Heatmaps

Rachael 'Rocky' Aikens, Voight Lab

June 19, 2017

A previously, when we have considered a kmer sequence context model, we have considered each of the $\frac{3(4^k)}{2}$ possible sequence contexts separately. In reality, however, it is likely that the mechinisms of mutation act similarly over certain types of contexts - for example, by recognizing certain shared motifs - so that those polymorphism types appear to have similar proportions in the data. Likewise, polymorphisms that have similar profiles across the globe are likely to be affected by the same mechanisms. For this reason, it is valuable to try and identify mutational 'signals' - groups of mutations which follow the same global patterns and are thus likely to be driven my the same mechanism.

Set-Up

To begin, we define the 'profile' of a given polymorphism to be its mutation rate inferred from each of the 20 nonadmixed subpopulation samples from the phase III 1,000 genomes release. The heatmaps here are constructed from normalizations of matricies which list the global rate profiles for each kmer mutation type. These dataframes of of rate profiles for the 3mer, 5mer, and 7mer models can be found in the "data" subdirectory.

Methodology

Normalizing the data

It is necessary to normalize the rows (mutation types) of the data or mutations will just cluster by bulk mutation rate instead of subtle differences in rate between populations. There are three ways I've written to do this:

- **L1:** Set the rates of any mutation m across the populations so that the total rate is 1. This is the method Segert and I used while developing this analysis.
- **Z:** Set the rates of mutation m to have mean 0 and variance 1 across all populations in the dataset. NOTE: This method *does not work* if you log-transform the data because you will naturally have negative values in your dataset.
- **fdiff** Normalize so that the data for mutation m is fold difference relative to the mean rate for mutation m over all pops. This is the reccommended method, which works well with a log transformation.

```
# helper function that normalizes a vector by z or L1 method
norm <- function(vec, method = c("z", "L1", "fdiff")){

  if (method == "L1"){ # normalize so vector sums to 1
    u <- vec/sum(abs(vec))}
  else if (method == "z"){ # normalize to mean 0, variance 1
    u <- (vec - mean(vec))/sd(vec)}
  else # normalize to fold change compared to mean
    u <- vec/mean(vec)
  return(u)
}

# normalizes a whole dataset by calling norm on each row
norm.byrow <- function(mat, m){
  data <- t(apply(mat,1,norm, method = m))
  return(data)
}
```

For the remainder of these analyses, we will use rate profiles which are normalized by fold difference.

Heatmap generators

This section builds some helper methods that I use to make heatmaps. Each method minimally requires a matrix of rates across populations (data), and a boolean (logunits) for whether a log transform should be used. As a default, logunits is set to true.

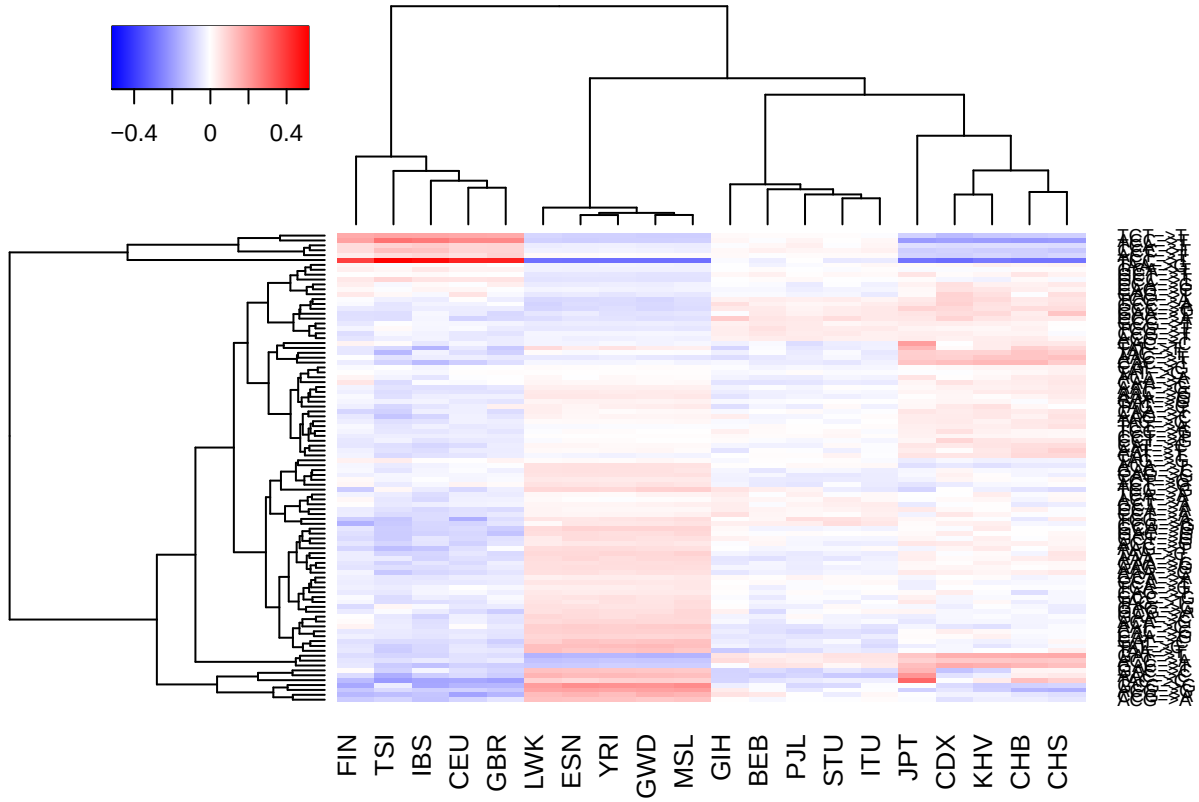
- **make.heatmap** makes a heatmap of the matrix passed in, and returns the row dendrogram (i.e. polymorphism clustering) produced in the process. A handy workhorse and the helper for the following functions.
- **subcontext.heatmap** makes a heatmap of all of the contexts with a 3mer or 5mer subcontext listed in "muts" (a vector of strings). Requires a reference dataframe (labels). NOTE: all polymorphism types in "muts" must be the same kmer type.
- **get.3mer.subcontext.data** is a helper function for subcontext.heatmap that finds the subset of the data with 3mer subcontexts in "muts", filters out invalid values (0 and na), and returns it.
- **get.5mer.subcontext.data** has the same functionality as get.3mer.subcontext.data, but works when muts is a vector of 5mer polymorphism types.

