nature genetics

A direct characterization of human mutation based on microsatellites

James X Sun[1,2], Agnar Helgason[3,4], Gisli Masson[3], Sigríður Sunna Ebenesersdóttir[3], Heng Li[2,5], Swapan Mallick[2], Sante Gnerre[5], Nick Patterson[5], Augustine Kong[3], David Reich[2,5] & Kari Stefansson[3,6]

Mutations are the raw material of evolution but have been difficult to study directly. We report the largest study of new mutations to date, comprising 2,058 germline changes discovered by analyzing 85,289 Icelanders at 2,477 microsatellites. The paternal-to-maternal mutation rate ratio is 3.3, and the rate in fathers doubles from age 20 to 58, whereas there is no association with age in mothers. Longer microsatellite alleles are more mutagenic and tend to decrease in length, whereas the opposite is seen for shorter alleles. We use these empirical observations to build a model that we apply to individuals for whom we have both genome sequence and microsatellite data, allowing us to estimate key parameters of evolution without calibration to the fossil record. We infer that the sequence mutation rate is 1.4–2.3 × 10^{-8} mutations per base pair per generation (90% credible interval) and that human-chimpanzee speciation occurred 3.7–6.6 million years ago.

The largest studies of human germline mutation to date have focused on whole-genome sequencing of nuclear families[1–3] and have identified more than a hundred new mutations. However, too few mutations were detected, and too few families were studied to provide a detailed characterization of the mutation process[4–7]. One outcome of understanding the mutation process would be a direct estimate of the rate of the molecular clock, which would make it possible to use genetic data to estimate dates of historically important events such as population separations without relying on the fossil record for calibration.

Here, we focus on microsatellites: 1- to 6-base-pair motifs that vary in the number of times they are repeated. Caused by DNA polymerase slippage during replication, mutations of microsatellites occur at a rate of approximately 1 × 10^{-4} to 1 × 10^{-3} mutations per locus per generation[8–12], which is far higher than existing estimates of the nucleotide substitution rate of around 1 × 10^{-8}. We analyzed 2,477 autosomal microsatellites that had been genotyped as part of linkage-based disease gene mapping studies and that were ascertained to be highly polymorphic[13]. The data set included microsatellite data from 85,289 Icelanders from 24,832 father-mother-child trios, after restricting analysis to individuals genotyped for at least half of these loci and without evidence of inaccurate parental assignment (Online Methods and **Supplementary Fig. 1**). The median genotype error rate was 1.8 × 10^{-3} mutations per allele (**Supplementary Fig. 2** and **Supplementary Note**), which is high compared to the mutation rate, and we thus took additional steps to reduce the error rate.

To distinguish genuine mutations from genotyping errors, we used two approaches (Online Methods and **Supplementary Note**). In the 'trio' approach (**Fig. 1a**), we identified 1,695 mutations in 5,085,672 transmissions by restricting analysis to instances in which each member of the trio was genotyped more than once. In the 'family' approach (**Fig. 1b**), we identified 363 mutations in 952,632 transmissions, 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 (in the family approach, we also used haplotypes of nearby microsatellites to determine the parental origin of mutations; Online Methods and **Supplementary Note**). The trio and family approaches produced indistinguishable inferences about the mutation process (**Table 1** and **Supplementary Figs. 3** and **4**), and, hence, we combined the data from these approaches for subsequent analysis (62 mutations were counted twice due to overlap).

To estimate the proportion of candidate mutations that are real, we regenotyped the individuals who had been used to discover 103 trio mutations and 99 family mutations, leading to false positive rate estimates of 2.9% and 2.6%, respectively (**Supplementary Fig. 5** and **Supplementary Table 1**). We also estimated the false positive rate due to errors in the allele-calling algorithm to be 4.3% by manually rescoring the electropherograms of 316 individuals from the family data set, declaring a false positive if there was disagreement. Combining the two sources of error, we estimated a 7.2% false positive rate (**Supplementary Table 1**). We also obtained an independent estimate of the false positive rate by analyzing next-generation sequencing data from the proband and the transmitting parent for 14 candidate mutations for which we had such data, allowing us to validate all but 1 and leading to an estimated false positive rate of 7.1% (**Supplementary Table 2**). The false negative rate (probability of an undetected real mutation) was estimated to be 9.0% by simulating mutations and recording the fraction that we did not detect (Online Methods).

[1]Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]deCODE Genetics, Reykjavik, Iceland. [4]Department of Anthropology, University of Iceland, Reykjavik, Iceland. [5]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [6]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. Correspondence should be addressed to J.X.S. (toxinsun@gmail.com), D.R. (reich@genetics.med.harvard.edu) or K.S. (kari.stefansson@decode.is).

