Limited role of generation time changes in driving the evolution of mutation spectrum in humans

Ziyue Gao[1+], Yulin Zhang[2], Nathan Cramer[3], Molly Przeworski[4,5], Priya Moorjani[2,3+]

[1]Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA
[2]Center for Computational Biology, University of California, Berkeley, CA
[3]Department of Molecular and Cell Biology, University of California, Berkeley, CA
[4]Department of Biological Sciences, Columbia University, New York, NY
[5]Department of Systems Biology, Columbia University, New York, NY
[+]To whom correspondence should be addressed:
Ziyue Gao (ziyuegao@pennmedicine.upenn.edu) and Priya Moorjani (moorjani@berkeley.edu)

Abstract

Recent studies have suggested that the human germline mutation rate and spectrum evolve rapidly. Variation in generation time has been linked to these changes, though its contribution remains unclear. We develop a framework to characterize temporal changes in polymorphisms within and between populations, while controlling for the effects of natural selection and biased gene conversion. Application to the 1000 Genomes Project dataset reveals multiple independent changes that arose after the split of continental groups, including a previously reported, transient elevation in TCC>TTC mutations in Europeans and novel signals of divergence in C>G and T>A mutation rates among population samples. We also find a significant difference between groups sampled in and outside of Africa, in old T>C polymorphisms that predate the out-of-Africa migration. This surprising signal is driven by TpG>CpG mutations, and stems in part from mis-polarized CpG transitions, which are more likely to undergo recurrent mutations. Finally, by relating the mutation spectrum of polymorphisms to parental age effects on *de novo* mutations, we show that plausible changes in the generation time cannot explain the patterns observed for different mutation types jointly. Thus, other factors—genetic modifiers or environmental exposures—must have had a non-negligible impact on the human mutation landscape.

Introduction

Recent advances in high throughput sequencing have enabled large-scale surveys of genetic variation in thousands of humans, providing a rich resource for understanding the source and mechanisms shaping the mutation landscape over time. Comparisons of polymorphism patterns across geographic population samples have uncovered numerous differences in the mutation rates and spectra (i.e., relative proportions of different types of mutations) (DeWitt et al., 2021; Goldberg and Harris, 2022; Harris, 2015; Harris and Pritchard, 2017; Hwang and Green, 2004; Mathieson and Reich, 2017; Moorjani et al., 2016a; Narasimhan et al., 2017; Speidel et al., 2019). A notable signal in humans is the enrichment of TCC>TTC variants in polymorphism data from Europeans relative to Africans and Asians (Harris and Pritchard, 2017). This signal is also observed in South Asians to a lesser degree and has been suggested to originate in ancient Neolithic farmers (Harris and Pritchard, 2017; Speidel et al., 2021). Many other subtle but statistically significant signals have also been detected; given the recent common ancestry of human populations, this finding indicates that the mutational spectrum in humans has been evolving rapidly.

Several genetic and non-genetic factors have been implicated as affecting mutation rates and acting as potential drivers of observed inter-population differences in the mutation spectrum of polymorphisms. First, some environmental exposures can increase mutation rates, especially of particular types. As humans in different geographic locations and environments may have experienced differential exposures over the past 50,000-100,000 years since the out-of-Africa (OOA) migration, rates of specific mutation types could have diverged between populations (Harris, 2015; Mathieson and Reich, 2017). Second, genetic modifiers of mutation rates, such as variants in genes that copy or repair DNA, could segregate at different frequencies across populations. Despite the deleterious effects of alleles that modify mutation rates, in recombining species, they could be nearly neutral and maintained for a long time, leading to genome-wide differences across populations (Milligan et al., 2022; Seoighe and Scally, 2017).

In addition, direct sequencing of human pedigrees has revealed the effects of the parental ages at reproduction on the relative fractions of mutation types (Goldmann et al., 2018; Jónsson et al., 2017). For example, as parents age, fathers pass on disproportionally more T>C mutations, and mothers contribute a higher fraction of C>G mutations (Jónsson et al., 2017). Thus, differences in the average reproductive ages, or equivalently “generation times”, alone can lead to differences in mutation spectrum across populations; indeed, such differences have been invoked to explain a large fraction of observed variation in types of polymorphisms among population samples (Macià et al., 2021).

The joint distribution of mutation type and frequency of polymorphisms, however, depends not only on the mutational input, but also on other evolutionary forces such as natural selection, biased gene conversion, and demography. In particular, natural selection distorts the allele frequency distribution and fixation probability of non-neutral variants, and the average effect of natural selection can differ across mutation types (Wakeley, 2010). As an example, genic regions tend to be more GC-rich, so mutations at G:C base pairs may be subject to stronger purifying or background selection compared to mutations at A:T base pairs (Lander et al., 2001; McVicker et al., 2009). GC-biased gene conversion (gBGC) is another process that exerts differential effects across mutation types by effectively acting like positive selection favoring mutations from weak alleles (A or T) to strong alleles (C or G) and negative selection against mutations from strong to weak alleles (Duret and Galtier, 2009). The strengths of selection and gBGC depend on the effective population size and thus on the demographic history of a population. Demographic history also influences allele frequencies, for a given allele age (Kimura, 1969). This poses a challenge in interpreting previous studies (Harris and Pritchard, 2017; Mathieson and Reich, 2017) aimed at learning about when changes in mutational processes may have occurred by using allele frequencies as mutations of the same frequency can have drastically different distributions of ages in distinct populations (e.g., doubletons in Africans are substantially older than doubletons in Europeans or Asians) (Mathieson and McVean, 2014) .

Beyond the biological processes that shape polymorphism data, the characterization of the mutational spectrum can be biased by many technical issues. For instance, a recent study showed that some inter-population differences discovered in low coverage 1000 Genomes data may be driven by cell line artifacts or errors in PCR amplification (Anderson-Trocmé et al., 2020). Further, comparisons of mutation patterns across datasets are sensitive to differences in the accessible genomic regions across studies. Because there is large variation in mutation rates and base pair composition across genomic regions, differences in the regions sequenced across studies can have a non-negligible impact on comparisons of mutation spectrum across datasets (Monroe et al., 2022; Seplyarskiy et al., 2021). In addition, the number of genomes surveyed, in combination with the specific population demographic history, influences the chance of observing repeated mutations at the same site, and thus the observed polymorphism patterns (Lek et al., 2016). Given these challenges, it remains unclear if the numerous observed differences across human populations stem from rapid evolution of the mutation process itself, other evolutionary processes, or technical factors.

Motivated by these considerations, we propose a new framework to compare the mutation spectrum over time and across human populations. First, we infer the age of each derived allele observed in a population using a newly developed approach, Relate, which

reconstructs local genealogies and estimates allele ages (Speidel et al., 2019). This approach allows us to perform more reliable comparisons across populations as well as to investigate changes in mutation processes across time. Next, we minimize confounding effects of selection by removing constrained regions and known targets of selection in the genome. Finally, we control the effects of biased gene conversion by focusing on comparison of pairs of mutations (e.g., T>C and T>G) that are subject to similar effects of gBGC. This pairwise comparison further mitigates the issue of interdependencies in comparing mutation fractions (i.e., an increased contribution of one mutation type necessarily lowers the contribution of other mutation types). Based on this new framework, we re-evaluate the evidence for evolution of the mutation spectrum in human populations and investigate when, how, and in which populations significant changes have occurred over the course of human evolution. Finally, by relating parental age effects on the mutation spectrum estimated in contemporary pedigrees to the observed patterns of polymorphisms of varying ages, we evaluate the role of changes in generation times in shaping the human mutation landscape.

Results

Variation in the spectrum of human polymorphisms over time

We analyzed single nucleotide polymorphisms (SNP) identified in high-coverage whole-genome sequencing data from the 1000 Genomes Project, including 178 individuals of West African ancestry living in Ibadan, Nigeria (YRI), 179 individuals of Northern European ancestry living in the United States (CEU), and 103 individuals of East Asian ancestry living in Beijing, China (CHB) (Byrska-Bishop et al., 2022). To focus on putatively neutral mutations, we removed exons and phylogenetically conserved regions as previous studies (Harris and Pritchard, 2017; Moorjani et al., 2016a). To perform reliable comparison between datasets in downstream analysis and to ensure the results are not driven by local genomic differences in mutation rate, we focused on regions that were accessible in both population and pedigree datasets (Materials and Methods; hereafter, referred to as “commonly accessible regions”).

We inferred the age of each derived variant (with respect to the ancestral allele determined based on the 6 primate EPO alignment) in YRI, CEU and CHB using Relate, a method to reconstruct local genealogies based on phased haplotype sequences (Speidel et al., 2019). We then divided all SNPs into 15 bins based on the ages of the derived allele inferred by Relate, accounting for uncertainty in the inferred mutation age by assuming a uniform distribution of ages between the inferred lower and upper bounds for each variant (Materials and Methods). We classified each SNP into six disjoint classes based on the type of base pair substitution: T>A, T>C, T>G, C>A, C>G, and C>T (each including the corresponding substitution on the reverse complement strand, such that,

153 e.g., T>C includes both T>C and A>G substitutions). Given the well-characterized
154 hypermutability of methylated CpG sites (Duncan and Miller, 1980; Kong et al., 2012), we
155 further divided C>T SNPs into sub-types occurring in CpG and nonCpG contexts by
156 considering the flanking base pair on either side of the variant.
157
158 We find marked differences in the relative proportions of different mutation types (i.e., the
159 mutation spectrum) across varying allele age bins within CEU (Figure 1) as well as in YRI
160 and in CHB (Figure 1—figure supplement 1), as seen earlier in low coverage 1000G data
161 (Speidel et al., 2019). We obtain qualitatively similar results when considering other
162 1000G populations of TSI, LWK and JPT (Figure 1—figure supplement 1). This
163 observation echoes previous findings about the evolution of the mutation spectrum
164 comparing polymorphisms across allele frequencies (Carlson et al., 2018; Harris and
165 Pritchard, 2017; Mathieson and Reich, 2017). As noted previously, however, differences
166 in mutation spectrum across frequencies alone are weak evidence for the evolution of the
167 mutation process itself, because patterns of standing polymorphisms can be affected by
168 repeat mutations and other evolutionary forces, including selection and gene conversion.
169
170 Notably, the infinite sites model is a reasonable assumption for small sample sizes
171 (Kimura, 1969), but recurrent mutations become highly likely in large datasets, especially
172 at sites with higher mutation rates (Harpak et al., 2016; Lek et al., 2016). Recurrent, multi-
173 allelic, and back mutations violate the model assumptions of Relate and are often
174 excluded from its output. For instance, given the higher mutation rate of transitions at
175 CpG sites, such SNPs are more likely to be subject to recurrent mutations in a large
176 sample and thus may map to multiple branches in the tree, leading to their exclusion from
177 Relate's output (Speidel et al., 2019). As expected from these considerations, the fraction
178 of CpG C>T SNPs in young mutations (i.e., those estimated to have occurred in the past
179 ~50 generations) is lower than proportions in *de novo* mutations in present-day pedigree
180 studies (Figure 1—figure supplement 2). Differences in mutation spectrum across age
181 bins in modern humans persist even after excluding CpG C>T mutations (Figure 1—figure
182 supplement 3), however, indicating that other mutation types are also changing in relative
183 frequency over time and the observed patterns are not driven solely by recurrent mutation
184 at CpG sites.
185
186 Next, we examined the effect of linked selection on different mutation types. While we
187 excluded direct targets of selection from analysis (i.e., exons and conserved regions),
188 much of the genome is linked to non-neutral variants and subject, notably, to background
189 selection (Charlesworth et al., 1993; McVicker et al., 2009; Murphy et al., 2022). A
190 common measure of the effects of background selection is the *B*-statistic or *B*-score that
191 estimates the reduction in nucleotide diversity levels compared to the neutral expectation
192 (McVicker et al., 2009). To characterize the impact of linked selection, we calculated the

average genome-wide *B*-score of each mutation type. We find nearly identical average *B*-scores and similar distributions for all mutation types (Figure 1—figure supplement 4). Further, comparing the mutation spectrum over time in CEU, YRI and CHB, we obtain qualitatively similar results when restricting to regions with weak background selection (*B-score* > 800, where the genetic diversity is reduced by <20% compared to the neutral expectation; Figure 1—figure supplement 5). These analyses suggest that although linked selection has pervasive effects, its average impact is relatively uniform across the seven mutation types in commonly accessible regions (Materials and Methods).

Gene conversion is another evolutionary process that can have a profound impact on the mutation spectrum of polymorphisms. GC-biased gene conversion (gBGC) acts like selection for certain mutation types, by causing the preferential transmission of strong (S) alleles (C or G) over weak (W) alleles (A or T) in heterozygotes (Duret and Galtier, 2009). Accordingly, we observe enrichments of W>S mutations (T>A and T>G) in common variants and of S>W mutations (C>A and C>T) in rare variants (Figure 1—figure supplement 6A). Moreover, gBGC violates model assumptions of Relate (for both neutrality and infinite-sites mutation model) and could lead to subtle biases in estimated allele ages (Speidel et al., 2019). Due to the effect of gBGC, W>S mutations are expected to be enriched in older variants compared to S>W variants, and this enrichment is expected to be stronger in regions with high recombination rates. Indeed, we observe such enrichment and the expected correlation with recombination rate (Figure 1—figure supplement 7A), supporting the effect of gBGC on the mutation spectrum of variants across mutation ages. Furthermore, the effect of gBGC is expected to vary across populations, as its strength depends on the effective population size. Accordingly, we observe that the trends of the ratio of W>S to S>W over time differ across human populations (Figure 1—figure supplement 6B; Figure 1—figure supplement 7B). These results highlight the need to account for gBGC in order to reliably interpret the source of observed differences within and between populations (whether using allele frequency bins or allele age estimates).

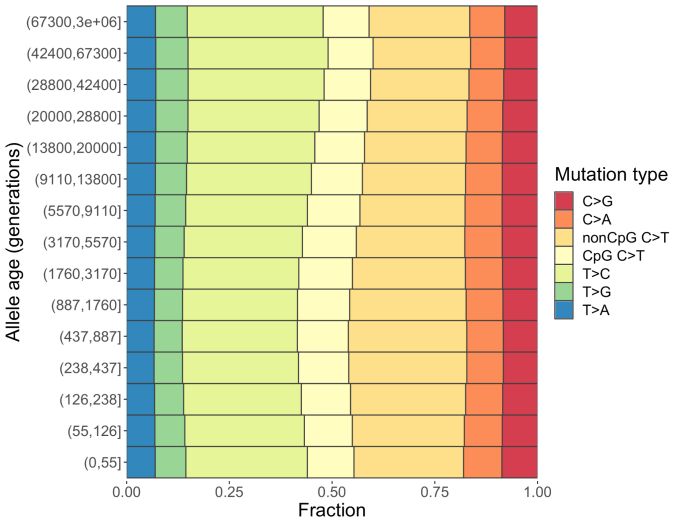


Figure 1. Changes in the mutation spectrum of polymorphisms in CEU over evolutionary time.

The following source data and figure supplements are available for figure 1:

Source data 1. Relate output for SNPs in commonly accessible region with additional annotation (one file for each population of YRI, LWK, CEU, TSI, CHB, JPT). Available at: doi:10.6078/D19B0H

Source data 2. Text file with (pseudo-)counts of different types of mutations in YRI, LWK, CEU, TSI, CHB, JPT in each time window.

Figure supplement 1. Mutation spectrum of polymorphisms stratified by allele age in six human populations (YRI, LWK, CEU, TSI, CHB, JPT).

Figure supplement 2. Mutation spectrum of CEU polymorphisms and DNMs from Icelandic trios (Halldorsson et al., 2019).

Figure supplement 3. Mutation spectrum of CEU polymorphisms and DNMs from Icelandic trios, excluding C>T transitions at CpG sites.

