Calibrating the Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes

Mark Lipson,[1,*] Po-Ru Loh,[2,3] Sriram Sankararaman,[1,3]

Nick Patterson,[3] Bonnie Berger,[3,4] David Reich[1,3,5,*]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

[2]Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA

[3]Medical and Population Genetics Program,

Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[4]Department of Mathematics and Computer Science and Artificial Intelligence Laboratory,

Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[5]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA

*To whom correspondence should be addressed

E-mail: mlipson@genetics.med.harvard.edu, reich@genetics.med.harvard.edu

Abstract

The human mutation rate is an essential parameter for studying the evolution of our species, interpreting present-day genetic variation, and understanding the incidence of genetic disease. Nevertheless, our current estimates of the rate are uncertain. Classical methods based on sequence divergence have yielded significantly larger values than more recent approaches based on counting *de novo* mutations in family pedigrees. Here, we propose a new method that uses the fine-scale human recombination map to calibrate the rate of accumulation of mutations. By comparing local heterozygosity levels in diploid genomes to the genetic distance scale over which these levels change, we are able to estimate a long-term mutation rate averaged over hundreds or thousands of generations. We infer a rate of $1.65 \pm 0.10 \times 10^{-8}$ mutations per base per generation, which falls in between phylogenetic and pedigree-based estimates, and we suggest possible mechanisms to reconcile our estimate with previous studies. Our results support intermediate-age divergences among human populations and between humans and other great apes.

Author Summary

The rate at which new heritable mutations occur in the human genome is a fundamental parameter in population and evolutionary genetics. However, recent direct family-based estimates of the mutation rate have consistently been much lower than previous results from comparisons with other great ape species. Because split times of species and populations estimated from genetic data are often inversely proportional to the mutation rate, resolving the disagreement would have important implications for understanding human evolution. In our work, we apply a new technique that uses mutations that have accumulated over many generations on either copy of a chromosome in an individual's genome.

Instead of an external reference point, we rely on fine-scale knowledge of the human recombination rate to calibrate the long-term mutation rate. Our procedure accounts for possible errors found in real data, and we also show that it is robust to a range of model violations. Using eight diploid genomes from non-African individuals, we infer a rate of $1.65 \pm 0.10 \times 10^{-8}$ single-nucleotide changes per base per generation, which is intermediate between most phylogenetic and pedigree-based estimates. Thus, our estimate implies reasonable, intermediate-age population split times across a range of time scales.

Introduction

All genetic variation—the substrate for evolution—is ultimately due to spontaneous heritable mutations in the genomes of individual germline cells. The most commonly studied mutations are point mutations, which consist of single-nucleotide changes from one base to another. The rate at which these changes occur, in combination with other forces, determines the frequency with which homologous nucleotides differ from one individual's genome to another.

A number of different approaches have previously been used to estimate the human mutation rate [1–3], of which we mention four categories here. The first method is to count the number of fixed genetic changes between humans and another species, such as chimpanzees [4]. Population genetic theory implies that neutral mutations (those that do not affect an organism's fitness) should accumulate between two genomes at a constant rate (the well-known "molecular clock" [5]). Thus, the mutation rate can be estimated based on the divergence time of the genomes, if this can be confidently inferred from fossil evidence. However, even if the age of fossil remains can be accurately determined, assigning their proper phylogenetic positions is often difficult. Moreover, because of shared ancestral polymorphism, the time to the most recent common ancestor is always older—and sometimes far older—than the time of species divergence, meaning that split-time calibrations cannot always be directly converted to genetic divergences.

A second common approach, which has only become possible within the last few years, is to count newly occurring mutations in deep sequencing data from family pedigrees, especially parent-child trios [6–9]. This approach provides a direct estimate but can be technically challenging, as it is sensitive to genotype accuracy and data processing from high-throughput sequencing. In particular, sporadic sequencing and alignment errors can be difficult to distinguish from true *de novo* mutations. Surprisingly, these sequencing-based estimates have consistently been much lower than those based on the first approach: in the neighborhood of 1–1.2 $\times 10^{-8}$ per base per generation, as opposed to 2–2.5 $\times 10^{-8}$ for those from long-term divergence [1–3].

A third method, which is also only now becoming possible, is to make direct comparisons between present-day samples and securely-dated ancient genomes. This method is similar to the first one, but by using two time-separated samples from the same species, it avoids the difficulty of needing an externally inferred split time. A recent study of a high-coverage genome sequence from a 45,000-year-old Upper Paleolithic modern human produced two estimates of this type [10]. Direct measurement of decreased mutational accumulation in this sample led to rate estimates of 0.44–0.63 $\times 10^{-9}$ per base per year (range of 14 estimates), or 1.3–1.8 $\times 10^{-8}$ per base per generation (assuming 29 years per generation [11]). An alternative technique, leveraging time shifts in historical population sizes, yielded an estimate of 0.38–0.49 $\times 10^{-9}$ per base per year (95% confidence interval), or 1.1–1.4 $\times 10^{-8}$ per base per generation, although a re-analysis of different mutational classes led to a total estimate of 0.44–0.59 $\times 10^{-9}$ per base per year (1.3–1.7 $\times 10^{-8}$), in better agreement with the first technique [10].

A fourth way to estimate the mutation rate is to calibrate the rate of accumulation of mutations using another clock that is better measured. For example, one study [12] used a model coupling single-nucleotide changes to the mutation of nearby microsatellite alleles to infer a single-nucleotide rate of 1.4–2.3 $\times 10^{-8}$ per base per generation (90% confidence interval).

In this study, we present a new approach that falls into this fourth category: we calibrate the mutation rate against the rate of meiotic recombination events, which has been measured with high precision in humans [13–15]. Intuitively, our method makes use of the following relationship between the mutation and recombination rates. At every site i in a diploid genome, the two copies of the base have some time to most recent common ancestor (TMRCA) T_i, measured in generations. The genome can be divided into blocks of sequence that have been inherited together from the same common ancestor, with different blocks separated by ancestral recombinations. If a given block has a TMRCA of T and a length of L bases, and if μ is the per-generation mutation rate per base, then the expected number of mutations that have accumulated in either copy of that block since the TMRCA is $2TL\mu$. This is the expected number of heterozygous sites that we observe in the block today (disregarding the possibility of repeat mutations). We also know that if the per-generation recombination rate is r per base, then the expected length of the block is $(2Tr)^{-1}$. Thus, the expected number of heterozygous sites per block (regardless of length or age) is μ/r.

This relationship allows us to estimate μ given a good prior knowledge of r. Our full method is more complex but is based on the same principle. We show below how we can capture the signal of heterozygosity per recombination to infer the historical per-generation mutation rate for non-African populations over approximately the last 50 thousand years (ky).

Results

Overview of methods

One difficulty of the simple method outlined above is that in practice we cannot accurately reconstruct the breakpoints between adjacent non-recombined blocks. Instead, we use an indirect statistic that captures information about the presence of breakpoints but can be computed in a simple way (without directly inferring blocks) and averaged over many loci in the genome (Figure 1).

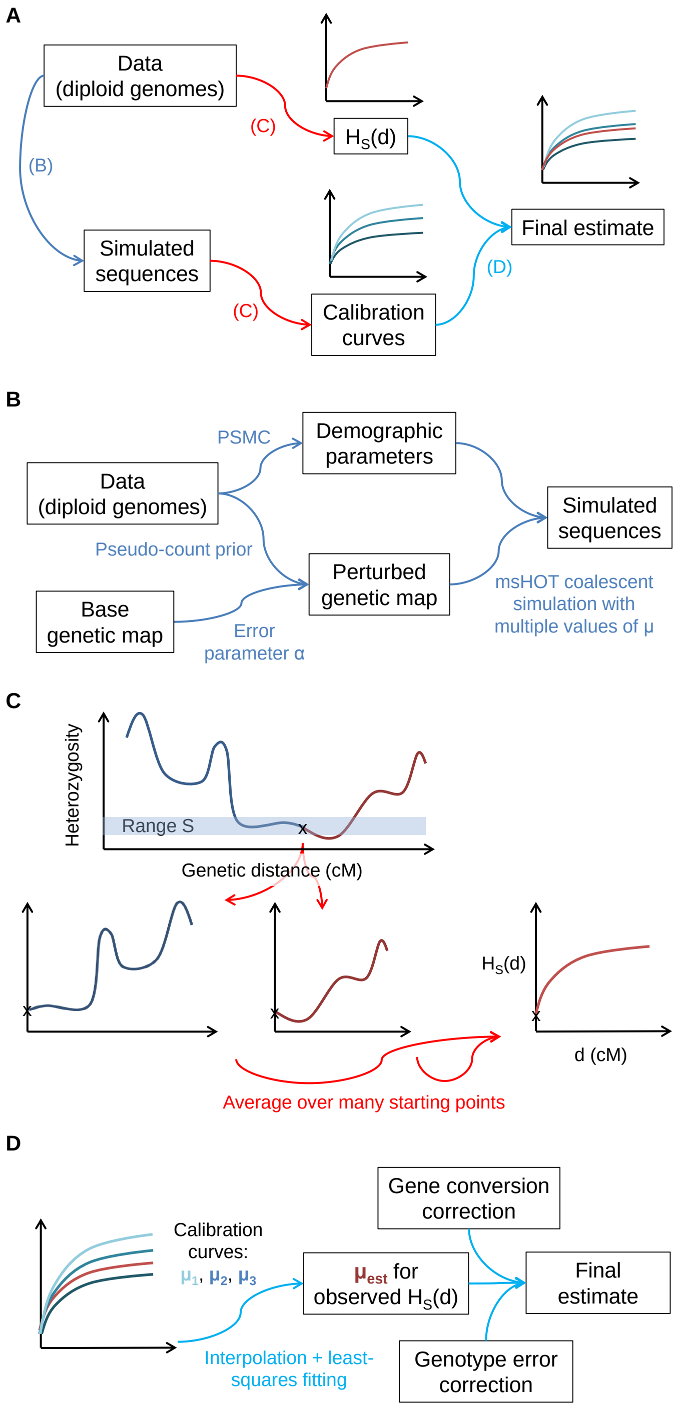

Figure 1. Illustration of the steps of our inference procedure. (A) Overview: from the data, we compute both the statistic $H_S(d)$ and other parameters necessary to create matching calibration curves with known values of μ. (B) Details of capturing aspects of the real data for the calibration data. (C) Computation of $H_S(d)$: the statistic captures the average heterozygosity as a function of genetic distance d from a starting point with heterozygosity in a defined range S, averaged over many such points. (D) For the final inferred value of μ, we compare matched $H_S(d)$ curves for the real data and calibration data (with known values of μ).

We define a statistic $H_S(d)$ equal to the average heterozygosity (local proportion of heterozygous sites) as a function of genetic distance d (measured in cM) from a set S of starting points within a collection of diploid genomes (Figure 1C). This statistic takes the form of a decay curve, with the rate of decay informative of the average TMRCA at the starting points (via the recombination rate), and the starting heterozygosity in turn informative of the mutation rate (see Methods). (While $H_S(d)$ is an increasing function of d in all of the examples we consider, we refer to the curve as a decay in the sense of an approach from the starting value $H_S(0)$ toward an asymptote at the genome-wide average heterozygosity.) We take S to be a set of points with similar local heterozygosities, indexed by the number of heterozygous sites per 100 kb (e.g., $H_{5\text{–}10}(d)$ when using starting points with a local total of 5–10 heterozygous sites per 100 kb). The TMRCAs of these points determine the time scale over which our inferred value of μ is measured. Since $H_S(d)$ cannot be expressed in closed form, our inference procedure involves creating matching "calibration data" with known values of μ with a coalescent simulator and then solving for the best-fit mutation rate for the test data (see Methods and Figure 1D).