Mutational signals acting at the 3mer level

This section provides an overview of some mutational signals that I think are boradly active at a 3mer level

All 3mers

The most basic heatmap we can imagine making first is a clustering of all 3mer polymorphism types. This can show us some groups of mutations which appear to be acting together at the 3mer level. You can see a couple separate blocks of signal below:



There are a couple observations to make about this plot:

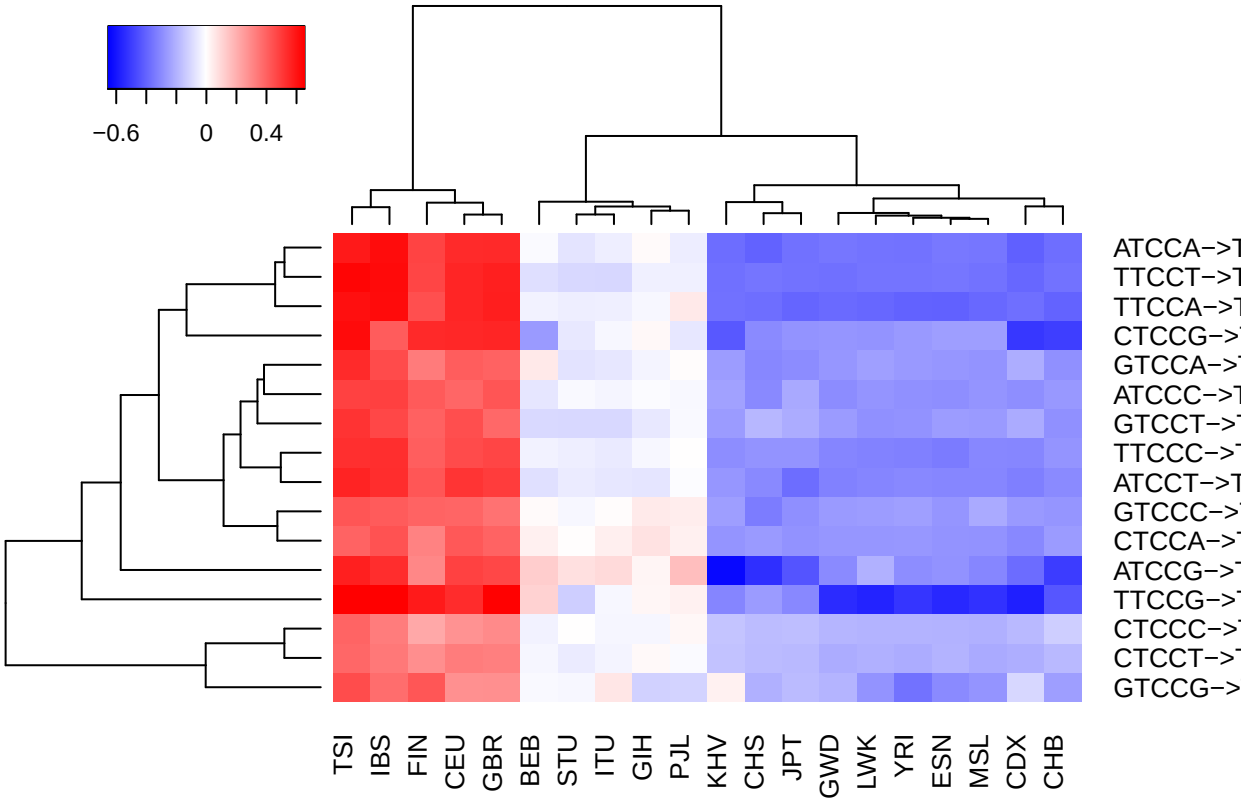
1. TCC->T and the rest of the EUR-elevated C->T signal separates out from the other mutations first
2. The second block of 3mers appears enriched in AFR compared to the rest of the continents. Among them are three *AC->C mutations (AAC->A, GAC->C, and GTA->G) which also appear enriched in JPT and certain other East Asian subpopulations. The other mutations in this block are all CpG transversions. Harris and Pritchard 2017 note that the proportions of these mutation types appear not to agree between 1kg and SGDP, suggesting that this may be an artifact (noteably CCG->A and ACG->G in Figure 2A of that report, and - to a lesser extent - ACG->A in Figure S22A).
3. The third block of mutations appears enriched in EAS across the continent. It is comprised of two mutations, ATC->A and ACC->A which were noted as highly significant and genome wide consistent in earlier experiments, in addition to a new mutation type, GAC->T, which shares a similar profile with the first two.

Beyond these first three blocks, the remainder of the data is difficult to interpret.

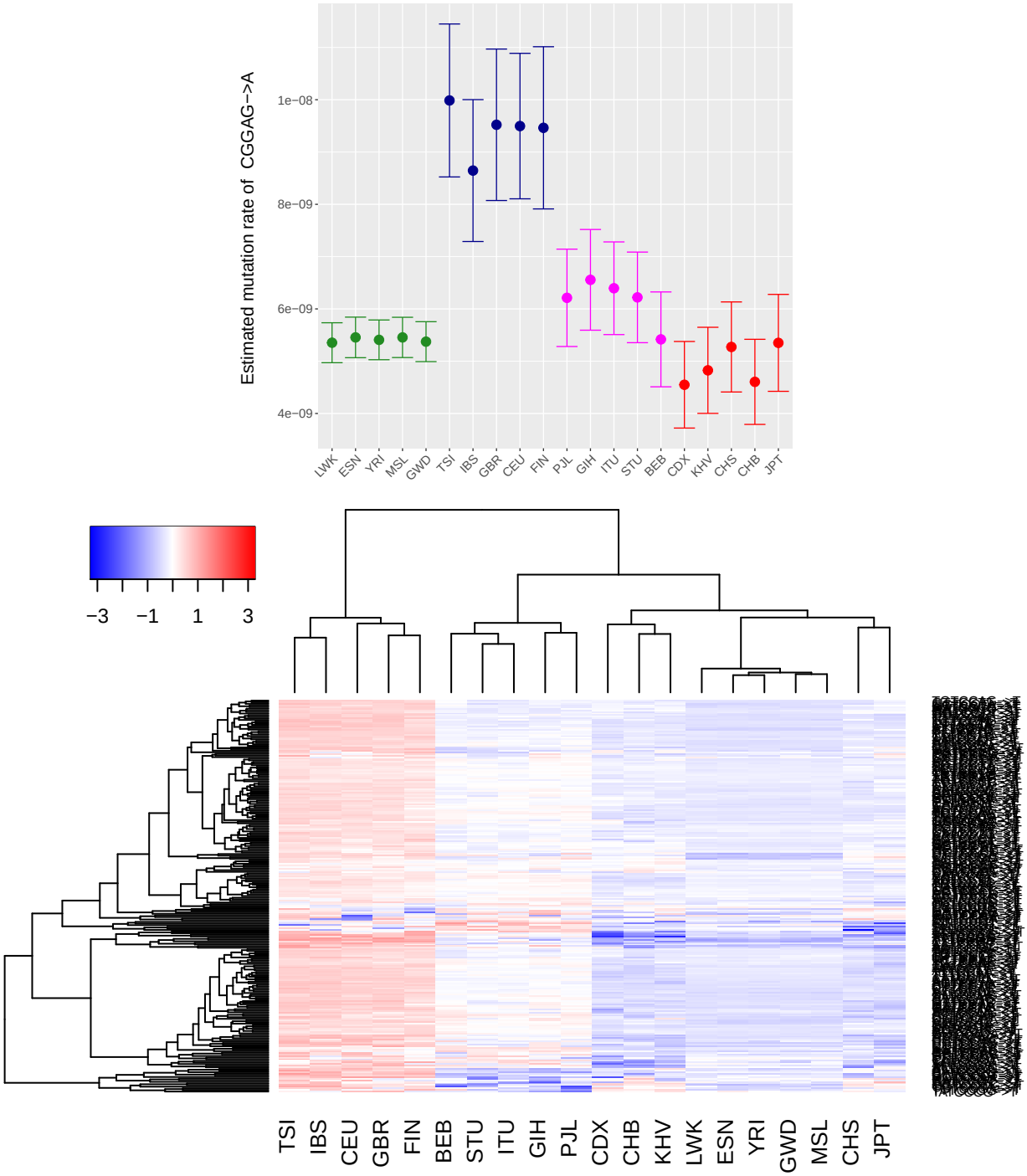
NOTE: heatmapping all 5mers and all 7mers yeilds pretty uninterpretable results (i.e. no clear mutational blocks).

