---

title: " Exploratory Data Analysis of High Heterozygosity Variants in Freeze2"

date: “July 14, 2016”

output: pdf_document

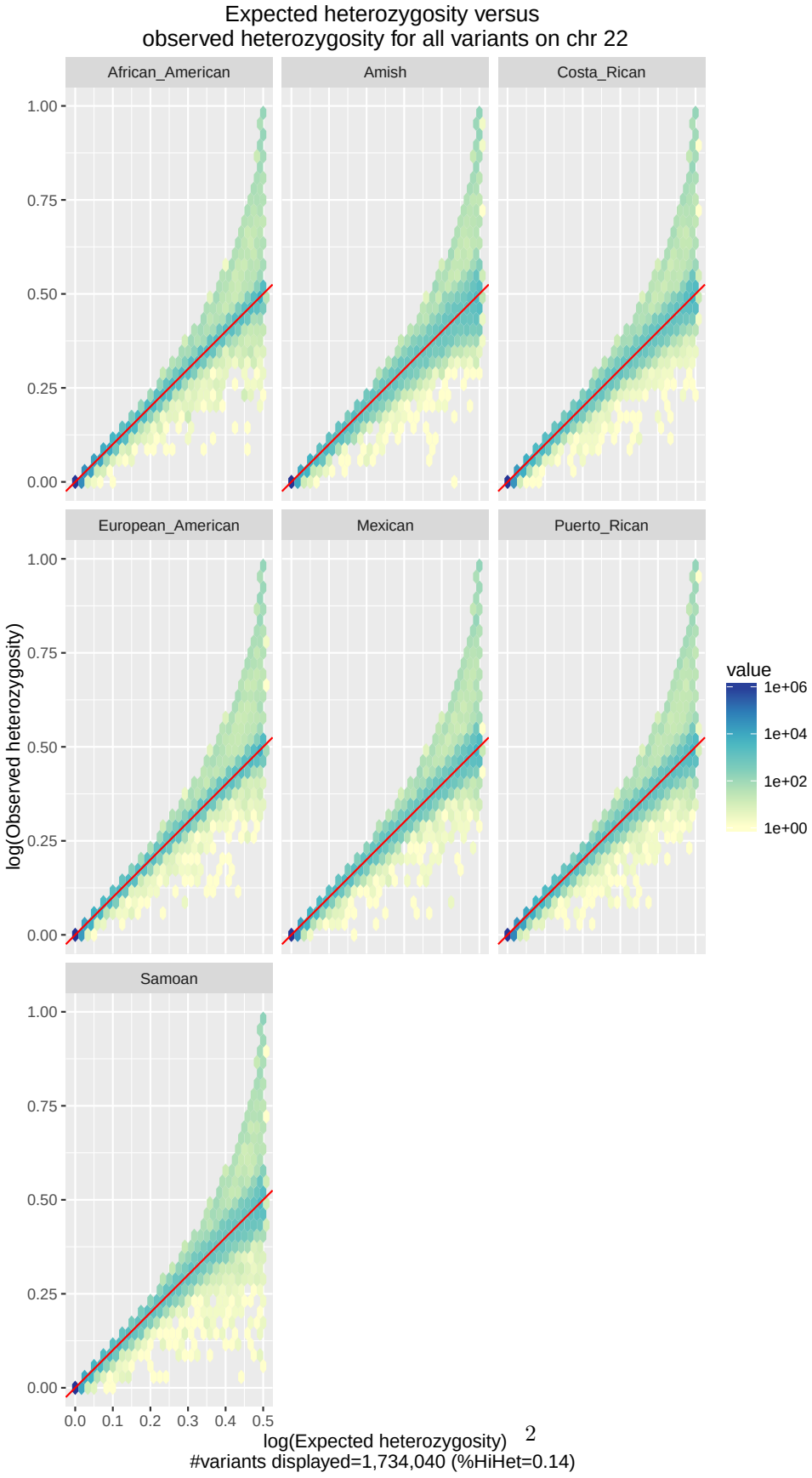
Files that were used in the analysis:

1. File with TOPMed InDel Annotations (courtesy of Xiaoming Liu)
2. Feeeze 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

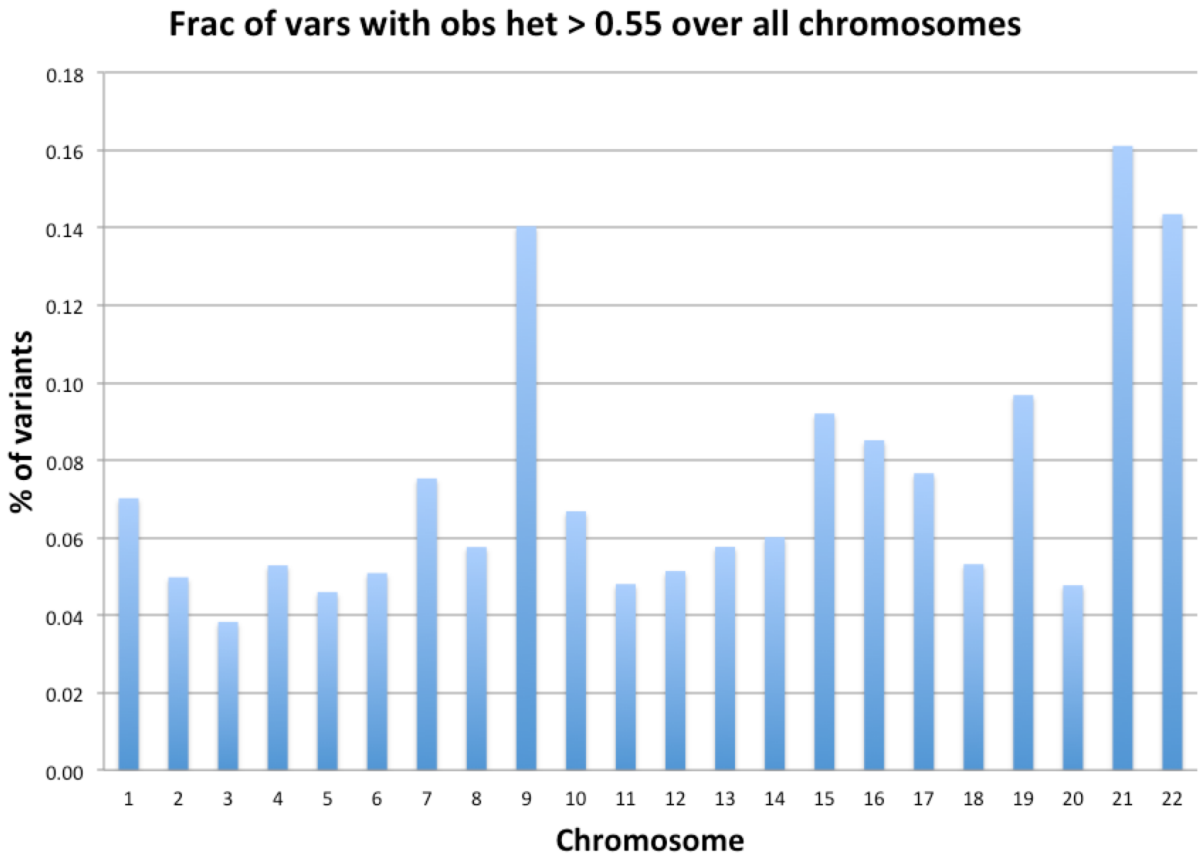


Figure 1:

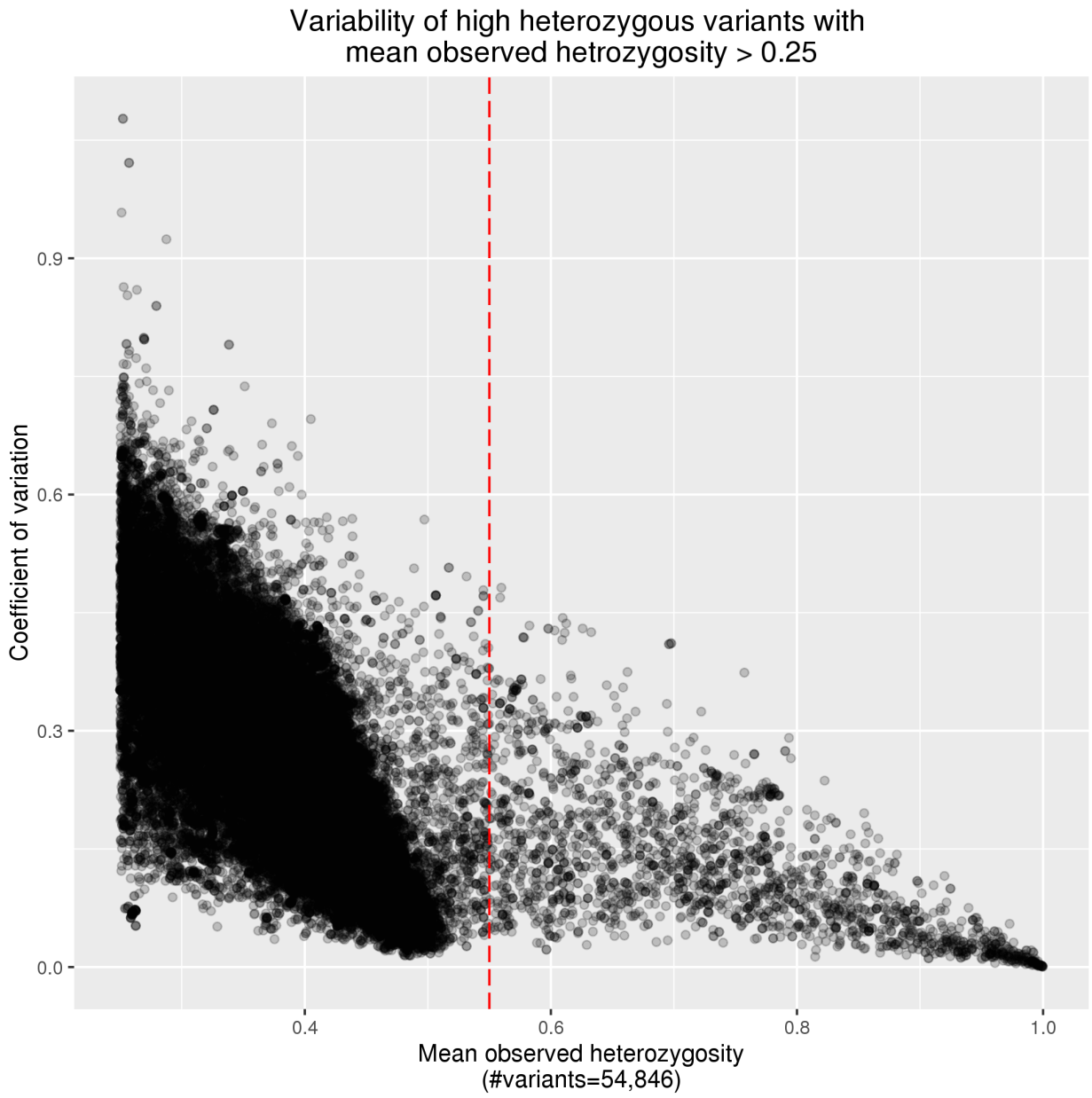

Variability of high heterozygous variants with
mean observed hetrozygosity > 0.25
Coefficient of variation
0.0
0.3
0.6
0.9
0.4
0.6
0.8
1.0
Mean observed heterozygosity
(#variants=54,846)

Figure 2:

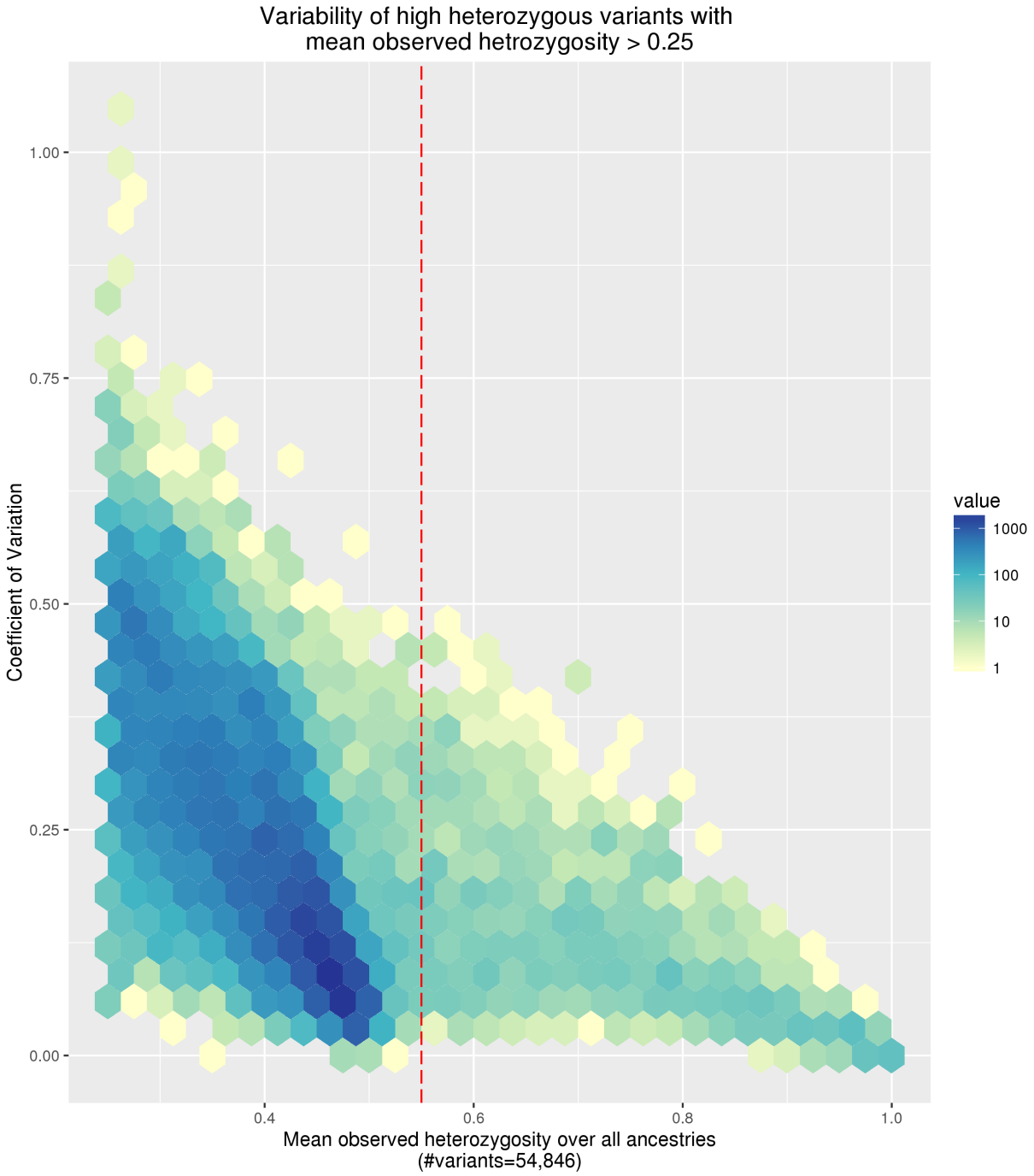

Variability of high heterozygous variants with
mean observed hetrozygosity > 0.25
Coefficient of Variation
1.00
0.75
0.50
0.25
0.00
0.4
0.6
0.8
1.0
Mean observed heterozygosity over all ancestries
(#variants=54,846)
value
1000
100
10
1

Figure 3:

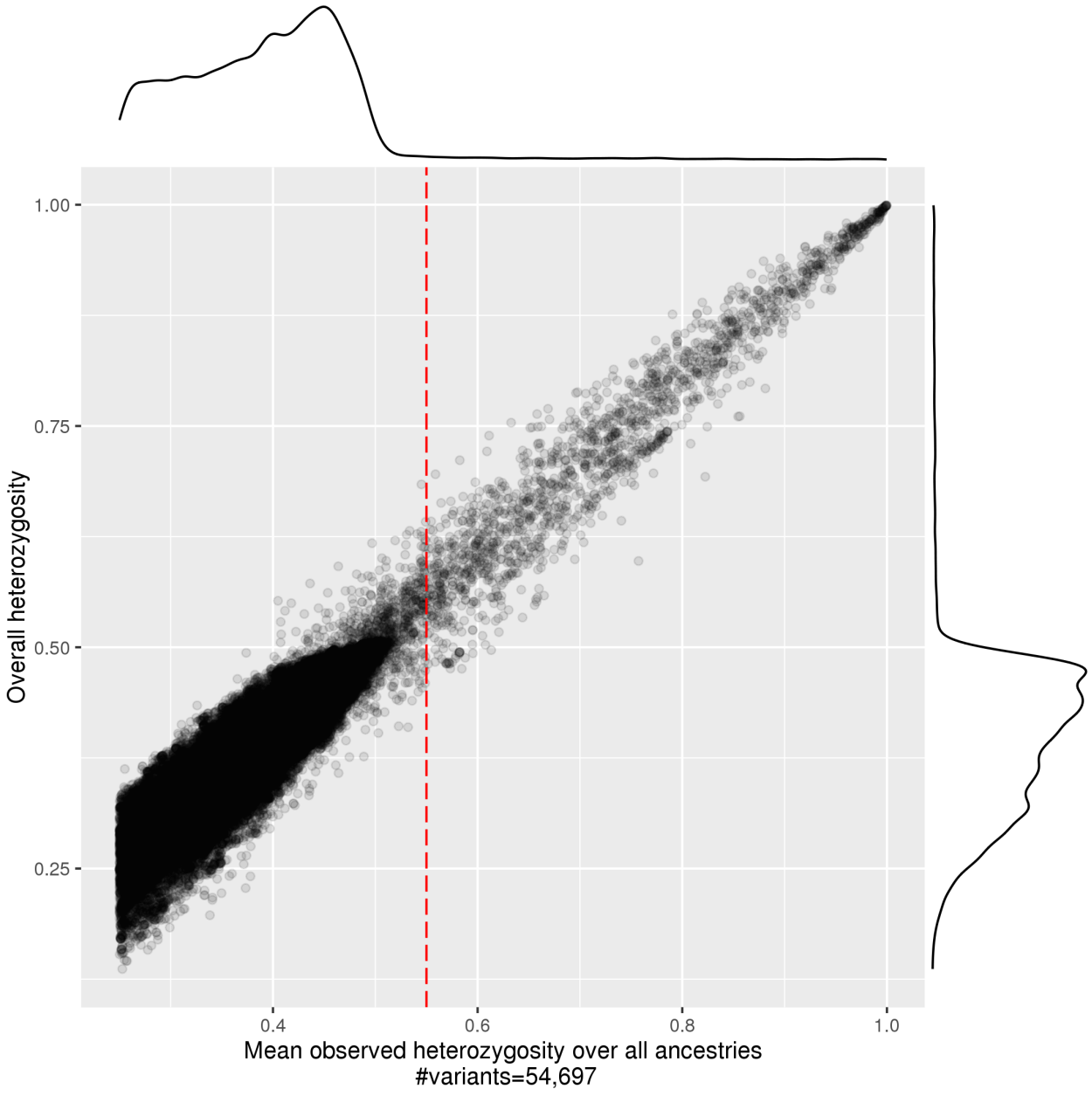

1.00
0.75
0.50
0.25
Overall heterozygosity
0.4
0.6
0.8
1.0
Mean observed heterozygosity over all ancestries
#variants=54,697

