Basic Plots and Data Gazing

Rachael 'Rocky' Aikens, Voight Lab

July 6, 2017

This document contains notes and visualizations of mutation rates for patterns of mutation on the 3mer, 5mer, and 7mer level which are interesting either because of their significance in heterogeneity test or the way that they cluster together in heatmaps.

Methodology

I use a handful of different plotting methods to gain different perpectives on the data. These are:

- **CI.plot** Given count dataframes for four populations and a polymorphism of interest m, plot the inferred mutation rates of m in each population with approximate confidence intervals.
- **chrom.box** Given count dataframes for four populations, a polymorphism of interest m, and a dataframe of genome wide context counts, plot the inferred mutation rates of m in each on each chromosome as a boxplot, labeling outliers.
- **subrate.scplot** Given count dataframes for two populations and a 3mer polymorphism type, find the rates of all expansions of that threemer in those populations and plot them against each other.
- **subrate.lplot** Given count dataframes for each non-admixed continental group and a 3mer polymorphism type, find the rates of all 5mer expansions of that threemer in those populations and plot them as as lines across the populations.

Mutation rate differences at the 3mer level

Recall the following list of the top 10 most highly significant 3mers:

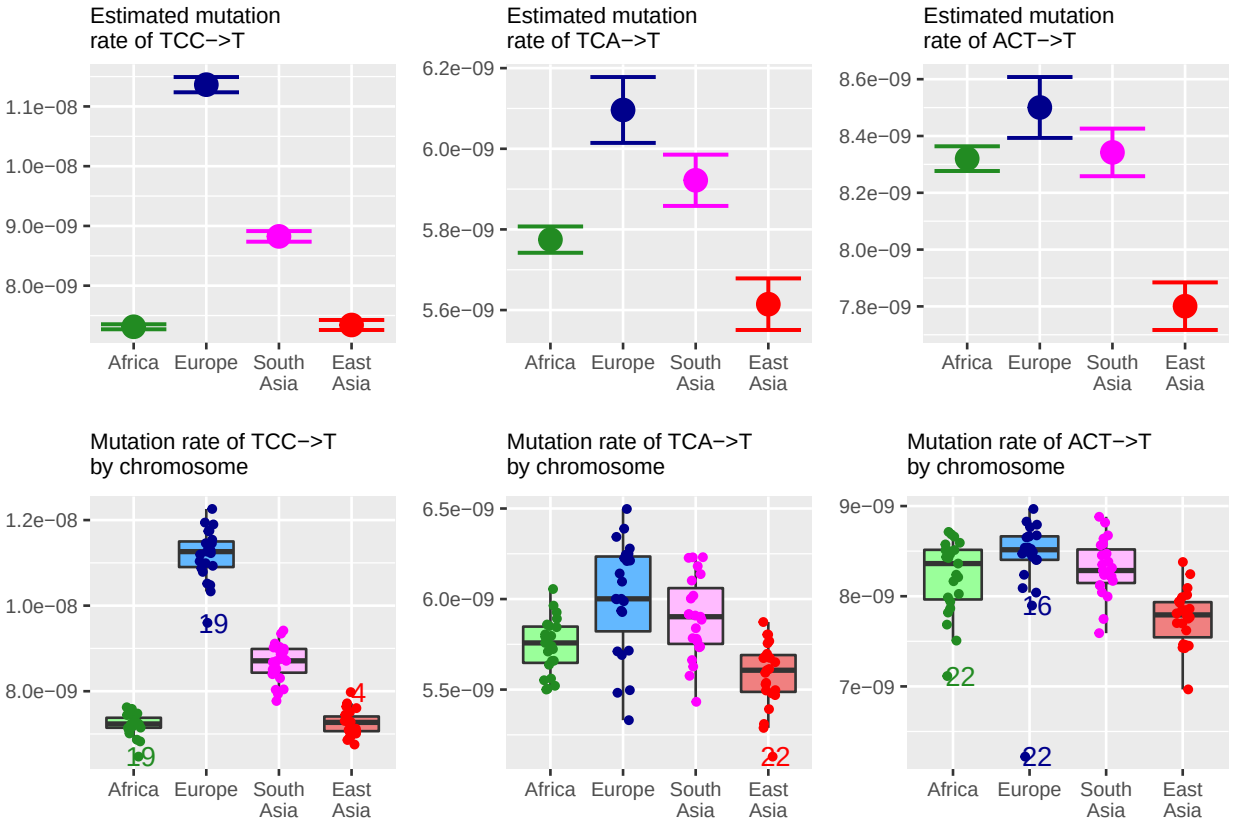
Table 1: 14 most highly significant 3mers

Context	X1mer	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
TCC->T	C->T	116126	30482	29565	37953	0.000000e+00
ACC->T	C->T	120823	24618	28466	35222	6.419786e-263
TCT->T	C->T	128789	25265	32178	37226	1.175837e-155
GAT->T	A->T	42007	7561	12866	12948	1.553628e-95
ACC->A	C->A	68375	12103	19959	20513	2.377759e-90
CCC->T	C->T	119810	21461	28885	33443	6.127063e-55
TCG->T	C->T	181089	31320	47553	51985	1.833727e-54
ACG->T	C->T	295109	50212	77743	83636	6.391329e-54
ACA->T	C->T	202388	32435	47413	53182	8.420241e-54
GCG->T	C->T	206633	35598	54787	58748	3.410460e-49
CCG->T	C->T	254277	42926	66164	71714	3.037409e-44
TCA->T	C->T	119977	21392	29578	33318	3.387884e-44
ACT->T	C->T	140014	24160	33285	38016	1.608621e-42
GCC->T	C->T	111025	20082	29128	31129	7.430059e-42

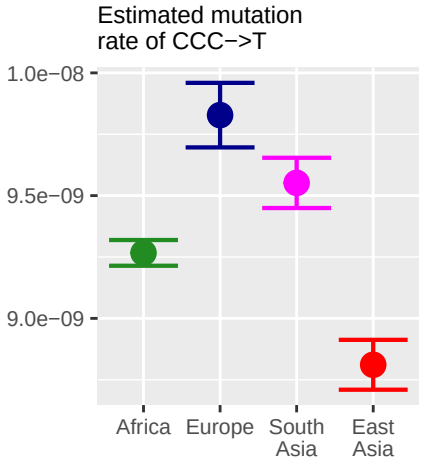
Here, I've chosen to show the top 14 results, which are all significant at $p < 10^{-40}$.

Signal 1: European C->T Elevation

Among the top polymorphisms, TCC->T, ACC->T, TCT->T, and CCC->T have already been noted as part of the signal of European C->T enrichment. In addition, #11 and #12, TCA->T and ACT->T share the same profile as the other Europe-enriched C->T mutations and cluster together with them in heatmap experiments. However, they have not been noted in previous studies. These two mutations are shown below, along with TCC->T for reference:



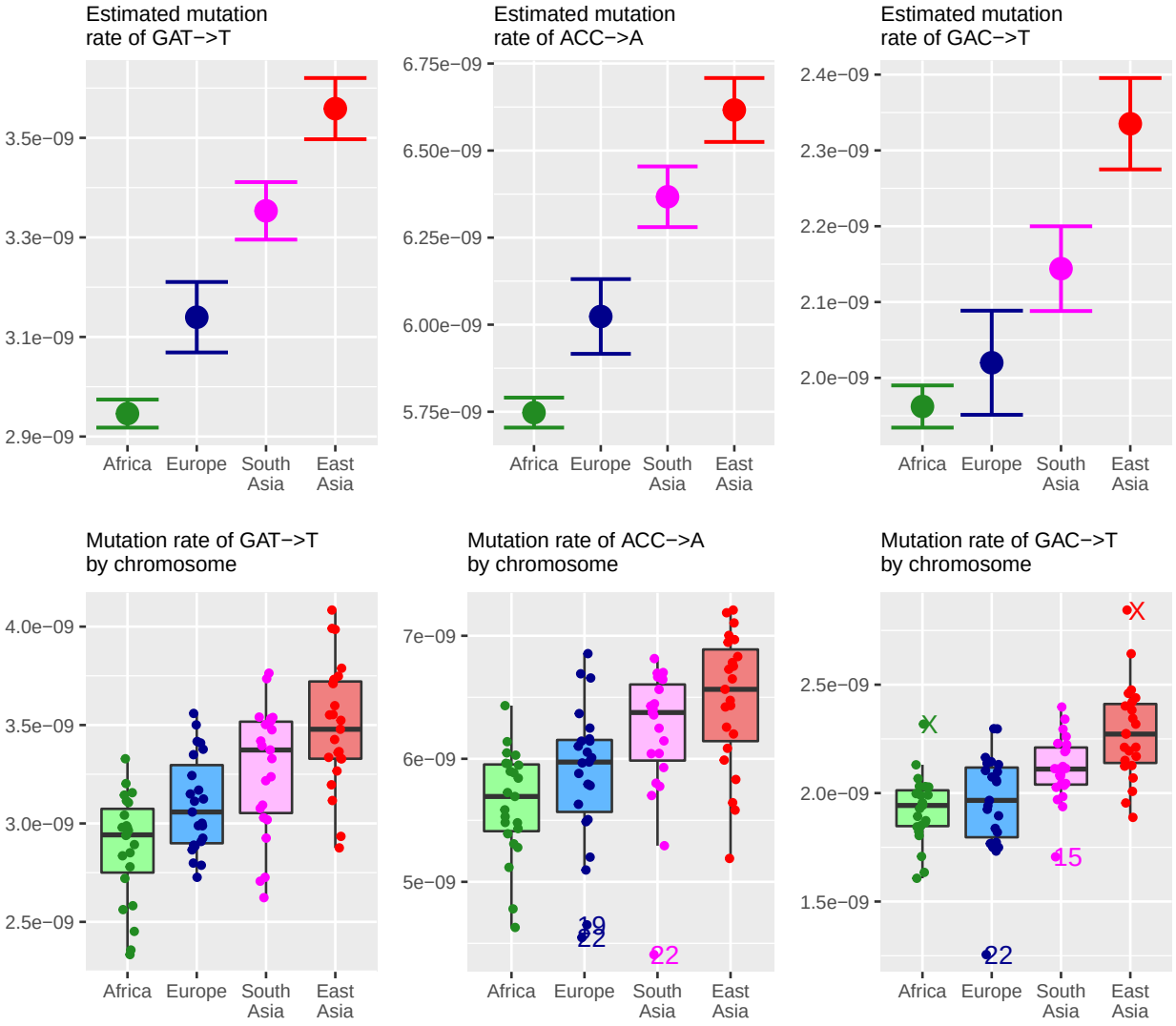
Note that although these patterns appear similar in EUR, EAS, and SAS, there are noteable differences between the relative rates in Africa. This seems to be common among the C->T signal polymorphisms. Consider for example CCC->T:



CCCT-1.bb

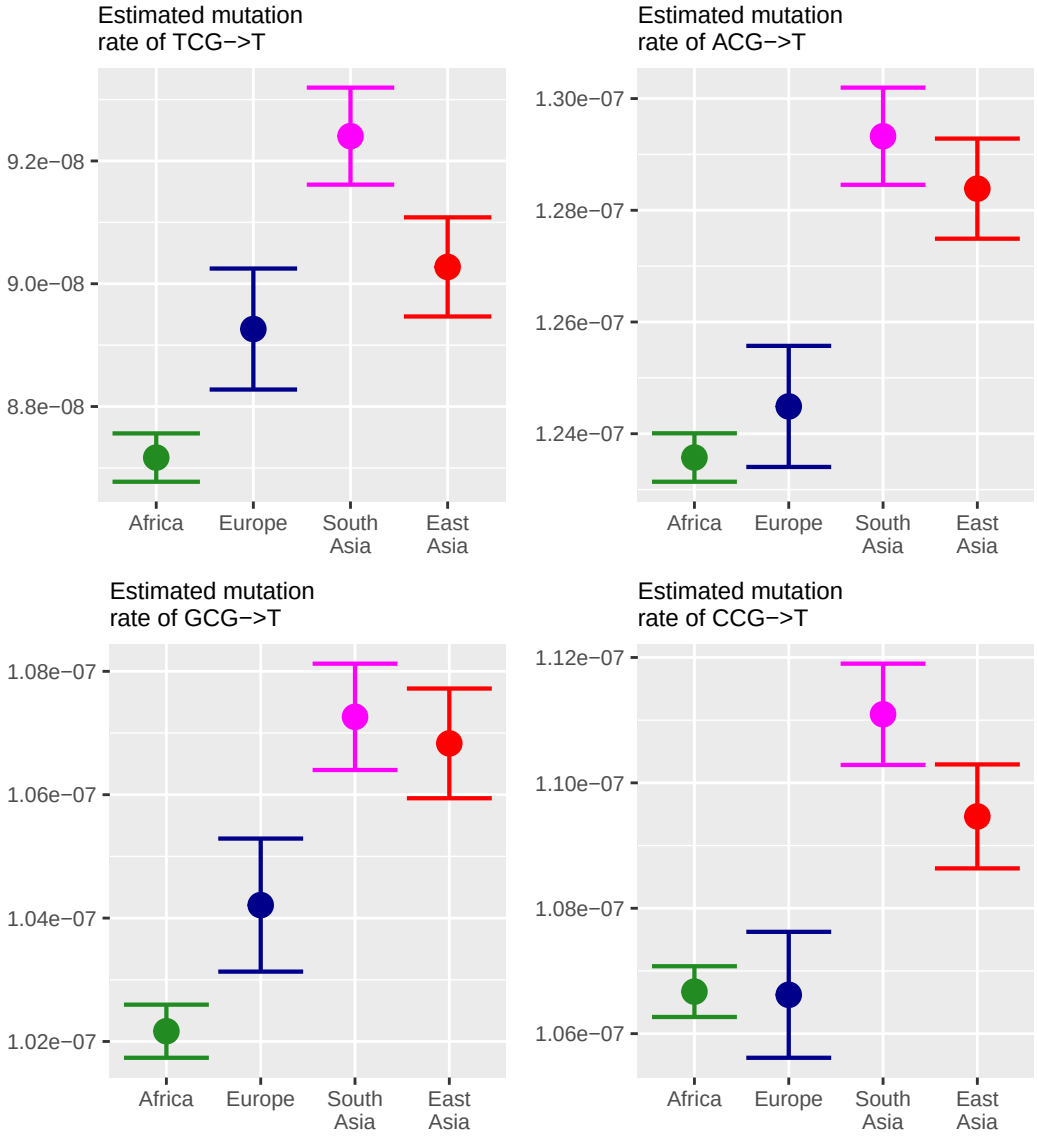
Signal 2: Enrichment of certain polymorphisms in Europe

The fourth and fifth most significant results, however, GAT->T and ACC->A have not previously been noted, and are highly significant at $p < 10^{-85}$. In addition, these mutation types are clustered with GAC->T, the 16th most highly significantly heterogeneous polymorphism type. The mutation rates for the three of these polymorphisms are shown below:



Signal 3: CpG transitions

All four CpG transitions appear among the top 11 most significant results. Ian Mathieson and David Riech have previously noted that there is some amount of variation in CpG enrichment between populations, but that this variation is slight relative to the overall rate of CpG transitions. Other studies have noted that CpGs are the most "clocklike" polymorphism types among humans and other primates. The CpG mutations are shown below:



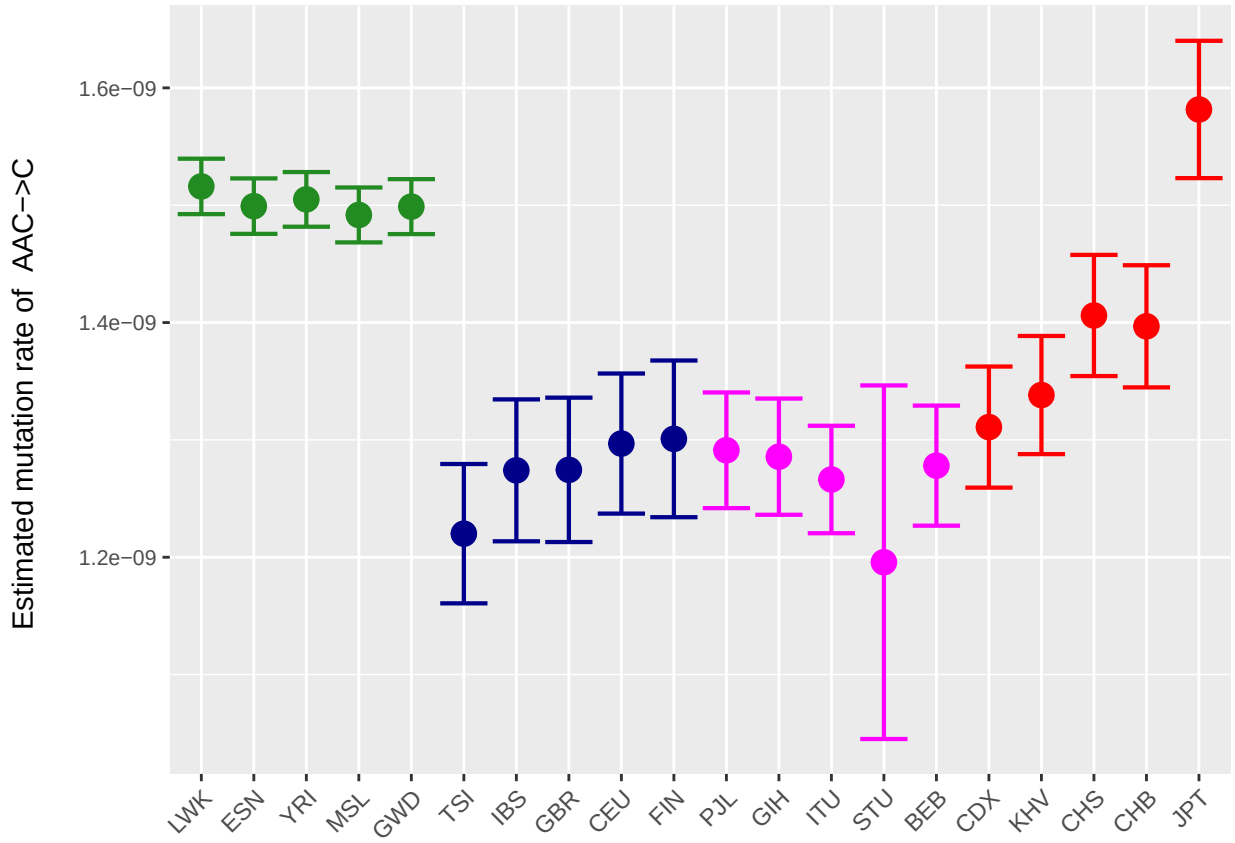
signal3-1.bb

It is worth nothing that the CpG mutations cluster together in heatmaps, even after normalization.

Signal 4: Heterogeneity within East Asia

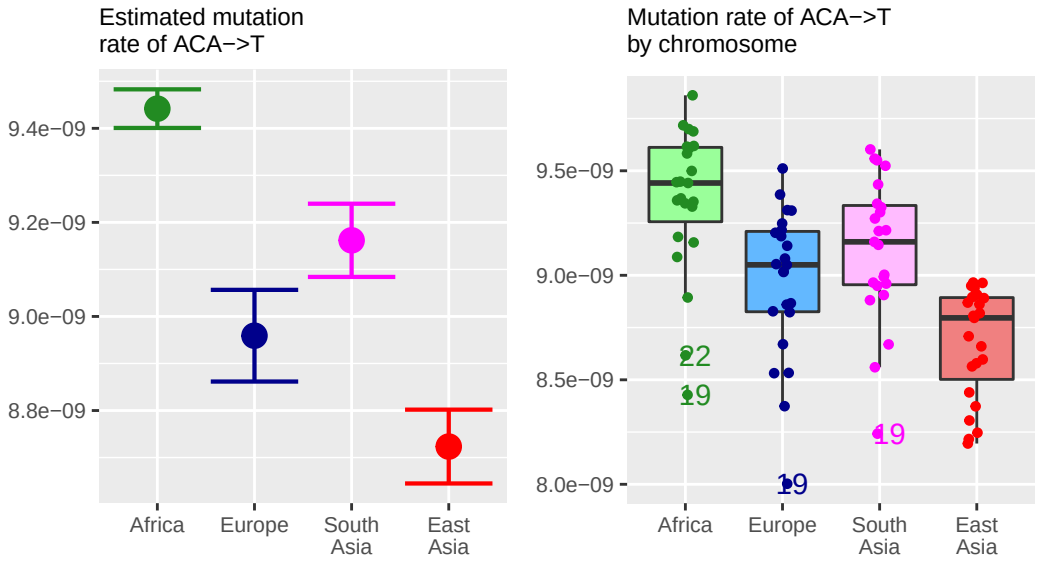
When we construct a heatmap of all 3mers, we find two clusters which appear enriched in Japan and other groups in East Asia. These clusters are comprised of the *AC->C polymorphisms, as well as TAT->T. When the mutational types are clustered using only the data from East Asia, excluding other continental groups, we find that these two clusters merge, and the additional polymorphism CAC->C is added. This group is in correspondence with results from Harris and Pritchard, who find that *AC->C, TAT->T, and CAC->C mutation types separate East Asians in a principal component analysis.

These mutation types have the global profiles shown below:

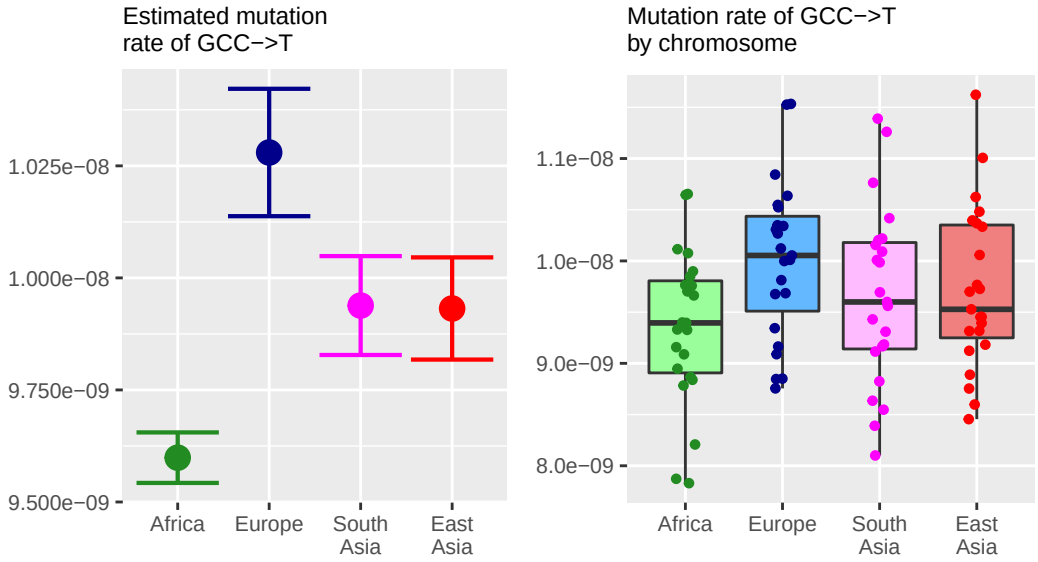


Additional Signals

Two remaining polymorphism types have not yet been mentioned. The first highly significant signal which has not been is ACA->T, which appears elevated in Africa.



The second is GCC->T,which is elevated in East Asia in addition to Europe and South Asia.

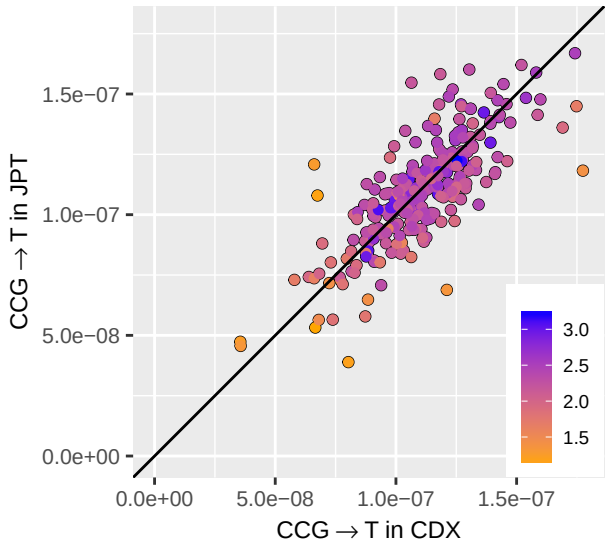


Finally, an additional pattern of interest is the shared profiles of certain CpG transversions, which appear to be enriched in Africa, and which cluster together in heatmaps. However, none of these polymorphism types are significant based on homogeneity tests (predictably, since CpG transversions are rare), and Harris and Pritchard have noted that the proportions of CpG transversions in 1,000 genomes and the Simmons Diversity Genome Project dataset tend not to agree, suggesting that this pattern may be driven by some sequencing artifact.

Heterogeneity of 3mer signals within higher order sequence context models

Now that we have identified several groups of 3mer polymorphisms which appear to vary across populations, we would like to know whether local sequence context (up to 3 bases from the substitution) plays a role in driving the variation we observe. To do this, we can bin the polymorphisms from any given 3mer (say, TCC->T) into 256 different 7mer expansions (e.g. ACTCCCT->A), and observe how their rates differ between populations.