Received 7 May; accepted 12 July; published online 23 August 2012; doi:10.1038/ng.2398

Table 1 Direct estimates of microsatellite mutation rates

	Mutations	Transmissions	Mutation rate ($\times 10^{-4}$)[a]	
			Mean	5th–95th percentile
Dinucleotide loci[b]				
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
Tetranucleotide 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

[a]The 90% credible interval was calculated on the basis of a Bayesian hierarchical beta-binomial model (**Supplementary Note**), which allows for the mutation rate to vary across loci. [b]A breakdown of the mutation rate by motif type for dinucleotides is given in **Supplementary Table 3**.

The estimated mutation rate for tetranucleotide microsatellites was 10.01×10^{-4} mutations per locus per generation, which is 3.7 times higher than the dinucleotide rate of 2.73×10^{-4} (**Table 1** and **Supplementary Table 3**). Estimates were nearly unchanged after correcting for false positives and false negatives by (1 –0.072)/(1 – 0.090), and, thus, we quote unadjusted rates here. Our estimate of the male-to-female mutation rate ratio (α) was 3.3 (95% credible interval (CI) 2.9–3.7; **Supplementary Table 4**), within the range of 2–7 that was previously inferred for sequence substitutions[3,4,14,15]. Paternal age was correlated with mutation rate ($P = 9.3 \times 10^{-5}$), whereas maternal age was not ($P = 0.47$) (**Fig. 2a** and **Supplementary Fig. 6**), consistent with observations based on disease-causing mutations and the fact that male germ cells undergo numerous mitoses as a man ages, whereas female oocytes do not undergo postnatal cell division[4].

These data allow the first high-resolution, direct characterization of the mutation process for the highly polymorphic di- and tetranucleotide microsatellites that are typically genotyped[8]. First, 32% of mutations at dinucleotide microsatellites were multistep, compared to 1% at tetranucleotide microsatellites (**Fig. 2b** and **Supplementary Fig. 3**) (this explains why the variance of allele-length distribution at the tetranucleotide microsatellites was similar to that of the dinucleotide microsatellites, despite their 3.7-fold higher mutation rate[16,17]). Second, the mutation rate increased with allele length[18,19], quadrupling between 30 and 70 bp for dinucleotide and 40 and 120 bp for tetranucleotide microsatellites (both tests significant, $P < 0.002$) (**Fig. 2c**). Third, loci with uniform repeat structures (for example, CACACACA) had a 40% higher rate ($P = 3 \times 10^{-7}$) than compound repeat structures (for example, CACATCACA), consistent with less DNA polymerase slippage for interrupted tandem repeats[8,20] (**Supplementary Fig. 7**). Fourth, we detected length constraints[21,22], with shorter alleles tending to mutate to become longer and vice versa ($P = 2 \times 10^{-15}$) (**Fig. 2d** and **Supplementary Figs. 8** and **9**)[20,23,24]. This pattern contrasts with that in trinucleotide repeat disorders, where long alleles tend to mutate to even longer lengths[19]. Fifth, the mutation rate correlated ($P < 1 \times 10^{-4}$) with motif length, repeat number, allele size, distance from exons, gender and age, but not with recombination rate, distance from telomeres, human-chimpanzee divergence and parental heterozygosity (**Supplementary Tables 5** and **6** and **Supplementary Note**).

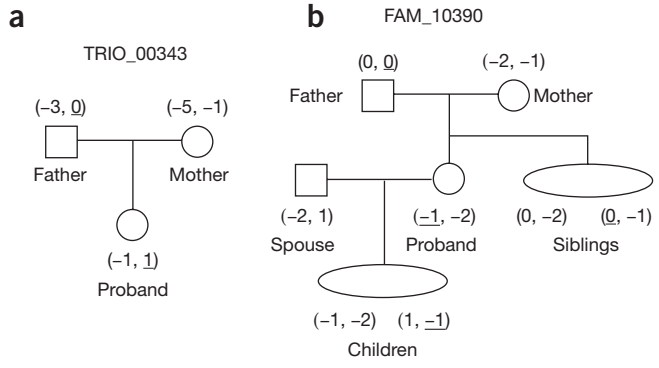

Figure 1 Examples of mutations in a trio and in a family. The proband is the individual inheriting a mutation, and all other individuals are named relative to the proband. All alleles are given in repeat units and are shifted so that the ancestral allele has length of 0. The mutating allele is underlined.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

Microsatellites have been widely used to make inferences about evolutionary history. However, the accuracy of these inferences has been limited by a poor understanding of the mutation process. We developed a new model of microsatellite evolution (**Supplementary Note**). This model can estimate the time to the most recent common ancestor (TMRCA) of two samples at a microsatellite by taking into account (i) the dependence of the mutation rate on allele length and parental age (**Fig. 2a,c**); (ii) the step size of mutations (**Fig. 2b**); (iii) the size constraints on allele length (**Fig. 2d** and **Supplementary Figs. 8** and **9**); and (iv) the variation in generation interval over history. In contrast to the generalized stepwise mutation model (GSMM), which predicts a linear increase of average squared distance (ASD) between microsatellite alleles over time, the new model predicts a sublinear increase (**Fig. 3**) and saturation of the molecular clock, due to the constraints on allele length. We also extended the model to estimate the sequence mutation rate, using the per-nucleotide diversity flanking each microsatellite as an additional datum. To implement the model, we used a Bayesian hierarchical approach, first generating global parameters common to all loci, followed by locus-specific parameters and finally the microsatellite alleles at each locus (Online Methods). We used Markov chain Monte Carlo to infer TMRCA and sequence mutation rate.

