**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 and qualitative insights regarding the mutation rate in human populations can facilitate the construction of increasingly informative models of human evolutionary history, targets of natural selection, and perhaps even genetic or environmental mechanisms that confer genomic stability and drive genetic change. Unfortunately, while the rates of DNA mutation and repair have been known to differ widely between certain individuals1, within and between chromosomes [cite], and down to specific local sequences2, the biological mechanisms underlying mutation rate variability across the genome are not yet completely known.

Recent work has suggested that the mutation rate in humans may itself have been in flux over recent evolutionary history3–5. Key evidence supporting this hypothesis stems from the observation that the relative proportions of certain types of polymorphisms vary across populations. Most notably, studies cite the strong enrichment of C→T substitutions at certain trinucleotide contexts in Europe and South Asia3–5. While some of these reports have also documented a number of additional polymorphism types that appear heterogeneous across populations[cite], the magnitude, extent, and distribution of this variation across continental and sub-continental groups have not been fully described. Moreover, since clusters of polymorphisms with similar global profiles of enrichment may be driven by a shared mechanism, is worthwhile to ask not only which polymorphism types vary across the globe but how variable polymorphism types group together as putative “signatures” of mutation rate variation. Developing a better understanding of these signatures of polymorphism variation is a necessary step to link such changes to a putative genetic or environmental cause.

Previous work also suggests that windows of sequence broader than the trinucleotide context may uncover additional detail in mutation rate variability across the genome [cite]. However, such approaches have not been explored in the context of population-specific mutation rate variability to date. Considering greater numbers of upstream and downstream base pairs of context could highlight new features of mutation rate variation. For example, a trinucleotide sequence context may fully capture signal of mutation rate heterogeneity. Alternatively, strong effects stemming from broader sequence context may drive signals of polymorphism, indicating that the underlying mechanism may rely on the local nucleotide configuration. As such, models that consider broader windows of local context may highlight subtle variation in polymorphism that might not have otherwise been detected. Addressing this question, however, requires mathematical models that incorporate both the effects of nucleotide context and population heterogeneity to determine what features ultimately best explain patterns of observed polymorphism data in human populations. Such models could then be applied to understand regions of the genome locally that may be subject to selective constraint quantified by the deficit of observed polymorphic sites relative to model predictions [cite].

For these reasons, we sought to expand upon previous studies at the trinucleotide 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 sequence context frameworks to analyze the current release of the 1,000 Genomes project, spanning >2,000 subjects across four continents. With this information in hand, we sought to catalog population-level heterogeneity in polymorphism across the spectrum of these contexts and use these insights to 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 (504, 503, 489, and 504 individuals, respectively) from the 1000 genomes phase III dataset6. Since genetic variants in the coding genome are likely to be under selection, we included only variants found in the non-coding genome (**Methods**). Our final sets consisted of 6,290,750 private African variants, 1,062,486 private European, and 1,595,092 and 1,703,548 private East Asian and South Asian variants, respectively.

**Identifying novel significantly variable 3-mers**

We first sought to compile a list of polymorphisms in trinucleotide (*i.e.*, ‘3-mer’) contexts that appear heterogeneous in their representation across the globe. To this end, we extended a previously described3,5 approach (**Methods**) to perform a test for homogeneity in private polymorphism across Africa, Europe, East Asia, and South Asia, to compare counts of polymorphisms between pairs of continental groups. In addition to reducing the required number of hypothesis tests compared to previous methods (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 (**Supplemental Note**), we performed our test for each 3-mer polymorphism type (96 total), applying a modified p-value correction (Pordered) as previously described5, and using Bonferroni correction for multiple hypothesis testing (nominal significance threshold Pordered < 5 x 10-4).

