**AASignals of variation in human mutation rate at multiple levels of sequence context**

Rachael Aikens1,2,3 and Benjamin Voight2,3,4

1Department of Mathematics, Swarthmore College, Swarthmore, PA 19081

2Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, PA 19104

3Department of Genetics, Perelman School of Medicine, University of Pennsylvania, PA 19104

4Institute for Translational Medicine and Therapeutics, University of Pennsylvania, PA 19104

Correspondence to:

Benjamin F. Voight, PhD

Associate Professor of Systems Pharmacology and Translational Therapeutics

Associate Professor of Genetics

University of Pennsylvania

Perelman School of Medicine

3400 Civic Center Boulevard

10-126 Smilow Center for Translational Research

Philadelphia, PA 19104

Email: [bvoight@upenn.edu](mailto:bvoight@upenn.edu)

**ABSTRACT**

Our understanding of mutation rate helps us build evolutionary models and make sense of genetic variation. However, recent work indicates that the frequencies of certain C→T mutations have been elevated in Europe, and that many more, subtler signatures of global polymorphism variation may yet remain unidentified. Here, we present an analysis of the 1,000 Genomes Project (phase 3), suggesting additional putative signatures of mutation rate variation and the extent to which they are shaped by local sequence context. First, we compiled a list of the most significantly variable polymorphism types in a cross-continental statistical test. Clustering polymorphisms together, we note multiple sets of substitution types that seem to correspond in relative rate across ancestral populations, and describe the patterns of these mutational clusters among continental groups. For the majority of these signatures, we found that a single flanking base pair of sequence context was sufficient to determine the enrichment or depletion of a mutation type. However, we find that local genetic context up to 2-3 base pairs away may shape a previously noted enrichment of certain polymorphism types in certain East Asians. Finally, we developed a statistical model for genome-wide polymorphism which captures population-specific mutation rate variation, and found that this model explains observed polymorphism data much more successfully than one which assumes no variation. Building our understanding of mutation rate in this way can help us to construct more accurate evolutionary models and better understand the mechanisms which underlie genetic change.

**Words: 242**

**INTRODUCTION**

The process of mutation is a formative force in molecular evolution because it generates the genetic variation that can be acted upon by natural selection. Quantitative insights into human mutation rate can help us construct more precise evolutionary models of past and present populations, while more qualitative observations may point us towards the genetic or environmental mechanisms which confer genomic stability and drive genetic change. Unfortunately, the factors affecting mutation rate across the genome are context-specific and by no means perfectly understood. In fact, rates of DNA mutation and repair have been known to differ widely between certain individuals1, within and between chromosomes, and down to specific local sequences2.

For example, when inferring the timing of evolutionary events, geneticists often rely on a “molecular clock” assumption, positing that genetic variants accumulate at a constant rate on a species-wide scale. However, recent findings3–5 have indicated that patterns of mutation vary across human populations, suggesting that mutation rate itself may be in flux. In fact, polymorphisms within certain local sequence contexts appear to be significantly enriched or depleted in certain continental groups5. Most notable among them is a group of C→T polymorphisms strongly enriched in Europe and South Asia3–5. This evidence of population-scale variation in mutation rate presents us with a call to retune our evolutionary models, but also may offer us an opportunity to identify novel factors shaping fundamental mutational processes.

While the search for pairwise enrichment or depletion of polymorphism types turns up a myriad of nominally significant results5, the difficult task remains to identify pertinent signals of mutation rate variation and the biological forces that drive them. To this end, it will be important to determine which trinucleotide polymorphism types are most strongly significantly variable at a global scale, and how those substitutions are distributed across continents and subpopulations. Additionally, while multiple studies have described a group of C→T mutation types which vary in concert across Europe and South Asia3–5, few other distinct groups of polymorphisms have yet been characterized4,5. If clusters of polymorphisms that behave together are driven by a shared mechanism, identifying such putative mutational “signatures” first step to uncover the genetic, or environmental forces driving these changes.

Additional insight may be gained from a consideration of broader sequence context. Previously published clinical and evolutionary studies of mutation rate have made use of a trinucleotide, or ‘3-mer’ sequence context model, which considers the flanking 5’ and 3’ bases of a given substitution6,7 (e.g. 5’ TCC 3’ → 5’ TTC ‘3, denoted hereafter as TCC→T). However, genetic context up to 3 base pairs away from the locus of a substitution can substantially shape the probability of a mutational event (a ‘7-mer’ level of sequence context)8. Moreover, it has been suggested that differences in polymorphism variety we observe between populations may be driven in part by variability in DNA repair or replication machinery5,9. Since this machinery appears to respond to a variety of cues from local genetic sequence8,10, we hypothesized that the factors driving mutation rate variation at the population-wide scale act simultaneously across many levels of sequence context. If this were the case, considering greater numbers of upstream and downstream base pairs of context would highlight new features of mutation rate variation. In some cases, a 3-mer sequence context view may capture most of the detail in a signal of mutation rate heterogeneity. In others, strong effects stemming from broader sequence context may drive signals of polymorphism variation at a 3-mer level, indicating that the responsible biological mechanism may rely on cues or motifs several base pairs away from the substitution. In these cases, models considering broader windows of local context may highlight subtle variation in polymorphism which could not have otherwise been detected.

For these reasons, we sought to expand upon previous studies at the 3-mer level by identifying highly significantly variable polymorphisms at multiple context levels and describing how they vary across populations. To this end, we have applied a combination of 3-mer, 5-mer, and 7-mer sequence context frameworks to analyze the current release of the 1,000 Genomes project, spanning >2,000 subjects across four continents. Armed with this information, we can begin to ask questions about the driving mechanisms of the patterns we observe and build more precise models for explaining human genetic variation.

**RESULTS**

In order to understand how mutational processes may have diverged over evolutionary time, we assembled sets of genetic variants specific to non-admixed Africans, Europeans, South Asians, and East Asians (783, 669, 661, and 617 individuals, respectively) from the 1000 genomes phase III dataset11. Since genetic variants in the coding genome are likely to be under selection, we included only variants found in the non-coding genome. Our final sets consisted of 6,290,750 private African variants, 1,062,486 private European, and 1,595,092 and 1,703548 private East Asian and South Asian variants, respectively.

**Identifying novel significantly variable 3-mers**

We first sought to compile a list of polymorphisms which show significant variation at a global scale. To this end, we performed a three degree of freedom chi-squared test for homogeneity in private polymorphism across Europe, Africa, East Asia, and South Asia. This method builds from the one degree of freedom chi squared test employed by Harris and Pritchard3,5 to compare counts of polymorphisms between pairs of continental groups. However, in addition to reducing the required number of hypothesis tests (important later for analyses with broader windows of sequence context), this statistical framework allowed us to rank order polymorphism types by the significance of their variation across multiple continental groups. After replicating previous results3,5 as a technical control (supplement), we performed this expanded chi-squared test for each 3-mer polymorphism type (96 total), applying a modified pordered p value correction put forth by Harris and Pritchard5, and conservative Bonferroni correction for multiple hypotheses (significance threshold 5×10-4).

