Replication in Simons Genome Diversity Project

Rachael Caelie (Rocky) Aikens

10/29/2018

Contents

Introduction **1**

0. Basic Checks of SGDP Data **2**
0.1 Visualizing dataset agreement 2
0.2 Agreement with Harris and Pritchard 4
0.3 Sample sizes Under the 7-mer model 8

1. Identifying novel 3-mer substitution classes that vary across continents **9**
1.1 Test for homogenity across all continental groups (table 1) 9
1.2 Inferred rates of top 15 3mers (supplementary figure 2) 10
1.3 Summary 12

2. Heirarchical Clustering of 3-mer signatures **13**
2.1 Signatures 1-4 13
2.2 Signature 5:CpG transversions 14
2.3 Summary 15

3. Broader Sequence Contexts of 3-mer Signatures **15**

4. Signatures of Variation at Broader Sequence Contexts **16**

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	3.6	0.7	0.5	0.3

0. Basic Checks of SGDP Data

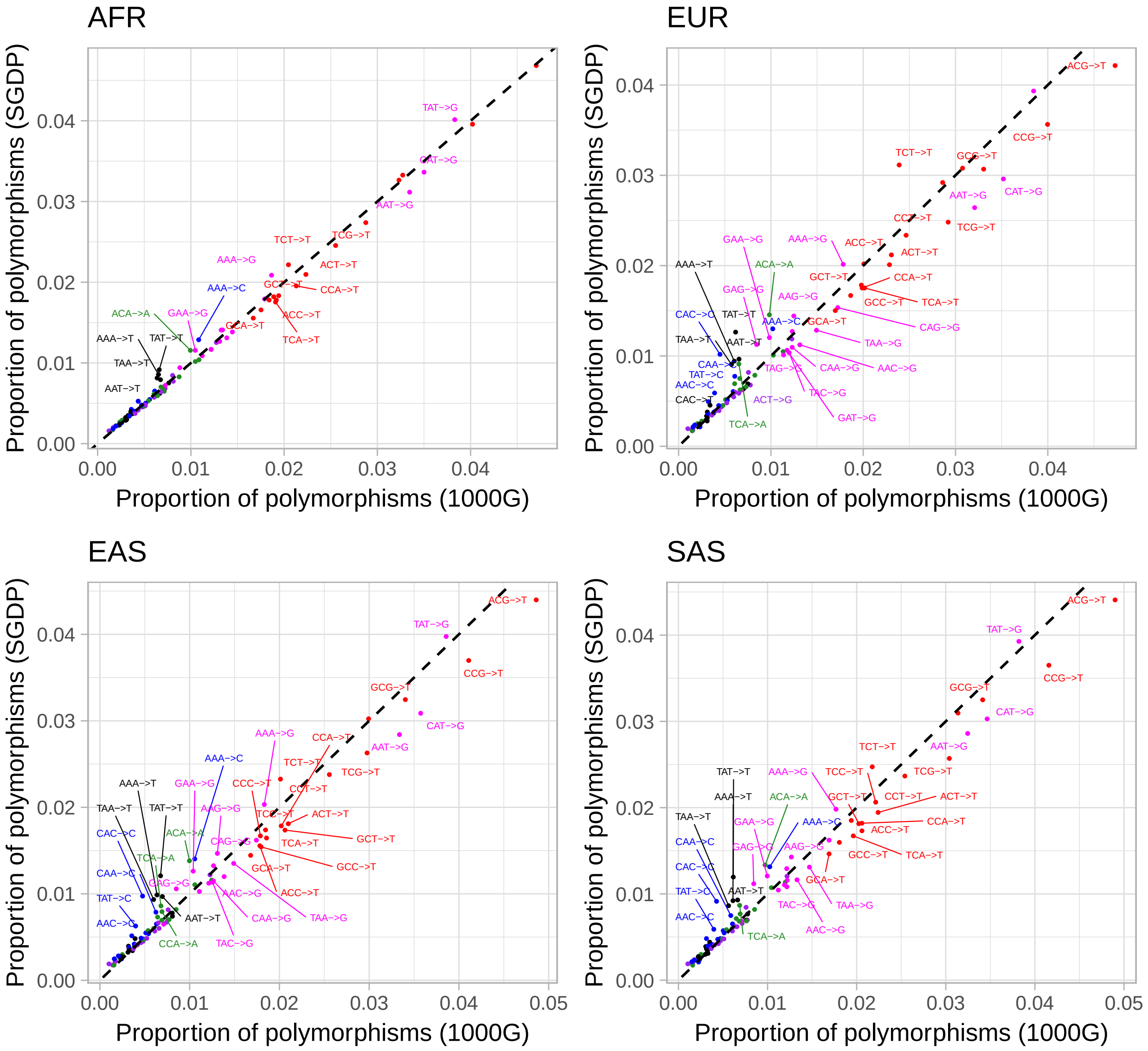
Here are some basic visualizations of the SGDP Data and their match up with 1KG and previous results.

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.



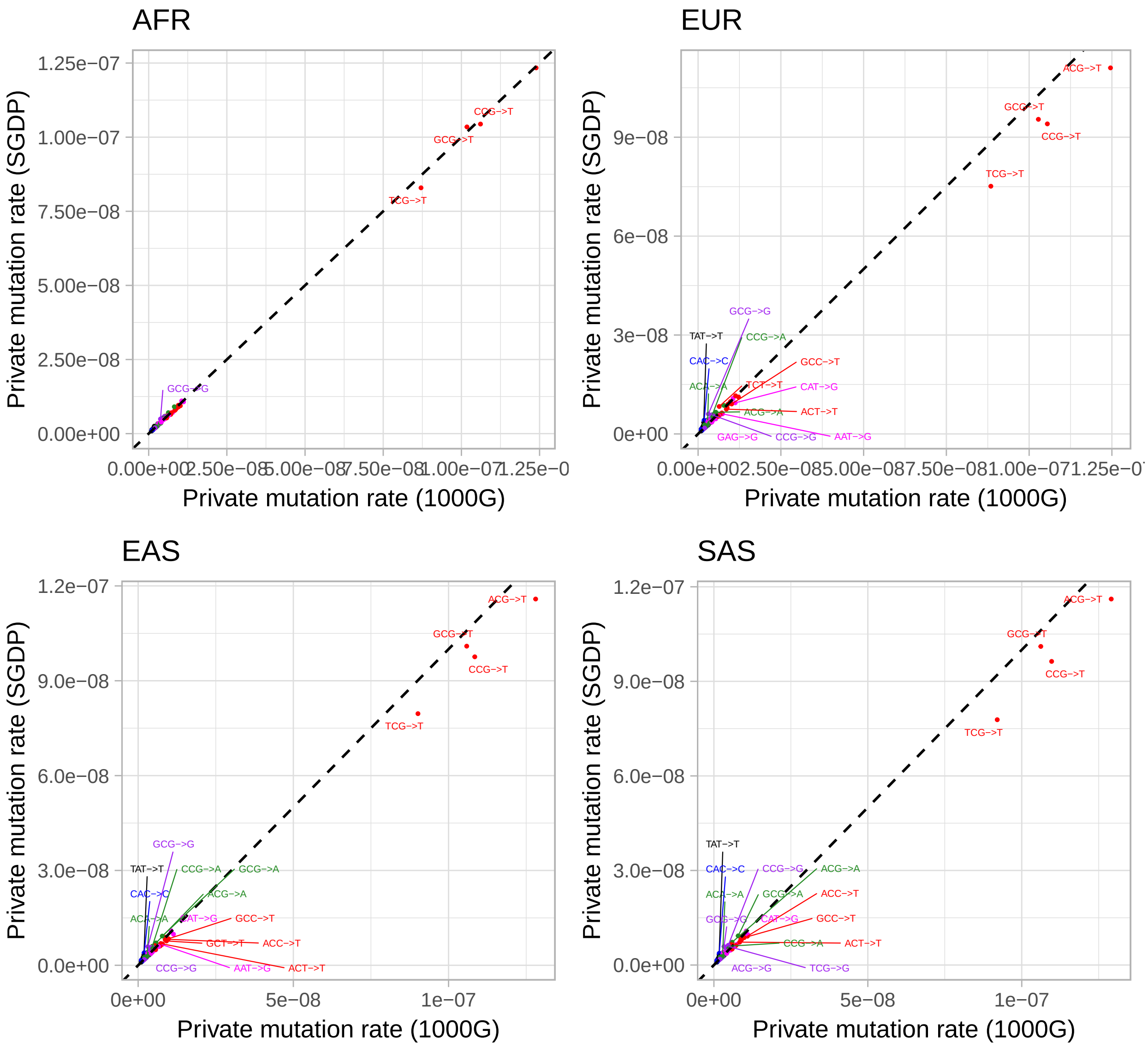
Qualitatively, there is relatively good agreement between the datasets. A few differences are noticeable:

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.

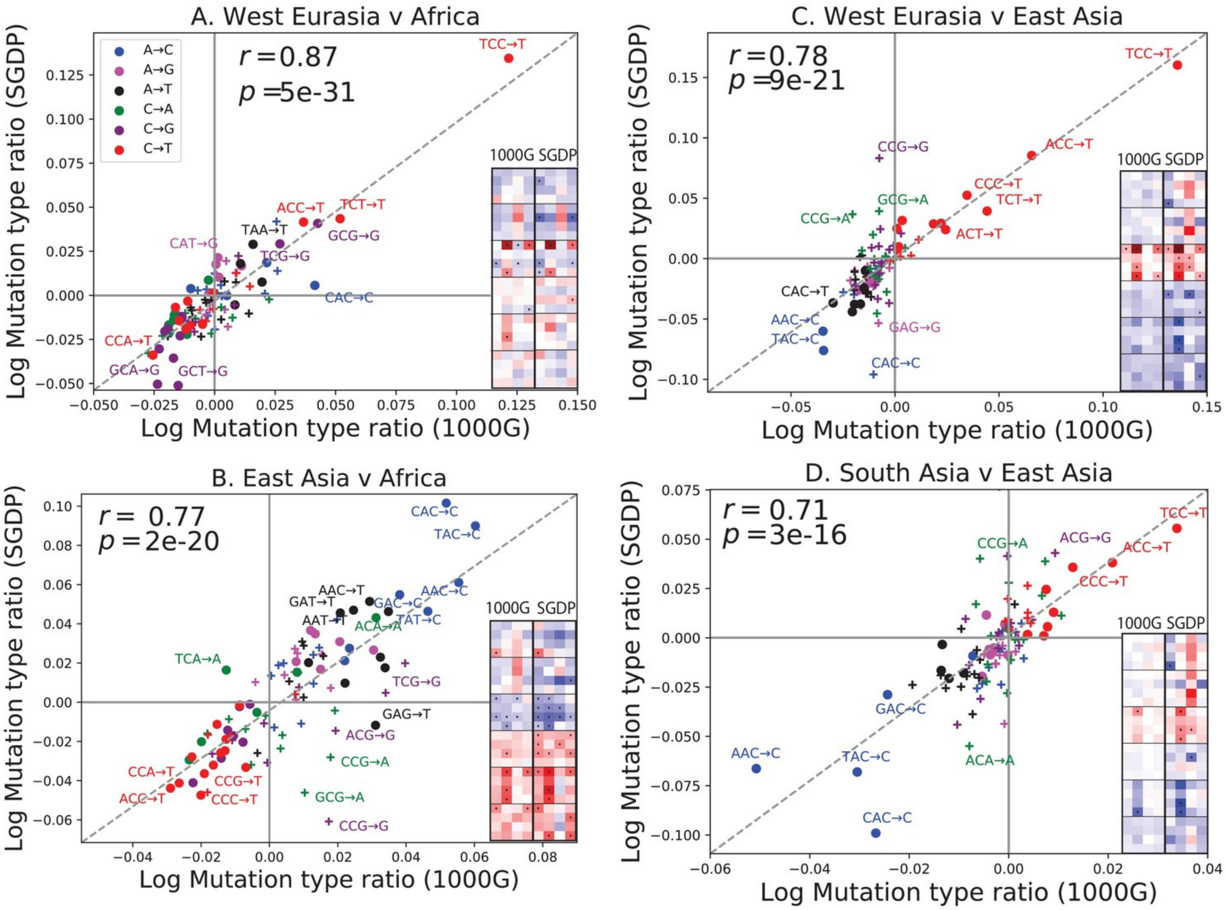
3) The following A→C substitutions are more abundant in SGDP in Europe, East Asia, and South Asia: CAA→C, AAA→C, CAC→C, and (somewhat) AAC→C. AAA→C also may be slightly more abundant in Africa in SGDP than in 1KG.

However, when we plot the agreement in mutation rate, it is immediately evident that the inferred private mutation rates of CpGs are much lower in SGDP than in 1KG in all populations except Africa