As expected, the most compelling group of contexts was composed of C→T polymorphism types previously reported to be enriched in Europe and South Asia (**Table 1**). All four types that have been previously noted as part of this signal - TCC→T, ACC→T, TCT→T, and CCC→T - were among the 6 most variable polymorphisms. We further observed that all four possible types of CpG transition mutations ranked among the top results (**Table 1**). Previous work has suggested that proportions of CpG substitutions are weakly variable between populations4, and we too note that the CpGs all appear to have a shared profile of enrichment in South and East Asia, but that this enrichment is smaller relative to the overall abundance of C/T polymorphism at CpG sites. Importantly, Mathieson and Reich have cautioned that an apparent CpG enrichment may be a signature of recurrent mutation in populations which have experienced recent exponential growth4. However, if this were the case, we would expect to see a strong excess of CpG polymorphisms at doubletons (i.e., allele count 2), which we did not observe (**Supplementary Note**).

We next examined the remaining variable polymorphisms for novel signatures of mutation rate variation. Surprisingly, we found that 62 of the 96 possible 3-mer types exceeded our Bonferroni correction for multiple tests across ancestral continental groups. Therefore, we opted first to consider the top 14 most heterogeneous polymorphism contexts (Pordered = 1 x 10-40, **Table 1**). To facilitate comparison of the relative enrichment between continental groups, we used all private polymorphisms to infer a mutation rate for each 3-mer (per generation per site) calibrated to the average estimated *de novo* mutation rate from Kong et. al7 (**Methods**).

In addition to the C→T polymorphisms mentioned above, we observed six additional contexts in our top ranked set that have not yet been specifically noted in previous studies (**Table 1**). Interestingly GAT→T and ACC→A polymorphisms displayed a similar profile of heterogeneity: highest relative rates in East and South Asia Africa and intermediate levels in Europe, relative to Africa. This suggests that GAT→T and ACC→A could represent a signature of polymorphism types enriched in Asia. A third substitution class, ACA→T, was highest in Africans, with lower rates in East Asians and Europeans, perhaps suggestive an Africa-enriched signature. One polymorphism - TCA→T showed enrichment in Europe and South Asia, a profile similar to the previously reported C→T polymorphism contexts in Europe (**Table 1**). The ACT→T context was similar to this, with greatest representation in Europe and South Asia, but with a higher proportion in Africa compared to other substitution classes from this group. A final polymorphism, GCC→T, was likewise enriched in Europe and South Asia, but also showed an elevation in East Asia.

As an additional validation, we estimated mutation rate separately on each chromosome and found that these patterns of enrichement hold relatively consistently across the genome for each of these contexts (**Supplementary Note**). Taken together, these results indicate that there may be several previously unreported signatures of variation in mutation rate observed at the 3-mer scale, beyond the previously reported signal of European C→T enrichment.

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

Since there appeared to be polymorphism types at the 3-mer level which varied between populations in different ways, we next sought to identify sets of substation classes that share similar profiles of enrichment or depletion across the globe, and which thus might be influenced by a common underlying mechanism. To this end, we generated a hierarchical clustering of 3-mer polymorphism types based upon their relative inferred mutation rates in each of the twenty 1,000 Genomes Project subpopulations comprising the non-admixed continental groups from our first analysis.

We highlight five “profiles” of substitution rates that emerge from the clusters of 3-mer substitution classes (**Figure 1A**). Profile 1 corresponds to European C→T enrichment, and includes all four 3-mers previously reported to comprise this signal4,5 (**Figure 1B**). The remaining two polymorphisms in this group, TCA→T and ACT→T are noted in the previous section (Table 1), and represent variable polymorphisms not previously highlighted as part of this group of Europe-enriched mutations.

The next profile (#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 for their cross-continental heterogeneity (**Table 1**). The third substitution class, GAC→T appeared heterogeneous across continents in a similar way (16th ranked 3-mer polymorphism, Pordered = 2 x 10-33). They are followed by profile #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 (profile 4a and 4b) that appeared enriched in Japan and other groups in East Asia, relative to other continental groups. 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 (**Supplementary Note**). 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, profiles #2 and #4 may represent two distinct signatures of enrichment for certain mutation types in Asia.

The final profile (#5) is comprised of CpG transversions which appear to be elevated in Africa. However, Harris and Pritchard5 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 this profile 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.

