Figures for paper

Rachael 'Rocky' Aikens, Voight Lab

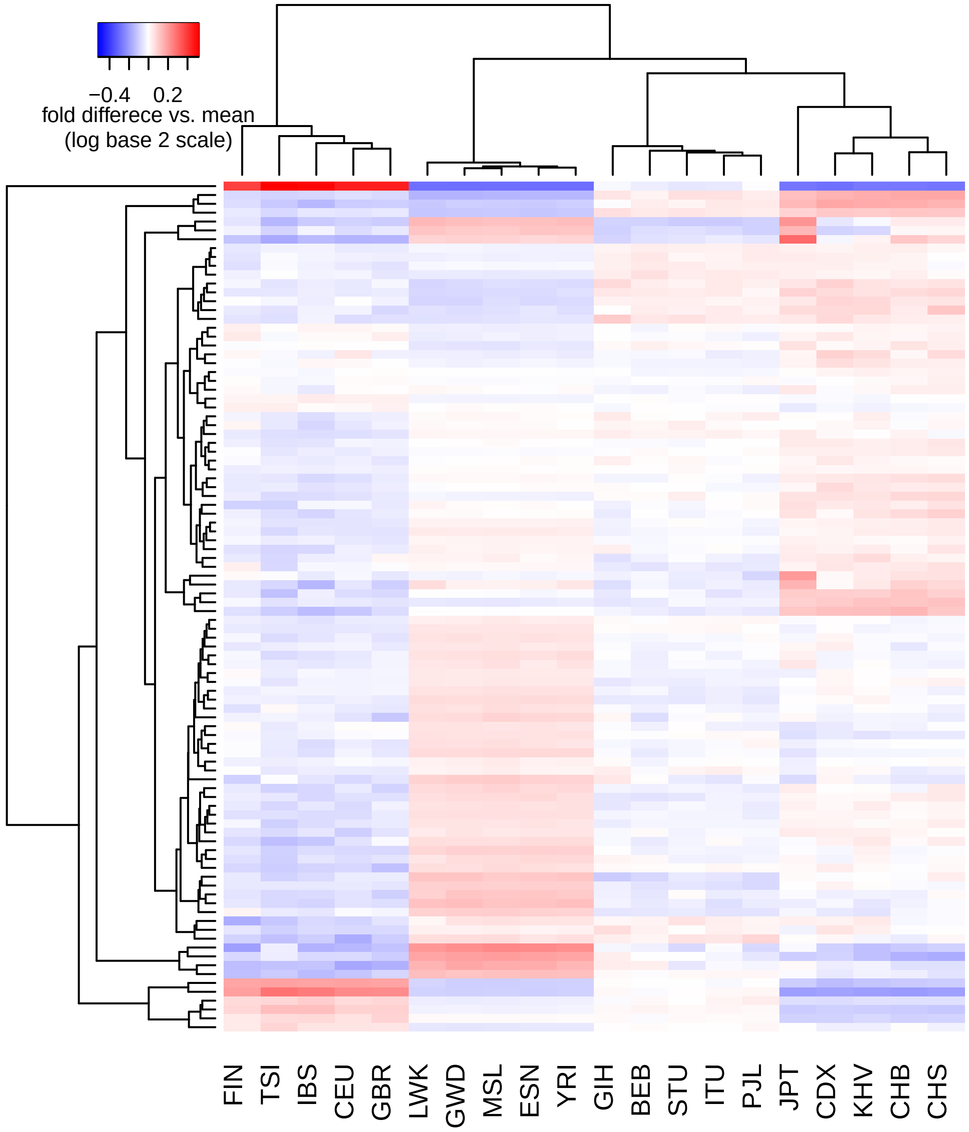
August 3, 2017

Figure 1

A: heatmap of all 3mer mutation types

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

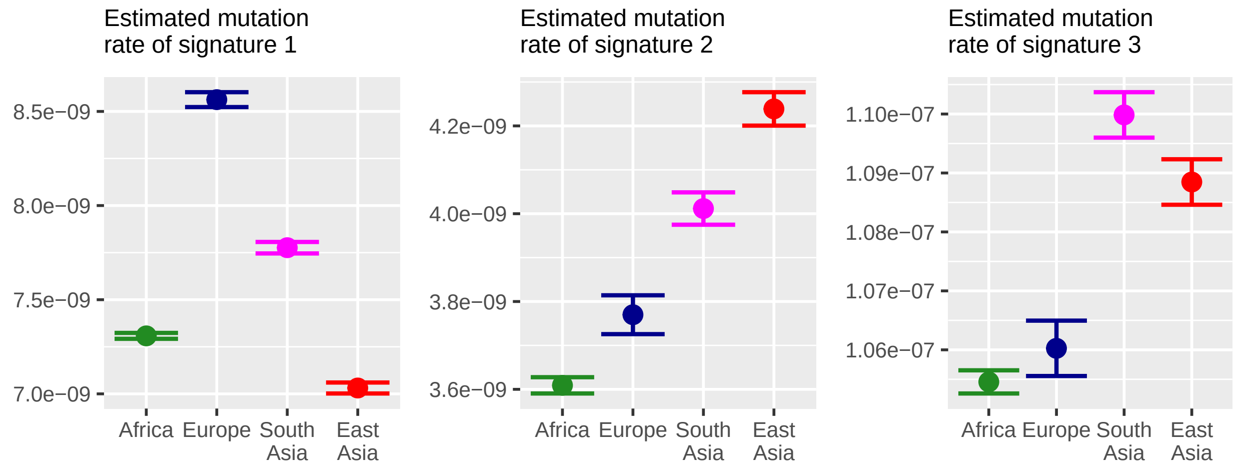
- **norm/norm.byrow** need these to normalize the data before making a heatmap
- **make.heatmap** calls heatmap2 to make a heatmap. The function defined herein is slightly different from the one in the heatmaps_report workflow; there are some small modifications to format the dendrogram and plot area 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*