Figure supplement 4. Mean and distribution of B-scores of different mutation types.

Figure supplement 5. Mutation spectra of human polymorphisms in genomic regions with weak background selection (*B*-score>800).

Figure supplement 6. Fractions of S>S, S>W, W>S, and W>W mutations stratified by (A) derived frequency and (B) allele age.

Figure supplement 7. Distribution of variants subject to biased gene conversion stratified by allele age and recombination rate, (A) Fraction of S>S, S>W, W>S, and W>W mutation types and (B) Ratio of W>S to S>W mutations.

Figure supplement 8. Mutation spectrum of human polymorphisms in six populations (YRI, CEU, CHB, LWK, TSI, JPT) stratified by allele age based on alternative binning strategies.

Pairwise comparisons of mutation types accounting for gBGC

In light of the impact of gBGC on the mutation spectrum, we focused on comparisons of pairs of mutation types subject to similar effects of gBGC (i.e., in which both are favored, disfavored, or unaffected by gBGC). Specifically, we focused on four pairwise comparisons including (1) C>T at nonCpGs vs. C>A at nonCpGs; (2) C>T CpGs vs. C>A CpGs (3) C>G vs. T>A; (4) T>C vs. T>G. In principle, it is possible that the strength of gBGC is distinct for different types of variants, e.g., S vs. W variants (Tsai-Wu et al., 1992). However, in mice, roughly similar conversion rates are observed for C>A and C>T non-crossover gene conversion events as well as for T>C and T>G events (Li et al., 2019), lending support to using pairwise mutation ratios for controlling the effects of gBGC at least to a first approximation.

Three of the four pairwise comparisons involve mutation types with the same mutational opportunity (e.g., both T>C and T>G mutations involve changes at ancestral T bases in the genome), which further minimizes the confounding effects of regional variation on the chance of recurrent mutation or strength of background selection. Moreover, the pairwise ratios impose no co-dependency among mutation types as the four comparisons are mathematically independent of each other (although they may be biologically dependent if multiple ratios are affected simultaneously by some change in the mutational process).

Investigating the mutation spectrum using these four pairwise comparisons, we observe marked differences in the ratios both over evolutionary time and across populations. Specifically, we find multiple independent signals of mutation rate evolution, reflected by both temporal variation within a population and differences between YRI, CHB and CEU ($P < 0.01$ by Chi-square test after correcting for multiple hypothesis testing; Materials and Methods) (Figure 2A). These differences may represent broader geographic or population differences, as we replicate these findings in other population samples from the same continents— LWK, TSI, and JPT— from 1000 Genomes Project (Figure 2—figure supplement 1).

We performed multiple sanity checks to rule out any technical artifacts or sources unrelated to the mutation process in contributing to the observed inter-population differences. We obtain qualitatively similar results when restricting the analysis to neutral regions largely unaffected by linked selection (*B*-score > 800) (Figure 2—figure supplement 2) or comparing regions with high and low recombination rates (Figure 2—figure supplement 3), confirming that the use of pairwise comparisons effectively controls for the effects of selection and gBGC. In turn, to account for potential inaccuracies in mutation ages estimated by Relate, we stratified variants by allele frequencies instead of inferred mutation ages and replicated the signals detected in mutation age analysis (Figure 2—figure supplement 4). We again observe similar results with different binning

strategies for allele age, based on inferred mutation ages in YRI or CHB (Figure 2—figure supplement 5). Together, these results provide strong evidence that the human germline mutation spectrum has evolved over time and differs across populations.

Below we discuss the timing, direction, and population origin of the mutation rate changes related to each of the signals we detected in detail.

Elevation of nonCpG C>T/C>A ratio in Europeans

The largest signal that we observed is the transient elevation in the ratio of C>T/C>A mutations at nonCpG sites in CEU compared to the ancestral state before the OOA migration; in contrast, the nonCpG C>T/C>A ratios of CHB and YRI do not exhibit a similar shift at recent timescales. This signal encompasses the previously reported enrichment of C>T polymorphisms in a TCC context in Europeans, as well as other tri-nucleotide contexts (Harris and Pritchard, 2017; Mathieson and Reich, 2017; Speidel et al., 2019). Investigating the temporal patterns in CEU, we find that the increase in the ratio of C>T/C>A mutations at nonCpG sites becomes discernible starting from the time window spanning the OOA migration (50,000–100,000 years ago or ~2,000–4,000 generations ago) (Schiffels and Durbin, 2014), peaks around 238–887 generations ago, and subsides in the most recent age bin of 0–55 generations (Figure 2A). Because there is large uncertainty in inferred allele ages and our binning approach spreads the contribution of each variant over two or more age bins (Materials and Methods), the timeline and magnitude of variation should be interpreted cautiously: the transient change in nonCpG C>T mutations was likely shorter-lived and possibly of higher intensity than our results suggest. However, the temporal and geographic enrichment patterns from our analysis are consistent with previous reports based on low coverage 1KG or other datasets (Harris and Pritchard, 2017; Mathieson and Reich, 2017; Speidel et al., 2019).

Among all nonCpG trinucleotide contexts, the inter-population differences are most pronounced in the four previously reported tri-nucleotide contexts (TCC, TCT, CCC, and ACC; (Harris and Pritchard, 2017)), but are detectable in other nonCpG contexts as well (Figure 2B). Previous analysis found that these mutational contexts are enriched in two of the mutational signatures extracted from somatic mutations in tumor samples: the Catalog of Somatic Mutations in Cancer (COSMIC) SBS7 and SBS11 associated with exposures to ultraviolet light and alkylating agents, respectively (Alexandrov et al., 2020; Harris, 2015; Mathieson and Reich, 2017). To test if one of these two mutational signatures may be responsible for the observed differences in polymorphism data, we recalculated the C>T/C>A mutation ratio at nonCpG sites after excluding the sequence contexts most affected by SBS7 or SBS11 (Materials and Methods). While we observe some reduction in the magnitude of nonCpG C>T/C>A ratio in Europeans, the inter-population differences remain significant (Figure 2B). These results suggest the transient

change in nonCpG C>T/C>A ratio is not fully driven by the mutational mechanisms corresponding to either COSMIC SBS7 or SBS11. Thus, the etiology of this signal in Europeans remains unclear.

Divergence of C>G/T>A ratio among populations

The second largest inter-population difference is in the C>G/T>A ratio (Figure 2A): following the OOA migration among all three populations. In the past 3000 generations, both YRI and CEU samples show an increase in the C>G/T>A ratio albeit of different magnitudes, while in the CHB, the ratio initially decreases and then stays relatively stable for roughly 900 generations (Figure 2A). Interestingly, unlike the previous signal, inter-population differences in C>G/T>A remain highly significant for the most recent variants as well (0–55 generations), pointing to ongoing factors differentiating the relative rates of C>G and T>A mutations at present.

The fraction of C>G in *de novo* germline mutations is particularly sensitive to parental ages, increasing rapidly with the mother's age at conception (Jónsson et al., 2017). This raises the possibility that the inter-population differences in C>G/T>A ratio are driven by different average maternal reproductive ages among populations (Macià et al., 2021). To test this hypothesis, we leveraged the regional enrichment of maternal C>G mutations—"C>G enriched regions"—defined as 10% of the genome with the highest C>G SNP density that contributes to one-third of the overall maternal age effect (i.e., the yearly increase in maternal DNMs with mother's age) (Jónsson et al., 2017). The C>G/T>A ratio within the C>G enriched regions does not show significant inter-population differences (Figure 2C), possibly reflecting reduced power due to the much lower SNP counts in these regions (<15% of all). Outside of the C>G enriched regions, the three populations differ as much as they do genome-wide (Figure 2C), indicating that the differential accumulation of C>G mutations with maternal ages is not the primary driver of the differences observed across these populations.

To determine whether the signal in C>G/T>A ratio is driven by differences between populations in C>G or T>A mutation rate, we performed two additional comparisons (T>G/T>A and C>G/C>A), substituting numerator or denominator in the ratio by another mutation type. Unlike previous comparisons, these two comparisons are sensitive to the effects of gBGC, so the variation across time bins and populations cannot be readily interpreted as evidence for an evolution of the mutation spectrum. However, if the inter-population differences are in the same direction (i.e., rates in CHB<CEU<YRI), we can reason that the mutation type that is not being substituted (C>G or T>A) contributes to the inter-population differences. For T>G/T>A ratio, we still observe highly significant inter-population differences across the three populations, with CEU and YRI converging in recent time windows (Figure 2—figure supplement 6). Considering the C>G/C>A ratio,

we also find subtle but significant differences during the period of 55–437 generations ago (Figure 2—figure supplement 6). These results suggest that the inter-population differences in the C>G/T>A ratio arise from differences in mutation rates in both numerator and denominator, with CHB having the highest T>A and lowest C>G rates, and YRI having the lowest T>A and highest C>G rates.

Differences in the T>C/T>G ratios at deep timescales in human evolution

The T>C/T>G ratios are higher in the three oldest bins (dated to >28,800 generations ago) than in more recent ones in all three population samples and the effect is more pronounced in YRI compared to CHB and CEU (Figure 2A). The difference between old polymorphisms in different contemporary populations is puzzling, because the majority of these variants long predate the OOA migration (~2,000–4,000 generations ago), and thus must have arisen in the common ancestor of the three contemporary populations. We also observed a significant though small difference in T>C/T>G ratio between CHB and other populations at recent timescales (Figure 2A), but the signal is no longer significant after removing singletons (Figure 2—figure supplement 13).

We performed additional analysis to verify that these apparent inter-population differences are not driven by biases or inaccuracies in dating mutations. For each variant, Relate inferred an initial estimate of mutation age based on the entire 1000 Genomes data across populations and a refined estimate for each population by applying an iterative Markov Chain Monte Carlo (MCMC) algorithm to the data from that population. Depending on the sample size and demographic history of each population, the refined population-specific estimates could be differentially biased or associated with varying degrees of uncertainty, which can produce spurious differences across populations. We find systematic differences in the initial and refined mutation ages, but the population-specific estimates for the same variant found in two populations have overlapping age ranges for over 90% of the variants (Figure 2– figure supplement 7). Moreover, the inter-population differences appear to be primarily driven by non-shared variants across populations (p-value$<10^{-8}$ by Chi-square tests on variants in the three oldest bins), while the evidence is not significant among variants shared by all three populations (p-value> 0.01 by Chi-square tests; Figure 2– figure supplement 8).

Next, we examined if the elevated T>C/T>G ratio is related to increased T>C or reduced T>G mutations. Using alternative pairwise comparisons (T>C/T>A and T>A/T>G), we infer that the signal is primarily driven by higher proportions of T>C mutations among older variants (Figure 2—figure supplement 9). We then investigated the sequence context in which the T>C signal is enriched by applying non-negative matrix factorization (NMF), which has been extensively used in analysis of somatic and germline mutations to reveal combinations of mutation types caused by the same mutational process (i.e.,

mutational signatures) (Alexandrov et al., 2020; Mathieson and Reich, 2017; Seplyarskiy et al., 2021). Two of the three significant signatures identified by NMF (based on standard diagnostic criteria, Methods and Materials) show the same trends as the nonCpG C>T/C>A and T>C/T>G ratios we observe. The mutational signature corresponding to the T>C/T>G ratio is characterized by NTG>NCG and ATW>ACW mutations (Figure 2—figure supplement 10).

The enrichment of signal in NTG contexts raises the possibility that some of the "old" TpG>CpG mutations may actually be mis-polarized CpG>TpG mutations that are hyper-mutable and more likely to undergo recurrent mutations in the 1000 Genomes samples. Consistent with this hypothesis, T>C variants dated to the last few age bins (>28,000 generations) contain substantially greater fractions of TpG>CpG mutations relative to those in the younger bins (Figure 2—figure supplement 11). Notably, when variants at TpG sites were excluded, both the elevation in T>C/T>G ratio in old variants and the inter-population differences disappear (Figure 2D).

Previous analyses have inferred that the genome-wide rate of mis-polarization of the ancestral alleles is ~1.5–2% (Hernandez et al., 2007a). To minimize the effect of ancestral misidentification errors on the T>C/T>G ratio, we applied several standard approaches. We repeated the analysis with the ancestral allele inferred (i) using only the high-confidence sites in EPO alignment, and (ii) using the chimpanzee reference genome (panTro2) that is equally distant to all modern human populations for the inference (Materials and Methods). In both cases, we find qualitatively similar results as Figure 2, though the population differences are more significant in (ii) (Figure 2—figure supplement 12A-B). Given that the human reference genome is derived from multiple individuals and has European, African and East Asian ancestry in different regions, we were concerned about the potential impact of reference bias on our results (Green et al., 2010). Thus, we stratified the human reference genome by the inferred local ancestry and again, obtained qualitatively similar results in regions of European or African ancestry (Figure 2—figure supplement 12C). Together, our analyses suggest that potential mis-polarization errors at hypermutable CpG sites may have a non-negligible impact on the T>C/T>G signal in ancient variants, but these are challenging to properly correct for using standard approaches. Out of caution, we therefore excluded TpG sites from downstream analysis.

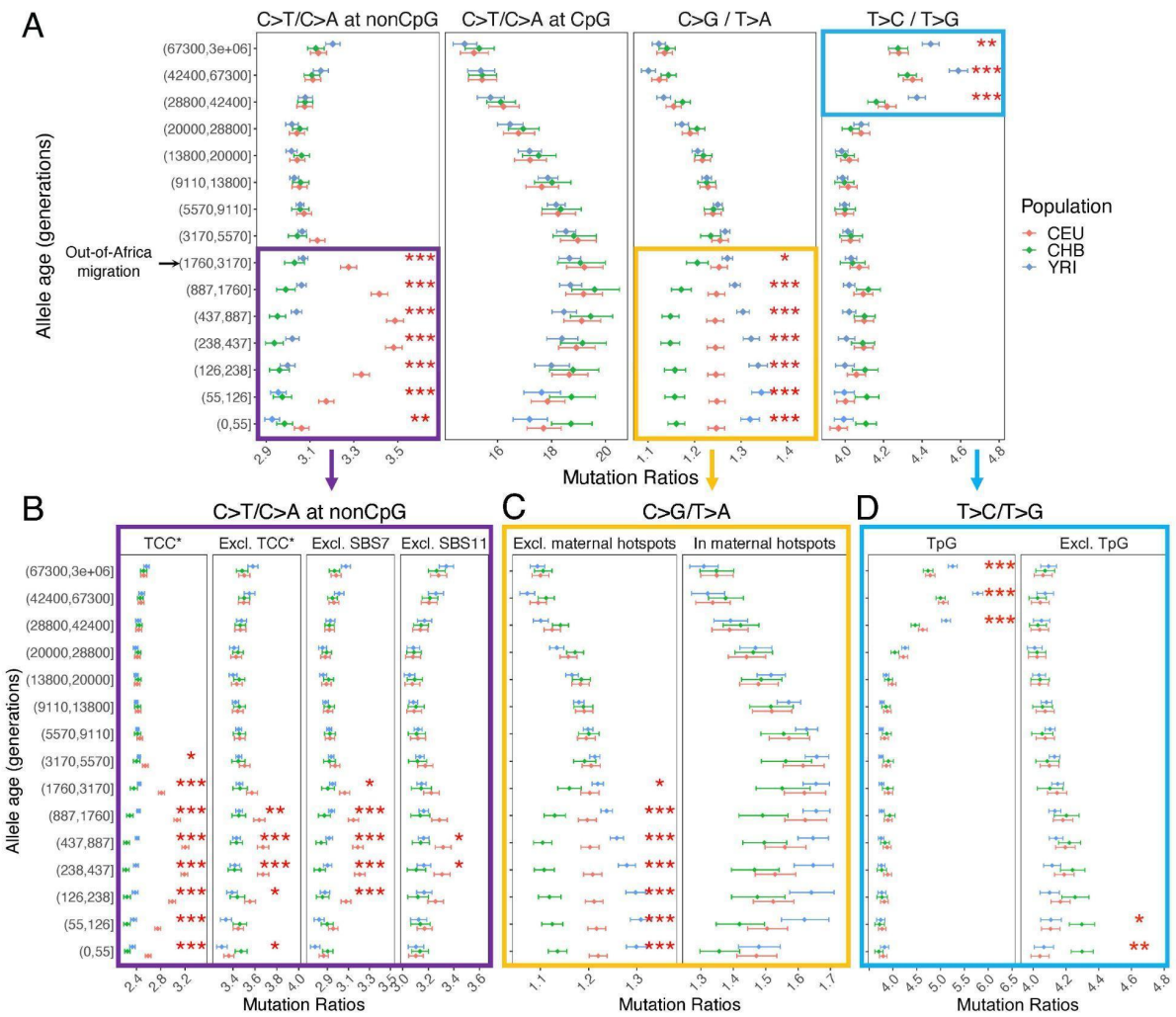


Figure 2. Comparison of pairwise mutation ratios for polymorphisms arising in different time windows. (A) Four pairwise mutation ratios are shown, each of which compares two mutation types that are matched for mutational opportunity and effects of gBGC. The black arrow indicates the window coinciding with the OOA migration. Highlighted in boxes are three ratios that show significant inter-population differences, with in-depth investigation into each shown in lower panels. Asterisks refer to the p-value obtained from a Chi-square test after a Bonferroni correction for 60 tests: * for $P < 0.01$, ** for $P<0.0001$ and *** for $P<10^{-8}$. (B) Elevation in C>T/C>A ratio in CEU at nonCpG sites, after excluding the four trinucleotide contexts (TCC, TCT, CCC, and ACC) previously identified to be associated with the TCC pulse in Europeans (denoted by TCC*; Harris and Pritchard, 2017), as well as contexts affected by COSMIC mutational signatures of SBS7 and SBS11 (Harris 2015; Mathieson and Reich 2017); (C) Post-OOA divergence in C>G/T>A ratio among three population groups; and (D) Higher T>C/T>G ratio in YRI than CEU and CHB samples among extremely old variants, driven by TpG variants.

