

Exploratory Data Analysis of High Heterozygosity Variants in Freeze2

July 14, 2016

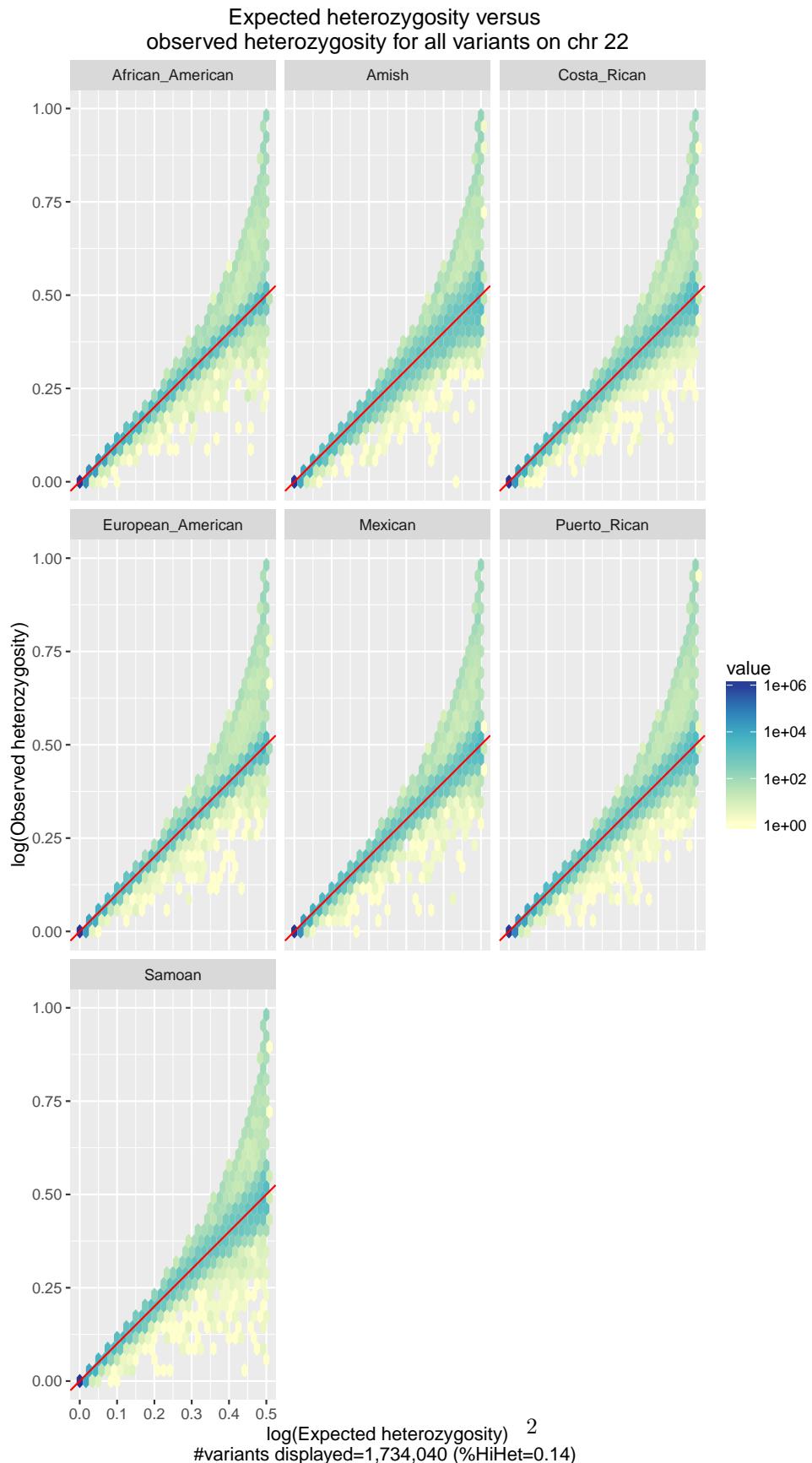
Files that were used in the analysis:

1. File with TOPMed InDel Annotations (courtesy of Xiaoming Liu)
2. Freeze 2 GDS GT only (includes all chromosomes)
3. HWE results for each ancestry (courtesy of Stephanie G.)

Pre-work that was done:

- * Created a dataframe with variant.id, chr, pos, ref, alt, MAP20 and MAP35 fields.
- * Extracted variants only for chromosome 22.

Plot 1. Expected heterozygosity versus observed heterozygosity for all variants on chr 22



Fraction of high hets over all chromosomes

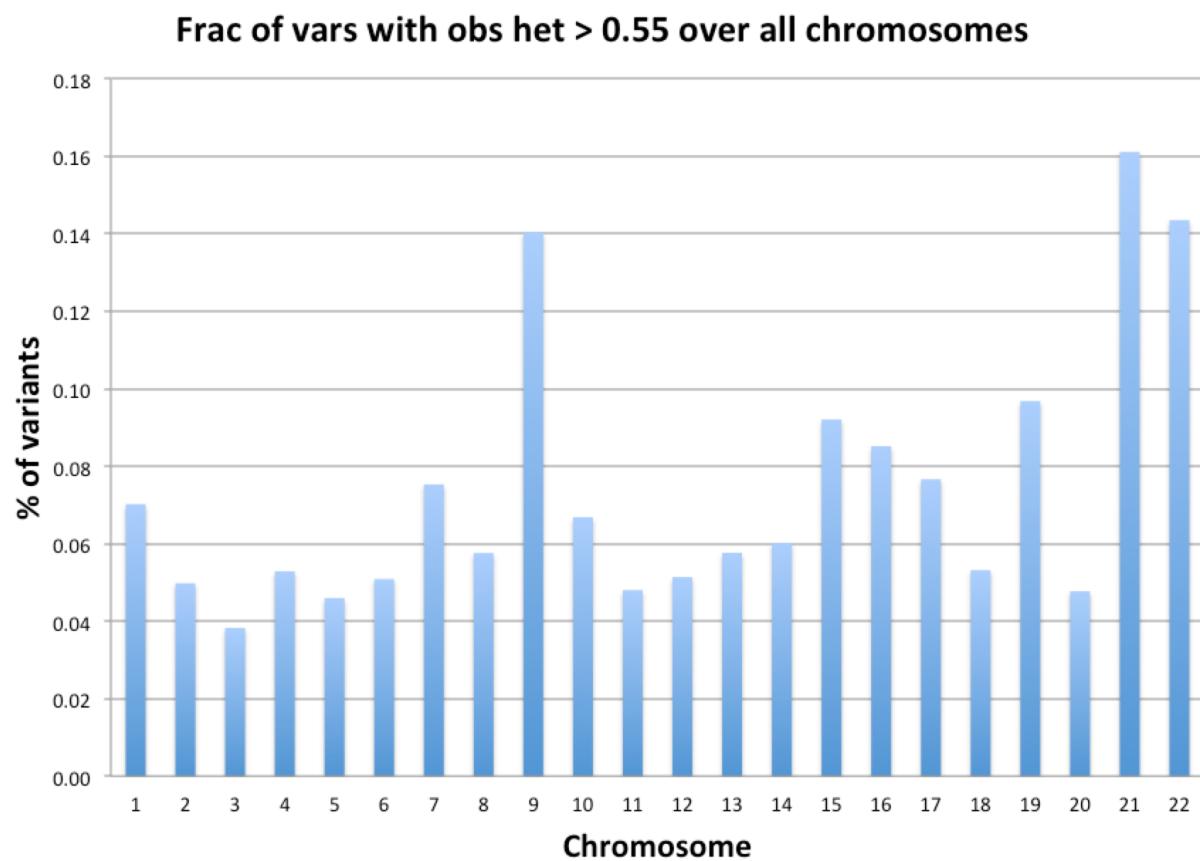
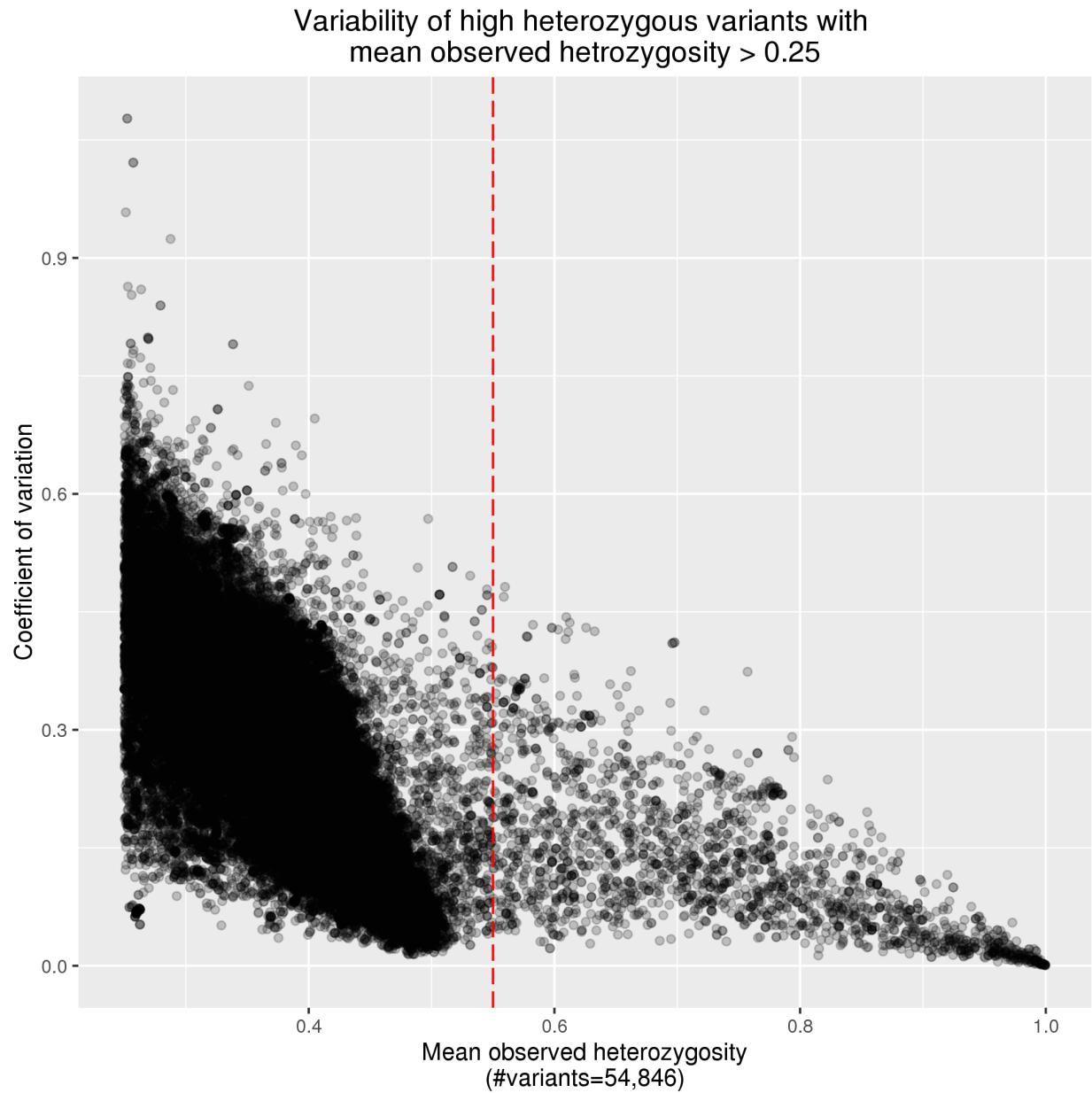
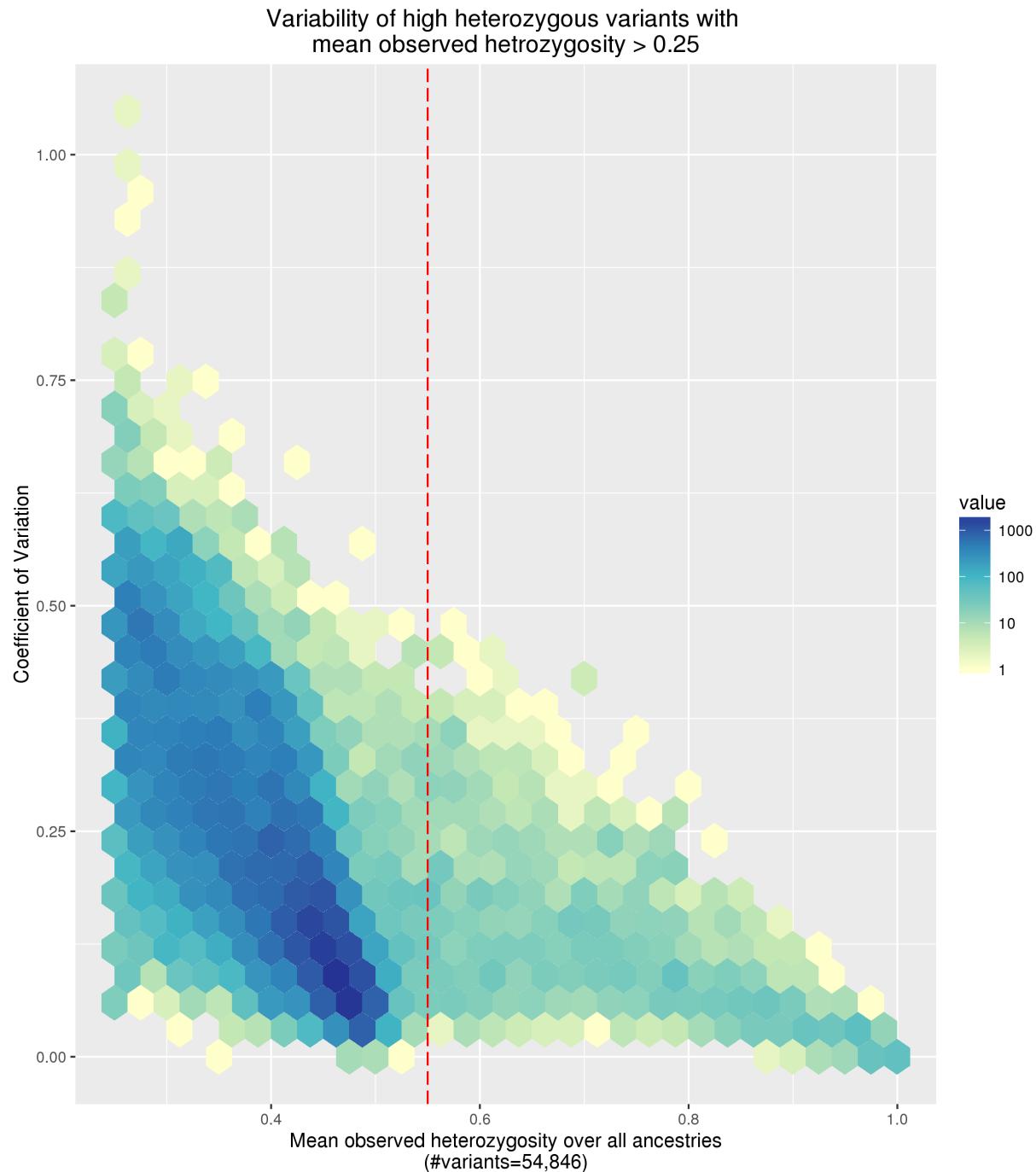


