Replication in Simons Genome Diversity Project

Rachael Caelie (Rocky) Aikens

10/29/2018

Contents

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.

Table 1: Sample sizes from 1,000 Genomes and SGDP

Study	Africans	Europeans	East.Asians	South.Asians
1,000 Genomes	504	503	504	489
SGDP	44	69	47	39

Table 2: Variant counts from 1,000 Genomes and SGDP (millions)

Study	Africans	Europeans	East.Asians	South.Asians
1,000 Genomes	7.0	1.3	2.0	2.0
SGDP	1.9	0.7	0.5	0.3

0. Basic Checks of SGDP Data

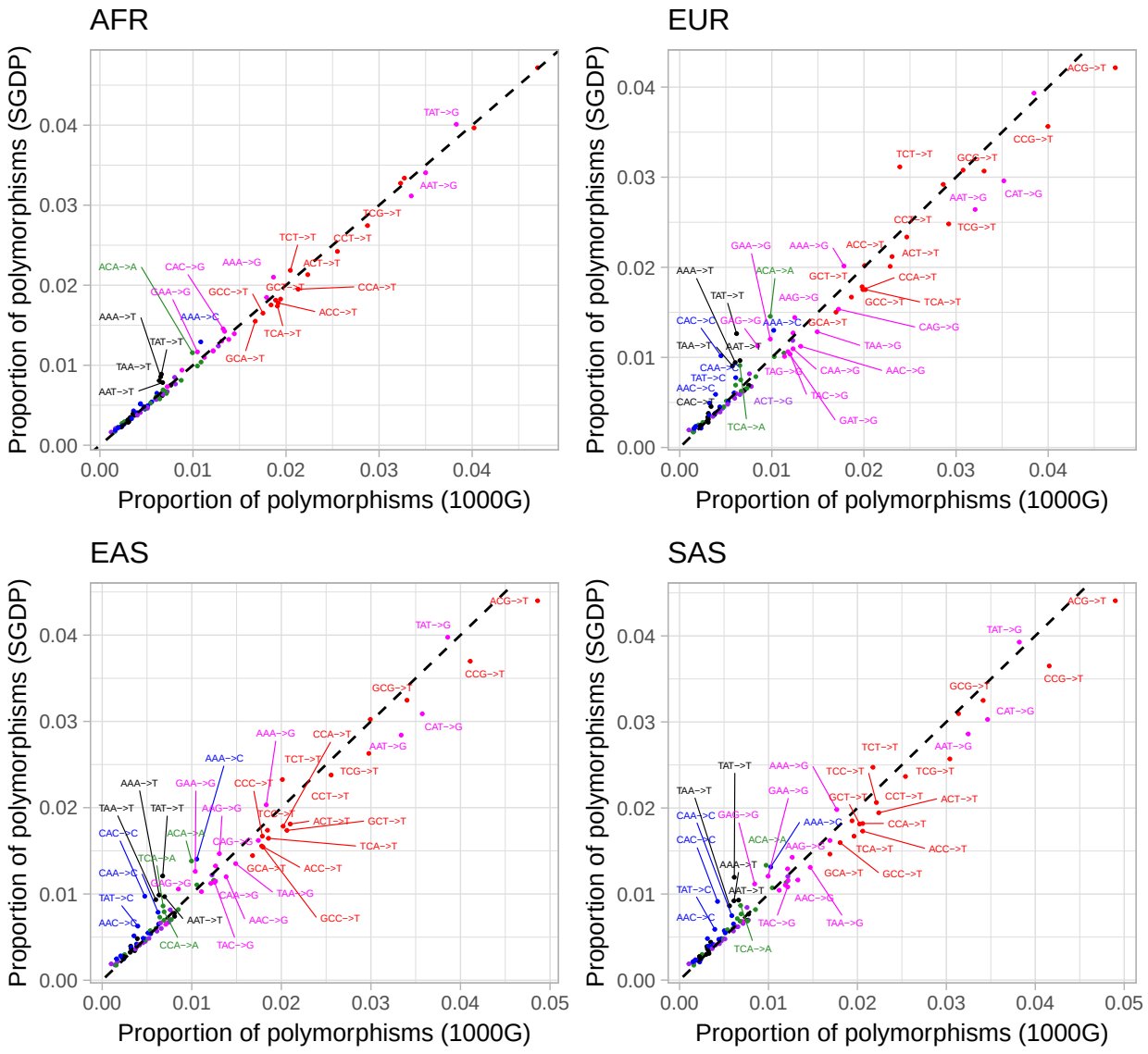
Here are some basic visualizations of the SGDP Data and their match up with 1KG and previous results.

0.1 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.



Qualitatively, there is relatively good agreement between the datasets. A few differences are noticeable:

1) There is a higher proportion of TCT→T substitutions in SGDP than 1KG in all populations, especially Europeans.