**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 mutation8,9, we next sought to determine which (if any) of the 3-mer signals that have been identified may actually be driven by effects at broader windows of sequence context. To achieve this, we reconsidered each interesting 3-mer substitution type identified in the previous section with two additional flanking nucleotide bases of local sequence context (i.e., ‘a 7-mer’ window). This subdivided each 3-mer substitution into 256 distinct ‘7-mer’ classes, allowing us to ask whether the population-specific heterogeneity was general to all 7-mer expansions or just a subset of specific contexts. To this end, we 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,** case I). 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,** case II). 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).

We found that nearly all of the 3-mers comprising signals 1, 2, and 3 matched case II (**Figure 2B** **and** **Supplementary Note**). Thus the global variation in European C→T elevation, the CpG transitions, and the Asian GAT→T, ACC→A, and GAC→T elevation was not obviously correlated with sequence context features beyond a single flanking nucleotide base. However, the polymorphisms comprising profile #4 more closely matched case III, indicating that the Japanese enrichment of the \*AC→C and TAT→T substitutions might be driven by a handful of 7-mer polymorphisms which are highly heterogeneous in East Asia (**Figure 3B and Supplementary Note**).

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 profile #4 polymorphism proportion) and Chinese Dai from Xishuangbana (CDX, lower profile #4 polymorphism proportion). We found nine 7-mer polymorphism contexts that were elevated in JPT relative to CDX (FDR-adjusted P < 0.05, **Figure 3A**). Seven of them have the shared motif XXXACAG→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** **and Supplementary Note**). Curiously, we also observed that four out of the nine polymorphism types were enriched on the X chromosome in East Asia, relative to the autosomes (the other five types had too few observed polymorphisms on the X chromosome to justify a valid statistical test).

**Variable polymorphism types among higher sequence contexts**

Motivated by this result and previous work8, we next hypothesized that additional signals of mutation rate variation might be observable only in specific pentanucloetide (*i.e.*, ‘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 homogeneity testing framework described above to each of the 1,536 possible 5-mers and the 24,576 possible 7-mer polymorphism types.

Within a 5-mer sequence context, we found that 142 out of 1,536 possible polymorphism surpassed Bonferroni multiple test correction (**Methods and Supplementary Note**). Of these, 39 represent expansions of Europe-elevated 3-mer polymorphisms which have been highlighted by previous analyses (e.g., CTCCA→T an expansion of the TCC→T 3-mer)3,4. An additional 48 represent expansions of 3-mers from profiles #1-3 described above (**Figure 1A**). However, the remaining 55 significantly variable 5-mer polymorphisms involve 3-mer contexts that have not yet been highlighted by previous work.

We next 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 surpassed Bonferroni multiple test correction (**Figure 4A, Methods, and Supplementary Note**). Of these, 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 Pordered = 2 x 10-34, **Figure 4**) corresponding to one of the Japanese-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 Japanese-enriched substitution types (**Figure 3C**). Interestingly, the third most significant unreported 7-mer polymorphism, AAACAAA→A (Pordered = 1 x 10-18) has a similar profile within East Asia as the other profile #3 polymorphisms (**Supplementary Note**). 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 groups. 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 (Pordered = 4 x 10-18), both of which were enriched in Africa (**Figure 4B and Supplementary Note**). These correspond to the 3-mer TAA→A, which is the 17th most significant polymorphism from our 3-mer-level heterogeneity analysis (Pordered = 2 x 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 (**Methods**). Our model is based on the idea of estimating contemporary (population-specific) polymorphism rates relative to ancestral (shared, or ‘cosmopolitan’) genetic variation. Compared to a model where population-specific rates of polymorphism are equivalent to rates that are not population specific, models that included additional parameters capturing population-specific mutation rate variation fit the observed data substantially better (Log-likelihood ratio test P < 1 x 10-100, **Methods, Supplementary Note**). In addition, including broader windows of sequence context with population-specific parameters further improved model fit (XXXXX –insert final model compared, **Methods, Supplementary Note**), 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 variability in polymorphism representation between human populations at a global scale. However, whether these patterns reflect a true difference in underlying mutational processes, and what those underlying causes might be, remains uncertain - 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 study4,10, 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 signature of variation, the large number of 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 groups have found various different patterns of heterogeneity in polymorphism levels that can be observed between substitution classes 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 heterogeneous 7-mers between Chinese Dai and Japanese in profile #4 contain the 7-mer motif XXXACAG→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 to segregate more substitutions than other TAA contexts (**Figures 4B and 4C**).

