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 divergent selection in tropical versus temperate latitudes on eQTLs that regulate the DNA damage response[1], as well as evidence that the human mutation rate per year has 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 30,000–50,000 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 development of evolutionary theory [7, 8, 9], but other adaptations have only been discovered within the past 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 germline mutation rate must somehow change the fidelity of DNA replication or repair, and most known genetic variants that affect replication fidelity have readily identifiable mutational signatures. Cancer cells, for example, exhibit a variety of mutator phenotypes, and the relative frequencies of the resulting mutations can be diagnostic of underlying carcinogen exposures [4, 17]. There is evidence of widespread latitude-correlated selection on SNPs that regulate the expression of proteins controlling DNA repair and the UV damage response [1], and such expression changes have the potential to elevate or depress the rates of specific mutations.

Results

Mutation spectra of continent-private variation

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 doubletons to minimize the impact of sequencing

error, there remain 2,114,916 private European SNPs (PE) that are variable in Europe but fixed in all non-admixed Asian and African populations, as well as 1,560,755 private Asian SNPs (PAs) that are variable in Asia but fixed 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 6,945,574 private African SNPs (PAf) that are variable in the Yorubans (YRI) and/or Luhya (LWK) but fixed in Europe and Asia. One notable feature of PE is the percentage of SNPs that are C→T transitions, which is high (41.70%) compared to the same percentage in PAs (38.74%) and PAf (39.24%).

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. For each mutation type $m = B_{5'}B_AB_{3'} \rightarrow B_{5'}B_DB_{3'}$ and each private SNP set P, I obtained the count $C_P(m)$ of type-m mutations in set P and computed $r_P(m) = C_P(m)/C_P(\text{5'-CCG-3'} \rightarrow \text{5'-CTG-3'})$, the relative abundance of mutation m to the fixed type 5'-CCG-3'→ 5'-CTG-3', which was chosen as the reference type because its frequency is similar across PE, PAs, and PAf. Supplemental Tables S1–S5 list the p-values and $r_m(P)$ values obtained for all mutation types and population comparisons. As shown in Figure 1A,B, no CpG-related transitions are overrepresented in PE relatively to PAf; instead, the single extreme outlier is the mutation type 5'-TCC-3'→5'-TTC-3' (hereafter called m_{TCC}). Several other C→T transitions are moderately more abundant in PE than PAf, in most cases flanked by either a 5' T or a 3' C. A few A→T transitions are more abundant in PAs than in PAf, but no difference approaches the significance level of $r_{\text{PE}}(m_{\text{TCC}})$ ($p < 4.38 \times 10^{-946}$, Pearson's $\chi^2 = 2176.8$). Combined with its reverse complement, m_{TCC} has frequency 3.12% in PE compared with 1.83% in PAf and 1.81% in PAs. As shown in Figure 1C, this 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. It is also reproducible in sequences obtained using Complete Genomics technology (see Supplemental Figure S1). The most parsimonious explanation for this result is that Europeans experienced a genetic change increasing the rate of m_{TCC} mutations.

Within each non-admixed population, there is relatively little variance in the frequency of m_{TCC}. To assess this, I let P_{total} denote the combined set of private variants from PE, PAs, and PAf, and for each haplotype h let $\text{P}_{\text{total}}(h)$ denote the subset of P_{total} whose derived alleles are found on haplotype h. $f_h(\text{TCC})$ then denotes the frequency of m_{TCC} within $\text{P}_{\text{total}}(h)$. 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.

Gradient of Asian ancestry within Europe

One pattern that is visible in Figure 2 is a north-to-south gradient of $f(\text{TCC})$ within Europe. The southern Spanish and Italian populations have the highest mean $f(\text{TCC})$ values (0.0335 and 0.0337, respectively), while the central British and CEU values (0.0325 and 0.0326) are intermediate and the northern Finnish value (0.0313) is lowest. A demographic event that may

be responsible for this gradient is gene flow into Europe from Asia.

Present-day Europeans have been inferred to share a complex recent demographic history, with genetic contributions from three differentiated ancient populations: western hunter-gatherers, eastern European farmers, and ancient north Eurasians (ANE) related to present-day Native Americans and to Upper Paleolithic hunter-gatherers from Siberia [19, 20, 21]. Using data from diverse ancient and modern European samples, Lazaridis, et al. inferred that the percentage of ANE ancestry ranges from $<1\%$ in Sardinia to $\sim 10\%$ in Italy and Southern Spain to 15–20% in Norwegians, Estonians, and Eastern Europeans. If TCC mutability increased in west Eurasians after they had already diverged from ANE, subsequent gene flow from ANE into west Eurasia could have lowered the frequency of segregating m_{TCC} non-uniformly across Europe.

Because East Asians share a more recent common ancestor with ANE than with west Eurasians [20], Europeans populations with high fractions of ANE ancestry should share more young SNPs with Asia than should European populations that have low ANE ancestry. This relationship can be verified by looking at the variant set PAsE of 913,662 SNPs that are fixed in Africa but variable in both Asia and Europe. For each haplotype h, I counted the number dPAsE(h) of derived alleles from the set PAsE that occur on haplotype h. As expected, dPAsE(h) is highest in Finland, lowest in Spain and Italy, and inversely correlated with $f_h(\mathrm{TCC})$ across all haplotypes h sampled in Europe (regression $p < 1.36 \times 10^{-14}$, Figure 3).