We'll begin with a simple null example. The CpG transition CCG->T is relatively the same within East Asia. In this case, we expect to see some variation in rate between CCG->T expanisions due to 7mer effects that are uniform across all populations. However, we expect the rates of any given 7mer to be equal between East Asian subpopulations. This is what we see below, in Chinese Dai versus Japanese from Tokyo:



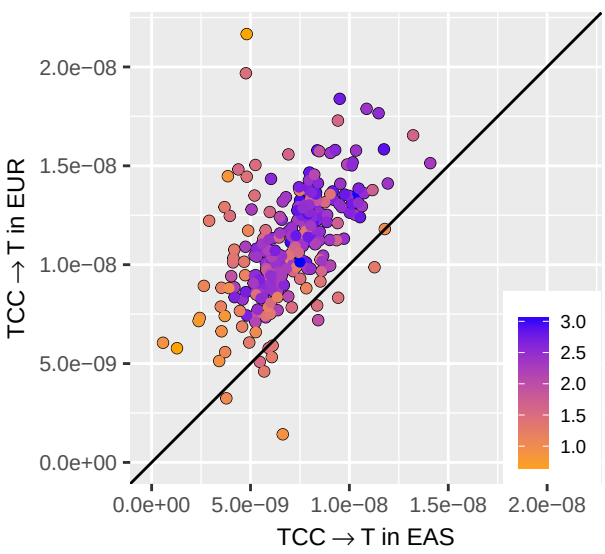
heterogeneity plot-1.bb

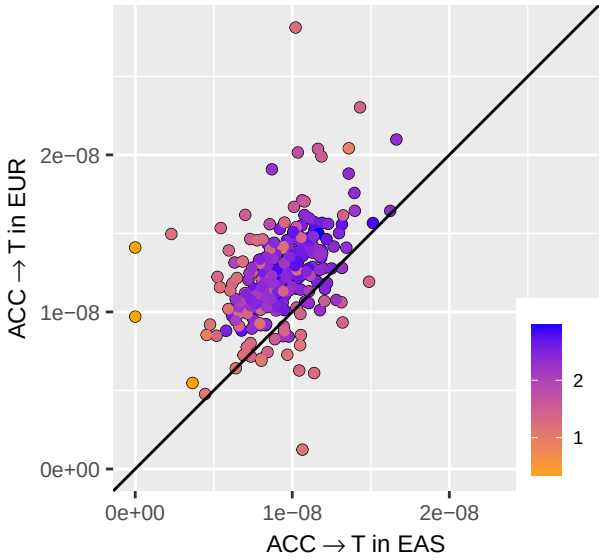
Here, each point represents a polymorphism, and the points are colored by the base 10 log of sample size (number of polymorphisms observed). We see some noise here, mostly among the yellow-colored (more uncertain) polymorphisms. However, most points lie along the y = x line.

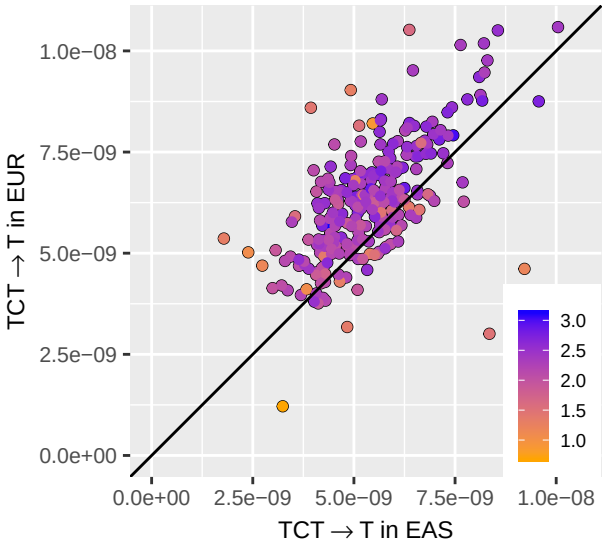
Signal 1: European C->T elevation

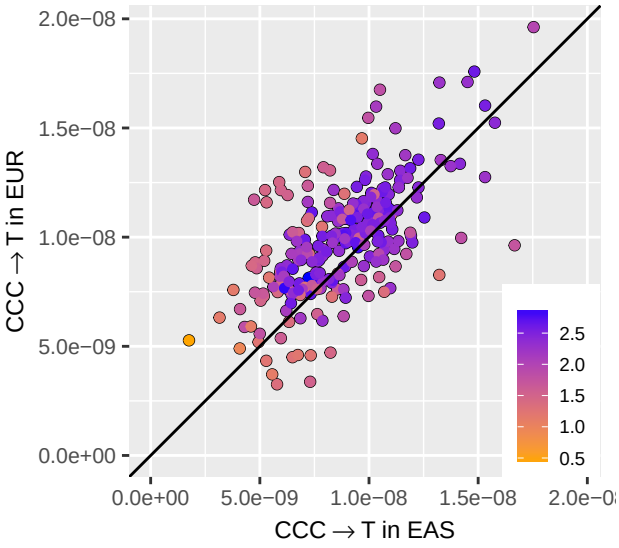
For each signal we've highlighted at the 3mer level, we'd like to know whether this is a true 3mer effect or whether this is driven by broader sequence context.

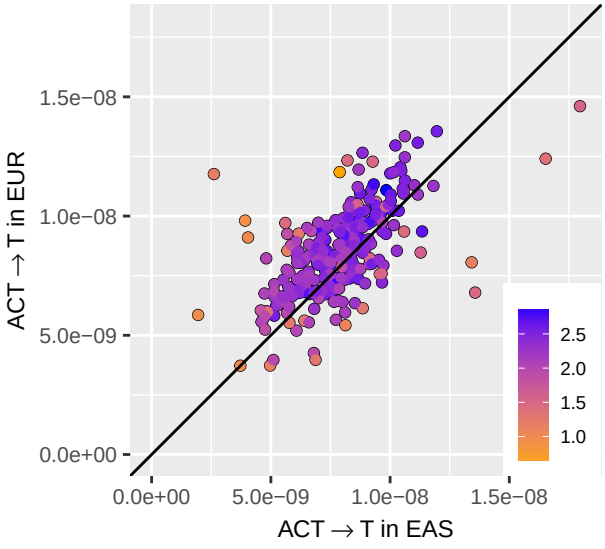
Let's consider the top 3mer polymorphisms for the European C->T elevation, shown on the next page. Here, we can see that, for each polymorphism type, the distribution of all 7mers lies slightly above the y = x line. This indicates that signal 1 is determined by local sequence context effects at 1 or fewer base pairs from the substitution locus.

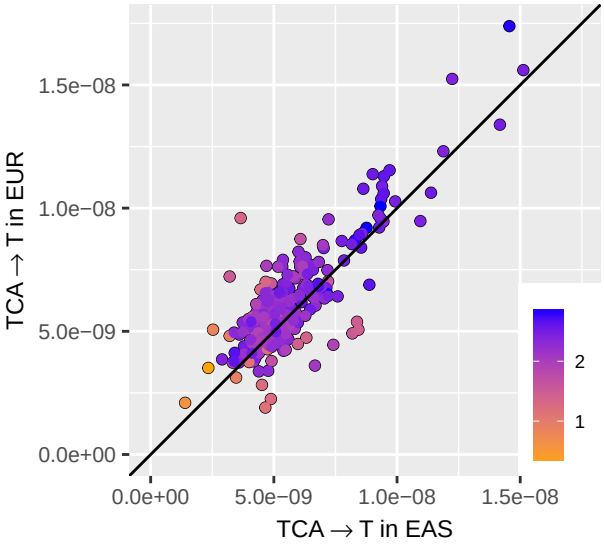

TCC → T in EUR
TCC → T in EAS
2.0e−08
1.5e−08
1.0e−08
5.0e−09
0.0e+00
3.0
2.5
2.0
1.5
1.0


ACC → T in EUR
ACC → T in EAS
2e−08
1e−08
0e+00
2
1


TCT → T in EUR
TCT → T in EAS
1.0e−08
7.5e−09
5.0e−09
2.5e−09
0.0e+00
3.0
2.5
2.0
1.5
1.0