TCC->T 7mers

Since TCC->T appears to have such a strong signature of global rate variation, one might ask what the mutational profiles are for 5mers and 7mers with this 3mer subcontext. Below we show the TCC->T subcontext expanded into 5mers, then 7mers.



This plot is less readily interpretable, since no obvious block structure emerges. The plot of 5mers suggests that CGGAG->A may be depleted in BEB, however this may just be noise (see CI plot of this mutation by ancestral population), and this signal does not appear as an evident pattern among 7mers.

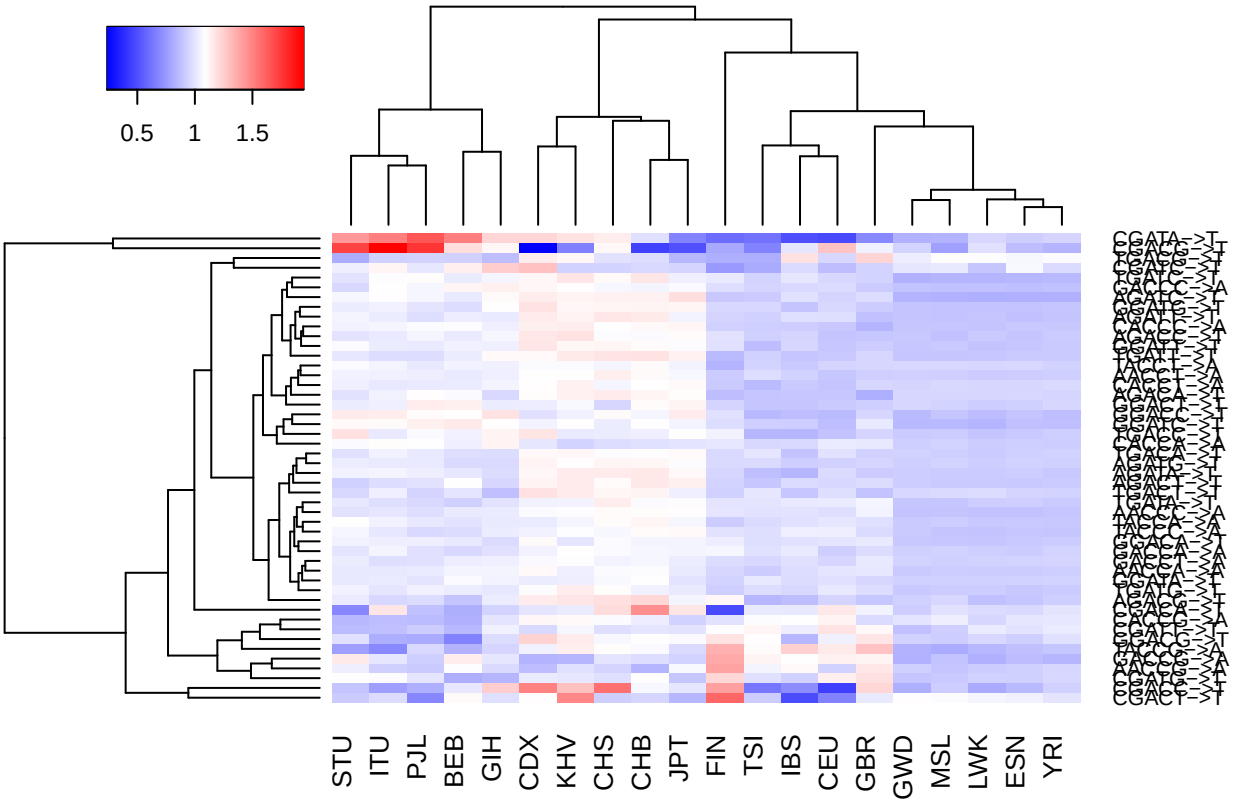


Another block that may be interesting is the first group of 7mers that separates in the second clustering, which appears to be depleted in South Asia.

