Supplementary Notes

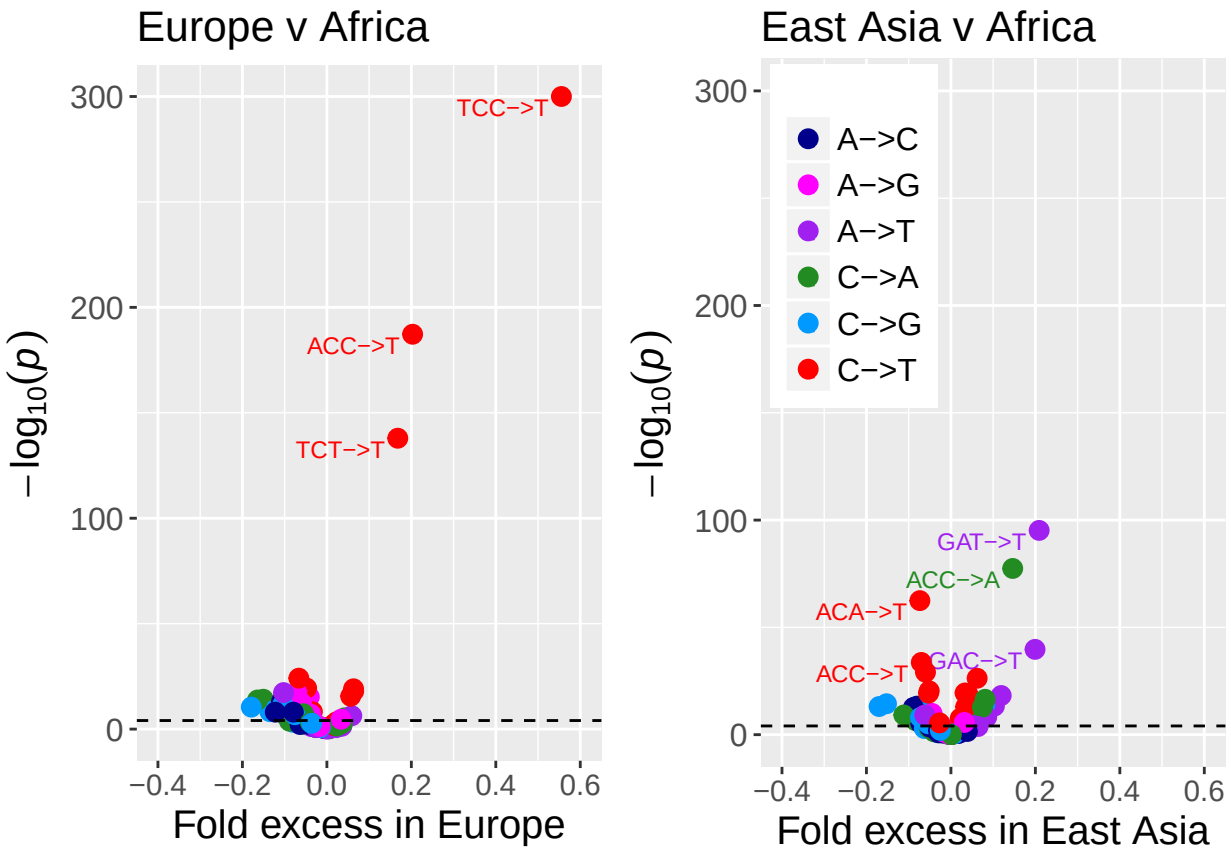
Rachael C. Aikens, Benjamin F. Voight

March 27, 2018

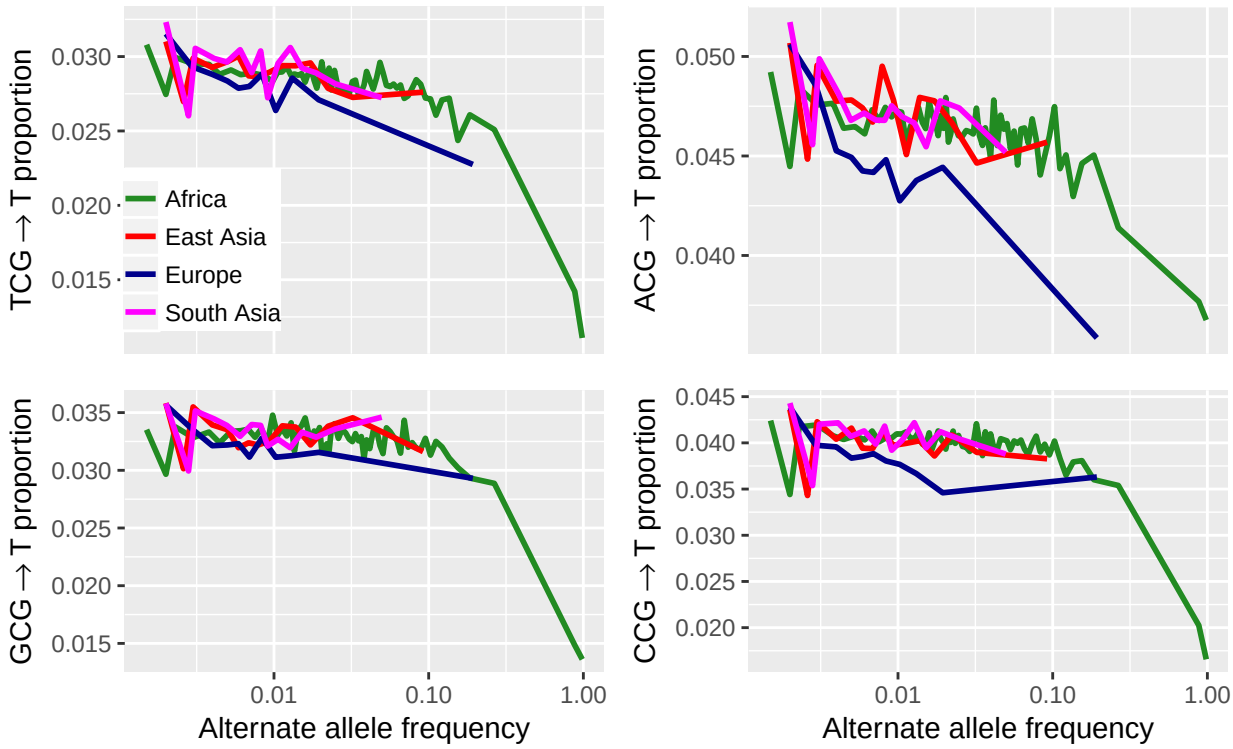
Supplementary Tables

Supplementary Table 1: All (A) 3-mer, (B) 5-mer, and (C) 7-mer subsitutions significantly heterogeneous across continental groups after ordered p-value correction. Relevant sub-contexts, counts in each population, and total number of contexts in the included regions of the genome are shown for each significant substitution type. Substitutions are listed in increasing order of significance.