Figure 1:

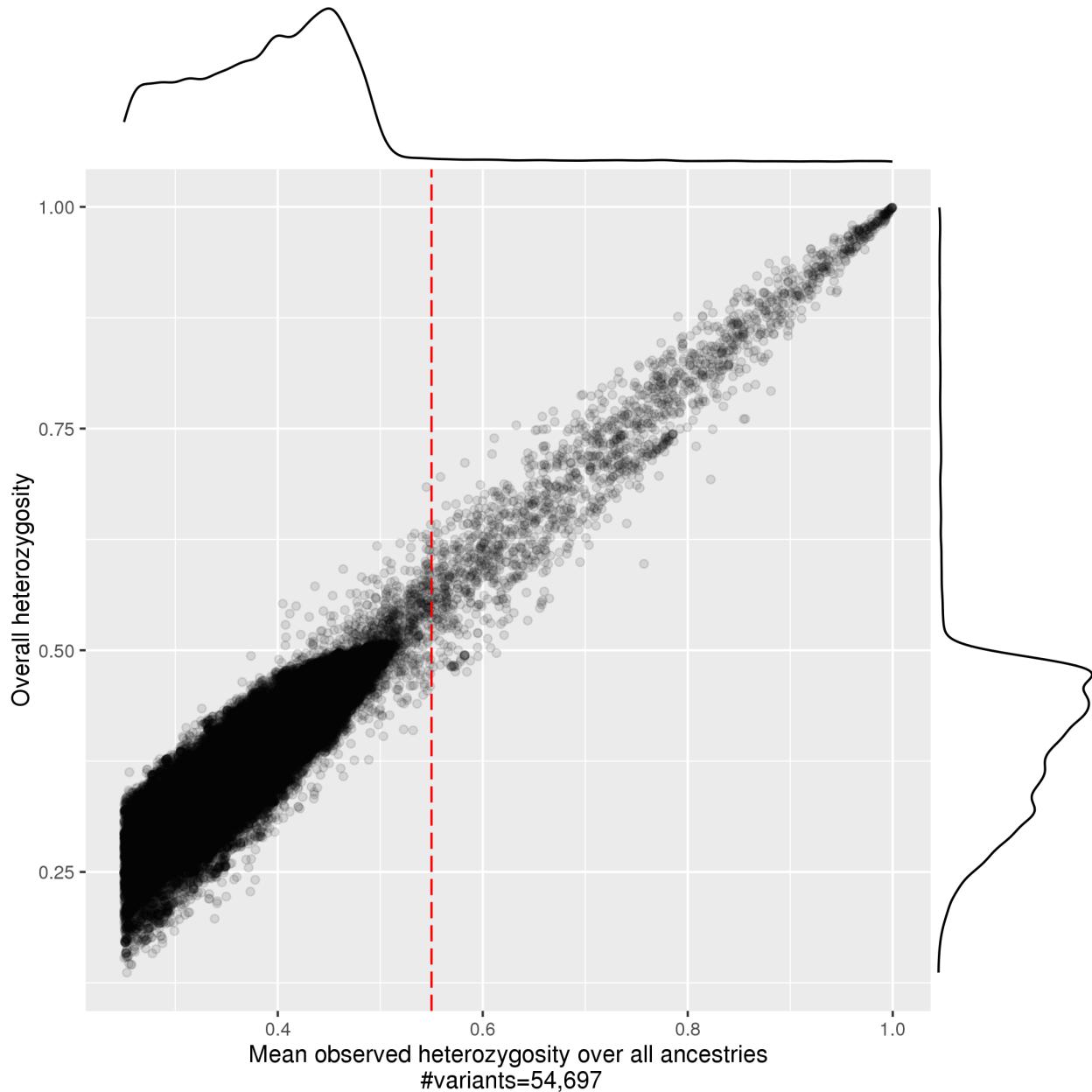
Plot 2. Mean observed heterozygosity versus coefficient of variation (among 7 ancestry groups) for all variants on chr 22 with observed heterozygosity > 0.25



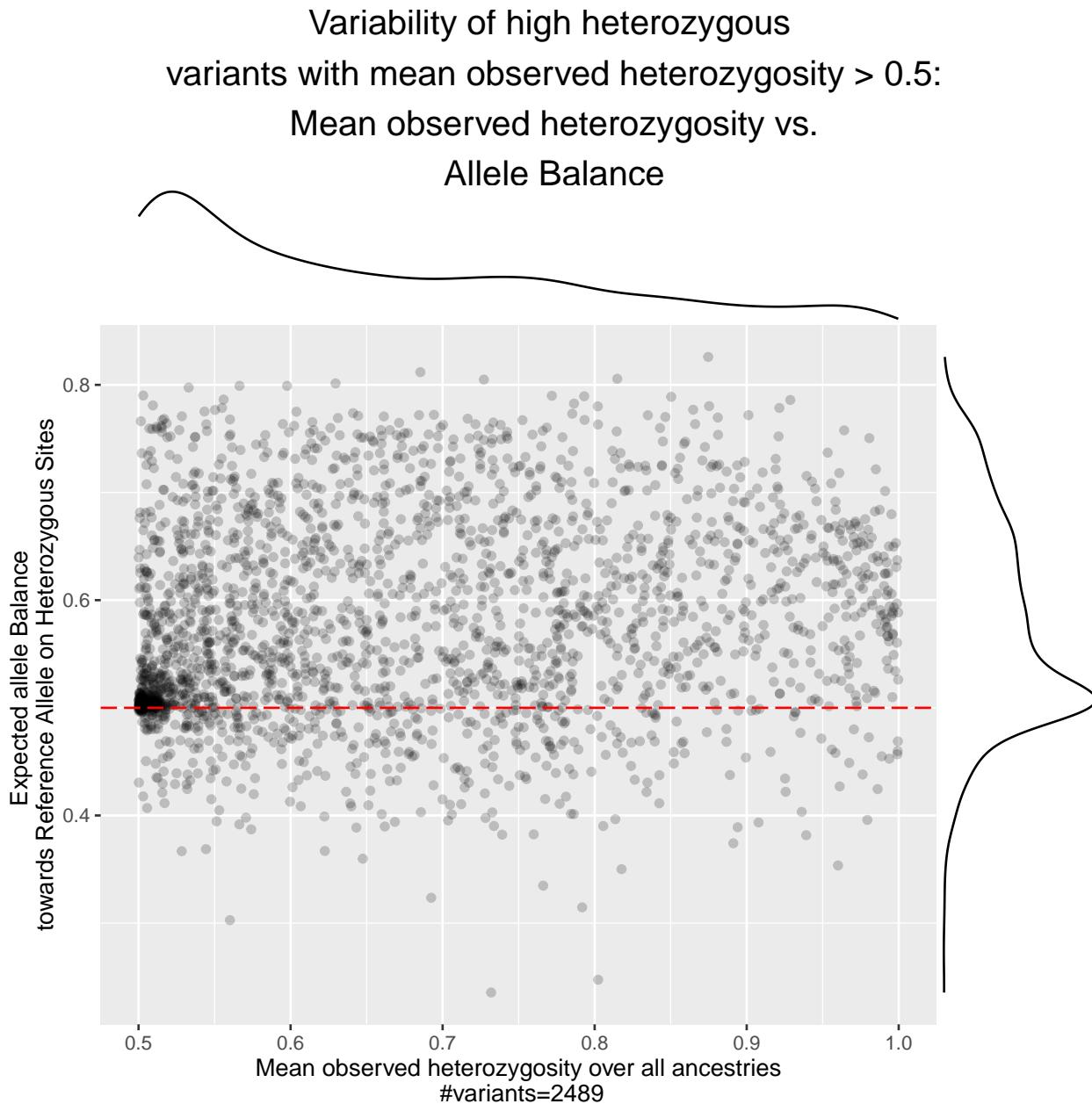
Plot 3. Mean observed heterozygosity versus coefficient of variation (among 7 ancestry groups) for all variants on chr 22 with observed heterozygosity > 0.25 (Density)



