Rapid evolution of the human mutation spectrum.

Kelley Harris[1] and Jonathan K. Pritchard[1,2,3]

[1]Department of Genetics, Stanford University

[2]Department of Biology, Stanford University

[3]Howard Hughes Medical Institute, Stanford University

Correspondence: kelleyh@stanford.edu, pritch@stanford.edu.

DNA is a remarkably precise medium for copying and storing biological information, with a mutation rate in humans of about 1×10^{-8} per base pair per generation. This extraordinary fidelity results from the combined action of hundreds of genes involved in DNA replication and proofreading, and repair of spontaneous damage. Recent studies of cancer have shown that mutation of specific genes often leads to characteristic mutational "signatures"–i.e., increased mutation rates within particular sequence contexts. We therefore hypothesized that more subtle variation in replication or repair genes within natural populations might also lead to differences in mutational signatures. As a proxy for mutational input, we examined SNV variation across human and other great ape populations. Remarkably we found that mutational spectra differ substantially among species, human continental groups and even, in some cases, between closely-related populations. Closer examination of one such signal, an increased rate of TCC$\rightarrow$TTC mutations reported previously in Europeans, indicates a burst of mutations from about 15,000 to 2,000 years ago, perhaps due to the appearance, drift, and ultimate elimination of a genetic modifier of mutation rate.

Main Text

Germline mutations provide the raw material for evolution, but also generate genetic load and inherited disease. Indeed, the vast majority of mutations that affect fitness are deleterious, and hence biological systems have evolved elaborate mechanisms for accurate DNA replication and repair of diverse types of spontaneous damage. Due to the combined action of hundreds of genes, mutation rates are extremely low–in humans, about 1 point mutation per 100MB or about 60 genome-wide per generation [1,2].

While the precise roles of most of the relevant genes have not been fully elucidated, research on somatic mutations in cancer has shown that defects in particular genes can lead to increased mutation rates within very specific sequence contexts [3,4]. For example, mutations in the proofreading exonuclease domain of DNA polymerase ϵ cause TCT→TAT and TCG→TTG mutations on the leading DNA strand [5]. Mutational shifts of this kind have been referred to as "mutational signatures". Specific signatures may also be caused by nongenetic factors such as chemical mutagens or UV damage.

Together, these observations imply a high degree of specialization of individual genes involved in DNA proofreading and repair. While the repair system has evolved to be extremely accurate overall, theory suggests that in such a system, natural selection may have limited ability to fine-tune the efficacy of individual genes [6,7]. If a variant in a repair gene increases or decreases the overall mutation rate by a small amount–for example, only in a very specific sequence context–then the net effect on fitness may fall below the threshold at which natural selection is effective. (Drift tends to dominate selection when the change in fitness is less than the inverse of effective population size). The limits of selection on mutation rate modifiers are especially acute in recombining organisms such as humans because a variant that increases the mutation rate can recombine away from deleterious mutations it generates elsewhere in the genome.

Given these theoretical predictions, we hypothesized that there may be substantial scope for modifiers of mutation rates to segregate within human populations, or between closely related species. If these affect specific pathways of proofreading or repair, then we may expect shifts in the abundance of mutations within particular sequence contexts. Indeed, one of us (K.H.) has recently identified a candidate signal of this type, namely an increase in TCC→TTC transitions in Europeans, relative to other populations [8]. A recent preprint has replicated that result, and reported an additional mutational signature enriched in South American populations [9]. Here we show that mutation spectrum change is much more widespread than these initial studies suggested: although the TCC→TTC rate increase in Europeans was unusually dramatic, smaller-scale changes are so commonplace that almost every great ape species and human continental group has its own distinctive mutational spectrum.

Results

To investigate the mutational processes in different human populations, we classified all single nucleotide variants (SNVs) in the 1000 Genomes Phase 3 data [10] in terms of their ancestral allele, derived allele, and

5' and 3' flanking nucleotides, collapsing strand-complements together to obtain 96 SNV categories. Since the detection of singletons may vary across samples, and because some singletons may result from cell-line or somatic mutations, we only considered variants seen in more than one copy. Using this scheme, we calculated the distribution of derived SNVs carried by each Phase 3 individual. We used this as a proxy for the mutational input spectrum in the ancestors of each individual, excluding annotated repeats and PhyloP conserved regions [11].

To explore global patterns of the mutation spectrum, we experimented with performing principal component analysis (PCA) in which each individual was characterized simply by the fraction of their derived alleles in each of the 96 SNV categories (Fig. 1A). PCA is commonly performed on individual-level genotypes, in which case the PCs are usually highly correlated with geography [12]. Although the triplet mutation spectrum is an extremely compressed summary statistic compared to typical genotype arrays, we found that it contains sufficient information to reliably classify individuals by their continent of origin. The first principal component separated Africans from non-Africans, and the second separated Europeans from East Asians, with South Asians and admixed native Americans (Fig. S2) appearing intermediate between the two.

