Recent evolution of the mutation rate and spectrum in Europeans

Kelley Harris[1]*

[1]Department of Mathematics; University of California, Berkeley; Berkeley, CA 94720 USA

*To whom correspondence should be addressed; E-mail: kharris@math.berkeley.edu.

Abstract

As humans dispersed out of Africa, they adapted to new environmental challenges including changes in exposure to mutagenic solar radiation. This raises the possibility that different populations experienced different selective pressures affecting genome integrity. Prior work has uncovered some evidence for local adaption of eQTLs that regulate the DNA damage response[1], as well as indications that the human mutation rate per year may have changed at least 2-fold since we shared a common ancestor with chimpanzees [2, 3]. Here, I present evidence that the rate of a particular mutation type has recently increased in the European lineage, rising in frequency by 50% during the 40,000–80,000 years since Europeans diverged from Asians. A comparison of single nucleotide polymorphisms (SNPs) private to Africa, Asia, and Europe in the 1000 Genomes data reveals that private European variation is enriched for the transition 5'-TCC-3'→5'-TTC-3'. Although it is not clear whether UV played a causal role in the changing the European mutational spectrum, 5'-TCC-3'→5'-TTC-3' is known to be the most common somatic mutation present in melanoma skin cancers [4], as well as the mutation most frequently induced *in vitro* by UV [5]. Regardless of its causality, this change indicates that DNA replication fidelity has not remained stable even since the origin of modern humans and might have changed numerous times during our recent evolutionary history.

Introduction

Anatomically moderns humans left Africa less than 200,000 years ago and have since dispersed into every habitable environment [6]. Since different habitats have presented humans with diverse environmental challenges, many local adaptations have caused human populations to diverge phenotypically. Some adaptations like light and dark skin pigmentation have been studied since the early days of evolutionary theory [7, 8, 9], but other adaptations have only been discovered within the last few years as a result of innovative phenotypic measurement strategies combined with genome-wide scans for natural selection [10, 11, 12, 13].

One phenotype that is notoriously hard to measure is the human germline mutation rate. It recently became possible to estimate this rate by sequencing parent-offspring trios and counting new mutations directly, but the resulting estimates are complicated by sequencing error and differ significantly from earlier indirect estimates based on the divergence between humans and chimpanzees [14, 15, 3, 2, 16]. It is an even harder problem to measure whether mutation rates differ between populations; however, it is becoming more straightforward and also more potentially illuminating to compare the relative frequencies of specific mutation types, e.g. transitions and transversions, across populations. Genetic variants that perturb the mutation rate must somehow change the fidelity of DNA replication or repair, or else alter the propensity of mutagens come into contact with the DNA and cause damage. Such perturbations tend to have unequal effects on various DNA motifs and mutation types. Different cancers, for example, exhibit different somatic mutation signatures that can be diagnostic of underlying carcinogen exposures and somatic driver mutations [4, 17].

Among eQTL SNPs that are variable in the human germline and affect the expression of genes from various functional categories, SNPs that affect regulation of the DNA damage response show some of the strongest evidence of local adaptation and recent positive selection, with frequency differentiation between populations that appears to be correlated with environmental UV exposure [1]. Such expression changes have the potential to alter the mutation spectrum in a more subtle, heritable way than occurs in cancer cells.

Results

To test for differences in the spectrum of mutagenesis between populations, I compiled sets of population-private variants from the 1000 Genomes Phase I panel of 1,092 human genome sequences [18]. Excluding singletons and SNPs with imputation quality lower than RSQ = 0.95, which might be misleadingly classified as population-private due to imputation error, there remain 462,876 private European SNPs (PE) that are variable in Europe but fixed ancestral in all non-admixed Asian and African populations, as well as 265,988 private Asian SNPs (PAs) that are variable in Asia but fixed ancestral in Africa and Europe. These SNPs should be enriched for young mutations that arose after humans had already left Africa and begun adapting to temperate latitudes. I compared PE and PAs to the set of 3,357,498 private African SNPs (PAf) that are variable in the Yorubans (YRI) and/or Luhya (LWK) but fixed ancestral in Europe and Asia. One notable feature of PE is the percentage of SNPs that are C→T transitions, which is high (41.01%) compared to the same percentage in PAs (38.99%) and PAf (38.29%).

Excess C→T transitions are characteristic of several different mutagenic processes including UV damage and cytosine deamination [4]. To some extent, these processes can be distinguished by partitioning SNPs into 192 different context-dependent classes, looking at the reference base pairs immediately upstream and downstream of the variable site [19]. For each mutation type $m = B_{5'}B_AB_{3'} \to B_{5'}B_DB_{3'}$ and each private SNP set P, I obtained the count $C_P(m)$ of type-m mutations in set P, the aim being to test whether type m-mutations have a significantly different frequency in PE or PAs compared to PAf. One complication is that mutation frequencies sum to 1 and are thus not independent of each other. For example, if mutation m occurs at a higher rate in Europe than Africa and thus has a higher frequency in PEu than PAf, one or more other mutations must have lower frequencies in PEu than PAf even if their rates do not differ between the two populations. To minimize this confounding effect, I tested for differences between populations of the ratio $r_P(m) = C_P(m)/C_P$(5'-CCG-3' $\to$ 5'-CTG-3'), the relative abundance of mutation m to the fixed type m_0 =5'-CCG-3'→ 5'-CTG-3'. The mutation m_0 was chosen as the reference type because its frequency is similar across PE, PAs, and PAf.

