Supporting Information: Recent Evolution of the Mutation Rate and Spectrum in Europeans

Kelley Harris

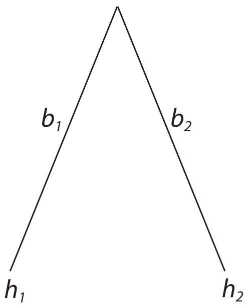
Contents

1 A branch length ratio test for mutation rate change

Consider two haplotypes h_1 and h_2 sampled from populations P_1 and P_2, respectively. At each locus, h_1 and h_2 are related by a very simple coalescent tree consisting of two branches b_1 and b_2 (Figure S1). If no mutation rate changes have occurred since the divergence of P_1 and P_2, the number of mutations falling on branches b_1 and b_2 are expected to be equal regardless of population demographic history. This reasoning motivates a simple branch length ratio test statistic (**BLR**).

Given a panel of sequences from P_1 and a panel of sequences from P_2, $\mathbf{BLR}(P_1, P_2)$ is computed by subsampling one haplotype from each panel, counting the derived alleles that appear on only the P_1 haplotype, and likewise counting the derived alleles that appear on only the P_2 haplotype. These counts can be summed across all haplotype pairs sampled from the two panels to yield derived allele counts D_1 and D_2. $\mathbf{BLR}(P_1, P_2)$ is then defined to be the ratio $\frac{D_1}{D_2}$, which is expected to equal 1 in the case of no mutation rate change.

Given a particular mutation type m (i.e. C→T transitions or TCC→T mutations), we can similarly define a branch length ratio $\mathbf{BLR}_m(P_1, P_2)$ that only counts derived alleles of mutation



Supplementary Figure S1: This simple two-lineage coalescent tree shows the branches b_1 and b_2 that lead backward in time from haplotypes h_1 and h_2 to their most recent common ancestor at a given locus. Branches b_1 and b_2 have equal length; in the absence of any mutation rate differences, h_1 and h_2 should contain equal numbers of derived alleles.

type m. It should hold that $\mathbf{BLR}_m(P_1, P_2) = 1$ for every mutation type m that has the same rate in P_1 and P_2.

Branch length ratios comparing Europeans (EUR) to Asians (ASN) were computed by subsampling pairs of haplotypes from the 1000 Genomes data and using the chimp PanTro4 reference to identify ancestral alleles. When $\mathbf{BLR}_m$(EUR, ASN) was computed separately for each transition and transversion type m, each derived allele count was greater in European sequences than Asian sequences, indicating that diverse mutation types appear to have higher rates in Europe than Asia. C→T mutations exhibited the largest apparent rate difference (Figure S2).

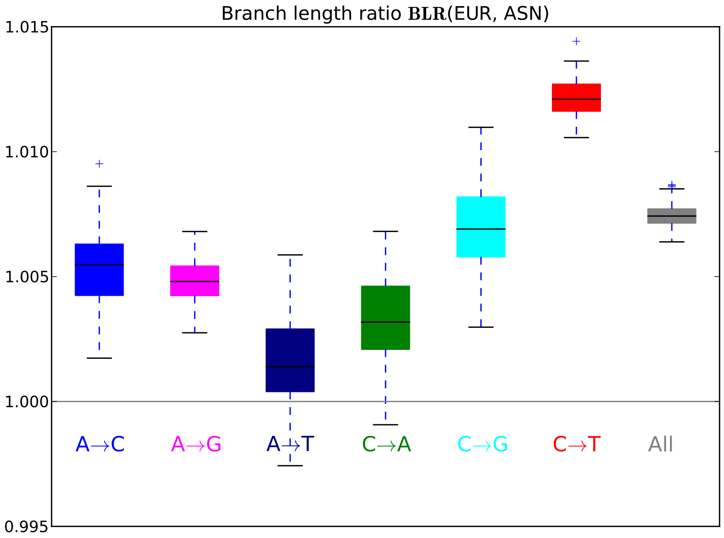
Nonparametric bootstrapping was used to estimate branch length ratio variance for each mutation type. The genome was divided into 100 bins with approximately equal SNP counts, and 100 replicates were generated by resampling 100 bins with replacement. A branch length ratio of 1 lies within the 95% confidence interval for only two mutation types, A→T and C→A. Each other mutation type appears to have a significantly higher rate in Europe than Asia.

Branch length ratios for context-dependent mutations yield a more complex picture, with numerous mutation types appearing to have equal rates in Asia and Europe and a few types (notably GAA→ GTA and CCG→CAG) appearing to have higher rates in Asia (Figure S3). As expected, TCC→T has the highest branch length ratio ($\mathbf{BLR}_{\mathrm{TCC}\to\mathrm{TTC}}$(EUR, ASN) > 1.04). Unexpectedly, the transversion type GAC→GCC has nearly as high a branch length ratio, indicating that this mutation is also a prime candidate for recent acceleration in Europe (or recent rate reduction in Asia). This signature merits further investigation, but since each transversion type is four times less common than a given transition type, GAC→GCC rate change makes less of an impact on total diversity than TCC→TTC rate change does.

2 Gradient of f(TCC) within Europe

One pattern that is visible in Figure 2 is a north-to-south gradient of f(TCC) within Europe. The southern Spanish and Italian populations have the highest mean f(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.

One demographic event that might have contributed to this f(TCC) gradient is gene flow from Asia into northeast Europe. Lazaridis, *et al.* and Hellenthal, *et al.* each inferred that the Finns have

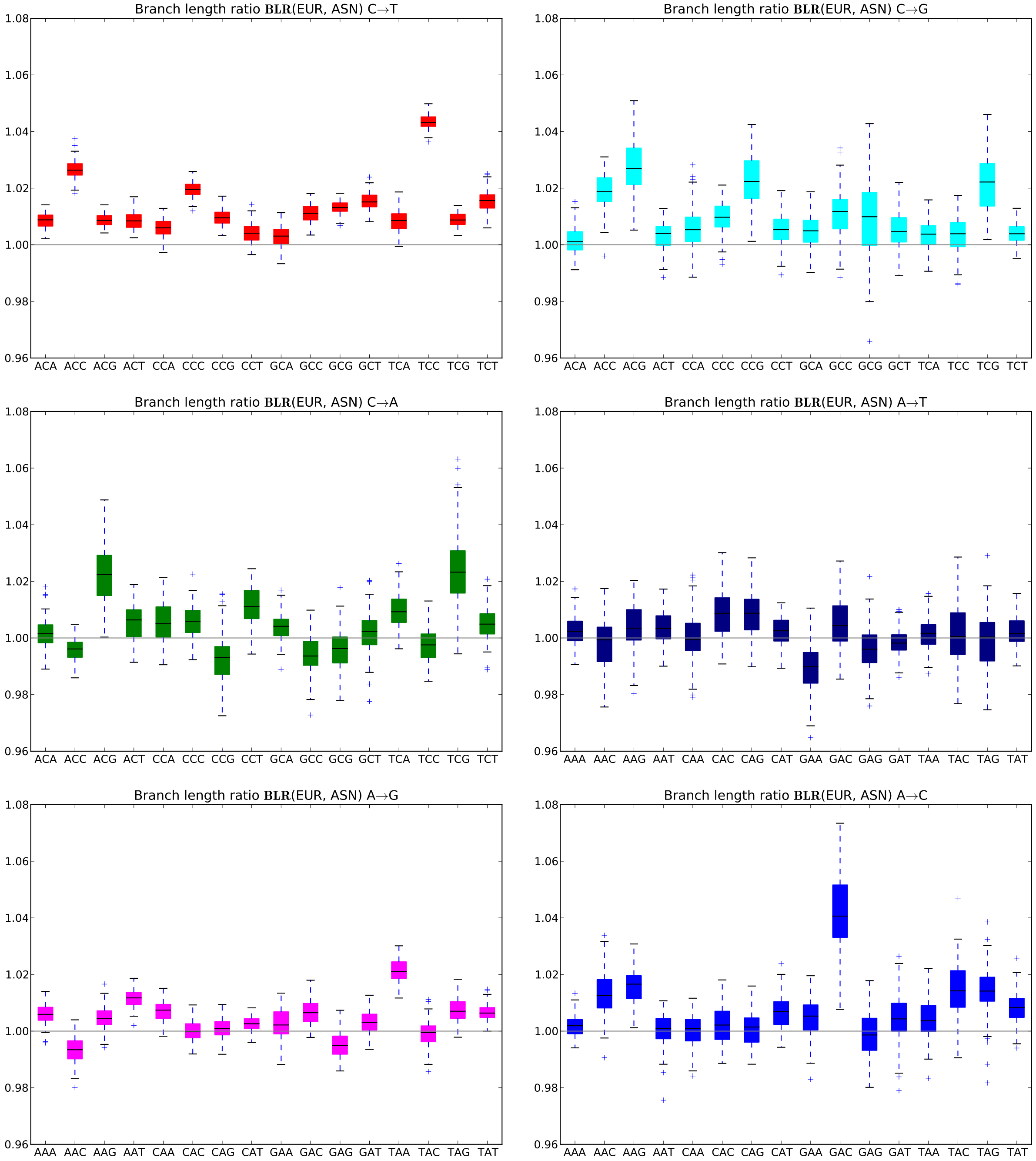


Supplementary Figure S2: This box plot shows branch length ratio distributions for each type of transition and transversion. For simplicity, each pair of complementary mutations (e.g. C→T and G→A) has been merged into a single category.

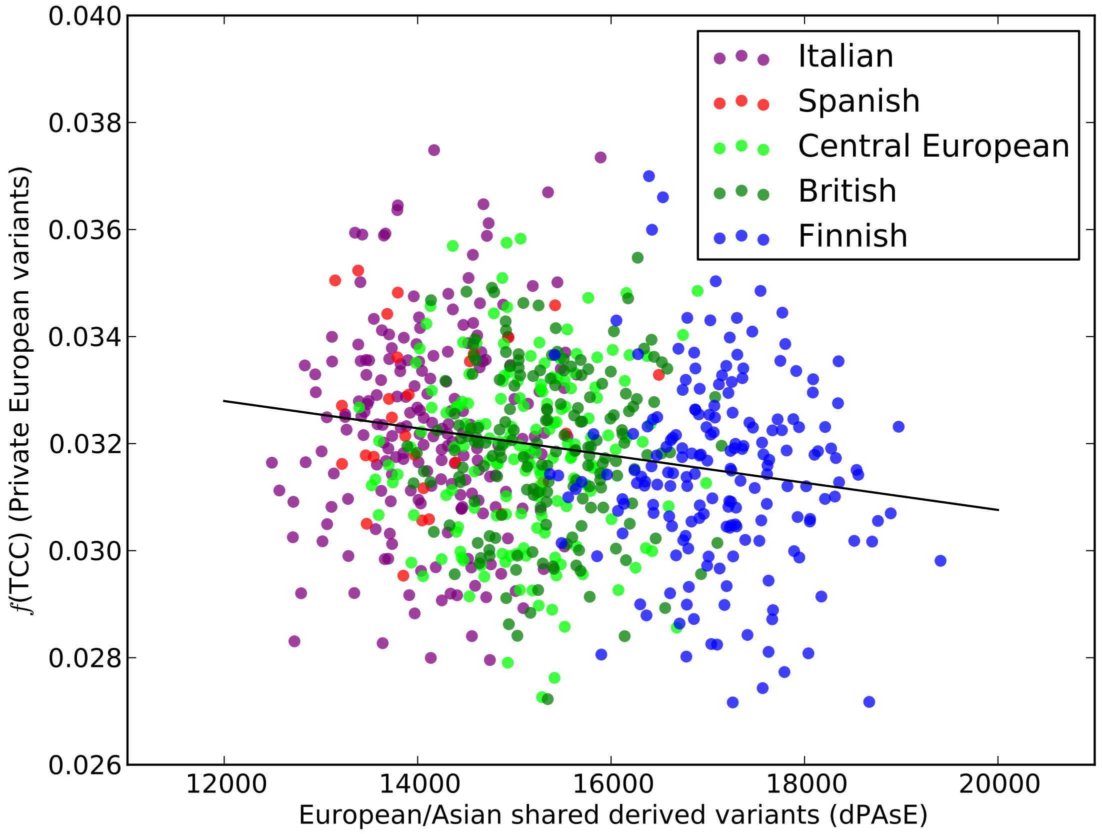
an Asian ancestry component, probably the result of gene flow from Siberia [1, 2]. In support of this, $f(\text{TCC})$ appears to be inversely correlated with European/Asian allele sharing. Specifically, I looked at the variant set PAsE of 913,662 SNPs that are fixed in Africa but variable in both Asia and Europe. For each European haplotype h, I counted the number dPAsE(h) of derived alleles from the set PAsE that occur on haplotype h. As expected given gene flow from Siberia into Finland, dPAsE(h) is highest in the Finns, lowest in the Spanish and Italians, and inversely correlated with $f_h(\text{TCC})$ across all haplotypes h sampled in Europe (regression $p < 2.17 \times 10^{-4}$, Figure 3).

