Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates

Pier Francesco Palamara,[1,2,*] Laurent C. Francioli,[3] Peter R. Wilton,[4] Giulio Genovese,[2,5,6] Alexander Gusev,[1,2] Hilary K. Finucane,[1,2,7] Sriram Sankararaman,[2,6] Genome of the Netherlands Consortium, Shamil R. Sunyaev,[2,8] Paul I.W. de Bakker,[3,9] John Wakeley,[4] Itsik Pe'er,[10] and Alkes L. Price[1,2,11]

The rate at which human genomes mutate is a central biological parameter that has many implications for our ability to understand demographic and evolutionary phenomena. We present a method for inferring mutation and gene-conversion rates by using the number of sequence differences observed in identical-by-descent (IBD) segments together with a reconstructed model of recent population-size history. This approach is robust to, and can quantify, the presence of substantial genotyping error, as validated in coalescent simulations. We applied the method to 498 trio-phased sequenced Dutch individuals and inferred a point mutation rate of 1.66×10^{-8} per base per generation and a rate of 1.26×10^{-9} for <20 bp indels. By quantifying how estimates varied as a function of allele frequency, we inferred the probability that a site is involved in non-crossover gene conversion as 5.99×10^{-6}. We found that recombination does not have observable mutagenic effects after gene conversion is accounted for and that local gene-conversion rates reflect recombination rates. We detected a strong enrichment of recent deleterious variation among mismatching variants found within IBD regions and observed summary statistics of local sharing of IBD segments to closely match previously proposed metrics of background selection; however, we found no significant effects of selection on our mutation-rate estimates. We detected no evidence of strong variation of mutation rates in a number of genomic annotations obtained from several recent studies. Our analysis suggests that a mutation-rate estimate higher than that reported by recent pedigree-based studies should be adopted in the context of DNA-based demographic reconstruction.

Introduction

Germline mutations represent a fundamental evolutionary force that shapes phenotypic variation and has a profound impact on heritable diversity. Precise estimation of mutation rates has several applications, including the interpretation of mutations implicated in diseases,[1–3] studies of natural selection,[4,5] the timing of demographic events inferred from genetic analysis,[6–8] and the study of several aspects of human mutagenesis.[9] High-throughput sequencing technologies have recently enabled the quantification of germline mutation rates, but the estimates obtained by these methods are inconsistent with those of previous studies. The source of these inconsistencies, whether biological or due to methodological biases, is at the center of recent debate,[7,10,11] and gaining additional insight into germline mutation rates will require new methods.

In this work, we propose a method for estimating mutation rates by using mutations occurring within identical-by-descent (IBD) haplotype blocks[12–17] transmitted through recent common ancestors who lived ~100 generations (~3,000 years) before the present. These IBD segments can be detected by several available methods[13,18,19] and reflect genetic relationships that are typically not known to the affected individuals but are found to be ubiquitous even in outbred populations.[15,20] IBD segments are defined in our work as contiguous chromosomal regions in which the most recent common ancestor (MRCA) for two sampled chromosomes is unchanged. Occasional mutations segregating along the lineages connecting a pair of IBD haplotypes to their MRCA will create mismatched sites on the shared haplotypes, and these sites can be used for inferring the rate at which new germline mutations appear. If the exact number of generations separating the IBD segments (via their MRCA) is known, one can infer the mutation rate by dividing the number of observed sequence mismatches by the number of generations and the physical length for all segments. A special case of this approach is used in trio-based analyses, where transmitted parental haplotypes and IBD offspring haplotypes are separated by a single generation. In this work, we instead use a reconstructed demographic model to infer the age of IBD segments.

The IBD-based approach we propose for inferring mutation rates is robust to, and can quantify, the presence of

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [3]Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, 3584 CX, the Netherlands; [4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA; [5]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [6]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; [7]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02142, USA; [8]Genetics Division, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [9]Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, 3584 CX, he Netherlands; [10]Department of Computer Science, Columbia University, New York City, NY 10027, USA; [11]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
*Correspondence: ppalama@hsph.harvard.edu

http://dx.doi.org/10.1016/j.ajhg.2015.10.006.

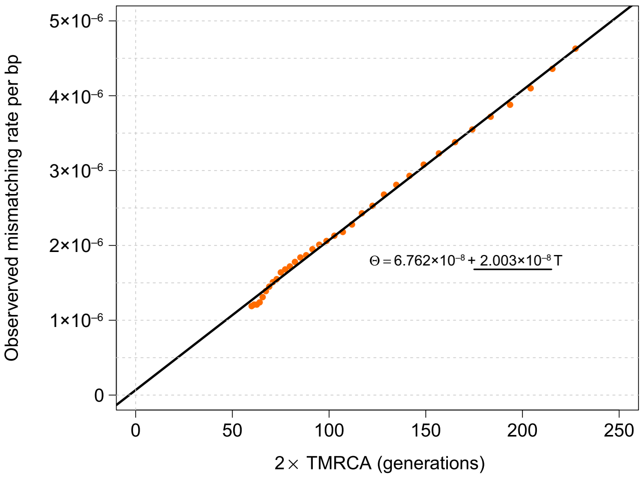

Figure 1. tMRCA Regression

We simulated a chromosome of 50 cM for 250 diploid samples by using $\mu = 2 \times 10^{-8}$ for the mutation rate and no genotyping error. We matched the allele-frequency spectrum of the simulated samples to the spectrum found in real data for IBD-segment detection with GERMLINE, and we used the IBD-segment detection parameters used in real data. The slope of this regression captures the simulated mutation rate; the intercept is proportional to genotyping error rate.

substantial amounts of genotyping error in the analyzed sequences and can be used for inferring the rate of non-crossover gene conversion. We applied the developed methodology to analyze 250 trio families from the Netherlands and infer mutation and gene-conversion rates. We further studied the rate of short indels and analyzed the relationship between recombination rates and mutation rates. We studied the enrichment of deleterious variation in mismatching variants within IBD regions, showed that the length of shared IBD segments along the genome closely reflects summary statistics of background selection, and explored enrichment or depletion of mutation rates in several specific genomic annotations.

Material and Methods

Overview of Methods

The method we propose is aimed at estimating the sex-averaged, genome-averaged, and time-averaged mutation and gene-conversion rates per base per generation by using mismatching genotype sites found on IBD haplotypes for a sample of individuals from a population of known demographic history. Note that these quantities are affected by several aspects of the study population, such as the length of the generation for males and females along the ancestral lineages in the past several generations (~100 generations in our analysis; see Discussion). We briefly describe solutions to three challenges. First, to estimate the number of generations separating two IBD segments (twice the time to the MRCA [tMRCA]), we use a recently developed method[14] that relies on the spectrum of observed IBD-segment lengths to infer demographic history. We then use this method to obtain a posterior mean estimate of the average tMRCA for pools of IBD segments of different lengths, as detailed in the real-data description (see GoNL Dataset) and Appendix A. Second, to deal with the presence of genotyping errors rather than rely on stringent filtering criteria (as in trio-based analyses[21–25]), we regress the observed sequence mismatches for several IBD-segment length thresholds on the estimated tMRCA; the slope of this regression reflects the rate at which new mutations accumulate per generation time unit, whereas the genotyping error rate is captured by the intercept. We refer to this procedure as tMRCA regression (illustrated in Figure 1). Finally, we correct for the occurrence of non-crossover gene-conversion events along the lineages leading to the MRCA by exploiting the relationship between an allele's frequency and the probability that it is involved in a gene-conversion event (see Controlling for Gene Conversion via MaAF-Threshold Regression below). This also allows us to estimate the rate at which a genomic locus is involved in a non-crossover gene-conversion event; this rate is proportional to the difference between corrected and uncorrected estimates for mutation rates. We have released open-source software (IBDMUT) implementing these methods (see Web Resources).

Estimating the Mutation Rate via tMRCA Regression

The proposed methodology for the inference of mutation rates requires the availability of haploid genotype data, a list of IBD segments that exist between pairs of haploid individuals and that are longer than a specified Morgan length threshold (including start and end positions), and a demographic model, which can be inferred from the spectrum of shared IBD segments as described in Palamara et al.[14] For each IBD segment *i*, we obtain an observed mismatch rate by counting the number of sequence differences m_i in the haploid genotypes within the region and dividing by the region size s_i in base pairs: $\theta_i = m_i/s_i$. We then obtain the observed mismatch rate by averaging all observations $\widehat{\theta}_u = n_u^{-1}\sum_{i=1}^{n_u}\theta_i$ for n_u segments longer than u Morgans. We repeat this measurement for several thresholds u to obtain a vector of observed mismatch rates $\widehat{\theta}$. Because of the lack of detailed pedigree structures at deep time scales, the exact number of meiotic events separating two individuals who share IBD segments is generally unknown. Using the reconstructed demographic model, we therefore infer the posterior mean age t_u of pooled IBD segments longer than a known genetic length threshold u by using recently developed coalescent theory[14,15] (details are summarized in Appendix A). Finally, we regress the observed mismatch rates $\widehat{\theta}_u$ on twice the posterior mean age (in generations) to the MRCA of the IBD segments: $\theta_u = \alpha + 2\mu t_u + \epsilon$. We refer to this regression as the tMRCA regression. Older segments will tend to harbor a larger number of sequence differences because mutation events have a higher chance of occurring along the lineages connecting extant individuals to their MRCA. The slope μ of this regression will capture the rate at which mutations arise per unit of time. Note that we are neglecting the uncertainty on the measurement in the regressor t_u, i.e., the inferred age of the pooled IBD segments. As shown in simulations, however, this only results in negligible biases for the estimated slope coefficient because of the large number of pooled segments.

If we assume a genotyping error model for which false-positive or -negative genotype calls are independent of the average coalescent time of pairs of individuals at a locus, the intercept α of this regression is expected to capture the rate at which genotyping errors occur on the considered range of IBD segments. Note that when performing the tMRCA regression, we rely on non-independent observations of mismatch rates (because we use

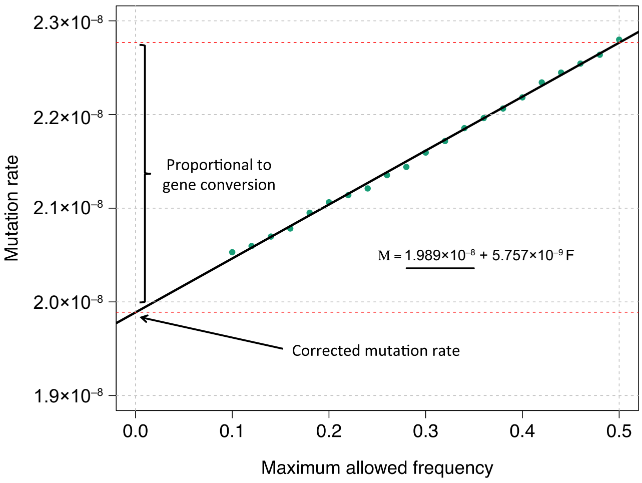

Figure 2. MaAF-Threshold Regression
We simulated 250 diploid samples as described in Figure 1 and a probability of 6 × 10^{-6} for a base pair to be involved in a non-crossover gene-conversion event. We performed the MaAF-threshold regression to correct for the occurrence of gene conversion. The regression intercept is used for estimating the corrected mutation rate, whereas the difference between the corrected and uncorrected mutation rates captures the effects of gene conversion, whose magnitude can be estimated with the observed population heterozygosity.

overlapping ranges for the length of the IBD segments), which corresponds to attributing larger weights to measurements obtained from long, more-reliable IBD segments. Although this violates independence assumptions in the regression, the reweighting of the data is not expected to result in biases for the estimated slope and intercept (Table S1), but it decreases heteroscedasticity. In order to estimate SEs of the resulting slope and intercept, which are expected to be biased because of non-independence, we rely in all cases on a block-weighted jackknife procedure,[26] which uses independent regions as resampling units.

Controlling for Gene Conversion via MaAF-Threshold Regression

Non-crossover gene-conversion events occur at a rate that is correlated to the recombination rate and have been observed to be more frequent than crossover recombination events.[27] In the coalescent process, gene conversion can be modeled as two consecutive recombination events that occur very close to each other,[28] at an average distance of ~300 bp.[29] These events introduce the possibility that polymorphisms segregating in the population might be assimilated into haplotypes within IBD regions. These polymorphisms can create sequence differences between individuals who share IBD segments. These sequence differences, however, are not due to newly arising mutations. Note that whereas gene-conversion events change the MRCA of the ~300 bp converted segment, here we do not consider this to break an IBD block. Furthermore, because the number of gene-conversion events is related to the number of meiotic events, short IBD regions will tend to exhibit more gene-conversion-driven mismatches than longer, more-recent IBD segments, therefore resulting in an upward bias when the mutation rate is estimated via the slope of tMRCA regression. The mismatching variants observed on IBD segments, therefore, will be due to at least two distinct sources of heterozygosity. The first, θ_p, which we hereafter call population heterozygosity, represents the effect of gene-conversion events, which introduce standing genetic variation onto IBD blocks. The second source of heterozygosity is due to newly arising point mutations on IBD blocks and will be referred to as θ_μ. For IBD segments of a chosen length, we can express the total observed mismatch rate as $\theta = \theta_\mu + \theta_p$. To estimate the mutation rate due to point mutations only, we need to exclude the effects of θ_p from our calculations. We make the following two observations:

1. The frequency of mutations that arise on long (e.g., ≥ 1 cM) IBD segments is typically low in the population (Figure S1), so that θ_μ is mostly due to rare variants.
2. If we divide the allele-frequency spectrum into bins of equal width, we find an approximately uniform contribution to θ_p for each frequency. This implies that if we compute the frequency-bounded population heterozygosity $\theta_{p,f}$ by using only variants of frequency up to f, we observe an approximately linear relationship between $\theta_{p,f}$ and f (Figure S2; see additional calculations in Appendix A).

Observation 1 implies that if we exclude high-frequency variants when we compute μ by using the proposed regression approach, the contribution of θ_μ to the observed mismatch rate on IBD segments will be largely unaffected. Furthermore, observation 2 suggests that if we estimate a frequency-bounded value of μ_f by ignoring variants of frequency higher than a threshold f, the contribution of population heterozygosity due to gene-conversion events, $\theta_{p,f}$, will be decreased to an extent that is approximately linear in f. Assuming that the contribution of θ_μ to μ_f is unaffected for values of f in the range $F = [F_{\min}, F_{\max}]$, we can therefore regress μ_f on F and observe a linear relationship. We refer to this regression as the MaAF-threshold regression (Figure 2). The intercept of this regression will then reflect an estimate of μ without the confounding effects of θ_p, whereas the contribution of θ_μ is left unchanged. We avoid computing values of μ_f corresponding to $F \in [0, F_{\min})$, for a sufficiently large $F_{\min}$ (e.g., >0.1), given that this might result in removing variants that are due to new point-mutation events on the IBD segments, which we use to estimate μ. Because this approach relies on the stochastic relationship among allele frequency, population heterozygosity, and gene conversion, it is not possible to fully determine whether the sequence mismatches that are found on IBD segments are due to a recent mutation event or a site involved in gene conversion, although those resulting from the latter are expected to have a substantially higher allele frequency. Finally, note that we neglect the possibility that point mutations arising on IBD segments are removed via gene conversion because this does not substantially affect the estimates. As in the tMRCA, the use of nested frequency bins in the MaAF regression results in non-independent observations in the performed regression. The use of nested frequency bins might improve the correction in cases where the relationship between MaAF cutoffs and population heterozygosity deviates from linearity as a result of recent demographic events. Simulations showed that his approach has no significant impact on the quality of the estimated mutation rates (Table S2). As in the case of tMRCA regression, we obtained reported SEs with the block-weighted jackknife method to avoid biases induced by the non-independent observations.

Estimating the Gene-Conversion Rate

The difference between the mutation rate computed without correction for gene-conversion events and the estimate obtained after removal of the effects of gene conversion can be used for quantifying the probability that a base pair within IBD segments is involved in a gene-conversion event during meiosis. This difference, which we indicate as μ_{GC}, represents the probability of observing a heterozygous site as a result of existing polymorphisms introduced via gene conversion in a single generation. This rate can be expressed as $\mu_{GC} = p(\mathrm{GC}) \times p(\theta_p \mid \mathrm{GC})$, i.e., the probability that a base pair is involved in a gene-conversion event can be multiplied by the probability of assimilating a heterozygous site given that the gene-conversion occurs at the locus. The quantity $p(\theta_p \mid \mathrm{GC})$ can be estimated with the genome-wide heterozygosity of the analyzed sample, and the value of μ_{GC} can be estimated with the previously described correction method. An estimate of $p(\mathrm{GC})$ is therefore obtained as $\widehat{p}(\mathrm{GC}) = \widehat{\mu}_{GC} \times \widehat{p}(\theta_p \mid \mathrm{GC})^{-1}$, and a confidence interval is obtained via block-weighted jackknife.[26]

Coalescent Simulations

We used extensive coalescent simulation to evaluate the proposed methodology. To this end, we used a publicly available coalescent simulator, COSI2[30] (which allows simulation of gene-conversion events), and our implementation of a coalescent simulator, inspired by the existing GENOME algorithm[31] (which enables simulation of a large number of samples and efficient extraction of information on IBD segments). The algorithm proceeds backward in time and, for each individual at generation g, samples a parent at the discrete time $g + 1$ in the past, occasionally resulting in coalescent events and sampling a new parent when a recombination event occurs. To speed up computation, the GENOME approach divides the simulated region into relatively large chunks that are not allowed to recombine, discretizing the recombination process and resulting in approximate linkage-disequilibrium (LD) structure at short genomic intervals. The version we developed enables substantial improvements of memory and run-time requirements while circumventing the original GENOME algorithm's simplifying assumption of non-recombining LD blocks. In brief, we sped up the original algorithm by sampling recombination breakpoints from an exponential distribution and by only storing chromosomal regions and individuals relevant for calculating the ancestral recombination graph (ARG) at each simulated generation. In addition, we applied several improvements to data structures and other algorithmic details. To evaluate our methodology, we further extended the program to allow efficient extraction of IBD segments from the ancestral recombination graph without requiring testing of differences in shared common ancestors for each marginal tree in the ARG, as done in previous works.[14,16,32] We have released open-source software (ARGON) implementing the simulator (see Web Resources).

To assess the impact of demographic history on our estimates, we simulated three plausible demographic scenarios, in addition to the reconstructed GoNL (Genome of the Netherlands; see GoNL Dataset below) demographic history. The simulated populations comprised an expanding population that experienced a severe founding event 30 generations before the present and a population that undergoes severe exponential contraction (referred to as Ashkenazi and Maasai, respectively, because they resemble recently studied groups[14]) and an exponentially expanding population (referred to as Europeans; see Figure S3). We used two types of recombination maps to simulate non-uniform recombination rates along the genome (Figure S4). To assess the impact of genotyping errors on our methodology, we simulated errors for which a previously unobserved variant is created ("de novo" errors) or false-positive or -negative calls on existing variants. To model frequency-dependent genotyping error rates, we used a beta distribution as a prior for sampling the frequency of planted genotyping errors[33] (Figure S5). For all simulations, we obtained posterior mean estimates for the age of IBD segments by using the coalescent distributions of the simulated models.

GoNL Dataset

We analyzed sequence data from a recent study of 250 trio families from the Netherlands (GoNL project,[34] release 4). The dataset consists of 748 individuals who passed quality control and were sequenced at an average of ~13× (details are provided in the analysis described elsewhere[34]). Combining the output of several detection algorithms detected indels (GoNL Release 5). In addition to using the quality-control filters applied in the original analysis of the data, we further excluded regions that did not meet several quality criteria derived from the 1000 Genomes Project phase 1, as described in Genovese et al.[35]

Trio phasing is expected to result in accurate estimation of haploid sequences in the GoNL data. Low-frequency variants, in particular, are unlikely to result in doubly heterozygous parents, so phasing of rare polymorphisms is generally trivial.

IBD segments and an inferred demographic model were obtained from the analysis described elsewhere[34] with the use of genetic maps from in the 1000 Genomes Project.[36] 26 regions were selected for the analysis reported in this paper; each was longer than 45 cM, which gave a total of 2,160 cM and an IBD density of 3.07×10^{-3} per site per pair. The B statistic of background selection[4] in these regions is slightly lower than the genome-wide average (0.78 versus 0.79; $p < 0.01$ based on 10,000 permutations). The B statistic, however, was not found to have a significant impact on our mutation-rate estimates (see Results). The recombination rate was not found to be significantly lower than the genome-wide average ($p = 0.37$). Informed by the density of de novo mutation events along the genome, a recent study[9] estimated a map of local variation in substitution rates. Out of the 14 types of substitutions reported in this map, four (C>T and G>A [$p < 0.01$]; A>T and T>A [$p = 0.032$]) were found to be depleted in these regions, although the differences were found to be minimal (−1.3% for C>T and G>A; −1.0% for A>T and T>A). This effect is probably mediated by the reduced B statistic in the regions, which is related to the substitution rates computed for primate sequences in Duret and Arndt,[37] on which the substitution map is based. Consistent with this hypothesis, we observed a small but significant correlation between these annotations and B statistic in these regions ($r = 0.014$ for C>T and G>A; $r = 0.017$ for A>T and T>A; $p < 10^{-6}$). The density of IBD-segment sharing along the analyzed regions is depicted in Figure S10 and is occasionally non-uniform, as expected given the deviations from neutrality along the genome.[20,38] No significant correlation was observed between our mutation-rate estimates and the density of IBD-segment sharing (see Results). To cope with imperfect detection of the IBD-segment boundaries, we excluded 0.5 cM on either side of the IBD segments from the analysis of mutations and gene-conversion rates, because we observed that inflation due to noisy boundary estimation plateaued for values larger than this threshold (Figure S11).