We validated the model in three ways (Online Methods). First, we simulated data sets in which we knew 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 (**Supplementary Note**). Second, we carried out sensitivity analyses by perturbing model parameters and found that our key inferences are robust (**Supplementary Fig. 10** and **Supplementary Note**). Third, we empirically validated the model by analyzing 23 individuals for whom we had both microsatellite genotypes and whole-genome sequencing data[2] and comparing the ASD to the surrounding sequence heterozygosity as a surrogate for TMRCA. The ASD predicted by our model is similar to the empirical curve (**Fig. 3** and **Supplementary Fig. 11**).

Our approach allows inference of 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 had whole-genome sequencing data (Online Methods), and comparing the ASD to local heterozygosity and human-macaque divergence (as a surrogate

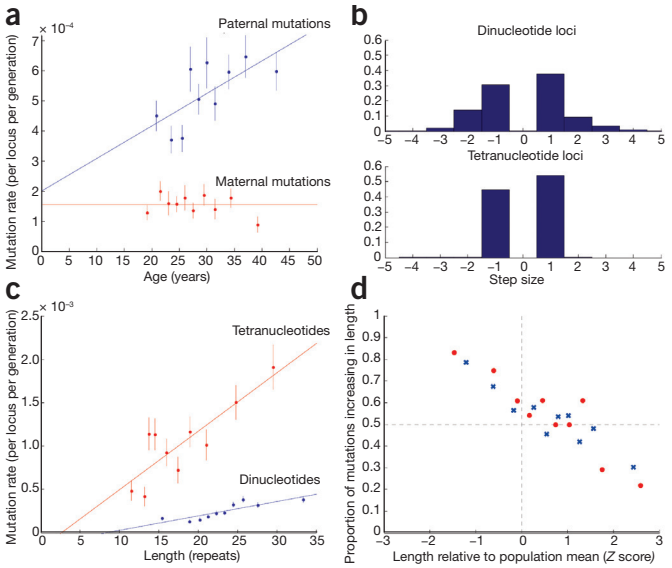

Figure 2 Characteristics of the microsatellite mutation process. (**a**) Paternal (blue) and maternal (red) mutation rates. The *x* axis shows the parental age at childbirth. Data points are grouped into ten bins (vertical bars show one standard error). The paternal rate shows a positive correlation with age (logistic regression of raw data: $P = 9.3 \times 10^{-5}$; slope = 1.1×10^{-5} mutations per year), 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 for dinucleotide (top) and tetranucleotide (bottom) microsatellites. Whereas the dinucleotide loci experience multistep mutations in 32% of instances, tetranucleotide loci mutate almost exclusively by a single step of 4 bases. (**c**) Mutation rate increases with allele length. Dinucleotide loci (blue) have a slope of 1.65×10^{-5} mutations per repeat unit ($P = 1.3 \times 10^{-3}$), and tetranucleotide loci (red) have a slope of 6.73×10^{-5} mutations per repeat unit ($P = 1.8 \times 10^{-3}$).(**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, mutations tend to decrease in length. Di- and tetranucleotide loci are shown as blue crosses and red circles, respectively. Probit regression of the combined di- and tetranucleotide data show highly significant evidence of an effect ($P = 2.8 \times 10^{-18}$).

for the local mutation rate; Online Methods), we inferred the sequence mutation rate and the TMRCA averaged across the genome (**Table 2**). We also inferred a 90% credible interval via a Bayesian approach integrating over uncertainty in the model parameters (Online Methods, **Supplementary Table 7** and **Supplementary Note**). Empirically, mutation rate estimates tend to be more similar within rather than between populations (**Supplementary Fig. 12** and **Supplementary Table 8**). The differences across populations are not likely to be due to poor modeling of demographic history, as, when we modeled more realistic histories involving two bottlenecks in non-Africans, we obtained the same results (**Supplementary Fig. 13**). The mutation rate differences between populations may be due to shared history, but they are not significant; therefore, we pooled our data across the 23 individuals to produce a sequence mutation rate estimate of 1.82×10^{-8} mutations per base pair per generation (90% CI $1.40–2.28 \times 10^{-8}$; **Table 2**) (the confidence interval takes into account correlations in the histories of the 23 individuals through a jackknife) (Online Methods).