A second possibility, not exclusive of the first, is that the $f(\text{TCC})$ gradient was created by ancient admixture among early European founder populations. Lazaridis, *et al.* have argued that Europeans are the admixed descendants of three genetically distinct groups: early European farmers (EEF), west European hunter-gatherers (WHG), and Ancient north Eurasians (ANE). The Italians, Spanish, and other southern European populations are inferred to have relatively high EEF ancestry fractions, compared to intermediate EEF ancestry levels in the English and low EEF ancestry in the Finns. This is consistent with a scenario where the TCC→TTC mutation rate change first occurred within the EEF population.

The light skin pigmentation alleles that are widespread in modern Europe are similarly believed to have originated in the EEF population, perhaps to compensate for low Vitamin D levels in a diet of cultivated grains. The hunter-gatherer populations that admixed with these early farmers show genetic indications of dark skin and hair [1, 3]. This is interesting in light of the association mentioned in the main text discussion between TCC→T mutations and melanoma, a skin cancer whose incidence is strongly predicted by light skin and European ancestry. There might also be some causal link between higher TCC→T frequencies and higher UV exposure at southern latitudes.



Supplementary Figure S3: These box plots shows branch length ratio distributions for context-dependent transition and transversion types. Each set of axes is restricted to a single mutation type (e.g. C→T transitions, C→G transversions, etc), and each individual bar plot is labeled with the ancestral nucleotide flaked by its 3' and 5' neighbors. Each complementary mutation pair (e.g. CCC→T and GGG→A) has been merged into a single mutation category.



Supplementary Figure S4: Each point in this scatterplot shows $f_h(\text{TCC})$ and $\text{dPAsE}(h)$ 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 (solid line, regression slope -2.54×10^{-7}, $p < 2.17 \times 10^{-4}$).

3 TCC→T mutations in Complete Genomics data

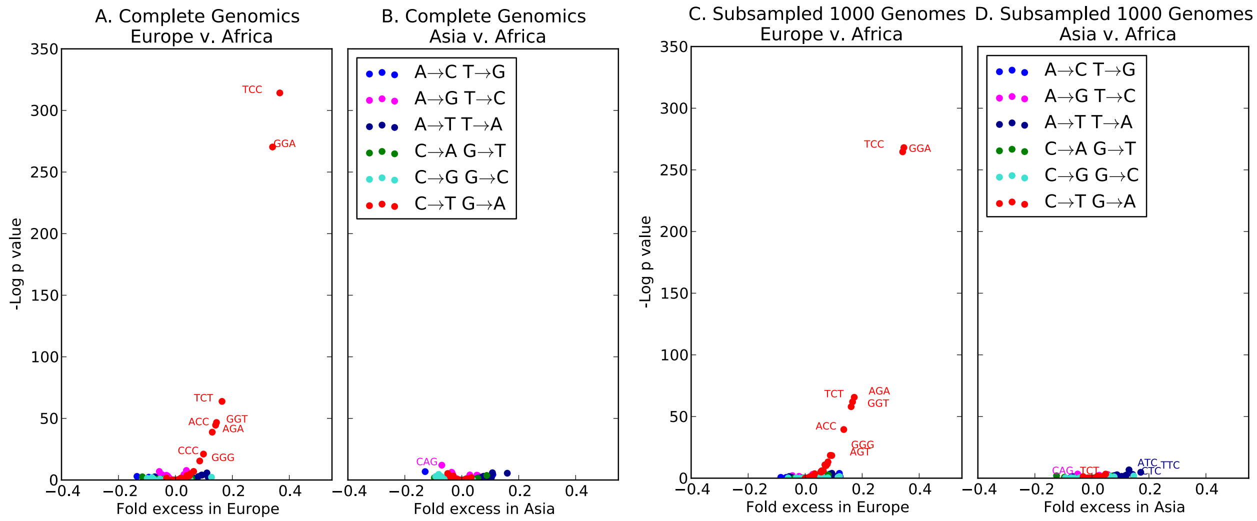
To ensure that the results presented in this paper are not specific to a single sequencing platform or consortium, 62 human genomes sequenced by by Complete Genomics (CG) were downloaded from `www.completegenomics.com /public-data/69-Genomes/` and analyzed [4]. These 62 unrelated individuals include 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). There are 13 Europeans, 12 Asians, 17 Africans, and 12 admixed individuals from the Americas.

Population-private SNP sets within the CG panel were defined independently of the 1000 Genomes data. Looking only at variation within the CG panel, 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 the 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. Singletons were excluded to minimize the impact of sequencing error. Using the private SNP sets PE(CG), PAs(CG), and PAf(CG), frequency differences $(f_{\mathrm{PE}}(m) - f_{\mathrm{PAf}}(m))/f_{\mathrm{PAf}}(m)$ and $(f_{\mathrm{PAs}}(m) - f_{\mathrm{PAf}}(m))/f_{\mathrm{PAf}}(m)$ were computed along with χ^2-based p values from the contingency tables in Figure1C,D of the main text.

The results are depicted in Figure S5A,B, a volcano plot analogous to Figure 1A,B from the main text. It can be seen that TCC→T 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. The minor outliers TCT→ TTT, AGA→ AAA, GGT→GAT, and ACC→ ATC are also significantly more abundant in Europeans than Asians or Africans in both CG and 1000 Genomes.

All outliers in Figure S5A have lower p-value significance than the corresponding outliers in Figure 1 of the main text. There are two reasons why the difference can be attributed to sample size. Intuitively, the CG data contains fewer individuals than the 1000 Genomes data and contains proportionately fewer SNPs. In addition, population-private SNPs in the CG data are ascertained with less certainty than population-private SNPs in 1000 Genomes. For example, if a particular SNP originated in Africa and is segregating today in both Africa and Europe, there is a chance that no Africans carrying the derived allele will be sampled, leading the SNP to be classified as private European variation. This misclassification should happen more often in a 62-genome panel than in a 1,092-genome panel.

To demonstrate that sample size can account for the difference between Figure 1 and Figure S5A,B, I subsampled 13 Europeans, 12 Asians, and 17 Africans from the 1000 Genomes data and ascertained the sets of mutations that appear to be continent-private with respect to this dataset. The population composition of the CG panel was mirrored except for substitution of 4 CHS individuals for GIH and 4 LWK individuals for MKK. Since the Complete Genomics data are not imputed and thus cannot be filtered for imputation quality, no imputation quality filtering was performed on the 1000 Genomes data for the purpose of this analysis. As shown in Figure S5C,D, all p values calculated from the subsampled 1000 Genomes panel are very close to the p values calculated from the CG panel.



Supplementary Figure S5: Panels A and B, analogous to Figure 1A,B of the main text, show differences in context-dependent mutation frequencies between the Complete Genomics populations. Panels C and D represent the same frequency differences in a panel subsampled from the 1000 Genomes data to mirror the sample size and population makeup of the CG panel.

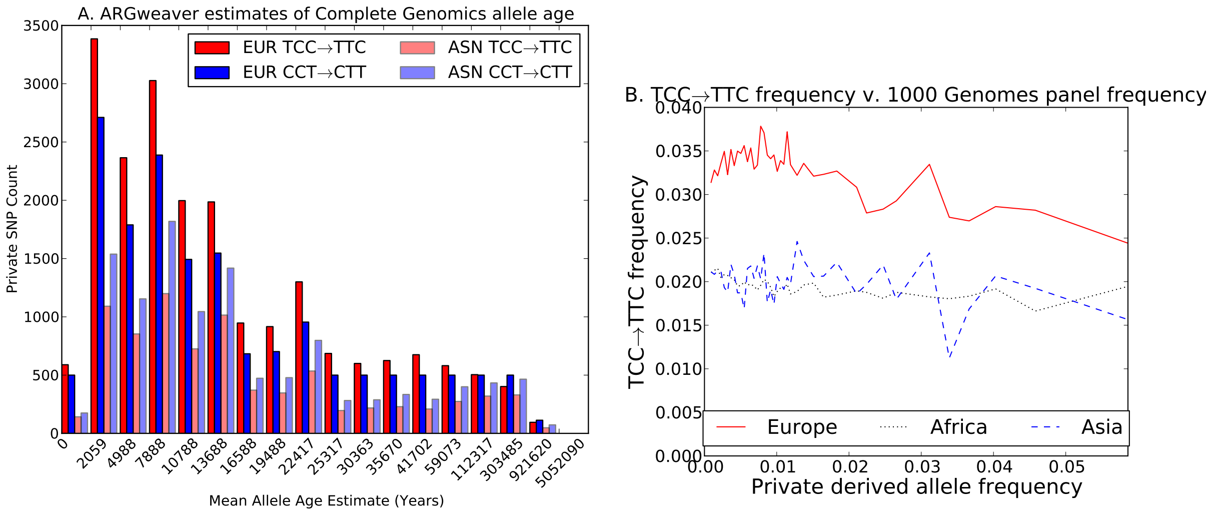
4 Estimating the time of TCC→TTC acceleration in Europe

The program ARGweaver recently developed by Rasmussen, et al. uses patterns of haplotype diversity to infer the approximate genealogical history of a collection of haplotypes [5]. This makes it possible to estimate the age of a particular derived allele based on the extent of divergence among the sequences that carry it. Rasmussen, et al. applied ARGweaver to the Complete Genomics data utilized in Section S3 and published the resulting estimates of genealogy and allele age on the server `http://compgen.bscb.cornell.edu/ARGweaver/CG_results/` .

Rather than estimating a single genealogy for the Complete Genomics data, ARGweaver generates a distribution of probable genealogies that reflect the uncertainty of estimating history from present-day diversity. This makes it possible to extract a mean derived allele age estimate for each SNP. ARGweaver estimates a posterior distribution of allele age by computing a posterior distribution of genealogies, identifying the branch of each genealogy on which the mutation must have occurred, and placing the mutation uniformly at random on this branch.