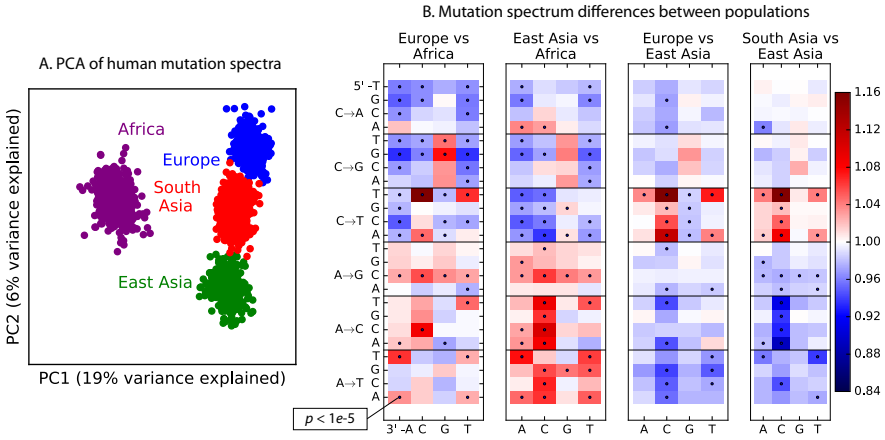

Figure 1 Global patterns of variation in SNV spectra. A. *Principal Component Analysis of individuals according to the fraction of derived alleles that each individual carries in each of 96 mutational types.* **B.** *Heatmaps showing, for pairs of continental groups, the ratio of the proportions of SNVs in each of the 96 mutational types. Each block corresponds to one mutation type; within blocks, rows indicate the 5' nucleotide, and columns indicate the 3' nucleotide. Red colors indicate a greater fraction of a given mutation type in the first-listed group relative to the second. Points indicate significant contrasts at* $p < 10^{-5}$. *See Figure S1 for comparisons between additional population pairs.*

Remarkably, we found that the mutation spectrum differences among continental groups are composed of small shifts in the abundance of many different mutation types (Fig. 1B). For example, comparing

Africans and Europeans, 43 of the 96 mutation types are significant at a $p < 10^{-5}$ threshold using a forward variable selection procedure. The previously described TCC→TTC signature partially drives the difference between Europeans and the other groups, but most other shifts are smaller in magnitude and appear to be spread over more diffuse sets of related mutation types. East Asians have excess A→T transversions in most sequence contexts, while Africans have proportionally more C→X mutations relative to A→X mutations.

One possible concern is that batch effects or other sequencing artifacts might contribute to differences in mutational spectra across individuals. Therefore we replicated our analysis using 201 genomes from the Simons Genome Diversity Project [13]. These genomes were sequenced at high coverage, independently from 1000 Genomes, using an almost non-overlapping panel of samples. We found extremely strong agreement between the mutational shifts in the two data sets (Fig. S3). Among the different continental comparisons, the median correlation between the pairwise enrichment/depletion ratios in the two data sets was 76% (Fig. S4).

These widespread differences may be footprints of allele frequency shifts affecting different mutator alleles. But in principle, other genetic and non-genetic processes may also impact the observed mutational spectrum. First, biased gene conversion (BGC) tends to favor C/G alleles over A/T, and BGC is potentially more efficient in populations of large effective size compared to populations of smaller effective size [14]. However, despite the bottlenecks that are known to have affected Eurasian diversity, there is no clear trend of an increased fraction of C/G→A/T relative to A/T→C/G in non-Africans vs. Africans, or with distance from Africa (Fig. S5), and previous studies have also found little evidence for a strong genome-wide effect of BGC on the mutational spectrum in humans and great apes [15,16].

It is also known that shifts in generation time or other life-history traits may affect mutational spectra, particularly for CpG transitions [17,18]. Most CpG transitions result from spontaneous methyl-cytosine deamination as opposed to errors in DNA replication. Hence the rate of CpG transitions is less affected by generation time than other mutations [19–21]. We observe that Europeans have a lower fraction of CpG variants compared to Africans, east Asians and south Asians (Fig. 1B), consistent with a recent report of a lower rate of *de novo* CCG→CTG mutations in European individuals compared to Pakistanis [22]. Such a pattern may be consistent with a shorter average generation time in Europeans [20], though it is unclear that a plausible shift in generation-time could produce such a large effect. Apart from this, the other patterns evident in Figure 1 do not seem explainable by known processes.