Discussion

It is beyond the scope of this article to pinpoint why the rate of m_{TCC} 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 m_{TCC} dominated the mutational spectra of single genes isolated from UV-irradiated cell cultures [5, 22]. Much more recently, it became possible to catalog the somatic mutations from cancer cells and

systematically identify the mutational signatures of particular mutagens. In this way, it was found that m_{TCC} is the most abundant mutation in the signature spectrum of melanoma, making up 28% of somatic point mutations on average [4, 23]. 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 m_{TCC} 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 m_{TCC} 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 reasonable 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 an individual rendered folate-deficient by sun exposure.

Although the data presented here do not reveal a clear underlying mechanism, they leave little doubt that the European population experienced a recent increase in the rate of one mutation type, which must have increased the total mutation rate unless many other mutation types decreased their rates in a zero-sum fashion. 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 [15, 14]. 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[15]. 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 [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, are not variable in YRI or LWK, and might or might not be variable in ASW, MXL, CLM, and PUR. Singleton

and doubleton variants are excluded to minimize the impact of possible sequencing error.

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. If $N_i(m)$ denotes the number of type-m SNPs in population P_i and M denotes the set of all possible SNPs, the frequency of $f_i(m)$ is defined as

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

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:

$$r_i(m) = \frac{N_i(m)}{N_i(m_0)}$$

The expected values of $N_1(m)$, $N_2(m)$, $N_1(m_0)$, and $N_2(m_0)$ under the expectation of no frequency difference are calculated as follows:

$$\begin{aligned}
\mathbb{E}(N_1(m)) &= \frac{(N_1(m)+N_1(m_0))(N_1(m)+N_2(m))}{N_1(m)+N_2(m)+N_1(m_0)+N_2(m_0)} \\
\mathbb{E}(N_2(m)) &= \frac{(N_2(m)+N_2(m_0))(N_1(m)+N_2(m))}{N_1(m)+N_2(m)+N_1(m_0)+N_2(m_0)} \\
\mathbb{E}(N_1(m_0)) &= \frac{(N_1(m)+N_1(m_0))(N_1(m_0)+N_2(m_0))}{N_1(m)+N_2(m)+N_1(m_0)+N_2(m_0)} \\
\mathbb{E}(N_2(m_0)) &= \frac{(N_2(m)+N_2(m_0))(N_1(m_0)+N_2(m_0))}{N_1(m)+N_2(m)+N_1(m_0)+N_2(m_0)}
\end{aligned}$$

To assess the significance of the difference $r_1(m) - r_2(m)$, a χ^2 value is calculated as

follows:

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

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 small and are bounded between 0 and 1.

Nonparametric bootstrap analysis

To assess the variance of $f(m_{\text{TCC}})$ 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 $f(m_{\text{TCC}})$ of the mutation m_{TCC} 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 m_{TCC} was computed for each SNP set restricted to the Y chromosome.

Complete Genomics data accession and analysis

To ensure that the results presented in this paper are not specific to a single sequencing platform or consortium, 62 human genomes sequenced by by Complete Genomics were downloaded from www.completegenomics.com /public-data/69-Genomes/ and analyzed. These 62 unrelated individuals include the 54-member CG diversity panel, the parents of the Yoruban trio and Puerto Rican trio, and the four grandparents of the 17-member CEPH pedigree. This dataset contains representatives of 11 populations: two European (CEU, TSI), three Asian (CHB, JPT, and GIH (Gujarati Indians from Houston)), three African (YRI, LWK, and MKK (Maasai from Kenya)), and three admixed (ASW, MXL, PUR). Using these data, population-private SNP sets were defined independently of the 1000 Genomes Data. Looking only at variation within the

Complete Genomics individuals, the private European set PE(CG) contains SNPs that are variable in CEU or TSI and not variable in CHB, JPT, GIH, YRI, LWK, or MKK. Similarly, the private Asian set PAs(CG) contains SNPs that are variable in CHB, JPT, or GIH and not variable in CEU, TSI, YRI, LWK, or MKK. The private African set PAf(CG) contains SNPs that are variable in YRI, LWK, or MKK and not variable in CEU, TSI, CHB, JPT, or GIH. Using the private SNP sets PE(CG), PAs(CG), and PAf(CG), fold differences $r_m(\text{PE(CG)}) - r_m(\text{PAf(CG)})$ and $r_m(\text{PAs(CG)}) - r_m(\text{PAf(CG)})$ were computed as described in "statistical analysis of frequency differences" along with their χ^2-based p values. The results are depicted in Figure S1, a volcano plot analogous to Figure 1A,B from the main text. It can be seen that m_{TCC} is again the major outlier in the comparison of Europe to Africa, with a significance that dwarfs that of all outliers in the Asia-to-Africa comparison.

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] Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).

[20] Lipson, M. *et al.* Efficient moment-based inference of of admixture parameters and sources of gene flow. *Mol Biol Evol* **30** (2013).

[21] Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).

[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).

Acknowledgements

I am grateful to Rasmus Nielsen for advice and manuscript comments, and to Stuart Linn, Elad Ziv, and members of the Nielsen lab for helpful discussions. This work was supported by a National Science Foundation Graduate Research Fellowship and NIH grant 2R14003229-07.

