Timing and causes of the evolution of the germline mutation spectrum in humans

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

[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 Biological Sciences, Columbia University, New York, NY
[4]Department of Systems Biology, Columbia University, New York, NY
[5]Department of Molecular and Cell Biology, University of California, Berkeley, CA
[+]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. When and why these changes occurred remains unclear, however. We develop a framework to characterize temporal changes in polymorphisms within and between human populations, while controlling for the effects of selection and biased gene conversion. Applying this approach to high-coverage, whole genome sequences from the 1000 Genomes Project, we detect significant changes in the mutation spectrum of alleles of different ages, notably two independent changes that arose after the split of the ancestors of African and non-African populations. We also find that the mutation spectrum differs significantly between populations sampled in and outside of Africa at old polymorphisms that predate the out-of-Africa migration; this seemingly contradictory observation is likely due to mutation rate differences in remote ancestors that contributed to varying degrees to the ancestry of contemporary human populations. Importantly, by relating the mutation spectrum of polymorphisms to the parental age effects on *de novo* mutations, we show that plausible changes in the age of reproduction over time cannot explain the joint patterns observed for different mutation types. 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) (Hwang and Green 2004; Harris 2015; Moorjani, Amorim, et al. 2016; Harris and Pritchard 2017; Mathieson and Reich 2017; Narasimhan et al. 2017; Goldberg and Harris 2022; DeWitt et al. 2021; 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). 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 mutator alleles, in recombining species, they could be nearly neutral and maintained for a long time, leading to genome-wide differences across populations (Seoighe and Scally 2017; Milligan, Amster, and Sella 2021).

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”, across populations alone can lead to differences in mutation spectrum and indeed, these have been suggested 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, not only depends 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 (McVicker et al. 2009; Lander et al. 2001). 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. Moreover, a challenge in interpreting previous studies, notably to learn about when changes in mutational processes may have occurred, is the reliance on allele frequencies (Harris and Pritchard 2017; Mathieson and Reich 2017). While frequencies are informative of the age of alleles (Kimura 1969), demographic history shapes the relationship between allele frequency and age. Thus, 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 spectrum can also 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 (Seplyarskiy et al. 2021; Monroe et al. 2022). 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 thus 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.

We therefore propose a new framework to compare the mutation spectrum within and across human populations. First, we infer the age of each derived allele observed in a population using a newly developed approach, Relate, to reconstruct local genealogies and estimate 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 at deeper timescales, beyond the split of contemporary human populations. 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 fraction of one mutation type necessarily lowers the fractions 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 evolutionary time. 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 generation time changes versus other biological factors 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., 2021). To focus on putatively neutral mutations, we removed exons and phylogenetically conserved regions (see Materials and Methods). 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 (polarized to the human ancestral allele) 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 roughly equal-sized bins based on the ages of the derived allele inferred by Relate, accounting for uncertainty in the inferred mutation age by assuming uniform distribution of age 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, e.g., T>C includes both T>C and A>G substitutions). Given the well-characterized hypermutability of methylated CpG sites (Duncan and Miller 1980; Kong et al. 2012), we further divided C>T SNPs into sub-types occurring in CpG and non-CpG contexts by considering the flanking base pair on either side of the variant.

We found marked differences in the relative proportions of different mutation types (i.e., the mutation spectrum) across varying allele age bins within CEU (Figure 1) as well as in YRI and in CHB (Figure 1—figure supplement 1), as seen earlier in low coverage 1000G data (Speidel et al. 2019). We obtained qualitatively similar results when considering other 1000G populations of TSI, LWK and JPT (Figure 1—figure supplement 1). This observation echoes previous findings about the evolution of the mutation spectrum comparing polymorphisms across allele frequencies (Harris and Pritchard 2017; Mathieson and Reich 2017; Carlson et al. 2018). As noted previously, however, differences in mutation spectrum across frequencies alone are weak evidence for the evolution of the mutation process itself, because patterns of standing polymorphisms can be affected by repeat mutations and other evolutionary forces, including selection and gene conversion.

Notably, the infinite-sites model is a reasonable assumption for small sample sizes (Kimura 1969), but recurrent mutations become highly likely in large datasets, especially at sites with higher mutation rates (Lek et al. 2016; Harpak, Bhaskar, and Pritchard 2016). Recurrent, multi-allelic, and back mutations violate the model assumptions of Relate and are often excluded from its output. For instance, given the higher mutation rate of transitions at CpG sites, such SNPs are

more likely to be subject to recurrent mutations in a large sample and thus may map to multiple branches in the tree, leading to their exclusion from Relate's output (Speidel et al. 2019). As expected from these considerations, the fraction of CpG C>T SNPs in young mutations (i.e., those estimated to have occurred in the past ~50 generations) is lower than proportions in *de novo* mutations in present-day pedigree studies (Figure 1—figure supplement 2). Differences in mutation spectrum across age bins in modern humans persist even after excluding CpG C>T mutations (Figure 1—figure supplement 3), however, indicating that other mutation types are also changing in relative frequency over time and the observed patterns are not driven solely by recurrent mutation at CpG sites.

Next, we examined the effect of linked selection on different mutation types. While we excluded direct targets of selection from analysis (i.e., exons and conserved regions), much of the genome is linked to non-neutral variants and notably subject to background selection (Charlesworth, Morgan, and Charlesworth 1993; McVicker et al. 2009; Murphy et al., 2021). A common measure of the effects of background selection is the *B*-statistic or *B*-score that measures the reduction in nucleotide diversity levels compared to neutral expectation (McVicker et al. 2009). To characterize the impact of linked selection, we calculated the average genome-wide *B*-score of each mutation type. We obtained 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 obtained qualitatively similar results when restricting to effectively neutral regions (*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 observed 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 (both neutrality and infinite-sites mutation model) of Relate and could lead to subtle biases in estimated allele age (Speidel et al. 2019). Assuming no systematic evolution in the relative rates of S>W mutations and W>S mutations and unbiased estimation of allele age under gBGC, we would expect similar fractions of S>W and W>S mutations across age bins; in contrast, the W>S variants are overall more enriched in older variants compared to S>W variants (Figure 1—figure supplement 6B; (Speidel et al. 2019)). As the strength of biased gBGC depends on the effective population size and thus may differ across human populations, 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).

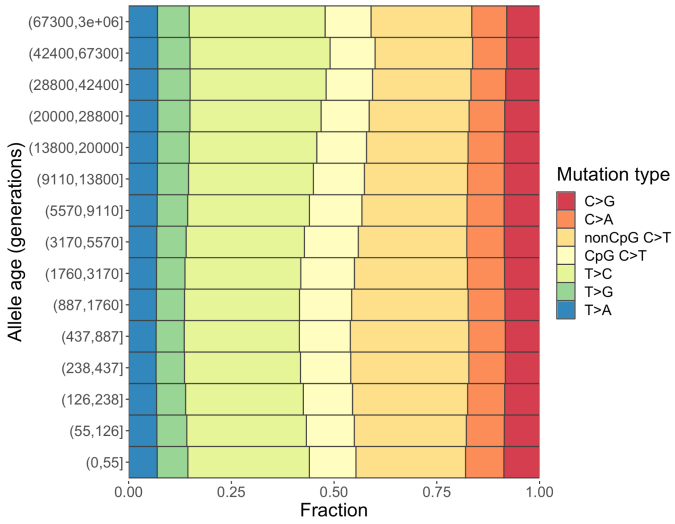

Figure 1. Changes in the mutation spectrum of polymorphisms in CEU over evolutionary time.