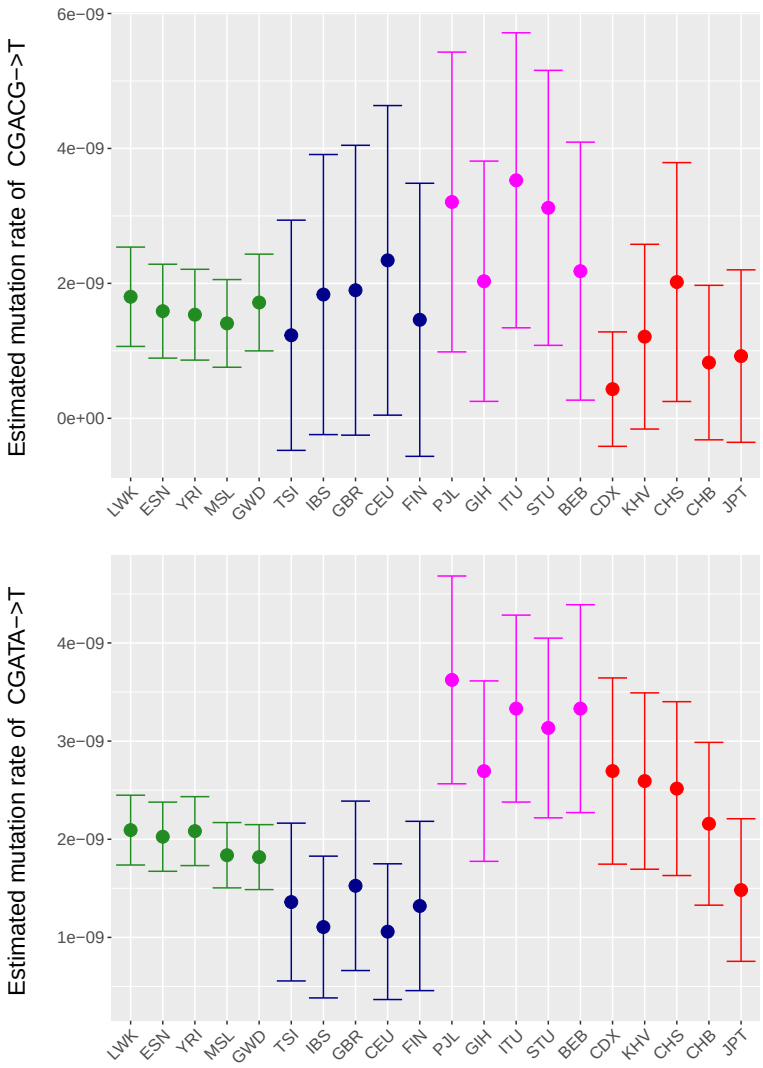
Overall, these plots suggest that the elevation of TCC->T is widespread among all polymorphisms with this 3mer subtype, rather than being driven by a handful of highly variable polymorphism types at the 5mer and 7mer level.

3mer EAS elevation

Since the first heatmap of all 3mers from this section highlighted a group of mutations (ATC->A, ACC->A, and GAC->T), which appear enriched across East Asia, it is reasonable to wonder how 5mers and 7mers with those subtypes vary. The plot below shows all 5mers associated with this 3mer signal. Here I have turned off the log transformation because otherwise there is very little signal to noise.



This plot also does not separate into clear blocks of signals. It would appear that the signal of enrichment in East Asia is spread across nearly all 7mers with these subcontexts, except for a few, which cluster apart first. However, it's hard to discern at a glance which of these signals are due to random noise and which are genuine signals. For example, one of the contexts from the first group, CGACG->T seems simply to be very rare, while the other, CGATA->T, *may* represent a real signal.

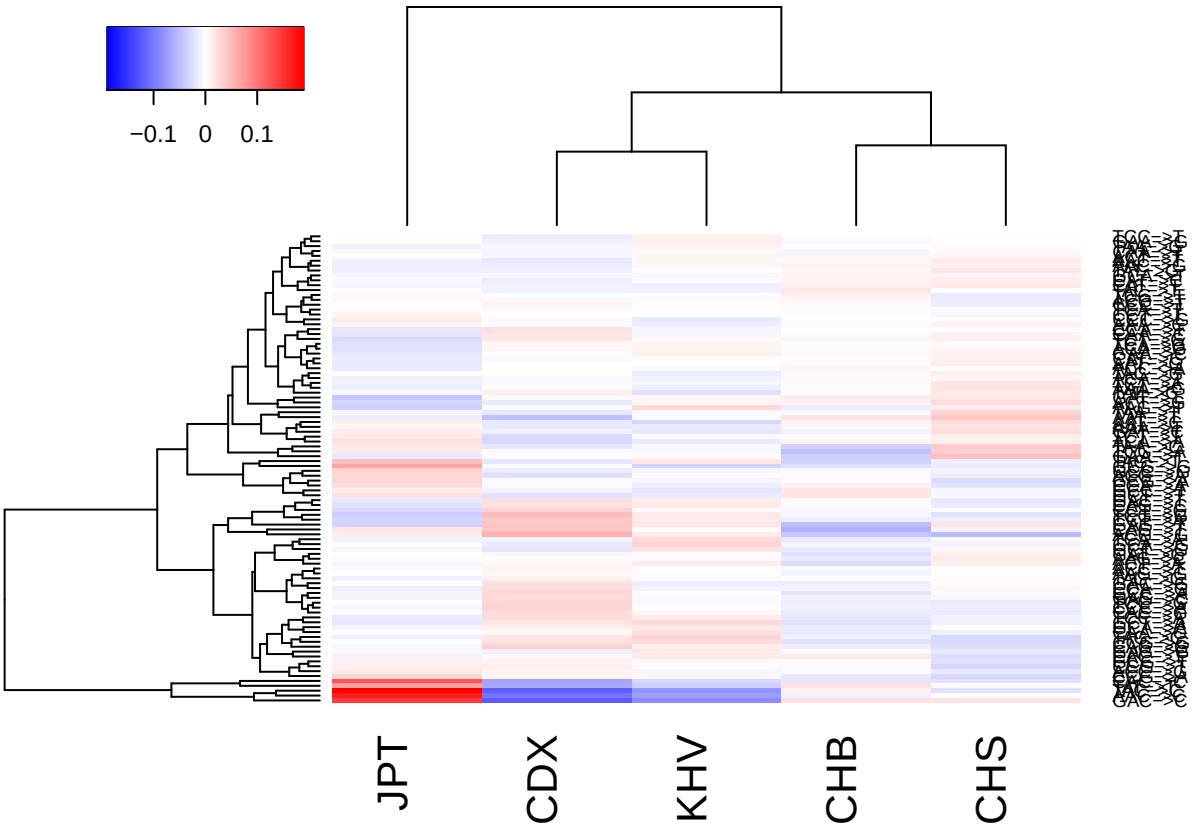

Estimated mutation rate of CGACG−>T
6e−09
4e−09
2e−09
0e+00
LWK ESN YRI MSL GWD TSI IBS GBR CEU FIN PJL GIH ITU STU BEB CDX KHV CHS CHB JPT
Estimated mutation rate of CGATA−>T
4e−09
3e−09
2e−09
1e−09
LWK ESN YRI MSL GWD TSI IBS GBR CEU FIN PJL GIH ITU STU BEB CDX KHV CHS CHB JPT

7mer signal - heterogeneity within East Asia

Since Harris, Pritchard, and ourselves have noted heterogeneity within 3mer types among East Asian subpopulations, we might be interested in better understanding which mutations show this variability. As we will see, it is likely that features at a higher sequence context level are driving these effects.

