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 most highly significant 3mers:

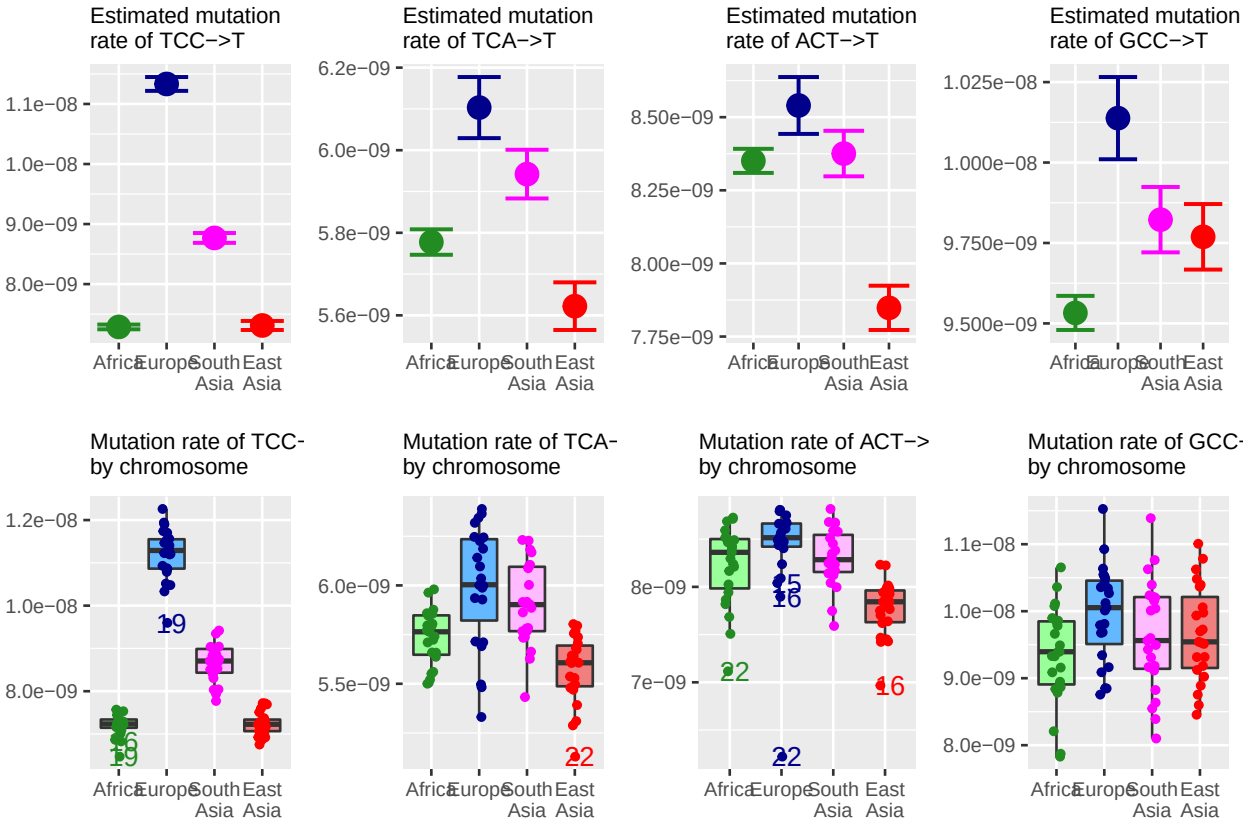
Table 1: 15 most highly significant 3mers

Context	X1mer	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
TCC->T	C->T	129676	37083	36252	43951	0.000000e+00
ACC->T	C->T	135088	29880	35002	40890	3.000000e-308
TCT->T	C->T	144254	30975	39522	43187	2.868119e-196
GAT->T	A->T	47152	9181	15883	15054	5.361344e-111
ACC->A	C->A	76646	14684	24481	23819	1.248600e-98
CCC->T	C->T	133143	26025	35129	38511	2.154370e-69
ACA->T	C->T	227803	39876	58806	62287	2.904922e-60
TCA->T	C->T	134514	26126	36472	38964	9.743309e-52
ACT->T	C->T	157475	29610	41237	44484	3.142140e-51
TCG->T	C->T	202752	37854	58472	60351	1.556530e-49
ACG->T	C->T	331572	61298	95484	97290	2.090019e-48
GCG->T	C->T	230676	42837	66845	67786	3.756211e-45
GCT->T	C->T	136895	25668	40513	40199	1.791085e-40
GAC->T	A->T	21353	4012	7135	6569	6.496890e-40
GCC->T	C->T	123557	24160	35281	35857	1.095624e-37

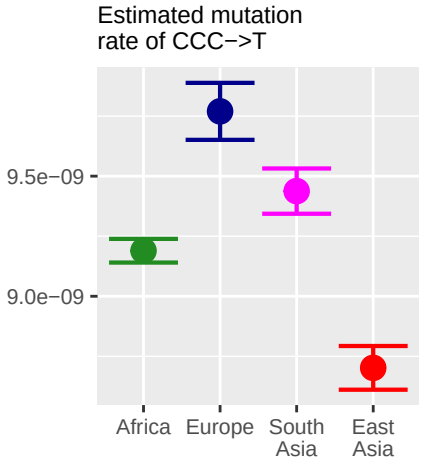
Here, I've chosen to show the top 15 results, which are all significant at $p < 10^{-40}$.

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, GCC->T, 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:



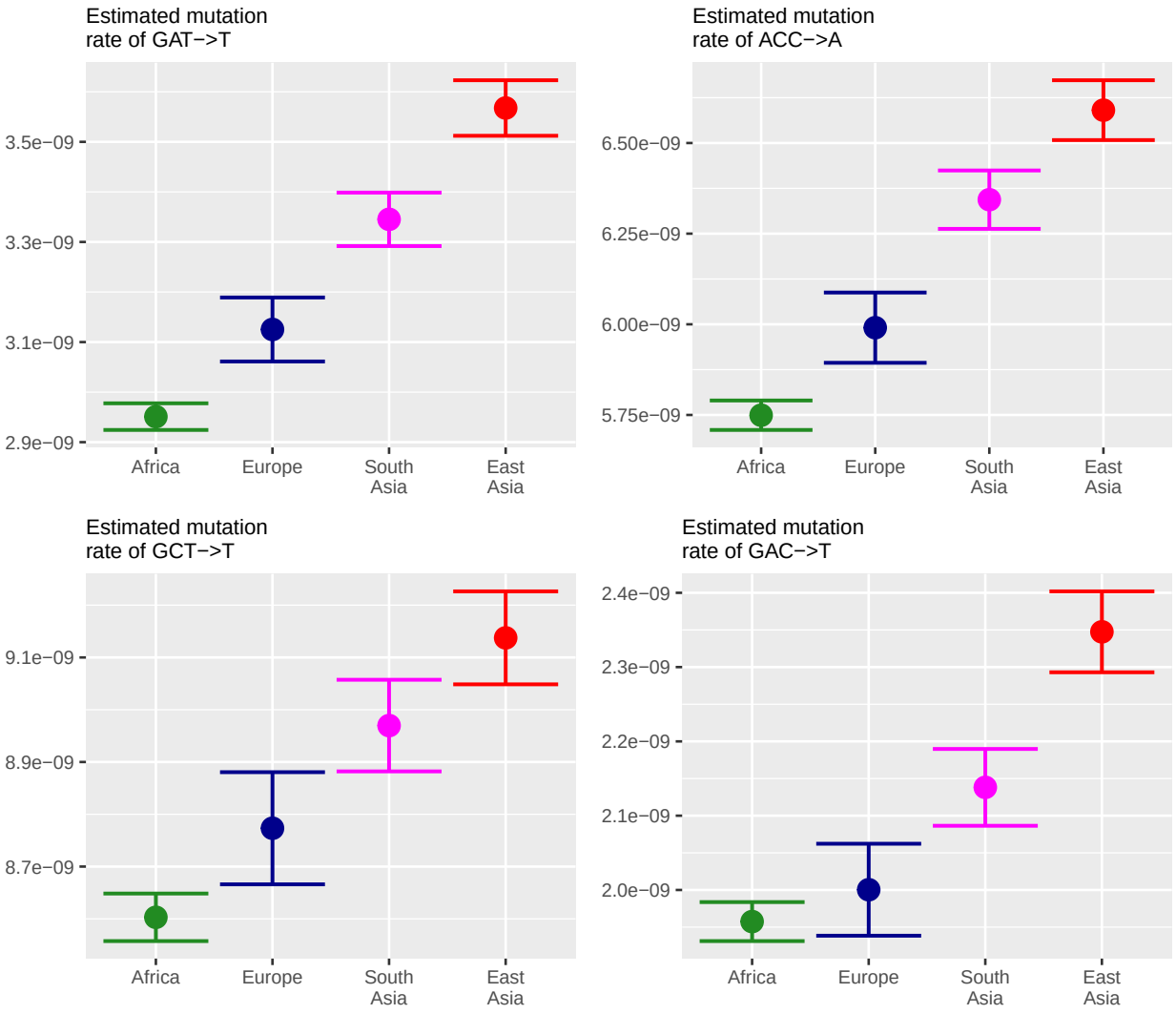
Note that although these patterns appear similar in EUR, EAS, and SAS, there are noteable differences between the relative rates in Africa. This seems to be common among the C->T signal polymorphisms. Consider for example CCC->T:

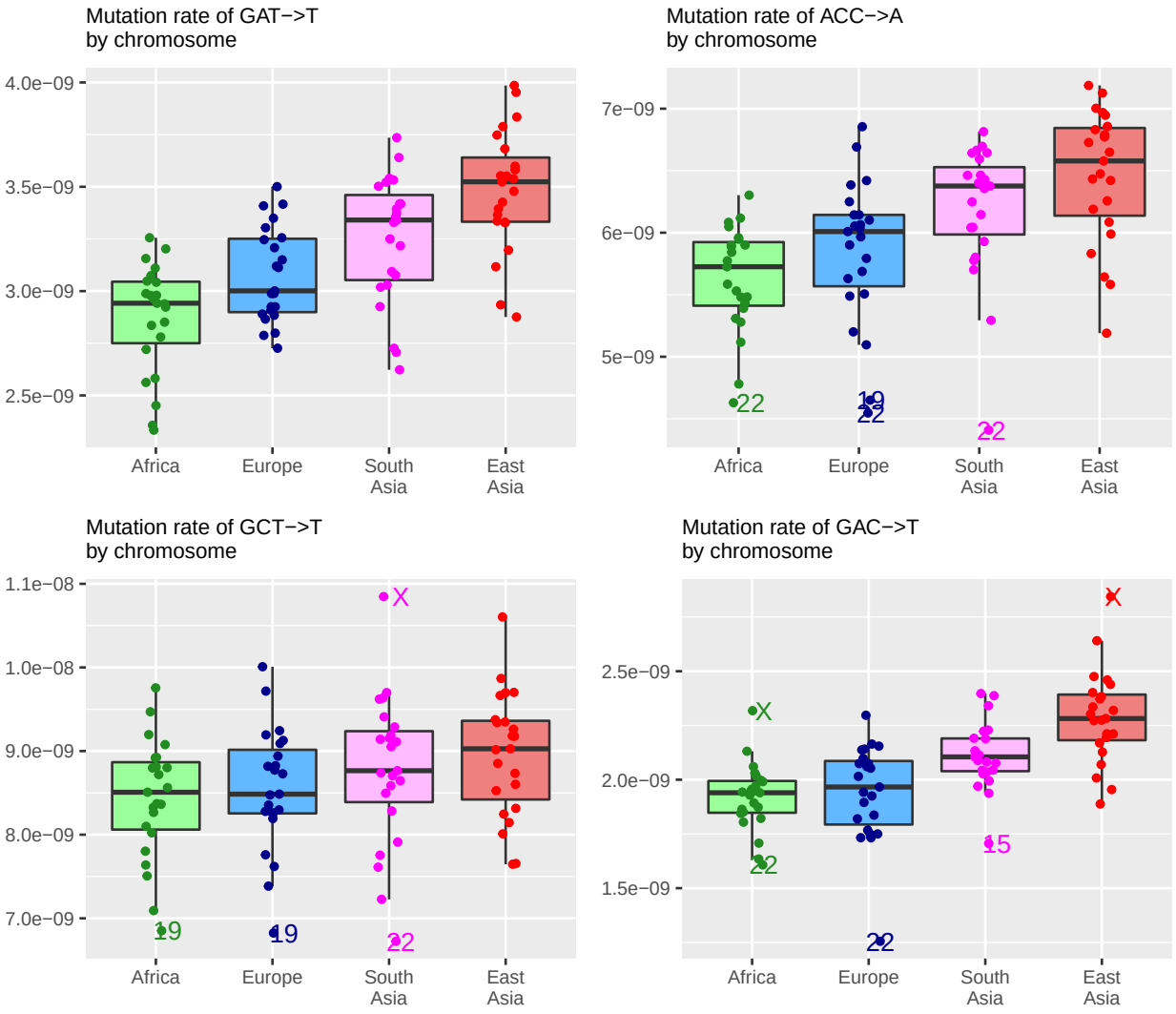


CCCT-1.bb

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 $p < 10^{-95}$. In addition, the 13th and 14th most significant results, GCT->T and GAC->T show a similar profile of enrichment in south and East Asia.

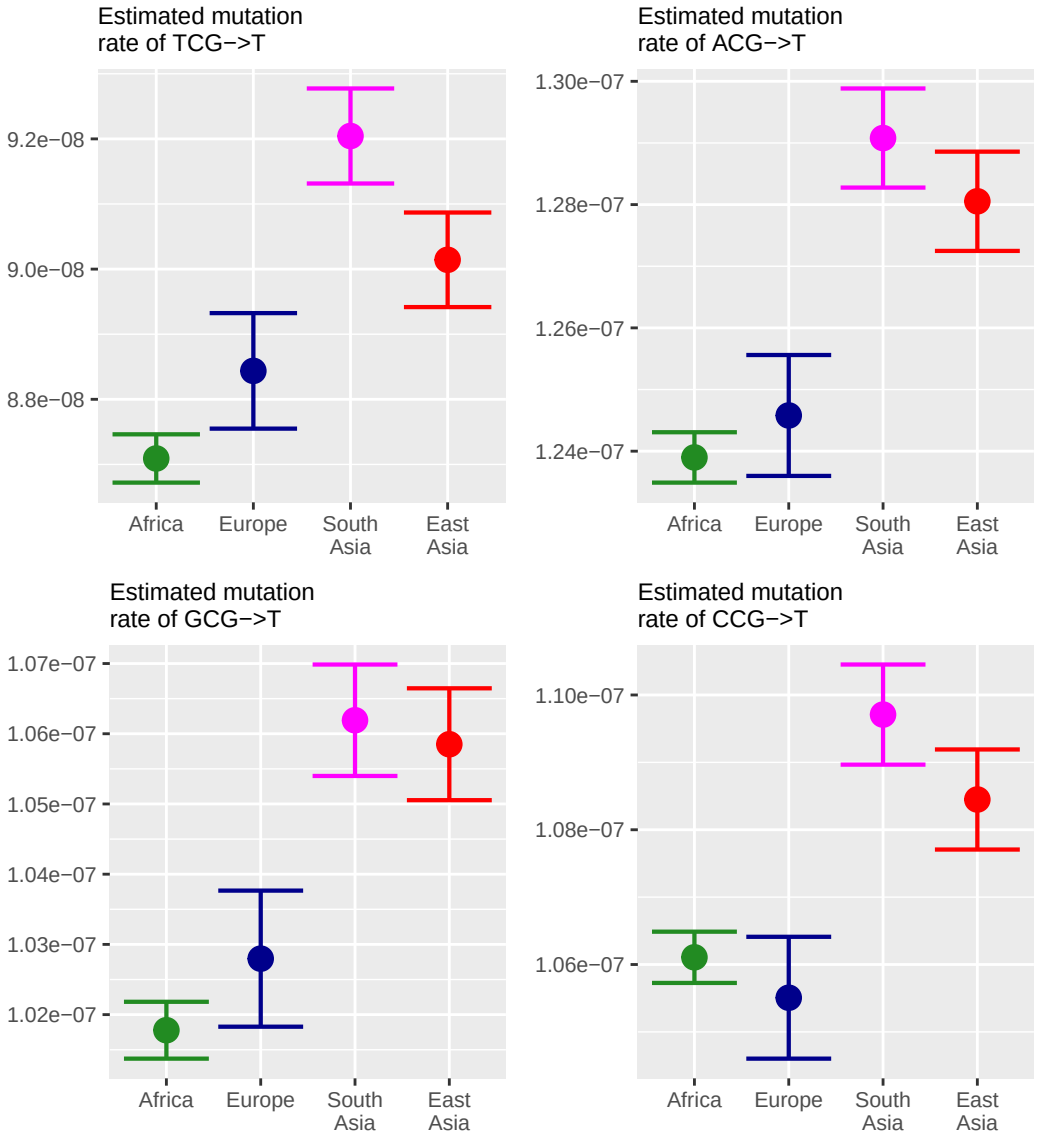




Interestingly, chromosomes 19 and 22 appear as outliers on the lower end, while the X chromosome appears as an outlier on the upper end.