Plot 4. Overall observed heterozygosity versus mean observed heterozygosity (over 7 ancestry groups) for all variants on chr 22 with mean observed heterozygosity > 0.25



Plot 5. Allele Balance versus mean heterozygosity



ABE is $(\text{reference allele count}) / (\text{reference allele count} + \text{alternate allele count})$, averaged over heterozygous genotypes

Plot 6. Allele Balance versus overall heterozygosity

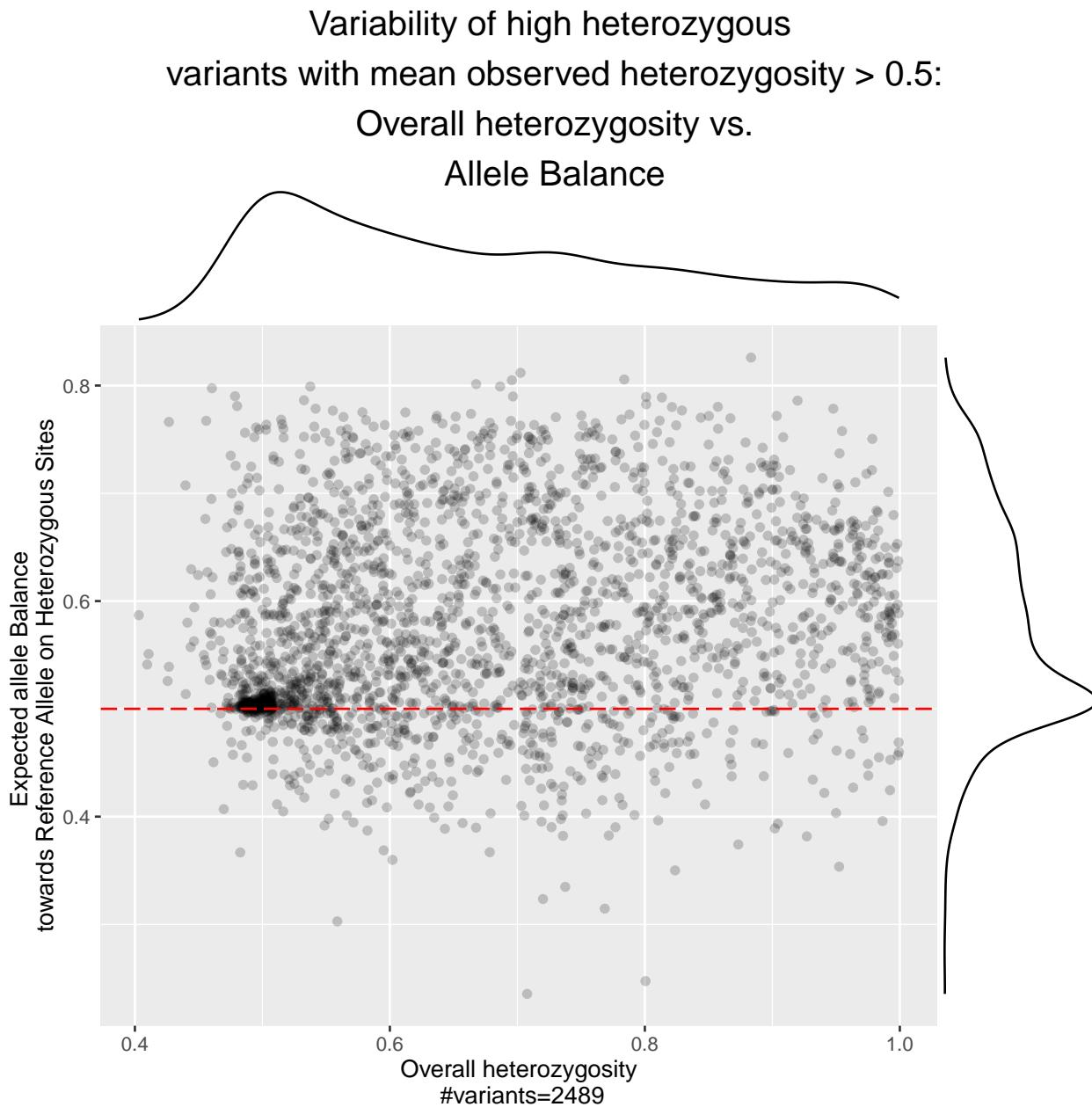


Figure 2:

ABE is (reference allele count)/(reference allele count + alternate allele count), averaged over heterozygous genotypes

MAP20 and MAP35 Definitions

MAP20 and MAP35 represent the average of Duke 20 and Duke 35 scores of the windows covering the variant.

MAP20 and MAP35 are the direct measures of sequence uniqueness throughout the reference genome. It displays how unique each sequence is on the positive strand starting at a particular base and of a particular length. Thus, the 20 bp track reflects the uniqueness of all 20 base sequences with the score being assigned to the first base of the sequence. Scores are normalized to between 0 and 1.

MAP20 = 1 completely unique sequence

MAP20=0 representing a sequence that occurs more than 4 times in the genome

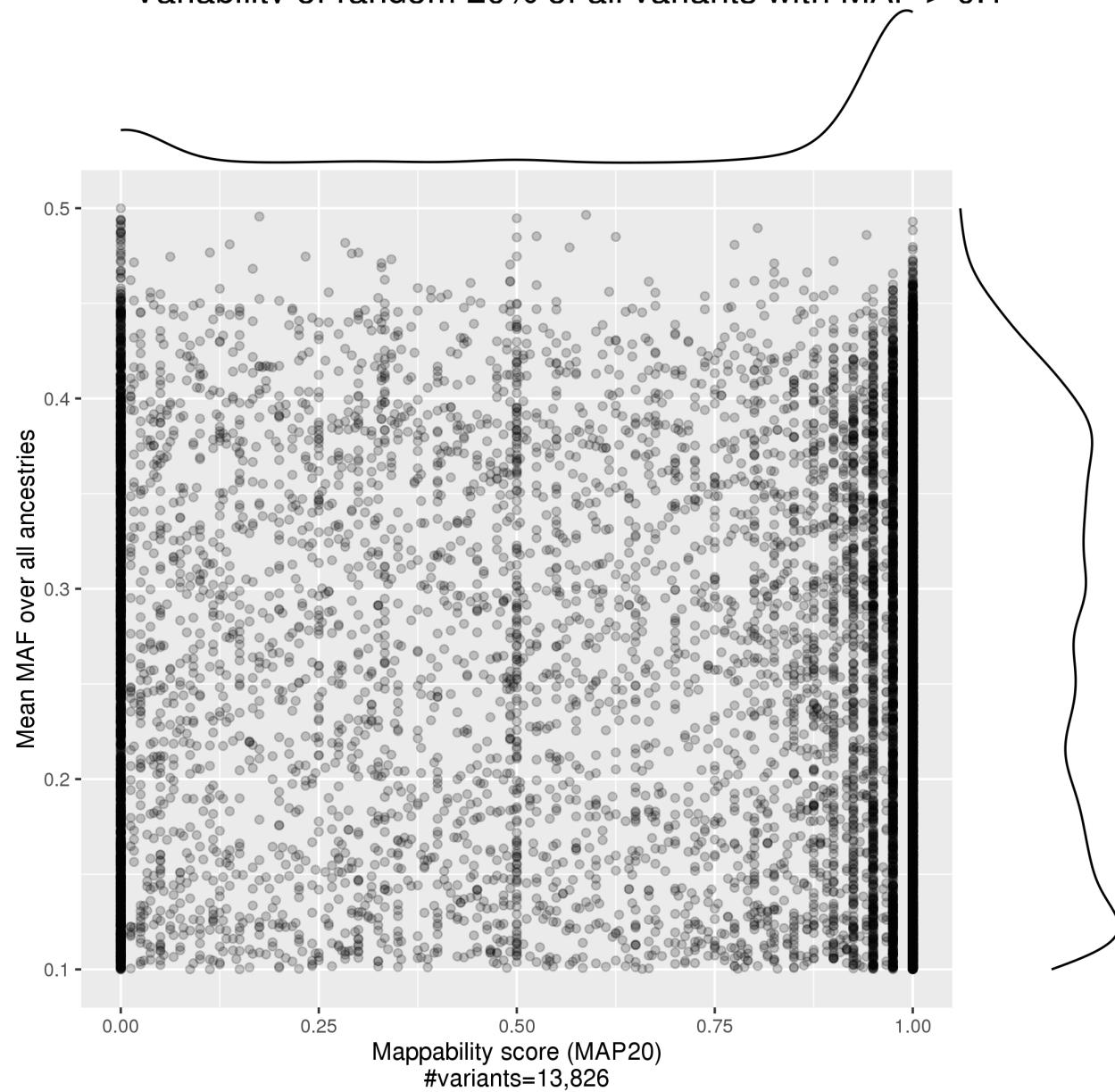
MAP20= 0.5 indicates the sequence occurs exactly twice

MAP20= 0.33 indicates the sequence occurs for three times

MAP20 = 0.25 indicates the sequence occurs for four times

Plot 7. Mean MAF vs MAP20 score

Variability of random 20% of all variants with $\text{MAF} > 0.1$



Plot 8. MAP20 score versus inbreeding coefficient for all variants on chromosome 22 with observed heterozygosity > 0.55 within each ancestry group

