Science Advances
AAAS

Supplementary Materials for

Human generation times across the past 250,000 years

Richard J. Wang *et al.*

Corresponding author: Richard J. Wang, rjwang@indiana.edu

Sci. Adv. **9**, eabm7047 (2023)
DOI: 10.1126/sciadv.abm7047

The PDF file includes:

Supplementary Methods
Figs. S1 to S15
Table S1
Legend for data S1
References

Other Supplementary Material for this manuscript includes the following:

Data S1

S1. Modeling the mutation spectrum as a function of parental age

S1.1. Data from Icelandic trios

We developed a parental age model for the mutation spectrum based on data from a large study of *de novo* mutations in an Icelandic population (*14*). We briefly summarize the findings from this study here as background for the development of our model. The study detected 101,377 single-nucleotide *de novo* mutations from 1,548 trios with known parental ages at conception. In general, they found an increasing number of mutations with both paternal and maternal age, with different rates of increase for different mutation classes. The parent-of-origin was determined for a subset of these mutations (*n* = 41,899), allowing inferences for the mutation spectrum to be made separately for mothers and fathers. Figure S1 summarizes these findings for each of the six different classes of single-nucleotide mutations (A→C, A→G, A→T, C→A, C→G, C→T; each class includes counts from their complements).

S1.2. Description of the Dirichlet-multinomial regression

The mutation spectrum is a form of compositional data: comparisons between spectra focus on differences in the relative abundance of each mutation class. Because of the small number of mutations produced by any one set of parents, observations from a single trio are insufficient to reliably determine the spectrum. A model for the mutation spectrum must therefore incorporate the probabilistic nature of mutation counts from a given trio while inferring the relationship between the underlying spectrum and given covariates. We apply a Dirichlet-multinomial regression to mutation count data to capture the relationship between the underlying mutation spectrum and parental ages, which are treated as covariates in the analysis.

Let $\mathbf{y}_i = \left(y_{i,\mathrm{A\to C}}, y_{i,\mathrm{A\to G}}, y_{i,\mathrm{A\to T}}, y_{i,\mathrm{C\to A}}, y_{i,\mathrm{C\to G}}, y_{i,\mathrm{C\to T}}\right)$ be the vector of mutation counts for each of the six respective mutation classes from trio *i*. The distribution for *m* mutation counts from a trio, $\mathbf{y}_\mathrm{i}$, is modeled as a multinomial, conditional on the probability vector $\mathbf{p}_i$,

$$\mathbf{y}_i \mid \mathbf{p}_i \sim \mathrm{Multinomial}(m, \mathbf{p}_i)$$

where $\mathbf{p}_i$ is defined on the 6-dimensional simplex, $S = \left\{(p_{\mathrm{A\to C}}, p_{\mathrm{A\to G}}, p_{\mathrm{A\to T}}, p_{\mathrm{C\to A}}, p_{\mathrm{C\to G}}, p_{\mathrm{C\to T}}) : p_j \geq 0, \sum_j p_j = 1\right\}$.

We then impose a conjugate Dirichlet prior on **p**, such that $\mathbf{p} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$, and $\boldsymbol{\alpha} = (\alpha_{\mathrm{A\to C}}, \alpha_{\mathrm{A\to G}}, \alpha_{\mathrm{A\to T}}, \alpha_{\mathrm{C\to A}}, \alpha_{\mathrm{C\to G}}, \alpha_{\mathrm{C\to T}})$, $\alpha_\mathrm{j} > 0$. The probability mass function for the count vector **y** over $m = \sum_i \mathbf{y}_i$ trials under this Dirichlet-multinomial model can be represented as

$$f(\mathbf{y} \mid \boldsymbol{\alpha}) = \binom{m}{\mathbf{y}} \frac{\prod_j (\alpha_j) y_j}{(\sum_j \alpha_j) m}$$

(see ref. (*28*)).

The parental ages for each trio are incorporated as covariates for the Dirichlet-multinomial regression, $\mathbf{x} = (\mathbf{x}_{\text{paternal}}, \mathbf{x}_{\text{maternal}})$, an $n \times 2$ matrix of parental ages. They are related to the Dirichlet parameter $\boldsymbol{\alpha}$ by the inverse link function,

$$\alpha_j = e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_j}$$

where $\boldsymbol{\beta}_j = (\beta_{j,\text{paternal}}, \beta_{j,\text{maternal}})$ is the vector of regression coefficients for each mutation class.

S1.3. Subset of mutations and trios for model fitting

For our main analysis we used a subset of mutations from the Icelandic dataset to model the mutation spectrum with parental age: we used only the set of phased mutations for which the parent-of-origin was determined by either read-tracing or transmission to a third generation. Further restrictions on the mutations used for modeling were made to mirror the filters placed on dated variants from the 1000 Genomes Project dataset. These include removing mutations at CpG sites and C→T transitions with a trinucleotide context associated with a putative mutation pulse (see section *S2.2*). We also restricted trios to those that had a minimum of at least 10 mutations. This was done to avoid matrix degeneracy when fitting the maximum likelihood mutation spectrum model (see below). After all filters, we fit the model on 27,902 mutations from 1,247 trios.

S1.4. Fitting the model to mutation data

We used the R package MGLM (*28*) to fit the Dirichlet-multinomial regression model to the filtered mutation dataset. MGLM implements several methods for multivariate generalized linear models, including the Dirichlet-multinomial. We used it to fit the regression coefficients for our covariates (parental age) that maximize the log-likelihood of our model. The result is a predictive model that gives the expected mutation spectrum for a set of parental ages. Figure S2 demonstrates a set of simple predictions from the fit model, showing the expected changes to the mutation spectrum when paternal and maternal age are individually adjusted.

To assess the accuracy of our model, we simulated mutations drawn from the previously fit Dirichlet-multinomial model with known parental ages. In general, our model assumes that the distribution of parental ages in the population is much less important to the mutation spectrum than the mean age. To further explore the fit of our model when we consider population variation in parental age, we drew a stochastic set of parental ages from a multivariate normal distribution. We parameterized this distribution with mean ages from the Icelandic dataset and a scalar product of the covariance matrix that we allowed to vary. Figure S9 shows increasing sum of squared error (SSE) from the underlying simulated spectrum with increasing population variance. Overall, there is very low error, but it increases steadily with variation in parental ages.

S2. Variants from the 1000 Genomes Project dated by GEVA

S2.1. GEVA and the Atlas of Variant Age

Human variants dated by the Genealogical Estimation of Variant Age (GEVA) approach are publicly available in the Atlas of Variant Age, an online database (https://human.genome.dating/). In order to jointly estimate the age of each derived allele, GEVA assumes a constant per-generation mutation rate through time (1.2×10^{-8} per base pair) and a constant per-generation recombination rate through time (varying by locus). Importantly, these age estimates are expressed as generations since the present, and consequently do not require the assumption of any particular generation time. We used dated variants in this database collected from the 1000 Genomes Project (Phase 3; GRCh37). This set includes autosomal variants sampled from 2,504 individuals in 26 worldwide subpopulations within 5 continental populations. Ancestral and derived states were determined in the Atlas of Variant Age through multispecies alignments from the Ensembl database (see ref. (*15*)). Throughout our main analysis, we use the median estimated allele age from the database as a point estimate of each variant's age. See section *S4.3* for an analysis that relaxes this assumption.

S2.2. Filtering dated variants

We took several additional filtering steps to ensure variants were appropriate for estimating generation time. We considered only biallelic single-nucleotide sites that were not singletons—variants that exist on only a single chromosome across samples. We also discarded variants with a derived allele frequency higher than 98% to reduce the possibility of ancestral state misidentification.

As mentioned above, CpG sites are more likely to have arisen more than once, and therefore to have been multiply mapped on genealogies; their frequency is much less consistent across time periods as a result (Fig. S15). As in the model for mutation spectrum with parental age, all variants at CpG sites were discarded from consideration.

Several C→T transitions have been inferred to be part of a recent mutation pulse, particularly in European populations (*19*, *20*). To reduce the potential effect of this mutation pulse on estimates of generation time, we discarded all triplet C→T transitions that have been found to be associated with this pulse. These include ACC→ATC, CCC→CTC, TCC→TTC, TCT→TTT, and their respective reverse complements.

S2.3. Binning data into time periods

After all filtering, there were 25.3 million variants from the Atlas of Variant Age for which there were estimates of allele age. Of these, 20.9 million were estimated to have arisen in the last 10,000 generations. Because there are very few young variants and a long tail for the number of older variants (Fig. S8), we estimated spectra in bins that were supported by equal numbers of variants rather than in equally spaced time periods. We divided the 20.9 million variants equally among 100 bins based on their estimated age. Bins were filled starting with the youngest variants, leaving a small number of the remainder of oldest variants unplaced.

The estimated spectrum for each bin was calculated as the count of variants in each of the six mutation classes divided by the total number of variants in the bin. The age of each bin was calculated as the mean of estimated ages from all variants in the bin. Figure S15 shows the spectra, as a frequency of each mutation class, across 100 bins from the past 10,000 generations. The same procedure was used to estimate historical spectra for each of the continental population groups, for which there were 11.0 (AFR), 4.3 (EAS), 4.4 (EUR), and 5.4 (SAS) million variants included after filtering (see section *S3.4*).

S3. Estimating generation times

S3.1. Fitting variant data to the Dirichlet-multinomial regression model

We jointly estimate separate male and female generation times from the historical mutation spectra calculated from the counts of variants in each time period. To do this, the parental ages in the Dirichlet-multinomial model were treated as parameters in a search for a predicted mutation spectrum that best fit the observed historical spectrum. We minimized the distance between each predicted spectrum and each observed historical spectrum.

Because a mutation spectrum is a composition underlain by count data, comparisons between spectra using simple Euclidean distance can be misleading. Like all compositional data, mutation spectra are mathematically constrained by the possible values for the frequency of each count class, distorting the simple Euclidean distance between compositions. To deal with this, we perform a centered log-ratio transformation (clr) on each spectrum before calculating the distance between them (*29*). The transformation can be obtained as

$$\mathrm{clr}(\mathbf{x}) = \left[\log\frac{x_1}{\mathrm{g}(\mathbf{x})}, \dots, \log\frac{x_\mathrm{D}}{\mathrm{g}(\mathbf{x})}\right]$$

for a composition vector $\mathbf{x}$ with D elements, where $\mathrm{g}(\mathbf{x})$ is the geometric mean of the composition. The Aitchison distance between two given spectra, $\mathbf{x}_1$ and $\mathbf{x}_2$, can then be calculated as $d = \|\mathrm{clr}(\mathbf{x}_1) - \mathrm{clr}(\mathbf{x}_2)\|$.