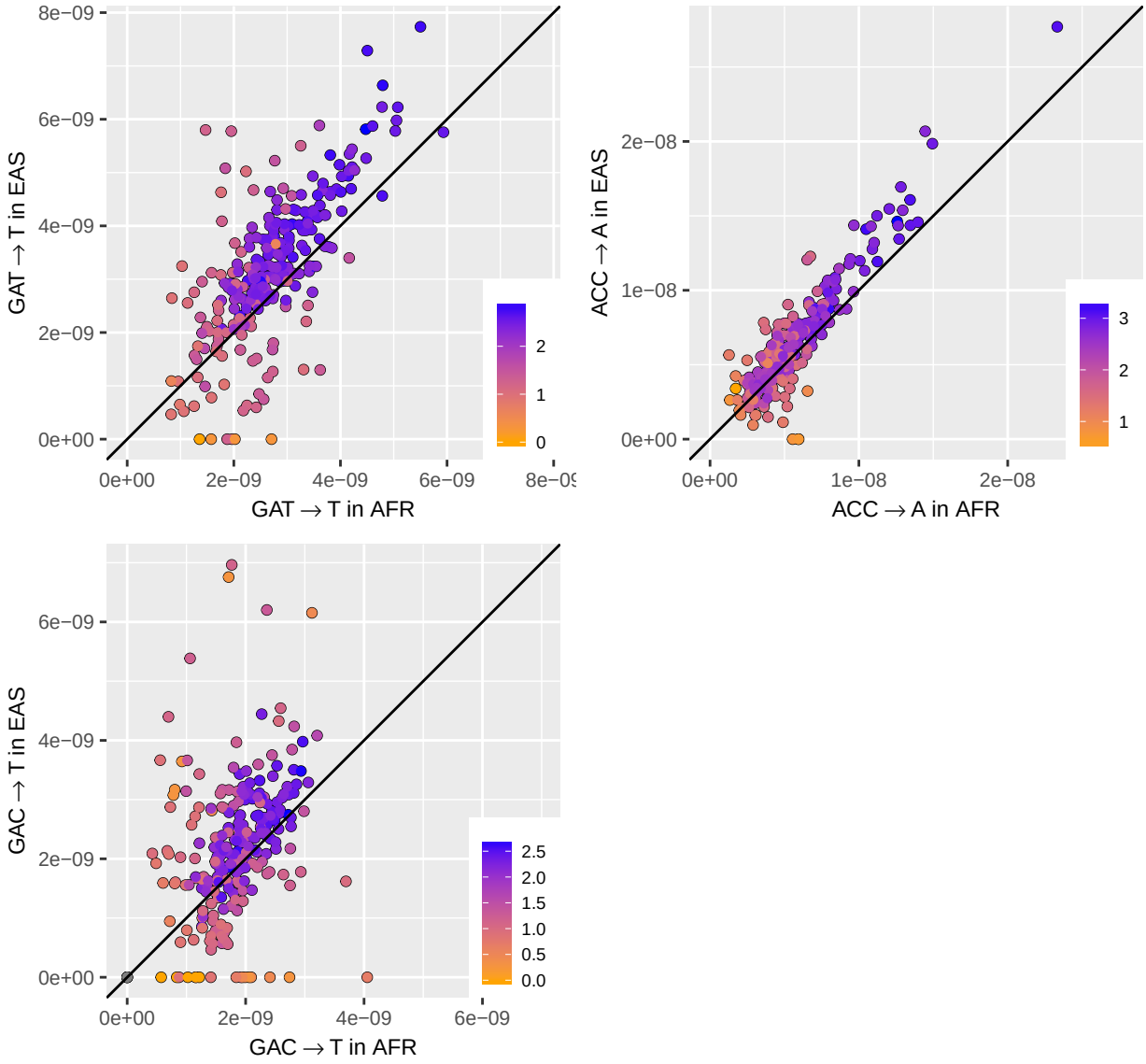

CCC → T in EUR
CCC → T in EAS
2.0e−08
1.5e−08
1.0e−08
5.0e−09
0.0e+00
2.5
2.0
1.5
1.0
0.5


ACT → T in EUR
ACT → T in EAS
1.5e−08
1.0e−08
5.0e−09
0.0e+00
2.5
2.0
1.5
1.0


TCA → T in EUR
TCA → T in EAS
1.5e−08
1.0e−08
5.0e−09
0.0e+00
2
1

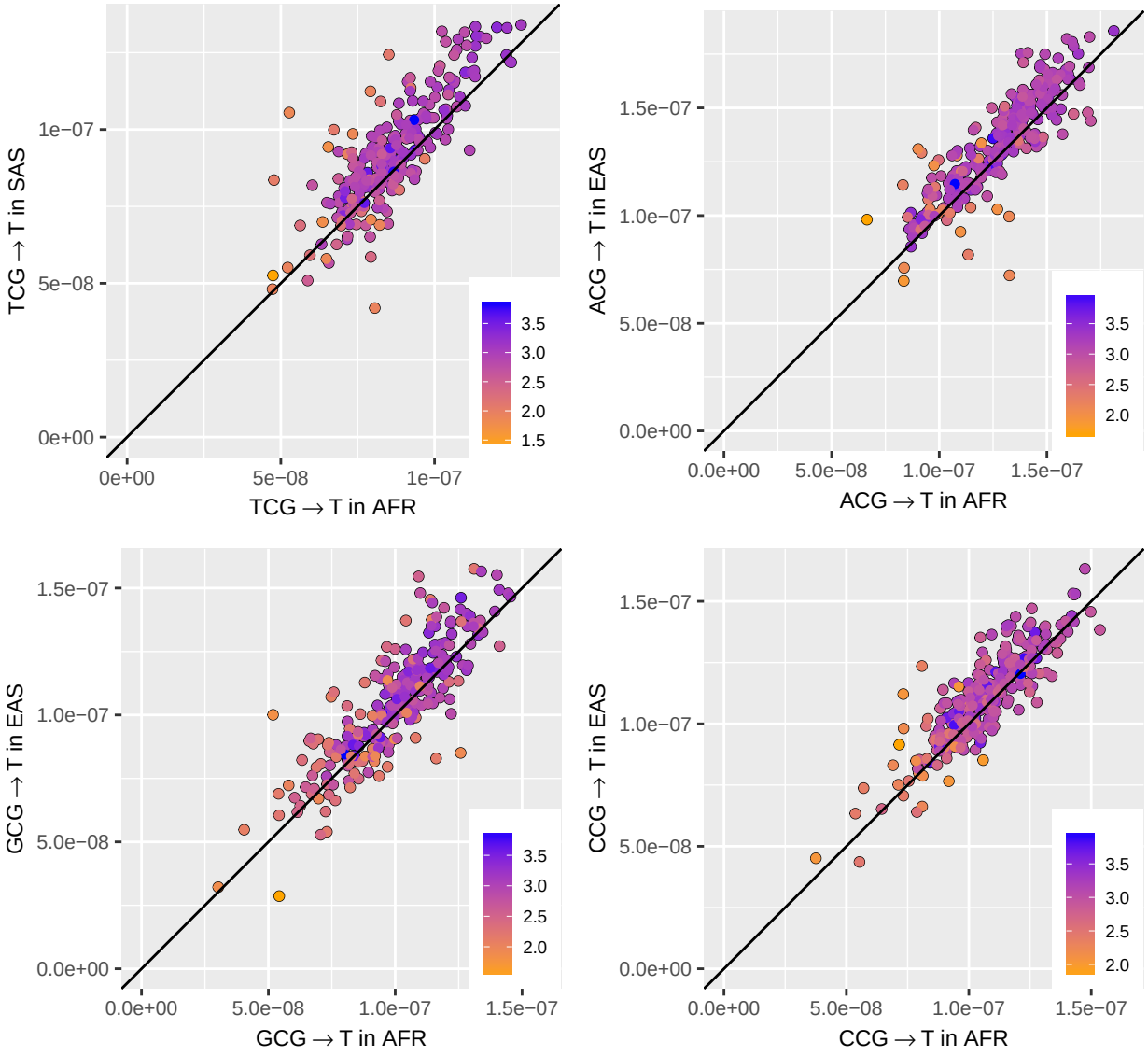
Signal 2: Enrichment of certain polymorphisms in East Asia

We come to a similar conclusion for Signal 2:



Signal 3: CpG polymorphisms

SAS vs AFR. Unsurprisingly, this appears to be a 3mer-level signal.



Signal 4: Heterogeneity within East Asia

Now we shift to discussing the fourth signal: *AC->C and TAT->T, which appear to be elevated in East Asia, most notably in certain individuals from Japan and China. In order to understand how this mutation type varies within East Asia, I will plot the rates of these polymorphisms in Japan versus Chinese Dai in Xishuangbana. These can be seen on the following page. In contrast to the previous plots, most points lie along the line y = x, with a few outliers. This indicates that some cues among the 7mer sequence context may be important.

Intrigued by these findings, we set out to begin to identify putative 7mer types responsible for this signature. To this end, we considered each of the 1280 possible 7mer 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).