Our inference of the sequence mutation rate is consistent with Nachman and Crowell's estimate of $1.3–2.7 \times 10^{-8}$, which is the average mutation rate since humans and chimpanzees diverged[6]. It is also consistent with Kondrashov's direct estimate of $\hat{\mu}_{seq}$ of 1.8×10^{-8} mutations per base pair per generation[25] from studies of disease-causing genes. However, the lower bound of our 90% CI is higher than those in two recent studies based on whole-genome sequencing data: $\hat{\mu}_{seq} = 1.1 \times 10^{-8}$ mutations per base pair per generation based on 28 sequence mutations detected in a 4-member family[1], and $\hat{\mu}_{seq} = 1.0 \times 10^{-8}$ and 1.2×10^{-8}, based on 84 sequence mutations detected in 2 trios[2,3]. We considered the possibility that this discrepancy might be due to ascertainment bias because the microsatellites we analyzed were selected to be highly polymorphic (for disease-associated gene mapping), which could cause ASD to be too high. However, this would overestimate TMRCA at the loci we analyzed and thus underestimate the mutation rate, opposite to what would be necessary to explain the discrepancy (**Supplementary Fig. 12** and **Supplementary Note**). We hypothesize that the lower mutation rate estimates from the whole-genome sequencing studies might be due to (i) the limited number of mutations detected in these studies, which explains why their confidence intervals overlap ours, (ii) possible underestimation of the false negative rate in the whole-genome sequencing studies or (iii) variability in the mutation rate across individuals, such that a few families cannot provide a reliable estimate of the population-wide rate. There is already empirical evidence for variability in the mutation process across individuals: in one trio analyzed in the 1000 Genomes Project study[8], the father transmitted 92% of the mutations, whereas, in the other trio, the father transmitted 36% of the mutations. 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 that are not accessible from microsatellite analysis.

Our direct estimation of the microsatellite mutation rate, combined with comparative genomics data, also allows us to estimate the date of human-chimpanzee speciation, τ_{HC}, which we define as the date of the last gene flow between human and chimpanzee ancestors[26,27]. We estimate a genome-wide average human-chimpanzee genetic divergence time t_{HC} of 5.80–9.77 million years ago[28] (**Table 2** and Online

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

	Mean	5th–95th percentile[a]	Mean	5th–95th percentile
Present-day mutation rates	(per generation per site)		(per year per site)	
Dinucleotide 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	(thousand generations ago)		(million years ago)	
t_{CEU} Northern and 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	717–1,220	19.80	15.20–25.9
τ_{HC} human-chimpanzee speciation time	233	176–309	4.97	3.75–6.57

[a]90% credible interval obtained from the Bayesian posterior distribution.

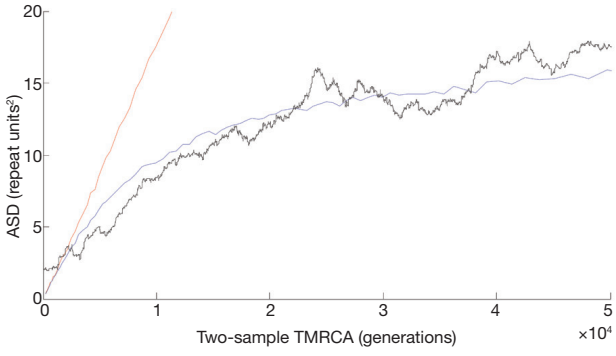

Figure 3 Empirical validation of our model with sequence-based estimates of TMRCA. Shown in red is the simulation of ASD as a function of TMRCA for the standard random walk (GSMM) model. In blue is the simulation of our model in which the nonlinearity compared to GSMM is primarily due to the length constraint 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 indicates that our model is consistent with the data and motivates the analysis in which we use the sequence substitution rate in small windows around the microsatellites to make inferences about evolutionary parameters such as the sequence mutation rate.

Methods). By definition, this divergence date must be older than the speciation date, τ_{HC}. We then inferred the human-chimpanzee speciation date, τ_{HC}, to be 3.75–6.57 million years ago by integrating our inferred distribution of t_{HC} with a Bayesian prior distribution of τ_{HC}/t_{HC} of 0.663 ± 0.041 (**Fig. 4**). We obtained the mean of the prior distribution from previous modeling studies that inferred τ_{HC}/t_{HC} = 0.61–0.68 (refs. 29,30) and the standard deviation of the prior distribution by setting the 95th percentile upper bound equal to 0.73, a value we obtained by analyzing human-chimpanzee sequence data in regions with a reduced divergence compared to the autosomal average due to being (i) on chromosome X; (ii) in proximity to genes; and (iii) near divergent sites that cluster humans and chimpanzees to the exclusion of gorillas (**Supplementary Note**). Our upper bound of τ_{HC} < 6.57 million years ago is lower than the estimate of 6.8–7.2 million years ago for *Sahelanthropus tchadensis*[31], a fossil that has been interpreted to be on the human lineage after the final separation of human and chimpanzee ancestors[32] because it shares derived features with other hominins, such as bipedal posture, reduced canines and expanded post-canines with thicker enamel[33] (**Fig. 4**). We also obtained an independent upper bound on the human-chimpanzee speciation date, τ_{HC}, of 6.3 million years ago on the basis of calibration to the fossil record of human-orangutan speciation[26] (**Supplementary Note**). If our date estimates are correct, then a possible explanation for the discrepancy between the genetic and fossil record dates is that *Sahelanthropus* was not a hominin but instead shared independently derived similarities (homoplasies)[34]. Alternatively, populations with hominin traits may have continued to exchange genes with chimpanzee ancestors after *Sahelanthropus*[26]. Finally, the age of *Sahelanthropus*[31] may be overestimated.