The generation time was then estimated from each historical mutation spectra by distance minimization as

$$\underset{t_\mathrm{p}, t_\mathrm{m}}{\mathrm{argmin}} \left\| \mathrm{clr}\big(\boldsymbol{F}(t_\mathrm{p}, t_\mathrm{m})\big) - \mathrm{clr}\big(\mathbf{x}_j - \Delta\big) \right\|$$

where $\boldsymbol{F}$ gives the predicted spectrum from the Dirichlet-multinomial model for a set of paternal and maternal ages, t_p and t_m, $\mathbf{x}_j$ is the historical mutation spectrum from a given time period, and Δ is the centering difference, the difference between the most recent bin and the average mutation spectrum, as described in the main text. The parental ages that minimized this distance were found by applying the L-BFGS-B optimization algorithm as implemented in the R *stats* package (*30*). We used the default convergence tolerance, default limit on number of iterations, and set bounds for both parental ages to be: [0, 100]. None of the searches returned a minimum distance at these bounds. The maternal and paternal ages that minimized the distance from each time period were taken to be the respective estimates of the generation time. These ages, as well as the sex-averaged generation time, for all time periods are provided in Supplementary Data S1.

S3.2. Calculating confidence intervals by double-bootstrap

There are two major sources of uncertainty in our estimates of the generation time: (1) the mutation data that specifies the Dirichlet-multinomial regression model, and (2) the dated variants that are used to calculate the variant spectrum in each time period. This led us to construct confidence intervals around the generation time estimates with a double-bootstrap resampling strategy.

The 1,247 trios from the Icelandic dataset were resampled with replacement and fit to the Dirichlet-multinomial regression model. We discarded cases where the likelihood search for the regression model failed to converge, but restricting the dataset to include only trios that had at least 10 mutations greatly reduced instances of failure to converge due to matrix singularity. The variants in each time period of the analysis were also resampled with replacement and the spectrum was recalculated for each bin. Finally, generation times were estimated by fitting the bootstrapped spectrum to the bootstrapped model by distance minimization as described above. The resampling steps were each repeated 100 times, resulting in a total 100 × 100 = 10,000 bootstrap estimates of generation time for each time period included in the analysis.

S3.3. Calculating averages and absolute generation times

The sex-averaged generation time was calculated as the mean of the maternal and paternal ages estimated for each time period. In figures plotting this sex-averaged estimate, we performed local polynomial regression (loess) to produce a smoothed curve across the past. We used the default smoothing parameter, $\alpha = 0.75$, in the R *stats* implementation of loess to smooth both sex-averaged estimates and their confidence intervals.

We calculated the absolute time scale (Fig 2A in main text) on which generation times change by integrating the estimated sex-averaged generation time across the age of mutations. We employed a Riemann sum, calculating the cumulative sum of estimated generation times in single generation steps from the smoothed sex-averaged curve. We added a small constant to this integration to account for the time between the present and the first estimate by assuming there has been no change to generation times in this short period.

A related strategy was used to calculate the average generation times across the period of our analysis. Because ranges for each time period were based on an equal number of variants, older bins span a greater amount of time. We weighted the estimate from each time period by the span of the bin when calculating the average generation times reported in the main text.

S3.4. Estimating population-specific generation times

We separated variants as belonging to one of four continental populations (AFR: Africa, EAS: East Asia, EUR: Europe, and SAS: Southeast Asia) based on their geographic sampling in the 1000 Genomes Project. Variants were placed into continental populations using an inclusive criterion: as long as more than one copy exists among samples from a population, it is included in that population. We analyzed each set of variants separately to arrive at population-specific estimates of generation times (Fig. S4). That is, we repeated each step of the previously described analysis with only the subset of variants included in each population.

In contrast to the broadly inclusive criteria, we also separated variants into each continental population by including only the private alleles exclusive to each population. The proportion of variants that are private to each continental population drops rapidly going back in time, and they make up a very small proportion of variants by 2,000 generations ago (Fig. S5B). Nevertheless, we estimated generation times after creating a subset of variants for each population using only the private alleles. Figure S5A shows the results of this analysis for the first 1,000 generations, before private variants for most populations disappear. These results are very similar to those found using the more inclusive criteria for variants (Fig. 3 in the main text).

S3.5. Goodness-of-fit through time

We took two approaches to quantify how well our generation time estimates fit with mutation spectra across human history. First, we calculated the sum of squared error (SSE) between the spectrum estimated in each bin and the spectrum predicted by the male and female generation times as estimated from our model. Lower SSE values indicate that our model better explains variation in the mutation spectrum. Figure S6A shows the SSE of the best-fit model for the full dataset and the range of SSE values across the double-bootstrap resampled datasets.

Our second approach was to calculate a composite likelihood for the predicted spectrum under a model that treats each mutation class as an independent Poisson regression (Fig. S6B). We previously used this simpler model to describe the mutation spectrum as a function of parental age (*23*). Here we calculate the likelihood for the predicted spectra in each time period under this alternative model to evaluate goodness-of-fit in a way that may better control for sparse data. The likelihood for each time period is calculated as,

$$\log L = \sum_c \frac{\lambda_c^{x_c} e^{-\lambda_c}}{x_c!}$$

where x_c is the number of observed variants in mutation class c, and λ_c is number of predicted counts for mutation class c. Here, λ_c is normalized to the total number of variants binned to each time period, that is,

$$\lambda_c = \frac{N_c(t_\mathrm{p}, t_\mathrm{m})}{\sum_c N_c(t_\mathrm{p}, t_\mathrm{m})} \sum_c x_c$$

where $N_c(t_p, t_m)$ is the number of mutations in class c predicted by the independent Poisson regression for t_p and t_m, the paternal and maternal generation intervals estimated under the Dirichlet-multinomial model (section *S3.1*).

Figure S7 shows both goodness-of-fit analyses run on each continental population separately. The SSE among these populations (Fig. S7A) is associated with the number of polymorphisms in each dataset, with lower error in datasets with more polymorphisms. Though the error in our model varies across time, it is notably not monotonic with time into the past, and the worst fit across populations is roughly coincident with the period with the greatest estimated difference in male-female generation time (Fig. 2B). Figure S7B also shows a relatively stable composite likelihood for predictions across the past 10,000 generations. Note that these likelihoods cannot be meaningfully compared across populations because they describe the fit to different data. Regardless, the composite likelihood of all populations does not increase monotonically with time into the past.

To help interpret SSE values over time, we also performed a cross-validation analysis using the original mutation dataset. We calculated the SSE between spectra estimated from a random subset of 20% of trios from the Icelandic dataset and the remaining 80%. The mean SSE from 100 cross-validation draws was 1.7×10^{-4}. The mean SSE in bins from the past 10,000 generations, 2.0×10^{-4}, is only slightly higher, demonstrating that there is little additional error in our fit model that is not attributable to sampling variance.

S4. Effects from relaxing filters and assumptions

S4.1. Effects of recombination rate on generation time estimates

Recombination could distort our generation time estimates if linked selection or biased gene conversion affect the inferred date of origin of variants in a way that nonuniformly changes historical spectra. Linked selection will change the shape of genealogies (*31*), especially in regions of low recombination. GC-biased gene conversion will change the population frequency of specific variants, but has a greater effect in regions of high recombination (*32, 33*). We carried out additional analyses to ensure the robustness of our results to the effects these processes may have had on the dating of variants within genealogies.

We first considered how differences in recombination rates may have affected parameterization of the model. Mutations identified in the Icelandic trios (GRCh38 positions) were divided into quintiles based on the recombination rate of their surrounding regions (GRCh38 map from https://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/, published 2018-09-25). The mutation spectra was then calculated separately for each quintile (Fig. S10A). We calculated the sum squared difference between the spectra in the first and last quintiles, and found them to be significantly different ($P < 0.005$, empirical CDF determined from cross-validation; see section *S3.5*). These data indicate a different mutation rate, and a slightly different mutation spectrum, associated with variation in recombination. When we estimated generation times based on the mutation spectra from each quintile, they produced a pattern of consistently increasing female generation time with increasing recombination rate (Q1: 24.5 y, Q5: 33.7 y).

That is, when stratified by recombination rate, the mutation data that originally parameterize the model predict longer generation times in strata with higher recombination rates.

Estimates of generation time based on polymorphisms from genomic regions stratified by recombination rate reveal the same patterns. We split variants (GRCh37 positions) into quintiles based on the interpolated human map of recombination (GRCh37 map from https://github.com/joepickrell/1000-genomes-genetic-maps, commit 73cbe92). While our estimates of generation time appear to show an increase with recombination rates (Fig. S10C), the pattern across history remains the same. The longer generation times seen in estimates from regions of higher recombination did not correspond with any bias toward, or accumulation of, G or C alleles in the inferred mutation spectra from high recombination quintiles (Fig. S11). Thus, while the mutation spectrum is affected by local recombination rates, this pattern is equally present in both the mutation data used to fit the model as well as the variants used to estimate historical spectra.

Since our estimates of the generation time are based on a model of the mutation spectrum fit to the whole genome, intra-genomic variation in recombination rate undoubtedly contributes to a proportion of unaccounted variance and error to our model. We considered whether our estimates of generation time may be affected by differences between the proportion of mutations identified as arising in regions with a given recombination rate and the proportion of variants identified in those same regions. We divided mutations and variants across the genome into quintiles based on recombination rate and found a significant difference between the proportion belonging to each bin (χ^2 test, $P < 1 \times 10^{-10}$; Fig S10E). Concerned about the effect this may have on our generation time estimates, we performed a jackknife resampling analysis that matched the proportion of variants in each recombination quintile to the proportion found among mutations. We sampled one-third of all variants in each of 100 replicates, exactly matching the proportion from each mutation quintile, and then re-estimated generation times across the past 10,000 generations. Figure S10G shows that despite significant differences in the recombination rate surrounding mutations versus variants, our estimates of historical generation times were not affected.

