Functional large non-coding RNAs in mammals

by

Mitchell Guttman

B.A. Molecular Biology and Computational Biology,
University of Pennsylvania, 2006

M.S. Computational Biology and Bioinformatics,
University of Pennsylvania, 2006

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Signature of the Author ....
Department of Biology
May 29, 2012

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

Accepted by ....
Stephen P. Bell, PhD
Professor of Biology
Chairman, Graduate Committee

Functional large non-coding RNAs in mammals

Mitchell Guttman

Submitted to the Department of Biology on May 29, 2012 in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

It is now clear that RNA is more than a messenger and performs vast and diverse functions. These functional RNAs include the ribosomal, transfer, and splicing-associated RNAs along with a cast of tiny RNAs, including microRNAs and other families. In addition to these classic examples, there were a handful of known functional large ncRNAs that play important biological roles.

To identify additional functional large ncRNAs we exploited a chromatin signature of actively transcribed genes to define discrete transcriptional units that do not overlap any known protein-coding genes. Using this approach we identified ~3,500 transcriptional units in the human and mouse genomes that produce multi-exonic RNAs that lack any coding potential. We termed these large intergenic non-coding RNAs (lincRNAs). Importantly, these lincRNAs exhibit strong purifying selection across various mammalian genomes.

To determine whether the lincRNA transcripts themselves have biological functions, we undertook systematic loss-of-function experiments on most lincRNAs defined in mouse embryonic stem cells (ESCs). We showed that knockdown of the vast majority of ESC-expressed lincRNAs has a strong effect on gene expression patterns in ESCs, of comparable magnitude to that seen for the well-known ESC regulatory proteins. We identify dozens of lincRNAs that upon loss-of-function cause an exit from the pluripotent state and dozens of additional lincRNAs that, while not essential for the maintenance of pluripotency, act to repress lineage-specific gene expression programs in ESCs.

Despite their important functional roles, how lincRNAs exert their influence was unknown. We showed that many lincRNAs physically interact with the Polycomb Repressive Complex. We systematically analyzed chromatin-modifying proteins that have been shown to play critical roles in ESCs and identified 11 additional chromatin complexes that physically interact with the ESC lincRNAs. Altogether, we found that ~30% of the ESC lincRNAs are associated with multiple chromatin complexes. These interactions are important for proper regulation of gene expression programs in ES cells.

Our data suggests a model whereby a distinct set of lincRNAs is transcribed in a cell type and interacts with ubiquitous regulatory protein complexes to give rise to cell-type-specific RNA-protein complexes that coordinate cell-type specific gene expression programs.

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

Acknowledgements

As I close this chapter of my life, there are so many people that I am indebted to for making my graduate career so wonderful. While I cannot do justice to all the people who have helped me, there are a few people who I especially want to thank.

First and foremost, I am extremely grateful and indebted to my advisor Eric Lander. Our scientific conversations were always extraordinarily thought provoking and exciting and he always challenged me to think big. Eric taught me so many things and molded me into the scientist that I am today. I am incredibly grateful for the trust that Eric placed in me, it was incredibly motivating to know that he trusted my judgment. Beyond science, Eric was a truly phenomenal mentor and the guidance he would provide was always impressive. There was never a problem too small for Eric to make time to discuss and help. Perhaps the greatest lesson he taught me was, by example, the value of mentorship.

I have been incredibly lucky to have the support and mentorship of many across the Broad community. I especially want to thank John Rinn who has been an extraordinary friend, mentor, and collaborator. I first met John while he was still at Stanford; we began working together on his first day in Boston. I am also incredibly lucky to count John as one of my closest collaborators and friends. I am indebted to John for his mentorship and support both in and outside of the lab.

I am also incredibly indebted to Aviv Regev who in many ways was my second advisor. I am so grateful for the generosity of her time and her extraordinary advice. Aviv always treated me as a colleague and despite my status always included me on important projects where she thought I might benefit or could contribute. I am extremely grateful to Aviv for her trust and support.

I am also incredibly grateful to Jill Mesirov who had brought me into the Broad community. Jill was always willing to talk through my ideas and provide advice and help. I am also grateful to many others at the Broad who have been wonderful mentors and colleagues including Dave Root, Brad Bernstein, and Alex Meissner. They all taught me so much and were always willing to work together or help me out.

I also want to thank my thesis committee members, Phil Sharp and Dave Bartel. Both Phil and Dave were always willing to provide scientific guidance. My committee meetings were always the highlight of my year. These meetings will be what I miss most about graduate school.

I am grateful to James Darnell who I had the great privilege of meeting as a 2nd year graduate student. Our scientific conversations were always thought provoking and incredibly rewarding. I am also extremely thankful to him for his support and encouragement throughout my graduate career.

I also want to thank the many people who I have had the privilege to work with every day at the Broad Institute. I am extremely thankful for having had the opportunity to work with Jesse Engreitz, Patrick McDonel, Pam Russell, Alex Shishkin, Klara Sirokman, Elena Stamenova, and

Christine Surka. For making work fun on a daily basis and for teaching me so much about molecular biology, biochemistry, computational biology, and stem cell biology I am grateful.

I am also extremely grateful for my close collaborators who were always willing to teach me new things and discuss new ideas. I am incredibly grateful to Julie Donaghey who was an extraordinary colleague and patient teacher without which I could not have done much of what is in this thesis. I am also thankful to Tarjei Mikkelsen who helped me get my bearings in the genomics world and who always made time to discuss ideas and answer questions no matter how naïve they may have been. I am thankful to Bryce Carey for always being willing to teach me new things in the lab or just meet and discuss ideas or techniques. I also owe a great debt of gratitude to Jim Robinson who was always willing to develop new visualization methods and to share code with me, both of which were absolutely essential to so much of what I did.

I am extraordinarily grateful for the support and friendship of both Manuel Garber and Ido Amit. Both Manuel and Ido were not just wonderful colleagues, teachers, and collaborators but are amongst my closest friends. Manuel taught me so much of what I know about computational biology and Ido taught me so much about experimental biology. I cannot even count how many ideas came about just through our various discussions in pairs or as a group. Their support both in and out of the lab was so instrumental to me scientifically and personally.

I want to thank my entire family for their unconditional love and support. I am extremely grateful to my father, George Guttman, who always trusted my judgment and always encouraged me to do what I love. I am eternally grateful and forever indebted to my grandfather, Alexander Guttman, who was my biggest supporter throughout my life. While not a scientist, the lessons that he taught me have always played a central role in shaping my thinking and my life. In many ways, he was the single greatest teacher and mentor that I have had.

Finally, I am incredibly lucky to have the love and support of Cheryl Hendricks. Cheryl has always been there to support me. Despite the fact that sometimes my work seemed to take priority, she was always understanding and encouraging. Her love and support is what keeps me going.

TABLE OF CONTENTS

Title page 1

Abstract 3

Acknowledgements 5

Introduction 9

Chapter 1: Chromatin signature reveals more than a thousand well conserved, large non-coding RNAs in mammals 45

Chapter 2: *Ab initio* reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincRNAs 85

Chapter 3: Expression patterns implicate lincRNAs in diverse biological processes 137

Chapter 4: Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression 177

Chapter 5: lincRNAs function in the molecular circuitry controlling pluripotency and differentiation in embryonic stem cells 203

Future Directions 259

Introduction

Parts of this chapter was first published as:

Guttman M and Rinn JL. (2012). Modular regulatory principles of large non-coding RNAs. ***Nature*** 482(7385):339-46

Overview

More than half a century after being placed as the intermediate in the central dogma it is now clear that RNA is more than a mere messenger and can perform diverse functional roles. Shortly after the discovery of messenger RNA (mRNA), a large class of heteronuclear RNAs (hnRNA)[1] was described much of which did not consist of mRNA nor were they associated with polyribosomes[2]. After years of sifting through these hnRNAs, the first RNA subfamilies emerged including the small nuclear RNAs involved in splicing[3], small nucleolar RNAs involved in ribosome biogenesis[4], and the 7SL RNA of the signal recognition particle involved in protein transport[5] adding to the ribosomal RNAs and transfer RNAs involved in translation[6-8]. More than a decade later, genetic studies identified a few tiny RNAs that act to silence mRNAs[9-11] leading to the discovery of a class of microRNAs[12-14] and other small RNA regulators[15-18].

The world of RNA genes became even more complex with the discovery of RNAs that resembled mRNA in length and splicing structure yet did not code for proteins. The first example, termed H19, was identified as an RNA induced during liver development in the mouse[19]. The mouse H19 transcript contained no large open reading frames (ORFs), but only small sporadic ORFs that were not evolutionarily conserved, could not template translation *in vivo,* and did not produce an identifiable protein product[20]. Shortly after, another ncRNA, termed XIST, was found to be expressed exclusively from the inactive X chromosome[21] and later shown to be required for X inactivation in mammals[22]. Over the next two decades a few additional large ncRNA genes were discovered including Air[23], Tug1[24], NRON[25], and HOTAIR[26].

Following the sequencing of the human genome, the next major hurdle was to define the genes it encoded. Studies probing global transcriptional activity yielded a surprising result: the mammalian genome is pervasively transcribed with nearly the entire genome being transcribed into RNA under some circumstance[27-30]. As the numbers of non-coding transcripts increased, so too did concerns that many of the transcripts were simply 'transcriptional noise' without a biological function[31,32]. The reasons for concern included the observation that many of the transcripts are expressed at extremely low levels and exhibit no evolutionary conservation. It was unclear whether the few functional examples represented quirky exceptions, or exemplified a major class of functional large ncRNAs. Distinguishing between these possibilities required additional biological information.

In this thesis, I present our contributions to the identification and functional characterization of a class of functional large intergenic non-coding RNA (lincRNA) genes in the mammalian genome. First, I describe the identification of thousands of well conserved lincRNAs in the human and mouse genomes by exploiting a chromatin signature of actively transcribed genes. Second, I describe the development of a statistical method for *de novo* reconstruction of a mammalian transcriptome from RNA-Seq data and its application to define the transcript structures of lincRNAs. Third, I describe systematic loss-of-function studies, which demonstrated that lincRNAs play a clear functional role in the cell and that many lincRNAs play an essential role in maintaining the pluripotent cell and repressing differentiation programs. Finally, I describe our work demonstrating that many lincRNAs act through their physical interaction with multiple chromatin protein complexes to regulate gene expression programs.

To provide context, I begin by briefly reviewing the key classes of regulatory RNAs that have emerged in the past 50 years. Next, I describe the identification of the first large ncRNAs and describe the few well characterized examples. I highlight the key regulatory principles and themes learned from these examples and explore parallel studies linking chromatin regulation and RNA. Finally, I review genome-wide efforts to annotate the mammalian transcriptome and describe the plethora of transcripts both small and large that have been identified from these efforts. I summarize the specific contributions of this thesis toward identifying and functionally characterizing large ncRNAs in mammals.

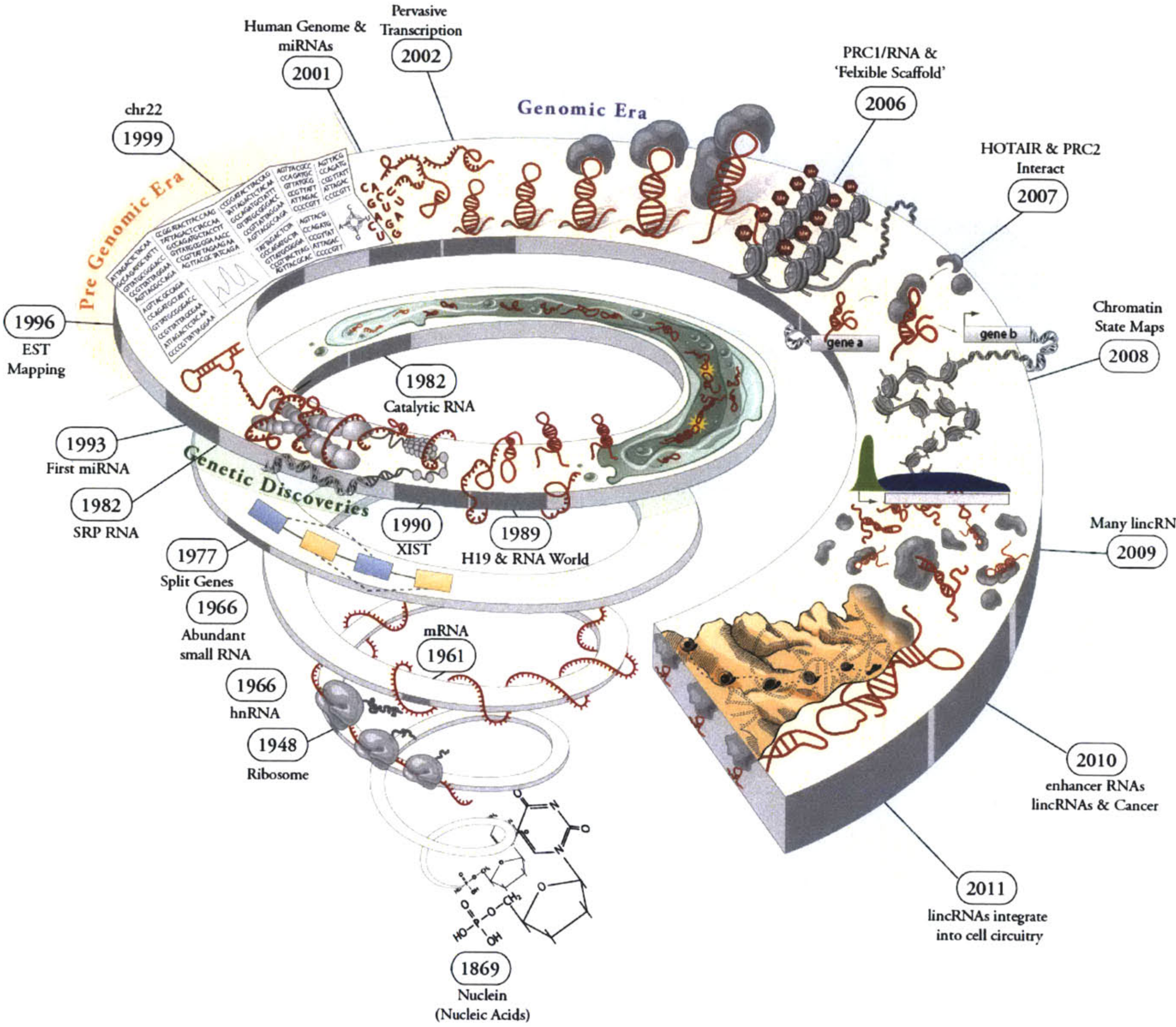


Figure 1: A timeline of major discoveries in RNA biology.

Credit: Sigrid Knemeyer, Mitchell Guttman, and John Rinn

RNA from messenger to regulator

Little more than a half a century ago, the importance of RNA was easily missed[33]. It was clear for several decades that proteins were functional actors[34] and it had recently been discovered that DNA was the genetic material[35,36]. Understanding how DNA encoded the genetic material and how this information was translated into protein products was a critical challenge. Pioneering genetic studies by Jacob and Monod suggested the importance of a "messenger" in bridging the genetic (DNA) to the functional (protein)[37]. Shortly after this suggestion, the elusive messenger was identified as an unstable RNA species that was quickly turned over in cells[38]. The identification of this messenger RNA (mRNA) led to the notion of a central dogma in the flow of genetic information; DNA gives rise to RNA which gives rise to protein[39].

With the discovery of mRNA, it was important to define precisely how the cell translated this information into proteins. Central to this entire process was Crick's "adaptor hypothesis", the notion that another molecule was responsible for bridging mRNA and amino acids[40]. A key component of this adaptor proposal was the thought that RNA would be uniquely able to function as this adaptor because of its ability to base pair with the mRNA[40]. With the discovery of the tRNAs[8] that match the sequence on an mRNA with the corresponding amino acids[41,42] it became clear that RNA itself was in fact this adaptor molecule. Beyond the adaptor tRNAs, it was already clear that the site of translation itself, the ribosome, was composed of RNA[43]. The centrality of RNA in the process of transcription and translation was now clear.

It did not take long before the notion that DNA→RNA→Protein was challenged. The first challenge was the discovery of viruses that encoded their genetic material as RNA, rather than DNA[44]. Soon after, an enzyme was identified that could convert the RNA genome into DNA using an encoded reverse transcriptase enzyme[45,46]. Not only did this establish that RNA could act as the genetic material but that the path from DNA to RNA was a reversible course.

The second challenge was the identification of RNA enzymes, termed ribozymes that demonstrated that RNA alone, in the complete absence of protein, contains catalytic activity. This discovery initially identified RNaseP in bacteria[47] and RNA self-splicing in *Tetrahymena*[48,49] demonstrated the role of RNA as a functional molecule. More recently it has become clear that the ribosome itself is made up of catalytic RNA where the RNA itself is critical for the peptidyl transferase reaction rather than the proteins within the ribosome[50-52]. Over the ensuing decades, the numbers of known catalytic RNAs have expanded and their roles broadened to include roles in diverse processes including splicing[49], translation[51], metabolite sensing[53], and gene regulation[54]. Taken together, these observations supported the notion of an ancient "RNA world", where RNA -- rather than DNA or protein -- was the original molecule given its ability to both encode genetic information and catalyze reactions[55-57].

Discovery of abundant small RNA species

The importance of RNA was becoming clearer with the continued discovery of an ever expanding collection of functional RNA molecules. Shortly after the discovery of mRNA, a large class of heteronuclear RNAs (hnRNA)[1] was described. While much of the hnRNAs would soon be explained by the discovery of splicing[58], there were still many hnRNAs that did not associate with mRNA[59]. These non-mRNA associated hnRNAs contained many of the same features as mRNA, including the 5' cap structure and in some cases a 3' poly-A tail[2,59]. Yet, most of these transcripts did not enter polyribosomes[2]. After years of sifting through these hnRNAs, the first RNA subfamilies emerged including the small nuclear RNAs involved in splicing[3], small nucleolar RNAs involved in ribosome biogenesis[4], and the 7SL RNA of the signal recognition particle (SRP) involved in protein transport[5], adding to the ribosomal RNAs and transfer RNAs involved in translation[6,7].

The small *nucleolar* RNAs (snoRNAs) and small *nuclear* RNAs (snRNAs) were originally discovered based on their abundance in the nuclei of mammalian cells[3]. These RNAs ranging in size from 90-300nts were distributed both in the nucleus, the site of transcription and splicing, as well as the nucleolus, the site of ribosomal RNA transcription, processing, and assembly[4]; some were even primarily cytoplasmic[5]. Importantly, these different small RNA species were shown to be conserved across different eukaryotes from yeast to mammals[3,60].

The first class of small RNAs in this population was the snoRNAs, identified in the nucleolus. Because the snoRNAs were identified in the nucleolus, it was immediately suspected that they would play a role in processing of the ribosomal RNAs[3,60]. This was confirmed with the identification that several of the snoRNAs base pair with the ribosomal RNA, providing the first clue as to how they may act[4,61]. Several of the snoRNAs were subsequently shown to play a direct role in rRNA processing, as depletion of snoRNAs in mouse extracts resulted in impaired cleavage of the ribosomal RNA[60]. Beyond their role in rRNA processing, other snoRNAs are important for guiding various modifications of the rRNA[60].

The second class was the snRNAs. Unlike the snoRNAs, at the time of their discovery, the function of these RNAs were unclear[3,4]. The first clues to their possible function required the discovery of mRNA splicing[58,62]. Initial clues to the function of snRNAs came from the discovery of sequence complementary between the dominant U1 snRNA and the canonical 5' splice site [63-65]. Beyond the U1 snRNA, other snRNAs interact with different splice sites including the branchpoint site, 3' splice site, and non-canonical splice sites[65,66]. These snRNAs interact with protein cofactors to give rise to the spliceosome, which interacts with mRNA precursors and is responsible for the splicing reaction[65,67,68]. Interestingly, the snRNAs are transcribed in the nucleus and immediately exported to the cytoplasm where they are processed and assembled into snRNPs, which are then reimported into the nucleus to function in the splicesome[66]. While the snRNAs themselves are not sufficient to catalyze the splicing reaction, it appears likely that the RNA may play some role, in conjunction with other protein cofactors, in the catalysis of the splicing reaction[68].

Initial discovery of the snRNA and snoRNAs also revealed another extremely abundant RNA, the 7SL RNA[3,4], which was later shown to be a component of the SRP[5]. The SRP is a large RNA-protein complex consisting of the 7SL RNA and as many as 6 different proteins[69]. The SRP is required for transport of nascent proteins from the ribosome to the endoplasmic reticulum, based on recognition of signal sequence in the nascent peptide[5]. The 7SL RNA itself is essential for the function of the SRP complex[5] and is important for stabilizing the structure of the SRP proteins[69] and kinetically enhances the interaction between the SRP and SRP-receptor[69,70]. The 7SL RNA promotes catalysis by acting as a transient tether between the SRP and associated receptor, thereby stabilizing the transition state and enhancing the interaction between the two complexes[70,71].

At the time, characterizing each RNA within a complex population was quite challenging. Accordingly, much of the RNA species within the hnRNA mix remained uncharacterized. A commonality across these few identified RNA species was their abundant expression and ubiquity across all cell types[3,4,60].

A class of tiny RNAs that regulate mRNA

Beyond the role of ncRNAs in constitutive cellular processes, we now know that there are many ncRNAs that are expressed in more cell-type specific ways and play a role in cell-type specific regulation. One of the first such classes identified were the microRNA genes[12-14]. The first miRNA was identified in a genetic screen in *C. elegans* for mutations that affect timing of development[10]. This screen turned up a mutation in the *lin-4* gene; subsequent characterization showed that this gene lacked an open reading frame (ORF) and instead encoded a 21-nucleotide RNA product processed from a larger 61-nucleotide precursor[10]. This tiny RNA product had strong sequence complementarity to the 3' untranslated regions (UTRs) in the *lin-14* protein-coding gene, which genetic experiments had previously shown it to repress[9,10]. A few years later, another example of a tiny RNA, termed *let-7*, was identified in *C. elegans* that seemed to work in a comparable way[11]. Homology mapping on the *let-7* RNA showed that it was conserved in all bilaterally symmetric animals from worms to humans[72]. Genomic strategies used to sequence tiny RNAs of similar size and sharing the chemical modifications of the previous RNAs revealed that these were not the exception and in fact there were hundreds of tiny RNAs, termed microRNAs[12-14,73].

The last decade has seen an explosion in our understanding of miRNAs, which have now been demonstrated to play critical roles in diverse biological processes and diseases all through a shared mechanism[73,74]. The mature miRNA associates with the Argonaute proteins to form the

RNA-induced silencing complex (RISC). The miRNA within RISC acts to provide target specificity through the six nucleotide miRNA seed region, which base pairs with an mRNA, usually in the 3' untranslated regions[74]. The complex primarily leads to mRNA destabalization[75], likely through mRNA de-adenylation[76-78].

Beyond miRNAs, there are several other classes of small RNA species that similarly associate with argonaute proteins and act as 'guides' to target genes. Another class of small RNAs, highly expressed in the testis, are the piwi-interacting RNA (piRNAs), so named because they interact with the Piwi-family of Argonaute proteins[15,79]. The piRNAs localize these complexes to active transposons in the germ line and silence them[80] through changes in chromatin and DNA[81].

A handful of functional large ncRNAs in mammals

The world of RNA genes became even more complex with the discovery of RNAs that resembled mRNA in length and splicing structure yet did not code for proteins. The first example, termed H19, was identified as an RNA induced during liver development in the mouse[19]. The mouse H19 transcript contained no large open reading frames (ORFs), but only small sporadic ORFs that were not evolutionarily conserved, could not template translation *in vivo,* and did not produce an identifiable protein product[20]. H19 opened up the possibility that many more messenger-like RNAs may in fact be non-coding RNAs. Shortly after, another ncRNA, termed XIST, was found to be expressed exclusively from the inactive X chromosome[21] and later shown to be required for X inactivation in mammals[22]. Over the next two decades a few additional large ncRNA genes were discovered including Air[23], Tug1[24], NRON[25], and HOTAIR[26].

Individual ncRNAs involved in specific processes have been functionally characterized (reviewed in reference 81). For example, XIST is critical for random inactivation of the X-chromosome[22], Air is critical for imprinting control at the Igf2r locus[23], HOTAIR affects expression of the HoxD gene cluster[26] as well as multiple genes throughout the genome[82-84], HOTTIP affects expression of the HoxA gene cluster[85], lincRNA-RoR affects reprogramming efficiency[86], NRON affects NFAT transcription factor activity[25], and Tug1 affects retina development through regulation of the cell cycle[24]. While there are now many examples of large ncRNAs that are required for proper regulation of gene expression, this is just one function for large ncRNAs, which play critical roles in telomere replication[87] and translation[88].

In parallel to the discovery of the first large ncRNAs, the relationship between RNA and chromatin was beginning to emerge[89,90]. A growing body of literature from yeast to mammals suggested that RNA plays an important role in chromatin-state formation[90]. In *Schizosaccharomyces pombe*, a process known as RNA Induced Transcriptional Silencing (RITS) has been shown to play an important role in heterochromatin formation over centromeric repeats[91]. Similarly, short RNAs have been shown to play an important role in the establishment of heterochromatic silencing in plants[90]. In *C. elegans*, genetic screens have identified Polycomb homologs to be required for proper gene silencing in an RNA-dependent manner[90]. In mammals, there is evidence that RNA plays a key role in shaping mammalian epigenetic landscapes. For example, depletion of single-stranded RNA (ssRNA) in mouse fibroblasts inhibits global heterochromatin formation by delocalizing the HP1 complex from genomic sites[92]. Similarly, ssRNA but not ssDNA is required for the maintenance of the histone modifications H3K27me3 and H3K9me3 through proper localization of proteins in the polycomb family[93]. Taken together, these studies suggested a role for RNA in maintenance and localization of chromatin regulatory complexes to genomic DNA targets. Yet, what these RNAs were remained uncharacterized.

The discovery of the XIST ncRNA made clear that an RNA can play a direct role in silencing large genomic regions. To determine how the XIST ncRNA works, several groups looked at the sequence of events follows induction of the XIST ncRNA as it nucleates the X-chromosome[94]. This led to the observation that alongside XIST accumulation on the chromosome, several histone modifications such as the H3K27me3 modification and the polycomb complex localize to the chromosome, suggesting that the XIST ncRNA may physically interact with the polycomb complex[95]. Recently, direct RNA-protein interaction mapping has demonstrated that the XIST ncRNA can in fact physically interact with the polycomb complex[96], leading to the condensation of chromatin and transcriptional repression of an entire X chromosome[94] and deletion of the domain of interaction allows proper localization to DNA but prevents X-chromosome silencing[97].

Similar to XIST, several other large ncRNAs have been identified that physically associate with chromatin regulatory complexes and ‘guide’ the associated complexes to specific genomic DNA region. These include the antisense transcript AIR[23,98,99] which is associated with the chromatin-modifying complex G9a, an H3K9me2 methyltransferase[100]; and the Kcnq1ot1 transcript[99] that binds both G9a and PRC2[101]. More recently, HOTAIR has been shown to contain distinct protein-interaction domains that can associate with PRC2[26] and the CoREST/LSD1 complex[83], which together enable its proper function[83].

In addition to their role in chromatin regulation, large ncRNAs can also modulate the regulatory activity of other protein complexes. As an example, a ncRNA upstream of cyclin D1 can bind to the TLS RNA binding protein, changing it from an inactive to active state[102]. Similarly, the NRON ncRNA can bind to the NFAT transcription factor rendering it inactive by preventing nuclear accumulation[25]. ncRNAs can also act as molecular 'decoys' by preventing proper regulation through competitive binding. For example, the GAS5 ncRNA binds to the glucocorticoid receptor and prevents the receptor from binding to its proper regulatory elements[103] and the PANDA ncRNA can prevent NF-Y localization leading to apoptosis[104]. Similarly, several recent studies report that large ncRNAs can act as 'decoys' to other RNA species, such as miRNAs, to control their levels[105,106].

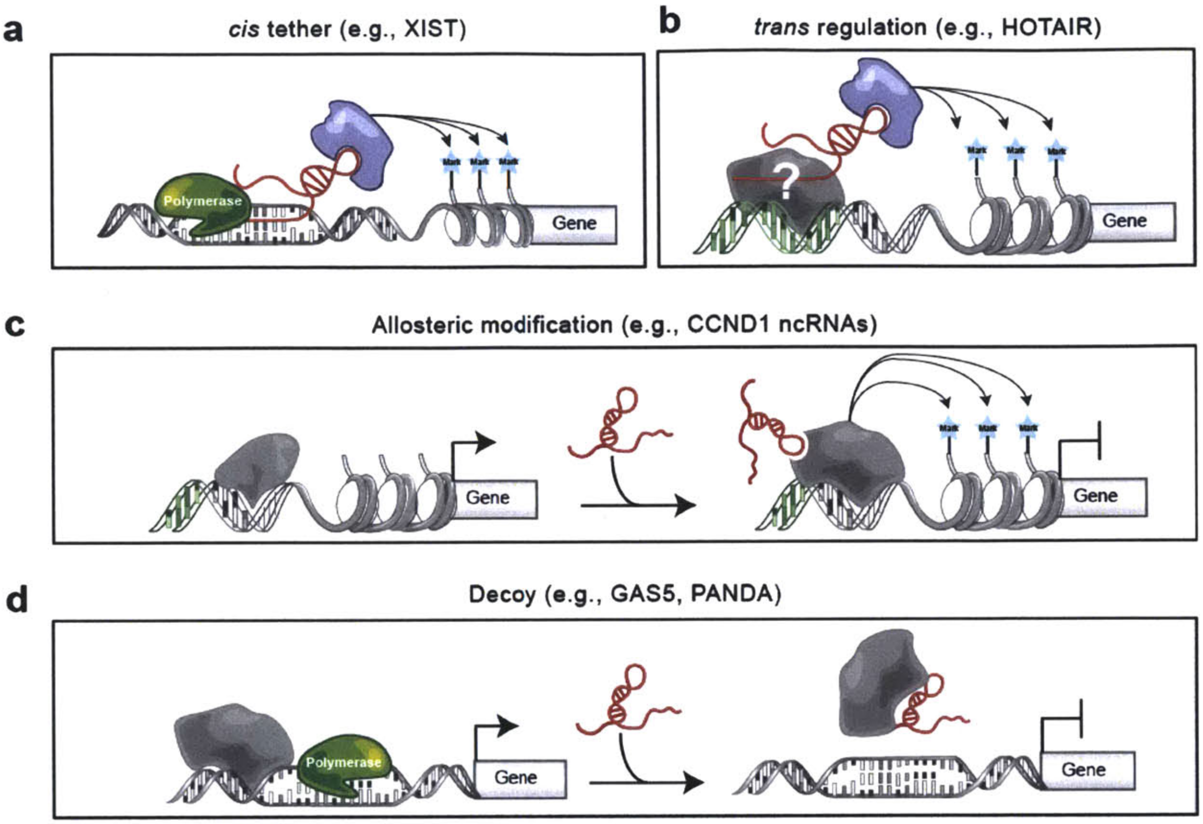


Figure 2: Examples of large ncRNA regulatory mechanisms. (a) A model of ncRNAs that act in *cis* while remaining tethered to their site of transcription. In this model, RNA polymerase (green) transcribes an RNA which can associate with regulatory proteins (blue) to affect it neighboring regions as proposed for XIST[21,107]. (b) One model for ncRNA regulation in *trans.* In this model a ncRNA can associate with DNA binding proteins (gray) and regulatory proteins to localize and affect expression of targets, as proposed for HOTAIR[83]. (c) A model for ncRNAs that bind regulatory proteins and change their activity in this case leading to change in modification state and expression of the target gene, as proposed for the CCND1 ncRNAs that interact with the TLS protein[102].(d) A model for ncRNAs that act as 'decoys'. In this model, ncRNAs bind protein complexes and prevent them from binding to their proper regulatory targets as proposed for GAS5[103].

Extensive transcription in the mammalian genome

While the human genome has now been sequenced[108] the functional elements encoded within it remain largely uncharted. More than 5% of the genome has clearly been under purifying selection over the past 100 million years and thus must be functional[109], but only ~1.2% encodes protein [108,109,110]. Within the remainder, an important and growing category consists of genes encoding functional RNA molecules. These include: classical examples such as ribosomal, transfer and splicing-associated RNAs; a recently discovered cast of tiny RNAs, including microRNAs and other families[73,111].

When we began this work, there were about a dozen well-characterized examples in mammals, with transcript size ranging from 2.3 to 17.2 kb [20, 21]. They each play distinctive biological roles through diverse molecular mechanisms, including functioning in X-chromosome inactivation (XIST, TSIX) [21,112], imprinting (H19, AIR)[20,98], *trans*-acting gene regulation (HOTAIR)[26] and regulation of nuclear import (NRON)[25]. It was unclear whether these few cases represented quirky exceptions, or whether they heralded an entire world of functional large intergenic non-coding RNAs.

Over the past decade, there have been extensive efforts to characterize the mammalian transcriptome. Following the sequencing of the human genome, several groups created tiling microarray spanning human chromosome 22, which probed the transcriptional activity across many tissues[113,114]. This led to the observation of widespread transcriptional activity across the chromosome with an estimate of ~10:1 non-coding to protein-coding transcription. This result was soon generalized to all human chromosomes with the creation of genome-wide tiling arrays[29,115]. In parallel, high-throughput sequencing of mouse cDNAs reported the identification of tens of thousands of transcripts of which only a small fraction were accounted for by protein-coding genes[27,116]. These studies, while limited to specific tissues and cell types, made clear that the mammalian genome is pervasively transcribed, giving rise to many thousands of non-coding transcripts. This culminated in the publication of the pilot Encyclopedia of DNA Elements (ENCODE) project which suggested that ~100% of the human genome is transcribed into RNA in some cell type[28].

As the numbers of discovered non-coding transcripts increased, so did the uncertainty about their functional role. It was unclear whether the majority of these transcripts have biological function or merely reflect "transcriptional noise" [31,32 31,32 31,32 31,32 31,32 31,32 31,32], spurious transcription due to random initiation by RNA polymerase[31,32], or incidental by-products of transcription from enhancer regions[117,118]. In support of these concerns, many of the reported transcripts occur at extremely low levels – orders of magnitude below the levels for protein-coding genes[119]. While some of the transcripts show tissue-specific expression, such expression patterns may simply reflect transcription from randomly-occurring RNA polymerase binding sites in open chromatin in a given cell type[120] , or possibly transcription whose purpose is to modify local chromatin structure[121].

Importantly, not all non-coding transcripts act as functional RNA molecules. Several examples of intergenic transcription have been identified where the act of transcription alone changes the chromatin and transcription factor binding landscape allowing activation and repression of neighboring genes[121-124]. An analogous mechanism has been suggested for the H19 ncRNA which through its transcription may affect allelic imprinting by competing for enhancer binding[125,126].

The only way to prove that an RNA is functional is to show that its disruption has a deleterious consequence. In principle, this can be done in two ways: (i) by demonstrating clear evolutionary conservation (that is, purifying selection) or (ii) by perturbing the gene in a laboratory experiment. The first approach is vastly more efficient, because it involves computational analysis that can be applied to an entire collection of RNAs, and more general because it does not require prior knowledge of the biological context in which the gene functions.

Attempts to prove the functionality of the observed non-coding RNAs through evolutionary conservation were discouraging. Initial studies found that non-coding transcripts show no better conservation than random intergenic sequence[127]. A more recent study reported that the conservation levels are slightly better than random – but still extremely low[128]. Some authors proposed that most non-coding RNAs are functional, but show little conservation because they are evolving rapidly[120]. Few, non-coding transcripts were identified to have a phenotypic effect in a large-scale screen performed with these sets[25]. A reasonable hypothesis is that amongst this pervasive transcription there are some *bona fide* functional non-coding RNAs hidden in a much larger background of transcriptional noise.

While some of these transcripts may indeed represent 'transcriptional noise'[31], within the remainder, we now know there are many distinct subclasses including processed small RNAs[30,129,130], promoter associated RNAs[30,131], transcripts from enhancer regions[117,118], and functional large ncRNAs[26,132] each of which vary in their expression and conservation properties[128,133]. Distinguishing between these classes of RNA transcripts required additional biological information including coding potential of the RNA and chromatin modifications of the corresponding genomic region (**Figure 3**).

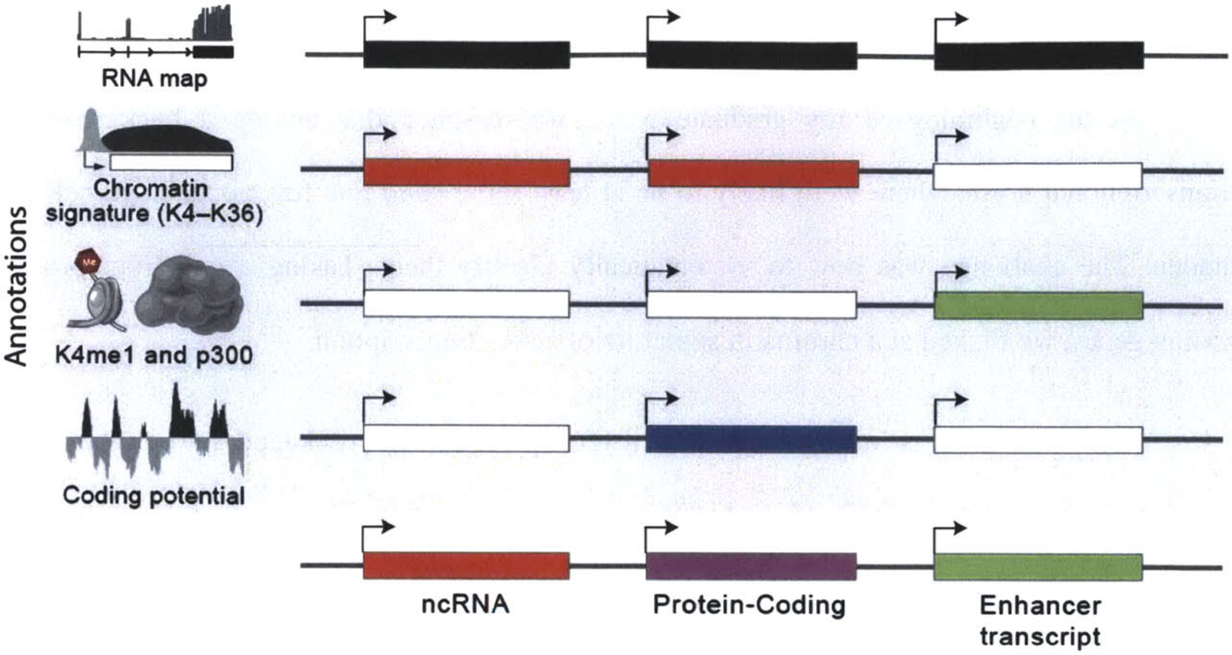


Figure 3: Genomic layers define subclasses of non-coding transcription. (a) Genomic regions are color coded by the presence of different genomic annotations. RNA transcription of a locus (black), K4-K36 chromatin signature (red), K4me1 and P300 modifications (green), and protein coding potential (blue). Overlaying this information reveals distinct transcripts including ncRNAs (red), protein-coding genes (purple), and transcripts from enhancer regions (green).

Contributions of this thesis

At the beginning of my graduate work, we reasoned that among a background of 'transcriptional noise', there were likely to be at least some *bona fide* functional large ncRNAs hidden. The challenge was how to systematically identify them. Taking a cue from protein-coding genes, we looked at a chromatin signature of active transcription.

Genomic DNA is wrapped around histone proteins and packaged into higher order structures termed chromatin[134]. These histones can be modified in different ways indicating the functional state of the underlying DNA region. Recent advances in sequencing technologies have enabled a comprehensive characterization of the chromatin modification landscape of mammalian genomes[135-138]. These studies revealed combinations of histone modifications, termed chromatin signatures, that correspond to various gene properties, including a signature of active transcription[135,138]. This signature consists of a short stretch of trimethylation of histone protein H3 at the lysine in position 4 (H3K4me3) corresponding to promoter regions followed by a longer stretch of trimethylation of histone protein H3 at the lysine in position 36 (H3K36me3) covering the entire transcribed region[135,138].

Chromatin maps revealed that, like protein-coding genes, many ncRNA genes also contain a 'K4-K36' signature[138]. We reasoned that by identifying K4-K36 domains that lay outside known protein-coding gene loci, we would be able to systematically discover functional large ncRNAs. To do this, we developed a computational algorithm that identifies K4-K36 domains from genome-wide chromatin datasets and excluded those that overlapped any annotated gene. This analysis yielded a set of ~1600 and ~2,500 unannotated intergenic K4-K36 domains in the mouse and human genomes. Using tiling microarrays, we demonstrated that the vast majority of the intergenic K4-K36 domains produced multi-exonic RNAs with many canonical features of RNA Polymerase II transcription. These transcripts had little capacity to encode a functional protein of any significant size. We termed the RNAs expressed from these 'K4-K36'domains large intergenic ncRNAs (lincRNAs) because identification by this chromatin signature required that they be contained within intergenic regions. Importantly, these lincRNAs demonstrated clear evolutionary conservation across mammalian genomes, providing strong evidence that the lincRNAs are biologically functional. This work demonstrated that the mammalian genome encodes thousands of functional lincRNAs.

Transcriptome reconstruction

Having identified a class of conserved lincRNAs, we next sought to determine the function of these lincRNAs. A critical pre-requisite for comprehensive experimental studies of lincRNA function is defining their precise sequences. While hybridization to tiling microarrays provided initial insights, it did not allow the precise identification of lincRNA sequences. Advances in massively-parallel cDNA sequencing, termed RNA-Seq, were allowing for the sequencing of cDNA at an unprecedented scale providing an unbiased way to collect data from a transcriptome, including both coding and non-coding genes. Yet, at the time there were only a few RNA-Seq studies which were limited to studying the expression and refining the splicing patterns of known genes. There were no studies to determine novel genes from RNA-Seq data.

Discovering lincRNA gene structures required reconstructing a mammalian transcriptome from scratch, a significant computational challenge as read lengths are significantly shorter than the size of the original RNA. To address this challenge, we developed a statistical method, called

Scripture, which was the first method to accurately reconstruct a mammalian transcriptome without prior gene models. We performed RNA-Seq on mouse ES cells, mouse lung fibroblasts, and mouse neural progenitor cells. Applying Scripture, we successfully recovered the gene structure of virtually all expressed protein-coding genes demonstrating the accuracy of the method. Importantly, Scripture identified the full-length gene structure of the vast majority of expressed lincRNA genes. This allowed us to pinpoint specific regions within each lincRNA that are under purifying selection demonstrating clear 'patches' of strong evolutionary constraint within the lincRNAs.

'Guilt-by-association' associates lincRNAs with diverse biological processes

Despite the identification of thousands of large ncRNAs it remained to be determined how these RNAs function. Determining the functional role of ncRNAs requires direct perturbation experiments such as loss of function and gain of function yet without a clear hypothesis of what phenotype to look for proved difficult in characterizing the functions of most ncRNAs[25]. To more globally classify putative functional roles of lincRNAs we developed a 'guilt-by-association' method to systematically associate functions based on correlation of gene expression[139]. This method associates ncRNAs with biological processes based on common expression patterns across cell types and tissues, and identifies groups of ncRNAs associated with specific cellular processes. Utilizing this approach allowed us to classify hundreds of ncRNAs across diverse biological processes such as stem cell pluripotency, immune response, neural processes, and cell cycle regulation[139].

While such correlations do not prove that ncRNAs function in these processes, they provide hypotheses for targeted loss-of-function experiments. To test these predictions, we performed targeted perturbations to determine the role of specific ncRNAs in the associated classes. As an example, we predicted 39 ncRNAs involved in the p53-mediated DNA damage response and showed that one of these candidates, termed lincRNA-p21, is a direct target of p53[140]. Perturbation of this ncRNA affected the apoptosis response upon exposure to DNA damage. Another lincRNA, lincEnc1[139], was predicted to have a roll in cell-cycle regulation in ESCs and shown in a distinct study to affect proliferation in ESCs[141]. Overall, our 'guilt-by-association' approach implicated lincRNAs in diverse biological processes[86,104,139,140,142].

lincRNAs regulate cell states in an embryonic stem cell

Using the lincRNA sequences, we set out to determine the functional role of lincRNAs using loss-of-function experiments. Unlike correlation analysis, these perturbation-based experiments provide evidence for the functional role of a ncRNA. We focused on mouse ES cells because the signalling, transcriptional, and chromatin regulatory networks controlling pluripotency have been well characterized providing an ideal system to determine how lincRNAs integrate into the molecular circuitry of the cell. We designed, cloned, and validated shRNAs targeting all lincRNAs expressed in mouse ES cells. To determine whether lincRNAs play an important function in the cell, we studied the effects of knocking down each lincRNA on global transcription. Upon knockdown, virtually all of the lincRNAs showed a significant impact on gene expression demonstrating that the lincRNAs are functionally important in the cell.

Next, we sought to determine if lincRNAs play a role in regulating the ESC state. We studied the effects of lincRNA knockdown on the expression of Nanog, a key transcription factor that is required to establish and uniquely marks the pluripotent state. We identified 26 lincRNAs that had major effects on endogenous Nanog levels along with other markers of the pluripotent state. To determine if lincRNAs play a role in repressing differentiation programs, we compared the overall gene expression patterns resulting from knockdown of the lincRNAs to gene expression patterns resulting from induced differentiation of ESCs. We identified 30 lincRNAs whose knockdown produced expression patterns similar to differentiation into specific lineages. This work demonstrated that many lincRNAs play important roles in regulating the ESC state.

Many lincRNAs interact with multiple chromatin regulatory proteins

Having demonstrated the functional importance of lincRNAs in the cell, we wanted to determine how lincRNAs affect gene expression. Motivated by the XIST and HOTAIR ncRNAs, which interact with the polycomb complex, we tested whether lincRNAs more generally associate with the polycomb complex. We found that ~20% of expressed human lincRNAs and ~10% of the ESC lincRNAs physically associate with the polycomb complex. Next, we systematically analyzed chromatin-modifying proteins that have been shown to play critical roles in ESCs. We screened antibodies against 28 chromatin complexes and identified 11 additional chromatin complexes that are strongly and reproducibly associated with the ESC lincRNAs. These chromatin complexes are involved in 'reading',. 'writing', and 'erasing' histone modifications. Altogether, we found that ~30% of the ESC lincRNAs are associated with at least one of these chromatin complexes. Interestingly, many of the lincRNAs physically interacted with multiple chromatin complexes.

Next, we sought to determine if the identified interactions are important for lincRNA mediated regulation. To do this, we examined the effects on gene expression resulting from knockdown of individual lincRNAs that are physically associated with particular chromatin complexes and from knockdown of the associated complex itself. For most of these lincRNA-protein interactions, we identified a significant overlap in affected gene expression programs. Together, these data demonstrate that many of the ESC lincRNAs physically associate with multiple different chromatin regulatory proteins to affect gene expression programs.

Outlook

Our data suggests a model whereby a distinct set of lincRNAs is transcribed in a cell type and interacts with ubiquitous regulatory protein complexes to give rise to cell-type-specific RNA-protein complexes that coordinate cell-type specific gene expression programs. Because many of the lincRNAs studied here interact with multiple different protein complexes, one hypothesis is that they act as cell-type specific 'flexible scaffolds'[87] to bring together protein complexes into larger functional units. This model has been previously demonstrated for the yeast telomerase RNA[87] and suggested for the XIST[97] and HOTAIR[83] lincRNAs. The hypothesis that lincRNAs serve as flexible scaffolds could explain the uneven patterns of evolutionary conservation seen across the length of lincRNA genes[133]: the more highly conserved patches could correspond to regions of interaction with protein complexes.

While a model of lincRNAs acting as 'flexible scaffolds' is attractive, it is far from proven. Testing the hypothesis for lincRNAs will require systematic studies, including defining all protein-complexes with which lincRNAs interact, determining where these protein interactions assemble on RNA, and ascertaining whether they bind simultaneously or alternatively. Moreover, understanding how lincRNA-protein interactions give rise to specific patterns of gene expression will require determination of the functional contribution of each interaction and possible localization of the complex to its genomic targets.

References

1 Warner, J. R., Soeiro, R., Birnboim, H. C., Girard, M. & Darnell, J. E. Rapidly labeled HeLa cell nuclear RNA. I. Identification by zone sedimentation of a heterogeneous fraction separate from ribosomal precursor RNA. *J Mol Biol* **19**, 349-361 (1966).

2 Salditt-Georgieff, M., Harpold, M. M., Wilson, M. C. & Darnell, J. E., Jr. Large heterogeneous nuclear ribonucleic acid has three times as many 5' caps as polyadenylic acid segments, and most caps do not enter polyribosomes. *Mol Cell Biol* **1**, 179-187 (1981).

3 Weinberg, R. A. & Penman, S. Small molecular weight monodisperse nuclear RNA. *J Mol Biol* **38**, 289-304 (1968).

4 Zieve, G. & Penman, S. Small RNA species of the HeLa cell: metabolism and subcellular localization. *Cell* **8**, 19-31 (1976).

5 Walter, P. & Blobel, G. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* **299**, 691-698 (1982).

6 Gesteland, R. F., Cech, T. & Atkins, J. F. *The RNA world : the nature of modern RNA suggests a prebiotic RNA world.* 3rd edn, (Cold Spring Harbor Laboratory Press, 2006).

7 Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* **2**, 919-929 (2001).

8 Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I. & Zamecnik, P. C. A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem* **231**, 241-257 (1958).

9 Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**, 855-862 (1993).

10 Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-854 (1993).

11 Reinhart, B. J. *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901-906 (2000).

12 Lee, R. C. & Ambros, V. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**, 862-864 (2001).

13 Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**, 858-862 (2001).

14 Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853-858 (2001).

15 Aravin, A. *et al.* A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**, 203-207 (2006).

16 Okamura, K. & Lai, E. C. Endogenous small interfering RNAs in animals. *Nat Rev Mol Cell Biol* **9**, 673-678 (2008).

17 Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539-543 (2008).

18 Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534-538 (2008).

19 Pachnis, V., Brannan, C. I. & Tilghman, S. M. The structure and expression of a novel gene activated in early mouse embryogenesis. *EMBO J* **7**, 673-681 (1988).

20 Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Mol Cell Biol* **10**, 28-36 (1990).

21 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**, 38-44 (1991).

22 Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379**, 131-137 (1996).

23 Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810-813 (2002).

24 Young, T. L., Matsuda, T. & Cepko, C. L. The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr Biol* **15**, 501-512 (2005).

25 Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**, 1570-1573 (2005).

26 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323 (2007).

27 Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563 (2005).

28 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

29 Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242-2246 (2004).

30 Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484-1488 (2007).

31 Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat Cell Biol* **10**, 1106-1113 (2008).

32 Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* **14**, 103-105 (2007).

33 Darnell, J. E. *RNA : life's indispensable molecule.* (Cold Spring Harbor Laboratory Press, 2011).

34 Sumner, J. B. The isolation and crystallization of the enzyme urease. *Journal of Biological Chemistry* (1926).

35 Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **79**, 137-158 (1944).

36 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

37 Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-356 (1961).

38 Brenner, S., Jacob, F. & Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**, 576-581 (1961).

39 Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

40 Crick, F. H. On protein synthesis. *Symp Soc Exp Biol* **12**, 138-163 (1958).

41 Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science* **147**, 1462-1465 (1965).

42 Holley, R. W., Everett, G. A., Madison, J. T. & Zamir, A. Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid. *J Biol Chem* **240**, 2122-2128 (1965).