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)

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 spectra of human polymorphisms stratified by allele age.

Figure supplement 2. Mutation spectrum of SNPs found in CEU in comparison to that of *de novo* mutations identified in trios from Iceland.

Figure supplement 3. Mutation spectra 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 in variants stratified by derived frequency (A) and allele age (B).

Figure supplement 7. Mutation spectra of human polymorphisms 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 mutation spectrum comparisons, we focused on comparisons of pairs of mutation types subject to similar effects of gBGC (i.e., both 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. Three of these comparisons involve mutation types with the same mutational opportunity (e.g., both T>C and T>G mutations arise from 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 observed marked differences in the ratios both over time and across populations. The contrasts across time windows and among populations point to the timing, direction, and population origin of mutation rate changes. Specifically, we found three key 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 they were replicated in other populations 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 sources unrelated to mutation or technical artifacts contributing to the inter-population differences. First, the signals are unlikely to be driven by gBGC and selection, as the pairwise comparisons are designed to control for their effects. In addition, we obtained 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). To account for potential inaccuracies in mutation ages estimated by Relate, we stratified variants by allele frequencies instead of inferred mutation age; the results were qualitatively similar (Figure 2—figure supplement 4). Lastly, we find qualitatively similar results with different boundaries for allele age binning (Figure 2—figure supplement 5). Together, these results provide unequivocal evidence that the human germline mutation spectrum has evolved over time and differs across populations.

We now discuss each of the three signals of mutation rate evolution in detail.

European-specific elevation of non-CpG C>T/C>A ratio

The largest signal that we observed is the CEU-specific transient elevation in the ratio of C>T/C>A mutations at nonCpG sites compared to the ancestral state before the OOA migration; in contrast, the nonCpG C>T/C>A ratios of CHB and YRI remain relatively constant through time. 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 found 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 effectively spreads the contribution of each variant to two or more 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 it is 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 (Harris 2015; Mathieson and Reich 2017; Alexandrov et al. 2020). 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 observed 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 transient increase in Europeans remains unclear.

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

The second largest inter-population difference was in the C>G/T>A ratio (Figure 2A): following the OOA migration, both YRI and CEU samples show a continuous 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. 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.

Notably, 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 (Jónsson et al. 2017): "C>G enriched regions", defined as 10% of the genome with the highest C>G SNP density, contribute one-third of the overall maternal age effect (i.e., the yearly increase in maternal DNMs with mother's age). The C>G/T>A ratio within the C>G enriched regions alone does not show significant inter-population differences (Figure 2C), probably 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 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 not being substituted (C>G or T>A) contributes to the inter-population differences. For T>G/T>A ratio, we still observed 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 found 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 the ancestors of all modern humans

The T>C/T>G ratios were higher in the three oldest bins (dated to >28,800 generations ago) than in more recent ones in all three population samples and that effect was more pronounced in YRI compared to CHB and CEU (Figure 2A). The difference seen between old polymorphisms observed 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 arose in the common ancestor of the three contemporary populations. To verify that these differences do not stem from inaccuracies in the polarization of ancestral and derived alleles, we repeated the analysis by determining the ancestral state on the basis of the chimpanzee reference genome (PanTro2; Materials and Methods); qualitatively similar results are obtained (Figure 2—figure supplement 7A). We further tested the effect of reference bias by stratifying the human reference genome by inferred local ancestry and found qualitatively similar results in regions of European or African ancestry (Figure 2—figure supplement 7B) (Green et al. 2010).

One known difference between Africans and non-Africans, which could contribute to the difference between YRI and the other two population samples, is that most non-Africans derive ~1-3% of their genomic ancestry from Neanderthals, and some groups also have an additional contribution from Denisovans (Mallick et al. 2016). In contrast, sub-Saharan African populations have minimal (<0.1%) ancestry from these archaic hominins (Prüfer et al. 2014; Chen et al. 2020). In principle, it is therefore possible that the increased T>C/T>G ratio in old age bins is related to variants that arose in known archaic groups (Neanderthal or Denisovan) and introgressed into ancestors of non-Africans after the OOA migration. However, several observations argue against this hypothesis. First, the similar T>C/T>G ratios in CHB and CEU, combined with the much higher Denisovan ancestry in East Asians (Mallick et al. 2016), rule out Denivosan introgression as the major driver of the observed differences between the YRI and CEU and CHB. Second, the inter-population differences persist when previously identified introgression tracts of Neanderthal ancestry were removed from CEU and CHB (Sankararaman et al. 2014). Lastly, even if the Neanderthals and/or Denisovans had higher T>C/T>G ratio compared to modern humans, their introgression into ancestors of non-Africans could only explain the inter-population differences, not the large elevation in this ratio in YRI (Figure 2).

To our surprise, when we removed variants with derived alleles observed in both Neanderthals and Denisovans (i.e., ND11 variants, where 0 and 1 denote ancestral and derived alleles respectively), the T>C/T>G ratios in the oldest bins reduced to similar levels as the younger variants for all three populations, and the differences between YRI and the other two populations disappeared (Figure 2D); the same is not seen, however, when removing only the derived variants shared with either Neanderthal or Denisovan but not both (i.e., ND01 or ND10 variants; Figure 2—figure supplement 8). Variants shared between Neanderthals, Denisovans, and modern humans (in particular YRI who have minimal archaic ancestry) likely arose in the common ancestral population before the split of the three groups (dated to 550,000-800,000 years ago; (Prüfer et al. 2014), consistent with the age estimates of >29,000 generations for variants in the three oldest bins. Therefore, we hypothesize that the T>C/T>G ratio was higher in one or more populations in the remote past and those ancient groups contribute variable amounts of ancestry to contemporary populations due to incomplete lineage sorting, complex population structure, or differential gene flow from unknown ghost hominins (Durvasula and Sankararaman 2020; Ragsdale et al. 2022; Hammer et al. 2011).

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 inferred that the signal is primarily driven by higher proportions of T>C mutations among older variants (Figure 2—figure supplement 9). Thus, it appears that the relative T>C mutation rate used to be higher in the ancestral population of modern humans (and some archaic humans) but decreased by ~29,000 generations ago and stayed relatively stable until present. Moreover, Neanderthal-introgressed segments have a slightly lower fraction of T>C mutations and a higher fraction of T>G mutations (~1% relative difference in each fraction) (Skov et al. 2020). Therefore, Neanderthal introgression into the ancestors of non-Africans could have the effect of further lowering the T>C/T>G ratio in non-Africans. Nonetheless, this effect alone would be insufficient to explain the ~4.4% difference genome-wide in the T>C/T>G ratio between YRI and non-African populations (CEU or CHB) in the older bins (Figure 2D), as only a small fraction (~1-3%) of non-Africans ancestry is derived from Neanderthals.