Supplemental Tables S**??**–S**??** list the p-values and $r_m(P)$ values obtained for all mutation types and population comparisons using a χ^2 test on a 2×2 contingency table (Figure 1B). As shown in Figure 1A, no CpG-related transitions appear to have elevated rates in Europe compared to Africa. Instead, the strongest candidate for rate change is the mutation type 5'-TCC-3'→5'-TTC-3' (hereafter abbreviated as TCC→T). Several other C→T transitions are also moderately more abundant in PE than PAf, in most cases flanked by either a 5' T or a 3' C. Combined with its reverse strand complement 5'-GGA-3'→5'-GAA-3', TCC→T has frequency 3.32% in PE compared with 1.98% in PAf and 2.04% in PAs.

As shown in Figure 1C, the TCC→T frequency difference holds genome-wide, evident on every chromosome except for chromosome Y, which has too little population-private variation to yield accurate measurements of context-dependent SNP frequencies. The most parsimonious explanation is that Europeans experienced a genetic change increasing the rate of TCC→T mutations. This claim is supported by the results of a branch length ratio test (Supporting Information Section S1), which finds that a single European haplotype contains more derived variants, particularly derived C→T variants, than a single Asian haplotype. This asymmetry cannot be explained by a demographic event such as a population bottleneck, and is not consistent with the assumption of equal mutation rates between populations.

C→T transitions may not be the only mutations that have experienced recent rate change. For example, TTA→TAA mutations appear to be less abundant in Europe than Africa, and several T→C transitions appear to have higher rates in Africa than in either Europe or Asia (Figure 1A,B). However, these subtle asymmetries do not consistently hold up in subsets of the data that have varied bioinformatic characteristics such as GC content, imputation quality, and sequencing depth. As described in the following section, C→T transitions, particularly TCC→T, show more robust indications of mutation rate change.

Robustness to sources of bioinformatic error

To rule out the possibility that the TCC→T excess is a bioinformatic artifact specific to the 1000 Genomes data, I reproduced Figure 1A,B in a set of human genomes sequenced at high coverage using Complete Genomics technology (Supporting Information Section S3). I also folded the context-dependent mutation frequency spectrum to check for effects of ancestral misidentification (Supporting Information Section S5). Finally, I partitioned the 1000 Genomes data into bins based on GC content and sequencing depth and found that the TCC→T excess in Europe was easily discernible within each bin (Supporting Information Sections S6 and S7). Three other C→T transitions (TCT→TTT, ACC→ATC, and CCC→CTC) are also more abundant in Europe than Africa across a broad range of GC contents and sequencing depths. In contrast, genomic regions that differ in GC content and/or sequencing depth show little consistency as to which mutation types have the greatest frequency differences between Africa and Asia.

As mentioned previously, singleton variants (minor allele count = 1) were excluded from all analyses. When singletons are included, they create spurious between-population differences that are not reproducible with non-singleton SNPs (Supporting Information Section S8). This is true of both the low coverage 1000 Genomes dataset and the smaller, higher coverage Complete Genomics dataset, suggesting that singletons are very error-prone even in high coverage data.

A particularly interesting class of singletons are the mutations appear to have arisen *de novo* in the offspring of parent-child trios. Barring bioinformatic problems, counting these mutations should yield an accurate estimate of the current human mutation rate and spectrum. When the mutation spectra of PE, PAf and PAs were compared to the spectrum of putative *de* *novo* mutations in 82 Icelandic trios [3], the Icelandic *de novo* TCC→T mutation rate appeared most similar to the European TCC→T rate based on relative abundances of all mutation types (Supplementary Information Section S9). To my knowledge, no large dataset of non-European trios is currently available for the purpose of a similar analysis.

Antiquity of the European mutation rate change

The 1000 Genomes Phase I dataset contains samples from five European sub-populations: Italians (TSI), Spanish (IBS), Utah residents of European descent (CEU), British (GBR), and Finnish (FIN). All of these populations have elevated TCC→T frequencies, suggesting that the European mutation rate changed before subpopulations diversified across the continent. To assess this, I let $\mathrm{P}_{\text{total}}$ denote the combined set of private variants from PE, PAs, and PAf, and for each haplotype h let $\mathrm{P}_{\text{total}}(h)$ denote the subset of $\mathrm{P}_{\text{total}}$ whose derived alleles are found on haplotype h. $f_h(\text{TCC})$ then denotes the frequency of TCC→T within $\mathrm{P}_{\text{total}}(h)$. For each 1000 Genomes population P, Figure 2 shows the distribution of $f_h(\text{TCC})$ across all haplotypes h sampled from P, and it can be seen that the distribution of $f(\text{TCC})$ values found in Europe does not overlap with the distributions from Asia and Africa. In contrast, the four admixed populations ASW (African Americans), MXL (Mexicans), PUR (Puerto Ricans), and CLM (Colombians) display broader ranges of $f(\text{TCC})$ with extremes overlapping both the European and non-European distributions. The African American $f(\text{TCC})$ values are only slightly higher on average than the non-admixed African values, but a few African American individuals have much higher $f(\text{TCC})$ values in the middle of the admixed American range, presumably because they have more European ancestry than the other African Americans who were sampled.