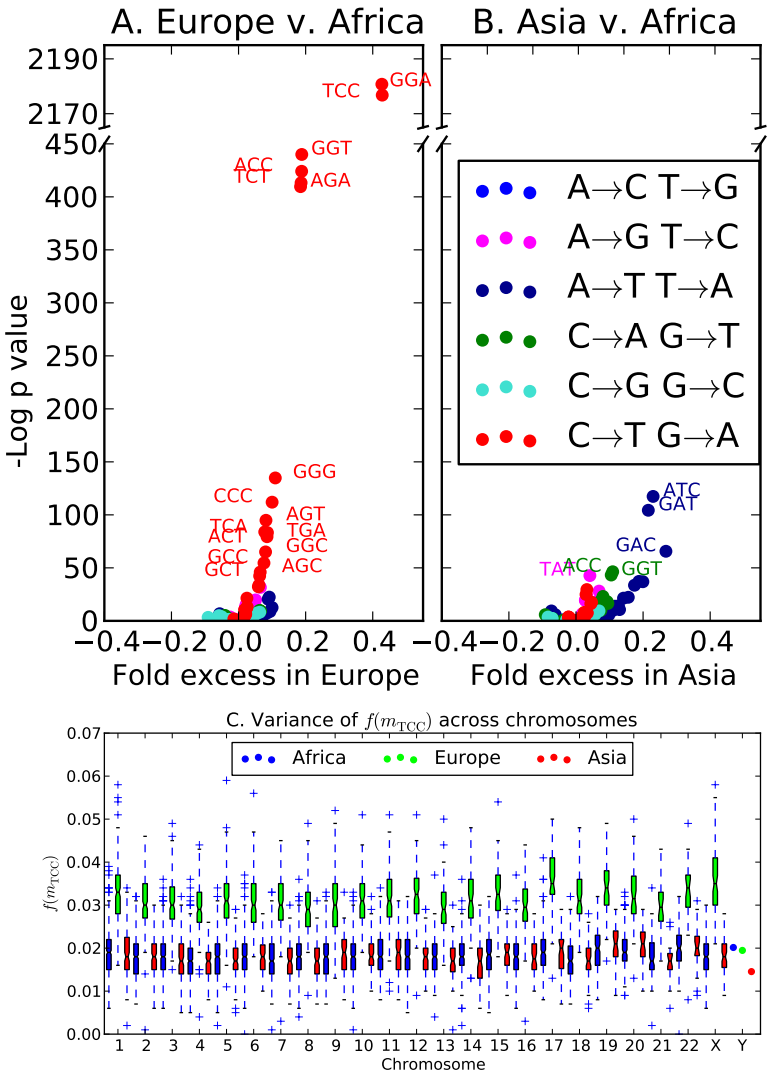


Fig. 1. Overrepresentation of 5'-TCC-3'→5'-TTC-3' within Europe. Panels A,B: The x coordinate of each point in gives the fold difference $r_m(\mathrm{PE}) - r_m(\mathrm{PAf})$ (resp. $r_m(\mathrm{PAs}) - r_m(\mathrm{PAf})$), while the y coordinate gives the Pearson's χ^2 p-value of its significance. 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. Panel C shows the distribution of m_{TCC} across bins of 1000 consecutive population-private SNPs. Only chromosome-wide frequencies are shown for Chromosome Y because of its low SNP count.

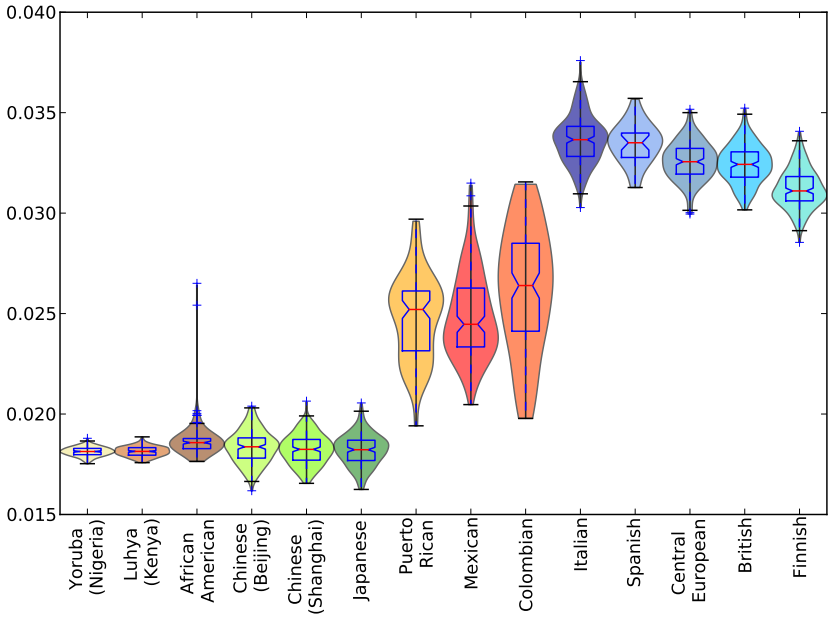


Fig. 2. Variation of $f(\mathbf{TCC})$ within and between populations. This violin plot shows the distribution of $f(\mathrm{TCC})$ within each 1000 Genomes population, i.e. the proportion of all derived variants from PA, PE, and PAf present in a particular genome that are m_{TCC} mutations. There is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