43 Palade, G. E. A small particulate component of the cytoplasm. *J Biophys Biochem Cytol* **1**, 59-68 (1955).

44 Kates, J. R. & McAuslan, B. R. Messenger RNA synthesis by a "coated" viral genome. *Proc Natl Acad Sci U S A* **57**, 314-320 (1967).

45 Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **226**, 1209-1211 (1970).

46 Temin, H. M. & Mizutani, S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* **226**, 1211-1213 (1970).

47 Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849-857 (1983).

48 Kruger, K. *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147-157 (1982).

49 Zaug, A. J., Grabowski, P. J. & Cech, T. R. Autocatalytic cyclization of an excised intervening sequence RNA is a cleavage-ligation reaction. *Nature* **301**, 578-583 (1983).

50 Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* **289**, 905-920 (2000).

51 Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920-930 (2000).

52 Muth, G. W., Ortoleva-Donnelly, L. & Strobel, S. A. A single adenosine with a neutral pKa in the ribosomal peptidyl transferase center. *Science* **289**, 947-950 (2000).

53 Baker, J. L. *et al.* Widespread genetic switches and toxicity resistance proteins for fluoride. *Science* **335**, 233-235 (2012).

54 Mandal, M. & Breaker, R. R. Gene regulation by riboswitches. *Nat Rev Mol Cell Biol* **5**, 451-463 (2004).

55 Gilbert, W. The RNA World. *Nature* **319**, 618 (1986).

56 Crick, F. H. The origin of the genetic code. *J Mol Biol* **38**, 367-379 (1968).

57 Orgel, L. E. Evolution of the genetic apparatus. *J Mol Biol* **38**, 381-393 (1968).

58 Berget, S. M., Berk, A. J., Harrison, T. & Sharp, P. A. Spliced segments at the 5' termini of adenovirus-2 late mRNA: a role for heterogeneous nuclear RNA in mammalian cells. *Cold Spring Harb Symp Quant Biol* **42 Pt 1**, 523-529 (1978).

59 Salditt-Georgieff, M. & Darnell, J. E., Jr. Further evidence that the majority of primary nuclear RNA transcripts in mammalian cells do not contribute to mRNA. *Mol Cell Biol* **2**, 701-707 (1982).

60 Maxwell, E. S. & Fournier, M. J. The small nucleolar RNAs. *Annu Rev Biochem* **64**, 897-934 (1995).

61 Beltrame, M. & Tollervey, D. Identification and functional analysis of two U3 binding sites on yeast pre-ribosomal RNA. *EMBO J* **11**, 1531-1542 (1992).

62 Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* **74**, 3171-3175 (1977).

63 Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L. & Steitz, J. A. Are snRNPs involved in splicing? *Nature* **283**, 220-224 (1980).

64 Rogers, J. & Wall, R. A mechanism for RNA splicing. *Proc Natl Acad Sci U S A* **77**, 1877-1879 (1980).

65 Maniatis, T. & Reed, R. The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature* **325**, 673-678 (1987).

66 Tycowski, K. T., Kolev, N. G., Conrad, N. K., Fok, V. & Steitz, J. A. *12 The Ever-Growing World of Small Nuclear Ribonucleoproteins.* (2006).

67 Will, C. L. & Lührmann, R. *13 Spliceosome Structure and Function.* (2006).

68 Collins, C. A. & Guthrie, C. The question remains: is the spliceosome a ribozyme? *Nat Struct Biol* **7**, 850-854 (2000).

69 Doudna, J. A. & Batey, R. T. Structural insights into the signal recognition particle. *Annu Rev Biochem* **73**, 539-557 (2004).

70 Peluso, P. *et al.* Role of 4.5S RNA in assembly of the bacterial signal recognition particle with its receptor. *Science* **288**, 1640-1643 (2000).

71 Shen, K. & Shan, S. O. Transient tether between the SRP RNA and SRP receptor ensures efficient cargo delivery during cotranslational protein targeting. *Proc Natl Acad Sci U S A* **107**, 7698-7703 (2010).

72 Pasquinelli, A. E. *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86-89 (2000).

73 Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297 (2004).

74 Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215-233 (2009).

75 Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835-840 (2010).

76 Giraldez, A. J. *et al.* Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312**, 75-79 (2006).

77 Wu, L., Fan, J. & Belasco, J. G. MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A* **103**, 4034-4039 (2006).

78 Eulalio, A. *et al.* Deadenylation is a widespread effect of miRNA regulation. *RNA* **15**, 21-32 (2009).

79 Lau, N. C. *et al.* Characterization of the piRNA complex from rat testes. *Science* **313**, 363-367 (2006).

80 Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**, 761-764 (2007).

81 Olovnikov, I., Aravin, A. A. & Fejes Toth, K. Small RNA in the nucleus: the RNA-chromatin ping-pong. *Curr Opin Genet Dev* (2012).

82 Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-11672 (2009).

83 Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689-693 (2010).

84 Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-1076 (2010).

85 Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-124 (2011).

86 Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**, 1113-1117 (2010).

87 Zappulla, D. C. & Cech, T. R. Yeast telomerase RNA: a flexible scaffold for protein subunits. *Proc Natl Acad Sci U S A* **101**, 10024-10029 (2004).

88 Korostelev, A. & Noller, H. F. The ribosome in focus: new structures bring new insights. *Trends Biochem Sci* **32**, 434-441 (2007).

89 Nickerson, J. A., Krochmalnic, G., Wan, K. M. & Penman, S. Chromatin architecture and nuclear RNA. *Proc Natl Acad Sci U S A* **86**, 177-181 (1989).

90 Bernstein, E. & Allis, C. D. RNA meets chromatin. *Genes Dev* **19**, 1635-1655 (2005).

91 Moazed, D. Small RNAs in transcriptional gene silencing and genome defence. *Nature* **457**, 413-420 (2009).

92 Maison, C. *et al.* Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nat Genet* **30**, 329-334 (2002).

93 Bernstein, E. *et al.* Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Mol Cell Biol* **26**, 2560-2569 (2006).

94 Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A. & Panning, B. Xist RNA and the mechanism of X chromosome inactivation. *Annu Rev Genet* **36**, 233-278 (2002).

95 Plath, K. *et al.* Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300**, 131-135 (2003).

96 Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750-756 (2008).

97 Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet* **30**, 167-174 (2002).

98 Wutz, A. *et al.* Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature* **389**, 745-749 (1997).

99 Koerner, M. V., Pauler, F. M., Huang, R. & Barlow, D. P. The function of non-coding RNAs in genomic imprinting. *Development* **136**, 1771-1783 (2009).

100 Nagano, T. *et al.* The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**, 1717-1720 (2008).

101 Pandey, R. R. *et al.* Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell* **32**, 232-246 (2008).

102 Wang, X. *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* **454**, 126-130 (2008).

103 Kino, T., Hurt, D. E., Ichijo, T., Nader, N. & Chrousos, G. P. Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal* **3**, ra8 (2010).

104 Hung, T. *et al.* Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet* **43**, 621-629 (2011).

105 Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell* **146**, 353-358 (2011).

106 Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147**, 358-369 (2011).

107 Lee, J. T. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev* **23**, 1831-1842 (2009).

108 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

109 Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

110 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

111 Grivna, S. T., Beyret, E., Wang, Z. & Lin, H. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* **20**, 1709-1714 (2006).

112 Lee, J. T., Davidow, L. S. & Warshawsky, D. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet* **21**, 400-404 (1999).

113 Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916-919 (2002).

114 Rinn, J. L. *et al.* The transcriptional activity of human Chromosome 22. *Genes Dev* **17**, 529-540 (2003).

115 Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149-1154 (2005).

116 Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563-573 (2002).

117 De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* **8**, e1000384 (2010).

118 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187 (2010).

119 Ravasi, T. *et al.* Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res* **16**, 11-19 (2006).

120 Pang, K. C., Frith, M. C. & Mattick, J. S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22**, 1-5 (2006).

121 Martens, J. A., Laprade, L. & Winston, F. Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature* **429**, 571-574 (2004).

122 Schmitt, S. & Paro, R. Gene regulation: a reason for reading nonsense. *Nature* **429**, 510-511 (2004).

123 Schmitt, S., Prestel, M. & Paro, R. Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev* **19**, 697-708 (2005).

124 Bumgarner, S. L. *et al.* Single-Cell Analysis Reveals that Noncoding RNAs Contribute to Clonal Heterogeneity by Modulating Transcription Factor Recruitment. *Mol Cell* **45**, 470-482 (2012).

125 Leighton, P. A., Ingram, R. S., Eggenschwiler, J., Efstratiadis, A. & Tilghman, S. M. Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature* **375**, 34-39 (1995).

126 Jones, B. K., Levorse, J. M. & Tilghman, S. M. Igf2 imprinting does not require its own DNA methylation or H19 RNA. *Genes Dev* **12**, 2200-2207 (1998).

127 Wang, J. *et al.* Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature* **431**, 1 p following 757; discussion following 757 (2004).

128 Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**, 556-565 (2007).

129 Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nat Genet* **41**, 572-578 (2009).

130 Wilusz, J. E., Freier, S. M. & Spector, D. L. 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135**, 919-932 (2008).

131 Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849-1851 (2008).

132 Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300 (2011).

133 Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-510 (2010).

134 Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).

135 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).

136 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858 (2009).

137 Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112 (2009).

138 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).

139 Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227 (2009).

140 Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409-419 (2010).

141 Ivanova, N. *et al.* Dissecting self-renewal in stem cells with RNA interference. *Nature* **442**, 533-538 (2006).

142 Cabili, M. N. *et al.* Integrative Annotation of Human Large Intergenic Non-Coding RNAs Reveals Global Properties and Specific Subclasses. *Genes Dev* (2011).

Chapter 1: Chromatin signature reveals more than a thousand well conserved, large non-coding RNAs in mammals

In this chapter, we describe the identification of a class of thousands of well conserved lincRNAs in the human and mouse genomes by exploiting a chromatin signature of actively transcribed genes.

Parts of this work were first published as:

Guttman M, Amit I, Garber M, French C, Lin M, Feldser D, Huarte M, Cabili M, Carey BW, Cassady J, Jaenisch R, Mikkelsen T, Jacks T, Hacohen N, Bernstein BEB, Kellis M, Regev A, Rinn JL, Lander ES. (2009) Chromatin structure reveals over a thousand highly conserved, large non-coding RNAs in mammals. ***Nature***. 458(7235):223-7

Khalil AM*, Guttman M*, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. ***Proceedings of the National Academies of Science*** 106(28):11667-72

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. (2009). Identifying Novel Constrained Elements by Exploiting Biased Substitution Patterns. Conference on Intelligent Systems for Molecular Biology (ISMB) ***Bioinformatics*** 25(12):i54-62

The role of large intergenic non-coding RNAs (lincRNAs) in mammals remains highly controversial. There is evidence of extensive transcription across the mammalian genome, based on large-scale cDNA sequencing and hybridization to DNA microarrays. However, the functional significance of most of these transcripts has been unclear because their expression levels are often very low and their sequence shows little evolutionary conservation. At present, only about a dozen functional lincRNAs have been well-characterized, with roles in diverse processes such as X-inactivation, HOX gene regulation and repression of nuclear import. Here, we report a new approach to identifying lincRNAs by using chromatin-state maps to define discrete transcriptional units that do not overlap any known protein-coding genes. By examining four mouse cell types and six human cell types, we identify ~3500 such transcriptional units. The majority produce large multi-exonic RNAs that show strong purifying selection in their genomic loci, exonic sequences, and promoter regions, but lack protein-coding potential. These lincRNAs are regulated by the canonical general transcriptional machinery and by specific transcription factors including Oct4, Sox2, and Nanog in ES cells. Together, these results demonstrate that functional lincRNAs are abundant in the mammalian genome.

Introduction

Although the human genome has now been sequenced[1] the functional elements encoded within it remain largely uncharted. More than 5% of the genome has clearly been under purifying selection over the past 100 million years and thus must be functional[2],but only ~1.2% encodes protein[1-3]. Within the remainder, an important and growing category consists of genes encoding functional RNA molecules. These include: classical examples such as ribosomal, transfer and splicing-associated RNAs; and a recently discovered cast of tiny RNAs, including microRNAs and other families[4-12].

There are tantalizing signs of another class of functional RNAs, which we will refer to as *large intergenic non-coding RNAs* (lincRNAs). There are currently about a dozen well-characterized examples in mammals, with transcript size ranging from 2.3 to 17.2 kb[13,14]. They each play distinctive biological roles through diverse molecular mechanisms, including functioning in X-chromosome inactivation (XIST, TSIX)[14,15], imprinting (H19, AIR)[13,16,17], *trans*-acting gene regulation (HOTAIR)[18] and regulation of nuclear import (NRON)[19]. Importantly, these well-characterized lincRNAs show clear evolutionary conservation confirming that they are functional. It has been unclear whether these few cases represent quirky exceptions, or whether they herald an entire world of functional large intergenic non-coding RNAs.

Over the past decade, there have been extensive efforts to characterize mammalian RNAs, both by massive shotgun sequencing of cDNAs[20] and by hybridization of RNA to 'tiling arrays' representing genomic sequence[21-25]. These studies have identified evidence of widespread transcription, leading to the recent suggestion that nearly the entire genome is transcribed into RNA under some circumstances[3].

It has been unclear, however, whether the majority of these transcripts are biologically functional or merely reflect "transcriptional noise"[3,26-29]. Many of the reported transcripts occur at extremely low levels – orders of magnitude below the levels for protein-coding genes. While some of the transcripts show tissue-specific expression, it has been argued that such expression patterns may simply reflect transcription from randomly-occurring RNA polymerase binding sites in open chromatin in a given cell type, or possibly transcription whose purpose is to modify local chromatin structure[30-32]. Given the ubiquity of RNA transcription, it has been hard to discern which of these transcripts, if any, are functional.

The only way to prove that an RNA is functional is to show that its disruption has a deleterious consequence. In principle, this can be done in two ways: (i) by demonstrating clear evolutionary conservation (that is, purifying selection) or (ii) by knocking out the gene in a laboratory experiment. The first approach is vastly more efficient, because it involves computational analysis that can be applied to an entire collection of RNAs, and more general because it does not require prior knowledge of the biological context in which the gene functions.

Attempts to prove the functionality of non-coding RNAs through evolutionary conservation have been discouraging. Initial studies found that non-coding transcripts show no better conservation than random intergenic sequence[33]. A recent study reported that the conservation levels are slightly better than random – but still extremely low[29]. Some authors have proposed that most non-coding RNAs are functional, but show little conservation because they are evolving rapidly[28]. A reasonable hypothesis is that the current transcript collections contain some *bona fide* functional non-coding RNAs hidden in a much larger background of transcriptional noise. But, there has been no systematic way to extract wheat from chaff.

We therefore decided to take an entirely different approach to discovering functional non-coding RNAs, which relies on exploiting chromatin structure. We recently developed an efficient technique to create genome-wide chromatin-state maps, using chromatin immunoprecipitation (ChIP) followed by massively parallel sequencing. We observed that genes actively transcribed by RNA polymerase II are marked by Histone 3-Lysine 4 trimethylation (H3K4me3) at their promoter and Histone 3-Lysine 36 trimethylation (H3K36me3) along the length of the transcribed region[34,35]. This distinctive structure, which we will refer to as a 'K4-K36 domain', is found both at known protein-coding genes as well as well-known ncRNAs, including miRNAs[34]. We hypothesized that, by identifying K4-K36 structures that lay outside known protein-coding gene loci, we could systematically discover large intergenic non-coding RNAs.

Here, we report the identification of ~3500 large non-coding RNAs in the mouse and human genomes using our chromatin-based approach in four cell types. These RNAs share many of the same features as protein coding transcripts but lack the evolutionary signature of protein coding potential. The ncRNAs are typically poly-adenylated, 5'-capped, and spliced. We characterize the conservation properties of the ncRNAs and find that the genomic locus, exons, and promoters of these regions are highly conserved. These ncRNAs are regulated in cell-type specific manners by key transcription factors. It is likely that many additional lincRNAs exist and can be identified by studying additional cell types.

Results

Discovery of novel large RNAs based on chromatin structure

We searched for K4-K36 domains in genome-wide chromatin-state maps of four cell types -- mouse embryonic stem cells (mES), mouse embryonic fibroblasts (mEF), mouse lung fibroblasts (mLF), and neural precursor cells (NPC) – created by chromatin immunoprecipitation-sequencing (ChIP-Seq)[34]. For this purpose, we developed a computational algorithm that identifies K4-K36 domains of at least 5Kb in size (Methods). Using this approach we identified 10,623 K4-K36 domains. We then excluded all regions that overlap a protein-coding gene annotated in the mouse genome, as well as those that are syntenic to protein-coding genes annotated in the human, rat or dog genomes. We also excluded all microRNA genes annotated in the MIRBASE database[36].

The analysis yielded a set of 1684 unannotated intergenic K4-K36 domains (**Figure 1a**). To exclude the possibility that any of the domains might represent extensions of known protein-coding genes, we then applied a stringent criterion to define a 'conservative' subset. Specifically, we excluded any K4-K36 domains that lie within any reported cDNA that extends into a region containing a protein-coding gene (even if the cDNA does not encode a protein-coding gene). This conservative subset contains 1109 K4-K36 domains. In our analysis below, we refer to the conservative set, but similar conclusions hold for the full set of 1,684 domains.

We tested whether the intergenic loci with K4-K36 domains produce RNA transcripts. We designed DNA microarrays containing oligonucleotides that tile across a random sample of 350 of the regions (50 base probes spaced every 10 bases) and various control regions, and we hybridized poly(A)$^+$-selected RNA from each of the four cell types to the arrays. We developed an algorithm (methods) to identify regions of significant hybridization and used it to define putative exons of transcripts detected at the loci. For ~70% of the intergenic loci with K4-K36 domains present in a cell type, we found clear evidence of RNA transcription in that cell type (**Figure 1b,c**). The proportion is similar to what we see for protein-coding genes: ~72% of K4-K36 domains corresponding to known protein-coding genes in these cell-types show similarly strong evidence of RNA transcription. For intergenic loci with K4-K36 domains present in multiple cell types, we typically observed corresponding patterns of hybridization in the various cell types, consistent with the presence of a reproducible transcript.

To independently validate the putative exons detected on tiling arrays, we employed RT-PCR. We confirmed the presence of 93 of 107 (87%) randomly selected exons, representing at least one exon from 19 of 20 K4-K36 domains tested. We also confirmed the connectivity of consecutive exons in 52 of 67 (78%) of cases, including one from each of 16 K4-K36 domains tested (**Figure 1c**). We further validated the presence of discrete transcripts by hybridization to RNA northern blots. We found that 15 of 17 tested loci show detectable distinct bands (**Figure 1b**). The cases that were not confirmed may reflect imperfect definition of the exons based on the tiling array data.

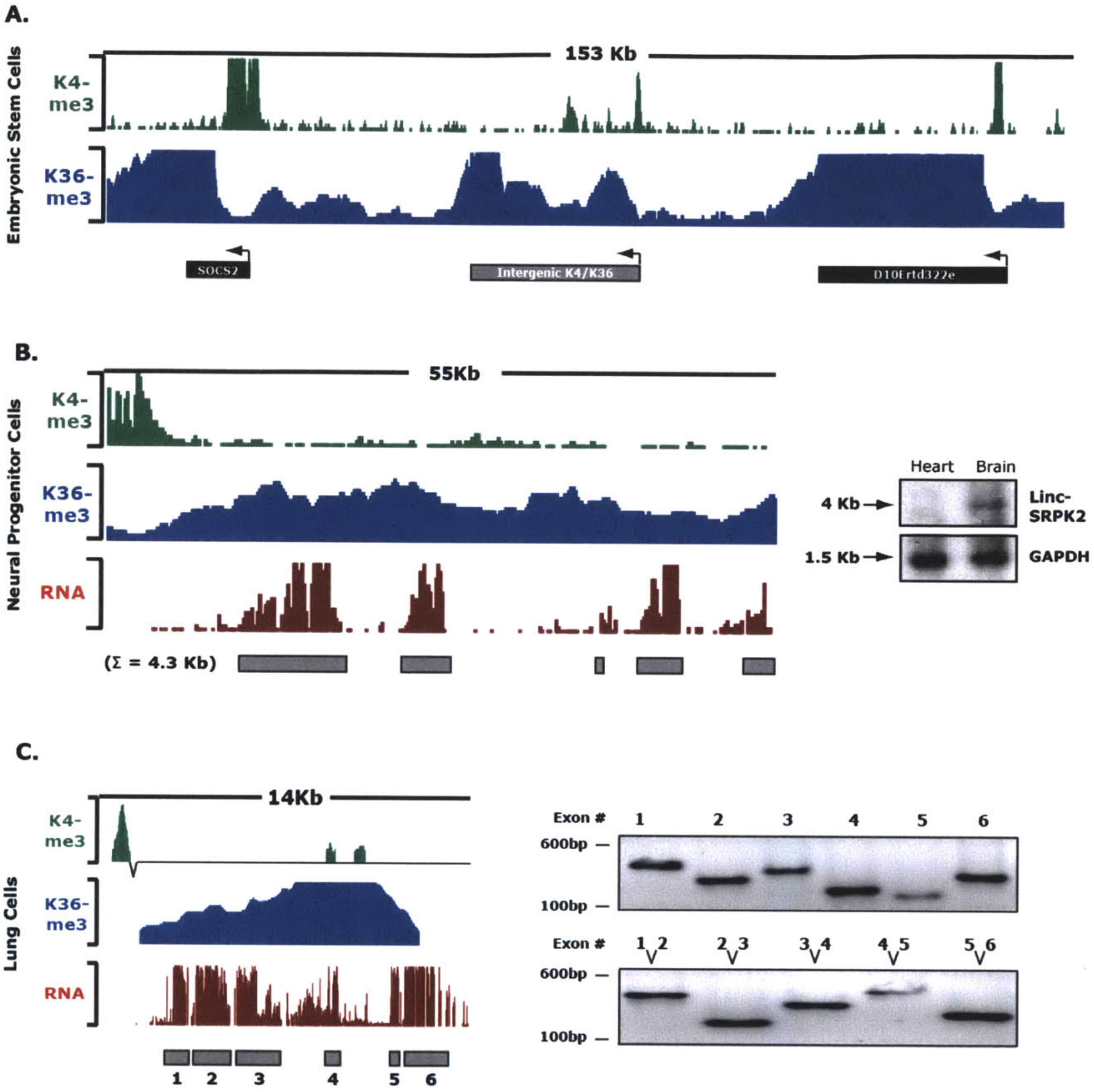


Figure 1: Intergenic K4-K36 domains produce multi-exonic RNAs. (A) Representative example of an intergenic K4-K36 domain and the K4-K36 domains of two flanking protein coding genes. This illustrates that novel K4-K36 domains have similar chromatin structures, at similar enrichment levels, as annotated K4-K36 protein coding domains. Each histone modification is plotted as the number of DNA fragments obtained by ChIP-Seq at each position divided by the average number for each position across the genome. Black boxes indicate known protein coding regions and gray boxes are intergenic K4-K36 domains. Arrowheads indicate the orientation of transcription, inferred from the position of the promoter (K4 domain) (B) Intergenic K4-K36 domains were interrogated for presence of transcription by hybridizing RNA to high-resolution DNA tiling arrays (10bp resolution). Each RNA track is plotted as a normalized hybridization intensity. RNA peaks were determined (Methods) and are represented

by gray boxes. The presence of a 4.5Kb spliced transcript, approximately the same size as predicted from the tiling array, was validated by hybridization to a Northern blot (right). (C) Connectivity between the inferred exons was validated by RT-PCR, indicating that intergenic K4-K36 domains produce spliced multi-exonic RNA molecules. Right top shows RT-PCR validation of each exon. Right bottom validates by RT-PCR the connectivity of each consecutive exon at the predicted sizes.

RNAs encoded by intergenic K4-K36 structures are not protein-coding

Determining whether a transcript is non-coding is challenging because a long non-coding transcript is likely to contain an ORF purely by chance[37]. Accordingly, the case for the absence of coding potential for the XIST and H19 genes rested on the lack of evolutionary conservation of the identified ORFs, the lack of homology to known protein domains, and the inability to template significant protein production[13,38]. We generalized these conservation properties to classify coding potential across thousands of transcripts by scoring conserved ORFs across dozens of species using the 'codon substitution frequency' (CSF) algorithm[39,40].

Computational methods such as CSF[39,40] leverage evolutionary information to determine whether an ORF is conserved across species and provides a general strategy for determining coding potential (**Figure 2a**). Due to the large number of available genome sequences, these methods have the ability to accurately determine conserved coding potential in regions as small as 5 amino acids[41] making them extremely sensitive to potential small peptides, such as the 11 amino acid peptide encoded by the tarsal-less (*tal*) gene[42,43] (**Figure 2b**).

In principle, the transcripts encoded by the intergenic loci with K4-K36 domains could either be novel protein-coding genes or non-coding RNAs. To explore this issue, we tested whether the genomic loci and the exons show the characteristic evolutionary signatures of protein coding genes. The CSF metric can accurately distinguish between a protein-coding exon (CSF > 0) and neighboring UTR or intronic sequence (CSF < 0) (**Figure 2b**). It has been shown to accurately distinguish between protein-coding and non-coding genes in mammalian genomes[44]. While the method was developed for assessing a specific candidate sequence, it can be adapted to scan regions by computing the CSF score in sliding windows and taking the maximum value (max-CSF score). To study evolutionary conservation, we analyzed aligned genomic sequence from 29 mammalian species[45].

Applying the approach to the genomic regions defined by the K4-K36 domains, 95% of protein-coding genes have a high score (max-CSF > 20) whereas 97.5% of untranslated regions (UTRs) and 100% of the known lincRNAs fall below this threshold. We found that 90% of the intergenic loci fall below this threshold as well, with a distribution that is similar to the set of known lincRNAs. We conclude that at least 90% of the intergenic loci with K4-K36 domains lack any significant protein-coding potential across the entire genomic region (**Figure 2c**).

We similarly applied the approach to the putative exons detected by hybridization to the tiling micorarrays (Methods). Whereas all protein-coding exons on our tiling array show a normalized CSF scores greater than 0, approximately 95% of the exons of the intergenic K4-K36 domains have negative scores, inconsistent with the evolutionary signatures of protein-coding genes. Consistent with this, fewer than 2.5% of the exons show any similarity to known protein-coding genes, using the BLASTX program (methods). The other 5% may encode protein-coding genes and are excluded in the analysis below.

Despite their sensitivity, conservation-based methods will fail to detect newly evolved proteins since they will not contain a conserved ORF[39,40]. However, three lines of evidence suggest that this is unlikely for the novel intergenic K4-K36 domains. First, a recent study[44] provides strong evidence that very few new protein-coding genes have appeared in mammals. Second, we show below that, while the intergenic K4-K36 loci lack evolutionary signatures of protein-coding genes, they do show strong evolutionary conservation at the nucleotide level across mammals – indicating that they do not represent recently arising functions. Finally, these domains do not contain conserved ORFs even within the rodent lineage and single nucleotide polymorphisms (SNPs) within mice do not retain the coding region (**Figure 2d**).

Because the majority of the intergenic loci with K4-K36 domains appear to encode large RNAs that lack significant coding potential and do not overlap known annotations, we will hereafter refer to the loci that lack protein-coding conservation as encoding large intergenic ncRNAs (lincRNAs).

It is important to note that a recent paper claims that the majority of lincRNAs are in fact translated [46]. This conclusion is drawn from experimental methods, such as ribosome profiling, which provides a strategy for identifying ribosome occupancy on RNA and has been proposed as a method to distinguish between coding and non-coding transcripts[46]. While this method can clearly distinguish between coding regions and untranslated regions occurring after the ribosome encounters a stop codon and ribosomes are released[47], it has yet to be demonstrated to accurately distinguish between coding and non-coding transcripts[46,48]. Demonstrating this will require identification of the predicted protein products at significant abundances *in vivo*. Importantly, the mere association of an RNA with the ribosome cannot be taken as evidence of protein-coding potential as both the RNase P ribozyme and telomerase RNA can be detected in the ribosome[46,49] despite clear mechanistic roles as ncRNAs[50,51].

An alternative explanation for these observed associations is "translational noise", spurious association that may lead to non-functional translation products[27]. Consistent with this, virtually all of the transcripts suggested to encode small peptides by ribosome profiling[46] lack evolutionary conservation of their proposed coding regions[41,52] even across different mice which is in striking contrast to almost all known protein-coding genes[44] including the few well-characterized functional small peptides[42,43,53 54] (**Figure 2b**). Accordingly, identification of any novel protein-coding gene requires a clear demonstration of the functionality of the protein product *in vivo*[42,43].

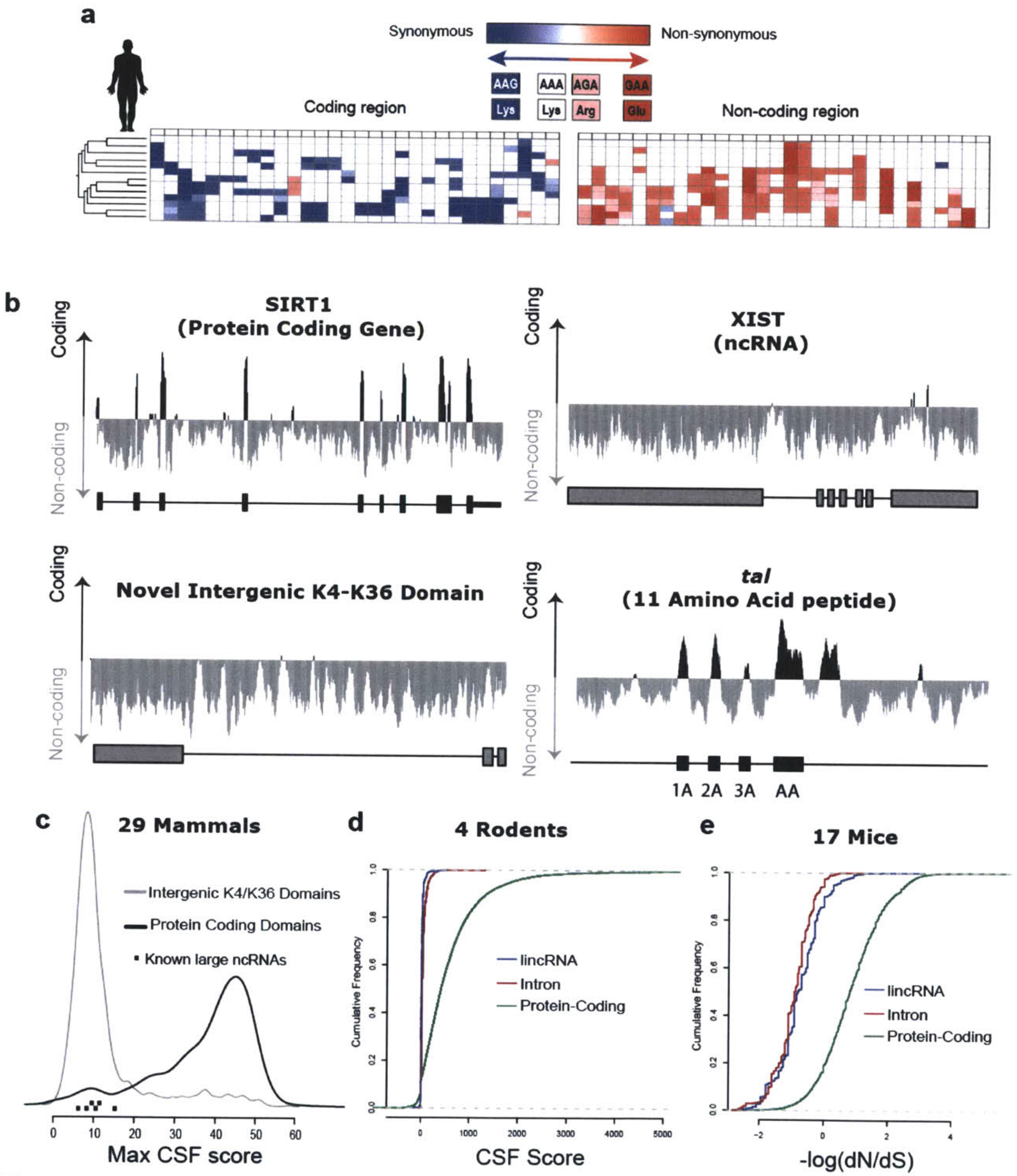


Figure 2: RNA molecules transcribed in intergenic K4-K36 domains do not have coding potential. (a) A cross-species alignment of a coding (left) and non-coding (right) gene. Boxes represent codons and each row represents a different aligned species. Blue boxes represent mutations that cause a synonymous substitution and red boxes represent mutations that cause a non-synonymous substitution. A score capturing the 'coding potential' of a sequence across

species, aligns sequences in all frames and scores mutations that maintain coding potential (blue boxes) relative to mutations that break coding potential (ie non-synonymous mutations, stop codons, and frame shifting insertion/deletions) (red boxes). (b) CSF score in sliding windows across the entire K4-K36 domains of known protein coding gene, Sirt1 (top panel). CSF values in each window that are greater than 0 are represented by black bars, CSF scores below 0 and likely non-coding RNA are shown in grey. The CSF score across the K4-K36 domain of a known ncRNA, XIST (middle panel). The CSF score of a novel lincRNA, linc-ZFAT1, across the K4-K36 domain (bottom panel). The CSF score of a known gene encoding four discrete small peptides of 11 and 32 amino acids (*tarsal-less*), contains positive scores over all known small peptides. Gene annotations are indicated below, showing that high CSF scores occur at the location of protein-coding regions, but not in the two lincRNAs. (c) Density plot of the maximum CSF score (Methods) across intergenic K4-K36 domains (grey) and a random sample of 1000 known protein-coding genes (black). The max-CSF scores for the handful of well characterized lincRNAs are indicated as points at the bottom of the figure. (d) A cumulative distribution of the CSF scores within the rodent lineage for introns (red), protein-coding exons (green), and lincRNAs (blue). (e) A cumulative distribution of the coding potential measured by the dN/dS ratio across 17 mouse strains for introns (red), protein-coding exons (green), and lincRNAs (blue).

Conservation properties of lincRNAs

To assess whether the lincRNA genes are likely functional, we studied their conservation properties. While the genes do not show the evolutionary signatures of protein-coding genes, they show striking and consistent evolutionary conservation in all other ways, and also have additional properties of *bona fide* genes.

(i) lincRNA transcripts show high levels of sequence conservation across mammals. The lincRNA genes show clear conservation by several methods both across the genomic loci and specifically in their exons. Using a method to assess evolutionary constraint on sequences that explicitly models the underlying substitution rate[55] (π, see methods), we found that the lincRNA loci show clear conservation across their transcribed region (defined as the region of K36me3 enrichment) when compared to other intergenic regions (**Figure 3A**). Using another approach based on conserved elements defined by the PhastCons program[56], we also found that the transcribed regions are highly enriched for conserved elements compared to other intergenic regions ($p<.0001$, permutation test).

The exons of lincRNAs show even stronger conservation (**Figure 3B**). The extent of exonic conservation is much higher than seen for random intergenic regions and is similar to that seen for known lincRNAs – although it is lower than that seen for protein-coding exons, likely reflecting a lower degree of constraint on RNA structures than on amino acid codons. We also note that lincRNA exons are significantly more conserved than the UTRs of protein coding genes (Supplemental Figure 4). The presence of strong purifying selection provides strong evidence that the vast majority of the lincRNAs are biologically functional in mammals.

(ii) lincRNA promoters show high levels of sequence conservation across mammals. We defined the promoter-proximal regions of the lincRNAs to be the peaks of K4me3 enrichment nearest to the transcribed unit (median size = 1.1 kb). Applying the conservation scoring methods used above to analyze the transcribed regions, the promoter regions show strong conservation relative to random intergenic regions ($p<.001$, permutation test). Moreover, the distribution of conservation scores is essentially indistinguishable from that seen for the promoters of known protein-coding genes (**Figure 3C**).

(iii) lincRNA loci show conserved chromatin structure in human. We constructed chromatin-state maps in human lung fibroblasts (hLF) and compared them to the chromatin structure seen in mouse lung fibroblasts (mLFs), to study the occurrence of K4-K36 domains in the two species. Interestingly, ~70% of the human K4-K36 domains also exhibited a K4-K36 domain in the syntenic region of the mouse genome. The proportion is similar to that seen for protein-coding genes (~80%). The lack of complete correspondence may simply reflect thresholds of detection.

lincRNAs promoters exhibit canonical features of general transcriptional initiation.

The lincRNA promoters show a striking enrichment of 'CAGE tags' that mark transcriptional start sites (TSS)[57] (**Figure 3D**). CAGE tags are obtained by shotgun sequencing of cDNAs prepared from RNA molecules captured on the basis of containing the 7-methylguanosine cap structure that marks the 5'-end of mRNAs[58]; they have been systematically catalogued in mouse. Most of the promoters regions of lincRNAs (85%) contain a significant cluster of CAGE tags, with the density tightly localized around the promoter. The proportion and localization are similar to that seen for protein-coding genes[57].

Additionally, we found that the lincRNA promoters show strong enrichment for binding of RNA PolII and Transcriptional initiation factor 2D (TF2D) in mouse ES cells, compared to random genomic regions ($p<2x10^{-16}$, Wilcoxon test). Together, these results suggest that the transcription and processing of lincRNAs is similar to that for protein-coding genes – including PolII transcription, 5'-capping and polyadenylation (inasmuch as the transcripts are detected in poly(A+)-selected RNA).

In embryonic stem cells, many lincRNA promoters are bound by Oct4 and Nanog.

The promoters of lincRNAs with K4-K36 domains in embryonic stem (ES) cells show strong enrichment for binding of the pluripotency-associated transcription factors Oct4 and Nanog compared to intergenic regions marked by K4me3 but not K36me3 in ES cells, based on analysis of ChIP-PET (Paired End Ditags) data in mouse ES cells[59]. Specifically, by analyzing ChIP-PET data from mouse ES cells we found 249 regions across the genome that are bound by either Oct4 or Nanog and do not correspond to known protein-coding genes. Of these regions, 118 coincide with the promoters of lincRNAs. Of the lincRNA promoters bound by Oct4 or Nanog, 86% occur at lincRNAs whose K4-K36 domain is present only in ES cells – consistent with the fact that Oct4 and Nanog are specific to the pluripotent state (**Figure 3e**).

Among 101 lincRNAs with K4-K36 domains present only in ES cells and containing an Oct4 or Nanog binding site, we noticed that one lincRNA was ~100 kb from the Sox2 locus, which encodes another key transcription factor associated with pluripotency. We cloned the promoter of this locus (which we will refer to as lincRNA-Sox2) upstream of a luciferase reporter gene and transfected the construct into mouse cells transiently expressing either Oct4, Sox2, or both, as well as several controls. We found that Sox2 and Oct4 were each sufficient to drive expression of this lincRNA promoter. We further found that Oct4 and Sox2 together synergistically increased expression of lincRNA-Sox2 (**Figure 3f**). Together, these results demonstrate that lincRNA promoters are directly regulated by key transcription factors in a cell-specific manner.

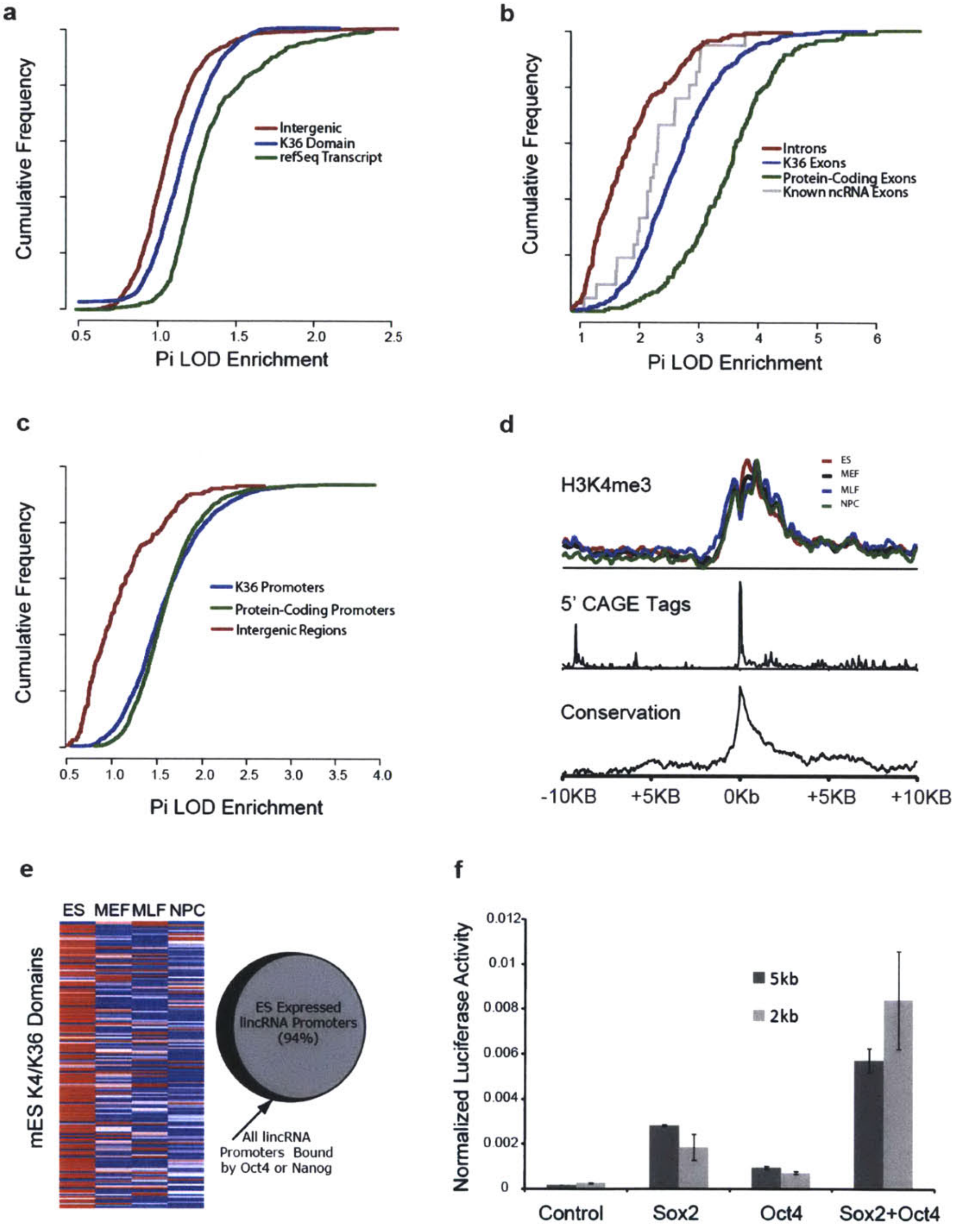


Figure 3: lincRNA K4-K36 domains, exons, and their promoters are highly conserved. (A) Cumulative distribution of sequence conservation across 21 mammalian species for the genomic region of the K4-K36 domain of each lincRNA (blue), protein coding gene (green), and random intergenic regions (red). (B) Cumulative distribution of sequence conservation across 21 mammals for lincRNA exons (blue), protein coding exons (green), and protein coding introns (red). (C) Cumulative distribution of sequence conservation across 21 mammals for the promoters of each lincRNA (blue) and protein coding promoters (green). The X-axis is the

conservation measure (Methods) normalized by random genomic regions; thus larger scores reflect higher conservation. (D) Enrichment of various promoter features plotted as the distance from the start of the K36me3 region averaged across all lincRNAs. (Top) Enrichment in each cell type of K4me3 domains across ES (red), MEF (black), MLF (blue), NPC (green) is shown. (Middle) Enrichment of 5' CAGE tag density representing 5' end of RNA molecules, indicating the K4me3 regions correspond to apparent transcriptional start site. (Bottom) Conservation scores within the K4me3 region. (E) K36me3 enrichment across 4 cell types for lincRNAs bound by Oct4 or Nanog. Red indicates high enrichment and blue low enrichment. The Venn diagram indicates the total number of ES specific lincRNAs bound by Oct4 or Nanog (Gray) compared with the total number of Oct4 or Nanog bound regions (Black). (F) Direct transcriptional regulation of lincRNA-Sox2 promoter by Sox2 and Oct4. The lincRNA-Sox2 promoter was cloned into a luciferase reporter construct and assayed for transcriptional activity with co-transfection with either Sox2 or Oct4 alone, in combination and the same reporter construct without the lincRNA-Sox2 promoter. The y-axis represents the transcriptional activity of this promoter relative to a renilla construct control to control for transfection efficiency.

Identification of human lincRNAs

While this studies clearly demonstrate that there are many functional lincRNAs, a key questions remains: How many lincRNAs are encoded in mammalian genomes? To further extend the catalog of lincRNAs, we sought to analyze chromatin-state maps of six human cell types: human embryonic stem cells (hESC)[60], two hematopoetic stem cells (CD133+ and CD36+)[61], T-cells [35], lung fibroblasts (hLF)[41], and normal embryonic kidney (hEK). Using our previous computational approach, we identified K4-K36 domains that are well-separated from (i) the regions containing known protein-coding genes and all known classes of small non-coding RNAs in human and (ii) the orthologous regions of known protein-coding genes in mouse, rat and dog. We also eliminated the orthologous regions of our previously identified mouse lincRNAs. (We previously showed that, for similar cell types in mouse and human, lincRNA loci show cross-species conservation not only at the level of nucleotide sequence, but also with respect to the presence of K4-K36 domains[41].

We found a total of 1833 novel intergenic K4-K36 domains in these six human cell types. We analyzed the coding potential of each such K4-K36 domain using the codon substitution frequency score (see Methods) and found that <8% showed any evidence of protein-coding capacity[39]. After eliminating these cases, we were left with 1703 loci encoding putative lincRNA genes.

To test whether these loci actually encode lincRNAs, we designed genomic tiling microarrays (at 10 base resolution) across 1147 of the 1703 loci (see Methods) to determine their exonic structure. We hybridized poly(A^+)-amplified RNA from hES, brain, breast, hEK, hFF, hLF, K562, ovary, skin, spleen, testis, thymus tissues. We analyzed the hybridization data using our previously reported peak-calling algorithm. This analysis revealed multi-exonic RNA transcripts in 74% of the K4-K36 domains examined. There was an average of 9 exons per K4-K36 domain (total of 7523 exons). We further focused on the 535 K4-K36 domains that were discovered in cell types in which RNA from the same cell type was hybridized. In these three cell types, RNA hybridization revealed multi-exonic RNA transcripts in 85% of the tested loci; this detection rate is similar to that previously seen for K4-K36 domains corresponding to known protein-coding genes and lincRNA loci in mouse[41]. Given that such a high proportion of the human K4-K36 domains tested were validated as encoding lincRNAs, we conclude that the vast majority of the full set of 1703 loci encode *bona fide* lincRNAs.

We then studied the evolutionary conservation of the novel lincRNA loci. For each exon, we calculated the extent of sequence conservation across 29 mammalian species as previously described (see Methods). The novel lincRNAs showed evolutionary conservation at levels similar to those seen for the lincRNAs in mouse.

Combining the 1586 human orthologs of the lincRNA genes reported in our previous study with the 1703 newly discovered human lincRNA genes identified in this study, our catalog of human lincRNA genes now includes 3289 distinct loci. This catalog is certain to be incomplete, because it is based on chromatin-state maps of only ten cell types (four mouse and six human). Nonetheless, it is possible to make a rudimentary estimate of the total number of human lincRNAs based on the observation that 73% of all protein-coding genes are expressed in at least one of the ten cell types analyzed here. If a similar proportion applies to lincRNAs, the total number of human lincRNAs would be estimated to be ~4500. If lincRNAs have expression patterns that are more tissue- or condition-specific, the total number could be considerably higher. Obtaining a complete catalog will require generating chromatin-state maps across many more tissues.

Discussion

Although it has become clear that abundant transcription in mammalian genomes generates many large RNAs without protein-coding capacity[62], the biological significance of these molecules remained in doubt because most of the cases identified by shotgun cDNA sequencing or microarray hybridization show little evidence of sequence conservation indicative of function. It is likely that many–perhaps most—of these RNAs represents transcriptional noise, but that hidden among them are important classes of ncRNAs.

We therefore developed an alternative approach to finding lincRNAs, based on identifying novel genomic regions carrying a distinctive chromatin structure associated with actively transcribed genes (K4-K36 domains). Studying chromatin-state maps across four mouse cell types and 6 human cell types revealed ~3500 lincRNA loci. Analysis of these loci show that most indeed encode large RNA transcripts that exhibit strong cross-species conservation at the level of sequence (in both exons and promoters), transcription and chromatin structure and show promoter binding by both general (PolII) and relevant specific transcription factors (Sox2 and Oct4 in ES cells). Yet, the vast majority shows no evidence of protein-coding capacity. Overall, these properties closely resemble those of the roughly dozen known lincRNAs (e.g. HOTAIR, and XIST)[62].

Our results show that chromatin structure can identify sets of lincRNAs that show a high degree of evolutionary conservation, implying that they are biologically functional. (We do not exclude the possibility that lincRNAs identified by shotgun sequencing that fail to show conservation are nonetheless functional, but other evidence will be required to establish this point.) Together, the results suggest that the mammalian genome may encode a large collection of functional lincRNAs. The precise number of lincRNAs is difficult to estimate from the current data. As a first approximation, we note that ~75% of all protein-coding genes are expressed in at least one of the four cell types. If a similar ratio pertains to lincRNAs, the total might exceed 4500. If lincRNAs are more tissue specific, however, the total could be much higher. We emphasize that our analysis focuses only on *large intergenic* RNAs. It thus excludes many other classes of non-coding RNAs, including those overlapping protein coding genes such as promoter associated RNAs[25], intronic and antisense transcription[13] and small RNAs, such as miRNAs, piRNAs and snoRNAs[4]. In addition, our analysis would likely not be sensitive to extremely low-level transcription; such low level transcription could be biologically important.

The next steps will require detailed characterization of structure and function of lincRNAs – including obtaining full-length cDNAs; expanding the gene-expression compendium to more tissues and more lincRNAs; performing RNAi-mediated knockout in appropriate settings; studying interactions with cellular proteins and DNA; and finally genetic deletion of lincRNAs in mouse model systems. Whatever their functions, the well-conserved lincRNAs represent an important new contingent in the growing population of the modern 'RNA world'.

METHODS

Chromatin Map Data

Chromatin data for H3K4me3 and H3K36me3, for mouse Embryonic Stem Cells (mES), mouse Embryonic Fibroblasts (MEF), and mouse neural precursor cells (NPC) were taken from Mikkelsen et al. 2007 and were downloaded from (ftp://ftp.broad.mit.edu/pub/papers/chipseq/). Chromatin data in mouse lung fibroblasts and human lung fibroblasts were generated as previously described[34].

Identifying K4-K36 Enriched Domains

To identify regions of enriched chromatin marks we employ a sliding window approach: we slide windows, score each window based on the number of ChIP fragments, compute a threshold for significance, and use the significant windows to define intervals. Specifically: **(i)** We fix a window size w and slide it across each position of the genome. For each position, we compute a score, S_w, as the number of reads aligned within the window. **(ii)** To identify windows that have significantly more reads than would be expected by chance, we define a null model based on the randomization of read locations across the genome. This null model is estimated as a Poisson distribution where λ is defined as the number of reads in the library divided by the number of possible non-overlapping windows of size w. **(iii)** Given the null model, we choose a threshold T on the score such that the genome-wide probability that the Score S_w exceeds the threshold T by chance is less than 0.01 ($\text{Prob}(S_w>T)<0.01$). We therefore cannot compute this probability exactly, since the scores S_w occur in overlapping windows they are not independent values or multiple testing corrected values. We therefore estimate it genome-wide across overlapping windows using the scan statistic procedure [63]. Therefore, windows that pass this threshold T are significantly enriched after multiple testing correction. **(iv)** We retain only windows that pass this threshold T, and merge overlapping significant windows into a single contiguous interval. We refine the boundaries of this interval by taking the maximum contiguous subsequence. (**vi**) To generalize for multiple window sizes, we compute a threshold for each window size separately and repeat the above procedure, merging windows of different sizes. (**v**) Finally, we score each interval and test if it is significantly enriched using the same scan statistic approach introduced above. The result is a set of intervals and their *p*-values.