Within Europe, Figure 2 shows a slight $f(\text{TCC})$ gradient running from North to South; the median $f(\text{TCC})$ is lowest in the Finns and highest in the Spanish and Italians. In this way, TCC→TTC frequency appears to correlate negatively with recent Asian co-ancestry (Supporting Information S2).

To roughly estimate the time when the TCC→T rate increased, I downloaded allele age estimates that were generated from the Complete Genomics data using the program ARGweaver [20]. Based on these estimates, the TCC→T rate acceleration appears to have occurred between 25,000 and 60,000 years ago, not long after Europeans diverged from Asians (Supporting Information Section S4). In the 1000 Genomes, data, TCC→T frequency differentiation is greatest for private alleles of frequency less than 0.02 (Supplementary Figure S6B).

Reversal of TCC→T transcription strand bias in Europe

In transcribed genomic regions, the mutation spectrum is shaped by forces of selection, repair, and mutagenesis that differ from the forces affecting non-transcribed regions. One such force is transcription-coupled repair (TCR) of damage that affects the template DNA strand. In mammalian genes, it has been observed that TCR repairs G/T→A/C mutations more efficiently than A/C→G/T mutations, producing a G+T excess on the transcribed strand relative to the non-transcribed strand [21].

TCC→T SNPs that are private to Asia or Africa exhibit the strand bias that is typical of G/T→A/C mutations, with TCC→T occurring less often on the transcribed strand than on the non-transcribed strand. In contrast, private European TCC→T SNPs exhibit no discernible strand bias, occurring equally often on both strands in transcribed regions (Figure 3A,B,C). This strand bias difference is significant according to a χ^2 test ($p < 7.87 \times 10^{-4}$). At the significance level $p < 0.01$, no other mutation type shows a difference in strand bias between Europe and Africa or Asia and Africa (Figure 3B).

This strand bias reversal suggests that decreased TCR efficiency might contribute to the high European TCC→T rate. In support of this, when PE is partitioned into genic SNPs and intergenic SNPs, the frequency of TCC→T is higher in the genic SNP set (Figure 3D), suggesting that TCR of TCC→T mutations is relatively inefficient. In contrast, when PAs and PAf are partitioned into genic and intergenic SNPs, the genic SNP sets have lower TCC→T frequencies, suggesting that TCR of this mutation type is relatively efficient in non-Europeans. However, TCR efficiency is not likely to be sole cause of the TCC→T rate differential, since this rate differential affects both genic and intergenic regions.

Discussion

It is beyond the scope of this article to pinpoint why the rate of TCC→T increased in Europe. However, some promising clues can be found in the literature on ultraviolet-induced mutagenesis. In the mid-1990s, Drobetsky, et al. and Marionnet, et al. each observed that TCC→T dominated the mutational spectra of single genes isolated from UV-irradiated cell cultures [5, 22]. Much more recently, Alexandrov, *et al.* systematically inferred “mutational signatures" from 7,042 different cancers and found that melanoma has a unique mutational signature not present in any other cancer type they studied [4]. Melanoma somatic mutations consist almost entirely of C→T transitions, 28% of which are TCC→T mutations [4, 23]. The mutation types CCC→CTC and TCT→TTT, two other candidates for rate acceleration in Europe, are also prominent in the spectrum of melanoma (Supporting Information Section S11). Incidentally, melanoma is not only associated with UV light exposure, but also with European ancestry, occurring at much lower rates in Africans, African Americans, and also lighter-skinned Asians [24, 25, 26]. A study of the California Cancer registry found that the annual age-adjusted incidence of melanoma cases per 100,000 people was 0.8-0.9 for Asians, 0.7-1.0 for African Americans, and 11.3–17.2 for Caucasians [27]. Melanoma incidence in admixed Hispanics is strongly correlated with European ancestry [27, 25, 26].

The association of TCC→T mutations with UV exposure is not well understood, but two factors appear to be in play: 1) the propensity of UV to cross-link the TC into a base dimer lesion and 2) poorer repair efficacy at TCC than at other motifs where UV lesions can form [28, 29]. Drobetsky, et al. compared the incidence of UV lesions to the incidence of mutations in irradiated cells and found that TCC motifs were not hotspots for lesion formation, but instead were disproportionately likely to have lesions develop into mutations rather than undergoing error-free repair [5].