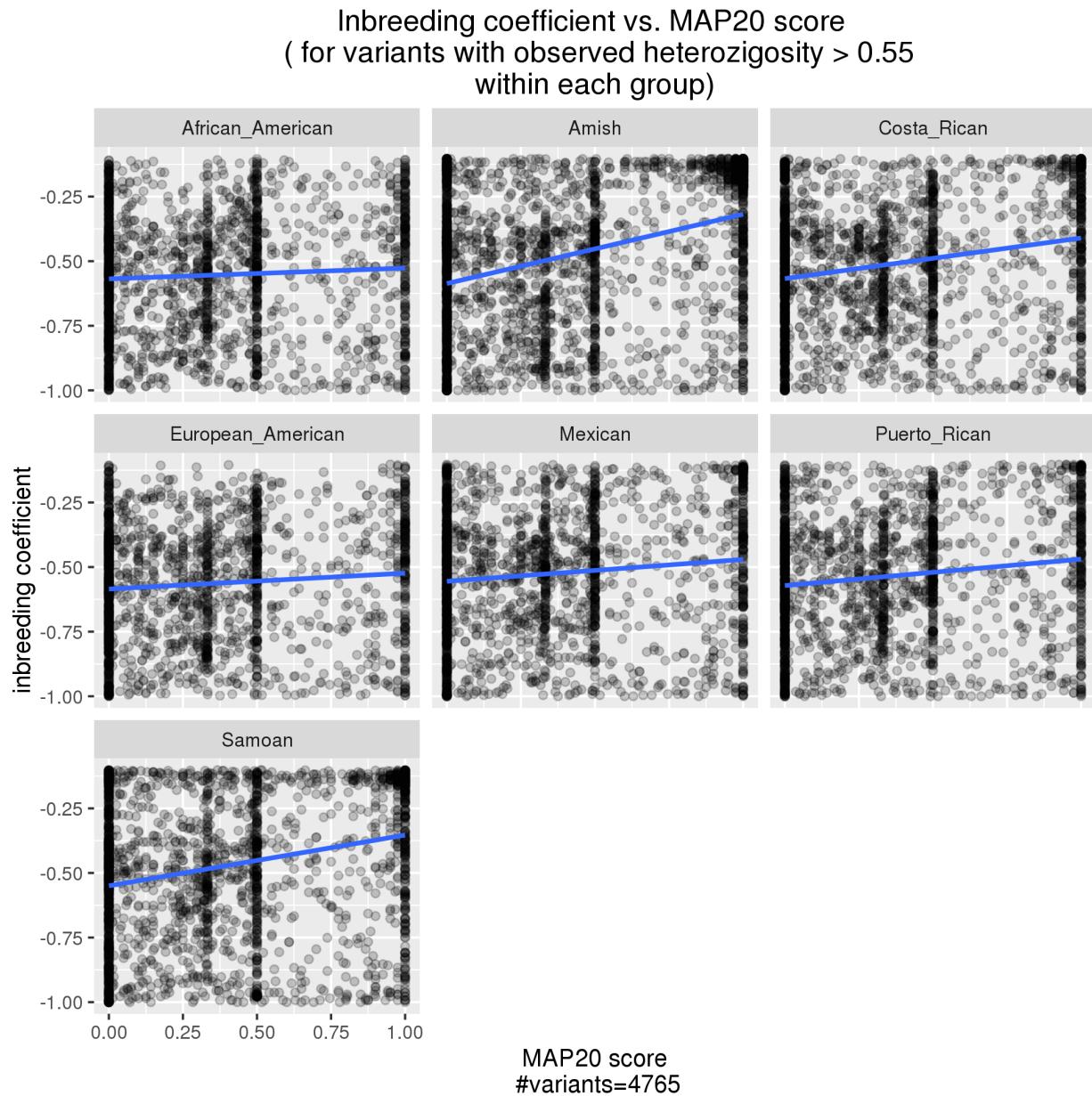
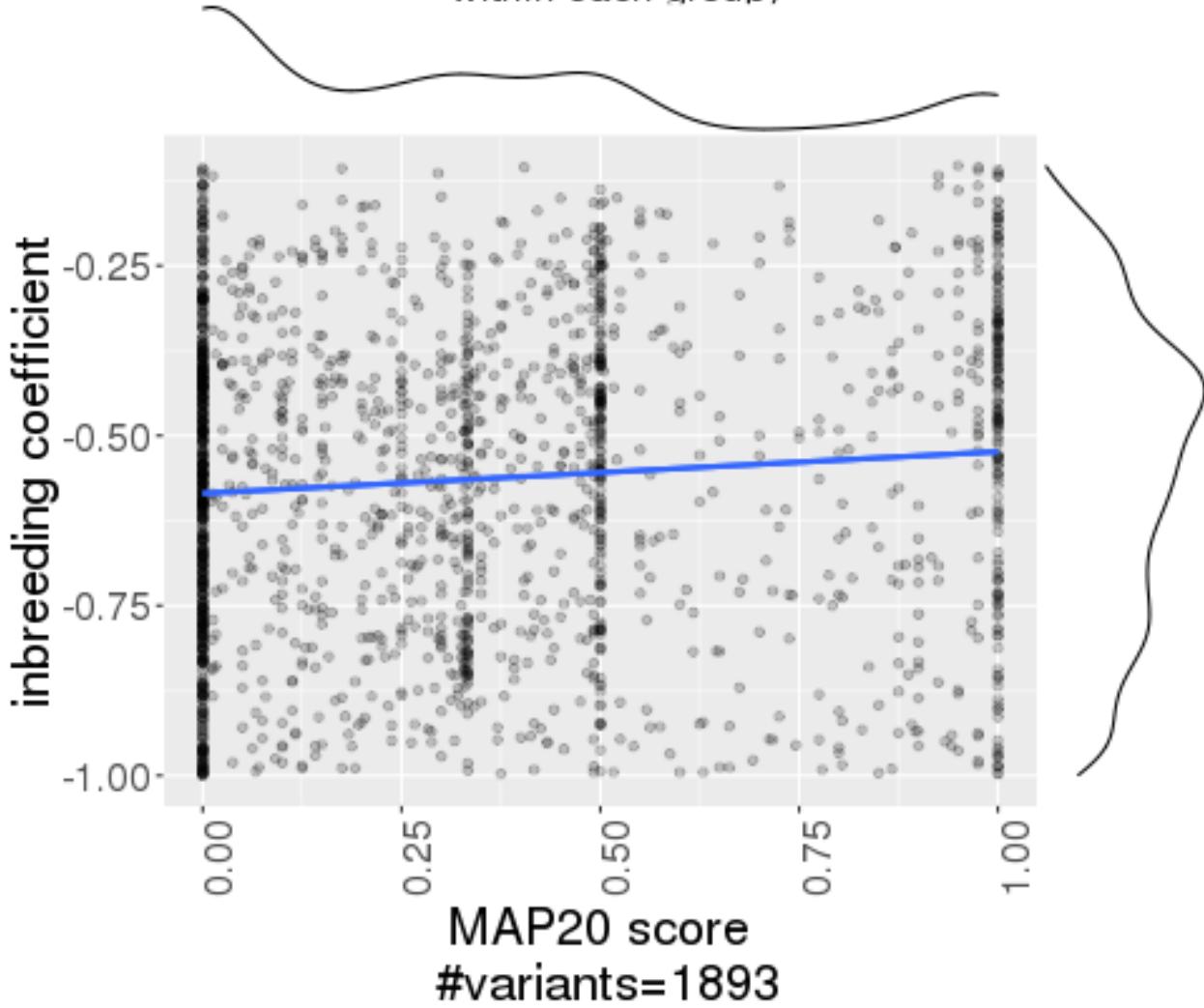
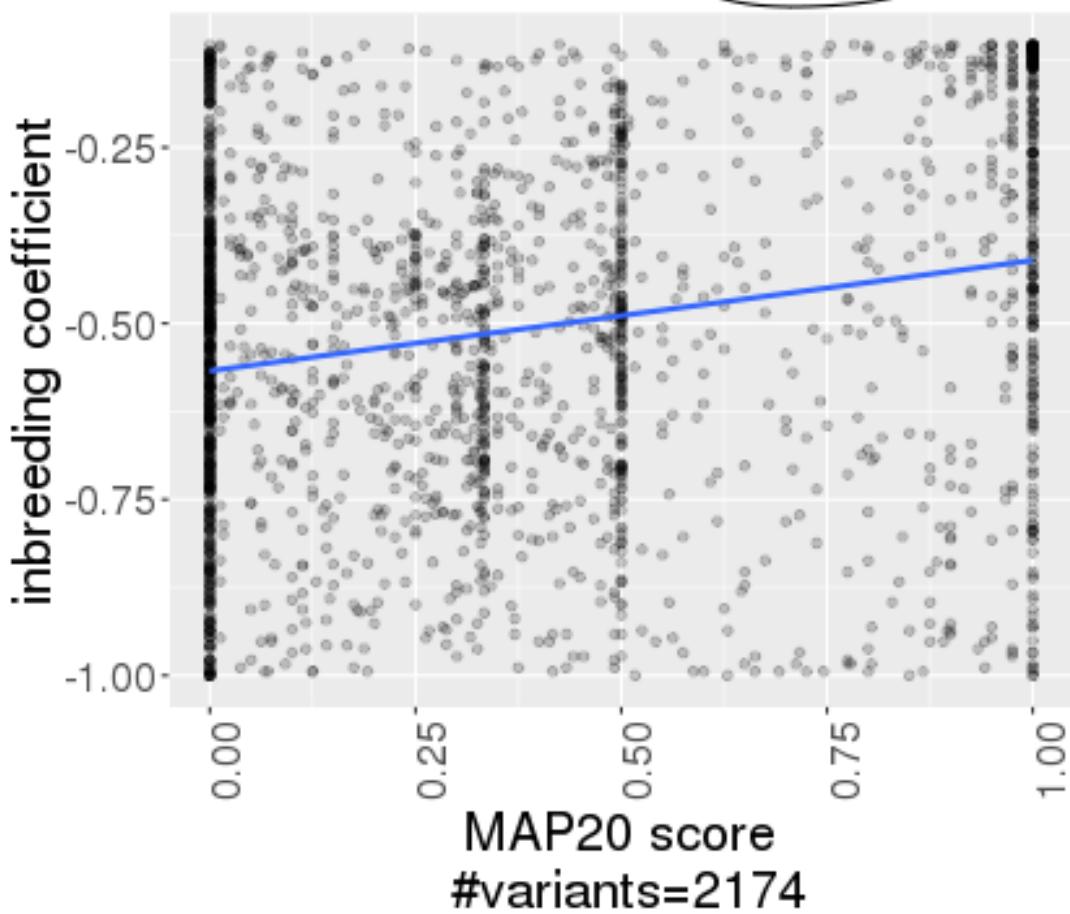


Figure 3:

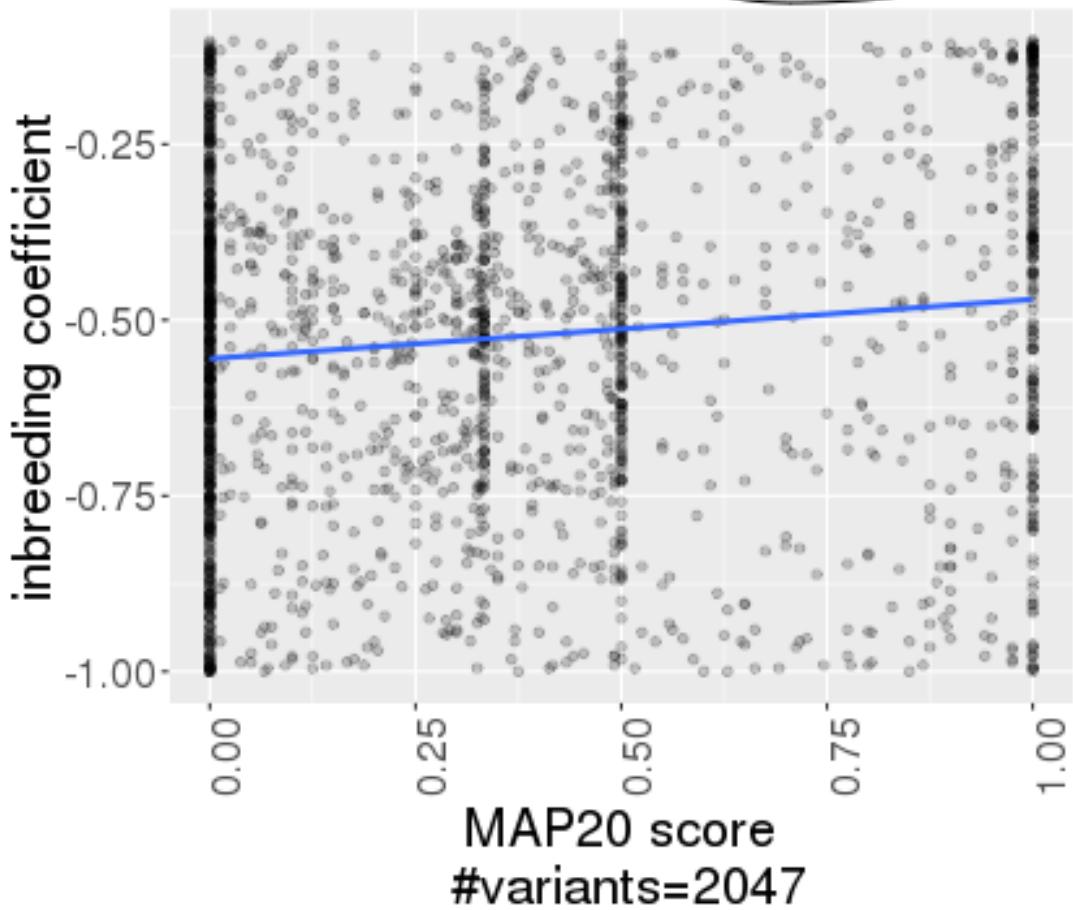
European_American
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozygosity > 0.55
within each group)