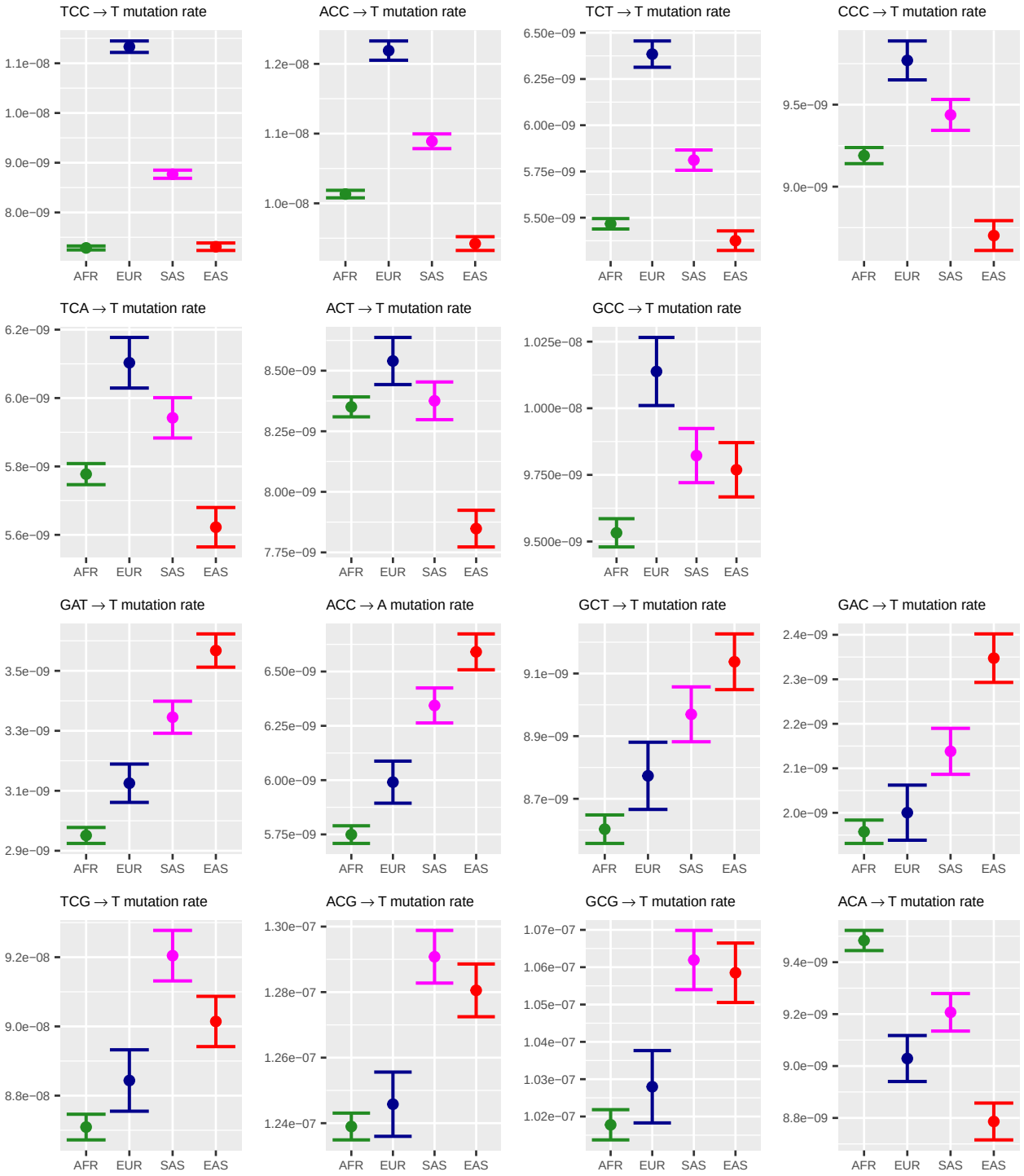
Supplementary Figures



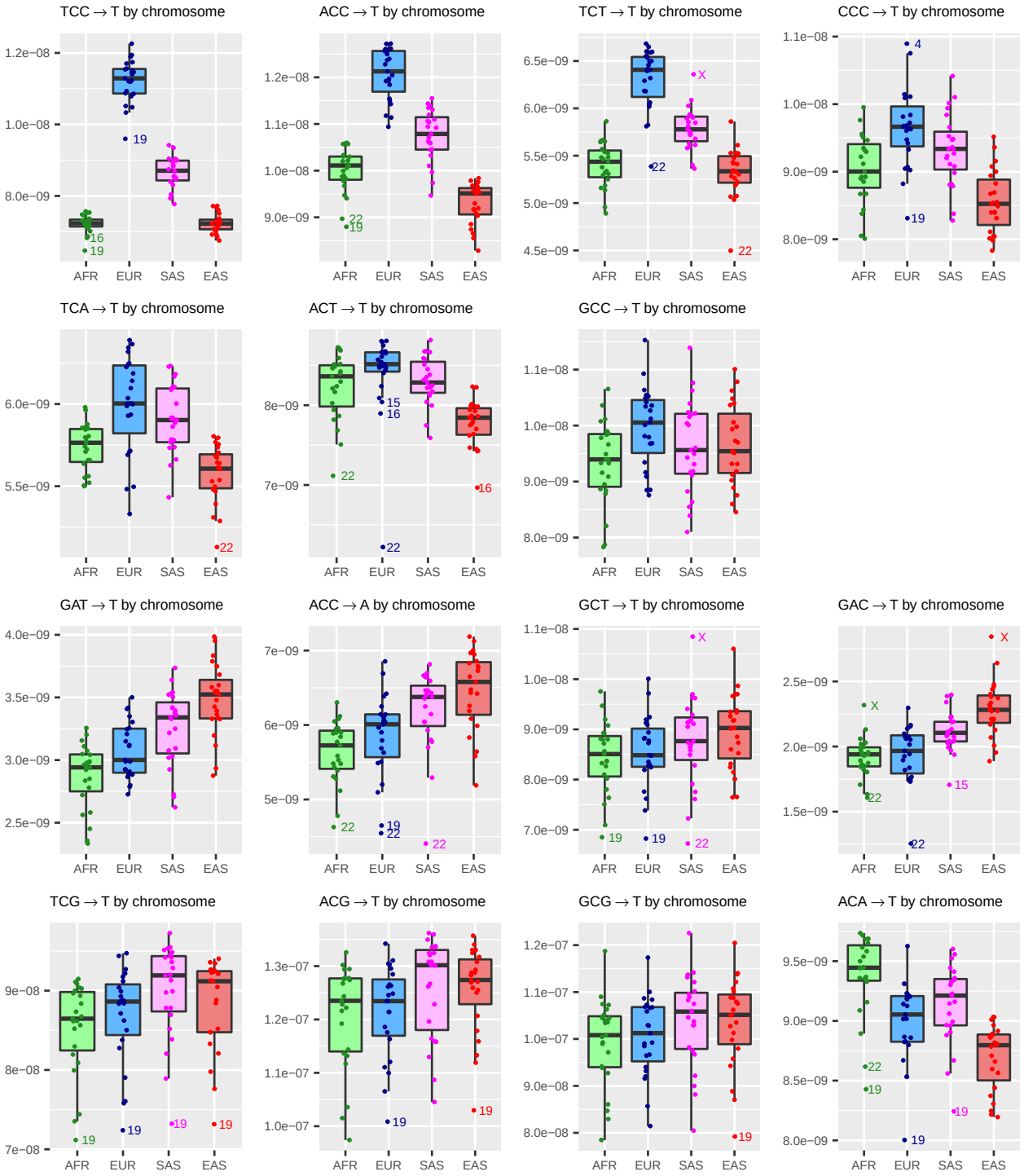
Supplementary Figure 1: *Replication of Figure 1 from Kelley Harris, 2015.* Using our variant filtration pipleline for extracting population-private variants from the Phase III 1,000 Genomes Dataset, we recapitulate Harris's previous findings from analysis of the Phase I Release. In keeping with Harris, 2015, the proportions of private 3mer substitutions were compared between pairs