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

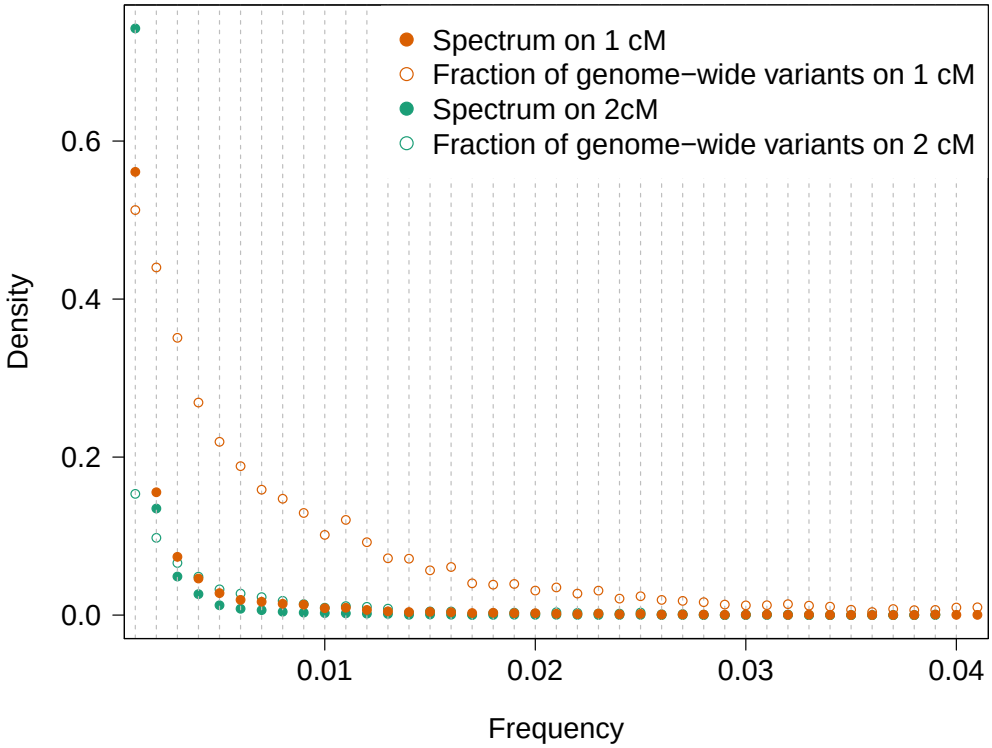

Figure S1: Simulated frequency spectra on IBD segments of different legnths. We computed the allele frequency spectrum of mismatching sites due to new mutation events occurring on IBD segments. Empty dots represent the fraction of the total genome-wide variants of a specific frequency that are found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

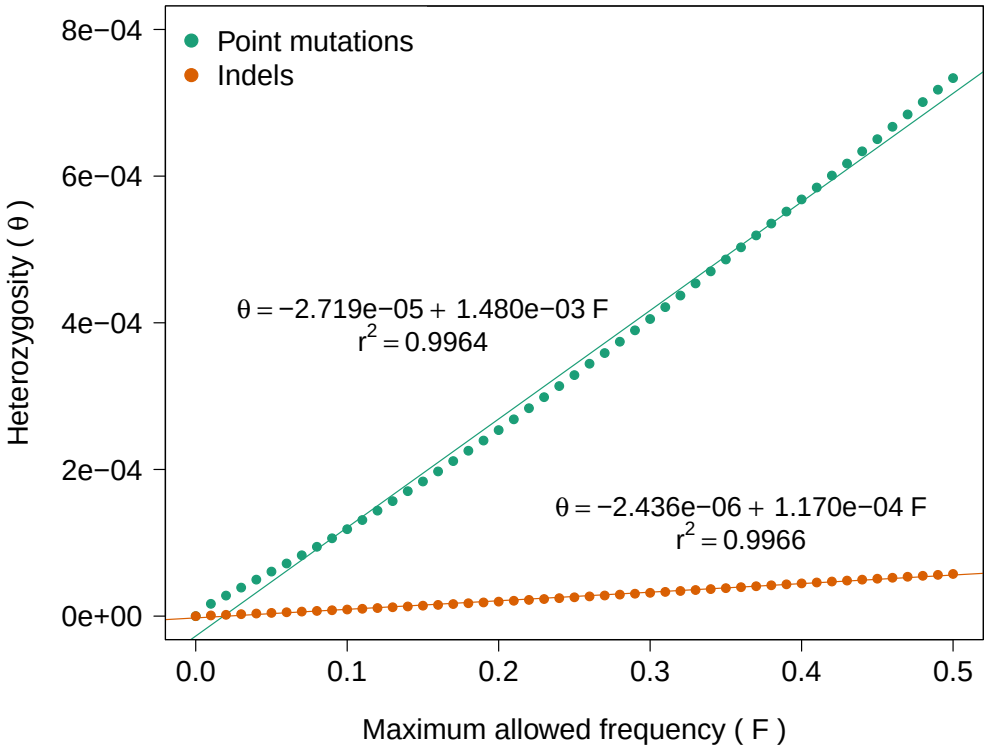

Figure S2: Approximately uniform contribution of variants of different frequencies to overall heterozygosity for both point mutations and indels in the GoNL dataset. Small deviations from linearity may be caused by demographic history (at both recent and remote time scales).

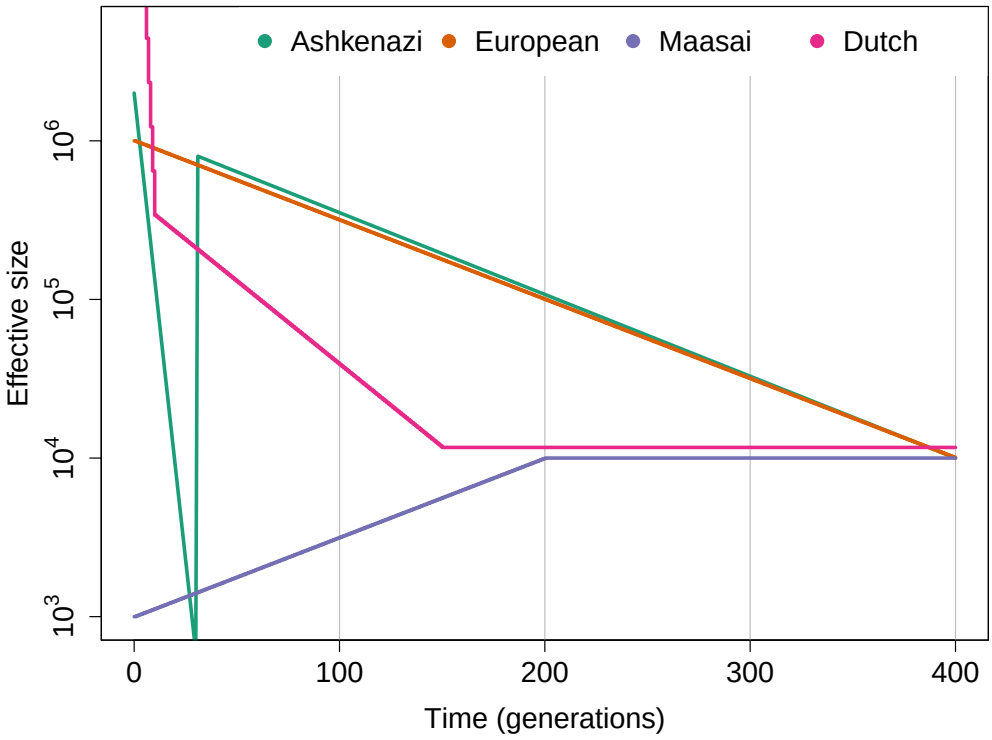

Figure S3: Demographic models inferred for the GoNL data or adopted in simulations.