As expected, the foremost group of significant results was composed of C→T polymorphism types previously reported to be strongly enriched in Europe. All four types which have already been noted as part of this signal, TCC→T, ACC→T, TCT→T, and CCC→T, were among the 6 most significantly variable polymorphisms (Table 1). Also of note, all four possible types of CpG transition mutations rank among the top 11 most significant results (Table 1). Mathieson and Reich4 have previously noted that CpG mutations show some variability between populations, however these differences are not very large compared to their overall mutation rate. In agreement with this, we find that the CpGs all appear to have a shared profile of enrichment in South and East Asia, but that this enrichment is not as heavily pronounced. Importantly, Mathieson and Reich have cautioned that an apparent CpG enrichment may in fact be a signature of recurrent mutation in populations which have experienced recent exponential growth. However, if this were the case, we would expect to see a strong peak in CpG polymorphisms at allele count 2, which we do not observe (supplement).

Encouraged by the replication of these expected results, we examined the remaining significantly variable polymorphisms for novel signatures of mutation rate variation. In total, we observed that 62 of the 96 possible 3-mer types were significantly heterogeneous across ancestral continental groups. However, in order to remain strongly conservative, we first chose to consider only the 14 most highly significantly variable polymorphisms (p < 10-40, shown in table 1). To more easily make comparisons of relative enrichment between continental groups, we used our private polymorphism data to infer mutation rates for each 3-mer (per generation per site) using the methodology from Kong et al12 (see “Mutation Rate Inference” in the Methods Section).

In addition to the C→T polymorphisms already mentioned, we noted two additional substitutions among this group, TCA→T and ACT→T, which were both significantly variable at p < 10-40 and which showed enrichment in Europe and South Asia. A final polymorphism, GCC→T, was likewise enriched in Europe and South Asia, but also showed an elevation in East Asia. One plausible explanation is that GCC→T is elevated in Europe and South Asia by the same mechanism as the other C→T polymorphisms listed here, but that it’s enrichment in East Asia is driven by some other biological force or artifact.

Among the remaining polymorphisms, GAT→T and ACC→A, were both significant at p < 1×10-85, and have not yet been specifically noted in previous studies (Table 1). Interestingly, we found that both GAT→T and ACC→A display a shared profile of enrichment in East and South Asia over Africa and Europe. In addition, by estimating mutation rate separately on each chromosome, we found that this pattern is relatively consistent across the genome (supplement). These findings suggest that GAT→T and ACC→A may represent a novel signal of polymorphism types enriched in Asia. The sole remaining polymorphism type, ACA→T, is enriched in Africans and depleted in East Asians and Europeans (Table 1, supplement). In addition, we again find that this pattern is consistent across chromosomes (supplement). Taken together, these results indicate that there may be serval strongly significant signatures of variable mutation rate operating at the 3-mer scale, in addition to the European C→T elevation.

**Hierarchical clustering of 3-mer mutational signatures**

Since there appear to be polymorphism types at the 3-mer level which vary significantly in a variety of ways, we hypothesized that polymorphism types with similar global profiles of enrichment and depletion may have a mechanism in common. This would motivate a need to catalogue “signatures” of mutational variation: sets of polymorphisms which show similar profiles of relative mutation rate across the globe. To this end, we hierarchically clustered 3-mer polymorphism types based upon their relative rates in each of the twenty 1,000 genomes subpopulations (excluding admixed populations) from our first analysis.

We highlight five clusters with consistent patterns in the profiles of substitution rates, enumerated in figure 1A. Signature 1 corresponds to European C→T enrichment, and includes all four 3-mers previously reported as part of this signature (figure 1B). The remaining two polymorphisms in this group, TCA→T and ACT→T are noted in the previous section, (Table 1), and represent significantly variable polymorphisms not previously highlighted as part of this group of Europe-enriched mutations.

The next signature (#2) consists of GAT→T, ACC→A, and GAC→T, which are elevated in East and South Asia (figure 1C). Two of these, GAT→T and ACC→A were noted above (Table 1) for their cross-continental heterogeneity (p<10-85) and similar profile. However, GAC→T is also heterogeneous across continents (16th most significant 3-mer polymorphism, p < 2×10-33). They are followed by signature 3, corresponding to the CpG transitions, which cluster together even after the data are normalized to show only relative mutation rates (figure 1D). Finally, we observed two clusters (4a and 4b) which appear enriched in Japan and other groups in East Asia. These are listed together because we find that 4a and 4b merge into a single grouping when we cluster polymorphism types based only on their profiles within East Asia (supplement). This unit is comprised of the \*AC→C polymorphisms, as well as ATA→A, corresponding with a previous report5 which documented that \*AC→C, and TAT→T mutation types separate East Asians in a principal component analysis. Together, polymorphism types 2 and 4 may represent two distinct signatures of enrichment for certain mutation types in Asia.

The final cluster (5) is made up of CpG transversions which appear to be elevated in Africa. However, Harris and Pritchard13 report that the proportions of these polymorphism types do not appear to agree between the Simons Diversity Genome Project and phase III 1,000 genomes dataset. This suggests that signature 5 may be the result of an experimental artifact, rather than a true divergence in mutation rate. In sum, the clusters identified here highlight sets of polymorphisms whose relative proportions tend to match across populations from 1,000 genomes, hinting at a shared underlying cause.

**Higher sequence contexts of 3-mer signatures**

Given that sequence context up to 3 flanking base pairs from a substitution can have a substantive effect on the probability of mutation8,10, we next sought to ask which (if any), of the 3-mer signals that have been identified may actually be driven by variation at the 5-mer or 7-mer sequence context level. To do this, we considered all 256 possible 7-mer expansions of each interesting 3-mer identified in the previous section, and plotted the relative inferred mutation rates of those polymorphisms in pairs of populations. If there were no signal of mutation rate difference between populations, we would expect all 7-mer expansions to be distributed along the diagonal y = x (e.g., Figure 2A). If the most important local features driving a mutational signal lay within a single nucleotide base of the substitution, then we would expect all 7-mers to lie together off the diagonal (e.g., Figure 2B). Alternatively, if a 3-mer signal were actually driven by a handful of highly variable 7-mer substitution types, only a handful of exceptional 7-mer types would lie far from the y=x line (case III).

Performing this experiment, we found that nearly all of the 3-mers comprising signals 1, 2, and 3 matched case II (Figure 2B, supplement). This suggested that the global variation in European C→T elevation, the CpG transitions, and the Asian GAT→T, ACC→A, and GAC→T elevation was primarily driven by sequence context features within one flanking nucleotide base. However, the polymorphisms comprising signature 4 had a profile which more closely matched case III, indicating that the Japanese enrichment of the \*AC→C and TAT→T substitutions may in fact be driven by a handful of 7-mer polymorphisms which are highly heterogeneous in East Asia (Figure 3B, supplement).