The following figure supplements are available for Figure 2:

465 **Source data 1.** Text file with (pseudo-)counts of mutations classified into eight types in
466 genomic regions including, excluding, and within the maternal C>G mutation hotspots,
467 in YRI, LWK, CEU, TSI, CHB, JPT in each time window.
468 **Figure supplement 1**. Pairwise polymorphism ratios in three additional populations from
469 Africa (LWK), Europe (TSI), and East Asia (JPT) in the 1000 Genomes project.
470 **Figure supplement 2**. Pairwise polymorphism ratios in genomic regions with a minimal
471 effect of background selection (B-score >800).
472 **Figure supplement 3**. Pairwise polymorphism ratios in 33% of the genome with the
473 lowest (A) and highest (B) regional recombination rates.
474 **Figure supplement 4.** Pairwise ratios of human polymorphisms stratified by allele
475 frequency.
476 **Figure supplement 5**. Pairwise polymorphism ratios in YRI, CEU, and CHB in commonly
477 accessible regions based on alternative binning strategies.
478 **Figure supplement 6.** Alternative pairwise polymorphism ratios (T>G/T>A and
479 C>G/C>A), used to investigate the cause of inter-population differences in C>G/T>A ratio.
480 **Figure supplement 7.** Comparison of allele age estimates inferred by Relate.
481 **Figure supplement 8.** The T>C/T>G signal in variants shared across continental groups
482 (A) and non-shared variants (B).
483 **Figure supplement 9**. Alternative pairwise polymorphism ratios (T>C/T>A and T>A/T>G)
484 to investigate the cause of inter-population differences in T>C/T>G ratio.
485 **Figure supplement 10.** Sequence context specificity of the T>C/T>G signal based on
486 mutational signature extraction by non-negative matrix factorization.
487 **Figure supplement 11.** Enrichment of TpG>CpG mutations in all T>C mutations in
488 ancient time bins.
489 **Figure supplement 12.** Persistence of the T>C/T>G signal with alternative methods for
490 derived allele polarization and stratification of the reference genome by local ancestry.
491 **Figure supplement 13.** Comparison of pairwise mutation ratios for polymorphisms
492 arising in different time windows after excluding singletons.

493 **Parental age effects on the mutation spectrum**

494 To explore whether the inter-population differences in polymorphism data could be driven
495 by changing mean generation times over evolution (Macià et al., 2021), we turned to
496 genomic data from present-day pedigrees and quantified the parental age effects on the
497 pairwise ratios of *de novo* mutations (DNMs) (Materials and Methods). To maximize the
498 power and precision, we focused on the largest published DNM dataset, which includes
499 200,435 DNMs from 2,976 Icelandic trios (Halldorsson et al., 2019). The inferred parental
500 age effects based on a previous, smaller DNM dataset were qualitatively similar (Figure
501 3—figure supplement 1), despite some significant differences across datasets, possibly
502 due to systematic differences in the criteria for identifying and filtering DNMs (Figure 3—
503 figure supplement 2).

Considering the four pairwise mutation ratios, which are mathematically independent, three show a significant dependence on parental age (Figure 3). As an illustration, if both parents reproduce at 40 years rather than at 20 years of age, the ratios of nonCpG C>T/C>A and nonTpG T>C/T>G decrease by 9.4% (90% confidence interval (CI): 4.3%-14.6%) and 7.5% (CI: 1.0%-13.8%) respectively, whereas the C>G/T>A ratio increases by 11.9% (90% CI: 4.1%-20.0%). In terms of sex-specific effects (Figure 3—figure supplement 3), nonCpG C>T/C>A and nonTpG T>C/T>G ratios are largely determined by the paternal age and much less so the maternal age, reflecting that the paternal age effect is 3–4-fold stronger than maternal age effect for these mutation types (Goldmann et al., 2018; Jónsson et al., 2017; Kong et al., 2012). For C>G/T>A ratio, however, the maternal age is nearly as important as the paternal age, consistent with the unusually strong maternal age effect on C>G mutations (Jónsson et al., 2017).

We were unable to directly quantify the dependence of CpG C>T/C>A ratio on parental ages, because the low count of C>A mutations at CpGs (on average 0.55 DNMs per trio) limits our ability to reliably infer the parental age effects (Materials and Methods). However, a previous study noted that the fraction of CpG C>T mutations among all DNMs depends strongly on parental age and decreases by 0.26% per year (Jónsson et al., 2017). Consistent with this finding, the ratio of the counts of CpG C>T and CpG C>A DNMs differs significantly: the 20%-tile of trios with the youngest parents have a significantly higher ratio than in the 20%-tile of trios with the oldest parents (21.0 vs. 17.4, P=0.03 by Chi-square test). This difference suggests the ratio of C>T to C>A mutations at CpG sites decreases with parental age. Overall, the significant age-dependency of three, and likely all, of the four pairwise mutation ratios highlights the pervasive influence of reproductive ages on the human germline mutation spectrum.

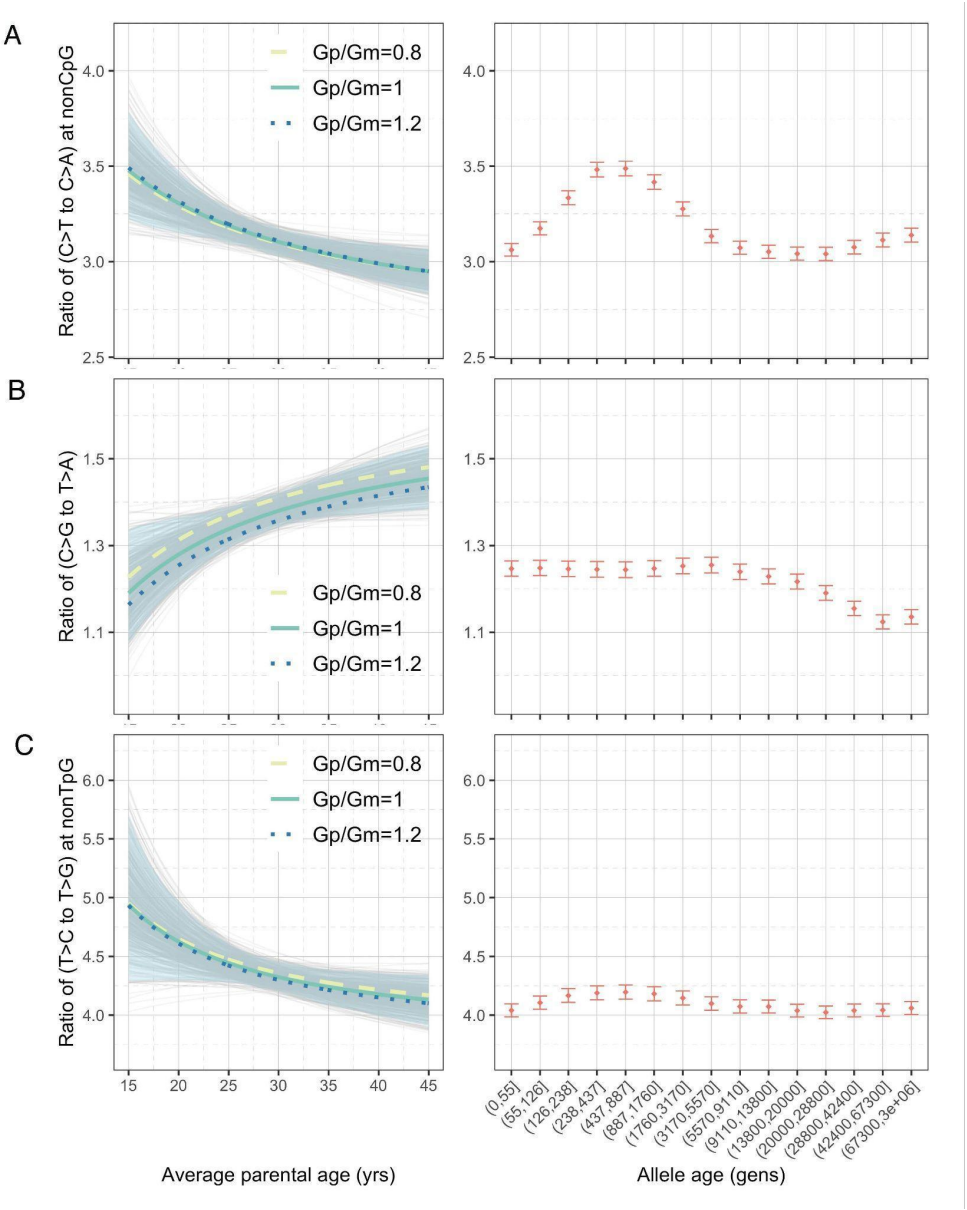


Figure 3 Effects of parental ages on three pairwise mutation ratios estimated from DNM data in 2,879 Icelandic trios (Halldorsson et al. 2019). The different curves reflect expectations for different ratios of paternal (Gp) to maternal (Gm) mean generation times. Each light gray curve represents the expected ratio for Gp/Gm=1 from one bootstrap resampling replicate (see Materials and Methods), with the lighter blue area denoting 90% confidence interval (CI) assessed from 500 replicates. For ease of comparison, ratios for polymorphisms of different ages identified in CEU are shown on the right of each panel. The points represent the observed polymorphism ratios, while the whiskers denote the 95% CI assuming a binomial distribution of polymorphism counts. All results are based on variants in the commonly accessible regions, excluding those with derived alleles observed in either Neanderthal or Denisovan.

The following source data and figure supplements are available for Figure 3:
Source data 1: Mutation parameters inferred from DNM data in 2,879 Icelandic trios with estimated uncertainty based on bootstrap resampling (one file for each mutation type for commonly accessible regions; *n*=500 replicates).
Figure supplement 1. Effects of parental ages on three pairwise mutation ratios estimated from an earlier DNM dataset (Jónsson et al., 2017).
Figure supplement 2. Discrepancies in the mutation spectrum between different DNM datasets and between DNMs and young polymorphisms.
Figure supplement 3. Sex-specific parental age effects on three pairwise mutation ratios.
Figure supplement 4. Power for detecting inter-population differences in C>G/T>A ratio driven by differences in generation time.

Shifts in generation times needed to explain the observed changes in the polymorphism data

Motivated by the strong dependency of DNM ratios on parental ages, we tested the hypothesis that changes in past generation times account for the observed mutation spectrum of polymorphism data, as suggested by a couple of recent studies (Macià et al., 2021; Wang et al., 2023). In particular, we asked whether the temporal shifts in the pairwise polymorphism ratios could be fully explained by shifts in average reproductive ages, i.e., without the need to invoke additional factors. As the mutation process may have evolved possibly separately in different human populations, we focused the comparison of DNMs to variants identified in CEU, who are most genetically similar to the Icelandic individuals (with $F_{ST} < 0.005$) for whom we have the largest DNM dataset (Halldorsson et al., 2019).

Assuming the observed changes in the mutation spectrum are solely driven by shifts in average reproductive ages, we inferred past generation times by relating observed pairwise ratios in DNM data and polymorphism data. Specifically, for a given pairwise mutation ratio in the polymorphism data, we asked what value of the generation time is compatible with the relationships to age estimated from pedigree data (assuming identical male to female mean generation times and a fixed onset of puberty). Accounting for uncertainty in the DNM data, we then inferred the 95% confidence interval of the generation time for each mutation ratio and time window. Given the complications with low numbers of CpG mutations in DNMs data and of recurrent mutations at CpG sites in polymorphism data, we excluded the pairwise ratio of C>T/C>A at CpGs for this analysis.

We inferred the generation times across mutation ages in CEU by solving for the parental ages that would give rise to the observed pairwise ratios in each time window. Surprisingly, the estimates are inconsistent for different mutation ratios within a single

583 mutation age bin. Moreover, the overall trends are inconsistent across time windows. For
584 instance, the steady increase in C>G/T>A ratio over time translates into a gradual
585 increase in reproductive age, with the ratio of the most recent bin corresponding to a
586 reproductive age under 23 years (Figure 3C; Figure 4). In contrast, the nonTpG T>C/T>G
587 polymorphism ratio appears to be lower than the ratio in *de novo* mutations across the
588 range of typical parental ages in pedigree studies and suggests a generation time of more
589 than 40 years (Figure 3C). Such a long population-average generation time is not only
590 inconsistent with the estimate of the C>G/T>A ratio, but it is also unrealistic for human
591 evolution (Fenner, 2005; Moorjani et al., 2016b). Finally, the transient elevation in
592 nonCpG C>T/C>A ratio suggests a drastic, rapid reduction in the generation time (Figure
593 3A; Figure 4). Specifically, the ratios of both ancient (>5,670 generations ago) and the
594 most recent polymorphisms (<55 generations ago) correspond to relatively old
595 reproductive ages of ~35 years, while the peak at around 238–887 generations ago
596 provides average reproductive age estimates of less than 20 years (Figure 4).
597
598 The incompatible patterns across different mutation ratios can potentially arise if male
599 and female generation times differed in the past. To explore this possibility, we varied the
600 ratio of male and female generation times between 0.8–1.2, as seen across a range of
601 contemporary human populations ((Fenner, 2005); Figure 4—figure supplement 1; Figure
602 4—figure supplement 2). We further allowed male and female generation times to vary
603 freely and inferred the combinations of paternal and maternal ages that could give rise to
604 the observed polymorphism ratios. Even after modeling sex-specific reproductive ages,
605 we find inconsistent generation time estimates and incompatible trends across mutation
606 ratios (Figure 4—figure supplement 3).
607
608 In addition to the temporal changes within CEU, the observed patterns in mutation spectra
609 across populations cannot be explained by differences in generation times. In particular,
610 given the extraordinarily strong maternal age effect on C>G mutations in maternal
611 hotspots (Jonsson et al 2017), we would expect more pronounced differences in the
612 C>G/T>A ratio in these regions than in the rest of the genome, should the mutation
613 patterns be driven by generation time. Simulations based on the parental age effects
614 observed in pedigree data suggest that the power to detect differences in C>G/T>A ratio
615 due to change in generation times is much greater within the maternal C>G mutation
616 hotspots than in the rest of the genome, despite the much smaller number SNPs in the
617 hotspots (Figure 3 – figure supplement 4). Therefore, if the inter-population differences in
618 C>G/T>A ratio were solely driven by differences in maternal age across populations, we
619 should be well-powered to observe a signal in the maternal hotspot regions alone. Not
620 seeing one (Figure 2C) provides another line of evidence against the premise that
621 generation time is a major driver of the observed mutation patterns across time or
622 between populations in humans.