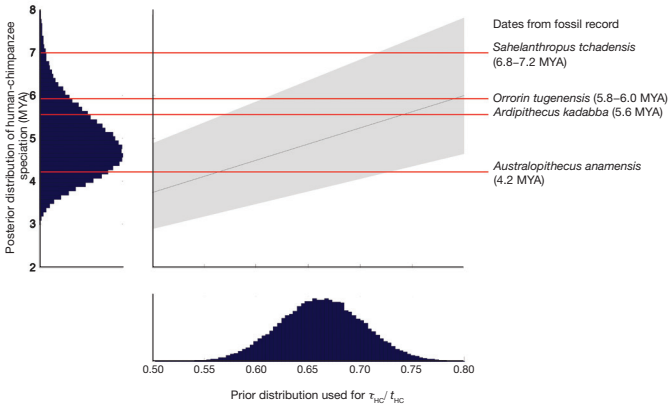

Figure 4 Human-chimpanzee speciation date inferred without calibration with the fossil record. The 90% Bayesian credible interval for human-chimpanzee speciation time (gray) for a range of values of the ratio of speciation time to divergence time (τ_{HC}/t_{HC}). The blue histogram shows our Bayesian prior distribution for τ_{HC}/t_{HC}, justified in the **Supplementary Note**. The red horizontal lines are the dates of fossils that are candidates for being on the hominin lineage after the speciation of humans and chimpanzees. *Australopithecus anamensis, Orrorin tugenensis* and *Ardipithecus kadabba* are within our plausible speciation times, whereas *Sahelanthropus tchadensis* predates the inferred speciation time for all plausible values of τ_{HC}/t_{HC}. Bottom histogram, our Bayesian prior distribution for τ_{HC}/t_{HC}; left histogram, our posterior distribution of human-chimpanzee speciation time. MYA, million years ago.

*Note added in proof: After this paper was accepted, another study[35] was published that independently estimates the human sequence mutation rate, using a direct measurement in contrast to the indirect measurement we report here. In spite of some key similarities between our results and those of Kong et al.[35] (the male-to-female mutation rate ratio and the absence of an effect of mother's age), they estimate a considerably stronger effect of father's age and an overall sequence mutation rate below the range we infer. The discrepancies in the sequence mutation rate may be in part due to the fact that Kong et al. focus on a more intensively filtered subset of the human genome than we analyze here, but other factors are also likely to be at work (**Supplementary Note**). As an initial attempt to compare the two studies in terms of their implications for evolutionary history, we ran the same Bayesian inference procedure we developed in this paper (integrating over uncertainty in unknown parameters), now using the sequence-based estimates rather than the microsatellite-based estimates as input (**Supplementary Note**). Notably, the inferred dates based on the measurement of the sequence mutation rate are older and no longer in direct conflict with the inference that S. tchadensis is on the human lineage since the split from chimpanzees. The sequence- and microsatellite-based data sets are very different, and an important direction for future research will be to understand why the direct sequence–based mutation rate estimate is lower than the one inferred on the basis of microsatellites.*

URLs. Complete Genomics data, http://www.completegenomics.com/public-data/69-Genomes/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The informed consents associated with the samples at deCODE Genetics specify that genotypes cannot be shared outside of Iceland. However, researchers who wish to reanalyze the data can visit deCODE Genetics to perform these analyses by arrangement with K.S.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D.F. Gudbjartsson for advice on running Allegro 2.0; J. Fenner, J. Hawks, K. Langergraber, D. Pilbeam and L. Vigilant for discussions that informed the Bayesian prior distributions on evolutionary parameters; and Y. Erlich, M. Gymrek, D. Lieberman, B. Payseur, D. Pilbeam, A. Siepel, S. Sunyaev and the anonymous reviewers for critiques. This work was supported by a Bioinformatics and Integrative Genomics PhD training grant (J.X.S.), a Burroughs Wellcome Travel Grant (J.X.S.), a Burroughs Wellcome Career Development Award in the Biomedical Sciences (D.R.), a HUSEC seed grant from Harvard University (D.R.), a SPARC award from the Broad Institute of Harvard and MIT (D.R.), a National Science Foundation HOMINID grant 1032255 (D.R.) and US National Institutes of Health grant R01HG006399 (D.R.).