2) The following A→T substitutions are more abundant in SGDP in all populations: AAA→T, TAA→T, TAT→T, and (somewhat), AAT→T.

3) The following A→C substitutions are more abundant in SGDP in Europe, East Asia, and South Asia: CAA→C, AAA→C, CAC→C, and (somewhat) AAC→C. AAA→C also may be slightly more abundant in Africa in SGDP than in 1KG.

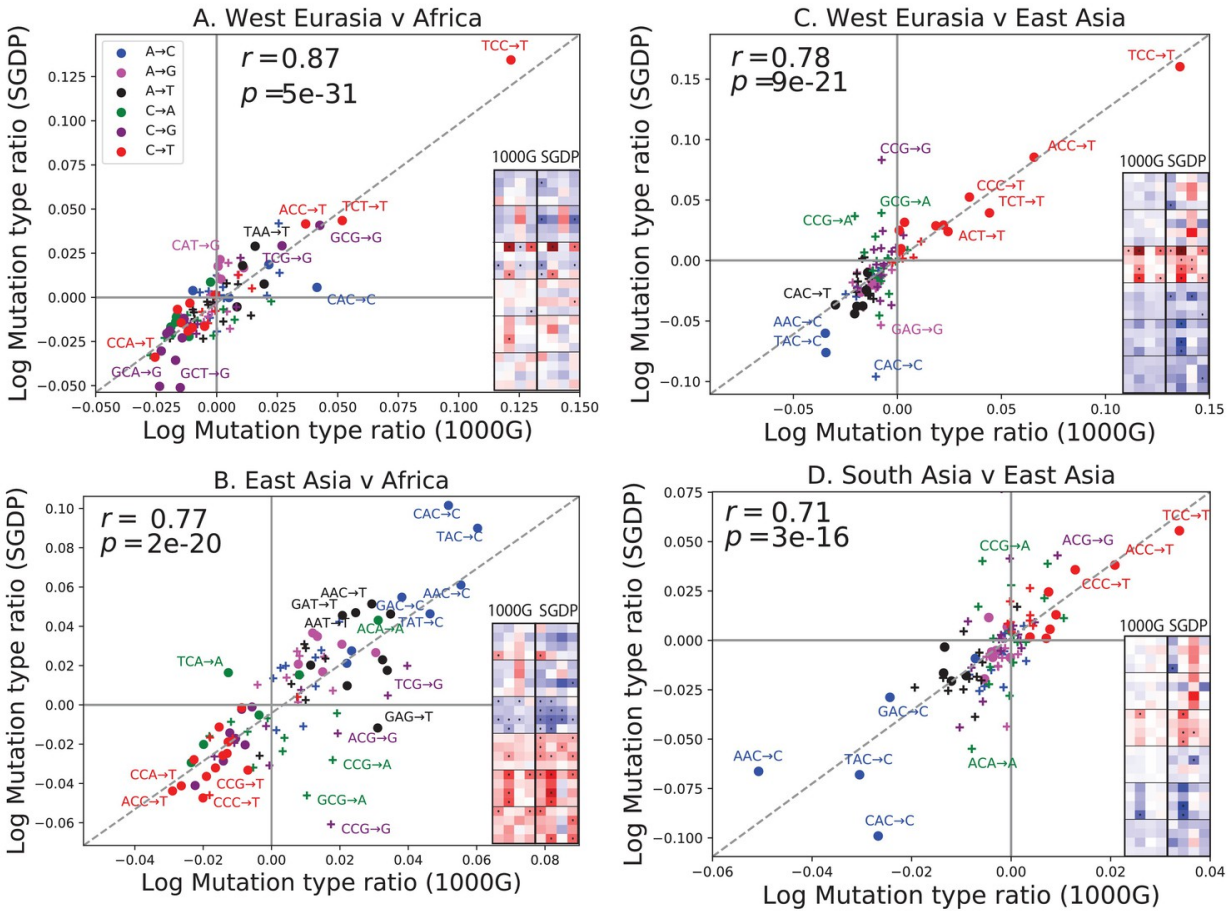
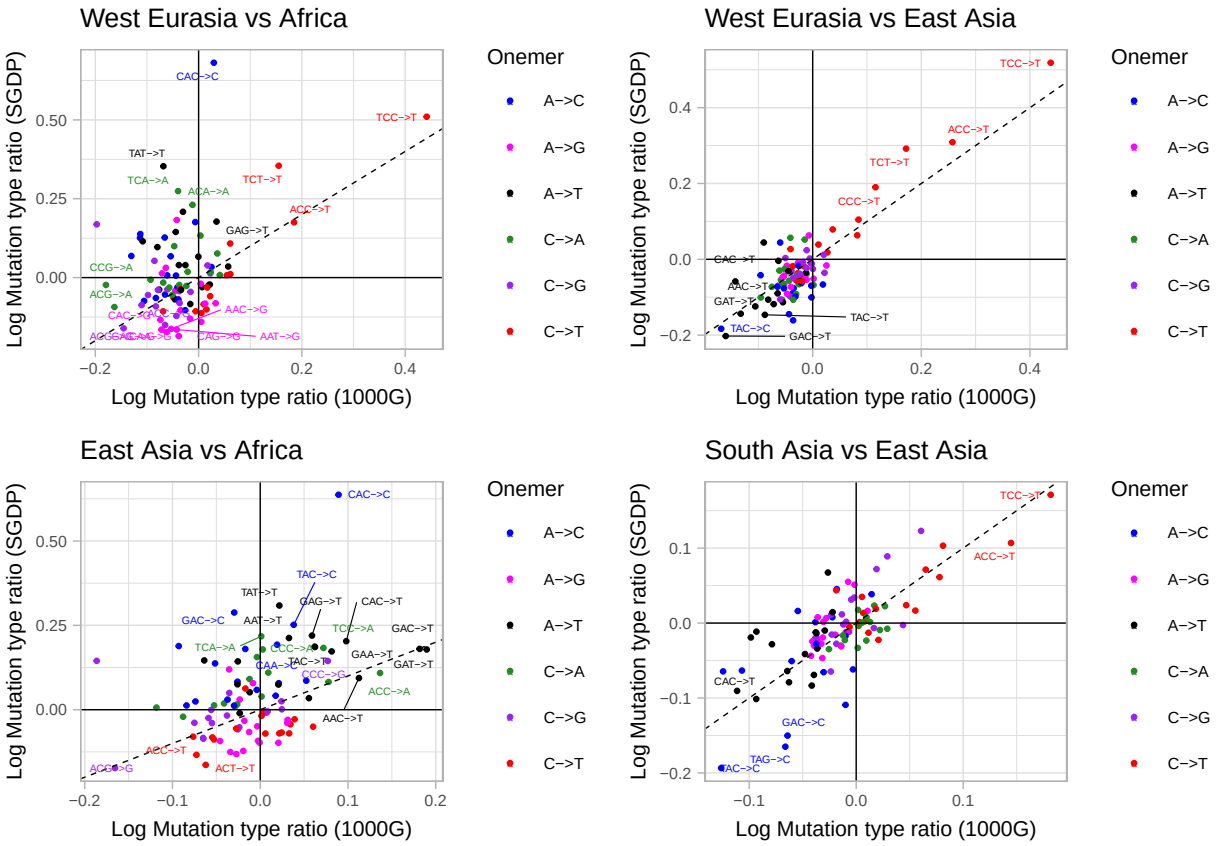