Signal 3: CpG transitions

Three of the four CpG transitions appear among the top results, while the forth ranks 18th ($p < 1E-30$). 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:



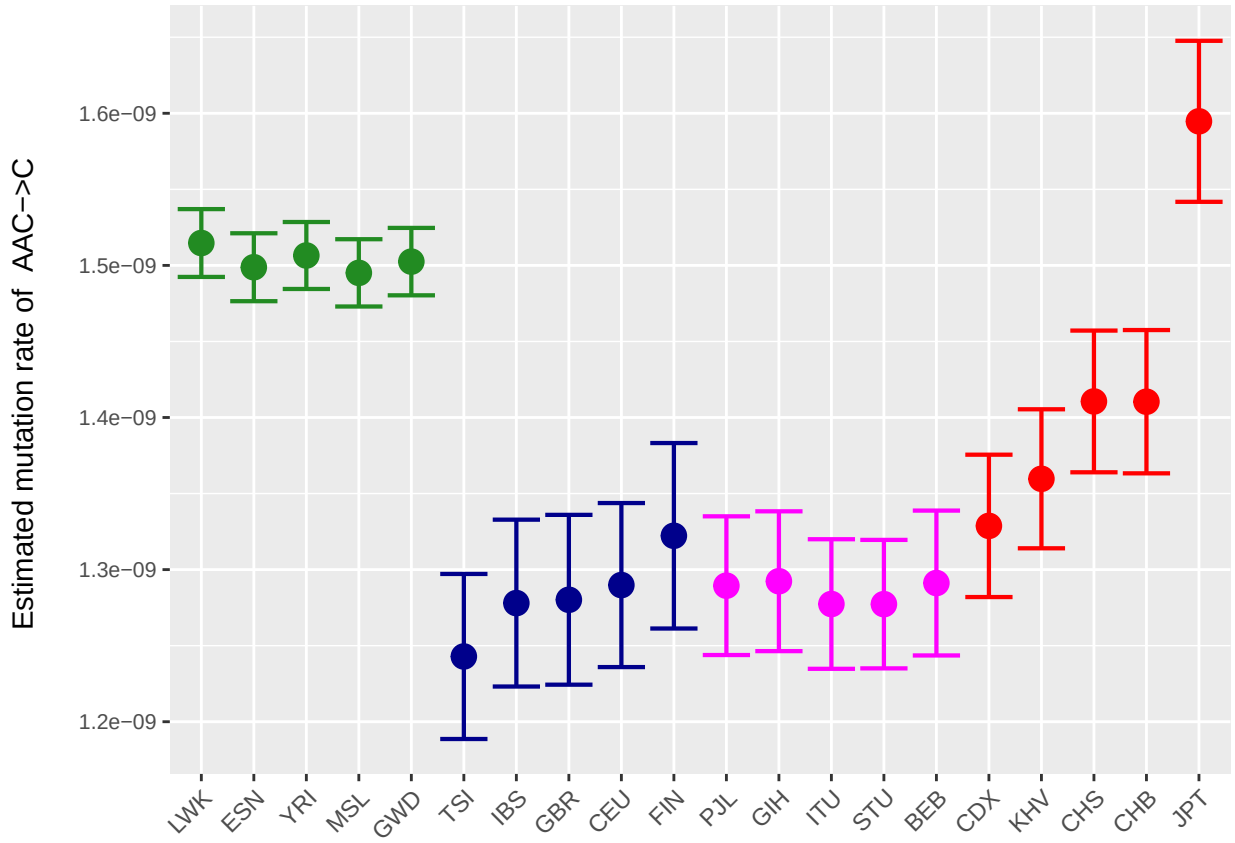
signal3-1.bb

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

Signal 4: Heterogeneity within East Asia

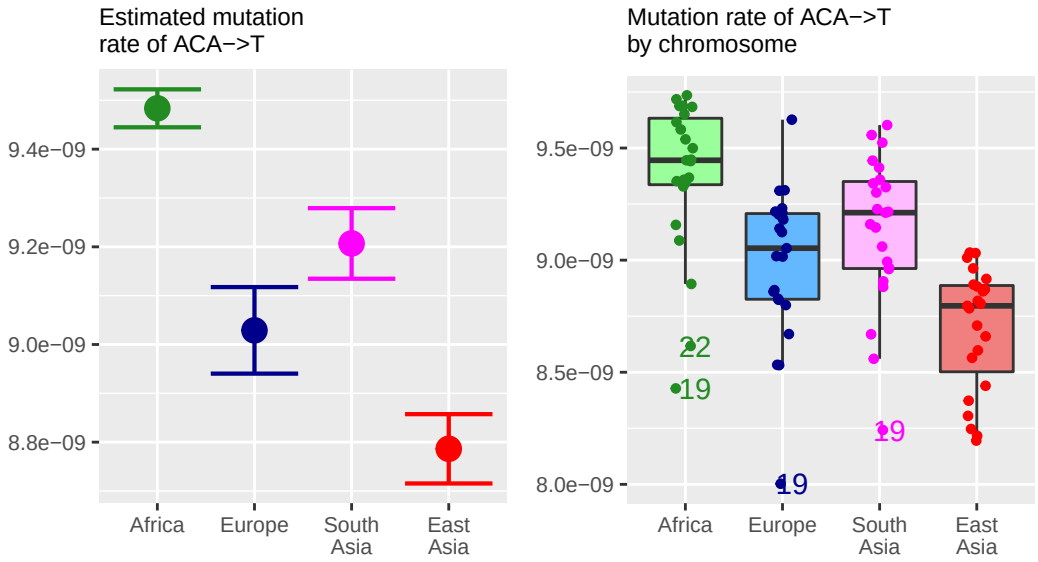
When we construct a heatmap of all 3mers, we find two clusters which appear enriched in Japan and other groups in East Asia. These clusters are comprised of the *AC->C polymorphisms, as well as TAT->T. When the mutational types are clustered using only the data from East Asia, excluding other continental 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

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

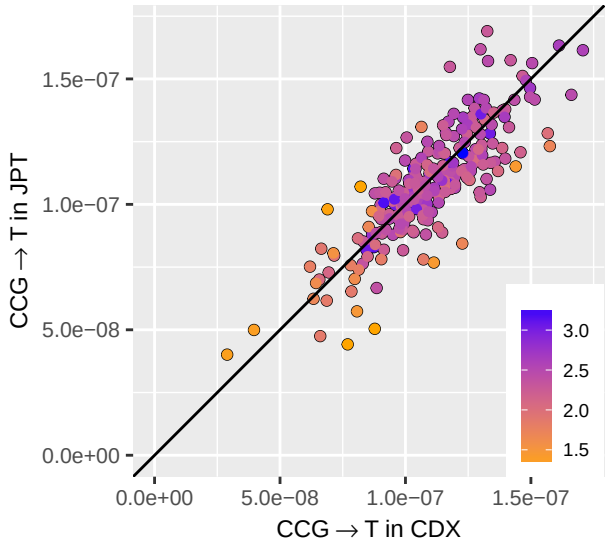


Finally, an additional pattern of interest is the shared profiles of certain CpG transversions, which appear to be enriched in Africa, and which cluster together in heatmaps. However, none of these polymorphism types are significant based on homogeneity tests (predictably, since CpG transversions are 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:



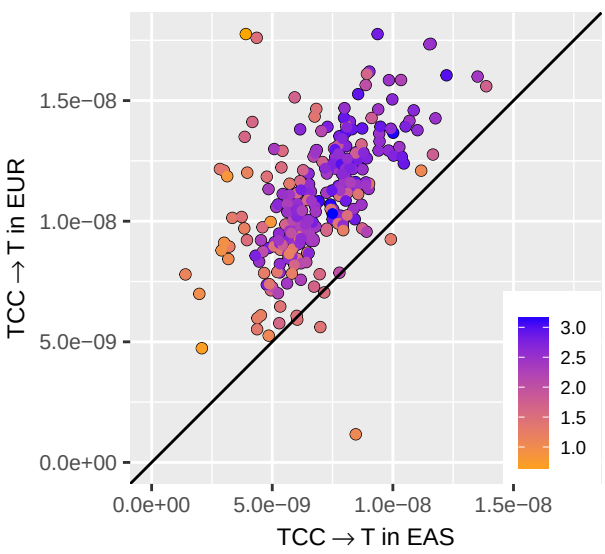
heterogeneity plot-1.bb

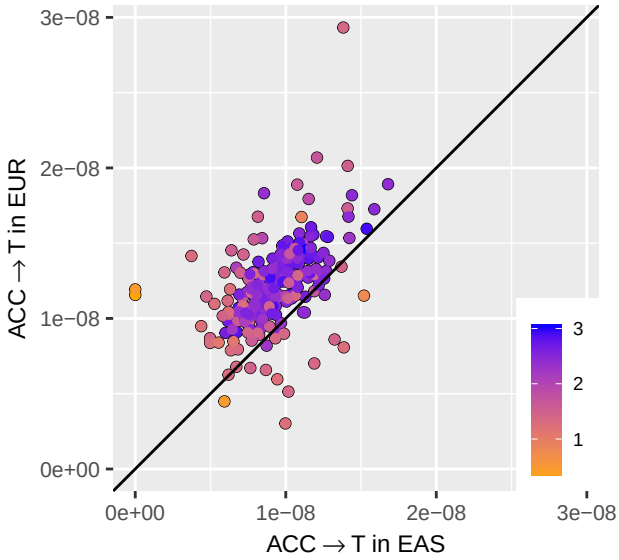
Here, each point represents a polymorphism, and the points are colored by the base 10 log of sample size (number of polymorphisms observed). We see some noise here, mostly among the yellow-colored (more uncertain) polymorphisms. However, most points lie along the y = x line.

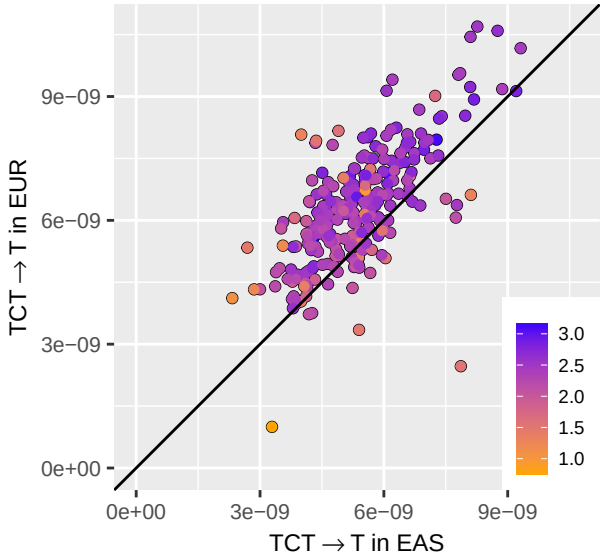
Signal 1: European C->T elevation

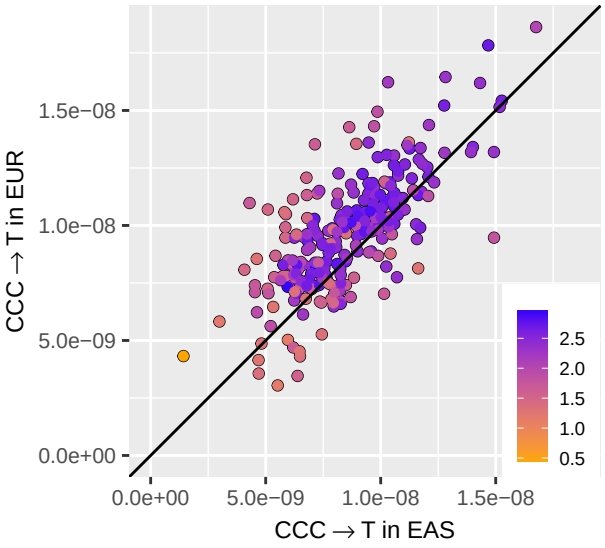
For each signal we've highlighted at the 3mer level, we'd like to know whether this is a true 3mer effect or whether this is driven by broader sequence context.

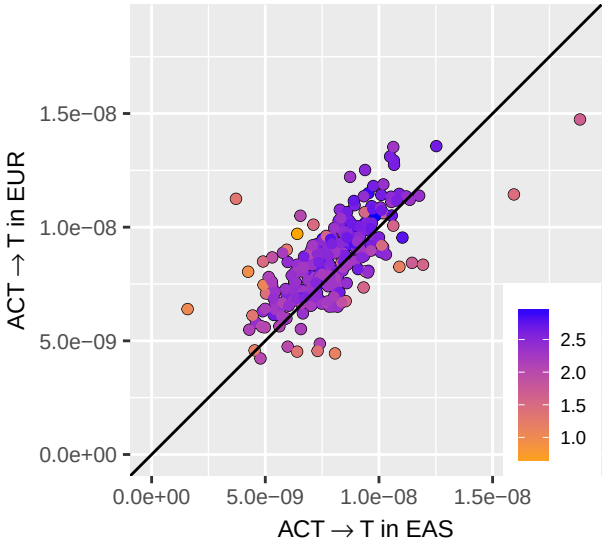
Let's consider the top 3mer polymorphisms for the European C->T elevation, shown on the next page. Here, we can see that, for each polymorphism type, the distribution of all 7mers lies slightly above the y = x line. This indicates that signal 1 is determined by local sequence context effects at 1 or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

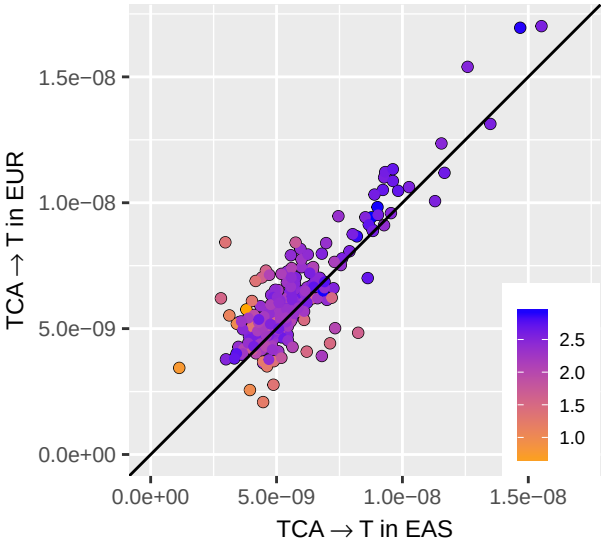

TCC → T in EUR
TCC → T in EAS
0.0e+00
5.0e−09
1.0e−08
1.5e−08
3.0
2.5
2.0
1.5
1.0