Costa_Rican
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozygosity > 0.55
within each group)



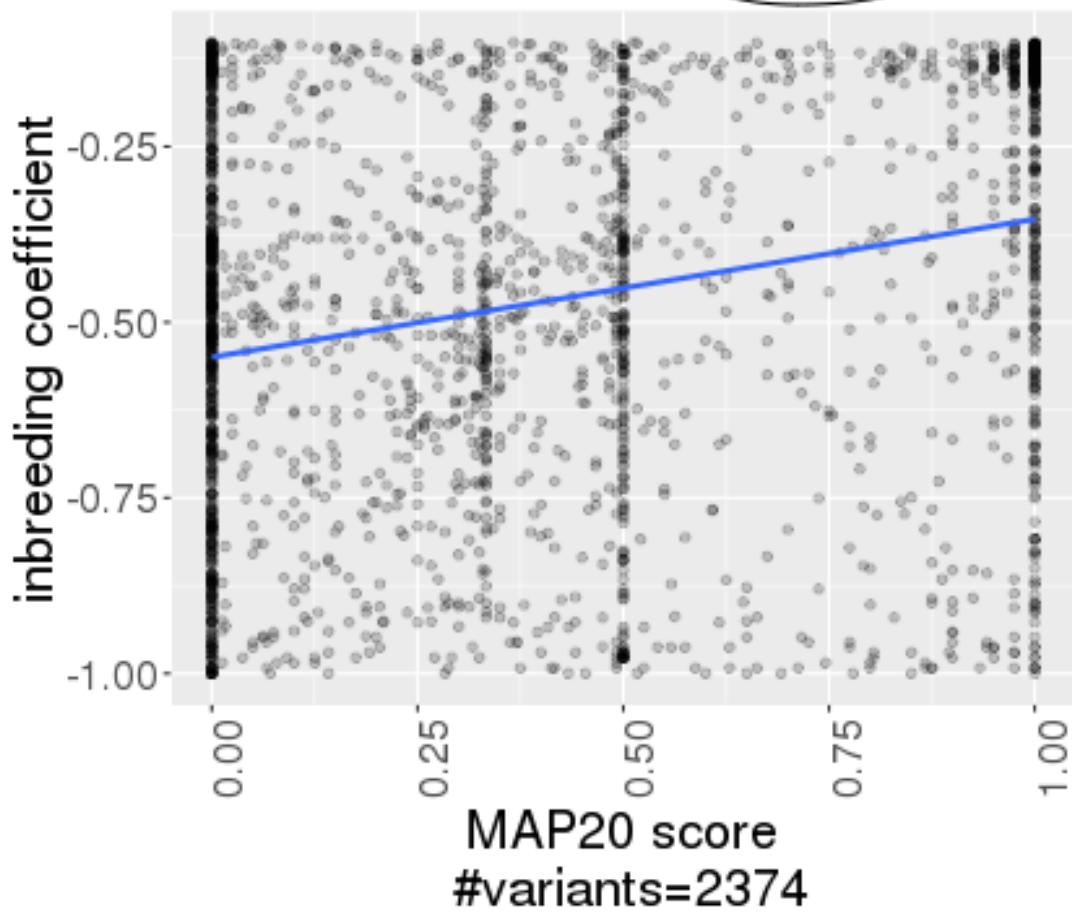
Mexican
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozygosity > 0.55
within each group)

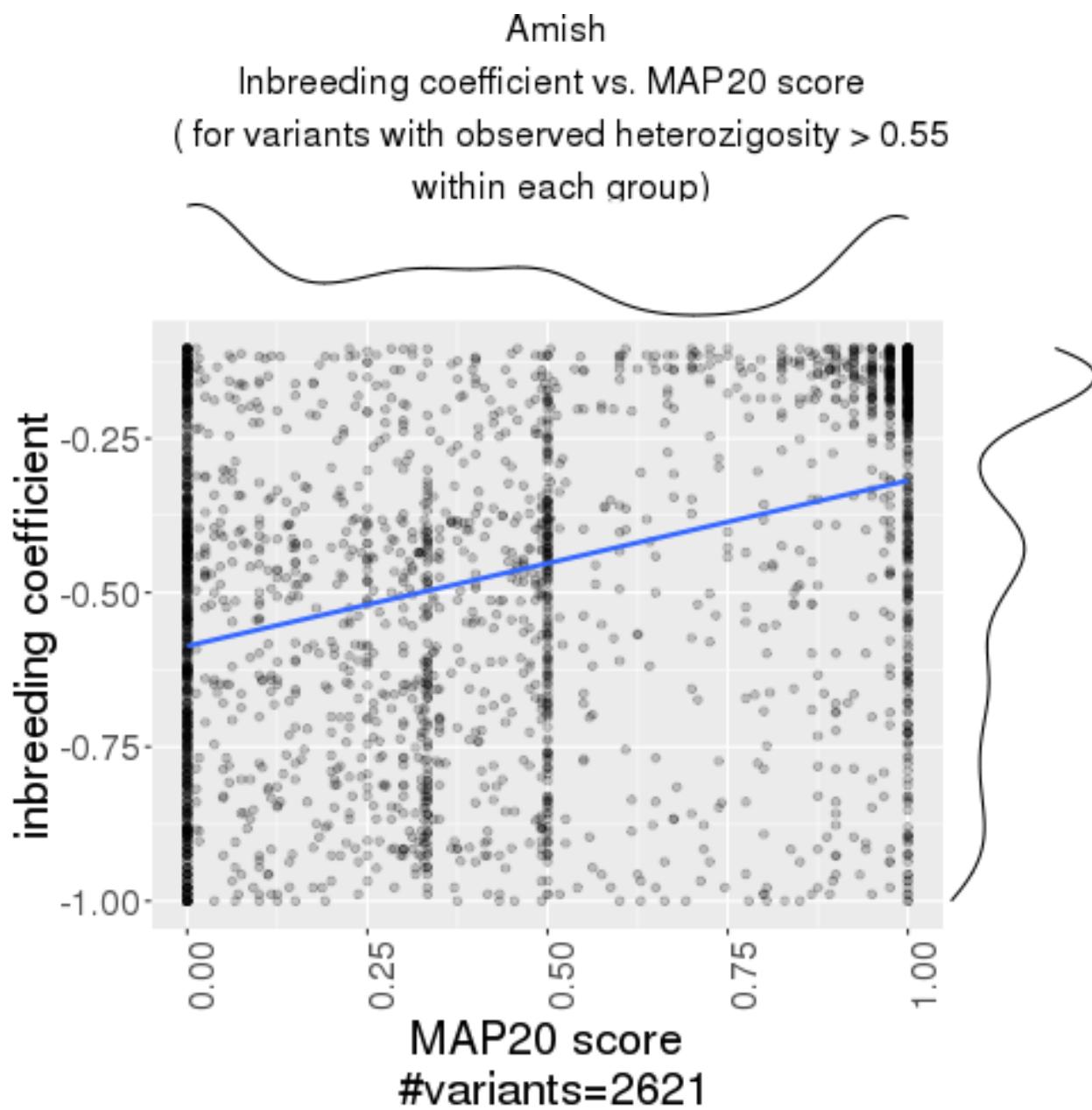


Samoan

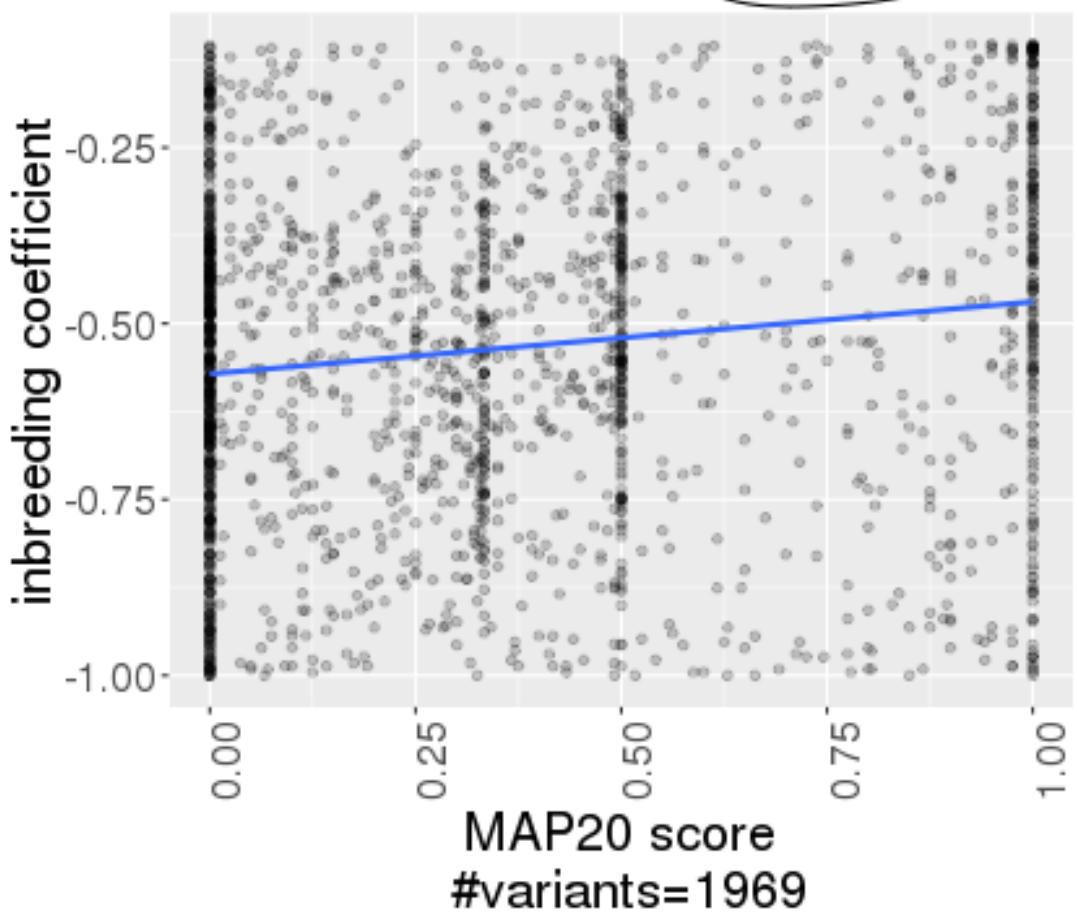
Inbreeding coefficient vs. MAP20 score

(for variants with observed heterozygosity > 0.55
within each group)