Together, our results suggest that shifts within a plausible range of human generation times–i.e., those that fall between the ages of puberty and reproductive cessation in contemporary humans–cannot explain the observed variation in the polymorphism data for CEU, and by extension are unlikely to explain the mutation ratios in CHB or YRI polymorphisms.

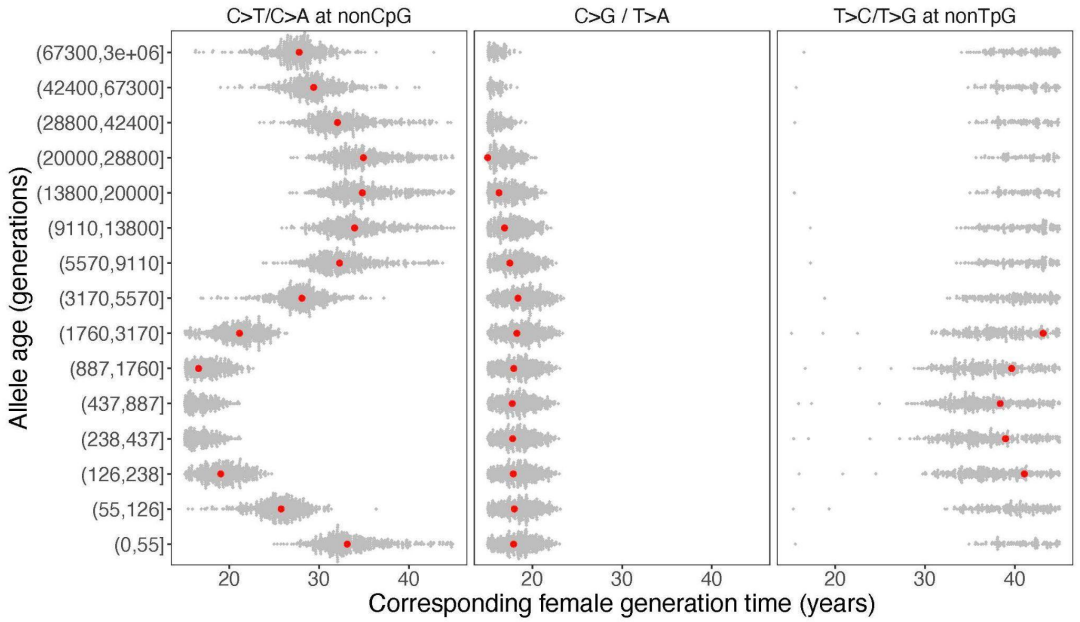


Figure 4 Past generation times corresponding to the observed polymorphism ratios in CEU, given parental age effects estimated from DNM data. Red points represent the point estimates based on maximum likelihood estimators of mutation parameters from the DNM data; gray dots show estimates from 500 bootstrap replicates by resampling trios with replacement. We assumed the same male to female generation times (Gp=Gm) for all time windows. Similar trends were obtained for other fixed values of Gp/Gm (between 0.8-1.2) or independently varying Gp and Gm (Figure 4—figure supplement 1; Figure 4—figure supplement 3)

The following source data and figure supplements are available for figure 4:

Source data 1: Past generation times inferred from each polymorphism ratio, assuming fixed ratios of male to female generation times (Gp/Gm=0.8, 1, 1.1, 1.2), with confidence intervals estimated using bootstrap resampling (*n*=500 replicates; one file for each mutation type).

Figure supplement 1. Past generation times inferred from the observed pairwise polymorphism ratios, assuming a fixed ratio of male to female generation times of (A) Gp/Gm=0.8 , or (B) Gp/Gm=1.1, or (C) Gp/Gm=1.2 .

Figure supplement 2. Past generation times that correspond to the observed pairwise polymorphism ratios, based on parental age effects estimated from an earlier DNM dataset (Jónsson et al., 2017).

Figure supplement 3. Combinations of paternal and maternal reproductive ages corresponding to the observed pairwise polymorphism ratios.

Discussion

Multiple changes in the germline mutation spectrum during the course of human evolution

We introduce a new framework to compare the mutation spectrum over time and across population samples, while controlling for the effects of selection and biased gene conversion. By applying this approach to multiple population samples from the 1000 Genomes Project dataset, we observe multiple independent signals of inter-population differences. Notably, we replicate the transient elevation in nonCpG C>T mutation (manifest in the C>T/C>A ratio) identified previously in Europeans compared to East Asians and Africans (DeWitt et al., 2021; Harris, 2015; Harris and Pritchard, 2017; Mathieson and Reich, 2017; Speidel et al., 2019). We find that this ratio also differs subtly between YRI and CHB, suggesting an additional change occurred in this mutation type. In both cases, the signal is enriched in the TCC, TCT, CCC, and ACC contexts (Figure 2 — figure supplement 11), and mutation types that are associated with exposure to UV and alkylating agents (Harris, 2015; Mathieson and Reich, 2017). While these contexts may contribute to this signal, we show that they do not fully explain the observed differences between contemporary populations. Thus, the etiology of this signal remains obscure, and it may not be specific to groups with west Eurasian ancestry. Further investigation into the extended sequence contexts of this mutation pulse may help elucidate the underlying molecular mechanism(s) (Aggarwala and Voight, 2016; Aikens et al., 2019).

We also observe two additional inter-population differences in the mutation spectrum. First, the ratio of C>G/T>A mutation rates differs between YRI, CEU and CHB. We find that this signal is related to an increase in T>A mutations and depletion of C>G mutations in CHB compared to CEU, as well as a depletion of T>A mutations in YRI. Some aspects of this observation (e.g., the enrichment of T>A mutations in East Asians) were previously noted (Harris and Pritchard, 2017), but our analysis adds information about the changes in additional populations and insights into the timing of this change. Given the distinct trends of T>A and C>G mutations with allele age (Figure 2—figure supplement 6), it appears that at least two changes in the mutational processes are needed to explain the inter-population differences. By comparing the ratios in inside and outside C>G hotspots, we confidently rule out a primary role of maternal generation times in driving these

differences. Interestingly, these differences are still observed in the most recent polymorphisms, indicating that—unlike the TCC mutation pulse—this process is likely ongoing. This finding therefore points to an opportunity to directly examine and map the underlying biological causes using large-scale *de novo* mutation datasets from a diverse set of populations.

We also detect a shift in the ratio of T>C/T>G mutations at old timescales (>28,000 generations) and differences among human populations. This signal is driven by T>C rather than T>G and enriched in TpG contexts; in fact, the signal disappears after excluding TpG sites. We hypothesize that some of the inferred ancient TpG>CpG mutations may be mis-polarized CpG>TpG mutations that are hyper-mutable and more likely to undergo recurrent mutations over the course of human evolution. While there can be evolution in T>C mutation rate at old timescales, controlling for the effects of mis-polarization errors is challenging and hence this signal remains tentative.

Effects of ancestral mis-polarization in the study of mutation spectrum

Misidentification of the ancestral alleles has been found to be more likely at sites with higher mutation rates and shown to substantially impact studies of natural selection and gene conversion (Hernandez et al., 2007a, 2007b). Accounting for ancestral misidentification is also critical in studies investigating variation in mutation rate and spectrum. Our analysis shows that current methods—using multi-species alignment or tree-based polarization approaches—still suffer from several limitations. We discuss some of these in turn.

For most of the analyses, we used the ancestral allele inferred by 6 primate EPO alignment. The EPO alignment relies on a model of continuous time DNA nucleotide substitutions. By default, it uses the HKY model, with the ratio of transitions to transversions set to 2, and a stationary GC frequency set to 40% (Paten et al., 2008). It does not take into account context-dependent, fine-scale variation in mutation rate, however, which can lead to higher error rate in assigning ancestral alleles for the hyper-mutable CpG sites (Harpak et al., 2016). Moreover, it considers only a single reference genome in each species; by not taking into account information about allele frequency of variants within a species, it provides less accurate inference of ancestral alleles (Hernandez et al., 2007a). As an alternative to the use of EPO, we used the chimpanzee reference genome to infer the ancestral state and obtained qualitatively similar results, likely because this approach is sensitive to similar biases, in particular at fast evolving sites.

In turn, the program Relate requires ancestral alleles as input and aims to identify mis-classified ancestral alleles based on whether the derived allele maps to a unique branch

of the gene genealogy. It assumes an infinite-site mutation model according to which all haplotypes carrying the derived allele should be descendents of a unique branch; if this is true for the ancestral allele but not the derived one, Relate 'flips' their assignments (the mutation is left "unmapped" if neither derived nor ancestral allele can be mapped to a unique branch). In this approach, mis-polarization of the ancestral allele still remains an issue, especially for sites that experience recurrent mutations in the sample. Because Relate assumes the mutation occurred exactly once in the genealogical history of the sample, the placement of mutations is bound to be inaccurate for repeat mutations that need to be mapped to multiple branches of the tree. This issue disproportionately impacts CpG transitions, as exemplified by the fact that CpG>TpG mutations are under-represented in the 'mapped' subset of mutations inferred by Relate (Speidel et al., 2019). These unmapped mutations are unevenly distributed across allele frequency and mutation age, likely in a demography-dependent manner, making it hard to predict or correct for this effect. Multiple independent CpG>TpG mutations at the same locus on different branches of the gene genealogy may be mis-classified as a single old TpG>CpG mutation (and the allele age inferred would also be inaccurate). Beyond CpGs, these effects are likely to impact other highly mutable sites (e.g., certain types of transitions) and may spuriously appear as signals of changes in the mutation spectrum.

Changes in generation times cannot explain the evolution of the mutation spectrum in humans

Across mammals, generation time is the strongest predictor of the yearly mutation rate and of some aspects of the mutation spectrum (Hwang and Green, 2004; Moorjani et al., 2016a; Wu and Li, 1985). Accordingly, our analysis of *de novo* mutations from pedigree studies shows significant effects of parental ages on all four pairwise mutation ratios that we examined. Recent studies have argued that changes in generation times can explain a large fraction of the differences in mutation spectrum observed across human populations (Macia et al 2021), and one study even leveraged the population-specific mutation spectrum to infer the historical generation time in humans (Wang et al., 2023).

It may seem like a promising idea to infer generation time based on changes in mutation spectrum, but in practice, several technical hurdles stand in the way. First, given the sampling noise associated with the limited number of *de novo* mutations per family, large numbers of pedigrees are required to characterize the parental age effects reliably, especially for very specific mutation types. In our analysis, we used the largest available pedigree dataset, but the parental age effects remain imprecisely estimated and should be revisited as larger datasets become available, ideally from a diverse set of populations. Second, technical issues, both molecular and computational, may affect the reliability of variant calls of different mutation types (Bergeron et al., 2021). Indeed, we find that the four pairwise ratios differ significantly across two recent pedigree datasets, as well as

between *de novo* mutations and young polymorphisms in 1000 Genomes dataset, which are unlikely to be reconciled by biological reasons (Figure 3—figure supplement 2). In addition, controlling for the effects of biased gene conversion is difficult, as its effects may differ to some extent by mutation type even within a class (for instance, T>C and T>G both classified as S>W mutations may be subject to differing strengths of BGC).

Importantly, fitting a generation time to observed differences in mutation spectrum relies on the assumption that changes in the generation time play the sole (or at least the predominant) role in the evolution of the mutation spectrum, an assumption that does not seem to hold in data. Specifically, we find that the generation time estimates inferred from different mutation ratios independently disagree with each other, within the same time window or over time. Notably, the temporal trends, which should be robust to most technical issues mentioned above, inferred from different mutation ratios are mutually inconsistent. In other words, changes in a single parameter—generation time—cannot explain the mutation patterns in humans (Figure 4). These inconsistencies persist after accounting for uncertainty in the parental age effects inferred from pedigrees and incorporating sex-specific reproductive ages. Moreover, the significant divergence in C>G/T>A ratio outside the maternal C>G hotspots but not within argues against parental age (in particular, maternal age) as a major driver of mutation spectrum differences across populations (Figure 3 – figure supplement 4). These findings thus establish that changes in generation time alone cannot account for all or even most of the observed variation in mutation spectrum over the course of human evolution.

Implications

The mutation spectrum of polymorphisms is a convolution of multiple evolutionary forces: mutation, recombination (including gene conversion), natural selection, and their interplay with demography. In this study, we investigated the contribution of these forces to differences in the mutation spectrum across contemporary human populations. For future studies aiming to understand the evolution of mutagenesis based on analyses of polymorphism patterns, it will be crucial to consider more realistic mutation models (including using context-dependent models) for ancestral allele reconstruction and model the impact of non-mutational evolutionary forces.

Our analysis demonstrates the limitations of inferring past generation times based on polymorphism patterns. We find that shifts in generation time alone cannot explain the observed variation in the mutation spectrum, leaving a non-negligible role for other factors—such as transient environmental exposures, genetic modifiers and other life history traits (e.g., changes in the onset of puberty)—in shaping the mutation landscape in human populations. This conclusion is in line with recent studies in model organisms that discovered naturally occurring genetic modifiers (Jiang et al., 2021; Sasani et al.,

2022) as well as a human pedigree study that identified individuals with germline hypermutation potentially due to genetic modifiers or exposures to chemotherapeutic agents (Kaplanis et al., 2021).

Although the three population samples that we focused on here were collected from three distinct continents, the observed differences among them are not necessarily generalizable to continental level. In particular, within the same continent, and notably in Africa, there is relatively deep genetic divergence between some populations (e.g., between Bantu groups and Khoe-San), often accompanied by long-term geographic isolation and environmental differences (Mallick et al., 2016). These different histories can lead to considerable variation in the mutation processes within a continent. In fact, even for closely related populations, we detect subtle but significant differences in the polymorphism ratios (e.g., between CEU and TSI in Figure 2—figure supplement 1). Genetic data from more diverse populations, in terms of both ancestry and geographic location, are needed to generate a more complete picture of past and ongoing variation in the mutation spectrum across human ancestries and to understand its evolution.

The variation in the mutation spectrum over the course of human evolution raises a fundamental puzzle about why the molecular clock works over long timescales and across species. Our analyses uncovered substantial variation in multiple pairwise mutation ratios at different time depths during human evolution. Since variation in each pairwise ratio suggests mutation rate variation for at least one (or both) of the mutation types involved, our findings suggest that the absolute mutation rate per year of several mutation types must have been evolving. For example, our result suggests the mutation rate for C>T mutations at nonCpG sites varied by ~15–20% in CEU over the past 3,000 generations. Over longer timescales, it is likely that all mutation types deviate from a strictly clock-like behavior. It is puzzling then that the mutation rates across species are strikingly similar over millions of years: for instance, the substitution rates differ by less than 10% percent for any mutation type across the human and chimpanzee lineages (Moorjani et al., 2016a). This observation suggests that although the mutation rate and spectrum can evolve over relatively short timescales, the fluctuations in yearly mutation rate often average out over longer timescales, possibly reflecting the effects of long-term stabilizing selection.