To explore this finding in more detail, we sought to identify the key 7-mer types underlying this 3-mer signature. To this end, we considered each of the 1280 possible 7-mer expansions \*AC →C and TAT→T 3-mer substitutions, testing for heterogeneity between Japanese from Tokyo (JPT, higher signature 4 polymorphism proportion) and Chinese Dai from Xishuangbana (CDX, lower signature 4 polymorphism proportion). We found nine 7mer substitution types that were elevated in JPT relative to CDX (FDR-adjusted P < 0.05, figure 3A). Seven of them have the shared motif \*\*\*ACAG→C, and all nine share a matching profile of enrichment in JPT and Han Chinese (CHB and CHS) compared to Vietnamese (KHV) and CDX (figure 3C, supplement). Curiously, we also observed that 4 of the 9 polymorphism types were enriched on the X chromosome in East Asia, relative to the autosomes (the remaining types had too few observed polymorphisms on the X chromosome to justify a statistical test).

**Variable polymorphism types among higher sequence contexts**

Motivated by this result and previous work8, we hypothesized that some signals of mutation rate variation might be active over only certain specific 5-mer or 7-mer polymorphism types. If this were true, considering a broader span of sequence context would highlight novel signals of mutation rate variation not evident from 3-mer level analyses. To this end, we applied the same homogeneity testing framework described above to each of the 1536 possible 5-mers and the 24,576 possible 7-mer polymorphism types.

Within a 5-mer sequence context, we found that 142 out of 1536 possible polymorphism types showed significant variation after Bonferroni correction (supplement). Of these, 39 represent expansions of Europe-elevated 3-mer polymorphisms which have been highlighted by previous analyses (for example, CTCCA→T, which is an expansion of the TCC→T 3-mer)3,4. An additional 48 represent expansions of 3mers from signatures 1-3 (figure 1A). However, the remaining 55 significantly variable 5-mer polymorphisms have not yet been highlighted by previous work, representing putative new variable substitution types which have not been underscored by 3-mer analyses.

Finally, we moved to our broadest, 7-mer sequence context model. Out of 20,076 possible 7-mer substitution types with sufficient data available for a statistical test, 115 vary significantly after Bonferroni correction (Figure 4A, supplement). This decrease in number of significant results from the 5-mer model is not surprising, since the increase in number of possible polymorphism types for the name number of private variants is expected to decrease statistical power for tests considering broader windows of sequence context.

Of these 115 significant contexts, 91 represent expansions of polymorphisms identified at the 3-mer level, while 24 have not been previously noted. The strongest effect at the 7-mer level was the CAAACCC→C substitution (homogeneity test p < 2×10-34, figure 4), one of the Japan-enriched 7-mers we identified above (Figure 3A). In total, 4 of the 24 previously unreported significant 7-mer polymorphisms were one of the seven Japan-enriched substitution types (figure 3C). Interestingly, the third most significant unreported 7-mer polymorphism, AAACAAA→A (p =1×10-18) has a similar profile within East Asia as the other signature 3 polymorphisms (supplement). In fact, we find that multiple 7-mer types with 3-mer subcontexts outside of \*AC →C or TAT→T are enriched in Han Chinese and Japanese. This suggests that there exist additional 7-mer mutations comprising this signature that have not yet been discovered.

Finally, two of the 7-mer polymorphism types with variable rates across populations are TTTAAAA→T and ATTAAAA→T (p < 4×10-18), both of which were enriched in Africa (figure 4B, supplement). These correspond to the 3-mer TAA→A, which is the 17th most significant polymorphism from our 3-mer-level heterogeneity analysis (p < 2×10-31). Examining the rates of the 7-mer expansions of TAA→A, we find that TTTAAAA→T and ATTAAAA→T are indeed outliers among other 7-mer expansions both in terms of their African enrichment and the overall number of mutations of those types (figure 4C). These results suggest that the heterogeneity we observe in proportions of TAA→A polymorphisms is in fact driven by an elevation of these two highly variable 7-mers in Africa.

**A Model for Polymorphism across Populations**

We next developed a statistical model for polymorphism on the 3-mer, 5-mer, and 7-mer levels, to capture the effects of putative population-specific mutation rate variation. To this end, we described polymorphism with a series of multinomial models capturing different levels of variability in polymorphism probability (see the Methods Section for details). If mutation rate has not changed over evolutionary time, it should be possible to predict contemporary (population-specific) polymorphism from ancestral (shared, or ‘cosmopolitan’) genetic variation patterns. Contrary to this null hypothesis, mutational models that captured population-specific mutation rate variation independent from ancestral polymorphism patterns fit the observed data substantially better than competing models (log likelihood ratio test, P < 1×10-100, Methods, Supplement). These findings hold true at the 3-mer, 5-mer, and 7-mer level, suggesting that private polymorphism is variable within and between mutation contexts at different levels of sequence context.

Then we did some other stuff to show that our model is interesting and good…

**DISCUSSION**

In this report, we describe a number of patterns of polymorphism variation at a global scale.

However, whether these patterns reflect a true difference in underlying mutational processes, and what those underlying causes might be, remains a difficult question. Even the most prominent signature, European C→T, is still poorly understood. Although it appears to correlate with mutational signatures linked to ultraviolet radiation or alkylating agents in one cancer study4,14, evidence supporting either of these causal mechanisms is limited4.

However, this is by no means the only signature of polymorphism variation that has been detected4,5. Although European C→T enrichment is by far the most prominent mark of variation on the global SNP spectrum, the multitude of significantly variable polymorphism types and the variety of patterns they follow at a global scale suggest that several different processes are at work in shaping the ratios of polymorphisms we observe. If this is correct, further scrutinizing these differences may present us with an opportunity to better understand the processes that shape genomic stability and genetic change. Moreover, quantifying and modeling polymorphism patterns as accurately as possible can help us fine-tune our predictions and interpretations of single nucleotide genetic variation, potentially advancing our understanding of evolution or genetic disease.

One tool that may help us in these efforts is the consideration of local genetic sequence. Other scientists have noticed that a multitude of different patterns in polymorphism can be observed between substitution types just within different 3-mer motifs5, illustrating the importance of a single flanking base pair of context in shaping substitution probability. In this report, we consider a broader window of local sequence information, noting that while certain signatures are fully shaped within a single flanking base pair or context, others appear to vary with sequence context up to 2-3 base pairs from the locus of substitution (figures 2, 3B, and 4C). This can give some suggestion about what local genetic information is important in shaping the patterns we observe. For example, we find that all but two of the nine 7mers significantly heterogeneous between Chinese Dai and Japanese in signature 4 contain the 7mer motif \*\*\*ACAG→C (Figure 3). In addition, we find that the apparent elevation of TAA→T 3-mer polymorphisms between Africa and Europe may in fact be driven by a strong enrichment of substitutions within WTTAAAA contexts (where ‘W’ represents a weak ‘A’ or ‘T’ base), which also appear more highly mutable than other TAA sites (Figure 4B and C).