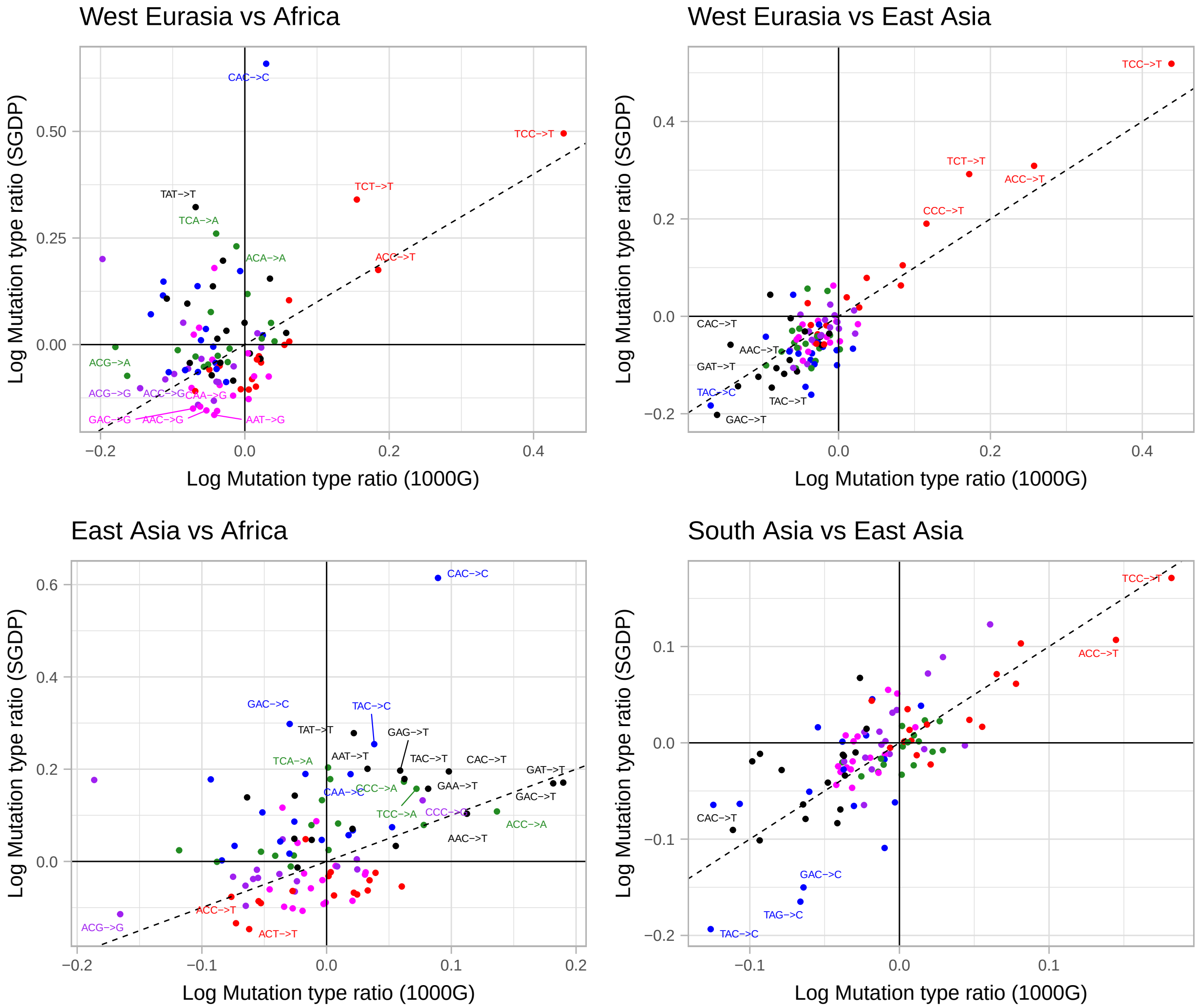
0.2 Agreement with Harris and Pritchard

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:



Note: Harris and Pritchard project the site frequency spectrum of the 1,000 genomes dataset onto the sample size of SGDP. It is not clear to me how this would change the results. We do not take this extra step.

Here is our attempt at the same plot:



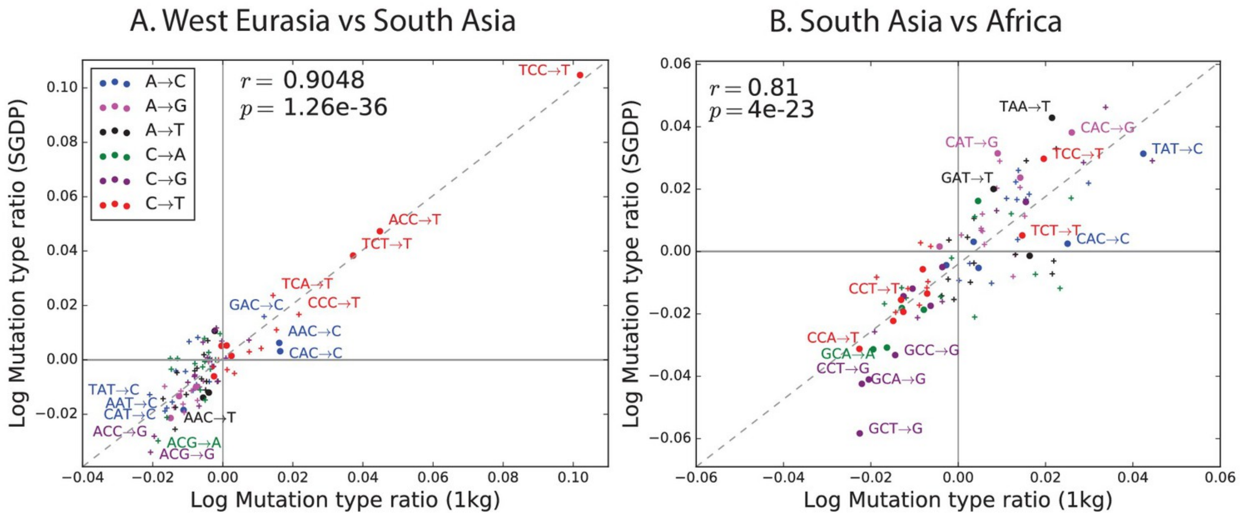
West Eurasia vs Africa

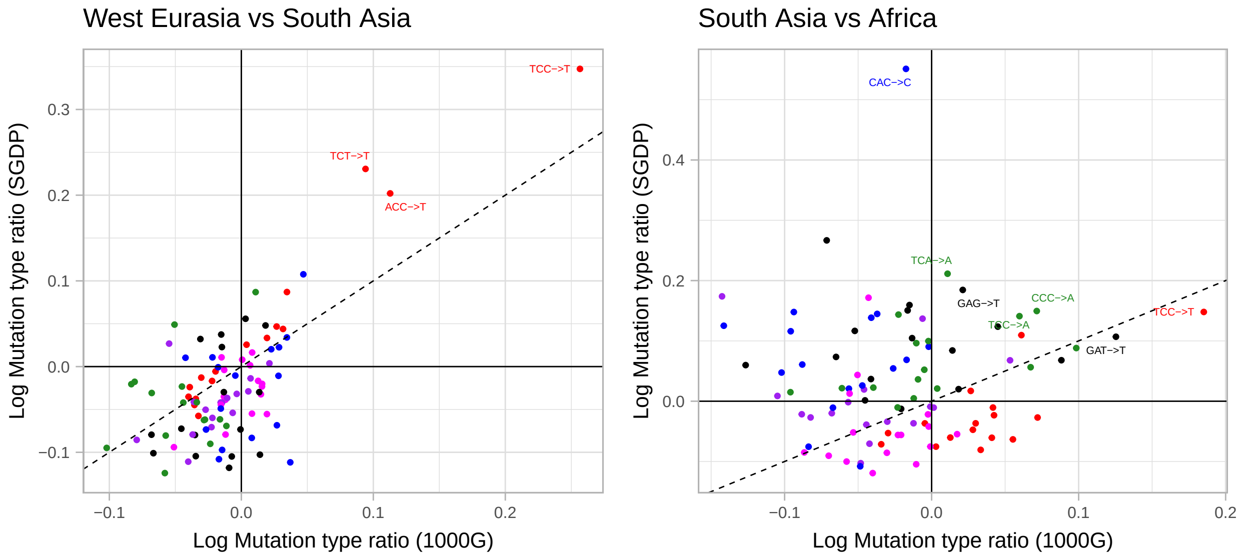
West Eurasia vs East Asia

East Asia vs Africa

South Asia vs East Asia

Also included are two supplementary comparisons:



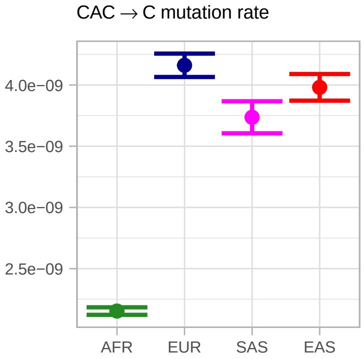


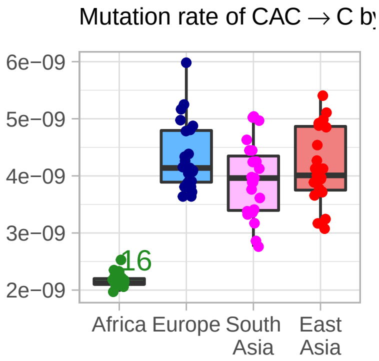
I make the following observations:

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.