Given the possible differences in mutation spectrum across ancestral populations that contributed differentially to contemporary populations, a potential concern is that some variants that actually arose before OOA could be incorrectly dated and lead to apparent differences between contemporary populations even at recent timescales. To address this concern, we excluded derived variants shared by either Neanderthals or Denisovans from further analysis. This filtering removes ancestral polymorphisms segregating in modern humans due to incomplete lineage sorting or introgression from Neanderthals, Denisovans or any unknown ghost population related to either of them. After removing these variants, we still found the elevation of C>T/C>A ratio at nonCpG sites in CEU and C>G/T>A divergence among populations (Figure 2—figure supplement 10). This result confirms that these two signals are not related to mis-dated ancient variants segregating in human populations.

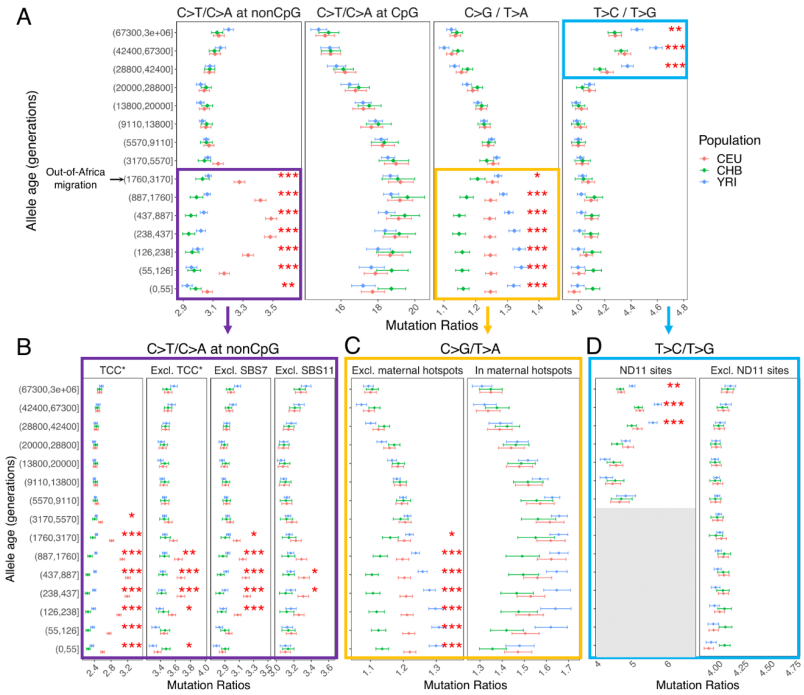

Figure 2. Pairwise comparisons of polymorphisms arising in different time windows. (A) Four pairwise mutation ratios are shown, each of which compares two mutation types that are matched for mutational opportunity and effects of gBGC. The black arrow indicates the window coinciding with the OOA migration. 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) CEU-specific elevation in C>T/C>A ratio 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 ND11 variants (data are not shown for ND11 variants in recent windows, as there are few such variants whose estimated ages fall in those windows).

The following figure supplements are available for Figure 2:

Figure supplement 1. Pairwise polymorphism ratios in three additional populations from Africa (LWK), Europe (TSI), and East Asia (JPT) in the 1000 Genomes project.

Figure supplement 2. Pairwise polymorphism ratios in genomic regions with a minimal effect of 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.

Figure supplement 5. Pairwise polymorphism ratios in YRI, CEU, and CHB in commonly accessible regions based on alternative binning strategies.

Figure supplement 6. Alternative pairwise polymorphism ratios (T>G/T>A and C>G/C>A), used to investigate the cause of inter-population differences in C>G/T>A ratio.
Figure supplement 7. Replication of the T>C/T>G signal with an alternative derived allele polarization and stratification by the local ancestry of the reference genome.
Figure supplement 8. T>C/T>G ratio in YRI, CEU and CHB excluding variants for which the derived alleles are also found in Denisovans (A), Neanderthals (B), and both (C).
Figure supplement 9. Alternative pairwise polymorphism ratios (T>C/T>A and T>A/T>G), used to investigate the cause of inter-population differences in T>C/T>G ratio.
Figure supplement 10. Pairwise polymorphism ratios in YRI, CEU, and CHB in commonly accessible regions excluding variants shared with Neanderthals and/or Denisovans (e.g., ND01, ND10 and ND11).

Parental age effects on the mutation spectrum

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

Considering the four pairwise mutation ratios, which are mathematically independent, three show a significant dependence on parental age (Figure 2). 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 T>C/T>G decrease by 9.8% (90% confidence interval (CI): 4.9%-14.7%) and 7.5% (CI: 1.8%-13.7%) respectively, whereas the C>G/T>A ratio increases by 11.6% (90% CI: 3.3%-19.8%). In terms of sex-specific effects (Figure 3—figure supplement 3), nonCpG C>T/C>A and T>C/T>G ratios were 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 (Kong et al. 2012; Jónsson et al. 2017; Goldmann et al. 2018). 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) limited 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 in the 20%-tile of trios with the youngest parents is significantly higher 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 the 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.

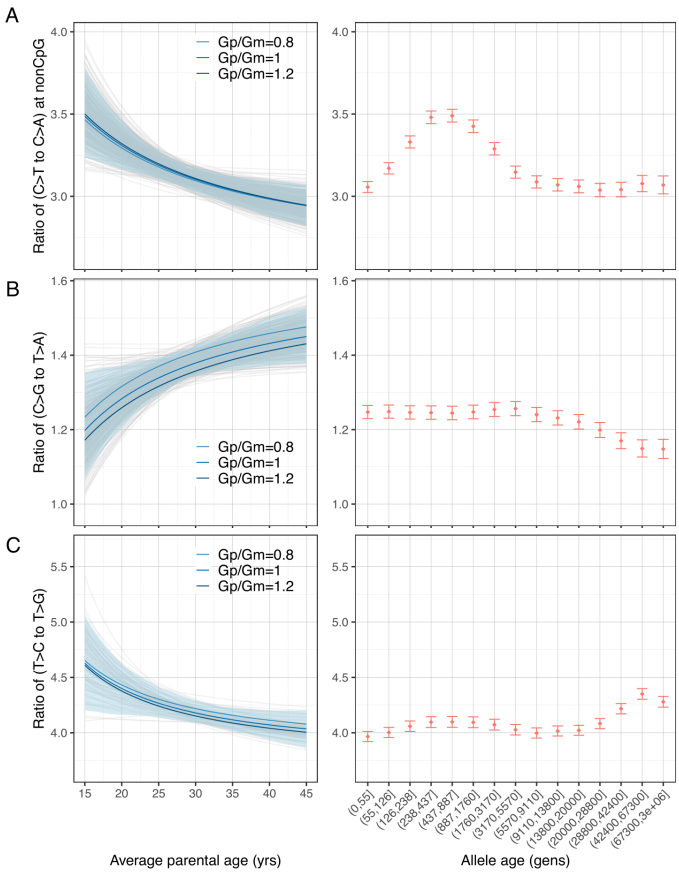

Figure 3 Effects of parental ages on three pairwise mutation ratios estimated from DNM data in 2,879 Icelandic trios (Halldorsson et al. 2019). The different curves reflect expectations for different ratios of paternal (Gp) to maternal (Gm) mean generation times. Each light gray curve represents the 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 3: 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 excluding archaic sites; *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.