In order for our inferences to be accurate, the calibration curves must recapitulate as closely as possible all aspects of the real data that could affect the shape of $H_S(d)$ (Figure 1B). Most notably, we apply techniques to infer fine-scale parameters for the demographic history and genetic map error of our samples, we correct for gene conversion and genotype errors, and we account for statistical uncertainty using a block jackknife (see Methods). We also test additional potential model violations through simulations (see Text S1 and Figure S1).

Simulations

First, for seven different scenarios, including a range of possible model violations (see Methods and Text S1), we generated 20 simulated diploid genomes with a known true mutation rate ($\mu = 2.5 \times 10^{-8}$ per generation except where otherwise specified) and ran our procedure as we would for real data, with perturbed genetic maps for both the test data and calibration data (variance parameter $\alpha = 3000\ \mathrm{M}^{-1}$; see Methods). In all cases, the $H_{5\text{–}10}(d)$ curves matched quite well between the test data and the calibration data (Figure 2). To measure the uncertainty in our estimates, we performed 25 independent trials of each simulation, and we also compared the standard deviations of the estimates across trials with jackknife-based standard errors (which are how we measure uncertainty for real data; see Methods).

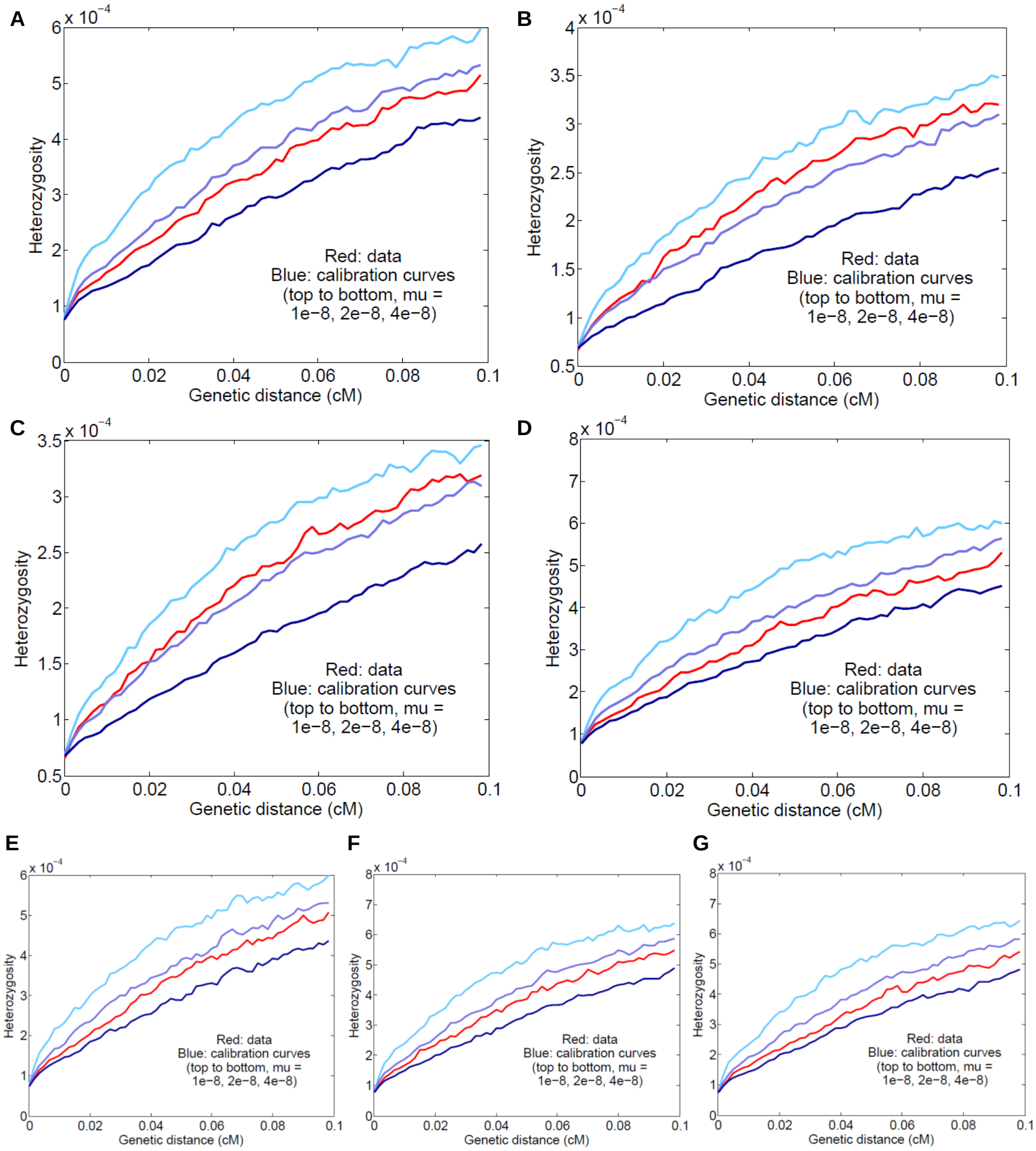

Figure 2. Results for simulated data. Means and standard deviations of 25 independent trials are given, and the curves displayed are for representative runs matching the 25-trial means. The true simulated rate is $\mu = 2.5 \times 10^{-8}$ unless otherwise specified. (A) Baseline simulated data; the inferred rate is $\mu = 2.47 \pm 0.05 \times 10^{-8}$. (B) Basic simulated data with a true rate of 1.5×10^{-8}; the inferred rate is $\mu = 1.61 \pm 0.05 \times 10^{-8}$. (C) Data with a true rate of 1.5×10^{-8} plus gene conversion; the inferred rate is $\mu = 1.60 \pm 0.05 \times 10^{-8}$ (corrected from a raw value of 1.70×10^{-8} with gene conversion included). (D) Data with simulated genotype errors; the inferred rate is $\mu = 2.39 \pm 0.06 \times 10^{-8}$ (corrected from a raw value of 2.71×10^{-8} with genotype errors included). (E) Data simulated with variable mutation rate; the inferred rate is $\mu = 2.61 \pm 0.08 \times 10^{-8}$. (F) Data from a simulated admixed population; the inferred rate is $\mu = 2.57 \pm 0.07 \times 10^{-8}$. (G) Simulated data with all three complications as in (D)–(F); the inferred rate is $\mu = 2.53 \pm 0.06 \times 10^{-8}$ (corrected from a raw value of 2.77×10^{-8}).

In all cases, our results were in good statistical agreement with the true rate (Figure 2), with only one inferred rate slightly outside two standard errors from the simulated parameter. Full details of the simulations can be found in Methods and Text S1. For the seven scenarios, we inferred values (mean ± standard error) of (a) $\mu = 2.47 \pm 0.05 \times 10^{-8}$ (basic model); (b) $\mu = 1.61 \pm 0.05 \times 10^{-8}$ (basic model, true rate of $\mu = 1.5 \times 10^{-8}$); (c) $\mu = 1.60 \pm 0.05 \times 10^{-8}$ (true rate of $\mu = 1.5 \times 10^{-8}$, plus gene conversion); (d) $\mu = 2.39 \pm 0.06 \times 10^{-8}$ (simulated genotype errors); (e) $\mu = 2.61 \pm 0.08 \times 10^{-8}$ (mutation rate variability); (f) $\mu = 2.57 \pm 0.07 \times 10^{-8}$ (admixture); and (g) $\mu = 2.53 \pm 0.06 \times 10^{-8}$ (all three complications (d)–(f)). Furthermore, our jackknife estimates of the standard error were comparable to the realized standard deviations and on average conservative, especially for the most complex simulation (g), despite not incorporating PSMC uncertainty (see Methods): 0.08×10^{-8}, 0.05×10^{-8}, 0.06×10^{-8}, 0.06×10^{-8}, 0.09×10^{-8}, 0.05×10^{-8}, and 0.11×10^{-8}, respectively, for (a)–(g). The fact that all of the inferred rates are close to the true value leads us to conclude that none of the aspects of the basic procedure or the tested model violations—or, by extension, natural selection, which has similar effects to rate and demographic heterogeneity (see Text S1)—create a substantial bias. We did find evidence of small biases for different starting points with all three model violations included; namely, we obtained point estimates of $\mu = 2.13 \times 10^{-8}$ (with 40 genomes' worth of calibration data) for $S = 1$–5 and $\mu = 2.81 \times 10^{-8}$ for $S = 10$–20. For real data, however, we restrict our primary analyses to starting points with local heterozygosities of 5–10 per 100 kb.

Error parameters

Before obtaining mutation rate estimates from real data, we quantified two important error parameters: the rate of false heterozygous genotype calls and the degree of inaccuracy in our genetic map.

We estimated the genotype error rate by comparing the ratio of observed transitions and transversions in windows of low heterozygosity to the genome as a whole, taking advantage of the excess of transversions among false heterozygous sites (Table 1; see Methods). As expected under the assumption that the error rate is independent of the local coalescent history, there are relatively more transversions (i.e., more errors as a fraction of the total heterozygous sites) for lower heterozygosities. The inferred error rates were slightly different for the different heterozygosity bins, which is not surprising given the ascertainment of the windows. However, all were in the range of 1 per 150–200 kb, consistent with previous results [16]. We used these numbers (the first three lines in Table 1), together with a previous estimate of the rate of gene conversion [17] (see Methods), to correct our estimates of μ below.

Table 1. Transition/transversion ratio and genotype error correction for ascertained genomic windows

Hets per 100 kb	Tr/tv ratio	Error rate	Correction factor
1–5	1.74	4.99×10^{-6}	0.93
5–10	1.87	6.37×10^{-6}	0.91
10–20	1.99	5.69×10^{-6}	0.92
1–20	1.88	5.70×10^{-6}	–
All	2.12	5.70×10^{-6}	–

Observed transition/transversion ratios, with inferred error rates (proportion of false heterozygous sites), for ascertained windows of eight genome sequences of non-African individuals. See Methods for details of calculations. The correction factor is the multiplier applied to the raw estimate of μ (see Methods); we note that this quantity varies slightly when using different sets of test data. See Table S1 for full site counts.

It was also necessary for us to estimate the accuracy of our genetic map (see Methods). We used the "shared" version of the African-American (AA) map from [15] as our base map and a modified version of the error model of [18]: $Z \sim \text{Gamma}(\alpha\gamma(g+\pi p), \alpha)$, where

Z is the true genetic length of a map interval, g is the observed genetic length, p is the physical length, α is the parameter measuring the accuracy of the map, and γ and π are constants. Based on pedigree crossover data from [19], we estimated $\alpha = 2802 \pm 14$ M^{-1} for the full AA map and $\alpha = 3414 \pm 13$ M^{-1} for the "shared" map. For our analyses, we took the intermediate value $\alpha = 3100$ M^{-1} (see Methods). This means that $1/\alpha \approx 0.03$ cM can be thought of as the length scale for the accuracy of genetic distances according to the base map (see Methods for details). We also inferred a pseudo-count prior of $\pi = 9 \times 10^{-5}$ cM/kb (see Methods).