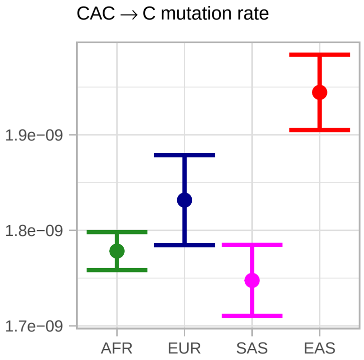
2. Whenever we make a comparison with Africa, the figures do not agree.

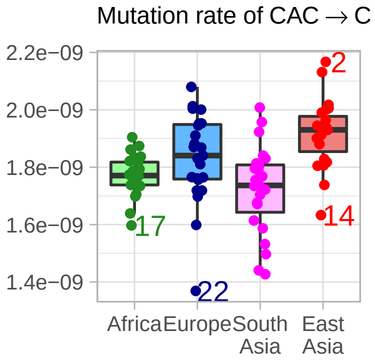
3. CAC -> C is consistently an outlier in comparisons with Africa. Here are the rates in SGDP:





And here they are in 1KG

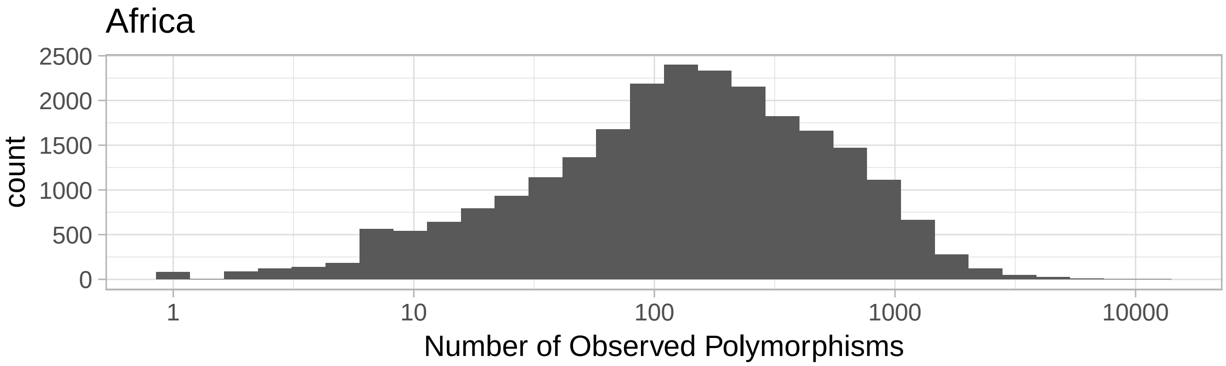


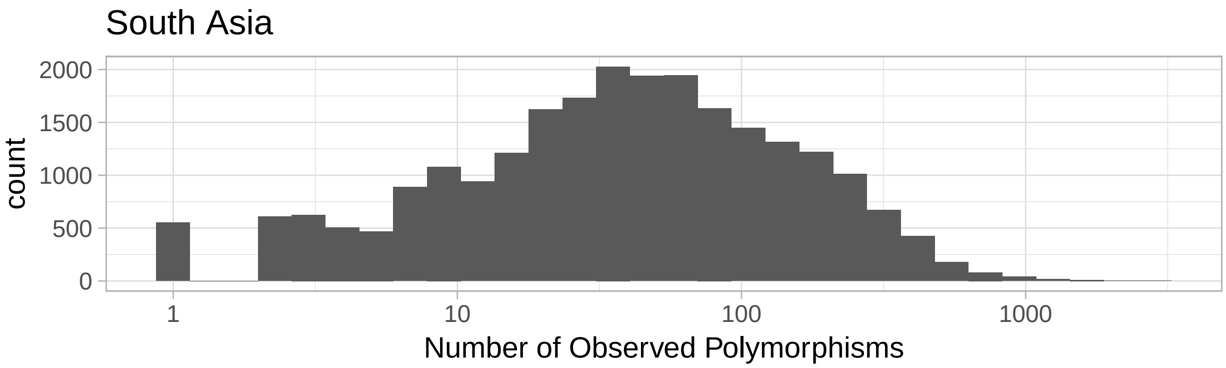


Looking through Harris and Pritchard's figures, their relative rates for CAC->C are: EAS ∼ EUR > SAS > AFR.

0.3 Sample sizes Under the 7-mer model

Before attempting to replicate any 7-mer analyses, we can first consider how many polymorphisms we have for each 7mer type in SGDP. Below are histograms of the number of observed polymorphisms for each of the 24,576 7-mer polymorphism classes. *Note that the x axis is on a log base 10 scale.* Africa and South Asia are shown as examples because they have the largest and smallest number of variants, respectively


Africa


South Asia

Clearly, while some 7-mers are fairly well-represented (>1000 polymorphisms), many are not (<100 polymorphisms). For these reasons be must be extra cautious when considering power.

1. Identifying novel 3-mer substitution classes that vary across continents

The major result for this section is Table 1, with supplementary figures 2 and 3. They are recapitulated below.

1.1 Test for homogenity across all continental groups (table 1)

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
TCC->T	1	1.64	1.16	0.98	0
ACC->T	1	1.19	0.97	0.87	2.30897118680569e-161
TCT->T	1	1.41	1.12	1.05	0
CCC->T	1	1.11	1.02	0.92	6.22382896953434e-68
TCG->T	1	0.91	0.94	0.96	1.48930212591532e-25
ACG->T	1	0.90	0.94	0.94	1.19530354955357e-43
GCG->T	1	0.92	0.98	0.98	4.34008861260168e-15
GAT->T	1	1.03	1.11	1.19	1.44648562849783e-23
ACC->A	1	1.01	1.09	1.11	4.69066491089771e-17
ACA->T	1	0.94	0.95	0.93	2.62241982201823e-30
TCA->T	1	1.00	0.95	0.94	1.66303149702796e-13
ACT->T	1	0.96	0.93	0.86	1.19404601491732e-45
GCT->T	1	0.97	0.99	0.95	7.89314714556592e-07
GAC->T	1	0.97	1.07	1.18	5.57326596649956e-10
GCC->T	1	1.01	0.96	0.93	3.29796264795211e-14

