The American Journal of Human Genetics, Volume *103*

Supplemental Data

Understanding the Hidden Complexity of Latin American Population Isolates

Jazlyn A. Mooney, Christian D. Huber, Susan Service, Jae Hoon Sul, Clare D. Marsden, Zhongyang Zhang, Chiara Sabatti, Andrés Ruiz-Linares, Gabriel Bedoya, Costa Rica/Colombia Consortium for Genetic Investigation of Bipolar Endophenotypes, Nelson Freimer, and Kirk E. Lohmueller

Supplementary Figures and Tables

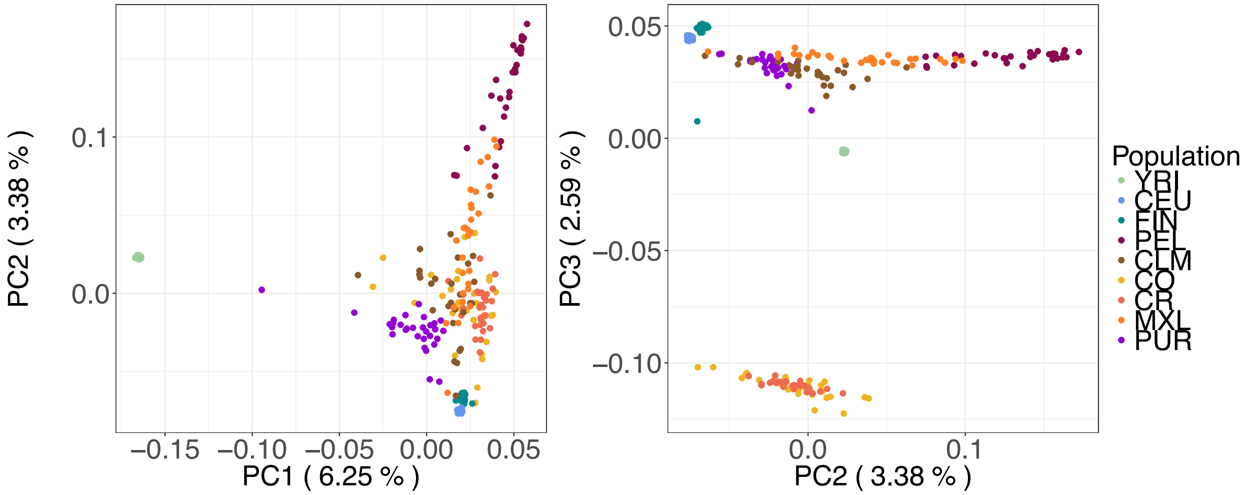

Figure S1. PCA of 30 unrelated individuals sampled from each population. Points correspond to individuals and colors correspond to populations. Africans and non-Africans separate from each other within the first principal component. The second principal component further separated European populations from American populations. The third principal component separated the admixed Latin American isolates from remaining admixed American populations. YRI: Yoruba 1000 Genomes; CEU: Ceph-European 1000 Genomes; FIN: Finnish 1000 Genomes; PEL: Peruvian 1000 Genomes; CLM: Colombian 1000 Genomes; CO: Colombia; CR: Costa Rica; MXL: Mexican from Los Angeles 1000 Genomes; and PUR: Puerto Rican 1000 Genomes.

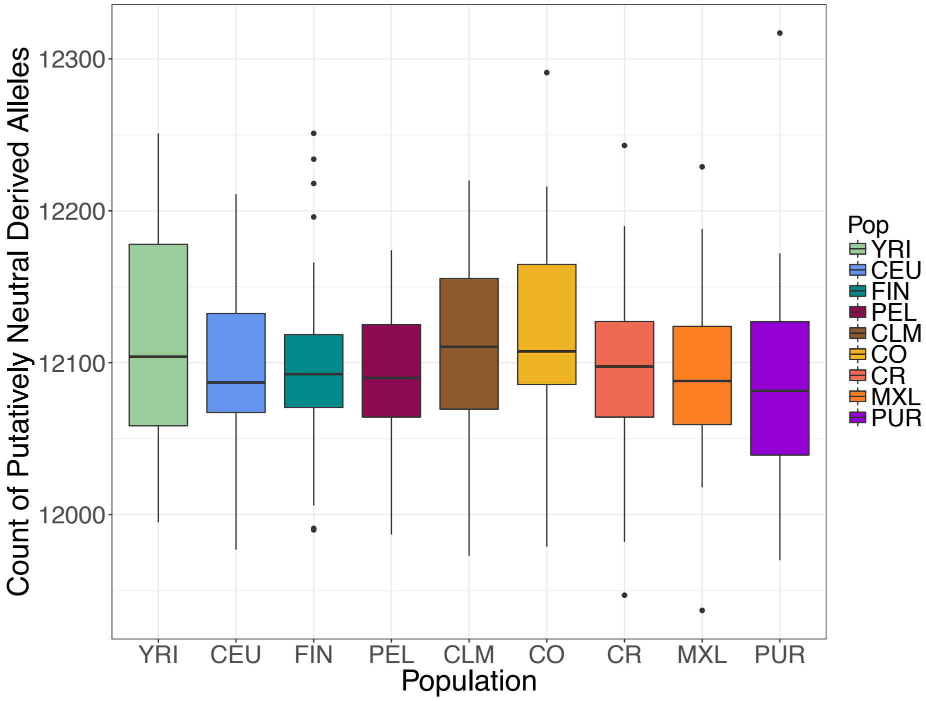

Figure S2. Comparison of the number of putatively neutral derived alleles per individual across populations. In the figure above we display the count of derived putatively neutral alleles that were identified using GERP[40]. A GERP score less than two was considered as putatively neutral. The median count of neutral alleles across the unrelated individuals in each population is relatively stable. The two populations with the highest median scores are the CLM and CO. The similar medians between these two populations is reassuring since the CLM were sequenced at a lower coverage than the CO. Further, this figure indicates that our merging pipeline produced comparable data. Population abbreviations are as in Figure S1.

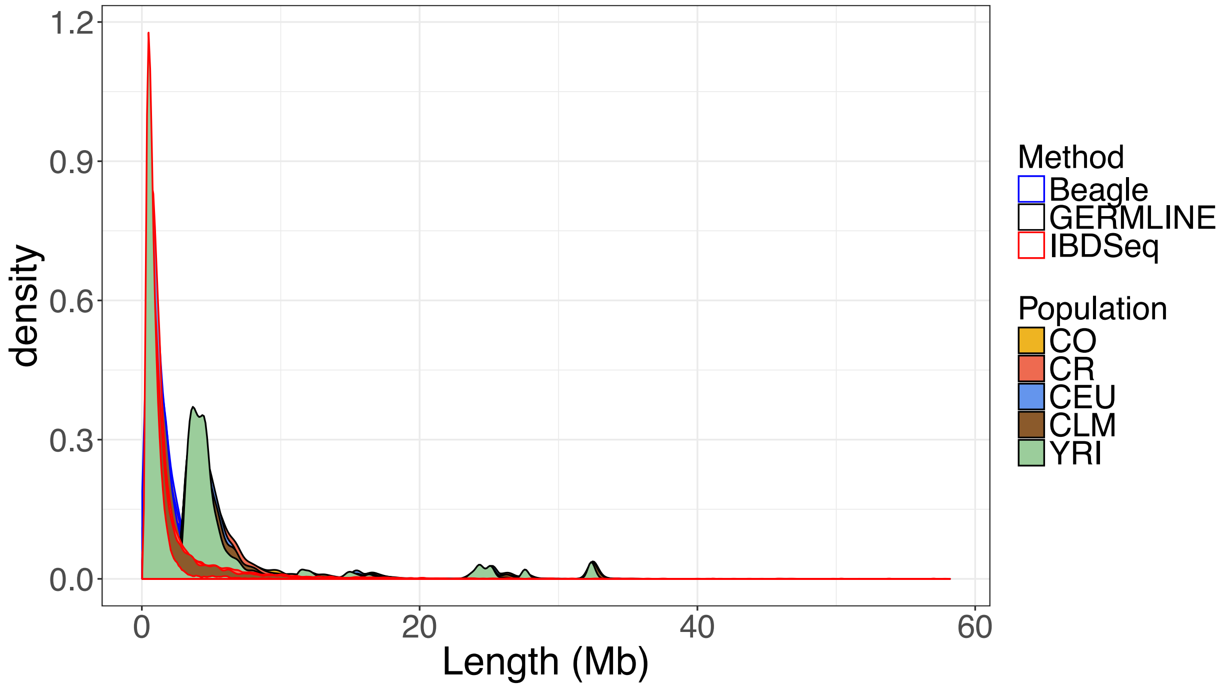

Figure S3. Comparison of methods to define IBD segments. We used three different methods to detect IBD segments our data Beagle[28] (blue) and GERMLINE[29] (black), and IBDSeq[27](red). The x-axis represents the length in megabases (Mb) of IBD segments and the y-axis depicts the proportion of IBD segments of a given length. For preliminary phasing and calling of IBD segments, we use only a subset of our sampled populations (CO, CR, CLM, CEU, and YRI). Beagle produces the shortest segments while GERMLINE segments tend to be much longer, and IBDSeq falls between the two. Population abbreviations are as in Figure S1.

