repeats within 500kb of the window. HCNEs were defined as windows with $S_{\text{HMD}} > 5.4$ that did not overlap a coding exon, as defined by the UCSC Known Genes track, and HCNE clusters were defined as all runs of overlapping

1Mb intervals (50kb step size) across the human genome with HCNE densities in the 90th percentile.

Gene set acceleration scores. Gene annotation was performed on CanFam1.0. A set of 13,816 human, mouse and dog orthologous genes were identified and compiled into 4950 gene sets containing genes related by functional annotations or microarray gene expression data. For each gene set S , the acceleration score $A(S)$ along a lineage is defined by (i) ranking all genes based on K_a/K_s within a lineage; (ii) calculating the rank-sum statistic for the set along each lineage (denoted $a_{\text{dog}}(S)$, $a_{\text{mouse}}(S)$, $a_{\text{human}}(S)$); (iii) calculating the rank-sum for the lineage minus the maximum rank-sum the other lineages (for example, $a_{\text{human}}(S) - \max(a_{\text{dog}}(S), a_{\text{mouse}}(S))$) and (iv) converting this rank-sum difference to a z-score by comparing it to the mean and standard deviation observed in 10,000 random sets of the same size. The expected number of sets at a given z-score threshold were estimated by repeating steps (i)-(iv) 10,000 times for groups of 4,950 randomly permuted gene sets.

References

1. Wayne, R. K. et al. Molecular systematics of the Canidae. *Syst Biol* 46, 622-53 (1997).
2. Vila, C. et al. Multiple and ancient origins of the domestic dog. *Science* 276, 1687-9 (1997).
3. Bardeleben, C., Moore, R. L. & Wayne, R. K. Isolation and molecular evolution of the selenocysteine tRNA (Cf TRSP) and RNase P RNA (Cf RPPH1) genes in the dog family, Canidae. *Mol Biol Evol* 22, 347-59 (2005).
4. Wayne, R. K. & Ostrander, E. A. Origin, genetic diversity, and genome structure of the domestic dog. *Bioessays* 21, 247-57 (1999).
5. American, K. C. *The Complete Dog Book* (eds. Crowley, J. & Adelman, B.) (Howell Book House, New York, NY, 1998).
6. Wayne, R. K. Limb morphology of domestic and wild canids: the influence of development on morphologic change. *J Morphol* 187, 301-19 (1986).
7. Ostrander, E. A., Galibert, F. & Patterson, D. F. Canine genetics comes of age. *Trends in Genetics* 16, 117-123 (2000).
8. Patterson, D. Companion animal medicine in the age of medical genetics. *J Vet Internal Med* 14, 1-9 (2000).
9. Sargan, D. R. IDID: inherited diseases in dogs: web-based information for canine inherited disease genetics. *Mamm Genome* 15, 503-6 (2004).
10. Chase, K. et al. Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. *Proc Natl Acad Sci U S A* 99, 9930-5. Epub 2002 Jul 11. (2002).
11. Breen, M. et al. Chromosome-specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes. *Genome Res* 11, 1784-95. (2001).
12. Breen, M., Bullerdiek, J. & Langford, C. F. The DAPI banded karyotype of the domestic dog (*Canis familiaris*) generated using chromosome-specific paint probes. *Chromosome Res* 7, 401-406 (1999).
13. Breen, M. et al. An integrated 4249 marker FISH/RH map of the canine genome. *BMC Genomics* 5, 65 (2004).
14. Hitte, C. et al. Opinion: Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nat Rev Genet* (2005).
15. Li, R. et al. Construction and characterization of an eightfold redundant dog genomic bacterial artificial chromosome library. *Genomics* 58, 9-17 (1999).
16. Kirkness, E. F. et al. The dog genome: survey sequencing and comparative analysis. *Science* 301, 1898-903. (2003).
17. Sutter, N. & Ostrander, E. Dog star rising: The canine genetic system. *Nat. Rev. Genet.* 5, 900-910 (2004).
18. Galibert, F., Andre, C. & Hitte, C. [Dog as a mammalian genetic model]. *Med Sci (Paris)* 20, 761-6 (2004).
19. Pollinger, J. P. et al. Selective sweep mapping of genes with large phenotypic effects. *Genome Research* (in press) (2005).
20. Sachidanandam, R. et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928-33 (2001).
21. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* 409, 860-921. (2001).
22. Venter, J. C. et al. The sequence of the human genome. *Science* 291, 1304-51. (2001).

23. Mikkelsen, T. et al. Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome. *Nature* 437, 69-87 (2005).
24. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-62. (2002).
25. Gibbs, R. A. et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428, 493-521 (2004).
26. Murphy, W. J. et al. Molecular phylogenetics and the origins of placental mammals. *Nature* 409, 614-8 (2001).
27. Thomas, J. W. et al. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424, 788-93 (2003).
28. Margulies, E. H. et al. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci U S A* 102, 4795-800 (2005).
29. Boffelli, D. et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299, 1391-4 (2003).
30. Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* 304, 1321-5 (2004).
31. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol* 3, e10 (2005).
32. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434, 338-45 (2005).
33. Dermitzakis, E. T. et al. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res* 14, 852-9 (2004).
34. Jaffe, D. B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13, 91-6 (2003).
35. H.U.G.O. (The Human Genome Organisation (HUGO), Bermuda, 1997).
36. Richterich, P. Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res* 8, 251-9 (1998).
37. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. & Eichler, E. E. Hotspots of mammalian chromosomal evolution. *Genome Biol* 5, R23 (2004).
38. Andelfinger, G. et al. Detailed four-way comparative mapping and gene order analysis of the canine *ctvm* locus reveals evolutionary chromosome rearrangements. *Genomics* 83, 1053-62 (2004).
39. Kirkness, E., et. al. *Genome Research* (submitted) (2005).
40. Mamedov, I. Z., Arzumanyan, E. S., Amosova, A. L., Lebedev, Y. B. & Sverdlov, E. D. Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach. *Nucleic Acids Res* 33, e16 (2005).
41. Lin, L. et al. The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* 98, 365-376 (1999).
42. Pele, M., Turet, L., Kessler, J. L., Blot, S. & Panthier, J. J. SINE exonic insertion in the *PTPLA* gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs. *Hum Mol Genet* 14, 1417-27 (2005).
43. Fondon, J. W., 3rd & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* 101, 18058-63 (2004).
44. Galtier, N. & Mouchiroud, D. Isochore evolution in mammals: a human-like ancestral structure. *Genetics* 150, 1577-84 (1998).

45. Belle, E. M., Duret, L., Galtier, N. & Eyre-Walker, A. The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *J Mol Evol* 58, 653-60 (2004).
46. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* 8, 1499-504 (1980).
47. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* 90, 11995-9 (1993).
48. Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S. & Sidow, A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res* 13, 813-20 (2003).
49. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A* 101, 13994-4001 (2004).
50. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci U S A* 90, 4087-91 (1993).
51. Gillooly, J. F., Allen, A. P., West, G. B. & Brown, J. H. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc Natl Acad Sci U S A* 102, 140-5 (2005).
52. Laird, C. D., McConaughy, B. L. & McCarthy, B. J. Rate of fixation of nucleotide substitutions in evolution. *Nature* 224, 149-54 (1969).
53. Li, W. H., Tanimura, M. & Sharp, P. M. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J Mol Evol* 25, 330-42 (1987).
54. Webber, C., Ponting, C. P. Hot spots of mutation and breakage in dog and human chromosomes. *Genome Research* (in press) (2005).
55. Hillier, L. W. et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695-716 (2004).
56. Marques-Bonet, T. & Navarro, A. Chromosomal rearrangements are associated with higher rates of molecular evolution in mammals. *Gene* 353, 147-54 (2005).
57. Miller, W., Makova, K. D., Nekrutenko, A. & Hardison, R. C. Comparative genomics. *Annu Rev Genomics Hum Genet* 5, 15-56 (2004).
58. Smith, N. G., Brandstrom, M. & Ellegren, H. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* 84, 806-13 (2004).
59. Woolfe, A. et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3, e7 (2005).
60. Ovcharenko, I. et al. Evolution and functional classification of vertebrate gene deserts. *Genome Res* 15, 137-45 (2005).
61. Walter, K., Abnizova, I., Elgar, G. & Gilks, W. R. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends Genet* 21, 436-40 (2005).
62. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* (2005).
63. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* 302, 413 (2003).
64. Kimura-Yoshida, C. et al. Characterization of the pufferfish *Otx2* cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development* 131, 57-71 (2004).
65. Uchikawa, M., Ishida, Y., Takemoto, T., Kamachi, Y. & Kondoh, H. Functional analysis of chicken *Sox2* enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev Cell* 4, 509-19 (2003).
66. de la Calle-Mustienes, E. et al. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res* 15, 1061-72 (2005).

67. Daly, M. J. Estimating the human gene count. *Cell* 109, 283-4 (2002).
68. Hogenesch, J. B. et al. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* 106, 413-5 (2001).
69. Emes, R. D., Goodstadt, L., Winter, E. E. & Ponting, C. P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum Mol Genet* 12, 701-9 (2003).
70. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* 25, 232-4 (2000).
71. Wolfe, K. H. & Li, W. H. Molecular evolution meets the genomics revolution. *Nat Genet* 33 Suppl, 255-65 (2003).
72. Bailey, J. A., Liu, G. & Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* 73, 823-34 (2003).
73. Hughes, A. L. The evolution of the type I interferon gene family in mammals. *J Mol Evol* 41, 539-48 (1995).
74. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* 18, 486 (2002).
75. Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc Natl Acad Sci U S A* 99, 16134-7 (2002).
76. Demetrius, L. Directionality theory and the evolution of body size. *Proc Biol Sci* 267, 2385-91 (2000).
77. Fay, J. C. & Wu, C. I. Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet* 4, 213-35 (2003).
78. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* 101, 4164-9 (2004).
79. Mootha, V. K. et al. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34, 267-73 (2003).
80. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proceedings of the National Academy of Sciences* (In Press) (2005).
81. Dorus, S. et al. Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell* 119, 1027-40 (2004).
82. Saetre, P. et al. From wild wolf to domestic dog: gene expression changes in the brain. *Brain Res Mol Brain Res* 126, 198-206 (2004).
83. Wyckoff, G. J., Wang, W. & Wu, C. I. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403, 304-9 (2000).
84. Birkhead, T. R. & Pizzari, T. Postcopulatory sexual selection. *Nat Rev Genet* 3, 262-73 (2002).
85. Dorus, S., Evans, P. D., Wyckoff, G. J., Choi, S. S. & Lahn, B. T. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. *Nat Genet* 36, 1326-9 (2004).
86. Ruiz-Pesini, E. et al. Correlation of sperm motility with mitochondrial enzymatic activities. *Clin Chem* 44, 1616-20 (1998).
87. Zeh, J. A. & Zeh, D. W. Maternal inheritance, sexual conflict and the maladapted male. *Trends Genet* 21, 281-6 (2005).
88. Grossman, L. I., Wildman, D. E., Schmidt, T. R. & Goodman, M. Accelerated evolution of the electron transport chain in anthropoid primates. *Trends Genet* 20, 578-85 (2004).
- 89-107. *Not referenced in this version of the text.*

108. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15-20 (2005).
109. Syvanen, A. C. Toward genome-wide SNP genotyping. *Nat Genet* 37 Suppl, S5-10 (2005).
110. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* 18, 440-5 (2002).
111. Schwartz, S. et al. Human-mouse alignments with BLASTZ. *Genome Res* 13, 103-7 (2003).
112. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14, 708-15 (2004).
113. Smit, A. F. A., & Green, P. in <http://ftp.genome.washington.edu/RM/RepeatMasker.html> (1999).
114. Yang, Z., Goldman, N. & Friday, A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol Biol Evol* 11, 316-24 (1994).
115. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res* 11, 1725-9 (2001).
116. Viterbi, A. J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Processing* 13, 260-269 (1967).
117. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263-5 (2005).
118. The International HapMap Project. *Nature* 426, 789-96 (2003).
119. Savolainen, P., Zhang, Y. P., Luo, J., Lundeberg, J. & Leitner, T. Genetic evidence for an East Asian origin of domestic dogs. *Science* 298, 1610-3. (2002).
120. Macdonald, D. W., Sillero-Zubiri, C. in *Biology and Conservation of Canids* (ed. Macdonald, D. W., and Sillero-Zubiri, C.) (Oxford University Press, Oxford, 2004).
121. Clamp, M., et al., Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* 2007. 104(49): p. 19428-33.

Supplementary notes: Generating a draft genome sequence

Data generation. We sequenced the genome of a female boxer, using a whole genome shotgun approach. A female dog was selected to obtain equal coverage of the X chromosome and the autosomes.

In order to facilitate the assembly process, we sought to identify a highly homozygous individual. Breed Clubs, Veterinary schools, and specific breeders were contacted via web sites, as well as by direct communiqué, to obtain potentially useful DNA samples. We estimated mean heterozygosity for each of 185 candidate dogs by resequencing 19,860 base pairs representing 60 unlinked loci [1]. A female boxer was selected based on the presence of only one heterozygous base pair within this sample. However, large scale data now shows that, after accounting for sample variance, this is within the expected rate of heterozygosity for the majority of dogs.

DNA was isolated from the chosen dog via peripheral blood samples. A BAC library was made by CHORI and 4 kb, 10 kb Fosmid libraries were made at the Broad Institute of Harvard and MIT. Both BAC and Fosmid clones can be ordered from the BACPAC Resource center at CHORI (<http://bacpac.chori.org/>). Sequence reads were generated from all four libraries. The BAC end reads were generated at Agencourt Biosciences. All other reads were generated at the Broad Institute of Harvard and MIT. A total of 35 million reads were attempted with 31.5 million passing reads used as input to the assembly process, resulting in ~7.5x coverage of the 2.4 Gb genome.

Genome assemblies. The most recent genome assembly (CanFam2.0) spans a total distance of 2.41 Gb, consisting of 2.38 Gb of nucleotide sequence with the remaining 1% in captured gaps. The assembly was carried out using an interim version of ARACHNE2+ [2].

Two consecutive assemblies were performed, the initial assembly CanFam1.0 was released in July 2004 and the CanFam2.0 assembly was released in May 2005. CanFam1.0 was used for the analyses described in this paper, unless otherwise specified.

Both assemblies were anchored to the canine chromosomes using data from RH and cytogenetic maps [3-5]. After comparisons to the maps, further algorithmic improvements resulted in an improved CanFam1.0 assembly. For this improved version only four discrepancies suggesting errors in the assembly were identified between CanFam1.0 and the RH map. These discrepancies were resolved using FISH [3-5] prior to the generation of CanFam2.0. Twenty-three BAC clones from four groups, each comprising four-seven clones that mapped to four different “problematic” regions of CFA 9, 11 and 16 were mapped. DNA for each of the 23 clones was prepared by routine alkaline lysis and then labeled for FISH with one of five

spectrally resolvable fluors according to previous methods [4]. To exclude any structural cytogenetic anomalies unique to the sequenced boxer, all groups were hybridized to chromosome preparations and interphase nuclei prepared from peripheral lymphocytes of the sequenced boxer and two other (mixed breed) dogs. All FISH techniques were as previously described [4].

Images were acquired using a semi-automated multicolor FISH workstation driven by SmartCapture2.3 software (Digital Scientific, Cambridge, UK). The cytogenetic location and order of all probes within each group were determined from metaphase and interphase analysis of no fewer than 20 cells and a consensus order derived.

Assessing base quality. We assessed base accuracy by comparing the CanFam2.0 assembly to finished sequence. Nine BACs, where both ends had passing reads, were randomly selected and finished using the standard finishing pipeline[6]. Only sequence from homozygous regions of the boxer genome, comprising in total 4 BACs (0.76 Mb), from the total of 9 BACs (1.65 Mb) were compared to the assembly. We did not use the heterozygous regions because the finished sequence represented only one of the two haplotypes seen intermingled in the assembly.

Explaining the high sequence continuity. The CanFam2.0 assembly has dramatically greater sequence continuity (180 kb) than the WGS assembly of the mouse genome (25 kb) obtained several years ago. Several factors contribute to the higher connectivity of the dog assembly:

1. Increased redundancy: We examined the influence of coverage by approximating the continuity of a 6.5X dog assembly by removing reads from the 7.5X assembly, understanding that somewhat lesser continuity would be expected for a de novo assembly having less reads. The 6.5X assembly had the same N50 of supercontig size (45 Mb), but lower N50 of contig size (122 kb) than the full 7.5X assembly.

2. Less duplication: We examined dog and mouse reads for the fraction of all observed k-mers that are represented more than $n \times c$ times in the sequence reads, where c is the assembly coverage level. Based on presence of overrepresented k-mers, the mouse genome appears to have roughly twice as much recently duplication sequence as dog. The contribution to these fractions arising from sampling variability of the reads (assuming random DNA) is negligible relative to the observed fractions ($\leq 0.005\%$).

3. Improved data quality: We examined the data quality for the mouse and dog data sets. The read length increased from 560 base pairs to 700 base pairs, the read pairing rate from 85% to 96% and the library spread (the standard deviation of insert length as a fraction of insert length for 4 kb plasmid libraries) decreased from 20% to 9%.

4. Algorithmic improvement: A comparison was made between the improved version of ARACHNE used for dog and the old version used to assemble the published mouse genome. Assembling the dog with the previous software decreases contig size from 180 kb to 61 kb, while assembling the mouse with the new version increased it from 25 kb to 35 kb.

Supplementary Notes: Genome landscape and evolution

Conserved synteny and large-scale rearrangements. We generated pair-wise synteny maps between dog (CanFam2.0) and human (hg17), mouse (mm5) and rat (rn3) using standard methods [7]. First, we aligned repeat masked assemblies using the PatternHunter program [8]. Next, we identified runs of alignments with well-defined order and orientation. These clusters of alignments were then merged hierarchically to create syntenic segments larger than a specified minimum size. Segments retain the dominant orientation of the alignments on which they are based. Finally, we defined syntenic blocks by merging adjacent syntenic segments that were contiguous in both genomes (orientation was ignored). For this reason, a local rearrangement of the segments within a syntenic block does not change the block boundaries. Different minimum segment sizes were used and yield different numbers of syntenic blocks and segments.

To assign breakpoints to particular lineages we created a four-way synteny map from the individual pair-wise maps as follows. We compiled a list of the locations in dog of all the pair-wise discontinuities. We then generated finer pair-wise maps to define orthologous segments in other genomes for each dog interval in this list (Figure 2). Breakpoints were assigned to particular branches of the evolutionary tree based on coincidence in the fourway synteny map. To simplify the nomenclature we called all breakpoints falling within syntenic segments “all breakpoints”, those falling within syntenic blocks were called “interchromosomal breakpoints” and the breakpoints not present among the interchromosomal breaks but present in the all breaks group were called “intrachromosomal breaks”. It is important to note that events identified as “coincident “ in two lineages by this approach need not be any closer together in the genomic sequence than the minimum size allowed in the syntenic maps.

Similar results are obtained when monodelphis is used as outgroup to dog, human, mouse and rat. We created pair-wise synteny maps at resolutions of 1,000, 500, 300 and 100 kb using a preliminary intermediate assembly of *Monodelphis Domesticata* as the reference genome. Because the scaffolds of this *Monodelphis* assembly are not yet anchored to chromosomes and we have not yet fully characterized the completeness and coverage properties of this assembly, we sought merely to confirm the overall trends in the assignment of breakpoints to lineages seen in the DHMR four-way analysis. In particular, we confirm that the human lineage has experienced a greater number of intrachromosomal breaks within syntenic blocks than the dog lineage. Furthermore, this tendency grows stronger for smaller segment size cutoffs.

Genomic insertion and deletion. The analysis of shared and lineage-specific sequence content across dog, human and mouse are based on BLASTZ whole-genome alignments

(<http://genome.ucsc.edu>) and RepeatMasker output. Nucleotides in the three genomes were first classified as belonging to lineage-specific interspersed repeats or 'other' sequence. The dog (canFam1), human (hg17) and mouse (mm5) assemblies were masked for known interspersed repeats using RepeatMasker [08/03/2002 library: RepBase Update 8.12, RM database version 20040306]. The final CanFam2.0 assembly was also masked with an updated repeat library [RepeatMasker 3.0.8 library: RepBase Update 9.04, RM database version 20040702]. The repeats were then subdivided into ancestral and lineage-specific instances based on the DateRepeats script and the presences or absence of orthologous sequences. Approximately 1-5% of repeat sequences classified as lineage-specific by DateRepeats in one of the genomes were aligned to at least one of the other species, although some of these alignments may be spurious. As noted previously [7], the activity of repeats in all classes except DNA transposons appears to have been distinctly higher in mouse relative to human. The average rates of SINE and LINE insertions have been similar in the human and dog lineages, but have contributed more sequence overall to the human genome, and the activity of DNA transposons and endogenous retrovirus have been significantly higher in human. The total amount of lineage-specific repeats may be a slight underestimate for each species as lineage-specific repeats inserted close after the mammalian radiation may be conservatively classified as ancestral.

The numbers of shared and lineage-specific nucleotides, as presented in Figure 1, were then counted directly from the BLASTZ alignments. The intersection between the human and the other genomes were computed from alignments using human as the reference sequence. The intersection between dog and mouse were computed from alignments using dog as the reference sequence. We note that asymmetries in the genome-wide alignment procedure leads to slight differences (on the order of ~50 Mb) in the estimated size of human-dog overlap depending on whether human or dog were used as the reference sequence. We also note that the size of the genome intersections depends on whether or not bases overlapping small indels within an otherwise orthologous sequence are counted as aligned. The numbers described in the main text count only bases present in all three genomes as aligning between all three species. If small indels are included as aligned, the number of lineage-specific bases fall by approximately 60 Mb per lineage, and the orthologous intersections increase correspondingly. This does not impact any conclusion presented in the main text.

We can obtain a rough estimate of lineage-specific deletion rates by assuming that ancestral nucleotides (other than the common, presumed functional 150 Mb) are deleted continuously and independently. Given the amount of extant ancestral sequence (X), total ancestral genome size (T), and divergence time (t), we have

$$X = T e^{-kt}$$

where k is the rate of deletion. Assuming $T = 2,800 \text{ Mb} - 150 \text{ Mb}$ and $t = 75 \text{ Myr}$ [7], we get $k = 0.0024 \text{ Myr}^{-1}$ for human and 0.0078 Myr^{-1} for mouse. And assuming $t = 90 \text{ Myr}$ (based on divergence rates the length of the BEA to rodent-primate branch is $\sim 10\%$ of the subsequent primate and rodent branches, i.e. 7.5 Myr which must be multiplied by 2 given the unrooted tree), we get $k = 0.0031 \text{ Myr}^{-1}$ for dog, suggesting that the relative rate of deletion between the three lineages are roughly proportional to the relative rates of substitutions. Similar estimates can be made for insertions, but transposon-derived insertions are unlikely to approximate a continuous process [9].

Active SINE family: RepeatMasker (3.0.8, RepBase Update 9.04, RM database version 20040702) identification of the dog-specific SINE element SINEC_Cf found 377,453 copies in the canFam2.0 assembly. To study only the most recent copies of SINEC_Cf, we restricted our analysis to those copies which matched the SINEC_Cf consensus from end to end with at most 10 mismatches and 20 indels, obtaining approximately 87,000 elements. Comparison of overlapping sequence reads within the assembly identified 8% of these as polymorphic within the sequenced boxer.

Rate of nucleotide divergence. Regional neutral divergence rates for the dog, mouse and human branches were estimated from alignments of ancestral repeat (AR) sequences (which are presumed to be largely non-functional), as classified by RepeatMasker and RepeatDater (<http://repeatmasker.org>), using the same approach as described previously [10]. The ancient L2 and L3 subclasses were excluded as inspection revealed a high frequency of apparent alignment artifacts, but the relative rates reported in the main text are robust to inclusion of these sequences.

The orthologous sequences for each AR fragment were extracted from the genome-wide alignment and realigned to the corresponding repeats consensus using ClustalW [11]. Branch-specific divergence rates were then estimated on a tree rooted by the AR consensus using PAML (<http://abacus.gene.ucl.ac.uk/software/paml.html>) with the REV model for all aligned sequences within each 1 Mb interval. Only intervals with more than 2,000 nucleotides aligned across all three species were used for estimates reported in the main text.

Male mutation bias. Given that autosomes spend equal time, and the X chromosome twice as long, in the female germ-line as in the male germ-line, the autosomal divergence rate (A), the X chromosome divergence rate (X), the male mutation rate (μ_m) and the female mutation rate (μ_f) are related by

$$A = 1/2 \mu_m + 1/2 \mu_f$$

$$X = 1/3 \mu_m + 2/3 \mu_f$$

which leads to

$$\mu_m / \mu_f = (4A - 3X)/(3X - 2A)$$

Taking A and X as the median lineage-specific divergence rates for each lineage from ancestral repeats in 1 Mb windows gives 0.193 and 0.162 for the dog lineage; 0.373 and 0.138 for the mouse lineage; and 0.138 and 0.108 for the human lineage; yielding $\mu = \mu_m / \mu_f = 2.8, 2.8$ and 4.6 for the dog, mouse and human lineages, respectively. Using the mean divergence rate, or changing the window size or the threshold for number of aligned bases per window, leads to correlated estimates in the 2.5-3.5 range on the dog and mouse lineages and in the 4.5-5.7 range on the human lineage.

Supplementary Notes: Proportion of genome under purifying selection

Low turnover of ancestral conserved elements. We estimated the fraction of orthologous sequence under selection from the excess of conservation observed in all sequence, relative to that seen in ancestral repeats only. The initial description and variations to this approach can be found in references [7, 12, 13]. Briefly, the human-dog conservation score for a 50 bp window (S_{HD}) is defined as

$$S_{HD} = (p - u) / (u(1 - u) / n)$$

where n is the number of aligned nucleotides between human and dog within the window, p is the fraction of identical nucleotides and u is the fraction of identical, aligned nucleotides in ancestral repeats in a 1 Mb interval centered on the window. S_{HD} was computed for all 50 bp windows with $n > 20$ (giving the distribution S_{genome}), and separately for the subset of such windows that completely overlap an ancestral repeat sequence (giving the distribution S_{ar}). The genome-wide distribution was then decomposed into selected ($S_{selected}$) and neutral ($S_{neutral}$) components defined by

$$S_{neutral} = p_0 S_{ar}$$

$$S_{selected} = S_{genome} - S_{neutral}$$

where the scaling coefficient p_0 is conservatively taken to be the minimum ratio between $S_{selected}(S_{HD})$ and $S_{ar}(S_{HD})$ for all values of S_{HD} . The probability that a particular window is under selection, given its conservation score

$$P_{selection}(S_{HD}) = 1 - p_0 S_{neutral}(S_{HMD}) / S_{genome}(S_{HMD}).$$

We first estimated the fraction of all windows under selection between human and dog. When all 50 bp windows for which human and dog align at ≥ 20 bp are selected, a total of 32.5 million windows are covered and $p_0 = 10.2\%$ and corresponding to 5.3% of all 50 bp windows along the human genome. When all windows where human, dog and mouse align at ≥ 20 bp are selected, 19.9 million windows are covered and $p_0 = 16\%$, corresponding to 5.2% of the human genome. For the remaining 12.6 million windows where mouse aligns over < 20 bp (the results are insensitive to this threshold, data not shown), $p_0 = 1\%$, corresponding to 0.1% of the human genome. This suggests that the vast majority of selected sequence between human and dog has also been retained in mouse.

We also devised a second approach by computing a combined human-mouse-dog conservation score S_{HMD} for each window along the human genome with orthologous sequence in both mouse and dog. S_{HMD} is defined as S_{HD} , but with n as the number of nucleotides aligned across all three genomes within the window, and p and the proportion of aligned nucleotides

identical across all three genomes. Repeating that analysis above on the 19.9 million eligible windows yields $p_0 = 18\%$, corresponding to 5.7% of the human genome. In principle, this estimate might not be directly comparable to the pairwise estimates from dog and mouse because the combined score may have increased power to detect weaker selection. However, the similarity of the combined and pair-wise estimates, taken together with the retention bias described above, strongly suggests that ~5% conserved bases between human-dog and between human-mouse are largely the same.

Distribution of highly conserved elements across the genome. We defined highly conserved elements as 50 bp windows on the human genome with more than 20 bp aligned sequence to both mouse and dog, and for which $P_{\text{selection}}(S_{\text{HMD}}) \geq 95\%$. We then computed the density of such windows across sliding 1 Mb windows (step size = 50 kb).

A cluster of highly conserved elements were defined as a run of one or more overlapping 1Mb windows with densities in the 90th percentile, a criteria chosen to identify all major peaks evident by visual analysis (Figure 6).

Potential regulatory targets were identified through manual inspection of functional annotations and related literature. Incomplete functional annotations prevent a fully automated significance test of the developmental gene enrichment in these clusters (the candidate genes were in several cases picked based on literature references and annotations of mouse orthologs that were reflected in the human databases). However, repeating the manual analysis of 100 randomly selected regions of the genome with a similar size distribution yielded putative developmental genes in only 26% of cases, compared with 92% of the highly conserved clusters ($p < 10^{-31}$ by Fisher's exact test).

Supplementary Notes: Genes

Gene predictions in dog and human. A dog gene set was produced based largely around the results of the Ensembl genebuild pipeline (CanFam1.0). The Ensembl set was augmented by additional, evidence based, genes from the Broad annotation pipeline and Goodstadt and Ponting's orthology pipeline [15]. Using this dog gene set and the Ensembl gene set for human (hg17), we used synteny relationships to establish clear 1-1 orthology relationship for ~14,500 genes. All other genes were further studied with a variety of tools as well as visual examination to confirm the presence or absence of genes in both human and dog. The exact methodology is described in [14].

Gene duplications. In order to distinguish genes duplicated in the dog or human lineage, since their common ancestor, we reconstructed the phylogeny of all Ensembl transcript sequences using the synonymous substitution rate KS as a distance metric (see [15]). This allowed predictions of gene orthology and paralogy, and of conserved synteny (gene order), whilst highlighting likely processed pseudogenes as predictions with multiple frameshift disruptions (as indicated by intron sizes less than, or equal to, 10bp) or else those in non-syntenic locations that also contain a single disruption or are single exon predictions.

Evolution of orthologous genes across three species. We compiled alignments of human-mouse-dog orthologs by extracting the orthologous sequences of human coding regions, as annotated by Ensembl and the UCSC “known genes” track (both downloaded December 30th 2004), from mouse and dog via MultiZ alignments of the human (hg17), mouse (mm5) and dog (canFam1) genomes (genome.ucsc.edu) after filtering out nonsyntenic alignment blocks. We removed all alignments with less than 80% coverage of the annotated coding regions or with one or more frame-shifting indels. If more than one transcript variant of a particular gene remained, we removed all but the longest such alignment, yielding 13,816 unique coding regions. Lineage-specific Ka and Ks were then estimated for each ortholog triplet on an unrooted tree using codeml from PAML with the F3x4 codon frequency model.

Gene sets. We compiled 4950 gene sets containing genes related by functional annotations or microarray gene expression data. For functional annotations, we used gene ontology (GO) to generate gene sets for each GO term based on annotations from Ensembl. We also used sets of genes involved in known metabolic and regulatory pathways (www.broad.mit.edu/gsea/molsigdb/molsigdb_index.html). For gene expression data, we used the tissue expression compendium containing gene expression profiles in 75 human tissues generated by Novartis [16]. From the compendium, we generated 75 tissue-specific gene sets containing

genes that were specifically expressed in certain tissues. We also extracted sets of genes that share similar expression profiles across this compendium to each of 2271 putative transcriptional regulators from TRANSFAC (www.gene-regulation.com), using Pearson correlation as the similarity measure. Finally, we included 1445 gene sets curated by [17]. The complete collection of gene sets used is available upon request.

Acceleration scores for gene sets. We adapted the Gene Set Enrichment Analysis approach (www.broad.mit.edu/gsea/) to search for sets enriched for genes with elevated Ka/Ks on one of the three branches, relative to the others. First, we ranked all considered orthologs by Ka/Ks on each branch. Second, for each set, we computed the rank sum of the constituent genes, r_H , r_M and r_D , for the human, mouse and dog branches, respectively. Figure 8 shows that these rank sums are highly correlated. Third, we computed the difference between each pair of rank sums for pair-wise comparisons (e.g. $r_H - r_D$ to identify sets accelerated in human relative to dog), and the difference between one rank sum and the maximum of the remaining two for three-way comparisons (e.g. $r_H - \max[r_M, r_D]$ to identify sets accelerated in human relative to both dog and mouse). Fourth, we converted the rank sum difference into a z-score using the mean and standard deviation observed from 10,000 random gene sets of the same size as the considered gene set.

In order to assess the significance of the observed acceleration scores we repeated the above procedure 10,000 times for the same collection of gene sets by randomly permuting the gene names (while keeping the three lineage-specific Ka/Ks ratios for each gene together and preserving the structure of the gene set intersections). For a given acceleration score threshold we then estimated the significance (p-value) of the observed number of gene sets above that threshold as the fraction of random trials yielding equal or more sets at the same threshold. We also estimated a conservative p-value for each individual gene set as the fraction of random trials yielding one or more gene sets at or above its acceleration score (the $p < 0.01$ threshold is shown in Figure 10). Sets related to testis-specific expression and the electron transport chain are significantly accelerated in humans according to both p-value definitions.

Comparisons of the observed distributions of pair-wise acceleration scores and the expected distributions are shown in Figure 9. There is a notably higher variance in the two comparisons involving mouse, whereas the human-dog distribution is highly symmetrical. This is reflected in the slight skew towards negative acceleration scores in Figure 10, and may signal more similar selective pressures or mutation biases on the human and dog lineages.

Acceleration in brain-related genes relative to rodents. Our database of gene sets contains multiple sets of genes expressed uniquely or non-uniquely in fetal and adult brains. Most showed weak positive acceleration scores for human or dog relative to mouse, but none were

significant. We therefore focused specifically on the 24 genes reported to show significantly faster evolution in primates relative to rodents in Table 1 of [18]. 18 of the genes were contained in our dataset (*MCPH1*, *CASP3*, *OPRM1*, *NRCAM*, *SHH*, *PYNX1*, *DRD2*, *GRIK4*, *CHRM5*, *CHRNA5*, *NTRK3*, *GRIN2A*, *PAFAH1B*, *LHX1*, *AANAT*, *ADCYAP1*, *TTRAP*, *CSPG3*). The median Ka/Ks estimated for these genes were 0.123, 0.0774 and 0.080, for the human, mouse and dog branches, respectively. This is 25% acceleration over the expected ratio for all orthologs on the human branch relative to the mouse branch, and 30% acceleration relative to the dog branch.

This enrichment for non-synonymous substitutions is somewhat lower than reported by [18], which is consistent with the expectation that positive selection on these genes increased during the more recent portion of branch leading to humans. Also consistent with [18], genes showing comparable acceleration on the mouse branch relative to the human branch appear to be relatively rare. However, it is simple to identify at least 18 genes in our brain-related gene sets showing strong acceleration on the dog branch (*MEIS2*, *DLG3*, *LHX2*, *NRXN1*, *NRXN2*, *NEUROD2*, *SH3GLB1*, *NAV2*, *NAV3*, *RTN1*, *LDB2*, *CDH10*, *DBN1*, *CSPG3*, *FOXP2*, *OLIG1*, *RBM9*, *INA*). The median Ka/Ks estimated for these genes were 0.0306, 0.0314 and 0.0546 for the human, mouse and dog branches, respectively. This is 110% acceleration over the expected ratio for all orthologs on the dog branch relative to the human branch, and 61% relative to the mouse branch.

Several of these genes are thought to be critical for nervous system development, suggesting that acceleration in the non-synonymous substitution rate of such genes is not unique to the lineage leading to humans.

References for Supplementary Notes

1. Parker, H.G., et al., Genetic structure of the purebred domestic dog. *Science*, 2004. 304(5674): p. 1160-4.
2. Jaffe, D.B., et al., Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res*, 2003. 13(1): p. 91-6.
3. Breen, M., et al., Chromosome-specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes. *Genome Res*, 2001. 11(10): p. 1784-95.
4. Breen, M., et al., An integrated 4249 marker FISH/RH map of the canine genome. *BMC Genomics*, 2004. 5(1): p. 65.
5. Hitte, C., et al., Opinion: Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nat Rev Genet*, 2005.
6. Consortium, I.H.G.S., Finishing the euchromatic sequence of the human genome. *Nature*, 2004. 431(7011): p. 931-45.
7. Waterston, R.H., et al., Initial sequencing and comparative analysis of the mouse genome. *Nature*, 2002. 420(6915): p. 520-62.
8. Ma, B., J. Tromp, and M. Li, PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 2002. 18(3): p. 440-5.
9. Liu, G., et al., Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res*, 2003. 13(3): p. 358-68.
10. Kirkness, E.F., et al., The dog genome: survey sequencing and comparative analysis. *Science*, 2003. 301(5641): p. 1898-903.
11. Thompson, J.D., et al., The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*, 1997. 25(24): p. 4876-82.
12. Chiaromonte, F., et al., The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harb Symp Quant Biol*, 2003. 68: p. 245-54.
13. Roskin, K.M., M. Diekhans, and D. Haussler, Score functions for determining regional conservation in two-species local alignments. *J Comput Biol*, 2004. 11(2-3): p. 395-411.
14. Clamp, M., et al., Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* 2007. 104(49): p. 19428-33.
15. Goodstadt, L. and C.P. Ponting, Analysis of gene orthology in the dog. *PLoS Comput Biol*, 2006. 2(9): e133
16. Su, A.I., et al., A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 2004. 101(16): p. 6062-7.
17. Segal, E., et al., A module map showing conditional activity of expression modules in cancer. *Nat Genet*, 2004. 36(10): p. 1090-8.
18. Dorus, S., et al., Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell*, 2004. 119(7): p. 1027-40.

Chapter 4: The opossum genome

In this chapter, we describe the first comprehensive comparative analysis of placental and marsupial genome sequences.

This work was first published as

Mikkelsen, T. S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167-177 (2007).

The full publication is attached as Appendix 3. Supplementary notes can be found at the end of the chapter. Supplementary data is available online from <http://www.nature.com/nature>

The text in this chapter was written with significant contributions from members of the analysis consortium.

We report a high-quality draft of the genome sequence of the grey, short-tailed opossum (*Monodelphis domestica*). As the first metatherian (‘marsupial’) species to be sequenced, the opossum provides a unique perspective on the organization and evolution of mammalian genomes. Distinctive features of the opossum chromosomes provide support for recent theories about genome evolution and function, including a strong influence of biased gene conversion on nucleotide sequence composition and a relationship between chromosomal characteristics and X inactivation. Comparison of opossum and eutherian genomes also reveals a sharp difference in evolutionary innovation between protein-coding and non-coding functional elements. True innovation in protein-coding genes appears to be relatively rare, with lineage-specific differences being largely due to diversification and rapid turnover in gene families involved in environmental interactions. By contrast, about 20% of eutherian conserved non-coding elements (CNEs) are recent inventions that postdate the divergence of Eutheria and Metatheria. A substantial proportion of these eutherian-specific CNEs arose from sequence inserted by transposable elements, pointing to transposons as a major creative force in the evolution of mammalian gene regulation.

Metatherians (‘marsupials’) comprise one of the three major groups of modern mammals and represent the closest outgroup to the eutherian (‘placental’) mammals (Figure 1). Metatherians and eutherians diverged ~180 million years ago (Mya), long before the radiation of the extant eutherian clades ~100 Mya^{1,2}. Although the metatherian lineage originally radiated from North America, only one extant species can be found there (the Virginia opossum), while all other species are found in South America (including more than 65 species of opossums and shrew opossums) and Australasia (~200 species, including possums, kangaroos, koalas and many small insectivores and carnivores)³.

All sequenced mammalian genomes to date have come from eutherian species. Although metatherians and eutherians (together, ‘therians’) share many ancient mammalian characteristics, they have each evolved distinctive morphological and physiological traits. Metatherians are particularly noted for the birth of young at a very early stage of development, followed by a lengthy and complex lactational period. Genomic analysis will help reveal the genetic innovations that underlie the distinctive traits of each lineage⁴⁻⁶.

Equally important, metatherian genomes can shed light on the human genome. Comparative analysis of eutherians has greatly improved our understanding of the architecture and functional organization of mammalian genomes⁷⁻¹⁰. Identification of sequence elements under purifying selection, based on cross-species sequence conservation, has led to increasingly refined inventories

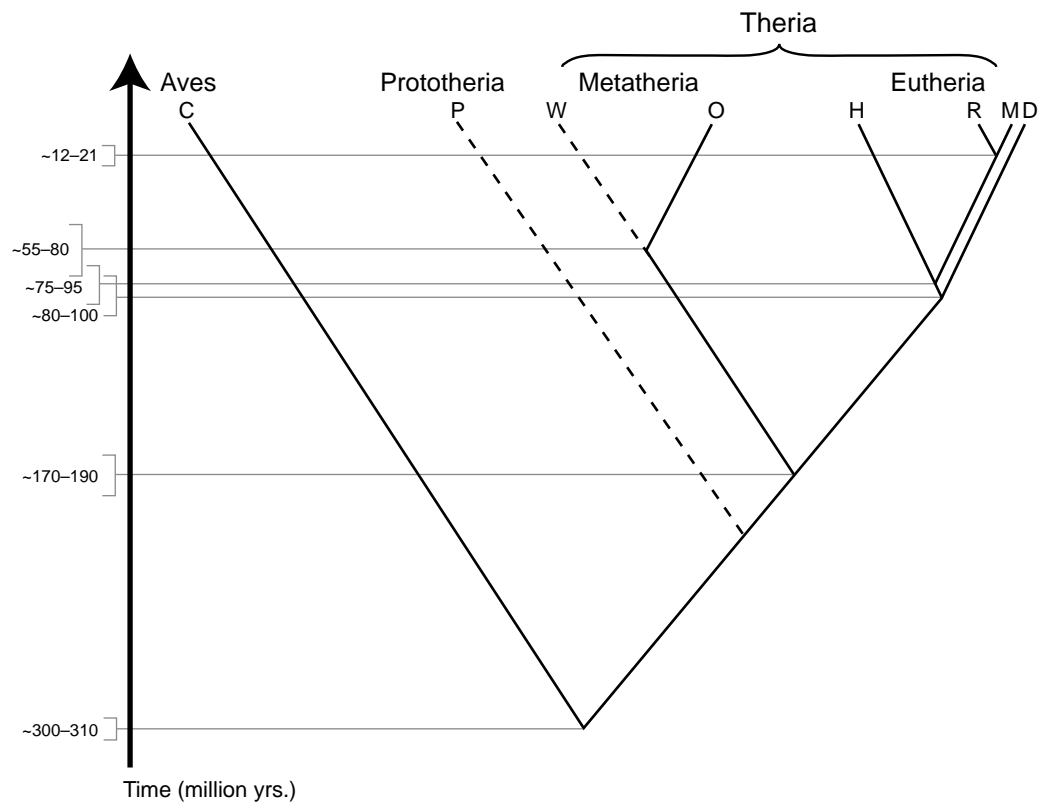


Figure 1. Simplified phylogeny of amniotes. Infraclass Eutheria is represented by human (H), mouse, (M), rat (R) and dog (D). Infraclass Metatheria is divided into two extant lineages: the Australasian and the American, represented by wallaby (W) and opossum (O), respectively. Infraclass Prototheria is represented by platypus (P). Aves is represented by chicken (C). The two dashed lines indicate major lineages that are not yet represented by complete genome sequences.

of protein-coding genes^{11,12}, proximal and distal regulatory elements^{13,14} and putative RNA genes¹⁵. Yet, we still know relatively little about the evolutionary dynamics of these and other functional elements. How stable is the complement of protein-coding genes? How rapidly do regulatory sequences appear and disappear? From what substrate do they evolve?

Comparison of the human genome to genomes from distant outgroups such as birds (~310 Mya) or fish (~450 Mya) has provided valuable information. When similarity between sequences from such distantly-related genomes can be detected, it surely signals functional importance. But the high specificity¹⁶ is offset by dramatically reduced sensitivity^{10,17,18}. Simulations have shown that the feasibility of aligning orthologous genomic sequences declines rapidly once their mean genetic distance exceeds one substitution per site¹⁹. The genome of chicken, the most closely related non-mammalian amniote genome available, is separated from the human genome by approximately 1.7 substitutions per site in orthologous, neutrally evolving sequences²⁰. Even moderately constrained functional elements may therefore be difficult to detect. By contrast, metatherian mammals are well positioned for addressing this issue; because unconstrained regions of their genomes are separated from human by only ~1 substitution per site (see below), most orthologous, constrained sequence should be readily aligned.

Here, we report the first high-quality draft of a metatherian genome sequence, which was derived from a female grey, short-tailed opossum, *Monodelphis domestica*. The species was chosen chiefly on the availability and utility of the organism for research purposes. *M. domestica*, is a small rapidly-breeding South American species that has been raised in pedigreed colonies for more than 25 years and developed as one of only two laboratory bred metatherians^{21,22}. *M. domestica* is being actively used as a model system for investigations in mechanisms of imprinting²³⁻²⁵, immunogenetics²⁶⁻²⁸, neurobiology, neoplasia, and developmental biology (reviewed in⁶). For example, newborn opossums are remarkable in that they can heal complete transections of the spinal cord²⁹. Elucidation of the molecular mechanisms underlying this ability promise important insights relevant to regenerative medicine for spinal cord or peripheral nerve injuries. Other than human, *M. domestica* is also the only mammal known in which ultraviolet radiation is a complete carcinogen for malignant melanoma³⁰, which has led to its establishment as a unique neoplasia model. All of these investigations will directly benefit from the development of genomic resources for this species.

Below, we describe the generation of the draft sequence of the opossum genome, analyze its large-scale characteristics, and compare it to previously sequenced amniote genomes. Our key findings include:

- The distinctive features of the opossum genome provide an informative test of current models of genome evolution and supports the hypothesis that biased gene conversion plays a key role in determining overall nucleotide composition.
- The evolution of random X inactivation in eutherians correlates with acquisition of *XIST*, elevation in LINE/L1 density and suppression of large-scale rearrangements.
- The opossum genome appears to contain 18,000-20,000 protein-coding genes, the vast majority of which have eutherian orthologs. Lineage-specific genes largely originate from expansion and rapid turnover in gene families involved in immunity, sensory perception and detoxification.
- Identification of orthologs of highly divergent immune genes and a novel T-cell receptor isotype challenge previous claims that metatherians possess a ‘primitive’ immune system.
- Of the non-coding sequences conserved among eutherians, ~20% appear to have evolved after the divergence from metatherians. Of protein-coding sequences conserved among eutherians, only ~1% appear to be absent in opossum.
- At least 16% of eutherian-specific conserved non-coding elements are clearly derived from transposons, implicating these elements as an important creative force in mammalian evolution.

Extensions to these findings, as well as additional topics, are reported in a series of companion papers³¹⁻⁴¹.

Genome assembly and SNP discovery

We sequenced the genome of a partially inbred female opossum using the whole-genome shotgun (WGS) method^{7,42}. The resulting WGS assembly has total length of 3,475 Mb, consistent with size estimates based on flow cytometry (~3.5-3.6 Gb; Supplementary Notes S1-S2). Approximately 97% of the assembled sequence has been anchored to eight large autosomes and one sex chromosome based on genetic markers mapped by linkage analysis³⁸ or fluorescence *in situ* hybridization⁴³ (FISH; Supplementary Note S3). The draft genome sequence has high continuity, coverage and accuracy (Table 1; Supplementary Note S4).

To enable genetic mapping studies of opossum, we also created a large catalog of candidate single nucleotide polymorphisms (SNPs). We identified ~775,000 heterozygous SNPs within the sequenced individual by analyzing assembled sequence reads. We identified an additional ~510,000 SNPs by generating and comparing ~300,000 sequence reads from three individuals from distinct,

partially outbred laboratory stocks maintained at the Southwest Foundation for Biomedical Research (SFBR; San Antonio, TX) ^{22,44} (Supplementary Note S5). The SNP rates between the different stocks range from 1/360 to 1/140 bases and correlate with the distance between their geographical origins.

Table 1: Genome assembly characteristics

WGS assembly (monDom5)	
Number of sequence reads	38.8 million
Sequence redundancy (Q20 bases)	6.8x
N50 contig size	108 kb
N50 scaffold size	59.8 Mb
Total anchored bases in the assembly	3,412 Mb
Total estimated euchromatic genome size ^a	3,475 Mb
Integration of physical mapping data	
Scaffolds anchored on chromosomes	216
Fraction of genome in anchored and oriented scaffolds	91%
Fraction of genome in anchored, but unoriented, scaffolds	6%
Quality control	
Bases with quality score ≥ 40	98%
Empirical error rate for bases with quality score ≥ 40 ^b	3×10^{-5}
Empirical euchromatic sequence coverage ^b	99%
Bases in regions with low probability of structural error ^c	98%

N50 is the size x such that 50% of the bases reside in contigs/scaffolds of length $\geq x$.

^a Includes anchored sequence and spanned gaps (~2%).

^b Based on comparison with 1.66 Mb of finished BAC sequence.

^c Based on ARACHNE assembly certification (see Supplementary Note S4).

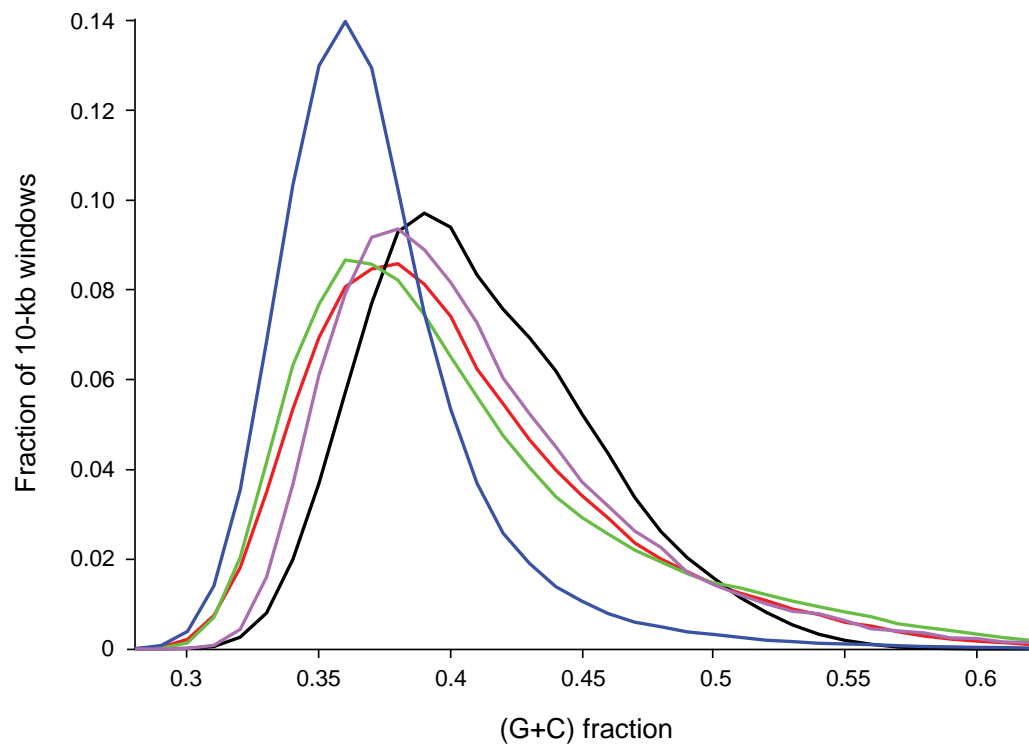


Figure 2. Sequence composition in the opossum genome. Distribution of (G+C)-content in 10-kb windows across the genome in opossum (blue), human (red), mouse (black), dog (green) and chicken (purple).

Genome landscape

The opossum genome has certain unusual properties that provide an opportunity to test recent models of genome evolution. The opossum autosomes are extremely large: they range from 257 Mb to 748 Mb, with the smallest being larger than the largest chromosome previously sequenced in any amniote (human chromosome 1). In contrast, the X chromosome is only ~76 Mb long; this is substantially less than the size of the X chromosome in any sequenced eutherian. Studies of G-banding and chromosome painting have also shown that karyotypes and basic chromosomal organization are extraordinarily conserved throughout Metatheria, even between the distantly related American and Australasian lineages (~55-80 Mya)^{5,45}.

Sequence composition. Recent analyses have uncovered two major trends in the evolution of sequence composition in amniote genomes: First, most modern lineages appear to be experiencing a gradual decline in total (G+C)-content relative to their common ancestors⁴⁶. Second, the local rate of recombination is positively correlated with local (G+C) content and, even more strongly, with the local density of CpG dinucleotides^{20,47}. These observations have led to a proposed model⁴⁸ whereby sequence composition reflects the balance between a genome-wide, AT-biased mutation process and a localized recombination-mediated GC-biased gene conversion (BGC) process. This model predicts that the sequence composition of a genomic region is a function of its historical rate of recombination, with the frequency of hypermutable CpG dinucleotides being a particularly sensitive indicator.

The opossum genome fits the predictions of this model well (see also^{34,35}). Current linkage data³⁸ show that the average recombination rates for the autosomes (~0.2-0.3 cM/Mb) are lower than in other sequenced amniotes (0.5 to more than 3 cM/Mb). Consistent with the proposed model, the mean autosomal (G+C)-content (37.7%) is also lower than in other sequenced amniotes (40.9-41.8%), and in particular, the mean autosomal density of CpGs (0.9%) is two-fold lower than in other amniotes (1.7-2.2%). Because large-scale patterns of recombination appear to be relatively stable in the absence of chromosomal rearrangements^{49,50}, the stability of the opossum karyotype suggests that the majority of the genome has experienced low recombination rates over an extended period. Consistent with this, the sequence composition is also more homogenous than seen in other amniotes (Figure 2).

The subtelomeric regions of autosomes are notable outliers with respect to sequence composition in the opossum genome, providing additional support for the BGC hypothesis. Cytological studies in opossum^{51,52} suggest that the rate of chiasmata formation (and hence meiotic recombination) is relatively uniform across each autosome in males, while it is strongly biased to

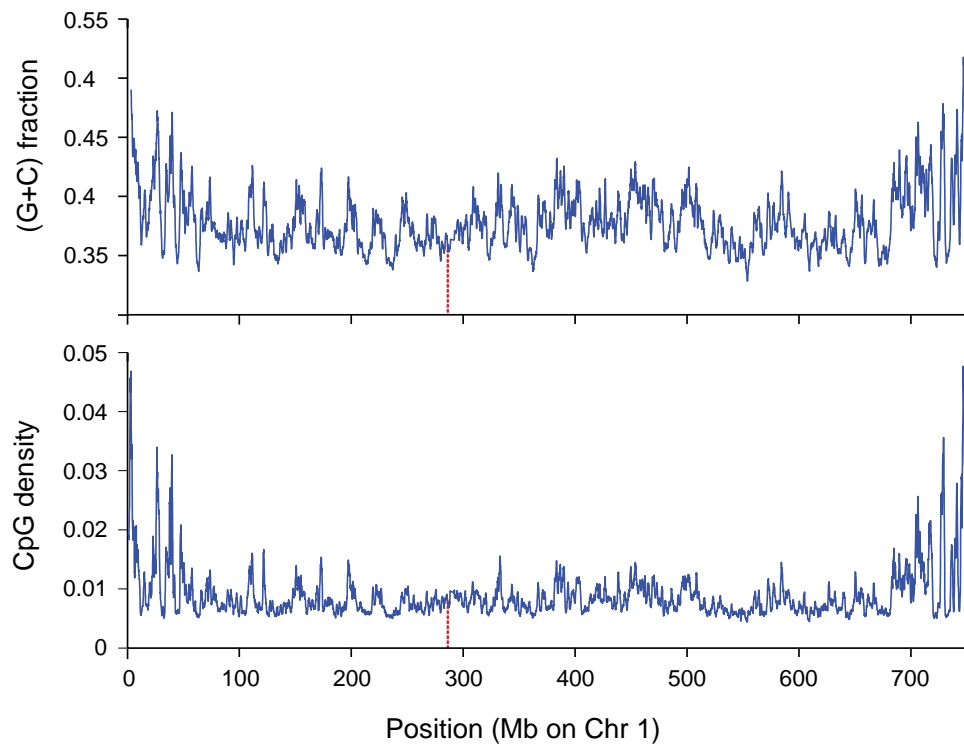


Figure 3. Sequence composition in the opossum genome. Distribution of (G+C)-content (upper) and CpG density (lower) in 1-Mb sliding windows across opossum chromosome 1. The centromere is indicated by the dotted red line. Increased (G+C)-content and CpG densities are evident in the subtelomeric regions.

Table 2: Comparative analysis of genome landscape in opossum and other amniotes

	Opossum	Human	Mouse	Dog	Chicken
Euchromatic genome size (Mb)	3475	2880	2550	2330	1050
Karyotype					
Haploid number	9	23	20	39	33
Autosomal size range (Mb)	258-748	47-247	61-197	27-125	5-201
X chromosome size (Mb)	76	155	167	127	NA
Segmental duplications					
Autosomal (%)	1.7	5.2	5.3	2.5	10.4
Intrachromosomal duplications (%)	76	46	84	ND	ND
Median length between duplications (Mb)	0.18	2.2	1.6	0.33	0.03
X chromosome (%)	3.3	4.1	13	1.7	NA
Interspersed repeats (%)					
Total	52.2	45.5	40.9	35.5	9.4
LINE/non-LTR retrotransposon	29.2	20.0	19.6	18.2	6.5
SINE	10.4	12.6	7.2	10.2	NA
Endogenous retrovirus	10.6	8.1	9.8	3.7	1.3
DNA transposon	1.7	2.8	0.8	1.9	0.8
G+C content (%)					
Autosomal	37.7	40.9	41.8	41.1	41.5
X chromosome	40.9	39.5	39.2	40.2	NA
CpG content (%)					
Autosomal	0.9	2.0	1.7	2.2	2.1
X chromosome	1.4	1.7	1.2	1.9	NA
Recombination rate (cM/Mb)					
Autosomal	~0.2-0.3	1-2	0.5-1	1.3-3.4 ^b	2.5-21
X chromosome ^c	≥0.44 ^d	0.8	0.3	ND	NA

NA, not applicable. ND, no or insufficient data.

^a Range of chromosome-averaged recombination rates.

^b http://www.vgl.ucdavis.edu/research/canine/projects/linkage_map/data/

^c Estimated as 2/3 of the female rate.

^d See text.

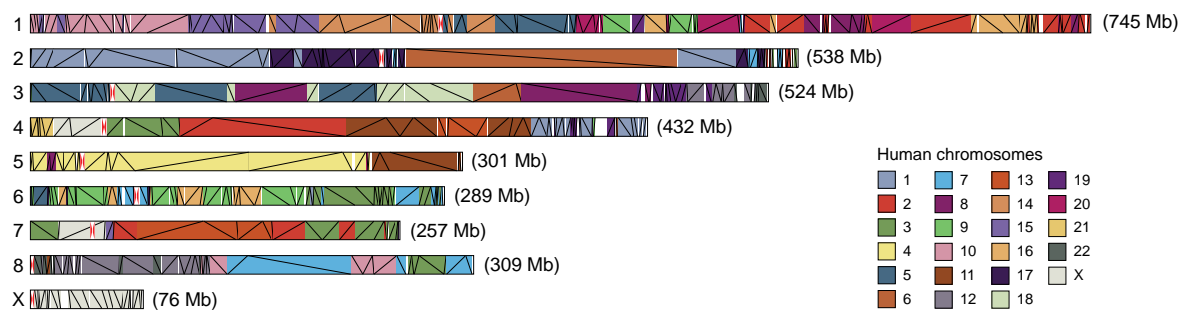


Figure 4. Opossum-human synteny map at 500 kb resolution. Segments on opossum chromosomes 1-8 and X are colored by their orthologous chromosomes in human. Diagonal lines show the extent of collinear syntenic segments. Large gaps in the map (white) typically correspond to extensive gene clusters where synteny is difficult to ascertain. Centromeres are indicated by the opposing red triangles. The estimated size of each chromosome is given on its right. The chromosomal assignments and size estimates reflect all available FISH data (assembly version monDom5).

subtelomeric regions in females. Consistent with a higher sex-averaged rate of recombination, mean (G+C)-content (41.6%) and CpG density (1.9%) are significantly elevated within ~10 Mb of the chromosome ends (Figure 3).

Similarly, the very short X chromosome also supports the BGC hypothesis. Although few linkage data are currently available for opossum X chromosome, the average effective recombination rate must be at least 0.44 cM/Mb, and thus larger than for the autosomes. (This estimate follows from the requirement of at least one meiotic crossover per bivalent in the female germ-line^{53,54}). The mean (G+C)-content (40.9%) and CpG density (1.4%) of the X chromosome are substantially higher than for any of the autosomes. The opossum pattern is thus the opposite of that seen in eutherians, in which the X chromosome has low recombination and low (G+C)-content and CpG density (Table 2).

Segmental duplication. In human and other eutherians, segmental duplications (defined as pairs of regions with $\geq 90\%$ sequence similarity over ≥ 1 kb) are associated with chromosomal fragility and syntenic breakpoints^{55,56}. The relative karyotypic stability of metatherians therefore suggests that they might have a low proportion of segmental duplications.

The overall proportion of segmental duplication in opossum (1.7%) was indeed substantially lower than in other sequenced amniotes (2.5-5.3%). The segmental duplications are also relatively short: only 22 exceed 100 kb in opossum as compared to 483 in human. Additionally, the segmental duplications are more locally distributed: 76% are intrachromosomal (vs. 46% for human) and the median distance between related duplications is 175 kb (vs. 2.2 Mb for human). We find no indication that correction for over-collapsed duplications in the assembly would significantly alter these estimates (Supplementary Note S6).

Transposable elements. Metatherian transposable elements (TEs) largely belong to families also found in eutherians, but can be divided into more than 500 subfamilies, many of which are lineage-specific (catalogued in Repbase⁵⁷). At least 52% of the opossum genome can be recognized as TEs and other interspersed repeats (Table 2)^{33,35}, which is more than in any of the other sequenced amniotes (34-43%). Notably, the opossum genome is significantly enriched in non-LTR retrotransposons (LINEs), with over 29% of the genome sequence comprising copies of various LINE subfamilies. Given the low abundance of segmental duplications, accumulation of TEs appears to be the primary reason for the relatively large opossum genome size. The total euchromatic sequence not recognized as TEs is rather similar in opossum and human (1638 Mb vs 1568 Mb). The enrichment of LINEs may be related to the overall low recombination rate in

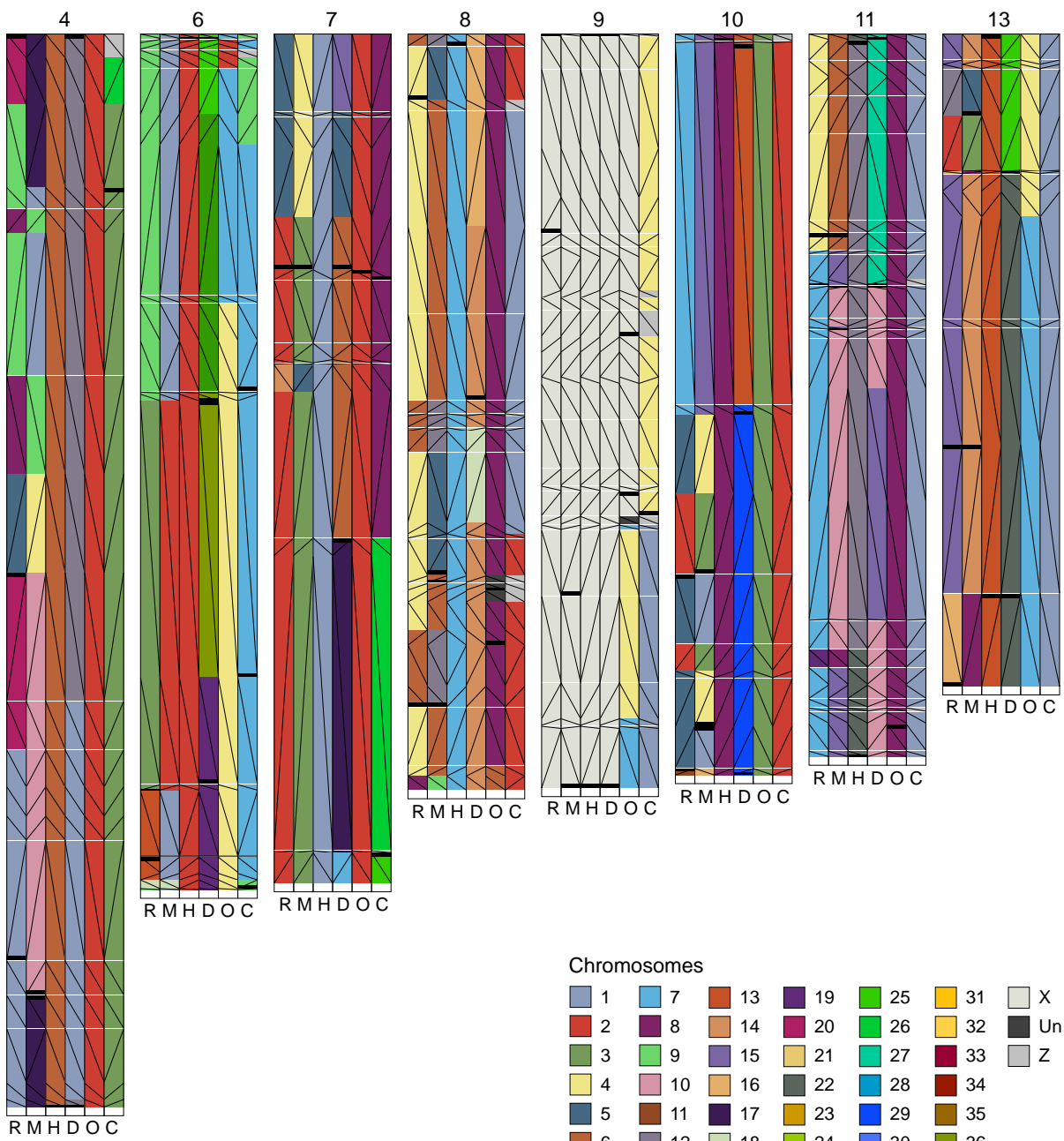
opossum, inasmuch as studies of eutherian genomes have shown that LINEs occur at elevated densities in regions with low local recombination rates ⁴⁷.

Conserved synteny

Identification of syntenic segments between related genomes can facilitate reconstruction of chromosomal evolution and identification of orthologous functional elements. Starting from nucleotide-level, reciprocal-best alignments ('synteny anchors'), we found that the opossum and human genomes can be subdivided (at a resolution of 500 kb) into 510 collinear segments with an N50 size of 19.7 Mb, which cover 93% of the opossum genome (Figure 4). If local rearrangements are disregarded, these segments can be further grouped into of 372 blocks of large-scale, conserved synteny.

Extending this analysis to additional eutherians (mouse, rat and dog), with chicken as an additional outgroup, we created a high-resolution synteny map that reveals 616 blocks of conserved synteny across the five fully sequenced mammals (Supplementary Note S7). Because the majority of synteny breakpoints between human, mouse, rat and dog are clearly lineage-specific (see also ¹⁰), genomic regions that were probably contiguous in the last common boreoeutherian ancestor can be inferred by parsimony (Supplementary Note S8). We found that the mammalian synteny blocks can be used to define 43 connected groups in the ancestral boreoeutherian genome (Figure 5). In fact, the largest 30 groups cover 95% of the human genome (see also ⁵⁸).

The resulting synteny map can be used to clarify chromosomal rearrangements during early mammalian evolution. For example, limited comparative mapping previously revealed that the eutherian X chromosome contains an 'X-conserved region' (XCR) that corresponds to the ancestral therian X chromosome, and an 'X-added region' (XAR), which was translocated from an autosome after the split from Metatheria ^{59,60}. The exact extent of the XCR has been unclear, however, due to unclear synteny to non-mammalian out-groups at its boundary ⁶¹. Using our high-resolution synteny map we can now confidently map the XAR-XCR fusion point to 46.85 Mb on human Xp11.3 (Figure 6).



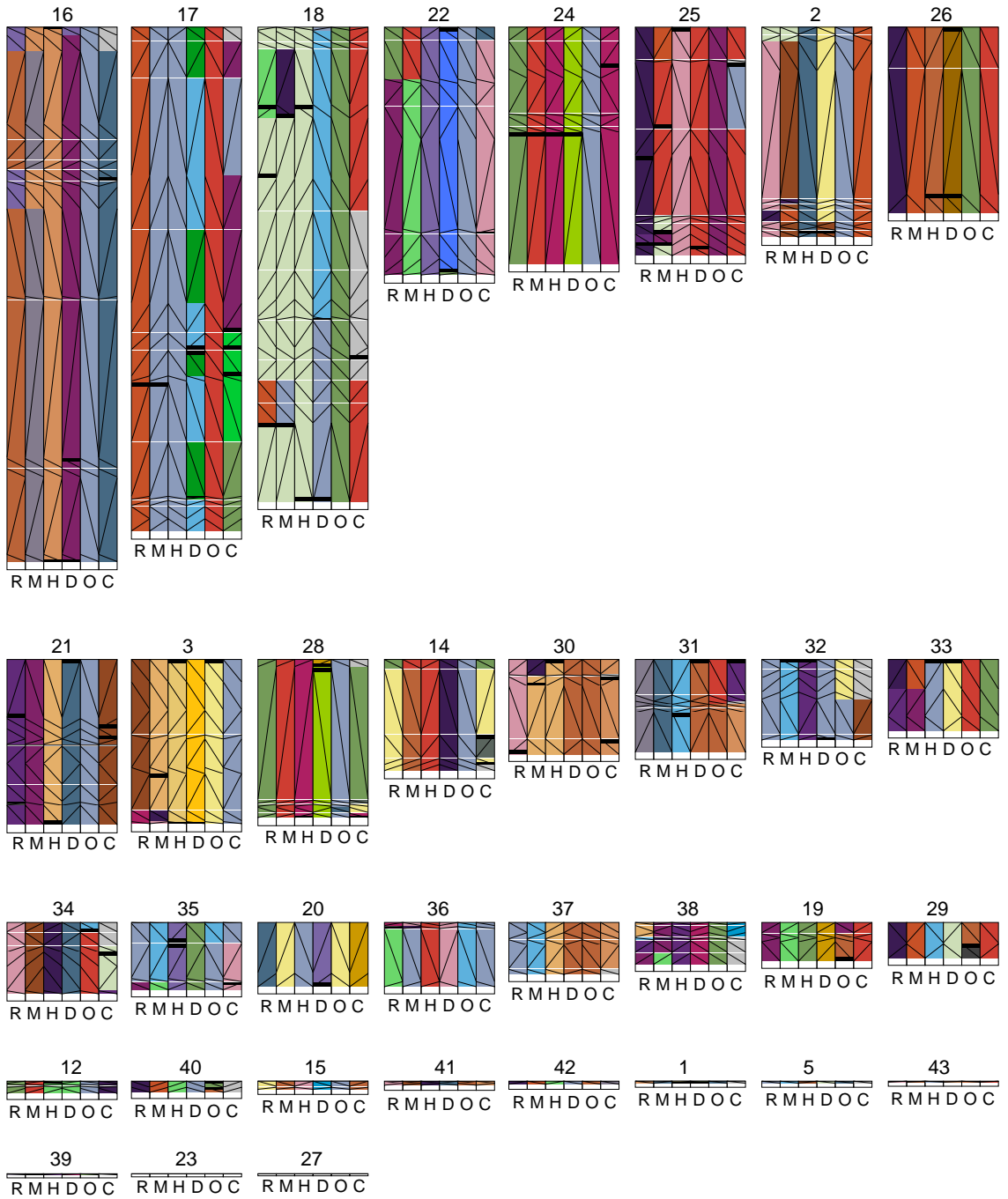


Figure 5. Reconstructed boreoeutherian synteny groups. All 42 ancestral synteny groups are shown sorted by size. For each group, the corresponding chromosomal assignments for mouse (M), rat (R), human (H), dog (D), opossum (O) and chicken (C) are shown.

X inactivation

In opossum and other metatherian mammals, dosage compensation for X-linked genes is achieved through inactivity of the paternally derived X chromosome in females⁶². In contrast, eutherian dosage compensation involves inactivation of the paternal X chromosome at spermatogenesis, reactivation in the early embryo, followed by random and clonally stable inactivation of one of the two X chromosomes in each cell of female embryos⁶³. The random inactivation step is controlled by a complex locus known as the X inactivation center (XIC). In the early female embryo, the non-coding *XIST* gene is transcribed from the XIC and coats one chromosome in *cis* to initiate silencing of the majority of its genes. It has been proposed that paternal X inactivation represents the ancestral therian dosage compensation system, and that random X inactivation is a recent innovation in the eutherian lineage^{64,65}. The opossum genome sequence provides the first opportunity to test major hypotheses about the evolution of this system.

No *XIST* homolog in opossum. We searched all assembled and unassembled opossum WGS sequence for homology to the human and mouse XIC non-coding genes but, in agreement with a recent report⁶⁶, did not find any significant alignments. (In particular, we found no match to the highly conserved 150 bp region overlapping the critical exon 4 of *XIST*; this region is so strongly conserved in Eutheria that it should be readily detectable if present⁴⁰). Analysis of synteny in the regions surrounding the eutherian XIC also revealed that it has been disrupted by large-scale rearrangements (Figure 6)^{40,41}. In eutherians, the XIC is flanked by the ancient protein-coding genes *CDX4-CHIC1* on one side and *SLC16A2-RNF12* on the other side. In both chicken and frog these four genes are clustered in autosomal XIC Homologous Regions (which do not contain homologs of the XIC non-coding genes⁶⁶). On the opossum X chromosome, however, these two pairs of genes are separated by ~29 Mb (compared to ~750 Kb in human). Taken together, the evidence strongly suggests that *XIST* is specific to eutherians^{40,41,66}.

The Lyon repeat hypothesis. LINE/L1 elements are of particular interest to the study of X inactivation. These TEs have been proposed to act as “boosters” for the spread of X inactivation in *cis* from the XIC (reviewed in⁶⁷). This hypothesis is supported in part by the observation that in human, LINE/L1 density is significantly elevated in the XCR (33%) where nearly all genes are inactivated, but approximates the autosomal density in the XAR (19%) where many genes escape inactivation (Figure 7)^{61,68}. In mouse, we found that the LINE/L1 density is elevated in both the XCR (35%) and the XAR (32%), which is consistent with the observation that genes that escape inactivation on the human XAR are often inactivated in mouse⁶⁹. As previously observed in human⁶⁸, the LINE/L1 elevation in mouse is particularly dramatic among recent, lineage-specific

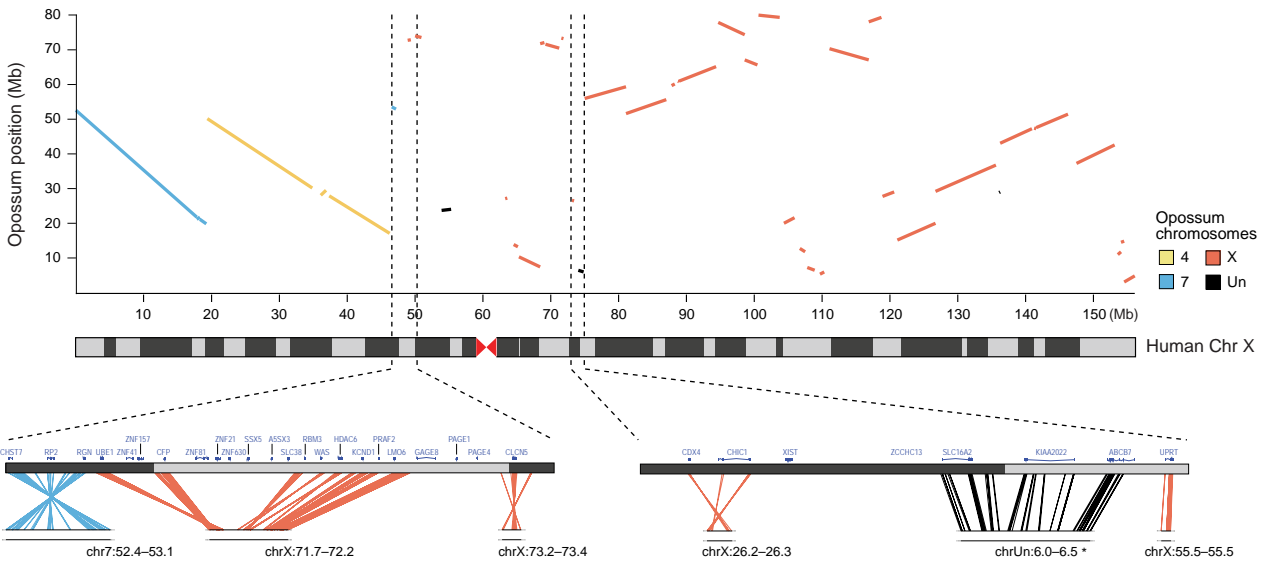


Figure 6. Opossum-human synteny for chromosome X. The dot plot shows correspondence between the human X chromosome and opossum at a resolution of 300 kb. Expanded views, at a resolution of 50-kb, of the XAR-XCR fusion and the XIC are shown on the bottom left and right, respectively. In the XIC region, the closest contig on the distal flank (*) was not anchored in the monDom5 assembly, but has been subsequently mapped near UPRT (opossum chromosome X ~55 Mb) by FISH [40].

subfamilies (Figure 8).

In contrast to human and mouse, the LINE/L1 density on the opossum X chromosome (22%) is significantly lower than in the eutherian XCR, and in fact slightly less than in the autosomal regions homologous to the eutherian XAR (23%). This difference between metatherian and eutherian X chromosomes is not readily explained by any simple correlation between LINE/L1 density, recombination or mutation rates. We therefore conclude that LINE/L1 density is unlikely to be a critical factor for X inactivation in the metatherian lineage, and that the approximately 2-fold increase on the eutherian X chromosome may be directly related to the acquisition of *XIST* and random X inactivation.

Suppression of large-scale rearrangements. Comparative analyses have revealed that the structure of the human X chromosome has remained essentially unchanged since the eutherian radiation^{10,20,61}. A possible reason is that the requirement for *XIST* to spread across the chromosome from a central location has led to selection against structural rearrangements. For example, translocation of LINE/L1-poor XAR segments into the XCR could potentially disrupt inactivation at more distal loci. Consistent with this hypothesis, our synteny map reveals that the XAR and XCR homologous regions have experienced several major rearrangements both in the opossum lineage (~15 lineage-specific synteny breakpoints) and in the eutherian lineage prior to the eutherian radiation (~9 lineage-specific breakpoints). The low rate of rearrangements in the human lineage is therefore unlikely to be due to functions or sequences that were present on the ancestral therian X chromosome, or in early eutherian evolution.

We note that unlike in human, the mouse X chromosome has experienced several rearrangements (with 15 lineage-specific synteny breakpoints), such that the XAR and XCR are no longer two separate segments. This would be consistent with the more comprehensive inactivation in the mouse imposing weaker constraints on rearrangement. Although little is known about the extent of X inactivation in dog or rat, their X chromosomes are also consistent with this hypothesis. The dog X chromosome is collinear with human and is enriched for LINE/L1 only in the XCR (33.4% vs. 16.8% for the XAR). The rat X chromosome has accumulated ~4 lineage-specific synteny breakpoints after the divergence from mouse⁶¹, and is similarly enriched for LINE/L1 in both the XCR (36.7%) and the XAR (34.5%).

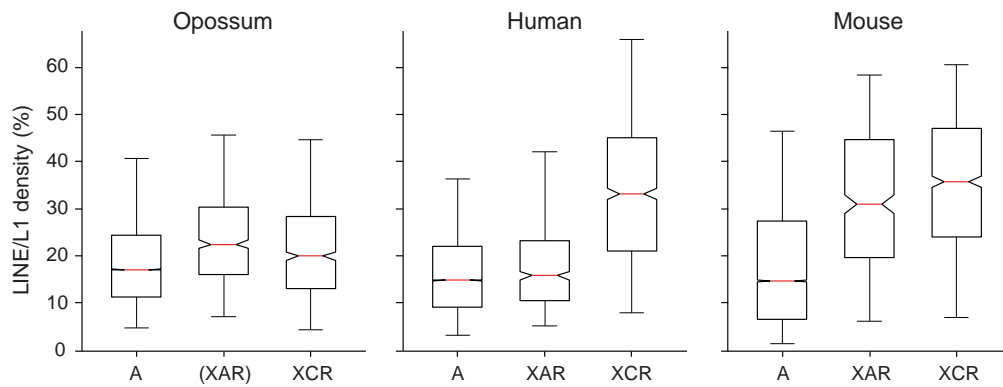


Figure 7. Enrichment of LINE/L1 correlates with random X inactivation. Box plot of LINE/L1 density in 500 kb intervals across the autosomes (A), the X-added region and its homologous regions in opossum (XAR) and the X conserved region (XCR). The red bar is the median. Box edges are the 25th and 75th percentiles. Whiskers show the range.

Genes

The gene content of metatherian and eutherian genomes provides key information about biological functions. We analyzed the gene content of the opossum genome and compared it with that of the human genome. We focused on instances of rapid divergence and duplication of protein-coding genes, which have led to lineage-specific gene complements⁷⁰.

Gene catalog. We generated an initial catalogue of 18,648 predicted protein-coding genes and 946 non-coding genes (primarily snRNA, snoRNA, miRNA and rRNA) in opossum³⁴ (Supplementary Note S9).

We next characterized orthology and paralogy relationships between predicted protein-coding genes in opossum and human¹¹ (Table 3). We could identify unambiguous human orthologs for 15,320 (82%) of the opossum predicted genes, with 12,898 cases having a single copy in each species (1:1 orthologs). Notably, we identified orthologs of key T cell lineage markers such as CD4 and CD8, which had not been successfully identified by cloning in metatherian species³⁹. Most (2,704) of the remaining genes are homologous to human genes, but could not be assigned to orthologous groups with certainty.

A small number (624) of predicted opossum genes have no clear homolog among the human gene predictions. Inspection revealed that most of these are short (median length = 120 amino acids, compared to 445 for 1:1 orthologs) and probably originate from pseudogenes or spurious open reading frames. Only 8 currently have strong evidence of representing functional genes without homologs in humans. These include CPD-photolyase, which is part of an ancestral photorepair system still active in opossum⁷¹, malate synthase⁷² and inosine/uridine hydrolase. The latter two are ancient genes not previously identified in a mammalian species.

Conversely, approximately ~1,100 current gene predictions from human have no clear homolog in the initial opossum catalog. Of these, ~620 can be at least partially aligned to the opossum genome and may not have been annotated as genes due to imperfections in the draft assembly or high sequence divergence. In particular, manual re-annotation identified orthologs of several rapidly evolving cytokines³⁹. The remaining predictions are dominated by gene families known to have undergone expansion and rapid evolution in the human lineage, such as beta-defensins and cancer-testis antigens. Based on our comparison, the opossum genome likely contains ~18,000-20,000 protein-coding genes with the vast majority having eutherian orthologs.

Divergence rates among orthologs. We calculated the synonymous substitution rate (K_S) of 1:1 opossum-human orthologs to estimate the unconstrained divergence rate between the species^{7,10}. The median value of K_S is 1.02. Consistent with expectation, this value is substantially

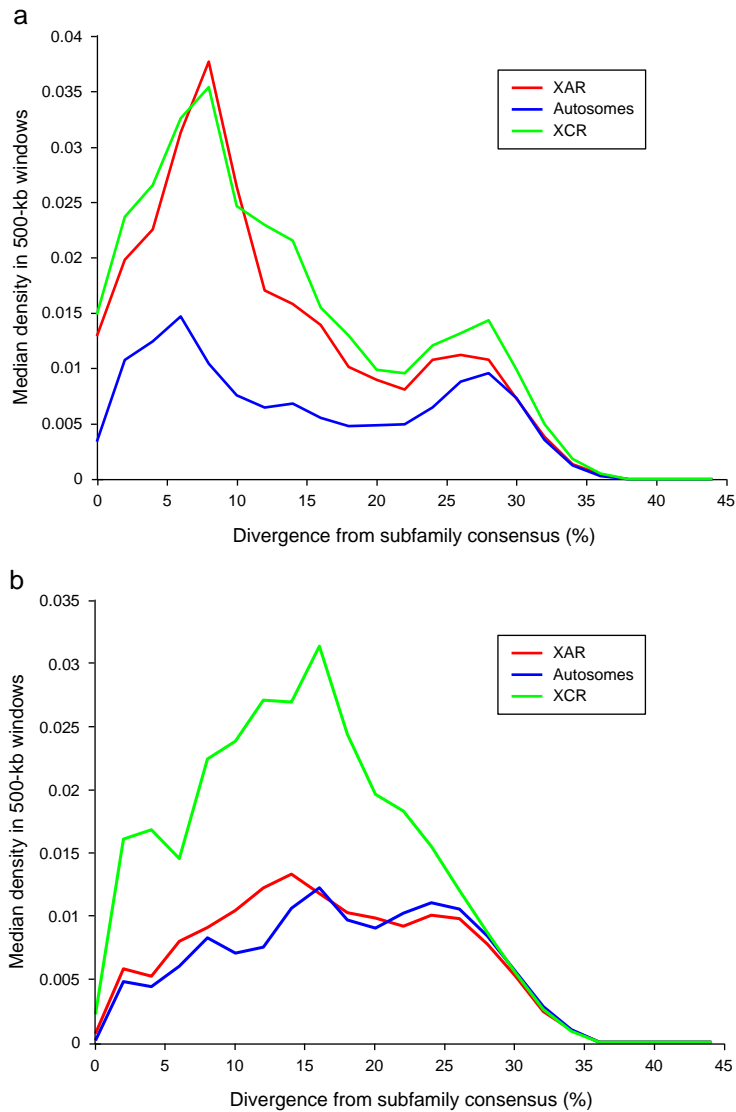


Figure 8. LINE/L1 density in human and mouse by approximate age of insertion. Curves show LINE/L1 density for autosomes, the XAR and the XCR in mouse (left) and human (right) as a function of divergence from the subfamily consensus, which approximates the age of insertion. In both species, the X chromosome enrichment is particularly strong for relatively recent insertions. The XAR enrichment unique to mouse extends almost as far back as the XCR enrichment.

Table 3: Opossum and human gene predictions and projected gene counts

Protein-coding genes	Opossum
Initial predictions	18,648
Orthologs in human ^a	15,320
1:1	12,898
Many:1	1,016
1:Many	451
Many:Many	582
Homologs in human, but unclear orthology ^b	2,704
No predicted homologs in human	624
Projected total ^c	18,000-20,000

^a Includes some cases where multiple transcripts have inconsistent phylogenies, or where the predicted ortholog is a putative pseudogene.

^b Includes members of highly duplicated gene families.

^c Accounting for missed annotations in opossum and removal of probable pseudogenes.

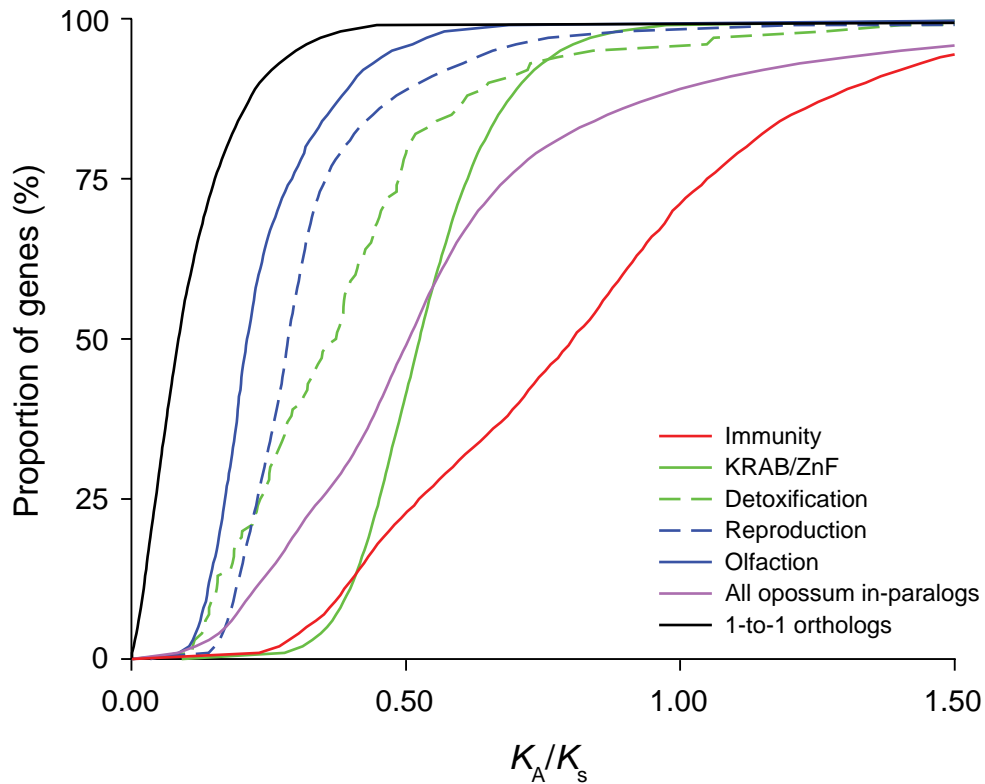


Figure 9. Cumulative distribution of Ka/Ks values for duplicated genes. Estimates are shown for opossum in-paralogs in the most commonly duplicated functional categories: immunity, KRAB zinc finger transcription factors, detoxification (including cytochrome P450, sulfotransferases), reproduction (including vomeronasal receptors, lipocalins and beta-seminoproteins) and olfaction. The total distributions for opossum in-paralogs and opossum-human 1:1 orthologs are shown for comparison.

smaller than the chicken-human K_S value (1.7), with the ratio being very close to the ratio of prior estimates of the divergence times for the two lineages (~180 Mya for opossum and ~310 Mya for chicken).

Notably, the median K_S for orthologs located on the XCR is significantly elevated relative to orthologs located on autosomes in both species (1.2 vs. 1.0; $p < 10^{-3}$) (see also ^{34,35}). This is opposite of what is observed within Eutheria ¹⁰, but is consistent with the expectation that the higher (G+C)-content and recombination rate on the opossum X chromosome relative to its autosomes implies a higher rate of mutation ⁴⁷. A similar elevation can also be detected in subtelomeric regions ³⁴.

Innovation and turnover in gene families. We next studied the evolution of gene family expansions in the metatherian lineage. The opossum gene catalog contains 2,743 (15%) genes that have probably been involved in one or more duplication or gene conversion event since the last common ancestor with eutherian mammals, as inferred from low K_S between the copies (median = 0.41). The number of duplications is one-third fewer than the number of human lineage-specific duplications (4037; 20%), which may reflect the lower rate of segmental duplication in the opossum genome.

We found a large number of lineage-specific copies of genes involved in sensory perception, such as the γ -crystallin family of eye lens proteins ⁷³, and taste, odorant ⁷⁴ and pheromone receptors. Other major lineage-specific duplications were found in the rapidly evolving KRAB zinc-finger family, and in genes related to toxin degradation and dietary adaptations, including cytochrome P450 and various gastric enzymes (see also ³⁴).

Innovation in the innate and adaptive immune systems is visible through substantial duplication or gene conversion involving the leukocyte receptor and natural killer complexes, immunoglobulins, type I interferons and defensins ^{32,39}. The opossum genome also contains a new T-cell receptor (TCR) isotype that is expressed early in ontogeny, prior to conventional TCR and may provide early immune function in the altricial young ³⁷.

The opossum also shows some surprising gene family expansions that are without precedent in other vertebrates. Notable among these are multiple duplications of the nonsense mediated decay factors SMG5 and SMG6, and the pre-mRNA splicing factors, CWC22 and PRP18. The opossum genome also harbors two adjacent paralogous copies of DNA (cytosine-5)-methyltransferase 1 (DNMT1), which catalyses methylation of CpG dinucleotides. It will be interesting to discover if specialized functions have been adopted by these paralogous genes.

The patterns of evolution among duplicated genes largely mirror those observed in eutherians^{34,70}. The set of opossum paralogs is strongly biased towards recent duplications ($K_S < 0.1$) and in general have accumulated a disproportionately high number of nonsynonymous mutations (Figure 9). The median intra-species ratio of nonsynonymous to synonymous substitution rates (K_A / K_S) between paralogs is 0.51, which is six-fold higher than the inter-species ratio seen for 1:1 orthologs (0.086). This is consistent with the rapid gene birth and death model⁷⁵, which predicts that duplicated genes either undergo functional divergence in response to positive selection or rapidly degenerate due to lack of evolutionary benefit.

Conserved sequence elements

The most surprising discovery to emerge from comparative analyses of eutherian genomes is the finding that the majority of evolutionarily conserved sequence does not represent protein-coding genes, but rather are conserved non-coding elements (CNEs)^{7,10}. The opossum genome provides a well-positioned outgroup to study the origin and evolution of these elements.

For simplicity, we will refer to sequence elements as ‘amniote conserved elements’ if they are conserved between chicken and at least one of opossum or human; ‘eutherian conserved elements’ if they are conserved between human and at least one of mouse, rat or dog; and ‘eutherian-specific elements’ if they are eutherian conserved sequence absent from both opossum and chicken. (‘Metatherian-specific elements’ surely also exist, but cannot be identified without additional metatherian genomes).

Loss of amniote conserved elements in mammals. We first studied the extent to which amniote conserved elements have been lost in the human lineage. We focused on ~133,000 conserved intervals between opossum and chicken (68 Mb), ~50% of which overlap protein-coding regions.

Nearly all (97.5%) of these amniote conserved elements can be aligned to the human genome (Figure 10a). We reasoned that some of the remainder might be orthologous to sequence that lies within gaps in the current human assembly, or which had been missed by the initial genome-wide alignment. We therefore repeated the analysis, focusing only on amniote elements present in opossum and occurring in ‘ungapped intervals’ (that is, syntenic intervals between human and opossum that have no sequence gaps whatsoever); the ungapped intervals contain 63% of all conserved elements.

We found that 99.0% of amniote elements in ungapped intervals could be unambiguously aligned to the human genome. The remaining 1.0% of amniote elements could not be found even by

a more sensitive alignment algorithm (Figure 10b), and thus appears to have been lost in the human lineage.

We also performed the converse analysis, by aligning the human and chicken genomes to identify amniote conserved elements potentially lost in opossum. The results were similar, with 99.4% of elements in ungapped intervals being readily aligned to opossum.

We conclude that the vast majority of amniote conserved elements encode such fundamental functions that they cannot be lost in either eutherians or metatherians. Nonetheless, the small fractions that have been lost correspond to more than 1,400 elements in total; it will be interesting to investigate their function and the consequence of their loss. Notably, although protein-coding sequence comprises 50% of all amniote conserved elements, they comprise only 4% of the elements lost in one of the lineages.

Eutherian-specific conserved elements. We next explored the appearance of novel conserved elements in the lineage leading from the common therian ancestor to the boreoeutherian ancestor, which could shed light on the origin of such elements in general. We identified a collection of eutherian conserved elements that cover 104 Mb (3.7%) of the human genome, ~29% of which overlap protein-coding sequence.

Only a small proportion of human conserved protein-coding sequences could not be aligned to the opossum genome (1.1% in ungapped regions; Figure 10c). By contrast, a much larger proportion of human non-coding elements appear to be eutherian-specific (20.5% in ungapped regions). Taking the results from ungapped syntenic intervals as a conservative estimate for the proportion of total innovation, we conclude that approximately 14.8 Mb (1.1% of 30 Mb of coding sequence and 20.5% of 74 Mb of CNEs) of the eutherian conserved elements are eutherian-specific.

The amount of apparent innovation is highest among short and moderately conserved elements (median length = 37 bp; median \log_2 -odds score = 22), probably reflecting that shorter elements may more readily diverge beyond recognition (see also ^{36,76}). Nonetheless, substantial innovation is apparent even among elements that are relatively long and unambiguously conserved within Eutheria. For example, the proportion of eutherian-specific elements is 8.1% among CNEs with phylo-HMM \log_2 -odds score ≥ 60 ($p_{\text{nominal}} \sim 10^{-18}$), which have a median length of 197 bp (Figure 10d).

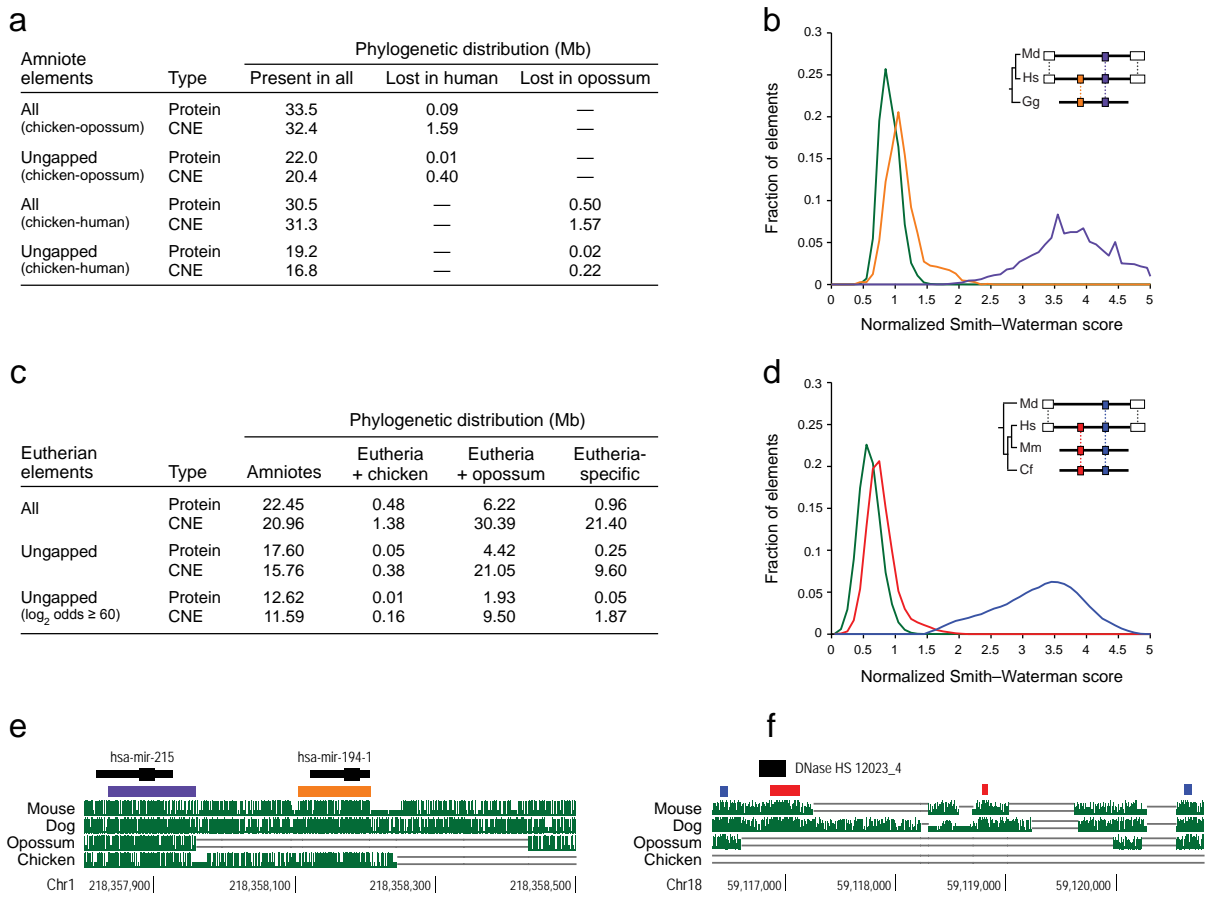


Figure 10. Lineage-specific conserved sequence elements. **a**, Phylogenetic distribution of amniote conserved elements. **b**, Distribution for alignment scores of amniote elements, represented by opossum (human), to ungapped syntenic intervals in the human (opossum) genome, for shared (purple) and lineage-specific (orange) elements, and randomly permuted sequences of the same length and base composition (green). Ungapped syntenic intervals are flanked by two synteny anchors (white) and contain no assembly gaps (insert). **c**, Phylogenetic distribution of eutherian conserved elements. **d**, Distribution of alignment scores for eutherian CNEs (\log_2 -odds ≥ 60), represented by human, to ungapped syntenic intervals in the opossum genome, for shared (blue) and eutherian-specific (red) elements, and randomly permuted sequences of the same length and base composition (green). The bimodal distribution of scores confirm that highly conserved eutherian-specific elements have no significant homology in the syntenic opossum sequence. **e**, The miRNA hsa-mir-194-1 corresponds to an amniote CNE lost in opossum (orange). It is flanked by an unrelated amniote miRNA that is present in opossum (purple). **f**, A eutherian-specific CNE in the intron of *BCL2* (red) overlaps a DNase hypersensitive site in human lymphocytes (black).

Lineage-specific CNEs correspond to functional elements. To establish the biological relevance of lineage-specific CNEs, we examined the overlap of eutherian and amniote CNEs with two disparate sets of experimentally identified functional elements. If the eutherian-specific CNEs were enriched for false positive predictions, we would expect them to be substantially underrepresented among these functional elements.

We first considered a set of known human miRNAs⁷⁷. Of the 51 miRNAs that overlap amniote CNEs, only one (*hsa-mir-194-1*⁷⁸) appears to have been lost in opossum (Figure 10e). (The mature form of this miRNA is identical to a second conserved miRNA, *hsa-mir-194-2*, which does have an opossum ortholog; this apparent redundancy may have made it more susceptible to lineage-specific loss). Of the 183 miRNAs that overlap eutherian CNEs in ungapped syntenic regions, 27 (15%) correspond to eutherian-specific elements. An example is an 87 bp eutherian-specific CNE corresponding to *hsa-mir-28*; it has previously been detected by Northern blot analysis in human and mouse, but not in any non-mammalian species⁷⁹.

We next considered a genome-wide set of DNase hypersensitive sites from human lymphocytes, which represent a variety of putative regulatory elements⁸⁰. Of the 290 sites that overlap amniote CNEs present in human, none overlap instances that are absent in opossum. Of the 2,041 sites that overlap eutherian CNEs in ungapped syntenic regions, 407 (20%) exclusively overlap eutherian-specific elements. An example is a 269 bp eutherian-specific CNE in intron 2 of the apoptosis regulator *BCL2*, which overlaps a DNase hypersensitive site, suggesting it has a *cis*-regulatory function (Figure 10f).

The fraction of eutherian CNEs overlapping DNase hypersensitive sites that are eutherian-specific is strikingly similar to the fraction of all conserved non-coding sequence that is eutherian-specific (20.5%). The fraction of miRNAs that correspond to eutherian-specific CNEs is slightly lower (15%), which is consistent with their higher average conservation scores. In particular, the results provide strong evidence that the majority of eutherian-specific CNEs are likely to be genuine functional elements.

Lineage-specific CNEs associated with key developmental genes. We next explored the distribution of lineage-specific CNEs across the human genome. Overall, there is a strong regional correlation between the density of eutherian CNEs shared with opossum and the density of eutherian-specific CNEs (Spearman's $\rho = 0.82$ for 1 Mb windows; Figure 11). The densities of amniote CNEs present or lost in opossum are also positively correlated (Spearman's $\rho = 0.30$).

Previous studies have shown that both eutherian and amniote CNEs are enriched in certain large, gene-poor regions surrounding genes that play key roles in development, primarily encoding

transcription factors, morphogens and axon guidance receptors^{10,81,82}. For example, 35% of all eutherian CNEs and 49% of all amniote CNEs (in ungapped syntenic regions) lie within the 204 largest clusters of CNEs in the human genome (described in¹⁰). The ~240 key developmental genes in these regions have relatively low rates of amino acid divergence (median $K_A/K_S = 0.03$) and show little evidence of lineage-specific loss or duplications. In contrast, we found that the rate of gain and loss of CNEs in the same regions is only moderately (~30%) lower than elsewhere in the genome. Indeed, we identified more than 37,000 lineage-specific CNEs in these developmentally important regions.

Because experimental studies of CNEs in these regions have frequently uncovered *cis*-regulatory functions affecting the nearby developmental genes^{16,82-85}, the substantial innovation in these regions are candidates for genetic changes underlying differential morphological and neurological evolution in mammalian lineages. This pattern would be consistent with the notion that modification of regulatory networks has been a major force in the evolution of animal diversity⁸⁶⁻⁸⁸.

Eutherian-specific CNEs derived from transposable elements. In general, each eutherian-specific element must have arisen by one of three mechanisms: (i) divergence of an ancestral functional element to such an extent that it is no longer detectably similar to its ortholog in other clades; (ii) duplication of an ancestral functional element giving rise to an element without a 1:1 ortholog in other clades; or (iii) evolution of a novel functional element from sequence that was absent or non-functional in the ancestral genome.

The first mechanism is not likely to account for most of the eutherian-specific CNE sequence, at least among those with high conservation scores: if an ancient functional element underwent such rapid divergence at some point in the eutherian lineage that it is no longer detectable, then there should be concomitant ‘loss’ of an amniote conserved element. But, lineage-specific loss appears to be relatively rare for both amniote elements, as shown above, and for eutherian elements¹⁰. The majority of eutherian-specific conserved elements therefore likely arose after the metatherian divergence, either by adaptive evolution of new or previously non-functional sequence, or by duplication of ancestral elements.

One intriguing source for eutherian-specific CNEs is transposable elements (TEs). A number of researchers have argued that TEs offer an obvious and ideal substrate for the evolution of lineage-specific functions⁸⁹⁻⁹³. TEs contain a variety of functional subunits that can be exapted and modified by the host genome^{89,91}, and they can mediate duplication of existing CNEs to distant genomic locations through transduction or chimerism⁹². Individual instances of CNEs derived from TEs have been described previously^{14,94,95}. However, these cases together comprise only a trivial

fraction of the CNEs in the human genome. It has thus been unclear whether the evolution of CNEs from TEs represents a general mechanism or a rare exception.

When we examined the set of eutherian-specific CNEs, we found a striking overlap with TEs. In ungapped syntenic intervals, at least 16% of eutherian-specific CNEs overlap currently recognized TEs in human. The fraction is similar (14%) if we focus only on the most highly conserved elements (phylo-HMM \log_2 -odds ≥ 60 , see above). The overlapping TEs originate from most major transposon families found in eutherians (Table 4), and are not clearly differentiated from other CNEs in terms of distribution across the genome. This implies that TE-mediated evolution has been a significant creative force in the emergence of recent CNEs. The fact that sequences from TEs themselves can be identified within these CNEs also implies that exaptation of at least a portion of the TE, rather than simply incidental transduction of adjacent sequence, has been a frequent occurrence.

In contrast, the eutherian CNEs that are present in opossum (and thus are more ancient) only rarely show overlap with recognizable TEs (~0.7%). We speculate that many of these CNEs also arose from TEs, but that they are difficult to recognize as such owing to substantial divergence. In fact, three large families of ancient paralogous CNEs have recently been discovered that were clearly distributed around the genome as part of TEs⁹⁶⁻⁹⁸. In each case, only a minority of the family members still retain evidence of transposon-like features. We also previously described ~100 smaller CNE families that pre-date the eutherian radiation, but which had no members associated with known TEs⁹⁸. For all but two of these families, we can find orthologs in the opossum genome for the majority of their members (Supplementary Note S10 and Figure 12). Moreover, closer inspection reveals previously unrecognized transposon-like features in several of these and other ancient CNE families³³.

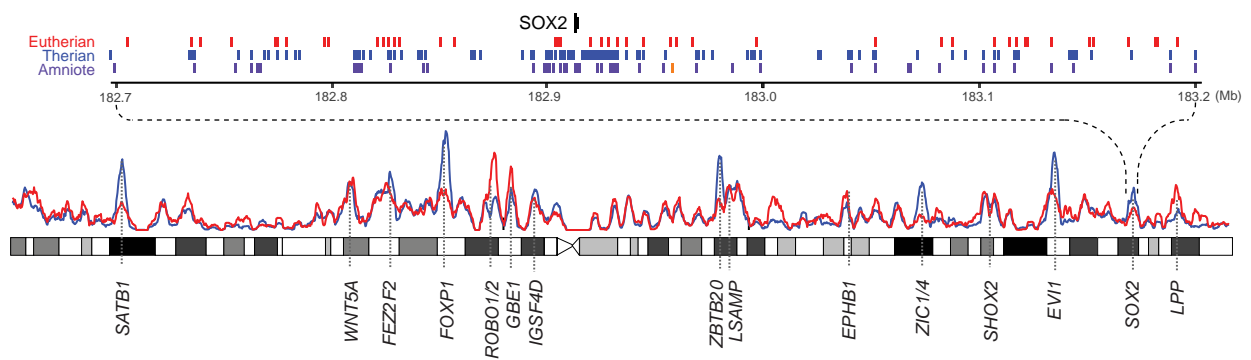


Figure 11. Lineage-specific CNEs near key developmental genes. The densities of eutherian CNEs present (blue) or absent (red) in opossum are plotted in 1 Mb sliding windows across human chromosome 3. Peaks in the distributions often correspond to key developmental genes. The expanded view shows positions of amniote CNEs (purple), eutherian CNEs not overlapping amniote CNEs (blue) and eutherian-specific CNEs (orange) across a 500-kb gene desert surrounding the SOX2 transcription factor. One amniote CNE present in human has been lost in opossum (orange).

Table 4: Eutherian-specific conserved non-coding elements derived from transposons.

Transposon family	All		Log ₂ -odds ≥ 60	
	Number of CNEs ^a	Overlapped length (kb) ^b	Number of CNEs ^a	Overlapped length (kb) ^b
SINE/MIR	9,617	364	363	49
LINE/L1	6,619	286	194	36
LINE/L2	7,616	303	290	47
LINE/CR1	2,520	136	203	36
LINE/RTE	867	48	56	11
LTR/MaLR	1,995	65	25	3.7
LTR/ERV1	140	5.1	1	0.2
LTR/ERVL	992	36	12	2.8
DNA/Tip100	242	9.3	2	0.6
DNA/MER1_type	2,427	93	54	9
DNA/MER2_type	113	5.3	4	0.9
DNA/Tc2	162	8.5	6	1.4
DNA/Mariner	250	14.6	20	3.3
DNA/AcHobo	151	5.1	3	0.3
Unknown (MER121)	49	4	10	1.6
Total	33,760	1,383	1,243	203
Fraction of overlapped CNEs		16%		14%

^a Number of eutherian-specific CNEs in ungapped syntenic regions overlapping annotated TEs.

^b Total length of annotated TE sequence overlapping the CNEs (this is less than the total length of CNEs overlapping TE sequence).

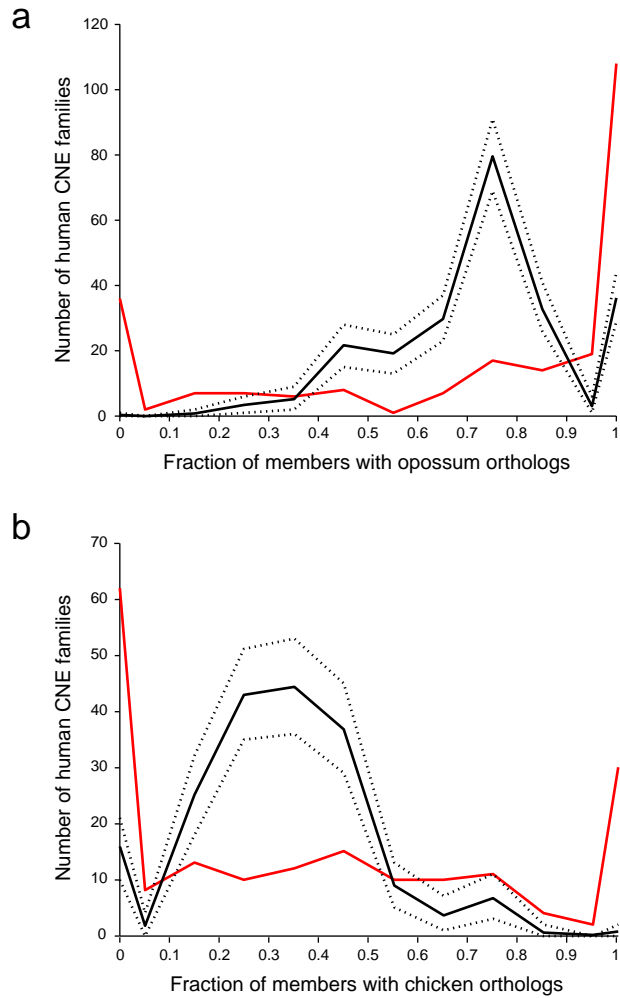


Figure 12. Phylogenetic distribution of paralogous CNE families with 4 or more members in human. a, The observed distribution of the fraction of members of eutherian CNE families with identifiable orthologs in opossum (red) is significantly biased towards 0 and 1, compared to the distribution expected if the presence and absence of opossum orthologs was uniformly distributed between the families (black). b, Discounting completely eutherian-specific CNE families, the observed distribution of the fraction of members of eutherian CNE families with identifiable orthologs in chicken is also significantly biased towards 0 and 1.

Strikingly, the proportion of eutherian-specific CNEs recognizable as TE-derived (16%) is very similar to the proportion of the total aligned sequence between the human, mouse and dog genomes recognizable as ancestral TEs (~17% of ~812 Mb; the vast majority of which is inactive)¹⁰. It is widely suspected that the latter proportion is a significant underestimate due to the difficulty of recognizing TEs that inserted more than ~100-200 Mya^{7,33}. In cases where the TE-related sequence hallmarks are not essential to the subsequent CNE, or where evolution of a new function did not follow immediately after the TE insertion, exapted sequences would be expected to have diverged to the point that they can no longer be readily recognized at a rate similar to inactive insertions. Since this appears to have occurred for most of the families of ancient CNEs described above, it is likely that the proportion of all eutherian (not just eutherian-specific) CNEs derived from TEs is substantially higher than the observed proportion of 16%.

Conclusions

The generation of the first complete genome sequence for a marsupial, *Monodelphis domestica*, provides an important resource for genetic analysis in this unique model organism, as well as the first reference sequence for metatherian mammals. Our initial results demonstrate the utility of this sequence for comparative analyses of the architecture and functional organization of mammalian genomes.

The relationship of sequence composition, segmental duplications and transposable element density with the large and stable karyotype of the opossum genome has provided new support for an emerging, general model of chromosome evolution in mammals. In addition, comparison of the opossum and eutherian X chromosomes revealed that the evolution of random X inactivation correlate with acquisition of *XIST*, elevation in LINE/L1 density and suppression of large-scale rearrangements.

Comparative analysis of protein-coding genes showed that the eutherian complement is largely conserved in opossum. Lineage-specific genes appear to be largely limited to gene families that are rapidly turning over in all mammals, although improved annotations that do not rely on homology to distant species will be required to complete the opossum gene catalog. Identification of a wide array of both conserved and lineage-specific immune genes is particularly notable because limited success in isolating these genes by cloning has led to claims that the metatherian immune system is relatively 'primitive'. Availability of the genome sequence facilitates generation of species-specific reagents, which can be used to gain a better understanding of the metatherian immune response³⁹.

At time-scales longer than the characteristic time of loss for gene duplications, it is clear that innovation in non-coding elements has been substantially more common relative to protein-coding sequences, at least during eutherian evolution. The opossum genome sequence has provided the first estimate of the genome-wide rate of CNE innovation in eutherian evolution, as well as identification of tens of thousands of lineage-specific elements. It has also provided evidence that exaptation of TEs plays a much greater role in the evolution of novel CNEs than has been previously realized.

Sequencing of additional metatherian genomes would be helpful for extending our results by allowing detection of metatherian-specific coding and non-coding elements. In addition, sampling of both the American and Australasian lineages would allow the reconstruction of the genome of their common ancestor, which would complement ongoing efforts for the boreoeutherian ancestral genome⁵⁸. The shorter genetic distance between the ancestral metatherian and boreoeutherian genomes (~0.6-7 substitutions per site) would facilitate a more comprehensive analysis of short and weakly conserved functional elements, for which the phylogenetic distribution and evolutionary origins are still difficult to ascertain.

Methods

WGS sequencing and assembly. Approximately 38.8 million high-quality sequence reads were derived from paired-end reads of 4- and 10-kb plasmids, fosmid and BAC clones, prepared from primary tissue DNA from a single female opossum. The reads were assembled using an interim version of ARACHNE2+ (<http://www.broad.mit.edu/wga/>). No comparative data was used in the assembly process. An intermediate assembly (monDom4) was used for the majority of the analyses reported here. The most recent version (monDom5) has identical sequence content and scaffold structure, but includes additional FISH data as described in Supplementary Note S2.

SNP discovery. The SNP discovery was performed using ARACHNE by comparison of the two haplotypes derived from the opossum assembly using only high-quality discrepancies supported by two or more reads each. Sequence reads from three additional individuals were also aligned to the reference assembly, and SNPs were discovered using SSAHA-SNP⁹⁹. Linkage disequilibrium was assessed using Haploview¹⁰⁰.

Genome alignment and comparisons. The assembly versions used in all comparative analyses were hg17 or hg18 (human), mm8 (mouse), rn4 (rat), canFam2 (dog), monDom4 or monDom5 (opossum) and galGal3 (chicken). The number of aligned nucleotides was counted directly from unfiltered, pairwise BLASTZ alignments (obtained from <http://genome.ucsc.edu>). Synteny maps were generated using standard methods^{7,10}, starting from 320,000 reciprocal-best syntenic anchors identified by PatternHunter¹⁰⁴ (see Supplementary Note S7). Reconstruction of the boreoeutherian ancestral karyotype is described in Supplementary Note S8.

Gene prediction and phylogeny. Opossum protein-coding and non-coding RNA genes were predicted using a modified version of the Ensembl genebuild pipeline¹⁰¹, followed by several rounds of refinement using Exonerate¹⁰² and manual curation. Orthology and paralogy were inferred using the PhyOP pipeline with all predicted opossum and human (Ensembl v40) gene transcripts as input and K_S as the distance metric^{11,34}. Coding regions were aligned according to their amino acid sequences using BLASTP. K_A and K_S were estimated using the codeml program¹⁰⁵, with default settings and the F3X4 codon frequency model. Functional categories were identified using the Gene Ontology¹⁰⁶.

Conserved element prediction. Amniote conserved elements were inferred directly from pairwise BLASTZ alignments of chicken to opossum or human. Every alignment block with more than 75% identity for ≥ 100 bp was classified as an amniote conserved element. Eutherian conserved elements

were inferred using `phastCons`¹⁴ on BLASTZ/MULTIZ^{107,108} alignments of human to mouse, rat and dog. The nonconserved model was fitted using `phyloFit` and the REV model on four-fold degenerate sites from 15,900 human RefSeqs projected onto the same alignments. A separate model was fitted for chromosome X. The scaling parameter for the conserved model was estimated by `phastCons`. Target coverage and expected element length were set to 0.125 and 12, respectively. Predicted eutherian conserved elements that did not fall within a 10-kb or longer synteny ‘net’¹⁰³ between human, mouse and dog were ignored. The coding status of each element was inferred from ≥ 1 nucleotide overlap with entries in the UCSC human ‘known genes’ track¹⁰⁹. Proportions are reported out of the total length of the elements considered. TE-derived eutherian CNEs were inferred from more than 20% nucleotide overlap (median = 100% for all elements, 54% for elements with \log_2 -odds ≥ 60) with human RepeatMasker annotations.

Phylogeny of conserved elements. For amniote conserved elements, pair-wise best-in-genome BLASTZ alignments of opossum to human and *vice versa* were used to infer their phylogenetic distributions. For eutherian conserved elements, concomitant BLASTZ/MULTIZ alignments to opossum and chicken were used. A conserved element was called absent from a species if it was not covered by a single aligned nucleotide in the relevant BLASTZ alignment.

Correction for assembly gaps and initial alignment artifacts. A conserved element was considered to be in an ungapped syntenic interval if it was flanked by two `PatternHunter` synteny anchors within 200-kb of each other on the same contigs in both the human and opossum assemblies. All conserved elements (represented by human or opossum, as appropriate) in ungapped syntenic intervals were realigned to the unmasked genome sequence (in opossum or human) using the `water` program (<http://emboss.sourceforge.net>) with default parameters and a gap extension penalty of 4. A randomly permuted version of each element was also realigned. For amniote conserved elements, only the longest interval with $\geq 75\%$ identity from within the originating alignment block (see above) was realigned. Amniote elements were called lost, and eutherian elements were called eutherian-specific if their Smith-Waterman realignment score, divided by the length of the element, did not exceed the corresponding score for the permuted element plus one. (Conservatively calling an element found if its score simply exceeded the score of the permuted element resulted in 15% of eutherian CNEs in ungapped regions and 8% of those with \log_2 -odds ≥ 60 being called eutherian-specific). Putatively eutherian-specific elements, including *XIST*, were also searched against all opossum sequencing reads using discontinuous `MegaBLAST`.

References

1. Kumar, S. and Hedges, S. B., A molecular timescale for vertebrate evolution. *Nature* 392 (6679), 917-920 (1998).
2. Woodburne, M. O., Rich, T. H., and Springer, M. S., The evolution of tribospheny and the antiquity of mammalian clades. *Mol Phylogenet Evol* 28 (2), 360-385 (2003).
3. Tyndale-Biscoe, C. H., *Life of marsupials*. (CSIRO Publishing, Collingwood, Vic., 2005).
4. Wakefield, M. J. and Graves, J. A. M., Marsupials and monotremes sort genome treasures from junk. *Genome Biol* 6 (5), 218 (2005).
5. Graves, J. A. M. and Westerman, M., Marsupial genetics and genomics. *Trends Genet* 18 (10), 517-521 (2002).
6. Samollow, P. B., Status and applications of genomic resources for the gray, short-tailed opossum, *Monodelphis domestica*, an American marsupial model for comparative biology. *Australian Journal of Zoology* 54 (3), 173-196 (2006).
7. Waterston, R. H. et al., Initial sequencing and comparative analysis of the mouse genome. *Nature* 420 (6915), 520-562 (2002).
8. Gibbs, R. A. et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428 (6982), 493-521 (2004).
9. Chimpanzee Sequencing and Analysis consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437 (7055), 69-87 (2005).
10. Lindblad-Toh, K. et al., Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438 (7069), 803-819 (2005).
11. Goodstadt, L. and Ponting, C. P., Phylogenetic Reconstruction of Orthology, Paralogy, and Conserved Synteny for Dog and Human. *PLoS Comput Biol* 2 (9) (2006).
12. Clamp, M. et al., *Proc Natl Acad Sci U S A* 104 (49), 19428-19433 (2006)
13. Xie, X. et al., Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434 (7031), 338-345 (2005).
14. Siepel, A. et al., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15 (8), 1034-1050 (2005).
15. Pedersen, J. S. et al., Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* 2 (4), e33 (2006).
16. Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M., Scanning human gene deserts for long-range enhancers. *Science* 302 (5644), 413 (2003).
17. Ovcharenko, I., Stubbs, L., and Loots, G. G., Interpreting mammalian evolution using Fugu genome comparisons. *Genomics* 84 (5), 890-895 (2004).
18. Prabhakar, S. et al., Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* 16 (7), 855-863 (2006).
19. Margulies, E. H. et al., An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci U S A* 102 (13), 4795-4800 (2005).
20. Hillier, L. W. et al., Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432 (7018), 695-716 (2004).

21. VandeBerg, J. L., The Gray Short-Tailed Opossum (*Monodelphis-Domestica*) as a Model Didelphid Species for Genetic Research. *Australian Journal of Zoology* 37 (2-4), 235-247 (1990).
22. VandeBerg, J. L., in *UFAW Handbook on the Management of Laboratory Animals. Vol. 1: Terrestrial Vertebrates*, edited by T. Poole and P. English (Blackwell Science, Oxford, 1999), pp. 193-209.
23. Murphy, S. K. and Jirtle, R. L., Imprinting evolution and the price of silence. *Bioessays* 25 (6), 577-588 (2003).
24. Rapkins, R. W. et al., Recent assembly of an imprinted domain from non-imprinted components. *PLoS Genet* 2 (10), e182 (2006).
25. Weidman, J. R. et al., Phylogenetic footprint analysis of IGF2 in extant mammals. *Genome Res* 14 (9), 1726-1732 (2004).
26. Deakin, J. E. et al., Evolution and comparative analysis of the MHC Class III inflammatory region. *BMC Genomics* 7, 281 (2006).
27. Deakin, J. E., Olp, J. J., Graves, J. A., and Miller, R. D., Physical mapping of immunoglobulin loci IGH, IGK, and IGL in the opossum (*Monodelphis domestica*). *Cytogenet Genome Res* 114 (1), 94H (2006).
28. Belov, K. et al., Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex. *PLoS Biol* 4 (3), e46 (2006).
29. Wintzer, M. et al., Strategies for identifying genes that play a role in spinal cord regeneration. *J Anat* 204 (1), 3-11 (2004).
30. VandeBerg, J. L. et al., Genetic analysis of ultraviolet radiation-induced skin hyperplasia and neoplasia in a laboratory marsupial model (*Monodelphis domestica*). *Arch Dermatol Res* 286 (1), 12-17 (1994).
31. Baker, M. L. et al., Analysis of a set of Australian northern brown bandicoot expressed sequence tags with comparison to the genome sequence of the south American grey short-tailed opossum. *BMC Genomics* 8, 50 (2007).
32. Belov, K. et al., Characterization of the opossum immune genome provides insights into the evolution of the mammalian immune system. *Genome Res* (2007).
33. Gentles, A. J. et al., Evolutionary dynamics and biological impact of transposable elements in the short-tailed opossum *Monodelphis domestica*. *Genome Res* (2007).
34. Goodstadt, L., Heger, A., Webber, C., and Ponting, C. P., An analysis of the genome of a marsupial *Monodelphis domestica*: Evolution of lineage-specific genes and giant chromosomes. *Genome Res* (2007).
35. Gu, W. et al., Phylogenetic detection, population genetics, and distribution of active SINEs in the genome of *Monodelphis domestica*. *Gene* (2007).
36. Mahony, S., Corcoran, D. L., and Benos, P. V., Evolutionary conservation of mammalian cis-regulatory regions in *Monodelphis*. *Genome Biol* 8, R84 (2007)
37. Parra, Z. E. et al., A new T cell receptor discovered in marsupials. *Proc Natl Acad Sci U S A*. 104, 9776-9781 (2007).
38. Samollow, P. B. et al., A microsatellite-based, physically anchored linkage map for the gray, short-tailed Opossum (*Monodelphis domestica*). *Chromosome Res* (2007).
39. Wong, E. S., Young, L. J., Papenfuss, A. T., and Belov, K., In silico identification of opossum cytokine genes suggests the complexity of the marsupial immune system rivals that of eutherian mammals. *Immunome Res* 2, 4 (2006).
40. Hore, T., Koina, E., Wakefield, M. J., and Graves, J. A. M., XIST is absent from the X chromosome, and its flanking region is disrupted in non-placental mammals. *Chromosome Res* 15, 147-161 (2007)..

41. Davidow, L. S. et al., The Search for a Marsupial XIC Reveals a Break with Vertebrate Synteny. *Chromosome Res* 15, 137-146 (2007).
42. Venter, J. C. et al., The sequence of the human genome. *Science* 291 (5507), 1304-1351 (2001).
43. Duke, S. E. et al., Integrated Cyteogenetic BAC Map of the Genome of the Gray Short-Tailed Opossum, *Monodelphis domestica*. *Chromosome Res* doi:10.1007/s10577-007-1131-4 (2007).
44. VandeBerg, J. L., The Laboratory opossum (*Monodelphis domestica*) in laboratory research. *ILAR Journal* 38, 4-12 (1997).
45. Rens, W. et al., Karyotype relationships between distantly related marsupials from South America and Australia. *Chromosome Res* 9 (4), 301-308 (2001).
46. Belle, E. M., Duret, L., Galtier, N., and Eyre-Walker, A., The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *J Mol Evol* 58 (6), 653-660 (2004).
47. Jensen-Seaman, M. I. et al., Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* 14 (4), 528-538 (2004).
48. Duret, L., Eyre-Walker, A., and Galtier, N., A new perspective on isochore evolution. *Gene* (2006).
49. Dumas, D. and Britton-Davidian, J., Chromosomal rearrangements and evolution of recombination: comparison of chiasma distribution patterns in standard and robertsonian populations of the house mouse. *Genetics* 162 (3), 1355-1366 (2002).
50. Myers, S. et al., A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310 (5746), 321-324 (2005).
51. Hope, R. M., Selected features of marsupial genetics. *Genetica* 90 (2-3), 165-180 (1993).
52. Sharp, P. J. and Hayman, D. L., An examination of the role of chiasma frequency in the genetic system of marsupials. *Heredity* 60 (Pt 1), 77-85 (1988).
53. Holm, P. B., Ultrastructural analysis of meiotic recombination and chiasma formation *Tokai J Exp Clin Med* 11 (6), 415-436 (1986).
54. Samollow, P. B. et al., First-generation linkage map of the gray, short-tailed opossum, *Monodelphis domestica*, reveals genome-wide reduction in female recombination rates. *Genetics* 166 (1), 307-329 (2004).
55. Bailey, J. A. et al., Hotspots of mammalian chromosomal evolution. *Genome Biol* 5 (4), R23 (2004).
56. Webber, C. and Ponting, C. P., Hotspots of mutation and breakage in dog and human chromosomes. *Genome Res* 15 (12), 1787-1797 (2005).
57. Jurka, J. et al., Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110 (1-4), 462-467 (2005).
58. Ma, J. et al., Reconstructing contiguous regions of an ancestral genome. *Genome Res* (2006).
59. Kohn, M. et al., Wide genome comparisons reveal the origins of the human X chromosome. *Trends Genet* 20 (12), 598-603 (2004).
60. Graves, J. A., Sex chromosome specialization and degeneration in mammals. *Cell* 124 (5), 901-914 (2006).
61. Ross, M. T. et al., The DNA sequence of the human X chromosome. *Nature* 434 (7031), 325-337 (2005).
62. Cooper, D. W., Johnston, P. G., Graves, J. A., and Watson, J. M., X-inactivation in marsupials and monotremes. *Sem. Dev. Biol.* 4, 117-128 (1993).
63. Heard, E., Recent advances in X-chromosome inactivation. *Curr Opin Cell Biol* 16 (3), 247-255 (2004).

64. Wakefield, M. J., Keohane, A. M., Turner, B. M., and Graves, J. A., Histone underacetylation is an ancient component of mammalian X chromosome inactivation. *Proc Natl Acad Sci U S A* 94 (18), 9665-9668 (1997).
65. Reik, W. and Lewis, A., Co-evolution of X-chromosome inactivation and imprinting in mammals. *Nat Rev Genet* 6 (5), 403-410 (2005).
66. Duret, L. et al., The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* 312 (5780), 1653-1655 (2006).
67. Lyon, M. F., Do LINEs Have a Role in X-Chromosome Inactivation? *J Biomed Biotechnol* 2006 (1), 59746 (2006).
68. Bailey, J. A., Carrel, L., Chakravarti, A., and Eichler, E. E., Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc Natl Acad Sci U S A* 97 (12), 6634-6639 (2000).
69. Disteche, C. M., Filippova, G. N., and Tsuchiya, K. D., Escape from X inactivation. *Cytogenet Genome Res* 99 (1-4), 36-43 (2002).
70. Emes, R. D., Goodstadt, L., Winter, E. E., and Ponting, C. P., Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum Mol Genet* 12 (7), 701-709 (2003).
71. Kato, T., Jr. et al., Cloning of a marsupial DNA photolyase gene and the lack of related nucleotide sequences in placental mammals. *Nucleic Acids Res* 22 (20), 4119-4124 (1994).
72. Kondrashov, F. A. et al., Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol Direct* 1, 31 (2006).
73. Wistow, G. et al., gammaN-crystallin and the evolution of the betagamma-crystallin superfamily in vertebrates. *Febs J* 272 (9), 2276-2291 (2005).
74. Grus, W. E., Shi, P., Zhang, Y. P., and Zhang, J., Dramatic variation of the vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals. *Proc Natl Acad Sci U S A* 102 (16), 5767-5772 (2005).
75. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* 431 (7011), 931-945 (2004).
76. Dermitzakis, E. T. and Clark, A. G., Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* 19 (7), 1114-1121 (2002).
77. Griffiths-Jones, S. et al., miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 34 (Database issue), D140-144 (2006).
78. Michael, M. Z. et al., Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol Cancer Res* 1 (12), 882-891 (2003).
79. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T., Identification of novel genes coding for small expressed RNAs. *Science* 294 (5543), 853-858 (2001).
80. Crawford, G. E. et al., Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* 16 (1), 123-131 (2006).
81. Sandelin, A. et al., Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* 5 (1), 99 (2004).
82. Woolfe, A. et al., Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3 (1), e7 (2005).
83. Bailey, P. J. et al., A global genomic transcriptional code associated with CNS-expressed genes. *Exp Cell Res* 312 (16), 3108-3119 (2006).

84. de la Calle-Mustienes, E. et al., A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res* 15 (8), 1061-1072 (2005).
85. Pennacchio, L. A. et al., In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444 (7118), 499-502 (2006).
86. Carroll, S. B., Evolution at two levels: on genes and form. *PLoS Biol* 3 (7), e245 (2005).
87. Davidson, E. H. and Erwin, D. H., Gene regulatory networks and the evolution of animal body plans. *Science* 311 (5762), 796-800 (2006).
88. Stathopoulos, A. and Levine, M., Genomic regulatory networks and animal development. *Dev Cell* 9 (4), 449-462 (2005).
89. Britten, R. J., Mobile elements inserted in the distant past have taken on important functions. *Gene* 205 (1-2), 177-182 (1997).
90. Britten, R. J. and Davidson, E. H., Gene regulation for higher cells: a theory. *Science* 165 (891), 349-357 (1969).
91. Brosius, J., Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* 107 (1-3), 209-238 (1999).
92. Kazazian, H. H., Jr., Mobile elements: drivers of genome evolution. *Science* 303 (5664), 1626-1632 (2004).
93. Marino-Ramirez, L., Lewis, K. C., Landsman, D., and Jordan, I. K., Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet Genome Res* 110 (1-4), 333-341 (2005).
94. Cooper, G. M. et al., Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15 (7), 901-913 (2005).
95. Silva, J. C. et al., Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet Res* 82 (1), 1-18 (2003).
96. Bejerano, G. et al., A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441 (7089), 87-90 (2006).
97. Nishihara, H., Smit, A. F., and Okada, N., Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res* 16 (7), 864-874 (2006).
98. Xie, X., Kamal, M., and Lander, E. S., A family of conserved noncoding elements derived from an ancient transposable element. *Proc Natl Acad Sci U S A* 103 (31), 11659-11664 (2006).
99. Ning, Z., Cox, A. J., and Mullikin, J. C., SSAHA: a fast search method for large DNA databases. *Genome Res* 11 (10), 1725-1729 (2001).
100. Barrett, J. C., Fry, B., Maller, J., and Daly, M. J., Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21 (2), 263-265 (2005).
101. Birney, E. et al., Ensembl 2006. *Nucleic Acids Res* 34 (Database issue), D556-561 (2006).
102. Slater, G. S. and Birney, E., Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 31 (2005).
103. Kent, W. J. et al., Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* 100 (20), 11484-11489 (2003).
104. Ma, B., Tromp, J., and Li, M., PatternHunter: faster and more sensitive homology search. *Bioinformatics* 18 (3), 440-445 (2002).
105. Yang, Z., PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13 (5), 555-556 (1997).

106. Ashburner, M. et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium *Nat Genet* 25 (1), 25-29 (2000).
107. Blanchette, M. et al., Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14 (4), 708-715 (2004).
108. Schwartz, S. et al., Human-mouse alignments with BLASTZ. *Genome Res* 13 (1), 103-107 (2003).
109. Hsu, F. et al., The UCSC Known Genes. *Bioinformatics* 22 (9), 1036-1046 (2006).

Supplementary Notes

S1 BAC library construction

A BAC library was constructed for the purpose of end-sequencing and subsequent anchoring of the WGS shotgun assembly. From the same partially inbred female that was selected for sequencing, a 15X coverage BAC library (384,000 total clones, 175 kb average insert size) was constructed using *EcoRI* partial digests of agarose-embedded DNA. This library was designated VMRC18.

For BAC library construction, nuclei were isolated from brain tissue using a Dounce homogenizer and sequential centrifugation steps. Nuclei were then embedded in Incert agarose and high molecular weight DNA was prepared *in situ* in the agarose. Generation of the library closely followed the cloning approach developed in P. de Jong's laboratory¹. DNA fragments from the appropriate size fraction were cloned into the pCC1BAC vector (Epicentre Technologies). The ligation products were then transformed into DH10B (T1 resistant) electro-competent cells (Invitrogen). The library was arrayed into 1000 384-well microtiter dishes. Analysis of 120 randomly selected BACs via pulsed field gel electrophoresis of *NotI*-digested DNA indicated that the average insert size was around 175 kb.