Figure 1: Figure 2 from KH JP.

0.2 Agreement with Harris and Pritchard

Next we attempt to replicate figure 2 from Harris and Pritchard, 2017 (above). 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)
Onemer
A−>C
A−>G
A−>T
C−>A
C−>G
C−>T
West Eurasia vs East Asia
East Asia vs Africa
South Asia vs East Asia

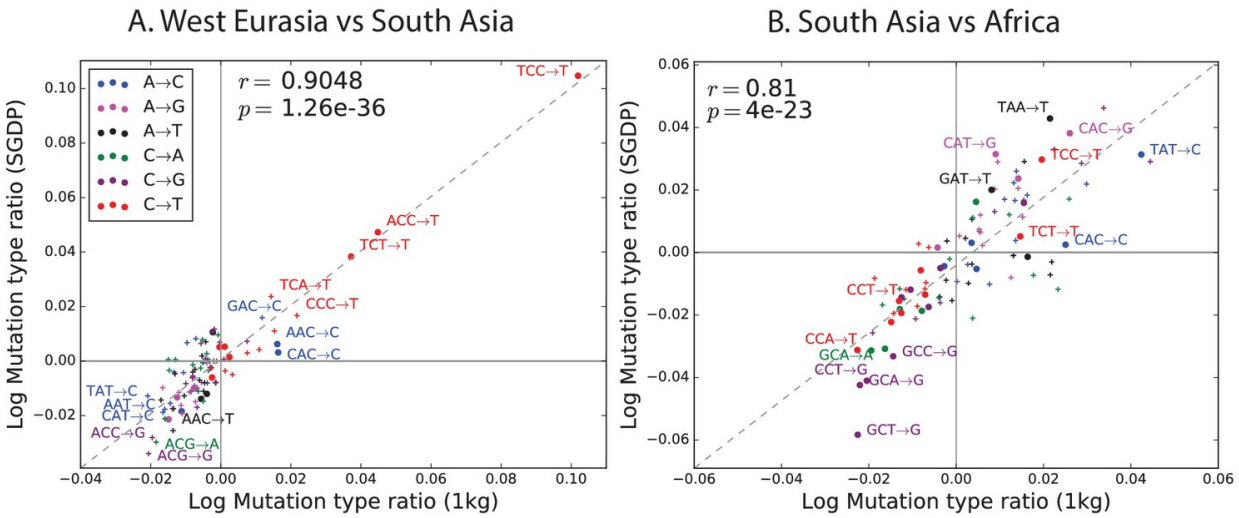
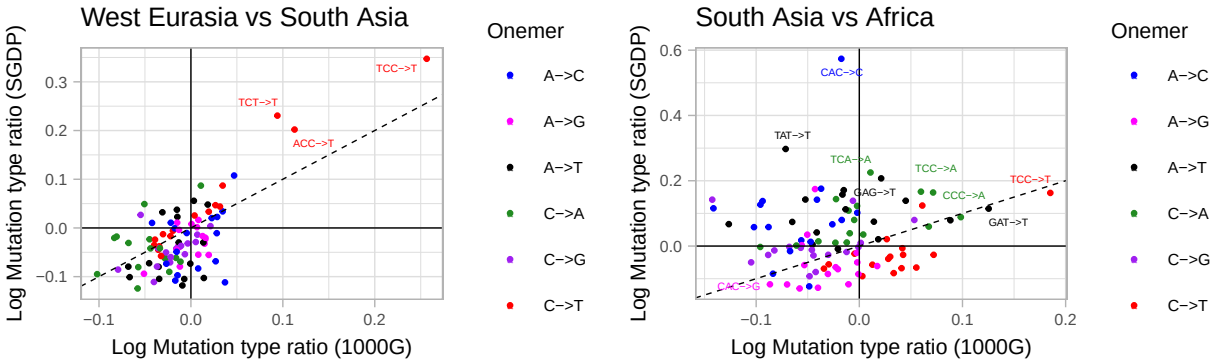