A roadblock to approaches like these, however, is sample size. While broader sequence context models can capture more information, they can also require much more total genetic data to be sufficiently well-powered for certain statistical approaches. This is made especially difficult because asking comprehensive questions about global mutation rate patterns requires a large and ethnically diverse dataset of genetic variation, the likes of which are only just becoming available to us. Further organized sequencing of diverse sets of people may enable us to make and test new, more targeted hypotheses. In figure 3A, for example, we find putative evidence suggesting that East Asian heterogeneity in \*AC→C and TAT→T mutations may be strongest on the X chromosome (figure 3C). Given this observation, it may be informative to examine the dispersion of these polymorphisms across the X-chromosome, since a genetic variant responsible for an increase in mutation rate is likely to be found in a genetic context with high polymorphism9. Unfortunately, however, a problem of power quickly emerges, since the total number of polymorphisms of any 7-mer type we observe on the X chromosome is still relatively small. As a result, analyses regarding this signature may be difficult until a larger amount of East Asian genetic data is made available.

It is a further complication that signals of polymorphism enrichment from population-level data may reflect some contemporary and some ancestral mutation rate variation, so that the biological mechanisms driving these phenomena may not be active today. Measurements of these enrichment of these polymorphism types in ancient DNA and across allele frequency bins suggest that this signal may correspond to an ancestral increase in mutation rate of certain C→T mutations ~15,000 years ago which may have subsided ~2,000 years ago4,5. Further analyses which consider polymorphism enrichment by allele frequency may help us piece together the timescale over while mutation rate variations have acted5. Signatures of population-specific polymorphism enrichment among common variants may result from ancestral mutation rate variation, while signatures among lower frequency variants are more-likely the result of contemporary mutation rate changes. One limitation to this approach again, however, is the number of genomic samples required to make effective inference, especially when broader sequence context is involved.

This report is not an exhaustive enumeration of all polymorphism variation at the 3-mer, 5-mer or 7-mer sequence context levels. It is likely that further investigation will reveal details of mechanism, evolutionary timing, and genome-wide or subpopulation-level patterns in mutation rate variation. Herein, we detail evidence suggesting that mutation rate variation acts in a variety of ways across human populations based on local sequence context cues at varying distances from the mutated locus. While some of these signals manifest at the 3-mer level, consideration of a broader context brings new patterns of variation to light.

**METHODS**

**Compilation of private variant sets**

Variants from the phase III 1,000 genomes release (downloaded 02/26/2016)11 were filtered to include only single nucleotide polymorphisms (SNPs) with minor allele count 2 or greater. Although including singleton variants (those observed only once in the dataset) in theory would provide more information about recent *de novo* mutation rate, previous efforts to analyze human polymorphism variation with singletons have proven difficult to replicate3. All multiallelic SNPs, and any variants with a filter tag other than “pass” were excluded from our analyses. Based on the exclusion regions from Aggarwala et al.8, we additionally omitted variants in coding regions, centromeres, telomeres and additional sections of the genome predicted to have low accessibility.

From these filtered variant lists, we compiled lists of variants ‘private’ to each nonadmixed continental group from the dataset: Africans from Africa (AFR), Europeans (EUR), East Asians (EAS), and South Asians (SAS). In doing so, we considered a SNP “private to population X” if it is observed in X but not in each of the other three continental groups. For all analyses, Americans of African Ancestry in Southwest USA (ASW) and African Carribeans from Barbados (ACB) were considered to be admixed American populations rather than ancestral African groups.

For subpopulation-level analyses, we then sorted the private polymorphisms for each continental group into subpopulation lists. For example, a polymorphism which was private to AFR and observed in both Kenya and Gambia would be added to the subpopulation lists for both LWK (Luhya in Webuye, Kenya) and GWD (Gambians in Western Divisions in Gambia). American admixed variant lists were compiled from all SNPs which were present in an American subpopulation but not present in more than one ancestral continental group. All filtration steps were carried out using vcftools and the vcf-isec tool (v0.1.12b)15.

From each continental or subpopulation list, we tallied counts of private variants by 3-mer, 5-mer, and 7-mer sequence context. During all this process, each mutation class is ‘folded’ to include its reverse-complimentary equivalent (e.g. TCC → T and GGA → A are always considered together). Sample code for these analyses is available online ([github.com/raikens1/mutatation\_rate](https://github.com/raikens1/mutatation_rate)).

**Statistical comparison with homogeneity test**

To replicate previous work by Harris and Pritchard3,5, we first performed pairwise chi-squared comparisons of polymorphism count between each possible pair of populations for each 3-mer polymorphism type (supplement). Next, to partially relieve the multiple testing burden of 6 pairwise population comparisons over each possible mutation type, we combined these tests into a single 2 by 4 contingency table for a three degrees-of-freedom chi-squared test for homogeneity. One issue with calculating such a chi-squared test result for each possible type of polymorphism is that the p values from these tests are non-independent; in fact, a polymorphism which is strongly heterogeneous across populations may alter the proportions of other polymorphism types. For these reasons, we used the *pordered* correction in keeping with Harris and Pritchard5. Using this procedure, each polymorphism type is initially tested and ranked according to increasing significance based on a simple homogeneity test using all the data. A corrected p-value is then calculated for each polymorphism. To do this, the least significant polymorphism is assigned its original p-value using all of the data. After this, the p-value for the *i*th least significant polymorphism type is recalculated using a homogeneity test with only the data for the *i* least significantly variable polymorphisms from the initial ranking. All chi squared comparisons were done using the chisq.test function in R (v3.3.2), and significance thresholds were determined based on a conservative Bonferroni correction with a nominal alpha value of 0.05.

**Mutation rate inference**

The probability of observing a given polymorphism in a population is determined by a composite of mutation rate, demography, and sample size16. To facilitate comparisons across populations, we calculated inferred mutation rate based on methods from Kong et al12. Assuming all populations have a total mutation rate of 1.2×10-8, we inferred the mutation rate of a specific type (say TCC→T) as

Where represents the inferred private germline TCC→T mutation rate per generation per site, represents the proportion of all TCC sites in the genome with private C/T polymorphism in the population, and represents the total proportion of all sites of any type in the genome which are private polymorphisms in the population. It can be shown that this formulation of gives an overall genome wide mutation rate of 1.2×10-8 when all mutation types are pooled. 95% confidence intervals for were calculated using the normal approximation to the binomial, assuming the variance in to be approximately zero.

**Clustering polymorphism types**

We used the heatmaps 2 hierarchical clustering methods from the basic stats package in R (v3.4.0) in order to heuristically identify mutation types which vary in similar ways across the globe. In doing so, we defined the “profile” of a mutation *m* across a set of populations as a vector of the inferred mutation rate of *m* in each population. Each profile was normalized by fold difference above or below the mean rate for that profile and compared using Euclidean distance in order to construct each heatmap. These methods for normalization and distance metric were selected above others because they gave the most clearly interpretable results which seemed to be in agreement with previous work3–5 (supplement)

**Testing for enrichment of signal 4 on the X chromosome**

In order to test the highly variable polymorphism types in figure 3A for enrichment on the X chromosome, we used a one-sided binomial test to check whether the observed proportion of privately polymorphic sites on the X chromosome was greater than , the expected proportion under the null hypothesis. For demographic and sampling reasons, we expect to observe fewer polymorphic sites of any given type on the X chromosome than on the autosomes, even if the mutation rate of that polymorphism type is identical across chromosomes. Thus, to estimate we first needed to calculate the ratio, , of X-chromosome to autosomal substitution probability across all other 7mer types. We could then use this as a scaling factor, estimating as , where represents the maximum likelihood estimate for the substitution probability for that polymorphism across all autosomes.