To identify the intervals that encode intergenic K4-K36 domains we applied this approach to independently find K4 and K36 regions. We filtered all K4 and K36 regions that overlapped with known annotations (as described below). We identified all K4 and K36 intervals that were adjacent. To define a K4-K36 domain we required that the interval from the K4 region through the end of the K36 region was significantly enriched for K36 using the same scan statistic approach. We then filter the list by regions that are at least 5Kb in length.

All results were produced in the March 2006 (MM8) freeze of the Mouse genome.

Filtering Gene Lists

We filtered the list of K4-K36 domains to eliminate all regions annotated as containing a protein coding gene in mouse or orthologous protein coding genes in human, rat, or dog.

We obtained the list of all human protein coding genes as determined by Clamp et al. 2007 in the Human genome (Hg17) from (http://www.broad.mit.edu/mammals/alpheus/data/) and used the liftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver) tool to identify their orthologous location in the mouse genome (MM8). We also used the list of allRefSeq protein coding genes (MM8) along with all RefSeq genes annotated in the Human (Hg18), Rat (Rn4), and Dog (canFam2) genomes. All refSeq gene lists were obtained from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html). The liftOver tool was similarly used to place genes from other species in the mouse genome (MM8). In our analysis, we eliminated all regions that overlapped any portion of a protein coding locus, including introns, exons, and UTRs. We also excluded all regions that overlap a known miRNA gene obtained from the MIRBASE database[64].

RNA Preparation and Sources

We purchased total RNA for mouse lung, brain, testes, and ovary (Ambion). We isolated RNA from Mouse whole embryo, forelimb, and hindlimb from developmental time points E9.5, 10.5 and 13.5. These mice were generated using timed mating embryo isolation and dissection. We obtained mES, mEF, and NPC RNA extracted from cell lines using the Qiagen RNAEasy Kit. Bone Marrow dendritic cells were extracted as previously described[65], and stimulated with various ligands (see below). We extracted RNA after 6 hours using the Qiagen RNAEasy Kit.

Tilling array design, hybridization and analysis.

High resolution DNA tiling arrays containing 2.1 million features were designed on the Nimblegen platform (HD2) to represent a random sampling of ~400 intergenic K4-K36 domains identified in the mouse genome. Total RNA from mES, mLF, NPC and mEF was amplified using poly-dT and labeled as described[18]. Arrays were hybridized and washed according to the Nimblegen protocols and kits (Roche/Nimblegen). Array image files were processed using Nimblescan (Roche/Nimblegen) and arrays were normalized by mean centering the data.

A second array was designed on the Nimblegen platform (HD1) arrays containing 300,000 and representing ~150 K4-K36 domains. We hybridized mES, mEF, mLF, NPC, BMDC, TLR2, TLR4, TLR9, lung (Ambion), brain (Ambion), testis (Ambion), ovary (Ambion), whole embryo, forelimb, and hindlimb to this array from developmental time points E9.5, 10.5 and 13.5. Total RNA was amplified and labeled for array as described[18]. For both arrays we tiled across all Hox genes as well as handful of other genes as controls.

Determining Transcribed Segments From Tiling Arrays

To identify transcribed regions of K4-K36 domains, we hybridized poly-A RNA to a tiling microarray. We developed a statistical algorithm to identify peaks in hybridization, representing likely exons in a mature transcript.

We normalized the data by dividing each probe value by the average probe intensity across the array. We scanned the K4-K36 domains using sliding windows of width w. We computed a score defined as the sum of the normalized probe intensities. To determine the significance of this score we permuted the intensity values assigned to each probe and recalculated the statistic. We took the value for each permutation as the maximum score obtained for any random region. We performed 1000 permutations and assigned a multiple testing corrected p-value to each region based on its rank within this distribution. All regions with a p-value less than 0.05 were retained. After determining the transcribed segments from each sample on the array, we defined exons as the union of all bases covered by a transcribed segment.

RNA blot analysis.

RNA blot analysis was performed on Ambion first choice RNA blots (Ambion). The blots contained RNA from various mouse tissues including heart, brain, liver, spleen, kidney, whole embryo, lung, thymus, testes, and ovary. Probes were designed to selected lincRNA exons, as determined by tiling arrays, and hybridized to the RNA blot. Probes were prepared by PCR of genomic regions followed by biotin incorporation using the North2South ® Biotin Labeling Kit (Pierce). Probes were hybridized to the RNA blot for 14-15 hrs using the North2South ® Hybridization Kit (Pierce). The resulting chemiluminescence was detected using a CCD camera.

RT-PCR

RT-PCR analysis was performed on cDNA libraries made from total RNA from mouse embryo (13.5), lung, brain, MEF, NPC, and ES cells reverse transcribed using Superscript II (Invitrogen) and a poly-dT /random hexamer primer mix.

To validate the presence of individual lincRNA exons and their connectivity within a locus we designed primers within and across exon boundaries using the *Primer3* computer program. PCR was performed as previously described[11] on reverse transcribed cDNAs. We performed a negative control using a no RT reaction and a positive control using the mouse GAPDH gene. The PCR products were analyzed by gel electrophoresis. To confirm splicing across exons, the PCR products were purified with QIAquick PCR Clean-up kits (Qiagen) and then sequenced, using the forward primer. To characterize apparent alternative splicing, the products were run on 2% NuSieve agarose (Lonza) gels and the multiple bands purified with a QIAquick Gel Extraction kit (Qiagen) and sequenced. The primers used are detailed in supplemental table 6.

Multiple Species Alignments

All multiple species alignments were the MULTIZ alignments obtained from the UCSC genome browser (build MM9, http://hgdownload.cse.ucsc.edu/goldenPath/mm9/multiz30way/).

Coding Potential

We tested for protein coding potential of K4-K36 domains by determining the maximum CSF[39,66] score observed across the entire genomic locus. We downloaded the alignments from UCSC and computed the CSF scores across sliding windows of 30 base pairs. We then scanned all 6 possible reading frames in each window. After computing a score for each window, we defined the 'max CSF score' for a K4-K36 domain to be the maximum observed score across the region.

We also computed a 'normalized CSF score' for each individual exon. The 'normalized CSF score' for each exon was defined to be the CSF score for each exon divided by the nucleotide length of the exon.

PhastCons Enrichment Within K4-K36 Domains and Promoter Regions

We downloaded the phastConsElements30way from the UCSC Genome Browser (MM9). We counted the number of phastCons elements within each K4-K36 domain as well as the number of these elements within random, size matched, genomic regions. We constructed a distribution based on the random genomic regions. A p-value was computed based on the rank of the K4-K36 domain's rank within the random genomic distribution. This statistic was similarly applied to the promoter regions of lincRNAs.

Computing Pi Constraint

To detect sequence constraint within large ncRNAs, we chose to use a method that explicitly models the rate of mutation as well as the level of constraint. This is especially relevant for detecting constrained sequences in noncoding regions of the genome since many of these sites are known to be degenerate and can tolerate mutations between certain nucleotides.

Briefly the method we used to identify purifying selection uses a probabilistic neutral model of evolution. Given a phylogenetic tree T and a substitution rate matrix Q, constrained regions will be evident because they are poor fits to the neutral model. In this framework, selection can be apparent in 2 ways either through contraction of the tree length that depends on the intensity of selection (ω) or through a mutation pattern (π) that is not concordant with the rate matrix.

We compute a log-odds score, Pi LOD score, which is the estimate of the sequence evolution compared to neutrally evolving sequences. Sitewise LOD score estimation provides low sensitivity to determine conservation, we therefore integrated across multiple bases. We chose 12mers based on empirically testing the tradeoff between sensitivity and specificity for various kmers. Since the estimation of functional constraint is site specific, we can determine the log-odds score for a region by adding the log-odds scores for each base contained in the region[55].

Exon Conservation and K4-K36 Pi LOD Enrichment

To identify functional constraint within exons of large ncRNAs, we analyzed each exon separately. We computed the Pi LOD score for each 12-mer contained within the exon. We took the maximum 12-mer value for each exon. In order to normalize for the size differences between different exons we computed a size matched random score. To do this we randomly generated size matched regions of the genome and divided the observed LOD score by the average LOD score from the random regions. This normalization procedure produces a score for each exon in the genome that reflects a size-independent level of constraint on each exon. The Observed/Expected score can be interpreted as an enrichment level of the LOD score compared with the genomic average. The distributions of this normalized score were then compared among multiple different classes of genomic units, specifically protein coding introns, exons, and untranslated regions (UTRs), as well as known large non-coding RNAs and non-coding cDNA sequences. This statistic is robust to detecting regions of the genome that, while highly constrained in sequence, are not neccessarily highly conserved over the entirety of the region. We performed the same analysis for the K4-K36 domain, using 75bp windows.

CAGE and RNA PolII Enrichment

For each promoter region, we computed the number of CAGE tags or ChIP-Seq reads for PolII. We compared the number of aligned reads in the promoters to the number of aligned reads in random regions of similar size (excluding repetitive regions of the genome). We computed enrichment with a wilcoxon rank sum statistic between the promoters and random genomic DNA.

CAGE data were downloaded from http://fantom31p.gsc.riken.jp/cage/download/mm5/ and the regions were mapped to the MM8 build using the liftOver tool (http://genome.ucsc.edu/) CAGE scores were computed by summing the number of reads in each tag cluster (Carninci et al. 2006).

RNA Polymerase II ChIP-Seq data was generated as previously described[34] in mES cells.

Oct4/Nanog Enrichments in ES-specific lincRNAs

We used data generated by Loh et al 2006[59]. Briefly, Chromatin Immunoprecipitation (ChIP) was performed using antibodies against Oct4 and Nanog in mES cells. The resulting library was sequenced using 454 sequencing and the 'paired end reads' were mapped to the genome. We downloaded the read clusters mapped on the mouse genome (build MM5) from http://www.nature.com/ng/journal/v38/n4/suppinfo/ng1760_S1.html. We used the liftOver tool (http://genome.ucsc.edu/) to place the reads on the MM8 build of the mouse genome. We defined binding events as clusters with at least 3 independent ChIP sequencing reads, as described in Loh et al. 2006.

In order to determine the enrichment of intergenic Oct4/Nanog binding sites we counted the number of intergenic Oct4/Nanog binding sites that overlapped with a K4me3 peak in the four cell types. Next we counted how many of these regions coincided with the promoter of a lincRNA in the four cell types. We then counted the number of these lincRNA promoter binding events in ES cells and the number that had strong enrichment levels specifically in ES cells. A hypergeometric statistic was applied to determine if the intergenic binding of Sox2 and Oct4 was enriched at lincRNA promoter regions (K4) compared to other intergenic non-lincRNA K4 regions.

Luciferase Reporter Assay

We amplified individual regions of the lincRNA-Sox2 promoter using AccuPrime *Pfx* polymerase (Invitrogen) and cloned the products into the pCR 2.1TOPO vector (Invitrogen). Each region was subsequently cloned into pGL3 firefly Luciferase Reporter Vector (Promega). 293T cells were transiently transfected in triplicate using FuGENE 6 transfection reagent (Roche) and analyzed 24 hours post-transfection by Promega Dual-Luciferase Reporter Assay kit. Analysis was performed using the Veritas Microplate Luminometer system. Expression of the promoter regions was detected by firefly luciferase activity and was determined by obtaining the relative value compared to the transfection control plasmid (CMV *Renilla* luciferase).

Comparison with Previous Transcript Maps

We downloaded the cDNAs sequenced by the FANTOM consortium from (ftp://fantom.gsc.riken.jp/FANTOM3/). We defined two sets of FANTOM transcripts: the first was the ncRNA conservative set, as provided on their site, and the second was all FANTOM cDNA transcripts. We computed significant overlap between the genomic locus of a lincRNA and a FANTOM unit by asking how much of a K4-K36 domain was covered by a FANTOM unit and how much of a FANTOM unit was covered by a K4-K36 unit. We identified all cases in which a transcript overlapped at least 25% of a K4-K36 domain or vice versa. We performed a similar analysis between exons determined by our tiling arrays and FANTOM exons.

References

1 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

2 Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

3 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

4 Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215-233 (2009).

5 Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297 (2004).

6 Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**, 858-862 (2001).

7 Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853-858 (2001).

8 Lee, R. C. & Ambros, V. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**, 862-864 (2001).

9 Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-854 (1993).

10 Grivna, S. T., Beyret, E., Wang, Z. & Lin, H. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* **20**, 1709-1714 (2006).

11 Aravin, A. *et al.* A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**, 203-207 (2006).

12 Girard, A., Sachidanandam, R., Hannon, G. J. & Carmell, M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**, 199-202 (2006).

13 Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Mol Cell Biol* **10**, 28-36 (1990).

14 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**, 38-44 (1991).

15 Lee, J. T., Davidow, L. S. & Warshawsky, D. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet* **21**, 400-404 (1999).

16 Lyle, R. *et al.* The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. *Nat Genet* **25**, 19-21 (2000).

17 Wutz, A. *et al.* Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature* **389**, 745-749 (1997).

18 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323 (2007).

19 Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**, 1570-1573 (2005).

20 Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563 (2005).

21 Rinn, J. L. *et al.* The transcriptional activity of human Chromosome 22. *Genes Dev* **17**, 529-540 (2003).

22 Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242-2246 (2004).

23 Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149-1154 (2005).

24 Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916-919 (2002).

25 Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484-1488 (2007).

26 Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat Cell Biol* **10**, 1106-1113 (2008).

27 Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* **14**, 103-105 (2007).

28 Pang, K. C., Frith, M. C. & Mattick, J. S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22**, 1-5 (2006).

29 Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**, 556-565 (2007).

30 Martens, J. A., Laprade, L. & Winston, F. Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature* **429**, 571-574 (2004).

31 Schmitt, S. & Paro, R. Gene regulation: a reason for reading nonsense. *Nature* **429**, 510-511 (2004).

32 Schmitt, S., Prestel, M. & Paro, R. Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev* **19**, 697-708 (2005).

33 Wang, J. *et al.* Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature* **431**, 1 p following 757; discussion following 757 (2004).

34 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).

35 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).

36 Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**, D140-144 (2006).

37 Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol* **4**, e1000176 (2008).

38 Brockdorff, N. *et al.* The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**, 515-526 (1992).

39 Lin, M. F., Deoras, A. N., Rasmussen, M. D. & Kellis, M. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comput Biol* **4**, e1000067 (2008).

40 Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-282 (2011).

41 Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227 (2009).

42 Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A. & Couso, J. P. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5**, e106 (2007).

43 Kondo, T. *et al.* Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science* **329**, 336-339 (2010).

44 Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* **104**, 19428-19433 (2007).

45 Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-482 (2011).

46 Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* (2011).

47 Kisselev, L. L. & Buckingham, R. H. Translational termination comes of age. *Trends Biochem Sci* **25**, 561-566 (2000).

48 Jiao, Y. & Meyerowitz, E. M. Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. *Mol Syst Biol* **6**, 419 (2010).

49 Li, Y. M. *et al.* The H19 transcript is associated with polysomes and may regulate IGF2 expression in trans. *J Biol Chem* **273**, 28247-28252 (1998).

50 Greider, C. W. & Blackburn, E. H. A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis. *Nature* **337**, 331-337 (1989).

51 Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849-857 (1983).

52 Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-510 (2010).

53 Kastenmayer, J. P. *et al.* Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Res* **16**, 365-373 (2006).

54 Hanada, K., Zhang, X., Borevitz, J. O., Li, W. H. & Shiu, S. H. A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Res* **17**, 632-640 (2007).

55 Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62 (2009).

56 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050 (2005).

57 Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626-635 (2006).

58 Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat Methods* **3**, 211-222 (2006).

59 Loh, Y. H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38**, 431-440 (2006).

60 Ku, M. *et al.* Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* **4**, e1000242 (2008).

61 Cui, K. *et al.* Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **4**, 80-93 (2009).

62 Amaral, P. P., Dinger, M. E., Mercer, T. R. & Mattick, J. S. The eukaryotic genome as an RNA machine. *Science* **319**, 1787-1789 (2008).

63 Glaz, J., Naus, J. I. & Wallenstein, S. *Scan statistics.* (Springer, 2001).

64 Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**, D154-158 (2008).

65 Palliser, D. *et al.* A role for Toll-like receptor 4 in dendritic cell activation and cytolytic CD8+ T cell differentiation in response to a recombinant heat shock fusion protein. *J Immunol* **172**, 2885-2893 (2004).

66 Lin, M. F. *et al.* Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Res* **17**, 1823-1836 (2007).

Chapter 2: *Ab initio* reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincRNAs

In this chapter, we describe the development of a method for *ab initio* reconstruction of a mammalian transcriptome from RNA-Seq data and used it to define the precise sequence of lincRNAs.

This work was first published as:

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A. (2010) Ab initio reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincRNAs. ***Nature Biotechnology*** 28(5):503-10

Recent studies have suggested that mammalian genomes encode thousands of large non-coding RNA (ncRNA) genes, including large intergenic ncRNAs (lincRNAs). Defining the gene structure of lincRNAs is essential for experimental and computational studies of their function. Recent advances in massively parallel cDNA sequencing (RNA-Seq) provide an unbiased way to study a transcriptome, including both coding and non-coding genes. To date, most RNA-Seq studies have critically depended on existing annotated genes, and thus focused on studying expression levels and structural variation in known transcripts. Here, we present Scripture, a method to reconstruct the transcriptome of a mammalian cell using only RNA-Seq reads and the unannotated genome sequence. We apply this approach to mouse embryonic stem cells, mouse neuronal precursor cells, and mouse lung fibroblasts to accurately reconstruct, for the vast majority of expressed genes, the full-length gene structures at single-base resolution, including different splice isoforms. We identify novel high-confidence biological variation in known protein-coding genes, including thousands of novel 5'-start sites and 3'-ends, and almost a thousand novel internal coding exons. We then determine the gene structures of over a thousand lincRNA loci, 27% of which show alternative isoforms. The gene structures demonstrate that lincRNAs show strong signatures of evolutionary conservation and pinpoint the specific regions under purifying selection. Finally, we also identify hundreds of large multi-exonic anti-sense transcripts, which show substantially lower conservation than the lincRNAs. Our results open the way to direct experimental manipulation of thousands of non-coding RNAs, and demonstrate the power of *ab initio* reconstruction to provide a comprehensive picture of mammalian transcriptomes.

INTRODUCTION

A critical task in understanding mammalian biology is defining a precise map of all the transcripts encoded in a genome. While much is known about protein-coding genes in mammals, relatively little is known about non-coding RNA (ncRNA) genes. Recent studies have suggested that the mammalian genome encodes many thousands of large ncRNA genes[1-3], including a class of large intergenic ncRNAs (lincRNAs)[4]. Recently, we used a chromatin signature, combining Histone 3 Lysine 4 tri-methylation modifications (H3K4me3) that mark the promoter region and Histone 3 Lysine 36 tri-methylation modifications (H3K36me3) that mark the entire transcribed region, to discover the genomic regions encoding ~1600 lincRNAs in four mouse cell types (mouse embryonic stem cells, embryonic fibroblast, lung fibroblasts, and neural progenitor cells)[4], and ~3300 lincRNAs across 6 human cell types[5].

Defining the complete gene structure of these lincRNAs is a pre-requisite for both experimental and computational studies of their function, including over-expression and knockdown experiments, site-directed mutagenesis, and analysis of functional sequence features and conservation. In our previous studies, we gained initial insights into the gene structure of lincRNAs by hybridizing total RNA to tiling microarrays defined across the K4-K36 region. This provided a coarse list of putative exonic locations and suggested that lincRNAs are likely to be multi-exonic, spliced transcripts. However, due to the limited resolution of tiling arrays, the precise gene structures of these lincRNAs – including 5’ and 3’ ends, intron-exon boundaries, and connectivity between different exons – have remained unclear.

Advances in massively-parallel cDNA sequencing (RNA-Seq) have opened the way to unbiased and efficient assays of the transcriptome of any mammalian cell[6],[7],[8,9]. In principle, RNA-Seq allows the identification of all expressed transcripts, both protein-coding and non-coding. Recent studies in mouse and human cells have mostly focused on using RNA-Seq to study *known* genes —for example, to quantify their expression level[6,8], assess the level of alternative splicing between known splice junctions[7,10], and identify fusions between known genes in cancers[11]. However, these studies have critically depended on existing annotated genes, and have focused on understanding variability within known transcripts. They were thus of limited utility for discovering the complete gene structure of large numbers of non-coding transcripts, such as lincRNAs.

An alternative strategy is to use an *ab initio* reconstruction approach[9,12-14] to learn the complete transcriptome of an individual sample from only the *unannotated* genome sequence and millions of relatively short sequence reads. A complete *ab initio* transcriptome reconstruction of a sample will (**1**) identify all exons within the transcriptome; (**2**) enumerate all the splicing events that connect these exons; (**3**) combine these connected sequences into complete transcriptional units; (**4**) determine all isoforms of these transcripts, including alternative 5' and 3' ends, and (**5**) discover novel transcriptional units. A successful *ab initio* method should be applicable to large and complex genomes such as in mammals, and should be able to reconstruct transcripts of variable sizes (short and long), expression levels (high and low) and protein-coding capacity (coding and non-coding). When applied to diverse cell types, *ab initio* reconstruction can thus render a comprehensive and unbiased picture of transcriptome variation including novel alternative splicing events, variation in existing annotations, and previously unknown genes.

Despite early successes in yeast[9], *ab initio* reconstruction of a mammalian transcriptome has remained an elusive and substantial computational challenge. There has been important recent progress in mammals, including (**1**) efficient gapped aligners (*e.g.*, TopHat) that can map short reads that span splice junctions ('spliced reads')[13]; (**2**) use of such gapped alignments to identify novel splicing events[9,13]; (**3**) exon identification methods that can be used in principle to piece together transcripts[14]; and (**4**) direct genome-independent assembly of the unmapped reads to create sequence contigs (*e.g.*, Abyss[12]). Each of these methods provides an important component towards reconstruction, but none can reconstruct the complete transcriptome of a mammalian cell. The *ab initio* approach applied in yeast[9] does not scale to mammalian genomes which are significantly larger, contain mostly multi-exonic transcripts, and have substantial and complex alternative splicing. Exon identification methods[14] treat each exon in isolation (as a 'transcribed region') but do not handle splicing directly. Thus, they are underpowered to identify short or low-expressed exons, can miss whole genes despite strong cumulative evidence, generate disconnected exons rather than (alternatively spliced) transcripts, and cannot resolve anti-sense from overlapping sense transcription. Finally, approaches for genome-independent, assembly-based reconstruction such as Abyss[12] (that assemble the reads directly without mapping to the genome) can currently only be applied to transcripts with immense coverage, and are hence partial and biased in practice.

Here, we present Scripture, a comprehensive method for *ab initio* reconstruction of the transcriptome of a mammalian cell, and apply it to transcriptomes of mouse embryonic stem (mES) cells, neural progenitor cells (NPCs), and mouse lung fibroblasts (MLFs) to discover the complete gene structures of 9738 protein coding genes, 1868 lincRNA genes (1073 from previously undescribed loci), and 446 large, multi-exonic anti-sense genes. Our approach uses gapped alignments of spliced reads followed by reconstruction of the reads into statistically significant transcript structures. For example, when applied to mES cells, we correctly identify at high confidence the complete annotated full-length gene structures from 5' to 3' end at single nucleotide resolution for 78% of mES expressed genes. Of the remainder that are expressed in ESC, we successfully build the annotated 5' start for 20% of genes and the annotated 3' end for 71% of genes. For many of the expressed genes, we reconstructed structures that differ from the reported annotation but we demonstrate that many of these alternative structures are supported by independent experimental data. The three reconstructed transcriptomes reveal substantial variation between cell types, including thousands of novel 5'-start sites for protein coding genes, hundreds of alternative 3'-ends, and thousands of additional coding exons spliced onto annotated protein-coding genes.

We also discover the gene structure and expression level of over 2000 non-coding transcripts. These include hundreds of transcripts from previously identified loci encoding mouse lincRNAs, more than a thousand additional previously unknown lincRNAs with similar properties, and hundreds of multi-exonic antisense ncRNAs transcribed from the opposite strand of an overlapping protein-coding gene locus. These detailed gene structures allow us to identify distinct alternatively spliced isoforms of lincRNAs in different cell types, definitively show that they have no significant coding potential, and show that they are evolutionary conserved, including identifying for the first time the specific regions of conserved sequence within these lincRNAs. These results open the way to direct experimental manipulation of this new class of genes. Finally, our sensitive approach can correctly determine the transcribed strand, allowing us to detect gene structures for hundreds of novel multi-exonic antisense ncRNAs transcribed from the opposite strand of an overlapping protein-coding gene locus. Our results highlight the power of RNA-seq along with an *ab initio* reconstruction to render a comprehensive picture of cell specific transcriptomes, and to identify novel genes and variation encoded in mammalian genomes.

RESULTS

RNA-Seq libraries

We used massively parallel (Illumina) sequencing to sequence a cDNA library from polyA(+) mRNA from mES, NPCs and MLFs cells. We used a cDNA preparation procedure that combines a random priming step with a fragmentsation step[8,9,15] and results in fragments of ~700 nucleotides in size (**Methods**). We previously found[9,15] that this protocol provides relatively uniform coverage of the whole transcript, thus assisting in *ab initio* reconstruction. We sequenced each library on three lanes of the Illumina Genome Analyzer. For example, for the mES library, we generated a total of 152 million paired-end reads of 76 bases in length. Using a gapped aligner[13], 93 million of these reads were uniquely alignable to the genome, providing 497Mb of aligned bases, at an average 262X fold coverage of the 38Mb within known protein coding genes expressed in mES cells. We obtained similar results for the NPCs and MLFs libraries (**Methods**).

Most uniquely aligned reads are consistent with the position of annotated protein coding genes and measured expression levels, supporting the quality of our dataset. For example, in mES cells, 76% of these reads map within the exonic regions of known protein-coding genes, 9% are in introns of known protein coding genes, and 15% map in intergenic regions. Furthermore, less than 0.001% of paired reads extend across multiple known protein-coding loci, indicating lack of chimeras. Finally, we found a strong correlation between expression levels of protein-coding genes as measured by the number of sequence reads obtained here compared to Affymetrix expression arrays (*r*=0.88 for all genes).

Scripture: a statistical method for *ab initio* reconstruction of a mammalian transcriptome

We next developed Scripture, a genome-guided method to reconstruct the transcriptome using only an RNA-Seq dataset and an (unannotated) reference genome sequence. Scripture consists of six steps (**Fig. 1**): (**1**) We use reads uniquely aligned to the genome, including those with gapped alignments spanning exon-exon junctions ('aligned spliced reads')[13] (**Fig. 1c**); (**2**) From the aligned spliced reads, we construct a *connectivity graph* representing spliced connections between base pairs in the genome (**Fig. 1d**); (**3**) Using all spliced and non-spliced (contiguous) read data, we use a statistical segmentation approach[4] to traverse the connectivity graph and identify significant paths (**Fig. 1e**); (**4**) From the paths, we construct a *transcript graph* connecting each exon in the transcript (**Fig. 1f**); (**5**) We augment the transcript graph with connections based on paired-end reads and their distance constraints, allowing us to join transcripts or remove unlikely isoforms (**Fig. 1g**); and (**6**) We generate a catalogue of transcripts defined by the transcript graph. We discuss each of these steps in detail below.

We first map our reads to the genome, using a gapped aligner[13] that efficiently handles reads that span splicing junctions (**Fig. 1a**). This step is critical since ~30% of 76 base reads are expected on average to span an exon-exon junction (**Methods**). Furthermore, 'spliced' reads provide direct information on the location of splice junctions within the transcript.

We next use only the spliced reads to infer a *connectivity graph* across the genome, where each base in the genome is connected to those bases in the genome that are its immediate neighbors either in the genomic sequence itself or within a spliced read (**Fig. 1d**). Furthermore, we use agreement with splicing motifs at each putative junction in the graph to orient the connection (edge) in the connectivity graph[9,13] (**Fig. 1d, Methods**).

To infer transcripts, we apply a statistical segmentation approach that identifies significantly enriched paths in the connectivity graph using both spliced and non-spliced reads (**Fig 1e**). Briefly, our segmentation approach identifies regions of mapped read enrichment compared to the genomic background. This is done by scoring a sliding window using a test statistic for each region, computing a threshold for genome-wide significance, and using the significant windows to define intervals (**Methods**). To define intervals, we scan short windows to identify consecutive coverage blocks that have a read coverage scoring above the genome-wide significance threshold we computed. This approach is based on our successful method for identification of chromatin modified regions in genomes[4], but is applied here to the *connectivity graph* rather than to the linear genome.

The result is a set of statistically significant directed *transcript graphs* (**Fig 1f**), each representing one or more splice isoforms of a transcript. Each node in a transcript graph is an exon and each edge is a splice junction. A path through the graph from an exon with no incoming edges (first exon) to an exon with no outgoing edges (last exon) represents one isoform of the gene. Since each graph is directed, all multi-exonic paths are oriented (*i.e.* strand-specific, **Fig. 1e**). Alternative spliced isoforms are identified by considering all possible paths in the transcript graph; since this number may be large and represent spurious paths, we refine it in the next step.

Paired-end reads in transcriptome reconstruction and resolution of alternative spliced isoforms

Paired-end information, consisting of reads that came from the two ends of the sequenced RNA fragment, can provide two kinds of valuable additional information in the reconstruction.

First, the presence of paired-ends linking two regions shows that they appear in the same transcript; such a connection might not otherwise be apparent because low expression levels or non-alignable sequence might prevent a continuous chain of overlapping sequence reads (spliced or unspliced) across the transcript. We thus augment the transcript graphs with paired-end information, where available, to (indirectly) link nodes in the graph. We use these indirect links (**Fig. 1g**) to add edges between disconnected graphs, add internal nodes (exons) that might have been missed within a path (transcript), and add extra support for existing edges. This refines the structure of our transcripts and increases our confidence in them, especially in lowly-expressed transcripts that are more likely to have coverage gaps.

Second, the distribution of library insert sizes constrains the distance between the paired end reads; these distance constraints can be used to infer the relative likelihood of some potential transcripts (for example, those in which the paired ends would be much closer or much further than typical). We infer the distribution of insert sizes for a given library from the position of read pairs on transcripts from those genes for which there is only a single transcript model (i.e., no detectable alternative splicing) (**Methods**). Indeed, for our ES library, for example, this estimated distribution matches extremely well with the experimentally determined size range (data not shown). We use this distribution to assign likelihoods to each read pair occurring within a transcript graph, and then remove transcripts that are too unlikely (**Methods**).

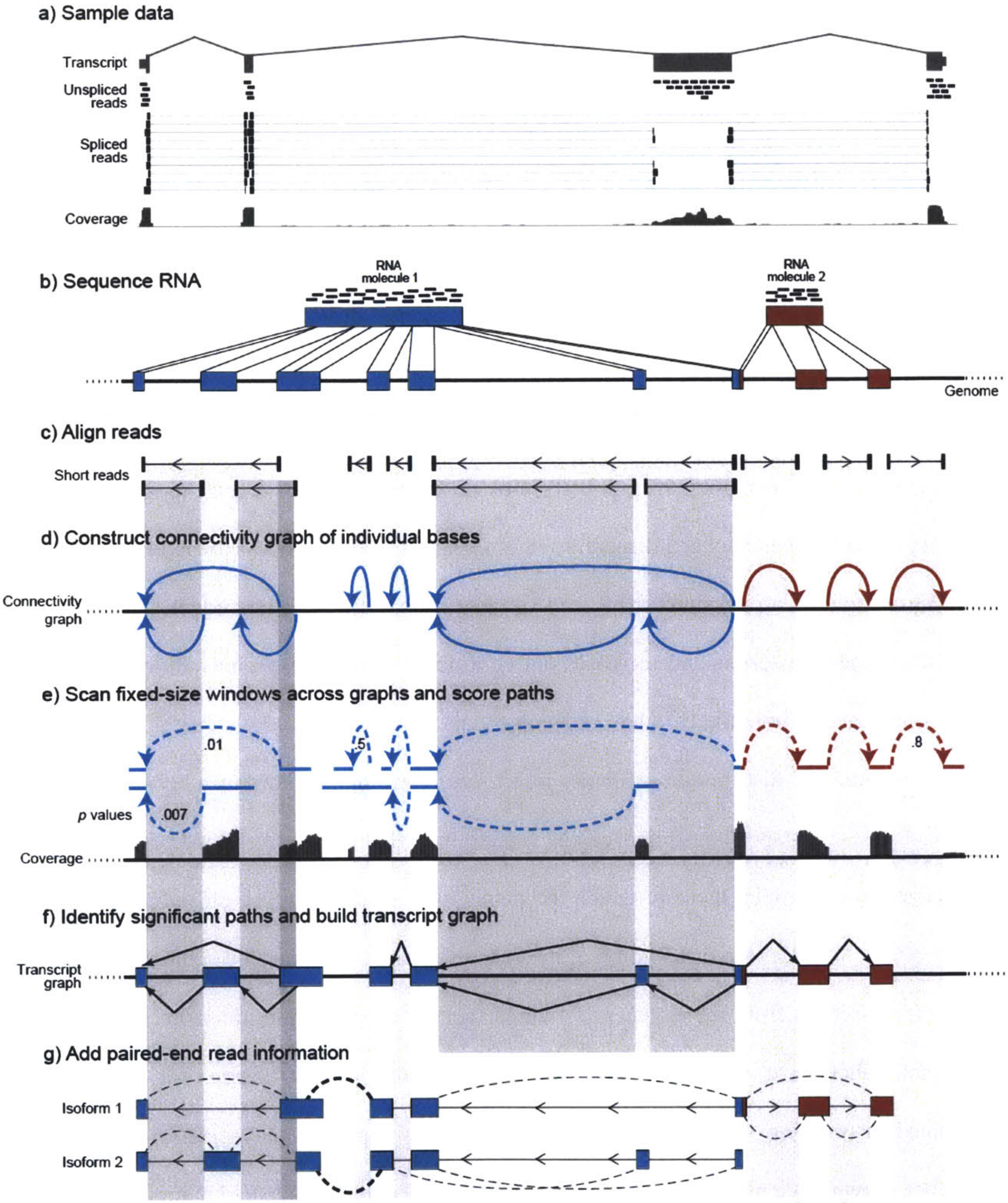


Figure 1: Scripture: a method for *ab initio* transcriptome reconstruction from RNA-Seq data. (a) Spliced and unspliced reads. Shown is a typical expressed 4-exon gene (1500032D16Rik, top, exons: grey boxes) with coverage from different type of reads. Unspliced reads (black bars) fall within a single exon, whereas splice reads (dumbbells) span exon-exon

junctions (thin horizontal lines connect the alignment of a read to the exons it spans). The coverage track (bottom) shows the aggregate coverage of both spliced and unspliced reads. **(b-g) A schematic description of Scripture. (b) A cartoon example.** Reads (black bars) originate from sequencing a contiguous RNA molecule. Shown are transcripts from two different genes (blue and red boxes), one with seven exons (blue boxes) and one with three exons (red boxes), which are adjacent in the genome (black line). The grayscale vertical shading in subsequent panels is shown for visual tracking. (**c**) **Spliced reads.** Scripture is initiated with a genome sequence and spliced aligned reads (dumbbells) with gaps in their alignment (thin horizontal lines). Scripture uses splice site information to orient splice reads (arrow heads). **(d) Connectivity graph construction**. Scripture builds a connectivity graph by drawing an edge (curved arrow) between any two bases that are connected by a spliced read gap. (Edges are color coded to relate to the original RNA and eventual transcript). (**e**) **Path scoring.** Scripture scans the graph with fixed-sized windows and uses coverage from all reads (spliced and non-spliced, bottom track) to score each path for significance (p-values shown as edge labels). (**f**) **Transcript graph construction.** Scripture merges all significant windows and uses the connectivity graph to give significant segments a graph structure (three graphs in this example). (**g**) **Refinement with paired-end data**. Scripture uses paired-end (dashed curved lines) to join previously disconnected graphs (Gene 1, bold dashed line), find break point regions within contiguous segments (*e.g.* no dashed lines between Gene 1 and 2), and eliminate isoforms that result in paired-end reads mapping at a distance with low likelihood.

Correct reconstruction of full-length gene structures at single-base resolution for the majority of protein-coding genes

We first applied Scripture to our mouse ES RNA-Seq dataset, and estimated the method's sensitivity and accuracy by comparing our reconstructions to known annotations of protein-coding genes. Scripture identified 15,246 nonoverlapping, multi-exonic transcript graphs which correspond to 16,122 known genes from the NCBI RefSeq project[16]. The average number of transcript graphs per gene is 1.3, with 76.6% of reconstructed genes covered by a single graph (single transcript in the reconstruction, though possibly with multiple paths for different splice isoforms) and 18.5% covered by two transcript graphs (the transcript is split to two separate pieces in the reconstruction).

For ~78% (10,355) of the expressed transcripts, Scripture reconstructed the full-length structure of the longest known splice isoform (from 5' to 3' end of the gene, including all exons and splice junctions) at single base resolution (*e.g.* **Fig. 2a**). All of our reconstructed transcripts for known multi-exonic transcripts also had the correct orientation (strand). In particular, Scripture was able to correctly reconstruct known genes that overlap one another on opposite strands (*e.g.* **Fig. 2a**).

Complete transcript structures are recovered across a very broad range of expression levels (**Fig. 2b,c**). For example, Scripture accurately reconstructs the full-length transcript of ~73% of the known protein-coding genes at the second quintile of expression (68.4X mean coverage), ~88% of genes from the third quintile (144X fold coverage), and 94% of the genes from the top quintile. Similarly, the average proportion of bases constructed for each transcript (considering both full and partial reconstructions) was high (**Fig. 2c**). For example, even for the bottom 5% of expressed genes (15X mean coverage), where we reconstruct the full length gene structure for only 19% of the transcripts (**Fig. 2b**), we do recover on average 62% of each of these transcripts' bases (**Fig. 2c**). This demonstrates the power of Scripture to reconstruct a substantial portion of lowly expressed transcripts.

In the remaining ~22% (3007 genes) of cases that do not correspond to annotated full-length transcripts, 71% (15% of the total) match at the annotated 3'-end, 20% (4% of the total) match at the annotated 5' start; and the remainder (3% of total reconstructed transcripts) match at neither end. Importantly, we show below that many of these variant transcripts likely represent true alternative isoforms.

We obtained similar results in the other two cell types, with 20,854 transcript graphs that correspond to 12,854 known genes in NPCs and 22,216 transcripts graphs corresponding to 13,257 known genes in MLFs. Taken together, our analysis shows that Scripture can accurately reconstruct full-length gene structures at nucleotide resolution for the majority of expressed genes. Since the minority of genes that are not reconstructed are those with low expression levels, it should be possible to reconstruct most of these genes simply by generating additional RNA-Seq data.

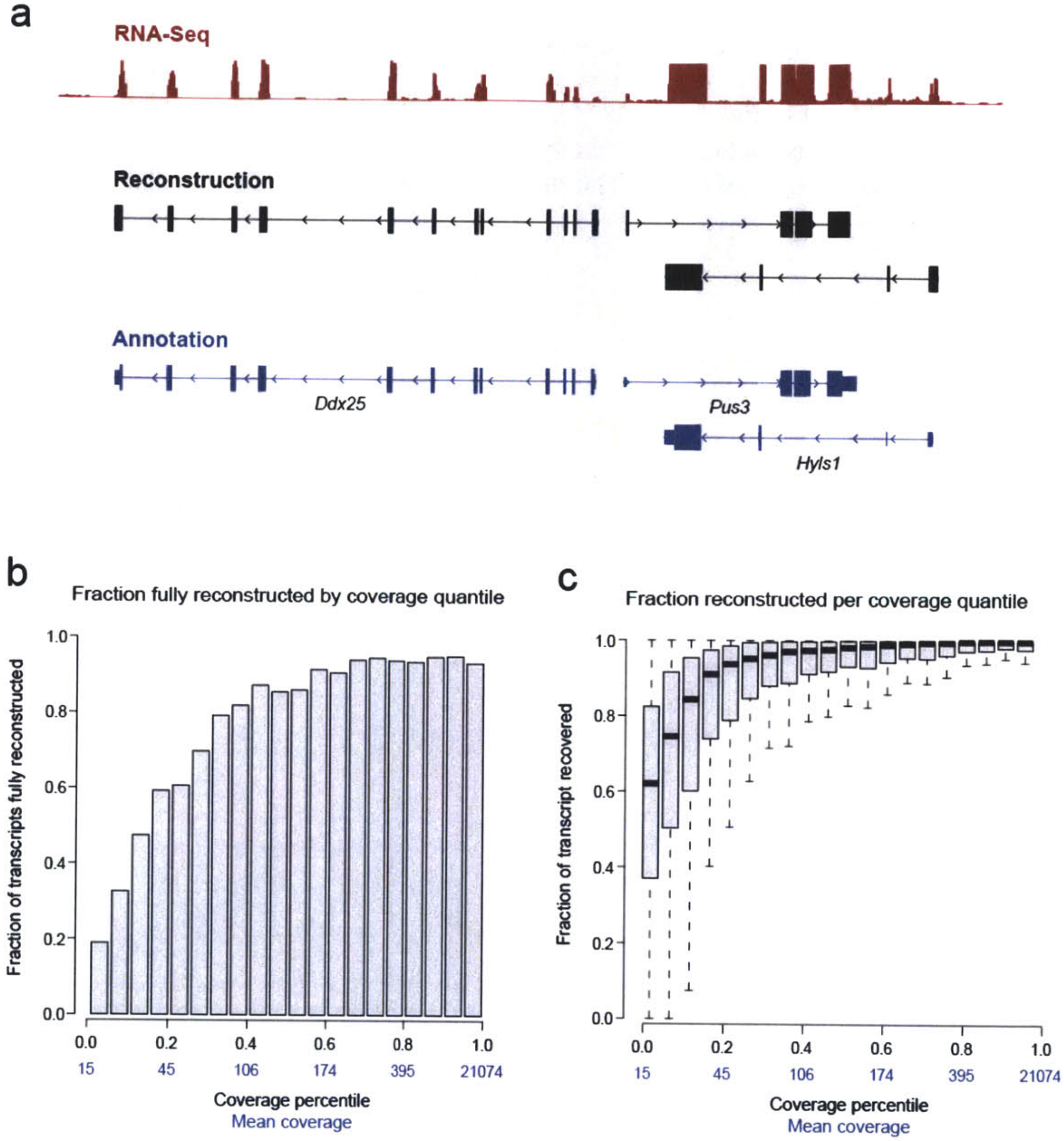


Figure 2: Scripture correctly reconstructs full length transcripts for the majority of annotated protein coding genes. (a) A typical Scripture reconstruction on mouse chr9. Top (red) – RNA-Seq read coverage (from both non-spliced and spliced reads); middle (black) – three transcripts reconstructed by Scripture, including exons (black boxes) and orientation (arrow heads); bottom (blue) –RefSeq annotations for this region. All three transcripts are fully reconstructed from 5' to 3' ends capturing all internal exons; notice that Scripture correctly reconstructed the overlapping transcripts Pus3 and Hyls1. (**b**) **Fraction of genes fully reconstructed in different expression quantiles (5% increments) in ESC.** Each bar represents a 5% quantile of read coverage for genes expressed (mean read coverage is noted in blue). The height of each bar is the fraction of genes in that quantile that were fully reconstructed. For example, ~20% of the transcripts at the bottom 5% of expression levels are fully reconstructed; ~94% of the genes at the top 95% of expression are fully reconstructed. (c) **Portion of gene**

length reconstructed in different expression quantiles in ESC. Shown is a box plot of the portion of each transcript's length that was covered by a Scripture reconstruction in each 5% coverage quantile. The black line in each box is at the median, the rectangle spans the 25% and 75% coverage quantiles; the whiskers depict the annotations in the quantile most and least covered by our reconstruction. For example, at the bottom 5% of expression, Scripture reconstruct a median length of 60% of the full length transcript.

Novel transcriptome variations in annotated protein-coding genes

Given that the vast majority of the significant *ab initio* reconstructions of protein-coding genes are extremely accurate, we next investigated the differences between the reconstructed mESC transcriptome and the known gene annotations. We focused on transcripts in mESC with (**i**) novel 5' start sites; (**ii**) novel 3' ends; and (**iii**) previously unidentified exons within the transcriptional units of known protein-coding genes. In each category, we first discuss below the reconstructed transcripts in mESC and then consider the results for the NPCs and MLFs.

(i) Alternative 5' start sites in mouse ES cells are supported by H3K4me3 marks

We found 1804 transcripts in mESCs that match the annotated 3'-end but have an alternative 5' start site. We distinguish between *internal* alternative 5' start sites (1397 cases, **Fig. 3a**) that occur downstream of the annotated start, and *external* alternative 5' start sites (407 cases, **Fig. 3b**) that occur upstream of the annotated start. These novel 5 start sites are derived from an additional exon (coding or UTR) not overlapping the annotated first exon.

We sought independent experimental support for the accuracy of 1397 internal 5'-start sites by examining the location of H3K4me3, a mark of the promoter region of genes[17]. We found that 1260 (90%) of the internal 5'-starts contain H3K4m3 marks, consistent with being actively transcribed promoters. Notably, in 63% of cases with an internal 5'-start site, our reconstructed transcriptome contained no isoform starting at the annotated 5'-start site.

For the 407 transcripts with external 5' start sites, we found that 75% are marked with H3K4me3. These alternative start sites are on average 21Kb upstream of the annotated site (48Kb SD), substantially revising the annotated promoters. For 66% (214 transcripts) of these cases, our reconstructions contain only the novel 5' start site and not the annotated start site.

We observed similar results from NPCs and MLFs cells (**Fig, 3a,b**) Altogether, we identified 2502 internal 5' start sites (2193 are supported by K4me3 in their respective tissues) in at least one cell type (1870 are specific to one cell type, 497 are present in two cell types, and 135 in all three), and 635 external 5' start sites in at least one cell type (396 are specific to one cell type, 149 are present in two cell types, and 90 in all three). In particular, 44% of these novel 5' ends are unique to the ESC state and are not present in either MLFs or NPCs.

(ii) Alternative 3' UTRs used in mES cells supported by polyadenylation motifs

Among our reconstructed transcripts in mES cells, there are 551 (~4%) cases with an alternative 3'-end downstream of the annotated 3'-end (mean distance 30 Kb ± 67Kb SD downstream, *e.g.* **Fig. 3c**). Of these, 275 (~50%) have evidence of a polyadenylation motif within the novel 3' exon, which is only slightly lower than for annotated 3' ends (60%), and much higher than for randomly chosen size-matched exons (6%). The frequency of the polyadenylation motif supports the accuracy of the reconstruction.

Accurately detecting upstream (early) termination is more challenging, because it is difficult to distinguish between early termination and incomplete reconstruction, especially in the case of genes with relatively low expression levels and sequence coverage. We therefore designated novel 3' ends only in those cases that did not overlap any of the known exons in the annotated transcript and required that all considered transcripts contain complete 5' start sites (further reducing the likelihood of incomplete reconstruction). We identified 759 transcripts with upstream 3'-ends in mESCs (**Fig. 3d**); the vast majority of them (90%) also had isoforms that contained the annotated 3' end. Of these upstream 3'-ends, 44% contain a poly-adenylation motif. This is lower than the ~60% for annotated 3'-ends, but much higher than the 6% for other size-matched exonic regions; it thus supports the biological relevance of many of the novel upstream 3'-ends. We note that the isoform with alternative 3' internal end tends to be expressed at a somewhat lower level than the isoform with the annotated 3' end (at a median 1.5 fold, $p<$ 0.002, paired t-test).

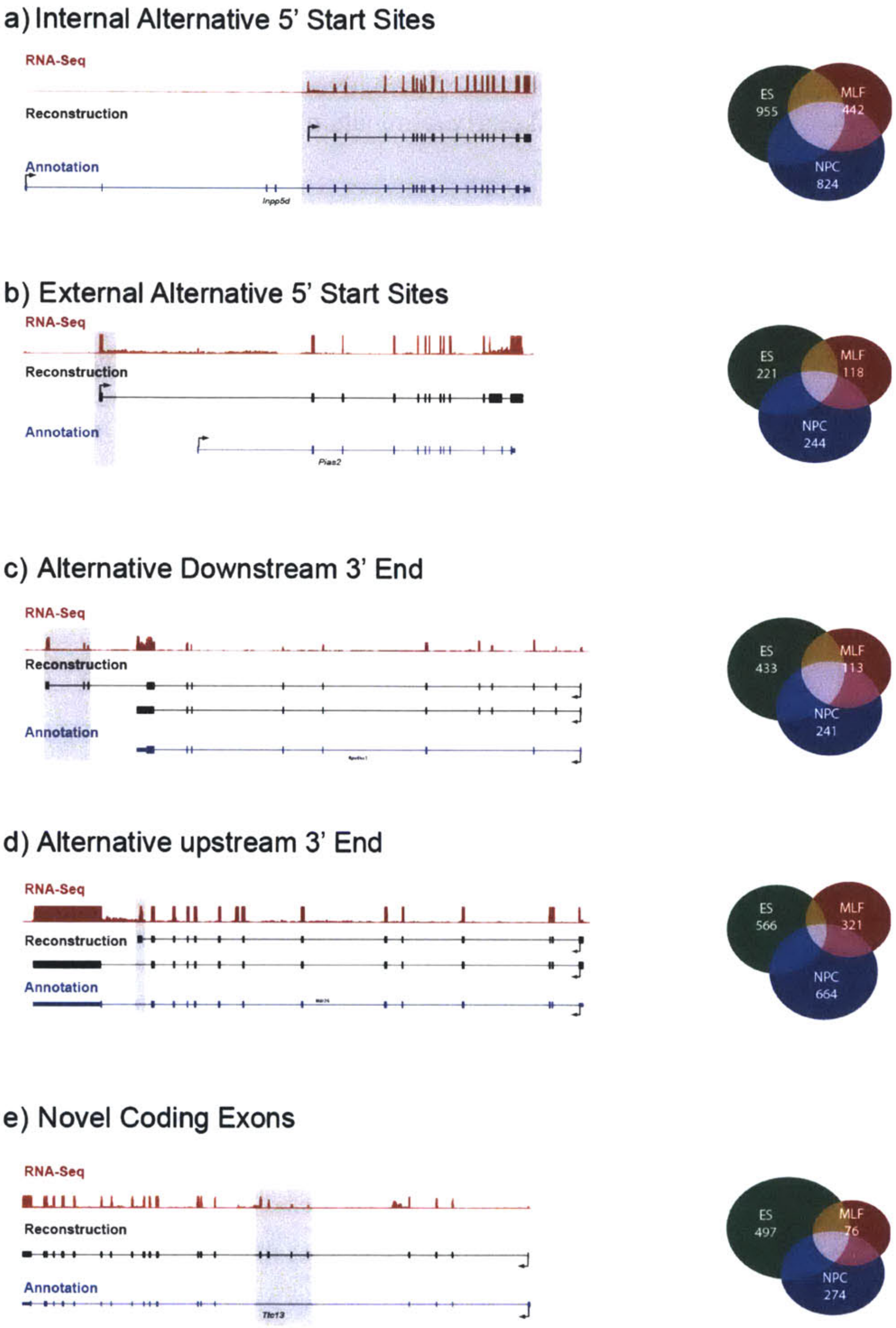
We observed similar results for NPCs and MLFs cells (**Fig. 3c,d**). Altogether, we identified 868 downstream 3' ends in at least one cell type (635 are specific to one cell type, 144 are present in two cell types, and 93 in all three) and 1609 upstream 3' ends in at least one cell type (1221 are specific to one cell type, 318 are present in two cell types, and 70 in all three).

