Figures for paper

Rachael 'Rocky' Aikens, Voight Lab

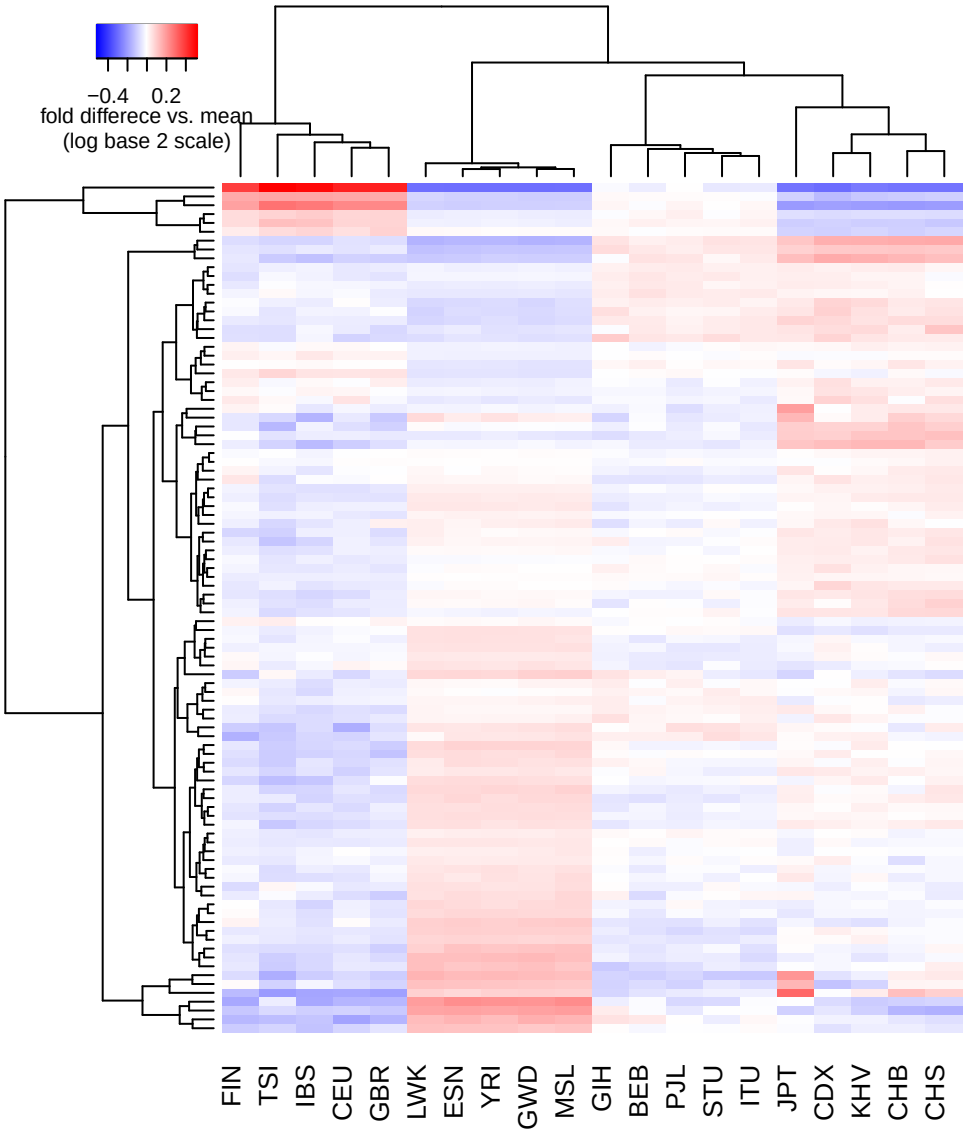
August 3, 2017

Figure 1

A: heatmap of all 3mer mutation types

To make this figure, we need the following **functions** and *datasets*:

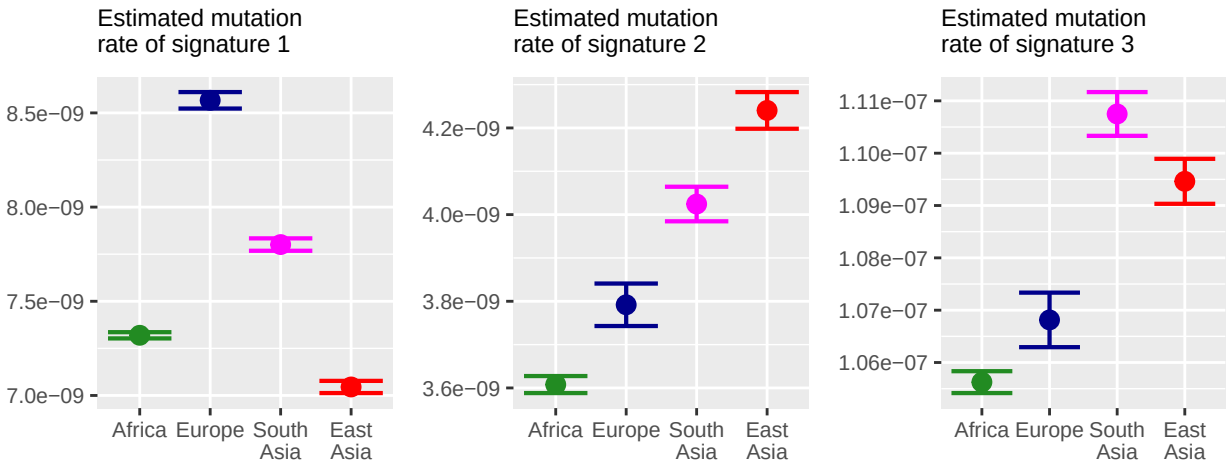
- **norm/norm.byrow** need these to normalize the data before making a heatmap
- **make.heatmap** calls heatmap2 to make a heatmap. The function defined herein is slightly different from the one in the heatmaps_report workflow; there are some small modifications to format the dendrogram and plot area specifically for this figure.
- *3mer rate matrix* saved in 'rate_profiles/rates_3mer.txt'



BCD: CI plots of polymorphism clusters

To make these panels, I need the following **functions** and *datasets*:

- **CI.plot.bygroup** Makes a plot of the rates of a group of mutations. Will bug out if the mutations are of the same context, although that's not a problem for these figures.
- *3mer count dataframes for all ancestral continental groups*



E: CI plots for signal 4

This figure was harder to make, and we'll probably revamp the way we do it. Once we've figured that out, I'll add that code here. For now, this figure requires the following:

- **CI.plotsubpop.bygroup** same as CI.plot.bygroup, but works for more populations than just the ancestral continental groups.
- *3mer count dataframes for all ancestral continental groups (except EAS)*
- *3mer count dataframes for all EAS subpopulations*

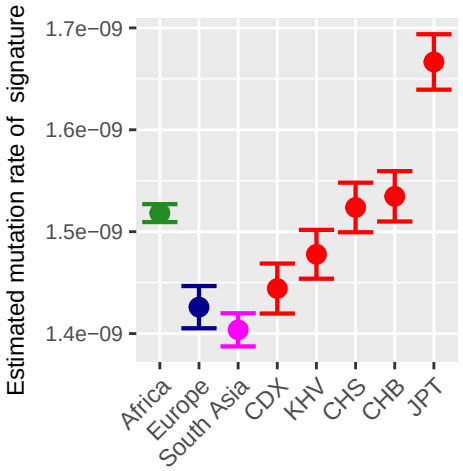


Figure 2

AB: scatter plot examples

To make these figures, I need

- **subrate.scplot** makes a scatterplot of all 7mers with a given 3mer subtype.
- *7mer count dataframes for JPT and CDX*
- *7mer count dataframes for East Asia and Europe*

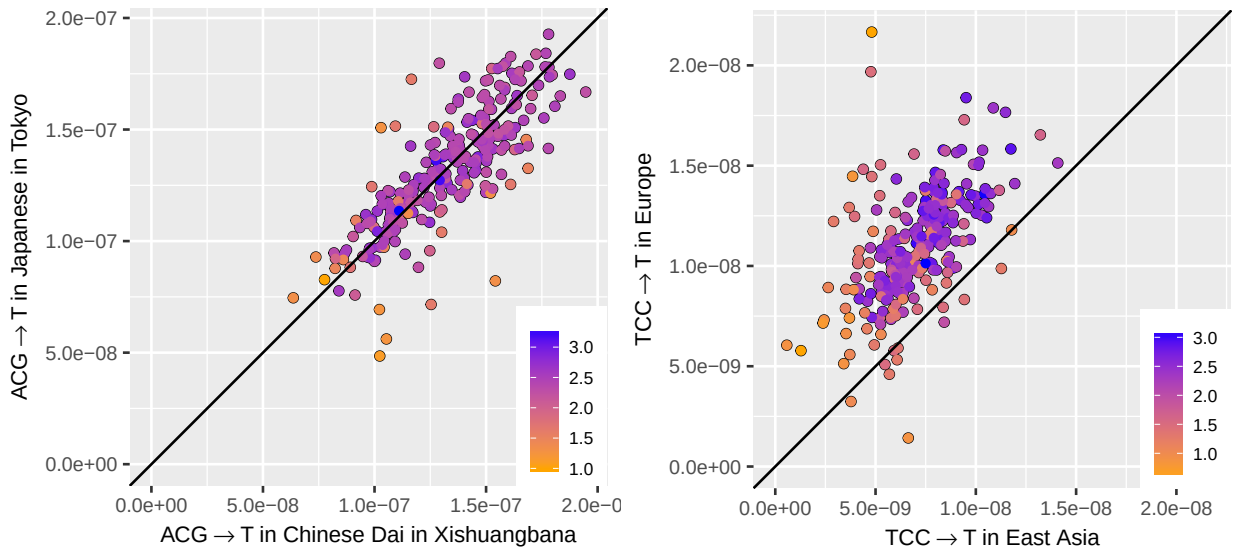


Figure 3

A: table of X enriched polymorphisms

We can obtain the first set of fdr-adjusted p values using a pairwise chi squared test between CDX and JPT 7mers whose 3mer subcontext is a part of signal 4.