We note that the values of α reported in [18] are substantially lower than ours, which we suspect is because our cross-checking data have much finer resolution than those used previously. (When using the same cross-checking data, the "shared" and Oxford LD maps appear to be relatively similar in accuracy.) If we substitute our new α values for the original application of inferring the date of Neanderthal gene flow into modern humans, we obtain a less distant time in the past, 28–65 ky (most likely 35–49 ky), versus 37–86 ky (most likely 47–65 ky) reported in [18]. While relatively recent, this date range is not in conflict with archaeological evidence or with an estimate of 49–60 ky (95% confidence interval) based on an Upper Paleolithic genome [10].

Estimates for Europeans and East Asians

Our primary results for real data (Figure 3) were obtained from European and East Asian individuals, a total of eight genomes (two each French, Sardinian, Han, and Dai; see Methods). As above, all results are given in the form of mean ± standard error. With all eight individuals combined to maximize the signal quality, we estimated a mutation rate of $\mu = 1.65 \pm 0.10 \times 10^{-8}$ per generation (Figure 3A) using our standard parameter settings and regions with 5–10 heterozygous sites per 100 kb.

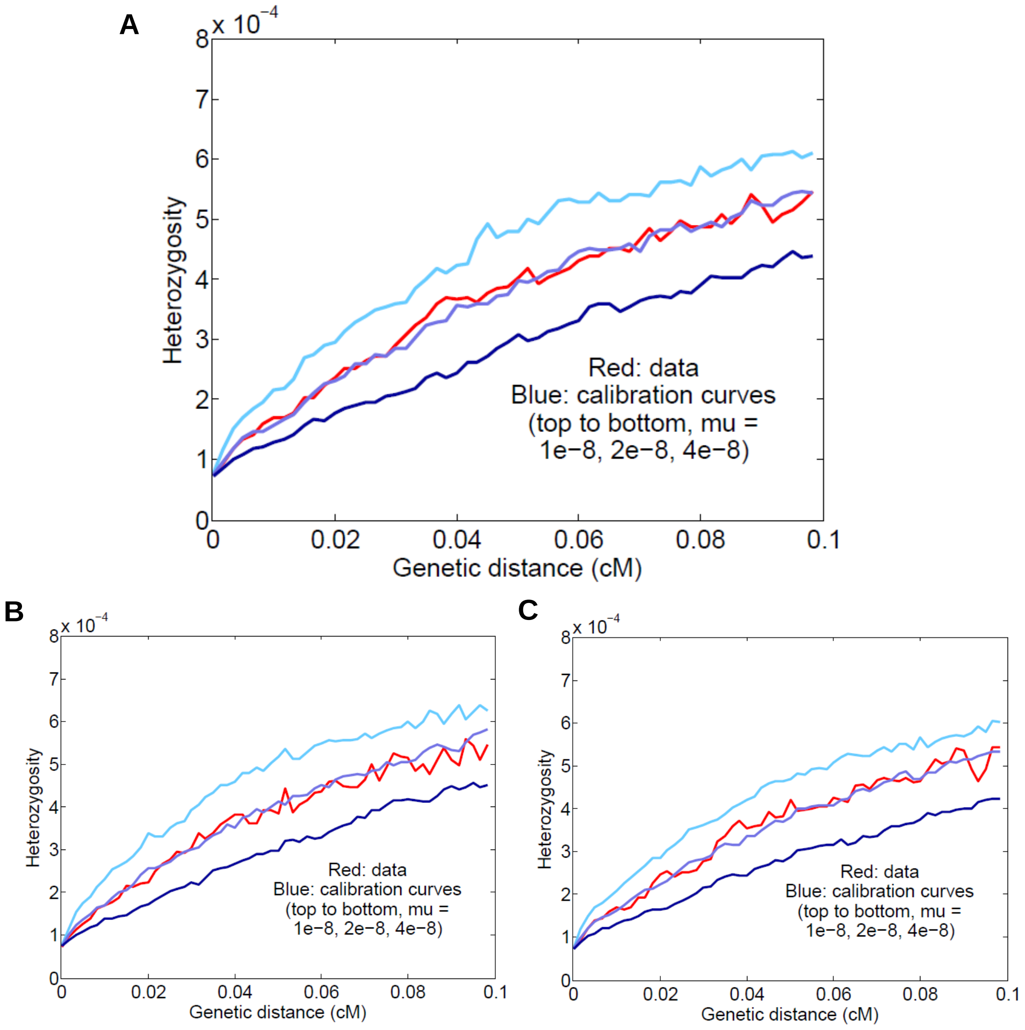

Figure 3. Results for Europeans and East Asians. All results for real data are corrected for gene conversion and genotype error, which explains the apparent discrepancy between the final estimates and the curves shown here. (A) All eight individuals together; the inferred rate is $\mu = 1.65 \pm 0.10 \times 10^{-8}$ per generation (corrected from a raw value of 1.95×10^{-8}). (B) Results for the four Europeans; the inferred rate is $\mu = 1.80 \pm 0.11 \times 10^{-8}$. (C) Results for the four East Asians; the inferred rate is $\mu = 1.65 \pm 0.12 \times 10^{-8}$.

It is possible that our full estimate could be slightly inaccurate due to population-level differences in either the fine-scale genetic map or demographic history (see Text S1). However, we expect Europeans and East Asians to be compatible in our procedure both because they are not too distantly related and because they have similar population size histories [20, 21]. To test empirically the effects of combining the populations, we estimated rates for the four Europeans and four East Asians separately (Figure 3B–C). Using the same genotype error corrections, we found that the $H_{5\text{–}10}(d)$ curves as well as the final inferred values were similar to those for the full data: $\mu = 1.80 \pm 0.11 \times 10^{-8}$ for Europeans and $\mu = 1.65 \pm 0.12 \times 10^{-8}$ for East Asians. Thus, we believe that the full eight-genome estimate is robust to the effects of population heterogeneity.

We also attempted to account for uncertainties in our genetic map error model. First, we repeated our eight-genome estimate with a range of alternative values of α and found that the inferred value of μ only varied on the order of 10% (Figure S2). We also used this set of inferred rates to derive an empirical estimate of $\alpha = 3760$ M^{-1} based on a second method for quantifying map error (see Text S1). The accompanying mutation rate estimate of $\mu = 1.79 \times 10^{-8}$ was only modestly higher than our primary value of 1.65×10^{-8}. Second, to mimic our uncertainty due to the exact form of the true genetic map, we generated 25 additional full-data estimates using different perturbed maps. We took the observed standard deviation of 0.034×10^{-8} to be an appropriate measure of this uncertainty and incorporated this quantity into all of our reported standard errors.

Finally, although we are most confident in our results with $S = 5$–10 per 100 kb, we obtained independent estimates of μ using different starting heterozygosities $S = 1$–5 and 10–20, reflecting average mutation rates over different time depths in the past. We inferred values of $\mu = 1.72 \pm 0.12 \times 10^{-8}$ and $\mu = 1.56 \pm 0.14 \times 10^{-8}$, respectively (Figure S3), in close agreement with our primary results.

Estimates for other populations

We also ran the procedure for three other non-African populations: aboriginal Australians, Karitiana (an indigenous group from Brazil), and Papua New Guineans. We used two genomes per population and computed curves for starting regions with 1–15 heterozygous sites per 100 kb. Since our genotype error estimates for $S = 1$–20 and 10–20 are almost identical (Table 1), we used this common value for the error correction. We inferred rates of $\mu = 2.00 \pm 0.17 \times 10^{-8}$, $\mu = 1.50 \pm 0.17 \times 10^{-8}$, and $\mu = 1.74 \pm 0.15 \times 10^{-8}$ for Australian, Karitiana, and Papuan, respectively (Figure 4). The latter two estimates differ only slightly from those given above, while Australians have a near-significantly higher per-generation value, which is consistent with the high average ages of fathers in many aboriginal Australian societies [11,22]. Overall, given the expected small differences for historical, cultural, and/or biological reasons (including, as mentioned above, our use of the same "shared" genetic map for all groups), we do not see evidence of substantial errors or biases in our procedure when applied to diverse populations.

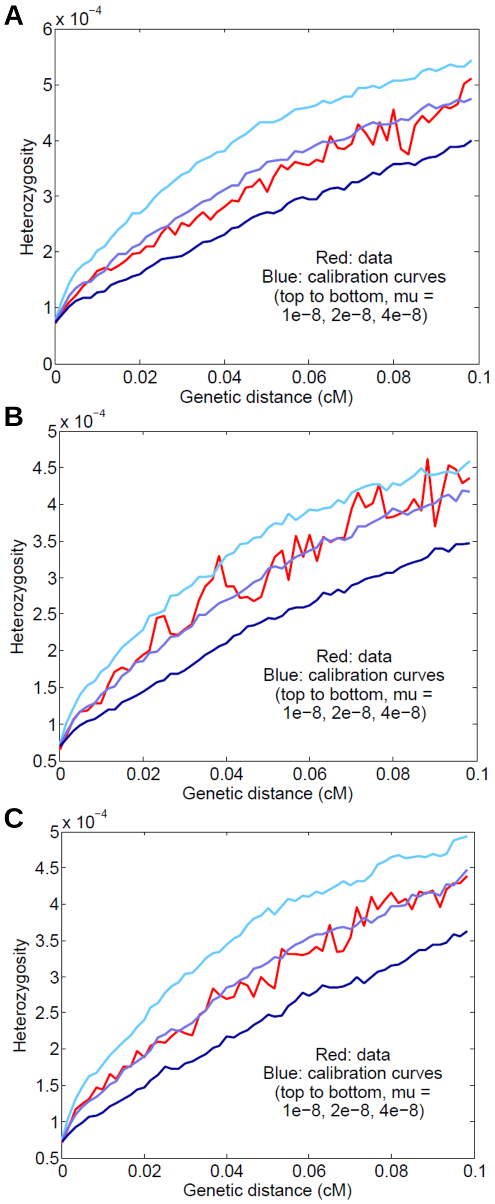

Figure 4. Results for other populations. (A) Australian, $\mu = 2.00 \pm 0.17 \times 10^{-8}$. (B) Karitiana, $\mu = 1.50 \pm 0.17 \times 10^{-8}$. (C) Papuan, $\mu = 1.74 \pm 0.15 \times 10^{-8}$.

Discussion

Using a new method for estimating the human mutation rate, we have obtained a genome-wide estimate of $\mu = 1.65 \pm 0.10 \times 10^{-8}$ single-nucleotide mutations per generation. Our approach counts mutations that have arisen over many generations (on the order of a thousand, or tens of thousands of years) and relies on our excellent knowledge of the human recombination rate to calibrate the length of the relevant time period.

We have shown that our estimate is robust to a large number of possible confounding factors (Figure S1). In addition to statistical noise in the data, our method directly accounts for ancestral gene conversion and for errors in genotype calls and in the genetic map. We have also demonstrated, based on simulations, that heterogeneity in demographic and genetic parameters, including the mutation rate itself, does not cause an appreciable bias, while empirical results for different populations and different starting points are likewise highly concordant. We acknowledge that there could still be some residual sources of error in our final estimates, but given our simulation results, we believe that our reported standard errors accurately capture our uncertainty.