Demographic inference was performed with the software tool DoRIS.[14,32] The resulting demographic history is one of exponential expansion starting with an ancestral population size of 11,500

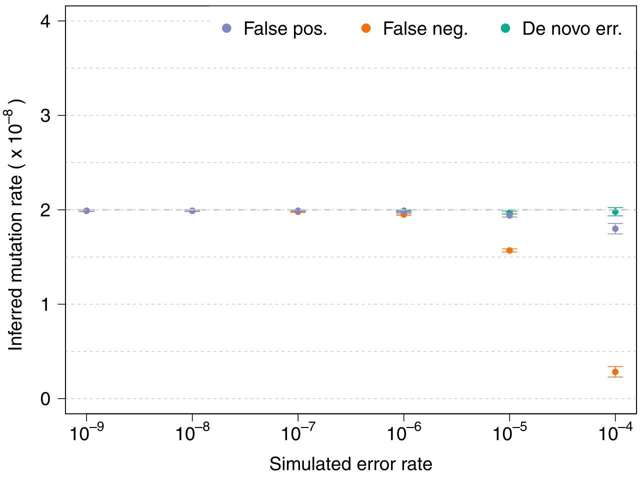

Figure 3. Inferred Mutation Rates under Several Values of Simulated Genotyping Error Rate for Three Types of Genotyping Errors

The simulated true underlying mutation rate was $\mu = 2 \times 10^{-8}$. All simulations involved a single chromosome of 250 cM for 200 haploid individuals from a GoNL-like population and used beta($\alpha = 0.5$, $\beta = 1$) as a prior for the allele frequency of erroneous variants. True IBD segments were extracted from the simulated ancestral recombination graph. Additional simulation results are shown in Figure S13. Error bars represent SE.

haploid individuals 150 generations in the past. Two periods of exponential expansion were inferred. The expansion rate between generations 150 and 10 was inferred to be 0.0146 and was followed by a strong expansion in the recent generations at a rate of 0.479 per generation. Because of the scarcity of extremely recent coalescent events, the magnitude of the latter expansion period was inferred with a high degree of uncertainty; however, this was observed to not have appreciable effects on the analysis described in the remainder of the paper (see Results).

Enrichment of Deleterious Variation in IBD Regions

We tested whether mutations arising between the present generation and the MRCA of IBD segments are enriched with deleterious variation. To this end, we ran the software tool ANNOVAR (version "2015Mar22"[39]) on the GoNL variants and obtained numeric scores for the PolyPhen-2 ("ljb23_pp2hvar"[40]) and Gerp++ ("gerp++gt2"[41]) annotations; we restricted the analysis to scores > 2 for the latter. To test for enrichment, we compared the average score of genome-wide variants to the average score of variants found to mismatch within IBD regions; we treated all variants as independent and reported Z test p values.

Analysis of Annotated Genomic Regions

Several sites along the genome were excluded from the analysis after application of the filtering criteria previously described. In addition, we analyzed mutation rates in specific regions described in several annotations (e.g., DNase I hypersensitive sites[42] and several others, as detailed in Table S3). It is sufficient to neglect regions that fall outside the genomic annotation at hand when computing the observed mismatch rate in the tMRCA regression. Annotations that are too small or too clustered in specific regions of the genome might result in downward biases of the estimated mutation rate because of the "inspection paradox" of the Poisson process underlying the model of IBD-segment sharing[14] (Figure S12). For this reason, our method cannot be used for inferring local mutation rates. Annotation-specific bias due to localization was computed and corrected with a permutation procedure (Table S3). Sequence context was accounted for with the trinucleotide context-specific mutation-rate matrix of Kryukov[43] (details in Table S3).

To derive mutation rates for different mutation categories (CpG or non-CpG and transition or transversion), we downloaded the ancestral alignment used in the 1000 Genomes Project[36] (see Web Resources). The ancestral allele for loci that were not present in this sequence (545,279 out of 12,181,714) was set to the major allele found in the 1000 Genomes dataset ($n = 300{,}503$) or set to the allele found in the human reference genome (UCSC Genome Browser hg19; see Web Resources) if monomorphic in the 1000 Genomes dataset ($n = 244{,}776$). We then computed mutation rates by using MaAF-threshold regression and excluding variants that did not match the analyzed mutation type (e.g., CpG transition) and scaled the resulting rate by the genomic fraction that might harbor the specific kind of mutation (e.g., CpG or non-CpG).

Results

Simulations

We evaluated the accuracy and robustness of the method via extensive coalescent simulation (see Material and Methods). To assess the impact of demographic history on our estimates, we simulated several plausible demographic scenarios and modeled genotyping errors by using a beta distribution with different parameters and specifying error rate at different allele frequencies (Figure S5). We extracted ground-truth shared IBD segments from the synthetic ancestral recombination graph and simulated three types of errors, referred to as de novo, false-positive, and false-negative errors (see Material and Methods; Figure 3). We observed that tMRCA regression is robust to the presence of substantial levels of de novo genotyping errors, consistent with the fact that IBD segments of different lengths are equally affected by the spurious sequence mismatches that result from errors of this kind. When we simulated false-positive genotyping errors, we observed our approach to be robust to errors up to a rate of $\sim 10^{-5}$ per base pair. False negatives were tolerated up to a frequency of $\sim 10^{-6}$. Very large values of false-positive or -negative genotyping error rates resulted in a downward bias of the estimates, which is due to the fact that IBD segments of different lengths harbor a slightly different spectrum of mismatching sites and are therefore not equally likely to be affected by spurious genotype calls (see Figure S1). Similar results were observed for several kinds of genotyping-error distributions, demographic models, and recombination maps, although the approach proved more robust for error distributions that are less concentrated on very rare variants (Figures S3–S5 and S13). The intercept of the tMRCA was observed to reflect genotyping error; average values were between 1× and 2× the simulated error rate (Figure S14), depending on the type of error and the parameters of the distribution used for selecting the frequency of affected alleles.

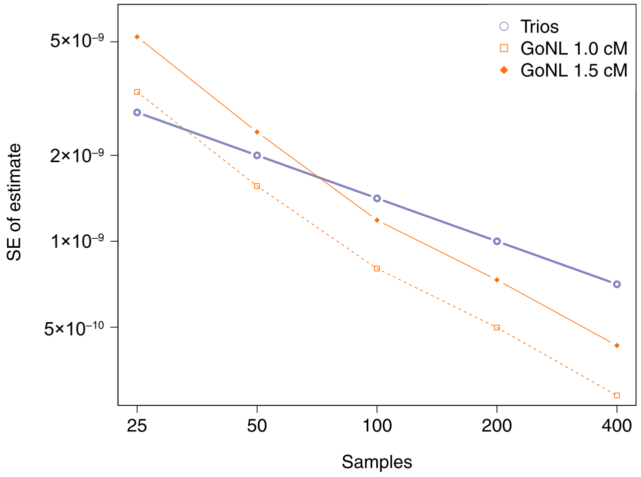

Figure 4. Comparison of the Estimated SE for Trios and tMRCA under Different Demographic Models and Minimum Cutoffs for IBD-Segment Length

We report the estimated SD from the analysis of several simulations of a single 100 Mb chromosome. For illustrative purposes, we show results of analyses using IBD-length cutoffs of 1.0 and 1.5 cM. Analysis of the GoNL data used a length cutoff of 1.6 cM.

We note that the proposed procedure estimates a historical sex-averaged mutation rate per base per generation, a quantity that might be affected by potential differences between the mutation and recombination rates of males and females over several generations in the past. We performed additional simulations to test whether sex-specific mutation and recombination rates and effective population sizes could bias our estimates. We determined that sex-specific variability of these parameters did not produce a bias (Table S4), given that the recovered estimate reflected a flat average of male and female mutation and recombination rates.

To compare the power of the proposed method to the power of trio-based mutation-rate inference, we simulated data at various sample sizes by using the GoNL demographic model. Because pairs sharing IBD segments increase quadratically as sample size increases, the proposed method results in smaller SEs than the trio-based approach, except at very small sample sizes (Figure 4). However, for demographic models that result in substantial IBD-segment sharing as a result of a small recent effective population size, higher sample size did not substantially decrease the SE (Figure S15). This is due to the fact that as new samples are added, early coalescent events result in overlapping ancestral lineages across pairs of individuals, so that limited new information is obtained from increasing the sample size.

We finally tested the MaAF-threshold-regression approach to correct biases introduced by non-crossover gene-conversion events and estimate the probability that a base pair is involved in gene conversion. We simulated realistic mutation and gene-conversion rates and used GERMLINE to detect IBD-segment sharing after subsampling synthetic SNPs in order to match the allele frequencies observed in the GoNL data. We observed good performance of the MaAF-threshold regression in recovering the simulated mutation-rate value (Figure 5) and observed a small downward bias when we recovered the gene-conversion rate by using the GERMLINE IBD-segment discovery parameters used in the real-data analysis (Figure S16).

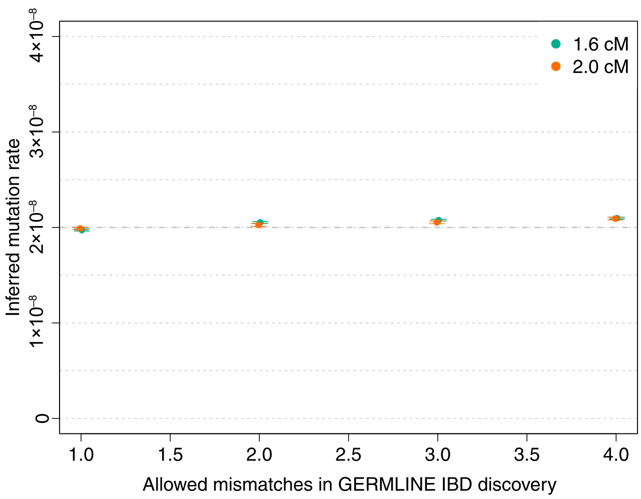

Figure 5. Inference of Gene-Conversion-Corrected Mutation Rate in Simulated Data

We simulated a chromosome of 50 cM for 250 diploid samples by using $\mu = 2 \times 10^{-8}$ for the mutation rate and a probability of 6×10^{-6} for a base pair to be involved in a non-crossover gene-conversion event. We matched the allele-frequency spectrum of the simulated samples to the spectrum found in real data for IBD-segment detection with GERMLINE. We used several values of the GERMLINE allowed mismatching sites ("-het") to assess the impact of this parameter in the results. Negligible biases were observed for the recovered mutation rate. Error bars represent SE.

Average Genome-wide Mutation Rate and Gene-Conversion Rate in the GoNL Dataset

We analyzed 498 founders that passed quality control in 250 trio families sequenced within the GoNL project (see Material and Methods). Because of the trio design of the GoNL study, the average ~13× sequencing depth is effectively doubled to ~26× for the transmitted haplotypes in the 498 analyzed founders. 248 trios and 2 duos passed sequencing quality control. In the remainder of the paper, we report results for the analysis of transmitted haplotypes only.

We estimated a mutation rate of $(2.08 \pm 0.06) \times 10^{-8}$ (Figure 6) by using IBD segments between 1.6 and 5.0 cM of length before correcting for gene-conversion events (hereafter, ± introduces a SE). For all analyses of mutation and gene-conversion rates, we discarded 0.5 cM on either edge of the segments and ignored variants with a trio-phasing and genotyping posterior value less than 1.0. Choosing more-conservative values for the minimum-length and edge-exclusion cutoffs resulted in compatible estimates (Figures S11, S17, and S18). As expected, including variants with lower trio-phasing and genotyping

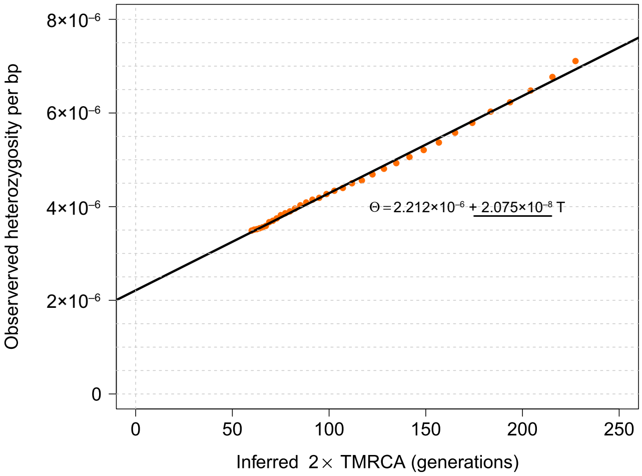

Figure 6. tMRCA Regression for Segments of Length ≥ 1.6 cM in the GoNL Dataset
The obtained slope is used for estimating mutation rate per generation per base pair before the effects of gene conversion are accounted for.

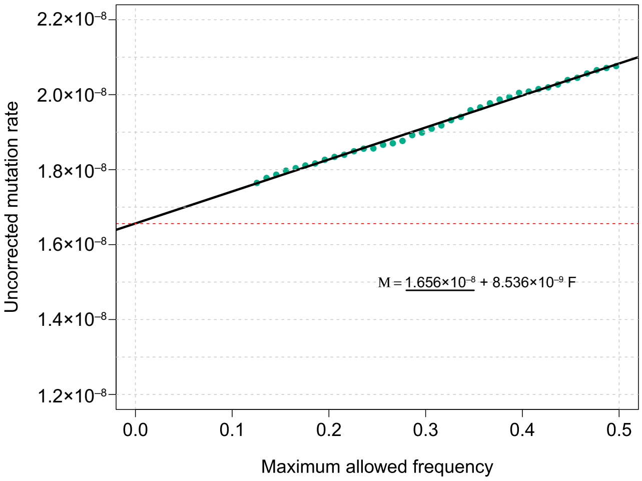

Figure 7. MaAF-Threshold Regression for Segments of Length ≥ 1.6 cM in the GoNL Dataset
We computed mutation rates for several allowed maximum-allele-frequency thresholds between 0.125 and 0.5 (green dots) and regressed the observed heterozygosity on the maximum allele frequency. The intercept of the resulting linear model reflects the corrected mutation-rate estimate.

posterior values resulted in higher estimates of genotyping error, but negligible effects were observed on the estimates of mutation rate (Figures S6–S9). The tMRCA-regression intercept, which reflects genotyping and phasing error rate (see Material and Methods), was estimated to be (2.21 ± 0.09) × 10^{-6}, within a range that is not expected to result in biases in the tMRCA-regression slope according to simulations (Figures 3, S6, S9, S13, and S14).

We then performed MaAF-threshold regression to correct for gene-conversion events (see Material and Methods; Figure 7). Using this approach, we estimated a genome-wide average mutation rate of (1.66 ± 0.04) × 10^{-8} per base per generation. Note that this represents a historical mutation rate, which includes effects such as average paternal age (see Discussion). Using segments up to 10 cM in length did not result in appreciable changes in our estimate: (1.66 ± 0.04) × 10^{-8} per base per generation (tMRCA regression is shown in Figure S19). Analysis performed with only the range of long IBD segments between 5.0 and 10.0 cM resulted in a compatible estimate (but a substantially larger SE). As expected given the simulated data (Tables S1 and S2), a concordant estimate was also obtained when the analysis was repeated with non-overlapping IBD-segment length bins in the tMRCA regression (and inverse-variance weighting of the observations) or with non-overlapping MaAF frequency bins. We truncated the MaAF regression to a conservative lower maximum allele frequency of 12.5% (see Material and Methods), given that including low MaAF values can result in downward biases as a result of the exclusion of recent mutation events (Figure S20). The mutation-rate estimates for each region and for regions of ~20 cM are shown in Tables S5 and S6. The difference between the corrected and uncorrected genome-wide estimates, (4.18 ± 0.48) × 10^{-9}, and the observed population heterozygosity of ~6.98 × 10^{-4} can be used for estimating the chance that a base pair is involved in a gene-conversion tract (see Material and Methods). We estimated that a base pair is involved in a gene-conversion event at a rate of (5.99 ± 0.69) × 10^{-6} per meiotic event. This rate is in good agreement with a recently published estimate of (5.9 ± 0.71) × 10^{-6}.[27]

We estimated the effects of uncertainty in the inferred model of demographic history on our estimates. When a genome-wide average mutation rate was inferred for a demographic model with ancestral population size perturbed by 10%, we observed a ~1.8% difference in the inferred average mutation rate. Larger variation in the ancestral population size was observed to have an approximately linear effect on our estimate (Table S8). We observed very limited effects on the mutation-rate estimate when we perturbed the present-day population size, which is inferred with uncertainty because of the scarcity of very recent coalescent events (Table S8).

Average Genome-wide Indel Rate

We applied the same procedure used for inferring mutation rates to infer the rate of <20 bp indels, which we estimated to be (1.26 ± 0.06) × 10^{-9}. This rate is higher than a recent estimate of 0.68 × 10^{-9}, reported in Kloosterman et al.,[44] but compatible with a second recent estimate of (1.5 ± 0.18) × 10^{-9},[25] both obtained via observation of de novo events in trios. We further divided the indels into insertions and deletions and estimated the rate of different classes as a function of their maximum length (Figure S21). We observed deletions to be about 50%–100% more frequent than insertions, depending on the length range. We additionally used our method to estimate the gene-conversion rate on the basis of indels and obtained a rate of (9.02 ± 2.91) × 10^{-6} per meiotic event, compatible with the rate obtained from point mutations.

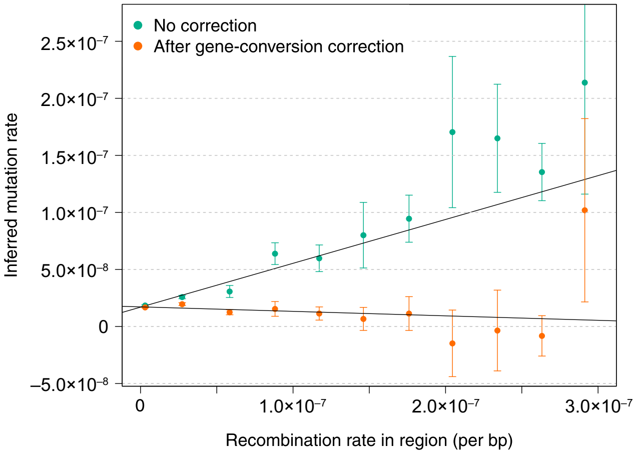

Figure 8. Association between Recombination Rate and Mutation Rate
We annotated the genome on the basis of uniform bins of recombination rate and estimated mutation rates for each obtained annotation. We observed a strong association between mutation and recombination rate before correcting for the occurrence of gene-conversion events. After applying the correction, we detected no significant association, which suggests that the linear relationship observed for the uncorrected estimates is induced by gene conversion (see Figure S22). Error bars represent SE.

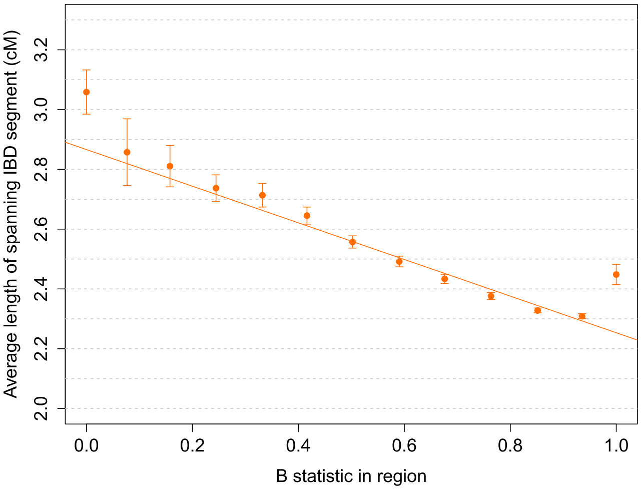

Figure 9. Relationship between Region-Specific Values of the B Statistic and the Average Length of ≥ 1.6 cM IBD Segments Spanning the Regions
Equally spaced bins of the B statistic were used. Reduced local effective population size has similar effects on the B statistic and the length of IBD haplotypes, which are longer in regions of strong background selection as a result of earlier average coalescent times between pairs of individuals. Error bars represent SE.

Recombination Does Not Strongly Affect Mutation Rate

We used our approach to analyze annotation-specific mutation rates (see Material and Methods). We looked for association between recombination rates and mutation rates, a relationship that has been previously detected and attributed to mutagenic properties of recombination.[45] Indeed, we found our tMRCA-regression estimates of mutation rate to be strongly associated with recombination rate ($\beta = 0.38 \pm 0.04$ mutations/recombination, F-test $p = 5.27 \times 10^{-6}$, $R^2 = 0.9$; Figure 8). As previously mentioned, however, an increased sequence-mismatch rate at loci that undergo frequent recombination might be a result of polymorphic variants introduced by gene-conversion events, which might increase the slope of the tMRCA regression. Consistently, after controlling for gene conversion, we observed no significant association between recombination rate and mutation rate ($\beta = -0.04 \pm 0.03$, F-test $p = 0.17$), suggesting a lack of observable mutagenic effects associated with recombination hotspots (Figures 8 and S22). A recent study reached similar conclusions.[5] Repeating the same analysis with indels, we detected no significant association between indel rate and recombination rate for either tMRCA-regression slope or MaAF-threshold-regression intercept ($\beta = -0.003 \pm 0.003$, F-test $p = 0.37$ after gene-conversion correction).