**Developing a Multinomial Model for Polymorphism**

In order to develop a formal statistical framework for understanding global polymorphism, we designed a series of multinomial models after Aggarwala and Voight8, which capture different levels of mutation rate variation. First, we defined cosmopolitan SNPs to be those which are shared between two or more of the African, European, South Asian, and East Asian 1,000 genomes samples. For a given population at a given sequence context, the probability of recurrent mutation is assumed to be zero. Under these assumptions, we have seven mutually exclusive possible events: either the site is not polymorphic, it is a private polymorphism for that population (with three possible alternate alleles), or it is a cosmopolitan polymorphism (with three possible alternate alleles). If the context appears N times in the genome, polymorphism in this population follows a multinomial distribution with size N and parameters c1, c2, and c3 for the probabilities of each cosmopolitan polymorphism, and p1, p2, and p3 for the three private polymorphisms.

If the mutation rate at this context had not changed in recent evolutionary time for this population, then we would expect the probabilities of each private polymorphism type to be proportional to the probabilities of the corresponding cosmopolitan polymorphism types. It remains to estimate this proportionality constant, which we denote α. In a null model (H0), mutation rate has not changed at any context, so α for a given population is just the ratio of total private polymorphisms to total cosmopolitan polymorphisms over all contexts. Alternatively (H1), if mutation rate has changed at specific contexts but the relative substitution probabilities for the alternative alleles is fixed, then a unique α must be estimated from the private to cosmopolitan ratio of polymorphisms at each context. Finally, in a model which allows for maximal polymorphism variation (H2), mutation rate may have changed even between different polymorphism types at same context (e.g. C/T, C/A, and C/G polymorphism at a C context). In this model, the private substitution rates are not proportional to the cosmopolitan rates even at a context-specific level, and the private rates must be estimated independently for each possible mutation.

Once we have estimated the necessary parameters for each of these three models from the 1,000 genomes dataset, we can compare the fit of the observed data under each model (H0,H1, or H2) using a log-likelihood ratio test. If Λ represents the ratio of the likelihoods of a null model to an alternative, the test statistic is known to approximately follow a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters in the null versus the alternative. This testing framework allows us to ask broad questions about what level of mutation rate variability best explains observed polymorphism data at a given sequence context level.

**ACKNOWLEDGEMENTS**

RCA is thankful for the guidance and feedback of Elizabeth Vallen, Nick Kaplinsky, Bradley Davidson, and Ameet Soni. The authors would also like to thank Varun Aggarwala, who offered helpful advice on the implementation of 5-mer and 7-mer sequence context analyses. B.F.V. is grateful for support from the US National Institutes of Health/National Institute of Diabetes and Digestive and Kidney Disorders (R01DK101478).

**SUPPLEMENT**

* All code and raw count data available on github.
* Replication of KH 2015 Figure 1
* Fdr and p values for all significant results for 3-mer, 5-mer, 7-mer
* Ranked p value figures: 3-mer/5-mer/7-mer
* TCC->T by subpopulations to show correlation with latitude
* CI and chromosome plots for top 3-mer signals not already in Figures 1 and 2.

**LITERATURE CITED**

1. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43,** 712–4 (2011).

2. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12,** 756–66 (2011).

3. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 3439–44 (2015).

4. Mathieson, I. *et al.* Differences in the rare variant spectrum among human populations. *PLOS Genet.* **13,** e1006581 (2017).

5. Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *Elife* **6,** (2017).

6. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 13994–4001 (2004).

7. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).

8. Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* **advance on,** (2016).

9. Seoighe, C. *et al.* Inference of Candidate Germline Mutator Loci in Humans from Genome-Wide Haplotype Data. *PLOS Genet.* **13,** e1006549 (2017).

10. Zhu, Y., Neeman, T., Yap, V. B. & Huttley, G. A. Statistical Methods for Identifying Sequence Motifs Affecting Point Mutations. *Genetics* **205,** (2017).

11. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

12. Kong, A. *et al.* Rate of de novo mutations and the importance of father’s age to disease risk. *Nature* **488,** 471–5 (2012).

13. Harris, K. & Pritchard, J. Rapid evolution of the human mutation spectrum. *bioRxiv* (2017).

14. Forbes, S. A. *et al.* COSMIC: exploring the world’s knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805–D811 (2015).

15. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

16. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7,** 256–276 (1975).

**TABLE LEGENDS**

**Table 1:** **A cross-continental test for polymorphism variability at the 3-mer level.** 14 SNPs, shown here, were highly significant (p < 10-40) according to a chi-squared test for heterogeneity across non-admixed continental groups. The Bonferroni significance threshold for this test would be 5×10-4. Boldface numbers indicate a significant difference in polymorphism proportion compared with Africa (p < 10-5) in a pairwise chi squared test. All p values from these tests apply the pordered correction put forward by Harris and Pritchard5. \*To facilitate comparison, approximate private mutation rates (per generation per site) were inferred based on the procedure by Kong et al12 then normalized relative to inferred rate in Africa.

**TABLES**

**Table 1**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Notes | 3-mer | Relative rate in Africa\* | Relative rate in Europe\* | Relative rate in South Asia\* | Relative rate in East Asia\* | p |
| Previously reported C→T elevation in Europe3–5 | TCC→T | 1 | **1.55** | **1.21** | 1.00 | ≈ 0 |
| ACC→T | 1 | **1.21** | **1.08** | **0.93** | 6×10-263 |
| TCT→T | 1 | **1.16** | **1.07** | 0.99 | 1×10-155 |
| CCC→T | 1 | **1.06** | **1.03** | **0.95** | 6×10-55 |
| CpG transition | TCG→T | 1 | **1.02** | **1.06** | **1.03** | 2×10-54 |
| ACG→T | 1 | 1.01 | **1.05** | **1.04** | 6×10-54 |
| GCG→T | 1 | 1.02 | **1.05** | **1.05** | 3×10-49 |
| CCG→T | 1 | 1.00 | **1.04** | **1.03** | 3×10-44 |
| Not previously highlighted | GAT→T | 1 | **1.06** | **1.13** | **1.21** | 2×10-95 |
| ACC→A | 1 | **1.04** | **1.11** | **1.15** | 2×10-90 |
| ACA→T | 1 | **0.95** | 0.97 | **0.92** | 8×10-54 |
| TCA→T | 1 | **1.06** | **1.03** | 0.97 | 3×10-44 |
| ACT→T | 1 | 1.02 | 1.00 | **0.94** | 2×10-42 |
| GCC→T | 1 | **1.07** | **1.04** | **1.03** | 7×10-42 |