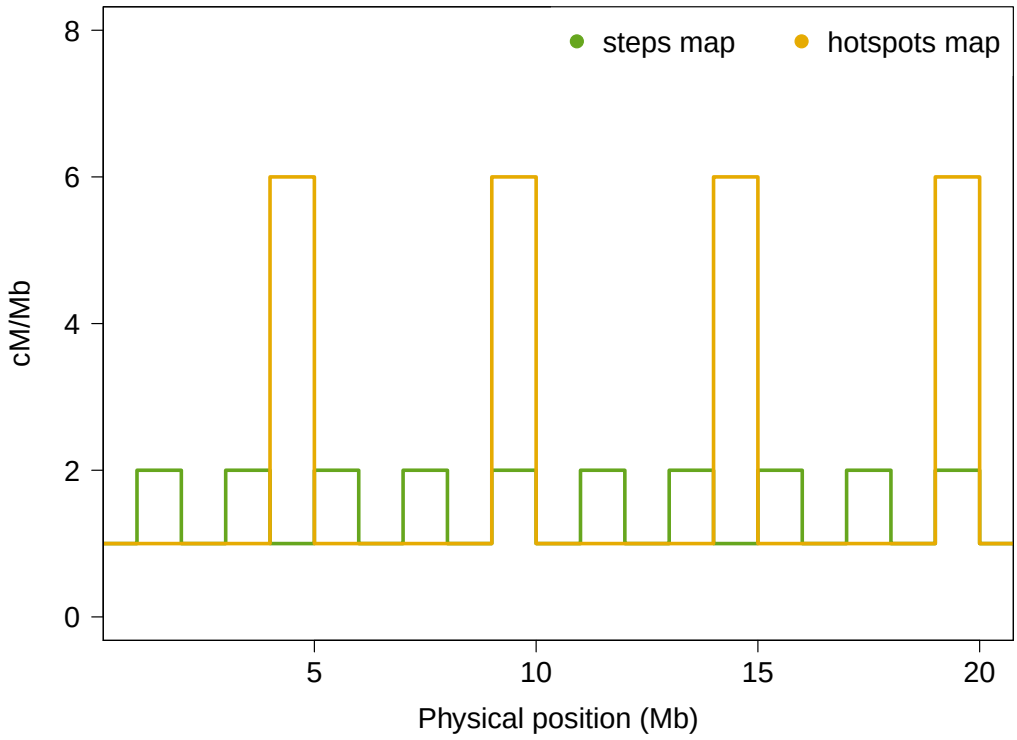

Figure S4: Genetic maps adopted in simulations.

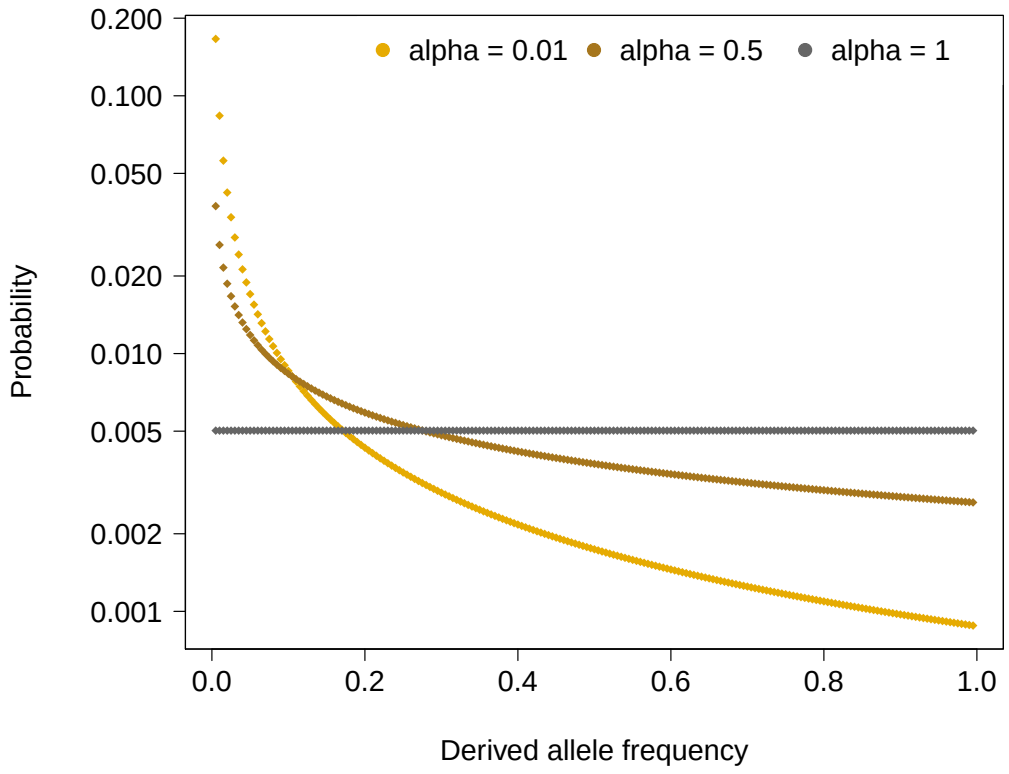

Figure S5: Distributions adopted to sample the frequency of spurious genotyping calls in simulated data. The beta distribution $Beta(\alpha, \beta)$ was used with $\beta = 1$ and α as specified in the Legend. For "de-novo" false positive errors, the frequency determines the number of individuals that 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.

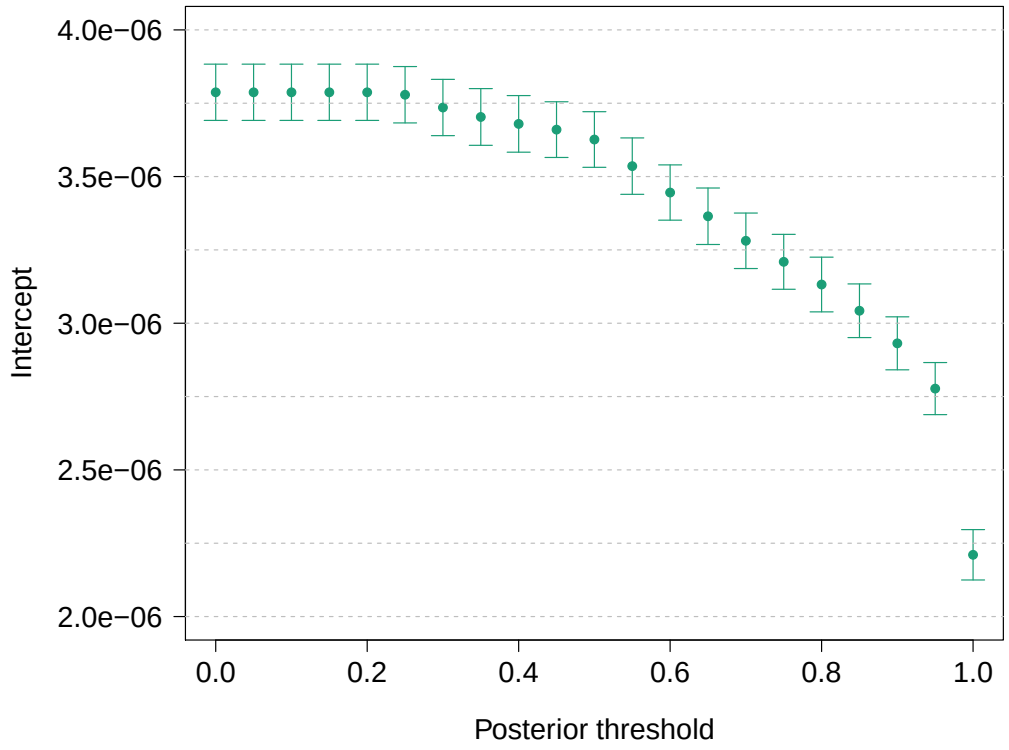

Figure S6: Estimated intercept of the tMRCA regression in the GoNL dataset for segments of length at least 1.6 cM, as a function of the minimum required MVNCall posterior quality for observed heterozygous sites. The MVNCall program used to trio-phase the analyzed data outputs posterior probabilities 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.

