The American Journal of Human Genetics, Volume *108*

Supplemental information

Summix: A method for detecting and adjusting for population structure in genetic summary data

Ian S. Arriaga-MacKenzie, Gregory Matesi, Samuel Chen, Alexandria Ronco, Katie M. Marker, Jordan R. Hall, Ryan Scherenberg, Mobin Khajeh-Sharafabadi, Yinfei Wu, Christopher R. Gignoux, Megan Null, and Audrey E. Hendricks

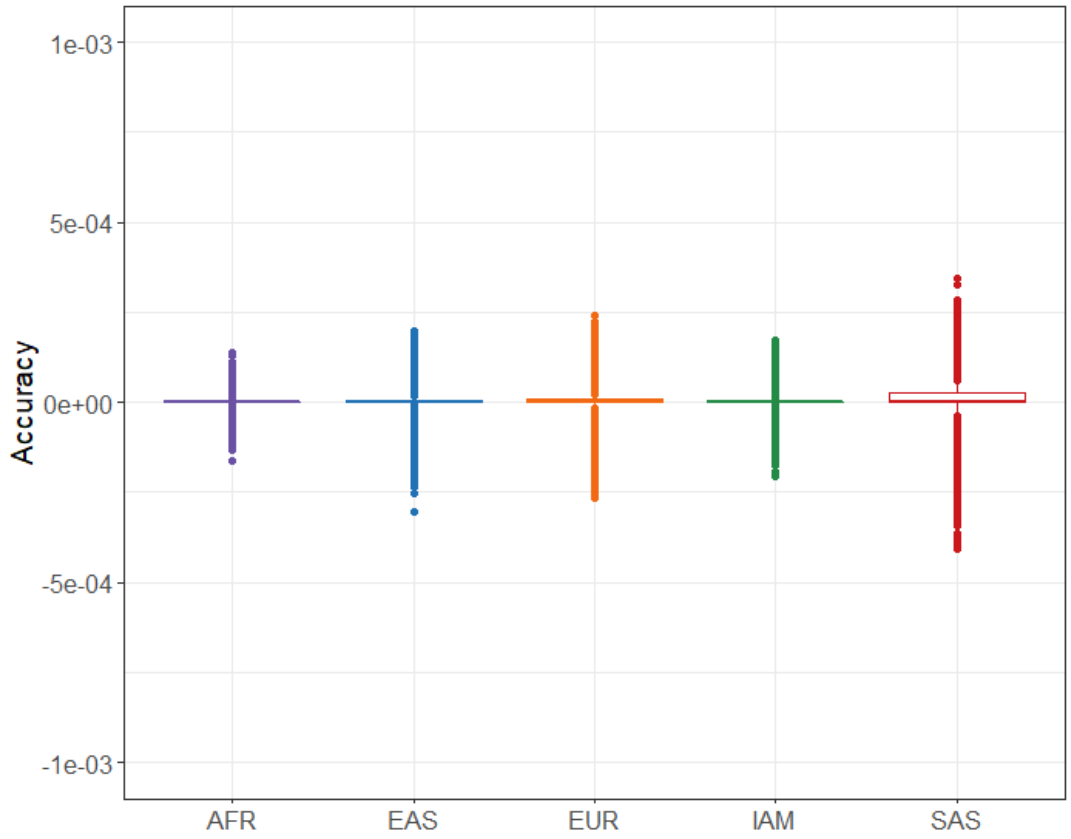

Figure S1. Simulation results for one ancestry parameters. Accuracy is defined as the absolute difference between the estimated ancestry proportions and given ancestry proportions within simulations. A single reference ancestry was used to simulate genotypes of a population.