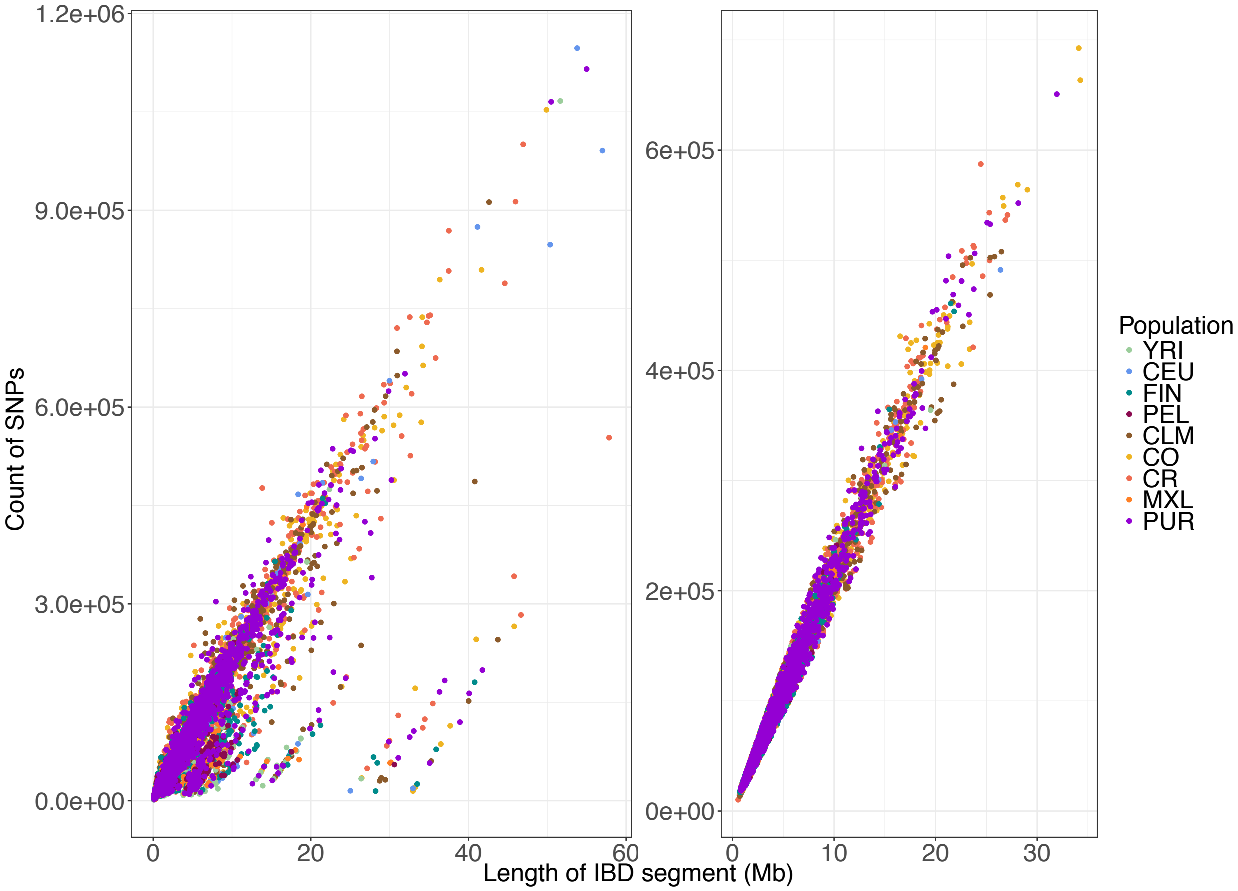

Figure S4. IBD segments before and after filtering. The x-axis represents the length of each IBD segment in megabases (Mb) and the y-axis is the total number of SNPs in each IBD segment. The graph on the left side depicts IBD segments before filtering. There are a considerable number of IBD segments with sparse SNP coverage in each population. IBD segments were removed if the proportion of the IBD segment covered by SNPs was not within one standard deviation of the mean proportion covered across all IBD segments (see **Methods**). On the right-hand graph, we depict IBD segments after filtering. One can see that IBD segments with sparse SNP coverage were removed, and these IBD segments were used in our IBD enrichment analyses and in IBDNe to estimate effective population size. Population abbreviations are as in Figure S1.

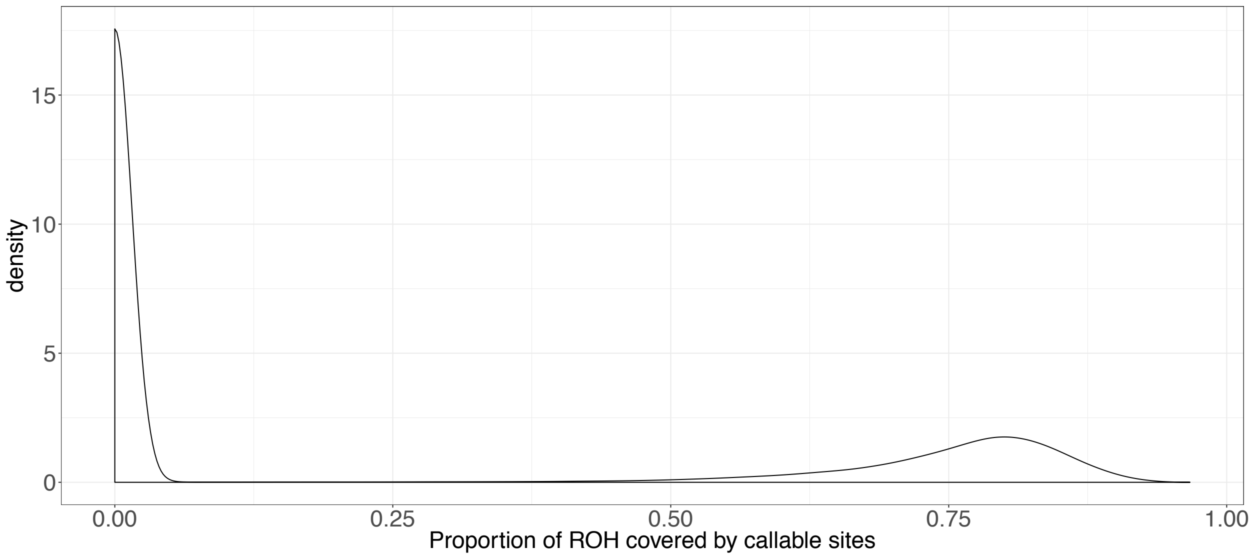

Figure S5. Proportion of a long run of homozygosity (ROH) covered by passing sites. This figure depicts the proportion of each ROH that is covered by sites that pass our depth filter and pass the 1000 Genomes Project strict mask. One can see that there is an appreciable number of ROH that are covered very sparsely by SNPs. If SNP coverage of an ROH was less than 60% the ROH was removed from our analyses.

