Replication in Simons Genome Diversity Project

Rachael Caelie (Rocky) Aikens

10/29/2018

Introduction

This document is meant to show all our efforts to replicate our study in the Simons Genome Diversity Project (SGDP). Since SGDP dataset is much smaller than the 1,000 genomes dataset, extra care must be taken to conserve statistical power. As a result, we will only replicate a subset of our discoveries from the main analysis, and restrict the number of hypothesis tests to a minimum where possible.

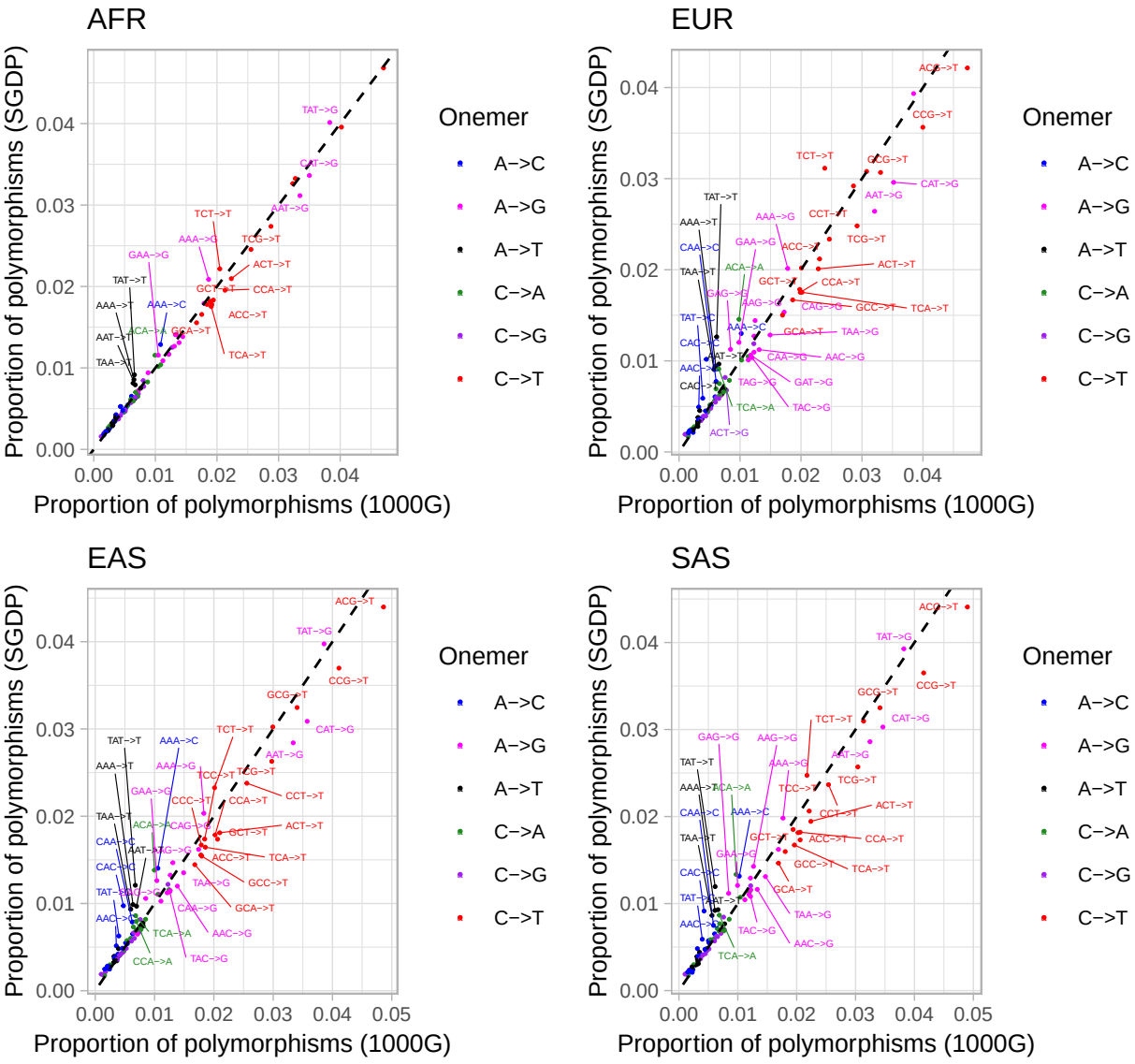
Preliminary Checks for Correctness

Visualizing dataset agreement

Below is a simple visualization of agreement between the datasets. For a given polymorphism c and population P, the proportion of that polymorphism in the population is defined as:

$$\frac{\text{Number of private polymorphisms of type c in population P}}{\text{Total number of private polymorphisms in population P}}$$

The plot below shows the agreement between polymorphism proportions from 1,000 genomes (x-axis) and SGDP (y-axis) for each population.