The meaning of an average rate

It is important to note that the mutation rate is not constant at all sites in the genome. As we have discussed, we believe that this variability does not cause a substantial bias in our inferences, but to the extent that some bases mutate faster than others, a rate is only meaningful when associated with the set of sites for which it is estimated. For example, methylated cytosines at CpG positions accumulate point mutations roughly an order of magnitude faster than other bases because of spontaneous deamination [3, 7, 8]. Such effects can lead to larger-scale patterns, such as the higher mutability of exons as compared to the genome as a whole [23],

In our work, we filter the data substantially, removing more than a third of the sites in the genome. The filters tend to reduce the heterozygosity of the remaining portions [21,24], which is to be expected if they have the effect of preferentially removing false heterozygous sites. It is important to remember that, strictly speaking, our rates only correspond to a subset of the genome. For reference, in Table S2, we present heterozygosity levels and human–chimpanzee divergence statistics for sites passing our filters.

Evolutionary implications and comparison to previous estimates

A key application of the mutation rate is for determining the divergence times of human populations from each other and of humans from other species [1]. Published mutation rate estimates are highly discrepant, however, by as much as a factor of 2 between those based on sequence divergence among great apes (2–2.5 $\times 10^{-8}$ per base per generation) versus *de novo* mutations in present-day families (generally 1–1.2 $\times 10^{-8}$) [1–3]. This uncertainty causes estimated population split times to be highly dependent on whether a high or low rate is assumed.

Given this context, our findings are notable for supporting an intermediate rate: $1.65 \pm 0.10 \times 10^{-8}$ per base per generation, or $0.57 \pm 0.04 \times 10^{-9}$ per base per year, assuming an average generation time of 29 years [11]. This value is also in good agreement with a previous estimate based on linked microsatellite mutations (0.5–0.8 $\times 10^{-9}$) [12] and at the high end of the range inferred from comparisons between modern samples and a Paleolithic modern human genome (0.4–0.6 $\times 10^{-9}$) [10]. Here we discuss some of the consequences of our results in relation to prior studies.

With regard to long-term sequence divergence, we believe that our inferred rate is consistent with the available paleontological evidence. Importantly, great ape species (e.g., human, chimpanzee, and gorilla) exhibit a large amount of incomplete lineage sorting, which is indicative of large population sizes and high rates of polymorphism in their common ancestors [25, 26]. This results in substantial differences between genetic divergence times and the final split times of species pairs. For example, according to a recent estimate [26], assuming a "fast" rate of 1.0×10^{-9} mutations per base per year results in an estimated population split time of ~ 3.8 million years (My) for humans and chimpanzees (with a genetic divergence time approximately 50% older), which seems too recent in light of the fossil record. By contrast, our rate of roughly 0.6×10^{-9} (acknowledging the complications in converting between per-generation and per-year rates [3]) results in a more reasonable split time of 6.6 ± 0.4 My (or 7.1 ± 0.5 My using our filtered subset of the genome; see above and Table S2). Perhaps a stronger constraint is the human–orangutan split, which is believed to be no older than about 16 My [1, 27]. In this case, our rate implies a split time of 19.6 ± 1.3 My (with a genetic divergence time substantially older, approximately 28 My) [26]. Although this appears to predate the fossil-inferred split (with some uncertainty), it is reasonable to expect modest changes in the biology (and, specifically, the mutation rate) of ancestral apes over this time scale (and likewise for older splits [1]). By comparison, however, a rate of 0.4×10^{-9} per year from *de novo* studies implies a much more discrepant split time of ~ 28 My.

Our results can also be assessed in terms of their implications for split times among modern human populations. It has been argued on the basis of results from demographic models that a slower rate fits better with our knowledge of human history [1, 28]. For example, a recent method for estimating population split times from coalescent rates placed the median split of African from non-African populations at 60–80 ky and the split of Native Americans from East Asians at ~ 20 ky, both assuming a per-generation mutation rate of 1.25×10^{-8} and an average generation interval of 30 years [28]. While both the model and the histories of the populations involved are somewhat complicated, it does seem unlikely that these dates could be half as old (30–40 ky and 10 ky), as would be required for a rate of 2.5×10^{-8}. Using our inferred rate also makes the dates more recent, but only by a factor of about 1.3 rather than 2, i.e., $\sim$ 46–61 and 15 ky (with some associated uncertainty both from the model and from our estimated rate), neither of which contradicts external evidence.

A possible explanation for the discrepancy between our results and those of trio sequencing studies is that because it is very difficult to separate true *de novo* mutations from genotype errors in single-generation data, some mutations have been missed in previous work. For example, three recent exome-sequencing studies [29–31], which estimated effective genome-wide mutation rates of approximately 1.5×10^{-8}, 1.2×10^{-8}, and 1.35×10^{-8} per base per generation, found in follow-up validation that, in addition to virtually all sites from their filtered data sets, many putative sites that did not pass all filters (roughly 70%, 20%, and 90% of tested sites, respectively) were confirmed as true *de novo* mutations. These results suggest that there may be a subset of *de novo* mutations having low quality metrics that are missed in trio-based counts as a result of the filtering that is necessary to remove genotype errors. It would be of interest to carry out larger-scale follow-up validation to test if this is the case. In theory, filtered sites can be accounted for by adjusting the denominator in the final rate calculation, but it seems possible that site-level filters preferentially remove *de novo* mutations or that the resulting denominators have not been fully corrected, or both.

Another possibility is that the mutation rate could have changed over time. It has been suggested, for example, that phylogenetic and pedigree-based estimates could be reconciled if the rate has recently slowed in extant great apes [1]. However, it seems unlikely that the underlying biology of the mutation process would have changed significantly over the last few tens of thousands of years, which is the time scale covered by our estimate. On the other hand, at least part of the discrepancy between our results and previous estimates could plausibly relate to changes in generation intervals. While the profile of germline mutations as a function of parental age is complicated and not fully understood [3], it is well established that most inherited mutations are paternal in origin and that they are more numerous for older fathers [8, 32]. As a result, changes in the sex-averaged generation interval only weakly impact per-year mutation rates, and moreover, present-day sex-averaged generation intervals are very similar among different countries and cultures, including hunter-gatherer societies [11]. However, surveyed hunter-gatherers, who may serve as the best comparison for our long-term rate estimate, on average have high paternal ages—about 32.3 years [11], as compared to a range in *de novo* studies from less than 27 [9] to around 30 [8, 30] to more than 33 [29]. This could lead to a higher long-term mutation rate. While this effect would likely only account for a portion of the discrepancy, it demonstrates the potential influence of cultural and demographic shifts. Going forward, we expect that new data, technologies, and analytical techniques will continue to add to our knowledge of the human mutation rate, both in the precision of estimates of its long-term average and in its variability over time, by age and sex, and in different populations.

Methods

Definition of the statistic $H_S(d)$

Our inferences are based on a statistic that allows us to compare the mutation rate to the (much better measured) recombination rate. Intuitively, we compare local levels of heterozygosity in diploid genomes to the distance scales over which these levels change;

the former are proportional to the mutation rate and the latter to the recombination rate, with the same constant of proportionality.

Starting from a certain position in the genome, the TMRCA of the two haploid chromosomes as a function of distance in either direction is a step function, with changes at ancestral recombination points (Figure 5A). Heterozygosity, being proportional to TMRCA in expectation (and directly observable), follows the same pattern on average (Figure 5B).

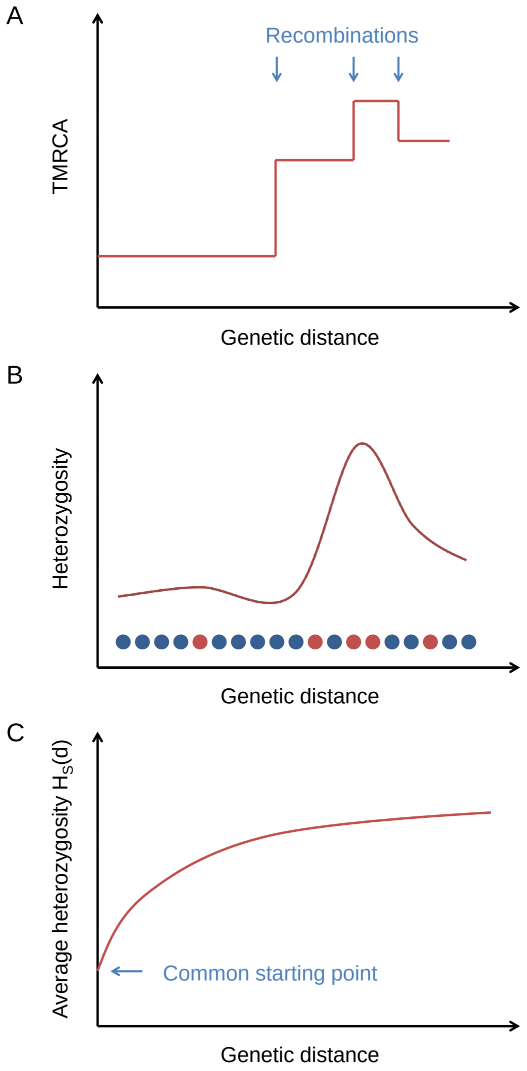

Figure 5. Explanation of the statistic $H_S(d)$. (A) Ancestral recombinations separate chromosomes into blocks of piecewise-constant TMRCA (and hence expected heterozygosity). (B) From the data, we measure local heterozygosity as a function of genetic distance; red and blue circles represent heterozygous and homozygous sites, respectively, along a diploid genome. (C) Our statistic $H_S(d)$ is an average heterozygosity as a function of genetic distance over many starting points with similar local heterozygosities, yielding a smooth decay toward the genome-wide average.

We can also compute heterozygosity by distance averaged over a collection of starting positions within the genome (and over a collection of multiple genomes). If these points have similar local heterozygosities, then as a function of the genetic distance d away from them, the average heterozygosity displays a smooth decay from the common starting value toward the global mean heterozygosity $\bar{H}$ as the probability increases of having encountered recombination points (Figure 5C). We define a statistic $H_S(d)$ that equals this average heterozygosity as a function of genetic distance from a set S of suitably ascertained starting points (see below for more details). In practice, we compute $H_S(d)$ in 60 distance bins from 0 to 0.1 cM.

Most importantly for our purposes, the probability of having encountered a recombination is a function of both d and the starting heterozygosity $H_S(0)$, since smaller values of $H_S(0)$ correspond to smaller TMRCAs, with less time for recombination to have occurred, and hence longer unbroken blocks. This relationship, with lower starting heterozygosity corresponding to a slower decay rate of $H_S(d)$, is what allows us to calibrate μ against the recombination rate r.

Locating starting points

In order to maximize signal quality, we would like to measure $H_S(d)$ using many starting points in the genome, but within a relatively narrow range of local heterozygosity at those points. For our primary analyses, we use a starting heterozygosity value $H_S(0) \approx 7.5 \times 10^{-5}$, corresponding to points with TMRCA roughly one-tenth of the genome-wide average for non-African individuals. This choice of S has two main advantages. First, there are relatively many such points in genomes of non-African individuals because it corresponds approximately to the age of the out-of-Africa bottleneck. Second, a smaller $H_S(0)$ corresponds to a slower and higher-amplitude (from a lower starting value to the same asymptote) decay of $H_S(d)$, making the curve easier to fit and less susceptible to genetic map error, an important consideration for our method (see below).