(iii) 903 additional coding exons within known gene structures are highly conserved and preserve ORFs

We found 534 high confidence transcripts in mESC with at least one additional previously unannotated internal coding exon (neither first nor last) spliced into annotated protein-coding transcripts (**Fig. 3e**). These transcripts contained 591 novel internal exons, ranging in length from 6bp to 3.5Kb (mean length of 217bp ± 388bp SD, comparable to annotated exons). Of these additional exons, 322 (60%) are present in all versions of the reconstructed transcript in mESC, whereas the remaining additional exons are part of some transcript isoforms but not others within the same cell type.

The vast majority (83%) of these novel exons retain the reading frame of the transcript, consistent with their being novel protein-coding exons. Moreover, the vast majority of these novel coding exons are as highly conserved as known coding exons, further supporting their functionality. We tested for the presence of the novel exons within 5 transcripts, using RT-PCR followed by Sanger sequencing (**Methods**), and validated all of these tested cases.

We observed similar results in MLFs (194 transcripts, with 212 exons) and NPCs (300 transcripts, 309 exons) (**Fig. 3e**). In ~70% of cases, the novel exons are present in all versions of the reconstructed transcript within a cell type. Altogether, we identified 903 novel internal exons in at least one cell type (739 are specific to one cell type, 128 are present in two cell types, and 36 in all three, **Fig. 3e**). The vast majorities of these retain the ORF and show clear evolutionary conservation.



a) Internal Alternative 5' Start Sites

b) External Alternative 5' Start Sites

c) Alternative Downstream 3' End

d) Alternative upstream 3' End

e) Novel Coding Exons

Figure 3: Alternative 5' ends, 3' ends and novel coding exons in transcripts reconstructed by Scripture. Shown are representative examples (tracks, left) and summary counts (Venn diagrams, right) of five categories of variations discovered in Scripture transcripts compared to the known annotations. In each representative example, shown is the coverage by RNA-Seq reads (top track, red), the reconstructed annotation (middle track, black), and the known annotation (bottom track, blue). The novel regions in the reconstruction are marked by gray

shading. In each proportional Venn diagram we show the number of transcripts in this class in each cell type (ESC – green, NPC – blue, MLF – red) and their overlap. **(a)** Internal alternative 5' start; **(b)** External alternative 5' start; **(c)** Alternative downstream 3' end (extended termination); **(d)** Alternative upstream 3' end (early termination); **(e)** Novel coding exons.

Discovery of the complete gene structures of hundreds of lincRNA loci

We next turned to identifying the gene structures of transcripts expressed from known lincRNAs loci. In our previous work, we had identified 317 lincRNAs based on K4-K36 domains in mES cells and using conservative filtering criteria[4]. In the mES RNA-Seq data, we were able to reconstruct multi-exonic gene structures for 250 (78.8%) of these 317 loci (*e.g.* **Fig. 4a**). (This is comparable to the proportion (78.5%) reconstructed for protein-coding genes with K4-K36 domains in mES cells.) We accurately reconstructed 88% (160/183) of mES lincRNAs for which we previously identified an RNA hybridization signal from tiling microarrays. In 11 cases we identified a single reconstructed lincRNA structure that spans across multiple K4-K36 regions and in 55 cases we identified a single K4-K36 locus that corresponded to two distinct lincRNA gene structures in opposite orientations. These discrepancies are likely due to the lower resolution of our chromatin maps compared with the base-pair resolution of our transcript maps.

We failed to reconstruct transcripts for the remaining 67 lincRNAs that had been previously identified based on K4-K36 domains in mESC. These lincRNA genes may not have been reconstructed because they are either expressed at lower levels, are single exons, are false positives of our chromatin signature, or are false negatives of the reconstruction approach. For example, 30 of our previously identified K4-K36 domains are reconstructed as likely connected to a new isoform of a neighbouring protein coding transcript and thus are no longer counted as lincRNAs in our refined catalogue. The principal reason we miss the remaining 37 K4-K36 domains is low expression levels. Nonetheless, 67% of these remaining lincRNA loci (25 lincRNAs) are significantly enriched for reads (average of 0.76 reads/bp compared to expected 0.01 reads/bp, nominal $p<0.001$, random permutation of reads against size matched random regions). This is consistent with these loci being transcribed. With higher coverage, it should be possible to reconstruct them.

The reconstructed lincRNA transcripts in mESCs have 3.7 exons ± 2.1 SD on average, an average exon size of 350 ± 465 bp, and an average mature spliced size of 3.2 ± 1.7 Kb (compared to 9.7 ± 9.5 exons, exon length of 291 ± 648 bp, total length of 2.9Kb ± 2.3 Kb for protein coding genes). Since lincRNAs contain canonical splice acceptor and splice donor sites at their exon-intron boundaries, we can use these to identify the strand information for >99% of reconstructed lincRNAs. The predicted strand is consistent with that inferred from the location of K4me3 modification, which marks the 5' end, and with the orientation determined from a strand-specific RNA-Seq library which we generated independently (below, **Methods**). At least 82% of lincRNAs in mESCs (205 lincRNAs) likely represent 5'-complete transcripts, as indicated by overlap between the reconstructed 5'-ends of lincRNAs and a site of H3K4me3 modification. Furthermore, the majority of the lincRNAs appear to be 3'-complete as well (since ~50% contain a polyadenylation motif, comparable to 60% for protein-coding genes and far above background of 6%).

We had similar success in reconstructing lincRNA gene structures for K4-K36 lincRNA loci in MLFs and NPCs. We identified 211 out of 264 multi-exonic lincRNAs in MLFs and 202 out of 245 in NPCs. 69% of lincRNAs in MLFs (145 lincRNAs) and 81% of lincRNAs in NPCs (163 lincRNAs) likely represent 5'-complete transcripts based on sites of H3K4me3 modification; 18% of lincRNAs in MLFs (37 lincRNAs) and 37% in NPCs (75 lincRNAs) have detectable 3' polyadenylation sites. In addition to these lincRNAs, we successfully reconstructed another 103 lincRNAs previously identified only in mouse embryonic fibroblasts but which were now reconstructed in at least one of the other three cell types (**Methods**). Altogether, we identified gene structures for 567 previously defined lincRNA loci in at least one of the three cell types (78% of those previously defined in these 3 cell type; 56% of those present in the previous catalogue in any cell type). 312 of the 567 lincRNAs are specific to one cell type, 174 are present in two cell types, and 80 are in all three.

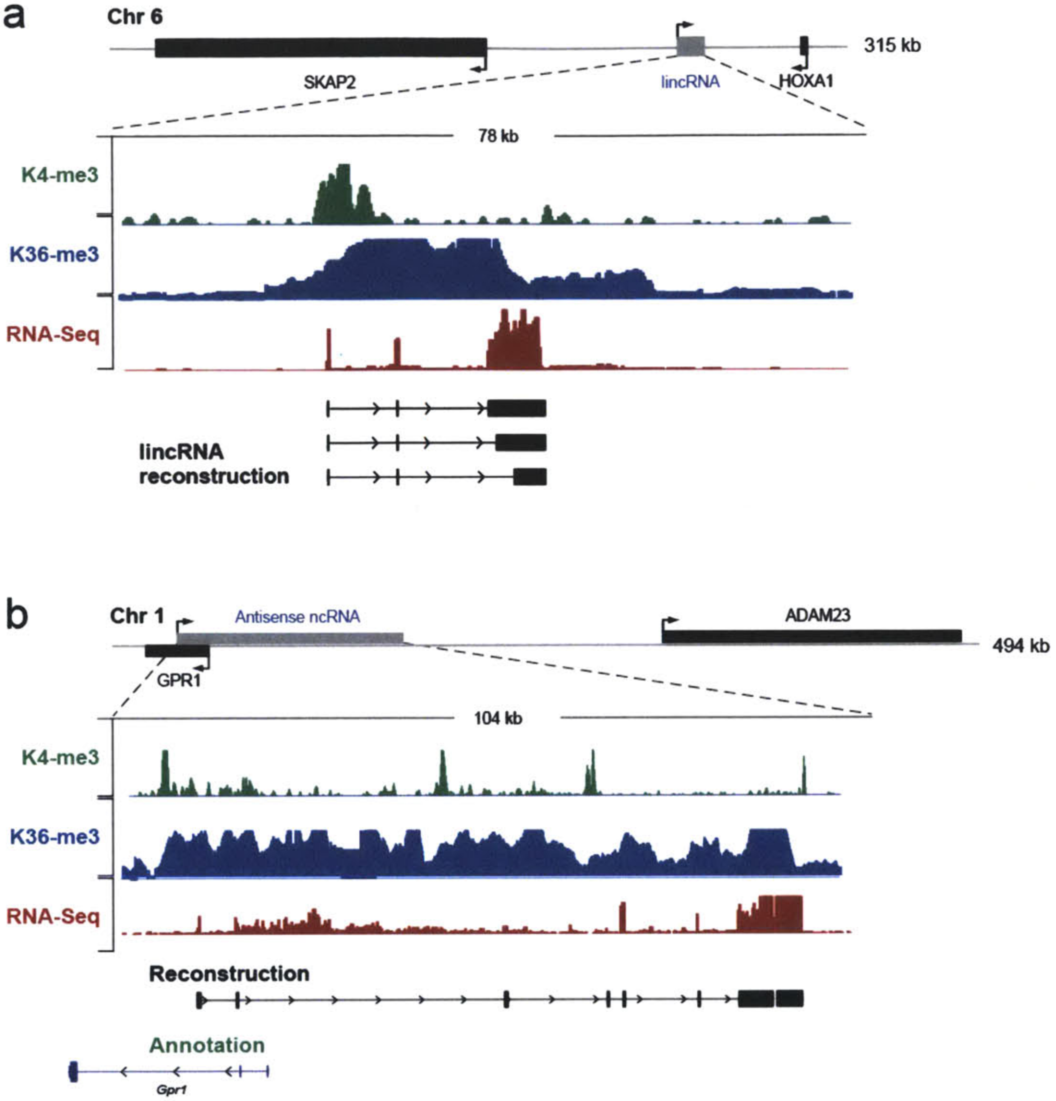


Figure 4: Non-coding transcripts reconstructed by Scripture. (a) A representative example of a lincRNA expressed in ESC. Top panel – mouse genomic locus containing the lincRNA and its neighbouring protein coding genes. Bottom panel – zoom in on the lincRNA locus showing the coverage of H3K4me3 (green track), H3K36me3 (blue track), and RNA-Seq reads (red track) overlapping the transcribed lincRNA locus, as well as its Scripture reconstructed transcript isoforms (black). **(b)** A representative example of a multi-exonic antisense ncRNA expressed in ESC. Top panel – mouse genomic locus containing the antisense transcript. Bottom panel – zoom in on the antisense locus showing the coverage of H3K4me3 (green track), H3K36me3 (blue track), and RNA-Seq reads (red track) overlapping the transcribed antisense locus, as well as its Scripture reconstructed gene structure (black).

Additional novel lincRNAs identified across mES cells, MLFs and NPCs

In addition to previously annotated protein-coding genes, pseudogenes, and lincRNA loci, our catalog contains another 1073 reconstructed multi-exonic transcripts that map to intergenic regions in at least one cell type (591 in mESCs, 369 in MLFs, and 445 in NPCs; 846 are cell type specific, 185 in two of the three, and 63 appear in all three cell types). These represent novel intergenic transcripts. In principle, they could be either protein coding or non-coding RNAs.

The majority (88%) of the 1073 novel intergenic transcripts do not appear to encode proteins, and can be designated as novel lincRNAs, based on the Codon Substitution Frequency (CSF) scores[18,19] (**Methods**) across the mature (spliced) RNA transcript (**Fig. 5a**). Furthermore, the vast majority (~80%) of the transcripts do not contain any open reading frame (ORF) larger than 100 amino acids (**Fig. 5b**). The remaining ~12% might reflect novel protein coding genes, ambiguous calls, and/or segmental duplications of protein coding loci. When we carefully reviewed the loci to eliminate possible artifacts, we identify 66 loci that are likely to be novel protein coding genes based on their high CSF score, a large open reading frame (>200 amino acids), and very high levels of evolutionary conservation, comparable to known protein-coding genes.

We investigated why the novel lincRNA loci had not been identified in our previous study that identified lincRNAs based on the presence of K4-K36 domains in mESCs. One of the reason appears to be that our previous study imposed stringent criteria to ensure that the novel loci were well separated from known protein-coding genes – for example, requiring that a K4-K36 domain extend over at least 5 Kb and be clearly separated from the nearest known gene locus[4]. Indeed, of the novel intergenic transcripts found in ES cells, 450 (76%) had enrichment for a K4-K36 domain (a comparable proportion as for protein-coding genes) but failed to meet the other two criteria or were too weak to be identified at a genome-wide significance (without knowing their locus *a priori*). The genomic loci of the novel lincRNAs tend to be much shorter than the previously identified set (average ~3.5Kb) and have shorter transcript lengths (859bp ± 1230bp SD vs. 3.2Kb ± 2.1Kb SD, with 3.4 exons vs 3.7 exons). On average, they are also closer to neighboring protein-coding genes (20Kb ± 157Kb SD). These results underscore the increased power of RNA-Seq to confidently identify lincRNAs compared to a chromatin-based method.

Most lincRNAs are evolutionarily conserved, with 22% of bases under purifying selection

The reconstructed full-length gene structures of lincRNAs allow us to accurately assess their evolutionary sequence conservation in each exon and in small windows. To this end, we identified the orthologous sequences for each lincRNA across 29 mammals and considered the total contraction of the branch length of the evolutionary tree connecting them. Specifically, we used a constraint metric ω[4] that reflects the 'contraction' of the branch length compared to the neutral tree based on the total number of substitutions per base for random genomic regions (**Methods**). We calculated ω over the entire lincRNA transcript, as well as over individual exons. Our previous estimates of conservation had relied on approximate definitions of the exons based on hybridization to tiling arrays[4], leaving open the possibility of mis-estimation.

Based on our high resolution gene structures, the lincRNA sequences show significantly greater conservation than random genomic regions or introns (**Fig. 5c**). Their conservation level is similar to that seen for 8 known functional lincRNAs, including XIST[20], HOTAIR[21] and NRON[22], and is lower than that seen for protein-coding exons, likely reflecting a difference in the constraints acting on protein coding sequences versus lincRNAs. The results are consistent with our previous estimates of conservation[4]. Interestingly, the conservation profiles are essentially identical between the chromatin defined lincRNAs from our previous study and the novel ones identified only in this study (**Fig. 5c**), further supporting our conclusion above that they are part of the same class of functional large ncRNA genes.

We also determined the specific regions within each lincRNA that are under purifying selection and thus likely to be functional. By computing ω within short windows (**Methods**), we found that on average, 22% of the bases within the lincRNAs lie within conserved patches, which is comparable to 25% for the 8 known functional lincRNAs cited above. In contrast, 7% of intronic bases and 77% of protein coding bases lie within conserved patches. These conserved patches provide a critical starting point for functional studies, by experimental manipulation and computational analysis. For example, one such conserved patch in the XIST lincRNA has been shown to allow the ncRNA to bind to the Polycomb complex[23].

lincRNAs are expressed at comparable levels to moderately expressed protein coding genes

On average, we found that lincRNAs are expressed at readily detectable levels, albeit slightly lower than those of protein-coding genes. We estimated the expression for each of our reconstructed transcripts using RPKM (**Methods**), and found that the median expression level of the lincRNAs is approximately 3-fold lower than that of protein-coding genes (**Fig. 5d**). The distributions show substantial overlap, with ~25% of lincRNAs having expression levels higher than the median level for protein-coding genes (**Fig. 5d**).

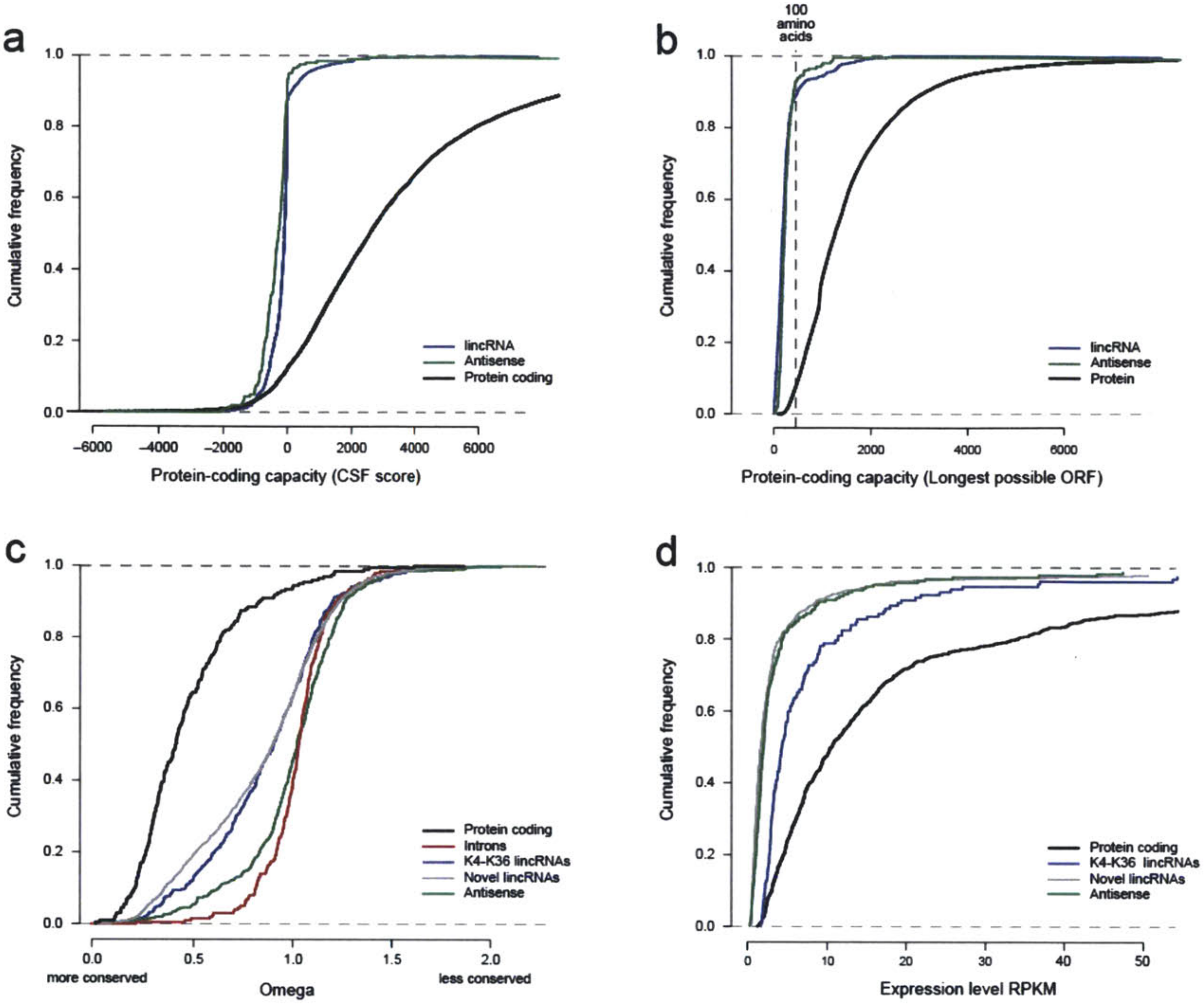


Figure 5: Protein coding capacity, conservation levels and expression of lincRNAs and multi-exonic antisense transcripts. (a-b) Coding capacity of protein coding, lincRNAs and multi-exonic antisense transcripts. Shown is the cumulative distribution of CSF scores (a) and maximal ORF length (b) for protein coding transcripts (black), lincRNAs (blue) and multi-exonic anti-sense transcripts (green). **(c)** Conservation levels for exons from protein coding transcripts, lincRNAs, multi-exonic antisense transcripts and introns. Shown is the cumulative distribution of sequence conservation across 29 mammals for exons from protein-coding exons (black), introns (red), exons from previously annotated lincRNA loci (blue), exons from newly annotated lincRNA transcripts (grey), and exons from multi-exonic antisense transcripts (green). **(d)** Expression levels of protein coding, lincRNAs and multi-exonic antisense transcripts. Shown is the cumulative distribution of expression levels (RPKM) in ESC for protein coding transcripts (black), transcripts from previously annotated lincRNA loci (blue), transcripts from newly annotated lincRNA loci (gray), and multi-exonic antisense transcripts (green).

Variations in lincRNA isoforms and expression

A substantial fraction (~41%) of the novel lincRNAs were identified in at least two of the three cell types. This is comparable to the 45% of the previously identified lincRNAs present in at least 2 out of the 3 cell types. In contrast, 80% of expressed protein coding genes are expressed across two of the three cell types. This suggests that lincRNAs are likely to be more tissue-specific than protein coding genes.

Despite the shared presence of hundreds of lincRNAs in all three cell types, there can be substantial differences in their expression levels. For example, of the 217 lincRNAs with detectable expression in all three cell types, 10% were expressed at least 3-fold higher in mESCs than in either of the other two cell types (3% were most highly expressed in MLFs, 29% were most highly expressed in NPCs). Conversely, 38% of lincRNAs were expressed at least 3-fold lower in mESCs than in MLFs and NPCs (11% in MLFs, 5% in NPCs).

A substantial portion of lincRNA loci produce alternative spliced isoforms. For example, within mESCs we identified two or more alternative splicing isoforms for 25% of lincRNA genes, comparable for 30% of protein coding genes (16% of lincRNAs in MLFs have alternative splice isoforms, and 13% in NPCs) Altogether, 27% of the 1640 lincRNA loci have evidence for alternative isoforms in any of the three cell types.

Discovery of hundreds of novel large antisense transcripts

Our transcriptome reconstruction also includes hundreds of transcripts that overlap known protein-coding gene loci but are transcribed in the opposite orientation and likely represent anti-sense transcripts. To focus on *novel* antisense transcripts, we required that a protein-coding gene has no known antisense protein-coding genes overlapping the region.

Furthermore, since Scripture used the orientation of the splice junctions to infer the transcript's orientation (strand), we required that any identified antisense transcript be multi-exonic.

Using these criteria, we identified 201 antisense multi-exonic transcripts in mESCs (*e.g.* **Fig. 4b**); these transcripts have an average 5 exons ± 8 SD per transcript, and an average transcript size of 1.7Kb ± 1.6Kb SD (min 121bp, max 15.8Kb). The antisense transcripts overlap the genomic locus of the overlapping protein coding gene by 1023 bp ± 1620bp SD (83% ± 29% SD) on average, but their overlap with the exons of the sense transcript is substantially lower at 766bp ± 1581 SD (48% ± 43% SD) on average.

We validated the reconstructed mESC anti-sense transcripts by three independent sets of experimental data. **First**, the majority of the anti-sense loci carry an H3K4me3 mark at their 5'-end, consistent with its independent and antisense transcription. Specifically, of the 164 transcripts where it is possible to detect an independent H3K4me3 mark (because the 5'-end of the anti-sense transcript does not overlap the 5'-ends of the sense gene), we found an independent H3K4me3 mark at the 5'-ends of the antisense locus in 64% of the cases (*e.g.* **Fig. 4b**). **Second**, we generated and sequenced a strand-specific library in mES cells (**Methods**) using a published RNA ligation protocol[24] and sequenced one lane of this library (17.5M reads, Illumina, **Methods**). In >90% of cases we were able to confirm the existence of a significant number of reads on the correct strand of these antisense transcripts. The remaining cases have lower average expression levels, and thus are likely less readily detected in the more limited amount of data in the strand-specific library. In no case did the strand-specific library indicate that Scripture had identified the wrong strand. **Finally**, we used RT-PCR to unique exons of the antisense transcript (**Methods**) followed by Sanger sequencing to individually confirm 5 of 5 anti-sense transcripts tested.

The majority of the anti-sense transcripts appear to be non-protein coding, by both ORF analysis and CSF scores (**Fig. 5a,b**). The novel antisense transcripts largely lack significant ORFs, with the maximum possible ORF less than 100 amino acids in >95% of cases (**Fig. 5b**). Furthermore, ~85% are likely to be non-protein-coding based on their CSF scores (**Methods, Fig. 5a**). Four of the newly identified antisense transcripts had a large, conserved open reading frame and are likely novel, previously unannotated protein coding genes.

We obtained similar results for anti-sense transcripts in MLFs and NPCs (159 and 168 multi-exonic anti-sense transcripts, respectively). Altogether, we identified 446 novel anti-sense transcripts expressed in at least one cell type (372 are cell type specific, 66 in two of the three, and 8 appear in all three cell types).

The 446 anti-sense transcripts are expressed at comparable levels to the novel lincRNAs (**Fig. 5d**), but show substantially lower sequence conservation. When we estimated the conservation of these genes by calculating the ω metric for the transcript (calculated over the portions that do not overlap protein-coding exons on the sense strand), we found that the antisense ncRNAs showed very little evolutionary conservation, suggesting that the antisense ncRNAs are a distinct class from the lincRNAs (**Fig. 5c**).

DISCUSSION

Despite the availability of the genome sequence of many mammals, a comprehensive understanding of the mammalian transcriptome has been an elusive goal. While the recent development of massively parallel RNA-Seq provides a systematic method to comprehensively sequence the transcriptome of a mammalian cell, the computational tools needed to reconstruct all full-length transcripts from the wealth of short read data were largely missing. Indeed, most methods for analyzing RNA-Seq have assumed the availability of a previously annotated catalog of protein-coding genes, and are thus not optimal for discovering novel transcript variations in known genes or for finding novel genes – especially moderately expressed, non-coding ones such as lincRNAs. A recent study proposed to overcome this limitation experimentally by using very long reads (e.g. 454 sequencing), as a scaffold for short read reconstruction[25]. This is applicable, albeit at a substantial cost, for highly expressed genes, but would require incredible depth to cover more lowly expressed ones.

Here, we present Scripture, a novel computational method to reconstruct a mammalian transcriptome with no prior knowledge of gene annotations. Scripture relies on spliced reads to traverse splice junctions and build a connectivity graph between discontiguous (spliced) segments, uses this graph and both spliced and non-spliced reads to identify transcripts as significant paths, resolves multiple splice isoforms as alternative paths, and leverages paired-end information to refine these transcripts, by removing low likelihood transcripts, breaking spurious connections and combining disconnected ones. At the heart of Scripture is a statistical segmentation approach, which provides a principled method to identify significant transcripts. By relying on a range of window sizes to drive the segmentation process, Scripture can identify both short but strongly expressed transcripts as well as much lower expressed transcripts for which there is aggregate (diffuse) evidence along the entire transcript length. This latter feature is critical for the discovery of relatively low expressed transcripts, such as many lincRNAs. Notably, Scripture does rely on a reference genome sequence (albeit unannotated), but many of its components can also be used in the development of methods for *de novo* assembly of transcripts from read data only (with no read mapping). This will be essential when a reference genome is not known, as in environmental samples, non-model organisms, and cancer.

We applied Scripture to RNA-Seq data from pluripotent ES cells and differentiated lineages and showed that we can accurately reconstruct the majority of expressed annotated protein coding genes, at a broad range of expression levels. Our results also uncover a large number of novel isoforms in the protein-coding transcriptome, including thousands of novel 5' start sites, hundreds of novel 3' ends and hundreds of novel coding exons. We provide strong evidence from chromatin modification states, polyadenylation signals and sequence conservation that these novel isoforms are biologically functional. This novel variation within known protein coding genes may play key regulatory roles. For example, most of the discovered alternative 5' start sites occur in a cell-type specific fashion, and thus involve cell-specific promoters likely with distinct regulatory mechanisms. Similarly, novel 3' ends define distinct 3'-UTRs and have the potential for distinct translational regulation (for example, through different miRNA binding sites). Finally, the novel tissue-specific protein-coding gene exons that we discover are highly conserved and preserve the ORF, suggesting that they encode cell type-specific protein products.

Going beyond protein-coding genes, we leverage Scripture's sensitivity and resolution to reconstruct the gene structures of hundreds of non-coding RNA genes in these 3 cell types, including both lincRNAs and multi-exonic antisense transcripts. It is clear that the mammalian genome encodes thousands of large ncRNAs, which are typically moderately expressed and are hence missed by traditional methods. We had previously identified many lincRNAs by searching for chromatin signatures of actively transcribed genes that are well-separated from known protein-coding genes to define novel loci, and by using RNA hybridization to tiling microarrays to approximately define the exons. In contrast, Scripture identified many additional lincRNAs (including those that are closer to protein-coding loci and those that have relatively low expression) and provided precise gene structures for each. In addition, Scripture identified hundreds of large antisense ncRNAs, which overlap protein coding gene transcripts and hence could not be effectively detected using chromatin data alone. Notably, our reconstructions (despite being based on non-strand specific libraries) resolve such overlapping sense and anti-sense transcripts at great accuracy, as subsequently validated with strand-specific RNA-Seq.

Overall, we find that there are thousands of large ncRNAs across the three cell types in our study. Our data show that the ratio of active protein-coding to non-coding genes in these cell types is roughly 10:1 (although the ratio may fall somewhat as additional cell types are studied, because ncRNAs appear to be somewhat more tissue specific). However, the total number of RNA molecules is more heavily biased toward the protein-coding fraction, with a proportion of ~100:1 coding to non-coding RNA.

Scripture identifies precise gene structures for the majority of previously found lincRNA loci (as well as for the newly discovered ones). These gene structures are a pre-requisite for further functional studies, both experimental and computational. For example, using these gene structures, we identified the specific regions within each lincRNA that are under purifying selection (conservation). These will provide a starting point for experimental manipulation (*e.g.* site directed mutagenesis) and computational investigation (*e.g.* identification of sequence or RNA structural elements).

Taken together our results highlight the power of *ab initio* reconstructions – using only read data and an unannotated reference genome – to discover novel transcriptional variation within known protein coding genes, and they provide a rich catalog of precise gene structures for novel non-coding RNAs. The next step is clearly to apply this approach to a wide range of cell types in human and mouse, to obtain a comprehensive picture of the complex and dynamic mammalian transcriptome.

METHODS

Cell culture

Mouse embryonic stem cells (V6.5) were co-cultured with irradiated MEFs (GlobalStem; GSC-6002C) on 0.2% gelatin coated plates in a culture media consisting of Knockout DMEM (Invitrogen; 10829018) containing 10% FBS (GlobalStem; GSM-6002), 1% pen-strep 1% Non-essential amino acids, 1% L-glutamine, 4ul Beta-mercaptoethanol, and .01% LIF (Millipore; ESG1106). mES cells were passaged once on gelatin without MEFs before RNA extraction. 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 described[26]. The cells uniformly express nestin and Sox2 and can differentiate into neurons, astrocytes and oligodendrocytes. Mouse lung fibroblasts (ATCC), were grown in DMEM with 10% fetal bovine serum and penicillin/streptomycin at 37°, 5% CO_2.

RNA Extraction & Library Preparation

RNA was extracted using the protocol outlined in the RNeasy kit (Qiagen). Extracts were treated with DNase (Ambion 2238). Polyadenylated RNAs was selected using Ambion's MicroPoly(A)Purist kit (AM1919M) and RNA integrity confirmed using Bioanalyzer (Agilent).

A 'regular' RNA sequencing library (non strand specific) was created as previously described[15], with the following modifications. 250 ng of polyA$^+$ RNA was fragmented by heating at 98°C for 33 minutes in 0.2 mM sodium citrate, pH 6.4 (Ambion). Fragmented RNA was mixed with 3 μg random hexamers, incubated at 70°C for 10 minutes, and placed on ice briefly before starting cDNA synthesis. First strand cDNA synthesis was performed using Superscript III (Invitrogen) for 1 hour at 55°C, and second strand using *E. coli* DNA polymerase and *E. coli* DNA ligase at

16°C for 2 hours. cDNA was eluted using Qiagen MiniElute kit with 30ul EB buffer. DNA ends were repaired using dNTPs and T4 polymerase, (NEB) followed by purification using the MiniElute kit. Adenine was added to the 3' end of the DNA fragments to allow adaptor ligation using dATP and Kelnow exonuclease (NEB; M0212S) and purified using MiniElute. Adaptors were ligated and incubated for 15 minutes at room temperature. Phenol/choloform/isoamyl alcohol (Invitrogen 15593-031) extraction followed to remove the DNA ligase. The pellet was then resuspend in 10ul EB Buffer. The sample was run on a 3% Agarose gel (Nusieve 3:1 Agarose) and a 160 - 380 base pair fragment was cut out and extracted. PCR was performed with Phusion High-Fidelity DNA Polymerase with GC buffer (New England Biolabs) and 2M Betaine (Sigma). [PCR conditions: 30 sec at 98°C, (10 sec at 98°C, 30 sec at 65°C, 30 sec at 72°C -16 cycles) 5 min at 72°C, forever at 4°C], and products were run on a poly-acrylamide gel for 60 minutes at 120 volts. The PCR products were cleaned up with Agencourt AMPure XP magnetic beads (A63880) to completely remove primers and product was submitted for Illumina sequencing.

The "strand-specific" library was created from 100 ng of polyA$^+$ RNA using the previously published RNA ligation method[24] with modifications from the manufacturer (Illumina, manuscript in preparation). The insert size was 110 to 170 bp.

RNA-Seq library sequencing

All libraries were sequenced using the Illumina Genome Analyzer (GAII). We sequenced 3 lanes for mouse ESC corresponding to 152 million reads, 2 lanes for MLFs corresponding to 161 million reads, and 2 lanes for NPCs corresponding to 180 million reads.

Alignments of reads to the genome

All reads were aligned to the mouse reference genome (NCBI 37, MM9) using the TopHat aligner[13]. Briefly, TopHat uses a two-step mapping process, first aligning all reads that map directly to the genome (with no gaps), and then attempting to map all the reads that were not aligned in the first step using gapped alignment. TopHat uses canonical and non-canonical splice sites to determine possible locations for gaps in the alignment. While all reported results rely on TopHat alignments, very similar results are obtained in practice using the BLAT algorithm[27], when allowing for gaps, and conservatively removing all gapped alignments that are aligned to less than 80% of the read and do not contain canonical or non-canonical splice sites at the locations of the gap. Since TopHat uses a global alignment strategy it is more suitable for short RNA-Seq reads and far more efficient than BLAT.

Generation of connectivity graph

Given a set of reads aligned to the genome, we first identified all spliced reads, as those whose alignment to the reference genome contains a gap. These reads and the reference genome are used to construct connectivity graphs. Each connectivity graph contains all bases from a single chromosome. The nodes in the graph are bases and the edges connect each base to the next base in the genome as well as to all bases to which it is connected through a 'spliced' read (**Fig. 1**). In the analysis presented, we defined an edge between any two bases in the chromosome that were connected by two or more spliced reads. The connectivity graph thus represents the contiguity that exists in the RNA but that is interrupted by intron sequences in the reference genome.

Identification of splice site motifs and directionality

We restricted our analysis to splice reads that mapped connecting donor/acceptor splice sites, either canonical (GT/AG) or non-canonical (GC/AG). We oriented each mapped spliced read using the orientation of the donor/acceptor sites it connected.

Construction of transcript graphs

The 'spliced' edges in the connectivity graph reflect bases that were connected in the original RNA but are not contiguous in the genome. To construct a transcript graph, we 'thread' the connectivity graph (which was constructed only from the genome and spliced reads) with the non-spliced (contiguous) reads, to provide a quantitative measure of the reads supporting each base and edge. We then use a statistical segmentation strategy to traverse the graph topology directly and determine *paths* through the connectivity graph that represent a contiguous path of significant enrichment over the background distribution (below). In this segmentation process, we scan variable sized windows across the graph and assign significance to each window. We then merge significant paths into a *transcript graph.* Specifically, for a window of fixed size, we slide the window across each base in the connectivity graph (after augmenting it with the non-spliced reads). If a window contains only contiguous non-spliced reads, then it represents a non-spliced part of the transcript. However, if the window hits an edge in the connectivity graph connecting two separate parts of the genome (based on two or more spliced reads), then the path follows this edge to a non-contiguous part of the genome, denoting a splicing event. Similarly, when alternative splice isoforms are present, if a base connects to multiple possible places, then all windows across these alternative paths are computed. Using a simple recursive procedure we can compute all paths of a fixed size across the graph.

Identification of significant segments

To assess the significance of each path, we first define a background distribution. We estimate a genomic defined background distribution by permuting the read alignments in the genome and counting the number of reads that overlap each region and the frequency by which they each occur. Specifically, if we are interested in computing the probability of observing alignment a (of length r) at position i (out of a total genome size of L) we can permute the alignments and ask how often read a overlaps position i. Under this uniform permutation model, the probability that read a overlaps position i is simply r/L. Extending this reasoning, we can compute the probability of observing k reads (of average length r) at position i as the binomial probability. Given the large number of reads and the large genome size, the binomial formula can be well approximated by a Poisson distribution where λ=np (or the number of reads/number of possible positions).

Given a distribution for the real number of counts over each position we scan the genome for regions that deviate from the expected background distribution. First consider a fixed window size w. We slide this window across each position (allowing for overlapping windows), and compute the probability of each observed window based on a Poisson distribution with $\lambda=wnp$. Since we are sliding this window across a genome of size L, we correct our nominal significance for multiple testing, by computing the maximum value observed for a window size (w) across a number of permutations of the data. This distribution controls the family-wise error rate, defined as the probability of observing at least one such value in the null distribution[28]. Notably, we can estimate this maximum permutation distribution well by a distribution known as the scan statistic distribution[29], which depends on the size of the genome that we scan, the window size used, and our estimate of the Poisson λ parameter. This method provides us with a general strategy to determine a multiple testing corrected P-value for a specified region of the genome in any given sample. We use this method to compute a corrected significance cutoff for any given region. Finally, to identify significant intervals, we scan the genome using variable sized windows, computing significance values for each and filtering by a 5% significance threshold. For each window size, we merge the significant regions that passed this cutoff into consecutive intervals. We trim the ends of the intervals as needed, since we are computing significant windows (rather than regions) and it is possible that an interval need not be fully contained within a significant region. Trimming is performed by computing a normalized read count for each base in the interval compared to the average number of reads in the genome. We then trim the interval to the maximum contiguous subsequence of this value. We test this trimmed interval using the scan procedure and retain it only if it passes our defined significance level.

We work with a range of different window sizes in order to detect paths (intervals) with variable support, Small windows have power to identify short regions of strong enrichment (e.g. short exon which is highly expressed), whereas long windows capture long contiguous regions with often lower and more 'diffuse' enrichment levels (*e.g.* a longer lower expression transcript, whose 'moderate evidence' aggregates along its entire length).

Estimation of library insert size

We estimated the insert size distribution by taking all reconstructed transcripts for which we only reconstructed a single isoform and computing the distribution of distances between the paired-end reads that aligned to them.

Weighting of isoforms using paired end edges

Using the size constraints imposed by the length of the paired ends, we assigned weights to each path in the transcript graph. We classified all paired ends overlapping a given path and assigned them to all possible paths that they overlapped. We then assigned a probability to each paired end of the likelihood that it was observed from this transcript given the inferred insert size for the pair in that path. We used an empirically determined distribution of insert sizes, estimated from single isoform graphs. We then scaled each value by the average insert size. We refer to this scaled value as our insert distribution. For each paired end in a path, we computed *I*, the inferred insert size (the distance between nodes following along the full path) minus the average insert size. We then determined the probability of *I* as the area in our insert distribution between $-I$, I. This value is the probability of obtaining the observed paired end insert distance given this distribution of paired end reads. To aggregate these into weights for each path, we simply weight each paired end by its probability of observing to the given path. Paired ends that equally support multiple isoforms will count equally for all, but paired ends with biases toward some isoforms and against others will provide weighted evidence for each isoform. We assign this weight to each isoform path. This score is normalized by the number of paired ends overlapping the path. We filter paths with little support (normalized score<0.1) of paired reads supporting it.

Determination of expression levels from RNA-Seq data

Expression levels are computed as previously described[8]. Briefly, the expression of a transcript is computed in Reads Per Kilobase of exonic sequence per Million aligned reads (RPKM) defined as: rpkm(transcript) = $\frac{10^9 r}{Rt}$, where r is the number of reads mapped to the exonic region of the transcript, *t* is the total exonic length of the transcript, and *R* is the total number of reads mapped in the experiment.

Array expression profiling in mES cells

Microarray hybridization data was obtained from our previous studies including ES and NPCs[17] and MLFs[4].

Comparisons to known annotation

The reconstructed transcripts were compared to the RefSeq genome annotation[16] (NCBI Release 39). To determine whether a known annotation of a protein coding gene from RefSeq was fully reconstructed, we first compared the 5' and 3' ends of the reconstructed *vs* the annotated transcript. If these overlapped, we further verified that all exons in the annotated transcript matched those in the reconstructed version. To score the portion of an annotated transcript covered by our reconstructions, we found the reconstructed transcript whose exons covered the largest fraction of the annotated transcript, and reported the portion of the annotation that it covered.

ChIP-seq profiles in mES cells and determination of K4 and K36 regions

To determine regions enriched in chromatin marks from ChIP-seq data we used our previously described method[4] applied to ESC, MLFs, and NPCs data[4,17].

Determination of external and internal 5' start sites

We identified alternative 5' start sites by comparing the 5' exon of our reconstructed transcripts to the location of the 5' exon of the annotated gene overlapping it. If the reconstructed 5' start site resided upstream to the annotated 5' we termed it 'external start site'. For the novel 5' ends that are downstream of the annotated 5' end (internal) we required a few additional criteria to avoid reconstruction biases due to low coverage. First, we required that the novel internal 5' end do not overlap any of the known exons within the known gene. Second, we required that the reconstructed gene contains a completed 3' end. To determine the presence of H3K4me3 modifications overlapping the promoter regions defined by these novel start sites, we computed regions of enriched K4me3 genome-wide (as previously described) and intersected the location of the novel 5' exon (both internal and external) with the location of a K4me3 peak.

Determination of premature/extended 3' end

To determine novel 3' ends, we compared the locations of the 3' exon of our reconstructed 3' ends and those of annotated genes. If the reconstruction extended past the annotated 3' end, we classified it as an extended 3' end. If the reconstruction ended before the annotated 3' end we required that it not overlap any known exon and have a fully reconstructed 5' start site.

Determination of sequence conservation levels

We used the SiPhy[30] algorithm and software package to estimate ω, the deviation ('contraction' or 'extension') of the branch length compared to the neutral tree based on the total number of substitutions estimated from the alignment of the region of interest across 20 placental mammals (build MM9, http://hgdownload.cse.ucsc.edu/goldenPath/mm9/multiz30way/). For global (whole transcript) conservation, we estimated ω for each protein coding, lincRNA and antisense transcript exon and compared it to similarly sized regions within introns. To identify local regions of conservation within a transcript, we computed ω for all 12-mers within the transcript sequence, and assigned a p-value for each 12-mer based on the chi-square distribution, as previously described[30]. We then took all 12-mers showing significance at $p< 0.05$, collapsed overlapping 12-mers, and identified constrained regions within the transcript.

ORF determination

We estimated maximal supported open reading frames (ORFs) for each transcript built by scanning for start codons and computing the length (in nucleotides) until the first stop codon was reached.

CSF Scores

To further estimate the coding potential of novel transcripts, we evaluated whether evolutionary sequence substitutions were consistent with the preservation of the reading frame of any detected peptide. In a nutshell, if a transcript encodes a protein, we expect a reduction in frame shifting indels, non-synonymous changes and, in general, any substitution that affects the encoded protein. To assess this, we used Codon Substitution Frequency (CSF) method as previously described[18,19].

RT-PCR validations

Primers were obtained for a randomly selected set of predicted lincRNA, protein coding genes, antisense transcripts, and intron primers; all begining with M13 primer sequence. RNA from mES cells was extracted using Qiagen's RNeasy kit (74106). A a one-step cDNA /RT-PCR reaction was run using Invitrogen's one-step RT-PCR kit (12574-018), following the manufacturer's instructions, with the following PCR protocol: 55°C for 30 minutes, 94°C for 2 minutes (94°C for 15 seconds, 64°C for 30 seconds, 68°C for 1 minute – 40 cycles) 68°C for 5 minutes, 4°C forever. Samples were separated on a 3% agarose gel, and all bands were cut out and gel extracted suing the QIAquick Gel Extraction Kit 28706. 30ng of DNA were mixed with 3.2pmol M13 forward or M13 reverse primer for sequencing.

References

1 Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563 (2005).

2 Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484-1488 (2007).

3 Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242-2246 (2004).

4 Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **in press** (2009).

5 Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-11672 (2009).

6 Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**, 613-619 (2008).

7 Wang, X. *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* **454**, 126-130 (2008).

8 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).

9 Yassour, M. *et al.* Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 3264-3269 (2009).

10 Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics* **40**, 1413-1415 (2008).

11 Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97-101 (2009).

12 Birol, I. *et al.* De novo transcriptome assembly with ABySS. *Bioinformatics (Oxford, England)* **25**, 2872-2877 (2009).

13 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)* **25**, 1105-1111 (2009).

14 Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol* **9**, R175 (2008).

15 Berger, M. F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res* **20**, 413-427 (2010).

16 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **35**, D61-65 (2007).

17 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).

18 Lin, M. F., Deoras, A. N., Rasmussen, M. D. & Kellis, M. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comput Biol* **4**, e1000067 (2008).

19 Lin, M. F. *et al.* Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Res* **17**, 1823-1836 (2007).

20 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**, 38-44 (1991).

21 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323 (2007).

22 Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**, 1570-1573 (2005).

23 Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750-756 (2008).

24 Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-536 (2008).

25 Wu, J. Q. *et al.* Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 5254-5259 (2010).

26 Conti, L. *et al.* Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol* **3**, e283 (2005).

27 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

28 Ewens, W. J. & Grant, G. R. *Statis[t]ical methods in bioinformatics : an introduction.* 2nd edn, (Springer, 2005).

29 Glaz, J., Naus, J. I. & Wallenstein, S. *Scan statistics.* (Springer, 2001).

30 Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62 (2009).

Chapter 3: Expression patterns implicate lincRNAs in diverse biological processes

In this chapter, we describe a functional genomics strategy for inferring putative roles for lincRNAs. The approach suggested functional roles for hundreds of lincRNAs in diverse biological processes.

Parts of this work were first published as:

Guttman M, Amit I, Garber M, French C, Lin M, Feldser D, Huarte M, Cabili M, Carey BW, Cassady J, Jaenisch R, Mikkelsen T, Jacks T, Hacohen N, Bernstein BEB, Kellis M, Regev A, Rinn JL, Lander ES. (2009) Chromatin structure reveals over a thousand highly conserved, large non-coding RNAs in mammals. ***Nature***. 458(7235):223-7

Huarte M, Guttman M, Feldser D, Garber M, Koziol M, Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi L, Regev A, Lander ES, Jacks T, Rinn JL. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 transcriptional response. ***Cell*** 142(3):409-19

We recently reported the identification of more than a thousand large intergenic ncRNAs (lincRNAs) that are evolutionarily conserved in mammalian genomes and thus likely to be functional. Yet, what these functional roles are remains largely uncharacterized. Here, we develop a novel functional genomics approach that assigns putative functions to each lincRNA, revealing a diverse range of roles for lincRNAs in processes from ES pluripotency to cell proliferation. We obtained independent functional validation for the predictions for over 100 lincRNAs, using cell-based assays. In particular, we identify several lincRNAs that are regulated by p53, based on two model systems and find that one of these lincRNAs, termed lincRNA-p21, serves as an important regulator in the p53-dependent transcriptional responses and is required for proper regulation of cellular apoptosis. Collectively, our 'guilt-by-association' approaches along with targeted perturbation studies have demonstrated that lincRNAs play critical regulatory roles across diverse biological processes.

INTRODUCTION

It has become increasingly clear that the mammalian genome encodes numerous large non-coding RNAs[1,2]. We have recently reported the identification of more than a thousand large intergenic non-coding RNAs (lincRNAs) in the mouse genome[3]. Our approach to identify lincRNAs was based on the fact that they contain a chromatin signature of actively transcribed genes, consisting of a histone 3–lysine 4 trimethylated (H3K4me3) promoter region and histone 3–lysine 36 trimethylation (H3K36me3) corresponding to the elongated transcript[3]. These lincRNAs show clear evolutionary conservation, implying that they are functional[3,4].

Despite the identification of thousands of ncRNAs, their functions remained largely unknown. Identifying the functional role of a ncRNA requires direct perturbation experiments, such as loss of function and gain of function. Individual ncRNAs involved in specific processes have been functionally characterized (Reviewed in[5]). For example, XIST is critical for random inactivation of the X-chromosome[6], Air is critical for imprinting control at the Igf2r locus[7], HOTAIR affects expression of the HoxD gene cluster[8] as well as multiple genes throughout the genome[9-11], HOTTIP affects expression of the HoxA gene cluster[12], lincRNA-RoR affects reprogramming efficiency[13], NRON affects NFAT transcription factor activity[14], and Tug1 affects retina development through regulation of the cell cycle[15]. While there are now many examples of large ncRNAs that are required for proper regulation of gene expression, this is just one function for large ncRNAs which play critical roles in processes ranging from telomere replication[16] to translation[17].

Yet, global characterization of ncRNA function is challenging because in most cases it is unclear what phenotype to explore[14]. One approach to classify the putative functional roles of ncRNAs utilizes 'guilt by association'[18]. These methods associate ncRNAs with biological processes based on common expression patterns across tissues and thereby identify groups of ncRNAs associated with specific cellular processes. We utilize expression patterns across 21 mouse cell types and tissues to classify the function of lincRNAs based on their coexpression with functional protein-coding genesets. Using this approach, we classified the putative functional roles of hundreds of lincRNAs across diverse biological processes such as stem cell pluripotency, immune responses, neural processes, and cell cycle regulation.

While such correlations cannot prove that lincRNAs function in these processes, they provide hypotheses for targeted follow-up using loss-of-function experiments. We demonstrate this principle using a group of lincRNAs that are strongly associated with and regulated by the p53 transcriptional pathway. p53 is an important tumor suppressor gene involved in maintaining genomic integrity[19]. In response to DNA damage, p53 becomes stabilized and triggers a transcriptional response that causes either cell arrest or apoptosis[20]. While p53 is known to activate numerous genes, the mechanisms by which p53 leads to gene repression has remained elusive. We recently reported evidence that many lincRNAs physically associate with repressive chromatin modifying complexes and suggested that they may serve as regulators in transcriptional regulatory networks[9]. We therefore hypothesized that lincRNAs may also be directly activated by p53 and subsequently regulate downstream transcriptional repression.

Here we show that lincRNAs play a key regulatory role in the p53 transcriptional response. By exploiting multiple independent cell-based systems, we identified lincRNAs that are misregulated upon p53 perturbations and showed that many are transcriptional targets of p53. Moreover, we find that one of these p53-activated lincRNAs (which we termed lincRNA-p21) serves as a transcriptional repressor in the p53 pathway and plays a role in the maintenance of apoptotic function. Together, these results implicate lincRNAs in diverse biological processes and reveal a direct role for a lincRNA in the p53 transcriptional response.