I downloaded the mean ARGweaver allele age estimates for each TCC→T variant that appears private to Europeans or Asians in both the Complete Genomics data and the 1000 Genomes data. As a control, I also downloaded mean allele age estimates for private CCT→CTT variants, which are similarly abundant to TCC→TTC variants but do not show much evidence of mutation rate divergence between Asia and Europe. These allele age estimates were grouped into bins such that each age bin spans at least 100 generations and contains at least 500 private European CCT→CTT variants. Figure S6A shows the number of private European and private Asian TCC→TTC and CCT→CTT variants that fall into each bin. ARGweaver's age estimates (given in generations) were converted to years before the present by assuming a human generation time of 29 years.

As shown in Figure S6A, each time bin contains slightly more private Asian CCT→CTT variants compared to private Asian TCC→TTC variants. However, the only bins containing more private European CCT→CTT variants than private European TCC→TTC variants are the two most ancient time bins ($>$ 303 thousand years ago (kya)). The next-most-ancient bin (112–303 kya)



Supplementary Figure S6: A. This bar plot shows the distribution of private European (EUR) and Asian (ASN) allele ages computed by Rasmussen, *et al.* from Complete Genomics data using ARGweaver [5]. For all but the most ancient alleles, TCC→TCC is more frequent in Europe in than the control mutation type CCT→CTT, indicating that TCC→TTC acceleration in Europe probably happened relatively soon after Europeans and Asians diverged. B. For this figure, private European, Asian, and African alleles from the 1000 Genomes data were partitioned into bins based on derived allele frequency (DAF). Within each DAF bin, the frequency of TCC→TTC (plus GGA→GAA) was calculated and plotted as a function of DAF. For private Asian and African alleles, TCC→TTC frequency appears to be independent of DAF (hovering around 2%). In contrast, private European TCC→TTC frequency is a decreasing function of DAF, indicating that higher DAF categories contain a lower percentage variants that arose later than the time of TCC→TTC acceleration in Europe.

contains 1% more private European TCC→TTC variants compared to CCT→CTT, and all other bins contain 16–37% more private European TCC→TTC variants compared to CCT→CTT, with the TCC→TTC-to-CCT→CTT ratio peaking around 25 kya. These ARGweaver estimates suggest that the European mutation rate changed very soon after Europeans diverged from Asians between 40,000 and 80,000 years ago [6], a result that is consistent with the relatively uniform distribution of excess TCC→TTC variants among diverse European populations.

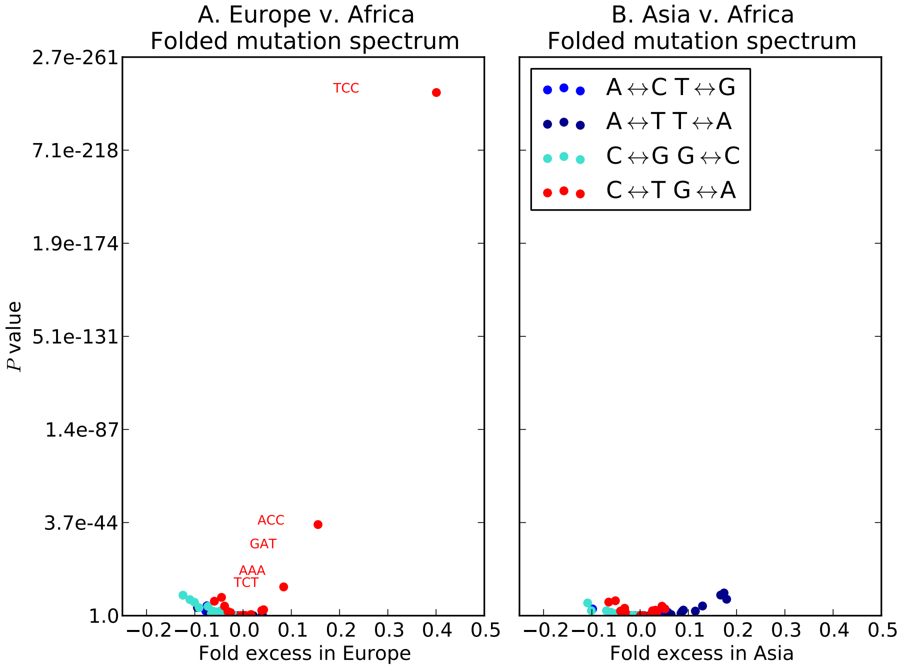
Due to sheer sample size, the 1000 Genomes dataset contains more information about allele age than the Complete Genomics dataset does. Figure S6B plots TCC→TTC frequency as a function of minor allele frequency in the 1000 Genomes data, showing that TCC→TTC SNPS comprise close to to 3.5% of private European variants that have less than 1% frequency in the entire 1000 Genomes panel. In contrast, TCC→TTC SNPs comprise fewer than 2.5% of older private European variants that occur in 5–6% of the 1000 Genomes haplotypes. Unfortunately, methods like ARGweaver are not yet scalable to datasets containing hundreds of haplotypes, making it hard to rigorously incorporate this information into a better estimate of the TCC→TTC acceleration time.

5 Controlling for ancestral misidentification

To polarize SNPs, we used the standard approach of inferring ancestral states from a human-chimpanzee reference genome alignment. This procedure is rooted in parsimony and is liable to error whenever the human and chimp populations are both variable at the same genomic site, violating the infinite sites model due to shared ancestral variation or separate coincident mutations. As noted by Hernandez, et al., different context-dependent mutations are liable to ancestral misidentification (AM) at different rates because nucleotide triplets with the highest mutation rates are most susceptible to coincident mutation on the human and chimp lineages [7]. This has the potential to confound the results in this paper if different populations are susceptible to AM at different rates, leading to erroneous inference of differences in context-dependent mutation rate between populations.

To verify that the results in this paper are not artifacts of AM, Figure 1 from the main text was replicated after "folding" the context-dependent mutation spectrum. Specifically, the set of 192 different mutation types $B_{5'}B_AB_{3'} \to B_{5'}B_DB_{3'}$ was collapsed into a set of 96 mutation classes, grouping $B_{5'}B_AB_{3'} \to B_{5'}B_DB_{3'}$ together with $B_{5'}B_DB_{3'} \to B_{5'}B_AB_{3'}$ into a class denoted $B_{5'}B_AB_{3'} \leftrightarrow B_{5'}B_DB_{3'}$. A χ^2 test was used to compare 95 of the mutation classes to the reference class 5'-CCG-3'↔5'-CTG-3', looking at the relative abundances of each mutation class in Europe and Asia versus Africa.

As shown in Figure S7, the main outliers from Figure 1 mostly correspond to outliers in the folded context-dependent frequency spectrum, implying that they are not artifacts of AM. The exceptions are transitions at CpG sites, which are known to be subject to higher rates of ancestral misidentification than other mutation types and appear to exhibit fewer frequency differences between populations once the mutation spectrum is folded. However, looking at either the folded data or unfolded data, categories TCC↔TTC, GAA↔GGA, GAT↔GGT, ACC↔ATC, TCT↔TTT, and AAA↔AGA, are all overrepresented in Europe compared to Africa, while AAC↔ATC, GAT↔GTT, and GAC↔GTC are all overrepresented in Asia compared to Africa.



Supplementary Figure S7: Panels A and B, analogous to Figure 1A,B of the main text, show differences in folded context-dependent mutation frequencies between African and non-African 1000 Genomes populations. The categories containing the main outliers from Figure1A,B are all outliers in this plot as well, indicating that the differences between populations are not artifacts of AM.

6 Stratifying the genome by GC content

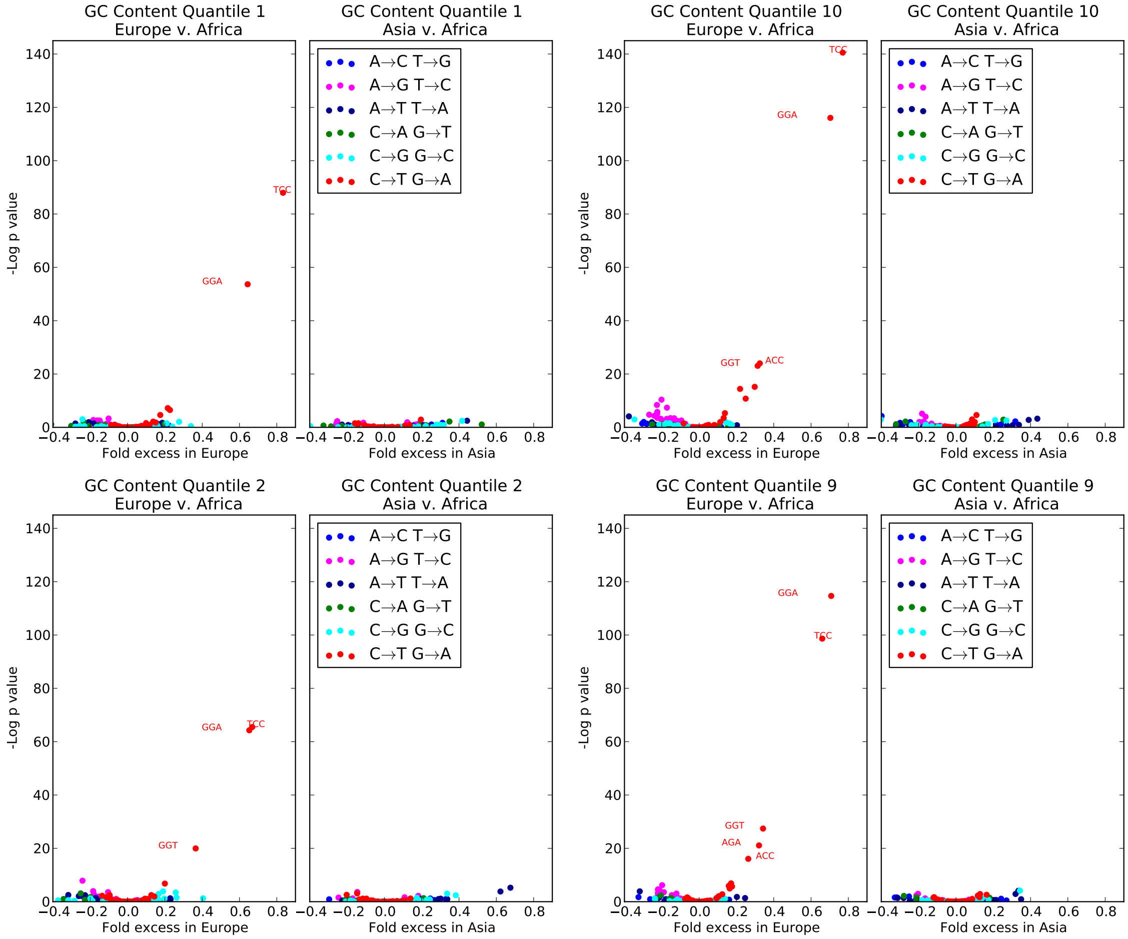
To assess the effect of GC content on mutation spectrum differences between the 1000 Genomes populations, the hg19 reference genome was partitioned into 100 kb bins. Within each bin, the ratio of G/C base pairs to total base pairs was computed, excluding sites annotated as N's. Bins containing more than 50% N's were excluded from the analysis. This yielded a distribution of GC content percentages ranging from 0% GC to 63.6% GC, which was partitioned into 10 quantiles containing the same number (2,862) of hg19 bins. Table S6 lists the upper and lower GC content percentages for each of the 10 quantiles.