Context	p	fdr
TTTATTT->T	0.0000000	0.0000000
AATACAG->C	0.0000925	0.0115197
AGTACAG->C	0.0000001	0.0000206
ATAACAG->C	0.0003625	0.0300886
ATGACAG->C	0.0001374	0.0146650
CAAACCC->C	0.0000000	0.0000000
CCCACAG->C	0.0000570	0.0106536
TCCACAG->C	0.0002214	0.0206713
TCAACAG->C	0.0000850	0.0115197

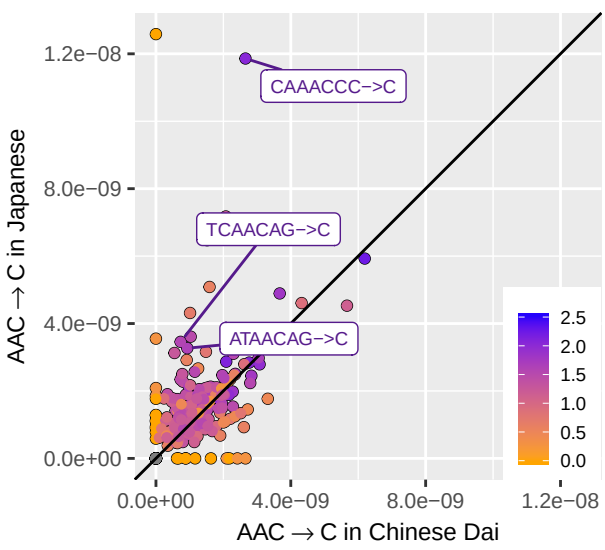
Among these polymorphisms, the motif ACAG->C appears to be very common.

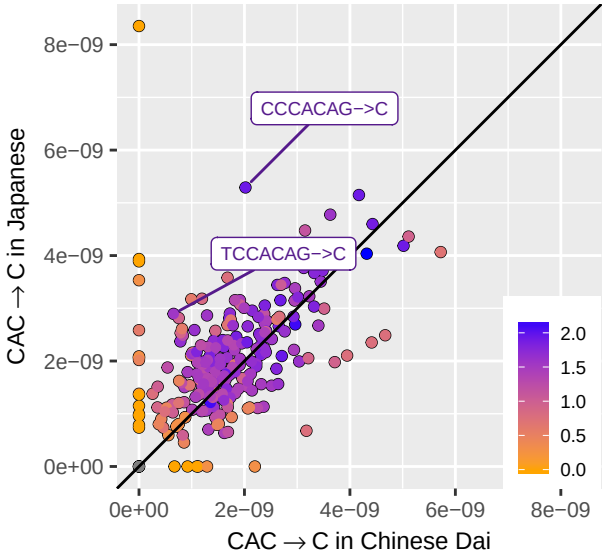
Testing for enrichment on the X chromosome, we find that 4 of 9 of these polymorphisms is significantly enriched on X (see below). The probability of observing this number of significant results by chance alone is:

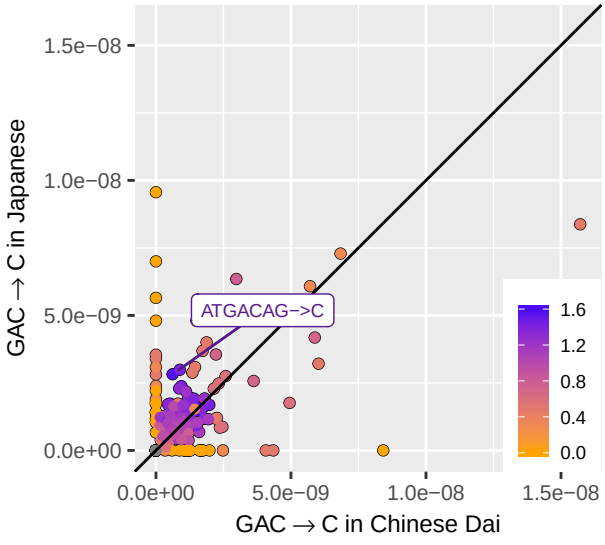
```
binom.test(4,9,0.05)$p.value
```

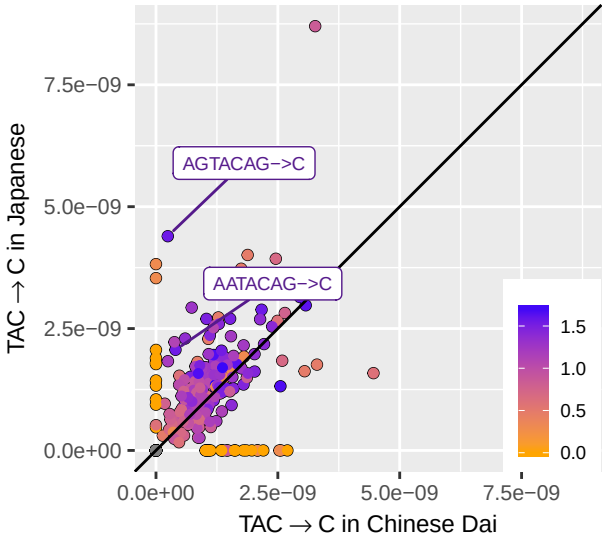
```
## [1] 0.0006425747
```

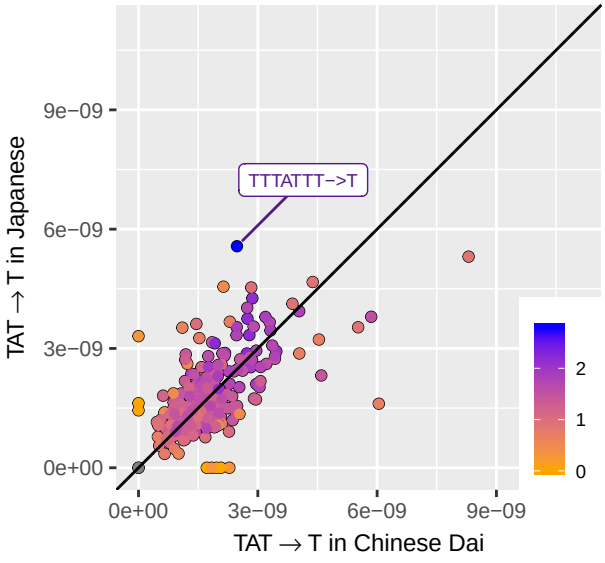
Context	Autosomes	X	Autosomal_sites	X_sites	alpha	p.0	p.MLE	p
TTTATTT->T	587	65	1444601	119969	0.961597	0.000391	0.000542	0.006997
AATACAG->C	33	1	258885	21463	0.961597	0.000123	0.000047	0.927992
AGTACAG->C	37	4	143524	10787	0.961597	0.000248	0.000371	0.280188
ATAACAG->C	41	4	185557	15802	0.961597	0.000212	0.000253	0.432330
ATGACAG->C	30	4	196359	15670	0.961597	0.000147	0.000255	0.201078
CAAACCC->C	80	27	136995	10993	0.961597	0.000562	0.002456	0.000000
CCCACAG->C	61	25	206875	15550	0.961597	0.000284	0.001608	0.000000
TCCACAG->C	35	3	186458	13398	0.961597	0.000181	0.000224	0.435089
TCAACAG->C	29	7	165015	13875	0.961597	0.000169	0.000505	0.010312


CAAACCC->C
TCAACAG->C
ATAACAG->C
AAC → C in Japanese
AAC → C in Chinese Dai
1.2e−08
8.0e−09
4.0e−09
0.0e+00
2.5
2.0
1.5
1.0
0.5
0.0


CCCACAG->C
TCCACAG->C
CAC → C in Japanese
CAC → C in Chinese Dai
8e−09
6e−09
4e−09
2e−09
0e+00
2.0
1.5
1.0
0.5
0.0

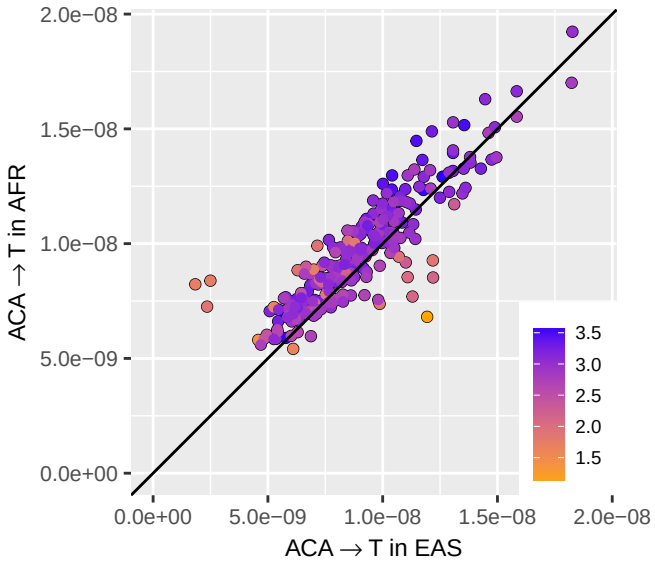

ATGACAG->C
GAC → C in Japanese
GAC → C in Chinese Dai
1.5e−08
1.0e−08
5.0e−09
0.0e+00
1.6
1.2
0.8
0.4
0.0