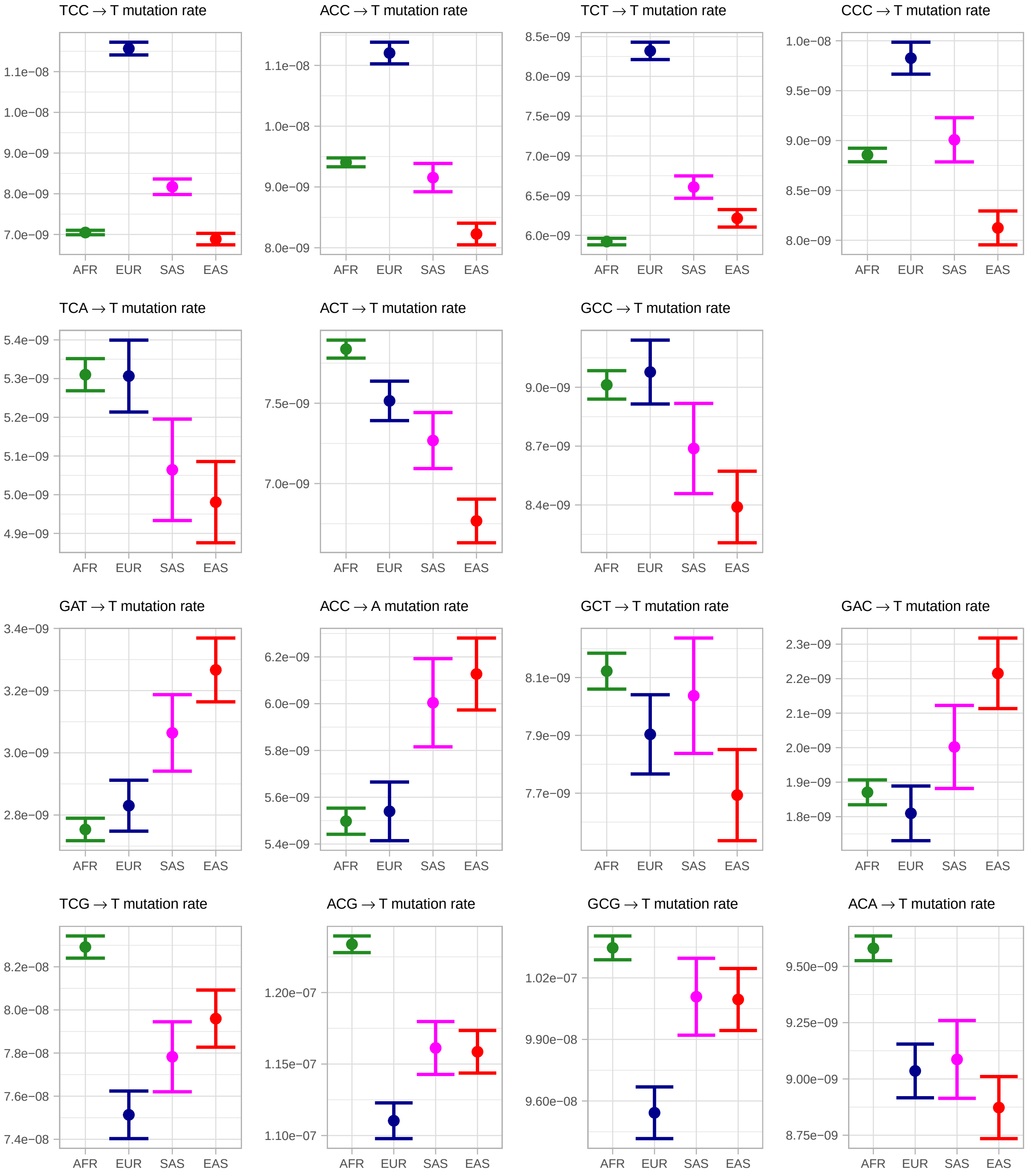
All p-values are nominally significant. In addition, the relative infered mutation rates agree for many polymorphism types, with some exceptions. Agreements and discrepancies are listed below:

Context	Agreement	Discrepancies	Notes
TCC->T	Agree		Prev. Reported C->T in EUR
ACC->T	Mostly Agree	SAS, AFR relationship	Prev. Reported C->T in EUR
TCT->T	Mostly Agree	EAS, AFR relationship	Prev. Reported C->T in EUR
CCC->T	Agree		Prev. Reported C->T in EUR
TCG->T	Disagree	Completely different	CpG
ACG->T	Mostly Disagree	AFR high in SGDP	CpG
GCG->T	Mostly Disagree	AFR high in SGDP	CpG
GAT->T	Agree		Not previously reported; Signature 2
ACC->A	Agree		Not previously reported; Signature 2
ACA->T	Mostly Agree	EUR, SAS relationship	Not previously reported; Not in a signature
TCA->T	Somewhat Agree	AFR high in SGDP	Not previously reported; Signature 1
ACT->T	Somewhat Agree	AFR high in SGDP	Not previously reported; Signature 1
GCT->T	Disagree	Completely different	Not previously reported; Not in a signature
GAC->T	Agree		Not previously reported; Signature 2
GCC->T	Somewhat Agree	AFR high in SGDP	Not previously reported; Signature 1

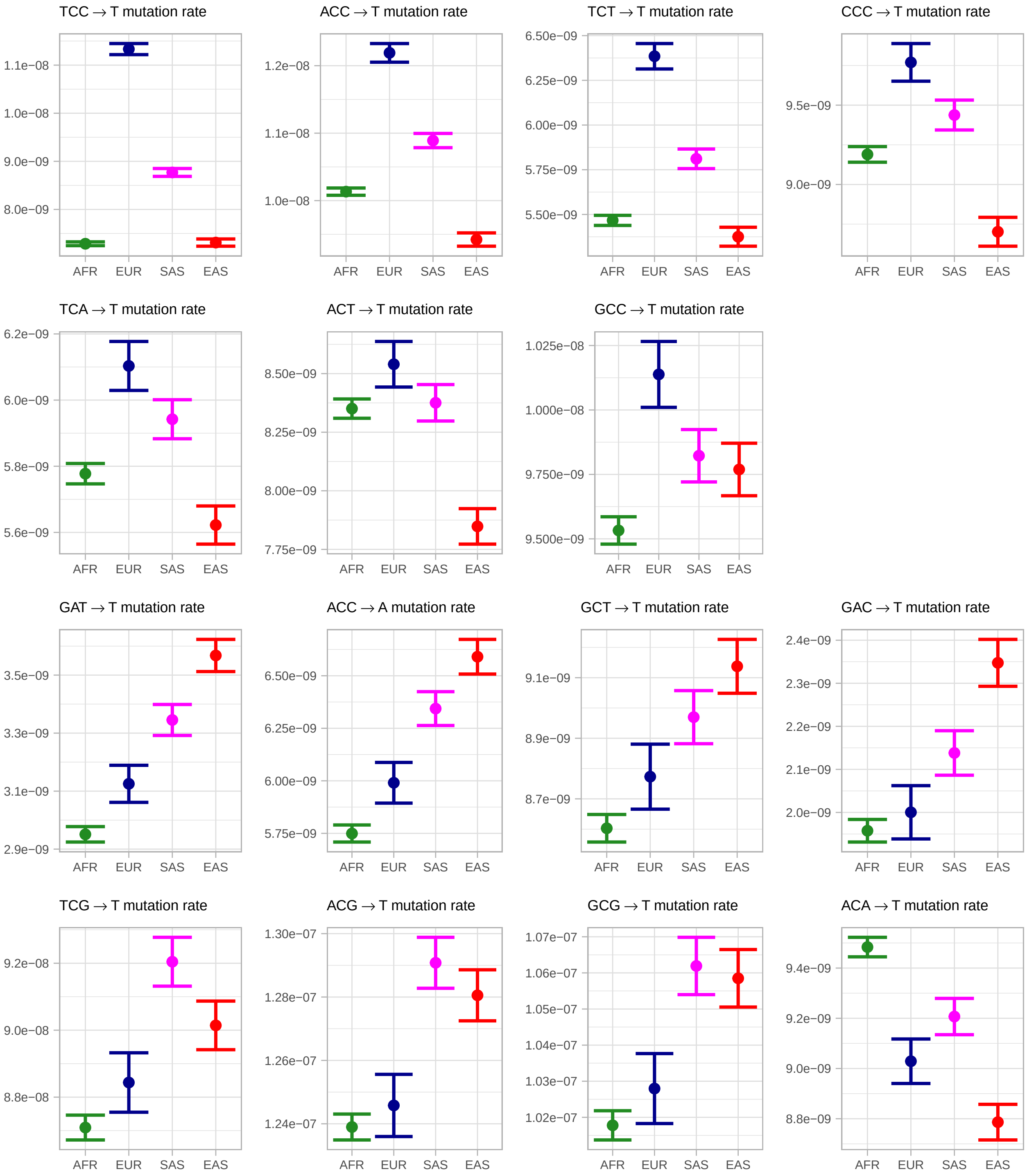
1.2 Inferred rates of top 15 3mers (supplementary figure 2)

To better visualize the differences between datasets, here are the inferred relative rates of the substitutions from the tables in the section above.