Despite the strong evidence that UV causes TCC→T mutations, the question remains how UV could affect germline cells that are generally shielded from solar radiation. Although the testes contain germline tissue that lies close to the skin with minimal shielding, to my knowledge it has not been tested whether UV penetrates this tissue effectively enough to induce spermatic mutations. Another possibility is that UV can indirectly cause germline mutations by degrading folate, a DNA synthesis cofactor that is transmitted through the bloodstream and required during cell division [30, 8, 9, 31]. Folate deficiency is known to cause DNA damage including uracil misincorporation and double-strand breaks, leading in some cases to birth defects and reduced male fertility [32, 33, 34]. It is therefore possible that folate depletion could cause some of the mutations observed in UV-irradiated cells, and that these same mutations might appear in the germline of a light-skinned individual rendered folate-deficient by sun exposure. It has also been hypothesized that, in a variety of species, differences in metabolic rate can drive latitudinal gradients in the rate of molecular evolution [**?**, **?**, **?**].

Although the data presented here do not reveal a clear underlying mechanism, they leave little doubt that the European population experienced a recent mutation rate increase. Even if the overall rate increase was small, it adds to a growing body of evidence that molecular clock assumptions break down on a faster timescale than generally assumed during population genetic analysis. It was once assumed that the human lineage's mutation rate had changed little since we shared a common ancestor with chimpanzees, but this assumption is losing credibility due to the conflict between direct mutation rate estimates and with molecular-clock-based estimates [14, 15]. One proposed reconciliation of this conflict is a “hominoid slowdown," a gradual decrease in the rate of germline mitoses per year as our ancestors evolved longer generation times [35, 36]. The results of this paper indicate that another force may have come into play: change in the mutation rate per mitosis. If the mutagenic spectrum was able to change during the last 100,000 years of human history, it might have changed numerous times during great ape evolution and beforehand. Given such a general challenge to the molecular clock assumption, it may be wise to infer demographic history from mutations such as CpG transitions that accumulate in a more clocklike way than other mutations [19, 14]. At the same time, less clocklike mutations may provide valuable insights into the changing biology of genome integrity.

Methods

1000 Genomes data accession

Publicly available VCF files containing the 1000 Genomes Phase I were downloaded from www.1000genomes.org/data. Ancestral states were inferred using the UCSC alignment of the chimp PanTro4 to the human reference genome hg19. These data were then subsampled to obtain four sets of SNPs: PE (private to Europe), PAs (private to Asia), PAf (private to Africa), and PAsE (fixed in Africa but variable in both Asia and Europe).

Construction of private SNP sets PE, PAs, PAf, and PAsE

The definitions of PE, PAs, and PAf differ slightly from the definitions of continent-private SNPs from the manuscript announcing the release of the 1000 Genomes Phase I data [18]. In that paper, a SNP is considered private to Africa if it is variable in at least one of the populations LWK (Luhya from Kenya), YRI (Yoruba from Nigeria), and ASW (African Americans from the Southwestern USA). In contrast, I consider a SNP to be private to Africa if it is variable in either LWK or YRI and is not variable in any of the following samples: CHB (Chinese from Beijing), CHS (Chinese from Shanghai), JPT (Japanese from Tokyo), CEU (Individuals of Central European descent from Utah), GBR (Great Britain), IBS (Spanish from the Iberian Peninsula), TSI (Italians from Tuscany), and FIN (Finnish). A private African SNP might or might not be variable in any of the admixed samples ASW, MXL (Mexicans from Los Angeles), CLM (Colombians from Medellin), and PUR (Puerto Ricans). Similarly, a private European SNP in PE is variable in one or more of the CEU, GBR, IBS, TSI, and FIN, is not variable in any of YRI, LWK, CHB, CHS, or JPT, and might or might not be variable in ASW, MXL, CLM, and PUR. The private Asian SNPs in PAs are variable in one or more of CHB, CHS, or JPT, are not variable in any of YRI, LWK, CEU, GBR, IBS, TSI, and FIN, and might or might not be variable in ASW, MXL, CLM, and PUR. These definitions are meant to select for mutations that have been confined to a single continent for most of their history except for possible recent transmission to the Americas. The shared European-Asian SNPs in PAsE are variable in one or more of CHB, CHS, or JPT plus one or more of CEU, GBR, IBS, TSI, and FIN and are not variable in YRI or LWK. Singletons are excluded to minimize the impact of possible sequencing error, and variants with imputation quality lower than RSQ $= 0.95$ are excluded to minimize erroneous designation of shared SNPs as population-private.

Statistical analysis of frequency differences

Given two populations P_1 and P_2 and one SNP type m, a Pearson's χ^2 value was used to assign a P-value to the significance of the frequency difference of m between P_1 and P_2. Using $C_{P_i}(m)$ to denote the number of type-m SNPs in population P_i and M to denote the set of all SNP types, the frequency of $f_i(m)$ is defined to be

$$f_i(m) = \frac{C_{P_i}(m)}{\sum_{m' \in M} C_{P_i}(m')}.$$

Since $f(m)$ depends on the abundances of every other SNP type m', I used a related measure $r(m)$ that normalizes the abundance of m by the number of SNPs sampled from P_1 but is less influenced by the relative abundances of other SNP types. I picked a single focal SNP type m_0 =5'-CCG-3'$\rightarrow$ 5'-CTG-3' whose frequency was similar across datasets and calculated the relative abundance of each other type to m_0:

$$r_i(m) = \frac{C_{P_i}(m)}{C_{P_i}(m_0)}$$

The expected values of $C_{P_1}(m)$, $C_{P_2}(m)$, $C_{P_1}(m_0)$, and $C_{P_2}(m_0)$ under the expectation of no frequency difference were calculated as follows based on a 4×4 contingency table:

$$\begin{aligned}\mathbb{E}(C_{P_i}(m)) &= \frac{(C_{P_i}(m) + C_{P_i}(m_0))(C_{P_i}(m) + C_{P_{3-i}}(m))}{C_{P_i}(m) + C_{P_{3-i}}(m) + C_{P_i}(m_0) + C_{P_{3-i}}(m_0)} \\ \mathbb{E}(C_{P_i}(m_0)) &= \frac{(C_{P_i}(m) + C_{P_i}(m_0))(C_{P_i}(m_0) + C_{P_{3-i}}(m_0))}{C_{P_i}(m) + C_{P_{3-i}}(m) + C_{P_i}(m_0) + C_{P_{3-i}}(m_0)}\end{aligned}$$

The following χ^2 value measures the significance of the difference $r_1(m) - r_2(m)$:

$$\chi^2 = \sum_{i=1}^{2} \frac{(C_{P_i}(m) - \mathbb{E}(C_{P_i}(m)))^2}{\mathbb{E}(C_{P_i}(m))} + \frac{(C_{P_i}(m_0) - \mathbb{E}(C_{P_i}(m_0)))^2}{\mathbb{E}(C_{P_i}(m_0))}$$

A p value is then obtained using the χ^2 distribution with 1 degree of freedom. The normality assumption of the χ^2 test is justified because differences are expected to be close to 0 and are bounded between 0 and 1.

Nonparametric bootstrapping within chromosomes

To assess the variance of $f(\text{TCC} \rightarrow \text{T})$ within each of the autosomes and the X chromosome, each private SNP To assess the variance of $f(\text{TCC})$ within each of the autosomes and the X chromosome, each private SNP set PE, PAs, and PAf was partitioned into non-overlapping bins of 1,000 consecutive SNPs. The frequency $f(\text{TCC})$ of the mutation TCC$\rightarrow$T was computed for each bin and used to generated the box plot in Figure 1C. No partitioning into separate bins was performed for chromosome Y because the entire chromosome has only 1,130 private European SNPs, 1,857 private Asian SNPs and 3,852 private African SNPs. Instead the global frequency of TCC$\rightarrow$T was computed for each SNP set restricted to the Y chromosome.

Quantifying strand bias

Gene locations and transcription directions for hg19 were downloaded from the UCSC Genome browser. Within each private SNP set P, each mutation type m was counted separately on transcribed and non-transcribed gene strands to obtain counts $\mathbf{T}(\text{P}, m)$ and $\mathbf{N}(\text{P}, m)$. A mutation with an A/C ancestral allele on the transcribed strand is equivalent to occurrence of the complementary G/T ancestral mutation on the non-transcribed strand. Here, the strand bias of each A/C ancestral mutation m is defined to be $\mathbf{S}(\text{P}, m) = \mathbf{N}(\text{P}, m)/\mathbf{T}(\text{P}, m)$. The significance of the strand bias differences $\mathbf{S}(\text{PAf}, m) - \mathbf{S}(\text{PE}, m)$ and $\mathbf{S}(\text{PAf}, m) - \mathbf{S}(\text{PAs}, m)$ were measured using a χ^2 test. For a general comparison between populations P_1 and P_2, the χ^2 test expected values are the following:

$$\mathbb{E}(\mathbf{T}(P_i, m)) = \frac{(\mathbf{T}(P_i, m) + \mathbf{T}(P_{3-i}, m))(\mathbf{T}(P_i, m) + \mathbf{N}(P_i, m))}{\mathbf{T}(P_i, m) + \mathbf{T}(P_{3-i}, m) + \mathbf{N}(P_i, m) + \mathbf{N}(P_{3-i}, m)}$$
$$\mathbb{E}(\mathbf{N}(P_i, m)) = \frac{(\mathbf{N}(P_i, m) + \mathbf{N}(P_{3-i}, m))(\mathbf{T}(P_i, m) + \mathbf{N}(P_i, m))}{\mathbf{T}(P_i, m) + \mathbf{T}(P_{3-i}, m) + \mathbf{N}(P_i, m) + \mathbf{N}(P_{3-i}, m)}$$

The χ^2 value measuring the significance between $\mathbf{N}(P_1, m)/\mathbf{T}(P_1, m)$ and $\mathbf{N}(P_2, m)/\mathbf{T}(P_2, m)$ is then computed as follows:

$$\chi^2 = \sum_{i=1}^{2} \frac{(\mathbf{T}(P_i, m) - \mathbb{E}(\mathbf{T}(P_i, m)))^2}{\mathbb{E}(\mathbf{T}(P_i, m))} + \frac{(\mathbf{N}(P_i, m) - \mathbb{E}(\mathbf{N}(P_i, m)))^2}{\mathbb{E}(\mathbf{N}(P_i, m))}$$

As before, a p value is then obtained using the χ^2 distribution with 1 degree of freedom. The normality assumption of the χ^2 test is justified because differences are expected to be close to 0 and are bounded between 0 and 1.