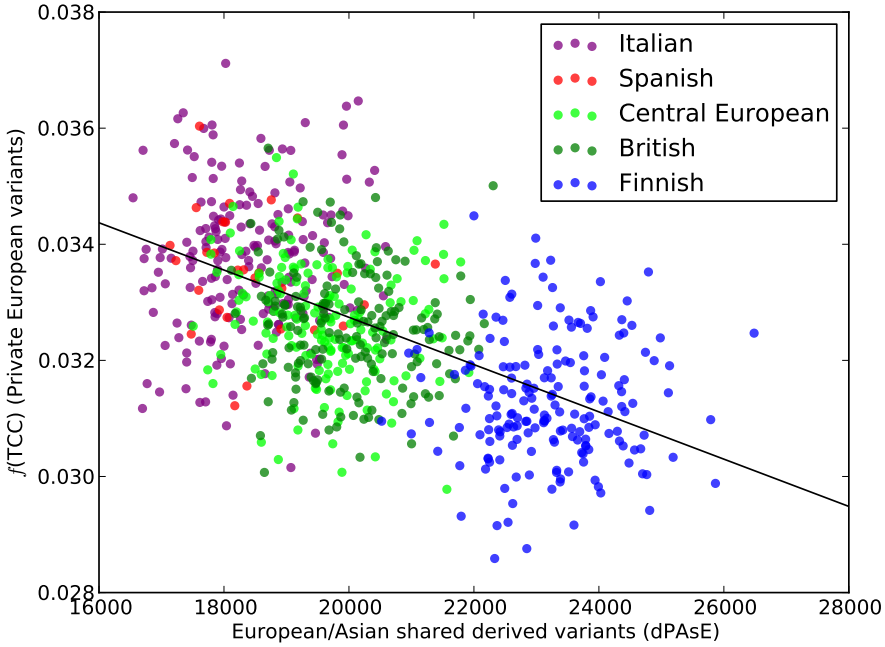
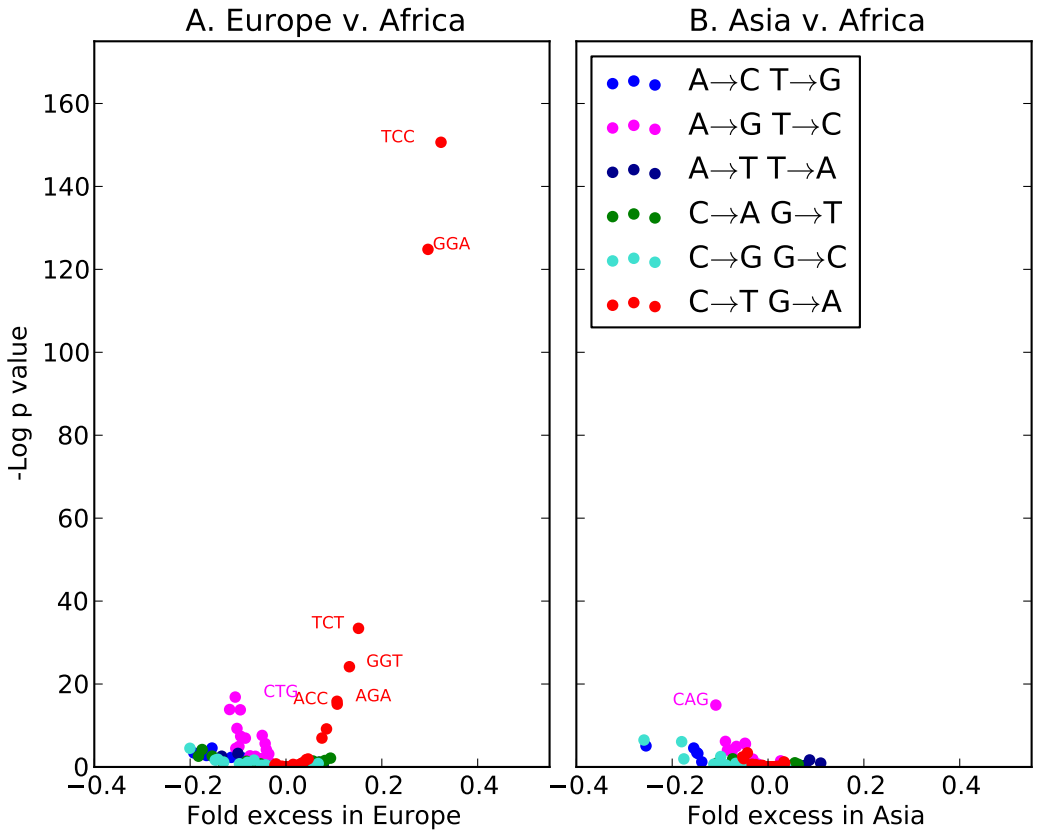


Fig. 3. Anticorrelation between $f(\mathbf{TCC})$ and number of young alleles shared with Asia. Each point in this scatterplot shows $f_h(\mathrm{TCC})$ and $\mathrm{dPAsE}(h)$ for a particular European haplotype h, revealing a negative correlation between the frequency of private European m_{TCC} variants and the number of SNPs shared with Asia but not with Africa (regression slope -4.07×10^{-7}, $p < 1.36 \times 10^{-14}$).