There are some limitations to note in our given report. First, 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, which are only recently becoming available. Additional, deeply sequenced samples from diverse populations would be ideal for further targeted hypothesis testing, validation, and improving the mathematical models designed to capture this variability. For example, in this report, we noted 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 polymorphism11. 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.

A second complication is 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 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 which mutation rates may have changed5.

It is likely that further investigation will reveal details of mechanism, evolutionary timing, and genome-wide or subpopulation-level patterns in mutation rate variation, and our report here is by no means exhaustive. 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 Project release (downloaded 02/26/2016)6 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 also 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 non-admixed 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 a continental group if it is observed in that group, but not in each of the other three 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)12.

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 utilized for each step 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 (**Supplementary Note**). 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 degree 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 as previously described5. 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 error rate (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 size13. To facilitate comparisons across populations, we calculated a mutation rate, calibrated to the average *de novo* mutation rate estimated by Kong et al7. Assuming all populations have a total mutation rate of 1.2 x 10-8, we inferred the mutation rate of a specific type (say TCC→T) as

μm = 1.2 x 10-8 x Θ-1 x Θm

Where μm represents the inferred private germline TCC→T mutation rate per generation per site, Θm represents the proportion of all TCC sites in the genome with private C/T polymorphism in the population, and Θrepresents 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 μm gives an overall genome wide mutation rate of 1.2 x 10-8 when all mutation types are pooled. 95% confidence intervals for μm were calculated using the normal approximation to the binomial, assuming the variance in Θ-1 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 that 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 pattern of rates across populations was normalized by fold difference above or below the mean rate for that profile. We used Euclidean distance to construct each heatmap and for comparison, selected because these methods gave the most clearly interpretable results and agreed the most closely with previous work3–5 (**Supplementary Note**)

**Testing for enrichment of profile #4 on the X chromosome**

In order to test the highly variable polymorphism types for enrichment on the X chromosome (**Figure 3A**), we used a one-sided binomial test to determine whether the observed proportion of privately polymorphic sites on the X chromosome was greater than *p0*, 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 *p0* we first needed to calculate the ratio, *ξ*, of X-chromosome to autosomal substitution probability across all other 7-mer types. We then used this as a scaling factor, estimating *p0* as *ξpA*, where *pA* 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 analogous to those devised in a previous report8, which capture different levels of mutation rate variation. First, we defined cosmopolitan SNPs to be those that are shared between two or more of the African, European, South Asian, and East Asian 1,000 Genomes Project samples. For a given population at a given sequence context, we assumed that the probability of recurrent mutation is 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 base 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 that 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 Project data, 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 -2ln(Λ)\_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**

R.C.A. 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 and Onur Yörük, 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. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

7. Kong, A. *et al.* Rate of de novo mutations and the importance of father’s age to disease risk. *Nature* **488,** 471–5 (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. Zhu, Y., Neeman, T., Yap, V. B. & Huttley, G. A. Statistical Methods for Identifying Sequence Motifs Affecting Point Mutations. *Genetics* **205,** (2017).

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

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

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

13. 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 polymorphism classes with 3-mer sequence contexts, shown here, were highly significant (Pordered < 1 x 10-40) according to a chi-squared test for heterogeneity across non-admixed continental groups. Boldface numbers indicate a significant difference in polymorphism proportion compared with Africa (P < 1 x 10-5) in a pairwise chi squared test. Pordered was calculated as previously described5 (**Methods**). \*To facilitate comparison, approximate private mutation rates (per generation per site) for each continent were inferred by normalizing estimate substitution probabilities using all private mutations to the *de novo* mutation rate estimated from Kong et al7, and then subsequently 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\* | Pordered |
| 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 twenty non-admixed 1,000 Genomes Project 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).