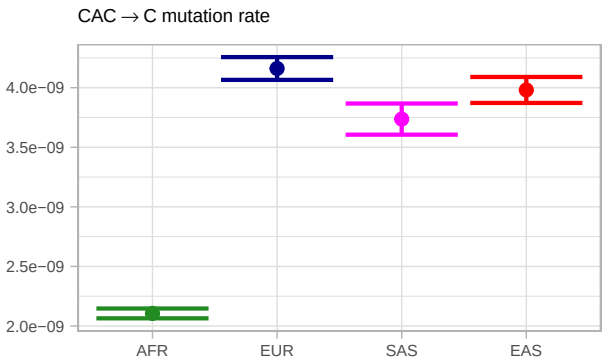
Figure 2: KH JP Figure 2S2

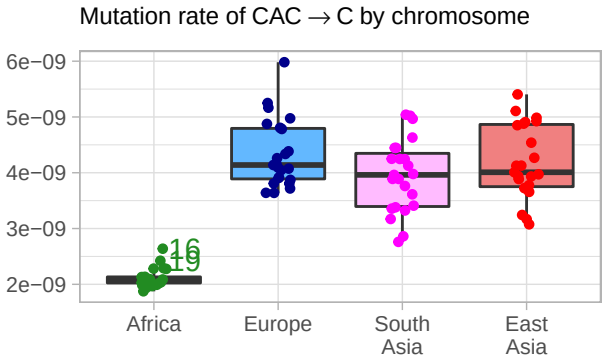
Also included are two supplementary comparisons:



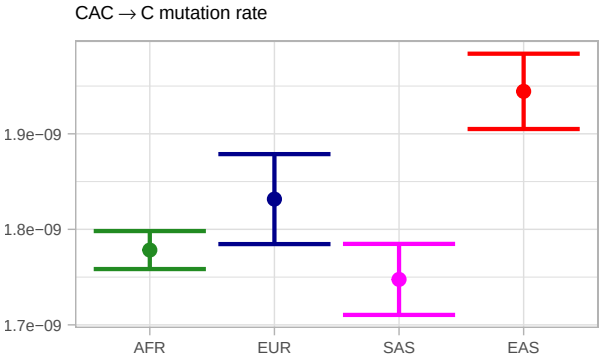
I make the following observations:

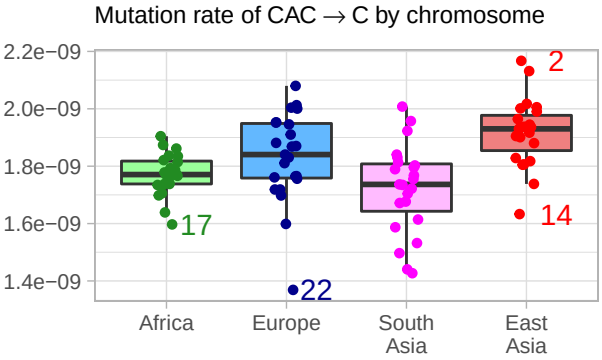
1. The scales of the plots do not agree. We have noted before that the trends and patterns we observe in polymorphism ratios between polymorphisms agree with Harris and Pritchard, but the exact numeric estimates do not. It is not clear why this is. One explanation may be that they use all polymorphism, but we use private variants only in our dataset. However, redoing the analysis with shared variants included does not fix the discrepancy. It is possible (though unlikely) that Harris and Pritchard were using logarithm base 10 rather than the natural log to make these figures, which would explain a discrepancy by a factor of three. However, they state in their manuscript that the natural logarithm was 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 $\sim$ EUR $>$ SAS $>$ AFR.