S4.2. Effects of replication time on generation time estimates

Another genomic property that has been shown to affect mutation rate and spectra is the timing of replication. Both early and late-replicating regions of the genome are associated with a higher mutation rate; late-replicating regions are also enriched for transversion mutations (*27*; Fig. S10B, S10F). We considered the effects that differences in genomic replication time may have on our generation time estimates with an analysis similar to the one performed for recombination rate variation (section *S4.1*). Dated variants were split into quintiles based on the genomic region's replication time (*27*). We found that estimates of the generation time were progressively longer with variants from increasingly early replication time (Fig. S10D). Generation times estimated separately from each quintile of replication time are more distinct than estimates from separate quintiles of recombination rate (Fig. S10C), showing the stronger effects of replication time on the mutation spectrum across the genome.

As with variation in recombination rate, we were concerned that the mutations used to fit our model may not have arisen in regions with the same replication time as the variants used to estimate historical mutation spectra. We found that the proportions of mutations in quintiles of replication time were significantly different from the proportions among variants (χ^2 test, $P < 1 \times 10^{-10}$; Fig S10F). To estimate the effect this may have had on our estimates, we once again performed a jackknife resampling. We matched the proportions from each quintile of replication time by sampling one-third of all variants in proportion to quintiles among mutations and re-estimating generation times from each of 100 replicates. Figure S10H shows that our estimates of generation times across the past 10,000 generations were not affected by differences in the proportion of replication time between mutations and variants in the genome.

S4.3. No significant effect on estimates from Neanderthal introgression

We considered the possible effects of Neanderthal introgression by repeating our analysis while masking all regions with evidence for introgression in any individual (*34*). This conservative approach, which removed sites regardless of allelic status, masked approximately 38% of the data. Figure S13 shows the results of our analysis with these regions removed. We find little effect on overall generation time estimates (Fig. S13A) or on estimates for non-African populations, the candidates for any effect from Neanderthal introgression (Fig. S13B).

S4.4. Additional effects of relaxing filters and assumptions

We examined several ways in which data or modeling choices might have affected our results. Rather than using only the set of high-quality phased mutations, we fit the Dirichlet-multinomial regression model to a much larger dataset that included unphased mutations from the Icelandic trios (n = 72,573 *de novo* mutations from 1,548 trios). The average age of parents in this dataset (males: 32.0, females: 28.2) is lower than in the smaller phased dataset (males: 33.4, females: 29.1). The results from using this model for analysis are shown in Figure S12A. The male-female difference is slightly accentuated, but the overall pattern for generation times remains the same. We also considered whether our estimates may have been affected by batch effects in the 1000 Genomes Project data, as identified by (*35*). To be conservative, we removed from this analysis all seven nucleotide-triplet mutation patterns identified as being associated with low quality scores. This includes *AC→*CC, TAT→TTT, TCT→TTT, TGT→TTT, and their reverse complements; C→T triplets that we had previously removed continued to be absent in this analysis. Figure S12B shows that the omission of these mutation patterns leads to slightly lower estimates for male generation times, but an overall pattern that remains very similar to our main findings.

As mentioned in the main text, our results were anchored by absolute generation time estimates from the most recent time period. We relaxed this assumption by anchoring to the mean spectrum across all dated variants. This effectively asserts that the Icelandic dataset has a generation time equivalent to the mean generation time across thousands of generations. While estimates of absolute parental age were slightly lower under this assumption, the patterns across human history were unaffected (Fig. S12C). Without any anchoring, estimates of the absolute generation time were much lower; across 10,000 generations, the mean (std. dev.) estimates were: males, 12.4 (2.8); females, 21.4 (1.1); sex-averaged, 16.9 (1.9).

We further investigated whether the difference between the mean spectrum of mutations from the Icelandic trios and the polymorphism data was specific to variants from the 1000 Genomes Project. We compared the spectrum to polymorphism data from a large set of extremely rare variants (ERVs; *22*). This stringently filtered high-quality set of over 35 million variants from 3560 whole-genome sequences is expected to closely resemble *de novo* mutations. Despite this, we find a subtle but significant difference in the mutation spectra ($P < 2 \times 10^{-16}$): the magnitude of this difference is comparable to what we found between the youngest bin and the average spectrum (sum of squared differences: 1.47×10^{-3} vs. 1.50×10^{-3}). Table S1 shows the count, spectrum, and difference between *de novo* mutations from the Icelandic trios and the extremely rare variants from ref. (*22*).

Table S1. Mutation spectrum from extremely rare variants (ERVs) versus *de novo* mutations

	A→C	**A→G**	**A→T**	**C→A**	**C→G**	**C→T**
ERVs	2596232	9686710	2483389	3625994	3135036	14047056
%	7.30	27.23	6.98	10.19	8.81	39.49
DNMs	7141	27178	6898	7697	9734	42729
%	7.04	26.81	6.80	7.59	9.60	42.15
Diff. (%)	-0.254	+0.421	+0.177	+2.600	-0.789	-2.662

Mutations from the Atlas of Variant Age included in our analysis were dated based on the median estimated allele age. To explore the effects of uncertainty in these estimates, we resampled allele ages based on the reported posterior distributions of age estimates. We drew new ages for each variant assuming a normal distribution around the reported 95% highest density interval (negative ages were set to zero). We then repeated the entire analysis, estimating generation times in the last 10,000 generations from 10 such resampled datasets (Fig. S14). Overall, the historical trajectory for human generation times estimated from resampled datasets is within our bootstrap confidence intervals. The exception to this pattern occurs in the earliest bins, where we estimate lower generation times in our resampled ages. This lower estimate is likely due to boundary effects: since we assume a normal distribution for allele ages, alleles close to the present have negative ages set to zero.

We also considered whether the mutation process could be significantly different among populations. Since we know generation times among continental populations in the present are very similar, we reasoned that any evolved differences in the mutation process between populations should be reflected in recent variant spectra. Figure S3 shows there is little difference in the recent spectra between populations, suggesting there are no significant differences in mutation processes between them. These inferences are supported by studies of *de novo* mutations in diverse populations (*36*), which show no differences in the mutation spectrum among parents of similar ages. Similarly, Figure S7 shows no effect of either genetic or geographic distance from Icelandic populations, as would be expected if differences in the mutation process among continental populations had a significant effect on model fit.

Supplementary Figures

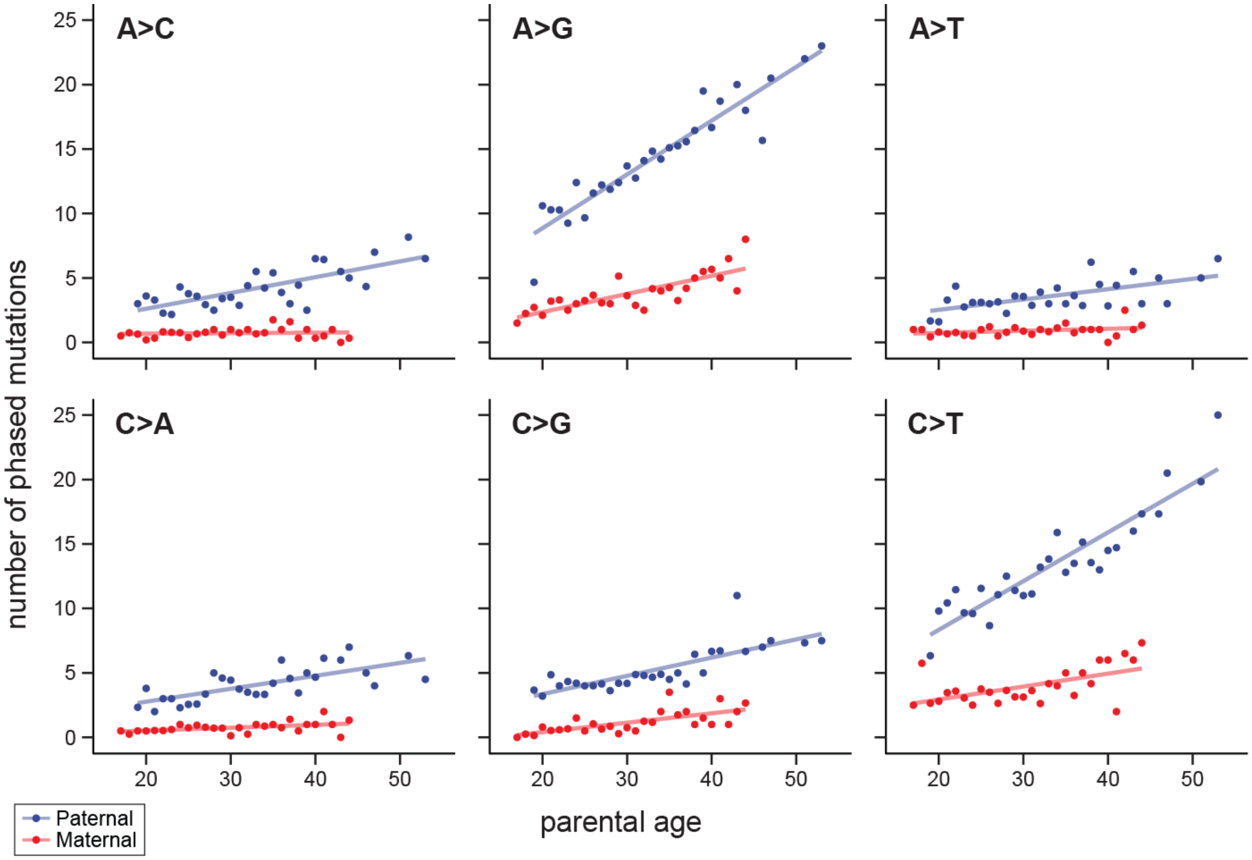

Figure S1. Frequency of mutation classes with parental age

A summary of the number of *de novo* mutations as a function of age. Phased mutations can be assigned to either the paternal or maternal lineage, so are shown separately for the six different types of single nucleotide changes (and their complement). Data from ref. (*14*).