AFR
EUR
EAS
SAS
Proportion of polymorphisms (SGDP)
Proportion of polymorphisms (1000G)
Onemer
A−>C
A−>G
A−>T
C−>A
C−>G
C−>T

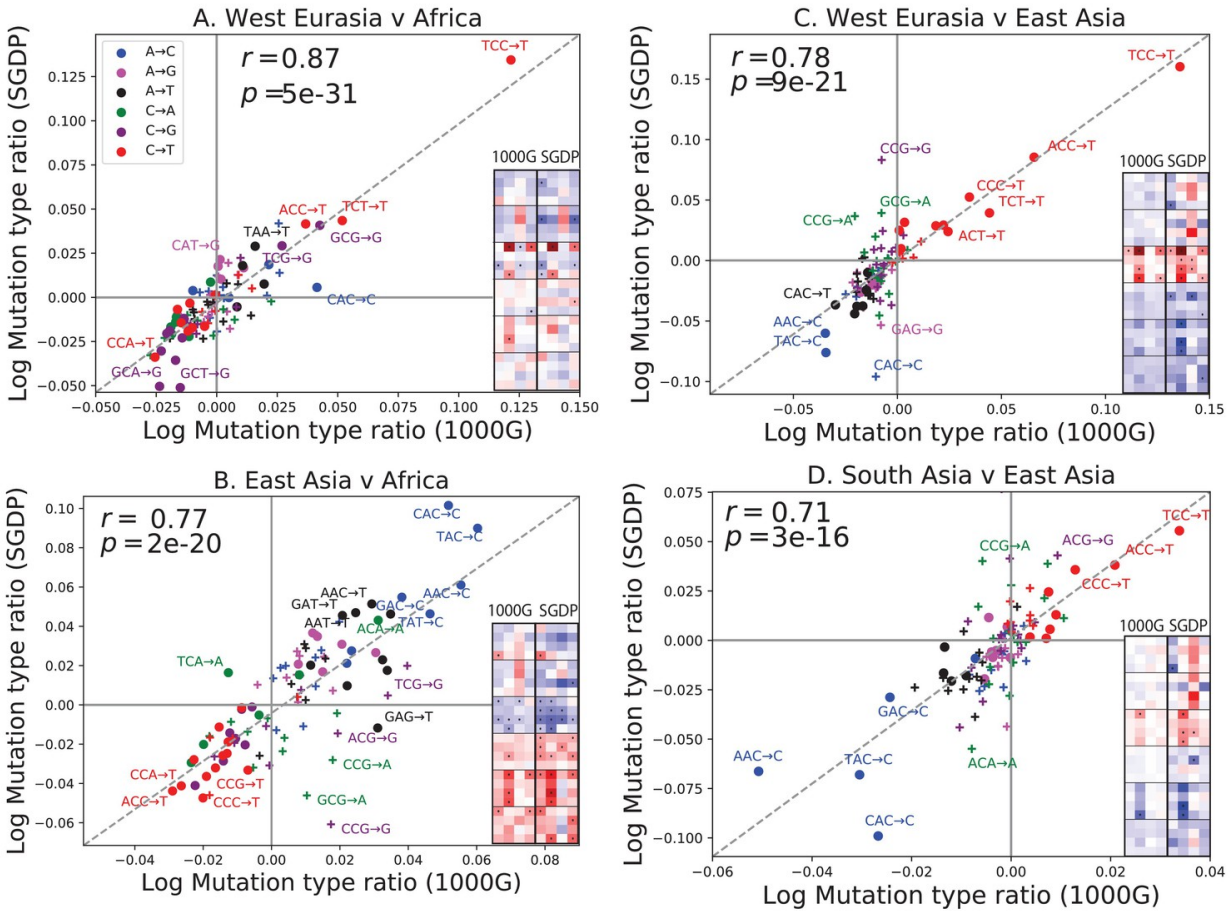
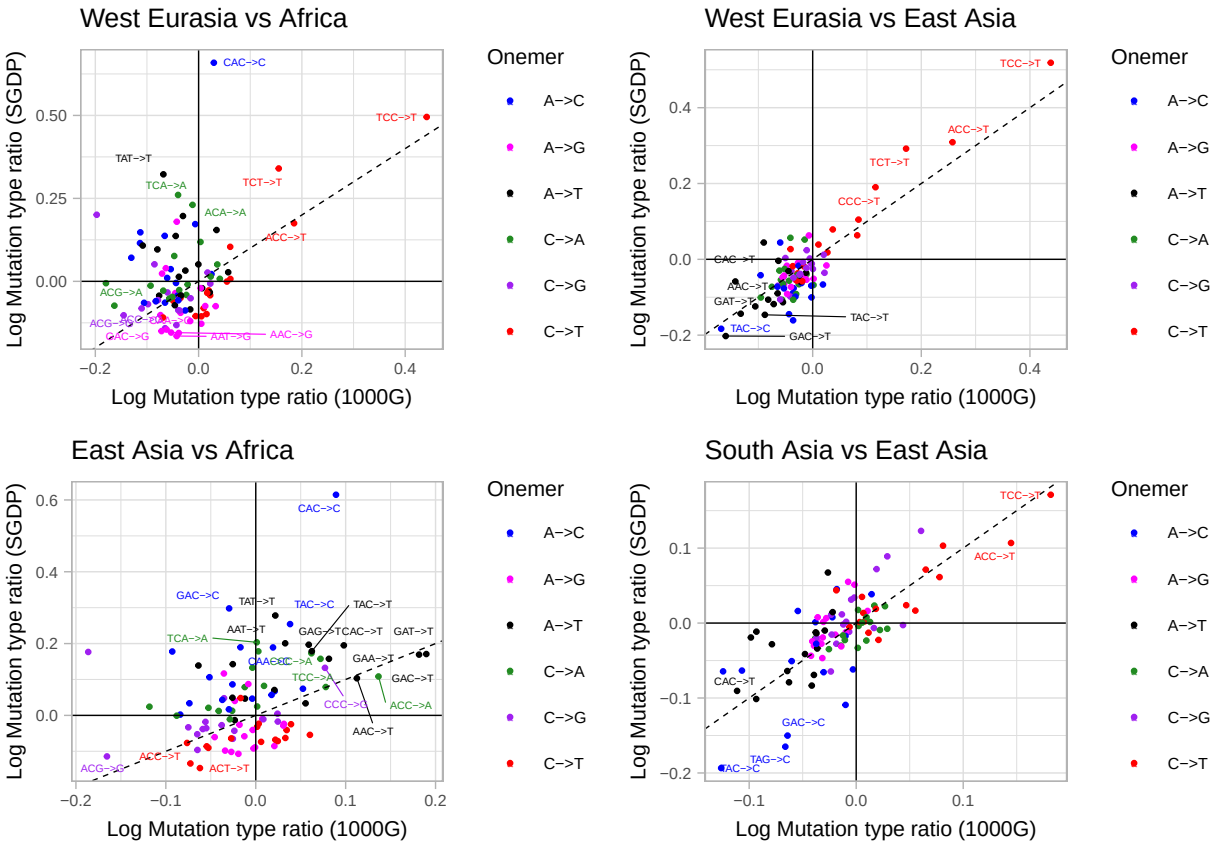