AUTHOR CONTRIBUTIONS

J.X.S., A.H., G.M. and D.R. conceived and performed the research. A.H., G.M., A.K., D.R. and K.S. jointly supervised the study, with A.H. acting as the coordinator at deCODE Genetics and D.R. at Harvard Medical School. A.H. and G.M. prepared the raw microsatellite data. J.X.S., A.H. and S.S.E. designed and analyzed the regenotyping, resequencing and electropherogram re-examination experiments; and A.H. analyzed next-generation sequencing data to independently validate mutations. J.X.S., A.H., N.P., A.K. and D.R. designed and analyzed the microsatellite modeling and the statistics. S.M., H.L. and J.X.S. processed and extracted sequence data for the 23 HapMap individuals. S.M., S.G. and D.R. performed the analyses of human-chimpanzee genetic divergence and developed the Bayesian prior distributions relevant to human-chimpanzee speciation. The manuscript was written primarily by J.X.S., A.H. and D.R. The supplementary information was prepared by J.X.S. and D.R.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Published online at http://www.nature.com/doifinder/10.1038/ng.2398.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
2. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
3. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).
4. Crow, J.F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1**, 40–47 (2000).
5. Crow, J.F. Age and sex effects on human mutation rates: an old problem with new complexities. *J. Radiat. Res.* **47 Suppl B**, B75–B82 (2006).
6. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
7. Arnheim, N. & Calabrese, P. Understanding what determines the frequency and pattern of human germline mutations. *Nat. Rev. Genet.* **10**, 478–488 (2009).
8. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5**, 435–445 (2004).
9. Weber, J.L. & Wong, C. Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**, 1123–1128 (1993).
10. Xu, X., Peng, M. & Fang, Z. The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* **24**, 396–399 (2000).
11. Whittaker, J.C. *et al.* Likelihood-based estimation of microsatellite mutation rates. *Genetics* **164**, 781–787 (2003).
12. Huang, Q.Y. *et al.* Mutation patterns at dinucleotide microsatellite loci in humans. *Am. J. Hum. Genet.* **70**, 625–634 (2002).
13. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
14. Makova, K.D. & Li, W.H. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624–626 (2002).
15. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* **107**, 961–968 (2010).
16. Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462 (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**, 463–471 (1995).
18. Ballantyne, K.N. *et al.* Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *Am. J. Hum. Genet.* **87**, 341–353 (2010).
19. Cummings, C.J. & Zoghbi, H.Y. Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum. Mol. Genet.* **9**, 909–916 (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. USA* **95**, 10774–10778 (1998).
21. Zhivotovsky, L.A., Feldman, M.W. & Grishechkin, S.A. Biased mutations and microsatellite variation. *Mol. Biol. Evol.* **14**, 926–933 (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**, 207–216 (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**, 383–395 (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. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2003).
26. Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103–1108 (2006).
27. Steiper, M.E. & Young, N.M. Primate molecular divergence dates. *Mol. Phylogenet. Evol.* **41**, 384–394 (2006).
28. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
29. 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**, 1979–1994 (2008).
30. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
31. Lebatard, A.E. *et al.* Cosmogenic nuclide dating of *Sahelanthropus tchadensis* and *Australopithecus bahrelghazali*: Mio-Pliocene hominids from Chad. *Proc. Natl. Acad. Sci. USA* **105**, 3226–3231 (2008).
32. Brunet, M. *et al.* A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418**, 145–151 (2002).
33. Lieberman, D.E. *The Evolution of the Human Head* (Belknap Press of Harvard University Press, Cambridge, Massachusetts, 2011).
34. Wood, B. & Harrison, T. The evolutionary context of the first hominins. *Nature* **470**, 347–352 (2011).
35. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).

ONLINE METHODS

Data sets. Microsatellite genotypes were obtained at deCODE Genetics using DNA extracted from blood and multiplexed capillary gel electrophoresis with automated allele calling[13]. We restricted analysis to 2,477 autosomal loci that were genotyped most heavily (all had a minimum repeat length of 5 units). We analyzed 85,289 individuals genotyped from at least half of these loci, from whom we identified 25,067 mother-father-offspring trios using the deCODE Genetics genealogical database (Íslendingabók). All participants in this study provided individual informed consent for this research consistent with protocols approved by the Data Protection Commission of Iceland and the National Bioethics Committee of Iceland. The personal identity of all samples is encrypted, with the key to the code held by the Data Protection Committee of Iceland.