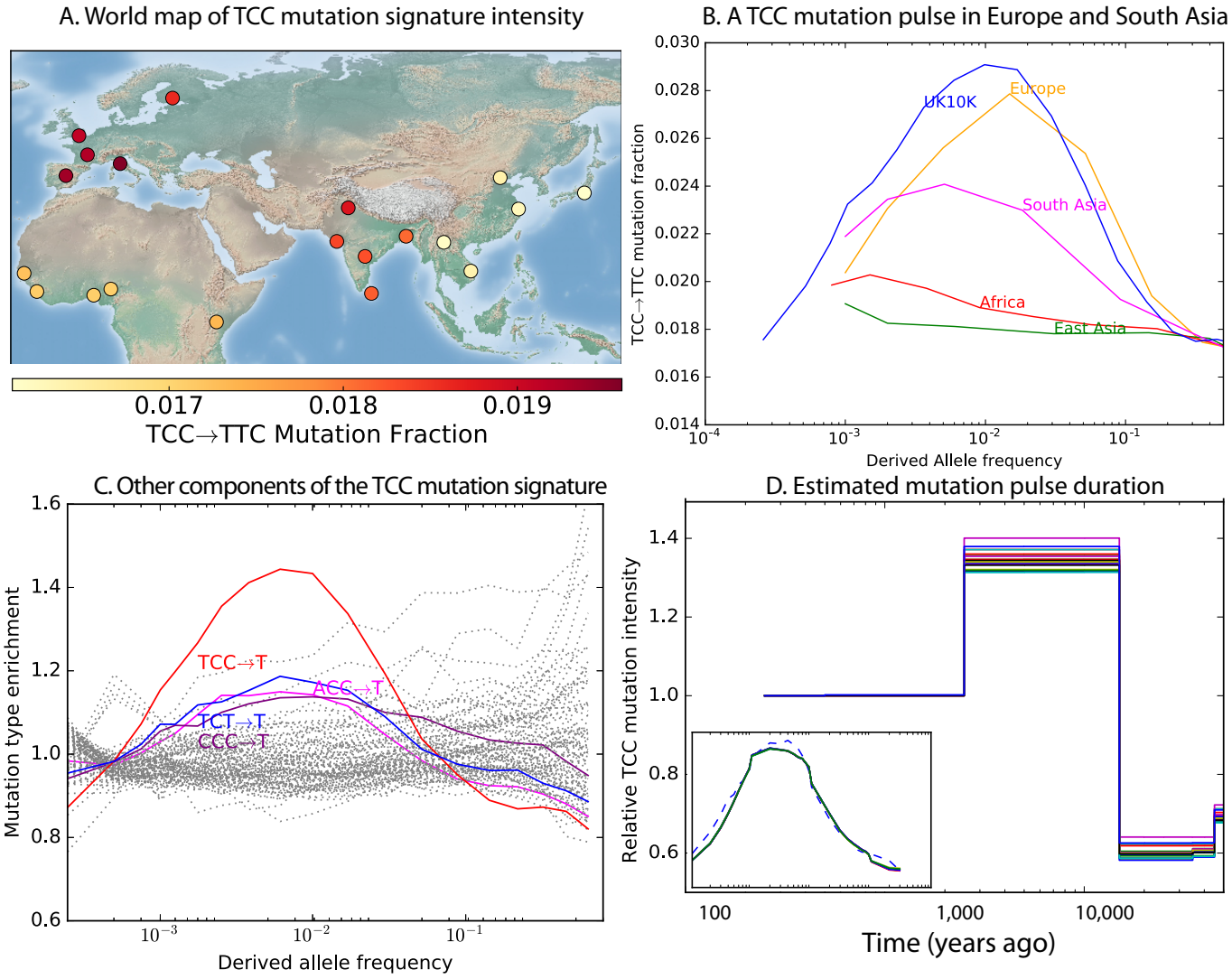

Figure 2 Geographic distribution and age of the TCC mutation pulse. **(A)** *Observed frequencies of TCC→TTC variants in 1000 Genomes populations.* **(B)** *Fraction of TCC→TTC variants as a function of allele frequency in different samples indicates that these peak around 1%. In the UK10K data, which has the largest sample size, the peak occurs at 0.6% allele frequency* **(C)** *Other enriched C→T mutations with similar context also peak at 0.6% frequency in UK10K.* **(D)** *A population genetic model supports a pulse of TCC→TTC mutations from 15,000–2,000 years ago. Inset shows the observed and predicted frequency distributions of this mutation under the inferred model.*

The most significant signal in Fig. 1B is for TCC→TTC mutations, which are highly enriched in Europeans and, to a lesser extent, in South Asians (Fig. 2A; $p < 1 \times 10^{-300}$ for Europe vs. Africa). To investigate when this mutational increase occurred, we plotted the allele frequency spectrum of this mutation type in data from 1000 Genomes, and from the larger UK10K sample [23]. As expected for a signal that is primarily European, we found particular enrichment of these mutations at low frequencies.

But surprisingly, the enrichment peaks around 1% frequency, and there is practically no enrichment among the very lowest frequency variants (Figures 2B and S6). In the larger UK10K dataset, signal peaks at 0.6% frequency. C→T mutations on other backgrounds, namely within TCT, CCC and ACC contexts, are also enriched in Europe, and show a similar peak around 0.6% frequency and declining among rarer variants (Fig. 2C). This suggests that these four mutation types comprise the signature of a single mutational pulse that is no longer active. No other mutation types show such a pulse-like distribution in UK10K, though several types show evidence of monotonic rate change over time (Figures S7 and S8).

We used the enrichment of TCC→TTC mutations as a function of allele frequency to estimate when this mutation pulse was active. Assuming a simple piecewise-constant model, we infer that the rate of TCC→TTC mutations increased dramatically ∼15,000 years ago and decreased again ∼2,000 years ago. This time-range is consistent with results showing this signal in a pair of prehistoric European samples from 7,000 and 8,000 years ago, respectively [9]. We hypothesize that this mutation pulse may have been caused by a mutator allele that drifted up in frequency starting 15,000 years ago, but that is now rare or absent from present day populations.

Encouraged by these results, we sought to find other signatures of recent mutation pulses. We generated heatmaps and PCA plots of mutation spectrum variation within each continental group, looking for fine-scale differences between closely related populations (Figures S9-S14). In some cases mutational spectra differ even between very closely related populations. For example, one notable signal is apparent in east Asia, where most Japanese individuals cluster separately from most other east Asians (Figures 3A and S12). These individuals carry elevated rates of *AC→*CC, ACA→AAA, and TAT→TTT mutations. This signature appears to be present in only a handful of Chinese individuals, and no Kinh or Dai individuals. As seen for the European TCC mutation, the enrichment of these mutation types peaks at low frequencies, i.e., ∼1%.

PCA reveals relatively little fine-scale structure within the mutational spectra of Europeans or South Asians (Figures S14, S13). However, Africans exhibit significant substructure (Fig. S11), and the Luhya have a particularly distinctive mutational spectrum. Unexpectedly, a closer examination of PC loadings reveals that the Luhya outliers are enriched for the same mutational signature identified in the Japanese. Even in Europeans and South Asians, the first PC is heavily weighted toward *AC→*CC, ACA→AAA, and TAT→TTT, although this signature explains less of the mutation spectrum variance within these more homogeneous populations. The sharing of this signature may suggest either parallel increases of a shared mutation modifier, or a shared aspect of environment or life history that affects the mutation spectrum.