Figure 1: Figure 2 from KH JP.

Replication of Harris and Pritchard, figure 2

Next we attempt to replicate figure 2 from Harris and Pritchard, 2017. They first calculate the ratio of the proportion of each polymorphism in a pair of populations, then plot the agreement between the log (base e) of the ratios:

Here is our attempt at the same plot:


West Eurasia vs Africa
Log Mutation type ratio (SGDP)
Log Mutation type ratio (1000G)
CAC−>C
TCC−>T
TAT−>T
TCA−>A
TCT−>T
ACA−>A
ACC−>T
ACG−>A
AAT−>G
AAC−>G
Onemer
A−>C
A−>G
A−>T
C−>A
C−>G
C−>T
West Eurasia vs East Asia
TCC−>T
ACC−>T
TCT−>T
CCC−>T
TAC−>T
GAC−>T
East Asia vs Africa
CAC−>C
GAC−>C
TAT−>T
TAC−>C
AAT−>T
GAT−>T
TCA−>A
GAA−>T
TCC−>A
CCC−>C
ACC−>A
AAC−>T
ACG−>C
ACT−>T
South Asia vs East Asia
TCC−>T
ACC−>T
CAC−>T
GAC−>C
TAG−>C
TAC−>C

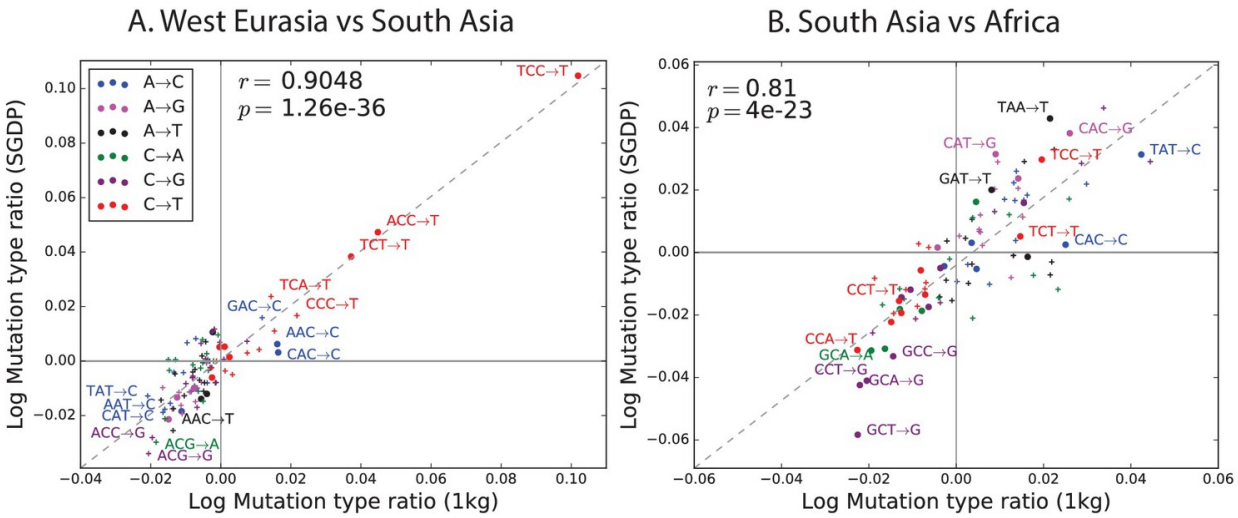
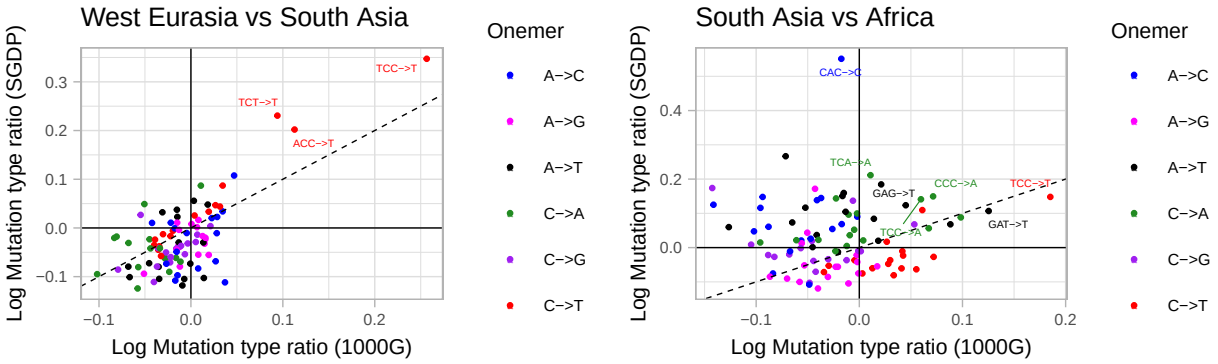