Our means of determining the local heterozygosity is very simple. Rather than try to compute heterozygosity precisely by delineating non-recombined blocks, we tile the genome with 100-kb regions and count the proportion of heterozygous sites within each. The starting points used to compute $H_S(d)$ are then the midpoints of the 100-kb regions having a heterozygosity at the desired level, for example 5–10 $\times 10^{-5}$ for out-of-Africa-age blocks with $H_S(0) \approx 7.5 \times 10^{-5}$. We use the notation (for example) $H_{5\text{–}10}(d)$ to denote an $H_S(d)$ curve computed for starting points with 5–10 heterozygous sites per 100 kb. This scheme may result in choosing starting points with unwanted true heterozygosity if there are recombinations within the 100-kb region, but 100 kb is long enough that most regions within a narrow range of heterozygosity on that scale should be similarly behaved. Additionally, any deviations from the desired heterozygosity range at the midpoints should, on average, be the same for real and simulated data (assuming that the simulator accurately models the genealogical process with recombination; see below), and hence would only cause noise rather than bias in the estimated mutation rate. Similarly, while the relationship between observed heterozygosity and TMRCA is non-linear because of randomness in the number of accumulated mutations, it is the same in real and simulated data. As an attempt to avoid certain kinds of undesirable behavior (for example, a very low heterozygosity over most of the region and a recombination near one end followed by high heterozygosity), we also require at least one heterozygous site in each half.

Inference strategy

As described above, $H_S(d)$ exhibits a decay as a function of d as a result of ancestral recombination events. Recombination can be modeled as a Poisson process (in units of genetic distance), but $H_S(d)$ does not have an exponential functional form, because the TMRCAs T_1 and T_2 at two loci separated by a recombination event are not independent

[20,33,34]. First, both T_1 and T_2 must be older than the time at which the recombination occurred, which imposes different constraints on T_2 for different values of T_1. This dependence becomes especially complicated when the population from which the chromosomes are drawn has changed in size over time. Second, the coalescence at time T_2 can involve additional lineages in the ancestral recombination graph, making the expected time different than would be true for two lineages in isolation. For example, with some probability, the two lineages split by the recombination can coalesce together more recently than this combined lineage coalesces with the second chromosome, in which case $T_1 = T_2$.

These complicating factors mean that $H_S(d)$ cannot be described as a closed-form function of d. However, we know that $H_S(d)$ decays from $H_S(0)$ toward the average heterozygosity $\bar{H}$, and the rate of decay is governed by the relationship between μ and r. Thus, our strategy is to infer the true value of μ by simulating sequence data matching our real data in all respects (see below for more detail) and with a range of different values of μ ($\mu = 1, 2, 4 \times 10^{-8}$). Then, we can compare the observed $H_S(d)$ curve to the same statistic calculated on each simulated data set and infer μ by finding which value gives the best match. Computationally, we interpolate the observed $H_S(d)$ curve between the simulated ones (from $d = 0$ to 0.1 cM), parametrized by the simulated μ values (we use the `MATLAB` "spline" interpolation method; for our main results, other methods differ by less than 1%). Finally, we perform variance-weighted least-squares to find the single best-fit value of μ.

Population size history

We estimate the historical population sizes for the sampled chromosomes with PSMC [20]. PSMC returns parameters in coalescent units: the scaled mutation rate $\theta = 4N\mu$, the scaled recombination rate $\rho = 4Nr$, and population sizes going back in (discretized) time, with both the sizes and time intervals in terms of the scaling factor N (the baseline total population size). We do not know N, but the inferred θ together with the population size history are exactly what we need in order to simulate matching data for the calibration curves. We do not use the inferred value of ρ but rather set $\rho = \theta r/\mu$, where r is the true per-base recombination rate and μ is the fixed mutation rate for a given calibration curve. This maintains the proper ratio between r and μ for that curve, as well as the proper diversity parameter θ. While we only use short regions of the genome in computing $H(d)$, we run PSMC on the full genome sequences. The exception is that when testing the method in simulations, we run PSMC on the simulated segments, as this is all that is available to us (see below).

Genetic map error

The statistic $H_S(d)$ is computed as a function of genetic distance, which we obtain from a previously-estimated genetic map. However, while map distances (i.e., local recombination rates) are known much more precisely than mutation rates, there is still some error in even the best maps, which we must account for in our inferences.

As with other variables, our approach is not to make a direct correction for map error but rather to include it in a matching fashion in the calibration data. We first select a baseline genetic map from the literature, and we plot $H_S(d)$ as a function of d using this base map as the independent variable. To make the calibration curves match the real data, whose intrinsic, true map does not match the base map exactly, we simulate the calibration data using a perturbed version of the base map, with the aim of capturing an equal amount of deviation as in the real data.

The base map we use is the "shared" version of an African-American (AA) genetic map published in [15]. The AA map was derived by tabulating switch points between local African and European ancestry in the genomes of African-Americans, which reflect recombination events since the time of admixture. The "shared" component of the map was estimated as the component of this recombination landscape that is active in non-Africans, particularly Europeans. From our experience, this map is the most accurate currently available for non-Africans.

The key to our approach is in the ability to quantify the amount of error in the base map. In what follows, we describe in detail our methods to measure the degree of genetic map error.

Basic model and previous estimates

Our basic error model is that of [18]. Consider a chromosomal interval whose true genetic length is Z. Given the measured genetic length g of the interval in our base map, we assume that $Z \sim \text{Gamma}(\alpha g, \alpha)$, so $E[Z|g] = g$ and $\text{var}(Z|g) = g/\alpha$. The gamma distribution has several desirable properties, in particular that it is scale-invariant. The parameter α measures the per-distance variance of the map (and hence of our perturbation), with a larger value of α corresponding to a smaller variance. It can be interpreted directly as the inverse of the length scale at which the coefficient of variation of the map equals 1: on average, intervals of length $1/\alpha$ have an equal mean and standard deviation, while longer intervals are relatively more accurately measured and shorter intervals are less accurate.

This model was previously used to estimate α for two genetic maps [18]: the 2010 deCODE map [14], which was estimated by observing crossovers in a large Icelandic pedigree cohort, and the Oxford LD map [13], which was estimated from variation in background LD levels in unrelated individuals. The authors used a validation data set with observed recombination events in a separate group of individuals [35] and specified a full probability model for those observations in terms of the error in the base map as well as other parameters. In per-Morgan units, they obtained $\alpha = 1400 \pm 100$ M^{-1} for the deCODE map and $\alpha = 1220 \pm 80$ M^{-1} for the Oxford LD map.

Here, we apply the same method to estimate α, but using a much larger set of validation data, consisting of 2.2 million crossovers from 71,000 meioses in Icelandic individuals [19] (versus 24,000 crossovers from 728 meioses in [35]). A potential complication is that the procedure to build the "shared" map [15] used information from the 2010 deCODE map, which is not independent from our validation set. However, we reasoned that we could constrain the true α by applying the estimation method first to the "shared" map as an upper bound (since the value will be inflated by the non-independence) and then to the full AA map as a lower bound (since the AA map includes African-specific hotspots not active in Europeans).

Modified prior distribution on Z

We also add one modification to the basic model of map error described above. For very short intervals in the base map, in particular those with estimated (genetic) length 0, the original model states that the true length of these intervals is 0 (since Z has mean $g = 0$ and $Z \geq 0$). In fact, though, the data used to build the map might simply have included no crossovers there by chance. Overall, very short intervals in the base map are likely underestimated, while very long intervals are likely overestimated.

To account for this effect, we modify the prior distribution on the true length Z by adding a pseudo-count adjustment, i.e., a small uniform prior on the true map length. In order for the model still to be additive, it is reasonable for the prior to be in units of genetic distance per physical distance.

Empirically, we observe that without the adjusted prior, the decay of $H_S(d)$ in the calibration curves is too slow at the smallest values of d and too fast at larger d (Figure S4). This would be expected if very short intervals in the base genetic map are underestimated, so that the calibration data have too few recombinations in that range compared to real data. By matching the curve shape of real data to the calibration data for different values of the pseudo-count, we can determine which value properly corrects for the underestimation of very short intervals in the "shared" map.

Perturbed genetic maps for calibration data

For our purposes, once we obtain a value for α, we use the gamma-distribution model to generate the randomized perturbed maps that we input to `msHOT` to generate the calibration data. Our complete error model is as follows: for an interval of physical length p, we take $Z \sim \text{Gamma}(\alpha g', \alpha)$, where $g' = \gamma(g + \pi p)$ for a pseudo-count π and the corresponding constant factor $\gamma < 1$ that preserves the total map length. Thus, for each interval in the map (between two adjacent SNPs on which the map is defined), if the physical length is p and the base genetic length is g, we create a new genetic length Z for that interval in the perturbed map by drawing from this distribution.

Also, we note that even if our map error model is properly specified and estimated, there could be a small bias in our final inferred μ if the exact form of the true map is such that the $H_S(d)$ curve decays slightly faster or slower than the calibration curves built with a perturbed map. Since this possible bias is analogous to variability in the inferred μ depending on the exact instantiation of the perturbed map, we quantify it by measuring the variability in the calibration results for different versions of the perturbed map.

Simulation of calibration data with `msHOT`

As discussed above, the decay of $H_S(d)$ reflects the decorrelation of heterozygosity as a function of genetic distance caused by recombination. However, in the sequence of TMRCAs for the recombination-separated blocks along a chromosome, successive values are not independent, and in fact the sequence is not Markovian, since even lineages that are widely separated along the chromosome can interact within the ancestral recombination graph [33, 34]. It is important, then, that our simulated data be generated according to an algorithm that captures all of the coalescent details that could impact the history of a real-data sample. For this reason we use `ms` [36] rather than a Markovian simulator, which would have had the advantage of greater speed. In fact, we run the extended software `msHOT` [37] to allow variable recombination rates matching the observed genetic map.

As result of using a non-Markovian simulator, it is computationally infeasible to generate entire simulated chromosomes. Thus, in practice we define wider "super-regions" around the 100-kb starting regions and simulate the super-regions independently of each other, matching the physical and genetic coordinates to the human genome. Since we compute $H_S(d)$ from $d = 0$ to 0.1 cM, we define the super-regions to include at least 0.1 cM on both sides of their internal starting point, which typically leads to a total length of several hundred kb per super-region.

Finally, in addition to matching the demographic and genetic map parameters of the calibration data to the test data, we also apply an adjustment to the calibration curves themselves to correct for residual unequal heterozygosity (not precisely captured by PSMC), which would cause the asymptotes of the curves to be mis-aligned. In particular, we multiply the decay portion of the calibration curves (i.e., $H_S(d) - H_S(0)$) by the ratio of the heterozygosity of the real data (over all of the super-regions) to that of the matching simulated data. In our experience, this correction ranges from 0–10%.

Gene conversion

Gene conversion is similar to recombination (as we have used the term; more precisely, crossing-over) in that it results from double-stranded breaks in meiosis and leads to the merging of genetic material between homologous chromosomes. However, whereas crossing-over creates large-scale blocks inherited from the recombining chromosomes, gene conversion occurs in very small tracts, on the order of 100 bases in humans [38]. For our purposes, gene conversion is significant primarily because it can introduce heterozygous sites in our test regions if one of the haplotypes has experienced a gene conversion event since the TMRCA.