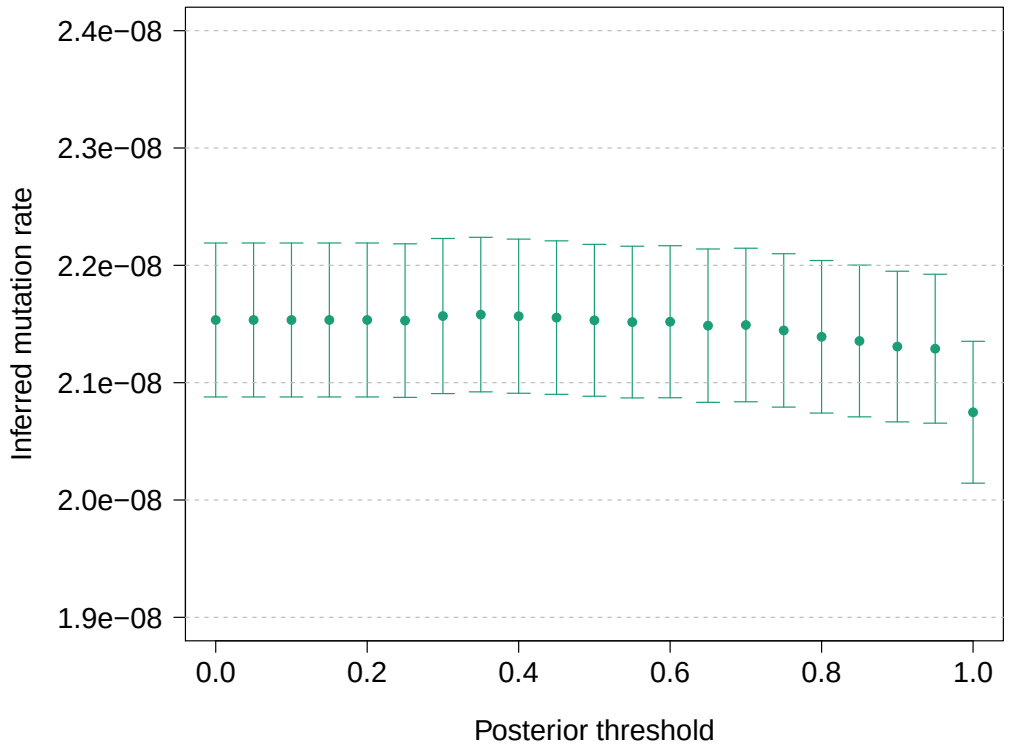

Figure S7: Estimated slope of the tMRCA regression in the GoNL dataset for segments of length at least 1.6 cM, as a function of the minimum required MVNCall posterior quality for observed heterozygous sites. Different values of the MVNCall posterior threshold, resulting in higher estimated genotyping 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.

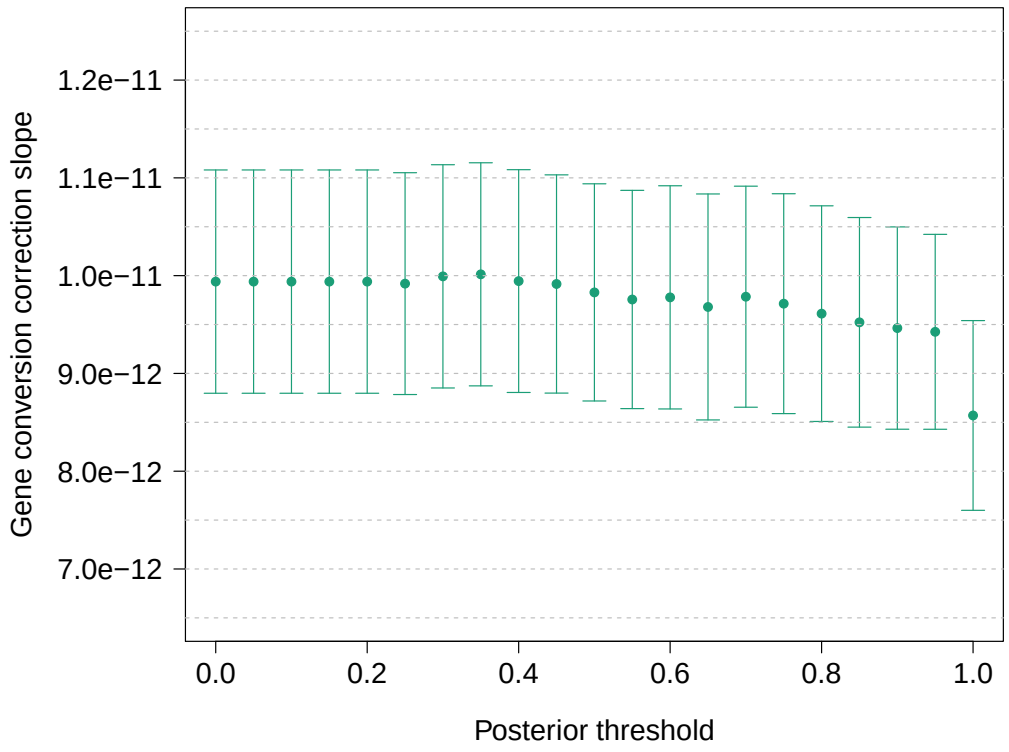

Figure S8: Estimated slope of the MaAF-threshold regression performed to correct for gene conversion in the GoNL dataset for segments of length at least 1.6 cM, as a function of the minimum required MVNCall posterior quality for observed heterozygous sites. Minimal variation is observed as the MVNCall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

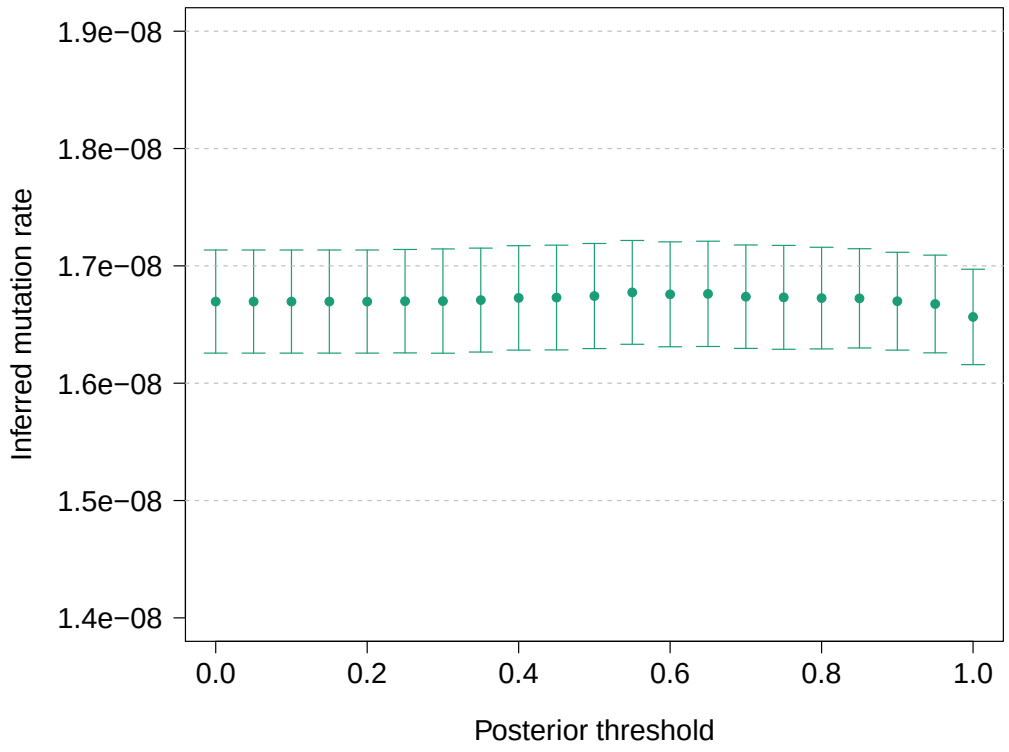

Figure S9: Estimated intercept of the MaAF-threshold regression in the GoNL dataset for segments of length at least 1.6 cM, as a function of the minimum required MVNCall posterior quality for observed heterozygous sites. We observed no significant impact of the chosen MVNCall posterior threshold on the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