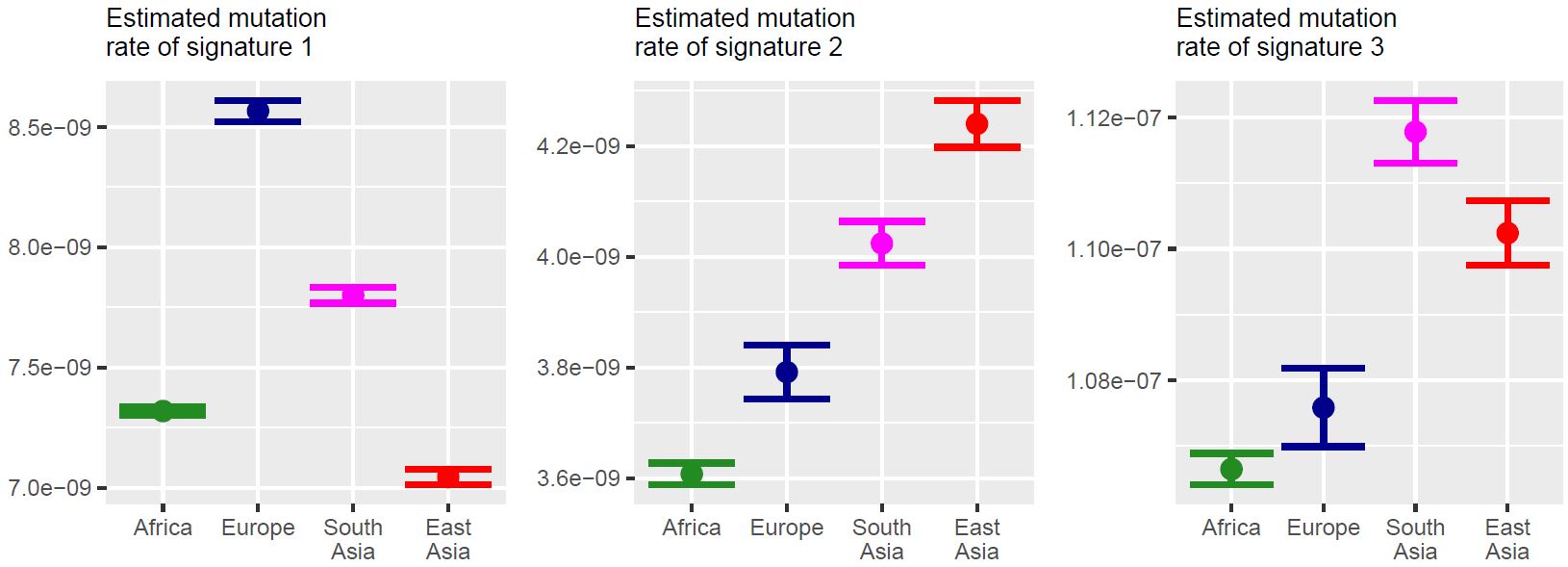
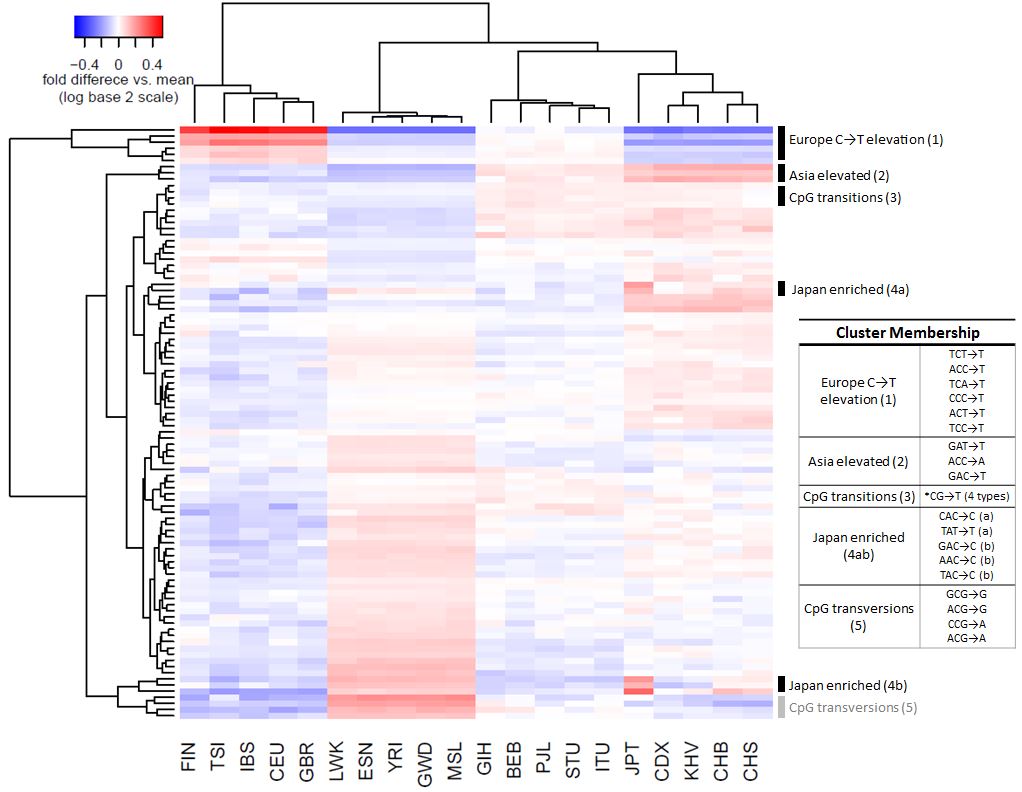
**Figure 2:** **Mutational signatures driven at the 3-mer level**. Each point represents a 7-mer expansion of the 3-mer subtype shown, plotted based on its estimated mutation rate in each of the two populations displayed. Colors indicate the log (base 10) of the number of substitutions observed for that 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.