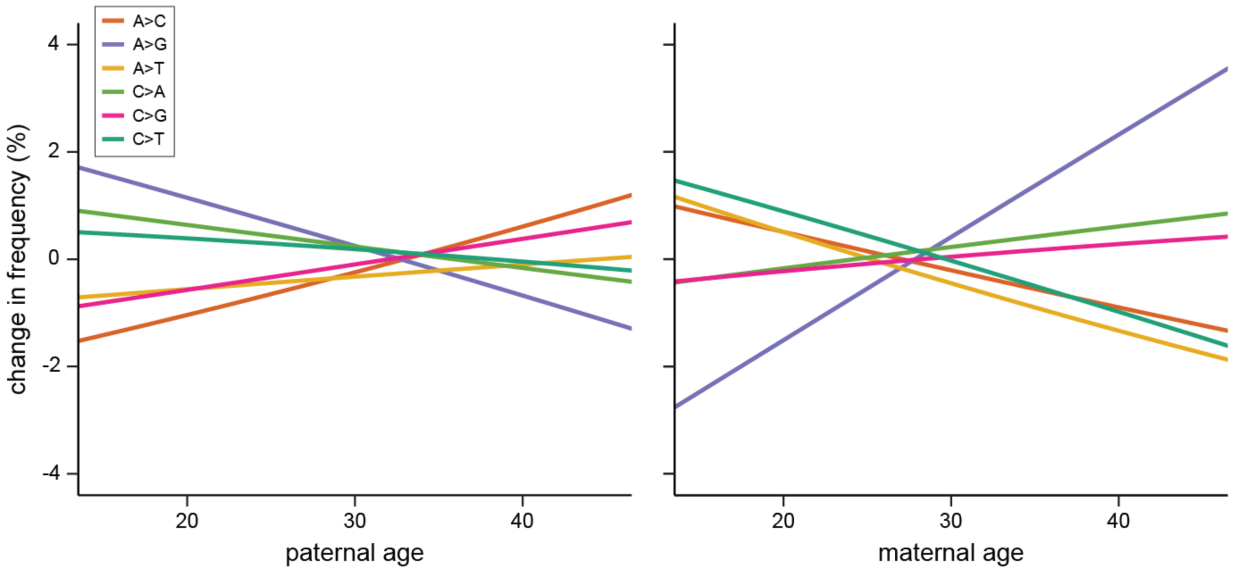

Figure S2. Predicted change in mutation frequency with paternal and maternal age

Data from Icelandic trios (Fig. S1) were used to parameterize the Dirichlet-multinomial model. Figures are centered on the average paternal and maternal ages among the trios (males: 32.0, females: 28.2), and show predicted changes with differences to only paternal (left) and only maternal (right) age. Predicted changes in frequency for each type of mutation are visualized as the difference relative to their frequency at the mean age.

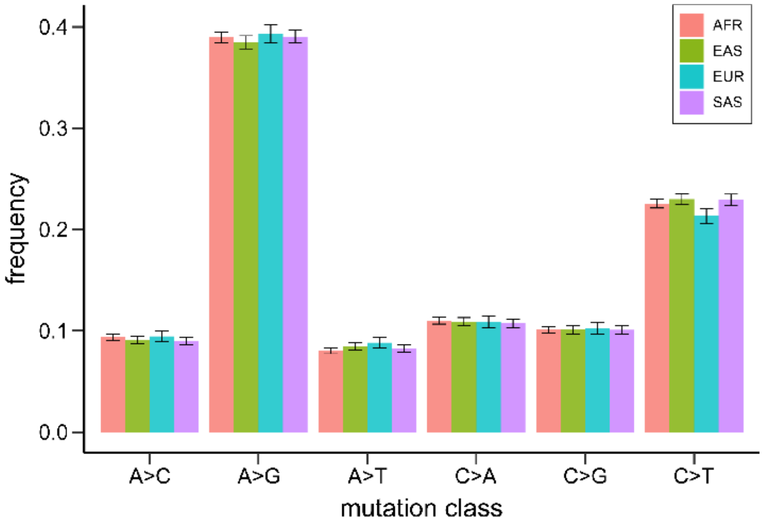

Figure S3. Variant spectrum of the most recent private alleles

The variant spectrum of private alleles for the most recent time period (average variant age of ~80 generations) are the same between continental populations. Error bars show 95% CI.

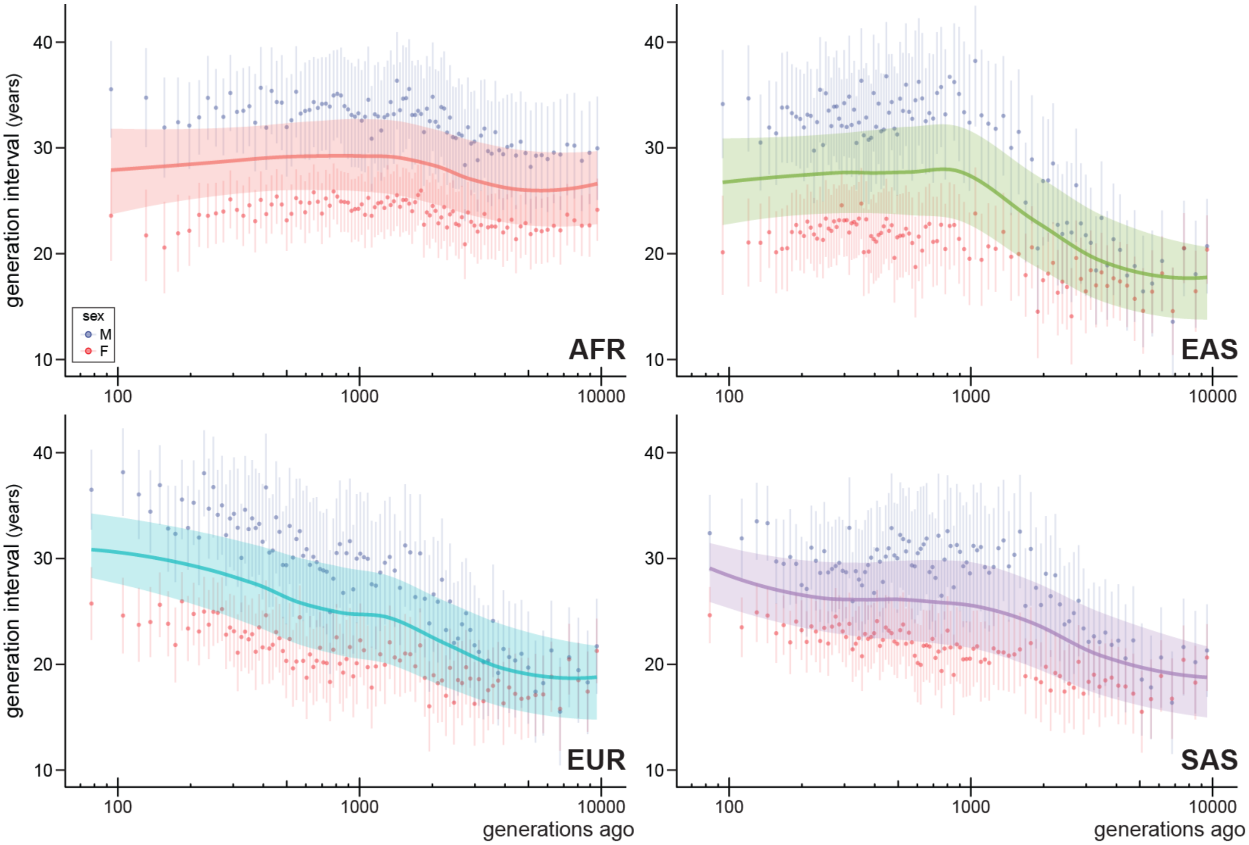

Figure S4. Population-specific estimates of male and female generation interval

Generation intervals were estimated for four major continental populations. These results are the same as those shown in Figure 3 in the main text, but with separate maternal and paternal generation times plotted.

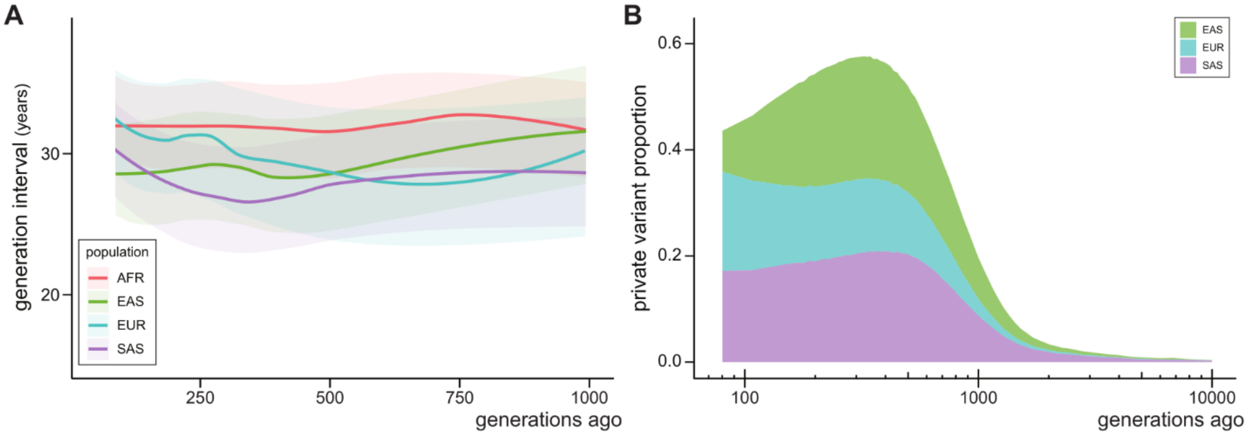

Figure S5. Population-specific estimates from private variants

(**A**) Estimates of the generation interval for each of the four major continental populations using only variants private to each population. These results can be compared to Figure 3 in the main text, but note that here we only plot estimates up to 1000 generations ago. (**B**) The proportion of all variation that is private to one continental population, as a function of time in the past. Almost all variation private to one of the non-African samples has arisen in the most recent 1000 generations.