Relative Rates in SGDP



Relative Rates in 1KG



TCC → T mutation rate

ACC → T mutation rate

TCT → T mutation rate

CCC → T mutation rate

TCA → T mutation rate

ACT → T mutation rate

GCC → T mutation rate

GAT → T mutation rate

ACC → A mutation rate

GCT → T mutation rate

GAC → T mutation rate

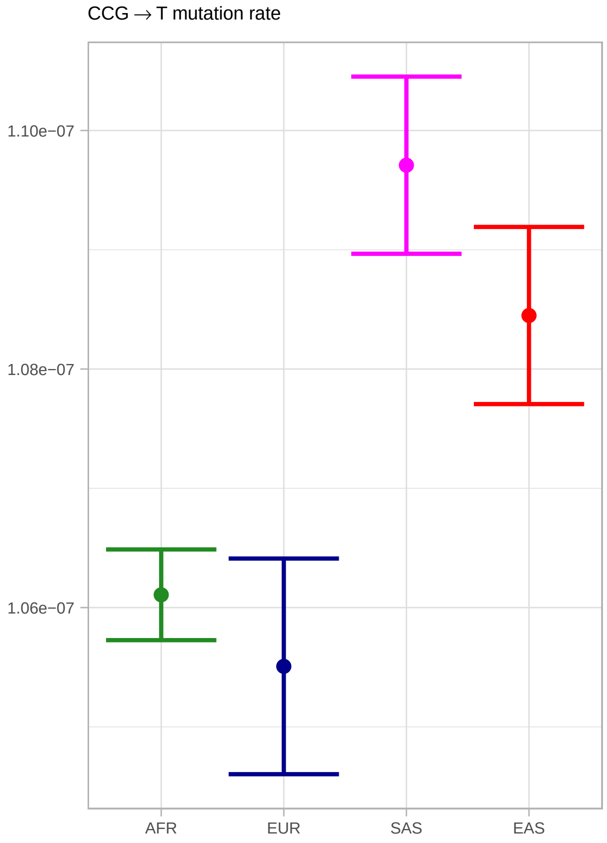
TCG → T mutation rate

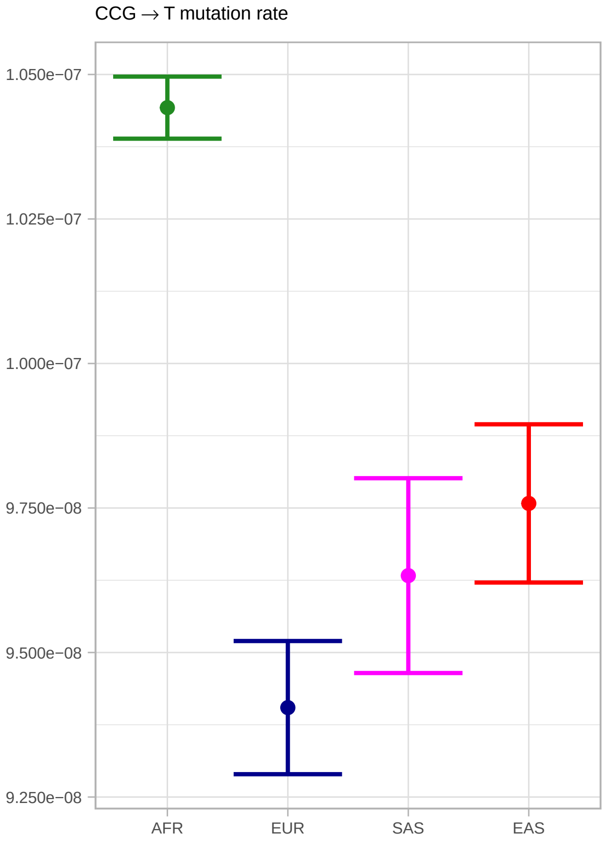
ACG → T mutation rate

GCG → T mutation rate

ACA → T mutation rate

For completenes, below are the CI plots for the last CpG 3-mer, CCG→T (1kg on the right, SGDP on the left). They also do not agree.





1.3 Summary

Most of the 3-mers from table 1 appear acceptably similar between SGDP and 1KG, allowing for some expected differences based on sampling and data collection discrepansies. The notable exceptions are (1) all three CpGs, and (2) GCT→T. GCT→ T is not in any signature discribed in results section 2, however in 1kg it does resemble signature 2.

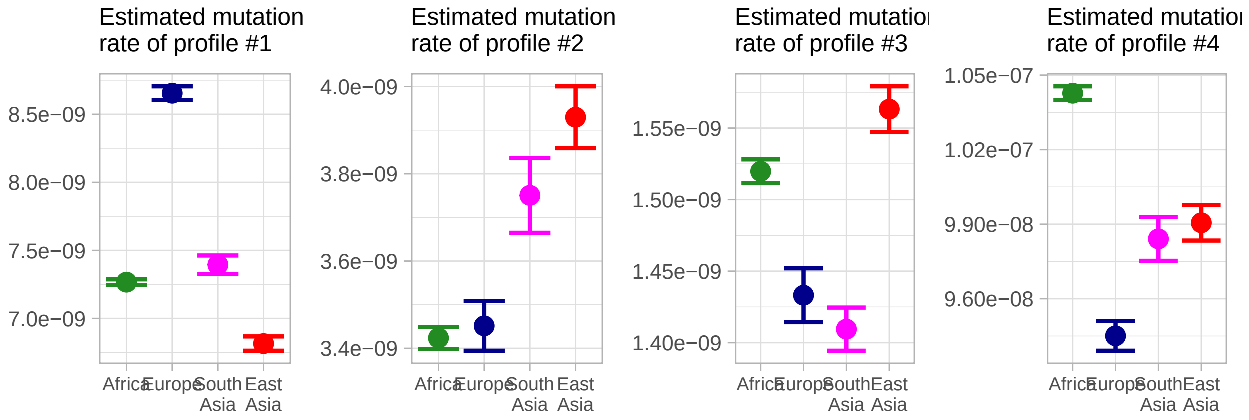
It is additionally is worth noting that while the approximate ranges of inferred mutation rate do agree between datasets, the actual numbers do not (see CI plots of inferred mutation rate, above). In general, when the rates disagree, African mutation rates in SGDP are closer to those in 1kg, while the mutation rates in the other populations are more discordant.