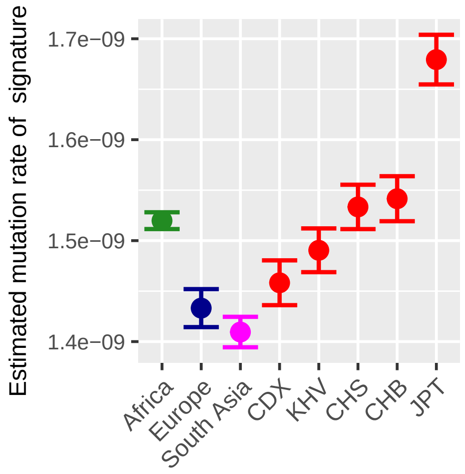


Figure 2

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*

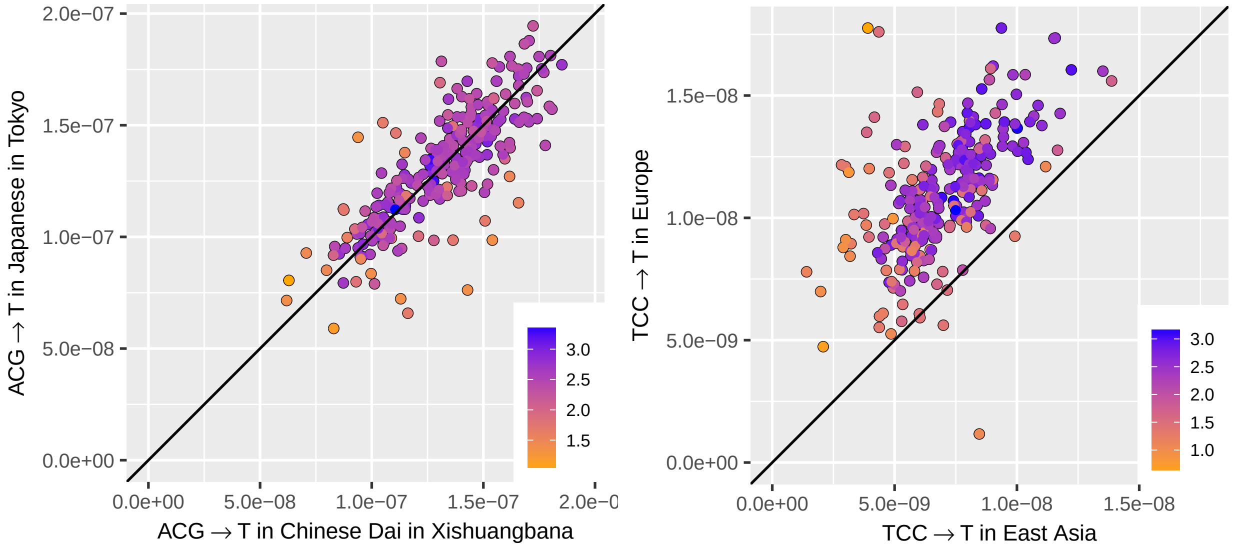


Figure 3

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

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

BC

These two panels use datasets and functions:

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

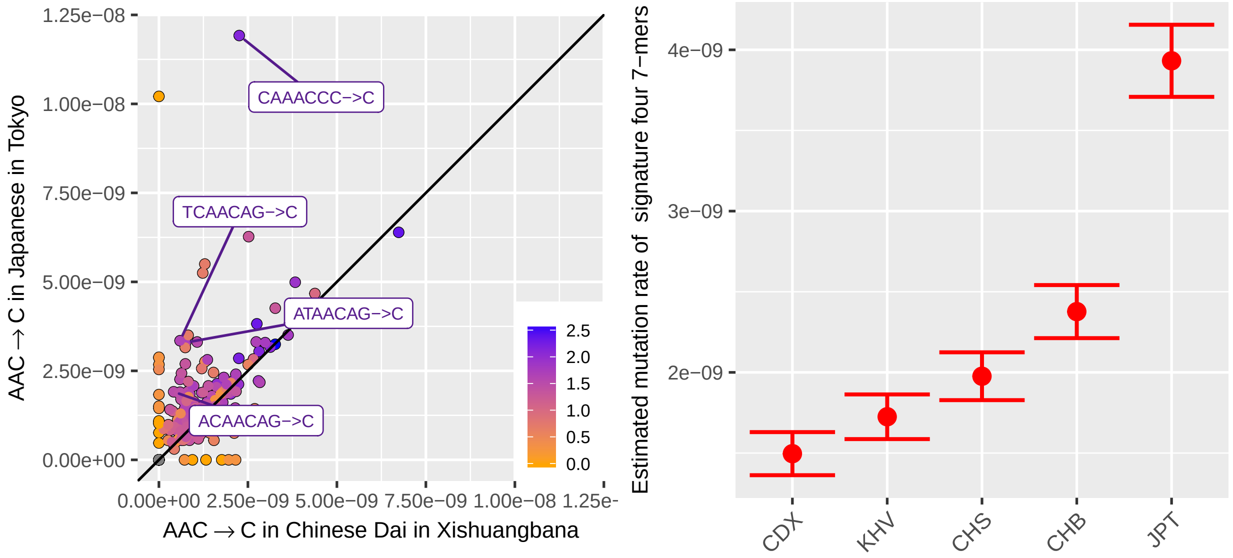


Figure 4

A

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

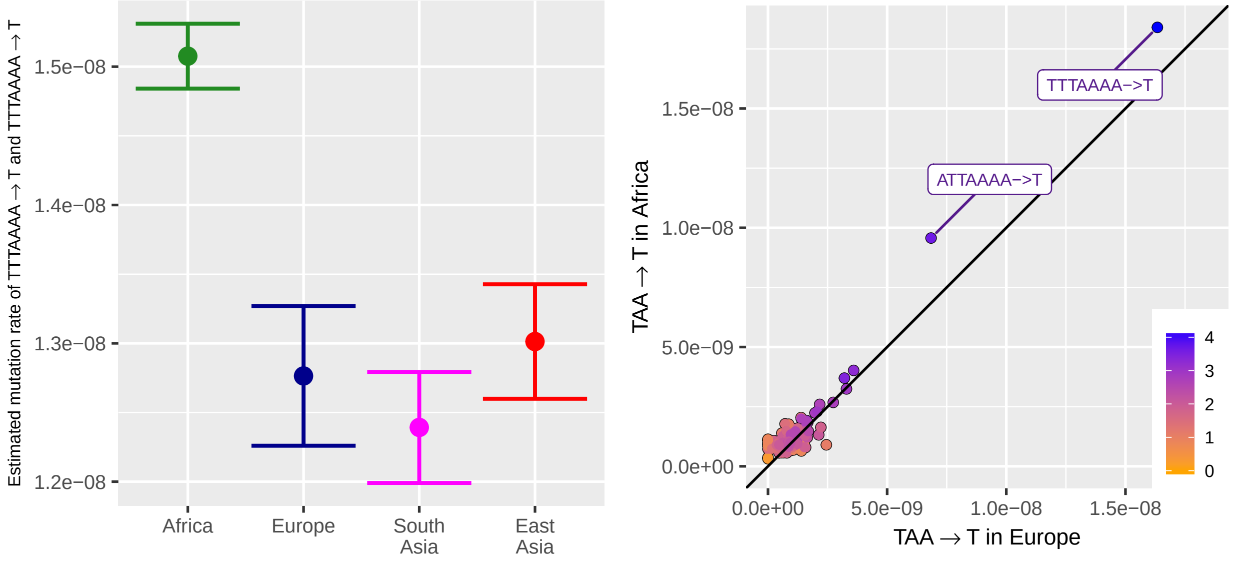
Context	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
CAAACCC->C	127	22	128	12	2.984752e-39
TTTATTT->T	2796	431	808	478	2.166804e-25
TTTAAAA->T	12011	1961	2939	2846	1.199912e-21
ATTAAAA->T	3773	496	857	808	1.968521e-21
AAACAAA->A	3108	446	766	578	2.110224e-21

Context	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
AGTACAG->C	51	14	55	9	2.375565e-15
ACTAAAA->G	2187	513	833	705	2.887438e-15
CTGCATA->G	72	19	63	12	7.903406e-14
TATATAT->G	7093	1181	1710	1724	3.338030e-11
AGGCTTT->T	1174	177	442	339	4.507439e-09

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*


Estimated mutation rate of TTTAAAA → T and TTTAAAA → T
1.5e−08
1.4e−08
1.3e−08
1.2e−08
Africa
Europe
South Asia
East Asia
TTTAAAA−>T
ATTAAAA−>T
TAA → T in Africa
1.5e−08
1.0e−08
5.0e−09
0.0e+00
0.0e+00
5.0e−09
1.0e−08
1.5e−08
TAA → T in Europe