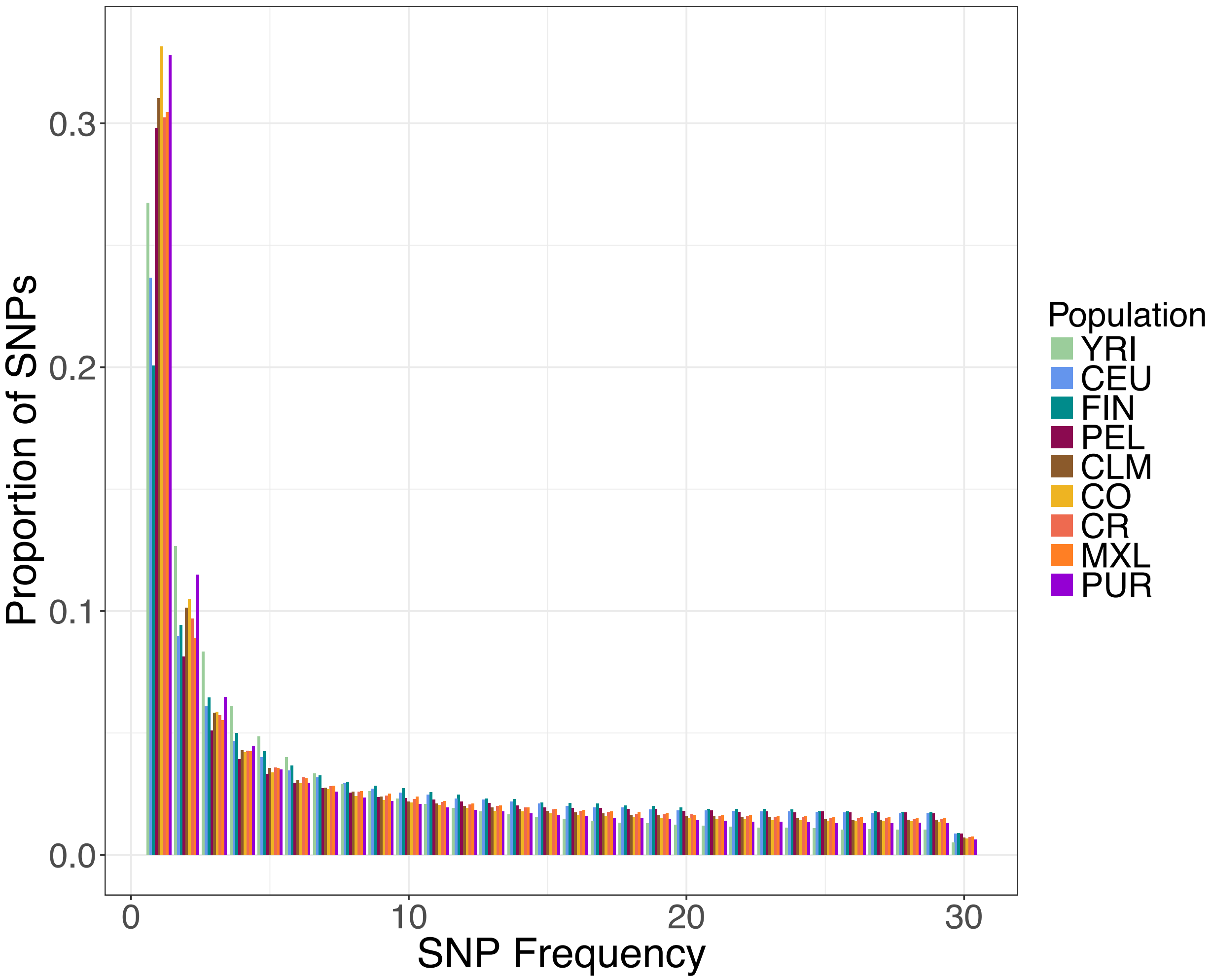

Figure S6. Whole genome folded SFS. Extended version of the SFS from Figure 1C where we truncated at a SNP frequency of 15. Population abbreviations are as in Figure S1.

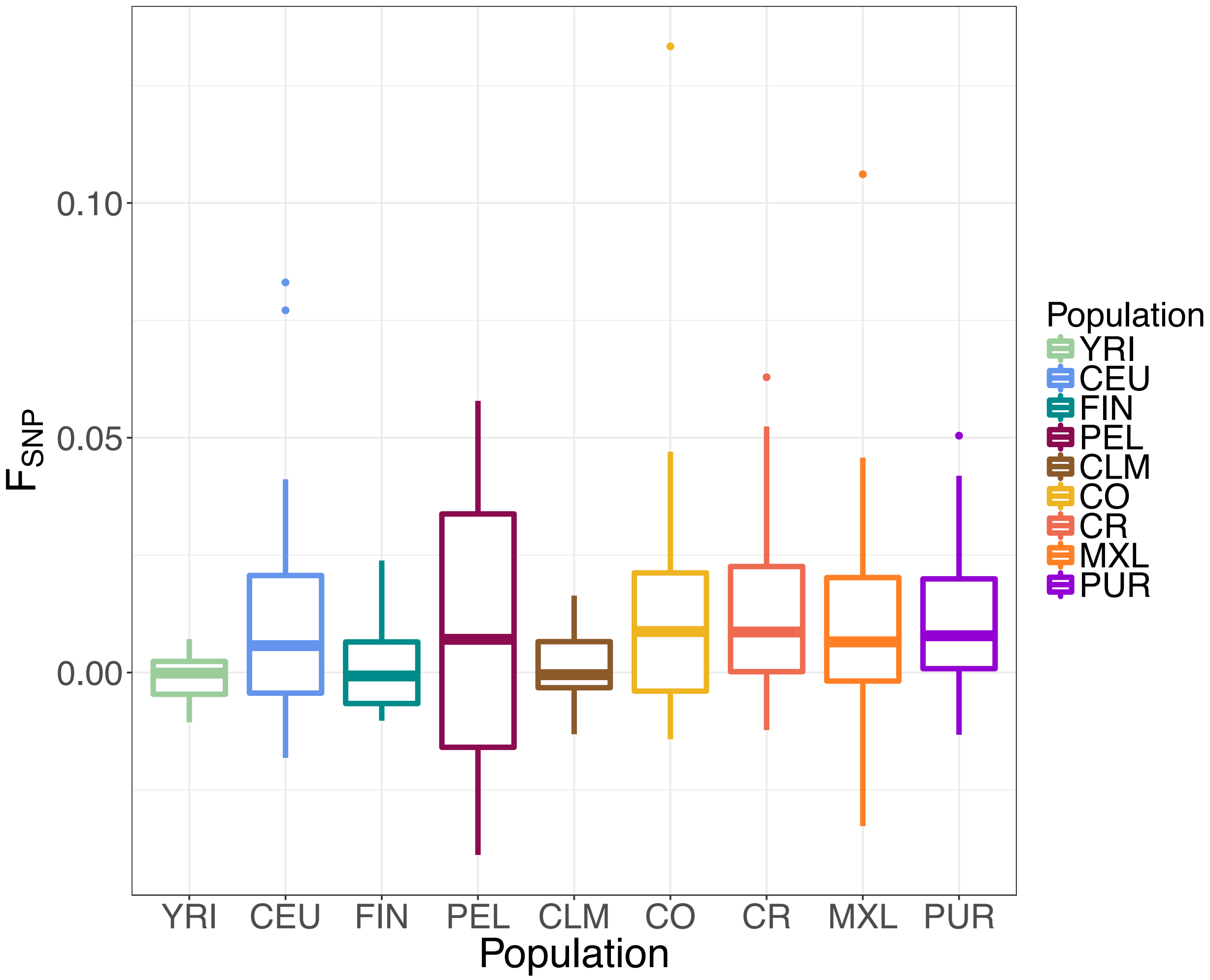

Figure S7. Boxplot depicting F_{SNP} per population. Each boxplot represents the distribution of F_{SNP} in the 30 unrelated individuals sampled in each population. Population abbreviations are as in Figure S1.

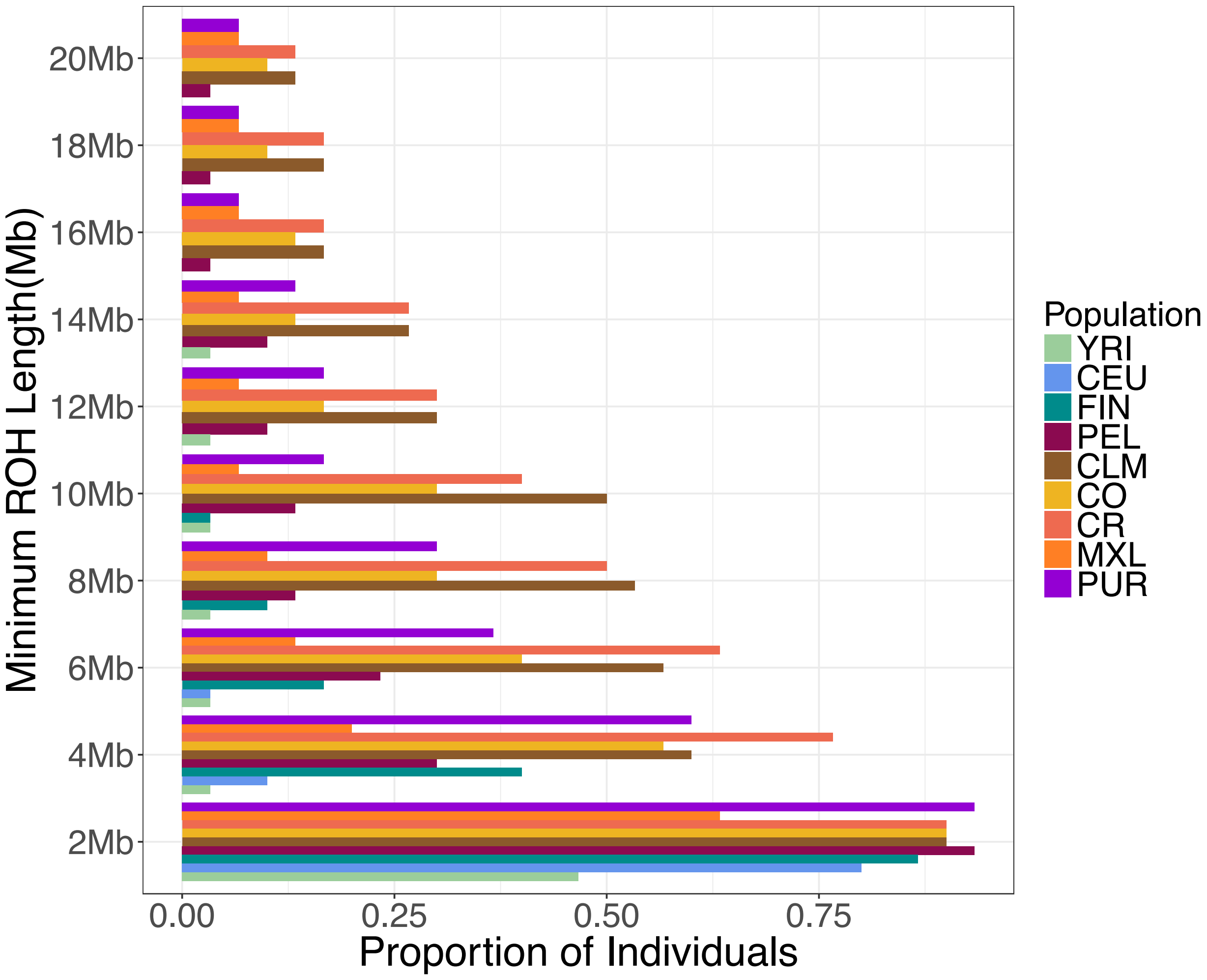

Figure S8. Proportion of individuals carrying a run of homozygosity of a given length. The x-axis represents the proportion of individuals from the 30 unrelated individuals that carry a ROH of a minimum length, given on the y-axis. Population abbreviations are as in Figure S1.