1. 3-mer Substitution Classes that Vary Across Continents

1.1 Test for Homogenity Across all Continental Groups

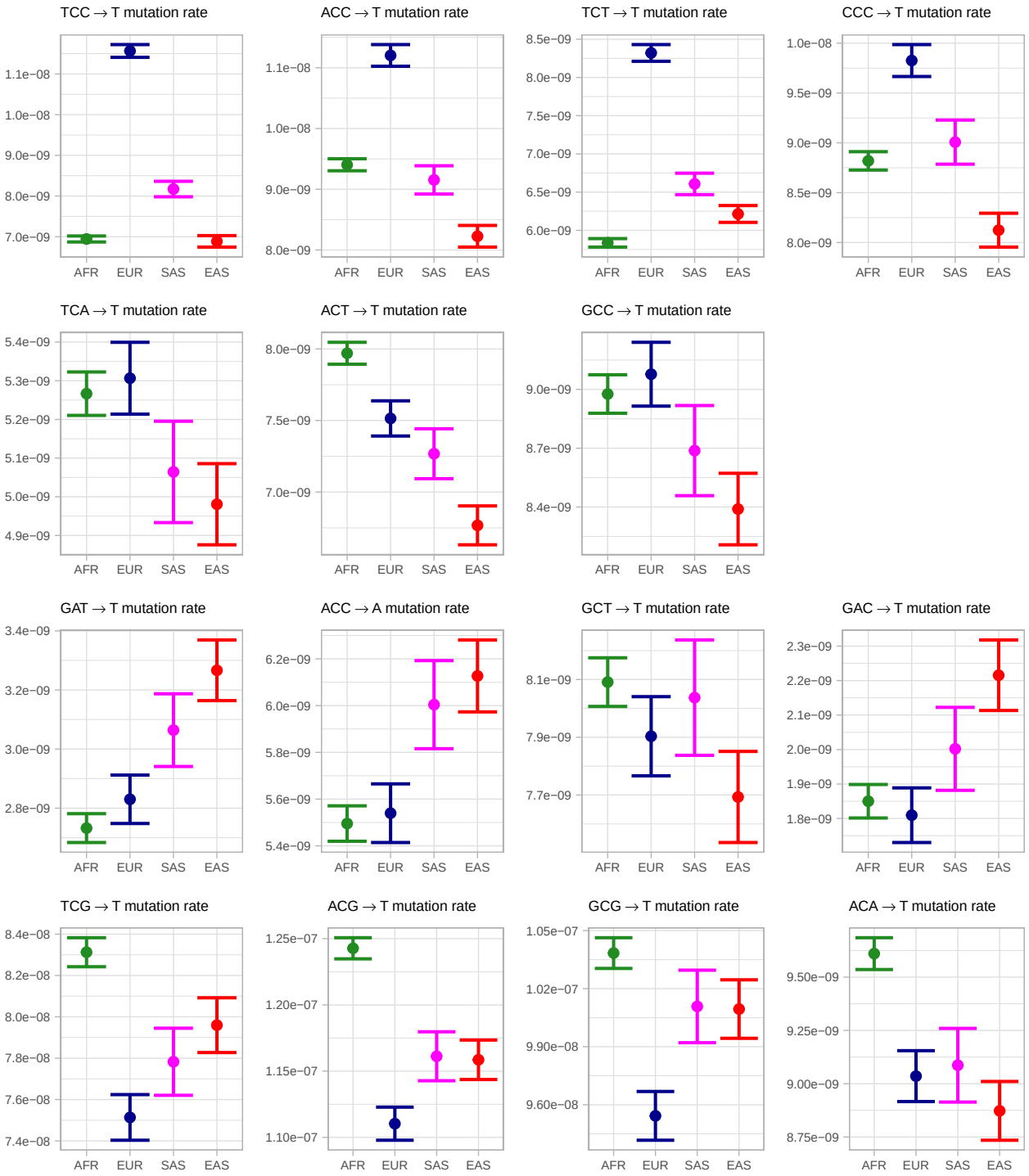
Here, we run the same p-ordered hypothesis test the we used on the 1,000 genomes dataset. However, rather than run the test for all 96 threemer contexts, we will avoid the multiple testing burden by testing for the significance of variation in only the top 15 polymorphism classes listed in Table 1.