We choose to account for the effect of gene conversion by applying a correction to our inferred mutation rates, reasoning that a subset of the observed heterozygous sites will be caused by gene conversion events rather than mutations. Since our method relies on the ratio between the recombination and mutation rates at the selected starting points, the key quantity is the proportion of heterozygous sites near those points that are due to gene conversion. Our correction amounts to subtracting the estimated rate of new gene conversion-derived heterozygous sites from the raw mutation rate estimate (assuming all heterozygous sites arose from mutation). This number is a combination of two factors: the probability that each base is involved in a gene conversion event and the conditional probability that a polymorphism is introduced. For the former, we use a recent estimate that gene conversion affects approximately 5.7×10^{-6} bases per generation (95% confidence interval 4.5–7.3 $\times 10^{-6}$) [17] and adjust for local recombination rate, and for the latter, we use differences between the heterozygosity in the test regions and in the genomes as a whole (see Text S1). The final correction ranges from 0.13–0.17 $\times 10^{-8}$ per base per generation (with a standard error of 0.02–0.03 $\times 10^{-8}$), approximately 7–10% of the total apparent mutational signal after accounting for genotype error (see next section). To confirm that this procedure is accurate, we also apply it to simulated data with gene conversion (see below and Text S1).

Genotype error

While our method is not as sensitive to genotype errors (by which we typically mean sites that are in fact homozygous but are mistakenly called as heterozygous) as are approaches based on counting *de novo* mutations, it is still important to consider their effects. False-positive heterozygous sites will artificially inflate our estimates of μ. This can be seen by considering matching $H_S(d)$ curves for real and simulated data, the former with genotype errors included. Since the decay rates of the curves as a function of d are equal, this means the underlying regions have had the same numbers of recombinations on average, and hence have the same average TMRCAs. The real and simulated regions also have the same average starting heterozygosity $H_S(0)$, but since the real data contain both true heterozygous sites and genotype errors, the calibration data must have a higher mutation rate per generation. Additionally, the upward bias in the estimates will be larger for smaller values of $H_S(0)$, since the local ratio of false to true heterozygous sites will be larger.

We have two main approaches for dealing with errors in genotype calls. First, we have taken a number of steps to filter the data, discussed below, such that the sites we analyze have high-quality calls and are as free from errors as possible. However, this is not sufficient to eliminate all false positives, and thus we also use local transition/transversion ratios to quantify the proportion of true homozygous sites that are called as heterozygous (for full details, see Text S1). For our final values of μ, we directly correct for these inferred levels of genotype error. Specifically, for a given starting heterozygosity value $H_S(0)$ and genotype error rate ϵ (inferred for the same set of windows S), we multiply the initial estimate of μ by a factor of $(H_S(0) - \epsilon)/H_S(0)$. To ensure that this correction is valid, we also perform simulations in which we randomly add false heterozygous sites to the simulated sequence data (see below).

Noise and uncertainty

Several of the steps in our procedure have some associated statistical uncertainty, while others rely on randomness. These include the computation of $H_S(d)$ from a finite number of genomes, the population size inference with PSMC, the random genetic map perturbation, and the simulation of calibration data. In order to capture this uncertainty, we use jackknife resampling to obtain standard errors for our estimates of μ, treating each autosome as a separate observation and leaving out one chromosome in each replicate. Our rationale for this scheme is that nearby regions of a chromosome are non-independent, and different individuals can also have correlated coalescent histories for a given locus, but a chromosomal unit encompasses most or all of the dependencies among the data. We note that we have found that the transformation from PSMC-inferred demographic parameters to calibration data via msHOT is discontinuous and is not properly captured by the jackknife. Thus, we use the same population size history (from the full data) for each replicate (see next section and Results).

Simulations

To test the accuracy of our procedure in a controlled setting, we first apply it to simulated data. When not otherwise specified, we create 20 sample genomes with an ancestral population size of 10,000 outside of a 10× bottleneck from 1000–2000 generations ago (approximating the age of the out-of-Africa bottleneck); the data are simulated with msHOT, using $\mu = 2.5 \times 10^{-8}$ and a perturbed version of the "shared" AA genetic map ($\alpha = 3000$ and $\pi = 9 \times 10^{-5}$). We run the full inference procedure as we would with real data, except with a default total of 30 genomes' worth of data per calibration curve versus 40 for real data. Also, for computational efficiency, when running PSMC on simulated test data, we only include a single copy of each chromosome (chosen at random from among the samples in the simulated data set).

As discussed in more detail in Text S1, in addition to this basic setup (a), we also run a number of additional simulations. First, we run the procedure (b) with a true rate of $\mu = 1.5 \times 10^{-8}$, (c) with a true rate of $\mu = 1.5 \times 10^{-8}$ plus gene conversion, and (d) with simulated genotype errors. Then, to test the effects of possible model violations, we simulate (e) samples from an admixed population, (f) mutation rate heterogeneity based on polymorphism levels in present-day African individuals, and (g) all three complications (d)–(f) simultaneously. For simulations (d) and (g), we add false heterozygous sites to the simulated diploid genomes (at a rate of 1 per 100 kb for (d) and 1 per 150 kb for (g)) and apply our standard correction, with one modification: because msHOT does not create individual nucleotides, we directly count the numbers of errors in ascertained regions instead of using transitions and transversions.

Data and filtering

As mentioned previously, we generate our estimates using genome sequences from non-African individuals, since the presence of a large number of relatively recently coalesced blocks arising from the out-of-Africa bottleneck gives us more data to work with at starting points with low heterozygosity. We use high-coverage sequences published in [21] and [24].

In order to remove as many genotype errors as possible, we use a filtering scheme based on the one applied to estimate heterozygosity in [24]. This consists of a tandem repeat filter, mapping quality threshold ($\mathrm{MQ} = 30$), genome alignability filter (all possible 35-mers overlapping a given base match uniquely to that position in the genome, with up to one mismatch), and coverage thresholds (central 95% of the depth distribution) [24]. We additionally apply a strict genotype quality threshold in order to preserve the highest-quality calls for analysis. From the GATK output, we compare the PL likelihood score of the heterozygous state to the minimum of the two PL scores of the homozygous states, imposing a quality threshold of 60 along with a prior of 31 (to reflect the genome-wide average heterozygosity). That is, if the heterozygote PL is at least $60+31 = 91$ lower than either homozygote PL, we call the site heterozygous; if it is at least $60 - 31 = 29$ higher, we call the site homozygous; and if it is in between, we mask the site as low-quality. Finally, we also remove all sites 1 or 2 bases away from any masked base under the five filters described.

While filtering is not necessary for the simulated calibration data, we still apply the same filters to the calibration data as to the real sequence data for consistency, on a genome-matching basis (e.g., for a sample of eight real genomes and our default real-data setting of 40 genomes' worth of calibration data, the base positions that are masked for each real sequence are also masked in five of the simulated sequences). In addition to filtering out individual sites, we impose a missing-data threshold for regions, ignoring any with more than 50% of sites masked (either of the super-region or the 100-kb central region).

Acknowledgments

We thank Shop Mallick and Heng Li for their assistance with technical details and Guy Amster, Priya Moorjani, George Tucker, Pier Palamara, Shai Carmi, and Molly Przeworski for helpful discussions. M.L. acknowledges support from the Simons Foundation and NIH (grant R01GM108348, to B.B.). P.L. was supported by NIH fellowship F32 HG007805 and S.S. by NIH grant K99 GM111744. N.P. and D.R. are grateful for support from NSF HOMINID grant #1032255 and NIH grant GM100233. D.R. is an Investigator at the Howard Hughes Medical Institute.

References

1. Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet 13: 745–753.

2. Campbell CD, Eichler EE (2013) Properties and rates of germline mutations in humans. Trends Genet 29: 575–584.

3. Ségurel L, Wyman MJ, Przeworski M (2014) Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet 15: 47–70.

4. Li WH, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. Nature 326: 93–96.

5. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–626.

6. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636–639.

7. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, et al. (2011) Variation in genome-wide mutation rates within and between human families. Nat Genet 43: 712–714.

8. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471–475.

9. Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, et al. (2012) Estimating the human mutation rate using autozygosity in a founder population. Nat Genet 44: 1277–1281.

10. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, et al. (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514: 445–449.

11. Fenner J (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128: 415–423.

12. Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, et al. (2012) A direct characterization of human mutation based on microsatellites. Nat Genet 44: 1161–1165.

13. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321–324.

14. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, et al. (2010) Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467: 1099–1103.

15. Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, et al. (2011) The landscape of recombination in African Americans. Nature 476: 170–175.

16. Li H (2014) Towards better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30: 2843–2851.

17. Williams A, Geneovese G, Dyer T, Truax K, Jun G, et al. (2015) Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. Preprint available : bioRxiv 009175. Accessed 5 February 2015.

18. Sankararaman S, Patterson N, Li H, Pääbo S, Reich D (2012) The date of interbreeding between Neandertals and modern humans. PLoS Genet 8: e1002947.

19. Kong A, Thorleifsson G, Frigge ML, Masson G, Gudbjartsson DF, et al. (2014) Common and low-frequency variants associated with genome-wide recombination rate. Nat Genet 46: 11–16.

20. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

21. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222–226.

22. Cochran G, Harpending H (2013) Paternal age and genetic load. Hum Biol 85: 515–528.

23. Neale BM, Kou Y, Liu L, Maayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485: 242–245.

24. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43–49.

25. Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, et al. (2012) A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet 8: e1003125.

26. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, et al. (2013) Great ape genetic diversity and population history. Nature 499: 471–475.

27. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, et al. (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169–175.

28. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nat Genet 46: 919–925.

29. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, et al. (2012) De novo gene disruptions in children on the autistic spectrum. Neuron 74: 285–299.

30. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, et al. (2014) De novo mutations in schizophrenia implicate synaptic networks. Nature 506: 179–184.

31. Iossifov I, ORoak BJ, Sanders SJ, Ronemus M, Krumm N, et al. (2014) The contribution of de novo coding mutations to autism spectrum disorder. Nature 515: 216–221.

32. of the Netherlands Consortium G, et al. (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46: 818–825.

33. McVean GA, Cardin NJ (2005) Approximating the coalescent with recombination. Philos Trans R Soc Lond B Biol Sci 360: 1387–1393.

34. Marjoram P, Wall JD (2006) Fast “coalescent” simulation. BMC Genet 7: 16.

35. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M (2008) High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science 319: 1395–1398.

36. Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

37. Hellenthal G, Stephens M (2007) mshot: modifying hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics 23: 520–521.

38. Jeffreys AJ, May CA (2004) Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat Genet 36: 151–156.

39. Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, et al. (2014) Recombination initiation maps of individual human genomes. Science 346: 1256442.

40. Harris K, Nielsen R (2014) Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Res 24: 1445–1454.

41. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, et al. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836–840.

42. Berg IL, Neumann R, Lam KWG, Sarbajna S, Odenthal-Hesse L, et al. (2010) PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet 42: 859–863.

43. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, et al. (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513: 409–413.

44. Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A (2013) A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. PLoS Genet 9: e1003684.

Text S1: Supplementary Methods

Technical details for population size history inference

PSMC runs on a reduced version of the genome, with consecutive sites grouped into bins of 100 and each bin marked as 1 or 0 depending on whether there is at least one heterozygous site in the bin or not. Bins can also be marked as "missing" if a certain number of the 100 sites have un-called genotypes (90 in the original PSMC publication). We find that two aspects of this procedure can affect the overall average heterozygosity of simulated data generated from a PSMC-estimated population size history. First, different values of the missing-bin threshold lead to different heterozygosity levels. Second, the PSMC program accepts a maximum of one heterozygous site per bin, whereas there can in fact be more than one. This effect is non-linear as a function of TMRCA; since heterozygosity varies substantially along the genome, the program will systematically underestimate the age of the most anciently coalesced regions.