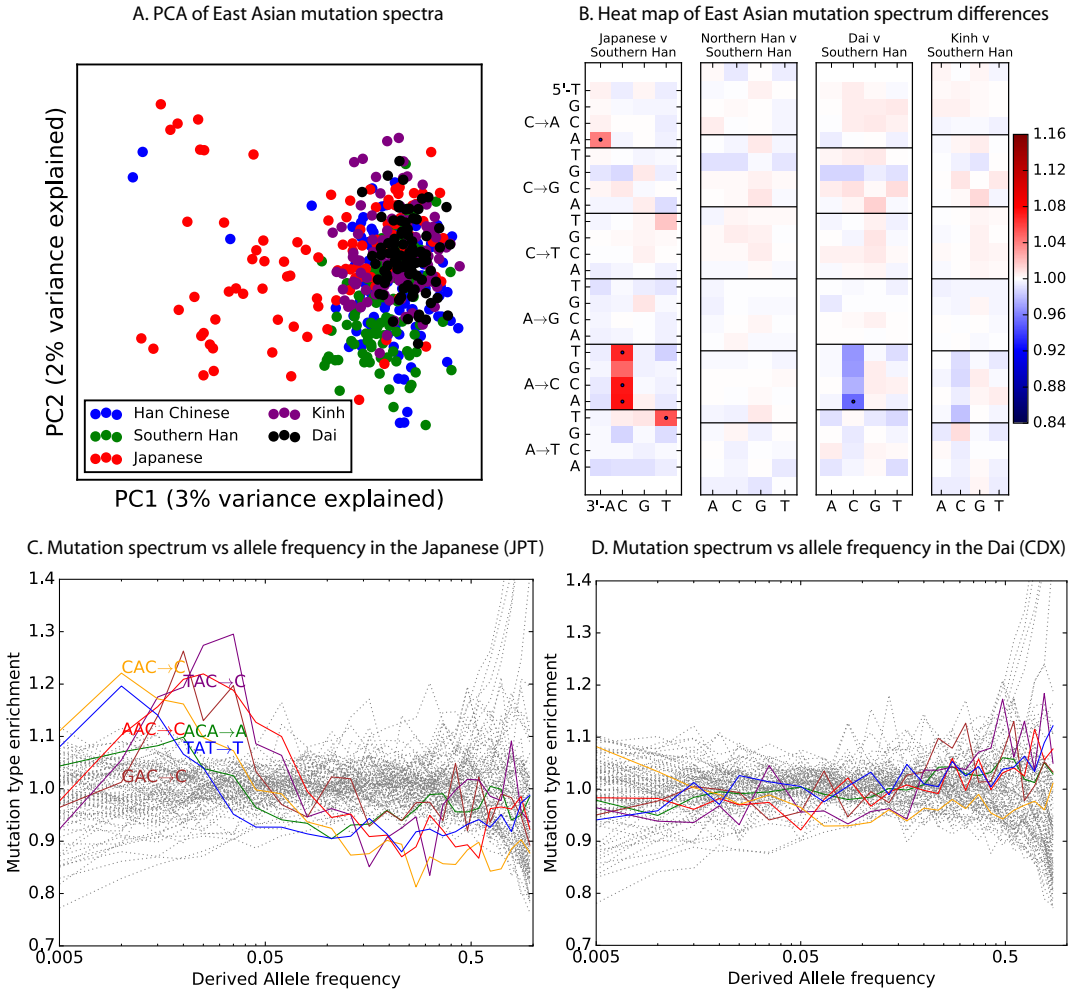

Figure 3 **Mutational variation among east Asian populations.** **(A)** *PCA of east Asian samples from 1000 Genomes, based on the relative proportions of each of the 96 mutational types.* **(B)** *Heatmaps showing, for pairs of east Asian samples, the ratio of the proportions of SNVs in each of the 96 mutational types. Points indicate significant contrasts at* $p < 10{-}5$. **(C)** *and* **(D)** *Relative enrichment of each mutational type in Japanese and Dai, respectively as a function of allele frequency. Six mutation types that are enriched in JPT are indicated. Populations: CDX=Dai, CHB=Han (Beijing); CHS=Han (south China); KHV=Kinh; JPT=Japanese.*

Together, these results suggest that modifiers of the mutation spectrum may segregate in human populations. It would be natural to perform genome-wide mapping for modifiers, although measurements of mutation spectrum in individual families are highly imprecise given the small number of de novo mutations per zygote. As an alternative, we explored the use of SNVs to test for mutator function at a candidate locus. Although a mutator allele would generate mutations genome-wide, recombination would quickly randomize the resulting variants with respect to genotypes at the mutator. However, at the mutator locus

itself, there should be a specific increase of the relevant mutation types on haplotypes that carry the mutator, compared to those that do not. A recent study reported that two common germline variants affect both cancer risk and the somatic mutational spectrum by altering regulation of APOBEC DNA-editing enzymes [24]. Given that APOBEC activity appears to cause germline mutations as well [25], we hypothesized that these two variants might also be germline mutators. Based on previous descriptions of the APOBEC signature [3,26], we classified TC→TT and TC→TG as potentially APOBEC-associated, all other mutations as non-APOBEC-associated, and tested for enrichment of APOBEC-associated variants on the candidate mutator haplotypes. We found modest enrichment of APOBEC mutations on the predicted mutator allele ($p = 0.03$) for SNP rs1014971. This associated variant lies in the APOBEC3B regulatory region and increases the risk of bladder cancer.