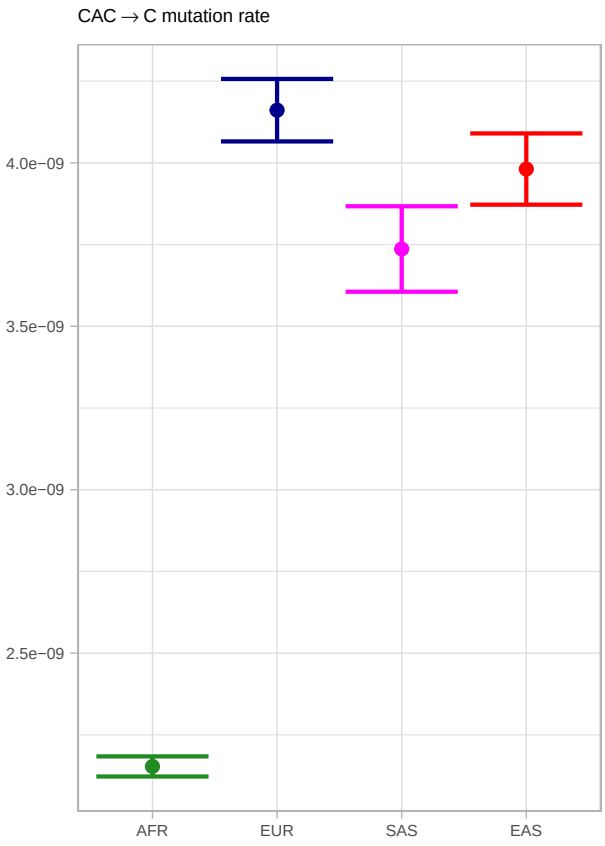
Figure 2: KH JP Figure 2S2

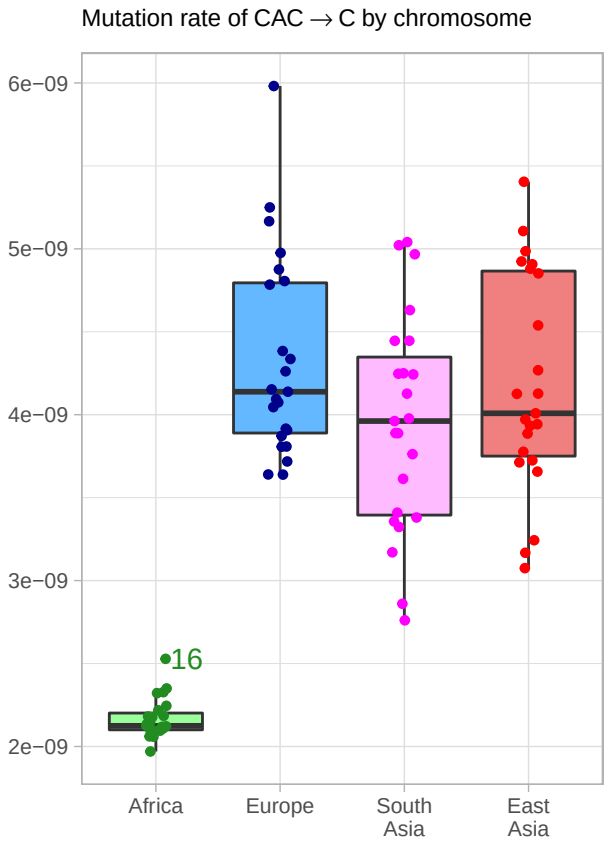
Also included are two supplementary comparisons:



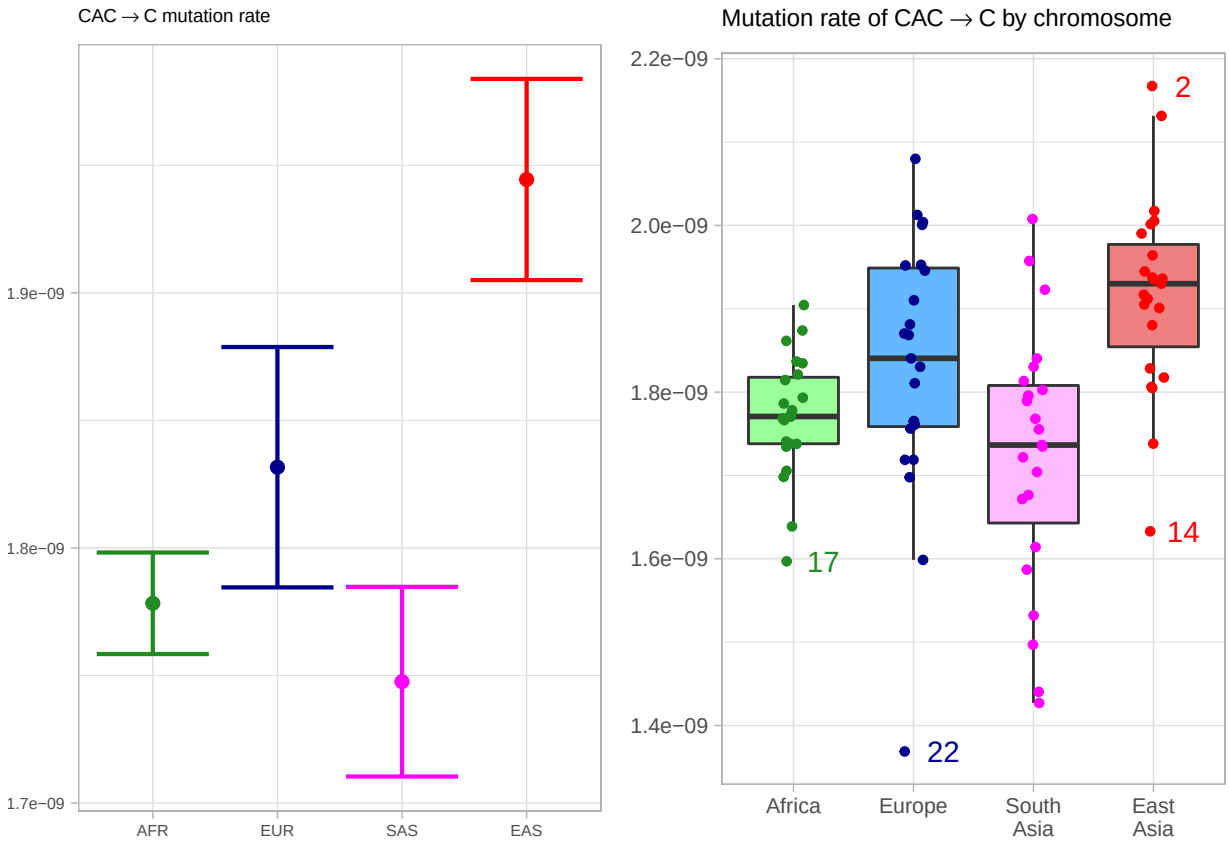
I make the following observations:

1. The scales of the plots do not agree. We have noted before that the trends and patterns we observe in polymorphism ratios between polymorphisms agree with Harris and Pritchard, but the exact numeric estimates do not. It is not clear why this is. One explanation may be that they use all polymorphism, but we use private variants only in our dataset. Our approach may cause differences between populations to be larger than when their approach is used.
2. Whenever we make a comparison with Africa, the figures do not agree.
3. CAC -> C is consistently an outlier in comparisons with Africa. Here are the rates in SGDP:





And here they are in 1KG



Looking through Harris and Pritchard's figures, their relative rates for CAC->C are:

EAS ∼ EUR > SAS > AFR.

3-mer Substitution Classes that Vary Across Continents

Test for Homogenity Across all Continental Groups

Here, we run the same p-ordered hypothesis test the we used on the 1,000 genomes dataset. However, we will test for the significance of variation in the top 15 polymorphism classes listed in Table 1.

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

- *3mer count dataframes for all ancestral continental groups*
- **Fourway.chi** Tests for heterogeneity among any four count dataframes
- **ordered.p** performs ordered p value correction

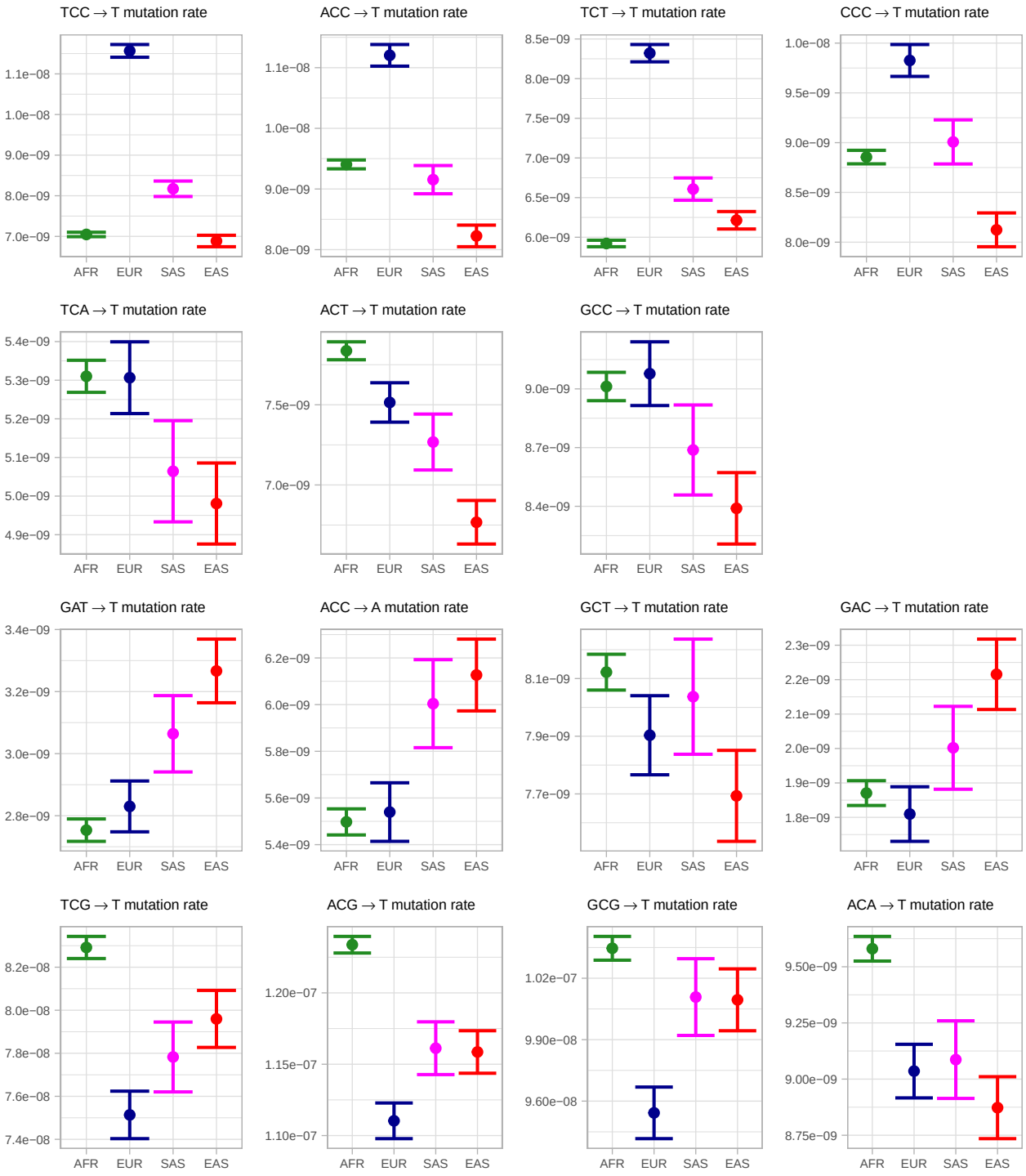
Table 1: Replication of Table 1 with data from SGDP