Supplementary Figure S1: This volcano plot, analogous to Figure 1A,B of the main text, represents differences between context-dependent mutation frequencies between the Complete Genomics populations. These differences have lower p-value significance from the 1000 Genomes-based values for two reasons. One reason is that the Complete Genomics data contains fewer individuals than the 1000 Genomes data and therefore has a smaller sample size of SNPs. The other reason is that population-private SNPs may be ascertained with less certainty, again due to the smaller sample size. For example, there is a higher chance that a polymorphism segregating in both Africa and Europe will appear to be private to Europe because no African variation happened to be sampled at that site. Despite these caveats, m_{TCC} is the clear outlier, with a 30% higher frequency in PE(CG) than in PAf(CG). The minor outliers TCT→ TTT, AGA→ AAA, GGT→GAT, and ACC→ ATC are also the most significant minor outliers from the 1000 Genomes analysis.

$B_{5'}B_AB_{3'}$	B_D	r(PE)	r(PAf)	r(PAs)	$-\log p$ value r(PE)$-r$(PAf)	$-\log p$ value r(PAs)$-r$(PAf)
GGA	A	8.56e-01	4.77e-01	4.90e-01	2.18e+03	2.05e+00
TCC	T	8.52e-01	4.75e-01	4.81e-01	2.18e+03	2.78e-01
GGT	A	6.81e-01	5.18e-01	5.00e-01	4.40e+02	3.65e+00
ACC	T	6.64e-01	5.06e-01	4.89e-01	4.24e+02	3.57e+00
TCT	T	6.59e-01	5.03e-01	5.28e-01	4.13e+02	8.05e+00
AGA	A	6.62e-01	5.06e-01	5.24e-01	4.10e+02	3.61e+00
GGG	A	5.81e-01	4.95e-01	4.80e-01	1.35e+02	3.19e+00
CCC	T	5.74e-01	4.96e-01	4.80e-01	1.12e+02	3.09e+00
AGT	A	6.72e-01	5.92e-01	5.93e-01	9.49e+01	5.72e-03
ACT	T	6.56e-01	5.81e-01	5.97e-01	8.39e+01	2.59e+00
TCA	T	5.68e-01	5.01e-01	5.12e-01	8.33e+01	1.43e+00
TGA	A	5.60e-01	4.94e-01	5.04e-01	7.94e+01	1.00e+00
GGC	A	5.20e-01	4.64e-01	4.78e-01	6.50e+01	2.83e+00
GCC	T	4.98e-01	4.47e-01	4.63e-01	5.46e+01	3.66e+00
AGC	A	5.52e-01	5.02e-01	5.36e-01	4.60e+01	1.62e+01
GCT	T	5.40e-01	4.93e-01	5.28e-01	4.24e+01	1.75e+01
GCA	T	4.79e-01	4.39e-01	4.61e-01	3.38e+01	7.44e+00
TGC	A	4.84e-01	4.45e-01	4.62e-01	3.22e+01	4.32e+00
TAA	G	4.13e-01	3.79e-01	4.16e-01	3.16e+01	2.81e+01
GAT	T	1.83e-01	1.66e-01	2.09e-01	2.24e+01	1.04e+02
ACG	T	1.20e+00	1.13e+00	1.21e+00	2.14e+01	2.49e+01
ATC	A	1.81e-01	1.64e-01	2.09e-01	2.04e+01	1.17e+02
TTA	C	4.23e-01	3.95e-01	4.24e-01	1.98e+01	1.70e+01
CAT	G	1.12e+00	1.07e+00	1.13e+00	1.53e+01	1.95e+01
TAT	G	1.08e+00	1.03e+00	1.13e+00	1.37e+01	4.27e+01
CGC	A	8.45e-01	8.07e-01	8.12e-01	1.31e+01	7.01e-02
TAG	G	3.13e-01	2.94e-01	3.06e-01	1.31e+01	3.70e+00
ATA	C	1.07e+00	1.03e+00	1.10e+00	1.27e+01	2.39e+01
GTT	A	9.68e-02	8.72e-02	1.05e-01	1.26e+01	3.72e+01
ACA	T	8.81e-01	8.43e-01	8.54e-01	1.20e+01	4.35e-01
CTA	C	3.13e-01	2.94e-01	3.14e-01	1.18e+01	1.09e+01
ATG	C	1.11e+00	1.07e+00	1.13e+00	1.12e+01	1.94e+01
CGT	A	1.18e+00	1.14e+00	1.22e+00	1.04e+01	2.95e+01
GGA	T	1.77e-01	1.65e-01	1.83e-01	9.87e+00	1.81e+01
CCC	A	1.61e-01	1.49e-01	1.62e-01	9.37e+00	9.89e+00
GGG	T	1.57e-01	1.46e-01	1.62e-01	9.28e+00	1.65e+01
GAG	T	8.04e-02	7.30e-02	8.00e-02	8.83e+00	6.17e+00
CCA	G	1.50e-01	1.39e-01	1.45e-01	8.15e+00	1.78e+00

Supplementary Table S1: Tables S1–S5 list all data associated with Figure 1A,B of the main text.