To filter out trios with inaccurate parental assignments, we computed the fraction of loci where both alleles differed between a parent and a child. We empirically set the threshold to filter out almost all known uncle-proband and aunt-proband pairs while retaining almost all known parent-proband pairs (**Supplementary Fig. 1**).

To estimate the per-locus genotyping error rate, we used discordance rates in cases of repeated genotypes (**Supplementary Note**).

Deep whole-genome sequencing data were obtained from two sources. We downloaded nine sequences generated using Illumina technology, mapping reads using Burrows-Wheeler Aligner (BWA)[36] and calling SNPs with SAMtools[37]. We also downloaded 20 Complete Genomics sequences, 6 of which overlapped with the Illumina sequences (**Supplementary Fig. 14** and **Supplementary Table 8**). To estimate heterozygosity around each microsatellite, we extracted data from a window centered on it (for the analyses reported in the main text, the window size was 0.001 cM, and we masked out the central 1 kb[38]; **Supplementary Fig. 15**).

Detecting mutations. For the trio approach, we restricted analysis to transmissions in which all members of the trio were genotyped at least twice and searched for Mendelian inheritance incompatibilities. There were some detected mutations for which the parental origin was ambiguous (**Supplementary Note**), and we included these for analyses of the mutation rate but not for analyses requiring knowledge of parental origin (**Fig. 2b**). For mutations of unambiguous parental origin, the ancestral allele was defined as the one that was closer in length to the mutant allele (we randomly chose the ancestral allele if both were equally close). We filtered out 49 loci that harbored many more mutations from homozygous parents to homozygous children than expected on the basis of Hardy-Weinberg equilibrium, a phenomenon that affected the trio but not the family data. We determined that this was a real error mode due to polymorphisms under the PCR primer sites[39–41] by sequencing primer sites from 15 mutations and identifying 5 with SNPs in the primer region (**Supplementary Note**).

For the family approach, we restricted analysis to transmissions where genotyping was available, not just for a proband's two parents, but also for at least one child and one sibling (**Fig. 1b**). We identified putative mutations by searching for Mendelian incompatibilities between the proband and their parents. We used Allegro 2.0 (ref. 42) to phase the family, masking out the mutant locus, using all available loci from the same chromosome. We then assigned the haplotype carrying the mutation to one of the parents of the proband (**Supplementary Note**). To validate the mutation, we required at least one sibling to carry the haplotype with the ancestral allele, no sibling to carry the mutant and at least one child to carry the haplotype with the mutant.

The trio and family approaches provide complementary information. A bias that only affects the trio approach is somatic mutations in the lineage of genotyped cells but not germline cells transmitted to offspring (this was minimized because the DNA we analyzed was extracted from blood but is still a concern). A bias that only affects the family based approach is 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. The fact that both approaches produced consistent inferences despite their different susceptibilities to bias increases our confidence in the results.

False positive and false negative rates. To estimate the false positive rate, we regenotyped the family members in whom candidate mutations had been found. In the trio data set, we randomly targeted 103 mutations. In the family data set, we targeted mutations that had a higher a priori chance of being in error (**Supplementary Table 1**). To provide an entirely independent estimate of the false positive rate, we identified 14 candidate mutations where we had at least sevenfold whole-genome sequencing data from the proband as well as (at least) the transmitting parent. We then manually examined the data, failing to validate only 1 of the 14 mutations (**Supplementary Table 2**).

To estimate the false negative rate (the proportion of genuine mutations that were missed), we randomly distributed mutations on the genealogy and then tested whether they gave rise to detectable inheritance errors. As an example, suppose that the father-mother-proband trio has genotypes of allele lengths (6, 10), (8, 10) and (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.

Statistical characterization of the microsatellite mutation process. To infer the standard error of the mutation rate, taking into account rate variation across loci, we used a hierarchical Bayesian model (**Supplementary Note**). To infer the number of microsatellite repeats, we started with the amplicon size, which includes not only the repeats but all the sequence between the PCR primers. We then subtracted the span of the flanking sequence inferred from the human genome reference (**Supplementary Fig. 16**). To compute the relative length of an allele, we measured the mean and standard deviation over all individuals at that locus and report the standard deviations from the mean (Z score). To estimate motif impurity (**Supplementary Fig. 7**), we applied Tandem Repeat Finder software to the human genome reference (**Supplementary Fig. 16**). To test for association between the microsatellite mutation process and genomic features (**Supplementary Table 5**), we performed logistic regression to mutation rate and directionality and Poisson regression to step size. To test for interaction, we performed multivariate logistic regression (**Supplementary Table 6**).