Context	AFR_relative_rate	EUR_relative_rate	SAS_relative_rate	EAS_relative_rate	p
TCT->T	1	1.41	1.12	1.05	0
TCC->T	1	1.64	1.16	0.98	0
ACC->T	1	1.19	0.97	0.87	2.30897118680569e-161
CCC->T	1	1.11	1.02	0.92	6.22382896953434e-68
ACT->T	1	0.96	0.93	0.86	1.19404601491732e-45
ACG->T	1	0.90	0.94	0.94	1.19530354955357e-43

Context	AFR_relative_rate	EUR_relative_rate	SAS_relative_rate	EAS_relative_rate	p
ACA->T	1	0.94	0.95	0.93	2.62241982201823e-30
TCG->T	1	0.91	0.94	0.96	1.48930212591532e-25
GAT->T	1	1.03	1.11	1.19	1.44648562849783e-23
ACC->A	1	1.01	1.09	1.11	4.69066491089771e-17
GCG->T	1	0.92	0.98	0.98	4.34008861260168e-15
GCC->T	1	1.01	0.96	0.93	3.29796264795211e-14
TCA->T	1	1.00	0.95	0.94	1.66303149702796e-13
GAC->T	1	0.97	1.07	1.18	5.57326596649956e-10
GCT->T	1	0.97	0.99	0.95	7.89314714556592e-07

Inferred Mutation Rate in top 15 Heterogeneous 3-mers

Here, we will calculate the inferred mutation rate of the 3-mers from table 1.

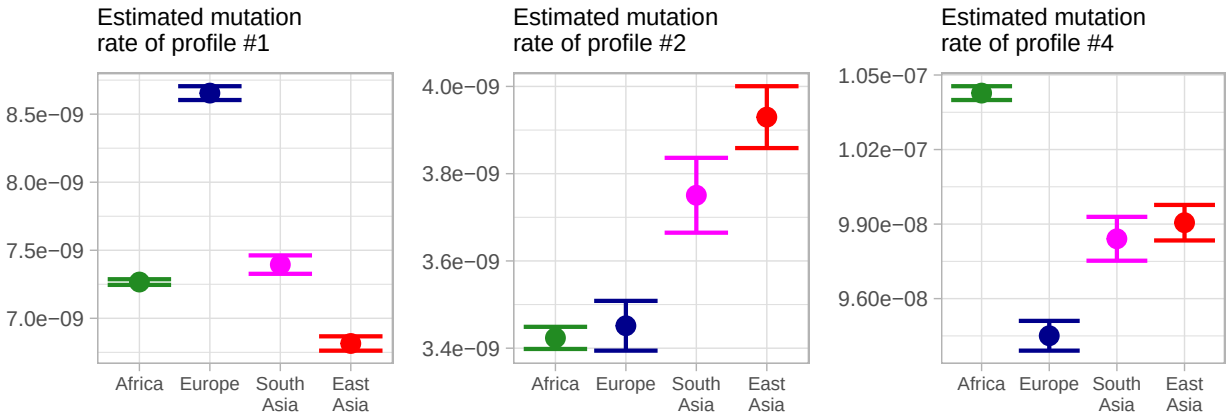


Signatures of Variation at the 3-mer Level

We will not attempt to construct a heatmap of the 3-mer signatures from the SGDP data, since this analysis on 1,000 genomes was mostly heuristic. However, we will plot the inferred mutation rate from SGDP for each of the signatures reported in figure 1.

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.