AGTACAG->C
AATACAG->C
TAC → C in Japanese
TAC → C in Chinese Dai
7.5e−09
5.0e−09
2.5e−09
0.0e+00
1.5
1.0
0.5
0.0


TTTATTT->T
TAT → T in Japanese
TAT → T in Chinese Dai
9e−09
6e−09
3e−09
0e+00
2
1
0

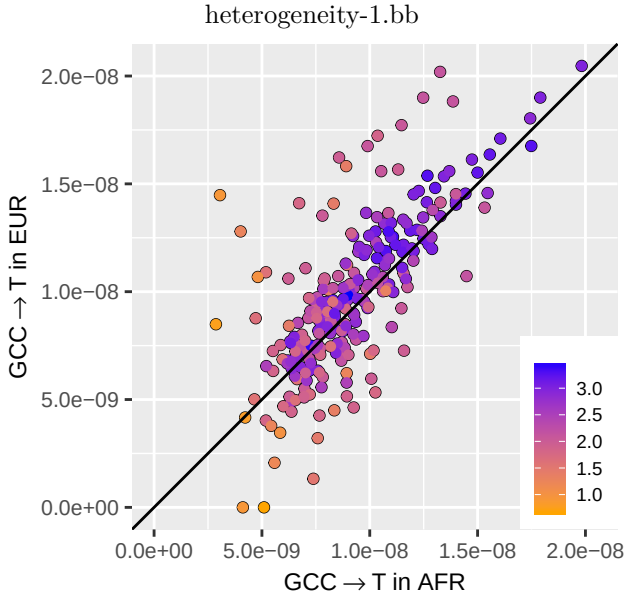
Additional 3mer signals

Here is ACA->T



This appears also to be a 3mer-level signal, although three mutation types at the bottom right appear to be outliers.

The last signal is GCC->T:



This one appears also to be a 3mer signal, but the whole story may be more complicated, since we are seeing some abnormal results at the 5mer level (see notebook).

Novel mutation rate differences at the 7mer level

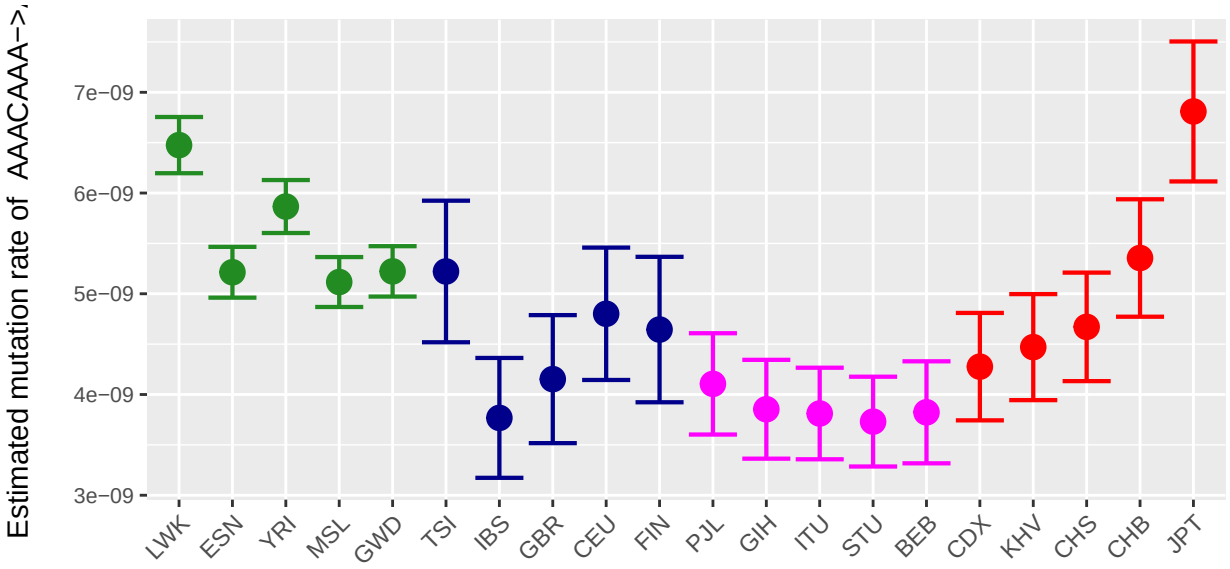
After removing known 3mer signals (TCC->T, ACC->T, TCT->T, ATC->C, ACC->C, and GAT->T), we have the following top significant results for 7mers:

Table 4: 10 most highly significant 7mers, after removing top 3mer signals

Context	X3mer	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
CAAACCC->C	AAC->C	118	16	107	10	1.628163e-34
TTTATTT->T	TAT->T	2494	344	652	410	5.112185e-22
TTTAAAA->T	TAA->T	10654	1565	2350	2424	1.816997e-20
AAACAAA->A	ACA->A	2773	377	620	489	1.116270e-18
ATTAAAA->T	TAA->T	3355	402	699	691	2.639385e-18
CTGCATA->G	GCA->G	65	13	51	12	4.348405e-11
TATATAT->G	TAT->G	6250	943	1349	1461	1.040820e-10
ACTAAAA->G	TAA->G	1956	410	662	590	1.294195e-10
AGTACAG->C	TAC->C	47	12	41	9	5.319299e-10
TATATTT->T	TAT->T	949	123	170	171	1.469412e-08

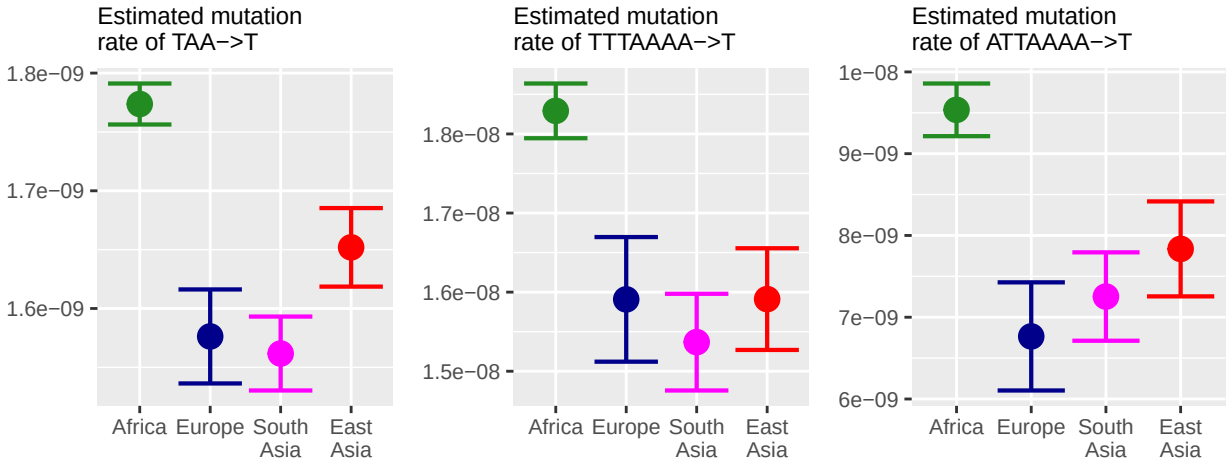
Additional polymorphisms heterogeneous in East Asia

Three of these, CAAACCC->C, TTTATTT->T, and AGTACAG->C are among the 7mers which we observe to be enriched in Japan. Interestingly, another two also appear enriched in Japan: AAACAAA->A, and CTGCATA->G. Their gloabel profiles are shown below.

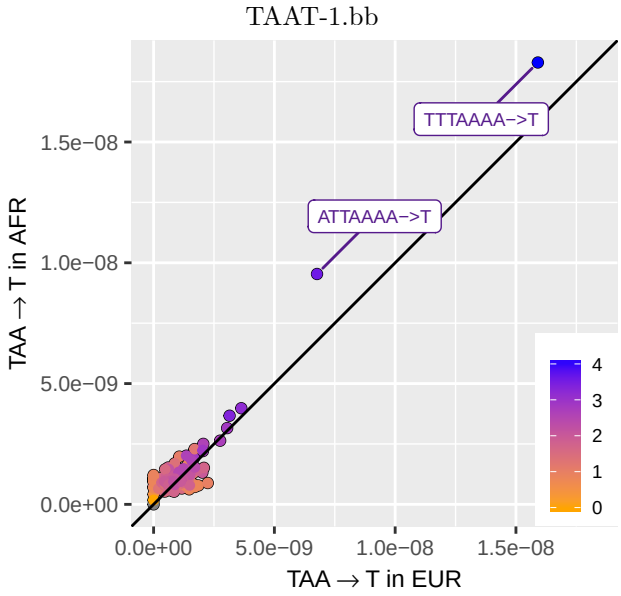


TAA->T

The remaining polymorphisms are all within A->T rich contexts. Here we examine the first, a pair of polymorphisms with the 3mer subcontext TAA->T.