This library and a previously constructed 10X coverage BAC library (VMRC6) from a male *Monodelphis domestica* specimen are available through the BACPAC Resources Center at the Children's Hospital Oakland Research Institute (<http://bacpac.chori.org/>).

S2 Assembly versions

Starting in October 2004, a series of draft assemblies were made publicly available by the Broad Institute. The monDom4 assembly (January 2006), provided the basis for many of the analyses described in the main manuscript, while an updated version was subsequently released as monDom5 (October 2006). All versions can be accessed from GenBank using NCBI whole-genome shotgun project accession ID AAFR00000000, and from <http://www.broad.mit.edu/ftp/pub/assemblies/mammals/monodelphis/>.

Both monDom4 and monDom5 have the same sequence content and contig/scaffold structure, but monDom5 incorporated additional FISH data (see S3), which increased the percentage of genome in anchored scaffolds from 94.6% to 96.5%. Three scaffolds were reassigned to different chromosomes: Scaffolds 141 (2.0 Mb) and 151 (1.8 Mb), which were erroneously mapped to chromosomes 4 and 7, respectively, in monDom4, were reassigned to chromosome X in monDom5. Scaffold 108 (4.1 Mb) was reassigned from chromosome 5 to chromosome 8. Nine

additional scaffolds that were unanchored in monDom4 could be anchored in monDom5: 1 to chromosome 3, 1 to chromosome 4, 1 to chromosome 6, and 6 to chromosome X. In addition there was minor shuffling of scaffold order and orientation within all chromosomes.

S3 Anchoring scaffolds

The assembly was anchored using primarily FISH mapping data. The linkage map data (200 SSLPs) was integrated with the FISH data and used to further order and orient supercontigs and to confirm the FISH data. For FISH mapping, two BAC clones were selected from every scaffold ≥ 1 Mb in size, and one clone each from scaffolds ≥ 500 kb. Additional clones were selected if the initial selections did not give unambiguous order and orientation.

In total, 402 BAC probes were prepared, hybridized and visualized using previously described protocols ². Metaphase chromosomes were prepared by mitotic stimulation of peripheral lymphocytes from each of two male and two female *Monodelphis domestica*. The precise cytogenetic band location of each probe was determined from analysis of no fewer than 10 metaphase spreads (20 chromosomes). Cytogenetic assignments were made according to the chromosome nomenclature of Pathak *et al.* ³.

Of the 395 clones mapped by FISH, 384 (97.2%) had a unique cytogenetic location and were assigned to cytogenetic band. The remaining 11 (2.8%) clones mapped to more than one cytogenetic location. To order clones that mapped within each band, the FISH analysis was repeated using groups of five clones. Clones that demonstrated overlapping signal in metaphase FISH were also analyzed using interphase FISH to resolve the precise clone order where possible. The full details of the cytogenetic anchoring of the monDom5 genome assembly and detailed cytogenetic BAC map is published elsewhere ⁴.

In addition, the assembly was compared to the fingerprint map generated in the laboratory of M. A. Marra. The order of BAC end read placement in the assembly was compared to the ordering of clone in the fingerprint map (no analysis of individual fingerprint bands was performed). The main purpose was to validate the order of BACs within scaffolds and to look for regions where a single fingerprint contig would contain multiple assembly scaffolds thereby enabling further ordering and orientation of the assembly scaffolds. No additional sequence could be ordered and oriented in this way.

S4 Assembly quality control

Several quality control steps were incorporated into the assembly process to ensure optimal continuity, structural integrity and nucleotide accuracy.

The continuity of the assembly is reflected in the large uninterrupted sequence ‘contigs’ (N50 length = 108 kb), which are joined into much larger ‘scaffolds’ (N50 length = 59.8 Mb). The high degree of continuity implies that the majority of genes should be present without sequence gaps. The coverage of the euchromatic genome is estimated to be ~98%, based on the small proportion of the assembly residing within spanned gaps.

The structural integrity of the assembly was assessed with a substantially revised version of the certification process first developed for the dog genome⁵. As before, internal inconsistencies in the assembly were used as evidence, but unlike the previous process which classified regions as either ‘certified’ (essentially equivalent to finished sequence) or ‘uncertified’ (containing assembly errors), the new process allows regions that are highly likely to be misassembled to be distinguished from those which are merely possibly incorrect.

The new certification model converts multiple forms of evidence into probabilities for the presence of different types of assembly errors across the genome: *structural problems*, where different regions of the genome have been joined (A simplified measure of this is the fraction of reads that are placed consistently with their partner. In the monDom5 genome assembly, 94.9% of reads have a consistent placement, which is similar to the 96.5% of reads with consistent placement in the dog genome.); *local deletions and insertions*, where either portions of the genome are missing or sequence from a different region has been inserted, but the overall structure is correct; and *high quality base errors*, where isolated bases are wrong, but reported as having a quality score of 50 or higher.

Approximately 87% of the opossum sequence resides in regions that are certified to be free of structural assembly errors in the monDom5 assembly. The remaining 13% are declared as ‘uncertified’, but only 2% of the genome is labeled as having a significant probability of containing an error. Of these latter regions, half reside in unanchored sequences. For comparison, by the new methodology 89% of the dog genome resides in certified regions in the most recent canFam2 assembly, while only 3% have a significant probability of error.

To assess nucleotide level accuracy, 11 finished BACs (1.66 Mb) from the sequenced opossum were obtained from Genbank. As each finished BAC contains only one haplotype, while the assembled genome contains a mixture of two haplotypes, we restricted the comparison to homozygous regions in the assembly. Such regions were located by grouping the SNPs identified

from the two assembled haplotypes (see S5) into consecutive blocks of either ≥ 0.3 SNP/kb, or ≤ 0.01 SNP/kb. Blocks of less than 100 kb were ignored, as were blocks with over 50% of the bases excluded due to gaps or uncertified sequences. Those blocks with rates ≤ 0.01 SNP/kb were considered homozygous and span 900-kb. The finished BACs were aligned to the assembly using an alignment tool built into ARACHNE. Those alignments falling within regions declared homozygous (0.9 Mb) were accepted for the base accuracy estimate. A somewhat higher error rate can be expected in polymorphic regions of the genome.

S5 Single Nucleotide Polymorphism discovery

SNPs were discovered in the *Monodelphis domestica* genome sequence as a standard part of the ARACHNE assembly process. Such SNPs represent cases where the single individual (I-1438) sequenced was heterozygous. Assembly SNPs were only called within certified regions (see S4), and if each allele were observed at least twice, implying that ≥ 4 separate shotgun reads aligned to the base containing the putative SNP.

Additional SNPs were discovered by low-coverage shotgun sequencing of three stocks frequently used in laboratory research: B5539 (SFBR Population 1; 104,042 passing reads), D3508 (SFBR Population 2; 100,000 reads), and C5181 (SFBR Population 5; 99,212 reads). Sequencing reads were trimmed to include only intervals of 250 bp or longer with a running 20 bp average quality score of at least 20. These intervals were aligned to the monDom4 assembly and SNPs were discovered using SSAHA-SNP⁶ with a window size of 500bp and a ceiling of 0.02 SNP/bp in any single comparison. SNPs were required to have a Phred quality score of 23 or higher with a minimum quality score of 15 for each of five bases either side⁷. SNPs are available at <http://www.broad.mit.edu/mammals/opossum/> and have been submitted to dbSNP.

To estimate the validation rate of the discovered SNPs, ten 200 kb regions (two regions each on chromosome 1 and 2 and one each on chromosome 3-8 and X) were chosen randomly from regions that included at least 40 SNPs and were more than 1 Mb away from a telomere. Within each region, 20 SNPs were selected for genotyping, with individual SNPs given preference if their source population was under-represented in the region. On the X chromosome, most available SNPs originated from the assembly. Validation genotyping was performed on a Sequenom Mass Spectrometry system. After removal of SNPs with low call rate ($<75\%$), 126 SNPs remained. The average validation rate for the autosomes was 94.9% ($n = 118$). On chromosome X the validation rate was only 62% ($n = 8$), but this was likely due to an unfortunate choice of region, which turned out to be highly repetitive. The applicability of the SNP set for different laboratory stocks was

assessed using representatives from SFBR Populations 1, 2, 3 and 5. The polymorphism rates within Populations 1 and 2 were 75% and 60% respectively, suggesting that SNPs can be randomly selected for mapping purposes within these populations. It should be noted, however, that of 18 representatives of Population 2, twelve had some Population 1 admixture (<25%) and in the six pure Population 2 individuals, the polymorphism rate was 42%. The two parentals from each of Population 3 and 5 were genotyped, capturing the entire polymorphism rate within those lines (Population 3 at 10% and Population 5 at 4%). Note, however, that SNPs identified from Population 5 are likely to be useful for crosses generated between it and either Population 1 or 2.

We assessed linkage disequilibrium by calculating r^2 across the 200kb regions both within and across populations, using Haploview⁸. SNPs with a minor allele frequency of less than 0.04 were excluded and SNPs at all distances were subjected to association analysis. The markers analyzed vary among populations, as Haploview automatically removes markers that are monomorphic within the group of individuals. Markers with a call rate of less than 75% were excluded. Linkage disequilibrium ($r^2 > 0.5$) extended less than 15 kb across populations.

S6 Representation of segmental duplications in the assembly

To elucidate the effect of any bias in the analysis due to the draft nature of the opossum assembly, we first examined its overall read coverage landscape to try to locate trace “pile-ups”. Such pile-ups might suggest an over-collapsed region in the assembly that concealed segmental duplication. The largest “pile-ups” are due to the incorporation of mitochondrial sequence into the assembly. However, excluding these, even if one assumed that every over-sampled base was actually part of an over-collapsed duplication (surely a conservative upper bound), the overall duplication rate would be boosted by less than half (to < 2.6%)- leaving it well under the duplication rate for human.

We next analyzed all traces that were omitted from the assembly. ~10% of the 38.8 million reads left after low quality and contaminating reads were not placed in the assembly. This is a similar fraction to that seen in other draft assemblies. The “Improved” ARACHNE WGS assembly of the mouse genome (see below) left ~13% of reads unplaced, and the most recent dog assembly (canFam2) left ~6% unplaced. For comparison, assembly of the highly repetitive and polymorphic fungus *Puccinia graminis* left 30-40% of reads unplaced (E. Mauceli, unpublished data). The opossum genome is therefore unlikely to contain large, euchromatic segmentally duplicated regions that completely failed to assemble.

Finally, we examined a historical series of mouse assemblies to chart how our ability to discover segmental duplications changes with the quality of an assembly. The MGSCv3 assembly

of the mouse genome⁹ was primarily derived from whole genome shotgun traces, though it had a tiny amount of directed sequencing incorporated as well. This assembly was done with an older version of the ARACHNE software. An “Improved” assembly was done with the newer ARACHNE version that was used to assemble opossum, using the same whole genome shotgun traces as the MGSCv3 assembly, but omitting all directed sequencing data. Finally, we analyzed the most recent finished assembly available for mouse, mm8 (NCBI Build 36; February 2006), which benefits from several rounds of directed sequencing and finishing work. Improvements made to the ARACHNE assembler allow for better recognition of segmental duplications, even without any finishing. 5.3% of the finished mm8 mouse assembly is covered by segmental duplication, as compared to only 3.5% in the “Improved” assembly most comparable to the current opossum assembly. Even if there were a roughly similar inflation of segmental duplication coverage in a finished opossum assembly (from 1.7% in the current assembly to 2.6% in a speculative finished assembly), opossum would remain far less duplicated than either mouse or human.

S7 Synteny maps and assignment of breakpoints to lineages

We performed full genomic alignments of repeat masked sequence from the opossum genome (monDom5) against the human (hg18), dog (canFam2), mouse (mm8), rat (rn4) and chicken (galGal3) genomes using *PatternHunter*¹⁰.

We sought to align the opossum genome sequence with the human genome. We found that ~10% (338 Mb) of the nucleotides can be directly aligned to the human sequence. This is much lower than for eutherians such as mouse (40%) or dog (~57%), likely reflecting both actual turnover of genomic sequence and some difficulty in detecting neutrally evolving segments due to nucleotide divergence^{11,12}.

Following established methods^{5,9}, we identified collinear clusters of the identified synteny anchors, which were used to form larger syntenic segments in a hierarchical fashion. Segments that are larger than a given size in both genomes, and are comprised of at least 4 anchors at a given stage in the merging process, define a resolution-dependent, pair-wise synteny map between the two genomes.

We next merged and reconciled the pair-wise synteny maps to create a 5-way multi-species map. We identified the locations along the opossum genome (the reference genome) where any of the four pair-wise maps indicated a synteny breakpoint, and then, from the opossum locations, identified the corresponding locations in each of the four non-reference genomes. This allowed each genomic interval in opossum bracketed by synteny breakpoints to be associated with, and oriented

with respect to, a genomic interval in each of the four non-reference genomes. The resulting multi-species synteny map represent blocks of homologous sequence that are present and contiguous in all five species.

The multi-species synteny map can be represented as a graph: each block is represented by two nodes (- and +). Each pair of nodes have exactly one incoming and one outgoing edge for each species, which connect to the previous and next neighboring block edge in that species, respectively, as defined by their species-specific genomic ordering. This representation greatly simplifies the bookkeeping needed to relate breakpoints to evolutionary events.

At a resolution of 500 kb, our 5-way maps consist of 616 blocks, representing 1,232 synteny breakpoints. Most (75%) of these breakpoints are simple cases where the outgoing (or incoming) edges connect to only one or two neighboring blocks (chromosomal ends are not counted, so if the end of a block is the most distal segment of a chromosome in one or more species, there may be only one neighboring block). In these cases, synteny breakpoints can be unambiguously associated with a lineage, using the principle of parsimony. The remaining breakpoints connected to either three (19%) or more than three (6%) neighboring blocks. At higher resolutions, we observe more breakpoints (2,126 at 100 kb), but also a higher percentage of simple breakpoints (80%).

We next focused specifically on the subset of simple cases involving purely internal breakpoints, *i.e.* those that do not involve a chromosome end in any of the five species (and therefore exactly two neighboring blocks). These represent 65% of all breakpoints in the 500 kb maps and 74% in the 100 kb maps. This subset is clearly associated with translocations and inversions (but not chromosome fusions or fissions), and can be unambiguously assigned to branches in the phylogenetic tree. This subset provides perhaps the most reliable estimate of the relative rates of such rearrangements between the eutherian lineages at any given resolution. Our 5-way synteny map largely confirms the trends reported previously⁵ and, in particular, the observation that although the human and dog lineages have generally experienced a similar number of rearrangements, the human lineage appears to have experienced a greater number of smaller rearrangements. We also confirm that most of the larger rearrangements (≥ 500 kb) that have occurred in the mouse and rat lineages appear to have occurred in their last common ancestor. Finally, we verified that these trends are qualitatively unchanged when internal breakpoints with 3 alternative neighboring blocks are included (as multiple events).

S8 Reconstruction of the ancestral boreoeutherian karyotype

Much of the structure of the ancestral eutherian karyotype can be inferred from syntenic relationships among extant mammals. Given that most of the synteny breakpoints in our 5-way map were clearly lineage-specific, we reasoned that it would be possible to reconstruct genomic regions that were once contiguous in the boreoeutherian ancestor (ancestral to dog, human, mouse and rat) by joining syntenic blocks into larger groups by following their links in a hierarchical fashion that is consistent with the phylogenetic tree.

References for Supplementary Notes

1. Osoegawa, K. et al., *Genomics* 52 (1), 1-8 (1998).
2. Breen, M. et al., *BMC Genomics* 5 (1), 65 (2004).
3. Pathak, S. et al., *Cytogenet Cell Genet* 63 (3), 181-184 (1993).
4. Duke, S. E. et al., *Chromosome Res* in press (2007).
5. Lindblad-Toh, K. et al., *Nature* 438 (7069), 803-819 (2005).
6. Ning, Z., Cox, A. J., and Mullikin, J. C., *Genome Res* 11 (10), 1725-1729 (2001).
7. Sachidanandam, R. et al., *Nature* 409 (6822), 928-933 (2001).
8. Barrett, J. C., Fry, B., Maller, J., and Daly, M. J., *Bioinformatics* 21 (2), 263-265 (2005).
9. Waterston, R. H. et al., *Nature* 420 (6915), 520-562 (2002).
10. Ma, B., Tromp, J., and Li, M., *Bioinformatics* 18 (3), 440-445 (2002).
11. Margulies, E. H., Chen, C. W., and Green, E. D., *Trends Genet* 22 (4), 187-193 (2006).
12. Margulies, E. H. et al., *Proc Natl Acad Sci U S A* 102 (13), 4795-4800 (2005).
13. Birney, E. et al., *Nucleic Acids Res* 34 (Database issue), D556-561 (2006).
14. Birney, E., Clamp, M., and Durbin, R., *Genome Res* 14 (5), 988-995 (2004).
15. Slater, G. S. and Birney, E., *BMC Bioinformatics* 6, 31 (2005).
16. Goodstadt, L. and Ponting, C. P., *PLoS Comput Biol* 2 (9) (2006).
17. Goodstadt, L., Heger, A., Webber, C., and Ponting, C. P., *Genome Res* doi:10.1101/gr.6093907 (2007)
18. Altschul, S. F. et al., *J Mol Biol* 215 (3), 403-410 (1990).
19. Xie, X., Kamal, M., and Lander, E. S., *Proc Natl Acad Sci U S A* 103 (31), 11659-11664 (2006).
20. Kumar, S. and Hedges, S. B., *Nature* 392 (6679), 917-920 (1998).
21. Lee, M. S., *J Mol Evol* 49 (3), 385-391 (1999).
22. Phillips, M. J. and Penny, D., *Mol Phylogenet Evol* 28 (2), 171-185 (2003).
23. Springer, M. S., Murphy, W. J., Eizirik, E., and O'Brien, S. J., *Proc Natl Acad Sci U S A* 100 (3), 1056-1061 (2003).
24. Woodburne, M. O., Rich, T. H., and Springer, M. S., *Mol Phylogenet Evol* 28 (2), 360-385 (2003).

Chapter 5: Proteomic analysis of conserved non-coding elements

In this chapter, we describe experiments designed to identify proteins that bind in a sequence-specific manner to mammalian conserved non-coding sequences.

Parts of this work has been published in

Xie, X., Mikkelsen, T. S. *et al.* Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A* **104**, 7145-7150 (2007)

This publication is attached as Appendix 4. Supplementary data is available online from <http://www.pnas.org>

The bait design and affinity capture protocols were initially formulated by Andreas Gnirke. The CNE motif discovery analysis was performed by Xiaohui Xie.

[This page is intentionally left blank]

Conserved noncoding elements (CNEs) constitute the majority of sequences under purifying selection in the human genome, yet their function remains largely unknown. Experimental evidence suggests that many of these elements play regulatory roles, but little is known about regulatory motifs contained within them or the proteins that interact with them. Here, we describe hypothesis-generating experiments designed to identify proteins that bind specifically to CNEs or sequence motifs within them using *in vitro* DNA affinity capture and mass spectrometry.

Purification of proteins that recognize specific DNA sequences was first achieved by development of DNA affinity chromatography in the late 1980s¹⁻³. This led to the purification and subsequent cloning of abundant regulatory factors, such as Sp1²⁻⁴. Recent advances in mass spectrometry has made it feasible to directly identify captured proteins. In 2004 Himeda and colleagues successfully identified Six4 as a regulator of the muscle creatine kinase enhancer using quantitative mass spectrometry and follow-up screening⁵. Öztürk and colleagues similarly identified AP2γ as a rat placental lactogen II trophoblast cell-specific enhancer binding protein⁶.

We have developed a variation of the DNA affinity chromatography method that relies on biotinylated DNA baits, streptavidin-coated magnetic beads and liquid-chromatography mass spectrometry (LCMS; see Methods). Composition of the capture buffer was at first optimized to reliably isolate recombinant *cro* protein using a lambda phage operator sequence and to capture sequence-specific transcription factors from ~1mg of nuclear extract using a human cytomegalovirus (CMV) promoter sequence (Table 1; A. Gnirke, *unpublished*). Here, we describe two experiments that relied on this method to gain insights into the putative functions of conserved non-coding sequences.

Proteins binding to ultraconserved non-coding elements

Ultraconserved elements (UCEs) are > 200 bp sequence elements with 100% nucleotide similarity between human, mouse and rat⁷. Although they are clearly critical elements in the genome, little is known about their functions.

In order to obtain clues about the possible functions of UCEs, we arbitrarily selected 6 ultraconserved elements that do not overlap known transcripts in the human or mouse genomes (Table1). We generated ~200-400 bp double-stranded DNA baits containing the full sequence of these elements by genomic PCR with 5' biotinylated primers. We also generated a control bait pool

Table 1: Baits for DNA affinity capture using highly conserved non-coding elements

Element	Genome coordinates (hg18)	Primers
UC47	chr2:7692005-7692429	F: CTCGCATTAGACATGTGC R: GCTTGCCTCAGTCATAACC
UC191	chr6:51184719-51185009	F: AGGATGGATACAGAAACC R: CCAGCCTAGTTTCAGCGA
UC249	chr9:8085650-8086080	F: GTGCCTTCTAAGGTGGAC R: TGTTAACAGCCAACCTCTG
UC320	chr11:8274340-8274781	F: ACAGCGTCCTTACCCTCT R: TGGAGTGAATTTCCCAAAC
UC337	chr12:40035411-40035716	F: ATGGGCCCTACCCTTTTC R: GTATCAATGTTCTTGTTTTAC
UC405	chr16:58132770-58133103	F: GGGAGGGGTAAAACCTATTG R: AATAAGCCTTTAATAAAAATCAG
AT1	chr16:58133721-58134073	F: CCTTTATCTGCGTGGACCAT R: TGTGCTACTGCCACACATCA
AT2	chr16:58146205-58146475	F: GGGGTAAATTGTTCCCTGAAGC R: TGGTGAACAATTGAAAGACTGTG
AT3	chr13:55090168-55090564	F: AAGTTGCTCTGCTGCTAAGG R: TGAACCTTCAAGTCAGGAACC
AT4	chr13:55088735-55089255	F: TGTGGTAAACATAACTGCTGCT R: TTGCTCAATGCACAAAGTCA
AT5	chr13: 89192977-89193491	F: GCAACAATTCTGCAACTCTGA R: TTGGGAAGTTTCTCACTTGGA
AT6	chr3:22251022-22251544	F: ACTGAAAAGCATCCCACCAT R: CCTTAGTGGCTCCACTTGAAA
CMV promoter	N/A	F: CGGGGTCATTAGTTCATAG R: CCACCGTACACGCCTACCGCC

UC, ultraconserved; AT, non-conserved AT-rich; CMV cytomegalovirus

by adding 5' biotin to sheared total human genomic DNA. Affinity capture experiments in HeLa nuclear extract, using each bait pool and mass spectrometric identification of bound proteins, were performed as described in Methods.

Table 2 shows proteins for which at least two different peptides were identified in the eluate from at least one of the six UCE baits, but not in the eluate from either the sheared total DNA control or a no-bait control. Each of the UCE baits captured a distinct set of sequence-specific and non-specific DNA binding proteins. Several of these interactions are consistent with the known sequence specificity of individual factors and the sequence of the individual baits. For example, UC320 captured a protein corresponding to 'Ets variant gene 6', also known as ETV6 or TEL. This transcription factor is known to bind to the sequence motif [C/T]TTCC[T/G]⁸. UC320 contains three exact matches to this motif, while only 2 of the other 5 UCE baits contain a single match each.

All 6 UCE baits captured many peptides corresponding to the CHD4 protein, also known as Mi2- β . CHD4 is a 218 kD protein that belongs to the SNF2/RAD54 helicase-like family. It contains a helicase/ATPase domain, a putative chromatin-binding chromodomain, and two PHD zinc finger domains thought to be involved in protein-protein interactions⁹. It is the major component of the abundant Mi2/NURD protein complex, which contains repressive chromatin remodeling and histone deacetylase activities¹⁰ (we note that none of the other subunits of the complex were identified in the experiment). Two interchangeable components of the Mi2/NURD complex, MBD2 and MBD3, appear to mediate recruitment to cytosine-methylated and sequence-specific targets, respectively. Notably, mouse embryonic stem (ES) cells lacking MBD3 show deficiencies in lineage-commitment when induced to differentiate^{11,12}. Because UCEs appear to be associated with the regulation of key developmental regulatory genes, we decided to further characterize the CHD4/UCE interaction.

In order to confirm the CHD4/UCE interaction we repeated the affinity capture experiment and probed resulting eluates with anti-CHD4 antibodies (Figure 1a). Because we noted that the UCE baits were relatively AT-rich (range: 59-72%) compared to the human genome sequence (mean: 58%), we also generated 6 control baits with similar AT-content from non-conserved regions of the human genome to rule out non-specific binding to AT-rich sequences (Table 1). Western blotting showed the presence of CHD4 in the HeLa nuclear extract and strong, specific bands from 3 out of 6 UCE baits (UC191, UC337, UC405). Weaker bands were also visible in eluates from the other 3 UCE baits, but only from 1 of the 6 AT-rich control baits (AT3), and not from the sheared total human DNA bait pool.

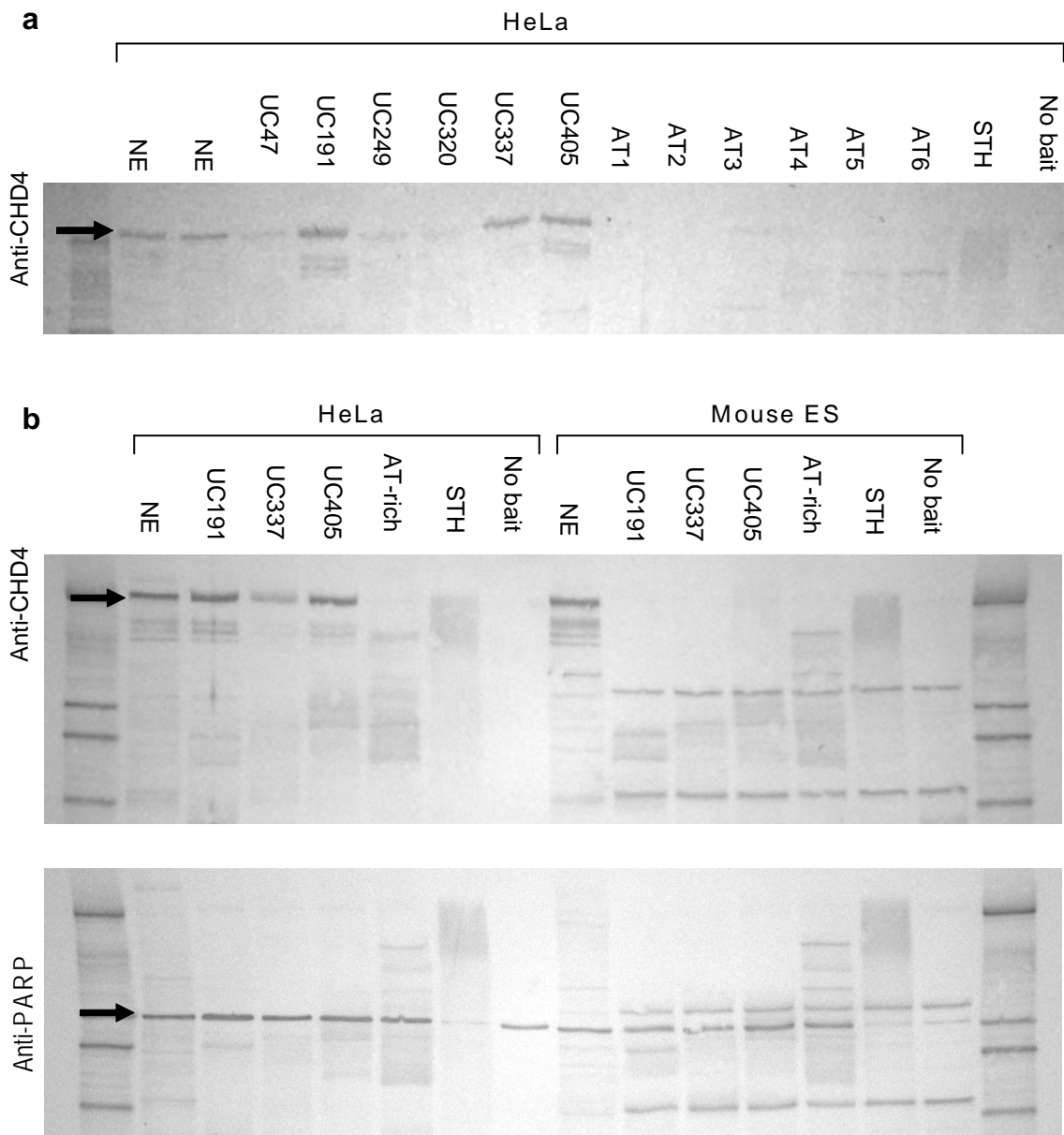


Figure 1. Sequence- and context-specific capture of CHD4 by ultraconserved elements in vitro. a, Anti-CHD4 Western blot in HeLa nuclear extract (NE) and eluates from affinity capture using ultraconserved (UC), non-conserved AT-rich (AT) or sheared total human DNA (STH) baits in HeLa NE. UC191, UC337 and UC405 show strong bands, and UC47, UC249 and UC320 show weak bands consistent with specific CHD4 capture. b, CHD4 is detected in NE from HeLa and mouse embryonic stem (ES) cells, but capture occurs only in HeLa NE, suggesting that a secondary factor is involved. Western blot of a non-specific protein (PARP) is shown for comparison. AT1-6 were pooled into one AT-rich control.

Table 2: Proteins captured by UCE/CMV baits in HeLa nuclear extract

UC47	UC191	UC249	UC320	UC337	UC405	CMV	Unique Peptides	Annotation
++		+				++	7	CCAAT/enhancer binding protein beta
++	+++	+	+	++	+		54	CHD4 protein
++						+++	8	Transcription factor ATF-a
+	++			+			3	Unnamed protein product
	+++		+	+++	++		47	AT motif-binding factor 1
	++	+++	+		+		24	BTB and CNC homology 1
	++	++	++				9	Purine-rich element binding protein B
	++	+					5	MAFF Transcription Factor
	++		+	++	++		6	AT-rich interactive domain-containing protein 3A
	++		++	++	+		3	LIM domain binding 1
	++			+	+		5	T-box 3 protein isoform 1
	+			+	+		3	POU2F1 protein
	+						2	Hypothetical protein
	+						2	Unnamed protein product
		++++		++	++		84	CUTL1 protein
		++				+++	19	Regulatory factor X1
		+			+		4	Replication protein A1, 70kDa
		+	++				2	Unnamed protein product
		+	+				2	Novel protein similar to annexin A2
			++		++		6	RNA-binding protein AUF1
			++				5	Ets variant gene 6
			++				3	Zinc finger protein 148
			++				3	Gag-Pro-Pol protein
			++				2	Nucleolar phosphoprotein p130
			+	+			4	GTF2I
			+				2	LGALS7 protein
			+				2	Homeobox protein A10
				++		+++	14	Nuclear factor 1 C-type
				+			2	ISL2 transcription factor
					++		5	Splicing factor proline/glutamine rich
					+		10	Transcription factor AP-2 alpha
						++	10	Nuclear receptor subfamily 2, group F, member 1
						++	9	Orphan nuclear receptor TR4
						++	7	Nuclear factor, interleukin 3 regulated
						++	6	Jun B proto-oncogene
						++	4	v-fos
						++	3	CREB
						++	3	FOS-like antigen 2
						++	3	CCCTC-binding factor (zinc finger protein)
						++	2	C-Rel proto-oncogene

+,++,+++ indicates three (qualitative) degrees of protein abundance, from weak to strong.

In order to determine whether CHD4 capture was context-specific, we repeated the affinity capture experiment using three UCE baits, a pooled AT-rich control (AT1-AT6) and sheared total human DNA in nuclear extracts from both HeLa cells and mouse ES cells (Figure 1b). We again detected specific capture of CHD4 by the UCE baits from HeLa extracts. In contrast, while CHD4 could be detected in mouse ES cell nuclear extract, it was not captured by any of the baits.

We conclude that the chromatin remodeling protein CHD4 can be recruited to ultraconserved non-coding elements in a sequence- and context-specific manner. Due to the lack of specific DNA binding domains in CHD4 itself, we speculate that the recruitment is mediated by a second factor that may not be present in mouse ES cells. Additional studies will be required to understand the relevance of these findings to the functions of UCEs *in vivo*.









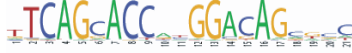

Proteins binding to conserved sequence motifs

It seems likely that conserved non-coding elements spanning all levels of constraint are involved in gene regulation, and transgenic experiments have identified some CNEs that are capable of driving highly specific spatiotemporal gene expression patterns¹³⁻¹⁶. However, little is known about regulatory motifs contained within CNEs or proteins that recognize these motifs.

We have used the recent availability of sequences of 12 mammalian genomes to search for motifs that are enriched in CNE regions relative to the rest of the genome¹⁷. We focused specifically on long regulatory motifs, between 12 and 22 bp, which provide a strong signal for motif discovery. We began by identifying k -mers (for $k \geq 12$) that occur at a significantly higher frequency in the CNE sequences than in the remainder of the genome. We focused only on relatively long k -mers, because the expected number N of random occurrences in the entire CNE database is small (for example, $n < 8$ for $k = 12$). We identified a total of 69,810 enriched k -mers. An example is 5'-GTTGCCATGGAAAC-3', which appears 698 times in the CNE data set, whereas only 27 sites are expected based on its genome-wide frequency (26-fold enrichment). We noticed that many of the enriched k -mers were closely related; therefore, we clustered them based on sequence similarity. The 69,810 enriched k -mers collapsed into 233 distinct groups, denoted LM1, LM2, etc. for "long motif." For each of these motifs we derived a positional weight matrix (PWM) representation reflecting the distribution of 4 nucleotides at each position. The enrichment of each motif in the CNE data set was expressed in an enrichment score. The top 10 motifs are shown in Table 3.

For each of the 233 discovered motifs, we searched the entire human genome to identify conserved instances; that is, we identified all human sites matching the PWM and then found those

Table 3: Properties of top 10 motifs discovered in conserved non-coding sequences

ID	Motif profile	No. of conserved instances	False positive rate*	Conservation rate, † %	Fold increase in conservation rate ‡	Correlation between cross-species conservation and motif profile	Positional bias around TSS [§]
LM1		5,332	0.050	29.3	9.5	0.92	
LM2		7,549	0.048	29.4	14.0	0.91	
LM3		844	0.048	40.1	14.3	0.94	
LM4		1,877	0.046	20.3	13.5	0.89	20.3
LM5		224	0.042	19.4	16.3	0.87	
LM6		79	0.026	20.1	10.1	0.81	25.5
LM7		6,302	0.048	21.6	10.3	0.72	
LM8		608	0.047	17.2	9.6	0.68	
LM9		1,443	0.039	11.8	8.4	0.90	6.1
LM10		5,914	0.050	14.5	6.6	0.77	

*The proportion of conserved instances expected to have occurred by chance.

†The proportion of instances detected in human that are also conserved in orthologous regions of other mammals.

‡Compared to the conservation rates of control motifs.

§Fold enrichment on the number of motif sites located within 1 kb of TSS over those for control motifs. Only motifs with fold enrichment above 4 are shown.

sites that show clear cross-species conservation. We found a total of 60,019 conserved instances, with roughly half residing within the CNE data set and roughly half in the remainder of the genome. Importantly, the approach of focusing on motifs enriched in the CNE data set identified many motif instances elsewhere in the genome.

The number of conserved instances is highly uneven across the motifs (range 37-7,549, with mean of 266 and median of 61). Most motifs (67%) have <100 conserved instances. But, remarkably, the two motifs with the highest enrichment scores, LM1 and LM2, both have >5,000 conserved instances in the human genome, suggesting a widespread functional role for these elements. We therefore focused on characterizing these two motifs.

LM1 defines RFX1 binding sites. The most highly enriched motif LM1 is similar to the X-box motif, which has been extensively studied in yeast and nematodes^{18,19}. In yeast, more than three dozen X-box sites have been identified, and these sites have been shown to be bound by the Crt1 protein, an effector of the DNA damage checkpoint pathway²⁰. In *Caenorhabditis elegans*, >700 X-box sites have been computationally predicted, and several dozen of these sites have been demonstrated to be recognized by the DAF-19 protein, which is known to regulate genes involved in the development of sensory cilia¹⁸.

Individual instances of the X-box motif in vertebrates have been reported, but no systematic survey of X-box motifs in the human genome has been conducted. Approximately three dozen such sites have been reported to be bound by RFX family proteins, which are homologous to both Crt1 and DAF-19 and contain a highly conserved winged helix DNA binding domain. The biochemically characterized consensus sequence for RFX binding shows similarity to the LM1 motif²¹, although it contains less information.

To identify proteins that bind LM1, we performed an affinity capture experiment. A biotinylated double-stranded DNA probe containing multiple copies of the LM1 motif and a degenerate control bait (see Methods) were incubated with HeLa nuclear extract. Table 4 shows the proteins for which at least 2 peptides were identified specifically in the LM1 eluate. RFX1 (Regulatory factor X1) was the only known sequence-specific transcription factor captured in each of two independent experiments. Western blots with anti-RFX1 on an independent eluate confirmed that the protein indeed specifically binds LM1 (Figure 2a).

LM2 defines a common insulator site across the human genome. The most interesting case among the 233 discovered motifs is LM2. It has the largest number of conserved instances

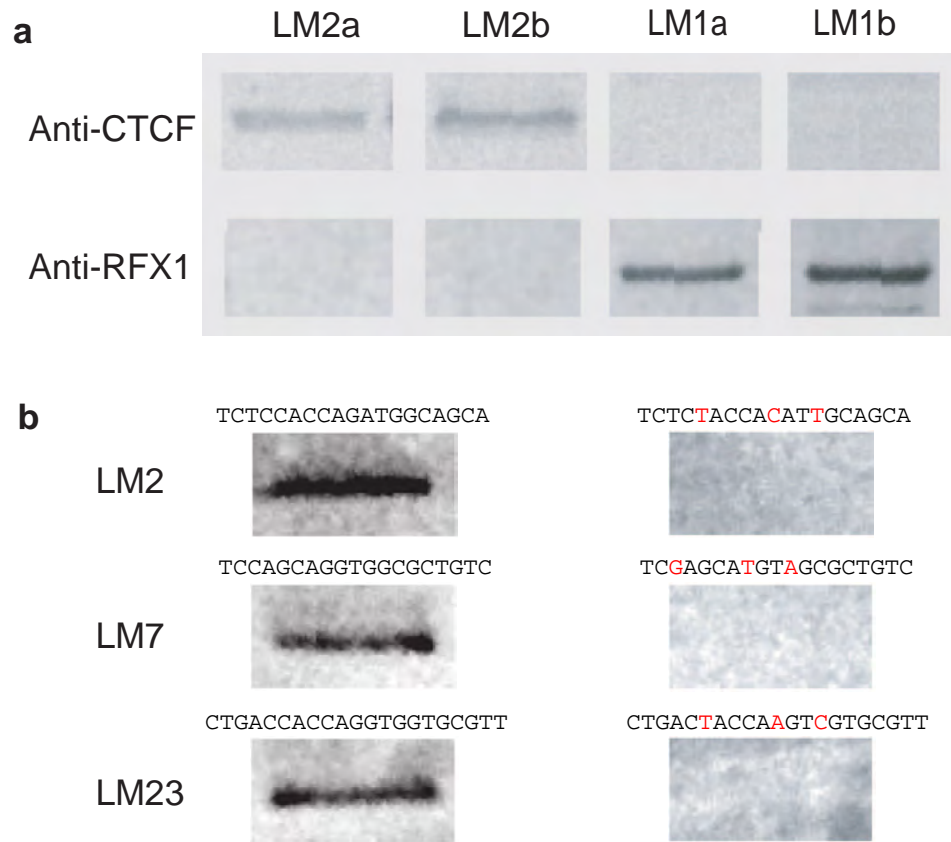


Figure 2. Confirmation of CTCF and RFX1 binding by in vitro affinity capture. a, CTCF was specifically captured by probes LM2a and LM2b constructed for the LM2 motif, whereas RFX1 was specifically captured by probes LM1a and LM1b constructed for the LM1 motif. b, The binding of CTCF to LM2, LM7 and LM23 (left), but not to their corresponding mutant motifs with three core bases altered (right).

(7,549) in the genome, with the vast majority being located far from TSSs. The LM2 motif is 19 bases in length and does not match the reported consensus sequence of any known motif.

To obtain a hint regarding the possible function of the LM2 motif, we again performed an affinity capture experiment. A biotinylated double-stranded DNA probe containing multiple copies of the LM2 motif and a degenerate control bait (see Methods) were incubated with HeLa nuclear extract. Table 4 shows the proteins for which at least 2 peptides were identified specifically in the LM2 eluate. CTCF (CCCTC-binding factor) was the only known sequence-specific transcription factor captured in each of two independent experiments.

CTCF, a protein containing 11 zinc-finger domains, is a major factor implicated in vertebrate insulator activities²²⁻²⁴. An insulator is a DNA sequence element that prevents a regulatory protein binding to the control region of one gene from influencing the transcription of neighboring genes. When placed between an enhancer and a promoter, an insulator can block the interaction between the two. Several dozen insulator sites have been characterized, and almost all have been shown to contain CTCF binding sites. In some cases, the CTCF site has been directly shown to be both necessary and sufficient for enhancer blocking activities in heterologous settings. The known CTCF sites show considerable sequence variation, and no clear consensus sequence has been derived²³. The well studied CTCF sites in the IGF2/H19 locus show similarity to the LM2 motif²⁵, although the similarity score is below our threshold used for detecting LM2 sites.

To confirm that CTCF binds the LM2 motif we repeated the affinity capture experiment and probed the captured material with anti-CTCF antibodies. The results confirmed that CTCF does indeed bind the LM2 motif (Figure 2a). By contrast, mutation of the three core positions with the highest information content (positions 5, 10, and 13 of LM2; Figure 2b) completely abolished the binding of the CTCF protein.

Given the sequence diversity among reported CTCF sites, we searched for additional motifs in our catalog that show substantial similarity to LM2. The motifs LM7 and LM23 are nearly identical in their first 14 positions, diverging only in the last four or five bases. The two additional motifs also have an unusually large number of conserved instances (6,302 for LM7 and 3,758 for LM23). Affinity-capture experiments using probes containing copies of the LM7 and LM23 motifs demonstrated that both motifs bind CTCF, whereas mutation of the three core positions with the highest information content completely abolish binding (Figure 2b). The three motifs, LM2, LM7, and LM23, will be referred to as a “supermotif,” LM2*.

Table 3: Proteins captured by LM motif baits in HeLa nuclear extract in two independent affinity capture experiments

LM1 (1)	LM1 (2)	LM2 (1)	LM2 (2)	Unique Peptides	Annotation
++++	++	++++		25	Unknown protein
+++		+++		27	GTBP-N protein
		+++	++	23	Bloom syndrome protein
++		+++	++	16	SMARCA3
++++	++	+++	+	16	RecQ protein-like isoform 1
+++	++	+++		17	XRCC1
++++	++			12	Regulatory factor X1
		+++		13	Topoisomerase (DNA) III alpha
+++	++	++	++	11	DNA topoisomerase
	++		+	10	Heterogenous nuclear ribonucleoprotein U
++		+++		8	C9orf76
+++				7	Polynucleotide kinase-3'-phosphatase
		++		5	KIAA1596
++				5	Werner syndrome protein
		+++		5	DNA-directed RNA polymerase III largest subunit
		++	+	4	Dermcidin
		+++	++	4	CCCTC-binding factor (zinc finger protein)
				4	Heterogeneous nuclear ribonucleoprotein D,
		+++		3	TPA: keratin 1b
			+	2	polymerase (RNA) III (DNA directed) polypeptide B

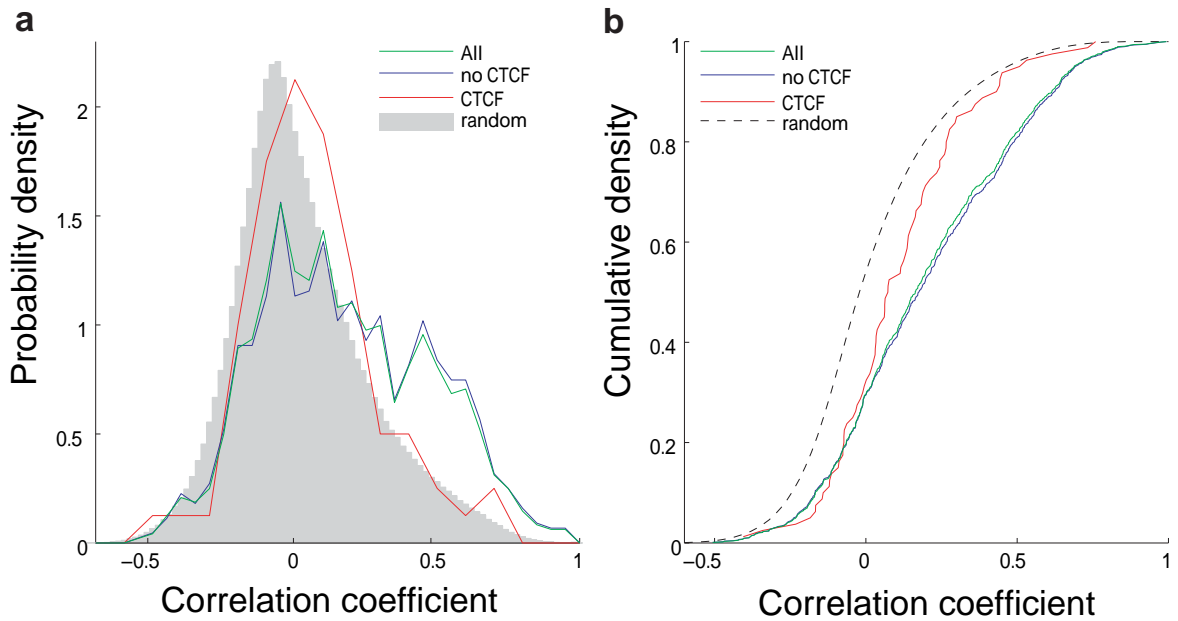


Figure 3. Genes separated by predicted CTCF sites are less correlated in gene expression. Correlation coefficients between neighboring gene pairs is shown in terms of a probability density (a) and a cumulative distribution (b). Green line, correlation between all neighboring genes; red line, correlation between genes separated by at least one CTCF site; gray shading, correlation between randomly chosen gene pairs.

Altogether the LM2* motif has 14,987 conserved instances in the human genome (which is 20-fold higher than for the corresponding control motifs). Strikingly, this comprises approximately one-fourth of the 60,019 sites for the complete catalog of 233 motifs. We propose that the vast majority of these sites are CTCF-binding sites and function as insulators.

Although the predicted CTCF sites tend to be located far from gene starts, they are not randomly distributed across the genome. Instead, their distribution closely follows the distribution of genes, with a correlation coefficient of 0.6.

We sought to test whether the predicted CTCF sites actually serve as functional insulators. Although it is possible to perform insulator assays on individual instances in a heterologous context, we were interested to assess the function of many CTCF sites in their natural context. If the predicted CTCF sites actually function as insulators, we reasoned that the presence of a CTCF site between two genes might “decouple” their gene expression.

It is known that divergent gene pairs, transcribed in opposite directions with transcription start sites close to each other, tend to show correlated gene expression patterns^{26,27}. We therefore assembled a data set of 963 divergent gene pairs with intergene distance <20 kb and with expression values measured across 75 human tissues²⁸. As expected, the divergent gene pairs are more closely correlated in gene expression than randomly chosen gene pairs (Figure 3). When the cases are divided into gene pairs separated by a CTCF site (CTCF pairs, 80 cases) and those not separated by a CTCF site (non-CTCF pairs, 883 cases), the former show correlations that are essentially equivalent to the random background. Overall, 37% of non-CTCF pairs are strongly correlated (correlation coefficient $\rho > 0.3$). This proportion is 2-fold higher than the proportion of random genes pairs (12%) showing similarly strong correlation. By contrast, the proportion of CTCF pairs with similarly strong correlation is 16%, which is close to that seen for random gene pairs. This difference persists after correcting for small difference in the lengths of CTCF-containing and CTCF-non-containing intergenic regions. This provides strong evidence that the majority of the predicted CTCF sites do indeed function as insulators.

Discussion

Our results show that DNA affinity capture followed by mass spectrometry can be an effective approach to hypothesis-generating identification of proteins that interact with conserved non-coding elements or motifs. We have identified sequence- and context-specific interactions between six UCEs and CHD4, a component of the Mi2/NURD chromatin remodeling complex, suggesting a

possible role for UCEs in epigenetic regulation. Moreover, we have identified thousands of conserved CTCF binding sites with putative insulator functions in the human genome.

The sensitivity and versatility of this method can be expected to increase as new proteomic technologies are developed and adapted. For example, performing affinity capture experiments in nuclear extracts obtained from stable isotope labeling of amino acids in culture (SILAC)²⁹ should yield a more quantitative representation of differential capture by target and control baits.

Methods

CNE motif analysis. Performed as described in ¹⁷.

Baits. UCE and AT-rich baits were generated by genomic PCR from total human DNA using 5'biotinylated primers (Table 1; Operon, Huntsville, AL). Sheared total human DNA control baits were generated by mechanical shearing, size-selection to 200-400 bp, end-repair and 5' biotinylation. CNE motif baits were generated by synthesizing, annealing and extending two 5'biotinylated oligos containing two instances of the motif and an overlapping M13 F site (Operon, Huntsville, AL). Probe sequences were as follows: LM1, TGTTGCTTAGCAACA; LM2, CCACTAGATGGCAGTGTT; degenerate control, NNNNNNNNNNNN (each position in each molecule contains a random nucleotide). For Western blot validation, probe sequences are as follows: LM1a, GCTGTTGCCATGGAAACCAG; LM1b, TGTTGCTTAGCAACA; LM2a, CCACCAGGTGGCAGCAGA; LM2b, CCACTAGATGGCAGTGTT. For testing the binding of CTCF to LM2, LM7, and LM23 and their mutants we used the following probes: LM2, TCTCCACCAGATGGCAGCA; LM2 mutant, TCTCTACCACATTGCAGCA; LM7, TCCAGCAGGTGGCGCTGTC; LM7 mutant, TCGAGCATGTAGCGCTGTC; LM23, CTGACCACCAGGTGGTGCTGTT; LM23 mutant, CTGACTACCAAGTCGTGCTGTT.

Affinity capture. In all experiments, 1 pmol of 5'-biotinylated bait was mixed with 1.4 mg of HeLa (Promega, Madison, WI) or mouse ES cell nuclear extract in binding buffer (50 mM Tris pH 8, 150 mM NaCl, 0.25 mM EDTA, 0.5 mM DTT, 0.1% Tween-20, 0.5 mg/mL BSA, 200 µg/mL poly-dI/dC) in a total volume of 700 µl. Mixtures were incubated for 1 hour at room temperature on a rotator. Ten µl of Streptavidin-coated magnetic beads (Dynal, Carlsbad, CA) was added to each tube, and the mixture was further incubated for 30 min at room temperature on a rotator. Beads were spun for 5 s at 1000g and then captured using a magnetic pull-down system. Beads were washed 3 X 700 µL in binding buffer without poly-dI/dC and then 4 X 700 µl in binding buffer without poly-dI/dC or BSA. The supernatant was discarded and proteins were eluted by boiling in Laemmli buffer (+10 mM DTT).

Mass spectrometry. Proteins were purified by non-resolving SDS-PAGE. The gel band between the loading well and the visible buffer/oligonucleotide fraction was cut out, reduced, alkylated, digested with trypsin according to standard proteomics practices ³⁰, and the resulting peptides were analyzed by LCMS on an Orbitrap (ThermoFisher) mass spectrometer as described in ³¹. Database searching was performed by the Mascot software package against the REFSEQ database of human

or mouse proteins. Reported hits were required to have at least 2 distinct peptide matches in the target bait eluate and none from the control or no-bait eluates.

Immunodetection. Captured proteins were eluted by boiling in XT reducing loading buffer (BioRad), separated by SDS/PAGE, and assayed by colorimetric Western blots using polyclonal antibodies (Santa Cruz Biotechnology, Santa Cruz, CA) for CHD4 (sc-11378), RFXI (sc-10652) or CTCF (sc-5916).

References

1. Rosenfeld, P.J. & Kelly, T.J., Purification of nuclear factor I by DNA recognition site affinity chromatography. *J Biol Chem* 261 (3), 1398-1408 (1986).
2. Kadonaga, J.T. & Tjian, R., Affinity purification of sequence-specific DNA binding proteins. *Proc Natl Acad Sci U S A* 83 (16), 5889-5893 (1986).
3. Wu, C. *et al.*, Purification and properties of Drosophila heat shock activator protein. *Science* 238 (4831), 1247-1253 (1987).
4. Kadonaga, J.T., Carner, K.R., Masiarz, F.R., & Tjian, R., Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. *Cell* 51 (6), 1079-1090 (1987).
5. Himeda, C.L. *et al.*, Quantitative proteomic identification of six4 as the trex-binding factor in the muscle creatine kinase enhancer. *Mol Cell Biol* 24 (5), 2132-2143 (2004).
6. Ozturk, A., Donald, L.J., Li, L., Duckworth, H.W., & Duckworth, M.L., Proteomic identification of AP2 gamma as a rat placental lactogen II trophoblast cell-specific enhancer binding protein. *Endocrinology* 147 (9), 4319-4329 (2006).
7. Bejerano, G. *et al.*, Ultraconserved elements in the human genome. *Science* 304 (5675), 1321-1325 (2004).
8. Matys, V. *et al.*, TRANSFAC and its module TRANSCmpel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34 (Database issue), D108-110 (2006).
9. Woodage, T., Basrai, M.A., Baxevanis, A.D., Hieter, P., & Collins, F.S., Characterization of the CHD family of proteins. *Proc Natl Acad Sci U S A* 94 (21), 11472-11477 (1997).
10. Denslow, S.A. & Wade, P.A., The human Mi-2/NuRD complex and gene regulation. *Oncogene* 26 (37), 5433-5438 (2007).
11. Kaji, K. *et al.*, The NuRD component Mbd3 is required for pluripotency of embryonic stem cells. *Nat Cell Biol* 8 (3), 285-292 (2006).
12. Kaji, K., Nichols, J., & Hendrich, B., Mbd3, a component of the NuRD co-repressor complex, is required for development of pluripotent cells. *Development* 134 (6), 1123-1132 (2007).
13. Pennacchio, L.A. *et al.*, In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444 (7118), 499-502 (2006).
14. Bejerano, G. *et al.*, A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441 (7089), 87-90 (2006).
15. Woolfe, A. *et al.*, Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3 (1), e7 (2005).
16. Dermitzakis, E.T., Reymond, A., & Antonarakis, S.E., Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet* 6 (2), 151-157 (2005).
17. Xie, X. *et al.*, Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A* 104 (17), 7145-7150 (2007).
18. Blacque, O.E. *et al.*, Functional genomics of the cilium, a sensory organelle. *Curr Biol* 15 (10), 935-941 (2005).
19. Zaim, J., Speina, E., & Kierzek, A.M., Identification of new genes regulated by the Crt1 transcription factor, an effector of the DNA damage checkpoint pathway in *Saccharomyces cerevisiae*. *J Biol Chem* 280 (1), 28-37 (2005).
20. Huang, M., Zhou, Z., & Elledge, S.J., The DNA replication and damage checkpoint pathways induce transcription by inhibition of the Crt1 repressor. *Cell* 94 (5), 595-605 (1998).

21. Emery, P. *et al.*, A consensus motif in the RFX DNA binding domain and binding domain mutants with altered specificity. *Mol Cell Biol* 16 (8), 4486-4494 (1996).
22. Bell, A.C., West, A.G., & Felsenfeld, G., The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* 98 (3), 387-396 (1999).
23. Ohlsson, R., Renkawitz, R., & Lobanenkov, V., CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* 17 (9), 520-527 (2001).
24. Gaszner, M. & Felsenfeld, G., Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 7 (9), 703-713 (2006).
25. Bell, A.C. & Felsenfeld, G., Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* 405 (6785), 482-485 (2000).
26. Trinklein, N.D. *et al.*, An abundance of bidirectional promoters in the human genome. *Genome Res* 14 (1), 62-66 (2004).
27. Li, Y.Y. *et al.*, Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comput Biol* 2 (7), e74 (2006).
28. Su, A.I. *et al.*, A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 101 (16), 6062-6067 (2004).
29. Ong, S.E. *et al.*, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1 (5), 376-386 (2002).
30. Kinter, M. & Sherman, N.E., *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, 1st ed. (Wiley, Chicester, UK, 2000).
31. Jaffe, J.D. *et al.*, Accurate inclusion mass screening: a bridge from unbiased discovery to targeted assay development for biomarker verification. *Mol Cell Proteomics* 7 (10), 1952-1962 (2008).

Chapter 6: Maps of histone methylation at key developmental loci

In this chapter, we describe the first ChIP-chip analysis of histone H3 lysine 4 and 27 trimethylation across developmental loci in pluripotent and differentiated cells.

This work was first published as part of

Bernstein, B. E., Mikkelsen, T. S. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315-326 (2006).

The full publication is attached as Appendix 5. Supplementary data is available online from <http://www.cell.com>

The text in this chapter was primarily written by Bradley E. Bernstein. Contributions from this thesis work includes experimental design and selection of target regions, analysis of H3K4me3/H3K27me3 overlap with transcription factors, CpG islands and transposable elements, and analysis of correlations between expression levels and chromatin state.

[This page is intentionally left blank]

The most highly conserved non-coding elements (HCNEs) in mammalian genomes cluster within regions enriched for genes encoding developmentally important transcription factors (TFs). This suggests that HCNE-rich regions may contain key regulatory controls involved in development. We explored this by examining histone methylation patterns in mouse embryonic stem (ES) cells across 56 large HCNE-rich loci. Here, we report the discovery of a novel chromatin modification pattern, termed ‘bivalent domains’, consisting of large regions of H3 lysine 27 methylation harboring smaller regions of H3 lysine 4 methylation. In ES cells, bivalent domains tend to coincide with TF genes expressed at low levels. Few bivalent domains are retained in differentiated cells. We propose that bivalent domains silence developmental genes in ES cells while keeping them poised to be activated upon differentiation. We also found a striking correspondence between genome sequence and histone methylation patterns in ES cells, which becomes notably weaker in differentiated cells. Together, these results highlight the importance of DNA sequence in defining the initial epigenetic landscape, and suggest a novel chromatin-based mechanism for maintaining pluripotency.

Epigenetic regulation of gene expression is mediated in part by post-translational modifications of histone proteins, which in turn modulate chromatin structure (Jenuwein and Allis, 2001; Margueron et al., 2005). The core histones H2A, H2B, H3 and H4 are subject to dozens of different modifications, including acetylation, methylation and phosphorylation. Histone H3 lysine 4 and lysine 27 tri-methylation (H3K4me3/H3K27me3) are of particular interest as these modifications are catalyzed, respectively, by trithorax- and polycomb-group proteins, which mediate mitotic inheritance of lineage-specific gene expression programs and have key developmental functions (Ringrose and Paro, 2004). H3K4me3 positively regulates transcription by recruiting nucleosome remodeling enzymes and histone acetylases (Pray-Grant et al., 2005; Santos-Rosa et al., 2003; Sims et al., 2005; Wysocka et al., 2005), while H3K27me3 negatively regulates transcription by promoting a compact chromatin structure (Francis et al., 2004; Ringrose et al., 2004).

Various observations suggest that chromatin undergoes important alterations during mammalian development (Delaval and Feil, 2004; Margueron et al., 2005; Sado and Ferguson-Smith, 2005). Embryonic stem (ES) cell differentiation is accompanied by changes in chromatin accessibility at several key developmental genes, including a large-scale opening of the HoxB locus (Chambeyron and Bickmore, 2004; Perry et al., 2004). Furthermore, polycomb-group proteins play an essential role in maintaining the pluripotent state of ES cells and show markedly reduced

expression upon differentiation (O'Carroll et al., 2001; Silva et al., 2003; Valk-Lingbeek et al., 2004). However, little is known about the overall structure of ES cell chromatin, how it is established, or how it contributes to the maintenance of pluripotency (Szutorisz and Dillon, 2005).

Large-scale studies of mammalian chromatin have recently become possible with the combination of chromatin immunoprecipitation (ChIP) and DNA microarrays. Initial studies in primary fibroblasts revealed thousands of genomic sites associated with H3K4me3 (Bernstein et al., 2005; Kim et al., 2005). The vast majority show a 'punctate' pattern, typically occurring at sites of ~1-2 kb near promoters of active genes. H3K27me3 is implicated in X-chromosome inactivation and imprinting (Plath et al., 2003; Umlauf et al., 2004). However, little is known about the overall genomic distribution of this repressive mark. Gene-specific and limited microarray studies have reported that H3K27me3 tends to occur at punctate sites near promoters of repressed genes (Cao and Zhang, 2004; Kimura et al., 2004; Kirmizis et al., 2004; Koyanagi et al., 2005). Based on such studies, the distributions of H3K4me3 and H3K27me3 have been thought to be non-overlapping.

A notable exception to the punctate pattern of histone modifications is evident at the Hox gene clusters: these loci contain large, cell type-specific H3K4me3 regions, up to 60 kb in length, that overlay multiple Hox genes (Bernstein et al., 2005; Guenther et al., 2005). These regions likely reflect accessible chromatin domains established during embryonic development to maintain Hox gene expression programs (Chambeyron and Bickmore, 2004). However, the extent to which large domains of chromatin modifications represent a general feature of mammalian genomes remains unclear.

Recent studies have revealed that the most highly conserved non-coding elements (HCNEs) in mammalian genomes cluster within ~200 HCNE-rich genomic loci, which include all four Hox clusters (Bejerano et al., 2004; Lindblad-Toh et al., 2005; Nobrega et al., 2003; Woolfe et al., 2005). These regions tend to be gene-poor, but are highly enriched for genes encoding transcription factors (TFs) implicated in embryonic development.

These findings suggested that the HCNE-rich regions or the TF genes within them may contain key epigenetic regulatory controls involved in development. We explored this by mapping histone methylation patterns in mouse ES cells across 61 large regions (~2.5% of the genome). The results reveal a novel chromatin modification pattern that we term 'bivalent domains', consisting of large regions of H3K27me3 harboring smaller regions of H3K4me3. In ES cells, bivalent domains frequently overlay developmental TF genes expressed at very low levels. Bivalent domains tend to resolve during ES cell differentiation and, in differentiated cells, developmental genes are typically marked by broad regions selectively enriched for either H3K27me3 or H3K4me3. We suggest that

bivalent domains silence developmental genes in ES cells while keeping them poised for activation. Finally, we analyzed the relationship between histone methylation patterns and the underlying DNA sequence in both ES and differentiated cells. This analysis suggests that DNA sequence largely defines the initial epigenetic state in ES cells, which is subsequently altered upon differentiation, presumably in response to lineage-specific gene expression programs and environmental cues.

Bivalent domains in ES cells contain repressive and activating histone modifications

H3K4me3 and H3K27me3 patterns in ES cells were examined across a subset of HCNE-rich loci using a combination of ChIP and tiling oligonucleotide arrays. The arrays tile 61 large genomic regions, totaling 60.3 Mb, at a density of approximately one probe per 30 bases. The regions consist of the four Hox clusters (1.3 Mb encoding 43 genes), 52 additional HCNE-rich regions (55 Mb encoding 169 genes) and five ‘control’ regions that do not show high HCNE density (4 Mb encoding 95 genes). We isolated genomic DNA associated with either H3K4me3 or H3K27me3 by immunoprecipitating cross-linked chromatin, and hybridized these DNA fractions to the tiling arrays (see Methods). We identified regions of H3K4me3 or H3K27me3 by comparing these hybridization results to those obtained for total genomic DNA. Experiments were performed in duplicate and analyzed using previously validated criteria (Bernstein et al., 2005). The resulting maps of ES cell chromatin (Figure 1) show a number of important features, many of which were unexpected.

We found a total of 343 sites of H3K4me3, ranging in size from 1 kb to 14 kb with a median size of 3.4 kb. Of these, 63% correspond to transcription start sites (TSSs) of known genes. Conversely, 80% of the TSSs are covered by H3K4me3 sites. Because H3K4me3 sites and TSSs each cover only ~2% of the genomic regions, this concordance is highly significant. We and others have previously noted a global concordance between H3K4me3 methylated sites and TSSs in differentiated mammalian cells (Bernstein et al., 2005; Kim et al., 2005).

We also found 192 H3K27me3 sites across these regions. These tend to affect much larger genomic regions than the H3K4me3 sites. The median H3K27me3 site is smaller in the control regions (5 kb), but is twice as large in the HCNE-rich regions (10 kb) and still larger in the Hox regions (18 kb). Overall, 75% of H3K27me3 sites are larger than 5 kb. We will refer to these large regions as ‘H3K27me3 domains’. There are 123 in the HCNE regions, 14 in the Hox regions and 7 in the control regions.

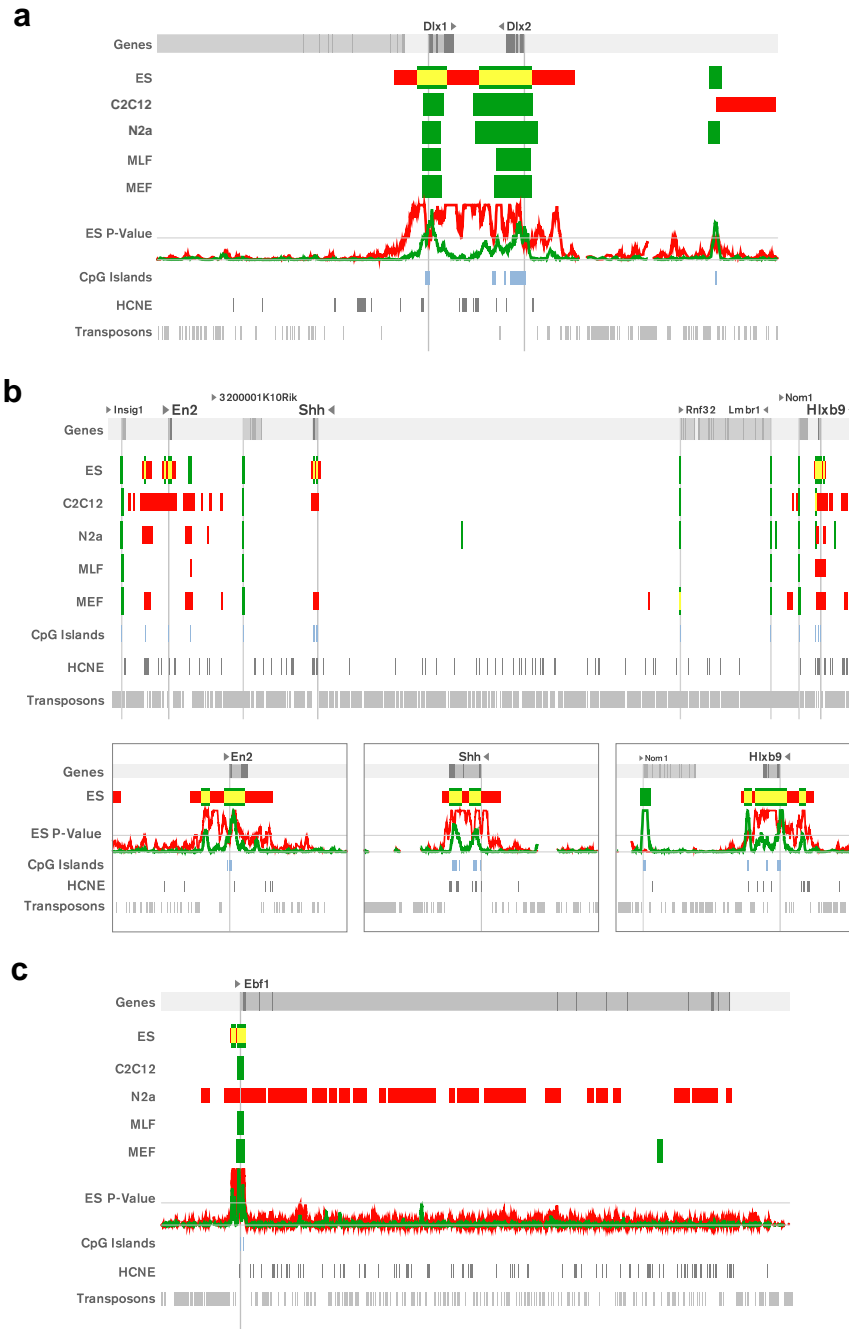


Figure 1. Representative views of histone methylation patterns across HCNE-rich regions in ES and differentiated cells. (a) Dlx1-Dlx2 gene cluster (Region 47, 112 kb). For each cell type, tracks show regions associated with H3K27me3 (red), H3K4me3 (green) or both modifications (yellow). For ES cells only, the raw p-value signals for H3K27me3 and H3K4me3 are also shown. Genes (TSSs indicated by long vertical lines; exons indicated as dark), CpG islands, HCNEs (Lindblad-Toh et al., 2005), and transposable elements are also shown. (b) En2, Shh, Hlx9 (Region 48, 1.5 Mb). Expanded views show 75 kb around En2, Shh and Hlx9. (c) Ebf1 (Region 31, 497 kb). Note expansive H3K27me3 methylated region in the Neuro2a cells.

Comparison of the two datasets revealed many instances of a previously undescribed pattern of chromatin modifications: three-quarters of the H3K27me3 domains contain H3K4me3 sites within them. These regions thus harbor both a ‘repressive’ and an ‘activating’ chromatin modification; we therefore termed them ‘bivalent domains’. There are 95 in the HCNE regions, 9 in the Hox regions and 5 in the control regions.

Bivalent domains overlay developmentally-important TF genes in HCNE-rich regions

Roughly three-quarters of the bivalent domains in the HCNE regions (69/95) overlap TSSs of known genes, with the H3K4me3 sites typically positioned directly at the TSS. Of these, a full 93% (64/69) occur at genes that encode TFs, including Sox, Fox, Pax, Irx and Pou gene family members, even though TF genes make up only half of the genes in the regions examined (Figure 2). The 26 bivalent domains that do not occur at known TSSs are also of interest: four occur at the 3’-ends of developmental genes (Npas3, Meis2, Pax2 and Wnt8b), and ten occur in locations that show strong evidence of encoding transcripts (including the presence of mRNA transcripts, CpG islands and high levels of sequence conservation). Among the non-TF genes associated with bivalent domains are genes implicated in neural development, such as Fgf8 and Prok1. In the Hox regions, the observed bivalent domains are especially large and overlap multiple TSSs of known genes, all of which are TFs. The 5 bivalent domains in the control regions are quite short; they all overlap gene starts, although these genes do not encode TFs.

The chromatin analysis thus reveals that ES cells contain many bivalent domains. In HCNE-rich regions, these domains are particularly large and highly enriched at developmentally important genes that establish cell identity. The bivalent nature of this novel epigenetic pattern raises the possibility that the associated genes are poised in a bipotential state, which may be resolved differently in different cell lineages. This hypothesis predicts that differentiated cells would contain few, if any, bivalent domains.

In differentiated cells, TF genes are marked by either repressive or activating modifications

We next examined H3K4me3 and H3K27me3 patterns across these same regions in a collection of differentiated cell types, including mouse embryonic fibroblasts (MEFs), mouse primary lung fibroblasts (MLFs), C2C12 myoblasts and Neuro2a neuroblastoma cells. We identified multiple H3K4me3 and H3K27me3 sites in each cell type, many of which are large. However, in marked contrast to the ES cell data, we found few bivalent domains in the differentiated cells (6 in MEFs, 1 in MLFs, 13 in myoblasts, 12 in the neuroblastoma cells).

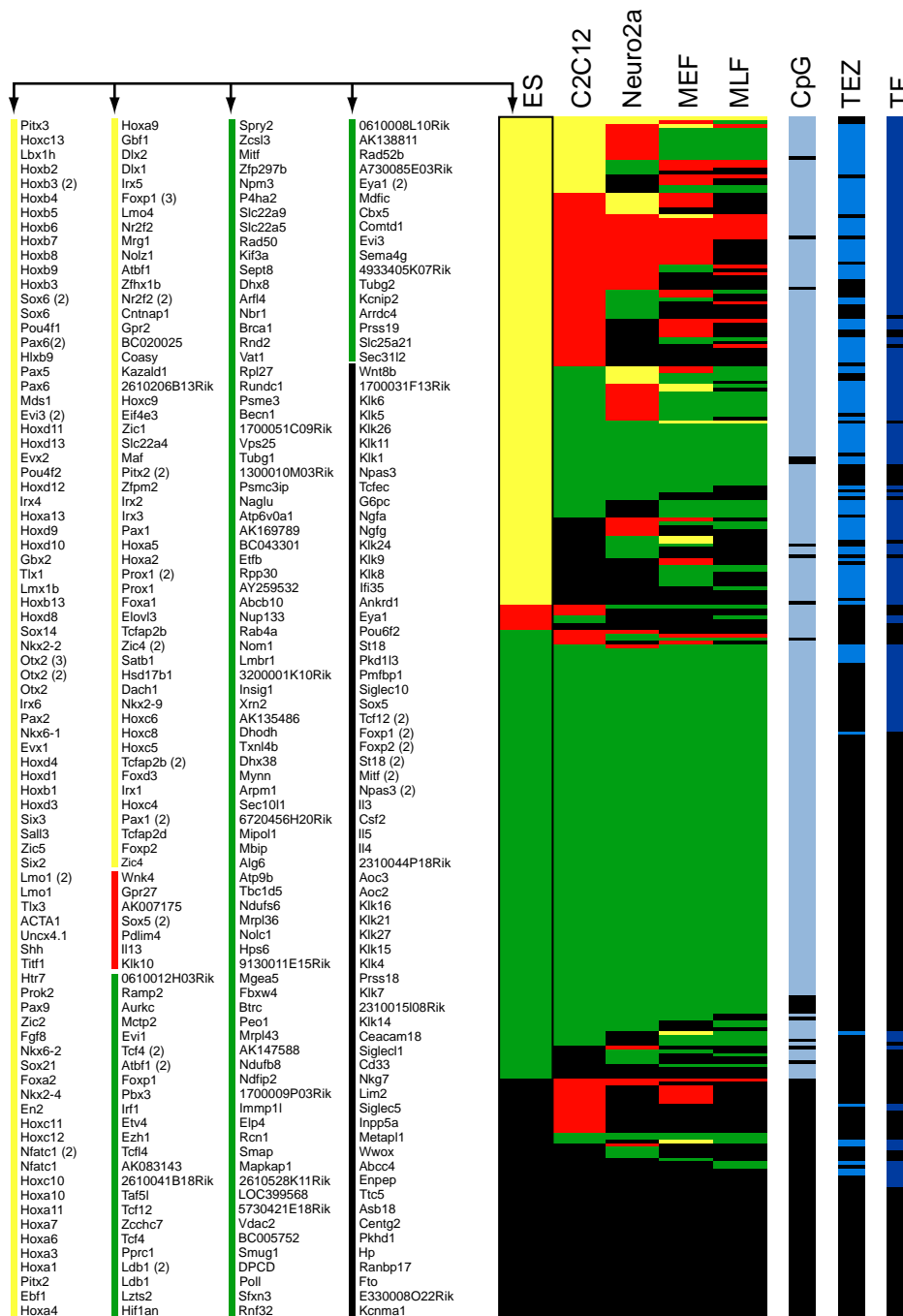


Figure 2. Histone methylation status of transcription start sites (TSSs). Methylation status of chromatin associated with each of the 332 known TSSs in the 61 examined regions in ES cells, C2C12 myoblasts, Neuro2a neuroblastoma cells, mouse embryonic fibroblasts (MEF), and mouse lung fibroblasts (MLF). Yellow indicates the presence of a bivalent domain, red indicates K27me3 only, green indicates K4me3 only, and black indicates no detected methylation. Blue rows in the three rightmost columns indicate which TSSs correspond to TF genes, contain CpG islands or coincide with transposon-exclusion zones (TEZs). This figure shows the strong correlation among bivalent domains, transcription factors, CpG islands and TEZs.

Thus, the majority of TSSs that show bivalent domains in ES cells do not show bivalent domains in the differentiated cells. The vast majority of these (93/97) instead show either a H3K27me3 or a H3K4me3 site in at least one of the differentiated cell types (Figure 2). These ‘monovalent’ sites tend to be large, with median sizes of 19.4 kb and 7.4 kb for H3K27me3 and H3K4me3, respectively (compared to 6.7 kb and 3.4 kb over all sites in the differentiated cells). Thus, bivalent domains appear largely specific to ES cells and, in differentiated cells, developmental genes are instead frequently organized within expansive regions showing either repressive or activating modifications.

Bivalent modification patterns in ES cells confirmed by alternate techniques

Given the novel nature of the bivalent domains, we sought to confirm our results using completely different reagents and protocols (see Methods). Specifically, we used an independent source of ES cells with a different genotype; we refer to the first source as ES1 and the second as ES2. We also used an alternative ChIP procedure carried out on micrococcal nuclease-digested nucleosomes that had not been subjected to cross-linking, and performed the immunoprecipitation with antisera from different sources. This alternative ChIP technique controls for nucleosome occupancy and is not subject to potential artifacts of cross-linking and sonication (O'Neill and Turner, 2003). The ES2 data also show a large number of bivalent domains, and these correspond closely to those seen in the ES1 data. Importantly, 94 of the 95 bivalent domains in the ES2 data correspond to bivalent domains in the ES1 cells.

We next sought to test whether the observed bivalent domain structure truly reflects the simultaneous presence of both H3K4me3 and H3K27me3 methylation on the same physical chromosomes. It is formally possible that the bivalent domains could instead reflect the presence of either two subpopulations with distinct character, or one population alternating between two states. To rule out this possibility, we carried out a sequential ChIP in which ES cell chromatin was immunoprecipitated first with H3K27me3 antibody and second with H3K4me3 antibody. This sequential purification is designed to retain only chromatin that concomitantly carries both kinds of modifications. Using real-time PCR, we tested three TSSs associated with bivalent domains (*Irx2*, *Dlx1* and *Hlxb9*). Each was significantly enriched relative to the controls (genes enriched for only H3K27me3 or only H3K4me3; Figure 3; Methods). For example, *Irx2* is enriched ~10-fold in the primary (H3K27me3) ChIP and further enriched >30-fold (relative to control) in the secondary (H3K4me3) ChIP. This shows that a large proportion of *Irx2* chromatin that contains H3K27me3 methylation also contains H3K4me3 – at least 30-fold more than the control. (Of course, the

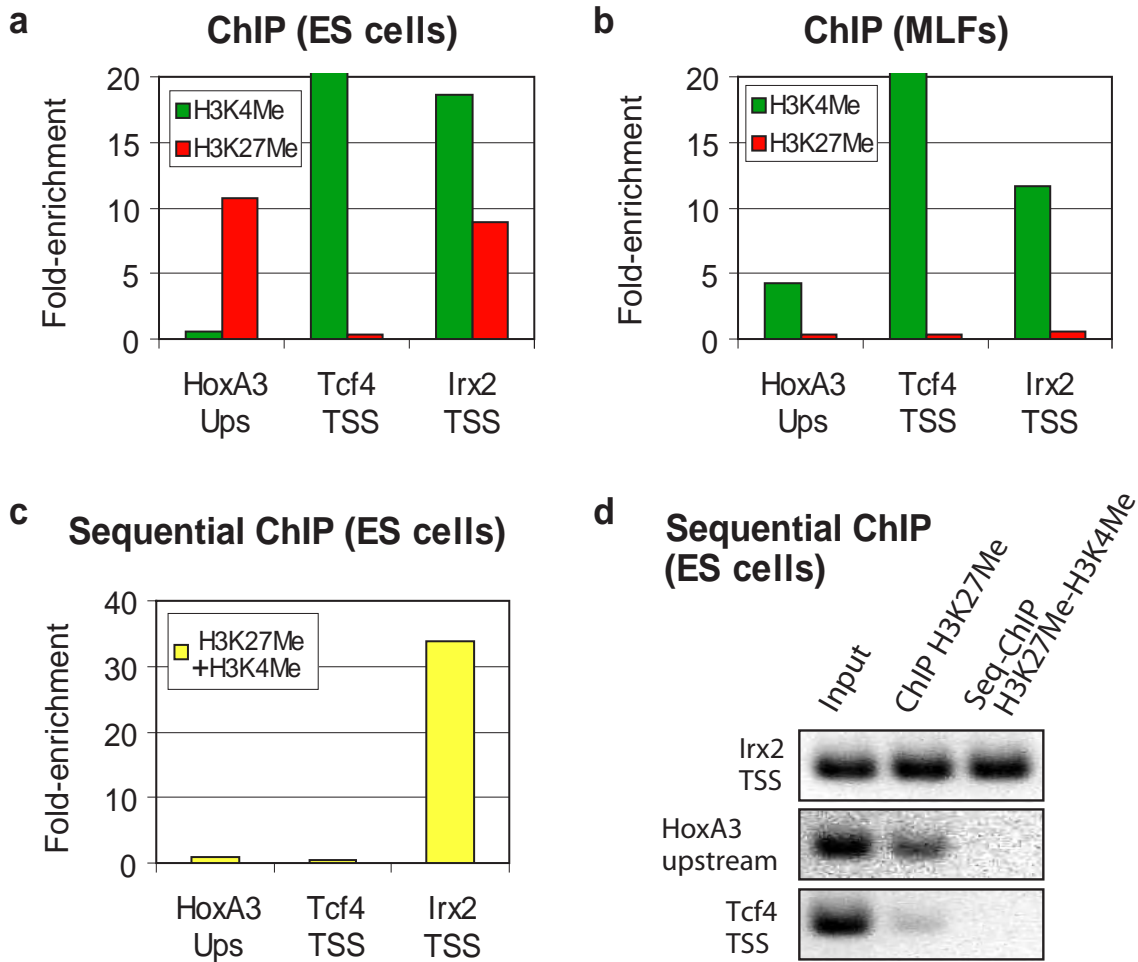


Figure 3. Characterization of the *Irx2* bivalent domain. ChIP and sequential ChIP were used to examine the methylation status of the *Irx2* TSS (bivalent), the *Tcf4* TSS (H3K4me3 only) and a site upstream of *HoxA3* (H3K27me3 only). (a) Real-time PCR ratios reflect the enrichment of indicated sites when ES cells are subjected to ChIP with H3K4me3 or H3K27me3 antibodies. (b) Corresponding data for mouse lung fibroblasts. (c) Real-time PCR ratios reflect the relative enrichment of indicated sites after sequential immunoprecipitations with H3K27me3 antibody and then H3K4me3 antibody (see Methods). (d) PCR products amplified from control (input) DNA, H3K27me3 ChIP DNA and H3K27me3-H3K4me3 sequential ChIP DNA are shown; the same samples were used as real-time PCR template in (c). The sequential chip results indicate that, in ES cells, the *Irx2* TSS is associated with chromatin marked by both H3K27me3 and H3K4me3.

technique cannot prove that 100% of all Irx2 species in ES cells carry both modifications). We also tested the Irx2 TSS by repeating the sequential ChIP with the order of the immunoprecipitations reversed, and again found significant enrichment (see Methods). Together, the experiments above suggest that the bivalent domains accurately represent the epigenetic state at many TF genes in ES cells.

Bivalent domains are associated with low levels of gene expression

To gain insight into the functional significance of bivalent domains, we examined gene expression patterns across the three cell types with at least 10 bivalent domains (ES cells, C2C12, Neuro2a) (Mogass et al., 2004; Perez-Iratxeta et al., 2005; Tomczak et al., 2004). Within each cell type, we found that genes marked by H3K4me3 tend to be expressed at significantly higher levels than those associated with H3K27me3 (Figure 4). A good example is the Ebf1 gene, which encodes a TF implicated in multiple differentiation pathways: in MEFs, MLFs and myoblasts it is expressed at relatively high levels (Koli et al., 2004; Schraets et al., 2003) and is associated with relatively large H3K4me3 sites (> 5 kb), while in neuroblastoma cells it is expressed at an essentially undetectable level and is associated with an expansive H3K27me3 domain.

We next examined the expression levels of genes marked by bivalent domains. These show low levels of expression, with the overall distribution being similar to that for genes marked by H3K27me3 alone (Figure 4). Thus, the presence of H3K4me3 at a TSS is typically associated with high gene activity when it occurs in the absence of H3K27me3, but with low gene activity when it occurs together with H3K27me3 (that is, the repressive effect of H3K27me3 appears to be epistatic to the activating effect of H3K4me3 in a bivalent domain). These results raise the possibility that bivalent domains function to silence developmental genes in ES cells while keeping them poised for induction upon initiation of specific developmental pathways.

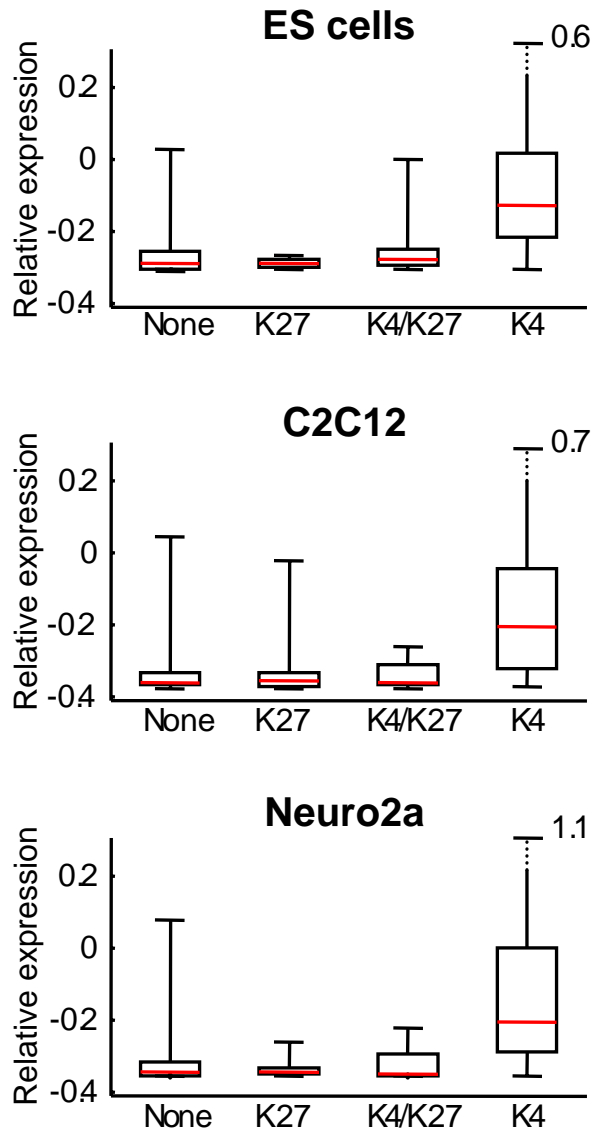


Figure 4. Gene expression as a function of histone methylation status. Box plot showing 25th, 50th and 75th percentile expression levels in ES cells, myoblasts and neuroblastoma cells for genes associated with no histone methylation, H3K27me3, bivalent domains or H3K4me3. Whiskers show 2.5th and 97.5th percentiles. Expression data (y-axis) were determined from published expression profiles (Mogass et al., 2004; Perez-Iratxeta et al., 2005; Tomczak et al., 2004), uniformly normalized to a mean of 0 and a standard deviation of 1 for all probes on each array.

Resolution of bivalent domains during ES cell differentiation

Our analysis of ES cells and four differentiated cell types suggests that bivalent domains are characteristic of pluripotent cells, that they silence developmental genes while keeping them poised, and that they tend to resolve upon ES cell differentiation into H3K4me3 or H3K27me3, in accordance with associated changes in gene expression. We sought to study whether the resolution of bivalent domains can be observed soon after ES cell differentiation, by examining a differentiated cell type obtained directly from ES cells. Specifically, we differentiated ES cells along a neural pathway in serum-free culture and generated a homogenous population of multipotent neural precursor cells maintained in FGF2- and EGF-containing media, as described previously (Conti et al., 2005). We focused on seven genes associated with bivalent domains in ES cells (Figure 5). These include three genes that are markedly induced during differentiation (Nkx2-2, Sox21 and Zfpn2), one that is weakly induced (Dlx1), and three that are not induced (Pax5, Lbx1h, and Evx1). Using ChIP and real-time PCR, we first confirmed that the TSS of each gene is indeed associated with both H3K4me3 and H3K27me3 in the original ES cells, and then examined the methylation status of these genes in the neural precursor cells. For the three genes whose expression is markedly induced, the TSS becomes specifically associated with H3K4me3 in these differentiated cells. For the three genes that are not induced, the TSS becomes specifically associated with H3K27me3. Interestingly, the TSS of the weakly induced gene, Dlx1, remains associated with both methylation marks in the neural precursor cells, although the H3K4me3 signal is significantly stronger. These data support a model in which bivalent domains are largely specific to ES cells and tend to resolve upon ES cell differentiation according to pathway-specific gene expression programs.

Epigenetic modifications in ES cells strongly correlate with underlying DNA sequence

Because the H3K4me3 and H3K27me3 sites in ES cells seem to represent important initial conditions for development, we searched for DNA sequence features that might underlie or predict the establishment of these epigenetic marks across the genome.

We found a strong positive correlation between the presence of H3K4me3 in ES cells and the density of CpG dinucleotides in the underlying DNA sequence (median 8% vs. 2% expected; Figure S2). Strikingly, 95% of TSSs with H3K4me3 sites have CpG islands, and 91% of TSSs with CpG islands also have H3K4me3 sites ($r_{\text{phi}} = 0.73$ see Methods). Moreover, the lengths of the two features are significantly correlated where they overlap ($r = 0.50$). By contrast, the correlation is weaker in the differentiated cells; this is primarily due to loss of H3K4me3 at 20-35% of CpG

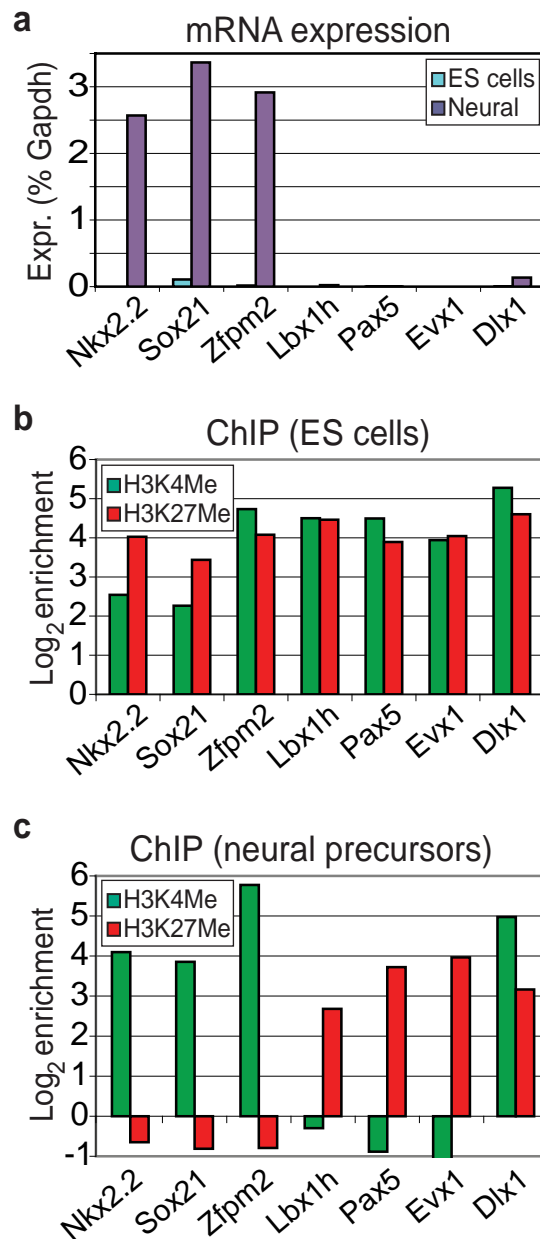


Figure 5. Resolution of bivalent domains during ES cell differentiation. ES cells were differentiated along a neural pathway in serum-free culture, and a homogenous population of multipotent neural precursor cells were maintained in FGF2- and EGF-containing media, as described (Conti et al., 2005). Several loci showing bivalent domains in ES cells were examined in the differentiated cells. (a) Expression levels (relative to Gapdh) were determined by RT-PCR for the indicated genes in ES cells and in neural precursors. The methylation states of the indicated genes were determined by ChIP and real-time PCR in ES cells (b) and in neural precursors (c). The data suggest that bivalent domains tend to resolve during ES cell differentiation in accordance with associated changes in gene expression.

islands ($r_{\text{phi}} = 0.40$ for MLFs). We note that a recent genome-wide study (Roh et al., 2005) of histone H3 acetylation in T-cells observed a correlation with CpG islands, at a similar level to that seen in the differentiated cells examined here.

We also found that H3K27me3 methylated regions in ES cells show a strikingly low density of transposon-derived sequence (median 6% vs. 22% expected; Figures 6, 7). The most extreme example is found at the Hox clusters, which are known to have the lowest density of transposon-derived sequence in the mouse and human genomes (Lander et al., 2001; Waterston et al., 2002) and which have the largest H3K27me3 domains (up to 141 kb) in our sample. Most of the H3K27me3 domains contain long stretches (>10 kb) with little or no identifiable transposon-derived sequence. We defined such regions as ‘transposon exclusion zones’ (TEZs; see Methods). Within the loci examined here, 89% of TSSs with a TEZ have a H3K27me3 domain in ES cells, and 73% of TSSs with a H3K27me3 domain have a TEZ ($r_{\text{phi}} = 0.69$). The lengths of these two features are significantly correlated where they overlap ($r = 0.78$). Interestingly, we note that the small number of H3K27me3 domains found only in differentiated cells do not appear to overlap particularly transposon-poor sequence (e.g., Figure 1c).

We tested if the TEZs represent conserved genomic features by examining the orthologous sequence in the human and dog genome. The frequency of lineage-specific repeats provides an independent test of whether transposons are tolerated in these regions. The TEZs show a clear deficit of lineage-specific repeats in both human (1.3% vs. 15.2% expected) and dog (1.0% vs. 9.1% expected), confirming that this property is strongly conserved across mammals.

We then searched for TEZs across the entire mouse genome. We identified 710 TEZs, of which 328 overlap TSSs of known genes. Strikingly, the vast majority of these genes encode developmental and tissue-specific TFs (189), proteins involved in axon guidance and neuronal function (65), and other cell signaling-related proteins such as growth factors (25), including Fgf8, Fgf10, Fgf14 and the imprinted gene Igf2. Notably, they include ~70% of the developmental regulators previously identified within 204 HCNE-rich loci (Lindblad-Toh et al., 2005). We predict that most of these genes will harbor H3K27me3 domains or bivalent domains in ES cells.

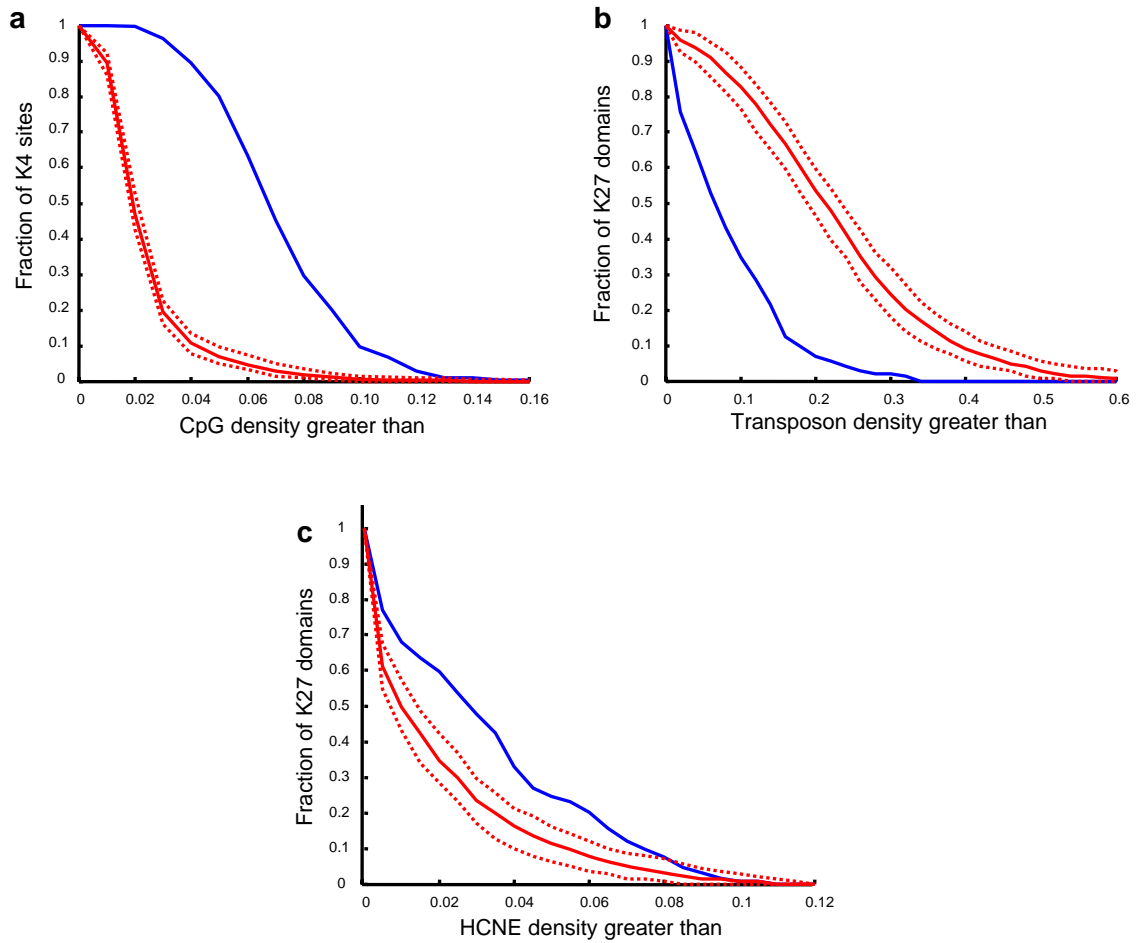


Figure 6. Cumulative distributions of density of CpG dinucleotides within H3K4me3 sites in ES cells (a), transposon-derived nucleotides within H3K27me3 domains in ES cells (b) and HCNE nucleotides within H3K27me3 domains in ES cells (c). The blue curves give the observed distributions. The red solid curves give the median, and the dotted red curves give the 2.5th and 97.5th percentiles over 10,000 randomized sites/domains.

Co-localization of bivalent domains with Oct4 and Nanog

Finally, we examined the relationship of bivalent domains to the reported binding sites of certain pluripotent TFs. A recent genomic analysis in human ES cells found that Oct4, Nanog and Sox2 are frequently associated with developmentally-important genes (Boyer et al., 2005). We mapped the Oct4, Nanog and Sox2 binding sites reported in that study to orthologous positions in the mouse genome and examined their overlap with bivalent domains. About 50% of bivalent domains coincide with binding sites of at least one of the pluripotent TFs, a highly significant correspondence ($p < 10^{-9}$). The correlation is primarily due to Oct4 and Nanog, and actually becomes more significant when Sox2 is removed from the analysis. Interestingly, although many of the genes targeted by these pluripotent factors are actively expressed in ES cells, those that are also associated with a bivalent domain tend to be silenced ($p < 5 \times 10^{-3}$). This suggests that the bivalent domains may override any activation potential these TFs might have, but also raises the possibility that the pluripotent TFs may help keep these genes in a poised state. Notably, a full 50% of bivalent domains are not associated with any of the three pluripotent TFs. It will be interesting to see if these coincide with binding sites of other important TFs.

Discussion

Our results shed light on chromatin structure in ES cells and raise intriguing hypotheses about its establishment and function during development. The bivalent domains reported here have many notable features: they combine both ‘repressive’ and ‘activating’ modifications; they are highly enriched in ES cells relative to differentiated cells; and they are associated with genes encoding TFs with roles in embryonic development and lineage specification. In differentiated cells, these TF genes instead tend to be associated with large regions carrying either an activating or a repressive methylation mark. We propose that bivalent domains silence developmental genes in ES cells, while preserving their potential to become activated upon initiation of specific differentiation programs. Bivalent domains may be related to a phenomenon observed at the bithorax complex in early fly development, where silenced polycomb response elements are nonetheless associated with trithorax-group proteins and low-level transcription. Remarkably, both of these activities appear to be required for subsequent gene activation during development (Orlando et al., 1998; Schmitt et al., 2005). By analogy, H3K4me3 within bivalent domains and associated trithorax activities may keep silenced developmental genes poised in ES cells. Our analyses of differentiated cells suggest that bivalent domains largely resolve during differentiation into large regions of either H3K27me3 or H3K4me3. These modified regions may provide a robust epigenetic memory to maintain lineage-

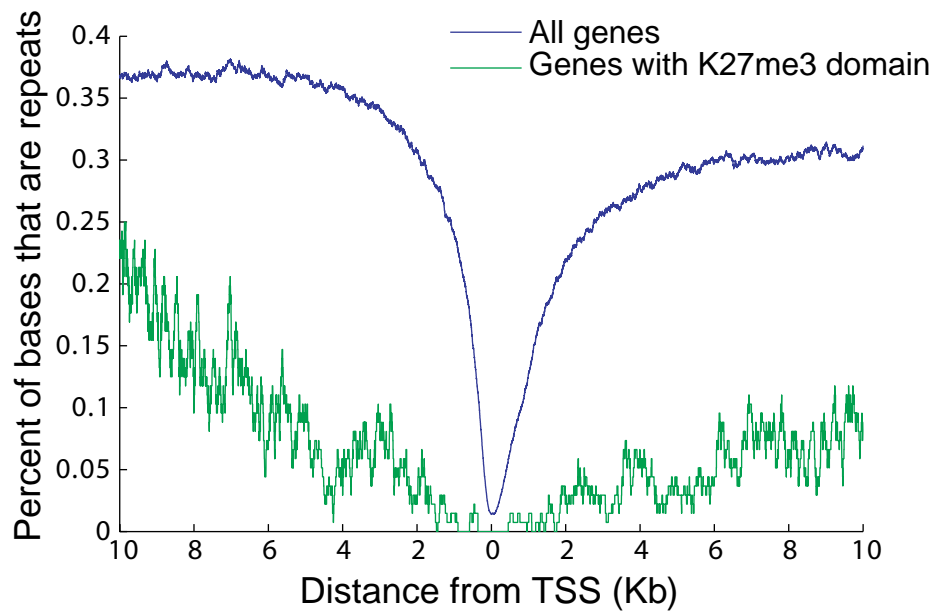


Figure 7. Density of transposable element-derived nucleotides surrounding TSSs of all genes (blue) and genes with H3K27me3 domains in ES cells (green).

specific expression or repression of these critical genes. Their large size would ensure that each daughter chromosome would likely inherit a substantial proportion of the modified histones, which could then promote similar modification of new histones in the immediate vicinity (Henikoff et al., 2004; van Steensel, 2005).

A fundamental issue remains to understand the mechanism by which the initial conditions are established in ES cells. The analysis here suggests that some of the answer can be read directly from the genome sequence. The strong association of H3K4me3 with CpG islands may well be directly causal, inasmuch as H3K4me3 methylases are known to associate with CpG-rich DNA (Ayton et al., 2004; Lee and Skalnik, 2005). The strong association of H3K27me3 with transposon-poor zones may instead reflect strong evolutionary pressure against the presence of transposon-derived sequence in these regions. Repetitive sequences are subject to repressive epigenetic modifications (Arnaud et al., 2000; Lippman et al., 2004; Martens et al., 2005), which might interfere with the function of the bivalent domains and thus be eliminated by selection. It has been reported previously that imprinted loci, while significantly depleted for short interspersed transposable elements (SINEs), are permissive to L1 long interspersed transposable elements (LINEs) (Greally, 2002). In contrast, we find that both classes of transposons tend to be excluded from regions associated with H3K27me3 or bivalent domains in ES cells. The direct signal for H3K27me3 remains unclear (although we cannot exclude the possibility that the deficit of transposon-related chromatin modifications in some fashion promotes the recruitment of H3K27me3 methylases). The correlations between the histone modifications and the genomic features are notably weaker in differentiated cells (Bernstein et al., 2005). We suggest that, while the embryonic state may be largely defined by DNA sequence, it is subsequently altered in response to lineage-specific transcriptional programs and environmental cues, and epigenetically maintained.

Our study was motivated by the suspicion that HCNE-rich regions might be particularly fruitful targets for studying chromatin structure in ES cells; this has indeed been borne out. However, the results here do not explain the functional role of the HCNEs themselves. Although HCNEs are markedly enriched at many of the H3K27me3 and H3K4me3 sites in both ES and differentiated cells, they tend overall to be distributed across much larger regions. One possibility is that some of the HCNEs dictate chromosome conformation or nuclear localization in a manner that facilitates robust gene regulation and/or epigenetic switching (Chambeyron and Bickmore, 2004; Kosak and Groudine, 2004).

Further studies will be needed to define bivalent domains and related features. It will be important to examine the entire genome in ES cells, as well as to follow their fate during

development and differentiation. In particular, it will be interesting to determine whether the bivalent domains that persist following ES cell differentiation correspond to genes that remain poised for later induction. In addition, it will be valuable to characterize the bivalent domains with respect to other epigenetic modifications and the binding sites of additional TFs. We note that preliminary studies of H3 Lys9 methylation show no evidence of association with bivalent domains.

A deeper understanding of bivalent domains may shed light on mechanisms that underlie the maintenance of pluripotency in ES cells and lineage fidelity in differentiated cells. Moreover, a comprehensive inventory of the presence or absence of bivalent domains over key developmental genes may provide valuable markers of cell identity and differentiation potential, both in normal and pathologic states.

Methods

Cell culture. The first source of ES cells (ES1 above) were V6.5 murine ES cells (genotype 129SvJae x C57BL/6; male; passages 10-15). They were cultivated in 5% CO₂ at 37 degrees on irradiated MEFs in DMEM containing 15% FCS, leukemia-inhibiting factor, penicillin/streptomycin, L-glutamine, and nonessential amino acids (Rideout et al., 2000). At least 2-3 passages under feeder-free conditions on 0.2% gelatin were used to exclude feeder contamination. The second source of ES cells (ES2 above, used for micrococcal nuclease-digestion ChIP) were SF1-1 murine ES cells (genotype C57BL/6 x M. spretus F1; male; passages 11-16) grown in the absence of feeder cells on gelatinized plates as described previously (Umlauf et al., 2004). Primary mouse lung fibroblasts (ATCC: # CCL-206), mouse embryonic fibroblasts (10.5 p.c.) immortalized with polyoma virus, C2C12 myoblasts (ATCC #CRL-1772) and Neuro2a neuroblastoma cells (ATCC #CCL-131) were grown in DMEM with 10% fetal bovine serum and penicillin/streptomycin at 37 degrees, 5% CO₂. ES1 cells were differentiated into pan-neural precursor cells through embryoid body formation for 4 days and selection in ITSFn media for 5-7 days (Okabe et al., 1996) and maintained in FGF2 and EGF2 (both from R&D Systems) containing, chemically defined media as described (Conti et al., 2005). These cells uniformly express nestin and Sox2 and upon growth factor withdrawal differentiate into neurons, astrocytes and oligodendrocytes (Brustle et al., 1999; Conti et al., 2005).

Chromatin immunoprecipitation. ChIP experiments for all cells except ES2 were carried out essentially as described in ref. (Bernstein et al., 2005) and at <http://www.upstate.com>. Briefly, ~5 x 10⁷ cells were trypsinized, fixed with 1% formaldehyde, resuspended in Lysis Buffer and sonicated with a Branson 250 Sonifier to fragment chromatin to a size range of 200 to 1000 bases. Solubilized chromatin was diluted 10-fold in ChIP dilution buffer and, after removal of a control aliquot (whole cell extract), incubated at 4°C overnight with antibodies against H3K4me3 (Abcam #8580), H3K27me3 (Upstate #07-449). Immune complexes were precipitated with Protein A-sepharose, washed sequentially with Low Salt Immune Complex Wash, LiCl Immune Complex Wash, and TE, and then eluted in Elution Buffer. After cross-link reversal and Proteinase K treatment, ChIP and control DNA samples were extracted with phenol-chloroform, precipitated under ethanol, treated with RNase and Calf Intestinal Alkaline Phosphatase, and then purified with a MinElute Kit (Qiagen).

Micrococcal nuclease-ChIP. Chromatin fragments of one to six nucleosomes were prepared from unfixed chromatin from ES2 cells by micrococcal nuclease digestion, and immunoprecipitated

using antibody against H3K27me3 (Plath et al., 2003) or H3K4me2 (Upstate #07-030) as described (Umlauf et al., 2004). Immunoprecipitated DNA fractions and a control DNA sample enriched using unrelated antisera (against chicken antibodies) were extracted and purified as described above for the immunoprecipitated cross-linked chromatin.

Sequential ChIP. Cross-linked chromatin from ES cells was immunoprecipitated with antibody against H3K27me3 as described above (see “Chromatin immunoprecipitation”) with the exception that chromatin was eluted from beads with a modified elution buffer containing 30mM DTT, 500mM NaCl and 0.1% SDS at 37 degrees. Eluted chromatin was diluted 50-fold, subjected to a second immunoprecipitation with antibody against tri-methyl H3K4me3, and eluted with standard Elution Buffer. Thus isolated DNA and unenriched control DNA were extracted and purified as described above. In addition, a ‘reverse’ sequential ChIP was carried out as above, except that chromatin was immunoprecipitated first with antibody against H3K4me3 and then with antibody against H3K27me3.

Real-time PCR. PCR primers were designed to amplify fragments 150 to 200 base pairs in size from the indicated genomic regions. Real-time PCR was carried out using Quantitect SYBR green PCR mix (Qiagen) in an MJ Research Opticon Instrument. For ChIP experiments, fold-enrichments were determined in real-time PCR reactions using either 0.5 ng ChIP DNA or 0.5 ng control DNA as template by the $2^{-\Delta\text{CT}}$ method described in the Applied Biosystems User Bulletin. For sequential ChIP experiments, 2 μl sequential ChIP DNA or 2 μl of a 1:100 dilution of control DNA were used as template, and relative fold-enrichments were determined by the $2^{-\Delta\Delta\text{CT}}$ method, using HoxA3 Ups as the normalizer. Each ratio was determined from two independent ChIP or sequential ChIP assays, each evaluated in duplicate by real-time PCR. Primers corresponding to the TSSs of Irx2, Dlx1 and Hlxb9 were used to test for enrichment of lysine 9 di- and tri-methylation, relative to a Gapdh control. RT-PCR was used to measure gene expression in ES cells and neural precursor cells. Briefly, RNA was isolated using an RNeasy mini kit (Qiagen), reverse transcribed and quantified on the 7000 ABI detection system using SYBR green PCR master mix.

Region selection and array design. We previously reported the identification of 204 HCNE-rich loci on the basis of sequence comparisons across the human, mouse and dog genomes (Lindblad-Toh et al., 2005). For the current study, we selected 56 HCNE-rich loci, including all four Hox clusters, as well as 5 control loci that do not show unusual HCNE density (ACTA locus, chromosome 19 gene desert, CD33r locus, BRCA1 locus, cytokine cluster). Each region was mapped to the mouse genome using coordinates from version mm5. Custom tiling arrays for these

regions were obtained from Affymetrix Inc. (Santa Clara, CA). These contain approximately 1.3 million probe pairs, each consisting of perfect match (PM) and single base mismatch (MM) 25-mer oligonucleotides, designed to interrogate the unique sequence in these regions at approximately 30 base intervals (Kapranov et al., 2002).

DNA amplification and array hybridization. ChIP and control DNA samples were amplified by *in vitro* transcription, converted into double-stranded cDNA with random primers, fragmented with DNase I, and end-labeled with biotin as described (Bernstein et al., 2005; Cawley et al., 2004; Kapranov et al., 2002). ChIP and control samples (5-10 μ g) were hybridized to separate oligonucleotide arrays. Arrays were hybridized 16-18 hours at 45°C, washed, stained, and scanned using an Affymetrix GeneChip Scanner 3000 7G as described in the Affymetrix Expression Analysis Technical Manual.

Analysis of ChIP tiling array data. Raw array data were quantile-normalized, scaled and analyzed as described (Bernstein et al., 2005; Cawley et al., 2004). Briefly, (PM, MM) intensity pairs were mapped to the genome using exact 25-mer matching to mm5. A Wilcoxon Rank Sum test was applied to the transformation $\log_2(\max(\text{PM}-\text{MM}, 1))$ for data from the ChIP and control (whole cell extract) arrays within a window of ± 500 base pairs, testing the null hypothesis that ChIP and control data come from the same probability distribution. Genomic positions belonging to enriched regions were defined by applying a high P-value cutoff of 10^{-4} . These regions were extended locally by merging adjacent windows with P-values of at least 10^{-2} , and resultant positions separated by < 2 kb were merged to form a predicted H3K4me3 or H3K27me3 site. We defined bivalent domains to be H3K27me3 methylated sites of at least 5 kb that overlap H3K4me3 sites of at least 1 kb.

Genomic analysis. We manually collated a list of known TSSs based on RefSeq and Genbank mRNAs aligned to the examined regions in mouse (mm5) and the orthologous regions in human (hg17; alignments obtained from the UCSC genome browser). The methylation status of each TSS was based on the presence of significantly enriched H3K4me3 or H3K27me3 sites, or bivalent domains within 2 kb upstream or downstream. The expected density of CpG and transposable elements at methylated sites were determined from random intervals of the same size, anchored in non-repetitive sequence. CpG islands were defined as in (Takai and Jones, 2002). TEZs (Transposon exclusion zones) were defined as regions satisfying one of two criteria: (a) regions of at least 10 kb without any transposable elements; (b) regions of at least 15 kb with no more than 250 bases annotated as transposable elements. Identified regions were then merged together as one TEZ if they were within distance of 1kb. For the genome-wide search, only criterion (a) was used.

References

- Arnaud, P., Goubely, C., Pelissier, T., and Deragon, J. M. (2000). SINE retroposons can be used in vivo as nucleation centers for de novo methylation. *Mol Cell Biol* 20, 3434-3441.
- Ayton, P. M., Chen, E. H., and Cleary, M. L. (2004). Binding to nonmethylated CpG DNA is essential for target recognition, transactivation, and myeloid transformation by an MLL oncoprotein. *Mol Cell Biol* 24, 10470-10478.
- Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* 304, 1321-1325.
- Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J., 3rd, Gingeras, T. R., *et al.* (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120, 169-181.
- Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., *et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947-956.
- Brustle, O., Jones, K. N., Learish, R. D., Karram, K., Choudhary, K., Wiestler, O. D., Duncan, I. D., and McKay, R. D. (1999). Embryonic stem cell-derived glial precursors: a source of myelinating transplants. *Science* 285, 754-756.
- Cao, R., and Zhang, Y. (2004). SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. *Mol Cell* 15, 57-67.
- Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., *et al.* (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116, 499-509.
- Chambeyron, S., and Bickmore, W. A. (2004). Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev* 18, 1119-1130.
- Conti, L., Pollard, S. M., Gorba, T., Reitano, E., Toselli, M., Biella, G., Sun, Y., Sanzone, S., Ying, Q. L., Cattaneo, E., and Smith, A. (2005). Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol* 3, e283.
- Delaval, K., and Feil, R. (2004). Epigenetic regulation of mammalian genomic imprinting. *Curr Opin Genet Dev* 14, 188-195.
- Francis, N. J., Kingston, R. E., and Woodcock, C. L. (2004). Chromatin compaction by a polycomb group protein complex. *Science* 306, 1574-1577.
- Greally, J. M. (2002). Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc Natl Acad Sci U S A* 99, 327-332.
- Guenther, M. G., Jenner, R. G., Chevalier, B., Nakamura, T., Croce, C. M., Canaani, E., and Young, R. A. (2005). Global and Hox-specific roles for the MLL1 methyltransferase. *Proc Natl Acad Sci U S A* 102, 8603-8608.
- Henikoff, S., Furuyama, T., and Ahmad, K. (2004). Histone variants, nucleosome assembly and epigenetic inheritance. *Trends Genet* 20, 320-326.
- Jenuwein, T., and Allis, C. D. (2001). Translating the histone code. *Science* 293, 1074-1080.
- Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P., and Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916-919.

- Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature* *436*, 876-880.
- Kimura, H., Tada, M., Nakatsuji, N., and Tada, T. (2004). Histone code modifications on pluripotential nuclei of reprogrammed somatic cells. *Mol Cell Biol* *24*, 5710-5720.
- Kirmizis, A., Bartley, S. M., Kuzmichev, A., Margueron, R., Reinberg, D., Green, R., and Farnham, P. J. (2004). Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes Dev* *18*, 1592-1605.
- Koli, K., Wempe, F., Sterner-Kock, A., Kantola, A., Komor, M., Hofmann, W. K., von Melchner, H., and Keski-Oja, J. (2004). Disruption of LTBP-4 function reduces TGF-beta activation and enhances BMP-4 signaling in the lung. *J Cell Biol* *167*, 123-133.
- Kosak, S. T., and Groudine, M. (2004). Gene order and dynamic domains. *Science* *306*, 644-647.
- Koyanagi, M., Baguet, A., Martens, J., Margueron, R., Jenuwein, T., and Bix, M. (2005). EZH2 and histone 3 trimethyl lysine 27 associated with Il4 and Il13 gene silencing in T(H)1 cells. *J Biol Chem* *280*, 31470-31477.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* *409*, 860-921.
- Lee, J. H., and Skalnik, D. G. (2005). CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J Biol Chem* *280*, 41725-41731.
- Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., 3rd, Zody, M. C., *et al.* (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* *438*, 803-819.
- Lippman, Z., Gendrel, A. V., Black, M., Vaughn, M. W., Dedhia, N., McCombie, W. R., Lavine, K., Mittal, V., May, B., Kasschau, K. D., *et al.* (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature* *430*, 471-476.
- Margueron, R., Trojer, P., and Reinberg, D. (2005). The key to development: interpreting the histone code? *Curr Opin Genet Dev* *15*, 163-176.
- Martens, J. H., O'Sullivan, R. J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P., and Jenuwein, T. (2005). The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *Embo J* *24*, 800-812.
- Mogass, M., York, T. P., Li, L., Rujirabanjerd, S., and Shiang, R. (2004). Genomewide analysis of gene expression associated with Tcof1 in mouse neuroblastoma. *Biochem Biophys Res Commun* *325*, 124-132.
- Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. *Science* *302*, 413.
- O'Carroll, D., Erhardt, S., Pagani, M., Barton, S. C., Surani, M. A., and Jenuwein, T. (2001). The polycomb-group gene *Ezh2* is required for early mouse development. *Mol Cell Biol* *21*, 4330-4336.
- Okabe, S., Forsberg-Nilsson, K., Spiro, A. C., Segal, M., and McKay, R. D. (1996). Development of neuronal precursor cells and functional postmitotic neurons from embryonic stem cells in vitro. *Mech Dev* *59*, 89-102.
- O'Neill, L. P., and Turner, B. M. (2003). Immunoprecipitation of native chromatin: NChIP. *Methods* *31*, 76-82.
- Orlando, V., Jane, E. P., Chinwalla, V., Harte, P. J., and Paro, R. (1998). Binding of trithorax and Polycomb proteins to the bithorax complex: dynamic changes during early *Drosophila* embryogenesis. *Embo J* *17*, 5141-5150.

- Perez-Iratxeta, C., Palidwor, G., Porter, C. J., Sanche, N. A., Huska, M. R., Suomela, B. P., Muro, E. M., Krzyzanowski, P. M., Hughes, E., Campbell, P. A., *et al.* (2005). Study of stem cell function using microarray experiments. *FEBS Lett* 579, 1795-1801.
- Perry, P., Sauer, S., Billon, N., Richardson, W. D., Spivakov, M., Warnes, G., Livesey, F. J., Merckenschlager, M., Fisher, A. G., and Azuara, V. (2004). A dynamic switch in the replication timing of key regulator genes in embryonic stem cells upon neural induction. *Cell Cycle* 3, 1645-1650.
- Plath, K., Fang, J., Mlynarczyk-Evans, S. K., Cao, R., Worringer, K. A., Wang, H., de la Cruz, C. C., Otte, A. P., Panning, B., and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. *Science* 300, 131-135.
- Pray-Grant, M. G., Daniel, J. A., Schieltz, D., Yates, J. R., 3rd, and Grant, P. A. (2005). Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. *Nature* 433, 434-438.
- Rideout, W. M., 3rd, Wakayama, T., Wutz, A., Eggan, K., Jackson-Grusby, L., Dausman, J., Yanagimachi, R., and Jaenisch, R. (2000). Generation of mice from wild-type and targeted ES cells by nuclear cloning. *Nat Genet* 24, 109-110.
- Ringrose, L., Ehret, H., and Paro, R. (2004). Distinct contributions of histone H3 lysine 9 and 27 methylation to locus-specific stability of polycomb complexes. *Mol Cell* 16, 641-653.
- Ringrose, L., and Paro, R. (2004). Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu Rev Genet* 38, 413-443.
- Roh, T. Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* 19, 542-552.
- Sado, T., and Ferguson-Smith, A. C. (2005). Imprinted X inactivation and reprogramming in the preimplantation mouse embryo. *Hum Mol Genet* 14 *Spec No 1*, R59-64.
- Santos-Rosa, H., Schneider, R., Bernstein, B. E., Karabetsou, N., Morillon, A., Weise, C., Schreiber, S. L., Mellor, J., and Kouzarides, T. (2003). Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. *Mol Cell* 12, 1325-1332.
- Schmitt, S., Prestel, M., and Paro, R. (2005). Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev* 19, 697-708.
- Schraets, D., Lehmann, T., Dingermann, T., and Marschalek, R. (2003). MLL-mediated transcriptional gene regulation investigated by gene expression profiling. *Oncogene* 22, 3655-3668.
- Silva, J., Mak, W., Zvetkova, I., Appanah, R., Nesterova, T. B., Webster, Z., Peters, A. H., Jenuwein, T., Otte, A. P., and Brockdorff, N. (2003). Establishment of histone H3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Dev Cell* 4, 481-495.
- Sims, R. J., 3rd, Chen, C. F., Santos-Rosa, H., Kouzarides, T., Patel, S. S., and Reinberg, D. (2005). Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. *J Biol Chem* 280, 41789-41792.
- Szutorisz, H., and Dillon, N. (2005). The epigenetic basis for embryonic stem cell pluripotency. *Bioessays* 27, 1286-1293.
- Takai, D., and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* 99, 3740-3745.
- Tomczak, K. K., Marinescu, V. D., Ramoni, M. F., Sanoudou, D., Montanaro, F., Han, M., Kunkel, L. M., Kohane, I. S., and Beggs, A. H. (2004). Expression profiling and identification of novel genes involved in myogenic differentiation. *Faseb J* 18, 403-405.
- Umlauf, D., Goto, Y., Cao, R., Cerqueira, F., Wagschal, A., Zhang, Y., and Feil, R. (2004). Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. *Nat Genet* 36, 1296-1300.

Valk-Lingbeek, M. E., Bruggeman, S. W., and van Lohuizen, M. (2004). Stem cells and cancer; the polycomb connection. *Cell* 118, 409-418.

van Steensel, B. (2005). Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat Genet* 37 *Suppl*, S18-24.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-562.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., *et al.* (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3, e7.

Wysocka, J., Swigut, T., Milne, T. A., Dou, Y., Zhang, X., Burlingame, A. L., Roeder, R. G., Brivanlou, A. H., and Allis, C. D. (2005). WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* 121, 859-872.

[This page is intentionally left blank]

Chapter 7: Genome-wide maps of histone methylation

In this chapter, we describe one of the first application of single molecule-based sequencing to generate genome-wide chromatin state maps for a mammalian species.

This work was first published as

Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).

This publication is attached as Appendix 6. Supplementary notes can be found at the end of the chapter. Supplementary data is available online from <http://www.nature.com/nature>

[This page is intentionally left blank]

We report the application of single molecule-based sequencing technology for high-throughput profiling of histone modifications in mammalian cells. By obtaining over 4 billion bases of sequence from chromatin immunoprecipitated DNA, we generated genome-wide chromatin state maps of mouse embryonic stem cells, neural progenitor cells and embryonic fibroblasts. We find that lysine 4 and lysine 27 tri-methylation effectively discriminate genes that are expressed, poised for expression, or stably repressed, and therefore reflect cell state and lineage potential. Lysine 36 tri-methylation marks primary coding and non-coding transcripts, facilitating gene annotation. Lysine 9 and lysine 20 tri-methylation are detected at satellite, telomeric and active long-terminal repeats, and can spread into proximal unique sequences. Lysine 4 and lysine 9 tri-methylation mark imprinting control regions. Finally, we show that chromatin state can be read in an allele-specific manner by using single nucleotide polymorphisms. This study provides a framework for the application of comprehensive chromatin profiling towards characterization of diverse mammalian cell populations.

One of the fundamental mysteries of biology is the basis of cellular state. Although their genomes are essentially identical, cell types in a multicellular organism maintain strikingly different behaviors that persist over extended periods. The most extreme case is lineage-commitment during development, where cells progress from totipotency to pluripotency to terminal differentiation; each step involves establishment of a stable state encoding specific developmental commitments that can be faithfully transmitted to daughter cells. Considerable evidence suggests that cellular state may be closely related to ‘chromatin state’ – that is, modifications to histones and other proteins that package the genome¹⁻³. Accordingly, it would be desirable to construct ‘chromatin state maps’ for a wide variety of cell types, showing the genome-wide distribution of important chromatin modifications.

Chromatin state can be studied by chromatin immunoprecipitation (ChIP), in which an antibody is used to enrich DNA from genomic regions carrying a specific epitope. The major challenge to generating genome-wide chromatin state maps lies in characterizing these enriched regions in a scalable manner. Enrichment at individual loci is commonly assayed by PCR, but this method does not scale efficiently. A more recent approach has been ChIP-chip, in which enriched DNA is hybridized to a microarray^{4,5}. This technique has been successfully used to study large genomic regions. However, ChIP-chip suffers from inherent technical limitations: (i) it requires large amounts (several micrograms) of DNA and thus involves extensive amplification, which introduces bias; (ii) it is subject to cross-hybridization which hinders study of repeated sequences

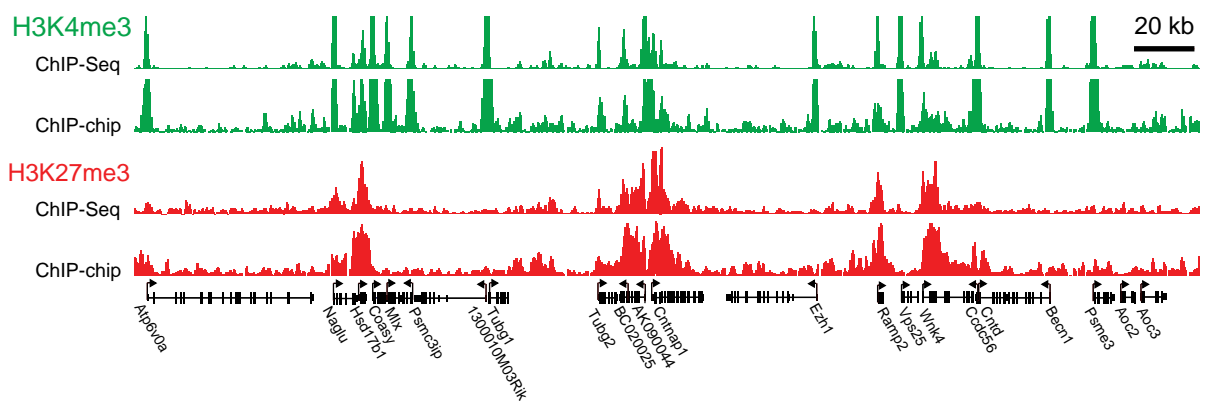


Figure 1. Comparison of ChIP-Seq and ChIP-chip data. Direct comparison of H3K4me3 (green) and H3K27me3 (red) ChIP data across a 300 kb region in mouse ESCs from independent experiments assayed by SMS (absolute fragment counts) or tiling arrays (log p-values for enrichment relative to whole-cell extracts [15]).

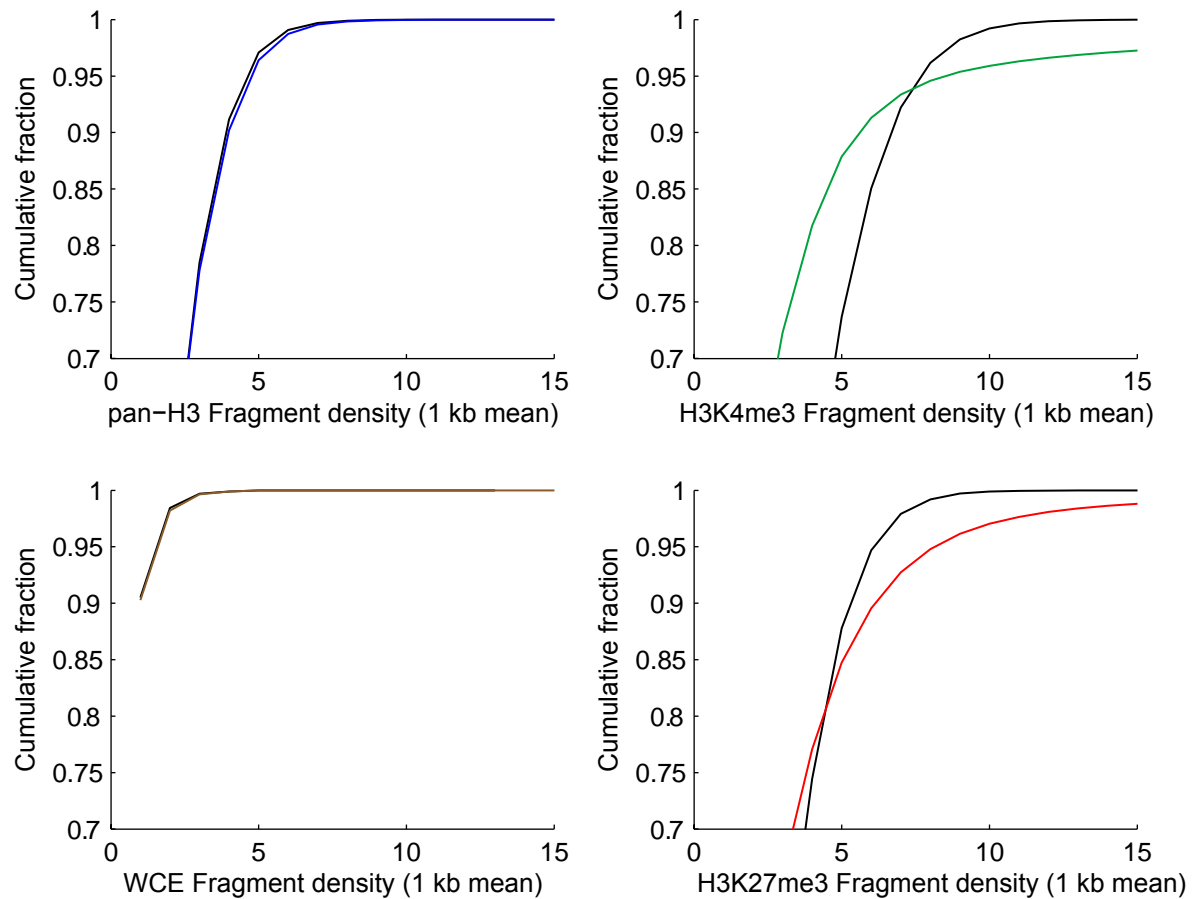


Figure 2. Cumulative distributions of fragment densities (averaged over 1-kb windows) across the mouse genome are shown for ES cell ChIPs of pan-H3 (blue), H3K4me3 (green) and H3K27me3 (red), and for unenriched whole-cell extracts (brown). The black curves show the distributions obtained from randomized placements of the of the same reads. The observed distributions for pan-H3 ChIP and whole-cell extract are virtually identical to the randomized distributions, indicating that ChIP-Seq generates unbiased data from unenriched samples. In contrast, the observed distributions for H3K4me3 and H3K27me3 enriched samples show clear excess of extreme values.

and allelic variants; and (iii) it is currently expensive to study entire mammalian genomes. Given these issues, only a handful of whole-genome ChIP-chip studies in mammals have been reported.

In principle, chromatin could be readily mapped across the genome by sequencing ChIP DNA and identifying regions that are over-represented among these sequences. Importantly, sequence-based mapping could require relatively small quantities of DNA and provide nucleotide-level discrimination of similar sequences, thereby maximizing genome coverage. The major limitation has been that high-resolution mapping requires millions of sequences (Supplementary Note). This is cost-prohibitive with traditional technology, even with concatenation of multiple sequence tags⁶. However, recent advances in single molecule-based sequencing (SMS) technology promise to dramatically increase throughput and decrease costs⁷. In the approach developed by Illumina/Solexa, DNA molecules are arrayed across a surface, locally amplified, subjected to successive cycles of primer-mediated single-base extension (using fluorescently-labeled reversible terminators) and imaged after each cycle to determine the inserted base. The ‘read length’ is short (25-50 bases), but tens of millions of DNA fragments may be read simultaneously.

Here, we report the development of a method for mapping ChIP enrichment by sequencing (ChIP-Seq) and describe its application to create chromatin-state maps for pluripotent and lineage-committed mouse cells. The resulting data (1) define three broad categories of promoters based on their chromatin state in ES cells, including a larger than anticipated set of ‘bivalent’ promoters; (2) reveal that lineage commitment is accompanied by characteristic chromatin changes at bivalent promoters that parallel changes in gene expression and transcriptional competence; (3) demonstrate the potential for using ChIP for genome-wide annotation of novel promoters and primary transcripts, active transposable elements, imprinting control regions and allele-specific transcription. This study provides a technological framework for comprehensive characterization of chromatin-state across diverse mammalian cell populations.

Genome-wide chromatin state maps

We created genome-wide chromatin state maps for three mouse cell types: ES cells, neural progenitor cells (NPCs)⁸ and embryonic fibroblasts (MEFs). For each cell type, we prepared and sequenced ChIP DNA samples for some or all of the following features: pan-H3, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H4K20me3 and RNA polymerase II.

In each case, we sequenced nanogram quantities of DNA fragments (~300 bp) on a Solexa 1GGA sequencer. We obtained an average of 10 million successful reads, consisting of the terminal 27-36 bases of each fragment. The reads were mapped to the genome and used to determine the

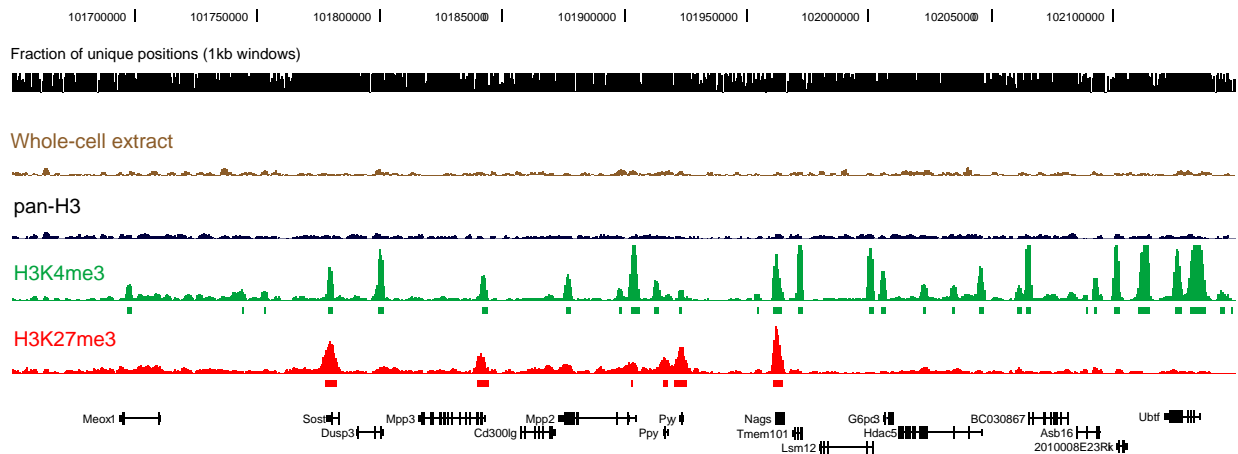


Figure 3. A representative comparison of ChIP-Seq fragment densities across a 500 kb interval on mouse chromosome 1. Rectangles beneath each density plot indicate significantly enriched intervals at the $p < 10e-5$ threshold (see Methods). Black bars at the top indicate the fraction of unique positions within 1kb windows at which ChIP-Seq reads can be uniquely aligned (at $k = 27$, $d = 2$; see Methods).

number of ChIP fragments overlapping any given position (Figure 1). Enriched intervals were defined as regions where this number exceeded a threshold defined by randomization (see Methods). The full data set consists of 18 chromatin-state maps, containing ~140 million uniquely aligned reads, representing over 4 billion bases of sequence.