To account for these factors, we first use an empirically-determined threshold of 35 un-called sites per 100 to call a bin as missing, which yields a more closely matching final heterozygosity. Second, we implement a multiple-het-per-bin adjustment, whereby we modify the PSMC output before using it to create the calibration data, as follows. The population size history inferred from PSMC consists of discretized time intervals $[T_{i_1} T_{i_2}]$ (in coalescent units), with the population size in each interval telling us how much of the genome falls within that level of TMRCA (and hence of per-bin heterozygosity). Given these heterozygosity levels, we use a binomial distribution to scale the times defining the endpoints from the output units of probability of at least one heterozygous site per 100 bp to the implied expected number of heterozygous sites per 100 bp. Creating calibration data according to these new values more accurately recapitulates the true distribution of heterozygosity levels across the genome, as well as the total genome-wide heterozygosity.

Additional notes on genetic map error

Effect of α on the inferred value of μ

The smaller the value of α used to generate the calibration data (i.e., the less accurate the genetic map is taken to be), the smaller the final inferred value of μ will be. This is because the genetic lengths used for simulation will be more discrepant from the base map, and as a result, the final value of $H_S(d)$ (computed on the calibration data) for a given d will reflect an average of the heterozygosity over a wider range of perturbed distances around d. Since $H_S(d)$ is a concave function of d, this smoothing will cause $H_S(d)$ to decrease, or in other words, to make the decay appear slower. Thus, in order to match the real data, a calibration curve with a smaller α would have to have a higher scaled recombination rate, and hence, with a fixed θ, a smaller μ.

Alternative method to infer α

We find that the dependence of our inferred μ on α is stronger for larger starting values $H_S(0)$, because a larger $H_S(0)$ leads to a steeper decay of $H_S(d)$ over a shorter genetic distance, so that the curve is more sensitive to the smoothing caused by map error. An illustration of this phenomenon can be seen in Figure S2. In addition to using Figure S2 to observe the sensitivity of our estimates to the chosen value of α, we can also derive from it an independent, empirical best-fit estimate of α. If we assume that the mutation rate has not changed appreciably in the last few thousand generations, then the estimates for the different starting point sets S should be equal. Thus, the correct value of α can be inferred by assuming that this equality holds. We fit linear regression lines—which, importantly, have different slopes—to the inferred values of μ as a function of α, and we find that the three lines very nearly intersect at a single point at $\alpha = 3760$ M^{-1} (with corresponding $\mu = 1.79 \times 10^{-8}$), only modestly higher than our primary estimate of 3100.

Interpolation within the genetic map

When computing $H(d)$, we measure all genetic distances in both the real and simulated data by linearly interpolating individual sites within the SNP marker grid on which the map is defined. In truth, genetic distances will not be uniform at sub-interval scales. We did attempt to test the importance of this effect by creating (randomized) unevenly subdivided maps and did not obtain noticeably different results (data not shown).

Details of gene conversion correction

As discussed in the Methods, the magnitude of the gene conversion effect depends on both the prevalence of gene conversion events and the rate at which they introduce heterozygous sites. We make use of a recent direct estimate that gene conversion affects 5.7×10^{-6} bases per generation [17], which is the product of the rate of events and their average length (on the order of 100 bases in humans [17, 38]). This rate is a genome-wide value, however, and the frequency of gene conversion is highly variable and correlated with local crossover rate [17, 39]. We assume that the gene conversion rate is proportional to the local recombination rate in our genetic map. Because of the way our starting points are chosen, they tend to lie in recombination-poor regions, and so for each of our estimates, we multiply the average 5.7×10^{-6} gene conversion rate by the ratio of the local recombination rate (measured in 10 kb windows around the starting points) to the genome-wide average (1.19 cM/Mb, as measured in the deCODE map [14], which was used in [17]). As an example, for our primary eight-genome $H_{5\text{–}10}(d)$ curve, the starting-point windows cover

96 Mb but only 51 cM, less than half the average.

The second component of the gene conversion effect is the probability that a gene conversion event introduces a polymorphism that we observe as a heterozygous site. A very simple model for this rate would be that it equals the probability that a randomly chosen base differs from the homologous base on another chromosome in the population (from which it might have been copied in the event of a gene conversion at some point in the past). This quantity can be estimated simply as the heterozygosity in a diploid genome. To this basic model, we add two additional details. First, although our test regions have relatively low heterozygosity (i.e., are relatively recently coalesced), there is also a chance that new mutations that have accumulated on either haplotype could be replaced by the ancestral base via gene conversion. For this probability, we use the heterozygosity at our starting points, $H_S(0)$, divided by 2 (because the new mutations have increased in number since the TMRCA). Second, the probability of mismatch with another chromosome could change over time, and we also find that because of correlated histories or other factors, the heterozygosity in other genomes is reduced around our ascertained starting points. For example, other European and East Asian genomes have heterozygosity 5.5–5.9 $\times 10^{-4}$ in the 5–10 het regions defined in French B, roughly 10–20% below the genome-wide average (see Table S2). We assume an average value of 5.7×10^{-4} both for European test genomes (which have the highest heterozygosity) and other populations, with the exception of the two least diverse populations, Karitiana and Papuan, for which we instead use their total genome-wide heterozygosity (5.1×10^{-4} and 5.5×10^{-4}, respectively).

Finally, we assign an uncertainty to our correction, which is incorporated into the standard errors we report for our real-data mutation rate estimates. For the rate of gene conversion, we translate the estimated 95% confidence interval of 4.5–7.3 $\times 10^{-6}$ per base per generation [17] into a standard deviation of 0.7×10^{-6}, and for the probability of introducing a new polymorphism, we use a standard deviation of 10% of the total (for example, $5.7 \pm 0.57 \times 10^{-4}$ for most applications).

To test the validity of this approach, we run simulations in which the test data are generated with gene conversion active and the rest of the procedure is carried out as normal, including the correction just described. For this scenario, we use a true mutation rate of 1.5×10^{-8} per base per generation, and for computational efficiency, we assume a uniform rate of gene conversion with $f = 2.5$ (i.e., 2.5 times the average recombination rate) and an average tract length of 100 bases. This results in a gene conversion rate of approximately 2.8×10^{-6} (about half the true rate in humans, but comparable for our starting points because there is no reduction in recombination-poor regions). When applying the final correction, we use the average heterozygosity over all simulated genomic segments (approximately 4.0×10^{-4}) as the "donor" polymorphism probability.

Details of genotype error rate estimation

To estimate the genotype error rate, we make use of the observed counts of transition and transversion mutations in different windows of the genome. We let $\beta_0 \approx 2$ [3] denote the true genome-wide average ratio of transitions to transversions, and we assume that false-positive heterozygous calls are equally likely for any nucleotides, and thus have a transition/transversion ratio $\beta_{fp} = 1/2$. Using the same data as for our final μ estimates, with the same filtering and ascertainment of starting points, we compile the numbers of transitions and transversions in windows with local heterozygosity rates of 1–5, 5–10, and 10–20 per 100 kb. We also compute genome-wide counts, which, assuming that the abundance of false positives is approximately the same everywhere, consist of $\sim 99\%$ true heterozygous sites. For our data, this yields a value of $\beta_0 = 2.15$. Given β_0 and β_{fp}, we can then solve for the numbers of true and false sites out of the total observed numbers of transitions and transversions. We perform this calculation separately for different categories of windows because the ascertainment scheme may cause them to have different error rates: the windows are selected based on their total heterozygosity levels (including errors), and the population history of the samples also leads to an excess of certain true TMRCAs. Also, we only count transitions and transversions in shorter windows (namely, 60 kb for the lowest heterozygosity, 30 kb for medium heterozygosity, and 15 kb for the highest heterozygosity) in order to minimize the numbers of recombinations within the intervals.

Mutation rate heterogeneity

In most of our theoretical and methodological discussion, we have assumed that μ is a constant parameter, but in fact different portions of the genome can have different local mutation rates (see Discussion). To learn about the effects this might have on $H_S(d)$, we use simulations in which we create data with variability in the mutation rate throughout the genome. Our goal is to approximate the true level of variability present in our data on length scales of several tens of kb, which we do by considering polymorphism rates from populations that are distantly related to those we are using to compute $H_S(d)$. Specifically, for each super-region sequence being simulated, we divide the length into even thirds and assign a mutation rate for each third proportional to the frequency of doubleton sites (those heterozygous in exactly two individuals) in that interval among eight full African genomes (two each San, Dinka, Mbuti, and Yoruba) [21, 24]. The exact rates are chosen in such a way that the overall average rate over the entire genome is equal to a specified value (e.g., 2.5×10^{-8}). When calling a doubleton, we require at most one individual to be masked (see Methods), and if there are fewer than 10,000 un-masked sites in any window, we assign the genome-wide average rate for that third of the super-region instead.

We note that while this diversity metric does not precisely capture the true variation in mutation rates across the genome, it should be sufficient for our purposes. The frequency of doubleton sites in the African individuals depends on the local coalescent trees, although this particular statistic should be relatively stable (as opposed to, for example, the total number of polymorphic sites) and will also be smoothed somewhat by considering windows typically on the order of 100 kb. Second, given a local tree, there is additional Poisson noise in the observed counts of doubleton mutations. However, because we are using simulations to gauge the effect of rate variability rather than attempting to add matching variability to the calibration data, we do not need the local estimates to be exact. This is especially true because both sources of stochasticity should cause the apparent variability to be too high, so that our procedure is conservative. Finally, the simulation approach avoids the issue of potential correlations between the African diversity and our non-African data as a result of deep shared ancestry.

Additionally, we note that our estimates should not be impacted by multiple-nucleotide mutation events. It has been estimated that a few percent of human point mutations involve nearby clustered nucleotide changes [31, 40], meaning that our inferred rate will correspond to the number of single-base changes per base per generation, rather than the number of independent mutational events per base per generation. However, this is simply a question of how the mutation rate is defined (and moreover, the two are very similar), and all previous mutation rate estimates, to our knowledge, also use the first definition.

Population heterogeneity and admixture

Our basic model assumes that all of the genomes used to calculate $H_S(d)$ are drawn from the same population. Thus, to the extent that our set of individuals are from groups with different historical sizes, the real data could have different TMRCA patterns in different genomes, whereas the calibration data will be based on the population size profile inferred from the aggregate of all of the samples. A similar phenomenon occurs on a within-genome scale if the test genomes are admixed, in which case individual chromosomes consist of alternating blocks derived from each ancestral mixing population. Although we avoid populations with substantial recent admixture, almost all human populations have experienced some degree of admixture at some point in their history.

Separately, our estimates could potentially be affected by inter-population differences in fine-scale recombination rates. However, the "shared" version of the AA map is intended to apply broadly to non-Africans [15], and it is reasonable to expect that recombination maps will be similar among non-African populations based on their relatively limited diversity of PRDM9 alleles [15, 41, 42].

As with mutation rate heterogeneity, it is difficult to quantify the exact divergence and admixture parameters for the populations under consideration, and hence we use empirical and simulation approaches to study their effects on our inferences. First, while our primary data set consists of a combination of European and East Asian individuals (see Results), we also apply our analyses to the continental groups separately. This allows us both to compare results for populations with different admixture histories and also to compare more homogeneous data sets to the full set with individuals from diverged populations.