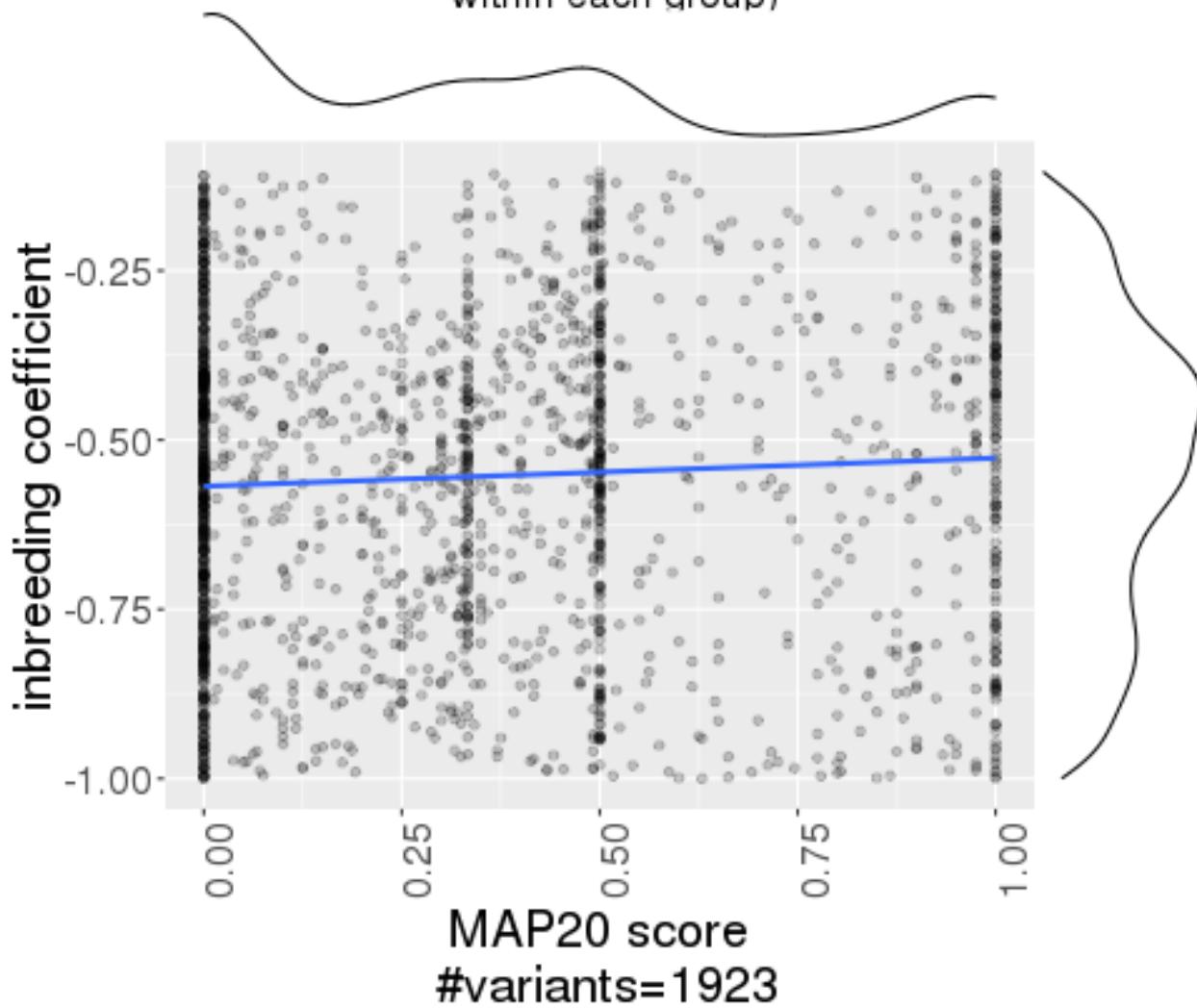




Puerto_Rican
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozygosity > 0.55
within each group)



African_American
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozygosity > 0.55
within each group)



Plot 9. MAP35 score versus inbreeding coefficient for all variants on chromosome 22 with observed heterozygosity > 0.55 within each ancestry group

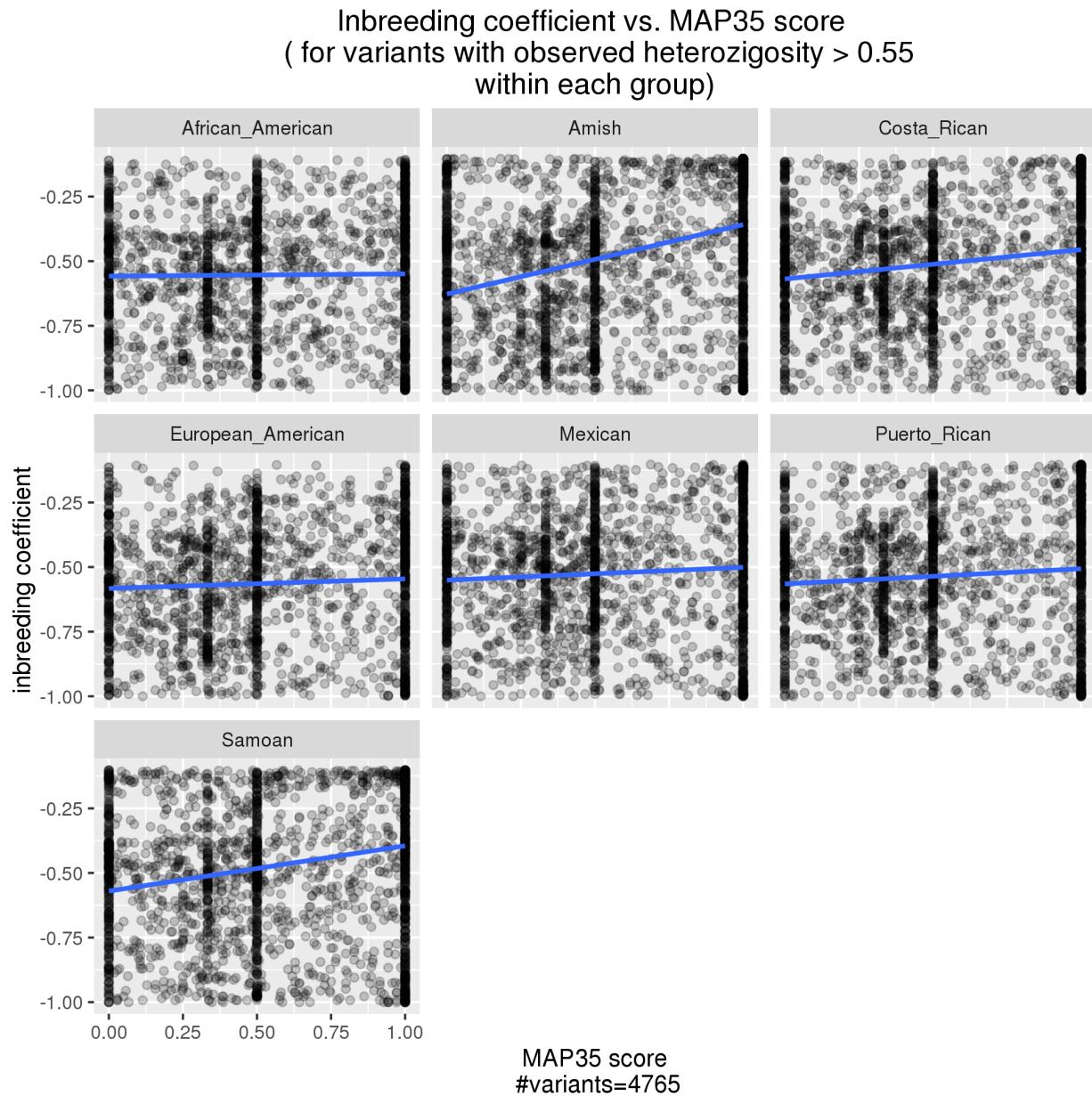
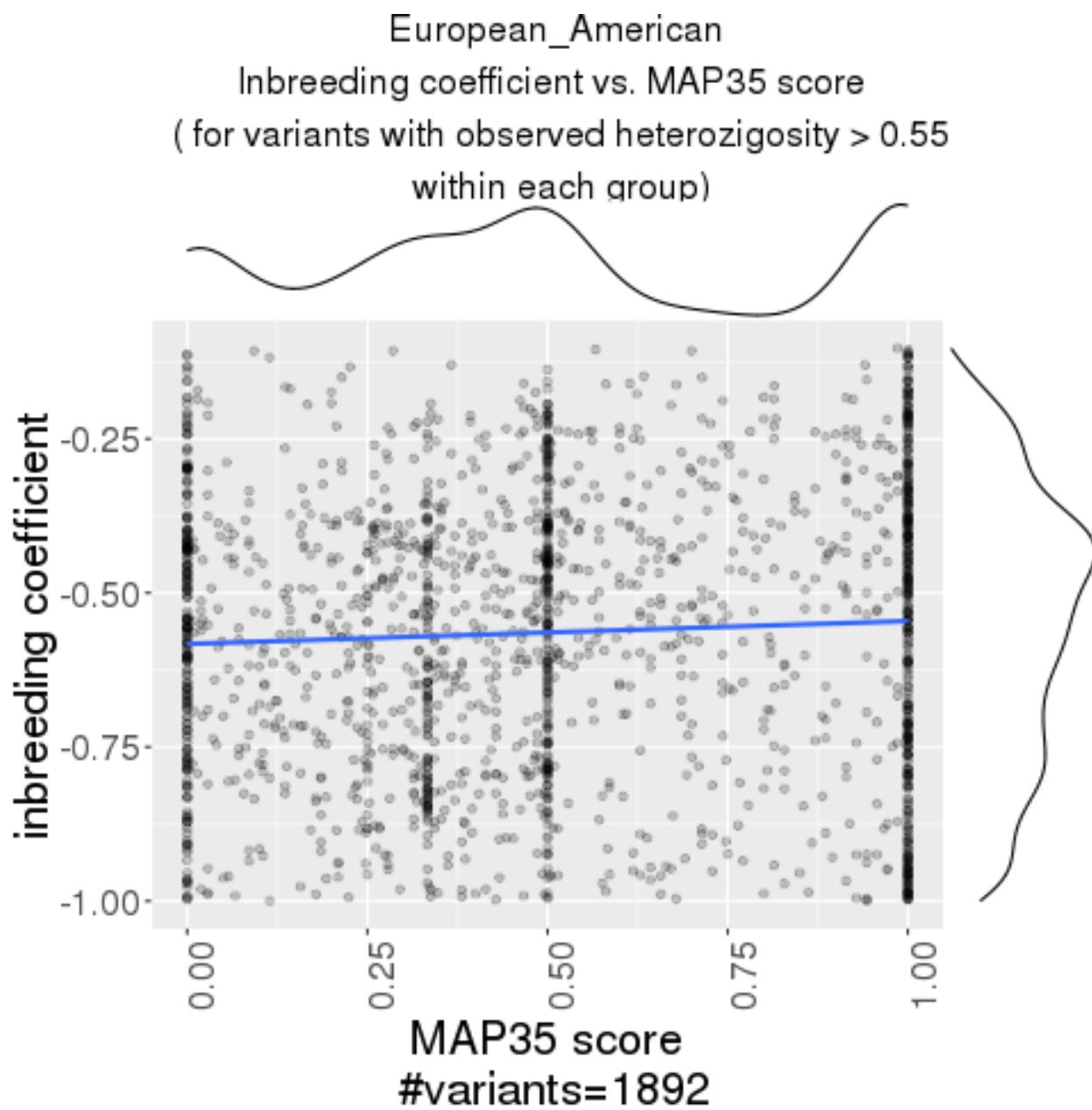
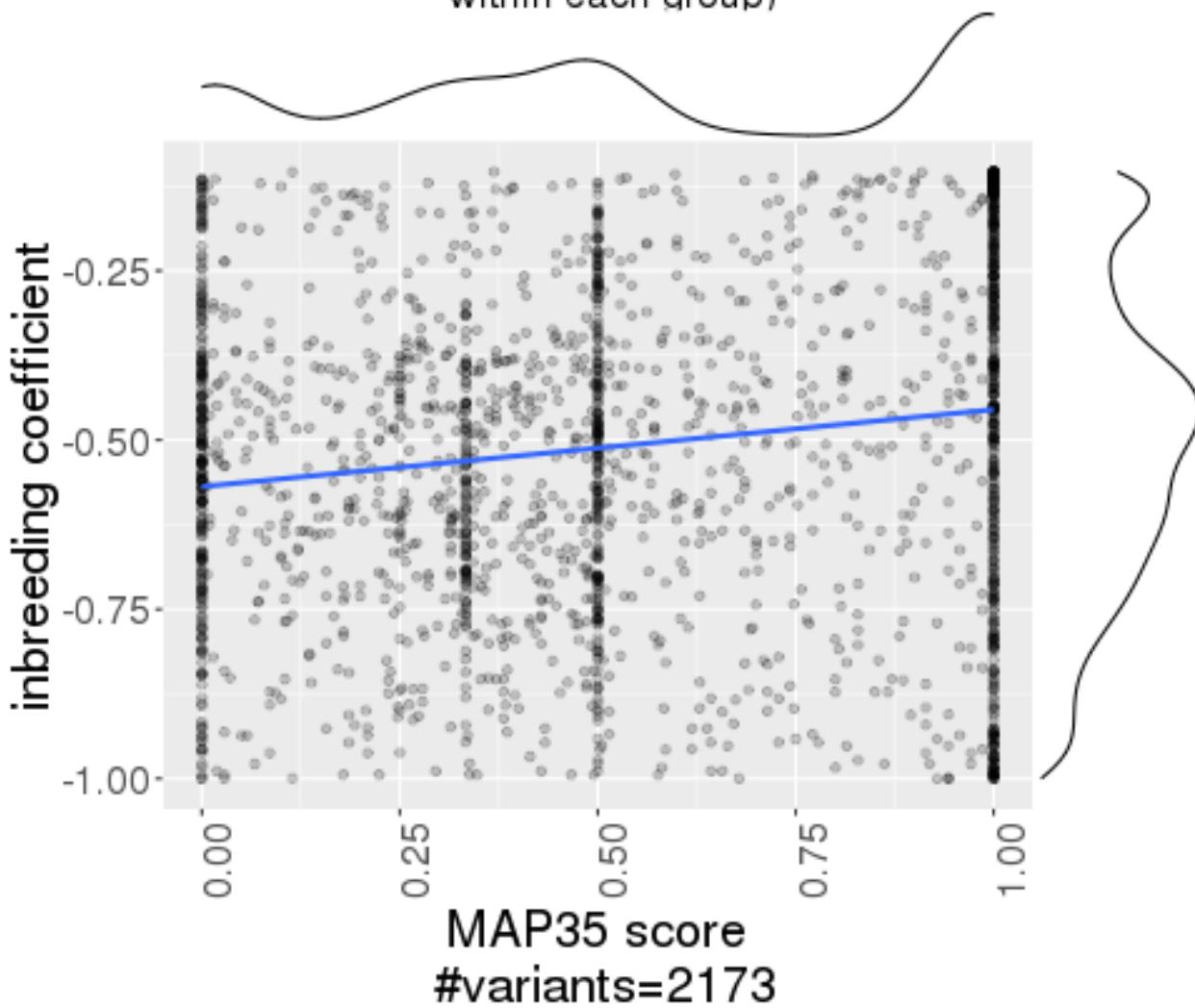
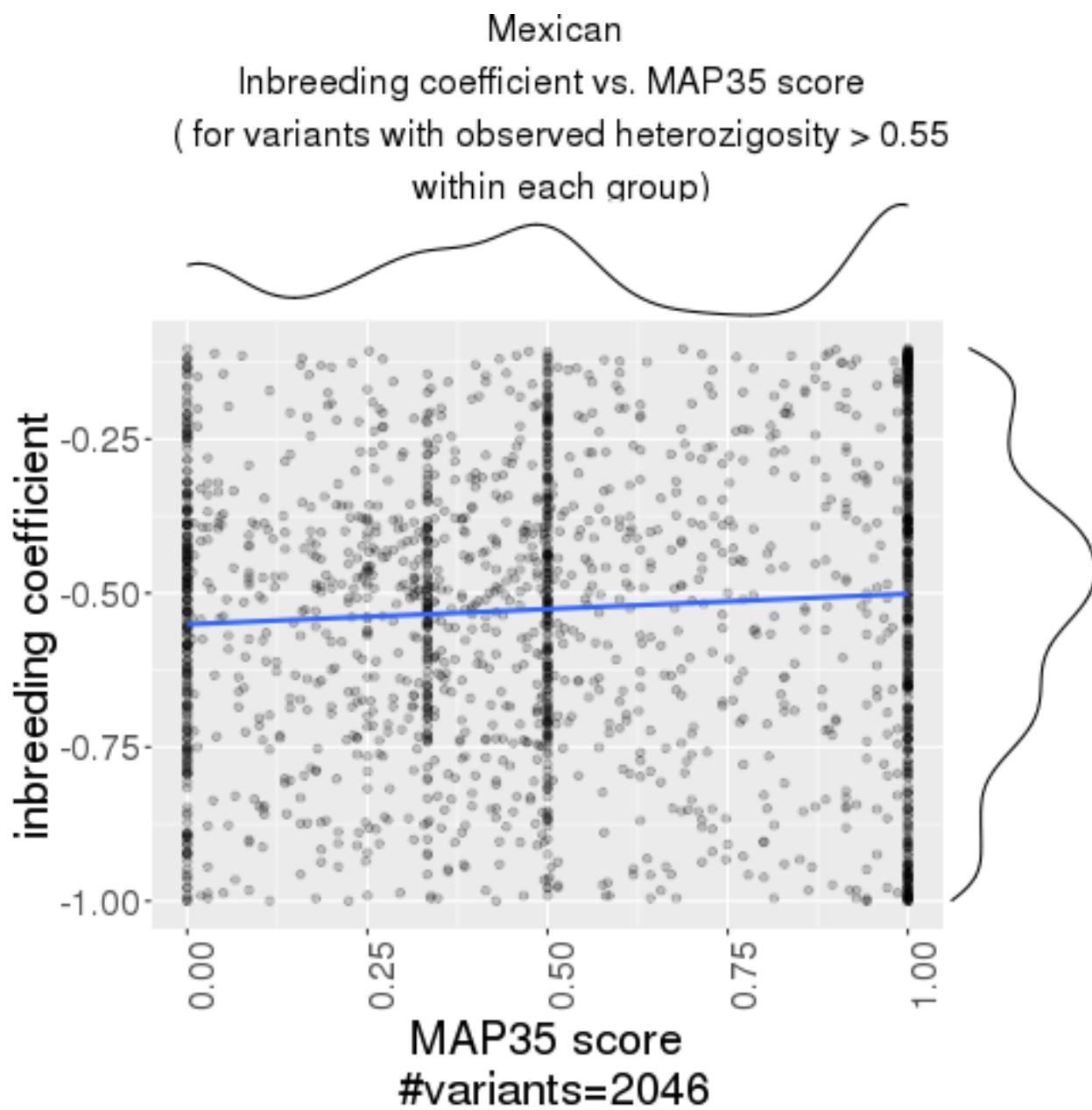