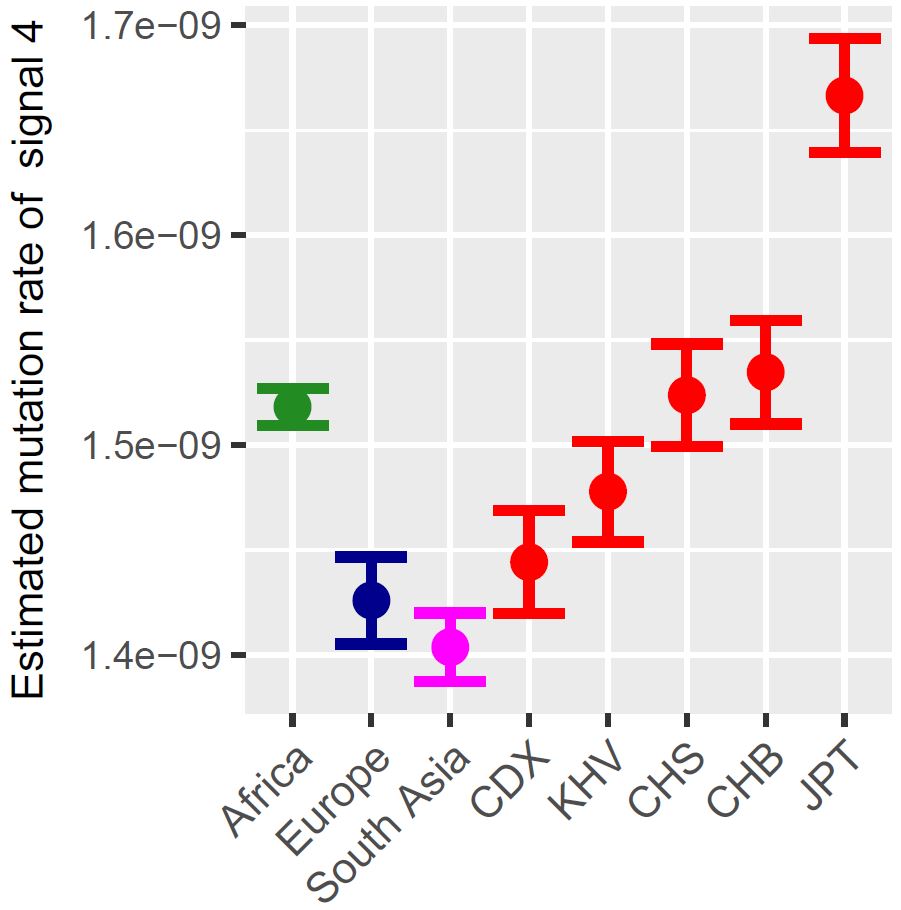
**Figure 3:** **Highly variable 7-mers within signature 4 in East Asia.** **(A)** Nine polymorphisms enriched in Japan compared to Chinese Dai (Chi-squared test FDR < 0.05). \*Fold increase in inferred mutation rate in Japan compared to CDX. Bold p values indicate nominal significance of enrichment on 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 < 1 x 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.