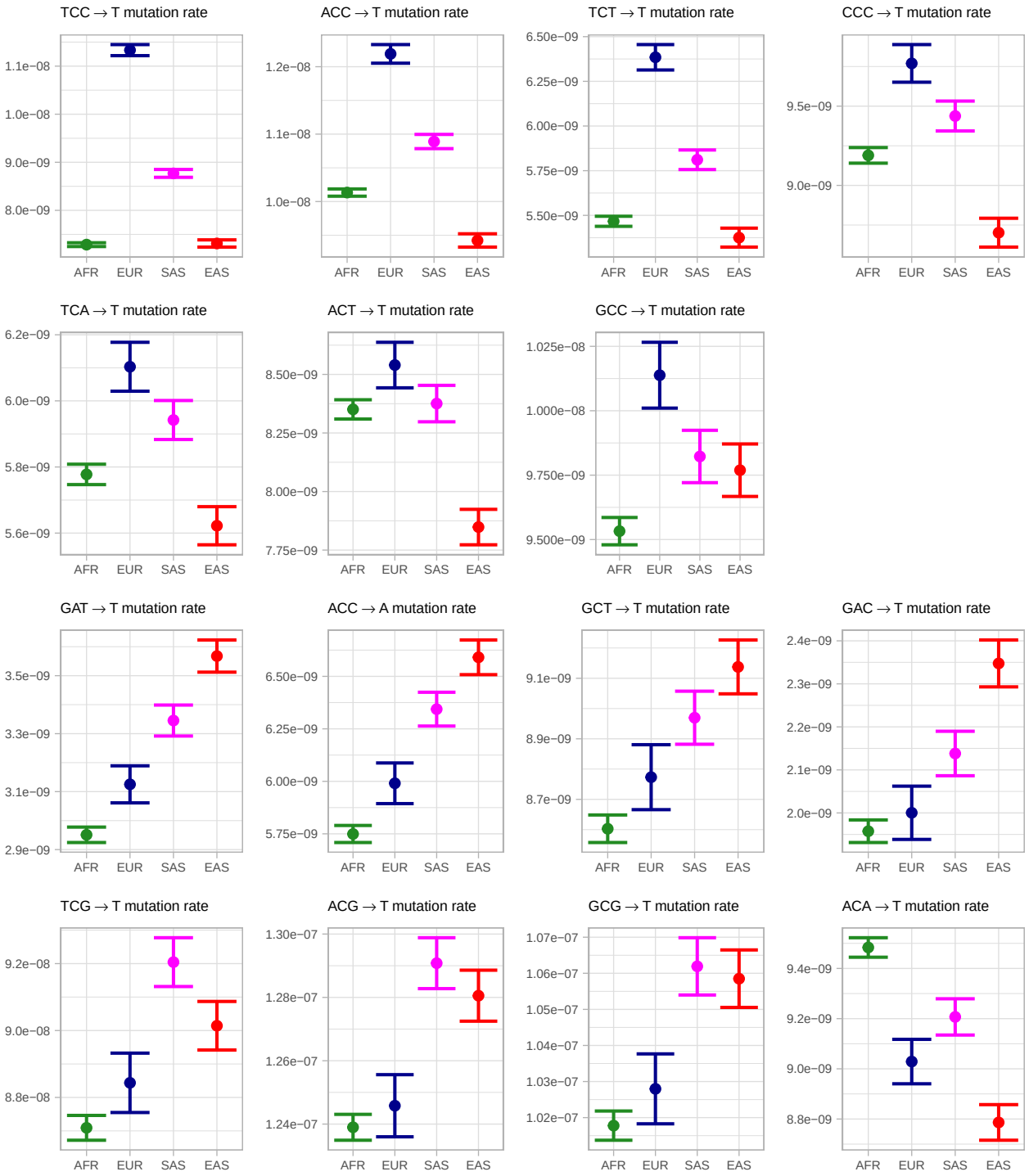
Table 3: Replication of Table 1 with data from SGDP

Context	AFR_relative_rate	EUR_relative_rate	SAS_relative_rate	EAS_relative_rate	p
TCT->T	1	1.43	1.13	1.06	0
TCC->T	1	1.67	1.18	0.99	0
ACC->T	1	1.19	0.97	0.87	1.09112283936947e-157
CCC->T	1	1.11	1.02	0.92	6.57420623915224e-68
ACT->T	1	0.94	0.91	0.85	2.040046983037e-50
ACG->T	1	0.89	0.93	0.93	1.23558962619583e-46
ACA->T	1	0.94	0.95	0.92	6.09656281752717e-31
TCG->T	1	0.90	0.94	0.96	3.44551514963053e-23
GAT->T	1	1.04	1.12	1.20	1.00489341040507e-18
GCG->T	1	0.92	0.97	0.97	1.3734869190287e-17
ACC->A	1	1.01	1.09	1.11	1.06988077060094e-13
GCC->T	1	1.01	0.97	0.93	3.73619463137433e-12
TCA->T	1	1.01	0.96	0.95	8.76482369347644e-12
GAC->T	1	0.98	1.08	1.20	2.16536478639532e-10
GCT->T	1	0.98	0.99	0.95	5.95545523910949e-05