Finally, given our finding of extensive fine-scale variation in mutational spectra between human populations, we hypothesized that mutational variation between species is likely to be even greater. To compare the mutation spectra of the great apes in more detail, we obtained SNV data from the Great Ape Diversity Panel, which includes 78 whole genome sequences from six great ape species including human [27]. Overall, we find dramatic variation in mutational spectra among the great ape species (Figures 4 and S15). As noted previously, one major trend is an increase in CpG proportion among the species closest to human, possibly reflecting lengthening generation time along the human lineage [16], consistent with previous indications that species closely related to humans have lower mutation rates than more distant species[28–30].

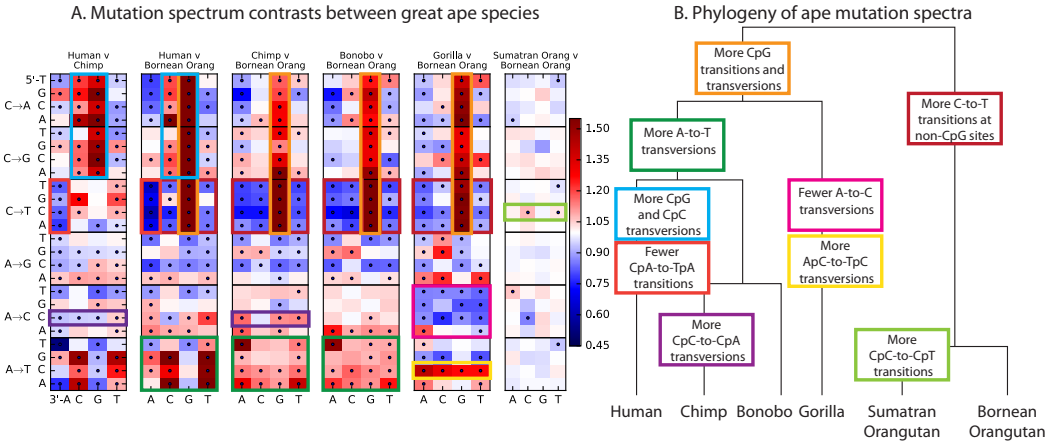

Figure 4 Mutational differences among the great apes. **(A)** *Relative abundance of SNV types in 5 ape species compared to Bornean Orangutan; data from [27]. Boxes indicate labels in (B). For additional comparisons see Fig. S15.* **(B)** *Schematic phylogeny of the great apes highlighting notable changes in SNV abundances.*

However, most other differences are not obviously related to known processes such as biased gene conversion and generation time change. The A→T mutation rate appears to have sped up in the common

ancestor of humans, chimpanzees, and bonobos, a change that appears consistent with a mutator variant that was fixed instead of lost. It is unclear whether this ancient A→T speedup is related to the A→T speedup in East Asians. Other mutational signatures appear on only a single branch of the great ape tree, such as a slowdown of A→C mutations in gorillas.

Discussion

In summary, we report here that, mutational spectra differ significantly among closely related human populations, and that they differ greatly among the great ape species. Our work shows that subtle, concerted shifts in the frequencies of many different mutation types are more widespread than dramatic jumps in the rate of single mutation types, although the existence of the European TCC→TTC pulse shows that both modes of evolution do occur [8,9,20].

At this time, we cannot exclude a role for nongenetic factors such as changes in life history or mutagen exposure in driving these signals. However given the sheer diversity of the effects reported here, it seems parsimonious to us to propose that most of this variation is driven by the appearance and drift of genetic modifiers of mutation rate. This situation is perhaps reminiscent of the earlier observation that genome-wide recombination patterns are variable among individuals [31], and ultimate discovery of PRDM9 [32]; although in this case it is unlikely that a single gene is responsible for all signals seen here. As large datasets of de novo mutations become available, it should be possible to map mutator loci genome-wide. In summary, our results suggest the likelihood that mutational modifiers are an important part of the landscape of human genetic variation.

Methods

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

1000 Genomes Phase 3	http://www.1000genomes.org/category/phase-3/
UK10K	http://www.uk10k.org/data-access.html
Simons Genome Diversity Panel	https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/

Human Mutation Spectrum Processing. Mutation spectra were computed using 1000 Genomes Phase 3 SNPs [10] that are biallelic, pass all 1000 Genomes quality filters, and are not adjacent to any N's in the hg19 reference sequence. Ancestral states were assigned using the UCSC Genome Browser alignment of hg19 to the PanTro2 chimpanzee reference genome; SNPs were discarded if neither the reference nor alternate allele matched the chimpanzee reference. To minimize the potential impact of ancestral misidentification errors, SNPs with derived allele frequency higher than 0.98 were discarded. We also filtered out regions annotated as "conserved" based on the 100-way PhyloP conservation score [11], download from http://hgdownload.cse. ucsc.edu/goldenPath/ hg19/phastCons100way/, as well as regions annotated as repeats by RepeatMasker [33], dowloaded from http://hgdownload.cse.ucsc.edu/ goldenpath/hg19/database/nestedRepeats.txt.gz. To be counted as part of the mutation spectrum of population P (which can be either a continental group or a finer-scale population from one city), a SNP should not be a singleton within population P–at least two copies of the ancestral and derived alleles must be present within that population.

An identical approach was used to extract the mutation spectrum of the UK10K ALSPAC panel [23], which is not subdivided into smaller populations. The data were filtered as described in [34]. The filtering procedure performed by Field, et al. reduces the ALSPAC sample size to 1927 individuals.

We also computed mutation spectra of the Simons Genome Diversity Panel (SGDP) populations. [9]. Four of the SGDP populations, West Eurasia, East Asia, South Asia, and Africa, were compared to their direct counterparts in the 1000 Genomes data. Three additional SGDP populations, Central Asia and Siberia, Oceania, and America, had no close 1000 Genomes counterparts and were not analyzed here (although each project contained a panel of people from the Americans, the composition of the American panels was extremely different, with the 1000 Genomes populations being much more admixed with Europeans and Africans). SGDP sites with more than 20% missing data were not utilized. All other

data processing was done the same way described for the 1000 Genomes data.