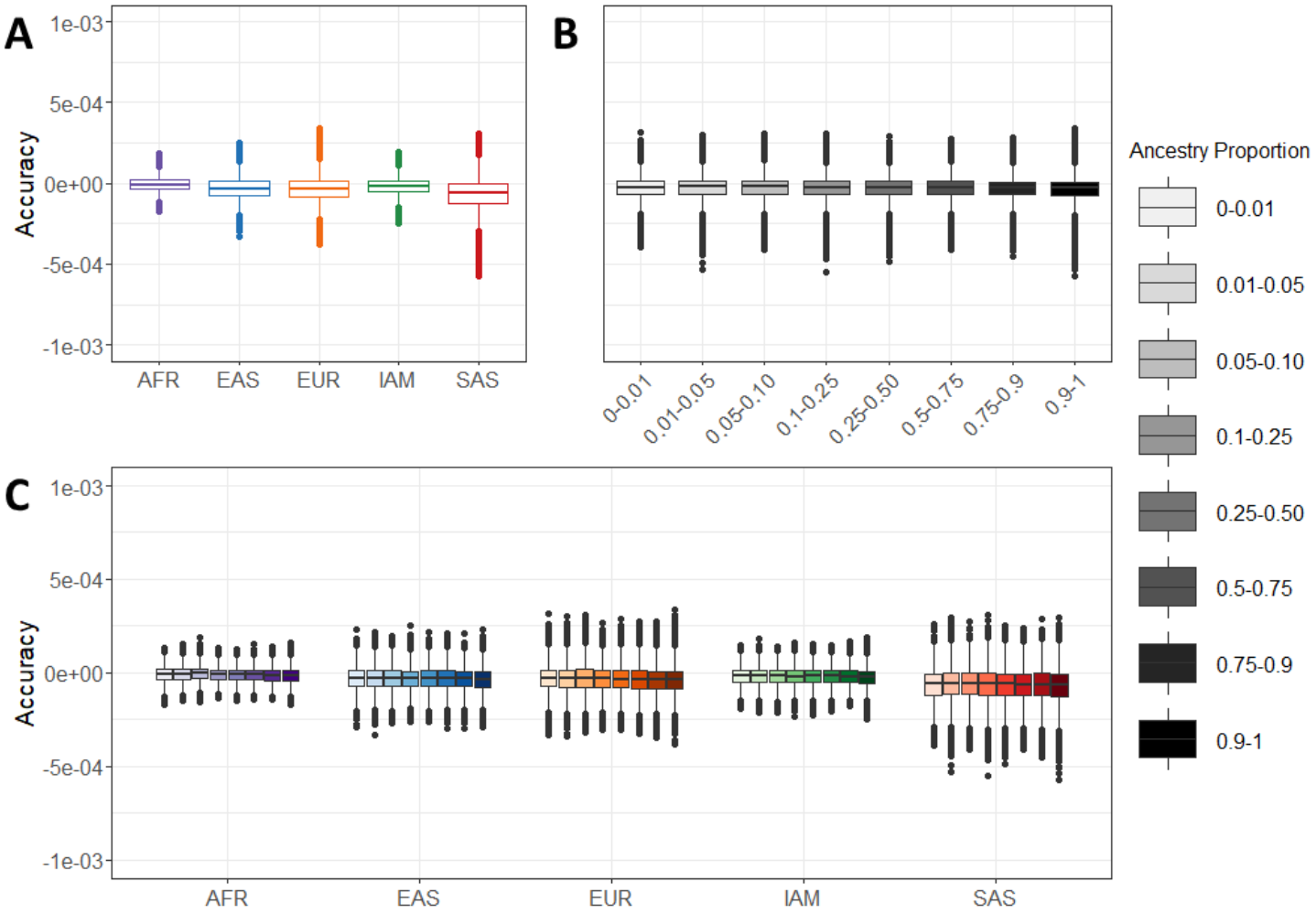

Figure S2. Simulation results for two ancestry parameters. Accuracy is defined as the absolute difference between the estimated ancestry proportions and given ancestry proportions within simulations. Two reference ancestries were used to simulate genotypes of an admixed population. **A)** Accuracy separated by ancestry. **B)** Accuracy separated by ancestry proportion. **C)** Accuracy separated by both ancestry and ancestry proportion.