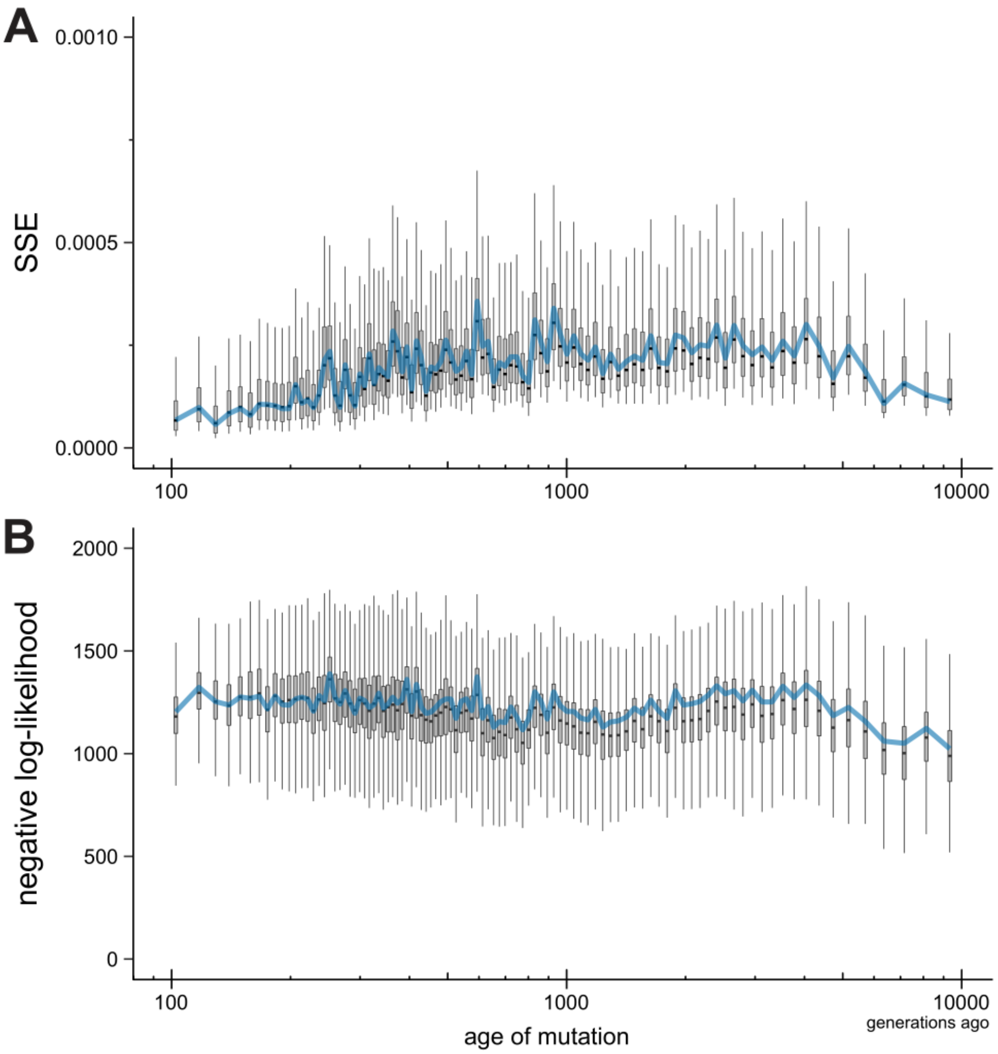

Figure S6. Goodness-of-fit through time

(A) Difference between spectra predicted by parental age estimates and spectra from 1000 Genomes data is shown as the sum of squared error (SSE) for each bin going back 10,000 generations (blue). Lower SSE values indicate that the model better explains variation in the mutation spectrum. **(B)** Composite likelihood of parental age estimates for spectra from each bin under an independent Poisson model for each mutation class (see Supplementary section *S3.5*). Boxplots show goodness-of-fit from analyses of the double-bootstrap datasets (bootstrap outliers not plotted).

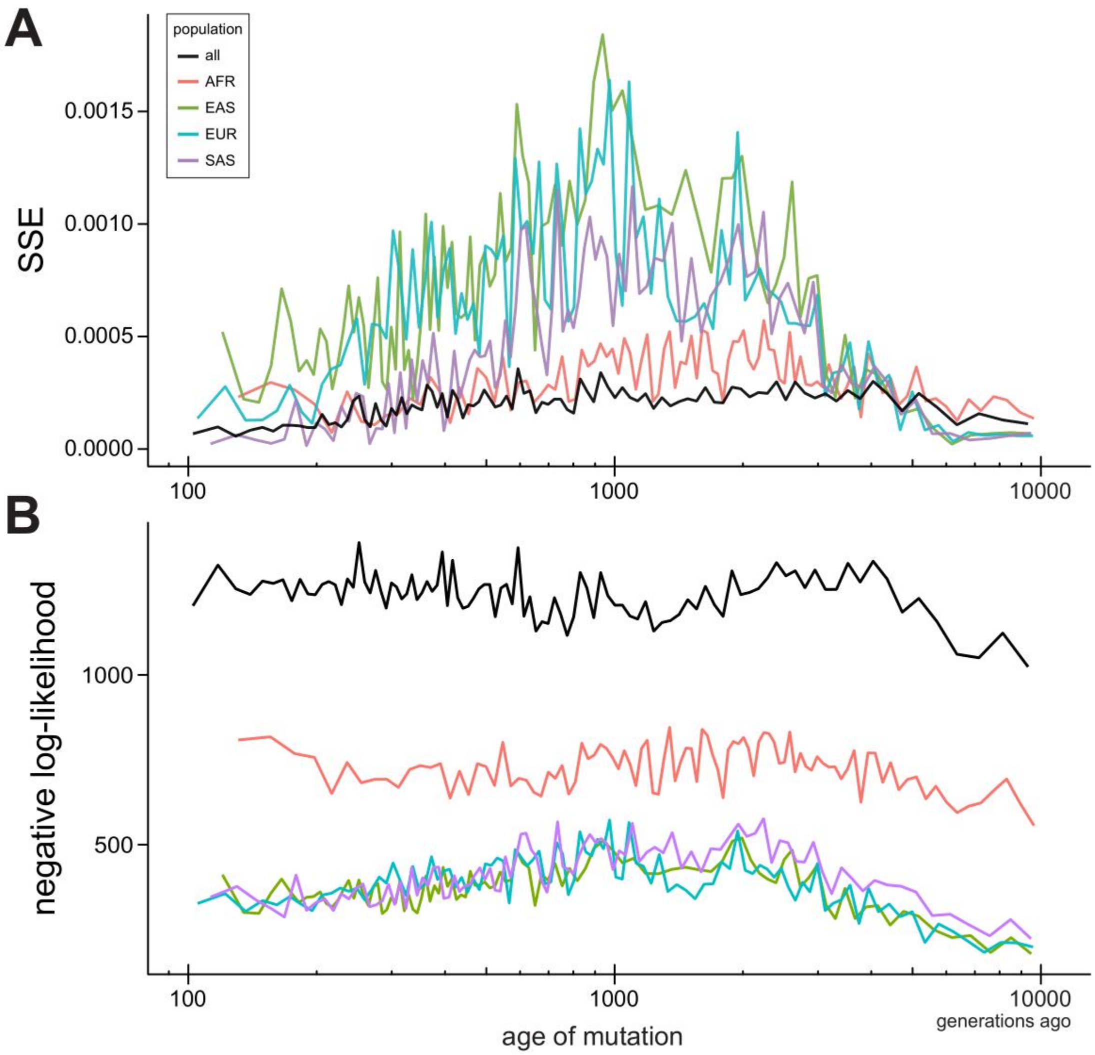

Figure S7. Goodness-of-fit for different continental populations

(A) Difference between spectra predicted by parental age estimates and spectra from analyzed 1000 Genomes continental populations is shown as the sum of squared error (SSE). Lower SSE values indicate better model fit to the observed mutation spectrum. Error among populations appears lower among datasets with more polymorphisms. **(B)** The Poisson composite likelihood is stable for generation time estimates from different continental populations across the past 10,000 generations. Note that, while shown on the same plot, likelihoods across populations cannot be meaningfully compared because they describe the fit to different underlying data.

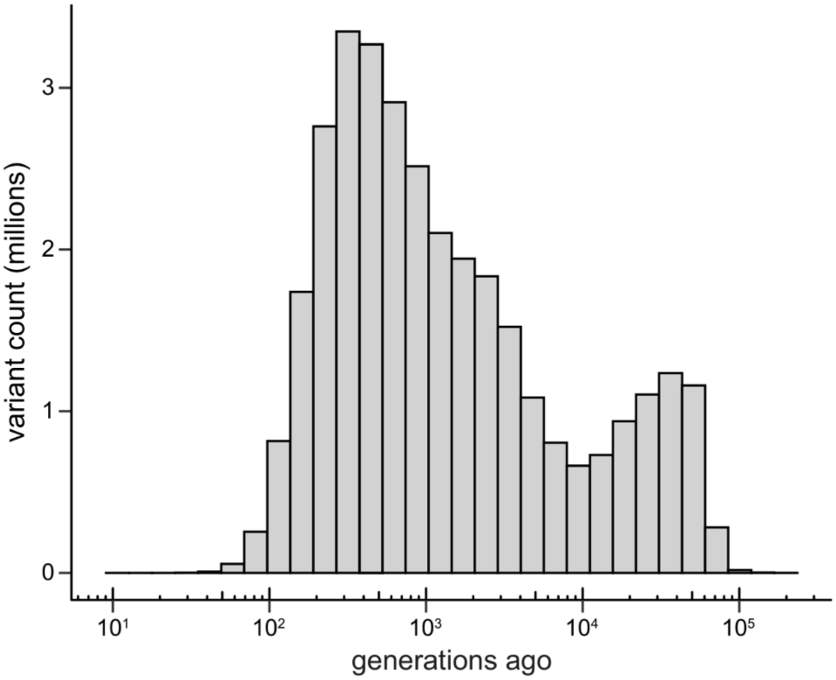

Figure S8. Density of variants by age of origin

Variants dated by GEVA (*15*) are plotted according to the time at which they are estimated to have arisen via mutation. The plot includes all data from the 1000 Genomes Project, regardless of which population(s) they are found in.