Results

Functional classification of lincRNAs

Having identified a large collection of conserved lincRNAs, we sought to gain insight into their biological function by profiling their expression across 16 mouse cell types and tissues. These samples consisted of the original four cell types (mES, NPC, mEF, mLF), a time course of embryonic development (Whole Embryo, Hindlimb and Forelimb at embryonic days 9.5, 10.5 and 13.5), and 4 normal tissues (brain, lung, ovary and testis). The expression compendium revealed numerous clusters of correlated and anticorrelated lincRNAs. Unsupervised clustering grouped the tissues and cells in our compendium in a biologically coherent way. For example, the developmental time course was robustly grouped according to their temporal regulation ($p < 10^{-5}$, bootstrap p-value).

The expression data contains a wealth of information about the lincRNAs. As an example, we searched for lincRNAs with an expression pattern opposite to the known lincRNA HOTAIR. Interestingly, we found that the most highly anti-correlated lincRNA in the genome lies in the HOXC cluster, in the same euchromatic domain as HOTAIR; we call this lincRNA, Frigidair (**Figure 1a**). This suggests that Frigidair may repress HOTAIR or perhaps activate genes in the HOXD cluster.

We then compared the expression profiles of the lincRNAs to the expression profiles of protein-coding genes (see methods). We first constructed a matrix *A* showing the correlation coefficient for the lincRNAs vs. ~22,000 protein coding genes, across the 16 cells and tissues; we found that numerous lincRNAs were associated with groups of protein coding genes. Using matrix *A*, we also constructed a matrix *B* showing the correlation of lincRNAs vs. 1700 functional gene sets; correlations were determined using the gene set enrichment analysis (GSEA) methodology and the functional gene sets were taken from the Molecular Signatures Database (MSigDB)[21]. We then performed biclustering on matrix *B* to identify groups of lincRNAs that are associated with various functional categories[22]. Each bicluster represents a group of lincRNAs with shared functional annotations. Finally, we used Gene Ontology (GO) analysis to infer functions associated with each bicluster.

This analysis revealed a wide range of lincRNAs that were associated with distinct and diverse biological processes. These include cell proliferation, RNA binding complexes, immune surveillance, neuronal processes, morphogenesis, gametogenesis, and muscle development (**Figure 1e**). The bicluster with the largest number of associated lincRNAs and highest significance level was associated with cell proliferation, cell-cycle regulation, and RNA binding genes. The second most significant bicluster was associated with innate immunity, response to biotic stimulus and inflammatory response genes.

To assess the validity of the inferred functional associations, we examined the gene sets associated with HOTAIR, a lincRNA of known function (repression of HOXD cluster). HOTAIR showed negative association with HOXD genes (FDR<0.018) and positive association with 'Chang Serum Response' (FDR <0.001) a known predictor of breast cancer metastasis[23]. Together, these associations suggest a role for HOTAIR in breast cancer. Consistent with this hypothesis, it has been experimentally shown that loss of HOXD expression is a signature of invasive breast cancer[24,25].

We then sought to obtain independent experimental validation of the inferred biological functions for many of the lincRNAs. We focused on three large clusters of lincRNAs associated with NFκB signalling, embryonic stem cell pluripotency, and the p53-mediated DNA damage response based on their expression patterns across tissues.

TLR-regulated lincRNAs

We used a similar strategy to investigate lincRNAs associated with the 'immune surveillance' cluster. To search for lincRNAs related to immune pathways, we stimulated sorted

CD11C+ bone marrow-derived dendritic cells with a specific agonist of the toll-like receptor TLR4, lipopolysaccharide (LPS). LPS is well known to cause dramatic activation of NFκB transcriptional activity[26] and induction of many immune genes. We found 20 lincRNAs whose expression changed dramatically upon TLR4-stimulation, representing ~5% of lincRNAs present on the array and comparable to the fraction of regulated protein-coding genes[27]. Interestingly, the greatest change in expression was observed in a lincRNA located ~51Kb upstream of the protein coding gene Cox2, a critical inflammation mediator that is induced by TLR4; we refer to this as linc-Cox2. Using quantitative PCR, we found that linc-Cox2 is induced ~1000 fold over the course of 12 hours following TLR4 stimulation (**Figure 1b**). In contrast, only weak induction of linc-Cox2 was seen upon stimulation of TLR3 (using polyI:C). ,which signals more strongly through IRF3 than NF-KB, suggesting that lincRNAs provide an additional regulatory layer in the signaling network defining the specificity of the innate immune response.

ES cell pluripotency and direct regulation by Sox2, Oct4, and Nanog

Using published data from mouse ES cells, we identified 118 lincRNAs whose promoter loci were bound by the core transcription factors Oct4 and Nanog[28]. Of those represented on our expression array 72% resided in the cluster associated with pluripotency, again supporting the validity of the functional inference. We noticed that one of these lincRNAs, which is only expressed in ES cells, is located ~100 kb from the Sox2 locus, which encodes another key transcription factor associated with pluripotency. This lincRNA is highly expressed only in the pluripotent cell state (**Figure 1c**). We cloned the promoter of this locus (which we will refer to as lincRNA-Sox2) upstream of a luciferase reporter gene and transfected the construct into mouse cells transiently expressing Oct4, Sox2, or both, as well as several controls. We found that Sox2 and Oct4 were each sufficient to drive expression of this lincRNA promoter, while expression of both Oct4 and Sox2 caused synergistic increases of expression. In addition to ES specific lincRNAs, we identified other lincRNAs that are exclusively expressed in other cell lineages such as the neural lineage (**Figure 1d**).

The ultimate proof of function will be to demonstrate that RNAi-mediated knock-out of each lincRNAs has the predicted phenotypic consequences. Toward this end, we examined a recently published shRNA screen of (presumed) protein-coding genes to identify genes that regulate cell proliferation rates in mouse ES cells[29]. The screen involved genes and some unidentified transcripts that had been identified as expressed in ES cells and showing rapid decrease in expression upon retinoic acid treatment. Of the top 10 hits in the screen, one corresponded to a gene of unknown function. We discovered that this gene corresponds to one of our lincRNAs (located ~181 kb from Enc1) contained in both the 'cell cycle and cell proliferation' cluster (FDR < 0.001) and the 'embryonic stem cell' cluster (FDR<0.001). This provides functional confirmation that this lincRNA plays a *direct* role in cell proliferation in ES cells, consistent with the analysis above.

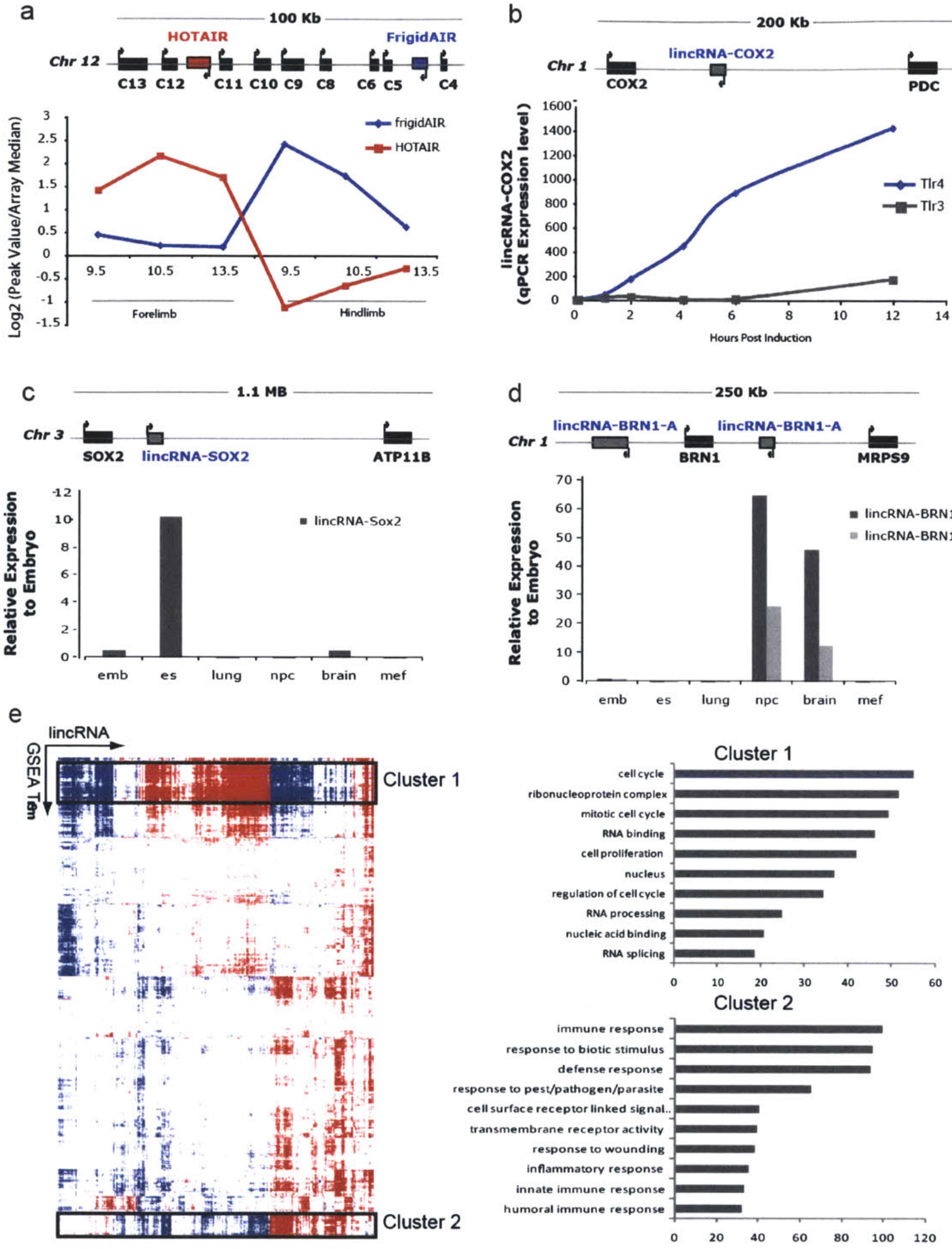


Figure 1: lincRNAs are differentially expressed in various conditions. (a) Map of mouse genomic locus (HOXC) containing HOTAIR, shows relative location of HOTAIR and FrigidAIR. HOTAIR and FrigidAIR show diametrically opposed expression patterns between mouse forelimb (anterior) and mouse hindlimb (posterior). (b) Map of genomic locus containing COX2, a key inflammation gene and a direct NFκB target, along with the location of lincRNA-

COX2. qRT-PCR shows that lincRNA-Cox2 is upregulated in TLR4 stimulated cells (NFκB mediated, green) but not TLR3 stimulated cells (IRF3 mediated, blue). (c) A map of the genomic locus containing SOX2 shows a lincRNA ~50Kb upstream that is similarly expressed specifically in ES cells. (d) Map of the genomic locus of Brn1, a key neural lineage transcription factor, is flanked by two lincRNAs. qRT-PCR shows lineage specific expression, similar to Bn1, in neural lineages. (e) Correlation matrix of lincRNA and functional gene sets. Each entry reflects the association between the lincRNA and the functional gene set based on Gene Set Enrichment Analysis (GSEA). Functional gene sets (columns) and lincRNAs (rows) are shown as either positively (red), negatively (blue) or not correlated (white) with lincRNA expression profiles. Two major cluster are highlighted, 'Cell-Cycle regulation and Cell Proliferation' and 'Immune Surveilence'. Gene ontology of the protein coding genes in these clusters is shown and plotted as the –log(p-value) for the enrichment of each GO term.

p53-dependent regulation of lincRNAs

We hypothesized that some of the lincRNAs associated with 'cell cycle and proliferation' might be regulated by p53. We decided to test this hypothesis in a well-defined DNA damage system that induces p53[30]. This system allows conditional restoration of wildtype Trp53 within an inducible cre-lox system[30]. We will refer to the recombined cells (Trp53 restored): $p53^{+/+}$ and non-recombined: $p53^{-/-}$ mEFs (**Figure 2a**). We performed several controls to demonstrate significant restoration of p53 function upon activation of cre. We obtained $p53^{+/+}$ and $p53^{-/-}$ MEFs and exposed them to a DNA damaging agent (doxorubicin). We then profiled lincRNA expression across a time course at 0, 1, 3, 6, and 9 hours post DNA damage induction. We found 38 lincRNAs that increased significantly across the induction time course in the $p53^{+/+}$ cells (**Figure 2b**). We found that the promoters of these lincRNAs were significantly enriched for the p53 *cis*-regulatory element (compared to all lincRNA promoters, p<.01 Wilcoxon Test). This suggests that p53 directly binds and regulates the expression of at least some these lincRNA genes. We then asked if these lincRNAs were also present in the "cell cycle and proliferation" cluster. Indeed, the p53-induced lincRNAs were strongly enriched in the "cell cycle and proliferation" cluster (p < 10e-7).

Several lincRNAs are transcriptional targets of p53

To validate the functional importance of these classifications, we focused on lincRNAs associated with the p53-mediated DNA damage response. We first sought to identify lincRNAs that could be canonical p53 target genes. We cloned two lincRNA promoter regions with highly conserved canonical p53-binding motifs [31,32] into a luciferase reporter vector. Both the lincRNA-p21 and lincRNA-Mkln1 constructs showed significant induction of luciferase driven in p53-wild type but not in p53-null cells (p<0.01, **Figure 2c**).

To determine if the canonical p53-binding motif is required for the observed transactivation we repeated these experiments in the absence of the p53 binding motif. Mutant promoters resulted in the abolition of the observed transactivation for lincRNA-p21 and lincRNA-Mkln1 in p53$^{+/+}$ cells. Finally, we performed Chromatin Immunoprecipitation (ChIP) experiments to determine if p53 directly binds to the consensus motif of lincRNA-p21 and lincRNA-Mknl1. Indeed, p53 is bound to the consensus motif in the promoters of both lincRNA-p21 and lincRNA-Mkln1 and is not enriched at negative control sites lacking a canonical p53 binding motif (**Fig 2d**). Together these results demonstrate that lincRNA-p21 and lincRNA-Mkln1 are *bona fide* transcriptional targets in the p53 pathway.

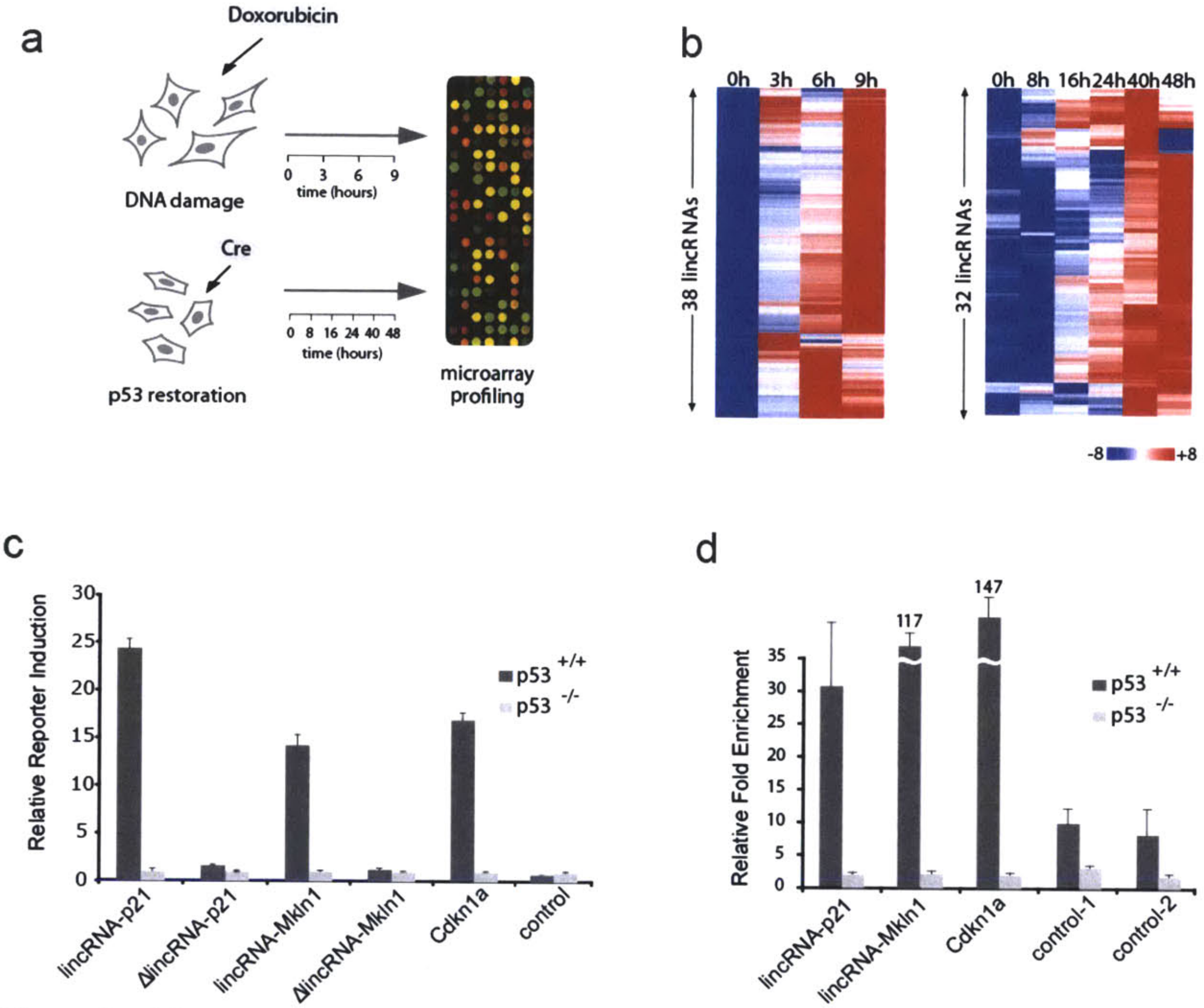


Figure 2: P53 regulated lincRNAs. (a) Experimental layout to monitor p53-dependent transcription. p53-restored (p53$^{+/+}$) and non-restored (p53$^{-/-}$) p53$^{LSL/LSL}$ MEFs were treated with 500nM dox for 0, 3, 6 and 9 hours (top left). KRAS (p53$^{LSL/LSL}$) tumor cells were treated with hydroxytamoxifen for p53 restoration for 0, 8, 16, 24, 40 or 48 hours (bottom left). RNA was subjected to microarray analysis of mRNAs and lincRNAs. (b) lincRNAs activated by p53 induction (FDR < 0.05) in MEF or KRAS system. Colors represent transcripts above (red) or below (blue) the global median scaled to 8 fold activation or repression, respectively. (c) p53-dependent induction of lincRNA promoters requires the consensus p53 binding elements. Relative firefly luciferase expression driven by promoters with p53 consensus motif (lincRNA-p21, lincRNA-Mkln1) or with deleted motif (DlincRNA-p21 and DlincRNA-Mkln1) in p53$^{+/+}$ or p53$^{-/-}$cells. Values are relative to p53$^{-/-}$ and normalized by renilla levels. (d) p53 specifically binds to p53 motifs in lincRNA promoters. p53 ChIP enrichment in p53$^{+/+}$ and p53$^{-/-}$ MEFs on regions with p53 motifs (lincRNA-p21, lincRNA-Mkln1, Cdkn1a) or two irrelevant regions (controls). Enrichment values are relative to IgG, and average of 3 technical replicates of a representative experiment.

lincRNA-p21 regulates gene expression in the p53 pathway

We next sought to determine the consequence of the loss of lincRNA-p21 function in the context of the p53 response. We reasoned that, if lincRNA-p21 plays an important role in orchestrating the p53 transcriptional response, then inhibition of lincRNA-p21 would show similar effects as inhibition of p53 itself. To test this hypothesis, we used RNAi-mediated depletion of lincRNA-p21 and p53 and monitored the resulting transcriptional changes by DNA microarray analysis.

Toward this end, we first designed three pools of siRNA duplexes targeting lincRNA-p21, p53 or non-targeting control sequences. We validated that they were effective at knocking down the intended target genes in p53$^{LSL/LSL}$ restored MEFs (**Figure 3a and 3b**). We then used microarray analysis to examine the broader transcriptional consequences of knockdown of p53 and lincRNA-p21 compared to the non-targeting control. We identified 1520 and 1370 genes that change upon knockdown of p53 or lincRNA-p21, respectively (relative to non-targeting control siRNA, FDR < 0.05). We observed a remarkable overlap of 930 genes in both the lincRNA-p21 and p53 knockdowns, vastly more than would be expected by chance ($p<10^{-200}$) (**Figures 3c**). Strikingly, 80% (745/930) of the common target genes are derepressed in response to both p53 and lincRNA-p21 knockdown; this proportion is much higher than expected by chance (**Figure 3c**). In contrast, the genes misregulated by the p53 knockdown alone showed no bias for upregulation or downregulation, suggesting that lincRNA-p21 participates in downstream p53 dependent transcriptional repression.

To further demonstrate that the observed derepression upon lincRNA-p21 loss-of-function is indeed p53-dependent and not due to off target effects of the RNAi mediated depletion experiments, we performed several additional experiments and analyses. First we repeated the depletion experiments with four individual siRNAs targeting lincRNA-p21, transfected separately rather than in a pool and confirmed the derepression effect across multiple duplexes for select target genes in the microarray experiment. Second we confirmed that the same genes that were derepressed in the lincRNA-p21 and p53 depletion experiments correspond to genes that are normally repressed upon p53 induction in both the KRAS and MEF systems (GSEA FDR < 0.002) (**Figure 3d**). Thus, derepressed genes in the siRNA depletion experiments are highly enriched for genes that exhibit temporal repression upon induction of p53 in both the KRAS and MEF systems in the absence of RNAi treatment. Third we demonstrated that forced over-expression of lincRNA-p21 (Methods) also perturbed the expression of genes that were affected upon depletion of lincRNA-p21. Finally, we did not observe derepression of these genes upon repeating the same siRNA depletion experiments in the absence of p53 (-AdCre). Collectively, these results suggest that lincRNA-p21 serves as a repressor in p53-dependent transcriptional responses.

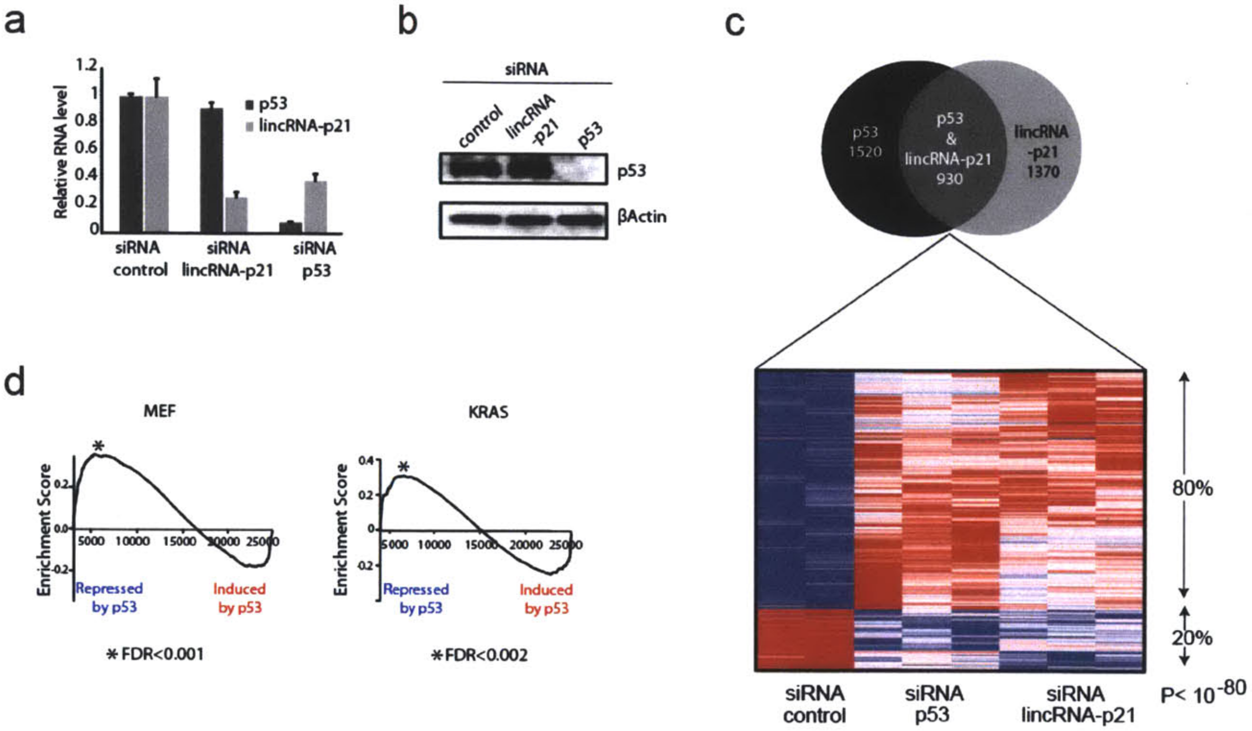


Figure 3: lincRNA-p21 is a global repressor of genes in the p53 pathway. (a) RNAi-mediated depletion of lincRNA-p21 and p53. Relative RNA levels determined by qRT-PCR in p53-reconstitued p53$^{LSL/LSL}$ MEFs transfected with the indicated siRNAs and treated with DOX. Values are the median of 4 technical replicates. (b) p53 protein levels after lincRNA-p21 and p53 depletion from cells treated as in (A). bActin levels are shown as loading control. (c) Many genes are corepressed by lincRNA-p21 and p53. Top: Venn diagram of differentially expressed genes (FDR < 0.05) upon p53 depletion (left) or lincRNA-p21 depletion (right); cells were treated as in (A) and subjected to microarray analysis. Bottom: expression level of genes in lincRNA-p21 and p53 siRNA-treated cells relative to control siRNA experiments. Expression values are displayed in shades of red or blue relative to the global median expression value across all experiments (linear scale). (d) Genes derepressed by lincRNA-p21 and p53 depletion overlap with the genes repressed by p53 restoration in the MEF and KRAS systems. The black line represents the observed enrichment score profile of genes in the lincRNA-p21/p53 derepressed gene set to the MEF or KRAS gene sets respectively.

lincRNA-p21 regulates apoptosis

. The activation of the p53 pathway has two major phenotypic outcomes: growth arrest and apoptosis [33]. Consistent with this, our microarray analysis demonstrates that p53 and lincRNA-p21 share regulation of several apoptosis and cell-cycle regulator genes. However, critical cell-cycle regulators, such as Cdkn1a/p21 Cdkn2a or Reprimo are regulated by p53 independently of lincRNA-p21. In contrast, p53 and lincRNA-p21 share regulation of common genes involved in apoptosis such as Apaf1, Noxa, G2e3 or Bcl2l3. Thus, we aimed to determine the physiological relevance of lincRNA-p21 in the p53 response.

Toward this end, we used RNAi mediated depletion of lincRNA-p21 in either dox treated or untreated primary MEFs. We similarly performed RNAi mediated depletion of p53 (as a positive control) or used the non-targeting siRNA pool (as a negative control) under the same conditions. We observed a significant increase in viability of cells treated with siRNAs targeting lincRNA-p21 or p53 in the presence of DNA damage compared to the control siRNA pool (**Figure 4a,b**). Such increase in viability was greater for depletion of p53, but still highly significant for depletion of lincRNA-p21 ($P < 0.01$). We observed similar results using three individual siRNA duplexes targeting lincRNA-p21, as well as two different control siRNA pools (**Figure 4b**). These results demonstrate that lincRNA-p21 plays a physiological role in regulating cell viability upon DNA damage in this system, although it does not distinguish whether the effect is due to misregulation of the cell cycle or apoptosis[33].

To distinguish between these two possibilities, we quantified the proportion of the cell population undergoing apoptosis by detection of Annexin-V by FACS analysis. We observed a significant decrease in the number of apoptotic cells in both the lincRNA-p21 and p53 depleted cells relative to siRNA control, when cells were subjected to DNA damage ($P<0.01$) (**Figure**

4d,e). Consistent with the lincRNA-p21 dependent reduction of apoptosis we observed a decrease in Caspase 3 cleavage relative to controls (**Figure 4f**). We note that MEFs typically respond to DNA damage with cellular arrest, rather than apoptosis [34]. However, we do observe a reproducible and similar reduction of apoptotic cells in response to DNA damage in both lincRNA-p21 and p53 experiments. We further determined that the observed apoptosis is dependent on the dosage of dox-induced DNA damage. Thus, the apoptosis response is clearly both p53-dependent and lincRNA-p21-dependent, with this dependence being confirmed in multiple settings (**Figures 4b,d,f**); The decrease of apoptotic cells in response to knockdown of lincRNA-p21 was comparable to that caused by knockdown of p53, suggesting that lincRNA-p21 is required for the p53-dependent induction of apoptosis under the experimental conditions used (**Figure 4c**).

We further tested whether depletion of lincRNAp-21 affects cell cycle regulation in response to DNA damage by measuring 5-bromo-2-deoxyuridine (BrdU) incorporation and propidium iodide staining of the cells. Consistent with the known function of p53, depletion of p53 caused a significant increase in BrdU incorporation in response to DNA damage ($P<0.01$). In contrast, depletion of lincRNA-p21 neither showed significant changes in BrdU levels nor in the percentages of cells in any of the cell cycle phases (S, G1 or G2) either treated or untreated with DOX (**Figure 4c**). These results suggest that lincRNA-p21 does not substantially contribute to cell cycle arrest upon DNA damage.

We next wanted to determine whether, conversely to lincRNA-p21 depletion, overexpression of lincRNA-p21 would result in increased apoptosis. Indeed, lincRNA-p21 overexpression in a lung cancer cell line harboring a KRAS mutation (referred to as LKR) and in 3T3 MEFs caused a significant decrease in cell viability (Experimental Procedures and Figures

4g). This decrease in viability was due to increased apoptosis in response to DNA damage (P<0.01) and not to an effect in cell cycle regulation (**Figure 4h,i**).

Several additional lines of evidence are consistent with the observed apoptosis phenotype. First, we observed that both lincRNA-p21 and p53 repress genes involved in the repression of apoptosis and the promotion of cell survival (Bcl2l3, Stat3, Atf2). Second, although lincRNA-p21 and p53 depletions exhibit derepression of cell-cycle regulators, some key cell-cycle genes are regulated by p53 independently of lincRNA-p21, including Ckn1a/p21. In fact, depletion of lincRNA-p21 does not perturb the transcript levels of Cdkn1a/p21 nor the protein stability; thus lincRNA-p21 is insufficient to mount a cell-cycle phenotype. Taken together, these observations demonstrate that lincRNA-p21 plays an important role in the p53-dependent induction of cell death under the experimental conditions used.

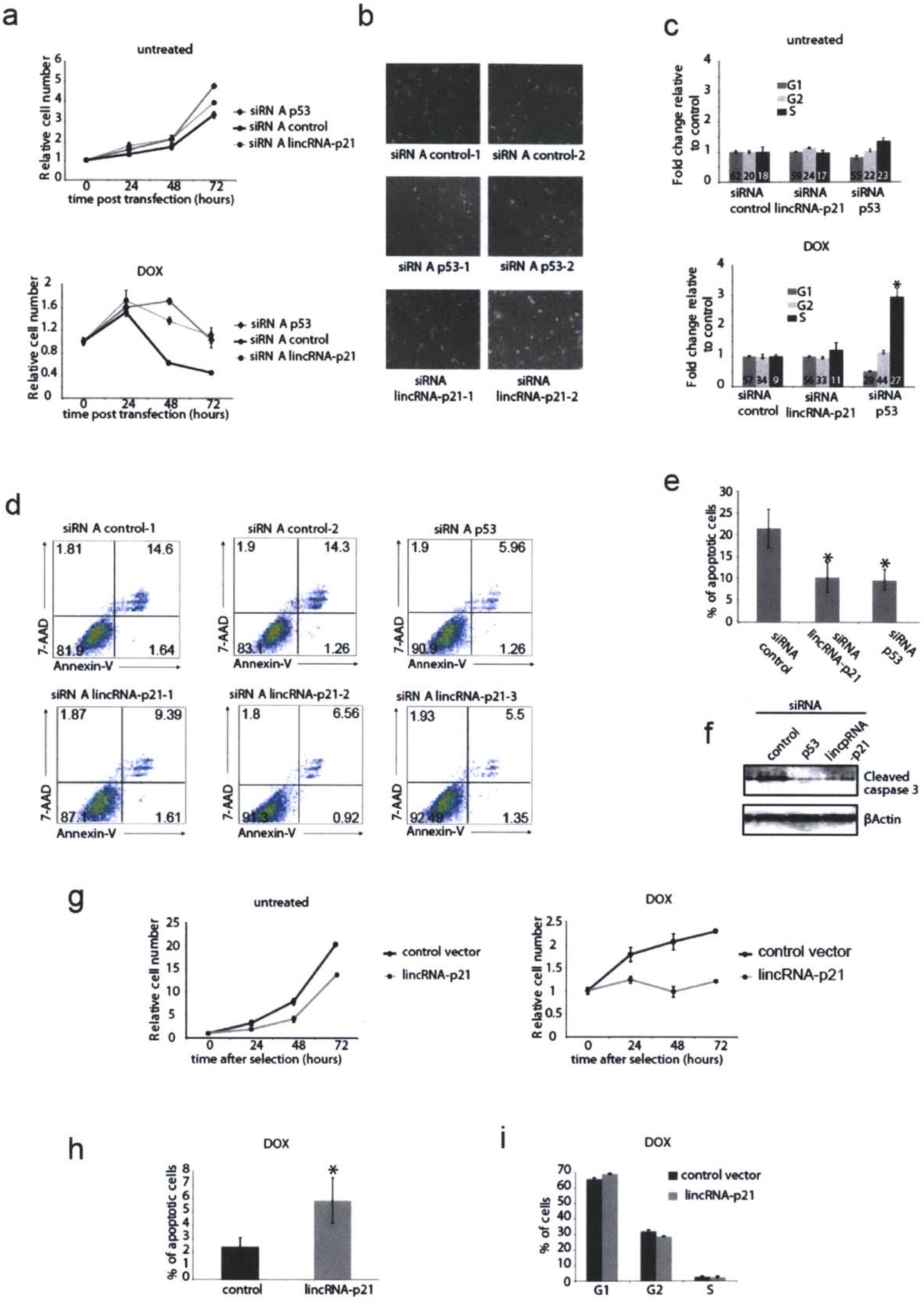

a
untreated
Relative cell number
time post transfection (hours)
siRN A p53
siRN A control
siRN A lincRNA-p21
DOX
b
siRN A control-1
siRN A control-2
siRN A p53-1
siRN A p53-2
siRNA lincRNA-p21-1
siRNA lincRNA-p21-2
c
untreated
Fold change relative to control
G1
G2
S
siRNA control
siRNA lincRNA-p21
siRNA p53
DOX
d
siRN A control-1
siRN A control-2
siRN A p53
siRN A lincRNA-p21-1
siRN A lincRNA-p21-2
siRN A lincRNA-p21-3
7-AAD
Annexin-V
e
% of apoptotic cells
siRNA control
siRNA lincRNA-p21
siRNA p53
f
siRNA
control
p53
lincpRNA -p21
Cleaved caspase 3
βActin
g
untreated
Relative cell number
time after selection (hours)
control vector
lincRNA-p21
DOX
h
DOX
% of apoptotic cells
control
lincRNA-p21
i
DOX
% of cells
control vector
lincRNA-p21
G1
G2
S

Figure 4: lincRNA-p21 is required for proper apoptotic induction. (a) Increased cell viability of lincRNA-p21 depleted cells. Relative number of siRNA-transfected MEFs treated with 400nM DOX from 24h after transfection (right) or untreated (left) determined by MTT assay. (b) Depletion of lincRNA-p21 with individual siRNAs increases cell viability. Images of MEFs treated with different individual siRNAs after 48 hours of DOX treatment (72h post transfection). (c) LincRNA-p21 depletion doesn't affect cell cycle regulation. Relative cell numbers in each cell cycle phase determined by FACS of BrdU incorporation and PI staining of MEFs treated as in (a). Numbers inside bars represent percentages of cells in each phase (average of 3 biological replicates). (d) LincRNA-p21 depletion causes a decrease in cellular apoptosis. P53-reconstituted $p53^{LSL/LSL}$ MEFs transfected with three individual siRNAs targeting lincRNA-p21 (bottom), two independent control siRNAs (top left and middle) or a siRNA pool targeting p53 (top right). 24 hours after transfection cells were treated with 400nM doxorubicin and 14 hours later harvested and subjected to FACS analysis. X-axis represents Annexin-V and y-axis 7-AAD staining. The percentage of cells in each quadrant are indicated. (e) Decreased apoptosis caused by lincRNA-p21 depletion. Quantification of the relative apoptosis levels by Annexin-V FACS detection at 38h post transfection (14h of 400nM DOX treatment) in MEFs treated as in (a). Values are average of 3 biological replicates of a representative experiment. Stars show significant difference ($P<0.01$) relative to control. (f) LincRNA-p21 depletion in p53-reconstituted $p53^{LSL/LSL}$ MEFs causes decrease in Caspase 3 cleavage. Levels of cleaved Caspase 3 or control bActin in p53 reconstituted-$p53^{LSL/LSL}$ MEFs treated with the indicated siRNA pools and 500nM DOX for 14 hours. (g) Decreased cell viability caused by lincRNA-p21 overexpression. Relative numbers of LKR cells overexpressing lincRNA-p21 or control plasmid determined by MTT assay. Values are average of 3 biological replicates. (h) Overexpression of lincRNA-p21 causes cellular apoptosis under DNA damage induction. Apoptosis quantification by Annexin-V FACS in LKR cells overexpressing lincRNA-p21 or control vector treated with 500nM DOX. Values are the average of 3 biological replicates. Star represents $P<0.01$. (i) LincRNA overexpression doesn't affect cell cycle regulation. Cell cycle analysis of DOX-treated LKR cells overexpressing lincRNAp21 or control plasmid. Values are average of 3 biological replicates.

DISCUSSION

We provide a functional genomics pipeline for inferring putative roles for lincRNAs. This 'guilt-by-association' approach associates lincRNAs with biological processes based on common expression patterns across tissues and thereby identify groups of ncRNAs associated with specific cellular processes. This approach suggested functional roles for 150 lincRNAs that we studied on microarrays, and the independent experiments provided support for the predicted pathways for ~85 lincRNAs. The pipeline thus provides a useful guide for hypothesis-driven functional studies.

Our analysis strongly suggests the existence of a rich world of functional lincRNAs with a diversity of biological roles that may resemble that seen for proteins. For example, the expression patterns of lincRNAs are as diverse as those of protein coding genes across the cell types and conditions studied, allowing us to associate lincRNAs with a myriad of biological processes at least at the level of correlated expression. We also show that some lincRNAs are directly regulated by the key master regulators of stem-cell pluripotency such as Sox2 and Oct4. And, we validated several functional associations between lincRNAs that are likely regulated by p53 in cancer and by NFκB during the immune response.

It is important to point out that not all non-coding transcripts act as functional RNA molecules. Several examples of intergenic transcription have been identified where the act of transcription alone changes the chromatin and transcription factor binding landscape allowing activation and repression of neighboring genes[35,36]. As such, these correlations cannot prove that lincRNAs function in these predicted processes but rather provide hypotheses for targeted loss-of-function experiments. Methods that degrade RNA after its transcription, such as RNAi, can distinguish between a functional RNA molecule and the act of transcription for which there should be no observable effect upon RNA degradation.

We utilize these predictions to show their value in identifying direct functional roles of a specific lincRNA predicted to play a role in the p53 process. For example, the lincRNA-p21 was predicted to be associated with the p53-mediated DNA damage response[18]. Indeed, lincRNA-p21 was found to be a target of p53 and upon perturbation was shown to regulate apoptosis in response to DNA damage[37]. Another lincRNA, lincEnc1[18], was predicted to have a role in cell-cycle regulation in embryonic stem cells (ESCs) and shown in a distinct study to affect proliferation in ESCs[29].

A fundamental issue will now be to elucidate the biological functions and determine the mechanisms by which lincRNAs act. One clue may come from our previous observation that HOTAIR[8] represses gene expression and is associated with chromatin remodeling proteins, together with recent similar observations for XIST[38]. Based on these observations, we speculate that many lincRNAs may play a role in transcriptional control, perhaps by guiding chromatin modifying proteins to target loci. Testing this hypothesis will require biochemical and genetic studies, including gene knock-down in appropriate settings. Whatever their mechanism, the lincRNAs appear to be involved in diverse biological process and likely play a direct regulatory role in these processes.

Methods

Protein Coding Gene Expression Profiles

We obtained Affymetrix 430 2.0 mouse gene expression data for all RNA samples profiled on our lincRNA array. For ES, MEF, NPC (GSE8024) and brain, lung, testis, and ovary (GSE9954) arrays were already available in the Gene Expression Omnibus (GEO) and in these cases we downloaded the data. For Forelimb, Hindlimb, and Whole Embryo for days 9.5, 10.5, and 13.5, we generated our own data using Affymetrrix 430 2.0 arrays. For dendridic cells we generated data for, unstimualted, TLR2 stimulated, TLR4 stimulated, and TLR9 stimulated cells using Affymetrrix 430A arrays (RNA isolated as mentioned above).

Correlation Matrix Clustering

We generated a correlation matrix between lincRNAs by computing the Pearson correlation coefficient between all pairs of lincRNAs. A matrix was constructed whoe entries are the correlation coefficients. This matrix was clustered and visualized using the Gene Pattern platform for integrative genomics (http://genepattern.broad.mit.edu/) using a Euclidian distance metric and complete linkage clustering[39]. The same procedures were used to produce, cluster, and visualize the lincRNA-Protein coding gene matrix and the lincRNA-Functional Term matrix.

Gene Set Enrichment Analysis and Functional Term Clustering

Gene Set Enrichment Analysis was performed as previously described[21]. Briefly, we used each lincRNA as a profile, computed the Pearson correlation for each protein coding gene and then ranked the protein coding genes by their correlation coefficient. The rank of these genes was used to identify significant gene sets, using the weighted Kolmogorov–Smirnov (KS) test[21]. Gene sets were permuted 1000 times to obtain FDR corrected p-values. We constructed an association matrix between lincRNAs and terms. We then performed biclustering on this matrix to identify significant lincRNAs associated with functional terms. Biclusters were obtained using the Samba algorithm implemented in the Expander software package.

Identifying Differentially Expressed Genes in DNA Damage Stimulated Cells

Tp53$^{LSL/+}$ heterozygous mice were intercrossed and fibroblasts were derived from p53$^{LSL/LSL}$ and p53$^{+/+}$ embryos as decribed previously[30]. Sub-confluent cultures were infected on two consecutive days with adenoviruses expressing green fluorescent protein (AdGFP) or Cre recombinase (AdCre) (University of Iowa Genetics Core Facility). Cells were then seeded overnight into 10 cm dishes and treated with 500 nM doxorubicin (Sigma) for the indicated time course. Cells were harvested into Trizol reagent (Invitrogen) and total RNA was extracted for subsequent analysis as described[40].

In parallel, cells were harvested for analysis of p53 protein expression. A monoclonal antibody to mouse p53 (Gift from Kristian Helin) was used for protein blotting and detected by enhanced chemiluminescence (GE Healthcare) per manufacturer's instructions. Hsp90 monoclonal antibody served as a loading control (BD Biosciences).

We identified differentially expressed genes, protein coding and lincRNA, using the Patterns from Gene Expression (http://www.cbil.upenn.edu/PaGE/) program[41]. Briefly, we determined differential expression between p53$^{+/+}$ MEFs compared to p53$^{-/-}$ MEFs at paired times (paired t-test). We filtered the list by genes that were specifically induced across the time points.

Motif Enrichments

Motifs were represented by Position Weight Matrix (PWM) downloaded from the TRANSFAC matrix database v8.3 (http://www.gene-regulation.com/pub/databases.html)[42]. Given a PWM, for each nucleotide position in a promoter, we calculated an affinity score defined as the log likelihood (LOD score) for observing the sequence given the PWM versus a given random genomic background. We then found the best conserved motif instance over the entire promoter region for each PWM. An instance was considered conserved if its conservation score was in the top 5% of the genome distribution.

We computed this score for each lincRNA promoter and computed enrichment of the motif for our experimentally determined set compared with all lincRNA promoters. To ensure that enrichment was not due to nucleotide bias within the promoter, we shuffled the PWM and computed enrichment for the true PWM compared to the shuffled PWMs. Enrichment was computed using a two-sided Wilcoxon rank-sum test between the set and the background. We then computed an FDR to correct for testing of multiple PWMs.

Bone marrow dendritic cell (BMDC) cultures

Bone marrow was harvested from 6-8 week old female mice and cultured for 6 days in GM-CSF[43] supplemented medium. Non-adherent cells were sorted using anti-CD11c-beads (Miltenyi Biotech) according to manufacturer's guidelines. CD11c positive cells where replated $1.5*10^6$ cells/plate on day 7. BMDCs were left untreated or stimulated with 100 ng/ml LPS for 6 hours or stimulated with 250 ng/ml Pam3CSK4 for 6 hours (TLR2 stimulation) or with CpG oligonucleotide 1uM for 6 hours (TLR9 stimulation) or with poly-inosine:cytosine (polyI:C) 2ug/ml for 6 hours (TLR3 stimulation) . Cells were then collected by scraping and RNA was purified using the miRNAEasy RNA isolation kit (Qiagen). RNA integrity was verified using bioanalyzer (Agilent).

Real-time quantitative PCR.

cDNA was generated by the use of High-Capacity cDNA Archive Kit (Applied Biosystems). Real-time PCR assays were performed using SYBR Green I as a fluorescent dye on a lightCycler 480 (Roche), according to the manufacturer's guidelines. Experiments were carried out in triplicate, and relative gene expression was normalized to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) RNA levels. Real-time PCR primer pairs for protein coding genes were designed using ProbeLibrary (https://www.roche-applied-science.com/sis/rtpcr/upl/index.jsp), primer pairs for lincRNA were designed using primer3 (http://frodo.wi.mit.edu/) with similar settings.

Cell lines and *in vivo* models

KRAS Lung tumor-derived cell lines were isolated from individual tumors . Isolation of matched p53$^{+/+}$ and p53$^{-/-}$ MEFs, p53$^{LSL/LSL}$ MEFs, Lymphomas and Sarcomas and p53 restoration as described [44]. Primary wt MEFs and 3T3 MEF cells were purchased from ATCC. Transfection, infection and treatment conditions are described in Supplemental Experimental Procedures.

Promoter reporter assays

LincRNA promoters were cloned into the pGL3-basic vector (Promega) and motif deletions were performed by mutagenesis. p53- reconstituted or control p53$^{LSL/LSL}$ MEFs were transfected with 800ng of pGL3 and 30ng of TK-Renilla plasmid per 24 well. 24 hours later cells were treated with 500nM dox for13 hours and cell extracts were assayed for firefly and renilla luciferase activities with Dual Luciferase Reporter Assay System (Promega E1910).

lincRNA and gene-expression profiling.

RNA isolation, lincRNA expression profiling (Nimblegen arrays) and analysis were performed as described [3] and Supplemental Experimental Procedures). Affymetrix, gene-expression profiling was performed as described [3].

Antibodies

Anti-p53:Novocastra (NCL-p53-CM5p) (western blot) and Vector Labs (CM-5) (ChIP). Anti-hnRNP-K: Santa Cruz Biotechnology (sc-25373) (western blot) and Abcam (Ab70492 and Ab39975) (ChIP and RIP). Control rabbit IgG Abcam (Ab37415-5) (RIP and ChIP-chip).

Viability and apoptosis assays and cell cycle analysis

MTT assays were performed using Cell Proliferation Kit I from Roche (11465007001) in 96-well plates with initial density of 2500 cells/well. For apoptosis quantification, the Apoptosis Detection Kit I from BD Biosciences (cat#559763) was used followed by FACS [45]. Cell cycle analysis was performed as described [46].

Cloning, RNA pull-down, deletion mapping, RIP, ChIP

5' and 3' RACE cloning of lincRNA-p21 was performed from total RNA of dox-treated MEFs using RLM-RACE Kit (Ambion AM1700). RNA pull-down and deletion mapping were performed as described [8] using 1mg of mES nuclear extract and 50 pmol of biotinylated RNA. RNA-bound proteins were resolved in a SDS-PAGE gel, bands were cutout and analyzed by Mass Spec as described [47] or detected by western blot. Native RIP was carried out as described [8]. For cross-linked RIP, cells were cross-linked with 1% formaldehyde, 6ug of antibody was added and incubated overnight, recovered with protein G magnetic Dynabeads and washed three times in RIPA buffer. After reverse-crosslink, RNA was analyzed by qRT-PCR. P53 ChIP and hnRNP-K ChIP-chip experiments were performed as previously described [8].

RNA interference and lincRNA-p21 overexpression

siRNA transfections were done in 6-well plates of subconfluent cells with 75nM of siRNA and 3ul of Lipofectamine 2000 (Invitrogen) per well following manufacturer's instructions. For overexpression, lincRNA-p21 was cloned into the pBABE vector and after transfection cells were selected with 2ug/ml puromycin for 8 days.

Cell lines, p53 restoration and DNA damage induction

Lung tumor-derived cell lines were isolated from individual tumors from *KrasLA2/+; Trp53LSL/LSL Rosa26CreERT2* animals (D.F. and T.J. manuscript in preparation). Lymphomas and Sarcomas were isolated when they formed in *Trp53LSL/LSL Rosa26CreERT2* animals as described (Ventura et al. 2007). For p53 restoration, cultured tumor cell lines were incubated with 500nM 4-hydroxytamoxifen (Sigma) for the indicated time points and $p53^{LSL/LSL}$ MEFs, were infected with AdenoCre virus or AdenoGFP for 24h (University of Iowa) at moi of 5. For DNA damage, cells were treated with 100 to 500nM doxorubicin hydrochloride (Sigma D1515).

lincRNA and Protein Coding Gene Expression Profiling

High resolution DNA tiling arrays were designed on the Nimblegen platform to represent a random sampling of ~400 lincRNAs identified in the mouse genome. Total RNA from different experimental conditions was amplified using poly-dT and labeled as described [3].

Identifying Differentially Expressed lincRNAs

We designed custom Nimbelgen tiling microarrys which tile the exonic regions of each mouse lincRNA at 10bp resolution. To identify lincRNAs that were differentially expressed in these conditions, we first determined which lincRNAs are significantly expressed in each sample. We then used this set of expressed lincRNAs to test for differential expression.

To determine expressed lincRNAs we used our previously developed statistical algorithm to identify peaks in hybridization. We first normalized the data by dividing each probe value by the average probe intensity across the array. We scanned each region and computed a score defined as the sum of the normalized probe intensities. To determine the significance of this score we permuted the intensity values assigned to each probe and recalculated the statistic. We took the value for each permutation as the maximum score obtained for any random region. We performed 1000 permutations and assigned a multiple testing corrected p-value to each region based on its rank within this distribution. All exons with a p-value less than 0.05 were retained.

We computed differentially expressed exons by extending the above strategy but computed a t-statistic between each group (ie 0hr vs 8hr). We assessed a multiple testing corrected p-value by permuting the probe values across all conditions and recomputing the t-statistic. We performed 1000 permutations and generated a maximum distribution for each permutation and assigned FWER corrected p-values. We retained all exons with p-values < 0.05.

We performed post-processing of these results to ensure robust differential lincRNAs. Specifically, for MEF time course we required that a lincRNA exon was differentially expressed between $P53^{+/+}$ and $P53^{-/-}$ cells and also differentially expressed between any time point and time

0. For the KRAS experiment we required that any differential exon be differentially expressed in 2 consecutive time points compared to time 0.

Protein Coding Gene Expression Profiles

We generated expression profiles for protein coding gene expression using Affymetrix 430 2.0 arrays. We identified differentially expressed genes using the Patterns from Gene Expression (http://www.cbil.upenn.edu/PaGE/) program. Briefly, we determined differential expression using a t-statistic between groups and permutation distribution to compute an FDR for each gene. We filtered all genes with an FDR<0.05 as significantly differentially expressed. We filtered the list by genes similar to the criteria used for the lincRNA (tiling arrays). We required differential expression between $P53^{+/+}$ and $P53^{-/-}$ for each time point and differential expression compared to time 0. For the RAS experiment we required differential expression of each gene for at least 2 consecutive time points.

Gene Set Enrichment Analysis and Functional Term Clustering

Gene Set Enrichment Analysis was performed as previously described (Grant et al., 2005, [21] Briefly, we used each condition as a group (ie siLincRNA-p21 vs siControl) and ranked the gene list based on differential expression between the groups. The rank of these genes was used to identify significant gene sets, using the weighted Kolmogorov–Smirnov (KS) test [41]. Gene sets were permuted 1000 times to obtain FDR corrected p-values. We used gene sets representing the Molecular Signatures Database or custom gene sets defined by other experiments.

p53 Motif Analysis

To scan for conserved motifs in putative P53 targets we used an extension of the a method that scores conservation at single nucleotide resolution based on the evolutionary substitution pattern inferred for the site [48]. Motifs were represented by Position Weight Matrix (PWM) downloaded from the TRANSFAC matrix database v8.3 (http://www.gene-regulation.com/pub/databases.html) [48]. Given a PWM, for each nucleotide position in a promoter, we calculated an affinity score defined as the log likelihood (LOD score) for observing the sequence given the PWM versus a given random genomic background. We then found the best conserved motif instance over the entire promoter region for each PWM. An instance was considered conserved if its conservation score was in the top 5% of the genome distribution. We computed this score for each lincRNA promoter and computed enrichment of the motif for our experimentally determined set compared with all lincRNA promoters. To ensure that enrichment was not due to nucleotide bias within the promoter, we shuffled the PWM and computed enrichment for the true PWM compared to the shuffled PWMs. Enrichment was computed using a two-sided Wilcoxon rank-sum test between the set and the background. We then computed an FDR to correct for testing of multiple PWMs.

RNA interference and lincRNA-p21 overexpression

siRNA oligos targeting lincRNA-p21 (#1 UGAAAAGAGCCGUGAGCUA, #2 AAAUAAAGAUGGUGGAAUG and #3 AGUCAAAGGCAAUGAGCAU) and hnRNP-K (siRNA smart pool M-048002) were purchased from Dharmacon. p53 siRNAs (#1 AGAAGAAAAUUUCCGCAAA and #2 ACAGCGUGGUGGUACCUUA) were purchased from Ambion. Non-targeting siRNAs were purchased from Dharmacon (D-001206-14) and Ambion (AM4636). siRNA transfections were done in 6-well plates of subconfluent cells with 75nM of siRNA and 3 ul of Lipofectamine 2000 (Invitrogen) per well following manufacturer's instructions. For overexpression, LincRNA-p21 was cloned into the pBABE vector and after transfection cells were selected with 2mg/ml puromycin for 8 days. For gene expression profiling of lincRNA-p21 overexpression, pBABE plasmid expressing lincRNA-p21 or empty vector were transfected into p53-reconstituted p53$^{LSL/LSL}$ MEFs and 24 hours later treated with 500nM doxorubicin. 14h after treatment total RNA was extracted for microarray analysis.

Nuclear fractionation

For nuclear fractionation 10^7 cells were harvested and resuspended in 1ml of PBS, 1ml of buffer C1 (cell lysis buffer, Qiagen) and 3ml of water, and incubated for 15 minutes on ice. Then cells were centrifuged for 15 minutes at 2,500 rpm, the supernatant was discarded and the nuclear pellet was kept for RNA extraction.

References

1 Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nature reviews* **10**, 155-159 (2009).

2 Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629-641 (2009).

3 Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227 (2009).

4 Ponjavic, J., Oliver, P. L., Lunter, G. & Ponting, C. P. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* **5**, e1000617 (2009).

5 Mattick, J. S. The genetic signatures of noncoding RNAs. *PLoS Genet* **5**, e1000459 (2009).

6 Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379**, 131-137 (1996).

7 Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810-813 (2002).

8 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323 (2007).

9 Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-11672 (2009).

10 Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689-693 (2010).

11 Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-1076 (2010).

12 Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-124 (2011).

13 Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**, 1113-1117 (2010).

14 Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**, 1570-1573 (2005).

15 Young, T. L., Matsuda, T. & Cepko, C. L. The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr Biol* **15**, 501-512 (2005).

16 Zappulla, D. C. & Cech, T. R. Yeast telomerase RNA: a flexible scaffold for protein subunits. *Proc Natl Acad Sci U S A* **101**, 10024-10029 (2004).

17 Korostelev, A. & Noller, H. F. The ribosome in focus: new structures bring new insights. *Trends Biochem Sci* **32**, 434-441 (2007).

18 Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227 (2009).

19 Vazquez, A., Bond, E. E., Levine, A. J. & Bond, G. L. The genetics of the p53 pathway, apoptosis and cancer therapy. *Nat Rev Drug Discov* **7**, 979-987 (2008).

20 Riley, T., Sontag, E., Chen, P. & Levine, A. Transcriptional control of human p53-regulated genes. *Nat Rev Mol Cell Biol* **9**, 402-412 (2008).

21 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

22 Tanay, A., Sharan, R. & Shamir, R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S136-144 (2002).

23 Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3738-3743 (2005).

24 Carrio, M., Arderiu, G., Myers, C. & Boudreau, N. J. Homeobox D10 induces phenotypic reversion of breast tumor cells in a three-dimensional culture model. *Cancer research* **65**, 7177-7185 (2005).

25 Charboneau, A., East, L., Mulholland, N., Rohde, M. & Boudreau, N. Pbx1 is required for Hox D3-mediated angiogenesis. *Angiogenesis* **8**, 289-296 (2005).

26 Kawai, T. & Akira, S. TLR signaling. *Seminars in immunology* **19**, 24-32 (2007).

27 Huang, Q. *et al.* The plasticity of dendritic cell responses to pathogens and their components. *Science* **294**, 870-875 (2001).

28 Loh, Y. H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38**, 431-440 (2006).

29 Ivanova, N. *et al.* Dissecting self-renewal in stem cells with RNA interference. *Nature* **442**, 533-538 (2006).

30 Ventura, A. *et al.* Cre-lox-regulated conditional RNA interference from transgenes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 10380-10385 (2004).

31 Funk, W. D., Pak, D. T., Karas, R. H., Wright, W. E. & Shay, J. W. A transcriptionally active DNA-binding site for human p53 protein complexes. *Molecular and cellular biology* **12**, 2866-2871 (1992).

32 el-Deiry, W. S., Kern, S. E., Pietenpol, J. A., Kinzler, K. W. & Vogelstein, B. Definition of a consensus binding site for p53. *Nature genetics* **1**, 45-49 (1992).

33 Levine, A. J., Hu, W. & Feng, Z. The P53 pathway: what questions remain to be explored? *Cell Death Differ* **13**, 1027-1036 (2006).

34 Kuerbitz, S. J., Plunkett, B. S., Walsh, W. V. & Kastan, M. B. Wild-type p53 is a cell cycle checkpoint determinant following irradiation. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 7491-7495 (1992).

35 Martens, J. A., Laprade, L. & Winston, F. Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature* **429**, 571-574 (2004).

36 Schmitt, S., Prestel, M. & Paro, R. Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev* **19**, 697-708 (2005).

37 Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409-419 (2010).

38 Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science (New York, N.Y* **322**, 750-756 (2008).

39 Reich, M. *et al.* GenePattern 2.0. *Nature genetics* **38**, 500-501 (2006).

40 Rinn, J. L., Bondre, C., Gladstone, H. B., Brown, P. O. & Chang, H. Y. Anatomic demarcation by positional variation in fibroblast gene expression programs. *PLoS genetics* **2**, e119 (2006).

41 Grant, G. R., Liu, J. & Stoeckert, C. J., Jr. A practical false discovery rate approach to identifying patterns of differential expression in microarray data. *Bioinformatics (Oxford, England)* **21**, 2684-2690 (2005).

42 Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research* **34**, D108-110 (2006).

43 Palliser, D. *et al.* A role for Toll-like receptor 4 in dendritic cell activation and cytolytic CD8+ T cell differentiation in response to a recombinant heat shock fusion protein. *J Immunol* **172**, 2885-2893 (2004).

44 Ventura, A. *et al.* Restoration of p53 function leads to tumour regression in vivo. *Nature* **445**, 661-665 (2007).

45 van Engeland, M., Ramaekers, F. C., Schutte, B. & Reutelingsperger, C. P. A novel assay to measure loss of plasma membrane asymmetry during apoptosis of adherent cells in culture. *Cytometry* **24**, 131-139 (1996).

46 Brugarolas, J. *et al.* Radiation-induced cell cycle arrest compromised by p21 deficiency. *Nature* **377**, 552-557 (1995).

47 Shevchenko, A. *et al.* A strategy for identifying gel-separated proteins in sequence databases by MS alone. *Biochem Soc Trans* **24**, 893-896 (1996).

48 Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62 (2009).

Chapter 4: Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression

In this chapter, we show that many lincRNAs bind to chromatin regulatory proteins and act through their physical interactions to regulate shared gene expression programs.

Parts of this work were first published as:

Khalil AM*, Guttman M*, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. ***Proceedings of the National Academies of Science*** 106(28):11667-72

We recently showed that the mammalian genome encodes more than a thousand large intergenic non-coding RNAs (lincRNAs) that are clearly conserved across mammals and thus functional. Gene expression patterns have implicated these lincRNAs in diverse biological processes including cell cycle regulation, immune surveillance, and embryonic stem cell pluripotency. However, the mechanism by which these lincRNAs function is unknown. Inspired by the observation that the well-characterized lincRNA HOTAIR bind the Polycomb Repressive Complex 2 (PRC2), we tested whether many lincRNAs are physically associated with PRC2. Remarkably, we observe that ~20% of lincRNAs expressed in various cell types are bound by PRC2, and that additional lincRNAs are bound by other chromatin-modifying complexes. Moreover, we show that siRNA-mediated depletion of certain lincRNAs associated with PRC2 leads to changes in gene expression and that the upregulated genes are enriched for those normally silenced by PRC2. We propose a model in which some lincRNAs guide chromatin–modifying complexes to specific genomic loci to regulate gene expression.

Introduction

Mammalian genomes produce a wide variety of non-coding RNA transcripts[1-3]. In addition to classical RNAs (such as ribosomal RNAs, transfer RNAs and others) and more recently discovered classes of small non-coding RNAs (such as microRNAs and promoter associated small RNAs)[4,5], there are many large non-coding RNAs of unknown function[3]. Several, such large non-coding RNAs have been biologically characterized (including XIST, TSIX, HOTAIR and AIR) (Reviewed in reference[3]), but shotgun cDNA sequencing and microarray hybridization have suggested that the vast majority of the mammalian genome can produce RNA transcripts under some circumstances[2,6,7]. The biological significance of these transcripts, however, has been highly controversial because most occur at extremely low levels and show little evolutionary conservation[8,9].

Recently, we developed a new approach for identifying large non-coding RNAs based on a distinctive chromatin signature that marks actively transcribed genes[1]. The signature consists of a short region with Histone 3 Lysine 4 trimethylation (H3K4me3) (corresponding to the promoter) and a longer region with Histone 3 Lysine 36 trimethylation (H3K36me3, corresponding the transcribed region)[1,10,11]. We refer to this chromatin signature as a K4-K36 domain. We generated chromatin-state maps across four mouse cell types, searched for K4-K36 domains and then eliminated those corresponding to known protein-coding genes. We found 1586 novel K4-K36 domains in the four mouse cell types and showed that the vast majority encode large intergenic non-coding RNAs (lincRNAs). These lincRNAs show similar expression levels as protein-coding genes, but lack any protein-coding capacity. Importantly, lincRNAs show significant evolutionary conservation relative to neutral sequences, providing strong evidence that they have been functional in the mammalian lineage[1]. This is in contrast to some recent catalogs of large non-coding RNAs obtained by shotgun sequencing, which show little or no evolutionary conservation within the RNA transcript[8,9]. (We note that these non-conserved RNAs could be functional, but biological evidence such as loss-of-function experiments would be needed to establish their functionality).

Our previous studies demonstrated that groups of lincRNAs exhibit expression patterns across cell types and tissues that correlate with patterns seen for protein-coding genes involved in cellular processes such as cell-cycle regulation, innate immunity responses, and stem cell pluripotency[1]. While these studies clearly demonstrate that there are many functional lincRNAs, key questions remain, including: How many lincRNAs are encoded in mammalian genomes? How do lincRNAs exert their functions? To begin to investigate the number of lincRNAs, we extended our approach of mapping K4-K36 domains to six human cell types. The results expand our catalog to 3289 lincRNAs, which show clear evolutionary conservation within their transcripts. Extrapolation suggests that the total number may approach ~5000 lincRNAs.

To examine the biochemical mechanism by which lincRNAs function, we drew inspiration from one of the few well-studied lincRNAs: HOTAIR. We previously reported HOTAIR as a lincRNA transcribed from within the HOXC cluster and showed that it acts to repress genes in the HOXD cluster, by binding to the Polycomb Repressive Complex 2 (PRC2) and recruiting it to the locus[12]. PRC2 is a methyltransferase that trimethylates H3K27 to repress transcription of specific genes[13,14]. Recently, several other large non-coding RNAs have been found to associate with chromatin modifying complexes – including a large non-coding RNA encoded within the 5' of XIST that can target PRC2 to the inactive X chromosome[15,16], the antisense transcript AIR that is associated with the chromatin-modifying complex G9a, an H3K9me2 methyltransferase[17]; and the Kcnq1ot1 transcript that binds both G9a and PRC2[18]. Some recent studies have demonstrated that large non-coding RNAs bind chromatin proteins that add activating modifications (e.g. Trithorax)[19,20].

These few examples raised the possibility that many lincRNAs might be physically associated with chromatin-modifying complexes and might potentially target them to specific genomic regions. To test this hypothesis, we performed RNA Co-immunoprecipitation (RIP) with antibodies directed against several proteins involved in chromatin-modifying complexes (PRC2 and CoREST), and found that this is indeed the case. We find that as many as 38% of the lincRNAs expressed in the cell types studied are reproducibly associated with one of these complexes. Moreover, we show that RNA-interference-based depletion of various PRC2-associated lincRNAs results in activation of genes known to be repressed by PRC2. Together, our results indicate that thousands of functional lincRNAs are encoded in the human genome and a significant proportion of lincRNAs are physically associated with chromatin-modifying complexes. We propose that some lincRNAs function by regulating the epigenetic landscape at distinctive target loci.

RESULTS

Many lincRNAs are associated with PRC2

We explored the mechanism by which lincRNAs function. As noted above, the lincRNA HOTAIR has been shown to physically associate with the Polycomb Repressive Complex PRC2[12]. This physical association was shown by an RNA immunoprecipitation-polymerase chain reaction (RIP-PCR) assay: total (non-crosslinked) nuclear extract was incubated with an antibody against the SUZ12 protein, a component of PRC2; the extract was precipitated with Protein-A-coupled beads; and the co-precipitated RNA was then subjected to locus-specific reverse transcriptase (RT)-PCR to demonstrate the presence of HOTAIR.

To test whether other lincRNAs are also associated with PRC2, we designed a 'RIP-Chip' assay (see Methods) to assay many lincRNAs simultaneously (**Figure 1**). Briefly, we used antibodies against the proteins SUZ12 and EZH2, components of PRC2[21]. The antibodies were incubated with non-crosslinked nuclear extracts from three human cell types: HeLa cells, lung fibroblasts (hLF) and foot fibroblasts (hFF); these cell types were chosen because they have previously been shown to have distinctive epigenetic landscapes and diverse gene expression patterns[22]. We analyzed the co-precipitated RNAs by hybridization to a custom 'exon-tiling' array (at 10 base resolution), containing exons from ~900 human lincRNA loci and ~1000 human protein-coding genes; the protein-coding genes were previously known to be expressed in at least one of the three cell types. In parallel, we carried out a mock control with a non-immune rabbit IgG polyclonal antibody to assess non-specific interactions that may occur in RIP.

To identify lincRNAs and protein-coding genes that are co-precipitated with each of the PRC2 components, we analyzed the hybridization data with a peak-calling algorithm that finds regions in which the signal from the RIP assay is significantly enriched over the signal from the mock controls (see Methods). Regions were defined based on a maximum family-wise error rate (FWER) < 0.05 (see Methods[1]). Given that RIP assays are known to show considerable variability (with typical reproducibility of ~60%[12]), we performed several biological replicates for each cell type. We observed that ~76% of the genes detected in one replicate are also detected in a second replicate (hLF: 70%, hFF: 75%, HeLa: 83%).

As a positive control, we checked whether HOTAIR and XIST were detectably co-precipitated in our RIP-Chip data. Consistent with previous reports, HOTAIR co-precipitated with PRC2 in both HeLa and foot fibroblasts, but not in lung fibroblasts. Similarly, XIST, which is expressed only in female cells, was detectably co-precipitated in the hLF cells (which came from a female source) but not the hFF cells (which came from a male source) (**Figure 1**). These results were consistent across all replicates.

In addition to the RIP assay, we also assayed expression patterns of lincRNAs and protein-coding genes on the custom exon-tiling array. We extracted total RNA from the same three human cell types (HeLa, hLF, hFF), prepared poly(A^+)-amplified cDNA and hybridized the product to the exon-tiling array. Of the lincRNA genes on the array, we found that 47% were detectably expressed in at least one of the three cell types (HeLa: 25%; hLF: 37%; and hFF:

33%). Consistent with the design of the tiling array, essentially all of the protein-coding genes were detectably expressed in the relevant cell type.

Analysis of the RIP-Chip results in conjunction with the expression analysis suggests that a significant proportion of all lincRNAs expressed in one of these 3 cell types are physically associated with PRC2. Specifically, we find that ~30% of expressed lincRNAs are detected in at least one of the replicates. As a conservative estimate, we only considered lincRNAs detected in at least two replicates. Using this criterion, we observe that 24% of lincRNAs (114 of 469) expressed in one of the three cell types is detected as physically associated with PRC2 (**Figure 1**).

As an independent validation of the association with PRC2, we selected five lincRNAs that were detected in our RIP-Chip data as associated with PRC2 in both HeLa and hFF and performed RIP-qPCR assays for these transcripts, using quantitative RT-PCR. In all 10 tests (five lincRNAs in two cell types), the results were confirmed. Notably, the RIP-qPCR assays showed a higher degree of enrichment than the RIP-Chip assays – consistent with the fact that arrays have a narrower dynamic range[22].

As a validation that the associations of lincRNAs with PRC2 are specific, we tested whether the enrichment in the RIP-Chip experiment was simply a reflection of transcript abundance (which would suggest non-specific interactions). We found no significant correlation between transcripts levels of the lincRNAs and their level of PRC2-enrichment (r = -0.109, p >.99).

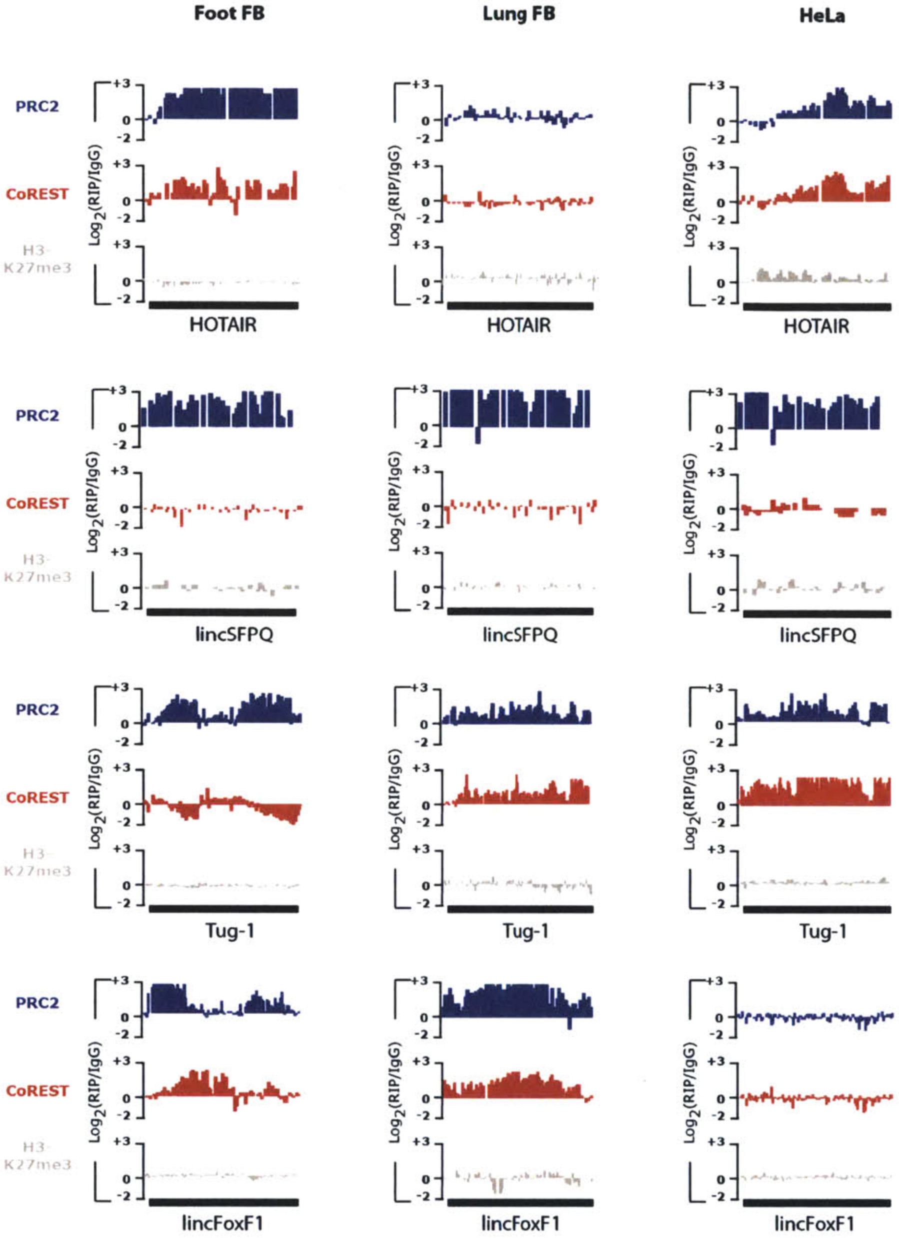


Figure 1: Numerous lincRNAs are physically associated with chromatin-modifying complexes. (A) Several examples of lincRNA exons (black box) that are enriched in RIP experiments relative to the IgG control in hFF (left column), hLF (middle column) and Hela (right column) cells. lincRNAs were in enriched in RIP experiments performed with antibodies recognizing the chromatin modifying complexes: PRC2 (blue), CoREST (red), but not with antibodies recognizing the chromatin protein H3K27me3 (grey). Coprecipitated RNA for each antibody and for the respective control (IgG) was hybridized to the DNA tiling arrays. The hybridization values for each probe within a lincRNA exon are plotted as the log2 values for RIP hybirdization intensity divided by control (IgG) hybridization intensity.

As a second approach to assess the specificity of PRC2 binding to lincRNAs, we examined the proportion of mRNAs bound to PRC2. In sharp contrast to the lincRNAs, very few of the protein-coding genes assayed in the RIP-Chip experiment showed physical association with PRC2. Of the 1000 protein-coding genes represented on the array, only 16 (<2%) were detected in two replicates (**Figure 2a**); we suspect that many of these 16 cases are artefacts, because only a small proportion (less than 1% of expressed mRNAs) are detected in three replicates. The proportion of transcripts associated with PRC2 is thus much higher for lincRNAs than for protein-coding mRNAs. To demonstrate that this result is not simply due to a low concentration of protein-coding mRNAs in the nucleus, we compared the concentration of lincRNAs and mRNAs in the nucleus (see Methods). While lincRNAs tend to have greater abundance in the nucleus than mRNAs, we find that the distributions of nuclear abundance of lincRNAs and mRNAs show substantial overlap, with at least 25% of mRNAs being expressed at levels comparable to the 50th percentile level for lincRNAs.

We also reasoned that lincRNAs associated with PRC2 should have significant representation in the nucleus. To test this, we examined the abundance of lincRNAs in the nucleus, and we found that PRC2-bound lincRNAs show a significantly higher abundance in the nucleus than non-PRC2-bound lincRNAs (methods). We also performed fluorescent *in situ* hybridization (FISH) on HOTAIR, XIST and four novel lincRNAs detected as associated with PRC2. In all cases, the lincRNAs showed either exclusively nuclear or nuclear and cytoplasmic localization (**Figure 2b**).

Finally, we explored whether a lincRNA that is expressed in two cell types (A and B) and associated with PRC2 in one cell type (A) is also associated with PRC2 in the second cell type (B). Considering all pairs of cell types, we found that this was the case for ~85% of lincRNAs (Supplemental Table 3). Collectively, these results provide strong evidence that a substantial portion (20-30%) of lincRNAs are specifically bound by PRC2 (**Figures 1 and 2A**).

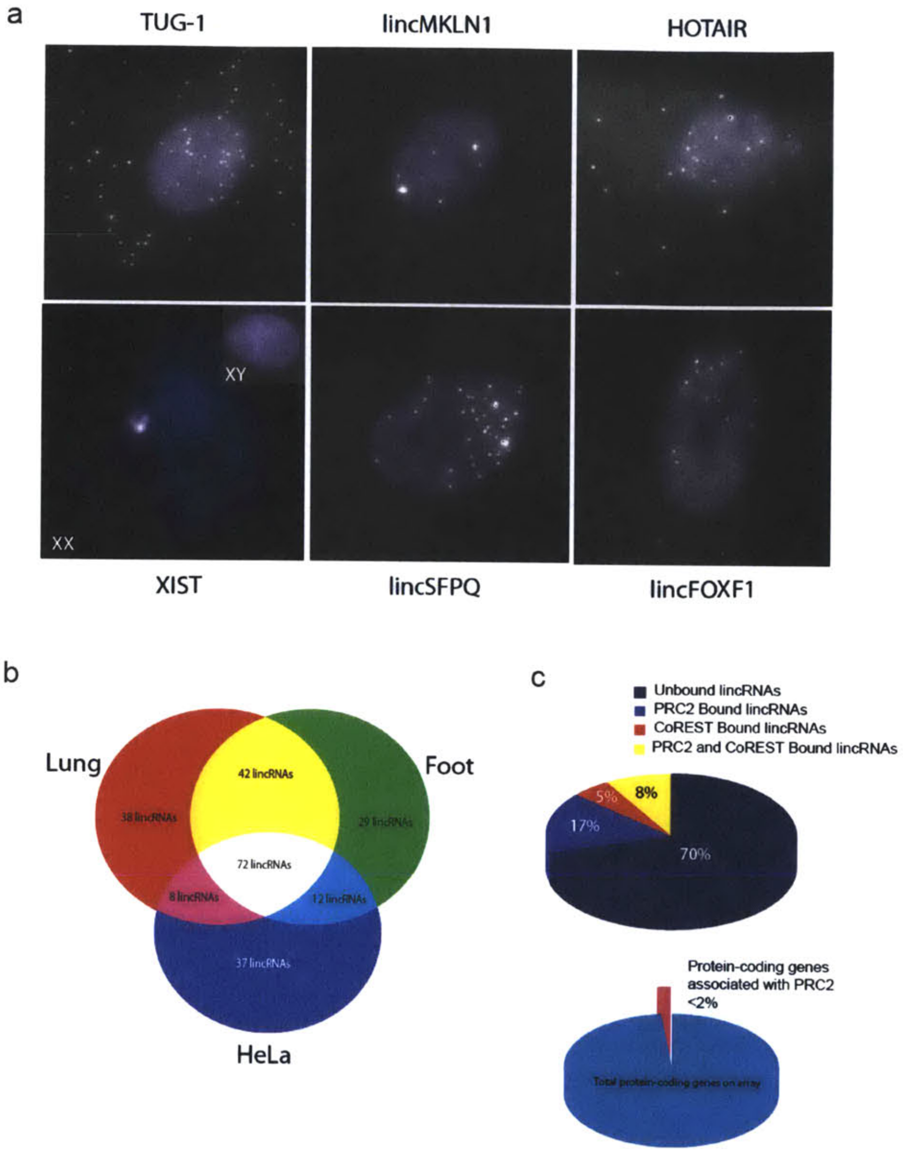


Figure 2: Diversity and nuclear localization of chromatin associated lincRNAs. (A) Subcellular localization analysis of lincRNAs by RNA fluorescence *in situ* hybridization (RNA FISH) demonstrates localization of a large majority of lincRNAs to the nucleus. Each panel represents the *in situ* hybridization of approximately 40 fluorescently labeled DNA oligos with complementarity to the interrogated lincRNA. RNA FISH experiments were performed in male hFF for each represented lincRNA (XIST, HOTAIR, TUG-1, lincMKLN-1, lincFOXF1, and lincSFPQ) and also in female hLF for XIST (XX). White 'speckles' indicate the subcellular localization of each lincRNA. The nuclear compartment is demarked by DAPI staining (purple). **(B)** Venn diagrams demonstrating the number of lincRNAs bound to PRC2 in hLF (red), hFF (green) and Hela (blue) cells. **(C)** Pie charts indicating many lincRNAs, but not protein-coding genes are physically associated with chromatin-modifying complexes. Top: pie chart representing the percentage of unique lincRNAs expressed in all three tested cell types (hFF, hLF and Hela) that are bound only to PRC2 (red), only to CoREST (green), bound by both PRC2

and CoREST (yellow) and those not bound by either complex (black). The percentage is calculated by adding the number of unique lincRNAs bound by each complex or those bound by both complexes in all three cell types and dividing by the total number of lincRNAs expressed in hLF, hFF and Hela cells. Bottom: pie chart indicating the percentage of protein-coding genes (black) reproducibly bound to PRC2 and or CoREST in all three cell types relative to the total number of expressed protein-coding genes (grey).

Association of lincRNAs with other chromatin-modifying complexes

Having found that many lincRNAs are associated with PRC2, we were interested to explore whether they might be associated with additional repressive chromatin-modifying complexes. We examined CoREST, a repressor of neuronal genes[23]. We performed RIP-Chip using an antibody against CoREST in the same three cell types (HeLa, hLF, and hFF). Applying the analysis above, we found that 63 of the 469 lincRNAs expressed in HeLa cells were reproducibly detected as bound to CoREST. As with PRC2, <2% of protein-coding genes co-precipitated with CoREST.

We note that about 60% of the lincRNAs associated with CoREST are not associated with PRC2 in HeLa cells, indicating that each complex has specific lincRNAs associated with it. The observation that 40% of the lincRNAs associated with CoREST are also associated with PRC2 may reflect the fact that the two complexes are known to share some regulatory targets [23,24].

Considering PRC2 and CoREST together, we find that ~38% of lincRNAs expressed in at least one of the cell types examined are reproducibly bound to at least one of the two complexes (180 of 469 expressed). This proportion is likely to be an underestimate, because we only count lincRNAs that were detected in at least two replicates; the proportion could be as high as 52%. These results raise the possibility that lincRNAs may be associated with additional chromatin-modifying proteins.

We also tested whether chromatin proteins themselves (rather than chromatin-modifying proteins) are associated with lincRNAs. Specifically, we performed RIP-Chip with antibodies against the modified histones H3K27me3 and H3K4me2. We found no significant enrichment of lincRNAs (**Figure 1**). These findings are consistent with other studies that identified XIST to co-precipitate with PRC2, but not H3K27me3 despite their immediate nuclear proximity[16].

Functional evidence that lincRNAs act through the PRC2 pathway

Having found that a substantial fraction of lincRNAs are physically associated with PRC2, we sought evidence that they play a functional role in polycomb-mediated repression. Previous studies have shown that depletion of HOTAIR and XIST causes up-regulation of genes normally repressed by PRC2[12,15]. To test whether other lincRNAs have a similar effect, we studied HOTAIR and six additional lincRNAs found to be associated with PRC2. For each of these seven lincRNAs, we designed pools consisting of four small-interfering RNAs (siRNAs) targeting each lincRNA (see Methods). We also used standard control siRNA pools that do not correspond to any human sequence. We transfected the siRNA pools into hLF (three pools), hFF (three pools) or both (one pool), with each experiment performed in duplicate. We measured the level of lincRNA knock-down by qRT-PCR and compared the results to the control siRNA pool; we only used experiments in which we achieved >2-fold depletion.

We hybridized the total RNA from these experiments to standard gene-expression arrays to measure the resulting changes in gene expressions. Specifically, for each of the seven lincRNAs, we determined the gene sets (S_1, S_2, ... , S_7) that were up-regulated relative to the control siRNA pools (at a false discovery rate (FDR) < 0.1). These sets contained between 30 and 134 genes (**Figure 3a**). The sets of genes affected by each lincRNA did not show significant overlap suggesting that each lincRNA has distinct target sets. We searched for, but found no common motifs enriched among the upregulated genes for each lincRNA. However, given the small number of target genes and the inability to distinguish between direct and indirect targets this result may simply reflect the low statistical power in analyzing a relatively small set of genes. Additionally, no lincRNA knock-down significantly affected the expression level of nearby genes (a window of at least 10 genes in either direction) suggesting that these lincRNAs are not likely to function via a *cis-acting* mechanism. This suggests that influence on gene regulation by PRC2 associated lincRNAs is likely exerted by a *trans* mechanism, similarly to what we have previously shown for HOTAIR[12].

We then sought to determine whether the up-regulated gene sets were highly enriched in genes normally repressed by PRC2 in human fibroblasts. Toward this end, we analyzed published data [13] in which the investigators measured gene expression changes in human embryonic fibroblasts in response to depletion of three key components of PRC2 (EZH2, SUZ12 and EED-1) with short-hairpin RNAs (shRNAs). For each component, we ranked all genes based on their change in expression level; the ranked lists are similar for each of the three components. We then used Gene Set Enrichment Analysis (GSEA)[25] to test whether the gene sets upregulated in response to depletion of the seven lincRNAs (S_1, S_2, … , S_7) were enriched among the genes up-regulated in response to depletion of the PRC2 components. The resulting enrichments were highly significant (FDR < 0.001) for each of the seven lincRNAs and each of the three PRC2 components (21 analyses in all, **Figure 3a**). As a negative control we examined the genes affected by the shRNA-mediated depletion of YY1[26], a transcription factor associated with chromatin. In contrast to depletion of the lincRNAs, we found no significant enrichment of PRC2 target genes. These results show that depletion of lincRNAs associated with PRC2 causes changes in gene expression and these genes are strongly enriched for genes normally repressed by PRC2. This provides functional evidence that many lincRNAs likely function through their interaction with PRC2.

An example: TUG1 represses p53-dependant cell cycle regulation

Finally, we decided to focus on a specific PRC2-associated lincRNA, TUG1. TUG1 was originally identified as a transcript upregulated by taurine, and siRNA-based depletion of TUG1 in the developing mouse eye was found to block retinal development [27,28]; the mechanism by which TUG1 depletion produces this phenotype is unknown. In our study, we found that TUG1 is ubiquitously expressed in human and mouse cell types and tissues and is bound to PRC2 in all three of the cell types examined. Previously, we studied regulation of lincRNAs in response to DNA damage and found that TUG1 was among the 39 lincRNA specifically induced in p53-wild type but not p53-mutant cells [1] (**Figure 3c**). Moreover, the TUG1 promoter contains many highly conserved binding sites for p53 (**Figure 3d**).

We selected TUG1 as one of the seven lincRNAs above that we depleted with siRNA pools. Depletion of TUG1 led to significant upregulation of 71 genes, which were strongly enriched for those involved in cell-cycle regulation (regulation of mitosis, spindle formation and cell-cycle phasing, **Figure 3b**). TUG1 thus is induced by p53, binds to PRC2 and plays a role in repressing specific genes involved in cell-cycle regulation. Interestingly, p53 is well known to cause both activation and repression of many genes. While p53 has been shown to be a direct activator of many genes, the mechanism of p53-induced repression remains unknown. Our results suggest the intriguing hypothesis that TUG1, and perhaps other lincRNAs, may function as downstream repressors in p53-mediated gene regulation.

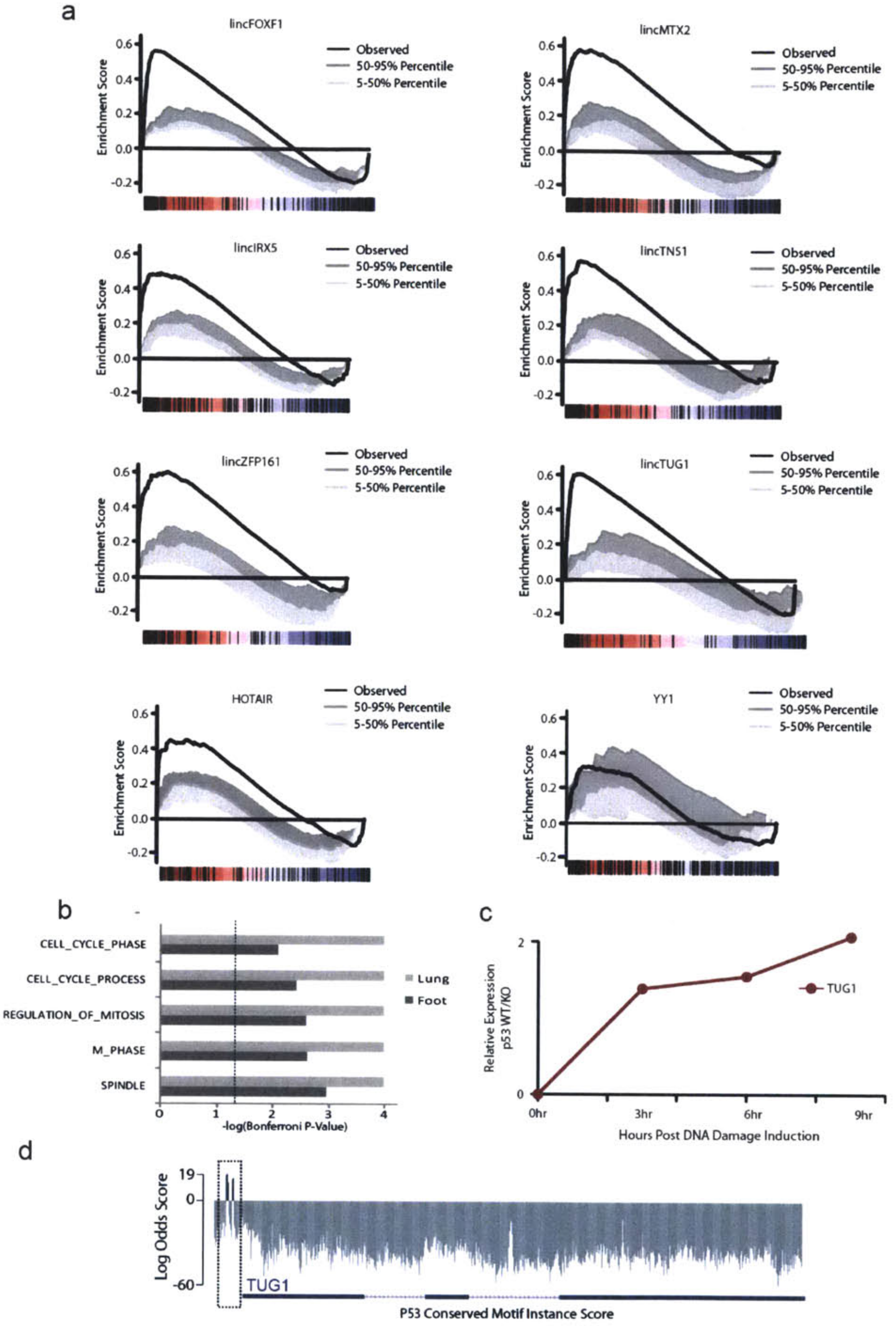


Figure 34: Genes repressed by PRC2 associated lincRNA overlap with genes repressed by PRC2. (A) Gene set enrichment analysis (GSEA) comparing the protein-coding genes that are upregulated upon depletion of a PRC2 bound lincRNA and those upregulated upon depletion of various components of PRC2. The black line represents the observed enrichment score profile of protein-coding genes in the lincRNA gene set to the PRC2 gene set. To represent the significance of the black line we permuted the enrichment score profiles for 100 random (size matched) gene sets. The dark grey region indicates distribution from the median to the 95th percentile and the light grey regions from the median to the 5th percentile; thus, results above the dark grey region

are significant at $p < 0.05$. The enrichment profiles for all lincRNAs tested were significant at $p < 0.05$, whereas as the enrichment profile for an unrelated protein depletion (YY-1) was not significant. The rank of each gene in the lincRNA gene set is indicated by tick marks (below each enrichment score plot) on a schematic color bar indicating levels of differential expression, upregulation in red and down regulation in blue. (B) Gene Ontology (GO) enrichment analysis identified numerous cell-cycle regulation pathways that were specifically derepressed upon knock down of lincRNA TUG1. The enrichment false discovery rate (FDR) is plotted as –log(FDR) on the x-axis. Results are shown from knockdown experiments in lung fibroblasts (grey) and in foot fibroblasts (black). Dashed line denotes FDR < 0.05. (C) lincRNA TUG1 is transcriptionally regulated by p53 in response to DNA damage. The y-axis indicates the log2 ratio of lincRNA TUG1 expression in p53 wild-type cells divided by the expression value in p53 knock-out cells. The x-axis indicates time after induction of DNA damage. (D) The lincRNA TUG1 promoter exhibits highly conserved p53 binding motifs (boxed region) whereas the transcriptional unit does not exhibit enrichment. The log odd conservation score (methods) is shown for the p53 binding motif at each position along the lincRNA TUG1 promoter.

Discussion

It is becoming clear that the mammalian genome encodes thousands of lincRNAs that are highly conserved and thus biologically functional[1]. Expression patterns suggest that these lincRNAs are involved in diverse biological processes, including cell cycle regulation, innate immunity, and ES pluripotency, but the mechanisms by which they play their roles were completely unknown.

Inspired by studies of the lincRNAs HOTAIR[12] and XIST[16], we investigated the idea that many lincRNAs are involved in the establishment of chromatin states. In this study, we report that a substantial proportion (24%) of lincRNAs expressed in a cell type are physically associated with the repressive chromatin modifying complex PRC2, and the proportion is even larger (38%) when additional chromatin modifying proteins (CoREST and SCMX) are included. It thus seems likely that significant fraction of lincRNAs will be associated with chromatin modifying proteins. Beyond the physical association, our functional analysis demonstrates that siRNA-mediated depletion of these lincRNAs results in preferential derepression of PRC2 regulated genes at distant loci, consistent with a *trans* acting mechanism. Together, these results suggest that many lincRNAs collaborate with chromatin modifying proteins to repress gene expression at specific loci.

There is a growing body of literature from yeast to mammals suggesting the non-coding RNAs play an important role in chromatin-state formation[29,30]. In *Schizosaccharomyces pombe*, a process known as RNA Induced Transcriptional Silencing (RITS) has been shown to play an important role in heterochromatin formation over centromeric repeats (reviewed in[31]). Similarly, short RNAs have been shown to play an important role in the establishment of heterochromatic silencing in plants. In *C. elegans*, genetic screens have identified polycomb homologs to be required for proper gene silencing in an RNA dependent manner[30]. In mammals, only a few specific RNAs (such as HOTAIR and XIST) have been implicated in directing chromatin modification. However, there is evidence that RNA plays a key role in shaping mammalian epigenetic landscapes. For example, depletion of single-stranded RNA (ssRNA) in mouse fibroblasts inhibits global heterochromatin formation[32]. Similarly, ssRNA but not ssDNA is required for the maintenance of the histone modifications H3K27me3 and H3K9me3[33].

Our results suggest an intriguing hypothesis: that lincRNAs bind to chromatin-modifying complexes to guide them to specific locations in the genome. Whereas chromatin-modifying proteins are often ubiquitously expressed, they establish epigenetic states that differ markedly among cell types and conditions[11]. Under our model, differentially expressed lincRNAs could bind to these complexes and help establish cell type specific epigenetic states. In particular, the PRC2 complex is involved in establishing repressive chromatin states involving H3K27me3. Together, PRC2 and a lincRNA might play the role of a transcriptional repressor by directing silencing to specific loci.

Such a mechanism could function within a larger regulatory program. Specifically, a newly induced transcription factor might establish a particular cellular state by (i) directly activating some downstream genes and (ii) activating lincRNAs that (with PRC2) repress genes involved in a previous or competing cellular state. Our observations concerning the lincRNA TUG1 suggest that it may function in such a program. Upon DNA damage, TUG1 is induced in a p53-dependent manner, likely through direct binding of p53, in view of many p53-binding sites in its promoter. It then binds PRC2 (based on our RIP-Chip data) and is involved in repressing important cell-cycle related genes (based on siRNA-based depletion of TUG1). Thus, we speculate that TUG1 may serve as a downstream transcriptional repressor in the p53 pathway to repress cell cycle progression in response to DNA damage.

Similarly, we have recently shown that HOTAIR serves as a transcriptional repressor of HOXD genes. We now know that HOXA13, the key distal regulator, directly transcribes HOTAIR to establish positional identity by repressing the appropriate HOX clusters[34]. Thus, HOTAIR serves as a downstream repressor in the HOXA13 transcriptional network (Presser et al. in preparation). This model raises many mechanistic questions, including (i) whether most lincRNAs associated with chromatin modifying complexes *directly* guide the complexes to specific loci and (ii) if so, how the guidance is accomplished (for example, by direct base pairing at specific sequence motifs). Future studies are needed to resolve the mechanism.

Our experiments have focused on chromatin-modifying complexes that add repressive chromatin marks. It is possible that many additional lincRNAs are associated with chromatin-modifying complexes that confer activating modifications, as has been recently reported in a few cases [19,20]. These questions can be addressed by performing RIP experiments with a wide range of antibodies across a wide range of cell types, to create a catalog of lincRNA-protein interactions.

Finally, while we have found that a substantial proportion of lincRNAs are associated with repressive chromatin modifying complexes, we do not mean to suggest that all lincRNAs necessarily function in this manner. There may be classes of lincRNAs that function in entirely different ways. For example, the lincRNAs NEAT1 and NEAT2 have been recently shown to be important in the formation of the nuclear speckle[35,36], and the lincRNA NRON plays a role in repressing nuclear import[37]. It is possible that additional lincRNAs play roles in these and numerous other cellular pathways. The full range of biological diversity of lincRNAs and their mechanisms clearly remains to be explored.

References

1 Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **in press** (2009).

2 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

3 Mattick, J. S. The genetic signatures of noncoding RNAs. *PLoS Genet* **5**, e1000459 (2009).

4 Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297 (2004).

5 Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849-1851 (2008).

6 Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563 (2005).

7 Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916-919 (2002).

8 Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**, 556-565 (2007).

9 Wang, J. *et al.* Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature* **431**, 1 p following 757; discussion following 757 (2004).

10 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).

11 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).

12 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323 (2007).

13 Bracken, A. P., Dietrich, N., Pasini, D., Hansen, K. H. & Helin, K. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes & development* **20**, 1123-1136 (2006).

14 Ku, M. *et al.* Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* **4**, e1000242 (2008).

15 Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet* **30**, 167-174 (2002).

16 Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750-756 (2008).

17 Nagano, T. *et al.* The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**, 1717-1720 (2008).

18 Pandey, R. R. *et al.* Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell* **32**, 232-246 (2008).

19 Dinger, M. E. *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* **18**, 1433-1445 (2008).

20 Sanchez-Elsner, T., Gou, D., Kremmer, E. & Sauer, F. Noncoding RNAs of trithorax response elements recruit Drosophila Ash1 to Ultrabithorax. *Science* **311**, 1118-1123 (2006).

21 Kirmizis, A. *et al.* Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes & development* **18**, 1592-1605 (2004).

22 Rinn, J. L., Bondre, C., Gladstone, H. B., Brown, P. O. & Chang, H. Y. Anatomic demarcation by positional variation in fibroblast gene expression programs. *PLoS genetics* **2**, e119 (2006).

23 Andres, M. E. *et al.* CoREST: a functional corepressor required for regulation of neural-specific gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 9873-9878 (1999).

24 Abuhatzira, L., Makedonski, K., Kaufman, Y., Razin, A. & Shemer, R. MeCP2 deficiency in the brain decreases BDNF levels by REST/CoREST-mediated repression and increases TRKB production. *Epigenetics* **2**, 214-222 (2007).

25 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

26 Affar el, B. *et al.* Essential dosage-dependent functions of the transcription factor yin yang 1 in late embryonic development and cell cycle progression. *Molecular and cellular biology* **26**, 3565-3581 (2006).

27 Altshuler, D., Lo Turco, J. J., Rush, J. & Cepko, C. Taurine promotes the differentiation of a vertebrate retinal cell type in vitro. *Development (Cambridge, England)* **119**, 1317-1328 (1993).

28 Young, T. L., Matsuda, T. & Cepko, C. L. The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr Biol* **15**, 501-512 (2005).

29 Mattick, J. S. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep* **2**, 986-991 (2001).

30 Bernstein, E. & Allis, C. D. RNA meets chromatin. *Genes Dev* **19**, 1635-1655 (2005).

31 Moazed, D. Small RNAs in transcriptional gene silencing and genome defence. *Nature* **457**, 413-420 (2009).

32 Maison, C. *et al.* Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nat Genet* **30**, 329-334 (2002).

33 Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326 (2006).

34 Rinn, J. L. *et al.* A dermal HOX transcriptional program regulates site-specific epidermal fate. *Genes & development* **22**, 303-307 (2008).

35 Sunwoo, H. *et al.* MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res* **19**, 347-359 (2009).

36 Clemson, C. M. *et al.* An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* **33**, 717-726 (2009).

37 Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**, 1570-1573 (2005).

Chapter 5: lincRNAs function in the molecular circuitry controlling pluripotency and differentiation in embryonic stem cells

In this chapter, we describe a systematic loss-of-function study, which demonstrated that lincRNAs play a clear functional role in the cell and that many lincRNAs play an essential role in maintaining the pluripotent cell and repressing differentiation programs.

Parts of this work were first published as:
Guttman M, Donaghey J, Carey BW, Garber M, Grenier J, Munson G, Young G, Lucas AB, Ach R, Bruhn L, Yang X, Amit I, Meissner A, Regev A, Rinn JL, Root DE, Lander ES. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. ***Nature*** 477(7364):295-300

Guttman M and Rinn JL. (2012). Modular regulatory principles of large non-coding RNAs. ***Nature*** 482(7385):339-46

While thousands of large intergenic non-coding RNAs (lincRNAs) have been identified in mammals, few have been functionally characterized leading to debate about their biological role. To address this, we performed loss-of-function studies on most lincRNAs expressed in mouse embryonic stem cells (ESC) and characterized the effects on gene expression. Here we show that knockdown of lincRNAs have major consequences on gene expression patterns, comparable to knockdown of well-known ESC regulators. Notably, lincRNAs primarily affect gene expression in *trans*. We identify dozens of lincRNAs whose knockdown causes an exit from the pluripotent state or upregulation of lineage commitment programs. We integrate lincRNAs into the molecular circuitry of ESCs and show that lincRNA genes are regulated by key transcription factors and that lincRNA transcripts physically bind to multiple chromatin regulatory proteins to affect shared gene expression programs. Together, the results demonstrate that lincRNAs have key roles in the circuitry controlling ESC state.

INTRODUCTION

The mammalian genome encodes many thousands of large non-coding transcripts[1-4] including a class of ~3500 large intergenic ncRNAs (lincRNAs) identified using a chromatin signature of actively transcribed genes[5-7]. These lincRNA genes have been shown to have interesting properties, including clear evolutionary conservation[5-9], expression patterns correlated with various cellular processes[5,10-12] and binding of key transcription factors to their promoters[5,11-13], and the lincRNAs themselves physically associate with chromatin regulatory proteins[7,10,14-16]. Yet, it remains unclear whether the RNA transcripts themselves have biological functions[8,17,18]. Few have been demonstrated to have phenotypic consequences by loss-of-function experiments[11]. As a result, the functional role of lincRNA genes has been widely debated. Various proposals include that lincRNA genes act as enhancer regions, with the RNA transcript simply being an incidental by-product[17-19], that lincRNA transcripts act in *cis* to activate transcription[20], and that lincRNA transcripts can act in *trans* to repress transcription[13,14].

We therefore sought to undertake systematic loss-of-function experiments on all lincRNAs known to be expressed in mouse embryonic stem cells (ESCs)[5,6]. ESCs are pluripotent cells that can self-renew in culture and can give rise to cells of any of the three primary germ layers including the germline[21]. The signalling[21-23] , transcriptional[24-29], and chromatin[24,30-34] regulatory networks controlling pluripotency have been well characterized providing an ideal system to determine how lincRNAs may integrate into these processes.

Here we show that knockdown of the vast majority of ESC-expressed lincRNAs has a strong effect on gene expression patterns in ESCs, of comparable magnitude to that seen for the well-known ESC regulatory proteins. We identify dozens of lincRNAs that upon loss-of-function cause an exit from the pluripotent state and dozens of additional lincRNAs that, while not essential for the maintenance of pluripotency, act to repress lineage-specific gene expression programs in ESCs. We integrate the lincRNAs into the molecular circuitry of ESCs by demonstrating that most lincRNAs are directly regulated by critical pluripotency-associated transcription factors and ~30% of lincRNAs physically interact with specific chromatin regulatory proteins to affect gene expression. Together, these results demonstrate a regulatory network in ESCs whereby transcription factors directly regulate the expression of lincRNA genes, many of which can physically interact with chromatin proteins, affect gene expression programs, and maintain the ESC state.

RESULTS

Functional effects of lincRNAs on gene expression

To perform loss-of-function experiments on lincRNAs, we generated five lentiviral-based shRNAs[35] targeting each of the 237 lincRNAs previously identified in ESCs[5,6] (see Methods).

These shRNAs successfully targeted 147 lincRNAs and reduced their expression by an average of ~75% compared to endogenous levels in ESCs (see Methods, **Figure 1a**). As positive controls, we generated shRNAs targeting ~50 genes encoding regulatory proteins, including both transcription factor and chromatin factor genes that have been shown to play critical roles in ESC regulation[29,32,36]; we obtained validated hairpins against 40 of these genes. As negative controls, we performed independent infections with lentiviruses containing 27 different shRNAs with no known cellular target RNA.

We then studied the effects of knocking down each lincRNA on global transcription. We infected each shRNA into ESCs, isolated RNA after 4 days, and profiled their effects by hybridization to genome-wide microarrays (**Figure 1a**, see Methods). We employed a stringent procedure to control for non-specific effects due to viral infection, generic RNAi responses, or 'off-target' effects. Expression changes were deemed significant only if they exceeded the maximum levels observed in any of the negative controls, showed a two-fold change in expression compared to the negative controls, and had a low false discovery rate (FDR) assessed across all genes based on permutation tests (**Figure 1b**, see Methods). This approach controls for the overall rate of non-specific effects by estimating the number and magnitude of observed effects in the negative control hairpins, where all effects are non-specific.

For 137 of the 147 lincRNAs (93%), knockdown caused a significant impact on gene expression, with an average of 175 protein-coding transcripts affected (range: 20-936) (**Figure 1c**). These results were similar to those obtained upon knockdown of the 40 well-studied ESC regulatory proteins: 38 (95%) showed significant effects on gene expression, with an average of 207 genes affected (range: 28 (for DNMT3L) to 1187 (for Oct4)) (**Figure 1c**). Although some individual lincRNAs have been found to lead primarily to gene repression[13,14], we find that knockdown of the lincRNAs studied here largely led to comparable numbers of activated and repressed genes. To further assess whether the expression changes were due to 'off-target' effects, we also profiled the effects of the second-best validated shRNA targeting 10 randomly selected lincRNA genes. In all cases, second shRNAs against the same target produced significantly similar expression changes (see Methods). Together, these results indicate that the vast majority of lincRNAs have functional consequences on overall gene expression of comparable magnitude (in terms of number of affected genes and impact on levels) to the known transcriptional regulators in ESCs.

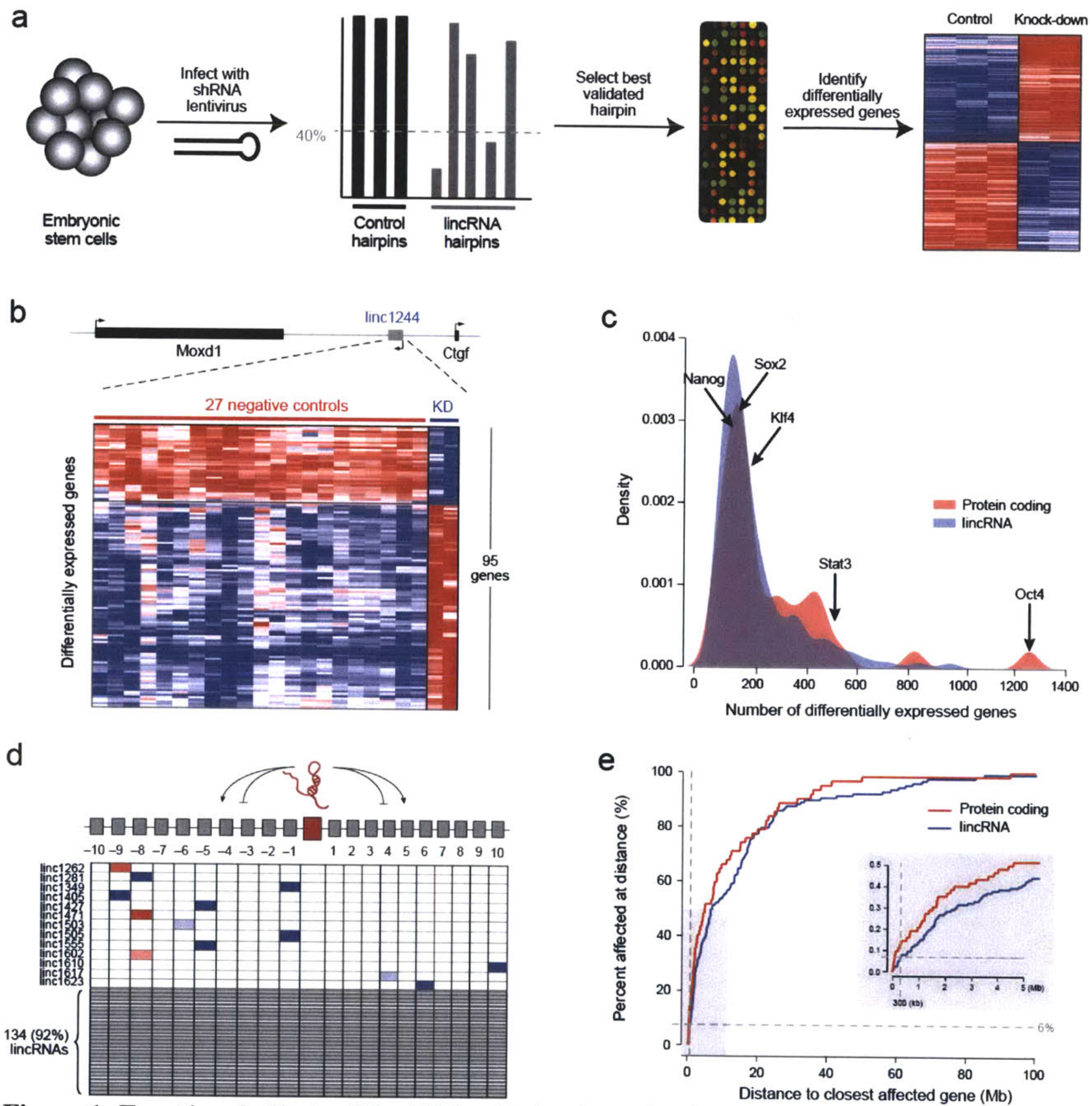


Figure 1: Functional effects of lincRNAs. (a) A schematic of lincRNA perturbation experiments. ESCs are infected with shRNAs targeting lincRNA genes. The knockdown level is computed compared to negative control hairpins. The best hairpin is selected and RNA levels are profiled on expression arrays. Differential gene expression is computed relative to negative control hairpins. (b) Representative example of a lincRNA knockdown effect on gene expression. Top: Genomic locus containing a lincRNA. Bottom: Heatmap of the 95 genes affected by knockdown of the lincRNA, expression for control hairpins (red line) and expression for lincRNA hairpins (blue line) are shown. (c) Distribution of number of affected genes upon

knockdown of 147 lincRNAs (blue) and 40 well-known ESC regulatory proteins (red). Points corresponding to five specific ESC regulatory proteins are marked. (d) Effects of knockdown of 13 lincRNAs on the 10 neighbour genes on each side (downregulation in blue, upregulation in red). For the remaining 134 lincRNA genes, no neighbouring genes are affected. (e) Distance to the closest affected gene upon knockdown of a lincRNA (blue) or protein-coding gene (red). Grey Inset: A close-up of the region from 0–5 Mb. The grey dashed line represents a distance of 300 kb in both panels.

lincRNAs affect gene expression in *trans*

The discovery of XIST as a ncRNA immediately suggested a model for how it can act in an allele specific manner[37]. In theory, a ncRNA possess intrinsic *cis* regulatory capacity since it can function while remaining tethered at its own locus[37,38] whereas an mRNA must be dissociated, exported, and translated to function. In this review, we define a *cis* regulator as one that exerts its function on a neighboring gene on the same allele from which it is transcribed and a *trans* regulator if it does not meet this criteria. Because of the unique *cis* regulatory capability of ncRNAs, it has been speculated that *cis* regulation may be a common mechanism for large ncRNAs[20,38]. However, global functional evidence strongly suggests that this is not the case (**see Figure 2**).

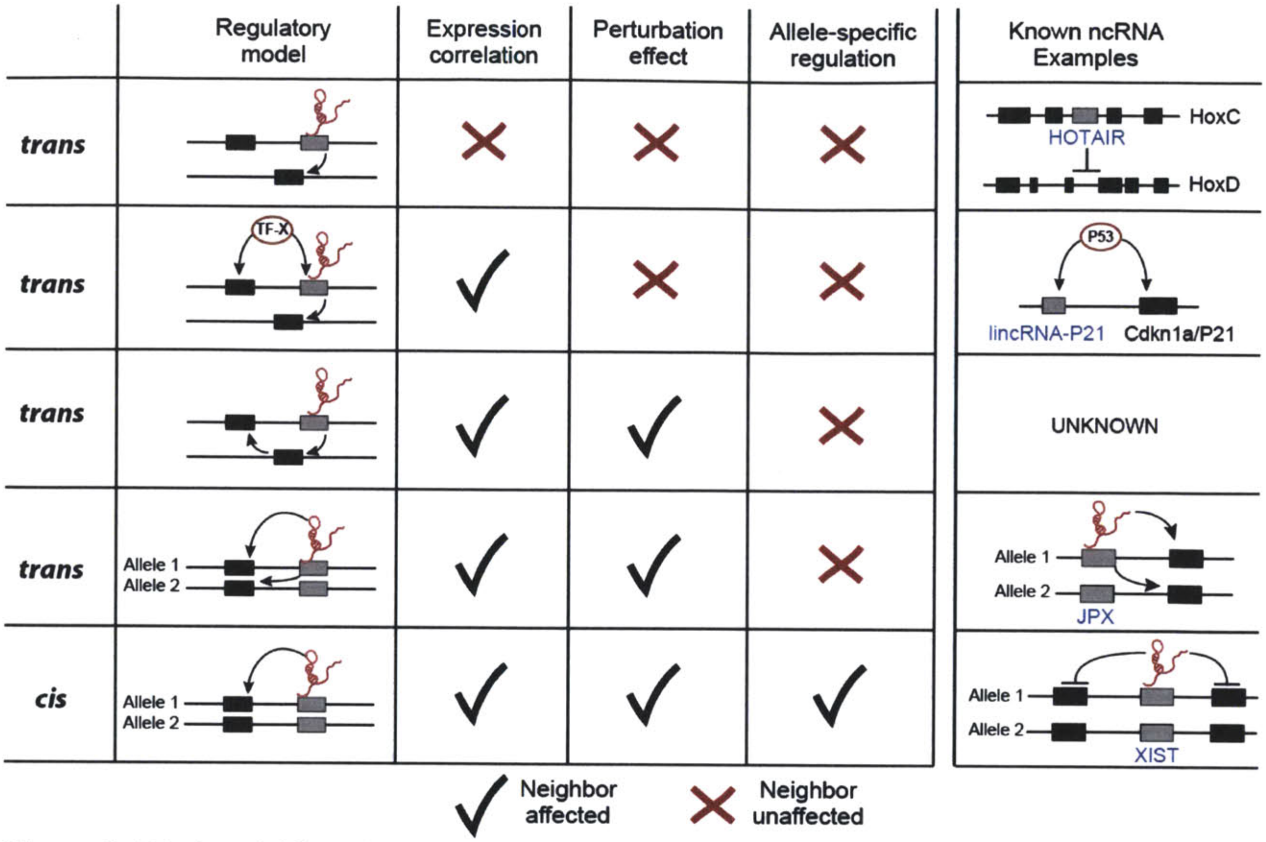
To distinguish *cis* and *trans* regulatory models, initial studies used correlation analysis and identified significant correlations of expression between ncRNAs and their neighboring protein-coding genes[19,39]. However, several of these cases have been demonstrated to be *trans* models and the apparent correlations are due to shared upstream regulation (such as lincRNA-p21[13] and lincRNA-Sox2[5]), positional correlation (such as HOTAIR[14]), transcriptional 'ripple effects'[19], and indirect regulation of neighboring genes (**Figure 2**). Consistent with these explanations, a recent study showed that the increased correlation of expression between ncRNAs and neighboring genes is comparable to that observed for protein-coding genes[40].

Some recent papers have claimed that most lincRNAs act primarily in *cis*[8,17,20,39]. We found no evidence to support this latter notion: knockdown of only 2 lincRNAs showed effects on a neighbouring gene, only 13 showed effects within a window of ten genes on either side (**Figure 1d**), and only 8 showed effects on genes within 300 kb (**Figure 1e**); these proportions are no greater than observed for protein-coding genes (**Figure 1e**). In short, lincRNAs appear to affect expression largely in *trans*.

Our results contrast with a recent study that concluded that lincRNAs act in *cis*, based on the observation that knockdown of 7 out of 12 lincRNAs affected expression of a gene within 300Kb[20]. The explanation appears to be that the threshold for significant changes in gene expression used in the study failed to account for multiple hypothesis testing within the local region. Accounting for this, the effects on neighbouring genes are no greater than expected by chance and are consistent with our observations here (see Methods).

While perturbation experiments can demonstrate that an RNA works in *trans,* evidence that an RNA works in *cis* is more difficult to obtain (**see Figure 2**). As an example, perturbation experiments demonstrate that the JPX ncRNA affects the expression of the neighboring XIST gene yet it was demonstrated to perform this role by acting in *trans*[41]. The ultimate proof of *cis* regulation requires demonstrating that an RNA regulates a neighbouring gene on the *same allele* (**Figure 2**). To date, few studies have performed such tests and it is unclear what percentage of ncRNAs suggested to act in *cis* by loss-of-function experiments[20,42] will pass this test.

While it is clear that some lincRNAs can regulate gene expression in *cis*[20,42-44], determining the precise proportion of *cis* regulators requires more direct experimental approaches. We note that our results are consistent with observed correlations between lincRNAs and neighbouring genes[5,39], which may represent shared upstream regulation[5,13] or local transcriptional effects[19,45]. In addition, the lincRNAs studied here should be distinguished from transcripts that are produced at enhancer sites[17,18], the function of which has yet to be determined.



	Regulatory model	Expression correlation	Perturbation effect	Allele-specific regulation	Known ncRNA Examples
trans		✗	✗	✗	HoxC HOTAIR HoxD
trans	TF-X	✓	✗	✗	P53 lincRNA-P21 Cdkn1a/P21
trans		✓	✓	✗	UNKNOWN
trans	Allele 1 Allele 2	✓	✓	✗	Allele 1 Allele 2 JPX
cis	Allele 1 Allele 2	✓	✓	✓	Allele 1 Allele 2 XIST

✓ Neighbor affected ✗ Neighbor unaffected

Figure 2: Distinguishing *cis* versus *trans* regulation. If a ncRNA is a *cis* regulator then several observations will be true: (i) the gene expression levels of a neighboring gene will be correlated with the RNA expression across conditions, (ii) loss-of-function of the RNA would affect expression of a neighboring gene, and (iii) the ncRNA would affect expression of a neighboring gene on the *same allele* that it is expressed from. The absence of any of these criteria supports regulation in *trans*. We illustrate this point using 5 common regulatory models (left), this table displays what would be observed using specific computational and experimental methods for each regulatory model. Check boxes (black) indicate observed effects on neighboring genes for each method and crosses (red) indicate no observed effects on neighboring genes. Known ncRNA examples of each of these regulatory models are shown to the right of the table.

lincRNAs are required to maintain the pluripotent state

We next sought to investigate whether lincRNAs play a role in regulating the ESC state. Regulation of the ESC state involves two components, maintaining the pluripotency program and repressing differentiation programs[24]. To determine whether lincRNAs play a role in the maintenance of the pluripotency program, we studied their effects on the expression of Nanog, a key transcription factor that is required to establish[46] and uniquely marks the pluripotent state[47,48]. We infected ESCs carrying a luciferase reporter gene expressed from the endogenous Nanog promoter[49] with shRNAs targeting lincRNAs or protein-coding genes. We monitored loss of reporter activity after 8 days relative to 25 negative control hairpins across biological replicates (see Methods). To ensure that the observed effects were not simply due to a reduction in cell viability, we excluded shRNAs that caused a reduction in cell numbers (see Methods). Altogether, we identified 26 lincRNAs that had major effects on endogenous Nanog levels with many at comparable levels to the knockdown of the known protein-coding regulators of pluripotency such as Oct4 and Nanog (**Figure 3a**). This establishes that these lincRNAs have a role in maintaining the pluripotent state.

To further validate the role of these 26 lincRNAs in regulating the pluripotent state, we knocked down these lincRNAs in wild-type ESCs and measured mRNA levels of Oct4 and Nanog after 8 days across biological replicates. For ~90% of these lincRNAs, we identified a significant decrease in both Oct4 and Nanog levels. For the 16 lincRNAs for which we had a second effective hairpin, we found comparable reductions in Oct4 expression levels upon knockdown using these hairpins (**Figure 3b**). Notably, >90% of lincRNA knockdowns affecting Nanog reporter levels led to loss of the ESC morphology (**Figure 3c**). In summary, inhibition of these 26 lincRNAs lead to an increased exit from the pluripotent state.

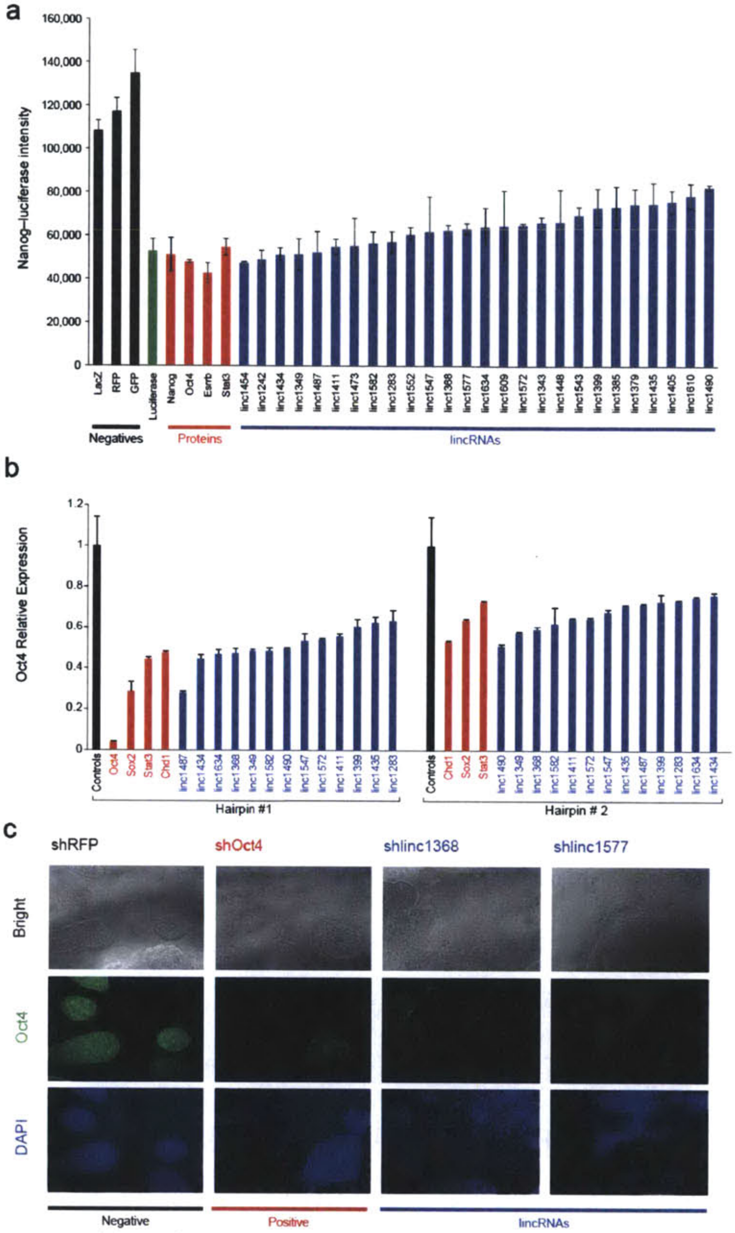


Figure 3: lincRNAs are critical for the maintenance of pluripotency. (a) Activity from a Nanog promoter driving luciferase, following treatment with control hairpins (black) or hairpins targeting luciferase (green), selected protein-coding regulators (red), and lincRNAs (blue). (b) Relative mRNA expression levels using qPCR of Oct4 following knockdown of selected protein-coding (red) and lincRNA (blue) genes affecting Nanog–luciferase levels. The best hairpin (black line) and second best hairpin (grey line) are shown. All knockdowns are significant with a p-value<0.001. Error bars represent standard error across replicate infections (n=4). (c) Morphology of ESCs and immunofluorescence staining of Oct4 for a negative control hairpin (black line), a hairpin targeting Oct4 (red line), and hairpins targeting two lincRNAs (blue line). The first row shows bright field images of infected ESCs. The second row shows immunofluorescence staining of the Oct4 protein and the third row shows DAPI staining of the nuclei.

lincRNAs repress lineage-specific gene expression programs

We next sought to explore the biological roles of ESC lincRNAs by classifying the overall gene expression patterns resulting from a lincRNA knockdown. To interpret the patterns, we compared them to a curated set of >100 publicly available gene expression profiles and signatures resulting from perturbations or differentiation of ESCs (see Methods)[50-54]. We mapped our observed profiles onto these previously identified states by comparing the expression changes induced by knockdown of the lincRNA with the expression changes in the previously studied states; we assessed significance using a permutation-derived FDR (see Methods)[55,56]. The states include differentiation into the endoderm, ectoderm, mesoderm, and trophectoderm lineages[50,51,53,54]. As a positive control for our analytical method, we confirmed the expected results that the expression pattern caused by Oct4 knockdown was strongly associated with the trophoectoderm lineage[57] and by Nanog knockdown with endoderm differentiation[48] (**Figure 4a**).

Using this approach, we identified 30 lincRNAs whose knockdown produced expression patterns similar to differentiation into specific lineages (**Figure 4a**). Amongst the lincRNAs associated with differentiation, 13 are associated with differentiation into the endoderm lineage, 7 with ectoderm differentiation, 5 with neuroectoderm differentiation, 7 with mesoderm differentiation, and 2 with the trophectoderm lineage (**Figure 4a**). Consistent with these functional assignments, we observe that the majority (>85%) of the 30 lincRNAs associated with specific differentiation lineages showed upregulation of the well-known marker genes for the identified states[29,53,54] upon knockdown (such as Sox17 (endoderm), Fgf5 (ectoderm), Pax6 (neuroectoderm), Brachyury (mesoderm), and Cdx2 (trophectoderm)) (**Figure 4b**).

The fact that knockdown of these 30 lincRNAs induce gene expression programs associated with specific early differentiation lineages suggests that these lincRNAs normally act as a barrier to such differentiation. Interestingly, most of the lincRNA knockdowns (~85%) that induce gene expression patterns associated with these lineages did not cause the cells to differentiate as determined by Nanog reporter levels (**Figure 3a**). This is consistent with observations for several critical ESC chromatin regulators, such as the polycomb complex; loss-of-function of these regulators similarly induce lineage-specific markers without causing differentiation[30,58,59].

Together, these data indicate that many lincRNAs play important roles in regulating the ESC state, including maintaining the pluripotent state and repressing specific differentiation lineages.

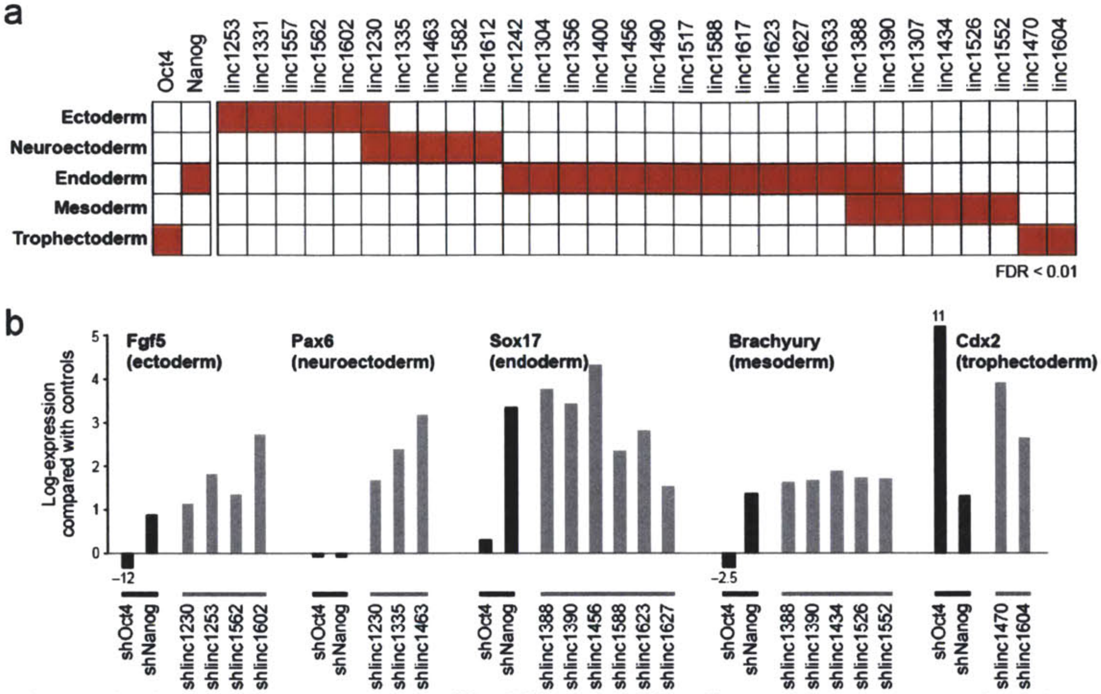


Figure 4: lincRNAs repress specific differentiation lineages. (a) Expression changes for each lincRNA compared to gene expression of five differentiation patterns. Shown are associations for Oct4 and Nanog (left) and for lincRNAs with a significant (FDR<0.01) association with these lineages (right). Each box shows significant positive association with known expression patterns (red). (b) Expression changes upon knockdown of Oct4 and Nanog (black bars) and representative lincRNAs (grey bars) for five lineage marker genes. The expression changes (FDR<0.05) are displayed on a log scale as the *t*-statistic compared to a panel of negative control hairpins for each lineage gene.

lincRNAs are direct regulatory targets of the ESC transcriptional circuitry

Having demonstrated a functional role for lincRNAs in ESCs, we sought to integrate the lincRNAs into the molecular circuitry controlling the pluripotent state. First, we explored how lincRNA expression is regulated in ESCs. Toward this end, we utilized published genome-wide maps of 9 pluripotency-associated transcription factors (TFs)[26,60] and determined whether they bind to the promoters of lincRNA genes expressed in ESCs. We observe that ~75% of the 237 lincRNA promoters are bound by at least one of 9 pluripotency-associated TFs (including Oct4, Sox2, Nanog, cMyc, nMyc, Klf4, Zfx, Smad, and Tcf3) with a median of 3 factors bound to each promoter (**Figure 5a**), comparable to the proportion reported for protein-coding genes[26]. Interestingly, the 3 core factors (Oct4, Sox2, and Nanog) bind to the promoters of ~12% of all lincRNAs and ~50% of lincRNAs involved in the regulation of the pluripotent state.

To determine if lincRNA expression is functionally regulated by the pluripotency-associated TFs, we used shRNAs to knock down the expression of 5 of the 9 pluripotency-associated TF genes for which we could obtain validated hairpins and profiled the resulting changes in lincRNA expression after 4 days. Upon knockdown of a TF, expression changes are seen at ~50% of lincRNAs genes whose promoters are bound by the TF (**Figure 5a**, bottom); the proportion is comparable to that seen for protein-coding genes whose promoters are bound by the TF. The strong but imperfect correlation between TF-binding and effect of TF-knockdown is consistent with previous observations[61-63] and may reflect regulatory redundancy in the pluripotency network[28,64]. In addition, we profiled the knockdown of an additional 7 pluripotency-associated transcription factors (including Esrrb, Zfp42, and Stat3). Altogether, for ~60% of the ESC lincRNAs, we identified a significant downregulation upon KD of one of these 11 TFs (**Figure 5b**).

We also characterized the expression of the ESC lincRNAs following retinoic-acid-induced differentiation of the ESCs. The ESC lincRNAs show temporal changes across the time course with ~75% showing a decrease in expression compared to untreated ESCs (**Figure 5c**).

Notably, all of the lincRNAs shown to regulate pluripotency are down-regulated upon retinoic acid treatment (**Figure 5c**). Our results establish that lincRNAs are direct transcriptional targets of the pluripotency-associated TFs and are dynamically expressed across differentiation. Collectively, these results demonstrate that lincRNAs are an important regulatory component within the ESC circuitry.

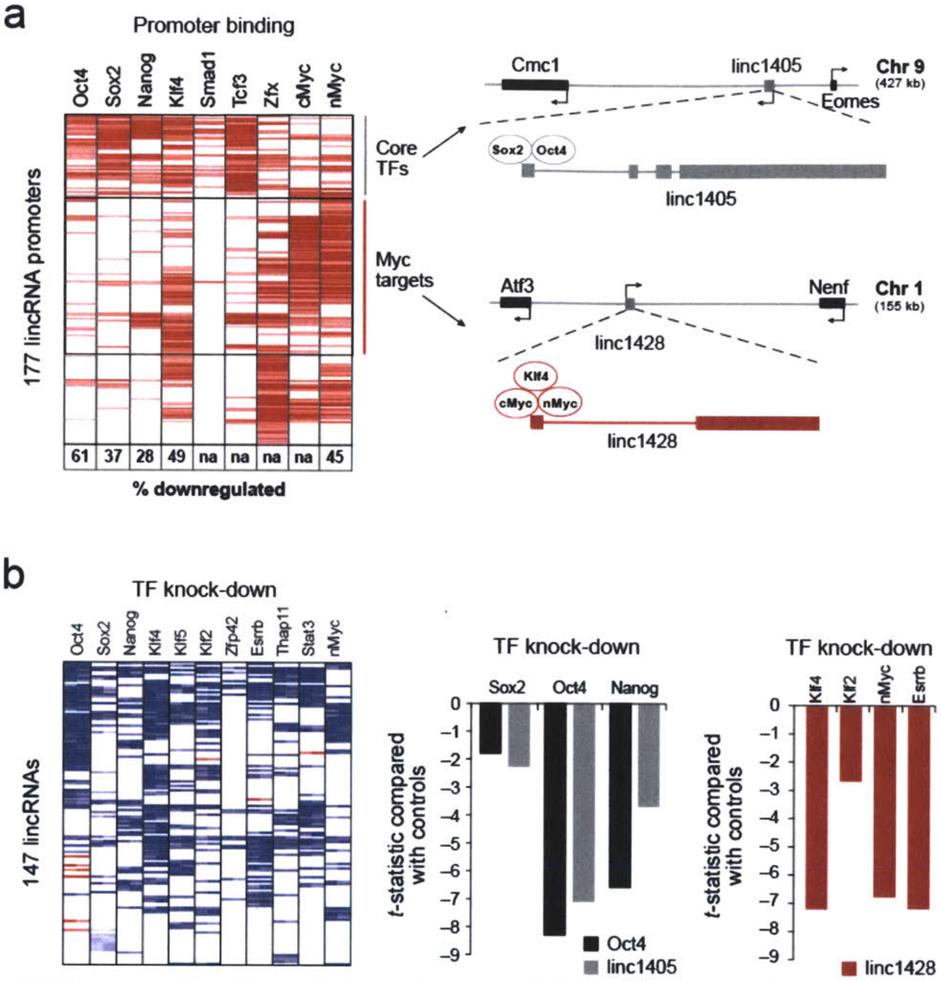


Figure 5: lincRNAs are direct regulatory targets of the ESC transcriptional circuitry. (a) A heatmap representing enrichment for lincRNA promoters (rows) by ChIP-Seq for 9 transcription factors (columns). The percentages of bound promoters downregulated upon knock-down of a TF, are indicated in boxes beneath the associated column ('na' were not measured). lincRNA promoters were grouped into two main clusters: one bound by Oct4, Sox2, and Nanog (Core regulated) and the other bound by cMyc and nMyc amongst other factors (Myc regulated). Right: Example lincRNAs in each cluster showing their genomic neighbourhood and TF binding. (b) A heatmap representing changes in lincRNA expression (rows) following knockdown of 11 TFs (columns), relative to negative control hairpins. Middle: Effect of knockdown of Sox2, Oct4 and Nanog on expression levels of linc1405 (gray) and Oct4 (black). Right: Effect of knockdown of Klf2, Klf4, nMyc, and Esrrb on expression levels of linc1428.

lincRNAs physically interact with diverse chromatin regulatory proteins

To explore how lincRNAs carry out their regulatory roles, we studied whether lincRNAs physically associate with chromatin modifying proteins in ESCs. We previously showed that many human lincRNAs can interact with the polycomb repressive complex[7], a complex that plays a critical functional role in the regulation of ESCs[30,31]. To determine whether the ESC lincRNAs physically associate with the polycomb complex, we crosslinked RNA-Protein complexes using formaldehyde, immunoprecipitated the complex using antibodies specific to both the Suz12 and Ezh2 components of Polycomb, and profiled the co-precipitated lincRNAs using a direct RNA quantification method[65] (see Methods). We performed immunoprecipitation of the Polycomb complex across 5 biological replicates and 8 mock-IgG controls, and we assessed significance using a permutation test (see Methods). Altogether, we identified 24 lincRNAs (~10% of the ESC lincRNAs) that were strongly enriched for both Polycomb components (**Figure 6b**).

To determine if lincRNAs interact with additional chromatin proteins, we systematically analysed chromatin-modifying proteins that have been shown to play critical roles in ESCs[30-34,66,67]. Specifically, we screened antibodies against 28 chromatin complexes (see Methods) and identified 11 additional chromatin complexes that are strongly and reproducibly associated with the lincRNAs (see Methods). These chromatin complexes are involved in 'reading' (PRC1, Cbx1, and Cbx3), 'writing' (Tip60/P400, PRC2, Setd8, ESET, and Suv39h1), and 'erasing' histone modifications (Jarid1b, Jarid1c, and HDAC1)[68], as well as a chromatin-associated DNA binding protein (YY1) (**Figure 6a**). Altogether, we found that 74 (~30%) of the ESC lincRNAs are associated with at least one of these 12 chromatin complexes (**Figure 6b**). While most of the identified interactions are with repressive chromatin regulators, this is likely due to limitations of our selection criteria and available antibodies.

Interestingly, we note that many lincRNAs are strongly associated with multiple chromatin complexes (**Figure 6b**). For example, we identified 8 lincRNAs that bind to the PRC2 H3K27 and ESET H3K9 methyltransferase complexes ('writers' of repressive marks) and the Jarid1c H3K4 demethylase complex (an 'eraser' of activating marks). Consistent with this, the PRC2 and ESET complexes have been reported to bind at many of the same 'bivalent' domains[33] and to functionally associate with the Jarid1c complex[69,70]. Similarly, we identified a distinct set of 17 lincRNAs that bind to the PRC2 complex ('writer' of K27 repressive marks), PRC1 complex ('reader' of K27 repressive marks), and Jarid1b complex ('eraser' of K4 activating marks) (**Figure 6c**), as well as other functionally consistent 'reader', 'writer', and 'eraser' combinations (**Figure 6c**). One of several potential models consistent with this data is that lincRNAs may bind to multiple distinct protein complexes, perhaps serving as 'flexible scaffolds' to bridge functionally related complexes as previously described for telomerase RNA[71].

To determine if the identified lincRNA-protein interactions have a functional role on gene expression, we examined the effects on gene expression resulting from knockdown of individual lincRNAs that are physically associated with particular chromatin complexes and from knockdown of genes encoding the associated complex itself (see Methods). For >40% of these lincRNA-protein interactions, we identified a highly significant overlap in affected gene expression programs compared to just ~6% for random lincRNA protein pairs (see Methods).

Other cases may reflect the limited power to detect the overlaps, because specific lincRNA-protein complexes may be related to only a fraction of the overall expression pattern mediated by the chromatin complex.

Together, these data suggest that many of the ESC lincRNAs physically associate with multiple different chromatin regulatory proteins and that many of these interactions are likely to be important for the regulation of gene expression programs.

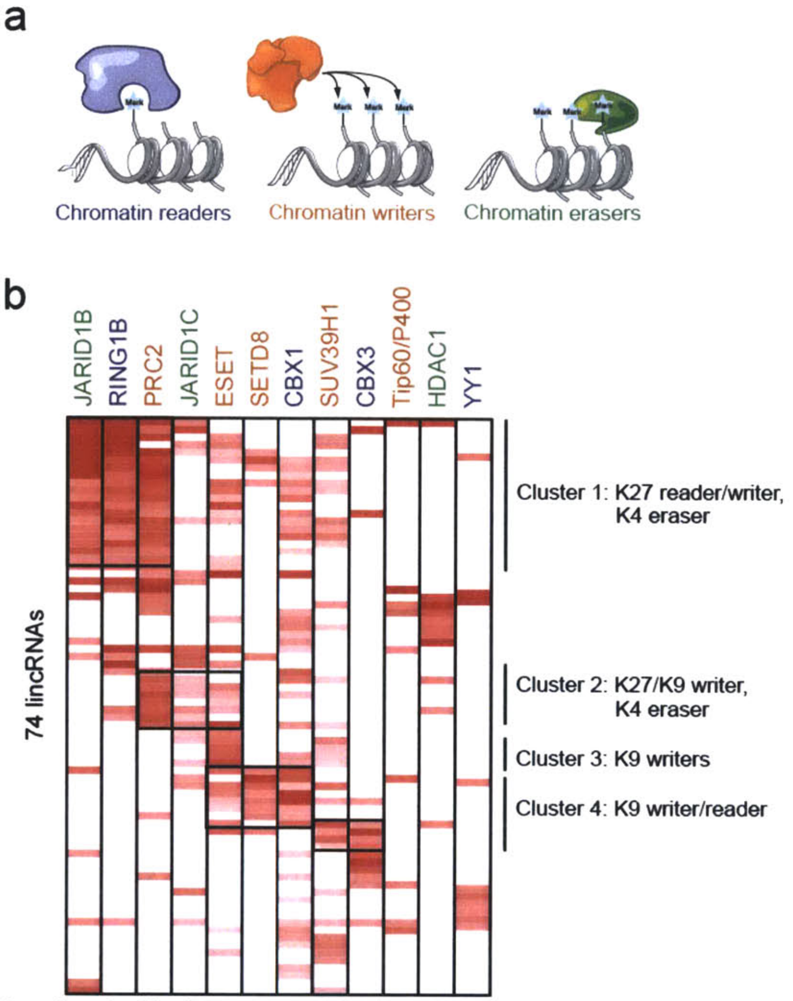


Figure 6: lincRNAs physically interact with chromatin regulatory proteins. (a) A schematic of the classes of chromatin regulators profiled: 'readers' (blue), 'writers' (orange), and 'erasers' (green). (b) A heatmap showing the enrichment of 74 lincRNAs (rows) for one of 12 chromatin regulatory complexes (columns). The names are color-coded by their chromatin-regulatory mechanism. Major clusters are indicated by vertical lines with a description of the chromatin components. (c) Examples of each cluster, enrichment levels for lincRNAs are shown for the indicated complexes. Enrichments are shown as the *t*-statistic compared to five mock-IGG controls. The colour of the bars corresponds to the chromatin regulatory mechanism of the immunoprecipitated protein.

DISCUSSION

While the mammalian genome encodes thousands of lincRNA genes, few have been functionally characterized. We performed an unbiased loss-of-function analysis of lincRNAs expressed in ESCs and show that lincRNAs are clearly functional and primarily act in *trans* to affect global gene expression. Our results establish that lincRNAs are key components of the ESC transcriptional network that are functionally important for maintaining the pluripotent state, and that many are down-regulated upon differentiation. The ESC lincRNAs physically interact with chromatin proteins, many of which have been previously implicated in the maintenance of the pluripotent state[30,32-34]. In addition to chromatin proteins, lincRNAs interact with other protein complexes including many RNA-binding proteins.

Our data suggests a model whereby a distinct set of lincRNAs is transcribed in a cell type and interacts with ubiquitous regulatory protein complexes to give rise to cell-type-specific RNA-protein complexes that coordinate cell-type specific gene expression programs (**Figure 7**). Because many of the lincRNAs studied here interact with multiple different protein complexes, one hypothesis is that they act as cell-type specific 'flexible scaffolds'[71,72] to bring together protein complexes into larger functional units (**Figure 7**). This model has been previously demonstrated for the yeast telomerase RNA[71] and suggested for the XIST[72-74] and HOTAIR[75] lincRNAs. The hypothesis that lincRNAs serve as flexible scaffolds could explain the uneven patterns of evolutionary conservation seen across the length of lincRNA genes[6]: the more highly conserved patches could correspond to regions of interaction with protein complexes.

While a model of lincRNAs acting as 'flexible scaffolds' is attractive, it is far from proven. Testing the hypothesis for lincRNAs will require systematic studies, including defining all protein-complexes with which lincRNAs interact, determining where these protein interactions assemble on RNA, and ascertaining whether they bind simultaneously or alternatively. Moreover, understanding how lincRNA-protein interactions give rise to specific patterns of gene expression will require determination of the functional contribution of each interaction and possible localization of the complex to its genomic targets.

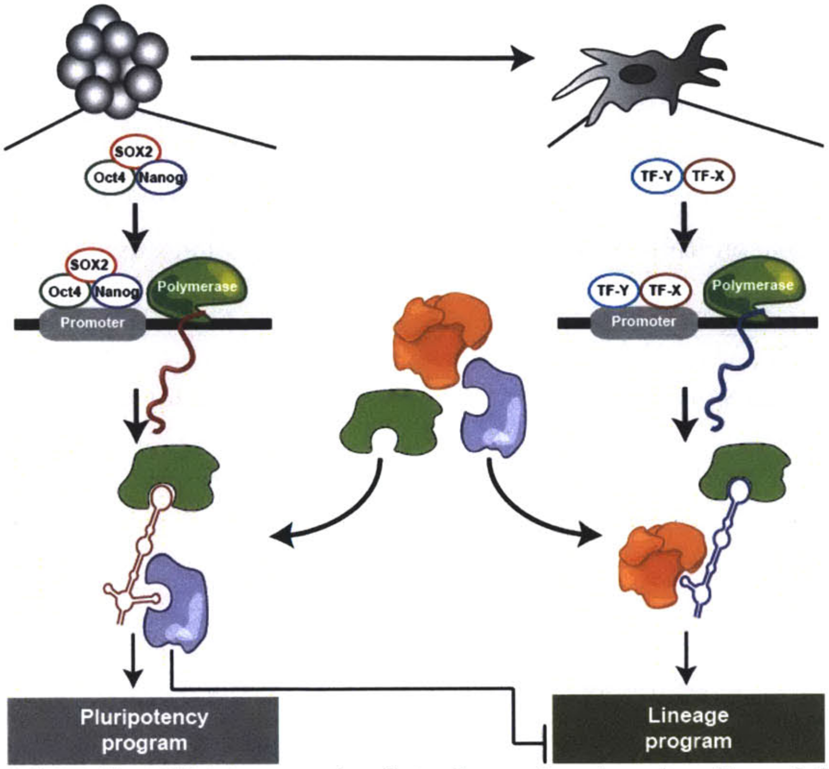


Figure 7: A model for lincRNA integration into the molecular circuitry of the cell. ESC-specific transcription factors (such as Oct4, Sox2, and Nanog) bind to the promoter of a lincRNA gene and drive its transcription. The lincRNA can then bind to different ubiquitous regulatory proteins, giving rise to cell-type specific RNA–protein complexes. Through different combinations of protein interactions, the lincRNA–protein complex can give rise to unique transcriptional programs. Right: A similar process may also work in other cell types with specific transcription factors regulating expression of lincRNAs, creating cell-type–specific RNA–protein complexes and regulating cell-type–specific expression programs.

METHODS

ES Cell Culture

V6.5 and Nanog-Luciferase[49] cells were co-cultured with irradiated C57BL/6 MEFs (GlobalStem; GSC-6002C) on pre-gelatinized plates as previously described[76]. Briefly, cells were cultured in mES media consisting of knock-out DMEM (Invitrogen; 10829018) supplemented with 10% FBS (GlobalStem; GSM-6002), 1% penicillin-streptomycin (Invitrogen; 15140-163), 1% L-glutamine (Invitrogen; 25030-164), 0.001% Beta-mercaptoethanol (Sigma; M3148-100ML) and 0.01% ESGRO (Millipore; ESG1106).

Picking lincRNA gene candidates

Using our previous catalogue of K4-K36 defined lincRNAs[5] along with the reconstructed full-length sequences we determined using RNA-Seq[6], we designed shRNA hairpins targeting each lincRNA identified in both sets. Specifically, we used the conservative K4-K36 definitions from our previous work[5] that were expressed in mouse ES cells. We further filtered the list to include only multi-exonic lincRNAs that were reconstructed in mouse ES cells[6]. Together, this yielded 237 lincRNA genes.

Picking protein-coding gene candidates

We selected protein coding gene controls consisting of both transcription factors and chromatin proteins. These proteins were selected based on their well-characterized role in regulating mouse ES cells and include Pou5f1 (Oct4)[57,77], Sox2[29,78], Nanog[48,79], Stat3[23], Klf4[80], and Zfp42

(Rex1)[81]. In addition, we selected additional transcriptional and chromatin regulators that were identified by RNAi screens as regulators of pluripotency[29,32,36] and/or were found in smaller focused studies to have critical roles in the maintenance of the pluripotent state (such as Carm1[82], Chd1[83], Thap11[84], Suz12[30,31,58], and Setdb1[33,34]). A full list is provided in Supplemental Table 4.

shRNA Design Rules

For each lincRNA we designed 5 hairpins by extending the previously described design rules[35] accounting for the sequence content of the hairpin, miRNA seed matches, uniqueness to the target compared to the transcriptome and the genome, and number of lincRNA isoforms covered.

For each lincRNA we enumerated all 21-mer sub-sequences as follows: (i) A "clamp score" was computed by looking at the nucleotides at positions 18, 19, and 20. If all three positions contained an A/T it was assigned a score of 4, if two positions were A/T it was assigned a score of 1.5 and if one was A/T it was assigned a score of 0.8. We then looked at positions 16, 17, and 21 if all 3 were A/T it was assigned a score of 1.25, if 2 were A/T it was assigned a score of 1.1, and if 1 was A/T is was assigned a score of 0.8. The clamp score was computed as the product of these two scores. (ii) A "GC score" was computed by looking at the total GC percentage of the 21-mer sequence. If the sequence was <25% GC it was assigned a score of 0.01 if it was <55% it was assigned a score of 3, if it was <60% it was assigned a score of 1, and if >60% it was assigned a score of 0.01. (iii) A "4-mer penalty" of 0.01 was assigned for any hairpin containing the same nucleotide in 4 subsequent nucleotides. (iv) A "7 GC penalty" of 0.01 was assigned to any hairpin containing any 7 consecutive G/C nucleotides. (v) We removed all hairpins containing an A in either position 1 or position 2 of the hairpin. (vi) We removed all hairpins containing a repeat masked nucleotide. (vii) Finally, we computed a "miRNA-seed penalty" by looking at the forward positions 11-17, 12-20, and 13-19 of the hairpin as well as the reverse complement of positions 14-20, 15-21, or 16-21 plus a 3' C. We then looked up whether these positions matched known miRNA seeds and with what frequency. We computed the scores for the forward and reverse positions and defined the score as the product of the forward and reverse scores. The final score for each hairpin sequence is defined as the product of all seven scores values.

We then sorted the candidate hairpin sequences by score, breaking high scoring ties by the total number of lincRNA isoforms that are covered by the hairpin. We then aligned each hairpin sequence against both the genome and the RefSeq-defined transcriptome (NCBI Release 39), and filtered any hairpin with fewer than three mismatches to any other gene or position in the genome. Candidate sequences were chosen for shRNA production by first picking the highest scoring candidate and then proceeding to successively lower scores. As each hairpin was selected, all other hairpins overlapping this hairpin were removed. We repeated this process until we identified 5 hairpins that covered each lincRNA.

shRNA cloning and virus prep

We designed 1,144 hairpins targeting 237 lincRNA genes. Of these, we successfully cloned 1087 hairpins targeting 223 lincRNAs. These hairpins were cloned into a vector containing a puromycin resistance gene and incorporated into a lentiviral vector as previously described[35]. Briefly, synthetic double stranded oligos that represent a stem-loop hairpin structure were cloned into the second-generation TRC (the RNAi Consortium) lentiviral vector, pLKO.5; the expression of a given hairpin produces shRNA that targets the gene of interest. Lentivirus was prepared as previously described[35]. Briefly, 100ng of shRNA plasmid, 100ng of packaging plasmid (psPAX2) and 10ng of envelope plasmid (VSV-G) were used to transfect packaging cells (293T) with TransIT-LT1 (Mirus Bio). Virus was harvested 48 and 70 hours post-transfection. Two harvests were combined. Virus titers were measured as previously described[35]. Briefly, we measured virus titers by infecting A549 cells with appropriately diluted viruses. 24 hours post infection, puromycin was added to a final concentration of 5ug/ml and the selection proceeded for 48 hours. The number of surviving cells, which is correlated to virus titer, was measured by alamarBlue (BioSource) staining utilizing the Envision 2103 Multilabel plate reader (PerkinElmer)

Infection and selection protocol

V6.5 ES cells were plated at a density of 5000 cells/well in 100ul mES media onto pre-gelatinized 96-well dishes (VWR; BD356689). Cells were infected with 5ul of a lentiviral shRNA stock and incubated at 37°C for 30 minutes. Puromycin resistant DR4 MEFs (GlobalStem; GSC-6004G) were then added to the plates at a density of ~6000/well and incubated overnight at 37°C, 5% CO2. After 24 hours, all media was removed from the cells and replaced with media containing 1ug/mL puromycin. Media was then changed every other day with fresh media containing 1ug/mL puromycin. The end-point depended on the assay and was either 4-days post infection (validation and microarrays) or 8-days (reporters and qPCR of marker genes).

RNA Extraction

ES cells were infected and lysed at day 4 with 150ul of Qiagen's RLT buffer and 3 replicates of each virus plate were pooled for RNA extraction using Qiagen's RNeasy 96-well columns (74181). RNA extraction was completed following Qiagen's RNeasy 96-well protocol with the following modifications: 450ul of 70% ethanol was added to 450ul total lysate prior to the first spin. An additional RPE wash was added to the protocol, for a total of 3 RPE washes.

lincRNA primer design and prescreen

lincRNA primers were designed using primer3 (http://frodo.wi.mit.edu/primer3/). Specifically, we designed primers spanning exon-exon junctions by specifying each of the regions as preferred inclusion regions in the primer3 program. When a low scoring primer pair (primer penalty <1) was available it was used. If none was available, we then identified all primers that contained amplicons that spanned an exon-exon junction. In a few cases, when we could not identify a primer pair spanning an exon-exon junction, we designed primers within an exon of the lincRNA. For each primer pair, we tested the specificity against the transcriptome[85] (Ref Seq NCBI Release 39) and the genome (Mouse MM9) using the isPCR (http://genome.ucsc.edu/cgi-bin/hgPcr) program. Specifically, we required that the primer pair amplify the lincRNA gene and no other genomic of gene amplicon.

For each primer pair, we validated the quantification and specificity prior to use. Specifically, we tested primers in qPCR reactions using a dilution series of mouse ES cDNA including a no reverse transcriptase (RT) sample. We excluded any primer that did not have robust quantification across a 64-fold dilution curve, had high signal in the no RT sample, or had low detectable expression in the undiluted sample (cycle number >34). For primers that failed this validation we redesigned and tested new primers.

Knockdown validation using qPCR

To determine if lincRNA hairpins were effective at knocking down the lincRNA of interest, we infected each hairpin into mouse embryonic stem cells, selected for lentiviral integration, and measured changes in the targeted lincRNA expression level. We isolated total cellular RNA after 4 days; this time-point was chosen to allow for identification of robust changes while minimizing secondary effects due to differentiation of the ES cells. We reasoned that this would allow us to determine more direct effects due to RNAi rather than to differentiation.

Gene panels were constructed that contained all 5 hairpins targeting a gene along with an empty vector control pLKO.5-nullT and the GFP-targeting hairpin clonetechGfp_437s1c1. cDNA was generated using 10ul of RNA and 10ul of 2x cDNA master mix containing 5x Transcriptor RT Reaction Buffer (Roche), DTT, MMLV-RT (Roche), dNTPs (Agilent; 200415-51), Random 9-mer oligos (IDT), Oligo-dT (IDT) and water. cDNA was diluted 1:9 and quantitative PCR was performed using 250 nM each primer in 2x Sybr green master mix (Roche) and run on a Roche Light-Cycler 480. Target lincRNA expression and GAPDH levels were computed for each panel. lincRNA expression levels were normalized by GAPDH levels and this normalized value was compared to the reference control hairpins within the panel. Knockdown levels were computed as the average of the fold decrease compared to the two control hairpins. Hairpins showing a knockdown greater than 60% of the endogenous level were considered validated and the best validated hairpin from a lincRNA panel was selected for subsequent studies.

Picking candidates for microarray analysis

To assess the effects of a lincRNA on gene expression, we profiled the changes in gene expression after knocking down each lincRNA gene. Specifically, for each lincRNA with at least 1 validated hairpin we profiled the genome-wide expression level changes after knockdown across 2 independent infections (see above). To control for expression changes due to viral infection, we performed five independent infections containing no RNAi hairpin (pLKO.5-nullT). This control hairpin was embedded in each RNA prep plate. To control for effects due to an off-target RNAi effect, we profiled 27 distinct negative control hairpins which do not have a known target in the cell. These hairpins included 6 RFP hairpins, 10 GFP hairpins, 6 Luciferase hairpins, and 5 LacZ hairpins. These hairpins provide a measurement of the variability of the RNAi response triggered due to non-specific effects. Furthermore, we profiled hairpins targeting 147 lincRNAs, including 10 with a second best hairpin, and 40 protein-coding genes in biological replicate. The hairpins and their replicates were randomly distributed across 7 96-well plates and prepared in batches. Each RNA preparation batch contained 1 pLKO hairpin and 1 clonetechGfp_437s1c1 hairpin in a random location on the plate. To minimize batch effects, the plate locations of the biological replicates were scrambled and the positions within the plates were scrambled for all hairpins and replicates.

Agilent Microarray hybridization

Using Agilent's One-Color Quick Amp Labelling kit (5190-0442), we amplified and labelled total RNA for hybridization to prototype mouse lincRNA arrays (G4140-90040) according to manufacturer's instructions with a few variations. The custom Agilent SurePrint G3 8x60K mouse array design used for this study (G4102A, AMADID 025725 G4852A) has probes to 21,503 Entrez genes and 2,230 lincRNA genes. A new updated version of this mouse design is commercially available that contains probes to 34,017 Entrez gene targets as well as 2,230 lincRNA genes (G4825A). The cRNA samples were prepared by diluting 200ng of RNA in 8.3ul water and adding positive control one-color RNA spike-in mix (Agilent, 5188-5282) that was diluted serially 1:20, then 1:25 and finally 1:10. We annealed the T7 promoter primer from the kit by incubating at 65°C for 10 minutes. We prepared the cDNA master mix and added to the annealed RNA and incubated at 40°C for 2 hours, followed by 65°C for 15 minutes. We prepared the cRNA transcription master mix and added it to the cDNA and incubated at 40°C for 2 hours protected from light. We purified the labeled cRNA using Qiagen's RNeasy 96-well columns (Qiagen, 74181) by adding 350ul of Qiagen RLT (without BME) to the cRNA followed by the addition of 250ul of 95% ethanol before applying to the plate column. After a 4 minute spin at 6000RPM, we washed the columns 3 times with 800ul buffer RPE. We dried the columns by spinning for 10 minutes and eluted the cRNA with 50ul of water. We measured the cRNA yield and dye incorporation using the Nanodrop 8000 Microarray measurement setting. We mixed 600ng of cRNA with blocking agent and fragmentation buffer (Agilent, 5190-0404) and fragmented for 30 minutes in the dark at 60°C. We added 2x Hybridization Buffer to each sample and loaded 40ul onto an 8-pack Hybridization gasket. We placed the microarray slides on top, sealed in the Hybridization Chamber, and incubated for 18 hours at 65°C. We washed the slides for 1 minute in room temperature GE Wash Buffer 1 and then for 1 minute in 37°C GE

Wash Buffer 2 (Agilent 5188-5327, no triton addition). We scanned the microarrays using an Agilent Scanner C (G2565CA) using the following settings: Dye Channel = Red&Green, Scan Region = ScanArea (61x21.6mm), Scan Resolution = 3μm. We prepared all of the samples simultaneously using homogenous master mixes to limit variability. Fragmentation and hybridization was staggered over time in batches of 3 to 4 slides (24 to 32 samples).

Array filtering, Normalization, and Probe filtering

Each array was processed and data extracted using the Agilent feature extraction software (G4462AA, Version 10.7.3). Samples were retained if they passed all the following quality control statistics:

AnyColorPrcntFeatNonUnifOL<1,

eQCOneColorSpikeDetectionLimit >0.01 and <2.0,

Metric_absGE1E1aSlope between 0.9 and 1.2,

Metric_gE1aMedCVProcSignal <8,

gNegCtrlAveBGSubSig >-10 and <5,

Metric_gNegCtrlAveNetSig <40,

gNegCtrlSDevBGSubSig <10,

Metric_gNonCntrlMedCVProcSignal <8,

Metric_gSpatialDetrendRMSFilterMinusFit <15, SpotAnalysis_PixelSkewCookiePct >0.8 and <1.2

Gene expression values were determined using the gProcessedSignal intensity values. Probes were flagged if they were not detectable well above background or had an expression level lower than the lowest detectable spike-in control value. The values were floored across all samples by taking the maximum of the minimum non-flagged values across all experiments. Any value less than this maximum value were set to the maximum. This conservatively eliminates any detection variability across the samples due to stringency or other array variables.

The result of this is a single value for each probe per array. To normalize expression values across arrays, we performed quantile normalization as previously described[86]. Briefly, we ranked each array from lowest to highest expression. For each rank, we computed the average expression and each experiment with this value at the associated rank. For each probe, we computed the difference between the second smallest expression value and the second largest expression value. If this difference was less than 2, we filtered the probe. This metric was chosen to eliminate bias due to single sample outliers.

Identifying significant gene expression hits from RNAi KDs

To control for effects due to non-specific effects of shRNAs, we profiled 27 distinct negative control hairpins which do not have a known target in the cell. These hairpins provide a measurement of the variability of the expression profiles due to random variability or triggered by 'off-target' effects of the shRNA lentiviruses. Assuming that any observed effects in the negative control hairpins are due to 'off-target' effects and observed effects in the targeting hairpins include a mix of both ‘off-target’ effects and ‘on-target’ effects, we use permutations of the negative controls to assign a false discovery rate (FDR) confidence level for being an ‘on-target’ hit to each gene. As such, a gene would only reach genome-wide significance if the number of genes and scale of the effect was much larger than would be observed randomly among all of the expression changes found for the negative control hairpin.

Specifically, for each gene we computed a t-statistic between shRNAs targeting the lincRNA and control shRNA samples. To assess the significance of each gene we permuted the sample and control groups retaining the relative sizes of the groups and computing the same t-statistic. We then assigned an FDR value to each gene by computing the average number of values in the permuted t-statistics that were greater than the observed value of interest and divided this by the number of all observed t-statistics that were greater than the observed value. We defined genes as significantly differentially expressed if the FDR was <5% and the fold-change compared to the negative controls was >2-fold. Using this approach, an effect would only reach a significant FDR if the scale is significantly larger than would be observed in the negative controls. Knockdown of a lincRNA was considered to have a significant effect of gene expression if we identified at least 10 genes that had an effect that passed all of the criteria.

Gene-Neighbour analysis

We identified neighbouring genes based on the RefSeq genome annotation[85] (NCBI Release 39). We excluded from analysis all RefSeq genes that corresponded to our lincRNA of interest but included all other coding and non-coding transcripts. We identified a significant hit as any lincRNA affecting a neighbour within 10 genes on either side with an FDR<.05 and 2-fold expression change. To compute the closest affected neighbour, we classified all genes affected upon knockdown of the lincRNAs using the same criteria above. We computed the distance between each affected gene and the locus of the lincRNA gene (and protein-coding gene) that was perturbed and took the minimum absolute distance across all affected genes.

Analysis of expected number of neighbouring genes that will change by chance

To determine the expected number of differentially expressed "neighbouring" genes occurring by chance assuming that the knockdown has no effect on gene expression, we calculated the average number of genes in a 300Kb window around a randomly selected gene in the human and mouse genome. We calculated this to be 11.2 (human) and 11.8 (mouse). For simplicity, we will conservatively round this down to 11. Assuming that no genes are changing between the knockdown and control, using a nominal p-value, which has a uniform distribution under the null hypothesis (nothing effected), we would expect to see a difference called in 5% of cases at a p-value of 0.05. If we test one locus, which has on average 11 neighbours we would expect to identify 0.55 hits by chance (11 x 0.05=0.55). However, if we now test 12 loci we would expect to see 6.6 (12 x 0.55) knockdowns which appear to have an effect under the null hypothesis.

Luciferase analysis of Nanog ES lines

ES cells containing a Nanog-Luciferase construct[49] were infected in biological duplicate and monitored after 7 days. Luciferase activity was measured using Bright-Glo (Promega). All reagents and cells were equilibrated to room temperature. 100ul Bright-Glo solution was added to each plate well. Plates were incubated in the dark at room temperature for 10 minutes and luciferase was measured on a plate reader. The luciferase units were normalized to the control hairpins and a z-score compared to the negative controls (excluding luciferase hairpins) was computed. For each hairpin, we computed a Z-Score relative to the negative control hairpins and identified hits reducing Luciferase levels more than 6 standard deviations ($Z<-6$) for both independent replicates. In all cases we were able to identify a significant reduction in luciferase levels when using distinct hairpins targeting luciferase. To exclude hits that were due to an overall reduction in proliferation (which would also cause a reduction of nanog positive cells in this read-out) we excluded all hairpins that caused a reduction in proliferation as measured by AlamarBlue incorporation (described below). AlamarBlue incorporation was measured in the same cells immediately before reading out Nanog-Luciferase levels.

AlamarBlue analysis of ES lines

After a 7 day infection, Nanog-Luciferase cell viability was measured using AlamarBlue (Invitrogen; DAL1025). AlamarBlue was mixed with mES media in a 1:10 ratio, added to the cells and incubated at 37°C for 1 hour. Absorbance readings at 570nm were taken. To control for possible effects due to virus titer, we measured AlamarBlue incorporation on both puromycin treated and non-puromycin treated samples for each infection.

Immunofluorescence

We crosslinked cells in 4% paraformaldehyde for 15 minutes, and washed in 1x PBS three times. To permeabilize the cells, we washed with 1x PBS + 0.1% Triton and then blocked in 1x PBS + 0.1% Triton + 1% BSA for 45 minutes at room temperature. We incubated cells with α-Pou5f1 antibody (Santa Cruz: SC-9081) at 1:100 dilution in blocking solution for 1.5 hours at room temperature and then washed in blocking solution three times. Next, we incubated cells in α-rabbit secondary antibody coupled to GFP (Jackson ImmunoResearch: 111-486-152) at a dilution of 1:1000 in blocking solution for 45 minutes. Finally, we thoroughly washed cells in blocking solution three times, and added vectashield containing DAPI (VWR: 101098-044) to each well.

Public Dataset curation

Traditionally, lineage markers are used to identify changes in phenotypic states. While these markers can be good indicators of differentiation potential, there are two major limitations with this approach. First, there are multiple genes that are associated with each lineage so simply looking at one can often be misleading. Second, this approach only works for classifying states with well-characterized marker genes but would not work for a comprehensive characterization of the function in the cell. Therefore, we decided to take a different approach and look at the entire gene expression profile of each lincRNA knockdown to determine what cell state each lincRNA resembles.

We curated a set of ES perturbations and differentiation states from publicly available sources. Specifically, we utilized the NCBI e-utils (http://eutils.ncbi.nlm.nih.gov/) to programmatically identify all published datasets containing keywords associated with embryonic stem cells. We filtered the list to only include mouse data sets that were generated across one of three commercial array platforms (Affymetrix, Agilent, and Illumina). Following this approach, we manually curated the list to include datasets associated with ESC perturbations (genetic deletions, RNAi, or chemical perturbations) and differentiation or induced differentiation profiles. This curation yielded 41 GEO datasets corresponding to >150 samples.

Specifically, we defined differentiation lineage states using the following datasets.

1. **Neuro-ectoderm**. We downloaded a dataset (GSE12982) corresponding to mouse ES cells containing a Sox1-GFP reporter construct. Upon differentiation of Sox1-GFP ES cells into Embryoid bodies (EBs), Sox1-GFP positive cells were collected and their global expression was profiled[50]. In addition, we downloaded a dataset (GSE4082)[87] corresponding to direct neuroectoderm differentiation[88].
2. **Mesoderm**. We downloaded the same dataset (GSE12982) as above, where the authors differentiated Brachyury-GFP reporter ES cells into EBs and sorted and profiled Brachyury-GFP positive cells[50].
3. **Endoderm**. We downloaded a dataset (GSE11523) corresponding to mouse ES cells which were engineered to overexpress GATA6[51]. GATA6 overexpression has been shown to drive ES cells into a primitive endoderm-like state[89].
4. **Ectoderm.** We downloaded a dataset (GSE4082)[87] corresponding to mouse ES cells differentiated into primitive ectoderm like cells with defined media[88].
5. **Trophectoderm**. We downloaded a dataset (GSE11523)[51] corresponding to mouse ES cells which were engineered to deplete Oct4[57]. These cells have been shown to enter a trophectoderm-like state[57]. To ensure specificity to the trophectoderm state, we also compared the expression effects to trophoblast stem cells[51]. For all lincRNAs identified, we required a significant enrichment for *both* induced Oct4 knock-out and trophoblast stem cell programs.

In addition, for all lineage states we utilized a curated discrete gene expression signature of differentiation which was previously functionally tested and shown to correspond specifically to differentiation into the associated states[54].

Continuous enrichment analysis and Phenotype-projection analysis

To determine relationships between lincRNA knockdowns and functional states, we employ a modified Gene Set Enrichment Analysis[55] approach that accounts for the continuous nature of the two datasets, similar to previously described extensions[55,56,90]. For each lincRNA knockdown by functional pair we compute a continuous enrichment score. Specifically, (i) for each lincRNA knockdown we compute a normalized score matrix compared to a panel of negative control hairpins by computing a t-statistic for each gene between the replicate lincRNA knockdown expression values and the control knockdown values. (ii) For each experiment, we sort the matrix by the normalized score such that the most differentially expressed upregulated gene is first and the most differentially expressed downregulated genes is last. Using this ordering we sort the functional dataset such that the ordering corresponds to the differential rank of the lincRNA knockdown set. (iii) We compute a score S_i as the running average of values from the first position to position *i*. We then define the enrichment score *E* as the maximum of the absolute value of S_i for all values of $i>10$. We require $i>10$ to avoid small fluctuations in the beginning of the ranked list causing fluctuations in the enrichment score. This score is computed for each lincRNA knockdown by functional set. Since we have many lincRNA knockdowns and functional sets, in reality we have a matrix of scores and we will refer to the enrichment score of the i^{th} knockdown and j^{th} functional set as E_{ij}.

To assess the significance of these scores, we compute a permutation derived false discovery rate and assign a confidence value for each projection. Specifically, to assess the significance of E_{ij}, we permute the lincRNA knockdown samples and control samples and compute the enrichment score for each pair across all permutations. To account for the false discovery rate associated with many lincRNAs and functional sets, we use the values of all permutations directly to assess the FDR level of E_{ij}. Specifically, to assess the FDR for each enrichment value E_{ij}, we accumulate all the permutation values for all lincRNA knockdowns and functional sets and compute the number of values greater than E_{ij} as well as a vector of values greater than E_{ij} corresponding to each permutation. The FDR is computed as the average number of permuted values greater than E_{ij} divided by the observed number greater than E_{ij}. Using this approach, we assign an FDR value to each lincRNA knockdown by functional set and identify significant hits as those with an FDR<0.01.

To highlight the accuracy of this approach, we observed that for publicly available gene perturbations for which we also perturbed the gene we were able to identify a significant association of target genes in ~75% of cases. While the remaining few did not pass our conservative significance criteria, they also showed increased enrichments consistent with their common effects. In addition, the projected effects are highly reproducible across distinct experiments originating from many groups and across multiple expression platforms. Highlighting the specificity of this approach, we note that there are many profiles for which no lincRNA had a similar effect.

Analysis of gene-expression overlaps between independent hairpin knockdowns

To determine whether independent hairpins targeting the same lincRNA gene share common gene targets, we computed a continuous enrichment score described above. Briefly, we computed a t-statistic for both hairpins against the negative controls. We then took the second best hairpin and sorted the genes. We scored the best hairpin affected genes based on this ranked order. We assessed the significance of this enrichment by permuting the samples and controls and assigned an FDR of the overlap of the expression effect (as described above).

Discrete gene set analysis

Discrete gene sets were analysed using the Gene Set Enrichment Analysis with a slight modification to the scoring procedure to be more analogous to our continuous scoring procedure (described above). Specifically, we computed the average of the expression changes (defined by the t-statistic) for all genes within the discrete gene set upon knockdown[54]. Significance was assessed by permuting the control and sample labels and recomputing the average statistic for each permutation. The FDR was assessed off of these values as described above.

Lineage marker gene analysis

We curated lineage marker gene sets from published work and publicly available sources[29,53,54]. We identified lineage marker genes as significantly upregulated using the differential expression criteria outlined above. We validated the expression of these lineage marker genes for a selected set of lineage marker genes using qPCR (as described above) after an 8-day infection.

Specifically, we looked at the expression of FGF5 (ectoderm), Sox1 (neuroectoderm), Sox17 (endoderm), Brachyury (mesoderm), and Cdx2 (trophectoderm). Expression estimates were normalized to GAPDH and compared to a panel of 25 negative control hairpins.

Identifying bound lincRNA promoters

We obtained genome-wide transcription factor binding data in mouse ES cells from 2 sources. The transcription factors Oct4, Sox2, Nanog, and Tcf3 were downloaded from the Gene Expression Omnibus (**GSE11724**) and the cMyc, nMyc, Zfx, Stat3, Smad1, Klf4, and Esrrb from GEO (**GSE11431**). For each ChIP-Seq dataset, the raw reads were obtained from the SRA (http://www.ncbi.nlm.nih.gov/sra) and processed as follows. (i) The reads were all aligned to the mouse genome assembly (build MM9) using the Bowtie aligner[91], requiring a single best placement of each read. All reads with multiple acceptable placements were removed from the analysis. (ii) Binding sites were determined from the aligned reads using the MACS[92] (http://liulab.dfci.harvard.edu/MACS/) algorithm using the default parameters with –mfold 8 to account for varying read counts in the libraries. (iii) lincRNA promoter regions were defined as previously described[5,6] using the location of the K4me3 peaks overlapping or within 5Kb of the transcriptional start site determined by RNA-Seq reconstruction. (iv) The transcription factor binding locations and lincRNA promoter locations were intersected and the enrichment level of the peak overlapping a lincRNA promoter was assigned transcription factor binding enrichment for each lincRNA. We defined transcription factor binding locations for protein-coding genes in a comparable way. (v) To exclude the possibility that some of this binding might be due to transcription factor binding at distal enhancers, we excluded all binding events that showed evidence of P300, a protein associated with active enhancers[93], localization. Altogether, we only identified ~5% of promoters overlapping with any P300 enrichment signal, a slightly lower percentage than identified for protein-coding gene promoters with detectable P300 signal.

Identifying TF-regulated lincRNA genes

lincRNA probes on the Agilent microarray were analysed using the differential expression methodology described above after knockdown of the transcription factor and comparison to the negative control hairpins. To confirm the expression changes of these lincRNAs, we hybridized 12 transcription factor knockdowns on a custom lincRNA codeset using the Nanostring nCounter assay[65]. The knockdowns were profiled in biological duplicate along with 15 negative controls. Regulated lincRNAs were identified using the differential expression approach described above.

Nanostring probeset design

Nanostring probes against lincRNA genes were designed following the standard nanostring design principles with the following modifications specifically for the lincRNA probes. (i) To exclude possible cross-hybridization, probes were screened for cross-hybridization against both the standard mouse transcriptome as well as a background database constructed from all the lincRNA sequences. (ii) To account for isoform coverage, a first pass design attempted to select a probe that would target as many isoforms as possible for each lincRNA. In cases where it was not possible to target all isoforms for a given lincRNA, the probe that targeted the largest number was selected, and additional probes were chosen when possible to target the remaining isoforms.

(iii) The standard restrictions on Tm and sequence composition were relaxed to include probes for as many lincRNAs as possible.

RA differentiation

V6.5 cells were cultured on gelatin-coated dishes in mES media in the absence of LIF. 5μM of retinoic acid was added daily and cell samples were taken daily for 6 days. RNA was extracted using Qiagen's RNeasy spin columns following the manufacturer's protocol.

Western blots

30ug of mESC nuclear protein extracts were run on 10% Bis-Tris gels (Invitrogen NP0316BOX) in MOPS buffer (Invitrogen NP0001) at 75 volts for 20 minutes followed by 120 volts for 1 hour. Gels were incubated for 30 minutes in 20% methanol transfer buffer (Invitrogen NP0006-1) and transferred onto PVDF membranes (Invitrogen 831605) at 20 volts for 1 hour using the Bio-Rad semi-dry transfer system (170-3940). Membranes were blocked in Blotto (Pierce, 37530) at room temperature for 1 hour. Antibodies were diluted in Blotto and membranes were incubated overnight at 4°C. Antibodies were diluted in using the following concentrations. Ezh2 1:2000, Suz12 1:5000, hnRNPH 1:1000, Ruvbl2 1:1000, Jarid1b 1:500, HDAC1 1:250, Cbx6 1:500, YY1 1:500. All antibodies tested were raised in rabbit. The next day, membranes were washed 3x in 0.1% TBST for 5 minutes each. The membranes were probed with anti-Rabbit-horse radish peroxidase (GE Healthcare; NA9340V) at a 1:10,000 dilution, washed 3x in 0.1% TBST, incubated in ECL reagent (GE Healthcare RPN2132), and exposed.

Crosslinked RNA immunoprecipitation

V6.5 mES cells were fixed with 1% formaldehyde for 10 minutes at room temperature, quenched with 2.5M glycine, washed with 1x PBS (3x) harvested by scraping, pelleting, and resuspended in modified RIPA lysis buffer (150mM NaCl, 50mM Tris, 0.5% Sodium deoxycholate, 0.2% SDS, 1% NP-40) supplemented with RNase inhibitors (Ambion, AM2694) and protease inhibitors. For UV crosslinking experiments, cells were irradiated with 254nm UV light. Cells were kept on ice and crosslinked in 1x PBS using 400,000 μjoules/cm^2.

Cell suspension was sonicated using Branson 250 Sonifier for 3 x 20 s cycles at 20% amplitude. 10ul of Turbo DNase (Ambion, AM2238) was added to sonicated material, incubated at 37°C for 10 minutes, and spun down at max speed for 10 minutes at 4°C. Protein-G beads were washed and pre-incubated with antibodies for 30 minutes at room temperature. Lysate and beads were incubated at 4°C for 2 hours. Beads were washed 3x using the following wash buffer (1x PBS, 0.1% SDS, 0.5% NP-40) and crosslinks were reversed and proteins were digested with 5ul proteinase-K (NEB, P8102S) at 65° for 2-4 hours. RNA was purified using phenol/chloroform/isoamyl alcohol and RNA was precipitated in isopropanol.

Nanostring hybridization

500ng of total RNA was hybridized for 17 hours using the lincRNA codeset. The hybridized material was loaded into the nCounter prep station followed by quantification on the nCounter Digital Analyzer following the manufacturer's protocol. For RNA immunoprecipitation experiments, we used a modified protocol. After reverse crosslinking, RNA was extracted using phenol/chloroform and ethanol precipitation methods and resuspended in 10ul of H2O. 5ul of the eluted material was hybridized for 17 hours using the lincRNA codeset.

Nanostring analysis

Probe values were normalized to negative control probes by dividing the value of the probe by the maximum negative control probe. Probe values were floored to a normalized value of 3 (3-fold higher than maximum negative control). Probes with no value greater than this floor across all samples were removed from the analysis. The values were log transformed. To control for variability between runs and different input material amounts, we normalized all samples simultaneously using the quantile normalization approach described above. The result is a set of normalized log-expression values for each probe normalized across all experiments.

Validation of RNA immunoprecipitation methods

To validate our formaldehyde based RNA immunoprecipitation method we immunoprecipitated the RNA binding protein hnRNPH, which plays a role in mRNA splicing[94] and identified the associated RNAs. Consistent with known interactions, we identified a strong enrichment for its binding to intronic regions of mRNA genes. We validated these observed results in mouse ES cells by performing UV-crosslinking experiments[95-97] and identified nearly identical results. We identified a similar correlation between the UV and formaldehyde crosslinked samples as for biological replicates of UV crosslinked samples and formaldehyde crosslinked samples and highly comparable enrichments (Supplemental Figure 13).

Antibody Selection

We selected chromatin proteins that have been implicated in regulation of the pluripotent state along with their known associated 'reader', 'writer', and 'eraser' complexes. Specifically, we tested antibodies against 40 chromatin proteins, corresponding to 28 chromatin complexes. In many cases, we tested multiple antibodies against the same target protein to try and identify an antibody that worked well for immunoprecipitation. A full list of tested complexes and their associated antibodies are listed in Supplemental Table 20.

Determining significant chromatin-lincRNA enrichments

We tested each antibody using formaldehyde crosslinked cells and had a two-step procedure for considering an antibody successful. (i) We tested all selected antibodies in batches, with each batch containing a mock-IGG (Santa Cruz) negative control and hnRNPH (Bethyl) positive control. Batches with variability in either the mock-IGG or hnRNPH controls were excluded and retested. For each successful batch, we computed enrichment for each lincRNA between the tested antibody and mock-IGG. We considered an antibody successful in the first step if the highest enrichment level exceeded a 5-fold change compared to the mock-IGG control and more than 10 lincRNAs exceeded this threshold. While this approach can yield false positives (antibodies that pass but are not efficient) it significantly reduced the number of antibodies to be tested in the next step. (ii) For all antibodies that successfully passed the first criteria, we performed immunoprecipitation on two additional biological replicates along with 4 mock-IGG controls. We computed a *t*-statistic for each lincRNA compared to the controls and assessed the significance using a permutation test, by permuting the samples and IGG samples (as above). Hits were considered significant if they exceed a *t*-statistic cutoff of 2 (log scale) compared to the controls and had an FDR<0.2. We allowed a slightly higher FDR cutoff since the number of permutations was far smaller yielding lower power to estimate the FDR. Only antibodies yielding significant lincRNAs were considered successful. In total, we identified 12 of the 28 complexes (55 antibodies) with at least one successful antibody.

Determining significant overlaps between lincRNA and chromatin protein knockdown effects

To determine the functional overlap between the lincRNA and the chromatin complexes it physically interacts with, we compared the effects on gene expression upon knockdown of the lincRNA and the associated protein complex. To do this, we utilized the gene expression profiles determined for each lincRNA knockdown and knockdowns of 9 of the 12 identified chromatin complexes for which we had good hairpins. We defined each interaction between a lincRNA and protein, and compute a continuous enrichment score, generated all permutations of the control hairpins and sample hairpins and assigned a false discovery rate to the scores (as described above). At an FDR<0.05 we identified 43% of the interactions to be significant. For 69% of the interactions, we were able to identify an overlap at an FDR<0.1.

References

1 Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629-641 (2009).

2 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

3 Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563 (2005).

4 Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916-919 (2002).

5 Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227 (2009).

6 Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-510 (2010).

7 Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-11672 (2009).

8 Marques, A. C. & Ponting, C. P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10**, R124 (2009).

9 Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**, 556-565 (2007).

10 Dinger, M. E. *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* **18**, 1433-1445 (2008).

11 Mattick, J. S. The genetic signatures of noncoding RNAs. *PLoS Genet* **5**, e1000459 (2009).

12 Ravasi, T. *et al.* Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res* **16**, 11-19 (2006).

13 Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409-419 (2010).

14 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323 (2007).

15 Nagano, T. *et al.* The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**, 1717-1720 (2008).

16 Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750-756 (2008).

17 De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* **8**, e1000384 (2010).

18 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187 (2010).

19 Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat Cell Biol* **10**, 1106-1113 (2008).

20 Orom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58 (2010).

21 Smith, A. G. Embryo-derived stem cells: of mice and men. *Annu Rev Cell Dev Biol* **17**, 435-462 (2001).

22 Ying, Q. L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519-523 (2008).

23 Niwa, H., Burdon, T., Chambers, I. & Smith, A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev* **12**, 2048-2060 (1998).

24 Jaenisch, R. & Young, R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* **132**, 567-582 (2008).

25 Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-956 (2005).

26 Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117 (2008).

27 Loh, Y. H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38**, 431-440 (2006).

28 Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049-1061 (2008).

29 Ivanova, N. *et al.* Dissecting self-renewal in stem cells with RNA interference. *Nature* **442**, 533-538 (2006).

30 Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349-353 (2006).

31 Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326 (2006).

32 Fazzio, T. G., Huff, J. T. & Panning, B. An RNAi screen of chromatin proteins identifies Tip60-p400 as a regulator of embryonic stem cell identity. *Cell* **134**, 162-174 (2008).

33 Bilodeau, S., Kagey, M. H., Frampton, G. M., Rahl, P. B. & Young, R. A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* **23**, 2484-2489 (2009).

34 Yuan, P. *et al.* Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. *Genes Dev* **23**, 2507-2520 (2009).

35 Moffat, J. *et al.* A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283-1298 (2006).

36 Hu, G. *et al.* A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes Dev* **23**, 837-848 (2009).

37 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**, 38-44 (1991).

38 Lee, J. T. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev* **23**, 1831-1842 (2009).

39 Ponjavic, J., Oliver, P. L., Lunter, G. & Ponting, C. P. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* **5**, e1000617 (2009).

40 Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-1927 (2011).

41 Tian, D., Sun, S. & Lee, J. T. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **143**, 390-403 (2010).

42 Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-124 (2011).

43 Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A. & Panning, B. Xist RNA and the mechanism of X chromosome inactivation. *Annu Rev Genet* **36**, 233-278 (2002).

44 Wang, X. *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* **454**, 126-130 (2008).

45 Sproul, D., Gilbert, N. & Bickmore, W. A. The role of chromatin structure in regulating the expression of clustered genes. *Nat Rev Genet* **6**, 775-781 (2005).

46 Silva, J. *et al.* Nanog is the gateway to the pluripotent ground state. *Cell* **138**, 722-737 (2009).

47 Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* **450**, 1230-1234 (2007).

48 Mitsui, K. *et al.* The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**, 631-642 (2003).

49 Brambrink, T. *et al.* Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell* **2**, 151-159 (2008).

50 Shen, X. *et al.* EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency. *Mol Cell* **32**, 491-502 (2008).

51 Aiba, K. *et al.* Defining developmental potency and cell lineage trajectories by expression profiling of differentiating mouse embryonic stem cells. *DNA Res* **16**, 73-80 (2009).

52 Nishiyama, A. *et al.* Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell Stem Cell* **5**, 420-433 (2009).

53 Sherwood, R. I. *et al.* Prospective isolation and global gene expression analysis of definitive and visceral endoderm. *Dev Biol* **304**, 541-555 (2007).

54 Bock, C. *et al.* Reference Maps of Human ES and iPS Cell Variation Enable High-Throughput Characterization of Pluripotent Cell Lines. *Cell* **144**, 439-452 (2011).

55 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

56 Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929-1935 (2006).

57 Niwa, H., Miyazaki, J. & Smith, A. G. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet* **24**, 372-376 (2000).

58 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).

59 Jiang, H. *et al.* Role for Dpy-30 in ES Cell-Fate Specification by Regulation of H3K4 Methylation within Bivalent Domains. *Cell* **144**, 513-525 (2011).

60 Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521-533 (2008).

61 Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**, 730-732 (2007).

62 Mikkelsen, T. S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156-169 (2010).

63 Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631-634 (2010).

64 Jiang, J. *et al.* A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol* **10**, 353-360 (2008).

65 Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol* **26**, 317-325 (2008).

66 Loh, Y. H., Zhang, W., Chen, X., George, J. & Ng, H. H. Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. *Genes Dev* **21**, 2545-2557 (2007).

67 Dey, B. K. *et al.* The histone demethylase KDM5b/JARID1b plays a role in cell fate decisions by blocking terminal differentiation. *Mol Cell Biol* **28**, 5312-5327 (2008).

68 Ruthenburg, A. J., Allis, C. D. & Wysocka, J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol Cell* **25**, 15-30 (2007).

69 Tahiliani, M. *et al.* The histone H3K4 demethylase SMCX links REST target genes to X-linked mental retardation. *Nature* **447**, 601-605 (2007).

70 Cloos, P. A., Christensen, J., Agger, K. & Helin, K. Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease. *Genes Dev* **22**, 1115-1140 (2008).

71 Zappulla, D. C. & Cech, T. R. Yeast telomerase RNA: a flexible scaffold for protein subunits. *Proc Natl Acad Sci U S A* **101**, 10024-10029 (2004).

72 Zappulla, D. C. & Cech, T. R. RNA as a flexible scaffold for proteins: yeast telomerase and beyond. *Cold Spring Harb Symp Quant Biol* **71**, 217-224 (2006).

73 Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet* **30**, 167-174 (2002).

74 Beletskii, A., Hong, Y. K., Pehrson, J., Egholm, M. & Strauss, W. M. PNA interference mapping demonstrates functional domains in the noncoding RNA Xist. *Proc Natl Acad Sci U S A* **98**, 9215-9220 (2001).

75 Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689-693 (2010).

76 Meissner, A., Eminli, S. & Jaenisch, R. Derivation and manipulation of murine embryonic stem cells. *Methods Mol Biol* **482**, 3-19 (2009).

77 Nichols, J. *et al.* Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **95**, 379-391 (1998).

78 Avilion, A. A. *et al.* Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev* **17**, 126-140 (2003).

79 Chambers, I. *et al.* Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113**, 643-655 (2003).

80 Nakatake, Y. *et al.* Klf4 cooperates with Oct3/4 and Sox2 to activate the Lefty1 core promoter in embryonic stem cells. *Mol Cell Biol* **26**, 7772-7782 (2006).

81 Brons, I. G. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191-195 (2007).

82 Torres-Padilla, M. E., Parfitt, D. E., Kouzarides, T. & Zernicka-Goetz, M. Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature* **445**, 214-218 (2007).

83 Gaspar-Maia, A. *et al.* Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature* **460**, 863-868 (2009).

84 Dejosez, M. *et al.* Ronin is essential for embryogenesis and the pluripotency of mouse embryonic stem cells. *Cell* **133**, 1162-1174 (2008).

85 Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**, D32-36 (2009).

86 Yang, Y. H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15 (2002).

87 Aiba, K. *et al.* Defining a developmental path to neural fate by global expression profiling of mouse embryonic stem cells and adult neural stem/progenitor cells. *Stem Cells* **24**, 889-895 (2006).

88 Ying, Q. L., Stavridis, M., Griffiths, D., Li, M. & Smith, A. Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nat Biotechnol* **21**, 183-186 (2003).

89 Morrisey, E. E. *et al.* GATA6 regulates HNF4 and is required for differentiation of visceral endoderm in the mouse embryo. *Genes Dev* **12**, 3579-3590 (1998).

90 Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108-112 (2009).

91 Langmead, B., Hansen, K. D. & Leek, J. T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol* **11**, R83 (2010).

92 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

93 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858 (2009).

94 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009-1015 (2010).

95 Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-469 (2008).

96 Ule, J. *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**, 1212-1215 (2003).

97 Wang, Z., Tollervey, J., Briese, M., Turner, D. & Ule, J. CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. *Methods* **48**, 287-293 (2009).

Chapter 6: Future Directions

In this chapter, we describe the outlook for the future of large ncRNA research.

Parts of this work were first published as:

Guttman M and Rinn JL. (2012). Modular regulatory principles of large non-coding RNAs. ***Nature*** 482(7385):339-46

Large ncRNAs as scaffolds of proteins

One emerging theme for many large ncRNAs is the formation of multiple distinct RNA-protein interactions to carry out their role (**Figure 1**). An initial clue for this phenomenon came from the discovery of telomerase[1]. Telomerase activity requires an RNA component, TERC[2], which serves as a template for telomeric regulation and as a scaffold for the polymerase enzyme around the RNA[3] (**Figure 1b**). Importantly, genetic studies demonstrated that the RNA plays a modular functional role as genetically swapping particular domains of the RNA retained the overall function[4]. This demonstrated that the RNA was made up of discrete functional modules that simply needed to be brought into proximity[4].

More recently, HOTAIR has been shown to contain distinct protein-interaction domains that can associate with PRC2[5] and the CoREST/LSD1 complex[6], which together enable its function (**Figure 1b**). XIST also contains discrete functional domains. Through a series of genetic deletions it was demonstrated that XIST contains at least two discrete domains responsible for silencing (RepA) and localization (RepC)[7] (**Figure 1b**). These functional domains could be independently deleted without affecting the role of the other domain, thus suggesting the modularity of the RNA[7]. The XIST ncRNA was also shown to have multiple protein interaction domains; the silencing domain (RepA) binds to PRC2 and the localization domain (RepC) binds to YY1[8] and hnRNPU[9]. These examples illustrate that large ncRNAs can act as molecular scaffolds for protein complexes. Importantly, this phenomenon may generalize, as our recent studies have demonstrated that ~30% of ESC lincRNAs associate with multiple regulatory complexes[10].

In addition to interacting with multiple proteins, several examples of ncRNAs have been reported to interact directly with both DNA and RNA. In a few cases, ncRNAs have been reported to form triplex structures with DNA[11,12] (**Figure 1a**) such as a ncRNA that binds to the ribosomal DNA promoter and interacts with the DNMT3b protein to silence expression[11]. Moreover, RNA can also form traditional duplex base pairing interactions with DNA, a property long speculated for large ncRNAs[13]. Finally, RNA can form base-pair interactions with RNA (**Figure 3a**), which are crucial for processes such as tRNA-mRNA anticodon recognition[14], miRNA targeting[15], ribosome structure as a ribozyme[16], and splicing[17] amongst many others. Despite these few examples, the interactions between large ncRNAs and genomic DNA and other RNAs are not well characterized.

A Potential Modular RNA Code

Collectively, these studies suggest an intriguing hypothesis for large ncRNAs as flexible modular scaffolds[4,6,7,10]. In this model, RNA contains discrete domains that interact with specific protein complexes. These RNAs, through a combination of domains, bring specific regulatory components into proximity resulting in the formation of a unique functional complex. These RNA regulatory complexes can include interactions with proteins but may also extend to RNA-DNA and RNA-RNA regulatory interactions.

RNA is well suited for such a role since RNA is a malleable evolutionary substrate compared to a protein, allowing for the selection of discrete interaction domains[14]. Specifically, RNA can be easily mutated, tested, and selected without breaking its core functionality[14]. This model of modular interactions may explain the observation that there are highly conserved 'patches' within large RNA genes[18-20] that may have evolved for specific protein interactions[7,21,22]. The remaining regions may be more evolutionarily flexible, allowing for formation of new functional domains by random mutation and selection. This is consistent with the observation, based on genetic deletion experiments, that non-constrained regions of telomerase are dispensable[4].

The model of a modular scaffold is not limited to protein interaction. RNA can also base-pair with DNA, which might be used to 'guide' complexes to specific DNA sequences. Alternatively, RNAs might 'guide' complexes by bridging together sets of DNA-binding proteins. Such a model could explain how the same protein complexes are 'guided' to different DNA loci in distinct cell types.

Large ncRNAs can also form RNA-RNA interactions, raising additional intriguing possibilities for future experimental exploration. For example, two large RNA scaffolds might be linked through RNA-RNA interactions. Another possibility is that RNA-RNA interactions may result in unique RNA structures that can interact with protein complexes not attainable by the individual units. This has been observed in the ribosome, where the combination of RNA-RNA and RNA-Protein interactions are required for proper complex formation.

Outlook

The mechanism by which large ncRNAs carry out their regulatory role is only beginning to emerge. While a modular RNA regulatory code is an attractive hypothesis, it remains to be tested. Specifically, how large ncRNA-Protein interactions occur and their molecular principles remain unknown. Determining these principles will require identifying the sites of RNA-protein interactions and the direct RNA binding proteins *in vivo*. Moreover, how large ncRNAs localize to targets genes is unknown, but may involve direct RNA-DNA interactions (**Figure 1a**) or interactions with proteins containing DNA recognition elements (as suggested for XIST[8] and HOTAIR[6]). To gain insight, it will be important to catalog the interactions that ncRNAs form with genomic DNA and RNAs. Together, this data will help elucidate the rules guiding interaction and the functional implications of these associations which can be tested experimentally.

If large ncRNAs are truly modular, then each individual domain would carry out a unique functional role independent of the other domains. Demonstrating modularity will require genetic deletions of domains and spacer regions as well as domain swapping experiments. Learning these principles would result in a defined 'modular RNA code' for how RNAs can affect cell states. A true understanding of the 'modular RNA code' will allow for the creation of synthetically engineered RNAs that can interact with both nucleic acids and protein modules to carry out engineered regulatory roles.

In addition, large ncRNAs may work by other mechanisms that may not fit neatly into this 'modular RNA code'. However, it is premature to dismiss large ncRNAs having other mechanisms that may not fit neatly into this 'modular RNA code'. In the meantime, it is clear that mammalian genomes encode a diverse cast of functional large important ncRNAs whose roles we are only beginning to understand.

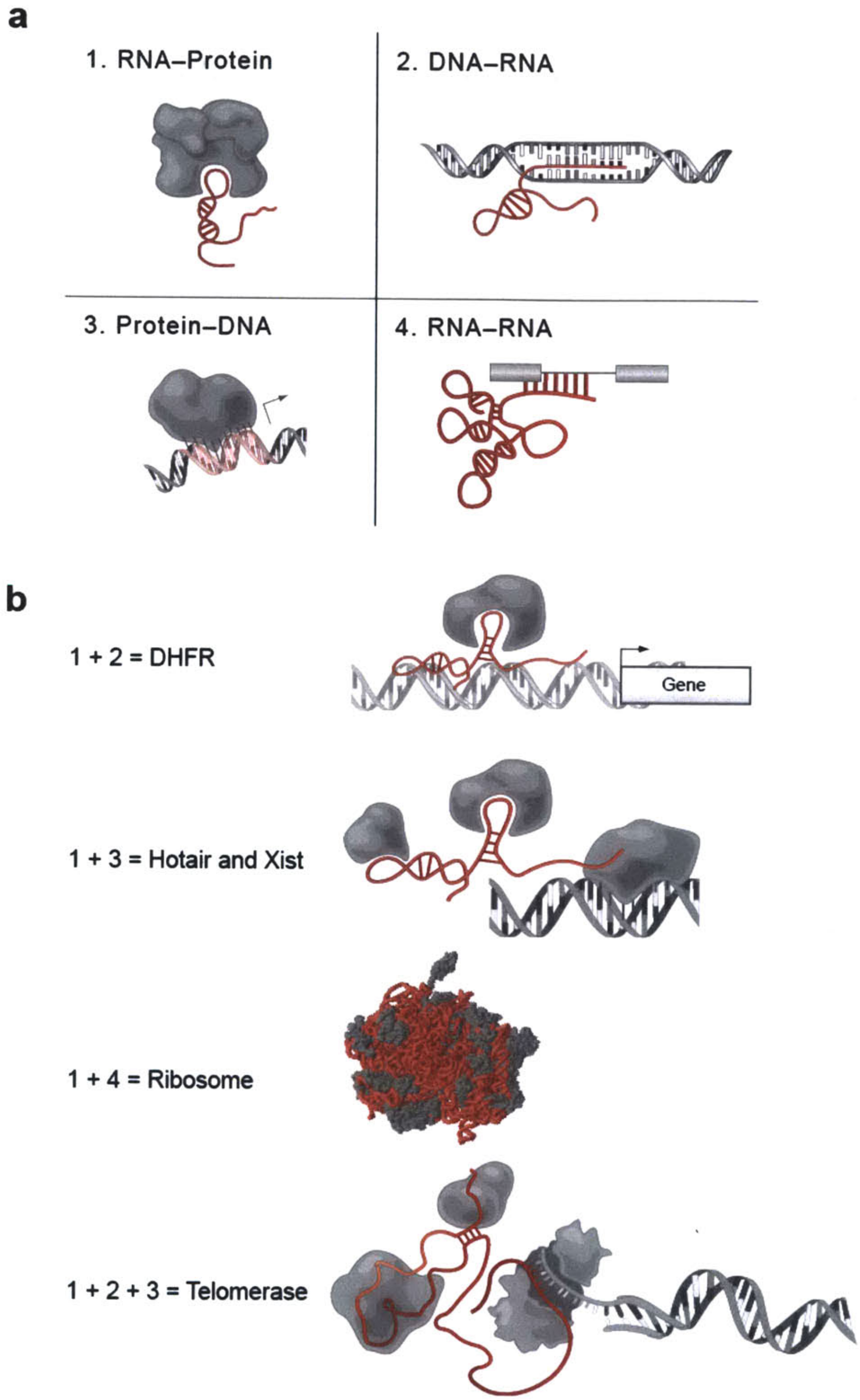


Figure 1: Modular principles of large RNAs. (a) Four principles of nucleic acid and protein interactions. (1) RNA-Protein interactions, (2) DNA-RNA hybridization based interactions, (3) DNA-Protein interactions, and (4) RNA-RNA hybridization based interactions. (b) Each of these principles can be combined to build distinct complexes. For example, combining RNA-Protein

and RNA-DNA interactions a protein complex can be localized to a specific DNA sequence in an RNA dependent manner as has been implicated for the DHFR[99] promoter and localization of DNMT3b[98]. Combining RNA-Protein and Protein-DNA principles can also localize a diverse set of proteins scaffolded by RNA to a specific DNA sequence in a protein dependent manner. The ribosome is a multifaceted combination of RNA-Protein interactions that facilitate proper RNA-RNA interactions for the ribozyme activity of the ribosome. The telomere replication activity of Telomerase is an example of combining RNA-Protein, RNA-DNA, and Protein-DNA interactions.

References

1 Greider, C. W. & Blackburn, E. H. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell* **43**, 405-413 (1985).

2 Feng, J. *et al.* The RNA component of human telomerase. *Science* **269**, 1236-1241 (1995).

3 Lingner, J. *et al.* Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* **276**, 561-567 (1997).

4 Zappulla, D. C. & Cech, T. R. Yeast telomerase RNA: a flexible scaffold for protein subunits. *Proc Natl Acad Sci U S A* **101**, 10024-10029 (2004).

5 Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323 (2007).

6 Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689-693 (2010).

7 Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet* **30**, 167-174 (2002).

8 Jeon, Y. & Lee, J. T. YY1 Tethers Xist RNA to the Inactive X Nucleation Center. *Cell* **146**, 119-133 (2011).

9 Hasegawa, Y., Brockdorff, N., Kawano, S., Tsutui, K. & Nakagawa, S. The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev Cell* **19**, 469-476 (2010).

10 Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300 (2011).

11 Schmitz, K. M., Mayer, C., Postepska, A. & Grummt, I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* **24**, 2264-2269 (2010).

12 Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666-670 (2007).

13 Lee, J. T. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev* **23**, 1831-1842 (2009).

14 Gesteland, R. F., Cech, T. & Atkins, J. F. *The RNA world : the nature of modern RNA suggests a prebiotic RNA world.* 3rd edn, (Cold Spring Harbor Laboratory Press, 2006).

15 Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215-233 (2009).

16 Korostelev, A. & Noller, H. F. The ribosome in focus: new structures bring new insights. *Trends Biochem Sci* **32**, 434-441 (2007).

17 Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* **2**, 919-929 (2001).

18 Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227 (2009).

19 Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-510 (2010).

20 Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**, 556-565 (2007).

21 Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750-756 (2008).

22 Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409-419 (2010).