We validated the chromatin state maps by computational analysis and by comparison to previous methods. ChIP-Seq maps of specific histone modifications show marked enrichment at specific locations in the genome, while the pan-H3 and unenriched samples show relatively uniform distributions (Figure 2, 3). The maps show close agreement with our previously reported ChIP-chip data from ~2.5% of the mouse genome⁹ (Figure 1). Also, ChIP-PCR assays of 50 sites chosen to represent a range of ChIP-Seq fragment counts showed 98% concordance and a strong, quantitative correlation (Figure 4).

Promoter state in ES and lineage-committed cells

We began our analysis by studying H3K4me3 and H3K27me3 patterns at known promoters. H3K4me3 is catalyzed by trithorax-group (trxG) proteins and associated with activation, while H3K27me3 is catalyzed by Polycomb-group (PcG) proteins and associated with silencing^{10,11}. Recently, we and others observed that some promoters in ES cells carry both H3K4me3 and H3K27me3^{9,12}. We termed this novel combination a ‘bivalent’ chromatin mark and proposed that it serves to poise key developmental genes for lineage-specific activation or repression.

We studied 17,762 promoters inferred from full-length cDNAs. Mammalian RNA Polymerase II promoters are known to occur in at least two major forms^{13,14}. CpG-rich promoters are associated with both ubiquitously expressed ‘housekeeping’ genes, and genes with more complex expression patterns, particularly those expressed during embryonic development. CpG-poor promoters are generally associated with highly tissue-specific genes. Accordingly, we divided our analysis to focus on high CpG promoters (HCP; n=11,410) and low CpG promoters (LCP; n=3,014) separately. To ensure a clean separation, we defined a set of intermediate CpG content promoters (ICP; n=3,338); this class shows properties consistent with being a mixture of the two major classes.

High CpG promoters in ES cells. Virtually all HCPs (99%) are associated with intervals of significant H3K4me3 enrichment in ES cells (Figure 5a). The modified histones are typically confined to a punctate interval of 1-2 kb (Figure 6). As observed previously^{15,16}, there is a strong correlation between the intensity of H3K4me3 and the expression level of the associated genes (Spearman’s $\rho=0.67$). However, not all promoters associated with H3K4me3 are active.

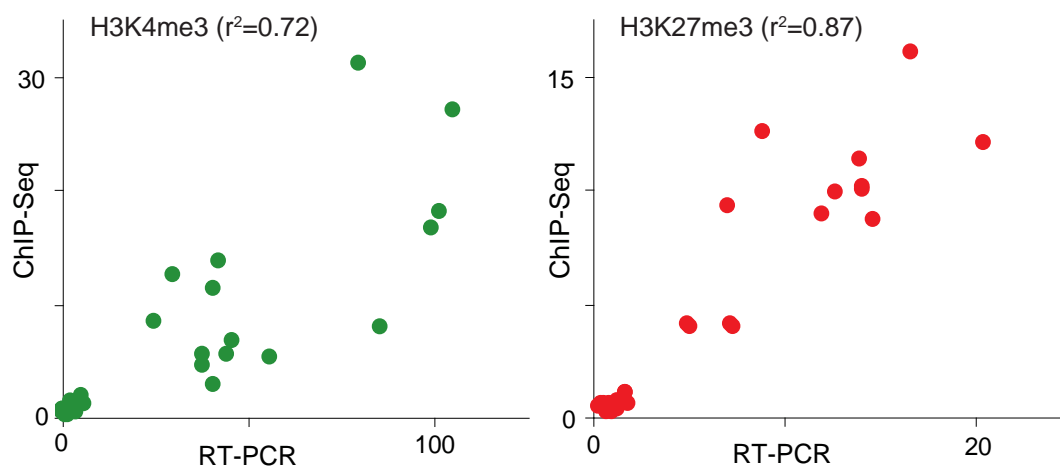


Figure 4. ChIP-Seq fragment densities (y-axis) are plotted against RT-PCR fold-enrichment (x-axis) for H3K4me3 (green) and H3K27me3 (red) at 60 selected sites in mouse ES cells. Notably 28 out of 29 sites (97%) identified as significantly enriched for one of the two modifications by ChIP-Seq were clearly differentiated from unenriched sites by RT-PCR, and 31 of 31 sites (100%) with no ChIP-Seq enrichment had no RT-PCR enrichment either.

The chromatin state maps reveal that ~22% of HCPs (n=2,525) are actually bivalent, exhibiting both H3K4me3 and H3K27me3 (Figure 5a). A minority (n=564) are ‘wide’ bivalent sites in which H3K27me3 extends over a region of at least 5 kb and resemble those described previously⁹. The majority (n=1,961) are ‘narrow’ bivalent sites, with more punctate H3K27me3, that correspond to many additional PcG target promoters¹⁷⁻¹⁹. Bivalent promoters show low activity despite the presence of H3K4me3, suggesting that the repressive effect of PcG activity is generally dominant over the ubiquitous trxB activity (Figure 7).

The different types of chromatin marks at HCP promoters are closely related to the nature of the associated genes. Monovalent promoters (H3K4me3) generally regulate genes with ‘housekeeping’ functions including replication and basic metabolism. By contrast, bivalent promoters are associated with genes with more complex expression patterns, including key developmental transcription factors, morphogens and cell surface molecules. In addition, several bivalent promoters appear to regulate transcripts for lineage-specific microRNAs.

High CpG promoters in NPCs and MEFs. The vast majority of HCPs marked with H3K4me3 alone in ES cells retain this mark both in NPCs and MEFs (92% in each; Figure 5b,5c,8a). This is consistent with the tendency for this sub-class of promoters to regulate ubiquitous housekeeping genes. A small proportion (~4%) of these promoters have H3K27me3 in MEFs, and are thus bivalent or marked by H3K27me3 alone. This correlates with lower expression levels and may reflect active recruitment of PcG proteins to new genes during differentiation²⁰. An example is the transcription factor Sox2, where the promoter is marked by H3K4me3 alone in ES cells and NPCs, but H3K27me3 alone in MEFs. Notably, this locus is flanked by CpG islands with bivalent markings in ES cells (see below), suggesting the locus may be poised for repression upon differentiation.

The majority of HCPs with bivalent marks in ES cells resolve to a monovalent status in the committed cells. In NPCs, 46% resolve to H3K4me3 only and these genes show increased expression (Figure 5b,5d,8b). Of the remaining promoters, 14% resolve to H3K27me3 alone and 32% lose both marks, with both outcomes being associated with low levels of expression. Importantly, 8% remain bivalent and these genes also continue to be repressed (Figure 5b,5d,8c). A somewhat less resolved pattern is seen in MEFs, with 32% marked by H3K4me3 alone, 22% marked by H3K27me3 alone, 3% without both marks, and the remaining (43%) still bivalent (Figure 5c). The relatively high number of bivalent promoters in MEFs may reflect a less differentiated state and/or heterogeneity in the population.

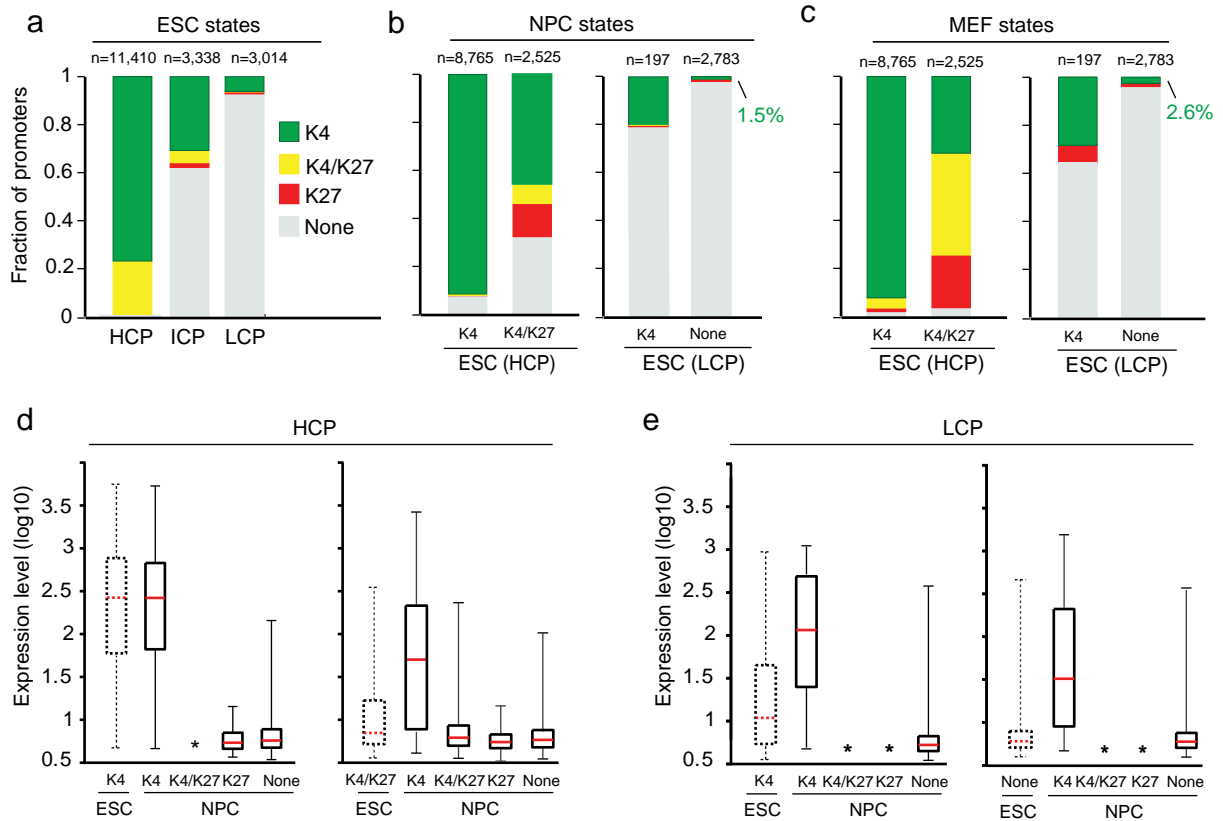


Figure 5. Histone tri-methylation state predicts expression of HCP and LCP promoters. (a) Mammalian promoters can be readily classified into sets with high (HCPs), intermediate (ICPs) or low (LCPs) CpG-content. In ES cells (ESCs), virtually all HCPs are marked by H3K4me3, either alone (green) or in combination with H3K27me3 (yellow). In contrast, most LCPs have neither mark (grey). Few promoters are only enriched for H3K27me3 (red). (b) Tri-methylation states of HCPs and LCPs in NPCs (indicated by colors), conditional on their ESC state (indicated below each bar). HCPs marked by H3K4me3 only in ESCs tend to retain this mark. HCPs marked by H3K4me3 and H3K27me3 tend to lose one or both marks, although some remain bivalent. Small, partially overlapping subsets of LCPs are marked by H3K4me3. (c) Tri-methylation states of HCPs and LCPs in MEFs. (d) Changes in expression levels of HCP genes with H3K4me3 alone (left) or also with H3K27me3 (right) upon differentiation to NPCs. Resolution of bivalent promoters to H3K4me3 is associated with increased expression. Boxplots show median (red bar), 25th and 75th percentile expression levels in ESCs. Whiskers show 2.5th and 97.5th percentiles. Asterisks indicate classes with less than 15 genes. (e) Changes in expression levels of LCP genes with H3K4me3 (left) or no mark (right) upon differentiation to NPCs. Gain of H3K4me3 is associated with increased expression.

Distinct regulation of Low CpG Promoters. The LCPs show a strikingly different pattern than the HCPs. Only a small minority (6.5%, n=207) of LCPs have significant H3K4me3 in ES cells and virtually none have H3K27me3 (Figure 5a). Most of these promoters have lost H3K4me3 in NPCs and MEFs, while a small number of other LCPs (1.5% and 2.6%, respectively) have gained the mark (Figure 5b,5c,8e). In all three cell types, the expression levels of the associated genes strongly correlate with presence or absence of H3K4me3 (Figure 5e, 7).

The genes with LCPs marked by H3K4me3 are closely related to tissue-specific functions. In NPCs, they include genes encoding several known markers of neural progenitors *in vivo* (such as *Fabp7*, *Cp*, *Gpr56*). In MEFs, they include genes encoding extracellular matrix components and growth factors (such as *Col3a1*, *Col6a1*, *Postn*, *Aspn*, *Hgf*, *Figf*), consistent with the mesenchymal origin of these cells (see below).

We conclude that HCPs and LCPs are subject to distinct modes of regulation. In ES cells, all HCPs appear to be targets of *trxG* activity, and may therefore drive transcription unless actively repressed by PcG proteins. In committed cell types, a subset of HCPs appear to lose the capacity to recruit *trxG* activity (possibly due to other epigenetic modifications, such as DNA methylation²¹). In contrast, CpG-poor promoters appear to be inactive by default, independent of repression by PcG proteins, and may instead be selectively activated by cell type- or tissue-specific factors.

Alternative promoter use. We note that genes with alternative promoters may have multiple, distinct chromatin states. An ‘active’ state at any one of these may be sufficient to drive expression. A common situation involves genes with one major HCP and one or more alternative LCPs. An example is the transcription factor *Foxp2*, which is expressed at moderate levels in both NPCs and MEFs (Figure 8f,g). The *Foxp2* HCP is marked by H3K4me3 in NPCs, but is bivalent in MEFs. However, an alternative LCP is marked by H3K4me3 exclusively in MEFs. The protocadherin- γ (*Pcdh- γ*) locus is a more extreme case: the N-terminal variable regions of this gene are transcribed from at least 20 different HCPs in neurons²², all of which carry bivalent chromatin marks in ES cells. *Pcdh- γ* expression is nevertheless detected by microarrays, possibly due to a single promoter in front of the C-terminal constant region marked by H3K4me3 alone (Figure 8h).

Although only ~10% of the genes analyzed here have more than one known promoter, recent ‘cap-trapping’ studies suggest that alternative promoter use may be substantially more common²³. The ability of ChIP-Seq to assess chromatin state at known promoters, as well as to identify novel promoters (see below), should prove valuable in analysis of transcriptional networks.

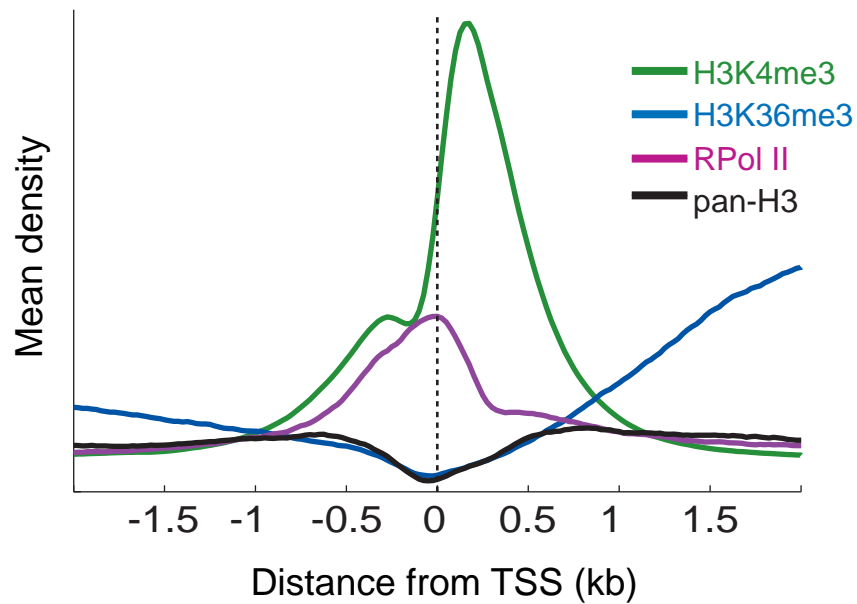


Figure 6. Composite profile of HCP promoters. Plots show mean ChIP-Seq fragment densities (scaled for comparison) of H3K4me3, H3K36me3, pan-H3 and RNA Polymerase II ChIP-Seq fragments over all analyzed high-CpG promoters. H3K4me3 marks a punctate interval peaking just downstream of the H3-depleted transcription start site, which is occupied by RNA Polymerase II. H3K36me3 marks begin roughly where H3K4me3 ends. H3K4me3 and H3K36me3 increasing in the negative direction likely represent bidirectional promoter activity.

Promoter state reflects lineage commitment and potential

Given their association with epigenetic memory, we next examined whether the patterns of H3K4me3 and H3K27me3 can reflect developmental potential. Both of the committed cell types studied here have been shown to be multipotent *ex vivo*. NPCs can be differentiated to glial and neuronal lineages⁸, while primary MEFs have been differentiated into adipocytes²⁴, chondrocytes²⁵ and osteoblast-like cells²⁶.

Lineage-specific resolution and retention of bivalent marks. We first examined a set of genes involved in *in vivo* differentiation pathways known to be, at least partially, recapitulated by MEFs, NPCs or neither. These genes all have bivalent promoters in ES cells. We found that their resolution in lineage-committed cells is closely related to their demonstrated developmental potential:

- Genes restricted to regulation or specialized functions in unrelated lineages, such as hematopoietic (Cdx4, PU.1), epithelial (Cncf, Krt2-4), endoderm (Gata6, Pdx1) or germ line (Tennr, Ctcf1), have generally resolved to monovalent H3K27me3 or carry neither mark in both NPCs and MEFs.

- Genes related to adipogenesis and chondro/osteogenesis often remain bivalent in MEFs, but not in NPCs. Examples include Ppar- γ , which is a key regulator of adipogenesis, and Sp7, which promotes chondro/osteogenic pathways. Early mesenchymal markers, such as Runx1 and Sox9 resolved to H3K4me3 alone in MEFs.

- Genes related to gliogenesis and neurogenesis often resolve to H3K4me3 alone or remain bivalent in NPCs, while resolving to H3K27me3 alone in the MEFs. Gliogenesis and neurogenesis are thought to be mutually opposing pathways²⁷, and we find that genes promoting gliogenesis are more likely to resolve to H3K4me3 in NPCs. Examples include Bmp2 and the miRNA mir-9-3, which promotes glial but inhibits neuronal differentiation²⁸. Several genes known to promote neuronal differentiation, such as Neurog1 and Neurog2, remain bivalent while others, such as Bmp6, appear to resolve to H3K27me3 alone. The NPCs differentiate to astrocytes with significantly higher efficiency than to neurons (M. Wernig, unpublished data). The observed chromatin patterns may reflect this gliogenic bias.

Correlation with expression in adult tissues. We next analyzed gene expression in adult tissues with major contributions from neuroectodermal or mesenchymal lineages. We reasoned that if H3K4me3 is generally not restored once lost, then differential loss of H3K4me3 at promoters early in these lineages (as represented by NPCs and MEFs, respectively) might be reflected in differential gene expression patterns in related adult tissues.

Strikingly, we observed a clear bias in relative expression levels between relevant adult tissues for genes that retain H3K4me3 in NPCs only versus genes that retain H3K4me3 in MEFs only. The former are strongly biased toward higher expression in various brain sections, while the latter are biased towards higher expression in bone, adipose and other mesenchyme-rich tissues (Figure 9).

These analyses are of course limited by alternative promoter usage, the cell models used, and the heterogeneity of the adult tissues. Nonetheless, the data show clear trends that support an important role for retention and resolution of bivalent chromatin in the regulation of hierarchical lineage commitment.

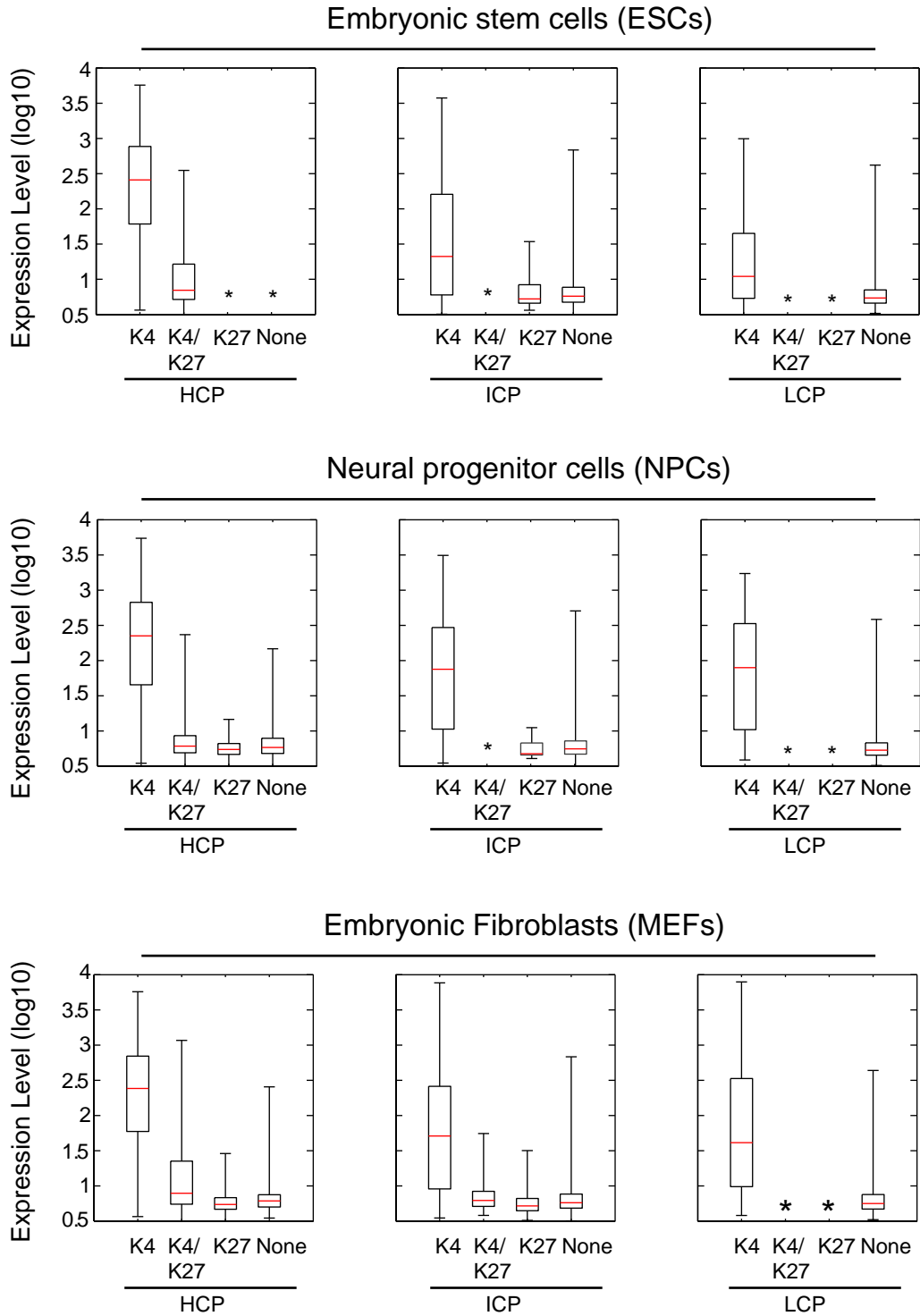


Figure 7. Boxplots showing the distributions of expression levels for genes in ES cells, NPCs and MEFs, according to promoter class and state. Red bar is median; box shows 25th and 75th percentiles; whiskers show 2.5th and 97.5th percentiles. Asterisks indicate class/state combinations with less than 15 genes.

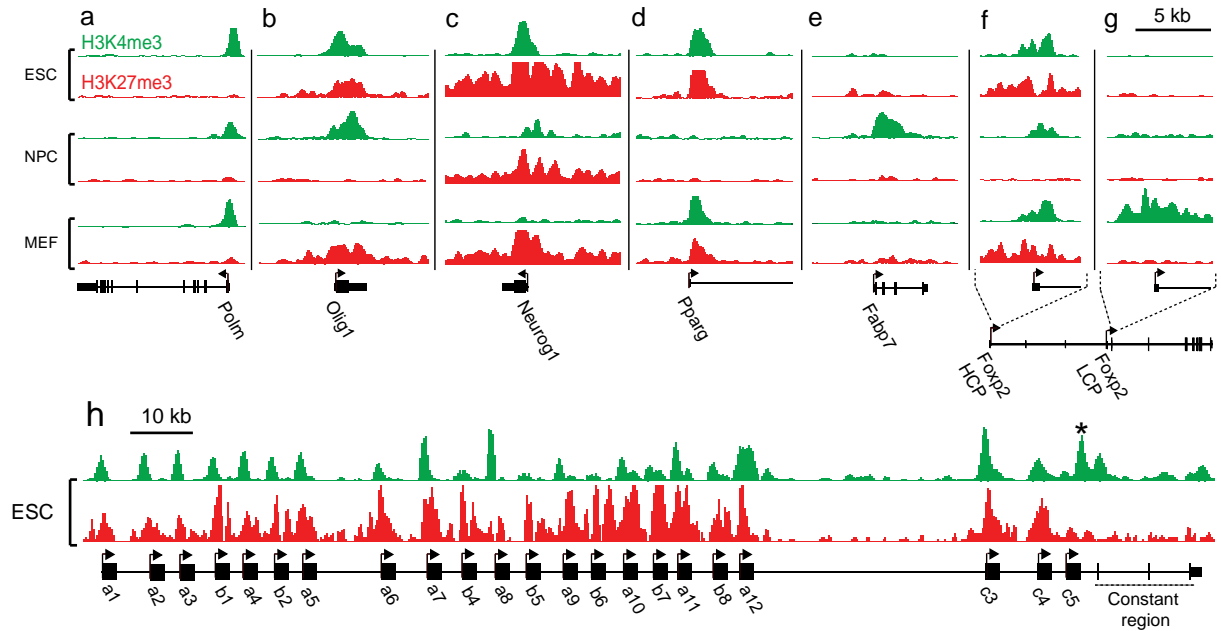


Figure 8. Cell type-specific chromatin marks at promoters. (a) Multiple ‘housekeeping genes’, such as DNA Polymerase mu (Polm), are associated with HCPs marked by H3K4me3 in all cell types. (b) The neural transcription factor gene Olig1 (HCP) is bivalent in ESCs, but resolves to H3K4me3 in NPCs and H3K27me3 in MEFs. (c) The neurogenesis transcription factor gene Neurog1 (HCP) remains bivalent upon differentiation to NPCs, but resolves to H3K27me3 in MEFs. (d) The adipogenesis transcription factor gene Ppar-? (HCP) remains bivalent in MEFs, but loses both marks in NPCs. (e) The neural progenitor marker gene Fabp7 (LCP) is marked by H3K4me3 in NPCs only. (f) The brain and lung expressed transcription factor gene Foxp2 is associated with an HCP that is bivalent in ES cells, but resolves to H3K4me3 in NPCs and remains bivalent in MEFs. (g) Foxp2 also has an LCP marked by H3K4me3 in MEFs only. (h) Multiple, distinct bivalent chromatin marks at the variable region promoters of protocadherin-gamma. A promoter proximal to the constant region exons (*) is marked by H3K4me3 only.

Genome-wide annotation of promoters and primary transcripts

We next considered genome-wide maps of H3K36me3. This mark has been linked to transcriptional elongation and may serve to prevent aberrant initiation within gene bodies²⁹⁻³³. Our chromatin maps reveal a global pattern of H3K36me3 in mammals similar to that previously observed in yeast²⁹.

In all three cell types, H3K36me3 is strongly enriched across the transcribed regions of active genes (Figure 10a), beginning immediately after the promoter H3K4me3 signal. The level of H3K36me3 is strongly correlated with the level of gene expression (Spearman's $\rho=0.77$), although the dynamic range is compressed (1-2 orders of magnitude for H3K36me3 vs 3-4 for expression levels; Figure 11). Genes with bivalent promoters rarely show H3K36me3, consistent with their low expression. Notably, there is essentially no overlap between intervals significantly enriched for H3K36me3 and for H3K27me3, consistent with a role for PcG complexes in the exclusion of polymerases¹¹.

The vast majority of intervals significantly enriched for H3K36me3 is associated with known genes (~92% in ESCs), but there are at least ~500 additional regions across the genome (median size ~2 kb), with most being adjacent to sites of H3K4me3. Inspection revealed a number of interesting cases, falling into three categories.

The first category corresponds to H3K36me3 that extends significantly upstream from the annotated start of a known gene, often until an H3K4me3 site. These appear to reflect the presence of unannotated alternate promoters. A notable example is the *Foxp1* locus. In ES cells, one annotated *Foxp1* promoter is marked by H3K4me3 and another CpG-rich region located ~500 kb upstream carries a bivalent mark. In MEFs, this CpG island is marked by H3K4me3 only, and H3K36me3 extends from this site to the 3' end of *Foxp1* (Figure 10a). Although no transcript extending across this entire region has been reported in mouse, the orthologous position in human has been shown to act as a promoter for the orthologous gene. The ChIP-Seq data contain many other examples where the combination of H3K36me3 and H3K4me3 appear to reveal novel promoters.

The second category corresponds to H3K36me3 that extends significantly downstream of a known gene. An example is the *Sox2* locus, which encodes a pluripotency-associated transcription factor that also functions during neural development. In ES cells, *Sox2* has an unusually large region of H3K4me3 (>20 kb) accompanied by H3K36me3 extending far beyond the annotated 3'-end (>15 kb); non-coding transcription throughout the locus has been noted previously³⁴ and may serve a regulatory role (Figure 10b).

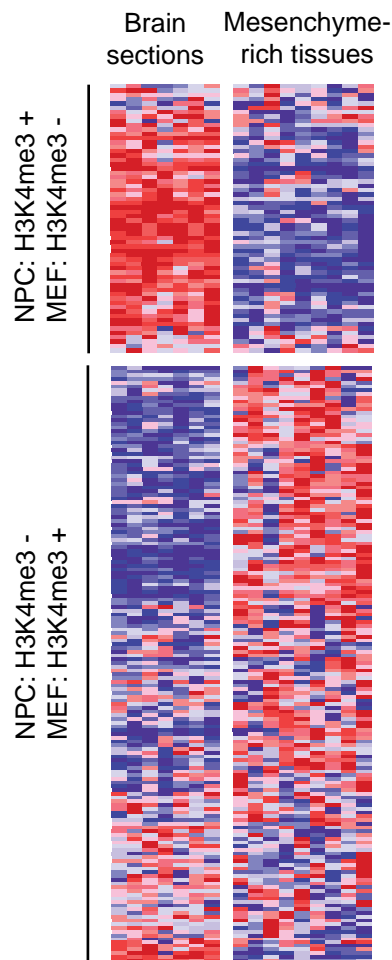


Figure 9. Correlation between chromatin state changes and lineage expression. Relative expression levels across adult mouse brain (frontal and cerebral cortex, substantia nigra, cerebellu, amygdale, hypothalamus, hippocampus) and relatively mesenchyme-rich tissues (bone, white fat, brown fat, trachea, digits, lung, bladder, uterus, umbilical cord) are shown for genes with bivalent chromatin marks in ES cells that retain H3K4me3 in NPCs but lose this mark in MEFs (n=62) or vice versa (n=160). Red, white and blue indicates higher, equal and lower relative expression, respectively.

The third category appears to reflect transcription of non-coding RNA genes. For example, two regions with H3K36me3 and adjacent H3K4me3 correspond to recently discovered nuclear transcripts with possible functions in mRNA processing³⁵ (Figure 10c). In addition, a number of these presumptive transcriptional units overlap microRNAs (Figure 10d). A striking example is a >200 kb interval within the Dlk1-Dio3 imprinted locus (Figure 12a). This region harbors over 40 non-coding RNAs, including clusters of microRNAs and small nucleolar RNAs³⁶. The CHIP-Seq data suggest that the entire region is transcribed as a single unit that initiates at an H3K4me3 marked HCP.

These findings suggest that genome-wide maps of H3K4me3 and H3K36me3 may provide a general tool for defining novel transcription units. The capacity to define the origins and extents of primary transcripts will be of particular value for characterizing the regulation of microRNAs and other non-coding RNAs that are rapidly processed from long precursors³⁷. Finally, the relatively narrow dynamic range of H3K36me3 may offer advantages over RNA-based approaches in assessing gene expression and defining cellular states.

H3K9 and H4K20 tri-methylation associated with specific repetitive elements

We next studied H3K9me3 and H4K20me3, both of which have been associated with silencing of centromeres, transposons and tandem repeats³⁸⁻⁴⁰. We sought first to assess the relative enrichments of H3K9me3 and H4K20me3 across different types of repetitive elements by aligning CHIP-Seq reads directly to consensus sequences for various repeat families (~40 million reads could be aligned this way).

H3K9me3 and H4K20me3 show nearly identical patterns of enrichment in ES cells. The strongest enrichments are observed for telomeric, satellite, and long terminal repeats (LTRs). The LTR signal primarily reflects enrichment of intracisternal A-particles (IAP), early transposon (ETn) elements, and the LTRIS sub-family (Figure 13).

IAP and ETn elements are active in murine ES cells and produce double-stranded RNAs^{41,42}. RNA has also been implicated in maintaining satellite and telomeric heterochromatin³⁸. Hence, these enrichment data are consistent with a global role for RNA in targeting repressive chromatin marks in mammalian ES cells, analogous to that observed in lower eukaryotes^{38,39}.

We next examined the distributions of H3K9me3 and H4K20me3 across unique sequence in the mouse genome. We identified ~1800 H3K9me3 sites (median size ~300 bp) in ES cells, with the vast majority also showing H4K20me3. Fully 78% of the sites lie within two kb of a satellite

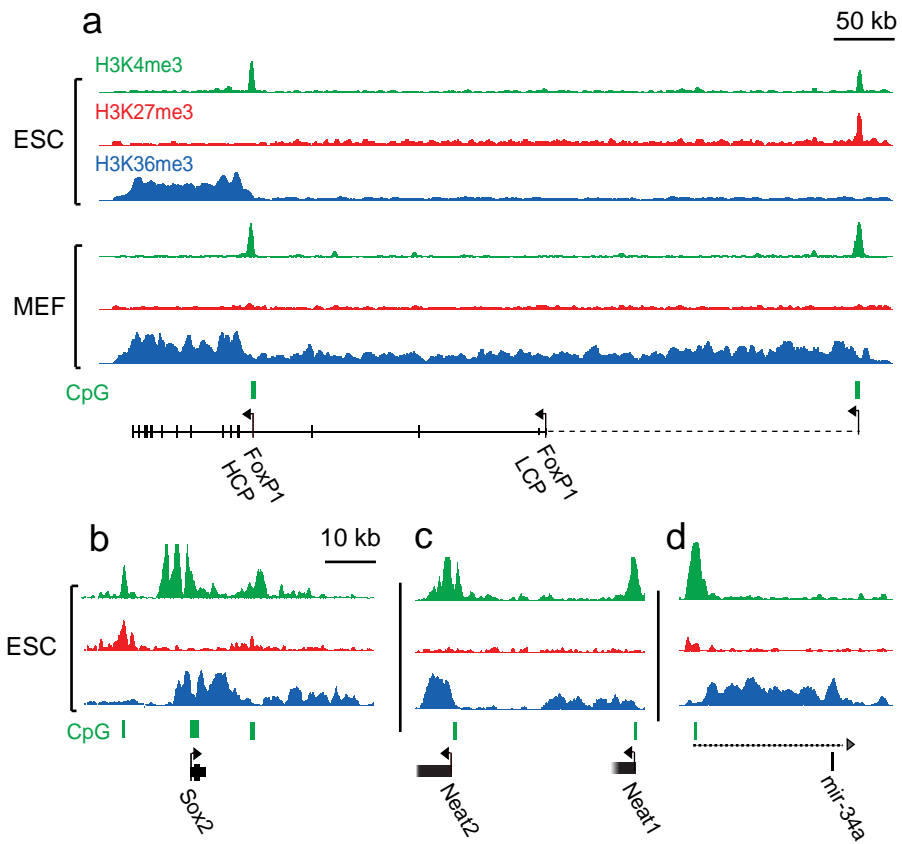


Figure 10. H3K4me3 and H3K36me3 annotate genes and non-coding RNA transcripts. (a) *Foxp1* has two annotated promoters (based on RefSeq and UCSC Known Genes), only one of which shows H3K4me3 in ES cells. The corresponding transcriptional unit is marked by H3K36me3. In MEFs, H3K36me3 extends an additional 500 kb upstream to an H3K4me3 site that appears to reflect an alternate promoter (this site is bivalent in ES cells). (b) H3K36me3 enrichment extends significantly downstream of *Sox2*. Though highly active in ES cells, *Sox2* is flanked by two bivalent CpG islands that may poise it for repression. (c) H3K4me3 and H3K36me3 indicate two highly expressed non-coding RNAs, and (d) the putative primary transcript (dashed line) for a single annotated microRNA.

repeat or LTR (primarily IAP and ETn elements). This suggests that repressive marks are capable of spreading from repeat insertions and could potentially regulate proximal unique sequence.

Recent studies have described a handful of active genes with H3K9me3 and H4K20me3, raising the possibility that these ‘repressive’ marks also function in transcriptional activation^{31,32}. One-third of the ~1800 H3K9me3 enriched sites reside within an annotated gene, which is roughly the proportion expected by chance. However, H3K9me3 sites that are larger and/or more distant from LTRs are more likely to occur within genes. The largest genic site in ES cells (~6 kb) coincides with the *Polrmt* gene (Figure 12d). This case is notable because the downstream gene (*Hcn2*) is convergent and contains a CpG island at its 3’ end. Transcription from 3’ promoters has been proposed as a potential mechanism of transcriptional interference by producing antisense transcripts²³. This example may therefore reflect a link between transcriptional interference and H3K9me3, as has been suggested for a few other mammalian loci^{43,44}. Our results thus confirm that the presence of H3K9me3 within genes is a general phenomenon, although the functional implications remain to be elucidated.

Imprinting control regions show overlapping H3K4 and H3K9 tri-methylation

We next studied chromatin marks associated with imprinting. This epigenetic process typically involves allele-specific DNA methylation of CpG-rich imprinting control regions (ICRs)⁴⁵. Several reports have also described allele-specific chromatin modification at a handful of ICRs, with H3K9me3 and H4K20me3 on the DNA methylated allele and H3K4me3 on the opposite allele^{46,47}.

We searched for regions showing overlapping H3K9me3 and H3K4me3 in ES cells. Strikingly, 13 of the top 20 sites, as ranked by enrichment of the two marks, are located within known imprinted regions, coincident with ICRs or imprinted gene promoters. An example is the *Peg13* promoter (Fig. 12c). Conversely, of the ~20 known and putative autosomal imprinted loci that contain ICRs, 17 have at least one with the overlapping chromatin marks. We conclude that overlapping H3K9me3 and H3K4me3 is a common signature of ICRs in ES cells.

Allele-specific histone methylation

To explore the feasibility of inferring allele-specific chromatin states, we constructed chromatin-state maps in male ES cells derived from a more distant cross (129 (maternal) x *M. castaneus* (paternal)), and used a catalog of ~3.5 million SNPs to assign ChIP-Seq reads to one of the two parental alleles.

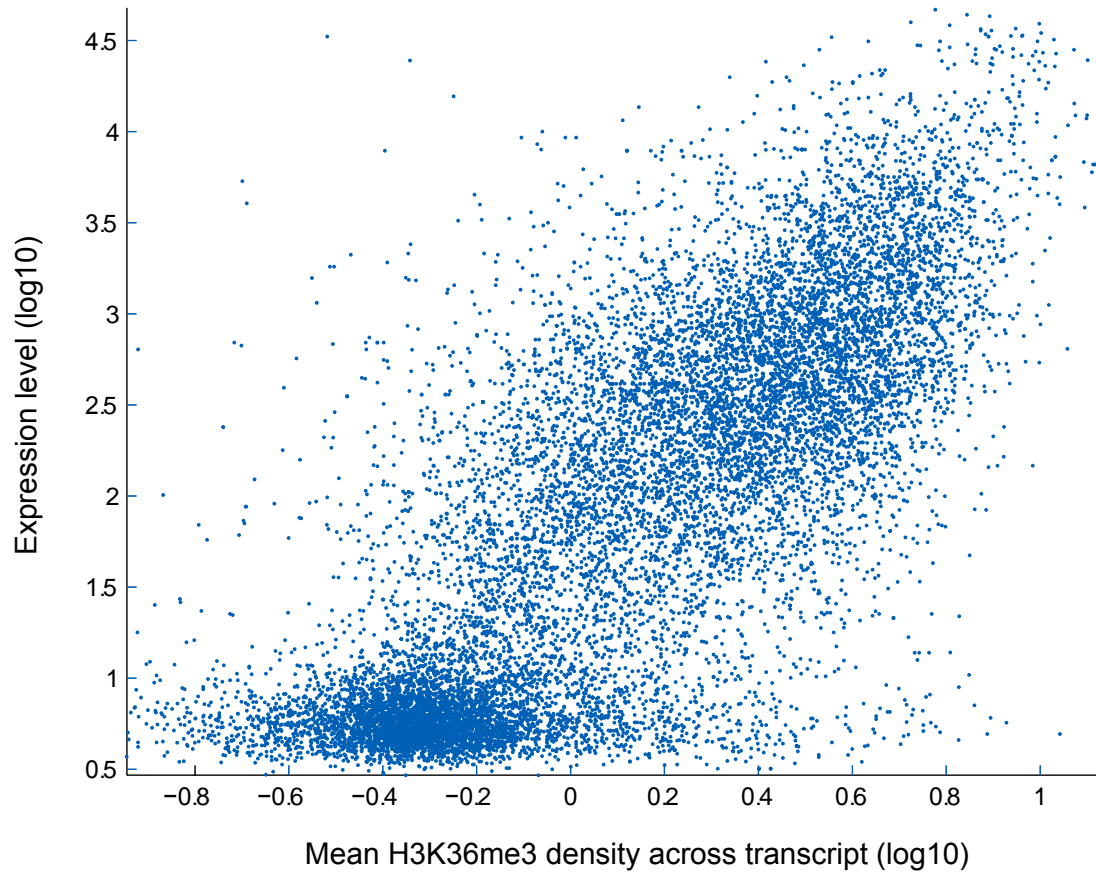


Figure 11. Scatter plot of H3K36me3 density across transcripts versus their expression levels as measured by Affymetrix GeneChips. The lower left-hand cluster corresponds to largely inactive genes. The range of H3K36me3 densities across most actively expressed genes spans ~ 1 order of magnitude, compared to ~ 3 for expression levels.

As a positive control, we first compared results for chromosome X and the autosomes for reads derived by H3K4me3 ChIP. Virtually all (97%) of ~3700 informative reads on chromosome X, and roughly half (57%) of the 178,000 informative reads on the autosomes, were assigned to the 129 strain. These proportions correspond roughly to the expected 100% and 50%.

We then examined the allelic distribution at overlapping H3K4me3 and H3K9me3 sites coincident with putative ICRs (see above). Six of the ICRs had enough reads (≥ 10) containing SNPs to assess allelic bias. In every case, the SNPs showed significant bias in the expected direction ($p < 0.02$; Figure 12c).

We applied the same approach to search for allelic imbalance in intervals with significant H3K36me3 enrichment, which would predict differential transcription of the two alleles. A striking interval corresponds to a microRNA cluster within the *Dlk1-Dio3* locus known to be imprinted in the embryo proper³⁶ (Fig. 12a-b). Of the additional imprinted genes with H3K36me3 enrichment, four (*Snrpn*, *Grb10*, *Impact*, *Peg3*) had enough reads containing SNPs to assess allelic bias. In every case, the data showed significant bias in the expected direction ($p < 0.02$). The data also revealed novel instances of allele-specific transcription. For example, a transcript of unknown function (BC054101), first identified in trophoblast stem cells⁴⁸, showed highly significant maternal bias for H3K36me3, as well as H3K4me3 ($p < 10^{-15}$; Figure 14).

The results suggest that, with sufficiently deep coverage and dense SNP maps, ChIP-Seq will provide a powerful means for identifying allele-specific chromatin modifications. With data from reciprocal crosses, it should be possible to discriminate novel cases of imprinting from strain-specific differences.

Conclusion

Genome-wide chromatin-state maps provide a rich source of information about cellular state, yielding insights beyond what is typically obtained by RNA expression profiling. Analysis of H3K4me3 and H3K36me3 allows recognition of promoters together with their complete transcription units. This should help define alternative promoters and their usage in specific cell types; identify the structure of genes encoding non-coding RNAs; detect gene expression (given the narrower dynamic range); and detect detecting allele-specific transcription. In addition, analysis of H3K9me3 and H4K20me3 should facilitate study of heterochromatin, spreading and imprinting mechanisms.

Most interestingly, analysis of H3K4me3 and H3K27me3 provides a rich description of cellular state. Our results suggest that promoters may be classified as active, repressed or poised for

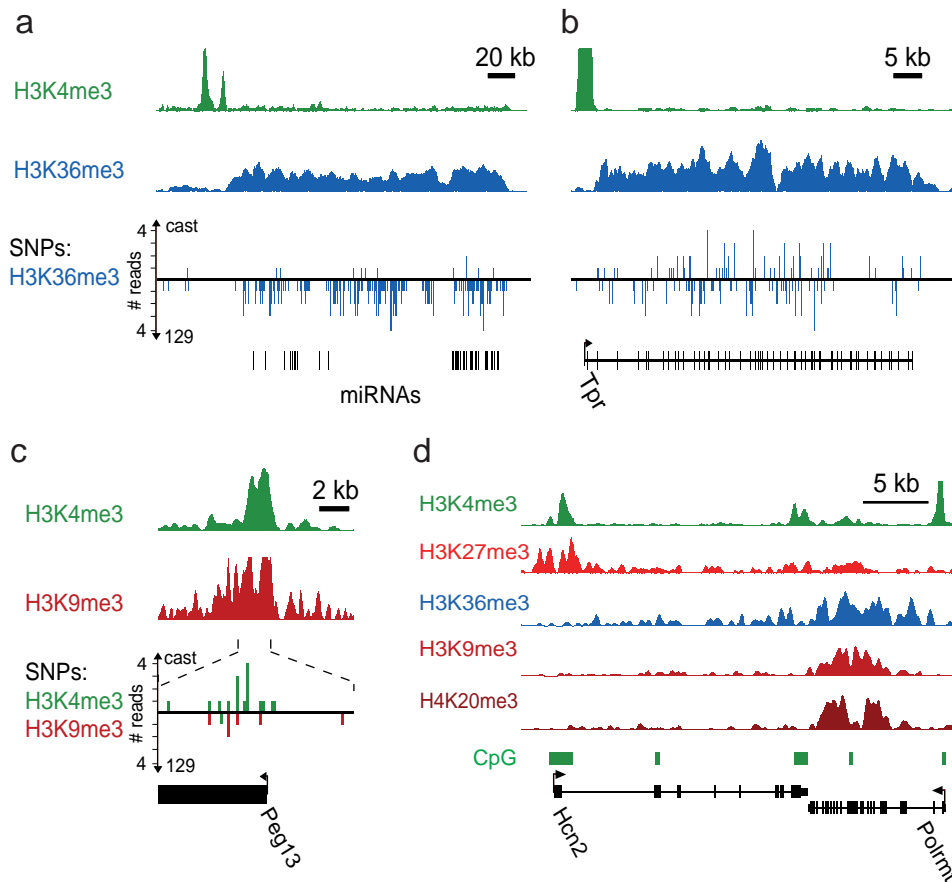


Figure 12. Allele-specific histone methylation and genic H3K9me3/H4K20me3 (a) H3K4me3 and H3K36me3 indicate a primary microRNA transcript in the *Dlk1-Dio3* locus. The allele-specificity of this transcript is read out using ChIP-Seq data for hybrid ES cells and a SNP catalogue. The H3K36me3 reads overwhelmingly correspond to maternal 129 alleles, consistent with the known maternal expression of these microRNAs³⁶. (b) In contrast, a non-imprinted transcript shows roughly equal proportions of reads assigned to 129 and castaneus alleles. (c) *Peg13* is marked by H3K4me3 and H3K9me3 in ES cells; 19 of 21 H3K4me3 reads correspond to the paternal castaneus allele, while 6 of 6 H3K9me3 reads correspond to the maternal 129 allele, consistent with paternal expression of this gene. (d) H3K9 me3 and H4K20me3 enrichment evident at the *Polrmt* gene may reflect transcriptional interference due to antisense transcription from the 3' UTR CpG island of *Hcn2* (see text).

alternative developmental fates. Conceivably, chromatin state at key regulatory genes may suffice to describe developmental commitment and potential.

Given the technical features of ChIP-Seq (high throughput, low cost and input requirement), it is now appropriate to contemplate projects to generate catalogs of chromatin-state maps representing a wide range of human and mouse cell types. These should include varied developmental stages and lineages, from totipotent to terminally differentiated, with the aim of precisely defining cellular states at the epigenetic level and observing how they change over the course of normal development. Chromatin-state maps should also be systematically cataloged from situations of abnormal development. Cancer cells are the most obvious targets, as they are frequently associated with epigenetic defects and many appear to have acquired characteristics of earlier developmental stages. A comprehensive public database of chromatin-state maps would be a valuable resource for the scientific community.

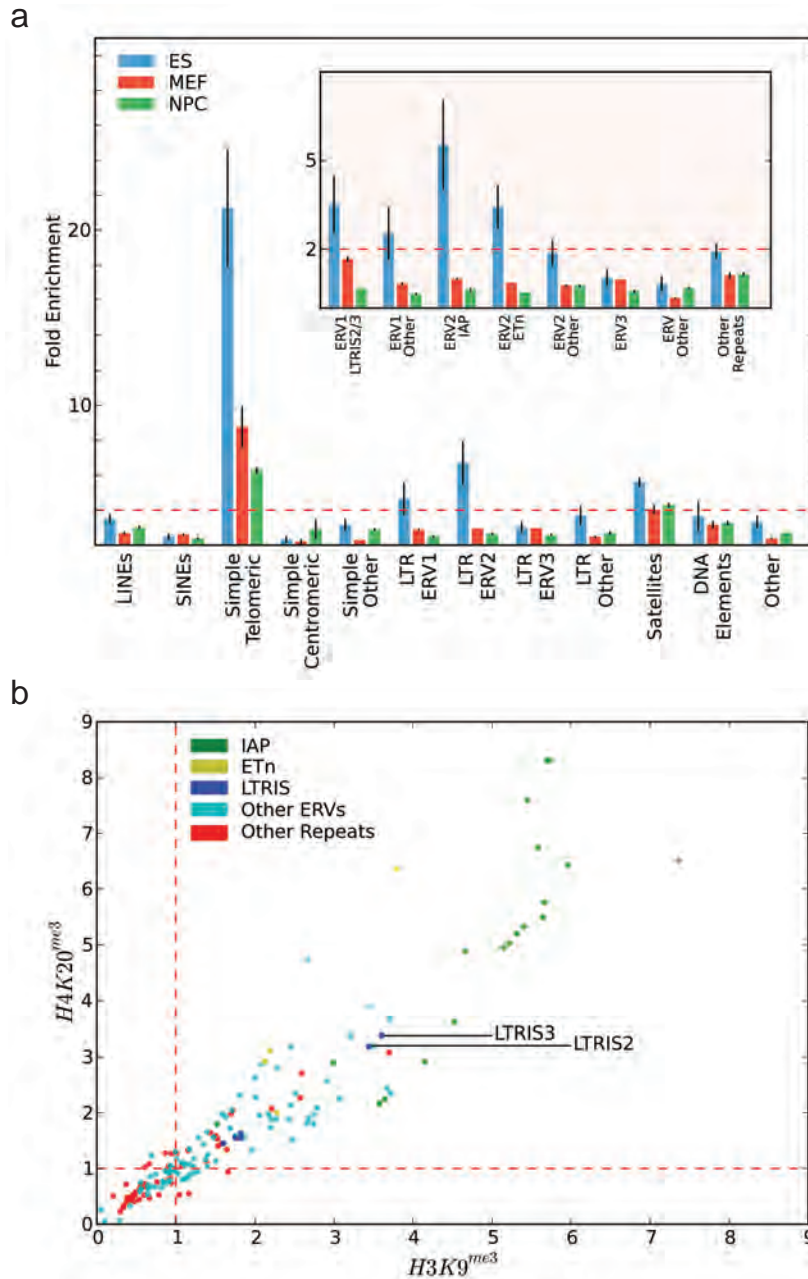


Figure 13. (a) Simple telomeric repeats, satellite repeats, and class II endogenous retroviruses (LTR ERV2) all show significant enrichment for H3K9me3 in ES cells (blue). A weaker signal is seen for class I endogenous retroviruses (LTR ERV1). Both ERV1 and ERV2 elements lose the H3K9me3 marking in MEFs (red) and NPCs (green). The dashed line indicates twofold enrichment; values below 1 indicate depletion. Error bars show the difference in signal observed between sample runs. Inset: Intracisternal A particles (ERV2 IAP) and Early Transposon associated elements (ERV2 ETn), both known to be active in mice, are largely responsible for the enrichment of H3K9me3 observed for ERV2s in ES cells. **(b)** H3K9me3 and H4K20me3 exhibit similar distribution at repeats, and are strongly enriched in active ERVs.

Methods

Cell Culture. V6.5 murine ES cells (genotype 129SvJae x C57BL/6; male; passages 10–15) and hybrid murine ES cells (genotype 129SvJae x *M. castaneus* F1; male; passages 4–6) were cultivated in 5% CO₂ at 37° on irradiated MEFs in DMEM containing 15% FCS, leukemia-inhibiting factor, penicillin/streptomycin, L-glutamine, nonessential amino acids and 2-mercaptoethanol. Cells were subject to at least two to three passages on 0.2% gelatin under feeder-free conditions to exclude feeder contamination. V6.5 ES cells were differentiated into neural precursor cells (NPCs) through embryoid body formation for 4 days and selection in ITSFn media for 5–7 days, and maintained in FGF2 and EGF2 (R&D Systems) as described⁸. The cells uniformly express nestin and Sox2 and can differentiate into neurons, astrocytes and oligodendrocytes. Mouse embryonic fibroblasts (genotype 129SvJae x C57BL/6; male; d13.5; passages 4–6), were grown in DMEM with 10% fetal bovine serum and penicillin/streptomycin at 37°, 5% CO₂.

Chromatin Immunoprecipitation (ChIP). ChIP experiments were carried out as described in Bernstein et al., 2005 and at www.upstate.com. Briefly, chromatin from fixed cells was fragmented to a size range of 200 to 700 bases with a Branson 250 Sonifier or a Diagenode Bioruptor. Solubilized chromatin was immunoprecipitated with antibody against H3K4me3 (Abcam #8580), H3K9me3 (Abcam #8898), H3K27me3 (Upstate #07-449), H3K36me3 (Abcam #9050), H4K20me3 (Upstate #07-463), pan-H3 (Abcam #1791) or RNA polymerase II (Covance MMS-126R). Antibody-chromatin complexes were pulled-down using Protein A-sepharose (or anti-IgM conjugated agarose for RNA polymerase II), washed and then eluted. After cross-link reversal and Proteinase K treatment, immunoprecipitated DNA was extracted with phenol-chloroform, ethanol precipitated, and treated with RNase. ChIP DNA was quantified using PicoGreen.

Library Preparation and Solexa sequencing. One to ten nanograms of ChIP DNA (or unenriched whole cell extract) were prepared for Solexa sequencing as follows: DNA fragments were repaired to blunt ends by T4 DNA polymerase and phosphorylated with T4 Polynucleotide kinase using the END-IT kit (Epicentre). Then, a single ‘A’ base was added to 3’ ends with Klenow (3’→5’ exo⁻, 0.3 U/μl). Double-stranded Solexa adaptors (75 bp with a ‘T’ overhang) were ligated to the fragments with DNA ligase (0.05 U/μl). Ligation products between 275 and 700 base pairs were gel purified on 2% agarose to remove unligated adaptors, and subjected to 18 PCR cycles. Completed libraries were quantified with PicoGreen.

DNA sequencing was carried out using Illumina’s Solexa sequencing system. Cluster amplification, linearization, blocking and sequencing primer reagents were provided in the Solexa

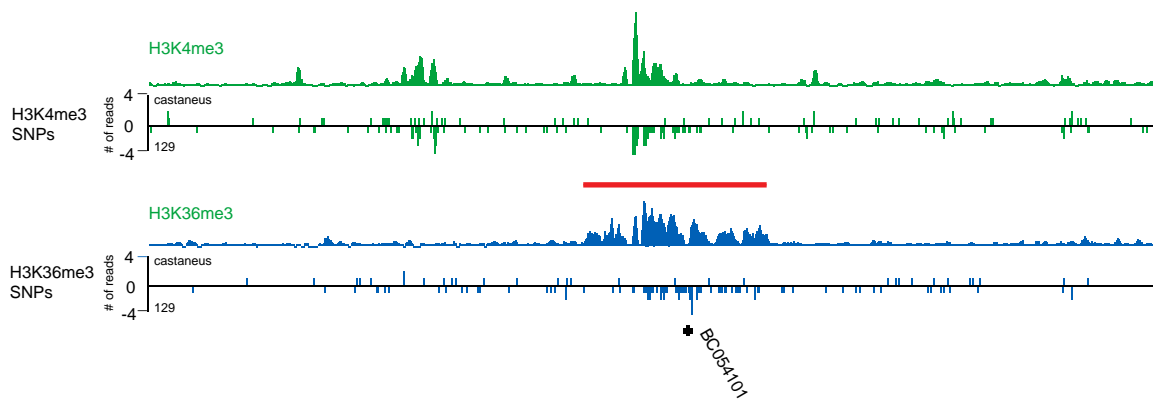


Figure 14. A transcribed region that exhibits allelic bias in the ChIP-Seq data from 129/castaneus hybrid ES cells is a candidate for imprinted or strain-specific expression. An interval of H3K36me3 enrichment (red bar) overlapping the transcript of unknown function BC054101. Of the 69 aligned reads within the enriched interval at the center of the locus, 64 were classified as 129 (maternal). Of the 61 aligned H3K4me3 reads in the same interval, 59 were classified as 129. This suggests near exclusive maternal transcription.

Cluster Amplification kits and were used according to the manufacturer's specifications as described here. To obtain single strand templates, the sample prep was first denatured in NaOH (0.1N final concentration) and diluted in Solexa hybridization buffer (4°C) to a final concentration of either 2 or 4pM. Sample loading was carried out as follows. A template sample was loaded into each lane of a Solexa flowcell mounted on a Solexa cluster station on which all subsequent steps were performed. The temperature was increased to 95°C for 1min and slowly decreased to 40°C to allow for annealing onto complementary adapter oligos on the flowcell surface. Cluster formation was then carried out as follows. The template strands were extended with Taq polymerase (0.25U/ul) to generate a fixed copy of the template on the flowcell. The samples were then denatured with formamide (Sigma-Aldrich, F-5786, >99.5% (GC)) and washed (Solexa Wash buffer) to remove the original captured template leaving behind a single stranded template ready for amplification. Clusters were then amplified under isothermal conditions (60°C) for 30 cycles using Solexa Amplification mix containing Bst I DNA polymerase (0.08U/ul). After each amplification cycle, the templates were denatured with formamide (as above). Fresh amplification mix was added after each denaturation step. Following amplification, the clusters were linearized with Solexa Linearization mix, and any unextended flowcell surface capture oligos were blocked with ddNTPs (2.4uM mix in the presence of 0.25U/ul terminal transferase). The linearized clusters were then denatured (0.1N NaOH) to remove and wash away the linearized strands. The single-stranded templates in the cluster were then annealed with the Solexa sequencing primer (10uM). The flowcells were removed from the cluster station and then transferred onto the 1G Genetic Analyzer which performed the sequencing according to its own standard protocols. We followed the protocol without any modifications.

Read alignment and generation of density maps and modified intervals. Sequence reads from each ChIP library are compiled, post-processed and aligned to the reference genome sequence using a general purpose computational pipeline. We first pre-compute a table that associates each possible 12-mer with all of its occurrences in the reference genome. Then, for each k-bp read, we scan both it and its reverse complement, and for each of its constituent 12-mers, we find each potential start point on the reference genome, and then compute the number of mismatches in the corresponding alignment. These computations are dynamically terminated so that only "unique" alignments are reported, according to the following rule: if an alignment A has only x mismatches, and if there is no alternative alignment having $\leq x + 2$ mismatches, then we call A unique.

For each ChIP (or control) experiment, we next estimate the number of end-sequenced ChIP fragments that overlap any given nucleotide position in the reference genome (here, at 25-bp

resolution). For each position, we count the number of aligned reads that are oriented towards it and closer than the average length of a library fragment (~300 bp).

To identify the portion of the mouse genome that can be interrogated with SMS reads of a given length (k) and alignment stringency, we aligned every k -mer that occurs in the reference sequence (mm8) using the same pipeline as for SMS reads. Nucleotide positions in the reference genome where less than 50% of the 200 flanking k -mers on each side had “unique” alignments, were masked as repetitive and disregarded from further analysis (<28% of the genome). Although we analyzed reads spanning 27-36 bp, all data were conservatively masked at $k=27$.

We identified genomic intervals enriched with a specific chromatin mark from the mean fragment count in 1kb sliding windows. To account for varying read numbers and lengths, we generated sample-specific expected distributions of fragment counts under the null hypothesis of no enrichment by moving each aligned read to a randomly chosen, “unique” position on the same chromosome. Nominal p -values for enrichment at a particular position were obtained by comparison to a randomized version of the same dataset (due to the large number of reads, multiple randomizations gave identical results). Genome-wide maps of enriched sites were created by identification of windows where the nominal p -value fell below 10^{-5} , and merging any enriched windows that were less than 1 kb apart, into continuous intervals. To improve sensitivity to the more diffuse enrichment observed from H3K9me3 and H4K20me3 near repetitive regions and from H3K36me3 across large transcripts, we also developed a Hidden Markov Model (HMM) to segment the reference genome into ‘enriched’ and ‘unenriched’ intervals (Koche, R. *unpublished*). The observed fragment densities were discretized to four categories, in a sample dependent manner (‘masked’, ‘sub-threshold’, ‘near-threshold’ and ‘above threshold’). Emission and transition probabilities were fitted using supervised learning on limited intervals (~10 Mb total) chosen to reflect diverse chromatin landscapes, and the resultant models were applied genome-wide.

Validation of ChIP-Seq by comparison to ChIP-chip and real-time PCR. ChIP-Seq data for H3K4me3 and H3K27me3 in ES cells were compared to published ChIP-chip profiles across ~2% of the mouse genome⁹. Significantly enriched sites in the ChIP-chip data were defined using a previously validated p -value threshold of 10^{-4} , and compared to the ChIP-Seq sites. In addition, a set of 50 PCR primer pairs was designed to amplify 100-140 bp fragments from genomic regions showing a wide range of signal for H3K4me3 and H3K27me3 by ChIP-Seq. Real-time PCR was carried out using Quantitect SYBR green PCR mix (Qiagen) on a 7000 ABI detection system, using 0.25 ng ChIP or WCE DNA as template. Fold-enrichments reflect two independent ChIP assays, each evaluated in duplicate by real-time PCR.

Promoter classification and definition of gene and transcript intervals. The analyzed promoters were based on transcription start sites inferred from full-length mouse RefSeqs (downloaded from the UCSC Genome Browser April 02, 2007). Promoters containing a 500 bp interval within -0.5 kb to +2 kb with a (G+C)-fraction ≥ 0.55 and a CpG observed to expected ratio (O/E) ≥ 0.6 were classified as HCPs. Promoters containing no 500 bp interval with CpG O/E ≥ 0.4 were classified as LCPs. The remainder were classified as ICPs. The chromatin states of promoters were determined by overlap with cell type specific H3K4me3 and H3K27me3 intervals. For comparison with expression levels, the chromatin states of genes with more than one known promoter were classified according to the most 'active' mark (i. e. a gene with an H3K4me3 marked promoter and a bivalent promoter, would be classified as 'H3K4me3'). Correlation between H3K4me3 enrichment and expression levels was calculated from the mean fragment density over promoter from -0.5 kb to +1 kb. Correlation between H3K36me3 and expression levels was calculated from the mean fragment density over each RefSeq transcript.

Expression data. RNA expression data for ES cells, NPCs and MEFs were generated from polyA RNA using GeneChip Mouse Genome 430 2.0 Arrays (Affymetrix). Expression data for adult tissues were downloaded from the Novartis Gene Expression Atlas at expression.gnf.org. Pre-processing, normalization (GC-RMA) and hierarchical clustering (Pearson, log-transformed, row-centered values) were performed using GenePattern [www.broad.mit.edu/cancer/software/].

Analysis of repetitive elements. Chromatin state at repetitive elements was evaluated by aligning SMS reads directly to a library of repetitive element consensus sequences [<http://www.girinst.org>]. The proportion of reads aligning to each class was calculated for H3K9me3 and H4K20me3, and enrichment determined by comparison to WCE and pan-H3. We also applied an orthogonal approach based on HMM intervals of H3K9me3 in unique sequence (see above). For each repetitive element type or class, we calculated the number of occurrences within 1 kb of a unique H3K9me3 site, controlling against a set of randomly placed sites of the same length distribution.

Allele-specific histone methylation. Mouse SNP between the 129 and *M. castaneus* strains were obtained from Perlegen at mouse.perlegen.com. Allele specific bias was evaluated by a binomial test of the null hypothesis that ChIP fragments were drawn uniformly from both alleles. (H3K4me3 and H3K9me3 reads were pooled prior to the test). We note that the 129 strain is closer to the B6-derived reference genome, and this may cause a slight bias towards assigning aligned reads to this strain.

References

1. Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and epigenetic regulators of pluripotency. *Cell* 128, 747-62 (2007).
2. Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* 128, 669-81 (2007).
3. Kouzarides, T. Chromatin modifications and their function. *Cell* 128, 693-705 (2007).
4. Buck, M. J. & Lieb, J. D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* 83, 349-60 (2004).
5. Mockler, T. C. et al. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* 85, 1-15 (2005).
6. Roh, T. Y., Cuddapah, S. & Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* 19, 542-52 (2005).
7. Service, R. F. Gene sequencing. The race for the \$1000 genome. *Science* 311, 1544-6 (2006).
8. Conti, L. et al. Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol* 3, e283 (2005).
9. Bernstein, B. E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125, 315-26 (2006).
10. Ringrose, L. & Paro, R. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu Rev Genet* 38, 413-43 (2004).
11. Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. & Cavalli, G. Genome regulation by polycomb and trithorax proteins. *Cell* 128, 735-45 (2007).
12. Azuara, V. et al. Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* 8, 532-8 (2006).
13. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* 103, 1412-7 (2006).
14. Weber, M. et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 39, 457-66 (2007).
15. Bernstein, B. E. et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120, 169-81 (2005).
16. Kim, T. H. et al. A high-resolution map of active promoters in the human genome. *Nature* 436, 876-80 (2005).
17. Boyer, L. A. et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* (2006).
18. Lee, T. I. et al. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell* 125, 301-13 (2006).
19. Squazzo, S. L. et al. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res* 16, 890-900 (2006).
20. Pasini, D., Bracken, A. P., Hansen, J. B., Capillo, M. & Helin, K. The Polycomb Group protein Suz12 is required for Embryonic Stem Cell differentiation. *Mol Cell Biol* (2007).
21. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci* 31, 89-97 (2006).
22. Wang, X., Su, H. & Bradley, A. Molecular mechanisms governing Pcdh-gamma gene expression: evidence for a multiple promoter and cis-alternative splicing model. *Genes Dev* 16, 1890-905 (2002).

23. Carninci, P. et al. The transcriptional landscape of the mammalian genome. *Science* 309, 1559-63 (2005).
24. Alexander, D. L., Ganem, L. G., Fernandez-Salguero, P., Gonzalez, F. & Jefcoate, C. R. Aryl-hydrocarbon receptor is an inhibitory regulator of lipid synthesis and of commitment to adipogenesis. *J Cell Sci* 111 (Pt 22), 3311-22 (1998).
25. Lengner, C. J. et al. Primary mouse embryonic fibroblasts: a model of mesenchymal cartilage formation. *J Cell Physiol* 200, 327-33 (2004).
26. Garreta, E., Genove, E., Borros, S. & Semino, C. E. Osteogenic differentiation of mouse embryonic stem cells and mouse embryonic fibroblasts in a three-dimensional self-assembling peptide scaffold. *Tissue Eng* 12, 2215-27 (2006).
27. Doetsch, F. The glial identity of neural stem cells. *Nat Neurosci* 6, 1127-34 (2003).
28. Krichevsky, A. M., Sonntag, K. C., Isacson, O. & Kosik, K. S. Specific microRNAs modulate embryonic stem cell-derived neurogenesis. *Stem Cells* 24, 857-64 (2006).
29. Rao, B., Shibata, Y., Strahl, B. D. & Lieb, J. D. Dimethylation of histone H3 at lysine 36 demarcates regulatory and nonregulatory chromatin genome-wide. *Mol Cell Biol* 25, 9447-59 (2005).
30. Bannister, A. J. et al. Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J Biol Chem* 280, 17732-6 (2005).
31. Kim, A., Kiefer, C. M. & Dean, A. Distinctive signatures of histone methylation in transcribed coding and noncoding human beta-globin sequences. *Mol Cell Biol* 27, 1271-9 (2007).
32. Vakoc, C. R., Sachdeva, M. M., Wang, H. & Blobel, G. A. Profile of histone lysine methylation across transcribed mammalian chromatin. *Mol Cell Biol* 26, 9185-95 (2006).
33. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* 128, 707-19 (2007).
34. Fantes, J. et al. Mutations in SOX2 cause anophthalmia. *Nat Genet* 33, 461-3 (2003).
35. Hutchinson, J. N. et al. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* 8, 39 (2007).
36. Seitz, H. et al. A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. *Genome Res* 14, 1741-8 (2004).
37. Cullen, B. R. Transcription and processing of human microRNA precursors. *Mol Cell* 16, 861-5 (2004).
38. Zaratiegui, M., Irvine, D. V. & Martienssen, R. A. Noncoding RNAs and gene silencing. *Cell* 128, 763-76 (2007).
39. Verdel, A. & Moazed, D. RNAi-directed assembly of heterochromatin in fission yeast. *FEBS Lett* 579, 5872-8 (2005).
40. Martens, J. H. et al. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *Embo J* 24, 800-12 (2005).
41. Baust, C. et al. Structure and expression of mobile ETnII retroelements and their coding-competent MusD relatives in the mouse. *J Virol* 77, 11448-58 (2003).
42. Svoboda, P. et al. RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos. *Dev Biol* 269, 276-85 (2004).
43. Cho, D. H. et al. Antisense transcription and heterochromatin at the DM1 CTG repeats are constrained by CTCF. *Mol Cell* 20, 483-9 (2005).
44. Feng, Y. Q. et al. The human beta-globin locus control region can silence as well as activate gene expression. *Mol Cell Biol* 25, 3864-74 (2005).

45. Edwards, C. A. & Ferguson-Smith, A. C. Mechanisms regulating imprinted genes in clusters. *Curr Opin Cell Biol* (2007).
46. Delaval, K. et al. Differential histone modifications mark mouse imprinting control regions during spermatogenesis. *Embo J* 26, 720-9 (2007).
47. Feil, R. & Berger, F. Convergent evolution of genomic imprinting in plants and mammals. *Trends Genet* 23, 192-9 (2007).
48. Strausberg, R. L. et al. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci U S A* 99, 16899-903 (2002).

Supplementary Note: ChIP-Seq read requirement, genome coverage and accuracy

The number of sequence reads required to map a chromatin feature can be estimated from a simple model. Suppose that the genome is divided into N non-overlapping bins of fixed size, that a fraction f of these bins contain a particular chromatin feature and that one performs ChIP-Seq with an antibody that enriches the sequence in these bins by a factor of e . If one collects a total of R sequence reads, the number of reads in a bin should approximately follow a Poisson distribution with mean eM for bins containing the feature and M for the other bins, where $M = R/N(e f + (1-f))$.

Theoretical specificity and sensitivity of ChIP-Seq, conditional on the number of reads, can be estimated from the overlap of the two distributions. For example, suppose that an epitope is present across 1% of the genome, and can be enriched 20-fold by an antibody. Mapping this epitope with 95% specificity and 95% sensitivity into bins of 500 bp would require ~2 million reads. Increasing the resolution to 200 base pairs would require ~5 million reads. Epitopes that enrich less efficiently require more reads (e.g. 10-fold enrichment and 200 base pair resolution would require ~10 million reads).

How much of the mouse genome can be interrogated by ChIP-Seq SMS reads? The proportion depends on the read length k and the mismatch tolerance d (where optimal read alignments are kept for analysis if they have no alternative alignment with $\leq d$ additional mismatches). In this report, we used $k=27$ and $d=2$ (although the actual read lengths varied from 27-36 bp). If we consider 500-bp windows in which at least half of the 27-mers are unique, then ~70% of all windows can be interrogated. Notably, this includes ~20% of all nucleotides in annotated interspersed repeats. Longer read lengths can provide over 80% theoretical coverage.

Moreover, the specificity of SMS read alignments is also high in practice: When reads from individual BAC clones are mapped onto the whole genome using our pipeline, >98% of mappable reads are placed correctly (at $k=27$, $d=2$). This implies that ChIP-Seq can accurately interrogate ~70% of the mammalian genome. By comparison, ChIP-chip yields data for at most ~50% because most repeats are ignored due to the problem of cross-hybridization.

Chapter 8: Genome-scale maps of DNA methylation

In this chapter, we describe the application of single molecule-based sequencing to generate genome-scale maps of DNA methylation for a mammalian species.

This work was first published as

Meissner, A.*, Mikkelsen, T. S.* *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770 (2008).

This publication is attached as Appendix 7. Supplementary notes can be found at the end of the chapter. Supplementary data is available online from <http://www.nature.com/nature>

[This page is intentionally left blank]

We report the generation and analysis of genome-scale DNA methylation profiles at nucleotide resolution in mammalian cells. Using high-throughput Reduced Representation Bisulfite Sequencing (RRBS) and single-molecule-based sequencing, we generated DNA methylation maps covering the vast majority of CpG islands, and a representative sampling of conserved non-coding elements, transposons and other genomic features, for murine embryonic stem (ES) cells, ES-derived and primary neural cells, and eight other primary tissues. Several key findings emerge from the data. First, DNA methylation patterns are better correlated with histone methylation patterns than with the underlying genome sequence context. Second, methylation of CpGs are dynamic epigenetic marks that undergo extensive changes during cellular differentiation, particularly in regulatory regions outside of core promoters. Third, analysis of ES-derived and primary cells reveals that ‘weak’ CpG islands associated with a specific set of developmentally regulated genes undergo aberrant hypermethylation during extended proliferation *in vitro*, in a pattern reminiscent of that reported in some primary tumors. More generally, the results establish RRBS as a powerful technology for epigenetic profiling of cell populations relevant to developmental biology, cancer and regenerative medicine.

Covalent epigenetic modifications to chromatin are thought to be essential for maintaining gene expression patterns and cellular states during development¹⁻³. Methylation of cytosines in CpG dinucleotides is generally associated with repressive chromatin contexts and stably propagated through cell division by DNA methyl-transferases (DNMTs). DNA methylation has been implicated in X inactivation, imprinting, silencing of germline-specific genes in somatic cells and transposon defense^{1,2,4,5}. Moreover, malignant cells frequently display seemingly aberrant DNA methylation patterns, including hypermethylation of CpG islands⁶⁻⁹.

Despite being the most extensively studied epigenetic modification in mammals, relatively little is known about the genome-wide distribution of DNA methylation, how it changes during cellular differentiation or how it relates to histone methylation and other chromatin modifications. Concerns have also been raised about discrepancies in methylation patterns of cells *in vivo* and cells propagated *in vitro*¹⁰⁻¹². Accordingly, a catalog of high-resolution DNA methylation maps from cells at different developmental stages and growth conditions would be a valuable resource for defining normal and abnormal patterns, and for elucidating their functional relevance.

Methylation can be detected by sequencing genomic DNA that has been treated with sodium bisulfite, which converts unmethylated cytosines to uracils by deamination¹³. The traditional approach has been to PCR-amplify targeted loci, using redundant sequence coverage to

estimate the methylation level of each cytosine in a cell population¹⁴⁻¹⁶. It has been impractical, however, to apply bisulfite sequencing at a genome-wide scale because PCR-based approaches are too labor intensive and whole-genome shotgun sequencing, while feasible for small genomes¹⁷, is currently too expensive for comparative analysis across multiple cell states in large mammalian genomes.

To facilitate comparative analysis of nucleotide-resolution DNA methylation levels across cell types, we recently developed Reduced Representation Bisulfite Sequencing (RRBS)¹⁸. This approach relies on restriction digestion and size-selection to isolate and sequence a defined fraction of a large genome. By choosing an enzyme with a recognition site including a CpG dinucleotide, one can enrich for 'CpG islands' while also sampling the remainder of the genome. Computational analysis indicated that digesting genomic DNA with the methylation-insensitive restriction enzyme *MspI* (recognition site: C/CGG), selecting fragments in the range 40-220 bp, and performing 36-bp end-sequencing would cover ~1 million distinct CpG dinucleotides (4.8% of all CpGs in the mouse genome) with roughly half located within CpG islands (including sequence from 90% of all annotated CpG islands in the mouse genome) and the rest distributed between other relatively CpG-poor sequence features (Figure 1; Table1). Notably, while CpGs are not distributed uniformly in the genome, every RRBS sequence read includes at least one informative CpG position (from the recognition site; Figure 2), making the approach highly efficient.

High-throughput bisulfite sequencing

We validated high-throughput RRBS by sequencing *MspI* fragments from wild-type (V6.5; 129SvJae/C57/B6 male) and methylation deficient (*Dnmt1^{kd},3a^{-/-},3b^{-/-}*) ES cells¹⁸, using an Illumina Genome Analyzer. We generated an initial set of ~21 million high quality, aligned RRBS reads. The reads from each cell type included ~97% of the predicted non-repetitive *MspI* fragments (median coverage 12x and 8x, respectively). This demonstrates that *MspI* based RRBS library construction is relatively unbiased (Figure 3) and insensitive to genome-wide CpG methylation levels (estimated by nearest-neighbor analysis (NNA) as 72% and 0.5%, respectively). Reads from methylation deficient cells showed 99% bisulfite conversion of CpGs, and reads from both cell types showed near complete (>99%) bisulfite conversion of non-CpG cytosines.

To investigate cell type-specific DNA methylation patterns, we generated 140 million additional RRBS reads (5.8 Gb of total sequence) from ES-derived neural precursor cells (NPCs) and various primary cell populations described below (Table 2). To study the relationship of DNA methylation and histone methylation patterns, we also generated new chromatin-state maps of H3



Figure 1. Reduced Representation Bisulfite Sequencing. **a**, Schematic overview of the RRBS approach. Genomic DNA is digested with methylation-insensitive MspI. Fragments between 40-220 bp are selected, treated with sodium bisulfite and 5' end-sequenced (see Figure 2 for more details). CpGs are represented as open circles and MspI cut sites are indicated above (v). Filled circles represent either unmethylated (green) or methylated (red) CpGs at each sampled molecule. The methylation level of each CpG is inferred from the number of unconverted sites in reads overlapping that site. The inferred methylation level is shown below each CpG site. The color of the box ranges from green (<20% methylation) to red (>80% methylation). **b**, The MspI-based reduced representation fraction contains ~4.8% of all CpGs in the mouse genome, but is significantly enriched for HCPs and other CpG-rich sequence features.

Table 1: RRBS coverage as a function of size selection

RR digest: Mouse (mm8)			Coverage		CpG islands		Enrichment ^b		
Enzyme	Range	Fragments	Mb ^a	CpGs	All	≥10 CpGs	CpG islands	TSS	Exons+ Introns
MspI	40-120	186,429	13.4	853,075	13,105	12,303	63.1	7.8	1.4
MspI	100-220	185,349	13.3	700,518	12,152	10,492	31.0	5.1	1.3
MspI	220-400	144,683	10.4	472,895	7,840	3,783	11.7	3.6	1.4
MspI	40-220	333,104	24.0	1,383,382	14,353	13,633	47.5	6.5	1.3
MspI	40-400	476,883	34.3	1,853,073	15,015	14,200	36.7	5.6	1.3

The RRBS strategy can be applied to any mammalian genome. Due to the higher CpG and CpG island content of the human genome, the same size fractions will result in approximately twice as many fragments:

RR digest: Human (hg18)			Coverage		CpG islands		Enrichment ^b		
Enzyme	Range	Fragments	Mb ^a	CpGs	All	≥10 CpGs	CpG islands	TSS	Exons+ Introns
MspI	40-120	369,554	23.4	1,808,076	22,434	21,069	41.7	6.9	1.5
MspI	100-220	337,756	24.3	1,463,283	21,064	18,206	19.2	5.3	1.4
MspI	220-400	232,189	16.7	843,688	14,415	7,542	8.6	3.8	1.4
MspI	40-220	647,902	43.5	2,985,666	24,633	23,303	30.0	6.9	1.5
MspI	40-400	878,491	60.1	3,823,195	25,783	24,336	24.1	6.1	1.4

^a Total unique and repetitive sequence covered, assuming 36 bp end reads

^b Relative to complete genome sequence

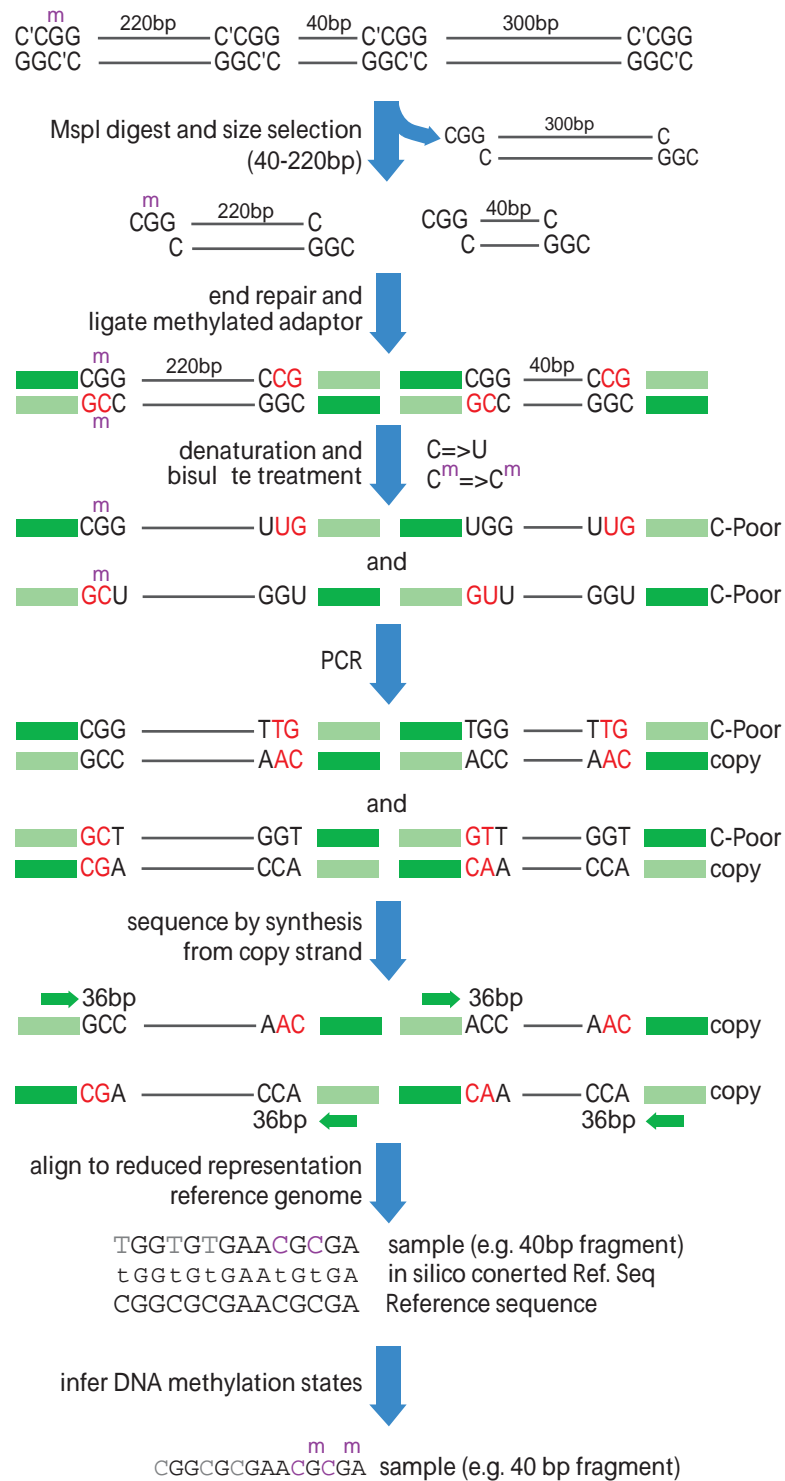


Figure 2. Overview of the RRBS process. Genomic DNA is digested with MspI, size selected, end-repaired and fitted with methylated Illumina/Solexa adapters prior to sodium bisulfite treatment and PCR enrichment. Sequenced reads are aligned to a reference genome digest to infer methylation levels.

Table 2: RRBS libraries sequenced in this study

RRBS Library source	Analyzed (aligned, high- quality) reads	Distinct CpGs	Median coverage (x)	Median CpG methylation level (%)
Astrocytes (in vitro, P18)	9,037,586	951,422	7	70
Astrocytes (primary, P11)	9,638,968	928,227	10	42
Astrocytes (primary, P2)	9,783,816	919,407	10	25
B cells	6,416,120	894,879	7	17
Brain	11,472,495	906,010	14	10
CD4+ T cells	7,312,532	874,811	9	11
CD8+ T cells	5,540,188	821,388	6	10
ES cells	13,298,707	950,671	12	14
ES cells (Dnmt deficient)	8,062,719	908,483	8	0
Liver	7,983,808	668,614	8	9
Lung	9,017,768	796,645	6	8
Embryonic fibroblasts	9,289,500	903,921	8	23
NPC (P18)	9,118,163	921,136	9	40
NPC (P9)	11,150,501	912,408	9	55
Sox1+	11,314,731	972,024	11	29
Sox1+-derived NPCs	12,872,974	996,991	11	35
Tail-tip fibroblasts	10,571,236	948,249	9	11

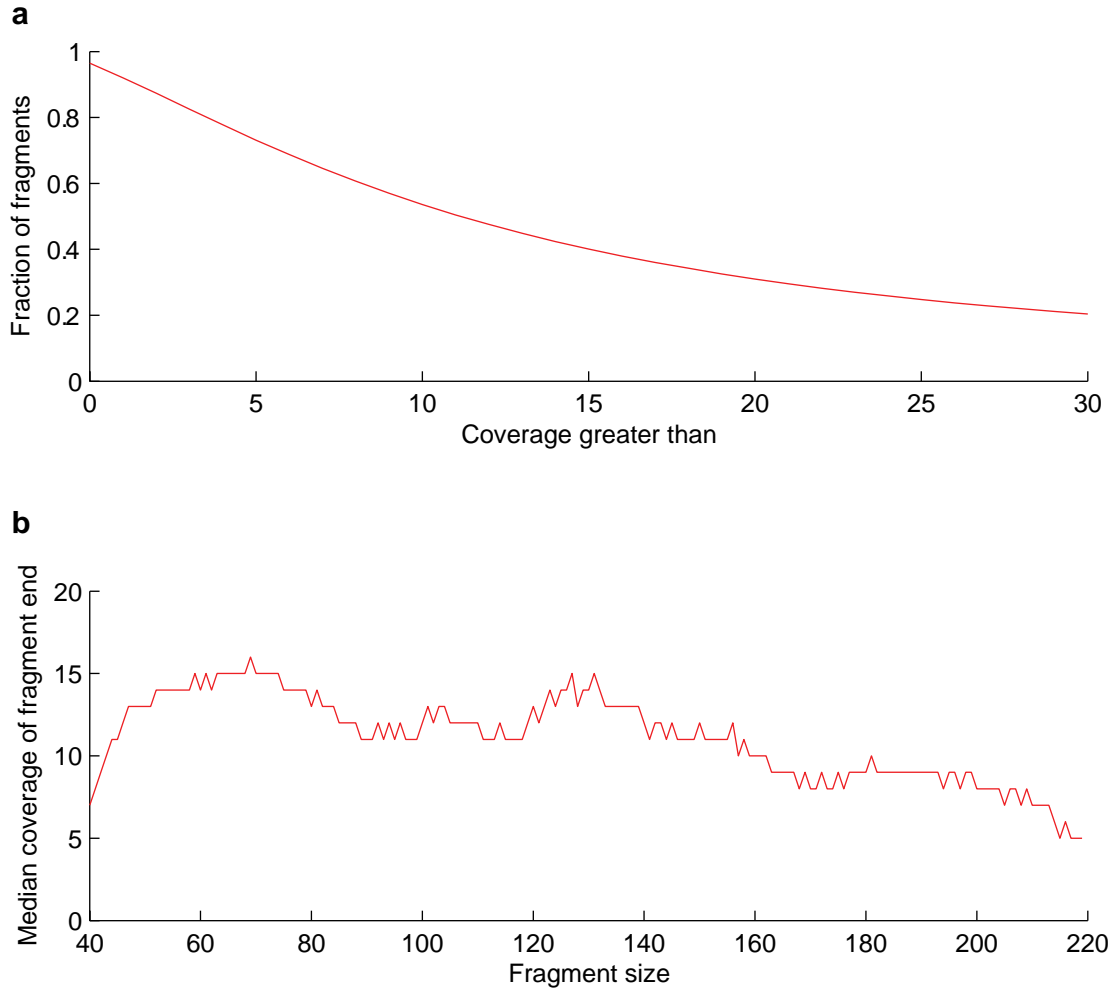


Figure 3. RRBS Library representation from ES cells. a, The majority (97%) of non-repetitive MspI fragment ends were observed at least once among 13 million aligned reads, and the median coverage was 12X. b, Median coverage was relatively similar for fragments of different lengths.

lysine 4 mono- and di-methylation (H3K4me1 and H3K4me2) from ES cells, NPCs and whole brain tissue, using ChIP-Seq¹⁹.

DNA methylation correlates with histone methylation

We began by analyzing the distribution of DNA methylation in wild-type ES cells. As the pluripotent *in vitro* counterpart of the inner cell mass, ES cells represent a key reference point for epigenomic studies^{3,19-21}.

The methylation levels of CpG dinucleotides display a bimodal distribution (Figure 4), with the vast majority being either ‘largely unmethylated’ (<20% of reads showing methylation) or ‘largely methylated’ (>80% of reads). As expected^{1,15,22}, CpGs in regions of high CpG density (>7% in a 300-bp window) tend to be unmethylated, while CpGs in low density regions (<5%) tend to be methylated. However, we noted that DNA methylation is not perfectly predicted by CpG density, particularly in regions of low density: ~10% of CpGs in low density regions were unmethylated, while ~0.3% of CpGs in high density regions were methylated. Because genomic features tend to be associated with distinct histone methylation patterns¹⁹, we analyzed these features separately. We found that DNA methylation patterns were better explained by histone methylation patterns than by CpG density.

High CpG-density promoters. In mammalian genomes, the vast majority of CpG islands are associated with two classes of genes: ‘housekeeping’ genes with ubiquitous expression and ‘key developmental’ genes with complex expression patterns²³. In ES cells, high CpG-density promoters (HCPs) at housekeeping genes are enriched with the transcription initiation mark H3K4me3 (‘univalent’) and are highly expressed, or at least primed for rapid activation, while those at developmental genes are enriched with both H3K4me3 and the repressive mark H3K27me3 (‘bivalent’) and are expressed at low levels^{19,21}. Both types of promoters are also enriched with H3K4me2, which is associated with an open chromatin conformation. Of the 10,299 HCPs sampled (on average, 19 distinct CpGs per promoter), we found that virtually all contain at least a core region of unmethylated CpGs, regardless of their level of expression or H3K27me3 enrichment (Figures 4 and 5a); this is consistent with previous reports^{21,22,24}.

Low CpG-density promoters. In contrast to HCPs, low CpG-density promoters (LCPs) are generally associated with highly tissue-specific genes. In ES cells, only a small subset of LCPs are enriched with H3K4me3 (~7%) or H3K4me2 (~3%), and essentially none are enriched with H3K27me3¹⁹. We found that while most CpGs located in sampled LCPs (990 sites with $\geq 10X$

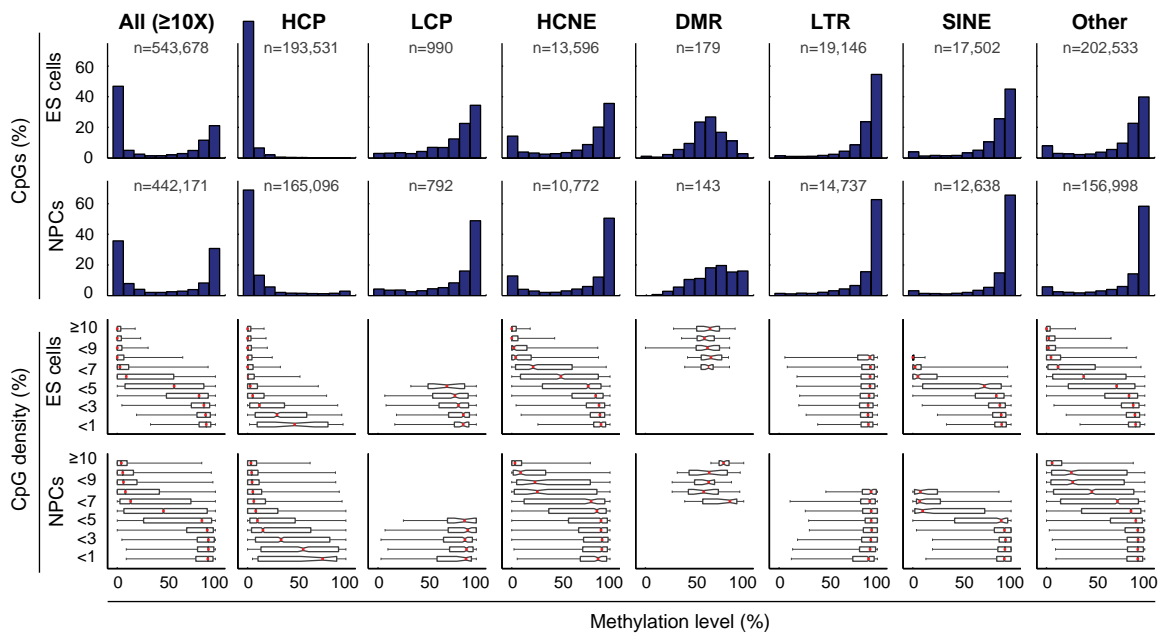


Figure 4. CpG methylation levels in ES cells and NPCs. Distribution of inferred methylation levels for all CpGs with $\geq 10X$ coverage in either ES cells or NPCs. The top histograms show the distribution of methylation levels (%) across all CpGs, high CpG density promoters (HCP), low CpG density promoters (LCP), highly conserved non-coding elements (HCNE), differentially methylated regions (DMR), long terminal repeats (LTR), short interspersed elements (SINE) and other genomic features (n gives the number of CpGs in each category). Methylation levels are bimodal (except at DMRs), and correlate with sequence features and local CpG density. The bottom box plots show the distribution of methylation levels conditional on local CpG density (defined as fraction of CpGs in a 300-bp window; %). The red lines denote medians, notches the standard errors, boxes the interquartile ranges, and whiskers the 2.5th and 97.5th percentiles.

coverage from 392 promoters) are methylated, those in LCPs enriched with H3K4me3 or H3K4me2 have significantly reduced methylation levels (Figure 6).

Distal regulatory regions. Establishment of correct gene expression patterns in mammalian cells often requires multiple *cis*-regulatory elements, such as enhancers, silencers and boundary elements²⁵. *Cis*-regulatory elements active in a particular cell type are often associated with markers of open chromatin, such as H3K4me2 or H3K4me1^{26,27}. Using our new ES cell chromatin state maps, we identified 25,051 punctuate sites of H3K4me2 enrichment in ES cells from 1 to >100 kb away from known promoters (the majority of these sites were also enriched with H3K4me1, but not with H3K4me3). We found that the CpGs sampled at the H3K4me2 enriched sites (outside of promoters and CpG islands) had significantly lower methylation levels than those at unenriched sites (Figure 5b). Interestingly, this relationship was particularly strong for CpGs located in highly conserved non-coding elements (HCNEs; Figure 5c).

Imprinting control regions. The epigenetic process of imprinting typically involves allele-specific histone and DNA methylation of CpG-rich regulatory elements known as imprinting control regions (ICRs)²⁸. Our RRBS library included sequence from 13 of the ~20 known ICRs (on average, 13 distinct CpGs per ICR). CpGs within these elements display a unimodal distribution of methylation levels, with a median close to 50%, which is consistent with hypomethylation of the active allele marked with H3K4me3 and hypermethylation of the silenced allele marked with H3K9me3 (Figure 4)¹⁹. The 50:50 ratio at ICRs also demonstrates the quantitative nature of RRBS.

Interspersed repeats. DNA methylation has been proposed to contribute to genome stability by suppressing the mobility of retrotransposons and other repetitive elements^{1,29}. Repeat families differ in their chromatin structure, with H3K9me3 enriched at active long terminal repeat (LTRs) and to a lesser extent long interspersed elements (LINEs), but not at short interspersed elements (SINEs). Notably, CpGs located in LTRs and LINEs are generally hypermethylated even in CpG-rich contexts (Figure 4). By contrast CpGs in SINEs show a correlation between methylation levels and CpG density that is comparable to non-repetitive sequences.

We conclude that the presence of H3K4 methylation and absence of H3K9 methylation are better predictors of unmethylated CpGs than sequence context alone. This is consistent with models where *de novo* methyl-transferases either specifically recognize sites with unmethylated H3K4³⁰ or are excluded by H3K4 methylation or associated factors. Similarly, H3K9me3 or associated factors may recruit DNMTs at ICRs and repetitive elements^{31,32}.

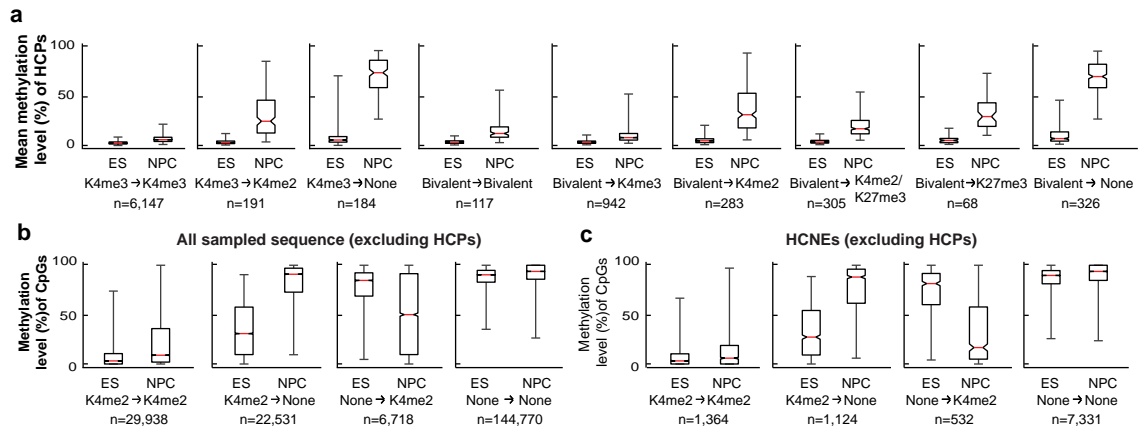


Figure 5. Correlation between DNA and histone methylation. a, Mean methylation levels across CpGs within each profiled HCP (requiring $\geq 5X$ coverage of ≥ 5 CpGs within the CpG island), conditional on their histone methylation state in ES cells and NPCs (n denotes the number of HCPs in each category; HCPs classified as enriched with H3K4me3 are generally also enriched for H3K4me2, but not vice versa). Loss of H3K4 methylation, and to a lesser extent H3K27me3, is strongly correlated with gain of DNA methylation. b, Methylation levels of individual CpGs outside of HCPs, conditional on overlapping enrichment of H3K4me2 (n denotes the number of distinct sites in each category). Changes in histone methylation state is strongly correlated with an inverse change in DNA methylation. c, Methylation levels of CpGs in HCNEs that do not overlap CpG islands, conditional on overlapping enrichment of H3K4me2. The red lines denote medians, notches the standard errors, boxes the interquartile ranges, and whiskers the 2.5th and 97.5th percentiles. All pair-wise comparisons of methylation levels at sites with changing chromatin states are significant ($p < 10e-20$; Mann-Whitney U test).

Dynamic changes in DNA methylation during differentiation

We next used RRBS to analyze how patterns of DNA methylation change when ES cells are differentiated *in vitro* into a homogeneous population of neural precursor cells (NPCs)³³. While CpG methylation levels are highly correlated between the two cell types ($\rho=0.81$), there were clear differences: ~8% of CpGs unmethylated in ES cells became largely methylated in NPCs, while ~2% of CpGs largely methylated in ES cells became unmethylated. We found that the changes in CpG methylation upon differentiation were strongly correlated with changes in histone methylation patterns.

High CpG-density promoters. At both univalent and bivalent HCPs, we found that the vast majority of CpGs remained unmethylated upon differentiation, particularly within their core CpG island, but that loss of H3K4me3 and retention of H3K4me2 or H3K27me3 correlated with partial increase in DNA methylation levels (median ~25%; 2.9% and 12% of univalent and bivalent HCPs, respectively) and complete loss of H3K4 and H3K27 methylation correlated with DNA hypermethylation (median ~75%; 2.8% and 32% of univalent and bivalent HCPs, respectively; Figure 5).

Low CpG-density promoters. Most LCPs marked by H3K4 methylation in ES cells lose this mark in NPCs; while LCPs associated with key genes expressed in NPCs (such as *Fabp7* and *Gpr56*) gain this mark. Loss or gain of H3K4 methylation is a strong predictor of inverse changes in CpG methylation levels at these promoters (Figure 6), resulting in methylation patterns that are cell-type specific.

Distal regulatory elements. Our chromatin state maps revealed that 18,899 (75%) of non-promoter sites enriched with H3K4me2 in ES cells lost this mark in NPCs, while 20,088 new H3K4me2 sites appeared, often in HCNE-rich regions surrounding activated developmental genes (Figure 7). Loss or gain of H3K4 methylation again correlated with a significant increase and decrease in CpG methylation levels, respectively (Figure 5b,c). In fact, these regions account for the majority of observed de-methylation events. Unlike for HCPs, the presence of H3K27me3 alone did not correlate with lower methylation levels in CpG-poor regions (Figure 8).

The data support the notion that CpG-rich and -poor regulatory elements undergo distinct modes of epigenetic regulation during cellular differentiation^{1,19,22}. The vast majority (>95%) of HCPs appear to be constitutively unmethylated and regulated by trithorax-group (trxG; associated with H3K4me3) and/or Polycomb-group (PcG; associated with H3K27me3) proteins, which may be recruited in part via non-specific unmethylated-CpG binding domains³⁴. Hypermethylation of these CpG-dense regions leads to exclusion of trxG/PcG activity, heterochromatin formation and

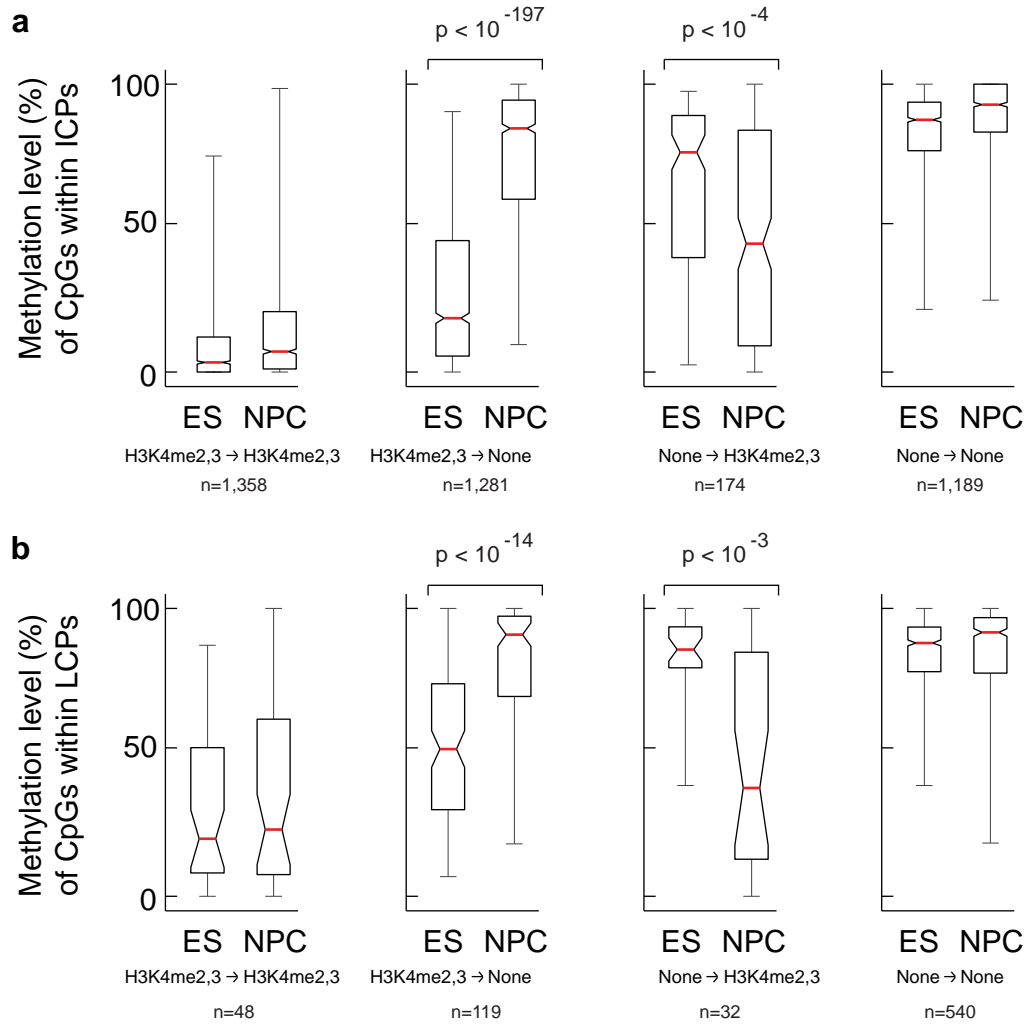


Figure 6. Distribution of CpG methylation levels for **(a)** intermediate CpG-density promoters (ICPs) and **(b)** low CpG-density promoters (LCPs), conditional on histone methylation states in ES cells and neural progenitor cells (NPCs). Changes in H3K4 methylation are significant correlated with inverse changes in DNA methylation levels (Mann-Whitney's U test).

essentially irreversible gene silencing^{1,4}. In contrast, promoters and other regulatory elements in CpG-poor sequence contexts appear to undergo extensive and dynamic methylation and demethylation. Hence, methylation of isolated CpGs may contribute to chromatin condensation or directly interfere with transcription factor binding^{1,3,4}, but do not necessarily prevent chromatin remodeling in response to activating signals.

Comparison of *in vitro* and *in vivo* DNA methylation patterns

As noted above, a small set of HCPs (n=252; ~3%) became hypermethylated (>75% mean methylation across sampled CpGs) upon *in vitro* differentiation of ES cells to NPCs. Because the role of HCP hypermethylation in normal development remains unclear¹, we next investigated whether the observed pattern reflects an *in vivo* regulatory mechanism (Figure9a-f).

We isolated NPCs from E13.5 embryos and differentiated them into Gfap-positive astrocytes (with no more than two passages *in vitro*). We similarly differentiated the *in vitro*-derived NPCs into astrocytes (with these cells having undergone at least 18 passage since isolation from embryoid bodies). We then compared the *in vivo*- and *in vitro*-derived astrocyte populations, using RRBS.

The methylation levels of CpGs were highly correlated ($\rho=0.85$), but astrocytes obtained from *in vivo* NPCs displayed substantially less HCP hypermethylation than those obtained from ES cells (Figure 9a). The *in vivo*-derived astrocytes showed hypermethylation only at 30 HCPs, largely associated with germline-specific genes (including testis-specific transcription factors and meiosis-related genes such as *Dazl*, *Hormad1*, *Sycp1*, *Sycp2* and *Taf7l*), several of which showed partial methylation even in ES cells. In contrast, the *in vitro*-derived astrocytes showed hypermethylation of ~305 additional HCPs, in addition to the germline-specific genes. This set includes some genes known to be expressed by at least some *in vivo* astrocytes (including *Isyna1*, *Gsn* and *Cldn5*;³⁵), but which were silent in the ES-cell derived astrocytes. However, the hypermethylated HCPs are significantly enriched for genes not normally expressed in neural progenitors or the astrocyte lineage (Tables 3-6). They include genes involved in neuronal (*Lhx8*, *Lhx9*, *Moxd1*, *Htf1*, *Slit1*) or ependymal (*Otx2*, *Kl*) differentiation and function, as well as developmental genes associated with unrelated lineages (muscle-specific *Myod1*, Sertoli- and Schwann cell-specific *Dhh* and prostate-specific *Nkx3-1*). In fact, we found that 'key developmental' HCPs that are bivalent in ES cells are six times more likely to be included in the hypermethylated set, compared to univalent 'housekeeping' HCPs. Moreover, univalent genes in the hypermethylated set are expressed at significantly lower levels in both ES cells and primary astrocytes, compared to those that remained

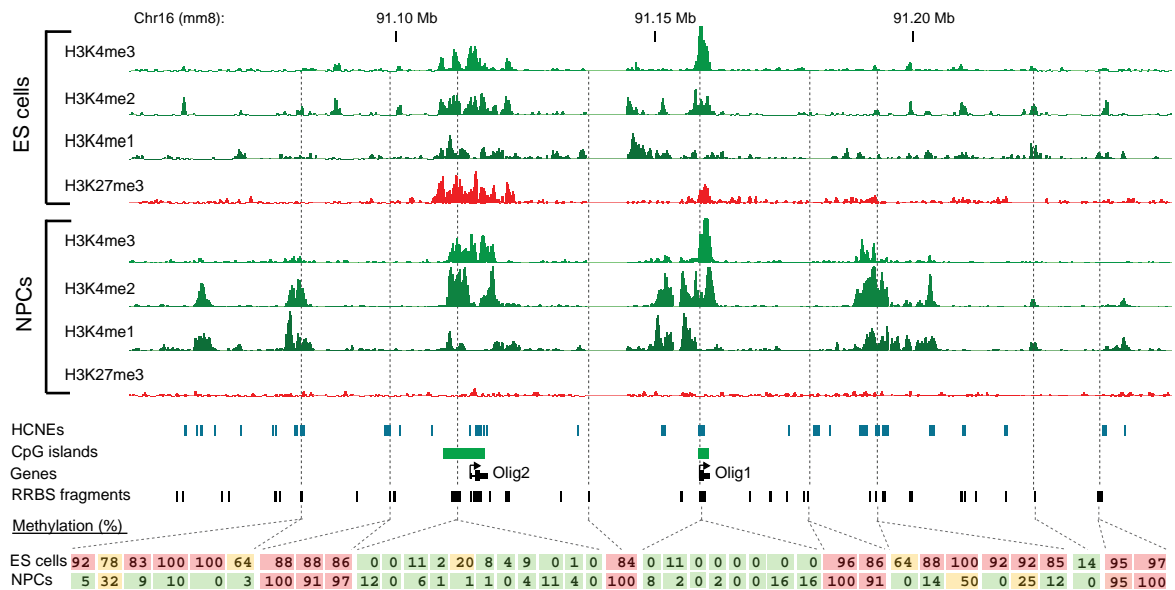


Figure 7. Developmentally regulated de-methylation of highly conserved non-coding elements. Comparison of histone and DNA methylation levels across the Olig1/2 neural-lineage transcription factor locus. Chip-Seq tracks for H3K4me1/2/3 and H3K27me3 in ES and NPCs are shown. The unmethylated CpG-rich promoters are bivalent and inactive in ES cells and resolve to univalent H3K4me3 as the genes are activated in NPCs. Several regions of H3K4me2 enrichment appear over HCNEs distal to the two genes, and this correlates with CpG de-methylation. Inferred methylation values for selected CpGs (40 out of 215 sampled by RRBS in the region) are shown and color-coded (red indicates largely methylated (>80%); green indicates largely unmethylated (<20%).

Table 3: GO Categories enriched for HCPs with > 75% mean methylation in ES-derived astrocytes

GO category	Description	p-value ^a	Genes associated with methylated HCPs ^b
GO:0007126	Meiosis	6.19E-05	Msh4,Sycp3,Sycp2,Spo11,Syce2,Sycp1,Dmc1
GO:0007155	Cell adhesion	0.000141	Dsc2,Cd97,Dsg2,Nlgn2,Cpxm2,Gp1bb,Lamc2,Scarf2,Pkp1,Pgm5,Ctgf,Col9a3,Parvb,Aebp1,Itga4,Col2a1,Cldn11
GO:0007165	Signal transduction	0.001517	Cd97,Gpr176,Cspg4,F2rl1,Lep,Gpr64,Ptger2,Plcd1,Sstr1,Oxtr,Tnfrsf25,Fgfr4,Galr2,Fgf20,Sstr4,Gpr83,Irak3,Fgf17,Prokr1,Prhr,Stat5a,
GO:0007283	Spermatogenesis	0.002592	... Taf7l,Sycp3,Spag6,Dazl,Dmc1,D1Pas1,Msh4,Cldn11,Spag16
GO:0009190	Cyclic nucleotide biosynthetic process	0.005906	Gucy2e,Npr1,Adcy7
GO:0006811	Ion transport	0.006922	Scnn1b,Grin2a,Clic6,Grik2,Atp2a3,Kcnj10,Slc34a2,Trpm6,Kcng1,Slc13a3,Kcna6,Bspry,Slc5a5,Pllp,Kcnb1,Tmem37,Mcoln2
GO:0001541	Ovarian follicle development	0.007887	Msh4,Dmc1,Spo11
GO:0006508	Proteolysis	0.008193	Mmp2,Ccdc79,Mmp23,A530088I07Rik,Agbl2,Mmp14,Casp8,Wdr31,Pgm5,Aebp1,Dhh,Adamts5,Npepl1
GO:0007275	Multicellular organismal development	0.00848	Taf7l,Cspg4,Myod1,Dkk3,Nnat,Hoxd12,Sema4b,Nodal,Lect1,Pgf,Dazl,Dhh,Amn,Dll3,Tnfaip2,Ddx4,Slit1,Nsd1,Hhat,Cdx1,Nkx3.1,
GO:0007218	Neuropeptide signaling pathway	0.011283	... Npb,Sstr1,Gal,Gpr64,Cd97
GO:0001525	Angiogenesis	0.014419	Ctgf,Casp8,Tnfaip2,Pgf,Cspg4,Plcd1
GO:0007517	Muscle development	0.031826	Des,Ky,Myod1
GO:0016477	Cell migration	0.03212	Ctgf,Mmp14,Nodal,Itga4
GO:0006836	Neurotransmitter transport	0.03672	Slc18a2,Slc6a2,Slc6a11

^a Nominal p-value of set enrichment based on Fisher's exact test (two-tailed)

^b Based on GO annotations obtained from <http://geneontology.org>

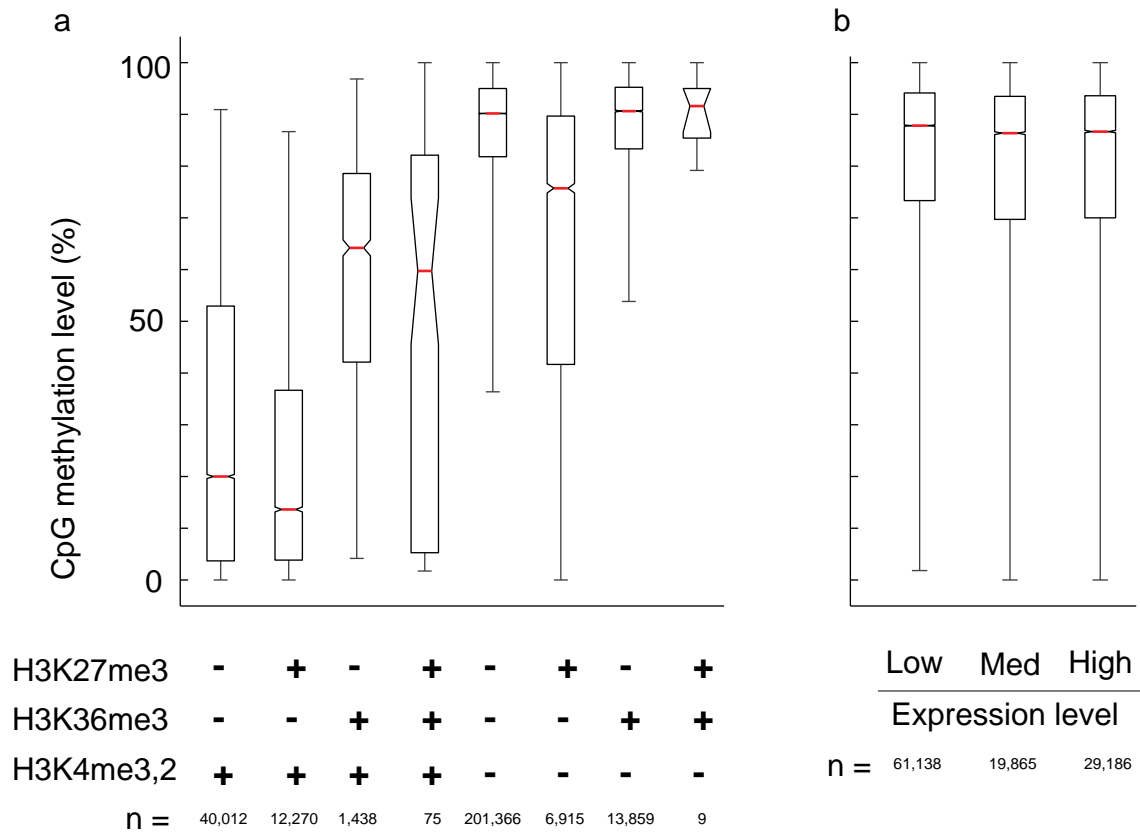


Figure 8. Correlations between histone methylation, expression levels and CpG methylation levels outside of annotated promoters and CpG islands in ES cells. a, H3K4me3 or H3K4me2 are correlated with low DNA methylation, whereas H3K36me3 and H3K27me3 alone is correlated with high DNA methylation levels. b, Distribution of methylation levels for CpGs overlapping known genes (excluding promoter regions), conditional on expression levels. Low = normalized absolute expression level < 50; Med ≥ 50 and < 200; High ≥ 200 .

Table 4: GO Categories enriched for HCPs with > 50% mean methylation in ES-derived astrocytes

GO category	Description	p-value ^a	Genes associated with methylated HCPs ^b
GO:0007165	Signal transduction	7.01E-06	Gpr37,Edaradd,Cd97,Grm8,Fgf15,Gpr176,Gpr101,Cspg4,F2rl1,Wnt3,Lep,Gpr64,Edg3,Ptger2,Plcd1,Gpr12,Sstr1,Oxtr,Grb10,Tnfrsf25,Gm266,Ltb4r2,Fgfr4,Galr2,Drd5,Gpr156,Hif3a, ...
GO:0007155	Cell adhesion	2.82E-05	Dsc2,Nlgn1,Cd97,Itgb4,Dsg2,Nlgn2,Cpxm2,Gp1bb,Lamc2,Pcdhac1,Scarf2,Sdk2,Pkp1,Pgm5,Ctgf,Col9a3,Parvb,Col18a1,Aebp1,Perp, ...
GO:0007275	Multicellular organismal development	2.82E-05	Ndrg2,Edaradd,Taf7l,Vamp5,Itgb4,Cspg4,Myod1,Lmx1b,Dkk3,Wnt3,Churc1,Nnat,Hoxd12,Bmp3,Sema4b,Snai1,Boll,Bmp8b, ...
GO:0006817	Phosphate transport	6.48E-05	Scara3,Col2a1,Emid1,Col25a1,Col18a1,Col12a1,Slc34a2,Gldn,Emid2
GO:0007126	Meiosis	0.00013	Msh4,Sycp3,Sycp2,Spo11,Syce2,Smc1b,Sycp1,Dmc1,Boll
GO:0006811	Ion transport	0.00115	Scnn1b,Kcnf1,Grin2a,Clic6,Grik2,Atp2a3,P2rx5,Kcnj10,Slc34a2,Kcnk13,Trpm6,Kcng1,Kcnc4,Slc13a3,Slc39a8,Kcna6,Bspry,Slc5a5, ...
GO:0030199	Collagen fibril organization	0.01084	Col2a1,Tnxb,Lox,Lmx1b
GO:0001525	Angiogenesis	0.01566	Ctgf,Col18a1,Casp8,Tnfaip2,Sox18,Htatip2,Pgf,Cspg4,Plcd1
GO:0051216	Cartilage development	0.01936	Bmp3,Gnas,Bmp8b,Lect1
GO:0007268	Synaptic transmission	0.02197	Chrna3,Grin2a,P2rx2,Nrxn2,Grik2,Grm8
GO:0001501	Skeletal development	0.02207	Rai1,Dll3,Hoxd10,Gnas,Hoxa11,Hoxd12,Pthlh
GO:0006629	Lipid metabolic process	0.02731	Fads3,Srebf1,Pcsk9,Acot12,Mlst1,Slc27a3,Tnxb,Lep,Acot6,Slc27a2,Plcd1
GO:0006508	Proteolysis	0.02961	Ctsf,Mmp2,Gpr26,Dnahc11,Ccdc79,Mmp23,A530088107Rik,Agbl2,Mmp14,Casp8,Wdr31,Pgm5,Aebp1,Pcsk9,Dhh,St14,Adamts5,Ctsh, ...
GO:0007283	Spermatogenesis	0.04606	Boll,Bmp8b,Taf7l,Sycp3,Spag6,Dazl,Dmc1,D1Pas1,Msh4,Cldn11,Spag16
GO:0007517	Muscle development	0.04622	Des,Ky,Tagln2,Myod1
GO:0001541	Ovarian follicle development	0.04828	Msh4,Dmc1,Spo11

^a Nominal p-value of set enrichment based on Fisher's exact test (two-tailed)

^b Based on GO annotations obtained from <http://geneontology.org>

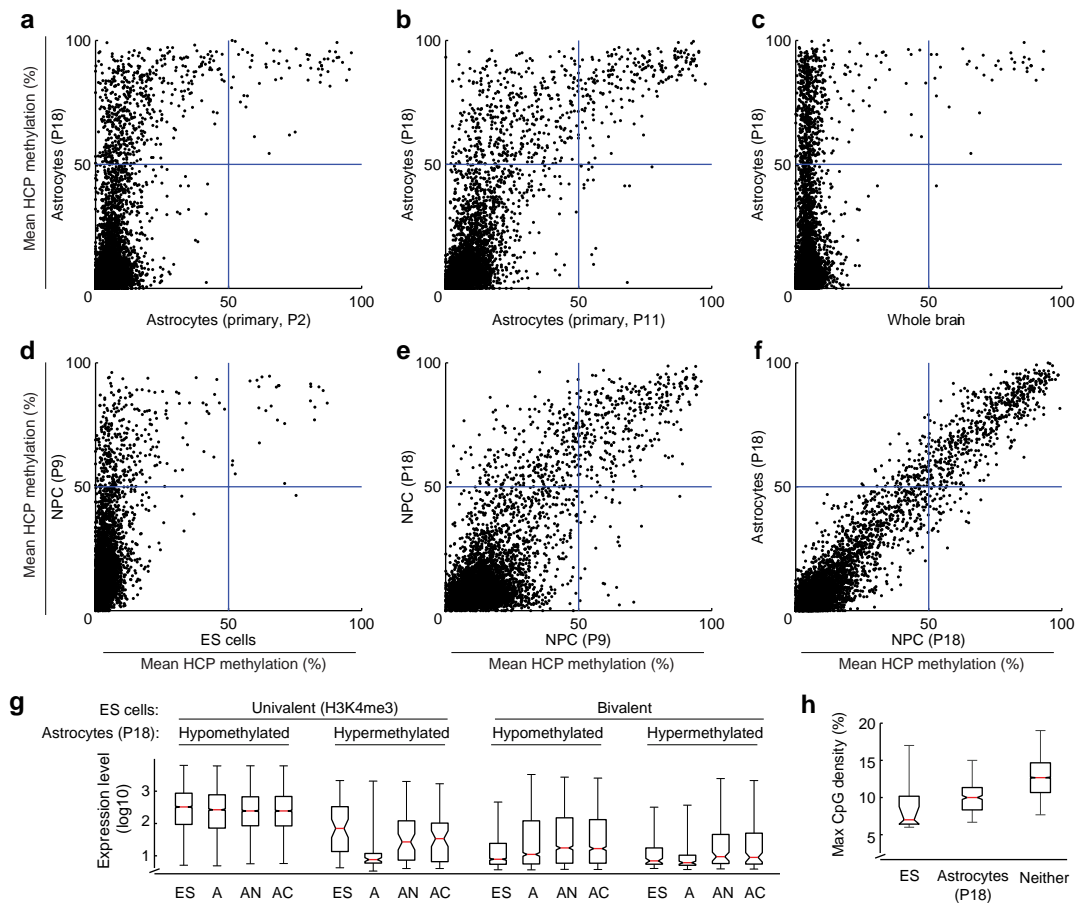


Figure 9. HCP hypermethylation of cultured cells. Inferred mean methylation levels (%) across autosomal HCPs (requiring $\geq 5X$ coverage of ≥ 5 CpGs within the CpG island) are compared between different cell populations. **a**, ES-derived astrocytes contains roughly 10 times more hypermethylated HCPs than primary astrocytes after 2 passages in culture. **b**, Continued passage of primary astrocytes lead to gradual hypermethylation of many of the same HCPs as in ES-derived astrocytes. **c**, Only a handful of mainly germ line-specific HCPs display significant methylation levels in a whole brain tissue sample. **d**, The vast majority of HCPs are unmethylated in ES cells, but a small subset gain significant methylation upon differentiation to NPCs. **e**, Continued proliferation of NPCs leads to additional HCPs becoming hypermethylated after 18 passages. **f**, Differentiation of late-stage NPCs into astrocytes by growth factor withdrawal does not lead to additional HCP hypermethylation. **g**, Expression levels of genes associated with profiled HCPs for ES cells, ES-derived astrocytes (A), primary neocortical (AN) and cerebellar (AC) astrocytes. Hypermethylation of HCPs is strongly correlated with low expression levels in ES-derived astrocytes. HCPs that are univalent in ES cells and become hypermethylated in ES-derived astrocytes are associated with lower expression levels in both ES cells and primary astrocytes. **h**, The maximal CpG density (300 bp window) of hypermethylated HCPs in ES cell or ES-derived astrocytes is significantly lower than for unmethylated HCPs. The red lines denote medians, notches the standard errors, boxes the interquartile ranges, and whiskers the 2.5th and 97.5th percentiles.

Table 5: GO Categories depleted for HCPs with > 75% mean methylation in ES-derived astrocytes

GO category	Description	p-value^a	Genes associated with methylated HCPs^b
GO:0006512	Ubiquitin cycle	0.000243	Parc
GO:0015031	Protein transport	0.001639	Lin7b,Rasef
GO:0006412	Translation	0.010517	Eef1a2
GO:0006886	Intracellular protein transport	0.016169	
GO:0008380	RNA splicing	0.023412	
GO:0006397	mRNA processing	0.031508	Papolb

^a Nominal p-value of set enrichment based on Fisher's exact test (two-tailed)

^b Based on GO annotations obtained from <http://geneontology.org>

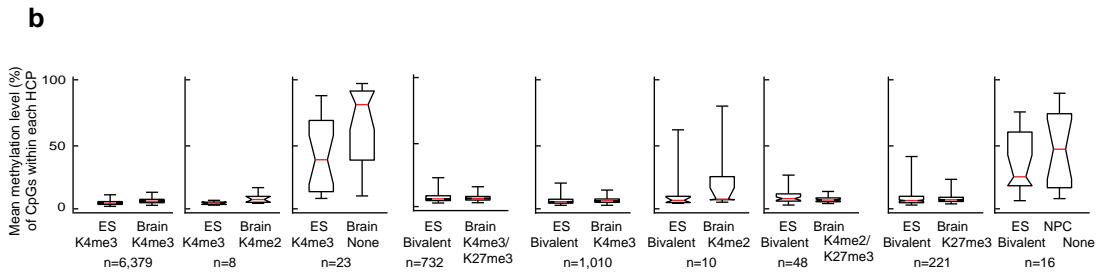
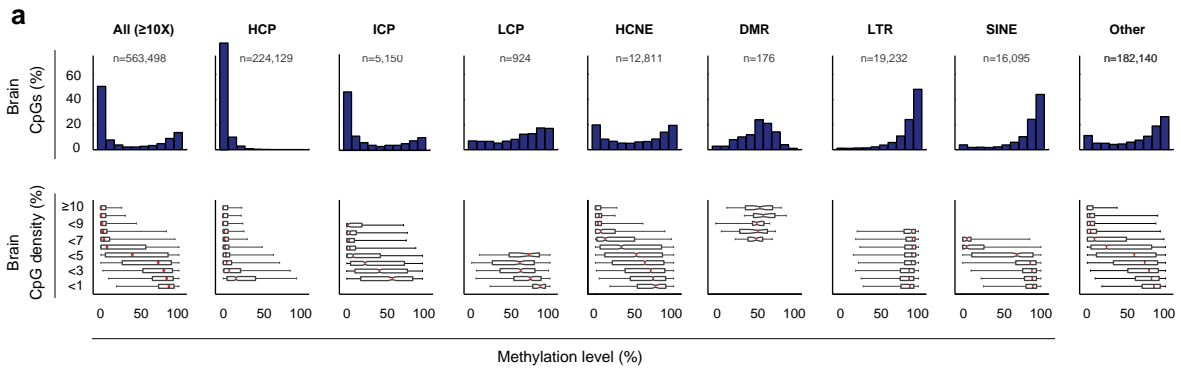


Figure 10. Distribution of CpG methylation levels inferred from a whole brain RRBS library. a, Distribution of inferred methylation levels for all CpGs with $\geq 10X$ coverage in either ES cells or NPCs. The top histograms show and the distribution of methylation levels (%) across all CpGs, high CpG density promoters (HCP), intermediate CpG density promoters (ICP), low CpG density promoters (LCP), highly conserved non-coding elements (HCNE), differentially methylated regions (DMR), long terminal repeats (LTR), short interspersed elements (SINE) and other genomic features (n gives the number of CpGs in each category). The distribution of methylation levels is bimodal and correlated with CpG density and genomic features in a pattern similar to the observed in ES cells. b, The distribution of CpG and histone methylation states for HCPs in ES cells and whole brain. The vast majority of HCPs that are univalent (H3K4me3) in ES cells also show this state in the brain sample. The vast majority of HCPs that are bivalent in ES cells, retain at least one of these marks in the brain sample (enrichment of H3K4me3 and H3K27me3 may not represent bivalency due to heterogeneity). The absence of both H3K4me3 and H3K27me3 correlates with hypermethylation. The red lines denote medians, notches the standard errors, boxes the interquartile ranges, and whiskers the 2.5th and 97.5th percentiles.

Table 6: GO Categories depleted for HCPs with > 50% mean methylation in ES-derived astrocytes

GO category	Description	p-value^a	Genes associated with methylated HCPs^b
GO:0006512	ubiquitin cycle	7.33E-08	Fbxo17,Parc
GO:0015031	protein transport	9.07E-06	Rab3b,Pitpnm1,Lin7b,Sec31b,Rasef
GO:0006412	translation	0.0001	Rps20,Eef1a2
GO:0006397	mRNA processing	0.00014	Papalb
GO:0008380	RNA splicing	0.00023	
GO:0006886	intracellular protein transport	0.001	Rab3b
GO:0006281	DNA repair	0.00444	Mpg
GO:0006974	response to DNA damage stimulus	0.00448	Mpg
GO:0006457	protein folding	0.0091	
GO:0042254	ribosome biogenesis and assembly	0.0141	
GO:0016568	chromatin modification	0.02006	Nsd1
GO:0051726	regulation of cell cycle	0.02915	Cdk11
GO:0006511	ubiquitin-dependent protein catabolic process	0.03008	Parc
GO:0051301	cell division	0.03397	Sycp3,Sycp1,Syce2,Sycp2

^a Nominal p-value of set enrichment based on Fisher's exact test (two-tailed)

^b Based on GO annotations obtained from <http://geneontology.org>

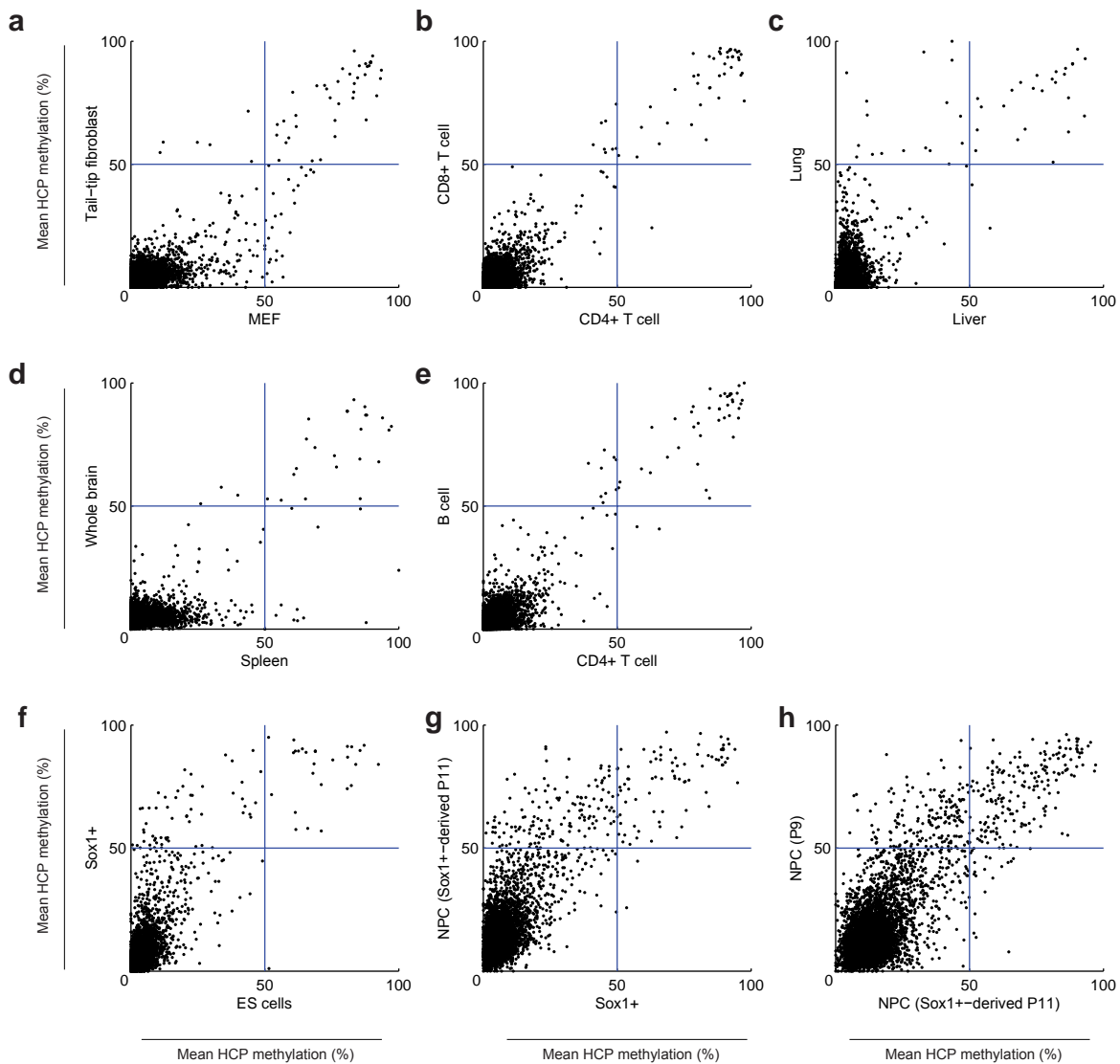


Figure 11. Inferred mean methylation levels (%) of autosomal HCPs compared across different primary and ES-derived cell populations. a-e, primary cell types contain only ~20-30 hypermethylated HCPs, largely associated with germline-specific genes. f-h, Progressive hypermethylation of HCPs during continued proliferation of Sox1+ progenitor cells. Sox1+ is the earliest known marker of neural progenitors and therefore allows isolation of a differentiated ES-derived population after minimal time in culture. There is initially little methylation in these cells, but after 11 passages in culture, many of the same HCPs that were methylated in the original NPC populations have also become methylated in Sox1+-derived NPCs.

hypomethylated (Figure 9g). We also found that the hypermethylated HCPs tend to have a somewhat lower CpG density (~15% lower; Figure 9h).