2. Heirarchical Clustering of 3-mer signatures

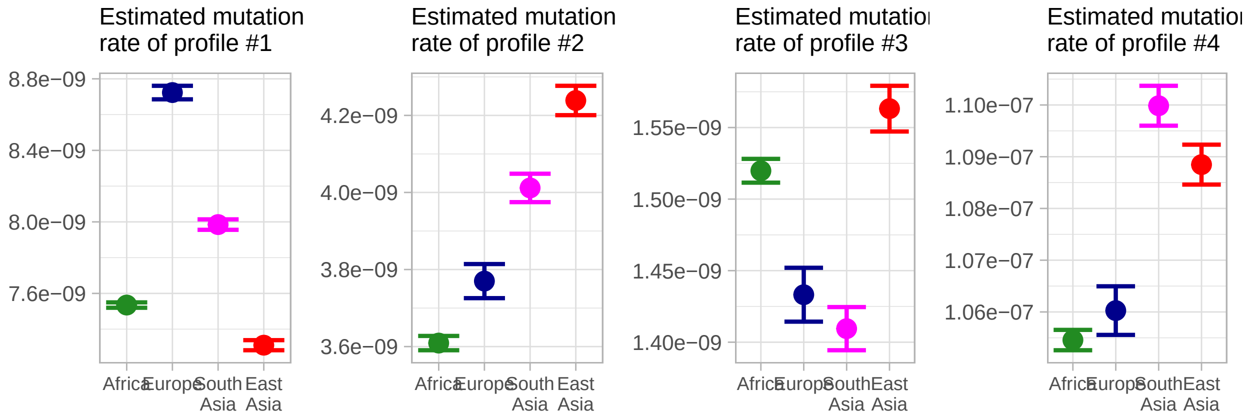
The primary result of this section of the paper was the heirarchical heatmap and the plots of mutation rate by signature. We will not attempt to construct a heatmap of the 3-mer signatures from the SGDP data, since this analysis on 1,000 genomes was mostly heuristic. However, we will plot the inferred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

2.1 Signatures 1-4

Mutational Signatures in SGDP



Mutational Signatures in 1kg

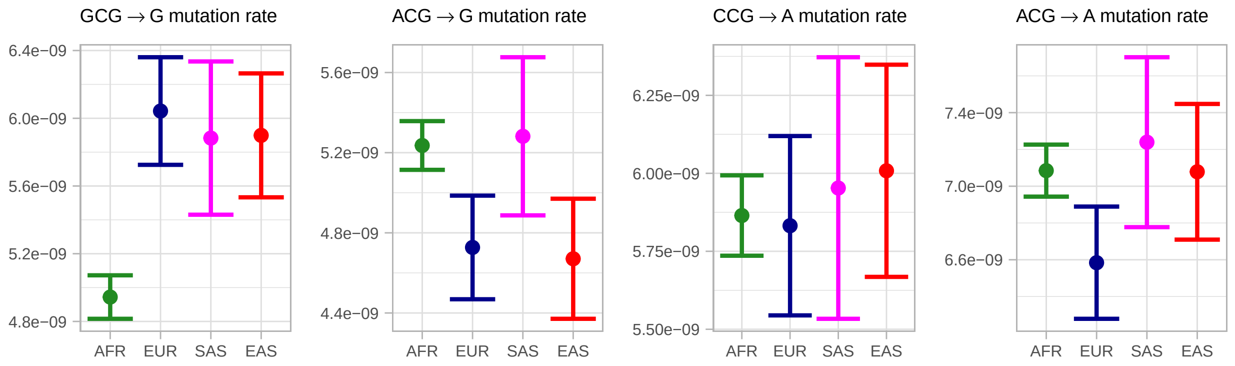


2.2 Signature 5:CpG transversions

We originally excluded signature 5 (CpG transversions) because Harris and Pritchard reported that the ratios of enrichment of CpG transversions do not agree between 1,000 genomes and SGDP. Since we now have an analysis of SGDP, we can compare this signature between datasets ourselves:

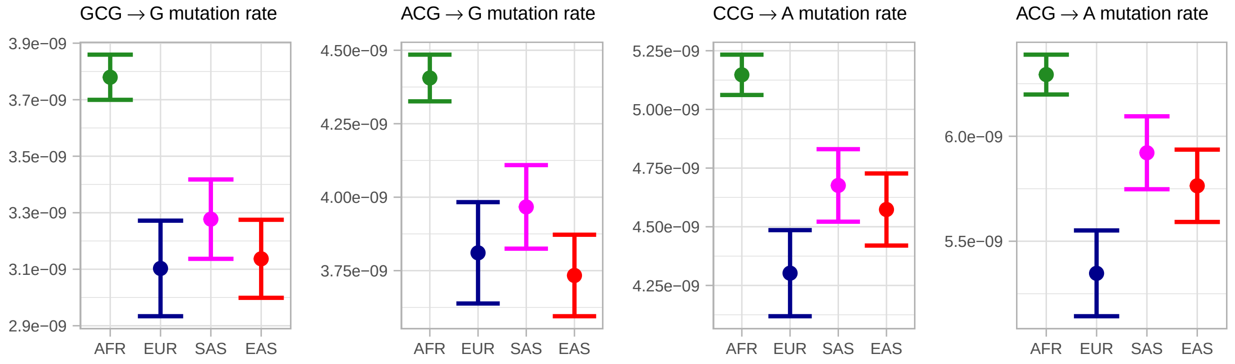
First, here are plots of rate for each of the subcontexts:

In SGDP:



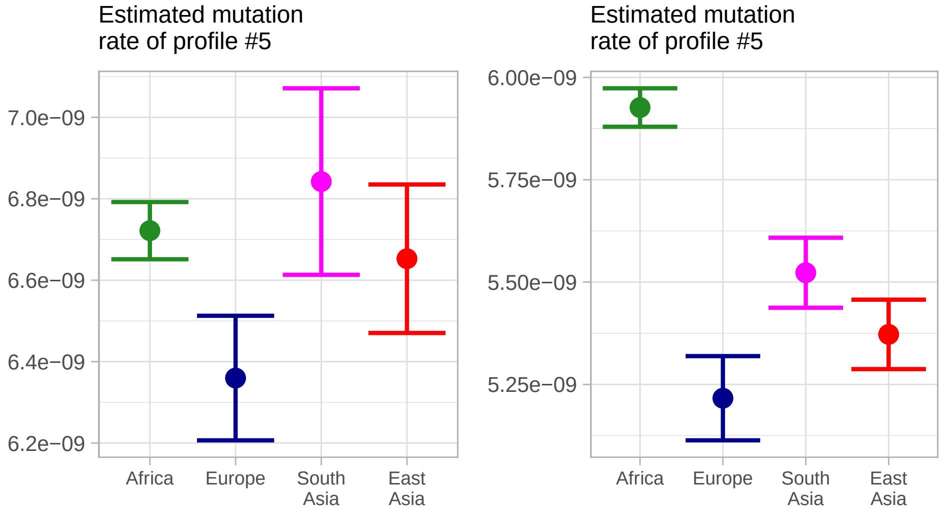
Frankly, these just seem noisy. This could be because transitions at CpG sites are rare.