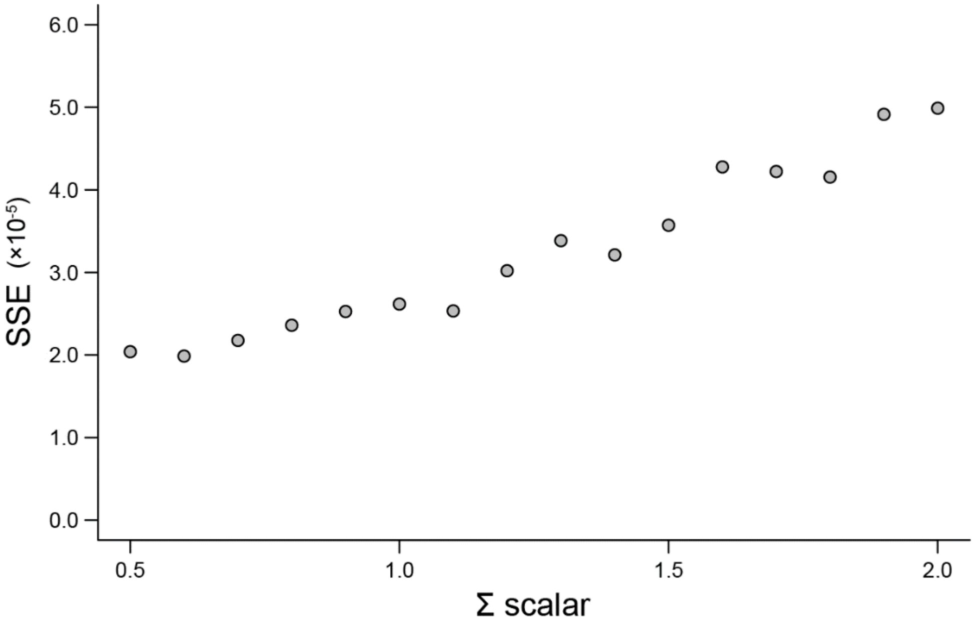

Figure S9. Effect of population variance in parental ages

Model error increases with variation in parental ages. Mutations were simulated as being from parents with a stochastic set of ages. The difference between the predicted spectra based on estimated ages and the simulated mutations is shown as the sum of squared error (SSE). Parental ages were drawn from a multivariate normal with mean and variance from the Icelandic dataset. Increasing variation in the distribution of parental ages was introduced by linear scaling of the covariance matrix. Each point represents the mean difference in SSE from 10,000 simulations.

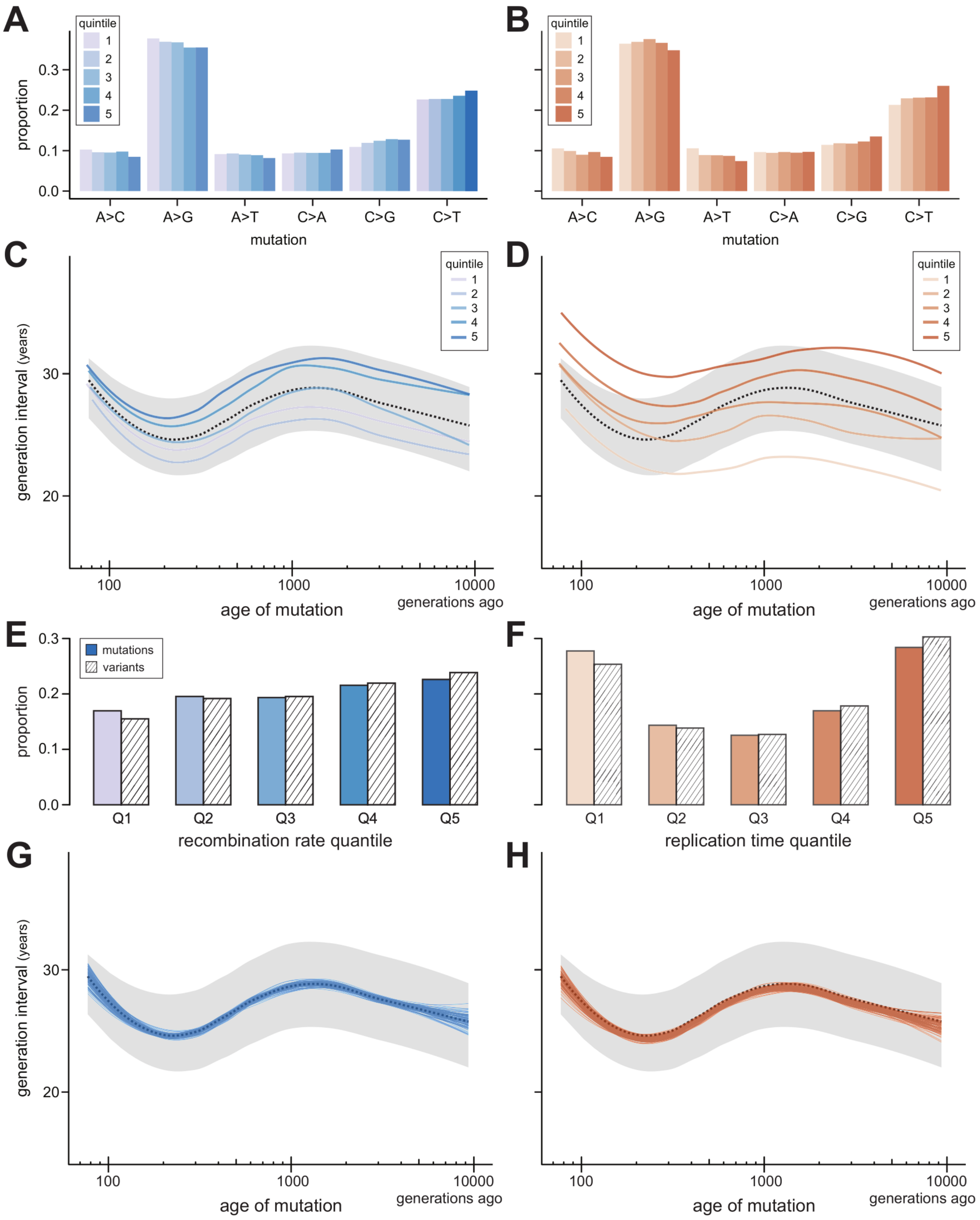
A
quintile
1
2
3
4
5
proportion
0.0
0.1
0.2
0.3
A>C
A>G
A>T
C>A
C>G
C>T
mutation
B
quintile
1
2
3
4
5
A>C
A>G
A>T
C>A
C>G
C>T
mutation
C
quintile
1
2
3
4
5
generation interval (years)
20
30
100
1000
10000
generations ago
age of mutation
D
quintile
1
2
3
4
5
100
1000
10000
generations ago
age of mutation
E
mutations
variants
proportion
0.0
0.1
0.2
0.3
Q1
Q2
Q3
Q4
Q5
recombination rate quantile
F
Q1
Q2
Q3
Q4
Q5
replication time quantile
G
generation interval (years)
20
30
100
1000
10000
generations ago
age of mutation
H
100
1000
10000
generations ago
age of mutation

Figure S10. Stratification of genomic regions by recombination rate and replication time

(A, B) Different mutation spectra in the Icelandic trio dataset are apparent when mutations are binned by the local genomic region's **(A)** recombination rate and **(B)** replication time. Quintiles (1-5) are ordered by increasing recombination rate and earliness of replication time.

(C, D) Different sex-averaged generation time trajectories are inferred when using only the dated variants from specific quintiles of **(C)** recombination rate or **(D)** replication time. Dashed line and shaded area show the estimate and confidence interval from the full dataset.

(**E**, **F**) The proportion of mutations (solid) and variants (hatched) found in different genomic quintiles of recombination rate and replication time. (**E**) An increasing proportion of both mutations and variants are found in regions of higher recombination rate; quintiles ordered by increasing recombination rate. (**F**) The proportion of both mutations and variants are bimodally distributed by replication time across the genome (*27*); quintiles ordered by increasing earliness of replication.

(G, H) Sex-averaged generation time estimates from 100 jackknife resamples where the proportion of variants from each quintile was matched to the proportion found among mutations (see Supplementary section *S4.1*, *S4.2*). These results indicate that there is little effect on estimates due to differences between the proportion of mutations and variants from each **(G)** recombination rate and **(H)** replication time quintile.

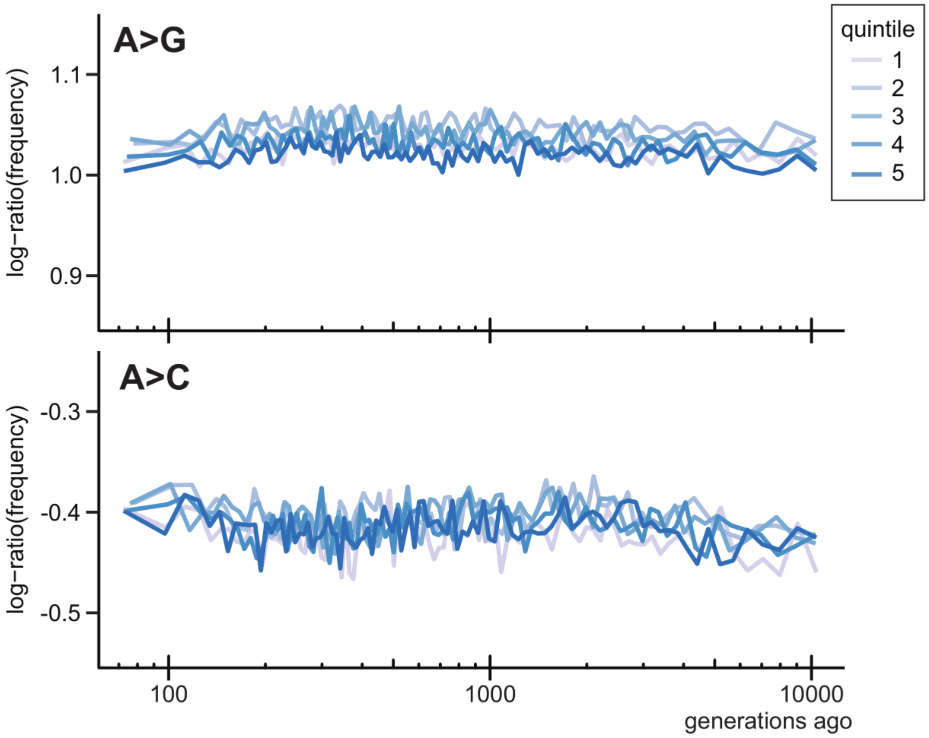

Figure S11. No accumulation of effects from biased gene conversion through time

The frequency of mutations subject to biased gene conversion (to G and C) do not accumulate over time, even across regions with different recombination rates. Frequencies have been center log-ratio transformed (clr; see Supplementary section *S3.1*) to makes differences more visible.