of populations using a pairwise chi-squared test. Ordered p-value correction was not applied. Unlike Harris, we choose to consider reverse-complimentary substitution classes as identical (e.g. TCC→TTC and GGA→ GAA are considered equivalent). The p-value for TCC→T in Europe versus Africa was too small to be represented in R; here it is shown rounded to 1×10^{-300}.



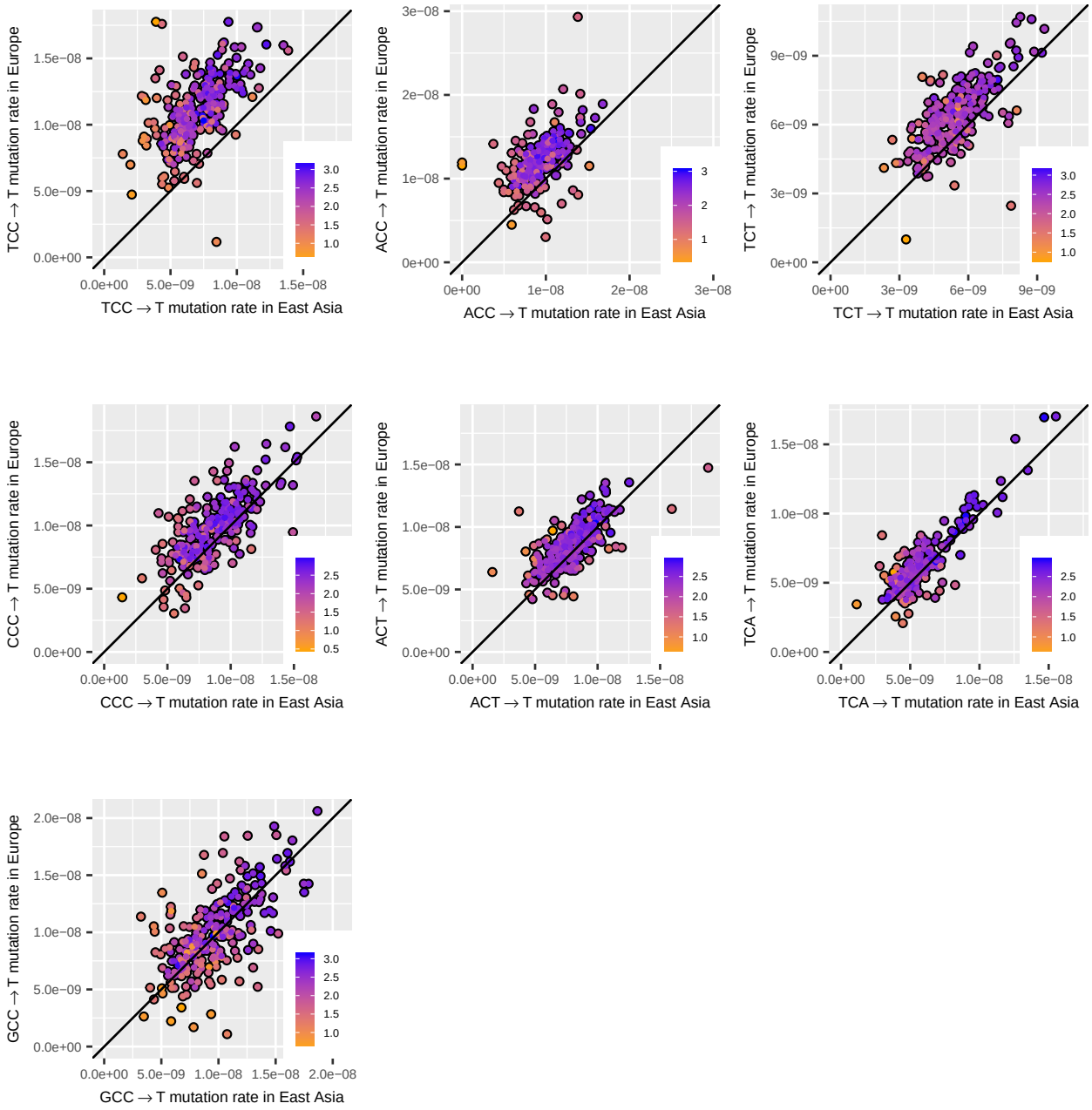
Supplementary Figure 2: *No doubleton excess in CpG mutation profile.* Plots of the proportion of each type of CpG substitution in each population across allele frequency bins. Since singletons are removed from this analysis, a doubleton excess would be reflected in an especially high proportion of CpG mutations in the lowest alternative allele frequency bin. While we do not observe a doubleton excess, we do see an unexplained depletion of tripleton CpG variants in Asia and Africa for most CpG substitutions, reflected in the immediate dip in substitution proportion in the second smallest allele frequency bin for African and Asian populations.