ACC → T in EUR
ACC → T in EAS
0e+00
1e−08
2e−08
3e−08
3
2
1


TCT → T in EUR
TCT → T in EAS
0e+00
3e−09
6e−09
9e−09
3.0
2.5
2.0
1.5
1.0

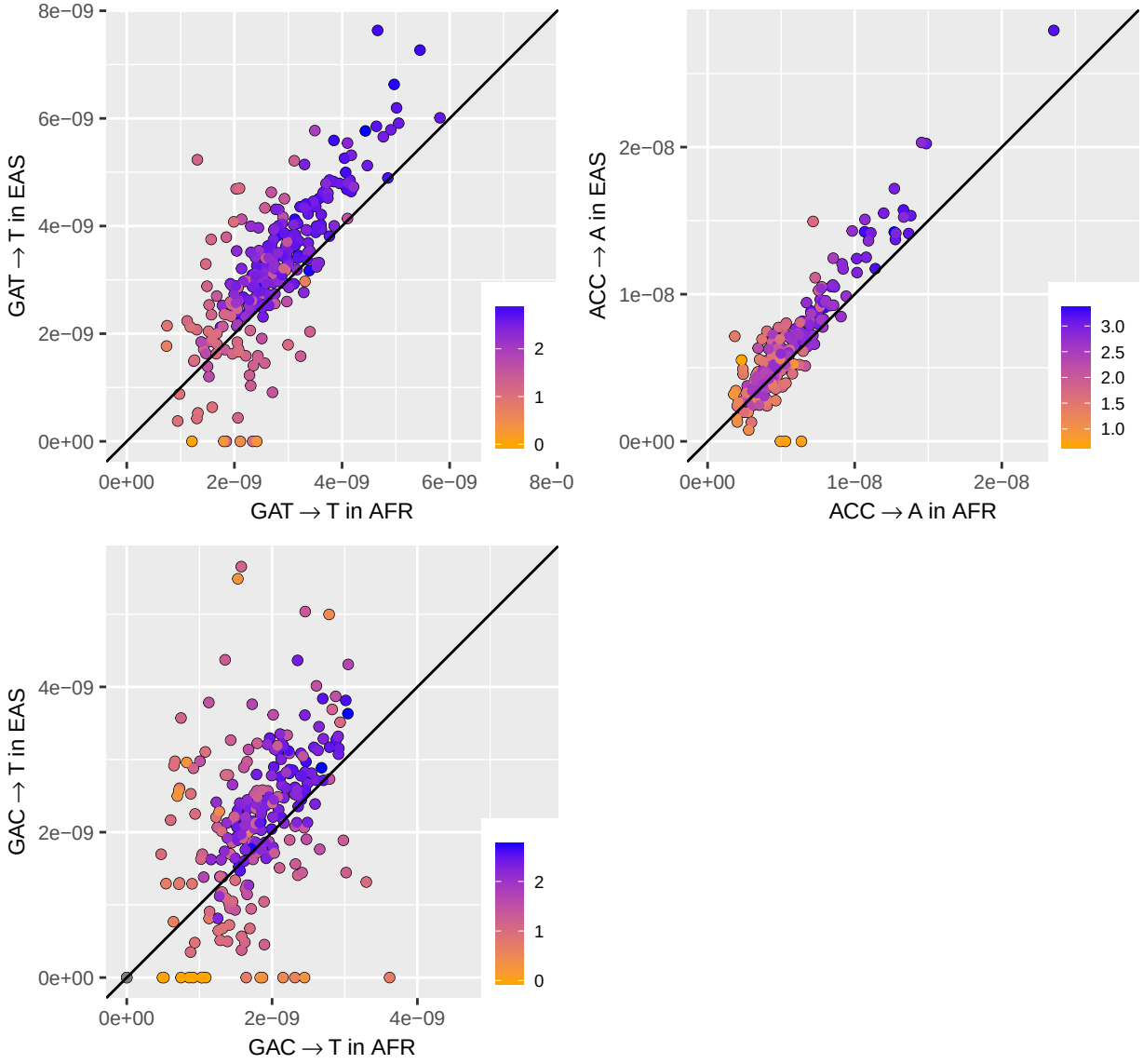

CCC → T in EUR
CCC → T in EAS
0.0e+00
5.0e−09
1.0e−08
1.5e−08
2.5
2.0
1.5
1.0
0.5


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


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

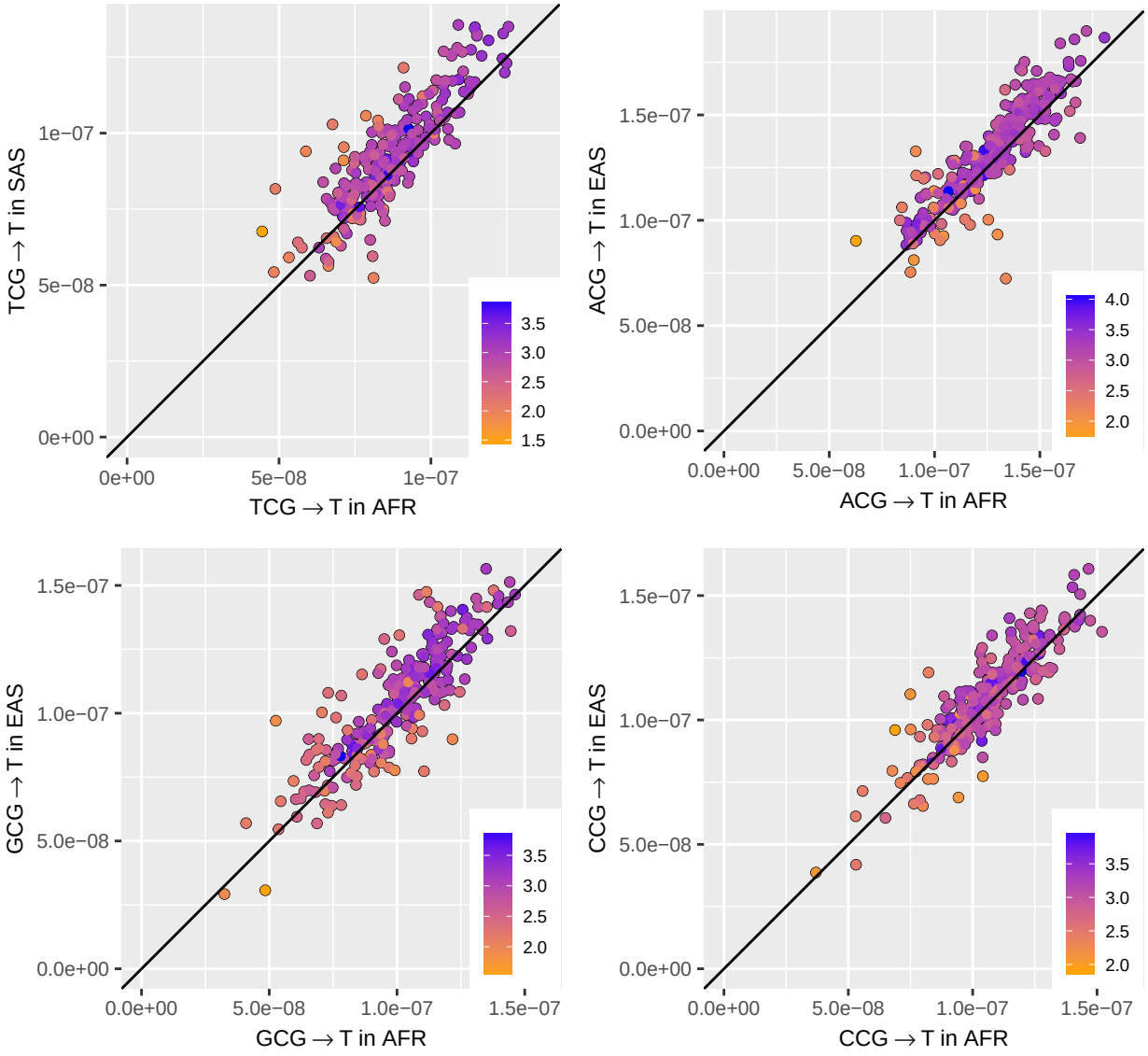
Signal 2: Enrichment of certain polymorphisms in East Asia