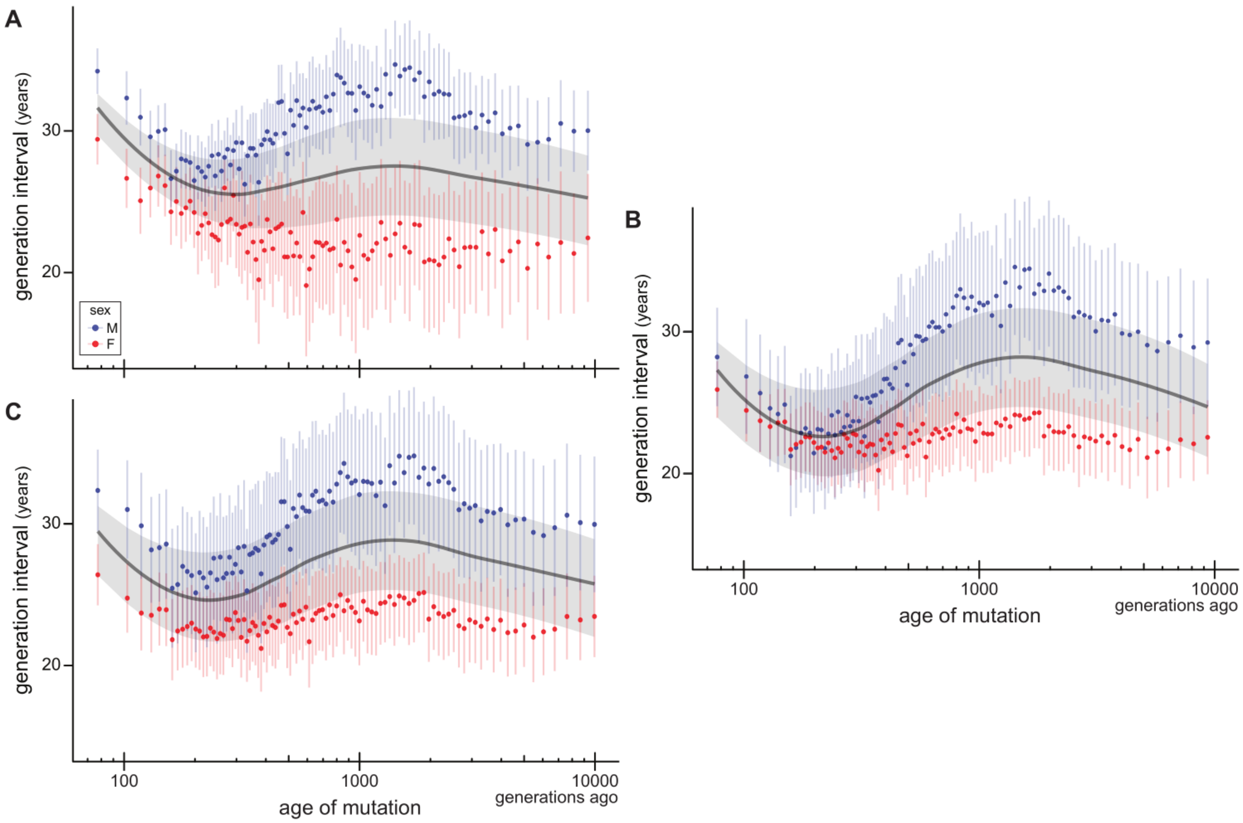

Figure S12. Generation intervals estimated with relaxed assumptions

(A) All de novo mutations from the Icelandic trio dataset (not just phased mutations, as in Fig. 2 in main text) were used to re-parameterize the Dirichlet-multinomial model, and then to re-estimate generation times. **(B)** All seven nucleotide-triplet mutation patterns associated with possible batch effects identified in (*35*) were removed from the analysis. **(C)** Generation times estimated by anchoring the Icelandic mutation frequency spectrum to the average frequency spectrum across all historical time periods.

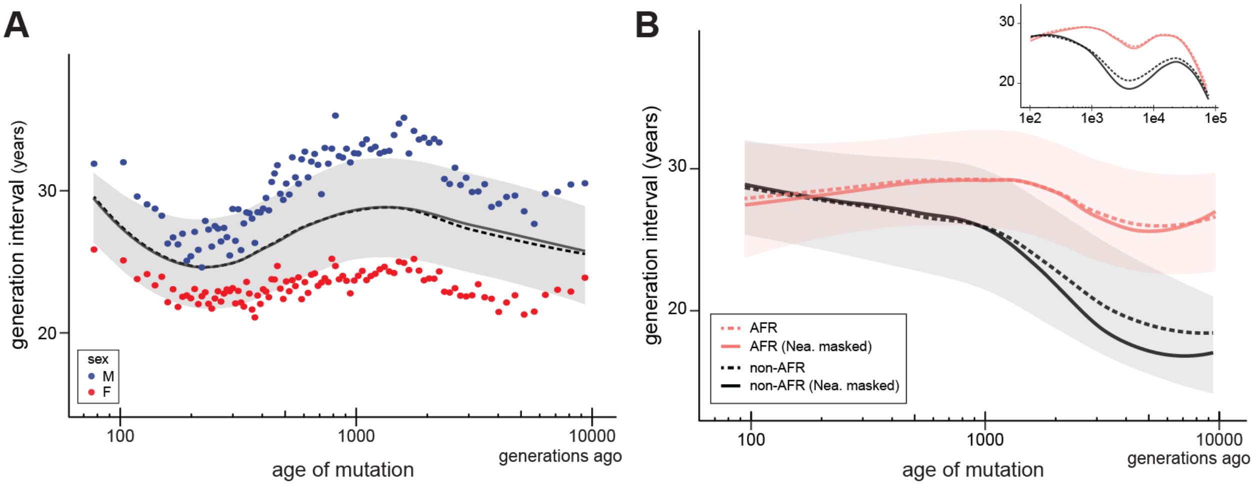

Figure S13. Estimates after masking tracts with evidence for Neanderthal introgression

(A) Estimates of the generation interval over the past 10,000 generations after omitting polymorphisms in tracts with any evidence for Neanderthal introgression. The sex-averaged generation interval with Neanderthal tracts masked (solid line) is little different from the estimate from the full dataset (dashed line, confidence interval shaded). **(B)** Estimates of the generation intervals between African (AFR) and non-African (non-AFR) continental populations were not significantly affected by masking Neanderthal tracts. Confidence intervals displayed are from bootstrap analyses using the full dataset (dashed lines). Inset shows results from including polymorphisms that date back to 78,000 generations ago.

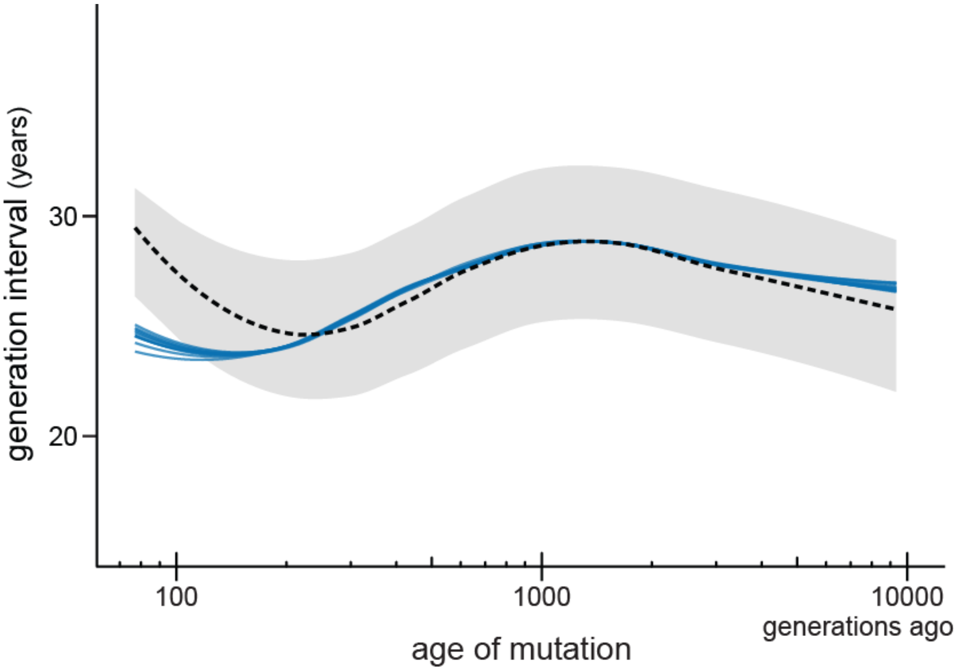

Figure S14. Estimated generation interval from resampled allele ages

Sex-averaged generation interval estimated from datasets using resampled allele ages (blue lines). New ages for each variant were drawn from a normal distribution parameterized by the reported posterior from the Atlas of Variant Age to create 10 resampled datasets. The trajectory of human generation intervals closely matches our estimates using median allele age (dashed line, bootstrap CI shaded), with the exception of the earliest bins where boundary effects dominate (see section *S4.3*).

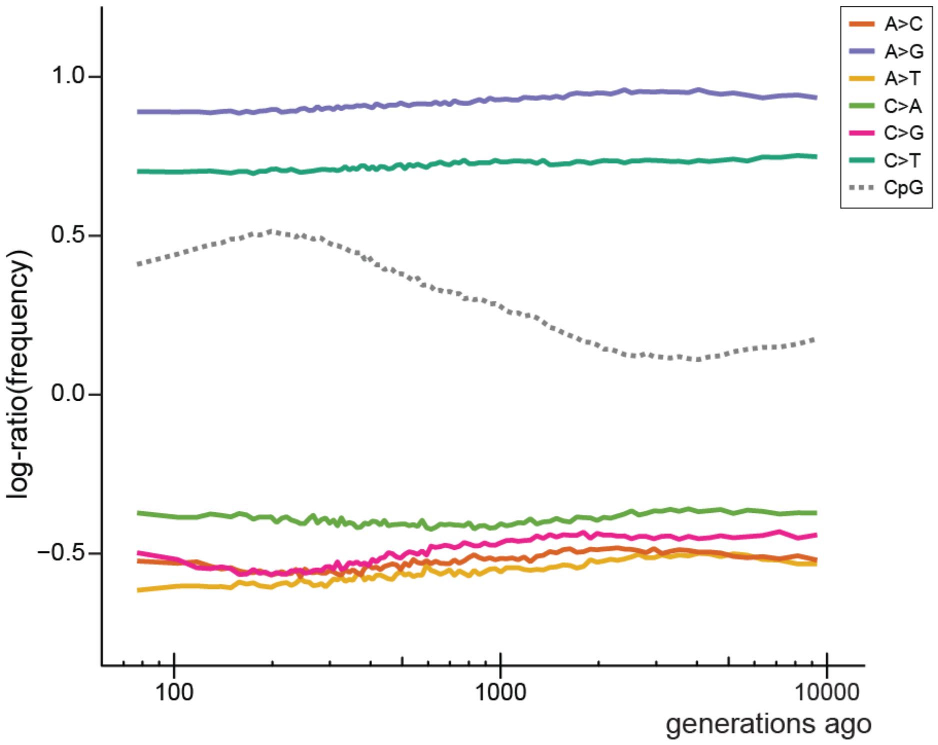

Figure S15. Mutation frequency by age of origin

For each of 100 time periods, the frequency of each type of mutation having been inferred to arise in that bin is plotted. Frequencies have been center log-ratio transformed (clr; see Supplementary section S3.1) to makes differences more visible. In addition to the six types of mutations used in the Dirichlet-multinomial model, we also show the behavior of CpG→CpT mutations for comparison (these were not used in the model).