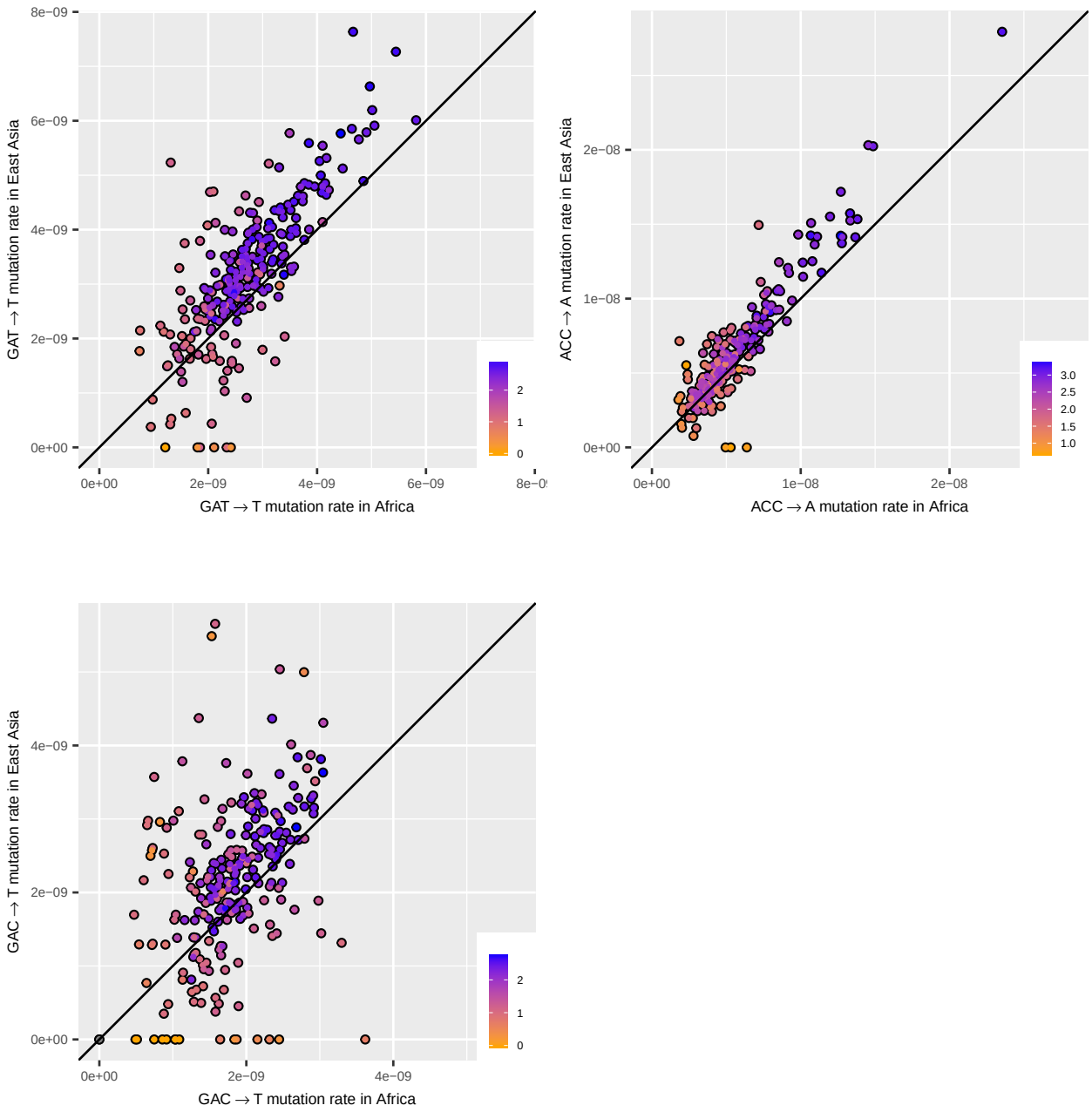
Supplementary Figure 3: Approximate 95% confidence interval estimates of inferred mutation rate for each highly significantly variable 3-mer type across Africa (AFR), Europe (EUR), South Asia (SAS), and East Asia (EAS).