Qtl	%GC	Qtl	%GC	Qtl	%GC	Qtl	%GC	Qtl	%GC
1	0–35.5	2	35.5–36.6	3	36.6–37.7	4	37.7–38.7	5	38.7–39.8
6	39.8–41.1	7	41.1–42.6	8	42.6–44.7	9	44.7–48.1	10	48.1–63.6

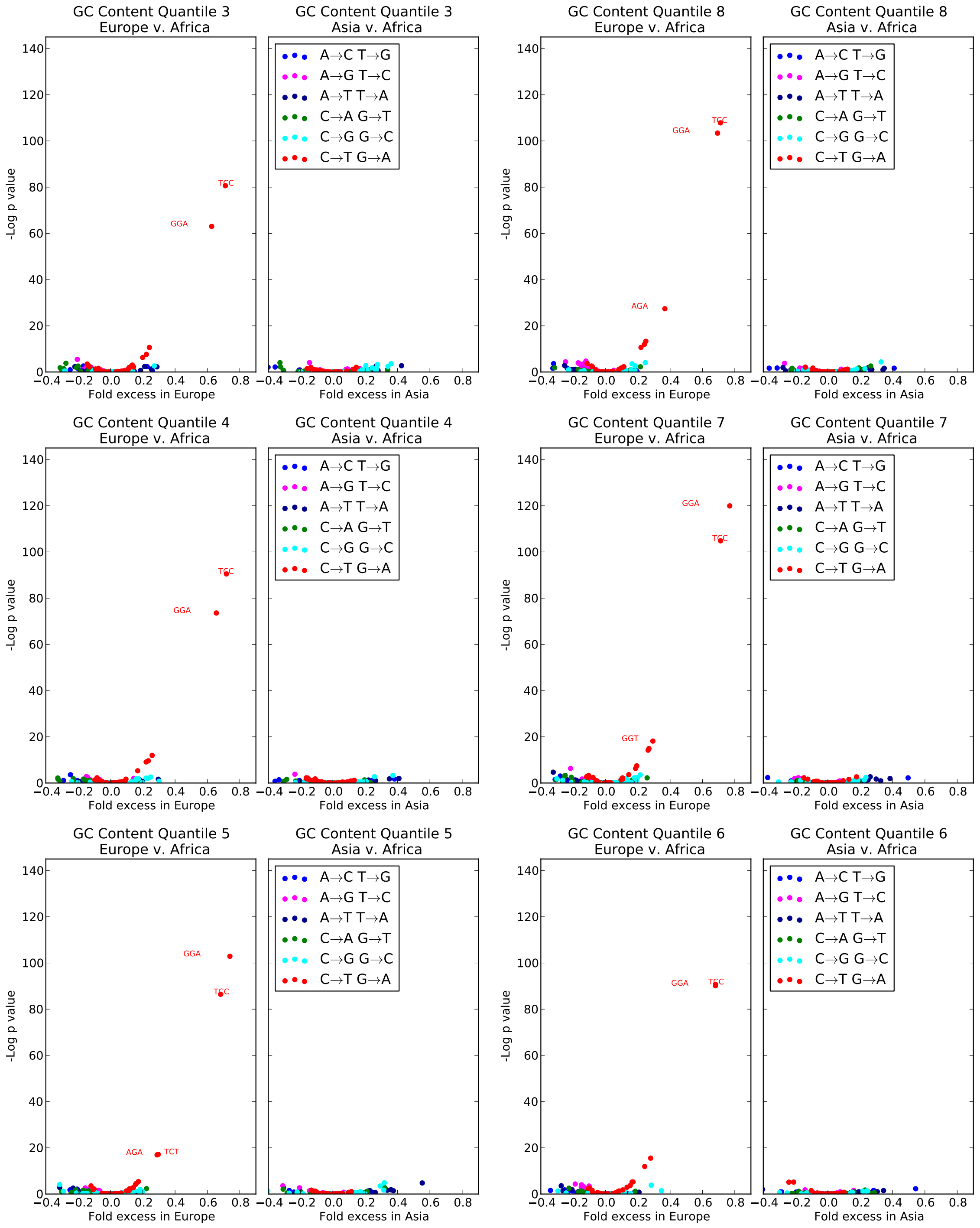
Lists of private European, Asian, and African SNPs were compiled within each GC quantile, and the χ^2 metric from Figure 1 of the main text was used to assess mutation spectrum differences. Tables S1–S10, one for each quantile, list the top-ranked 10 SNPs that show frequency differences between Europe and Africa. Similarly, Tables S11–S20 contain lists of the SNPs that show the most significant frequency differences between Asia and Africa.

The Europe v. Africa results show much more consistency across GC-content bins than do the Asia v. Africa results. Almost all Europe v. Africa outliers are C→T/G→A transitions, whereas the Asia v. Africa outliers contain a variety of transitions and transversions. In addition, 6 of the top 10 outliers for Europe v. Africa Quantile 1, including TCC→TTC/GGA→GAA, appear in the top 10 outliers for at least 9 out of 10 Europe v. Africa GC quantiles. In contrast, GAT→GTT is the only mutation that appears in the top 10 outliers for every Asia vs. Africa quantile. In each table, the column "T10-EUR" records the number of distinct GC quantiles for which a given mutation appears in the Europe v. Africa top 10; similarly, "T10-ASN" records the number of GC quantiles for which the mutation appears in the Europe v. Asia top 10. The Europe v. Africa outliers are mostly disjoint from the Asia v. Africa outliers.

Figures S8 and S9 contain a series of 10 volcano plots, each summarizing data from a single GC-content quantile. These plots are analogous to Figure 1A,B of the main text. The excess of TCC→TTC in Europe is the clearest feature of every plot, and the excess increases in magnitude with increasing GC content. A similar trend can be seen for minor outliers ACC→ATC, TCT→TTT, CCC→CTC and their reverse complements. Each C→T transition is more common in Europe than Africa across the genome, and the magnitude of this excess increases with increasing GC content.



Supplementary Figure S8: Mutation frequency differences between populations versus GC content, part 1.



Supplementary Figure S9: Mutation frequency differences between populations versus GC content, part 2.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) $-$ r(PAf)	T10-EUR	T10-ASN
TCC	T	1.464e-02	>	7.986e-03	1.833e+00	6.33e-39	10	0
GGA	A	1.336e-02	>	8.131e-03	1.644e+00	5.00e-24	10	0
AGA	A	1.279e-02	>	1.052e-02	1.216e+00	7.49e-04	9	0
ACC	T	1.066e-02	>	8.683e-03	1.227e+00	0.001	9	2
TCT	T	1.256e-02	>	1.068e-02	1.175e+00	0.010	10	0
ATT	C	2.050e-02		2.282e-02	8.983e-01	0.037	2	1
TCA	G	2.866e-03		3.779e-03	7.584e-01	0.048	1	0
GAA	G	4.950e-03		6.061e-03	8.167e-01	0.064	1	0
CAA	G	6.200e-03		7.400e-03	8.379e-01	0.074	1	0
GTT	C	7.347e-03		8.632e-03	8.511e-01	0.078	1	0

Supplementary Table S1: Quantile 1: 0–35.5% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) $-$ r(PAf)	T10-EUR	T10-ASN
TCC	T	1.418e-02	>	8.501e-03	1.668e+00	3.62e-29	10	0
GGA	A	1.444e-02	>	8.740e-03	1.653e+00	1.21e-28	10	0
GGT	A	1.207e-02	>	8.844e-03	1.365e+00	2.12e-09	9	0
CTG	C	6.096e-03		8.035e-03	7.586e-01	3.83e-04	2	1
TCT	T	1.283e-02	>	1.069e-02	1.200e+00	0.001	10	0
AAG	G	6.143e-03		7.531e-03	8.157e-01	0.019	1	2
TCT	G	8.562e-03	>	7.193e-03	1.190e+00	0.021	1	1
AAT	G	1.924e-02		2.148e-02	8.957e-01	0.028	1	2
CCT	G	4.506e-03	>	3.582e-03	1.258e+00	0.032	1	1
ACT	A	2.467e-03		3.294e-03	7.489e-01	0.044	1	0

Supplementary Table S2: Quantile 2: 35.5–36.6% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) $-$ r(PAf)	T10-EUR	T10-ASN
TCC	T	1.484e-02	>	8.671e-03	1.711e+00	9.55e-36	10	0
GGA	A	1.418e-02	>	8.717e-03	1.626e+00	4.24e-28	10	0
TCT	T	1.296e-02	>	1.045e-02	1.241e+00	2.41e-05	10	0
GGT	A	1.105e-02	>	9.036e-03	1.222e+00	4.87e-04	9	0
ACC	T	1.129e-02	>	9.412e-03	1.199e+00	0.002	9	2
CTT	C	5.997e-03		7.544e-03	7.948e-01	0.004	1	0
CCA	A	2.183e-03		3.013e-03	7.244e-01	0.023	1	0
TGG	A	8.752e-03		1.023e-02	8.560e-01	0.032	1	1
AGA	A	1.208e-02	>	1.064e-02	1.135e+00	0.049	9	0
GCT	T	1.082e-02	>	9.525e-03	1.136e+00	0.068	1	0

Supplementary Table S3: Quantile 3: 36.6–37.7% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
TCC	T	1.552e-02	>	9.034e-03	1.718e+00	5.00e-40	10	0
GGA	A	1.457e-02	>	8.800e-03	1.655e+00	1.11e-32	10	0
ACC	T	1.208e-02	>	9.602e-03	1.258e+00	6.45e-06	9	2
GGT	A	1.175e-02	>	9.519e-03	1.234e+00	6.72e-05	9	0
AGA	A	1.241e-02	>	1.016e-02	1.221e+00	1.09e-04	9	0
TCT	T	1.245e-02	>	1.066e-02	1.168e+00	0.005	10	0
CCG	A	3.108e-04		7.300e-04	4.257e-01	0.012	1	0
ATT	G	2.528e-03		3.360e-03	7.524e-01	0.028	1	1
TAC	G	6.278e-03		7.383e-03	8.503e-01	0.064	1	1
AAC	G	6.692e-03		7.804e-03	8.576e-01	0.074	1	2

Supplementary Table S4: Quantile 4: 37.7–38.7% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
GGA	A	1.648e-02	>	9.475e-03	1.739e+00	2.08e-45	10	0
TCC	T	1.569e-02	>	9.329e-03	1.682e+00	2.88e-38	10	0
TCT	T	1.302e-02	>	1.005e-02	1.296e+00	3.49e-08	10	0
AGA	A	1.343e-02	>	1.042e-02	1.288e+00	4.33e-08	9	0
GGT	A	1.179e-02	>	1.005e-02	1.173e+00	0.005	9	0
TGA	A	1.179e-02	>	1.017e-02	1.160e+00	0.010	2	0
ACC	T	1.151e-02	>	9.961e-03	1.155e+00	0.015	9	2
TGC	C	1.555e-03		2.267e-03	6.858e-01	0.017	1	0
CGG	A	1.212e-02		1.377e-02	8.797e-01	0.030	1	2
CCC	T	9.732e-03	>	8.526e-03	1.141e+00	0.060	3	1