3mers in EAS

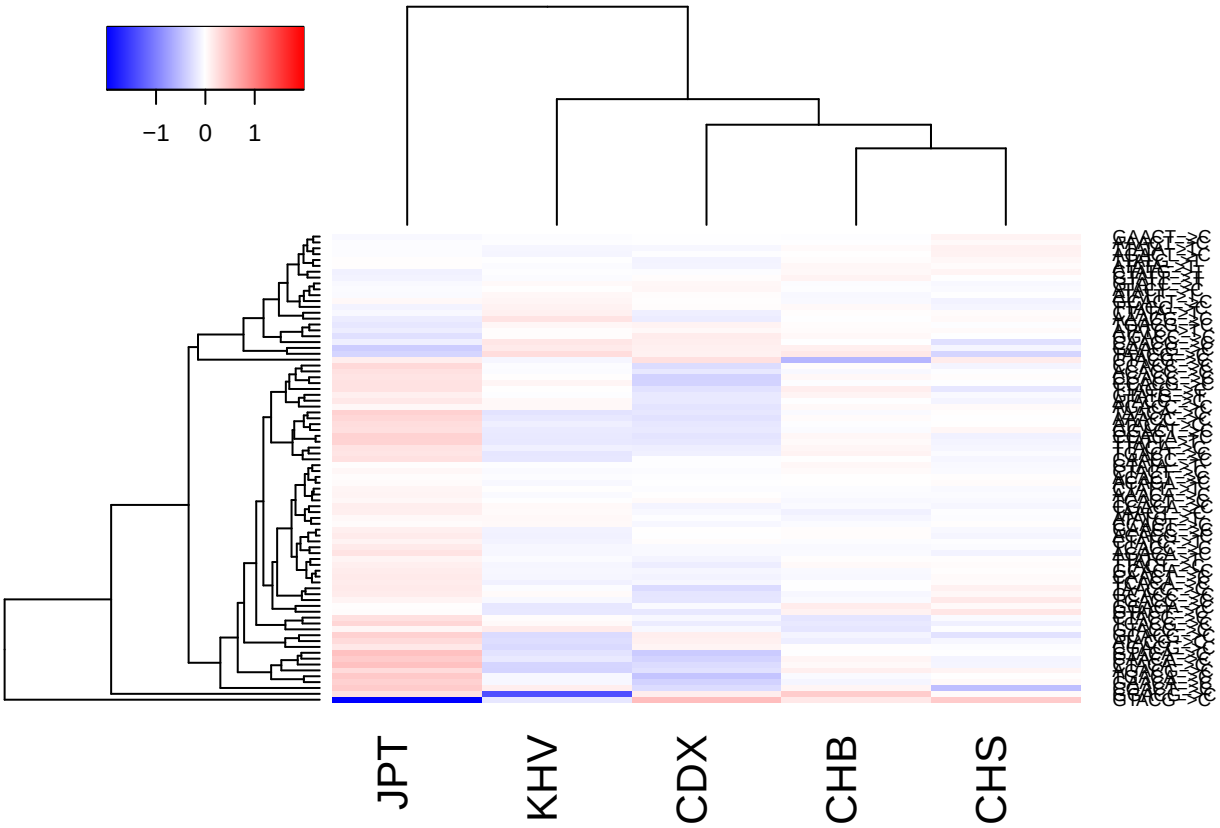
It is simplest to begin with 3mers, since we know that our first heatmap of all 3mers identified a few contexts which appear to vary within East Asia. To try and maximize our sensitivity, the heatmap below performs a similar clustering using only the data from within East Asia.



This clustering sets apart a group of mutations enriched in Japan compared to Vietnamese and Chinese from Xishuangbana. Among these are AAC->C, ATC->G, and GAC->C (the three mutations grouped in the first heatmap, CAC->C (the remaining type of *AC->C mutation), and ATA->A.

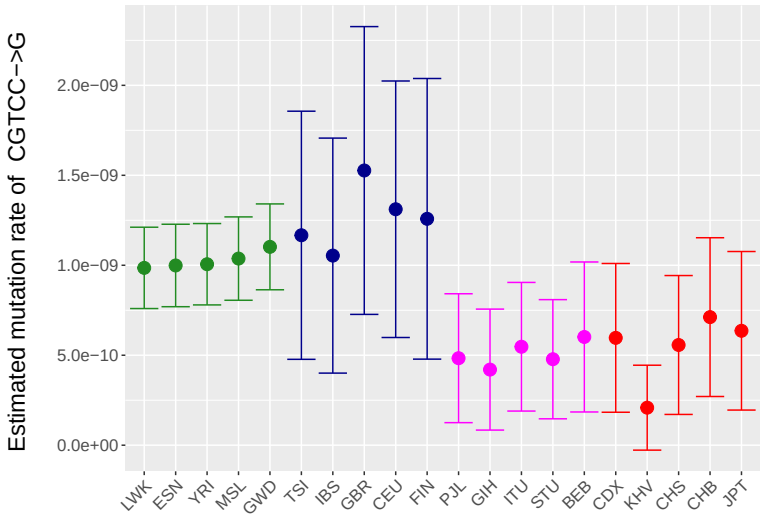
EAS heterogeneity at a higher sequence context level