Effects of Background Selection

Natural selection affecting new mutations can reduce genomic variation, leading to downward bias in our mutation-rate estimates. Because our analysis is limited to mutation events that occurred in the past ~100 generations, given the length of IBD segments, we expect the effects of natural selection on our genome-wide average mutation rate estimate to be small. Genomic regions with functional or regulatory roles, however, might be under selective pressures that might result in a measurable impact even at these short time scales.

To estimate the impact of selective pressures on our estimates, we divided the genome on the basis of the B statistic proposed in McVicker et al.[4] The B statistic measures the impact of background selection on a genomic region by estimating the ratio between local effective population size and the effective population size expected under neutrality, such that small values of the B statistic correspond to higher selective pressures (see page 11 in McVicker et al.[4] for details on the computation of the B statistic). Similarly, a local reduction in effective population size affects the spectrum of shared IBD segments, which are expected to be longer on average, as a result of early coalescent events in populations of smaller effective size.[20,38] Indeed, we observed a strong correspondence between small values of the B statistic and the average length of IBD segments (F-test $p = 8.43 \times 10^{-7}$; Figure 9). As expected, the effect is such that smaller values of the B statistic correspond to longer average shared IBD segments as a result of reduced local effective population size. This effect is remarkably strong up to the measured genome-wide average value of the B statistic. We observed longer average IBD segments for large values of the B statistic, a result that might be explained by biases in either of the two measures or by the fact that additional evolutionary forces, such as selection acting on standing genetic variation,[38] are being captured by IBD-segment lengths. When we measured the impact of different values of the B statistic on our estimates

Table 1. Analyses of PolyPhen-2 and Gerp++ Annotated Variants: Genome-wide versus Mismatching within IBD Segments

	Genome-wide	Mismatching in IBD Segments
PolyPhen-2 Results		
Annotated variants	54,960	1,843
Mean score	0.41 ± 0.0018	0.45 ± 0.0099
Gerp++ (>2) Results		
Annotated variants	948,782	27,900
Mean score	3.08 ± 0.00098	3.11 ± 0.0059

of mutation rate, however, we found the effect to not be significant ($\beta = [2.17 \pm 1.55] \times 10^{-9}$ mutations per generation per unit of B statistic, F-test $p = 0.19$; Figure S23). The genome-wide average B statistic was estimated to be 0.78. If we were to correct the estimated average genome-wide mutation rate to account for this, we would obtain an updated average mutation rate of $(1.7 \pm 0.05) \times 10^{-8}$, which is not significantly different from the uncorrected estimate. In addition to these analyses, we tested for significant correlation with the mutation rate inferred for each region or for sub-regions of size 10, 20, 30, or 40 cM. The correlation was found to not be significant for the average value of the B statistic in the region, the recombination rate, or the average density of IBD sharing.

Sequence Differences in IBD Segments Are Enriched with Deleterious Variation

Mutation events occurring within the analyzed IBD regions are expected to have arisen within the past ~100 generations and are therefore on average substantially younger than variants randomly sampled along the genome. Several recent studies have outlined the recent origin of a large fraction of functionally relevant variants.[46–48] We therefore tested whether the presence of recent mutations on IBD segments results in more deleterious variants than in the average genome-wide locus by contrasting average scores obtained from PolyPhen-2[40] and Gerp++[41] annotations (see Material and Methods). Of the analyzed GoNL variants, 54,960 were annotated with PolyPhen-2, and 948,782 were annotated with Gerp++; of these, 1,843 and 27,900, respectively, were found to be mismatched on IBD segments of 1 cM or longer. When we compared average scores, we found that mismatching sites within IBD regions were strongly enriched with higher scores in both annotations (PolyPhen-2 Z-test $p = 2.8 \times 10^{-5}$; Gerp++ Z-test $p = 9.03 \times 10^{-10}$; Table 1). We further found a marginal association between PolyPhen-2 scores and the B statistic of background selection ($\beta = -0.074 \pm 0.025$, $p = 0.014$, $R^2 = 0.39$) and a strong association between Gerp++ scores and regional B statistics ($\beta = -0.734 \pm 0.068$, $p = 3.55 \times 10^{-7}$, $R^2 = 0.91$; Figure S24), which is expected because both measures rely on metrics related to sequence conservation.

We finally tested for enrichment or depletion of the mutation rate in several genomic annotations that have recently been extracted from several studies (Material and Methods; Table S3). None of the annotations were significantly enriched with or depleted of mutation rates after we controlled for trinucleotide context and multiple hypothesis testing. A recent paper[49] found that cell-specific chromatin features are a strong determinant of cancer mutations. On the other hand, our estimated mutation rate of $(1.66 \pm 0.05) \times 10^{-8}$ in DNase I hypersensitive regions suggests that the germline mutation rate is not substantially different from the genome-wide average in these regions, in line with recent analyses.[9] We further computed estimates of the rate of mutations at CpG and non-CpG sites (Table S7) and found them to be higher than in previous reports according to trio analysis, consistent with a higher genome-wide rate (see Table 2 in Kong et al.[24]).

Discussion

We propose a method for estimating mutation and gene-conversion rates from genealogical relationships across the past tens to few hundreds of generations. This approach is robust to substantial amounts of genotyping error, which is an important confounder for many recent mutation-rate estimators based on trio data. Using this method, we inferred a genome-wide average point mutation rate of $(1.66 \pm 0.04) \times 10^{-8}$ per base per generation, which is significantly higher than several recent family-based estimates ranging from 1.0×10^{-8} to 1.2×10^{-8} per base per generation.[7,10,11] Family-based methods have the advantage of relying on direct observation of de novo mutation events while making minimal modeling assumptions but are currently affected by the need to rely on strict filtering criteria to deal with false-positive and -negative genotype calls, and this could in part or entirely explain the discrepancy with our results. We note that the approach of Campbell et al.[23] is similar in spirit to ours, because it uses de novo mutations on long stretches of recently arisen autozygosity within individuals from a known pedigree. However, the autozygosity reflects identity by descent at a more recent time scale, and the authors still mainly rely on stringent filtering criteria to avoid false-positive genotype calls, thereby incurring the same potential biases as other trio-based studies. Phylogenetic methods, on the other hand, fall within the range of 2.0×10^{-8} to 2.5×10^{-8} per base per generation.[50,51] These estimates rely on several underlying modeling assumptions, which provide a possible explanation for the higher inferred rates, although some have suggested the possibility that these analyses might capture the results of evolutionary changes of the mutation rate across populations or the effects of a varying length of generation times.[7,10,52] Converting between per-year and per-generation estimates requires making assumptions on the sex-averaged generation length.[10] This is generally done with indirect evidence, which complicates the comparison of different estimates. If we assume a sex-averaged long-term generation length of 29 years,[53] we can convert our inferred sex-averaged per-generation rate to (5.71 ± 0.14) × 10^{-10} per base per year. Fu et al.[54] used ancient DNA to estimate a range of 0.4 × 10^{-9} to 0.6 × 10^{-9} per year, which is slightly lower than our result, but the reported confidence intervals are compatible. Similarly, Sun et al.[55] computed a rate of 1.4 × 10^{-8} to 2.3 × 10^{-8} per base per generation on the basis of point mutations near microsatellites, which is also compatible with our estimate. A contemporary study[56] related in spirit to ours used simulation-based calibration of the decay of heterozygosity along the genome to infer an average genome-wide mutation rate of (1.61 ± 0.13) × 10^{-8}, which matches our estimated value. The authors discuss several implications of this mutation rate on the ability to reconcile demographic events inferred with DNA analysis and fossil records, which apply to our analysis as well.

Because conversion between sequence divergence and phylogenetic split times across different primate species relies on an estimate of the per-year mutation rate, different values of this rate have a direct impact on our ability to reconstruct the timing of these events.[7,10] If we assume no significant effects of generation time and no changes in mutation rates, our estimate (in conjunction with additional data from Table S5 of Prado-Martinez et al.[57]) implies that the split between humans and chimpanzees occurred ~6.6 million years ago and that the split between humans and orangutans occurred about ~19.5 million years ago. When we used our estimate of mutation rate to interpret recently reported split times across human populations,[8] we found dates that are compatible with what has been inferred by methods other than DNA-based reconstruction. The split of African and non-African populations is estimated to have occurred 46,000–61,000 years ago, whereas a split time of 15,000 years ago is inferred for the separation of East Asians and Native American populations. These estimates are lower than those obtained under the assumption of a smaller mutation rate, but they do not contradict current fossil evidence.

Note that the several different available estimates might disagree not only because of statistical uncertainty and possible biases induced by violations of modeling assumptions but also as a result of differences in the underlying quantity being estimated. Our approach aims at measuring the sex-averaged, genome-averaged, and time-averaged mutation and gene-conversion rates per base per generation. As pointed out in several recent studies,[10,23,24,55,58] paternal age at conception is an important determinant of sex-averaged mutation rates, and it is interesting to ask whether variation in historical paternal age might at least partially explain the discrepancy between our estimate and that obtained in recent pedigree studies. In Kong et al.,[24] the authors reported that the paternal age in Iceland between 1650 and 1900 was ~36 years, significantly higher than the average paternal age of ~30 years for the contemporary samples they analyzed. We found that even if we conservatively assume the per-year paternal-age effect β_y from Kong et al.[24]—which is higher than the β_y value from other studies[10,23,55,58]—and a drop from the historical paternal age of 36 years to the contemporary paternal age of 30 years, then our extrapolated estimate decreases to 1.43 × 10^{-8} (Table S9). Thus, in the absence of additional evidence for historical average age variation in the analyzed samples, this observation alone might not fully explain the difference between our estimate and those reported in recent pedigree studies, although it outlines the importance of taking this additional source of variation into account in a comparison of estimates obtained from different methods. We note that paternal-age-related differences in per-generation estimates of the mutation rate might also affect the previously described conversion between the per-generation and per-year scales.

In addition to estimating the rate of point mutations, we report a gene-conversion rate of (5.99 ± 0.69) × 10^{-6} per base per generation, in close agreement with a recent report,[27] and have found that recombination is not associated with mutation rates, supporting recent findings.[5] A recent sperm-typing study further dissected the relationship among mutation, recombination, and gene conversion and found evidence of both higher mutational load in regions of high recombination and repairing mechanisms associated with gene conversion.[59] These lead to a higher prevalence of GC alleles than of AT alleles. Overall, these effects might be counteracting each other in a way that results in minimal differences in the total number of observed mutations in recombination-rich regions while affecting sequence composition. Interestingly, a recent study has reported that recombination rate affects the distribution of putatively deleterious variants along the genome but found no evidence suggesting a role of biased gene conversion in this observation.[60]

Finally, we applied our method to estimate the rate of short (<20 bp) indels, which have not thus far been extensively characterized. We inferred a rate of (1.26 ± 0.06) × 10^{-9}, compatible with two previous estimates of (1.5 ± 0.18) × 10^{-9} from Besenbacher et al.[25] and (1.06 ± 0.1) × 10^{-9} from Ramu et al.[61] but higher than the estimate of 0.68 × 10^{-9} reported in Kloosterman et al.[44] Although these analyses are most likely affected by difficulties related to detection of short indels, collectively they suggest that insertions and deletions occur at a significantly lower rate than do single point mutations.

In addition to analyzing genome-wide average rates, we looked for enrichment or depletion of mutation rates in a number of genomic annotations that were recently derived from several studies. Although we cannot exclude significant deviations from genome-wide averages, we found no evidence of changes in overall mutation rates for the analyzed regions. Notably, although the distribution of shared IBD haplotypes closely reflects the effects of background selection along the genome, we observed a negligible effect on our estimated mutation rates, suggesting that estimating mutation rates by using mutation events under the effects of ~100 generations of natural selection does not significantly bias local mutation-rate estimates in European populations. Consistent with the idea that mutations on IBD segments are recent and under the effects of selective forces,[46–48] we found a strong enrichment of deleterious variants within IBD regions.

Our method provides a way of studying mutation and gene-conversion events in large samples of unrelated individuals because it is robust to substantial amounts of genotyping error, which limits other approaches. A main limitation is the need to rely on two fundamental components—namely, detecting shared IBD segments and inferring the recent demographic history for the analyzed population—that are potential sources of bias. Our analysis of mutation rates in the GoNL dataset relies on IBD detection and demographic inference performed in a previous study,[34] but it is possible that additional sources of uncertainty in these two components affect our results. Our conservative exclusion of substantial portions of IBD segments, together with our sensitivity analysis for changes in the demographic model, however, suggests that these biases, if present, should not be substantial.

Several potential directions for improvement of the proposed methodology and analysis can be outlined. First, additional developments of the coalescent calculations used in this work can remove the requirement of estimating a demographic model for the analyzed samples.[62] Second, it might be possible to devise more-sophisticated weighting schemes for dealing with heteroscedasticity in the regressions and develop additional modeling for dealing with any small deviations from linearity that demographic variation might induce in the MaAF regression. Third, alternative genotype-calling strategies (e.g., individual-based calling) can be employed for reducing these effects of the relationship between allele frequency and genotyping error rates. Finally, applying the tMRCA regression approach proposed in this paper might make it possible to analyze multi-generation pedigrees (e.g., Campbell et al.[23]) while controlling for substantial genotyping error. In this scenario, in fact, IBD calling and the inference of tMRCA for IBD segments are substantially simplified.

Future improvements of sequencing technologies and methods for downstream analysis will lead to accurate and direct characterization of biological properties of the processes leading to mutation and gene-conversion events. These advances will also shed light on the discrepancy between previous pedigree-based mutation-rate estimates and those obtained by our methods and will enable testing whether cross-population differences exist. In particular, it will be possible to test whether false-negative de novo genotype calls due to stringent filtering criteria lead to systematically lower mutation-rate estimates in pedigree-based studies (we currently believe this to be the most plausible explanation for the observed discrepancy). Accordingly, we expect that improved sequence quality and analysis will lead trio-based studies to detect a higher number of de novo mutations. Because our methodology relies on evidence from several generations in the past, it is sensitive to additional historical parameters, such as variation in the average paternal age or changes of the mutation rate itself. Although we cannot exclude that historical variation in these quantities might play a role in the higher mutation-rate estimate we obtained, our method relies on evidence from a relatively small number of generations, and it seems less plausible that substantial variation might be observed in such a short time span. Future methodological developments, however, might enable testing of these hypotheses. On the basis of the analysis we described, we believe that several pedigree-based estimates available to date might not accurately reflect the historical mutation rate, particularly in the context of demographic reconstruction, where a higher rate should be assumed.

Appendix A

The Age of IBD Segments

If a pair of chromosomes share a common ancestor at time t generations before present, the probability that a single site is spanned by an IBD segment of length l at least u Morgans can be expressed as

$$\sigma(t) = \int_u^\infty l(2t)^2 e^{-2tl} dl = e^{-2tu}(2tu+1). \qquad \text{(Equation A1)}$$

The distribution $l(2t)^2 e^{-2tl}$ represents the sum of two exponential random variables with parameter $2t$, which is the rate at which a recombination occurs on either side of the chosen site. Note that this assumes that an IBD segment is delimited by the occurrence of recombination events, which is equivalent to assuming an underlying sequentially Markovian coalescent (SMC) model. For very short IBD segments (e.g., <0.3 cM) and in populations that experience substantial and long-lasting isolation (e.g., $N_e < 1{,}000$), the slightly more-complex SMC′ model[63] provides more-accurate calculations.[8,64] This is, however, unnecessary given the demographic history and length ranges here considered. It follows from the linearity of the expectation operator that the expected genomic fraction $f(t)$ shared identically by descent by a pair of individuals whose ancestral lineages coalesce at time t can be obtained from the probability that a single site is spanned by an IBD segment of length at least u Morgans, which we write $f(t) = \sigma(t)$. The expected length of an IBD segment transmitted from a common ancestor living at time t is therefore

$$\ell(t) = \int_u^\infty l \times 2te^{2t(u-l)}\, dl = 1/(2t) + u. \qquad \text{(Equation A2)}$$

To determine the expected number of IBD segments obtained if the lineages of two individuals coalesce at time t, we therefore divide the expected total amount of genome shared identically by descent by the expected length of an IBD segment co-inherited from an ancestor living at time t. This yields

$$\begin{aligned} n_u(t) &= Lf_u(t)/\ell(t) \\ &= \frac{Le^{-2tu}(2tu+1)}{1/(2t)+u} \\ &= L2e^{-2tu}t, \end{aligned} \qquad \text{(Equation A3)}$$

where L is the size, in Morgans, of the considered genomic region. To obtain the expected number of IBD segments longer than u Morgans for the average pair of individuals in the population, we marginalize over the distribution of pairwise coalescence times, $c(t)$, which depends on the demographic history,

$$n_u = \int_0^\infty c(t)n_u(t)dt. \qquad \text{(Equation A4)}$$

This quantity has a closed-form expression if we assume that the population size becomes constant at an arbitrarily remote point in time, and we can use it to obtain the posterior age distribution of IBD-segment ages,

$$p_u(t) = \frac{c(t)n_u(t)}{n_u}. \qquad \text{(Equation A5)}$$

Contribution of Individual Variants to Heterozygosity

For a sample of K homologous sequences from a population, the heterozygosity per site can be estimated by

$$\widehat{\theta} = \frac{1}{s}\sum_{i=1}^{s}\frac{K}{K-1}2\frac{x_i}{K}\left(1-\frac{x_i}{K}\right), \qquad \text{(Equation A6)}$$

(see Nei[65]), where s is the number of sites in each sequence, x_i is the number of samples carrying a derived allele at site i, and $K/(K-1)$ is a bias-correction factor. Defining $M(x)$ as the total number of sites in the sample for which exactly x sequences carry a derived allele, we can rewrite this equation as a sum over x:

$$\widehat{\theta} = \sum_{x=1}^{K-1}\frac{M(x)}{s}\frac{2x(K-x)}{K(K-1)}. \qquad \text{(Equation A7)}$$

The term $M(x)/s$ is the proportion of sites at which x of the K sequences carry a derived allele, and the term

$$\frac{2x(K-x)}{K(K-1)} \qquad \text{(Equation A8)}$$

is the probability of discovering such a polymorphic site when just two sequences are sampled without replacement from the K sequences. Note that this probability is the same for sites with x copies of a derived allele as it is for sites with $K-x$ copies. Thus, we can also write

$$\widehat{\theta} = \sum_{x=1}^{[K/2]}\frac{M(x)+M(K-x)}{s}\frac{2x(K-x)}{K(K-1)}, \qquad \text{(Equation A9)}$$

where $[K/2]$ is the largest integer that is less than or equal to $K/2$, and x is now the count of the minor allele.

This allows us to consider the average contribution of different kinds of polymorphic sites to overall heterozygosity. Under the model of constant population size and neutral evolution of Watterson,[66]

$$E[M(x)] = \frac{s\theta}{x} \qquad \text{(Equation A10)}$$

(see Fu[67]), in which $\theta = 4N\mu$ is the diploid population-scaled mutation rate per site, or the expected per-site heterozygosity of the population. Using Equation A10 together with Equation A9 and simplifying gives

$$E\left[\widehat{\theta}\right] = \sum_{x=1}^{(K-1)/2}\theta\frac{2}{K-1}, \qquad \text{(Equation A11)}$$

in which we assume that K is odd for simplicity. The sum in Equation A11 evaluates to θ, as expected for an unbiased estimator.

Equation A11 shows that, on average, the different kinds of polymorphic sites, categorized by minor allele frequency, contribute uniformly to heterozygosity, as noted previously by Kruglyak and Nickerson.[68] Another way of stating this is that polymorphisms discovered by screening in samples of size two will be uniformly distributed among classes of minor allele frequencies. This depends on genotype-calling criteria and a constant population size over time and is also not true for derived allele-frequency classes. Figure S2 shows that contributions to heterozygosity are close to uniform for the GoNL site-frequency spectrum.

Supplemental Data

Supplemental Data include 24 figures and 9 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.10.006.