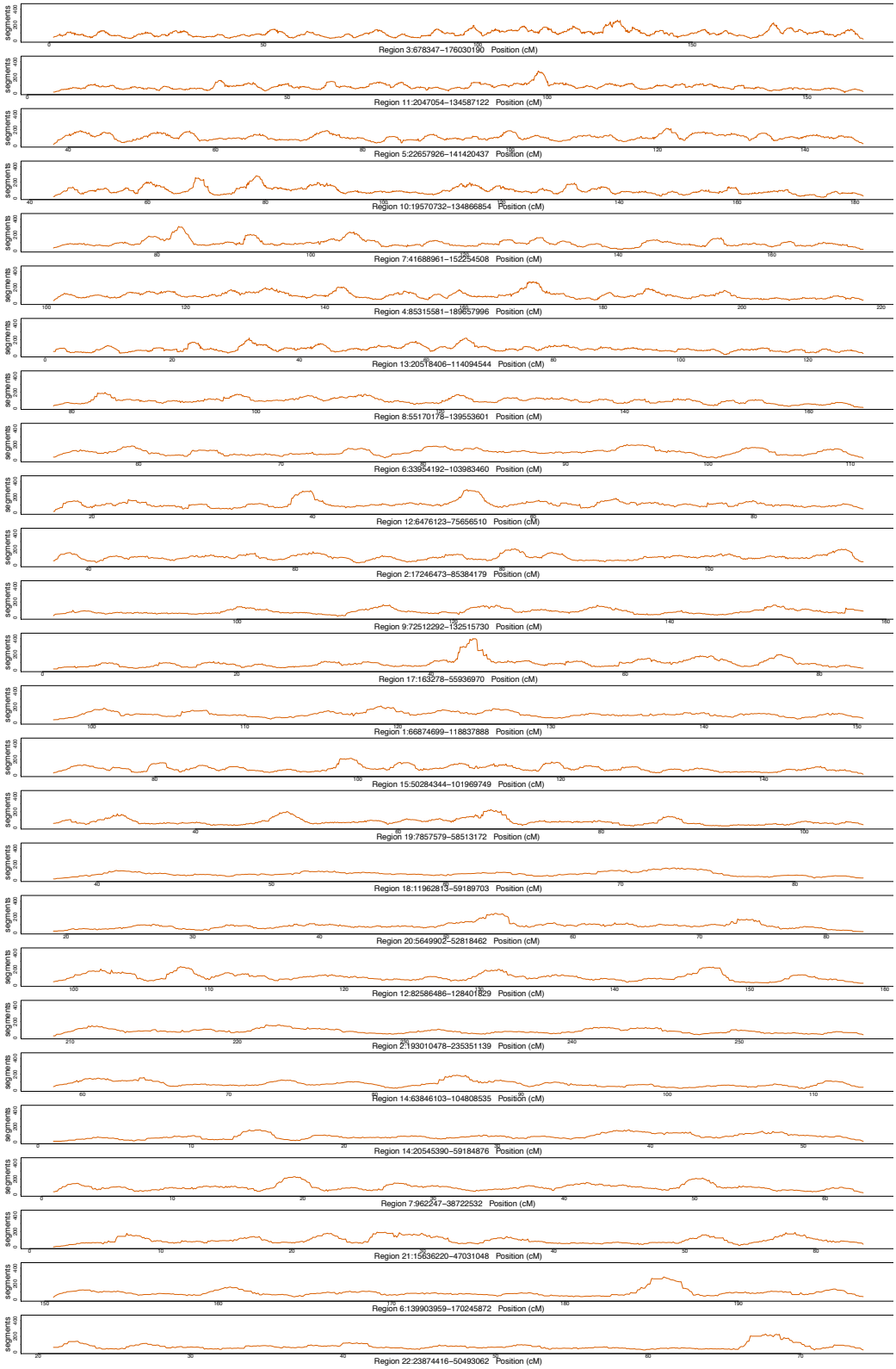

Figure S10: Region-specific density of ≥ 1.6 IBD segments.

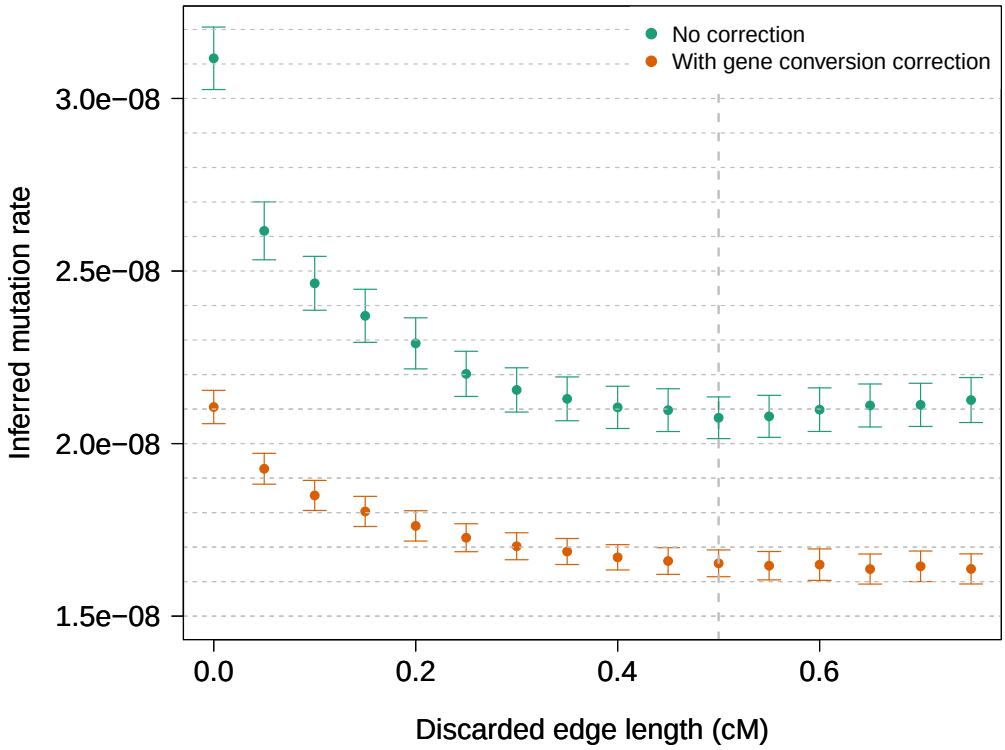

Figure S11: Gene conversion-corrected and uncorrected mutation rates inferred for segments longer than 1.6 cM in the GoNL data set, as a function of the size of discarded IBD segment edge. Inferred values become stable when > 0.5 cM edges are exluded.

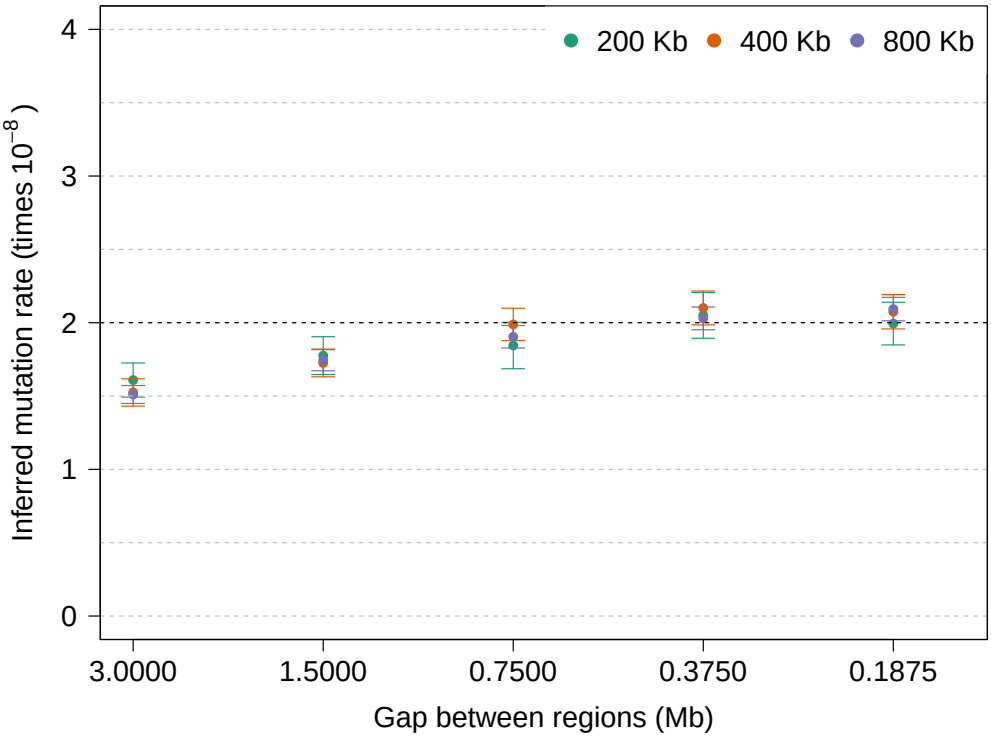

Figure S12: We observed a downward bias when we simulated annotations that are extremely localized, with a large average distance between the analyzed regions. This occurs due to the fact that in our approach we are estimating the age of chromosome-wide IBD segments of a specified length, rather than 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).