$B_{5'}B_AB_{3'}$	B_D	r(PE)	r(PAf)	r(PAs)	$-\log p$ value r(PE)$-r$(PAf)	$-\log p$ value r(PAs)$-r$(PAf)
TCC	A	1.77e-01	1.66e-01	1.81e-01	7.71e+00	1.13e+01
GAC	T	8.25e-02	7.56e-02	9.78e-02	7.34e+00	6.56e+01
GGG	C	1.75e-01	1.64e-01	1.77e-01	7.33e+00	9.39e+00
GCG	T	8.29e-01	8.01e-01	8.03e-01	6.84e+00	2.49e-03
TTA	A	1.36e-01	1.46e-01	1.34e-01	6.74e+00	9.36e+00
TGT	A	8.79e-01	8.50e-01	8.62e-01	6.55e+00	5.75e-01
CCT	T	6.82e-01	6.58e-01	6.80e-01	6.41e+00	4.33e+00
GCC	A	1.90e-01	1.80e-01	1.96e-01	6.23e+00	1.26e+01
CCC	G	1.75e-01	1.65e-01	1.77e-01	5.85e+00	6.32e+00
ACC	A	2.78e-01	2.65e-01	3.03e-01	5.81e+00	4.63e+01
GGT	T	2.81e-01	2.68e-01	3.05e-01	5.62e+00	4.32e+01
AGG	A	6.79e-01	6.57e-01	6.80e-01	5.30e+00	4.59e+00
CTC	A	7.91e-02	7.32e-02	8.20e-02	5.28e+00	9.93e+00
AAC	T	9.76e-02	9.10e-02	1.09e-01	5.26e+00	3.36e+01
TGC	T	1.99e-01	2.09e-01	2.02e-01	5.15e+00	1.60e+00
TGC	C	9.91e-02	1.06e-01	1.06e-01	4.96e+00	2.05e-09
CGA	A	7.39e-01	7.18e-01	7.40e-01	4.52e+00	3.80e+00
GTC	A	8.29e-02	7.75e-02	9.45e-02	4.10e+00	3.69e+01
TAA	T	1.27e-01	1.34e-01	1.25e-01	4.03e+00	5.78e+00
GCC	G	1.08e-01	1.14e-01	1.11e-01	3.80e+00	3.99e-01
AAC	G	3.51e-01	3.63e-01	3.72e-01	3.63e+00	1.07e+00
GGC	T	1.92e-01	1.84e-01	2.00e-01	3.61e+00	1.17e+01
TCG	T	7.43e-01	7.23e-01	7.55e-01	3.52e+00	8.01e+00
GCG	G	2.98e-02	3.28e-02	3.02e-02	3.31e+00	1.71e+00
TTC	G	1.10e-01	1.05e-01	1.13e-01	3.10e+00	5.32e+00
GTC	C	1.72e-01	1.79e-01	1.81e-01	2.98e+00	1.21e-01
CGT	T	6.05e-02	6.46e-02	5.83e-02	2.79e+00	5.81e+00
GCA	A	1.99e-01	2.07e-01	2.01e-01	2.55e+00	7.58e-01
TAC	C	4.64e-02	4.98e-02	5.17e-02	2.54e+00	4.34e-01
GAT	G	3.31e-01	3.21e-01	3.42e-01	2.52e+00	1.00e+01
GAC	G	1.75e-01	1.82e-01	1.81e-01	2.34e+00	1.69e-02
AAA	C	2.63e-01	2.71e-01	2.74e-01	2.26e+00	5.14e-02
GAA	T	8.19e-02	7.78e-02	9.11e-02	2.17e+00	2.22e+01
ATT	G	1.36e-01	1.41e-01	1.40e-01	2.12e+00	5.21e-02
ACG	A	6.57e-02	6.94e-02	7.08e-02	2.06e+00	1.20e-01
TTA	G	1.24e-01	1.29e-01	1.39e-01	1.89e+00	6.60e+00
TGG	A	5.48e-01	5.60e-01	5.44e-01	1.82e+00	2.78e+00
ATC	G	6.33e-02	6.67e-02	6.69e-02	1.75e+00	1.15e-03

Supplementary Table S2: Tables S1–S5 list all data associated with Figure 1A,B of the main text.