Consortia

The members of the Genome of the Netherlands Consortium are Laurent C. Francioli, Androniki Menelaou, Sara L. Pulit, Freerk van Dijk, Pier Francesco Palamara, Clara C. Elbers, Pieter B.T. Neerincx, Kai Ye, Victor Guryev, Wigard P. Kloosterman, Patrick Deelen, Abdel Abdellaoui, Elisabeth M. van Leeuwen, Mannis van Oven, Martijn Vermaat, Mingkun Li, Jeroen F.J. Laros, Lennart C. Karssen, Alexandros Kanterakis, Najaf Amin, Jouke Jan Hottenga, Eric-Wubbo Lameijer, Mathijs Kattenberg, Martijn Dijkstra, Heorhiy Byelas, Jessica van Setten, Barbera D.C. van Schaik, Jan Bot, Isac J. Nijman, Ivo Renkens, Tobias Marschall, Alexander Schnhuth, Jayne Y. Hehir-Kwa, Robert E Handsaker, Paz Polak, Mashaal Sohail, Dana Vuzman, Fereydoun Hormozdiari, David van Enckevort, Hailiang Mei, Vyacheslav Koval, Matthijs H. Moed, K. Joeri van der Velde, Fernando Rivadeneira, Karol Estrada, Carolina Medina-Gomez, Aaron Isaacs, Steven A. McCarroll, Marian Beekman, Anton J.M. de Craen, H. Eka D. Suchiman, Albert Hofman, Ben Oostra, Andr G. Uitterlinden, Gonneke Willemsen, LifeLines Cohort Study, Mathieu Platteel, Jan H. Veldink, Leonard H. van den Berg, Steven J. Pitts, Shobha Potluri, Purnima Sundar,

David R. Cox, Shamil R. Sunyaev, Johan T. den Dunnen, Mark Stoneking, Peter de Knijff, Manfred Kayser, Qibin Li, Yingrui Li, Yuanping Du, Ruoyan Chen, Hongzhi Cao, Ning Li, Sujie Cao, Jun Wang, Jasper A. Bovenberg, Itsik Pe'er, P. Eline Slagboom, Cornelia M. van Duijn, Dorret I. Boomsma, Gert-Jan B van Ommen, Paul I.W. de Bakker, Morris A. Swertz, and Cisca Wijmenga.

Acknowledgments

We express our gratitude to Nick Patterson, Priya Moorjani, Mark Lipson, Amy Williams, and two anonymous reviewers for useful scientific discussions and comments on an early draft and to Ilya Shlyakhter for support with the COSI2 simulator. This research was funded by NIH grant R01 MH101244. H.K.F. was supported by the Fannie and John Hertz Foundation. S.S. was supported by NIH grant K99 GM111744.

Received: June 12, 2015
Accepted: October 13, 2015
Published: November 12, 2015

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes ancestral alignments, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/

ARGON, https://github.com/pierpal/ARGON

IBDMUT, https://github.com/pierpal/IBDMUT

UCSC Genome Browser hg19 assembly, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/

References

1. Crow, J.F. (2000). The origins, patterns and implications of human spontaneous mutation. Nat. Rev. Genet. *1*, 40–47.
2. Arnheim, N., and Calabrese, P. (2009). Understanding what determines the frequency and pattern of human germline mutations. Nat. Rev. Genet. *10*, 478–488.
3. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.
4. McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. *5*, e1000471.
5. Schaibley, V.M., Zawistowski, M., Wegmann, D., Ehm, M.G., Nelson, M.R., St Jean, P.L., Abecasis, G.R., Novembre, J., Zöllner, S., and Li, J.Z. (2013). The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome Res. *23*, 1974–1984.
6. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.
7. Scally, A., and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. *13*, 745–753.
8. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. *46*, 919–925.
9. Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn, C.M., Swertz, M., Wijmenga, C., van Ommen, G., et al.; Genome of the Netherlands Consortium (2015). Genome-wide patterns and properties of de novo mutations in humans. Nat. Genet. *47*, 822–826.
10. Ségurel, L., Wyman, M.J., and Przeworski, M. (2014). Determinants of mutation rate variation in the human germline. Annu. Rev. Genomics Hum. Genet. *15*, 47–70.
11. Campbell, C.D., and Eichler, E.E. (2013). Properties and rates of germline mutations in humans. Trends Genet. *29*, 575–584.
12. Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. Nat. Genet. *40*, 1068–1075.
13. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res. *19*, 318–326.
14. Palamara, P.F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. Am. J. Hum. Genet. *91*, 809–822.
15. Ralph, P., and Coop, G. (2013). The geography of recent genetic ancestry across Europe. PLoS Biol. *11*, e1001555.
16. Browning, B.L., and Browning, S.R. (2013). Detecting identity by descent and estimating genotype error rates in sequence data. Am. J. Hum. Genet. *93*, 840–851.
17. Gudbjartsson, D.F., Sulem, P., Helgason, H., Gylfason, A., Gudjonsson, S.A., Zink, F., Oddson, A., Magnusson, G., Halldorsson, B.V., Hjartarson, E., et al. (2015). Sequence variants from whole genome sequencing a large group of Icelanders. Sci Data *2*, 150011.
18. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
19. Browning, B.L., and Browning, S.R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics *194*, 459–471.
20. Gusev, A., Palamara, P.F., Aponte, G., Zhuang, Z., Darvasi, A., Gregersen, P., and Pe'er, I. (2012). The architecture of long-range haplotypes shared within and across populations. Mol. Biol. Evol. *29*, 473–486.
21. Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science *328*, 636–639.
22. Conrad, D.F., Keebler, J.E., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., et al.; 1000 Genomes Project (2011). Variation in genome-wide mutation rates within and between human families. Nat. Genet. *43*, 712–714.
23. Campbell, C.D., Chong, J.X., Malig, M., Ko, A., Dumont, B.L., Han, L., Vives, L., O'Roak, B.J., Sudmant, P.H., Shendure, J., et al. (2012). Estimating the human mutation rate using autozygosity in a founder population. Nat. Genet. *44*, 1277–1281.
24. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature *488*, 471–475.

25. Besenbacher, S., Liu, S., Izarzugaza, J.M., Grove, J., Belling, K., Bork-Jensen, J., Huang, S., Als, T.D., Li, S., Yadav, R., et al. (2015). Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat. Commun. *6*, 5969.
26. Busing, F.M., Meijer, E., and Van Der Leeden, R. (1999). Delete-m jackknife for unequal m. Stat. Comput. *9*, 3–8.
27. Williams, A.L., Genovese, G., Dyer, T., Altemose, N., Truax, K., Jun, G., Patterson, N., Myers, S.R., Curran, J.E., Duggirala, R., et al.; T2D-GENES Consortium (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. eLife *4*, e04637.
28. Wiuf, C., and Hein, J. (2000). The coalescent with gene conversion. Genetics *155*, 451–462.
29. Odenthal-Hesse, L., Berg, I.L., Veselis, A., Jeffreys, A.J., and May, C.A. (2014). Transmission distortion affecting human noncrossover but not crossover recombination: a hidden source of meiotic drive. PLoS Genet. *10*, e1004106.
30. Shlyakhter, I., Sabeti, P.C., and Schaffner, S.F. (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. Bioinformatics *30*, 3427–3429.
31. Liang, L., Zöllner, S., and Abecasis, G.R. (2007). GENOME: a rapid coalescent-based whole genome simulator. Bioinformatics *23*, 1565–1567.
32. Palamara, P.F., and Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. Bioinformatics *29*, i180–i188.
33. He, Z., Li, X., Ling, S., Fu, Y.-X., Hungate, E., Shi, S., and Wu, C.-I. (2013). Estimating DNA polymorphism from next generation sequencing data with high error rate by dual sequencing applications. BMC Genomics *14*, 535.
34. Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. *46*, 818–825.
35. Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487.
36. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
37. Duret, L., and Arndt, P.F. (2008). The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. *4*, e1000071.
38. Albrechtsen, A., Moltke, I., and Nielsen, R. (2010). Natural selection and the distribution of identity-by-descent in the human genome. Genetics *186*, 295–308.
39. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
40. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
41. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.
42. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.
43. Kryukov, G.V., Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am. J. Hum. Genet. *80*, 727–739.
44. Kloosterman, W.P., Francioli, L.C., Hormozdiari, F., Marschall, T., Hehir-Kwa, J.Y., Abdellaoui, A., Lameijer, E.-W., Moed, M.H., Koval, V., Renkens, I., et al.; Genome of Netherlands Consortium (2015). Characteristics of de novo structural changes in the human genome. Genome Res. *25*, 792–801.
45. Hellmann, I., Ebersberger, I., Ptak, S.E., Pääbo, S., and Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. *72*, 1527–1535.
46. Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science *337*, 100–104.
47. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al.; NHLBI Exome Sequencing Project (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.
48. Kiezun, A., Pulit, S.L., Francioli, L.C., van Dijk, F., Swertz, M., Boomsma, D.I., van Duijn, C.M., Slagboom, P.E., van Ommen, G.J., Wijmenga, C., et al.; Genome of the Netherlands Consortium (2013). Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS Genet. *9*, e1003301.
49. Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M.S., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature *518*, 360–364.
50. Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics *156*, 297–304.
51. Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87.
52. Harris, K. (2015). Evidence for recent, population-specific evolution of the human mutation rate. Proc. Natl. Acad. Sci. USA *112*, 3439–3444.
53. Fenner, J.N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. *128*, 415–423.
54. Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L., Aximu-Petri, A., Prüfer, K., de Filippo, C., et al. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. Nature *514*, 445–449.
55. Sun, J.X., Helgason, A., Masson, G., Ebenesersdóttir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., and Stefansson, K. (2012). A direct characterization of human mutation based on microsatellites. Nat. Genet. *44*, 1161–1165.
56. Lipson, M., Loh, P.-R., Sankararaman, S., Patterson, N., Berger, B., and Reich, D. (2015). Calibrating the human mutation rate via ancestral recombination density in diploid genomes. bioRxiv, http://dx.doi.org/10.1101/015560.

57. Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., et al. (2013). Great ape genetic diversity and population history. Nature *499*, 471–475.

58. Jiang, Y.H., Yuen, R.K., Jin, X., Wang, M., Chen, N., Wu, X., Ju, J., Mei, J., Shi, Y., He, M., et al. (2013). Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. Am. J. Hum. Genet. *93*, 249–263.

59. Arbeithuber, B., Betancourt, A.J., Ebner, T., and Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proc. Natl. Acad. Sci. USA *112*, 2109–2114.

60. Hussin, J.G., Hodgkinson, A., Idaghdour, Y., Grenier, J.-C., Goulet, J.-P., Gbeha, E., Hip-Ki, E., and Awadalla, P. (2015). Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nat. Genet. *47*, 400–404.

61. Ramu, A., Noordam, M.J., Schwartz, R.S., Wuster, A., Hurles, M.E., Cartwright, R.A., and Conrad, D.F. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. Nat. Methods *10*, 985–987.

62. Palamara, P.F. (2014). Population genetics of identity by descent. PhD thesis (Columbia University).

63. Marjoram, P., and Wall, J.D. (2006). Fast "coalescent" simulation. BMC Genet. *7*, 16.

64. Harris, K., and Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. *9*, e1003521.

65. Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics *89*, 583–590.

66. Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. *7*, 256–276.

67. Fu, Y.-X. (1995). Statistical properties of segregating sites. Theor. Popul. Biol. *48*, 172–197.

68. Kruglyak, L., and Nickerson, D.A. (2001). Variation is the spice of life. Nat. Genet. *27*, 234–236.

The American Journal of Human Genetics

Supplemental Data

Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates

Pier Francesco Palamara, Laurent C. Francioli, Peter R. Wilton, Giulio Genovese, Alexander Gusev, Hilary K. Finucane, Sriram Sankararaman, Genome of the Netherlands Consortium, Shamil R. Sunyaev, Paul I.W. de Bakker, John Wakeley, Itsik Pe'er, and Alkes L. Price

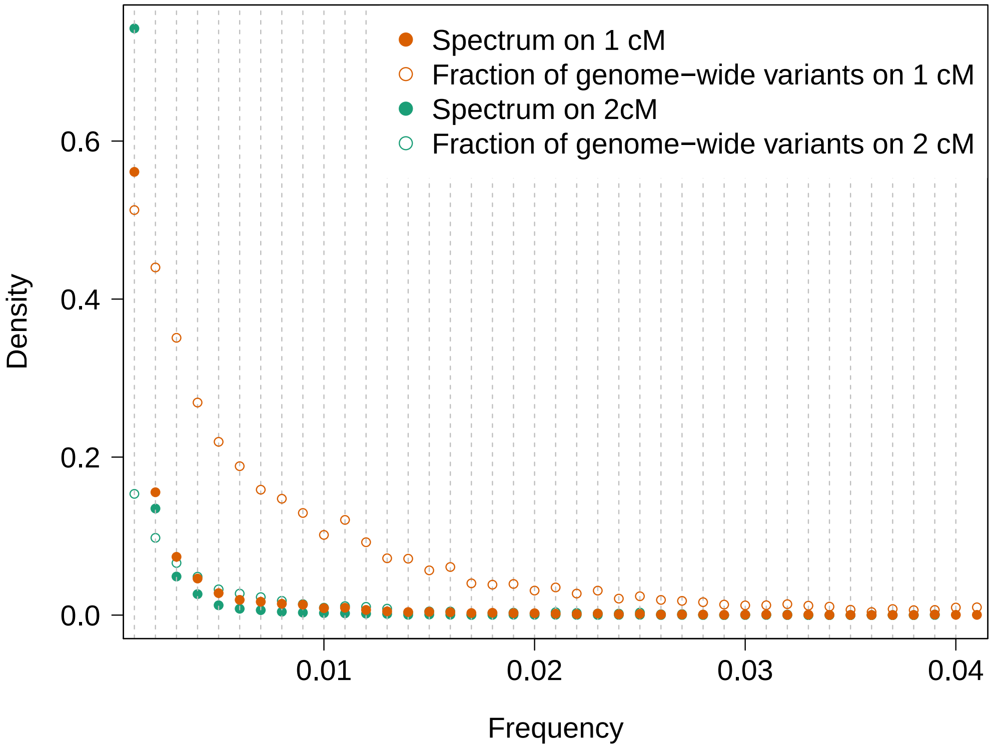

Figure S1: Simulated frequency spectra on IBD segments of different legnths. We computed the allele frequency spectrum of mismatching sites due to new mutation events occurring on IBD segments. Empty dots represent the fraction of the total genome-wide variants of a specific frequency that are found heterozygous on the IBD segments. Simulations were performed using the reconstructed GoNL demographic model.

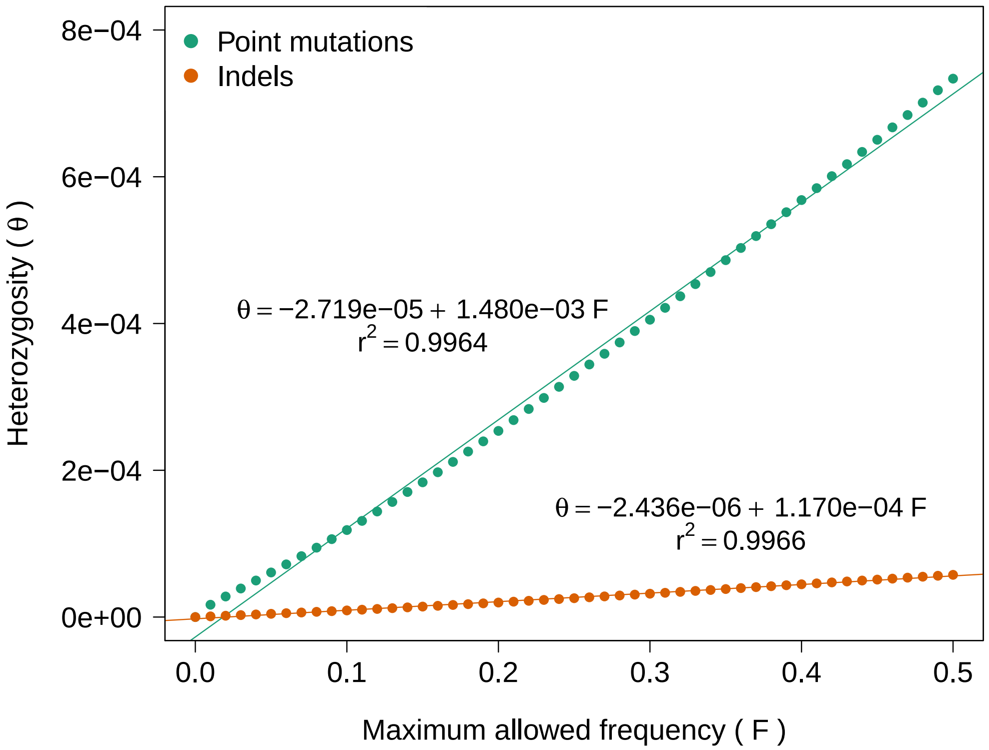

Figure S2: Approximately uniform contribution of variants of different frequencies to overall heterozygosity for both point mutations and indels in the GoNL dataset. Small deviations from linearity may be caused by demographic history (at both recent and remote time scales).

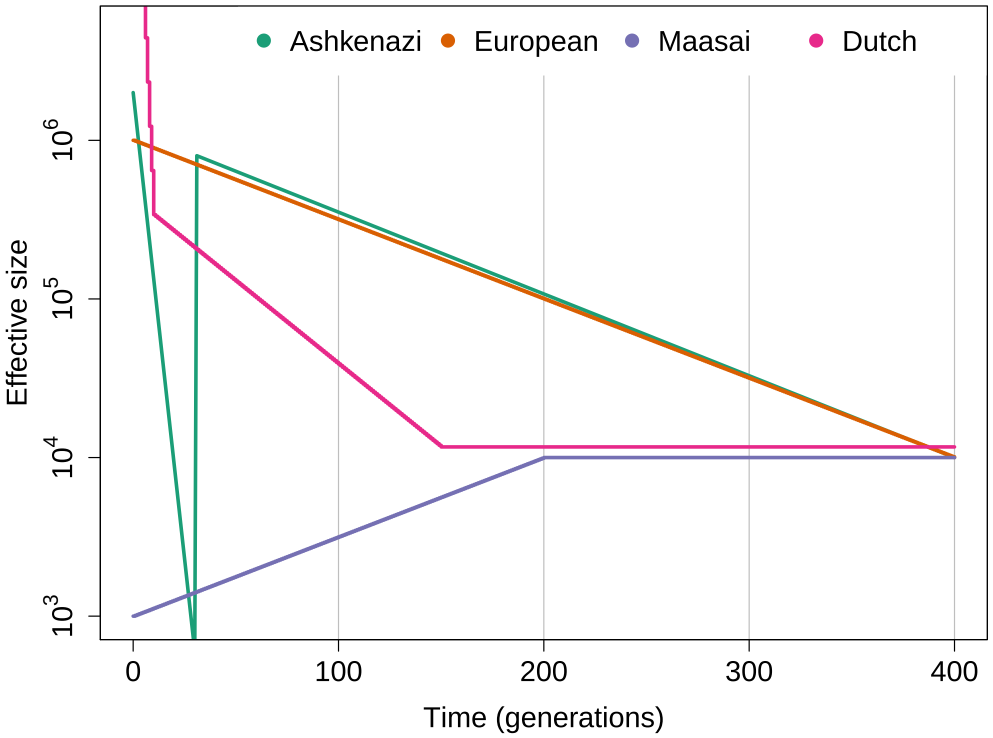

Figure S3: Demographic models inferred for the GoNL data or adopted in simulations.

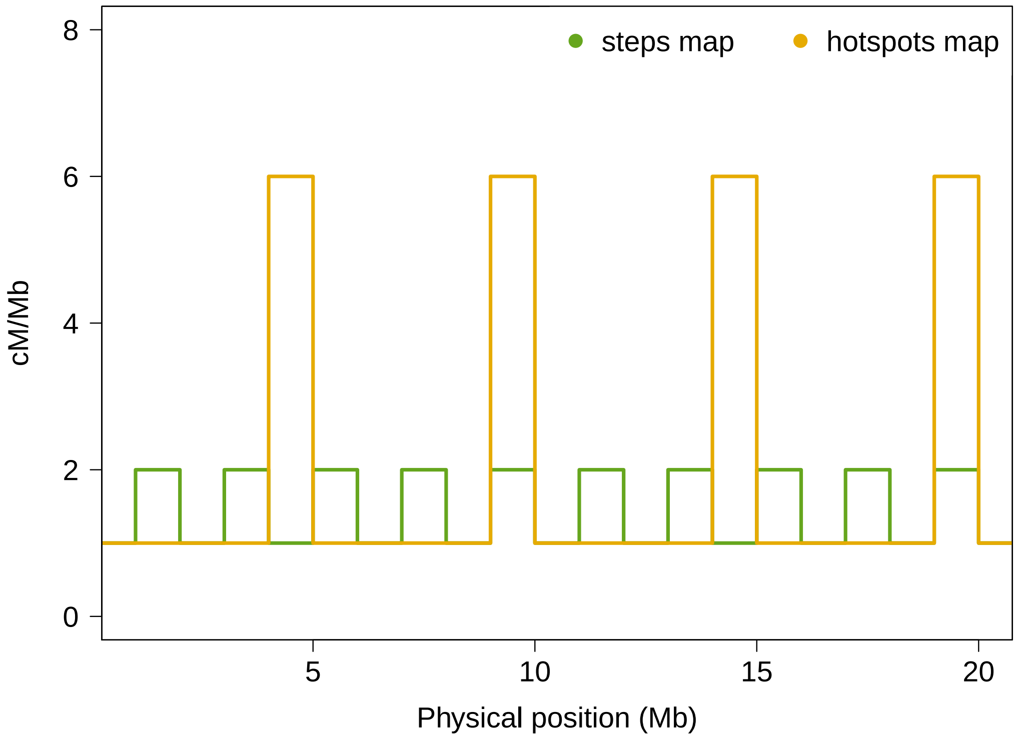

Figure S4: Genetic maps adopted in simulations.