Figure 4:

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

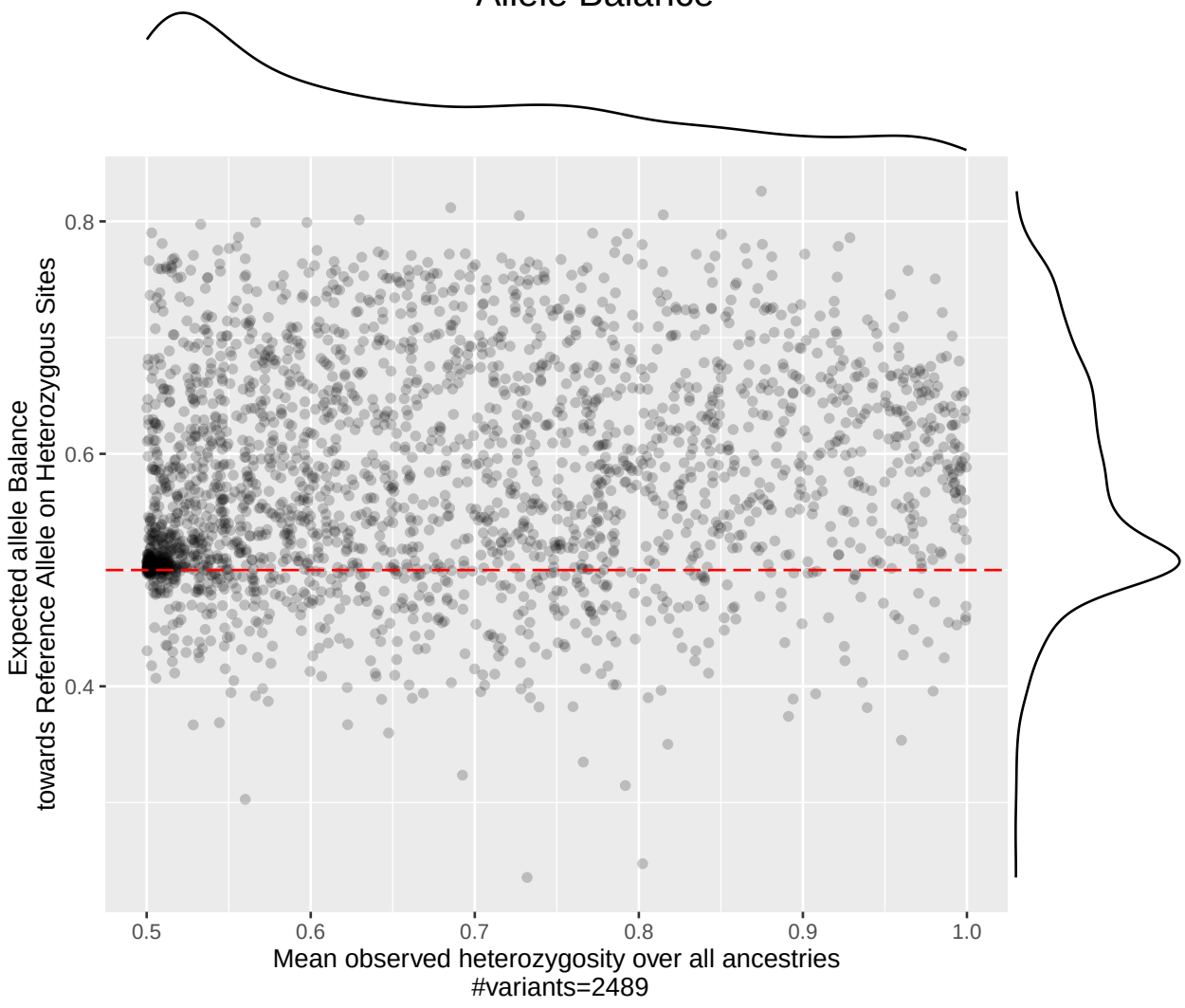

Variability of high heterozygous
variants with mean observed heterozygosity > 0.5:
Mean observed heterozygosity vs.
Allele Balance
Expected allele Balance
towards Reference Allele on Heterozygous Sites
0.8
0.6
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Mean observed heterozygosity over all ancestries
#variants=2489

Figure 5:

Plot 6. Allele Balance versus overall heterozygosity

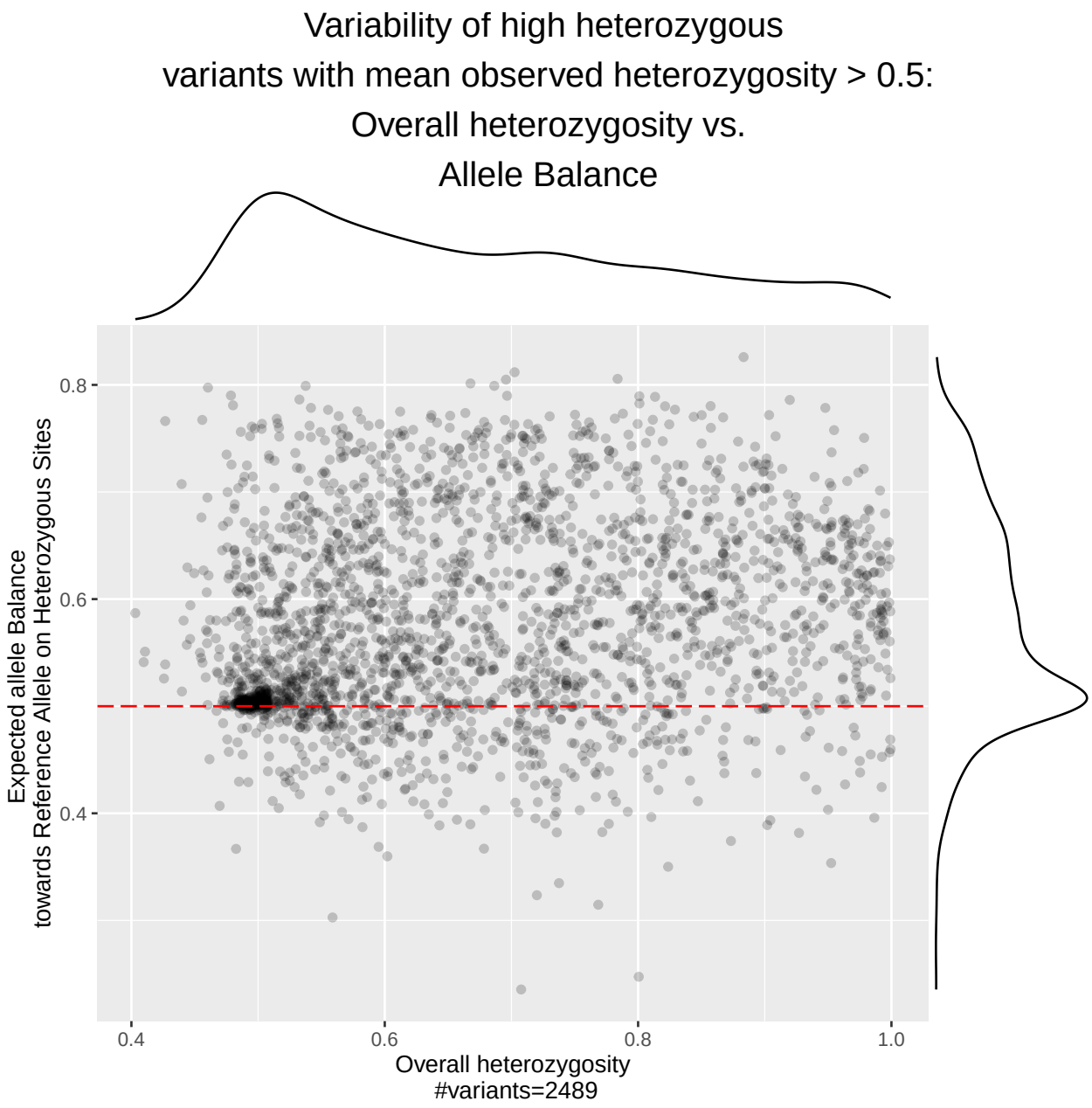


Figure 6:

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

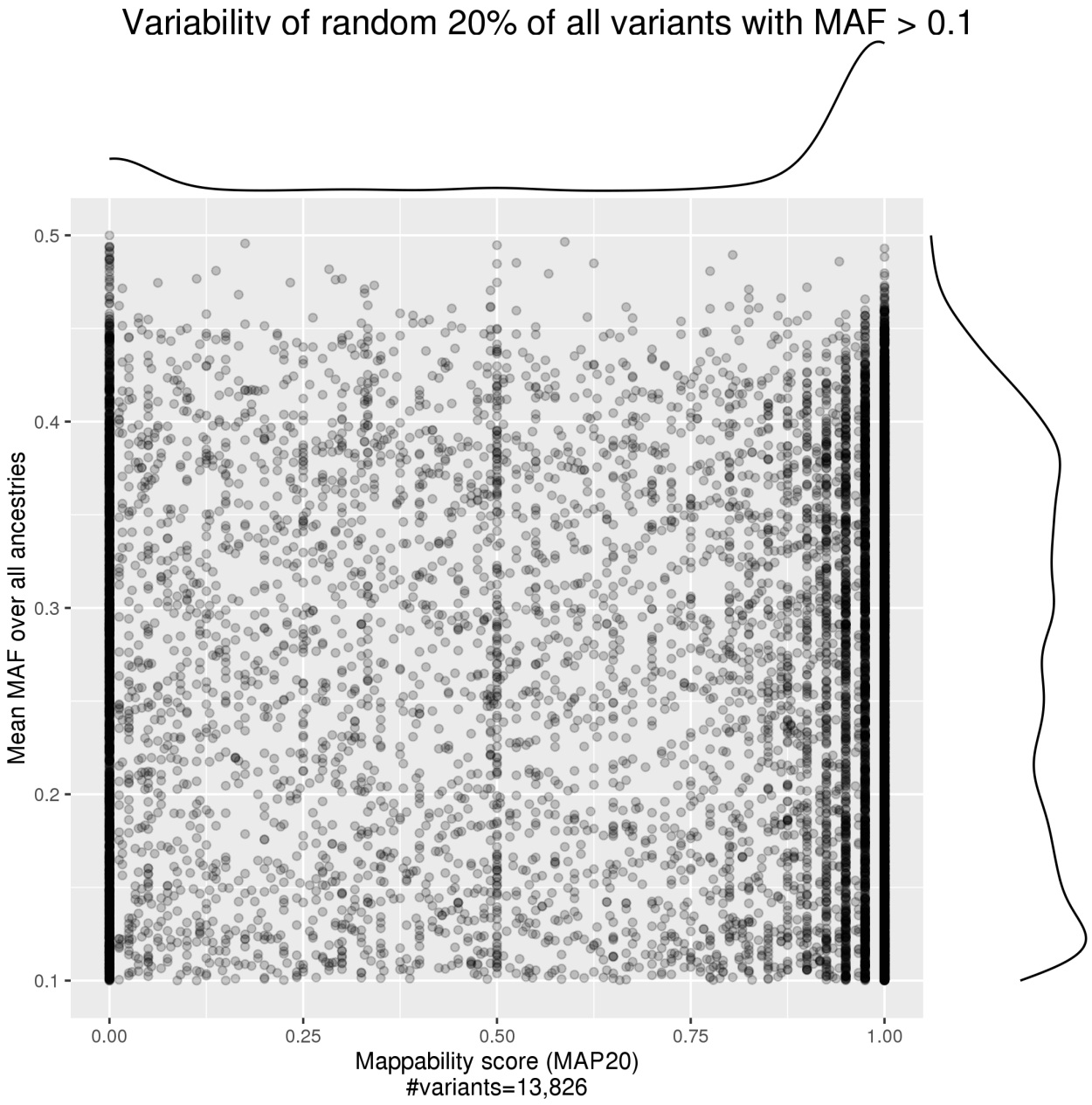


Figure 7:

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

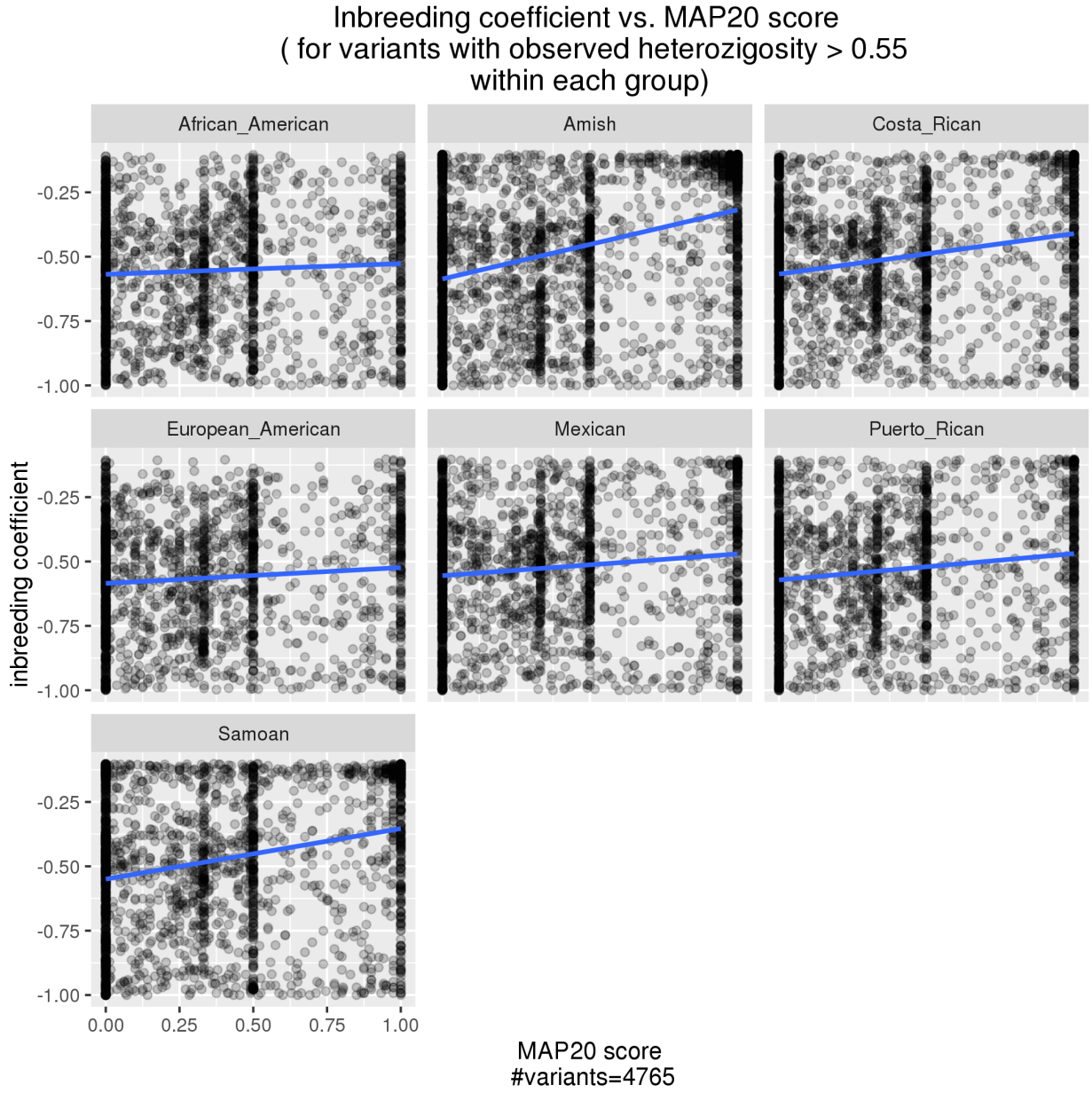


Figure 8:

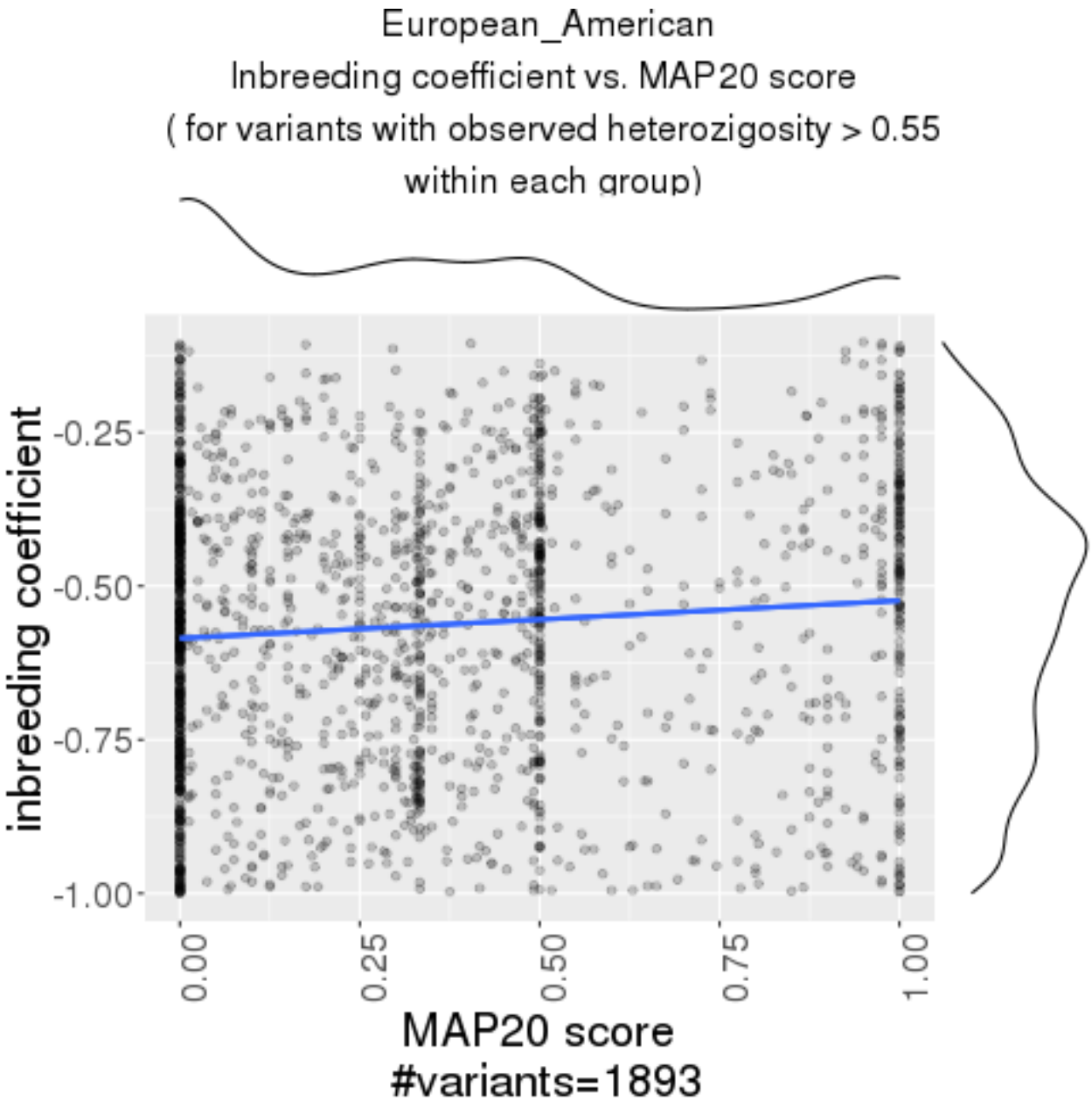

European_American
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP20 score
#variants=1893

Figure 9:

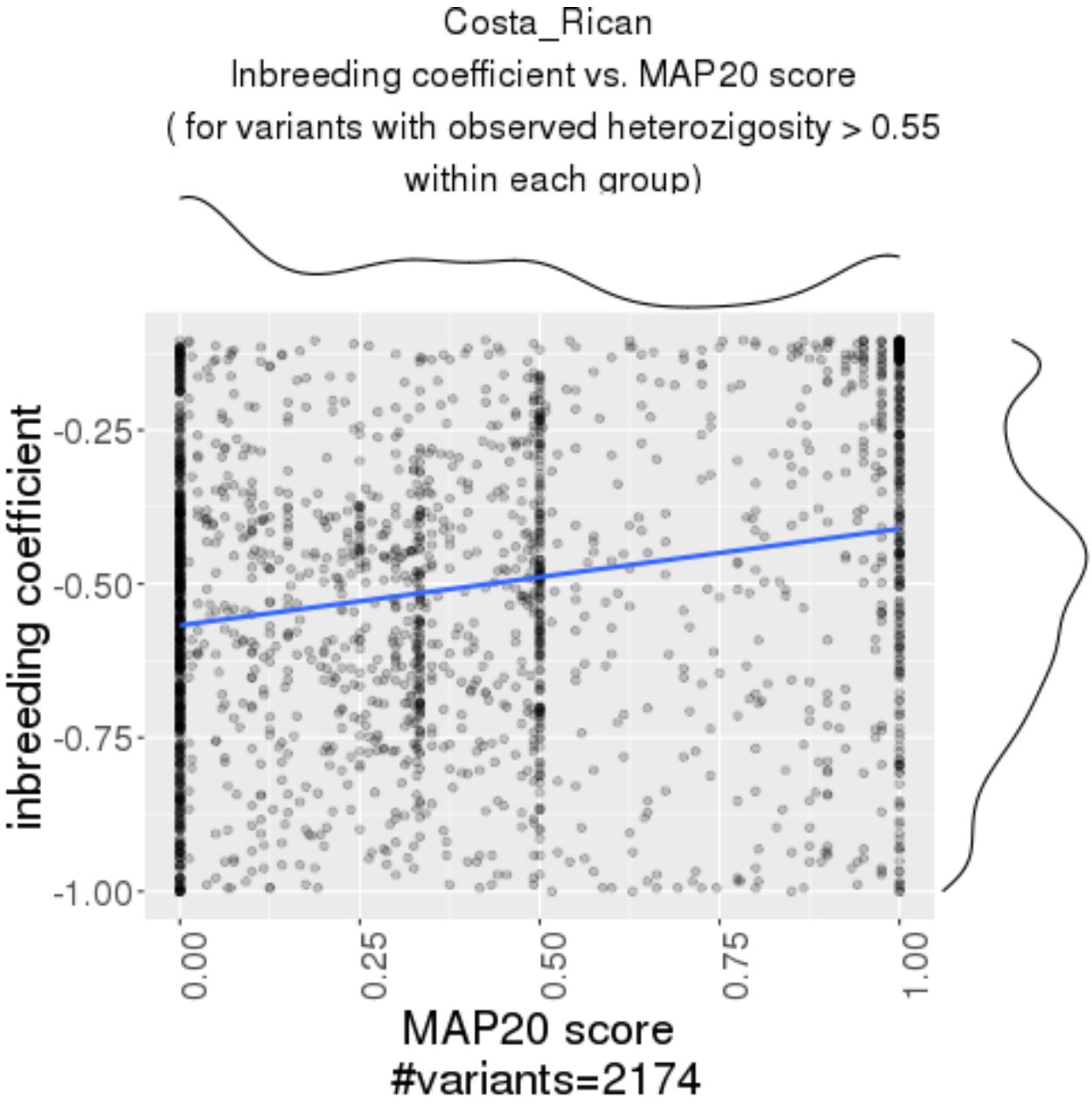

Costa_Rican
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP20 score
#variants=2174

Figure 10:

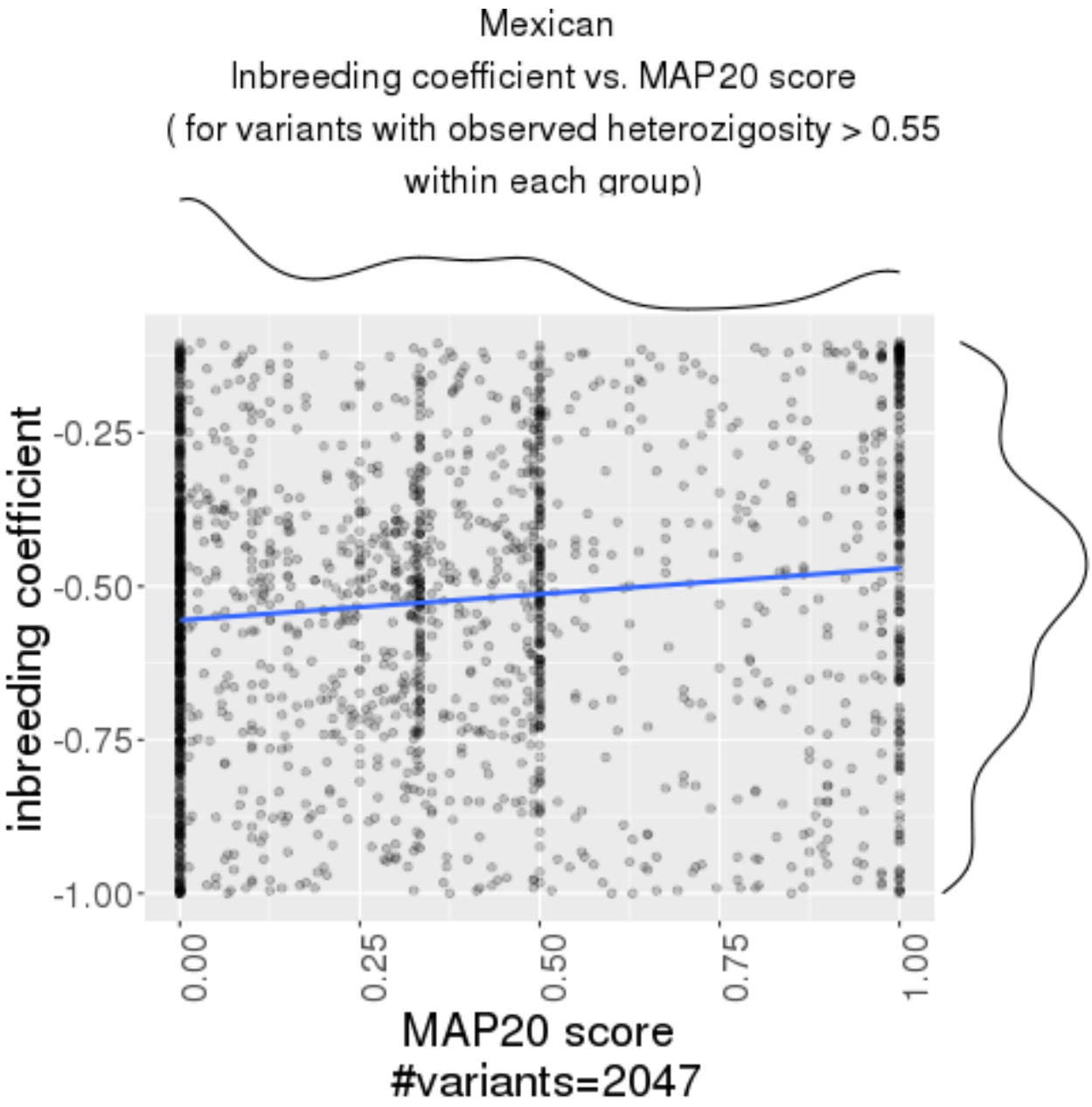

Mexican
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP20 score
#variants=2047

Figure 11:

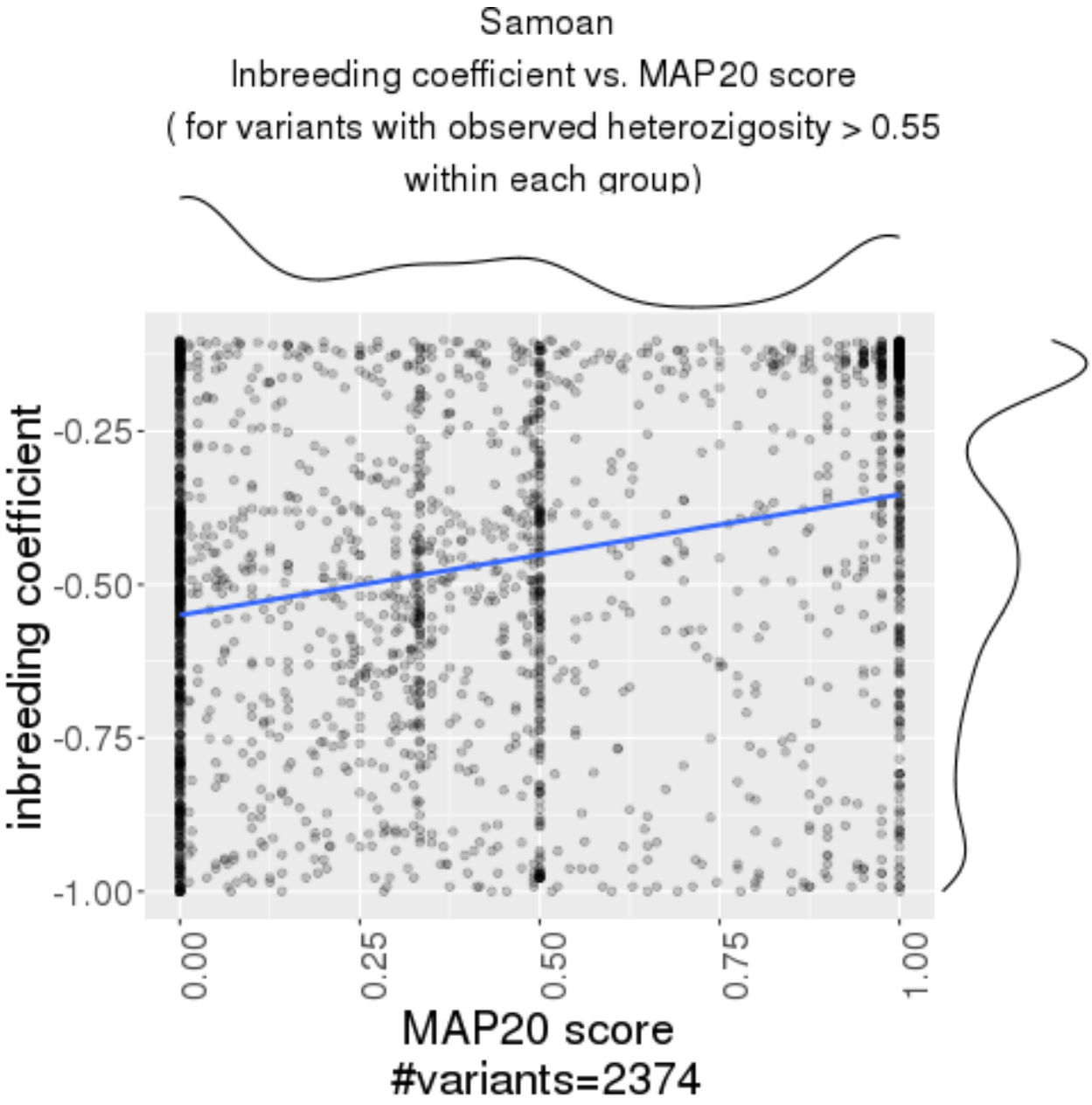

Samoan
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP20 score
#variants=2374

Figure 12:

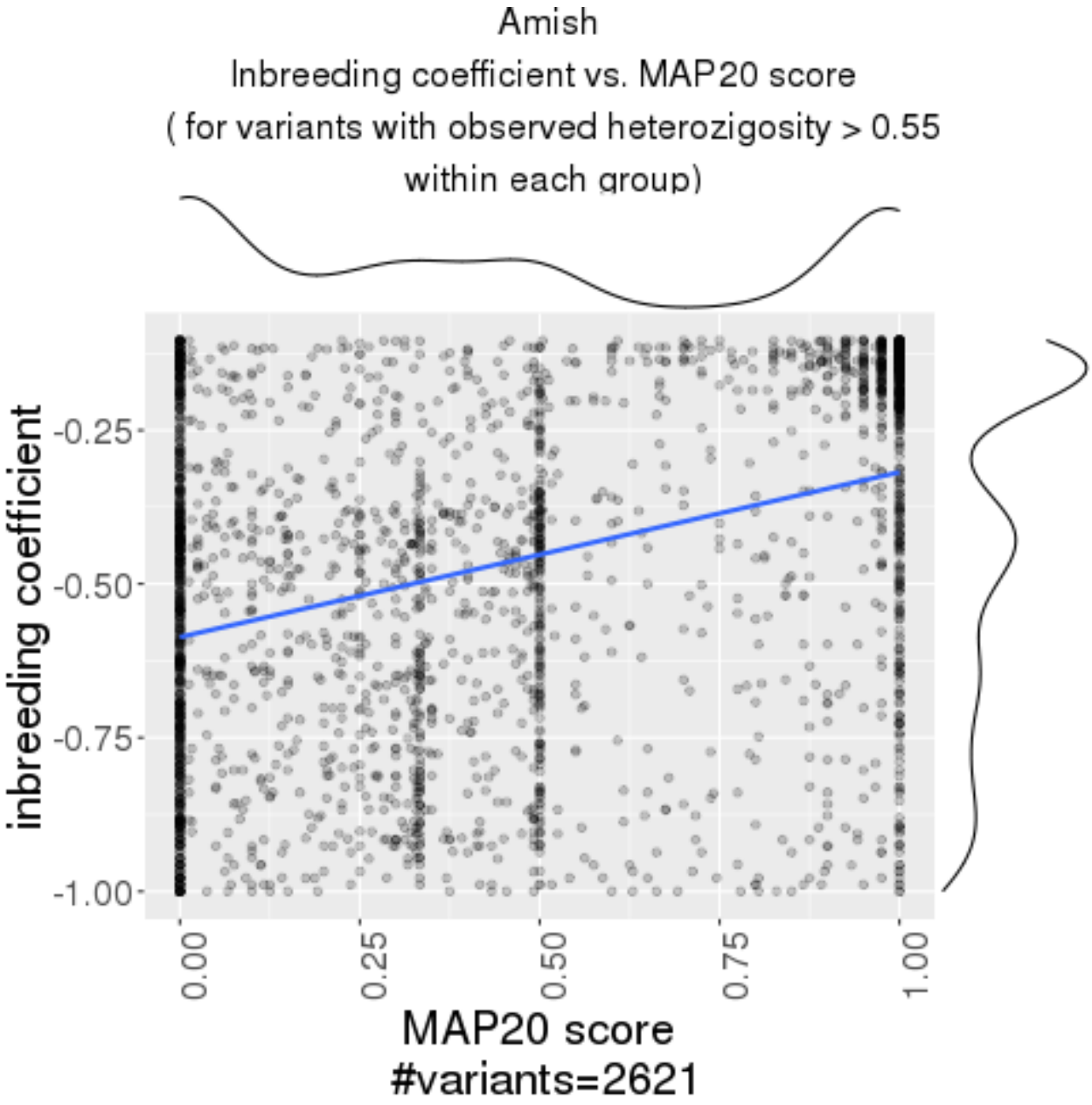

Amish
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP20 score
#variants=2621

Figure 13:

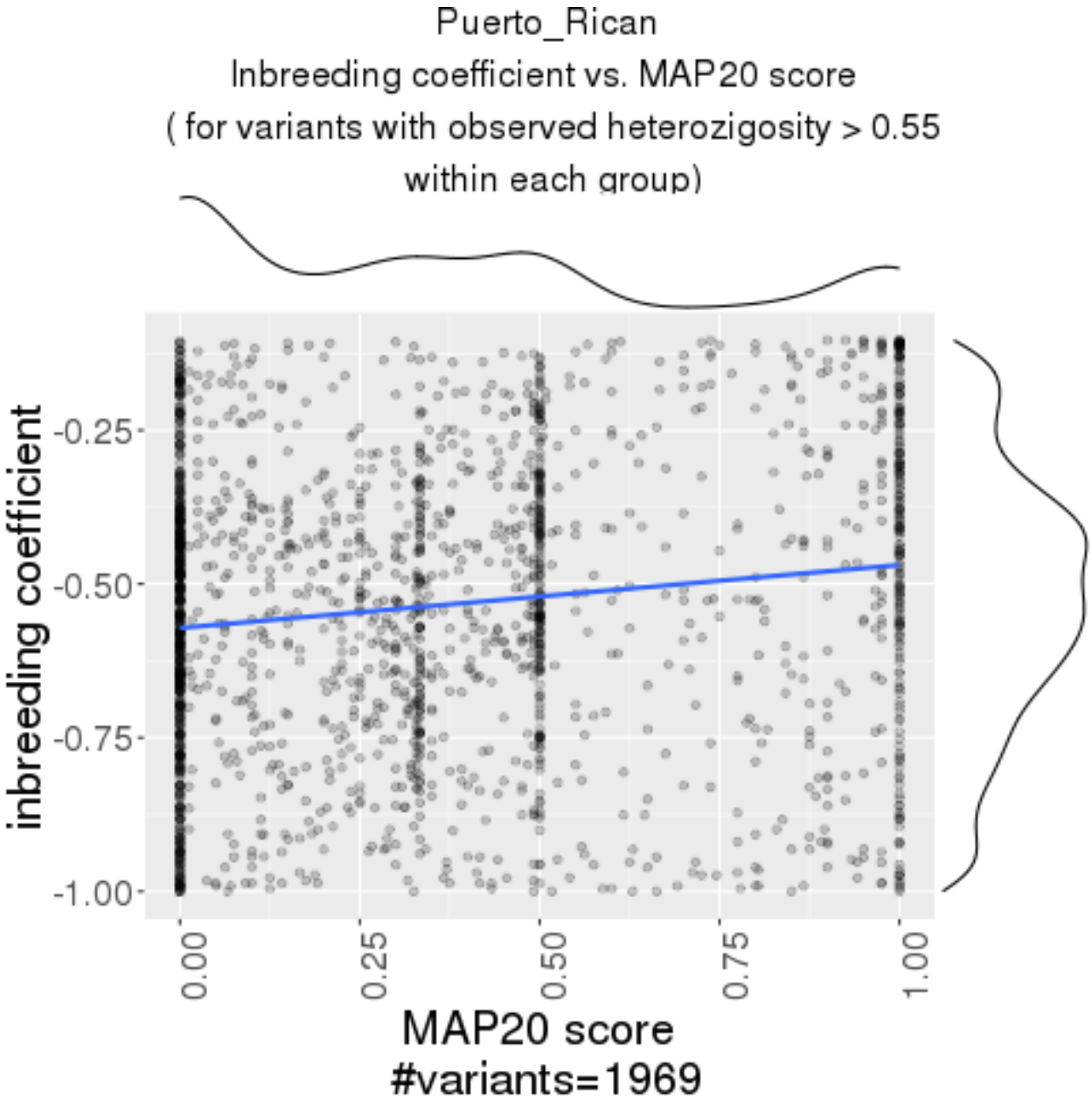

Puerto_Rican
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP20 score
#variants=1969

Figure 14:

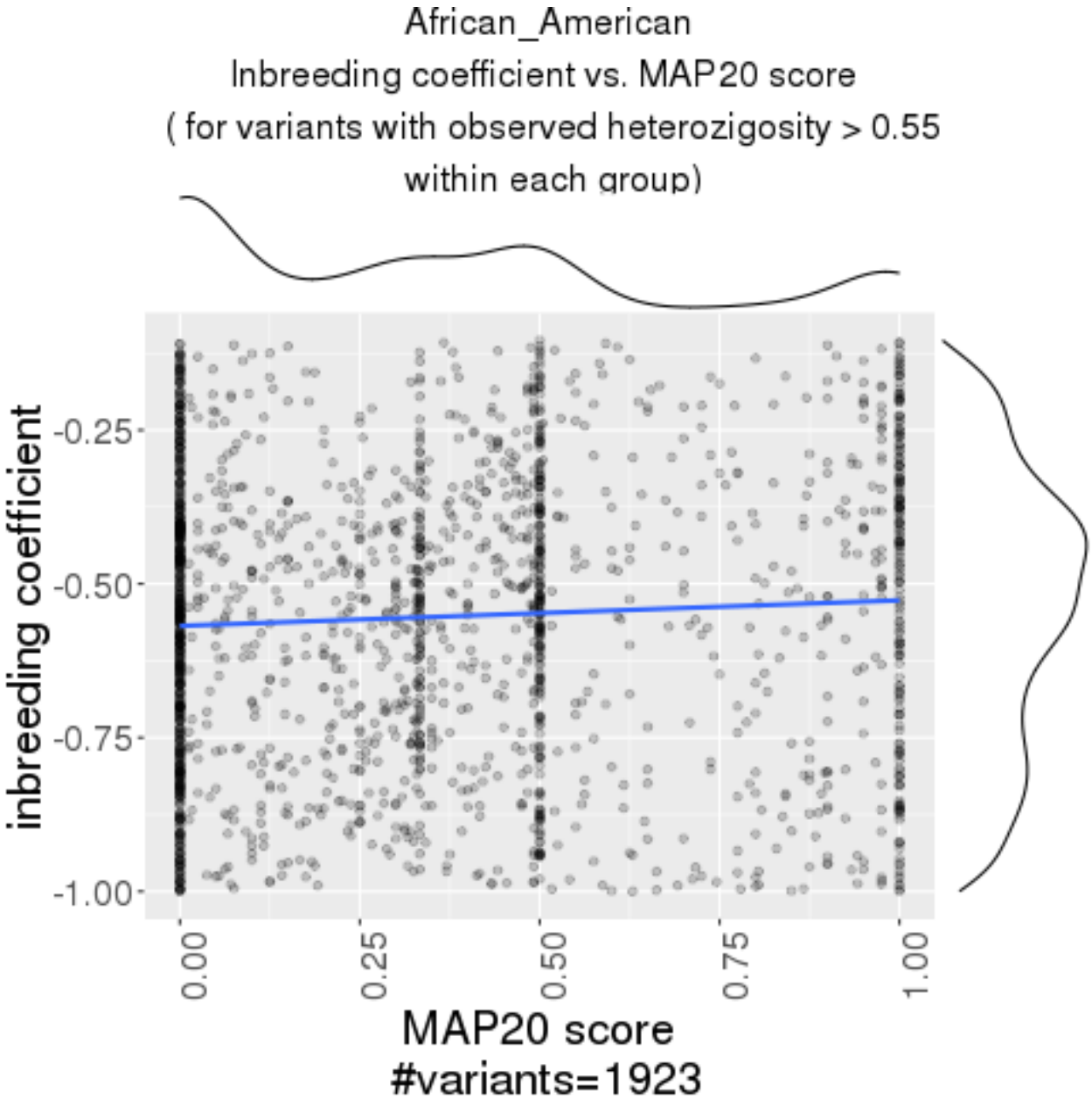

African_American
Inbreeding coefficient vs. MAP20 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP20 score
#variants=1923

Figure 15:

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

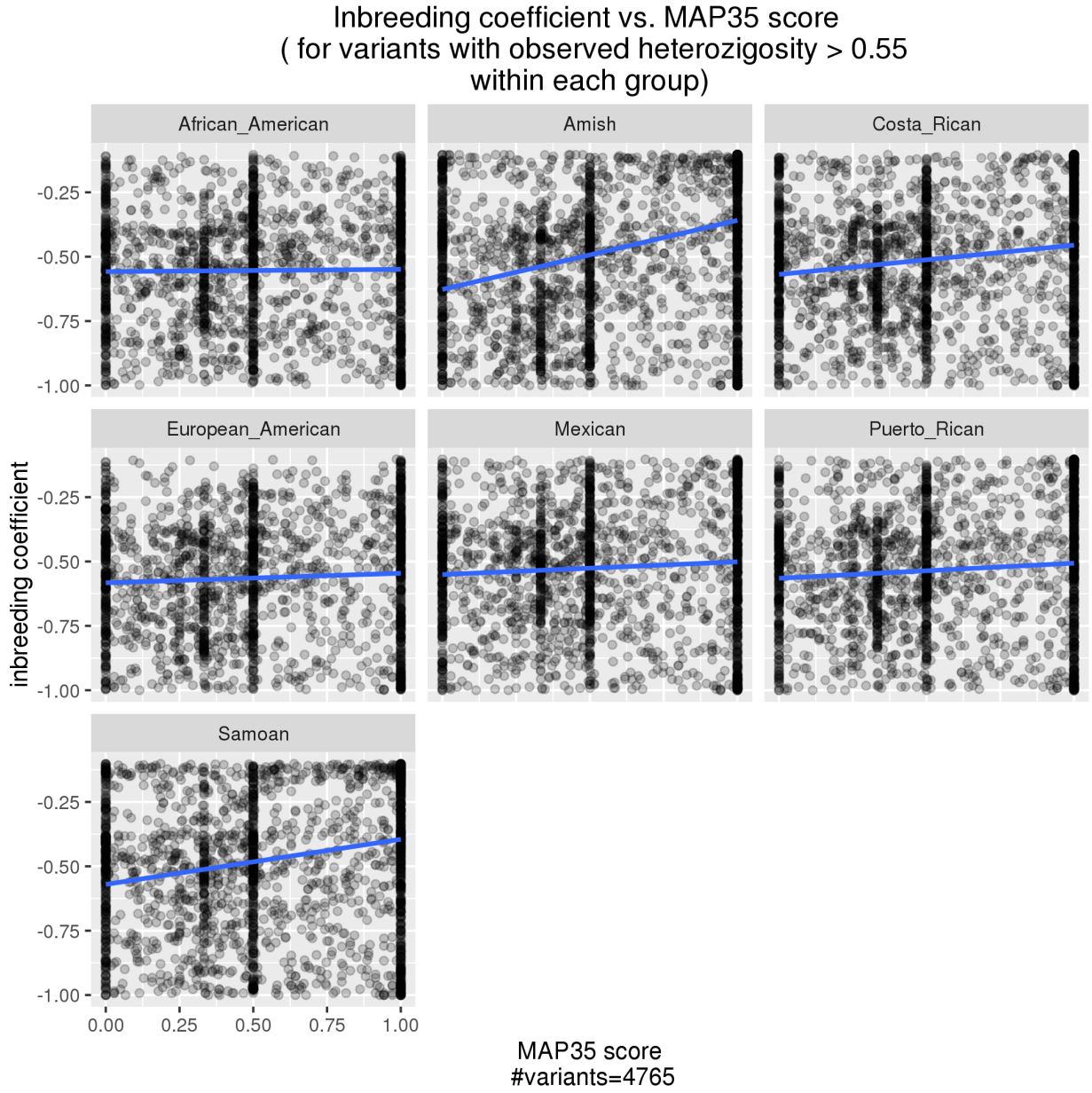


Figure 16:

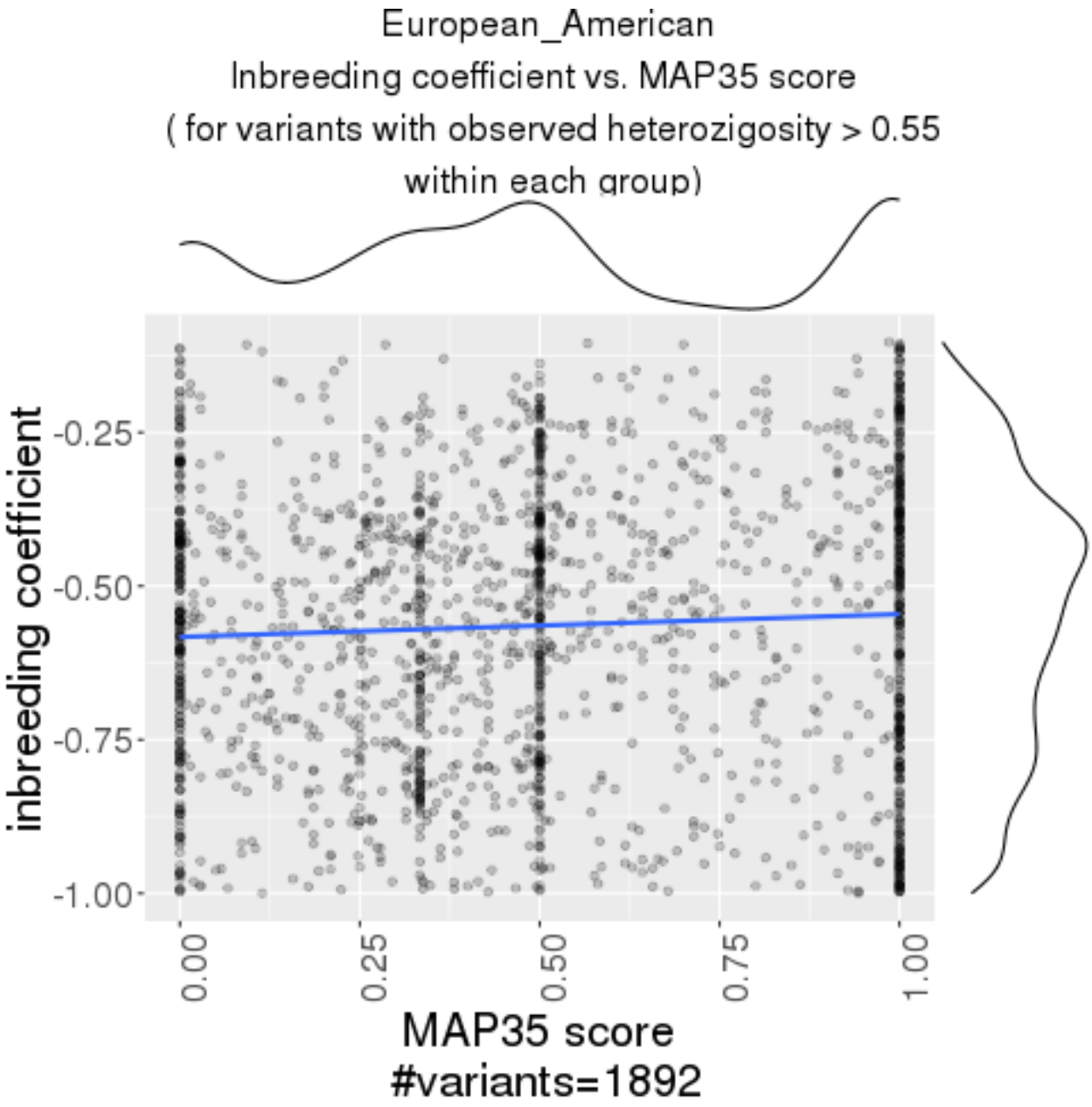

European_American
Inbreeding coefficient vs. MAP35 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP35 score
#variants=1892

Figure 17:

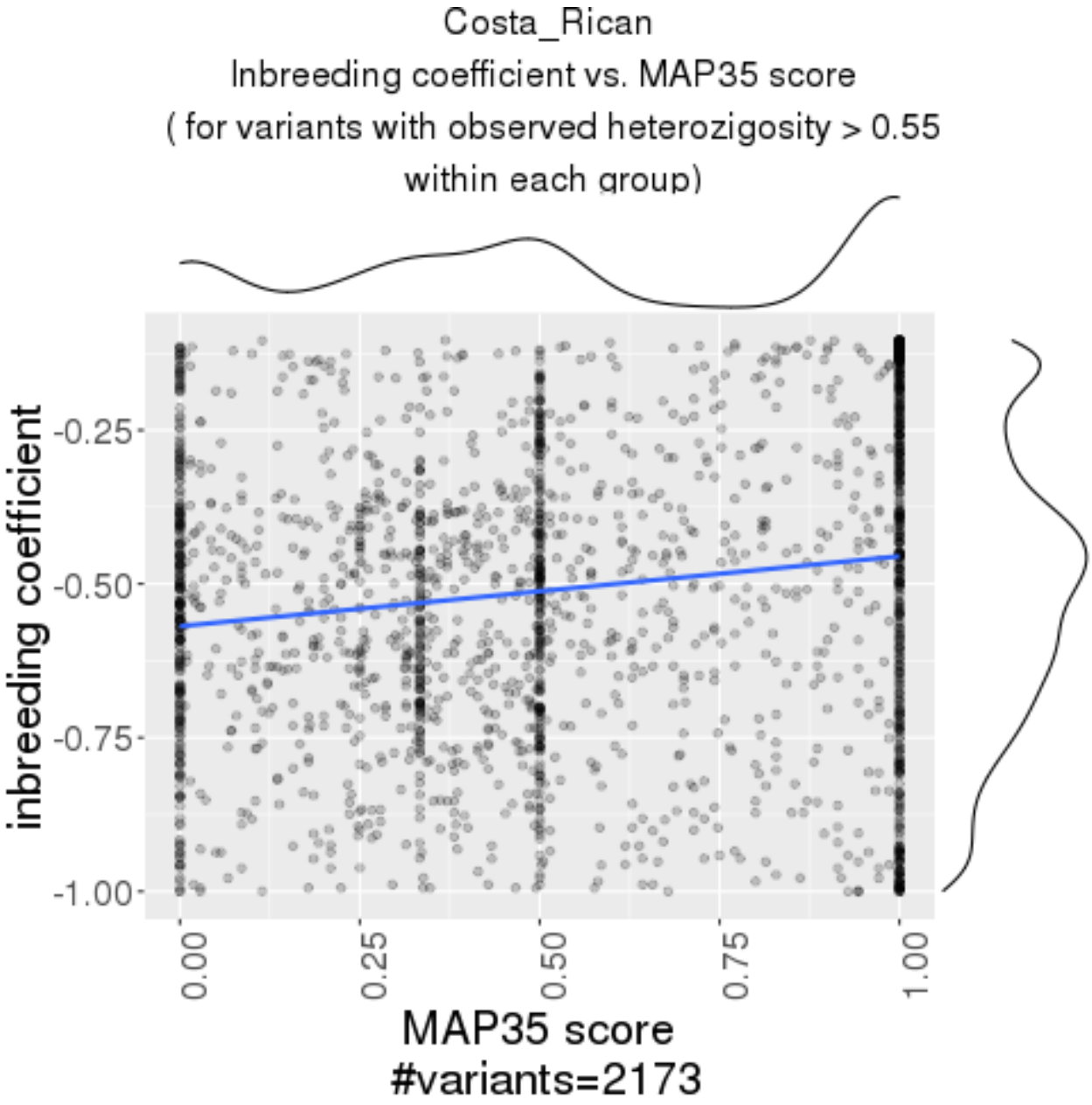

Costa_Rican
Inbreeding coefficient vs. MAP35 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP35 score
#variants=2173