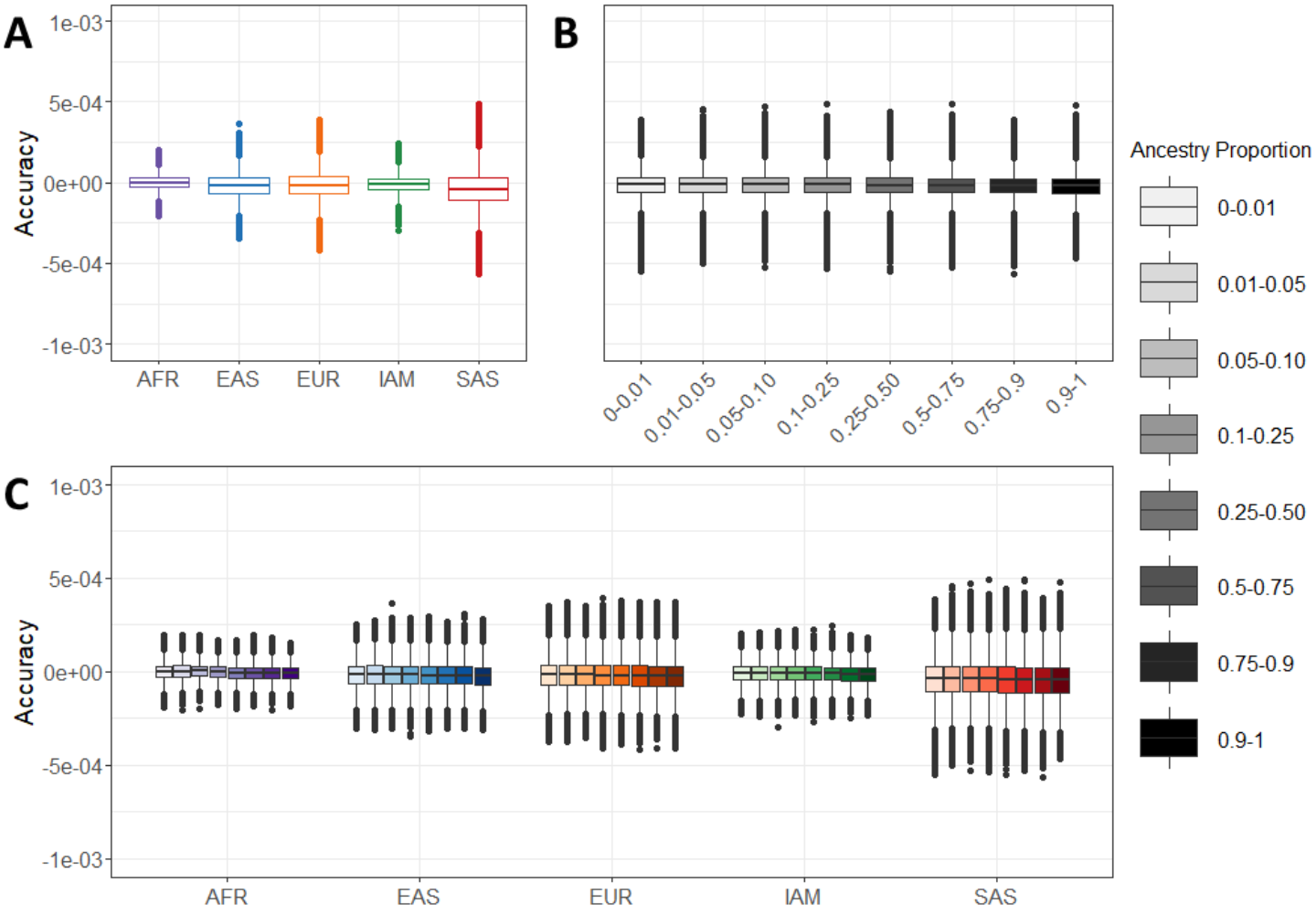

Figure S3. Simulation results for three ancestry parameters. Accuracy is defined as the absolute difference between the estimated ancestry proportions and given ancestry proportions within simulations. Three reference ancestries were used to simulate genotypes of an admixed population. **A)** Accuracy separated by ancestry. **B)** Accuracy separated by ancestry proportion. **C)** Accuracy separated by both ancestry and ancestry proportion.