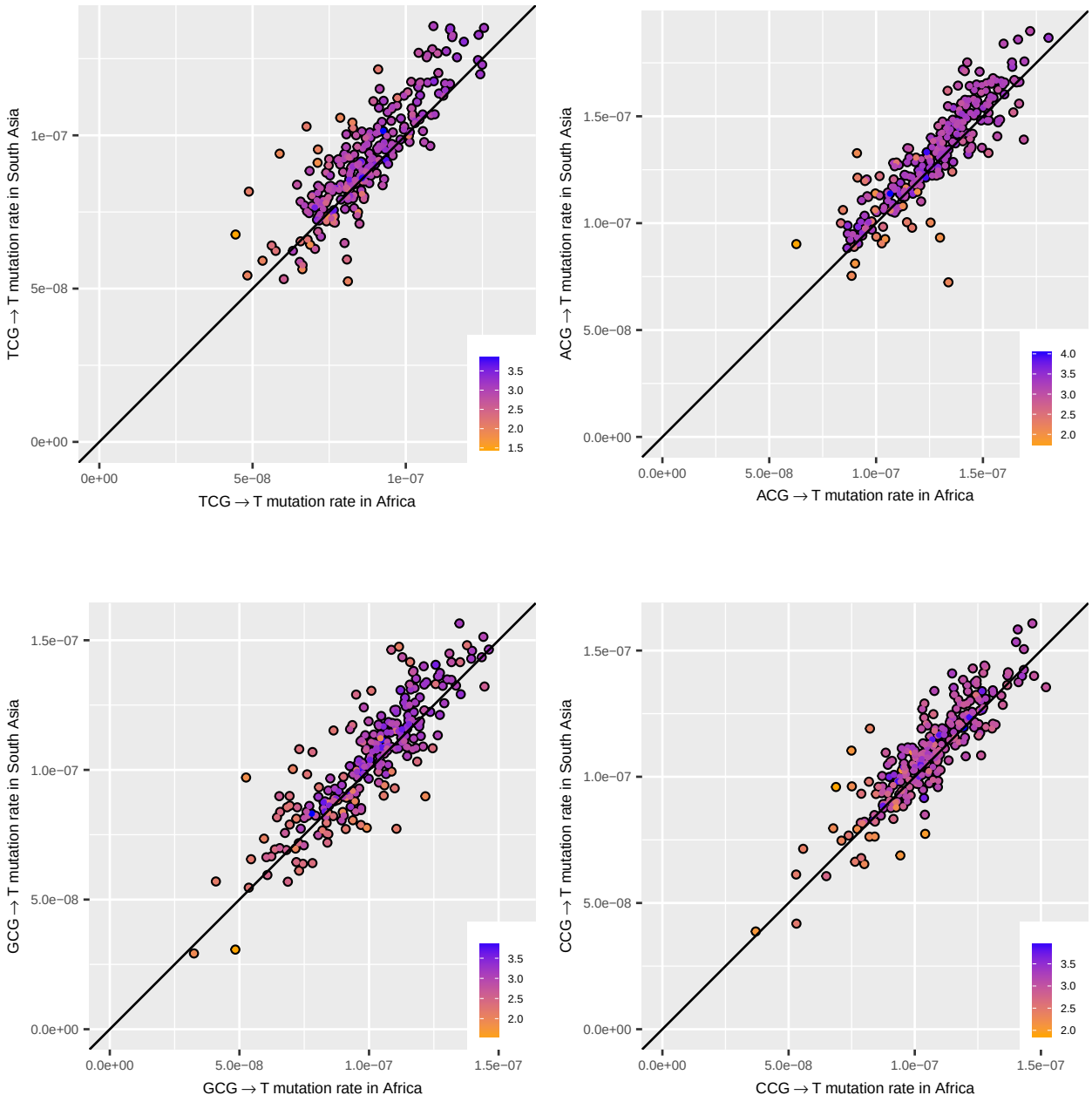
Supplementary Figure 4: Box plots of inferred private mutation rate across chromosomes for each highly significantly variable 3-mer type. Outlier points are labeled with chromosome.