The following table gives the same size of each population panel, as well as the total number of SNPs segregating in the panel that are used to compute mutation type ratios:

Dataset	Population	Number of individuals	Number of SNPs
1kg	Africa	504	16,870,400
1kg	Europe	503	8,508,040
1kg	East Asia	504	7,895,925
1kg	South Asia	489	9,552,781
SGDP	Africa	45	6,569,658
SGDP	West Eurasia	69	4,201,571
SGDP	East Asia	49	3,312,645
SGDP	South Asia	38	3,449,624

Great Ape Diversity Panel Data Processing. Biallelic great ape SNPs were extracted from the Great Ape Diversity Panel VCF [27], which is aligned to the hg18 human reference sequence. Ancestral states were assigned using the Great Ape Genetic Diversity project annotation, which used the Felsenstein pruning algorithm to assign states to internal nodes in the great ape tree (Fig. S1). For each site, node 18 is assumed to encode the human ancestral state, while node 17 encodes the chimpanzee and bonobo ancestral state, node 19 encodes the gorilla ancestral state, and node 15 encodes the orangutan ancestral state. A SNP was discarded whenever the ancestral node was assigned an uncertain or polymorphic ancestral state. As with the human data, SNPs with derived allele frequency higher than 0.98 were not used, and both repeats and PhyloP-annotated conserved regions were filtered away.

Visual representation of mutation spectra. The mutation type of a SNP is defined in terms of its ancestral allele, its derived allele, and its two immediate 5' and 3' neighbors. Two mutation types are considered equivalent if they are strand-complementary to each other (e.g. ACG→ATG is equivalent to CGT→CAT). This scheme classifies SNPs into 96 different mutation types, each that can be represented with an A or C ancestral allele.

To compute the frequency $f_P(m)$ of SNP m in population P, we count up all SNPs of type m where

the derived allele is present in at least one representative of population P (which can be either a specific population such as YRI or a broader continental group such as AFR). After obtaining this count $C_P(m)$, we define $f_P(m)$ to be the ratio $C_P(m)/\sum_{m'} C_P(m')$, where the sum in the denominator ranges over all 96 mutation types m'. The enrichment of mutation type m in population P_1 relative to population P_2 is defined to be $f_{P_1}(m)/f_{P_2}(m)$; these enrichments are visualized as heat maps in Figures 1B, 3B, and 4A.

To track changes in the mutational spectrum over time, we compute $f_P(m)$ in bins of restricted allele frequency. This involves counting the number of SNPs of type m that are present at frequency ϕ in population P to obtain counts $C_P(m,\phi)$ and frequencies $f_P(m,\phi) = C_P(m,\phi)/\sum_{m'} C_P(m'\phi)$. Deviation of the ratio $f_P(m,\phi)/f_P(m)$ from 1 indicates that the rate of m has fluctuated recently in the history of population P. To make the sampling noise approximately uniform across alleles of different frequencies, alleles of derived count greater than 5 were grouped into approximately log-spaced bins that each contained similar numbers of UK10K SNPs. More precisely, we defined a set of bin endpoints $b_1, b_2, ...$ such that the total number of SNPs ranging in derived allele count between b_i and $b_{i+1} - 1$ is greater than or equal to the number of 5-ton SNPs, while the total number of SNPs ranging in derived allele count from b_i to $b_{i+1} - 2$ is less than the number of 5-ton SNPs.

Significance testing. Let S_i denote the total number of SNPs segregating in population P_i, and let $S_i^{(m)}$ denote the number of SNPs of mutation type m. If mutation type m is more prevalent in population P_1 than in population P_2, a chi-square test provides a natural way of assessing the significance of this difference. As described in [8], this test is performed on the following 2-by-2 contingency table:

$S_1^{(m)}$	$P_1 - S_1^{(m)}$
$S_2^{(m)}$	$P_2 - S_2^{(m)}$

If we were to perform 96 different chi-square tests of this type, one for each mutation considered with triplet context, these tests would not be independent. A sufficiently large increase in the rate of one mutation type m_1 in population P_1 after divergence from P_2 could cause another mutation type m_2, whose rate has remained constant, to comprise significantly different fractions of the SNPs from P_1 and P_2. To minimize this effect, we use the following iterative procedure: first, compute a chi-square significance value $p_{\text{unordered}}(m)$ for each mutation type m using the 2-by-2 chi-square table above. We then use these values to order the SNPs from lowest p value to highest and compute a set of ordered p values $p_{\text{ordered}}(m)$. For the mutation type m_0 with the lowest unordered p value, $p_{\text{unordered}}(m_0) = p_{\text{ordered}}(m_0)$. For mutation type m_i, which has the ith lowest unordered p value and $i < 96$, $p_{\text{ordered}}(m_i)$ is computed from the following contingency table:

$S_1^{(m_i)}$	$\sum_{j=i+1}^{96} S_1^{(m_j)}$
$S_2^{(m_i)}$	$\sum_{j=i+1}^{96} S_2^{(m_j)}$

For mutation type m_{96}, which has the highest unordered p value, the ordered p value is computed from the contingency table

$S_1^{(m_{96})}$	$S_1^{(m_{95})}$
$S_2^{(m_{96})}$	$S_2^{(m_{95})}$

Principal component analysis (PCA). The python package matplotlib.mlab.PCA was used to perform PCA on the complete set of 1000 Genomes haplotypes, each haplotype h represented by a 96-element vector encoding the mutation frequencies $(f_h(m))_m$ of the non-singleton derived alleles present on that haplotype. In the same way, a separate PCA was performed on each of the 5 continental groups to reveal finescale components of mutation spectrum variation.