Broader Sequence Contexts of 3-mer Signatures

We will not attempt to replicate scatter plots as in Figure 2 because it is not likely that inferred mutation rate for 7-mers in SGDP will be accurate enough for these plots to be meaningful. Likewise, the suggestion that certain 7-mers are driving the *AC→C enrichment in Japan compared to East Asia is an interesting result, but since there are 5 Chinese Dai and 3 Japanese individuals in SGDP, attempting to replicate this result may not be appropriate. Ideally, this preliminary finding could be replicated and perhaps further explored in a large Asian genomic dataset, the likes of which, to our knowledge, are currently not publically available.

Signatures of Variation at Broader Sequence Contexts

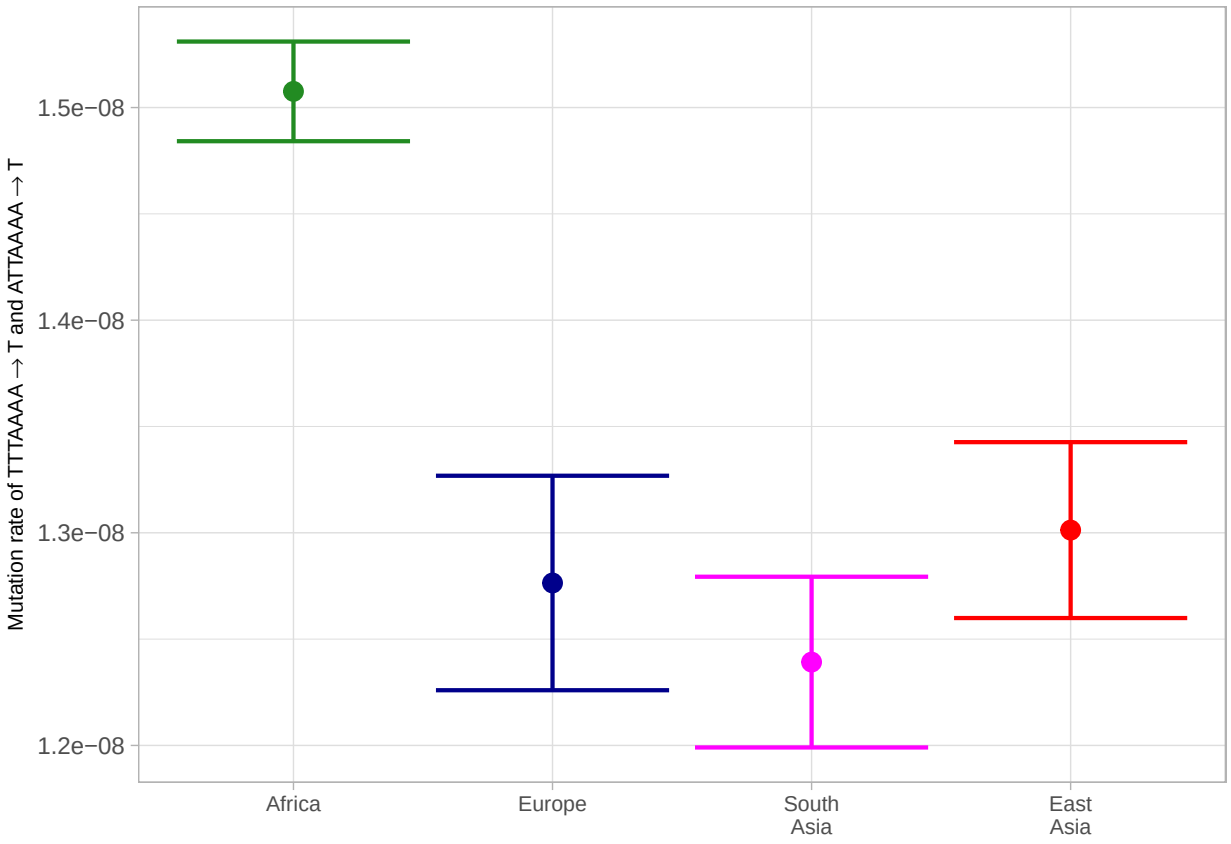
We will not attempt hypothesis testing across all 5-mer and 7-mer polymorphism classes, since many of these tests are sure to invlolve too few observations to be carried out, and the hypothesis testing burden would massively reduce statistical power. Rather, we will repeat the hypothesis tests for only the 7-mer classes shown in table 3.

Table: Table 3 recalculated with SGDP data

Context AFR_relative_rate EUR_relative_rate SAS_relative_rate EAS_relative_rate p

---

We will additionally plot the inferred mutation rate of the WTTAAA→T 7-mers across continents. We can't use my usual graphing function to make This figure because there are '→' characters that we need to insert in the plot text.


Mutation rate of TTTAAAA → T and ATTAAAA → T
1.5e−08
1.4e−08
1.3e−08
1.2e−08
Africa
Europe
South Asia
East Asia