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
AAGACAG->C	0.0003943	0.0277929
AATACAG->C	0.0003943	0.0277929
ACAACAG->C	0.0003970	0.0277929
ACCACCA->C	0.0002571	0.0256823
AGTACAG->C	0.0000000	0.0000001
ATAACAG->C	0.0000093	0.0011780
ATCACAG->C	0.0004174	0.0277929
ATGACAG->C	0.0000103	0.0011780
CAAACCC->C	0.0000000	0.0000000
CCCACAG->C	0.0000097	0.0011780
GTGACAG->C	0.0004664	0.0286643
TCAACAG->C	0.0000045	0.0008978
TTTATTA->T	0.0007946	0.0453474

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

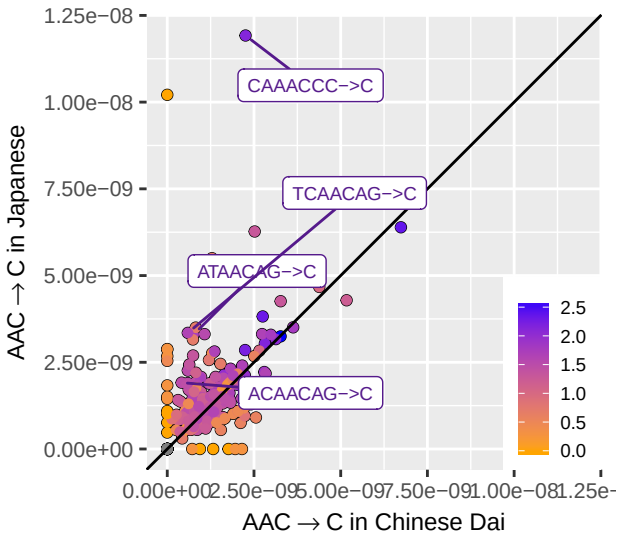
Testing for enrichment on the X chromosome, we find that 4 of 9 of these polymorphisms is significantly enriched on X (see below). The probability of observing this number of significant results by chance alone is:

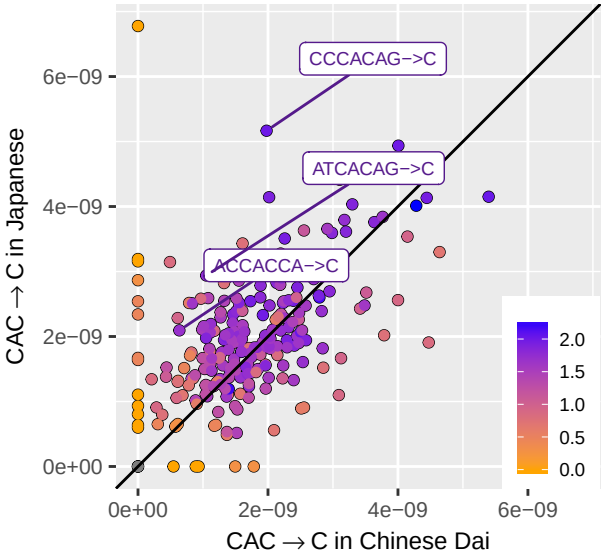
```
binom.test(4,9,0.05)$p.value
```

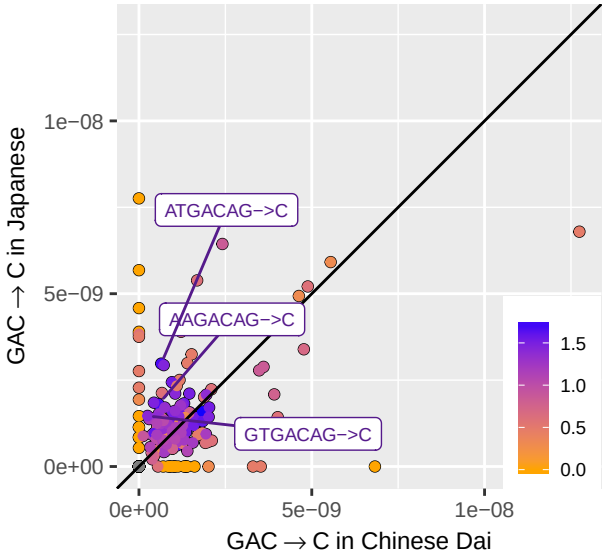
```
## [1] 0.0006425747
```

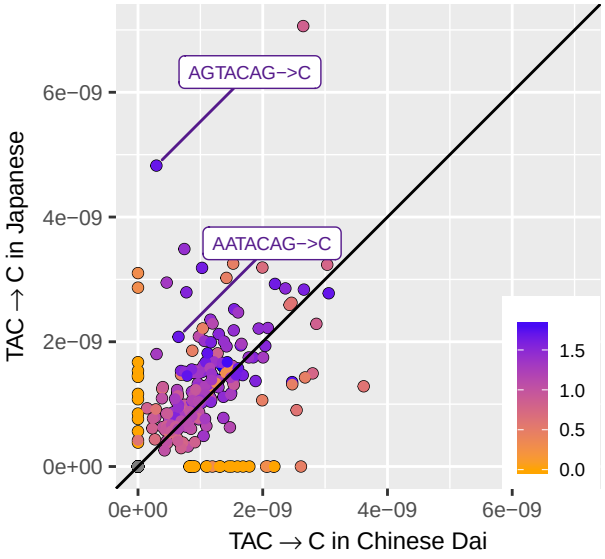
Context	Autosomes	X	Autosomal_sites	X_sites	alpha	p.0	p.MLE	p
TTTATTT->T	743	65	1444601	119969	0.770283	0.000396	0.000542	0.009215
AAGACAG->C	50	4	298423	24142	0.770283	0.000129	0.000166	0.378682
AATACAG->C	44	1	258885	21463	0.770283	0.000131	0.000047	0.939798
ACAACAG->C	34	3	202230	17800	0.770283	0.000130	0.000169	0.405338
ACCACCA->C	48	3	266068	19888	0.770283	0.000139	0.000151	0.521860
AGTACAG->C	51	4	143524	10787	0.770283	0.000274	0.000371	0.342137
ATAACAG->C	50	4	185557	15802	0.770283	0.000208	0.000253	0.415215
ATCACAG->C	57	4	216095	15185	0.770283	0.000203	0.000263	0.371870
ATGACAG->C	39	4	196359	15670	0.770283	0.000153	0.000255	0.220716
CAAACCC->C	101	27	136995	10993	0.770283	0.000568	0.002456	0.000000
CCCACAG->C	80	25	206875	15550	0.770283	0.000298	0.001608	0.000000
GTGACAG->C	24	1	228147	15040	0.770283	0.000081	0.000066	0.704399
TCAACAG->C	35	7	165015	13875	0.770283	0.000163	0.000505	0.008691

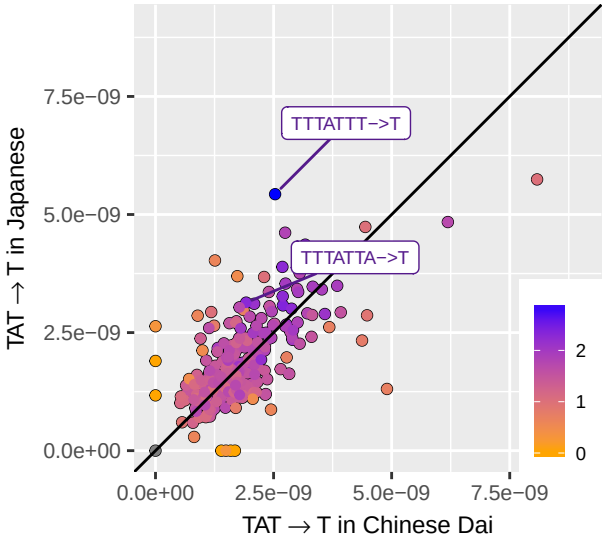
Context	Autosomes	X	Autosomal_sites	X_sites	alpha	p.0	p.MLE	p
TTTATTA->T	209	15	577465	47331	0.770283	0.000279	0.000317	0.344927


CAAACCC−>C
TCAACAG−>C
ATAACAG−>C
ACAACAG−>C
AAC → C in Japanese
AAC → C in Chinese Dai