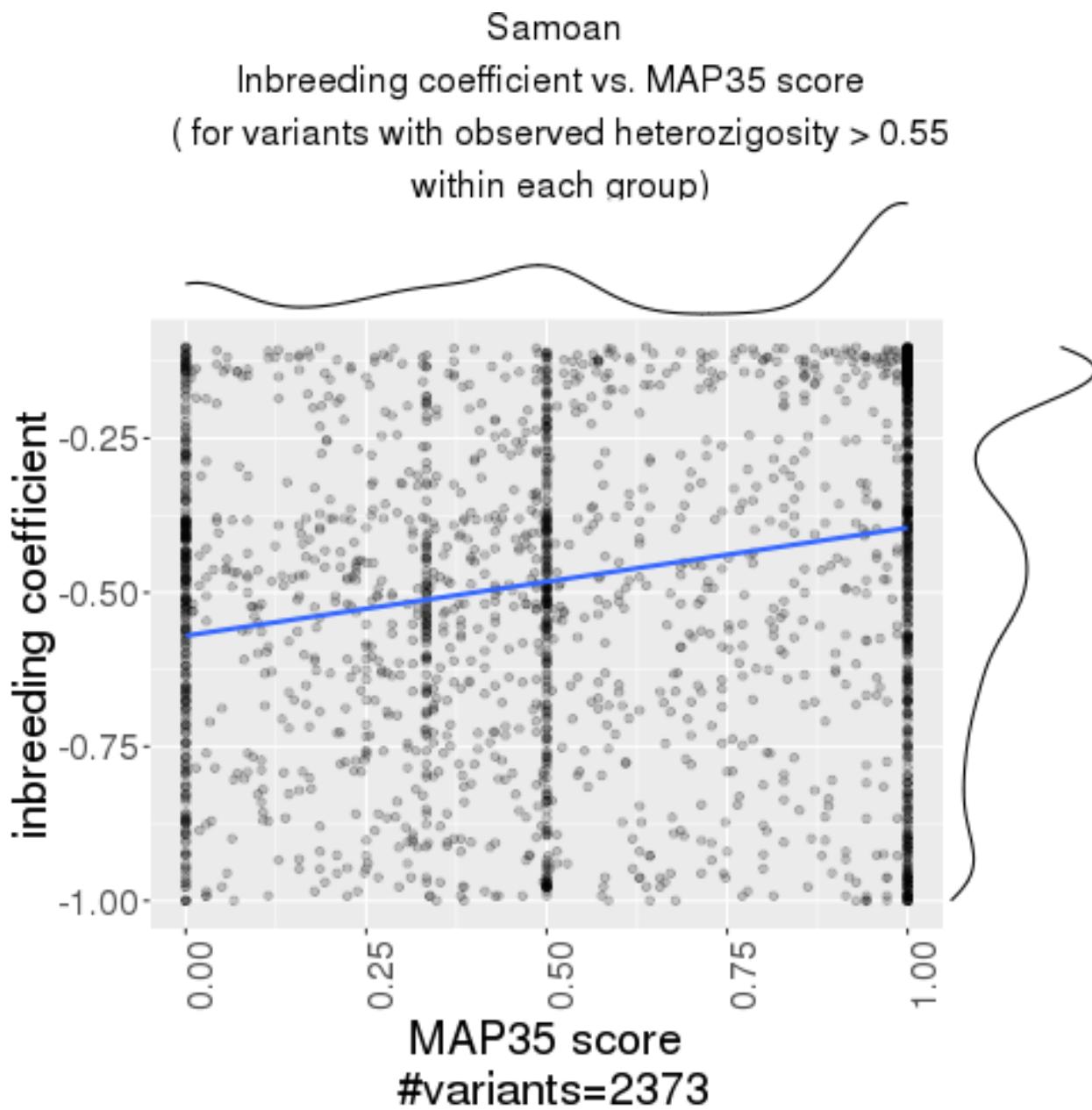
Figure 4:



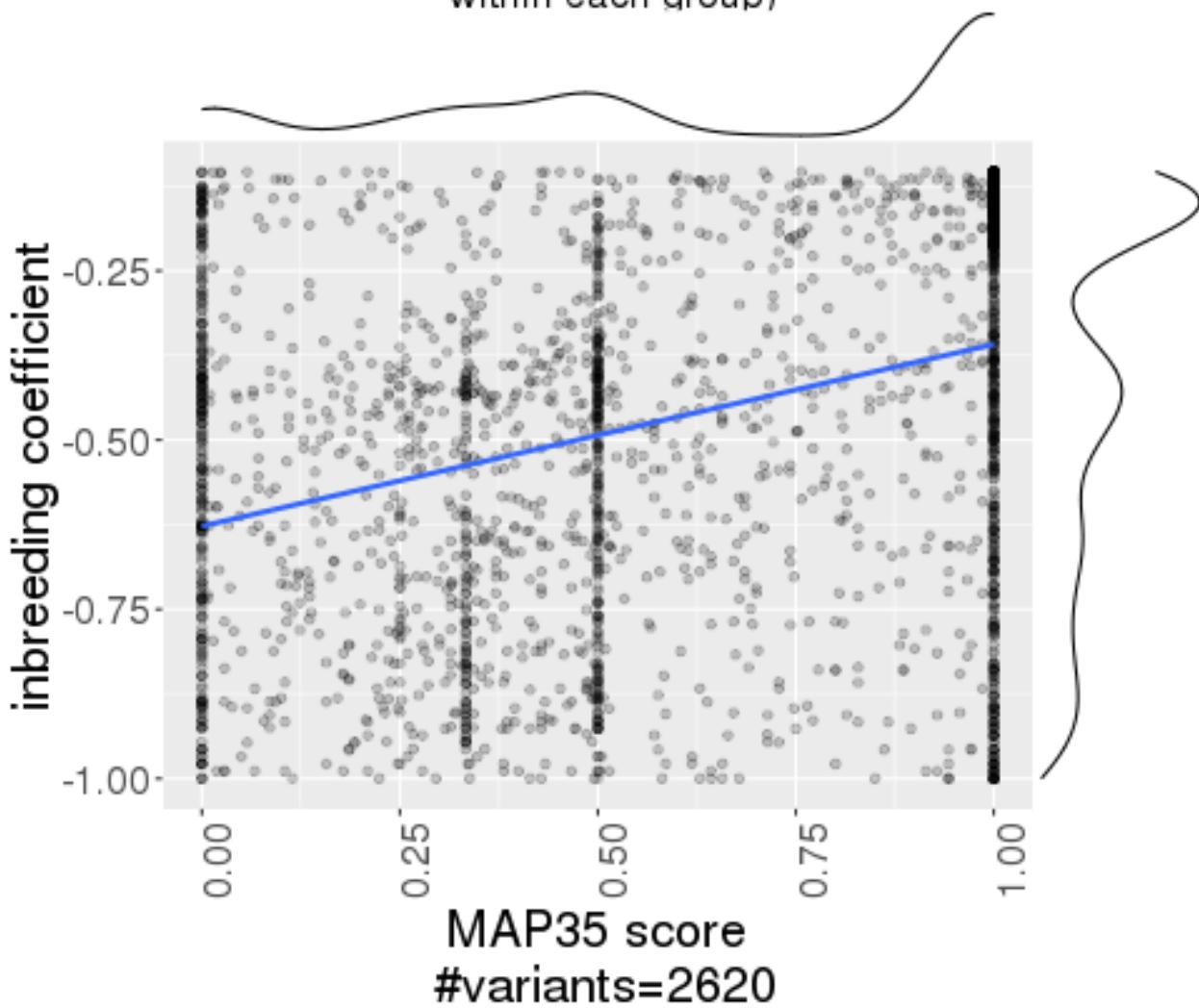
Costa_Rican
Inbreeding coefficient vs. MAP35 score
(for variants with observed heterozygosity > 0.55
within each group)

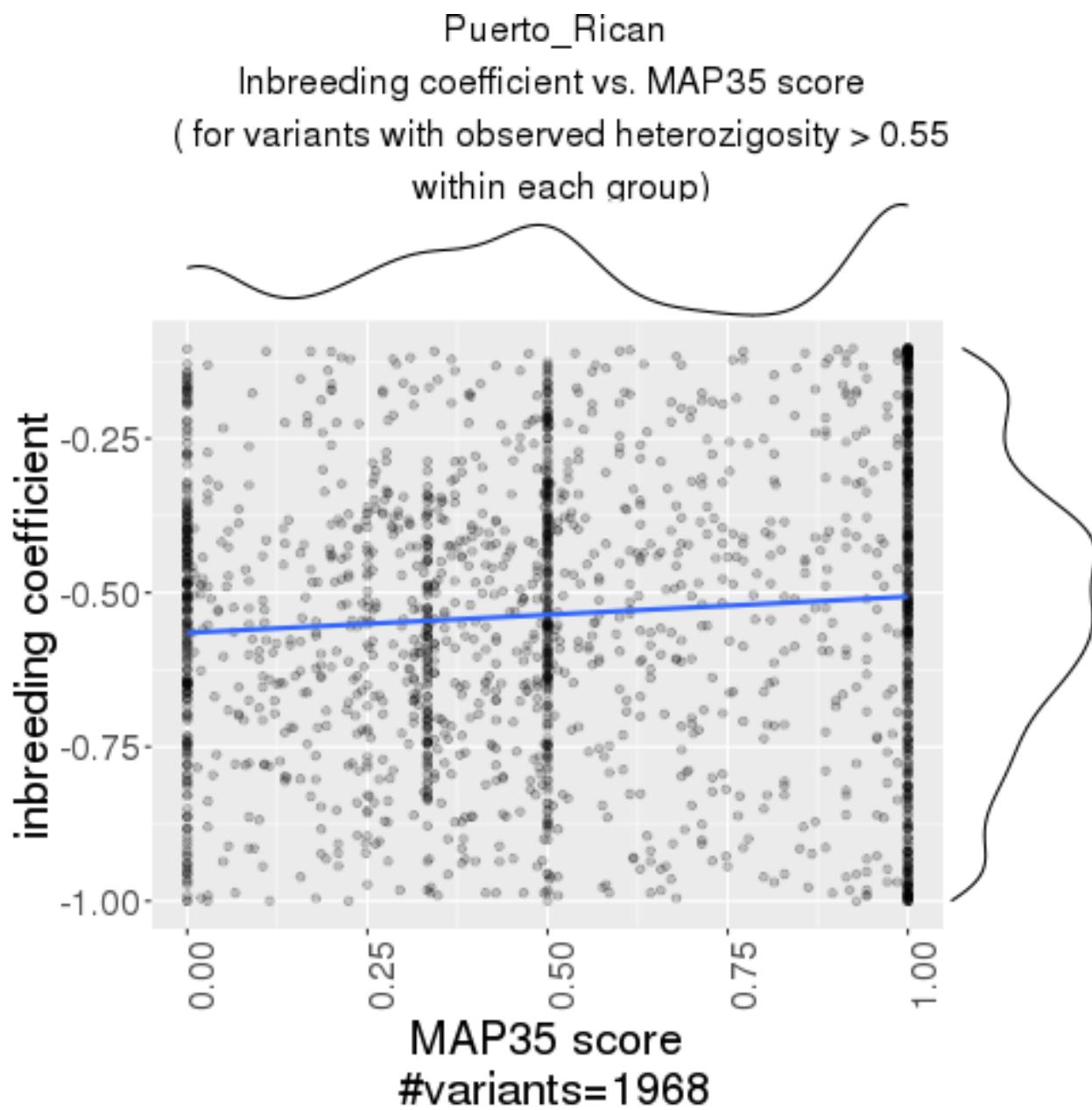


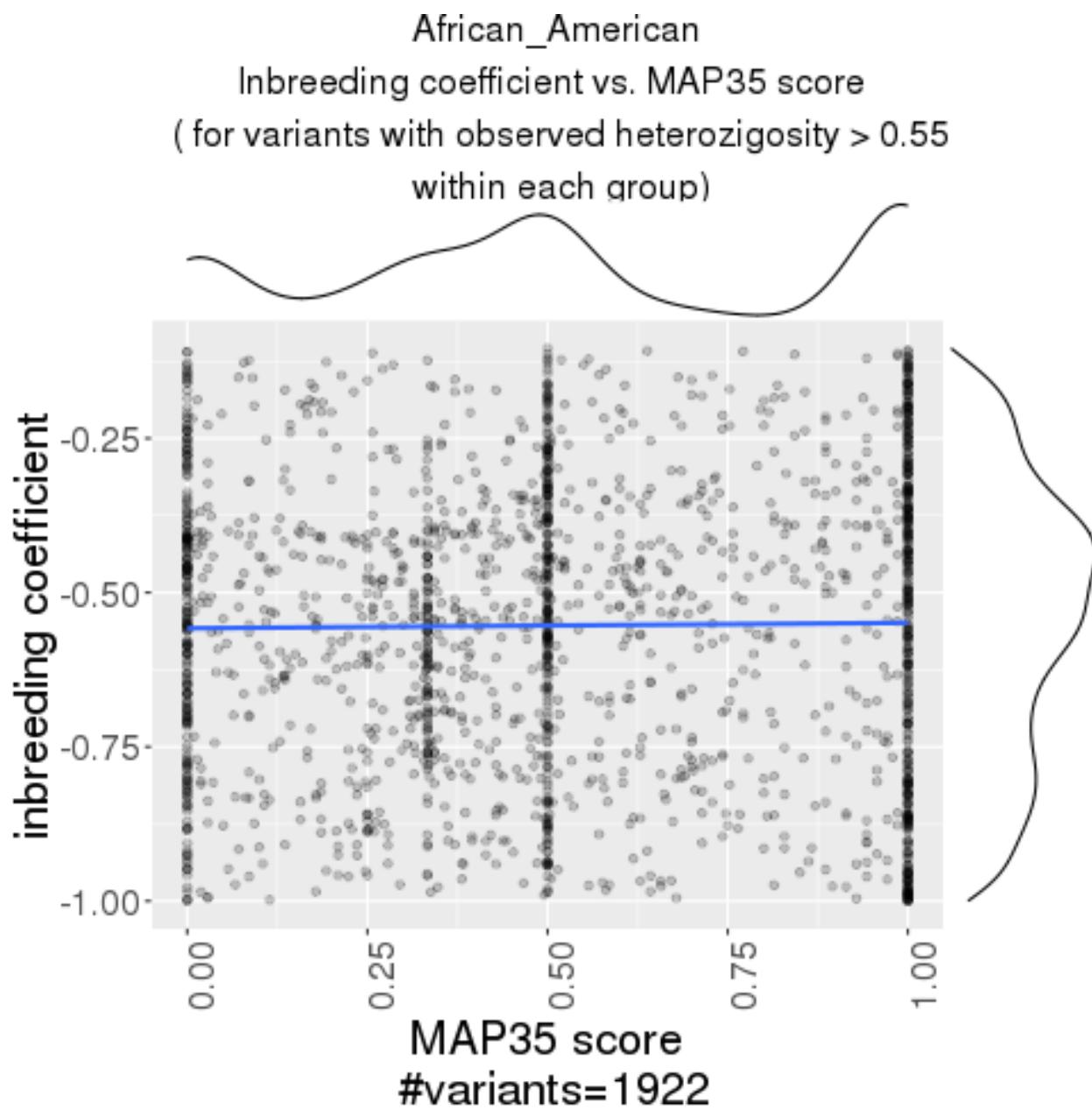




Amish
Inbreeding coefficient vs. MAP35 score
(for variants with observed heterozygosity > 0.55
within each group)







Plot 10. Distribution of high heterozygous SNPs with obs. het > 0.55 on chromosomes