To further investigate the differences between *in vitro* and *in vivo* cell populations, we also constructed RRBS and ChIP-Seq libraries directly from whole brain tissue (representing cells of mainly glial origin). We found that virtually all (>99%) of sampled HCPs were unmethylated (Figure 9c) and enriched with H3K4me3 and/or H3K27me3 (Figure 10), with ~20 germline-specific HCPs being the only clear exceptions. RRBS libraries from other *in vivo* sources (CD4+ and CD8+ T-cells, B-cells, lung, liver, and embryonic and tail-tip fibroblasts) also showed low levels of hypermethylated promoters (Figure 11). This strongly suggests that – apart from silencing germline specific genes²², imprinted genes and X-inactivated (Figure 12) genes in somatic tissues – hypermethylation of HCPs is not a major mechanisms of normal developmental regulation *in vivo*.

To test for a correlation between passage number and HCP hypermethylation, we examined cell populations derived from intermediate degrees of passage *in vitro*. We studied independently derived, early stage NPCs collected after only 9 passages; these cells displayed HCP hypermethylation at approximately half of the HCPs that are hypermethylated in the late stage NPCs (Figure 9d,e). To further reduce time in culture, we used a Sox1-GFP knock-in ES cell line (Sox1-GFP 129/129)³⁶. Sox1 is the earliest known marker for neural progenitors and allows isolation by FACS of a homogenous population of very early progenitor cells. These cells initially displayed virtually no HCP hypermethylation. However, after continued culturing they acquired hypermethylation at many of the same HCPs as the previous NPC populations (Figure 11). Finally, we grew the *in vivo*-derived astrocyte population discussed above for 11 passages *in vitro* and then examined its methylation pattern. Strikingly, these cells had also begun to acquire hypermethylation at a largely similar set of HCPs (Figure 9a,b).

These results imply that independently derived NPC populations from both *in vitro* and *in vivo* sources and different genetic backgrounds reproducibly undergo hypermethylation at a characteristic set of HCPs, with the process beginning gradually and becoming more prominent with increased passage number.

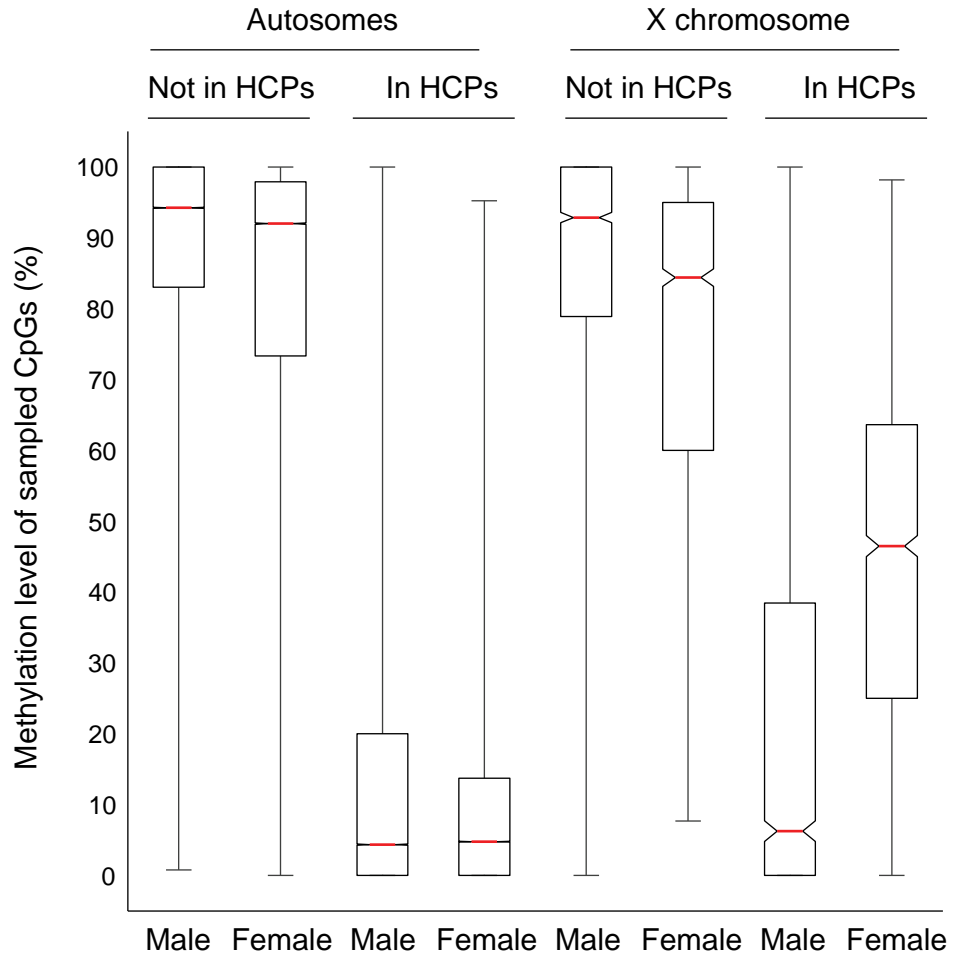


Figure 12. Comparison of methylation levels (%) for CpGs within and outside of HCPs in male and female cell populations (ES-derived and primary astrocytes, respectively). CpG islands show an average of ~50% methylation in the female population, consistent with hypermethylation of HCPs on the inactivated X-chromosome.

Discussion

The RRBS method makes it feasible to perform genome-wide bisulfite sequencing on large-mammalian genomes, providing a valuable tool for epigenetic profiling of cell populations. Bisulfite sequencing is highly reproducible and provides absolute quantitation of methylation levels at nucleotide resolution. As sequencing capacity increases, genome coverage can be readily scaled in step by adding additional restriction enzymes or increasing the selected size range.

Comparative analysis of DNA and histone methylation profiles in ES cells and NPCs reveals novel insights into the relationship between these epigenetic marks and the genome sequence. While various studies have attempted to predict DNA methylation levels from sequence context^{37,38}, we find that cell type-specific histone H3K4 methylation is a significantly better predictor of DNA methylation levels. In regions where H3K4 methylation levels change upon differentiation from ES cells to NPCs, CpG methylation shows a significant inverse correlation. Notably, while all previous genome-scale DNA methylation analyses have focused on CpG islands or promoter regions, our analysis reveals that the majority of differentiation-associated changes take place in relatively low CpG density distal regulatory regions.

Comparison of *in vitro* and *in vivo* cell populations show that key developmental regulators that are bivalent in ES cells and expressed at low levels in the cell type or lineage studied, and also HCPs that have lower than average CpG density, are particularly susceptible to culture-induced hypermethylation. These observations have several implications.

First, generating cellular models by directed differentiation of pluripotent cells is of central interest for developmental biology and regenerative medicine. Aberrant epigenetic regulation in culture have raised concern over the accuracy of such models¹⁰⁻¹². For example, it is well known that both primary and transformed cell lines often lose developmental potency after continued proliferation in culture¹⁰. The efficiency of neuronal differentiation of the multipotent NPC populations studied here declines with increased passage number^{19,39,40}. Susceptibility to hypermethylation at key regulatory genes that are normally activated upon differentiation could explain this phenomenon.

Second, malignant cells are often found to harbor hypermethylated CpG islands⁷⁻⁹. Recently, genes known to undergo frequent hypermethylation in adult cancers were noted to be significantly enriched for genes with bivalent promoters in ES cells⁴¹⁻⁴³. It was therefore suggested that hypermethylation of genes during tumorigenesis may be initiated by the presence of Polycomb group proteins (associated with H3K27 methylation), although functional studies have been inconclusive⁴⁴. The similarities between hypermethylation both in culture and in cancer (and

potentially aging ⁹) may provide a useful *in vitro* model for studying a common underlying mechanism.

Finally, the gradual hypermethylation of lower density HCPs hints at underlying kinetics. Since H3K4 methylases are targeted to HCPs, at least in part by non-specific CpG-binding domains ³⁴, lower CpG density likely contributes to a lower rate of H3K4 methylation. Such HCPs may therefore be particularly sensitive to imbalanced chromatin modifying factors or other culture-induced perturbations.

Methods

Cell culture and ES cell differentiation. V6.5 ES cells and Sox1-EGFP ES cells³⁶ were expanded on γ -irradiated mouse embryonic fibroblasts in DMEM plus 15% fetal bovine serum (FBS; Hyclone) supplemented with 1x MEM-nonessential amino acids (Life Technologies), 0.1mM 2-mercaptoethanol, and 10^3 U/ml leukemia inhibitory factor (LIF). After passaging onto gelatin-coated dishes (0.1% gelatin; Sigma), ES cells were trypsinized and transferred to bacterial dishes allowing embryoid body (EB) formation. EBs were propagated for 4 days in the same medium in the absence of LIF and subsequently plated onto tissue culture dishes. One day after plating, the medium was replaced by ITSFn, i.e. DMEM/F12 (Life Technologies) supplemented with 5 μ g/ml insulin, 50 μ g/ml human APO transferrin, 30 nM sodium selenite (all Sigma), 2.5 μ g/ml fibronectin and penicillin/streptomycin (both Life Technologies). After 5-7 days, cells were trypsinized, triturated to a single cell suspension, replated on laminin-coated dishes (1 μ g/ml; Life Technologies) and further propagated in N3 medium composed of DMEM/F12, 25 μ g/ml insulin, 50 μ g/ml transferrin, 30 nM sodium selenite, 20 nM progesterone, 100 nM putrescine (Sigma), 10 ng/ml FGF2 (R&D Systems, Wiesbaden-Nordenstadt, Germany) and penicillin/streptomycin. Neural precursor cell proliferation was maintained by daily additions of FGF2. Sox1-EGFP positive neural precursors were isolated and FACS-purified (FACS Aria, Becton Dickinson) either from ITSFn cultures or after short-term expansion in FGF2. Growth factor withdrawal of these cultures results in terminal differentiation into primarily neuronal cell populations⁴⁵. Neural precursor cell lines were obtained by sequential passaging and propagation in the presence of 20ng/ml EGF and 10 ng/ml FGF2 (both R&D Systems). Differentiation into astrocytes was induced by growth factor withdrawal and addition of 5% FBS for 5 days.

Primary tissues and cell types. Primary tissues were isolated from 4-6 week old male 129SvJae/C57/B6 mice. Mouse embryonic fibroblasts (MEFs) and primary neural precursors were isolated from 129SvJae/C57/B6 E14.5 embryos. MEFs were generated according to standard protocols. *In vivo* neural precursors were isolated by disaggregating the whole brain and plating the suspension under the conditions described above. Established lines were differentiated into astrocytes by growth factor withdrawal and addition of serum (see above).

MspI RRBS library construction. 1-10 μ g mouse genomic DNA was digested with 10-100 U of MspI (NEB) in a 30-500 μ l reaction overnight at 37°C. Digested DNA was phenol extracted, ethanol precipitated and size selected on a 4% NuSieve 3:1 Agarose gel (Lonza). DNA marker lanes were excised from the gel and stained with SYBR Green (Invitrogen). For each sample, two slices

containing DNA fragments of approximately 40-120 bp and 120-220bp, respectively, were excised from the unstained preparative portion of the gel. DNA was recovered using Easy Clean DNA spin filters (Primm labs, Boston, MA, USA), phenol extracted and ethanol precipitated. The two size fractions were kept apart throughout the procedure including the final sequencing. Size-selected *MspI* fragments were filled in and 3'-terminal A extended in a 50µl reaction containing 20 U Klenow exo⁻ (NEB), 0.4 mM dATP, 0.04 mM dGTP, and 0.04 mM 5-methyl-dCTP (Roche) in 1X NEB buffer 2 (15 min at room temperature followed by 15 min at 37°C), phenol extracted and ethanol precipitated with 10 µg glycogen (Roche) as a carrier. Ligation to pre-annealed Illumina adapters containing 5'-methyl-cytosine instead of cytosine (Illumina) was performed using the Illumina DNA preparation kit and protocol. QIAquick (QIAGEN) cleaned-up, adapter-ligated fragments were bisulfite-treated using the EpiTect Bisulfite Kit (QIAGEN) with minor modifications: The bisulfite conversion time was increased to approximately 14 hours by adding 3 cycles (5 min of denaturation at 95°C followed by 3 hours at 60 °C). After bisulfite conversion, the single-stranded uracil-containing DNA was eluted in 20 µl of EB buffer. Analytical (25 µl) PCR reactions containing 0.5 µl of bisulfite-treated DNA, 5 pmol each of genomic PCR primers 1.1 and 2.1 (Illumina) and 2.5 U PfuTurboC_x Hotstart DNA polymerase (Stratagene) were set up to determine the minimum number of PCR cycles required to recover enough material for sequencing. Preparative scale (8 x 25 µl) PCR was performed using the same PCR profile: 5 min at 95°C, n x (30 s at 95°C, 20 s at 65°C, 30 s at 72 °C) followed by 7 min at 72°C, with n ranging from 18 to 24 cycles. QIAquick purified PCR products were subjected to a final size selection on a 4% NuSieve 3:1 Agarose gel. SYBR Green-stained gel slices containing adapter-ligated fragments of 130-210 bp or 210-310 bp in size were excised. RRBS library material was recovered from the gel (QIAquick) and sequenced on an Illumina 1G Genome Analyzer.

Sequence alignments and data analysis. Sequence reads from bisulfite-treated Solexa libraries were analyzed using a custom computational pipeline. Residual Cs in each read were first converted to Ts, with each such conversion noted for subsequent analysis. A reference sequence database was constructed from the 36 bp ends of each computationally predicted *MspI* fragment in the 40-220 bp size range. All Cs in each fragment end was then converted to Ts (only the C-poor strands are sequenced in the RRBS process; Figure 2).

The converted reads were aligned to the converted reference by finding all 12 bp perfect matches and then extending to both ends of the treated read, not allowing gaps (reverse complement alignments are not considered). The number of mismatches in the induced alignment were then counted between the unconverted read and reference, ignoring cases where a T in the unconverted

read is matched to a C in the unconverted reference. For a given read, the best alignment was kept if the second best alignment had ≥ 2 more mismatches, otherwise the read was discarded as non-unique. Low quality reads were identified and discarded if $\sum_{q \in Q} 10^{q/10} > 1000$, where Q denotes the read quality scores at each mismatched position. The methylation level of each sampled cytosine was estimated as the number of reads reporting a C, divided by the total number of reads reporting a C or T, counting only reads with quality scores of ≥ 20 at the position.

HCP, ICP and LCP annotations were taken from ¹⁹. CpG island and other annotations were downloaded from the UCSC browser (mm8). Estimation of methylation levels from individual CpGs was limited to those with $\geq 10X$ coverage. The methylation level of an HCP promoter was estimated as the mean methylation level across all CpGs with $\geq 5X$ coverage overlapping the annotated CpG island(s) in the promoter, requiring at least 5 such CpGs. HCPs were classified as hypermethylated if this mean methylation level was $\geq 75\%$.

Chromatin immunoprecipitation. H3K4me1 (ab8895), H3K4me2 (ab7766) and H3K4me3 (ab8580) antibodies were purchased from Abcam. ChIP experiments on mouse ES cells (H3K4me1/2), NPCs (H3K4me1/2) and whole brain tissue (H3K4me1/2/3), Illumina/Solexa sequencing, alignments and identification of significantly enriched regions (using 1 kb sliding windows and correction for alignability) were carried out as described previously ¹⁹.

Expression data. RNA expression for ES-derived astrocytes were generated as described previously ¹⁹. Primary astrocyte data was obtained from ³⁵.

References

1. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev* 16, 6-21 (2002).
2. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 33 Suppl, 245-54 (2003).
3. Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* 128, 669-81 (2007).
4. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447, 425-32 (2007).
5. Bestor, T. H. The DNA methyltransferases of mammals. *Hum Mol Genet* 9, 2395-402. (2000).
6. Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 3, 415-28. (2002).
7. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* 128, 683-92 (2007).
8. Feinberg, A. P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet* 7, 21-33 (2006).
9. Esteller, M. Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum Mol Genet* 16 Spec No 1, R50-9 (2007).
10. Jones, P. A., Wolkowicz, M. J., Harrington, M. A. & Gonzales, F. Methylation and expression of the Myo D1 determination gene. *Philos Trans R Soc Lond B Biol Sci* 326, 277-84 (1990).
11. Smiraglia, D. J. et al. Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies. *Hum Mol Genet* 10, 1413-9 (2001).
12. Shen, Y., Chow, J., Wang, Z. & Fan, G. Abnormal CpG island methylation occurs during in vitro differentiation of human embryonic stem cells. *Hum Mol Genet* 15, 2623-35 (2006).
13. Frommer, M. et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* 89, 1827-31 (1992).
14. Rakyan, V. K. et al. DNA Methylation Profiling of the Human Major Histocompatibility Complex: A Pilot Study for the Human Epigenome Project. *PLoS Biol* 2, e405 (2004).
15. Eckhardt, F. et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 38, 1378-85 (2006).
16. Taylor, K. H. et al. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res* 67, 8511-8 (2007).
17. Cokus, S. J. et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452, 215-9 (2008).
18. Meissner, A. et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 33, 5868-77 (2005).
19. Mikkelsen, T. S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553-60 (2007).
20. Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* 292, 154-6 (1981).
21. Bernstein, B. et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* 125, 315-326 (2006).
22. Weber, M. et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 39, 457-66 (2007).

23. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* 103, 1412-7 (2006).
24. Illingworth, R. et al. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol* 6, e22 (2008).
25. West, A. G. & Fraser, P. Remote control of gene transcription. *Hum Mol Genet* 14 Spec No 1, R101-11 (2005).
26. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39, 311-8 (2007).
27. Bernstein, B. E. et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120, 169-81 (2005).
28. Edwards, C. A. & Ferguson-Smith, A. C. Mechanisms regulating imprinted genes in clusters. *Curr Opin Cell Biol* 19, 281-9 (2007).
29. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* 20, 116-7. (1998).
30. Ooi, S. K. et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* 448, 714-7 (2007).
31. Esteve, P. O. et al. Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. *Genes Dev* 20, 3089-103 (2006).
32. Tamaru, H. & Selker, E. U. A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa*. *Nature* 414, 277-83 (2001).
33. Conti, L. et al. Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol* 3, e283 (2005).
34. Voo, K. S., Carlone, D. L., Jacobsen, B. M., Flodin, A. & Skalnik, D. G. Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol Cell Biol* 20, 2108-21 (2000).
35. Sharma, M. K. et al. Distinct genetic signatures among pilocytic astrocytomas relate to their brain region origin. *Cancer Res* 67, 890-900 (2007).
36. Aubert, J. et al. Screening for mammalian neural genes via fluorescence-activated cell sorter purification of neural precursors from Sox1-gfp knock-in mice. *Proc Natl Acad Sci U S A* 100 Suppl 1, 11836-41 (2003).
37. Bock, C. et al. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet* 2, e26 (2006).
38. Das, R. et al. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci U S A* 103, 10713-6 (2006).
39. Bouhon, I. A., Joannides, A., Kato, H., Chandran, S. & Allen, N. D. Embryonic stem cell-derived neural progenitors display temporal restriction to neural patterning. *Stem Cells* 24, 1908-13 (2006).
40. Brustle, O. et al. Embryonic stem cell-derived glial precursors: a source of myelinating transplants. *Science* 285, 754-6 (1999).
41. Ohm, J. E. et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet* 39, 237-42 (2007).
42. Widschwendter, M. et al. Epigenetic stem cell signature in cancer. *Nat Genet* 39, 157-8 (2007).
43. Schlesinger, Y. et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat Genet* 39, 232-6 (2007).

44. McGarvey, K. M., Greene, E., Fahrner, J. A., Jenuwein, T. & Baylin, S. B. DNA methylation and complete transcriptional silencing of cancer genes persist after depletion of EZH2. *Cancer Res* 67, 5097-102 (2007).
45. Okabe, S., Forsberg-Nilsson, K., Spiro, A. C., Segal, M. & McKay, R. D. Development of neuronal precursor cells and functional postmitotic neurons from embryonic stem cells in vitro. *Mech Dev* 59, 89-102 (1996).

[This page is intentionally left blank]

Chapter 9: Integrative analysis of cellular reprogramming

In this chapter, we describe an integrative analysis of changes in gene expression, histone methylation and DNA methylation during direct reprogramming of somatic cells to a pluripotent state.

This work was first published as

Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis
Nature **454**, 49-55 (2008).

This publication is attached as Appendix 8. Supplementary notes can be found at the end of the chapter. Supplementary data is available online from <http://www.nature.com/nature>

[This page is intentionally left blank]

Somatic cells can be reprogrammed to a pluripotent state through the ectopic expression of defined transcription factors. Understanding the mechanism and kinetics of this remarkable transformation may shed light on the nature of developmental potency and suggest strategies with improved efficiency or safety. Here we report an integrative genomic analysis of reprogramming of murine fibroblasts and B lymphocytes. Lineage-committed cells show a complex response to the ectopic expression involving induction of genes downstream of individual reprogramming factors. Fully reprogrammed cells show gene expression and epigenetic states that are highly similar to embryonic stem cells. In contrast, stable partially reprogrammed cell lines show reactivation of a distinctive subset of stem cell-related genes, incomplete repression of lineage-specifying transcription factors, and DNA hypermethylation at pluripotency-related loci. These observations suggest that (i) some cells may become trapped in partially reprogrammed states due to incomplete repression of transcription factors, and (ii) DNA de-methylation is an inefficient step in the transition to pluripotency. We demonstrate that RNA inhibition of transcription factors can facilitate reprogramming, and that treatment with DNA methyltransferase inhibitors can improve the overall efficiency of the reprogramming process.

Mouse and human cells can be reprogrammed to pluripotency through ectopic expression of defined transcription factors¹⁻⁸ (“direct reprogramming”). Generation of such induced pluripotent stem (iPS) cells may provide an attractive source of patient-specific stem cells^{2,4,6,7,9} (reviewed in refs. 10,11). However, the mechanism and nature of molecular changes underlying the process of direct reprogramming remain largely mysterious¹¹. It is a slow and inefficient process that currently requires weeks, with the vast majority of cells failing to reprogram^{2,9,12-14}. A clearer understanding of the process would enable development of safer and more efficient reprogramming strategies, and it might shed light on fundamental questions concerning the establishment of cellular identity.

To identify possible obstacles to reprogramming and to use this knowledge to devise ways to accelerate the transition to full pluripotency, we undertook a comprehensive genomic characterization of cells at various stages of the reprogramming process. The characterization involved gene expression profiling, chromatin state maps of key activating and repressive marks (histone H3 K4me3 and K27me3), and DNA methylation analysis.

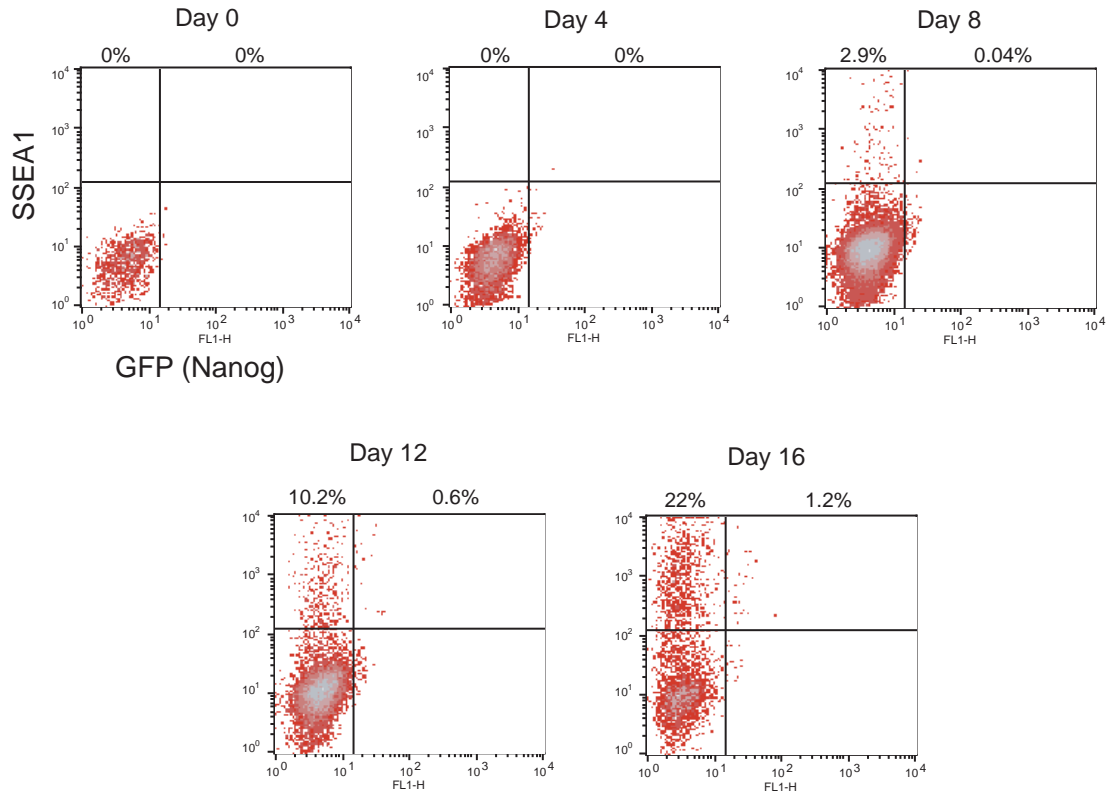


Figure 1. Nanog-GFP MEFs were induced with doxycycline and monitored without passaging (parallel plates were induced and harvested at different time points). At day 8, SSEA1 positive cells and some GFP positive cells appear. At day 16, about 22% of the cells had become SSEA1 positive and around 1.2% SSEA1/GFP double positive.

Response to reprogramming factors in lineage-committed cells

We first studied the response of lineage-committed cells to ectopic expression of the four reprogramming factors Oct4, Sox2, Klf4 and c-Myc. Because the vast majority of induced cells fail to achieve successful reprogramming, we reasoned that genomic characterization might yield insights into the basis of the low overall efficiency of the method.

To eliminate heterogeneity due to differential viral integration, we studied mouse embryonic fibroblasts (MEFs) isolated from chimeric mice that had been generated from an iPS cell line, carrying integrated doxycycline (Dox)-inducible lentiviral vectors with the four reprogramming factors and a Nanog-GFP reporter gene^{13,15}. We induced the expression of the reprogramming factors and obtained gene expression profiles at days 4, 8, 12 and 16. Fluorescence activated cell sorting (FACS) analysis on day 16 showed that ~20% of the cells stained positive for the stem-cell marker SSEA1, but only ~1.2% had achieved complete reprogramming, as indicated by activation of the Nanog-GFP reporter (Figure 1) and consistent with previous reports^{13,14}.

The immediate response to induction of the reprogramming factors (>3-fold change by day 4) is characterized by de-differentiation from the wild-type MEF state and up-regulation of proliferative genes. De-differentiation is evident in significant decrease (5-40 fold) in expression levels of typical mesenchymal genes expressed in MEFs (for example, *Snai1*, *Snai2*). The proliferative response is evident in up-regulation of genes with functions such as DNA replication (*Poli*, *Rfc4*, *Mcm5*) and cell cycle progression (*Ccnd1*, *Ccnd2*); this response may be consistent with expression of reprogramming factor c-Myc^{10,16}.

We also detect a significant increase in expression of stress-induced and anti-proliferative genes. In particular, we detect a sustained 5-10 fold up-regulation of *Cdkn1a* and *Cdkn2a*, which encode cyclin-dependent kinase (CDK) inhibitors that are key effectors of multiple differentiation and tumor suppressor pathways. *Cdkn1a* is a downstream target of the reprogramming factor Klf4¹⁷, while *Cdkn2a* is known to be activated by deregulated c-Myc expression¹⁸. This response was followed by gradual up-regulation of genes associated with differentiating MEFs (*Pparg*, *Fabp4*, *Mgp*) on days 12-16. This suggests that induction of the reprogramming factors triggers normal 'fail-safe' mechanisms that act to prevent uncontrolled proliferation, which may prevent the majority of cells from reaching a stably de-differentiated state.

Interestingly, we also detect strong up-regulation of lineage-specific genes from unrelated lineages. These include axon guidance factors (*Epha7*, *Ngef*), epidermal (*Krt14*, *Krt16*, *Ivl*, *Sprr1a*) and glomerular proteins (*Podxl*). We speculate that this gene activation reflects responses to the

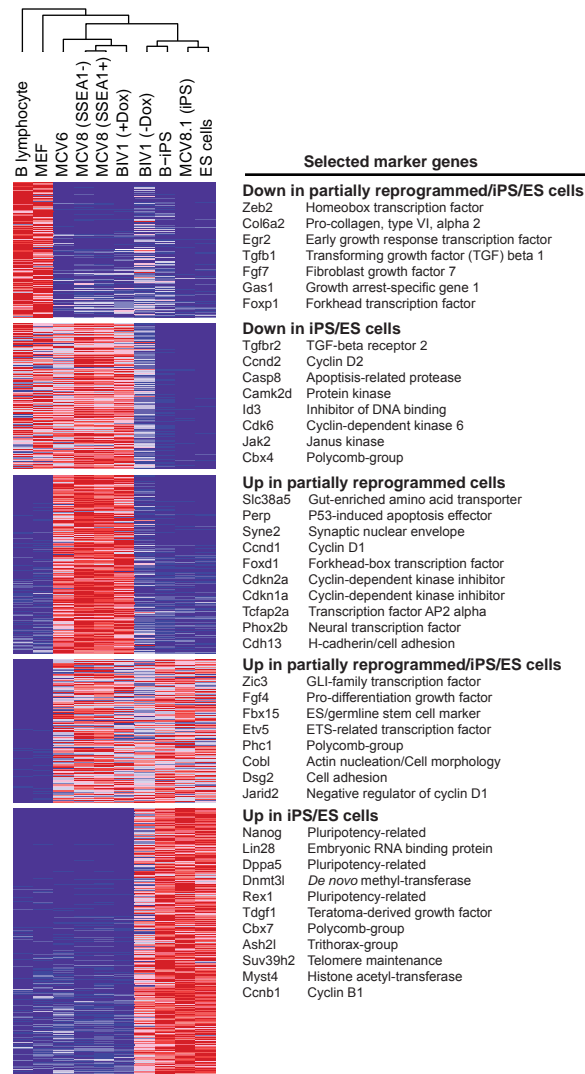


Figure 2. Gene expression profiling. Relative expression levels across differentiated, partially reprogrammed and pluripotent cell populations. The dendrogram was generated by complete linkage hierarchical clustering using Pearson correlation on all measured genes. Only genes with at least 2-fold difference between any pair of samples from different classes are shown in the heatmap. Red, white and blue indicate higher, identical and lower relative expression, respectively.

reprogramming factors Sox2 and Klf4, which, independent of their roles in ES cell regulation, function in neural, epidermal and kidney differentiation^{10,17}.

Pluripotent cell lines

We next studied the changes to gene expression patterns and epigenetic states seen in successfully reprogrammed iPS cells. We analyzed three cell lines: MEF derived iPS cells carrying an Oct4-GFP reporter (MCV8.1; corresponding to subclone 8.1 in¹²); mature B lymphocyte-derived iPS cells carrying a Nanog-GFP reporter (B-iPS)¹⁵; and wild-type ES cells (V6.5)¹⁹.

We found that the genome-wide expression profiles of Oct4- or Nanog-iPS cells derived from different cell types and systems are highly similar, but not identical, to wild-type ES cells (Figure 2), consistent with recent studies of independent cell lines^{2,4,9,20}. For example, the iPS and ES cell lines share high expression levels of genes related to maintenance of pluripotency and self-renewal such as *Oct4*, *Sox2*, *Nanog*, *Lin28*, *Zic3*, *Fgf4*, *Tdgf1* and *Rex1*, and low expression levels for most lineage-specifying transcription factors and other developmental genes. Consistent with the characteristically short cell cycle of ES cells, the iPS cells show low expression of cyclin D (*Ccnd1*, *Ccnd2*)²¹.

To determine whether iPS cells have also regained ES cell-like chromatin states, we generated genome-wide maps showing the location of H3K4me3 and H3K27me3 from the MEF-derived MCV8.1 cell line, using ChIP-Seq. We previously described the differences in these chromatin modifications between wild-type ES cells and MEFs²². In ES cells, virtually all high-CpG promoters (HCPs) are enriched with H3K4me3; a subset of these HCPs, associated with repressed developmental genes, are also enriched with H3K27me3 ('bivalent'). In MEFs, the majority of HCPs that are bivalent in ES cells resolve to become monovalent (H3K4me3- or H3K27me3-only). Some pluripotency- and germline-specific genes show loss of both H3K4me3 and H3K27me3 in somatic cells and this correlates with DNA hypermethylation^{23,33}.

The chromatin state maps of the iPS cell line MCV8.1 are strikingly similar to those of ES cells both near promoters and in intergenic regions (Figures 3-5). The vast majority (>97%) of HCPs that lack H3K4me3-enrichment in MEFs have regained this mark in MCV8.1 cells. At all pluripotency- and germline-specific genes examined, the promoters have regained H3K4me3-enrichment and show DNA hypomethylation (Figure 6). At genes encoding lineage-specific transcription factors that are bivalent and transcriptionally silent in ES cells, the bivalent pattern is typically re-established (~80% of HCPs classified as bivalent in wild-type ES cells; and ~95% of loci encoding key developmental transcription factors; Figure 3b-d,g).

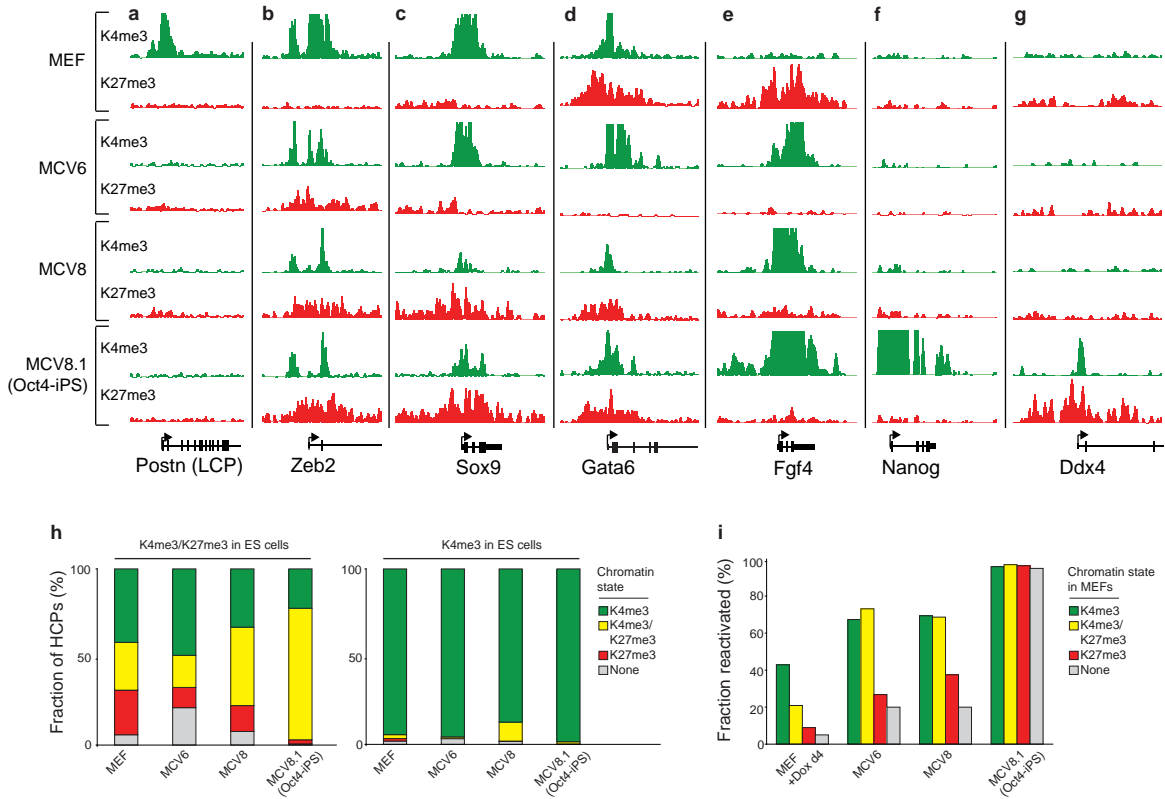


Figure 3. Chromatin state maps. a, Loss of H3K4me3 correlates with inactivation of MEF-specific low-CpG promoters (LCPs) during reprogramming. b, The transcription factor Zeb2 is marked by H3K4me3 and expressed in MEFs, but gains H3K27me3 and is silenced in partially and fully reprogrammed cells. c, The mesoderm/neural-crest transcription factor Sox9 is marked by H3K4me3 only and remains active in MCV6. d, The endodermal transcription factor Gata6 inappropriately lost H3K27me3 and is activated in MCV6 cells. e, The autocrine growth factor Fgf4 loses H3K27me3, gains H3K4me3 and becomes highly expressed in both partially and fully reprogrammed cells. f, The pluripotency gene Nanog gains H3K4me3 and is active only in iPS cells. g, The germline-specific gene Ddx4 gains H3K4me3 and H3K27me3 in iPS cells only, and remains poised for activation in germ cells. h, Chromatin states for high-CpG promoters (HCPs) in MEFs and reprogrammed cells, conditional on their state in ES cells. i, Fraction of genes with HCPs expressed in ES cells, but not wild-type MEFs, that have been re-activated in cells at various stages of reprogramming, conditional on their chromatin state in MEFs. Most HCPs marked by H3K27me3 only or neither mark are not re-activated in partially reprogrammed cells.

We conclude that direct reprogramming to a pluripotent state involves re-activation of endogenous pluripotency-related genes, establishment of an ‘open’ chromatin state (as indicated by genome-wide H3K4me3 enrichment and DNA de-methylation), and comprehensive Polycomb-mediated repression of lineage-specifying genes (as indicated by bivalent chromatin states involving H3K27me3-enrichment).

Partially reprogrammed cell lines

Only a subset of the stably de-differentiated cells obtained in the absence of drug selection show evidence of complete reprogramming to a pluripotent state. We previously derived clonal cell lines that can be maintained in relatively stable, “partially reprogrammed” states in the absence of drug selection¹². We reasoned that characterizing such cells might help identify key barriers in the late stages of the process. Accordingly, we studied three partially reprogrammed independent cell lines established during attempts to reprogram MEFs or mature B lymphocytes (Figures 2-6).

MCV8. This cell line, which corresponds to subclone 8 from¹², was established during our attempt to reprogram MEFs carrying an Oct4-GFP reporter with constitutive retroviruses. It produces heterogeneous cultures of cells with mainly fibroblast-like morphology, with ~20-30% positive for the stem cell marker SSEA1 (Figure 7, 8), and occasional interspersed ES-like colonies at late passages. Multiple secondary subclones from these ES-like colonies have been shown to establish homogeneous GFP positive iPS cell lines (including the MCV8.1 line characterized above¹²). Proviral integration patterns showed that the same parental cells in the MCV8 population gave rise to both GFP-positive and -negative cells, suggesting that complete reprogramming depends on stochastic epigenetic events^{11,12}.

The gene expression patterns of MCV8 cells are clearly distinct from both MEFs and iPS cells (Figure 2). MCV8 cultures show significant down-regulation of both structural genes (*Colla1*, *Colla2*) and regulatory factors (*Snai1*, *Snai2*, *Zeb2*) expressed in MEFs, up-regulation of some lineage-specific genes with neural, epidermal or endodermal functions (presumably as a consequence of Sox2 and Klf4 expression), and particularly high expression of proliferative genes. Interestingly, high levels of expression can also be detected for several of the CDK inhibitors (*Cdkn1a*, *Cdkn2a*) induced by the reprogramming factors. It is unclear how the partially reprogrammed cells have escaped the presumed anti-proliferative effects of these genes, but possible explanations include compensation by overexpression of proliferative genes, repression of differentiation pathways (MCV8 is cultured in the presence of the differentiation inhibitor LIF and

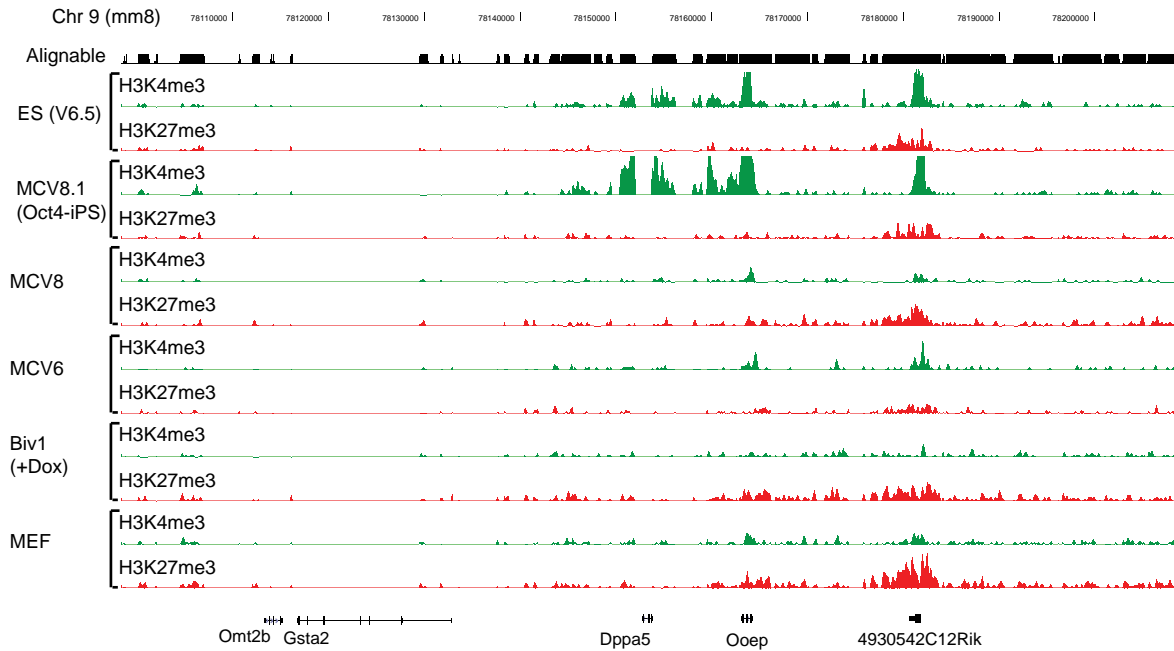


Figure 4. Chromatin state maps of the pluripotency-related *Dppa5* locus, covering H3K4me3 and H3K27me3 in differentiated, partially reprogrammed and pluripotent cells. “Alignable” shows subintervals that support unique Illumina read alignments and can therefore be queried by ChIP-Seq.

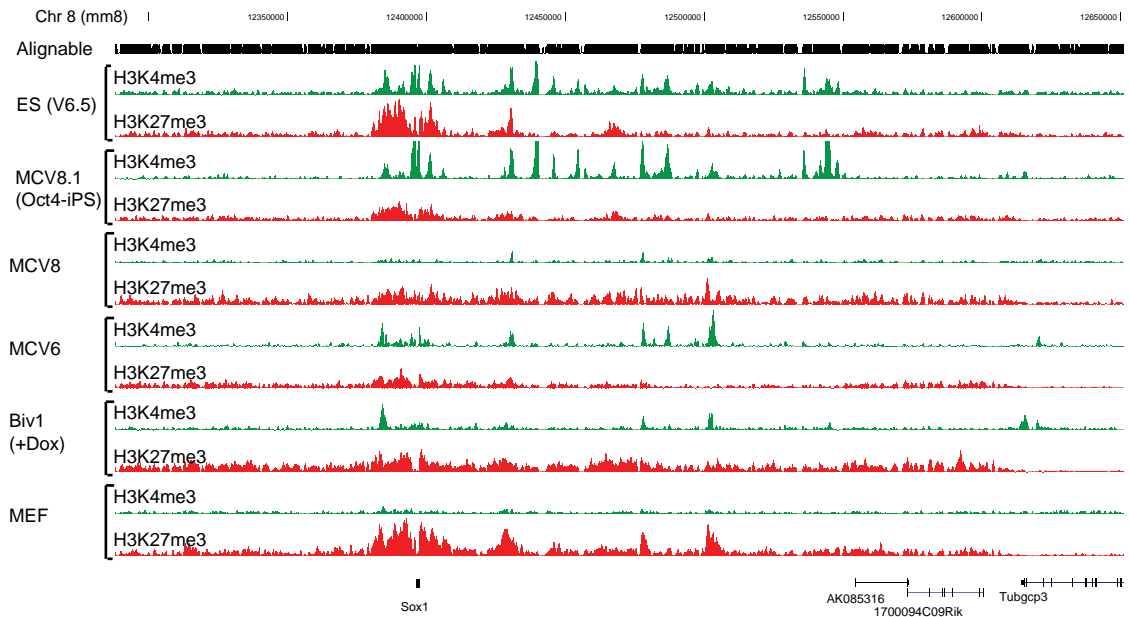


Figure 5. Chromatin state maps of the key developmental Sox1 locus, covering H3K4me3 and H3K27me3 in differentiated, partially reprogrammed and pluripotent cells. “Alignable” shows subintervals that support unique Illumina read alignments and can therefore be queried by ChIP-Seq.

expresses the LIF receptor at 2-3-fold higher levels than ES cells), or transformation (but we note that MCV8 cells have not lost the ability to re-differentiate, see below).

The pattern of gene re-activation of genes expressed in ES cells in MCV8 is strongly correlated with chromatin state in MEFs (Figure 3i). Several genes related to self-renewal and proliferation of embryonic and adult stem cells show re-activation, including the autocrine growth factor *Fgf4*²⁴ and the transcription factor *Zic3*²⁵, but genes directly related to pluripotency show low or undetectable expression. Of HCPs that are enriched with H3K4me3 in MEFs but not expressed at detectable levels, the majority (~70%) are re-activated in MCV8. In contrast, transcriptionally silent HCPs that are enriched in MEFs for H3K27me3 only or neither mark are significantly less likely to be re-activated (~35% and ~20%, respectively; $P_{\text{Fisher}} < 10^{-6}$).

There are notable differences in the chromatin states of MCV8, MEFs and MCV8.1 iPS cells (Figure 3). Examining HCPs that are bivalent in ES cells, MCV8 cells show bivalent chromatin structures at 70% more of these loci (n=1,467) than seen in the MEFs (n=859), but ~40% fewer than in MCV8.1 iPS cells (n=2,360), which is consistent with partial de-differentiation (~88% of the bivalent loci in MCV8 are also bivalent in MCV8.1). There are many more HCPs that lack H3K4me3 and H3K27me3 in MCV8 than in MCV8.1 (n=311 vs. 31), and these genes include the majority of pluripotency- and germ cell-specific loci. Using bisulfite sequencing, we confirmed that this chromatin state correlates with DNA hypermethylation (Figure 6).

We note that we initially sorted MCV8 cells into SSEA1-positive and -negative cells and analyzed them separately. However, we found no major differences in expression levels or DNA methylation patterns between the two fractions (Figure 2, 6). Moreover, when the two subpopulations were cultured separately, both reverted to a heterogeneous state within 1-2 passages. Similar results were obtained from sorting by MHC surface expression, which decreases upon reprogramming. Thus, while these surface markers may provide some enrichment for cells that are amenable to full reprogramming¹⁴, they do not appear to discriminate between significantly different cell states within MCV8 cultures.

MCV6. This cell line was also established during our attempt to reprogram Oct4-GFP MEFs (subclone 6 from¹²). It produces homogeneous cultures with compact colonies and ES cell-like morphology (Figure 8). It differs from MCV8 in that it has different proviral integrations and has never spontaneously given rise to fully reprogrammed cells (Figure 7).

The gene expression profile and chromatin state maps from MCV6 are largely similar to MCV8, but we found two notable differences. First, MCV6 has fewer genes with bivalent chromatin signatures, and a disproportionately large fraction of HCPs without neither H3K4me3- or

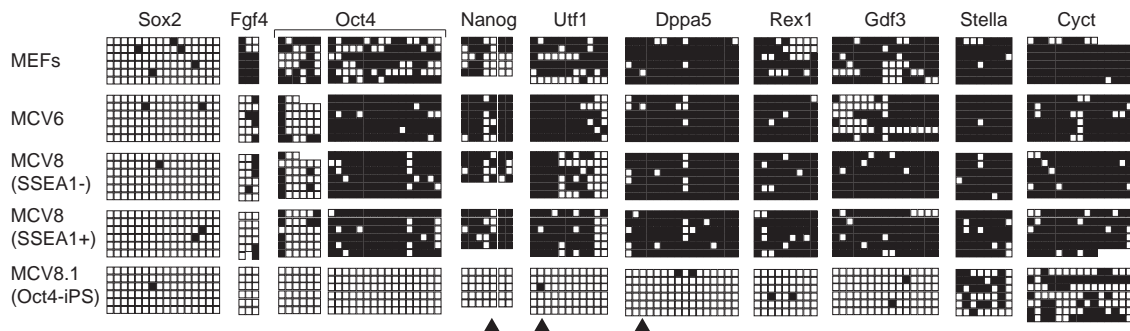


Figure 6. DNA methylation analysis. Bisulfite sequencing of promoters or enhancers with Oct4/Sox2 binding sites near pluripotency-related and germ cell specific (Stella, Cyct) genes, as cataloged in 23. Empty squares indicate unmethylated and filled squares methylated CpG dinucleotides. The majority of assayed sites are hypermethylated in differentiated and partially reprogrammed cells. Sox2 is enriched with H3K27me3 in non-pluripotent cells and accordingly hypomethylated in all cell types. Triangles show sites used for COBRA analysis (see text).

H3K27me3-enrichment (7% vs. ~2.5% in MEFs and MCV8). Second, MCV6 expresses high levels of several lineage-specifying transcription factors that are expressed at low or undetectable levels in MCV8 or iPS cells, including *Sox9* (Figure 3c), and *Gata6* (Figure 3d). The latter observation suggests that MCV6 may have become trapped in more differentiated state than MCV8.

BIV1. This cell line was established during our attempt to reprogram B lymphocytes with inducible lentiviral vectors¹⁵. It had lost surface expression of all common lymphoid markers and did not require any lymphoid cytokines for growth, but also showed no evidence of achieving complete reprogramming during 50 days of continuous Dox-mediated viral expression (as judged by the absence of SSEA1 or GFP-positive cells). After Dox withdrawal and loss of any detectable viral expression (see below), the cells continued to proliferate with a more fibroblast-like morphology and, after more than 10 additional days in culture, spontaneously gave rise to some GFP positive ES-like colonies, but at a lower frequency than MCV8 (Figure 8, 9).

The gene expression profile and chromatin state maps from BIV1 cells grown with Dox show striking similarities to those of MCV8, including: down-regulation of lineage-specific genes, such as the B lymphocyte master regulator *Pax5*; high expression of proliferative genes; activation of neural and epidermal genes; low levels of H3K4me3 and H3K27me3 enrichment relative to ES cells, consistent with DNA hypermethylation (see below); and incomplete activation of pluripotency-related loci (Figures 2-6). Notably, the expression profiles of BIV1, MCV8 and MCV6 are more similar to each other ($r^2 > 0.9$ for any pair) than to the lineage-committed cell types from which they originated or to any of the pluripotent cell types ($r^2 < 0.8$ for any pair; Figure 2). This suggests that the three cell lines may represent relatively common intermediate states induced by the four reprogramming factors. (The three lines also show expression of *Fbx15*, suggesting that they may be similar to the *Fbx15*-selected cells obtained during initial attempts to generate iPS cells⁷.)

Comparing the expression profiles of BIV1 cultures before and after Dox withdrawal, we found that Dox withdrawal resulted in: up-regulation of mesenchymal extracellular matrix genes (*Coll1a1*, *Col2a1*), consistent with the shift to a more fibroblast-like morphology; down-regulation of most inappropriately expressed neural and epidermal genes, which is consistent with these genes being induced by over-expression of Sox2 or Klf4; and up-regulation of some iPS/ES-specific genes (*Dppa5*, *Lin28*, *Dnmt3l*), which is consistent with the eventual emergence of rare GFP positive colonies. Thus, continuous over-expression of the reprogramming factors may paradoxically have stabilized BIV1 cells in its partially reprogrammed state.

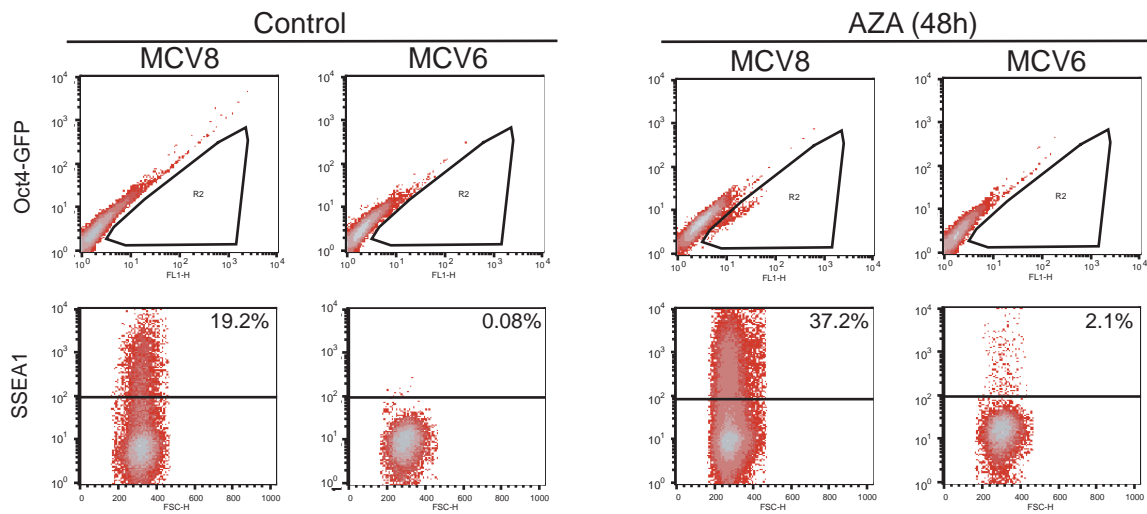


Figure 7. MCV8 and MCV6 cells were sorted for GFP and SSEA1 by FACS. The left panels show cells in ES cell medium only. The right panels show cells that were treated with AZA for 48 hours. FACS was done 48-72 hours after terminating the AZA treatment (passage 1).

In summary, the three partially reprogrammed cell lines appear to represent similar (but distinct) cell states that emerge at an intermediate stage in the direct reprogramming process. The states are characterized by re-activation of genes related to stem cell renewal and maintenance, but not pluripotency; incomplete repression of lineage-specific transcription factors; and incomplete epigenetic remodeling, including persistent DNA hypermethylation.

Inhibition of DNA methyltransferase accelerates reprogramming

Because the partially reprogrammed cell lines show DNA hypermethylation at pluripotency-related genes, we hypothesized that loss of DNA methylation (or a closely linked epigenetic mark, such as H3K9 methylation²⁶) is a critical and inefficient step in the transition from a partially reprogrammed state to pluripotency.

Overcoming the block in partially reprogrammed cell lines. We tested this notion by treating cells with the DNA methyltransferase inhibitor 5-aza-cytidine (AZA) and found that it induced a rapid and stable transition to a fully reprogrammed iPS state. We initially studied SSEA1-positive MCV8 cells, treating them with AZA for 48 hours and monitoring the subsequent appearance of GFP-positive cells (Figure 7, 10). GFP-positive cells appeared at a frequency of 7.5% after one passage, compared to 0.25% in untreated cells. After five passages, GFP-positive cells comprised 77.8% of the treated population, while the proportion in untreated cells remained stably low (0.41%). We obtained similar results when treating the SSEA1-negative fraction. (When untreated cells from the fifth passage were subsequently treated with AZA, GFP-positive cells appeared at a similar rate as in the initial treatment; Figure 10b). We also found robust induction of the GFP reporter after AZA treatment of BIV1 (-Dox) cells (Figure 10a, 11).

We evaluated the cellular state and developmental potency of the GFP-positive MCV8 and BIV1 cells obtained after AZA treatment and FACS. Both populations stained positive for the stem-cell marker SSEA1. Combined bisulfite restriction analysis (COBRA) revealed significant demethylation of CpGs near the pluripotency-related genes *Dppa5*, *Nanog* and *Utf1* (Figure 12), implying that re-activation was not limited to the GFP-tagged reporters. The viral transgenes showed low or undetectable expression levels (Figure 10c,d) indicating that AZA treatment did not interfere with viral silencing, which is required for full reprogramming⁹, and that the emergence of GFP-positive cells was not caused by viral re-activation. Finally, subcutaneous injection into severe combined immunodeficiency (SCID) mice led to teratoma formation in 3-4 weeks (Figure 10e), demonstrating that the GFP-positive cells had undergone a stable transition to the pluripotent state. (Untreated MCV8 or BIV1 cells did not generate teratomas in the same timeframe).

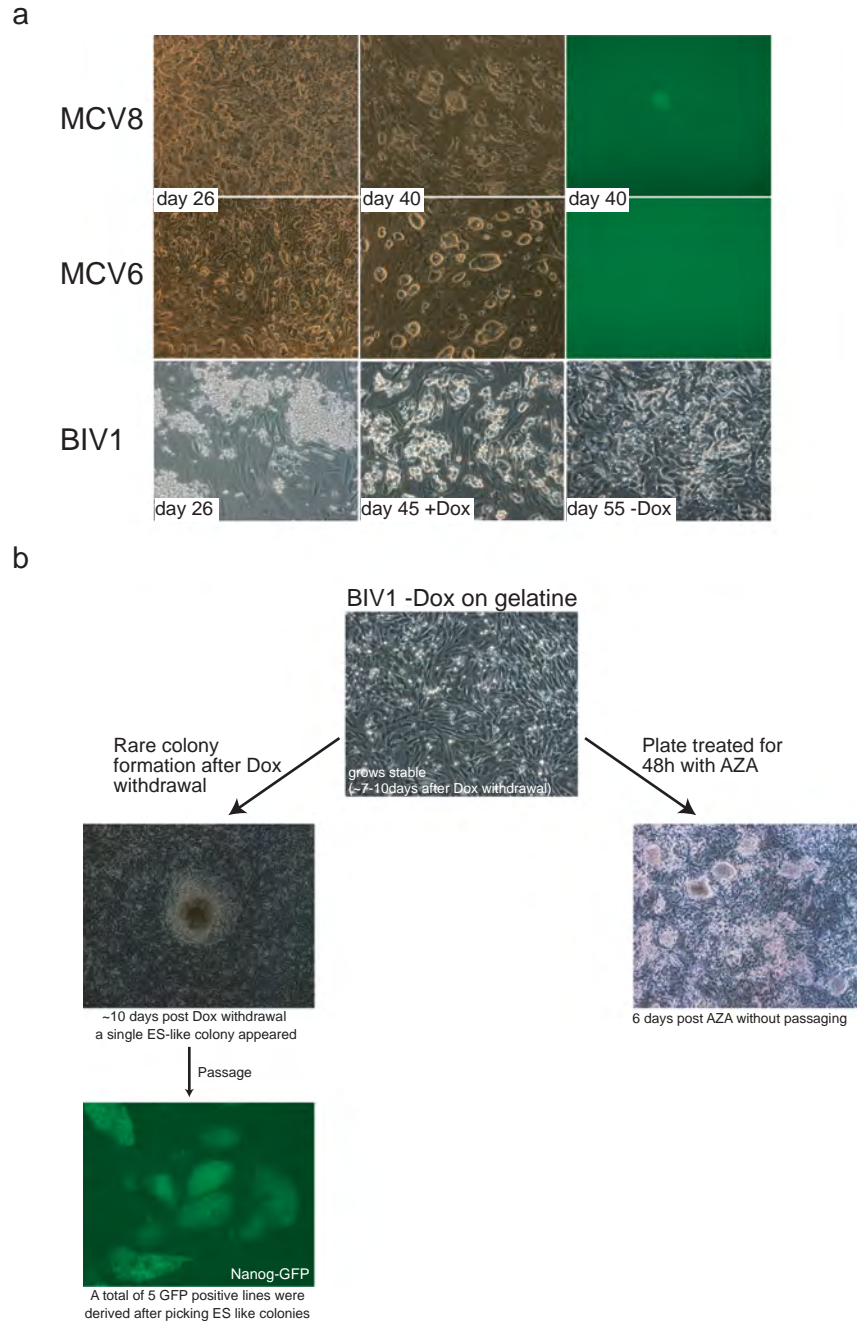


Figure 8. a, Morphology of the three partially reprogrammed cell lines (MCV8, MCV6 and BIV1). After extended passaging we observed positive colonies in MCV8 at very low frequencies, but never in MCV6. **b**, BIV1 -Dox (but not BIV1 +Dox) yielded a few ES like colonies after extensive culturing. In total we were able to isolate 5 Nanog-GFP positive cell lines from this clone (although we cannot rule out that all of them are identical due to the passaging). Using this inducible secondary system, Hanna et al. were able to generate additional B lymphocyte-derived iPS lines at much higher frequencies by adding an additional transcription factor (Hanna et al., 2008).

To exclude nonspecific effects of AZA we treated MCV8 cells with siRNAs or lentiviral shRNAs against *Dnmt1*, which also led to the appearance GFP-positive cells within one passage (up to 1.7%; Figure 11b,c,d). We conclude that transient inhibition of DNA methyltransferase is sufficient to rapidly transition MCV8 and BIV1 cells from a partially reprogrammed state to a pluripotent state.

Populations of lineage-committed cells. We next used the chimera-derived Nanog-GFP MEFs (described above) to test whether AZA treatment could increase the overall reprogramming efficiency. The cells were grown in the presence of Dox from day 1, and AZA was administered for 48 hours starting on day 4, 6, or 8. The reprogramming efficiency was determined by counting ES-like colonies at day 14 (Figure 10f,g).

We found that starting AZA treatment on days 4 and 6 led to high cell death and no overall gain in efficiency. The cell death may reflect the fact that most cells are still in a differentiated state: genome-wide hypomethylation is known to induce apoptosis in differentiated cells, while ES cells are resistant^{27, 28, 29}.

In contrast, there was a consistent 4-fold increase in the number of ES-like colonies in the cultures treated with AZA starting on day 8 ($P_T < 0.007$). Moreover, the vast majority (>95%) of the ES-like colonies were GFP-positive in the treated cells, whereas only a minority (<25%) were GFP-positive in the untreated controls (a proportion consistent with^{9,12-14}). While early AZA treatment is counter productive to reprogramming, there may be a sufficient number of partially reprogrammed cells in the population to outweigh its cytotoxic effect.

We conclude that de-methylation of one or more (unknown) loci is a critical step in the late stages of direct reprogramming, and that inhibition of DNA methyltransferase lowers this kinetic barrier, thereby facilitating transition to pluripotency. A similar role for DNA demethylation has been recently reported during *in vivo* reprogramming in the germ line³⁰.

Transcription factor knockdown facilitates reprogramming

In contrast to the other partially reprogrammed cell lines, MCV6 did not respond to AZA treatment (Figure 7). We also noted above that MCV6 cells never show spontaneous appearance of GFP-positive colonies. We hypothesized that expression of one or more lineage-specifying transcription factors may have stabilized these cells in a more differentiated state than MCV8 or BIV1.

To test this hypothesis, we studied our genome-wide maps and identified lineage-specifying transcription factors that are expressed at low or undetectable levels in MCV8 or iPS cell populations. We transfected MCV6 cells with siRNAs against four transcription factors with >5-

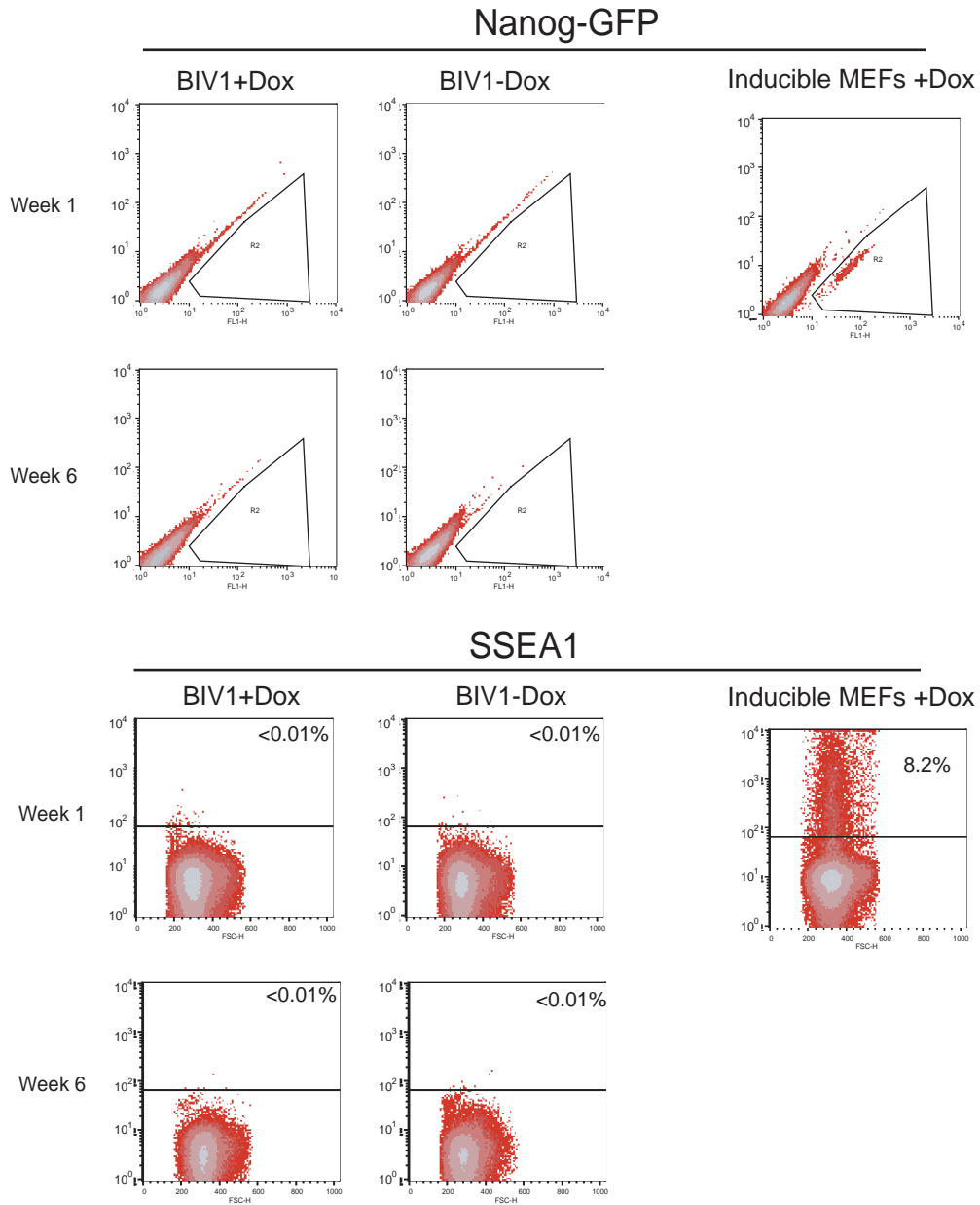


Figure 9. SSEA1 and GFP (Nanog-GFP) were monitored in BIV1 in the presence or absence of doxycycline for 6 weeks. No GFP or SSEA1 was detectable in either population. As an induction control we used Nanog-GFP MEFs (with the same viral integrations as BIV1) that show appearance of SSEA1 and GFP positive after induction with doxycycline.

fold higher expression in MCV6 relative to MCV8 (*Gata6*, *Pax7*, *Pax3* and *Sox9*). This resulted in no significant response. However, when transfection of siRNA targeting any one of the factors was followed by treatment with AZA for 48 hours, GFP-positive cells appeared at a significant frequency in all examined populations (16 independent transfections; Figure 13, 14). For example, targeting the primitive endoderm marker *Gata6*³¹ generated ~2% GFP-positive cells within one passage of the subsequent AZA treatment. By contrast, no GFP-positive cells appeared in populations transfected with negative control siRNAs, or siRNAs targeted against transcription factors not expressed in MCV6 (*Zic1*, *Meox2*) or against *Dnmt1* (7 control populations; $P_{\text{MWU}} < 4 \times 10^{-4}$).

We conclude that re-activation or incomplete repression of lineage-specifying transcription factors during the reprogramming process blocks activation of the endogenous pluripotency regulatory network in MCV6. Transient silencing of one or more of these factors, combined with inhibition of DNA methyltransferase, appears to shift the regulatory balance towards the pluripotent state, which may then be stabilized by autoregulatory feedback¹¹.

Discussion

Several insights emerge from our integrative genomic analyses. First, the Oct4/Sox2/Klf4/c-Myc-based reprogramming process appears to be fairly general, with two independent strategies (constitutive retrovirus or inducible lentivirus) and two distinct cell types (MEFs and B lymphocytes) yielding similar immediate responses, partially reprogrammed states and a similar mechanism for the final transition to pluripotency. Second, cells may fail to successfully reprogram for several apparent reasons: the cells may induce anti-proliferative genes in response to proliferative stress; they may inappropriately activate or fail to repress endogenous or ectopic transcription factors, and become ‘trapped’ in differentiated states; and they may fail to reactivate hypermethylated pluripotency genes. Third, complete reprogramming can be facilitated by direct intervention against these failure modes, such as transient inhibition of DNA methyltransferase and expressed transcription factors.

We expect that further characterization of intermediate states and alternative small molecule treatments will yield critical insights that help facilitate the desired transitions, making reprogramming efficient and safe for use in regenerative medicine. More generally, our data are consistent with a model of development where cellular states are defined by transcription factors and stabilized by epigenetic remodeling. Integrative gene expression and epigenomic profiling provides a powerful tool for defining and guiding directed transitions between these states.

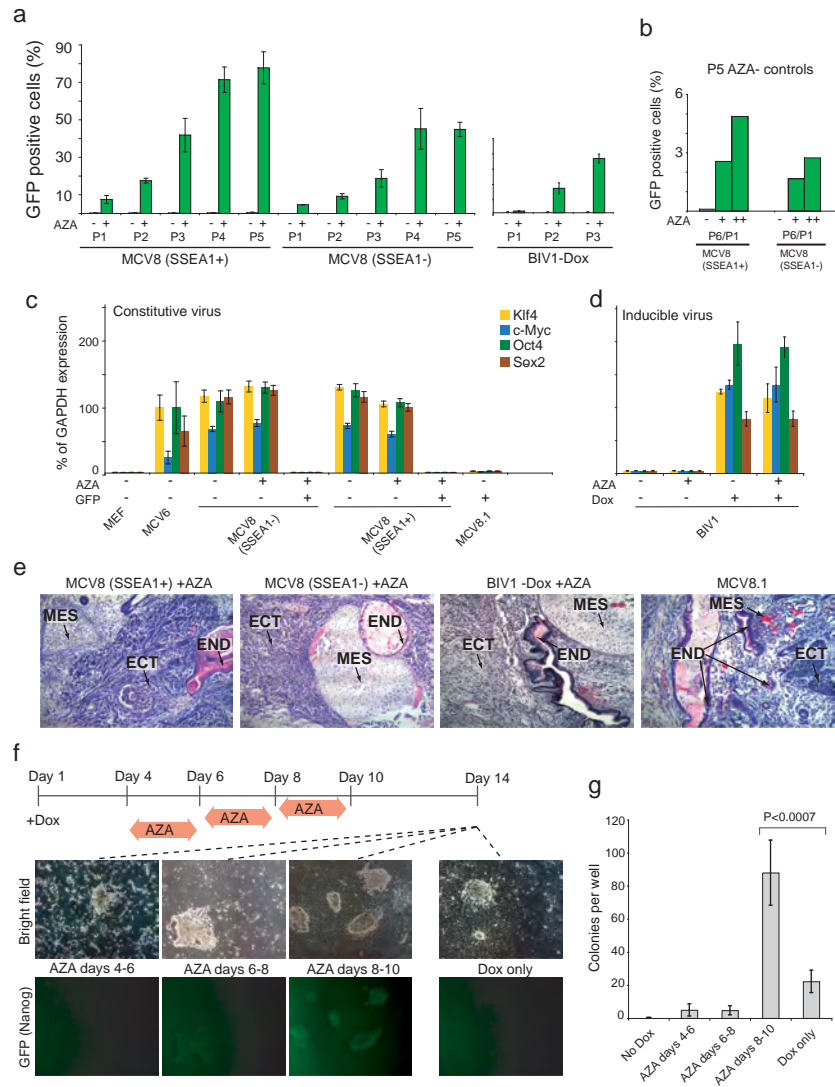


Figure 10. Inhibition of DNMT1 accelerates the transition to pluripotency. a, MCV8 and BIV1 (-Dox) cells were sorted by FACS on SSEA1 status and either exposed to AZA for 48 hours (green) or kept in regular ES medium (grey). The number of Oct4-GFP-positive cells was analyzed over multiple passages (P) by FACS. b, Untreated MCV8 control cells from passage 5 were subsequently subjected to AZA treatment for 48 (+) or 120 hours (++), and resulting Oct4-GFP positive cells were counted after one passage. c, AZA treatment does not influence retroviral expression levels. d, AZA treatment has no influence on lentiviral expression in uninduced or induced BIV1 cells. e, Pluripotency of all AZA treated lines and MCV8.1 was demonstrated by teratoma formation. ECT, ectoderm; MES, mesoderm; END, endoderm. f, Overall efficiency of AZA treatment. Nanog-GFP MEFs were plated on 6-well plates (4 wells per time point with Dox and 2 wells without). Cells were treated with AZA during one of the indicated intervals. On day 14, colony formation was analyzed by fluorescence microscopy (representative panels are show). g, Number of AP-positive, ES-like colonies obtained from each treatment. AZA treatment during days 8-10 resulted in a ~4-fold increase in efficiency over untreated controls. For a, c, d and g, error bars show standard deviations (n = 2, 2, 2 and 4, respectively).

Methods

Viral infections and cell lines: MEFs used to derive primary iPS cell lines by infections with inducible lentiviruses were harvested at 13.5dpc from F1 matings between ROSA26-M2rtTA mice³² and Nanog-GFP mice¹³. Secondary Nanog-GFP MEFs were isolated using neomycin selection. Lentiviral preparation and infection with Doxycycline inducible lentiviruses encoding Oct4, Klf4, c-Myc and Sox2 cDNA driven by the TetO/CMV promoter, were previously described¹³. MCV6 and MCV8 were generated by retroviral infection of Oct4-GFP MEFs as described previously¹².

Cell culture: Infected MEFs or secondary inducible MEFs¹⁵ were cultured and expanded in standard ES medium and conditions¹². Culture and viral induction was done as described^{13,15} and BIV1 was obtained as a stable line and grown under regular ES conditions in the presence or absence of 2 µg/ml of Doxycycline. AZA treatment was performed for 48 hours or as indicated at a concentration of 0.5µM. Higher doses showed similar effects but increased toxicity.

Expression profiling: RNA was isolated using TRIZOL followed by a second round of purification using RNeasy Columns (Qiagen). RNA was then processed and analyzed as described elsewhere²². Absolute expression values were RMA normalized, truncated to ≥ 20 , and visualized using GenePattern (<http://www.broad.mit.edu/cancer/software/genepattern/>).

Chromatin IP and Illumina/Solexa Sequencing. Cells were harvested and cross-linked with formaldehyde (final concentration 1%) for 10 min at 37°C. Washed twice with cold PBS (plus Protease inhibitors), frozen and kept at -80°C. Chromatin IP, library construction, sequencing, identification of enriched intervals and chromatin state classification was done as described previously²².

Bisulfite Sequencing and COBRA: Genomic DNA was isolated and bisulfite conversion was performed in a thermocycler using the Qiagen EpiTect Kit according to manufacturers instructions with two additional cycles (5min 99°C and 3h 60°C) at the end. When using 2µg genomic DNA as starting material, converted DNA was eluted in 40µl EB (Qiagen) and 2µl were used and amplified with previously described primer sets²³ and the following additional primer pairs (Cyc2F: GAAGGATTAATAGATGTATAAGAAAATAT; Cyc2R: AAACCCTAATTATAACAAATACAAC; Sox2F: GGTTTAGGAAAAGGTTGGGAATA; Sox2R: AACCAAATAAAAACAAAACCCATAA). PCR was done in 25µl reactions using EpiTect MSP Kit (Qiagen) mastermix according to manufactures instructions with a 45sec annealing step at 50°C (35cycles). PCR products were gel purified, TOPO cloned (Invitrogen) and sequenced. Combined bisulfite restriction analysis (COBRA) for Dppa5, Nanog and UTF1 was

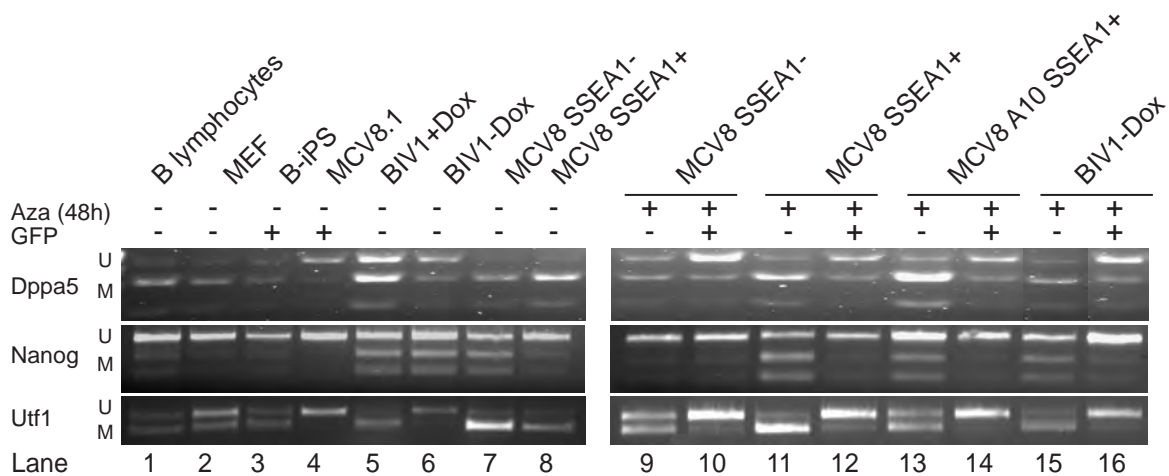


Figure 12. Combined Bisulfite Restriction Analysis (COBRA) of CpG dinucleotides near Dppa5, Nanog and Utf1. The bisulfite treated and PCR amplified product of Dppa5 was digested with Taq1 (TCGA). The products for Nanog and Utf1 were digested with HpyCHIV (ACGT). The top band indicates unmethylated (uncut; U) CpGs and the lower band(s) methylated CpGs in the recognition sequence of the respective enzyme. The left panel shows several donor, partially reprogrammed and reprogrammed cell lines, including the BIV1 +/-Dox. As can be seen in lane 6, simply withdrawing doxycycline (Dox) resulted in further reprogramming and loss of methylation, in particular at Utf1. Nanog however did not change, which is consistent with the fact that these cells remain Nanog-GFP negative. Notably, MEFs often show mixed methylation patterns (compare Figure 6 and (Imamura et al. 2006)). The right panel shows analysis of the GFP positive and negative fractions post AZA treatment. Clear loss of methylation can be seen in the positive fraction, but not in the negative fraction. A10 is a subclone picked from the SSEA1 positive fraction prior to AZA treatment.

done using 15µl of the gel purified DNA. Dppa5 was digested for 4h at 65°C with Taq1 (TCGA). Nanog and UTF1 were digested with HpyCHIV (ACGT) for 4h at 37°C. Digested products were run on 2% agarose gels.

Knockdown of transcription factors and Dnmt1: Reverse transfections were done in 24 well dishes according to manufacturers instructions using the siPORT NeoFX Transfection Agent (Ambion). The following Silencer Select (Ambion/ABI) siRNAs were used: Negative Control siRNA (#4390843), positive control Cy3 GAPDH siRNA (#AM4649), Pax 3 siRNA (s71259, s71260), Pax 7 siRNA (s71271, s71272), Gata6 siRNA (s66489, s66490), Sox9 siRNA (s74192, s74193), Meox2 siRNA (s69792, s69793), Zic1 siRNA (s76384, s76385) and Dnmt1 siRNA (s65071, s65072). Dnmt1 was stably knocked down using five independent shRNAs from the RNAi consortium (TRC; http://www.broad.mit.edu/genome_bio/trc/). shRNA1 (TRCN0000039024; target: GCTGACACTAAGCTGTTTGTA), shRNA2 (TRCN0000039025; target: GCCTTTACTTTCAACATCAAA), shRNA3 (TRCN0000039026; target: CCGCACTTACTCCAAGTTCAA), shRNA4 (TRCN0000039027; target: CCCGAAGATCAACTCACAAA) and shRNA5 (TRCN0000039028; target: GCAAAGAGTATGAGCCAATAT). MCV8 cells were infected overnight and selected in puromycin (final: 2µg/ml) for 48h.

Quantitative RT-PCR. Total RNA was isolated using Rneasy Kit (Qiagen). Three micrograms of total RNA was treated with DNase I to remove potential contamination of genomic DNA using a DNA Free RNA kit (Zymo Research, Orange, CA). Retroviral expression levels were determined as described previously⁹. For inducible lentiviral expression one microgram of DNase I-treated RNA was reverse transcribed using a First Strand Synthesis kit (Invitrogen) and ultimately resuspended in 100 µl of water. Quantitative PCR analysis was performed in triplicate using 1/50 of the reverse transcription reaction in an ABI Prism 7000 (Applied Biosystems, Foster City, CA) with Platinum SYBR green qPCR SuperMix-UDG with ROX (Invitrogen). Primers used for amplification were as follows: c-Myc: F, 5'-ACCTAACTCGAGGAGGAGCTGG-3' and R, 5'-TCCACATAGCGTAAAAGGAGC-3'; Klf4: F, 5'-ACACTGTCTTCCCACGAGGG-3' and R, 5'-GGCATTAAAGCAGCGTATCCA-3'; Sox2: F, 5'-CATTAACGGCACACTGCCC-3' and R, 5'-GGCATTAAAGCAGCGTATCCA-3'; Oct4: F, 5'-AGCCTGGCCTGTCTGTCCTC-3' and R, 5'-GGCATTAAAGCAGCGTATCCA-3'. To ensure equal loading of cDNA into RT reactions, GAPDH mRNA was amplified using the following primers: F, 5'-TTCACCACCATGGAGAAGGC-3'; and R, 5'-CCCTTTTGGCTCCACCCT-3'. Data were extracted from the linear range of amplification. All graphs of qRT-PCR data shown represent

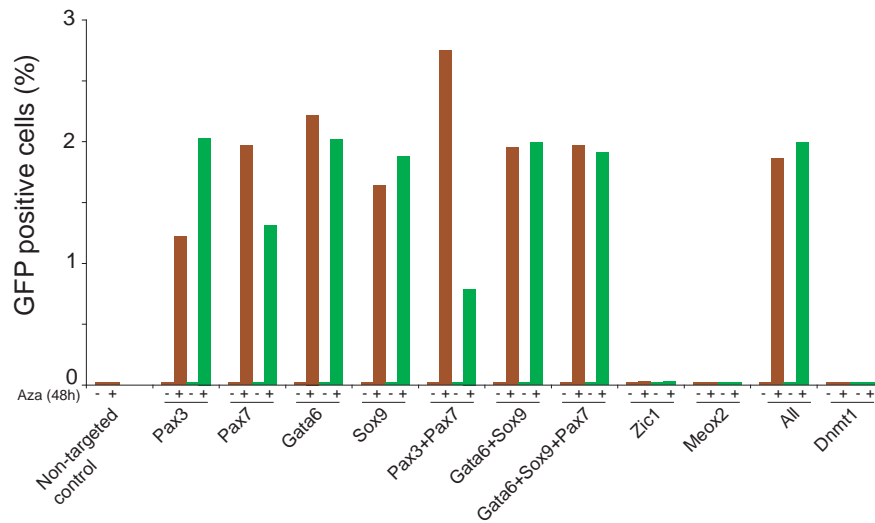


Figure 13. Transcription factor knockdown facilitates reprogramming. MCV6 cells were plated onto 24-well dishes and transfected with siRNAs targeting expressed (Pax7, Pax3, Gata6, Sox9) or non-expressed (Zic1, Meox2) transcription factors. One plate was kept in ES medium and the second was exposed to AZA for 48 hours. Two independent siRNA sequences were used for duplicate experiments (red and green). FACS analysis was performed 48 hours after AZA treatment (96 hours after transfection) without passaging. The transfection efficiency was estimated as ~20% using Cy3-coupled GADPH control siRNA.

samples of RNA that were DNase treated, reverse transcribed, and amplified in parallel to avoid variation inherent in these procedures.

Flow cytometry analysis and cell sorting. The following fluorescently conjugated antibodies (PE, FITC, Cy-Chrome or APC labeled) were used for FACS analysis and cell sorting: anti-SSEA1 (RnD systems), anti-Ig κ , anti-Ig λ 1,2,3, anti-CD19, anti-B220, anti-sIgM, anti-sIgD (all obtained from BD-Biosciences). Cell sorting was performed by using FACS-Aria (BD-Biosciences), and consistently achieved cell sorting purity of >97%. For determining GFP-positive cell numbers by FACS we counted >50,000 cells.

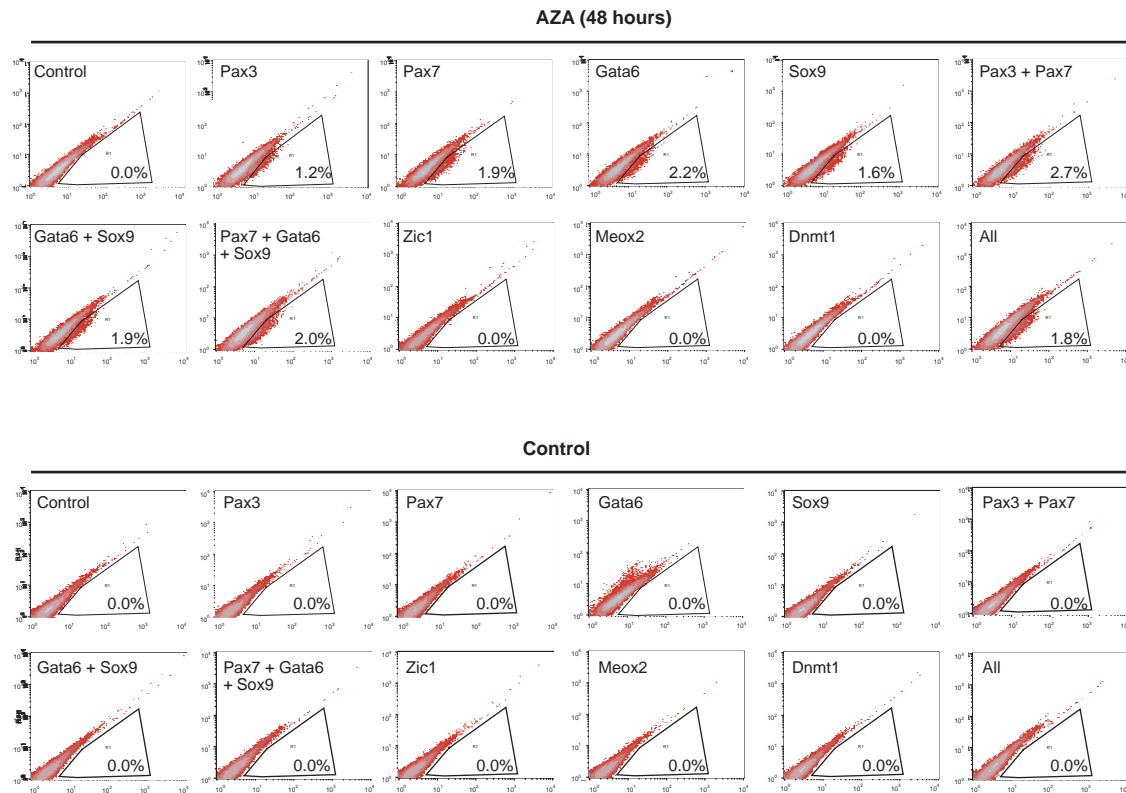


Figure 14. FACS analysis for GFP positive cells after siRNA-mediated transcription factor knockdown. Top panels show transfections with the indicated single siRNAs or combinations 96 hours after transfection and 48 hours after AZA treatment. The percent GFP positive cells are shown in the gate. Bottom panel shows that the same siRNAs without the subsequent AZA treatment show no GFP positive cells. While Pax3, Pax7, Gata6 and Sox9 are expressed in MCV6, Zic1 and Meox2 are not. The lack of GFP positive cells in the latter wells serves as an additional control.

References

1. T. Aoi, K. Yae, M. Nakagawa et al., Generation of Pluripotent Stem Cells from Adult Mouse Liver and Stomach Cells. *Science* (2008). DOI: 10.1126/science.1154884
2. N. Maherali, R. Sridharan, W. Xie et al., Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cells* (1), 55-77 (2007).
3. M. Nakagawa, M. Koyanagi, K. Tanabe et al., Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat Biotechnol* 26 (1), 101-106 (2008).
4. K. Okita, T. Ichisaka, and S. Yamanaka, Generation of germline-competent induced pluripotent stem cells. *Nature* 448 (7151), 313-317 (2007).
5. I. H. Park, R. Zhao, J. A. West et al., Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* 451 (7175), 141-146 (2008).
6. K. Takahashi, K. Tanabe, M. Ohnuki et al., Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131 (5), 861-872 (2007).
7. K. Takahashi and S. Yamanaka, Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126 (4), 663-676 (2006).
8. J. Yu, M. A. Vodyanik, K. Smuga-Otto et al., Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318 (5858), 1917-1920 (2007).
9. M. Wernig, A. Meissner, R. Foreman et al., In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* 448 (7151), 318-324 (2007).
10. S. Yamanaka, Strategies and new developments in the generation of patient-specific pluripotent stem cells. *Cell Stem Cells* 1, 39-49 (2007).
11. R. Jaenisch and R. Young, Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* 132 (4), 567-582 (2008).
12. A. Meissner, M. Wernig, and R. Jaenisch, Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat Biotechnol* 25 (10), 1177-1181 (2007).
13. T. Brambrink, R. Foreman, G. Welstead et al., Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell* 2, 151-159 (2008).
14. M. Stadtfeld, N. Maherali, D. Breault et al., Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. *Cell Stem Cell* 2, 230-240 (2008).
15. J. Hanna, S. Markoulaki, P. Schorderet et al., Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. *Cell* 133 (2), 250-264 (2008).
16. S. Adhikary and M. Eilers, Transcriptional regulation and transformation by Myc proteins. *Nat Rev Mol Cell Biol* 6 (8), 635-645 (2005).
17. B. D. Rowland and D. S. Peeper, KLF4, p21 and context-dependent opposing forces in cancer. *Nat Rev Cancer* 6 (1), 11-23 (2006).
18. M. A. Gregory, Y. Qi, and S. R. Hann, The ARF tumor suppressor: keeping Myc on a leash. *Cell Cycle* 4 (2), 249-252 (2005).
19. W. M. Rideout, 3rd, T. Wakayama, A. Wutz et al., Generation of mice from wild-type and targeted ES cells by nuclear cloning. *Nat Genet* 24 (2), 109-110 (2000).
20. W. E. Lowry, L. Richter, R. Yachechko et al., Generation of human induced pluripotent stem cells from dermal fibroblasts. *Proceedings of the National Academy of Sciences of the United States of America* 105 (8), 2883-2888 (2008).

21. K. W. Orford and D. T. Scadden, Deconstructing stem cell self-renewal: genetic insights into cell-cycle regulation. *Nat Rev Genet* 9 (2), 115-128 (2008).
22. T. S. Mikkelsen, M. Ku, D. B. Jaffe et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448 (7153), 553-560 (2007).
23. M. Imamura, K. Miura, K. Iwabuchi et al., Transcriptional repression and DNA hypermethylation of a small set of ES cell marker genes in male germline stem cells. *BMC Dev Biol* 6, 34 (2006).
24. J. Silva and A. Smith, Capturing pluripotency. *Cell* 132 (4), 532-536 (2008).
25. L. S. Lim, Y. H. Loh, W. Zhang et al., *Zic3* is required for maintenance of pluripotency in embryonic stem cells. *Mol Biol Cell* 18 (4), 1348-1358 (2007).
26. B. E. Bernstein, A. Meissner, and E. S. Lander, The mammalian epigenome. *Cell* 128 (4), 669-681 (2007).
27. L. Jackson-Grusby, C. Beard, R. Possemato et al., Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nat Genet* 27 (1), 31-39 (2001).
28. H. Lei, S. P. Oh, M. Okano et al., De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* 122 (10), 3195-3205 (1996).
29. A. Meissner, A. Gnirke, G. W. Bell et al., Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 33 (18), 5868-5877 (2005).
30. P. Hajkova, K. Ancelin, T. Waldmann et al., Chromatin dynamics during epigenetic reprogramming in the mouse germ line. *Nature* 452 (7189), 877-881 (2008).
31. A. M. Singh, T. Hamazaki, K. E. Hankowski et al., A heterogeneous expression pattern for *Nanog* in embryonic stem cells. *Stem Cells* 25 (10), 2534-2542 (2007).
32. C. Beard, K. Hochedlinger, K. Plath et al., Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. *Genesis* 44 (1), 23-28 (2006).
33. A. Meissner, T. S. Mikkelsen, et al., Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454(7205), 766-770 (2008).

[This page is intentionally left blank]

Chapter 10: Future directions

[This page is intentionally left blank]

The contributions described in this thesis suggest several future directions for research on the structure, evolution and function of the human genome.

Our comparative analysis of mammalian genome sequences suggest that the rate of turnover of ancestral functional elements has been low within the infraclass of eutherian mammals. Sequencing additional eutherian genomes should therefore yield increasing specificity for detection of functional elements without a significant loss of sensitivity. The highly correlated and nearly-neutral patterns of evolution in protein-coding sequences also suggest that new methods for detecting positive selection in regulatory elements must be developed to help define the molecular basis of morphological evolution. Moreover, the significant contribution of transposon-derived sequences to the evolution of eutheria-specific conserved elements suggests that primate- or human-specific regulatory elements might have similar origins. Development of new analysis methods might be required to study these highly repetitive sequences.

Our studies on genome-wide patterns of chromatin modifications provide a framework for comprehensive characterization of chromatin state across a variety of mammalian cell populations. We focused on DNA methylation and a small number of relatively well-understood histone lysine methylation marks. The genome-wide distributions of additional histone modifications, such as acetylation, ubiquitination, arginine methylation, as well as the use of histone variants, remain to be explored. Generating a more comprehensive compendium of chromatin state maps should help define signatures that clearly differentiate between pluripotent and differentiated cells or between ‘normal’ and ‘abnormal’ epigenetic regulation.

Projects devoted to sequencing additional mammalian genomes and generating chromatin state maps from various mammalian cell and tissue types are already in progress. It is important to note, however, that comparative sequence analysis is largely limited to detecting the *location* of functional sequence elements; similarly, generation of chromatin state maps is largely limited to detecting the *location* of specific chromatin modifications. Understanding the *function* or specific role of each sequence element or chromatin mark requires directed experimentation in relevant biological contexts. Here, we describe a few potential research projects designed to assign specific biological functions to sequence elements involved in morphological development or cellular differentiation, and to further explore the direct relationship between chromatin state and cell state.

Dissect the regulatory architecture of key developmental loci

Embryonic patterning and morphological development is driven, at least in part, by combinatorial expression of transcription factors and signaling molecules. Our results show that more than half of

highly conserved non-coding sequences, and more than a quarter of all conserved non-coding sequences, in mammalian genomes are clustered in long ‘gene deserts’ surrounding ~200 coding sequences for transcription factors, signaling molecules and axon guidance receptors. It seems reasonable to hypothesize that many of these conserved non-coding sequences are regulatory elements that control the complex developmental expression patterns of the nearby coding sequences. If this is the case, associating each discrete functional element in these regions with the anatomical location or developmental stage at which they influence gene expression would be a major step towards describing how morphology is encoded in our genome sequence.

Understanding the architecture of these loci would be critical to designing an effective experimental strategy. Regulatory elements might act independently to activate gene expression at specific anatomical locations or developmental stages; multiple redundant or mutually reinforcing elements might be required to activate expression; or gene expression levels might be determined by the sum of interactions of multiple synergistic and antagonistic elements spread across the locus. Regulatory elements on one allele might also influence the promoter and protein-coding sequences on the other allele through transvection.

Systematic functional studies in the context of human development is intractable. It seems likely that the basic regulatory architecture of these regions are conserved and amenable to rapid dissection in vertebrate model systems such as zebrafish and frog. However, the low sensitivity of non-coding sequence alignments between mammals and non-mammalian vertebrates and the presumed high rate of evolutionary innovation across these loci suggest that studies in distantly related vertebrates would not be effective for uncovering the specific functions of most regulatory elements found in the mammalian genomes. Focusing on mouse development might therefore be the best trade-off between evolutionary divergence and experimental tractability.

The ease and high efficiency of genetic engineering in murine ES cells suggest a general experimental approach. An appropriate reporter gene can first be inserted into a locus of interest using homologous recombination. Embryos at different stages of development can then be generated from clonal ES cells containing the reporter (“unmodified clones”), potentially using tetraploid complementation. The expression pattern of the reporter can be determined using high-resolution *in situ* hybridization on tissue sections or on whole specimens using optical projection tomography. A second round of recombination can then be used to generate a series of sub-clones with deletions across the flanking regulatory regions (“deletion clones”). Tiling 20 kb deletions with 10 kb overlap would require a few hundred successful targeting events to cover a typical locus. Obtaining heterozygous deletions linked to the reporter gene should be sufficient, but homozygous

deletions could be generated if transvection occurs. In parallel, each of the deleted regions can be inserted along with a reporter gene at a different chromosomal location (“insertion clones”). Embryos can then be generated from each of these modified clones and the expression patterns of their reporter genes can be compared to that of the unmodified clone to infer regulatory activities and architecture. For example, if regulatory elements act independently, a deletion clone should show loss of reporter expression at a particular location, while a corresponding insertion clone should show specific expression at the same location. If there are redundant regulatory elements, a deletion clone might show no change in reporter expression, while the corresponding insertion clone should still show specific expression.

Once a course-grained map of regulatory elements and their interactions across a locus has been generated, regions of particular interest could be fine-mapped to identify and study the minimal sequences required for their regulatory activities. Sequence analysis and perturbation experiments can be used to identify the *trans*-acting factors that interact with each element. Regions that appear to control anatomical structures that differ significantly between mice and humans can be replaced with orthologous human DNA to test whether it can induce a more ‘human-like’ expression pattern in the mouse background.

If pilot studies on one or a small number of loci turn out to be informative, this approach can in principle be extended to all key developmental loci in the mammalian genome. Covering 200 loci at 10 kb resolution would require generation of ~40,000 insertion and deletion clones, which is not much different from the scale of ongoing efforts to generate comprehensive gene-trap or knockout clone collections.

Dissect regulatory networks that control cellular differentiation

Once cells have organized themselves in anatomical structures, they differentiate into the specialized cell types that make up our various tissues. Cellular differentiation typically involves (1) induction of tissue-specific effector genes and (2) silencing of genes that drive cell growth and proliferation. In each cell type, these processes are controlled by specialized transcriptional regulatory networks that are composed of *trans*-acting proteins or RNA molecules and *cis*-acting regulatory elements.

Cellular differentiation is, at least in some cases, simpler to study than developmental patterning and morphology. Good *in vitro* models have been developed for several differentiation processes in both human and mouse, such as adipogenesis, myogenesis and some variations of neurogenesis. These models can be used to obtain relatively homogeneous cell populations from

multiple stages of the differentiation process. Comparison of *in vitro* differentiated cells to their *in vivo* equivalents should help to filter out any cell culture-related artifacts in the models. In some cases, such as hematopoiesis, cells at different stages can also be sorted directly from tissue samples. Over the last few decades, candidate gene studies and expression profiling have successfully identified key transcription factors and other *trans*-acting factors in all of these differentiation processes. Identification of the *cis*-regulatory elements they interact with have been significantly more challenging, however. The methods developed in this thesis, in particular ChIP-Seq, have the potential to accelerate identification and functional annotation of differentiation-related regulatory elements.

Our results show that ChIP-Seq can be used to generate chromatin state maps from cell populations at different stages of a differentiation process. If enough stages are covered, these maps should locate essentially all *cis*-regulatory elements utilized in the process. In principle, ChIP-Seq can also be used to map all binding sites for every expressed sequence-specific transcription factor. Combining data from such experiments with genome-wide expression profiling using microarrays or RNA sequencing should yield an essentially complete inventory of the *trans*- and *cis*-acting components of the regulatory network.

Once all *trans*- and *cis*-acting components of the regulatory network have been identified, standard techniques can be used to define their specific functions and interactions. RNA interference, over-expression and knockouts can be used to examine the role of individual *trans*-acting factors. Transfection of reporter constructs or modification of the endogenous loci can be used to examine the function and composition of individual *cis*-regulatory elements. Most of these techniques should be amenable to high-throughput experimentation and allow efficient interrogation of every component and interaction in a regulatory network.

Probe the relationship between chromatin state and cell state

The regulatory networks that control morphology and differentiation are specified by the genome sequence, but various evidence suggest that chromatin structure can influence when and where individual components of these networks are accessible. Nuclear transfer and direct reprogramming experiments have demonstrated that differentiation-induced changes in chromatin state are, within the resolution of our assays, completely reversible given the right signals. However, the relatively low efficiency of these methods and the observation that inhibiting chromatin modifying enzymes can improve the efficiency suggest that chromatin modifications can contribute to stabilizing a

differentiated cell state. A key open question is whether knowing the chromatin state of a cell can help predict how it will respond to new environmental signals or other perturbations.

Our results show that, in any given cell type, gene promoters display one of a limited number of distinct chromatin states. CpG-rich promoters display at least four different states: H3K4 methylated, H3K27 methylated, bivalent or DNA methylated. CpG-poor promoters display at least two different states: H3K4 methylated or DNA methylated. Additional states involving H3K9 methylation and other modifications might also exist. These states are closely correlated with gene expression levels and cellular differentiation – but do they also predict how genes respond to new regulatory signals? For example, if cells committed to muscle differentiation are exposed to a signal that induces pluripotent cells to commit to a neural fate, will genes with H3K27 methylated promoters be less responsive to this signal than genes with H3K4 methylated or bivalent promoters? Will DNA hypermethylated promoters be less responsive than H3K27 methylated promoters? Will there be differences in the responsiveness of genes with CpG-rich and CpG-poor promoters? And if there are any such differences, are they a direct consequence of differences in chromatin accessibility? Our analysis of direct reprogramming suggest positive answers to these questions, but more systematic investigations are needed to establish the generality of these results.

One experimental approach to these questions would be to subject cells with different chromatin states to identical perturbations and then measure the resulting changes in gene expression patterns or chromatin states. Perturbations might involve introducing ectopic transcription factors or changing the extracellular environment. If chromatin accessibility is a critical factor in stabilizing gene expression patterns, there should be a statistical correlation between the magnitude or kinetics of the response of each gene and its pre-perturbation chromatin state across different cell types. If a correlation is detected, repeating the experiment while interfering with the implicated regulatory pathways should help differentiate between direct and indirect effects. For example, if bivalent promoters are found to be less responsive than H3K4 methylated promoters, one might predict that interfering with the function of Polycomb group proteins should result in an increased response from genes in the former state. Small molecules have been or are being developed to inhibit various families of chromatin modifying and remodeling enzymes. RNA interference and genetic ablation can also be used to control specific enzymes.

A variation on this approach would be to introduce ectopic transcription factors and directly assay their binding sites across the genome using ChIP-Seq. If a correlation between differential binding sites and chromatin states across different cell types is detected, small molecules or RNA interference could again be used to gauge whether it represents a direct effect. A related experiment

would be to examine correlations between retro- or lentiviral integration patterns and chromatin state, which would be of relevance to gene therapy and related applications.

There are at least three potential technical challenges to this approach. First, chromatin state maps generated by ChIP-Seq are generated by averaging signals from a relatively large population of cells (at least several hundred thousand). Any variation in chromatin state between cells in the population, due to cell cycle progression, cryptic differentiation, stochastic effects or other causes, would introduce noise and reduce any observable correlation with perturbation responses. It will therefore be critical to develop quantitative models to estimate and control for population heterogeneity. Second, differences in the expression of signaling receptors or transcriptional co-factors between different cell types can be expected to influence their gene expression responses, independent of any effect of chromatin accessibility. It might therefore be important to choose perturbations for which such dependencies are well-understood. Finally, because the same regulatory pathways often control both differentiation- and growth-related genes, it can be difficult to interfere with chromatin modifying pathways without compromising the viability or well-being of the perturbed cells. The development of methods for modulating chromatin state at specific loci would therefore be highly desirable. If these challenges can be overcome, integrated analysis of gene expression patterns and chromatin state maps before and after relevant cellular perturbations should become a powerful tool for elucidating the direct, functional relationship between chromatin state and cell state.

Appendices

1. The Chimpanzee Sequencing and Analysis Consortium (Mikkelsen, T. S. *et al.*). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).
2. Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).
3. Mikkelsen, T. S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167-177 (2007).
4. Xie, X., Mikkelsen, T. S. *et al.* Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A* **104**, 7145-7150 (2007)
5. Bernstein, B. E., Mikkelsen, T. S. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315-326 (2006).
6. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).
7. Meissner, A.*, Mikkelsen, T. S.* *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770 (2008).
8. Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis *Nature* **454**, 49-55 (2008).

[This page is intentionally left blank]

Initial sequence of the chimpanzee genome and comparison with the human genome

The Chimpanzee Sequencing and Analysis Consortium*

Here we present a draft genome sequence of the common chimpanzee (*Pan troglodytes*). Through comparison with the human genome, we have generated a largely complete catalogue of the genetic differences that have accumulated since the human and chimpanzee species diverged from our common ancestor, constituting approximately thirty-five million single-nucleotide changes, five million insertion/deletion events, and various chromosomal rearrangements. We use this catalogue to explore the magnitude and regional variation of mutational forces shaping these two genomes, and the strength of positive and negative selection acting on their genes. In particular, we find that the patterns of evolution in human and chimpanzee protein-coding genes are highly correlated and dominated by the fixation of neutral and slightly deleterious alleles. We also use the chimpanzee genome as an outgroup to investigate human population genetics and identify signatures of selective sweeps in recent human evolution.

More than a century ago Darwin¹ and Huxley² posited that humans share recent common ancestors with the African great apes. Modern molecular studies have spectacularly confirmed this prediction and have refined the relationships, showing that the common chimpanzee (*Pan troglodytes*) and bonobo (*Pan paniscus* or pygmy chimpanzee) are our closest living evolutionary relatives³. Chimpanzees are thus especially suited to teach us about ourselves, both in terms of their similarities and differences with human. For example, Goodall's pioneering studies on the common chimpanzee revealed startling behavioural similarities such as tool use and group aggression^{4,5}. By contrast, other features are obviously specific to humans, including habitual bipedality, a greatly enlarged brain and complex language⁵. Important similarities and differences have also been noted for the incidence and severity of several major human diseases⁶.

Genome comparisons of human and chimpanzee can help to reveal the molecular basis for these traits as well as the evolutionary forces that have moulded our species, including underlying mutational processes and selective constraints. Early studies sought to draw inferences from sets of a few dozen genes^{7–9}, whereas recent studies have examined larger data sets such as protein-coding exons¹⁰, random genomic sequences^{11,12} and an entire chimpanzee chromosome¹³.

Here we report a draft sequence of the genome of the common chimpanzee, and undertake comparative analyses with the human genome. This comparison differs fundamentally from recent comparative genomic studies of mouse, rat, chicken and fish^{14–17}. Because these species have diverged substantially from the human lineage, the focus in such studies is on accurate alignment of the genomes and recognition of regions of unusually high evolutionary conservation to pinpoint functional elements. Because the chimpanzee lies at such a short evolutionary distance with respect to human, nearly all of the bases are identical by descent and sequences can be readily aligned except in recently derived, large repetitive regions. The focus thus turns to differences rather than similarities. An observed difference at a site nearly always represents a single event, not multiple indepen-

dent changes over time. Most of the differences reflect random genetic drift, and thus they hold extensive information about mutational processes and negative selection that can be readily mined with current analytical techniques. Hidden among the differences is a minority of functionally important changes that underlie the phenotypic differences between the two species. Our ability to distinguish such sites is currently quite limited, but the catalogue of human–chimpanzee differences opens this issue to systematic investigation for the first time. We would also hope that, in elaborating the few differences that separate the two species, we will increase pressure to save chimpanzees and other great apes in the wild.

Our results confirm many earlier observations, but notably challenge some previous claims based on more limited data. The genome-wide data also allow some questions to be addressed for the first time. (Here and throughout, we refer to chimpanzee–human comparison as representing hominids and mouse–rat comparison as representing murids—of course, each pair covers only a subset of the clade.) The main findings include:

- Single-nucleotide substitutions occur at a mean rate of 1.23% between copies of the human and chimpanzee genome, with 1.06% or less corresponding to fixed divergence between the species.
- Regional variation in nucleotide substitution rates is conserved between the hominid and murid genomes, but rates in subtelomeric regions are disproportionately elevated in the hominids.
- Substitutions at CpG dinucleotides, which constitute one-quarter of all observed substitutions, occur at more similar rates in male and female germ lines than non-CpG substitutions.
- Insertion and deletion (indel) events are fewer in number than single-nucleotide substitutions, but result in ~1.5% of the euchromatic sequence in each species being lineage-specific.
- There are notable differences in the rate of transposable element insertions: short interspersed elements (SINEs) have been threefold more active in humans, whereas chimpanzees have acquired two new families of retroviral elements.

*Lists of participants and affiliations appear at the end of the paper.

- Orthologous proteins in human and chimpanzee are extremely similar, with ~29% being identical and the typical orthologue differing by only two amino acids, one per lineage.
- The normalized rates of amino-acid-altering substitutions in the hominid lineages are elevated relative to the murid lineages, but close to that seen for common human polymorphisms, implying that positive selection during hominid evolution accounts for a smaller fraction of protein divergence than suggested in some previous reports.
- The substitution rate at silent sites in exons is lower than the rate at nearby intronic sites, consistent with weak purifying selection on silent sites in mammals.
- Analysis of the pattern of human diversity relative to hominid divergence identifies several loci as potential candidates for strong selective sweeps in recent human history.

In this paper, we begin with information about the generation, assembly and evaluation of the draft genome sequence. We then explore overall genome evolution, with the aim of understanding mutational processes at work in the human genome. We next focus on the evolution of protein-coding genes, with the aim of characterizing the nature of selection. Finally, we briefly discuss initial insights into human population genetics.

In recognition of its strong community support, we will refer to chimpanzee chromosomes using the orthologous numbering nomenclature proposed by ref. 18, which renumbers the chromosomes of the great apes from the International System for Human Cytogenetic Nomenclature (ISCN; 1978) standard to directly correspond to their human orthologues, using the terms 2A and 2B for the two ape chromosomes corresponding to human chromosome 2.

Genome sequencing and assembly

We sequenced the genome of a single male chimpanzee (Clint; Yerkes pedigree number C0471; Supplementary Table S1), a captive-born descendant of chimpanzees from the West Africa subspecies *Pan troglodytes verus*, using a whole-genome shotgun (WGS) approach^{19,20}. The data were assembled using both the PCAP and ARACHNE programs^{21,22} (see Supplementary Information 'Genome sequencing and assembly' and Supplementary Tables S2–S6). The former was a *de novo* assembly, whereas the latter made limited use of human genome sequence (NCBI build 34)^{23,24} to facilitate and confirm contig linking. The ARACHNE assembly has slightly greater continuity (Table 1) and was used for analysis in this paper. The draft genome assembly—generated from ~3.6-fold sequence redundancy of the autosomes and ~1.8-fold redundancy of both sex chromosomes—covers ~94% of the chimpanzee genome with >98% of the sequence in high-quality bases. A total of 50% of the sequence (N50) is contained in contigs of length greater than 15.7 kilobases (kb) and supercontigs of length greater than 8.6 megabases (Mb). The assembly represents a consensus of two haplotypes, with one allele from each heterozygous position arbitrarily represented in the sequence.

Assessment of quality and coverage. The chimpanzee genome assembly was subjected to rigorous quality assessment, based on comparison to finished chimpanzee bacterial artificial chromosomes (BACs) and to the human genome (see Supplementary Information

'Genome sequencing and assembly' and Supplementary Tables S7–S16).

Nucleotide-level accuracy is high by several measures. About 98% of the chimpanzee genome sequence has quality scores²⁵ of at least 40 (Q40), corresponding to an error rate of $\leq 10^{-4}$. Comparison of the WGS sequence to 1.3 Mb of finished BACs from the sequenced individual is consistent with this estimate, giving a high-quality discrepancy rate of 3×10^{-4} substitutions and 2×10^{-4} indels, which is no more than expected given the heterozygosity rate (see below), as 50% of the polymorphic alleles in the WGS sequence will differ from the single-haplotype BACs. Comparison of protein-coding regions aligned between the WGS sequence, the recently published sequence of chimpanzee chromosome 21 (ref. 13; formerly chromosome 22 (ref. 18)) and the human genome also revealed no excess of substitutions in the WGS sequence (see Supplementary Information 'Genome sequencing and assembly'). Thus, by restricting our analysis to high-quality bases, the nucleotide-level accuracy of the WGS assembly is essentially equal to that of 'finished' sequence.

Structural accuracy is also high based on comparison with finished BACs from the primary donor and other chimpanzees, although the relatively low level of sequence redundancy limits local contiguity. On the basis of comparisons with the primary donor, some small supercontigs (most <5 kb) have not been positioned within large supercontigs (~1 event per 100 kb); these are not strictly errors but nonetheless affect the utility of the assembly. There are also small, undetected overlaps (all <1 kb) between consecutive contigs (~1.2 events per 100 kb) and occasional local misordering of small contigs (~0.2 events per 100 kb). No misoriented contigs were found. Comparison with the finished chromosome 21 sequence yielded similar discrepancy rates (see Supplementary Information 'Genome sequencing and assembly').

The most problematic regions are those containing recent segmental duplications. Analysis of BAC clones from duplicated ($n = 75$) and unique ($n = 28$) regions showed that the former tend to be fragmented into more contigs (1.6-fold) and more supercontigs (3.2-fold). Discrepancies in contig order are also more frequent in duplicated than unique regions (~0.4 versus ~0.1 events per 100 kb). The rate is twofold higher in duplicated regions with the highest sequence identity (>98%). If we restrict the analysis to older duplications ($\leq 98\%$ identity) we find fewer assembly problems: 72% of those that can be mapped to the human genome are shared as duplications in both species. These results are consistent with the described limitations of current WGS assembly for regions of segmental duplication²⁶. Detailed analysis of these rapidly changing regions of the genome is being performed with more directed approaches²⁷.

Chimpanzee polymorphisms. The draft sequence of the chimpanzee genome also facilitates genome-wide studies of genetic diversity among chimpanzees, extending recent work^{28–31}. We sequenced and analysed sequence reads from the primary donor, four other West African and three central African chimpanzees (*Pan troglodytes troglodytes*) to discover polymorphic positions within and between these individuals (Supplementary Table S17).

A total of 1.66 million high-quality single-nucleotide polymorphisms (SNPs) were identified, of which 1.01 million are heterozygous within the primary donor, Clint. Heterozygosity rates were estimated to be 9.5×10^{-4} for Clint, 8.0×10^{-4} among West African chimpanzees and 17.6×10^{-4} among central African chimpanzees, with the variation between West and central African chimpanzees being 19.0×10^{-4} . The diversity in West African chimpanzees is similar to that seen for human populations³², whereas the level for central African chimpanzees is roughly twice as high.

The observed heterozygosity in Clint is broadly consistent with West African origin, although there are a small number of regions of distinctly higher heterozygosity. These may reflect a small amount of central African ancestry, but more likely reflect undetected regions of segmental duplications present only in chimpanzees.

Table 1 | Chimpanzee assembly statistics

Assembler	PCAP	ARACHNE
Major contigs*	400,289	361,782
Contig length (kb; N50)†	13.3	15.7
Supercontigs	67,734	37,846
Supercontig length (Mb; N50)	2.3	8.6
Sequence redundancy: all bases (Q20)	$5.0 \times (3.6 \times)$	$4.3 \times (3.6 \times)$
Physical redundancy	20.7	19.8
Consensus bases (Gb)	2.7	2.7

*Contigs >1 kb.

†N50 length is the size x such that 50% of the assembly is in units of length at least x .

Genome evolution

We set out to study the mutational events that have shaped the human and chimpanzee genomes since their last common ancestor. We explored changes at the level of single nucleotides, small insertions and deletions, interspersed repeats and chromosomal rearrangements. The analysis is nearly definitive for the smallest changes, but is more limited for larger changes, particularly lineage-specific segmental duplications, owing to the draft nature of the genome sequence.

Nucleotide divergence. Best reciprocal nucleotide-level alignments of the chimpanzee and human genomes cover ~2.4 gigabases (Gb) of high-quality sequence, including 89 Mb from chromosome X and 7.5 Mb from chromosome Y.

Genome-wide rates. We calculate the genome-wide nucleotide divergence between human and chimpanzee to be 1.23%, confirming recent results from more limited studies^{12,33,34}. The differences between one copy of the human genome and one copy of the chimpanzee genome include both the sites of fixed divergence between the species and some polymorphic sites within each species. By correcting for the estimated coalescence times in the human and chimpanzee populations (see Supplementary Information ‘Genome evolution’), we estimate that polymorphism accounts for 14–22% of the observed divergence rate and thus that the fixed divergence is ~1.06% or less.

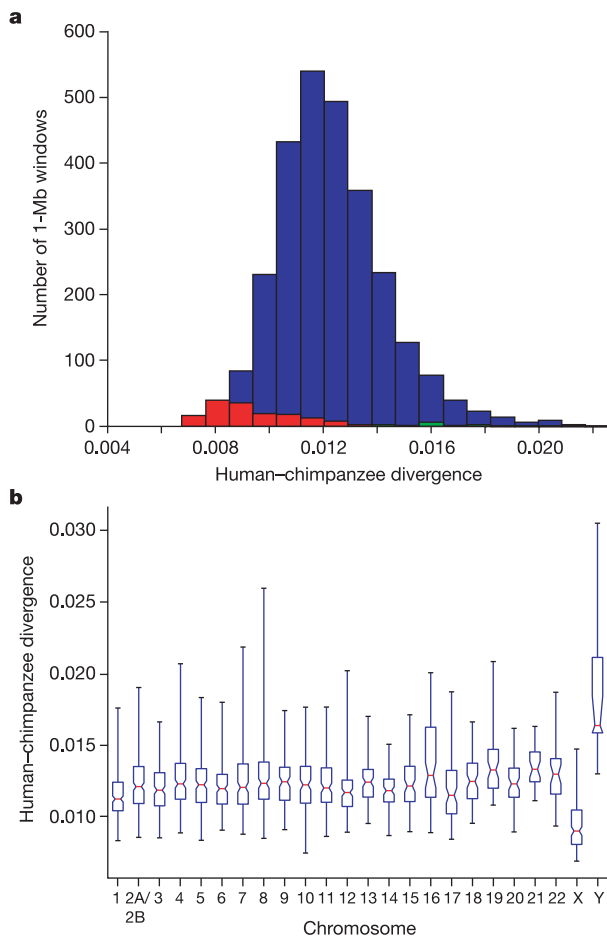


Figure 1 | Human-chimpanzee divergence in 1-Mb segments across the genome. **a**, Distribution of divergence of the autosomes (blue), the X chromosome (red) and the Y chromosome (green). **b**, Distribution of variation by chromosome, shown as a box plot. The edges of the box correspond to quartiles; the notches to the standard error of the median; and the vertical bars to the range. The X and Y chromosomes are clear outliers, but there is also high local variation within each of the autosomes.

Nucleotide divergence rates are not constant across the genome, as has been seen in comparisons of the human and murid genomes^{16,17,24,35,36}. The average divergence in 1-Mb segments fluctuates with a standard deviation of 0.25% (coefficient of variation = 0.20), which is much greater than the 0.02% expected assuming a uniform divergence rate (Fig. 1a; see also Supplementary Fig. S1).

Regional variation in divergence could reflect local variation in either mutation rate or other evolutionary forces. Among the latter, one important force is genetic drift, which can cause substantial differences in divergence time across loci when comparing closely related species, as the divergence time for orthologues is the sum of two terms: t_1 , the time since speciation, and t_2 , the coalescence time for orthologues within the common ancestral population³⁷. Whereas t_1 is constant across loci (~6–7 million years³⁸), t_2 is a random variable that fluctuates across loci (with a mean that depends on population size and here may be on the order of 1–2 million years³⁹). However, because of historical recombination, the characteristic scale of such fluctuations will be on the order of tens of kilobases, which is too small to account for the variation observed for 1-Mb regions⁴⁰ (see Supplementary Information ‘Genome evolution’). Other potential evolutionary forces are positive or negative selection. Although it is more difficult to quantify the expected contributions of selection in the ancestral population^{41–43}, it is clear that the effects would have to be very strong to explain the large-scale variation observed across mammalian genomes^{16,44}. There is tentative evidence from in-depth analysis of divergence and diversity that natural selection is not the major contributor to the large-scale patterns of genetic variability in humans^{45–47}. For these reasons, we suggest that the large-scale variation in the human–chimpanzee divergence rate primarily reflects regional variation in mutation rate.

Chromosomal variation in divergence rate. Variation in divergence rate is evident even at the level of whole chromosomes (Fig. 1b). The most striking outliers are the sex chromosomes, with a mean divergence of 1.9% for chromosome Y and 0.94% for chromosome X. The likely explanation is a higher mutation rate in the male compared with female germ line⁴⁸. Indeed, the ratio of the male/female mutation rates (denoted α) can be estimated by comparing the divergence rates among the sex chromosomes and the autosomes and correcting for ancestral polymorphism as a function of population size of the most recent common ancestor (MRCA; see Supplementary Information ‘Genome evolution’). Estimates for α range from 3 to 6, depending on the chromosomes compared and the assumed ancestral population size (Supplementary Table S18). This is significantly higher than recent estimates of α for the murids (~1.9) (ref. 17) and resolves a recent controversy based on smaller data sets^{12,24,49,50}.

The higher mutation rate in the male germ line is generally attributed to the 5–6-fold higher number of cell divisions undergone by male germ cells⁴⁸. We reasoned that this would affect mutations resulting from DNA replication errors (the rate should scale with the number of cell divisions) but not mutations resulting from DNA damage such as deamination of methyl CpG to TpG (the rate should scale with time). Accordingly, we calculated α separately for CpG sites, obtaining a value of ~2 from the comparison of rates between autosomes and chromosome X. This intermediate value is a composite of the rates of CpG loss and gain, and is consistent with roughly equal rates of CpG to TpG transitions in the male and female germ line^{51,52}.

Significant variation in divergence rates is also seen among autosomes (Fig. 1b; $P < 3 \times 10^{-15}$, Kruskal–Wallis test over 1-Mb windows), confirming earlier observations based on low-coverage WGS sampling¹². Additional factors thus influence the rate of divergence between chimpanzee and human chromosomes. These factors are likely to act at length scales significantly shorter than a chromosome, because the standard deviation across autosomes (0.21%) is comparable to the standard deviation seen in 1-Mb windows across the genome (0.13–0.35%). We therefore sought to

understand local factors that contribute to variation in divergence rate.

Contribution of CpG dinucleotides. Sites containing CpG dinucleotides in either species show a substantially elevated divergence rate of 15.2% per base; they account for 25.2% of all substitutions while constituting only 2.1% of all aligned bases. The divergence at CpG sites represents both the loss of ancestral CpGs and the creation of new CpGs. The former process is known to occur at a rapid rate per base due to frequent methylation of cytosines in a CpG context and their frequent deamination^{53,54}, whereas the latter process probably proceeds at a rate more typical of other nucleotide substitutions. Assuming that loss and creation of CpG sites are close to equilibrium, the mutation rate for bases in a CpG dinucleotide must be 10–12-fold higher than for other bases (see Supplementary Information ‘Genome evolution’ and ref. 51).

Because of the high rate of CpG substitutions, regional divergence rates would be expected to correlate with regional CpG density. CpG density indeed varies across 1-Mb windows (mean = 2.1%, coefficient of variation = 0.44 compared with 0.0093 expected under a Poisson distribution), but only explains 4% of the divergence rate variance. In fact, regional CpG and non-CpG divergence is highly correlated ($r = 0.88$; Supplementary Fig. S2), suggesting that higher-order effects modulate the rates of two very different mutation processes (see also ref. 47).

Increased divergence in distal regions. The most striking regional pattern is a consistent increase in divergence towards the ends of most chromosomes (Fig. 2). The terminal 10 Mb of chromosomes (including distal regions and proximal regions of acrocentric chromosomes) averages 15% higher divergence than the rest of the genome (Mann–Whitney U -test; $P < 10^{-30}$), with a sharp increase towards the telomeres. The phenomenon correlates better with physical distance than relative position along the chromosomes and may partially explain why smaller chromosomes tend to have higher divergence (Supplementary Fig. S3; see also ref. 15). These observations suggest that large-scale chromosomal structure, directly or indirectly, influences regional divergence patterns. The cause of this effect is unclear, but these regions (~15% of the genome) are

notable in having high local recombination rate, high gene density and high G + C content.

Correlation with chromosome banding. Another interesting pattern is that divergence increases with the intensity of Giemsa staining in cytogenetically defined chromosome bands, with the regions corresponding to Giemsa dark bands (G bands) showing 10% higher divergence than the genome-wide average (Mann–Whitney U -test; $P < 10^{-14}$) (see Fig. 2). In contrast to terminal regions, these regions (17% of the genome) tend to be gene poor, (G + C)-poor and low in recombination^{55,56}. The elevated divergence seen in two such different types of regions suggests that multiple mechanisms are at work, and that no single known factor, such as G + C content or recombination rate, is an adequate predictor of regional variation in the mammalian genome by itself (Fig. 3). Elucidation of the relative contributions of these and other mechanisms will be important for formulating accurate models for population genetics, natural selection, divergence times and the evolution of genome-wide sequence composition⁵⁷.

Correlation with regional variation in the murid genome. Given that sequence divergence shows regional variation in both hominids (human–chimpanzee) and murids (mouse–rat), we asked whether the regional rates are positively correlated between orthologous regions. Such a correlation would suggest that the divergence rate is driven, in part, by factors that have been conserved over the ~75 million years since rodents, humans and apes shared a common ancestor. Comparative analysis of the human and murid genomes has suggested such a correlation^{58–60}, but the chimpanzee sequence provides a direct opportunity to compare independent evolutionary processes between two mammalian clades.

We compared the local divergence rates in hominids and murids across major orthologous segments in the respective genomes (Fig. 4). For orthologous segments that are non-distal in both hominids and murids, there is a strong correlation between the divergence rates ($r = 0.5$, $P < 10^{-11}$). In contrast, orthologous segments that are centred within 10 Mb of a hominid telomere have disproportionately high divergence rates and G + C content relative to the murids (Mann–Whitney U -test; $P < 10^{-11}$ and

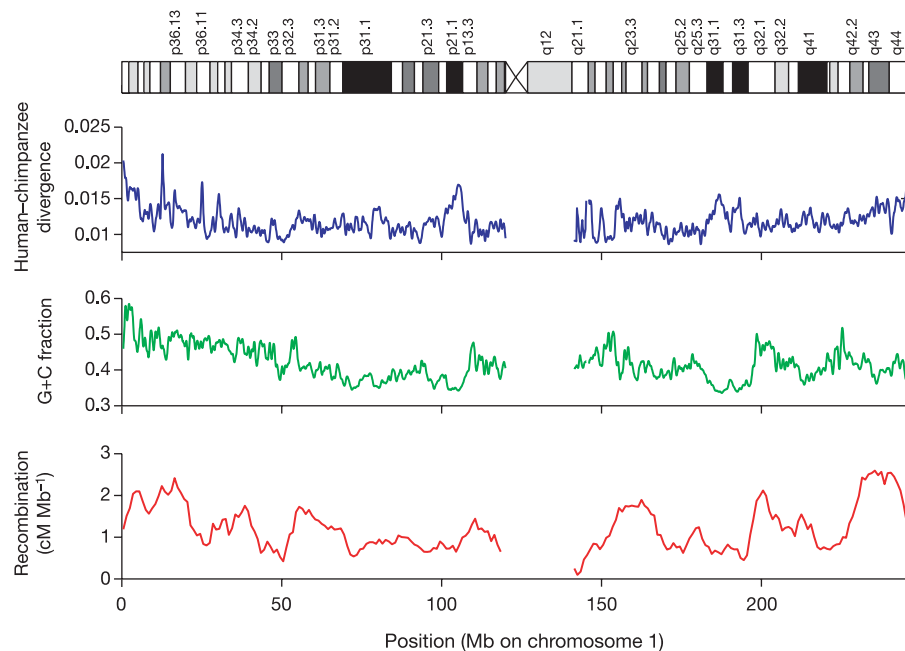


Figure 2 | Regional variation in divergence rates. Human–chimpanzee divergence (blue), G + C content (green) and human recombination rates¹⁷³ (red) in sliding 1-Mb windows for human and chimpanzee chromosome 1. Divergence and G + C content are noticeably elevated near the 1p telomere,

a trend that holds for most subtelomeric regions (see text). Internally on the chromosome, regions of low G + C content and high divergence often correspond to the dark G bands.

$P < 10^{-4}$), implying that the elevation in these regions is, at least partially, lineage specific. The same general effect is observed (albeit less pronounced) if CpG dinucleotides are excluded (Supplementary Fig. S4). Increased divergence and G + C content might be explained by 'biased gene conversion'⁶¹ due to the high hominid recombination rates in these distal regions. Segments that are distal in murids do not show elevated divergence rates, which is consistent with this model, because the recombination rates of distal regions are not as elevated in mouse and rat⁶².

Taken together, these observations suggest that sequence divergence rate is influenced by both conserved factors (stable across mammalian evolution) and lineage-specific factors (such as proximity to the telomere or recombination rate, which may change with chromosomal rearrangements).

Insertions and deletions. We next studied the indel events that have occurred in the human and chimpanzee lineages by aligning the genome sequences to identify length differences. We will refer below to all events as insertions relative to the other genome, although they may represent insertions or deletions relative to the genome of the common ancestor.

The observable insertions fall into two classes: (1) 'completely covered' insertions, occurring within continuous sequence in both species; and (2) 'incompletely covered' insertions, occurring within sequence containing one or more gaps in the chimpanzee, but revealed by a clear discrepancy between the species in sequence length. Different methods are needed for reliable identification of modest-sized insertions (1 base to 15 kb) and large insertions (>15 kb), with the latter only being reliably identifiable in the human genome (see Supplementary Information 'Genome evolution').

The analysis of modest-sized insertions reveals ~32 Mb of human-specific sequence and ~35 Mb of chimpanzee-specific sequence, contained in ~5 million events in each species (Supplementary Information 'Genome evolution' and Supplementary Fig. S5). Nearly all of the human insertions are completely covered, whereas only half of the chimpanzee insertions are completely covered. Analysis of the completely covered insertions shows that the vast majority are small (45% of events cover only 1 base pair (bp), 96% are <20 bp and 98.6% are <80 bp), but that the largest few contain most of the

sequence (with the ~70,000 indels larger than 80 bp comprising 73% of the affected base pairs) (Fig. 5). The latter indels >80 bp fall into three categories: (1) about one-quarter are newly inserted transposable elements; (2) more than one-third are due to microsatellite and satellite sequences; (3) and the remainder are assumed to be mostly deletions in the other genome.

The analysis of larger insertions (>15 kb) identified 163 human regions containing 8.3 Mb of human-specific sequence in total (Fig. 6). These cases include 34 regions that involve exons from known genes, which are discussed in a subsequent section. Although we have no direct measure of large insertions in the chimpanzee genome, it appears likely that the situation is similar.

On the basis of this analysis, we estimate that the human and chimpanzee genomes each contain 40–45 Mb of species-specific euchromatic sequence, and the indel differences between the genomes thus total ~90 Mb. This difference corresponds to ~3% of both genomes and dwarfs the 1.23% difference resulting from nucleotide substitutions; this confirms and extends several recent studies^{63–67}. Of course, the number of indel events is far fewer than the number of substitution events (~5 million compared with ~35 million, respectively).

Transposable element insertions. We next used the catalogue of lineage-specific transposable element copies to compare the activity of transposons in the human and chimpanzee lineages (Table 2).

Endogenous retroviruses. Endogenous retroviruses (ERVs) have become all but extinct in the human lineage, with only a single retrovirus (human endogenous retrovirus K (HERV-K)) still active²⁴. HERV-K was found to be active in both lineages, with at least 73 human-specific insertions (7 full length and 66 solo long terminal repeats (LTRs)) and at least 45 chimpanzee-specific insertions (1 full length and 44 solo LTRs). A few other ERV classes persisted in the human genome beyond the human–chimpanzee split, leaving ~9 human-specific insertions (all solo LTRs, including five HERV9 elements) before dying out.

Against this background, it was surprising to find that the chimpanzee genome has two active retroviral elements (PtERV1 and PtERV2) that are unlike any older elements in either genome;

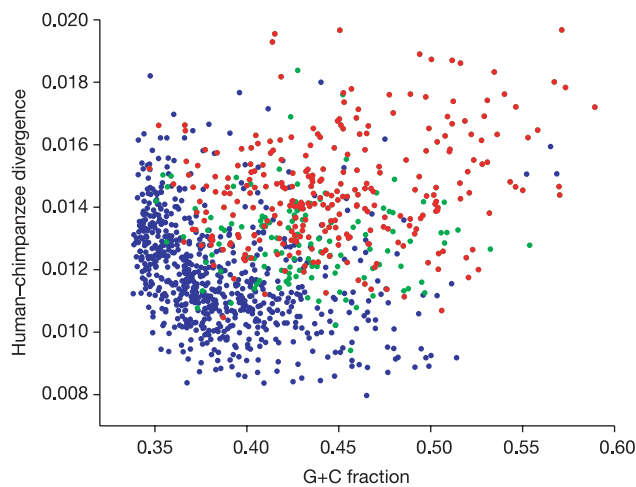


Figure 3 | Divergence rates versus G + C content for 1-Mb segments across the autosomes. Conditional on recombination rate, the relationship between divergence and G + C content varies. In regions with recombination rates less than 0.8 cM Mb^{-1} (blue), there is an inverse relationship, where high divergence regions tend to be (G + C)-poor and low divergence regions tend to be (G + C)-rich. In regions with recombination rates greater than 2.0 cM Mb^{-1} , whether within 10 Mb (red) or proximal (green) of chromosome ends, both divergence and G + C content are uniformly high.

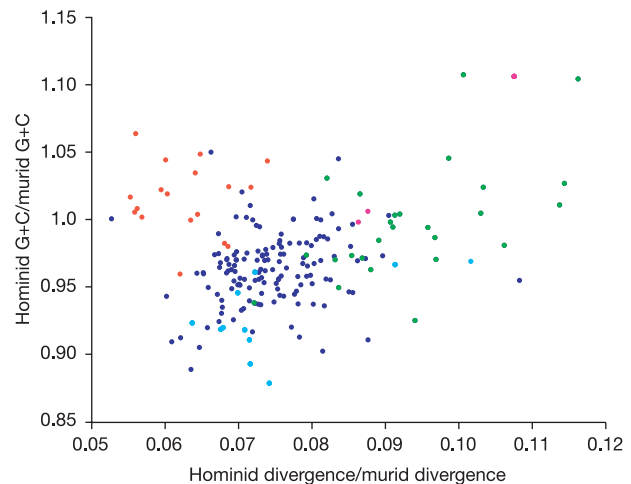


Figure 4 | Disproportionately elevated divergence and G + C content near hominid telomeres. Scatter plot of the ratio of human–chimpanzee divergence over mouse–rat divergence versus the ratio of human G + C content over mouse G + C content across 199 syntenic blocks for which more than 1 Mb of sequence could be aligned between all four species. Blocks for which the centre is within 10 Mb of a telomere in hominids only (green) or in hominids and murids (magenta), but not in murids only (light blue), show a significant trend towards higher ratios than internal blocks (dark blue). Blocks on the X chromosome (red) tend to show a lower divergence ratio than autosomal blocks, consistent with a smaller difference between autosomal and X divergence in murids than in hominids (lower α).

these must have been introduced by infection of the chimpanzee germ line. The smaller family (PtERV2) has only a few dozen copies, which nonetheless represent multiple (~5–8) invasions, because the sequence differences among reconstructed subfamilies are too great (~8%) to have arisen by mutation since divergence from human. It is closely related to a baboon endogenous retrovirus (BaEV, 88% ORF2 product identity) and a feline endogenous virus (ECE-1, 86% ORF2 product identity). The larger family (PtERV1) is more homogeneous and has over 200 copies. Whereas older ERVs, like HERV-K, are primarily represented by solo LTRs resulting from LTR–LTR recombination, more than half of the PtERV1 copies are still full length, probably reflecting the young age of the elements. PtERV1-like elements are present in the rhesus monkey, olive baboon and African great apes but not in human, orang-utan or gibbon, suggesting separate germline invasions in these species⁶⁸.