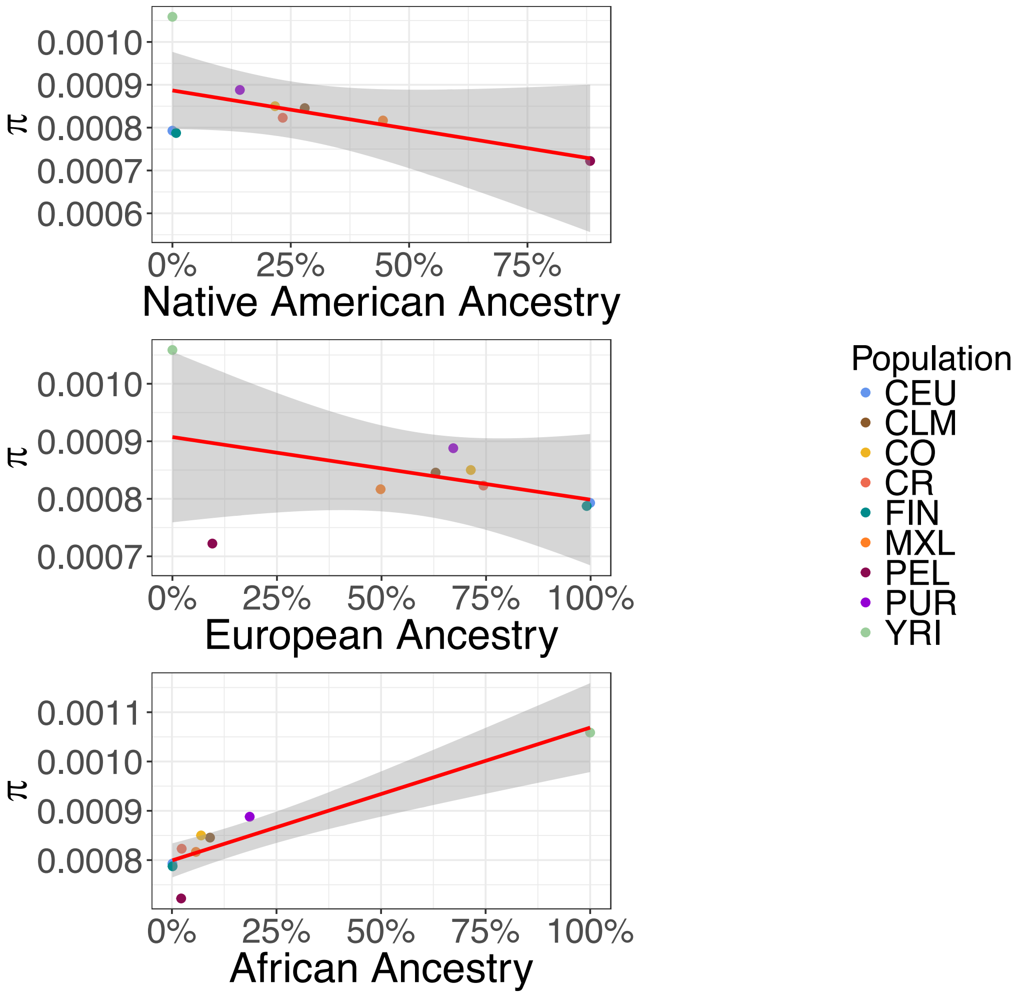

Figure S9. Correlation between ancestry and pi (*n* = 30). Each data point represents average intergenic diversity across the autosomes for the thirty sampled individuals per population. Ancestry proportions for each population are listed in Table 1. Population abbreviations are as in Figure S1.

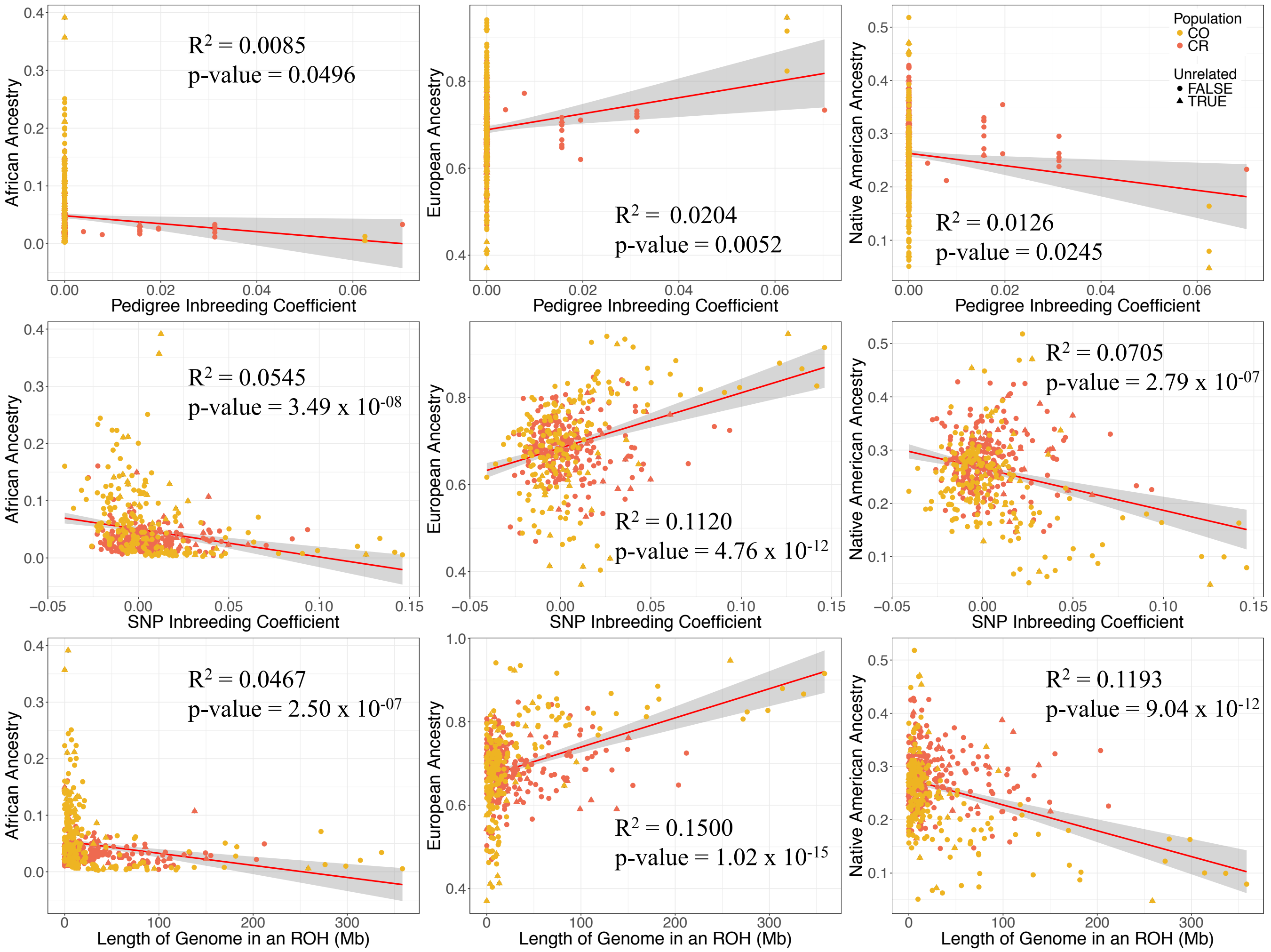

Figure S10. Correlation between ancestry, F_{PED}, F_{SNP}, and the length of the autozygous genome (*n* = 449). Each data point represents an individual. Triangles represent the individuals that were sampled in the unrelated data set (*n* = 30). The top row shows the correlation between global ancestry proportion and each Latin American individual's pedigree inbreeding coefficient, the middle row corresponds to ancestry proportion and SNP inbreeding coefficient, and the bottom row depicts the correlation between ancestry proportion and length of the genome within an ROH. Population abbreviations are as in Figure S1.