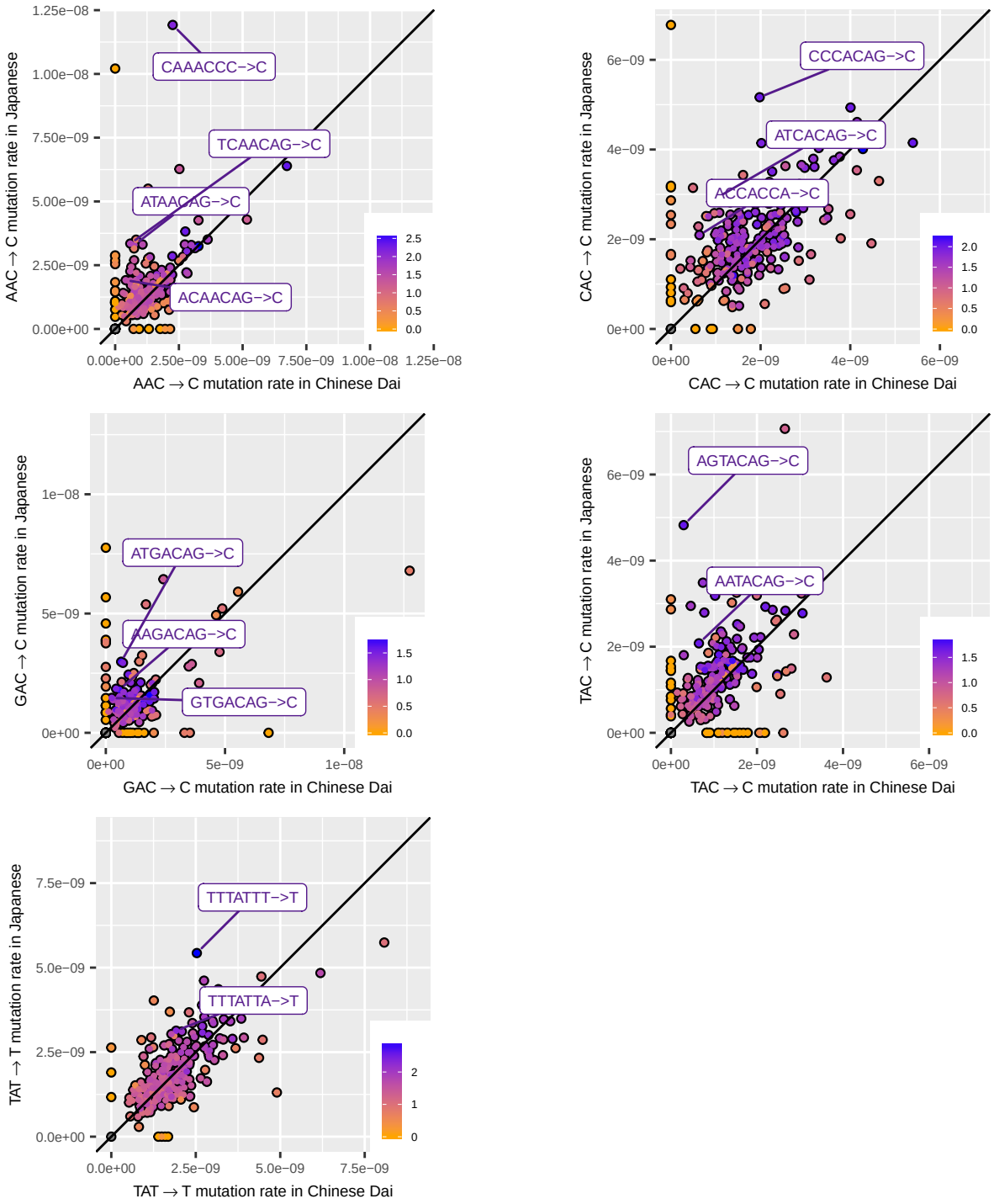
Supplementary Figure 5: Rates of all 7-mer expansions of each profile #1 signature in Europe versus East Asia. Each point represents a 7-mer expansion of the 3-mer subtype shown, plotted based on its estimated mutation rate in each of the two populations displayed. Colors indicate the log (base 10) of the number of substitutions observed for that 7-mer class. Europe and East Asia were selected to visualize this comparison because the difference in mutation rate for profile #1 3-mers is most strong and consistent between these populations. Generally, these polymorphisms match case II (see main text): enrichment is consistent across 7mers, with some expected amount of noise.