**FIGURE CAPTIONS**

**Figure 1: Putative signatures of mutation rate variability at the 3-mer level.** **(A)** Heat map of all 3-mer polymorphisms, clustered based on their relative rates in each of 20 nonadmixed 1,000 genomes populations. Clusters of interest are labeled, and their membership is detailed in the table to the right. Polymorphisms are clustered and colored based on fold elevation over the mean mutation rate for each mutation type. All units are log base 2 transformed, with red color corresponding to enrichment and blue to depletion. **(B-D)** Approximate 95% confidence interval estimates of inferred mutation rate across continental groups for signatures 1-3. **(E)** Inferred mutation rates for signature 4 shown across Europe, Africa, South Asia, and five East Asian subpopulations: Chinese Dai (CDX), Vietnamese (KHV), Han Chinese from Beijing and Southern China (CHB and CHS), and Japanese in Tokyo (JPT).

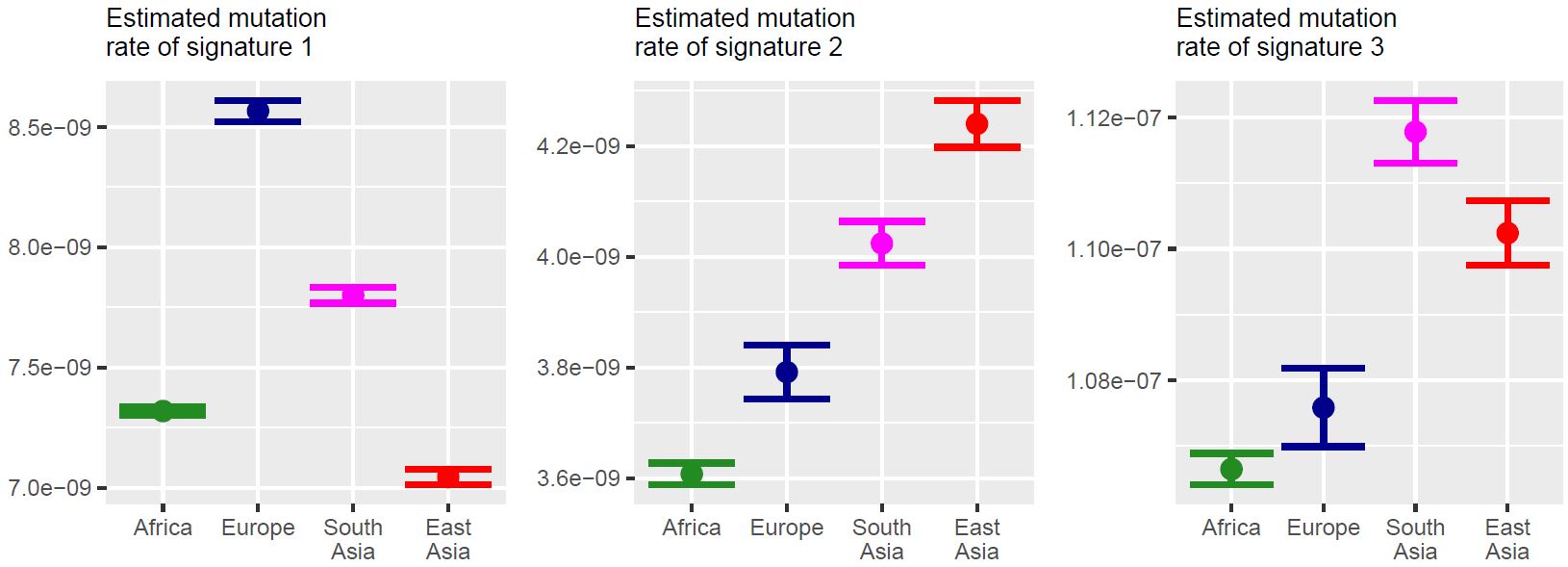
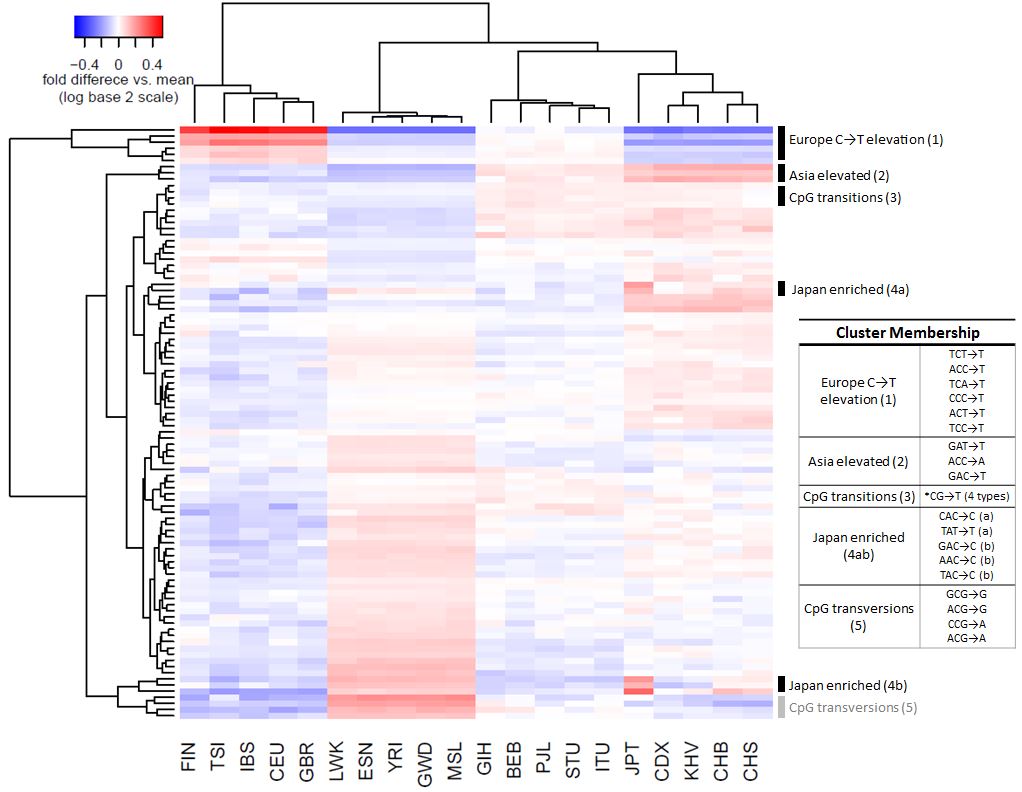
**Figure 2:** **Mutational signatures driven at the 3-mer level**. Each point represents a 7-mer expansion of the 3-mer subtype shown, plotted based on its estimated mutation rate in each of the two populations displayed. Colors indicate the log (base 10) of the number of substitutions observed for that 7-mer class. **(A)** When a 3-mer substitution type occurs at equal rates in two related populations, most of the 256 7-mer expansions of this 3-mer appear along the diagonal y=x line. **(B)** For TCC→T and the other C→T polymorphism types elevated in Europe, the bulk of the 7-mer expansions lie above the y=x diagonal, indicating that there has been a substantial difference in mutation rate between Europe and East Asia, and this difference is driven by effects at the 3-mer, rather than the 7-mer level.