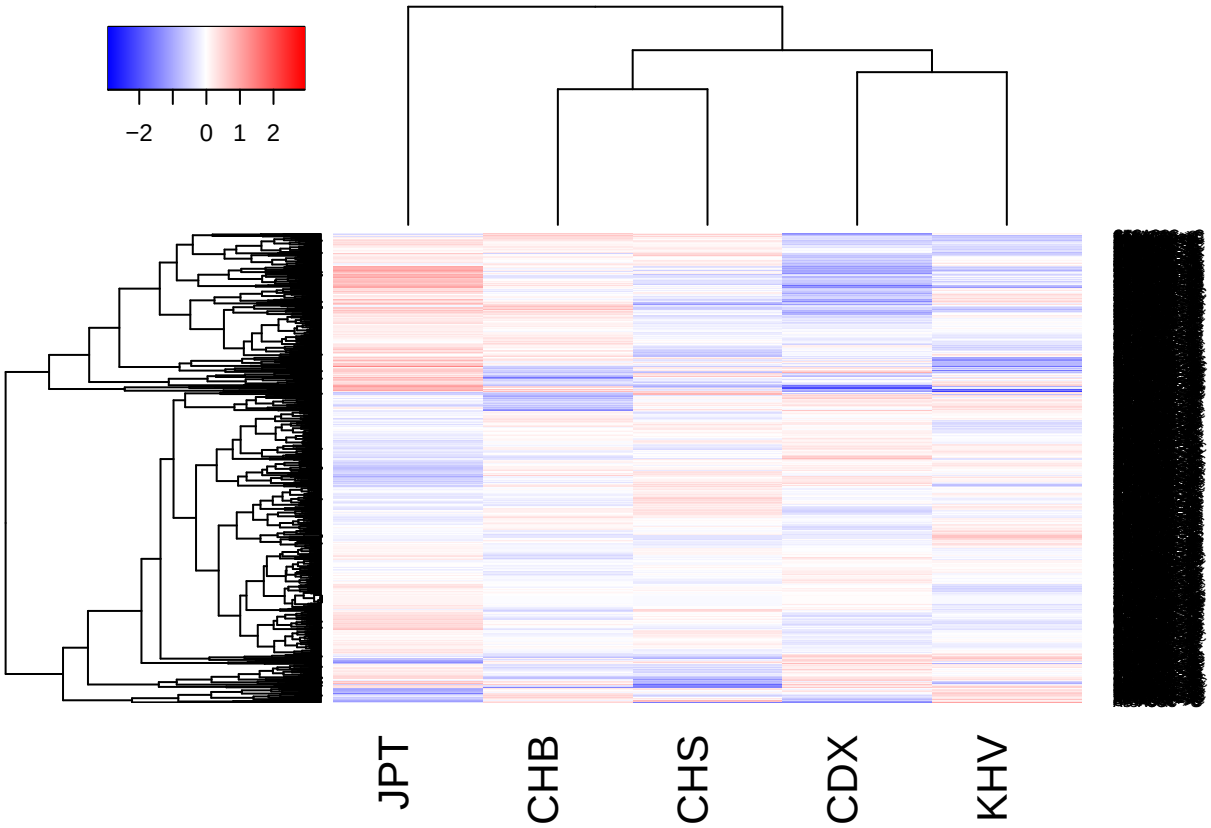
Since the heatmap above highlights a block of mutations elevated in Japan, we'd like to visualize how this signal fluctuates at higher context levels. First, we can start with 5mers



This plot delineates a group of polymorphisms which appear enriched in JPT and depleted in CDX. What is most notable about this plot is that it demonstrates that not all 5mers with the "*AC->C" or "ATA->A" subcontexts show the same heterogeneity. This hints that there may be some feature at a greater distance of sequence context that is driving this heterogeneity.

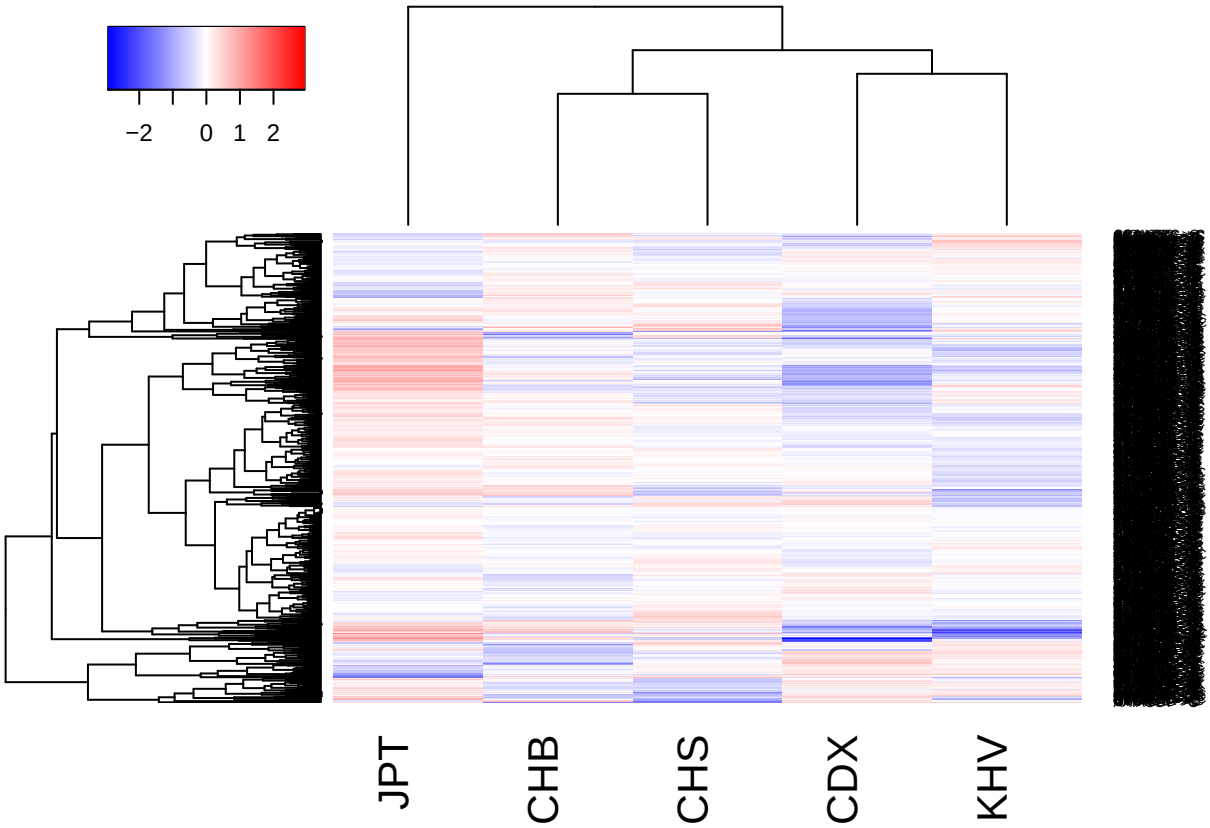
NOTE: It is not clear whether the mutation type which separates first, CGTCC->G represents a real signal - while it is depleted in KHV, the estimated rate in KHV overlaps with the CIs for all but one (CHB) of the remaining East Asian populations:





Among the 7mers, the first group that separates out appears similarly promising. One interesting feature is that these groups separate out not only by JPT enrichment but by CDX depletion. This makes sense based on the fact that rows are all normalized. Even if the enrichment is seen to varying degrees in CHS, CHB, and JPT, this will appear as a depletion in the population which is consistently not enriched for these mutations: CDX.

With subcontext.heatmap, we can visualize all 7mers which have 5mer contexts that fall into the first cluster in the first heatmap of this section.



Cluster properties

Overlap

Now that we have a few possible clusterings of 5mers and 7mers which we believe to be enriched in JPT, we can begin to ask questions about what these polymorphism types have in common. To begin, I've constructed five clusterings: One from each heatmap in this section, plus one (7mer - 3) which is the intersection of the two 7mer clusters.