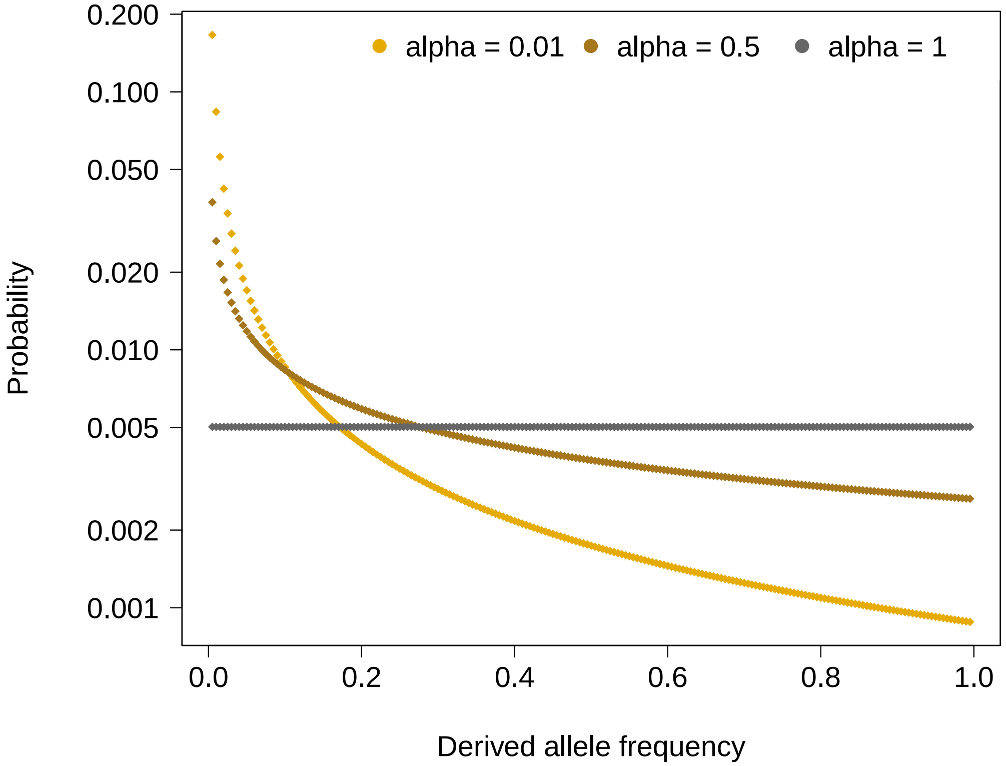

Figure S5: Distributions adopted to sample the frequency of spurious genotyping calls in simulated data. The beta distribution $Beta(\alpha, \beta)$ was used with $\beta = 1$ and α as specified in the Legend. For "de-novo" false positive errors, the frequency determines the number of individuals that are affected by an erroneous genotype call. For false-positive/negative genotyping errors, the sampled frequency corresponds to the frequency of the allele that is chosen to add/remove erroneous genotype calls. Three shape parameters were tested for the beta distribution: $\alpha = 0.01$, $\alpha = 0.5$, resulting in a strong preference for rare variants being erroneously called, and $\alpha = 1$, resulting in a uniform distribution.

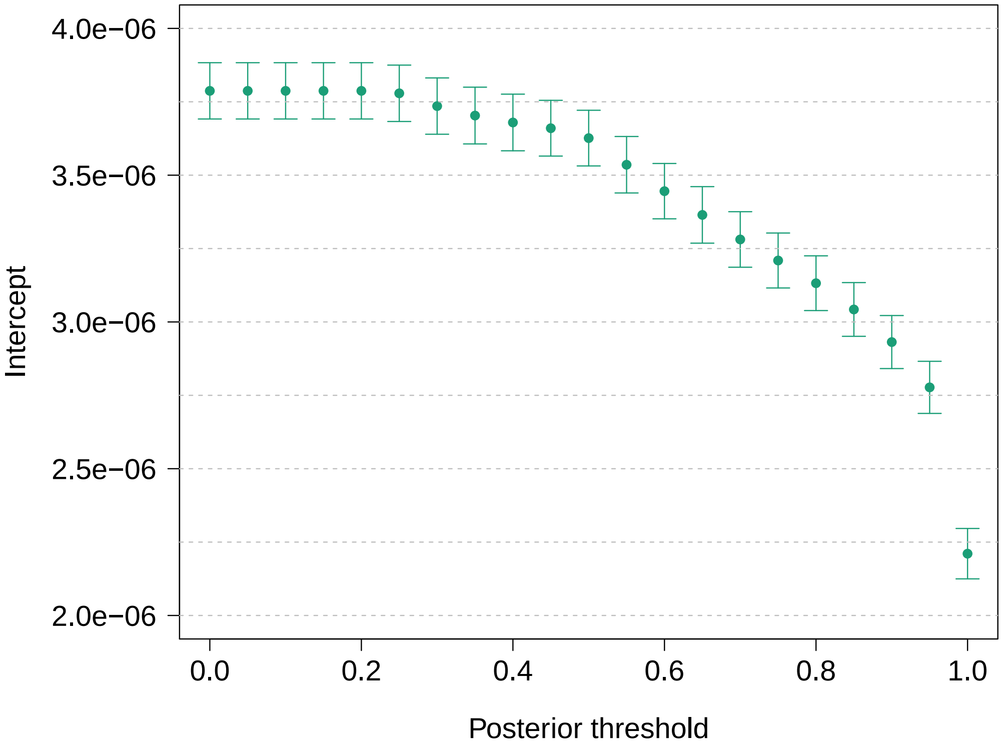

Figure S6: Estimated intercept of the tMRCA regression in the GoNL dataset for segments of length at least 1.6 cM, as a function of the minimum required MVNCall posterior quality for observed heterozygous sites. The MVNCall program used to trio-phase the analyzed data outputs posterior probabilities that capture uncertainty about genotyping and phasing calls. To test the robustness of our approach to the effects of genotype uncertainty, we computed mutation rates excluding from the analysis variants for which the MVNCall posterior was lower than a chosen threshold in IBD regions. Lower values of the posterior threshold resulted in a larger intercept of the tMRCA regression, reflecting higher genotyping error.

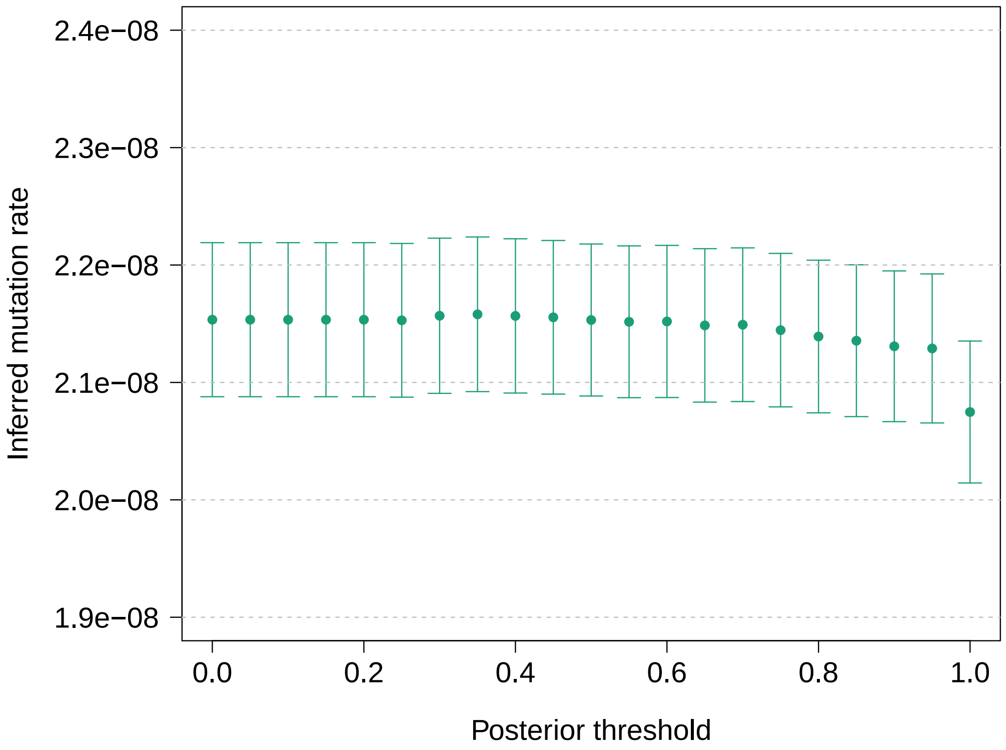

Figure S7: Estimated slope of the tMRCA regression in the GoNL dataset for segments of length at least 1.6 cM, as a function of the minimum required MVNCall posterior quality for observed heterozygous sites. Different values of the MVNCall posterior threshold, resulting in higher estimated genotyping error rates (Figure S8), did not significantly affect the estimated mutation rate. tMRCA estimates are inflated due to uncorrected effects of gene conversion, for which MaAF-threshold regression is adopted.

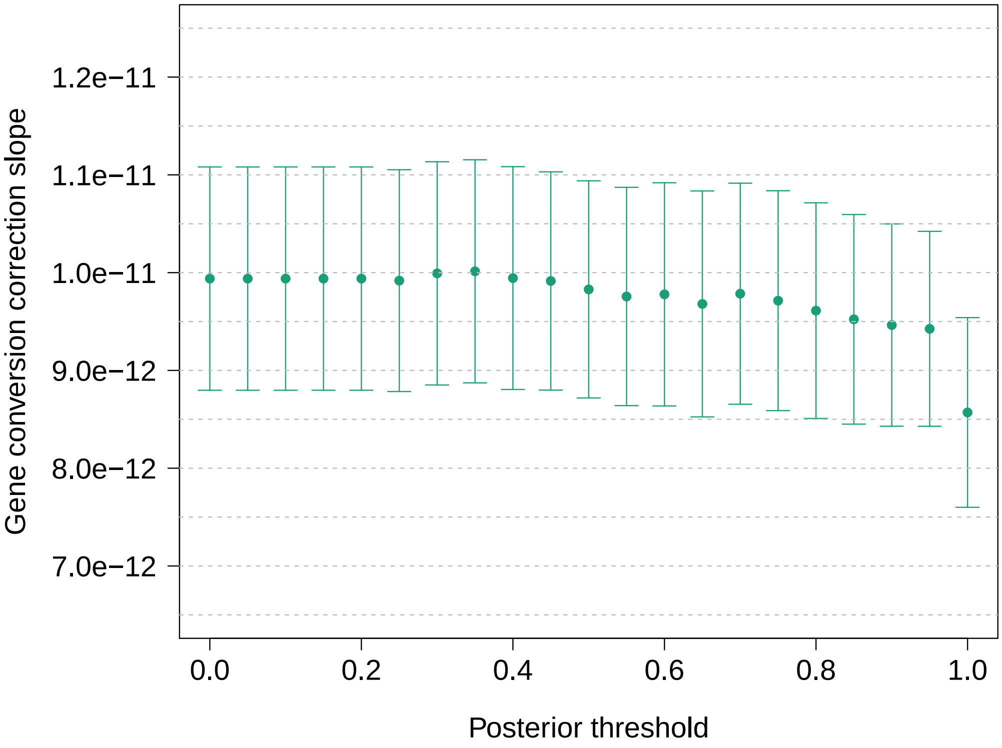

Figure S8: Estimated slope of the MaAF-threshold regression performed to correct for gene conversion in the GoNL dataset for segments of length at least 1.6 cM, as a function of the minimum required MVNCall posterior quality for observed heterozygous sites. Minimal variation is observed as the MVNCall posterior threshold is changed.

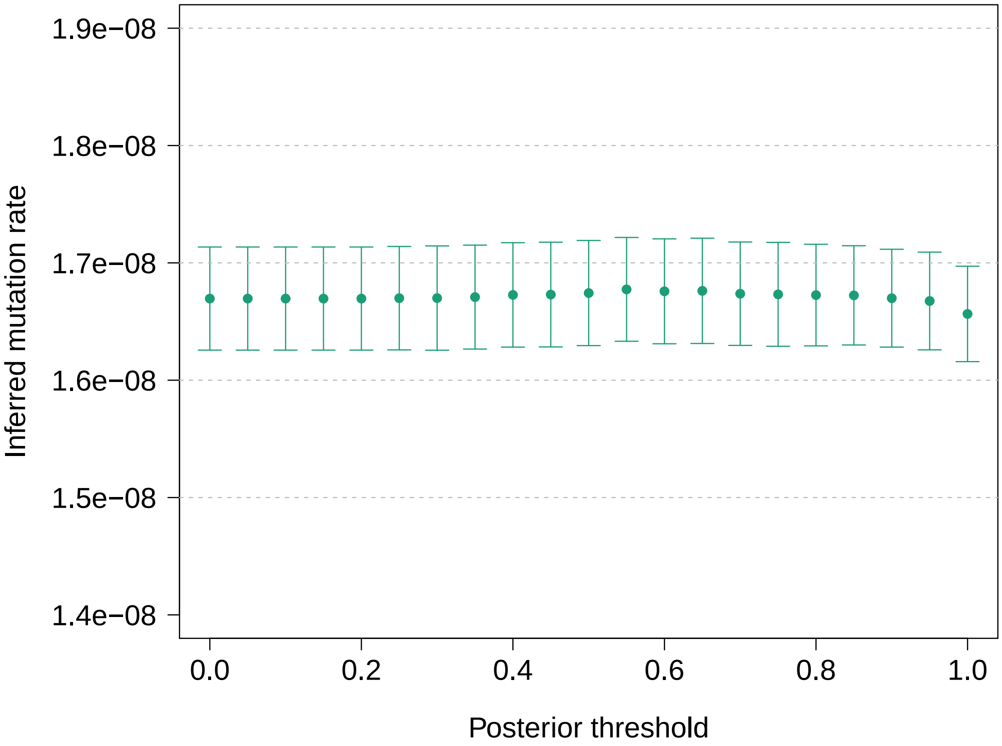

Figure S9: Estimated intercept of the MaAF-threshold regression in the GoNL dataset for segments of length at least 1.6 cM, as a function of the minimum required MVNCall posterior quality for observed heterozygous sites. We observed no significant impact of the chosen MVNCall posterior threshold on the inferred average genome-wide mutation rate.

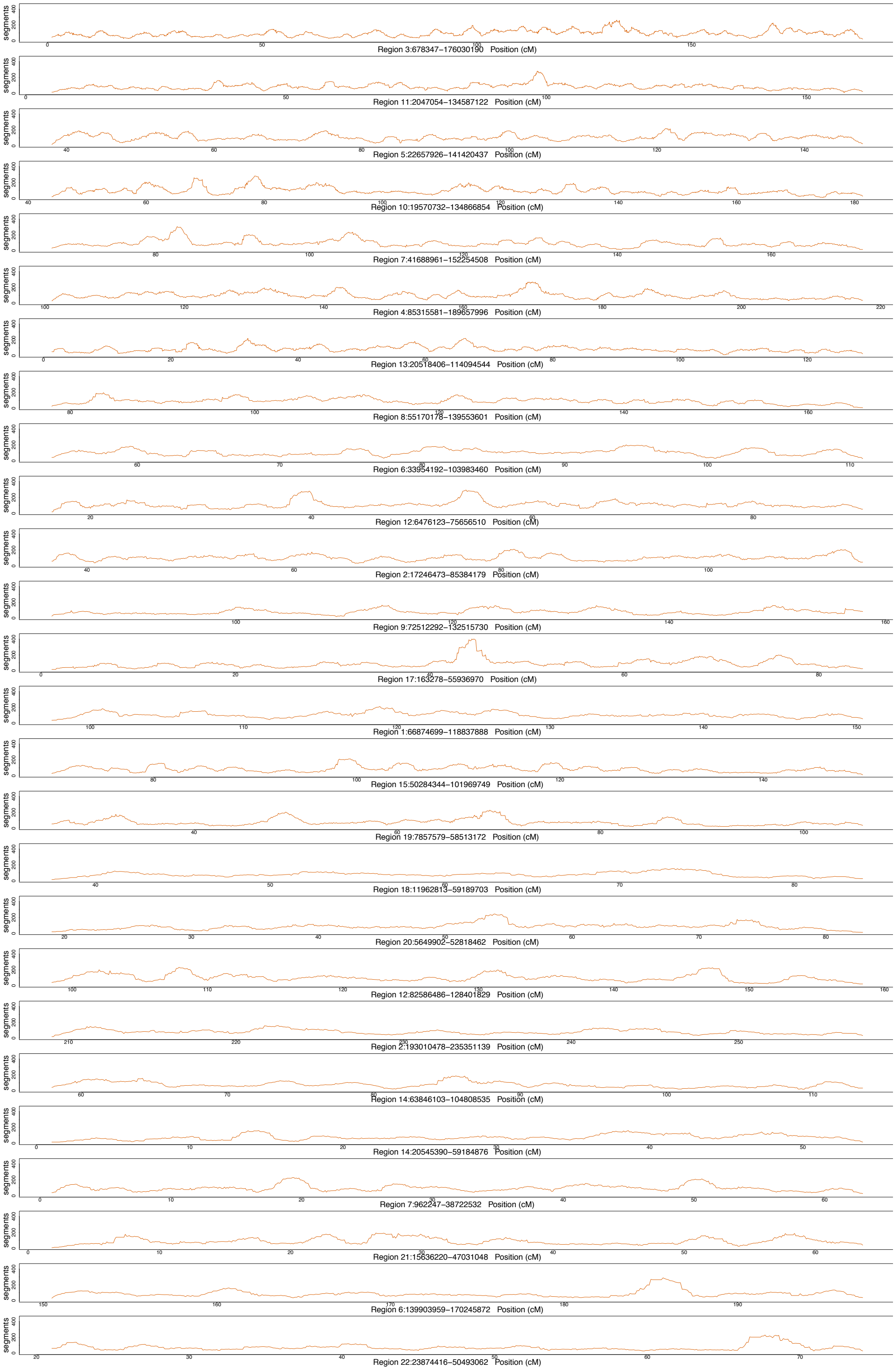

Figure S10: Region-specific density of ≥ 1.6 IBD segments.

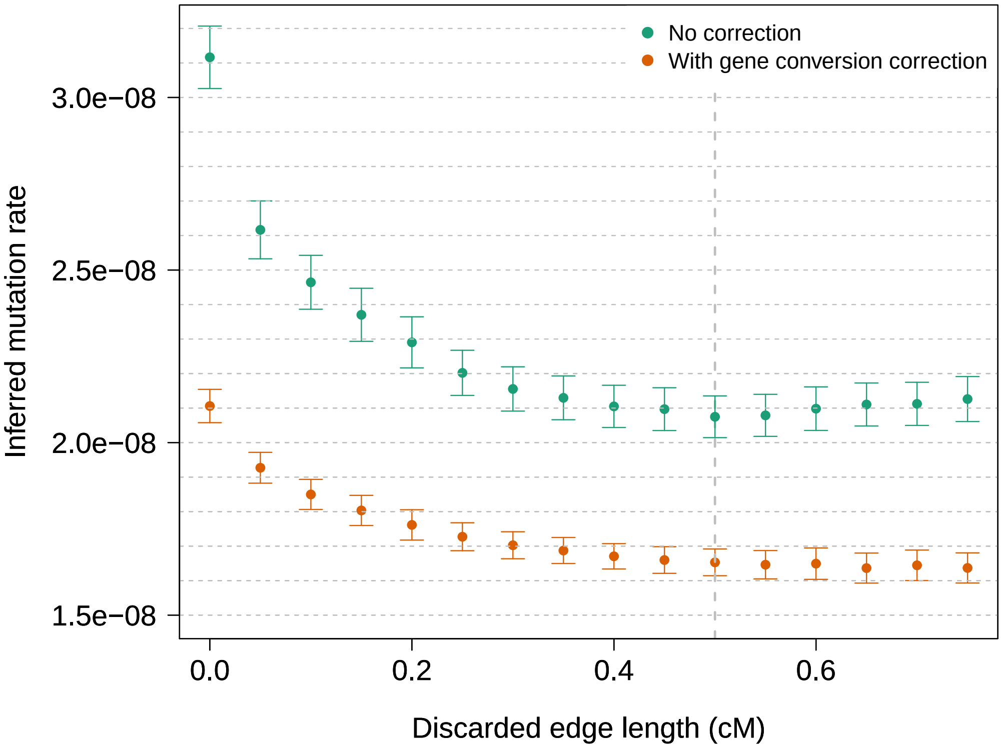

Figure S11: Gene conversion-corrected and uncorrected mutation rates inferred for segments longer than 1.6 cM in the GoNL data set, as a function of the size of discarded IBD segment edge. Inferred values become stable when > 0.5 cM edges are exluded.