Context	p	fdr
TTTATTT->T	0.0000000	0.0000000
CAAACCC->C	0.0000000	0.0000000
AGTACAG->C	0.0000001	0.0000206
CCCACAG->C	0.0000570	0.0106536
AATACAG->C	0.0000925	0.0115197
TCAACAG->C	0.0000850	0.0115197
ATGACAG->C	0.0001374	0.0146650
TCCACAG->C	0.0002214	0.0206713
ATAACAG->C	0.0003625	0.0300886
Next, we have	to run a te	st for X enrichment, described in the paper.

Context	Autosomes	X	Autosomal_sites	X_sites	alpha	p.0	p.MLE	p
TTTATTT->T	587	65	1446006	120078	0.961597	0.000390	0.000541	0.006988
AATACAG->C	33	1	259131	21488	0.961597	0.000122	0.000047	0.928033
AGTACAG->C	37	4	143681	10790	0.961597	0.000248	0.000371	0.279709
ATAACAG->C	41	4	185728	15810	0.961597	0.000212	0.000253	0.432024
ATGACAG->C	30	4	196565	15689	0.961597	0.000147	0.000255	0.201155
CAAACCC->C	80	27	137120	11001	0.961597	0.000561	0.002454	0.000000
CCCACAG->C	61	25	207077	15570	0.961597	0.000283	0.001606	0.000000
TCCACAG->C	35	3	186629	13409	0.961597	0.000180	0.000224	0.435029
TCAACAG->C	29	7	165171	13893	0.961597	0.000169	0.000504	0.010330

BC

These two panels use datasets and functions:

- **subrate.scplot**
- **CI.plot.subpop.bygroup**
- *7mer count dataframes for all EAS subpops*

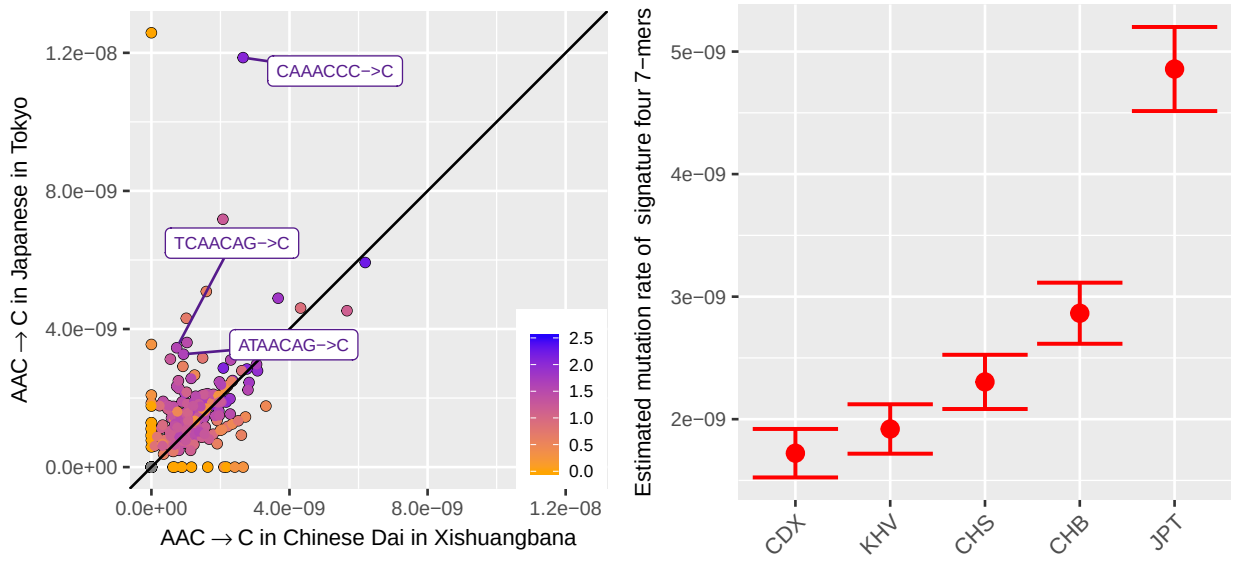


Figure 4

A

Table 3: 10 most significant new 7mers using ordered p value correction

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

BC

We can't use my usual graphing function to make Figure 4B because there are '→' characters that we need to insert in the plot text. We will also need the following:

- **subrate.scplot**

- *7mer count dataframes for all nonadmixed continental populations*