Supplementary Table S5: Quantile 5: 38.7–39.8% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
TCC	T	1.606e-02	>	9.553e-03	1.681e+00	4.03e-40	10	0
GGA	A	1.598e-02	>	9.509e-03	1.680e+00	6.99e-40	10	0
GGT	A	1.274e-02	>	9.954e-03	1.279e+00	1.82e-07	9	0
AGA	A	1.275e-02	>	1.026e-02	1.243e+00	6.73e-06	9	0
ACC	T	1.152e-02	>	9.839e-03	1.170e+00	0.005	9	2
TCT	T	1.217e-02	>	1.045e-02	1.164e+00	0.006	10	0
CAC	G	5.405e-03		6.648e-03	8.130e-01	0.013	2	2
TTT	C	8.352e-03		9.805e-03	8.518e-01	0.019	1	0
CCA	G	3.538e-03	>	2.757e-03	1.283e+00	0.022	1	1
GGG	A	1.028e-02	>	8.917e-03	1.153e+00	0.026	3	1

Supplementary Table S6: Quantile 6: 39.8–41.1% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
GGA	A	1.740e-02	>	9.840e-03	1.769e+00	8.11e-53	10	0
TCC	T	1.709e-02	>	9.983e-03	1.712e+00	2.97e-46	10	0
GGT	A	1.341e-02	>	1.037e-02	1.293e+00	1.32e-08	9	0
TCT	T	1.296e-02	>	1.022e-02	1.269e+00	3.36e-07	10	0
AGA	A	1.283e-02	>	1.015e-02	1.264e+00	6.72e-07	9	0
ACC	T	1.234e-02	>	1.034e-02	1.193e+00	5.93e-04	9	2
GTG	C	5.353e-03		6.830e-03	7.838e-01	0.002	2	1
CCC	T	1.139e-02	>	9.599e-03	1.186e+00	0.002	3	1
TAA	T	1.577e-03		2.330e-03	6.769e-01	0.009	1	0
AGC	A	1.119e-02	>	9.781e-03	1.144e+00	0.027	1	0

Supplementary Table S7: Quantile 7: 41.1–42.6% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
TCC	T	1.696e-02	>	9.915e-03	1.711e+00	1.47e-47	10	0
GGA	A	1.691e-02	>	9.985e-03	1.693e+00	1.25e-45	10	0
AGA	A	1.295e-02	>	9.472e-03	1.367e+00	1.27e-12	9	0
ACC	T	1.278e-02	>	1.023e-02	1.250e+00	1.70e-06	9	2
TCT	T	1.237e-02	>	9.970e-03	1.241e+00	6.26e-06	10	0
GGT	A	1.306e-02	>	1.071e-02	1.220e+00	2.41e-05	9	0
TAC	T	5.039e-04		9.842e-04	5.120e-01	0.008	1	0
ATT	C	1.433e-02		1.631e-02	8.790e-01	0.009	2	1
GCG	G	3.546e-04		7.752e-04	4.574e-01	0.009	1	0
GAC	G	2.743e-03		3.645e-03	7.527e-01	0.012	1	0

Supplementary Table S8: Quantile 8: 42.6–44.7% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
GGA	A	1.813e-02	>	1.061e-02	1.708e+00	1.61e-50	10	0
TCC	T	1.741e-02	>	1.049e-02	1.660e+00	1.45e-43	10	0
GGT	A	1.456e-02	>	1.085e-02	1.343e+00	1.24e-12	9	0
AGA	A	1.271e-02	>	9.617e-03	1.321e+00	6.91e-10	9	0
ACC	T	1.386e-02	>	1.097e-02	1.264e+00	1.06e-07	9	2
GGG	A	1.377e-02	>	1.175e-02	1.171e+00	0.001	3	1
GTG	C	6.186e-03		7.713e-03	8.020e-01	0.002	2	1
CCC	T	1.377e-02	>	1.187e-02	1.159e+00	0.003	3	1
TGA	A	1.105e-02	>	9.396e-03	1.176e+00	0.003	2	0
TCT	T	1.116e-02	>	9.592e-03	1.164e+00	0.007	10	0

Supplementary Table S9: Quantile 9: 44.7–48.1% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
TCC	T	2.019e-02	>	1.141e-02	1.770e+00	9.36e-62	10	0
GGA	A	1.913e-02	>	1.123e-02	1.704e+00	4.00e-51	10	0
ACC	T	1.449e-02	>	1.093e-02	1.325e+00	3.83e-11	9	2
GGT	A	1.483e-02	>	1.129e-02	1.314e+00	9.66e-11	9	0
TCT	T	1.098e-02	>	8.455e-03	1.298e+00	2.56e-07	10	0
GGG	A	1.783e-02	>	1.462e-02	1.220e+00	5.49e-07	3	1
AGA	A	1.090e-02	>	8.722e-03	1.250e+00	2.11e-05	9	0
CAG	G	9.767e-03		1.223e-02	7.986e-01	3.09e-05	1	2
CAC	G	6.198e-03		8.007e-03	7.741e-01	2.28e-04	2	2
CTG	C	1.017e-02		1.229e-02	8.277e-01	6.29e-04	2	1

Supplementary Table S10: Quantile 10: 48.1–100% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
GAT	C	7.286e-04		1.559e-03	4.674e-01	0.052	0	1
CCT	T	1.384e-02	>	1.157e-02	1.196e+00	0.055	0	1
GAA	T	2.914e-03	>	2.019e-03	1.443e+00	0.084	0	1
TGG	C	3.195e-03	>	2.255e-03	1.417e+00	0.087	0	1
CAC	G	4.259e-03		5.699e-03	7.474e-01	0.098	2	2
GGA	T	3.923e-03	>	2.908e-03	1.349e+00	0.114	0	1
CAT	G	2.303e-02	>	2.052e-02	1.123e+00	0.149	0	2
AAT	G	1.962e-02		2.210e-02	8.875e-01	0.179	1	2
CTT	G	3.867e-03	>	2.949e-03	1.311e+00	0.186	0	1
AAG	G	8.911e-03	>	7.499e-03	1.188e+00	0.210	1	2

Supplementary Table S11: Quantile 1: 0–35.5% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
GTC	A	2.490e-03	>	1.486e-03	1.676e+00	0.005	0	3
GTA	A	2.149e-03	>	1.325e-03	1.622e+00	0.022	0	1
ATT	C	1.875e-02		2.193e-02	8.549e-01	0.027	2	1
CGT	A	1.626e-02		1.904e-02	8.542e-01	0.045	0	1
AGG	C	4.834e-03	>	3.617e-03	1.337e+00	0.050	0	2
GCG	T	6.788e-03		8.507e-03	7.979e-01	0.076	0	1
GGT	C	3.223e-03	>	2.331e-03	1.383e+00	0.090	0	1
TAA	G	1.109e-02	>	9.382e-03	1.182e+00	0.112	0	1
CGT	T	5.860e-04		1.127e-03	5.199e-01	0.159	0	1
CAT	G	2.261e-02	>	2.040e-02	1.109e+00	0.193	0	2

Supplementary Table S12: Quantile 2: 35.5–36.6% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
TGC	T	2.762e-03		4.107e-03	6.724e-01	0.017	0	1
AAT	G	1.778e-02		2.080e-02	8.552e-01	0.018	1	2
TCC	G	4.352e-03	>	3.199e-03	1.360e+00	0.027	0	1
AGT	C	6.277e-03	>	4.921e-03	1.276e+00	0.040	0	2
TCT	G	9.373e-03	>	7.788e-03	1.204e+00	0.065	1	1
TTG	A	2.594e-03	>	1.821e-03	1.424e+00	0.066	0	1
CAT	T	5.565e-03	>	4.391e-03	1.267e+00	0.073	0	2
GGG	C	3.431e-03	>	2.559e-03	1.341e+00	0.087	0	3
CCT	A	1.674e-03		2.478e-03	6.755e-01	0.112	0	2
GTT	G	1.297e-03		2.019e-03	6.426e-01	0.115	0	2

Supplementary Table S13: Quantile 3: 36.6–37.7% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
TAC	G	5.646e-03		7.383e-03	7.647e-01	0.022	1	1
AGC	C	3.723e-03	>	2.715e-03	1.371e+00	0.038	0	1
ACT	G	5.973e-03	>	4.750e-03	1.257e+00	0.067	0	1
ACC	T	8.019e-03		9.602e-03	8.351e-01	0.106	9	2
TGG	A	8.673e-03		1.029e-02	8.429e-01	0.114	1	1
GAC	T	2.046e-03	>	1.452e-03	1.409e+00	0.142	0	3
GGG	A	6.873e-03		8.214e-03	8.368e-01	0.164	3	1
GTT	A	2.496e-03	>	1.849e-03	1.350e+00	0.165	0	1
TGT	A	1.591e-02		1.787e-02	8.904e-01	0.168	0	1
TAG	C	2.086e-03	>	1.511e-03	1.381e+00	0.177	0	1

Supplementary Table S14: Quantile 4: 37.7–38.7% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
AGT	C	6.272e-03	>	4.757e-03	1.318e+00	0.008	0	2
GAC	T	2.464e-03	>	1.587e-03	1.553e+00	0.008	0	3
GTC	C	2.539e-03		3.678e-03	6.902e-01	0.029	0	1
AGG	C	5.414e-03	>	4.189e-03	1.292e+00	0.032	0	2
TCC	A	3.771e-03	>	2.854e-03	1.321e+00	0.065	0	1
ATT	G	2.091e-03		3.011e-03	6.943e-01	0.066	1	1
CAC	G	5.190e-03		6.538e-03	7.938e-01	0.069	2	2
CGG	T	3.360e-04		7.468e-04	4.499e-01	0.118	0	1
CCT	A	1.643e-03		2.373e-03	6.922e-01	0.124	0	2
CTT	A	2.240e-03	>	1.648e-03	1.359e+00	0.157	0	1

Supplementary Table S15: Quantile 5: 38.7–39.8% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
CCA	T	8.341e-03		1.059e-02	7.876e-01	0.006	0	1
GGC	A	6.155e-03		8.119e-03	7.582e-01	0.006	0	1
GTC	G	1.311e-03	>	8.498e-04	1.543e+00	0.102	0	1
CTA	A	7.649e-04		1.284e-03	5.958e-01	0.137	0	1
CAG	G	8.377e-03		9.756e-03	8.586e-01	0.165	1	2
GCA	A	4.480e-03	>	3.641e-03	1.230e+00	0.180	0	1
AAC	G	8.341e-03	>	7.196e-03	1.159e+00	0.199	1	2
GAA	C	2.914e-03	>	2.271e-03	1.283e+00	0.205	0	1
GTC	A	1.967e-03	>	1.463e-03	1.344e+00	0.226	0	3
CGG	A	1.362e-02		1.519e-02	8.967e-01	0.236	1	2

Supplementary Table S16: Quantile 6: 39.8–41.1% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
ATG	A	5.017e-03	>	3.976e-03	1.262e+00	0.068	0	1
GCA	T	1.003e-02	>	8.525e-03	1.177e+00	0.070	0	2
CCT	G	5.578e-03	>	4.516e-03	1.235e+00	0.087	1	1
GTT	G	1.052e-03		1.677e-03	6.274e-01	0.095	0	2
GTG	C	5.578e-03		6.830e-03	8.167e-01	0.102	2	1
ACC	T	8.806e-03		1.034e-02	8.515e-01	0.103	9	2
GTA	G	1.438e-03	>	9.615e-04	1.496e+00	0.108	0	1
GAC	T	1.965e-03	>	1.419e-03	1.384e+00	0.143	0	3
GAG	G	3.578e-03		4.513e-03	7.929e-01	0.155	0	2
AAG	G	6.104e-03		7.288e-03	8.376e-01	0.158	1	2