Second, we perform simulations in which we apply our method to an admixed population designed to emulate present-day Europeans [28, 43]. We simulate a population that experienced a 50/50 mixture 200 generations ago between two populations that had been diverged for 1000 generations. The population sizes are 10,000 in the shared ancestral population before a bottleneck beginning 2000 generations ago; then 1000 during the bottleneck, which continued until 1000 generations ago (so that the last 200 generations are separate in the two mixing populations); then 15,000 and 7500 in the two mixing populations after the bottleneck; and finally 25,000 after the admixture.

Natural selection

A similar issue to within-genome variation in demography and in the mutation rate is that of heterogeneity in selective effects. Until now we have assumed implicitly that all loci are neutral with respect to fitness, and thus their TMRCAs follow the distributions implied by the standard coalescent model. However, if some sites have non-zero effects on fitness, then the local genealogies there will have different properties from the genome-wide average, whereas all loci in the calibration data will be drawn from the average distribution. Since our PSMC inferences still capture the true genome-wide ancestral population size history, though, the local variation in this profile will likely only be of second-order importance. Moreover, the effects of selection (and also of GC-biased gene conversion, which acts similarly to natural selection [44]) should be similar to those of rate and demographic heterogeneity, since different local genealogical properties caused by selection are analogous to different local mutation rates and/or different local ancestry. Thus, because all three phenomena lead to changes in local levels of diversity, any overall impact of natural selection should be captured in our rate-heterogeneity and admixture simulations described above.

Effect	Figure references
Finite sample size	**2A, 2B**
Genetic map error	**2A, 2B**
Coalescent simulation	**2A, 2B**
Interpolation and least-squares fitting	**2A, 2B**
Demographic parameter estimation	**2A, 2B**
Gene conversion	**2C**
Genotype error	**2D, 2G**
Uncertainty in α	*S2*
Within-genome rate heterogeniety	**2E, 2G**
Admixture (population heterogeneity)	***2F, 2G, 3A-C***

Figure S1. Guide to potential sources of uncertainty associated with our method. Blue shading: included in procedure and in jackknife standard error; red shading: included in procedure; bold: tested with simulations; italic: tested empirically with real data.

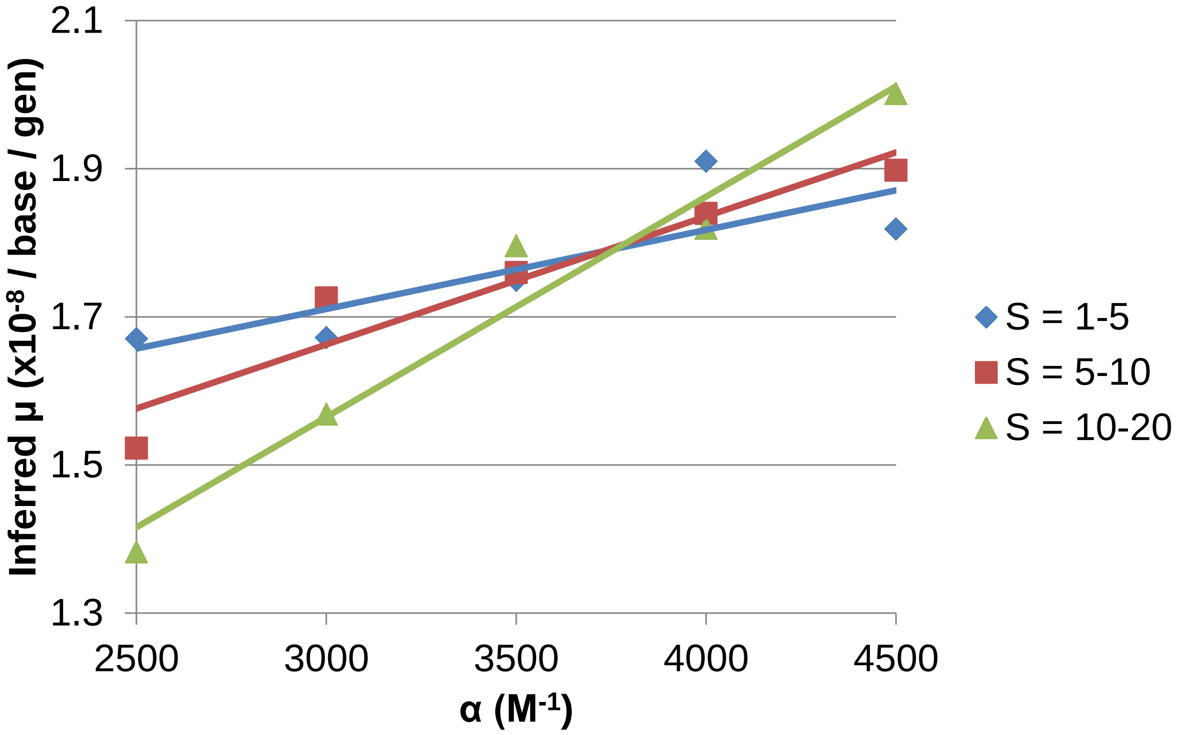

Figure S2. Inferred mutation rates for a range of values of the genetic map error parameter α and starting heterozygosity S. All estimates use our standard data set of eight non-African genomes. Data points represent the inferred rates (independent point estimates), and the lines are linear regression fits for each of the three choices of S as a function of α.

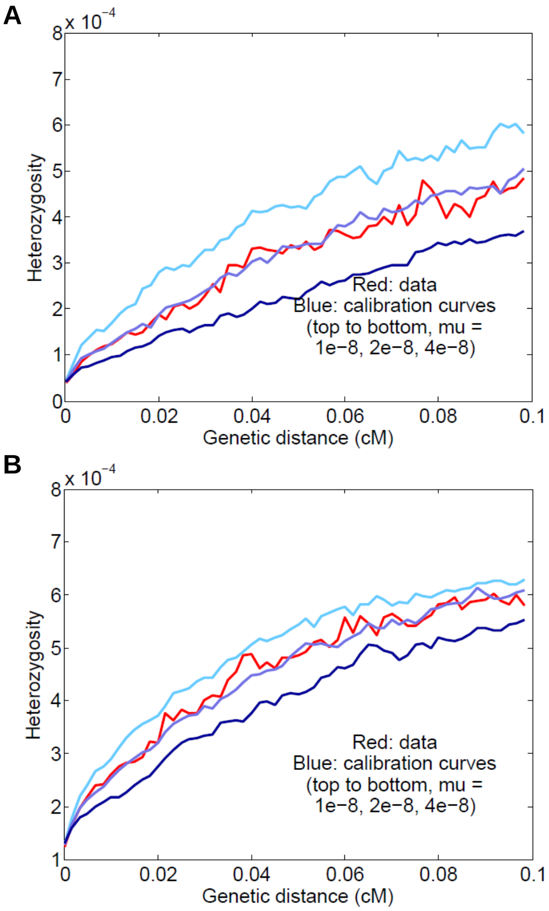

Figure S3. Results for Europeans plus East Asians with alternative choices of the starting point S. (A) Curves with 1–5 heterozygous sites per 100 kb (using 56 genomes' worth of calibration data); the inferred rate is $\mu = 1.72 \pm 0.12 \times 10^{-8}$. (B) Curves with 10–20 heterozygous sites per 100 kb; the inferred rate is $\mu = 1.56 \pm 0.14 \times 10^{-8}$.

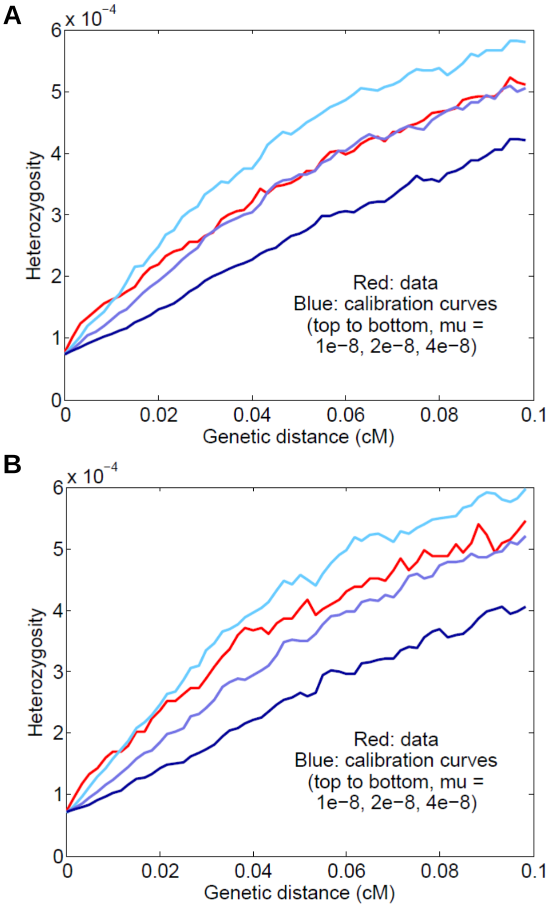

Figure S4. $H_{5\text{–}10}(d)$ curves without the pseudo-count prior. (A) Simulated data: we create test data using the prior but omit it for the calibration data. The curve shapes are markedly different, as the calibration curves decay too slowly at the smallest values of d. It is also apparent that the inferred value of μ is lower than the true value of 2.5×10^{-8}. (B) Real data for eight non-African genomes. We observe a very similar discrepancy between the real-data and calibration curves (compare Figure 3A).

Table S1. Transition and transversion counts for ascertained genomic windows

Hets per 100 kb	Total tr	Total tv	Estimated true tr	Estimated true tv	Estimated errors	Total sites
1–5	4114	2358	3823	1776	873	174863733
5–10	8690	4648	8292	3852	1195	187581501
10–20	6789	3404	6626	3078	489	86074882
1–20	19593	10410	18740	8705	2558	448520116
All	6339965	2985086	6313794	2932744	78513	13768383829

Observed transitions and transversions, with estimated genotype error counts, for eight genome sequences of non-African individuals. We assume that errors occur equally frequently for all nucleotide pairs and solve for the transition/transversion ratio $\beta_0 = 2.15$ among true heterozygous sites by assuming that the error rate is equal for the combined 1–20 het regions (fourth line) and the genome as a whole (last line). See Methods for complete details. Tr, transitions; tv, transversions.

Table S2. Sequence divergence for sites passing filters

Sequence comparison	Per-base diff	Divergence time (My)
French A	7.03×10^{-4}	0.62 ± 0.04
French B	7.01×10^{-4}	0.62 ± 0.04
Sardinian A	6.98×10^{-4}	0.61 ± 0.04
Sardinian B	6.76×10^{-4}	0.59 ± 0.04
Han A	6.65×10^{-4}	0.58 ± 0.04
Han B	6.50×10^{-4}	0.57 ± 0.04
Dai A	6.66×10^{-4}	0.59 ± 0.04
Dai B	6.59×10^{-4}	0.58 ± 0.04
Human–chimpanzee	1.23×10^{-2}	10.9 ± 0.7

Sequence comparisons, with implied divergence times (mean ± standard deviation, in millions of years, using our inferred mutation rate of $1.65 \pm 0.10 \times 10^{-8}$ per base per generation and an average generation interval of 29 years), for sites in the genome passing filters. The first eight lines represent divergence between the two chromosomes within the individual genomes in our primary data set (suffix “A” from [21] and suffix “B” from [24]), based on genome-specific filtering, and are an average over all super-regions (see Methods). Human–chimpanzee statistics are averaged over the filters for all eight genomes; we note that the third column represents the TMRCA of the two species’ reference sequences rather than the population split time (see Discussion).