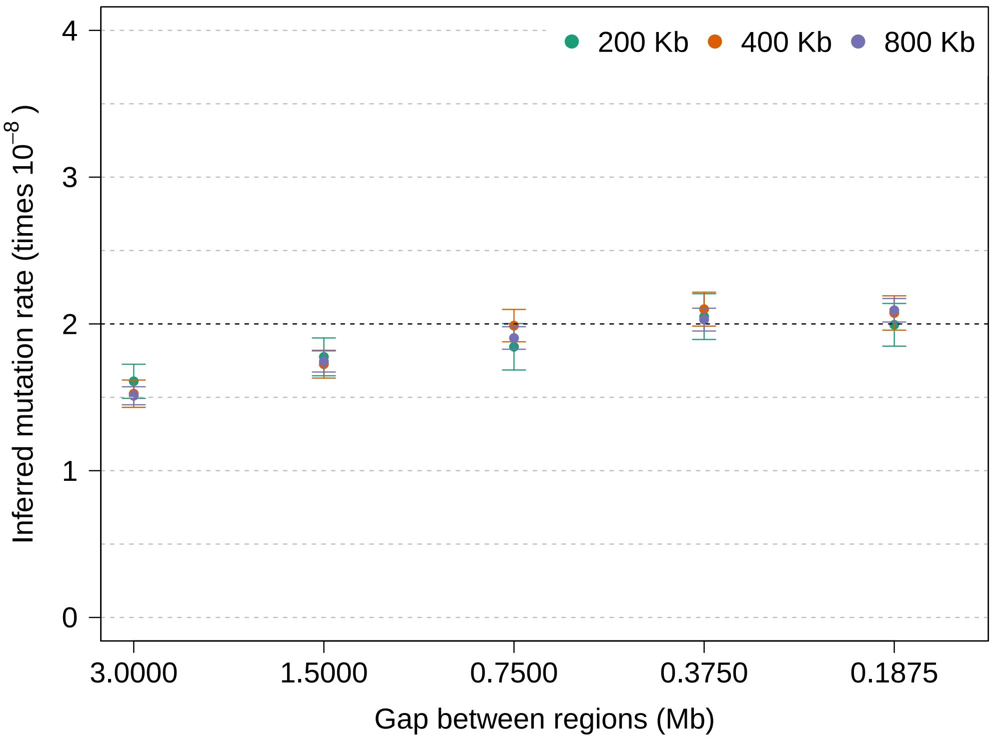

Figure S12: We observed a downward bias when we simulated annotations that are extremely localized, with a large average distance between the analyzed regions. This occurs due to the fact that in our approach we are estimating the age of chromosome-wide IBD segments of a specified length, rather than the age of segments spanning a small genomic region. Due to the "inspection paradox" of Poisson processes, the length distribution of IBD segments spanning individual sites differs from that of segments spanning large regions such as chromosomes. To quantify and correct the resulting bias, we randomly shifted the tested annotation along the analyzed chromosomal regions and computed the ratio between the mutation rate obtained from random shifting and the genome-wide mutation estimate. The computed correction factor was used to correct for the observed bias in real data analysis (see Table S3).

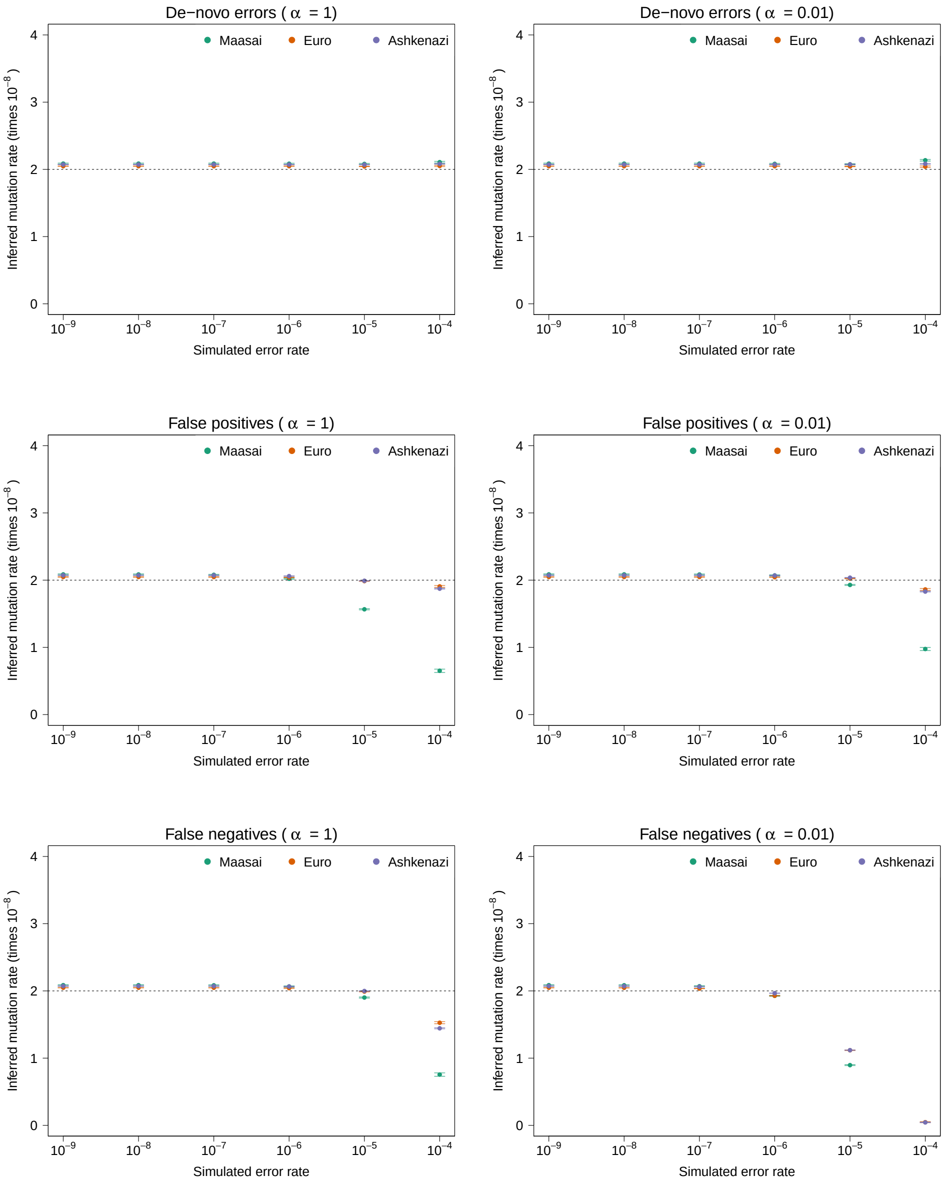

Figure S13: Inferred mutation rates for several values of simulated genotyping error rate, for several types of genotyping errors, demographic history and prior distribution for the frequency of spurious calls. The simulated true underlying mutation rate was $\mu = 2 \times 10^{-8}$. All simulations involved a single chromosome of 250 cM for 100 diploid individuals. The "steps" recombination map was adopted (Figure S4). Analogous results, omitted from this summary, were obtained for the "hotspots" recombination map.

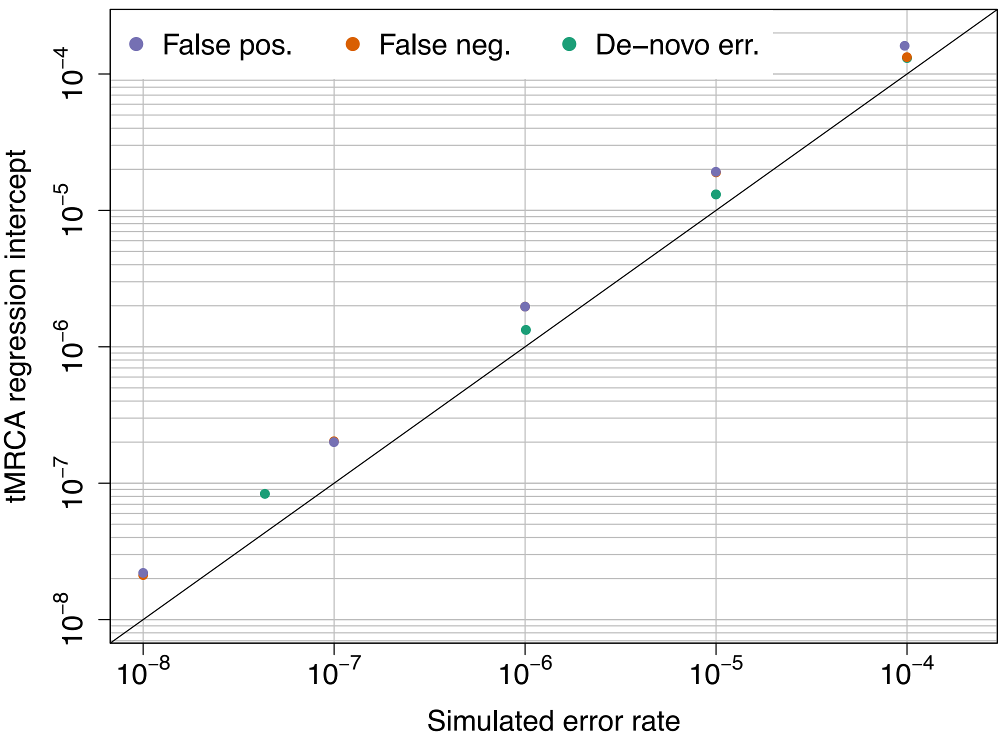

Figure S14: We simulated a chromosome of 250 cM for 100 diploid samples and introduced several types and magnitudes of sequencing errors using the GoNL demographic model. In all cases we used the beta distribution with parameter 0.5 as a prior for the frequency of simulated errors (Figure S5), and the "steps" recombination map (Figure S4). We report the intercept from the tMRCA regression as a function of simulated error.

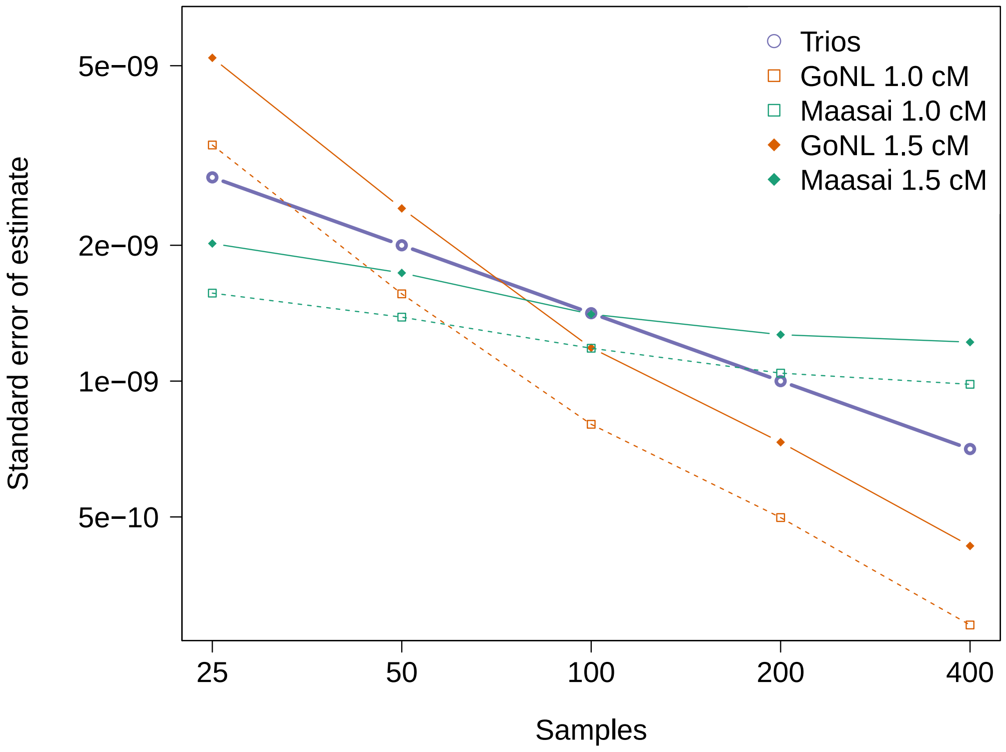

Figure S15: Comparison of the estimate standard error for trios and tMRCA under different demographic models and minimum IBD segment length cutoffs. We report the estimated standard deviation from the analysis of several simulations of a single 100 Mb chromosome.

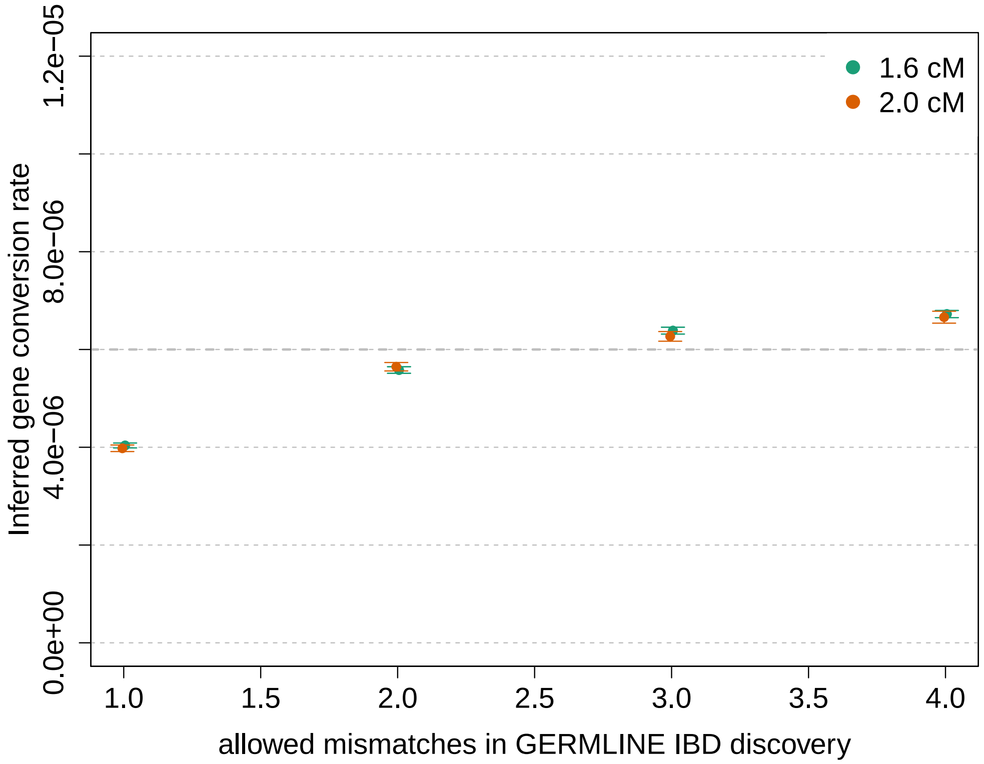

Figure S16: We simulated a chromosome of 50 cM for 250 diploid samples, using $\mu = 2 \times 10^{-8}$ for the mutation rate and a probability of 6×10^{-6} for a basepair to be involved in a non-crossover gene conversion event. We matched the allele frequency spectrum of the simulated samples to the spectrum found in real data for IBD detection with GERMLINE. We used several values of the GERMLINE allowed mismatching sites ("-het") to asses the impact of this parameter in the results. Using a stringent "-het" value of 1, we observed a downwards bias in the estimated gene converion rate. A small bias is observed for higher values, including "-het 2" used in the real data analysis.

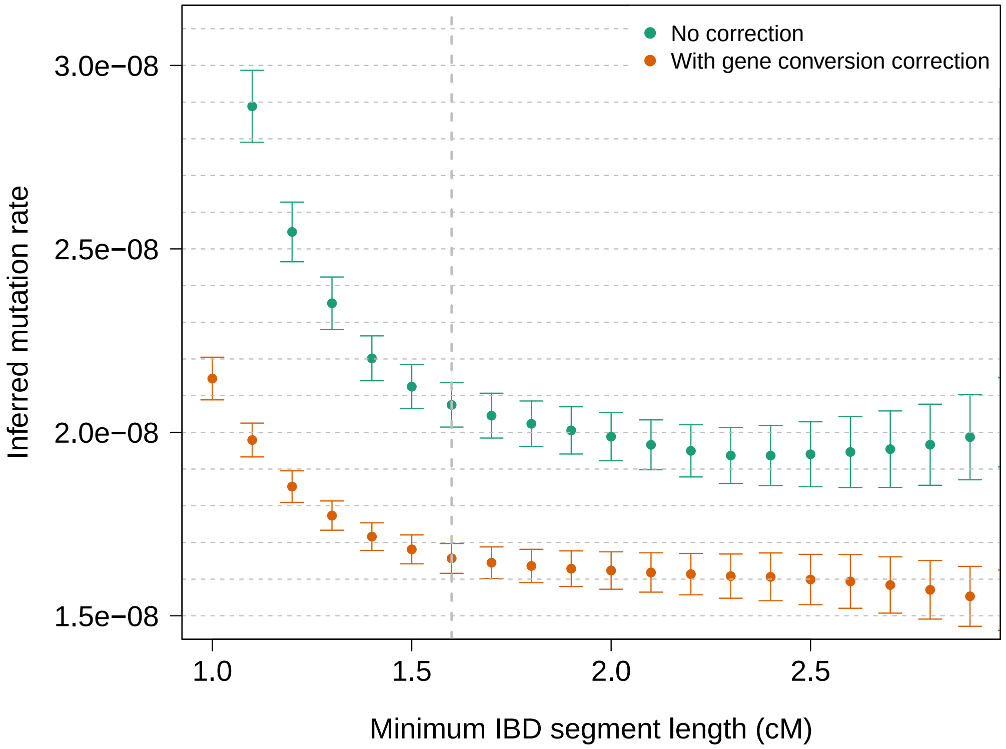

Figure S17: Gene conversion-corrected and uncorrected mutation rates inferred for segments longer than several length thresholds in the GoNL data set.

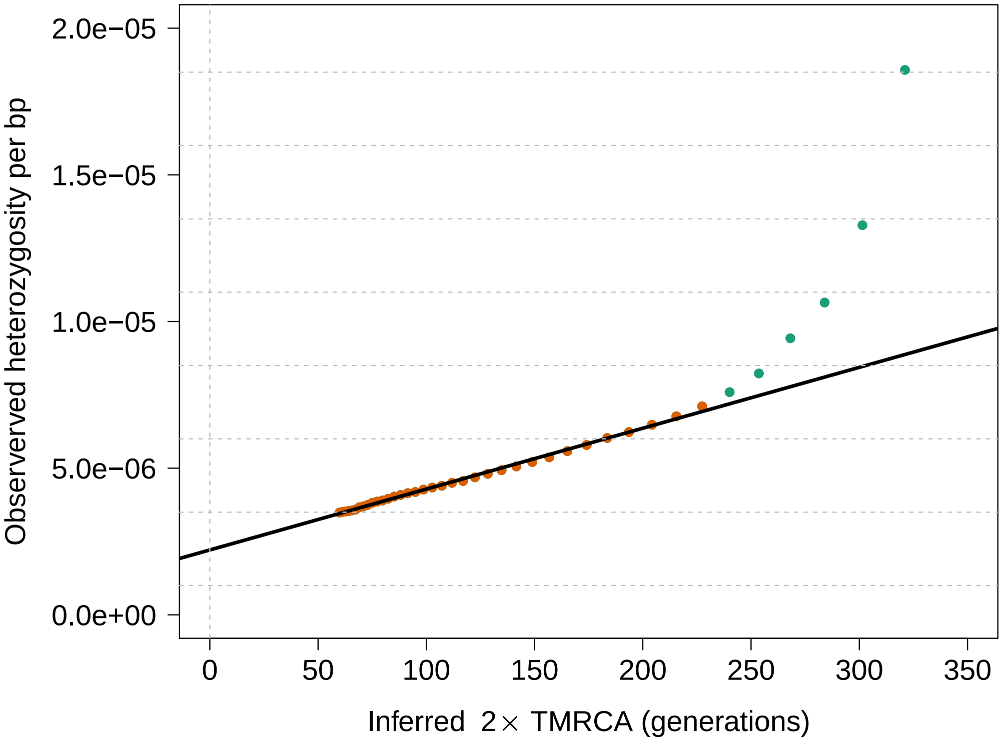

Figure S18: tMRCA regression for segments of length ≥ 1.0 cM in the GoNL data set. The obtained slope is used to estimate mutation rate per generation per base pair, before the effects of gene conversion are accounted for. Segments shorter than 1.6 cM (green) result in mismatching estimates that appear non-linear when compared to segments longer than 1.6 cM. This is likely due to inaccuracies of the underlying demographic model and noisy IBD detection for short segments.

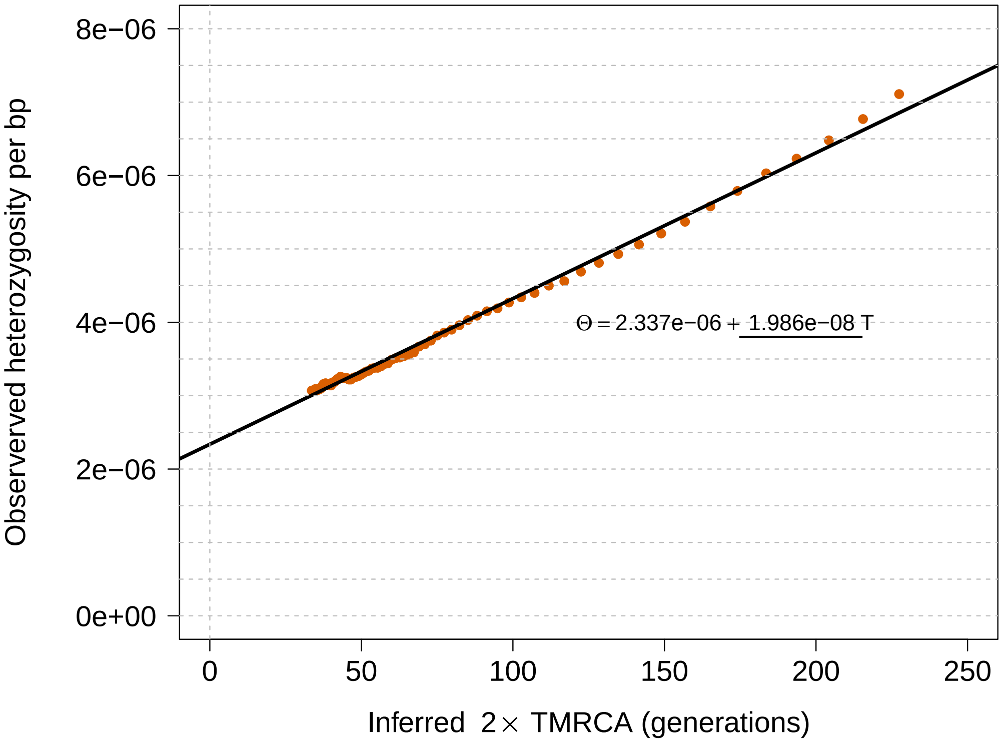

Figure S19: tMRCA regression using segments up to 10 cM.