Materials and Methods

Data availability

All datasets analyzed here are publicly available at the following websites:

Dataset	Source	Reference
High coverage 1000 Genomes Project	https://www.internationalgenome.org/data-portal/data-collection/30x-grch38	(Byrska-Bishop et al., 2022)
Decode *de novo* mutations 2019 (2,976 families)	https://www.science.org/doi/10.1126/science.aau1043	(Halldorsson et al., 2019)
Decode *de novo* mutations 2017 (1,548 families)	https://www.nature.com/articles/nature24018#additional-information	(Jónsson et al., 2017)

Data filtering and partitions used in the analysis

Commonly accessible regions: In order to reliably compare mutation patterns across datasets, we generated a list of genomic regions that were "accessible" or assayed by the study after accounting for the constraints of the study design. To generate this list, we first followed the variant calling procedure described in Jonsson et al. 2017 to identify the accessible genome for de novo studies. This yielded an accessible genome length of 2.7 Gb similar to the estimate reported in the original study (Jónsson et al., 2017). We intersected this dataset with the 1000 Genomes Strict Mask (see Resources below). We used the strict mask generated using low coverage 1000 Genomes dataset as it encompasses a larger set of low complexity regions and thus may port well across datasets. Further, to focus on putatively neutral regions, we removed exons and phylogenetically conserved regions sources). The combined set of accessible autosomal regions contained 2.15 Gb. Unless otherwise stated, we present all results generated for this subset of the autosomal genome, which we refer to as the "commonly accessible" regions.

Regions of high and low recombination rate: To study the impact of recombination rate on mutation patterns, we divided the genome into bins sorted by recombination rate using the HapMap recombination map (1000 Genomes Project Consortium et al., 2015). We then sorted all genomic sites by recombination rate and divided the genome into three

discrete bins with recombination rates of (0, 0.0717], (0.0717, 0.422], and (0.422, ∞) cM/Mb, each containing roughly 33% of the genomic bases.

B-statistic or B-scores: To focus on regions of the genome that are minimally affected by linked selection, we assigned a *B*-score to each variant site in a population. The *B*-score measures the expected reduction in diversity levels at a site due to background selection, with smaller values implying greater effects of background selection. We used the *B*-score values provided by (McVicker et al., 2009). We then compared the mutation patterns within windows of different values of B-scores. Additionally, where specified, we used the list of effectively neutral regions that contains the commonly accessible regions with a B-score > 800. This subset of the genome includes 1.33 Gb.

Human ancestral allele reconstruction: For most analyses, we used the ancestral allele reconstruction based on the 6 primate EPO alignment (1000 Genomes Project Consortium et al., 2015). In the EPO (Enredo-Pecan-Ortheus) pipeline, Ortheus infers ancestral states from the Pecan alignments (Paten et al., 2008). The confidence in the ancestral call is determined by comparing the call to the ancestor of the ancestral sequence as well as the 'sister' sequence of the query species. High confidence sites are annotated with capital letters in the alignment. For some analysis, we also used the chimpanzee (panTro2) reference genome to infer the ancestral allele (mapped to human reference genome coordinates, *hg19*).

Resources

Dataset	**Source**
1000 Genomes Strict Mask	https://www.internationalgenome.org/announcements/genome-accessibility-masks/
Recombination rate map	https://www.science.org/doi/10.1126/science.aau1043
B-scores	https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000471
Conserved regions	http://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/phastConsElements46wayPrimates.txt.gz

Coding regions	http://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/refGene.txt.gz
Human ancestral genome	https://ftp.ensembl.org/pub/release-86/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh38_e86.tar.gz
Chimpanzee reference genome (panTro2) in hg19 coordinates	https://reichdata.hms.harvard.edu/pub/datasets/sgdp/
COSMIC signatures (v3.2, GRCh38)	https://cancer.sanger.ac.uk/signatures/documents/453/COSMIC_v3.2_SBS_GRCh38.txt

Relate analysis

We applied Relate v1.1.5 (Speidel et al., 2019) to phased whole genome sequences from the 1000 Genomes Project (see Datasets). The 1000 Genomes data was phased and imputed statistically (Byrska-Bishop et al., 2022). As a result most singletons, which are missing in the phased data, were not included in Relate analysis. We focused on biallelic SNPs only using VCFtools (--remove-indels --min-alleles 2 --max-alleles 2) (Danecek et al., 2011). We then converted VCFs to haps/sample format using RelateFileFormats (--mode ConvertFromVcf) and prepared the input files using PrepareInputFiles.sh provided by Relate. We used 1000 Genomes Pilot Mask as the genome accessibility filter and polarized each allele to ancestral or derived state using the 6 primate EPO alignment. We assumed a mutation rate (*m*) of 1.25 x 10^{-8} per base pair per generation and an effective population size (*N*) of 30,000 (Jónsson et al., 2017). We used the Hapmap II genetic map (1000 Genomes Project Consortium et al., 2015). We first inferred the mutation ages using the entire dataset. Following Relate's guidelines, we inferred a refined estimate for each population by splitting the Relate output genealogies into subtrees using RelateExtract (--mode SubTreesForSubpopulation) and re-estimated the branch lengths (using EstimatePopulationSize.sh) to obtain the final mutation ages and the associated uncertainty (upper and lower bounds that reflect the start and end points of the branch that the mutation falls on in the reconstructed genealogical tree). For each mutation, we then inferred the upstream and downstream base pair using the 6 primate EPO alignment. We excluded sites where either the upstream or downstream base in the human ancestral genome was missing or ambiguous. Unless otherwise stated, we present the results for the commonly accessible regions.

Using chimpanzee reference genome to infer the ancestral allele: To test the robustness of our results to the inferred ancestral allele, we reran the Relate analysis with the chimpanzee (panTro2) reference genome mapped to the human reference genome (hg19). Because we only had access to the chimpanzee genome in hg19 coordinates, we first lifted over the vcf files of 1000 Genomes high coverage data from hg38 to hg19 coordinates using *CrossMap (Zhao et al., 2014)*. We then polarized the ancestral and derived alleles with respect to the allele in the chimpanzee reference genome and used the Hapmap II genetic map in hg19 coordinates (1000 Genomes Project Consortium et al., 2015). The rest of the parameters and setup in Relate was identical as previously described.

We note that for all the analyses reported in the manuscript, we focused on SNPs for which Relate successfully inferred the allele ages. Thus, unphased, multiallelic and unmapped SNPs were excluded. This likely introduces some biases in comparison of mutation spectrum especially for young variants, so the Relate results of very recent bins should be interpreted with caution.

Classification of shared and non-shared variants across continental groups

Considering the six populations under study (YRI, LWK, CEU, TSI, CHB, JPT), we operationally defined “shared variants” as SNPs in Relate outputs that have both alleles observed in samples from at least one population from each of the three continental groups, i.e., variants segregating in (YRI or LWK) and (CEU or TSI) and (CHB or JPT). Conversely, SNPs that do not meet the above criteria were classified as “non-shared variants”, although many of them may actually be shared by populations from two (but not three) of the continental groups.

Binning of polymorphisms based on mutation age

Among SNPs with Relate-inferred allele ages, we filtered out those with extremely old ages (i.e., upper bounds of allele age greater than 3,000,000 generations), as those were too old to be compatible with reasonable human evolutionary history. There is large uncertainty in the mutation ages estimated by Relate, such that the estimated lower and upper bounds often differ by an order(s) of magnitude or more. We took a two-step approach to bin the polymorphisms by age, accounting for this uncertainty. First, we determined the boundaries of age bins by sorting all SNPs segregating in CEU into 15 bins of roughly equal sizes with a Monte Carlo method (i.e., randomly selecting a point estimate by sampling a point uniformly between the upper and lower bounds of inferred allele age by Relate). We then calculated the *pseudo-counts* of each mutation type in each bin by summing up the probability densities across all variants, assuming a uniform distribution of each variant within the inferred age intervals. For example, if a T>A SNP

has an estimated age range of (500, 1300) generations, which overlaps with three of the predetermined age bins (312, 545], (545, 1160], and (1160, 2970], we would assign the T>A SNP to three bins with the following weights (545-500)/(1300-500)=0.056,(1160-545)/(1300-800)=0.769, and (1300-1160)/(1300-800)=0.175, respectively. We note that since the allele age distribution differs across populations due to differences in their demographic history, there is no way to bin variants equally for all populations simultaneously. For results shown in main figures, we based our binning into equal sizes on the age estimates of variants observed in CEU. Results were qualitatively similar when the bin boundaries were determined based on variants observed in YRI and CHB samples (Figure 1—figure supplement 8; Figure 2—figure supplement 5).

Calculating confidence intervals of polymorphism ratios and the statistical significance of inter-population differences

To assess the confidence intervals (CIs) of the mutation ratios in polymorphism data, we assumed the pseudo-counts of the two mutation types being compared follow a binomial distribution conditional on the total count. In practice, we used the normal approximation for calculating the 95% CI of the proportion for a given mutation type based on the observed counts of two types, using $\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1+n_2}}$, where n_1 and n_2 are the pseudo-counts of two mutation types, $\hat{p} = \frac{n_1}{n_1+n_2}$ is the point estimate of the probability of success, and *z*=1.96 is the *Z*-score corresponding to the upper 2.5%-tile. We then transformed the CI of fraction of one mutation type into that of the ratio of the two mutation types using $(\alpha_{lower}, \alpha_{upper}) = (\frac{p_{lower}}{1-p_{lower}}, \frac{p_{upper}}{1-p_{upper}})$.

We performed chi-square tests to evaluate the statistical significance of inter-population differences in observed mutation ratios. Specifically, for each mutation ratio in each age bin, we constructed a $2xN_{pop}$ contingency table, where each entry is the pseudo-count of observed polymorphisms of one of the two mutation types in a population. We then calculated the *p*-value of the χ^2 statistic and corrected for multiple hypothesis testing by Bonferroni correction by multiplying the *p*-value by 15x4, which represents the product of the number of age bins and the number of mutation ratios studied (for Figure 2—figure supplement 4, we substituted the first number by the number of derived allele frequency bins).

Non-negative matrix factorization (NMF) analysis

To investigate the sequence context in which the T>C signal is enriched, we applied NMF analysis. We applied NMF to the 96 x 45 dimensions matrix containing the normalized allele counts for 96 mutation types (considering the flanking 5' and 3' base nucleotides neighboring each SNP) for 15 mutation age bins in the three populations (CEU, CHB and

993 YRI); normalization was done by dividing the count of alleles by the number of SNPs in
994 each age bin (i.e., within each column). We applied NMF using the R package
995 *MutationalPatterns* and the *brunet* algorithm (Brunet et al., 2004; Manders et al., 2022)
996 with factorization ranks of 2–15. We chose the factorization rank of three as it explains
997 >99% variance and has the highest cophenetic correlation coefficient (which starts
998 decreasing after $K > 3$) (Figure 2 — figure supplement 10). Two of the three signatures
999 identified by NMF align well with inter-population differences identified in our pairwise
1000 ratio analysis (Figure 2). In particular, signature 1 corresponds to the nonCpG C>T/C>A,
1001 signature 2 corresponds to the T>C/T>G ratios, and signature 3 is a mirror image of
1002 signatures 1 and 2 likely due to the constraint imposed by performing the analysis using
1003 fractions of allele counts and using a small K value for the analysis (Figure 2 – figure
1004 supplement 10). The NMF signatures 1 and 2 are robust to removal of CpG sites and/ or
1005 singletons.
1006
1007 **Sequence contexts related to COSMIC mutational signatures SBS7 and SBS11**
1008 We downloaded loadings of the single base substitution (SBS) reference signatures on
1009 the 96 trinucleotide mutation types from COSMIC website (v3.2, GRCh38; link provided
1010 under *Resources*). We found that both SBS7a/b and SBS11 consist of nearly exclusive
1011 C>T mutations, with 86.7% mutations caused by SBS7a/b concentrated in YCN contexts
1012 while 70.0% SBS11 mutations are in NCY contexts, where Y represents pyrimidine (i.e.,
1013 C or T) and N represents any base. Therefore, as proxies for mutations potentially
1014 affected by SBS7a/b and SBS11, we removed C>T mutations in YCN and NCY contexts
1015 in analysis corresponding to Figure 2B.
1016
1017 **Quantification of parental age effects on DNM counts and ratios**
1018 We used a model-based approach to quantify the effects of paternal and maternal ages
1019 jointly by leveraging information from all phased and unphased DNMs. In short, as
1020 described in (Gao et al., 2019), we modeled the expected number of DNMs inherited from
1021 a parent as a linear function of parental age at conception, and assumed that the
1022 observed number of DNMs follows a Poisson distribution. Using a maximum likelihood
1023 approach, we estimated the sex-specific slopes and intercepts (at age zero) for each
1024 mutation ratio. Confidence intervals of the slopes and intercepts were assessed by
1025 bootstrap resampling of trios. With these estimated parental age effects, we then
1026 predicted the expected count of each mutation type and the pairwise ratios under given
1027 combinations of maternal and paternal ages, such as shown in the left panel of Figure 3.
1028
1029 For analysis corresponding to Figures 3 and 4, we inferred the parental age effects based
1030 on a DNM dataset from 2,976 Icelandic trios (Halldorsson et al., 2019). Five trios have
1031 exceedingly large numbers of DNMs given the parental ages (Proband IDs: 24496,
1032 71657, 8008, 64783, 126025) and were removed in our analysis. Given the evidence for

a non-linear effect of maternal age (i.e., a more rapid increase in maternal mutations at older ages) (Gao et al., 2019), we further excluded 92 trios with maternal ages above 40 in our analysis. Overall, DNM data from 2,879 trios were used for inference of (linear) parental age effects on DNMs. We also replicated the analysis (Figure 3—figure supplement 1; Figure 4—figure supplement 2) with an earlier dataset of DNMs from 1,548 Icelandic trios, excluding 73 trios with maternal ages above 40 (Jónsson et al., 2017).

Power simulations for C>G mutations

We performed simulations to estimate the power of detecting significant differences in the C>G/T>A ratio within and outside the maternal C>G mutation hotspots, assuming that all differences in the mutation spectrum are driven by parental ages. Specifically, we assumed two populations with different generation times (G=20, 25, 30, 35, 40) and predicted the expected fractions of C>G and T>A mutations among all *de novo* single nucleotide mutations, using point estimates of parental age effects estimated as described above. Then, we simulated the numbers of C>G and T>A SNPs assuming two independent binomial distributions with the expected fractions and the observed number of SNPs in CEU (taking the largest number of the 15 age bins). We then applied Chi-square test on the 2x2 contingency table (two mutation types x two populations) to evaluate significance of inter-population difference. We performed this simulation 10,000 times for the maternal C>G mutation hotspots and other regions separately, and estimated the detection power as the fraction of replicates with p<0.001 in the two regions (Figure 3 – figure supplement 4).

Inference of generation time corresponding to the observed polymorphism ratios

Under the scenario of co-varying paternal and maternal reproductive ages, we inferred the generation time by solving the system of linear equations:

$\frac{Gp}{Gm} = \gamma$, where γ=0.8, 1, 1.1, or 1.2 is the assumed ratio of paternal to maternal ages; and

$\frac{(\beta_p^1 Gp+\alpha_p^1)+(\beta_m^1 Gm+\alpha_m^1)}{(\beta_p^2 Gp+\alpha_p^2)+(\beta_m^2 Gm+\alpha_m^2)} = R_{1,2}$, where β and α are the slopes and intercepts estimated from DNM data for maternal (*m*) or paternal age (*p*) effects and *R* is the observed ratio of pseudo-counts of two mutation types (indicated with superscript 1,2) in an age bin.

To evaluate the uncertainty in the generation time estimates, we solved the equation system with maximum likelihood estimates from each bootstrap replicate of pedigree data and obtained 90% CIs of the inferred generation times from the overall distribution of estimates across all replicates.

Under the scenario of independently varying paternal and maternal reproductive ages, the combinations of (*Gp*, *Gm*) that satisfy $\frac{(\beta_p^1 Gp+\alpha_p^1)+(\beta_m^1 Gm+\alpha_m^1)}{(\beta_p^2 Gp+\alpha_p^2)+(\beta_m^2 Gm+\alpha_m^2)} = R_{1,2}$ follow a simple linear constraint, when other parameters are set. Therefore, we plotted in a two-dimensional plot the linear combinations of (*Gp*, *Gm*) corresponding to each observed polymorphism ratio and the slope and intercept estimates from all bootstrap replicates and compared the distribution of linear constraints across mutation ratios (Figure 4—figure supplement 2).

Acknowledgements

We thank Guy Amster, Monty Slatkin, David Reich, Nick Patterson, Chrisopher Adams and Iain Mathieson for helpful discussions. We thank Monty Slatkin, Ben Voight, Laurits Skov and Giulio Genovese for their comments on the manuscript. We thank the editors and reviewers, in particular Laurent Duret, for their valuable comments that led to significant improvement of the manuscript.

Funding

ZG was supported by a Sloan Research Fellowship and NIH R35GM146810. PM was supported by a Sloan Research Fellowship, the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics and NIH R35GM142978. YL was supported by the Hellman Fellows Fund to PM. NC was supported by the NSF GRFP fellowship. MP was supported by NIH grant GM122975.

Competing interests

The authors declare that no competing interests exist.

Additional files

Figure 1—figure supplements.

Figure supplement 1. Mutation spectrum of polymorphisms stratified by allele age in six human populations (YRI, LWK, CEU, TSI, CHB, JPT). The ancestral allele of each SNP is determined based on the human ancestral allele, and the allele age is inferred using Relate (Speidel et al., 2019). Populations are coded the same way as in the 1000 Genomes project (same below): YRI-Yoruba in Ibadan, Nigeria; LWK-Luhya in Webuye, Kenya; CEU-Utah Residents (CEPH) with Northern and Western European ancestry; TSI-Toscani in Italia; CHB-Han Chinese in Beijing, China; and JPT: Japanese in Tokyo, Japan.

1106 **Figure supplement 2. Mutation spectrum of CEU polymorphisms and DNMs from**
1107 **Icelandic trios** (Halldorsson et al., 2019)**.** "DNM 2019" denotes DNM data from 2,879
1108 Icelandic trios (Halldorsson et al., 2019), from which five trios with exceedingly large
1109 numbers of DNMs and 92 trios with maternal age above 40 were excluded. "DNM 2017"
1110 denotes DNM data from 1,475 Icelandic trios from (Jónsson et al., 2017), excluding 73
1111 trios with maternal age above 40. All polymorphisms and DNMs were filtered to the
1112 commonly accessible region.
1113
1114 **Figure supplement 3**. **Mutation spectrum of CEU polymorphisms and DNMs from**
1115 **Icelandic trios, excluding C>T transitions at CpG sites.**
1116
1117 **Figure supplement 4. Mean and distribution of B-scores of different mutation**
1118 **types.** We show the mean and the distribution of B-scores (measures the reduction in
1119 nucleotide diversity levels compared to neutral expectation (McVicker et al., 2009) (0-
1120 high constraint; 1000-low constraint) for different mutation types.
1121
1122 **Figure supplement 5. Mutation spectra of human polymorphisms in genomic**
1123 **regions with weak background selection (*B*-score>800).** The box represents the
1124 interquartile range, with the centerline showing the median value.
1125
1126 **Figure supplement 6. Fractions of S>S, S>W, W>S, and W>W mutations stratified**
1127 **by (A) derived frequency and (B) allele age.** The enrichment of weak (W)> strong (S)
1128 mutations and depletion of S>W mutations in variants with higher derived allele
1129 frequencies and old allele ages support profound effects of GC-biased gene conversion
1130 (gBGC) on human polymorphisms and biases in allele age dating by Relate, which
1131 ignores gBGC.
1132
1133 **Figure supplement 7. Distribution of variants subject to biased gene conversion**
1134 **stratified by allele age and recombination rate, (A) Fraction of S>S, S>W, W>S, and**
1135 **W>W mutation types and (B) Ratio of W>S to S>W mutations.**
1136
1137 **Figure supplement 8. Mutation spectrum of human polymorphisms in six**
1138 **populations (YRI, CEU, CHB, LWK, TSI, JPT) stratified by allele age based on**
1139 **alternative binning strategies.** The age bin boundaries were determined based on allele
1140 age distribution of variants observed in YRI (A) and CHB (B) samples, respectively.
1141
1142 **Figure 2—figure supplements.**
1143 **Figure supplement 1**. **Pairwise polymorphism ratios in three additional populations**
1144 **of African (LWK), European (TSI), and East Asian (JPT) ancestry in the 1000**
1145 **Genomes project.** Panel (A) shows the results based on all variants dated by Relate,

whereas (B) excludes singleton variants. The three signals of differences between YRI, CEU and CHB, are observed in LWK, TSI and JPT, with slight differences in the timing and magnitude of differences. Additionally, TSI shows a slightly elevated T>C/T>G ratio in recent time windows, which is absent in CEU and possibly represents another signal.

Figure supplement 2. Pairwise polymorphism ratios in genomic regions with weak background selection (*B*-score >800).

Figure supplement 3. **Pairwise polymorphism ratios in 33% of the genome with the lowest (A) and highest (B) regional recombination rates**.

Figure supplement 4. Pairwise ratios of human polymorphisms stratified by allele frequency. Unlike Figure 2 and other figure supplements, this analysis was performed on all non-singleton variants, including unphased SNPs and those unmapped by Relate. Singletons were removed because they usually have a higher false positive rate. We note that the derived allele frequency is a poor proxy for allele age, and variants at the same (sample) frequency can have drastically different ages within and across populations, which renders the inter-population comparisons difficult to interpret and not directly comparable with Figure 2. However, for recent changes in the mutation spectrum, we expect the mutation ratios of low-frequency variants to differ across populations, as those variants are likely to be young. Consistent with this expectation, the two post-OOA signals in nonCpG C>T/C>A and C>G/T>A ratios are replicated in variants at low and intermediate frequencies. The signal of T>C/T>G in ancient variants is also discernible in high-frequency variants with DAF>90%.

Figure supplement 5. **Pairwise polymorphism ratios in YRI, CEU, and CHB in commonly accessible regions based on alternative binning strategies**. The age bin boundaries were determined based on allele age distribution of variants observed in YRI (A) and CHB (B) samples, respectively.

Figure supplement 6. Alternative pairwise polymorphism ratios (T>G/T>A and C>G/C>A) to investigate the cause of inter-population differences in C>G/T>A ratio.

Figure supplement 7. Comparison of allele age estimates inferred by Relate. Relate infers the allele age of a mutation in two rounds: one initial estimate based on the entire dataset and then a more refined estimate for each population after branch length re-estimation (as recommended by Relate). Panels A-C show the difference in the initial and refined population-specific mutation ages for each population (CEU, YRI and CHB). Panels D-F show the comparison population-specific estimates for the same variant found in pairs of populations. Mutation ages with overlapping mutation age estimates across the

two populations are shown in grey and non-overlapping ranges are shown in red. Numbers in the legend indicate the total number and proportion of variants in each class.

Figure supplement 8. The T>C/T>G signal in variants shared across continental groups (A) and non-shared variants (B). Considering only six populations (YRI, LWK, CEU, TSI, CHB, JPT), we operationally defined shared variants as SNPs with both alleles observed in samples from at least one population from each of the three continental groups (i.e., variants segregating in (YRI or LWK) and (CEU or TSI) and (CHB or JPT)). For shared variants, we observed no significant inter-population differences in the T>C/T>G ratio ($p>0.05$), despite some elevation in old variants compared to younger variants in all populations (A); in contrast, the inter-population differences for non-shared variants are highly significant (B), suggesting that the T>C / T>G signal is driven mostly by non-shared variants instead of bias or inaccuracy in allele age estimation.

Figure supplement 9. **Alternative pairwise polymorphism ratios (T>C/T>A and T>A/T>G) to investigate the cause of inter-population differences in T>C/T>G ratio.** The significant differences in recent time windows (<3000 generations) in T>C/T>A and T>A/T>G comparison are driven by differences in T>A mutation rate (CHB >CEU>YRI), which is consistent with findings in Figure 2—figure supplement 6.

Figure supplement 10. Analysis of Mutational Signatures Across Allele Age Bins Using non-negative matrix factorization (NMF): We ran NMF on the normalized allele counts for 96 mutation types (considering the flanking 3' and 5' base nucleotides neighboring each SNP) for 15 mutation age bins in the three populations (CEU, CHB and YRI). (A) shows the fractions of each mutation context for each of the three significant signatures. (B) shows the contributions of each of three Signatures 1-3 in each population, YRI, CEU and CHB, across allele age bins. (C) shows the percent variance explained by NMF analysis for factorization rank K (ranging from K = 2–15), and (D) represents the cophenetic correlation coefficients calculated over 200 independent NMF runs with a constant factorization rank K (ranging from K = 2–15).

Figure supplement 11. Enrichment of TpG>CpG mutations in all T>C mutations in ancient time bins. T>C mutations at CpG sites are greatly enriched in ancient variants relative to T>C mutations at nonCpG sites, consistent with the hypothesis that some of the ancient TpG>CpG mutations are mis-polarized CpG>TpG mutations. Moreover, the temporal trend of (T>C at CpG)/(T>C at nonCpG) ratio mirrors that of T>C/T>G ratio closely.

Figure supplement 12. Persistence of the T>C/T>G signal with alternative methods for derived allele polarization and stratification of the reference genome by local

ancestry. Panels (A) and (B) show the results when the ancestral alleles were polarized to the chimpanzee reference allele or high-confidence sites in EPO. Panel (C) shows the results in subsets of genomic regions stratified based on local ancestry of human reference genome (African or European; results not shown for the small fraction of reference genome of Asian ancestry). The T>C/T>G is no longer significant in the two subsets of the genome due to reduction in sample size, but YRI still shows a higher ratio than CEU and CHB in both subsets with similar magnitude of elevation.

Figure supplement 13. Comparison of pairwise mutation ratios for polymorphisms arising in different time windows after excluding singletons. Unlike Figure 2 and other figure supplements, this analysis was performed after excluding singletons because they usually have a higher false positive rate. We show four pairwise mutation ratios, each of which compares two mutation types that are matched for mutational opportunity and effects of gBGC. Asterisks refer to the p-value obtained from a Chi-square test after a Bonferroni correction for 60 tests: * for $P < 0.01$, ** for $P<0.0001$ and *** for $P<10^{-8}$.

Figure 3—figure supplements.

Figure supplement 1. **Effects of parental ages on three pairwise mutation ratios estimated from an earlier DNM dataset** (Jónsson et al., 2017)**.** Shown on the right are the observed ratios of polymorphisms in CEU stratified by allele age.

Figure supplement 2. **Discrepancies in the mutation spectrum between DNM datasets and between DNMs and young polymorphisms**. Panel (A) shows the fractions of seven mutation types in two DNM datasets (Halldorsson et al. 2019; (Jónsson et al., 2017)) as well as in CEU SNPs in the three most recent time windows. Young polymorphisms are depleted of C>T transitions at CpG sites compared to DNMs, consistent with the expectation that recurrent mutations are undated and ignored by Relate. As we previously noted (Gao et al., 2019), the fraction of C>A mutation in 2017 DNM dataset is substantially lower than that in polymorphisms, which indicates under-detection and is somewhat ameliorated in the 2019 DNM dataset. Panel (B) shows that, in addition to differences in C>T/C>A ratios at both CpG and nonCpG sites, the two DNM datasets also differ significantly in the T>C/T>G ratio, suggesting additional technical differences in mutation identification (note that the 2017 dataset is a subset of the 2019 dataset). Furthermore, the C>G/T>A ratios of both DNM datasets are significantly higher than that in young polymorphisms, highlighting technical differences between variant detection in DNM study and population dataset.

Figure supplement 3. **Sex-specific parental age effects on three pairwise mutation ratios.** In the upper panel, the background color represents the expected DNM ratio given the paternal (x-axis) and maternal (y-axis) age, with darker colors representing greater

values. Each colored line represents the linear combinations of paternal and maternal ages corresponding to a certain mutation ratio observed in polymorphisms in an age bin. The vertical patterns for nonCpG C>T/C>A and T>C/T>G ratios suggest that these two ratios are insensitive to the maternal age. The lower panel shows the observed polymorphism ratios ordered by allele age (same data as in Figure 2—figure supplement 10), with the colors matching those of lines in the upper panel.

Figure supplement 4. **Power for detecting inter-population differences in C>G/T>A ratio driven by differences in generation time.** We performed simulations based on the parental age effects on the mutation spectrum estimated from the DNM dataset and the observed number of SNPs (see Materials and Methods). 10,000 simulation replicates were performed for maternal C>G mutation hotspots and non-hotspots, respectively. The simulated C>G/T>A ratios were compared between two populations with different generation times (G=20, 25, 30, 35, or 40), and the power was estimated from the percentage of replicates with p-value<0.001 based on Chi-square test.

Figure 4—figure supplements.

Figure supplement 1. **Past generation times corresponding to the observed pairwise polymorphism ratios, assuming fixed ratio of male to female generation times of 0.8 (A), 1.1 (B) and 1.2 (C).** The parental age effects were inferred from DNM data of 2,879 Icelandic trios from (Halldorsson et al., 2019).

Figure supplement 2. Past generation times corresponding to the observed pairwise polymorphism ratios, based on parental age effects estimated from an earlier DNM data (Jónsson et al., 2017). In each panel, the ratio of male to female generation times is assumed to be fixed throughout time at 0.8 (A), 1 (B) and 1.2 (C), respectively.

Figure supplement 3. **Combinations of paternal and maternal reproductive ages corresponding to the observed pairwise polymorphism ratios.** Under the assumption of linear parental age effects on the mutation rate of each mutation type, a specific value of pairwise polymorphism ratio places a linear constraint on the values of paternal and maternal reproductive ages (denoted by Gp and Gm), which is represented by a line in each plot. Blue lines represent predicted constraints based on the maximum likelihood estimators of mutation parameters estimated from DNM data ((Halldorsson et al., 2019)); gray lines show constraints from 500 bootstrap replicates by resampling trios with replacement. Panels (A) and (B) show the same results but with Gm and Gp shown on the x- and y-axes respectively, in order to illustrate their temporal trends. Consistent with results shown in Figure 3—figure supplement 3, the T>C/T>G and nonCpG C>T/C>A ratios are relatively insensitive to Gm and largely determined by Gp, so the temporal

trends of these two mutation ratios in panel (B) mirror those of Figure 4. The C>G/T>A depends on both Gp and Gm, so its temporal trend in both panels (A) and (B) mirrors that shown in Figure 4. Note that for each time window, the gray areas predicted by three polymorphism ratios barely, if at all, overlap, suggesting that no combination of (Gp, Gm) values can explain the three observed polymorphism ratios simultaneously. In addition, the temporal trends predicted by the three polymorphism ratios disagree with each other.

Supplementary files

The source data can be downloaded from:

Figure 1 – source data 1. Relate output for SNPs in commonly accessible regions with additional annotation (one file for each population of YRI, LWK, CEU, TSI, CHB, JPT).

Figure 1 – source data 2. Text file with (pseudo-)counts of different types of mutations in YRI, LWK, CEU, TSI, CHB, JPT in each time window.

Figure 2 – source data 1. Text file with (pseudo-)counts of mutations classified into eight types in genomic regions including, excluding, and within the maternal C>G mutation hotspots, in YRI, LWK, CEU, TSI, CHB, JPT in each time window.