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. 2021). 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 somewhat 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 an identical male to female generation time 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.

Solving each of the three pairwise ratios over time, we inferred inconsistent values of generation times across different mutation types within a single time window and discrepant trends across time windows. For instance, the steady increase in C>G/T>A ratio over time translates into a gradual increase in reproductive age, with the ratio of the most recent bin corresponding to a reproductive age under 23 years (Figure 3C; Figure 4). In contrast, the T>C/T>G polymorphism ratio appears to be lower than the ratio in *de novo* mutations across the range of typical parental ages in pedigree studies and suggests a generation time of more than 40 years (Figure 3C). Such a long population-average generation time is not only inconsistent with the estimate of the C>G/T>A ratio, but it is also unrealistic for human evolution (Fenner 2005; Moorjani, Sankararaman, et al. 2016). Finally, the transient elevation in nonCpG C>T/C>A ratio suggests a drastic, rapid reduction in the generation time (Figure 3A; Figure 4). Specifically, the ratios of both ancient (>5,670 generations ago) and the most recent polymorphisms (<55 generations ago)

correspond to relatively old reproductive ages of ~35 years, while the peak at around 238-887 generations ago provides average reproductive age estimates of less than 20 years (Figure 4).

The incompatible patterns across different mutation ratios could potentially arise if male and female generation times differed in the past. To explore this possibility, we varied the ratio of male and female generation times between 0.8–1.2, as seen across a range of contemporary human populations ((Fenner 2005); Figure 4—figure supplement 1; Figure 4—figure supplement 2). We further allowed male and female generation times to vary freely and inferred the combinations of paternal and maternal ages that could give rise to the observed polymorphism ratios. Even after modeling sex-specific reproductive ages, we found inconsistent generation time estimates and incompatible trends across mutation ratios (Figure 4—figure supplement 3). Overall, our results suggest that shifts within a plausible range of human generation times–i.e., those that fall within the range of puberty and reproductive cessation in humans–cannot explain the observed variation in the polymorphism data for CEU, and by analogy are unlikely to explain the mutation ratios in CHB or YRI polymorphisms.

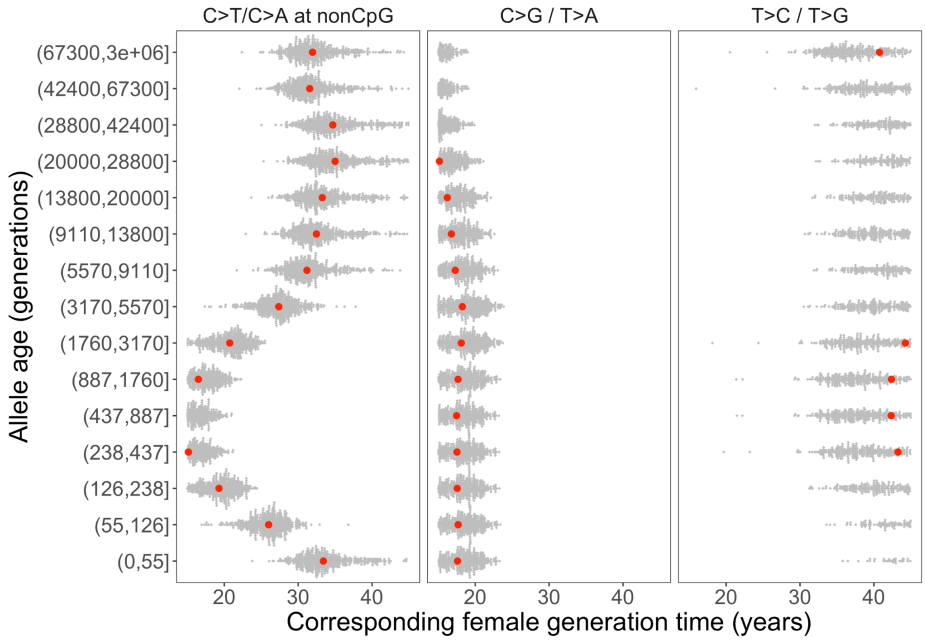

Figure 4 Past generation times corresponding to the observed polymorphism ratios in CEU, given parental age effects estimated from DNM data. Red points represent the point estimates based on maximum likelihood estimators of mutation parameters from the DNM data; gray dots show estimates from 500 bootstrap 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 independent 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 4: Past generation times inferred from each polymorphism ratio, assuming fixed ratios of male to female generation times (Gp/Gm=0.8, 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 corresponding 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 introduced 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 identified three signatures of inter-population differences. Notably, we replicated the transient elevation in non-CpG C>T mutation (manifest in the C>T/C>A ratio) identified previously in Europeans compared to East Asians and Africans (Harris 2015; Harris and Pritchard 2017; Mathieson and Reich 2017; Speidel et al. 2019). We found that this ratio also differs weakly 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, and previously identified COSMIC signatures of SBS7 and SBS11 which are associated with exposure to UV and alkylating agents. 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 Europeans. Further investigation into the extended sequence contexts of this mutation pulse may help elucidate the underlying molecular mechanism(s) (Aggarwala and Voight 2016; Aikens, Johnson, and Voight 2019).

We also identified two novel inter-population differences in the mutation spectrum. First, the ratio of C>G/T>A mutation rates differs between the three populations considered. We found 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), and our analysis added information about the timing of this variation. Given the distinct trends of T>A and C>G mutations with allele age (Figure 2—figure supplement 6), at least two changes in the mutational processes are needed to explain the inter-population 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 these populations.

Unexpectedly, we detected significant differences between YRI and the other two populations, CEU and CHB, in the mutation spectra of polymorphisms that are estimated to long predate the OOA migration. Specifically, the T>C/T>G mutation ratio is elevated in the very old allele age bins compared to more recent bins for all populations, with a significantly higher ratio seen in YRI than in CEU and CHB. We showed that the inter-population differences cannot be explained by differential gene flow from sequenced archaic hominins—Neanderthals or Denisovans—into the ancestors of non-Africans and such introgression alone cannot explain the shift in the older bins in all modern human populations.