CCCACAG−>C
ATCACAG−>C
ACCACCA−>C
CAC → C in Japanese
CAC → C in Chinese Dai

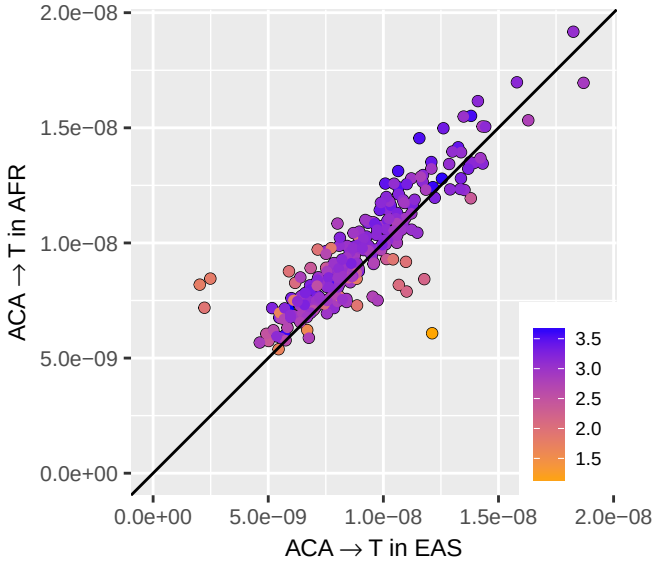

ATGACAG−>C
AAGACAG−>C
GTGACAG−>C
GAC → C in Japanese
GAC → C in Chinese Dai


AGTACAG−>C
AATACAG−>C
TAC → C in Japanese
TAC → C in Chinese Dai


TTTATTT−>T
TTTATTA−>T
TAT → T in Japanese
TAT → T in Chinese Dai

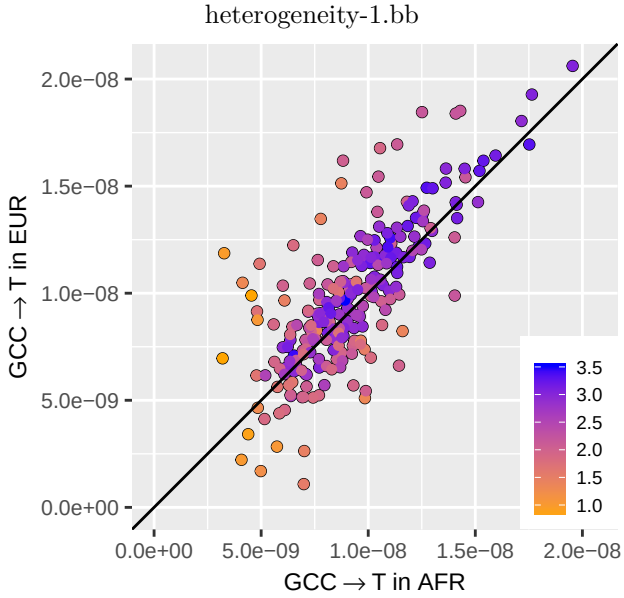
Additional 3mer signals

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

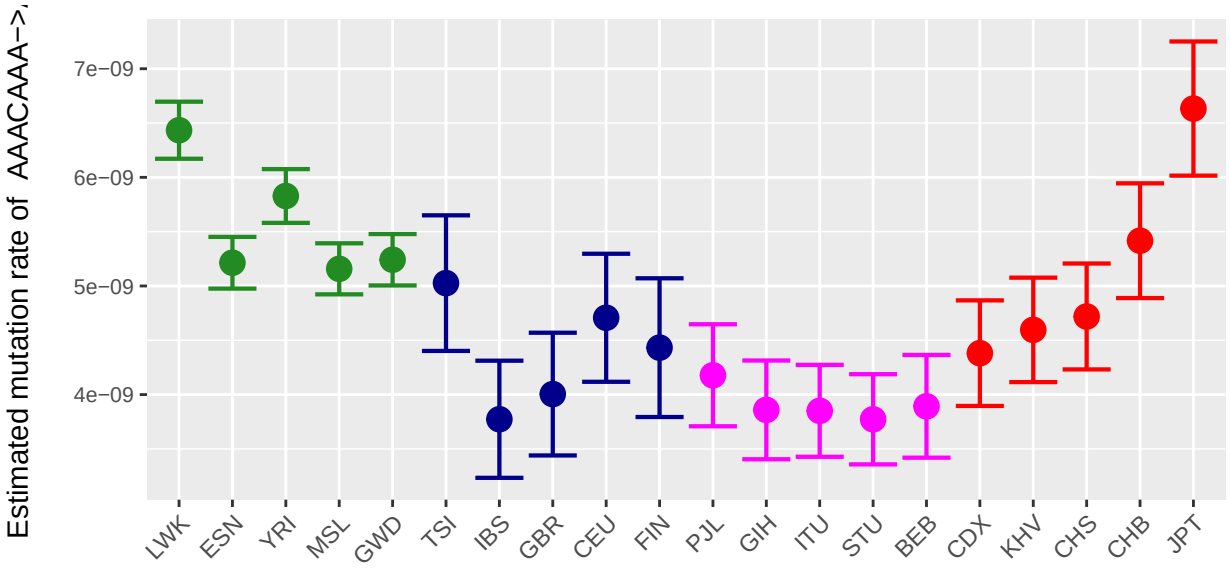
After removing known 3mer signals (TCC->T, ACC->T, TCT->T, ATC->C, ACC->C, and GAT->T), we have the following top significant results for 7mers:

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	127	22	128	12	2.984752e-39
TTTATTT->T	TAT->T	2796	431	808	478	2.166804e-25
TTTAAAA->T	TAA->T	12011	1961	2939	2846	1.199912e-21
ATTAAAA->T	TAA->T	3773	496	857	808	1.968521e-21
AAACAAA->A	ACA->A	3108	446	766	578	2.110224e-21
AGTACAG->C	TAC->C	51	14	55	9	2.375565e-15
ACTAAAA->G	TAA->G	2187	513	833	705	2.887438e-15
CTGCATA->G	GCA->G	72	19	63	12	7.903406e-14
TATATAT->G	TAT->G	7093	1181	1710	1724	3.338030e-11
AGGCTTT->T	GCT->T	1174	177	442	339	4.507439e-09

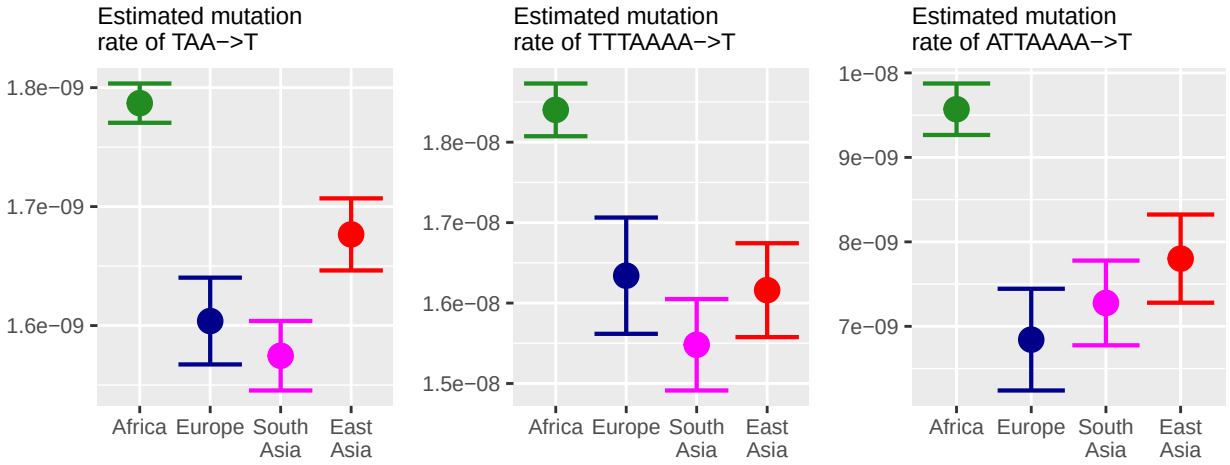
Additional polymorphisms heterogeneous in East Asia

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.