Figure 18:

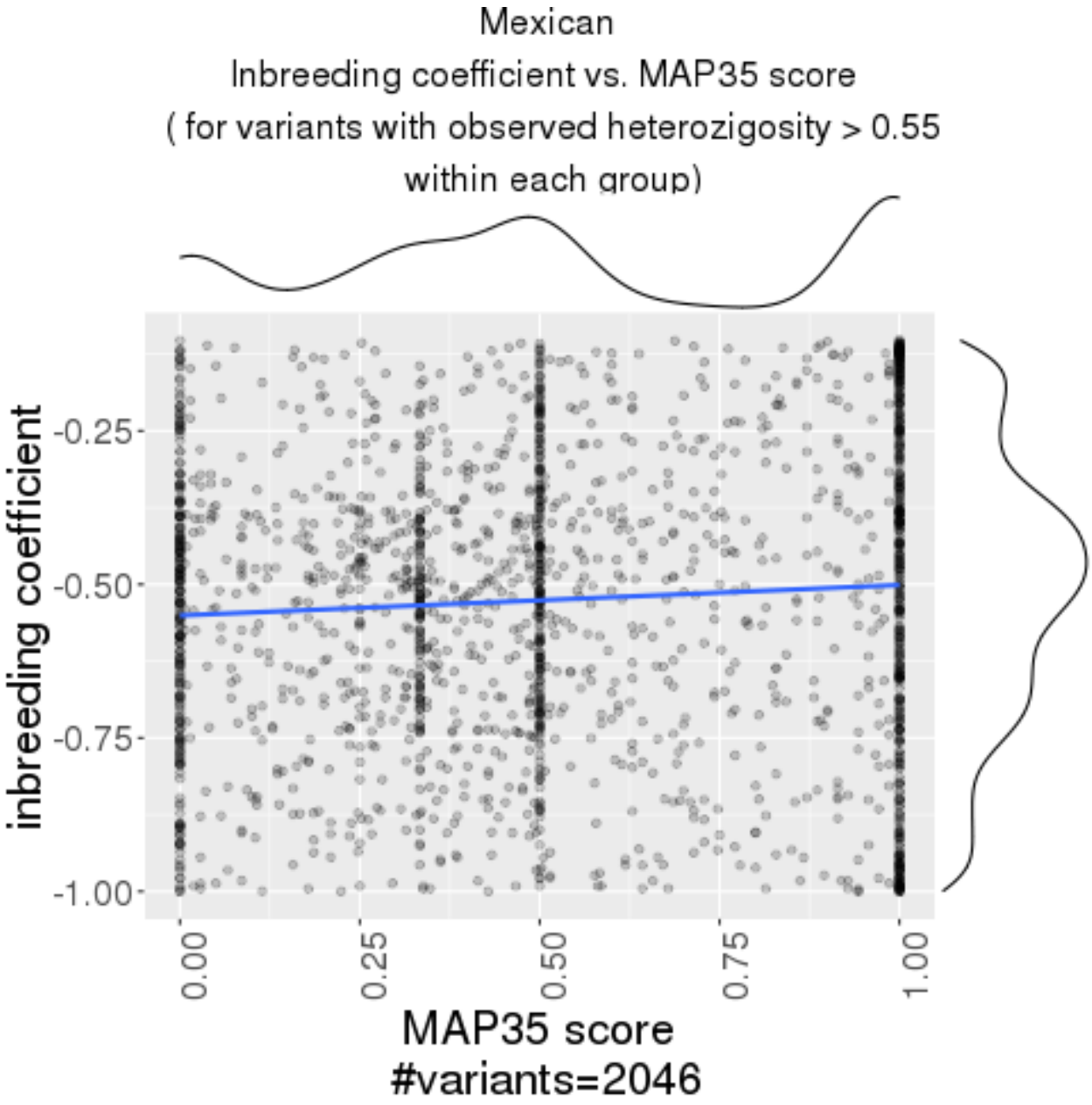

Mexican
Inbreeding coefficient vs. MAP35 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP35 score
#variants=2046

Figure 19:

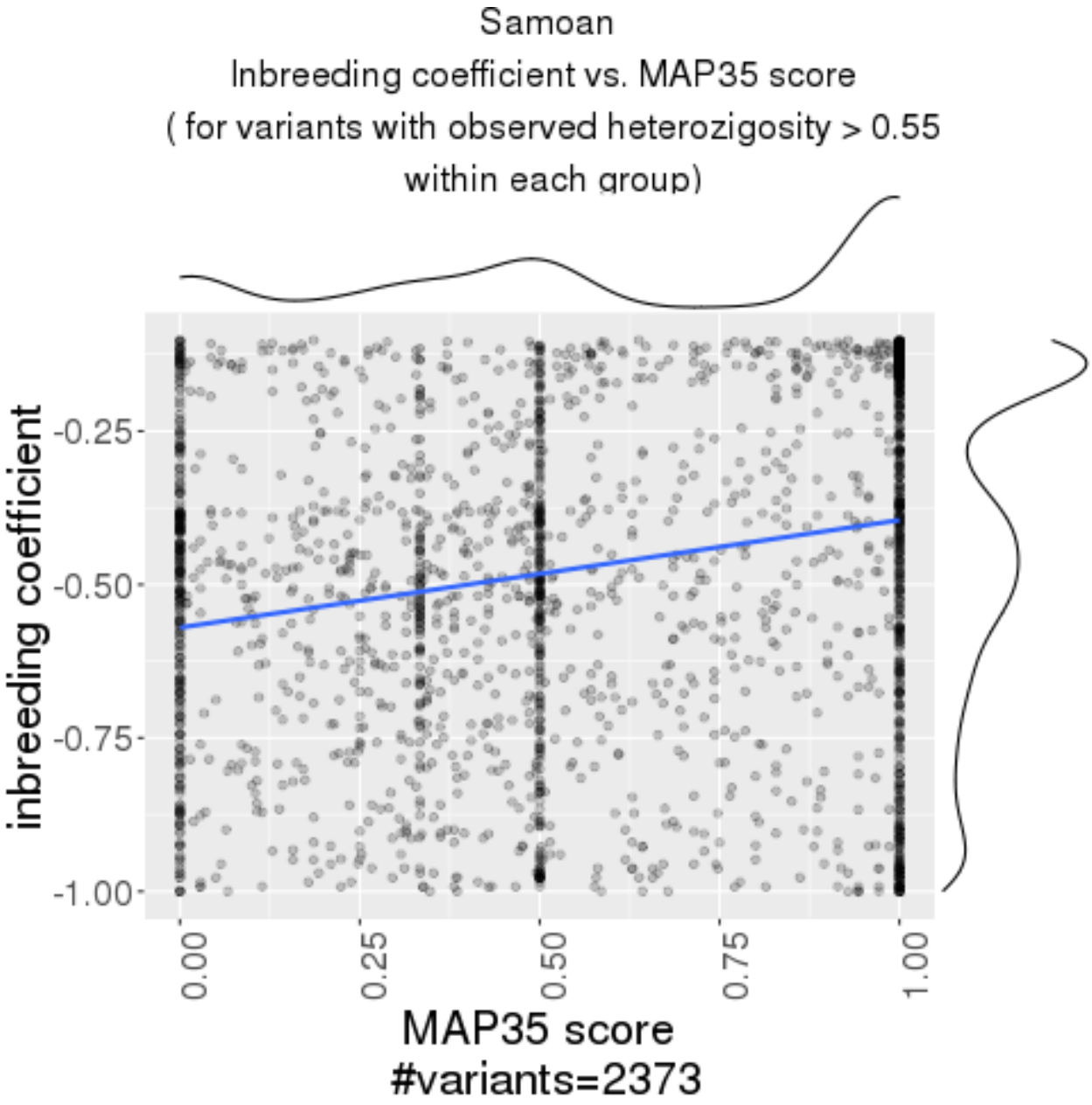

Samoan
Inbreeding coefficient vs. MAP35 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP35 score
#variants=2373

Figure 20:

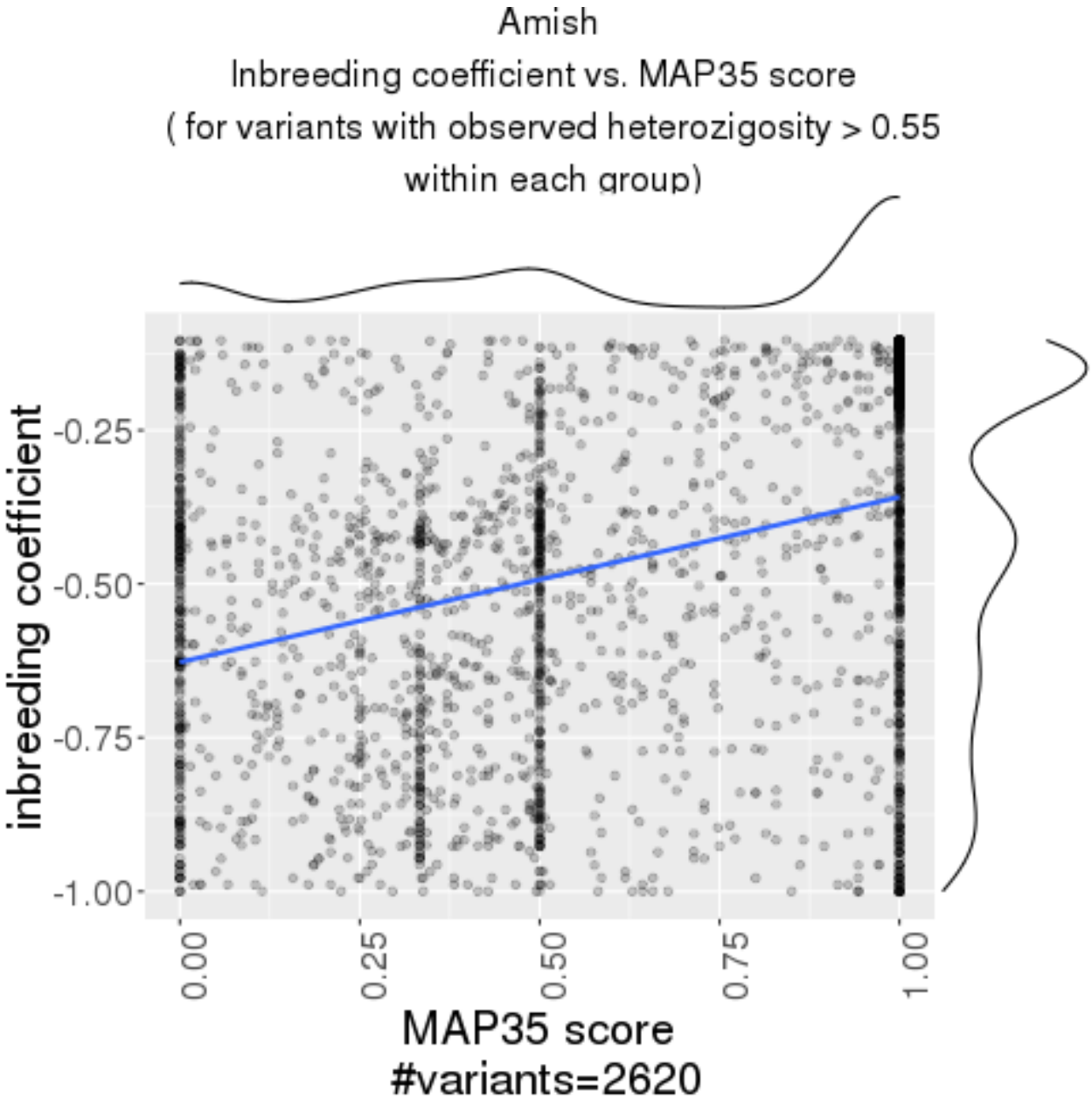

Amish
Inbreeding coefficient vs. MAP35 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP35 score
#variants=2620

Figure 21:

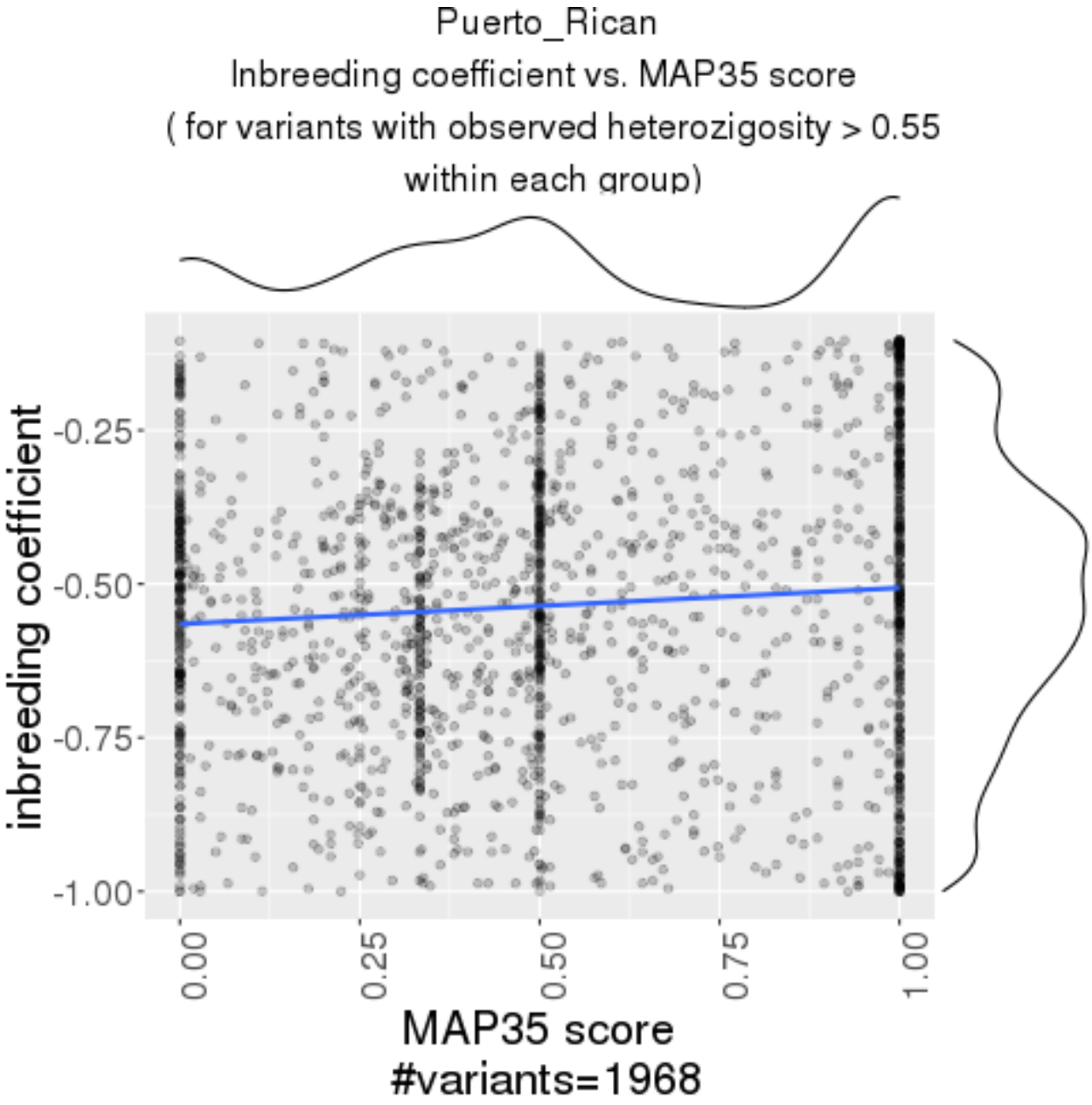

Puerto_Rican
Inbreeding coefficient vs. MAP35 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP35 score
#variants=1968

Figure 22:

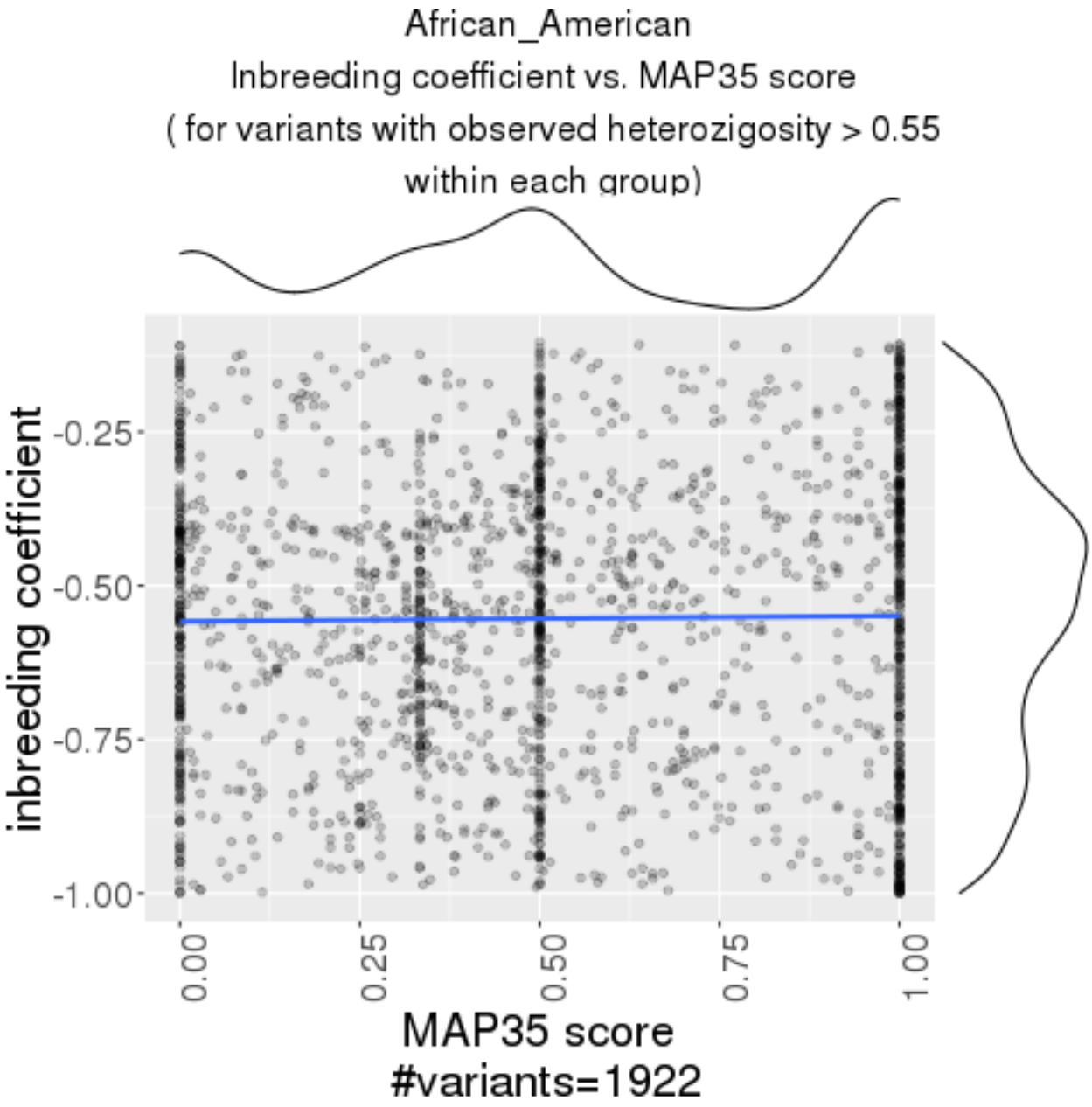

African_American
Inbreeding coefficient vs. MAP35 score
(for variants with observed heterozigosity > 0.55
within each group)
inbreeding coefficient
-0.25
-0.50
-0.75
-1.00
0.00
0.25
0.50
0.75
1.00
MAP35 score
#variants=1922

Figure 23:

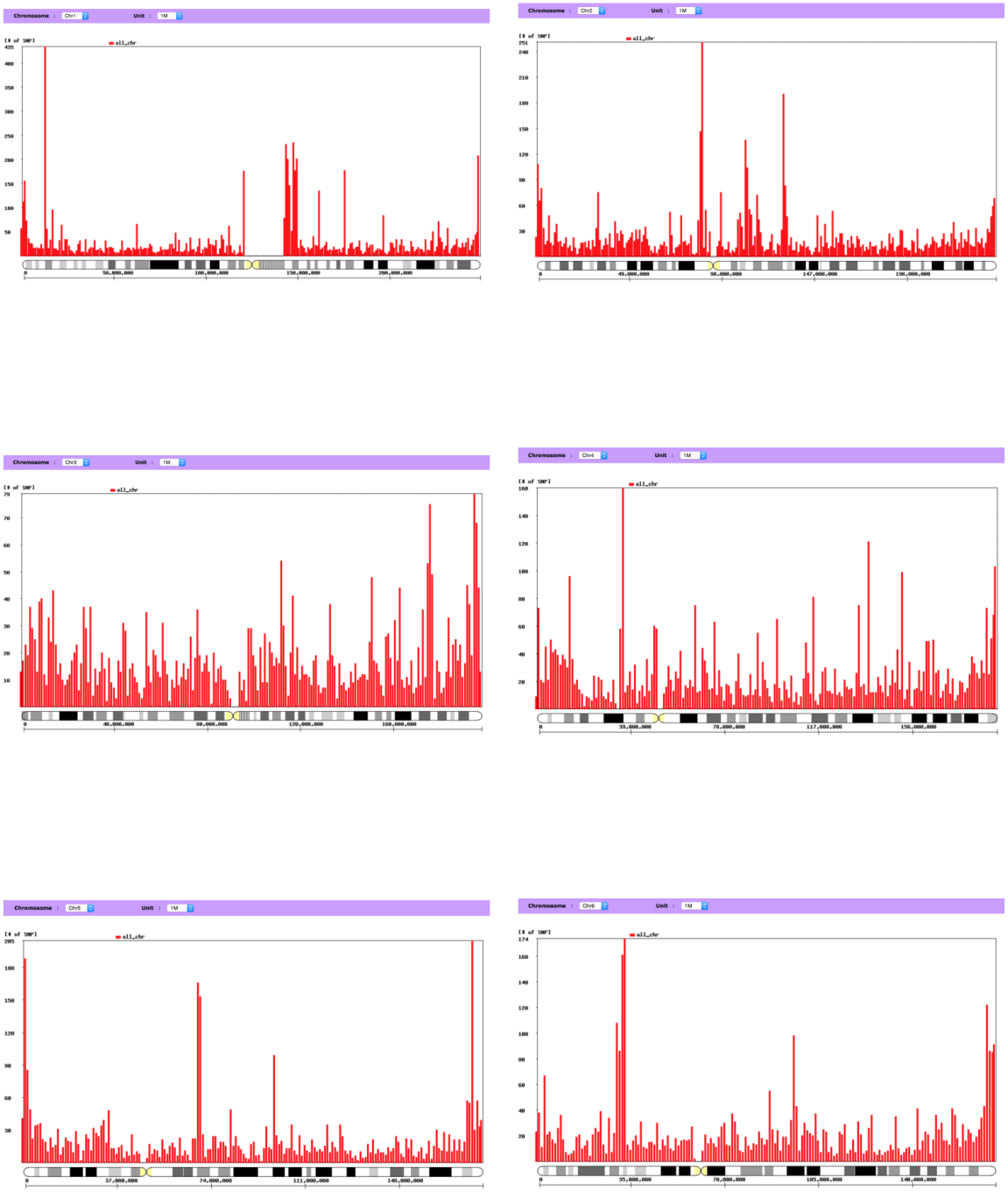

Chromosome : Chr1 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr2 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr3 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr4 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr5 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr6 Unit : 1M
[# of SNP]
all_chr

Figure 24:

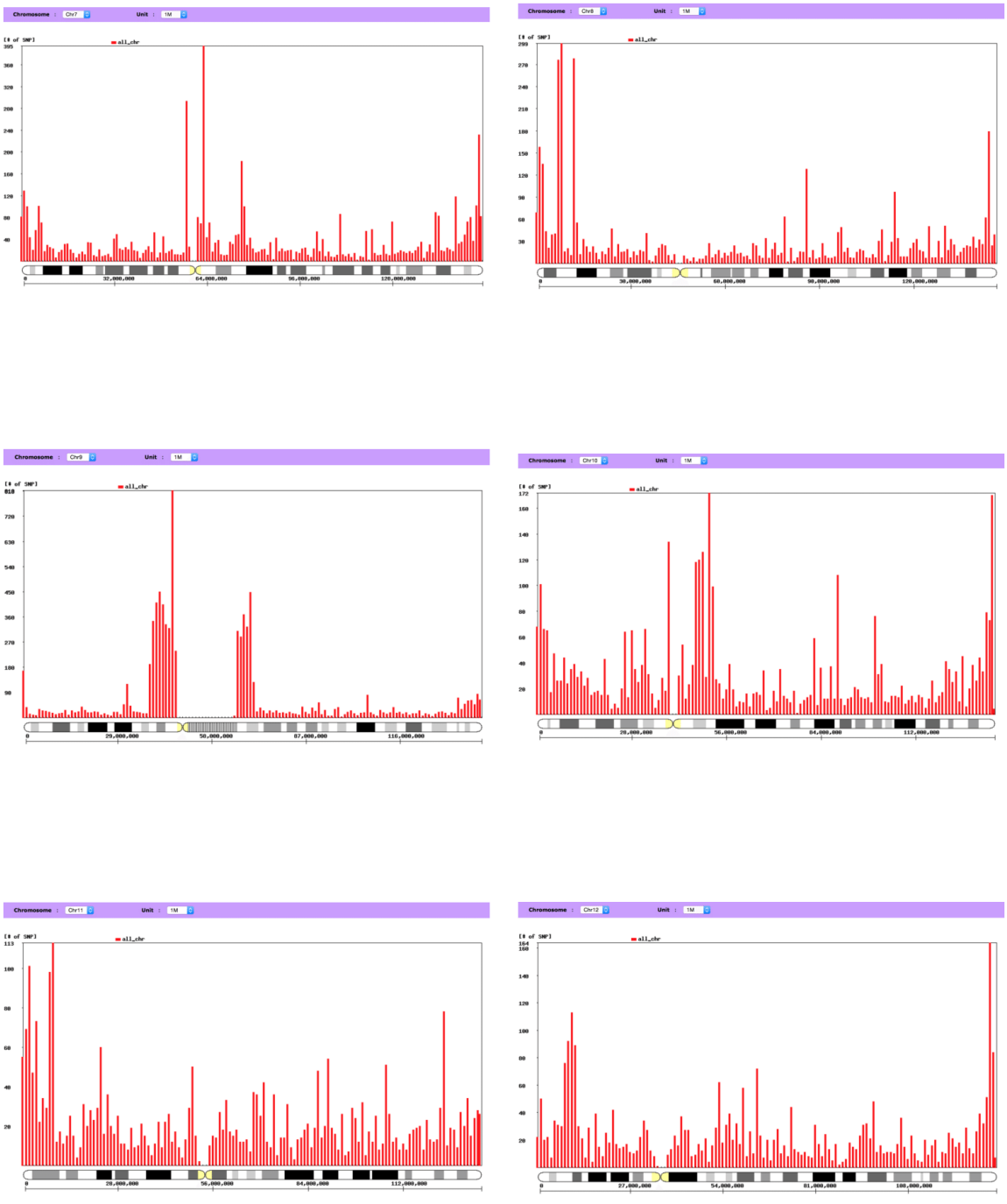

Chromosome : Chr7 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr8 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr9 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr10 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr11 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr12 Unit : 1M
[# of SNP]
all_chr

Figure 25:


Chromosome : Chr13 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr14 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr15 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr16 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr17 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr18 Unit : 1M
[# of SNP]
all_chr

Figure 26:


Chromosome : Chr19 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr20 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr21 Unit : 1M
[# of SNP]
all_chr
Chromosome : Chr22 Unit : 1M
[# of SNP]
all_chr

Figure 27:

Plot 10. Distribution of high heterozygous SNPs with obs. het > 0.55 on physical chromosomal map

Plot 11. Fraction of reference allele reads vs total unfiltered depth (sum over "AD" for all alleles) for 10 random variants on chr 22 with observed heterozyfosity > 0.55

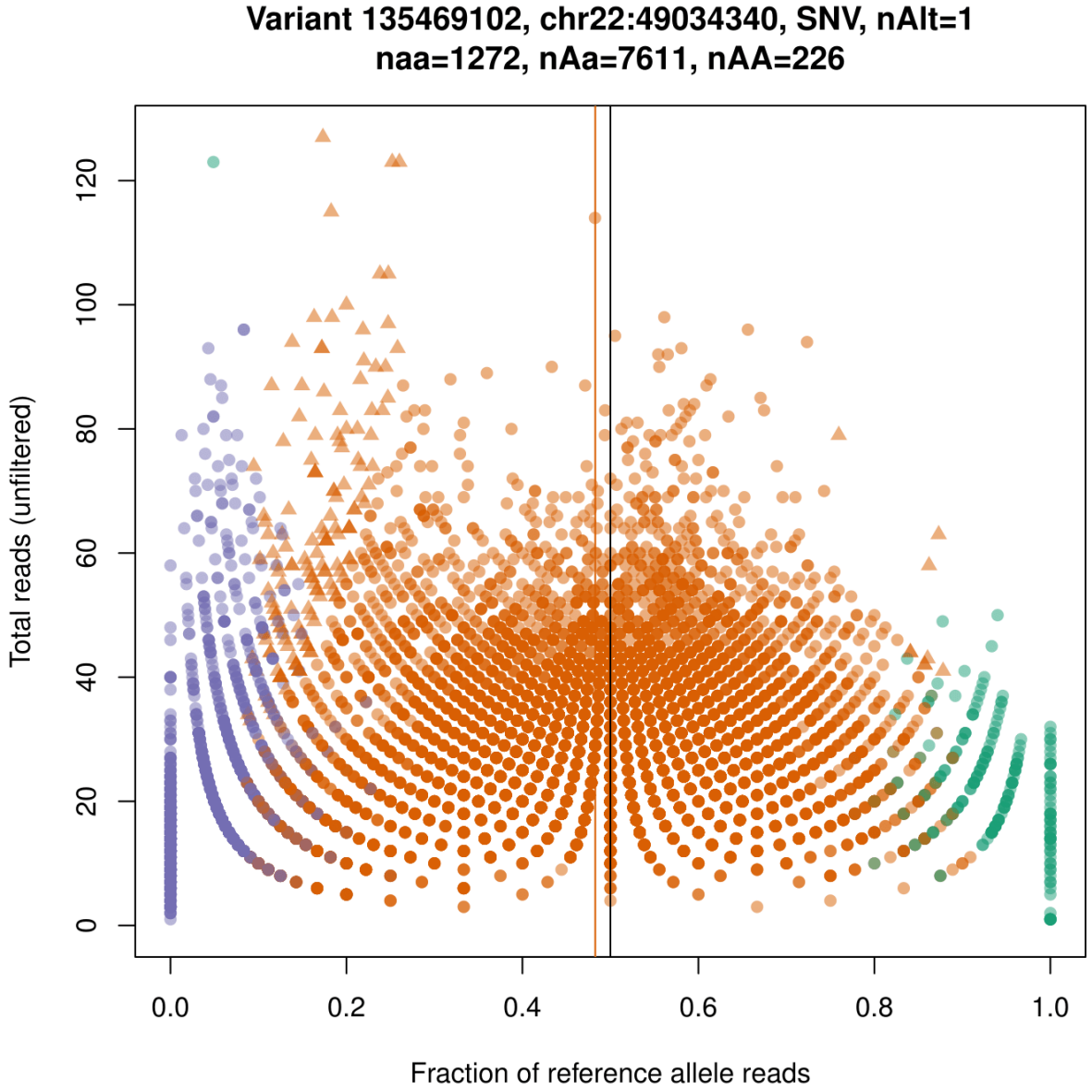


Figure 28:

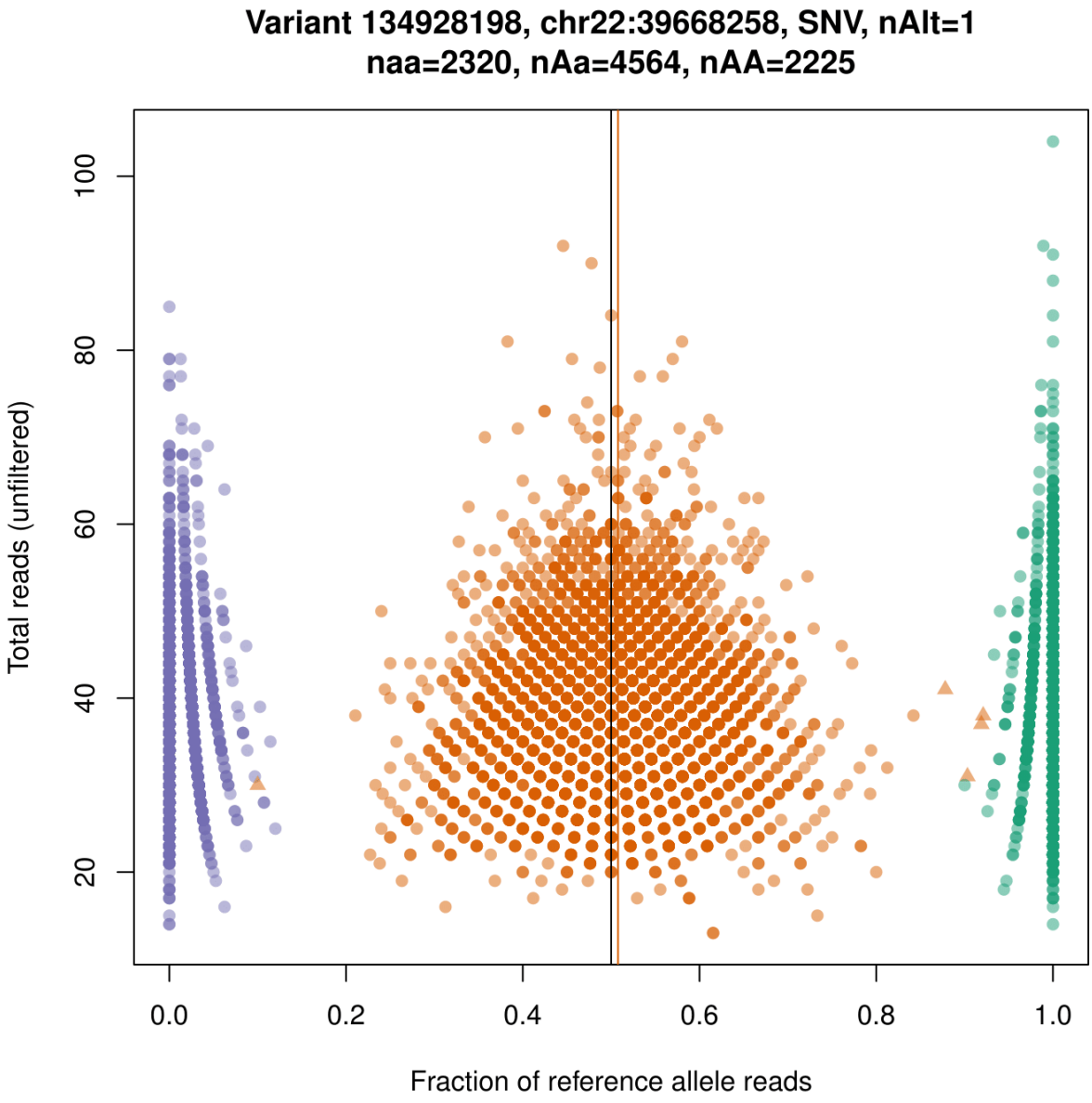

Variant 134928198, chr22:39668258, SNV, nAlt=1
naa=2320, nAa=4564, nAA=2225
Total reads (unfiltered)
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of reference allele reads