Higher Alu activity in humans. SINE (Alu) elements have been threefold more active in humans than chimpanzee (~7,000 compared with ~2,300 lineage-specific copies in the aligned portion), refining the rather broad range (2–7-fold) estimated in smaller studies^{13,67,69}. Most chimpanzee-specific elements belong to a subfamily (AluYc1) that is very similar to the source gene in the common ancestor. By contrast, most human-specific Alu elements belong to two new subfamilies (AluYa5 and AluYb8) that have evolved since the chimpanzee–human divergence and differ substantially from the ancestral source gene⁶⁹. It seems likely that the resurgence of Alu elements in humans is due to these potent new source genes. However, based on an examination of available finished sequence, the baboon shows a 1.6-fold higher Alu activity relative to human new insertions, suggesting that there may also have been a general decline in activity in the chimpanzee⁶⁷.

Some of the human-specific Alu elements are highly diverged (92 with >5% divergence), which would seem to suggest that they are much older than the human–chimpanzee split. Possible explanations include: gene conversion by nearby older elements; processed pseudogenes arising from a spurious transcription of an older element; precise excision from the chimpanzee genome; or high local mutation rate. In any case, the presence of such anomalies suggests that caution is warranted in the use of single-repeat elements as homoplasy-free phylogenetic markers.

New Alu elements target (A + T)-rich DNA in human and chimpanzee genomes. Older SINE elements are preferentially found in gene-rich,

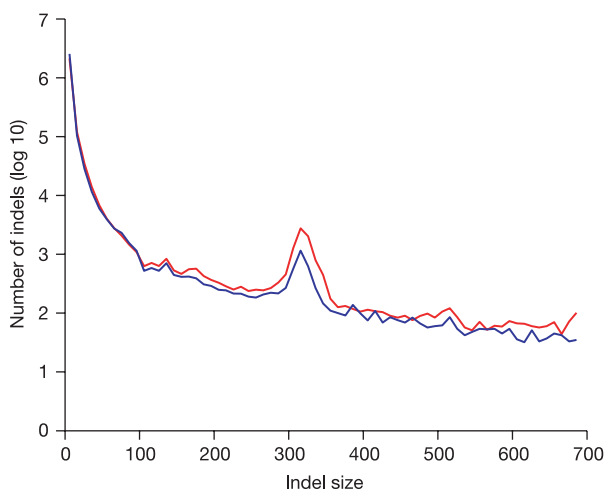


Figure 5 | Length distribution of small indel events, as determined using bounded sequence gaps. Sequences present in chimpanzee but not in human (blue) or present in human but not in chimpanzee (red) are shown. The prominent spike around 300 nucleotides corresponds to SINE insertion events. Most of the indels are smaller than 20 bp, but larger indels account for the bulk of lineage-specific sequence in the two genomes.

(G + C)-rich regions, whereas younger SINE elements are found in gene-poor, (A + T)-rich regions where long interspersed element (LINE)-1 (L1) copies also accumulate^{24,70}. The latter distribution is consistent with the fact that Alu retrotransposition is mediated by L1 (ref. 71). Murid genomes revealed no change in SINE distribution with age¹⁷.

The human pattern might reflect either preferential retention of SINEs in (G + C)-rich regions, due to selection or mutation bias, or a recent change in Alu insertion preferences. With the availability of the chimpanzee genome, it is possible to classify the youngest Alu copies more accurately and thus begin to distinguish these possibilities.

Analysis shows that lineage-specific SINEs in both human and chimpanzee are biased towards (A + T)-rich regions, as opposed to even the most recent copies in the MRCA (Fig. 7). This indicates that SINEs are indeed preferentially retained in (G + C)-rich DNA, but comparison with a more distant primate is required to formally rule out the possibility that the insertion bias of SINEs did not change just before speciation.

Equal activity of L1 in both species. The human and chimpanzee genomes both show ~2,000 lineage-specific L1 elements, contrary to previous estimates based on small samples that L1 activity is 2–3-fold higher in chimpanzee⁷².

Transcription from L1 source genes can sometimes continue into 3' flanking regions, which can then be co-transposed^{73,74}. Human–chimpanzee comparison revealed that ~15% of the species-specific insertions appear to have carried with them at least 50 bp of flanking sequence (followed by a poly(A) tail and a target site duplication). In principle, incomplete reverse transcription could result in insertions of the flanking sequence only (without any L1 sequence), mobilizing gene elements such as exons, but we found no evidence of this.

Retrotransposed gene copies. The L1 machinery also mediates retrotransposition of host messenger RNAs, resulting in many intronless (processed) pseudogenes in the human genome^{75–77}. We identified 163 lineage-specific retrotransposed gene copies in human and 246 in chimpanzee (Supplementary Table S19). Correcting for incomplete sequence coverage of the chimpanzee genome, we estimate that there are ~200 and ~300 processed gene copies in human and chimpanzee, respectively. Processed genes thus appear to have arisen at a rate of ~50 per million years since the divergence of human and chimpanzee; this is lower than the estimated rate for early primate evolution⁷⁵, perhaps reflecting the overall decrease in L1 activity. As expected⁷⁸, ribosomal protein genes constitute the largest class in both species. The second largest class in chimpanzee corresponds to zinc finger C2H2 genes, which are not a major class in the human genome.

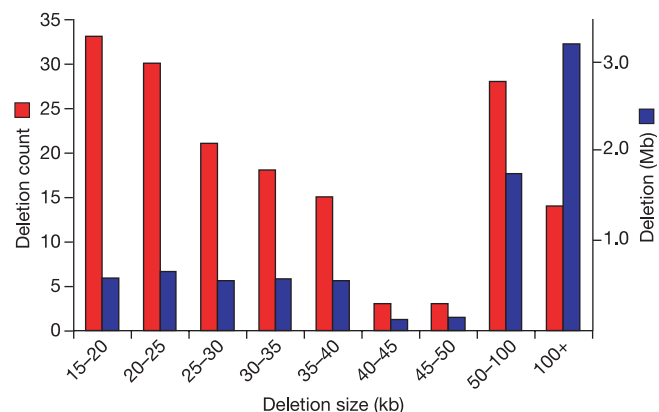


Figure 6 | Length distribution of large indel events (>15 kb), as determined using paired-end sequences from chimpanzee mapped against the human genome. Both the total number of candidate human insertions/chimpanzee deletions (blue) and the number of bases altered (red) are shown.

The retrotransposon SVA and distribution of CpG islands by transposable elements. The third most active element since speciation has been SVA, which created about 1,000 copies in each lineage. SVA is a composite element (~1.5–2.5 kb) consisting of two Alu fragments, a tandem repeat and a region apparently derived from the 3' end of a HERV-K transcript; it is probably mobilized by L1 (refs 79, 80). This element is of particular interest because each copy carries a sequence that satisfies the definition of a CpG island⁸¹ and contains potential transcription factor binding sites; the dispersion of 1,000 SVA copies could therefore be a source of regulatory differences between chimpanzee and human (Supplementary Table S20). At least three human genes contain SVA insertions near their promoters (Supplementary Table S21), one of which has been found to be differentially expressed between the two species^{82,83}, but additional investigations will be required to determine whether the SVA insertion directly caused this difference.

Homologous recombination between interspersed repeats. Human–chimpanzee comparison also makes it possible to study homologous recombination between nearby repeat elements as a source of genomic deletions. We found 612 deletions (totalling 2 Mb) in the human genome that appear to have resulted from recombination between two nearby Alu elements present in the common ancestor; there are 914 such events in the chimpanzee genome. (The events are not biased to (A + T)-rich DNA and thus would not explain the preferential loss of Alu elements in such regions discussed above.) Similarly, we found 26 and 48 instances involving adjacent L1 copies and 8 and 22 instances involving retroviral LTRs in human and chimpanzee, respectively. None of the repeat-mediated deletions removed an orthologous exon of a known human gene in chimpanzee.

The genome comparison allows one to estimate the dependency of homologous recombination on divergence and distance. Homologous recombination seems to occur between quite (>25%) diverged copies (Fig. 8), whereas the number of recombination events (n) varies inversely with the distance (d , in bases) between the copies (as $n \approx 6 \times 10^6 d^{-1.7}$; $r^2 = 0.9$).

Large-scale rearrangements. Finally, we examined the chimpanzee genome sequence for information about large-scale genomic alterations. Cytogenetic studies have shown that human and chimpanzee chromosomes differ by one chromosomal fusion, at least nine pericentric inversions, and in the content of constitutive heterochromatin⁸⁴. Human chromosome 2 resulted from a fusion of two ancestral chromosomes that remained separate in the chimpanzee lineage (chromosomes 2A and 2B in the revised nomenclature¹⁸, formerly chimpanzee chromosomes 12 and 13); the precise fusion point has been mapped and its duplication structure described in detail^{85,86}. In accord with this, alignment of the human and chimpanzee genome sequences shows a break in continuity at this point.

We searched the chimpanzee genome sequence for the precise locations of the 18 breakpoints corresponding to the 9 pericentric inversions (Supplementary Table S22). By mapping paired-end sequences from chimpanzee large insert clones to the human genome, we were able to identify 13 of the breakpoints within the

assembly from discordant end alignments. The positions of five breakpoints (on chromosomes 4, 5 and 12) were tested by fluorescence *in situ* hybridization (FISH) analysis and all were confirmed. Also, the positions of three previously mapped inversion breakpoints (on chromosomes 15 and 18) matched closely those found in the assembly^{87,88}. The paired-end analysis works well in regions of unique sequence, which constitute the bulk of the genome, but is less effective in regions of recent duplication owing to ambiguities in mapping of the paired-end sequences. Beyond the known inversions, we also found suggestive evidence of many additional smaller inversions, as well as older segmental duplications (<98% identity; Supplementary Fig. S6). However, both smaller inversions and more recent segmental duplications will require further investigations.

Gene evolution

We next sought to use the chimpanzee sequence to study the role of natural selection in the evolution of human protein-coding genes. Genome-wide comparisons can shed light on many central issues, including: the magnitude of positive and negative selection; the variation in selection across different lineages, chromosomes, gene families and individual genes; and the complete loss of genes within a lineage.

We began by identifying a set of 13,454 pairs of human and chimpanzee genes with unambiguous 1:1 orthology for which it was possible to generate high-quality sequence alignments covering virtually the entire coding region (Supplementary Information 'Gene evolution' and Table S23). The list contains a large fraction of the entire complement of human genes, although it under-represents gene families that have undergone recent local expansion (such as olfactory receptors and immunoglobulins). To facilitate comparison with the murid lineage, we also compiled a set of 7,043 human, chimpanzee, mouse and rat genes with unambiguous 1:1:1:1 orthology and high-quality sequence alignments (Supplementary Table S24).

Average rates of evolution. To assess the rate of evolution for each gene, we estimated K_A , the number of coding base substitutions that result in amino acid change as a fraction of all such possible sites (the non-synonymous substitution rate). Because the background

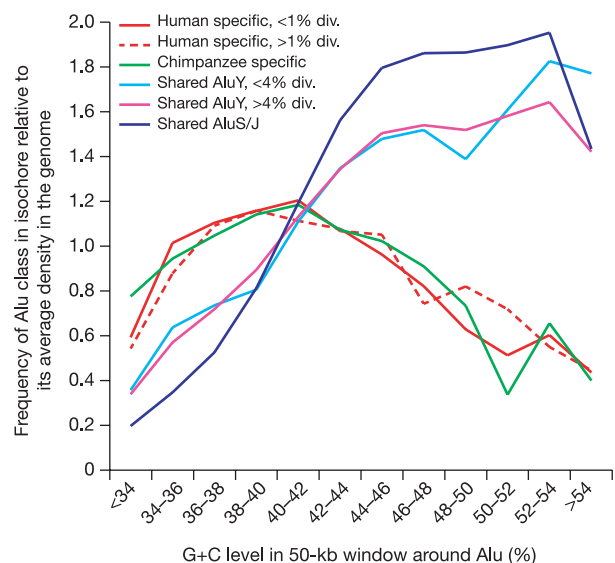


Figure 7 | Correlation of Alu age and distribution by G + C content. Alu elements that inserted after human–chimpanzee divergence are densest in the (G + C)-poor regions of the genome (peaking at 36–40% G + C), whereas older copies, common to both genomes, crowd (G + C)-rich regions. The figure is similar to figure 23 of ref. 24, but the use of chimpanzee allows improved separation of young and old elements, leading to a sharper transition in the pattern.

Table 2 | Transposable element activity in human and chimpanzee lineages

Element	Chimpanzee*	Human*
Alu	2,340 (0.7 Mb)	7,082 (2.1 Mb)
LINE-1	1,979 (>5 Mb)	1,814 (5.0 Mb)
SVA	757 (>1 Mb)	970 (1.3 Mb)
ERV class 1	234 (>1 Mb)†	5 (8 kb)‡
ERV class 2	45 (55 kb)§	77 (130 kb)§
(Micro)satellite	7,054 (4.1 Mb)	11,101 (5.1 Mb)

*Number of lineage-specific insertions (with total size of inserted sequences indicated in brackets) in the aligned parts of the genomes.

†P1ERV1 and P1ERV2.

‡HERV9.

§Mostly HERV-K.

mutation rate varies across the genome, it is crucial to normalize K_A for comparisons between genes. A striking illustration of this variation is the fact that the mean K_A is 37% higher in the rapidly diverging distal 10 Mb of chromosomes than in the more proximal regions. Classically, the background rate is estimated by K_S , the synonymous substitution rate (coding base substitutions that, because of codon redundancy, do not result in amino acid change). Because a typical gene has only a few synonymous changes between humans and chimpanzees, and not infrequently is zero, we exploited the genome sequence to estimate the local intergenic/intronic substitution rate, K_I , where appropriate. K_A and K_S were also estimated for each lineage separately using mouse and rat as outgroups (Fig. 9).

The K_A/K_S ratio is a classical measure of the overall evolutionary constraint on a gene, where $K_A/K_S \ll 1$ indicates that a substantial proportion of amino acid changes must have been eliminated by purifying selection. Under the assumption that synonymous substitutions are neutral, $K_A/K_S > 1$ implies, but is not a necessary condition for, adaptive or positive selection. The K_A/K_I ratio has the same interpretation. The ratios will sometimes be denoted below by ω with an appropriate subscript (for example, ω_{human}) to indicate the branch of the evolutionary tree under study.

Evolutionary constraint on amino acid sites within the hominid lineage. Overall, human and chimpanzee genes are extremely similar, with the encoded proteins identical in the two species in 29% of cases. The median number of non-synonymous and synonymous substitutions per gene are two and three, respectively. About 5% of the proteins show in-frame indels, but these tend to be small (median = 1 codon)

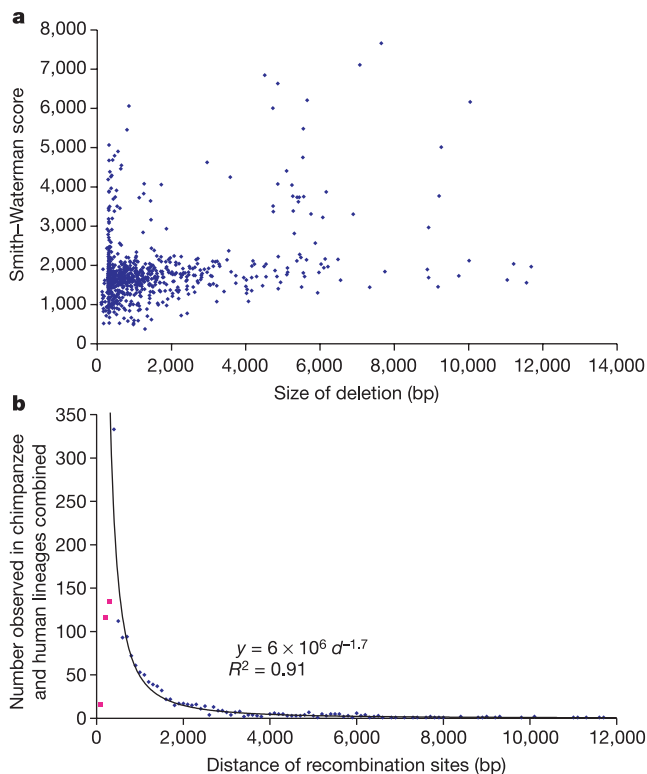


Figure 8 | Dependency of homologous recombination between Alu elements on divergence and distance. **a**, Whereas homologous recombination occurs between quite divergent (Smith–Waterman score <1,000), closely spaced copies, more distant recombination seems to favour a better match between the recombining repeats. **b**, The frequency of Alu–Alu-mediated recombination falls markedly as a function of distance between the recombining copies. The first three points (magenta) involve recombination between left or right arms of one Alu inserted into another. The high number of occurrences at a distance of 300–400 nucleotides is due to the preference of integration in the A-rich tail; exclusion of this point does not change the parameters of the equation.

and to occur in regions of repeated sequence. The close similarity of human and chimpanzee genes necessarily limits the ability to make strong inferences about individual genes, but there is abundant data to study important sets of genes.

The K_A/K_S ratio for the human–chimpanzee lineage (ω_{hominid}) is 0.23. The value is much lower than some recent estimates based on limited sequence data (ranging as high as 0.63 (ref. 7)), but is consistent with an estimate (0.22) from random expressed-sequence-tag (EST) sequencing⁴⁵. Similarly, K_A/K_I was also estimated as 0.23.

Under the assumption that synonymous mutations are selectively neutral, the results imply that 77% of amino acid alterations in hominid genes are sufficiently deleterious as to be eliminated by natural selection. Because synonymous mutations are not entirely neutral (see below), the actual proportion of amino acid alterations with deleterious consequences may be higher. Consistent with previous studies⁸, we find that K_A/K_S of human polymorphisms with frequencies up to 15% is significantly higher than that of human–chimpanzee differences and more common polymorphisms (Table 3), implying that at least 25% of the deleterious amino acid alterations may often attain readily detectable frequencies and thus contribute significantly to the human genetic load.

Evolutionary constraint on synonymous sites within hominid lineage. We next explored the evolutionary constraints on synonymous sites, specifically fourfold degenerate sites. Because such sites have no effect on the encoded protein, they are often considered to be selectively neutral in mammals.

We re-examined this assumption by comparing the divergence at fourfold degenerate sites with the divergence at nearby intronic sites. Although overall divergence rates are very similar at fourfold degenerate and intronic sites, direct comparison is misleading because the former have a higher frequency of the highly mutable CpG dinucleotides (9% compared with 2%). When CpG and non-CpG sites are considered separately, we find that both CpG sites and non-CpG sites show markedly lower divergence in exonic synonymous sites than in introns (~50% and ~30% lower, respectively). This result resolves recent conflicting reports based on limited data sets^{45,89} by showing that such sites are indeed under constraint.

The constraint does not seem to result from selection on the usage of preferred codons, which has been detected in lower organisms⁹⁰ such as bacteria⁹¹, yeast⁹² and flies⁹³. In fact, divergence at fourfold

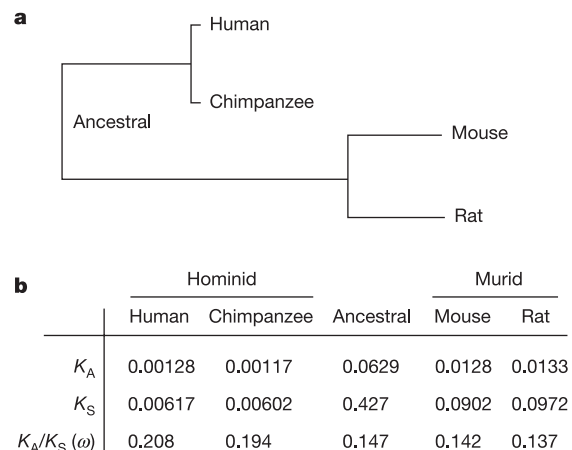


Figure 9 | Human–chimpanzee–mouse–rat tree with branch-specific K_A/K_S (ω) values. **a**, Evolutionary tree. The branch lengths are proportional to the absolute rates of amino acid divergence. **b**, Maximum-likelihood estimates of the rates of evolution in protein-coding genes for humans, chimpanzees, mice and rats. In the text, ω_{hominid} is the K_A/K_S of the combined human and chimpanzee branches and ω_{murid} of the combined mouse and rat branches. The slight difference between ω_{human} and $\omega_{\text{chimpanzee}}$ is not statistically significant; masking of some heterozygous bases in the chimpanzee sequence may contribute to the observed difference (see Supplementary Information ‘Gene evolution’).

degenerate sites increases slightly with codon usage bias (Kendall's $\tau = 0.097$, $P < 10^{-14}$). Alternatively, the observed constraint at synonymous sites might reflect 'background selection'—that is, the indirect effect of purifying selection at amino acid sites causing reduced diversity and thereby reduced divergence at closely linked sites⁴². Given the low rate of recombination in hominid genomes (a 1 kb region experiences only ~ 1 crossover per 100,000 generations or 2 million years), such background selection should extend beyond exons to include nearby intronic sites⁹⁴. However, when the divergence rate is plotted relative to exon–intron boundaries, we find that the rate jumps sharply within a short region of ~ 7 bp at the boundary (Fig. 10). This pattern strongly suggests that the action of purifying selection at synonymous sites is direct rather than indirect, suggesting that other signals, for example those involved in splice site selection, may be embedded in the coding sequence and therefore constrain synonymous sites.

Comparison with murids. An accurate estimate of K_A/K_S makes it possible to study how evolutionary constraint varies across clades. It was predicted more than 30 years ago⁹⁵ that selection against deleterious mutations would depend on population size, with mutations being strongly selected only if they reduce fitness by $s \gg 1/4N$ (where N is effective population size). This would predict that genes would be under stronger purifying selection in murids than hominids, owing to their presumed larger population size. Initial analyses (involving fewer than 50 genes⁹⁶) suggested a strong effect, but the wide variation in estimates of K_A/K_S in hominids^{7,8,97} and murids⁹⁸ has complicated this analysis⁴⁵.

Using the large collection of 7,043 orthologous quartets, we calculated mean K_A/K_S values for the various branches of the four-species evolutionary tree (human, chimpanzee, mouse and rat; Fig. 9). The K_A/K_S ratio for hominids is 0.20. (This is slightly lower than the value of 0.23 obtained with all human–chimpanzee orthologues, probably reflecting slightly greater constraint on the class of proteins with clear orthologues across hominids and murids.)

The K_A/K_S ratio is markedly lower for murids than for hominids ($\omega_{\text{murid}} \approx 0.13$ compared with $\omega_{\text{hominid}} \approx 0.20$) (Fig. 9). This implies that there is an $\sim 35\%$ excess of the amino-acid-changing mutations in the two hominids, relative to the two murids. Excess amino acid divergence may be explained by either increased adaptive evolution or relaxation of evolutionary constraints. As shown in the next section, the latter seems to be the principal explanation.

Relaxed constraints in human evolution. The K_A/K_S ratio can be used to make inferences about the role of positive selection in human evolution^{99,100}. Because alleles under positive selection spread rapidly through a population, they will be found less frequently as common human polymorphisms than as human–chimpanzee differences⁸. Positive selection can thus be detected by comparing the K_A/K_S ratio for common human polymorphisms with the K_A/K_S ratio for

hominid divergence. These ratios have been estimated as $\omega_{\text{polymorphism}} \approx 0.20$ based on an initial collection of common SNPs in human genes and $\omega_{\text{divergence}} \approx 0.34$ based on comparison of human and Old World monkey genes⁸. Thus, the proportion of amino acid changes attributable to positive selection was inferred to be $\sim 35\%$ (ref. 8). This would imply a huge quantitative role for positive selection in human evolution.

With the availability of extensive data for both human polymorphism and human–chimpanzee divergence, we repeated this analysis (using the same set of genes for both estimates). We find that $\omega_{\text{polymorphism}} \approx 0.21\text{--}0.23$ and $\omega_{\text{divergence}} \approx 0.23$ are statistically indistinguishable (Table 3). Although some of the amino acid substitutions in human and chimpanzee evolution must surely reflect positive selection, the results indicate that the proportion of changes fixed by positive selection seems to be much lower than the previous estimate⁸. (Because the previous results involved comparison to Old World monkeys, it is possible that they reflect strong positive selection earlier in primate evolution; however, we suspect that they reflect the fact that relatively few genes were studied and that different genes were used to study polymorphism and divergence.)

Relaxed negative selection pressures thus primarily explain the excess amino acid divergence in hominid genes relative to murids. Moreover, because both ω_{human} and $\omega_{\text{chimpanzee}}$ are similarly elevated this explanation applies equally to both lineages.

We next sought to study variation in the evolutionary rate of genes within the hominid lineage by searching for unusually high or low levels of constraint for genes and sets of genes.

Rapid evolution in individual genes. We searched for individual genes that have accumulated amino acid substitutions faster than expected given the neutral substitution rate; we considered these genes as potentially being under strong positive selection. A total of 585 of the 13,454 human–chimpanzee orthologues (4.4%) have observed $K_A/K_S > 1$ (see Supplementary Information 'Gene evolution'). However, given the low divergence, the K_A/K_S statistic has large variance. Simulations show that estimates of $K_A/K_S > 1$ would be expected to occur simply by chance in at least 263 cases if purifying selection is allowed to act non-uniformly across genes (Supplementary Fig. S7).

Nonetheless, this set of 585 genes may be enriched for genes that are under positive selection. The most extreme outliers include glycoprotein C, which mediates one of the *Plasmodium falciparum* invasion pathways in human erythrocytes¹⁰¹; granulysin, which mediates antimicrobial activity against intracellular pathogens such as *Mycobacterium tuberculosis*¹⁰²; as well as genes that have previously been shown to be undergoing adaptive evolution, such as the protamines and semenogelins involved in reproduction¹⁰³ and the Mas-related gene family involved in nociception¹⁰⁴. With similar

Table 3 | Comparison of K_A/K_S for divergence and human diversity

Substitution type	ΔA	ΔS	K_A/K_S	Per cent excess*	Confidence interval†
Human–chimpanzee divergence	38,773	61,737	0.23	–	–
HapMap (European ancestry)‡					
Rare derived alleles (<15%)	1,614	1,540	0.39	67	[59, 75]
Common alleles	1,199	1,907	0.23	0	[–5, 6]
Frequent derived alleles (>85%)	209	356	0.22	–7	[–19, 7]
HapMap (African ancestry)‡					
Rare derived alleles (<5%)	849	842	0.36	61	[50, 72]
Common alleles	495	803	0.22	–2	[–10, 7]
Frequent derived alleles (>85%)	59	82	0.26	15	[–11, 48]
Affymetrix 120K (multi-ethnic)§					
Rare derived alleles (<15%)	74	82	0.33	44	[14, 80]
Common alleles	77	137	0.21	–11	[–28, 12]
Frequent derived alleles (>85%)	10	15	0.25	6	[–42, 95]

ΔA , Number of observed non-synonymous substitutions. ΔS , Number of observed synonymous substitutions.

* A negative value indicates excess of non-synonymous divergence over polymorphism.

† 95% confidence intervals assuming non-synonymous substitutions are Poisson distributed.

‡ Source: <http://www.hapmap.org> (Public Release no. 13).

§ Source: <http://www.affymetrix.com>.

follow-up studies on candidates from this list, one may be able to draw conclusions about positive selection on other individual genes. In subsequent sections, we examine the rate of divergence for sets of related genes with the aim of detecting subtler signals of accelerated evolution.

Variation in evolutionary rate across physically linked genes. We explored how the rate of evolution varies regionally across the genome. Several studies of mammalian gene evolution have noted that the rate of amino acid substitution shows local clustering, with proteins encoded by nearby genes evolving at correlated rates^{16,105–107}. *Variation across chromosomes.* On the basis of an analysis of ~100 genes¹⁰⁸, it was recently reported that the normalized rate of protein evolution is greater on the nine chromosomes that underwent major structural rearrangement during human evolution (chromosomes 1, 2, 5, 9, 12, 15, 16, 17 and 18); it was suggested that such rearrangements led to reduced gene flow and accelerated adaptive evolution. A subsequent study of a collection of chimpanzee ESTs gave contradictory results^{109,110}. With our larger data set, we re-examined this issue and found no evidence of accelerated evolution on chromosomes with major rearrangements, even if we considered each rearrangement separately (Supplementary Table S25).

Among all hominid chromosomes, the most extreme outlier is chromosome X with a mean K_A/K_I of 0.32. The higher mean seems to reflect a skewed distribution at both high and low values, with the median value (0.17) being more in line with other chromosomes (0.15). The excess of low values may reflect greater purifying selection at some genes, owing to the hemizyosity of chromosome X in males. The excess of high values may reflect increased adaptive selection also resulting from hemizyosity, if a considerable proportion of advantageous alleles are recessive¹¹¹. Interestingly, the higher K_A/K_I value on the X chromosome versus autosomes is largely restricted to genes expressed in testis⁸³.

Variation in local gene clusters. We next searched for genomic neighbourhoods with an unusually high density of rapidly evolving genes. Specifically, we calculated the median K_A/K_I for sliding windows of ten orthologues and identified extreme outliers ($P < 0.001$ compared to random ordering of genes; see Supplementary Information 'Gene evolution'). A total of 16 such neighbourhoods were found, which greatly exceeds random expectation (Table 4). Repeating the analysis with larger windows (25, 50 and 100 orthologues) did not identify additional rapidly diverging regions.

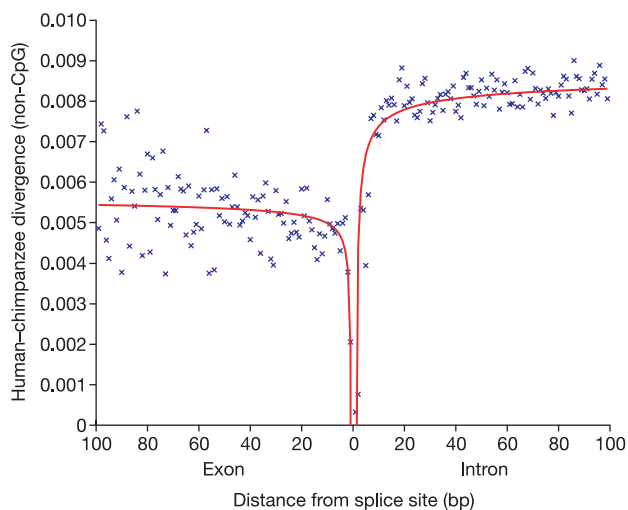


Figure 10 | Purifying selection on synonymous sites. Mean divergence around exon boundaries at non-CpG, exonic, fourfold degenerate sites and intronic sites, relative to the closest mRNA splice junction. The divergence rate at exonic, fourfold degenerate sites is significantly lower than at nearby intronic sites (Mann–Whitney U -test; $P < 10^{-27}$), suggesting that purifying selection limits the rate of synonymous codon substitutions.

In nearly all cases, the regions contain local clusters of phylogenetically and functionally related genes. The rapid diversification of gene families, postulated by ref. 112, can thus be readily discerned even at the relatively close distance of human–chimpanzee divergence. Most of the clusters are associated with functional categories such as host defence and chemosensation (see below). Examples include the epidermal differentiation complex encoding proteins that help form the cornified layer of the skin barrier (Supplementary Fig. S8), the WAP-domain cluster encoding secreted protease inhibitors with antibacterial activity, and the Siglec cluster encoding *CD33*-related genes. Rapid evolution in these clusters does not seem to be unique to either human or chimpanzee^{113,114}.

Variation in evolutionary rate across functionally related genes. We next studied variation in the evolutionary rate of functional categories of genes, based on the Gene Ontology (GO) classification¹¹⁵.

Rapidly and slowly evolving categories within the hominid lineage. We started by searching for sets of functionally related genes with exceptionally high or low constraint in humans and chimpanzees. For each of the 809 categories with at least 20 genes, K_A/K_S was calculated by concatenating the gene sequences. The category-specific ratios were compared to the average across all orthologues to identify extreme outliers using a metric based on the binomial test (Supplementary Information 'Gene evolution' and Supplementary Tables S26–S29). The numbers of observed outliers below a specific threshold (test statistic < 0.001) were then compared to the expected distribution of outliers given randomly permuted annotations.

A total of 98 categories showed elevated K_A/K_S ratios at the specified threshold (Table 5). Only 30 would be expected by chance, indicating that most (but not all) of these categories undergo significantly accelerated evolution relative to the genome-wide average ($P < 10^{-4}$). The rapidly evolving categories within the hominid lineage are primarily related to immunity and host defence, reproduction, and olfaction, which are the same categories known to be undergoing rapid evolution within the broader mammalian lineage, as well as more distantly related species^{15,16,116}. Hominids thus seem to be typical of mammals in this respect (but see below).

A total of 251 categories showed significantly low K_A/K_S ratios (compared with ~32 expected by chance; $P < 10^{-4}$). These include a wide range of processes including intracellular signalling, metabolism, neurogenesis and synaptic transmission, which are evidently under stronger-than-average purifying selection. More generally, genes expressed in the brain show significantly stronger average constraint than genes expressed in other tissues⁸³.

Differences between hominid and murid lineages. Having found gene categories that show substantial variation in absolute evolutionary rate within hominids, we next examined variation in relative rates

Table 4 | Rapidly diverging gene clusters in human and chimpanzee

Location (human)	Cluster	Median K_A/K_I *
1q21	Epidermal differentiation complex	1.46
6p22	Olfactory receptors and HLA-A	0.96
20p11	Cystatins	0.94
19q13	Pregnancy-specific glycoproteins	0.94
17q21	Hair keratins and keratin-associated proteins	0.93
19q13	CD33-related Siglecs	0.90
20q13	WAP domain protease inhibitors	0.90
22q11	Immunoglobulin- λ /breakpoint critical region	0.85
12p13	Taste receptors, type 2	0.81
17q12	Chemokine (C-C motif) ligands	0.81
19q13	Leukocyte-associated immunoglobulin-like receptors	0.80
5q31	Protocadherin- β	0.77
1q32	Complement component 4-binding proteins	0.76
21q22	Keratin-associated proteins and uncharacterized ORFs	0.76
1q23	CD1 antigens	0.72
4q13	Chemokine (C-X-C motif) ligands	0.70

*Maximum median K_A/K_I if the cluster stretched over more than one window of ten genes.

Table 5 | GO categories with the highest divergence rates in hominids

GO categories within 'biological process'	Number of orthologues	Amino acid divergence	K_A/K_S
GO:0007606 sensory perception of chemical stimulus	59	0.018	0.590
GO:0007608 perception of smell	41	0.018	0.521
GO:0006805 xenobiotic metabolism	40	0.013	0.432
GO:0006956 complement activation	22	0.013	0.428
GO:0042035 regulation of cytokine biosynthesis	20	0.011	0.402
GO:0007565 pregnancy	34	0.014	0.384
GO:0007338 fertilization	24	0.010	0.371
GO:0008632 apoptotic programme	36	0.010	0.358
GO:0007283 spermatogenesis	80	0.008	0.354
GO:0000075 cell cycle checkpoint	27	0.006	0.354

Listed are the ten categories in the taxonomy biological process with the highest K_A/K_S ratios, which are not significant solely due to significant subcategories.

between murids and hominids. The K_A/K_S of each of the GO categories are highly correlated between the hominid and murid orthologue pairs, suggesting that the selective pressures acting on particular functional categories have been largely proportional in recent hominid and recent murid evolution (Fig. 11). However, there are several categories with significantly accelerated non-synonymous divergence on each of the lineages, which might represent functions that have undergone lineage-specific positive selection or a lineage-specific relaxation of constraint (Supplementary Information 'Gene evolution' and Supplementary Tables S30–S39).

A total of 59 categories (compared with 11 expected at random, $P < 0.0003$) show evidence of accelerated non-synonymous divergence in the murid lineage. These are dominated by functions and processes related to host defence, such as immune response and lymphocyte activation. Examples include genes encoding interleukins and various T-cell surface antigens (*Cd4*, *Cd8*, *Cd80*). Combined with the recent observation that genes involved in host defence have undergone gene family expansion in murids^{16,17}, this suggests that the immune system has undergone extensive lineage-specific innovation in murids. Additional categories that also show relative acceleration in murids include chromatin-associated proteins and proteins involved in DNA repair. These categories may have similarly undergone stronger adaptive evolution in murids or, alternatively, they may contain fewer sites for mutations with slightly deleterious effects (with the result that the K_A/K_S ratios are less affected by the differences in population size^{96,117}).

Another 58 categories (versus 14 expected at random, $P < 0.0005$) show evidence of accelerated evolution in hominids, with the set dominated by genes encoding proteins involved in transport (for example, ion transport), synaptic transmission, spermatogenesis and perception of sound (Table 6). Notably, some outliers include genes with brain-related functions, compatible with a recent finding¹¹⁸. Potential positive selection on spermatogenesis genes in the hominids was also recently noted¹¹⁹. However, as above, it is possible that these categories could have more sites for slightly deleterious mutations and thus be more affected by population size differences. Sequence information from more species and from individuals

within species will be necessary to distinguish between the possible explanations.

Differences between the human and chimpanzee lineage. One of the most interesting questions is perhaps whether certain categories have undergone accelerated evolution in humans relative to chimpanzees, because such genes might underlie unique aspects of human evolution.

As was done for hominids and murids above, we compared non-synonymous divergence for each category to search for relative acceleration in either lineage (Fig. 12). Seven categories show signs of accelerated evolution on the human lineage relative to chimpanzee, but this is only slightly more than the four expected at random ($P < 0.22$). Intriguingly, the single strongest outlier is 'transcription factor activity', with the 348 human genes studied having accumulated 47% more amino acid changes than their chimpanzee orthologues. Genes with accelerated divergence in human include homeotic, forkhead and other transcription factors that have key roles in early development. However, given the small number of changes involved, additional data will be required to confirm this trend. There was no excess of accelerated categories on the chimpanzee lineage.

We also compared human genes with and without disease associations, including mental retardation, for differences in mutation rate when compared to chimpanzee. Briefly, no significant differences were observed in either the background mutation rate or in the ratio of human-specific changes to chimpanzee-specific amino acid changes (see Supplementary Information 'Gene evolution' and Supplementary Tables S40 and S41).

We thus find minimal evidence of acceleration unique to either the human or chimpanzee lineage across broad functional categories. This is not simply due to general lack of power resulting from the small number of changes since the divergence of human and chimpanzee, because one can detect acceleration of categories in either hominid relative to either murid. For example, 29 accelerated categories versus 9 expected at random ($P < 0.02$) can be detected on the human lineage, and 40 categories versus 11 expected at random ($P < 0.007$) on the chimpanzee lineage, relative to mouse. But the

Table 6 | GO categories with accelerated divergence rates in hominids relative to murids

GO categories within 'biological process'	Number of orthologues	Amino acid divergence in hominids	Amino acid divergence in murids	K_A/K_S in hominids	K_A/K_S in murids
GO:0007283 spermatogenesis	43	0.0075	0.054	0.323	0.188
GO:0006869 lipid transport	22	0.0081	0.051	0.306	0.120
GO:0006865 amino acid transport	24	0.0058	0.033	0.218	0.084
GO:0015698 inorganic anion transport	29	0.0061	0.027	0.195	0.072
GO:0006486 protein amino acid glycosylation	50	0.0056	0.040	0.166	0.100
GO:0019932 second-messenger-mediated signalling	58	0.0049	0.036	0.159	0.083
GO:0007605 perception of sound	28	0.0052	0.033	0.158	0.085
GO:0016051 carbohydrate biosynthesis	27	0.0047	0.028	0.147	0.067
GO:0007268 synaptic transmission	93	0.0040	0.025	0.126	0.069
GO:0006813 potassium ion transport	65	0.0035	0.022	0.113	0.056

Listed are the ten categories in the taxonomy biological process with the strongest evidence for accelerated evolution in hominids relative to murids, which are not significant solely due to significant subcategories.

outliers are largely the same for both human and chimpanzee, indicating that the fraction of amino acid mutations that have contributed to human- and chimpanzee-specific patterns of evolution must be small relative to the fraction that have contributed to a common hominid and, to a large extent, mammalian pattern of evolution.

It was recently reported¹⁰ that several functional categories are enriched for genes with evidence of positive selection in the human lineage or the chimpanzee lineage, and that these categories are largely different between the two lineages. These results and ours differ in ways that will require further investigation. With the potential exception of some developmental regulators, the categories that ref. 10 reported as showing the strongest enrichment of positive selection in one lineage (including cell adhesion, ion transport and perception of sound) are among those that we show as having accelerated divergence in both human and chimpanzee. This suggests that positive selection and relaxation of constraints may be correlated, or alternatively, that the results of ref. 10 may be enriched for false positives in categories that have experienced particularly strong relaxation of constraints in the hominids. Data from additional primates, as well as advances in analytical methods, will be necessary to distinguish between these alternatives. At present, strong evidence of positive selection unique to the human lineage is thus limited to a handful of genes¹²⁰.

Our analysis above largely omitted genes belonging to large gene families, because gene family expansion makes it difficult to define 1:1:1 orthologues across hominids and murids. One of the largest such families, the olfactory receptors, is known to be undergoing rapid divergence in primates. Directed study of these genes in the draft assembly has suggested that more than 100 functional human olfactory receptors are likely to be under no evolutionary constraint¹²¹. Our analysis also omitted the majority of very recently duplicated genes owing to their lower coverage in the current chimpanzee assembly. However, recent human-specific duplications can be readily identified from the finished human genome sequence, and have previously been shown to be highly enriched for the same categories found to have high absolute rates of evolution in 1:1 orthologues here; that is, olfaction, immunity and reproduction²³.

Gene disruptions in human and chimpanzee. Whereas most genes have undergone only subtle substitutions in their amino acid sequence, a few dozen have suffered more marked changes. We found a total of 53 known or predicted human genes that are either deleted entirely (36) or partially (17) in chimpanzee (Supplementary

Table S42). We have so far tested and confirmed 15 of these cases by polymerase chain reaction (PCR) or Southern blotting. An additional eight genes have sustained large deletions (>15 kb) entirely within an intron. Some genes may have been missed in this count owing to limitations of the draft genome sequence. In addition, some genes may have suffered chain termination mutations or altered reading frames in chimpanzee, but accurate identification of these will require higher-quality sequence. The sensitivity of the reciprocal analysis of genes disrupted in human is currently limited by the small number of independently predicted gene models for the chimpanzee. Some of the gene disruptions may be related to interesting biological differences between the species, as discussed below. **Genetic basis for human- and chimpanzee-specific biology.** Given the substantial number of neutral mutations, only a small subset of the observed gene differences is likely to be responsible for the key phenotypic changes in morphology, physiology and behavioural complexity between humans and chimpanzees. Determining which differences are in this evolutionarily important subset and inferring their functional consequences will require additional types of evidence, including information from clinical observations and model systems¹²². We describe some novel examples of genetic changes for which plausible functional or physiological consequences can be suggested.

Apoptosis. Mouse and human are known to differ with respect to an important mediator of apoptosis, caspase-12 (refs 123–125). The protein triggers apoptosis in response to perturbed calcium homeostasis in mice, but humans seem to lack this activity owing to several mutations in the orthologous gene that together affect the protein produced by all known splice forms; the mutations include a premature stop codon and a disruption of the SHG box required for enzymatic activity of caspases. By contrast, the chimpanzee gene encodes an intact open reading frame and SHG box, indicating that the functional loss occurred in the human lineage. Intriguingly, loss-of-function mutations in mice confer increased resistance to amyloid-induced neuronal apoptosis without causing obvious developmental or behavioural defects¹²⁶. The loss of function in humans may contribute to the human-specific pathology of Alzheimer's disease, which involves amyloid-induced neurotoxicity and deranged calcium homeostasis.

Inflammatory response. Human and chimpanzee show a notable difference with respect to important mediators of immune and inflammatory responses. Three genes (*IL1F7*, *IL1F8* and *ICEBERG*)

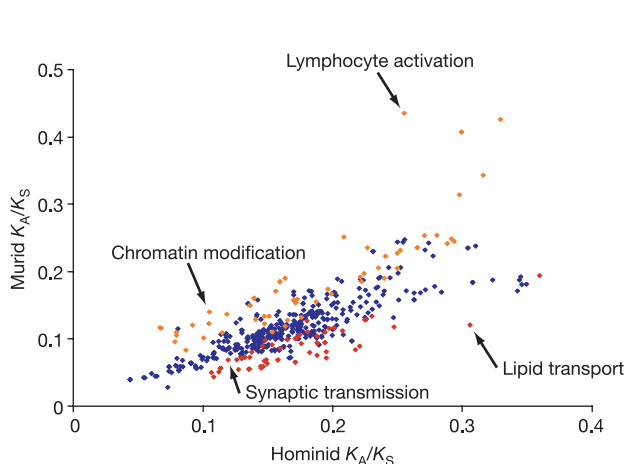


Figure 11 | Hominid and murid K_A/K_S (ω) in GO categories with more than 20 analysed genes. GO categories with putatively accelerated (test statistic <0.001; see Methods) non-synonymous divergence on the hominid lineages (red) and on the murid lineages (orange) are highlighted. Owing to the hierarchical nature of GO, the categories do not all represent independent data points. A non-redundant list of significant categories is provided in Table 8 and a complete list in Supplementary Table S30.

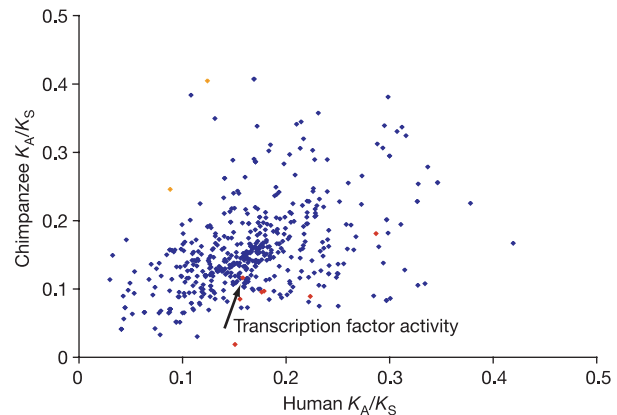


Figure 12 | Human and chimpanzee K_A/K_S (ω) in GO categories with more than 20 analysed genes. GO categories with putatively accelerated (test statistic <0.001; see Methods) non-synonymous divergence on the human lineage (red) and on the chimpanzee lineage (orange) are highlighted. The variance of these estimates is larger than that seen in the hominid–murid comparison owing to the small number of lineage-specific substitutions. Owing to the hierarchical nature of the GO ontology, the categories do not all represent independent data points. A complete list of categories is provided in Supplementary Table S30.

that act in a common pathway involving the caspase-1 gene all appear to be deleted in chimpanzee. *ICEBERG* is thought to repress caspase-1-mediated generation of pro-inflammatory *IL1* cytokines, and its absence in chimpanzee may point to species-specific modulation of the interferon- γ - and lipopolysaccharide-induced inflammatory response¹²⁷.

Parasite resistance. Similarly, we found that two members of the primate-specific *APOL* gene cluster (*APOL1* and *APOL4*) have been deleted from the chimpanzee genome. The *APOL1* protein is associated with the high-density lipoprotein fraction in serum and has recently been proposed to be the lytic factor responsible for resistance to certain subspecies of *Trypanosoma brucei*, the parasite that causes human sleeping sickness and the veterinary disease nagana¹²⁸. The loss of the *APOL1* gene in chimpanzees could thus explain the observation that human, gorilla and baboon possess the trypanosome lytic factor, whereas the chimpanzee does not¹²⁹.

Sialic acid biology related proteins. Sialic acids are cell-surface sugars that mediate many biological functions¹³⁰. Of 54 genes involved in sialic acid biology, 47 were suitable for analysis. We confirmed and extended findings on several that have undergone human-specific changes, including disruptions, deletions and domain-specific functional changes^{113,131,132}. Human- and chimpanzee-specific changes were also found in otherwise evolutionarily conserved sialyl motifs in four sialyl transferases (*ST6GAL1*, *ST6GALNAC3*, *ST6GALNAC4* and *ST8SIA2*), suggesting changes in donor and/or acceptor binding¹³⁰. Lineage-specific changes were found in a complement factor H (*HF1*) sialic acid binding domain associated with human disease¹³³. Human *SIGLEC11* has undergone gene conversion with a nearby pseudogene, correlating with acquisition of human-specific brain expression and altered binding properties¹³⁴.

Human disease alleles. We next sought to identify putative functional differences between the species by searching for instances in which a human disease-causing allele appears to be the wild-type allele in the chimpanzee. Starting from 12,164 catalogued disease variants in 1,384 human genes, we identified 16 cases in which the altered sequence in a disease allele matched the chimpanzee sequence, and had plausible support in the literature (Table 7; see also Supplementary Table S43). Upon re-sequencing in seven chimpanzees, 15 cases were confirmed homozygous in all individuals, whereas one (*PONI* I102V) appears to be a shared polymorphism (Supplementary Table S44).

Six cases represent *de novo* human mutations associated with simple mendelian disorders. Similar cases have also been found in comparisons of more distantly related mammals¹³⁵, as well as

between insects¹³⁶, and have been interpreted as a consequence of a relatively high rate of compensatory mutations. If compensatory mutations are more likely to be fixed by positive selection than by neutral drift¹³⁶, then the variants identified here might point towards adaptive differences between humans and chimpanzees. For example, the ancestral Thr 29 allele of cationic trypsinogen (*PRSSI*) causes autosomal dominant pancreatitis in humans¹³⁷, suggesting that the human-specific Asn 29 allele may represent a digestion-related molecular adaptation¹³⁸.

The remaining ten cases represent common human polymorphisms that have been reported to be associated with complex traits, including coronary artery disease and diabetes mellitus. In all of these cases we confirmed that the disease-associated allele in humans is indeed the ancestral allele by showing that it is carried not only by chimpanzee but also by outgroups such as the macaque. These ancestral alleles may thus have become human-specific risk factors due to changes in human physiology or environment, and the polymorphisms may represent ongoing adaptations. For example, *PPARG* Pro 12 is the wild-type allele in chimpanzee but has been clearly associated with increased risk of type 2 diabetes in human¹³⁹. It is tempting to speculate that this allele may represent an ancestral 'thrifty' genotype¹⁴⁰.

The current results must be interpreted with caution, because few complex disease associations have been firmly established. The fact that the human disease allele is the wild-type allele in chimpanzee may actually indicate that some of the putative associations are spurious and not causal. However, this approach can be expected to become increasingly fruitful as the quality and completeness of the disease mutation databases improve.

Human population genetics

The chimpanzee has a special role in informing studies of human population genetics, a field that is undergoing rapid expansion and acquiring new relevance to human medical genetics¹⁴¹. The chimpanzee sequence allows recognition of those human alleles that represent the ancestral state and the derived state. It also allows estimates of local mutation rates, which serve as an important baseline in searching for signs of natural selection.

Ancestral and derived alleles. Of ~7.2 million SNPs mapped to the human genome in the current public database, we could assign the alleles as ancestral or derived in 80% of the cases according to which allele agrees with the chimpanzee genome sequence¹⁴² (see Supplementary Information 'Human population genetics'). For the remaining cases, no assignment could be made because of the following: the orthologous chimpanzee base differed from both human alleles (1.2%); was polymorphic in the chimpanzee sequences obtained (0.4%); or could not be reliably identified with the current draft sequence of the chimpanzee (18.8%), with many of these occurring in repeated or segmentally duplicated sequence. The first two cases arise presumably because a second mutation occurred in the chimpanzee lineage. It should be possible to resolve most of these cases by examining a close outgroup such as gorilla or orang-utan.

Mutations in the chimpanzee may also lead to the erroneous assignment of human alleles as derived alleles. This error rate can be estimated as the probability of a second mutation resulting in the chimpanzee sequence matching the derived allele (see Supplementary Information 'Human population genetics'). The estimated error rate for typical SNPs is 0.5%, owing to the low nucleotide substitution rate. The exceptions are those SNPs for which the human alleles are CpG and TpG and the chimpanzee sequence is TpG. For these, a non-negligible fraction may have arisen by two independent deamination events within an ancestral CpG dinucleotide, which are well-known mutational hotspots⁵¹ (also see above). Human SNPs in a CpG context for which the orthologous chimpanzee sequence is TpG account for 12% of the total, and have an estimated error rate of 9.8%. Across all SNPs, the average error rate, ϵ , is thus estimated to be ~1.6%.

We compared the distribution of allele frequencies for ancestral

Table 7 | Candidate human disease variants found in chimpanzee

Gene	Variant*	Disease association	Ancestral†	Frequency‡
<i>AIRE</i>	P252L ¹⁵⁹	Autoimmune syndrome	Unresolved	0
<i>MKKS</i>	R518H ¹⁶⁰	Bardet-Biedl syndrome	Wild type	0
<i>MLH1</i>	A441T ¹⁶¹	Colorectal cancer	Wild type	0
<i>MYOC</i>	Q48H ¹⁶²	Glaucoma	Wild type	0
<i>OTC</i>	T125M ¹⁶³	Hyperammonaemia	Wild type	0
<i>PRSS1</i>	N29T ¹³⁷	Pancreatitis	Disease	0
<i>ABCA1</i>	I883M ¹⁶⁴	Coronary artery disease	Unresolved	0.136
<i>APOE</i>	C130R ¹⁶⁵	Coronary artery disease and Alzheimer's disease	Disease	0.15
<i>DIO2</i>	T92A ¹⁶⁶	Insulin resistance	Disease	0.35
<i>ENPP1</i>	K121Q ¹⁶⁷	Insulin resistance	Disease	0.17
<i>GSTP1</i>	I105V ¹⁶⁸	Oral cancer	Disease	0.348
<i>PON1S</i>	I102V ¹⁶⁹	Prostate cancer	Wild type	0.016
<i>PONI</i>	Q192R ¹⁷⁰	Coronary artery disease	Disease	0.3
<i>PPARG</i>	A12P ¹³⁹	Type 2 diabetes	Disease	0.85
<i>SLC2A2</i>	T110I ¹⁷¹	Type 2 diabetes	Disease	0.12
<i>UCP1</i>	A64T ¹⁷²	Waist-to-hip ratio	Disease	0.12

* This takes the following format: benign variant, codon number, disease/chimpanzee variant.

† Ancestral variant as inferred from closest available primate outgroups (Supplementary Information).

‡ Frequency of the disease allele in human study population.

§ Polymorphic in chimpanzee.

and derived alleles using a database of allele frequencies for $\sim 120,000$ SNPs (see Supplementary Information ‘Human population genetics’). As expected, ancestral alleles tend to have much higher frequencies than derived alleles (Supplementary Fig. S9). Nonetheless, a significant proportion of derived alleles have high frequencies: 9.1% of derived alleles have frequency $\geq 80\%$.

An elegant result in population genetics states that, for a randomly interbreeding population of constant size, the probability that an allele is ancestral is equal to its frequency¹⁴³. We explored the extent to which this simple theoretical expectation fits the human population. We tabulated the proportion $p_a(x)$ of ancestral alleles for various frequencies of x and compared this with the prediction $p_a(x) = x$ (Fig. 13).

The data lie near the predicted line, but the observed slope (0.83) is substantially less than 1. One explanation for this deviation is that some ancestral alleles are incorrectly assigned (an error rate of ε would artificially decrease the slope by a factor of $1-2\varepsilon$). However, with ε estimated to be only 1.6%, errors can only explain a small part of the deviation. The most likely explanation is the presence of bottlenecks during human history, which tend to flatten the distribution of allele frequencies. Theoretical calculations indicate that a recent bottleneck would decrease the slope by a factor of $(1-b)$, where b is the inbreeding coefficient induced by the bottleneck (see Supplementary Information ‘Human population genetics’ and Supplementary Fig. S10). This suggests that measurements of the slope in different human groups may shed light on population-specific bottlenecks. Consistent with this, preliminary analyses of allele frequencies in several regions for SNPs obtained by systematic uniform sampling indicate that the slope is significantly lower than 1 in European and Asian samples and close to 1 in an African sample (see Supplementary Information ‘Human population genetics’ and Supplementary Fig. S11).

Signatures of strong selective sweeps in recent human history. The pattern of human genetic variation holds substantial information about selection events that have shaped our species. Strong positive selection creates the distinctive signature of a ‘selective sweep’, whereby a rare allele rapidly rises to fixation and carries the haplotype on which it occurs to high frequency (the ‘hitchhiking’ effect). The surrounding region should show two distinctive signatures: a significant reduction of overall diversity, and an excess of derived alleles with high frequency in the population owing to hitchhiking of

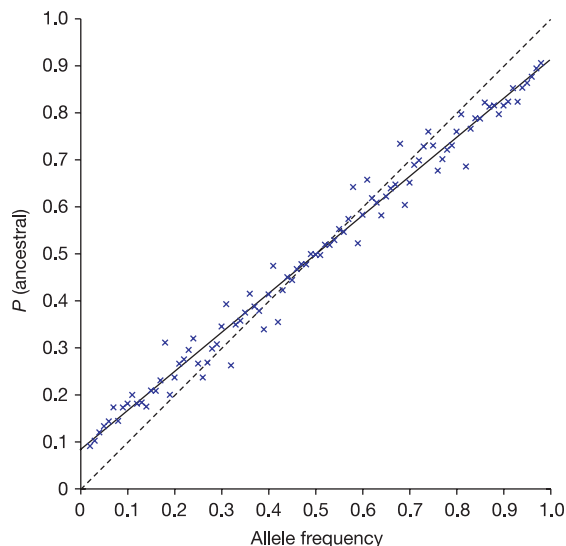


Figure 13 | The observed fraction of ancestral alleles in 1% bins of observed frequency. The solid line shows the regression ($b = 0.83$). The dotted line shows the theoretical relationship $p_a(x) = x$. Note that because each variant yields a derived and an ancestral allele, the data are necessarily symmetrical about 0.5.

derived alleles on the selected haplotype (see Supplementary Information ‘Human population genetics’). The pattern might be detectable for up to 250,000 years after a selective sweep has ended¹⁴⁴. Notably, the chimpanzee genome provides crucial baseline information required for accurate assessment of both signatures.

The size of the interval affected by a selective sweep is expected to scale roughly with s , the selective advantage due to the mutation. Simulations can be used to study the distribution of the interval size (see Supplementary Information ‘Human population genetics’). With $s = 1\%$, the interval over which heterozygosity falls by 50% has a modal size of 600 kb and a probability of greater than 10% of exceeding 1 Mb.

We undertook an initial scan for large regions (>1 Mb) with the two signatures suggestive of strong selective sweeps in recent human history. We began by identifying regions in which the observed human diversity rate was much lower than the expectation based on the observed divergence rate with chimpanzee. The human diversity rate was measured as the number of occurrences from a database of 1.92 million SNPs identified by shotgun sequencing in a panel of African–American individuals (see Supplementary Information ‘Genome sequencing and assembly’). The comparison with the chimpanzee eliminates regions in which low diversity simply reflects a low mutation rate in the region. Regions were identified based on a simple statistical procedure (see Supplementary Information ‘Human population genetics’). Six genomic regions stand out as clear outliers that show significantly reduced diversity relative to divergence (Table 8; see also Supplementary Fig. S12).

We next tested whether these six regions show a high proportion of SNPs with high-frequency derived alleles (defined here as alleles with frequency $\geq 80\%$). Within each region, we focused on the 1-Mb interval with the greatest discrepancy between diversity and divergence and compared it to 1-Mb regions throughout the genome. For the database of 120,000 SNPs with allele frequencies discussed above, the typical 1-Mb region in the human genome contains ~ 40 SNPs, and the proportion p_h of SNPs with high-frequency derived alleles is $\sim 9.1\%$. All six regions identified by our scan for reduced diversity have a higher than average fraction of high-frequency derived alleles; all six fall within the top 10% genome-wide and three fall within the top 1%. Although this is not definitive evidence for any particular region, the joint probability of all six regions randomly scoring in the top 10% is 10^{-6} . The results indicate that the six regions are candidates for strong selective sweeps during the past 250,000 years¹⁴⁴. The regions differ notably with respect to gene content, ranging from one containing 57 annotated genes (chromosome 22) to another with no annotated genes whatsoever (chromosome 4). We have no evidence to implicate any individual functional element as a target of recent selection at this point, but the regions contain a number of interesting candidates for follow-up studies. Intriguingly, the chromosome 4 gene desert, which flanks a protocadherin gene and is conserved across vertebrates¹⁵, has been implicated in two independent studies as being associated with obesity^{145,146}.

In addition to the six regions, one further genomic region deserves mention: an interval of 7.6 Mb on chromosome 7q (see Supplementary Information ‘Human population genetics’). The interval contains several regions with high scores in the diversity–divergence analysis (including the seventh highest score overall) as well as in the proportion of high-frequency derived alleles. The region contains the *FOXP2* and *CFTR* genes. The former has been the subject of much interest as a possible target for selection during human evolution¹⁴⁷ and the latter as a target of selection in European populations¹⁴⁸.

Convincing proof of past selection will require careful analysis of the precise pattern of genetic variation in the region and the identification of a likely target of selection. Nonetheless, our findings suggest that the approach outlined here may help to unlock some of the secrets of recent human evolution through a combination of within-species and cross-species comparison.

Table 8 | Human regions with strongest signal of selection based on diversity relative to divergence

Chromosome	Start (Mb)	End (Mb)	Regression log-score	Skew P-value	Genes
1	48.58	52.58	103.3	0.071	Fourteen known genes from <i>ELAVL4</i> to <i>GPX7</i>
2	144.35	148.47	84.8	0.074	<i>ARHGAP15</i> (partial), <i>GTDC1</i> and <i>ZFH1B</i>
22	36.15	40.22	81.8	0.00022	Fifty-seven known genes from <i>CARD10</i> to <i>PMM1</i>
12	84.69	89.01	80.9	0.031	Ten known genes from <i>PAMC1</i> to <i>ATP2B1</i>
8	34.91	37.54	76.9	0.00032	<i>UNC5D</i> and <i>FKSG2</i>
4	32.42	35.62	55.9	0.00067	No known genes or Ensembl predictions

Discussion

Our knowledge of the human genome is greatly advanced by the availability of a second hominid genome. Some questions can be directly answered by comparing the human and chimpanzee sequences, including estimates of regional mutation rates and average selective constraints on gene classes. Other questions can be addressed in conjunction with other large data sets, such as issues in human population genetics for which the chimpanzee genome provides crucial controls. For still other questions, the chimpanzee genome simply provides a starting point for further investigation.

The hardest such question is: what makes us human? The challenge lies in the fact that most evolutionary change is due to neutral drift. Adaptive changes comprise only a small minority of the total genetic variation between two species. As a result, the extent of phenotypic variation between organisms is not strictly related to the degree of sequence variation. For example, gross phenotypic variation between human and chimpanzee is much greater than between the mouse species *Mus musculus* and *Mus spretus*, although the sequence difference in the two cases is similar. On the other hand, dogs show considerable phenotypic variation despite having little overall sequence variation (~0.15%). Genomic comparison markedly narrows the search for the functionally important differences between species, but specific biological insights will be needed to sift the still-large list of candidates to separate adaptive changes from neutral background.

Our comparative analysis suggests that the patterns of molecular evolution in the hominids are typical of a broader class of mammals in many ways, but distinctive in certain respects. As with the murids, the most rapidly evolving gene families are those involved in reproduction and host defence. In contrast to the murids, however, hominids appear to experience substantially weaker negative selection; this probably reflects their smaller population size. Consequently, hominids accumulate deleterious mutations that would be eliminated by purifying selection in murids. This may be both an advantage and a disadvantage. Although decreased purifying selection may tend to erode overall fitness, it may also allow hominids to 'explore' larger regions of the fitness landscape and thereby achieve evolutionary adaptations that can only be reached by passing through intermediate states of inferior fitness^{149,150}.

Although the analyses presented here focus on protein-coding sequences, the chimpanzee genome sequence also allows systematic analysis of the recent evolution of gene regulatory elements for the first time. Initial analysis of both gene expression patterns and promoter regions suggest that their overall patterns of evolution closely mirror that of protein-coding regions. In an accompanying paper⁸³, we show that the rates of change in gene expression among different tissues in human and chimpanzee correlate with the nucleotide divergence in the putative proximal promoters and even more interestingly with the average level of constraint on proteins in the same tissues. Another study¹⁵¹ has similarly used the chimpanzee sequence described here to show that gene promoter regions are also evolving under markedly less constraint in hominids than in murids.

The draft chimpanzee sequence here is sufficient for initial analyses, but it is still imperfect and incomplete. Definitive studies of gene and genome evolution—including pseudogene formation, gene family expansion and segmental duplication—will require high-

quality finished sequence. In this regard, we note that efforts are already underway to construct a BAC-based physical map and to increase the shotgun sequence coverage to approximately sixfold redundancy. The added coverage alone will not affect the analysis greatly, but plans are in place to produce finished sequence for difficult to sequence and important segments of the genome.

Our close biological relatedness to chimpanzees not only allows unique insights into human biology, it also creates ethical obligations. Although the genome sequence was acquired without harm to chimpanzees, the availability of the sequence may increase pressure to use chimpanzees in experimentation. We strongly oppose reducing the protection of chimpanzees and instead advocate the policy positions suggested by an accompanying paper¹⁵². Furthermore, the existence of chimpanzees and other great apes in their native habitats is increasingly threatened by human civilization. More effective policies are urgently needed to protect them in the wild. We hope that elaborating how few differences separate our species will broaden recognition of our duty to these extraordinary primates that stand as our siblings in the family of life.

METHODS

Sequencing and assembly. Approximately 22.5 million sequence reads were derived from both ends of inserts (paired end reads) from 4-, 10-, 40- and 180-kb clones, all prepared from primary blood lymphocyte DNA. Genomic resources available from the source animal include a lymphoid cell line (S006006) and genomic DNA (NS06006) at Coriell Cell Repositories (<http://locus.umdj.edu/ccr/>), as well as a BAC library (CHORI-251)¹⁵³ (see also Supplementary Information 'Genome sequencing and assembly').

Genome alignment. BLASTZ¹⁵⁴ was used to align non-repetitive chimpanzee regions against repeat-masked human sequence. BLAT¹⁵⁵ was subsequently used to align the more repetitive regions. The combined alignments were chained¹⁵⁶ and only best reciprocal alignments were retained for further analysis.

Insertions and deletions. Small insertion/deletion (indel) events (<15 kb) were parsed directly from the BLASTZ genome alignment by counting the number and size of alignment gaps between bases within the same contig. Sites of large-scale indels (>15 kb) were detected from discordant placements of paired sequence reads against the human assembly. Size thresholds were obtained from both human fosmid alignments on human sequence (40 ± 2.58 kb) and chimpanzee plasmid alignments against human chromosome 21 (4.5 ± 1.84 kb). Indels were inferred by two or more pairs surpassing these thresholds by more than two standard deviations and the absence of sequence data within the discordancy.

Gene annotation. A total of 19,277 human RefSeq transcripts¹⁵⁷, representing 16,045 distinct genes, were indirectly aligned to the chimpanzee sequence via the genome alignment. After removing low-quality sequences and likely alignment artefacts, an initial catalogue containing 13,454 distinct 1:1 human–chimpanzee orthologues was created for the analyses described here. A subset of 7,043 of these genes with unambiguous mouse and rat orthologues were realigned using Clustal W¹⁵⁸ for the lineage-specific analyses. Updated gene catalogues can be obtained from <http://www.ensembl.org>.

Rates of divergence. Nucleotide divergence rates were estimated using baseml with the REV model. Non-CpG rates were estimated from all sites that did not overlap a CG dinucleotide in either human or chimpanzee. K_A and K_S were estimated jointly for each orthologue using codeml with the F3x4 codon frequency model and no additional constraints, except for the comparison of divergent and polymorphic substitutions where K_A/K_S for both was estimated as $(\Delta A/N_A)/(\Delta S/N_S)$, with N_S/N_A , the ratio of synonymous to non-synonymous sites, estimated as 0.36 from the orthologue alignments. Unless otherwise specified, K_A/K_S for a set of genes was calculated by summing the number of substitutions and the number of sites to obtain K_A and K_S for the concatenated set before taking

the ratio. Hominid and murid pairwise rates were estimated independently from codons aligned across all four species. Human and chimpanzee lineage-specific K_A and K_S were estimated on an unrooted tree with both mouse and rat included. Lineage-specific rates were also estimated by parsimony, with essentially identical results (see Supplementary Information). K_1 was estimated from all interspersed repeats within 250 kb of the mid-point of each gene.

Accelerated evolution in GO categories. The binomial probability of observing X or more non-synonymous substitutions, given a total of $X + Y$ substitutions and the expected proportion x from all orthologues, was calculated by summing substitutions across the orthologues in each GO category. For the absolute rate test, Y = the number of synonymous substitutions in orthologues in the same category. For the relative rate tests, Y = the number of non-synonymous substitutions on the opposite lineage. Note that this binomial probability is simply a metric designed to identify potentially accelerated categories, it is not a P -value that can be used to reject directly the null hypothesis of no acceleration in that particular category. For each test, the observed number of categories with a binomial probability less than 0.001 was compared to the expected distribution of such outliers by repeating the procedure 10,000 times on randomly permuted GO annotations. The significance of the number of observed outliers n was estimated as the proportion of random trials yielding n or more outliers.

Detection of selective sweeps. The observed number of human SNPs, u_i , human bases, m_i , human–chimpanzee substitutions, v_i , and chimpanzee bases, n_i , within each set of non-overlapping 1-Mb windows along the human genome were used to generate two random numbers, x_i (adjusted human diversity) and y_i (adjusted human–chimpanzee divergence), from the two beta-distributions:

$$x_i \approx \text{Beta}(u_i + a, m_i - u_i + b)$$

$$y_i \approx \text{Beta}(v_i + c, n_i - v_i + d)$$

where $a = 1$, $b = 1,000$, $c = 1$ and $d = 100$. These numbers were then fit to a linear regression:

$$x|y \approx N(\alpha_0 + \alpha_1 y, \beta^2)$$

A P -value for each window was calculated for each window based on (x_i, y_i) and the regression line. This was repeated 100 times and the average of the P -values taken as the P -value for diversity given divergence in each window. Overlapping windows with $P < 0.1$ containing at least one window of $P < 0.05$ were coalesced and scored as the sum of their $-\log(p)$ scores.

Received 21 March; accepted 20 July 2005.

- Darwin, C. *The Descent of Man, and Selection in Relation to Sex* (D Appleton and Company, New York, 1871).
- Huxley, T. H. *Evidence as to Man's Place in Nature* (Williams and Norgate, London, 1863).
- Goodman, M. The genomic record of humankind's evolutionary roots. *Am. J. Hum. Genet.* **64**, 31–39 (1999).
- Goodall, J. Tool-using and aimed throwing in a community of free-living chimpanzees. *Nature* **201**, 1264–1266 (1964).
- Whiten, A. *et al.* Cultures in chimpanzees. *Nature* **399**, 682–685 (1999).
- Olson, M. V. & Varki, A. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nature Rev. Genet.* **4**, 20–28 (2003).
- Eyre-Walker, A. & Keightley, P. D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347 (1999).
- Fay, J. C., Wyckoff, G. J. & Wu, C. I. Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234 (2001).
- King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
- Clark, A. G. *et al.* Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
- Hellmann, I. *et al.* Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13**, 831–837 (2003).
- Ebersberger, I., Metzler, D., Schwarz, C. & Paabo, S. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70**, 1490–1497 (2002).
- Watanabe, H. *et al.* DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429**, 382–388 (2004).
- Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946–957 (2004).
- Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
- Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
- McConkey, E. H. Orthologous numbering of great ape and human chromosomes is essential for comparative genomics. *Cytogenet. Genome Res.* **105**, 157–158 (2004).
- Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* **162**, 729–773 (1982).
- Myers, G. Whole-genome DNA sequencing. *Comput. Sci. Eng.* **1**, 33–43 (1999).
- Huang, X., Wang, J., Aluru, S., Yang, S. P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res.* **13**, 2164–2170 (2003).
- Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
- International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–920 (2001).
- Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
- She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
- Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* doi:10.1038/nature04000 (this issue).
- Fischer, A., Wiebe, V., Paabo, S. & Przeworski, M. Evidence for a complex demographic history of chimpanzees. *Mol. Biol. Evol.* **21**, 799–808 (2004).
- Yu, N. *et al.* Low nucleotide diversity in chimpanzees and bonobos. *Genetics* **164**, 1511–1518 (2003).
- Kaessmann, H., Wiebe, V., Weiss, G. & Paabo, S. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genet.* **27**, 155–156 (2001).
- Kitano, T., Schwarz, C., Nickel, B. & Paabo, S. Gene diversity patterns at 10 X-chromosomal loci in humans and chimpanzees. *Mol. Biol. Evol.* **20**, 1281–1289 (2003).
- The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
- Chen, F. C. & Li, W. H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444–456 (2001).
- Fujiyama, A. *et al.* Construction and analysis of a human-chimpanzee comparative clone map. *Science* **295**, 131–134 (2002).
- Hardison, R. C. *et al.* Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**, 13–26 (2003).
- Webster, M. T., Smith, N. G., Lercher, M. J. & Ellegren, H. Gene expression, synteny, and local similarity in human noncoding mutation rates. *Mol. Biol. Evol.* **21**, 1820–1830 (2004).
- Rosenberg, H. F. & Feldmann, M. W. *The Relationship Between Coalescence Times and Population Divergence Times* (Oxford Univ. Press, Oxford, 2002).
- Vignaud, P. *et al.* Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* **418**, 152–155 (2002).
- Wall, J. D. Estimating ancestral population sizes and divergence times. *Genetics* **163**, 395–404 (2003).
- Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).
- Maynard Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
- Hudson, R. R. & Kaplan, N. L. Deleterious background selection with recombination. *Genetics* **141**, 1605–1617 (1995).
- Charlesworth, B. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* **63**, 213–227 (1994).
- Birky, C. W. Jr & Walsh, J. B. Effects of linkage on rates of molecular evolution. *Proc. Natl Acad. Sci. USA* **85**, 6414–6418 (1988).
- Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S. & Przeworski, M. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**, 1527–1535 (2003).
- Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18**, 337–340 (2002).
- Hellmann, I. *et al.* Why do human diversity levels vary at a megabase scale? *Genome Res.* **15**, 1222–1231 (2005).
- Li, W. H., Yi, S. & Makova, K. Male-driven evolution. *Curr. Opin. Genet. Dev.* **12**, 650–656 (2002).
- Bohossian, H. B., Skaletsky, H. & Page, D. C. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* **406**, 622–625 (2000).
- Makova, K. D. & Li, W. H. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624–626 (2002).
- Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl Acad. Sci. USA* **101**, 13994–14001 (2004).
- Taylor, J., Tyekucheva, S., Zody, M., Ciaromonte, F. & Makova, K. D. Strong and weak male mutation bias at different sites in the primate genomes: Insights from the human-chimpanzee comparison. *Mol. Biol. Evol.* (submitted).
- Bulmer, M., Wolfe, K. H. & Sharp, P. M. Synonymous nucleotide substitution

- rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. *Proc. Natl Acad. Sci. USA* **88**, 5974–5978 (1991).
54. Ehrlich, M., Zhang, X. Y. & Inamdar, N. M. Spontaneous deamination of cytosine and 5-methylcytosine residues in DNA and replacement of 5-methylcytosine residues with cytosine residues. *Mutat. Res.* **238**, 277–286 (1990).
 55. Craig, J. M. & Bickmore, W. A. Chromosome bands—flavours to savour. *Bioessays* **15**, 349–354 (1993).
 56. Holmquist, G. P. Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.* **51**, 17–37 (1992).
 57. Ellegren, H., Smith, N. G. & Webster, M. T. Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* **13**, 562–568 (2003).
 58. Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S. & Sidow, A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13**, 813–820 (2003).
 59. Cooper, G. M. *et al.* Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* **14**, 539–548 (2004).
 60. Yang, S. *et al.* Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. *Genome Res.* **14**, 517–527 (2004).
 61. Birdsall, J. A. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.* **19**, 1181–1197 (2002).
 62. Jensen-Seaman, M. I. *et al.* Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14**, 528–538 (2004).
 63. Fortna, A. *et al.* Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* **2**, E207 (2004).
 64. Britten, R. J. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl Acad. Sci. USA* **99**, 13633–13635 (2002).
 65. Frazer, K. A. *et al.* Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* **13**, 341–346 (2003).
 66. Locke, D. P. *et al.* Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* **13**, 347–357 (2003).
 67. Liu, G. *et al.* Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13**, 358–368 (2003).
 68. Yohn, C. T. *et al.* Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLoS Biol.* **3**, 1–11 (2005).
 69. Hedges, D. J. *et al.* Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res.* **14**, 1068–1075 (2004).
 70. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663 (1999).
 71. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nature Genet.* **35**, 41–48 (2003).
 72. Mathews, L. M., Chi, S. Y., Greenberg, N., Ovchinnikov, I. & Swergold, G. D. Large differences between LINE-1 amplification rates in the human and chimpanzee lineages. *Am. J. Hum. Genet.* **72**, 739–748 (2003).
 73. Pickeral, O. K., Makalowski, W., Boguski, M. S. & Boeke, J. D. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10**, 411–415 (2000).
 74. Goodier, J. L., Ostertag, E. M. & Kazazian, H. H. Jr Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000).
 75. Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13**, 2541–2558 (2003).
 76. Torrents, D., Suyama, M., Zdobnov, E. & Bork, P. A genome-wide survey of human pseudogenes. *Nature Genet.* **13**, 2559–2567 (2003).
 77. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24**, 363–367 (2000).
 78. Zhang, Z., Harrison, P. & Gerstein, M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* **12**, 1466–1482 (2002).
 79. Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H. Jr SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am. J. Hum. Genet.* **73**, 1444–1451 (2003).
 80. Shen, L. *et al.* Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *J. Biol. Chem.* **269**, 8466–8476 (1994).
 81. Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA* **99**, 3740–3745 (2002).
 82. Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343 (2002).
 83. Khaitovich, P. *et al.* Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* (in the press).
 84. Yunis, J. J., Sawyer, J. R. & Dunham, K. The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. *Science* **208**, 1145–1148 (1980).
 85. Fan, Y., Linardopoulou, E., Friedman, C., Williams, E. & Trask, B. J. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13-2q14.1 and paralogous regions on other human chromosomes. *Genome Res.* **12**, 1651–1662 (2002).
 86. Fan, Y., Newman, T., Linardopoulou, E. & Trask, B. J. Gene content and function of the ancestral chromosome fusion site in human chromosome 2q13-2q14.1 and paralogous regions. *Genome Res.* **12**, 1663–1672 (2002).
 87. Locke, D. P. *et al.* Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol.* **4**, R50 (2003).
 88. Dennehey, B. K., Gutches, D. G., McConkey, E. H. & Krauter, K. S. Inversion, duplication, and changes in gene context are associated with human chromosome 18 evolution. *Genomics* **83**, 493–501 (2004).
 89. Subramanian, S. & Kumar, S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* **13**, 838–844 (2003).
 90. Duret, L. Detecting genomic features under weak selective pressure: the example of codon usage in animals and plants. *Bioinformatics* **18** (suppl. 2), S91 (2002).
 91. Sharp, P. M. & Li, W. H. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res.* **14**, 7737–7749 (1986).
 92. Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G. & Peden, J. F. DNA sequence evolution: the sounds of silence. *Phil. Trans. R. Soc. Lond. B* **349**, 241–247 (1995).
 93. Moriyama, E. N. & Powell, J. R. Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. *J. Mol. Evol.* **45**, 378–391 (1997).
 94. McVean, G. A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
 95. Ohta, T. Slightly deleterious mutant substitutions during evolution. *Nature* **246**, 96–98 (1973).
 96. Ohta, T. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* **40**, 56–63 (1995).
 97. Eyre-Walker, A., Keightley, P. D., Smith, N. G. & Gaffney, D. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.* **19**, 2142–2149 (2002).
 98. Makalowski, W. & Boguski, M. S. Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. *J. Mol. Evol.* **47**, 119–121 (1998).
 99. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
 100. Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176 (1992).
 101. Maier, A. G. *et al.* *Plasmodium falciparum* erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. *Nature Med.* **9**, 87–92 (2003).
 102. Stenger, S. *et al.* An antimicrobial activity of cytolytic T cells mediated by granulysin. *Science* **282**, 121–125 (1998).
 103. Swanson, W. J. & Vacquier, V. D. The rapid evolution of reproductive proteins. *Nature Rev. Genet.* **3**, 137–144 (2002).
 104. Choi, S. S. & Lahn, B. T. Adaptive evolution of MRG, a neuron-specific gene family implicated in nociception. *Genome Res.* **13**, 2252–2259 (2003).
 105. Hardison, R. C. *et al.* Global predictions and tests of erythroid regulatory regions. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 335–344 (2003).
 106. Lercher, M. J., Chamary, J. V. & Hurst, L. D. Genomic regionalism in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res.* **14**, 1002–1013 (2004).
 107. Williams, E. J. & Hurst, L. D. The proteins of linked genes evolve at similar rates. *Nature* **407**, 900–903 (2000).
 108. Navarro, A. & Barton, N. H. Chromosomal speciation and molecular divergence—accelerated evolution in rearranged chromosomes. *Science* **300**, 321–324 (2003).
 109. Zhang, J., Wang, X. & Podlaha, O. Testing the chromosomal speciation hypothesis for humans and chimpanzees. *Genome Res.* **14**, 845–851 (2004).
 110. Lu, J., Li, W. H. & Wu, C. I. Comment on "Chromosomal speciation and molecular divergence—accelerated evolution in rearranged chromosomes". *Science* **302**, 988 (2003).
 111. Charlesworth, B., Coyne, J. A. & Orr, H. A. Meiotic drive and unisexual hybrid sterility: a comment. *Genetics* **133**, 421–432 (1993).
 112. Ohno, S. *Evolution by Gene Duplication* (Springer, New York, 1970).
 113. Angata, T., Margulies, E. H., Green, E. D. & Varki, A. Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. *Proc. Natl Acad. Sci. USA* **101**, 13251–13256 (2004).
 114. Teumer, J. & Green, H. Divergent evolution of part of the involucrin gene in the hominoids: unique intragenic duplications in the gorilla and human. *Proc. Natl Acad. Sci. USA* **86**, 1283–1286 (1989).
 115. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
 116. Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496–503 (2000).
 117. Weinreich, D. M. The rates of molecular evolution in rodent and primate mitochondrial DNA. *J. Mol. Evol.* **52**, 40–50 (2001).
 118. Dorus, S. *et al.* Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell* **119**, 1027–1040 (2004).

119. Neilsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).
120. Vallender, E. J. & Lahn, B. T. Positive selection on the human genome. *Hum. Mol. Genet.* **13** (suppl. 2), R245–R254 (2004).
121. Gilad, Y., Man, O. & Glusman, G. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res.* **15**, 224–230 (2005).
122. Enard, W. & Paabo, S. Comparative primate genomics. *Annu. Rev. Genomics Hum. Genet.* **5**, 351–378 (2004).
123. Saleh, M. *et al.* Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature* **429**, 75–79 (2004).
124. Fischer, H., Koenig, U., Eckhart, L. & Tschachler, E. Human caspase 12 has acquired deleterious mutations. *Biochem. Biophys. Res. Commun.* **293**, 722–726 (2002).
125. Puente, X. S., Sanchez, L. M., Overall, C. M. & Lopez-Otin, C. Human and mouse proteases: a comparative genomic approach. *Nature Rev. Genet.* **4**, 544–558 (2003).
126. Nakagawa, T. *et al.* Caspase-12 mediates endoplasmic-reticulum-specific apoptosis and cytotoxicity by amyloid- β . *Nature* **403**, 98–103 (2000).
127. Humke, E. W., Shriver, S. K., Starovasnik, M. A., Fairbrother, W. J. & Dixit, V. M. ICEBERG: a novel inhibitor of interleukin-1 β generation. *Cell* **103**, 99–111 (2000).
128. Vanhamme, L. *et al.* Apolipoprotein L-1 is the trypanosome lytic factor of human serum. *Nature* **422**, 83–87 (2003).
129. Seed, J. R., Sechelski, J. B. & Loomis, M. R. A survey for a trypanocidal factor in primate sera. *J. Protozool.* **37**, 393–400 (1990).
130. Angata, T. & Varki, A. Chemical diversity in the sialic acids and related α -keto acids: an evolutionary perspective. *Chem. Rev.* **102**, 439–469 (2002).
131. Varki, A. How to make an ape brain. *Nature Genet.* **36**, 1034–1036 (2004).
132. Sonnenburg, J. L., Altheide, T. K. & Varki, A. A uniquely human consequence of domain-specific functional adaptation in a sialic acid-binding receptor. *Glycobiology* **14**, 339–346 (2004).
133. Pangburn, M. K. Host recognition and target differentiation by factor H, a regulator of the alternative pathway of complement. *Immunopharmacology* **49**, 149–157 (2000).
134. Hayakawa, T. *et al.* Human-specific gene in microglia. *Science* (in the press).
135. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99**, 14878–14883 (2002).
136. Kulathinal, R. J., Bettencourt, B. R. & Hartl, D. L. Compensated deleterious mutations in insect genomes. *Science* **306**, 1553–1554 (2004).
137. Pflutzer, R. *et al.* Novel cationic trypsinogen (PRSS1) N29T and R122C mutations cause autosomal dominant hereditary pancreatitis. *Gut* **50**, 271–272 (2002).
138. Chen, J. M., Montier, T. & Ferec, C. Molecular pathology and evolutionary and physiological implications of pancreatitis-associated cationic trypsinogen mutations. *Hum. Genet.* **109**, 245–252 (2001).
139. Altshuler, D. *et al.* The common PPAR γ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26**, 76–80 (2000).
140. Neel, J. V. Diabetes mellitus: a “thrifty” genotype rendered detrimental by “progress”? *Am. J. Hum. Genet.* **14**, 353–362 (1962).
141. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
142. Hacia, J. G. *et al.* Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Genet.* **22**, 164–167 (1999).
143. Watterson, G. A. & Guess, H. A. Is the most frequent allele the oldest? *Theor. Popul. Biol.* **11**, 141–160 (1977).
144. Przeworski, M. The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189 (2002).
145. Stone, S. *et al.* A major predisposition locus for severe obesity, at 4p15-p14. *Am. J. Hum. Genet.* **70**, 1459–1468 (2002).
146. Arya, R. *et al.* Evidence of a novel quantitative-trait locus for obesity on chromosome 4p in Mexican Americans. *Am. J. Hum. Genet.* **74**, 272–282 (2004).
147. Enard, W. *et al.* Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
148. Schroeder, S. A., Gaughan, D. M. & Swift, M. Protection against bronchial asthma by *CFTR* Δ F508 mutation: a heterozygote advantage in cystic fibrosis. *Nature Med.* **1**, 703–705 (1995).
149. Ohta, T. Evolution by nearly-neutral mutations. *Genetica* **102–103**, 83–90 (1998).
150. Hayakawa, T., Altheide, T. K. & Varki, A. Genetic basis of human brain evolution: accelerating along the primate speedway. *Dev. Cell* **8**, 2–4 (2005).
151. Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3**, e42 (2005).
152. Gagneux, P., Moore, J. J. & Varki, A. The ethics of research on great apes. *Nature* **437**, 27–29 (2005).
153. Osoegawa, K. *et al.* An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* **52**, 1–8 (1998).
154. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
155. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
156. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution’s cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100**, 11484–11489 (2003).
157. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.* **31**, 34–37 (2003).
158. Higgins, D. G., Thompson, J. D. & Gibson, T. J. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266**, 383–402 (1996).
159. Meloni, A. *et al.* Delineation of the molecular defects in the AIRE gene in autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy patients from Southern Italy. *J. Clin. Endocrinol. Metab.* **87**, 841–846 (2002).
160. Beales, P. L. *et al.* Genetic and mutational analyses of a large multiethnic Bardet-Biedl cohort reveal a minor involvement of *BBS6* and delineate the critical intervals of other loci. *Am. J. Hum. Genet.* **68**, 606–616 (2001).
161. Cunningham, J. M. *et al.* The frequency of hereditary defective mismatch repair in a prospective series of unselected colorectal carcinomas. *Am. J. Hum. Genet.* **69**, 780–790 (2001).
162. Mukhopadhyay, A. *et al.* Mutations in MYOC gene of Indian primary open angle glaucoma patients. *Mol. Vis.* **8**, 442–448 (2002).
163. Tuchman, M., Jaleel, N., Morizono, H., Sheehy, L. & Lynch, M. G. Mutations and polymorphisms in the human ornithine transcarbamylase gene. *Hum. Mutat.* **19**, 93–107 (2002).
164. Clee, S. M. *et al.* Common genetic variation in *ABCA1* is associated with altered lipoprotein levels and a modified risk for coronary artery disease. *Circulation* **103**, 1198–1205 (2001).
165. Fullerton, S. M. *et al.* Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.* **67**, 881–900 (2000).
166. Mentuccia, D. *et al.* Association between a novel variant of the human type 2 deiodinase gene Thr92Ala and insulin resistance: evidence of interaction with the Trp64Arg variant of the β -3-adrenergic receptor. *Diabetes* **51**, 880–883 (2002).
167. Pizzuti, A. *et al.* A polymorphism (K121Q) of the human glycoprotein PC-1 gene coding region is strongly associated with insulin resistance. *Diabetes* **48**, 1881–1884 (1999).
168. Katoh, T. *et al.* Human glutathione S-transferase P1 polymorphism and susceptibility to smoking related epithelial cancer; oral, lung, gastric, colorectal and urothelial cancer. *Pharmacogenetics* **9**, 165–169 (1999).
169. Marchesani, M. *et al.* New paraoxonase 1 polymorphism I102V and the risk of prostate cancer in Finnish men. *J. Natl Cancer Inst.* **95**, 812–818 (2003).
170. Humbert, R. *et al.* The molecular basis of the human serum paraoxonase activity polymorphism. *Nature Genet.* **3**, 73–76 (1993).
171. Barroso, I. *et al.* Candidate gene association study in type 2 diabetes indicates a role for genes involved in β -cell function as well as insulin action. *PLoS Biol.* **1**, E20 (2003).
172. Herrmann, S. M. *et al.* Uncoupling protein 1 and 3 polymorphisms are associated with waist-to-hip ratio. *J. Mol. Med.* **81**, 327–332 (2003).
173. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Generation of the *Pan troglodytes* sequence at Washington University School of Medicine’s Genome Sequencing Center and the Broad Institute of MIT and Harvard was supported by grants from the National Human Genome Research Institute (NHGRI). We would like to thank the entire staff of both of those institutions. For work from other groups, we acknowledge the support of the European Molecular Biology Laboratory, Ministerio de Educacion y Ciencia (Spain), Howard Hughes Medical Institute, NHGRI, National Institutes of Health and National Science Foundation. Resources for exploring the sequence and annotation data are available on browser displays available at UCSC (<http://genome.ucsc.edu>), Ensembl (<http://www.ensembl.org>) and the NCBI (<http://www.ncbi.nlm.nih.gov>). We thank L. Gaffney for graphical help.

Author Contributions The last three authors co-directed the work.

Author Information This *Pan troglodytes* whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accessions ARACHNE, AADA01000000 and PCAP, AACZ01000000. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.H.V. (waterston@gs.washington.edu) E.S.L. (lander@broad.mit.edu) or R.K.W. (rwilson@watson.wustl.edu).

The Chimpanzee Sequencing and Analysis Consortium Tarjei S. Mikkelsen^{1,2}, LaDeana W. Hillier³, Evan E. Eichler⁴, Michael C. Zody¹, David B. Jaffe¹, Shiaw-Pyng Yang³, Wolfgang Enard⁵, Ines Hellmann⁵, Kerstin Lindblad-Toh¹, Tasha K. Altheide⁶, Nicoletta Archidiacono⁷, Peer Bork^{8,9}, Jonathan Butler¹, Jean L. Chang¹, Ze Cheng⁴, Asif T. Chinwalla³, Pieter deJong¹⁰, Kimberley D. Delehaunty³, Catrina C. Fronick³, Lucinda L. Fulton³, Yoav Gilad¹¹, Gustavo Glusman¹², Sante Gnerre¹, Tina A. Graves³, Toshiyuki Hayakawa⁶, Karen E. Hayden¹³, Xiaohu Huang¹⁴, Hongkai Ji¹⁵, W. James Kent¹⁶, Mary-Claire King⁴, Edward J. Kulbokas III¹, Ming K. Lee⁴, Ge Liu¹³, Carlos Lopez-Otin¹⁷, Kateryna D. Makova¹⁸, Orna Man¹⁹, Elaine R. Mardis³, Evan Mauceli¹, Tracie L. Miner³, William E. Nash³, Joanne O. Nelson³, Svante Pääbo⁵, Nick J. Patterson¹, Craig S. Pohl³, Katherine S. Pollard¹⁶, Kay Prüfer⁵, Xose S. Puente¹⁷, David Reich^{1,20}, Mariano Rocchi⁷, Kate Rosenbloom¹⁶, Maryellen Ruvolo²¹, Daniel J. Richter¹, Stephen F. Schaffner¹, Arian F. A. Smit¹², Scott M. Smith³, Mikita Suyama⁸, James Taylor¹⁸, David Torrents⁸, Eray Tuzun⁴, Ajit Varki⁶, Gloria Velasco¹⁷, Mario Ventura⁷, John W. Wallis³, Michael C. Wendt³, Richard K. Wilson³, Eric S. Lander^{1,22,23,24} & Robert H. Waterston⁴

Affiliations for participants: ¹Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141, USA. ²Division of Health Sciences and Technology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. ³Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. ⁴Genome Sciences, University of Washington School of Medicine, 1705 NE Pacific Street, Seattle, Washington 98195, USA. ⁵Max Planck Institute of Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany. ⁶University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. ⁷Department of Genetics and Microbiology, University of Bari, 70126 Bari, Italy. ⁸EMBL, Meyerhofstrasse 1, Heidelberg D-69117, Germany. ⁹Max Delbrück Center for Molecular Medicine (MDC), Robert-Rössle-Strasse 10, D-13125 Berlin, Germany. ¹⁰Children's Hospital Oakland Research Institute, 747 52nd Street, Oakland, California 94609, USA. ¹¹Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06520, USA. ¹²Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103, USA. ¹³Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA. ¹⁴Department of Computer Science, Iowa State University, 226 Atanasoff Hall, Ames, Iowa 50011, USA. ¹⁵Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138, USA. ¹⁶University of California, Santa Cruz, Center for Biomolecular Science and Engineering, 1156 High Street, Santa Cruz, California 95064, USA. ¹⁷Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología del Principado de Asturias, Universidad de Oviedo, C/Fernando Bongera s/n, 33006 Oviedo, Spain. ¹⁸The Pennsylvania State University, Center for Comparative Genomics and Bioinformatics and Department of Biology, University Park, Pennsylvania 16802, USA. ¹⁹Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel. ²⁰Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. ²¹Departments of Anthropology and of Organismic and Evolutionary Biology, Harvard University, 11 Divinity Avenue, Cambridge, Massachusetts 02138, USA. ²²Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. ²³Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA. ²⁴Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

[This page is intentionally left blank]

Genome sequence, comparative analysis and haplotype structure of the domestic dog

Kerstin Lindblad-Toh¹, Claire M Wade^{1,2}, Tarjei S. Mikkelsen^{1,3}, Elinor K. Karlsson^{1,4}, David B. Jaffe¹, Michael Kamal¹, Michele Clamp¹, Jean L. Chang¹, Edward J. Kulbokas III¹, Michael C. Zody¹, Evan Mauceli¹, Xiaohui Xie¹, Matthew Breen⁵, Robert K. Wayne⁶, Elaine A. Ostrander⁷, Chris P. Ponting⁸, Francis Galibert⁹, Douglas R. Smith¹⁰, Pieter J. deJong¹¹, Ewen Kirkness¹², Pablo Alvarez¹, Tara Biagi¹, William Brockman¹, Jonathan Butler¹, Chee-Wye Chin¹, April Cook¹, James Cuff¹, Mark J. Daly^{1,2}, David DeCaprio¹, Sante Gnerre¹, Manfred Grabherr¹, Manolis Kellis^{1,13}, Michael Kleber¹, Carolyn Bardeleben⁶, Leo Goodstadt⁸, Andreas Heger⁸, Christophe Hitte⁹, Lisa Kim⁷, Klaus-Peter Koepfli⁶, Heidi G. Parker⁷, John P. Pollinger⁶, Stephen M. J. Searle¹⁴, Nathan B. Sutter⁷, Rachael Thomas⁵, Caleb Webber⁸, Broad Institute Genome Sequencing Platform* & Eric S. Lander^{1,15}

Here we report a high-quality draft genome sequence of the domestic dog (*Canis familiaris*), together with a dense map of single nucleotide polymorphisms (SNPs) across breeds. The dog is of particular interest because it provides important evolutionary information and because existing breeds show great phenotypic diversity for morphological, physiological and behavioural traits. We use sequence comparison with the primate and rodent lineages to shed light on the structure and evolution of genomes and genes. Notably, the majority of the most highly conserved non-coding sequences in mammalian genomes are clustered near a small subset of genes with important roles in development. Analysis of SNPs reveals long-range haplotypes across the entire dog genome, and defines the nature of genetic diversity within and across breeds. The current SNP map now makes it possible for genome-wide association studies to identify genes responsible for diseases and traits, with important consequences for human and companion animal health.

Man's best friend, *Canis familiaris*, occupies a special niche in genomics. The unique breeding history of the domestic dog provides an unparalleled opportunity to explore the genetic basis of disease susceptibility, morphological variation and behavioural traits. The position of the dog within the mammalian evolutionary tree also makes it an important guide for comparative analysis of the human genome.

The history of the domestic dog traces back at least 15,000 years, and possibly as far back as 100,000 years, to its original domestication from the grey wolf in East Asia^{1–4}. Dogs evolved through a mutually beneficial relationship with humans, sharing living space and food sources. In recent centuries, humans have selectively bred dogs that excel at herding, hunting and obedience, and in this process have created breeds rich in behaviours that both mimic human behaviours and support our needs. Dogs have also been bred for desired physical characteristics such as size, skull shape, coat colour and texture⁵,

producing breeds with closely delineated morphologies. This evolutionary experiment has produced diverse domestic species, harbouring more morphological diversity than exists within the remainder of the family Canidae⁶.

As a consequence of these stringent breeding programmes and periodic population bottlenecks (for example, during the World Wars), many of the ~400 modern dog breeds also show a high prevalence of specific diseases, including cancers, blindness, heart disease, cataracts, epilepsy, hip dysplasia and deafness^{7,8}. Most of these diseases are also commonly seen in the human population, and clinical manifestations in the two species are often similar⁹. The high prevalence of specific diseases within certain breeds suggests that a limited number of loci underlie each disease, making their genetic dissection potentially more tractable in dogs than in humans¹⁰.

Genetic analysis of traits in dogs is enhanced by the close relationship between humans and canines in modern society.

¹Broad Institute of Harvard and MIT, 320 Charles Street, Cambridge, Massachusetts 02141, USA. ²Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA. ³Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ⁴Program in Bioinformatics, Boston University, 44 Cummings Street, Boston, Massachusetts 02215, USA. ⁵Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, 4700 Hillsborough Street, Raleigh, North Carolina 27606, USA. ⁶Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California 90095, USA. ⁷National Human Genome Research Institute, National Institutes of Health, 50 South Drive, MSC 8000, Building 50, Bethesda, Maryland 20892-8000, USA. ⁸MRC Functional Genetics, University of Oxford, Department of Human Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK. ⁹UMR 6061 Genetique et Developpement, CNRS—Université de Rennes 1, Faculté de Médecine, 2, Avenue Leon Bernard, 35043 Rennes Cedex, France. ¹⁰Agencourt Bioscience Corporation, 500 Cummings Center, Suite 2450, Beverly, Massachusetts 01915, USA. ¹¹Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr Way, Oakland, California 94609, USA. ¹²The Institute for Genomic Research, Rockville, Maryland 20850, USA. ¹³Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. ¹⁴The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ¹⁵Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA.

*A list of participants and affiliations appears at the end of the paper.

Through the efforts of the American Kennel Club (AKC) and similar organizations worldwide, extensive genealogies are easily accessible for most purebred dogs. With the exception of human, dog is the most intensely studied animal in medical practice, with detailed family history and pathology data often available⁸. Using genetic resources developed over the past 15 years^{11–16}, researchers have already identified mutations in genes underlying ~25 mendelian diseases^{17,18}. There are also growing efforts to understand the genetic basis of phenotypic variation such as skeletal morphology^{10,19}.

The dog is similarly important for the comparative analysis of mammalian genome biology and evolution. The four mammalian genomes that have been intensely analysed to date (human^{20–22}, chimpanzee²³, mouse²⁴ and rat²⁵) represent only one clade (Euarchontoglires) out of the four clades of placental mammals. The dog represents the neighbouring clade, Laurasiatheria²⁶. It thus serves as an outgroup to the Euarchontoglires and increases the total branch length of the current tree of fully sequenced mammalian genomes, thereby providing additional statistical power to search for conserved functional elements in the human genome^{24,27–33}. It also helps us to draw inferences about the common ancestor of the two clades, called the boreoeutherian ancestor, and provides a bridge to the two remaining clades (Afrotheria and Xenarthra) that should be helpful for anchoring low-coverage genome sequence currently being produced from species such as elephant and armadillo²⁸.

Here we report a high-quality draft sequence of the dog genome covering ~99% of the euchromatic genome. The completeness, nucleotide accuracy, sequence continuity and long-range connectivity are extremely high, exceeding the values calculated for the recent draft sequence of the mouse genome²⁴ and reflecting improved algorithms, higher-quality data, deeper coverage and intrinsic genome properties. We have also created a tool for the formal assessment of assembly accuracy, and estimate that >99% of the draft sequence is correctly assembled.

We also report an initial compendium of SNPs for the dog population, containing >2.5 million SNPs derived primarily from partial sequence comparison of 11 dog breeds to a reference sequence. We characterized the polymorphism rate of the SNPs across breeds and the long-range linkage disequilibrium (LD) of the SNPs within and across breeds.

We have analysed these data to study genome structure, gene evolution, haplotype structure and phylogenetics of the dog. Our key findings include:

- The evolutionary forces molding the mammalian genome differ among lineages, with the average transposon insertion rate being lowest in dog, the deletion rate being highest in mouse and the nucleotide substitution rate being lowest in human.
- Comparison between human and dog shows that ~5.3% of the human genome contains functional elements that have been under purifying selection in both lineages. Nearly all of these elements are confined to regions that have been retained in mouse, indicating that they represent a common set of functional elements across mammals.
- Fifty per cent of the most highly conserved non-coding sequence in the genome shows striking clustering in ~200 gene-poor regions, most of which contain genes with key roles in establishing or maintaining cellular identity, such as transcription factors or axon guidance receptors.
- Sets of functionally related genes show highly similar patterns of evolution in the human and dog lineages. This suggests that we should be careful about interpreting accelerated evolution in human relative to mouse as representing human-specific innovations (for example, in genes involved in brain development), because comparable acceleration is often seen in the dog lineage.
- Analysis across the entire genome of the sequenced boxer and across 6% of the genome in ten additional breeds shows that linkage disequilibrium (LD) within breeds extends over distances of several megabases, but LD across breeds only extends over tens of kilobases.

These LD patterns reflect two principal bottlenecks in dog history: early domestication and recent breed creation.

- Haplotypes within breeds extend over long distances, with ~3–5 alleles at each locus. Portions of these haplotypes, as large as 100 kilobases (kb), are shared across multiple breeds, although they are present at widely varying frequencies. The haplotype structure suggests that genetic risk factors may be shared across breeds.
- The current SNP map has sufficient density and an adequate within-breed polymorphism rate (~1/900 base pairs (bp) between breeds and ~1/1,500 bp within breeds) to enable systematic association studies to map genes affecting traits of interest. Genotyping of ~10,000 SNPs should suffice for most purposes.
- The genome sequence can be used to select a small collection of rapidly evolving sequences, which allows nearly complete resolution of the evolutionary tree of nearly all living species of Canidae.

Generating a draft genome sequence

We sequenced the genome of a female boxer using the whole-genome shotgun (WGS) approach^{22,24} (see Methods and Supplementary Table S1). A total of 31.5 million sequence reads, providing ~7.5-fold sequence redundancy, were assembled with an improved version of the ARACHNE program³⁴, resulting in an initial assembly (CanFam1.0) used for much of the analysis below, and an updated assembly (CanFam2.0) containing minor improvements (Table 1 and Supplementary Table S2).

Genome assembly. The recent genome assembly spans a total distance of 2.41 Gb, consisting of 2.38 Gb of nucleotide sequence with the remaining 1% in captured gaps. The assembly has extremely high continuity. The N50 contig size is 180 kb (that is, half of all bases reside in a contiguous sequence of 180 kb or more) and the N50 supercontig size is 45.0 Mb (Table 1). In particular, this means that most genes should contain no sequence gaps and that most canine chromosomes (mean size 61 Mb) have nearly all of their sequence ordered and oriented within one or two supercontigs (Supplementary Table S2). Notably, the sequence contigs are ~50-fold larger than the earlier survey sequence of the standard poodle¹⁶.

The assembly was anchored to the canine chromosomes using data from both radiation hybrid and cytogenetic maps^{11,13,14}. Roughly 97% of the assembled sequence was ordered and oriented on the chromosomes, showing an excellent agreement with the two maps. There were only three discrepancies, which were resolved by obtaining additional fluorescence *in situ* hybridization (FISH) data from the sequenced boxer. The 3% of the assembly that could not be anchored consists largely of highly repetitive sequence, including eight supercontigs of 0.5–1.0 Mb composed almost entirely of satellite sequence.

The nucleotide accuracy and genome coverage of the assembly is high (Supplementary Table S3). Of the bases in the assembly, 98% have quality scores exceeding 40, corresponding to an error rate of less than 10^{-4} and comparable to the standard for the finished human sequence³⁵. When we directly compared the assembly to 760 kb of finished sequence (in regions where the boxer is

Table 1 | Assembly statistics for CanFam1.0 and 2.0

	CanFam1.0	CanFam2.0
N50 contig size	123 kb	180 kb
N50 supercontig size	41.2 Mb	45.0 Mb
Assembly size (total bases)	2.360 Gb	2.385 Gb
Number of anchored supercontigs	86	87
Percentage of genome in anchored supercontigs	96	97
Sequence in anchored bases	2.290 Gb	2.309 Gb
Percentage of assembly in gaps	0.9	0.8
Estimated genome size*	2.411 Gb	2.445 Gb
Percentage of assembly in 'certified regions', without assembly inconsistency	99.3	99.6

*Includes anchored bases, spanned gaps (21 Mb in CanFam1.0, 18 Mb in CanFam2.0) and centromeric sequence (3 Mb for each chromosome).

homozygous, to eliminate differences attributable to polymorphisms; see below), we found that the draft genome sequence covers 99.8% of the finished sequence and that bases with quality scores exceeding 40 have an empirical error rate of 2×10^{-5} (Supplementary Table S3). **Explaining the high sequence continuity.** The dog genome assembly has superior sequence continuity (180 kb) than the WGS assembly of the mouse genome (25 kb) obtained several years ago²⁴. At least three factors contribute to the higher connectivity of the dog assembly (see Supplementary Information). First, we used a new version of ARACHNE with improved algorithms. Assembling the dog genome with the previous software version decreased N50 contig size from 180 kb to 61 kb, and assembling the mouse genome with the new version increased N50 contig size from 25 kb to 35 kb. Second, the amount of recently duplicated sequence is roughly twofold lower in dog than mouse (Supplementary Table S4); this improves contiguity because sequence gaps in both organisms tend to occur in recently duplicated sequence. Third, the dog sequence data has both higher redundancy (7.5-fold versus 6.5-fold) and higher quality (in terms of read length, pairing rate and tight distribution of insert sizes) compared with mouse. The contig size for the dog genome drops by about 32% when the data redundancy is decreased from 7.5-fold to 6.5-fold. A countervailing influence is that the dog genome contains polymorphism, whereas the laboratory mouse is completely inbred. **Assembly certification.** Although ‘quality scores’ have been developed to indicate the nucleotide accuracy of a draft genome sequence³⁶, no analogous measures have been developed to reflect the long-range assembly accuracy. We therefore sought to develop such a measure on the basis of two types of internal inconsistencies (see Supplementary Information). The first is haplotype inconsistency, involving clear evidence of three or more distinct haplotypes within an assembled region from a single diploid individual. The second is linkage inconsistency, involving a cluster of reads for which the placement of the paired-end reads is illogical. This includes cases in which: (1) one end cannot be mapped to the region, (2) the linkage relationships are inconsistent with the sequence within contigs, or (3) distance constraints imply overlap between non-overlapping sequence contigs. The linkage inconsistency tests are most powerful when read pairs are derived from clone libraries with tight constraints on insert size. A region of assembly is defined as ‘certified’ if it is free of inconsistencies, and is otherwise ‘questionable’.

Approximately 99.6% of the assembly resides in certified regions, with the N50 size of certified regions being ~12 Mb or about one-fifth of a chromosome. The remaining questionable regions are typically small (most are less than 40 kb), although there are a handful of regions of several hundred kilobases (Supplementary Fig. S1 and Supplementary Tables S5, S6). The questionable regions typically contain many inconsistencies, probably reflecting mis-assembly or overcollapse owing to segmental duplication. Chromosomes 2, 11 and 16 have 1.0–2.0% of their sequence in questionable regions. The certified and questionable regions are annotated in the public release of the dog genome assembly. With the concept of assembly certification, the scientific community can have appropriate levels of confidence in the draft genome sequence.

Genome landscape and evolution

Our understanding of the evolutionary processes that shape mammalian genomes has greatly benefited from the comparative analysis of sequenced primate^{21,23} and rodent^{24,25} genomes. However, the rodent genome is highly derived relative to that of the common ancestor of the eutherian mammals. As the first extensive sequence from an outgroup to the clade that includes primates and rodents, the dog genome offers a fresh perspective on mammalian genome evolution. Accordingly, we examined the rates and correlations of large-scale rearrangement, transposon insertion, deletion and nucleotide divergence across three major mammalian orders (primates, rodents and carnivores).

Conserved synteny and large-scale rearrangements. We created multi-species synteny maps from anchors of unique, unambiguously aligned sequences (see Supplementary Information), showing regions of conserved synteny among dog, human, mouse and rat genomes. Approximately 94% of the dog genome lies in regions of conserved synteny with the three other species (Supplementary Figs S2–S4 and Supplementary Table S7).

Given a pair of genomes, we refer to a ‘syntenic segment’ as a region that runs continuously without alterations of order and orientation, and a ‘syntenic block’ as a region that is contiguous in two genomes but may have undergone internal rearrangements. Syntenic breakpoints between blocks reflect primarily interchromosomal exchanges, and breakpoints between syntenic segments reflect intrachromosomal rearrangements. In the analysis below, we focus on syntenic segments of at least 500 kb.

We identified a total of 391 syntenic breakpoints across dog, human, mouse and rat genomes (Fig. 1 and Supplementary Figs S2, S5). With data for multiple species, it is possible to assign events to specific lineages (Fig. 1 and Supplementary Table S8). We counted the total number of breakpoints along the human, dog, mouse and rat lineages, with the values for each rodent lineage reflecting all breakpoints since the common ancestor with human (Fig. 1). The total number of breakpoints in the human lineage is substantially smaller than in the dog, mouse or rat lineages (83 versus 100, 161 or 176, respectively). However, there are more intrachromosomal breakpoints in the human lineage than in dog (52 versus 33).

Although the overall level of genomic rearrangement has been much higher in rodent than in human, comparison with dog shows that there are regions where the opposite is true. In particular, of the many intrachromosomal rearrangements previously observed between human chromosome 17 and the orthologous mouse

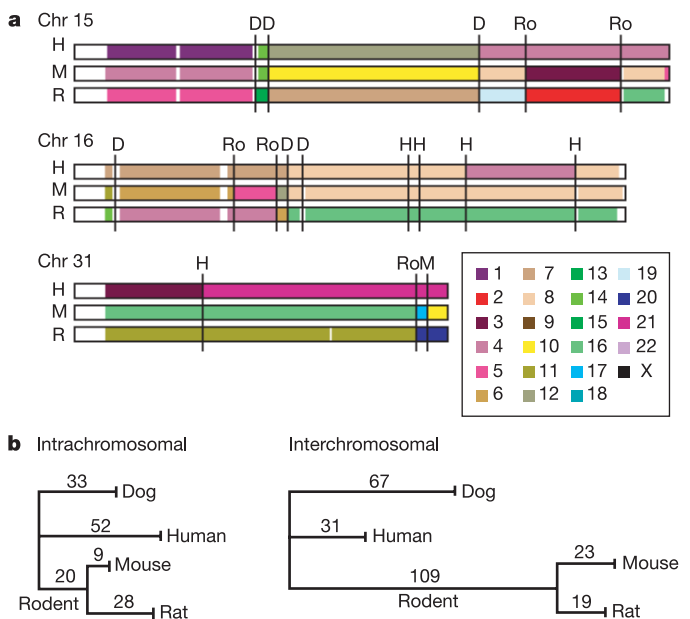


Figure 1 | Conserved synteny among the human, dog, mouse and rat genomes. **a**, Diagram of syntenic blocks (>500 kb) along dog chromosomes (Chr) 15, 16 and 31, with colours indicating the chromosome containing the syntenic region in other species. Syntenic breakpoints were assigned to one of five lineages: dog (D), human (H), mouse (M), rat (R) or the common rodent ancestor (Ro). **b**, Lineage-specific intrachromosomal and interchromosomal breaks displayed on phylogenetic trees. Intrachromosomal breaks are seen more frequently in the human lineage than in mouse and rat, whereas interchromosomal breaks are somewhat more common in dog and considerably more common in rodents than in humans.

sequence²⁴, most have occurred in the human lineage (see Supplementary Information). Human chromosome 17 is rich in segmental duplications and gene families²¹, which may contribute to its genomic fragility^{37,38}.

Genomic insertion and deletion. The euchromatic genome of the dog is ~150 Mb smaller than in mouse, and ~500 Mb smaller than in human. The smaller total size is reflected at the local level, with 100-kb blocks of conserved synteny in dog corresponding to regions for which the median size is ~3% larger in mouse and ~15% larger in human.

To understand the balance of forces that determine genome size, we studied the alignments of the human, mouse and dog genomes (Fig. 2). In particular, we identified the lineage-specific interspersed repeats within each genome, which consist of particular families of short interspersed elements (SINEs), long interspersed elements (LINEs) and other transposable elements that are readily recognized by sequence analysis (Supplementary Tables S9, S10). The remaining sequence was annotated as 'ancestral', consisting of both ancestral unique sequence and ancestral repeat sequence; these two categories were combined because the power to recognize ancient transposon-derived sequences degrades with repeat age, particularly in the rapidly diverging mouse lineage²⁴.

This comparative analysis indicates that different forces account for the smaller genome sizes in dog and mouse relative to human. The smaller size of the dog genome is primarily due to the presence of substantially less lineage-specific repeat sequence in dog (334 Mb) than in human (609 Mb) or mouse (954 Mb). This reflects a lower activity of endogenous retroviral and DNA transposons (~26,000 extant copies in dog versus ~183,000 in human), as well as the fact that the SINE element in dog is smaller than in human (although of similar length to that in mouse). As a consequence, the total proportion of repetitive elements (both lineage-specific and ancestral) recognizable in the genome is lower for dog (34%) than for mouse (40%) or human (46%). In contrast, the smaller size of the mouse genome is primarily due to a higher deletion rate. Specifically, the amount of extant 'ancestral sequence' is much lower in mouse (1,474 Mb) than in human (2,216 Mb) or dog (1,997 Mb). Assuming an ancestral genome size of 2.8 Gb (ref. 24) and also that deletions occur continuously, we suggest that the rate of genomic deletion in the rodent lineage has been approximately 2.5-fold higher than in the

dog and human lineages (see Supplementary Information). As a consequence, the human genome shares ~650 Mb more ancestral sequence with dog than with mouse, despite our more recent common ancestor with the latter.

Active SINE family. Despite its relatively low proportion of transposable element-derived sequence, the dog genome contains a highly active carnivore-specific SINE family (defined as SINEC_Cf; RepBase release 7.11)¹⁶. The element is so active that many insertion sites are still segregating polymorphisms that have not yet reached fixation. Of ~87,000 young SINEC_Cf elements (defined by low divergence from the consensus sequence), nearly 8% are heterozygous within the draft genome sequence of the boxer. Moreover, comparison of the boxer and standard poodle genome sequences reveals more than 10,000 insertion sites that are bimorphic, with thousands more certain to be segregating in the dog population^{16,39}. In contrast, the number of polymorphic SINE insertions in the human genome is estimated to be fewer than 1,000 (ref. 40).

The biological effect of these segregating SINE insertions is unknown. SINE insertions can be mutagenic through direct disruption of coding regions or through indirect effects on regulation and processing of messenger RNAs³⁹. Such SINE insertions have already been shown to be responsible for two diseases in dog: narcolepsy and centronuclear myopathy^{41,42}. It is conceivable that the genetic variation resulting from these segregating SINE elements has provided important raw material for the selective breeding programmes that have produced the wide phenotypic variations among modern dog breeds^{16,43}.

Sequence composition. The human and mouse genomes differ markedly in sequence composition, with the human genome having slightly lower average G+C content (41% versus 42% in mouse) but much greater variation across the genome. The dog genome closely resembles the human genome in its distribution of G+C content (Fig. 3a; Spearman's $\rho = 0.85$ for dog–human and 0.76 for dog–mouse comparisons), even if we consider only nucleotides that can be aligned across all three species (Supplementary Fig. S6). The wider distribution of G+C content in human and dog is thus likely to reflect the boreoeutherian ancestor^{44,45}, with the more homogeneous composition in rodents having arisen primarily through lineage-specific changes in substitution patterns^{46,47} rather than deletion of sequences with high G+C content.

Rate of nucleotide divergence. We estimated the mean nucleotide divergence rates in 1-Mb windows along the dog, human and mouse lineages on the basis of alignments of all ancestral repeats, using the consensus sequence for the repeats as a surrogate outgroup (Fig. 3b; see also Supplementary Information).

The dog lineage has diverged more rapidly than the human lineage (median relative divergence rate of 1.18, longer branch length in 95% of windows), but at only half the rate of the mouse lineage (median relative rate of 0.48, shorter branch length in 100% of windows). The absolute divergence rates are somewhat sensitive to the evolutionary model used and the filtering of alignment artefacts (data not shown), but the relative rates appear to be robust and are consistent with estimates from smaller sequence samples with multiple outgroups^{28,48,49}. The lineage-specific divergence rates (human < dog < mouse) are probably explained by differences in metabolic rates^{50,51} or generation times^{52,53}, but the relative contributions of these factors remain unclear⁴⁹.

Correlation in nucleotide divergence. As seen in other mammalian genomes^{23–25}, the average nucleotide divergence rate across 1-Mb windows varies significantly across the dog genome (coefficient of variation 0.11, compared with 0.024 expected under a uniform distribution). This regional variation shows significant correlation in orthologous windows across the dog, human and mouse genomes, but the strength of the correlation seems to decrease with total branch length (pair-wise correlation for orthologous 1-Mb windows: Spearman's $\rho = 0.49$ for dog–human and 0.24 for dog–mouse comparisons). Lineage-specific variation in the regional divergence

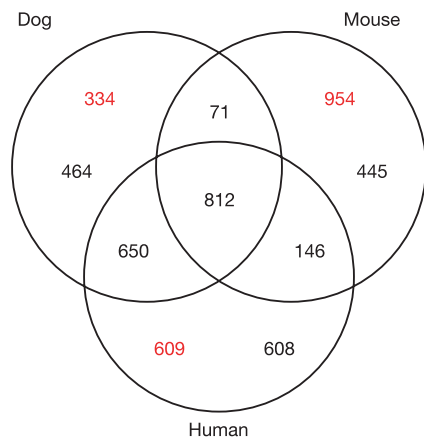


Figure 2 | Venn diagram showing the total lengths of aligned and unique sequences in the euchromatic portions of the dog, human and mouse genomes. Lengths shown in Mb, as inferred from genome-wide BLASTZ alignments (see Methods and Supplementary information). Overlapping partitions represent orthologous ancestral sequences. Each lineage-specific partition is further split into the total length of sequence classified as either lineage-specific interspersed repeats (red) or ancestral sequence (black). The latter is assumed to primarily represent ancestral sequences deleted in the two other species.

rates may be coupled with changes in factors such as sequence composition or chromosomal position^{23,54}. Consistent with this, the ratios of lineage-specific divergence rates in orthologous windows are positively correlated with the ratios of current G+C content in the same windows (Spearman's $\rho = 0.16$ for dog–human, 0.24 for dog–mouse).

Male mutation bias. Comparison of autosomal and X chromosome substitution rates can be used to estimate the relative mutation rates in the male and female germ lines (α), because the X chromosome is present in females twice as often as in males. Using the lineage-specific rates from ancestral repeats, we estimate α as 4.8 for the lineage leading to human, and 2.8 for the lineages leading to both mouse and dog. These values fall between recent estimates from murids^{24,25} and from hominids²³, and suggest that male mutation bias may have increased in the lineage leading to humans.

Mutational hotspots and chromosomal fission. Genome comparisons of human with both chicken⁵⁵ and chimpanzee²³ have previously revealed that sequences close to a telomere tend to have increased divergence rates and G + C content relative to interstitial sequences. It has been unclear whether these increases are inherent characteristics of the subtelomeric sequence itself or derived characteristics causally connected with its chromosomal position. We find a similar increase in both divergence (median increase 15%, $P < 10^{-5}$; Mann-Whitney U -test) and G+C content (median increase 9%, $P < 10^{-9}$) for subtelomeric regions along the dog lineage, with a sharp increase towards the telomeres (Supplementary Fig. S7).

This phenomenon is manifested at other synteny breaks, not only those at telomeres. We also observed a significant increase in divergence and G+C content in interstitial regions that are sites of syntenic breakpoints^{54,56} (Supplementary Fig. S7). These properties therefore seem correlated with the susceptibility of regions to chromosomal breakage.

Proportion of genome under purifying selection

One of the striking discoveries to emerge from the comparison of the human and mouse genomes^{21,24} was the inference that ~5.2% of the human genome shows greater-than-expected evolutionary conservation (compared with the background rate seen in ancestral repeat elements, which are presumed to be nonfunctional). This proportion greatly exceeds the 1–2% that can be explained by protein-coding regions alone. The extent and function of the large fraction of non-coding conserved sequence remain unclear⁵⁷, but this sequence is likely to include regulatory elements, structural elements and RNA genes.

Low turnover of conserved elements. We repeated the analysis of conserved elements using the human and dog genomes. Briefly, the

analysis involves calculating a conservation score S_{HD} , normalized by the regional divergence rate, for every 50-bp window in the human genome that can be aligned to dog. The distribution of conservation scores for all genomic sequences is compared to the distribution in ancestral repeat sequences (which are presumed to diverge at the local neutral rate), showing a clear excess of sequences with high conservation scores. By subtracting a scaled neutral distribution from the total distribution, one can estimate the distribution of conservation scores for sequences under purifying selection. Moreover, for a given sequence with conservation score S_{HD} , one can also assign a probability $P_{\text{selection}}(S_{HD})$ that the sequence is under purifying selection (see ref. 24 and Supplementary Information).

The human–dog genome comparison indicates that ~5.3% of the human genome is under purifying selection (Fig. 4a), which is equivalent to the proportion estimated from human–rodent analysis. The obvious question is whether the bases conserved between human and dog coincide with the bases conserved between humans and rodents^{25,58}. Because the conservation scores do not unambiguously assign sequences as either selected or neutral (but instead only assign probability scores for selection), we cannot directly compare the conserved bases. We therefore devised the following alternative approach.

We repeated the human–dog analysis, dividing the 1462 Mb of orthologous sequence between human and dog into those regions with (812 Mb) or without (650 Mb) orthologous sequence in mouse (Fig. 2). The first set shows a clear excess of conservation relative to background, corresponding to ~5.2% of the human genome (Fig. 4b). In contrast, the second set shows little or no excess conservation, corresponding to at most 0.1% of the human genome (Fig. 4c). This implies that hardly any of the functional elements conserved between human and dog have been deleted in the mouse lineage (see also Supplementary Information).

The results strongly suggest that there is a common set of functional elements across all three mammalian species, corresponding to ~5% of the human genome (~150 Mb). These functional elements reside largely within the 812 Mb of ancestral sequence common to human, mouse and dog. If we eliminate ancestral repeat elements within this shared sequence as largely non-functional, most functional elements can be localized to 634 Mb, and constitute approximately 24% of this sequence.

It should be noted that the estimate of ~5% pertains to conserved elements across distantly related mammals. It is possible that there are additional weakly constrained or recently evolved elements within narrow clades (for example, primates) that can only be detected by genomic sequencing of more closely related species²⁹.

Clustering of highly conserved non-coding elements. We next

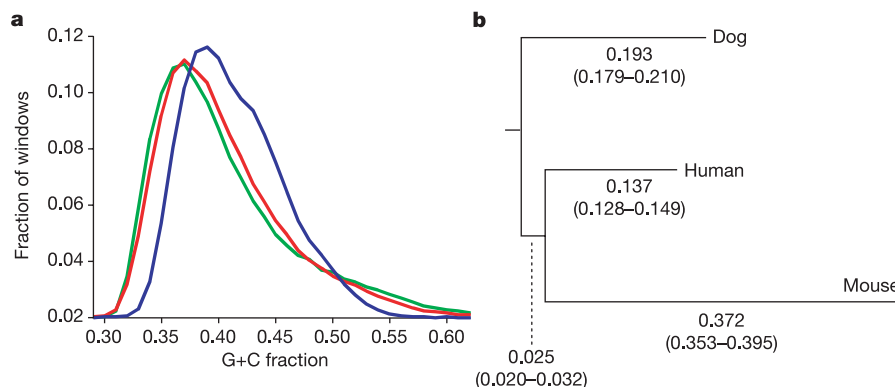


Figure 3 | Sequence composition and divergence rates. **a**, Distribution of G + C content in 10-kb windows across the genome in dog (green), human (red) and mouse (blue). **b**, Median lineage-specific substitution rates based on analysis of ancestral repeats aligning across all three genomes. Analysis was performed in non-overlapping 1-Mb windows across the dog genome

that contained at least 2 kb of aligned ancestral repeat sequence (median 8.8 kb). The tree was rooted with the consensus sequences from the ancestral repeats. Numbers in parentheses give the 20–80th percentile range across the windows studied.

explored the distribution of conserved non-coding elements (CNEs) across mammalian genomes. For this purpose, we calculated a conservation score S_{HMD} based on simultaneous conservation across all three species (see Methods). We defined highly conserved non-coding elements (HCNEs) to be 50-bp windows that do not overlap coding regions and for which $P_{\text{selection}}(S_{\text{HMD}})$, the probability of being under purifying selection given the conservation score, is at least 95%. We identified $\sim 140,000$ such windows (6.5 Mb total sequence), comprising $\sim 0.2\%$ of the human genome and representing the most conserved $\sim 5\%$ of all mammalian CNEs.

The density of HCNEs shows striking peaks when plotted in 1-Mb windows across the genome (Fig. 4d and Supplementary Figs S8 and S9), with 50% lying in 204 regions that span less than 14% of the human genome (Supplementary Table S11). These regions are generally gene-poor, together containing only $\sim 6\%$ of all protein-coding sequence.

The genes contained within these gene-poor regions are of particular interest. At least 182 of the 204 regions contain genes with key roles in establishing or maintaining cellular 'state'. At least 156 of the regions contain one or, in a few cases, several transcription factors involved in differentiation and development⁵⁹. Another 26 regions contain a gene important for neuronal specialization and growth, including several axon guidance receptors. The proportion of developmental regulators is far greater than expected by chance ($P < 10^{-31}$; see Supplementary Information).

We then tested whether the HCNEs within these regions tend to cluster around the genes encoding regulators of development. Analysis of the density of HCNEs in the intronic and intergenic sequences flanking every gene in the 204 regions revealed that the 197 genes encoding developmental regulators show an average of ~ 10 -fold enrichment for HCNEs relative to the full set of 1,285 genes

in the regions (Fig. 4e and Supplementary Fig. S10). The enrichment sometimes extends into the immediately flanking genes.

We note that the 204 regions include nearly all of the recently identified clusters of conserved elements between distantly related vertebrates such as chicken and pufferfish^{55,59–62}. For example, they overlap 56 of the 57 large intervals containing conserved non-coding sequence identified between human and chicken⁵⁵. The mammalian analysis, however, detects vastly more CNEs (>100 -fold more sequence than with pufferfish⁵⁹ and 2–3-fold more than with chicken) and identifies many more clusters. The limited sensitivity of these more distant vertebrate comparisons may reflect the difficulty of aligning short orthologous elements across such large evolutionary distances or the emergence of mammal-specific regulatory elements. In any case, mammalian comparative analysis may be a more powerful tool for elucidating the regulatory controls across these important regions.

Although the function of conserved non-coding elements is unknown, on the basis of recent studies^{59,63–66} it seems likely that many regulate gene expression. If so, the above results suggest that $\sim 50\%$ of all mammalian HCNEs may be devoted to regulating $\sim 1\%$ of all genes. In fact, the distribution may be even more skewed, as there are additional genomic regions with only slightly lower HCNE density than the 204 studied above (Supplementary Fig. S8). All of these regions clearly merit intensive investigation to assess indicators of regulatory function. We speculate that these regions may harbour characteristic chromatin structure and modifications that are potentially involved in the establishment or maintenance of cellular state.

Genes

Accurate identification of the protein-coding genes in mammalian genomes is essential for understanding the human genome, including its cellular components, regulatory controls and evolutionary

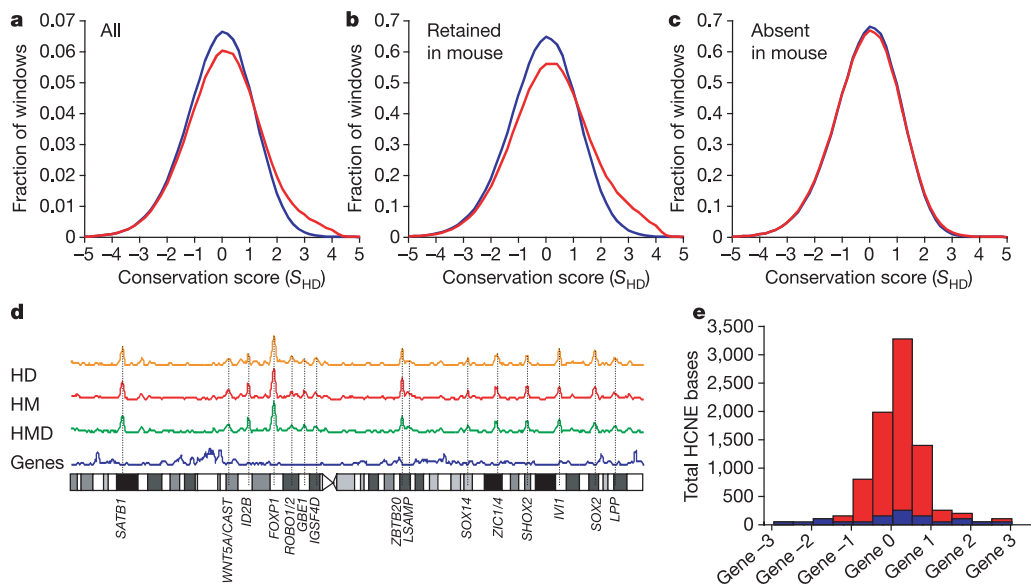


Figure 4 | Conservation of orthologous sequence between human and dog. **a**, Histogram of conservation scores, S , for all 50-bp windows across the human genome with at least 20 bases of orthologous sequence aligning to the dog genome, for all aligning sequences (red) and for ancestral sequence only (blue). **b**, Conservation scores for the subset of windows that also have at least 20 bases of orthologous sequence aligning to the mouse genome. **c**, Conservation scores of the complementary subset of windows lacking such orthologous sequence in mouse. **d**, Density of 50-bp windows not overlapping known coding regions, for which $P_{\text{selection}}(S) > 95\%$, based on comparisons between human and dog (HD), human and mouse (HM), or between human, mouse and dog (HMD), and the density of known genes, all in 1-Mb sliding windows across human chromosome 3. **e**, Enrichment

of HCNEs in the immediate neighbourhood of genes encoding developmental regulators in the 204 highly conserved regions. The histogram shows the median number of HCNE bases in the intronic and surrounding intergenic sequence, for the 197 known or putative development regulators (indicated by top of red bar) and for all of the 1,285 genes (blue bar). The histogram is centred at the 5'-end of the gene (marked 0) and each bin corresponds to half of the normalized distance to the flanking consecutive upstream genes (marked -1, -2 and -3) or consecutive downstream genes (1, 2 and 3) as indicated. The sequences surrounding the developmental genes are typically longer, have more HCNE sequence and have a higher density of HCNE sequence than other genes in the regions (see Supplementary Information).

constraints. The number of protein-coding genes in human has been a topic of considerable debate, with estimates steadily falling from ~100,000 to 20,000–25,000 over the past decade^{21,22,67–70}. We analysed the dog genome in order to refine the human gene catalogue and to assess the evolutionary forces shaping mammals. (In the Genes section, 'gene' refers only to a protein-coding gene.)

Gene predictions in dog and human. We generated gene predictions for the dog genome using an evidence-based method (see Supplementary Information). The resulting collection contains 19,300 dog gene predictions, with nearly all being clear homologues of known human genes.

The dog gene count is substantially lower than the ~22,000-gene models in the current human gene catalogue (Ensembl build 26). For many predicted human genes, we find no convincing evidence of a corresponding dog gene. Much of the excess in the human gene count is attributable to spurious gene predictions in the human genome (M. Clamp, personal communication).

Gene duplications. Gene duplication is thought to contribute substantially to functional innovation^{69,71}. We identified 216 gene duplications that are specific to the dog lineage and 574 that are specific to the human lineage, using the synonymous substitution rate K_S as a distance metric and taking care to discard likely pseudogenes. (The CanFam 2.0 assembly contains approximately 24 additional gene duplications, mostly olfactory receptors.) Human genes are thus 2.7-fold more likely to have undergone duplication than are dog genes over the same time period. This may reflect increased repeat-mediated segmental duplication in the human lineage⁷².

Although gene duplication has been less frequent in dog than human, the affected gene classes are very similar. Prominent among the lineage-specific duplicated genes are genes that function in adaptive immunity, innate immunity, chemosensation and reproduction, as has been seen for other mammalian genomes^{24,25,69,71}. Reproductive competition within the species and competition against parasites have thus been major driving forces in gene family expansion.

The two gene families with the largest numbers of dog-specific genes are the histone H2B family and the α -interferons, which cluster in monophyletic clades when compared to their human homologues. This is particularly notable for the α -interferons, for which the gene families within the six species (human, mouse, rat, dog, cat and horse) are apparently monophyletic. This may be due either to coincidental independent gene duplication in each of the six lineages or to ongoing gene conversion events that have homogenized ancestral gene duplicates⁷³.

Evolution of orthologous genes across three species. The dog genome sequence allows us for the first time to characterize the large-scale patterns of evolution in protein-coding genes across three major mammalian orders. We focused on a subset of 13,816 human, mouse and dog genes with 1:1:1 orthology. For each, we inferred the number of lineage-specific synonymous (K_S) and non-synonymous (K_A) substitutions along each lineage and calculated the K_A/K_S ratio (Table 2 and Supplementary Information), a traditional measure of the strength of selection (both purifying and directional) on proteins⁷⁴.

The median K_A/K_S ratio differs sharply across the three lineages ($P < 10^{-44}$, Mann-Whitney U -test), with the dog lineage falling

between mouse and human. Population genetic theory predicts⁷⁵ that the strength of purifying selection should increase with effective population size (N_e). The observed relationship (mouse < dog < human) is thus consistent with the evolutionary prediction, given the expectation that smaller mammals tend to have larger effective population sizes⁷⁶.

We next searched for particular classes of genes showing deviations from the expected rate of evolution for a species. Such variation in rate (heterotachy) may point to lineage-specific positive selection or relaxation of evolutionary constraints⁷⁷. We developed a statistical method similar to the recently described Gene Set Enrichment Analysis (GSEA)^{78–80} to detect evidence of heterotachy for sets of functionally related genes (see Supplementary Information). Briefly, the approach involves ranking all genes by K_A/K_S ratio, testing whether the set is randomly distributed along the list and assessing the significance of the observed deviations by comparison with randomly permuted gene sets. In contrast to previous studies, which focused on small numbers of genes with prior hypotheses of selection, this approach detects signals of lineage-specific evolution in a relatively unbiased manner and can provide context to the results of more limited studies.