Instead, we found evidence that the signals come from extremely old variants that emerged prior to the split of modern humans and archaic hominins at least ~550,000 years ago (Prüfer et al. 2014). This suggests that the observed differences between contemporary populations could have arisen from the complex demographic history of ancestral populations. Based on observed polymorphism patterns in contemporary African populations and using simulations, several recent studies have suggested that one or more ghost archaic populations may have introgressed into the ancestors of Africans and possibly into the common ancestors of all modern humans (Hammer et al. 2011; Ragsdale and Gravel 2019; Speidel et al. 2019; Durvasula and Sankararaman 2020). After the ancestors of non-Africans migrated out of Africa, the ghost archaic group(s) may have continued interbreeding with remaining populations in Africa, leading to higher ancestry in YRI. An alternative model is deep population structure in modern humans. Under this model, two or more long-lasting, weakly differentiated ancestral populations contributed differentially to contemporary human populations through continuous gene flow or multiple merger events (Ragsdale et al. 2022). In both models, a greater contribution from a group with a higher T>C/T>G ratio to the ancestors of African individuals would explain differences between YRI and non-African population samples as well as the elevated ratio in old variants for all three contemporary human populations. Our analysis further showed that the T>C/T>G signal comes from T>C mutations rather than T>G mutations, suggesting that one or more of the remote ancestral populations had a higher T>C mutation rate relative to their contemporaries as well as to modern humans.

Interpretation of inter-population differences in the mutation spectrum

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 the 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 could lead to considerable variation in the mutation processes within a continent. In fact, even for closely related populations, we detected 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 among human populations and to understand its evolution.

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

Across mammals, the generation time is the strongest predictor of the yearly mutation rate and of some aspects of the mutation spectrum (Wu and Li 1985; Hwang and Green 2004; Moorjani, Amorim, et al. 2016). Recent studies have suggested that changes in generation times can also explain the differences in mutation spectrum observed between human populations (Macia et al 2021). Indeed, our analysis of *de novo* mutations from pedigree studies shows significant effects of parental ages on all four pairwise mutation ratios that we examined.

On that basis, it has been suggested that the population-specific mutation spectrum can be leveraged to infer the historical generation time in humans (Macià et al. 2021; Wang et al. 2021). In practice, there are several technical hurdles. 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 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 diverse populations. Second, technical issues, both molecular and computational, may affect the reliability of variant calls of different mutation types to varying degrees (Bergeron et al. 2021). Indeed, we found that the four pairwise ratios differ significantly across three 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). Finally, 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 (e.g., T>C and T>G both classified as S>W mutations may be subject to differing strengths of BGC).

While these technical limitations likely influence the absolute values of the inferred generation times, the temporal trends, whether increasing or decreasing, should be robust to technical biases. In that regard, it is striking that the trends in generation times estimated from different mutation ratios are mutually inconsistent (Figure 4). These inconsistencies persist after accounting for uncertainty in the parental age effects inferred from pedigrees and incorporating sex-specific reproductive ages. Our findings therefore reveal that there is no single value or trajectory of generation times that can account for the observed patterns at all mutation ratios across time windows simultaneously. In other words, generation time alone cannot be the sole driver of the 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 demography. In this study, we investigated the contribution of these forces to differences in the mutation spectrum between contemporary human populations. Notably, we identified an unexpected role of demographic history in shaping variants that long predate the OOA migration, due either to deep population structure or to introgression from unknown ghost hominin population(s). Complex

demographic history alone cannot lead to inter-population differences in the mutation spectrum, so changes in the mutational processes must have happened in the ancestral populations; however, incorrect or incomplete demographic models may misguide interpretation of the timing and source of mutation rate evolution. For future studies aiming to understand the evolution of mutagenesis based on analyses of polymorphism patterns, it will be crucial to consider and model the impact of these non-mutational evolutionary forces.

Our analysis further demonstrates the limitations in inferring past generation times based on polymorphism patterns. The approach relies on the assumption that changes in the generation time play the sole (or a predominant) role in the evolution of the mutation spectrum. Yet we found that shifts in generation time could not explain the observed variations in the mutation spectrum on their own, 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. 2021) 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).

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 rate differs by less than 10% percent for any mutation type in the human and chimpanzee lineages (Moorjani, Amorim, et al. 2016). This observation suggests that although the mutation rate and spectrum can evolve over relatively short timescales, the fluctuations in 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., 2021)
Low coverage 1000 Genomes Project	http://www.1000genomes.org/category/phase-3/	(1000 Genomes Project Consortium et al. 2015)
Archaic genomes: Neanderthal and Denisovans	http://cdna.eva.mpg.de/neandertal/	(Meyer et al. 2012; Prüfer et al. 2014)
Introgressed regions from Neanderthals identified in non-Africans	https://reich.hms.harvard.edu/datasets/landscape-neandertal-ancestry-present-day-humans	(Sankararaman et al. 2014)
Introgressed regions from unknown ghost archaic hominin identified in Africans	https://drive.google.com/file/d/1Q7iEdjCf7d4HnI65_JzjIYnsDP8G0Iy8/view	(Durvasula and Sankararaman 2020)
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)

Goldman *de novo* mutations 2016 (806 families)	https://pubmed.ncbi.nlm.nih.gov/27322544/	(Goldmann et al. 2016)

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 high-resolution recombination map inferred from over 126,000 meioses in Icelandic pedigrees (Halldorsson et al. 2019). Briefly, for each position in the genome, we interpolated the recombination rate (excluding regions that fell outside the span of regions provided in (Halldorsson et al. 2019)). We then sorted all genomic sites by interpolated recombination rate and divided the genome into three discrete bins with recombination rates of (0, 2.6×10^{-4}], (2.6×10^{-4}, 6.3×10^{-2}], and ($\times10^{-2}$, ∞) 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.

Archaic regions: To explore the source of differences across human populations and their relationship to archaic hominins, we annotated variants in modern humans according to whether

771 they shared the derived allele with Neanderthals (Vindija or Altai) or Denisovans (Altai), where 0
772 and 1 denote ancestral and derived alleles respectively. We generated three sets of variants:
773 ND10: Neanderthals have the derived allele, but Denisovans have the ancestral allele. This set
774 includes ~2,500,000 sites (not all of which are segregating variants in contemporary human
775 populations).
776 ND01: Neanderthals have the ancestral allele, but Denisovans have derived allele. This set
777 includes ~2,100,000 sites.
778 ND11: Both Neanderthals and Denisovans both have the derived alleles. This set includes
779 ~15,200,000 sites.
780 ND: Either Neanderthals or Denisovans have a derived allele. This set includes ~19,800,000 sites.
781 ND00: Neither Neanderthals nor Denisovans have a derived allele, while modern humans have
782 derived alleles (i.e., all accessible sites excluding ND sites). Most of these sites have the ancestral
783 allele in archaic hominins, though some derived variants may be missing or of low quality in
784 archaic genomes (thus not annotated in the VCF file).
785
786 Resources
787

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	http://web.corral.tacc.utexas.edu/WGSAdownload/resources/homo_sapiens_ancestor_GRCh38_e86

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., 2021). As a result most singletons are missing in the phased data. 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 the 6-EPO human ancestral genome to identify the most likely ancestral allele for each SNP. We assumed a mutation rate (m) of 1.25×10^{-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). For each population, we inferred the mutation ages by splitting the Relate output genealogies into subtrees for each population 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). For each mutation, we then inferred the upstream and downstream base pair based on the human ancestral genome. We annotated if the variant has a derived allele in Neanderthals and Denisovans using the publicly available genomes for Vindija Neanderthal, Altai Neanderthal and Altai Denisovan genome (see Datasets) and if it was previously identified as in a region that introgressed from Neanderthals (Sankararaman et al. 2014). Unless otherwise stated, we present the results for the commonly accessible regions.