Figure 29:

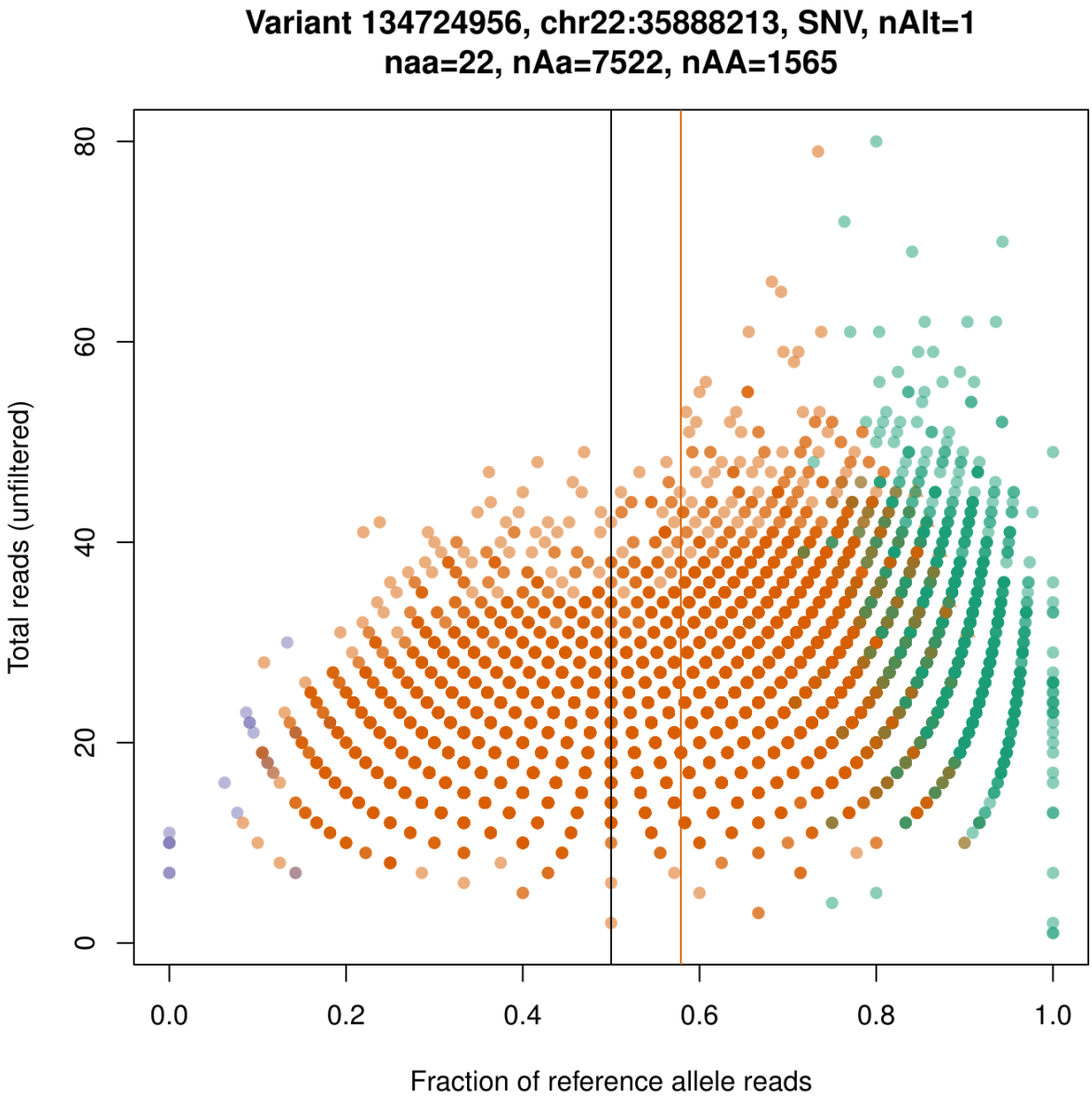

Variant 134724956, chr22:35888213, SNV, nAlt=1
naa=22, nAa=7522, nAA=1565
Total reads (unfiltered)
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of reference allele reads

Figure 30:

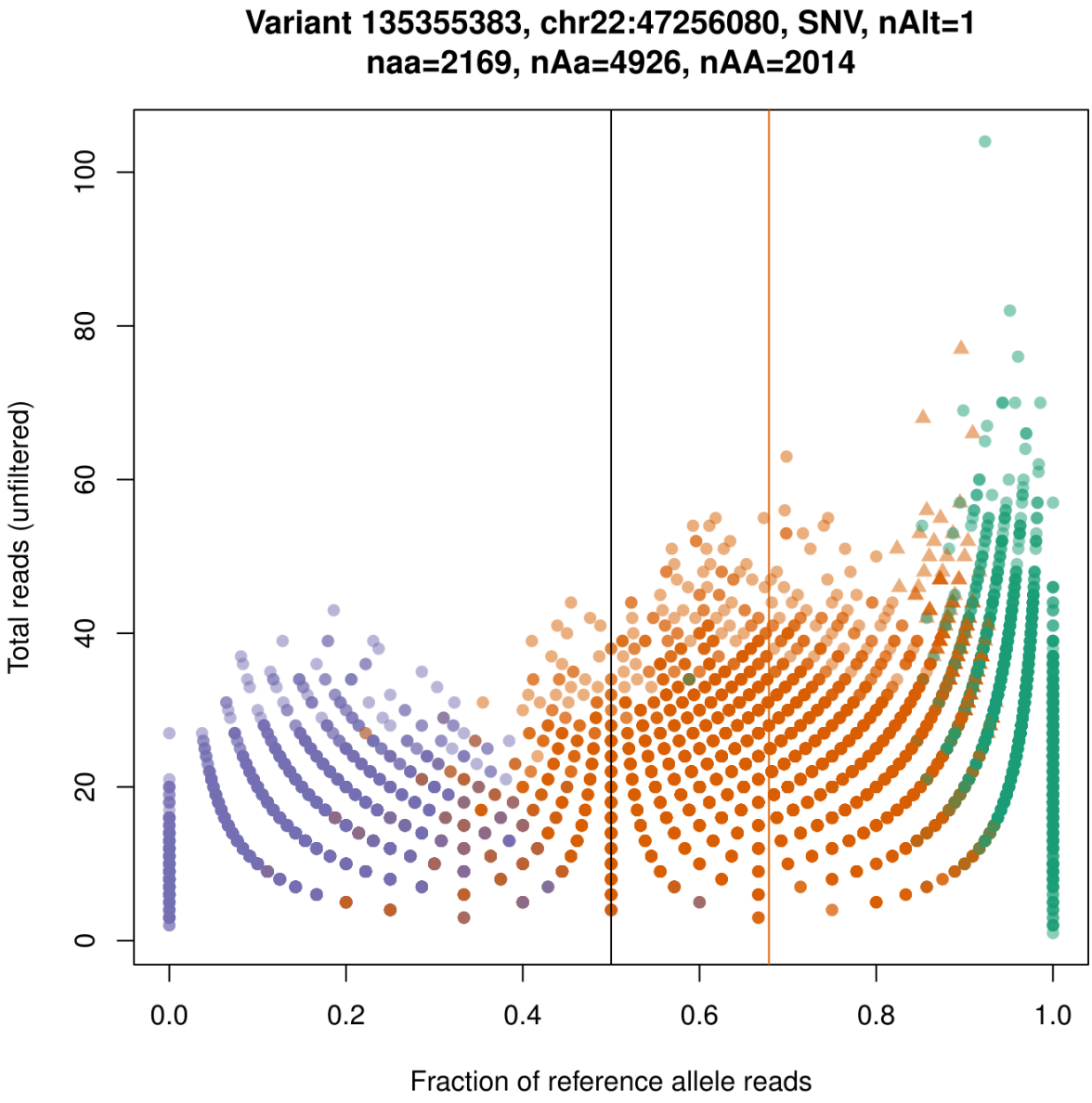

Variant 135355383, chr22:47256080, SNV, nAlt=1
naa=2169, nAa=4926, nAA=2014
Total reads (unfiltered)
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of reference allele reads

Figure 31:

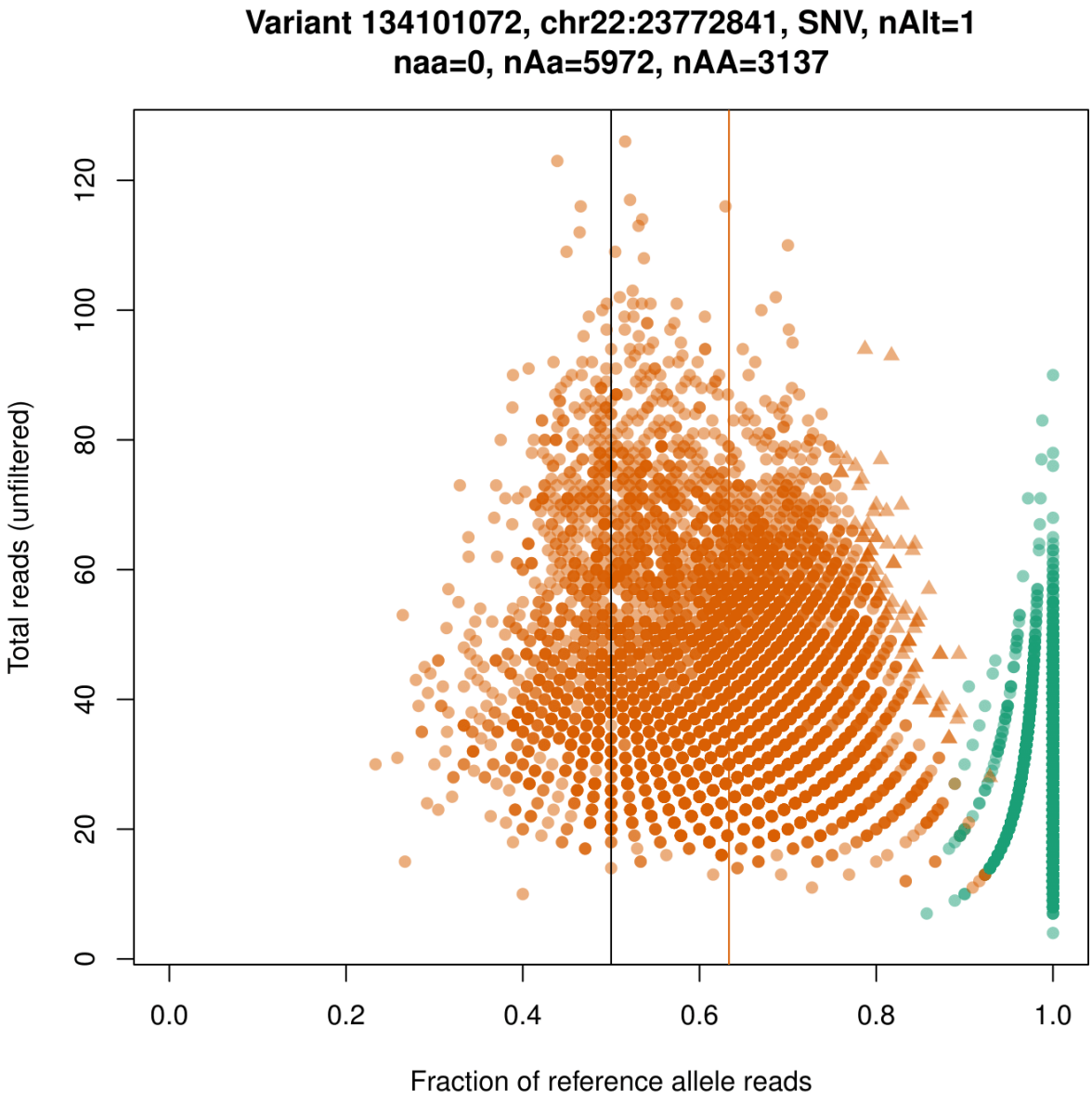

Variant 134101072, chr22:23772841, SNV, nAlt=1
naa=0, nAa=5972, nAA=3137
Total reads (unfiltered)
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of reference allele reads

Figure 32:

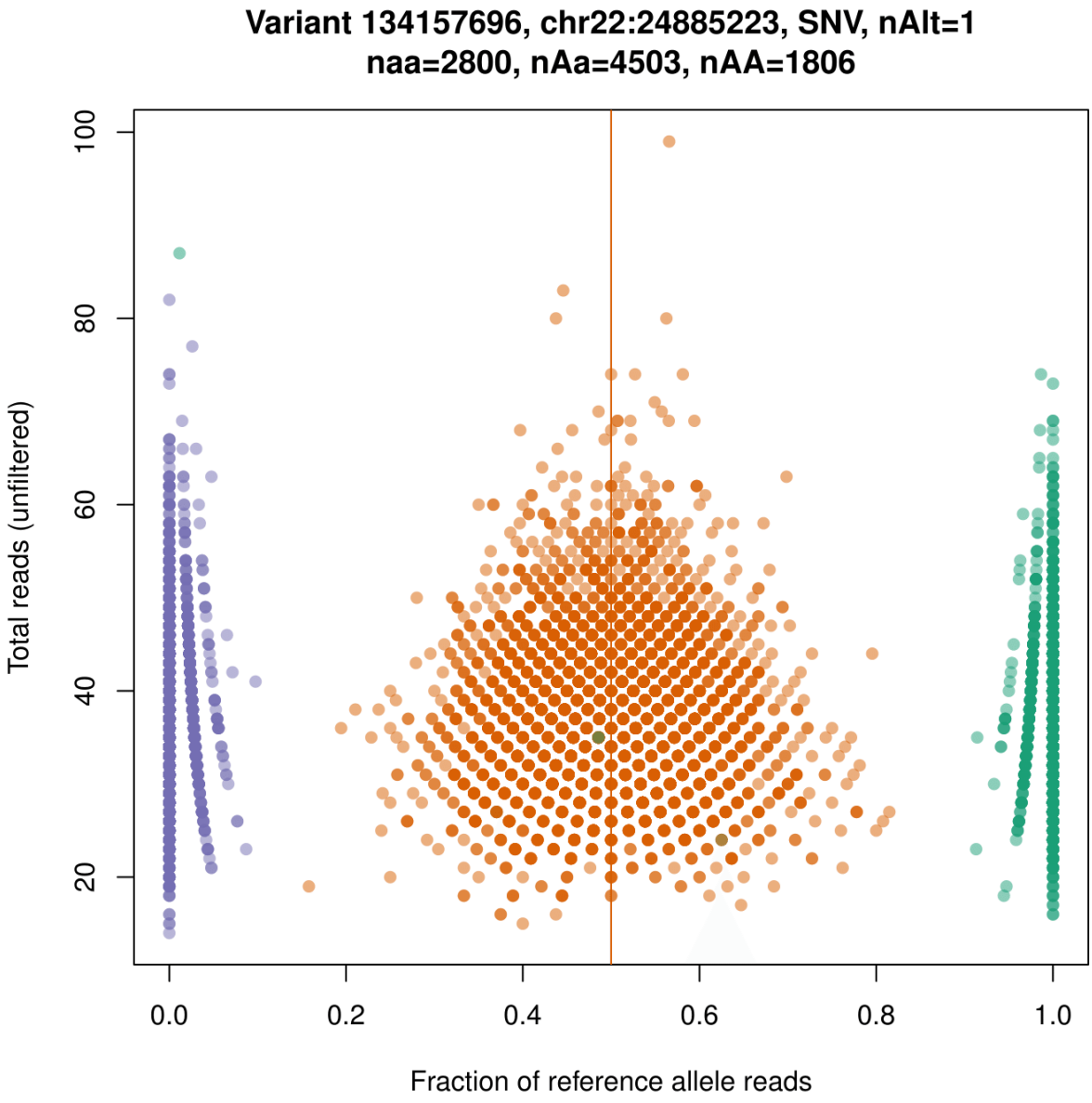

Variant 134157696, chr22:24885223, SNV, nAlt=1
naa=2800, nAa=4503, nAA=1806
Total reads (unfiltered)
100
80
60
40
20
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of reference allele reads

Figure 33:

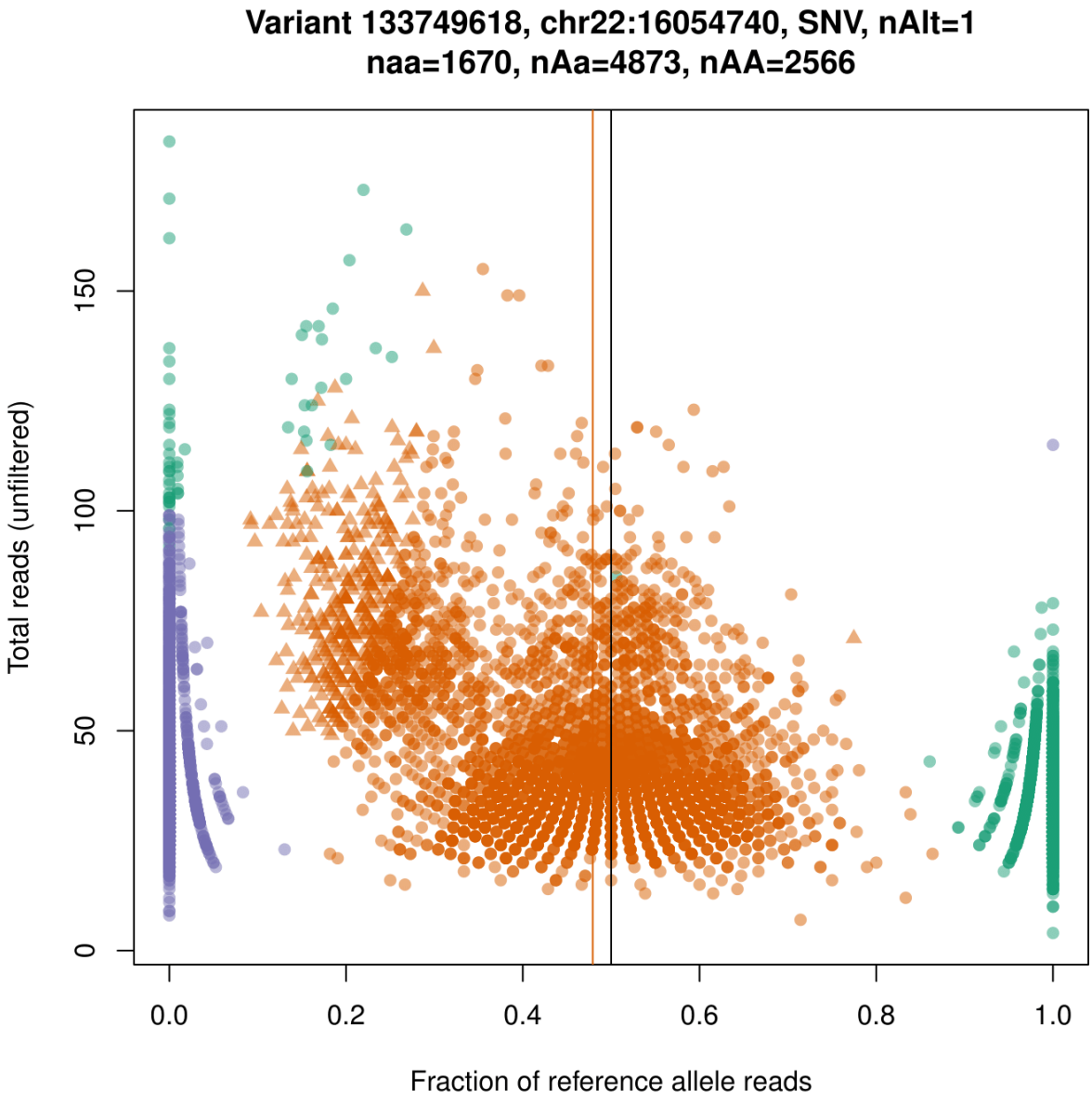

Variant 133749618, chr22:16054740, SNV, nAlt=1
naa=1670, nAa=4873, nAA=2566
Total reads (unfiltered)
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of reference allele reads

Figure 34:

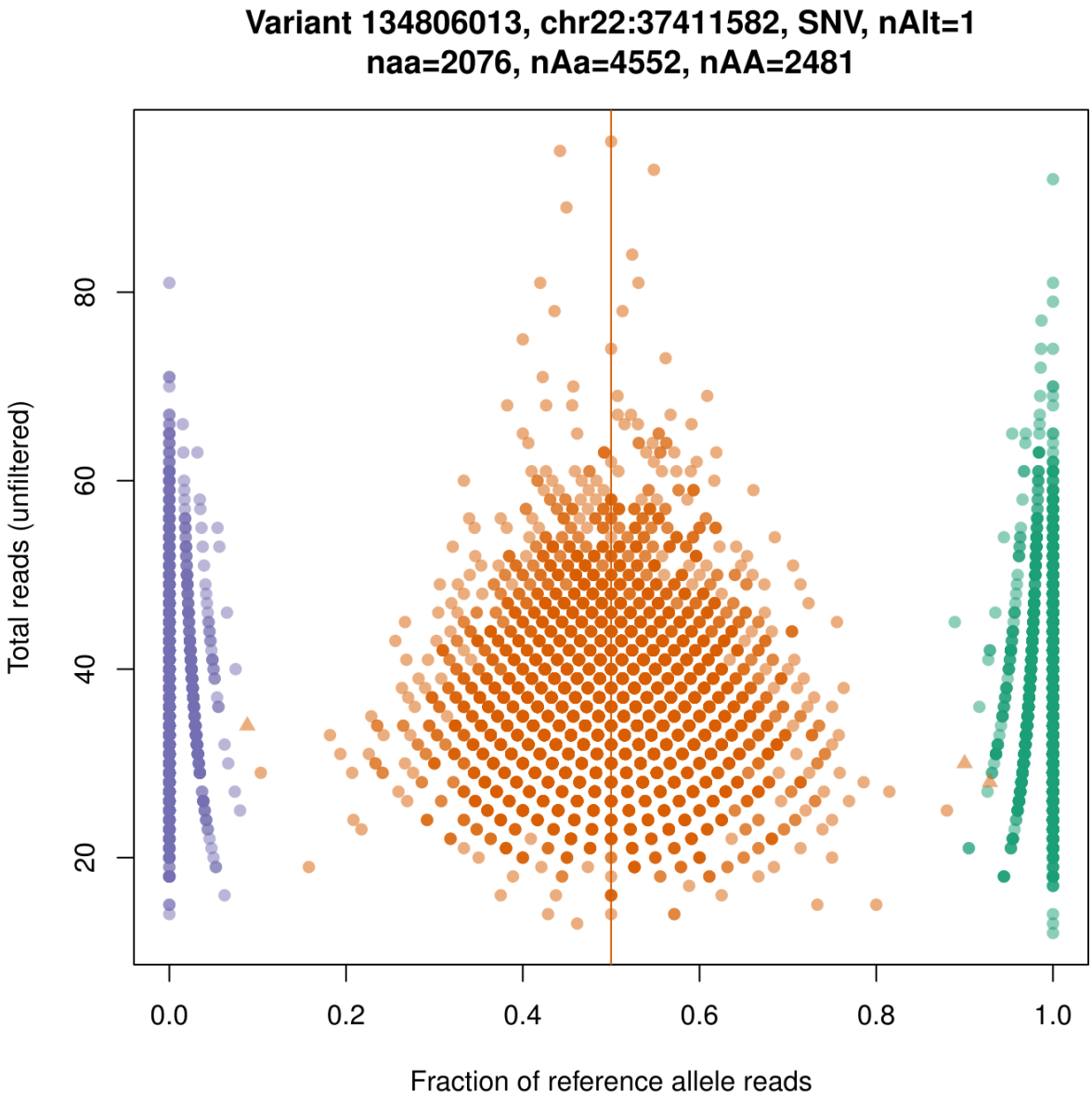

Variant 134806013, chr22:37411582, SNV, nAlt=1
naa=2076, nAa=4552, nAA=2481
Total reads (unfiltered)
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of reference allele reads

Figure 35:

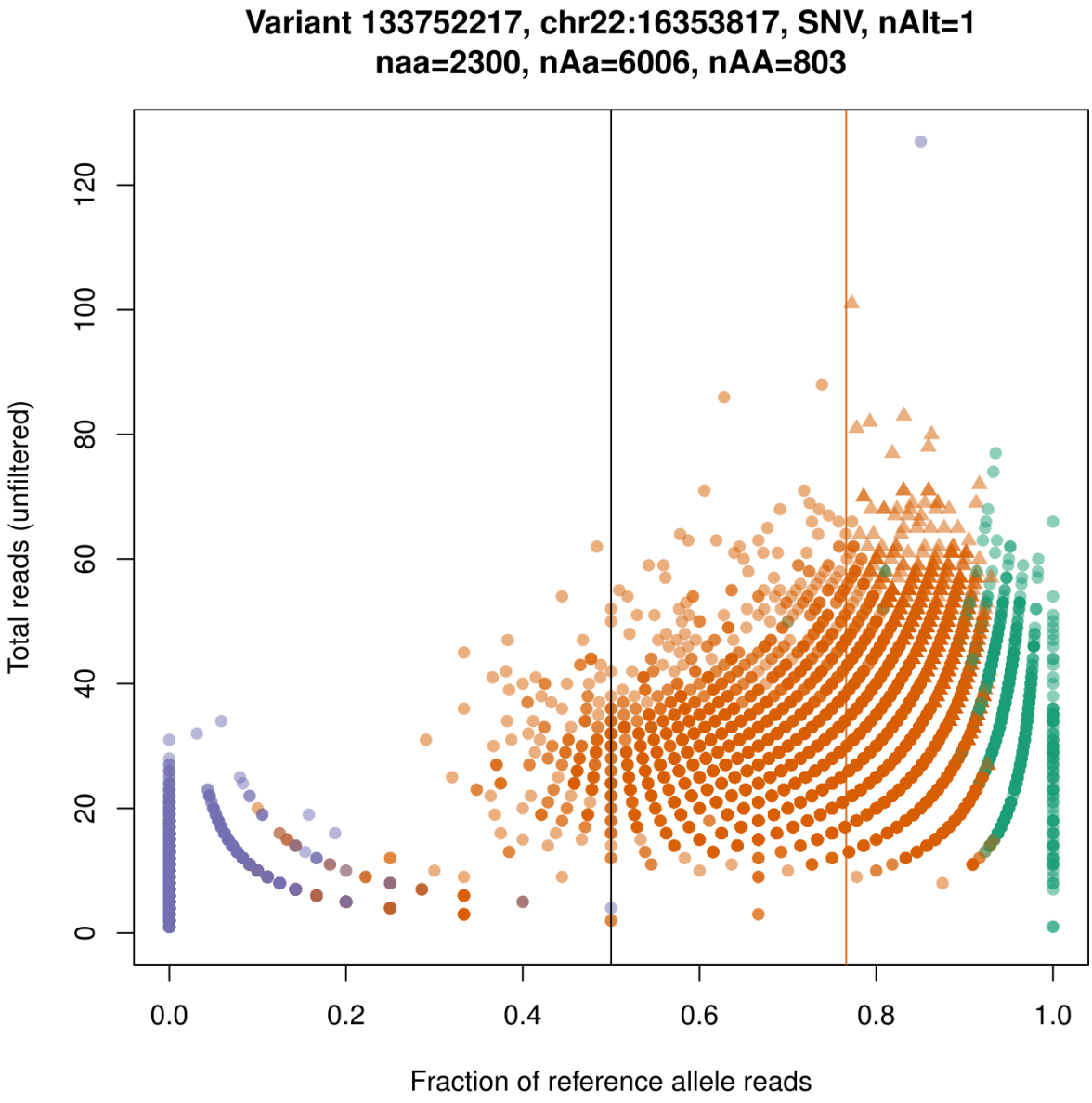

Variant 133752217, chr22:16353817, SNV, nAlt=1
naa=2300, nAa=6006, nAA=803
Total reads (unfiltered)
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of reference allele reads

Figure 36:

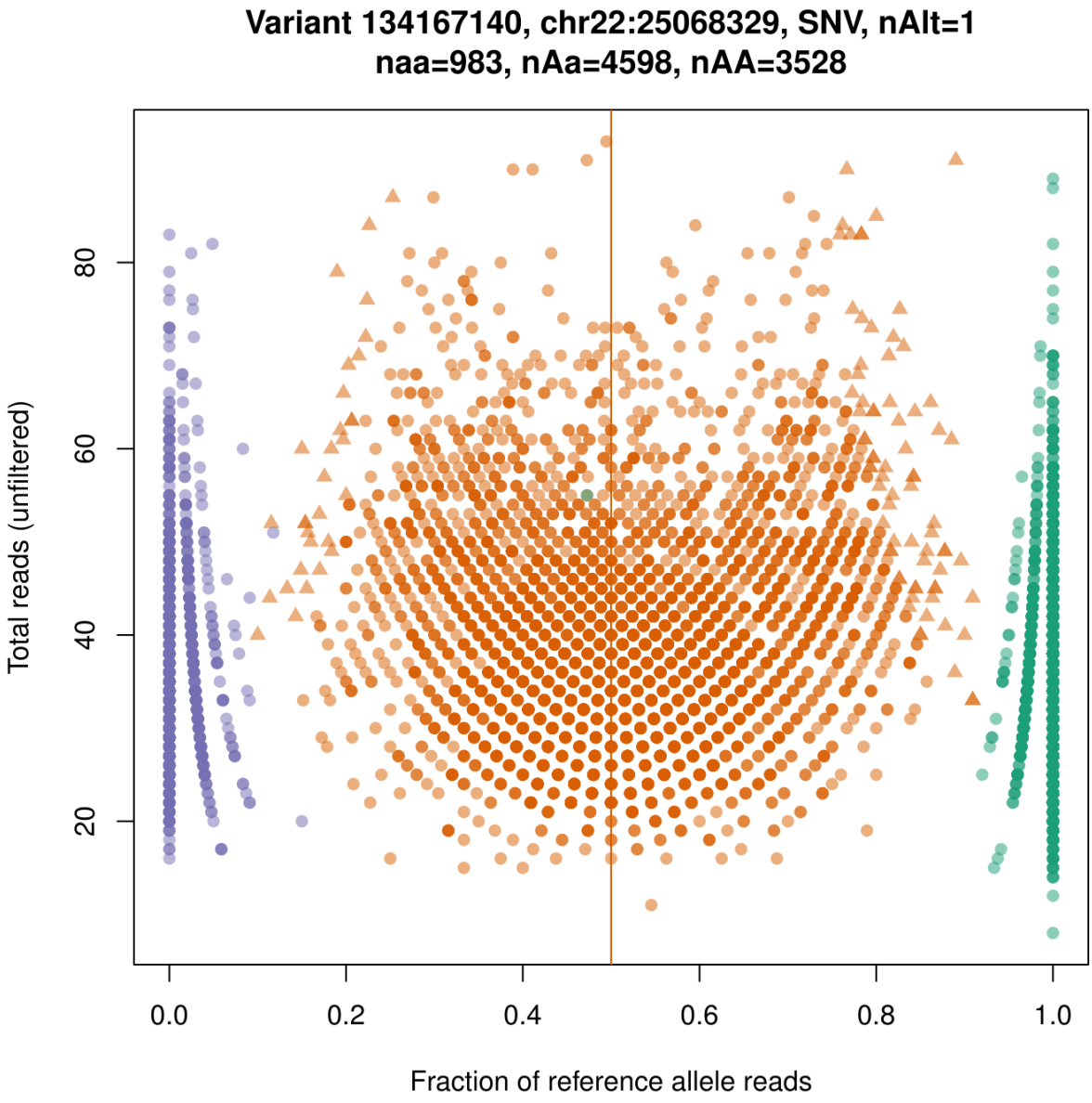

Variant 134167140, chr22:25068329, SNV, nAlt=1
naa=983, nAa=4598, nAA=3528
Total reads (unfiltered)
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of reference allele reads

Figure 37:

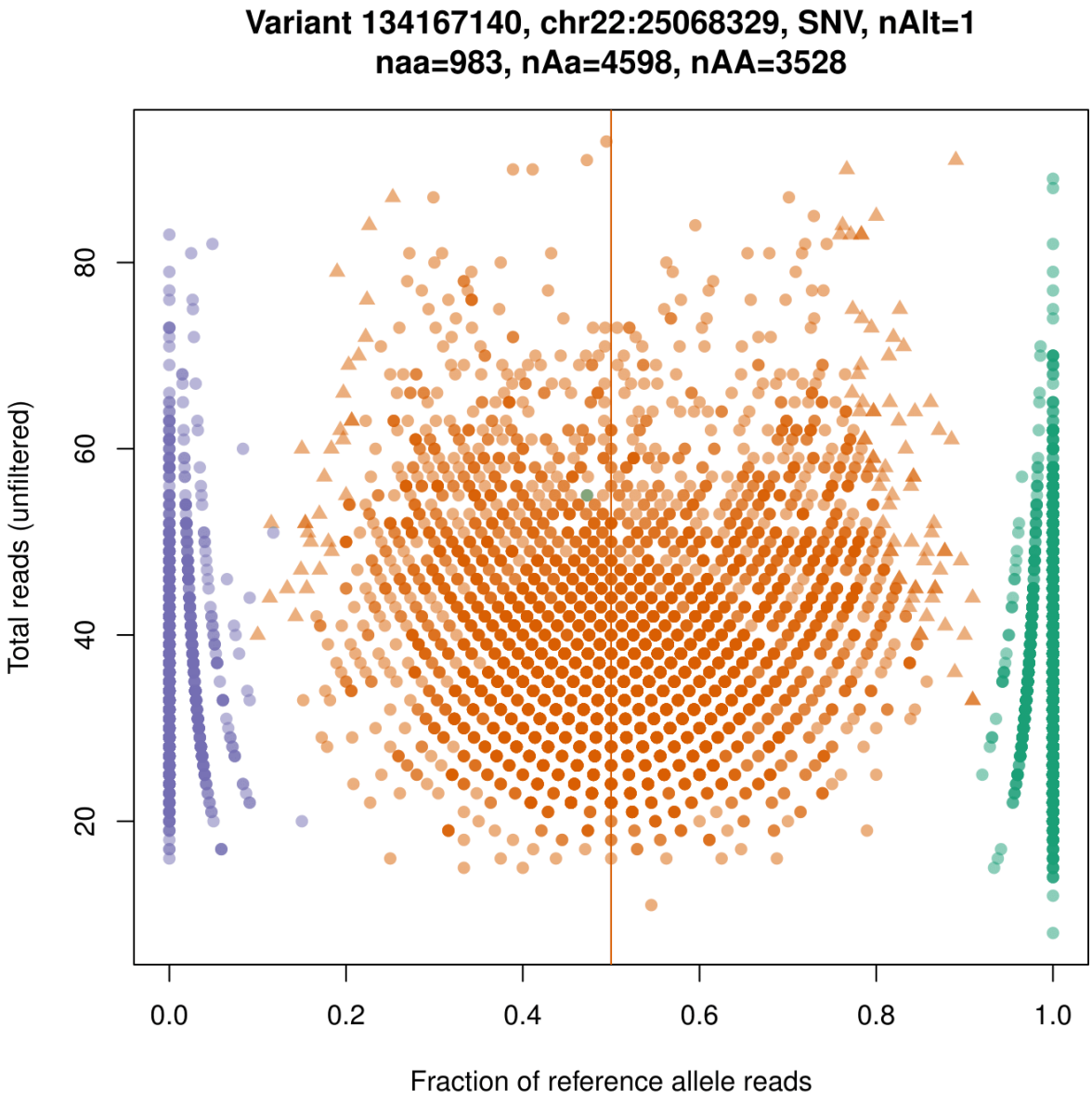

Variant 134167140, chr22:25068329, SNV, nAlt=1
naa=983, nAa=4598, nAA=3528
Total reads (unfiltered)
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of reference allele reads

Figure 38: