The Human Molecular Clock and Mutation Process: A Characterization Using Microsatellite DNA

by

James Xin Sun

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

DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING AND BIOINFORMATICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2012

Signature of Author: _____
Harvard-MIT Division of Health Sciences and Technology
May 2012

Certified by: _____
David Reich, PhD
Professor of Genetics, Department of Genetics, Harvard Medical School
Thesis Supervisor

Certified by: _____
Nick Patterson, PhD
Senior Computational Biologist, Broad Institute
Thesis Supervisor

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

The Human Molecular Clock and Mutation Process: A Characterization Using Microsatellite DNA

by
James Xin Sun

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on May 7, 2012, in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN
ELECTRICAL ENGINEERING AND BIOINFORMATICS

Abstract

In the past decade, thousands of human genomes have been catalogued, either by whole-genome sequencing or by targeted genotyping. The variability between human genomes encodes invaluable information about human traits and genetic diseases, as well as human migration patterns and population interactions. A key challenge is to understand and characterize the evolution of the variability between human genomes. In this thesis, I focus on studying human evolution through the use of microsatellites, which are simple repetitive sections of DNA of typically 1-6bp motifs (e.g. CACACACACA) that are highly polymorphic and highly mutable.

The first aim is to establish that microsatellites are useful as reliable molecular clocks, such that its evolution highly correlates to time, especially when applied to the time range appropriate for human history. Using existing models of microsatellites, we examine microsatellite data from populations around the world to demonstrate that microsatellites are accurate molecular clocks for coalescent times of at least two million years. These results raise the prospect of using microsatellite data sets to determine parameters of population history.

In order to calibrate genetic distances into time, the mutation rate must be known. This leads to the second aim, which is to directly measure the microsatellite mutation rate from large-scale pedigree genetics data and provide a precision that is unprecedented. To do so, we use data from over 95,000 individuals in Icelandic pedigrees, genotyped in over 3000 microsatellite loci. Using trio and extended-family based approaches, we discover 2058 denovo mutations. In addition, we also attempt to capture many features that are covariates with the mutation rate, such as parental gender and age.

The third aim takes our empirical observations of the microsatellite mutation process to build a new model of microsatellite evolution. This model improves upon the standard random walk model with features we have captured from aim 2. We use a Bayesian coalescent approach to provide a model that estimates the sequence mutation rate, European genetic divergence times, and human-chimpanzee speciation time.

Thesis Supervisor: David Reich
Title: Professor of Genetics
Department of Genetics, Harvard Medical School

Acknowledgments

This research would not have been possible without the mentoring and support of many people. First, I would like to thank my thesis supervisors, David Reich and Nick Patterson, for being exceptional mentors. Their approach to research: creativity, mathematical rigor, and enthusiasm towards science have taught me and influenced me tremendously. Their guidance for the research has made my time at the lab productive and enjoyable.

I would also like to thank my thesis committee members Manolis Kellis and Shamil Sunyaev for their guidance and valuable critics of the research.

I would like to thank my collaborators at deCODE Genetics, in particular Agnar Helgason and Gisli Masson, for all their guidance and support while I spent my six trips in Iceland. Without them, two-thirds of this thesis would not have been possible.

I would like to thank all members of Reich lab for the technical discussions and of course, all the entertainment they have brought, whether it's for good fun at the lab, ping pong matches at the triangle table, or companionship at conferences.

I would like to thank friends and family: To my mom, dad, and Andrew, for their encouragement, love, and support; To Frank, Tamara, and Susie, for their life-long friendship; and of course, to my love, Dandi, for bringing so much happiness to my life.

Contents

Chapter 1

Introduction

In the past decade, thousands of human genomes have been catalogued, either by whole-genome sequencing or by targeting specific regions. The variability between human genomes encodes invaluable information about human traits and genetic diseases, as well as human migration patterns and population interactions. A key challenge is to understand and characterize the evolution of the variability between human genomes.

In this thesis, I focus on studying human evolution through the use of microsatellites, which are simple repetitive sections of DNA of typically 1-6bp motifs (e.g. CACACACACA) that are highly polymorphic and highly mutable. In particular, we focus on (1) studying the effectiveness of microsatellites as a molecular clock, (2) capturing microsatellite germline mutations in families, (3) characterizing microsatellite mutation process, (4) building a novel model of microsatellite evolution, and (5) using the model in conjunction with genomic sequence data to infer properties of human-chimpanzee speciation and human coalescent dates.

In this chapter, I introduce and review the concepts of microsatellites, molecular clocks, and mutation rates. The chapter is concluded with three specific aims, addressing the research focus of this thesis in more detail.

Microsatellites: Polymorphic tandemly repeated DNA

Microsatellites, also known as short tandem repeats (STR) or simple sequence repeats (SSR), are repetitive sections of DNA of typically 1-6bp motifs (e.g. CACACACACA, or shorthand $(CA)_5$).[1] In humans, at least 150,000 polymorphic microsatellites exist. A unique feature of microsatellites is their high mutation rate: due to DNA polymerase slippage during replication, the mutation rate is estimated to be around 10^{-3} to 10^{-4} per locus per generation, which is about 5 orders of magnitude higher than the nucleotide substitution rate of 10^{-8} per bp per generation.[2] As a result, a microsatellite locus genotyped in a population usually have a large number of alleles, distinguished by their variable allele lengths.

The hallmark feature of microsatellites, hypermutability, only becomes prominent when the repeat length is long enough to cause DNA polymerase slippage. However, there is no consensus as to the formal minimal length in defining a microsatellite. Assaying efforts have typically focused on on the most polymorphic microsatellite loci, producing lengths that are at least 10bp long. Upper limits of a microsatellite is usually a few hundred basepairs, where the locus starts to become impure: for example $(CA)_{20}TA(CA)_{20}$. Again, there is no consensus on defining a microsatellite with respect to the level of impurity tolerated in the repeat sequence. However, it has been hypothesized that the impurities reduce DNA polymerase slippage, and drastically lowers the mutation rate.

The technology to efficiently genotype microsatellites — using PCR followed by length separation on gel — has sparked an enormous amount of effort on using them in making inferences on genetic variation. They have been extensively analyzed in the context of constructing genetic linkage maps in a wide range of species, from humans to zebrafish to wheat [3-5]. Using linkage maps and family-based linkage analysis, microsatellites have been used to discover regions of identity by descent in related individuals, which in turn have been used to localize the search for disease genes.

There was also great interest in using microsatellites to study evolution[6-15]. Based on the microsatellite differences between individuals, and having a model of evolution that translates the differences into meaningful values such as the time of separation, one could in principle,

learn about the human past. In order to establish the usefulness of microsatellites in studying evolution, one must first demonstrate that these markers have the properties of a molecular clock.

Molecular clocks

First proposed by Zuckerkandl and Pauling in 1962[16], molecular clocks measure the time that has elasped since the two molecules shared a common ancestor. Used in genetics, the concept is as follows: first, measure the genetic distance between the two pieces of DNA, which may come from distinct individuals of a population, or from distinct species. The measured genetic distance is linear with respect to time. Then a calibration step is done to convert genetic distance into time in years.

In the simplest form, for DNA nucleotide substitutions, one can measure the genetic distance as the percentage of discrepant nucleotides, and if the mutation rate is known, time is readily calculated[17]. For example, humans and chimpanzees differ in nucleotide substitutions by $d_{seq} = 1.2\%$, and if we assume that the sequence mutation rate is $\mu_{seq} = 7 \times 10^{-10}$ per base per year, then the last time they shared a common ancestor is $t_{MRCA} = \frac{d_{seq}}{2\mu_{seq}} = 8.6$ million years ago.

For microsatellites, the molecular was based on preliminary evidence that microsatellites mutate approximately according to a random walk, whereby alleles undergo length changes during DNA replication due to polymerase slippage[1,18]. The simplest model was the single step symmetric stepwise mutation model (SMM)[19,20], whereby microsatellites mutate to one motif length shorter or longer with equal probability. Assuming that SMM holds, the microsatellite-based genetic distance between two samples is computed as the square distance[21], which is proportional to the time to the most recent common ancestor (t_{MRCA})[22]. For example, at a particular locus, if the two samples differ by 3 repeat units, then the squared distance is $d_{msat} = 9$. If we assume that the microsatellite mutation rate is $\mu_{msat} = 10^{-5}$ per locus per year, then $t_{MRCA} = \frac{d_{msat}}{2\mu_{msat}} = 450$ thousand years ago.

However, the simple clocks described above have been controversial[23]. One concern is with precision: unlike quartz clocks that directly measure time using an oscillator that vibrates precisely at 32,768 hertz, molecular clocks' tick rate is probabilistic: If the average mutation rate

is μ, the occurrence of mutations is not evenly spaced but rather modeled as a Poisson or over-dispersed Poisson distribution. Furthermore, the average mutation rate μ can be time-varying, and factors such as generation-time are hypothesized to influence this rate. The probabilistic nature of a molecular clock reduces precision.

The second concern is with accuracy: The clock can be biased if the evolution process is modelled incorrectly or if the mutation rate measured incorrectly. For example, the DNA nucleotide substiution clock presented above assumes that there are no multiple hits at any basepair, i.e. mutations that occur several times at the same site. This phenomena becomes important to correct for when comparing species far apart, enough such that these recurrent mutations become a significant concern. For microsatellites, accuracy of the model becomes even more crucial due to its complex behavior in length changes. Despite the initial excitement in using microsatellites to make inferences about history, this interest has waned because experimental evidence has revealed instances where the standard random walk model is violated. In the context of boundary constraints on microsatellite allele lengths, for example, ASD can lose accuracy for separations beyond 10,000 generations (assuming the range of alleles is constrained to 20) [24], which is well within the depth of human genetic variation. Researchers have also explored more complex models of microsatellite mutation that include boundary constraints [8,24] and length-dependent mutation [25-28], where ASD is also inappropriate. Perhaps the greatest concern for using microsatellites as molecular clocks is the concern that each locus would have to be characterized experimentally and individually modeled. Due to doubts about the ability to accurately model the microsatellite mutation process, recent studies have eschewed the use of microsatellite data to infer parameters of human history, though there are some important exceptions [29,30]. Thus, while large-scale microsatellite datasets have recently been collected in many human populations — in particular ~700 microsatellite loci were genotyped in approximately 3,000 individuals from 147 populations, including the Human Genome Diversity Panel (HGDP) [31-33], South Asians [34], Native Americans [35], Latinos [15], and Pacific Islanders [14]— only 2 of 8 studies [13,32] attempted to make time inferences with these data. Most studies have instead focused on using microsatellite data to detect and analyze population structure.

Therefore, it is exceptionally important to characterize the microsatellite mutation rate and mutation process with extreme precision, in order to use it as a molecular clock.

Mutation rates

Germline mutations provide the raw material for evolution and are the ultimate source of genetic variation. The rate of mutations (μ) is a fundamental quantity of evolution and is ubiquitous in modeling evolution. Despite its core importance, there are surprisingly little empirical data supporting how μ varies with phenotypes such as parental gender and age, and the variation of the μ with genomic features[36-39].

Previous studies of mutations have been primarily indirect species comparisons, disease-gene focused in humans, or in inbred lines of model organisms, leading to the confounding of mutations with other evolutionary forces such as drift and natural selection. Single-generation genome-wide mutation measurements from parent to offspring minimize the confounders and hence are most desirable. However, even with the latest sequencing technology, discovery of parent-offspring mutations via whole-genome sequencing is still severely affected by sequencing errors, leading to an imprecise mutation rate and a miniscule quantity of true mutations that is intractable for mutation variation characterization. While whole-genome sequencing of three nuclear families[40-42] revealed dozens of new mutations, there were too few individuals to provide a detailed characterization of the mutation process. Moreover, the studied families may be atypical and it is essential to study many families to obtain a population-wide estimate. One outcome of an understanding of the mutation process would be to provide a direct estimate of the rate of the molecular clock, which would make it possible to estimate the divergence time of humans and our closest living relatives, chimpanzees, without relying on the fossil record for time calibrations.

Thesis goals

Chapter 2: Aim 1: Microsatellites are accurate molecular clocks

The first aim is to establish that microsatellites are useful as reliable molecular clocks, such that its evolution highly correlates to time, especially when applied to the time range appropriate for the span of human history since the speciation from chimpanzees. To do this, we take advantage of newly available genome sequencing data sets that permit empirical assessments of the

microsatellite molecular clock. We compare a popular microsatellite distance statistic known as the Average Squared Distance (ASD) to genomic sequence divergence using datasets from both humans and chimpanzees, and show that the averaged microsatellite clock over all loci applies with remarkable accuracy to time depths that are about 10-fold greater than previous simulations. We discuss potential applications of this molecular clock, and show that after combining the clock with modeling analysis, microsatellites can be used to accurately estimate not only mean coalescent times between populations, but also to correct for ascertainment bias in SNP data. These results raise the prospect of using microsatellite data sets to determine parameters of population history.

Chapter 3: Aim 2: Characterizing the denovo microsatellite mutation rate

Aim 1 showed the molecular clock potential of microsatellites. However, in order to calibrate the microsatellite genetic distances into time, the mutation rate and mutation process must be known and characterized precisely. This leads to the second aim, which is to directly measure the microsatellite mutation rate from large-scale pedigree genetics data and provide a precision that is unprecedented. A large microsatellite dataset has been collected at deCODE Genetics over the past 15 years: about a third of the Icelandic population has been genotyped in over 5000 loci. Coupled with genotypes of multiple coverage and the full genealogy of Iceland, deCODE's data allows for the capture of the largest quantity of verifiable mutations across many families and loci in humans to date. We searched for mutations in 2,477 autosomal microsatellite loci in 24,832 father-mother-child trios as well as in 2,406 extended families. Using trio and extended-family based approaches, we discover 2058 denovo mutations in 6.04 million transmissions. In addition, we also capture many features that are covariates with the mutation rate, such as parental gender, age, and allele length.

Chapter 4: Aim 3: A model of microsatellite evolution

The third aim takes our empirical observations of the microsatellite mutation process to build a new model of microsatellite evolution. This model improves upon the standard random walk model with features we have captured from aim 2. We use a Bayesian coalescent approach to provide a model that estimates the sequence mutation rate, European genetic divergence times, and human-chimpanzee speciation time.

References

1. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**, 435-45 (2004).
2. Weber, J.L. & Wong, C. Mutation of human short tandem repeats. *Hum Mol Genet* **2**, 1123-8 (1993).
3. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152-4 (1996).
4. Shimoda, N. *et al.* Zebrafish genetic map with 2000 microsatellite markers. *Genomics* **58**, 219-32 (1999).
5. Roder, M.S. *et al.* A microsatellite map of wheat. *Genetics* **149**, 2007-23 (1998).
6. Bowcock, A.M. *et al.* High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455-7 (1994).
7. Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L. & Feldman, M.W. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci U S A* **92**, 6723-7 (1995).
8. Nauta, M.J. & Weissing, F.J. Constraints on allele size at microsatellite loci: implications for genetic differentiation. *Genetics* **143**, 1021-32 (1996).
9. Amos, W. & Rubinstzein, D.C. Microsatellites are subject to directional evolution. *Nat Genet* **12**, 13-4 (1996).
10. Goldstein, D.B. & Pollock, D.D. Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. *J Hered* **88**, 335-42 (1997).
11. Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J. & Rolf, B. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* **62**, 1408-15 (1998).
12. Zhivotovsky, L.A., Bennett, L., Bowcock, A.M. & Feldman, M.W. Human population expansion and microsatellite variation. *Mol Biol Evol* **17**, 757-67 (2000).
13. Becquet, C., Patterson, N., Stone, A.C., Przeworski, M. & Reich, D. Genetic structure of chimpanzee populations. *PLoS Genet* **3**, e66 (2007).
14. Friedlaender, J.S. *et al.* The genetic structure of Pacific Islanders. *PLoS Genet* **4**, e19 (2008).
15. Wang, S. *et al.* Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* **4**, e1000037 (2008).
16. Zuckerkandl, E. & Pauling, L. Molecular disease, evolution, and genetic heterogeneity. in *Horizons in Biochemistry* (eds. Kasha, M. & Pullman, B.) 189-225 (Academic Press, New York, 1962).
17. Hartl, D.L. & Clark, A.G. *Principles of population genetics*, xv, 652 p. (Sinauer Associates, Sunderland, Mass., 2007).
18. Levinson, G. & Gutman, G.A. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. *Nucleic Acids Res* **15**, 5323-38 (1987).
19. Ohta, T. & Kimura, M. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet Res* **22**, 201-4 (1973).
20. Valdes, A.M., Slatkin, M. & Freimer, N.B. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **133**, 737-49 (1993).
21. Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L. & Feldman, M.W. An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**, 463-71 (1995).

22. Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457-62 (1995).
23. Bromham, L. & Penny, D. The modern molecular clock. *Nat Rev Genet* **4**, 216-24 (2003).
24. Feldman, M.W., Bergman, A., Pollock, D.D. & Goldstein, D.B. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. *Genetics* **145**, 207-16 (1997).
25. Sainudiin, R., Durrett, R.T., Aquadro, C.F. & Nielsen, R. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics* **168**, 383-95 (2004).
26. Di Rienzo, A. *et al.* Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci U S A* **91**, 3166-70 (1994).
27. Kruglyak, S., Durrett, R.T., Schug, M.D. & Aquadro, C.F. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A* **95**, 10774-8 (1998).
28. Xu, X., Peng, M. & Fang, Z. The direction of microsatellite mutations is dependent upon allele length. *Nat Genet* **24**, 396-9 (2000).
29. Ramachandran, S., Rosenberg, N.A., Feldman, M.W. & Wakeley, J. Population differentiation and migration: Coalescence times in a two-sex island model for autosomal and X-linked loci. *Theor Popul Biol* (2008).
30. Szpiech, Z.A., Jakobsson, M. & Rosenberg, N.A. ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* **24**, 2498-504 (2008).
31. Rosenberg, N.A. *et al.* Genetic structure of human populations. *Science* **298**, 2381-5 (2002).
32. Zhivotovsky, L.A., Rosenberg, N.A. & Feldman, M.W. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am J Hum Genet* **72**, 1171-86 (2003).
33. Rosenberg, N.A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet* **1**, e70 (2005).
34. Rosenberg, N.A. *et al.* Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet* **2**, e215 (2006).
35. Wang, S. *et al.* Genetic variation and population structure in native Americans. *PLoS Genet* **3**, e185 (2007).
36. Crow, J.F. The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* **1**, 40-7 (2000).
37. Crow, J.F. Age and sex effects on human mutation rates: an old problem with new complexities. *J Radiat Res (Tokyo)* **47 Suppl B**, B75-82 (2006).
38. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).
39. Arnheim, N. & Calabrese, P. Understanding what determines the frequency and pattern of human germline mutations. *Nat Rev Genet* **10**, 478-88 (2009).
40. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-9 (2010).
41. Durbin, R.M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
42. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature genetics* **43**, 712-714 (2011).

Chapter 2

Microsatellites are molecular clocks that support accurate inferences about history

James X. Sun, James C. Mullikin, Nick Patterson, and David Reich

This chapter originally appeared in Molecular Biology and Evolution (2009) 26 (5): 1017-1027

Abstract

Microsatellite length mutations are often modeled using the generalized stepwise mutation process, which is a type of random walk. If this model is sufficiently accurate, one can estimate the coalescence time between alleles of a locus after a mathematical transformation of the allele lengths. When large-scale microsatellite genotyping first became possible, there was substantial interest in using this approach to make inferences about time and demography, but that interest has waned because it has not been possible to empirically validate the clock by comparing it to data in which the mutation process is well understood. We analyzed data from 783 microsatellite loci in human populations and 292 loci in chimpanzee populations, and compared them to up to one gigabase of aligned sequence data, where the molecular clock based upon nucleotide substitutions is believed to be reliable. We empirically demonstrate a remarkable linearity (r^2>0.95) between the microsatellite average squared distance (ASD) statistic and sequence divergence. We demonstrate that microsatellites are accurate molecular clocks for coalescent times of at least two million years. We apply this insight to confirm that the African populations San, Biaka Pygmy, and Mbuti Pygmy have the deepest coalescent times among populations in the Human Genome Diversity Project. Furthermore, we show that microsatellites support unbiased estimates of population differentiation (F_{ST}) that are less subject to ascertainment bias

than single nucleotide polymorphism (SNP) F_{ST}. These results raise the prospect of using microsatellite data sets to determine parameters of population history. When genotyped along with SNPs, microsatellite data can also be used to correct for SNP ascertainment bias.

Introduction

To be useful as a molecular clock, a polymorphic genetic locus needs to accumulate mutations in a predictable way, so that with an appropriate statistical transformation, the differences between two alleles present in the population can be used to obtain an unbiased estimate of the time that has elapsed since their last common genetic ancestor (Zuckerkandl and Pauling 1962). When loci dispersed throughout the genome are combined, this molecular clock can in principle provide accurate estimtes of genetic divergence times, and with further analysis, can also estimate ancestral population sizes and population migration histories.

Microsatellites (or short tandem repeats) are simple repetitive sections of DNA of typically 2-5bp motifs (e.g. CACACACACA). They possess several features suitable for a molecular clock. First, microsatellites are widely dispersed throughout the genome. In humans, an estimated 150,000 informative (sufficiently polymorphic) loci exist, of which tens of thousands have been genotyped (Weber and Broman 2001). Second, in humans the mutation rate at these markers is estimated to be around 10^{-3} to 10^{-4} per locus per generation (Ellegren 2000), which is orders of magnitude larger than the genome-wide average nucleotide mutation rate of around 10^{-8} per base per generation. The higher mutation rate means that a much smaller fraction of the genome needs to be sampled to make inferences with microsatellite data than with sequence data. Third, microsatellites are largely free of ascertainment bias compared to that of single nucleotide polymorphisms (SNPs) (Conrad et al. 2006). The extraordinarily high mutation rate at microsatellites means that they are primarily discovered not based on their polymorphism pattern in any one population (they are essentially guaranteed to be polymorphic) but instead based on their sequence. Thus, the population in which they are first studied is not expected to substantially bias inferences based on the data. By contrast, SNP allele frequency in

the population in which it is discovered has a dramatic influence on the probability that it will be included in a study, and thus SNP data sets are deeply affected by ascertainment bias (Clark et al. 2005). The great majority of SNPs on human genome-wide scanning arrays have been ascertained in a complex way that is very difficult to model, confounding the interpretation of allele frequency distributions for inferences about history.

The technology to efficiently genotype microsatellites — using PCR followed by length separation on gel — has sparked an enormous amount of effort on using them in making inferences on genetic variation. They have been extensively analyzed in the context of constructing genetic linkage maps in a wide range of species, from humans to zebrafish to wheat (Dib et al. 1996; Roder et al. 1998; Shimoda et al. 1999). Using linkage maps and family-based linkage analysis, microsatellites have been used to discover regions of identity by descent in related individuals, which in turn have been used to localize the search for disease genes.

Initially, there was great interest in using microsatellites to make inferences about history, not only in humans but also in other species (Bowcock et al. 1994; Paetkau et al. 1997). The idea that inferences about history were possible using these markers was based on preliminary evidence that microsatellites mutate approximately according to a random walk, whereby alleles undergo length changes during DNA replication due to polymerase slippage (Levinson and Gutman 1987; Ellegren 2004). The simplest model was the single step symmetric stepwise mutation model (SMM) (Ohta and Kimura 1973; Valdes, Slatkin, and Freimer 1993), whereby microsatellites mutate to one motif length shorter or longer with equal probability. In the generalized stepwise mutation model (GSMM) (Kimmel and Chakraborty 1996), the length changes can also be multi-step (Di Rienzo et al. 1994) or involve directional asymmetry (Amos and Rubinstzein 1996). Assuming that the GSMM holds, the average square distance (ASD) (Goldstein et al. 1995a) between microsatellite allele lengths of two individuals provides an unbiased estimate of the coalescence time between alleles across the genome, also known as the time to the most recent common ancestor (t_{MRCA}) (Slatkin 1995). The establishment of the microsatellite molecular clock using the GSMM led researchers to infer average coalescent times (Goldstein et al. 1995a; Goldstein et al. 1995b; Goldstein and Pollock 1997; Zhivotovsky 2001), population differentiation (F_{ST} for microsatellites) (Slatkin 1995), and patterns of population size expansion and contraction (Kimmel et al. 1998; Reich and Goldstein 1998).

Despite the initial excitement in using microsatellites to make inferences about history, this interest has waned because experimental evidence has revealed instances where the GSMM is violated. In the context of boundary constraints on microsatellite allele lengths, for example, ASD can lose accuracy for separations beyond 10,000 generations (assuming the range of alleles is constrained to 20) (Feldman et al. 1997), which is well within the depth of human genetic variation. Researchers have also explored more complex models of microsatellite mutation that include boundary constraints (Nauta and Weissing 1996; Feldman et al. 1997) and length-dependent mutation (Di Rienzo et al. 1994; Kruglyak et al. 1998; Xu, Peng, and Fang 2000; Sainudiin et al. 2004), where ASD is also inappropriate. Perhaps the greatest concern for using microsatellites as molecular clocks is the concern that each locus would have to be characterized experimentally and individually modeled.

Due to doubts about the ability to accurately model the microsatellite mutation process, recent studies have eschewed the use of microsatellite data to infer parameters of human history, though there are some important exceptions (Ramachandran et al. 2008; Szpiech, Jakobsson, and Rosenberg 2008). Thus, while large-scale microsatellite datasets have recently been collected in many human populations — in particular ~700 microsatellite loci were genotyped in approximately 3,000 individuals from 147 populations, including the Human Genome Diversity Panel (HGDP) (Rosenberg et al. 2002; Zhivotovsky, Rosenberg, and Feldman 2003; Rosenberg et al. 2005), South Asians (Rosenberg et al. 2006), Native Americans (Wang et al. 2007), Latinos (Wang et al. 2008), and Pacific Islanders (Friedlaender et al. 2008)—only 2 of 8 studies (Zhivotovsky, Rosenberg, and Feldman 2003; Becquet et al. 2007) attempted to make time inferences with these data. Most studies have instead focused on using microsatellite data to detect and analyze population structure.

In this study, we revisit the hypothesis that reliable inferences about history can be obtained using microsatellite data. To do this, we take advantage of newly available genome sequencing data sets that permit empirical assessments of the microsatellite molecular clock. We compare ASD to genomic sequence divergence using datasets from both humans and chimpanzees, and show that despite the known presence of deviations from the GSMM at many individual loci, the averaged microsatellite clock over all loci applies with remarkable accuracy to time depths that are about 10-fold greater than previous simulations. We discuss potential applications of this molecular clock, and show that after combining the clock with modeling

analysis, microsatellites can be used to accurately estimate not only mean coalescent times between populations, but also to correct for ascertainment bias in SNP data. It is likely that the microsatellite molecular clock can be useful to the analysis of population history for many populations and closely-related species, beyond the humans and chimpanzees analyzed here.

It is important to note that microsatellite ASD, like sequence divergence between two samples (the number of nucleotide differences per base pair), is expected to be proportional to the average time since the common ancestor (t_{MRCA}) of two alleles diverged across the genome, and does not provide any direct information about population split times. We focus on ASD here because we can directly plot it against average sequence divergence for population pairs and test whether the molecular clock holds, without making any assumptions about demographic history. Only after having demonstrated that ASD is an accurate molecular clock do we discuss its potential applications in estimating population split times, historical population sizes, and historical migrations, which are more complicated inferences that can only be done with the help of population genetics modeling.

Materials and Methods

Microsatellite data

For humans, we used 783 autosomal microsatellites from Rosenberg et al (Rosenberg et al. 2005). From this set, we found that 2 loci were almost perfectly correlated and removed the locus (*D2S1334*) with more missing data. We used Rosenberg's H952 set of individuals, who are expected to be less related than second cousins (Rosenberg 2006). To match individuals to the sequence datasets, we pooled individuals according to population (Table S1). For chimpanzees, we used the 292 autosomal microsatellites generated by Becquet et al (Becquet et al. 2007). We only used chimpanzees (Table S1) that have no population ambiguity based on geographic and genetic clustering information.

Sequence data

We used 3 sequence datasets (Table 1): The first was generated by Keinan et al. (Keinan et al. 2008), which used whole genome shotgun sequencing (WGS) (Weber and Myers 1997) to sequence 4 East Asians (Chinese and Japanese), 5 North European, 5 West Africans (Yoruba), and 1 Biaka Pygmy. The second dataset was experimentally generated in our own laboratory using a reduced representation shotgun (RRS) library (Altshuler et al. 2000) to sequence 1 San, 1 Australian aborigine, and 1 Mbuti Pygmy, and has not been previously published. Unlike WGS, which fragments the genome at random, RRS produces fragments that cut at specific restriction enzymes, constraining sequences to specific regions of the genome (see details of RRS sequencing below). WGS data from Yoruba, Europeans, and East Asians from WGS were aligned to the sequence from the 3 RRS individuals, allowing for a larger number of pairwise comparisons across populations than was possible with WGS. The third dataset was generated by Caswell et al (Caswell et al. 2008), and consisted of WGS sequence data from 1 Bonobo, 3 Western Chimpanzees (including "Clint", the individual used to generate the chimpanzee reference sequence (2005)), 3 Central Chimpanzees, and 1 Eastern Chimpanzee. We converted divergence values from Caswell et al. into absolute units of substitutions per Kb by assuming

that the Western-Western chimpanzee divergence is approximately equal to WGS European-European divergence (Patterson et al. 2006).

RRS Sequencing. We used restriction enzymes PmeI (5′-GTTT^AAAC-3′) and EcoRI (5′-G^AATTC-3′) to fully digest DNA extracted from cell lines of 5 diverse human DNA samples, using a Reduced Representation Shotgun (RRS) protocol similar to that described in (Altshuler et al. 2000). We ran the products of the 2 restriction enzyme digests on a gel, and cut out a 2-3Kb band, which we expected would isolate approximately the same subset of the genome in each of the samples. Finally, we cloned the fragments into a pUC19 vector using a protocol that required a PmeI overhang on one side and an EcoRI overhang on the other.

We calculated that the same ~30 Mb, or ~1% of the genome, would be isolated in the 5 samples by this experimental protocol. Given the human genome GC content of 41%, PmeI sites are expected to occur every 36 Kb ($0.205^{-2} \times 0.295^{-6}$) for a total of ~86,000 fragments, and EcoRI are expected to occur every 3.1 Kb ($0.205^{-2} \times 0.295^{-4}$), for a total of ~1,000,000 fragments. Given the human genome size of 3.1 Gb, and assuming a Poisson distribution of restriction sites flanked by PmeI and EcoRI, we expect that there will be approximately $2 \times 86{,}000 \times (1{,}000{,}000 - 86{,}000)/(1{,}000{,}000) = 157{,}000$ such fragments in the genome. Of these, we carried out an integral to infer that the proportion of these fragments that are expected to be in the 2-3Kb range is ~15%, which translates to an expectation of ~23,000 fragments of 2-3Kb for sequencing in each sample. Since each fragment we analyzed was sequenced from both ends with an expected 500-800bp per read, the total amount of sequence that we expected in our "reduced representation" of the genome was about $23{,}000 \times 1.3$ kb = 30Mb. The advantage of RRS over WGS is that with deterministic fragmentation of the genome, the sequences that we obtained in distinct individuals were expected to overlap with greatly increased probability, so that we required substantially less sequencing to obtain genome overlaps from different samples.

We carried out RRS sequencing on two San male samples from HGDP (HGDP_988 and HGDP_991), two Mbuti Pygmy females from the Coriell Cell Repositories (NA10493 and NA10496), and one Australian Aborigine female from the European Collection of Cell Cultures (ECCAC_9118). We attempted to sequence 15,360 reads (7,680 paired ends) from each sample,

and then aligned the reads to the reference human genome sequence, NCBI Build 35, using ssahaSNP (Ning, Cox, and Mullikin 2001) with stringent NQS parameters of Qsnp>=40, Qflank>=15, Nflank=5, maxFlankDiff=1, and maxSNPs/kb<15. Reads that map to multiple places in the genome with nearly identical scores are removed from further analysis. After alignment and filtering, we had data from 11,687 reads in HGDP_998 (5,656,804 bp meeting neighborhood quality score thresholds), 11,500 reads in HGDP_991 (5,359,356 bp), 11,848 reads in NA10493 (5,702,532 bp), 11,905 reads in NA10496 (5,486,017 bp), and 12,193 reads in ECCAC_9118 (6,034,676 bp).

We note that in this study we do not examine overlaps of RRS libraries, even though such comparisons were the original intention the RRS data collection strategy. This is because we found that if the same section of the genome passes through the RRS process in two or more chromosomes, they are in practice biased to be too closely related to each other in time (the inferred t_{MRCA} was systematically lower than the value obtained based on microsatellite ASD). We hypothesize that this reflects the fact that to enable a comparison between two RRS libraries, two haplotypes must be identical at both the PmeI (8 bp) and EcoR1 (6 bp) restriction cut sites, which requires identity at 14 = 8 + 6 bases. By requiring that pairs of haplotypes match at 14 bases, we are biasing the haplotypes that we analyze to be ones with fewer mutations separating them, and thus to be more closely related to each other (in time) than the average pair of sequences in the genome. It is straightforward to show that this generates an appreciable (if small) downward bias in the divergence time estimate, which we in fact observed.

SNP data

We used the HGDP autosomal 650K SNPs (Li et al. 2008).

Computation of genetic distances for microsatellites and sequences

For microsatellites, we computed the unbiased sample statistic of ASD, which is theoretically proportional to t_{MRCA} assuming that that the GSMM is valid (Goldstein et al. 1995a). It is important to realize that the average t_{MRCA} across the genome can be estimated directly from genetic data (using either microsatellite ASD or per base pair sequence divergence). It is a

property of the samples that are being analyzed, and can be estimated empirically without making any assumptions about the demographic history of populations.

For a single locus, ASD works as follows: Suppose we have population A with n_A individuals ($2n_A$ alleles) and population B with n_B individuals ($2n_B$ alleles). We take an allele from each population, perform a subtraction, and square the result. Then, the single locus ASD is the average of all allele pairs defined as follows:

$$ASD = \frac{1}{2n_A \cdot 2n_B} \sum_{i=1}^{2n_A} \sum_{j=1}^{2n_B} (A_i - B_j)^2$$

It can be shown that ASD is very similar to the total variance of all samples between two populations. Furthermore, the within population ASD (not explicitly shown) is equal to twice the variance of the sampled population.

Next, we averaged ASD over multiple loci. We assumed that the microsatellite loci are independent since they were selected for the purpose of linkage analysis to be distantly-spaced across the genome. Thus, the standard error is simply the standard deviation of ASD across all loci divided by the square root of the number of loci. We did not correct for mutation rate heterogeneities across loci, because their empirical values were unknown. More importantly, we did not normalize across loci to equalize the t_{MRCA} of each locus, because in general t_{MRCA} are different for each locus due to different gene genealogies (Rosenberg 2002).

To compute genetic distances for pairwise aligned sequences, we counted nucleotide differences to obtain sequence divergences. Assuming that the molecular clock hypothesis is true for sequence divergence (in which the genome-average nucleotide substitution rate is constant since human-chimpanzee speciation), then sequence divergence is strictly proportional to t_{MRCA}. Because of linkage disequilibrium, nearby divergent sites are dependent, and standard errors of sequence divergence were computed via a block jackknife approach (Keinan et al. 2007).

Computation of F_{ST} for microsatellites and SNPs

While there are multiple methods to compute F_{ST}, our goal is to have an unbiased F_{ST} statistic for microsatellites that is also coherent with SNP F_{ST}. F_{ST} is defined as

$$F_{ST} = 1 - \frac{H_S}{H_T}$$

H_S is the average heterozygosity across all populations. H_T is the heterozygosity of all populations pooled together. Slatkin (1995) showed that in the context of the generalized stepwise mutation model heterozygosity is simply the variance of the allelic distribution at a particular locus. However, we do not use his sample statistic directly because he requires equal sample sizes, and instead use one that we derived that allows for unequal sample sizes.

A pairwise F_{ST} estimator at a single microsatellite locus. Suppose we have two populations, each with allelic distributions described by random variables A and B. H_S is trivial:

$$H_S = \frac{1}{2}\mathbf{var}(A) + \frac{1}{2}\mathbf{var}(B)$$

H_T is found using the law of total variance, yielding:

$$H_T = \frac{1}{2}\mathbf{var}(A) + \frac{1}{2}\mathbf{var}(B) + \frac{1}{4}(\mathbf{E}[A] - \mathbf{E}[B])^2$$

Combining terms, we have an F_{ST} estimator:

$$F_{ST} = 1 - \frac{H_S}{H_T} = \frac{(\mathbf{E}[A] - \mathbf{E}[B])^2}{2\mathbf{var}(A) + 2\mathbf{var}(B) + (\mathbf{E}[A] - \mathbf{E}[B])^2}$$

Coherence with SNP F_{ST}. SNP loci are biallelic, and hence random variables A and B are Bernoulli distributed with minor allele frequency (MAF) parameters p_A and p_B. SNP F_{ST} becomes:

$$\begin{aligned} \text{SNP } F_{ST} &= \frac{(p_A - p_B)^2}{2p_A(1-p_A) + 2p_B(1-p_B) + (p_A - p_B)^2} \\ &= \frac{d^2}{P(1-P)} \end{aligned}$$

This is a classical definition for SNP F_{ST}, where *P* is the MAF of the two populations combined and *d* is the difference between the MAF of a population and *P*:

$$p_A = P + d$$
$$p_B = P - d$$

Hence, SNP F_{ST} is just a special case of microsatellite F_{ST}.

Unbiased sample statistic for F_{ST}. We compute unbiased sample statistics (which we refer to using a "hat" notation) separately for the numerator and denominator, then calculated the ratio.

$$\widehat{F_{ST}} = \frac{\widehat{N}}{\widehat{D}}$$

Given sample sizes and unbiased sample statistics for mean and variance, the numerator becomes:

$$\widehat{N} = (\widehat{\mu_A} - \widehat{\mu_B})^2 - \frac{\widehat{\mathbf{var}(A)}}{n_A} - \frac{\widehat{\mathbf{var}(B)}}{n_B}$$

Similarly, the denominator becomes:

$$\widehat{D} = 2\widehat{\mathbf{var}(A)} + 2\widehat{\mathbf{var}(B)} + \widehat{N}$$

Multiple loci. All discussion so far has been for a single microsatellite locus. For K loci, we first compute K unbiased sample statistics, each for numerator and denominator. Then we separately average the numerator and denominator, and finally compute the ratio. This strategy avoids numerical instability issues of averaging ratios (namely, when denominators are small at certain loci).

$$\widehat{F_{ST}} = \frac{\sum_i \widehat{N_i}}{\sum_i \widehat{D_i}}$$

Standard error across loci is computed via the jackknife method (Efron and Gong 1983). SNP F_{ST} quantities and standard errors were computed using EIGENSOFT (Patterson, Price, and Reich 2006).

Relating F_{ST} and ASD in microsatellites

F_{ST} and ASD are closely related. From the above it is clear that F_{ST} is a function of first and second-order moments of allelic distributions. Furthermore, it is known (Goldstein et al. 1995a) that the ASD estimator is:

$$ASD = \mathbf{var}(A) + \mathbf{var}(B) + (\mathbf{E}[A] - \mathbf{E}[B])^2$$

Define X as the sum of intra-population variances. Define Y as inter-population variance.

$$X = \mathbf{var}(A) + \mathbf{var}(B)$$
$$Y = (\mathbf{E}[A] - \mathbf{E}[B])^2$$

$$ASD = X + Y$$
$$F_{ST} = \frac{Y}{2X + Y}$$

Now the relationship of F_{ST} and ASD is clear. ASD closely resembles the total variance of allelic distributions of populations A and B combined. F_{ST} is the ratio of inter-population variance to total variance.

Results

Microsatellite ASD and sequence divergence are linearly related

To test empirically whether the microsatellite ASD statistic (Goldstein et al. 1995a) can be an unbiased estimate of t_{MRCA}, we used genomic sequence divergence as a "gold standard", and assessed how closely the microsatellite inferences matched this number. We restricted our analysis to pairs of populations for which we had both extensive genome sequence alignments and large scale microsatellite data. We first used sequence datasets to compute autosomal sequence divergence, which was assumed to be proportional to the average t_{MRCA}. This formed our gold standard molecular clock. For the same pairs of populations, we then computed ASD using microsatellite data. Comparing sequence divergence to ASD provided a metric for the accuracy of the microsatellite molecular clock, assessed in terms of linearity (correlation coefficient) and standard errors.

Figure 1 plots sequence divergence against microsatellite ASD. For WGS humans (Panel A), the correlation coefficient is *r=0.989 (P=4.9e-7, 95% CI 0.946-0.998)*. For RRS humans (Panel B), *r=0.979 (P=2.2e-10, 95% CI 0.937-0.993)*. For chimpanzees (Panel C), *r=0.986 (P=2.7e-4, 95% CI 0.877-0.999)*. Figure 1 suggests the following:

- Sequence divergence and microsatellite ASD are linearly related: The regressions have correlation coefficients all greater than 0.97. Since sequence divergence is known to be proportional to t_{MRCA}, microsatellite ASD is linear to t_{MRCA}. Interestingly, however, the regression lines do not intersect the origin, a point we return to below.
- Combining microsatellite loci yields a reasonably precise molecular clock, and in principle supports precise inferences about history. Examining the standard errors in Figure 1A, the 783 human microsatellite loci are approximately 2.5 times less precise than that of Biaka Pygmy sequence alignments. Thus, 783 microsatellite loci correspond to about 7.2 Mb of alignment of two WGS sequences (Table 1). In turn, 1 microsatellite is "worth" approximately 10 Kb of shotgun sequencing, which is expected to contain 10 nucleotide mutations between 2 modern humans.

- The microsatellite molecular clock appears to be linear for at least 2 million years: It has been shown theoretically that in the presence of severe range constraints, microsatellite ASD should lose its linear behavior after about 10,000 generations (Feldman et al. 1997), which is 250,000 years assuming 25 years per generation. Bonobos are a distinct species from chimpanzees, and are thought to have t_{MRCA} of around 2.2 million years (Caswell et al. 2008) averaged across the genome, yet the linearity in Figure 1C still applies to bonobo-chimpanzee divergence. Therefore, encouragingly, the duration of ASD linearity is at least 10 times that of theoretical predictions, suggesting range constraints are not as severe as previously imagined.

Non-zero y-intercept in Figure 1. While these results demonstrate microsatellites' usefulness in estimating t_{MRCA}, there is a non-zero y-intercept (Figure S1), oddly suggesting that zero sequence divergence (t_{MRCA}=0) is associated with a positive ASD. We used simulations to investigate the possibility that microsatellite genotyping error caused the elevated ASD relative to its true value. Assuming a typical genotype error rate of 1% with error being randomly distributed at ±1 repeat length (Weber and Broman 2001), we can only explain 10% of the offset. It is possible, however, that the most pertinent error in microsatellite genotyping is not miscalling microsatellite lengths by a single repeat length, but instead, miscalling heterozygous genotypes as homozygous, which can easily occur with microsatellites (Weber and Broman 2001). Missing of heterozygotes would have the effect of producing multi-step mutation errors, which would result in a much larger inflation in the ASD (due to the squaring of the difference in allele lengths), and could plausibly explain our significantly non-zero y-intercept. Alternatively, the relationship between ASD and t_{MRCA} could be globally nonlinear, but easily linearizable in our time window. Whatever the cause for our observations, these results indicate that for population genetic analysis, it is important to use a calibration curve (such as Figure 1) to convert ASD to sequence divergence, correcting for the inflated estimate of divergence time from microsatellite ASD.

The microsatellite clock reveals deep lineages of human genetic variation. The microsatellite data show that the San, Biaka Pygmy, and Mbuti Pygmy Africans are more diverged in their pairwise t_{MRCA} from non-African populations than are Yoruba West Africans. These results are

consistent with an analysis of microsatellite data by Zhivotovsky et al. (Zhivotovsky, Rosenberg, and Feldman 2003), but strengthen their result because microsatellite and sequence divergence concur (Figure 1A,1B). It was already known based on mitochondrial DNA and Y chromosome data that the San and Mbuti contain deeply diverged lineages, but our results and those of Zhivotovsky et al. using autosomal microsatellite data show definitively that these populations are outgroups to all other populations.

Inferred pairwise sequence divergence of HGDP populations. An immediate application of the regressions from Figure 1 is to infer sequence divergences for the remaining HGDP populations in which we lack sequence data. Figure 2 is a matrix plot showing the inferred divergences (hence inferred t_{MRCA}). In this plot, the San and Pygmy Africans are the only populations equidistant to all other populations, further suggesting that these populations are the most deeply diverged.

Microsatellite F_{ST} accurately estimates allele frequency differentiation

F_{ST} measures the degree of differentiation between populations. Given genetic diversity data for 2 populations, F_{ST} (a quantity between 0 and 1) is the ratio of inter-population variance to total variance. When F_{ST} is appropriately transformed (Slatkin 1991; Patterson 2007), one can infer the genetic drift that occurred between two populations since they split. In particular, one can estimate the population split time (t_{pop}) in units of 2N, where N is the effective population size, under the assumption that populations have been constant in size since their divergence. We note that in human populations, t_{pop} and t_{MRCA} are different by an order of magnitude: for Africans versus non-Africans, the average t_{MRCA} is thought to be ~500,000 years ago, while t_{pop} is thought to be 40,000-80,000 years ago (Keinan et al. 2008). As we have shown that the microsatellite molecular clock works for time depths of at least 2 million years, we can be confident that it also works for time separations that are an order of magnitude less.

F_{ST} is usually estimated based on SNP and sequencing data when available, because uncertainties of the complex microsatellite mutation process confound the interpretation of a microsatellite F_{ST} in terms of history. Assuming the GSMM of microsatellite evolution, however, Slatkin derived a microsatellite-based F_{ST} estimator (Slatkin called it R_{ST}) (Slatkin 1995) that

should be identical to SNP-based F_{ST}. The empirical analyses using Slatkin's estimator have been encouraging. For example, based on <300 SNPs (Fischer et al. 2006) and <300 microsatellites in four chimpanzee populations, Becquet et al. (Becquet et al. 2007) showed that the SNP F_{ST} and microsatellite F_{ST} were concordant.

As of today, the richest data sets with both genome-wide SNPs and large numbers of microsatellites are those from HGDP (Rosenberg et al. 2002; Li et al. 2008). We computed and compared F_{ST} based on SNPs and microsatellites in these samples. An important distinction between the comparison we present here and that of the previous section (where we examined ASD) is that we do not assume SNP-based F_{ST} as gold-standard.

Empirical relationship between microsatellite and SNP F_{ST}. Figure 3A plots SNP F_{ST} on the horizontal axis and microsatellite F_{ST} on the vertical axis. There are 53 populations in HGDP, and hence 1,378 data points (53 choose 2) with standard errors. The linearity is clear and the regression lines intersect the origin. However, there are two distinct lines for F_{ST}>0.1. The 1,035 pairwise comparisons of non-Africans populations (46 choose 2) have regression line slope of 0.91 and correlation coefficient *r=0.983 (95% CI 0.982-0.986)*. The African vs non-African comparisons have a distinctly smaller slope of 0.73 and *r=0.969 (95% CI 0.962-0.975)*. It is evident that for F_{ST}>0.1, SNP-based quantities are larger than microsatellite quantities when Africans are involved. We next investigate the possible reasons for this discrepancy.

SNP ascertainment bias can explain the discrepancy between the two F_{ST} measurements. To investigate whether SNP ascertainment bias can explain the phenomena in Figure 3A, we simulated SNP ascertainment as follows:

1. Demographic model 1 (Figure S2A): The goal of this model is to generate a wide range of F_{ST} values, larger than that of real human populations. As shown in Figure S2A, the size of population A is fixed at N_0=10,000. The size of population B varies from $0.01N_0$ to N_0, enabling an F_{ST}(A,B) range of 0.01 to 0.45. t_{AB}, the population separation time, is fixed at 400 generations.
2. Coalescent simulation and mutation generation: Given demographic model 1, we used Hudson's ms coalescent simulator (Hudson 2002) to generate trees and mutations assuming

the infinite-sites model. Microsatellite alleles were then generated according to the SMM. Thus, each mutation is added or subtracted, at random, to the microsatellite lengths.

3. Ascertaining SNPs: To generate ascertainment bias-free SNPs, we recorded the derived allele frequency of each population across all loci. To generate SNPs affected by ascertainment bias, for each locus we took 2 samples and examined the allele. If and only if they are different, we recorded the data from the locus, excluding the 2 used for ascertaining. We ascertain in three ways: (1) 2 samples from population A, (2) 2 samples from population B, and (3) 1 sample from each population.
4. F_{ST} calculation: With the data sets generated from simulated microsatellites and SNPs, we calculated F_{ST}. We examined if any of the three ascertainment schemes could generate the same directionality of bias as such in Figure 3A.
5. Enhanced demographic model (Figure S2B): The goals of this model are to more closely mimic real human history, and to apply the appropriate ascertainment scheme to all populations simultaneously and observe if ascertainment can cause the bias in Figure 3A. As shown in Figure S2B, populations A, B, C, D are approximately Africans, Europeans, East Asians, and Native Americans, respectively. We used the same ascertainment scheme as above and estimated F_{ST}.

Simulations can replicate the effect of ascertainment bias on SNPs. For demographic model 1, we denoted population A (the one with the larger effective population size) as "Africans" and population B as "non-Africans". The simulation results are shown in Figure 3B. Without ascertainment, both F_{ST} are identical. Ascertainment using 2 Africans showed negligible bias. Ascertainment using 2 non-Africans negatively biased SNP F_{ST}. Ascertainment using 1 sample from each population positively biased SNP F_{ST}. Compared to the real HGDP data (Figure 3A), ascertaining from 1 African and 1 non-African generated the same directional effect. This result is reasonable, because SNPs on medical genetics arrays were discovered as differences between a non-African chromosome and the reference human genome. The reference human genome sequence has a substantial amount of African ancestry because RPCI-11, the Bacterial Artificial Chromosome library that has contributed ~74% of the human genome reference sequence (Lander et al. 2001), is likely to be derived from an African American (Reich et al. in preparation).

We applied the 1 African 1 non-African ascertainment scheme to demographic model 2. There are 4 populations in the model, producing 6 F_{ST} values in total (4 choose 2). As shown in Figure 3C, the non-African vs non-African comparisons show little bias. The African vs non-African comparisons show a positively biased SNP F_{ST}. Thus, we have demonstrated that SNP ascertainment bias can generate the discrepancy in Figure 3A.

A unifying view of ASD and microsatellite F_{ST}

Having established the accuracy of both microsatellite ASD and F_{ST}, we next show a two dimensional view of HGDP microsatellite data that illustrates important historical events.

Just as sequence variation data contains information on both divergence time and genetic drift, it can be shown (Materials and Methods) that microsatellite ASD and F_{ST} are functions of two independent quantities: inter-population variance and intra-population variance. Using the HGDP microsatellite data as previously described, in Figure 4 we projected the data onto the 2 orthogonal statistics: inter-population variance (horizontal axis) and intra-population variance (vertical axis). Again we have 1,378 data points, and lines of constant ASD and F_{ST} are marked. Above the thick black line are Africans vs all populations, and below are non-Africans vs non-Africans. This figure suggests the following:

- With the exception of Native American to Native American comparisons, lines of constant ASD have slopes similar to slopes of the data points. Africans populations are equidistant from non-Africans. This is expected from the "out-of-Africa" migration hypothesis in which all non-African populations form a clade (Cavalli-Sforza and Feldman 2003).
- Projecting onto lines of constant ASD, we see a clear gap (thick black line) between Africans and non-Africans. This confirms that there is a time difference between the out-of-Africa event and the rest of migration events. There is a second gap for the Native Americans, confirming that migration into America is a significantly more recent event (Cavalli-Sforza and Feldman 2003).
- Examining Africans vs all populations, F_{ST} projections show the drift out-of-Africa: The top left rectangle shows Africans vs Africans, followed by Europeans and Asians, then Pacific Islanders, and finally Native Americans (the rectangle crossing the largest F_{ST}

values). The series of events is in agreement with progressive bottleneck events leading out of Africa (Ramachandran et al. 2005).

Discussion

The fact that microsatellites are useful as molecular clocks has immediate applications: First, as described above (and in Figure S3), we were able to use the clocklike nature of microsatellites to provide clear evidence that the San, Biaka, and Mbuti Pygmy branch off near the root of the tree of human populations, with all other populations (including West Africans) forming a clade. Note that all of our analyses are restricted to population average coalescent time, a quantity distinctly different and much more ancient than population split time. Second, we can use microsatellite data to correct inferences about F_{ST} based on high density SNP array data. SNP F_{ST} values can be precise, but they are affected by ascertainment bias. Potentially, we can use microsatellite F_{ST} to correct most of this bias. For example, based on Figure 3, we estimate that all pairwise autosomal F_{ST}'s between African and non-African populations in the Li et al. HGDP data (Li et al. 2008) are too large by a factor of 1.25 for F_{ST} values >0.1. By deflating all these F_{ST} values by this factor, we can obtain a pairwise F_{ST} matrix that is likely to be more accurate.

We finally note that our results are intriguing because in principle, they offer a way to obtain a direct estimate of the human per nucleotide mutation rate for sequence divergence data. To date, it has been impossible to obtain a direct estimate of the human per base pair mutation rate because the rate is too low (about 2×10^{-8} per nucleotide per generation) to permit observation in real human data. However, the microsatellite mutation rate is sufficiently high (10^{-3} to 10^{-4} per generation) that novel mutations are frequently directly observed in families (Weber and Wong 1993). By directly estimating the microsatellite mutation rate and mutation process in families, and then extrapolating to sequence divergence, we should be able to estimate the human per base pair mutation rate, and infer the dates of important historical events, like the divergence times of human and chimpanzees (Patterson et al. 2006).

Acknowledgments

We thank Alon Keinan for his suggestions about the design of the SNP ascertainment bias simulations. DR was supported by a Burroughs Wellcome Career Development Award in the Biomedical Sciences. JS was supported by the Bioinformatics and Integrative Genomics PhD training grant by NIH. JCM was supported by the Intramural Research Program of the National

Human Genome Research Institute, NIH. We are grateful to Nicole Stange-Thomann and Julie Neubauer for preparing the Reduced Representation Shotgun data.

References

Altshuler, D., V. J. Pollara, C. R. Cowles, W. J. Van Etten, J. Baldwin, L. Linton, and E. S. Lander. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature **407**:513-516.

Amos, W., and D. C. Rubinstzein. 1996. Microsatellites are subject to directional evolution. Nat Genet **12**:13-14.

Becquet, C., N. Patterson, A. C. Stone, M. Przeworski, and D. Reich. 2007. Genetic structure of chimpanzee populations. PLoS Genet **3**:e66.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, and L. L. Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368**:455-457.

Caswell, J. L., S. Mallick, D. J. Richter, J. Neubauer, C. Schirmer, S. Gnerre, and D. Reich. 2008. Analysis of chimpanzee history based on genome sequence alignments. PLoS Genet **4**:e1000057.

Cavalli-Sforza, L. L., and M. W. Feldman. 2003. The application of molecular genetic approaches to the study of human evolution. Nat Genet **33 Suppl**:266-275.

Chimpanzee Sequencing and Analysis Consortium 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**:69-87.

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen. 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res **15**:1496-1502.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall, N. A. Rosenberg, and J. K. Pritchard. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet **38**:1251-1260.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer. 1994. Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci U S A **91**:3166-3170.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380**:152-154.

Efron, B., and G. Gong. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician **37**:36-48.

Ellegren, H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet **16**:551-558.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet **5**:435-445.

Feldman, M. W., A. Bergman, D. D. Pollock, and D. B. Goldstein. 1997. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics **145**:207-216.

Fischer, A., J. Pollack, O. Thalmann, B. Nickel, and S. Paabo. 2006. Demographic history and genetic differentiation in apes. Curr Biol **16**:1133-1138.

Friedlaender, J. S., F. R. Friedlaender, F. A. Reed, K. K. Kidd, J. R. Kidd, G. K. Chambers, R. A. Lea, J. H. Loo, G. Koki, J. A. Hodgson, D. A. Merriwether, and J. L. Weber. 2008. The genetic structure of Pacific Islanders. PLoS Genet **4**:e19.

Goldstein, D. B., and D. D. Pollock. 1997. Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. J Hered **88**:335-342.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza, and M. W. Feldman. 1995a. An evaluation of genetic distances for use with microsatellite loci. Genetics **139**:463-471.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza, and M. W. Feldman. 1995b. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci U S A **92**:6723-6727.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**:337-338.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.

Keinan, A., J. C. Mullikin, N. Patterson, and D. Reich. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet **39**:1251-1255.

Keinan, A., J. C. Mullikin, N. Patterson, and D. Reich. 2008. Accelerated genetic drift on chromosome X during the human disperal out of Africa. Nat Genet (in press).

Kimmel, M., and R. Chakraborty. 1996. Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor Popul Biol **50**:345-367.

Kimmel, M., R. Chakraborty, J. P. King, M. Bamshad, W. S. Watkins, and L. B. Jorde. 1998. Signatures of population expansion in microsatellite repeat data. Genetics **148**:1921-1930.

Kruglyak, S., R. T. Durrett, M. D. Schug, and C. F. Aquadro. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A **95**:10774-10778.

Levinson, G., and G. A. Gutman. 1987. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. Nucleic Acids Res **15**:5323-5338.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science **319**:1100-1104.

Nauta, M. J., and F. J. Weissing. 1996. Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics **143**:1021-1032.

Ning, Z., A. J. Cox, and J. C. Mullikin. 2001. SSAHA: a fast search method for large DNA databases. Genome Res **11**:1725-1729.

Ohta, T., and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res **22**:201-204.

Paetkau, D., L. P. Waits, P. L. Clarkson, L. Craighead, and C. Strobeck. 1997. An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. Genetics **147**:1943-1957.

Patterson, N. 2007. Notes on Fst.

Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PLoS Genet **2**:e190.

Patterson, N., D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature **441**:1103-1108.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A **102**:15942-15947.

Ramachandran, S., N. A. Rosenberg, M. W. Feldman, and J. Wakeley. 2008. Population differentiation and migration: Coalescence times in a two-sex island model for autosomal and X-linked loci. Theor Popul Biol.

Reich, D. E., and D. B. Goldstein. 1998. Genetic evidence for a Paleolithic human population expansion in Africa. Proc Natl Acad Sci U S A **95**:8119-8123.

Roder, M. S., V. Korzun, K. Wendehake, J. Plaschke, M. H. Tixier, P. Leroy, and M. W. Ganal. 1998. A microsatellite map of wheat. Genetics **149**:2007-2023.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. Theor Popul Biol **61**:225-247.

Rosenberg, N. A. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet **70**:841-847.

Rosenberg, N. A., S. Mahajan, C. Gonzalez-Quevedo, M. G. Blum, L. Nino-Rosales, V. Ninis, P. Das, M. Hegde, L. Molinari, G. Zapata, J. L. Weber, J. W. Belmont, and P. I. Patel. 2006. Low levels of genetic divergence across geographically and linguistically diverse populations from India. PLoS Genet **2**:e215.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet **1**:e70.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. 2002. Genetic structure of human populations. Science **298**:2381-2385.

Sainudiin, R., R. T. Durrett, C. F. Aquadro, and R. Nielsen. 2004. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. Genetics **168**:383-395.

Shimoda, N., E. W. Knapik, J. Ziniti, C. Sim, E. Yamada, S. Kaplan, D. Jackson, F. de Sauvage, H. Jacob, and M. C. Fishman. 1999. Zebrafish genetic map with 2000 microsatellite markers. Genomics **58**:219-232.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457-462.

Slatkin, M. 1991. Inbreeding coefficients and coalescence times. Genet Res **58**:167-175.

Szpiech, Z. A., M. Jakobsson, and N. A. Rosenberg. 2008. ADZE: a rarefaction approach for counting alleles private to combinations of populations. Bioinformatics **24**:2498-2504.

Valdes, A. M., M. Slatkin, and N. B. Freimer. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133**:737-749.

Wang, S., C. M. Lewis, M. Jakobsson, S. Ramachandran, N. Ray, G. Bedoya, W. Rojas, M. V. Parra, J. A. Molina, C. Gallo, G. Mazzotti, G. Poletti, K. Hill, A. M. Hurtado, D. Labuda, W. Klitz, R. Barrantes, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, L. T. Tsuneto, E. Llop, F. Rothhammer, L. Excoffier, M. W. Feldman, N. A. Rosenberg, and A. Ruiz-Linares. 2007. Genetic variation and population structure in native Americans. PLoS Genet **3**:e185.

Wang, S., N. Ray, W. Rojas, M. V. Parra, G. Bedoya, C. Gallo, G. Poletti, G. Mazzotti, K. Hill, A. M. Hurtado, B. Camrena, H. Nicolini, W. Klitz, R. Barrantes, J. A. Molina, N. B. Freimer, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, L. T. Tsuneto, J. E. Dipierri, E. L. Alfaro, G. Bailliet, N. O. Bianchi, E. Llop, F. Rothhammer, L. Excoffier, and A. Ruiz-Linares. 2008. Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet **4**:e1000037.

Weber, J. L., and K. W. Broman. 2001. Genotyping for human whole-genome scans: past, present, and future. Adv Genet **42**:77-96.

Weber, J. L., and E. W. Myers. 1997. Human whole-genome shotgun sequencing. Genome Res **7**:401-409.

Weber, J. L., and C. Wong. 1993. Mutation of human short tandem repeats. Hum Mol Genet **2**:1123-1128.

Xu, X., M. Peng, and Z. Fang. 2000. The direction of microsatellite mutations is dependent upon allele length. Nat Genet **24**:396-399.

Zhivotovsky, L. A. 2001. Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. Mol Biol Evol **18**:700-709.

Zhivotovsky, L. A., N. A. Rosenberg, and M. W. Feldman. 2003. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet **72**:1171-1186.

Zuckerkandl, E., and L. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. Pp. 189-225 *in* M. Kasha, and B. Pullman, eds. Horizons in Biochemistry. Academic Press, New York.

Figure 1

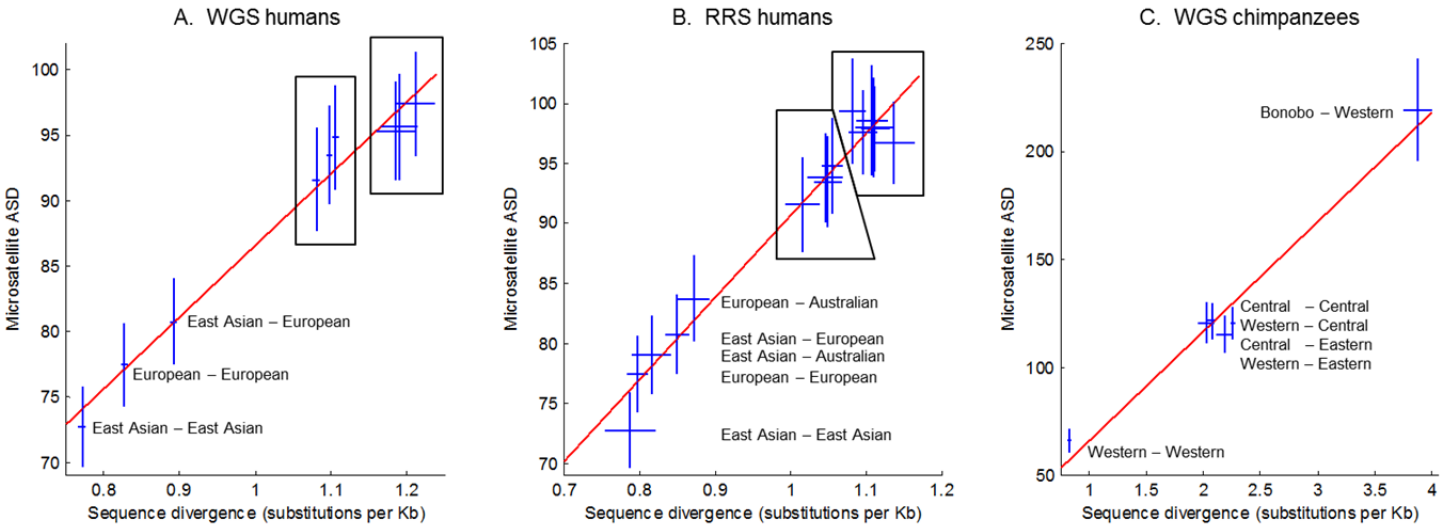


Figure 1. Microsatellite ASD is linear with sequence divergence. Horizontal axes are sequence divergences measured in substitutions per kilobase (Kb), which we assume is an accurate "gold standard". Vertical axes are microsatellite ASD values. Crosshairs are data with standard errors for each population pair. The linear regression line is shown. For WGS humans (A), the correlation coefficient is r=0.989 (P=4.9e-7, 95% CI 0.946-0.998). In the left box are Yoruba versus (top to bottom): European, East Asian, and Yoruba. In the right box are Biaka Pygmy versus (top to bottom): European, Yoruba, East Asian. For RRS humans (B), r=0.983 (P=5.3e-11, 95% CI 0.949-0.995). In the left box are Yoruba versus (top to bottom): European, Australian Aborigine, East Asian, and Yoruba. In the right box is Biaka Pygmy versus: European, Yoruba, East Asian; also are San versus: Yoruba, European, East Asian. For chimpanzees (C), r=0.986 (P=2.7e-4, 95% CI 0.877-0.999).

Figure 2

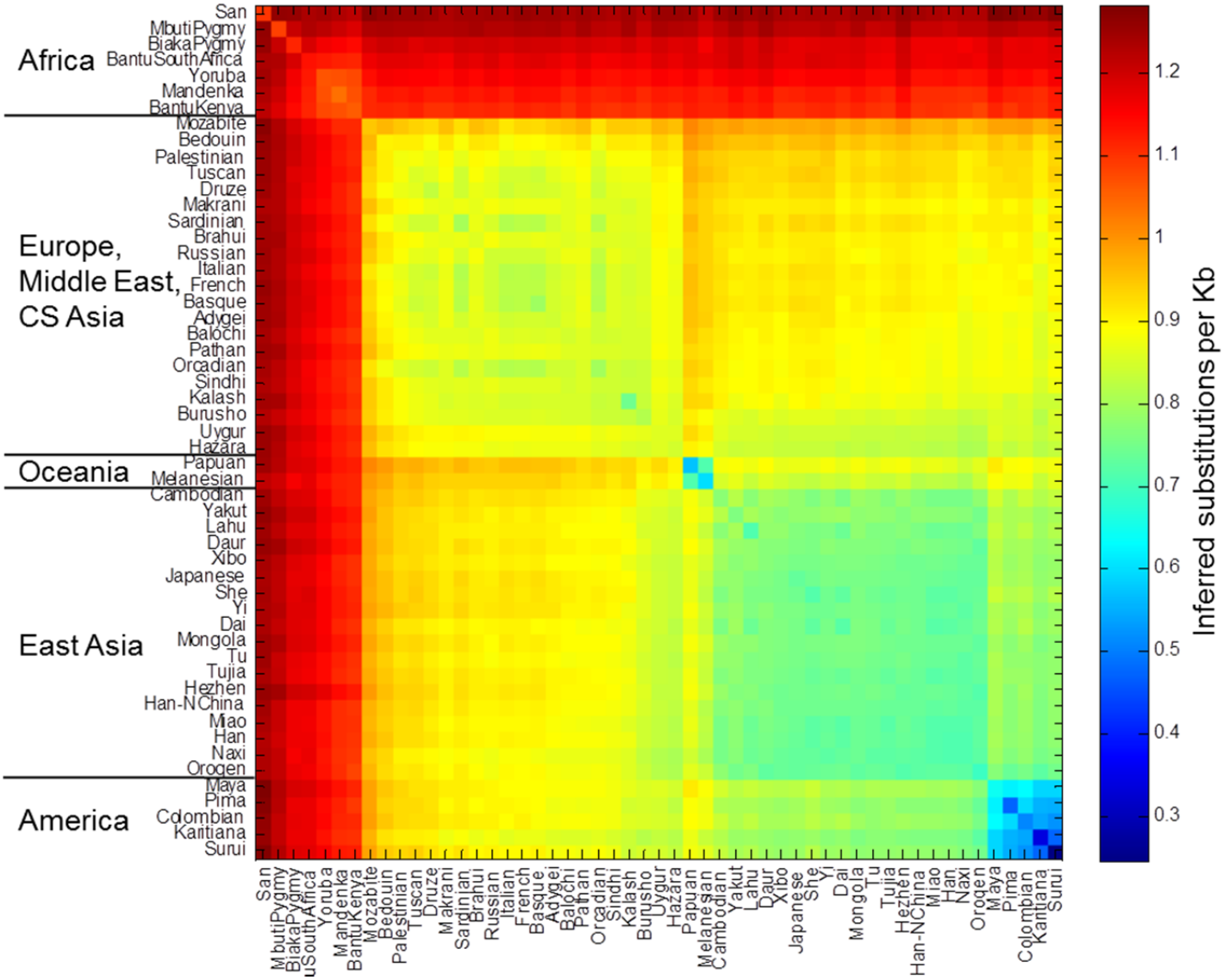


Figure 2. Inferred pairwise sequence divergences of HGDP populations. Microsatellite ASD for each pair of populations in HGDP is computed. Then using regression from Figure 1A, we inferred the divergence of each population pair in substitutions per Kb. The grayscale intensities display the range of divergences. As shown, San and Pygmy Africans are equidistant from all other populations, suggesting that they have the largest t_{MRCA} to any other human population.

Figure 3

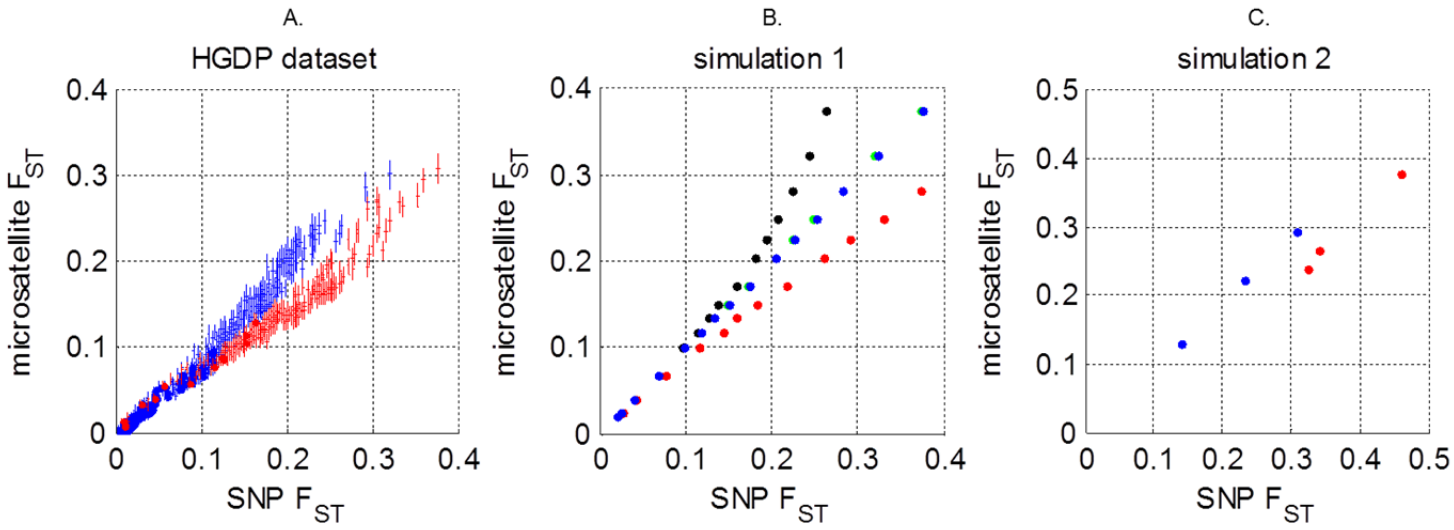


Figure 3. Microsatellite and SNP F_{ST} are almost equivalent, with the discrepancy likely due to SNP ascertainment. Horizontal axes are the SNP F_{ST}. Vertical axes are the microsatellite F_{ST}. In Panel A are F_{ST} computed from real HGDP data. There are (53 choose 2) = 1,378 pairwise population comparisons (data points). Circles and plus signs are data for each population pair. The linearity is clear and the regression lines (not shown) intersect the origin. However, there are two distinct slopes for F_{ST}>0.1. In circles are 1,035 (46 non-African populations, choose 2) non-Africans vs. non-Africans, with regression line slope=0.91 and correlation coefficient 0.983 (P<1e-10, 95% CI 0.982-0.986). In plus signs are Africans vs all populations, with regression line slope=0.73 and correlation coefficient 0.969 (P<1e-10, 95% CI 0.962-0.975). In Panel B are simulated data (demographic model in Figure S2A) with different SNP ascertainment schemes: No ascertainment in circles, ascertaining using 2 samples from population A ("African") in dots, ascertaining using 2 samples from population B ("European") in crosses, and ascertaining using 1 sample from each population in plus signs. In Panel C are simulated data (demographic model in Figure S2B) of 4 populations resembling Africans, Europeans, East Asians, and Native Americans. We used the European-African ascertainment scheme (see text). In circles are non-Africans vs non-Africans. In plus signs are Africans vs non-Africans. For panels B and C, enough loci were simulated such that standard errors are of negligible magnitude.

Figure 4

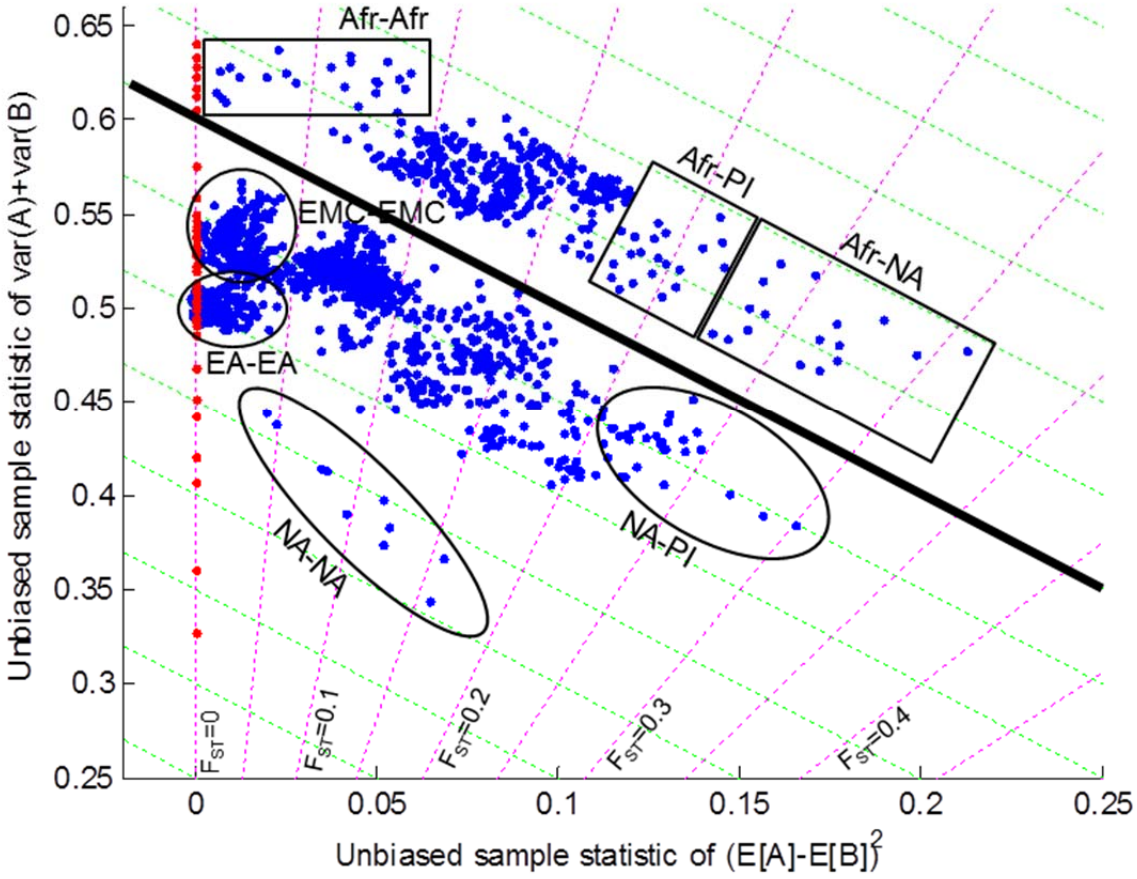


Figure 4. A unifying view of ASD and microsatellite F_{ST}. The horizontal axis is inter-population variance. The vertical axis is intra-population variance. Afr=Africans, NA=Native Americans, PI=Pacific Islanders, EA=East Asians, EMC=Europeans, Middle Easterners, Central South Asians. It is shown (Materials and Methods) that microsatellite F_{ST} and ASD are functions of these two variances. Lines of constant ASD are dashed lines with negative slope. Lines of constant F_{ST} are dashed lines with positive slope. The data are (53 choose 2)=1,378 pairwise HGDP population comparisons. Clearly, this picture segregates populations into distinguishable clusters. Africans vs all are above the thick black line. Non-Africans vs. non-Africans are below the line. Distinguishable clusters are demarcated in ovals and squares.

Figure S1

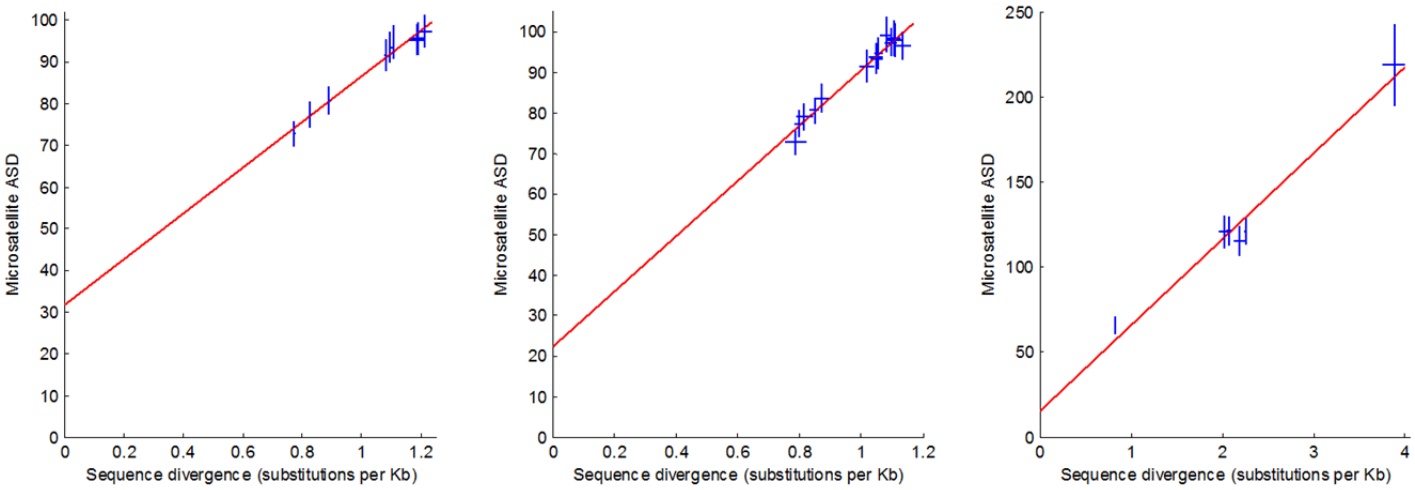


Figure S1. Non-zero Y-intercept of microsatellite ASD vs sequence divergence. This is a zoomed out view of Figure 1. The significant offset implies (strangely) that zero sequence divergence would yield non-zero ASD. Potential explanations for this phenomenon are presented in the main text.

Figure S2

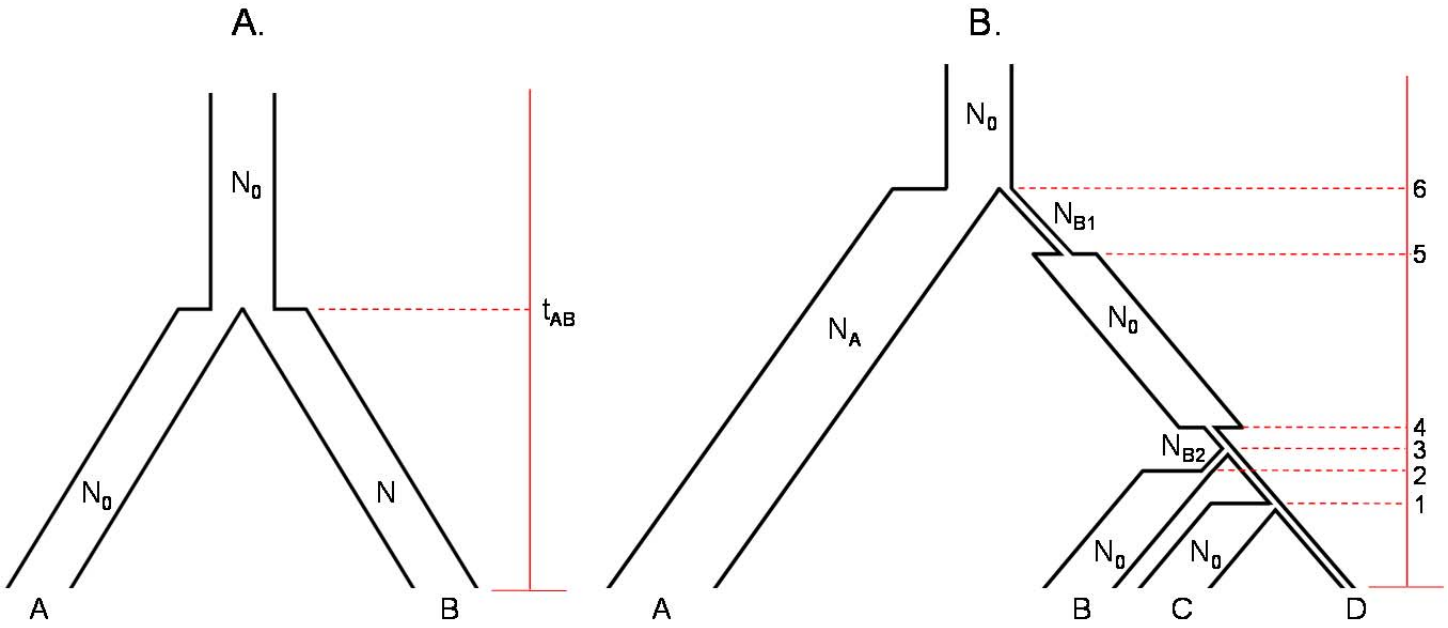


Figure S2. Demographic models for SNP ascertainment bias simulation. In Panel A is a model of 2 populations, with parameters N_0=10,000, t_{AB}=400 generations, N varies from $0.01N_0$ to N_0, enabling an F_{ST} range of 0.01 to 0.45, which is a superset of the real data. Populations A and B can be roughly thought of, as Africans and non-Africans, respectively. In Panel B is a model of 4 populations, with A ≈ Africans, B ≈ Europeans, C ≈ East Asians, D ≈ Native Americans. This is a 2 bottleneck model suggested by Keinan et al. (Keinan et al. 2008), with N_0=10,000, $N_A=1.6N_0$, $N_{B1}=0.02N_0$, $N_{B2}=0.05N_0$, $t_1=0.014\times4N_0$, $t_2=0.016\times4N_0$, $t_3=0.018\times4N_0$, $t_4=0.019\times4N_0$, $t_5=0.107\times4N_0$, $t_6=0.109\times4N_0$.

Figure S3

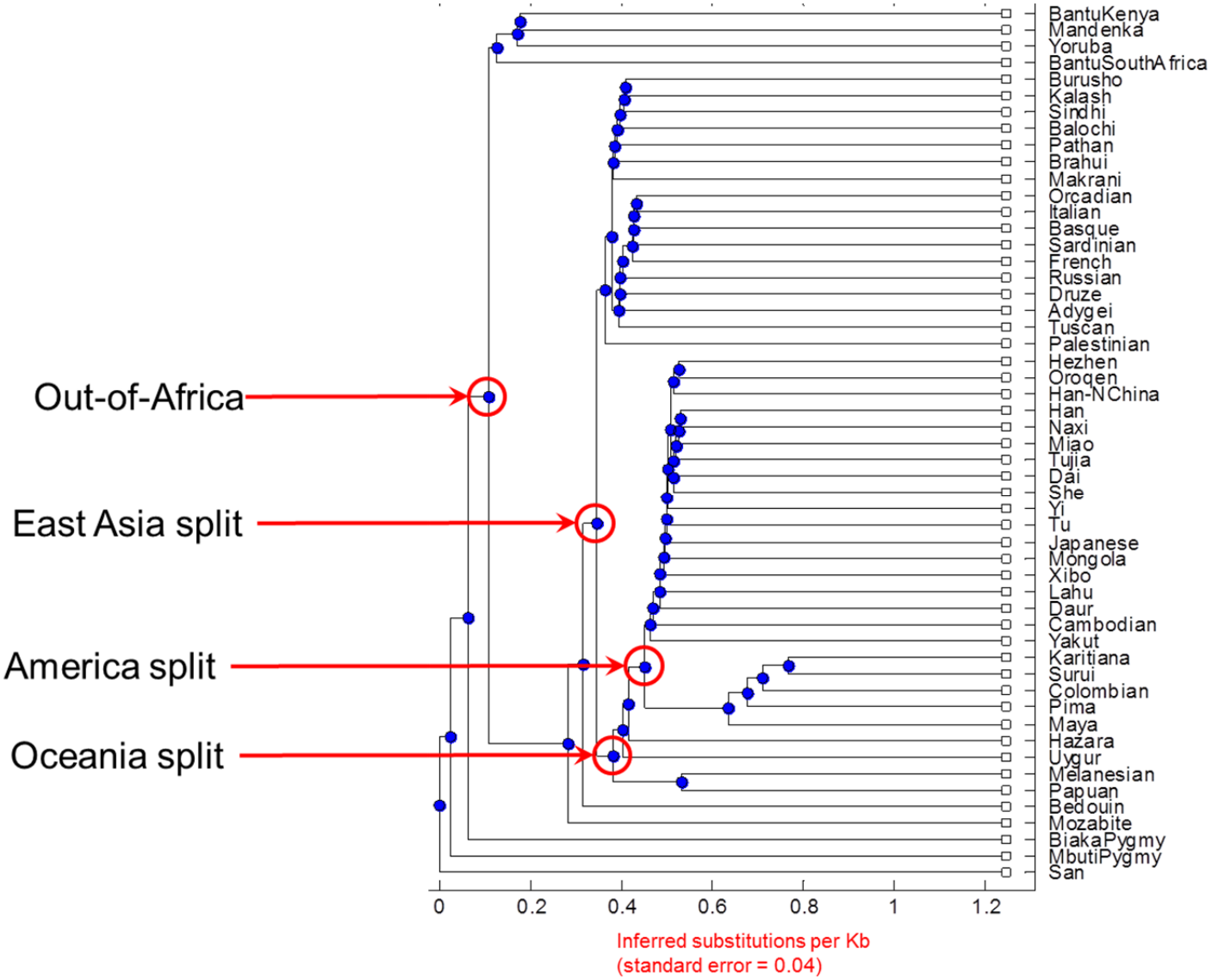


Figure S3. Ultrametric tree. UPGMA tree constructed from the pairwise divergence matrix in Figure 2. The tree is mostly sensible, with clear demarcations of major historical events in a timeline: (1) out-of-Africa, (2) East Asia split, (3) Oceania split, (4) America split. While this tree is interesting and recapitulates the main patterns of human migration that have been observed in previous studies, we note that when there is gene flow in among populations, tree representations are not appropriate.

Tables

Table 1. Gold-standard sequence divergences

Human WGS data set

	Yoruba	European	East Asian	Biaka Pygmy	
	1.081	1.106	1.098	1.190	Divergence (sites per Kb)
Yoruba	0.005	0.004	0.004	0.024	Standard error of divergence
	641.7	1117.0	814.7	18.5	Number of pairwise aligned bases (Mb)
		0.827	0.892	1.212	
European		0.004	0.004	0.025	
		657.2	848.2	22.6	
			0.772	1.186	
East Asian			0.005	0.027	
			296.8	18.1	

Human RRS data set

	Yoruba	European	East Asian	Australian	Mbuti Pygmy	San
	1.017	1.056	1.050	1.047	1.108	1.113
Yoruba	0.023	0.014	0.019	0.024	0.021	0.020
	4.1	11.1	5.3	3.0	4.5	4.5
		0.798	0.850	0.873	1.082	1.096
European		0.015	0.016	0.021	0.018	0.019
		7.1	7.0	3.8	5.8	5.7
			0.788	0.817	1.111	1.137
East Asian			0.034	0.026	0.025	0.027
			1.3	1.9	2.9	2.9

Chimpanzee WGS data set

	Central	Eastern	Western
	2.072	2.023	2.254
Central	0.032	0.069	0.019
	5.0	1.0	13.7
		2.185	0.827
Western		0.069	0.012
		1.0	13.7
			3.875
Bonobo			0.126
			0.6

Chapter 3

Characterizing the microsatellite mutation rate and process

James X. Sun, Agnar Helgason, Gisli Masson, Sigríður Sunna Ebenesersdóttir, Heng Li, Swapan Mallick, Sante Gnerre, Nick Patterson, Augustine Kong, David Reich & Kari Stefansson

Mutation provides the raw material of evolution. This study reports the largest study of new mutations to date: 2,058 germline mutations discovered by analyzing 85,289 Icelanders at 2,477 microsatellites. We find that the paternal-to-maternal mutation rate ratio is 3.3, that the mutation rate in fathers doubles between the ages of 20 to 58 whereas there is no association to age in mothers, and that strong length constraints apply: longer alleles tend to mutate more often and decrease in length, whereas shorter alleles tend to mutate less and increase in length.

Germline mutations provide the raw material for evolution. However, there is limited empirical data about how the mutation rate (μ) in humans varies with parental gender and age, or with genomic features[1-4]. While whole-genome sequencing of three nuclear families[5-7] revealed dozens of new mutations, there were too few individuals to provide a detailed characterization of the mutation process. Moreover, the studied families may be atypical and it is essential to study many families to obtain a population-wide estimate. One outcome of an understanding of the

mutation process would be to provide a direct estimate of the rate of the molecular clock, which would make it possible to estimate the divergence time of humans and our closest living relatives, chimpanzees, without relying on the fossil record for time calibrations.

To characterize the mutation process in humans at unprecedented resolution, we focused on microsatellites: 1-6 base pair motifs that vary in the number of times they are repeated. Due to DNA polymerase slippage during replication, the mutation rate of microsatellites is on the order of 10^{-4} to 10^{-3} per locus per generation[8-12], far higher than the nucleotide substitution rate of around 10^{-8}. We analyzed data from 2,477 autosomal microsatellite loci that had been genotyped as part of linkage-based disease gene mapping studies and thus had been specifically ascertained to be highly polymorphic[13]. The data set included 85,289 Icelanders after restricting to individuals genotyped for at least half of the microsatellites; this included 24,832 father-mother-child trios, after removing those with evidence of inaccurate parental assignment (Methods, Fig. S1). The median genotype error rate was 1.8×10^{-3} per allele (Fig. S2, Note S1), which is high compared to the mutation rate, and thus we took additional steps to reduce the error.

To distinguish genuine mutations from genotype errors, we used both 'trio' and 'family' approaches (Methods, Notes S2-3). The trio approach (Fig. 1A) identified 1,695 mutations in 5,085,672 transmissions (total number of alleles transmitted over all individuals over all loci), validating new mutations by restricting to transmissions in which every member of the trio was genotyped more than once. The family approach (Fig. 1B) identified 363 mutations in 952,632 transmissions in 2,406 families, validating new mutations by requiring them to be seen in at least one of the proband's children, and validating ancestral alleles by requiring them to be seen in all of the proband's siblings. We also traced haplotypes of linked microsatellite alleles through families to determine parental origin (Methods, Note S3). The trio and family approaches produced statistically indistinguishable inferences about mutation rate and the mutational process (Table 1, Fig. S3, Fig. S16), and hence we combined them to obtain 2,058 mutations in 6,038,304 transmissions (62 mutations were counted twice due to overlap).

To estimate the proportion of the 2,058 candidate mutations that are real, we re-genotyped a random sample of 103 trio and 99 family mutations, and estimated false-positive rates of 2.9%

and 2.6%, respectively (Table S1, Fig. S4). We also estimated the rate of false-positives due to errors in the allele-calling algorithm. By manually re-scoring the electropherograms of 316 individuals from the family dataset, and declaring a false-positive wherever there was a disagreement with the algorithm's results, we estimate a false-positive rate of 4.3%. Combining the two modes of false-positives, we estimated a rate of 7.2% (Table S1). We estimate the false-negative rate (the probability of an undetected real mutation) to be 9.0% by first generating mutations at random and then using our approaches to detect the simulated mutations (Methods).

The estimated mutation rate of tetra-nucleotides is 10.01×10^{-4} per locus per generation, 3.7 times higher than the di-nucleotide rate of 2.73×10^{-4} (Table 1). Estimates are nearly unchanged after correcting for false-positives and false-negatives by (1-0.072)/(1-0.090), and thus we quote unadjusted rates in what follows. Our estimate of the male-to-female mutation rate ratio is α=3.3 (95% CI 2.9-3.7) (Table S2), within the range of 2-7 previously inferred for sequence substitutions by comparative genomic methods[1,14] and direct observation in families[7,15]. Paternal age is highly correlated with mutation rate ($P=9.3\times10^{-5}$), whereas maternal age is not (P=0.47; Fig. 2A, S5), consistent with observations based on disease-causing mutations, and the biology of germ cell production in which male germ cell precursors undergo numerous mitoses as a man ages, whereas female oocytes do not undergo postnatal cell division[1].

These data allow the first high resolution characterization of the microsatellite mutation process, at least for the highly polymorphic di- and tetra-nucleotide microsatellites that are typically used for disease gene mapping studies and population genetics[8]. First, we find that 32% of mutations at di-nucleotide microsatellites are multi-step, compared to only 1% in tetra-nucleotides (Fig. 2B, S3). An implication is that the predicted variance of the allele length distribution in tetra-nucleotides is almost identical to that of di-nucleotides despite their 3.7-fold higher mutation rate[16,17]. Second, longer microsatellite alleles have a greater rate as previously seen on the Y-chromosome[18] and in tri-nucleotide repeat disorders[19]. At di-nucleotides, the mutation rate of 70 bp alleles is four times that of 30 bp alleles (P=0.0013), and at tetra-nucleotides, 120 bp alleles have a four times higher mutation rate than 40 bp alleles (P=0.0018) (Fig. 2C). Third, loci with uniform repeat structures (e.g. CACACACA) have a 40% higher rate ($P=3x10^{-7}$) than those with compound repeat structures (e.g. CACATCACA), supporting the hypothesis that DNA

polymerase slips less for interrupted tandem repeats[8,20] (Fig. S6, S15). Fourth, we provide the first direct demonstration of length constraints at microsatellites[21,22]—shorter alleles tend to become longer and vice versa ($P=2x10^{-15}$) (Fig. 2D, S7-8)—and thus our observations provide empirical documentation of a phenomenon that has previously been inferred only through comparative genomics work[20,23,24]. This length constraint is different from the mutation process in tri-nucleotide repeat disorders[19] such as Huntington's disease, where longer repeats tend to mutate to even longer alleles. Fifth, we observe correlations ($P<10^{-4}$) between mutation rate and motif length, repeat length, allele-size, distance from exons, parental gender, and paternal age, but none to recombination rate, distance from telomeres, human-chimpanzee divergence, and parental heterozygosity (Table S3-4; Note S4).

Methods

Data collection and filters

All genotyping was carried out at deCODE Genetics. We identified 2,477 autosomal microsatellite loci that were most heavily genotyped, and 85,289 individuals in whom at least 50% of the loci were genotyped. These loci were previously selected at deCODE Genetics to be highly polymorphic and thus useful for disease gene mapping studies. All microsatellites were genotyped using multiplexed capillary gel electrophoresis with automated allele calling. Details of the genotyping process can be found in Kong et al 2002[13]. All microsatellites had a minimum repeat length of 5 units. Using the deCODE Genetics genealogical database (Íslendingabók), we assembled 25,067 mother-father-offspring trios from 85,289 individuals.

To filter out trios with potentially inaccurate parental assignments that would result in false inference of mutations, we computed the genome-wide identity-by-state (IBS) fraction for each trio. In a trio, for each locus, the IBS status between a parent-child pair is '0' if none of the diploid alleles are identical, and '1' otherwise. To determine a suitable IBS threshold, we compared against IBS values of known uncle-child and aunt-child pairs, and eliminated any trio below the threshold (Fig. S1). This resulted in the final set of 24,832 trios.

To determine the per-locus genotyping error rate, we estimated the error probability of a single allele for each locus based upon the discordance of multiple genotypes. Let $\hat{p}$ be the probability of a genotype error, let k be the number of times an allele is repeatedly genotyped, let n_k be the total number of individuals who were each genotyped k-times, and let y_k be the number of individuals with inconsistent genotypes. Then, the probability of error is $\hat{p} = \frac{\sum_k y_k}{\sum_k 2kn_k}$ (Note S1). Fig. S2 shows the histogram of the estimated genotype error rate for the analyzed loci.

For the 23 HapMap individuals, microsatellite genotyping was performed at deCODE Genetics, using a methodology identical to that of Icelandic individuals. These include a trio of European (CEU) ancestry (NA12878, NA12891, NA12892) and a trio of African (YRI) ancestry (NA19238, NA19239, NA19240). Both have been sequenced to deep coverage by the 1000 genomes project using Illumina technology, where BWA[25] was used for alignment and SAMtools[26] used for consensus calling. Three additional YRI genomes (NA18506, NA18507 and NA18508) were similarly processed providing genome-wide heterozygous calls for 9 of the HapMap individuals. Additionally, 20 of the individuals were sequenced to deep coverage by Complete Genomics who provide genotype calls at known variant dbSNP positions (dbSNP release 130, which includes calls from the 1000 genomes project). These were downloaded from http://www.completegenomics.com/. Six of the individuals overlap between the two sequencing technologies (Table S6, Fig. S11). To extract sequence heterozygosity data around each microsatellite, we located the physical position of the microsatellite and extracted a window of data around the microsatellite with genetic distance thresholds of 0.001, 0.002, and 0.004 centimorgans. The central 1kb segment in which the microsatellite lies is masked out. Then, for each of the three genetic distance windows, heterozygous calls were recorded. Based on the results of matching empirical observations to our model simulations, we used sequence heterozygosity calculated from genetic windows with a threshold of 0.001 centimorgans (Fig. S17).

Detecting mutations via the trio approach

When identifying mutations in trios, we restricted to loci genotyped many times and searched for any Mendelian inheritance incompatibilities. Once identified, we assigned the unmatched proband allele to be the mutant allele, and attempted to identify the parent in whom the mutation arose. Cases where we needed to invoke simultaneous mutations from both parents were

excluded from analysis. There were some cases in trios where the parental origin was ambiguous (Note S2), and these were excluded from analyses in which parental origin was required, such as in the mutation length distribution (Fig 2B). However, these cases were included for the total mutation count and for mutation rate analyses. For cases with unambiguous parental origin, the ancestral allele was defined as the one that was closer in length to the mutant allele. If both parental alleles were different by the same mutational step size, e.g. the parent is (6,10) and offspring is 8, we randomly chose the ancestral allele. After all mutations were identified, we removed loci that were observed to harbor many more mutations from homozygous parents to homozygous children than would be expected based on Hardy-Weinberg equilibrium, a phenomenon that we found affected the trio, but not the family data. We hypothesized that these loci might harbor false mutations due to polymorphisms under the PCR primer sites, leading to allele-specific PCR mis-amplification[27-29]. This was confirmed by sequencing the primer sites from 15 mutations and identifying 5 with SNPs in the primer site region. Other possible explanations for homozygous-homozygous false mutations, such as deletions, were harder to test. However, based on the primer site experiments, a cautious approach was evidently warranted and we thus removed 49 loci from further analysis (Note S2).

A potential pitfall for analyses of the trio data are somatic mutations: those that occurred post-natally in the lineage of genotyped cells, but not in germline cells transmitted to offspring. However, the family-based approach is immune to somatic mutations and is nevertheless consistent with the results of the trio dataset, which increases confidence in these results. Moreover, the overall effect of somatic mutation on this study is minimized by the fact that all DNA was extracted directly from blood, rather than from immortalized cell lines.

Detecting mutations via the family approach

To discover mutations, we applied the following procedure: (1) We identified Mendelian inheritance errors for a locus in a trio, where the proband has at least one child and one sibling that have been genotyped (Fig. 1B). (2) We used Allegro 2.0[30] to haplotype (i.e. phase) the family, using all available loci from the same chromosome. Since Allegro cannot determine haplotypes in the presence of loci with inheritance errors, we initially masked out such loci (including the putative mutant locus). Based on neighboring loci, Allegro was then used to impute alleles into the masked loci. (3) The optimal assignment of haplotypes to alleles from the

locus with the putative mutation was solved as a constraint satisfaction problem (CSP) (Note S3). (4) In order to call the inheritance error as a confirmed mutation, we required at least one sibling to carry the haplotype with the ancestral allele, none of the siblings to carry the mutant allele, and at least one child to carry the same haplotype, but with the mutant allele. In this way, we obtained independent confirmation of the ancestral and mutant alleles. To obtain the total number of transmissions, we repeated the same process as above for all loci genotyped for each family (including loci with no inheritance errors), i.e. (1) masked out the allele, (2) imputed the haplotype, (3) assigned haplotypes to alleles, and (4) required a randomly chosen haplotype to be present in at least one child and one sibling.

The family-based approach has the potential problem that mutations in progenitor germ cells might cause a mutation to be observed simultaneously in the proband and its siblings, causing us to reject a real mutation. However, this problem does not affect the trio approach. The trio and family approaches produce consistent results despite being differently affected by these potential biases, suggests that the overall impact of these biases may be small, and increases confidence in our results.

Experimental validation

To estimate the rate of false-positives, we began by re-genotyping a subset of mutations. (Our "false discovery rate" is defined as the fraction of identified mutations that are later shown to be false positives, which is different from the usual statistical definition.) For the trio dataset, we randomly re-genotyped 103 mutations using capillary gel electrophoresis, using the same primer sites as the original genotypes. For the family dataset, to maximize our discovery of false-positives, we targeted our re-genotyping efforts toward the mutations that had a higher *a priori* chance of being in error. A candidate mutation was flagged as error prone if (1) both parent and offspring were homozygous, (2) the mutation length was a non-integer multiple of the motif size, or (3) the mutation length was longer than 6 nucleotides. Altogether, these error-prone categories were responsible for 13% of candidate mutations (Table S1).

To obtain an estimate of the proportion of genuine mutations that were missed by our methods, we simulated mutations by randomly distributing them on the genealogy and then tested whether they gave rise to detectable inheritance errors. At a locus, the procedure is as follows: (1) For the parents, randomly sample 2 alleles from the allelic distribution generated

from all Icelandic individuals genotyped at the locus. (2) For the non-mutating allele of the offspring, randomly draw a parent and an allele. For the mutating allele, draw an allele from the other parent, mask out this allele from the allelic distribution, and sample from the remaining alleles of the distribution. (3) For this simulated trio, determine whether there is a detectable inheritance error. The failure to detect an inheritance error corresponds to a false-negative mutation, and thus we can use the fraction of sites in this class as an estimate of the expected false-negative rate. To be consistent, the number of mutations we simulated at a locus was proportional to the number of transmissions (denominator) observed from the mutation detection process. As an example of a real mutation that would be missed by our method, suppose that the father-mother-proband trio has genotypes of allele-lengths (6 10), (8 10), (8 10), respectively. If the mother passed allele 10 to the proband, and the father passed a 6 → 8 mutation, then this mutation would not be detected by either the trio or family based approach.

Software and computation

Data were mined and analyzed using Perl. Simulations and statistical analyses were written using C++, Matlab, and Octave. Computationally intensive analyses were performed using the Orchestra shared research cluster at Harvard Medical School.

Statistical analyses of the microsatellite mutation rate

To estimate the standard error of the mutation rate, taking into account rate variation across loci, we developed a hierarchical Bayesian model. The model assumes that that the mutation rate at each locus is governed by a beta-binomial distribution. The mutation rate is sampled from a beta distribution. Given the sampled rate, mutations are then generated from the binomial (Note S9).

All the microsatellite genotypes used for this study were reported based on amplicon size, which includes all the sequence between the PCR primers and not just the microsatellite repeat units. To obtain the absolute length of repeat units for each microsatellite locus, we used Tandem Repeat Finder[31] from the UCSC genome browser to obtain the start and stop coordinates for the repeat units in the human reference genome (Fig. S15). The start and stop coordinates of each amplicon in the human reference sequence were then used to calculate the size of the flanking sequence. For any locus, the absolute length of an allele was calculated by subtracting the flanking sequence size from its measured amplicon size.

To compute the relative length of an allele in a given locus, we estimated the mean length and standard deviation of all individuals genotyped at the locus, and then reported the relative length of the allele in terms of the number of standard deviations from the mean (Z-score).

The estimates of motif impurity from Fig. S6 are based on the application of the Tandem Repeat Finder software to the human genome reference sequence (Fig. S15), and thus there is no guarantee that the same level of impurity applies to the entire population. Nevertheless, we expect that the human genome reference sequence is correlated in its motif impurity to that of the general population, an expectation that is validated by the fact that the estimated motif impurity is strongly correlated to the observed mutation rate.

To evaluate whether features of the microsatellites were predictors of the mutational process (Table S3), we regressed each feature at a time. After scaling each of the tested variables, we performed a logistic regression to the mutation rate and directionality, and a Poisson regression to the step size. The P-values reported in Table S3 are the P-values of the regression coefficients. To determine whether the tested variables interact, we performed a multivariate logistic regression with interactions, for every pair of variables, i.e. $logit(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$, where y is the mutation rate (Table S4).

References

1. Crow, J.F. The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* **1**, 40-7 (2000).
2. Crow, J.F. Age and sex effects on human mutation rates: an old problem with new complexities. *J Radiat Res (Tokyo)* **47 Suppl B**, B75-82 (2006).
3. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).
4. Arnheim, N. & Calabrese, P. Understanding what determines the frequency and pattern of human germline mutations. *Nat Rev Genet* **10**, 478-88 (2009).
5. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-9 (2010).
6. Durbin, R.M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
7. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature genetics* **43**, 712-714 (2011).
8. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**, 435-45 (2004).
9. Weber, J.L. & Wong, C. Mutation of human short tandem repeats. *Hum Mol Genet* **2**, 1123-8 (1993).
10. Xu, X., Peng, M. & Fang, Z. The direction of microsatellite mutations is dependent upon allele length. *Nat Genet* **24**, 396-9 (2000).
11. Whittaker, J.C. *et al.* Likelihood-based estimation of microsatellite mutation rates. *Genetics* **164**, 781-7 (2003).
12. Huang, Q.Y. *et al.* Mutation patterns at dinucleotide microsatellite loci in humans. *Am J Hum Genet* **70**, 625-34 (2002).
13. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat Genet* **31**, 241-7 (2002).
14. Makova, K.D. & Li, W.H. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624-6 (2002).
15. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961-8 (2010).

16. Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457-62 (1995).
17. Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L. & Feldman, M.W. An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**, 463-71 (1995).
18. Ballantyne, K.N. *et al.* Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *Am J Hum Genet* **87**, 341-53 (2010).
19. Cummings, C.J. & Zoghbi, H.Y. Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum Mol Genet* **9**, 909-16 (2000).
20. Kruglyak, S., Durrett, R.T., Schug, M.D. & Aquadro, C.F. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A* **95**, 10774-8 (1998).
21. Zhivotovsky, L.A., Feldman, M.W. & Grishechkin, S.A. Biased mutations and microsatellite variation. *Mol Biol Evol* **14**, 926-33 (1997).
22. Feldman, M.W., Bergman, A., Pollock, D.D. & Goldstein, D.B. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. *Genetics* **145**, 207-16 (1997).
23. Sainudiin, R., Durrett, R.T., Aquadro, C.F. & Nielsen, R. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics* **168**, 383-95 (2004).
24. Garza, J.C., Slatkin, M. & Freimer, N.B. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol Biol Evol* **12**, 594-603 (1995).
25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
26. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
27. Weber, J.L. & Broman, K.W. Genotyping for human whole-genome scans: past, present, and future. *Adv Genet* **42**, 77-96 (2001).
28. Johansson, A.M. & Sall, T. The effect of pedigree structure on detection of deletions and other null alleles. *Eur J Hum Genet* **16**, 1225-34 (2008).
29. Callen, D.F. *et al.* Incidence and origin of "null" alleles in the (AC)n microsatellite markers. *Am J Hum Genet* **52**, 922-7 (1993).
30. Gudbjartsson, D.F., Thorvaldsson, T., Kong, A., Gunnarsson, G. & Ingolfsdottir, A. Allegro version 2. *Nat Genet* **37**, 1015-6 (2005).

31. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
32. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics* **5**, e1000471 (2009).
33. Amos, W., Flint, J. & Xu, X. Heterozygosity increases microsatellite mutation rate, linking it to demographic history. *BMC Genet* **9**, 72 (2008).

Tables

Table 1. Direct estimates of microsatellite mutation rates

	Mutations	Transmissions	Mutation rate ($x10^{-4}$)* mean	5^{th} – 95^{th} percentile
di-nucleotide loci†				
Trio-approach	1,218	4,578,348	2.66	2.47 – 2.85
Family-approach	269	861,204	3.12	2.65 – 3.59
Combined	1,487	5,439,552	2.73	2.56 – 2.91
tetra-nucleotide loci				
Trio-approach	380	393,072	9.67	8.44 – 10.89
Family-approach	86	72,516	11.86	8.70 – 15.02
Combined	466	465,588	10.01	8.86 – 11.15

* The 90% Bayesian credible interval is calculated based on a Bayesian hierarchical beta-binomial model (Note S9), which allows for the mutation rate to vary across loci.

† The breakdown of the mutation rate by motif type, for di-nucleotides, can be found in Table S8.

Table S1. Experimental validation of mutations

Mutations from family data set	Mutation Counts	Targeted re-genotyping			Electropherogram review			Intersection of sites		
		TP	FP	FP/(TP+FP)	TP	FP	FP/(TP+FP)	TP	FP	FP/(TP+FP)
Class 1 mutations	326	74	2	0.026	262	8	0.030	57	2	0.034
Class 2 mutations										
Homozygous parent and offspring	21	10	2	0.167	20	0	0.000	9	2	0.182
Non-integer multiple of motif length	10	0	2	1.000	6	3	0.333	0	2	1.000
Excessively long (>6bp)	18	7	2	0.222	13	3	0.188	6	3	0.333
More than 1 of the above	1	0	0	N/A	1	0	0.000	0	0	N/A
Total	376			0.058			0.043			0.072

Experimental validation of mutations from the family data are shown here. See Figure S4 for validation of the trio data.

TP = True Positives, i.e. candidate mutations that are verified to be true.
FP = False Positives, i.e. candidate mutations that are rejected by the verification.

Class 1 mutations are the ones that do not belong to Class 2, which are likely to have a higher false identification rate. Class 2 mutations include: (1) both parent and offspring were homozygous, (2) the mutation length was a non-integer multiple of the motif size, or (3) the mutation length was longer than 6 nucleotides.

In our re-genotyping efforts, to maximize our discovery of false-positives, we targeted our re-genotyping efforts toward Class 2. No such sampling bias was used in the electropherogram review. In combining the results of re-genotyping and electropherogram review, we examined only overlap data, calling a candidate mutation as a false-positive if either method rejects the mutation.

In obtaining the total false identification rate, due to sampling bias towards the Class 2 mutations, we calculated an overall rate that weights the number of Class 1 and Class 2 candidate mutations, i.e. to obtain the final value of 0.072, we have:

$$\frac{50}{376} \cdot \frac{7}{22} + \frac{326}{376} \cdot \frac{2}{59} = 0.072$$

Table S2. Differences in α

Mutation class	Trio data				Family data			
	Paternal	Maternal	α	[95% CI]	Paternal	Maternal	α	[95% CI]
homozygous to homozygous	123	81	1.52	[1.15 2.04]	13	8	1.63	[0.62 4.25]
homozygous to heterozygous	146	43	3.40	[2.50 4.91]	57	21	2.71	[1.69 4.57]
heterozygous to homozygous	104	42	2.48	[1.75 3.56]	25	14	1.79	[0.95 3.88]
heterozygous to heterozygous	471	82	5.74	[4.59 7.38]	184	41	4.49	[3.25 6.50]
Total	844	248	3.40	[2.97 3.94]	279	84	3.32	[2.63 4.26]

α is the ratio of the paternal mutation rate to the maternal mutation rate. Since we are only examining full trios and families (i.e. probands that have both parents genotyped), the paternal and maternal transmissions are the same, hence α is just the ratio of the mutations.

We split our mutations by trio/family data and by mutation class. A "homozygous to homozygous" mutation is when a parent with homozygous alleles transmits a mutation to a child with homozygous alleles, e.g. parent = (6,6) and child = (8,8).

To construct the 95% confidence interval for α, we assume that the partition of paternal and maternal events is generated via a binomial distribution. For example, in the total mutations for trio data, assume that the paternal counts are generated with $Binomial(n, p)$, where $n = 844 + 248 = 1092$ and $p = \frac{844}{1092} = 0.773$. α is simulated enough times to suppress Monte Carlo noise, and then the 95% CI is obtained. Note that although we have 1,695 mutations from the trio data, only 1,092 are used here, because the parent transmitting the mutation is ambiguous for the rest (Note S2).

Comparing the trio data to the family data, α is not significantly different, as the 95% CI significantly overlap for each mutation class.

Table S3. Predictors of the mutation process

	p-values for assessing significance in the tested variable		
Tested variable†	**mutation rate**	**magnitude in step size***	**directionality***
motif length (di- vs. tetra-)	$<10^{-12}$	1.78×10^{-9}	0.58
absolute length‡	$<10^{-12}$	0.19	0.16
variance in allele length distribution in Icelanders	$<10^{-12}$	0.70	0.11
repeat impurity	3.1×10^{-7}	0.12	0.26
distance from exons (measured by B-statistic††)	2.2×10^{-6}	0.71	0.74
DNA replication timing	0.005	0.07	0.69
recombination rate	0.02	0.49	0.59
sequence divergence, human-chimp (10Kb window)	0.24	0.61	0.67
recombination hotspot	0.42	0.83	0.79
physical distance from telomeres	0.86	0.24	0.40
Heterozygosity	$<10^{-12}$	0.28	0.46
parental gender	$<10^{-12}$	0.04	0.01
paternal age	9.3×10^{-5}	0.67	0.18
maternal age	0.47	0.33	0.66
relative length***	N/T**	1.41×10^{-7}	$<10^{-12}$

† Because our data are mostly di-nucleotides, and di and tetra-nucleotides show major differences in their characteristics, all tested variables excluding motif length, are tested only using di-nucleotides.

†† The B-statistic predicts the intensity of background selection, according to McVicker et al.[32]

‡ When regressing to mutation rate, absolute length is the mean absolute length of each locus. When regressing to step-size variance and directionality, absolute length is defined as that of the parental allele.

* For each mutation, if the mutational length is X, then the magnitude in step size is defined as the absolute value of X, and the directionality is defined as the sign of X.

** Not testable.

*** Relative length is the Z-score of the allele length, relative to the allelic distribution at the microsatellite locus. See the Methods of the main manuscript for a formal definition.

Table S4. Interactions between covariates

Covariate x_1	Covariate x_2	r^2	P-value x_1	P-value x_2	P-value $x_1 \cdot x_2$
Genotype error rate	absolute length	0.004	2.37E-01	2.01E-04	9.20E-03
human-chimp divergence	absolute length	0.000	8.51E-01	6.04E-01	9.75E-01
human-chimp divergence	Genotype error rate	0.002	2.12E-01	5.50E-02	1.67E-01
recombination rate	absolute length	0.001	3.30E-01	1.48E-13	5.62E-01
recombination rate	Genotype error rate	0.002	2.97E-01	5.02E-10	4.88E-01
recombination rate	human-chimp divergence	0.053	2.17E-03	2.56E-01	1.03E-03
DNA replication time	absolute length	0.000	1.56E-03	7.05E-14	7.47E-03
DNA replication time	Genotype error rate	0.004	1.98E-01	3.34E-12	1.10E-01
DNA replication time	human-chimp divergence	0.006	5.48E-02	4.31E-01	3.73E-02
DNA replication time	recombination rate	0.005	4.11E-01	3.77E-03	7.69E-03
ASD	absolute length	0.045	1.07E-04	1.41E-04	1.75E-01
ASD	Genotype error rate	0.019	5.80E-01	7.83E-06	4.45E-02
ASD	human-chimp divergence	0.000	4.80E-01	9.04E-01	8.55E-01
ASD	recombination rate	0.000	3.35E-33	5.94E-04	3.15E-03
ASD	DNA replication time	0.001	5.90E-33	3.91E-01	9.21E-01
B-stat	absolute length	0.000	1.60E-01	2.14E-05	6.46E-01
B-stat	Genotype error rate	0.000	4.71E-02	8.96E-03	1.49E-02
B-stat	human-chimp divergence	0.188	1.03E-01	4.20E-01	4.98E-02
B-stat	recombination rate	0.155	1.33E-01	5.69E-02	7.69E-02
B-stat	DNA replication time	0.103	1.65E-03	2.14E-03	3.83E-03
B-stat	ASD	0.000	8.35E-01	2.98E-15	2.64E-01
recombination hotspot	absolute length	0.002	1.08E-02	3.36E-14	9.32E-03
recombination hotspot	Genotype error rate	0.000	8.70E-01	1.31E-13	9.93E-01
recombination hotspot	human-chimp divergence	0.005	2.20E-01	3.14E-01	1.87E-01
recombination hotspot	recombination rate	0.220	2.94E-01	1.84E-02	2.25E-01
recombination hotspot	DNA replication time	0.002	1.66E-01	1.45E-02	6.33E-01
recombination hotspot	ASD	0.001	1.16E-01	8.76E-31	1.75E-01
recombination hotspot	B-stat	0.015	1.65E-01	2.98E-04	2.17E-01
physical position	absolute length	0.000	1.28E-01	9.51E-11	1.24E-01
physical position	Genotype error rate	0.000	7.45E-01	1.38E-06	8.24E-01
physical position	human-chimp divergence	0.007	4.98E-01	2.95E-01	4.69E-01
physical position	recombination rate	0.001	8.39E-01	1.88E-02	2.88E-01
physical position	DNA replication time	0.005	6.33E-01	9.53E-02	7.88E-01
physical position	ASD	0.001	4.46E-03	3.38E-07	3.49E-03

physical position	B-stat	0.004	3.40E-01	7.38E-03	5.01E-01
physical position	recombination hotspot	0.002	3.15E-01	8.80E-02	2.23E-01
repeat impurity	absolute length	0.180	5.82E-01	5.68E-31	9.37E-03
repeat impurity	Genotype error rate	0.001	8.29E-04	1.53E-06	4.12E-04
repeat impurity	human-chimp divergence	0.000	4.14E-01	2.37E-01	3.40E-01
repeat impurity	recombination rate	0.000	1.12E-01	1.43E-02	6.32E-01
repeat impurity	DNA replication time	0.002	1.20E-02	9.11E-03	1.31E-01
repeat impurity	ASD	0.014	9.70E-01	1.27E-28	6.92E-01
repeat impurity	B-stat	0.003	3.60E-06	1.19E-06	7.12E-06
repeat impurity	recombination hotspot	0.001	5.09E-02	3.25E-01	2.63E-01
repeat impurity	physical position	0.000	3.31E-01	7.44E-01	7.45E-01
Heterozygosity	absolute length	0.099	3.89E-03	2.11E-01	8.00E-01
Heterozygosity	Genotype error rate	0.014	6.49E-01	6.87E-06	1.68E-02
Heterozygosity	human-chimp divergence	0.001	3.75E-01	7.18E-01	7.71E-01
Heterozygosity	recombination rate	0.000	8.50E-48	5.55E-02	3.00E-01
Heterozygosity	DNA replication time	0.005	1.48E-53	2.47E-02	6.83E-02
Heterozygosity	ASD	0.416	2.31E-02	3.95E-04	9.65E-13
Heterozygosity	B-stat	0.002	3.13E-19	2.20E-01	7.60E-01
Heterozygosity	recombination hotspot	0.000	1.44E-52	1.13E-01	2.89E-01
Heterozygosity	physical position	0.000	3.14E-16	2.95E-02	3.22E-02
Heterozygosity	repeat impurity	0.019	1.79E-48	5.31E-01	4.03E-01

Figures

Figure 1. Examples of verified mutations from a trio and a family.

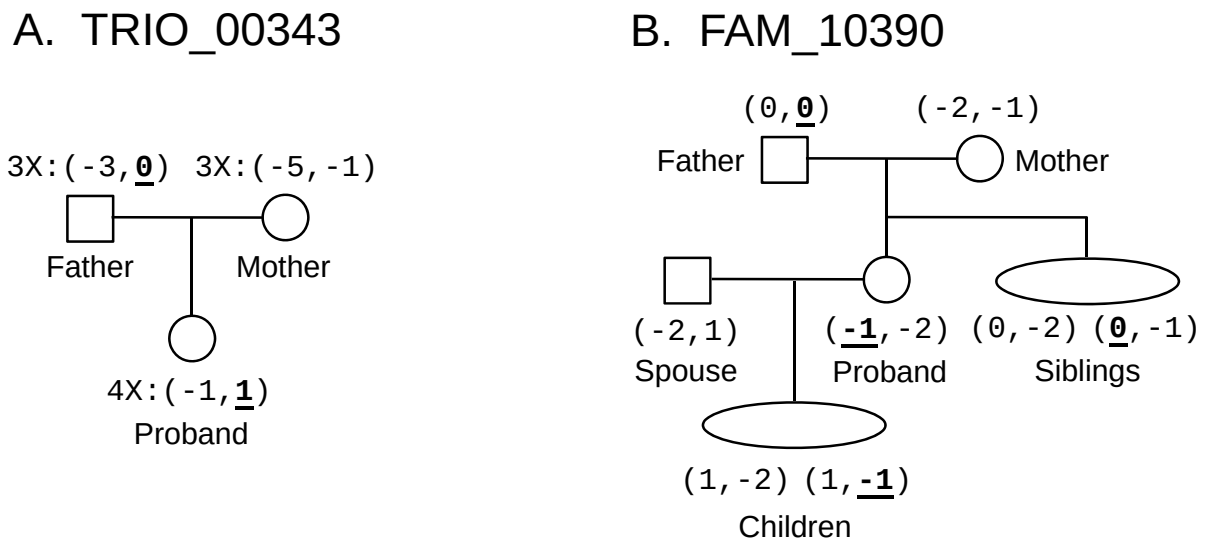


Figure 1. Examples of verified mutations from a trio and a family. The proband is the individual inheriting a mutation, and all individuals are named relative to the proband. All alleles are given in repeat units and shifted so that the ancestral allele has length 0. The mutating allele is underlined. (A) We show a mutation detected using the trio approach. Confirmation of the mutation is from multiple genotyping of the trio: the father, mother, and proband are genotyped 3×, 3×, and 4×, respectively. (B) We show a mutation detected using the family approach. One sibling verified the ancestral allele, and one child verified the mutant allele. The phasing of alleles from the mutant locus and other loci from the same chromosome shows that the sibling with alleles (0,-2) did not inherit the ancestral '0' but rather the other '0' allele from the father.

Figure 2. Characteristics of the microsatellite mutation process

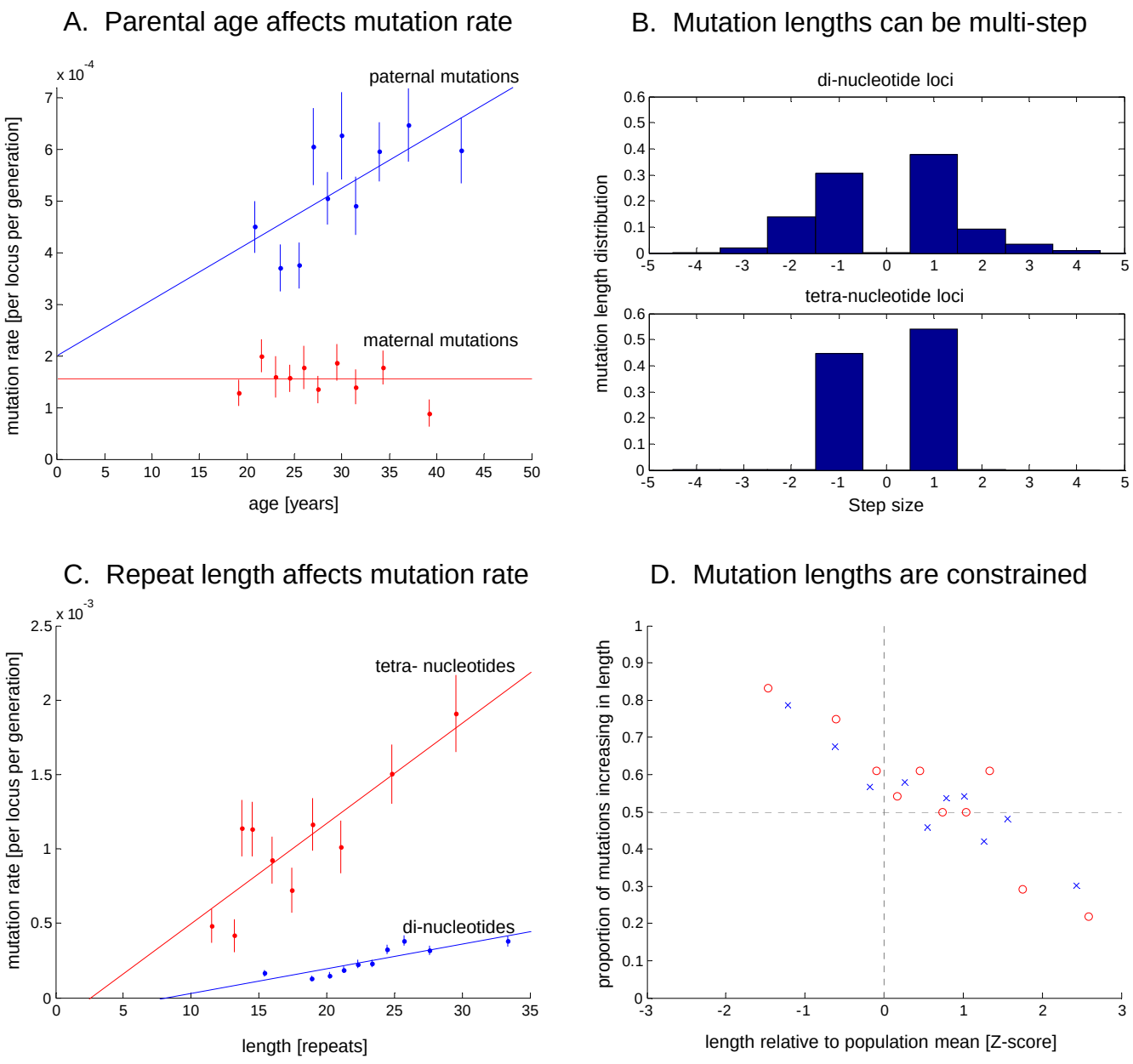


Figure 2. Characteristics of the microsatellite mutation process. (A) Paternal (blue) and maternal (red) mutation rates. The x-axis shows the parental age at child-birth. The data points are grouped into 10 bins (vertical bars show standard errors). The paternal rate shows a positive correlation with age (logistic regression of raw data: $P=9.3\times10^{-5}$; slope = 1.1×10^{-5}/yr), with an estimated doubling of the rate from age 20 to 58. The maternal rate shows no evidence of increasing with age (P=0.47). (B) Mutation length distributions differ between di- and tetra-nucleotides (upper and lower histograms), with x-axis in units of step-size. While the di-nucleotide loci experience multi-step mutations in 32% of instances, tetra-nucleotides mutate almost exclusively by a single-step of 4 bases. (C) Mutation rate increases with allele length: di-nucleotides (blue) have a slope of 1.65×10^{-5} per repeat unit ($P=1.3x10^{-3}$) and tetra-nucleotides (red) have a slope of 6.73×10^{-5} per repeat unit ($P=1.8\times10^{-3}$). (D) Constraints on allele lengths: When the parental allele is relatively short, mutations tend to increase in length, and when the parental allele is relatively long, the mutations tend to decrease in length. Di- and tetra-nucleotides are shown in blue crosses and red circles, respectively. Probit regression of the combined di- and tetra- data shows highly significant evidence of an effect ($P=2.8\times10^{-18}$).

Figure S1. Removal of trios due to potential false-parenthood

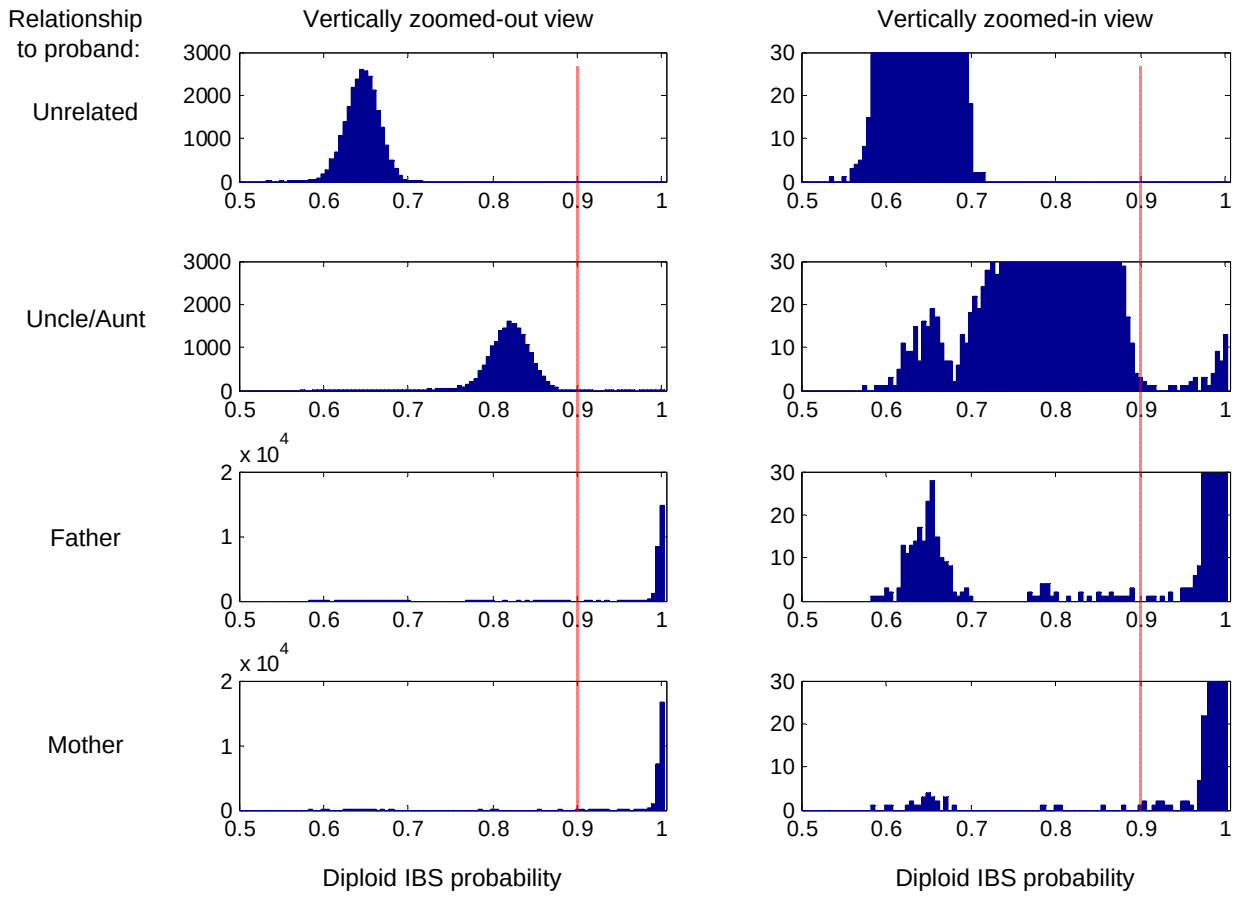


Figure S1. Removal of trios due to potential false-parenthood. Trios were removed based on identity-by-state (IBS) probabilities between a parent and the proband, using all available microsatellite loci. In the figure, the first row is the empirically sampled IBS between pairs of unrelated individuals. The second row shows IBS between the proband and his/her uncle or aunt, allowing us to set a threshold that removes such trios as well. The 3rd and 4th rows are the IBS from the trios, assembled using the Icelandic genealogy. Based on the "null hypothesis" from the first two rows, the threshold for removal of trios was set at 0.9 (red line). A trio is removed if either the Father or the Mother falls below the threshold. Out of 25,067 trios, 235 were removed with this filter.

Definition of diploid IBS: Given individuals *A* and *B*, assume that *n* loci have been genotyped in both. At locus *i*, let the diploid genotype of *A* be A_i, and that of *B* be B_i. We call $A_i = B_i$ if any of the alleles match. For example, if A_i = (4,6) and B_i = (4,8), they are considered equal. Let $\mathbb{I}(A_i = B_i)$ be the indicator variable that is 1 if they are equal and 0 otherwise. Then, the IBS probability is defined as $pIBS(A, B) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i = B_i)$.

Figure S2. Estimated genotype error rate per locus

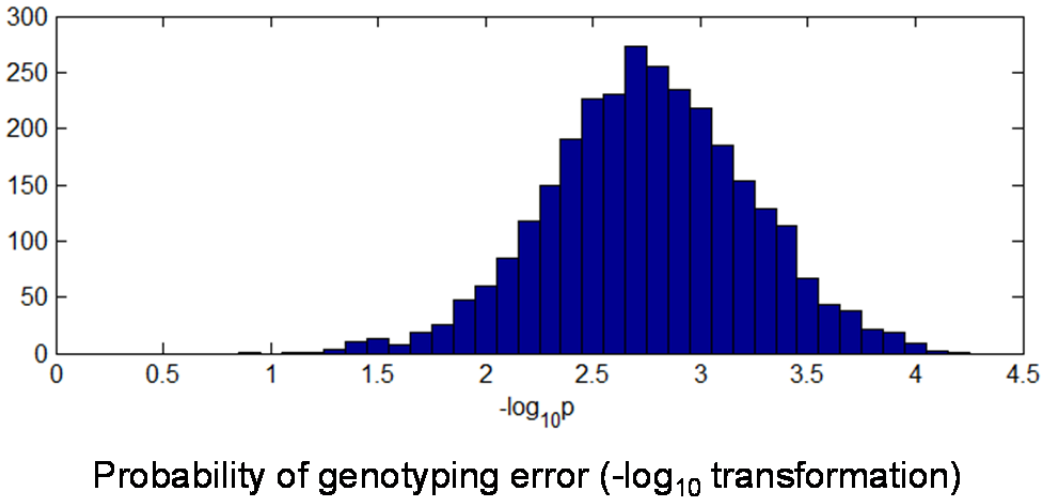


Figure S2. Estimated genotype error rate per locus. Distribution of genotype errors across loci is shown. The genotype error rate is defined as the probability that a single allele will be erroneous after genotyping. The horizontal axis shows the $-\log_{10}$ of the error rate. The median genotype error rate is 1.8×10^{-3}, with a 95% CI of 1.7×10^{-4} to 1.4×10^{-2}.

Definition of genotype error rate at a given locus: Let $\hat{p}$ be the estimated probability of a genotype error when a single allele is observed, let k be the number of times an allele is repeatedly genotyped, let n_k be the total number of individuals who were each genotyped k-times, and let y_k be the number of individuals with inconsistent genotypes. For example, if an individual is genotyped 10 times, 9 times yielding the genotype (4,6) and once yielding (5,6), this would be regarded as an inconsistent genotype. Then, the estimated probability of error is

$$\hat{p} = \frac{\sum_k y_k}{\sum_k 2kn_k}$$

Note S1 describes the derivation of this expression and its assumptions.

Figure S3. Similarity between trio and family data in mutational length distribution

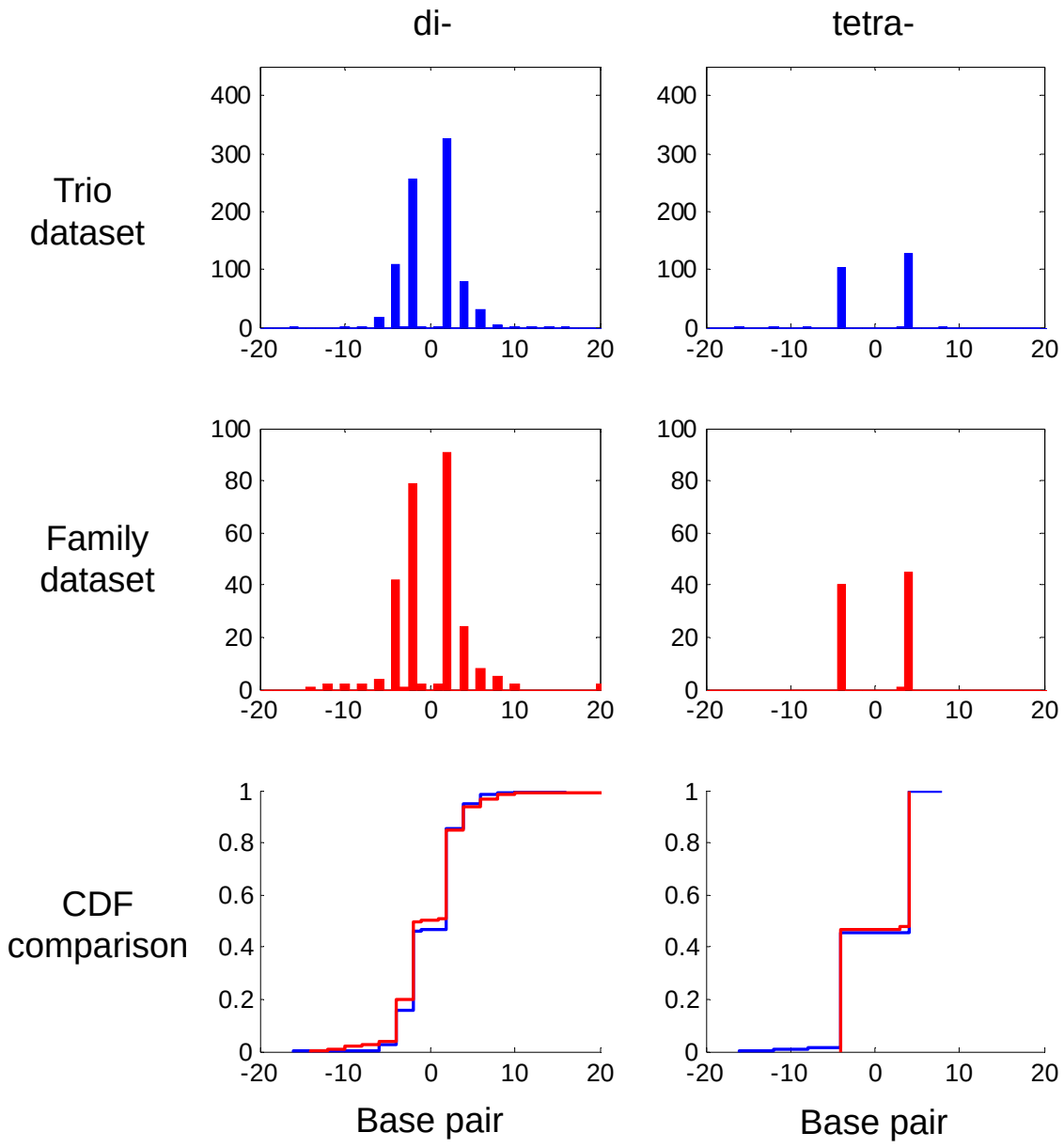


Figure S3. Similarity between trio and family data in mutational length distribution. This figure separates the trio and family datasets from main text Fig. 2B. Additionally, the bottom row compares the CDF between the datasets. The two-sample Kolmogorov-Smirnov test gives P-values of 0.807 and 1 for the di- and tetra- comparisons, respectively. Thus, in the mutational length distribution, there are no significant differences between the two datasets.

Figure S4. False-positive mutations from the trio approach

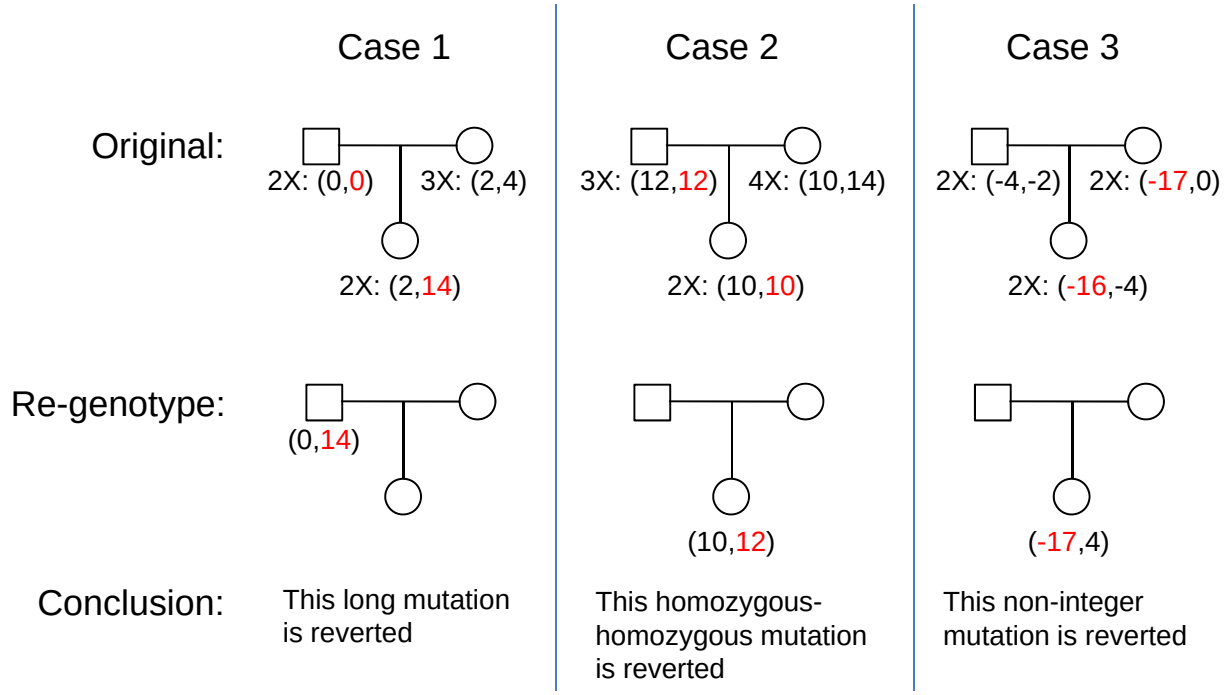


Figure S4. False-positive mutations from re-genotyping in the trio approach. From the set of trio mutations identified, we randomly chose 103 mutations and re-genotyped them. 3 false-positives were identified, which are shown here. All genotypes are in units of base pairs. The 1st case is an apparent mutation that is unusually long, with a mutational length of 14 bp. The 2nd case involves a homozygous parent transmitting to a homozygous child, which we believe is a more error-prone class as discussed in the text. The 3rd case is an apparent mutation of a single base pair, which is a non-integer multiple of the motif length (2 base pairs in this case).

See Note S2 and Table S1 for a more elaborate analysis of false-positive rates when a mutation is either (1) excessively long, (2) a transmission from a homozygous parent to a homozygous child, or (3) a non-integer multiple of the motif length.

Note that allele lengths illustrated above are relative lengths, which is an offset (in units of base pairs) based upon the absolute length of a reference individual's allele.

Figure S5. Predictors of mutation rate and direction (logistic regression)

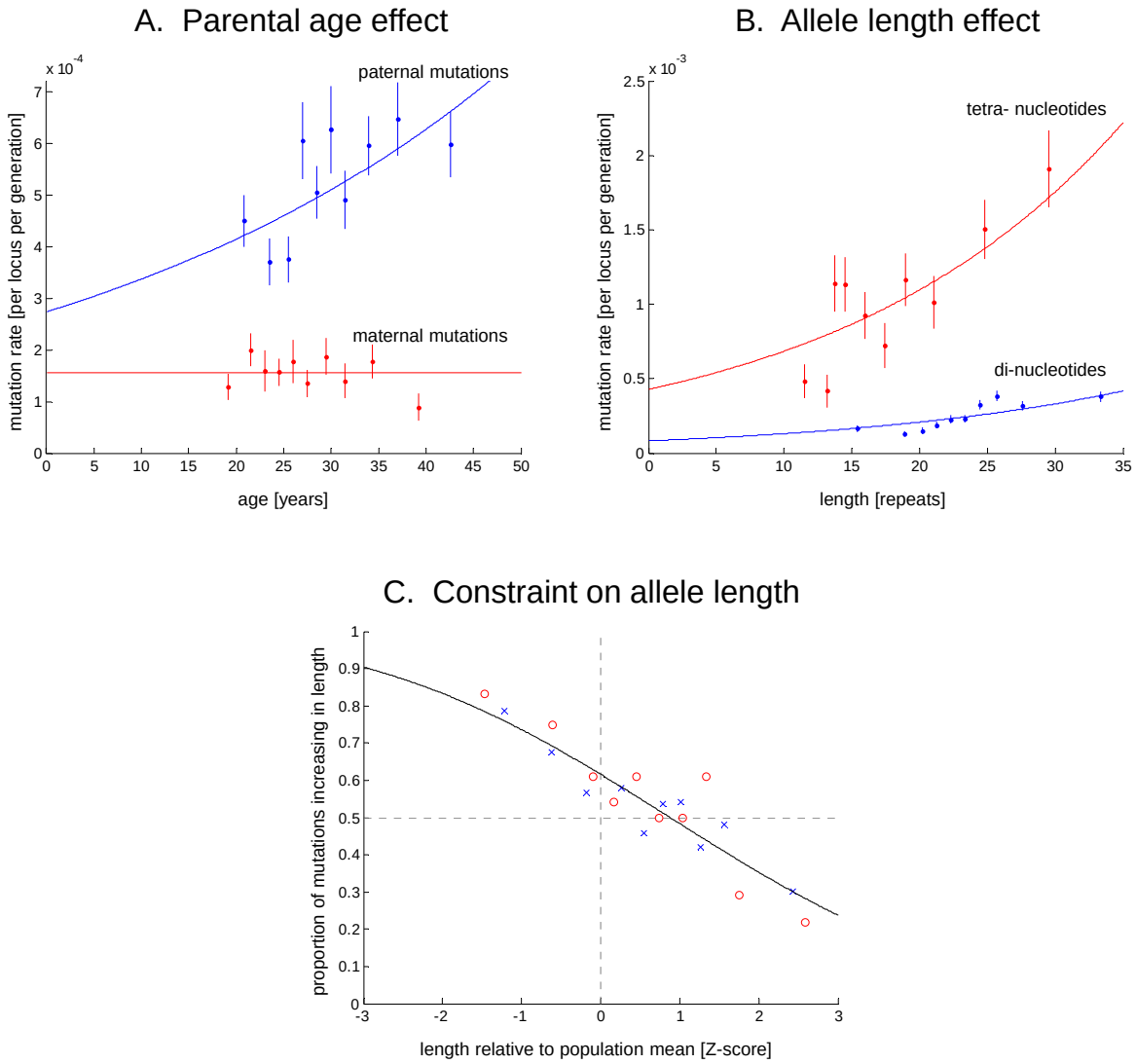


Figure S5. Predictors of mutation rate and direction (logistic regression). Same as main text Fig 2, but with logistic regression curve fits. Note that while the data points shown here are from binning the data, as described in Fig 2, the logistic regressions are performed over the raw data, in which a binomial model of generating mutations (response variable) is assumed. Logistic regression over the raw data has more statistical power than linear regression over the binned data and is constrained to have non-negative mutation rates. The P-values in the main text are reported based on the logistic regression analysis.

Figure S6. Imperfect repeats have a lower mutation rate

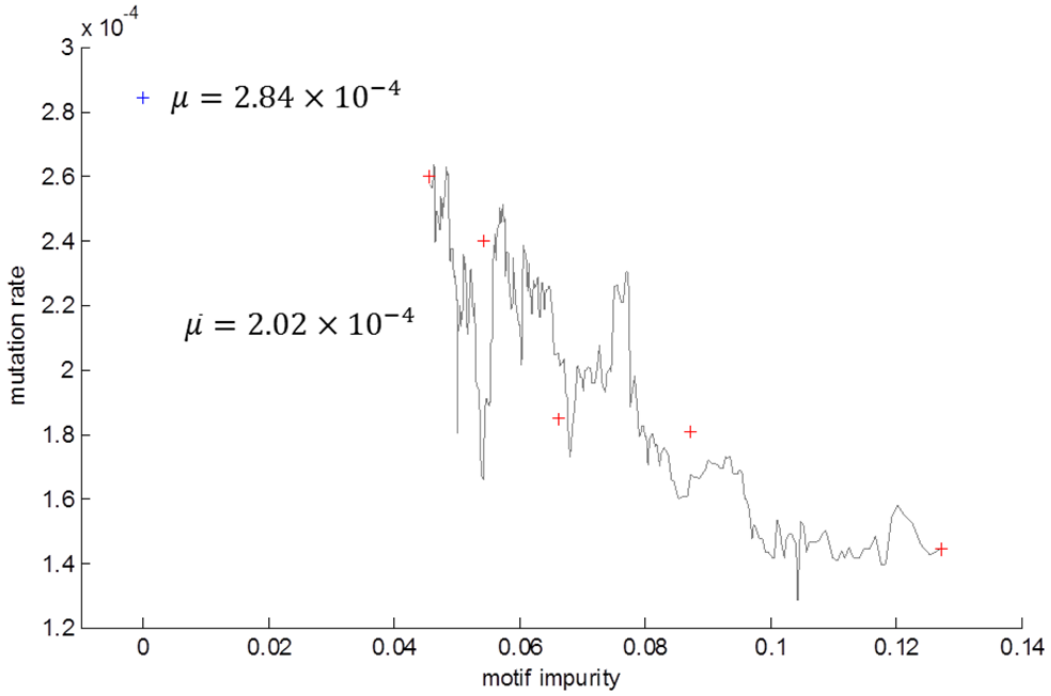


Figure S6. Imperfect repeats have a lower mutation rate. The purity of a motif is computed using the human reference sequence *hg19* from the UCSC genome browser, and downloading data for "simple repeats", in which the "perMatch" column gives the percentage match of the human-genome reference microsatellite to the pure repeat. We define "motif impurity" as one minus this statistic. In blue is the aggregate of 1,036 di-nucleotide loci in which the repeats are perfect (e.g. CACACACACA), without any interrupting bases in the pattern. In red are the imperfect repeats (e.g. CACACA**T**CACA), binned according to the level of repeat impurity. In gray is the window-averaged mutation rate of the imperfect repeats. There are a total of 396 di-nucleotide loci with imperfect repeats. Logistic regression shows that the level of repeat impurity regresses significantly ($P = 3.1 \times 10^{-7}$) with mutation rate. The evidence here is compatible with the hypothesis that when a tandem repeat is interrupted, DNA polymerase slippage is less likely to occur, and hence the mutation rate becomes lower.

Figure S7. Length constraints in microsatellites (raw)

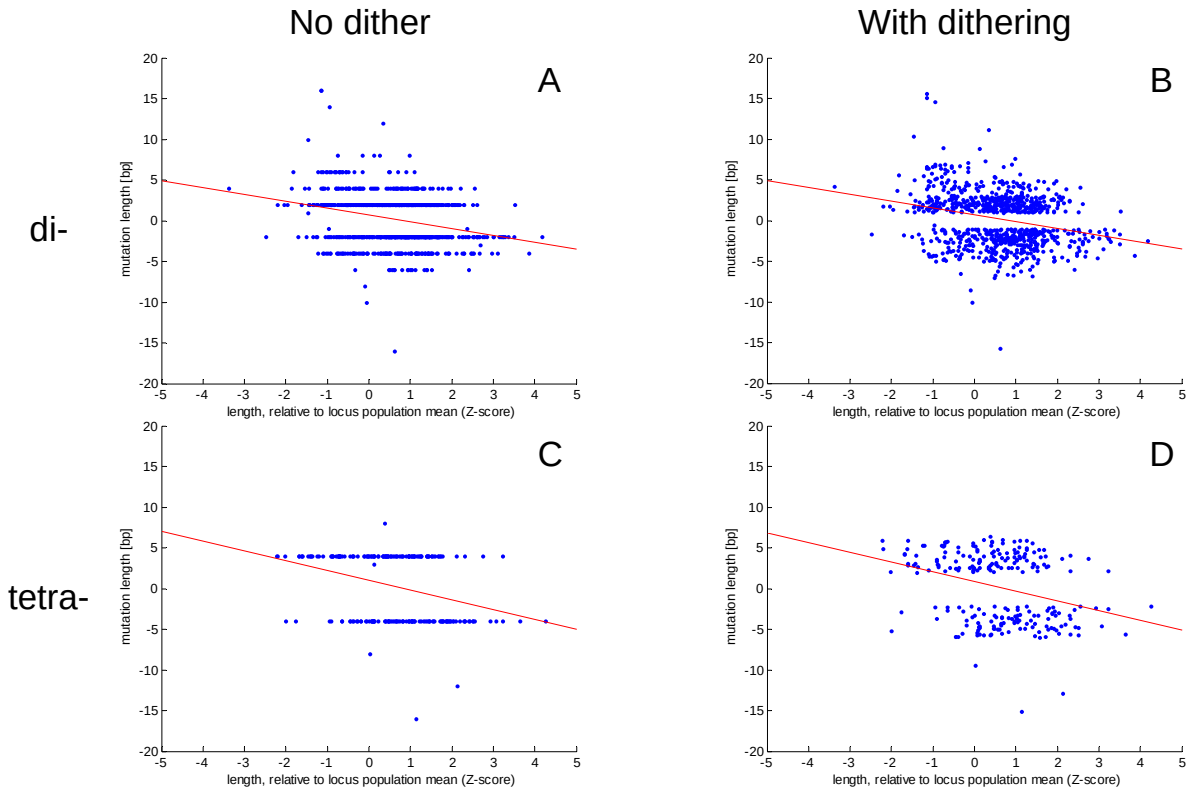


Figure S7. Length constraints in microsatellites (raw). Relative length (x-axis) is in units of Z-scores, the number of standard deviations from the mean length at a given locus. The left panels plot relative length against the mutation length, in base pairs. The right panels provide dithering using a uniform distribution from -0.5 to 0.5 bp to reduce quantization on each mutation length. There is a significant negative correlation.

For di-nucleotides, panel A has: r^2=0.0739, slope=-0.838, P=1.48x10^{-15}.
For tetra-nucleotides, panel C has: r^2=0.106, slope= -1.202, P=3.33x10^{-7}.

Figure S8. Length constraints in microsatellites (binned)

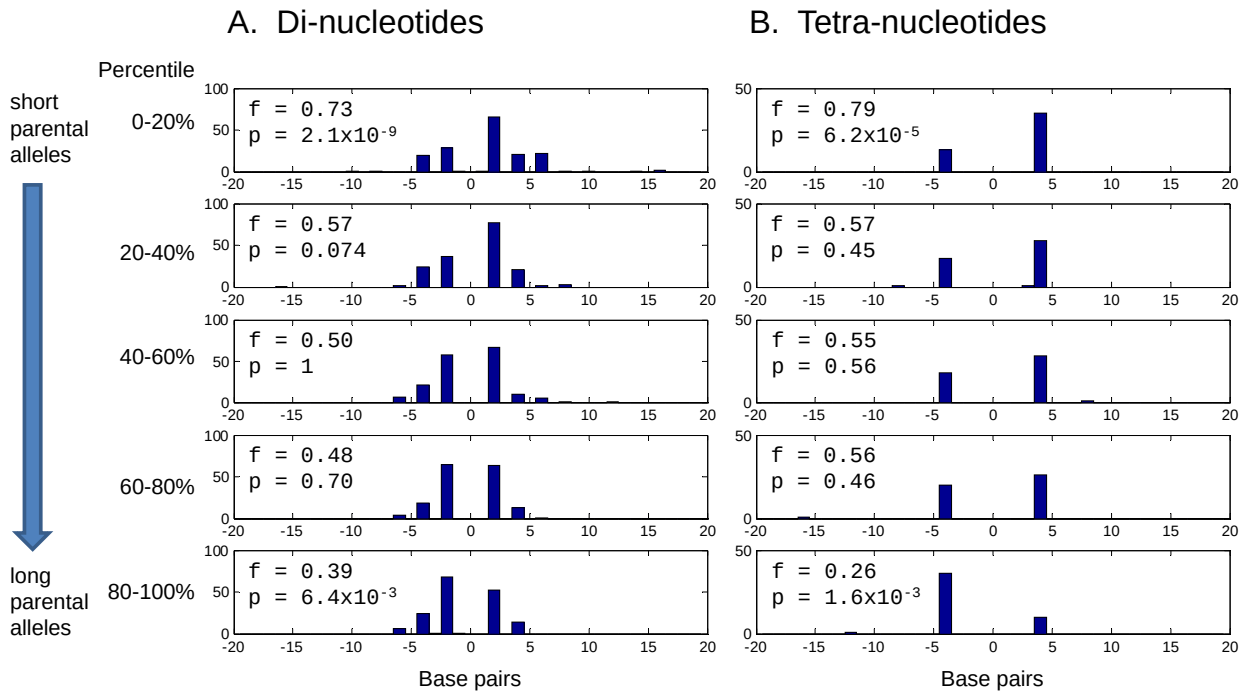


Figure S8. Length constraints in microsatellites (binned version). This figure shows the mutation length distributions as a function of the length of the parental allele, relative to the mean length of a locus. When the parental allele is short (percentiles are displayed on the left), mutation length is biased towards the positive direction. When the parental allele is long, the mutation length is biased towards the negative direction. The fraction (f) of length expansions and the P-value (p) using a two-sided binomial test (the null hypothesis is that microsatellites have no directional bias), are shown in each histogram.

Note S1. Estimating the genotype error rate

Based on the inconsistency rate of multiple-genotyped alleles, we estimated the per-allele genotype error rate for each locus. Formally, at a particular microsatellite locus, a single allele is observed after genotyping. There is a non-zero probability that the genotyping yielded an erroneous allele length. What is this probability of error?

Let $\hat{p} =$ Our goal. $0 \leq \hat{p} \leq 1$.

$k =$ Number of times an allele is repeatedly genotyped.

$n_k =$ Total number of individuals who were each genotyped k-times.

$y_k =$ Number of individuals that resulted in inconsistent genotypes.

For a given individual at a given locus, suppose the true bi-allelic genotype is **a**, and after genotyping, $\mathbf{b}_i$ is observed.

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \longrightarrow \mathbf{b}_i = \begin{pmatrix} a_0 + \epsilon_{i0} \\ a_1 + \epsilon_{i1} \end{pmatrix}$$

To further simplify, suppose that after repeatedly genotyping k times (k is a known quantity), with ε_{ij} IID (independent and identically distributed) with probability p of being nonzero, we only observe the indicator random variable X:

$$X = 1 - \mathbb{I}(\boldsymbol{b_1} = \boldsymbol{b_2} = \cdots = \boldsymbol{b_k})$$

Assuming that the probability of making k identical errors is negligibly small, then

$$X \sim \text{Bernoulli}[\theta] = \text{Bernoulli}[1 - (1 - p)^{2k}]$$

Suppose for n individuals genotyped k times at this particular locus, p is unknown but constant. Our goal is to find the optimal estimate for parameter p.

Thus, our data is modeled as IID $X_1 \ldots X_n \sim \text{Bernoulli}[\theta]$.

By using the maximum likelihood estimate (MLE) for the Bernoulli family, and applying the invariance property of MLE, the MLE for *p* is

$$\begin{aligned}\widehat{p} &= 1 - (1 - \bar{X})^{\frac{1}{2k}} \\ &\approx \frac{\bar{X}}{2k}\end{aligned}$$

The approximation is a 1st-order Taylor expansion around $\bar{X} = 0$, and hence is good only for sufficiently small genotype error probabilities, which we expect in this case. With this approximation, $\theta \approx 2kp$. We use this approximation for all subsequent analyses.

Above we gave the derivation of a single *k*. For multiple *k*, what is the best estimate of *p*, assuming *p* is constant for all *k*? To derive the correct MLE, let $Y_k = n_k \bar{X}_k$, where the subscript *k* emphasizes the dependence on *k*. It can be shown that Y_k is a sufficient statistic for *p*, and

$$Y_k \sim \text{Binomial}[n_k, \theta_k] \approx \text{Binomial}[n_k, 2kp]$$

Importantly, Y_k are independent for different *k*, but clearly not identically distributed.

$$\begin{aligned}l(p|\mathbf{Y}) &= \ln \prod_k \binom{n_k}{y_k} (2kp)^{y_k} (1 - 2kp)^{n_k - y_k} \\ &= \sum_k \ln \binom{n_k}{y_k} + y_k \ln(2kp) + (n_k - y_k) \ln(1 - 2kp)\end{aligned}$$

Differentiating and setting equal to 0 yields:

$$\frac{1}{\widehat{p}} \sum_k y_k = \sum_k \frac{2k(n_k - y_k)}{1 - 2k\widehat{p}}$$

Unfortunately, p cannot be expressed explicitly. A numerical algorithm such as Newton's Method is needed to find p. However, if we use the Poisson approximation to the binomial, i.e. n_k is large and $2kp$ is small, then an analytical solution can be found:

$$Y_k \sim \text{Poisson}[\lambda_k] \approx \text{Poisson}[n_k 2kp]$$

$$\begin{aligned} l(p|\mathbf{Y}) &= \ln \prod_k e^{-n_k 2kp} (n_k 2kp)^{y_k} / y_k! \\ &= \sum_k y_k \ln(n_k 2kp) - n_k 2kp - \ln y_k! \end{aligned}$$

Differentiating and setting equal to 0 yields:

$$\widehat{p} = \frac{\sum_k y_k}{\sum_k 2kn_k}$$

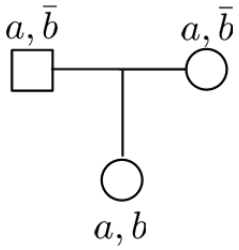
We use this formula to estimate the per allele genotype error rate at each microsatellite locus. Figure S2 shows the distribution of error rates across the 2,477 loci. The median rate is 1.8×10^{-3}, with a 95% CI of 1.7×10^{-4} to 1.4×10^{-2}. Since this number is comparable to the expected microsatellite mutation rate, a simple search for mutations using trios genotyped at 1× coverage will lead to many erroneous mutations. Thus, we developed the "trio approach" and "family approach" to obtain mutations that are most likely to be genuine.

Note S2. Details of the trio approach in mutation detection

Mutations with ambiguous parental origin

In the trio approach, since we do not phase the alleles using neighboring microsatellites, there are cases in which the parental origin is ambiguous. Below we describe the how this scenario occurs.

Let *a* and *b* be distinct alleles. Let $\bar{b}$ be any allele that is not *b*. If there are multiple instances of $\bar{b}$, they are not required to be equal. Then, the following mutant case has ambiguous parental origin:

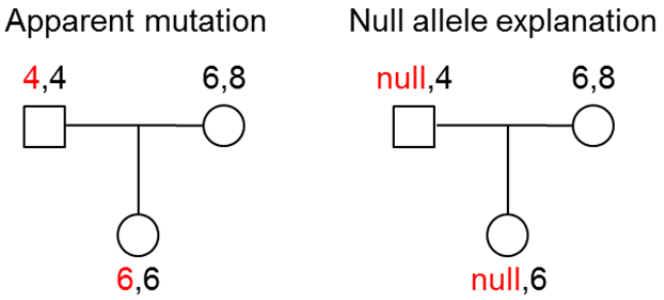


In this pattern, allele *a* is the allele that is also present in the parents, and allele *b* is the mutant. However, since we cannot identify the parental origin of *a*, that of *b* is also ambiguous. Note that we do not attempt to assign *b* to the parent who has a smaller delta in the mutational length, if such a parent exists.

Excessive mutations from homozygous-parent to homozygous-child

After identifying mutations, we discovered that certain loci exhibited many more *de novo* mutations from homozygous parents to homozygous children than would be expected based on Hardy-Weinberg equilibrium. We suspected that these loci might be generating false mutations due to polymorphisms under PCR primer sites, leading to allele-specific PCR mis-amplification.

An example is shown below (left panel), in which there is an apparent mutation from father's allele 4 to child's allele 6. Alternatively, this can be explained by a null allele (right panel). This could be due to (1) a polymorphism in the PCR primer site, resulting in mis-amplification, or (2) a deleted allele, both of which would mean that there is no real mutation.



We removed loci that have an excess rate of homozygous-to-homozygous mutations, compared with the expectation from Hardy-Weinberg equilibrium. To do this, for each locus we compare the observed homozygosity of all alleles to the observed homozygosity of the mutations. We perform a one-sided binomial test and remove any locus with a p-value < 0.05 (plus a Bonferroni correction by a factor of 2477, the number of loci examined). Formally, for each locus let

$p =$ Observed homozygosity of all alleles genotyped. $0 \leq p \leq 1$.

$n =$ Number of mutations observed.

$k =$ Number of mutations that are from a homozygous-parent to a homozygous-child

$$\text{P-value} = \sum_{i=k}^{n} \binom{n}{i} p^{2i} (1 - p^2)^{n-i}$$

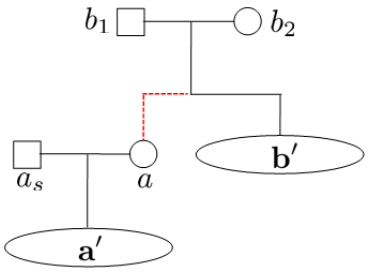
Note that we have p^2 instead of p because we are observing two homozygous genotypes simultaneously. In this manner, 49 loci were removed from the trio approach.

Note S3. Details of the family approach in mutation detection

Assigning alleles to haplotypes: a constraint satisfaction problem

Since Allegro cannot determine haplotypes in the presence of a mutation (a Mendelian inheritance error), we initially mask out any locus that generates inheritance errors. Based on neighboring loci, Allegro imputes haplotypes into the masked loci. To optimally assign haplotypes to alleles, this problem can now be posed as a constraint satisfaction problem (CSP) and solved.

Goal: Given the family structure below, a set of haplotypes, and a set of alleles at a particular locus, assign haplotypes to alleles in a way that is consistent with the family structure.



Solution:

We formulate this problem in terms of a constraint satisfaction problem (CSP). Suppose we have individuals $I_1, I_2, \ldots, I_m$ and haplotypes $H_1, H_2, \ldots, H_n$, where n is even. Then, we can write the alleles in a sparse matrix format, as shown below. Each row is an individual, each column is a haplotype, and each matrix entry is the pair of alleles of the corresponding individual. Since each individual has 2 haplotypes, we have 2 matrix entries per row. The CSP problem is then to find the suitable unique number for each matrix entry.

Formally, the set of variables is the non-empty entries of the matrix, denoted as X_{ij}. In the example below, there are 6 variables. Each variable has a domain of values. Since loci are diploid, we have 2 values per domain. There are two constraints for this CSP: (1) The non-empty entries of each column must be equal. (2) The non-empty entries of each row must be

different, unless the domain is a homozygote, such as "7, 7". The desired outcome of the CSP is shown below.

CSP in the presence of mutation. Without mutations, we simply run the algorithm over the entire family in one batch. However, suppose that there is a candidate mutant in the proband, then a single batch CSP would yield an empty solution. To resolve this, we instead use the following steps: (1) Run CSP over b_1, b_2, and b'. This group should carry the ancestral allele. (2) Run CSP over a, a_s, and a'. This group should carry the mutant allele. At this point, we should have the 6 six haplotypes assigned to the alleles, with 1 haplotype assigned inconsistently between the two groups. Thus, in combining the results, we have successfully identified the haplotype carrying the mutant, the mutant allele, and the ancestral allele.

Example. In this family, we have 2 members of *a'* and 2 members of *b'*. We first run CSP over the ancestral group, yielding:

$$\begin{array}{cccccc} & H_1 & H_2 & H_3 & H_4 & \\ b_1 & 4,8 & 4,8 & & & \\ b_2 & & & 2,8 & 2,8 & \longrightarrow \\ b' & 2,8 & & 2,8 & & \\ b' & & 4,8 & & 4,8 & \end{array} \qquad \begin{array}{ccccc} & H_1 & H_2 & H_3 & H_4 \\ b_1 & 8 & 4 & & \\ b_2 & & & 2 & 8 \\ b' & 8 & & 2 & \\ b' & & 4 & & 8 \end{array}$$

This yields a haplotype assignment of

$$\{H_1 = 8, H_2 = 4, H_3 = 2, H_4 = 8\}$$

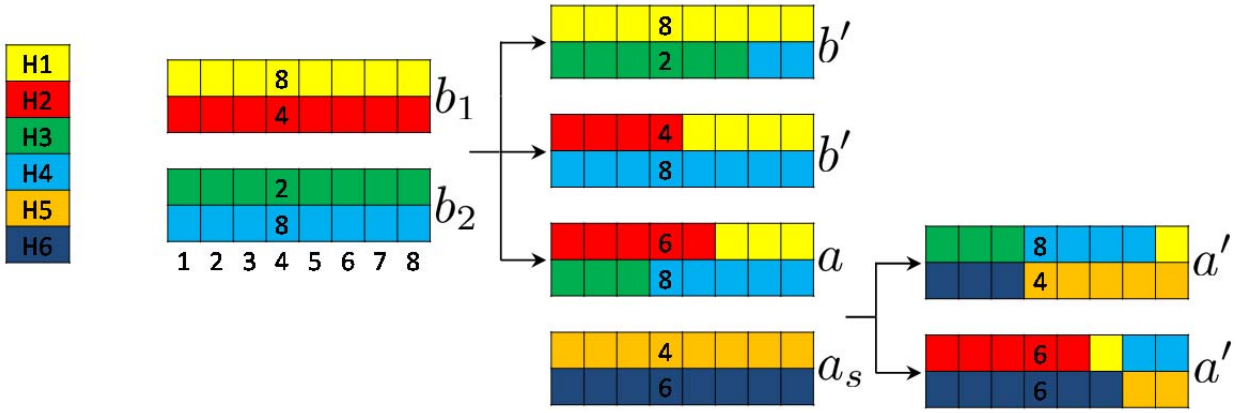
Next, we run CSP of the mutant group, yielding:

$$\begin{array}{cccccc} & H_2 & H_4 & H_5 & H_6 & \\ a & 6,8 & 6,8 & & & \\ a_s & & & 4,6 & 4,6 & \longrightarrow \\ a' & & 4,8 & 4,8 & & \\ a' & 6,6 & & & 6,6 & \end{array} \qquad \begin{array}{ccccc} & H_2 & H_4 & H_5 & H_6 \\ a & 6 & 8 & & \\ a_s & & & 4 & 6 \\ a' & & 8 & 4 & \\ a' & 6 & & & 6 \end{array}$$

This yields a haplotype assignment of

$$\{H_2 = 6, H_4 = 8, H_5 = 4, H_6 = 6\}$$

We see that haplotype 2 is inconsistent between the two sets of assignments. Therefore, haplotype 2 is the one of interest, carrying ancestral allele 4 and mutant allele 6. Below is the full haplotype of the entire region and the 4th microsatellite locus as the mutating one:



Note that in this example, if we instead used the trio approach, i.e. we are limited to the data of $b_1 = (4, 8)$, $b_2 = (2, 8)$, $a = (6, 8)$. The mutant allele of 6 would be detected, but we would not be able to find the parental origin of the mutation. Thus, by using additional family members and neighboring loci, the family approach allows parental assignment of the mutation.

Note S4. Testing the Amos Hypothesis

Amos et al.[33] suggested that if the parental allele is heterozygous, the mutation rate will be elevated compared to homozygous parental alleles. This would have significant implications as population size (N) is related to heterozygosity, and thus $\mu = f(N)$ would significantly undermine the population genetics assumption that N and μ are independent.

We tested the Amos hypothesis as follows:

> The Amos Hypothesis: If the parent is more heterozygous (i.e. length differences of alleles are large), then the mutation rate is higher.
>
> Prediction of the hypothesis: For each microsatellite mutation, the magnitude of length difference in the parent who transmitted the mutation is expected to be larger than that of an individual randomly sampled at the same microsatellite locus.
>
> Definitions:
>
> Ω The entire sample space of individuals genotyped.
> S' The subspace of parents who transmitted mutations.
> S The subspace of individuals who do not belong to S' (complement of S').
>
> A_j B_j A random sample of a pair of alleles from S at locus j.
> A'_j B'_j Likewise, but sampled from S'.
>
> L_j The length difference of the alleles, i.e. $L_j = |A_j - B_j|$
> L'_j Likewise, but sampled from S'.

Formalized hypothesis: Given the definitions, and assuming the hypothesis is true, then $L' - L > 0$ is true over the set of loci J.

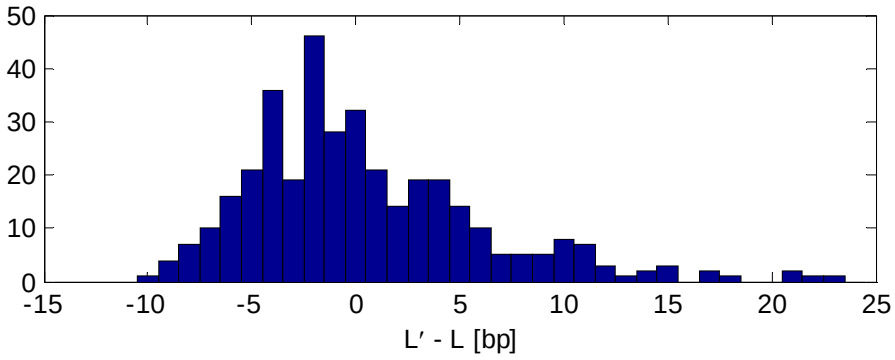
Testing the hypothesis:

Dataset: 363 mutations from the family approach. We do not use trio mutations for this analysis, because in trio mutations we have directly filtered based on the excessive homozygosity of certain mutant loci. Since the filter directly influences the parameter we are trying to estimate, we cannot use the larger trio dataset.

Sampling L': We use the parents who transmitted the mutations. Thus, $l'_j =$ parental allele difference for case j.

Sampling L: For each mutation case, we take that locus' allelic distribution, and independently sample n length differences and take the average. More precisely, at case j, we sample and compute $l_j = \frac{1}{n}\sum_{i=1}^{n} l_{j,i}$..

Results: Below is the histogram for the 363 data points of $l'_j - l_j$, with $n = 1000$. To test whether the mean is significantly different from 0, we perform a one-sample two-sided t-test, as was done by Amos et al., and obtain $t_{362} = 1.48,\ p = 0.14$. Therefore, our data provide no significant support for the Amos hypothesis.



Chapter 4

A model of microsatellite evolution

James X. Sun, Agnar Helgason, Gisli Masson, Sigríður Sunna Ebenesersdóttir, Heng Li, Swapan Mallick, Sante Gnerre, Nick Patterson, Augustine Kong, David Reich & Kari Stefansson

Based on observations of microsatellite mutation from the previous chapter, we build a model to estimate key parameters of evolution without calibration to the fossil record. The sequence mutation rate is estimated at 1.4×10^{-8} – 2.3×10^{-8} per base pair per generation (90% credible interval), and human-chimpanzee speciation at 3.7-6.6 million years ago (Mya).

Microsatellites have been widely used for making inferences about evolutionary history, because they are highly polymorphic and relatively unaffected by ascertainment biases that can skew inferences based on single nucleotide polymorphism (SNP) data. However, the accuracy of these inferences has been limited by a poor understanding of the mutation process. Using the empirically observed mutations, we developed a new model of microsatellite evolution that can estimate the time to the most recent common ancestor (TMRCA) given data (Methods, Note S5). This model accounts for: (1) the length dependence of mutation rate on allele length and parental age (Fig. 2A,C); (2) the step-size of mutations (Fig. 2B); (3) the size constraints on allele length (Fig. 2D, S7-8); and (4) the variation in generation interval (which affects parental age and thus mutation rate) over history. In contrast to the Generalized Stepwise Mutation Model (GSMM), which predicts a linear increase of average squared distance (ASD; see Methods for definition)

over time, the new model predicts a sub-linear increase (Fig. 3) and saturation of the molecular clock, due to the size constraints on allele lengths. To implement the model, we used a Bayesian hierarchical approach, where we first generated global parameters common to all loci, followed by locus-specific parameters, and finally the microsatellite alleles at each locus (Methods). We used Markov Chain Monte Carlo to infer model parameters such as the TMRCA.

We validated the model in three ways (Methods). First, we simulated datasets in which we know the true sequence mutation rate and TMRCA, and found that our model is unbiased in estimating sequence mutation rate while producing accurate estimates of the standard error (Methods, Note S6). Second, we carried out sensitivity analyses in which we perturbed model parameters and found that our key inferences are robust (Note S6, Fig. S9). Third, we empirically validated the model by analyzing 23 individuals for whom whole genome sequence (WGS) data was available[1], and comparing the ASD to the surrounding sequence heterozygosity which is proportional to the TMRCA. The ASD predicted by our model is similar to the empirical curve that combines the 23 individuals (Fig. 3, S10); in contrast, the predictions of the GSMM deviate.

Our direct measurement of the mutation rate and inference of a microsatellite mutation model allows us to infer evolutionary parameters without calibration to the fossil record. Using the empirical ASD at the dinucleotide microsatellites in each of the 23 individuals of European, East Asian and sub-Saharan African ancestry for whom we also had whole genome sequence data (Methods), we inferred a sequence mutation rate as well as estimates of the genome-wide average time since the most recent common ancestor (Table 2, S6), inferring a 90% credible interval (CI) based on a Bayesian approach that integrates over uncertainty in the parameters of the model (Methods; Note S5, Table S5). Our mutation rate estimates for each individual are shown in Table S6. Empirically, we find that the mutation rates tend to be more similar for within than across populations, which may be due to shared history (Fig. S12) (this observation is not likely to be an artifact of differences in demographic history across populations, as it persists when we use a more accurate fit to the demography via a 2-bottleneck model; Methods; Fig. S13). The mutation rate differences across populations are not statistically significant, and so we pooled our data across the 23 individuals to produce a maximally precise estimate of 1.82×10^{-8} per bp per generation (90% CI 1.40-2.28$\times10^{-8}$/bp/generation; Table 2).

Our inference of the sequence mutation rate is consistent with Nachman and Crowell's estimate of $\bar{\mu}_{seq}$= 1.3-2.7×10^{-8}/bp/generation based on calibration to the fossil record[2]. Our mutation rate estimate is also consistent with Kondrashov's direct estimate of $\hat{\mu}_{seq}$ = 1.8×10^{-8}/bp/generation[3] from studies of disease causing genes. However, the lower bound of our 90% CI is higher than two recent studies based on whole-genome sequencing (WGS): Roach et al.[4] identified 28 sequence mutations between two parents and their two offspring and used them to estimate a mutation rate of $\hat{\mu}_{seq}$= 1.1×10^{-8}/bp/generation and the 1000 Genomes Project[1,5] identified 84 sequence mutations in two father-mother-child trios and used them to estimate mutation rates of $\hat{\mu}_{seq}$ = 1.0×10^{-8} and 1.2×10^{-8}/bp/generation. We considered the possibility that this discrepancy might be due to ascertainment bias in our data because the microsatellites we analyzed were selected to be highly polymorphic (for disease gene mapping) which could cause a too-high ASD; however, this would overestimate of TMRCA and consequently underestimate mutation rate, opposite to what is necessary to explain the discrepancy (Fig. S12 and Note S6). We hypothesize that the lower estimates from the WGS studies are due to a combination of: (i) the limited number of mutations detected in the WGS studies which means that their confidence intervals are in fact consistent with ours, (ii) underestimation of the false-negative rate in the WGS studies, and (iii) variability in the mutation rate across individuals so that a few families cannot provide a reliable estimate of the population-wide rate. There is already empirical evidence for high variability in the sequence mutation process across individuals: in one family from the 1000 Genomes study[8], the father transmitted 92% of mutations but in the other 36%. Studies of sequence substitution in many families are important, as they will make it possible to measure population-wide rates and study features of the sequence substitution process not accessible to microsatellite-based analysis.

Our direct estimation of the microsatellite mutation rate, combined with comparative genomics data, also allows us to obtain an estimate the date of human-chimpanzee speciation τ_{HC}, defined as the time of last gene flow between human and chimpanzee ancestors, without relying on a calibration to the fossil record[6,7]. We estimate a genome-wide genetic divergence time t_{HC} = 5.80-9.77 Mya[8] (Methods; Table 2). By definition, this must be older than the speciation date τ_{HC}.

We then inferred the human-chimpanzee speciation date to be τ_{HC} = 3.75-6.57 Mya by integrating our estimate of t_{HC} with a prior distribution on τ_{HC}/t_{HC} of 0.663±0.041, obtained from published point estimates of τ_{HC}/t_{HC} = 0.61-0.68[9,10], and an upper bound of $\tau_{HC}/t_{HC} < 0.73$ that we newly obtained by analyzing human-chimpanzee sequence data in regions with a reduced divergence compared to the autosomal average due to being (1) on chromosome X, (2) in proximity to genes, and (3) near divergent sites that cluster humans and chimpanzees to the exclusion of gorilla (Note S8). Both our upper bound of $\tau_{HC} < 6.57$ Mya, and a completely independent upper bound of $\tau_{HC} < 6.3$ Mya that we obtain in Note S7 by calibrating to the fossil record of human-orangutan speciation, are lower than the date of 6.8-7.2 Mya[11] for *Sahelanthropus tchadensis,* a fossil that is often interpreted as being on the human lineage after the final separation of human and chimpanzee ancestors[12] because it shares derived features with other hominins such as bipedal posture, reduced canines and expanded post-canines with thicker enamel[13]. If our speciation date is correct, a possibility is that *Sahelanthropus* was not a hominin, but instead shared independently-derived similarities (homoplasies), as suggested in a review that cautioned against interpreting fossils close to human-chimpanzee speciation as on the human lineage[14]. Alternatively, populations with hominin traits may have continued to exchange genes with chimpanzee ancestors after *Sahelanthropus*, thus explaining why fossils with hominin traits predate the time of final human-chimpanzee speciation[6]. A final possibility is that the dates for *Sahelanthropus*, based primarily on cosmogenic nuclide beryllium isotopes[11], are too old.

Methods

Parameters used in inference of sequence mutation rate and human-ape divergence times

For microsatellite evolution modeling, inference of sequence mutation rate, and inference of human-ape divergence times, many parameters need to be fit to the data. Our Bayesian procedure takes into account uncertainty in each parameter by inferring a distribution for its possible value conditional on the data. An abbreviated description is below. Full details are given in Note S5, and a summary of parameters and prior distributions is given in Table S5.

Generation interval: We assume that the generation interval changed over time according to the logistic function $g(t) = g_{anc} + \frac{g_{now} - g_{anc}}{1+\exp\left(\frac{t-t_0}{t_{0/4}}\right)}$, with ancestral generation-time g_{anc}, modern-

day generation-time g_{now}, and switching time t_0. Based on studies of generation interval in humans and chimpanzees, we used Bayesian prior distributions that incorporate the uncertainties. Based on interviews with experts on chimpanzee and gorilla demographic structure (Linda Vigilant and Kevin Langergraber, personal communication), we assume that g_{anc} is sampled from a normal distribution of 22.5±4.2 (mean ± SD) years, covering a 95% confidence interval of generation times from 15 to 30 years to reflect our uncertainty. Based on a reading of the literature on present-day humans[15] as well as Icelanders[16] (Fig. S14), and discussions with an expert in human generation interval (Jack Fenner, personal communication) we sample g_{now} to be 29±2 years. We sample t_0 to be a mixture of 3 equally weighted exponential distributions, with means of 50Kya, 200Kya, and 2Mya, corresponding to hypothetical changes in human life history around the time of the Upper Paleolithic revolution, evolution of anatomically modern humans, and evolution of *Homo erectus*. We then obtain paternal and maternal-specific curves: $g_{pat}(t) = g(t) + \Delta(t)/2$ and $g_{mat}(t) = g(t) - \Delta(t)/2$, where $\Delta(t) = \Delta_{anc} + \frac{\Delta_{now} - \Delta_{anc}}{1+\exp\left(\frac{t-t_0}{t_0/4}\right)}$ for the parental age difference (paternal minus maternal age): Δ_{now} is sampled to be 6.0±2.0 years and Δ_{anc} as 0.5±3.3 years (based on suggestions from Linda Vigilant and Kevin Langergraber, personal communication). The switching time t_0 is obtained from the same sample as that of the generation-time curve. All other parameters are sampled independently of each other.

Mutation rate adjusted for generation interval $\left(\mu_g\right)$: We model the paternal rate as $\mu_{pat}(t) = \beta_{0,pat} + \beta_{1,pat} \cdot g_{pat}(t)$, the maternal rate as $\mu_{mat}(t) = \beta_{0,mat} + \beta_{1,mat} \cdot g_{mat}(t)$, and the overall rate $\mu_g(t)$ is the average of the two. To take into account the stochasticity of the regressions in Fig 2A, the slopes$\left(\beta_1\right)$ and intercepts $\left(\beta_0\right)$ are sampled from bivariate student-*t* distributions from standard linear regression analyses.

Human-chimpanzee divergence time (t_{HC}): A human-chimpanzee genetic divergence time can be obtained as $t_{HC} = \frac{\pi_{HC}/\pi_E}{\bar{\mu}_{HC}/\bar{\mu}_E} \cdot t_E$, where π_{HC}/π_E is estimated from the literature[8] to be 15.400±0.356. We use this formulation instead of the more straightforward $t_{HC} = \pi_{HC}/2\bar{\mu}_{HC}$ calculation because the ratio π_{HC}/π_E is genomic-region agnostic: π_{HC} fluctuates depending on the set of genomic filters used, but the ratio is more robust to genomic filters. Assuming that dependence of the sequence mutation rate on generation interval is the same as that of

microsatellites, $\frac{\bar{\mu}_{HC}}{\bar{\mu}_E} = \frac{1/t_{HC} \cdot \int_0^{t_{HC}} \mu_g(t)dt}{1/t_E \cdot \int_0^{t_E} \mu_g(t)dt}$. Since t_{HC} is much larger than t_E, the numerator is approximated as $\mu_g(\infty)$. Thus, the final expression for the human-chimpanzee divergence time becomes $t_{HC} = \frac{\pi_{HC}}{\pi_E} \cdot \frac{1}{\mu_g(\infty)} \cdot \int_0^{t_E} \mu_g(t)dt$.

Human-orangutan divergence time (t_{HO}): A human-orangutan genetic divergence time can be obtained as $t_{HO} = \frac{\pi_{HO}}{\pi_{HC}} \cdot t_{HC}$, making the assumption that the mutation rate averaged over the history of human-orangutan is the same as that of human-chimpanzee. From the literature, we estimate $\frac{\pi_{HO}}{\pi_{HC}}$ as a random sample from a normal distribution of 2.650±0.075.[6,9] We note that our inferences about human-orangutan genetic divergence time will be biased if mutation rates have changed over great ape history.

Human-chimpanzee speciation time (τ_{HC}): A human-chimpanzee speciation time can be obtained as $r \cdot t_{HC}$, where r is the ratio of human-chimpanzee speciation time to that of the genetic divergence time. We estimate r as a random sample from a normal distribution with mean 0.663, with the range of inferences of 0.61-0.68 from model-based analyses[9,10] (Note S8). We used a wide standard deviation of 0.041 (conservatively much wider than emerges from the model-based analysis) to reflect the fact that the ancestral human-chimpanzee population may have deviated from a model of size-constancy[33]. This choice is motivated by an analysis suggesting that a conservative upper bound on the ratio of human-chimpanzee speciation time to human-chimpanzee genetic divergence time is 0.73. This is chosen to be exactly 1.65 standard deviations above the mean, so that only 5% of the density is above the upper bound.

Definition of average squared distance (ASD)

At a particular locus, given microsatellite allele lengths $x_1, x_2, \ldots, x_n$, the ASD is defined as $ASD = \frac{1}{n(n-1)} \cdot \sum_{i,j} (x_i - x_j)^2$.

Building a model of microsatellite evolution assisted by flanking sequence heterozygosity

The model we used simulates the evolution of a pair of chromosomes from a common ancestor, over multiple loci and individuals. The model is hierarchical: At the top level, global parameters (Table S5) common to all loci are simulated, such as the genome-wide present-day sequence and microsatellite mutation rates, and generation-time effects. One level down, locus-specific

mutation rates are computed based on global parameters and locus-specific information (see below). At the third level, for each individual, a two-sample coalescent tree is generated.

For an individual whose genome sequence is available, diploid microsatellites genotypes are simulated as follows:

1. **Generate 1 set of genome-wide parameters** (Table S5), which are common across loci, sampling from the prior distributions obtained from the literature and our direct measurements in this study. This includes the genome-wide sequence mutation rate and microsatellite mutation rate.
2. **At locus $i = 1$, generate locus-specific mutation rate $\mu_{msat,i}$.** The local microsatellite mutation rate is the genome-wide rate multiplied by l_i/l_{genome}, where l_{genome} is the genome-wide mean microsatellite length, and l_i is the locus-specific length (averaged across individuals). The local variation in microsatellite mutation rate is modeled to be purely due to allele length variation, which strongly influences mutation rate (Fig. 2C).
3. **At the locus, generate locus-specific mutation rate $\mu_{seq,i}$.** Analogous to step 2, the local sequence mutation rate is the genome-wide rate multiplied by D_i/D_{genome}, where D_i is the local human-macaque divergence, and D_{genome} is the genome-wide human-macaque divergence. The local variation in sequence mutation rate is modeled to be purely due to human-macaque divergence variation, which is known to strongly influence mutation rate.
4. **At the locus, generate coalescent time t_i,** using local sequence heterozygosity if available. The key is that the coalescent tree is shared between microsatellites and sequence, and if the local sequence heterozygosity is highly precise, it puts a strong constraint on the local TMRCA. The coalescent time is drawn from a gamma distribution with mean: $\frac{N_i+1}{\lambda_i+1/\tau_{genome,i}}$, where $\lambda_i = 2\mu_{seq,i}D_i$, N_i/D_i is the local heterozygosity, and $\tau_{genome,i} = \theta_{genome}/2\mu_{seq,i}$ is the genome-wide average TMRCA. Note that if D_i is small, we revert to the genome-wide TMRCA, but if D_i is large, the locus-specific heterozygosity overwhelms the genome-wide estimate. The gamma distribution is demography-free: If D_i is small, the distribution converges to an exponential with mean $\tau_{genome,i}$. To test our inference's robustness to demographic differences across populations, we use a 2-bottleneck demographic model (Fig. S13) and sample the coalescent time using rejection sampling with the following steps: (1)

Sample $\tau_{genome,i}$ with demography (distributions for each population shown in Fig. S13B); (2) calculate the importance ratio of $r = \exp\left[(N_i - \lambda_i t) \cdot \ln(\lambda_i t) - \sum_{i=1}^{N} \ln i + \sum_{i=1}^{\lfloor \lambda_i t \rfloor} \ln i\right]$; (3) accept t with probability r; (4) If rejected, go to step (1).

5. **Simulate mutations.** Mutations are sequentially generated from the root of the coalescent tree, using our model of microsatellite evolution which has length constraints and time-varying mutation rate as follows: At time t on the coalescent tree, the mutation rate is determined using parental length $y(t)$, mutation rate μ_i, and the mutation rate relative to the present, taking into account variation in generation-time: $\mu_g(t)/\mu_g(0)$. We model this as: $\mu_i(t) = \left(m_\mu \cdot y(t) + \mu_i\right) \cdot \mu_g(t)/\mu_g(0)$. The slope parameter m_μ is empirically determined from Fig. 2C. The waiting time until a mutation is sampled from an exponential distribution with mean of $1/\mu(t)$ generations. Once a mutation event occurs, its length is $l_{child} = (1 + m/\sigma)\, l_{parent} + X$, where m is the negative slope reflecting the length constraint in Fig. S7, σ is the standard deviation of the allelic distribution at a locus, l_{parent} is the parent allele length, and X is the mutational length, sampled from the histogram in Fig. 2B. At the root of the tree, without-loss-of-generality the absolute length is set to be 0. Using this scheme of generating mutation events and mutation lengths, we begin at the root of the tree and iterate until the leaves are reached. The leaves are the sets of sampled microsatellite alleles, which are used to compute ASD. To obtain time in units of years, we rescale branch lengths of the coalescent tree and mutation rates by $g(t)$, which is the generation-interval logistic function described above.
6. **Record ASD between the two microsatellites**, and go to Step 2, with i incremented by 1.

Inferences of sequence mutation rate and TMRCA using the microsatellite evolution model

We use a Markov Chain Montel Carlo (MCMC) approach to obtain the posterior distribution for present-day sequence mutation rate in a single diploid individual. This algorithm is a variation of "algorithm F" of Marjoram et al[17], and is as follows:

1. Sample a set of global parameters λ from their prior distribution (Table S5).
2. Propose a move of the sequence mutation rate from μ_{seq} to μ'_{seq}. We use μ'_{seq} as a random walk, sampled from a normal distribution with mean μ_{seq}, and standard deviation 0.5×10^{-8}.

3. At locus i:
    a. Generate 1000 pairs of microsatellite alleles using our evolution model with parameters μ'_{seq} and λ.
    b. Calculate ASD. Thus, we now have 1000 samples of simulated ASD.
    c. Compute the error distance $d_i = (mean(ASD_{sim}) - ASD_{real})^2$ between the simulated ASD and the real ASD of the individual.
4. Sum the error distance across all loci: $d_{total} = \sqrt{\sum_i d_i}$. If $d_{total} < \epsilon$, accept and set μ_{seq} to be μ'_{seq} and go to step 2. Otherwise, reject μ'_{seq}. We choose ϵ such that the overall acceptance rate of the MCMC is between 10% and 50%. (Note that since the proposal function is symmetric, and we choose a flat prior on μ_{seq}, we do not need to calculate the ratio as described in Step F4 of Marjoram et al., because the ratio is always 1.)

The result of MCMC is a correlated $\mu_{seq}|\lambda$ chain. To collect independent samples, the autocorrelation function of the chain is calculated and the correlogram is plotted. The first lag in which the correlation coefficient drops below 0.1 is recorded. Call this n_{lag}. Then, we thin the chain and collect at every n_{lag}-th sample. Finally, we run 1000 independently sampled $\mu_{seq}|\lambda$ and combine the thinned samples to produce the overall posterior distribution for μ_{seq}.

The above is for a single individual. To combine the individuals, we first treat all individuals as independent, conditioned upon each locus. Because of genealogy-sharing between individuals, especially when deeper in the coalescent tree (say, the past 100 thousand years), the individuals are expected to have shared mutations, and therefore may not be independent samples. Therefore, to obtain proper standard errors for the combined mutation rate, we performed a jackknife procedure[18], where each locus (assumed to be independent due to the distant spacing of microsatellite loci) is removed at a time. This gives the final set of standard errors for the sequence mutation rate in Table 2.

References

1. Durbin, R.M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
2. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).
3. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* **21**, 12-27 (2003).
4. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-9 (2010).
5. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature genetics* **43**, 712-714 (2011).
6. Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103-8 (2006).
7. Steiper, M.E. & Young, N.M. Primate molecular divergence dates. *Molecular phylogenetics and evolution* **41**, 384-94 (2006).
8. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-22 (2010).
9. Burgess, R. & Yang, Z. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol Biol Evol* **25**, 1979-94 (2008).
10. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics* **5**, e1000471 (2009).
11. Lebatard, A.E. *et al.* Cosmogenic nuclide dating of Sahelanthropus tchadensis and Australopithecus bahrelghazali: Mio-Pliocene hominids from Chad. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 3226-31 (2008).
12. Brunet, M. *et al.* A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418**, 145-51 (2002).
13. Lieberman, D.E. *The Evolution of the Human Head*, (Belknap Press of Harvard University Press, 2011).
14. Wood, B. & Harrison, T. The evolutionary context of the first hominins. *Nature* **470**, 347-52 (2011).

15. Fenner, J.N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* **128**, 415-23 (2005).
16. Helgason, A., Hrafnkelsson, B., Gulcher, J.R., Ward, R. & Stefansson, K. A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am J Hum Genet* **72**, 1370-88 (2003).
17. Marjoram, P., Molitor, J., Plagnol, V. & Tavare, S. Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci U S A* **100**, 15324-8 (2003).
18. Efron, B. & Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* **37**, 36-48 (1983).
19. Keinan, A., Mullikin, J.C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* **39**, 1251-5 (2007).
20. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814-28 (2008).
21. Leakey, M.G., Feibel, C.S., McDougall, I. & Walker, A. New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. *Nature* **376**, 565-71 (1995).
22. MacLatchy, L., Gebo, D., Kityo, R. & Pilbeam, D. Postcranial functional morphology of Morotopithecus bishopi, with implications for the evolution of modern ape locomotion. *J Hum Evol* **39**, 159-83 (2000).
23. Dutheil, J.Y. *et al.* Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* **183**, 259-74 (2009).
24. Yang, Z. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome biology and evolution* **2**, 200-11 (2010).
25. Mallick, S., Gnerre, S., Muller, P. & Reich, D. The difficulty of avoiding false positives in genome scans for natural selection. *Genome research* **19**, 922-33 (2009).
26. Presgraves, D.C. & Yi, S.V. Doubts about complex speciation between humans and chimpanzees. *Trends in ecology & evolution* **24**, 533-40 (2009).
27. Gelman, A., Carlin, J., Stern, H. & Rubin, D. *Bayesian Data Analysis*, (2004).

Table 2. Estimates of mutation rates and human-ape divergence times

	Mean	$5^{th} – 95^{th}$ percentile*	mean	$5^{th} – 95^{th}$ percentile
Present-day mutation rates†	*units: per generation per site*		*units: per year per site*	
di-nucleotide microsatellite rate (per locus)	2.73×10^{-4}	$2.56 – 2.91 \times 10^{-4}$	9.47×10^{-6}	$8.29 – 10.82 \times 10^{-6}$
$\hat{\mu}_{seq}$: nucleotide substitution rate (per base)	1.82×10^{-8}	$1.40 – 2.28 \times 10^{-8}$	6.76×10^{-10}	$5.11 – 8.41 \times 10^{-10}$
Genetic divergence times††	*units: thousand generations ago*		*units: million years ago*	
t_{CEU}: Western Europeans	22.8	17.8 – 29.6	0.546	0.426 – 0.709
t_{YRI}: Yoruba (African)	30.2	23.6 – 39.2	0.720	0.562 – 0.933
t_{HC}: human-chimpanzee	352	272 – 459	7.49	5.80 – 9.77
t_{HO}: human-orangutan	932	717 – 1220	19.8	15.2 – 25.9
τ_{HC}: human-chimpanzee speciation time	233	176 – 309	4.97	3.75 – 6.57

* 90% Bayesian credible interval obtained from the Bayesian posterior distribution shown in Fig. S12-S13.

† The microsatellite mutation rate is directly measured. The sequence mutation rate is first time-averaged through human population history, and then the present-day inferred rate is reported. There is a difference between the present-day and time-averaged rates due to generation-time difference in human history. Most experts believe that the ancestral human generation interval was less than that of the present, leading to a present-day mutation rate per generation that is higher than the time-averaged value (which we estimate as mean: 2.54×10^{-8}, 90% CI: $2.03 – 2.97 \times 10^{-8}$), and a present-day mutation rate per year that is lower than the time-averaged value (which we estimate as mean: 1.07×10^{-9}, 90% CI: $0.82 – 1.47 \times 10^{-9}$).

†† The Western European genetic divergence time was estimated from the sequence substitution rate as $t_{CEU} = \frac{\theta_{CEU}}{2 \cdot \mu_{seq}}$, where θ_{CEU} has mean 8.12×10^{-4} and standard error 0.26×10^{-4} from the Illumina WGS data (Table S6). We further performed an ascertainment bias correction, which corrects for the fact that on average the microsatellite loci we analyzed have about a 4% increased TMRCA compared with random loci in the genome (Methods, Table S7). Similarly, the Yoruba genetic divergence time is obtained as $t_{YRI} = \frac{\theta_{YRI}}{2 \cdot \mu_{seq}}$, where θ_{CEU} has mean 1.08×10^{-3} and standard error 0.29×10^{-4} from the Illumina data. The other divergence times were obtained from a scaling of t_{CEU} (see Methods for details).

Table S5. Bayesian parameters for evolution modeling

Class	Description	Sampling distribution	Mean (SD)	Units
Generation interval	g_{anc} Generation time in the human-chimp ancestor	Normal	22.5 (4.24)	years
	g_{now} Present-day human generation time	Normal	29.0 (2.04)	years
	t_0 Inflection point of the logistic curve	Mixture of 3 exponentials of equal probability	50 200 2000	thousand years
Parental age difference (paternal minus maternal)	Δ_{anc} Age difference in the human-chimp ancestor	Normal	0.50 (3.33)	years
	Δ_{now} Present-day human parental age difference	Normal	6.00 (2.04)	years
Mutation rate as a function of generation interval	$\beta_{0,pat}$ Paternal mutation rate, baseline (at age 0)	multivariate t (sampled from Fig 2A)	see Fig 2A	μ
	$\beta_{0,mat}$ Maternal mutation rate, baseline (at age 0)			μ
	$\beta_{1,pat}$ Slope of paternal mutation rate with age			μ per year
	$\beta_{1,mat}$ Slope of maternal mutation rate with age			μ per year
Mutation rate with length	m_μ Slope of mutation rate vs. absolute allele length	Normal	1.66 (0.30) x10^{-5}	μ per repeat unit
Length constraint	Slope of mutational direction vs. relative allele length	Normal	-0.419 (0.060)	repeat units per SD
For human-chimp divergence time	π_{HC}/π_E Ratio of human-chimp to Western European sequence divergence	Normal	15.4 (0.356)	dimensionless
For human-chimp speciation time	τ_{HC}/t_{HC} Ratio of human-chimp speciation time to genetic divergence time	Normal	0.663 (0.041)	dimensionless
For human-orangutan divergence time	π_{HO}/π_{HC} Ratio of human-orangutan to human-chimp sequence divergence	Normal	2.65 (0.075)	dimensionless

Note: This table gives the prior distributions used in our Bayesian modeling analysis, obtained from surveys of the literature and discussions with experts in relevant fields (our approach to obtain these priors is also discussed in the Methods section). The experts we consulted were John Hawks and David Pilbeam regarding the ape fossil record; Kevin Langergraber and Linda Vigilant regarding primate generation intervals and plausible generation intervals in the ancestral population; and Jack Fenner regarding the recent human generation interval. We thank all these colleagues for useful discussions and advice.

The parameters above the thick black line are "global parameters" used for microsatellite evolution modeling, in which the same set of parameter values apply to all loci, per simulation. The parameters below the line are used after the posterior TMRCA of Western Europeans has been obtained.

Table S6. Mutation rate estimates and sequence heterozygosities in 23 individuals

Illumina dataset		**Sequence heterozygosity**		**Mutation rate estimates (x 10^{-8})**			
Population	ID	mean	std error	mean	std error	5th percentile	95th percentile
CEU	**NA12891**	0.000860	0.000026	1.65	0.44	1.00	2.43
CEU	**NA12892**	0.000838	0.000026	1.92	0.37	1.33	2.56
CEU	**NA12878**	0.000838	0.000026	1.42	0.34	0.91	2.01
YRI	**NA19239**	0.001112	0.000027	1.80	0.44	1.12	2.55
YRI	**NA19238**	0.001048	0.000027	2.46	0.53	1.65	3.38
YRI	**NA18508**	0.001174	0.000028	1.18	0.35	0.64	1.79
YRI	NA19240	0.001168	0.000028	2.57	0.56	1.68	3.53
YRI	NA18507	0.001077	0.000031	2.12	0.53	1.33	3.04
YRI	NA18506	0.001141	0.000030	2.13	0.54	1.33	3.09

Complete Genomics dataset		**Sequence heterozygosity**		**Mutation rate estimates (x 10^{-8})**			
Population	ID	mean	std error	mean	std error	5th percentile	95th percentile
CEU	**NA12891**	0.000804	0.000025	1.36	0.31	0.90	1.90
CEU	**NA12892**	0.000804	0.000025	1.58	0.30	1.11	2.10
CEU	**NA12878**	0.000780	0.000026	1.15	0.25	0.77	1.58
CEU	NA06985	0.000800	0.000027	1.06	0.28	0.65	1.54
CEU	NA06994	0.000850	0.000029	0.91	0.20	0.61	1.25
CEU	NA07357	0.000794	0.000027	1.12	0.31	0.66	1.67
CEU	NA10851	0.000848	0.000029	1.00	0.23	0.66	1.40
CEU	NA12004	0.000841	0.000028	1.13	0.29	0.69	1.63
YRI	**NA19239**	0.001035	0.000026	1.50	0.35	0.96	2.08
YRI	**NA19238**	0.000980	0.000026	2.09	0.42	1.44	2.81
YRI	**NA18508**	0.001089	0.000027	1.06	0.29	0.62	1.57
YRI	NA18501	0.001062	0.000026	1.52	0.37	0.95	2.14
YRI	NA18502	0.001062	0.000027	2.86	0.53	1.98	3.72
YRI	NA18504	0.001059	0.000026	1.31	0.31	0.84	1.84
YRI	NA18505	0.001076	0.000027	1.27	0.29	0.82	1.77
YRI	NA18517	0.001083	0.000027	1.32	0.41	0.72	2.08
CHB	NA18526	0.000798	0.000027	1.89	0.36	1.32	2.50
CHB	NA18537	0.000766	0.000026	1.51	0.32	1.02	2.06
CHB	NA18555	0.000779	0.000026	1.38	0.28	0.94	1.88
CHB	NA18558	0.000770	0.000027	1.25	0.36	0.72	1.90

Mutation rates (in units of X*1e-8 /bp/generation) and Bayesian posterior intervals for each individual are shown here. In bold are individuals that overlap between the two datasets. See Figure S12 for a graphical representation.

Table S7. Ascertainment bias around microsatellite loci

Population	HapMap ID	msat region	random region	ratio
CEU	NA12891	0.088	0.085	1.037
CEU	NA12892	0.087	0.082	1.067
CEU	NA12878	0.090	0.085	1.057
YRI	NA19239	0.118	0.113	1.041
YRI	NA19238	0.110	0.105	1.046
YRI	NA18508	0.119	0.116	1.025
YRI	NA19240	0.121	0.114	1.059
YRI	NA18507	0.112	0.107	1.043
YRI	NA18506	0.118	0.114	1.036
human-chimp		2.347	2.248	1.044
human-macaque		7.978	7.884	1.012

We compared sequence heterozygosity (in units of $X*10^{-2}$) of regions surrounding our set of microsatellites to that of a random region. On average, the sequence heterozygosity was about 4% higher, suggesting that we have a slight bias towards the deeper trees in the human genome. Our modeling of evolutionary parameters explicitly corrects for such biases in two ways. First, we correct for unusual mutation rates around microsatellites by normalizing inferences by the ratio of local human-macaque sequence divergence to genome-wide average human-macaque sequence divergence. Second, we correct for unusual gene tree depths around microsatellites by making all inferences based on the comparison of local microsatellite ASD to heterozygosity in the flanking sequence data.

Table S8. Di-nucleotide microsatellite mutations by motif type

Repeat-type, by motif	mutations	transmissions	rate	std error
AC/CA/GT/TG	1102	4063534	2.71	0.08
AG/GA/CT/TC	27	93352	2.89	0.56
AT/TA	12	8760	13.70	3.95
CG/GC	0	0	N/A	N/A

Figures

Figure 3. Model validation using sequence-based estimates of TMRCA.

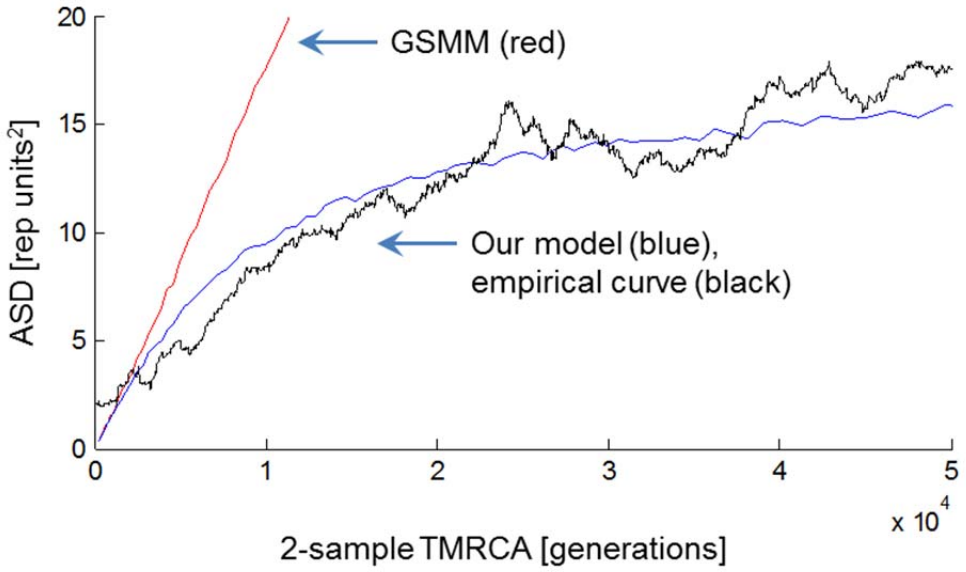


Figure 3. Empirical validation of our model with sequence-based estimates of TMRCA. In red is the simulation of ASD as a function of TMRCA for the standard random walk (GSMM) model. In blue is the simulation of our model, in which the non-linearity compared to GSMM is primarily due to the length constraint that we empirically observed in microsatellites. In black is the empirically observed ASD at microsatellites in 23 HapMap individuals as a function of sequence-based estimates of TMRCA, which is estimated using $\theta_{seq}/2\mu_{seq}$, where θ_{seq} is the local sequence diversity surrounding each microsatellite locus, and μ_{seq} is 1.82×10^{-8} (obtained from Table 2). The close match of the empirical curve to our model simulations suggests that our model works, and motivates the analysis in which we use the sequence substitution rate in small windows around the microsatellites we analyze to make inferences about evolutionary parameters like the sequence mutation rate.

Figure 4. Human-chimp speciation date inferred without a fossil calibration.

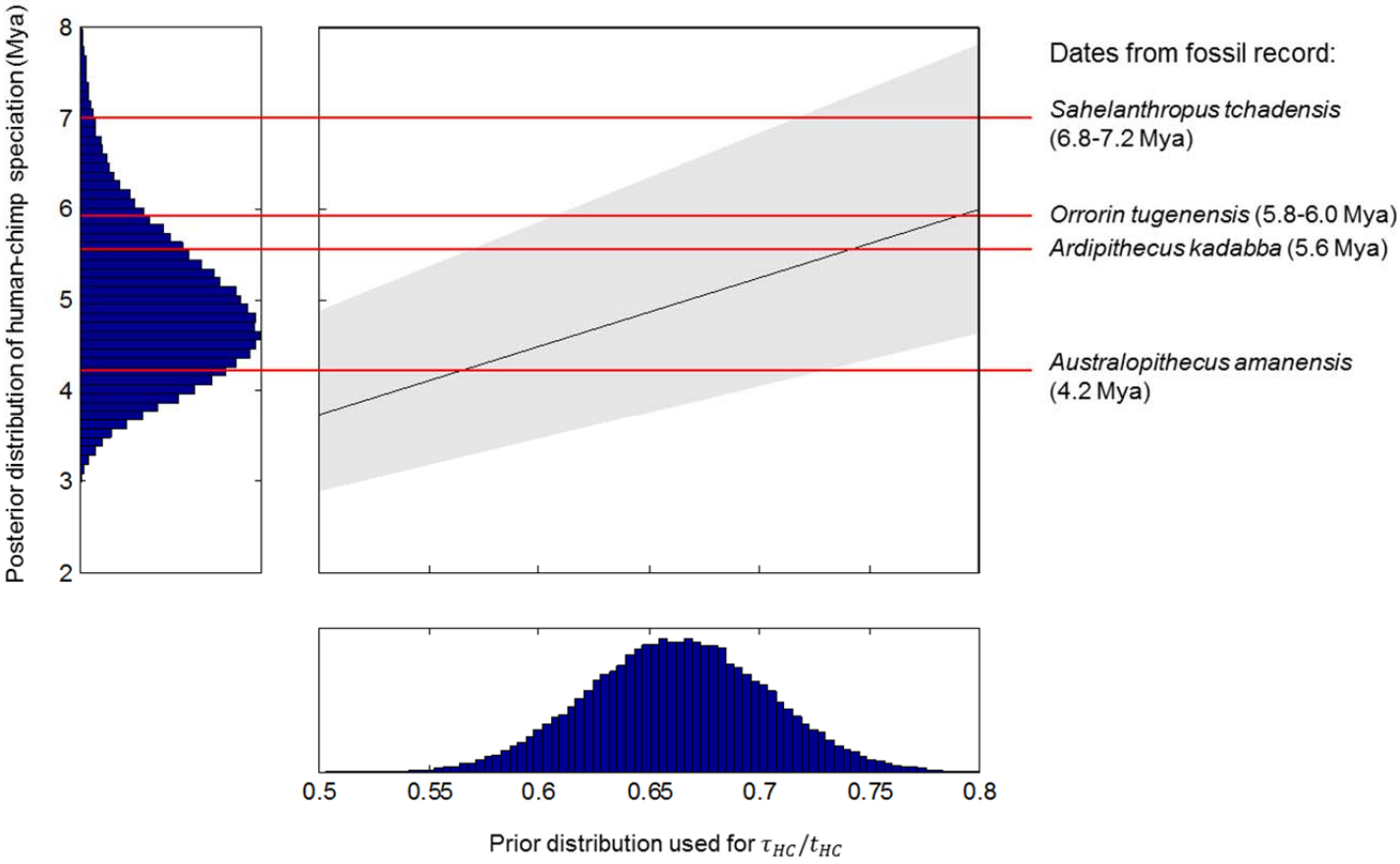


Figure 4. Human-chimpanzee speciation date inferred without a fossil calibration. In the square panel, we give the 90% Bayesian credible interval for human-chimpanzee speciation time (gray), for a range of plausible point values of the ratio of speciation time to divergence time τ_{HC}/t_{HC}. The blue curve shows our prior probability distribution for τ_{HC}/t_{HC}, justified in Note S8. The red horizontal lines are the dates of fossils that are candidates for being on the hominin lineage post-dating the speciation of humans and chimpanzees. *Australopithecus amanensis, Orrorin tugenensis* and *Ardipithecus kadabba* is within our plausible speciation times, while *Sahelanthropus tchadensis* pre-dates the inferred speciation time for all plausible values of τ_{HC}/t_{HC}. Our prior distribution for τ_{HC}/t_{HC} is shown in the bottom histogram, and our posterior distribution of human-chimpanzee speciation time is shown in the left histogram.

Figure S9. Sensitivity analysis of evolution model

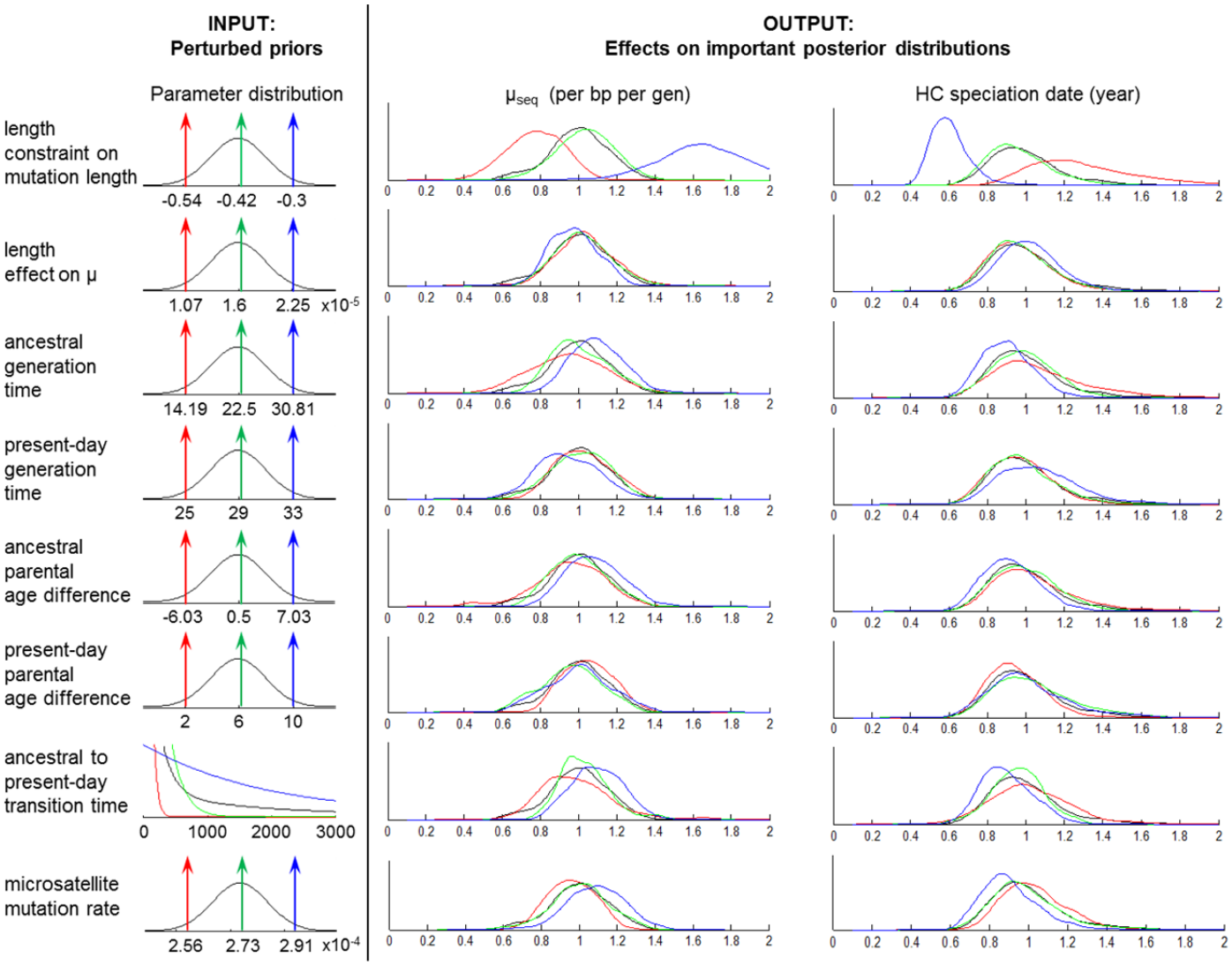
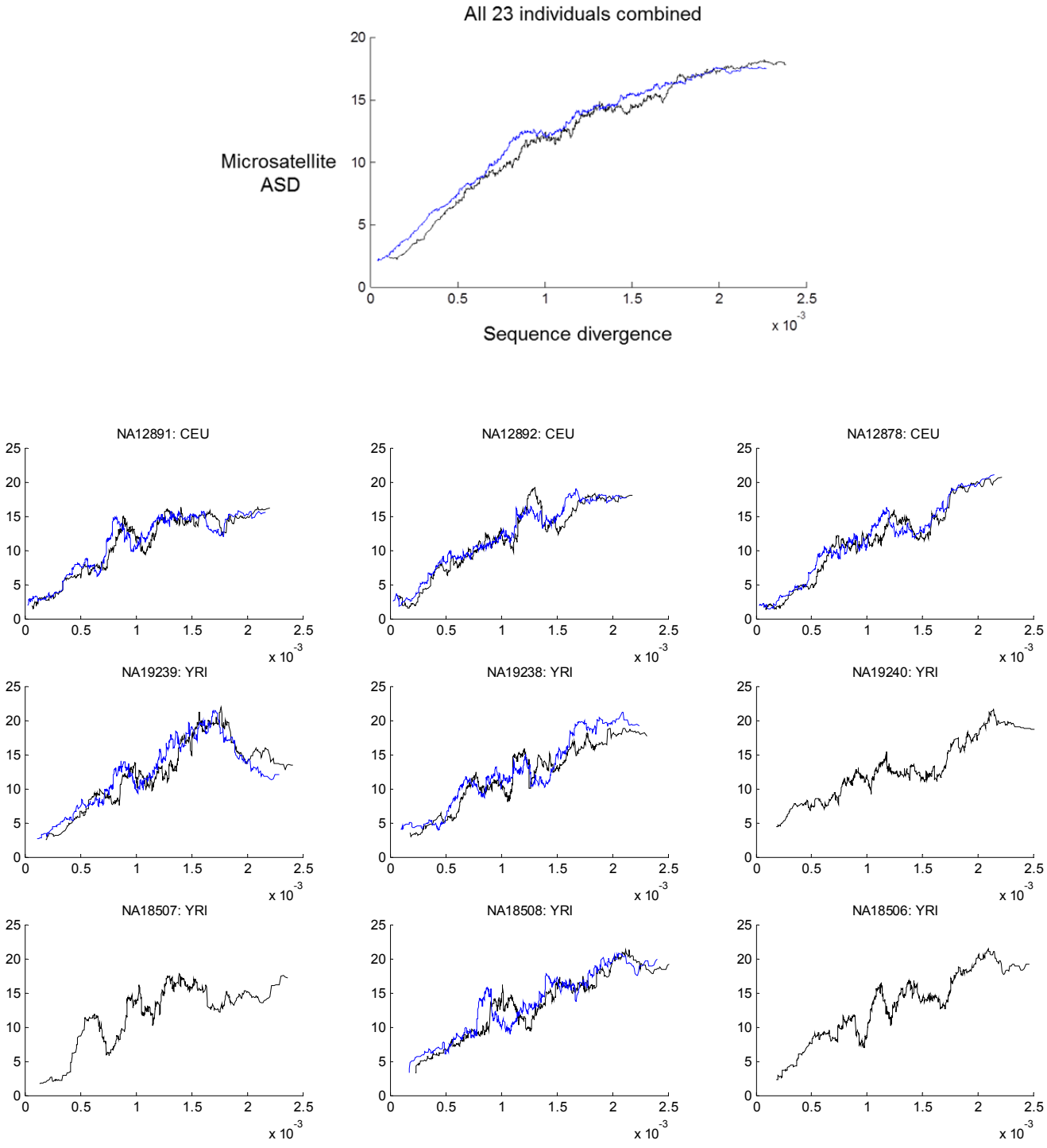


Figure S9. Sensitivity analysis of the evolution model. Our model of evolution is robust to changes in the prior distributions. Eight parameters that we use as priors are in the left column, with the default distributions in black. We tested robustness by setting each prior to have different point values (the mean, 5th percentile, and 95th percentile of the default distribution in black), and exploring how this changes the posterior distributions (the coloring of the posteriors correspond to the respective priors, all scaled by the mean of the black posterior). In the case of the "ancestral to present-day transition" in the generation time (t_0 in Note S5), the parameter distribution is a mixture of 3 exponentials (see Methods), and we test

robustness by sampling from each separately. Our posterior estimates are not much affected by the input parameters as long as they fall within the range of the priors. The exception is the length constraint (top row) that governs the non-linear mapping between TMRCA and ASD (Fig. 3), where we observe substantial differences. Note, however, that we obtain essentially the same posterior distribution when we use a point estimate corresponding to the mean of the prior distribution and the full prior distribution, which demonstrates the robustness of our inference procedure. Our evolutionary modeling updates its inference of the length constraint directly from comparing the microsatellite ASD to flanking sequence diversity; it is not solely based on our direct measurements. Thus, as long as we include the true value within the prior, we get robust results even for the length constraint parameter (see also Note S6).

Figure S10. Sequence divergence versus microsatellite ASD for 23 HapMap individuals



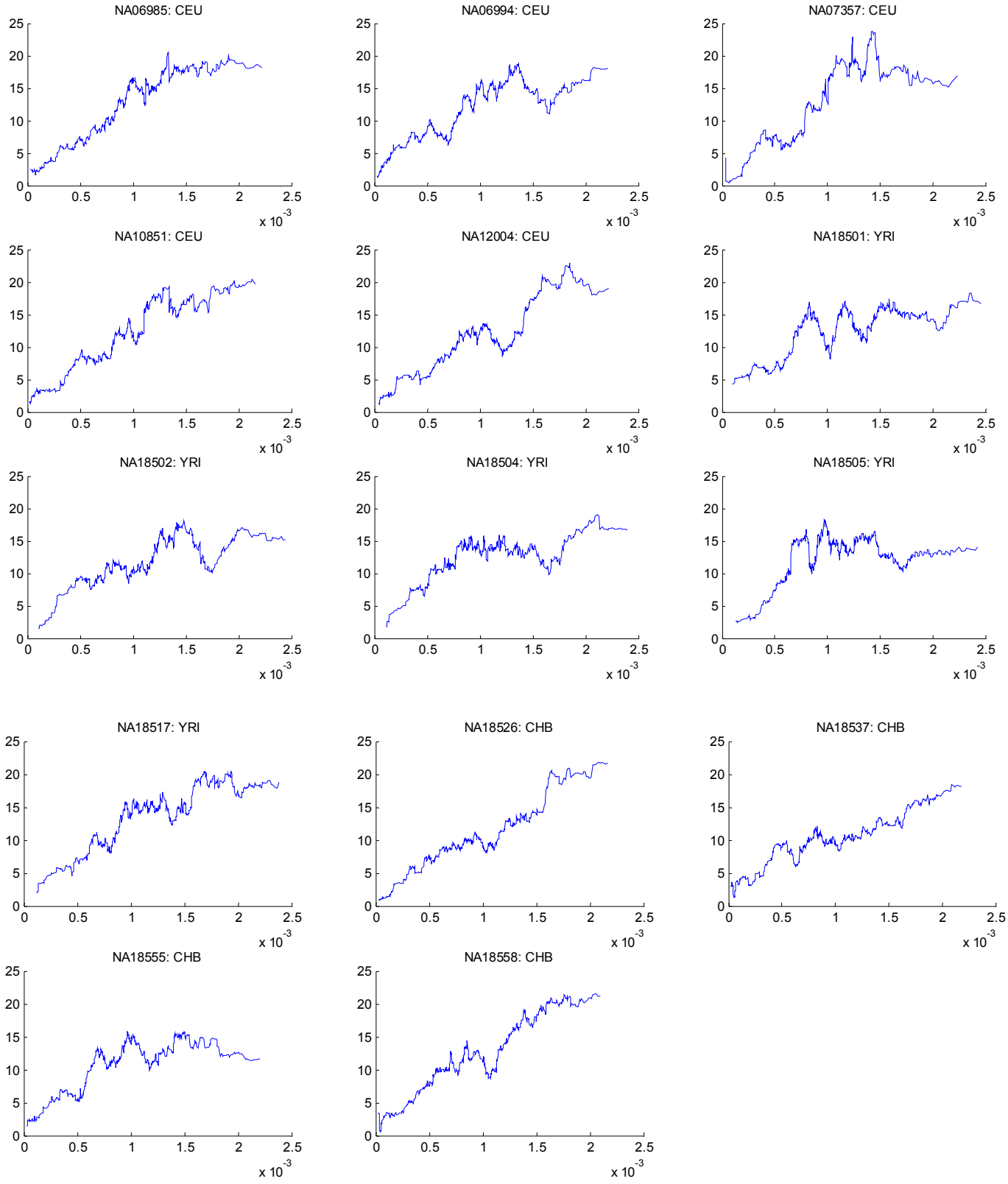

NA06985: CEU
NA06994: CEU
NA07357: CEU
NA10851: CEU
NA12004: CEU
NA18501: YRI
NA18502: YRI
NA18504: YRI
NA18505: YRI
NA18517: YRI
NA18526: CHB
NA18537: CHB
NA18555: CHB
NA18558: CHB
x 10^-3

Figure S10. Sequence divergence versus microsatellite ASD. These plots are similar to that of Fig. 3 but with the x-axis un-rescaled to TMRCA. The combined plot and separate plots for the 23 HapMap individuals are shown. We empirically validate the non-linear behavior predicted by our model by exploiting the fact that there exists considerable variability in sequence heterozygosity (hence TMRCA) across the genome. The x-axis shows the pairwise sequence heterozygosities from sequence data. The y-axis shows the ASD statistic from microsatellite data. In blue are sequence data from Complete Genomics (20 individuals), and in black are data generated using Illumina technology (9 individuals). Microsatellite ASD at each di-nucleotide locus and heterozygosity were computed for each individual and then combined and smoothed using a sliding-window average. We computed the local sequence heterozygosity based on the sequence flanking each microsatellite over a genetic distance window of 0.1 centimorgans in either direction and excluding a 1kb region where the microsatellite itself lies. The result shows a non-linear relationship between microsatellite ASD and sequence heterozygosity which is assumed to increase linearly with time, empirically demonstrating that our model of microsatellite evolution is more appropriate than the GSMM model.

Figure S11. Heterozygosity: CGI versus Illumina

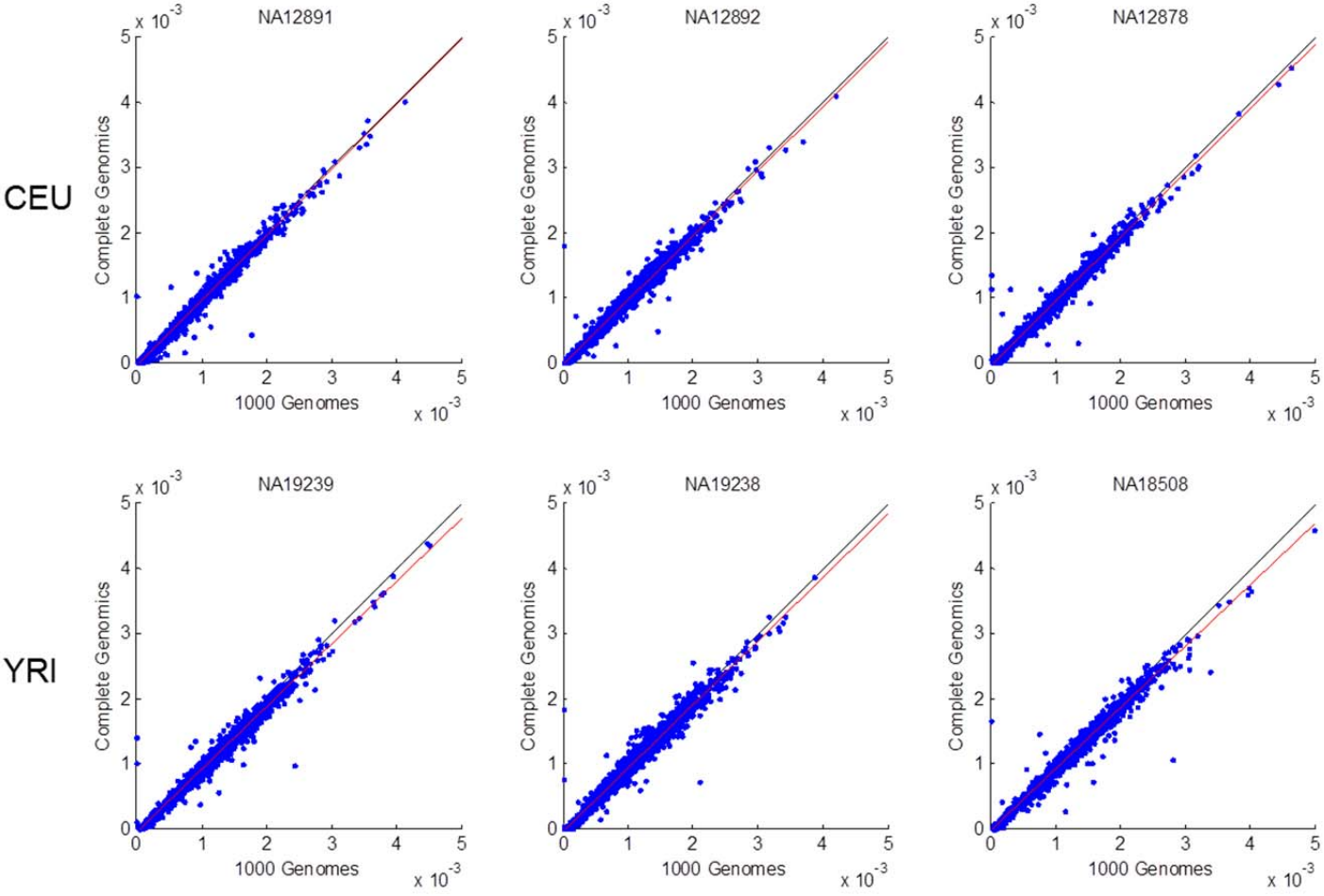


Figure S11. Heterozygosity: CGI versus Illumina. Six individuals have sequence data from both CGI and Illumina. Here we compare heterozygosities. The Illumina heterozygosity is slightly higher than that of CGI.

Figure S12. Inferred sequence mutation rate of 23 individuals

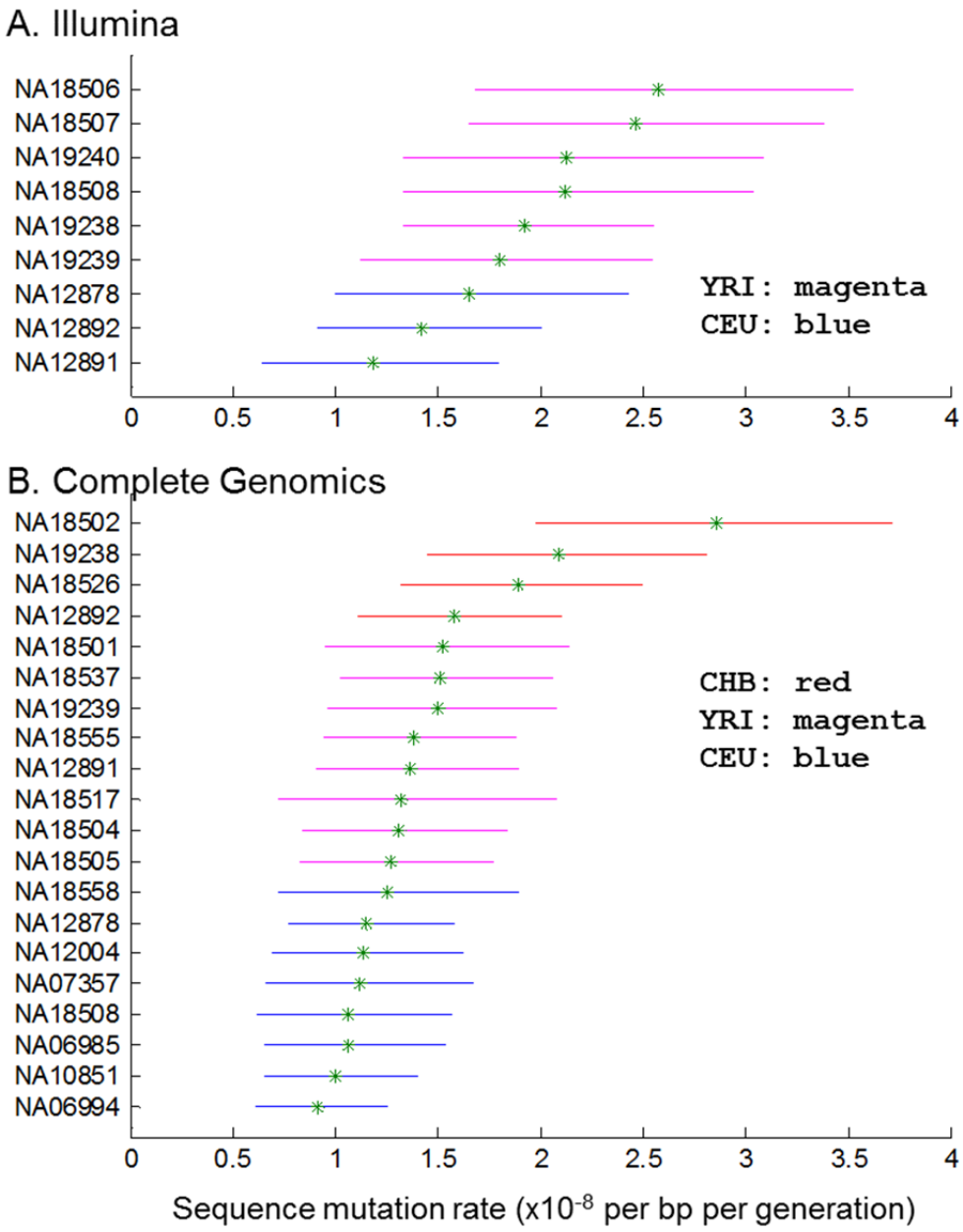


Figure S12. Inferred sequence mutation rate of 23 individuals. This is a graphical representation of Table S6. The asterisk is the mean mutation rate, and the bars are the 90% Bayesian credible intervals. Populations are coded by color. Note that while the individual

mutation rates are not significantly different from each other, the populations do exhibit some clustering, where CEU Europeans have a lower mutation rate than either YRI Africans or CHB Han Chinese. We see two possible explanations for non-random clustering within populations. (1) One possibility is random fluctuation: the differences are not statistically significant, and the clustering within populations could thus simply reflect correlated histories within populations. (2) A second possibility is ascertainment bias for microsatellites with high heterozygosity in Europeans (to make them more useful for disease gene mapping). To understand how this bias could cause underestimation of the mutation rate especially in Europeans, we note that ascertaining for highly polymorphic microsatellites is expected to inflate the measured ASD compared with the expectation based on the true mutation rate, thus overestimating the TMRCA. This in turn results in an underestimate the sequence heterozygosity, since if we infer that more time elapsed in the process of generating the observed mutations, we will estimate a lower mutation rate. Such an ascertainment bias would be expected to be strongest in people of European ancestry as we observe (since they are most closely related to Icelanders), while it would be more mild in more distant populations (CHB and YRI).

Figure S13. Demographic model for coalescent simulation

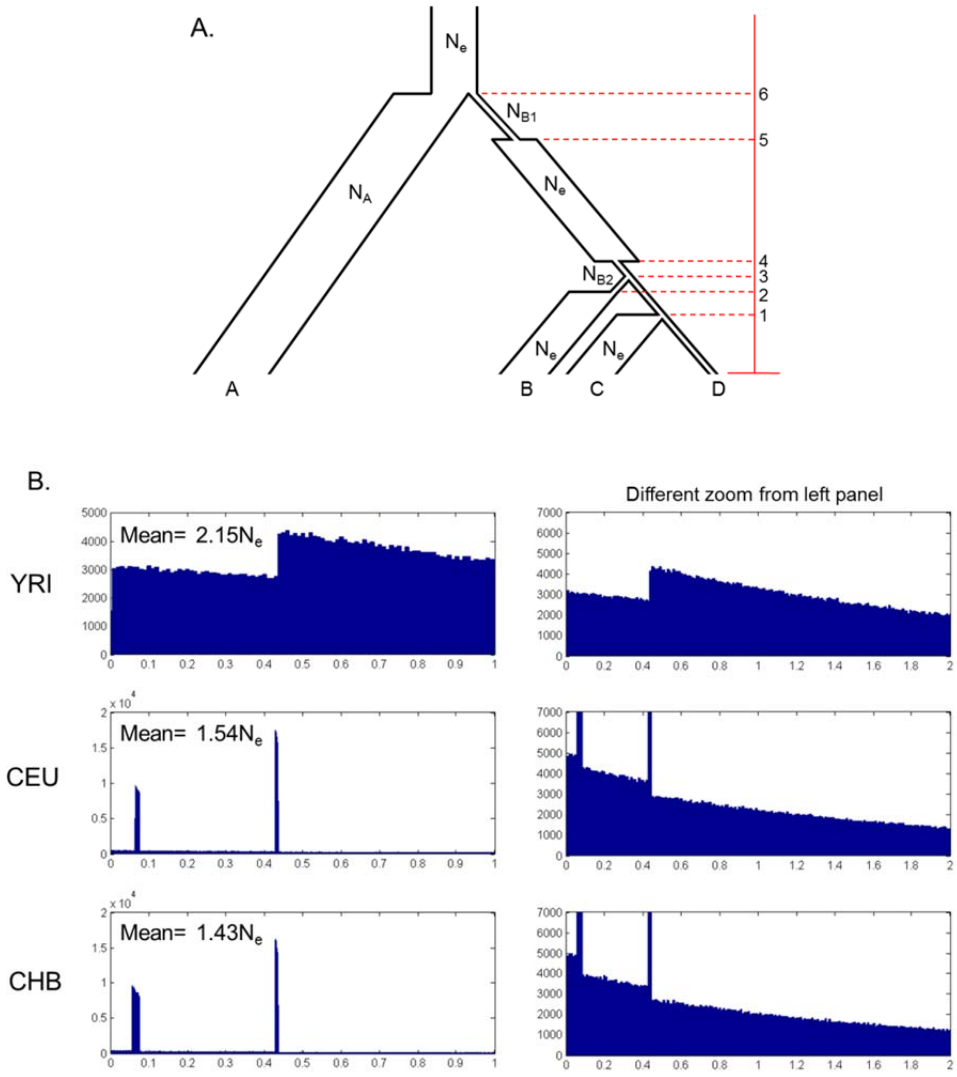


Figure S13. Demographic model for coalescent simulation. In panel A is a model of 4 populations, with A=West Africans (YRI), B=Western Europeans (CEU), C=Han Chinese (CHB), D=Native Americans. This is a 2-bottleneck model suggested by Keinan et al.[19], with N_e=10,000, N_A=1.1N_e, N_{B1}=0.02N_e, N_{B2}=0.05N_e, t_1=0.0147*4N_e, t_2=0.016*4N_e, t_3=0.018*4N_e, t_4=0.019*4N_e, t_5=0.107*4N_e, t_6=0.109*4N_e. Panel B shows the distribution of within-population 2-sample coalescent times, scale by N_e. There are more coalescent events within bottlenecks, as shown by the peaks in the distribution for CEU and CHB. We use this model to verify that our inference of mutation rates is robust to differences in demographic histories across populations.

Figure S14. Distribution of parental age at child-birth

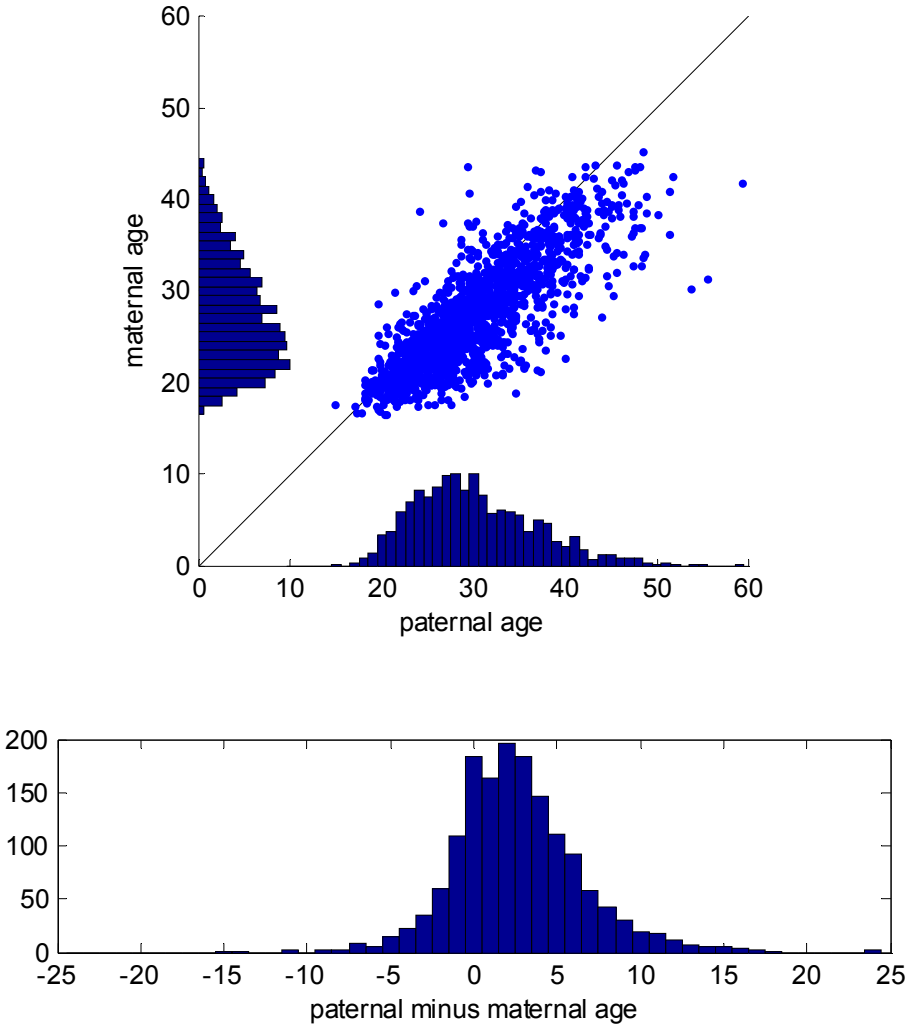
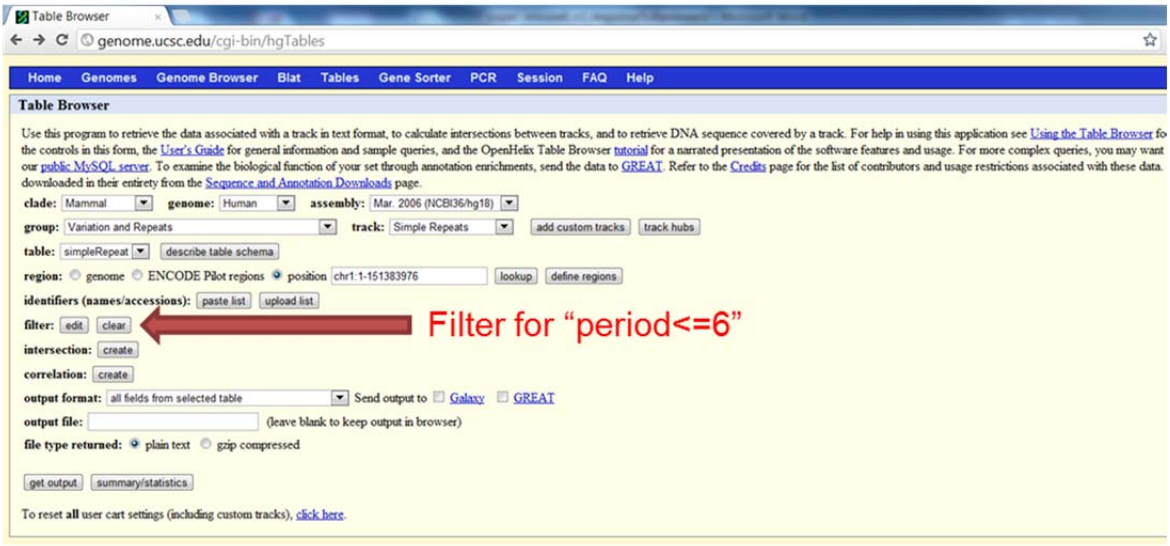


Figure S14. Distribution of parental age at child birth. These are the parental age of trios used in our mutation rate analyses. The paternal age has a mean and standard deviation of 30.1 and 6.5 years, while the maternal age has a mean and standard deviation of 27.4 and 5.9 years. Combining parents, the generation-time has a mean and standard deviation of 28.8 and 6.4 years.

Figure S15. UCSC web query for obtaining microsatellite information

A.



B.

#chrom	chromStart	chromEnd	name	period	copyNum	perMatch	sequence
chr1	0	468	trf	6	77.2	95	TAACCC
chr1	20725	20822	trf	2	47.5	75	TC
chr1	34698	34739	trf	4	10	94	AAAT
chr1	40344	40376	trf	2	16	100	GT
chr1	44575	44680	trf	4	25.8	87	TTTC
chr1	56023	56493	trf	2	262	71	TA
chr1	56067	56495	trf	5	87.4	73	ATATA
chr1	61991	62026	trf	4	8.8	87	ATAC
chr1	73654	73904	trf	4	64.8	86	AAAG
chr1	73726	73844	trf	6	18.2	69	AAAGAA
chr1	88862	88905	trf	4	10.8	100	TTTA
chr1	88909	88979	trf	1	70	76	T

Figure S15. UCSC web query for obtaining microsatellite information. To obtain information for repeat motif (column: "sequence"), repeat length (column: "copyNum"), motif purity (column: "perMatch"), we obtained the output of Tandem Repeat Finder from the UCSC genome browser, with settings shown in panel A, and an excerpt of the output in panel B.

Figure S16. Mutations by locus and by trio

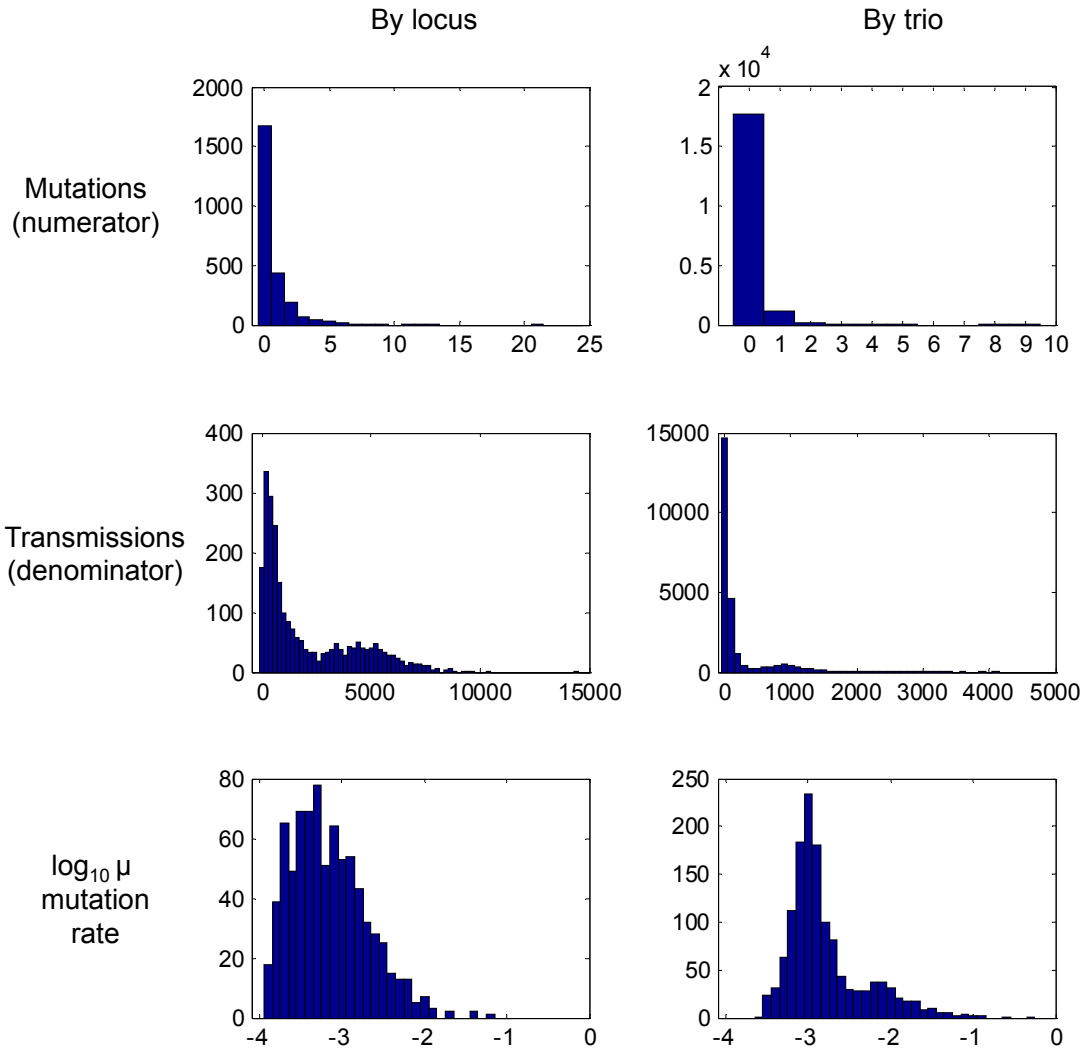


Figure S16. Mutations by locus and by trio. The rows show histograms of mutations, transmissions, and the mutation rate per locus. Of the 2,477 loci, most loci do not contain any mutations. For the loci with at least 1 mutation, the histogram of $\log_{10}$ of the mutation rate resembles a truncated normal distribution, since our denominator is limited to at most about 10,000 per locus. The right column shows the corresponding plots by trio. Of the 24,832 trios, most do not contain a mutation. Due to the sparseness of mutations by locus and by trio, we combine locus and trio data as appropriate to perform our analyses.

Figure S17. Genetic windows for sequence heterozygosity

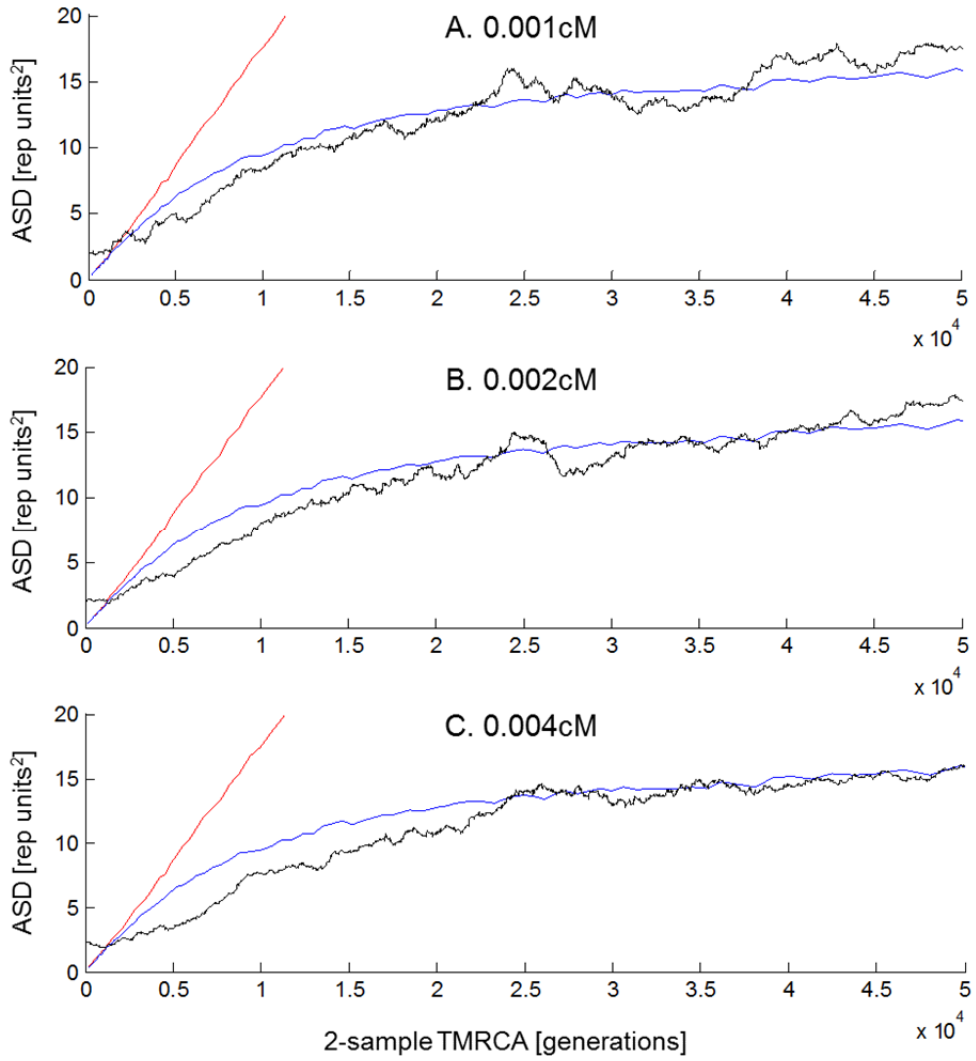


Figure S17. Varying genetic windows for sequence heterozygosity. To extract sequence heterozygosity around each microsatellite, a suitable window length is required. If this window size is too short, sequence heterozygosity becomes imprecise. If the window is too large, crossing multiple recombination events, then the sequence heterozygosity approaches the genome-wide average, rather than local. We tried 3-different window sizes with thresholds at 0.001, 0.002, and 0.004 cM. Shown in black is the empirical curve of microsatellite ASD versus sequence-based 2-sample TMRCA, averaged across the 23 HapMap individuals. The TMRCA is estimated from sequence heterozygosity using a sequence mutation rate of 1.82×10^{-8}, which is

the value we inferred (main manuscript Table 2). The red and blue curves are simulations: in red is the standard random walk (GSMM) model, and in blue is our evolution model. As shown in the figure, all 3 window sizes clearly show a saturation of the ASD curve, closely matching our model. The threshold with 0.001cM is noisier due to less sequence data, however, the fit seems slightly better. Thus, this is the threshold we use, and panel A is the one used for Figure 3 of the main manuscript.

Note S5. Microsatellite evolution modeling to infer TMRCA

I. Overview

Using the mutational characteristics that we observed, we can build a model of microsatellite evolution through time. Given additional parameters summarizing evolutionary history, such as the coalescent time (t_{MRCA}) of modern-day Western Europeans, we can simulate allelic distributions of microsatellites at any genotyped locus. By optimally matching statistics (such as ASD) of the simulated allelic distribution to that of the empirically observed data, we can infer parameters of interest such as t_{MRCA}.

Given any local region of the genome, t_{MRCA} between individuals in that region (assuming no recombinations occurred in the region) must be constant, regardless of whether the genomic features examined are microsatellites or nucleotide substitutions. Therefore, once we have determined t_{MRCA} at each microsatellite locus, we can use that value in conjunction with neighboring sequence divergence to infer parameters such as the sequence mutation rate. Furthermore, given a ratio of human-chimpanzee t_{MRCA} to Western-European divergence, we can use our Western-European t_{MRCA} to estimate the genetic divergence of present-day humans to chimpanzees. A key point is that all inferences here are performed without a calibration to the fossil record.

II. Model design and simulation

At a particular microsatellite locus, a single run consists of simulating a coalescent tree, adding mutations onto the branches of the tree, and finally collecting simulated data at the leaf nodes. By default, the coalescent tree has time in units of generations. When conducting inferences that require time in years, we rescale the branch lengths into years following a generation-time function, as described below.

1. Demography: Generating the coalescent tree

We use the 2-bottleneck model from Keinan et al.[19] (Fig. S13). Coalescent trees are sampled using this demography.

2. Variation of generation-time in history

In modern-day human populations, the average time per generation is about 29 years[16]. However, this number is likely to have been different in the past. To simulate variation in generation-time, we use the logistic curve

$$g(t) = g_{anc} + \frac{g_{now} - g_{anc}}{1 + \exp\left(\frac{t - t_0}{t_0/4}\right)}$$

Where we define

g_{anc}	Generation time of the common ancestor of humans and chimpanzees
g_{now}	Generation time of present-day humans
t_0	Inflection point of an assumed rapid change between g_{anc} and g_{now}

These 3 parameters are stochastic. The shapes of the distribution, means, and variances are given in Table S5. To determine $g(t)$, we first sample these 3 parameters from their distributions.

3. Scale coalescent tree into units of years

The $g(t)$ logistic function is the transformation factor from generations to years. When it is necessary to make inferences in years, we use $g(t)$ to rescale branch lengths as follows: The mean generation-time between a node and its parent is analytically calculated as

$$\bar{g}(t_1, t_2) = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} g(t)dt \; = g_{now} + \frac{g_{now} - g_{anc}}{r_2 - r_1}\log\frac{1 + \exp(r_1 - 4)}{1 + \exp(r_2 - 4)}$$

Where we define

t_1	Time of current node, in units of generations
t_2	Time of parental node, in units of generations
r_1	$4t_1/t_0$
r_2	$4t_2/t_0$

Once $\bar{g}(t_1, t_2)$ is calculated, that particular branch length is trivially scaled into time in units of years.

4. Mutation generation

Mutations are added onto the coalescent tree, sequentially from the root to the leaves. We first generate the baseline mutation rate, which is governed by the mean number of repeats of the microsatellite locus (Fig. 2C). Furthermore, using our empirical observations, we build into our model that the mutation rate changes dynamically as generation-time and allele length change (Fig 2A,C) as we propagate from the root to the leaves of the tree. Finally, as mutations are generated, there is a constraint on allele length (Fig 2D). The details are given below.

(a) The locus-specific baseline mutation rate: For a given locus, we first establish the mutation rate μ_0, which is constant throughout the coalescent tree. This baseline mutation rate is determined using the mean absolute length.

(b) Generation-time effect: In Fig 2A we observed that parental age affects mutation rate. Since generation-time $g(t)$ is modeled as varying as we travel down the coalescent tree, $g(t)$ causes a dynamic change in the mutation rate. In Fig 2A we demonstrated a difference in the paternal and maternal behavior, and we therefore first split generation-time into paternal time $g_{pat}(t)$ and maternal time $g_{mat}(t)$:

$$g_{pat}(t) = g(t) + 0.5 \cdot \Delta(t)$$
$$g_{mat}(t) = g(t) - 0.5 \cdot \Delta(t)$$

$\Delta(t)$ is the mean difference between paternal and maternal age, at time *t*. Note that this is a time-varying quantity too, as Δ of present-day humans could be different from that of the human-chimp common ancestor. In particular, we model $\Delta(t)$ as entirely analogous to the logistic function of $g(t)$.

$$\Delta(t) = \Delta_{anc} + \frac{\Delta_{now} - \Delta_{anc}}{1 + \exp\left(\frac{t - t_0}{t_0/4}\right)}$$

Δ_{now} and Δ_{anc} are sampled values. (See Table S5 for the distributions, means, and variances used.) t_0 uses the same value sampled from $g(t)$ and hence is not a new sample. Once $g_{pat}(t)$ and $g_{mat}(t)$ are determined, we can obtain the gender-specific mutation rates and the gender-averaged mutation rate:

$$\mu_{pat}(t) = \beta_{0,pat} + \beta_{1,pat} \cdot g_{pat}(t)$$
$$\mu_{mat}(t) = \beta_{0,mat} + \beta_{1,mat} \cdot g_{mat}(t)$$
$$\mu_g(t) = \left(\mu_{pat}(t) + \mu_{mat}(t)\right)/2$$

Where we define

$\beta_{0,pat}$, $\beta_{0,mat}$ The intercepts of regressions in Fig 2A

$\beta_{1,pat}$, $\beta_{1,mat}$ The slopes of regressions in Fig 2A

To take into account the stochasticity of the slopes and intercepts, these quantities are sampled from the data, using a Bayesian analysis of simple linear regression (or equivalently, a draw from the multivariate student-*t* distribution).

We can summarize $\mu_g(t)$ using the matrix notation below:

$$\mu_g(t) = \frac{1}{2}[1 \quad 1] \cdot \left(\begin{bmatrix} \beta_{1,pat} & 0 \\ 0 & \beta_{1,mat} \end{bmatrix} \begin{bmatrix} 1 & 1/2 \\ 1 & -1/2 \end{bmatrix} \begin{bmatrix} g_{anc} & g_{now} \\ \Delta_{anc} & \Delta_{now} \end{bmatrix} \begin{bmatrix} 1-f(t) \\ f(t) \end{bmatrix} + \begin{bmatrix} \beta_{0,pat} \\ \beta_{0,mat} \end{bmatrix} \right)$$

Where $f(t) = \frac{1}{1+\exp\left(\frac{t-t_0}{t_{0/4}}\right)}$

We highlight two special cases:

i. If mutations are entirely generation-like, i.e. β_1 for both parents are 0, then the expression simplifies to $\mu_g(t) = (\beta_{0,pat} + \beta_{0,mat})/2$. Thus, as expected in this case, the mutation rate does not vary as a function of generation interval.

ii. If mutations are entirely year-like, i.e. β_0 for both parents are 0 and $\beta_{1,pat} = \beta_{1,mat}$, then the expression simplifies to $\mu_g(t) = \beta_1 \cdot g(t)$. Hence the mutation rate per generation perfectly correlates with generation-time. However, the mutation rate per year, $\mu_g(t)/g(t)$, becomes a constant.

(c) Generating the instantaneous mutation rate: At any point along the coalescent tree, the instantaneous mutation rate is a function of the baseline rate, generation-time, and allele length. We combine these three factors to generate the mutation rate $\mu(t)$:

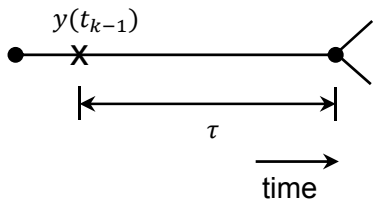
$$\mu(t) = (m \cdot y(t) + \mu_0) \cdot \frac{\mu_g(t)}{\mu_g(0)}$$

Where we define

$y(t)$	The allelic length of the branch at time *t*
m	The slope in Fig 2C that relates allelic length to mutation rate
μ_0	The baseline mutation rate described in part (a)
$\mu_g(t)$	The mutation rate as a function of generation time, as described in (b)
$\mu_g(0)$	The present-day mutation rate, as determined by the $\mu_g(t)$

Note that this mutation rate model simplifies to that of the generalized stepwise mutation model (GSMM) if $m = 0$ and $\mu_g(t) = \mu_g(0)$.

(d) Generating mutation events: Suppose we are on a branch (shown below) where the *(k-1)*-th mutation occurred at t_{k-1}, which is marked by the "X". The allele length immediately following that event is $y(t_{k-1})$ and the generation-time is $g(t_{k-1})$. Mutation events are simulated forward in time, from the root of the tree, using an exponential distribution with mean $\mu(t_{k-1})$, which is determined from the equation in part (c). After a random sample $T \sim Exp(\mu(t_{k-1}))$ is drawn, if $T < \tau$, generate a mutation with length $Y(t_k)$ and update τ to be $\tau - T$. Otherwise, there are no more mutations in the branch and move on to the next branch. Details for generating $y(t_k)$ are described in the next section.



The process for generating mutation events for a coalescent tree re-scaled into units of years is very similar, except that the mutation rate at any point in time is divided by the generation-time, e.g. we set the mutation rate per year to be $\mu(t_{k-1})/g(t_{k-1})$.

(e) Generating microsatellite lengths for each mutation event: In the GSMM, the microsatellite length $y(t_k)$ is the parental length plus the mutational length, which is an independent random sample from the mutation length distribution, defined as x for the *k*-th mutation event. However, using our empirical observations (Fig 2D, S7), we model the fact that longer microsatellites tend to mutate to a shorter length, and vice versa, as a linear function:

$$y(t_k) = y(t_{k-1}) + x(t_k) + \frac{y(t_{k-1})}{\sigma} m$$
$$= \left(1 + \frac{m}{\sigma}\right) y(t_{k-1}) + x(t_k)$$

Where we define

$x(t_k)$	The mutation length, drawn randomly from the mutation length distribution in Fig 2B
$y(t_{k-1})$	The microsatellite allele length, just prior to the mutation
$y(t_k)$	The microsatellite allele length, just after the mutation
m	The slope in Fig S7A. This quantity is negative, generating the length constraint.
σ	The standard deviation of the allelic distribution of the locus, based on empirical data

Observations:

- Note that while σ is locus specific, m was obtained from the combined mutational data of all loci.
- If $m = 0$, this equation reduces to the GSMM.

- At the root of the coalescent tree, we begin with allele length of x_0, which is determined from the empirical allele length distribution. However, we set $y(t_{root}) = 0$ when propagating mutations. When collecting allele lengths at the leaf nodes, x_0 is added back in.
- $y(t_{k-1})/\sigma$ produces a Z-score (horizontal axis of Fig S7A) showing the degree of deviation from the mean length, and through multiplication with slope m, gives the strength of the return-to-mean length constraint.

Note S6. Testing the microsatellite evolution model

Overview

To test our procedure for using the microsatellite mutation model to estimate evolutionary parameters, we use two approaches. First, we show that our inferences based on the model produce unbiased sequence mutation rate estimates. To do this, we simulate microsatellite alleles and sequence heterozygosity using a 2-bottleneck demographic model (Fig S13), with a known sequence mutation rate and effective population size. Then, with the simulated sequence and microsatellite data, we infer the sequence mutation rate and compare it to the truth.

Second, we show that the model is robust to each parameter's prior probability distribution: we use different parameter values for our prior and show that our inferences of the sequence mutation rate and human-chimpanzee speciation time are not greatly affected (Fig. S9).

I. Simulated data shows that the model is unbiased

Procedure:

1. Choose a sequence mutation rate to use in simulation: [1.0, 1.5, 2.0, 2.5, 3.0] $x10^{-8}$ per bp per generation. Use N_e of 12,500 for the 2-bottleneck demography model (Fig S13). Generate a set of global parameters (Table S5).

2. Based on the demographic model and mutation rate chosen for the simulation, generate the local TMRCA for each individual at each locus, followed by the local sequence heterozygosity and microsatellite ASD. Generate the local sequence heterozygosity using a Poisson process, and the local microsatellite ASD using our model of evolution.

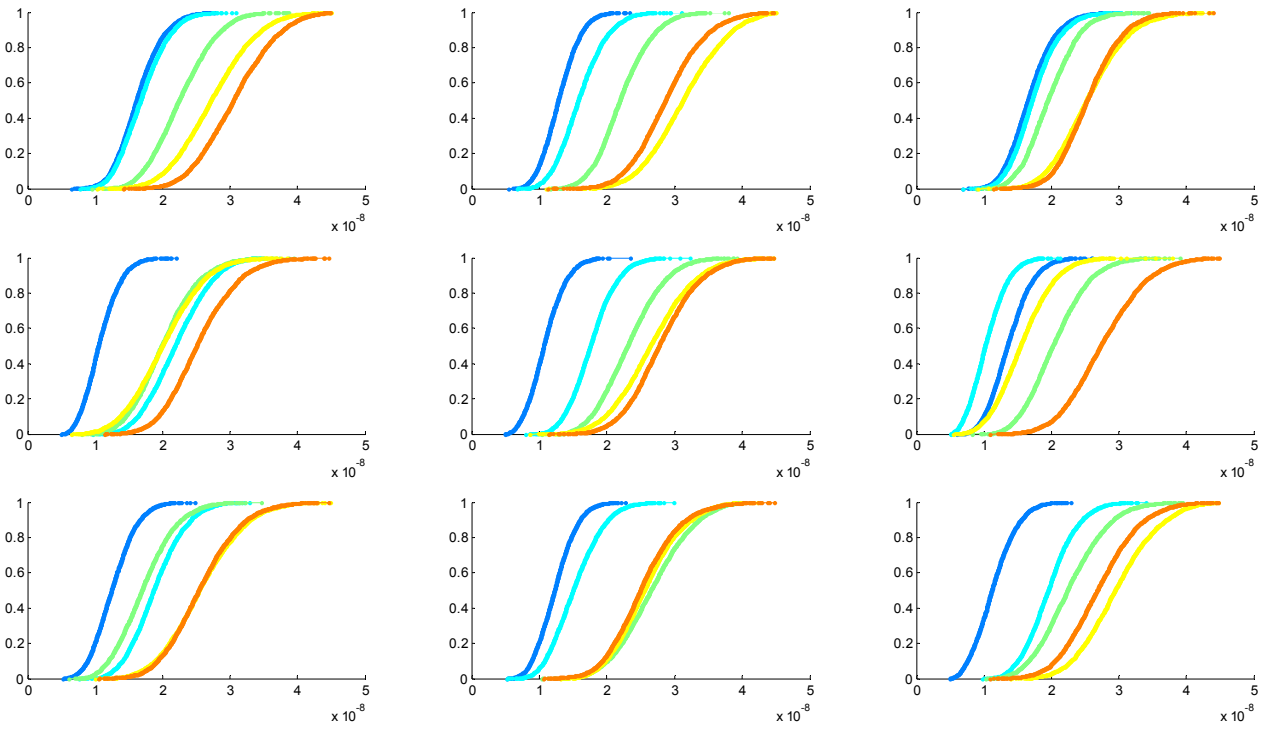
3. Run the Markov Chain Monte Carlo inference to obtain a posterior sequence mutation rate estimate for each individual, without any knowledge of the values from Step 1 used in generating the data (we also do not use knowledge about the values of the global parameters used in the simulations).

4. Obtain inferences for 9 individuals, for each of 5 mutation rates, resulting in 45 posterior distributions for sequence mutation rate. With these results, we can report the fraction of simulations in which the true TMRCA falls in the 90% Bayesian credible interval.

Results:

The CDFs (cumulative distribution function) of posterior sequence mutation rate are shown below, one panel per individual. There are 5 curves for each individual, each corresponding to a different true mutation rate: [Blue=1.0, Cyan=1.5, Green=2.0, Yellow=2.5, Red=3.0] $x10^{-8}$. The table summarizes the results by the percentile (of the posterior distribution) in which the true mutation rate lies. Only in 3 of 45 cases (6.7%) does the true mutation rates fall outside the 90% Bayesian credible interval.

	True sequence mutation rate				
	1.0E-08	**1.5E-08**	**2.0E-08**	**2.5E-08**	**3.0E-08**
Person 1	0.018	0.317	0.297	0.349	0.462
Person 2	0.137	0.412	0.302	0.123	0.607
Person 3	0.011	0.247	0.553	0.485	0.846
Person 4	0.427	0.055	0.514	0.826	0.815
Person 5	0.399	0.214	0.253	0.398	0.670
Person 6	0.107	0.944	0.485	0.983	0.675
Person 7	0.208	0.166	0.759	0.470	0.802
Person 8	0.211	0.502	0.101	0.461	0.838
Person 9	0.347	0.102	0.312	0.199	0.727



II. The model is robust to changes in the parameter prior distributions

Each parameter in our evolution model has a prior distribution governing its uncertainty. We therefore explored how changing the value of the parameter—within the plausible range given by the prior—influences our inferences about the sequence mutation rate and human-chimpanzee speciation time.

To test for robustness of our priors, for each of 8 parameters (Fig. S9), instead of using the default prior distribution, we set them to point values at three different points: the lower 95% CI, the mean, and the upper 95% CI. Then, this altered set of parameters was fed through our inference process. The primary purpose of this exercise was to see whether an extreme value of the prior, if used, would cause our inferences to change greatly. Reasonable extreme values are at the boundary of our prior distribution specifications. The second purpose is to see whether

shrinkage in the variance (to zero) of any prior would cause a significant shrinkage in the variance of the posterior estimates. Note that we only perturb one parameter at a time.

As shown in Fig S9, using our model of evolution, our inference of sequence mutation rate and human-chimpanzee speciation date is reasonably robust to changes in the prior, both in the mean and in the standard error of the inferred distributions. We observe the following:

- Aside from the length constraint parameter, when we use extreme values, the inference on the sequence mutation rate does not change significantly. This suggests that (1) our priors are reasonably tight such that no significant changes are observed, or (2) the model is not heavily dependent on that parameter. For example, case (1) holds for the microsatellite mutation rate parameter: although the microsatellite mutation rate can in principle affect our inferences greatly since it has a linear effect on ASD, it is determined with high precision by our direct observations of mutations, with a 95% CI of 2.56-2.91 $x10^{-4}$; thus, the extreme values of this prior do not affect our inferences substantially.

- The length constraint governs the non-linearity mapping between TMRCA and ASD (Fig. 3), and changes to it (Fig. S9) can cause large changes to our inferences on the sequence mutation rate. Our prior distribution for this parameter was determined entirely based on the direct observation of mutations (Fig. S7, Table S5), and not on comparisons between microsatellite ASD and sequence heterozygosity (Fig. 3, S10). As a result, the length constraint prior was not determined to a high level of precision. This is in fact desirable, because in the inference machinery, we use the empirical data of Fig S10 (comparison to flanking sequence data) to further infer the length constraint parameter, rather than being extremely precise about the prior. The result from Fig. S9 is as expected: If we give the length constraint parameter the default prior, the resulting sequence mutation rate distribution is not different from the green spike prior, and this is because the data of Fig S10 down-weighted any sampling of the default prior away from its mean. On the other hand, if we actually forced an unreasonable prior, such as the red or blue spikes, the data of Fig S10 could not influence the length constraint in any way, and since this is such an important parameter in our model, the resulting inferences are inaccurate.

Note S7. Constraints on sequence mutation rate from calibration to the fossil record

(i) Overview

We were interested in obtaining constraints on the sequence substitution rate based on calibration to the fossil record, to which we could compare our absolute estimate based on direct measurement of the mutation rate at microsatellites.

(ii) Assumptions

For the analyses in this note, we make a number of simplifying assumptions:

• d_{HC}, the divergence per base pair between human and chimpanzee, is 0.0130. This number is derived from the Enredo-Pecan-Ortheus (EPO) 6-way primate whole genome alignments[20].

• d_{HO}/d_{HC} the divergence per base pair between human and orangutan divided by that between human and chimpanzee at aligned bases is 2.65, as argued in the main text.

• τ_{HC}, human-chimpanzee speciation time, is >4.2 Mya, based on the date of the *Australopithecus amanensis* fossil which is believed to be on the hominin lineage since the split from chimpanzee[21].

• τ_{HC}/t_{HC}, the ratio of human-chimpanzee time of last gene flow to human-chimpanzee average autosomal divergence time, is <0.73. This bound (also discussed in the text) is based on human-chimpanzee genetic divergence near genes on chromosome X, close to sites where humans and

chimpanzees share an allele not seen in gorilla, orangutan and macaque. Here, the ratio τ_{HC}/t_{HC} is 0.73. Thus, the time of most recent gene flow between humans and chimpanzees is <0.73.

• t_{HO}, human-orangutan genetic divergence time is <23 Mya. This is based on a view that the *Proconsul* fossil places an upper bound on human-orangutan speciation time of $\tau_{HO} < 18\ Mya$[6,22]. We assume that $t_{HO} - \tau_{HO} < 5\ Mya$, that is, the human-orangutan average autosomal genetic divergence time is at most 5 Mya older than human-orangutan speciation time.

• The mutation rate per year has been constant since human-orangutan genetic divergence. (For the upper bound on the mutation rate, we only require the assumption that it has been constant since human-chimpanzee genetic divergence).

• The present-day human generation time has a lower bound 25.6 years per generation and an upper bound of 32.4 years per generation. This range is derived from our prior distribution of present-day generation time of 29 ± 2.04 from Table S5, and using the 90% confidence interval.

(iii) Upper bound on mutation rate: <3.7×10^{-8} /bp/gen. from *Australopithecus anamensis*

τ_{HC}>4.2 Mya (since *Australopithecus anamensis* is a hominin)

$\Rightarrow$ t_{HC} > 5.8 Mya (since τ_{HC}/t_{HC} <0.73)

$\Rightarrow$ $\mu_{year}^{seq} < 1.1 \times 10^{-9}$ (since $\mu_{year}^{seq} = d_{HC}/2t_{HC}$ = 0.0130/(2×5.8×10^6)

$\Rightarrow$ $\mu_{generation}^{seq} < 3.7 \times 10^{-8}$ (since $\mu_{generation}^{seq} < 32.3\mu_{year}^{seq}$)

(iv) Lower bound on mutation rate: >1.9×10^{-8} /bp/generation from *Proconsul*

τ_{HO}<18 Mya (from *Proconsul*)

$\Rightarrow$ t_{HO} < 23 Mya (since we assume that $t_{HC} < \tau_{HC} + 5$ Mya)

$\Rightarrow$ $\mu_{year}^{seq} > 7.5 \times 10^{-10}$ (since $\mu_{year}^{seq} = d_{HC}(\frac{d_{HO}}{d_{HC}})/2t_{HO}$ = 0.0130(2.65)/(2×23×10^6))

$\Rightarrow$ $\mu_{generation}^{seq} > 1.9 \times 10^{-8}$ (since $\mu_{generation}^{seq} > 25.6\mu_{year}^{seq}$)

The most likely way that this lower bound could be in error would be if the mutation rate were not constant over time since human-orangutan genetic divergence. For example, if the mutation rate slowed down on the African great ape lineage (and perhaps also on the orangutan lineage) since the two diverged—perhaps associated with the increase in their body size as documented in the fossil record—the lower bound would be substantially less.