In 1,000 genomes:



Pooled signature:

Second, we can look at the CI plot for the whole signature (SGDP on the left, 1KG on the right):

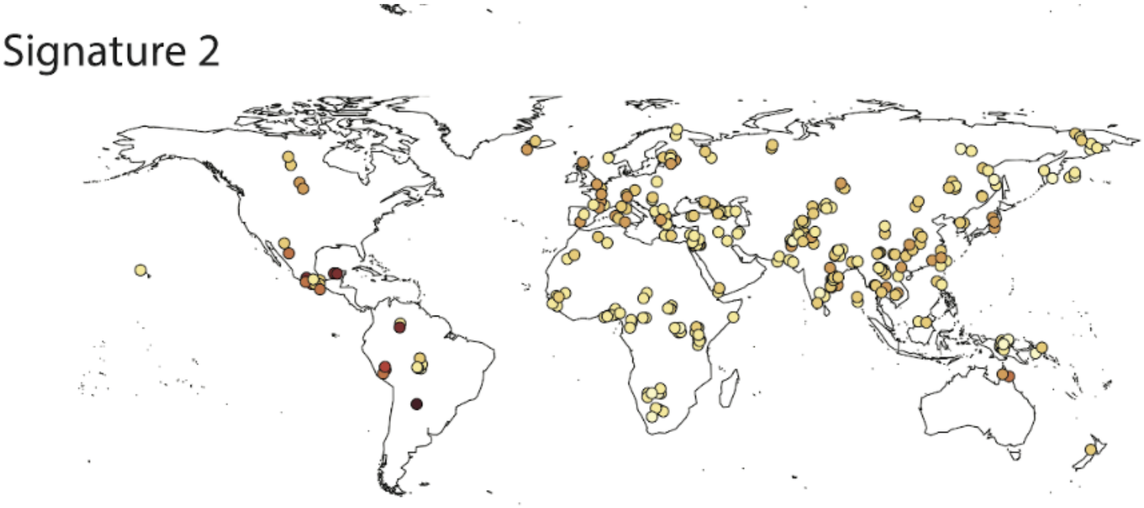


2.3 Summary

Signature 1 (known C→T), 2 (new Asia enriched), and 3 (*AC→C Japan enriched) appear to replicate. Signature 4 (CpGs) does not. The rates of CpG mutation we estimate from SGDP are much lower than in 1KG; it is not clear why. Here are all the previous analyses of CpGs I know of:

1) Harris and Pritchard: Not all comparisons involving each CpG is significant. The patterns of depletion and enrichment disagree slightly between (a) 1,000 genomes, (b) SGDP, and (c) 1,000 genomes projected onto the sample size of SGDP. For example, in (a) all CpGs are significantly *depleted* in Europe relative to East Asia while in (b) two CpGs are significantly *enriched* in Europe relative to East Asia. In (b) and (c) Africa appears to have the highest rate of CpGs, while in (a) some CpGs are enriched in East Asia relative to Africa. In part, their analysis in 1kg and SGDP agree with ours, although it is difficult to tell precisely because of differences in our visualizations.

2) Mathieson and Reich also report a CpG signature, enriched in Native American populations in SGDP. However, they conclude that this signature is driven by recent population increase, rather than a change in the underlying mutation rate. They add that because the rate of CpG mutation is so much higher than most, this mutation type is more sensitive to differences in demographic history between populations. Although this is not explicitly reported, their plot of enrichment for this signature (signature 2, below) appears depleted in Africa. (This is confusing because Harris and Pritchards analaysis of SGDP and our own found CpGs enriched in Africa.)



3. 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 (See Section 0.3). 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.

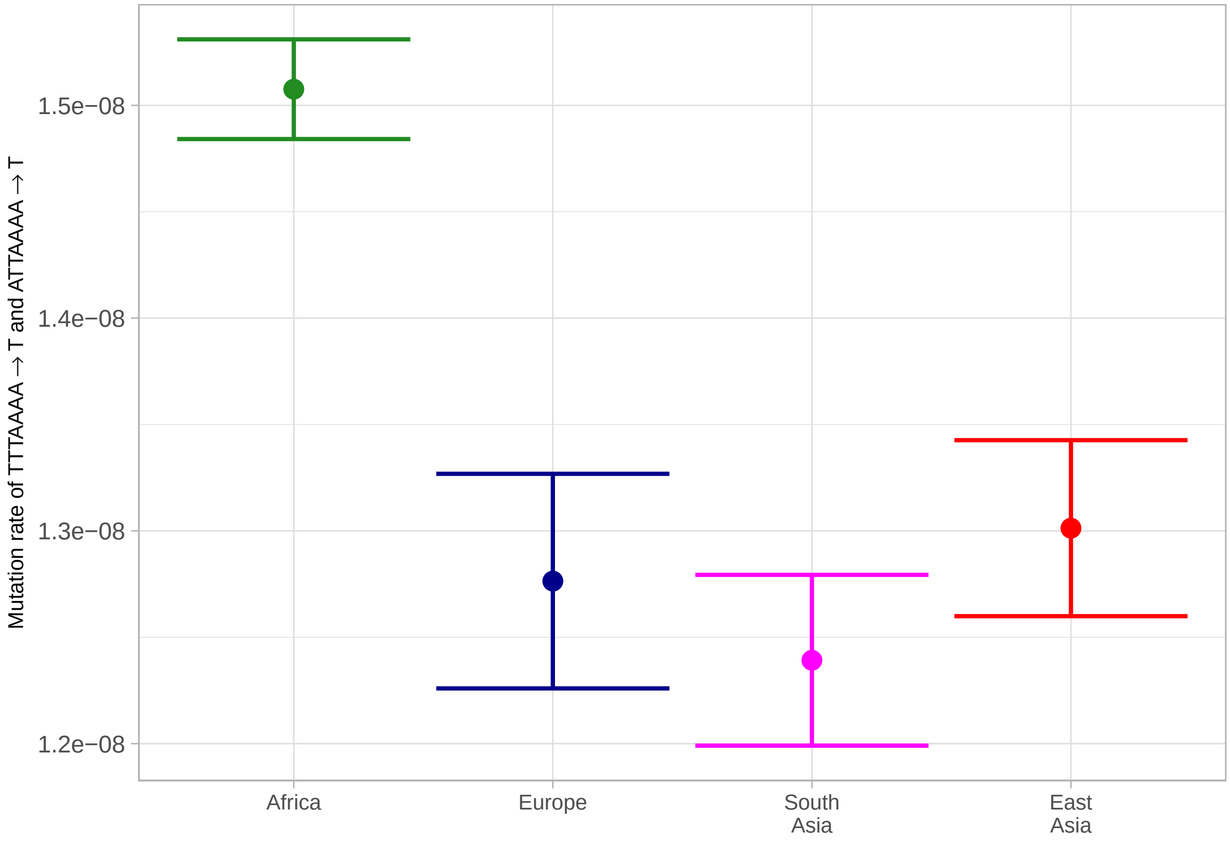
4. Signatures of Variation at Broader Sequence Contexts

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 5: Table 3 recalculated with SGDP data

Context	AFR relative rate	EUR relative rate	SAS relative rate	EAS relative rate	p
AAACAAA->A	1	0.89	0.96	1.14	2.1102236232382e-21
TTTATTT->T	1	1.01	1.03	1.02	2.16680412196906e-25
ATTAAAA->T	1	1.41	1.24	1.05	1.96852149154353e-21
CAAACCC->C	1	0.49	0	0.65	2.98475200264798e-39
TTTAAAA->T	1	1.21	1.18	1.6	1.19991229881911e-21
AAACAAA->A	1	1.16	1.01	0.95	2.1102236232382e-21
TTTATTT->T	1	1.21	0.91	0.87	2.16680412196906e-25

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