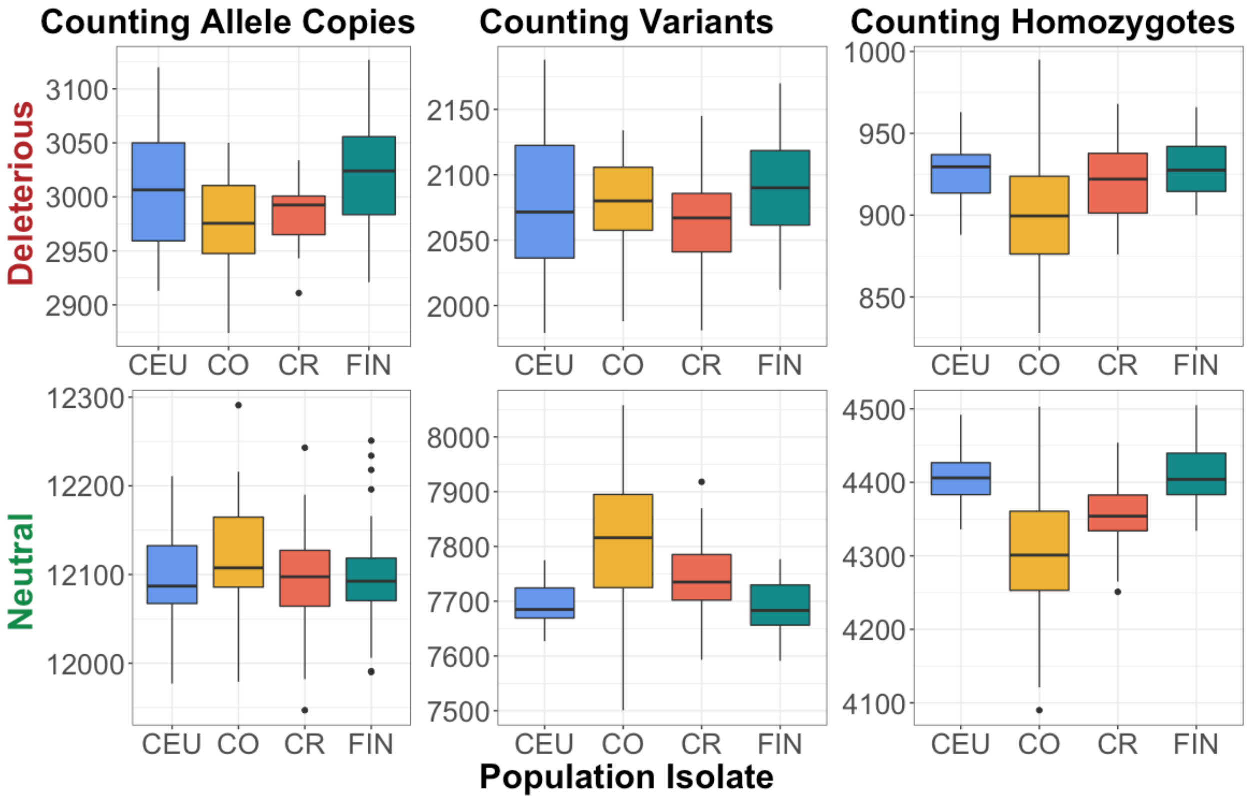

Figure S11. Patterns of deleterious variation in in the Colombian, Costa Rican, and Finnish samples (*n* = 30). Absolute counts of mutations per individual for the Colombian (CO), Costa Rican (CR) and Finnish (FIN) population at putatively deleterious nonsynonymous and putatively neutral synonymous sites. We have also included the absolute counts from CEU as a reference. Each column represents a particular counting method (see **Methods**): number of derived alleles per individual, number of variants per individual, and the number of homozygous derived genotypes per individual. The top row corresponds to putatively deleterious variation and the bottom row corresponds to putatively neutral variation. Population abbreviations are as in Figure S1.

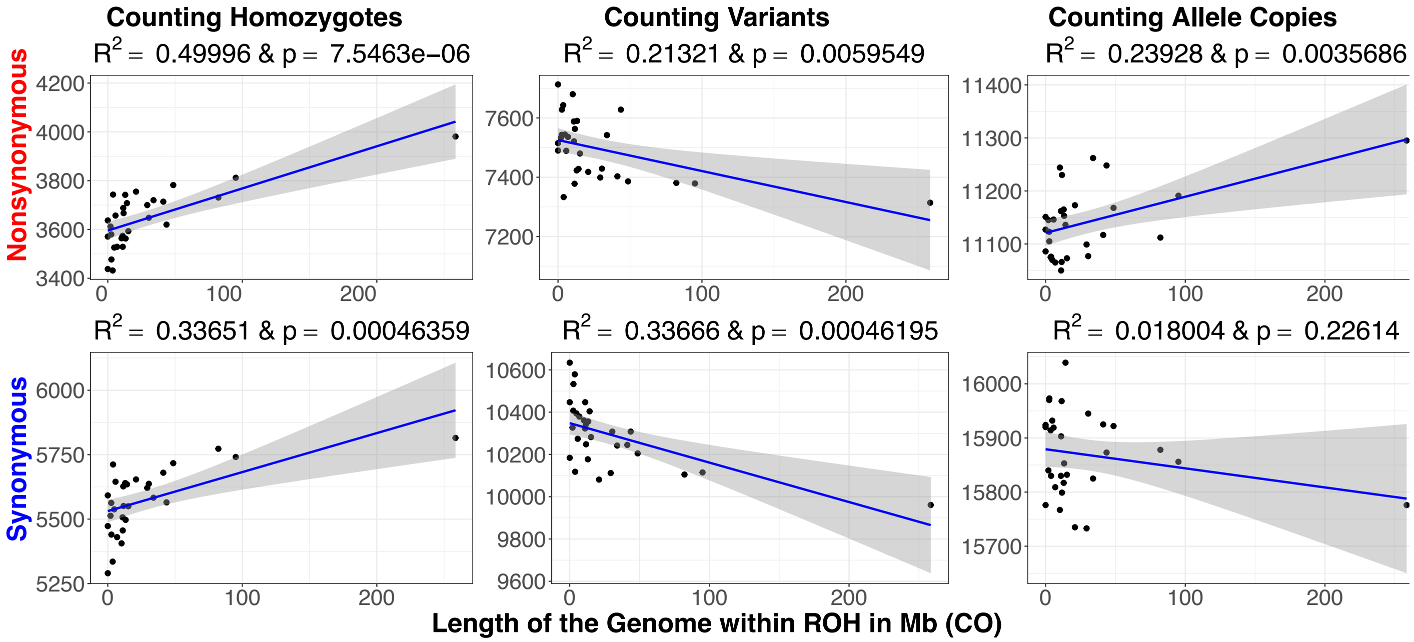

Figure S12. Correlation between length of the genome in an ROH and each counting method using nonsynonymous and synonymous sites in unrelated Colombians (*n* = 30). Each column represents a particular counting method (see **Methods**). The top row shows the correlation between nonsynonymous mutations and the length of the genome within an ROH. The bottom row shows the correlation between synonymous mutations and the length of the genome within an ROH. This figure is a zoomed in version of the correlations for CO depicted in Figure 6. Population abbreviations are as in Figure S1.