Binning of polymorphisms based on mutation age

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 7; 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).

Sequence contexts related to COSMIC mutational signatures SBS7 and SBS11

We downloaded loadings of the single base substitution (SBS) reference signatures on the 96 trinucleotide mutation types from COSMIC website (v3.2, GRCh38; link provided under *Resources*). We found that both SBS7a/b and SBS11 consist of nearly exclusive C>T mutations, with 86.7% mutations caused by SBS7a/b concentrated in YCN contexts while 70.0% SBS11 mutations are in NCY contexts, where Y represents pyrimidine (i.e., C or T) and N represents any base. Therefore, as proxies for mutations potentially affected by SBS7a/b and SBS11, we removed C>T mutations in YCN and NCY contexts in analysis corresponding to Figure 2B.

Quantification of parental age effects on DNM counts and ratios

We used a model-based approach to quantify the effects of paternal and maternal ages jointly by leveraging information from all phased and unphased DNMs. In short, as described in (Gao et al. 2019), we modeled the expected number of DNMs inherited from a parent as a linear function of parental age at conception, and assumed that the observed number of DNMs follows a Poisson distribution. Using a maximum likelihood approach, we estimated the sex-specific slopes and intercepts (at age zero) for each mutation ratio. Confidence intervals of the slopes and intercepts were assessed by bootstrap resampling of trios. With these estimated parental age effects, we then predicted the expected count of each mutation type and the pairwise ratios under given combinations of maternal and paternal ages, such as shown in the left panel of Figure 3.

For analysis corresponding to Figures 3 and 4, we inferred the parental age effects based on a DNM dataset from 2,976 Icelandic trios (Halldorsson et al. 2019). Five trios have exceedingly large numbers of DNMs given the parental ages (Proband IDs: 24496, 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).

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.

Funding

ZG was supported by a Sloan Research Fellowship. PM was supported by a Sloan Research Fellowship, the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics and NIH R35GM142978. 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 spectra of human polymorphisms stratified by allele age. The ancestral allele of each SNP is determined based on the human ancestral allele, and the allele age inferred by 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 Bejing, China; and JPT: Japanese in Tokyo, Japan.

Figure supplement 2. Mutation spectrum of SNPs found in CEU in comparison to that of *de novo* mutations identified in trios from Iceland. "DNM 2019" denotes DNM data from 2,879 Icelandic trios (Halldorsson et al. 2019), from which five trios with exceedingly large numbers of DNMs and 92 trios with maternal age above 40 were excluded. "DNM 2017" denotes DNM data from 1,475 Icelandic trios from (Jónsson et al. 2017), excluding 73 trios with maternal age above 40. All polymorphisms and DNMs were filtered to the commonly accessible region.

Figure supplement 3. **Mutation spectra 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. We show the mean and the distribution of B-scores (measures the reduction in nucleotide diversity levels compared to neutral expectation (McVicker et al. 2009) (0-high constraint; 1000-low contraint) for different mutation types.

Figure supplement 5. **Mutation spectra of human polymorphisms in genomic regions with weak background selection (B-score>800)**. The box represents the interquartile range, with the centerline showing the median value.

Figure supplement 6. Fractions of S>S, S>W, W>S, and W>W mutations in variants stratified by derived frequency (A) and allele age (B). The enrichment of weak (W)> strong (S) mutations and depletion of S>W mutations in variants with higher derived allele frequencies and old allele ages support profound effects of GC-biased gene conversion (gBGC) on human polymorphisms and biases in allele age dating by Relate, which ignores gBGC.

Figure supplement 7. Mutation spectra of human polymorphisms stratified by allele age 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 2—figure supplements.

Figure supplement 1. **Pairwise polymorphism ratios in three additional populations of African (LWK), European (TSI), and East Asian (JPT) ancestry in the 1000 Genomes project.** 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 its figure supplements, this analysis was performed on all non-singleton variants, including unphased SNPs and those undated by Relate. Singletons were removed because they showed very different mutation ratios compared to all other variants, indicating higher false positive rates. 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. 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 discernible in low-frequency variants (DAF<5%) but weak in high frequency variants, possibly because most of those ancient variants are segregating at low frequencies.

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. Replication of the T>C/T>G signal with alternative derived allele polarization and region stratification by ancestry of reference genome. (A) The ancestral alleles polarized to the chimpanzee reference allele. (B) Genomic regions are stratified based on ancestry of human reference genome (African or European). Results are 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 8. **T>C/T>G ratio in YRI, CEU and CHB excluding variants for which the derived alleles are found in Denisovans (A), Neanderthals (B), and both (C).** The absence of significant signals in non-ND11 variants (C) is not purely due to a reduction in sample size, as the set of ND11 variants is much smaller but shows highly significant differences between YRI and the other two populations (Figure 2D).

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. Pairwise polymorphism ratios in YRI, CEU, and CHB in commonly accessible regions excluding variants shared with Neanderthals and/or Denisovans.

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. All results are based on variants in the commonly accessible regions, excluding those with derived alleles observed in either Neanderthal or Denisovan or both.

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. (TBD add goldmann 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 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:
https://www.dropbox.com/s/552lhlsm5tplp0i/MutSpecGentime2022_SourceData.zip?dl=0

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

Source data 3: 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 excluding archaic sites; *n*=500 replicates).

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

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Aggarwala, Varun, and Benjamin F. Voight. 2016. “An Expanded Sequence Context Model Broadly Explains Variability in Polymorphism Levels across the Human Genome.” *Nature Genetics* 48 (4): 349–55.

Aikens, Rachael C., Kelsey E. Johnson, and Benjamin F. Voight. 2019. “Signals of Variation in Human Mutation Rate at Multiple Levels of Sequence Context.” *Molecular Biology and Evolution* 36 (5): 955–65.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. “The Repertoire of Mutational Signatures in Human Cancer.” *Nature* 578 (7793): 94–101.

Anderson-Trocmé, Luke, Rick Farouni, Mathieu Bourgey, Yoichiro Kamatani, Koichiro Higasa, Jeong-Sun Seo, Changhoon Kim, Fumihiko Matsuda, and Simon Gravel. 2020. “Legacy Data Confound Genomics Studies.” *Molecular Biology and Evolution* 37 (1): 2–10.