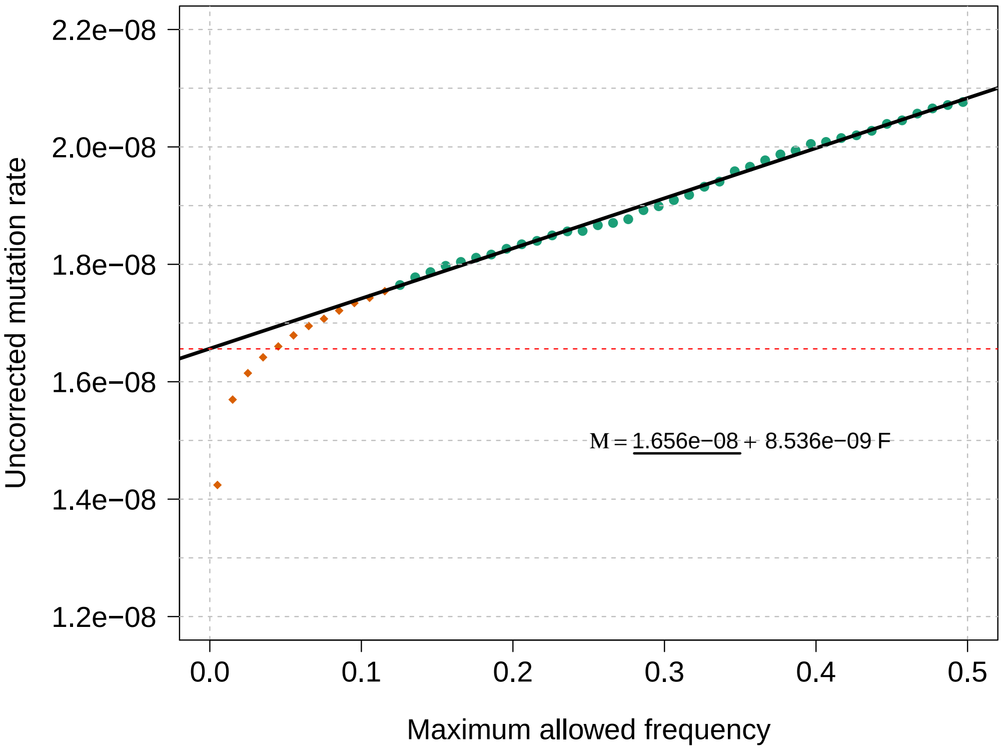

Figure S20: MaAF regression. Red dots show mutation rates for low MaAF values, not used in the regression.

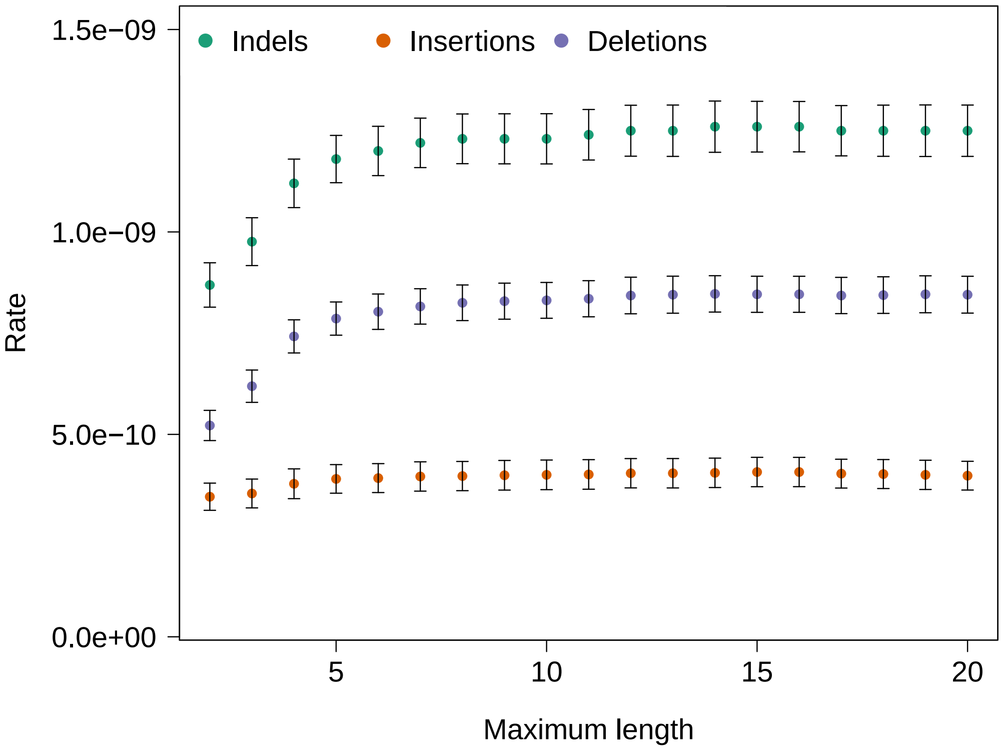

Figure S21: Inferred rate for indels, insertions and deletions, as a function of maximum length.

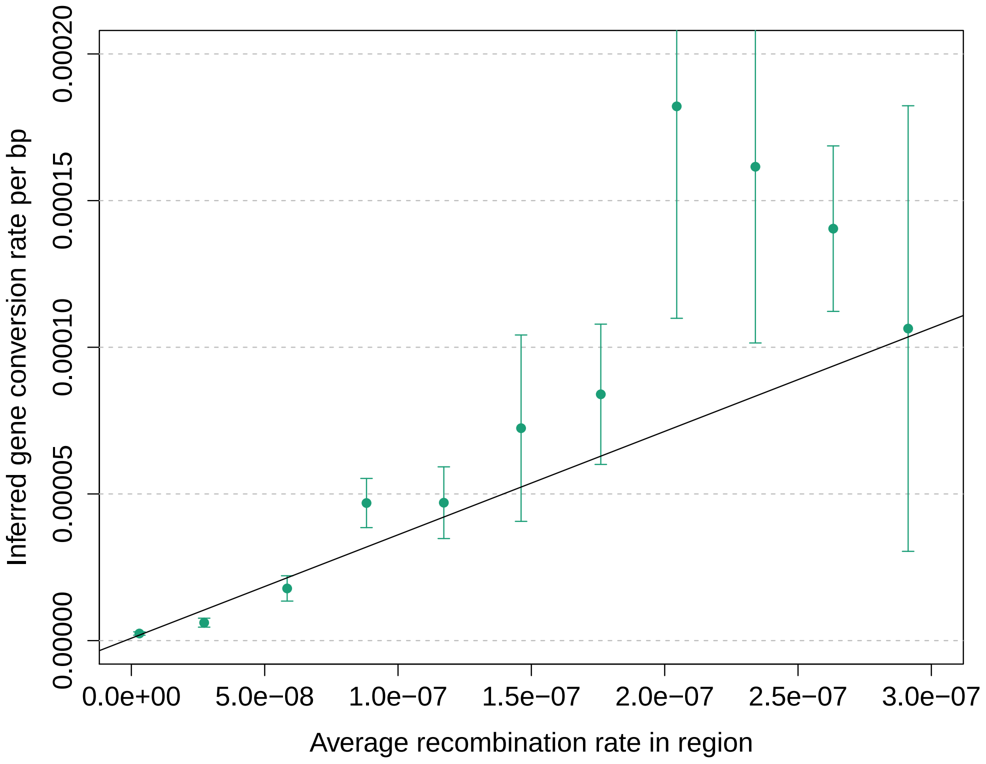

Figure S22: Association between recombination rate and gene conversion rate. We annotated the genome based on uniform bins of recombination rate (per base, per generation), and estimated gene conversion rates for each obtained annotation. We observed association between gene conversion and recombination rate ($R = 0.91$; slope = 353.6, s.e. = 56.5, $p = 1.52 \times 10^{-4}$; intercept $= 8.107 \times 10^{-7}$, s.e. $= 9.208 \times 10^{-7}$, $p = 0.401$).

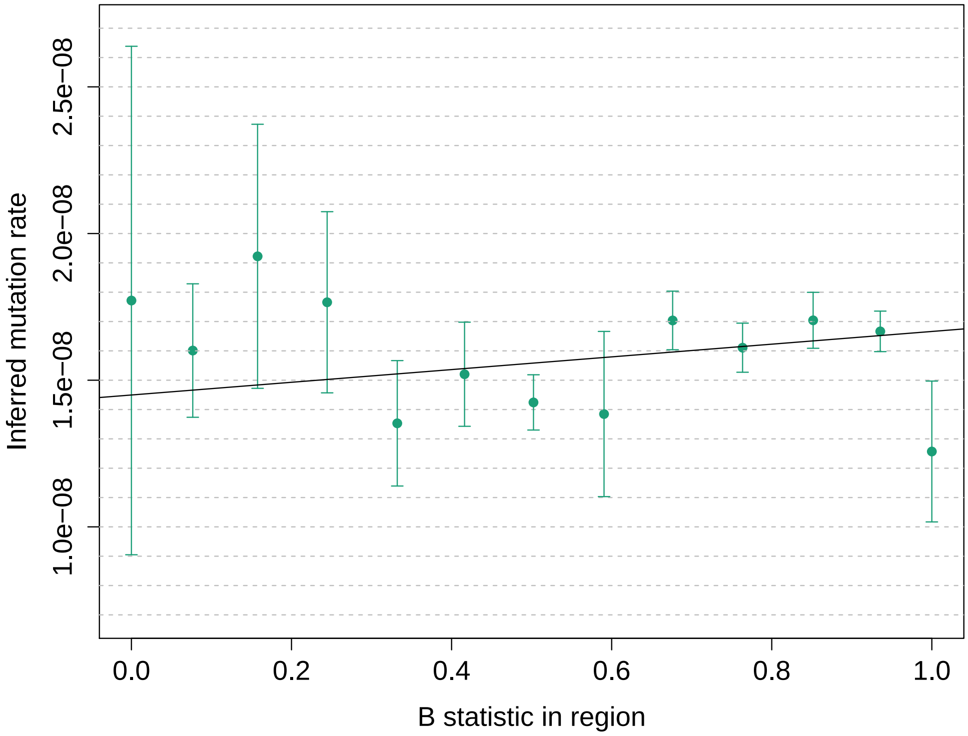

Figure S23: Despite a strong association between average IBD segment length and McVicker B statistic, no significant association is detected between the B statistic and the inferred mutation rates, indicating that the change in local coalescent distributions does not significantly affect the posterior mean IBD segment age used in this analysis.

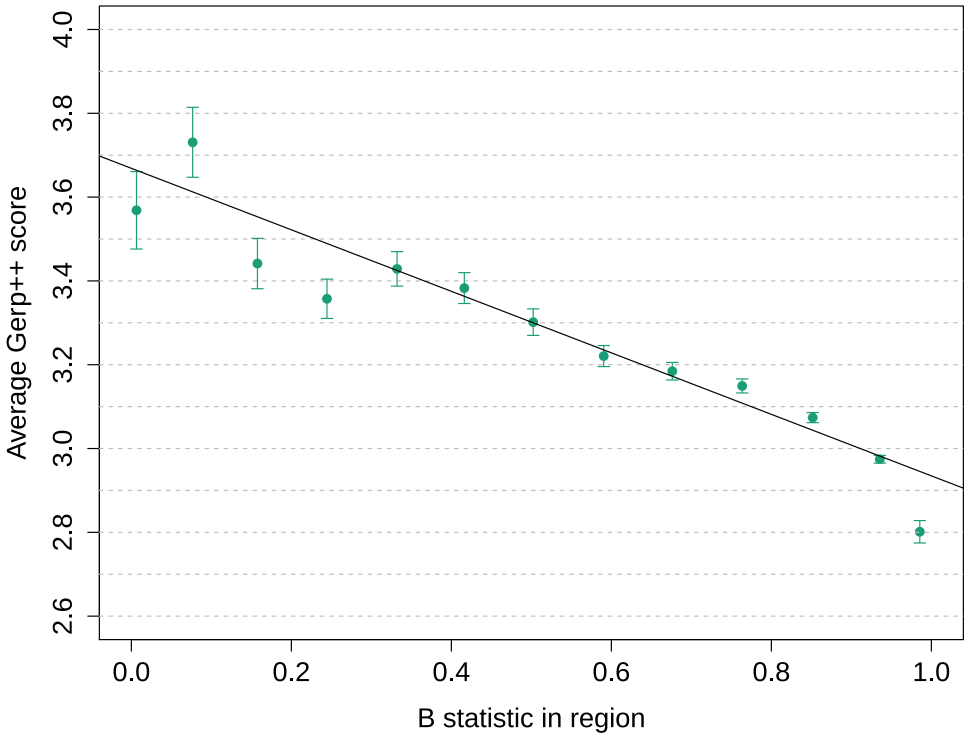

Figure S24: Association between Gerp++ scores of mismatching variants found on IBD segments and average B statistic in IBD regions.

estimator	estimate $\times 10^8$
$\hat{\mu}_o$	$1.981 \pm .172$
$\hat{\mu}_n$	$2.018 \pm .197$
$\hat{\mu}_{o,w}$	$1.985 \pm .178$
$\hat{\mu}_{n,w}$	$2.012 \pm .180$

(a) Estimates in simulation.

estimator	estimate $\times 10^8$
$\hat{\mu}_o$	1.64 ± 0.0396
$\hat{\mu}_n$	1.65 ± 0.0397
$\hat{\mu}_{o,w}$	1.63 ± 0.0441
$\hat{\mu}_{n,w}$	1.67 ± 0.0394
$\hat{\mu}_{o,long}$	1.73 ± 0.1928

(b) Estimates in GoNL.

Table S1: Effects of non-independent observations on tMRCA regression. We performed tMRCA regression using either overlapping (o) or non-overlapping (n) IBD length bins for segments between 1.6 and 10 cM, with intervals of 0.1 cM. (a) We simulated a mutation rate of 2×10^{-8} in a sample of 200 individuals for a chromsome of 100 cM using the GoNL demographic model. We report the inferred average mutation rate and observed standard deviation across 500 independent simulations. We estimated mutation rate using overlapping length bins ($\hat{\mu}_o$), non-overlapping length bins ($\hat{\mu}_n$), overlapping length bins weighted by inverse-variance ($\hat{\mu}_{o,w}$), non-overlapping length bins weighted by inverse-variance ($\hat{\mu}_{n,w}$). We report the mean and standard deviation empirically determined across independent simulations. The overlapping length bins estimator performed as well or better than other estimators. Very small biases were observed, consistent with other analyses. (b) We used the same four estimators in the GoNL data, observing negligible differences. We report the estimate and the standard error determined via block-weighted jackknife. An estimate ($\hat{\mu}_{o,long}$) obtained using overlapping bins and very long segments ($5-10$ cM) was compatible but resulted in large standard error. Inverse-variance weights were inferred using block-weighted jackknife.

estimator	estimate $\times 10^8$
$\hat{\mu}_{o,intercept}$	2.05 ± 0.103
$\hat{\mu}_{o,slope}$	2.05 ± 0.103
$\hat{\mu}_{n,slope}$	2.07 ± 0.103

(a) Estimates in simulations.

estimator	estimate $\times 10^8$
$\hat{\mu}_{o,slope}$	1.64 ± 0.0408
$\hat{\mu}_{n,slope}$	1.65 ± 0.0404

(b) Estimates in GoNL.

Table S2: Effects of non-independent observations on MaAF regression. (a) We performed 500 independent simulations of 200 samples for 100 cM using the GoNL demographic model, a mutation rate of 2×10^{-8} and a gene conversion rate of 6×10^{-6} per generation, per base. IBD was detected using GERMLINE (het=2), as described in Figure 5. The gene conversion corrected mutation rate is inferred using three estimators. The estimator $\hat{\mu}_{o,intercept}$ is obtained from the MaAF regression intercept, as detailed in the main text. The estimator $\hat{\mu}_{o,slope}$ is obtained by first computing the MaAF regression slope, $\hat{\beta}$, and then subtracting $0.5 \times \hat{\beta}$ from the uncorrected mutation rate estimate, which is inflated by gene conversion events. Note that this is closely related to the intercept of the regression (estimator $\hat{\mu}_{o,intercept}$), and has the same performance. Both estimators use overlapping frequency bins, due to the use of maximum allele frequency cutoffs. $\hat{\mu}_{n,slope}$ is obtained the same way, but the MaAF slope is calculated by taking the average of non-overlapping allele frequency cutoffs, where mutation rates are only computed using mismatching sites for which the allele frequency is contained within a frequency range. For all simulations, we used MaAF frequency values from 0.1 to 0.5, with intervals of 0.02. Consistent with Figure 5, a small upward bias is obtained for (het=2). Because the allele frequency spectrum in the simulations reflects recent exponential expansion, the $\hat{\mu}_{o,intercept}$ estimator provides a slightly better correction than $\hat{\mu}_{n,slope}$, although by a minimal amount, as the slope is inferred with more weight on low frequency cutoffs. Note that the demographic model reconstructed using IBD reflects expansion in the recent (~ 100) generations, but neglects demographic events at deeper time scales. The full GoNL spectrum, however, presents small deviations from linearity at intermediate MaAF values, likely due to demographic events (e.g. bottlenecks) at deeper times scales (Figure S2). (b) $\hat{\mu}_{o,slope}$ and $\hat{\mu}_{n,slope}$ estimates for segments between $1.6 - 10$ cM in real data are negligibly different. Real data estimators rely on nested length bins for the tMRCA regression (see Table S1).