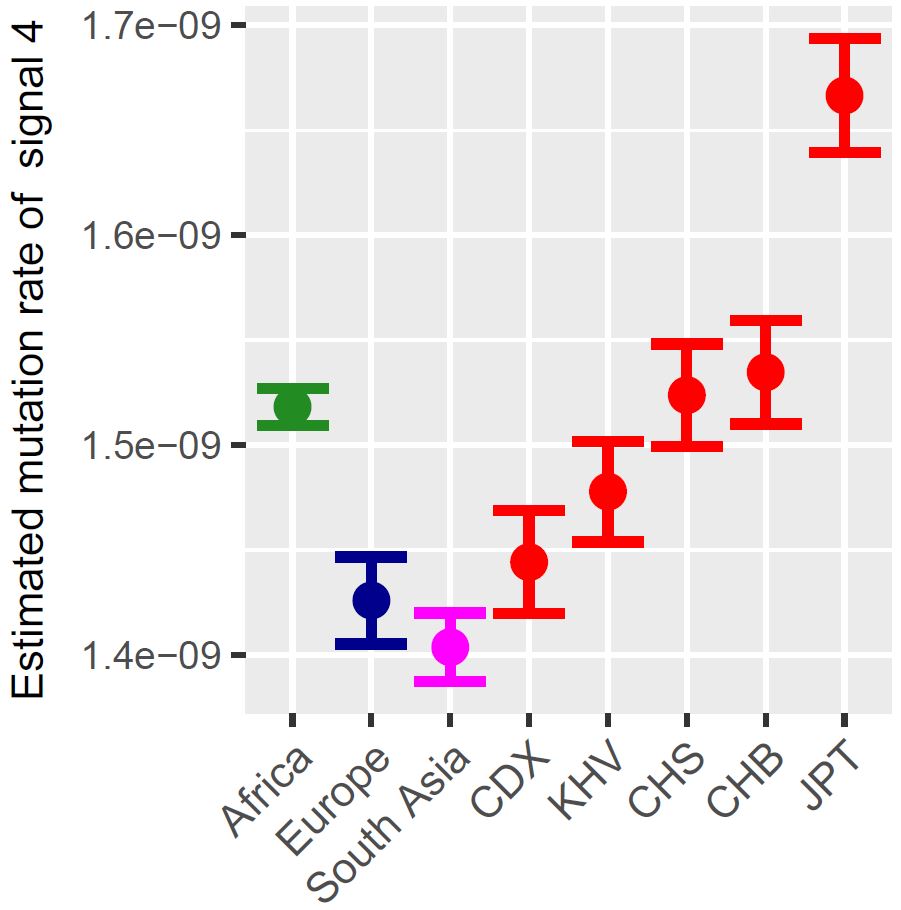
**Figure 3:** **Highly variable 7-mers within signature 4 in East Asia.** **(A)** Nine polymorphisms enriched in Japan compared to Chinese Dai (chi squared test, fdr < 0.05). \*Fold increase in inferred mutation rate in Japan compared to CDX. Bold p values indicate nominal significance of enrichment on the X chromosome in East Asia according to a one-sided binomial test (see methods).The significance values of tests for contexts with 5 or fewer observations on the X chromosome were not calculated. **(B)** Most 7-mer expansions of AAC→C are the same in Chinese Dai versus Japanese, with the exception of some highly variable 7-mer polymorphism types. Polymorphisms significantly heterogeneous between Japan and Chinse Dai are labeled. **(C)** Estimated private mutation rate of the nine 7-mer polymorphisms shown in part A displayed across each East Asian subpopulation. Brackets indicate approximate 95% confidence intervals.

**Figure 4:** **Signatures of mutation rate variation detected at 7-mer level.** **(A)** 5 most significantly heterogeneous polymorphism types at a 7-mer level, removing expansions of known 3-mer signals. Each p value was calculated from a chi-squared test for heterogeneity across non-admixed continental groups. The Bonferroni significance for this test would be 2.5×10-6. Boldface numbers indicate a significant difference in polymorphism proportion compared with Africa (p < 10-7) in a pairwise chi squared test with pordered correction. \*Approximate mutation rates (per generation per site), inferred assuming a total mutation rate of 1.2×10-8 mutations per base pair per generation. **(B)** Estimated mutation rate of TTTAAAA→T across continental groups, with approximate 95% confidence intervals shown. **(C)** TTTAAAA→T and ATTAAAA→T appear to be both more variable between continental groups and more common than other 7-mer expansions of TAA→T.