Cluster	Number of Polymorphisms
3mer	5
5mer	56
7mer - 1	323
7mer - 2	436
7mer - 3	174

One interesting observation is that CAAACCC->C, the 7mer that is most highly significantly variable in heterogeneity tests across continents, can be found in all three 7mer clusterings.

X enrichment: Chi-squared

We have previously noted that CAAACCC->C and some of the substitution type with similar profiles of polymorphism counts tend to be enriched on the X chromosome. For this reason, it would be interesting to

check whether the mutation types in the clusters above are enriched on X compared to the autosomes. A naive way of doing this involves running a Fisher's exact test for each mutation type. I have written two functions to make this work:

- **X.preprocessing**: given a count dataframe, and a list of polymorphism types, returns a dataframe with the count of those polymorphisms on X vs the autosomes.
- **proptest.byrow**: given the output from X.proprocessing and a dataframe of genomewide counts, runs a fishers exact test for enrichment on X versus the autosomes for each polymorphism and returns a dataframe with p values appended.

Since this signal appears to be strongest in Japan, I first chose to use only the data from JPT. The table below shows a test for X enrichment in those 3mers which Harris and Pritchard have pointed out as heterogeneous in EUR and enriched in JPT.

Table 2: X enrichment among 3mer cluster

Context	Autosomes	X	Autosomal_sites	X_sites	p	x.enrichment
AAC->C	2592	214	32865196	2698506	0.9666770	1.0055221
TAT->T	4649	405	48432566	4159639	0.8037729	1.0143254
CAC->C	3364	255	32917149	2481964	0.9606271	1.0053353
GAC->C	1183	66	20803493	1591644	0.0142008	0.7292048
TAC->C	1594	101	25416447	2115811	0.0087295	0.7611514

It may seem surprising that GAC->C and TAC->C appear depleted on the X chromosome. In fact, this is not uncommon. X is depleted for polymorphisms in all populations, as we would expect since there are fewer X chromosomes in every population and so they would accumulate less genetic variation.

The next section will set up a binomial test which may be more sensitive to this particular issue.

Binomial test

One important problem with the analysis above is that it assumes under the null hypothesis that the probability of observing a polymorphism on the X chromosome is equal to the probability of oberving a polymorphism on an Autosome. Instead, it is more reasonable to expect that fewer polymorphisms will be observed on the X chromosome than an equally sized autosome, even when the mutation rate is the same between both chromosomes.

- Since males lack an X chromosome, there are fewer X's than autosomes in any given population that is not 100% female. Recall that for autosomes in a diploid population,

  $$\theta = 4N_e\mu.$$

  For X chromosomes, the left hand side of this equation must be smaller even when μ is the same, thus fewer polymorphisms should be expected.

- For related reasons, in a sample of mixed sexes, it must be true that we are sequencing fewer X chromosomes than any given autosome, which means that our power to detect genetic variation is reduced. Again, this will cause us to expect fewer polymorphisms to be observed on X than an autosome with the same size and mutation rate.

Based on the information above, a simple chi squared test is likely to give us a high rate of false negatives when we test for X enrichment. Instead, we can set up a binomial hypothesis test as follows.

Suppose we are interested in a specific polymorphism type, m. Under the null hypothesis, the ratio between polymorphisms on X and the Autosomes is the same for m as for all other mutations. Therefore, one way to estimate p_0 the null-hypothesis probability of observing an m-type polymorphism at a given site on X, is:

$$p_0 = \alpha p_m^A,$$

where

$$\alpha = \frac{\frac{\text{polymorphsims on X}}{\text{sites on X}}}{\frac{\text{polymorphisms on Autosomes}}{\text{sites on Autosomes}}}$$

for all of the data, and

$$p_m^A = \frac{\text{m polymorphisms on A}}{\text{m sites on A}}.$$

We would then like to test the hypotheses:

$$H_0 : p = p_0 \qquad H_1 : p \neq p_0$$

Table 3: X enrichment among 3mer cluster

	Context	Autosomes	X	Autosomal_sites	X_sites	alpha	p.0	p.MLE	p
4	AAC->C	2592	214	32865196	2698506	0.9197293	7.25e-05	0.0000793	0.1979863
34	TAT->T	4649	405	48432566	4159639	0.9197293	8.83e-05	0.0000974	0.0503093
46	CAC->C	3364	255	32917149	2481964	0.9197293	9.40e-05	0.0001027	0.1590298
76	GAC->C	1183	66	20803493	1591644	0.9197293	5.23e-05	0.0000415	0.0620526
90	TAC->C	1594	101	25416447	2115811	0.9197293	5.77e-05	0.0000477	0.0571605

As you can see in the table above, in JPT, the proportion of polymorphic sites on X is 0.9197293 the proportion of polymorphic sites on Autosomes. None of the tests above are nominally significant, showing that this signal is not expecially strong at the 3mer level.

It would be natural to run such a test for the 5mer and 7mer polymorphism lists we've developed. However, upon closer inspection, I've found that the vast majority of 5mers and 7mers in these groupings are not observed enough times on the X chromosome in Japan for a statistical test to be informative (i.e. many mutation types have 0-3 polymorphisms present in Japan on X). Below, you can see an example:

Table 4: X enrichment among 5mer cluster (first ten 5mers)

	Context	Autosomes	X	Autosomal_sites	X_sites	alpha	p.0	p.MLE	p
13	AAACA->C	452	42	5434067	451202	0.9197294	0.0000765	0.0000931	0.2008985
16	AAACC->C	367	46	2750815	230010	0.9197294	0.0001227	0.0002000	0.0017864
142	TTATT->T	1028	92	6440393	537846	0.9197294	0.0001468	0.0001711	0.1431163
148	CTATT->T	225	23	2903557	251193	0.9197294	0.0000713	0.0000916	0.2348148
199	ACACA->C	376	29	4289374	339687	0.9197294	0.0000806	0.0000854	0.7023021
202	ACACC->C	180	18	1812856	140125	0.9197294	0.0000913	0.0001285	0.1587076
205	ACACG->C	45	4	453562	30875	0.9197294	0.0000913	0.0001296	0.3716818
208	ACACT->C	341	21	2354105	188856	0.9197294	0.0001332	0.0001112	0.4843193
373	AGACA->C	188	7	3477557	282581	0.9197294	0.0000497	0.0000248	0.0605646
376	AGACC->C	103	12	1855998	140623	0.9197294	0.0000510	0.0000853	0.0876847

We find only two polymorphisms significant at fdr < 0.1:

Table 5: 5mers significantly enriched on X

Context	p.0	p.MLE	p	fdr
AAACC->C	0.0001227	0.0002000	0.0017864	0.0500203
CCACA->C	0.0000859	0.0002098	0.0000001	0.0000039