Annotation	**Short name**	**Reference**	**Size (Mb)**	**bias**	**Raw μ**	**s.e.**	**Z-score**	**Trinucleotide factor**	**Trinucleotide-corrected μ**	**s.e.**	**Z-score**
Coding regions	Coding	[1]	29	0.99	1.71E-08	1.30E-09	0.42	1.23	1.40E-08	1.06E-09	-2.30
Conserved-unconserved regions	ConservedUnconserved	[2]	1177	1.00	1.66E-08	4.07E-10	0.03	1.00	1.66E-08	4.07E-10	0.05
Conserved regions	Conserved	[2]	68	1.03	1.47E-08	8.41E-10	-1.95	1.04	1.41E-08	8.07E-10	-2.69
Digital Genomic Footprinting assay	DGF-ENCODE	[3, 4]	192	1.00	1.69E-08	7.05E-10	0.44	1.08	1.57E-08	6.52E-10	-1.18
DNAseI hyper sensitivity sites (Maurano)	DHS-Maurano	[5]	556	1.00	1.66E-08	5.27E-10	0.11	1.04	1.60E-08	5.08E-10	-0.82
DNAseI hyper sensitivity sites (Trynka)	DHS-Trynka	[6]	262	1.01	1.74E-08	7.06E-10	0.98	1.08	1.61E-08	6.54E-10	-0.63
DNAseI hyper sensitivity sites, peaks	DHS-peaks	[7]	175	1.00	1.91E-08	1.51E-09	1.63	1.11	1.73E-08	1.37E-09	0.50
DNAseI hyper sensitivity sites, Promoter	DHSPromoter	[5]	37	1.00	1.81E-08	1.19E-09	1.21	1.06	1.71E-08	1.12E-09	0.41
Enhancers (Andersson)	Enhancer-And	[7]	6	1.00	1.68E-08	4.45E-09	0.05	1.12	1.50E-08	3.97E-09	-0.40
Enhancers (Hoffman)	Enhancer-Hoff	[8]	87	1.00	1.63E-08	1.27E-09	-0.22	1.10	1.48E-08	1.16E-09	-1.45
Fetal DNAseI hyper sensitivity sites	fetal-DHS	[6]	135	1.00	1.71E-08	8.42E-10	0.54	1.09	1.57E-08	7.72E-10	-1.04
Histone modification H3K27ac-Hnisz	H3K27ac-Hnisz	[9]	493	1.00	1.66E-08	4.42E-10	0.12	1.04	1.60E-08	4.26E-10	-0.89
Histone modification H3K27ac-PGC2	H3K27ac-PGC2	[10]	356	1.00	1.60E-08	5.00E-10	-0.86	1.05	1.53E-08	4.79E-10	-1.98
Histone modification H3K4me1, peaks	H3K4me1-peaks	[6]	250	1.01	1.75E-08	7.50E-10	1.13	1.08	1.63E-08	6.96E-10	-0.37
Histone modification H3K4me1	H3K4me1	[6]	592	1.00	1.65E-08	4.20E-10	-0.06	1.04	1.59E-08	4.05E-10	-1.07
Histone modification H3K4me3, peaks	H3K4me3-peaks	[6]	57	1.00	1.67E-08	1.76E-09	0.10	1.17	1.43E-08	1.50E-09	-1.46
Histone modification H3K4me3	H3K4me3	[6]	180	1.00	1.61E-08	6.39E-10	-0.58	1.10	1.47E-08	5.81E-10	-2.68
Histone modification H3K9ac, peaks	H3K9ac-peaks	[6]	55	1.01	1.81E-08	1.84E-09	0.84	1.17	1.55E-08	1.57E-09	-0.65
Histone modification H3K9ac	H3K9ac	[6]	176	1.00	1.62E-08	5.82E-10	-0.54	1.11	1.46E-08	5.25E-10	-2.97
Intron	Intron	[1]	513	0.99	1.64E-08	6.60E-10	-0.18	1.00	1.64E-08	6.57E-10	-0.27
Late replication	LateReplication	[11]	14	0.98	2.19E-08	3.75E-09	1.43	1.13	1.95E-08	3.33E-09	0.87
Large intergenic non-coding RNAs	lincRNAs-transcripts	[12]	55	0.96	1.73E-08	2.10E-09	0.34	0.99	1.75E-08	2.12E-09	0.42
Neanderthal-depleted in Europeans	NeanderthalDepleted	[13]	21	1.04	1.37E-08	3.72E-09	-0.77	0.95	1.43E-08	3.90E-09	-0.57
Neanderthal-enriched in Europeans	NeanderthalEnriched	[13]	1181	1.00	1.66E-08	4.07E-10	0.06	1.00	1.66E-08	4.07E-10	0.06
Promoter	Promoter	[1]	38	0.98	1.57E-08	2.04E-09	-0.40	1.15	1.37E-08	1.78E-09	-1.57
Constrained genes	ConstrainedGenes	[14]	1	0.89	1.06E-08	1.80E-08	-0.33	1.20	8.81E-09	1.50E-08	-0.52
Segway-chromHMM CTCF Binding Site	segment.CTCF	[8]	28	1.00	1.51E-08	1.80E-09	-0.78	1.09	1.39E-08	1.66E-09	-1.56
Segway-chromHMM enhancer	segment.E	[8]	58	1.00	1.34E-08	1.46E-09	-2.09	1.11	1.21E-08	1.32E-09	-3.27
Segway-chromHMM promoter flanking	segment.PF	[8]	12	1.01	1.49E-08	2.22E-09	-0.71	1.06	1.41E-08	2.10E-09	-1.15
Segway-chromHMM repressed/inactive region	segment.R	[8]	532	1.00	1.61E-08	4.64E-10	-0.80	0.97	1.66E-08	4.79E-10	0.08
Segway-chromHMM transcribed region	segment.T	[8]	424	1.00	1.70E-08	5.74E-10	0.59	1.01	1.68E-08	5.69E-10	0.39

Segway-chromHMM transcription start site	segment.TSS	[8]	21	0.99	9.50E-09	4.15E-09	-1.70	1.33	7.15E-09	3.12E-09	-2.99
Segway-chromHMM weak enhancer	segment.WE	[8]	29	1.00	2.01E-08	2.64E-09	1.33	1.10	1.83E-08	2.40E-09	0.72
Transcription factor binding sites	TFBS	[3]	179	1.00	1.67E-08	7.65E-10	0.16	1.08	1.54E-08	7.06E-10	-1.41
Untranslated regions 3’	UTR-3	[1]	18	0.99	1.32E-08	1.88E-09	-1.74	1.08	1.22E-08	1.73E-09	-2.47
Untranslated regions 5’	UTR-5	[1]	7	1.00	1.69E-08	3.41E-09	0.10	1.26	1.34E-08	2.70E-09	-1.16
Untranslated regions	UTR	[1]	17	0.99	1.55E-08	1.85E-09	-0.55	1.13	1.38E-08	1.64E-09	-1.65

Table S3: List of annotations, mutation rates and bias/trinucleotide factors used to correct estimates. Trinucleotide factors were computed to control for trinucleotide substitution rate heterogeneity [15, 16]. When analyzing mutation rates within different genomic regions, we computed annotation-specific correction factors to account for the differences in mutation rates that are expected as a result of trinucleotide context variation. We used the trinucleotide context-specific mutation-rate matrix of Kryukov [16]. We denote the substitution rate of trinucleotides of the form XYZ as $\phi_{XYZ} = \sum_{V \in \{A,C,G,T\} | V \neq Y} \phi_{XYZV}$, where ϕ_{XYZV} is the substitution rate of $XYZ \rightarrow XVZ$ and $\{A, C, G, T\}$ represent the four possible bases. We then use the Human Genome h19 consensus sequence from the UCSC Genome Browser to determine the trinucleotide context of the considered annotations. Denoting the fraction of trinucleotides XYZ contained in annotation α as f_α^{XYZ}, we compute a correction factor $\lambda_\alpha = \left(\sum_{XYZ \in \Gamma} f_\alpha^{XYZ} \phi_{XYZ}\right) / \left(\sum_{XYZ \in \Gamma} f_{GW}^{XYZ} \phi_{XYZ}\right)$, where GW denotes the genome-wide annotation and Γ is the set of 64 possible trinucleotide combinations. We then scaled the obtained local mutation rate by $1/\lambda_\alpha$ to obtain a context-corrected estimate of the mutation rate. To correct for the small-annotation bias (see Figure S12) reported in the table, permutations were computed until a standard error smaller than 10^{-10} was obtained for all annotations. We then scaled the annotation-specific mutation rate by the inverse of the computed bias to correct the estimate. 95% confidence intervals for genome-wide and annotation specific rates were computed based on standard errors estimated using weighted block jackknife, using the 26 independent chromosomal regions obtained as previously described. For almost all considered annotations, the computed bias was found to be extremely small.

$10^8\mu_m$	$10^8\mu_f$	$10^8\rho_m$	$10^8\rho_f$	N_m	N_f	$10^8\hat{\mu}_a$	$10^8\hat{\alpha}$	f_m
2	2	1.5	1.5	10	4990	1.989 ± 0.004	1.657 ± 0.13	0.553
2	2	1.5	1.5	100	4900	1.993 ± 0.003	1.790 ± 0.15	0.536
2	2	1.5	1.5	1500	3500	2.001 ± 0.003	1.538 ± 0.16	0.514
2	2	1.5	1.5	2500	2500	2.001 ± 0.003	1.389 ± 0.16	0.500
3	1	1.5	1.5	10	4990	1.960 ± 0.004	3.798 ± 0.13	0.553
3	1	1.5	1.5	100	4900	1.999 ± 0.003	3.379 ± 0.16	0.536
3	1	1.5	1.5	1500	3500	2.000 ± 0.003	2.297 ± 0.17	0.514
3	1	1.5	1.5	2500	2500	1.998 ± 0.003	1.513 ± 0.16	0.500
3	1	1	2	10	4990	1.946 ± 0.004	4.519 ± 0.13	0.554
3	1	1	2	100	4900	1.991 ± 0.003	4.236 ± 0.16	0.537
3	1	1	2	1500	3500	2.000 ± 0.003	2.812 ± 0.16	0.516
3	1	1	2	2500	2500	2.002 ± 0.003	1.952 ± 0.16	0.500

Table S4: Effects of sex-averaging on inferred rates. We simulated IBD segments from a population composed of N_m males and N_f females, which have mutation and recombination rates μ_m, μ_f and ρ_m, ρ_f, respectively. The simulated differences in male/female mutation and recombination rates are similar to those of Table S9 and [17]. Given two randomly chosen individuals from the population, the simulation iteratively samples ancestral lineages from generation t to generation $t+1$ in the past. Each ancestor is sampled male or female with probability 1/2. At each generation, and for both lineages, the closest recombination event on either side of the site is sampled from a geometric distribution using the sex-specific recombination rate, and the distance to the first recombination event in either direction is stored. The physical length of IBD segments is then used to obtain a length in units of sex-averaged recombination. The sampling proceeds until either a MRCA is found, or the IBD segment becomes smaller than the detectable threshold. The number of mutations on IBD segments is determined by sampling a Poisson distribution with rate $\mu = T_m\mu_m + T_f\mu_f$, where T_m is the number of meioses occurring in males. tMRCA regression is then performed using sex-averaged genetic lengths and observed mutation rates on the sampled segments, as described in the Methods section. In this model, coalescence occurs if both individual select the same ancestor, at rate $\frac{1}{4} \times \frac{1}{N_f} + \frac{1}{4} \times \frac{1}{N_m} + \frac{1}{2} \times 0 = \frac{N_f+N_m}{4N_fN_m}$, implying an effective population size of $N_e = \frac{4N_fN_m}{N_f+N_m}$, [18], which we use to compute the posterior mean tMRCA estimate. We report the mean and standard error for the inferred mutation rates, $\hat{\mu}_a$, the tMRCA regression intercept $\hat{\alpha}$, and the fraction f_m of meiotic events occurring in males in the ancestral lineages of segments longer than 1.6 cM. We omit the s.e. for the latter, which was $\sim 10^{-4}$ for all entries. 300 independent simulations were run, each sampling $50,000$ IBD segments. A small but significant difference between the flat average of sex-specific mutation rates and the tMRCA slope is observed only for very extreme differences between male and female effective population sizes ($N_m/(N_m + N_f) = 0.002$). The tMRCA intercept increases with larger mutation rate and effective population size differences.

chromosome	from bp	to bp	estimate $(\times 10^8)$
1	66,874,699	118,837,888	1.53
2	17,246,473	85,384,179	1.95
2	193,010,478	235,351,139	1.80
3	678,347	176,030,190	1.62
4	85,315,581	189,657,996	1.43
5	22,657,926	141,420,437	1.60
6	33,954,192	103,983,460	1.62
6	139,903,959	170,245,872	1.89
7	962,247	38,722,532	1.85
7	41,688,961	152,254,508	1.79
8	55,170,178	139,553,601	1.54
9	72,512,292	132,515,730	1.30
10	19,570,732	134,866,854	2.00
11	2,047,054	134,587,122	1.53
12	6,476,123	75,656,510	1.57
12	82,586,486	128,401,829	1.80
13	20,518,406	114,094,544	1.51
14	20,545,390	59,184,876	1.29
14	63,846,103	104,808,535	1.63
15	50,284,344	101,969,749	1.73
17	163,278	55,936,970	1.89
18	11,962,813	59,189,703	1.21
19	7,857,579	58,513,172	1.70
20	5,649,902	52,818,462	1.52
21	15,636,220	47,031,048	1.93
22	23,874,416	50,493,062	1.72

Table S5: Region-specific estimates of mutation rate (mean: 1.65×10^{-8}, s.e.: 0.04×10^{-8}).

chromosome	from bp	to bp	estimate ($\times 10^8$)
1	66,874,699	88,238,750	1.61
1	88,238,751	108,526,486	9.91
2	17,246,473	34,280,051	1.30
2	34,280,052	49,009,386	2.11
2	49,009,387	69,293,237	1.44
2	193,010,478	216,555,564	2.01
2	216,555,565	230,068,380	1.60
3	678,347	7,867,058	1.05
3	7,867,059	21,680,325	1.80
3	21,680,326	36,948,001	1.55
3	36,948,002	61,394,898	1.63
3	61,394,899	73,519,262	1.42
3	73,519,263	109,288,895	1.94
3	109,288,896	127,471,868	1.85
3	127,471,869	147,679,411	2.62
3	147,679,412	171,161,266	1.53
4	85,315,581	109,663,976	1.30
4	109,663,977	132,801,458	1.39
4	132,801,459	153,995,617	1.90
4	153,995,618	171,817,565	1.41
4	171,817,566	183,599,323	1.36
5	22,657,926	37,949,446	1.62
5	37,949,447	67,185,960	1.58
5	67,185,961	82,957,503	1.77
5	82,957,504	110,480,596	1.50
5	110,480,597	128,743,448	2.04
6	33,954,192	48,250,743	1.50
6	48,250,744	84,668,623	1.51
6	139,903,959	155,635,584	1.04
6	155,635,585	166,874,299	2.49
7	962,247	11,388,991	1.19
7	11,388,992	23,827,910	1.95
7	23,827,911	37,498,171	1.61
7	41,688,961	68,729,788	1.96
7	68,729,789	89,724,984	1.72
7	89,724,985	109,644,709	1.29
7	109,644,710	135,508,955	1.49
7	135,508,956	149,826,715	1.80

8	55,170,178	73,892,270	1.83
8	73,892,271	99,400,617	9.96
8	99,400,618	122,503,061	1.65
8	122,503,062	134,271,328	2.27
9	72,512,292	87,943,421	1.34
9	87,943,422	106,603,815	1.36
9	106,603,816	120,062,948	1.20
10	19,570,732	35,924,606	2.22
10	35,924,607	61,715,654	2.38
10	61,715,655	79,857,311	1.52
10	79,857,312	97,321,680	1.97
10	97,321,681	117,955,613	2.11
10	117,955,614	128,006,669	1.45
11	20,470,54	12,359,828	1.65
11	12,359,829	25,940,249	2.02
11	25,940,250	44,965,371	1.65
11	44,965,372	76,910,242	1.51
11	76,910,243	96,579,605	1.50
11	96,579,606	116,325,155	1.44
11	116,325,156	127,550,767	1.55
12	6,476,123	20,195,998	1.17
12	20,195,999	42,284,690	1.65
12	42,284,691	63,497,675	1.80
12	82,586,486	101,536,560	2.27
12	101,536,561	116,921,218	1.90
13	20,518,406	28,691,009	1.09
13	28,691,010	40,724,913	1.84
13	40,724,914	62,072,103	1.50
13	62,072,104	82,940,941	1.43
13	82,940,942	102,214,268	1.95
13	102,214,269	110,883,495	5.53
14	20,545,390	29,913,958	1.31
14	29,913,959	47,564,047	1.21
14	63,846,103	83,501,046	1.07
14	83,501,047	96,262,155	2.12
15	50,284,344	66,967,905	1.70
15	66,967,906	86,564,188	1.82
15	86,564,189	94,855,437	1.79
17	163,278	8,583,495	1.07
17	8,583,496	15,014,380	1.92

17	15,014,381	35,509,268	2.49
17	35,509,269	54,833,347	2.20
18	11,962,813	35,726,545	9.99
18	35,726,546	55,512,688	1.28
19	7,857,579	19,249,992	1.64
19	19,249,993	41,845,871	1.59
19	41,845,872	52,143,902	1.55
20	5,649,902	16,025,762	1.76
20	16,025,763	39,217,325	1.48
20	39,217,326	50,714,875	1.28
21	15,636,220	25,900,943	2.36
21	25,900,944	38,711,179	1.81
21	38,711,180	46,359,224	2.12
22	23,874,416	35,756,706	1.54
22	35,756,707	46,950,433	1.98

Table S6: Estimates of mutation rate for regions of ~ 20 cM (mean: 1.64×10^{-8}, s.e.: 0.04×10^{-8}).

Type	Mutation rate
Transition at non-CpG	$9.28 \pm 0.27 \times 10^{-9}$
Transition at CpG	$1.68 \pm 0.12 \times 10^{-7}$
Transversion at non-CpG	$4.93 \pm 0.15 \times 10^{-9}$
Transversion at CpG	$1.30 \pm 0.13 \times 10^{-8}$

Table S7: Mutation rates for CpG/non-CpG transitions/transversions.

Perturbation of demographic parameter	Effect on mutation rate estimate
Ancestral size decreased by 50%	−10.7%
Ancestral size decreased by 30%	−5.9%
Ancestral size decreased by 10%	−1.8%
Ancestral size increased by 10%	+1.7%
Ancestral size increased by 30%	+4.9%
Ancestral size increased by 50%	+7.9%
Current size changed by 10%	less than 0.01% difference
Current size divided by 100	−0.4%

Table S8: Effects of changes in the reconstructed demographic model on the estimated mutation rate in GoNL.

β_y	G	$10^8\hat{\mu}_{f,g}$	$10^8\hat{\mu}_{m,g}$	$10^8\hat{\mu}_{a,g\to 28}$	$10^8\hat{\mu}_{a,g\to 30}$	$10^8\hat{\mu}_{a,g\to 32}$
$1.0/(2.681 \times 10^9)$	28	1.09	2.22	**1.66**	1.69	1.73
	30	1.06	2.25	1.62	**1.66**	1.69
	32	1.03	2.28	1.58	1.62	**1.66**
	36	0.97	2.34	1.51	1.54	1.58
$2.0/(2.681 \times 10^9)$	28	0.87	2.44	**1.66**	1.73	1.81
	30	0.81	2.50	1.58	**1.66**	1.73
	32	0.75	2.56	1.51	1.58	**1.66**
	36	0.63	2.68	1.36	1.43	1.51
$3.0/(2.681 \times 10^9)$	28	0.65	2.66	**1.66**	1.77	1.88
	30	0.56	2.75	1.54	**1.66**	1.77
	32	0.47	2.84	1.43	1.54	**1.66**
	36	0.29	3.02	1.21	1.32	1.43

Table S9: Effects of historical paternal age. We express the sex-averaged per generation, per base mutation rate as $\mu_{a,g} = \frac{1}{2}(\mu_{m,g} + \mu_{f,g})$, where $\mu_{m,g}$ and $\mu_{f,g}$ are the per generation male and female mutation rates, respectively. We assume the linear model $\mu_{m,g} = C\mu_{f,g} + \beta_y(G - P)$ for the paternal mutation rate [19], where β_y represents the per year, per base paternal age effect on mutation rate, G represents the father's age at reproduction, $P = 13$ represents puberty onset [20], and $C = 35/23$ is a scaling constant to account for the different number of cell divisions in males and females at birth [21]. Using this model, $\mu_{a,g} = \frac{1}{2}\left[\mu_{f,g}(1 + C) + \beta_y(G - P)\right]$. Given our estimate of historical sex-averaged mutation rate $\hat{\mu}_{a,g} = 1.66 \times 10^{-8}$, and an estimate of the per year paternal age effect β_y, we compute the maternal and paternal contributions to the sex-averaged rate as $\hat{\mu}_{f,g} = \frac{1}{1+C}\left[2\hat{\mu}_{a,g} - \beta_y(G - P)\right]$ and $\hat{\mu}_{m,g} = C\hat{\mu}_{f,g} + \beta_y(G{-}P)$. For β_y, [22] reported an effect of ~ 2 mutations for a haploid genome of $\sim 2.681 \times 10^9$. We report results for values in $\{1.0, 2.0, 3.0\}/(2.681 \times 10^9)$. We then compute a projected sex-averaged mutation rate which assumes a reproductive paternal age different from the historical average for which $\hat{\mu}_{a,g}$ was measured. To this end, we use the same linear model, $\hat{\mu}_{m,g} = C\hat{\mu}_{f,g} + \beta_y(G - P)$, but using $G \in \{28, 30, 32\}$.

References

[1] Ucsc genome browser. http:\genome.ucsc.edu.

[2] Ward, L. D. and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science *337*, 1675–1678.

[3] Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. Nature *489*, 57–74.

[4] Gusev, A., Lee, S. H., Neale, B. M., Trynka, G., Vilhjalmsson, B. J., Finucane, H., Xu, H., Zang, C., Ripke, S., Stahl, E., et al. (2014). Regulatory variants explain much more heritability than coding variants across 11 common diseases. bioRxiv pp. 004309.

[5] Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory dna. Science *337*, 1190–1195.

[6] Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B. E., Han, B., and Raychaudhuri, S. (2014). Disentangling effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex trait loci. bioRxiv pp. 009258.

[7] Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

[8] Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., et al. (2012). Integrative annotation of chromatin elements from encode data. Nucleic acids research pp. gks1284.

[9] Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., and Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

[10] of the Psychiatric Genomics Consortium, S. W. G. et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.

[11] Koren, A., Polak, P., Nemesh, J., Michaelson, J. J., Sebat, J., Sunyaev, S. R., and McCarroll, S. A. (2012). Differential relationship of dna replication timing to different forms of human mutation and variation. The American Journal of Human Genetics *91*, 1033–1040.

[12] Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. Genes & development *25*, 1915–1927.

[13] Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of neanderthal ancestry in present-day humans. Nature *507*, 354–357.

[14] Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nature Genetics *46*, 944–950.

[15] Krawczak, M., Ball, E. V., and Cooper, D. N. (1998). Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. The American Journal of Human Genetics *63*, 474–488.

[16] Kryukov, G. V., Pennacchio, L. A., and Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. The American Journal of Human Genetics *80*, 727–739.

[17] Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–1103.

[18] Wright, S. (1931). Evolution in mendelian populations. Genetics *16*, 97.

[19] Ségurel, L., Wyman, M. J., and Przeworski, M. (2014). Determinants of mutation rate variation in the human germline. Annual Review of Genomics and Human Genetics *15*, 47–70.

[20] Nielsen, C. T., SKAKKEBAeK, N. E., Richardson, D. W., Darling, J. A. B., Hunter, W. M., Jorgensen, M., Nielsen, A., Ingerslev, O., Keiding,

N., and Muller, J. (1986). Onset of the Release of Spermatozia (Supermarche) in Boys in Relation to Age, Testicular Growth, Pubic Hair, and Height. J Clin Endocrinol Metab *62*, 532–535.

[21] Crow, J. F. (2000). The origins, patterns and implications of human spontaneous mutation. Nature Reviews Genetics *1*, 40–47.

[22] Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of fathers age to disease risk. Nature *488*, 471–475.