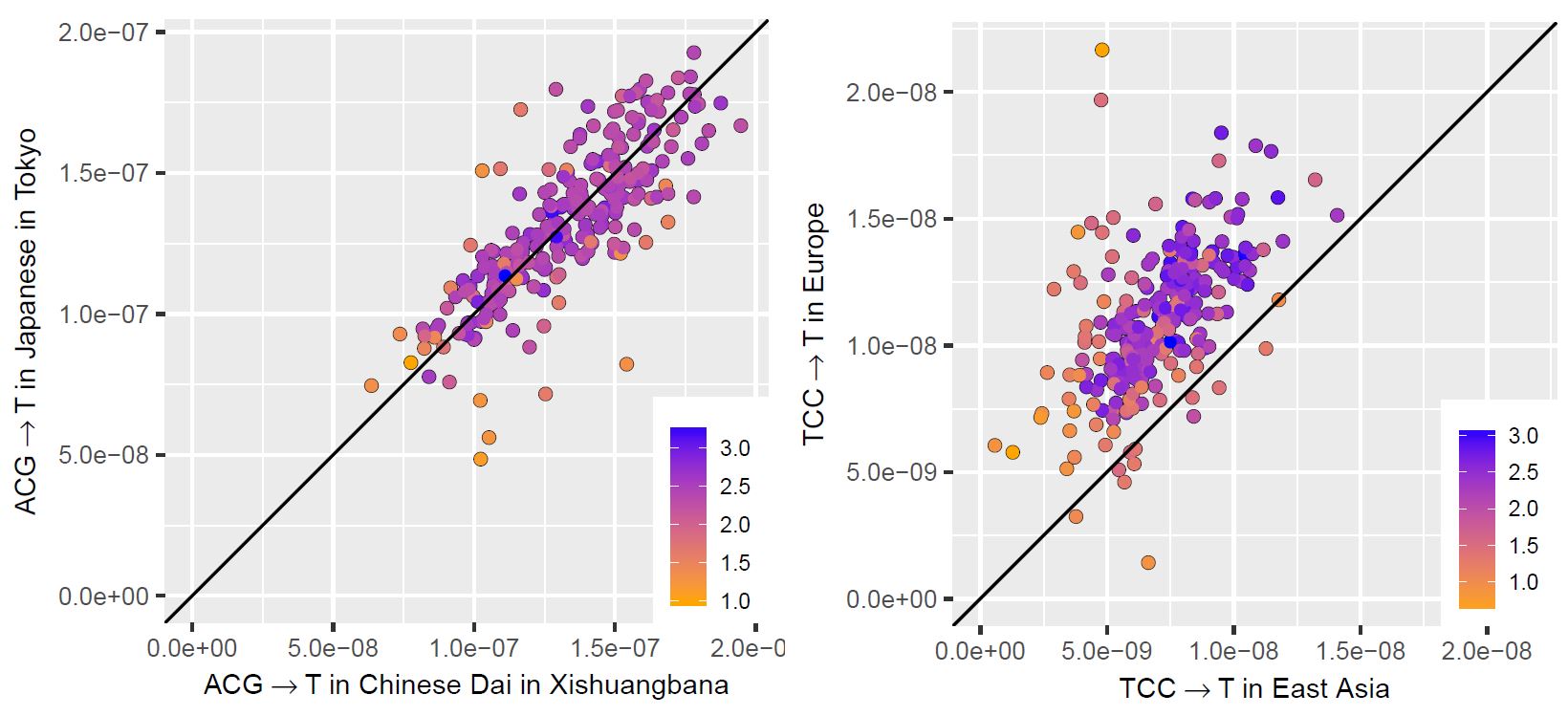
**FIGURES**

**Figure 1: ABCDE**

****

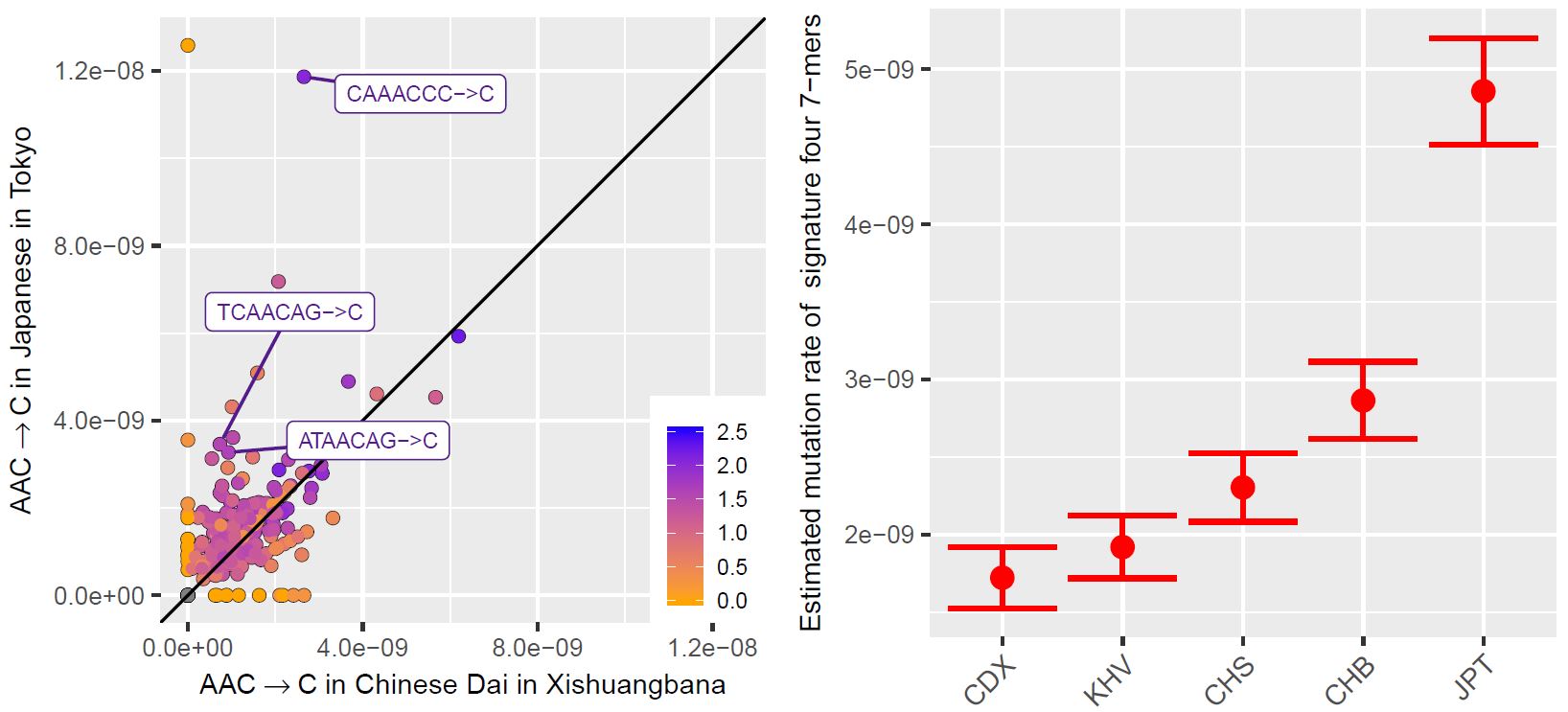


**Figure 2: AB**



**Figure 3: ABC**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 7-mer | Fold  enrichment  in Japan\* | FDR-adjusted p  (enrichment in Japan) | Expected  polymorphic  sites on X | Observed polymorphic sites on X | p  (X enrichment) |
| TTTATTT→T | 2.25 | 5.6×10-20 | 47 | 65 | **0.007** |
| CAAACCC→C | 4.47 | 1.0×10-8 | 6.2 | 27 | **5.3×10-10** |
| AGTACAG→C | 18.1 | 2×10-5 | 2.7 | 4 | - |
| CCCACAG→C | 2.62 | 0.01 | 4.4 | 25 | **1.2×10-11** |
| AATACAG→C | 5.15 | 0.01 | 2.6 | 1 | - |
| TCAACAG→C | 4.72 | 0.01 | 2.3 | 7 | **0.01** |
| ATGACAG→C | 4.57 | 0.01 | 2.3 | 4 | - |
| TCCACAG→C | 4.42 | 0.02 | 2.4 | 3 | - |
| ATAACAG→C | 3.52 | 0.03 | 3.4 | 4 | - |



**Figure 4: ABC**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 7-mer | Africa  rate\* | Europe rate\* | South Asia rate\* | East Asia rate\* | p |
| CAAACCC→C | 1 | 0.80 | 0.31 | **3.58** | 2×10-34 |
| TTTATTT→T | 1 | 0.82 | **0.61** | 1.03 | 5×10-22 |
| TTTAAAA→T | 1 | **0.87** | **0.84** | **0.86** | 2×10-20 |
| AAACAAA→A | 1 | 0.80 | **0.65** | 0.88 | 1×10-18 |
| ATTAAAA→T | 1 | **0.71** | **0.76** | 0.82 | 3×10-18 |