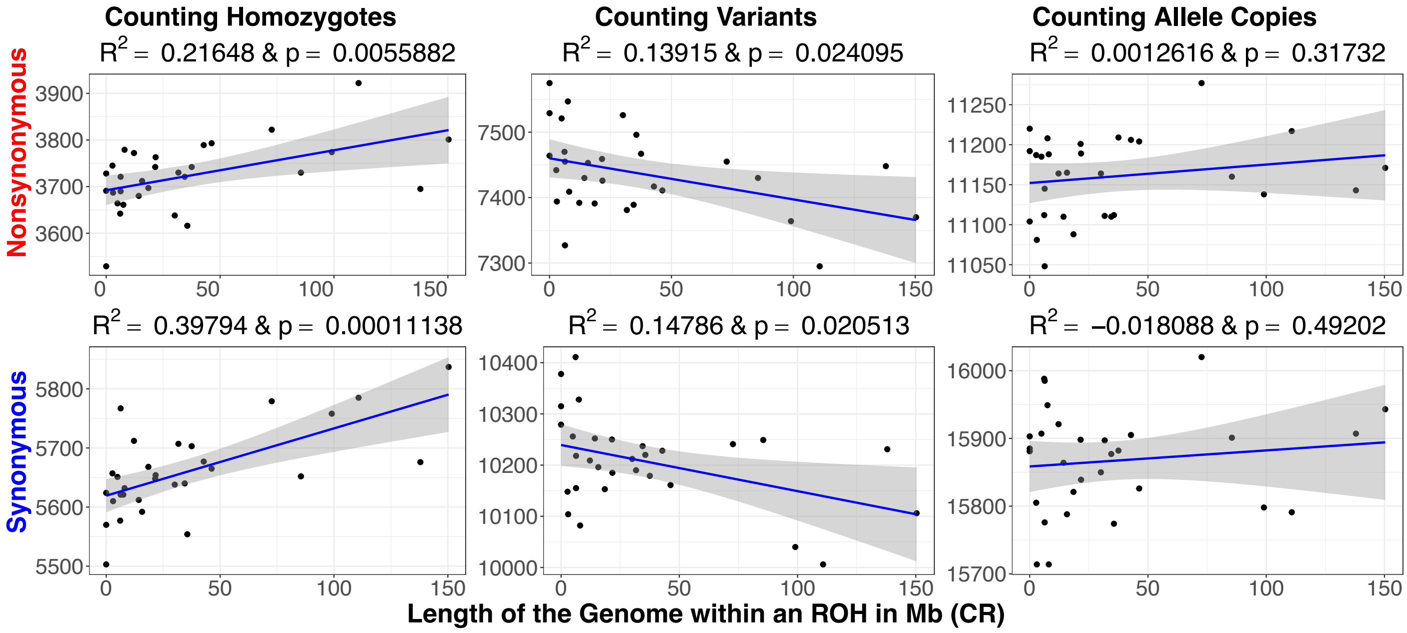

Figure S13. Correlation between length of the genome in an ROH and each counting method using nonsynonymous and synonymous sites in Costa Ricans (*n* = 30). Each column represents a particular counting method (see **Methods**). The top row shows the correlation between nonsynonymous mutations and the length of the genome within an ROH. The bottom row shows the correlation between synonymous mutations and the length of the genome within an ROH. This figure is a zoomed in version of the correlations for CR depicted in Figure 6. Population abbreviations are as in Figure S1.

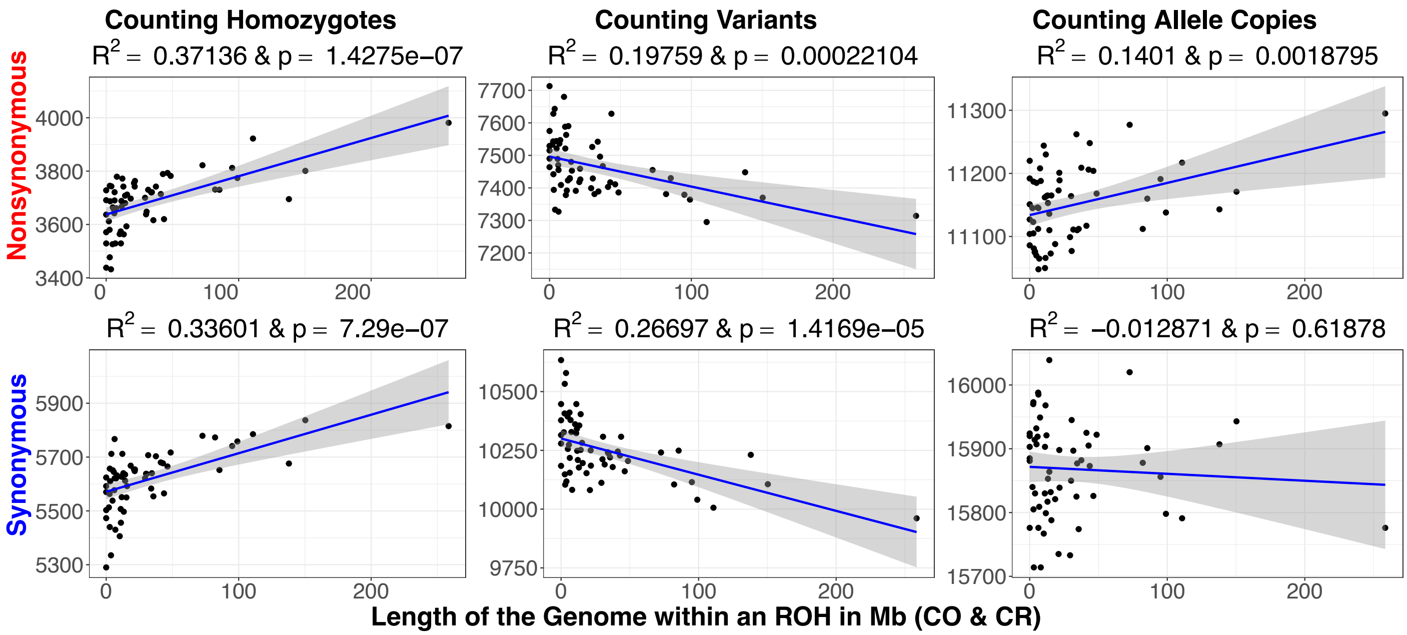

Figure S14. Correlation between length of the genome in an ROH and each counting method using nonsynonymous and synonymous mutations in a combined super-population of Colombians and Costa Ricans (*n* = 60). Each column represents a particular counting method (see **Methods**). The top row shows the correlation between nonsynonymous mutations and the length of the genome within an ROH. The bottom row shows the correlation between synonymous mutations and the length of the genome within an ROH. Population abbreviations are as in Figure S1.

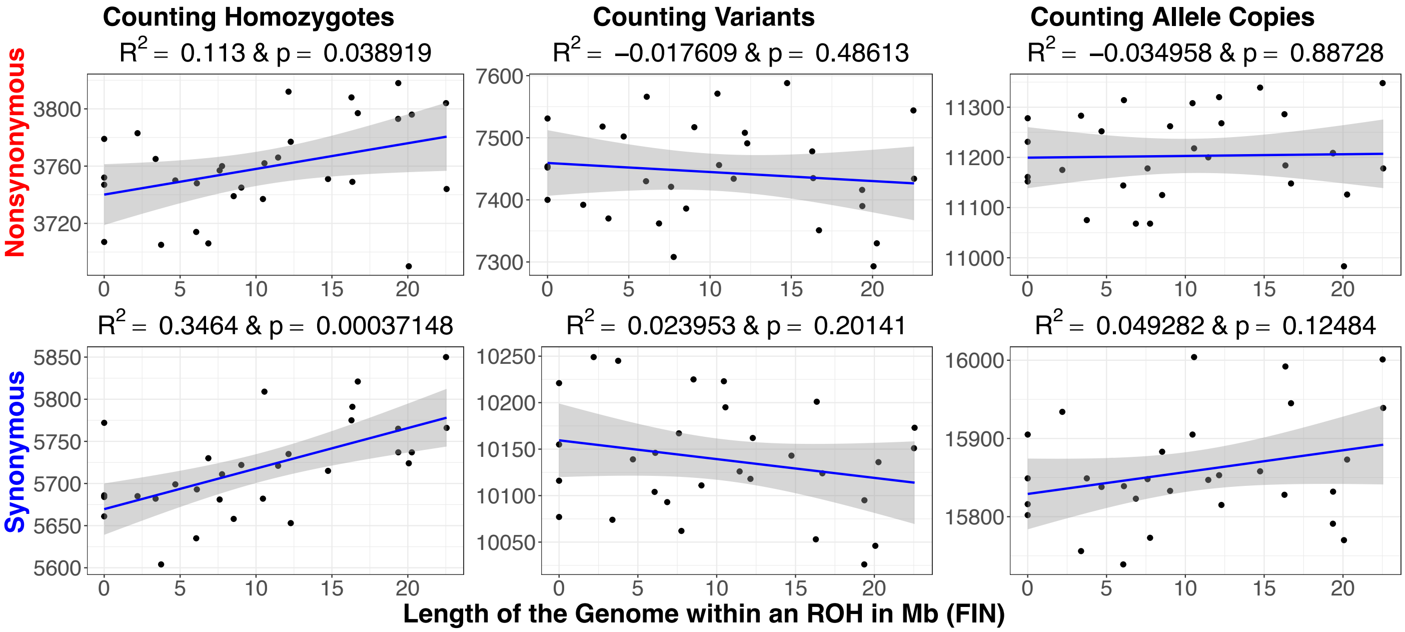

Figure S15. Correlation between length of the genome in an ROH and each counting method using nonsynonymous and synonymous mutations in the Finnish (*n* = 30). Each column represents a particular counting method (see **Methods**). The top row shows the correlation between nonsynonymous mutations and the length of the genome within an ROH. The bottom row shows the correlation between synonymous mutations and the length of the genome within an ROH. This figure is a zoomed in version of the correlations for FIN depicted in Figure 6. Population abbreviations are as in Figure S1.