Figure 3 – source data 1. Mutation parameters inferred from DNM data in 2,879 Icelandic trios with estimated uncertainty based on bootstrap resampling (one file for each mutation type for commonly accessible regions; *n*=500 replicates).

Figure 4 – source data 1. Past generation times inferred from each polymorphism ratio, assuming fixed ratios of male to female generation times (Gp/Gm=0.8, 1, 1.1, 1.2), with confidence intervals estimated using bootstrap resampling (*n*=500 replicates; one file for each mutation type).

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. *Nature* **526**:68–74.

Aggarwala V, Voight BF. 2016. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat Genet* **48**:349–355.

Aikens RC, Johnson KE, Voight BF. 2019. Signals of Variation in Human Mutation Rate at Multiple Levels of Sequence Context. *Mol Biol Evol* **36**:955–965.

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, Islam SMA, Lopez-Bigas N, Klimczak LJ, McPherson JR, Morganella S, Sabarinathan R, Wheeler DA, Mustonen V, PCAWG Mutational Signatures Working Group, Getz G, Rozen SG, Stratton MR, PCAWG Consortium. 2020. The repertoire of mutational signatures in human cancer. *Nature* **578**:94–101.

Anderson-Trocmé L, Farouni R, Bourgey M, Kamatani Y, Higasa K, Seo J-S, Kim C,

Matsuda F, Gravel S. 2020. Legacy Data Confound Genomics Studies. *Mol Biol Evol* **37**:2–10.

Bergeron LA, Besenbacher S, Turner TN, Versoza CJ, Wang RJ, Price AL, Armstrong E, Riera M, Carlson J, Chen H-Y, Hahn MW, Harris K, Lo Natalie M Kleppe AS, López-Nandam EH, Moorjani P, Pfeifer SP, Tiley GP, Yoder AD, Zhang G, Schierup MH. 2021. Mutationathon: towards standardization in estimates of pedigree-based germline mutation rates. *bioRxiv*. doi:10.1101/2021.08.30.458162

Brunet J-P, Tamayo P, Golub TR, Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* **101**:4164–4169.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, Fairley S, Runnels A, Winterkorn L, Lowy E, Human Genome Structural Variation Consortium, Paul Flicek, Germer S, Brand H, Hall IM, Talkowski ME, Narzisi G, Zody MC. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**:3426–3440.e19.

Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, Boehnke M, Kang HM, Scott LJ, Li JZ, Zöllner S, BRIDGES Consortium. 2018. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat Commun* **9**:3753.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**:1289–1303.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* **27**:2156–2158.

DeWitt WS, Harris KD, Ragsdale AP, Harris K. 2021. Nonparametric coalescent inference of mutation spectrum history and demography. *Proc Natl Acad Sci U S A* **118**. doi:10.1073/pnas.2013798118

Duncan BK, Miller JH. 1980. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**:560–561.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**:285–311.

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

Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, Jorde LB, Amster G, Przeworski M. 2019. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc Natl Acad Sci U S A* **116**:9491–9500.

Goldberg ME, Harris K. 2022. Mutational Signatures of Replication Timing and Epigenetic Modification Persist through the Global Divergence of Mutation Spectra across the Great Ape Phylogeny. *Genome Biol Evol* **14**. doi:10.1093/gbe/evab104

Goldmann JM, Seplyarskiy VB, Wong WSW, Vilboux T, Neerincx PB, Bodian DL, Solomon BD, Veltman JA, Deeken JF, Gilissen C, Niederhuber JE. 2018. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat Genet* **50**:487–492.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Jensen JD,

1394 Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R,
1395 Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A,
1396 Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan
1397 P, Brajkovic D, Kucan Ž, Gušic I, Doronichev VB, Golovanova LV, Lalueza-Fox C,
1398 de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush
1399 D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D,
1400 Pääbo S. 2010. A draft sequence of the Neandertal genome. *Science* **328**:710–
1401 722.
1402 Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson
1403 HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, Gudjonsson SA, Frigge
1404 ML, Thorleifsson G, Sigurdsson A, Stacey SN, Sulem P, Masson G, Helgason A,
1405 Gudbjartsson DF, Thorsteinsdottir U, Stefansson K. 2019. Characterizing
1406 mutagenic effects of recombination through a sequence-level genetic map.
1407 *Science*. doi:10.1126/science.aau1043
1408 Harpak A, Bhaskar A, Pritchard JK. 2016. Mutation Rate Variation is a Primary
1409 Determinant of the Distribution of Allele Frequencies in Humans. *PLoS Genet*
1410 **12**:e1006489.
1411 Harris K. 2015. Evidence for recent, population-specific evolution of the human mutation
1412 rate. *Proc Natl Acad Sci U S A* **112**:3439–3444.
1413 Harris K, Pritchard JK. 2017. Rapid evolution of the human mutation spectrum. *Elife* **6**.
1414 doi:10.7554/eLife.24284
1415 Hernandez RD, Williamson SH, Bustamante CD. 2007a. Context dependence,
1416 ancestral misidentification, and spurious signatures of natural selection. *Mol Biol*
1417 *Evol* **24**:1792–1800.
1418 Hernandez RD, Williamson SH, Zhu L, Bustamante CD. 2007b. Context-dependent
1419 mutation rates may cause spurious signatures of a fixation bias favoring higher GC-
1420 content in humans. *Mol Biol Evol* **24**:2196–2202.
1421 Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis
1422 reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl*
1423 *Acad Sci U S A* **101**:13994–14001.
1424 Jiang P, Ollodart AR, Sudhesh V, Herr AJ, Dunham MJ, Harris K. 2021. A modified
1425 fluctuation assay reveals a natural mutator phenotype that drives mutation
1426 spectrum variation within Saccharomyces cerevisiae. *Elife* **10**.
1427 doi:10.7554/eLife.68285
1428 Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT,
1429 Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, Ward LD, Arnadottir GA,
1430 Helgason EA, Helgason H, Gylfason A, Jonasdottir A, Jonasdottir A, Rafnar T,
1431 Frigge M, Stacey SN, Th Magnusson O, Thorsteinsdottir U, Masson G, Kong A,
1432 Halldorsson BV, Helgason A, Gudbjartsson DF, Stefansson K. 2017. Parental
1433 influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*
1434 **549**:519–522.
1435 Kaplanis J, Ide B, Sanghvi R, Neville M, Danecek P, Coorens T, Prigmore E, Short P,
1436 Gallone G, McRae J, Others. 2021. Genetic and chemotherapeutic causes of
1437 germline hypermutation. *bioRxiv* 2021–2006.
1438 Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite
1439 population due to steady flux of mutations. *Genetics* **61**:893–903.

1440 Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson
1441 SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WSW, Sigurdsson G, Bragi
1442 Walters G, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason
1443 A, Magnusson OT, Thorsteinsdottir U, Stefansson K. 2012. Rate of de novo
1444 mutations and the importance of father's age to disease risk. *Nature*.
1445 doi:10.1038/nature11396
1446 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K,
1447 Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L,
1448 Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C,
1449 Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C,
1450 Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J,
1451 Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman
1452 R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S,
1453 Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L,
1454 Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims
1455 S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER,
1456 Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty
1457 KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx
1458 PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S,
1459 Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier
1460 M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives
1461 CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock
1462 GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C,
1463 Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave
1464 F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm
1465 L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M,
1466 Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L,
1467 Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers
1468 RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C,
1469 Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe
1470 BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la
1471 Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L,
1472 Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB,
1473 Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler
1474 EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D,
1475 Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A,
1476 Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM,
1477 McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler
1478 G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-
1479 Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP,
1480 Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos
1481 A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen
1482 YJ, Szustakowki J, International Human Genome Sequencing Consortium. 2001.
1483 Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.
1484 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria
1485 AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA,

1486 Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN,
1487 Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta
1488 N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso
1489 GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K,
1490 Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-
1491 H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R,
1492 Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M,
1493 McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A,
1494 Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang
1495 MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation
1496 Consortium. 2016. Analysis of protein-coding genetic variation in 60,706 humans.
1497 *Nature* **536**:285–291.
1498 Li R, Bitoun E, Altemose N, Davies RW, Davies B, Myers SR. 2019. A high-resolution
1499 map of non-crossover events reveals impacts of genetic diversity on mammalian
1500 meiotic recombination. *Nat Commun* **10**:3900.
1501 Macià MC, Skov L, Peter BM, Schierup MH. 2021. Different historical generation
1502 intervals in human populations inferred from Neanderthal fragment lengths and
1503 mutation signatures. *Nature Communications*. doi:10.1038/s41467-021-25524-4
1504 Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N,
1505 Nordenfelt S, Tandon A, Skoglund P, Lazaridis I, Sankararaman S, Fu Q, Rohland
1506 N, Renaud G, Erlich Y, Willems T, Gallo C, Spence JP, Song YS, Poletti G, Balloux
1507 F, van Driem G, de Knijff P, Romero IG, Jha AR, Behar DM, Bravi CM, Capelli C,
1508 Hervig T, Moreno-Estrada A, Posukh OL, Balanovska E, Balanovsky O,
1509 Karachanak-Yankova S, Sahakyan H, Toncheva D, Yepiskoposyan L, Tyler-Smith
1510 C, Xue Y, Abdullah MS, Ruiz-Linares A, Beall CM, Di Rienzo A, Jeong C,
1511 Starikovskaya EB, Metspalu E, Parik J, Villems R, Henn BM, Hodoglugil U, Mahley
1512 R, Sajantila A, Stamatoyannopoulos G, Wee JTS, Khusainova R, Khusnutdinova E,
1513 Litvinov S, Ayodo G, Comas D, Hammer MF, Kivisild T, Klitz W, Winkler CA,
1514 Labuda D, Bamshad M, Jorde LB, Tishkoff SA, Watkins WS, Metspalu M, Dryomov
1515 S, Sukernik R, Singh L, Thangaraj K, Pääbo S, Kelso J, Patterson N, Reich D.
1516 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse
1517 populations. *Nature* **538**:201–206.
1518 Manders F, Brandsma AM, de Kanter J, Verheul M, Oka R, van Roosmalen MJ, van der
1519 Roest B, van Hoeck A, Cuppen E, van Boxtel R. 2022. MutationalPatterns: the one
1520 stop shop for the analysis of mutational processes. *BMC Genomics* **23**:134.
1521 Mathieson I, McVean G. 2014. Demography and the age of rare variants. *PLoS Genet*
1522 **10**:e1004528.
1523 Mathieson I, Reich D. 2017. Differences in the rare variant spectrum among human
1524 populations. *PLoS Genet* **13**:e1006581.
1525 McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of
1526 natural selection in hominid evolution. *PLoS Genet* **5**:e1000471.
1527 Milligan WR, Amster G, Sella G. 2022. The impact of genetic modifiers on variation in
1528 germline mutation rates within and among human populations. *Genetics*
1529 **221**:iyac087.
1530 Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M,
1531 Klein M, Hildebrandt J, Neumann M, Kliebenstein D, Weng M-L, Imbert E, Ågren J,

Rutter MT, Fenster CB, Weigel D. 2022. Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature* **602**:101–105.

Moorjani P, Amorim CEG, Arndt PF, Przeworski M. 2016a. Variation in the molecular clock of primates. *Proc Natl Acad Sci U S A* **113**:10607–10612.

Moorjani P, Sankararaman S, Fu Q, Przeworski M, Patterson N, Reich D. 2016b. 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.

Murphy D, Elyashiv E, Amster G, Sella G. 2022. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. *eLife*. doi:10.7554/eLife.76065

Narasimhan VM, Rahbari R, Scally A, Wuster A, Mason D, Xue Y, Wright J, Trembath RC, Maher ER, van Heel DA, Auton A, Hurles ME, Tyler-Smith C, Durbin R. 2017. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat Commun* **8**:303.

Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, Birney E. 2008. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* **18**:1829–1843.

Sasani TA, Ashbrook DG, Beichman AC, Lu L, Palmer AA, Williams RW, Pritchard JK, Harris K. 2022. A natural mutator allele shapes mutation spectrum variation in mice. *Nature* **605**:497–502.

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

Seoighe C, Scally A. 2017. Inference of Candidate Germline Mutator Loci in Humans from Genome-Wide Haplotype Data. *PLoS Genet* **13**:e1006549.

Seplyarskiy VB, Soldatov RA, Koch E, McGinty RJ, Goldmann JM, Hernandez RD, Barnes K, Correa A, Burchard EG, Ellinor PT, McGarvey ST, Mitchell BD, Vasan RS, Redline S, Silverman E, Weiss ST, Arnett DK, Blangero J, Boerwinkle E, He J, Montgomery C, Rao DC, Rotter JI, Taylor KD, Brody JA, Chen Y-DI, de Las Fuentes L, Hwu C-M, Rich SS, Manichaikul AW, Mychaleckyj JC, Palmer ND, Smith JA, Kardia SLR, Peyser PA, Bielak LF, O'Connor TD, Emery LS, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Population Genetics Working Group, Gilissen C, Wong WSW, Kharchenko PV, Sunyaev S. 2021. Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science* **373**:1030–1035.

Speidel L, Cassidy L, Davies RW, Hellenthal G, Skoglund P, Myers SR. 2021. Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies. *Mol Biol Evol* **38**:3497–3511.

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

Tsai-Wu JJ, Liu HF, Lu AL. 1992. Escherichia coli MutY protein has both N-glycosylase and apurinic/apyrimidinic endonuclease activities on A.C and A.G mispairs. *Proceedings of the National Academy of Sciences* **89**:8779–8783.

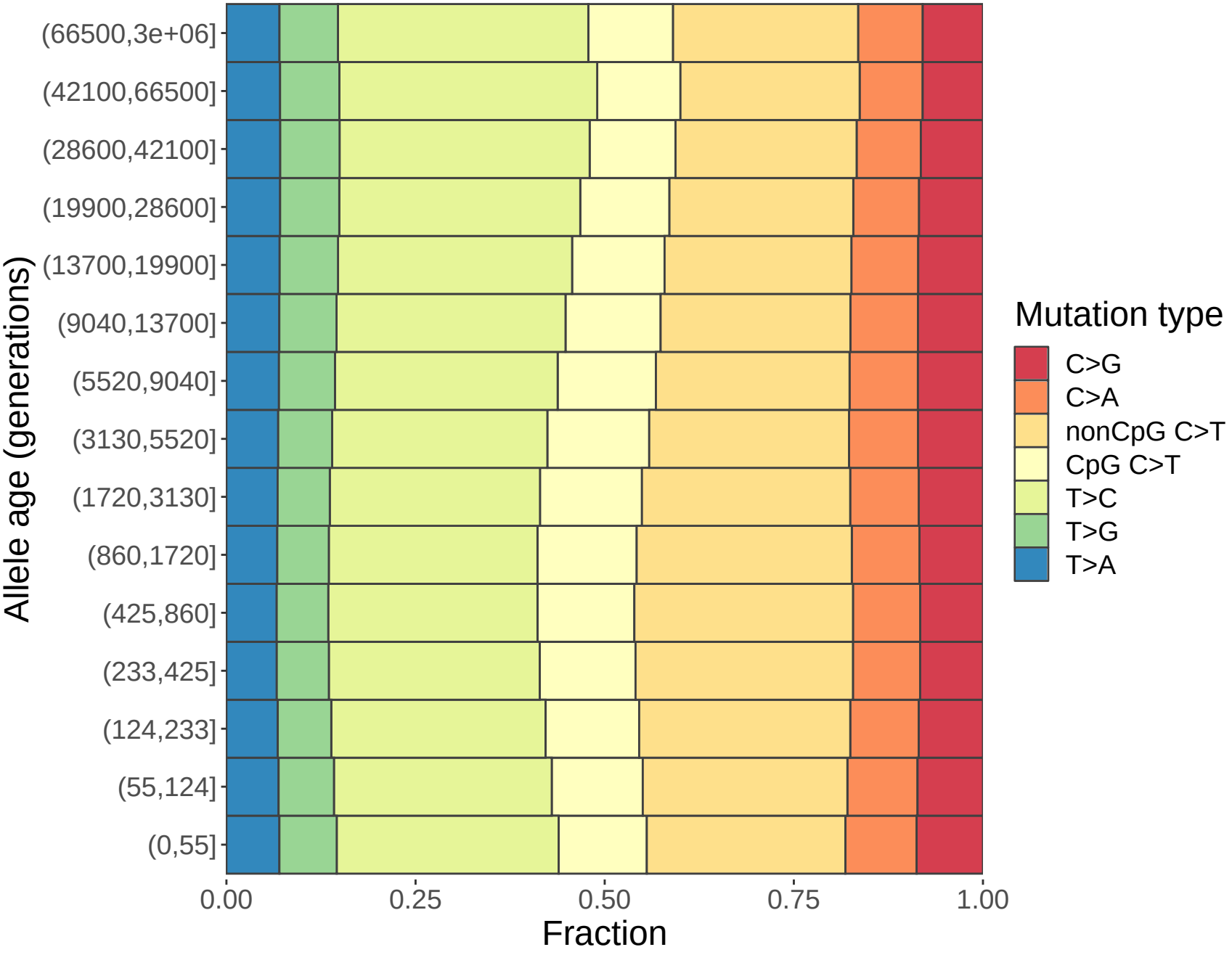
Wakeley J. 2010. Natural selection and coalescent theory. *Evolution since Darwin: the first* **150**:119–149.