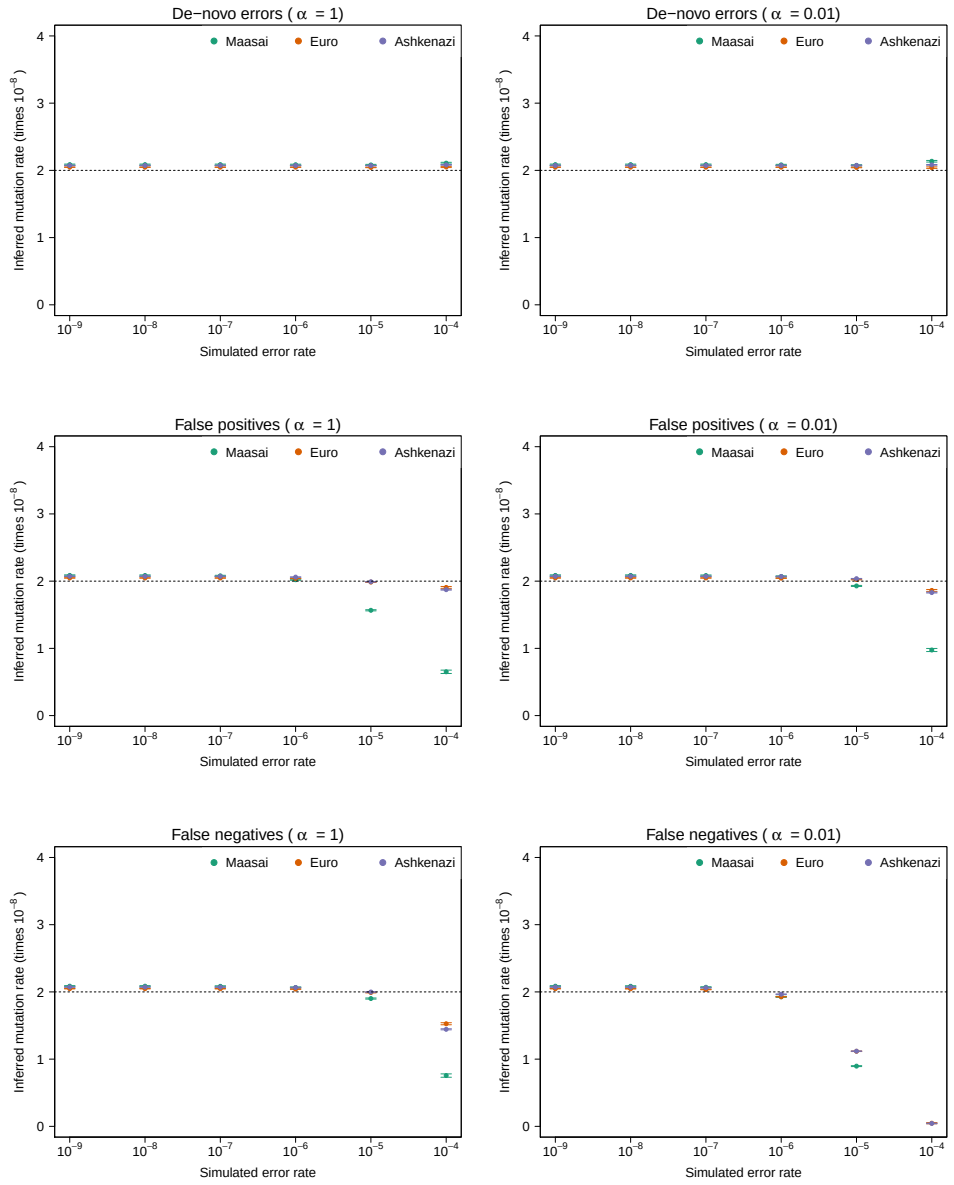

Figure S13: Inferred mutation rates for several values of simulated genotyping error rate, for several types of genotyping errors, demographic history and prior distribution for the frequency of spurious calls. The simulated true underlying mutation rate was $\mu = 2 \times 10^{-8}$. All simulations 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.

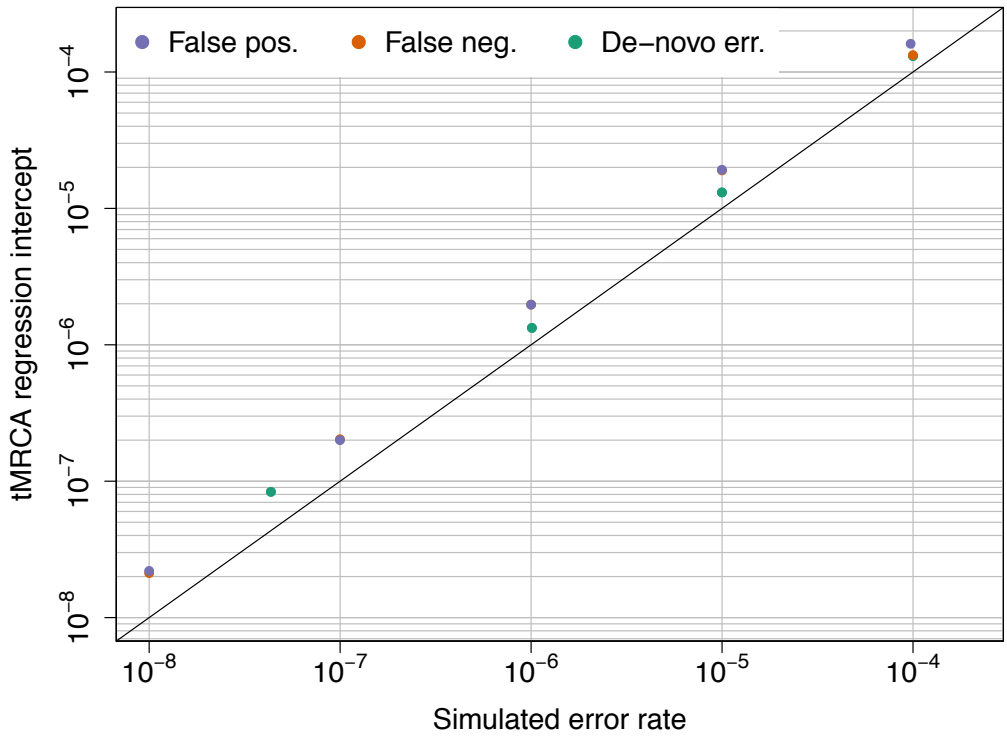

Figure S14: We simulated a chromosome of 250 cM for 100 diploid samples and introduced several types and magnitudes of sequencing errors using the GoNL demographic model. In all cases we used the beta distribution with parameter 0.5 as a prior for the frequency of simulated errors (Figure S5), and the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