Supplementary Table S17: Quantile 7: 41.1–42.6% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
GGG	C	4.816e-03	>	3.625e-03	1.329e+00	0.013	0	3
CTC	C	3.365e-03		4.599e-03	7.317e-01	0.023	0	1
ACA	A	4.486e-03	>	3.546e-03	1.265e+00	0.075	0	1
CCC	T	8.940e-03		1.039e-02	8.604e-01	0.121	3	1
TTT	A	1.748e-03		2.421e-03	7.221e-01	0.146	0	1
ATC	A	3.794e-03	>	3.038e-03	1.249e+00	0.157	0	1
GGT	T	5.872e-03	>	4.931e-03	1.191e+00	0.171	0	1
ATA	G	1.188e-03		1.732e-03	6.858e-01	0.177	0	1
GTA	C	4.618e-03		5.595e-03	8.254e-01	0.180	0	1
CCG	T	1.768e-02		1.950e-02	9.069e-01	0.180	0	1

Supplementary Table S18: Quantile 8: 42.6–44.7% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
CCA	G	4.220e-03	>	3.141e-03	1.343e+00	0.016	1	1
CAT	T	4.350e-03	>	3.260e-03	1.334e+00	0.017	0	2
AAC	G	5.096e-03		6.396e-03	7.968e-01	0.053	1	2
GAT	T	3.506e-03	>	2.657e-03	1.319e+00	0.055	0	1
CGA	A	1.772e-02	>	1.569e-02	1.129e+00	0.057	0	1
GCA	T	1.045e-02	>	8.976e-03	1.164e+00	0.075	0	2
GAG	G	3.928e-03		4.979e-03	7.889e-01	0.091	0	2
TCG	T	1.733e-02	>	1.551e-02	1.117e+00	0.105	0	1
TGA	T	1.883e-03		2.611e-03	7.212e-01	0.115	0	1
CGC	C	4.869e-04		9.045e-04	5.383e-01	0.127	0	1

Supplementary Table S19: Quantile 9: 44.7–48.1% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
CTG	C	1.006e-02		1.229e-02	8.184e-01	0.006	2	1
ACG	T	3.382e-02	>	3.046e-02	1.110e+00	0.010	0	1
GTG	G	1.288e-03		2.140e-03	6.017e-01	0.014	0	1
CAG	G	1.024e-02		1.223e-02	8.372e-01	0.019	1	2
GTC	A	2.177e-03	>	1.516e-03	1.436e+00	0.039	0	3
CGG	A	3.642e-02	>	3.352e-02	1.087e+00	0.051	1	2
CGT	C	9.198e-04		1.537e-03	5.985e-01	0.053	0	1
CCC	A	4.967e-03	>	3.960e-03	1.254e+00	0.056	0	1
GTG	A	2.299e-03	>	1.654e-03	1.390e+00	0.062	0	1
GGG	C	6.714e-03	>	5.550e-03	1.210e+00	0.064	0	3

Supplementary Table S 20: Quantile 10: 48.1–100% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

7 Filtering the 1000 Genomes data for sequencing depth

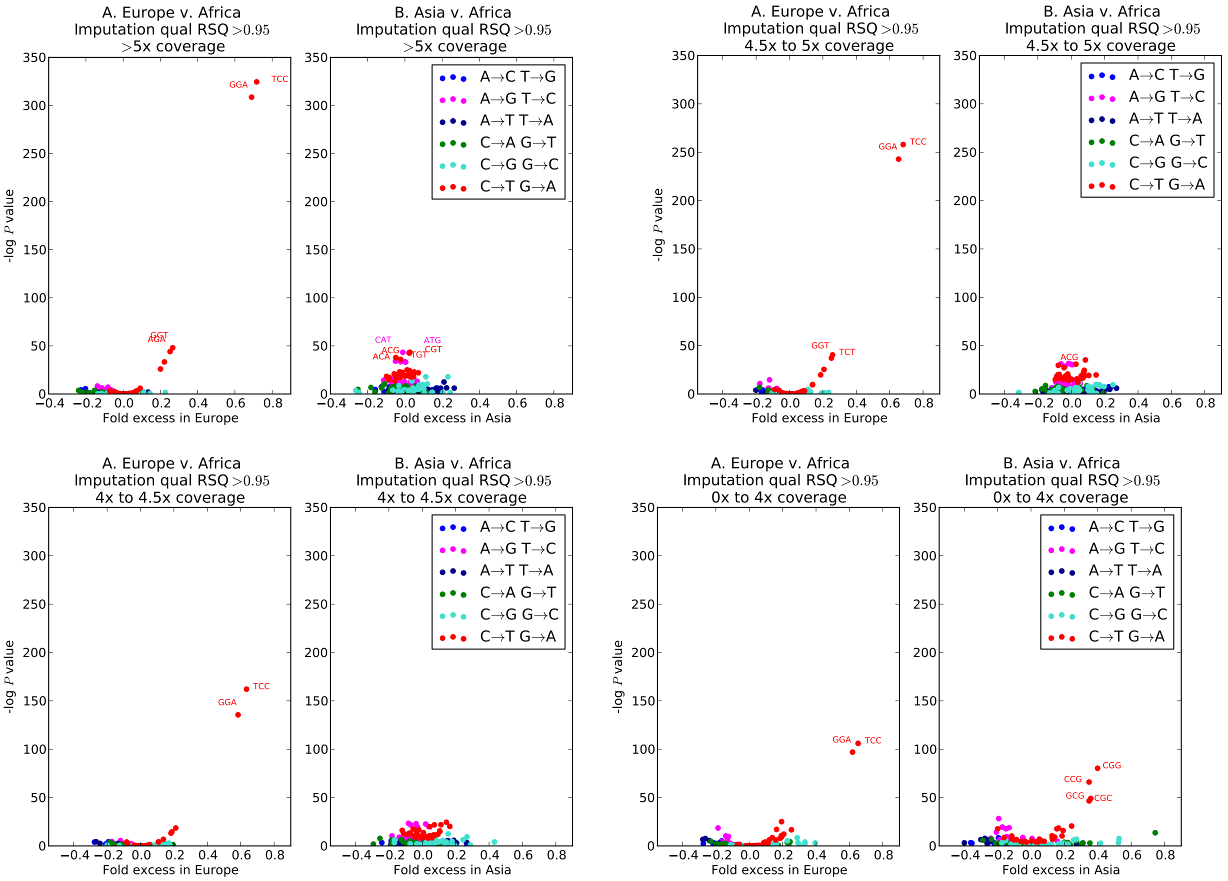
The 1000 Genomes Phase I samples were sequenced at low coverage (2x–6x per genome) [8]. Illumina read coverage can have biases related to GC content and other factors [9], which could lead to uneven discovery rates across the set of context-dependent SNP types.

To minimize the chance that low/uneven read depth has confounded the results in this paper, I partitioned the private SNP sets PE, PAs, and PAf into three separate bins based on read depth and assessed mutation spectrum differences within each bin separately. Each SNP was assigned to a bin as follows: First, the read depth for every individual at each SNP was extracted from the 1000 Genomes BAM files using the `samtools` "depth" command [10]. This made it possible to calculate the average depth for each SNP across European individuals, African individuals, and Asian individuals separately. Each SNP was placed into bin 1 (>5x), bin 2 (4.5–5x), bin 3 (4–4.5x), or bin 4 (0–4x) based on the minimum average depth per individual across the three continental groups. By these definitions, high-depth bin 1 contains 34.4% of PE, 35.7% of PAs and 28.2% of PAf. Bin 2 contains 29.3% of PE, 28.9% of PAs, and 28.6% of PAf, while bin 3 contains 21.6% of PE, 20.8% of PAs, and 24.2% of PAf. Low-depth bin 4 contains the remaining 14.7%, 14.6%, and 19.0% of PE, PAs and PAf, respectively.

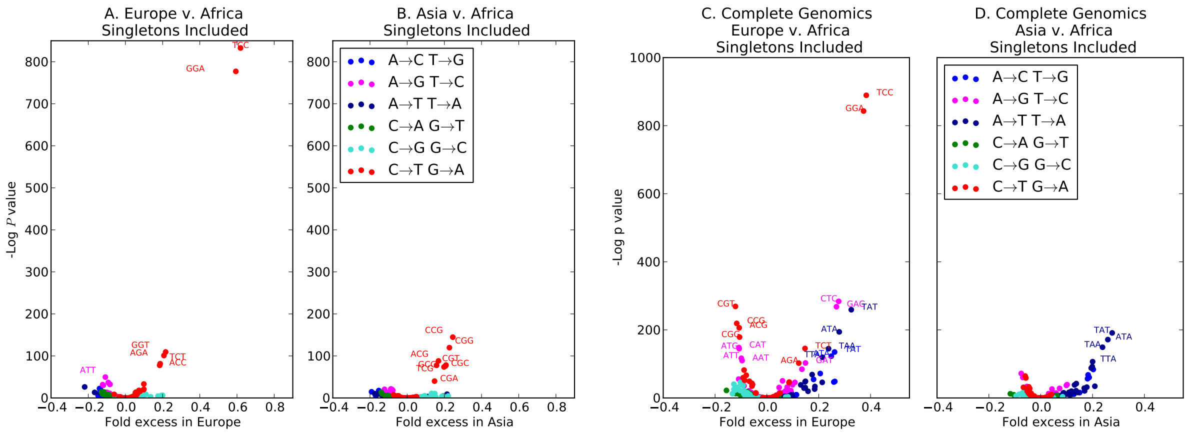
Figure S10 shows volcano plots similar to Figure1A,B that describe frequency differences between populations for each depth category separately. The three high-depth and medium-depth categories closely resemble Figure1A,B, with little differentiation aside from a few C→T transitions showing greater abundance in Europe than Africa, most notably TCC→TTC. Although the lowest-depth bin 4 also shows an excess of TCC→TTC in Europe, this signal is somewhat obscured by a large number of differences between populations that are not reproducible in higher-depth data. These results suggest that C→T transitions are truly more frequent in Europe compared to Asia and Africa, but that care must be taken to avoid calling additional spurious frequency differences when coverage is lower than 4x per genome.

8 Singleton variants in 1000 Genomes and Complete Genomics

Singleton variants that occur within only a single genome were excluded from the analyses in this paper because of concerns about their quality. Such concerns appear significant given the alterations to Figure 1A,B and Figure S1A,B that are produced by including singletons in the analysis. The volcano plots in Figure S11A,B, which compare the 1000 Genomes populations with singletons included, show many apparent frequency differences between populations that do not show up when singletons are excluded, particularly in the Asia versus Africa comparison. The plots in Figure S11C,D show that the same is true for the Complete Genomics data, despite its higher coverage. Figure S11A,B is qualitatively very different from Figure S11C,D, suggesting that the Illumina/SOLID 1000 Genomes pipeline and the Complete Genomics pipeline have very different error patterns with regard to calling singletons.



Supplementary Figure S10: These four volcano plots show mutation spectrum differences between populations within four different categories of sequencing coverage. They show that the TCC→TTC excess in Europe is strongest among high-depth SNPs, and that additional spurious frequency differences show up when the mean coverage in any population is lower than 4x per individual.