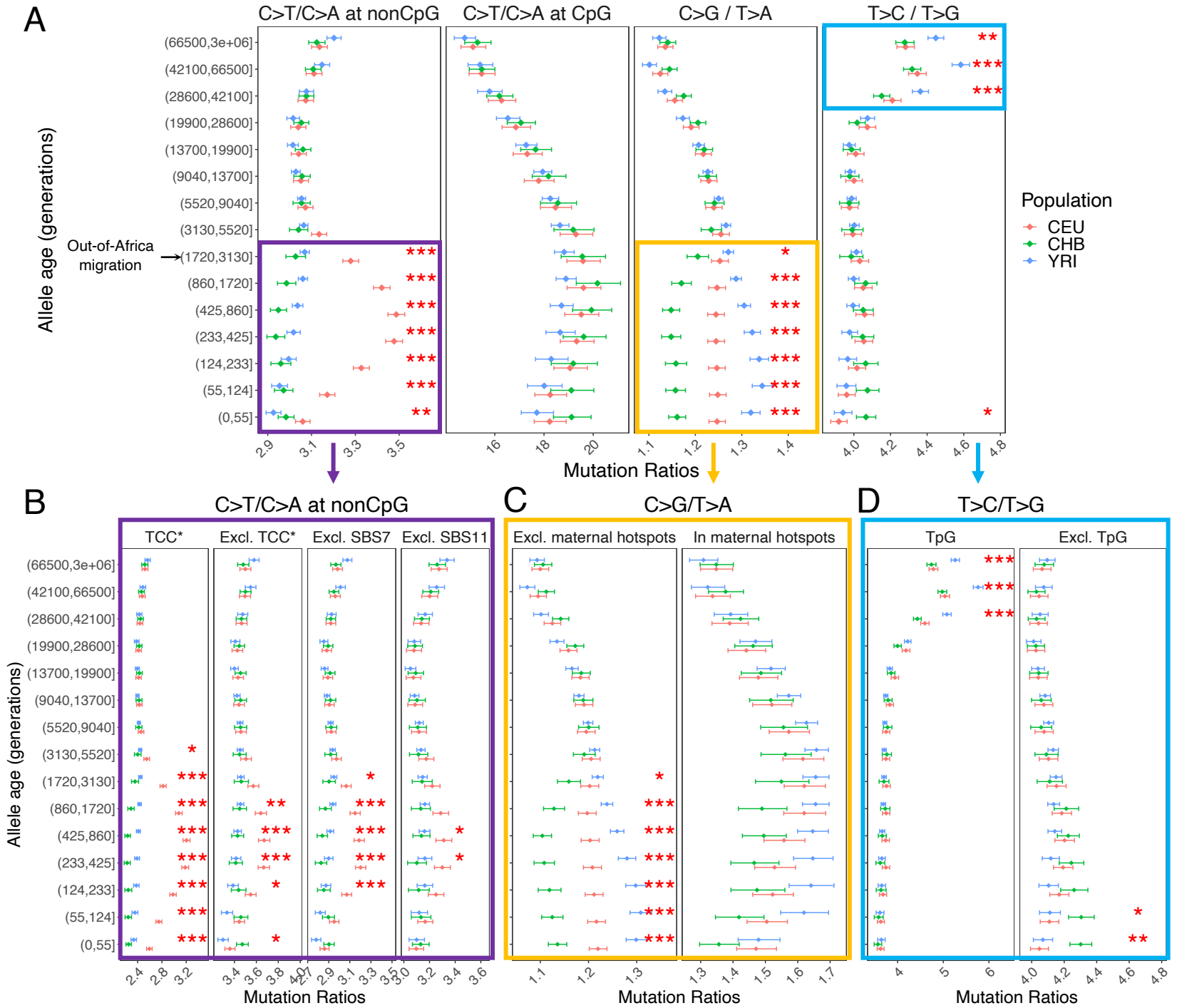
Wang RJ, Al-Saffar SI, Rogers J, Hahn MW. 2023. Human generation times across the

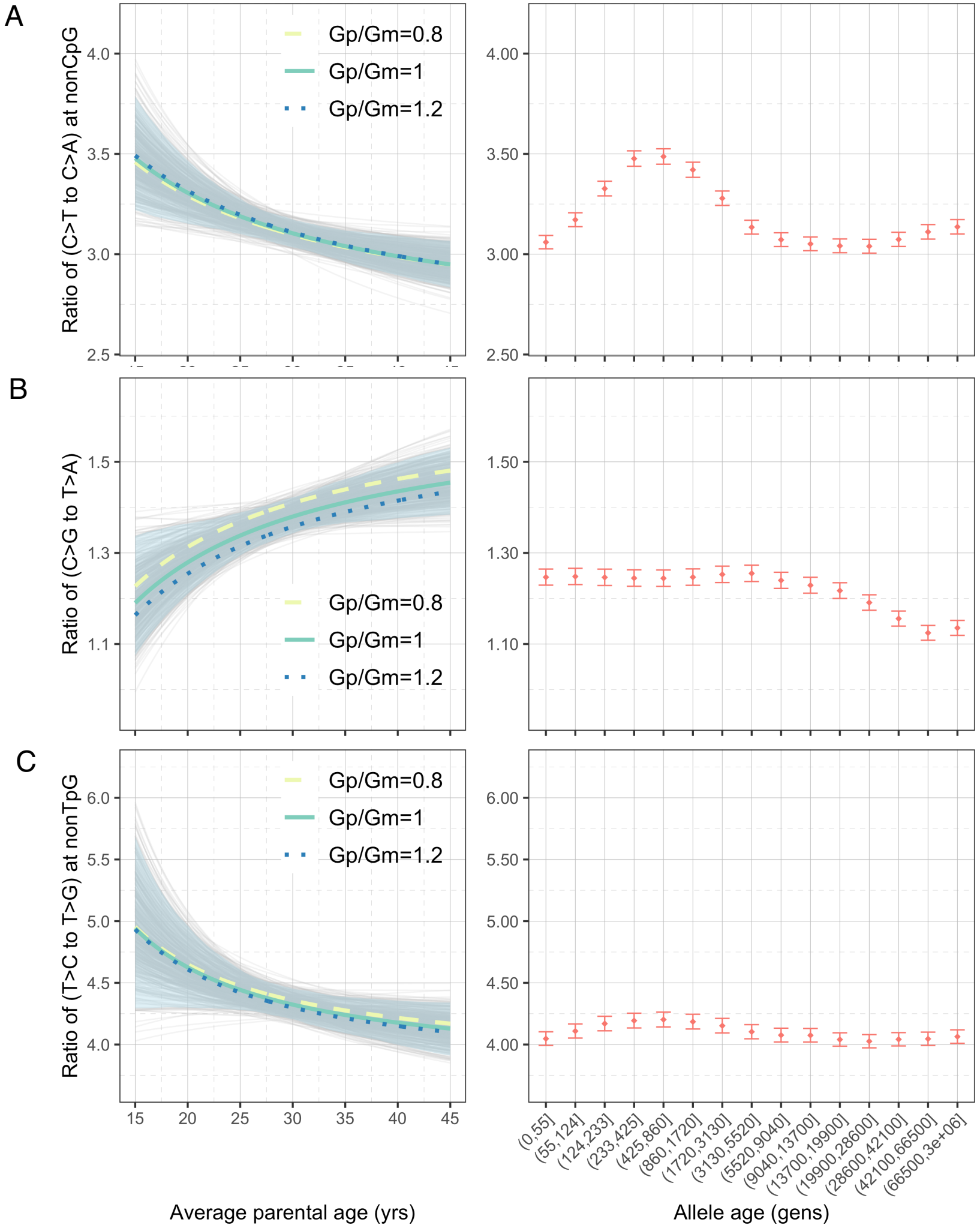
1578 past 250,000 years. *Science Advances* **9**:eabm7047.
1579 Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than
1580 in man. *Proc Natl Acad Sci U S A* **82**:1741–1745.
1581 Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. 2014. CrossMap: a versatile
1582 tool for coordinate conversion between genome assemblies. *Bioinformatics*
1583 **30**:1006–1007.

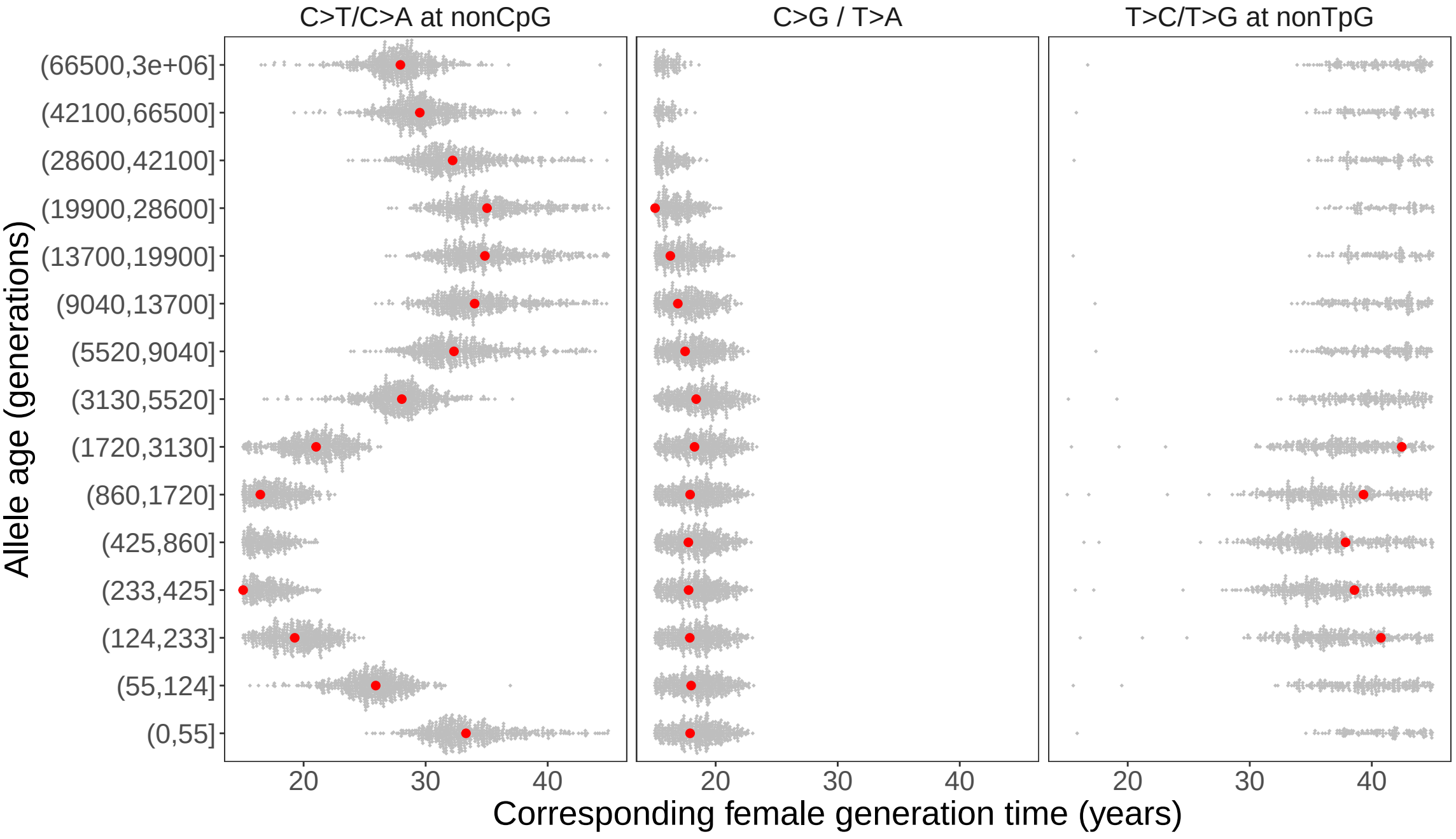
1584

1585


Allele age (generations)
(66500,3e+06]
(42100,66500]
(28600,42100]
(19900,28600]
(13700,19900]
(9040,13700]
(5520,9040]
(3130,5520]
(1720,3130]
(860,1720]
(425,860]
(233,425]
(124,233]
(55,124]
(0,55]
0.00
0.25
0.50
0.75
1.00
Fraction
Mutation type
C>G
C>A
nonCpG C>T
CpG C>T
T>C
T>G
T>A


A
C>T/C>A at nonCpG
C>T/C>A at CpG
C>G / T>A
T>C / T>G
Allele age (generations)
(66500,3e+06]
(42100,66500]
(28600,42100]
(19900,28600]
(13700,19900]
(9040,13700]
(5520,9040]
(3130,5520]
Out-of-Africa migration
(1720,3130]
(860,1720]
(425,860]
(233,425]
(124,233]
(55,124]
(0,55]
Population
CEU
CHB
YRI
Mutation Ratios
B
C>T/C>A at nonCpG
TCC*
Excl. TCC*
Excl. SBS7
Excl. SBS11
C
C>G/T>A
Excl. maternal hotspots
In maternal hotspots
D
T>C/T>G
TpG
Excl. TpG


A
B
C
Ratio of (C>T to C>A) at nonCpG
Ratio of (C>G to T>A)
Ratio of (T>C to T>G) at nonTpG
Gp/Gm=0.8
Gp/Gm=1
Gp/Gm=1.2
4.0
3.5
3.0
2.5
4.00
3.50
3.00
2.50
1.5
1.3
1.1
1.50
1.30
1.10
6.0
5.5
5.0
4.5
4.0
6.00
5.50
5.00
4.50
4.00
15
20
25
30
35
40
45
(0,55]
(55,124]
(124,233]
(233,425]
(425,860]
(860,1720]
(1720,3130]
(3130,5520]
(5520,9040]
(9040,13700]
(13700,19900]
(19900,28600]
(28600,42100]
(42100,66500]
(66500,3e+06]
Average parental age (yrs)
Allele age (gens)


C>T/C>A at nonCpG
C>G / T>A
T>C/T>G at nonTpG
Allele age (generations)
(66500,3e+06]
(42100,66500]
(28600,42100]
(19900,28600]
(13700,19900]
(9040,13700]
(5520,9040]
(3130,5520]
(1720,3130]
(860,1720]
(425,860]
(233,425]
(124,233]
(55,124]
(0,55]
20
30
40
Corresponding female generation time (years)