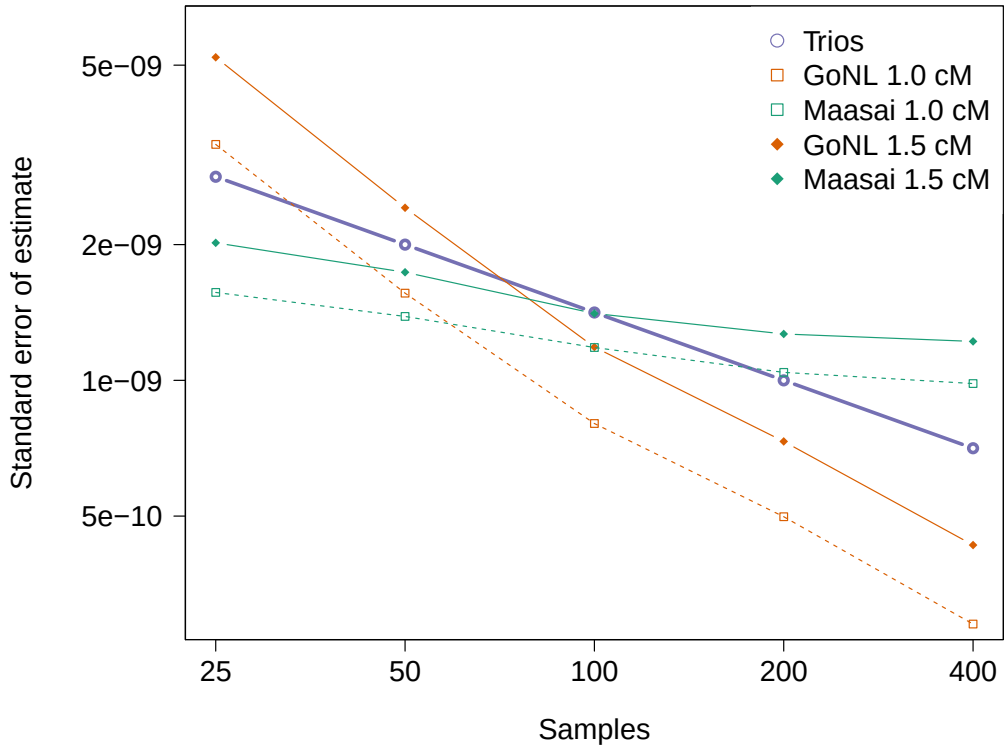

Figure S15: Comparison of the estimate standard error for trios and tMRCA under different demographic models and minimum IBD segment length cut-offs. We report the estimated standard deviation from the analysis of several simulations of a single 100 Mb chromosome.

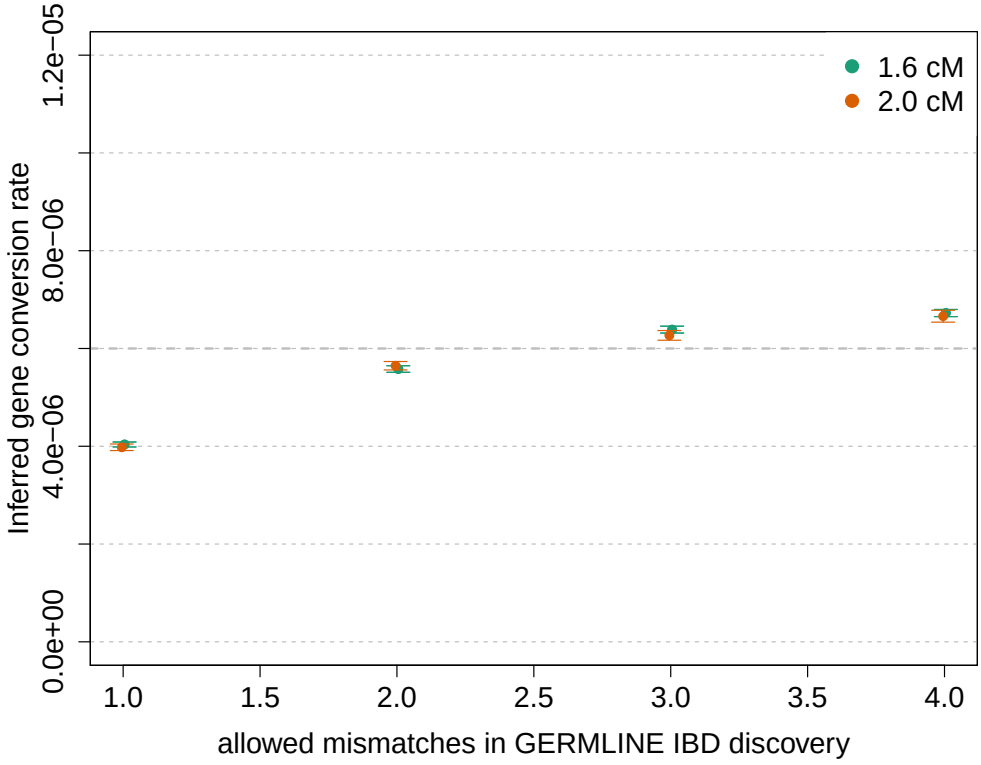

Figure S16: We simulated a chromosome of 50 cM for 250 diploid samples, using $\mu = 2 \times 10^{-8}$ for the mutation rate and a probability of 6×10^{-6} for a basepair to be involved in a non-crossover gene conversion event. We matched the allele frequency spectrum of the simulated samples to the 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.

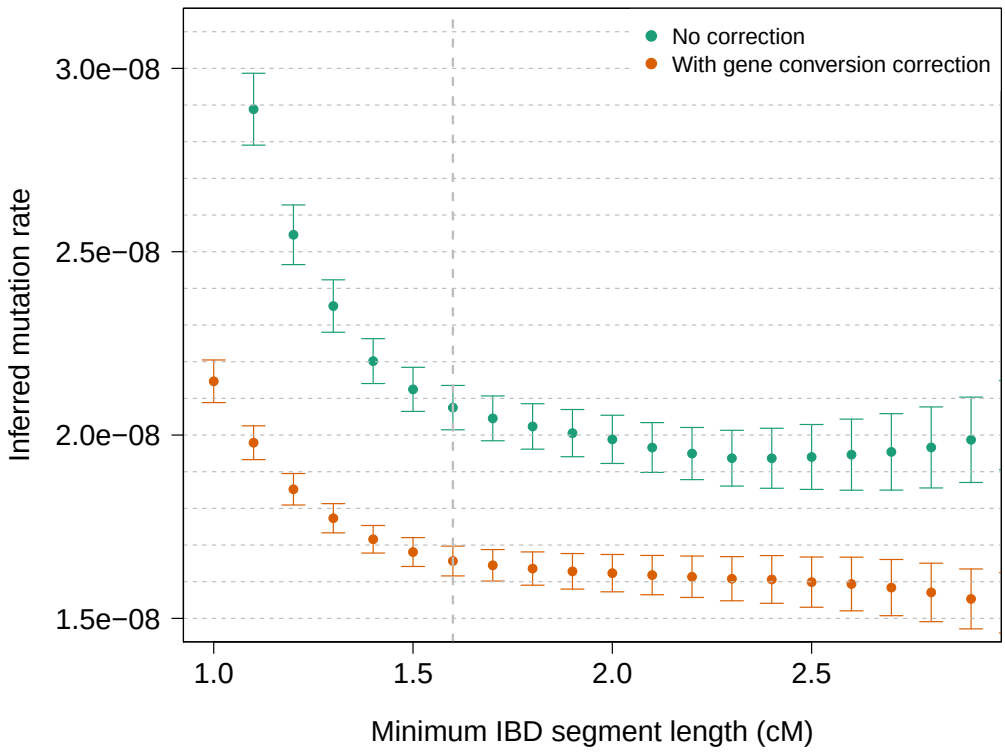

Figure S17: Gene conversion-corrected and uncorrected mutation rates inferred for segments longer than several length thresholds in the GoNL data set.

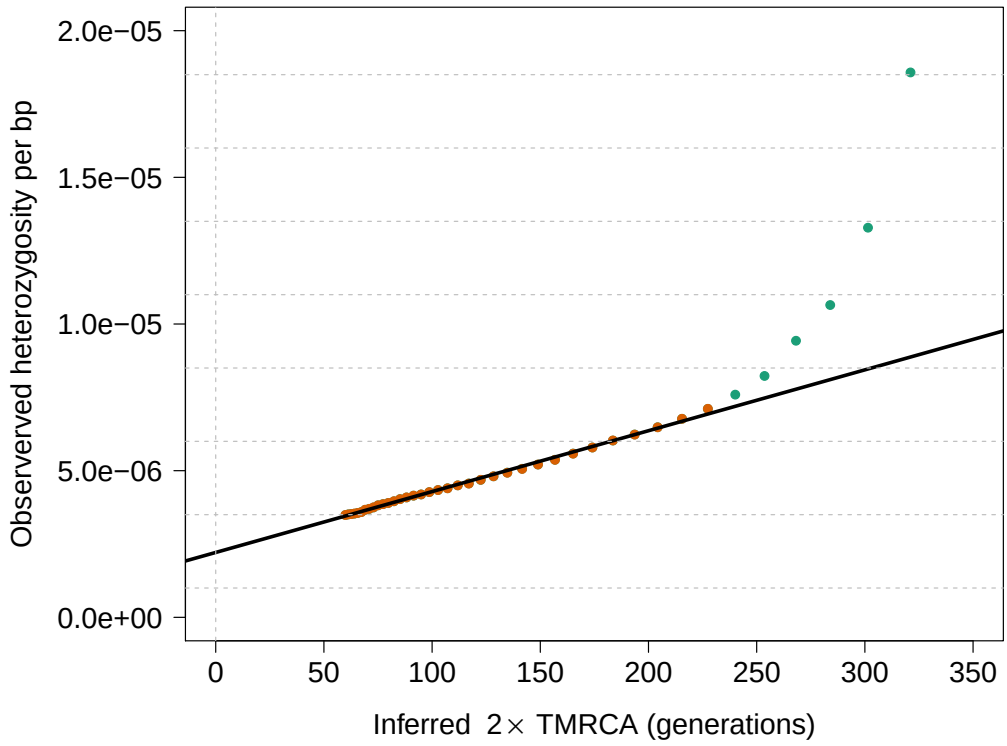

Figure S18: tMRCA regression for segments of length ≥ 1.0 cM in the GoNL data set. The obtained slope is used to estimate mutation rate per generation per base pair, before the effects of gene conversion are accounted for. Segments shorter than 1.6 cM (green) result in mismatching estimates that appear 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.