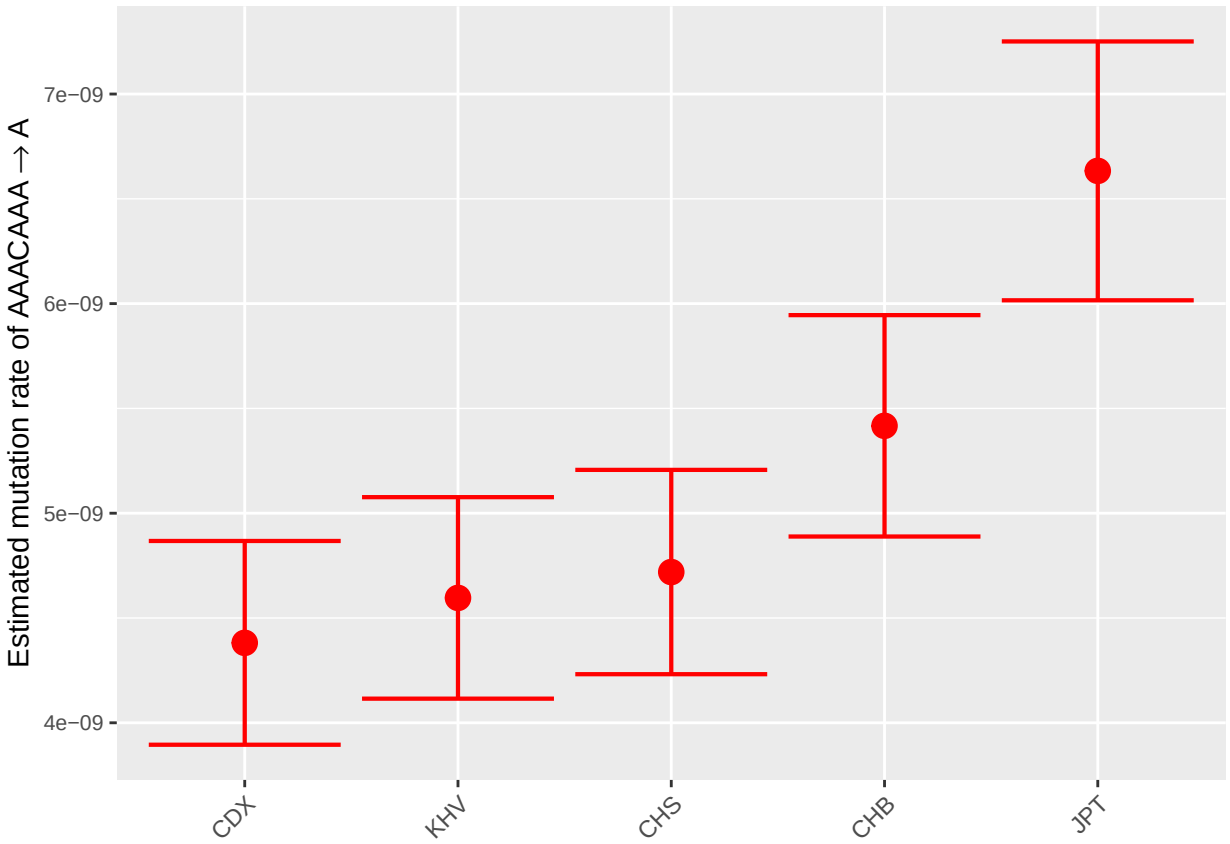
Supplementary Figure 6: Rates of all 7-mer expansions of each profile #2 signature in Africa versus East Asia. Again, the population-specific enrichment is consistent across 7mers, consistent with case II.



Supplementary Figure 7: Rates of all 7-mer expansions of each profile #4 signature in Africa versus South Asia. The relative mutation rate variability between the CpG mutations is small compared to other profiles noted in this report. However, we still observe that most 7-mer expansions of these CpG substitutions lie above the diagonal, consistent with case II.



Supplementary Figure 8: Rates of all 7-mer expansions of each profile #3 signature in Japanese versus Chinese Dai. Labeled points are those which were significantly variable between Japanese and Chinese Dai (fdr < 0.05 among all profile #3 7-mer expansions).



Supplementary Figure 9: Approximate 95% confidence interval estimates of inferred mutation rate for AAACAAA→A across East Asian subpopulations: Chinese Dai in Xishuangbanna (CDX); Kinh in Ho Chi Minh City, Vietnam (KHV); Southern Han Chinese (CHS); Han Chinese in Beijing (CHB); and Japanese in Tokyo. This pattern resembles that of 3-mer profile #3.