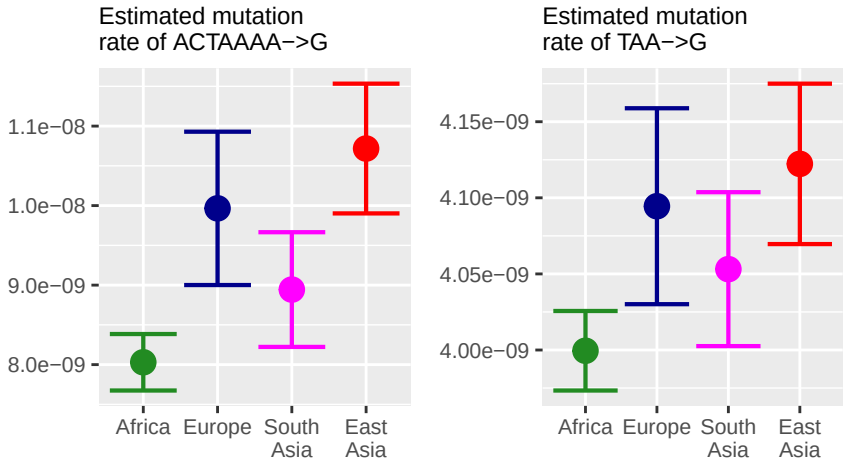


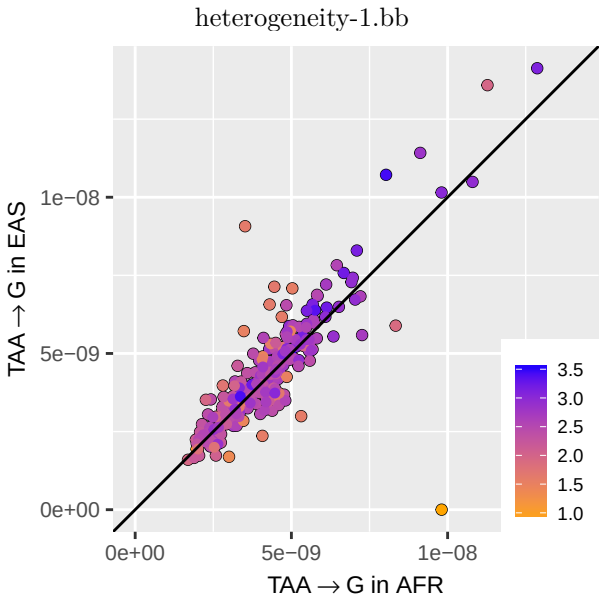
For the most part, the profiles of TTTAAAA->T and ATTAAAA->T match that of the broader 3mer subcontext. However, upon closer examination we find that TTTAAAA->T may be outlier contexts. For most other TAA->T exansions, in fact, the rates in Africa are in agreement with the rates in Europe. Only these two highly variable contexts appear to be driving the signal on the 3mer level.



TAA->G

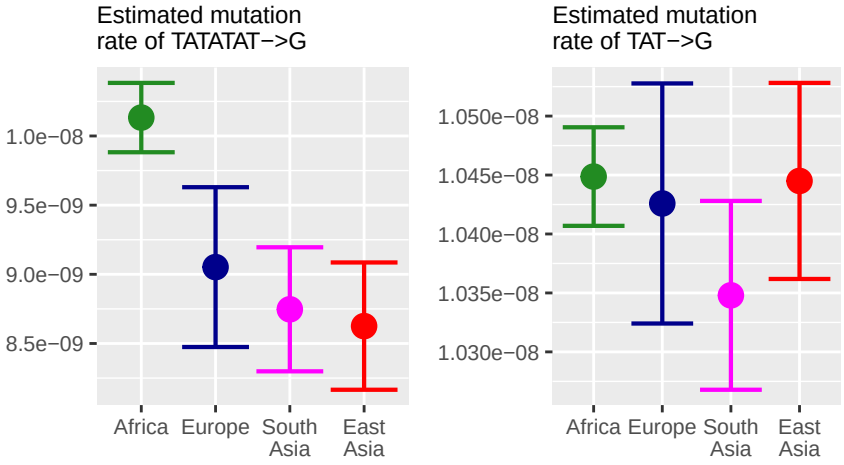
Surprisingly, TAA->G has an altogether different pattern than TAA->T. Again, we see that the 3mer subcontext is more or less in agreement with the profile of this 7mer expansion.



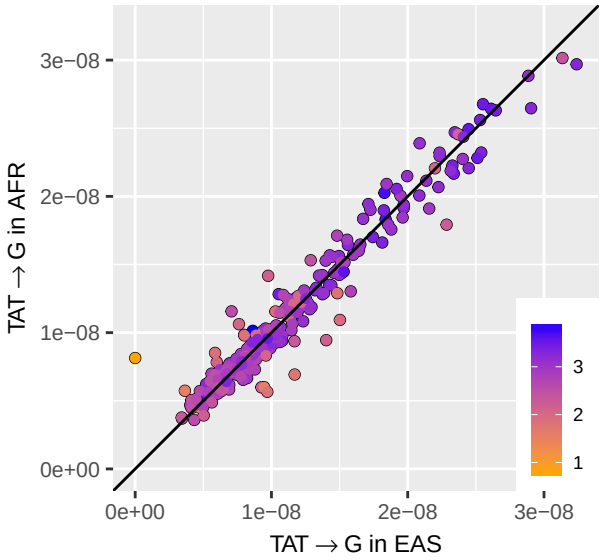


Based on the scatterplot above, it is possible that the profile of TAA->G is actually shaped by a small handful of 7mer outliers.

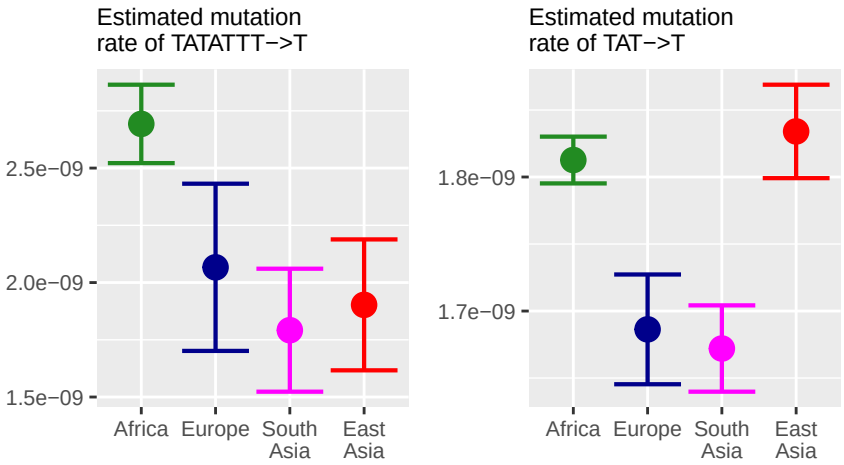
Remaining TAT signals

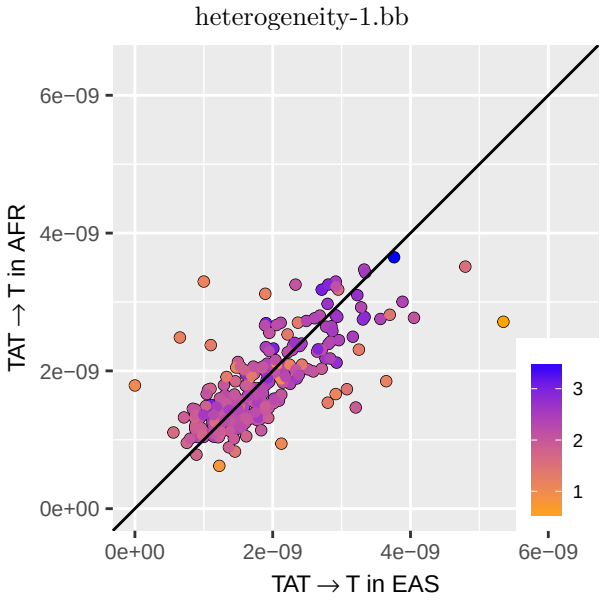


heterogeneity-1.bb



To be honest, I'm not really sure what the correct interpretation of this is.




heterogeneity-1.bb
6e−09
4e−09
2e−09
0e+00
TAT → T in AFR
0e+00
2e−09
4e−09
6e−09
TAT → T in EAS
3
2
1