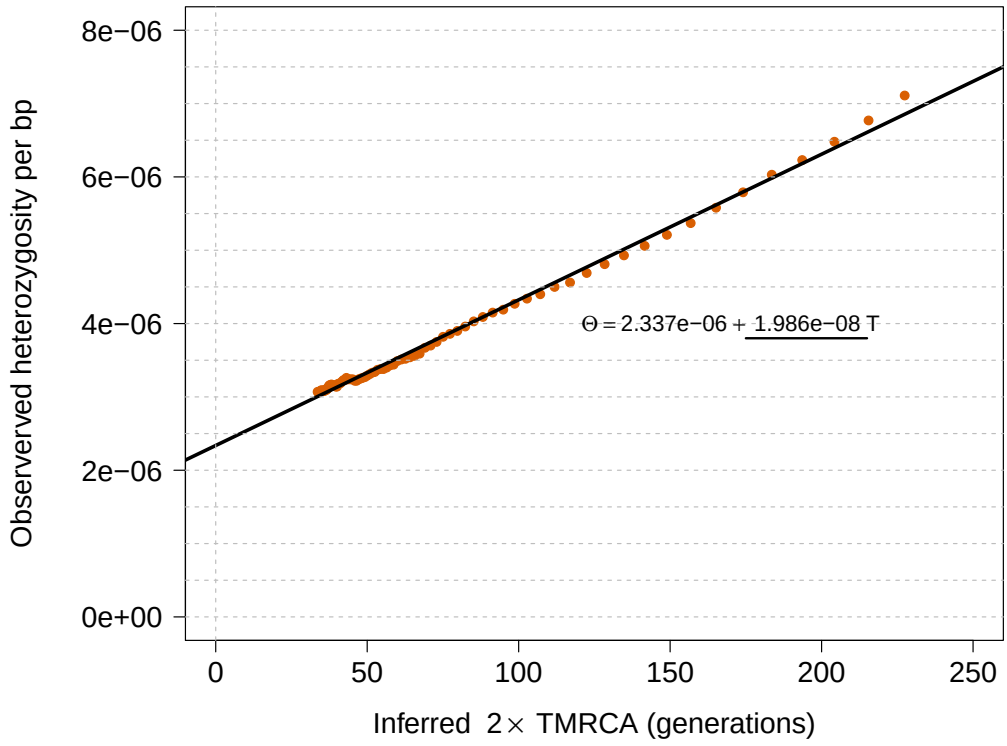

Figure S19: tMRCA regression using segments up to 10 cM.

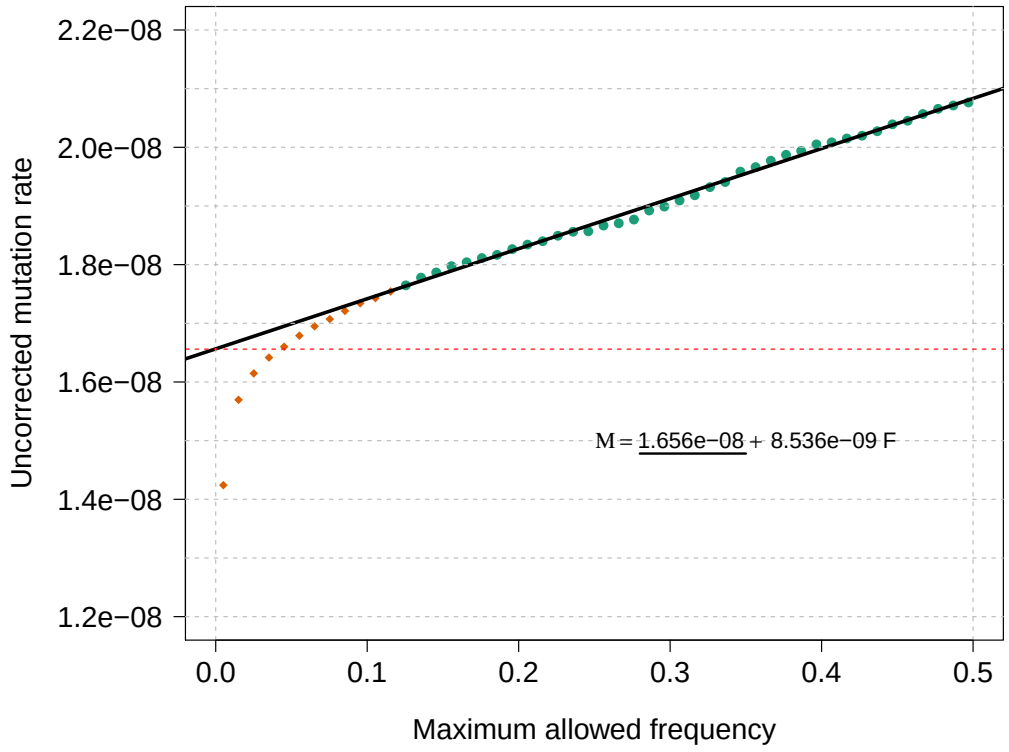

Figure S20: MaAF regression. Red dots show mutation rates for low MaAF values, not used in the regression.

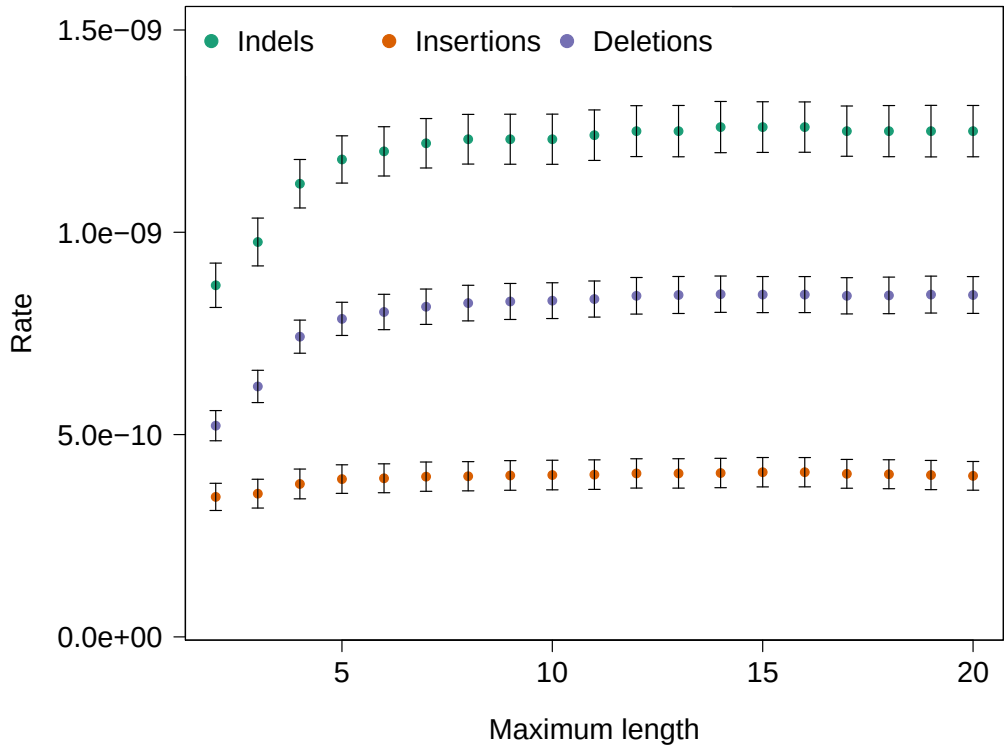

Figure S21: Inferred rate for indels, insertions and deletions, as a function of maximum length.