$B_{5'}B_AB_{3'}$	B_D	r(PE)	r(PAf)	r(PAs)	$-\log p$ value r(PE)$-r$(PAf)	$-\log p$ value r(PAs)$-r$(PAf)
CCG	A	4.90e-02	5.20e-02	5.14e-02	1.70e+00	1.74e-02
TTC	A	8.25e-02	7.89e-02	8.82e-02	1.57e+00	1.04e+01
ATT	C	8.87e-01	9.03e-01	9.36e-01	1.55e+00	6.28e+00
TTT	G	2.65e-01	2.72e-01	2.83e-01	1.50e+00	3.08e+00
AGC	C	1.34e-01	1.39e-01	1.39e-01	1.41e+00	2.89e-03
ACC	G	1.22e-01	1.26e-01	1.23e-01	1.33e+00	5.91e-01
GGT	C	1.23e-01	1.28e-01	1.29e-01	1.31e+00	1.02e-01
ATC	C	3.24e-01	3.16e-01	3.30e-01	1.22e+00	4.50e+00
GTT	G	7.60e-02	7.92e-02	7.60e-02	1.20e+00	8.88e-01
TCT	G	3.64e-01	3.56e-01	3.59e-01	1.15e+00	5.36e-02
AGA	T	2.61e-01	2.54e-01	2.81e-01	1.12e+00	2.30e+01
GTA	C	3.28e-01	3.21e-01	3.48e-01	1.10e+00	1.84e+01
GTC	G	4.00e-02	3.79e-02	3.98e-02	1.10e+00	6.33e-01
AAT	T	1.61e-01	1.57e-01	1.74e-01	1.03e+00	1.68e+01
AGT	C	2.18e-01	2.23e-01	2.30e-01	8.96e-01	1.45e+00
GAG	G	2.12e-01	2.17e-01	2.13e-01	8.50e-01	3.61e-01
TGG	C	1.44e-01	1.40e-01	1.47e-01	8.17e-01	2.47e+00
GCA	G	1.02e-01	1.05e-01	1.01e-01	7.90e-01	6.56e-01
CTA	G	6.13e-02	6.37e-02	6.46e-02	7.77e-01	4.04e-02
CGG	A	9.93e-01	9.80e-01	9.77e-01	7.65e-01	9.47e-03
CCA	T	5.60e-01	5.68e-01	5.58e-01	7.59e-01	8.92e-01
GAG	C	8.83e-02	9.12e-02	8.96e-02	7.58e-01	1.21e-01
GCT	A	1.25e-01	1.21e-01	1.24e-01	7.48e-01	2.41e-01
AAT	C	1.43e-01	1.46e-01	1.42e-01	7.31e-01	8.01e-01
AGG	C	2.12e-01	2.07e-01	2.08e-01	7.08e-01	4.14e-04
CGG	T	4.79e-02	4.99e-02	4.63e-02	6.67e-01	2.21e+00
TCC	G	1.75e-01	1.71e-01	1.76e-01	6.33e-01	8.38e-01
CGC	C	2.86e-02	3.01e-02	2.76e-02	5.95e-01	1.60e+00
GAA	C	1.08e-01	1.05e-01	1.08e-01	5.38e-01	5.49e-01
ATT	A	1.54e-01	1.51e-01	1.64e-01	5.35e-01	1.14e+01
AGT	T	1.47e-01	1.43e-01	1.52e-01	5.34e-01	4.05e+00
CGA	T	4.15e-02	4.32e-02	4.41e-02	5.08e-01	7.74e-02
CGC	T	6.95e-02	7.16e-02	6.88e-02	4.95e-01	7.55e-01
TAT	C	1.06e-01	1.09e-01	1.11e-01	4.83e-01	8.37e-02
ACA	G	2.12e-01	2.08e-01	2.12e-01	4.81e-01	4.11e-01
AAC	C	8.04e-02	8.27e-02	8.39e-02	4.61e-01	5.47e-02
TGT	T	2.28e-01	2.24e-01	2.34e-01	4.54e-01	2.95e+00
TAA	C	1.23e-01	1.26e-01	1.31e-01	4.51e-01	1.91e+00

Supplementary Table S3: Tables S1–S5 list all data associated with Figure 1A,B of the main text.

$B_{5'}B_AB_{3'}$	B_D	r(PE)	r(PAf)	r(PAs)	$-\log p$ value r(PE)$-r$(PAf)	$-\log p$ value r(PAs)$-r$(PAf)
TGG	T	1.52e-01	1.55e-01	1.60e-01	4.17e-01	1.16e+00
CAG	T	9.91e-02	1.02e-01	1.05e-01	4.12e-01	9.82e-01
GTT	C	3.45e-01	3.50e-01	3.72e-01	4.06e-01	1.01e+01
TTG	G	1.58e-01	1.55e-01	1.66e-01	4.06e-01	6.93e+00
TCA	G	1.91e-01	1.88e-01	1.94e-01	3.90e-01	1.38e+00
GGC	C	1.14e-01	1.16e-01	1.13e-01	3.43e-01	4.57e-01
CTG	A	9.98e-02	1.02e-01	1.04e-01	3.41e-01	1.19e-01
AAG	G	3.47e-01	3.51e-01	3.63e-01	3.32e-01	2.76e+00
TCT	A	2.61e-01	2.57e-01	2.81e-01	3.02e-01	1.83e+01
GGA	C	1.73e-01	1.70e-01	1.73e-01	2.96e-01	1.41e-01
TGA	T	1.67e-01	1.64e-01	1.71e-01	2.78e-01	2.14e+00
GTA	G	4.81e-02	4.95e-02	5.23e-02	2.65e-01	1.13e+00
AAT	G	9.17e-01	9.25e-01	9.59e-01	2.49e-01	6.65e+00
GTG	C	3.90e-01	3.95e-01	3.92e-01	2.43e-01	5.39e-02
CAA	G	3.71e-01	3.67e-01	3.74e-01	2.24e-01	5.14e-01
CGG	C	4.76e-02	4.88e-02	4.52e-02	2.04e-01	2.31e+00
GAC	C	3.98e-02	3.87e-02	3.95e-02	1.76e-01	4.52e-02
TAG	T	5.78e-02	5.65e-02	5.85e-02	1.70e-01	4.11e-01
TGT	C	2.11e-01	2.13e-01	2.14e-01	1.64e-01	2.47e-03
TAT	T	1.47e-01	1.49e-01	1.51e-01	1.51e-01	2.11e-01
CCT	G	2.10e-01	2.08e-01	2.12e-01	1.50e-01	3.74e-01
CAT	T	1.94e-01	1.91e-01	2.08e-01	1.46e-01	1.33e+01
TTC	C	2.58e-01	2.60e-01	2.67e-01	1.26e-01	9.91e-01
CAC	C	1.06e-01	1.08e-01	1.13e-01	1.09e-01	1.58e+00
CAC	G	3.91e-01	3.94e-01	3.91e-01	9.70e-02	8.71e-02
GTA	A	5.64e-02	5.75e-02	6.54e-02	9.63e-02	1.06e+01
CCG	G	5.12e-02	5.22e-02	4.76e-02	9.27e-02	3.61e+00
AAG	T	7.94e-02	7.82e-02	8.33e-02	8.21e-02	2.78e+00
CAG	G	5.40e-01	5.43e-01	5.34e-01	7.71e-02	8.00e-01
TCG	G	4.74e-02	4.83e-02	4.83e-02	7.46e-02	1.72e-05
CTG	G	1.41e-01	1.40e-01	1.40e-01	7.42e-02	5.92e-03
CTT	A	7.31e-02	7.42e-02	8.00e-02	7.27e-02	4.01e+00
TAG	C	6.55e-02	6.45e-02	6.53e-02	7.16e-02	2.11e-02
GAT	C	6.50e-02	6.60e-02	6.64e-02	7.09e-02	3.33e-03
GCG	A	7.63e-02	7.52e-02	7.44e-02	6.75e-02	2.08e-02
CAT	C	1.61e-01	1.63e-01	1.75e-01	6.24e-02	7.60e+00
TCG	A	4.50e-02	4.58e-02	4.71e-02	5.77e-02	1.53e-01
CGT	C	4.43e-02	4.51e-02	4.17e-02	5.28e-02	2.13e+00
CTA	A	5.63e-02	5.72e-02	5.92e-02	4.95e-02	4.36e-01