Data S1.

Estimated male and female generation times for each time period and population.

REFERENCES AND NOTES

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

2. L. Speidel, M. Forest, S. Shi, S. R. Myers, A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).

3. A.-S. Malaspinas, M. C. Westaway, C. Muller, V. C. Sousa, O. Lao, I. Alves, A. Bergström, G. Athanasiadis, J. Y. Cheng, J. E. Crawford, T. H. Heupink, E. Macholdt, S. Peischl, S. Rasmussen, S. Schiffels, S. Subramanian, J. L. Wright, A. Albrechtsen, C. Barbieri, I. Dupanloup, A. Eriksson, A. Margaryan, I. Moltke, I. Pugach, T. S. Korneliussen, I. P. Levkivskyi, J. V. Moreno-Mayar, S. Ni, F. Racimo, M. Sikora, Y. Xue, F. A. Aghakhanian, N. Brucato, S. Brunak, P. F. Campos, W. Clark, S. Ellingvåg, G. Fourmile, P. Gerbault, D. Injie, G. Koki, M. Leavesley, B. Logan, A. Lynch, E. A. Matisoo-Smith, P. J. McAllister, A. J. Mentzer, M. Metspalu, A. B. Migliano, L. Murgha, M. E. Phipps, W. Pomat, D. Reynolds, F.-X. Ricaut, P. Siba, M. G. Thomas, T. Wales, C. M. Wall, S. J. Oppenheimer, C. Tyler-Smith, R. Durbin, J. Dortch, A. Manica, M. H. Schierup, R. A. Foley, M. M. Lahr, C. Bowern, J. D. Wall, T. Mailund, M. Stoneking, R. Nielsen, M. S. Sandhu, L. Excoffier, D. M. Lambert, E. Willerslev, A genomic history of Aboriginal Australia. *Nature* **538**, 207–214 (2016).

4. E. Huerta-Sánchez, X. Jin, Asan, Z. Bianba, B. M. Peter, N. Vinckenbosch, Y. Liang, X. Yi, M. He, M. Somel, P. Ni, B. Wang, X. Ou, Huasang, J. Luosang, Z. X. P. Cuo, K. Li, G. Gao, Y. Yin, W. Wang, X. Zhang, X. Xu, H. Yang, Y. Li, J. Wang, J. Wang, R. Nielsen, Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).

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

6. S. Matsumura, P. Forster, Generation time and effective population size in Polar Eskimos. *Proc. Biol. Sci.* **275**, 1501–1508 (2008).

7. J. P. Bocquet-Appel, When the world's population took off: The springboard of the Neolithic Demographic Transition. *Science* **333**, 560–561 (2011).

8. P. Jordan, in *The Oxford Handbook of the Archaeology and Anthropology of Hunter-Gatherers*, V. Cummings, P. Jordan, M. Zvelebil, Eds. (Oxford Univ. Press, 2014).

9. K. E. Langergraber, K. Prüfer, C. Rowney, C. Boesch, C. Crockford, K. Fawcett, E. Inoue, M. Inoue-Muruyama, J. C. Mitani, M. N. Muller, M. M. Robbins, G. Schubert, T. S. Stoinski, B. Viola, D. Watts, R. M. Wittig, R. W. Wrangham, K. Zuberbühler, S. Pääbo, L. Vigilant, Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15716–15721 (2012).

10. P. Moorjani, S. Sankararaman, Q. Fu, M. Przeworski, N. Patterson, D. Reich, A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 5652–5657 (2016).

11. M. Coll Macià, L. Skov, B. M. Peter, M. H. Schierup, Different historical generation intervals in human populations inferred from Neanderthal fragment lengths and mutation signatures. *Nat. Commun.* **12**, 5317 (2021).

12. A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. W. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, K. Stefansson, Rate of de novo mutations and the importance of father’s age to disease risk. *Nature* **488**, 471–475 (2012).

13. R. Rahbari, A. Wuster, S. J. Lindsay, R. J. Hardwick, L. B. Alexandrov, S. A. Turki, A. Dominiczak, A. Morris, D. Porteous, B. Smith, M. R. Stratton; UK10K Consortium, M. E. Hurles, Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).

14. H. Jónsson, P. Sulem, B. Kehr, S. Kristmundsdottir, F. Zink, E. Hjartarson, M. T. Hardarson, K. E. Hjorleifsson, H. P. Eggertsson, S. A. Gudjonsson, L. D. Ward, G. A. Arnadottir, E. A.

Helgason, H. Helgason, A. Gylfason, A. Jonasdottir, A. Jonasdottir, T. Rafnar, M. Frigge, S. N. Stacey, O. Th. Magnusson, U. Thorsteinsdottir, G. Masson, A. Kong, B. V. Halldorsson, A. Helgason, D. F. Gudbjartsson, K. Stefansson, Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).

15. P. K. Albers, G. McVean, Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* **18**, e3000586 (2020).

16. A.-S. Parent, G. Teilmann, A. Juul, N. E. Skakkebaek, J. Toppari, J.-P. Bourguignon, The timing of normal puberty and the age limits of sexual precocity: Variations around the world, secular trends, and changes after migration. *Endocr. Rev.* **24**, 668–693 (2003).

17. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

18. E. M. L. Scerri, M. G. Thomas, A. Manica, P. Gunz, J. T. Stock, C. Stringer, M. Grove, H. S. Groucutt, A. Timmermann, G. P. Rightmire, F. d’Errico, C. A. Tryon, N. A. Drake, A. S. Brooks, R. W. Dennell, R. Durbin, B. M. Henn, J. Lee-Thorp, P. deMenocal, M. D. Petraglia, J. C. Thompson, A. Scally, L. Chikhi, Did our species evolve in subdivided populations across Africa, and why does it matter? *Trends Ecol. Evol.* **33**, 582–594 (2018).

19. K. Harris, Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3439–3444 (2015).

20. K. Harris, J. K. Pritchard, Rapid evolution of the human mutation spectrum. *eLife* **6**, e24284 (2017).

21. I. Mathieson, D. Reich, Differences in the rare variant spectrum among human populations. *PLOS Genet.* **13**, e1006581 (2017).

22. J. Carlson, W. S. DeWitt, K. Harris, Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation. *Curr. Opin. Genet. Dev.* **62**, 50–57 (2020).

23. R. J. Wang, M. Raveendran, R. A. Harris, W. J. Murphy, L. A. Lyons, J. Rogers, M. W. Hahn, *De novo* mutations in domestic cat are consistent with an effect of reproductive longevity on both the rate and spectrum of mutations. *Mol. Biol. Evol.* **39**, msac147 (2021).

24. A. Keinan, A. G. Clark, Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).

25. A. Helgason, B. Hrafnkelsson, J. R. Gulcher, R. Ward, K. Stefánsson, A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: Evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am. J. Hum. Genet.* **72**, 1370–1388 (2003).

26. F. K. Mendes, M. W. Hahn, Gene tree discordance causes apparent substitution rate variation. *Syst. Biol.* **65**, 711–721 (2016).

27. A. Koren, R. E. Handsaker, N. Kamitaki, R. Karlić, S. Ghosh, P. Polak, K. Eggan, S. A. McCarroll, Genetic variation in human DNA replication timing. *Cell* **159**, 1015–1026 (2014).

28. J. Kim, Y. Zhang, J. Day, H. Zhou, MGLM: An R package for multivariate categorical data analysis. *R J.* **10**, 73–90 (2018).

29. J. Aitchison, *The Statistical Analysis of Compositional Data* (Monographs on Statistics and

Applied Probability, Chapman and Hall, 1986).

30. J. Nocedal, S. J. Wright, *Numerical Optimization* (Springer Series in Operations Research and Financial Engineering, Springer, 1999), vol. 35.

31. N. H. Barton, The effect of hitch-hiking on neutral genealogies. *Genet. Res.* **72**, 123–133 (1998).

32. S. Glémin, P. F. Arndt, P. W. Messer, D. Petrov, N. Galtier, L. Duret, Quantification of GC-biased gene conversion in the human genome. *Genome Res.* **25**, 1215–1228 (2015).

33. J. Lachance, S. A. Tishkoff, Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *Am. J. Hum. Genet.* **95**, 408–420 (2014).

34. M. Steinrücken, J. P. Spence, J. A. Kamm, E. Wieczorek, Y. S. Song, Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Mol. Ecol.* **27**, 3873–3888 (2018).

35. L. Anderson-Trocmé, R. Farouni, M. Bourgey, Y. Kamatani, K. Higasa, J.-S. Seo, C. Kim, F. Matsuda, S. Gravel, Legacy data confound genomics studies. *Mol. Biol. Evol.* **37**, 2–10 (2020).

36. M. D. Kessler, D. P. Loesch, J. A. Perry, N. L. Heard-Costa, D. Taliun, B. E. Cade, H. Wang, M. Daya, J. Ziniti, S. Datta, J. C. Celedón, M. E. Soto-Quiros, L. Avila, S. T. Weiss, K. Barnes, S. S. Redline, R. S. Vasan, A. D. Johnson, R. A. Mathias, R. Hernandez, J. G. Wilson, D. A. Nickerson, G. Abecasis, S. R. Browning, S. Zöllner, J. R. O'Connell, B. D. Mitchell; National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Consortium; TOPMed Population Genetics Working Group, T. D. O'Connor, De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proc. Natl. Acad. Sci.* **117**, 2560–2569 (2020).