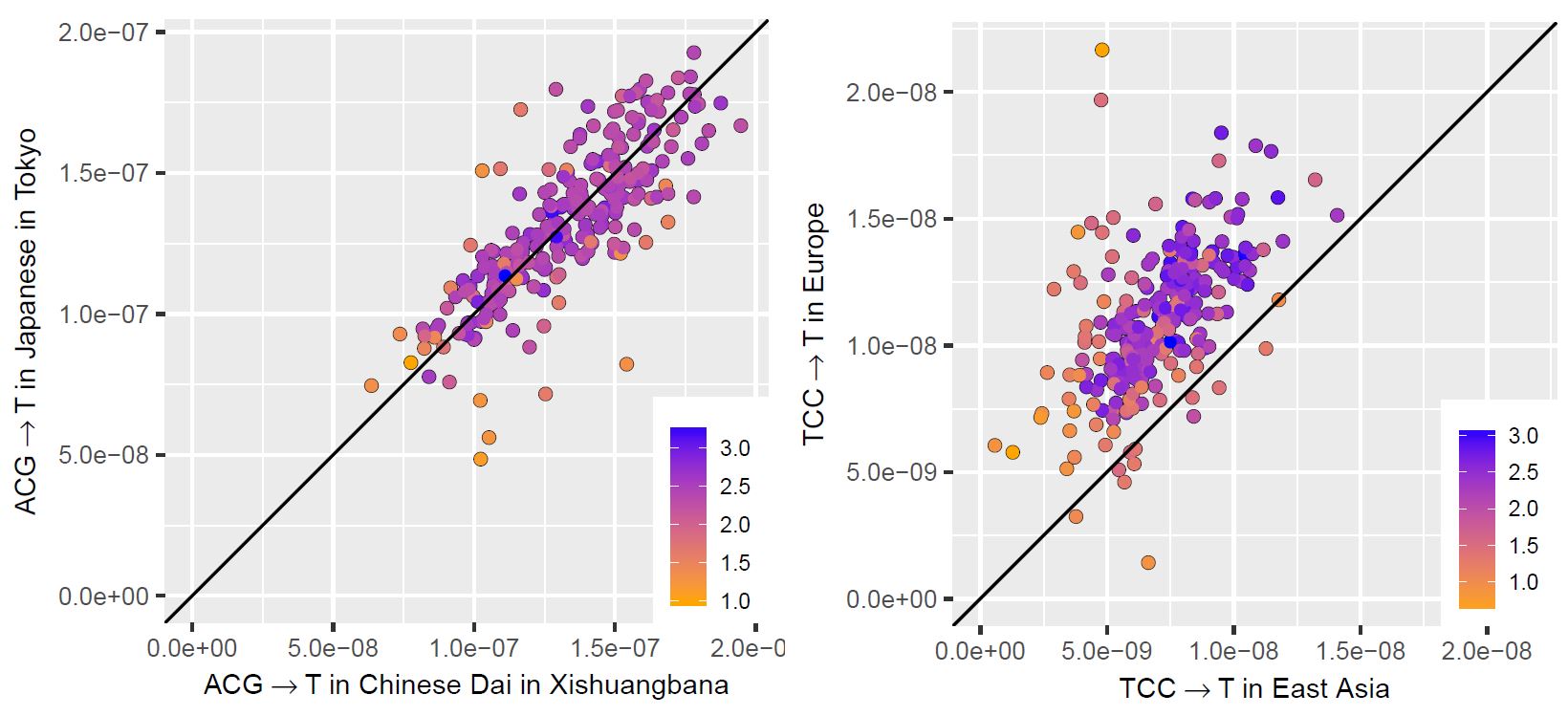
**FIGURES**

**Figure 1: ABCDE**

****

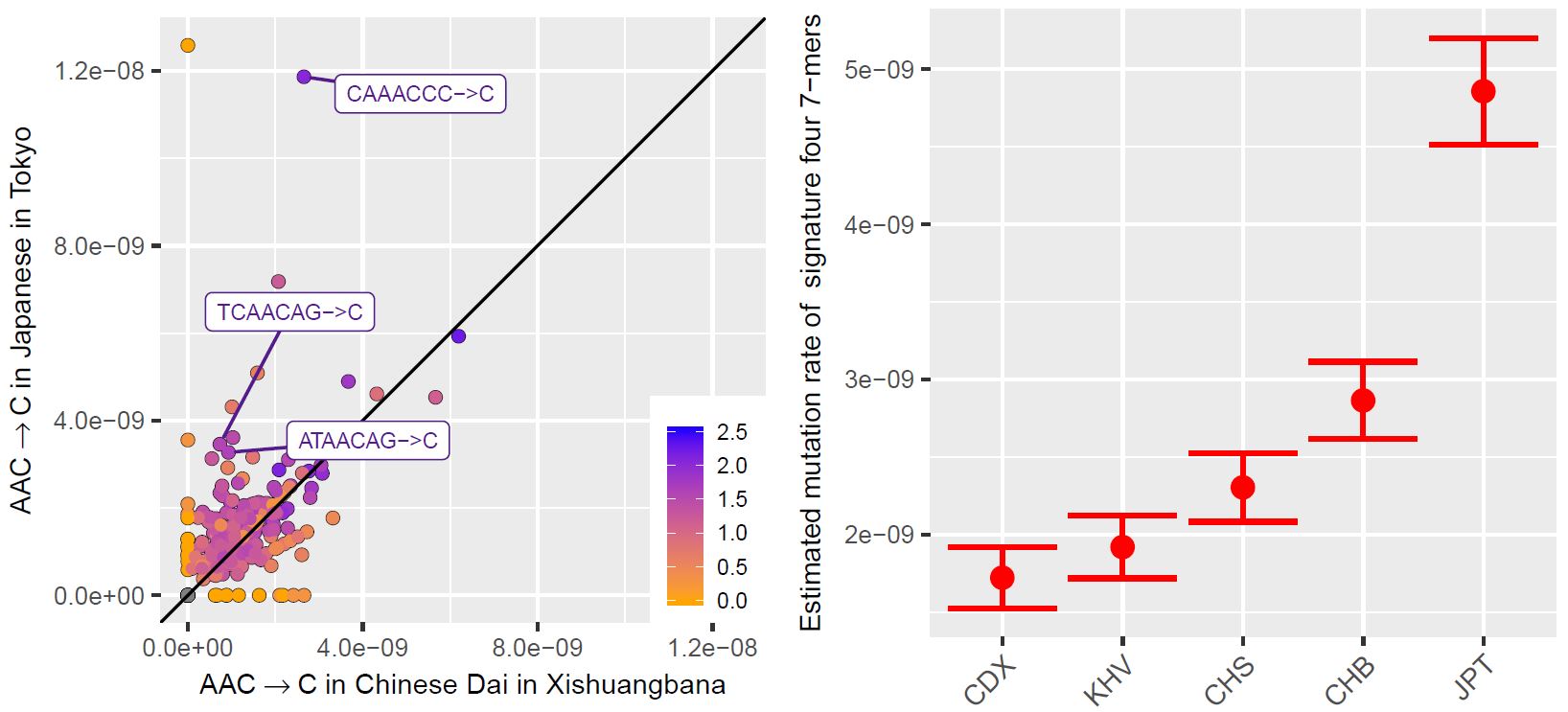


**Figure 2: AB**



**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 |