Dating of the TCC→T mutation pulse. We estimated the duration and intensity of TCC→T rate acceleration in Europe by fitting a simple piecewise-constant rate model to the UK10K frequency data. To specify the parameters of the model, we divide time into discrete log-spaced intervals bounded by time points $t_1, ..., t_d$, assigning each interval a TCC→T mutation rate $r_0, ...r_d$. In units of generations before the present, the time discretization points were chosen to be: 20, 40, 200, 400, 800, 1200, 1600, 2000, 2400, 2800, 3200, 3600, 4000, 8000, 12000, 16000, 20000, 24000, 28000, 32000, 36000, 40000. We assume that the total rate r of mutations other than TCC→T stays constant over time (a first-order approximation).

In terms of these rate variables, we can calculate the expected shape of the TCC→T pulse shown in Figure 2B of the main text. The shape of this curve depends on both the mutation rate parameters r_i and the demographic history of the European population, which determines the joint distribution of allele frequency and allele age. To account for the effects of demography, we use Hudson's ms program to simulate 10,000 random coalescent trees under a realistic European demographic history inferred from allele frequency data [35] and condition our inference upon this collection of trees as follows: Let $A(m,t)$ be the function for which $\int_{t_i}^{t_{i+1}} A(m,t)dt$ equals the coalescent tree branch length, averaged over the sample of simulated trees, that is ancestral to exactly m lineages and falls between time t_i and t_{i+1}. Given this function, which can be empirically estimated from a sample of simulated trees, the expected frequency

spectrum entry k/n is

$$E(k/n) = \frac{\sum_{i=1}^{d} \int_{t_{i-1}}^{t_i} A(k,t)dt}{\sum_{j=1}^{n} \sum_{i=1}^{d} \int_{t_{i-1}}^{t_i} A(j,t)dt}$$

and the expected fraction of TCC→T mutations in allele frequency bin k/n is

$$E(f_{\mathrm{TCC}\to\mathrm{T}}(k/n)) = \frac{\sum_{i=1}^{d} r_i \int_{t_{i-1}}^{t_i} A(k,t)dt}{r \sum_{i=1}^{d} \int_{t_{i-1}}^{t_i} A(k,t)dt}.$$

The expected value of the TCC→T enrichment ratio being plotted in Figure 2B is

$$E(r_{\mathrm{TCC}\to\mathrm{T}}(k/n)) = \frac{\sum_{i=1}^{d} r_i \int_{t_{i-1}}^{t_i} A(k,t)dt \cdot \sum_{j=1}^{n} \sum_{i=1}^{d} \int_{t_{i-1}}^{t_i} A(j,t)dt}{\sum_{i=1}^{d} \int_{t_{i-1}}^{t_i} A(k,t)dt \cdot \sum_{j=1}^{n} \sum_{i=1}^{d} r_i \int_{t_{i-1}}^{t_i} A(j,t)dt}$$

In Figure 2B, enrichment ratios are not computed for every allele frequency in isolation, but for allele frequency bins that each contain similar numbers of SNPs. Given integers $1 \leq k_m < k_{m+1} \leq n$, the expected TCC→T enrichment ratio averaged over all SNPs with allele frequency between k_m/n and k_{m+1}/n is:

$$E(r_{\mathrm{TCC}\to\mathrm{T}}(k_m/n)) = \frac{\sum_{i=1}^{d} r_i \int_{t_{i-1}}^{t_i} \sum_{k=k_m}^{k_{m+1}} A(k,t)dt \cdot \sum_{j=1}^{n} \sum_{i=1}^{d} \int_{t_{i-1}}^{t_i} A(j,t)dt}{\sum_{i=1}^{d} \int_{t_{i-1}}^{t_i} \sum_{k=k_m}^{k_{m+1}} A(k,t)dt \cdot \sum_{j=1}^{n} \sum_{i=1}^{d} r_i \int_{t_{i-1}}^{t_i} A(j,t)dt}$$

We optimize the mutation rates $r_1, \ldots, r_d$ using a log-spaced quantization of allele frequencies $k_1/n, \ldots, k_m/n$ defined such that all bins contain similar numbers of SNPs. The chosen allele count endpoints $k_1, \ldots, k_m$ are: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000. Given this quantization of allele frequencies, we optimize $r_1, \ldots, r_d$ by using the BFGS algorithm to minimize the least squares distance $D(r_0, \ldots, r_d)$ between $E(r_{\mathrm{TCC}\to\mathrm{T}}(k_m/n))$ and the empirical ratio $r_{\mathrm{TCC}\to\mathrm{T}}(k_m/n)$ computed from the UK10K data. This optimization is subject to a regularization penalty that minimizes the jumps between adjacent mutation rates r_i and r_{i+1}:

$$D(r_0, \ldots, r_d) = \sum_{m=1}^{d} \left(E(r_{\mathrm{TCC}\to\mathrm{T}}(k_m/n)) - r_{\mathrm{TCC}\to\mathrm{T}}(k_m/n)\right)^2 + 0.25\sqrt{\sum_{i=1}^{d}(r_{i-1} - r_i)^2}$$

Testing candidate loci for APOBEC-associated mutagenicity. A recent study found that two common germline variants affect both cancer risk and the somatic mutational spectrum by altering the regulation of APOBEC DNA-editing enzymes [24]. We tested each variant for germline mutator activity in the ALSPAC cohort of the UK10K data by looking at mutations that fall within 20 kB of the candidate mutator and have likely not had time to recombine away. We posited that a variant arose on the mutator background if

at least 80% of its derived alleles occur on a haplotype containing the candidate mutator; conversely, we posited that it arose on the ancestral background if less than 20% of its derived alleles are currently linked to the candidate mutator. In genomes that are heterozygous for the putative mutator allele, singletons were assigned uniformly at random to the ancestral and derived allelic backgrounds. Based on previous descriptions of the APOBEC signature [3,26], we classified TC→TT and TC→TG mutations as potentially APOBEC-associated, all other mutations as non-APOBEC-associated, and tested for an enrichment of APOBEC-associated variants on the candidate mutator background.

The derived allele of SNP rs1014971, in the APOBEC3B regulatory region, is associated with an increased burden of somatic APOBEC mutations as well as increased risk of bladder cancer. We counted 34 APOBEC-type mutations linked to the derived allele in UK10K, as well as 206 non-APOBEC-type mutations. On the ancestral allele background, we counted 15 APOBEC-type mutations and 189 non-APOBEC-type mutations. By a χ^2 test, the derived allele is significantly associated with an increased proportion of APOBEC-type mutations ($p = 0.03$).

A second APOBEC variant is associated with an increased risk of breast cancer: a 30 kB deletion that elides APOBEC3A and APOBEC3B into a single upregulated chimera. We attempted to test this deletion for germline mutagenicity by applying our testing procedure to the tag SNP rs12628403. One complicating factor is that the APOBEC deletion is rare in Europeans compared to East Asians, Amerindians, and Oceanians [36]. In addition, low coverage sequencing datasets like the UK10K are vulnerable to assembly difficulties and a dearth of SNP calls near sites of structural variation. As a result, only 19 SNPs, 1 APOBEC-type and 18 non-APOBEC-type, were found linked to the deletion within 20 kB, too few to reliably test for germline mutator activity.

Code availability. Contact K.H. (kelleyh@stanford.edu) to request copies of the python scripts that were used to perform the analyses in this paper.

References

1. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).

2. Ségurel, L., Wyman, M. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15**, 19.1–19.24 (2014).

3. Alexandrov, L. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

4. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics* **15**, 585–598 (2014).

5. Shinbrot, E. *et al.* Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome research* **24**, 1740–1750 (2014).

6. Lynch, M. The lower bound to the evolution of mutation rates. *Genome biology and evolution* **3**, 1107–1118 (2011).

7. Sung, W., Ackerman, M., Miller, S., Doak, T. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci USA* **109**, 18488–18492 (2012).

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

9. Mathieson, I. & Reich, D. E. Variation in mutation rates among human populations. *bioRxiv* 063578 (2016).

10. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

11. Pollard, K., Hubisz, M., Rosenbloom, K. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20**, 110–121 (2010).

12. November, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

13. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

14. Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**, 907–911 (2001).

15. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genetics* **47**, 126–131 (2015).

16. Moorjani, P., Amorim, C., Arndt, P. & Przeworski, M. Variation in the molecular clock of primates. *Proc Natl Acad Sci USA* **113**, 10607–10612 (2016).

17. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences* **90**, 4087–4091 (1993).

18. Amster, G. & Sella, G. Life history effects on the molecular clock of autosomes and sex chromosomes. *Proceedings of the National Academy of Sciences* **113**, 1588–1593 (2016).

19. Hwang, D. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA* **101**, 13994–14001 (2004).

20. Moorjani, P., Gao, Z. & Przeworski, M. Human germline mutation and the erratic molecular clock. *bioRxiv* 058024 (2016).

21. Gao, Z., Wyman, M., Sella, G. & Przeworski, M. Interpreting the dependence of mutation rates on age and time. *PLoS Biology* **14**, e1002355 (2016).

22. Narasimhan, V. *et al.* A direct multi-generational estate of the human mutation rate from autozygous segments seen in thousands of parentally related individuals. *bioRxiv preprint* http://dx.doi.org/10.1101/059436 (206).

23. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

24. Middlebrooks, C. *et al.* Association of germline variants in APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nature Genetics* **early edition**, 1–9 (2016).

25. Pinto, Y. *et al.* Clustered mutations in hominid genome evolution are consistent with APOBEC3G enzymatic activity. *Genome Res* **26**, 579–587 (2016).

26. Nik-Zainal, S. *et al.* Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat Genetics* **46**, 487–491 (2014).

27. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013).

28. Goodman, M. The role of immunochemical differences in the phyletic development of human behavior. *Human Biol* **33**, 131–162 (1961).

29. Li, W. & Tanimura, M. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326**, 93–96 (1987).

30. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature Rev Genetics* **13**, 745–753 (2012).

31. Coop, G., Wen, X., Ober, C., Pritchard, J. K. & Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319**, 1395–1398 (2008).

32. Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**, 836–840 (2010).

33. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013-15).

34. Field, Y. *et al.* Detection of human adaptation during the past 2,000 years. *Science* **13**, 10.1126/science.aag0776 (2016).

35. Tennessen, J. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **6**, 64–69 (2012).

36. Kidd, J., Newman, T., Tuzun, E., Kaul, R. & Eichler, E. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genetics* **3**, e63 (2007).

Acknowledgements This work was funded by NIH grants GM116381 and HG008140, and by the Howard Hughes Medical Institute. We thank Jeffrey Spence and Yun S. Song for technical assistance and Ziyue Gao, Arbel Harpak, Molly Przeworski and Joshua Schraiber for comments and discussion.

Competing Interests The authors declare that they have no competing financial interests.