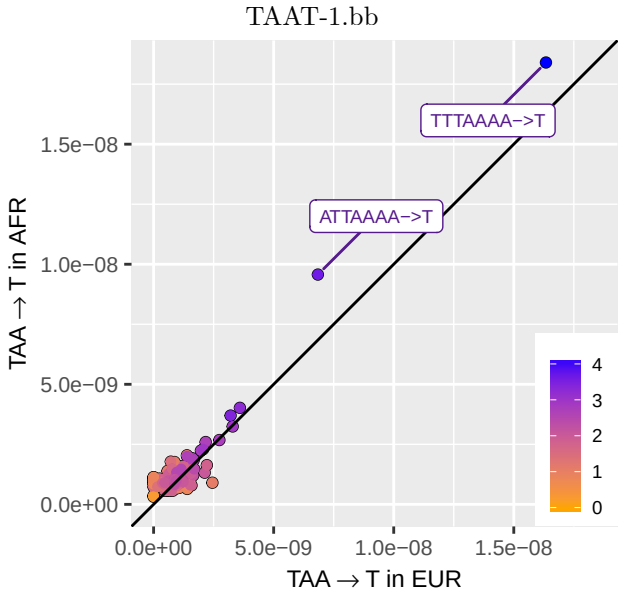


TAA->T

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.

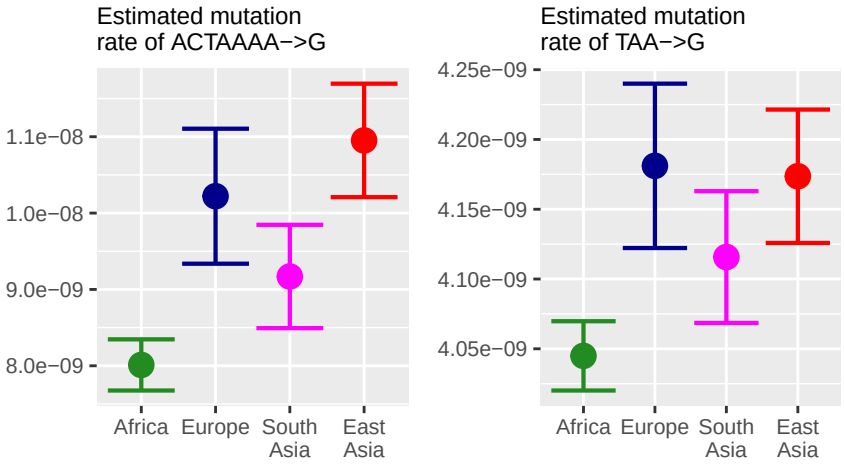


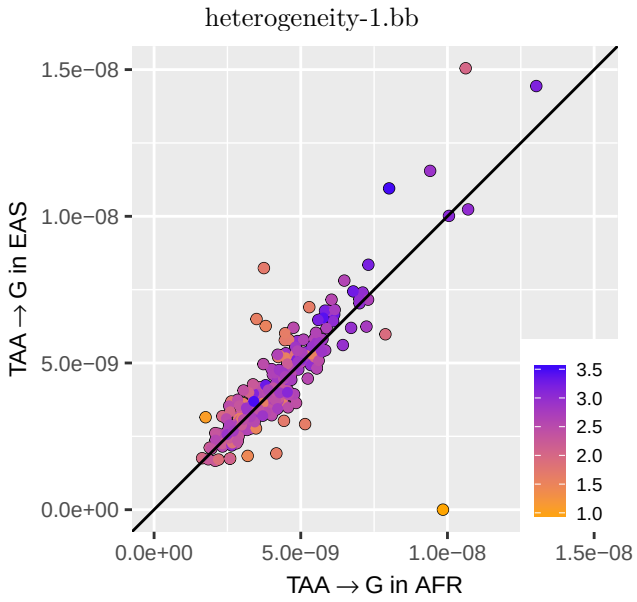
For the most part, the profiles of TTTAAAA->T and ATTAAAA->T match that of the broader 3mer subcontext. However, upon closer examination we find that TTTAAAA->T may be outlier contexts. For most other TAA->T exansions, in fact, the rates in Africa are in agreement with the rates in Europe. Only these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



TAA->G

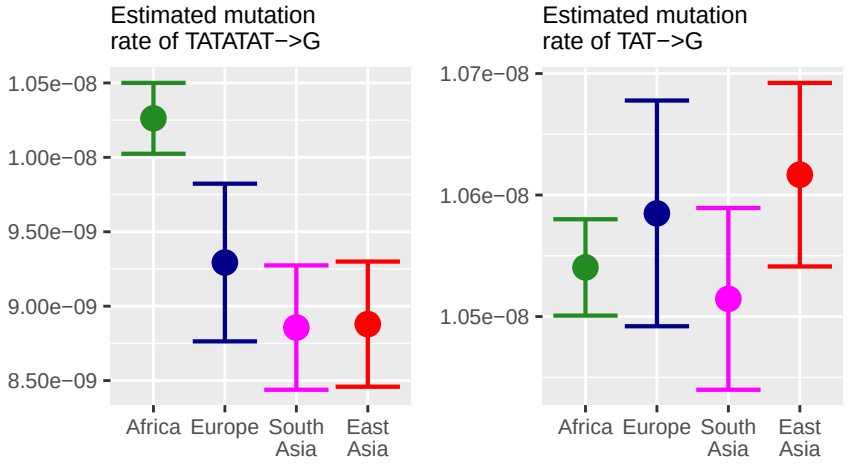
Surprisingly, TAA->G has an altogether different pattern than TAA->T. Again, we see that the 3mer subcontext is more or less in agreement with the profile of this 7mer expansion.

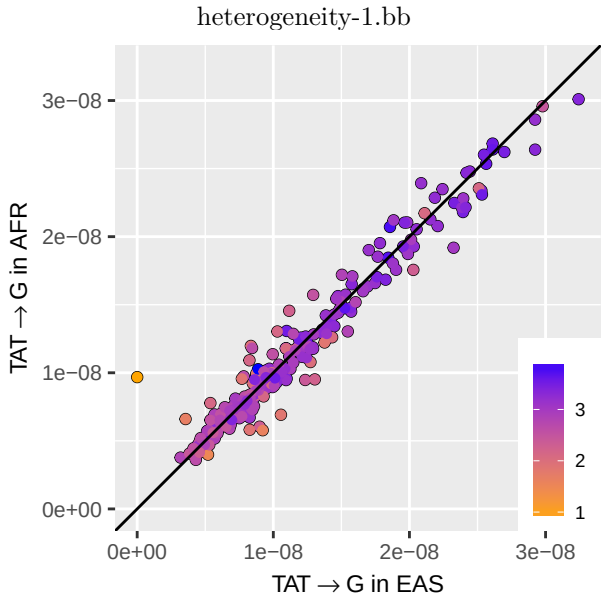




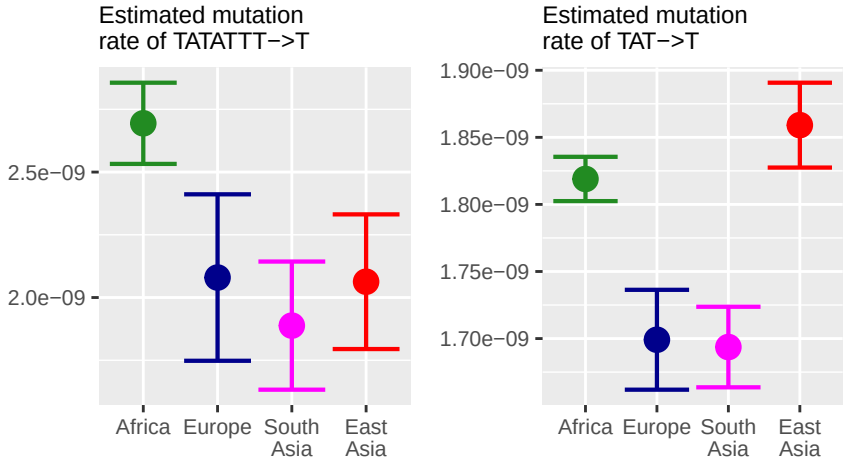
Based on the scatterplot above, it is possible that the profile of TAA->G is actually shaped by a small handful of 7mer outliers.

Remaining TAT signals





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



heterogeneity-1.bb