Supplementary Table S4: Tables S1–S5 list all data associated with Figure 1A,B of the main text.

$B_{5'}B_AB_{3'}$	B_D	r(PE)	r(PAf)	r(PAs)	$-\log p$ value r(PE)$-r$(PAf)	$-\log p$ value r(PAs)$-r$(PAf)
ACA	A	2.25e-01	2.27e-01	2.28e-01	4.27e-02	1.69e-03
TAC	T	5.62e-02	5.70e-02	6.29e-02	3.88e-02	5.69e+00
TTG	A	7.78e-02	7.69e-02	8.75e-02	3.76e-02	1.41e+01
GTG	G	1.05e-01	1.06e-01	1.12e-01	3.35e-02	2.43e+00
TTG	C	3.83e-01	3.81e-01	3.88e-01	2.38e-02	4.80e-01
CTT	C	3.41e-01	3.43e-01	3.47e-01	2.32e-02	2.33e-01
CAG	C	1.39e-01	1.40e-01	1.43e-01	2.11e-02	2.16e-01
ATG	G	1.62e-01	1.61e-01	1.70e-01	1.84e-02	3.84e+00
ATA	G	1.08e-01	1.08e-01	1.10e-01	1.69e-02	9.07e-02
CGA	C	4.56e-02	4.51e-02	4.43e-02	1.50e-02	5.88e-02
CCA	A	1.58e-01	1.59e-01	1.64e-01	1.50e-02	8.65e-01
CTC	C	2.17e-01	2.18e-01	2.14e-01	1.34e-02	2.77e-01
ACT	A	1.44e-01	1.45e-01	1.52e-01	1.20e-02	3.28e+00
AAA	T	1.50e-01	1.50e-01	1.57e-01	1.03e-02	2.94e+00
TTT	C	4.88e-01	4.86e-01	4.96e-01	9.47e-03	9.60e-01
AGC	T	1.24e-01	1.25e-01	1.21e-01	6.79e-03	5.89e-01
CAC	T	8.95e-02	8.90e-02	1.00e-01	5.68e-03	1.38e+01
ACG	G	4.90e-02	4.86e-02	4.74e-02	5.52e-03	1.40e-01
CTC	G	9.05e-02	9.10e-02	8.87e-02	5.50e-03	3.01e-01
GAA	G	2.49e-01	2.48e-01	2.53e-01	4.60e-03	5.20e-01
TCA	A	1.67e-01	1.67e-01	1.71e-01	3.17e-03	5.51e-01
AGG	T	1.28e-01	1.27e-01	1.35e-01	2.83e-03	3.87e+00
GTG	A	9.01e-02	8.97e-02	1.04e-01	1.98e-03	2.14e+01
TGA	C	1.89e-01	1.88e-01	1.92e-01	1.38e-03	2.54e-01
CAA	T	7.35e-02	7.32e-02	7.89e-02	1.16e-03	3.72e+00
AAA	G	4.86e-01	4.87e-01	4.95e-01	1.06e-03	6.13e-01
TTT	A	1.49e-01	1.48e-01	1.56e-01	6.85e-04	3.72e+00
AGA	C	3.53e-01	3.54e-01	3.57e-01	4.23e-04	6.09e-02
GCT	G	1.35e-01	1.35e-01	1.34e-01	2.24e-04	1.01e-02
ATA	A	1.48e-01	1.48e-01	1.58e-01	1.55e-04	5.11e+00
CCT	A	1.29e-01	1.29e-01	1.31e-01	7.19e-05	1.10e-01
CTT	G	1.31e-01	1.31e-01	1.32e-01	5.77e-05	5.37e-02
AAG	C	1.32e-01	1.32e-01	1.37e-01	5.55e-05	8.26e-01
ATG	A	1.92e-01	1.92e-01	2.10e-01	2.62e-05	1.41e+01
ACT	G	2.20e-01	2.20e-01	2.17e-01	1.81e-05	7.98e-02
CAA	C	1.54e-01	1.54e-01	1.61e-01	5.88e-06	2.20e+00
TAC	G	3.26e-01	3.26e-01	3.52e-01	1.22e-06	1.68e+01
CTG	C	5.43e-01	5.43e-01	5.43e-01	1.17e-08	3.80e-04

Supplementary Table S5: Tables S1–S5 list all data associated with Figure 1A,B of the main text.