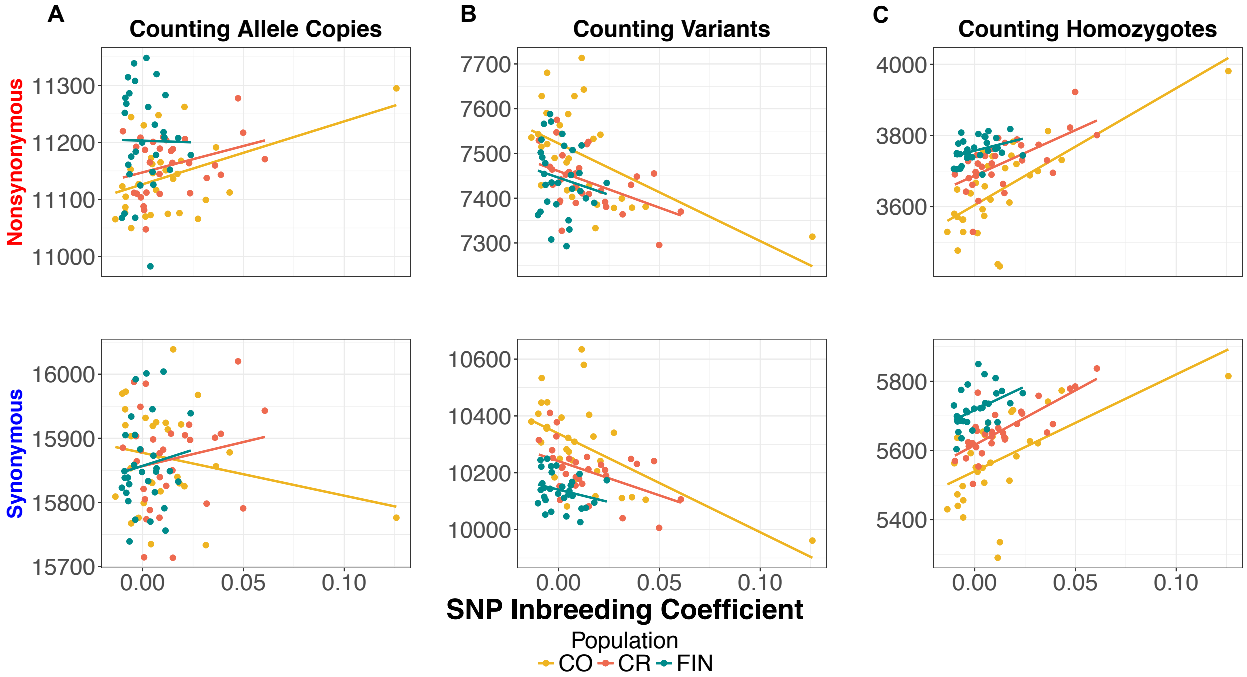

Figure S16. The relationship between the number of mutations per individual and the SNP inbreeding coefficient (F_{SNP}) using nonsynonymous and synonymous mutations in Colombians (CO), Costa Ricans (CR), and Finnish (FIN). This figure is comparable to Figure 6D-E where we examine the relationship between segregating sites and the length of the genome within an ROH. In this figure, we are using the same variant sites and instead examining the relationship with F_{SNP}. (A) Number of derived alleles per individual. (B) Number of variants per individual. (C) Number of homozygous derived genotypes per individual. The top row shows the correlation between nonsynonymous mutations and F_{SNP} across all population isolates, and the bottom row shows the correlation between synonymous mutations and F_{SNP} across all population isolates. Population abbreviations are as in Figure S1.

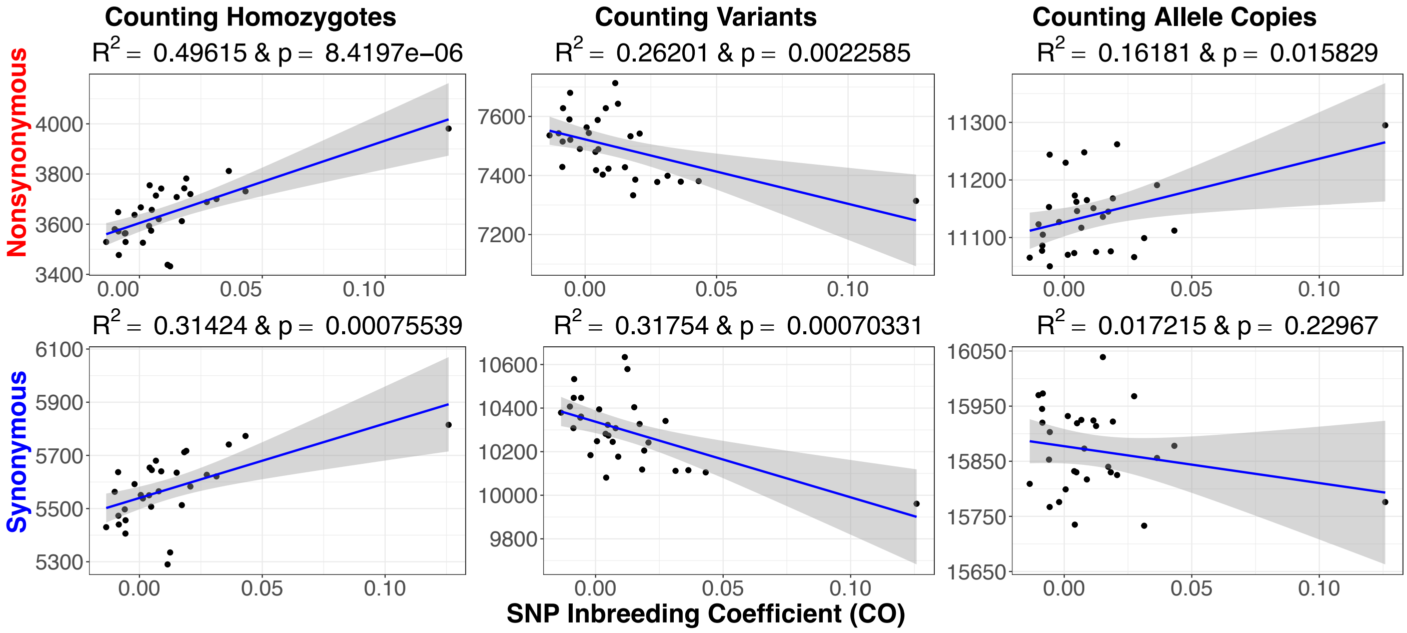

Figure S17. Correlation between F_{SNP} and each counting method using nonsynonymous and synonymous mutations in Colombians (*n* = 30). Each column represents a particular counting method (see **Methods**). The top row shows the correlation between nonsynonymous mutations and F_{SNP}. The bottom row shows the correlation between synonymous mutations and F_{SNP}. This figure is a zoomed in version of the correlations for CO depicted in Figure S16. Population abbreviations are as in Figure S1.

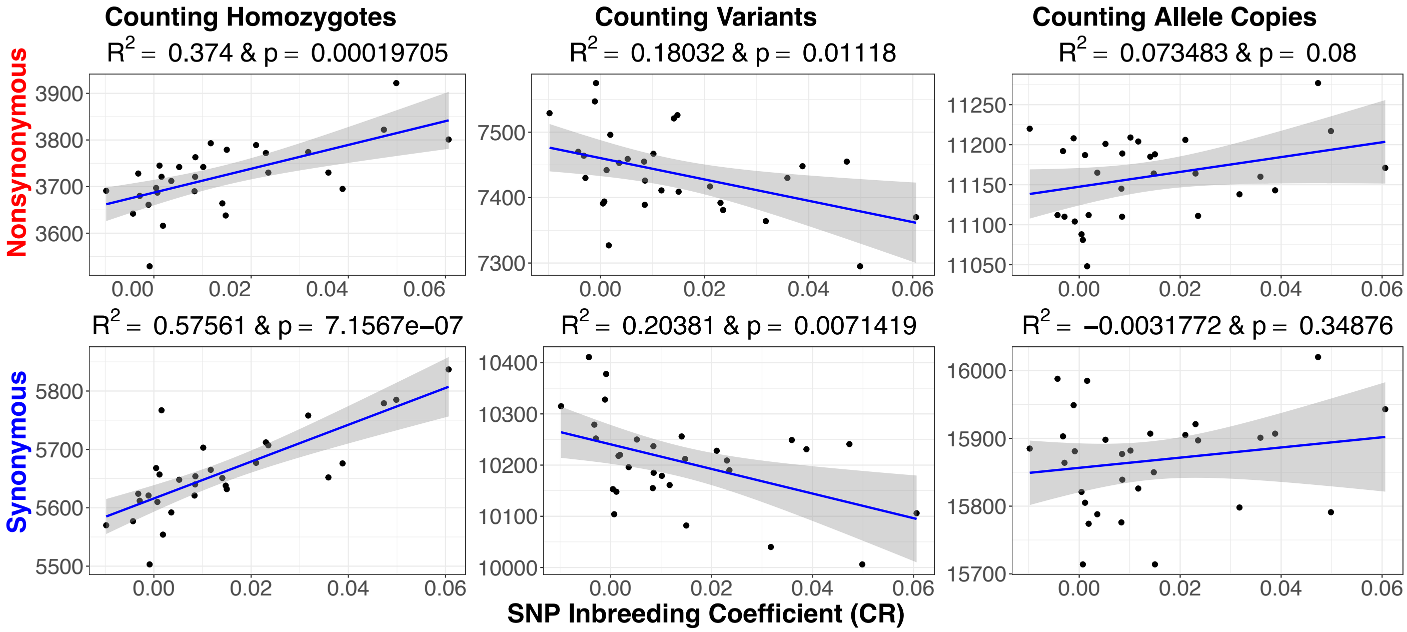

Figure S18. Correlation between F_{SNP} and each counting method using nonsynonymous and synonymous mutations in Costa Ricans (*n* = 30). Each column represents a particular counting method (see **Methods**). The top row shows the correlation between nonsynonymous mutations and F_{SNP}. The bottom row shows the correlation between synonymous mutations and F_{SNP}. This figure is a zoomed in version of the correlations for CR depicted in Figure S16. Population abbreviations are as in Figure S1.