Bergeron, Lucie A., Søren Besenbacher, Tychele N. Turner, Cyril J. Versoza, Richard J. Wang, Alivia Lee Price, Ellie Armstrong, et al. 2021. “Mutationathon: Towards Standardization in Estimates of Pedigree-Based Germline Mutation Rates.” *bioRxiv*. https://doi.org/10.1101/2021.08.30.458162.

Byrska-Bishop, Marta, Uday S. Evani, Xuefang Zhao, Anna O. Basile, Haley J. Abel, Allison A. Regier, André Corvelo, et al. 2021. “High Coverage Whole Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios.” https://doi.org/10.1101/2021.02.06.430068.

Carlson, Jedidiah, Adam E. Locke, Matthew Flickinger, Matthew Zawistowski, Shawn Levy, Richard M. Myers, Michael Boehnke, et al. 2018. “Extremely Rare Variants Reveal Patterns of Germline Mutation Rate Heterogeneity in Humans.” *Nature Communications* 9 (1): 3753.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. “The Effect of Deleterious Mutations on Neutral Molecular Variation.” *Genetics* 134 (4): 1289–1303.

Chen, Lu, Aaron B. Wolf, Wenqing Fu, Liming Li, and Joshua M. Akey. 2020. “Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals.” *Cell* 180 (4): 677–87.e16.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58.

DeWitt, William S., Kameron Decker Harris, Aaron P. Ragsdale, and Kelley Harris. 2021. “Nonparametric Coalescent Inference of Mutation Spectrum History and Demography.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (21). https://doi.org/10.1073/pnas.2013798118.

Duncan, B. K., and J. H. Miller. 1980. “Mutagenic Deamination of Cytosine Residues in DNA.” *Nature* 287 (5782): 560–61.

Duret, Laurent, and Nicolas Galtier. 2009. “Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes.” *Annual Review of Genomics and Human Genetics* 10: 285–311.

Durvasula, Arun, and Sriram Sankararaman. 2020. “Recovering Signals of Ghost Archaic Introgression in African Populations.” *Science Advances* 6 (7): eaax5097.

Fenner, Jack N. 2005. “Cross-Cultural Estimation of the Human Generation Interval for Use in Genetics-Based Population Divergence Studies.” *American Journal of Physical Anthropology* 128 (2): 415–23.

Gao, Ziyue, Priya Moorjani, Thomas A. Sasani, Brent S. Pedersen, Aaron R. Quinlan, Lynn B. Jorde, Guy Amster, and Molly Przeworski. 2019. “Overlooked Roles of DNA Damage and Maternal Age in Generating Human Germline Mutations.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (19): 9491–9500.

Goldberg, Michael E., and Kelley Harris. 2022. “Mutational Signatures of Replication Timing and Epigenetic Modification Persist through the Global Divergence of Mutation Spectra across the Great Ape Phylogeny.” *Genome Biology and Evolution* 14 (1). https://doi.org/10.1093/gbe/evab104.

Goldmann, Jakob M., Vladimir B. Seplyarskiy, Wendy S. W. Wong, Thierry Vilboux, Pieter B. Neerincx, Dale L. Bodian, Benjamin D. Solomon, et al. 2018. “Germline de Novo Mutation Clusters Arise during Oocyte Aging in Genomic Regions with High Double-Strand-Break Incidence.” *Nature Genetics* 50 (4): 487–92.

Goldmann, Jakob M., Wendy S. W. Wong, Michele Pinelli, Terry Farrah, Dale Bodian, Anna B. Stittrich, Gustavo Glusman, et al. 2016. “Parent-of-Origin-Specific Signatures of de Novo Mutations.” *Nature Genetics* 48 (8): 935–39.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. “A Draft Sequence of the Neandertal Genome.” *Science* 328 (5979): 710–22.

Halldorsson, Bjarni V., Gunnar Palsson, Olafur A. Stefansson, Hakon Jonsson, Marteinn T. Hardarson, Hannes P. Eggertsson, Bjarni Gunnarsson, et al. 2019. “Characterizing Mutagenic Effects of Recombination through a Sequence-Level Genetic Map.” *Science*. https://doi.org/10.1126/science.aau1043.

Hammer, Michael F., August E. Woerner, Fernando L. Mendez, Joseph C. Watkins, and Jeffrey D. Wall. 2011. “Genetic Evidence for Archaic Admixture in Africa.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (37): 15123–28.

Harpak, Arbel, Anand Bhaskar, and Jonathan K. Pritchard. 2016. “Mutation Rate Variation Is a Primary Determinant of the Distribution of Allele Frequencies in Humans.” *PLoS Genetics* 12 (12): e1006489.

Harris, Kelley. 2015. “Evidence for Recent, Population-Specific Evolution of the Human Mutation Rate.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (11): 3439–44.

Harris, Kelley, and Jonathan K. Pritchard. 2017. “Rapid Evolution of the Human Mutation Spectrum.” *eLife* 6 (April). https://doi.org/10.7554/eLife.24284.

Hwang, Dick G., and Phil Green. 2004. “Bayesian Markov Chain Monte Carlo Sequence Analysis Reveals Varying Neutral Substitution Patterns in Mammalian Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (39): 13994–1.

Jiang, Pengyao, Anja R. Ollodart, Vidha Sudhesh, Alan J. Herr, Maitreya J. Dunham, and Kelley Harris. 2021. “A Modified Fluctuation Assay Reveals a Natural Mutator Phenotype That Drives Mutation Spectrum Variation within Saccharomyces Cerevisiae.” *eLife* 10 (September). https://doi.org/10.7554/eLife.68285.

1182 Jónsson, Hákon, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdottir, Florian Zink, Eirikur
1183 Hjartarson, Marteinn T. Hardarson, et al. 2017. “Parental Influence on Human Germline de
1184 Novo Mutations in 1,548 Trios from Iceland.” *Nature* 549 (7673): 519–22.
1185 Kaplanis, Joanna, Benjamin Ide, Rashesh Sanghvi, Matthew Neville, Petr Danecek, Tim
1186 Coorens, Elena Prigmore, et al. 2021. “Genetic and Chemotherapeutic Causes of Germline
1187 Hypermutation.” *bioRxiv*, 2021–2006.
1188 Kimura, M. 1969. “The Number of Heterozygous Nucleotide Sites Maintained in a Finite
1189 Population due to Steady Flux of Mutations.” *Genetics* 61 (4): 893–903.
1190 Kong, Augustine, Michael L. Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli
1191 Magnusson, Sigurjon A. Gudjonsson, et al. 2012. “Rate of de Novo Mutations and the
1192 Importance of Father’s Age to Disease Risk.” *Nature*. https://doi.org/10.1038/nature11396.
1193 Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al.
1194 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–
1195 921.
1196 Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy
1197 Fennell, Anne H. O’Donnell-Luria, et al. 2016. “Analysis of Protein-Coding Genetic Variation
1198 in 60,706 Humans.” *Nature* 536 (7616): 285–91.
1199 Macià, Moisès Coll, Laurits Skov, Benjamin Marco Peter, and Mikkel Heide Schierup. 2021.
1200 “Different Historical Generation Intervals in Human Populations Inferred from Neanderthal
1201 Fragment Lengths and Mutation Signatures.” *Nature Communications*.
1202 https://doi.org/10.1038/s41467-021-25524-4.
1203 Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo,
1204 Mengyao Zhao, et al. 2016. “The Simons Genome Diversity Project: 300 Genomes from
1205 142 Diverse Populations.” *Nature* 538 (7624): 201–6.
1206 Mathieson, Iain, and Gil McVean. 2014. “Demography and the Age of Rare Variants.” *PLoS*
1207 *Genetics* 10 (8): e1004528.
1208 Mathieson, Iain, and David Reich. 2017. “Differences in the Rare Variant Spectrum among
1209 Human Populations.” *PLoS Genetics* 13 (2): e1006581.
1210 McVicker, Graham, David Gordon, Colleen Davis, and Phil Green. 2009. “Widespread Genomic
1211 Signatures of Natural Selection in Hominid Evolution.” *PLoS Genetics* 5 (5): e1000471.
1212 Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan
1213 Mallick, Joshua G. Schraiber, et al. 2012. “A High-Coverage Genome Sequence from an
1214 Archaic Denisovan Individual.” *Science* 338 (6104): 222–26.
1215 Milligan, William R., Guy Amster, and Guy Sella. 2021. “The Impact of Genetic Modifiers on
1216 Variation in Germline Mutation Rates within and among Human Populations.” *bioRxiv*.
1217 https://doi.org/10.1101/2021.08.25.457718.
1218 Monroe, J. Grey, Thanvi Srikant, Pablo Carbonell-Bejerano, Claude Becker, Mariele Lensink,
1219 Moises Exposito-Alonso, Marie Klein, et al. 2022. “Mutation Bias Reflects Natural Selection
1220 in Arabidopsis Thaliana.” *Nature* 602 (7895): 101–5.
1221 Moorjani, Priya, Carlos Eduardo G. Amorim, Peter F. Arndt, and Molly Przeworski. 2016.
1222 “Variation in the Molecular Clock of Primates.” *Proceedings of the National Academy of*
1223 *Sciences of the United States of America* 113 (38): 10607–12.
1224 Moorjani, Priya, Sriram Sankararaman, Qiaomei Fu, Molly Przeworski, Nick Patterson, and
1225 David Reich. 2016. “A Genetic Method for Dating Ancient Genomes Provides a Direct
1226 Estimate of Human Generation Interval in the Last 45,000 Years.” *Proceedings of the*
1227 *National Academy of Sciences of the United States of America* 113 (20): 5652–57.
1228 Murphy, David, Eyal Elyashiv, Guy Amster, and Guy Sella. 2021. “Broad-Scale Variation in
1229 Human Genetic Diversity Levels Is Predicted by Purifying Selection on Coding and Non-
1230 Coding Elements.” https://doi.org/10.1101/2021.07.02.450762.
1231 Narasimhan, Vagheesh M., Raheleh Rahbari, Aylwyn Scally, Arthur Wuster, Dan Mason, Yali
1232 Xue, John Wright, et al. 2017. “Estimating the Human Mutation Rate from Autozygous

1233 Segments Reveals Population Differences in Human Mutational Processes." *Nature*
1234 *Communications* 8 (1): 303.
1235 Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna
1236 Sawyer, Anja Heinze, et al. 2014. "The Complete Genome Sequence of a Neanderthal from
1237 the Altai Mountains." *Nature* 505 (7481): 43–49.
1238 Ragsdale, Aaron P., and Simon Gravel. 2019. "Models of Archaic Admixture and Recent History
1239 from Two-Locus Statistics." *PLOS Genetics*. https://doi.org/10.1371/journal.pgen.1008204.
1240 Ragsdale, Aaron P., Timothy D. Weaver, Elizabeth G. Atkinson, Eileen Hoal, Marlo Möller,
1241 Brenna M. Henn, and Simon Gravel. 2022. "A Weakly Structured Stem for Human Origins
1242 in Africa." *bioRxiv*. https://doi.org/10.1101/2022.03.23.485528.
1243 Sankararaman, Sriram, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante
1244 Pääbo, Nick Patterson, and David Reich. 2014. "The Genomic Landscape of Neanderthal
1245 Ancestry in Present-Day Humans." *Nature* 507 (7492): 354–57.
1246 Sasani, Thomas A., David G. Ashbrook, Annabel C. Beichman, Lu Lu, Abraham A. Palmer,
1247 Robert W. Williams, Jonathan K. Pritchard, and Kelley Harris. 2021. "A Natural Mutator
1248 Allele Shapes Mutation Spectrum Variation in Mice." *bioRxiv*. bioRxiv.
1249 https://doi.org/10.1101/2021.03.12.435196.
1250 Schiffels, Stephan, and Richard Durbin. 2014. "Inferring Human Population Size and Separation
1251 History from Multiple Genome Sequences." *Nature Genetics* 46 (8): 919–25.
1252 Seoighe, Cathal, and Aylwyn Scally. 2017. "Inference of Candidate Germline Mutator Loci in
1253 Humans from Genome-Wide Haplotype Data." *PLoS Genetics* 13 (1): e1006549.
1254 Seplyarskiy, Vladimir B., Ruslan A. Soldatov, Evan Koch, Ryan J. McGinty, Jakob M.
1255 Goldmann, Ryan D. Hernandez, Kathleen Barnes, et al. 2021. "Population Sequencing
1256 Data Reveal a Compendium of Mutational Processes in the Human Germ Line." *Science*
1257 373 (6558): 1030–35.
1258 Skov, Laurits, Moisès Coll Macià, Garðar Sveinbjörnsson, Fabrizio Mafessoni, Elise A. Lucotte,
1259 Margret S. Einarsdóttir, Hakon Jonsson, et al. 2020. "The Nature of Neanderthal
1260 Introgression Revealed by 27,566 Icelandic Genomes." *Nature* 582 (7810): 78–83.
1261 Speidel, Leo, Marie Forest, Sinan Shi, and Simon R. Myers. 2019. "A Method for Genome-Wide
1262 Genealogy Estimation for Thousands of Samples." *Nature Genetics* 51 (9): 1321–29.
1263 Wakeley, John. 2010. "Natural Selection and Coalescent Theory." *Evolution since Darwin: The*
1264 *First* 150: 119–49.
1265 Wang, Richard J., Samer I. Al-Saffar, Jeffrey Rogers, and Matthew W. Hahn. 2021. "Human
1266 Generation Times across the Past 250,000 Years." *bioRxiv*.
1267 https://doi.org/10.1101/2021.09.07.459333.
1268 Wu, C. I., and W. H. Li. 1985. "Evidence for Higher Rates of Nucleotide Substitution in Rodents
1269 than in Man." *Proceedings of the National Academy of Sciences of the United States of*
1270 *America* 82 (6): 1741–45.

1271