(v) Upper bound on human-chimpanzee speciation date from fossil record <6.3 Mya

For comparison to the upper bound on human-speciation obtained by direct calibration to the microsatellite-based molecular clock, we also use the fossil record of human-orangutan divergence to produce a complementary bound based on the fossil record. As in (iv), we write:

τ_{HO}<18 Mya	(from *Proconsul*)
$\Rightarrow$ t_{HO} < 23 Mya	(since we assume that $t_{HO} < \tau_{HO} + 5$ Mya)
$\Rightarrow$ t_{HC} < 8.7 Mya	(since $t_{HC} = t_{HO}/(\frac{d_{HO}}{d_{HC}})$ = (23 Mya)/2.65)
$\Rightarrow$ τ_{HC}< 6.3 Mya	(since $\tau_{HC} = t_{HC}(\tau_{HC}/t_{HC})$, and $\tau_{HC}/t_{HC} << 0.73$, Note S8)

As in (iv), the most plausible way that this lower bound could be in error would be if the mutation rate were not constant over time since human-orangutan genetic divergence.

Note S8. Constraints on human-chimpanzee speciation date

(i) Motivation for estimating the ratio of human-chimpanzee speciation to divergence

Our calibration of the molecular clock allows us to estimate the genetic divergence time of humans and chimpanzees $\bar{t}_{HC}$, averaged across the autosomes. However, the speciation date τ_{HC}—defined in this study as the date of last gene flow between the ancestors of humans and chimpanzees—is also of biological interest. To infer τ_{HC}, we require a Bayesian prior distribution on the ratio of these two quantities: $\tau_{HC}/\bar{t}_{HC}$. This is the most difficult of our prior distributions to formulate, and the following note describes how we construct our distribution based on obtaining a number of point estimates of the ratio, as well as conservative upper bounds.

(ii) A point estimate of $\tau_{HC}/\bar{t}_{HC}$ = 0.61 from modeling of a simple demographic history

Burgess and Yang 2008

For a best estimate of the ratio $\tau_{HC}/\bar{t}_{HC}$, we use the results from Burgess and Yang 2008, who analyzed a data set of 7.4 Mb of aligned sequence from human, chimpanzee, gorilla, orangutan and macaque across "neutral" autosomal loci using the MCMCcoal software[9]. This software analyzes the 5-species alignment data under the simplifying assumptions that:

(i) The phylogeny is ((((human, chimpanzee),gorilla),orangutan),macaque)
(ii) The speciation events were instantaneous.
(iii) The populations in the intervening periods were constant in size and panmictic.
(iv) All the analyzed loci are unlinked, neutral and free of recombination

Under these assumptions, MCMCcoal estimates the ancestral population sizes and speciation times, conditional on the observed divergent site pattern. On page 7 of Burgess and Yang 2008, the authors estimate that the fraction of human-chimpanzee coalescences that occurred prior to human-chimpanzee speciation is $1 - \tau_{HC}/\bar{t}_{HC}$ = 0.39 (thus, $\tau_{HC}/\bar{t}_{HC}$ = 0.61) under a model of no gene flow after initial speciation.

Dutheil et al. 2009

Dutheil et al. 2009 made inferences under the same demographic assumptions, but using a different approach based on a coalescent Hidden Markov Model (CoalHMM) that also exploits information from recombination between adjacent loci[23]. We inferred $\tau_{HC}/\overline{t}_{HC}$ for the four autosomal loci ("targets") that Dutheil et al. analyzed, using their "bias-corrected" estimates of demographic parameters in their Table 2. After translating the quantities to estimates of $\tau_{HC}/\overline{t}_{HC}$, we obtained results in the range of Burgess and Yang 2008: 0.67 (Target 1), 0.57 (Target 106), 0.60 (Target 121) and 0.66 (Target 122). We use the Burgess and Yang 2008 estimate of $\tau_{HC}/\overline{t}_{HC}$ = 0.61 for our primary calculations because it is based on more data and because it falls within the range of the Dutheil et al. estimates.

(iii) Conservative upper bound on the ratio: $\tau_{HC}/\overline{t}_{HC}$ < 0.73

Analyzing subsets of the genome to obtain a conservative upper bound on $\tau_{HC}/\overline{t}_{HC}$

The published studies infer demographic parameters for human-chimpanzee speciation under a simplified model that assumes constant population size, sudden speciation, and no impact of natural selection on the genome. However, the truth likely differs from this model, as Yang found in 2010 when he carried out a formal test of the fit of the data from Burgess and Yang 2008 to the model assumed in that study[24]. Thus, while the simplified models provide a useful initial estimate, deviations from the assumptions might mean that the time of last gene flow between humans and chimpanzee was more ancient or more recent.

To obtain a conservative upper bound on the ratio $\tau_{HC}/\overline{t}_{HC}$, we take advantage of an idea of Patterson et al. 2006[6]. The idea is to compute human-chimpanzee genetic divergence (dividing by human-macaque divergence to correct for variation in the local mutation rate across the genome) in subsets of the genome where the genetic divergence is expected to be less than the genome-wide average for population genetic reasons. Human-chimpanzee genetic divergence at all loci in the genome must be older than the speciation time (by definition, if we define speciation as the time of last gene flow). Thus, the ratio of the local divergence at any subset of the genome to the genome-wide average provides an upper bound on the speciation date τ_{HC}.

A new 5-way alignment of human-chimpanzee-gorilla-orangutan-macaque (HCGOM)

Overview of a 100x larger dataset generated for studying human-chimpanzee-gorilla speciation

Patterson et al. 2006 analyzed datasets consisting of about 9 Mb of aligned DNA rom human, chimpanzee, gorilla, orangutan and macaque[6]. Here we describe how we generated a similar dataset with about 100x more data. In brief, we restricted to data generated using traditional Sanger long-read sequencing data from five genomes, and used an alignment and filtering procedure described in Mallick et al. 2009[25] (the detailed filters we applied are given below). In comparison to other multi-species alignments methodologies (e.g. EPO[20]), which have as a goal the maximization of the number of covered nucleotides, our alignment procedure filters out a larger fraction of the data, since for the purpose of making inferences about population history, we do not mind losing data as long as what is left is of high reliability. These filters resulted in 849.6 Mb of 5-species genomic alignment on the autosomes (48.58 million bi-allelic divergent sites passing filters), and 32.6 Mb on chromosome X (1.62 million bi-allelic divergent sites passing filters). These datasets are available on request from the authors.

Genome assemblies used as input

The raw data consisted of 5 whole genome assemblies based on Sanger long-read sequencing data. These consisted of the human genome reference sequence (*hg18*), and four assisted assemblies that we built ourselves so as to have full control over the data: chimpanzee (7.3× coverage), orangutan (6.2× coverage), macaque (6.3× coverage) and gorilla (1.8× coverage). Since we assembled the genomes ourselves, we had a sequence quality score at each nucleotide that did not automatically assign low quality to bases overlapping at within-species single nucleotide polymorphisms (SNPs), which is a feature of some genome assemblies that makes it difficult to carry out population genetic analyses.

Generating local alignments

We applied a stringent local alignment procedure that took advantage of the long range synteny information available from the genome assemblies[25], and then applied the following filters:

- Restrict to loci that have alignments of all 5 species over at least 100 bp
- Restrict to loci for which a unique consensus sequence is available from all 5 species

Identifying divergent sites for analysis

We identified sites that were divergent across the species after applying the following filters:

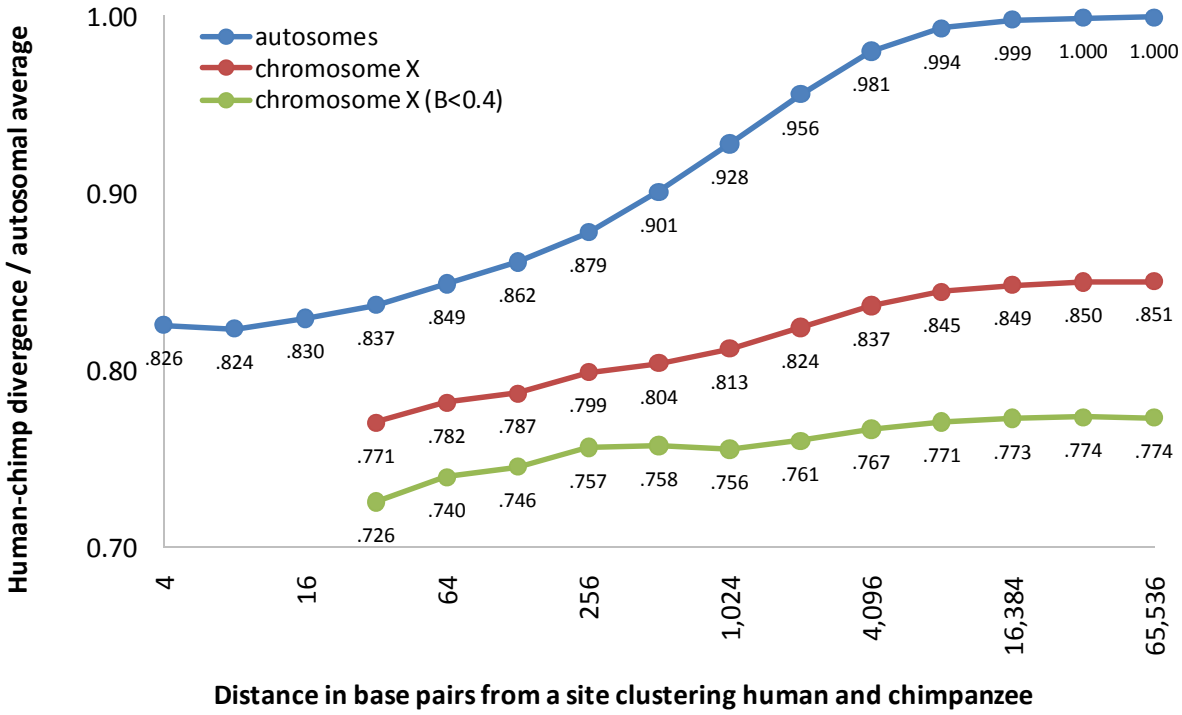
- Filter out sites with 3 or more alleles across species

- Filter out sites where any species has a Phred sequence quality score of <30
- Filter out sites where any species has a Phred score of <15 within 5 bp on either side.
- Filter out sites within 1 bp of an insertion/deletion in any of the species.
- Filter out sites within 5 bp of the end of an alignment
- Filter out sites within 1 bp of any other divergent site, as these sites have consistently different properties indicating that they are determined less reliably
- Filter out divergent sites that could potentially reflect a C→T mutation in the first base of a hyper-mutable CpG dinucleotide on either DNA strand (these are subject to high rates of recurrent mutation, which could complicate tests of relative divergence time).

Post-processing to remove potential misalignments

We filtered out entire alignments where the pattern of divergent sites showed evidence of an extreme excess on a single lineage compared with genome-wide pattern, which could reflect erroneous alignment due to low copy number repeats (paralogs). For 7 species pairs—Human-chimpanzee, Human-gorilla, Chimp-gorilla, Human-orang, Chimp-orang, Orang-macaque—we counted the number of divergent sites reflecting changes on one lineage or the other, using the other species to polarize. We compared the ratio of sites on the tested lineage to the average genome-wide (performing the analysis separately for chromosome X and the autosomes), and removed alignments with P <0.001 by a chi-square test for any of the seven comparisons

Figure S8.1: Bounds on human-chimp speciation based on proximity to sites clustering humans and chimps. (Blue curve) We stratify the autosomal data based on the distance to the closest site clustering humans and chimps to the exclusion of gorilla. Within 4bp, the divergence is 0.826 of the autosomal average. (Red curve) Repeating the same computation on chromosome X, the average divergence as a fraction of the autosomes is 0.851, and within 32 bp of a human-chimp clustering site is 0.771. (Green curve) We again present data for the X chromosome, but now restrict to the quarter of the data with B-statistic <0.4 reflecting an expectation of further reduced divergence due to directional selection in the ancestral population. The average X chromosome divergence in this subset of the data is 0.774, and within 32 bp of human-chimp clustering sites, it is 0.726.



Bound B: Genetic divergence on chromosome X divided by the autosomes ($\tau_{HC}/\overline{t}_{HC} < 0.851$)

The second upper bound on ratio of human-chimpanzee speciation time also exploits a strategy first described in Patterson et al. 2006, and is based on dividing the human-chimpanzee genetic divergence as a fraction of human-macaque on chromosome X by that on the autosomes. The motivation is that there is an *a priori* reason to expect that genetic divergence on chromosome X will be lower than on the autosomes. In a constant-sized, freely mixing population, there are 3 copies of chromosome X for every 4 copies of the autosomes, leading to a lower predicted coalescence time at X chromosome loci in the common ancestral population of humans and chimpanzees. In addition, selection operates differently on chromosome X and the autosomes (because of the exposure of recessive alleles in males), further motivating a search to explore whether the genetic divergence is unusually low.

In our new dataset, we computed the ratio of human-chimpanzee to human-macaque divergence on chromosome X divided by that on the autosomes, filtering out the pseudo-autosomal regions of chromosome X (<2.710 Mb and >154.585 Mb). After applying the correction for recurrent mutation (nearly identical results are obtained without the correction), we obtained an upper

bound of $\tau_{HC}/t_{HC} < 0.851$. This is one standard error from the estimate of $\tau_{HC}/t_{HC} < 0.835 \pm 0.016$ from Patterson et al. 2006, and so the two inferences are statistically consistent.

Bound C: Chromosome X loci close to sites clustering humans and chimps ($\tau_{HC}/\overline{t}_{HC} < 0.771$)

We combined the two ideas from Patterson et al. 2006 (bounds A and B) to obtain an even more stringent upper bound. Using our 32.6 Mb of X chromosome alignment, we computed the ratio of human-chimpanzee to human-macaque divergence close to sites that cluster humans and chimpanzees to the exclusion of gorilla. Figure S8.1 (blue curve) shows that just as on the autosomes, the closer one is to a human-chimpanzee clustering site, the lower the normalized human-chimpanzee divergence. We compute the human-chimpanzee divergence divided by human-macaque divergence in the vicinity of these sites, and divide by the autosomal average after correction for recurrent mutation, resulting in a bound of $\tau_{HC}/t_{HC} < 0.771$ based on data from <32 bp away from informative sites. (We focus on the <32 bp distance because of noisy estimates in lower bin sizes, although the estimates are qualitatively consistent for smaller bin sizes as well: 0.773 (<16 bp), 0.752 (<8 bp) and 0.725 (<4 bp).)

Bound D: Chr X loci subject to directional selection close to HC sites ($\tau_{HC}/\overline{t}_{HC} < 0.726$)

We next studied genetic divergence between humans and chimpanzees at a subset of the genome that was not exploited in Patterson et al. 2006: loci that are at increased likelihood of having been subject to directional selection in the ancestral population of humans and chimpanzees (due to hitchhiking and selection at linked sites), thus reducing the average genetic divergence between the two species. McVicker et al. 2009 showed that loci that are close to exons or conserved non-coding sequences have a reduced genetic divergence between humans and chimpanzees compared with the average in the genome, which is likely to reflect directional selection in the ancestral population (either positive selective sweeps or negative background selection)[10]. For each nucleotide, they also computed a quantity, B, which predicts the genetic divergence without using any information from genetic variation and comparative genomics at all, and only using its proximity to functional elements. We confirmed that the B statistic is strongly predictive of divergence in our data by stratifying human-chimpanzee genetic divergence along chromosome X by the B-statistic (Figure S8.2). Figure S8.2 shows long regions of low divergence on chromosome X where B is low (and which further bound the human-chimpanzee speciation

time), interspersed with regions of high divergence where the B is high. The pattern in this plot can only be explained by strong directional natural selection in the ancestral population of humans and chimpanzees prior to human-chimpanzee speciation. The cause remains a mystery. Possibilities include an increased rate of background selection in the ancestral population of humans and chimpanzee, an increased rate of positive selection, or selection to remove Dobzhansky-Muller incompatibilities following hybridization[6]. Determining which factors are responsible is outside the scope of this note.

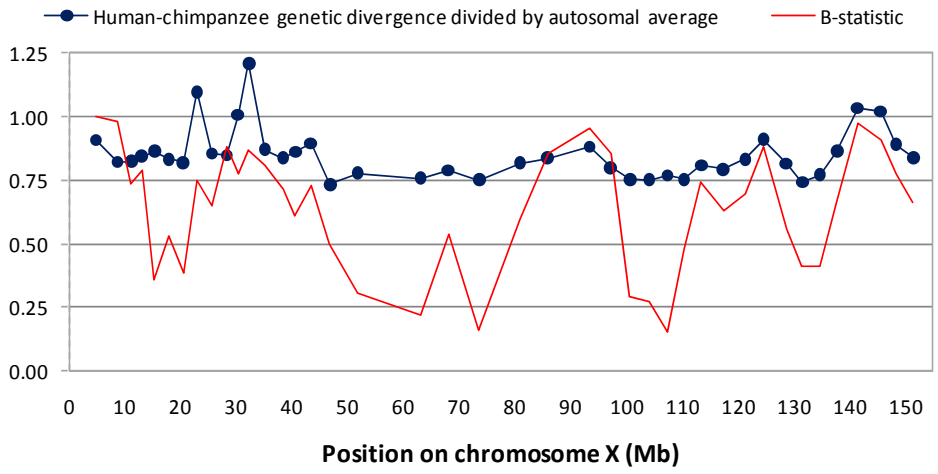


Figure S8.2: B-statistic predicts chromosome X divergence. We analyzed 41 equally sized bins of 40,000 sites excluding pseudoautosomal regions, and plotted human-chimp divergence as a fraction of human-macaque genetic divergence. This strongly correlates to the B-statistic, and there are large regions (e.g. 46.6-86.7 Mb, and 95.6-136.1 Mb) with low average B that also have low average divergence.

To take advantage of the correlation of divergence with selection to set a new constraint on the date of human-chimpanzee speciation, we stratified human-chimpanzee genetic divergence along chromosome X into ten approximately equal-sized bins based on the B-statistic, performing the analysis separately for chromosome X and the autosomes. Figure S8.3 shows that the bin with the smallest B-statistic on the X chromosome gives a new upper bound on $\tau_{HC}/\overline{t}_{HC}$<0.82, even without using the additional information from proximity to human-chimpanzee clustering sites.

Table S8.2: Summary of the bounds on human-chimpanzee genetic divergence

Bound	Description	$\tau_{HC}/\overline{t}_{HC}$
A	Genetic divergence near sites clustering humans and chimpanzees	< 0.826
B	Genetic divergence on chromosome X divided by the autosomes	< 0.851
C	Chromosome X loci close to HC sites (A+B)	< 0.771
D	X loci close to HC sites and B<0.4 (C + B-statistic)	< 0.726

Motivated by the power of the B-statistics to predict human-chimpanzee genetic divergence, we combined all three ideas for finding segments of the genome with reduced divergence to produce an even more stringent (but still conservative) upper bound on human chimpanzee speciation compared with any of the approaches by themselves: (i) Restriction to chromosome X, (ii) Restriction to loci strongly affected by directional selection (B<0.4, where the genetic divergence in Figure S8.3B appears to asymptote), and (iii) Restriction to sites that are within 32 bp of a divergent site that clusters human and chimpanzee to the exclusion of gorilla. From this subset of the data, we obtain a new upper bound of $\tau_{HC}/\bar{t}_{HC} < 0.726$ (green curve in Figure S8.1). For completeness the numbers for the even lower bin sizes are: 0.742 (<16 bp), 0.730 (<8 bp) and 0.671 (<4 bp).) Table S9.2 lists the various bounds. In what follows and the main text, we use the strongest (D), conservatively rounding it off to $\tau_{HC}/\bar{t}_{HC} < 0.73$.

The upper bound of $\tau_{HC}/\bar{t}_{HC} < 0.73$ is conservative and robust

We conclude this section by noting that the true value of the ratio is likely to be less than 0.73.

(a) Upper bounds using X chromosome data are conservative: Our upper bound on human-chimpanzee speciation based on data from the X chromosome is conservative. The reason is that we are dividing by human-macaque divergence to normalize for differences in the mutation rate across loci in the genome, assuming that the average time since the most recent common ancestor (TMRCA) between humans and macaques is identical across the genome. In fact, the TMRCA varies, and is expected to be less on chromosome X than on the autosomes, since in the ancestral population of humans and macaques, the ancestral effective population size is expected to have been less on chromosome X than the autosomes (3/4). As discussed in Patterson et al. 2006, the true TMRCA could plausibly be 0-5% lower on average on chromosome X due to this effect, which will result in an overestimate of our upper bound by the same amount[6].

(b) Upper bounds using X data are not strongly affected by changes in male-to-female mutation rate. In 2009, Presgraves and Yi suggested that the finding of Patterson et al. 2006 of a greatly reduced genetic divergence time on chromosome X relative to the autosomes might be an artifact of changing male-to-female mutation rates among great apes, for example, due to an acceleration of the male mutation rate on the chimpanzee lineage due to more male competition for mates

leading to larger numbers of sperm cell divisions and a higher male mutation rate[26]. To evaluate whether there is evidence that this might affect our inferences, we computed the human-chimpanzee genetic divergence as a fraction of human-macaque divergence across the X chromosome, after separating the data by mutations on the human lineage and chimpanzee lineage since divergence. The inference on the human-specific lineage is $\tau_{HC}/\bar{t}_{HC} < 0.850$, and on the chimpanzee-specific lineage is $\tau_{HC}/\bar{t}_{HC} < 0.852$, suggesting that this is not a major effect.

(c) Although $\tau_{HC}/\bar{t}_{HC} < 0.73$ is a hard bound we conservatively treat it as a soft bound. While $\tau_{HC}/\bar{t}_{HC} < 0.73$ is in principle a hard upper bound—in the sense that we have found loci where the genetic divergence is 72.6% of the autosomal average making this a maximum on human-chimpanzee speciation time—in fact we conservatively treat it as a soft bound in the main text, where we use it as the upper 5% bound of a 90% Bayesian prior probability distribution on the ratio $\tau_{HC}/\bar{t}_{HC}$. Thus, with 5% probability, we allow for the possibility that the true ratio is larger, which means that our quoted upper bound on human-chimpanzee speciation reported in the main text is actually somewhat less stringent than it should be.

(iv) Point estimates of $\tau_{HC}/\bar{t}_{HC}$ = 0.61-0.68 from modeling of background selection

In this section, we obtain new point estimates of the ratio $\tau_{HC}/\bar{t}_{HC}$ that take advantage of the modeling analyses in McVicker et al. 2009[10] taking into account the impact of directional selection on human-chimpanzee genetic divergence to obtain not just an upper bound, but also a best estimate of the ratio. This kind of modeling analysis is important, since as shown in Figure S8.2-S8.3, in our data directional selection is clearly having an important impact.

We first used the modeling of autosomal data directly reported in the McVicker et al. 2009 paper[10]. In Table 1 of their paper (page 7), they give parameter estimates under their model taking into account a fitted model of background selection on the autosomes, which translate to an estimate of $\tau_{HC}/\bar{t}_{HC} = 0.61$, matching the estimate from Burgess and Yang.

As an additional estimate using >100 times more data than was analyzed by McVicker et al. 2009, we examined the correlation of B-statistic with genetic divergence in our own data. If the model underlying the B-statistic is correct, then the value of B (on its scale of 0-1) predicts the

reduction in genetic diversity in the human-chimpanzee ancestral population at a locus, compared with the expectation if there were no selection at all. Assuming that the B-statistics are measured with perfect accuracy and the model is correct, if we measure human-chimpanzee genetic divergence as a fraction of the autosomal average in ten bins of B-statistic, and fit a line, then the y-intercept gives the expected human-chimpanzee genetic divergence at loci in the genome where the time to the common ancestor in the ancestral population was zero; that is, they give the date of human-chimpanzee speciation.

Figure S8.3: Human-chimpanzee divergence divided by the autosome average, stratified by B. We divided (A) the autosomal and (B) chromosome X data into 10 equally sized bins, based on McVicker B-statistics. Blue lines show least squares fits to all ten data points, and red lines leave out three points that contribute to non-linearity and may reflect model failure (the two points with the lowest B and the one point with the highest B). The y-intercepts provide an estimate of human-chimp speciation as a fraction of the autosomal divergence; that is, the expected genetic divergence assuming no genetic variation in the ancestors.

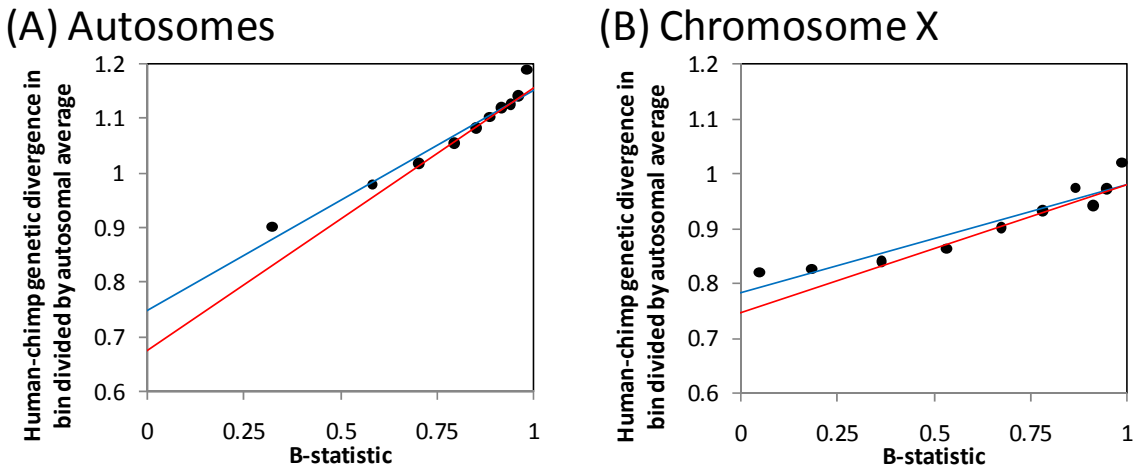


Figure S8.3 shows the empirical relationship of genetic divergence between human and chimpanzee to B-statistics on the autosomes and chromosome X separately. There is evident non-linearity, mostly in the two bins with the lowest B-statistics. A potential explanation (even if the model is correct) is "regression to the mean". The assignment of B-statistics to individual nucleotides is noisy and thus the bin of nucleotides with the lowest B-statistics is likely to contain a substantial fraction of nucleotides that are not in fact so constrained by selection as indicated by their assigned B-statistic. Thus, the observed human-chimpanzee divergence in these bins is not as reduced as predicted. We therefore fit lines not just to all ten bins, but also to

a subset of seven bins that exclude the two with the lowest B-statistics, and the highest bin (which appears to be an outlier perhaps due to structural variation). In the middle seven bins, the points appear linear. The extrapolated y-intercept from the fitted (red) regression line is $\tau_{HC}/\overline{t}_{HC}$ = 0.68 on the autosomes, giving a new point estimate. (On chromosome X, it is $\tau_{HC}/\overline{t}_{HC}$ = 0.75 (Figure S8.3), but we focus here on the autosomes since McVicker et al. 2009 had much better autosomal data to use in their modeling analysis and obtained a much better fit of their B-statistic model to the data on the autosomes. Moreover, the best estimate of the ration on chromosome X is clearly too high, as it exceeds the upper bound of section (iii).)

(v) Prior distribution on $\tau_{HC}/\overline{t}_{HC}$

Above, we described several inferences about the ratio of human-chimpanzee speciation to average human-chimpanzee genetic divergence:

(a) We described a point estimate of $\tau_{HC}/\overline{t}_{HC}$ (0.61) based on the modeling analyses under neutral evolution from Burgess and Yang, which is consistent with Dutheil and colleagues.

(b) We described a conservative upper bound of <0.73.

(c) We described point estimates of $\tau_{HC}/\overline{t}_{HC}$ (0.61-0.68) from modeling analyses that take into account background selection using insights from McVicker et al. 2009.

Taking these various inferences into account, we propose a prior distribution on $\tau_{HC}/\overline{t}_{HC}$ that is normally distributed, and that allows 5% of its density above 0.73 and 10% of its density below 0.61. Thus, its mean is 0.663, and its standard deviation is 0.041 (Figure S8.4). This distribution captures the observation that none of the point estimates are substantially below 0.61, and that we have a strong upper bound at 0.73 (which conservatively, we treat as a soft upper bound, although in fact it would be very surprising if the true value was higher).

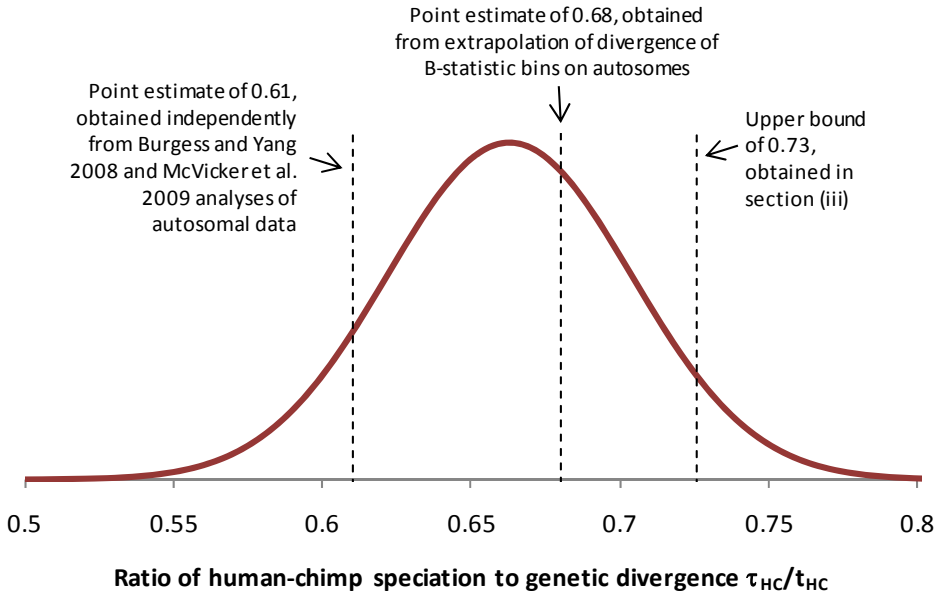


Figure S8.4: Prior distribution on the ratio of human-chimp speciation to genetic divergence, $\tau_{HC}/\overline{t}_{HC}$. This distribution has a mean of 0.663 and a standard deviation of 0.041, set so that 10% of the density is below 0.61 and 5% of the density is above 0.73. The inferences that we use to inform this prior are indicated by dashed lines.

We conclude by discussing what the effect on our inferences would be if the true value of the ratio was below 0.61, which is especially relevant since two of the point estimates were at this value. Lower values would reduce the posterior estimate of the human-chimpanzee speciation date, which is already lower in our paper than would be consistent with some interpretations of the fossil record. Figure 4 of the paper allows readers to ignore our prior, and instead infer the speciation date that would be obtained for any choice of $\tau_{HC}/\overline{t}_{HC}$. This analysis shows that speciation dates above 6.8 Mya (the current minimum date of the *Sahelanthropus* fossil) require a ratio of $\tau_{HC}/\overline{t}_{HC}$ >0.70.

Note S9. Hierarchical Bayes Model

Because of inter-locus variation in mutation rate, statistics such as the standard error of the mutation rate, pooled across loci, become non-trivial. To estimate such statistics, and to find out the degree of inter-locus variation in mutation rate, we model the data using a Hierarchical Bayes Model (HBM).

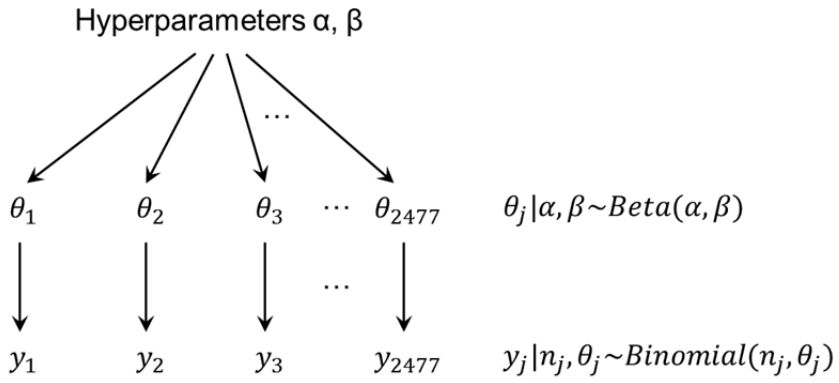
The framework of the HBM is as follows: (1) Describe the data generative process using a set of equations, that is, the method to generate data (mutation events) given the parameters. (2) Derive the posterior distribution, which is conditioned upon the data. (3) Using the set of posterior equations with the empirical data as input, sample the posterior distribution using direct-sampling or MCMC techniques. (4) Perform extensive model-checking to ensure that the HBM performs appropriately.

II. Methods

Hierarchical model of the mutation process

1. Data generative process

   For loci $j = 1, \ldots, J$, the numbers of mutations y_j are modeled as independent binomial samples: $y_j|n_j, \theta_j \sim Bin(n_j, \theta_j)$, where n_j is the number of observations and assumed to be known. θ_j is the mutation rate. We use a conjugate distribution $\theta_j|\alpha, \beta \sim Beta(\alpha, \beta)$ with hyperparameters α, β that are the same for all θ_j.



2. The joint posterior density $p(\theta, \alpha, \beta|y)$ is as follows:

$$p(\theta, \alpha, \beta|y) = \frac{p(\theta, \alpha, \beta)p(y|\theta, \alpha, \beta)}{p(y)} \tag{1}$$

$$\propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta) \tag{2}$$

$$= p(\alpha, \beta)\prod_j p(\theta_j|\alpha, \beta)p(y_j|\theta_j) \tag{3}$$

$$= p(\alpha, \beta)\prod_j Beta(\alpha, \beta)Bin(n_j, \theta_j) \tag{4}$$

$$\propto p(\alpha, \beta)\prod_j \frac{1}{B(\alpha, \beta)}\theta_j^{\alpha-1+y_j}(1-\theta_j)^{\beta-1+n_j-y_j} \tag{5}$$

Line 1 is by Bayes rule.

Line 2 is the product of the hyper-prior distribution, the parameter distribution, and the likelihood.

Line 3 follows by conditional independence of the parameter and data.

Lines 4 and 5 follow from our data generative model. $B(\alpha, \beta)$ is the beta function.

3. In order to sample from the posterior, we first find $p(\alpha, \beta|y)$ by integrating over each θ_j from 0 to 1, obtaining:

$$p(\alpha, \beta|y) \propto p(\alpha, \beta)\prod_j \frac{B(\alpha + y_j, \beta + n_j - y_j)}{B(\alpha, \beta)}$$

4. A suitable hyper-prior distribution $p(\alpha, \beta)$: We would like to choose a diffuse prior. However, an improper prior such as $p(\alpha, \beta) = 1$ doesn't work because $p(\alpha, \beta|y)$ cannot integrate to 1. This is because

$$\lim_{\alpha,\beta\to\infty} \frac{B(\alpha + y_j, \beta + n_j - y_j)}{B(\alpha, \beta)} = 1$$

Instead, we choose a diffuse (uniform) density on $(\frac{\alpha}{\alpha+\beta}, (\alpha + \beta)^{-\frac{1}{2}})$, which are the mean and approximately proportional to the standard deviation of $\theta_j|\alpha, \beta \sim Beta(\alpha, \beta)$. From equation 5.9 of Gelman et al[27], this leads to $p(\alpha, \beta) \propto (\alpha + \beta)^{-\frac{5}{2}}$. Hence,

$$p(\alpha, \beta|y) \propto (\alpha+\beta)^{-5/2} \prod_j \frac{B(\alpha+y_j, \beta+n_j-y_j)}{B(\alpha,\beta)}$$

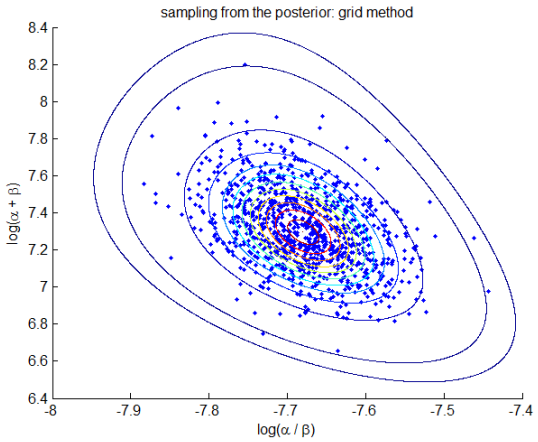
Drawing simulations from the posterior distributions

1. The first step is to crudely estimate the parameters θ, α, β. From the data, we find $mean\left(\frac{y_j}{n_j}\right) = 5 \times 10^{-4}$ and $var\left(\frac{y_j}{n_j}\right) = 5 \times 10^{-6}$, obtaining estimates of $(\theta, \alpha, \beta) = (5 \times 10^{-4}, 0.05, 99)$.

2. Next, we look for the posterior mode of $p(\alpha, \beta|y)$. When calculating values of the posterior, to avoid numerical issues, we compute the log posterior, then exponentiate at the end. We can use the EM algorithm to find the mode, using our crude estimates as a starting point. Alternatively, for this 2 dimensional problem, we can simply use a grid of (α, β) to look for $\max_{\alpha,\beta} p(\alpha, \beta|y)$ in the vicinity of the crude estimates. We find that the posterior mode is located at $(\alpha, \beta) = (0.68, 1480)$. At the mode, this would correspond to $E[\theta|\alpha, \beta] = 4.6 \times 10^{-4}$ and $var[\theta|\alpha, \beta] = 3 \times 10^{-7}$. Our variance here is about 10 times smaller than that of our crude estimates. This is because $var\left(\frac{y_j}{n_j}\right) = 5 \times 10^{-6}$ was estimating $var(\theta)$, taking into account variability in (α, β).

   Below is a contour plot of $p(\alpha, \beta|y)$, re-parameterized in terms of $\left(\log\frac{\alpha}{\beta}, \log\alpha+\beta\right)$, with contours at 0.0001, 0.001, and at 0.05, 0.15, 0.25, …, 0.95 of the modal value.

3. Given our sense of how $p(\alpha, \beta|y)$ behaves, we now sample from the posterior. We directly sample via grids. This method is feasible because we are sampling only in 2 dimensions. Using the contour plot above, we compute the grid of points where most of the density lies. Then, we numerically sum one dimension to obtain the marginal distribution, say $p(\alpha|y)$. α is then sampled using the inverse-CDF method. Then we sample β using the inverse-CDF method again, this time on $p(\beta|\alpha, y)$. 1000 samples of (α, β) are shown below.



4. After sampling from $p(\alpha, \beta|y)$, we sample θ using $p(\theta|\alpha, \beta, y)$. Note that the posterior for θ is beta distributed, and has parameters that combine the data and the hyper-parameters:

$$p(\theta_j|\alpha, \beta, y_j) = Beta(\alpha + y_j, \beta + n_j - y_j)$$

With the hierarchical framework, for each sample of (α, β), we sample the entire set of 2,477 θ_j. This is one experiment. Since we have 1,000 samples of (α, β), we run 1,000 experiments and obtain a confidence bound for each θ_j. The plot below shows our posterior for θ_j. The horizontal axis gives the 2,477 mutation rates, taken as the raw ratio of mutant to observed events. The vertical axis gives the posterior. Crosses "x" are the median. Gray vertical bars show the 95% posterior confidence interval. The y=x line is in red. The red vertical line on the left shows the median and confidence interval of a locus that has $n_j = 0$, an uninformative locus. Note that the slope of a regression line through the crosses would be substantially less than 1. This is the effect of "smoothing" the raw mutation rates, using the combined information from all loci.