All p-values are nominally significant. In addition, the relative infered mutation rates agree for many polymorphism types, with some exceptions:

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


TCC → T mutation rate
1.1e−08
1.0e−08
9.0e−09
8.0e−09
7.0e−09
AFR EUR SAS EAS
ACC → T mutation rate
1.1e−08
1.0e−08
9.0e−09
8.0e−09
AFR EUR SAS EAS
TCT → T mutation rate
8.5e−09
8.0e−09
7.5e−09
7.0e−09
6.5e−09
6.0e−09
AFR EUR SAS EAS
CCC → T mutation rate
1.0e−08
9.5e−09
9.0e−09
8.5e−09
8.0e−09
AFR EUR SAS EAS
TCA → T mutation rate
5.4e−09
5.3e−09
5.2e−09
5.1e−09
5.0e−09
4.9e−09
AFR EUR SAS EAS
ACT → T mutation rate
8.0e−09
7.5e−09
7.0e−09
AFR EUR SAS EAS
GCC → T mutation rate
9.0e−09
8.7e−09
8.4e−09
AFR EUR SAS EAS
GAT → T mutation rate
3.4e−09
3.2e−09
3.0e−09
2.8e−09
AFR EUR SAS EAS
ACC → A mutation rate
6.2e−09
6.0e−09
5.8e−09
5.6e−09
5.4e−09
AFR EUR SAS EAS
GCT → T mutation rate
8.1e−09
7.9e−09
7.7e−09
AFR EUR SAS EAS
GAC → T mutation rate
2.3e−09
2.2e−09
2.1e−09
2.0e−09
1.9e−09
1.8e−09
AFR EUR SAS EAS
TCG → T mutation rate
8.4e−08
8.2e−08
8.0e−08
7.8e−08
7.6e−08
7.4e−08
AFR EUR SAS EAS
ACG → T mutation rate
1.25e−07
1.20e−07
1.15e−07
1.10e−07
AFR EUR SAS EAS
GCG → T mutation rate
1.05e−07
1.02e−07
9.90e−08
9.60e−08
AFR EUR SAS EAS
ACA → T mutation rate
9.50e−09
9.25e−09
9.00e−09
8.75e−09
AFR EUR SAS EAS