Bayesian prior distributions on evolutionary parameters. For Bayesian modeling of sequence mutation rate and genetic divergence times, we required Bayesian prior distributions on evolutionary parameters (**Supplementary Table 7**), including:

1. Generation interval. On the basis of interviews with experts on chimpanzee and gorilla demographic structure (L. Vigilant and K. Langergraber), we assumed that the ancestral generation time was 22.5 ± 4.2 (mean ± s.d.) years. On the basis of the literature[43,44] (**Supplementary Fig. 17**), we assumed that present-day generation time is 29 ± 2 years. We also assumed that the difference between the male and female generation time was 0.5 ± 3.3 years in the ancestral population and 6.0 ± 2.0 today (**Supplementary Note**). We sampled the transition from ancestral to present-day generation time to be a mixture of three equally weighted exponential distributions, with means of 50,000 years ago, 200,000 years ago and 2 million years ago, corresponding to hypothetical changes around the Upper Paleolithic

revolution, evolution of modern humans and evolution of *Homo erectus*, respectively.

2. Human-ape relative genetic divergences. From the literature, we assumed that the ratio of human-chimpanzee to European-European genetic divergence per base pair was 15.400 ± 0.356 (ref. 28) and that the ratio of human-orangutan to human-chimpanzee genetic divergence was 2.650 ± 0.075 (refs. 26,29). We assumed that the molecular process of mutation has been constant over great-ape history.

3. Human-chimpanzee speciation time τ_{HC}. Human-chimpanzee speciation time is by definition less than human-chimpanzee genetic divergence time. Our Bayesian prior distribution on τ_{HC}/t_{HC} was set to have a normal distribution with mean 0.663, within the range of 0.61–0.68 from model-based analyses[29,30]. The Bayesian prior distribution was also set to have a standard deviation of 0.041, based on an analysis in the **Supplementary Note** that places a conservative upper bound on the ratio of human-chimpanzee speciation of 0.73 (in particular, our standard deviation implies that 95% of the density of our Bayesian prior distribution is below 0.73).

Model of microsatellite evolution assisted by flanking sequence heterozygosity. As a metric of microsatellite allelic divergence between two samples, we used average squared distance (ASD).

$$\text{Given allele lengths } x_1, x_2, \ldots, x_n, \text{ ASD} = \frac{1}{n(n-1)} \times \sum_{i,j} (x_i - x_j)^2.$$

To model ASD along with flanking sequence heterozygosity, we simulated 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 (**Supplementary Table 7**) common to all loci were simulated, such as the genome-wide present-day sequence and microsatellite mutation rates, and generation-time effects. One level down, locus-specific mutation rates were computed on the basis of global parameters and locus-specific information. At the third level, for each individual, a two-sample coalescent tree was generated (**Supplementary Note**).

A potential pitfall in inferring TMRCA with our data is that the microsatellites we analyzed were ascertained to be highly polymorphic in Europeans. This raises two complications. First, the sequence flanking the microsatellites may have a different mutation rate than the genome average, and, to correct for this, we compared ASD to the ratio of sequence heterozygosity and human-macaque divergence at each locus (as a surrogate for local mutation rate). Second, ascertainment of highly polymorphic microsatellites can bias toward deeper genealogies than the genome average, which in turn can bias average TMRCA to be too high. By studying the sequence flanking the microsatellites, we determined that the trees were on average 1.04 times deeper than the genome average, and we corrected the estimate of genome-wide average TMRCA in **Table 2** by this factor (**Supplementary Table 9** and **Supplementary Note**).

To infer sequence mutation rate and TMRCA using the microsatellite evolution model, we used Markov chain Monte Carlo (MCMC) following the method described in ref. 45 (**Supplementary Note**). Combining data across individuals is not trivial because of shared history across individuals. To obtain proper standard errors for the combined mutation rate, we performed a jackknife[46], where each locus was removed at a time, and we studied the empirical variance of the inferred mutation rate. The statistical theory of the jackknife allows us to compute an appropriate standard error based on this procedure.

36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
37. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. Hinch, A.G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170–175 (2011).
39. Weber, J.L. & Broman, K.W. Genotyping for human whole-genome scans: past, present, and future. *Adv. Genet.* **42**, 77–96 (2001).
40. Johansson, A.M. & Sall, T. The effect of pedigree structure on detection of deletions and other null alleles. *Eur. J. Hum. Genet.* **16**, 1225–1234 (2008).
41. Callen, D.F. *et al.* Incidence and origin of "null" alleles in the (AC)n microsatellite markers. *Am. J. Hum. Genet.* **52**, 922–927 (1993).
42. Gudbjartsson, D.F., Thorvaldsson, T., Kong, A., Gunnarsson, G. & Ingolfsdottir, A. Allegro version 2. *Nat. Genet.* **37**, 1015–1016 (2005).
43. Fenner, J.N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
44. Helgason, A., Hrafnkelsson, B., Gulcher, J.R., Ward, R. & Stefansson, K. A population-wide 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**, 1370–1388 (2003).
45. Marjoram, P., Molitor, J., Plagnol, V. & Tavare, S. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100**, 15324–15328 (2003).
46. Efron, B. & Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Stat.* **37**, 36–48 (1983).

doi:10.1038/ng.2398