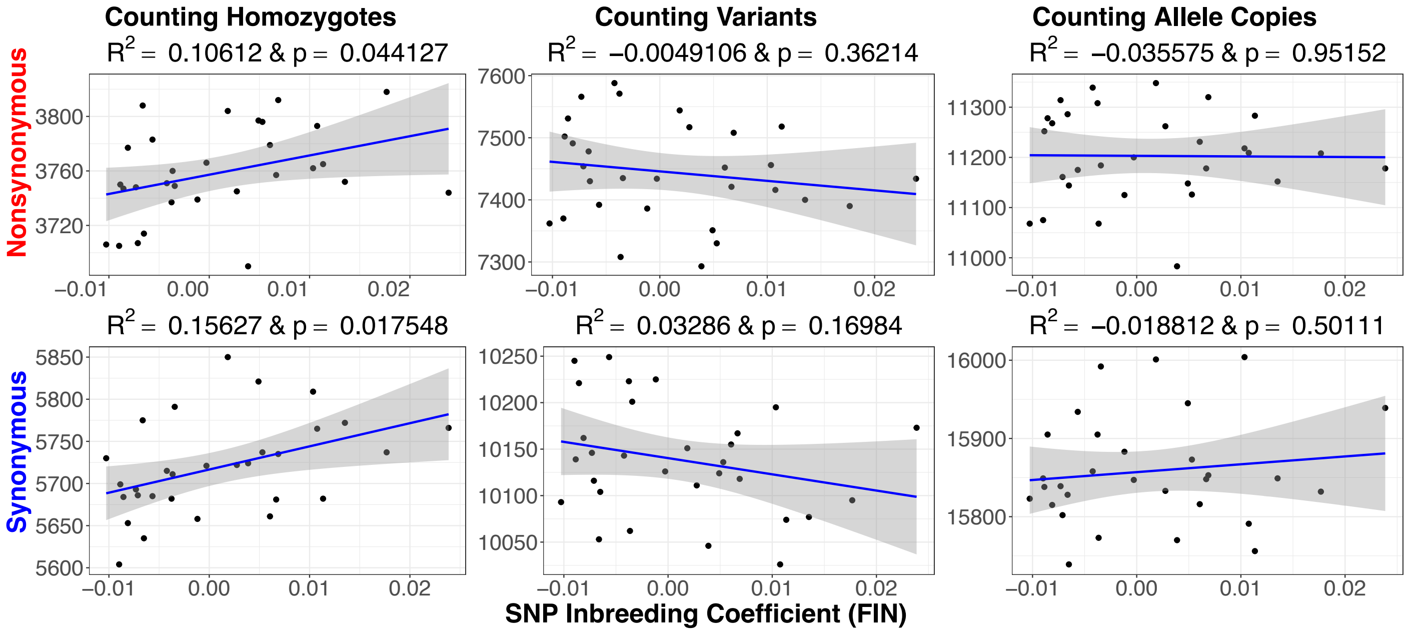

Figure S19. Correlation between F_{SNP} and each counting method using nonsynonymous and synonymous mutations in Finnish (*n* = 30). Each column represents a particular counting method (see **Methods**). The top row shows the correlation between nonsynonymous mutations and F_{SNP}. The bottom row shows the correlation between synonymous mutations and F_{SNP}. This figure is a zoomed in version of the correlations for FIN depicted in Figure S16. Population abbreviations are as in Figure S1.

Table S1. GenABEL correlations using both kinship matrix and genetic relatedness matrix (GRM).

Correlation	p-value (kinship matrix)	p-value (GRM)
F_{SNP} and F_{PED}	2.00e^{-16}	2.00e^{-16}
Length of Genome within ROH (>2Mb) and F_{PED}	2.00e^{-16}	2.00e^{-16}
Length of Genome within ROH (>2Mb) and F_{SNP}	2.00e^{-16}	2.00e^{-16}
European Ancestry and F_{PED}	5.21e^{-03}	3.56e^{-03}
Native American Ancestry and F_{PED}	2.45e^{-02}	1.48e^{-02}
African Ancestry and F_{PED}	4.96e^{-02}	4.26e^{-02}
European Ancestry and F_{SNP}	4.76e^{-12}	4.50e^{-12}
Native American Ancestry and F_{SNP}	2.79e^{-07}	1.71e^{-07}
African Ancestry and F_{SNP}	3.49e^{-08}	2.36e^{-08}
European Ancestry and Length of Genome within ROH (>2Mb)	1.02e^{-15}	3.40e^{-16}
Native American Ancestry and Length of Genome within ROH (>2Mb)	9.04e^{-12}	1.54e^{-12}
African Ancestry and Length of Genome within ROH (>2Mb)	2.50e^{-07}	1.64e^{-07}
Count Derived Deleterious Alleles and F_{PED}	4.32e^{-01}	5.84e^{-01}
Count Derived Deleterious Variants and F_{PED}	6.02e^{-06}	1.70e^{-05}
Count Derived Deleterious Homozygotes and F_{PED}	1.00e^{-06}	7.43e^{-07}
Count Neutral Alleles and F_{PED}	5.89e^{-02}	4.25e^{-02}
Count Neutral Variants and F_{PED}	2.26e^{-10}	4.99e^{-10}
Count Neutral Homozygotes and F_{PED}	1.03e^{-08}	3.38e^{-08}

The first column is the correlation being examined using GenABEL[47]. The second column contains the p-value from the tested correlation using a kinship matrix obtained from the Costa Rican and Colombian isolates pedigrees' using kinship2[34], and these are the p-values reported in the manuscript. The third column contains the p-value for the tested correlation using a GRM created using PC-AiR[21] and PC-Relate[22].

Table S2. Average values of pairwise differences (π) separated by genomic region using unrelated individuals in each population.

Population	Exonic	Intronic	Intergenic	Genome Wide
YRI	4.790e^{-04}	9.570e^{-04}	1.059e^{-03}	9.830e^{-04}
CEU	3.640e^{-04}	7.130e^{-04}	7.930e^{-04}	7.350e^{-04}
FIN	3.640e^{-04}	7.080e^{-04}	7.870e^{-04}	7.300e^{-04}
PUR	4.060e^{-04}	8.030e^{-04}	8.880e^{-04}	8.250e^{-04}
CO	3.870e^{-04}	7.700e^{-04}	8.500e^{-04}	7.890e^{-04}
CLM	3.850e^{-04}	7.630e^{-04}	8.460e^{-04}	7.840e^{-04}
CR	3.780e^{-04}	7.430e^{-04}	8.230e^{-04}	7.630e^{-04}
MXL	3.750e^{-04}	7.380e^{-04}	8.170e^{-04}	7.580e^{-04}
PEL	3.310e^{-04}	6.520e^{-04}	7.220e^{-04}	6.690e^{-04}

Each column represents the average value for a given region of the genome, or the whole genome. Averages were computed across the autosomes for each population. Population abbreviations are as in Figure S1.

Table S3. Average value of Watterson's theta (θ_w) separated by genomic region using unrelated individuals in each population.

Population	Exonic	Intronic	Intergenic	Genome Wide
YRI	0.00066	1.124e^{-03}	1.216e^{-03}	1.147e^{-03}
CEU	0.000419	6.900e^{-04}	7.520e^{-04}	7.070e^{-04}
FIN	0.000402	6.620e^{-04}	7.220e^{-04}	6.790e^{-04}
PUR	0.000555	9.330e^{-04}	1.006e-03	9.490e^{-04}
CO	0.000506	8.720e^{-04}	9.400e^{-04}	8.850e^{-04}
CLM	0.000496	8.330e^{-04}	8.980e^{-04}	8.470e^{-04}
CR	0.000468	7.920e^{-04}	8.580e^{-04}	8.080e^{-04}
MXL	0.00047	7.760e^{-04}	8.410e^{-04}	7.920e^{-04}
PEL	0.000394	6.580e^{-04}	7.160e^{-04}	6.730e^{-04}

Each column represents the average value for a given region of the genome, or the whole genome. Averages were computed across the autosomes for each population. Population abbreviations are as in Figure S1**.**