Supplementary Figure S11: Panels A and B display frequency differences between continental groups in the 1000 Genomes Phase I data. These figures were produced in the same way as Figure 1A,B of the main text except that singletons (variants of minor allele count 1) were included. Similarly, Panels C and D show differences between the same groups in the Complete Genomics data. These panels were produced in the same way as Supplementary Figure S1A,B except that singletons were included. In both cases, singletons show extensive frequency differentiation that is not reproducible in non-singletons.

9 Comparison to *de novo* mutations

Mutation rates can be can be estimated in a seemingly straightforward way by sequencing parents and offspring to identify *de novo* mutations. If the TCC→TTC mutation rate did recently increase in Europeans, it should be possible to observe that TCC→TTC makes up a higher percentage of *de novo* mutations in European trios than in non-European trios. However, identifying *de novo* mutations is a formidable technical challenge because next generation sequencing error rates are on par with mutation rates; new mutations should almost always generate singleton variants, which appear to be much more error prone than non-singletons (as seen in Section S8). This raises some concern about the ability of current technologies to accurately estimate the total germline mutation rate [11], much less to discern subtle differences between context-dependent mutation rates in different populations.

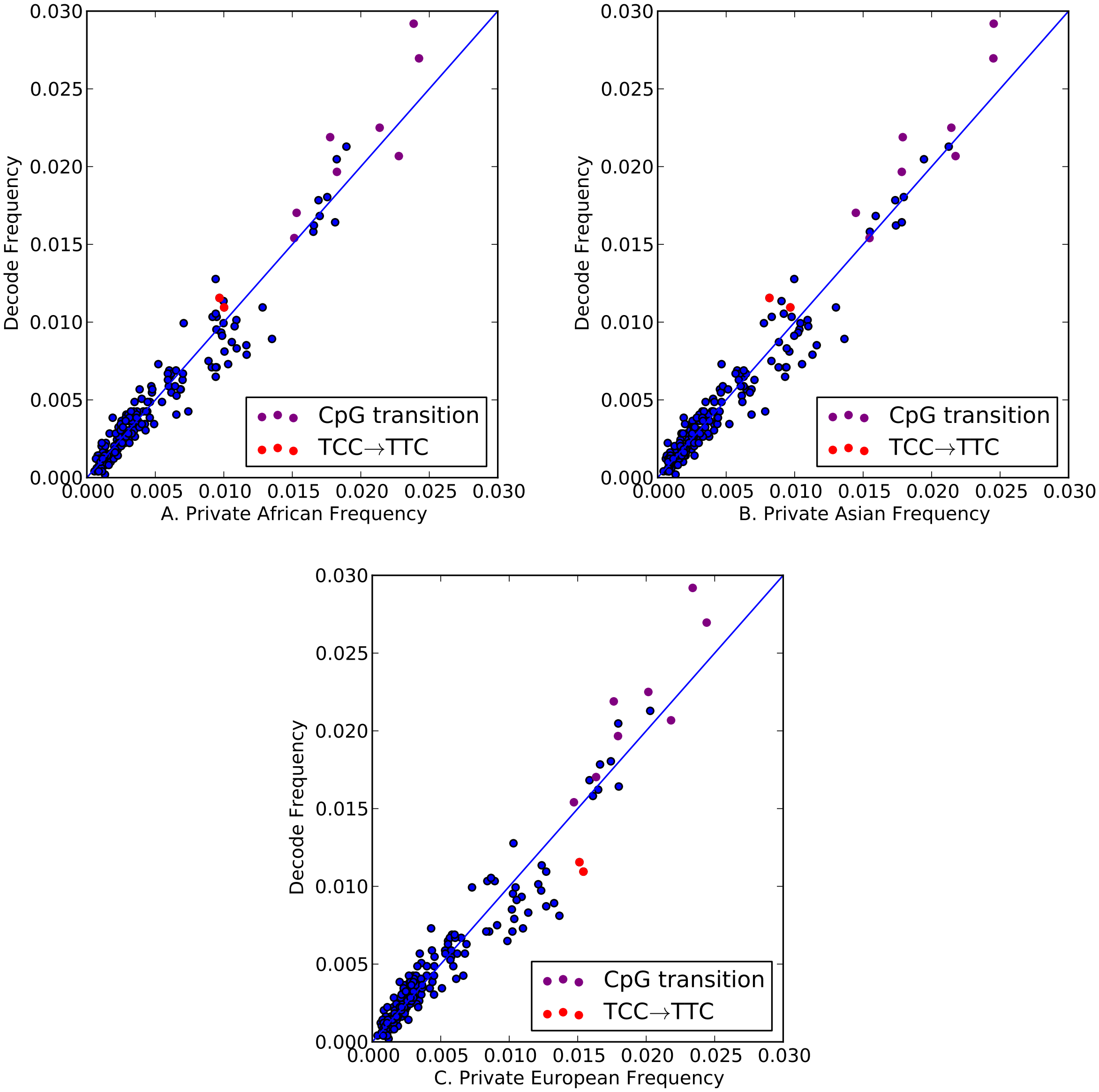
After filtering to remove putative sequencing errors and somatic mutations, human trios are each thought to contain about 40–80 *de novo* germline mutations [12, 11]. This implies that tens of whole-genome trio sequences are needed to distinguish a non-European TCC→TTC rate of 2% from a higher European rate of 3%. Segurél, *et al.* recently complied a comprehensive list of papers that estimate the human mutation rate by counting *de novo* mutations in trios [11], six by sequencing whole exomes [13, 14, 15, 16, 17, 18] and six others by sequencing whole genomes [19, 20, 21, 12, 22, 23]. Of these twelve papers, none sequenced a cohort of non-European genomes that was large enough for estimating the *de novo* TCC→TTC rate. Fromer, *et al.* identified their cohort as Bulgarian [18], and none of the other papers that studied disease cohorts provided ethnicity information for their either their cases or controls [13, 14, 15, 16, 17, 22, 23]. Conrad, *et al.* utilized one European trio and one Yoruban trio, an insufficient sample size for comparing mutation rates between populations [20]. The remaining studies by Roach, *et al.*, Campbell, *et al.*, and Kong, *et al.* identified their samples as European in origin [19, 21, 12]. The Genome of the Netherlands trios, published later than the Segurél, *et al.* review, also represent a European cohort [24]. The Complete Genomics diversity panel included a Yoruban trio and a Puerto Rican trio [4], but these sample sizes are also not sufficient for estimating Yoruban or Puerto Rican *de novo* TCC→T rates.

Although there was insufficient published data to estimate a *de novo* TCC→TTC rate from non-European trios, it was possible to estimate a context-dependent mutation rate spectrum from 78 Icelandic trios sequenced by DECODE Genetics [12]. Of the 4,932 germline mutations calls in this dataset, 54 (1.09%) were TCC→TTC and 57 (1.16%) were GGA→GAA. These values are intermediate between Asian SNP frequencies (0.968% and 0.816%, respectively) and European SNP frequencies (1.54% and 1.51%, respectively) (Figure S12). However, the DECODE *de novo* mutations are more similar to the private European mutations in terms of mutation type rank order. In the DECODE data, TCC→TCC and GGA→GAA are the 20th and 18th most abundant SNP types, respectively. They are similarly ranked 16th and 17th among private European SNPs, but are only the 30th and 42nd most abundant private Asian SNPs and the 27th and 32nd most abundant private African SNPs (Figure S13). This suggests that, in the Icelandic population, the TCC→TTC *de novo* mutation rate is better predicted by the private European spectrum than the private Asian or African spectra.

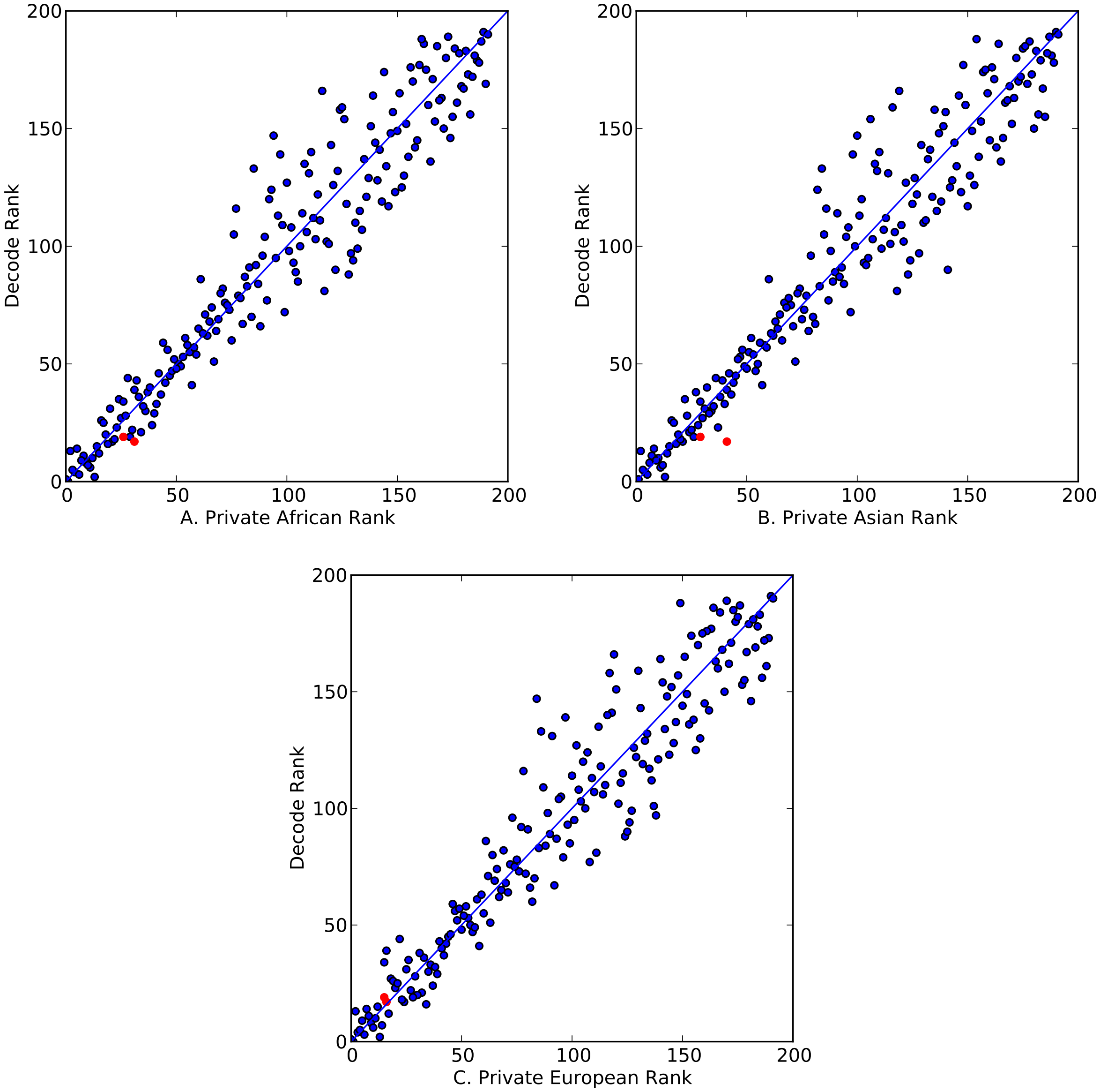
One fact that is apparent from Figure S12 is that CpG transition frequencies appear systematically higher in the DECODE data than either the private European or private Asian SNP sets. This discrepancy might indicate that the DECODE pipeline has either higher recall or a higher false positive rate at CpG sites than non-CpG sites. Since calling *de novo* mutations presents a greater technical challenge than calling SNPs, bioinformatic errors might be confounding this attempt to compare TCC→TTC frequencies across the DECODE and 1000 Genomes datasets.