Non-parametric bootstrapping was used to estimate strand bias variance within each population. The exome was partitioned into 100 bins containing approximately equal numbers of SNPs, and 100 exome replicates were generated each by sampling 100 bins with replacement. For each replicate, the frequency of TCC→T was calculated on the transcribed and non-transcribed strands. These two frequencies were added together to obtain the cumulative

TCC→T frequency within genic regions. The distribution of strand bias $\mathbf{S}(\mathrm{TCC} \to \mathrm{T})$ across replicates for each population is shown in Figure 3C.

Bootstrapping was similarly applied to intergenic SNPs by partitioning the non-genic portion of the genome into 100 bins with similar SNP counts. 100 bootstrap replicates were generated by sampling 100 bins with replacement, and the intergenic TCC→T frequency was computed for each replicate. The ratio of genic to intergenic $f(\mathrm{TCC})$ was calculated for each replicate (Figure 3D).

References

[1] Fraser, H. Gene expression drives local adaptation in humans. *Genome Res* **23**, 1089–1096 (2013).

[2] 1000 Genomes Project. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

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

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

[5] Drobetsky, E. & Sage, E. UV-induced G:C→A:T transitions at the APRT locus of Chinese hamster ovary cells cluster at frequently damaged 5'-TCC-3' sequences. *Mut Res* **289**, 131–138 (1993).

[6] Cann, R., Stoneking, M. & Wilson, A. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).

[7] Loomis, W. Skin-pigment regulation of vitamin-D biosynthesis in man. *Science* **157**, 501–506 (1967).

[8] Jablonski, N. & Chaplin, G. The evolution of human skin coloration. *J Hum Evol* **39**, 57–106 (2000).

[9] Jablonski, N. & Chaplin, G. Human skin pigmentation as an adaptation to UV radiation. *Proc Natl Acad Sci USA* **107**, 8962–8968 (2010).

[10] Akey, J. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res* **19**, 711–722 (2009).

[11] Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74**, 1111–1120 (2004).

[12] Sabeti, P. *et al.* Positive natural selection in the human lineage. *Science* **16**, 1614–1620 (2006).

[13] Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).

[14] 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).

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

[16] Nachman, M. & Crowell, S. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).

[17] Lawrence, M. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

[18] 1000 Genomes Project. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

[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] Rasmussen, M., Hubisz, M., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics* **10**, e1004342 (2014).

[21] Green, P. *et al.* Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genetics* 514–517 (2003).

[22] Marionnet, C., Benoit, A., Benhamou, S., Sarasin, A. & Stary, A. Characteristics of UV-induced mutation spectra in human XP-D/ERCC2 gene-mutated xeroderma pigmentosum and trichothiodystrophy cells. *J Mol Biol* **252**, 550–562 (1995).

[23] Pleasance, E. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).

[24] Crombie, I. Racial differences in melanoma incidence. *Br J Cancer* **40**, 185–193 (1979).

[25] Hu, D., Yu, G., McCormick, S., Schneider, S. & Finger, P. Population-based incidence of uveal melanoma in various races and ethnic groups. *Am J Ophthalmology* **140**, 612.e1–612.e6 (2005).

[26] Bakos, L. *et al.* European ancestry and cutaneous melanoma in southern Brazil. *JEADV* **23**, 304–307 (2009).

[27] Cress, R. & Holly, E. Incidence of cutaneous melanoma among non-Hispanic whites, Hispanics, Asians, and blacks: an analysis of California Cancer Registry data, 1988–93. *Cancer Causes and Control* **8**, 246–252 (1997).

[28] Brash, D., Seetharam, S., Kraemer, K., Seidman, M. & Bredberg, A. Photoproduct frequency is not the major determinant of UV base substitution hot spots or cold spots in human cells. *Proc Natl Acad Sci USA* **84**, 3782–3786 (1987).

[29] Drobetsky, E., Grosovsky, A. & Glickman, B. The specificity of UV-induced mutations at an endogenous locus in mammalian cells. *Proc Natl Acad Sci USA* **84**, 9103–9107 (1987).

[30] Branda, R. & Eaton, J. Skin color and nutrient photolysis: an evolutionary hypothesis. *Science* **201**, 625–626 (1978).

[31] Off, M. *et al.* Ultraviolet photo degradation of folic acid. *J Photochem Photobiol B* **82**, 47–55 (2005).

[32] Blount, B. *et al.* Folate deficiency causes uracil disincorporation into human DNA and chromosomal breakage: implications for cancer and neuronal damage. *Proc Natl Acad Sci USA* **94**, 3290–3295 (1997).

[33] Wallock, L. *et al.* Low seminal plasma folate concentrations are associated with low sperm density and count in male smokers and nonsmokers. *Fertility and Sterility* **75**, 252–259 (2001).