As you can see, some polymorphism types are found in large enough numbers for hypothesis testing, and some among those are significant. However, there are not many more significant results than we would expect by random chance, and many of the polymorphism types are not observed in sufficient numbers for a statistical test to be well-powered. Clearly, we are having issues with power. One solution is to pool across clustering groups.

```
#3mer groups
chi.bycluster(data = counts.JPT.3mer, gw = gw.3mer,dr = dr.3mer.EASonly, k = 2, n = 1, exclude = c()) #
```

```
## Chisquared p        alpha          P_0          MLE
## 0.9799195829 0.9197293181 0.0001900281 0.0001900494
```

```
chi.bysubset(data = counts.JPT.3mer, gw = gw.3mer, muts = JPT3mers) # p = 0.2, JPT enriched group
```

```
## Chisquared p        alpha          P_0          MLE
## 2.060635e-01 9.197293e-01 7.671536e-05 7.978501e-05
```

```
#5mer groups
chi.bycluster(data = counts.JPT.5mer, gw = gw.5mer,dr = dr.5.JPT3mers, k = 5, n = 1, exclude = c()) # p
```

```
## Chisquared p        alpha          P_0          MLE
## 7.540027e-02 9.197294e-01 7.839317e-05 8.356453e-05
```

```
#7mer groups
chi.bycluster(data = counts.JPT.7mer, gw = gw.7mer,dr = dr.7.JPT3mers, k = 4, n = 2) # p = 0.15
```

```
## Chisquared p        alpha          P_0          MLE
## 1.535500e-02 9.197295e-01 8.612156e-05 9.813113e-05
```

```
chi.bycluster(data = counts.JPT.7mer, gw = gw.7mer,dr = dr.7.JPT3mers, k = 4, n = 2, # p = 0.65
              exclude = c("CAAACCC->C", "CCCACAG->C", "TTTATTT->T"))
```

```
## Chisquared p        alpha          P_0          MLE
## 6.528760e-01 9.197295e-01 7.921031e-05 7.688538e-05
```

```
chi.bycluster(data = counts.JPT.7mer, gw = gw.7mer,dr = dr.7.JPT5mers, k = 7, n = 1) # p = 0.003
```

```
## Chisquared p        alpha          P_0          MLE
## 2.313558e-03 9.197295e-01 8.957819e-05 1.015212e-04
```

```
chi.bycluster(data = counts.JPT.7mer, gw = gw.7mer,dr = dr.7.JPT5mers, k = 7, n = 1, # p = 0.36
              exclude = c("CAAACCC->C", "CCCACAG->C", "TTTATTT->T"))
```

```
## Chisquared p        alpha          P_0          MLE
## 3.623553e-01 9.197295e-01 8.556759e-05 8.900450e-05
```

```
chi.bysubset(counts.JPT.7mer, gw.7mer, JPT7mers.3) # p = 0.000044
```

```
## Chisquared p        alpha          P_0          MLE
## 4.415153e-05 9.197295e-01 9.533125e-05 1.236967e-04
```

```
muts.7.filtered <- subset(JPT7mers.3, !is.element(JPT7mers.3, c("CAAACCC->C", "CCCACAG->C", "TTTATTT->T
chi.bysubset(counts.JPT.7mer, gw.7mer, muts.7.filtered) # p = 0.49
```

```
## Chisquared p        alpha          P_0          MLE
## 4.885926e-01 9.197295e-01 8.403065e-05 8.847552e-05
```

Polymorphisms highly enriched on X

Since we have all these pieces set up, I wanted to run the following test on X:

```
# 3mers
x.3mers <- binomtest.byrow(counts.JPT.3mer, gw.3mer)
kable(subset(x.3mers, x.3mers$p < 0.05/length(x.3mers$Context)))
```

	Context	Autosomes	X	Autosomal_sites	X_sites	alpha	p.0	p.MLE	p
6	AAC->T	2750	264	32865196	2698506	0.9197293	0.0000770	0.0000978	0.0001752
13	ACA->A	6711	631	45612549	3701201	0.9197293	0.0001353	0.0001705	0.0000000
28	GCT->T	14013	1141	30414124	2253335	0.9197293	0.0004238	0.0005064	0.0000000
37	GAT->T	5285	467	30337805	2465111	0.9197293	0.0001602	0.0001894	0.0004238
41	CAT->G	23689	1557	41986297	3435057	0.9197293	0.0005189	0.0004533	0.0000001
64	GCG->T	23382	1530	4384814	268329	0.9197293	0.0049045	0.0057020	0.0000000
79	GCA->A	5564	462	31719717	2416995	0.9197293	0.0001613	0.0001911	0.0003889

```
# 5mers
x.5mers <- binomtest.byrow(counts.JPT.5mer, gw.5mer)
kable(subset(x.5mers, x.5mers$p < 0.05/length(x.5mers$Context)))
```

	Context	Autosomes	X	Autosomal_sites	X_sites	alpha	p.0	p.MLE	p
820	CCACA->C	266	47	2846685	224060	0.9197294	8.59e-05	0.0002098	1e-07

```
# 7mers
x.7mers <- binomtest.byrow(counts.JPTCHBCHS.7mer, gw.7mer)
kable(subset(x.7mers, x.7mers$p < 0.05/length(x.7mers$Context)))
```

	Context	Autosomes	X	Autosomal_sites	X_sites	alpha	p.0	p.MLE	p
6151	TGGCTCT->T	192	36	201906	14608	0.9593272	0.0009123	0.0024644	2e-07
10801	CAAACCC->C	78	27	137120	11001	0.9593272	0.0005457	0.0024543	0e+00
13522	CCCACAG->C	56	25	207077	15570	0.9593272	0.0002594	0.0016057	0e+00
17765	CTGCATA->G	32	17	168260	15895	0.9593272	0.0001824	0.0010695	0e+00

Future directions

Additionally, CAAACCC->C, the very strong signature of heterogeneity in East Asia that we identified in our first analysis, is enriched on the X chromosome. Two simple experiments come out of this example:

1. If we remove the X chromosome, do we see the same signal?
2. Among the mutations that we've identified as interesting in this analysis, how many are X enriched? Is this more than we would expect from random chance (Binomial(n, p = 0.05))?

Weaknesses

There are a few broad weaknesses to this analysis

1. There is no correction for uncertainty. This causes mutation types with relatively few observations (and thus, usually a high variance in rate across populations), to obscure real results.
2. Plotting more than a couple hundred polymorphism types tends to get messy, fast. This makes it harder to make sense of patterns at higher sequence context levels (although, for the moment, it's the best we've got).
3. The default distance function for creating these clusterings is Euclidean. For our purposes, it might make more sense to use Mahattan distance.