10 Imputation accuracy of TCC→TTC mutations

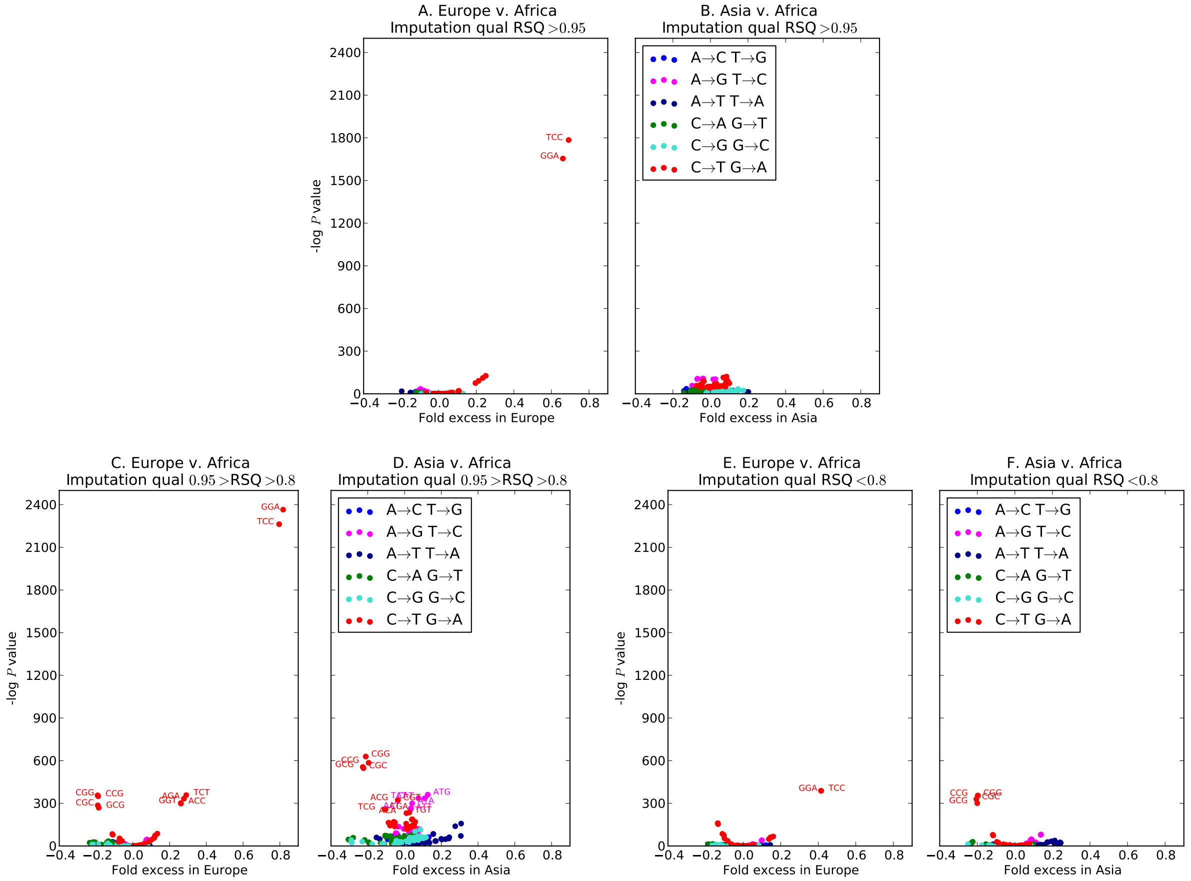
Each genotype call in the 1000 Genomes Phase I data is associated with an RSQ quality score, the estimated correlation coefficient between true and imputed genotypes at a given locus [25]. To assess the effect of imputation error on mutation frequency differences between 1000 Genomes populations, I repeated the volcano plot analysis from Figure 1 on medium imputation quality SNPs (RSQ between 0.8 and 0.95) as well as low imputation quality SNPs (RSQ less than 0.8). As shown in Figure S14, excess TCC→TTC mutations in Europe are evident across all imputation quality ranges. However, the medium-quality and low-quality volcano plots show many more minor outliers, mutation type frequency differences between Europe and Africa or Asia and Africa that are not reproducible in the high-quality 1000 Genomes SNPs or Complete Genomics data. These minor outliers might be indicative of real mutation rate change that occurred very recently and affects only alleles of very low frequency, too low-frequency to appear in the Complete Genomics dataset or to have high average imputation quality. However, these outliers might also be bioinformatic artifacts, particularly the CpG outliers that appear susceptible to ancestral misidentification errors.



Supplementary Figure S12: Each of the points plotted in panel A represents a context-dependent mutation type m. The point's y coordinate is the frequency of m among DECODE *de novo* mutation calls, whereas its x coordinate is the frequency of m in the private African SNP set PAf. All points lie fairly close to the line $x = y$, but there is still substantial variation between DECODE and PAf frequencies, for example, CpG mutations (plotted in purple) being more frequent in the DECODE data. TCC→TTC and its complement GGA→GAA, both plotted in red, are more frequent in the DECODE data than in PAf, but this difference is no greater than the general extent of the variability between the two datasets. The same can be said of panel B, where DECODE frequencies are plotted against private Asian frequencies, and also of panel C, where DECODE frequencies are plotted against private European frequencies. Together, these panels suggest that context-dependent frequency distributions of *de novo* mutations are not yet accurate enough for detection of mutation frequency differences between populations.



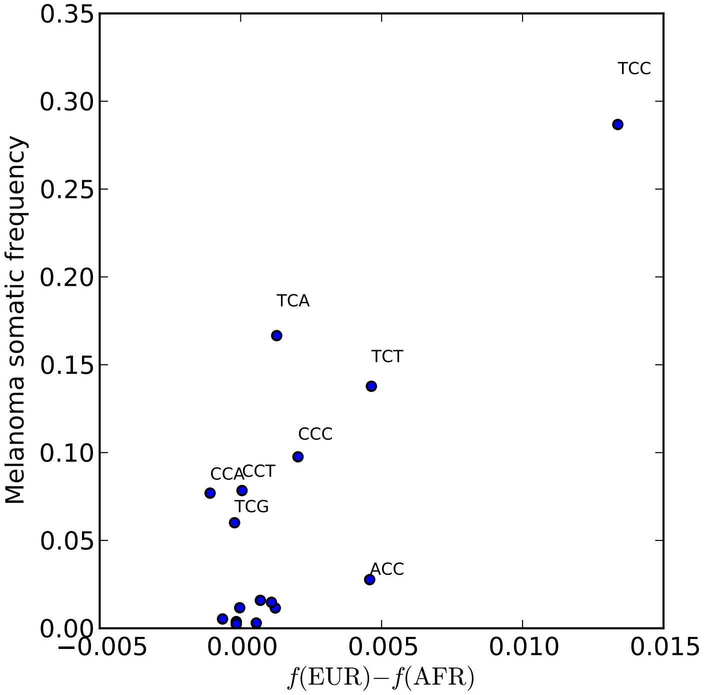
Supplementary Figure S13: Within each of the four datasets PAf, PAs, PE, and DECODE, each context-dependent mutation type was assigned a rank between 1 and 192 such that the most abundant mutation is ranked 1, the second most abundant mutation is ranked 2, and so on. In panel A, each point represents a mutation type m. The x and y coordinates of the point are the ranks of m within PAf and DECODE, respectively. As in Figure S13, the DECODE rank ordering is well correlated but not perfectly correlated with the rank orderings of PAf, PAs, and PE. Again, the rank variation of TCC→TTC and its complement GGA→GAA (shown in red) lie within the general rank variation of each of the three datasets. However, the PE rank ordering of TCC→TTC and GGA→GAA clearly fits the DECODE rank ordering better than the PAf and PAs rank orderings do, which makes sense given that the DECODE individuals are of European (Icelandic) descent.



Supplementary Figure S14: Panels A and B show differently scaled versions of Figure 1A and 1B from the main text, illustrating that high imputation quality SNPs (RSQ $>$ 0.95) support little frequency differentiation other than a C→T transition excess in Europe compared to Africa. In contrast, Panels C and D show volcano plots generated in the same way except that medium-imputation-quality SNPs (RSQ between 0.8 and 0.95) were used instead of high-quality SNPs. These medium-quality SNPs reveal an even clearer excess of various C→T transitions in Europe, but they also suggest evidence of other frequency differences between Europe and Africa and between Asia and Africa that are not reproducible in higher-quality data. Finally, Panels E and F reproduce the same analysis using only low-imputation-quality SNPs (RSQ $<$ 0.8). Again, the low-quality SNPs show evidence of differences that are not reproducible in the highest-quality data. Further work will be required to assess whether these differences are real or artifactual.

11 Comparison to somatic mutations in melanoma

In 2013, Alexandrov, *et al.* introduced the concept of cancer mutational signatures: collections of mutation types that are each characteristic of one or more cancer types and sometimes associated with exposure to a known carcinogen [26]. They discovered that melanoma has a unique mutational signature composed almost entirely of C→T transitions, almost 30% of which are TCC→TTC mutations. I downloaded the mutational spectrum of a melanoma skin cancer described in [27] to make a more detailed comparison between its mutational signature and the differences between PE and PAf. As shown in the scatterplot below, CCC→CTC mutations and TCT→TTT mutations are candidates for mutation rate acceleration in Europe that also contribute substantially to both the spectrum of melanoma (Figure S15). To a lesser extent, the same is true of ACC→ATC and TCA→TTA. However, the correlation between melanoma and European mutation rate change is far from perfect overall.



Supplementary Figure S15: Each dot in this scatterplot represents a different type of C→T transition, merged with its G→A strand complement for simplicity. The y coordinate of the dot representing mutation type m is the frequency of m in melanoma, while the x coordinate is the difference between the frequency of m in PE and the frequency of m in PAf.

References

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

[2] Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).

[3] Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225–228 (2014).

[4] Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).

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

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

[7] Hernandez, R., Williamson, S. & Bustamante, C. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol* **24**, 1792–1800 (2007).

[8] 1000 Genomes Project. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

[9] Benjamini, Y. & Speed, T. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucl Acids Res* **10**, 1–14 (2012).

[10] Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2078–2079 (2009).

[11] 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).

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

[13] Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).

[14] Neale, B. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).

[15] O'Roak, B. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).

[16] Sanders, S. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 237–241 (2012).

[17] Zaidi, S., Choi, M., Wakimoto, H., Ma, L. & Jiang, J. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–223 (2013).

[18] Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).

[19] Roach, J. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).

[20] Conrad, D. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genetics* **43**, 712–715 (2011).

[21] Campbell, C. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* **44**, 1277–1281 (2012).

[22] Michaelson, J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431–42 (2012).

[23] Jiang, Y. *et al.* Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am J Hum Genet* **93**, 249–263 (2013).

[24] Francioli, L. *et al.* Whole-genome sequence variation, population structure and demographic histroy of the Dutch population. *Nat Genet* **46**, 818–825 (2014).

[25] Li, Y., Willer, C., Ding, J., Scheet, P. & Abecasis, G. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidem* **34**, 816–834 (2010).

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

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