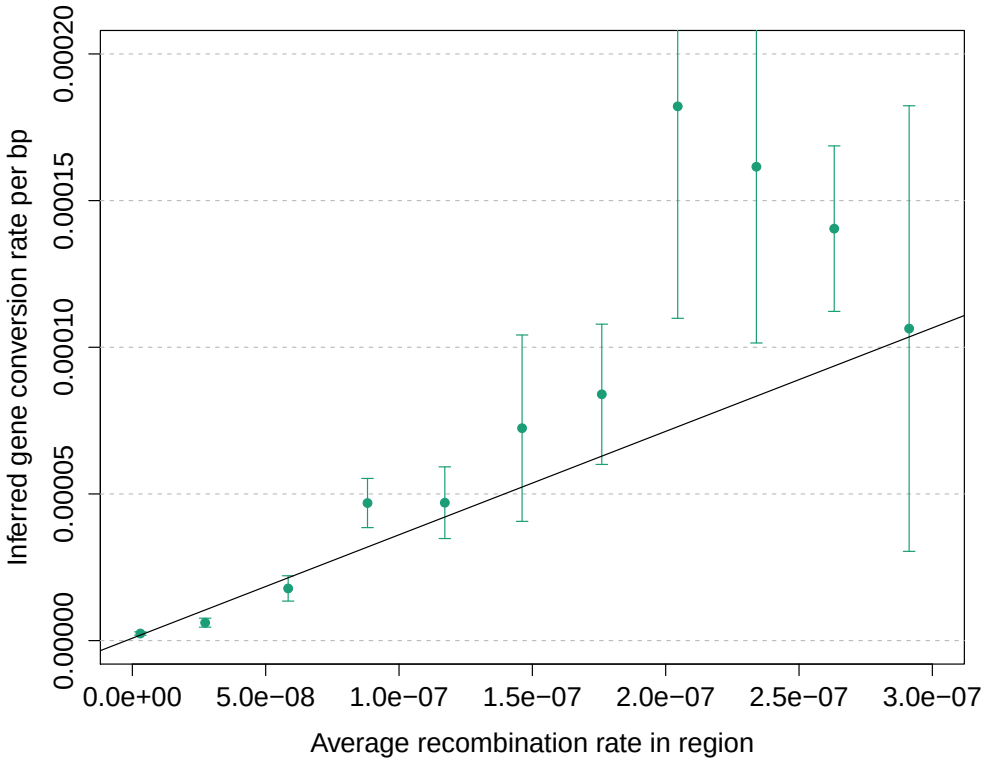

Figure S22: Association between recombination rate and gene conversion rate. We annotated the genome based on uniform bins of recombination rate (per base, per generation), and estimated gene conversion rates for each obtained annotation. We observed association between gene conversion and recombination 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. $= 9.208 \times 10^{-7}$, $p = 0.401$).

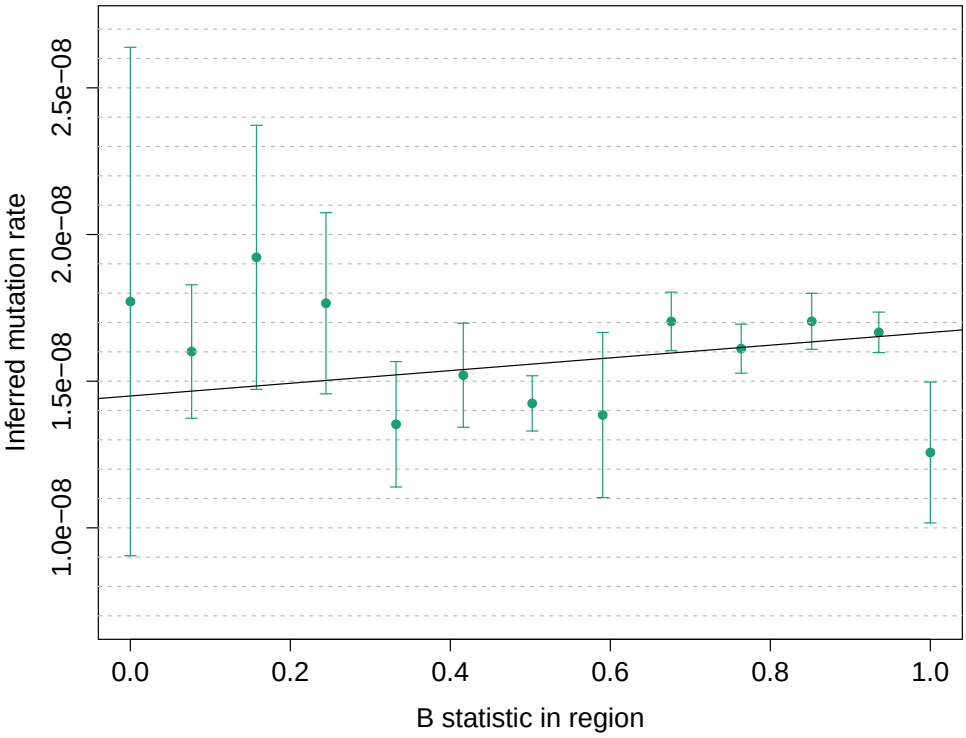

Figure S23: Despite a strong association between average IBD segment length and McVicker B statistic, no significant association is detected between the B statistic and the inferred mutation rates, indicating that the change in local coalescent distributions does not significantly affect the posterior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

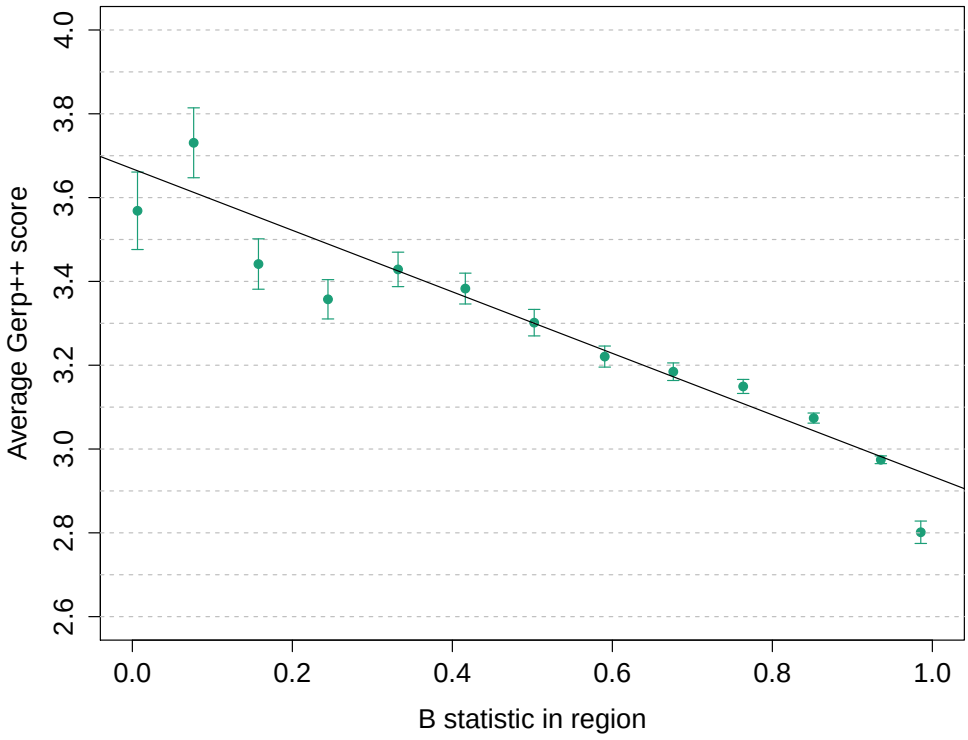

Figure S24: Association between Gerp++ scores of mismatching variants found on IBD segments and average B statistic in IBD regions.

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 $\sim$ 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.