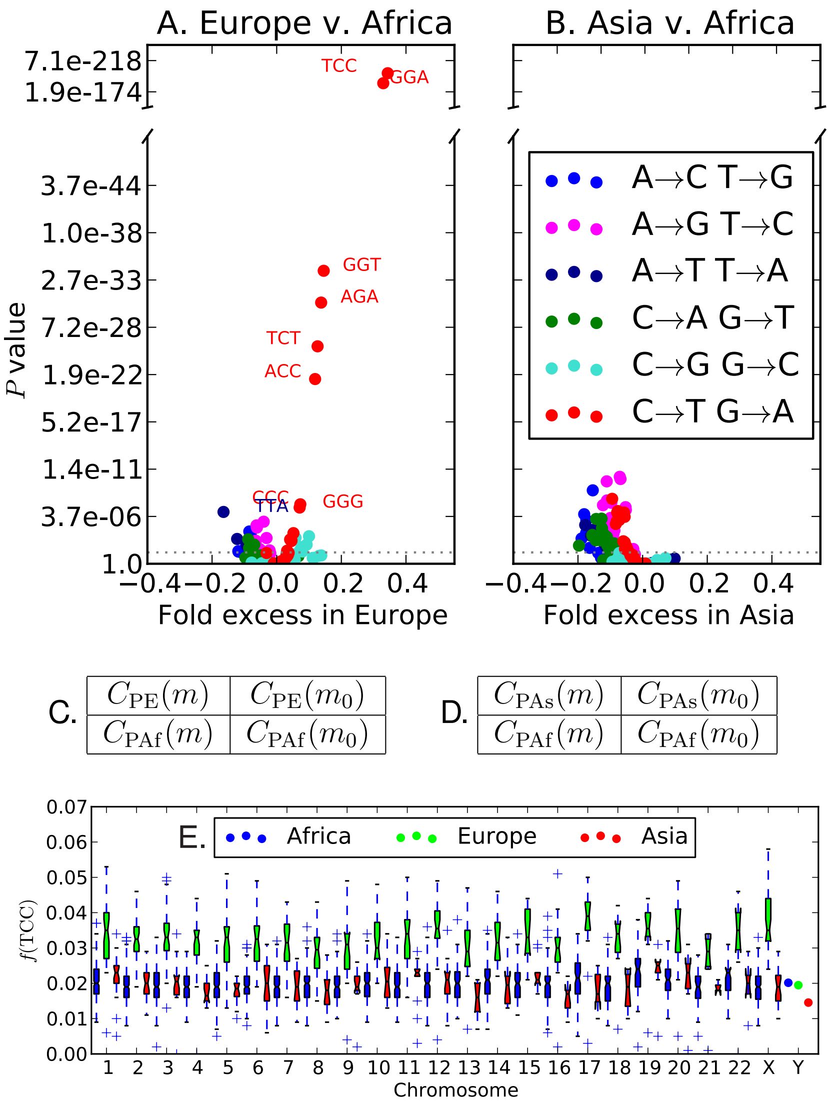
[34] Stover, P. One-carbon metabolism-genome interactions in folate-associated pathologies. *J Nutr* **139**, 2402–2405 (2009).

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

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

Acknowledgements

I am grateful to Rasmus Nielsen for advice and manuscript comments, and to two anonymous reviewers for providing feedback that improved upon an earlier draft. David Reich shared valuable insights into gene flow between early European farmers and hunter gatherers, and Richard Durbin, Stuart Linn, and Elad Ziv contributed additional helpful comments and suggestions. This work was supported by a National Science Foundation Graduate Research Fellowship (awarded to K.H.) and NIH grant IR01GM109454-01 (awarded to Rasmus Nielsen, Yun Song, and Steve Evans).



C.

$C_{\mathrm{PE}}(m)$	$C_{\mathrm{PE}}(m_0)$
$C_{\mathrm{PAf}}(m)$	$C_{\mathrm{PAf}}(m_0)$

D.

$C_{\mathrm{PAs}}(m)$	$C_{\mathrm{PAs}}(m_0)$
$C_{\mathrm{PAf}}(m)$	$C_{\mathrm{PAf}}(m_0)$

Fig. 1. Overrepresentation of 5'-TCC-3'→5'-TTC-3' within Europe. Panels A,B: The x coordinate of each point in gives the fold difference $r_m(\mathrm{PE}) - r_m(\mathrm{PAf})$ (resp. $r_m(\mathrm{PAs}) - r_m(\mathrm{PAf})$), while the y coordinate gives the Pearson's χ^2 p-value of its significance. Outlier points are labeled with the ancestral state of the mutant nucleotide flanked by two neighboring bases, and the color of the point specifies the ancestral and derived alleles of the mutant site. Panels C and D show the χ^2 contingency tables used to compute the p values plotted on the y axes of Panels A and B, respectively. Each value compares the abundance of a test mutation type m to the control mutation type $m_0 = \mathrm{CCG} \rightarrow \mathrm{CTG}$. Panel E shows the distribution of $f(\mathrm{TCC})$ across bins of 1000 consecutive population-private SNPs. Only chromosome-wide frequencies are shown for Chromosome Y because of its low SNP count.