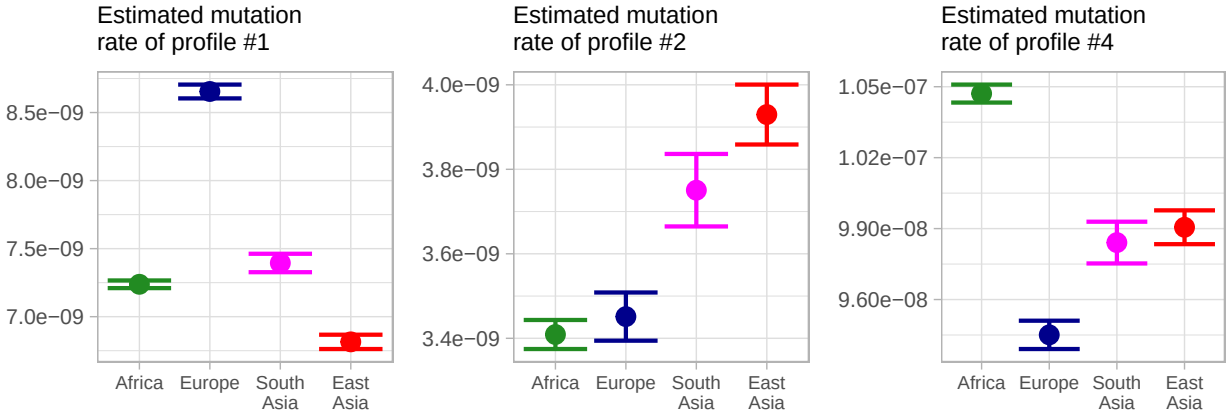

TCC → T mutation rate
1.1e−08
1.0e−08
9.0e−09
8.0e−09
AFR EUR SAS EAS
ACC → T mutation rate
1.2e−08
1.1e−08
1.0e−08
AFR EUR SAS EAS
TCT → T mutation rate
6.50e−09
6.25e−09
6.00e−09
5.75e−09
5.50e−09
AFR EUR SAS EAS
CCC → T mutation rate
9.5e−09
9.0e−09
AFR EUR SAS EAS
TCA → T mutation rate
6.2e−09
6.0e−09
5.8e−09
5.6e−09
AFR EUR SAS EAS
ACT → T mutation rate
8.50e−09
8.25e−09
8.00e−09
7.75e−09
AFR EUR SAS EAS
GCC → T mutation rate
1.025e−08
1.000e−08
9.750e−09
9.500e−09
AFR EUR SAS EAS
GAT → T mutation rate
3.5e−09
3.3e−09
3.1e−09
2.9e−09
AFR EUR SAS EAS
ACC → A mutation rate
6.50e−09
6.25e−09
6.00e−09
5.75e−09
AFR EUR SAS EAS
GCT → T mutation rate
9.1e−09
8.9e−09
8.7e−09
AFR EUR SAS EAS
GAC → T mutation rate
2.4e−09
2.3e−09
2.2e−09
2.1e−09
2.0e−09
AFR EUR SAS EAS
TCG → T mutation rate
9.2e−08
9.0e−08
8.8e−08
AFR EUR SAS EAS
ACG → T mutation rate
1.30e−07
1.28e−07
1.26e−07
1.24e−07
AFR EUR SAS EAS
GCG → T mutation rate
1.07e−07
1.06e−07
1.05e−07
1.04e−07
1.03e−07
1.02e−07
AFR EUR SAS EAS
ACA → T mutation rate
9.4e−09
9.2e−09
9.0e−09
8.8e−09
AFR EUR SAS EAS

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

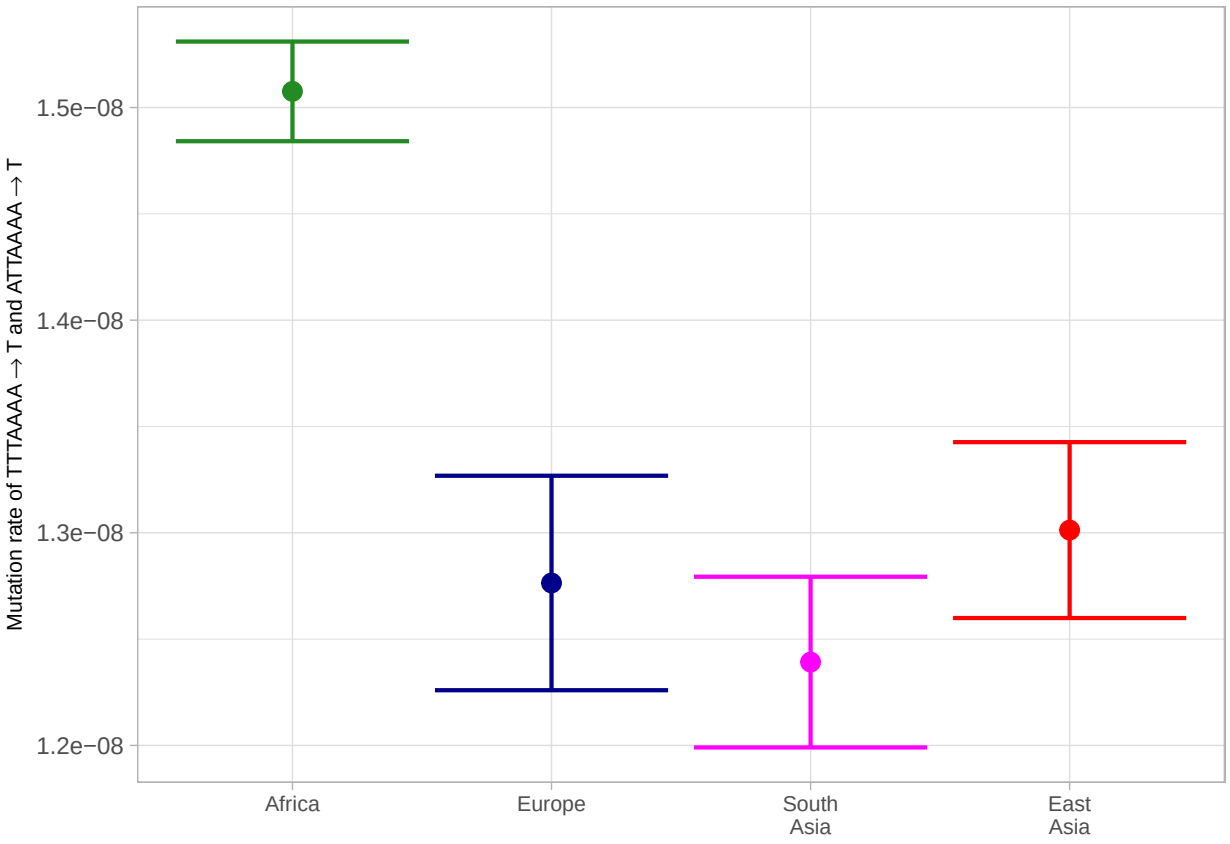
We will not attempt hypothesis testing across all 5-mer and 7-mer polymorphism classes, since many of these tests are sure to invlolve too few observations to be carried out, and the hypothesis testing burden would massively reduce statistical power. Rather, we will repeat the hypothesis tests for only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