A total of 4,950 overlapping gene sets were studied, defined by such criteria as biological function, cellular location or co-expression (see Supplementary Information). Overall, the deviations between the three lineages are small, and median K_A/K_S ratios for particular gene sets are highly correlated for each pair of species (Supplementary Fig. S11). However, there is greater relative variation in human–mouse and dog–mouse comparisons than in human–dog comparisons (Supplementary Fig. S12).

This suggests that observed heterotachy between human and mouse must be interpreted with caution. For example, there is a great interest in the identification of genetic changes underlying the unique evolution of the human brain. A recent study⁸¹ highlighted 24 genes involved in brain development and physiology that show signs of accelerated evolution in the lineage leading from ancestral primates to humans when compared to their rodent orthologues. We observe the same trend for the 18 human genes that overlap with the genes studied here, but find at least as many genes with higher relative acceleration in the dog lineage (see Supplementary Information). Heterotachy relative to mouse therefore does not appear to be a distinctive feature of the human lineage. It may reflect decelerated evolution in the rodent lineage, or possibly independent adaptive evolution in the human and dog lineages⁸².

A small number of gene sets show evidence of significantly accelerated evolution in the human lineage, relative to both mouse and dog (32 sets at $z \geq 5.0$ versus zero sets expected by chance, $P < 10^{-4}$; Fig. 5a). These sets fall into two categories: genes expressed exclusively in testis, and (nuclear) genes encoding subunits of the mitochondrial electron transport chain (ETC) complexes. The former are believed to undergo rapid evolution as a consequence of sperm competition across a wide range of species^{83–85}, and lineage-specific acceleration suggests that sexual selection may have been a particularly strong force in primate evolution. The selective forces acting on the latter category are less obvious. Because of the importance of mitochondrial ATP generation for sperm motility⁸⁶, and the potentially antagonistic co-evolution of these genes with maternally inherited mitochondrial DNA-encoded subunits⁸⁷, we

Table 2 | Evolutionary rates for 1:1:1 orthologues among dog, mouse and human

	Median (20–80th percentile range)			Spearman's ρ		
	Dog*	Mouse	Human	Dog-human	Dog-mouse	Human-mouse
K_S	0.210 (0.138–0.322)	0.416 (0.310–0.558)	0.139 (0.0928–0.214)	0.47	0.50	0.52
K_A	0.021 (0.006–0.051)	0.038 (0.013–0.087)	0.017 (0.005–0.040)	0.87	0.87	0.86
K_A/K_S	0.095 (0.030–0.221)	0.088 (0.031–0.197)	0.112 (0.034–0.272)	0.80	0.85	0.82

*Estimates are based on unrooted tree. The dog branch thus includes the branch from the boreoeutherian ancestor to the primate–rodent split.

propose that sexual selection may also be the primary force behind the rapid evolution of the primate ETC genes. Given the ubiquitous role of mitochondrial function, however, such sexual selection may have led to profound secondary effects on physiology⁸⁸.

We found no gene sets with comparably strong evidence for dog-specific accelerated evolution. There is, however, a small excess of sets with moderately high acceleration scores (19 sets at $z \geq 3.0$ versus 5 sets expected by chance, $P < 0.02$; Fig. 5b). These sets, which are primarily related to metabolism, may contain promising candidates for follow-up studies of molecular adaptation in carnivores.

Polymorphism and haplotype structure in the domestic dog

The modern dog has a distinct population structure with hundreds of genetically isolated breeds, widely varying disease incidence and distinctive morphological and behavioural traits^{89,90}. Unlocking the full potential of the dog genome for genetic analysis requires a dense SNP map and an understanding of the structure of genetic variation both within and among breeds.

Generating a SNP map. We generated a SNP map of the dog genome containing >2.5 million distinct SNPs mapped to the draft genome sequence, corresponding to an average density of approximately one SNP per kb (Table 3). The SNPs were discovered in three complementary ways (see Supplementary Information). (1) We identified SNPs within the sequenced boxer genome (set 1; $\sim 770,000$ SNPs) by searching for sites at which alternative alleles are supported by at least two independent reads each. We tested a subset ($n = 40$ SNPs) by genotyping and confirmed all as heterozygous sites. (2) We compared the 1.5 \times sequence from the standard poodle¹⁶ with the draft genome sequence from the boxer (set 2; $\sim 1,460,000$ SNPs). (3) We generated shotgun sequence data from nine diverse dog breeds ($\sim 100,000$ reads each, 0.02 \times coverage), four grey wolves and one coyote ($\sim 22,000$ reads each, 0.004 \times coverage) and compared it to the boxer (set 3; $\sim 440,000$ SNPs). We tested a subset ($n = 1,283$ SNPs) by genotyping and confirmed 96% as true polymorphisms.

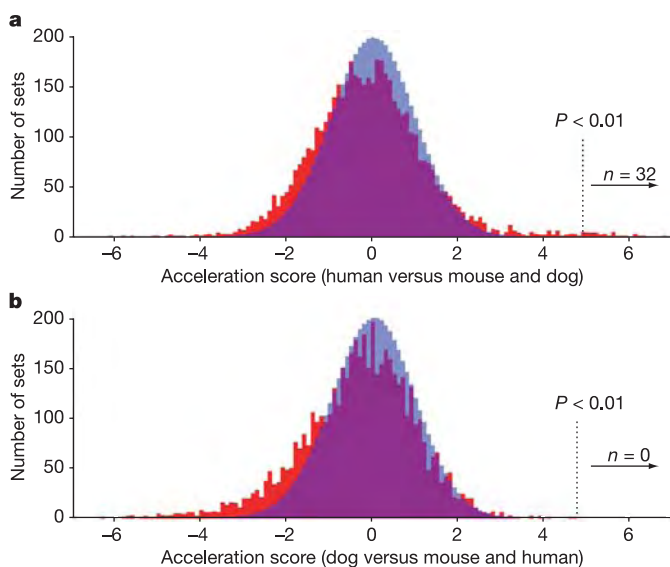


Figure 5 | Gene sets showing accelerated evolution along the human and dog lineages. **a**, Distribution of acceleration scores along the human lineage relative to both mouse and dog, observed for 4,950 gene sets (red). The expected distribution based on 10,000 randomized trials is shown in blue. The dotted line shows the acceleration score for which the probability of observing even a single set by random chance (out of the 4,950 sets tested) is less than 1%. In fact, 32 sets show acceleration scores on the human lineage exceeding this threshold. **b**, The observed (red) and expected (blue) distribution of acceleration scores for the dog lineage, relative to both human and mouse.

Table 3 | SNPs discovered in dogs, wolves and coyotes compared to the boxer assembly

Set number	Breed or species	Number of SNPs	SNP rate (one per x bases)
1	Boxer versus boxer	768,948	3,004 (observed) 1,637 (corrected)
2	Boxer versus poodle	1,455,007	894
3a	Boxer versus breeds*		
	German shepherd	45,271	900
	Rottweiler	44,097	917
	Bedlington terrier	44,168	913
	Beagle	42,572	903
	Labrador retriever	40,730	926
	English shepherd	40,935	907
	Italian greyhound	39,390	954
	Alaskan malamute	45,103	787
	Portuguese water dog	45,457	896
	Total distinct SNPs	373,382	900
3b	Boxer versus Canids†		
	China grey wolf	12,182	580
	Alaska grey wolf	13,888	572
	India grey wolf	14,510	573
	Spanish grey wolf	10,349	587
	California coyote	20,270	417
	Total distinct SNPs	71,381	
3	Set 3 total distinct SNPs	441,441	
Total	Total distinct SNPs	2,559,519	

*Based on $\sim 100,000$ sequence reads per breed.

†Based on $\sim 20,000$ sequence reads per wolf.

The SNP rate between the boxer and any of the different breeds is one SNP per ~ 900 bp, with little variation among breeds (Table 3). The only outlier ($\sim 1/790$ bp) is the Alaskan malamute, which is the only breed studied that belongs to the Asian breed cluster⁹¹. The grey wolf ($\sim 1/580$ bp) and coyote ($\sim 1/420$ bp) show greater variation when compared with the boxer, supporting previous evidence of a bottleneck during dog domestication, whereas that the SNP rate is lower in the grey wolf than in the coyote reflects the closer relationship of the grey wolf to the domestic dog^{1-3,92} (see section 'Resolving canid phylogeny').

The observed SNP rate within the sequenced boxer assembly is $\sim 1/3,000$ bp. This underestimates the true heterozygosity owing to the conservative criterion used for identifying SNPs within the boxer assembly (requiring two reads containing each allele); correcting for this leads to an estimate of $\sim 1/1,600$ bp (see Supplementary Information). This low rate reflects reduced polymorphism within a breed, compared with the greater variation of $\sim 1/900$ bp between breeds.

To assess the utility of the SNPs for dog genetics, we genotyped a subset from set 3a ($n = 1,283$) in 20 dogs from each of ten breeds (Supplementary Table S16). Within a typical breed, $\sim 73\%$ of the SNPs were polymorphic. The polymorphic SNPs have minor allele frequencies that are approximately evenly distributed between 5% and 50% (allele frequencies less than 5% are not reliable with only 40 chromosomes sampled). In addition, the SNPs from sets 2 and 3 have a roughly uniform distribution across the genome (Fig. 6a, see below concerning set 1). The SNP map thus has high density, even distribution and high cross-breed polymorphism, indicating that it should be valuable for genetic studies.

Expectations for linkage disequilibrium and haplotype structure.

Modern dog breeds are the product of at least two population bottlenecks, the first associated with domestication from wolves ($\sim 7,000$ – $50,000$ generations ago) and the second resulting from intensive selection to create the breed (~ 50 – 100 generations ago). This population history should leave distinctive signatures on the patterns of genetic variation both within and across breeds. We might expect aspects of both the long-range LD seen in inbred mouse strains, with strain-specific haplotypes extending over multiple megabases, and the short-range LD seen in humans, with ancestral haplotype blocks typically extending over tens of kilobases. Specifically,

long-range LD would be expected within dog breeds and short-range LD across breeds.

Preliminary evidence of long-range LD within breeds has been reported⁹⁰. Five genome regions were examined (~1% of the genome) in five breeds using ~200 SNPs with high minor allele frequency. LD seemed to extend 10–100-fold further in dog than in human, with relatively few haplotypes per breed.

With the availability of a genome sequence and a SNP map, we sought to undertake a systematic analysis of LD and haplotype structure in the dog genome.

Haplotype structure within the boxer assembly. We first analysed the structure of genetic variation within the sequenced boxer genome by examining the distribution of the ~770,000 SNPs detected between homologous chromosomes. Strikingly, the genome is a mosaic of long, alternating regions of near-total homozygosity and high heterozygosity (Fig. 6b, c), with observed SNP rates of ~14 per Mb and ~850 per Mb, respectively. (The latter is close to that seen within breeds and is indistinguishable when one corrects for the conservative criterion used to identify SNPs within the boxer assembly; see Supplementary Information.) The homozygous regions have an N50 size of 6.9 Mb and cover 62% of the genome, and the heterozygous regions have an N50 size of 1.1 Mb and cover

38% of the genome. The results imply that the boxer genome is largely comprised of vast haplotype blocks. The long stretches of homozygosity indicate regions in which the sequenced boxer genome carries the same haplotype on both chromosomes. The proportion of homozygosity (~62%) reflects the limited haplotype diversity within breeds.

Long-range haplotypes in different breeds. We sought to determine whether the striking haplotype structure seen in the boxer genome is representative of most dog breeds. To this end, we randomly selected ten regions of 15 Mb each (~6% of the genome) and examined linkage disequilibrium in these regions in a collection of 224 dogs, consisting of 20 dogs from each of ten breeds and one dog from each of 24 additional breeds (see Supplementary Tables S17–S19).

The ten breeds were chosen to represent all four clusters described in ref. 91. The selected breeds have diverse histories, with varying population size and bottleneck severity. For example, the Basenji is an ancient breed from Africa that has a small breeding population in the United States descending from dogs imported in the 1930s–1940s (refs 93, 94). The Irish wolfhound suffered a severe bottleneck two centuries ago, with most dogs today being descendents of a single dog in the early 1800s (refs 5, 94). In contrast, the Labrador retriever and golden retriever have long been, and remain, extremely popular dogs

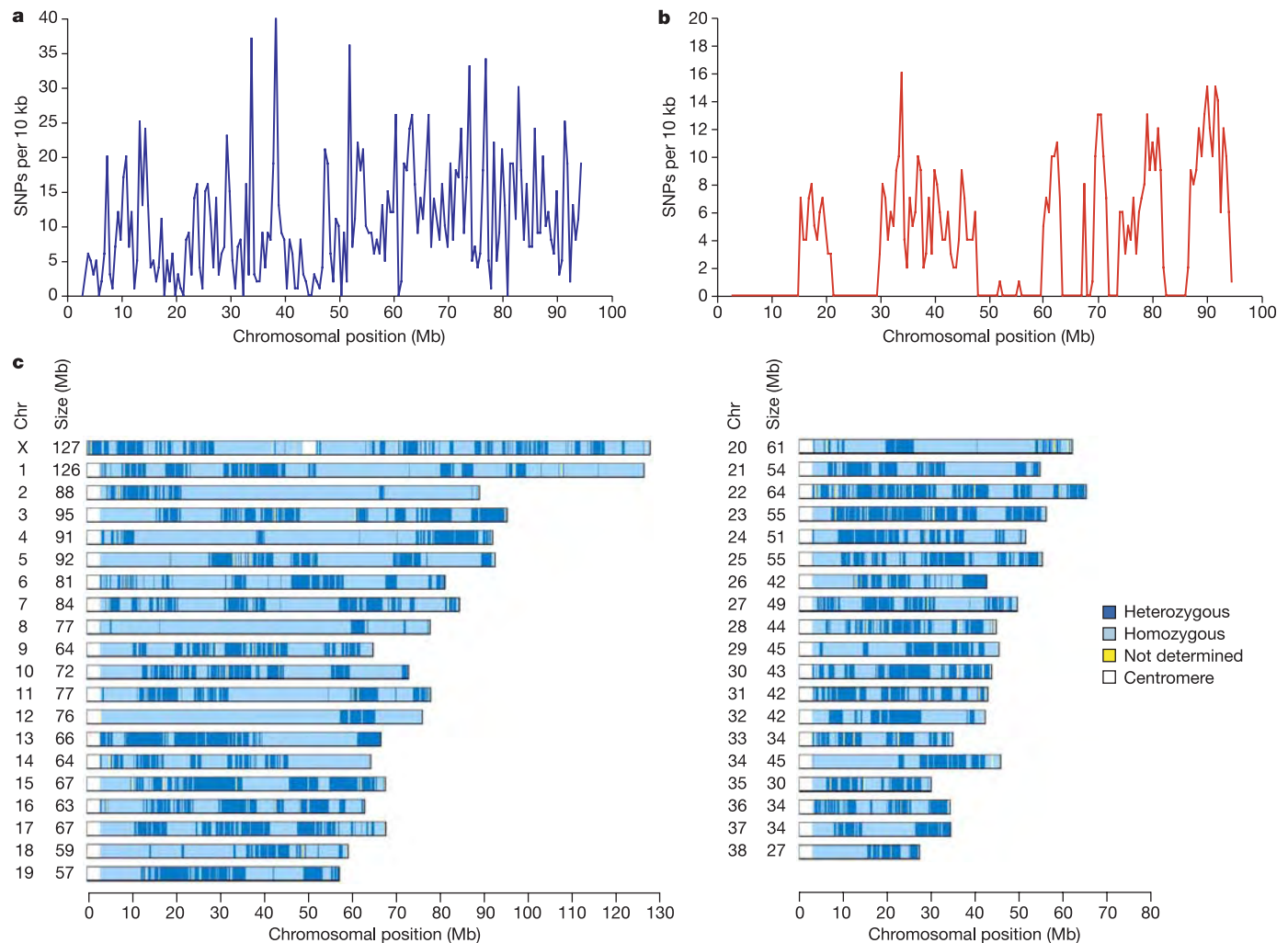


Figure 6 | The distribution of SNPs is fairly uniform across breeds, but non-uniform within the sequenced boxer assembly. **a**, SNPs across chromosome 3, generated by comparing the boxer assembly with WGS reads from nine breeds. **b**, The SNPs on chromosome 3 of the boxer assembly show an uneven distribution (plotted in 500-kb windows). Note that boxer SNPs were identified using a more conservative method, lowering the observed

SNP rate by roughly twofold. **c**, An alternating pattern of large homozygous (light blue, ~62% of genome; N50 size 6.9 Mb) and large heterozygous (dark blue ~38% of genome; N50 size 1.1 Mb) blocks indicates large identical or divergent haplotypes across the boxer genome. White indicates centromeric sequence.

(with ~150,000 and ~50,000 new puppies registered annually, respectively). They have not undergone such recent severe bottlenecks, but some lines have lost diversity because of the repeated use of popular sires⁸⁹. The Glen of Imaal terrier represents the opposite end of the popularity spectrum, with fewer than 100 new puppies registered with the American kennel Club each year.

The 224 dogs were genotyped for SNPs across each of the ten regions, providing 2,240 cases in which to assess long-range LD. The SNPs ($n = 1,219$; Supplementary Table S19) were distributed along the regions to measure the fall-off of genetic correlation, with higher density at the start of the region and lower densities at further distances (Fig. 7a). In 645 cases, we also examined the first 10 kb in

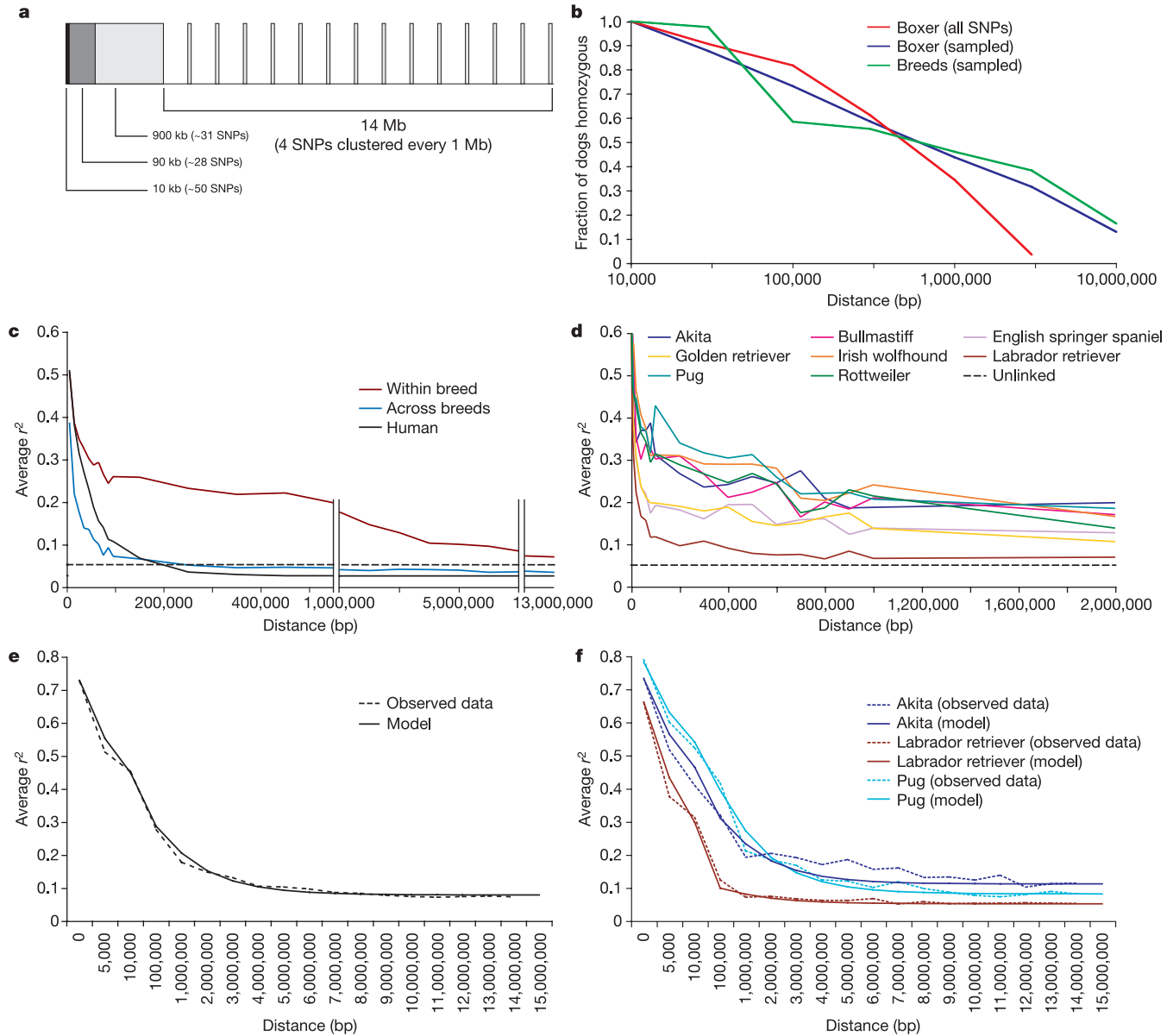


Figure 7 | Homozygous regions and linkage disequilibrium are nearly 100-fold longer within dog breeds than across the dog or human populations. **a**, Sampling design for ten random regions of 15 Mb each, used to assess the haplotype structure of ~6% of the genome (see Supplementary Information). For each region, we examined the first 10 kb through resequencing and dense genotyping. To detect long haplotypes, we genotyped SNPs distributed throughout the next 1 Mb and sampled SNPs at intervals of 1 Mb for the next 14 Mb. In total we genotyped 1,219 SNPs across the ten regions in a collection of 224 dogs (20 dogs from each of 10 breeds and one dog from each of 24 breeds). **b**, Conditional on a dog being homozygous for the initial 10-kb region ($n = 245$), we assessed the probability that the dog was homozygous for all SNPs within a given distance. The average proportion remaining homozygous is compared for the various breeds (green), for the boxer when sampled in the same ways as the breeds (blue) and for the boxer using all SNPs found in the genome sequence (red). About 50% of the individuals seem to be homozygous throughout 1 Mb both in the boxer and other breeds, indicating that other

breeds have comparable long-range homozygosity. **c**, Linkage disequilibrium (LD) as a function of distance is shown as the r^2 statistic within individual breeds (red), across various breeds (blue) and a human population (black) taken from the CEPH collection genotyped as part of the ENCODE component of the International HapMap Project¹¹⁸. For the overall dog and human populations, LD falls rapidly, reaching the baseline level seen for unlinked loci by ~200 kb. In contrast, LD for individual breeds falls initially but then stays at a moderately high level across several megabases. **d**, The LD curves are broadly similar for most breeds, but the proportion of long-range LD is correlated with known breed history. **e**, The observed within-breed LD curve (averaged across breeds) is well fitted by a simple model with a domestication bottleneck 10,500 generations ago and a breed-creation bottleneck occurring 50 generations ago (see Supplementary Information). **f**, LD curves for individual dog breeds can be fitted by models with different breed-creation bottlenecks. The poorest fit is obtained for the akita, the breeding history of which is known to involve two separate breed-creation bottlenecks.

greater detail by denser genotyping (with ~ 2 SNPs per kb) in 405 cases and complete resequencing in 240 cases. The resequencing data yielded a heterozygosity rate of ~ 1 SNP per 1,500 bp, essentially equivalent to the rate seen in the sequenced boxer genome.

On the basis of examining the first 10 kb, we found that $\sim 38\%$ of instances seem to be completely homozygous and that all dogs seem to be homozygous for at least one of the ten regions. We then measured the distance over which homozygosity persisted. Of instances homozygous in the initial 10-kb segment, 46% were homozygous across 1 Mb and 17% were still homozygous across 10 Mb (Fig. 7b). The fall-off in homozygosity is essentially identical to that seen in the boxer genome, provided that the boxer data are sampled in an equivalent manner (see Supplementary Information). This indicates that the long-range haplotype structure seen in the boxer is typical of most dog breeds, although the precise haplotypes vary with breed and the locations of homozygous regions vary between individuals.

We also assessed long-range correlations by calculating r^2 , a traditional measure of LD, across the 15-Mb regions. The r^2 curve representing the overall dog population (one dog from each of 24 breeds) drops rapidly to background levels. This is in sharp contrast to the r^2 curves within each breed. Within breeds, LD is biphasic, showing a sharp initial drop within ~ 90 kb followed by an extended shoulder that gradually declines to the background (unlinked) level by 5–15 Mb in most breeds (Fig. 7c). The basic pattern is similar in all ten regions (Supplementary Fig. S13) and in all breeds (Fig. 7d). (Labrador retrievers show the shortest LD, probably due to their mixed aetiology and large population size.)

The biphasic r^2 curves within each breed thus consist of two components (Fig. 7e), at scales differing by ~ 100 -fold. The first component matches the fall-off in the general dog population and is likely to represent the short-range de-correlation of local haplotype blocks in the ancestral dog population. The second component represents long-range breed-specific haplotypes (Fig. 8a). Notably, the first component falls off nearly twice as quickly as the LD in the human population (~ 200 kb), and the second component falls off slightly slower than seen in laboratory mouse strains⁹⁵.

Modelling the effects of population history. We tested this interpretation by performing mathematical simulations on a dog population that underwent an ancient bottleneck and recent breed-creation bottlenecks, using the coalescent approach⁹⁶ (see Supplementary Information). Our experimental results were well fitted by models assuming an ancient bottleneck (effective domesticated population size 13,000, inbreeding coefficient $F = 0.12$) occurring $\sim 9,000$ generations ago (corresponding to $\sim 27,000$ years) and subsequent breed-creation bottlenecks of varying intensities occurring 30–90 generations ago⁹⁷ (Supplementary Fig. S14). The model closely reproduces the observed r^2 curves and the observed polymorphism rates within breeds, among breeds and between dog and grey wolf. The model also yields estimates of breed-specific bottlenecks that are broadly consistent with known breed histories. For example, Labrador retrievers, and to a lesser extent golden retrievers and English springer spaniels, show less severe bottlenecks.

Deterministically modelled results (Fig. 7e, f) indicate that a simple, two-bottleneck model provides a close fit to the data for the breeds. They do not rule out a more complex population history, such as multiple domestication events, low levels of continuing gene flow between domestic dog and grey wolf^{97,98} or multiple bottlenecks within breeds. Notably, the akita yields the poorest fit to the model, with an r^2 curve that appears to be triphasic. This may reflect the initial creation of the breed as a hunting dog in Japan ~ 450 generations ago, and a consecutive bottleneck associated with its introduction into the United States during the 1940s (ref. 99).

Haplotype diversity. We next studied haplotype diversity within and among breeds, using the dense genotypes from the 10-kb regions. Across the 645 cases examined, there is an average of ~ 10 distinct haplotypes per region. Within a breed, we typically see four of

these haplotypes, with the average frequency of the most common haplotype being 55% and the average frequency of the two most common being 80% (Fig. 8c and Supplementary Fig. S18). The haplotypes and their frequencies differ sharply across breeds. Nonetheless, 80% of the haplotypes seen with a frequency of at least 5% in one breed are found in other breeds as well (Supplementary Table S26). This extends previous observations of haplotype sharing across breeds⁹⁰. In particular, the inclusion of all SNPs with a minor allele frequency $\geq 5\%$ across all breeds provides a more accurate picture of haplotype sharing, because the analysis includes haplotypes that are rare within a single breed but more common across the population.

We then inferred the ancestral haplotype block structure in the ancestral dog population (before the creation of modern breeds) by combining the data across breeds and applying methods similar to those used for haplotype analysis in the human genome¹⁰⁰ (see Supplementary Information). In the 10-kb regions studied, one or two haplotype blocks were typically observed. Additional data across 100-kb regions suggest that the ancestral blocks have an average size of ~ 10 kb. The blocks typically have ~ 4 – 5 distinct haplotypes across the entire dog population (Fig. 8b). The overall situation closely resembles the structure for the human genome, although with slightly smaller block size (Supplementary Figs S15–S19 and Supplementary Table S24–26).

Ancestral and breed-specific haplotypes. A clear picture of the population genetic history of dogs emerges from the results detailed above:

- The ancestral dog population had short-range LD. The haplotype blocks were somewhat shorter than in modern humans (~ 10 kb versus ~ 20 kb in human), consistent with the dog population being somewhat older than the human population ($\sim 9,000$ generations versus $\sim 4,000$ generations). Haplotype blocks at large distances were essentially uncorrelated (Fig. 8a).
- Breed creation introduced tight breed-specific bottlenecks, at least for the breeds examined. From the great diversity of long-range haplotype combinations carried in the ancestral population, the founding chromosomes emerging from the bottleneck represented only a small subset. These became long-range breed-specific haplotypes (Fig. 8a).
- Although the breed-specific bottlenecks were tight, they did not cause massive random fixation of individual haplotypes. Only 13% of the small ancestral haplotypes are monomorphic within a typical breed, consistent with the estimated inbreeding coefficient of $\sim 12\%$. Across larger regions (≥ 100 kb), we observed no cases of complete fixation within a breed (Supplementary Fig. S20).
- There is notable sharing of 100-kb haplotypes across breeds, with $\sim 60\%$ seen in multiple breeds although with different frequencies. On average, the probability of sampling the same haplotype on two chromosomes chosen from different breeds is roughly twofold lower than for chromosomes chosen within a single breed (Supplementary Fig. S21).

Implications for genetic mapping. These results have important implications for the design of dog genetic studies. Although early efforts focused on cross-breeding of dogs for linkage analysis^{101–103}, it is now clear that within-breed association studies offer specific advantages in the study of both monogenic and polygenic diseases. First, they use existing dogs coming to medical attention and do not require the sampling of families with large numbers of affected individuals. Such studies should be highly informative, because dog breeds have retained substantial genetic diversity. Moreover, they will require a much lower density of SNPs than comparable human association studies, because the long-range LD within breeds extends ~ 50 -fold further than in humans^{90,104,105}.

Whereas human association studies require $>300,000$ evenly spaced SNPs^{100,106,107}, the fact that LD extends over at least 50-fold greater distances in dog suggests that dog association studies would require perhaps $\sim 10,000$ evenly spaced SNPs. To estimate the

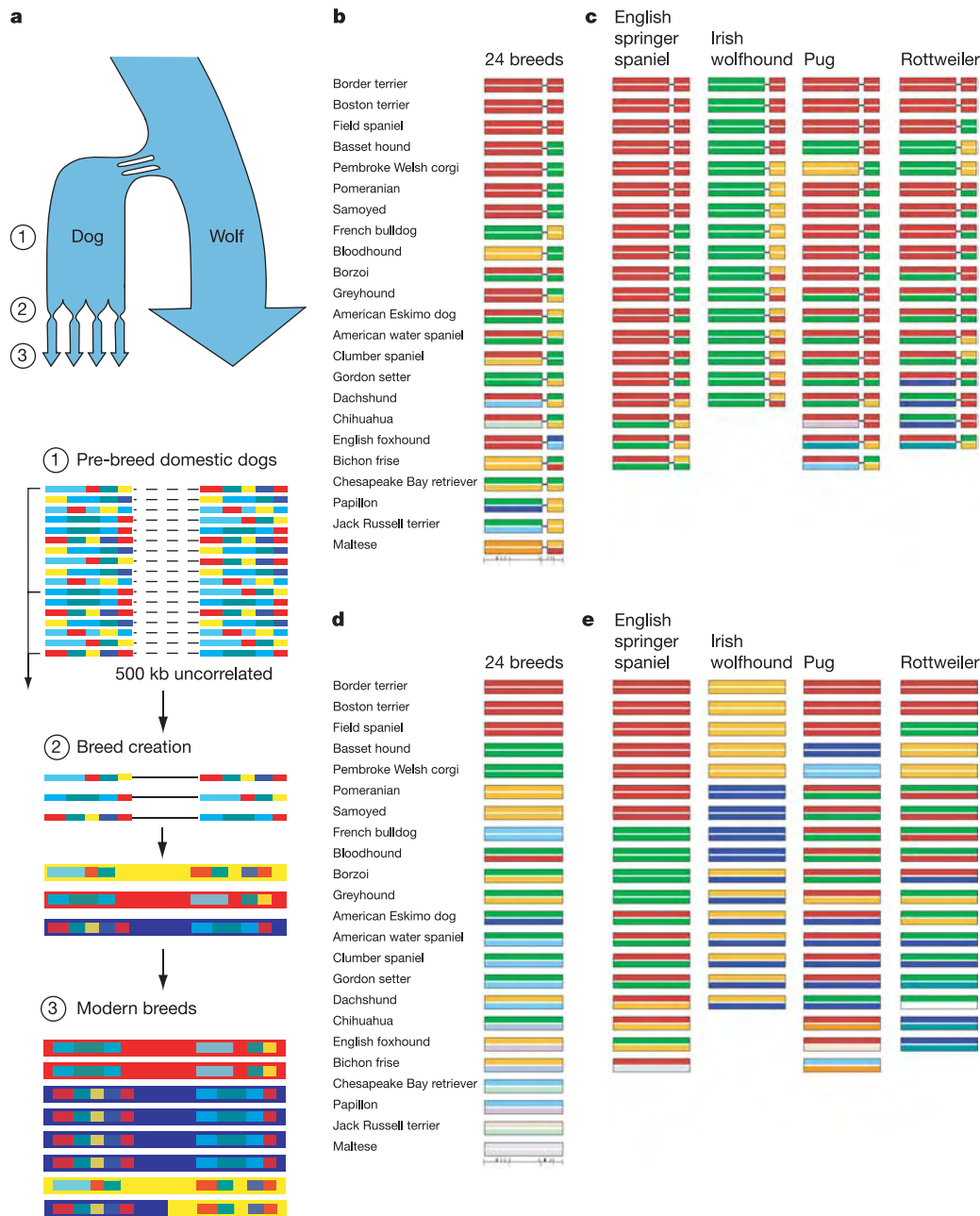


Figure 8 | Two bottlenecks, one old and one recent, have shaped the haplotype structure and linkage disequilibrium of canine breeds.

a, Modern haplotype structure arose from key events in dog breeding history. The domestic dog diverged from wolves 15,000–100,000 years ago^{97,119}, probably through multiple domestication events⁹⁸. Recent dog breeds have been created within the past few hundred years. Both bottlenecks have influenced the haplotype pattern and LD of current breeds. (1) Before the creation of modern breeds, the dog population had the short-range LD expected on the basis of its large size and time since the domestication bottleneck. (2) In the creation of modern breeds, a small subset of chromosomes was selected from the pool of domestic dogs. The long-range patterns that happened to be carried on these chromosomes became common within the breed, thereby creating long-range LD. (3) In the short time since breed creation, these long-range patterns have not yet been substantially broken down by recombination. Long-range haplotypes, however, still retain the underlying short-range ancestral haplotype blocks from the domestic dog population, and these are revealed when one examines chromosomes across many breeds. **b, c**, Distribution of ancestral haplotype blocks in a 10-kb window on chromosome 6 at ~31.4 Mb across

24 breeds (**b**) and within four breeds (**c**). Ancestral haplotype blocks are 5–15 kb in size (which is shorter than the ~25-kb blocks seen in humans) and are shared across breeds. Typical blocks show a spectrum of ~5 haplotypes, with one common major haplotype. Blocks were defined using the modified four-gamete rule (see Supplementary Information) and each haplotype (minor allele frequency (maf) > 3%) within a block was given a unique colour. **d, e**, Distribution of breed-derived haplotypes across a 10-kb window on chromosome 6 at ~31.4 Mb across 24 breeds (**d**) and within four breeds (**e**). Each colour denotes a distinct haplotype (maf > 3%) across 11 SNPs in the 10-kb window for each of the analysed dogs. Pairs of haplotypes have an average of 3.7 differences. Most haplotypes can be definitively identified on the basis of homozygosity within individual dogs. Grey denotes haplotypes that cannot be unambiguously phased owing to rare alleles or missing data. Within each of the four breeds shown, there are 2–5 haplotypes, with one or two major haplotypes accounting for the majority of the chromosomes. Across the 24 breeds, there are a total of seven haplotypes. All but three are seen in multiple breeds, although at varying frequencies.

number of SNPs required, we generated SNP sets from ten 1-Mb regions by coalescent simulations using the bottleneck parameters that generate SNP rates and LD curves equivalent to the actual data (Supplementary Fig. S14 and Supplementary Table S20). We then selected individual SNPs as ‘disease alleles’ and tested our ability to map them by association analysis with various marker densities (Fig. 9a).

For disease alleles causing a simple mendelian dominant trait with high penetrance and no phenocopies, there is overwhelming power to map the locus (Fig. 9a). Using ~15,000 evenly spaced SNPs and a log likelihood odds ratio (LOD score) score threshold of 5, the probability of detecting the locus is over 99% given a collection of 100 affected and 100 unaffected dogs. (The LOD score threshold corresponds to a false positive rate of 3% loci per genome.)

For a multigenic trait, the power to detect disease alleles depends on several factors, including the relative risk conferred by the allele, the allele frequency and the interaction with other alleles. We investigated a simple model of an allele that increases risk by a multiplicative factor (λ) of 2 or 5 (see Supplementary Information). Using the above SNP density and LOD score threshold, the power to detect a locus with a sample of 100 affected and 100 unaffected dogs is 97% for $\lambda = 5$ and 50% for $\lambda = 2$ (Fig. 9b, c). Although initial mapping will be best done by association within breeds, subsequent fine-structure mapping to pinpoint the disease gene will probably benefit from cross-breed comparison. Given the genetic relationships across breeds described above, it is likely that the same risk allele will be carried in multiple breeds. By comparing risk-associated haplotypes in multiple breeds, it should be possible to substantially narrow the region containing the gene.

Resolving canid phylogeny

The dog family, Canidae, contains 34 closely related species that diverged within the last ~10 million years¹. Resolving the evolutionary relationships of such closely related taxa has been difficult because a great quantity of genomic sequence is typically required to yield enough informative nucleotide sites for the unambiguous reconstruction of phylogenetic trees. We sought to streamline the process of evolutionary reconstruction by exploiting our knowledge of the dog genome to select genomic regions that would maximize the amount of phylogenetic signal per sequenced base. Specifically, we sought regions of rapidly evolving, unique sequence.

We first compared the coding regions of 13,816 dog genes with human–dog–mouse 1:1:1 orthologues to find those with high neutral evolutionary divergence (comparing K_S and K_A/K_S). We selected 12 exons (8,080 bp) for sequencing, based on the criteria that their sequences (1) are consistent with the known phylogeny of human,

dog, mouse and rat, (2) have a high percentage of bases ($\geq 15\%$) that are informative for phylogenetic reconstruction in the human, dog, mouse and rat phylogenies, and (3) could be successfully amplified in all canids. The chosen exons contain 3.3-fold more substitutions than random exonic sequence. Using our SNP database, we also evaluated introns to identify those with high variation between dog and coyote. We selected four introns (3,029 bp) that contained ~5-fold more SNPs than the background frequency. We sequenced these exons and introns (11,109 bp) in 30 out of 34 living wild canids, and we combined the data with additional sequences (3,839 bp) from recent studies^{3,92}.

The resulting evolutionary tree has a high degree of statistical support (Fig. 10), and uniquely resolves the topology of the dog’s closest relatives. Grey wolf and dog are most closely related (0.04% and 0.21% sequence divergence in nuclear exon and intron sequences, respectively), followed by a close affiliation with coyote, golden jackal and Ethiopian wolf, three species that can hybridize with dogs in the wild (Fig. 10). Closest to this group are the dhole and African wild dog, two species with a uniquely structured meat-slicing tooth, suggesting that this adaptation was later lost. The molecular tree supports an African origin for the wolf-like canids, as the two African jackals are the most basal members of this clade. The two other large groupings of canids are (1) the South American canids, which are clearly rooted by the two most morphologically divergent canids, the maned wolf and bush dog; and (2) the red fox-like canids, which are rooted by the fennec fox and Blanford’s fox, but now also include the raccoon dog and bat-eared fox with higher support. Together, these three clades contain 93% of all living canids. The grey fox lineage seems to be the most primitive and suggests a North American origin of the living canids about 10 million years ago¹.

These results demonstrate the close kinship of canids. Their limited sequence divergence suggests that many molecular tools developed for the dog (for example, expression microarrays) will be useful for exploring adaptation and evolutionary divergence in other canids as well.

Conclusions

Genome comparison is a powerful tool for discovery. It can reveal unknown—and even unsuspected—biological functions, by sifting the records of evolutionary experiments that have occurred over 100 years or over 100 million years. The dog genome sequence illustrates the range of information that can be gleaned from such studies.

Mammalian genome analysis is helping to develop a global picture of gene regulation in the human genome. Initial comparison with rodents revealed that ~5% of the human genome is under purifying selection, and that the majority of this sequence is not protein-

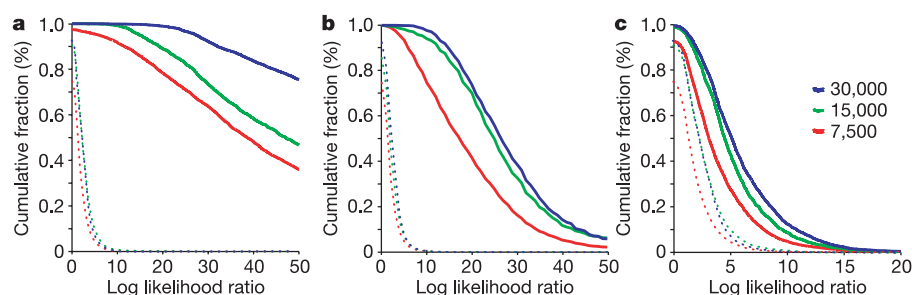


Figure 9 | Power to detect a disease locus by association mapping. One SNP was designated as a disease allele under one of three genetic models: (a) simple mendelian dominant, (b) fivefold multiplicative increase in risk and (c) twofold multiplicative increase in risk. SNP genotypes across surrounding chromosomal regions of 1 Mb were simulated, using the coalescent model corresponding to observed within-breed variation (see text). Diploid genotypes across the chromosomal region were then generated for 100 affected and 100 unaffected dogs, based on the disease model, and association analysis was performed to detect the presence of the

disease allele. The distribution of the maximum LOD score across the 1-Mb region is shown for analyses based on multi-SNP haplotypes (solid lines) with SNP densities equivalent to a genome-wide map with a total of 7,500 (red), 15,000 (green) or 30,000 (blue) SNPs. Dotted curves show the null distribution for a genome-wide search in which no disease locus is present (see Supplementary Information). A LOD score of 5 corresponds to <3% chance of a false positive across the genome. For this threshold, the power to detect a disease allele that increases risk by twofold using haplotype analysis and a map with 15,000 SNPs is ~50%.

coding. The dog genome is now further clarifying this picture, as our data suggest that this ~5% represents functional elements common to all mammals. The distribution of these elements relative to genes is highly heterogeneous, with roughly half of the most highly conserved non-coding elements apparently devoted to regulating ~1% of human genes; these genes have important roles in development, and understanding the regulatory clusters that surround them may reveal how cellular states are established and maintained. In recent papers^{32,108}, the dog genome sequence has been used to greatly expand the catalogue of mammalian regulatory motifs in promoters and 3'-untranslated regions. The dog genome sequence is also being used to substantially revise the human gene catalogue. Despite these advances, it is clear that mammalian comparative genomics is still in its early stages. Progress will be markedly accelerated by the availability of many additional mammalian genome sequences, initially with light coverage²⁸ but eventually with near-complete coverage.

In addition to its role in studies of mammalian evolution, the dog has a special role in genomic studies because of the unparalleled phenotypic diversity among closely related breeds. The dog is a testament to the power of breeding programmes to select naturally occurring genetic variants with the ability to shape morphology, physiology and behaviour. Genome comparison within and across breeds can reveal the genes that underlie such traits, informing basic research on development and neurobiology. It can also identify disease genes that were carried along in breeding programmes. Potential benefits include insights into disease mechanism, and the possibility of clinical trials in disease-affected dogs to accelerate new therapeutics that would improve health in both dogs and humans. The SNP map of the dog genome confirms that dog breeds show the long-range haplotype structure expected from recent intensive breeding. Moreover, our analysis shows that the current collection of >2.5 million SNPs should be sufficient to allow association studies of

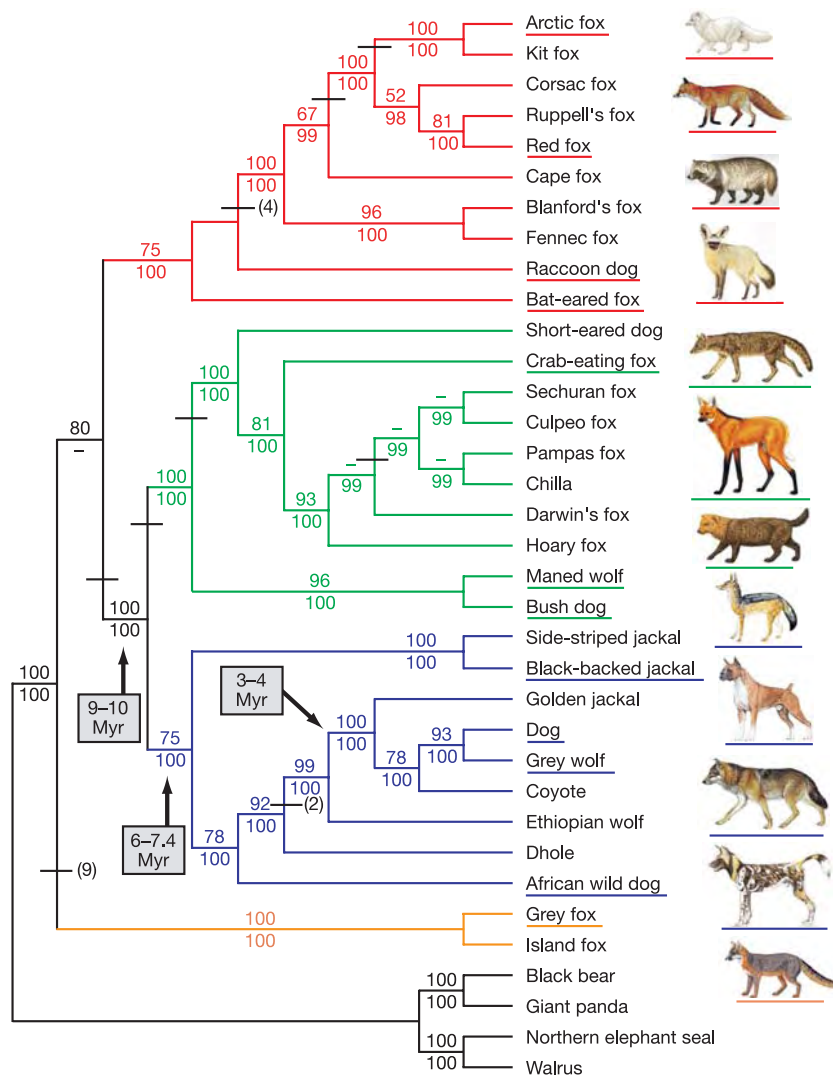


Figure 10 | Phylogeny of canid species. The phylogenetic tree is based on ~15 kb of exon and intron sequence (see text). Branch colours identify the red-fox-like clade (red), the South American clade (green), the wolf-like clade (blue) and the grey and island fox clade (orange). The tree shown was constructed using maximum parsimony as the optimality criterion and is the single most parsimonious tree. Bootstrap values and bayesian posterior probability values are listed above and below the internodes, respectively; dashes indicate bootstrap values below 50% or bayesian posterior probability values below 95%. Horizontal bars indicate indels, with the number of indels shown in parentheses if greater than one. Underlined

species names are represented with corresponding illustrations. (Copyright permissions for illustrations are listed in the Supplementary Information.) Divergence time, in millions of years (Myr), is indicated for three nodes as discussed in ref. 1. For scientific names and species descriptions of canids, see ref. 119. A tree based on bayesian inference differs from the tree shown in two respects: it groups the raccoon dog and bat-eared fox as sister taxa, and groups the grey fox and island fox as basal to the clade containing these sister taxa. However, neither of these topological differences is strongly supported (see text and Supplementary Information).

nearly any trait in any breed. Realizing the full power of dog genetics now awaits the development of appropriate genotyping tools, such as multiplex 'SNP chips'¹⁰⁹—this is already underway. For millennia, dogs have accompanied humans on their travels. It is only fitting that the dog should also be a valued companion on our journeys of scientific discovery.

METHODS

Detailed descriptions of all methods are provided in the Supplementary Information. Links to all of the data can be obtained via the Broad Institute website (<http://www.broad.mit.edu/tools/data.html>).

WGS sequencing and assembly. Approximately 31.5 million sequence reads were derived from both ends of inserts (paired-end reads) from 4-, 10-, 40- and 200-kb clones, all prepared from primary blood lymphocyte DNA from a single female boxer. This particular animal was chosen for sequencing because it had the lowest heterozygosity rate among ~120 dogs tested at a limited set of loci; subsequent analysis showed that the genome-wide heterozygosity rate in this boxer is not substantially different from other breeds⁹¹. The assembly was carried out using an interim version of ARACHNE2+ (<http://www.broad.mit.edu/wga/>). **Genome alignment and comparison.** Synteny maps were generated using standard methods²⁴ from pair-wise alignments of repeat masked assemblies using PatternHunter¹¹⁰ on CanFam2.0. All other comparative analyses were performed on BLASTZ/MULTIZ^{111,112} genome-wide alignments obtained from the UCSC genome browser (<http://genome.ucsc.edu>), based on CanFam1.0. Known interspersed repeats were identified and dated using RepeatMasker and DateRepeats¹¹³. The numbers of orthologous nucleotides were counted directly from the alignments using human (hg17) as the reference sequence for all overlaps except the dog–mouse overlap, for which pair-wise (CanFam1.0, mm5) alignments were used.

Divergence rate estimates. Orthologous ancestral repeats were excised from the genome alignment and realigned with the corresponding RepBase consensus using ClustalW. Nucleotide divergence rates were estimated from concatenated repeat alignments using baseml with the REV substitution model¹¹⁴. Orthologous coding regions were excised from the genome alignments using the annotated human coding sequences (CDS) from Ensembl and the UCSC browser Known Genes track (October 2004) as reference. K_A and K_S were estimated for each orthologue triplet using codeml with the F3 × 4 codon frequency model and no additional constraints.

Detection and clustering of sequence conservation. Pair-wise conservation scores and the fraction of orthologous sequences under purifying selection were estimated as in ref. 24. The three-way conservation score S_{HMD} was defined as $S_{\text{HMD}} = (p - u) / \sqrt{(u(1 - u))/n}$, where n is the number of nucleotides aligned across all three genomes (human, mouse, dog) for each non-overlapping 50-bp window with more than 20 aligned bases, p is the fraction of nucleotides identical across all three genomes, and u is the mean identity of ancestral repeats within 500 kb of the window. HCNEs were defined as windows with $S_{\text{HMD}} > 5.4$ that did not overlap a coding exon, as defined by the UCSC Known Genes track, and HCNE clusters were defined as all runs of overlapping 1-Mb intervals (50-kb step size) across the human genome with HCNE densities in the 90th percentile.

Gene set acceleration scores. Gene annotation was performed on CanFam1.0. A set of 13,816 orthologous human, mouse and dog genes were identified and compiled into 4,950 gene sets containing genes related by functional annotations or microarray gene expression data. For each gene set S , the acceleration score $A(S)$ along a lineage is defined by (1) ranking all genes based on K_A/K_S within a lineage, (2) calculating the rank-sum statistic for the set along each lineage (denoted $a_{\text{dog}}(S)$, $a_{\text{mouse}}(S)$, $a_{\text{human}}(S)$), (3) calculating the rank-sum for the lineage minus the maximum rank-sum the other lineages, for example, $a_{\text{human}}(S) - \max(a_{\text{dog}}(S), a_{\text{mouse}}(S))$ and (4) converting this rank-sum difference to a z -score by comparing it to the mean and standard deviation observed in 10,000 random sets of the same size. The expected number of sets at a given z -score threshold was estimated by repeating steps (1)–(4) 10,000 times for groups of 4,950 randomly permuted gene sets.

SNP discovery. The SNP discovery was performed on CanFam2.0. Set 1 SNPs were discovered by comparison of the two haplotypes derived from the boxer assembly using only high-quality discrepancies supported by two reads. SNPs in sets 2 and 3 were discovered by aligning reads or contigs to the boxer assembly and using the SSAHA SNP algorithm¹¹⁵.

Haplotype structure. The SNPs within the sequenced boxer genome (CanFam2.0) were assigned to homozygous or heterozygous regions using a Viterbi algorithm¹¹⁶. To determine whether the haplotype structure seen in the boxer is representative of most dog breeds, we randomly selected ten regions of 15 Mb each (~6% of the CanFam2.0 genome) and examined the extent of homozygosity and linkage disequilibrium in these regions in a collection of 224

dogs, consisting of 20 dogs from each of 10 breeds (akita, basenji, bullmastiff, English springer spaniel, Glen of Imaal terrier, golden retriever, Irish wolfhound, Labrador retriever, pug and rottweiler) and one dog from each of 24 additional breeds (see Supplementary Information). For each instance in which a dog was homozygous in a particular 10-kb region, we measured the distance from the beginning of the 10-kb region to the first heterozygous SNP in the adjoining 100-kb, 1-Mb and 15-Mb data. This distance was used as the extent of homozygosity. The boxer sequence was sampled in an identical manner to the actual breed data. Linkage disequilibrium (represented by r^2) across the ten 15-Mb regions was assessed using Haploview¹¹⁷.

Received 9 August; accepted 11 October 2005.

- Wayne, R. K. *et al.* Molecular systematics of the Canidae. *Syst. Biol.* **46**, 622–653 (1997).
- Vila, C. *et al.* Multiple and ancient origins of the domestic dog. *Science* **276**, 1687–1689 (1997).
- Bardeleben, C., Moore, R. L. & Wayne, R. K. Isolation and molecular evolution of the selenocysteine tRNA (*Cf TRSP*) and RNase P RNA (*Cf RPPH1*) genes in the dog family, Canidae. *Mol. Biol. Evol.* **22**, 347–359 (2005).
- Savolainen, P., Zhang, Y. P., Luo, J., Lundeberg, J. & Leitner, T. Genetic evidence for an East Asian origin of domestic dogs. *Science* **298**, 1610–1613 (2002).
- American Kennel Club. *The Complete Dog Book* (eds Crowley, J. & Adelman, B.) (Howell Book House, New York, 1998).
- Wayne, R. K. Limb morphology of domestic and wild canids: the influence of development on morphologic change. *J. Morphol.* **187**, 301–319 (1986).
- Ostrander, E. A., Galibert, F. & Patterson, D. F. Canine genetics comes of age. *Trends Genet.* **16**, 117–123 (2000).
- Patterson, D. Companion animal medicine in the age of medical genetics. *J. Vet. Intern. Med.* **14**, 1–9 (2000).
- Sargan, D. R. IDID: inherited diseases in dogs: web-based information for canine inherited disease genetics. *Mamm. Genome* **15**, 503–506 (2004).
- Chase, K. *et al.* Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. *Proc. Natl Acad. Sci. USA* **99**, 9930–9935 (2002).
- Breen, M. *et al.* Chromosome-specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes. *Genome Res.* **11**, 1784–1795 (2001).
- Breen, M., Bullerdiel, J. & Langford, C. F. The DAPI banded karyotype of the domestic dog (*Canis familiaris*) generated using chromosome-specific paint probes. *Chromosome Res.* **7**, 401–406 (1999).
- Breen, M. *et al.* An integrated 4249 marker FISH/RH map of the canine genome. *BMC Genomics* **5**, 65 (2004).
- Hitte, C. *et al.* Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nature Rev. Genet.* **6**, 643–648 (2005).
- Li, R. *et al.* Construction and characterization of an eightfold redundant dog genomic bacterial artificial chromosome library. *Genomics* **58**, 9–17 (1999).
- Kirkness, E. F. *et al.* The dog genome: survey sequencing and comparative analysis. *Science* **301**, 1898–1903 (2003).
- Sutter, N. & Ostrander, E. Dog star rising: The canine genetic system. *Nature Rev. Genet.* **5**, 900–910 (2004).
- Galibert, F., Andre, C. & Hitte, C. Dog as a mammalian genetic model [in French]. *Med. Sci. (Paris)* **20**, 761–766 (2004).
- Pollinger, J. P. *et al.* Selective sweep mapping of genes with large phenotypic effects. *Genome Res.* doi:10.1101/gr.4374505 (in the press).
- Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
- Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
- Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
- Murphy, W. J. *et al.* Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614–618 (2001).
- Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
- Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102**, 4795–4800 (2005).
- Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).

30. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
31. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3**, e10 (2005).
32. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
33. Dermitzakis, E. T. *et al.* Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14**, 852–859 (2004).
34. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
35. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
36. Richterich, P. Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res.* **8**, 251–259 (1998).
37. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. & Eichler, E. E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5**, R23 (2004).
38. Andelfinger, G. *et al.* Detailed four-way comparative mapping and gene order analysis of the canine *ctvm* locus reveals evolutionary chromosome rearrangements. *Genomics* **83**, 1053–1062 (2004).
39. Wang, W. & Kirkness, E. F. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Res.* doi:10.1101/gr.3765505 (in the press).
40. Mamedov, I. Z., Arzumanyan, E. S., Amosova, A. L., Lebedev, Y. B. & Sverdlov, E. D. Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach. *Nucleic Acids Res.* **33**, e16 (2005).
41. Lin, L. *et al.* The sleep disorder canine narcolepsy is caused by a mutation in the *hypocretin (orexin) receptor 2* gene. *Cell* **98**, 365–376 (1999).
42. Pele, M., Tiret, L., Kessler, J. L., Blot, S. & Panthier, J. J. SINE exonic insertion in the *PTPLA* gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs. *Hum. Mol. Genet.* **14**, 1417–1427 (2005).
43. Fondon, J. W. III & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl Acad. Sci. USA* **101**, 18058–18063 (2004).
44. Galtier, N. & Mouchiroud, D. Isochore evolution in mammals: a human-like ancestral structure. *Genetics* **150**, 1577–1584 (1998).
45. Belle, E. M., Duret, L., Galtier, N. & Eyre-Walker, A. The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *J. Mol. Evol.* **58**, 653–660 (2004).
46. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
47. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
48. Cooper, G. M., Brudno, M., Green, E. D., Batzoglu, S. & Sidow, A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13**, 813–820 (2003).
49. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl Acad. Sci. USA* **101**, 13994–14001 (2004).
50. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl Acad. Sci. USA* **90**, 4087–4091 (1993).
51. Gillooly, J. F., Allen, A. P., West, G. B. & Brown, J. H. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc. Natl Acad. Sci. USA* **102**, 140–145 (2005).
52. Laird, C. D., McConaughy, B. L. & McCarthy, B. J. Rate of fixation of nucleotide substitutions in evolution. *Nature* **224**, 149–154 (1969).
53. Li, W. H., Tanimura, M. & Sharp, P. M. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* **25**, 330–342 (1987).
54. Webber, C. & Ponting, C. P. Hot spots of mutation and breakage in dog and human chromosomes. *Genome Res.* doi:10.1101/gr.3896805 (in the press).
55. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
56. Marques-Bonet, T. & Navarro, A. Chromosomal rearrangements are associated with higher rates of molecular evolution in mammals. *Gene* **353**, 147–154 (2005).
57. Miller, W., Makova, K. D., Nekrutenko, A. & Hardison, R. C. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5**, 15–56 (2004).
58. Smith, N. G., Brandstrom, M. & Ellegren, H. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* **84**, 806–813 (2004).
59. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
60. Ovcharenko, I. *et al.* Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15**, 137–145 (2005).
61. Walter, K., Abnizova, I., Elgar, G. & Gilks, W. R. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends Genet.* **21**, 436–440 (2005).
62. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
63. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
64. Kimura-Yoshida, C. *et al.* Characterization of the pufferfish *Otx2* cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development* **131**, 57–71 (2004).
65. Uchikawa, M., Ishida, Y., Takemoto, T., Kamachi, Y. & Kondoh, H. Functional analysis of chicken *Sox2* enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev. Cell* **4**, 509–519 (2003).
66. de la Calle-Mustienes, E. *et al.* A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate *Iroquois* cluster gene deserts. *Genome Res.* **15**, 1061–1072 (2005).
67. Daly, M. J. Estimating the human gene count. *Cell* **109**, 283–284 (2002).
68. Hogenesch, J. B. *et al.* A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106**, 413–415 (2001).
69. Emes, R. D., Goodstadt, L., Winter, E. E. & Ponting, C. P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12**, 701–709 (2003).
70. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
71. Wolfe, K. H. & Li, W. H. Molecular evolution meets the genomics revolution. *Nature Genet.* **33** (suppl.), 255–265 (2003).
72. Bailey, J. A., Liu, G. & Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**, 823–834 (2003).
73. Hughes, A. L. The evolution of the type I interferon gene family in mammals. *J. Mol. Evol.* **41**, 539–548 (1995).
74. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486 (2002).
75. Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl Acad. Sci. USA* **99**, 16134–16137 (2002).
76. Demetrius, L. Directionality theory and the evolution of body size. *Proc. Biol. Sci.* **267**, 2385–2391 (2000).
77. Fay, J. C. & Wu, C. I. Sequence divergence, functional constraint, and selection in protein evolution. *Annu. Rev. Genomics Hum. Genet.* **4**, 213–235 (2003).
78. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101**, 4164–4169 (2004).
79. Mootha, V. K. *et al.* PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34**, 267–273 (2003).
80. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
81. Dorus, S. *et al.* Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell* **119**, 1027–1040 (2004).
82. Saetre, P. *et al.* From wild wolf to domestic dog: gene expression changes in the brain. *Brain Res. Mol. Brain Res.* **126**, 198–206 (2004).
83. Wyckoff, G. J., Wang, W. & Wu, C. I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**, 304–309 (2000).
84. Birkhead, T. R. & Pizzari, T. Postcopulatory sexual selection. *Nature Rev. Genet.* **3**, 262–273 (2002).
85. Dorus, S., Evans, P. D., Wyckoff, G. J., Choi, S. S. & Lahn, B. T. Rate of molecular evolution of the seminal protein gene *SEMG2* correlates with levels of female promiscuity. *Nature Genet.* **36**, 1326–1329 (2004).
86. Ruiz-Pesini, E. *et al.* Correlation of sperm motility with mitochondrial enzymatic activities. *Clin. Chem.* **44**, 1616–1620 (1998).
87. Zeh, J. A. & Zeh, D. W. Maternal inheritance, sexual conflict and the maladapted male. *Trends Genet.* **21**, 281–286 (2005).
88. Grossman, L. I., Wildman, D. E., Schmidt, T. R. & Goodman, M. Accelerated evolution of the electron transport chain in anthropoid primates. *Trends Genet.* **20**, 578–585 (2004).
89. Ostrander, E. A. & Kruglyak, L. Unleashing the canine genome. *Genome Res.* **10**, 1271–1274 (2000).
90. Sutter, N. B. *et al.* Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Res.* **12**, 2388–2396 (2004).
91. Parker, H. G. *et al.* Genetic structure of the purebred domestic dog. *Science* **304**, 1160–1164 (2004).
92. Bardeleben, C., Moore, R. L. & Wayne, R. K. A molecular phylogeny of the Canidae based on six nuclear loci. *Mol. Phylogenet. Evol.* **37**, 815–831 (2005).
93. Fogel, B. *The Encyclopedia of the Dog* (D.K. Publishing, New York, 1995).
94. Wilcox, B. & Walkowicz, C. *The Atlas of Dog Breeds of the World* (T.H.F. Publications, Neptune City, New York, 1995).
95. Frazer, K. A. *et al.* Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 mb of mouse genome. *Genome Res.* **14**, 1493–1500 (2004).
96. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* Vol. 7 (eds Futuyma, D. & Antonovics, J.) 1–44 (Oxford Univ. Press, Oxford, 1990).
97. Vila, C., Seddon, J. & Ellegren, H. Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends Genet.* **21**, 214–218 (2005).

98. Leonard, J. A. *et al.* Ancient DNA evidence for Old World origin of New World dogs. *Science* **298**, 1613–1616 (2002).
99. Kajiwaru, N. & Japanese Kennel Club in Akita (eds Kariyabu, T. & Kaluzniacki, S.) 1–103 (Japan Kennel Club, Tokyo, 1998).
100. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
101. Werner, P., Raducha, M. G., Prociuk, U., Henthorn, P. S. & Patterson, D. F. Physical and linkage mapping of human chromosome 17 loci to dog chromosomes 9 and 5. *Genomics* **42**, 74–82 (1997).
102. Todhunter, R. J. *et al.* Power of a Labrador Retriever-Greyhound pedigree for linkage analysis of hip dysplasia and osteoarthritis. *Am. J. Vet. Res.* **64**, 418–424 (2003).
103. Sidjanin, D. J. *et al.* Canine *CNGB3* mutations establish cone degeneration as orthologous to the human achromatopsia locus *ACHM3*. *Hum. Mol. Genet.* **11**, 1823–1833 (2002).
104. Lou, X. Y. *et al.* The extent and distribution of linkage disequilibrium in a multi-hierarchical outbred canine pedigree. *Mamm. Genome* **14**, 555–564 (2003).
105. Hyun, C. *et al.* Prospects for whole genome linkage disequilibrium mapping in domestic dog breeds. *Mamm. Genome* **14**, 640–649 (2003).
106. Cardon, L. R. & Abecasis, G. R. Using haplotype blocks to map human complex trait loci. *Trends Genet.* **19**, 135–140 (2003).
107. Tsui, C. *et al.* Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. *Nucleic Acids Res.* **31**, 4910–4916 (2003).
108. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
109. Syvanen, A. C. Toward genome-wide SNP genotyping. *Nature Genet.* **37** (suppl.), S5–10 (2005).
110. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445 (2002).
111. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
112. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
113. Smit, A. F. A. & Green, P. RepeatMasker (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>).
114. Yang, Z., Goldman, N. & Friday, A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**, 316–324 (1994).
115. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
116. Viterbi, A. J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Inform. Process.* **13**, 260–269 (1967).
117. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
118. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
119. Macdonald, D. W. & Sillero-Zubiri, C. in *Biology and Conservation of Canids* (eds Macdonald, D. W. & Sillero-Zubiri, C.) 1–30 (Oxford Univ. Press, Oxford, 2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are indebted to the canine research community, and in particular D. Patterson, G. Acland and K. G. Lark, whose vision and research convinced the NIH of the importance of generating a canine genome sequence. We also thank all those who shared insights at the Dog Genome Community meetings, including G. Acland, G. D. Aguirre, M. Binns, U. Giger, P. Henthorn, F. Lingaas, K. Murphy and P. Werner. We thank our many colleagues (G. Acland, G. D. Aguirre, C. Andre, N. Fretwell, G. Johnson, K. G. Lark and J. Modiano), as well as the dog owners and breeders who provided us with samples. We thank colleagues at the UCSC browser for providing data (such as BLASTZ alignments), A. Smit for providing the RepeatMasker annotations used in our analyses and N. Manoukis for providing Unix machines for the phylogenetic analyses. Finally, we thank L. Gaffney and K. Siang Toh for editorial and graphical assistance. The genome sequence and analysis was supported in part by the National Human Genome Research Institute. The radiation hybrid map was supported in part by the Canine Health Foundation. Sample collection was supported in part by the Intramural Research Program of the National Human Genome Research Institute and the Canine Health Foundation.

Author Information The draft genome sequence has been deposited in public databases under NCBI accession codes AAEX01000000 (CanFam1.0) and AAEX02000000 (CanFam2.0). SNPs have been deposited in the dbSNP database (<http://www.ncbi.nlm.nih.gov/projects/SNP/>). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.L.T. (kersli@broad.mit.edu) or E.S.L. (lander@broad.mit.edu).

Broad Sequencing Platform members Jennifer Baldwin¹, Adal Abebe¹, Amr Abouelleil¹, Lynne Aftuck¹, Mostafa Ait-zahra¹, Tyler Aldredge¹, Nicole Allen¹, Peter An¹, Scott Anderson¹, Claudel Antoine¹, Harindra Arachchi¹, Ali Aslam¹, Laura Ayotte¹, Pasang Bachantsang¹, Andrew Barry¹, Tashi Bayul¹, Mostafa Benamara¹, Aaron Berlin¹, Daniel Besette¹, Berta Blitshteyn¹, Toby Bloom¹, Jason Blye¹, Leonid Boguslavskiy¹, Claude Bonnet¹, Boris Boukhgalter¹, Adam Brown¹, Patrick Cahill¹, Nadia Calixte¹, Jody Camarata¹, Yama Cheshatsang¹, Jeffrey Chu¹, Mieke Citroen¹, Alville Collymore¹, Patrick Cooke¹, Tenzin Dawoe¹, Riza Daza¹, Karin Decktor¹, Stuart DeGray¹, Norbu Dhargay¹, Kimberly Dooley¹, Kathleen Dooley¹, Passang Dorje¹, Kunsang Dorjee¹, Lester Dorris¹, Noah Duffey¹, Alan Dupes¹, Osebhajajeme Egbiremolen¹, Richard Elong¹, Jill Falk¹, Abderrahim Farina¹, Susan Faro¹, Diallo Ferguson¹, Patricia Ferreira¹, Sheila Fisher¹, Mike FitzGerald¹, Karen Foley¹, Chelsea Foley¹, Alicia Franke¹, Dennis Friedrich¹, Diane Gage¹, Manuel Garber¹, Gary Gearin¹, Georgia Giannoukos¹, Tina Goode¹, Audra Goyette¹, Joseph Graham¹, Edward Grandbois¹, Kunsang Gyaltzen¹, Nabil Hafez¹, Daniel Hagopian¹, Birhane Hagos¹, Jennifer Hall¹, Claire Healy¹, Ryan Hegarty¹, Tracey Honan¹, Andrea Horn¹, Nathan Houde¹, Leanne Hughes¹, Leigh Hunnicutt¹, M. Husby¹, Benjamin Jester¹, Charlien Jones¹, Asha Kamat¹, Ben Kanga¹, Cristyn Kells¹, Dmitry Khazanovich¹, Alix Chinh Kieu¹, Peter Kisner¹, Mayank Kumar¹, Krista Lance¹, Thomas Landers¹, Marcia Lara¹, William Lee¹, Jean-Pierre Leger¹, Niall Lennon¹, Lisa Leuper¹, Sarah LeVine¹, Jinlei Liu¹, Xiaohong Liu¹, Yeshi Lokyitsang¹, Tashi Lokyitsang¹, Annie Lui¹, Jan Macdonald¹, John Major¹, Richard Marabella¹, Kebede Maru¹, Charles Matthews¹, Susan McDonough¹, Teena Mehta¹, James Meldrim¹, Alexandre Melnikov¹, Louis Meneus¹, Atanas Mihalev¹, Tanya Mihova¹, Karen Miller¹, Rachel Mittelman¹, Valentine Mlenga¹, Leonidas Mulrain¹, Glen Munson¹, Adam Navidi¹, Jerome Naylor¹, Tuyen Nguyen¹, Nga Nguyen¹, Cindy Nguyen¹, Thu Nguyen¹, Robert Nicol¹, Nyima Norbu¹, Choe Norbu¹, Nathaniel Novod¹, Tenchoe Nyima¹, Peter Olandt¹, Barry O'Neill¹, Keith O'Neill¹, Sahal Osman¹, Lucien Oyono¹, Christopher Patti¹, Danielle Perrin¹, Pema Phunkhang¹, Fritz Pierre¹, Margaret Priest¹, Anthony Rachupka¹, Sujaa Raghuraman¹, Rayale Rameau¹, Verneda Ray¹, Christina Raymond¹, Filip Rege¹, Cecil Rise¹, Julie Rogers¹, Peter Rogov¹, Julie Sahalie¹, Sampath Settipalli¹, Theodore Sharpe¹, Terrance Shea¹, Mechele Sheehan¹, Ngawang Sherpa¹, Jianying Shi¹, Diana Shih¹, Jessie Sloan¹, Cherylyn Smith¹, Todd Sparrow¹, John Stalker¹, Nicole Stange-Thomann¹, Sharon Stavropoulos¹, Catherine Stone¹, Sabrina Stone¹, Sean Sykes¹, Pierre Tchuinga¹, Pema Tenzing¹, Senait Tesfaye¹, Dawa Thoultsang¹, Yama Thoultsang¹, Kerri Topham¹, Ira Topping¹, Tsamla Tsamla¹, Helen Vassiliev¹, Vijay Venkataraman¹, Andy Vo¹, Tsering Wangchuk¹, Tsering Wangdi¹, Michael Weiland¹, Jane Wilkinson¹, Adam Wilson¹, Shailendra Yadav¹, Shuli Yang¹, Xiaoping Yang¹, Geneva Young¹, Qing Yu¹, Joanne Zainoun¹, Lisa Zembek¹ & Andrew Zimmer¹

[This page is intentionally left blank]

Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences

Tarjei S. Mikkelsen^{1,2}, Matthew J. Wakefield³, Bronwen Aken⁴, Chris T. Amemiya⁵, Jean L. Chang¹, Shannon Duke⁶, Manuel Garber¹, Andrew J. Gentles^{7,8}, Leo Goodstadt⁹, Andreas Heger⁹, Jerzy Jurka⁸, Michael Kamal¹, Evan Mauceli¹, Stephen M. J. Searle⁴, Ted Sharpe¹, Michelle L. Baker¹⁰, Mark A. Batzer¹¹, Panayiotis V. Benos¹², Katherine Belov¹³, Michele Clamp¹, April Cook¹, James Cuff¹, Radhika Das¹⁴, Lance Davidow¹⁵, Janine E. Deakin¹⁶, Melissa J. Fazzari¹⁷, Jacob L. Glass¹⁷, Manfred Grabherr¹, John M. Grealley¹⁷, Wanjun Gu¹⁸, Timothy A. Hore¹⁶, Gavin A. Huttley¹⁹, Michael Kleber¹, Randy L. Jirtle¹⁴, Edda Koina¹⁶, Jeannie T. Lee¹⁵, Shaun Mahony¹², Marco A. Marra²⁰, Robert D. Miller¹⁰, Robert D. Nicholls²¹, Mayumi Oda¹⁷, Anthony T. Papenfuss³, Zuly E. Parra¹⁰, David D. Pollock¹⁸, David A. Ray²², Jacqueline E. Schein²⁰, Terence P. Speed³, Katherine Thompson¹⁶, John L. VandeBerg²³, Claire M. Wade^{1,24}, Jerilyn A. Walker¹¹, Paul D. Waters¹⁶, Caleb Webber⁹, Jennifer R. Weidman¹⁴, Xiaohui Xie¹, Michael C. Zody¹, Broad Institute Genome Sequencing Platform*, Broad Institute Whole Genome Assembly Team*, Jennifer A. Marshall Graves¹⁶, Chris P. Ponting⁹, Matthew Breen^{6,25}, Paul B. Samollow²⁶, Eric S. Lander^{1,27} & Kerstin Lindblad-Toh¹

We report a high-quality draft of the genome sequence of the grey, short-tailed opossum (*Monodelphis domestica*). As the first metatherian ('marsupial') species to be sequenced, the opossum provides a unique perspective on the organization and evolution of mammalian genomes. Distinctive features of the opossum chromosomes provide support for recent theories about genome evolution and function, including a strong influence of biased gene conversion on nucleotide sequence composition, and a relationship between chromosomal characteristics and X chromosome inactivation. Comparison of opossum and eutherian genomes also reveals a sharp difference in evolutionary innovation between protein-coding and non-coding functional elements. True innovation in protein-coding genes seems to be relatively rare, with lineage-specific differences being largely due to diversification and rapid turnover in gene families involved in environmental interactions. In contrast, about 20% of eutherian conserved non-coding elements (CNEs) are recent inventions that postdate the divergence of Eutheria and Metatheria. A substantial proportion of these eutherian-specific CNEs arose from sequence inserted by transposable elements, pointing to transposons as a major creative force in the evolution of mammalian gene regulation.

Metatherians ('marsupials') comprise one of the three major groups of modern mammals and represent the closest outgroup to the eutherian ('placental') mammals (Supplementary Fig. 1). Metatherians

and eutherians diverged ~180 million years (Myr) ago, long before the radiation of the extant eutherian clades ~100 Myr ago¹². Although the metatherian lineage originally radiated from North

¹Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ²Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ³Bioinformatics Division, The Walter & Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville Victoria 3050, Australia. ⁴The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ⁵Molecular Genetics Program, Benaroya Research Institute at Virginia Mason, 1201 Ninth Avenue, Seattle, Washington 98101, USA. ⁶Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, 4700 Hillsborough Street, Raleigh, North Carolina 27606, USA. ⁷Stanford University School of Medicine, P060 Lucas Center, Stanford, California 94305, USA. ⁸Genetic Information Research Institute, 1925 Landings Drive, Mountain View, California 94043, USA. ⁹MRC Functional Genetics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK. ¹⁰Department of Biology, Center for Evolutionary and Theoretical Immunology, University of New Mexico, Albuquerque, New Mexico 87131, USA. ¹¹Department of Biological Sciences, Biological Computation and Visualization Center, Center for Bio-Modular Multi-Scale Systems, Louisiana State University, 202 Life Sciences Building, Baton Rouge, Louisiana 70803, USA. ¹²Department of Computational Biology, University of Pittsburgh, 3501 Fifth Avenue, Suite 3064, BST3, Pittsburgh, Pennsylvania 15260, USA. ¹³Faculty of Veterinary Science, University of Sydney, New South Wales 2006, Australia. ¹⁴Department of Radiation Oncology, Duke University Medical Center, Box 3433, Durham, North Carolina 27710, USA. ¹⁵Department of Molecular Biology, Hughes Medical Institute, Massachusetts General Hospital, and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA. ¹⁶ARC Centre for Kangaroo Genomics, Research School of Biological Sciences, The Australian National University, Canberra, ACT 2601, Australia. ¹⁷Department of Medicine (Hematology) and Molecular Genetics, Albert Einstein College of Medicine, Ullmann 911, 1300 Morris Park Avenue, Bronx, New York 10461, USA. ¹⁸Department of Biochemistry and Molecular Genetics, University of Colorado Health Sciences Center, MS 8101, 12801 17th Avenue, Aurora, Colorado 80045, USA. ¹⁹John Curtin School of Medical Research, The Australian National University, Canberra, ACT 0200, Australia. ²⁰Genome Sciences Centre, British Columbia Cancer Agency, 570 West 7th Avenue, Vancouver, British Columbia V5Z 4S6, Canada. ²¹Department of Pediatrics, Research Center Children's Hospital of Pittsburgh, 3460 Fifth Avenue, Room 2109, Rangos, Pittsburgh, Pennsylvania 15213, USA. ²²Department of Biology, West Virginia University, Morgantown, West Virginia 26505, USA. ²³Department of Genetics and Southwest National Primate Research Center, Southwest Foundation for Biomedical Research, San Antonio, Texas 78245, USA. ²⁴Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA. ²⁵Center for Comparative Medicine and Translational Research, North Carolina State University, 4700 Hillsborough Street, Raleigh, North Carolina 27606, USA. ²⁶Department of Veterinary Integrative Biosciences, Texas A&M University, 4458 TAMU, College Station, Texas 77843, USA. ²⁷Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA.

*Lists of participants and affiliations appear at the end of the paper.

America, only one extant species can be found there (the Virginia opossum), whereas all other species are found in South America (including more than 65 species of opossums and shrew opossums) and Australasia (~200 species, including possums, kangaroos, koalas and many small insectivores and carnivores)³.

All sequenced mammalian genomes until now have come from eutherian species. Although metatherians and eutherians (together, 'therians') share many ancient mammalian characteristics, they have each evolved distinctive morphological and physiological traits. Metatherians are particularly noted for the birth of young at a very early stage of development, followed by a lengthy and complex lactational period. Genomic analysis will help reveal the genetic innovations that underlie the distinctive traits of each lineage^{4–6}.

Equally important, metatherian genomes can shed light on the human genome. Comparative analysis of eutherians has greatly improved our understanding of the architecture and functional organization of mammalian genomes^{7–10}. Identification of sequence elements thought to be under purifying selection, on the basis of cross-species sequence conservation, has led to increasingly refined inventories of protein-coding genes^{11,12}, proximal and distal regulatory elements^{13,14} and putative RNA genes¹⁵. Yet, we still know relatively little about the evolutionary dynamics of these and other functional elements: how stable is the complement of protein-coding genes? How rapidly do regulatory sequences appear and disappear? From what substrate do they evolve?

Comparison of the human genome with genomes from distant outgroups such as birds (divergence ~310 Myr ago) or fish (~450 Myr ago) has provided valuable information. When similarity between sequences from such distantly related genomes can be detected, it surely signals functional importance; but the high specificity of these signals¹⁶ is offset by dramatically reduced sensitivity^{10,17,18}. Simulations have shown that the feasibility of aligning orthologous genomic sequences declines rapidly once their mean genetic distance exceeds 1 substitution per site¹⁹. The genome of chicken, the most closely related non-mammalian amniote genome available, is separated from the human genome by approximately 1.7 substitutions per site in orthologous, neutrally evolving sequences²⁰. Even moderately constrained functional elements may therefore be difficult to detect. In contrast, metatherian mammals are well positioned to address this issue: because unconstrained regions of their genomes are separated from that of human by only ~1 substitution per site (see below), most orthologous, constrained sequence should be readily aligned.

Here we report the first high-quality draft of a metatherian genome sequence, which was derived from a female, grey, short-tailed opossum—*Monodelphis domestica*. The species was chosen chiefly on the availability and utility of the organism for research purposes. *M. domestica* is a small rapidly breeding South American species that has been raised in pedigreed colonies for more than 25 years and developed as one of only two laboratory bred metatherians^{21,22}. *M. domestica* is being actively used as a model system for investigations in mechanisms of imprinting^{23–25}, immunogenetics^{26–28}, neurobiology, neoplasia and developmental biology (reviewed in ref. 6). For example, newborn opossums are remarkable in that they can heal complete transections of the spinal cord²⁹. Elucidation of the molecular mechanisms underlying this ability promise important insights relevant to regenerative medicine concerning spinal cord or peripheral nerve injuries. Other than human, *M. domestica* is also the only mammal known in which ultraviolet radiation is a complete carcinogen for malignant melanoma³⁰, and this has led to its establishment as a unique neoplasia model. All of these investigations will directly benefit from the development of genomic resources for this species.

Below we describe the generation of the draft sequence of the opossum genome, analyse its large-scale characteristics, and compare it to previously sequenced amniote genomes. Our key findings include:

- The distinctive features of the opossum genome provide an informative test of current models of genome evolution and support the hypothesis that biased gene conversion has a key role in determining overall nucleotide composition.

- The evolution of random inactivation of the X chromosome in eutherians correlates with acquisition of X-inactive-specific transcript (*XIST*), elevation in long interspersed element (LINE)/L1 density and suppression of large-scale rearrangements.

- The opossum genome seems to contain 18,000–20,000 protein-coding genes, the vast majority of which have eutherian orthologues. Lineage-specific genes largely originate from expansion and rapid turnover in gene families involved in immunity, sensory perception and detoxification.

- Identification of orthologues of highly divergent immune genes and a novel T-cell receptor isotype challenge previous claims that metatherians possess a 'primitive' immune system.

- Of the non-coding sequences conserved among eutherians, ~20% seem to have evolved after the divergence from metatherians. Of protein-coding sequences conserved among eutherians, only ~1% seems to be absent in opossum.

- At least 16% of eutherian-specific conserved non-coding elements are clearly derived from transposons, implicating these elements as an important creative force in mammalian evolution.

Extensions to these findings, as well as additional topics, are reported in a series of companion papers^{31–41}.

Genome assembly and single nucleotide polymorphism discovery

We sequenced the genome of a partially inbred female opossum using the whole-genome shotgun (WGS) method^{7,42}. The resulting WGS assembly has a total length of 3,475 megabases (Mb), consistent with size estimates based on flow cytometry (~3.5–3.6 Gb; Supplementary Notes 1–2 and Supplementary Fig. 2). Approximately 97% of the assembled sequence has been anchored to eight large autosomes and one sex chromosome on the basis of genetic markers mapped by linkage analysis³⁸ or fluorescence *in situ* hybridization⁴³ (FISH; Supplementary Note 3). The draft genome sequence has high continuity, coverage and accuracy (Table 1; Supplementary Note 4 and Supplementary Tables 1–7).

To enable genetic mapping studies of opossum, we also created a large catalogue of candidate single nucleotide polymorphisms (SNPs). We identified ~775,000 SNPs within the sequenced individual by analysing assembled sequence reads. We identified an additional ~510,000 SNPs by generating and comparing ~300,000 sequence reads from three individuals from distinct, partially outbred laboratory stocks maintained at the Southwest Foundation for Biomedical Research (San Antonio, Texas)^{22,44} (Supplementary Note 5). The SNP rates between the different stocks range from

Table 1 | Genome assembly characteristics

WGS assembly (monDom5)	
Number of sequence reads	38.8 × 10 ⁶
Sequence redundancy (Q20 bases)	6.8 ×
Contig length (kb; N50*)	108
Scaffold length (Mb; N50)	59.8
Anchored bases in the assembly (Mb)	3,412
Estimated euchromatic genome size† (Mb)	3,475
Integration of physical mapping data	
Scaffolds anchored on chromosomes	216
Fraction of genome in anchored and oriented scaffolds (%)	91
Fraction of genome in anchored, but unoriented, scaffolds (%)	6
Quality control	
Bases with quality score ≥40 (%)	98
Empirical error rate for bases with quality score ≥40‡ (%)	3 × 10 ^{–5}
Empirical euchromatic sequence coverage‡ (%)	99
Bases in regions with low probability of structural error§ (%)	98

* N50 is the size *x* such that 50% of the assembly reside in contigs/scaffolds of length at least *x*.

† Includes anchored bases and spanned gaps (~2%).

‡ Based on comparison with 1.66 Mb of finished bacterial artificial chromosome (BAC) sequence.

§ Based on ARACHNE assembly certification (see Supplementary Note 4).

1 per 360 to 1 per 140 bases and correlate with the distance between their geographical origins (Supplementary Table 8–10 and Supplementary Fig. 3).

The data from this study, including the draft genome assembly and SNPs, are freely available on our website (<http://www.broad.mit.edu/mammals/opossum/>) and have been deposited in appropriate public databases.

Genome landscape

The opossum genome has certain unusual properties that provide an opportunity to test recent models of genome evolution. The opossum autosomes are extremely large: they range from 257 Mb to 748 Mb, with the smallest being larger than the largest chromosome previously sequenced in any amniote (human chromosome 1). In contrast, the X chromosome is only ~76 Mb long; this is substantially less than the size of the X chromosome in any sequenced eutherian. Studies of G-banding and chromosome painting have also shown that karyotypes and basic chromosomal organization are extraordinarily conserved throughout Metatheria, even between the distantly related American and Australasian lineages (~55–80 Myr ago)^{5,45}.

Sequence composition. Recent analyses have uncovered two major trends in the evolution of sequence composition in amniote genomes: first, most modern lineages seem to be experiencing a gradual decline in total G+C content relative to their common ancestors⁴⁶; second, the local rate of recombination is positively correlated with local G+C content and, even more strongly, with the local density of CpG dinucleotides^{20,47}. These observations have led to a proposed model⁴⁸ whereby sequence composition reflects the balance between a genome-wide, (A+T)-biased mutation process and a localized recombination-mediated (G+C)-biased gene conversion process. This model predicts that the sequence composition of a genomic region is a function of its historical rate of recombination, with the frequency of hypermutable CpG dinucleotides being a particularly sensitive indicator.

The opossum genome fits the predictions of this model well (see also refs 34, 35). Current linkage data³⁸ show that the average recombination rate for the autosomes (~0.2–0.3 cM Mb⁻¹) is lower than in other sequenced amniotes (0.5–>3 cM Mb⁻¹). Consistent with the proposed model, the mean autosomal G+C content (37.7%) is also lower than in other sequenced amniotes (40.9–41.8%) and, in particular, the mean autosomal density of CpGs (0.9%) is twofold lower than in other amniotes (1.7–2.2%). Because large-scale patterns of recombination seem to be relatively stable in the absence of chromosomal rearrangements^{49,50}, the stability of the opossum karyotype suggests that the majority of the genome has experienced low recombination rates over an extended period. Indeed, the sequence composition is also more homogeneous than seen in other amniotes (Fig. 1).

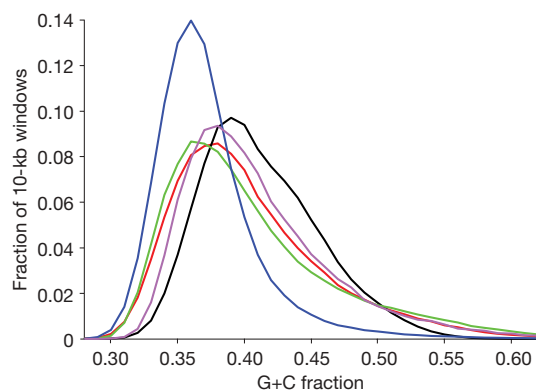


Figure 1 | Sequence composition in the opossum genome. Distribution of G+C content in 10-kb windows across the genome in opossum (blue), human (red), mouse (black), dog (green) and chicken (purple).

The subtelomeric regions of autosomes are notable outliers with respect to sequence composition in the opossum genome, providing additional support for the biased gene conversion hypothesis. Cytological studies in opossum^{51,52} suggest that the rate of chiasmata formation (and hence meiotic recombination) is relatively uniform across each autosome in males, whereas it is strongly biased to subtelomeric regions in females. Consistent with a higher sex-averaged rate of recombination, mean G+C-content (41.6%) and CpG density (1.9%) are significantly elevated within ~10 Mb of the chromosome ends (Supplementary Fig. 4).

Similarly, the very short X chromosome also supports the biased gene conversion hypothesis. Although few linkage data are currently available for opossum X chromosome, the average effective recombination rate must be at least 0.44 cM Mb⁻¹, and thus larger than for the autosomes. (This estimate follows from the requirement of at least one meiotic crossover per bivalent in the female germ-line^{53,54}.) The mean G+C content (40.9%) and CpG density (1.4%) of the X chromosome are substantially higher than for any of the autosomes (Supplementary Table 11). The opossum pattern is thus the opposite of that seen in eutherians, in which the X chromosome has low recombination and low G+C content and CpG density (Table 2).

Segmental duplication. In human and other eutherians, segmental duplications (defined as pairs of regions with ≥90% sequence similarity over ≥1 kb) are associated with chromosomal fragility and syntenic breakpoints^{55,56}. The relative karyotypic stability of metatherians therefore indicated that they might have a low proportion of segmental duplications.

The overall proportion of segmental duplication in opossum (1.7%) is indeed substantially lower than in other sequenced amniotes (2.5–5.3%). The segmental duplications are also relatively short: only 22 exceed 100 kb in opossum as compared with 483 in human (Supplementary Table 12). Additionally, the segmental duplications are more locally distributed: 76% are intrachromosomal (versus 46% for human) and the median distance between related duplications is 175 kb (versus 2.2 Mb for human). We find no indication that correction for over-collapsed duplications in the assembly

Table 2 | Comparative analysis of genome landscape in opossum and other amniotes

	Opossum	Human	Mouse	Dog	Chicken
Euchromatic genome size (Mb)	3,475	2,880	2,550	2,330	1,050
Karyotype					
Haploid number	9	23	20	39	33
Autosomal size range (Mb)	258–748	47–247	61–197	27–125	5–201
X chromosome size (Mb)	76	155	167	127	NA
Segmental duplications					
Autosomal (%)	1.7	5.2	5.3	2.5	10.4
Intrachromosomal duplications (%)	76	46	84	ND	ND
Median length between duplications (Mb)	0.18	2.2	1.6	0.33	0.03
X chromosome (%)	3.3	4.1	13	1.7	NA
Interspersed repeats (%)					
Total	52.2	45.5	40.9	35.5	9.4
LINE/non-LTR retrotransposon	29.2	20.0	19.6	18.2	6.5
SINE	10.4	12.6	7.2	10.2	NA
Endogenous retrovirus	10.6	8.1	9.8	3.7	1.3
DNA transposon	1.7	2.8	0.8	1.9	0.8
G+C content (%)					
Autosomal	37.7	40.9	41.8	41.1	41.5
X chromosome	40.9	39.5	39.2	40.2	NA
CpG content (%)					
Autosomal	0.9	2.0	1.7	2.2	2.1
X chromosome	1.4	1.7	1.2	1.9	NA
Recombination rate (cM Mb ⁻¹)					
Autosomal*	~0.2–0.3	1–2	0.5–1	1.3–3.4†	2.5–21
X chromosome‡	≥0.44§	0.8	0.3	ND	NA

NA, not applicable; ND, no or insufficient data.

* Range of chromosome-averaged recombination rates.

† See (http://www.vgl.ucdavis.edu/research/canine/projects/linkage_map/data/)

‡ Estimated as 2/3 of the female rate.

§ See text.

would significantly alter these estimates (Supplementary Note 6 and Supplementary Table 13).

Transposable elements. Metatherian transposable elements largely belong to families also found in eutherians, but can be divided into more than 500 subfamilies, many of which are lineage specific (catalogued in Repbase⁵⁷). At least 52% of the opossum genome can be recognized as transposable elements and other interspersed repeats (Table 2)^{33,35}, which is more than in any of the other sequenced amniotes (34–43%). Notably, the opossum genome is significantly enriched in non-long terminal repeat (LTR) retrotransposons (LINEs, 29%), comprising copies of various LINE subfamilies. Given the low abundance of segmental duplications, accumulation of transposable elements seems to be the primary reason for the relatively large opossum genome size. The total euchromatic sequence that is not recognized as transposable elements is rather similar in opossum and human (1638 Mb versus 1568 Mb, respectively). The enrichment of LINEs may be related to the overall low recombination rate in opossum, inasmuch as studies of eutherian genomes have shown that LINEs occur at elevated densities in regions with low local recombination rates⁴⁷.

Conserved synteny

Identification of syntenic segments between related genomes can facilitate reconstruction of chromosomal evolution and identification of orthologous functional elements. Starting from nucleotide-level, reciprocal-best alignments ('synteny anchors'), we found that the opossum and human genomes can be subdivided (at a resolution of 500 kb) into 510 collinear segments with an N50 length (size x such that 50% of the assembly is in units of length at least x) of 19.7 Mb, which cover 93% of the opossum genome (Supplementary Fig. 5). If local rearrangements are disregarded, these segments can be further grouped into 372 blocks of large-scale, conserved synteny.

Extending this analysis to additional eutherians (mouse, rat and dog), with chicken as an additional outgroup, we created a high-resolution synteny map that reveals 616 blocks of conserved synteny across the five fully sequenced mammals (Supplementary Note 7, Supplementary Figs 6–7 and Supplementary Table 14). Because the majority of synteny breakpoints between human, mouse, rat and dog are clearly lineage specific (see also ref. 10), genomic regions that were

probably contiguous in the last common boreoeutherian ancestor can be inferred by parsimony (Supplementary Note 8). We found that the mammalian synteny blocks can be used to infer 43 connected groups in the ancestral boreoeutherian genome (Supplementary Fig. 8). In fact, the largest 30 groups cover 95% of the human genome (see also ref. 58).

The resulting synteny map can be used to clarify chromosomal rearrangements during early mammalian evolution. For example, limited comparative mapping previously revealed that the eutherian X chromosome contains an 'X-conserved region' (XCR) that corresponds to the ancestral therian X chromosome, and an 'X-added region' (XAR), which was translocated from an autosome after the split from Metatheria^{59,60}. The exact extent of the XCR has been unclear, however, owing to unclear synteny with non-mammalian out-groups at its boundary⁶¹. Using our high-resolution synteny map we can now confidently map the XAR–XCR fusion point to 46.85 Mb on human chromosome band Xp11.3 (Fig. 2).

X chromosome inactivation

In opossum and other metatherian mammals, dosage compensation for X-linked genes is achieved through inactivity of the paternally derived X chromosome in females⁶². In contrast, eutherian dosage compensation involves inactivation of the paternal X chromosome at spermatogenesis, reactivation in the early embryo, followed by random and clonally stable inactivation of one of the two X chromosomes in each cell of female embryos⁶³. The random inactivation step is controlled by a complex locus known as the X inactivation centre (XIC). In the early female embryo, the non-coding *XIST* gene is transcribed from the XIC and coats one chromosome, *in cis*, to initiate silencing of the majority of its genes. It has been proposed that paternal X chromosome inactivation represents the ancestral therian dosage compensation system, and that random X chromosome inactivation is a recent innovation in the eutherian lineage^{64,65}. The opossum genome sequence provides the first opportunity to test major hypotheses about the evolution of this system.

No *XIST* homologue in opossum. We searched all assembled and unassembled opossum WGS sequence for homology to the human and mouse XIC non-coding genes but, in agreement with a recent report⁶⁶, did not find any significant alignments. (In particular, we

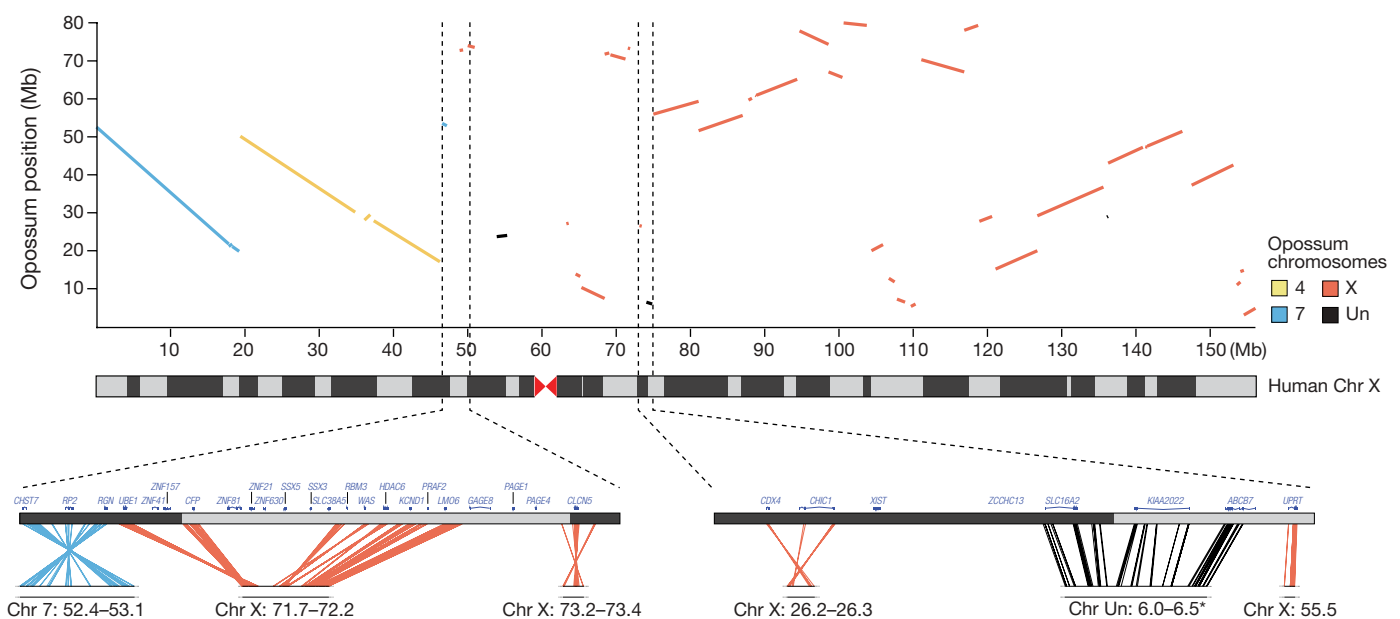


Figure 2 | Opossum–human synteny for the X chromosome. The dot plot shows correspondence between the human chromosome (Chr X) and opossum chromosomes at a resolution of 300 kb. Expanded views, at a resolution of 50 kb, of the XAR–XCR fusion and the XIC are shown on the

bottom left and right, respectively. In the XIC region, the closest contig on the distal flank (*) was not anchored in the monDom5 assembly (see Methods), but has been subsequently mapped near *UPRT* (opossum X chromosome ~55 Mb) by FISH⁴⁰.

found no match to the highly conserved 150-bp region overlapping the critical exon 4 of *XIST*; this region is so strongly conserved in the Eutheria that it should be readily detectable if present⁴⁰. Analysis of synteny in the regions surrounding the eutherian XIC also revealed that it has been disrupted by large-scale rearrangements (Fig. 2)^{40,41}. In eutherians, the XIC is flanked by the ancient protein-coding genes *CDX4-CHIC1* on one side and *SLC16A2-RNF12* on the other side. In both chicken and frog these four genes are clustered in autosomal XIC homologous regions (which do not contain homologues of the XIC non-coding genes⁶⁶). On the opossum X chromosome, however, these two pairs of genes are separated by ~29 Mb (compared with ~750 Kb in human). Taken together, the evidence strongly suggests that *XIST* is specific to eutherians^{40,41,66}.

The Lyon repeat hypothesis. LINE/L1 elements are of particular interest to the study of X chromosome inactivation. These transposable elements have been proposed to act as 'boosters' for the spread of X chromosome inactivation in *cis* from the XIC (reviewed in ref. 67). This hypothesis is supported in part by the observation that in human, LINE/L1 density is significantly elevated in the XCR (33%), where nearly all genes are inactivated, but approximates the autosomal density in the XAR (19%), where many genes escape inactivation (Fig. 3)^{61,68}. In mouse, we found that the LINE/L1 density is elevated in both the XCR (35%) and the XAR (32%), which is consistent with the observation that genes that escape inactivation on the human XAR are often inactivated in mouse⁶⁹. As previously observed in human⁶⁸, the LINE/L1 elevation in mouse is particularly dramatic among recent, lineage-specific subfamilies (Supplementary Fig. 9).

In contrast to human and mouse, the LINE/L1 density on the opossum X chromosome (22%) is significantly lower than in the eutherian XCR, and is in fact slightly less than in the autosomal regions homologous to the eutherian XAR (23%). This difference between metatherian and eutherian X chromosomes is not readily explained by any simple correlation between LINE/L1 density, recombination or mutation rates. We therefore conclude that LINE/L1 density is unlikely to be a critical factor for X chromosome inactivation in the metatherian lineage, and that the approximately twofold increase on the eutherian X chromosome may be directly related to the acquisition of *XIST* and random X chromosome inactivation.

Suppression of large-scale rearrangements. Comparative analyses have revealed that the structure of the human X chromosome has remained essentially unchanged since the eutherian radiation^{10,20,61}. A possible reason is that the requirement for *XIST* transcripts to spread across the chromosome from a central location has led to selection against structural rearrangements. For example, translocation of LINE/L1-poor XAR segments into the XCR could potentially disrupt inactivation at more distal loci. Consistent with this hypothesis, our synteny map reveals that the XAR and XCR homologous regions have experienced several major rearrangements both in the opossum lineage (~15 lineage-specific synteny breakpoints) and in the eutherian lineage before the eutherian radiation (~9 lineage-specific breakpoints; Supplementary Table 15). The low rate of rearrangements in the human lineage is therefore unlikely to be due to functions or

sequences that were present on the ancestral therian X chromosome, or in early eutherian evolution.

We note that unlike in human, the mouse X chromosome has experienced several rearrangements (with 15 lineage-specific synteny breakpoints), such that the XAR and XCR are no longer two separate segments. This would be consistent with the more comprehensive inactivation in the mouse imposing weaker constraints on rearrangement. Although little is known about the extent of X chromosome inactivation in dog or rat, their X chromosomes are also consistent with this hypothesis. The dog X chromosome is collinear with human and is enriched for LINE/L1 only in the XCR (33.4% versus 16.8% for the XAR). The rat X chromosome has accumulated ~4 lineage-specific synteny breakpoints after the divergence from mouse⁶¹, and is similarly enriched for LINE/L1 in both the XCR (36.7%) and the XAR (34.5%).

Genes

The gene content of metatherian and eutherian genomes provides key information about biological functions. We analysed the gene content of the opossum genome and compared it with that of the human genome. We focused on instances of rapid divergence and duplication of protein-coding genes, which have led to lineage-specific gene complements⁷⁰.

Gene catalogue. We generated an initial catalogue of 18,648 predicted protein-coding genes and 946 non-coding genes (primarily small nuclear RNA, small nucleolar RNA, microRNA and ribosomal RNA) in opossum³⁴ (Supplementary Note 9 and Supplementary Data). Regularly updated annotations can be obtained from public databases (<http://www.ensembl.org> and <http://genome.ucsc.edu>).

We next characterized orthology and paralogy relationships between predicted protein-coding genes in opossum and human¹¹ (Table 3). We could identify unambiguous human orthologues for 15,320 (82%) of the opossum predicted genes, with 12,898 cases having a single copy in each species (1:1 orthologues). Notably, we identified orthologues of key T-cell lineage markers such as CD4 and CD8, which had not been successfully identified by cloning in metatherian species³⁹. Most (2,704) of the remaining genes are homologous to human genes, but could not be assigned to orthologous groups with certainty.

A small number (624) of predicted opossum genes have no clear homologue among the human gene predictions. Inspection revealed that most of these are short (median length of 120 amino acids, compared with 445 for 1:1 orthologues) and probably originate from pseudogenes or spurious open reading frames. Only eight currently have strong evidence of representing functional genes without homologues in humans (Supplementary Table 16). These include CPD-photolyase, which is part of an ancestral photorepair system still active in opossum⁷¹, malate synthase⁷² and inosine/uridine hydrolase. The latter two are ancient genes not previously identified in a mammalian species.

Conversely, approximately ~1,100 current gene predictions from human have no clear homologue in the initial opossum catalogue (Supplementary Data). Of these, ~620 can be at least partially

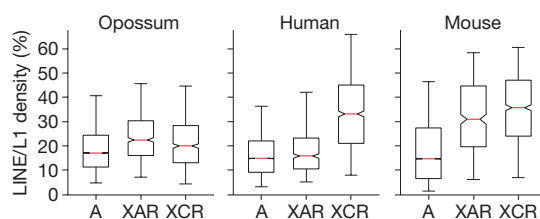


Figure 3 | Enrichment of LINE/L1 correlates with random X chromosome inactivation. Box plot of LINE/L1 density in 500-kb intervals across the autosomes (A), the X-added region (XAR) and its homologous regions in opossum, and the X conserved region (XCR). Red bar, median; box edges, 25th and 75th percentiles; whiskers, range.

Table 3 | Opossum and human gene predictions and projected gene counts

Protein-coding genes	Opossum
Initial predictions	18,648
Orthologues in human*	15,320
1:1	12,898
Many:1	1,016
1:Many	451
Many:Many	582
Homologues in human, but unclear orthology†	2,704
No predicted homologues in human	624
Projected total‡	18,000–20,000

* Includes some cases where multiple transcripts have inconsistent phylogenies, or where the predicted orthologue is a putative pseudogene.

† Includes members of highly duplicated gene families.

‡ Accounting for missed annotations in opossum and removal of probable pseudogenes.

aligned to the opossum genome and may not have been annotated as genes owing to imperfections in the draft assembly or high sequence divergence. In particular, manual re-annotation identified orthologues of several rapidly evolving cytokines³⁹. The remaining predictions are dominated by gene families known to have undergone expansion and rapid evolution in the human lineage, such as β -defensins and cancer-testis antigens. On the basis of our comparison, we conclude that the opossum genome probably contains ~18,000–20,000 protein-coding genes, with the vast majority having eutherian orthologues.

Divergence rates among orthologues. We calculated the synonymous substitution rate (K_S ; substitutions that do not result in amino acid change because of codon redundancy) of 1:1 opossum–human orthologues to approximate the unconstrained divergence rate between the species^{7,10}. The median value of K_S is 1.02. Consistent with expectation, this value is substantially smaller than the chicken–human K_S value (1.7), with the ratio being very close to the ratio of prior estimates of the divergence times for the two lineages (~180 Myr ago for opossum and ~310 Myr ago for chicken).

Notably, the median K_S for orthologues located on the XCR is significantly elevated relative to orthologues located on autosomes in both species (1.2 versus 1.0; $P < 10^{-3}$; see also refs 34, 35). This is the opposite to what is observed within Eutheria¹⁰, but is consistent with the expectation that the higher G+C-content and recombination rate on the opossum X chromosome relative to its autosomes implies a higher rate of mutation⁴⁷. A similar elevation can also be detected in subtelomeric regions³⁴.

Innovation and turnover in gene families. We next studied the evolution of gene family expansions in the metatherian lineage. The opossum gene catalogue contains 2,743 (15%) genes that have probably been involved in one or more duplication or gene conversion event since the last common ancestor with eutherian mammals, as inferred from low K_S between the copies (median = 0.41). The number of duplications is one-third fewer than the number of human lineage-specific duplications (4,037; 20%), which may reflect the lower rate of segmental duplication in the opossum genome.

We found a large number of lineage-specific copies of genes involved in sensory perception, such as the γ -crystallin family of eye lens proteins⁷³, and taste, odorant⁷⁴ and pheromone receptors. Other major lineage-specific duplications were found in the rapidly evolving KRAB zinc-finger family, and in genes related to toxin degradation and dietary adaptations, including cytochrome P450 and various gastric enzymes (see also ref. 34).

Innovation in the innate and adaptive immune systems is visible through substantial duplication or gene conversion involving the leukocyte receptor and natural killer complexes, immunoglobulins, type I interferons and defensins^{32,39}. The opossum genome also contains a new T-cell receptor isotype that is expressed early in ontogeny, before conventional T-cell receptors, and may provide early immune function in the altricial young³⁷.

The opossum also shows some surprising gene family expansions that are without precedent in other vertebrates. Notable among these are multiple duplications of the nonsense-mediated decay factors SMG5 and SMG6, and the pre-mRNA splicing factors, KIAA1604 and PRP18. The opossum genome also harbours two adjacent paralogous copies of DNA (cytosine-5)-methyltransferase 1 (DNMT1), which catalyses methylation of CpG dinucleotides. It will be interesting to discover if specialized functions have been adopted by these paralogous genes.

The patterns of evolution among duplicated genes largely mirror those observed in eutherians^{34,70}. The set of opossum paralogues is strongly biased towards recent duplications ($K_S < 0.1$) and in general have accumulated a disproportionately high number of non-synonymous mutations (Fig. 4). The median intraspecific ratio of nonsynonymous to synonymous substitution rates (K_A/K_S) between paralogues is 0.51, which is sixfold higher than the interspecies ratio seen for 1:1 orthologues (0.086). This is consistent with the rapid

gene birth and death model⁷⁵, which predicts that duplicated genes either undergo functional divergence in response to positive selection or rapidly degenerate owing to lack of evolutionary benefit.

Conserved sequence elements

The most surprising discovery to emerge from comparative analyses of eutherian genomes is the finding that the majority of evolutionarily conserved sequence does not represent protein-coding genes, but rather are conserved non-coding elements (CNEs)^{7,10}. The opossum genome provides a well-positioned outgroup to study the origin and evolution of these elements.

For simplicity, we will refer to sequence elements as ‘amniote conserved elements’ if they are conserved between chicken and at least one of opossum or human; ‘eutherian conserved elements’ if they are conserved between human and at least one of mouse, rat or dog; and ‘eutherian-specific elements’ if they are eutherian conserved sequence absent from both opossum and chicken. (‘Metatherian-specific elements’ surely also exist, but cannot be identified without additional metatherian genomes.)

Loss of amniote conserved elements in mammals. We first studied the extent to which amniote conserved elements have been lost in the human lineage. We focused on ~133,000 conserved intervals between opossum and chicken (68 Mb), ~50% of which overlaps protein-coding regions (Supplementary Data).

Nearly all (97.5%) of these amniote conserved elements can be aligned to the human genome (Fig. 5a). We reasoned that some of the remainder might be orthologous to sequence that lies within gaps in the current human assembly, or which had been missed by the initial genome-wide alignment. We therefore repeated the analysis, focusing only on amniote elements present in opossum and occurring in ‘ungapped intervals’ (that is, syntenic intervals between human and opossum that have no sequence gaps); the ungapped intervals contain 63% of all conserved elements.

We found that 99.0% of amniote elements in ungapped intervals could be unambiguously aligned to the human genome. The remaining 1.0% of amniote elements could not be found even by a more sensitive alignment algorithm (Fig. 5b), and thus seem to have been lost in the human lineage.

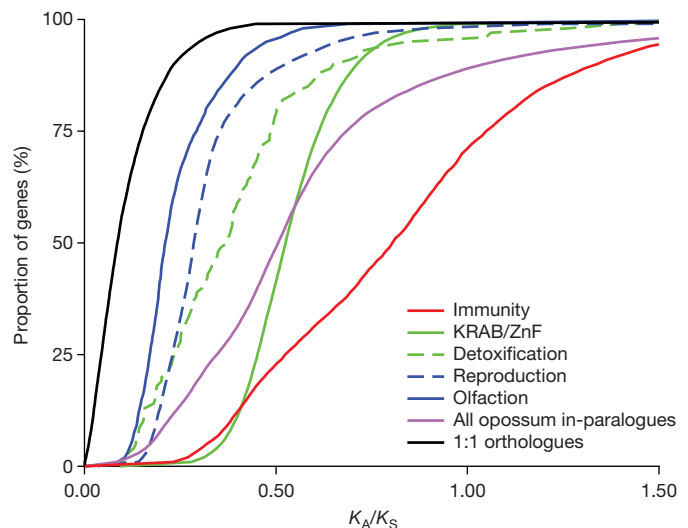


Figure 4 | Cumulative distribution of K_A/K_S values for duplicated genes. Estimates are shown for pairs of genes duplicated in opossum (in-paralogues) in the most common functional categories: immunity, KRAB zinc finger (ZnF) transcription factors, detoxification (including cytochrome P450, sulphotransferases), reproduction (including vomeronasal receptors, lipocalins and β -seminoproteins) and olfaction. The total distributions for opossum in-paralogues and opossum–human 1:1 orthologues are shown for comparison.

We also performed the converse analysis, by aligning the human and chicken genomes to identify amniote conserved elements potentially lost in opossum. The results were similar, with 99.4% of elements in ungapped intervals being readily aligned to opossum.

We conclude that the vast majority of amniote conserved elements encode such fundamental functions that they cannot be lost in either eutherians or metatherians. Nonetheless, the small fractions that have been lost correspond to more than 1,400 elements in total; it will be interesting to investigate their function and the consequence of their loss. Notably, although protein-coding sequence comprises 50% of all amniote conserved elements, they comprise only 4% of the elements lost in one of the lineages.

Eutherian-specific conserved elements. We next explored the appearance of novel conserved elements in the lineage leading from the common therian ancestor to the boreoeutherian ancestor, which could shed light on the origin of such elements in general. We identified a collection of eutherian conserved elements that cover 104 Mb (3.7%) of the human genome, using the phylo-HMM approach¹⁴; ~29% of them overlap protein-coding sequence (Supplementary Data).

Only a small proportion of human conserved protein-coding sequences could not be aligned to the opossum genome (1.1% in ungapped regions; Fig. 5c). In contrast, a much larger proportion of human non-coding elements seem to be eutherian specific (20.5% in ungapped regions). Taking the results from ungapped syntenic intervals as a conservative estimate for the proportion of total innovation, we conclude that approximately 14.8 Mb (1.1% of 30 Mb of coding sequence and 20.5% of 74 Mb of CNEs) of the eutherian conserved elements are eutherian specific.

The amount of apparent innovation is highest among short and moderately conserved elements (median length of 37 bp; median \log_2 -odds score = 22), probably reflecting, in part, that shorter elements may more readily diverge beyond recognition (see also refs 36, 76). Nonetheless, substantial innovation is apparent even among elements that are relatively long and unambiguously conserved within Eutheria. For example, the proportion of eutherian-specific elements is 8.1% among CNEs with \log_2 -odds score ≥ 60 , which have a median length of 197 bp (Fig. 5d).

Lineage-specific CNEs correspond to functional elements. To establish the biological relevance of lineage-specific CNEs, we examined the overlap of eutherian and amniote CNEs with two disparate sets of experimentally identified functional elements. If the eutherian-specific CNEs were enriched for false-positive predictions, we would expect them to be substantially under-represented among these functional elements.

We first considered a set of known human microRNAs (miRNAs)⁷⁷. Of the 51 miRNAs that overlap amniote CNEs, only one (*hsa-mir-194-1*; ref. 78) seems to have been lost in opossum (Fig. 5e). (The mature form of this miRNA is identical to a second conserved miRNA, *hsa-mir-194-2*, which does have an opossum orthologue; this apparent redundancy may have made it more susceptible to lineage-specific loss.) Of the 183 miRNAs that overlap eutherian CNEs in ungapped syntenic regions, 27 (15%) correspond to eutherian-specific elements (Supplementary Data). An example is an 87-bp eutherian-specific CNE corresponding to *hsa-mir-28*; it has previously been detected by northern blot analysis in human and mouse, but not in any non-mammalian species⁷⁹.

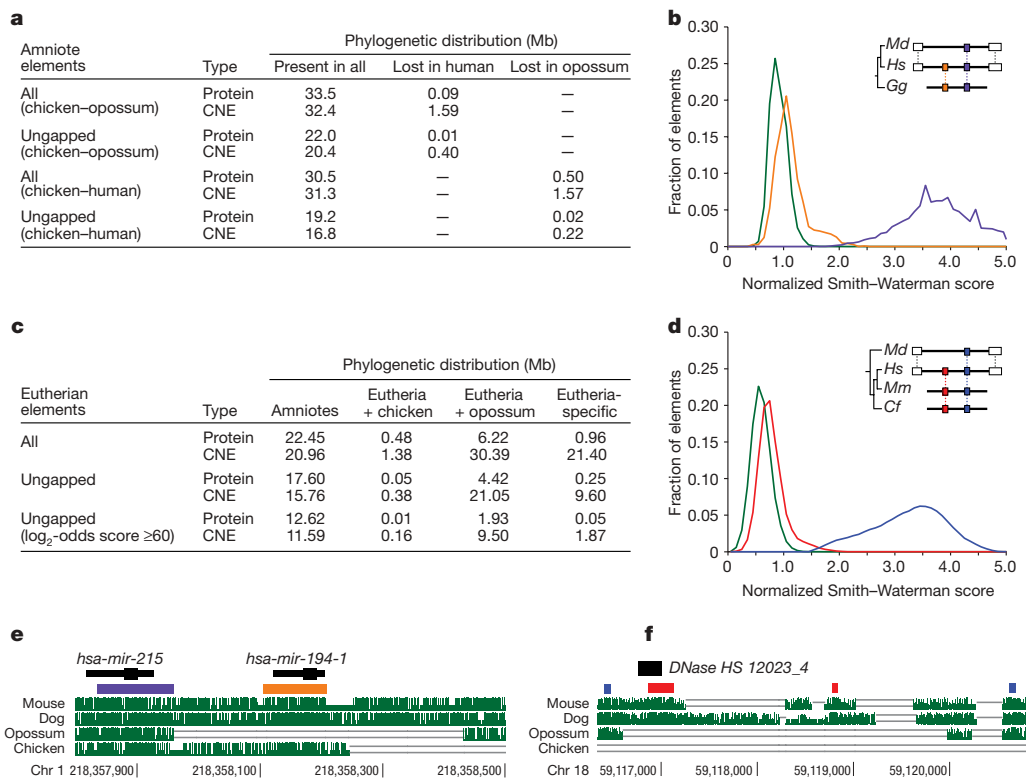


Figure 5 | Lineage-specific conserved sequence elements. **a**, Phylogenetic distribution of amniote conserved elements. **b**, Distribution for alignment scores of amniote elements, represented by opossum (human), to ungapped syntenic intervals in the human (opossum) genome, for shared (purple) and lineage-specific (orange) elements, and randomly permuted sequences of the same length and base composition (green). Ungapped syntenic intervals are flanked by two synteny anchors (white) and contain no assembly gaps (inset). *Md*, *Monodelphis domestica*; *Hs*, *Homo sapiens*; *Gg*, *Gallus gallus*. **c**, Phylogenetic distribution of eutherian conserved elements. **d**, Distribution of alignment scores for eutherian CNEs (\log_2 -odds

score ≥ 60), represented by human, to ungapped syntenic intervals in the opossum genome, for shared (blue) and eutherian-specific (red) elements, and randomly permuted sequences of the same length and base composition (green). The bimodal distribution of scores confirms that highly conserved eutherian-specific elements have no significant homology in the syntenic opossum sequence. *Mm*, *Mus musculus*; *Cf*, *Canis familiaris*. **e**, The miRNA *hsa-mir-194-1* corresponds to an amniote CNE lost in opossum (orange). It is flanked by an unrelated amniote miRNA that is present in opossum (purple). **f**, A eutherian-specific CNE in the intron of the *BCL2* gene (red) overlaps a DNase hypersensitive site in human lymphocytes (black).

We next considered a genome-wide set of DNase hypersensitive sites from human lymphocytes, which represent a variety of putative regulatory elements⁸⁰. Of the 290 sites that overlap amniote CNEs present in human, none overlaps instances that are lost in opossum. Of the 2,041 sites that overlap eutherian CNEs in ungapped syntenic regions, 407 (20%) exclusively overlap eutherian-specific elements (Supplementary Data). An example is a 269-bp eutherian-specific CNE in intron 2 of the apoptosis regulator *BCL2*, which overlaps a DNase hypersensitive site, suggesting it has a *cis*-regulatory function (Fig. 5f).

The fraction of eutherian CNEs overlapping DNase hypersensitive sites that are eutherian specific is strikingly similar to the fraction of all conserved non-coding sequence that is eutherian specific (20.5%). The fraction of miRNAs that correspond to eutherian-specific CNEs is slightly lower (15%), which is consistent with their higher average conservation scores. In particular, the results provide strong evidence that the majority of eutherian-specific CNEs are likely to be genuine functional elements.

Lineage-specific CNEs associated with key developmental genes.

We next explored the distribution of lineage-specific CNEs across the human genome. Overall, there is a strong regional correlation between the density of eutherian CNEs shared with opossum and the density of eutherian-specific CNEs (Spearman's $\rho = 0.82$ for 1-Mb windows; Fig. 6). The densities of amniote CNEs present or lost in opossum are also positively correlated (Spearman's $\rho = 0.30$).

Previous studies have shown that both eutherian and amniote CNEs are enriched in certain large, gene-poor regions surrounding genes that have key roles in development, primarily encoding transcription factors, morphogens and axon guidance receptors^{10,81,82}. For example, 35% of all eutherian CNEs and 49% of all amniote CNEs (in ungapped syntenic regions) lie within the 204 largest clusters of CNEs in the human genome (described in ref. 10). The ~240 key developmental genes in these regions have relatively low rates of amino acid divergence (median $K_A/K_S = 0.03$) and show little evidence of lineage-specific loss or duplications. In contrast, we found that the rate of gain and loss of CNEs in the same regions is only moderately (~30%) lower than elsewhere in the genome. Indeed, we identified more than 37,000 lineage-specific CNEs in these developmentally important regions.

Because experimental studies of CNEs in these regions have frequently uncovered *cis*-regulatory functions affecting the nearby developmental genes^{16,82–85}, the substantial innovations in these regions are candidates for genetic changes underlying differential morphological and neurological evolution in mammalian lineages. This pattern would be consistent with the notion that modification of regulatory networks has been a major force in the evolution of animal diversity^{86–88}.

Eutherian-specific CNEs derived from transposable elements. In general, each eutherian-specific element must have arisen by one of three mechanisms: (1) divergence of an ancestral functional element

to such an extent that its similarity is no longer detectable; (2) duplication of an ancestral functional element giving rise to an element without a 1:1 orthologue in other clades; or (3) evolution of a novel functional element from sequence that was absent or non-functional in the ancestral genome.

The first mechanism is not likely to account for most of the eutherian-specific CNE sequence, at least among those with high conservation scores—if an ancient functional element underwent such rapid divergence at some point in the eutherian lineage that it is no longer detectable, then there should be concomitant ‘loss’ of an amniote conserved element. But, lineage-specific loss seems to be relatively rare for both amniote elements, as shown above, and for eutherian elements¹⁰. The majority of eutherian-specific conserved elements therefore probably arose after the metatherian divergence, either by adaptive evolution of new or previously non-functional sequence, or by duplication of ancestral elements.

One intriguing source for eutherian-specific CNEs is transposable elements. A number of researchers have argued that transposable elements offer an obvious and ideal substrate for the evolution of lineage-specific functions^{89–93}. Transposable elements contain a variety of functional subunits that can be exapted and modified by the host genome^{89,91}, and they can mediate duplication of existing CNEs to distant genomic locations through transduction or chimaerism⁹². Individual instances of CNEs derived from transposable elements have been described previously^{14,94,95}. However, these cases together comprise only a trivial fraction of the CNEs in the human genome. It has thus been unclear whether the evolution of CNEs from transposable elements represents a general mechanism or a rare exception.

When we examined the set of eutherian-specific CNEs, we found a striking overlap with transposable elements. In ungapped syntenic intervals, at least 16% of eutherian-specific CNEs overlap currently recognized transposable elements in human. The fraction is similar (14%) if we focus only on the most highly conserved elements (phylo-HMM \log_2 -odds score) ≥ 60 , see above). The overlapping transposable elements originate from most major transposon families found in eutherians (Table 4), and are not clearly differentiated from other CNEs in terms of distribution across the genome. This implies that transposable-element-mediated evolution has been a significant creative force in the emergence of recent CNEs. The fact that sequences from transposable elements themselves can be identified within these CNEs also implies that exaptation of at least a portion of the transposable element, rather than simply incidental transduction of adjacent sequence, has been a frequent occurrence.

In contrast, the eutherian CNEs that are present in opossum (and thus are more ancient) only rarely show overlap with recognizable transposable elements (~0.7%). We speculate that many of these CNEs also arose from transposable elements, but that they are difficult to recognize as such owing to substantial divergence. In fact, three large

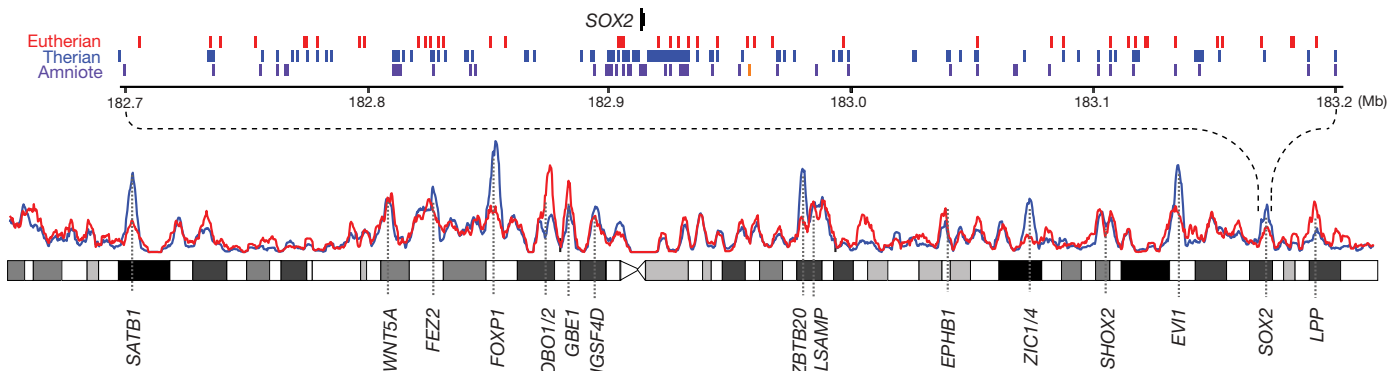


Figure 6 | Lineage-specific CNEs near key developmental genes. The densities of eutherian CNEs present (blue) or absent (red) in opossum are plotted in 1-Mb sliding windows across human chromosome 3. Peaks in the distributions often correspond to key developmental genes. The expanded

view shows positions of amniote CNEs (purple), eutherian CNEs not overlapping amniote CNEs (blue) and eutherian-specific CNEs (red) across a 500-kb gene desert surrounding the *SOX2* transcription factor gene. One amniote CNE present in human has been lost in opossum (orange).

families of ancient paralogous CNEs have recently been discovered that were clearly distributed around the genome as parts of transposable elements^{96–98}. In each case, only a minority of the family members still retain evidence of transposon-like features. We also previously described ~100 smaller CNE families that pre-date the eutherian radiation, but which had no members associated with known transposable elements⁹⁸. For all but two of these families, we can find orthologues in the opossum genome for the majority of their members (Supplementary Note 10 and Supplementary Fig. 10). Moreover, closer inspection reveals previously unrecognized transposon-like features in several of these and other ancient CNE families³³.

Strikingly, the proportion of eutherian-specific CNEs recognizable as transposable-element-derived (16%) is very similar to the proportion of the total aligned sequence between the human, mouse and dog genomes recognizable as ancestral transposable elements (~17% of ~812 Mb; the vast majority of which is inactive)¹⁰. It is widely suspected that the latter proportion is a significant underestimate owing to the difficulty of recognizing transposable elements that inserted more than ~100–200 Myr ago^{7,33}. In cases where the transposable-element-related sequence hallmarks are not essential to the subsequent CNE, or where evolution of a new function did not follow immediately after the transposable element insertion, exapted sequences would be expected to have diverged to the point that they can no longer be readily recognized at a rate similar to inactive insertions. Because this seems to have occurred for most of the families of ancient CNEs described above, it is likely that the proportion of all eutherian (not just eutherian-specific) CNEs derived from transposable elements is substantially higher than the observed proportion of 16%.

Conclusions

The generation of the first complete genome sequence for a marsupial, *Monodelphis domestica*, provides an important resource for genetic analysis in this unique model organism, as well as the first reference sequence for metatherian mammals. Our initial results demonstrate the usefulness of this sequence for comparative analyses of the architecture and functional organization of mammalian genomes.

The relationship of sequence composition, segmental duplications and transposable element density with the large and stable karyotype

of the opossum genome has provided new support for an emerging, general model of chromosome evolution in mammals. In addition, comparison of the opossum and eutherian X chromosomes revealed that the evolution of random X chromosome inactivation correlates with acquisition of *XIST*, elevation in LINE/L1 density and suppression of large-scale rearrangements.

Comparative analysis of protein-coding genes showed that the eutherian complement is largely conserved in opossum. Lineage-specific genes seem to be largely limited to gene families that are rapidly turning over in all mammals, although improved annotations that do not rely on homology to distant species will be required to complete the opossum gene catalogue. Identification of a wide array of both conserved and lineage-specific immune genes is particularly notable because limited success in isolating these genes by cloning has led to claims that the metatherian immune system is relatively 'primitive'. Availability of the genome sequence now facilitates more systematic study of the metatherian immune response³⁹.

At timescales longer than the characteristic time of loss for gene duplications, it is clear that innovation in non-coding elements has been substantially more common relative to protein-coding sequences, at least during eutherian evolution. The opossum genome sequence has provided the first estimate of the genome-wide rate of CNE innovation in eutherian evolution, as well as identification of tens of thousands of lineage-specific elements. It has also provided evidence that exaptation of transposable elements has a much greater role in the evolution of novel CNEs than has been previously realized.

Sequencing of additional metatherian genomes would be helpful for extending our results by allowing detection of metatherian-specific coding and non-coding elements. In addition, sampling of both the American and Australasian lineages would allow the reconstruction of the genome of their common ancestor, which would complement ongoing efforts for the boreoeutherian ancestral genome⁵⁸. The shorter genetic distance between the ancestral metatherian and boreoeutherian genomes (~0.6–0.7 substitutions per site) would facilitate a more comprehensive analysis of short and weakly conserved functional elements, for which the phylogenetic distribution and evolutionary origins are still difficult to ascertain.

METHODS SUMMARY

WGS sequencing and assembly. Approximately 38.8 million high-quality sequence reads were assembled using an interim version of ARACHNE2+ (<http://www.broad.mit.edu/wga/>).

SNP discovery. The SNP discovery was performed using ARACHNE and SSAHA-SNP⁹⁹. Linkage disequilibrium was assessed using Haploview¹⁰⁰.

Genome alignment and comparisons. Synteny maps were generated using standard methods^{7,10}.

Gene prediction and phylogeny. Opossum protein-coding and non-coding RNA genes were predicted using a modified version of the Ensembl genebuild pipeline¹⁰¹, followed by several rounds of refinement using Exonerate¹⁰² and manual curation. Orthology and paralogy were inferred using the PhyOP pipeline^{11,34}.

Conserved element prediction. Amniote conserved elements were inferred from pairwise BLASTZ alignment blocks with more than 75% identity for ≥100 bp. Eutherian conserved elements were inferred using phastCons¹⁴. Eutherian elements that did not fall within a 10-kilobase or longer synteny 'net'¹⁰³ were ignored.

Phylogeny of conserved elements. For amniote conserved elements, pairwise best-in-genome BLASTZ alignments of opossum to human and vice versa were used to infer their phylogenetic distributions. For eutherian conserved elements, concomitant BLASTZ/MULTIZ alignments to opossum and chicken were used. A conserved element was called absent from a species if it was not covered by a single aligned nucleotide in the relevant alignment.

Correction for assembly gaps and initial alignment artefacts. A conserved element was considered to be in an ungapped syntenic interval if it was flanked by two synteny anchors within 200 kb on the same contigs in both the human and opossum assemblies. All conserved elements in ungapped syntenic intervals were realigned using water (<http://emboss.sourceforge.net>). Putatively eutherian-specific elements, including *XIST*, were also searched against all opossum sequencing reads using MegaBLAST.

Table 4 | Eutherian-specific conserved non-coding elements derived from transposons

Transposon family	All		\log_2 -odds score ≥ 60	
	Number of CNEs*	Overlapped length (kb)†	Number of CNEs*	Overlapped length (kb)†
SINE/MIR	9,617	364	363	49
LINE/L1	6,619	286	194	36
LINE/L2	7,616	303	290	47
LINE/CR1	2,520	136	203	36
LINE/RTE	867	48	56	11
LTR/MaLR	1,995	65	25	3.7
LTR/ERV1	140	5.1	1	0.2
LTR/ERV2	992	36	12	2.8
DNA/Tip100	242	9.3	2	0.6
DNA/MER1_type	2,427	93	54	9
DNA/MER2_type	113	5.3	4	0.9
DNA/Tc2	162	8.5	6	1.4
DNA/Mariner	250	14.6	20	3.3
DNA/AcHobo	151	5.1	3	0.3
Unknown (MER121)	49	4	10	1.6
Total	33,760	1,383	1,243	203
Fraction of overlapped CNEs	16%		14%	

* Number of eutherian-specific CNEs in ungapped syntenic regions overlapping annotated transposable elements.

† Total length of annotated transposable element sequence overlapping the CNEs (this is less than the total length of CNEs overlapping transposable element sequence).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 5 December 2006; accepted 3 April 2007.

- Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
- Woodburne, M. O., Rich, T. H. & Springer, M. S. The evolution of tribospheny and the antiquity of mammalian clades. *Mol. Phylogenet. Evol.* **28**, 360–385 (2003).
- Tyndale-Biscoe, C. H. *Life of Marsupials* (CSIRO Publishing, Collingwood, Victoria, 2005).
- Wakefield, M. J. & Graves, J. A. M. Marsupials and monotremes sort genome treasures from junk. *Genome Biol.* **6**, 218 (2005).
- Graves, J. A. M. & Westerman, M. Marsupial genetics and genomics. *Trends Genet.* **18**, 517–521 (2002).
- Samollow, P. B. Status and applications of genomic resources for the gray, short-tailed opossum, *Monodelphis domestica*, an American marsupial model for comparative biology. *Aust. J. Zool.* **54**, 173–196 (2006).
- Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
- Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
- Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
- Goodstadt, L. & Ponting, C. P. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.* **2**, e133 (2006).
- Clamp, M. *et al.* Gene content of the human genome. *Nature* (submitted).
- Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
- Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
- Pedersen, J. S. *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2**, e33 (2006).
- Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
- Ovcharenko, I., Stubbs, L. & Loots, G. G. Interpreting mammalian evolution using Fugu genome comparisons. *Genomics* **84**, 890–895 (2004).
- Prabhakar, S. *et al.* Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res.* **16**, 855–863 (2006).
- Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102**, 4795–4800 (2005).
- Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
- VandeBerg, J. L. The gray short-tailed opossum (*Monodelphis domestica*) as a model didelphid species for genetic research. *Aust. J. Zool.* **37**, 235–247 (1990).
- VandeBerg, J. L. in *UFAW Handbook on the Management of Laboratory Animals*. Vol. 1 *Terrestrial Vertebrates* (eds Poole, T. & English, P.) 193–209 (Blackwell Science, Oxford, 1999).
- Murphy, S. K. & Jirtle, R. L. Imprinting evolution and the price of silence. *Bioessays* **25**, 577–588 (2003).
- Rapkins, R. W. *et al.* Recent assembly of an imprinted domain from non-imprinted components. *PLoS Genet.* **2**, e182 (2006).
- Weidman, J. R. *et al.* Phylogenetic footprint analysis of IGF2 in extant mammals. *Genome Res.* **14**, 1726–1732 (2004).
- Deakin, J. E. *et al.* Evolution and comparative analysis of the MHC Class III inflammatory region. *BMC Genomics* **7**, 281 (2006).
- Deakin, J. E., Olp, J. J., Graves, J. A. & Miller, R. D. Physical mapping of immunoglobulin loci *IGH@*, *IGK@*, and *IGL@* in the opossum (*Monodelphis domestica*). *Cytogenet. Genome Res.* **114**, 94H (2006).
- Belov, K. *et al.* Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex. *PLoS Biol.* **4**, e46 (2006).
- Wintzer, M. *et al.* Strategies for identifying genes that play a role in spinal cord regeneration. *J. Anat.* **204**, 3–11 (2004).
- VandeBerg, J. L. *et al.* Genetic analysis of ultraviolet radiation-induced skin hyperplasia and neoplasia in a laboratory marsupial model (*Monodelphis domestica*). *Arch. Dermatol. Res.* **286**, 12–17 (1994).
- Baker, M. L. *et al.* Analysis of a set of Australian northern brown bandicoot expressed sequence tags with comparison to the genome sequence of the south American grey short-tailed opossum. *BMC Genom.* **8**, 50 (2007).
- Belov, K. *et al.* Characterization of the opossum immune genome provides insights into the evolution of the mammalian immune system. *Genome Res.* doi:10.1101/gr.6121807 (2007).
- Gentles, A. J. *et al.* Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. *Genome Res.* doi:10.1101/gr.6070707 (2007).
- Goodstadt, L., Heger, A., Webber, C. & Ponting, C. P. An analysis of the gene complement of a marsupial *Monodelphis domestica*: Evolution of lineage-specific genes and giant chromosomes. *Genome Res.* doi:10.1101/gr.6093907 (2007).
- Gu, W. *et al.* Phylogenetic detection, population genetics, and distribution of active SINES in the genome of *Monodelphis domestica*. *Gene* doi:10.1016/j.gene.2007.02.028 (2007).
- Mahony, S., Corcoran, D. L., Feingold, E. & Benos, P. V. Regulatory conservation of protein coding and miRNA genes in vertebrates: lessons from the opossum genome. *Genome Biol. (in the press)*.
- Parra, Z. E. *et al.* A new T-cell receptor discovered in marsupials. *Proc. Natl Acad. Sci. USA* (submitted).
- Samollow, P. B. *et al.* A microsatellite-based, physically anchored linkage map for the gray, short-tailed opossum (*Monodelphis domestica*). *Chromosome Res.* advance online publication, doi:10.1007/s10577-007-1123-4 (25 February 2007).
- Wong, E. S., Young, L. J., Papenfuss, A. T. & Belov, K. *In silico* identification of opossum cytokine genes suggests the complexity of the marsupial immune system rivals that of eutherian mammals. *Immunome Res.* **2**, 4 (2006).
- Hore, T., Koina, E., Wakefield, M. J. & Graves, J. A. M. The region homologous to the X-chromosome inactivation centre has been disrupted in marsupial and monotreme mammals. *Chromosome Res.* **15**, 147–161 (2007).
- Davidow, L. S. *et al.* The search for a marsupial XIC reveals a break with vertebrate synteny. *Chromosome Res.* **15**, 137–146 (2007).
- Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- Duke, S. E. *et al.* Integrated cytogenetic BAC map of the genome of the gray short-tailed opossum, *Monodelphis domestica*. *Chromosome Res.* advance online publication, doi:10.1007/s10577-007-1131-4 (6 April 2007).
- VandeBerg, J. L. The laboratory opossum (*Monodelphis domestica*) in laboratory research. *ILAR J.* **38**, 4–12 (1997).
- Rens, W. *et al.* Karyotype relationships between distantly related marsupials from South America and Australia. *Chromosome Res.* **9**, 301–308 (2001).
- Belle, E. M., Duret, L., Galtier, N. & Eyre-Walker, A. The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *J. Mol. Evol.* **58**, 653–660 (2004).
- Jensen-Seaman, M. I. *et al.* Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14**, 528–538 (2004).
- Duret, L., Eyre-Walker, A. & Galtier, N. A new perspective on isochore evolution. *Gene* **385**, 71–74 (2006).
- Dumas, D. & Britton-Davidian, J. Chromosomal rearrangements and evolution of recombination: comparison of chiasma distribution patterns in standard and robertsonian populations of the house mouse. *Genetics* **162**, 1355–1366 (2002).
- Myers, S. *et al.* A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).
- Hope, R. M. Selected features of marsupial genetics. *Genetica* **90**, 165–180 (1993).
- Sharp, P. J. & Hayman, D. L. An examination of the role of chiasma frequency in the genetic system of marsupials. *Heredity* **60**, 77–85 (1988).
- Holm, P. B. Ultrastructural analysis of meiotic recombination and chiasma formation. *Tokai J. Exp. Clin. Med.* **11**, 415–436 (1986).
- Samollow, P. B. *et al.* First-generation linkage map of the gray, short-tailed opossum, *Monodelphis domestica*, reveals genome-wide reduction in female recombination rates. *Genetics* **166**, 307–329 (2004).
- Bailey, J. A. *et al.* Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5**, R23 (2004).
- Webber, C. & Ponting, C. P. Hotspots of mutation and breakage in dog and human chromosomes. *Genome Res.* **15**, 1787–1797 (2005).
- Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
- Ma, J. *et al.* Reconstructing contiguous regions of an ancestral genome. *Genome Res.* **16**, 1557–1565 (2006).
- Kohn, M. *et al.* Wide genome comparisons reveal the origins of the human X chromosome. *Trends Genet.* **20**, 598–603 (2004).
- Graves, J. A. Sex chromosome specialization and degeneration in mammals. *Cell* **124**, 901–914 (2006).
- Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).
- Cooper, D. W., Johnston, P. G., Graves, J. A. & Watson, J. M. X-inactivation in marsupials and monotremes. *Sem. Dev. Biol.* **4**, 117–128 (1993).
- Heard, E. Recent advances in X-chromosome inactivation. *Curr. Opin. Cell Biol.* **16**, 247–255 (2004).
- Wakefield, M. J., Keohane, A. M., Turner, B. M. & Graves, J. A. Histone underacetylation is an ancient component of mammalian X chromosome inactivation. *Proc. Natl Acad. Sci. USA* **94**, 9665–9668 (1997).
- Reik, W. & Lewis, A. Co-evolution of X-chromosome inactivation and imprinting in mammals. *Nature Rev. Genet.* **6**, 403–410 (2005).
- Duret, L. *et al.* The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312**, 1653–1655 (2006).
- Lyon, M. F. Do LINEs have a role in X-chromosome inactivation? *J. Biomed. Biotechnol.* **2006**, 59746 (2006).
- Bailey, J. A., Carrel, L., Chakravarti, A. & Eichler, E. E. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc. Natl Acad. Sci. USA* **97**, 6634–6639 (2000).
- Disteche, C. M., Filippova, G. N. & Tsuchiya, K. D. Escape from X inactivation. *Cytogenet. Genome Res.* **99**, 36–43 (2002).
- Emes, R. D., Goodstadt, L., Winter, E. E. & Ponting, C. P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12**, 701–709 (2003).

71. Kato, T. Jr *et al.* Cloning of a marsupial DNA photolyase gene and the lack of related nucleotide sequences in placental mammals. *Nucleic Acids Res.* **22**, 4119–4124 (1994).
72. Kondrashov, F. A. *et al.* Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol. Direct* **1**, 31 (2006).
73. Wistow, G. *et al.* γ N-crystallin and the evolution of the β -crystallin superfamily in vertebrates. *FEBS J.* **272**, 2276–2291 (2005).
74. Grus, W. E., Shi, P., Zhang, Y. P. & Zhang, J. Dramatic variation of the vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals. *Proc. Natl Acad. Sci. USA* **102**, 5767–5772 (2005).
75. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
76. Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19**, 1114–1121 (2002).
77. Griffiths-Jones, S. *et al.* miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34** (database issue), D140–D144 (2006).
78. Michael, M. Z. *et al.* Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol. Cancer Res.* **1**, 882–891 (2003).
79. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853–858 (2001).
80. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16**, 123–131 (2006).
81. Sandelin, A. *et al.* Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genom.* **5**, 99 (2004).
82. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
83. Bailey, P. J. *et al.* A global genomic transcriptional code associated with CNS-expressed genes. *Exp. Cell Res.* **312**, 3108–3119 (2006).
84. de la Calle-Mustienes, E. *et al.* A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate *Iroquois* cluster gene deserts. *Genome Res.* **15**, 1061–1072 (2005).
85. Pennacchio, L. A. *et al.* *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
86. Carroll, S. B. Evolution at two levels: on genes and form. *PLoS Biol.* **3**, e245 (2005).
87. Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800 (2006).
88. Stathopoulos, A. & Levine, M. Genomic regulatory networks and animal development. *Dev. Cell* **9**, 449–462 (2005).
89. Britten, R. J. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205**, 177–182 (1997).
90. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357 (1969).
91. Brosius, J. Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107**, 209–238 (1999).
92. Kazazian, H. H. Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
93. Marino-Ramirez, L., Lewis, K. C., Landsman, D. & Jordan, I. K. Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet. Genome Res.* **110**, 333–341 (2005).
94. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
95. Silva, J. C. *et al.* Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet. Res.* **82**, 1–18 (2003).
96. Bejerano, G. *et al.* A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90 (2006).
97. Nishihara, H., Smit, A. F. & Okada, N. Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res.* **16**, 864–874 (2006).
98. Xie, X., Kamal, M. & Lander, E. S. A family of conserved noncoding elements derived from an ancient transposable element. *Proc. Natl Acad. Sci. USA* **103**, 11659–11664 (2006).
99. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
100. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
101. Birney, E. *et al.* Ensembl 2006. *Nucleic Acids Res.* **34** (database issue), D556–D561 (2006).
102. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
103. Kent, W. J. *et al.* Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100**, 11484–11489 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Generation of the *Monodelphis domestica* sequence at the Broad Institute of MIT and Harvard was supported by grants from the National Human Genome Research Institute (NHGRI). For work from other members of the Opossum Genome Sequencing Consortium, we acknowledge the support of the National Institutes of Health (NHGRI, NIAID, NLM), the National Science Foundation, the Robert J. Kleberg Jr and Helen C. Kleberg Foundation, the State of Louisiana Board of Regents Support Fund, State of Colorado support funds, the Pittsburgh Foundation, TATRC/DoD, the UK Medical Research Council and the Australian Research Council. We thank colleagues at the UCSC genome browser for providing data (BLASTZ/MULTIZ alignments, synteny nets, and annotations). We thank L. Gaffney for assistance in preparing the manuscript and figures, and J. Danke for flow cytometry data.

Author Information All analysed data sets can be obtained from <http://www.broad.mit.edu/mammals/opossum/>. This *Monodelphis domestica* whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under NCBI accession code AAFR00000000. SNPs have been deposited in the dbSNP database (<http://www.ncbi.nlm.nih.gov/projects/SNP/>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.L.-T. (kersli@broad.mit.edu), T.S.M. (tarjei@broad.mit.edu) and E.S.L. (lander@broad.mit.edu).

Broad Institute Genome Sequencing Platform members Jennifer Baldwin¹, Amr Abdouelleil¹, Jamal Abdulkadir¹, Adal Abebe¹, Brikti Abera¹, Justin Abreu¹, St Christophe Acer¹, Lynne Aftuck¹, Allen Alexander¹, Peter An¹, Erica Anderson¹, Scott Anderson¹, Harindra Arachi¹, Marc Azer¹, Pasang Bachantsang¹, Andrew Barry¹, Tashi Bayul¹, Aaron Berlin¹, Daniel Bessette¹, Toby Bloom¹, Jason Blye¹, Leonid Boguslavskiy¹, Claude Bonnet¹, Boris Boukhgalter¹, Imane Bourzgui¹, Adam Brown¹, Patrick Cahill¹, Sheridan Channer¹, Yama Cheshatsang¹, Lisa Chuda¹, Mieke Citroen¹, Alville Collymore¹, Patrick Cooke¹, Maura Costello¹, Katie D'Acò¹, Riza Daza¹, Georgius De Haan¹, Stuart DeGray¹, Christina DeMaso¹, Norbu Dhargay¹, Kimberly Dooley¹, Erin Dooley¹, Missole Dorcent¹, Passang Dorje¹, Kunsang Dorjee¹, Alan Dupes¹, Richard Elong¹, Jill Falk¹, Abderrahim Farina¹, Susan Faro¹, Diallo Ferguson¹, Sheila Fisher¹, Chelsea D. Foley¹, Alicia Franke¹, Dennis Friedrich¹, Loryn Gadbois¹, Gary Gearin¹, Christina R. Gearin¹, Georgia Giannoukos¹, Tina Goode¹, Joseph Graham¹, Edward Grandbois¹, Sharleen Grewal¹, Kunsang Gyaltzen¹, Nabil Hafez¹, Birhane Hagos¹, Jennifer Hall¹, Charlotte Henson¹, Andrew Hollinger¹, Tracey Honan¹, Monika D. Huard¹, Leanne Hughes¹, Brian Hurhula¹, M. Erii Husby¹, Asha Kamat¹, Ben Kanga¹, Seva Kashin¹, Dmitry Khazanovich¹, Peter Kisner¹, Krista Lance¹, Marcia Lara¹, William Lee¹, Niall Lennon¹, Frances Letendre¹, Rosie LeVine¹, Alex Lipovsky¹, Xiaohong Liu¹, Jinlei Liu, Shangtao Liu¹, Tashi Lokyitsang¹, Yeshi Lokyitsang¹, Rakela Lubonja¹, Annie Lui¹, Pen MacDonald¹, Vasilisa Magnisalis¹, Kebede Maru¹, Charles Matthews¹, William McCusker¹, Susan McDonough¹, Teena Mehta¹, James Meldrim¹, Louis Meneus¹, Oana Mihai¹, Atanas Mihalev¹, Tanya Mihova¹, Rachel Mittelman¹, Valentine Mlenga¹, Anna Montmayeur¹, Leonidas Mulrain¹, Adam Navidi¹, Jerome Naylor¹, Tamrat Negash¹, Thu Nguyen¹, Nga Nguyen¹, Robert Nico¹, Choe Norbu¹, Nyima Norbu¹, Nathaniel Novod¹, Barry O'Neill¹, Sahal Osman¹, Eva Markiewicz¹, Otero L. Oyono¹, Christopher Patti¹, Pema Phunkhang¹, Fritz Pierre¹, Margaret Priest¹, Sujaa Raghuraman¹, Filip Rege¹, Rebecca Reyes¹, Cecil Rise¹, Peter Rogov¹, Keenan Ross¹, Elizabeth Ryan¹, Sampath Settipalli¹, Terry Shea¹, Ngawang Sherpa¹, Lu Shi¹, Diana Shih¹, Todd Sparrow¹, Jessica Spaulding¹, John Stalker¹, Nicole Stange-Thomann¹, Sharon Stavropoulos¹, Catherine Stone¹, Christopher Strader¹, Senait Tesfaye¹, Talene Thomson¹, Yama Thoulutsang¹, Dawa Thoulutsang¹, Kerri Topham¹, Ira Topping¹, Tsamla Tsamla¹, Helen Vassiliev¹, Andy Vo¹, Tsering Wangchuk¹, Tsering Wangdi¹, Michael Weiland¹, Jane Wilkinson¹, Adam Wilson¹, Shailendra Yadav¹, Geneva Young¹, Qing Yu¹, Lisa Zembek¹, Danni Zhong¹, Andrew Zimmer¹ & Zac Zwirko¹

Broad Institute Whole Genome Assembly Team members David B. Jaffe¹, Pablo Alvarez², Will Brockman¹, Jonathan Butler¹, CheeWhye Chin¹, Sante Gnerre¹ & Iain MacCallum

Affiliation for participants: ¹Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.

METHODS

WGS sequencing and assembly. Approximately 38.8 million high-quality sequence reads were derived from paired-end reads of 4- and 10-kb plasmids, fosmid and BAC clones, prepared from primary tissue DNA from a single female opossum. The reads were assembled using an interim version of ARACHNE2+ (<http://www.broad.mit.edu/wga/>). No comparative data were used in the assembly process. An intermediate assembly (monDom4) was used for the majority of the analyses reported here. The most recent version (monDom5) has identical sequence content and scaffold structure, but includes additional FISH data as described in Supplementary Note 2.

SNP discovery. The SNP discovery was performed using ARACHNE by comparison of the two haplotypes derived from the opossum assembly using only high-quality discrepancies supported by two or more reads each. Sequence reads from three additional individuals were also aligned to the reference assembly, and SNPs were discovered using SSAHA-SNP⁹⁹. Linkage disequilibrium was assessed using Haploview¹⁰⁰.

Genome alignment and comparisons. The assembly versions used in all comparative analyses were hg17 or hg18 (human), mm8 (mouse), rn4 (rat), canFam2 (dog), monDom4 or monDom5 (opossum) and galGal3 (chicken). The number of aligned nucleotides was counted directly from unfiltered, pairwise BLASTZ alignments (obtained from <http://genome.ucsc.edu>). Synteny maps were generated using standard methods^{7,10}, starting from 320,000 reciprocal-best syntenic anchors identified by PatternHunter¹⁰⁴ (see Supplementary Note 7). Reconstruction of the boreoeutherian ancestral karyotype is described in Supplementary Note 8.

Gene prediction and phylogeny. Opossum protein-coding and non-coding RNA genes were predicted using a modified version of the Ensembl genebuild pipeline¹⁰¹, followed by several rounds of refinement using Exonerate¹⁰² and manual curation. Orthology and paralogy were inferred using the PhyOP pipeline with all predicted opossum and human (Ensembl v40) gene transcripts as input and K_S as the distance metric^{11,34}. Coding regions were aligned according to their amino acid sequences using BLASTP. K_A and K_S were estimated using the codeml program¹⁰⁵, with default settings and the F3X4 codon frequency model. Functional categories were identified using the Gene Ontology¹⁰⁶.

Conserved element prediction. Amniote conserved elements were inferred directly from pairwise BLASTZ alignments of chicken to opossum or human. Every alignment block with more than 75% identity for ≥ 100 bp was classified as an amniote conserved element. Eutherian conserved elements were inferred using phastCons¹⁴ on BLASTZ/MULTIZ^{107,108} alignments of human to mouse, rat and dog. The nonconserved model was fitted to fourfold degenerate sites from 15,900 human RefSeqs projected onto the same alignments, using phyloFit and REV. A separate model was fitted for the X chromosome. The scaling parameter for the conserved model was estimated by phastCons. Target coverage and expected element length were set to 12.5% and 12 bp, respectively. Predicted

eutherian conserved elements that did not fall within a 10-kb or longer synteny 'net'¹⁰³ between human, mouse and dog were ignored. The coding status of each element was inferred from ≥ 1 nucleotide overlap with entries in the UCSC human 'known genes' track¹⁰⁹. Proportions are reported out of the total length of the elements considered. Eutherian CNEs were classified as transposable-element-derived if they showed more than 20% nucleotide overlap (median = 100% for all elements, 54% for elements with \log_2 -odds score ≥ 60) with human RepeatMasker annotations.

Phylogeny of conserved elements. For amniote conserved elements, pairwise best-in-genome BLASTZ alignments of opossum to human and vice versa were used to infer their phylogenetic distributions. For eutherian conserved elements, concomitant BLASTZ/MULTIZ alignments to opossum and chicken were used. A conserved element was called absent from a species if it was not covered by a single aligned nucleotide in the relevant BLASTZ alignment.

Correction for assembly gaps and initial alignment artefacts. A conserved element was considered to be in an ungapped syntenic interval if it was flanked by two PatternHunter synteny anchors within 200-kb of each other on the same contigs in both the human and opossum assemblies. All conserved elements (represented by human or opossum, as appropriate) in ungapped syntenic intervals were realigned to the unmasked genome sequence (in opossum or human) using the water program (<http://emboss.sourceforge.net>) with default parameters and a gap extension penalty of 4. A randomly permuted version of each element was also realigned. For amniote conserved elements, only the longest interval with $\geq 75\%$ identity from within the originating alignment block (see above) was realigned. Amniote elements were called lost, and eutherian elements were called eutherian-specific if their Smith–Waterman realignment score, divided by the length of the element, did not exceed the corresponding score for the permuted element plus one. (Conservatively calling an element found if its score simply exceeded the score of the permuted element resulted in 15% of eutherian CNEs in ungapped regions and 8% of those with \log_2 -odds score ≥ 60 being called eutherian-specific.) Putatively eutherian-specific elements, including *XIST*, were also searched against all opossum sequencing reads using discontinuous MegaBLAST.

104. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445 (2002).

105. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).

106. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).

107. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).

108. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).

109. Hsu, F. *et al.* The UCSC known genes. *Bioinformatics* **22**, 1036–1046 (2006).

Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites

Xiaohui Xie[†], Tarjei S. Mikkelsen^{†‡}, Andreas Gnirke[†], Kerstin Lindblad-Toh[†], Manolis Kellis^{†§}, and Eric S. Lander^{†||††}

[†]Broad Institute of MIT and Harvard, Massachusetts Institute of Technology and Harvard Medical School, Cambridge, MA 02142; [‡]Division of Health Sciences and Technology, [§]Computer Science and Artificial Intelligence Laboratory, and ^{||}Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and ^{††}Whitehead Institute for Biomedical Research, Cambridge, MA 02142

Contributed by Eric S. Lander, March 3, 2007 (sent for review January 26, 2007)

Conserved noncoding elements (CNEs) constitute the majority of sequences under purifying selection in the human genome, yet their function remains largely unknown. Experimental evidence suggests that many of these elements play regulatory roles, but little is known about regulatory motifs contained within them. Here we describe a systematic approach to discover and characterize regulatory motifs within mammalian CNEs by searching for long motifs (12–22 nt) with significant enrichment in CNEs and studying their biochemical and genomic properties. Our analysis identifies 233 long motifs (LMs), matching a total of $\approx 60,000$ conserved instances across the human genome. These motifs include 16 previously known regulatory elements, such as the histone 3'-UTR motif and the neuron-restrictive silencer element, as well as striking examples of novel functional elements. The most highly enriched motif (LM1) corresponds to the X-box motif known from yeast and nematode. We show that it is bound by the RFX1 protein and identify thousands of conserved motif instances, suggesting a broad role for the RFX family in gene regulation. A second group of motifs (LM2*) does not match any previously known motif. We demonstrate by biochemical and computational methods that it defines a binding site for the CTCF protein, which is involved in insulator function to limit the spread of gene activation. We identify nearly 15,000 conserved sites that likely serve as insulators, and we show that nearby genes separated by predicted CTCF sites show markedly reduced correlation in gene expression. These sites may thus partition the human genome into domains of expression.

comparative genomics | conserved noncoding element

Comparative analysis of the human and several other mammalian genomes has revealed that 5% of the human genome is under purifying selection, with less than one-third of the sequences under selection encoding proteins. The vast majority lies in hundreds of thousands of conserved noncoding elements (CNEs). The functional significance of these CNEs is largely unknown. It seems likely that many are involved in gene regulation, and transgenic experiments have identified some CNEs that are capable of driving highly specific spatiotemporal gene expression patterns (1–4). However, little is known about regulatory motifs contained within CNEs or proteins that recognize these elements.

We and others have previously undertaken large-scale efforts to discover conserved motifs in limited subsets of the human genome (5–8), specifically, gene promoters and 3'-UTRs. The approach has been to search for motifs that are preferentially conserved in these regions by using syntenic alignments of human, mouse, rat, and dog sequences (5). Using this approach we have discovered 174 motifs in promoter regions (within 2 kb of the transcriptional start), most of which are involved in transcriptional regulation and in tissue-specific gene expression control, and 105 motifs in 3'-UTRs, implicated in posttranscriptional regulation with half related to microRNA targeting. These studies were limited in scope because gene promoters and 3'-UTRs contain only a small fraction ($\approx 6\%$) of the CNEs in the genome. In addition, they were limited in power

because they involved comparison with only three non-human mammals.

Here we use the recent availability of sequences of 12 mammalian genomes to extend our motif discovery efforts to the entire human genome. We focus specifically on long regulatory motifs, between 12 and 22 nt, which provide a strong signal for motif discovery. We searched for motifs that are enriched in CNE regions relative to the rest of the genome.

We discovered >200 motifs showing striking enrichment in CNE regions. The analysis automatically rediscovered a dozen previously known regulatory elements. More importantly, most of the discovered motifs are new and show properties distinct from typical promoter elements. In particular, one of the novel motifs defines $\approx 15,000$ potential insulator elements in the human genome, highlighting the diverse role of the CNEs in gene regulation.

Results

Creating a Motif Catalog. We began by compiling a data set of 829,730 CNEs in the human genome (totaling 62 Mb or $\approx 2\%$ of the euchromatic genome), consisting of sequences showing strong conservation in syntenic regions in comparisons involving 12 mammalian genomes [see supporting information (SI) *Text*]. The vast majority of these elements are located at a considerable distance from the transcriptional start sites (TSS) of protein-coding genes (SI Fig. 4). Approximately 95% are located >2 kb away from the TSS of any gene, and half are >100 kb from a TSS. This suggests that only a small portion of the CNEs serve functions specific to core or proximal promoters.

We sought to create a catalog of sequence motifs enriched in the CNEs (SI Fig. 5). We began by identifying k -mers (for $k \geq 12$) that occur at a significantly higher frequency in the CNE sequences than in the remainder of the genome. We focused only on relatively long k -mers, because the expected number N of random occurrences in the entire CNE database is small (for example, $n < 8$ for $k = 12$; SI Fig. 6). We identified a total of 69,810 enriched k -mers. An example is 5'-GTTGCCATGGAAAC-3', which appears 698 times in the CNE data set, whereas only 27 sites are expected based on its genome-wide frequency (26-fold enrichment). We noticed that many of enriched k -mers were closely related; therefore, we clustered them based on sequence similarity. The 69,810 enriched k -mers collapsed into 233 distinct groups, denoted LM1, LM2, etc.

Author contributions: X.X. and E.S.L. designed research; X.X. and E.S.L. performed research; X.X., T.S.M., A.G., K.L.-T., M.K., and E.S.L. contributed new reagents/analytic tools; X.X., T.S.M., A.G., K.L.-T., M.K., and E.S.L. analyzed data; and X.X., T.S.M., A.G., K.L.-T., M.K., and E.S.L. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Abbreviations: CNE, conserved noncoding element; TSS, transcriptional start site; PWM, positional weight matrix; NRSE, neuron-restrictive silencer element.

^{††}To whom correspondence should be addressed. E-mail: lander@broad.mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0701811104/DC1.

© 2007 by The National Academy of Sciences of the USA

Table 1. List of top 50 most highly enriched motifs

ID	k-mer sequence	No. of allowed mismatches	No. of initial sites in CNEs	Fold enrichment	Z-score	Known motif
LM1	GTTGCCATGGAAC	1	698	25.9	130.0	X-box
LM2	ACCACTAGATGGCA	1	305	22.9	80.4	
LM3	GTTGCTAGGCAACC	1	204	30.7	76.9	
LM4	GCCTGCTGGGAGTTGTAGTT	3	143	26.3	59.2	
LM5	AACTCCCATTAGCGTTAATGG	3	43	68.1	53.5	
LM6	AAAGGCCCTTTAAGGGCCAC	3	48	46.2	46.3	Histone 3'-UTR
LM7	CAGCAGATGGCGCTGTT	2	97	22.1	44.4	
LM8	ATGAATTATTCATG	1	280	8.8	44.3	
LM9	TCAGCACCACGGACAG	1	82	25.6	44.2	NRSE
LM10	CTGTTTCCTTGGAAACCAG	3	165	9.3	35.2	
LM11	GAAATGCTGACAGACCCTTAA	3	41	30.7	34.5	
LM12	TGGCCTGAAAGAGTTAATGCA	3	51	22.8	32.7	
LM13	TGCTAATTAGCA	0	82	13.1	30.4	CHX10
LM14	ATCCAGATGTTGGCA	1	33	27.0	28.9	RP58
LM15	CATTTGCATGCAAATGA	2	124	8.5	28.8	
LM16	TTGAGATCCTTAGATGAAAG	3	64	14.6	28.6	
LM17	CATCTGGTTTGCAT	1	117	8.8	28.5	
LM18	CATTTGCATCTGATTGTCAT	3	80	11.8	28.2	
LM19	TGCTAATTAGCAGC	1	88	10.8	28.1	
LM20	TGACAGCTGTCAA	1	118	8.5	28.0	
LM21	ATTTGCATCTCATTTC	2	123	8.2	27.9	
LM22	CAGCTGTTAAACAGCTG	2	80	11.4	27.6	
LM23	AGCACCACCTGGTGGTA	2	65	13.4	27.5	
LM24	AGAACAGATGGC	0	70	12.1	26.8	TAL1BETAIF2
LM25	AAAAGCAATTCCT	1	202	5.3	26.7	
LM26	TAAACACAGCTG	0	83	10.2	26.3	
LM27	CATTTGCATCTCATTAGCA	3	110	8.0	26.1	
LM28	AGAACATCTGTTTC	1	144	6.3	25.5	
LM29	GCTAATTGCAAATG	1	98	8.4	25.3	
LM30	CTTTGAAATGTCAA	1	182	5.3	25.3	
LM31	CTTTTCATCTTCAAAGCACTT	3	57	13.0	25.2	
LM32	CTGACATTTCCAAA	1	174	5.4	25.0	
LM33	GTAATTGGAAACAGCTG	2	69	10.7	24.8	
LM34	GATTTGCATTGCAAATG	2	84	8.8	24.1	
LM35	ACTTCAAAGGGAGC	1	87	8.5	24.1	
LM36	GAAATGCAATTTGC	1	125	6.4	24.1	
LM37	ATGCAAATGAGCCC	1	85	8.5	23.9	
LM38	GCAAATTAGCAGCT	1	82	8.5	23.4	
LM39	GTCTCCTAGGAAAC	1	84	8.4	23.4	
LM40	TCCCATTGACTTCAATGGGA	3	44	14.2	23.4	
LM41	TTTGAAATGCTAATG	1	80	8.6	23.2	
LM42	AAGCCTAATTAGCA	1	69	9.6	23.1	
LM43	CAGGAAATGAAA	0	141	5.6	23.1	
LM44	GTGTAATTGGAAACAGCTG	3	75	8.9	23.0	
LM45	GCTAATTGGATTG	1	76	8.7	22.9	
LM46	AACAGCTGTTGAAA	1	128	5.9	22.9	
LM47	AGAGTGCCACCTACTGAAT	3	65	9.8	22.7	
LM48	TAATGAGCTCATTA	1	108	6.5	22.6	
LM49	GTAATTAGCAGCTG	1	68	9.3	22.5	
LM50	TGGGTAATTACATTCTG	2	65	9.6	22.5	

for “long motif.” For each of these motifs we derived a positional weight matrix (PWM) representation reflecting the distribution of 4 nt at each position. The enrichment of each motif in the CNE data set was expressed in an enrichment score (see *Methods*). The top 50 motifs are shown in Table 1, and a full list of 233 motifs is given in *SI Table 3*. The motifs range in size from 12 to 22 bases.

For each of 233 discovered motifs, we searched the entire human genome to identify conserved instances; that is, we identified all human sites matching the PWM and then found those sites that show clear cross-species conservation (see *SI Text*). We found a total of 60,019 conserved instances, with roughly half residing within the CNE data set and roughly half in the remainder of the genome. Importantly, the approach of focusing on motifs enriched in the CNE data set identified many motif instances elsewhere in the genome.











To assess the significance of these results, the procedure was repeated with matched control motifs. For each of the 233 motifs

we created a control motif by permuting the columns of the PWM while preserving the occurrence of CpG dinucleotides. These control motifs have only 3,081 conserved instances, which is 20-fold lower than for the discovered motifs. These results indicate that only a small fraction of the 60,019 instances of the discovered motifs are likely to have occurred purely by chance.

The number of conserved instances is highly uneven across the motifs (range 37–7,549, with mean of 266 and median of 61). Most motifs (67%) have <100 conserved instances (*SI Table 3*). But, remarkably, the two motifs with the highest enrichment scores, LM1 and LM2, both have >5,000 conserved instances in the human genome (Table 2), suggesting a widespread functional role for these elements.

Characterizing the Discovered Motifs. Known regulatory elements. Among the 233 discovered motifs, 16 match known regulatory elements (Table 1). For example, the LM9 motif is nearly identical

Table 2. Properties of the top 10 discovered motifs

ID	Motif profile	No. of conserved instances	False positive rate*	Conservation rate, † %	Fold increase in conservation rate‡	Correlation between cross-species conservation and motif profile	Positional bias around TSS§
LM1		5,332	0.050	29.3	9.5	0.92	
LM2		7,549	0.048	29.4	14.0	0.91	
LM3		844	0.048	40.1	14.3	0.94	
LM4		1,877	0.046	20.3	13.5	0.89	20.3
LM5		224	0.042	19.4	16.3	0.87	
LM6		79	0.026	20.1	10.1	0.81	25.5
LM7		6,302	0.048	21.6	10.3	0.72	
LM8		608	0.047	17.2	9.6	0.68	
LM9		1,443	0.039	11.8	8.4	0.90	6.1
LM10		5,914	0.050	14.5	6.6	0.77	

*The proportion of conserved instances expected to have occurred by chance.

†The proportion of instances detected in human that are also conserved in orthologous regions of other mammals.

‡Compared to the conservation rates of control motifs.

§Fold enrichment on the number of motif sites located within 1 kb of TSS over those for control motifs. Only motifs with fold enrichment above 4 are shown.

to the consensus sequence of the neuron-restrictive silencer element (NRSE). The NRSE is recognized by the transcription factor REST (RE1 silencing transcription factor), which plays a pivotal role in repressing the expression of neuronal genes in nonneuronal tissues (9–11). Between 800 and 1,900 NRSE sites have been estimated to exist in the human genome (12, 13), which is consistent with our count of the number of conserved instances (1,443). It is reassuring to note that our procedure identified the LM9 motif without any prior knowledge, recovering the correct size of NRSE and showing nearly perfect similarity to NRSE along all of its 21 positions (SI Fig. 7).

Another example is LM6, which is a well studied RNA motif that is present exclusively in the 3'-UTRs of genes encoding histone proteins. In histone mRNAs this sequence is known to fold into a stem-loop structure involved in posttranscriptional regulation, playing a role similar to poly(A) tails on typical mRNAs (14, 15).

Conservation properties. The discovered motifs have two notable conservation properties. First, they show a much higher conservation rate than the control motifs, even outside the CNEs where they were discovered. The conservation rate was defined as the ratio of conserved instances to total instances in the human genome. All of the discovered motifs have a conservation rate that is 2-fold higher than for their matched controls, and 65% have a rate that is 5-fold

higher (SI Table 4). If the conservation rate is computed based only on motif instances outside the CNEs, 96% of the discovered motifs have a conservation rate that is 2-fold higher than for their controls, and 63% have a rate that is 5-fold higher.

Second, the motifs show similar patterns of cross-species conservation and within-species conservation (Fig. 1 *a* and *b*). For each motif we asked whether the most conserved positions across the various motif instances within human (and thus those most likely to be involved in motif recognition) are also the positions within individual instances that show the highest conservation across species (and thus are most constrained in their evolution). To measure the within-species conservation of a motif we used the information content (I_k) of its PWM at the position k . To quantify its cross-species conservation we identified its instances located within the CNEs and calculated the proportion (M_k) of the instances with bases not mutated in the orthologous regions of the mouse or dog genomes at position k of the motif. The correlation coefficient between I and M for each of the discovered motifs is shown in Fig. 1c. We found that nearly all motifs (95%) show a positive correlation, and 53% have correlation coefficient >0.5 . This suggests that the discovered motifs are indeed functional. The results also suggest that these motifs retain similar recognition properties across species.

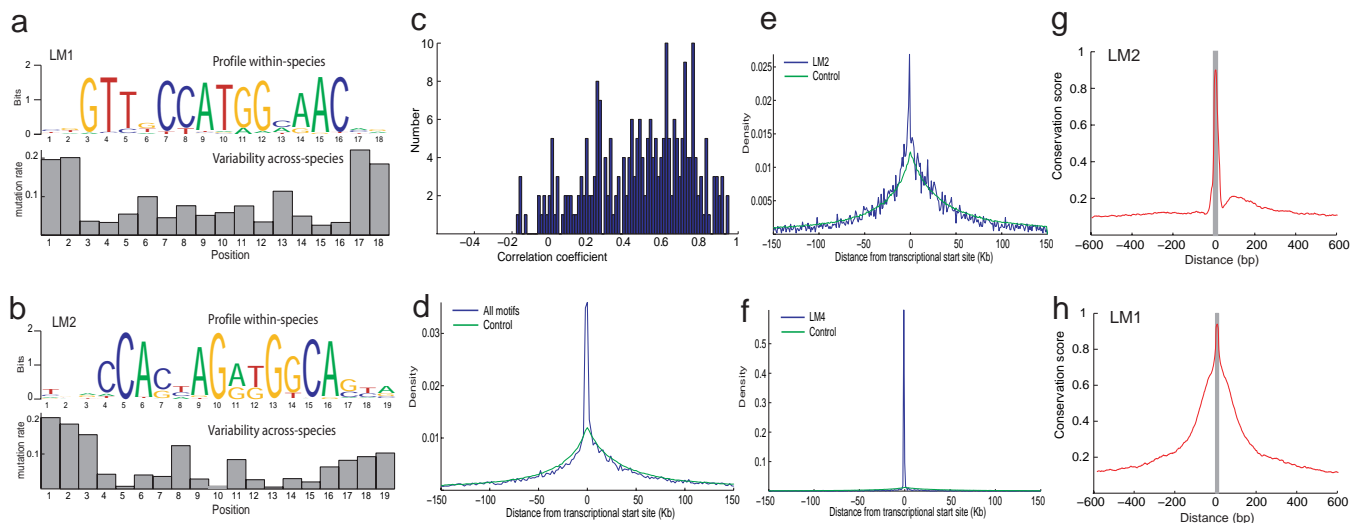


Fig. 1. A summary on properties of the discovered motifs. (a and b) Motif profile within species and variability across species for LM1 (a) and LM2 (b). Positions with high information content are less variable than those with low information content in cross-species comparison. (c) The correlation coefficients between motif profile and across-species conservation pattern for all discovered motifs. (d–f) The location of motif sites relative to TSS for all motifs (d), LM2 (e), and LM4 (f). They demonstrate that most of the discovered motifs, in particular LM2, are broadly distributed relative to TSS, not much different from the distribution of control sites randomly drawn from the genome (green lines). (h) Sequences surrounding LM1 sites are also conserved, in contrast to those surrounding LM2 sites (g). Gray bars show locations of the motifs. Conservation scores are phastCons scores (28) averaged over motif sites.

Palindromes. A significant proportion (17%) of the 233 motifs are palindromes, forming perfect or nearly perfect matches to their reverse complement over nearly their entire length. For example, LM3 consists of GTTGCY juxtaposed with its reverse complement, RGCAAC, with a central W, itself a self-palindrome ($W = A/T$). The proportion of palindromes is much higher than for random control sequences (0.13%) (see *SI Text*) and is similar to the proportion seen for the 16 known motifs (18%). The enrichment is especially pronounced among the 20 top-scoring motifs, with 45% being palindromic. Notably, the palindromic motifs are also symmetric in the information content of each base, and weakly specified positions are symmetrically placed with weakly specified positions on the two motif halves. Palindromicity can be indicative of DNA sequences that bind by a protein homodimer. Alternatively, palindromicity can sometimes reflect RNA sequences that form stem-loop structures, as illustrated by the LM6 motif in the 3'-UTRs of histone genes.

Distance from transcriptional starts. Most of the discovered motifs show little or no enrichment near genes. More than 93% have 80% of their conserved instances located >10 kb away from the TSS of any gene (Fig. 1d). A typical example is the LM2 motif (Fig. 1e). Most of these motifs are likely not to be related to core and proximal promoter functions, but may instead encode distal regulators, insulators, or other functions.

There are five cases, however, with a strong preference for being located near gene starts. A striking example is the LM4 motif, for which $\approx 60\%$ of the conserved instances lie within 1 kb of the TSS (26-fold enriched over random expectation) and the modal distance is 75 bases upstream of the TSS (Fig. 1f). Another example is LM100, a palindromic sequence for which 45% of conserved instances lie within 2 kb of a TSS. These motifs are likely to be related to core and proximal promoter functions.

Local conservation context. We studied the conservation context of the discovered motifs. Because CNE sequences used to discover the motifs tend to occur in large blocks (N50 length = 110 bases, where N50 length is the length x such that 50% of all CNE bases lie in CNEs with the size $\geq x$), conserved motif occurrences lying within CNEs would be expected to be embedded within blocks of conserved sequence. This is indeed the case. For each motif M we examined the block of conserved sequence surrounding the each

conserved occurrence in a CNE and defined $d_1(M)$ to be the N50 length of the block. The median value of $d_1(M)$ is 112 bases, with an interquartile range of 88–140 bases.

More revealingly, we examined the corresponding value $d_2(M)$ defined for conserved motif instances that lie outside the CNEs data set. The median value of $d_2(M)$ is 96 bases (interquartile range of 61–133 bases), which is similar to $d_1(M)$. This indicates that the discovered motifs typically function as part of regulatory modules containing many other regulatory elements. These results suggest that CNE motifs here may provide a useful initial entry point for studying the function of diverse large CNEs, including ultra-conserved elements that have been shown to have enhancer function.

Although most motifs appear to function in concert with others, we found eight striking examples among the 233 motifs that appear to act in isolation. This is true both for conserved occurrences within and outside the CNE data set. These motifs are LM9 (NRSE), LM6 (the histone 3'-UTR element), LM4 (a promoter-proximal motif), and four unknown motifs: LM2, LM7, LM23, and LM194. (We show below that LM2, LM7, and LM23 correspond to CTCF binding sites.) The median lengths of surrounding conserved sequences for these motifs are all less than five flanking bases on each side. For example, LM2 has only a median two flanking conserved bases on each side (Fig. 1g), whereas LM1 (Fig. 1h) has a median of 31 flanking conserved bases on each side.

LM1 Defines RFX Binding Sites. The most highly enriched motif LM1 is similar to the X-box motif, which has been extensively studied in yeast and nematodes (16–18). In yeast, more than three dozen X-box sites have been identified, and these sites have been shown to be bound by the Crt1 protein, an effector of the DNA damage checkpoint pathway (19). In *Caenorhabditis elegans*, >700 X-box sites have been computationally predicted, and several dozen of these sites have been demonstrated to be recognized by the DAF-19 protein, which is known to regulate genes involved in the development of sensory cilia (16, 18).

Individual instances of the X-box motif in vertebrates have been reported, but no systematic survey of X-box motifs in the human genome has been conducted. Approximately three dozen such sites have been reported to be bound by RFX family proteins, which are

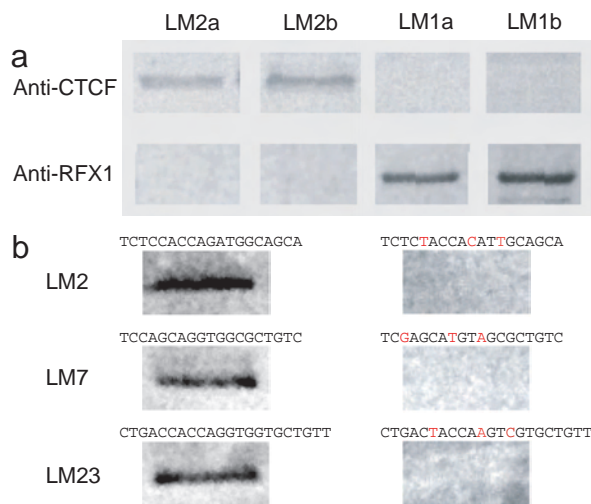


Fig. 2. Confirmation of CTCF and RFX1 binding by *in vitro* affinity capture. (a) CTCF was specifically captured by probes LM2a and LM2b constructed for the LM2 motif, whereas RFX1 was specifically captured by probes LM1a and LM1b constructed for the LM1 motif. (b) The binding of CTCF to LM2, LM7, and LM23 (Left), but not to their corresponding mutant motifs with three core bases altered (Right). See *Methods* for probes used in the experiments.

homologous to both Crt1 and DAF-19 and contain a highly conserved winged helix DNA binding domain. The biochemically characterized consensus sequence for RFX binding shows similarity to the LM1 motif (20), although it contains less information.

To test whether LM1 binds RFX proteins, we performed an affinity-capture experiment (see *SI Text*). A biotinylated double-stranded DNA probe containing multiple copies of the LM1 motif was incubated with HeLa cell nuclear extract and then captured with streptavidin. The bound protein was electrophoresed, blotted, and probed with an antibody against RFX1, a prototypical member of the RFX family, revealing that the protein indeed specifically binds LM1 (Fig. 2*a*).

LM2 Defines a Common Insulator Site Across the Human Genome. The most interesting case among the 233 discovered motifs is LM2. It has the largest number of conserved instances (7,549) in the genome, with the vast majority being located far from TSSs (Fig. 1*e*). The LM2 motif is 19 bases in length and does not match the reported consensus sequence of any known motif.

We obtained a hint regarding the possible function of the LM2 motif by using proteomic experiments in which HeLa cell nuclear extract was subjected to affinity capture with a biotinylated double-stranded DNA probe containing multiple copies of the LM2 motif, and the resulting material was analyzed by protease digestion and mass spectrometry. These affinity-capture experiments suggested that the CTCF protein binds the LM2 motif (unpublished data).

CTCF, a protein containing 11 zinc-finger domains, is a major factor implicated in vertebrate insulator activities (21–23). An insulator is a DNA sequence element that prevents a regulatory protein binding to the control region of one gene from influencing the transcription of neighboring genes. When placed between an enhancer and a promoter, an insulator can block the interaction between the two. Several dozen insulator sites have been characterized, and almost all have been shown to contain CTCF binding sites. In some cases, the CTCF site has been directly shown to be both necessary and sufficient for enhancer blocking activities in heterologous settings. The known CTCF sites show considerable sequence variation, and no clear consensus sequence has been derived (22). The well studied CTCF sites in the IGF2/H19 locus show similarity to the LM2 motif (24), although the similarity score is below our threshold used for detecting LM2 sites.

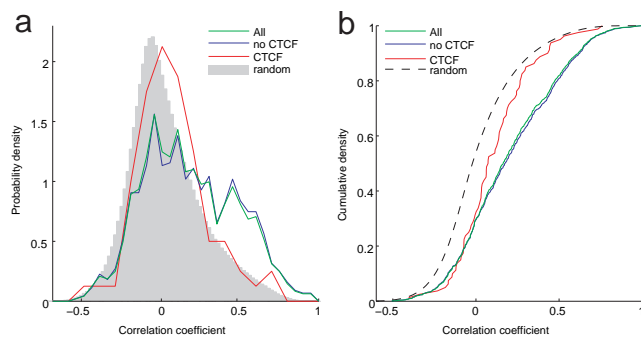


Fig. 3. Genes separated by predicted CTCF sites are less correlated in gene expression. Correlation coefficient between neighboring gene pairs is shown in terms of probability density (a) and cumulative distribution (b). Green line, correlation between all neighboring genes; red line, correlation between genes separated by at least one CTCF site; gray shading, correlation between randomly chosen gene pairs.

To test directly whether CTCF binds the LM2 motif we analyzed the material obtained by affinity capture with a biotinylated double-stranded DNA probe containing multiple copies of the LM2 motif by immunoblotting with an antibody against the human CTCF protein (see *SI Text*). The results confirmed that CTCF does indeed bind the LM2 motif (Fig. 2). By contrast, mutation of the three core positions with the highest information content (positions 5, 10, and 13 of LM2) (Fig. 1*b*) completely abolished the binding of the CTCF protein.

Given the sequence diversity among reported CTCF sites, we searched for additional motifs in our catalog that show substantial similarity to LM2. The motifs LM7 and LM23 are nearly identical in their first 14 positions, diverging only in the last four or five bases (*SI Fig. 8*). The two additional motifs also have an unusually large number of conserved instances (6302 for LM7 and 3758 for LM23). Affinity-capture experiments using probes containing copies of the LM7 and LM23 motifs demonstrated that both motifs bind CTCF, whereas mutation of the three core positions with the highest information content completely abolish binding (Fig. 2*b*). The three motifs, LM2, LM7, and LM23, will be referred to as a “supermotif,” LM2*.

Altogether the LM2* motif has 14,987 conserved instances in the human genome (which is 20-fold higher than for the corresponding control motifs). Strikingly, this comprises approximately one-fourth of the 60,019 sites for the complete catalog of 233 motifs. We propose that the vast majority of these sites are CTCF-binding sites and function as insulators.

Although the predicted CTCF sites tend to be located far from gene starts, they are not randomly distributed across the genome. Instead, their distribution closely follows the distribution of genes, with a correlation coefficient of 0.6 (*SI Fig. 9*). This is consistent with the notion that the sites are related to gene regulation, rather than, for example, chromosomal structure.

We sought to test whether the predicted CTCF sites actually serve as functional insulators. Although it is possible to perform insulator assays on individual instances in a heterologous context, we were interested to assess the function of many CTCF sites in their natural context. If the predicted CTCF sites actually function as insulators, we reasoned that the presence of a CTCF site between two genes might “decouple” their gene expression.

It is known that divergent gene pairs, transcribed in opposite directions with transcription start sites close to each other, tend to show correlated gene expression patterns (25, 26). We therefore assembled a data set of 963 divergent gene pairs with intergene distance <20 kb and with expression values measured across 75 human tissues (27). As expected, the divergent gene pairs are more closely correlated in gene expression than randomly chosen gene

pairs (Fig. 3). When the cases are divided into gene pairs separated by a CTCF site (CTCF pairs, 80 cases) and those not separated by a CTCF site (non-CTCF pairs, 883 cases), the former show correlations that are essentially equivalent to the random background. Overall, 37% of non-CTCF pairs are strongly correlated (correlation coefficient $\rho > 0.3$). This proportion is 2-fold higher than the proportion of random genes pairs (12%) showing similarly strong correlation. By contrast, the proportion of CTCF pairs with similarly strong correlation is 16%, which is close to that seen for random gene pairs. This difference persists after correcting for small difference in the lengths of CTCF-containing and CTCF-non-containing intergenic regions (SI Fig. 10). This provides strong evidence that the majority of the predicted CTCF sites do indeed function as insulators.

Finally, we examined the frequency of the CTCF motif LM2* across various vertebrate genomes. The three motifs all occurred frequently in all eutherian mammals, opossum, chicken, and the pufferfish *Tetraodon*. The motif shows a similar total number of instances across all vertebrate species despite a 5-fold variation in genome size (SI Fig. 11). This is consistent with the LM2* motif being related to gene number (which is fairly constant across these species) rather than genome size.

Discussion

Our analysis provided an initial systematic catalog of regulatory motifs in the conserved regions of the entire human genome. The 233 discovered motifs are highly enriched in the CNE sequence, with all being at least 5-fold enriched relative to the rest of the genome. These motifs match 60,019 conserved instances in the human genome, with a typical motif having ≈ 100 conserved instances. Among the 233 discovered motifs, only 16 could be recognized as previously known regulatory elements, indicating that much more still remains to be learned about the function of CNE.

The most interesting unknown motif is LM2, which has $\approx 7,500$ conserved instances in the genome, more abundant than any other discovered motif. We used affinity-capture assays to demonstrate that LM2, as well as two other closely related motifs, LM7 and LM23, are specifically bound by the CTCF protein, which is involved in insulator function. Together, the three motifs match nearly 15,000 conserved instances in the human genome, corresponding to approximately one-fourth of all matching instances for the entire set of discovered motifs. Although we cannot rule out that CTCF protein can also bind to other, highly dissimilar sites, our findings suggest that a few dominant CTCF motifs are extremely enriched throughout the human genome.

The results here are, of course, only a step toward comprehensive catalog of regulatory motifs across the human genome. In particular, our analysis used stringent threshold to identify only the most highly enriched motifs in the CNEs and therefore have omitted short motifs (e.g., 6–8 nt). Additionally, the current study focused primarily on motifs present in most mammals, and therefore many lineage-specific motifs, such as those unique to primates, still remain to be discovered. The power of motif discovery can be boosted not just by considering the enrichment of sequences in the human CNEs, but by exploiting their detailed conservation pattern across different species. With the availability of genome sequences from an increasing number of related mammals, it should be possible to create a complete dictionary of human motifs in the years ahead.

Methods

We started by enumerating a list of candidate k -mers with $12 \leq k \leq 22$ and counting the number (C) of matching instances of each k -mer present in the CNE data set (SI Fig. 5). A sequence was declared a match to a k -mer if the number of mismatches between the sequence and the k -mer was less than a threshold M (where $M = 0$ for $k \leq 13$; $M = 1$ for $k = 14, 15$, or 16 ; $M = 2$ for $k = 17$ or 18 ; and $M = 3$ for $k \geq 19$). For k -mers with $C \geq 30$ we identified all matching instances in the entire human genome and assessed their enrichment in the CNEs using two scores: (i) fold enrichment: $SNR = C/\mu$; and (ii) Z -score = $(C - \mu)/\sqrt{\mu}$, quantifying the significance of the enrichment. Here, μ is the expected number of matching instances within the CNE data set based on the observed frequency of matches in the overall genome. Finally, we collected all k -mers with $SNR \geq 5$ and Z -score ≥ 10 , resulting in a total of 69,810 k -mers significantly enriched in the CNEs. The probability of a k -mer having a Z -score ≥ 10 at random is $< 2 \times 10^{-12}$, and, after Bonferroni correction, the expected number of such k -mers is $< 10^{-4}$. These k -mers were further clustered and grouped into 233 distinct motifs according to the procedure described in ref. 5. For each motif we selected the k -mer with the highest Z -score to represent the motif.

Please see SI Text for additional methods.

We thank David Jaffe, Manuel Garber, and Sarah Calvo for insightful comments and suggestions. We gratefully acknowledge Xiaolan Zhang, Li Wang, Jacob Jaffe, and Steve Carr for assistance with affinity-capture experiments. This work was supported in part by grants from the National Human Genome Research Institute (to E.S.L.) and from the Broad Institute.

- Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. (2006) *Nature* 444:499–502.
- Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W, Haussler D (2006) *Nature* 441:87–90.
- Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. (2005) *PLoS Biol* 3:e7.
- Dermitzakis ET, Reymond A, Antonarakis SE (2005) *Nat Rev Genet* 6:151–157.
- Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) *Nature* 434:338–345.
- Elemento O, Tavazoie S (2005) *Genome Biol* 6:R18.
- Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, Birney E (2005) *Genome Biol* 6:R104.
- Jones NC, Pevzner PA (2006) *Bioinformatics* 22:e236–242.
- Ballas N, Mandel G (2005) *Curr Opin Neurobiol* 15:500–506.
- Chong JA, Tapia-Ramirez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC, Altshuler YM, Frohman MA, Kraner SD, Mandel G (1995) *Cell* 80:949–957.
- Schoenherr CJ, Anderson DJ (1995) *Science* 267:1360–1363.
- Mortazavi A, Thompson EC, Garcia ST, Myers RM, Wold B (2006) *Genome Res* 16:1208–1221.
- Bruce AW, Donaldson IJ, Wood IC, Yerbury SA, Sadowski MI, Chapman M, Gottgens B, Buckley NJ (2004) *Proc Natl Acad Sci USA* 101:10458–10463.
- Williams AS, Marzluff WF (1995) *Nucleic Acids Res* 23:654–662.
- Pandey NB, Marzluff WF (1987) *Mol Cell Biol* 7:4557–4559.
- Blacque OE, Perens EA, Boroevich KA, Inglis PN, Li C, Warner A, Khattri J, Holt RA, Ou G, Mah AK, et al. (2005) *Curr Biol* 15:935–941.
- Zaim J, Speina E, Kierzek AM (2005) *J Biol Chem* 280:28–37.
- Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR, Ruvkun G, Thomas JH, Swoboda P (2005) *Development (Cambridge, UK)* 132:1923–1934.
- Huang M, Zhou Z, Elledge SJ (1998) *Cell* 94:595–605.
- Emery P, Strubin M, Hofmann K, Bucher P, Mach B, Reith W (1996) *Mol Cell Biol* 16:4486–4494.
- Bell AC, West AG, Felsenfeld G (1999) *Cell* 98:387–396.
- Ohlsson R, Renkawitz R, Lobanov V (2001) *Trends Genet* 17:520–527.
- Gaszner M, Felsenfeld G (2006) *Nat Rev Genet* 7:703–713.
- Bell AC, Felsenfeld G (2000) *Nature* 405:482–485.
- Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otilar RP, Myers RM (2004) *Genome Res* 14:62–66.
- Li YY, Yu H, Guo ZM, Guo TQ, Tu K, Li YX (2006) *PLoS Comput Biol* 2:e74.
- Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. (2004) *Proc Natl Acad Sci USA* 101:6062–6067.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. (2005) *Genome Res* 15:1034–1050.

A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells

Bradley E. Bernstein,^{1,2,3,*} Tarjei S. Mikkelsen,^{3,4} Xiaohui Xie,³ Michael Kamal,³ Dana J. Huebert,¹ James Cuff,³ Ben Fry,³ Alex Meissner,⁵ Marius Wernig,⁵ Kathrin Plath,⁵ Rudolf Jaenisch,⁵ Alexandre Wagschal,⁶ Robert Feil,⁶ Stuart L. Schreiber,^{3,7} and Eric S. Lander^{3,5}

¹Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA 02129, USA

²Department of Pathology, Harvard Medical School, Boston, MA 02115, USA

³Broad Institute of Harvard and MIT, Cambridge, MA 02139, USA

⁴Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA

⁵Whitehead Institute for Biomedical Research, MIT, Cambridge, MA 02139, USA

⁶Institute of Molecular Genetics, CNRS UMR-5535 and University of Montpellier-II, Montpellier, France

⁷Howard Hughes Medical Institute at the Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA

*Contact: bbernstein@partners.org

DOI 10.1016/j.cell.2006.02.041

SUMMARY

The most highly conserved noncoding elements (HCNEs) in mammalian genomes cluster within regions enriched for genes encoding developmentally important transcription factors (TFs). This suggests that HCNE-rich regions may contain key regulatory controls involved in development. We explored this by examining histone methylation in mouse embryonic stem (ES) cells across 56 large HCNE-rich loci. We identified a specific modification pattern, termed “bivalent domains,” consisting of large regions of H3 lysine 27 methylation harboring smaller regions of H3 lysine 4 methylation. Bivalent domains tend to coincide with TF genes expressed at low levels. We propose that bivalent domains silence developmental genes in ES cells while keeping them poised for activation. We also found striking correspondences between genome sequence and histone methylation in ES cells, which become notably weaker in differentiated cells. These results highlight the importance of DNA sequence in defining the initial epigenetic landscape and suggest a novel chromatin-based mechanism for maintaining pluripotency.

INTRODUCTION

Epigenetic regulation of gene expression is mediated in part by posttranslational modifications of histone proteins,

which in turn modulate chromatin structure (Jenuwein and Allis, 2001; Margueron et al., 2005). The core histones H2A, H2B, H3, and H4 are subject to dozens of different modifications, including acetylation, methylation, and phosphorylation. Histone H3 lysine 4 (Lys4) and lysine 27 (Lys27) methylation are of particular interest as these modifications are catalyzed, respectively, by trithorax- and Polycomb-group proteins, which mediate mitotic inheritance of lineage-specific gene expression programs and have key developmental functions (Ringrose and Paro, 2004). Lys4 methylation positively regulates transcription by recruiting nucleosome remodeling enzymes and histone acetylases (Santos-Rosa et al., 2003; Pray-Grant et al., 2005; Sims et al., 2005; Wysocka et al., 2005), while Lys27 methylation negatively regulates transcription by promoting a compact chromatin structure (Francis et al., 2004; Ringrose et al., 2004).

Various observations suggest that chromatin undergoes important alterations during mammalian development (Delaval and Feil, 2004; Margueron et al., 2005; Sado and Ferguson-Smith, 2005). Embryonic stem (ES) cell differentiation is accompanied by changes in chromatin accessibility at several key developmental genes, including a large-scale opening of the HoxB locus (Chambeyron and Bickmore, 2004; Perry et al., 2004). Furthermore, Polycomb-group proteins play an essential role in maintaining the pluripotent state of ES cells and show markedly reduced expression upon differentiation (O'Carroll et al., 2001; Silva et al., 2003; Valk-Lingbeek et al., 2004). However, little is known about the overall structure of ES cell chromatin, how it is established, or how it contributes to the maintenance of pluripotency (Szutorisz and Dillon, 2005).

Large-scale studies of mammalian chromatin have recently become possible with the combination of

chromatin immunoprecipitation (ChIP) and DNA microarrays. Initial studies in primary fibroblasts revealed thousands of genomic sites associated with Lys4 methylation (Bernstein et al., 2005; Kim et al., 2005). The vast majority show a “punctate” pattern, typically occurring at sites of ~1-2 kb near promoters of active genes. Lys27 methylation is implicated in X chromosome inactivation and imprinting (Plath et al., 2003; Umlauf et al., 2004). However, little is known about the overall genomic distribution of this repressive mark. Gene-specific and limited microarray studies have reported that Lys27 methylation tends to occur at punctate sites near promoters of repressed genes (Cao and Zhang, 2004; Kimura et al., 2004; Kirmizis et al., 2004; Koyanagi et al., 2005). Based on such studies, the distributions of Lys4 and Lys27 methylation have been thought to be nonoverlapping.

A notable exception to the punctate pattern of histone modifications is evident at the Hox gene clusters: These loci contain large, cell type-specific Lys4 methylated regions, up to 60 kb in length, that overlay multiple Hox genes (Bernstein et al., 2005; Guenther et al., 2005). These regions likely reflect accessible chromatin domains established during embryonic development to maintain Hox gene expression programs (Chambeyron and Bickmore, 2004). However, the extent to which large domains of chromatin modifications represent a general feature of mammalian genomes remains unclear.

Recent studies have revealed that the most highly conserved noncoding elements (HCNEs) in mammalian genomes cluster within ~200 HCNE-rich genomic loci, which include all four Hox clusters (Nobrega et al., 2003; Bejerano et al., 2004; Lindblad-Toh et al., 2005; Woolfe et al., 2005). These regions tend to be gene-poor but are highly enriched for genes encoding transcription factors (TFs) implicated in embryonic development.

These findings suggest that the HCNE-rich regions or the TF genes within them may contain key epigenetic regulatory controls involved in development. We explored this by mapping histone methylation patterns in mouse ES cells across 61 large regions (~2.5% of the genome). The results reveal a novel chromatin modification pattern that we term “bivalent domains,” consisting of large regions of Lys27 methylation harboring smaller regions of Lys4 methylation. In ES cells, bivalent domains frequently overlay developmental TF genes expressed at very low levels. Bivalent domains tend to resolve during ES cell differentiation and, in differentiated cells, developmental genes are typically marked by broad regions selectively enriched for either Lys27 or Lys4 methylation. We suggest that bivalent domains silence developmental genes in ES cells while keeping them poised for activation. Finally, we analyzed the relationship between histone methylation and the underlying DNA sequence in both ES and differentiated cells. This analysis suggests that DNA sequence largely defines the initial epigenetic state in ES cells, which is subsequently altered upon differentiation, presumably in response to lineage-specific gene expression programs and environmental cues.

RESULTS

Bivalent Domains in ES Cells Contain Repressive and Activating Histone Modifications

Histone H3 Lys4 and Lys27 methylation patterns in ES cells were examined across a subset of HCNE-rich loci using a combination of ChIP and tiling oligonucleotide arrays. The arrays tile 61 large genomic regions, totaling 60.3 Mb, at a density of approximately one probe per 30 bases (Table S1). The regions consist of the four Hox clusters (1.3 Mb encoding 43 genes), 52 additional HCNE-rich regions (55 Mb encoding 169 genes), and five “control” regions that do not show high HCNE density (4 Mb encoding 95 genes). We isolated genomic DNA associated with either trimethylated Lys4 or trimethylated Lys27 by immunoprecipitating cross-linked chromatin, and we then hybridized these DNA fractions to the tiling arrays (see [Experimental Procedures](#)). We identified regions of Lys4 or Lys27 methylation by comparing these hybridization results to those obtained for total genomic DNA. Experiments were performed in duplicate and analyzed using previously validated criteria (Bernstein et al., 2005). The resulting maps of ES cell chromatin (Figures 1 and S1) show a number of important features, many of which were unexpected.

We found a total of 343 sites of Lys4 methylation, ranging in size from 1 kb to 14 kb with a median size of 3.4 kb (Table S2). Of these, 63% correspond to transcription start sites (TSSs) of known genes. Conversely, 80% of the TSSs are covered by Lys4 sites. Because Lys4 sites and TSSs each cover only ~2% of the genomic regions, this concordance is highly significant. We and others have previously noted a global concordance between Lys4-methylated sites and TSSs in differentiated mammalian cells (Bernstein et al., 2005; Kim et al., 2005).

We also found 192 Lys27-methylated sites across these regions. These tend to affect much larger genomic regions than the Lys4 sites. The median Lys27 site is smaller in the control regions (5 kb) but twice as large in the HCNE-rich regions (10 kb) and still larger in the Hox regions (18 kb). Overall, 75% of Lys27 sites are larger than 5 kb. We will refer to these large regions as “Lys27 domains.” There are 123 in the HCNE regions, 14 in the Hox regions, and 7 in the control regions (Table S2).

Comparison of the two datasets revealed many instances of a previously undescribed pattern of chromatin modifications: Three-quarters of the Lys27 domains contain Lys4 sites within them. These regions thus harbor both a “repressive” and an “activating” chromatin modification; we therefore termed them “bivalent domains.” There are 95 in the HCNE regions, 9 in the Hox regions, and 5 in the control regions (Table S2).

Bivalent Domains Overlay Developmentally Important TF Genes in HCNE-Rich Regions

Roughly three-quarters of the bivalent domains in the HCNE regions (69/95) overlap TSSs of known genes, with the Lys4 sites typically positioned directly at the

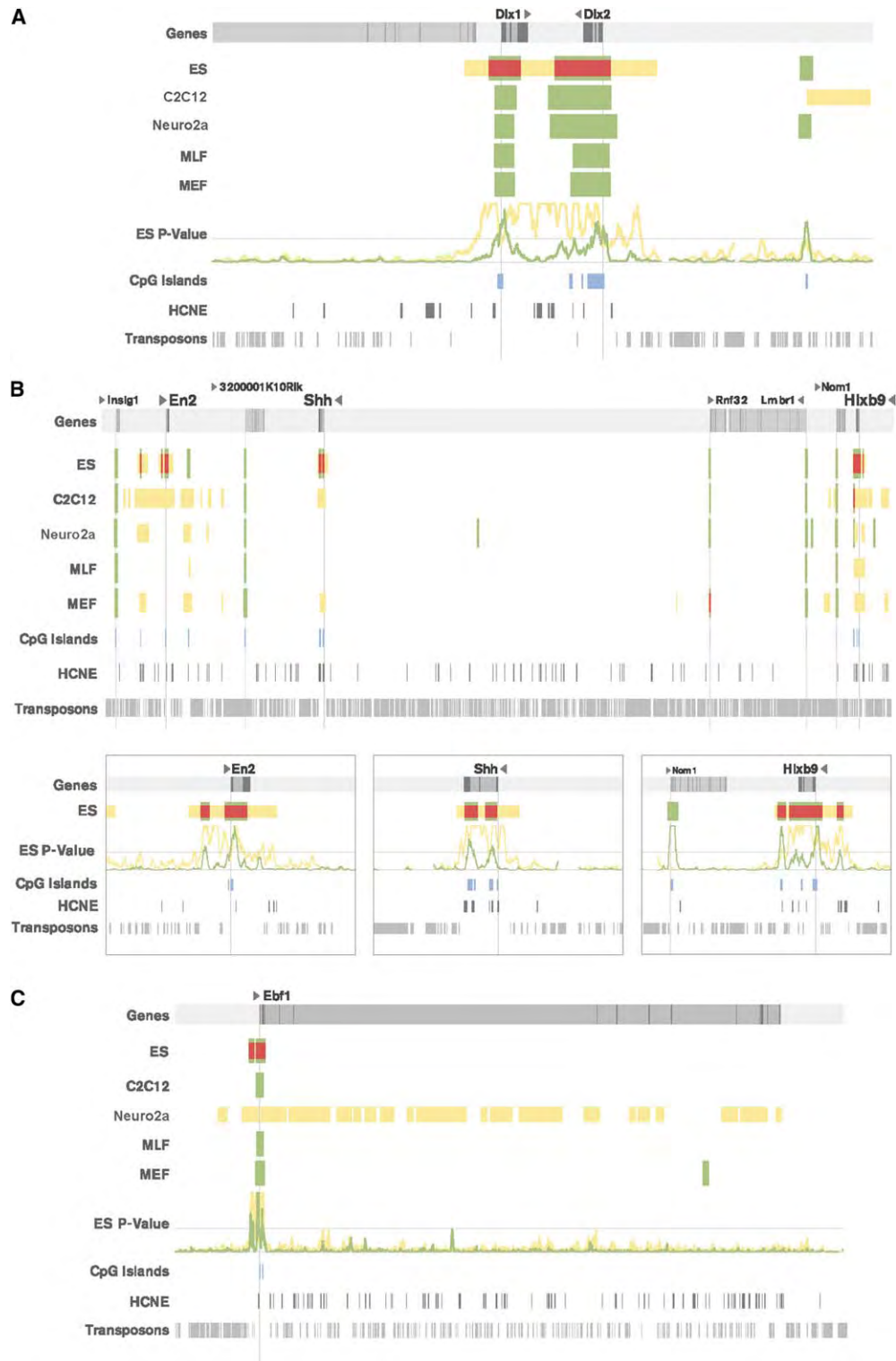


Figure 1. Representative Views of Histone Methylation Patterns across HCNE-Rich Regions in ES and Differentiated Cells

(A) Dlx1-Dlx2 gene cluster (Region 47, 112 kb). For each cell type, tracks show regions associated with Lys27 methylation (yellow), Lys4 methylation (green), or both modifications (red). For ES cells only, the raw p-value signals for Lys27 (yellow) and Lys4 methylation (green) are also shown. Genes (TSSs indicated by long vertical lines; exons indicated as dark), CpG islands, HCNEs (Lindblad-Toh et al., 2005), and transposable elements are also shown.

(B) En2, Shh, Hlxb9 (Region 48, 1.5 MB). Expanded views show 75 kb around each gene.

(C) Ebf1 (Region 31, 497 kb). Note expansive Lys27 methylated region in the Neuro2a cells.

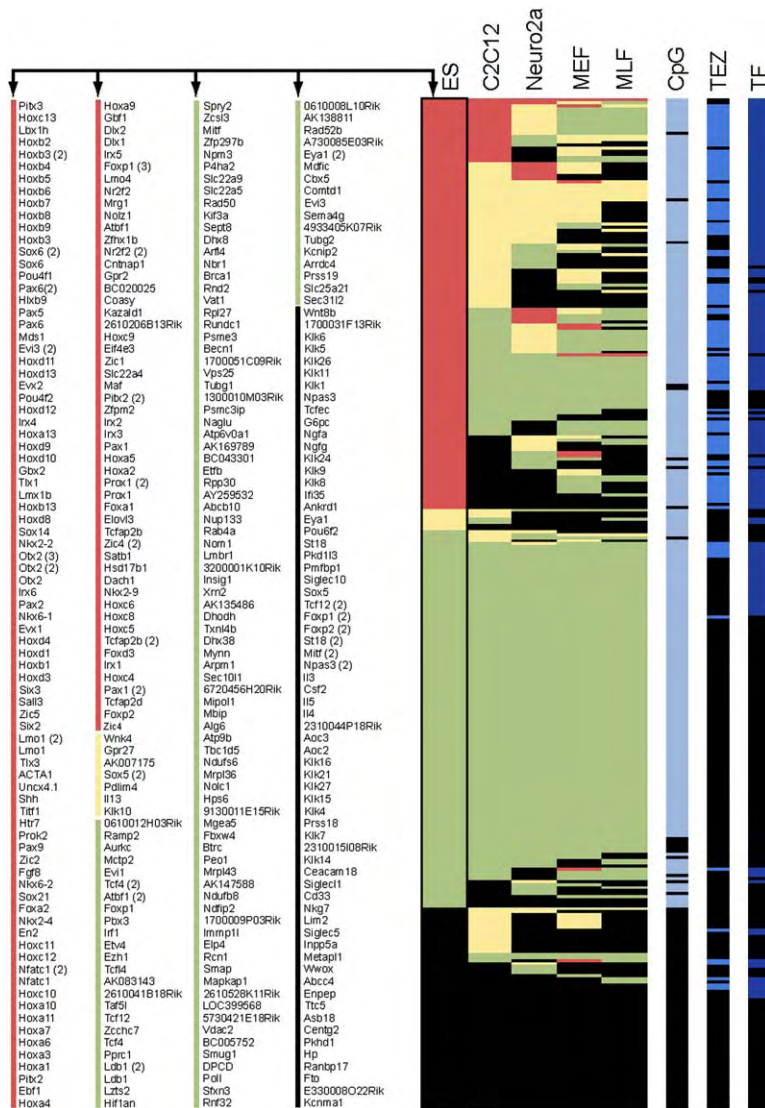


Figure 2. Histone Methylation Status of Transcription Start Sites

Methylation status is shown for the 332 known TSSs in the 61 examined regions in ES cells, C2C12 myoblasts, Neuro2a neuroblastoma cells, mouse embryonic fibroblasts (MEF), and mouse lung fibroblasts (MLF). Red indicates the presence of a bivalent domain; yellow indicates Lys27 methylation only; green indicates Lys4 methylation only; and black indicates no detected methylation. Blue rows in the three rightmost columns indicate TSSs that correspond to TF genes, contain CpG islands, or coincide with transposon-exclusion zones (TEZs). This figure and the corresponding Table S3 show the strong correlation among bivalent domains, TF genes, CpG islands, and TEZs.

TSS. Of these, a full 93% (64/69) occur at genes that encode TFs, including Sox, Fox, Pax, Irx, and Pou gene family members, even though TF genes make up only half of the genes in the regions examined (Figure 2). The 26 bivalent domains that do not occur at known TSSs are also of interest: Four occur at the 3'-ends of developmental genes (Npas3, Meis2, Pax2, and Wnt8b), and ten occur in locations that show strong evidence of encoding transcripts (including the presence of mRNA transcripts, CpG islands, and high levels of sequence conservation). Among the nonTF genes associated with bivalent domains are genes implicated in neural development, such as Fgf8 and Prok1. In the Hox regions, the observed bivalent domains are especially large and overlap multiple TSSs of known genes, all of which are TFs. The five bivalent domains in the control regions are quite short; they all overlap gene starts, although these genes do not encode TFs.

The chromatin analysis thus reveals that ES cells contain many bivalent domains. In HCNE-rich regions, these domains are particularly large and highly enriched at developmentally important genes that establish cell identity. The bivalent nature of this novel epigenetic pattern raises the possibility that the associated genes are poised in a bi-potential state, which may be resolved differently in differentiated cell lineages. This hypothesis predicts that differentiated cells would contain few, if any, bivalent domains.

In Differentiated Cells, TF Genes Are Marked by Either Repressive or Activating Modifications

We next examined Lys4 and Lys27 methylation patterns across these same regions in a collection of differentiated cell types, including mouse embryonic fibroblasts (MEFs), mouse primary lung fibroblasts (MLFs), C2C12 myoblasts, and Neuro2a neuroblastoma cells. We identified multiple

Lys4 and Lys27 methylated sites in each cell type, many of which are large. However, in marked contrast to the ES cell data, we found few bivalent domains in the differentiated cells (6 in MEFs, 1 in MLFs, 13 in myoblasts, and 12 in the neuroblastoma cells).

Thus, the majority of TSSs that show bivalent domains in ES cells do not show bivalent domains in the differentiated cells. The vast majority of these (93/97) instead show either a Lys27- or a Lys4-methylated site in at least one of the differentiated cell types (Figure 2 and Table S3). These “monovalent” sites tend to be large, with median sizes of 19.4 kb and 7.4 kb for Lys27 and Lys4, respectively (compared to 6.7 kb and 3.4 kb over all sites in the differentiated cells). Thus, bivalent domains appear largely specific to ES cells and, in differentiated cells, developmental genes are instead frequently organized within expansive regions showing either repressive or activating modifications.

Bivalent Modification Patterns in ES Cells Confirmed by Alternate Techniques

Given the novel nature of the bivalent domains, we sought to confirm our results using completely different reagents and protocols (see [Experimental Procedures](#)). Specifically, we used an independent source of ES cells with a different genotype; we refer to the first source as ES1 and the second as ES2. We also used an alternative ChIP procedure carried out on micrococcal nuclease-digested nucleosomes that had not been subjected to cross-linking and performed the immunoprecipitation with antisera from different sources. This alternative ChIP technique controls for nucleosome occupancy and is not subject to potential artifacts of cross-linking and sonication (O'Neill and Turner, 2003). The ES2 data also show a large number of bivalent domains, and these correspond closely to those seen in the ES1 data. Importantly, 94 of the 95 bivalent domains in the ES2 data correspond to bivalent domains in the ES1 cells.

We next sought to test whether the observed bivalent domain structure truly reflects the simultaneous presence of both Lys4 and Lys27 methylation on the same physical chromosomes. It is formally possible that the bivalent domains could instead reflect the presence of either two subpopulations with distinct character or one population alternating between two states. To rule out this possibility, we carried out a sequential ChIP in which ES cell chromatin was immunoprecipitated first with Lys27 tri-methyl antibody and second with Lys4 tri-methyl antibody. This sequential purification is designed to retain only chromatin that concomitantly carries both kinds of modifications. Using real-time PCR, we tested three TSSs associated with bivalent domains (*Irx2*, *Dlx1*, and *Hlxb9*). Each was significantly enriched relative to the controls (genes enriched for only Lys27 or only Lys4) (see [Figure 3](#) and [Experimental Procedures](#)). For example, *Irx2* is enriched ~10-fold in the primary (Lys27) ChIP and further enriched >30-fold (relative to control) in the secondary (Lys4) ChIP. This shows that a large proportion of *Irx2* chromatin that contains Lys27 methylation also contains Lys4 methyla-

tion—at least 30-fold more than the control. (Of course, the technique cannot prove that 100% of all *Irx2* species in ES cells carry both modifications because of inherent limitations due to background). We also tested the *Irx2* TSS by repeating the sequential ChIP with the order of the immunoprecipitations reversed and again found significant enrichment (see [Experimental Procedures](#)). Together, the experiments above suggest that the bivalent domains accurately represent the epigenetic state at many TF genes in ES cells.

Bivalent Domains Are Associated with Low Levels of Gene Expression

To gain insight into the functional significance of bivalent domains, we examined gene expression patterns across the three cell types with at least ten bivalent domains (ES cells, C2C12, and Neuro2a) (Mogass et al., 2004; Tomczak et al., 2004; Perez-Iratxeta et al., 2005). Within each cell type, we found that genes marked by Lys4 methylation tend to be expressed at significantly higher levels than those associated with Lys27 methylation (Figure 4). A good example is the *Ebf1* gene, which encodes a TF implicated in multiple differentiation pathways: In MEFs, MLFs, and myoblasts, it is expressed at relatively high levels (Schraets et al., 2003; Koli et al., 2004) and is associated with relatively large Lys4 sites (>5 kb), while in neuroblastoma cells it is expressed at an essentially undetectable level and is associated with an expansive Lys27 domain.

We next examined the expression levels of genes marked by bivalent domains. These show low levels of expression, with the overall distribution being similar to that for genes marked by Lys27 methylation alone (Figure 4). Thus, the presence of Lys4 methylation at a TSS is typically associated with high gene activity when it occurs in the absence of Lys27 methylation but with low gene activity when it occurs together with Lys27 methylation (that is, the repressive effect of Lys27 appears to be epistatic to the activating effect of Lys4 methylation in a bivalent domain). These results raise the possibility that bivalent domains function to silence developmental genes in ES cells while keeping them poised for induction upon initiation of specific developmental pathways.

Resolution of Bivalent Domains during ES Cell Differentiation

Our analysis of ES cells and four differentiated cell types suggests that bivalent domains are characteristic of pluripotent cells, that they silence developmental genes while keeping them poised, and that they tend to resolve upon ES cell differentiation into Lys4 or Lys27 methylation, in accordance with associated changes in gene expression. We sought to study whether the resolution of bivalent domains can be observed soon after ES cell differentiation by examining a differentiated cell type obtained directly from ES cells. Specifically, we differentiated ES cells along a neural pathway in serum-free culture and generated a homogenous population of multipotent neural precursor cells maintained in FGF2- and EGF-containing media, as

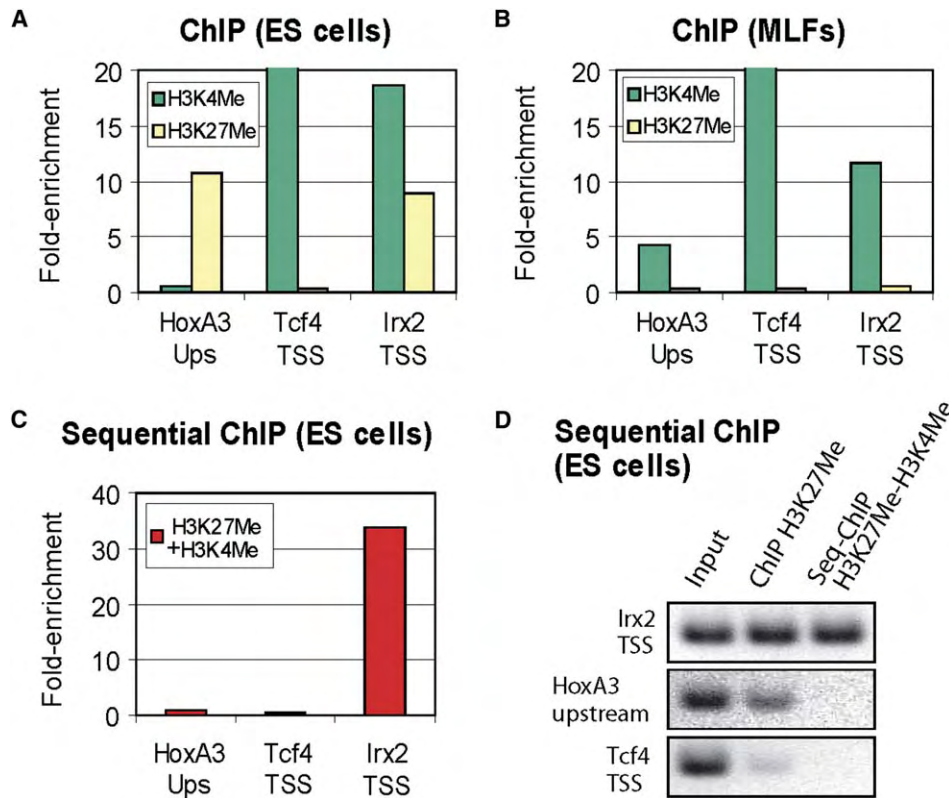


Figure 3. Characterization of the *Irx2* Bivalent Domain

ChIP and sequential ChIP were used to examine the methylation status of the *Irx2* TSS (bivalent), the *Tcf4* TSS (Lys4 only), and a site upstream of *HoxA3* (Lys27 only).

(A) Real-time PCR ratios reflect the enrichment of indicated sites when ES cells are subjected to ChIP with trimethyl Lys4 antibody or trimethyl Lys27 antibody.

(B) Corresponding data for mouse lung fibroblasts.

(C) Real-time PCR ratios reflect the relative enrichment of indicated sites after sequential immunoprecipitations with trimethyl Lys27 antibody and then trimethyl Lys4 antibody (see [Experimental Procedures](#)).

(D) PCR products amplified from control (input) DNA, Lys27 ChIP DNA, and Lys27-Lys4 sequential ChIP DNA are shown; the same samples were used as real-time PCR template in (C). The sequential ChIP results indicate that, in ES cells, the *Irx2* TSS is associated with chromatin marked by both Lys27 and Lys4 methylation.

described previously (Conti et al., 2005). We focused on seven genes associated with bivalent domains in ES cells (Figure 5). These include three genes that are markedly induced during differentiation (*Nkx2.2*, *Sox21*, and *Zfp21*), one that is weakly induced (*Dlx1*), and three that are not induced (*Pax5*, *Lbx1h*, and *Evx1*). Using ChIP and real-time PCR, we first confirmed that the TSS of each gene is indeed associated with both Lys4 and Lys27 methylation in the original ES cells, and we then examined the methylation status of these genes in the neural precursor cells. For the three genes whose expression is markedly induced, the TSS becomes specifically associated with Lys4 methylation in these differentiated cells. For the three genes that are not induced, the TSS becomes specifically associated with Lys27 methylation. Interestingly, the TSS of the weakly induced gene, *Dlx1*, remains associated with both methylation marks in the neural precursor cells, although the Lys4 signal is significantly stronger. These

data support a model in which bivalent domains are largely specific to ES cells and tend to resolve upon ES cell differentiation according to pathway-specific gene expression programs.

Epigenetic Modifications in ES Cells Strongly Correlate with Underlying DNA Sequence

Because the Lys4- and Lys27-methylated sites in ES cells seem to represent important initial conditions for development, we searched for DNA sequence features that might underlie or predict the establishment of these epigenetic marks across the genome.

We found a strong positive correlation between the presence of Lys4 methylation in ES cells and the density of CpG dinucleotides in the underlying DNA sequence (median 8% versus 2% expected) (Figure S2). Strikingly, 95% of TSSs with Lys4 sites have CpG islands, and 91% of TSSs with CpG islands also have Lys4 sites

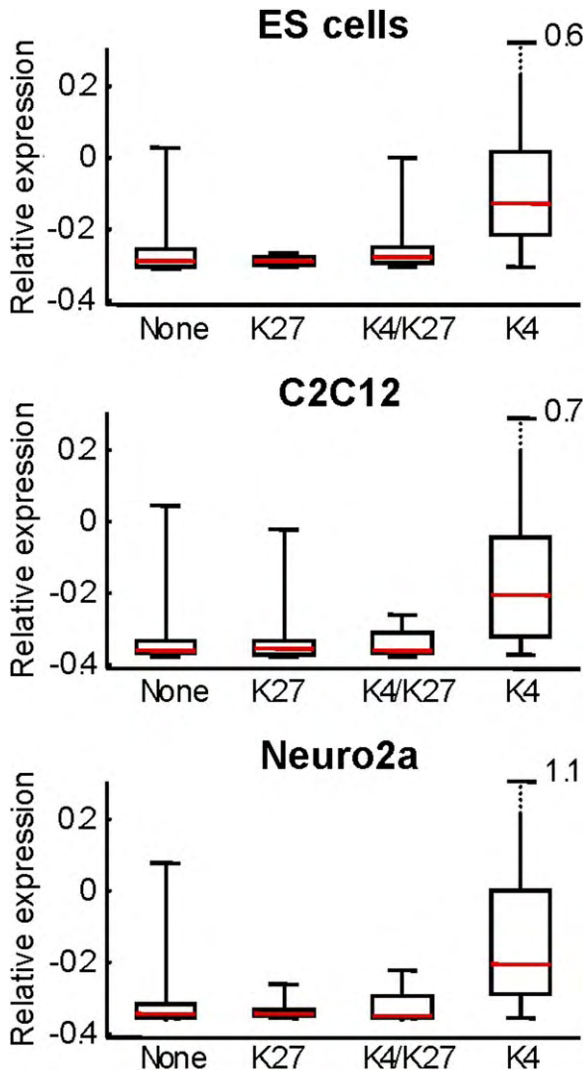


Figure 4. Gene Expression as a Function of Histone Methylation Status

Box plot showing 25th, 50th, and 75th percentile expression levels in ES cells, myoblasts, and neuroblastoma cells for genes associated with no histone methylation, Lys27 methylation, bivalent domains, or Lys4 methylation. Whiskers show 2.5th and 97.5th percentiles. Expression data (y axis) were determined from published expression profiles (Mogass et al., 2004; Tomczak et al., 2004; Perez-Iratxeta et al., 2005), uniformly normalized to a mean of 0 and a standard deviation of 1 for all probes on each array.

($r_{\text{phi}} = 0.73$) (see Table S4 and Experimental Procedures). Moreover, the lengths of the two features are significantly correlated where they overlap ($r = 0.50$). By contrast, the correlation is weaker in the differentiated cells; this is primarily due to loss of Lys4 methylation at 20%–35% of CpG islands ($r_{\text{phi}} = 0.40$ for MLFs) (Table S4). We note that a recent genome-wide study (Roh et al., 2005) of histone H3 acetylation in T cells observed a correlation with CpG islands at a similar level to that seen in the differentiated cells examined here.

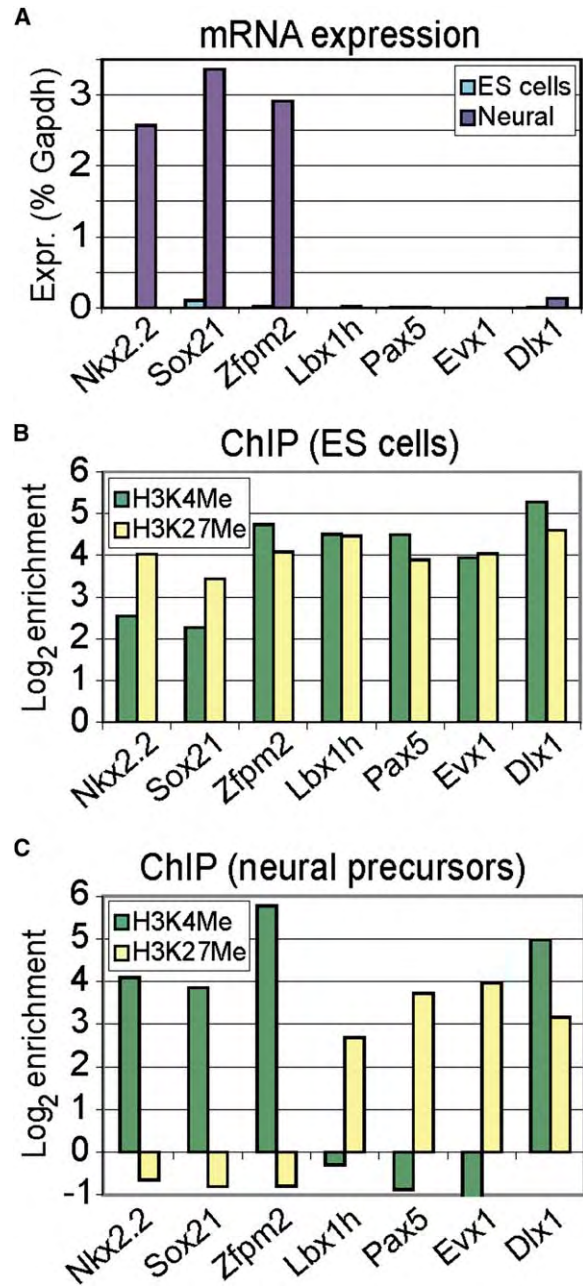


Figure 5. Resolution of Bivalent Domains during ES Cell Differentiation

ES cells were differentiated along a neural pathway in serum-free culture, and a homogenous population of multipotent neural precursor cells were maintained in FGF2- and EGF-containing media, as described in Conti et al., 2005. Several loci showing bivalent domains in ES cells were examined in the differentiated cells.

(A) Expression levels (relative to Gapdh) were determined by RT-PCR for the indicated genes in ES cells and in neural precursors. The methylation states of the indicated genes were determined by ChIP and real-time PCR in ES cells (B) and in neural precursors (C). The data suggest that bivalent domains tend to resolve during ES cell differentiation in accordance with associated changes in gene expression.

We also found that Lys27-methylated regions in ES cells show a strikingly low density of transposon-derived sequence (median 6% versus 22% expected) (Figures S2 and S3). The most extreme example is found at the Hox clusters, which are known to have the lowest density of transposon-derived sequence in the mouse and human genomes (Lander et al., 2001; Waterston et al., 2002) and which have the largest Lys27 domains (up to 141 kb) in our sample. Most of the Lys27 domains contain long stretches (>10 kb) with little or no identifiable transposon-derived sequence. We defined such regions as “transposon exclusion zones” (TEZs) (see [Experimental Procedures](#)). Within the loci examined here, 89% of TSSs with a TEZ have a Lys27 domain in ES cells, and 73% of TSSs with a Lys27 domain have a TEZ ($r_{\text{phi}} = 0.69$) (Table S4). The lengths of these two features are significantly correlated where they overlap ($r = 0.78$). Interestingly, we note that the small number of Lys27 domains found only in differentiated cells does not appear to overlap particularly transposon-poor sequence (Figure 1C and Table S4).

We tested if the TEZs represent conserved genomic features by examining the orthologous sequence in the human and dog genome. The frequency of lineage-specific repeats provides an independent test of whether transposons are tolerated in these regions. The TEZs show a clear deficit of lineage-specific repeats in both human (1.3% versus 15.2% expected) and dog (1.0% versus 9.1% expected), confirming that this property is strongly conserved across mammals.

We then searched for TEZs across the entire mouse genome. We identified 710 TEZs, of which 328 overlap TSSs of known genes (Table S5). Strikingly, the vast majority of these genes encode developmental and tissue-specific TFs (189), proteins involved in axon guidance and neuronal function (65), and other cell signaling-related proteins such as growth factors (25), including Fgf8, Fgf10, Fgf14, and the imprinted gene Igf2. Notably, they include ~70% of the developmental regulators previously identified within 204 HCNE-rich loci (Lindblad-Toh et al., 2005). We predict that most of these genes will harbor Lys27 domains or bivalent domains in ES cells.

Colocalization of Bivalent Domains with Oct4 and Nanog

Finally, we examined the relationship of bivalent domains to the reported binding sites of certain pluripotent TFs. A recent genomic analysis in human ES cells found that Oct4, Nanog, and Sox2 are frequently associated with developmentally important genes (Boyer et al., 2005). We mapped the Oct4, Nanog, and Sox2 binding sites reported in that study to orthologous positions in the mouse genome and examined their overlap with bivalent domains. About 50% of bivalent domains coincide with binding sites of at least one of the pluripotent TFs, a highly significant correspondence ($p < 10^{-9}$). The correlation is primarily due to Oct4 and Nanog and actually becomes more significant when Sox2 is removed from the analysis.

Interestingly, although many of the genes targeted by these pluripotent factors are actively expressed in ES cells, those that are also associated with a bivalent domain tend to be silenced ($p < 5 \times 10^{-3}$). This suggests that the bivalent domains may override any activation potential these TFs might have but also raises the possibility that the pluripotent TFs may help keep these genes in a poised state. Notably, a full 50% of bivalent domains are not associated with any of the three pluripotent TFs. It will be interesting to see if these coincide with binding sites of other important TFs.

DISCUSSION

Our results shed light on chromatin structure in ES cells and raise intriguing hypotheses about its establishment and function during development. The bivalent domains reported here have many notable features: They combine both “repressive” and “activating” modifications; they are highly enriched in ES cells relative to differentiated cells; and they are associated with genes encoding TFs with roles in embryonic development and lineage specification. In differentiated cells, these TF genes instead tend to be associated with large regions carrying either an activating or a repressive methylation mark. We propose that bivalent domains silence developmental genes in ES cells while preserving their potential to become activated upon initiation of specific differentiation programs. Bivalent domains may be related to a phenomenon observed at the bithorax complex in early fly development, where silenced Polycomb response elements are nonetheless associated with trithorax-group proteins and low-level transcription. Remarkably, both of these activities appear to be required for subsequent gene activation during development (Orlando et al., 1998; Schmitt et al., 2005). By analogy, Lys4 methylation within bivalent domains and associated trithorax activities may keep silenced developmental genes poised in ES cells. Our analyses of differentiated cells suggest that bivalent domains largely resolve during differentiation into large regions of either Lys27 or Lys4 methylation. These modified regions may provide a robust epigenetic memory to maintain lineage-specific expression or repression of these critical genes. Their large size would ensure that each daughter chromosome would likely inherit a substantial proportion of the modified histones, which could then promote similar modification of new histones in the immediate vicinity (Henikoff et al., 2004; van Steensel, 2005).

A fundamental issue that remains is to understand the mechanism by which the initial conditions are established in ES cells. The analysis here suggests that some of the answer can be read directly from the genome sequence. The strong association of Lys4 methylation with CpG islands may well be directly causal, inasmuch as the trithorax complexes that methylate Lys4 are reported to associate with CpG-rich DNA (Ayton et al., 2004; Lee and Skalnik, 2005). The strong association of Lys27 methylation with transposon-exclusion zones

may instead reflect strong evolutionary pressure against the presence of transposon-derived sequence in these regions. Repetitive sequences are subject to repressive epigenetic modifications (Arnaud et al., 2000; Lippman et al., 2004; Martens et al., 2005), which might interfere with the function of the bivalent domains and thus be eliminated by selection. It has been reported previously that imprinted loci, while significantly depleted for short interspersed transposable elements (SINEs), are permissive to L1 long interspersed transposable elements (LINEs) (Greally, 2002). In contrast, we find that both classes of transposons tend to be excluded from regions associated with Lys27 methylation or bivalent domains in ES cells. The direct signal for Lys27 methylation remains unclear (although we cannot exclude the possibility that the deficit of transposon-related chromatin modifications in some fashion promotes the association of the Polycomb complex that methylates Lys27). The correlations between the histone modifications and the genomic features are notably weaker in differentiated cells (Bernstein et al., 2005). We suggest that while the embryonic state may be largely defined by DNA sequence, it is subsequently altered in response to lineage-specific transcriptional programs and environmental cues and epigenetically maintained.

Our study was motivated by the suspicion that HCNE-rich regions might be particularly fruitful targets for studying chromatin structure in ES cells; this has indeed been borne out. However, the results here do not explain the functional role of the HCNEs themselves. Although HCNEs are markedly enriched at many of the Lys27 and Lys4 sites in both ES and differentiated cells, they tend overall to be distributed across much larger regions. One possibility is that some of the HCNEs dictate chromosome conformation or nuclear localization in a manner that facilitates robust gene regulation and/or epigenetic switching (Chambeyron and Bickmore, 2004; Kosak and Groudine, 2004).

Further studies will be needed to define bivalent domains and related features. It will be important to examine the entire genome in ES cells as well as to follow their fate during development and differentiation. In particular, it will be interesting to determine whether the bivalent domains that persist following ES cell differentiation correspond to genes that remain poised for later induction. In addition, it will be valuable to characterize the bivalent domains with respect to other epigenetic modifications and the binding sites of additional TFs. We note that preliminary studies of H3 Lys9 methylation show no evidence of association with bivalent domains.

A deeper understanding of bivalent domains may shed light on mechanisms that underlie the maintenance of pluripotency in ES cells and lineage fidelity in differentiated cells. Moreover, a comprehensive inventory of the presence or absence of bivalent domains over key developmental genes may provide valuable markers of cell identity and differentiation potential, both in normal and pathologic states.

EXPERIMENTAL PROCEDURES

Cell Culture

The first source of ES cells (ES1) were V6.5 murine ES cells (genotype 129SvJae × C57BL/6; male; passages 10–15). They were cultivated in 5% CO₂ at 37° on irradiated MEFs in DMEM containing 15% FCS, leukemia-inhibiting factor, penicillin/streptomycin, L-glutamine, and non-essential amino acids (Rideout et al., 2000). At least two to three passages under feeder-free conditions on 0.2% gelatin were used to exclude feeder contamination. The second source of ES cells (ES2, used for micrococcal nuclease-ChIP) was SF1-1 murine ES cells (genotype C57BL/6 × M. spretus F1; male; passages 11–16) grown in the absence of feeder cells on gelatinized plates as described previously (Umlauf et al., 2004). Primary mouse lung fibroblasts (ATCC #CCL-206), mouse embryonic fibroblasts (10.5 p.c.) immortalized with polyoma virus, C2C12 myoblasts (ATCC #CRL-1772), and Neuro2a neuroblastoma cells (ATCC #CCL-131) were grown in DMEM with 10% fetal bovine serum and penicillin/streptomycin at 37°, 5% CO₂. ES1 cells were differentiated into pan-neural precursor cells through embryoid body formation for 4 days and selection in ITSFn media for 5–7 days (Okabe et al., 1996), and they were maintained in FGF2 and EGF2 (both from R&D Systems) containing chemically defined media as described (Conti et al., 2005). These cells uniformly express nestin and Sox2 and upon growth factor withdrawal differentiate into neurons, astrocytes, and oligodendrocytes (Brustle et al., 1999; Conti et al., 2005).

Chromatin Immunoprecipitation

ChIP experiments for all cells except ES2 were carried out as described in Bernstein et al., 2005 and at <http://www.upstate.com>. Briefly, $\sim 5 \times 10^7$ cells were trypsinized, fixed with 1% formaldehyde, resuspended in Lysis Buffer, and fragmented with a Branson 250 Sonifier to a size range of 200 to 1,000 bases. Solubilized chromatin was diluted 10-fold in ChIP dilution buffer and, after removal of a control aliquot, incubated at 4°C overnight with antibody against trimethyl Lys4 (Abcam #8580) or trimethyl Lys27 (Upstate #07-449). Immune complexes were precipitated with Protein A-sepharose, washed sequentially with Low Salt Immune Complex Wash, LiCl Immune Complex Wash, and TE, and then eluted in Elution Buffer. After cross-link reversal and Proteinase K treatment, ChIP and control DNA samples were extracted with phenol-chloroform, precipitated under ethanol, treated with RNase and Calf Intestinal Alkaline Phosphatase, and purified with a MinElute Kit (Qiagen).

Micrococcal Nuclease-ChIP

Chromatin fragments of one to six nucleosomes were prepared from unfixed chromatin from ES2 cells by micrococcal nuclease digestion and immunoprecipitated using antibody against trimethyl Lys27 (Plath et al., 2003) or dimethyl Lys4 (Upstate #07-030) as described (Umlauf et al., 2004). Immunoprecipitated DNA fractions and a control DNA sample enriched using unrelated antisera (against chicken antibodies) were extracted and purified as described above.

Sequential ChIP

Cross-linked chromatin from ES cells was immunoprecipitated with antibody against trimethyl Lys27 as described above (see "Chromatin Immunoprecipitation") except that chromatin was eluted in a solution of 30 mM DTT, 500 mM NaCl, and 0.1% SDS at 37°. Eluted chromatin was diluted 50-fold, subjected to a second immunoprecipitation with antibody against trimethyl Lys4, and then eluted with standard Elution Buffer. The isolated DNA was extracted and purified as above. In addition, a "reverse" sequential ChIP was carried out in which chromatin was immunoprecipitated first with antibody against trimethyl Lys4 and then with antibody against trimethyl Lys27.

Real-Time PCR

PCR primers for evaluating ChIP assays were designed to amplify 150–200 base pair fragments from the indicated genomic regions. Real-time PCR was carried out using Quantitect SYBR green PCR mix (Qiagen) in an MJ Research Opticon Instrument. For ChIP experiments, either 0.5 ng ChIP DNA or 0.5 ng control DNA was used as template, and fold-enrichments were determined by the $2^{-\Delta\text{CT}}$ method described in the Applied Biosystems User Bulletin. For sequential ChIP experiments, 2 μl sequential ChIP DNA or 2 μl of a 1:100 dilution of input DNA was used as template, and relative fold-enrichments were determined by the $2^{-\Delta\Delta\text{CT}}$ method, using HoxA3 Ups as the normalizer. Ratios were determined from two independent ChIP or sequential ChIP assays, each evaluated in duplicate by real-time PCR. Primers corresponding to the *Irx2*, *Dlx1*, or *Hlx9* TSSs were used to test for enrichment of Lys9 methylation (see Supplemental Data). RT-PCR was used to measure gene expression in ES cells and neural precursor cells. Briefly, RNA was isolated using an RNeasy mini kit (Qiagen), reverse transcribed, and quantified using SYBR green PCR master mix on a 7000 ABI detection system. Primer sequences are available in Supplemental Data.

Region Selection and Array Design

The identification of 204 HCNE-rich loci on the basis of sequence comparisons across the human, mouse, and dog genomes was reported previously (Lindblad-Toh et al., 2005). For the current study, we selected 56 HCNE-rich loci, including all four Hox clusters, as well as 5 control loci that do not show unusual HCNE density (ACTA locus, chromosome 19 gene desert, CD33r locus, BRCA1 locus, cytokine cluster). Each region was mapped to the mouse genome using mm5 coordinates (Table S1). Custom tiling arrays for these regions were obtained from Affymetrix Inc. (Santa Clara, CA). They contain approximately 1.3 million probe pairs, each consisting of perfect match (PM) and single base mismatch (MM) 25-mer oligonucleotides, designed to interrogate the unique sequence in these regions at approximately 30 base intervals (Kapranov et al., 2002).

DNA Amplification and Array Hybridization

ChIP and control DNA samples were amplified by in vitro transcription, converted into double-stranded cDNA with random primers, fragmented with DNase I, and end-labeled with biotin as described (Kapranov et al., 2002; Liu et al., 2003; Cawley et al., 2004; Bernstein et al., 2005). ChIP and control samples (5–10 μg) were hybridized to separate oligonucleotide arrays. Arrays were hybridized 16–18 hr at 45°C, washed, stained, and scanned using an Affymetrix GeneChip Scanner 3000 7G as described in the Affymetrix Expression Analysis Technical Manual.

Analysis of ChIP Tiling Array Data

Raw array data were quantile-normalized, scaled, and analyzed as described (Cawley et al., 2004; Bernstein et al., 2005). Enrichment was quantified using a Wilcoxon Rank Sum test applied to the transformation $\log_2(\max[\text{PM}-\text{MM}, 1])$ for data from ChIP and control arrays within a window of ± 500 base pairs, testing the null hypothesis that ChIP and control data come from the same probability distribution. Genomic positions belonging to enriched regions were defined by applying a high P-value cutoff of 10^{-4} . These regions were extended locally by merging adjacent windows with P-values of at least 10^{-2} , and resultant positions separated by < 2 kb were merged to form a predicted Lys4 or Lys27 methylated site. Bivalent domains were defined as Lys27 methylated regions > 5 kb that overlap Lys4 sites > 1 kb. Histone methylation data are available as interactive tracks at: http://www.broad.mit.edu/cell/hcne_chromatin.

Genomic Analysis

We collated a list of known TSSs based on RefSeq and Genbank mRNAs aligned to the examined regions in mouse (mm5) and the orthologous regions in human (hg17; alignments obtained from the

UCSC genome browser). The methylation status of each TSS was based on the presence of significantly enriched Lys4 or Lys27 sites or bivalent domains within 2 kb upstream or downstream. The expected density of CpG and transposable elements at methylated sites was determined from random intervals of the same size, anchored in nonrepetitive sequence. CpG islands were defined as in Takai and Jones, 2002. TEZs (transposon exclusion zones) were defined as regions satisfying one of two criteria: (1) regions of at least 10 kb without any transposable elements; (2) regions of at least 15 kb with no more than 250 bases annotated as transposable elements. Identified regions were then merged together as one TEZ if they were within 1 kb. For the genome-wide search, only criterion (1) was used. Supplemental Data are available at: http://www.broad.mit.edu/cell/hcne_chromatin.

Supplemental Data

Supplemental Data include three figures and five tables and can be found with this article online at <http://www.cell.com/cgi/content/full/125/2/315/DC1/>.

ACKNOWLEDGMENTS

We thank Aisling O'Donovan, Jen Couget, Phil Kapranov, Rob Schneider, Andi Gnirke, Christina Hughes, Leslie Gaffney, and Michelle Clamp for insightful comments and suggestions. We gratefully acknowledge Kathryn Coser and Toshi Shioda at the Massachusetts General Hospital Cancer Center microarray core facility for assistance with hybridizations and Claire Reardon at the Bauer Center for Genomics Research for assistance with real-time PCR. B.E.B. is supported by a K08 Development Award from the National Cancer Institute. A.M. was supported by a Boehringer Ingelheim Fonds (BIF) PhD fellowship. M.W. is supported by a longterm fellowship provided by the Human Frontier Science Program Organization (HFSP). R.F. acknowledges grant funding from the European Science Foundation (EuroSTELLS) and from ARC (France). S.L.S. is an investigator in the Howard Hughes Medical Institute. This work was supported in part by grants from the National Cancer Institute (CA84198 to R.J.), the National Institute of Child Health and Human Development (HD045022 to R.J.), the National Institute for General Medical Sciences (GM38627 to S.L.S.), the National Human Genome Research Institute (to E.S.L.), and by funds from the Broad Institute of MIT and Harvard.

Received: November 4, 2005

Revised: January 18, 2006

Accepted: February 23, 2006

Published: April 20, 2006

REFERENCES

- Arnaud, P., Goubely, C., Pelissier, T., and Deragon, J.M. (2000). SINE retroposons can be used in vivo as nucleation centers for de novo methylation. *Mol. Cell. Biol.* 20, 3434–3441.
- Ayton, P.M., Chen, E.H., and Cleary, M.L. (2004). Binding to nonmethylated CpG DNA is essential for target recognition, transactivation, and myeloid transformation by an MLL oncoprotein. *Mol. Cell. Biol.* 24, 10470–10478.
- Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* 304, 1321–1325.
- Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120, 169–181.
- Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al.

- (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947–956.
- Brustle, O., Jones, K.N., Learish, R.D., Karram, K., Choudhary, K., Wiestler, O.D., Duncan, I.D., and McKay, R.D. (1999). Embryonic stem cell-derived glial precursors: a source of myelinating transplants. *Science* 285, 754–756.
- Cao, R., and Zhang, Y. (2004). SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. *Mol. Cell* 15, 57–67.
- Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., et al. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116, 499–509.
- Chambeyron, S., and Bickmore, W.A. (2004). Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev.* 18, 1119–1130.
- Conti, L., Pollard, S.M., Gorba, T., Reitano, E., Toselli, M., Biella, G., Sun, Y., Sanzone, S., Ying, Q.L., Cattaneo, E., and Smith, A. (2005). Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol.* 3, e283.
- Delaval, K., and Feil, R. (2004). Epigenetic regulation of mammalian genomic imprinting. *Curr. Opin. Genet. Dev.* 14, 188–195.
- Francis, N.J., Kingston, R.E., and Woodcock, C.L. (2004). Chromatin compaction by a polycomb group protein complex. *Science* 306, 1574–1577.
- Greally, J.M. (2002). Short interspersed transposable elements (SINES) are excluded from imprinted regions in the human genome. *Proc. Natl. Acad. Sci. USA* 99, 327–332.
- Guenther, M.G., Jenner, R.G., Chevalier, B., Nakamura, T., Croce, C.M., Canaani, E., and Young, R.A. (2005). Global and Hox-specific roles for the MLL1 methyltransferase. *Proc. Natl. Acad. Sci. USA* 102, 8603–8608.
- Henikoff, S., Furuyama, T., and Ahmad, K. (2004). Histone variants, nucleosome assembly and epigenetic inheritance. *Trends Genet.* 20, 320–326.
- Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. *Science* 293, 1074–1080.
- Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919.
- Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature* 436, 876–880.
- Kimura, H., Tada, M., Nakatsuji, N., and Tada, T. (2004). Histone code modifications on pluripotential nuclei of reprogrammed somatic cells. *Mol. Cell. Biol.* 24, 5710–5720.
- Kirmizis, A., Bartley, S.M., Kuzmichev, A., Margueron, R., Reinberg, D., Green, R., and Farnham, P.J. (2004). Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes Dev.* 18, 1592–1605.
- Koli, K., Wempe, F., Sterner-Kock, A., Kantola, A., Komor, M., Hofmann, W.K., von Melchner, H., and Keski-Oja, J. (2004). Disruption of LTBP-4 function reduces TGF-beta activation and enhances BMP-4 signaling in the lung. *J. Cell Biol.* 167, 123–133.
- Kosak, S.T., and Groudine, M. (2004). Gene order and dynamic domains. *Science* 306, 644–647.
- Koyanagi, M., Baguet, A., Martens, J., Margueron, R., Jenuwein, T., and Bix, M. (2005). EZH2 and histone 3 trimethyl lysine 27 associated with Il4 and Il13 gene silencing in T(H)1 cells. *J. Biol. Chem.* 280, 31470–31477.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Lee, J.H., and Skalnik, D.G. (2005). CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J. Biol. Chem.* 280, 41725–41731.
- Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., 3rd, Zody, M.C., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438, 803–819.
- Lippman, Z., Gendrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., et al. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430, 471–476.
- Liu, C.L., Schreiber, S.L., and Bernstein, B.E. (2003). Development and validation of a T7 based linear amplification for genomic DNA. *BMC Genomics* 4, 19.
- Margueron, R., Trojer, P., and Reinberg, D. (2005). The key to development: interpreting the histone code? *Curr. Opin. Genet. Dev.* 15, 163–176.
- Martens, J.H., O'Sullivan, R.J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P., and Jenuwein, T. (2005). The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J.* 24, 800–812.
- Mogass, M., York, T.P., Li, L., Rujibanjerd, S., and Shiang, R. (2004). Genomewide analysis of gene expression associated with Tcof1 in mouse neuroblastoma. *Biochem. Biophys. Res. Commun.* 325, 124–132.
- Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. (2003). Scanning human gene deserts for long-range enhancers. *Science* 302, 413.
- O'Carroll, D., Erhardt, S., Pagani, M., Barton, S.C., Surani, M.A., and Jenuwein, T. (2001). The polycomb-group gene *Ezh2* is required for early mouse development. *Mol. Cell. Biol.* 21, 4330–4336.
- Okabe, S., Forsberg-Nilsson, K., Spiro, A.C., Segal, M., and McKay, R.D. (1996). Development of neuronal precursor cells and functional postmitotic neurons from embryonic stem cells in vitro. *Mech. Dev.* 59, 89–102.
- O'Neill, L.P., and Turner, B.M. (2003). Immunoprecipitation of native chromatin: NChIP. *Methods* 31, 76–82.
- Orlando, V., Jane, E.P., Chinwalla, V., Harte, P.J., and Paro, R. (1998). Binding of trithorax and Polycomb proteins to the bithorax complex: dynamic changes during early *Drosophila* embryogenesis. *EMBO J.* 17, 5141–5150.
- Perez-Iratxeta, C., Palidwor, G., Porter, C.J., Sanche, N.A., Huska, M.R., Suomela, B.P., Muro, E.M., Krzyzanowski, P.M., Hughes, E., Campbell, P.A., et al. (2005). Study of stem cell function using microarray experiments. *FEBS Lett.* 579, 1795–1801.
- Perry, P., Sauer, S., Billon, N., Richardson, W.D., Spivakov, M., Warnes, G., Livesey, F.J., Merckenschlager, M., Fisher, A.G., and Azuara, V. (2004). A dynamic switch in the replication timing of key regulator genes in embryonic stem cells upon neural induction. *Cell Cycle* 3, 1645–1650.
- Plath, K., Fang, J., Mlynarczyk-Evans, S.K., Cao, R., Worringer, K.A., Wang, H., de la Cruz, C.C., Otte, A.P., Panning, B., and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. *Science* 300, 131–135.
- Pray-Grant, M.G., Daniel, J.A., Schieltz, D., Yates, J.R., 3rd, and Grant, P.A. (2005). Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. *Nature* 433, 434–438.
- Rideout, W.M., 3rd, Wakayama, T., Wutz, A., Eggan, K., Jackson-Grusby, L., Dausman, J., Yanagimachi, R., and Jaenisch, R. (2000). Generation of mice from wild-type and targeted ES cells by nuclear cloning. *Nat. Genet.* 24, 109–110.

- Ringrose, L., Ehret, H., and Paro, R. (2004). Distinct contributions of histone H3 lysine 9 and 27 methylation to locus-specific stability of polycomb complexes. *Mol. Cell* 16, 641–653.
- Ringrose, L., and Paro, R. (2004). Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet.* 38, 413–443.
- Roh, T.Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* 19, 542–552.
- Sado, T., and Ferguson-Smith, A.C. (2005). Imprinted X inactivation and reprogramming in the preimplantation mouse embryo. *Hum. Mol. Genet.* 14, R59–R64.
- Santos-Rosa, H., Schneider, R., Bernstein, B.E., Karabetsou, N., Morillon, A., Weise, C., Schreiber, S.L., Mellor, J., and Kouzarides, T. (2003). Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. *Mol. Cell* 12, 1325–1332.
- Schmitt, S., Prestel, M., and Paro, R. (2005). Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev.* 19, 697–708.
- Schraets, D., Lehmann, T., Dinger, T., and Marschalek, R. (2003). MLL-mediated transcriptional gene regulation investigated by gene expression profiling. *Oncogene* 22, 3655–3668.
- Silva, J., Mak, W., Zvetkova, I., Appanah, R., Nesterova, T.B., Webster, Z., Peters, A.H., Jenuwein, T., Otte, A.P., and Brockdorff, N. (2003). Establishment of histone H3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Dev. Cell* 4, 481–495.
- Sims, R.J., 3rd, Chen, C.F., Santos-Rosa, H., Kouzarides, T., Patel, S.S., and Reinberg, D. (2005). Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. *J. Biol. Chem.* 280, 41789–41792.
- Szutorisz, H., and Dillon, N. (2005). The epigenetic basis for embryonic stem cell pluripotency. *Bioessays* 27, 1286–1293.
- Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* 99, 3740–3745.
- Tomczak, K.K., Marinescu, V.D., Ramoni, M.F., Sanoudou, D., Montanaro, F., Han, M., Kunkel, L.M., Kohane, I.S., and Beggs, A.H. (2004). Expression profiling and identification of novel genes involved in myogenic differentiation. *FASEB J.* 18, 403–405.
- Umlauf, D., Goto, Y., Cao, R., Cerqueira, F., Wagschal, A., Zhang, Y., and Feil, R. (2004). Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. *Nat. Genet.* 36, 1296–1300.
- Valk-Lingbeek, M.E., Bruggeman, S.W., and van Lohuizen, M. (2004). Stem cells and cancer; the polycomb connection. *Cell* 118, 409–418.
- van Steensel, B. (2005). Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat. Genet. Suppl.* 37, S18–S24.
- Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562.
- Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3, e7.
- Wysocka, J., Swigut, T., Milne, T.A., Dou, Y., Zhang, X., Burlingame, A.L., Roeder, R.G., Brivanlou, A.H., and Allis, C.D. (2005). WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* 121, 859–872.

Genome-wide maps of chromatin state in pluripotent and lineage-committed cells

Tarjei S. Mikkelsen^{1,2}, Manching Ku^{1,4}, David B. Jaffe¹, Biju Issac^{1,4}, Erez Lieberman^{1,2}, Georgia Giannoukos¹, Pablo Alvarez¹, William Brockman¹, Tae-Kyung Kim⁵, Richard P. Koche^{1,2,4}, William Lee¹, Eric Mendenhall^{1,4}, Aisling O'Donovan⁴, Aviva Presser¹, Carsten Russ¹, Xiaohui Xie¹, Alexander Meissner³, Marius Wernig³, Rudolf Jaenisch³, Chad Nusbaum¹, Eric S. Lander^{1,3*} & Bradley E. Bernstein^{1,4,6*}

We report the application of single-molecule-based sequencing technology for high-throughput profiling of histone modifications in mammalian cells. By obtaining over four billion bases of sequence from chromatin immunoprecipitated DNA, we generated genome-wide chromatin-state maps of mouse embryonic stem cells, neural progenitor cells and embryonic fibroblasts. We find that lysine 4 and lysine 27 trimethylation effectively discriminates genes that are expressed, poised for expression, or stably repressed, and therefore reflect cell state and lineage potential. Lysine 36 trimethylation marks primary coding and non-coding transcripts, facilitating gene annotation. Trimethylation of lysine 9 and lysine 20 is detected at satellite, telomeric and active long-terminal repeats, and can spread into proximal unique sequences. Lysine 4 and lysine 9 trimethylation marks imprinting control regions. Finally, we show that chromatin state can be read in an allele-specific manner by using single nucleotide polymorphisms. This study provides a framework for the application of comprehensive chromatin profiling towards characterization of diverse mammalian cell populations.

One of the fundamental mysteries of biology is the basis of cellular state. Although they have essentially identical genomes, the different cell types in a multicellular organism maintain markedly different behaviours that persist over extended periods. The most extreme case is lineage commitment during development, where cells progress from totipotency to pluripotency to terminal differentiation; each step involves establishment of a stable state encoding specific developmental commitments that can be faithfully transmitted to daughter cells. Considerable evidence suggests that cellular state may be closely related to 'chromatin state'—that is, modifications to histones and other proteins that package the genome^{1–3}. Accordingly, it would be desirable to construct 'chromatin-state maps' for a wide variety of cell types, showing the genome-wide distribution of important chromatin modifications.

Chromatin state can be studied by chromatin immunoprecipitation (ChIP), in which an antibody is used to enrich DNA from genomic regions carrying a specific epitope. The major challenge to generating genome-wide chromatin-state maps lies in characterizing these enriched regions in a scalable manner. Enrichment at individual loci is commonly assayed by polymerase chain reaction (PCR), but this method does not scale efficiently. A more recent approach has been ChIP-chip, in which enriched DNA is hybridized to a microarray^{4,5}. This technique has been successfully used to study large genomic regions. However, ChIP-chip suffers from inherent technical limitations: (1) it requires large amounts (several micrograms) of DNA and thus involves extensive amplification, which introduces bias; (2) it is subject to cross-hybridization, which hinders the study of repeated sequences and allelic variants; and (3) it is currently expensive to study entire mammalian genomes. Given these issues, only a handful of whole-genome ChIP-chip studies in mammals have been reported.

In principle, chromatin could be readily mapped across the genome by sequencing ChIP DNA and identifying regions that are over-represented among these sequences. Notably, sequence-based mapping could require relatively small quantities of DNA and provide nucleotide-level discrimination of similar sequences, thereby maximizing genome coverage. The major limitation has been that high-resolution mapping requires millions of sequences (Supplementary Note 1). This is cost-prohibitive with traditional technology, even with concatenation of multiple sequence tags⁶. However, recent advances in single-molecule-based sequencing (SMS) technology promise to increase throughput and decrease costs markedly⁷. In the approach developed by Illumina/Solexa, DNA molecules are arrayed across a surface, locally amplified, subjected to successive cycles of primer-mediated single-base extension (using fluorescently labelled reversible terminators) and imaged after each cycle to determine the inserted base. The 'read length' is short (25–50 bases), but tens of millions of DNA fragments may be read simultaneously.

Here, we report the development of a method for mapping ChIP enrichment by sequencing (ChIP-Seq) and describe its application to create chromatin-state maps for pluripotent and lineage-committed mouse cells. The resulting data define three broad categories of promoters based on their chromatin state in embryonic stem (ES) cells, including a larger than anticipated set of 'bivalent' promoters; reveal that lineage commitment is accompanied by characteristic chromatin changes at bivalent promoters that parallel changes in gene expression and transcriptional competence; and demonstrate the potential for using ChIP for genome-wide annotation of novel promoters and primary transcripts, active transposable elements, imprinting control regions and allele-specific transcription. This study provides a technological framework for comprehensive characterization of chromatin state across diverse mammalian cell populations.

¹Broad Institute of Harvard and MIT, ²Division of Health Sciences and Technology, MIT, and ³Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA. ⁴Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA. ⁵Department of Neurology, Children's Hospital, and ⁶Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA.

*These authors contributed equally to this work.

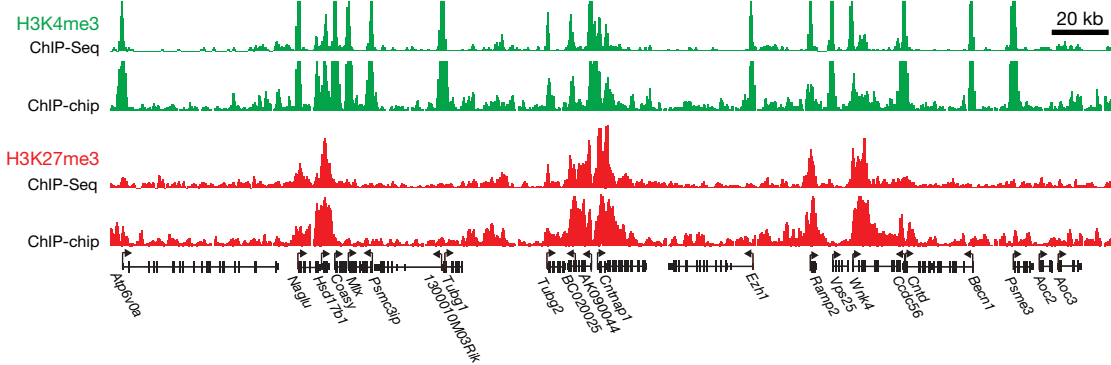


Figure 1 | Comparison of ChIP-Seq and ChIP-chip data. Direct comparison of H3K4me3 (green) and H3K27me3 (red) ChIP data across a 300-kb region in mouse ES cells from independent experiments assayed by SMS (absolute

fragment counts) or tiling arrays (log *P*-values for enrichment relative to whole-cell extracts¹⁵).

Genome-wide chromatin-state maps

We created genome-wide chromatin-state maps for three mouse cell types: ES cells, neural progenitor cells (NPCs) and embryonic fibroblasts (MEFs). For each cell type, we prepared and sequenced ChIP DNA samples for some or all of the following features: pan-H3, trimethylated histone H3 lysine 4 (H3K4me3), H3K9me3, H3K27me3, H3K36me3, H4K20me3 and RNA polymerase II (Supplementary Table 1).

In each case, we sequenced nanogram quantities of DNA fragments (~300 base pairs (bp)) on an Illumina/Solexa sequencer. We obtained an average of 10 million successful reads, consisting of the terminal 27–36 bases of each fragment. The reads were mapped to the genome and used to determine the number of ChIP fragments

overlapping any given position (Fig. 1). Enriched intervals were defined as regions where this number exceeded a threshold defined by randomization (see Methods). The full data set consists of 18 chromatin-state maps, containing ~140 million uniquely aligned reads, representing over 4 billion bases of sequence.

We validated the chromatin-state maps by computational analysis and by comparison to previous methods. ChIP-Seq maps of specific histone modifications show marked enrichment at specific locations in the genome, whereas the pan-H3 and unenriched samples show relatively uniform distributions (Supplementary Figs 1 and 2). The maps show close agreement with our previously reported ChIP-chip data from ~2.5% of the mouse genome⁹ (Fig. 1). Also, ChIP-PCR assays of 50 sites chosen to represent a range of ChIP-Seq fragment

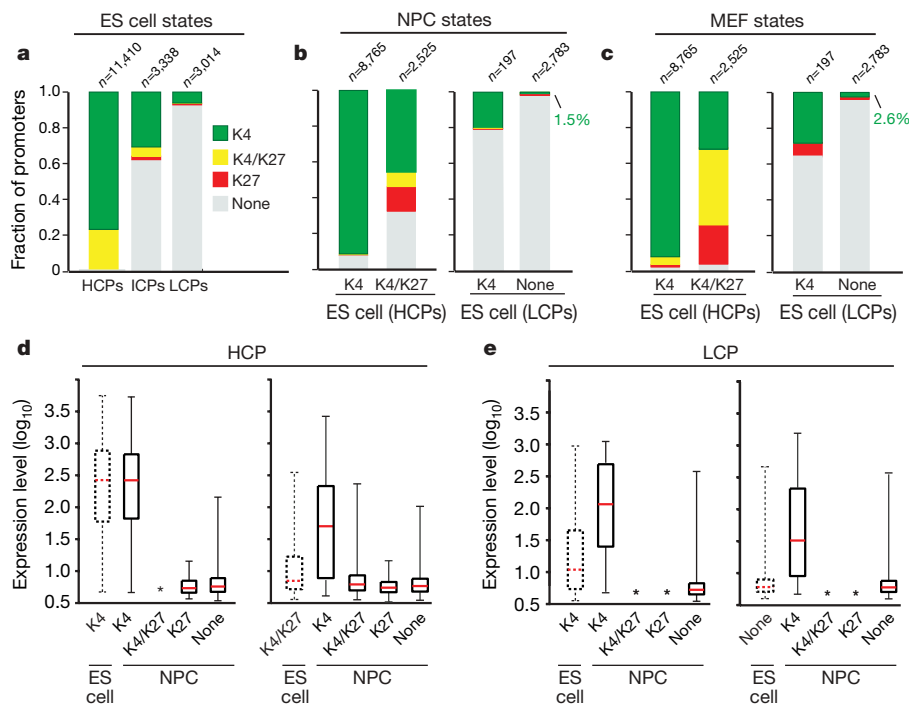


Figure 2 | Histone trimethylation state predicts expression of HCPs and LCPs. **a**, Mammalian promoters can be readily classified into sets with high (HCPs), intermediate (ICPs) or low (LCPs) CpG-content. In ES cells, virtually all HCPs are marked by H3K4me3, either alone (green) or in combination with H3K27me3 (yellow). In contrast, most LCPs have neither mark (grey). Few promoters are only enriched for H3K27me3 (red). **b**, Trimethylation states of HCPs and LCPs in NPCs (indicated by colours), conditional on their ES cell state (indicated below each bar). HCPs marked by H3K4me3 only in ES cells tend to retain this mark. HCPs marked by H3K4me3 and H3K27me3 tend to lose one or both marks, although some

remain bivalent. Small, partially overlapping subsets of LCPs are marked by H3K4me3. **c**, Trimethylation states of HCPs and LCPs in MEFs. **d**, Changes in expression levels of HCP genes with H3K4me3 alone (left) or also with H3K27me3 (right) upon differentiation to NPCs. Resolution of bivalent promoters to H3K4me3 is associated with increased expression. **e**, Changes in expression levels of LCP genes with H3K4me3 (left) or no mark (right) upon differentiation to NPCs. Gain of H3K4me3 is associated with increased expression. For **d** and **e**, boxplots show median (red bar), 25th and 75th percentile expression levels in ES cells. Whiskers show 2.5th and 97.5th percentiles. Asterisks indicate classes with less than 15 genes.

counts showed 98% concordance and a strong, quantitative correlation (Supplementary Fig. 3 and Supplementary Table 2).

Promoter state in ES and lineage-committed cells

We began our analysis by studying H3K4me3 and H3K27me3 patterns at known promoters. H3K4me3 is catalysed by trithorax-group (trxG) proteins and associated with activation, whereas H3K27me3 is catalysed by polycomb-group (PcG) proteins and associated with silencing^{10,11}. Recently, we and others observed that some promoters in ES cells carry both H3K4me3 and H3K27me3^{9,12}. We termed this novel combination a 'bivalent' chromatin mark and proposed that it serves to poise key developmental genes for lineage-specific activation or repression.

We studied 17,762 promoters inferred from full-length transcripts (Supplementary Table 3). Mammalian RNA polymerase II promoters are known to occur in at least two major forms^{13,14} (Supplementary Fig. 4). CpG-rich promoters are associated with both ubiquitously expressed 'housekeeping' genes, and genes with more complex expression patterns, particularly those expressed during embryonic development. CpG-poor promoters are generally associated with highly tissue-specific genes. Accordingly, we divided our analysis to focus on 'high' CpG promoters (HCP; $n = 11,410$) and 'low' CpG promoters (LCP; $n = 3,014$) separately. To ensure a clean separation, we defined a set of intermediate CpG content promoters (ICP; $n = 3,338$); this class shows properties consistent with being a mixture of the two major classes.

High CpG promoters in ES cells. Virtually all HCPs (99%) are associated with intervals of significant H3K4me3 enrichment in ES cells (Fig. 2a). The modified histones are typically confined to a punctate interval of 1–2 kilobases (kb) (Supplementary Fig. 5). As observed previously^{15,16}, there is a strong correlation between the intensity of H3K4me3 and the expression level of the associated genes (Spearman's $\rho = 0.67$). However, not all promoters associated with H3K4me3 are active.

The chromatin-state maps reveal that ~22% of HCPs ($n = 2,525$) are actually bivalent, exhibiting both H3K4me3 and H3K27me3 (Fig. 2a). A minority of these ($n = 564$) are 'wide' bivalent sites in which H3K27me3 extends over a region of at least 5 kb and resemble those described previously⁹. The majority ($n = 1,961$) are 'narrow' bivalent sites, with more punctate H3K27me3, that correspond to many additional PcG target promoters^{17–19}. Bivalent promoters show low activity despite the presence of H3K4me3, suggesting that the repressive effect of PcG activity is generally dominant over the ubiquitous trxG activity (Supplementary Fig. 6 and Supplementary Table 4).

The different types of chromatin marks at HCPs are closely related to the nature of the associated genes (Supplementary Table 5). Monovalent promoters (H3K4me3) generally regulate genes with 'housekeeping' functions including replication and basic metabolism. By contrast, bivalent promoters are associated with genes with more complex expression patterns, including key developmental transcription factors, morphogens and cell surface molecules. In addition, several bivalent promoters appear to regulate transcripts for lineage-specific microRNAs.

High CpG promoters in NPCs and MEFs. Most HCPs marked with H3K4me3 alone in ES cells retain this mark both in NPCs and MEFs (92% in each; Figs 2b, c and 3a). This is consistent with the tendency for this sub-class of promoters to regulate ubiquitous housekeeping genes. A small proportion (~4%) of these promoters have H3K27me3 in MEFs, and are thus bivalent or marked by H3K27me3 alone. This correlates with lower expression levels and may reflect active recruitment of PcG proteins to new genes during differentiation²⁰. An example is the transcription factor gene *Sox2*, where the promoter is marked by H3K4me3 alone in ES cells and NPCs, but H3K27me3 alone in MEFs. Notably, this locus is flanked by CpG islands with bivalent markings in ES cells (see below), suggesting that the locus may be poised for repression upon differentiation.

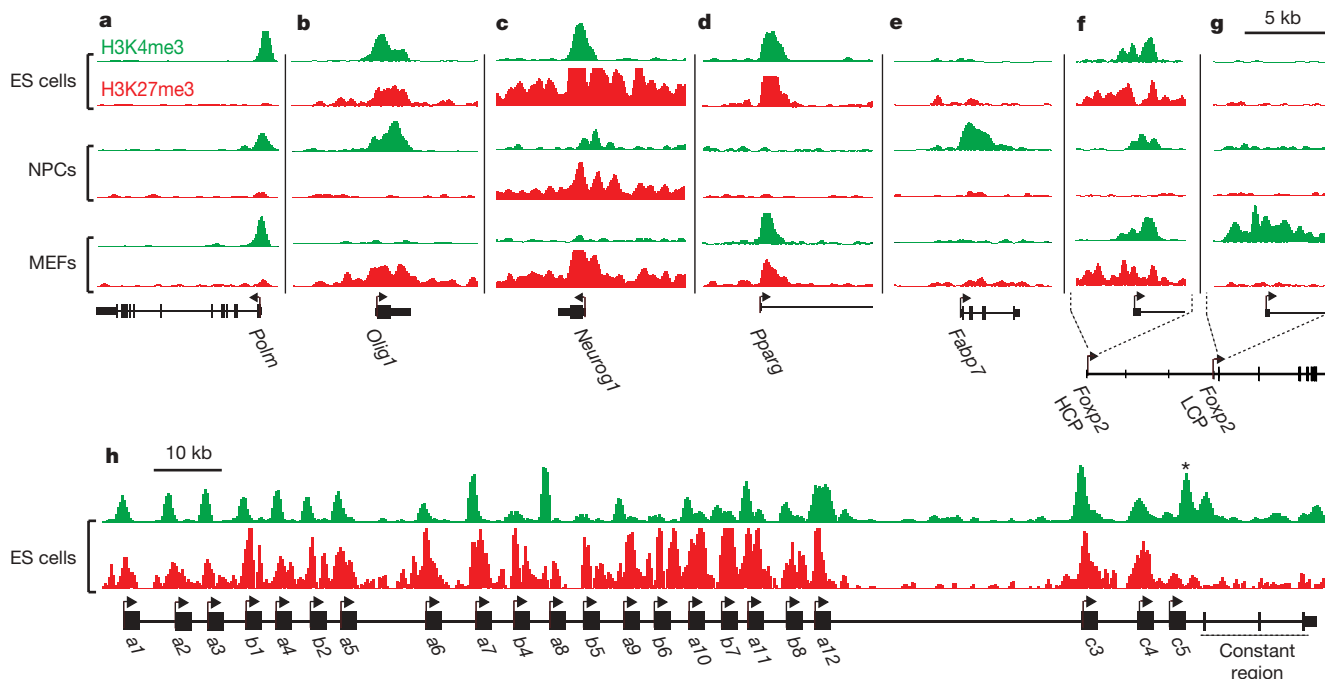


Figure 3 | Cell-type-specific chromatin marks at promoters. **a**, Multiple 'housekeeping genes', such as DNA polymerase μ (*Polm*), are associated with HCPs marked by H3K4me3 in all cell types. **b**, The neural transcription factor gene *Olig1* (HCP) is bivalent in ES cells, but resolves to H3K4me3 in NPCs and H3K27me3 in MEFs. **c**, The neurogenesis transcription factor gene *Neurog1* (HCP) remains bivalent upon differentiation to NPCs, but resolves to H3K27me3 in MEFs. **d**, The adipogenesis transcription factor gene *Pparg* (HCP) remains bivalent in MEFs, but loses both marks in NPCs.

e, The neural progenitor marker gene *Fabp7* (LCP) is marked by H3K4me3 in NPCs only. **f**, The brain and lung expressed transcription factor gene *Foxp2* is associated with an HCP that is bivalent in ES cells, but resolves to H3K4me3 in NPCs and remains bivalent in MEFs. **g**, *Foxp2* also has an LCP marked by H3K4me3 in MEFs only. **h**, Multiple, distinct bivalent chromatin marks at the variable region promoters of *Pcdhg*. A promoter proximal to the constant region (asterisk) is marked by H3K4me3 only.

The majority of HCPs with bivalent marks in ES cells resolve to a monovalent status in the committed cells. In NPCs, 46% resolve to H3K4me3 only and these genes show increased expression (Figs 2b, d and 3b). Of the remaining promoters, 14% resolve to H3K27me3 alone and 32% lose both marks, with both outcomes being associated with low levels of expression. Notably, 8% remain bivalent and these genes also continue to be repressed (Figs 2b, d and 3c). A less resolved pattern is seen in MEFs, with 32% marked by H3K4me3 alone, 22% marked by H3K27me3 alone, 3% without both marks, and the remaining (43%) still bivalent (Fig. 2c). The relatively high number of bivalent promoters in MEFs may reflect a less differentiated state and/or heterogeneity in the population.

Distinct regulation of low CpG promoters. The LCPs show a very different pattern than the HCPs. Only a small minority (6.5%, $n = 207$) of LCPs have significant H3K4me3 in ES cells and virtually none have H3K27me3 (Fig. 2a). Most of these promoters have lost H3K4me3 in NPCs and MEFs, whereas a small number of other LCPs (1.5% and 2.6%, respectively) have gained the mark (Figs 2b, c and 3e). In all three cell types, the expression levels of the associated genes strongly correlate with the presence or absence of H3K4me3 (Fig. 2e and Supplementary Fig. 6).

The genes with LCPs marked by H3K4me3 are closely related to tissue-specific functions. In NPCs, they include genes encoding several known markers of neural progenitors *in vivo* (such as *Fabp7*, *Cp*, *Gpr56*). In MEFs, they include genes encoding extracellular matrix components and growth factors (such as *Col3a1*, *Col6a1*, *Postn*, *Aspn*, *Hgf*, *Fgf*), consistent with the mesenchymal origin of these cells (see below).

We conclude that HCPs and LCPs are subject to distinct modes of regulation. In ES cells, all HCPs seem to be targets of trxB activity, and may therefore drive transcription unless actively repressed by PcG proteins. In committed cell types, a subset of HCPs appears to lose the capacity to recruit trxB activity (possibly due to other epigenetic modifications, such as DNA methylation²¹). In contrast, CpG-poor promoters seem to be inactive by default, independent of repression by PcG proteins, and may instead be selectively activated by cell-type- or tissue-specific factors.

Alternative promoter use. We note that genes with alternative promoters may have multiple, distinct chromatin states. An 'active' state at any one of these promoters may be sufficient to drive expression. A common situation involves genes with one major HCP and one or more alternative LCPs. An example is the transcription factor *Foxp2*, which is expressed at moderate levels in both NPCs and MEFs (Fig. 3f, g). The *Foxp2* HCP is marked by H3K4me3 in NPCs, but is bivalent in MEFs. However, an alternative LCP is marked by H3K4me3 exclusively in MEFs. The protocadherin- γ (*Pcdhg*) locus is a more extreme case: the amino-terminal variable regions of this gene are transcribed from at least 20 different HCPs in neurons²², all of which carry bivalent chromatin marks in ES cells. *Pcdhg* expression is nevertheless detected by microarrays, possibly owing to a single promoter in front of the carboxy-terminal constant region marked by H3K4me3 alone (Fig. 3h).

Although only ~10% of the genes analysed here have more than one known promoter, recent 'cap-trapping' studies suggest that alternative promoter use may be substantially more common²³. The ability of ChIP-Seq to assess chromatin state at known promoters, as well as to identify novel promoters (see below), should prove valuable in analysis of transcriptional networks.

Promoter state reflects lineage commitment and potential

Given their association with epigenetic memory, we next examined whether the patterns of H3K4me3 and H3K27me3 can reflect developmental potential. Both of the committed cell types studied here have been shown to be multipotent *ex vivo*. NPCs can be differentiated to glial and neuronal lineages⁸, whereas primary MEFs have been differentiated into adipocytes²⁴, chondrocytes²⁵ and osteoblast-like cells²⁶.

Lineage-specific resolution and retention of bivalent marks. We first examined a set of genes involved in *in vivo* differentiation pathways known to be, at least partially, recapitulated by MEFs, NPCs, or neither. These genes all have bivalent promoters in ES cells. We found that their resolution in lineage-committed cells is closely related to their demonstrated developmental potential (Supplementary Table 6): (1) genes restricted to regulation or specialized functions in unrelated lineages, such as haematopoietic (*Cdx4*, *PU.1* (also called *Sfp1*)), epithelial (*Cnfn*, *Krt2-4*), endoderm (*Gata6*, *Pdx1*) or germ line (*Tenr* (*Adad1*), *Ctcf*), generally resolved to monovalent H3K27me3 or carry neither mark in both NPCs and MEFs. (2) Genes related to adipogenesis and chondro/osteogenesis often remain bivalent in MEFs, but not in NPCs. Examples include *Pparg*, which is a key regulator of adipogenesis, and *Sp7*, which promotes chondro/osteogenic pathways. Early mesenchymal markers, such as *Runx1* and *Sox9*, resolved to H3K4me3 alone in MEFs. (3) Genes related to gliogenesis and neurogenesis often resolved to H3K4me3 alone or remain bivalent in NPCs, whereas they resolved to H3K27me3 alone in the MEFs. Gliogenesis and neurogenesis are thought to be mutually opposing pathways²⁷, and we find that genes promoting gliogenesis are more likely to resolve to H3K4me3 in NPCs. Examples include *Bmp2* and the microRNA *mir-9-3*, which promotes glial but inhibits neuronal differentiation²⁸. Several genes known to promote neuronal differentiation, such as *Neurog1* and *Neurog2*, remain bivalent whereas others, such as *Bmp6*, appear to resolve to H3K27me3 alone. In our hands, the NPCs differentiate to astrocytes with significantly higher efficiency than to neurons (M.W., unpublished data). The observed chromatin patterns may reflect this gliogenic bias.

Correlation with expression in adult tissues. We next analysed gene expression in adult tissues with major contributions from neuroectodermal or mesenchymal lineages. We reasoned that if H3K4me3 is generally not restored once lost, then differential loss of H3K4me3 at promoters early in these lineages (as represented by NPCs and MEFs, respectively) might be reflected in differential gene expression patterns in related adult tissues.

Notably, we observed a clear bias in relative expression levels between relevant adult tissues for genes that retain H3K4me3 in NPCs only versus genes that retain H3K4me3 in MEFs only. The former are strongly biased towards higher expression in various brain sections, whereas the latter are biased towards higher expression in bone, adipose and other mesenchyme-rich tissues (Fig. 4).

These analyses are limited by alternative promoter usage, the cell models used, and the heterogeneity of the adult tissues. Nonetheless, the data show clear trends that support an important role for retention and resolution of bivalent chromatin in the regulation of hierarchical lineage commitment.

Genome-wide annotation of promoters and primary transcripts

We next considered genome-wide maps of H3K36me3. This mark has been linked to transcriptional elongation and may serve to prevent aberrant initiation within gene bodies^{29–33}. Our chromatin maps reveal a global pattern of H3K36me3 in mammals similar to that previously observed in yeast²⁹.

In all three cell types, H3K36me3 is strongly enriched across the transcribed regions of active genes (Fig. 5a), beginning immediately after the promoter H3K4me3 signal. The level of H3K36me3 is strongly correlated with the level of gene expression (Spearman's $\rho = 0.77$), although the dynamic range is compressed (1–2 orders of magnitude for H3K36me3 versus 3–4 for expression levels; Supplementary Fig. 7). Genes with bivalent promoters rarely show H3K36me3, consistent with their low expression. Notably, there is essentially no overlap between intervals significantly enriched for H3K36me3 and for H3K27me3, consistent with a role for PcG complexes in the exclusion of polymerases¹¹.

The vast majority of intervals significantly enriched for H3K36me3 is associated with known genes (~92% in ES cells), but there are

at least ~500 additional regions across the genome (median size ~2 kb), with most being adjacent to sites of H3K4me3. Inspection revealed a number of interesting cases, falling into three categories.

The first category corresponds to H3K36me3 that extends significantly upstream from the annotated start of a known gene, often until an H3K4me3 site. These seem to reflect the presence of unannotated alternative promoters. A notable example is the *Foxp1* locus. In ES cells, one annotated *Foxp1* promoter is marked by H3K4me3 and another CpG-rich region located ~500 kb upstream carries a bivalent mark. In MEFs, this CpG island is marked by H3K4me3 only, and H3K36me3 extends from this site to the 3' end of *Foxp1* (Fig. 5a). Although no transcript extending across this entire region has been reported in mouse, the orthologous position in human has been shown to act as a promoter for the orthologous gene. The ChIP-Seq data contain many other examples where the combination of H3K36me3 and H3K4me3 seems to reveal novel promoters.

The second category corresponds to H3K36me3 that extends significantly downstream of a known gene. An example is the *Sox2* locus, which encodes a pluripotency-associated transcription factor that also functions during neural development. In ES cells, *Sox2* has an unusually large region of H3K4me3 (>20 kb) accompanied by H3K36me3 extending far beyond the annotated 3' end (>15 kb); non-coding transcription throughout the locus has been noted previously³⁴ and may serve a regulatory role (Fig. 5b).

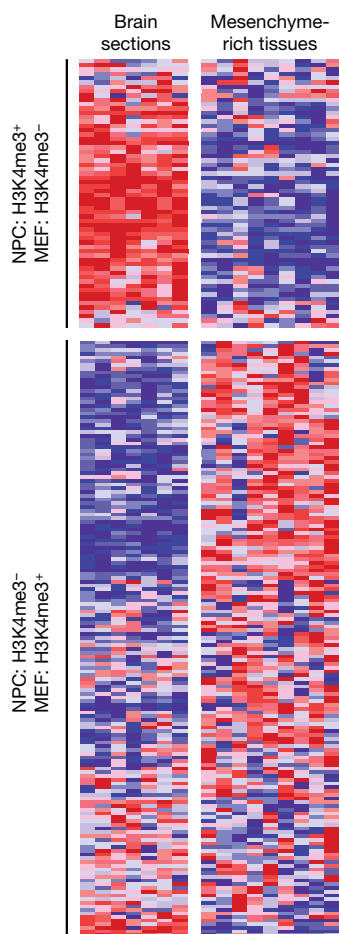


Figure 4 | Correlation between chromatin-state changes and lineage expression. Relative expression levels across adult mouse brain (frontal and cerebral cortex, substantia nigra, cerebellum, amygdale, hypothalamus, hippocampus) and relatively mesenchyme-rich tissues (bone, white fat, brown fat, trachea, digits, lung, bladder, uterus, umbilical cord) are shown for genes with bivalent chromatin marks in ES cells that retain H3K4me3 in NPCs but lose this mark in MEFs ($n = 62$) or vice versa ($n = 160$). Red, white and blue indicate higher, equal and lower relative expression, respectively.

The third category seems to reflect transcription of non-coding RNA genes. For example, two regions with H3K36me3 and adjacent H3K4me3 correspond to recently discovered nuclear transcripts with possible functions in messenger RNA processing³⁵ (Fig. 5c). In addition, a number of these presumptive transcriptional units overlap microRNAs (Fig. 5d). A striking example is a >200-kb interval within the *Dlk1-Dio3* imprinted locus (Fig. 6a). This region harbours over 40 non-coding RNAs, including clusters of microRNAs and small nucleolar RNAs³⁶. The ChIP-Seq data suggest that the entire region is transcribed as a single unit that initiates at a H3K4me3-marked HCP.

These findings suggest that genome-wide maps of H3K4me3 and H3K36me3 may provide a general tool for defining novel transcription units. The capacity to define the origins and extents of primary transcripts will be of particular value for characterizing the regulation of microRNAs and other non-coding RNAs that are rapidly processed from long precursors³⁷. Finally, the relatively narrow dynamic range of H3K36me3 may offer advantages over RNA-based approaches in assessing gene expression and defining cellular states.

H3K9 and H4K20 trimethylation mark specific repetitive elements

We next studied H3K9me3 and H4K20me3, both of which have been associated with silencing of centromeres, transposons and tandem repeats^{38–40}. We sought first to assess the relative enrichments of H3K9me3 and H4K20me3 across different types of repetitive elements by aligning ChIP-Seq reads directly to consensus sequences for various repeat families (~40 million reads could be aligned this way).

H3K9me3 and H4K20me3 show nearly identical patterns of enrichment in ES cells. The strongest enrichments are observed for telomeric, satellite and long terminal repeats (LTRs). The LTR signal

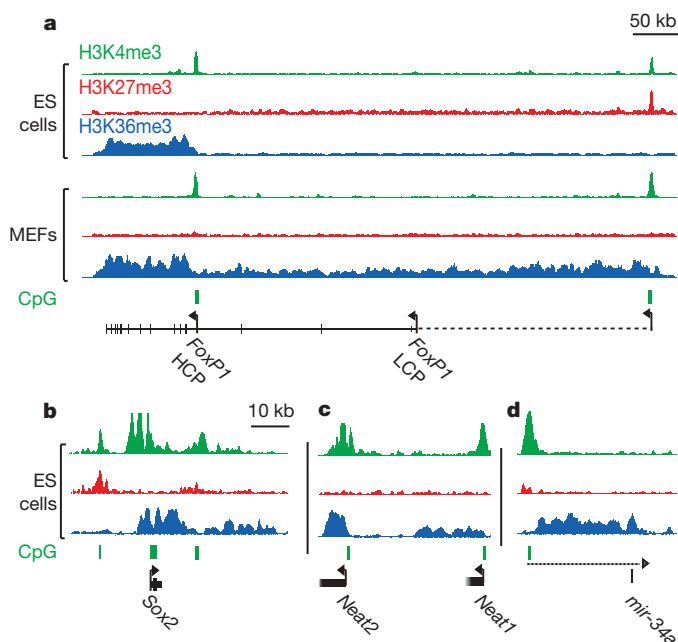


Figure 5 | H3K4me3 and H3K36me3 annotate genes and non-coding RNA transcripts. **a**, *Foxp1* has two annotated promoters (based on RefSeq and UCSC 'known genes'), only one of which shows H3K4me3 in ES cells. The corresponding transcriptional unit is marked by H3K36me3. In MEFs, H3K36me3 extends an additional 500 kb upstream to an H3K4me3 site that seems to reflect an alternative promoter (this site is bivalent in ES cells). **b**, H3K36me3 enrichment extends significantly downstream of *Sox2*. Although highly active in ES cells, *Sox2* is flanked by two bivalent CpG islands that may poise it for repression. **c**, **d**, H3K4me3 and H3K36me3 indicate two highly expressed non-coding RNAs (**c**), and the putative primary transcript (dashed line) for a single annotated microRNA (**d**).

primarily reflects enrichment of intracisternal A-particle (IAP) and early transposon (ETn) elements (Supplementary Fig. 8).

IAP and ETn elements are active in murine ES cells and produce double-stranded RNAs^{41,42}. RNA has also been implicated in maintaining satellite and telomeric heterochromatin³⁸. Hence, these enrichment data are consistent with a global role for RNA in targeting repressive chromatin marks in mammalian ES cells, analogous to that observed in lower eukaryotes^{38,39}.

We next examined the distributions of H3K9me3 and H4K20me3 across unique sequences in the mouse genome. We identified ~1,800 H3K9me3 sites (median size ~300 bp) in ES cells, with the vast majority also showing H4K20me3. Fully 78% of the sites lie within 2 kb of a satellite repeat or LTR (primarily IAP and ETn elements). This suggests that repressive marks are capable of spreading from repeat insertions and could potentially regulate proximal unique sequence.

Recent studies have described a handful of active genes with H3K9me3 and H4K20me3, raising the possibility that these 'repressive' marks also function in transcriptional activation^{31,32}. One-third of the ~1,800 H3K9me3-enriched sites reside within an annotated gene, which is roughly the proportion expected by chance. However, H3K9me3 sites that are larger and/or more distant from LTRs are more likely to occur within genes (Supplementary Fig. 9). The largest genic site in ES cells (~6 kb) coincides with the *Polrmt* gene (Fig. 6d). This case is notable because the downstream gene (*Hcn2*) is convergent and contains a CpG island at its 3' end. Transcription from 3' promoters has been proposed as a potential mechanism of transcriptional interference by producing antisense transcripts²³. This example may therefore reflect a link between transcriptional interference and H3K9me3, as has been suggested for a few other mammalian loci^{43,44}. Our results thus confirm the presence of H3K9me3

within a subset of genes, although the functional implications remain to be elucidated.

H3K4 and H3K9 trimethylation at imprinted loci

We next studied chromatin marks associated with imprinting. This epigenetic process typically involves allele-specific DNA methylation of CpG-rich imprinting control regions⁴⁵. Several reports have also described allele-specific chromatin modification at a handful of imprinting control regions, with H3K9me3 and H4K20me3 on the DNA methylated allele and H3K4me3 on the opposite allele^{46,47}.

We searched for regions showing overlapping H3K9me3 and H3K4me3 in ES cells. Notably, 13 of the top 20 sites, as ranked by enrichment of the two marks, are located within known imprinted regions, coincident with imprinting control regions or imprinted gene promoters. An example is the *Peg13* promoter (Fig. 6c). Conversely, of the ~20 known and putative autosomal imprinted loci that contain imprinting control regions, 17 have at least one with the overlapping chromatin marks (Supplementary Table 7). We conclude that overlapping H3K9me3 and H3K4me3 is a common signature of imprinting control regions in ES cells.

Allele-specific histone methylation

To explore the feasibility of inferring allele-specific chromatin states, we constructed chromatin-state maps in male ES cells derived from a more distant cross (129SvJae (maternal) x *Mus musculus castaneus* (paternal)), and used a catalogue of ~3.5 million single nucleotide polymorphisms (SNPs) to assign ChIP-Seq reads to one of the two parental alleles.

As a positive control, we first compared results for chromosome X and the autosomes for reads derived by H3K4me3 ChIP. Virtually all

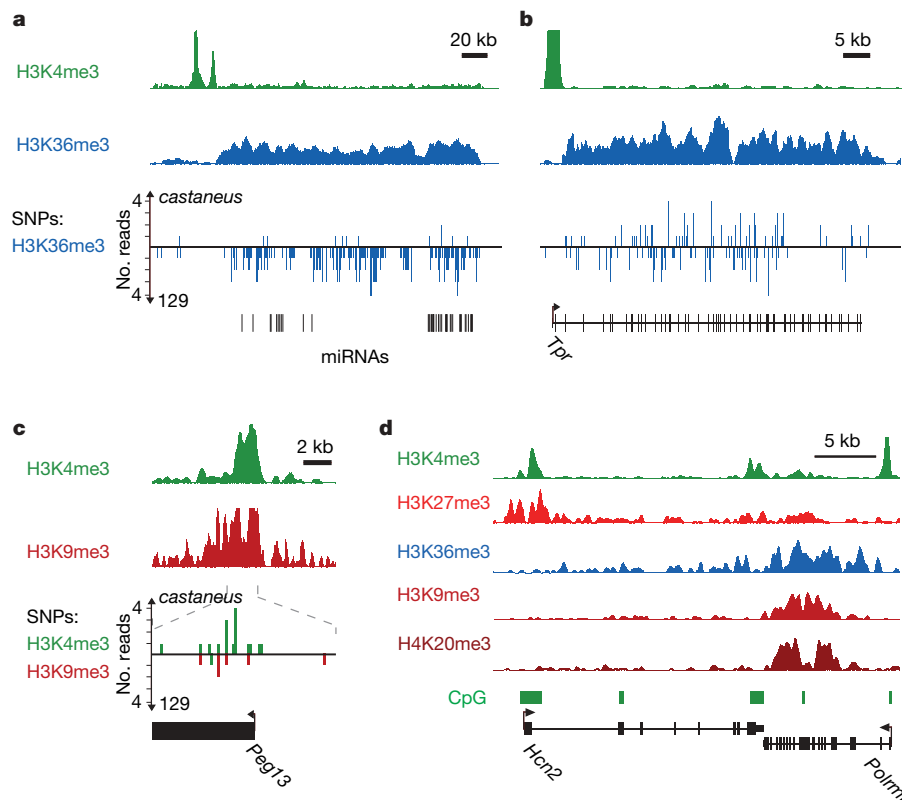


Figure 6 | Allele-specific histone methylation and genic H3K9me3/H4K20me3. **a**, H3K4me3 and H3K36me3 indicate a primary microRNA transcript in the *Dlk1-Dio3* locus. The allele specificity of this transcript is read out using ChIP-Seq data for hybrid ES cells and a SNP catalogue. The H3K36me3 reads overwhelmingly correspond to maternal 129SvJae alleles, consistent with the known maternal expression of these microRNAs³⁶. **b**, In contrast, a non-imprinted transcript shows roughly equal proportions of

reads assigned to 129SvJae and *M. m. castaneus* alleles. **c**, *Peg13* is marked by H3K4me3 and H3K9me3 in ES cells; 19 out of 21 H3K4me3 reads correspond to the paternal *M. m. castaneus* allele, whereas 6 out of 6 H3K9me3 reads correspond to the maternal 129SvJae allele, consistent with paternal expression of this gene. **d**, H3K9me3 and H4K20me3 enrichment evident at the *Polrmt* gene may reflect transcriptional interference owing to antisense transcription from the 3' UTR CpG island of *Hcn2* (see text).

(97%) of ~3,700 informative reads on chromosome X, and roughly half (57%) of the 178,000 informative reads on the autosomes, were assigned to the 129SvJae strain. These proportions correspond roughly to the expected 100% and 50%.

We then examined the allelic distribution at overlapping H3K4me3 and H3K9me3 sites coincident with putative imprinting control regions (see above). Six of the imprinting control regions had enough reads (≥ 10) containing SNPs to assess allelic bias. In every case, the SNPs showed significant bias in the expected direction ($P < 0.02$; Fig. 6c and Supplementary Table 7).

We applied the same approach to search for allelic imbalance in intervals with significant H3K36me3 enrichment, which would predict differential transcription of the two alleles. A striking interval corresponds to a microRNA cluster within the *Dlk1-Dio3* locus known to be imprinted in the embryo proper³⁶ (Fig. 6a, b). Of the additional imprinted genes with H3K36me3 enrichment, four (*Snrpn*, *Grb10*, *Impact*, *Peg3*) had enough reads containing SNPs to assess allelic bias. In every case, the data showed significant bias in the expected direction ($P < 0.02$). The data also revealed novel instances of allele-specific transcription. For example, a transcript of unknown function (*BC054101*), first identified in trophoblast stem cells⁴⁸, showed highly significant maternal bias for H3K36me3, as well as H3K4me3 ($P < 10^{-15}$; Supplementary Fig. 10).

The results suggest that, with sufficiently deep coverage and dense SNP maps, ChIP-Seq will provide a powerful means for identifying allele-specific chromatin modifications. With data from reciprocal crosses, it should be possible to discriminate novel cases of imprinting from strain-specific differences.

Discussion

Genome-wide chromatin-state maps provide a rich source of information about cellular state, yielding insights beyond what is typically obtained by RNA expression profiling. Analysis of H3K4me3 and H3K36me3 allows recognition of promoters together with their complete transcription units. This should help to define alternative promoters and their usage in specific cell types; identify the primary structure of genes encoding non-coding RNAs; detect gene expression (given the narrower dynamic range); and detect allele-specific transcription. In addition, analysis of H3K9me3 and H4K20me3 should facilitate the study of heterochromatin, spreading and imprinting mechanisms.

Most interestingly, analysis of H3K4me3 and H3K27me3 provides a rich description of cellular state. Our results suggest that promoters may be classified as active, repressed or poised for alternative developmental fates. Conceivably, chromatin state at key regulatory genes may suffice to describe developmental commitment and potential.

Given the technical features of ChIP-Seq (high throughput, low cost and input requirement), it is now appropriate to contemplate projects to generate catalogues of chromatin-state maps representing a wide range of human and mouse cell types. These should include varied developmental stages and lineages, from totipotent to terminally differentiated, with the aim of precisely defining cellular states at the epigenetic level and observing how they change over the course of normal development. Chromatin-state maps should also be systematically catalogued from situations of abnormal development. Cancer cells are the most obvious targets, as they are frequently associated with epigenetic defects and many appear to have acquired characteristics of earlier developmental stages. A comprehensive public database of chromatin-state maps would be a valuable resource for the scientific community.

METHODS SUMMARY

Murine V6.5 ES cells (129SvJae \times C57BL/6; male), hybrid ES cells (129SvJae \times *M. m. castaneus* F₁; male) and NPCs were cultured as described^{8,9}. Primary MEFs (129SvJae \times C57BL/6; male) were obtained at embryonic day (E)13.5.

ChIP experiments were carried out as described¹⁵. Sequencing libraries were generated from 1–10 ng of ChIP DNA by adaptor ligation, gel purification and 18

cycles of PCR. Sequencing was carried out using the Illumina/Solexa Genome Analyzer system according to the manufacturer's specifications.

Reads were aligned to the reference genome, and the fragment count at any given position (25-bp resolution) was estimated as the number of uniquely aligned reads oriented towards it and within 300 bp. Enriched intervals were identified by comparison of the mean fragment count in 1-kb windows against a sample-specific expected distribution estimated by randomization (H3K4me3, H3K27me3), or using a supervised Hidden Markov Model (H3K36me3, H3K9me3, H4K20me3).

Promoters were inferred from full-length mouse RefSeqs. HCPs contain a 500-bp interval within -0.5 kb to $+2$ kb with a (G+C)-fraction ≥ 0.55 and a CpG observed to expected ratio (O/E) ≥ 0.6 . LCPs contain no 500-bp interval with CpG O/E ≥ 0.4 . Chromatin states of promoters were determined by overlap with H3K4me3- and H3K27me3-enriched intervals. Correlations with expression levels were calculated from the mean fragment count over each promoter or transcript.

ES cell, NPC and MEF expression data were generated using GeneChip arrays (Affymetrix) and GenePattern (<http://www.broad.mit.edu/cancer/software/>). Expression data for adult tissues were downloaded from Novartis (<http://symatlas.gnf.org>).

Repeat class enrichments were determined by aligning reads to consensus sequences (<http://www.girinst.org>). Mouse SNP maps were obtained from Perlegen (<http://mouse.perlegen.com>). Allele-specific bias was evaluated by a binomial test of the null hypothesis that ChIP fragments were drawn uniformly from both alleles.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 May; accepted 13 June 2007.

Published online 1 July 2007.

- Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and epigenetic regulators of pluripotency. *Cell* **128**, 747–762 (2007).
- Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* **128**, 669–681 (2007).
- Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
- Buck, M. J. & Lieb, J. D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**, 349–360 (2004).
- Mockler, T. C. et al. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85**, 1–15 (2005).
- Roh, T. Y., Cuddapah, S. & Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* **19**, 542–552 (2005).
- Service, R. F. Gene sequencing. The race for the \$1000 genome. *Science* **311**, 1544–1546 (2006).
- Conti, L. et al. Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol.* **3**, e283 (2005).
- Bernstein, B. E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
- Ringrose, L. & Paro, R. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet.* **38**, 413–443 (2004).
- Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. & Cavalli, G. Genome regulation by polycomb and trithorax proteins. *Cell* **128**, 735–745 (2007).
- Azuara, V. et al. Chromatin signatures of pluripotent cell lines. *Nature Cell Biol.* **8**, 532–538 (2006).
- Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA* **103**, 1412–1417 (2006).
- Weber, M. et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genet.* **39**, 457–466 (2007).
- Bernstein, B. E. et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169–181 (2005).
- Kim, T. H. et al. A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).
- Boyer, L. A. et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349–353 (2006).
- Lee, T. I. et al. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell* **125**, 301–313 (2006).
- Squazzo, S. L. et al. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res.* **16**, 890–900 (2006).
- Pasini, D., Bracken, A. P., Hansen, J. B., Capillo, M. & Helin, K. The Polycomb Group protein Suz12 is required for Embryonic Stem Cell differentiation. *Mol. Cell Biol.* **27**, 3769–3779 (2007).
- Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31**, 89–97 (2006).

22. Wang, X., Su, H. & Bradley, A. Molecular mechanisms governing *Pcdh-γ* gene expression: evidence for a multiple promoter and *cis*-alternative splicing model. *Genes Dev.* **16**, 1890–1905 (2002).
23. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
24. Alexander, D. L., Ganem, L. G., Fernandez-Salguero, P., Gonzalez, F. & Jefcoate, C. R. Aryl-hydrocarbon receptor is an inhibitory regulator of lipid synthesis and of commitment to adipogenesis. *J. Cell Sci.* **111**, 3311–3322 (1998).
25. Lengner, C. J. *et al.* Primary mouse embryonic fibroblasts: a model of mesenchymal cartilage formation. *J. Cell. Physiol.* **200**, 327–333 (2004).
26. Garreta, E., Genove, E., Borros, S. & Semino, C. E. Osteogenic differentiation of mouse embryonic stem cells and mouse embryonic fibroblasts in a three-dimensional self-assembling peptide scaffold. *Tissue Eng.* **12**, 2215–2227 (2006).
27. Doetsch, F. The glial identity of neural stem cells. *Nature Neurosci.* **6**, 1127–1134 (2003).
28. Krichevsky, A. M., Sonntag, K. C., Isacson, O. & Kosik, K. S. Specific microRNAs modulate embryonic stem cell-derived neurogenesis. *Stem Cells* **24**, 857–864 (2006).
29. Rao, B., Shibata, Y., Strahl, B. D. & Lieb, J. D. Dimethylation of histone H3 at lysine 36 demarcates regulatory and nonregulatory chromatin genome-wide. *Mol. Cell. Biol.* **25**, 9447–9459 (2005).
30. Bannister, A. J. *et al.* Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J. Biol. Chem.* **280**, 17732–17736 (2005).
31. Kim, A., Kiefer, C. M. & Dean, A. Distinctive signatures of histone methylation in transcribed coding and noncoding human β -globin sequences. *Mol. Cell. Biol.* **27**, 1271–1279 (2007).
32. Vakoc, C. R., Sachdeva, M. M., Wang, H. & Blobel, G. A. Profile of histone lysine methylation across transcribed mammalian chromatin. *Mol. Cell. Biol.* **26**, 9185–9195 (2006).
33. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128**, 707–719 (2007).
34. Fantes, J. *et al.* Mutations in *SOX2* cause anophthalmia. *Nature Genet.* **33**, 461–463 (2003).
35. Hutchinson, J. N. *et al.* A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8**, 39 (2007).
36. Seitz, H. *et al.* A large imprinted microRNA gene cluster at the mouse *Dlk1-Gtl2* domain. *Genome Res.* **14**, 1741–1748 (2004).
37. Cullen, B. R. Transcription and processing of human microRNA precursors. *Mol. Cell* **16**, 861–865 (2004).
38. Zaratigui, M., Irvine, D. V. & Martienssen, R. A. Noncoding RNAs and gene silencing. *Cell* **128**, 763–776 (2007).
39. Verdel, A. & Moazed, D. RNAi-directed assembly of heterochromatin in fission yeast. *FEBS Lett.* **579**, 5872–5878 (2005).
40. Martens, J. H. *et al.* The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J.* **24**, 800–812 (2005).
41. Baust, C. *et al.* Structure and expression of mobile *ETnII* retroelements and their coding-competent *MusD* relatives in the mouse. *J. Virol.* **77**, 11448–11458 (2003).
42. Svoboda, P. *et al.* RNAi and expression of retrotransposons *MuERV-L* and *IAP* in preimplantation mouse embryos. *Dev. Biol.* **269**, 276–285 (2004).
43. Cho, D. H. *et al.* Antisense transcription and heterochromatin at the *DM1* CTG repeats are constrained by CTCF. *Mol. Cell* **20**, 483–489 (2005).
44. Feng, Y. Q. *et al.* The human β -globin locus control region can silence as well as activate gene expression. *Mol. Cell. Biol.* **25**, 3864–3874 (2005).
45. Edwards, C. A. & Ferguson-Smith, A. C. Mechanisms regulating imprinted genes in clusters. *Curr. Opin. Cell Biol.* **19**, 281–289 (2007).
46. Delaval, K. *et al.* Differential histone modifications mark mouse imprinting control regions during spermatogenesis. *EMBO J.* **26**, 720–729 (2007).
47. Feil, R. & Berger, F. Convergent evolution of genomic imprinting in plants and mammals. *Trends Genet.* **23**, 192–199 (2007).
48. Strausberg, R. L. *et al.* Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA* **99**, 16899–16903 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Fisher, M. Kellis, B. Birren and M. Zody for technical assistance and constructive discussions. We acknowledge L. Zagachin in the MGH Nucleic Acid Quantitation core for assistance with real-time PCR. E.M. was supported by an institutional training grant from NIH. M.W. was supported by fellowships from the Human Frontiers Science Organization Program and the Ellison Foundation. This research was supported by funds from the National Human Genome Research Institute, the National Cancer Institute, the Burroughs Wellcome Fund, Massachusetts General Hospital, and the Broad Institute of MIT and Harvard.

Author Information All analysed data sets can be obtained from http://www.broad.mit.edu/seq_platform/chip/. Microarray data have been submitted to the GEO repository under accession number GSE8024. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.S.L. (lander@broad.mit.edu) or B.E.B. (bbernstein@partners.org).

METHODS

Cell culture. V6.5 murine ES cells (genotype 129SvJae × C57BL/6; male; passages 10–15) and hybrid murine ES cells (genotype 129SvJae × *M. m. castaneus* F₁; male; passages 4–6) were cultivated in 5% CO₂ at 37 °C on irradiated MEFs in DMEM containing 15% FCS, leukaemia-inhibiting factor, penicillin/streptomycin, L-glutamine, nonessential amino acids and 2-mercaptoethanol. Cells were subjected to at least two to three passages on 0.2% gelatin under feeder-free conditions to exclude feeder contamination. V6.5 ES cells were differentiated into neural progenitor cells (NPCs) through embryoid body formation for 4 days and selection in ITSFn media for 5–7 days, and maintained in FGF2 and EGF2 (R&D Systems) as described⁸. The cells uniformly express nestin and *Sox2* and can differentiate into neurons, astrocytes and oligodendrocytes. Mouse embryonic fibroblasts (genotype 129SvJae × C57BL/6; male; E13.5; passages 4–6), were grown in DMEM with 10% fetal bovine serum and penicillin/streptomycin at 37 °C, 5% CO₂.

Chromatin immunoprecipitation. ChIP experiments were carried out as described previously¹⁵ and at <http://www.upstate.com>. Briefly, chromatin from fixed cells was fragmented to a size range of 200–700 bases with a Branson 250 Sonifier or a Diagenode Bioruptor. Solubilized chromatin was immunoprecipitated with antibody against H3K4me3 (Abcam 8580), H3K9me3 (Abcam 8898), H3K27me3 (Upstate 07-449), H3K36me3 (Abcam 9050), H4K20me3 (Upstate 07-463), pan-H3 (Abcam 1791) or RNA polymerase II (Covance MMS-126R). Antibody–chromatin complexes were pulled-down using protein A-sepharose (or anti-IgM-conjugated agarose for RNA polymerase II), washed and then eluted. After cross-link reversal and proteinase K treatment, immunoprecipitated DNA was extracted with phenol-chloroform, ethanol precipitated, and treated with RNase. ChIP DNA was quantified using PicoGreen.

Library preparation and Solexa sequencing. One to ten nanograms of ChIP DNA (or unenriched whole-cell extract) were prepared for Solexa sequencing as follows: DNA fragments were repaired to blunt ends by T4 DNA polymerase and phosphorylated with T4 polynucleotide kinase using the END-IT kit (Epicentre). Then, a single ‘A’ base was added to 3′ ends with Klenow (3′→5′ exo⁻, 0.3 U μl⁻¹). Double-stranded Solexa adaptors (75 bp with a ‘T’ overhang) were ligated to the fragments with DNA ligase (0.05 U μl⁻¹). Ligation products between 275 and 700 bp were gel purified on 2% agarose to remove unligated adaptors, and subjected to 18 PCR cycles. Completed libraries were quantified with PicoGreen.

DNA sequencing was carried out using the Illumina/Solexa Genome Analyzer sequencing system. Cluster amplification, linearization, blocking and sequencing primer reagents were provided in the Solexa Cluster Amplification kits and were used according to the manufacturer’s specifications as described here. To obtain single strand templates, the sample preparation was first denatured in NaOH (0.1 N final concentration) and diluted in Solexa hybridization buffer (4 °C) to a final concentration of either 2 or 4 pM. Sample loading was carried out as follows. A template sample was loaded into each lane of a Solexa flowcell mounted on a Solexa cluster station on which all subsequent steps were performed. The temperature was increased to 95 °C for 1 min and slowly decreased to 40 °C to allow for annealing onto complementary adaptor oligonucleotides on the flowcell surface. Cluster formation was then carried out as follows. The template strands were extended with Taq polymerase (0.25 U μl⁻¹) to generate a fixed copy of the template on the flowcell. The samples were then denatured with formamide (Sigma-Aldrich, F-5786, >99.5% (GC)) and washed (Solexa Wash buffer) to remove the original captured template, leaving behind a single-stranded template ready for amplification. Clusters were then amplified under isothermal conditions (60 °C) for 30 cycles using Solexa amplification mix containing *Bst*I DNA polymerase (0.08 U μl⁻¹). After each amplification cycle, the templates were denatured with formamide (as above). Fresh amplification mix was added after each denaturation step. After amplification, the clusters were linearized with Solexa linearization mix, and any unextended flowcell surface capture oligonucleotides were blocked with ddNTPs (2.4 μM mix in the presence of 0.25 U μl⁻¹ terminal transferase). The linearized clusters were then denatured (0.1 N NaOH) to remove and wash away the linearized strands. The single-stranded templates in the cluster were then annealed with the Solexa sequencing primer (10 μM). The flowcells were removed from the cluster station and then transferred onto the 1G Genetic Analyser which performed the sequencing according to its own standard protocols. We followed the protocol without any modifications.

Read alignment and generation of density maps and modified intervals. Sequence reads from each ChIP library are compiled, post-processed and aligned to the reference genome sequence using a general purpose computational pipeline. We first pre-compute a table that associates each possible 12-mer with all of its occurrences in the reference genome. Then, for each SMS read, we scan both it and its reverse complement, and for each of its constituent 12-mers, we find each

potential start point on the reference genome, and then compute the number of mismatches in the corresponding alignment. These computations are dynamically terminated so that only ‘unique’ alignments are reported, according to the following rule: if an alignment *A* has only *x* mismatches, and if there is no alternative alignment having $\leq x + 2$ mismatches, then we call *A* unique. To minimize the risk of amplification bias, only one read was kept if multiple reads aligned to the same start point.

For each ChIP (or control) experiment, we next estimate the number of end-sequenced ChIP fragments that overlap any given nucleotide position in the reference genome (here, at 25-bp resolution). For each position, we count the number of aligned reads that are oriented towards it and closer than the average length of a library fragment (~300 bp).

To identify the portion of the mouse genome that can be interrogated with SMS reads of a given length (*k*) and alignment stringency, we aligned every *k*-mer that occurs in the reference sequence (mm8) using the same pipeline as for SMS reads. Nucleotide positions in the reference genome where less than 50% of the 200 flanking *k*-mers on each side had ‘unique’ alignments were masked as repetitive and disregarded from further analysis (<28% of the genome). Although we analysed reads spanning 27–36 bp, all data were conservatively masked at *k* = 27.

We identified genomic intervals enriched with a specific chromatin mark from the mean fragment count in 1-kb sliding windows. To account for varying read numbers and lengths, we generated sample-specific expected distributions of fragment counts under the null hypothesis of no enrichment by moving each aligned read to a randomly chosen ‘unique’ position on the same chromosome. Nominal *P*-values for enrichment at a particular position were obtained by comparison to a randomized version of the same data set (due to the large number of reads, multiple randomizations gave identical results). Genome-wide maps of enriched sites were created by identification of windows where the nominal *P*-value fell below 10⁻⁵, and merging any enriched windows that were less than 1-kb apart into continuous intervals. To improve sensitivity to the more diffuse enrichment observed from H3K9me3 and H4K20me3 near repetitive regions and from H3K36me3 across large transcripts, we also developed a Hidden Markov Model (HMM) to segment the reference genome into ‘enriched’ and ‘unenriched’ intervals (R.P.K., manuscript in preparation). The observed fragment densities were divided into four categories, in a sample-dependent manner (‘masked’, ‘sub-threshold’, ‘near-threshold’ and ‘above threshold’). Emission and transition probabilities were fitted using supervised learning on limited intervals (~10 Mb total) chosen to reflect diverse chromatin landscapes, and the resultant models were applied genome wide.

Validation of ChIP-Seq by comparison to ChIP-chip and real-time PCR. ChIP-Seq data for H3K4me3 and H3K27me3 in ES cells were compared to published ChIP-chip profiles across ~2% of the mouse genome⁹. Significantly enriched sites in the ChIP-chip data were defined using a previously validated *P*-value threshold of 10⁻⁴, and compared to the ChIP-Seq sites. In addition, a set of 50 PCR primer pairs (Supplementary Table 2) was designed to amplify 100–140-bp fragments from genomic regions showing a wide range of signals for H3K4me3 and H3K27me3 by ChIP-Seq. Real-time PCR was carried out using Quantitect SYBR green PCR mix (Qiagen) on a 7000 ABI detection system, using 0.25 ng ChIP or WCE DNA as template. Fold enrichments reflect two independent ChIP assays, each evaluated in duplicate by real-time PCR.

Promoter classification and definition of gene and transcript intervals. The analysed promoters were based on transcription start sites inferred from full-length mouse RefSeqs (downloaded from the UCSC Genome Browser 2 April 2007). Promoters containing a 500-bp interval within -0.5 kb to +2 kb with a (G+C)-fraction ≥ 0.55 and a CpG observed to expected ratio (O/E) ≥ 0.6 were classified as HCPs. Promoters containing no 500-bp interval with CpG O/E ≥ 0.4 were classified as LCPs. The remainder were classified as ICPs. The chromatin states of promoters were determined by overlap with cell-type-specific H3K4me3 and H3K27me3 intervals. For comparison with expression levels, the chromatin states of genes with more than one known promoter were classified according to the most ‘active’ mark (that is, a gene with an H3K4me3 marked promoter and a bivalent promoter would be classified as ‘H3K4me3’). Correlation between H3K4me3 enrichment and expression levels was calculated from the mean fragment density over each promoter from -0.5 kb to +1 kb. Correlation between H3K36me3 and expression levels was calculated from the mean fragment density over each RefSeq transcript.

Expression data. RNA expression data for ES cells, NPCs and MEFs were generated from polyA RNA using GeneChip Mouse Genome 430 2.0 Arrays (Affymetrix). Expression data for adult tissues were downloaded from the Novartis Gene Expression Atlas at <http://symatlas.gnf.org>. Pre-processing, normalization (GC-RMA) and hierarchical clustering (Pearson, log-transformed, row-centred values) were performed using GenePattern (<http://www.broad.mit.edu/cancer/software/>).

Analysis of repetitive elements. Chromatin state at repetitive elements was evaluated by aligning SMS reads directly to a library of repetitive element consensus sequences (<http://www.girinst.org>). The proportion of reads aligning to each class was calculated for H3K9me3 and H4K20me3, and enrichment determined by comparison to WCE and pan-H3. We also applied an orthogonal approach based on HMM intervals of H3K9me3 in unique sequences (see above). For each repetitive element type or class, we calculated the number of occurrences within 1 kb of a unique H3K9me3 site, controlling against a set of randomly placed sites of the same length distribution.

Allele-specific histone methylation. SNPs between the 129S1/SvlmJ (used as proxy for 129SvJae) and *M. m. castaneus* mouse strains were obtained from Perlegen at <http://mouse.perlegen.com>. Allele-specific bias was evaluated by a binomial test of the null hypothesis that ChIP fragments were drawn uniformly from both alleles. (H3K4me3 and H3K9me3 reads were pooled before the test, see Supplementary Table 7.) We note that the 129SvJae strain is closer to the C57BL/6-derived reference genome, and this may cause a slight bias towards assigning aligned reads to this strain. To minimize this bias, aligned reads were kept for analysis if no alternative alignment had the same number of mismatches to the reference sequence.

LETTERS

Genome-scale DNA methylation maps of pluripotent and differentiated cells

Alexander Meissner^{1,2,3*}, Tarjei S. Mikkelsen^{2,4*}, Hongchang Gu², Marius Wernig¹, Jacob Hanna¹, Andrey Sivachenko², Xiaolan Zhang², Bradley E. Bernstein^{2,5,6}, Chad Nusbaum², David B. Jaffe², Andreas Gnirke², Rudolf Jaenisch^{1,7} & Eric S. Lander^{1,2,7,8}

DNA methylation is essential for normal development^{1–3} and has been implicated in many pathologies including cancer^{4,5}. Our knowledge about the genome-wide distribution of DNA methylation, how it changes during cellular differentiation and how it relates to histone methylation and other chromatin modifications in mammals remains limited. Here we report the generation and analysis of genome-scale DNA methylation profiles at nucleotide resolution in mammalian cells. Using high-throughput reduced representation bisulphite sequencing⁶ and single-molecule-based sequencing, we generated DNA methylation maps covering most CpG islands, and a representative sampling of conserved non-coding elements, transposons and other genomic features, for mouse embryonic stem cells, embryonic-stem-cell-derived and primary neural cells, and eight other primary tissues. Several key findings emerge from the data. First, DNA methylation patterns are better correlated with histone methylation patterns than with the underlying genome sequence context. Second, methylation of CpGs are dynamic epigenetic marks that undergo extensive changes during cellular differentiation, particularly in regulatory regions outside of core promoters. Third, analysis of embryonic-stem-cell-derived and primary cells reveals that ‘weak’ CpG islands associated with a specific set of developmentally regulated genes undergo aberrant hypermethylation during extended proliferation *in vitro*, in a pattern reminiscent of that reported in some primary tumours. More generally, the results establish reduced representation bisulphite sequencing as a powerful technology for epigenetic profiling of cell populations relevant to developmental biology, cancer and regenerative medicine.

DNA methylation can be detected by sequencing genomic DNA that has been treated with sodium bisulphite⁷. It has been impractical to apply bisulphite sequencing at a genome-wide scale because polymerase chain reaction (PCR)-based⁸ and whole-genome shotgun⁹ approaches are currently too inefficient for comparative analysis across multiple cell states in large mammalian genomes. However, reduced representations can be generated to sequence a defined fraction of a large genome^{6,10}. Computational analysis indicated that digesting mouse genomic DNA with the methylation-insensitive restriction enzyme MspI, selecting 40–220-base pair (bp) fragments, and performing 36-bp end-sequencing would cover ~1 million distinct CpG dinucleotides (4.8% of all CpGs), with roughly half located within ‘CpG islands’ (including sequences from 90% of all CpG islands) and the rest distributed between other relatively CpG-poor sequence features (Supplementary Fig. 1 and Supplementary Table

1). Notably, although CpGs are not distributed uniformly in the genome, every MspI reduced representation bisulphite sequencing (RRBS) sequence read includes at least one informative CpG position (Supplementary Fig. 2), making the approach highly efficient.

We validated high-throughput RRBS by sequencing MspI fragments from wild-type and methylation-deficient embryonic stem (ES) cells⁶, using an Illumina Genome Analyser. We generated an initial set of ~21 million high quality, aligned RRBS reads. The reads from each cell type included ~97% of the predicted non-repetitive MspI fragments (12-fold and 8-fold median coverage, respectively). This demonstrates that RRBS library construction is relatively unbiased (Supplementary Fig. 3) and is insensitive to genome-wide CpG methylation levels (estimated by nearest-neighbour analysis as 72% and 0.5%, respectively). Reads from both cell types showed near complete (>99%) bisulphite conversion of non-CpG cytosines.

To investigate cell-type-specific DNA methylation patterns, we generated 140 million additional RRBS reads (5.8 gigabase (Gb); Supplementary Information) from ES-derived neural precursor cells (NPCs) and various primary cell populations (Supplementary Table 2). We also generated new chromatin-state maps of H3 lysine 4 mono- and di-methylation (H3K4me1 and H3K4me2) from ES cells, NPCs and whole brain tissue (Supplementary Table 3 and Supplementary Information), using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq)¹¹.

The methylation levels of CpG dinucleotides in wild-type ES cells display a bimodal distribution (Fig. 1), with most being either ‘largely unmethylated’ (<20% of reads showing methylation) or ‘largely methylated’ (>80% of reads). As expected^{2,8,12}, CpGs in regions of high CpG density (>7% over 300 bp) tend to be unmethylated, whereas CpGs in low-density regions (<5%) tend to be methylated. However, we noted that ~10% of CpGs in low-density regions were unmethylated, whereas ~0.3% of CpGs in high-density regions were methylated. We found that DNA methylation patterns were better explained by histone methylation patterns than by CpG density. Because genomic features tend to be associated with distinct histone methylation patterns¹¹, we analysed these features separately.

High-CpG-density promoters (HCPs) are associated with two classes of genes: ubiquitous ‘housekeeping’ genes and highly regulated ‘key developmental’ genes¹³. In ES cells, HCPs at housekeeping genes are enriched with the transcription initiation mark H3K4me3 (‘univalent’) and are generally highly expressed, whereas those at developmental genes are enriched with both H3K4me3 and the repressive mark H3K27me3 (‘bivalent’) and are generally silent^{11,14}.

¹Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. ²Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ³Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ⁴Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ⁵Molecular Pathology Unit and Center for Cancer Research, MGH, Charlestown, Massachusetts 02129, USA. ⁶Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁷Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ⁸Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02114, USA.

*These authors contributed equally to this work.

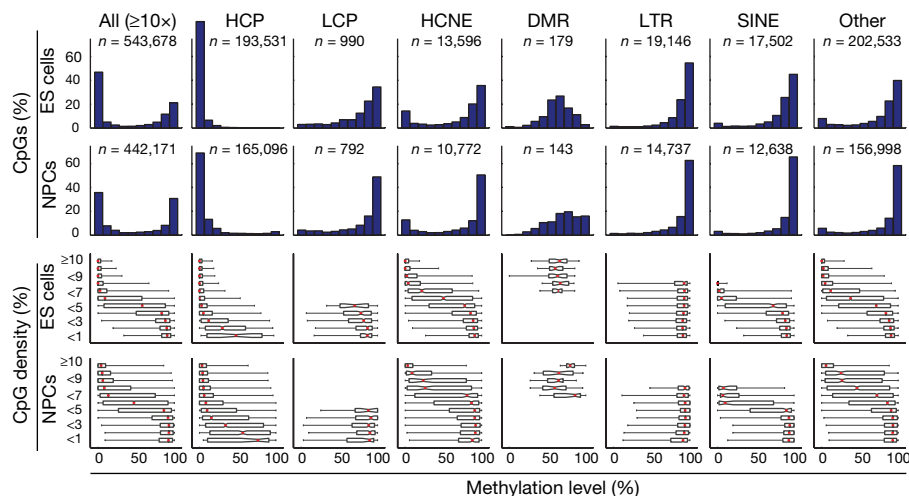


Figure 1 | CpG methylation levels in ES cells and NPCs for CpGs with ≥ 10 -fold coverage. The top histograms show the distribution of methylation levels (%) across all CpGs, HCPs, LCPs, HCNEs, differentially methylated regions (DMRs), LTRs, SINEs and other genomic features (n , number of CpGs). Methylation levels are bimodal (except at DMRs, which have a unimodal distribution largely consistent with uniform sampling from the

maternal and paternal alleles in ES cells and partial hypermethylation in NPCs). The bottom box plots show the distribution of methylation levels conditional on local CpG density (defined as fraction of CpGs in a 300-bp window; shown as percentage). The red lines denote medians, notches the standard errors, boxes the interquartile ranges, and whiskers the 2.5th and 97.5th percentiles.

Both types of promoters are also enriched with H3K4me2, which is associated with an open chromatin confirmation. Out of the 10,299 HCPs sampled (on average, 19 distinct CpGs per promoter), we found that virtually all contain a core region of unmethylated CpGs, regardless of their level of expression or H3K27me3 enrichment (Figs 1 and 2a)^{12,14,15}.

Low-CpG-density promoters (LCPs) are generally associated with tissue-specific genes. In ES cells, a small subset of LCPs are enriched with H3K4me3 (~7%) or H3K4me2 (~3%), and essentially none are enriched with H3K27me3 (ref. 11). We found that whereas most CpGs located in sampled LCPs (990 sites from 392 promoters) are methylated, those in LCPs enriched with H3K4me3 or H3K4me2 have significantly reduced methylation levels (Supplementary Fig. 4).

Distal regulatory regions such as enhancers, silencers and boundary elements are often required to establish correct gene expression patterns in mammalian cells¹⁶. *Cis*-regulatory elements active in a particular cell type are often associated with markers of open chromatin, such as H3K4me2 or H3K4me1 (refs 17, 18). We identified 25,051 sites of H3K4me2 enrichment in ES cells from 1 kb to >100 kb

away from known promoters (most were also enriched with H3K4me1, but not with H3K4me3). CpGs sampled at H3K4me2-enriched sites (outside of promoters and CpG islands) had significantly lower methylation levels than those at unenriched sites (Fig. 2b). This relationship was particularly strong for CpGs located in highly conserved non-coding elements (HCNEs; Fig. 2c).

Imprinting control regions (ICRs) are CpG-rich regulatory regions that display allele-specific histone and DNA methylation¹⁹. Our RRBS library included sequences from 13 of ~20 known ICRs (on average, 13 distinct CpGs per ICR). CpGs within these elements display a unimodal distribution of methylation levels, with a median close to 50%, which is consistent with hypomethylation of the active allele marked with H3K4me3 and hypermethylation of the silenced allele marked with H3K9me3 (Fig. 1)¹¹.

Interspersed repeat families differ in their chromatin structure, with H3K9me3 enriched at active long terminal repeats (LTRs) and to a lesser extent at long interspersed elements (LINEs), but not at short interspersed elements (SINEs). Notably, CpGs located in LTRs and LINEs are generally hypermethylated, even in CpG-rich contexts

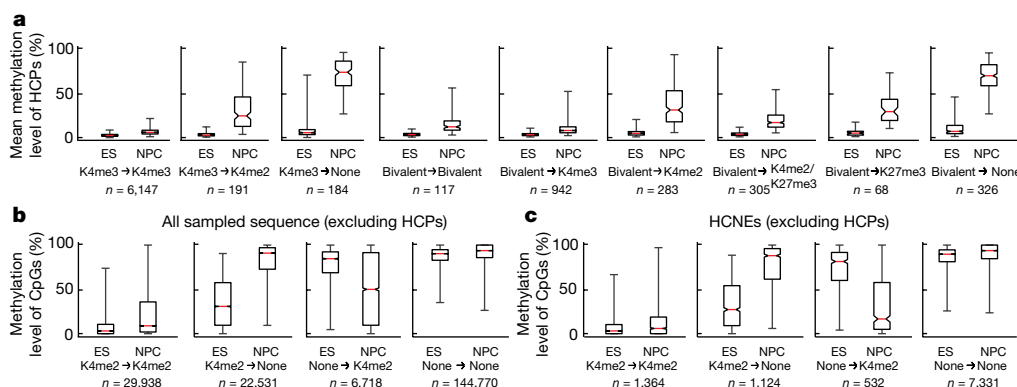


Figure 2 | Correlation between DNA and histone methylation. **a**, Mean methylation levels across CpGs within each profiled HCP (requiring ≥ 5 -fold coverage of ≥ 5 CpGs), conditional on their histone methylation state in ES cells and NPCs (n , number of HCPs; those enriched with H3K4me3 are generally also enriched for H3K4me2, but not vice versa). Loss of H3K4 methylation, and to a lesser extent of H3K27me3, is correlated with gain of DNA methylation. **b**, Methylation levels of individual CpGs outside of HCPs, conditional on enrichment of H3K4me2 (n , number of distinct sites

in each category). Changes in histone methylation state are inversely correlated with changes in DNA methylation. **c**, Methylation levels of CpGs in HCNEs not overlapping CpG islands, conditional on H3K4me2 enrichment. For **a–c**, the red lines denote medians, notches the standard errors, boxes the interquartile ranges, and whiskers the 2.5th and 97.5th percentiles. All pair-wise comparisons of methylation levels at sites with changing chromatin states are significant ($P < 10^{-20}$, Mann–Whitney U test).

(Fig. 1). In contrast, CpGs in SINEs show a correlation between methylation levels and CpG density that is comparable to non-repetitive sequences.

We conclude that in ES cells the presence of H3K4 methylation and the absence of H3K9 methylation are better predictors of unmethylated CpGs than sequence context alone. This is consistent with models in which *de novo* methyl-transferases either specifically recognize sites with unmethylated H3K4 (ref. 20) or are excluded by H3K4 methylation or associated factors. Similarly, H3K9me3 or associated factors may recruit methyl-transferases at ICRs and repetitive elements²¹.

We next used RRBS to analyse how DNA methylation patterns change when ES cells are differentiated *in vitro* into a homogeneous population of NPCs (Supplementary Fig. 4)²². Whereas CpG methylation levels are highly correlated between the two cell types ($\rho = 0.81$), there were clear differences: ~8% of CpGs unmethylated in ES cells became largely methylated in NPCs, whereas ~2% of CpGs methylated in ES cells became unmethylated; these changes were strongly correlated with changes in histone methylation patterns.

At both univalent and bivalent HCPs, we found that most CpGs remained unmethylated on differentiation, particularly within their core CpG island, but that loss of H3K4me3 and retention of H3K4me2 or H3K27me3 correlated with a partial increase in DNA methylation levels (median, ~25%; 2.9% and 32% of univalent and bivalent HCPs, respectively) and complete loss of H3K4 and H3K27 methylation correlated with DNA hypermethylation (median, ~75%; 2.8% and 16% of univalent and bivalent HCPs, respectively; Fig. 2).

Most LCPs marked by H3K4 methylation in ES cells lose this mark in NPCs; however, LCPs associated with genes expressed in NPCs gain this mark. Loss or gain of H3K4 methylation is a strong predictor of inverse changes in CpG methylation levels at these promoters (Supplementary Fig. 5).

Our chromatin-state maps revealed that 18,899 (75%) of putative distal regulatory elements enriched with H3K4me2 in ES cells lost this mark in NPCs, whereas 20,088 new H3K4me2 sites appeared, often in HCNE-rich regions surrounding activated developmental genes (Fig. 3). Loss or gain of H3K4 methylation were again inversely correlated with CpG methylation levels (Fig. 2b, c). In fact, these regions account for most observed de-methylation events. The presence of H3K27me3 alone did not correlate with lower methylation levels in CpG-poor regions (Supplementary Fig. 6).

The data support the notion that CpG-rich and -poor regulatory elements undergo distinct modes of epigenetic regulation^{2,11,12}. Most (>95%) HCPs seem to be constitutively unmethylated and regulated by trithorax-group (trxG; associated with H3K4me3) and/or Polycomb-group (PcG; associated with H3K27me3) proteins, which may be recruited in part by means of non-specific unmethylated-CpG binding domains²³. Hypermethylation of these CpG-dense regions leads to exclusion of trxG/PcG activity, heterochromatin formation and essentially irreversible gene silencing². In contrast, regulatory elements in CpG-poor sequence contexts seem to undergo extensive and dynamic methylation and de-methylation. Hence, methylation of isolated CpGs may contribute to chromatin condensation or directly interfere with transcription factor binding², but does not necessarily prevent chromatin remodelling in response to activating signals.

As noted above, a small set of HCPs ($n = 252$; ~3%) became hypermethylated (>75% mean methylation across sampled CpGs) on *in vitro* differentiation of ES cells to NPCs. To investigate whether the observed pattern reflects an *in vivo* regulatory mechanism, we isolated NPCs from embryonic day (E)13.5 embryos and differentiated them into glial fibrillary acidic protein (Gfap)-positive astrocytes (with no more than two passages *in vitro*). We similarly differentiated the *in vitro*-derived NPCs into astrocytes (with these cells having undergone at least 18 passages; Supplementary Fig. 4), and compared the two populations using RRBS (Fig. 4a–f).

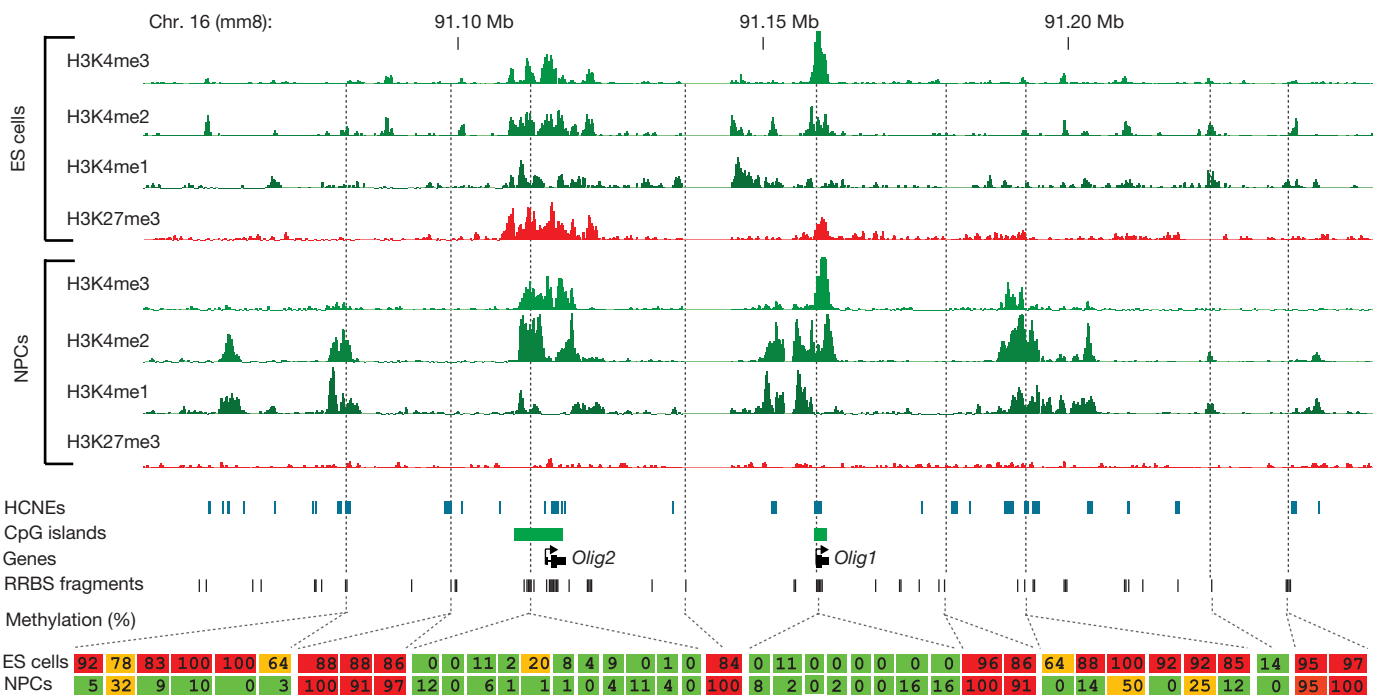


Figure 3 | Developmentally regulated de-methylation of highly conserved non-coding elements. Comparison of histone and DNA methylation levels across the *Olig1/Olig2* neural-lineage transcription factor locus. ChIP-Seq tracks for H3K4me1/2/3 and H3K27me3 in ES cells and NPCs are shown. The unmethylated CpG-rich promoters are bivalent and inactive in ES cells and resolve to univalent H3K4me3 on activation in NPCs. H3K4me2

enrichment appears over HCNEs distal to the two genes, and this correlates with CpG de-methylation. Inferred methylation levels for 40 out of 215 sampled CpGs are shown and colour-coded. Red indicates largely methylated (>80%); green indicates largely unmethylated (<20%), and orange indicates intermediate levels ($\geq 20\%$ and $\leq 80\%$).

The methylation levels of CpGs were highly correlated ($\rho = 0.85$), but astrocytes obtained from *in vivo* NPCs displayed substantially less HCP hypermethylation than those obtained from ES cells (Fig. 4a). The *in vivo*-derived astrocytes showed hypermethylation at only 30 HCPs, largely associated with germline-specific genes (including *Dazl*, *Hormad1*, *Sycp1*, *Sycp2* and *Taf7l*), several of which also showed partial methylation in ES cells. In contrast, the *in vitro*-derived astrocytes showed hypermethylation of these and ~305 additional HCPs. This set includes some genes known to be expressed by at least some *in vivo* astrocytes (including *Isynal*, *Gsn* and *Cldn5*; ref. 24) but that were silent in the ES-cell-derived astrocytes (Supplementary Information). However, the hypermethylated HCPs are significantly enriched for genes not expressed in NPCs or in the astrocyte lineage (Supplementary Tables 4–7). They include genes involved in development and differentiation of neuronal (*Lhx8*, *Lhx9*, *Moxd1*, *Htr1f* and *Slit1*), ependymal (*Otx2* and *Kl*) and unrelated lineages (including *Myod1*, *Dhh* and *Nkx3-1*). In fact, we found that ‘key developmental’ HCPs that are bivalent in ES cells are six times more likely to be included in the hypermethylated set compared to univalent HCPs. Moreover, univalent genes in the hypermethylated set are expressed at significantly lower levels in both ES cells and primary astrocytes, compared to those that remained hypomethylated (Fig. 4g). We also found that the hypermethylated HCPs tend to have a ~15% lower CpG density (Fig. 4h).

To investigate further the differences between *in vitro* and *in vivo* cell populations, we analysed whole brain tissue (representing cells of mainly glial origin). Virtually all (>99%) of sampled HCPs were

unmethylated (Fig. 4c) and enriched with H3K4me3 and/or H3K27me3 (Supplementary Fig. 7), with ~20 germline-specific HCPs being the only clear exceptions. RRBS libraries from other *in vivo* sources (T cells, B cells, spleen, lung, liver and fibroblasts) also showed few hypermethylated HCPs (Supplementary Fig. 8). This suggests that—apart from silencing germline-specific¹², imprinted and X-inactivated (Supplementary Fig. 9) genes in somatic tissues—hypermethylation of HCPs is not a major mechanism of developmental regulation *in vivo*.

To test for a correlation between passage number and HCP hypermethylation, we examined independently derived *in vitro* NPCs collected after only 9 passages. These cells displayed hypermethylation at approximately half of the HCPs that are hypermethylated in the NPCs after 18 passages (Fig. 4d, e). To reduce time in culture further, we used Sox1–GFP (green fluorescent protein) ES cells²⁵ to isolate very early NPCs. These cells initially displayed virtually no HCP hypermethylation. However, after continued culturing they acquired hypermethylation at many of the same HCPs as the previous NPC populations (Supplementary Fig. 8). Finally, we grew the *in vivo*-derived NPCs for 11 passages *in vitro*, differentiated them into astrocytes and then examined the methylation pattern. Notably, these cells had also begun to acquire hypermethylation at a largely similar set of HCPs (Fig. 4a, b).

These results show that independently derived NPC populations from both *in vitro* and *in vivo* sources and different genetic backgrounds reproducibly undergo gradual hypermethylation at a characteristic set of HCPs. These observations have several implications.

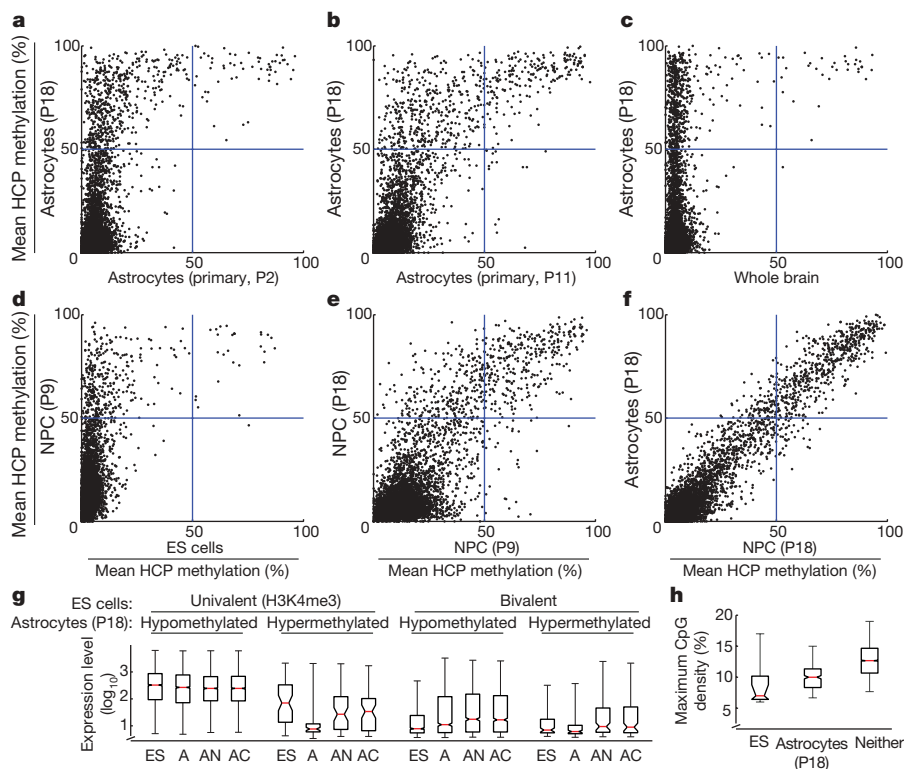


Figure 4 | HCP hypermethylation of cultured cells. Inferred mean methylation levels (%) across autosomal HCPs (requiring ≥ 5 -fold coverage of ≥ 5 CpGs within the CpG island). **a**, ES-derived astrocytes contains roughly 10 times more hypermethylated HCPs than primary NPC-derived astrocytes after two passages (P) in culture. **b**, Continued passage of the primary cells lead to gradual hypermethylation of many of the same HCPs. **c**, Only a handful of mainly germline-specific HCPs display hypermethylation in a whole brain tissue sample. **d**, Most HCPs are unmethylated in ES cells, but a small subset gain significant methylation on differentiation to NPCs. **e**, Continued proliferation of NPCs leads to additional HCPs becoming hypermethylated after 18 passages. **f**, Differentiation of late-stage NPCs into astrocytes by growth factor

withdrawal does not lead to additional HCP hypermethylation. **g**, Expression levels of genes associated with profiled HCPs for ES cells (ES), ES-derived astrocytes (A), primary neocortical astrocytes (AN) and cerebellar astrocytes (AC). Hypermethylation of HCPs is correlated with low expression levels in ES-derived astrocytes. HCPs that are univalent in ES cells and become hypermethylated in ES-derived astrocytes are associated with lower expression levels in both ES cells and primary astrocytes. **h**, The maximal CpG densities (300-bp window) of hypermethylated HCPs in ES cells or ES-derived astrocytes are significantly lower than for unmethylated HCPs. For **g** and **h**, the red lines denote medians, notches the standard errors, boxes the interquartile ranges, and whiskers the 2.5th and 97.5th percentiles.

First, aberrant epigenetic regulation in culture has raised concern over the accuracy of cellular models generated by *in vitro* differentiation or manipulation^{26–28}. Both primary and transformed cell lines, including ES-derived NPC populations, tend to lose developmental potency after continued proliferation in culture^{26,29}. Susceptibility to hypermethylation at key regulatory genes that are normally activated on differentiation could explain this phenomenon. Second, malignant cells are often found to harbour hypermethylated CpG islands^{4,5}. Recently, genes known to undergo frequent hypermethylation in adult cancers were noted to be significantly enriched for genes with bivalent promoters in ES cells (reviewed in ref. 30). The similarities between hypermethylation in culture and in cancer may provide a useful *in vitro* model for studying a common underlying mechanism. Finally, the gradual hypermethylation of ‘weak’ HCPs hints at underlying kinetics. Because H3K4 methylases are targeted, at least in part, by non-specific CpG-binding domains²³, such HCPs may be particularly sensitive to imbalanced chromatin-modifying factors or other cancer- or culture-related perturbations.

More generally, RRBS makes it feasible to perform genome-scale bisulphite sequencing on large-mammalian genomes, providing a valuable tool for epigenetic profiling of cell populations. As sequencing capacity increases, genome coverage can be readily scaled in step by adding restriction enzymes, increasing the selected size range or using hybridization-based reduced representation strategies.

METHODS SUMMARY

ES cells and ES-derived neural cells were cultured as described previously^{11,25}. Primary tissues were isolated from 4–6-week-old male 129SvJae/C57/B6 mice. Mouse embryonic fibroblasts (MEFs) and primary neural precursors were isolated from 129SvJae/C57/B6 E14.5 embryos.

RRBS libraries were prepared from 1–10 µg mouse genomic DNA digested with 10–100 Units MspI (NEB). Size-selected MspI fragments (40–120 bp and 120–220 bp) were filled in and 3′-terminal-A extended, extracted with phenol and precipitated with ethanol. Ligation to pre-annealed adapters containing 5′-methyl-cytosine instead of cytosine (Illumina) was performed using the Illumina DNA preparation kit and protocol. QIAquick (Qiagen) cleaned-up, adaptor-ligated fragments were bisulphite-treated using the EpiTect Bisulphite Kit (Qiagen). Preparative-scale PCR was performed and QIAquick-purified PCR products were subjected to a final size selection on a 4% NuSieve 3:1 agarose gel. SYBR-green-stained gel slices containing adaptor-ligated fragments of 130–210 bp or 210–310 bp in size were excised. Library material was recovered from the gel (QIAquick) and sequenced on an Illumina 1G genome analyser.

Sequence reads from bisulphite-treated Solexa libraries were identified using standard Illumina base-calling software and then analysed using a custom computational pipeline. ChIP-Seq experiments, sequencing, alignments and identification of significantly enriched regions were carried out as described previously¹¹.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 March; accepted 21 May 2008.

Published online 6 July 2008.

- Bestor, T. H. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* **9**, 2395–2402 (2000).
- Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
- Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425–432 (2007).
- Feinberg, A. P. The epigenetics of cancer etiology. *Semin. Cancer Biol.* **14**, 427–432 (2004).
- Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
- Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).

- Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl Acad. Sci. USA* **89**, 1827–1831 (1992).
- Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genet.* **38**, 1378–1385 (2006).
- Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
- Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
- Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
- Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genet.* **39**, 457–466 (2007).
- Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA* **103**, 1412–1417 (2006).
- Bernstein, B. *et al.* A bivalent chromatin structure marks key Developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
- Illingworth, R. *et al.* A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.* **6**, e22 (2008).
- West, A. G. & Fraser, P. Remote control of gene transcription. *Hum. Mol. Genet.* **14** (Spec No 1) R101–R111 (2005).
- Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39**, 311–318 (2007).
- Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169–181 (2005).
- Edwards, C. A. & Ferguson-Smith, A. C. Mechanisms regulating imprinted genes in clusters. *Curr. Opin. Cell Biol.* **19**, 281–289 (2007).
- Ooi, S. K. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to *de novo* methylation of DNA. *Nature* **448**, 714–717 (2007).
- Esteve, P. O. *et al.* Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. *Genes Dev.* **20**, 3089–3103 (2006).
- Conti, L. *et al.* Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol.* **3**, e283 (2005).
- Voo, K. S., Carlone, D. L., Jacobsen, B. M., Flodin, A. & Skalik, D. G. Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol. Cell. Biol.* **20**, 2108–2121 (2000).
- Sharma, M. K. *et al.* Distinct genetic signatures among pilocytic astrocytomas relate to their brain region origin. *Cancer Res.* **67**, 890–900 (2007).
- Aubert, J. *et al.* Screening for mammalian neural genes via fluorescence-activated cell sorter purification of neural precursors from *Sox1-gfp* knock-in mice. *Proc. Natl Acad. Sci. USA* **100** (Suppl 1), 11836–11841 (2003).
- Jones, P. A., Wolkowicz, M. J., Harrington, M. A. & Gonzales, F. Methylation and expression of the Myo D1 determination gene. *Phil. Trans. R. Soc. Lond. B* **326**, 277–284 (1990).
- Smiraglia, D. J. *et al.* Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies. *Hum. Mol. Genet.* **10**, 1413–1419 (2001).
- Shen, Y., Chow, J., Wang, Z. & Fan, G. Abnormal CpG island methylation occurs during *in vitro* differentiation of human embryonic stem cells. *Hum. Mol. Genet.* **15**, 2623–2635 (2006).
- Bouhous, I. A., Joannides, A., Kato, H., Chandran, S. & Allen, N. D. Embryonic stem cell-derived neural progenitors display temporal restriction to neural patterning. *Stem Cells* **24**, 1908–1913 (2006).
- Ohm, J. E. & Baylin, S. B. Stem cell chromatin patterns: an instructive mechanism for DNA hypermethylation? *Cell Cycle* **6**, 1040–1043 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the staff of the Broad Institute Genome Sequencing Platform for assistance with data generation and B. Ramsahoye for the nearest neighbour analysis. This research was supported by funds from the National Human Genome Research Institute, the National Cancer Institute, and the Broad Institute of MIT and Harvard.

Author Information All primary sequencing data have been submitted to the NCBI GEO repository under accession numbers GSE11034 (RRBS), GSE11172 (ChIP-Seq) and GSE11483 (gene expression microarrays). Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.J. (jaenisch@wi.mit.edu) or E.S.L. (lander@broad.mit.edu).

METHODS

Cell culture and ES cell differentiation. V6.5 (129/B6), Sox1-EGFP knock-in (Sox1-GFP 129/129)²⁵ and methylation-deficient (Dnmt1^{kd}, 3a^{-/-}, 3b^{-/-}) ES cells were expanded on γ -irradiated MEFs in DMEM plus 15% fetal bovine serum (FBS, Hyclone) supplemented with 1 \times MEM-nonessential amino acids (Life Technologies), 0.1 mM 2-mercaptoethanol and 10³ Units ml⁻¹ leukaemia inhibitory factor (LIF). After passaging onto gelatin-coated dishes (0.1% gelatin, Sigma), ES cells were trypsinized and transferred to bacterial dishes allowing embryoid body formation. Embryoid bodies were propagated for 4 days in the same medium in the absence of LIF and subsequently plated onto tissue culture dishes. One day after plating, the medium was replaced by ITSFn; that is, DMEM/F12 (Life Technologies) supplemented with 5 μ g ml⁻¹ insulin, 50 μ g ml⁻¹ human APO transferrin, 30 nM sodium selenite (all Sigma), 2.5 μ g ml⁻¹ fibronectin and penicillin/streptomycin (both Life Technologies). After 5–7 days, cells were trypsinized, triturated to a single cell suspension, replated on laminin-coated dishes (1 μ g ml⁻¹, Life Technologies) and further propagated in N3 medium composed of DMEM/F12, 25 μ g ml⁻¹ insulin, 50 μ g ml⁻¹ transferrin, 30 nM sodium selenite, 20 nM progesterone, 100 nM putrescine (Sigma), 10 ng ml⁻¹ Fgf2 (R&D Systems) and penicillin/streptomycin. Neural precursor cell proliferation was maintained by daily additions of Fgf2. Sox1-EGFP-positive neural precursors were isolated and FACS-purified (FACS Aria, Becton Dickinson) either from ITSFn cultures or after short-term expansion in Fgf2. Growth factor withdrawal of these cultures results in terminal differentiation into primarily neuronal cell populations. Neural precursor cell lines were obtained by sequential passaging and propagation in the presence of 20 ng ml⁻¹ Egf and 10 ng ml⁻¹ Fgf2 (both R&D Systems). Differentiation into astrocytes was induced by growth factor withdrawal and addition of 5% FBS for 5 days.

Primary tissues and cell types. Primary tissues were isolated from 4–6-week-old male 129SvJae/C57/B6 mice. MEFs and primary neural precursors were isolated from 129SvJae/C57/B6 E14.5 embryos. MEFs were generated according to standard protocols. *In vivo* neural precursors were isolated by disaggregating the whole brain and plating the suspension under the conditions described previously. Established lines were differentiated into astrocytes by growth factor withdrawal and addition of serum (see previously).

MspI RRBS library construction. 1–10 μ g mouse genomic DNA was digested with 10–100 Units of MspI (NEB) in a 30–500 μ l reaction 16–20 h at 37 °C. Digested DNA was extracted with phenol, precipitated with ethanol and size-selected on a 4% NuSieve 3:1 agarose gel (Lonza). DNA marker lanes were excised from the gel and stained with SYBR Green (Invitrogen). For each sample, two slices containing DNA fragments of approximately 40–120 bp and 120–220 bp, respectively, were excised from the unstained preparative portion of the gel. DNA was recovered using Easy Clean DNA spin filters (Primm Labs), extracted with phenol and precipitated with ethanol. The two size fractions were kept apart throughout the procedure, including during the final sequencing. Size-selected MspI fragments were filled in and 3'-terminal A extended in a 50 μ l reaction containing 20 U Klenow exo⁻ (NEB), 0.4 mM dATP, 0.04 mM dGTP and 0.04 mM 5-methyl-dCTP (Roche) in 1 \times NEB buffer 2 (15 min at 25 °C followed by 15 min at 37 °C), extracted with phenol and precipitated with ethanol using 10 μ g glycogen (Roche) as a carrier. Ligation to pre-annealed Illumina adapters containing 5'-methyl-cytosine instead of cytosine (Illumina) was performed using the Illumina DNA preparation kit and protocol. QIAquick (Qiagen) cleaned-up, adaptor-ligated fragments were bisulphite-treated using the EpiTect Bisulphite Kit (Qiagen) with minor modifications:

the bisulphite conversion time was increased to approximately 14 h by adding three cycles (5 min of denaturation at 95 °C followed by 3 h at 60 °C). After bisulphite conversion, the single-stranded uracil-containing DNA was eluted in 20 μ l of elution buffer. Analytical (25 μ l) PCR reactions containing 0.5 μ l of bisulphite-treated DNA, 5 pmol each of genomic PCR primers 1.1 and 2.1 (Illumina) and 2.5 U PfuTurboC_x Hotstart DNA polymerase (Stratagene) were set up to determine the minimum number of PCR cycles required to recover enough material for sequencing. Preparative-scale (8 \times 25 μ l) PCR was performed using the same PCR profile: 5 min at 95 °C, n \times (30 s at 95 °C, 20 s at 65 °C, 30 s at 72 °C), followed by 7 min at 72 °C, with n ranging from 18 to 24 cycles. QIAquick-purified PCR products were subjected to a final size selection on a 4% NuSieve 3:1 agarose gel. SYBR-green-stained gel slices containing adaptor-ligated fragments of 130–210 bp or 210–310 bp in size were excised. RRBS library material was recovered from the gel (QIAquick) and sequenced on an Illumina 1G genome analyser.

Sequence alignments and data analysis. Sequence reads from bisulphite-treated Solexa libraries were identified using standard Illumina base-calling software and then analysed using a custom computational pipeline. Residual cytosines (Cs) in each read were first converted to thymines (Ts), with each such conversion noted for subsequent analysis. A reference sequence database was constructed from the 36-bp ends of each computationally predicted MspI fragment in the 40–220-bp size range. All Cs in each fragment end were then converted to Ts (only the C-poor strands are sequenced in the RRBS process; Supplementary Fig. 2).

The converted reads were aligned to the converted reference by finding all 12-bp perfect matches and then extending to both ends of the treated read, not allowing gaps (reverse complement alignments were not considered). The number of mismatches in the induced alignment was then counted between the unconverted read and reference, ignoring cases in which a T in the unconverted read is matched to a C in the unconverted reference. For a given read, the best alignment was kept if the second-best alignment had ≥ 2 more mismatches, otherwise the read was discarded as non-unique. Low-quality reads were identified and discarded if $\sum_{q \in Q} 10^{q/10} > 1,000$, where Q denotes the read quality scores at each mismatched position. The methylation level of each sampled cytosine was estimated as the number of reads reporting a C, divided by the total number of reads reporting a C or T, counting only reads with quality scores of ≥ 20 at the position.

HCP, ICP and LCP annotations were taken from ref. 11. CpG island and other annotations were downloaded from the UCSC browser (mm8). Estimation of methylation levels from individual CpGs was limited to those with ≥ 10 -fold coverage. The methylation level of an HCP promoter was estimated as the mean methylation level across all CpGs with ≥ 5 -fold coverage overlapping the annotated CpG island(s) in the promoter, requiring at least 5 such CpGs. HCPs were classified as hypermethylated if this mean methylation level was $\geq 75\%$.

Chromatin immunoprecipitation. H3K4me1 (ab8895), H3K4me2 (ab7766) and H3K4me3 (ab8580) antibodies were purchased from Abcam. ChIP experiments on mouse ES cells (H3K4me1/2), NPCs (H3K4me1/2) and whole brain tissue (H3K4me1/2/3), Illumina/Solexa sequencing, alignments and identification of significantly enriched regions (using 1 kb sliding windows and correction for alignability) were carried out as described previously¹¹.

Expression data. RNA expression data for ES-derived astrocytes were generated as described previously¹¹ and analysed using GenePattern (<http://www.broad.mit.edu/cancer/software/genepattern/>). Primary astrocyte data were obtained from ref. 24.

Dissecting direct reprogramming through integrative genomic analysis

Tarjei S. Mikkelsen^{1,2}, Jacob Hanna⁴, Xiaolan Zhang¹, Manching Ku⁵, Marius Wernig⁴, Patrick Schorderet⁴, Bradley E. Bernstein^{1,5,6}, Rudolf Jaenisch^{3,4}, Eric S. Lander^{1,3,4,7} & Alexander Meissner^{1,8}

Somatic cells can be reprogrammed to a pluripotent state through the ectopic expression of defined transcription factors. Understanding the mechanism and kinetics of this transformation may shed light on the nature of developmental potency and suggest strategies with improved efficiency or safety. Here we report an integrative genomic analysis of reprogramming of mouse fibroblasts and B lymphocytes. Lineage-committed cells show a complex response to the ectopic expression involving induction of genes downstream of individual reprogramming factors. Fully reprogrammed cells show gene expression and epigenetic states that are highly similar to embryonic stem cells. In contrast, stable partially reprogrammed cell lines show reactivation of a distinctive subset of stem-cell-related genes, incomplete repression of lineage-specifying transcription factors, and DNA hypermethylation at pluripotency-related loci. These observations suggest that some cells may become trapped in partially reprogrammed states owing to incomplete repression of transcription factors, and that DNA de-methylation is an inefficient step in the transition to pluripotency. We demonstrate that RNA inhibition of transcription factors can facilitate reprogramming, and that treatment with DNA methyltransferase inhibitors can improve the overall efficiency of the reprogramming process.

Mouse and human cells can be reprogrammed to pluripotency through ectopic expression of defined transcription factors^{1–9} ('direct reprogramming'). Generation of such induced pluripotent stem (iPS) cells may provide an attractive source of patient-specific stem cells (reviewed in refs 10, 11). However, the mechanism and nature of molecular changes underlying the process of direct reprogramming remain largely mysterious¹¹. It is a slow and inefficient process that currently requires weeks, with most cells failing to reprogramme^{2,9,12–14}. A clearer understanding of the process would enable development of safer and more efficient reprogramming strategies, and might shed light on fundamental questions concerning the establishment of cellular identity.

To identify possible obstacles to reprogramming and to use this knowledge to devise ways to accelerate the transition to full pluripotency, we undertook a comprehensive genomic characterization of cells at various stages of the reprogramming process. The characterization involved gene expression profiling, chromatin state maps of key activating and repressive marks (histone H3 K4me3 and K27me3) and DNA methylation analysis.

Response to reprogramming factors

We first studied the response of lineage-committed cells to ectopic expression of the four reprogramming factors Oct4 (also known as Pou5f1), Sox2, Klf4 and c-Myc. Because most induced cells fail to achieve successful reprogramming, we reasoned that genomic characterization might yield insights into the basis of the low overall efficiency of the method.

To eliminate heterogeneity caused by differential viral integration, we studied mouse embryonic fibroblasts (MEFs) isolated from chimaeric mice that had been generated from an iPS cell line carrying

integrated doxycycline (Dox)-inducible lentiviral vectors with the four reprogramming factors and a *Nanog*-GFP (green fluorescent protein) reporter gene^{13,15}. We induced the expression of the reprogramming factors and obtained gene expression profiles at days 4, 8, 12 and 16 (Supplementary Data). Fluorescence-activated cell sorting (FACS) analysis on day 16 showed that ~20% of the cells stained positive for the stem-cell marker SSEA1, but only ~1.2% had achieved complete reprogramming, as indicated by activation of the *Nanog*-GFP reporter (Supplementary Fig. 1) and consistent with previous reports^{13,14}.

The immediate response to induction of the reprogramming factors (>3-fold change by day 4) is characterized by de-differentiation from the wild-type MEF state and upregulation of proliferative genes. De-differentiation is evident in a significant decrease (5–40-fold) in expression levels of typical mesenchymal genes expressed in MEFs (for example, *Snai1* and *Snai2*). The proliferative response is evident in upregulation of genes with functions such as DNA replication (*Poli*, *Rfc4* and *Mcm5*) and cell cycle progression (*Ccnd1* and *Ccnd2*); this response may be consistent with expression of reprogramming factor c-Myc^{10,16}.

We also detected a strong increase in the expression of stress-induced and anti-proliferative genes. In particular, we detected a sustained 5–10-fold upregulation of *Cdkn1a* and *Cdkn2a*, which encode cyclin-dependent kinase (CDK) inhibitors that are key effectors of multiple differentiation and tumour suppressor pathways. *Cdkn1a* is a downstream target of the reprogramming factor Klf4 (ref. 17), whereas *Cdkn2a* is known to be activated by deregulated c-Myc expression¹⁸. This response was followed by gradual upregulation of genes associated with differentiating MEFs (*Pparg*, *Fabp4* and *Mgp*) on days 12–16. This suggests that induction of the reprogramming factors

¹Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ²Division of Health Sciences and Technology, ³Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ⁴Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. ⁵Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA. ⁶Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁷Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02114, USA. ⁸Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

triggers normal 'fail-safe' mechanisms that act to prevent uncontrolled proliferation, which may prevent the majority of cells from reaching a stably de-differentiated state.

We also detected strong upregulation of lineage-specific genes from unrelated lineages. These include axon guidance factors (*Epha7* and *Ngef*), epidermal proteins (*Krt14*, *Krt16*, *Ivl* and *Sprr1a*) and glomerular proteins (*Podxl*). We speculate that this gene activation reflects responses to the reprogramming factors Sox2 and Klf4, which, independent of their roles in embryonic stem cell regulation, function in neural, epidermal and kidney differentiation^{10,17}.

Pluripotent cell lines

We next studied the changes to gene expression patterns and epigenetic states seen in successfully reprogrammed iPS cells. We analysed three cell lines: MEF-derived iPS cells carrying an Oct4-GFP reporter (MCV8.1; corresponding to subclone 8.1 in ref. 12); mature-B-lymphocyte-derived iPS cells carrying a Nanog-GFP reporter (B-iPS)¹⁵; and wild-type embryonic stem cells (V6.5)¹⁹.

We found that the genome-wide expression profiles of Oct4- or Nanog-iPS cells derived from different cell types and systems are highly similar, but not identical, to wild-type embryonic stem cells (Fig. 1), consistent with recent studies of independent cell lines^{2,4,9,20}. For example, the iPS and embryonic stem cell lines share high expression levels of genes related to maintenance of pluripotency and self-renewal such as *Oct4*, *Sox2*, *Nanog*, *Lin28*, *Zic3*, *Fgf4*, *Tdgf1* and *Rex1* (also known as *Zfp42*), and low expression levels for most lineage-specifying transcription factors and other developmental genes. Consistent with the characteristically short cell cycle of embryonic stem cells, the iPS cells show low expression of cyclin D (*Ccnd1* and *Ccnd2*)²¹.

To determine whether iPS cells have also regained embryonic-stem-cell-like chromatin states, we generated genome-wide maps showing the location of H3K4me3 and H3K27me3 from the MEF-derived MCV8.1 cell line using ChIP-Seq. Previously we described the differences in these chromatin modifications between wild-type embryonic stem cells and MEFs²². In embryonic stem cells, virtually all high-CpG promoters (HCPs) are enriched with H3K4me3; a subset of these HCPs, associated with repressed developmental genes, are also enriched with H3K27me3 ('bivalent'). In MEFs, most HCPs that are bivalent in embryonic stem cells resolve to become monovalent (H3K4me3- or H3K27me3-only). Some pluripotency- and germline-specific genes show loss of both H3K4me3 and H3K27me3 in somatic cells, and this correlates with DNA hypermethylation (ref. 23, and A.M. *et al.*, unpublished observations).

The chromatin state maps of the iPS cell line MCV8.1 are markedly similar to those of embryonic stem cells both near promoters and in intergenic regions (Fig. 2 and Supplementary Figs 2–6). Most (>97%) HCPs that lack H3K4me3-enrichment in MEFs have regained this mark in MCV8.1 cells. At all pluripotency- and germline-specific genes examined, the promoters have regained H3K4me3-enrichment and show DNA hypomethylation (Fig. 3). At genes encoding lineage-specific transcription factors that are bivalent and transcriptionally silent in embryonic stem cells, the bivalent pattern is typically re-established (~80% of HCPs classified as bivalent in wild-type embryonic stem cells, and ~95% of loci encoding key developmental transcription factors; Fig. 2b–d, g).

We conclude that direct reprogramming to a pluripotent state involves re-activation of endogenous pluripotency-related genes, establishment of an 'open' chromatin state (as indicated by genome-wide H3K4me3 enrichment and DNA de-methylation), and comprehensive Polycomb-mediated repression of lineage-specifying genes (as indicated by bivalent chromatin states involving H3K27me3-enrichment).

Partially reprogrammed cell lines

Only a subset of the stably de-differentiated cells obtained in the absence of drug selection show evidence of complete reprogramming

to a pluripotent state. Previously we derived clonal cell lines that can be maintained in relatively stable 'partially reprogrammed' states in the absence of drug selection¹². We reasoned that characterizing such cells might help to identify key barriers in the late stages of the process. Accordingly, we studied three partially reprogrammed independent cell lines established during attempts to reprogramme MEFs or mature B lymphocytes (Figs 1–3).

MCV8. This cell line, which corresponds to subclone 8 from ref. 12, was established during our attempt to reprogramme MEFs carrying an Oct4-GFP reporter with constitutive retroviruses. It produces heterogeneous cultures of cells with mainly fibroblast-like morphology, with ~20–30% positive for the stem cell marker SSEA1 (Supplementary Figs 7 and 8) and occasional interspersed embryonic-stem-cell-like colonies at late passages. Multiple secondary subclones from these embryonic-stem-cell-like colonies have been shown to establish homogeneous GFP-positive iPS cell lines

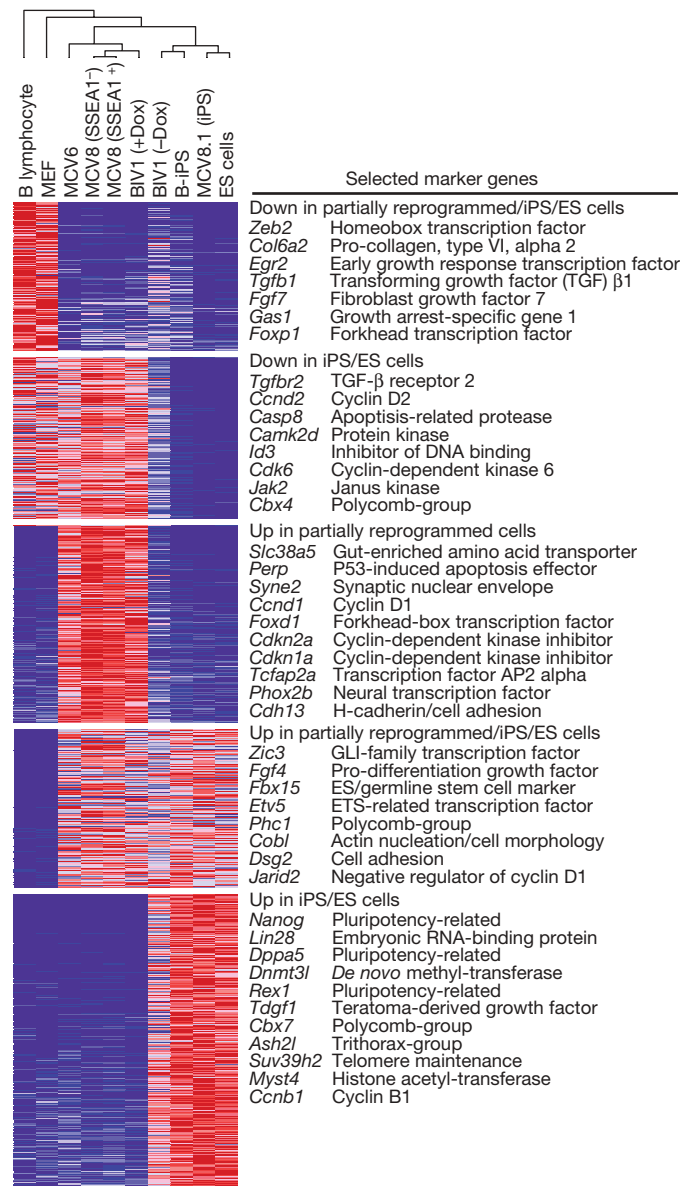


Figure 1 | Gene expression profiling. Relative expression levels across differentiated, partially reprogrammed and pluripotent cell populations. The dendrogram was generated by complete linkage hierarchical clustering using Pearson correlation on all measured genes. Only genes with at least twofold difference between any pair of samples from different classes are shown in the heat map. Red, white and blue indicate higher, identical and lower relative expression, respectively. ES cells, embryonic stem cells.

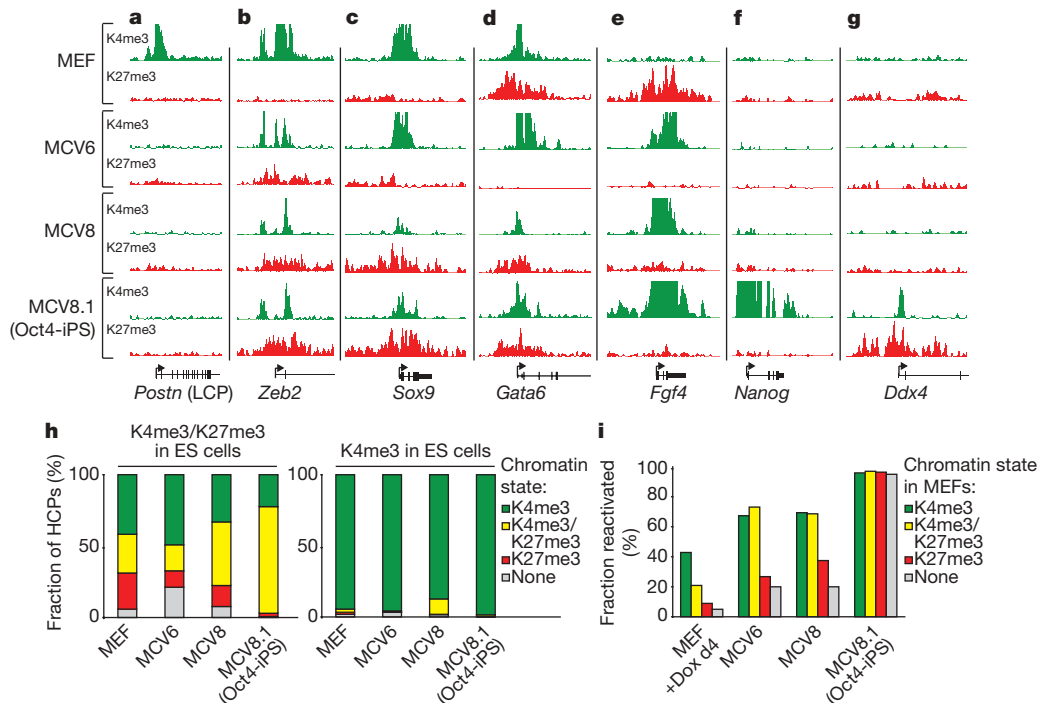


Figure 2 | Chromatin state maps. **a**, Loss of H3K4me3 correlates with inactivation of MEF-specific low-CpG promoters (LCPs), such as that of *Postn* (periostin), during reprogramming. **b**, The transcription factor *Zeb2* is marked by H3K4me3 and expressed in MEFs, but gains H3K27me3 and is silenced in partially and fully reprogrammed cells. **c**, The mesoderm/neural-crest transcription factor *Sox9* is marked by H3K4me3 only and remains active in MCV6. **d**, The endodermal transcription factor *Gata6* inappropriately lost H3K27me3 and is activated in MCV6 cells. **e**, The autocrine growth factor *Fgf4* loses H3K27me3, gains H3K4me3 and becomes highly expressed in both partially and fully reprogrammed cells. **f**, The

pluripotency gene *Nanog* gains H3K4me3 and is active only in iPS cells. **g**, The germline-specific gene *Ddx4* gains H3K4me3 and H3K27me3 in iPS cells only, and remains poised for activation in germ cells. **h**, Chromatin states for high-CpG promoters (HCPs) in MEFs and reprogrammed cells, conditional on their state in embryonic stem cells. **i**, Fraction of genes with HCPs expressed in embryonic stem cells, but not wild-type MEFs, that have been re-activated in cells at various stages of reprogramming, conditional on their chromatin state in MEFs. Most HCPs marked by H3K27me3 only or by neither mark are not re-activated in partially reprogrammed cells. d4, day 4.

(including the MCV8.1 line characterized above¹²). Proviral integration patterns showed that the same parental cells in the MCV8 population gave rise to both GFP-positive and -negative cells, suggesting that complete reprogramming depends on stochastic epigenetic events^{11,12}.

The gene expression patterns of MCV8 cells are clearly distinct from both MEFs and iPS cells (Fig. 1). MCV8 cultures show down-regulation of both structural genes (*Coll1a1* and *Coll1a2*) and regulatory factors (*Snai1*, *Snai2* and *Zeb2*) expressed in MEFs, upregulation of some lineage-specific genes with neural, epidermal or endodermal functions (presumably as a consequence of *Sox2* and *Klf4* expression), and particularly high expression of proliferative genes. Interestingly, high levels of expression can also be detected for several of the CDK inhibitors (*Cdkn1a* and *Cdkn2a*) induced by the reprogramming factors. It is unclear how the partially reprogrammed cells

have escaped the presumed anti-proliferative effects of these genes, but possible explanations include compensation by overexpression of proliferative genes, repression of differentiation pathways (MCV8 is cultured in the presence of the differentiation inhibitor LIF and expresses the LIF receptor at 2–3-fold higher levels than embryonic stem cells) or transformation (but we note that MCV8 cells have not lost the ability to re-differentiate, see below).

The pattern of re-activation of genes expressed in embryonic stem cells in MCV8 is strongly correlated with chromatin state in MEFs (Fig. 2i). Several genes related to self-renewal and proliferation of embryonic and adult stem cells show re-activation, including the autocrine growth factor *Fgf4* (ref. 24) and the transcription factor *Zic3* (ref. 25), but genes directly related to pluripotency show low or undetectable expression. Of HCPs that are enriched with H3K4me3 in MEFs but are not expressed at detectable levels, most (~70%) are

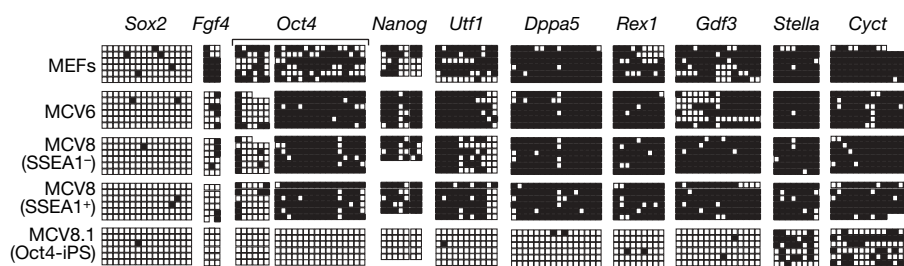


Figure 3 | DNA methylation analysis. Bisulphite sequencing of promoters or enhancers with Oct4/Sox2 binding sites near pluripotency-related and germ-cell-specific (*Stella* and *Cyct*) genes, as catalogued in ref. 23. Empty squares indicate unmethylated and filled squares methylated CpG

dinucleotides. Most assayed sites are hypermethylated in differentiated and partially reprogrammed cells. *Sox2* is enriched with H3K27me3 in non-pluripotent cells and accordingly hypomethylated in all cell types. Triangles show sites used for COBRA analysis (see text).

re-activated in MCV8. In contrast, transcriptionally silent HCPs that are enriched in MEFs for H3K27me3 only or for neither mark are significantly less likely to be re-activated ($\sim 35\%$ and $\sim 20\%$, respectively; $P_{\text{Fisher}} < 10^{-6}$).

There are notable differences in the chromatin states of MCV8, MEFs and MCV8.1 iPS cells (Fig. 2). Examining HCPs that are bivalent in embryonic stem cells demonstrates that MCV8 cells show bivalent chromatin structures at 70% more of these loci ($n = 1,467$) than seen in the MEFs ($n = 859$), but at $\sim 40\%$ fewer than in MCV8.1 iPS cells ($n = 2,360$); this is consistent with partial de-differentiation ($\sim 88\%$ of the bivalent loci in MCV8 are also bivalent in MCV8.1). There are many more HCPs that lack H3K4me3 and H3K27me3 in MCV8 than in MCV8.1 ($n = 311$ versus 31), and these genes include the majority of pluripotency- and germ-cell-specific loci. Using bisulphite sequencing, we confirmed that this chromatin state correlates with DNA hypermethylation (Fig. 3).

We initially sorted MCV8 cells into SSEA1-positive and -negative cells and analysed them separately. However, we found no major differences in expression levels or DNA methylation patterns between the two fractions (Figs 1 and 3; Supplementary Data). Moreover, when the two subpopulations were cultured separately, both reverted to a heterogeneous state within 1–2 passages (Supplementary Fig. 9). Similar results were obtained from sorting by major histocompatibility complex surface expression, which decreases on reprogramming (Supplementary Fig. 10; Supplementary Data). Thus, although these surface markers may provide some enrichment for cells that are amenable to full reprogramming¹⁴, they do not seem to discriminate between significantly different cell states within MCV8 cultures.

MCV6. This cell line was also established during our attempt to reprogramme Oct4–GFP MEFs (subclone 6 from ref. 12). It produces homogeneous cultures with compact colonies and embryonic-stem-cell-like morphology (Supplementary Fig. 8). It differs from MCV8 in that it has different proviral integrations and has never spontaneously given rise to fully reprogrammed cells (Supplementary Fig. 7).

The gene expression profile and chromatin state maps from MCV6 are largely similar to those of MCV8, but we found two notable differences. First, MCV6 has fewer genes with bivalent chromatin signatures, and a disproportionately large fraction of HCPs with neither H3K4me3- nor H3K27me3-enrichment (7% versus $\sim 2.5\%$ in MEFs and MCV8). Second, MCV6 expresses high levels of several lineage-specifying transcription factors that are expressed at low or undetectable levels in MCV8 or iPS cells, including *Sox9* (Fig. 2c) and *Gata6* (Fig. 2d). The latter observation suggests that MCV6 may have become trapped in a more differentiated state than MCV8.

BIV1. This cell line was established during our attempt to reprogramme B lymphocytes with inducible lentiviral vectors¹⁵. It had lost surface expression of all common lymphoid markers and did not require any lymphoid cytokines for growth, but also showed no evidence of achieving complete reprogramming during 50 days of continuous Dox-mediated viral expression (as judged by the absence of SSEA1- or GFP-positive cells). After Dox withdrawal and loss of any detectable viral expression (see below), the cells continued to proliferate with a more fibroblast-like morphology and, after more than ten additional days in culture, spontaneously gave rise to some GFP-positive embryonic-stem-cell-like colonies, but at a lower frequency than MCV8 (Supplementary Figs 8 and 11).

The gene expression profile and chromatin state maps from BIV1 cells grown with Dox show notable similarities to those of MCV8, including: downregulation of lineage-specific genes, such as the B lymphocyte master regulator *Pax5*; high expression of proliferative genes; activation of neural and epidermal genes; low levels of H3K4me3 and H3K27me3 enrichment relative to embryonic stem cells, consistent with DNA hypermethylation (see below); and incomplete activation of pluripotency-related loci (Fig. 1; Supplementary Figs 2–6 and 12). Notably, the expression profiles

of BIV1, MCV8 and MCV6 are more similar to each other ($r^2 > 0.9$ for any pair) than to the lineage-committed cell types from which they originated or to any of the pluripotent cell types ($r^2 < 0.8$ for any pair; Fig. 1). This suggests that the three cell lines may represent relatively common intermediate states induced by the four reprogramming factors (Oct4, Sox2, Klf4 and c-Myc). (The three lines also show expression of *Fbx15*, suggesting that they may be similar to the *Fbx15*-selected cells obtained during initial attempts to generate iPS cells⁷.)

By comparing the expression profiles of BIV1 cultures before and after Dox withdrawal, we found that Dox withdrawal resulted in: upregulation of mesenchymal extracellular matrix genes (*Col1a1* and *Col2a1*), consistent with the shift to a more fibroblast-like morphology; downregulation of most inappropriately expressed neural and epidermal genes, which is consistent with these genes being induced by overexpression of Sox2 or Klf4; and upregulation of some iPS and embryonic-stem-cell-specific genes (*Dppa5* (also known as *Dppa5a*), *Lin28* and *Dnmt3l*), which is consistent with the eventual emergence of rare GFP-positive colonies. Thus, continuous overexpression of the reprogramming factors may paradoxically have stabilized BIV1 cells in its partially reprogrammed state.

In summary, the three partially reprogrammed cell lines appear to represent similar (but distinct) cell states that emerge at an intermediate stage in the direct reprogramming process. The states are characterized by: re-activation of genes related to stem cell renewal and maintenance, but not pluripotency; incomplete repression of lineage-specific transcription factors; and incomplete epigenetic remodelling, including persistent DNA hypermethylation.

Inhibition of Dnmt1 accelerates reprogramming

Because the partially reprogrammed cell lines show DNA hypermethylation at pluripotency-related genes, we hypothesized that loss of DNA methylation (or a closely linked epigenetic mark, such as H3K9 methylation²⁶) is a critical and inefficient step in the transition from a partially reprogrammed state to pluripotency.

Partially reprogrammed cell lines. We tested this notion by treating cells with the DNA methyltransferase inhibitor 5-aza-cytidine (AZA) and found that it induced a rapid and stable transition to a fully reprogrammed iPS state. We initially studied SSEA1-positive MCV8 cells, treating them with AZA for 48 h and monitoring the subsequent appearance of GFP-positive cells (Fig. 4a and Supplementary Fig. 7). GFP-positive cells appeared at a frequency of 7.5% after one passage, compared to 0.25% in untreated cells. After five passages, GFP-positive cells comprised 77.8% of the treated population, whereas the proportion in untreated cells remained stably low (0.41%). We obtained similar results when treating the SSEA1-negative fraction. (When untreated cells from the fifth passage were subsequently treated with AZA, GFP-positive cells appeared at a similar rate as in the initial treatment; Fig. 4b.) We also found robust induction of the GFP reporter after AZA treatment of BIV1 (–Dox) cells (Fig. 4a and Supplementary Fig. 13a).

We evaluated the cellular state and developmental potency of the GFP-positive MCV8 and BIV1 cells obtained after AZA treatment and FACS. Both populations stained positive for the stem-cell marker SSEA1. Combined bisulphite restriction analysis (COBRA) revealed significant de-methylation of CpGs near the pluripotency-related genes *Dppa5*, *Nanog* and *Utf1* (Supplementary Fig. 14), implying that re-activation was not limited to the GFP-tagged reporters. The viral transgenes showed low or undetectable expression levels (Fig. 4c, d) indicating that AZA treatment did not interfere with viral silencing, which is required for full reprogramming⁹, and that the emergence of GFP-positive cells was not caused by viral re-activation. Finally, subcutaneous injection into severe combined immunodeficiency (SCID) mice led to teratoma formation in 3–4 weeks (Fig. 4e), demonstrating that the GFP-positive cells had undergone a stable transition to the pluripotent state. (Untreated MCV8 or BIV1 cells did not generate teratomas in the same time frame.)

To exclude nonspecific effects of AZA, we treated MCV8 cells with small interfering RNAs (siRNAs) or lentiviral short hairpin RNAs (shRNAs) against *Dnmt1*, which also led to the appearance GFP-positive cells within one passage (up to 1.7%; Supplementary Fig. 13b–d). We conclude that transient inhibition of *Dnmt1* is sufficient to transition MCV8 and BIV1 cells rapidly from a partially reprogrammed state to a pluripotent state.

Populations of lineage-committed cells. We next used the chimera-derived Nanog–GFP MEFs (described previously) to test

whether AZA treatment could increase the overall reprogramming efficiency. The cells were grown in the presence of Dox from day 1, and AZA was administered for 48 h starting on day 4, 6 or 8. The reprogramming efficiency was determined by counting embryonic-stem-cell-like colonies at day 14 (Fig. 4f, g).

We found that starting AZA treatment on days 4 and 6 led to high cell death and no overall gain in efficiency. The cell death may reflect the fact that most cells are still in a differentiated state: genome-wide hypomethylation is known to induce apoptosis in differentiated cells,

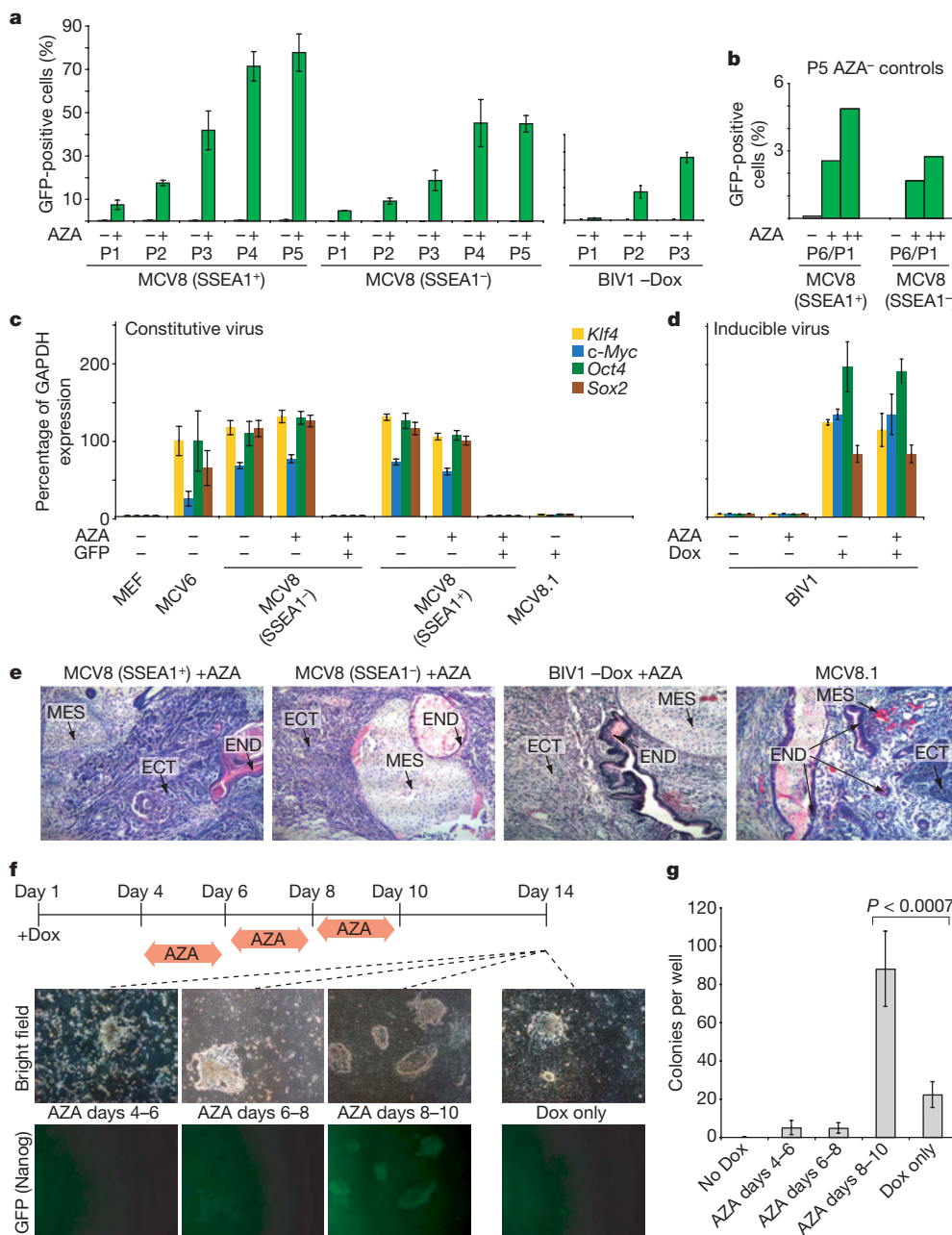


Figure 4 | Inhibition of *Dnmt1* accelerates the transition to pluripotency. **a**, MCV8 (sorted by FACS using a SSEA1-specific antibody) and BIV1 (–Dox) were either exposed to AZA for 48 h (green) or kept in regular embryonic stem cell medium (grey). The number of Oct4–GFP-positive cells was analysed over multiple passages (P) by FACS. **b**, Untreated MCV8 control cells from passage 5 were subsequently subjected to AZA treatment for 48 h (+) or 120 h (++), and resulting Oct4–GFP-positive cells were counted after one passage. P6/P1, total passage 6/passages 1 after AZA treatment. **c**, AZA treatment does not influence retroviral expression levels. **d**, AZA treatment has no influence on lentiviral expression in uninduced or induced BIV1 cells. **e**, Pluripotency of all AZA-treated lines and MCV8.1 was

demonstrated by teratoma formation. ECT, ectoderm; END, endoderm; MES, mesoderm. **f**, Overall efficiency of AZA treatment. Nanog–GFP MEFs were plated on 6-well plates (4 wells per time point with Dox, and 2 wells without). Cells were treated with AZA during one of the indicated intervals. On day 14, colony formation was analysed by fluorescence microscopy (representative panels are shown). **g**, Number of alkaline-phosphatase-positive, embryonic-stem-cell-like colonies obtained from each treatment. AZA treatment during days 8–10 resulted in a ~4-fold increase in efficiency over untreated controls. For **a**, **c**, **d** and **g**, error bars show standard deviations ($n = 2, 2, 2$ and 4 , respectively).

whereas embryonic stem cells are resistant^{27–29}. In contrast, there was a consistent fourfold increase in the number of embryonic-stem-cell-like colonies in the cultures treated with AZA starting on day 8 ($P < 0.007$; *t*-test). Moreover, most (>95%) embryonic-stem-cell-like colonies were GFP-positive in the treated cells, whereas only a minority (<25%) were GFP-positive in the untreated controls (a proportion consistent with refs 9, 12–14). Whereas early AZA treatment is counter-productive to reprogramming, there may be a sufficient number of partially reprogrammed cells in the population to outweigh its cytotoxic effect.

We conclude that de-methylation of one or more (unknown) loci is a critical step in the late stages of direct reprogramming, and that inhibition of Dnmt1 lowers this kinetic barrier, thereby facilitating the transition to pluripotency. A similar role for DNA demethylation has been reported recently during *in vivo* reprogramming in the germ line³⁰.

Transcription-factor-knockdown

In contrast to the other partially reprogrammed cell lines, MCV6 did not respond to AZA treatment (Supplementary Fig. 7). We also noted previously that MCV6 cells never show spontaneous appearance of GFP-positive colonies. We hypothesized that expression of one or more lineage-specifying transcription factor may have stabilized these cells in a more differentiated state than MCV8 or BIV1.

To test this hypothesis, we studied our genome-wide maps and identified lineage-specifying transcription factors that are expressed at low or undetectable levels in MCV8 or iPS cell populations. We transfected MCV6 cells with siRNAs against four transcription factors with >5-fold higher expression in MCV6 than in MCV8 (*Gata6*, *Pax7*, *Pax3* and *Sox9*). This resulted in no significant response. However, when transfection of siRNA targeting any one of the factors was followed by treatment with AZA for 48 h, GFP-positive cells appeared at a significant frequency in all examined populations (16 independent transfections; Fig. 5 and Supplementary Fig. 15). For example, targeting the primitive endoderm marker *Gata6* (ref. 31) generated ~2% GFP-positive cells within one passage of the subsequent AZA treatment. In contrast, no GFP-positive cells appeared in populations transfected with negative control siRNAs, or siRNAs targeted against transcription factors not expressed in MCV6 (*Zic1*

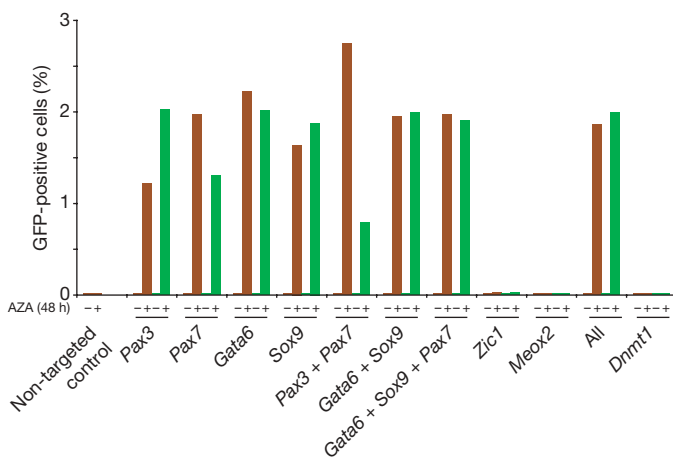


Figure 5 | Transcription factor knockdown facilitates reprogramming. MCV6 cells were plated onto 24-well dishes and transfected with siRNAs targeting expressed (*Pax7*, *Pax3*, *Gata6*, *Sox9*) or non-expressed (*Zic1*, *Meox2*) transcription factors. One plate was kept in embryonic stem cell medium and the second was exposed to AZA for 48 h. Two independent siRNA sequences were used for duplicate experiments (red and green). FACS analysis was performed 48 h after AZA treatment (96 h after transfection) without passaging. The transfection efficiency was estimated as ~20% using Cy3-coupled GADPH control siRNA.

and *Meox2*) or against *Dnmt1* (7 control populations; $P < 4 \times 10^{-4}$; Mann–Whitney U-test).

We conclude that re-activation or incomplete repression of lineage-specifying transcription factors during the reprogramming process blocks activation of the endogenous pluripotency regulatory network in MCV6. Transient silencing of one or more of these factors, combined with inhibition of Dnmt1, seems to shift the regulatory balance towards the pluripotent state, which may then be stabilized by autoregulatory feedback¹¹.

Discussion

Several insights emerge from our integrative genomic analyses. First, the Oct4/Sox2/Klf4/c-Myc-based reprogramming process appears to be fairly general, with two independent strategies (constitutive retrovirus or inducible lentivirus) and two distinct cell types (MEFs and B lymphocytes) yielding similar immediate responses, partially reprogrammed states and a similar mechanism for the final transition to pluripotency. Second, cells may fail to reprogramme successfully for several apparent reasons: the cells may induce anti-proliferative genes in response to proliferative stress; they may inappropriately activate or fail to repress endogenous or ectopic transcription factors, and become ‘trapped’ in differentiated states; and they may fail to reactivate hypermethylated pluripotency genes. Third, complete reprogramming can be facilitated by direct intervention against these failure modes, such as transient inhibition of Dnmt1 and expressed transcription factors.

We expect that further characterization of intermediate states and alternative small molecule treatments will yield critical insights that will help facilitate the desired transitions, making reprogramming efficient and safe for use in regenerative medicine. More generally, our data are consistent with a model of development in which cellular states are defined by transcription factors and stabilized by epigenetic remodelling. Integrative gene expression and epigenomic profiling provides a powerful tool for defining and guiding directed transitions between these states.

Note added in proof: The work by A.M. *et al.* cited in the text as unpublished observations has now been accepted for publication³².

METHODS SUMMARY

Embryonic stem and iPS cells were cultivated on irradiated MEFs. MEFs were infected for 16–20 h with the Moloney-based retroviral vector pLIB (Clontech) containing the complementary DNAs of *Oct4*, *Sox2*, *Klf4* and *c-Myc*. Cell lines containing the inducible lentiviruses and a ROSA26-targeted M2rTA were induced with $2 \mu\text{g ml}^{-1}$ of Dox. AZA treatment was performed for 48 h or as indicated at a concentration of 0.5 mM.

Bisulphite treatment was performed with the Qiagen EpiTect Kit. For chromatin immunoprecipitation, cells were harvested and cross-linked with formaldehyde (final concentration 1%) for 10 min at 37 °C, were washed twice with cold PBS (plus protease inhibitors), frozen and kept at –80 °C. Chromatin immunoprecipitation, library construction, sequencing, identification of enriched intervals and chromatin state classification were performed as described previously²². RNA was isolated using Trizol followed by a second round of purification using RNeasy columns (Qiagen). RNA was then processed and analysed as described elsewhere²².

Reverse transfections were performed in 24-well dishes according to manufacturer’s instructions using the siPORT NeoFX transfection agent (Ambion) and Silencer Select (Ambion/ABI) siRNAs for the respective targets.

Fluorescently conjugated antibodies were used for FACS analysis and cell sorting. Cell sorting was performed by using FACS-Aria (BD-Biosciences), and consistently achieved cell sorting purity of >97%. For determining GFP-positive cell numbers by FACS, we counted >50,000 cells.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 20 March; accepted 8 May 2008.

Published online 28 May 2008.

1. Aoi, T. *et al.* Generation of pluripotent stem cells from adult mouse liver and stomach cells. *Science*. doi:10.1126/science.1154884 (2008).

2. Maherali N. *et al.* Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cells* **1**, 55–77 (2007).
3. Nakagawa, M. *et al.* Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nature Biotechnol.* **26**, 101–106 (2008).
4. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448**, 313–317 (2007).
5. Park, I. H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141–146 (2008).
6. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
7. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
8. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).
9. Wernig, M. *et al.* *In vitro* reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448**, 318–324 (2007).
10. Yamanaka, S. Strategies and new developments in the generation of patient-specific pluripotent stem cells. *Cell Stem Cells* **1**, 39–49 (2007).
11. Jaenisch, R. & Young, R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* **132**, 567–582 (2008).
12. Meissner, A., Wernig, M. & Jaenisch, R. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nature Biotechnol.* **25**, 1177–1181 (2007).
13. Brambrink, T. *et al.* Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell* **2**, 151–159 (2008).
14. Stadtfeld, M. *et al.* Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. *Cell Stem Cell* **2**, 230–240 (2008).
15. Hanna, J. *et al.* Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. *Cell* **133**, 250–264 (2008).
16. Adhikary, S. & Eilers, M. Transcriptional regulation and transformation by Myc proteins. *Nature Rev. Mol. Cell Biol.* **6**, 635–645 (2005).
17. Rowland, B. D. & Peeper, D. S. KLF4, p21 and context-dependent opposing forces in cancer. *Nature Rev. Cancer* **6**, 11–23 (2006).
18. Gregory, M. A., Qi, Y. & Hann S. R.. The ARF tumor suppressor: keeping Myc on a leash. *Cell Cycle* **4**, 249–252 (2005).
19. Rideout, W. M. III *et al.* Generation of mice from wild-type and targeted ES cells by nuclear cloning. *Nature Genet.* **24**, 109–110 (2000).
20. Lowry, W. E. *et al.* Generation of human induced pluripotent stem cells from dermal fibroblasts. *Proc. Natl Acad. Sci. USA* **105**, 2883–2888 (2008).
21. Orford, K. W. & Scadden, D. T. Deconstructing stem cell self-renewal: genetic insights into cell-cycle regulation. *Nature Rev. Genet.* **9**, 115–128 (2008).
22. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
23. Imamura, M. *et al.* Transcriptional repression and DNA hypermethylation of a small set of ES cell marker genes in male germline stem cells. *BMC Dev. Biol.* **6**, 34 (2006).
24. Silva, J. & Smith, A. Capturing pluripotency. *Cell* **132**, 532–536 (2008).
25. Lim, L. S. *et al.* Zic3 is required for maintenance of pluripotency in embryonic stem cells. *Mol. Biol. Cell* **18**, 1348–1358 (2007).
26. Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* **128**, 669–681 (2007).
27. Jackson-Grusby, L. *et al.* Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nature Genet.* **27**, 31–39 (2001).
28. Lei, H. *et al.* *De novo* DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* **122**, 3195–3205 (1996).
29. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).
30. Hajkova, P. *et al.* Chromatin dynamics during epigenetic reprogramming in the mouse germ line. *Nature* **452**, 877–881 (2008).
31. Singh, A. M. *et al.* A heterogeneous expression pattern for Nanog in embryonic stem cells. *Stem Cells* **25**, 2534–2542 (2007).
32. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* (in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the staff of the Broad Institute Genome Sequencing Platform, Genetic Analysis Platform and RNAi Platform for assistance with reagents and data generation. This research was supported by funds from the National Institutes of Health, the National Human Genome Research Institute, the National Cancer Institute, and the Broad Institute of MIT and Harvard.

Author Information All analysed data sets can be obtained from http://www.broad.mit.edu/seq_platform/chip/. Microarray and sequence data have been submitted to the NCBI GEO database under accession numbers GSE10871 and GSE11074, respectively. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.M. (alex@broad.mit.edu).

METHODS

Viral infections and cell lines. MEFs used to derive primary iPS cell lines by infections with inducible lentiviruses were harvested at 13.5 days post coitum from F₁ matings between ROSA26–M2rtTA mice³³ and Nanog–GFP mice¹³. Secondary Nanog–GFP MEFs were isolated using neomycin selection. Lentiviral preparation and infection with Dox-inducible lentiviruses encoding *Oct4*, *Klf4*, *c-Myc* and *Sox2* cDNA driven by the tetracycline operator (TetO) and a minimal cytomegalovirus (CMV) promoter were described previously¹³. MCV6 and MCV8 were generated by retroviral infection of Oct4–GFP MEFs as described previously¹².

Cell culture. Infected MEFs or secondary inducible MEFs¹⁵ were cultured and expanded in standard embryonic stem medium and conditions¹². Culture and viral induction were performed as described^{13,15} and BIV1 was obtained as a stable line and grown under regular embryonic stem cell conditions in the presence or absence of 2 µg ml⁻¹ Dox. AZA treatment was performed for 48 h or as indicated at a concentration of 0.5 mM. Higher doses showed similar effects but increased toxicity.

Expression profiling. RNA was isolated using Trizol followed by a second round of purification using RNeasy Columns (Qiagen). RNA was then processed and analysed as described elsewhere²². Absolute expression values were Robust Multi-Array (RMA)-normalized, truncated to absolute intensity values ≥ 20 , and visualized using GenePattern (<http://www.broad.mit.edu/cancer/software/genepattern/>).

Chromatin immunoprecipitation and Illumina/Solexa sequencing. Cells were harvested and cross-linked with formaldehyde (final concentration 1%) for 10 min at 37 °C. They were washed twice with cold PBS (plus protease inhibitors), frozen and kept at –80 °C. Chromatin immunoprecipitation, library construction, sequencing, identification of enriched intervals and chromatin state classification were performed as described previously²².

Bisulphite sequencing and COBRA. Genomic DNA was isolated and bisulphite conversion was performed in a thermocycler using the Qiagen EpiTect Kit according to manufacturer's instructions with two additional cycles (5 min at 99 °C and 3 h at 60 °C) at the end. When using 2 µg genomic DNA as starting material, converted DNA was eluted in 40 µl elution buffer (Qiagen) and 2 µl were used and amplified with previously described primer sets²³ and the following additional primer pairs (CycT: GAAGGATTAATAGATGTATAAGA AAATAT; CycT: AAACCCTAATTATAAACAATAACAAC; Sox2F: GGTTTA GGAAAAGGTTGGGAATA; Sox2R: AACCAAAATAAAACAAAACCCATAA). PCR was performed in 25-µl reactions using EpiTect MSP Kit (Qiagen) mastermix according to the manufacturer's instructions with a 45 s annealing step at 50 °C (35 cycles). PCR products were gel-purified, TOPO-cloned (Invitrogen) and sequenced. COBRA for *Dppa5*, *Nanog* and *Utf1* was performed using 15 µl of the gel-purified DNA. *Dppa5* was digested for 4 h at 65 °C with Taq1 (TCGA). *Nanog* and *Utf1* were digested with HpyCHIV (ACGT) for 4 h at 37 °C. Digested products were run on 2% agarose gels.

Knockdown of transcription factors and *Dnmt1*. Reverse transfections were performed in 24-well dishes according to the manufacturer's instructions using the siPORT NeoFX transfection agent (Ambion). The following Silencer Select (Ambion/ABI) siRNAs were used (the sequence shown is the sense strand): negative control siRNA (4390843: sequence not provided), positive control *Cy3 GAPDH* siRNA (AM4649: sequence not provided), *Pax3* siRNA (s71259,

GCCCACGUCUAUCCACAA; s71260, GCUCGGAUUAUUGACUCUGA), *Pax7* siRNA (s71271, CCCUCAGUGAGUUCGAUUA; s71272, CCACAUCU GUCACAAGUA), *Gata6* siRNA (s66489, CAAAAUACUUCUCCUUCU; s66490, CCUCUGCACGCUUCCCUA), *Sox9* siRNA (s74192, AGACU CACAUCUCUCCUAA; s74193, AAGUUGAUCUGAAGCGAGA), *Meox2* siRNA (s69792, GCAGUGAAUCUAGACCUCU; s69793, GCCCAUCAU AAUUAUCUGA), *Zic1* siRNA (s76384, CAAAAGUCUGCAACAAA; s76385, GGGACUUUCUGUCCGCA) and *Dnmt1* siRNA (s65071, GGU AGAGAGUUACGACGAA; s65072, CAACGGAUCCUAUCACACU). *Dnmt1* was stably knocked down using five independent shRNAs from the RNA interference consortium (TRC; http://www.broad.mit.edu/genome_bio/trc/). shRNA1 (TRCN0000039024; target: GCTGACACTAAGCTGTTTGTGA), shRNA2 (TRCN0000039025; target: GCCTTACTTTCAACATCAAAA), shRNA3 (TRCN0000039026; target: CCGCACTTACTCCAAGTTCAA), shRNA4 (TRCN0000039027; target: CCCGAAGATCAACTCACAAA) and shRNA5 (TRCN0000039028; target: GCAAAGAGTATGAGCCAATAT). MCV8 cells were infected overnight and selected in puromycin (final, 2 µg ml⁻¹) for 48 h.

Quantitative RT–PCR. Total RNA was isolated using RNeasy Kit (Qiagen). Three micrograms of total RNA was treated with DNase I to remove potential contamination of genomic DNA using a DNA-Free RNA kit (Zymo Research). Retroviral expression levels were determined as described previously⁹. For inducible lentiviral expression, 1 µg of DNase I-treated RNA was reverse transcribed using a First Strand Synthesis kit (Invitrogen) and ultimately resuspended in 100 µl of water. Quantitative PCR analysis was performed in triplicate using 1/50 of the reverse transcription reaction in an ABI Prism 7000 (Applied Biosystems) with Platinum SYBR green qPCR SuperMix-UDG with ROX (Invitrogen). Primers used for amplification were as follows: *c-Myc*: F, 5'-ACCTAATCGAGGAGGAGCTGG-3', and R, 5'-TCCACATAGCGTAAAGGAGC-3'; *Klf4*: F, 5'-ACACTGTCTTCCCACGAGGG-3', and R, 5'-GGCATTAAAGCAGCGTATCCA-3'; *Sox2*: F, 5'-CATTAAACGGCACACTG CCC-3', and R, 5'-GGCATTAAAGCAGCGTATCCA-3'; *Oct4*: F, 5'-AGCCTGGCCTGTCTGTCACCT-3', and R, 5'-GGCATTAAAGCAGC GTATCCA-3'. To ensure equal loading of cDNA into qRT–PCR reactions, *GAPDH* messenger RNA was amplified using the following primers: F, 5'-TTCACCACCATGGAGAAGGC-3', and R, 5'-CCCTTTGGCTCCACCCT-3'. Data were extracted from the linear range of amplification. All graphs of qRT–PCR data shown represent samples of RNA that were DNase-treated, reverse transcribed, and amplified in parallel to avoid variation inherent in these procedures.

Flow cytometry analysis and cell sorting. The following fluorescently conjugated antibodies (PE, FITC, Cy-Chrome or APC-labelled) were used for FACS analysis and cell sorting: anti-SSEA1 (RnD Systems), anti-Igκ, anti-Igλ1,2,3, anti-CD19, anti-B220, anti-sIgM and anti-sIgD (all obtained from BD-Biosciences). Cell sorting was performed by using FACS-Aria (BD-Biosciences), and consistently achieved cell sorting purity of >97%. For determining GFP-positive cell numbers by FACS, we counted >50,000 cells.

33. Beard, C. *et al.* Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. *Genesis* **44**, 23–28 (2006).