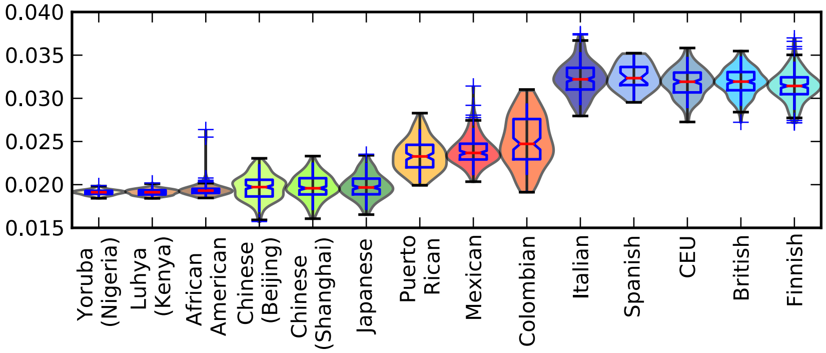


Fig. 2. Variation of $f(\text{TCC})$ within and between populations. This plot shows the distribution of $f(\text{TCC})$ within each 1000 Genomes population, i.e. the proportion of all derived variants from PA, PE, and PAf present in a particular genome that are m_{TCC} mutations. There is a clear division between the low $f(\text{TCC})$ values of African and Asian genomes and the high $f(\text{TCC})$ values of European genomes. The slightly admixed African Americans and more strongly admixed Latin American populations have intermediate $f(\text{TCC})$ values reflecting partial European ancestry.

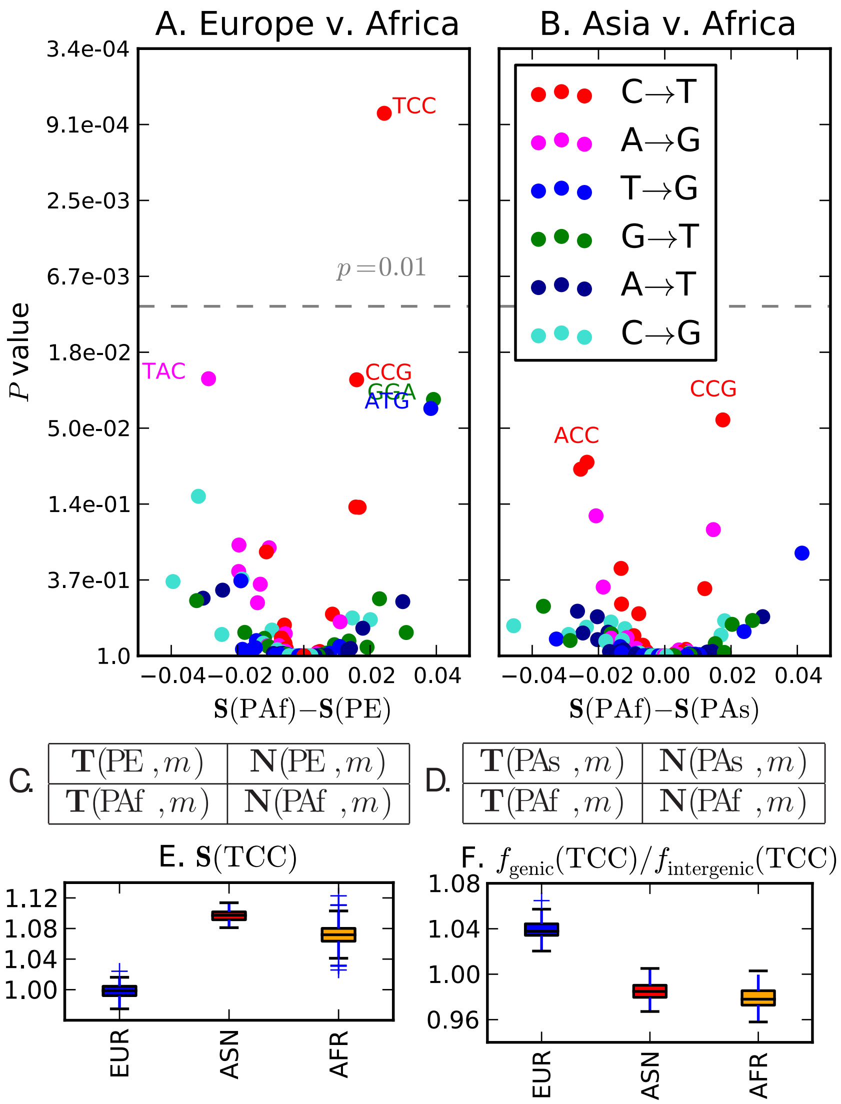


Fig 3. Differences in transcriptional strand bias. Each point in panels A and B represents a mutation type with an A or C ancestral allele. The strand bias for each mutation is defined to be the frequency on the nontranscribed strand divided by the frequency on the transcribed strand. The x coordinate of each point in panel A is the PAf strand bias minus the PE strand bias; similarly, the x coordinates in panel B describe the PAf strand bias minus the PAs strand bias. The y coordinate of each point is the χ^2 p value of the strand bias difference. At the $p = 0.01$ significance level (grey dashed line), only TCC→T has a significant strand bias difference between Europe and Africa, while no mutation type differs in strand bias between Asia and Africa. Panel C shows the variance of strand bias in each population across 100 bootstrap replicates. Similarly, Panel D shows the distribution across bootstrap replicates of the ratio between genic f(TCC) and intergenic f(TCC).