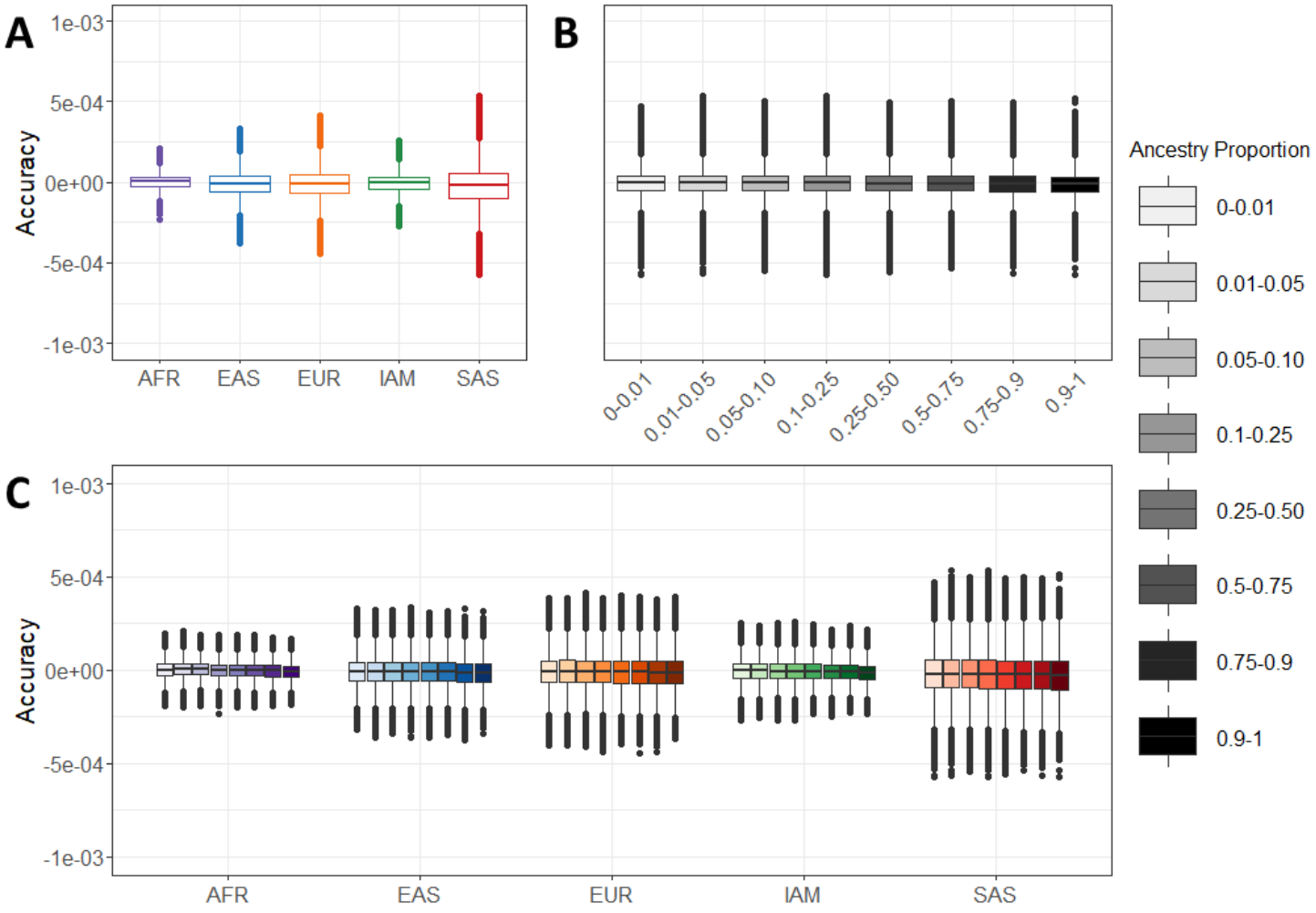

Figure S4. Simulation results for four ancestry parameters. Accuracy is defined as the absolute difference between the estimated ancestry proportions and given ancestry proportions within simulations. Four reference ancestries were used to simulate genotypes of an admixed population. **A)** Accuracy separated by ancestry. **B)** Accuracy separated by ancestry proportion. **C)** Accuracy separated by both ancestry and ancestry proportion.

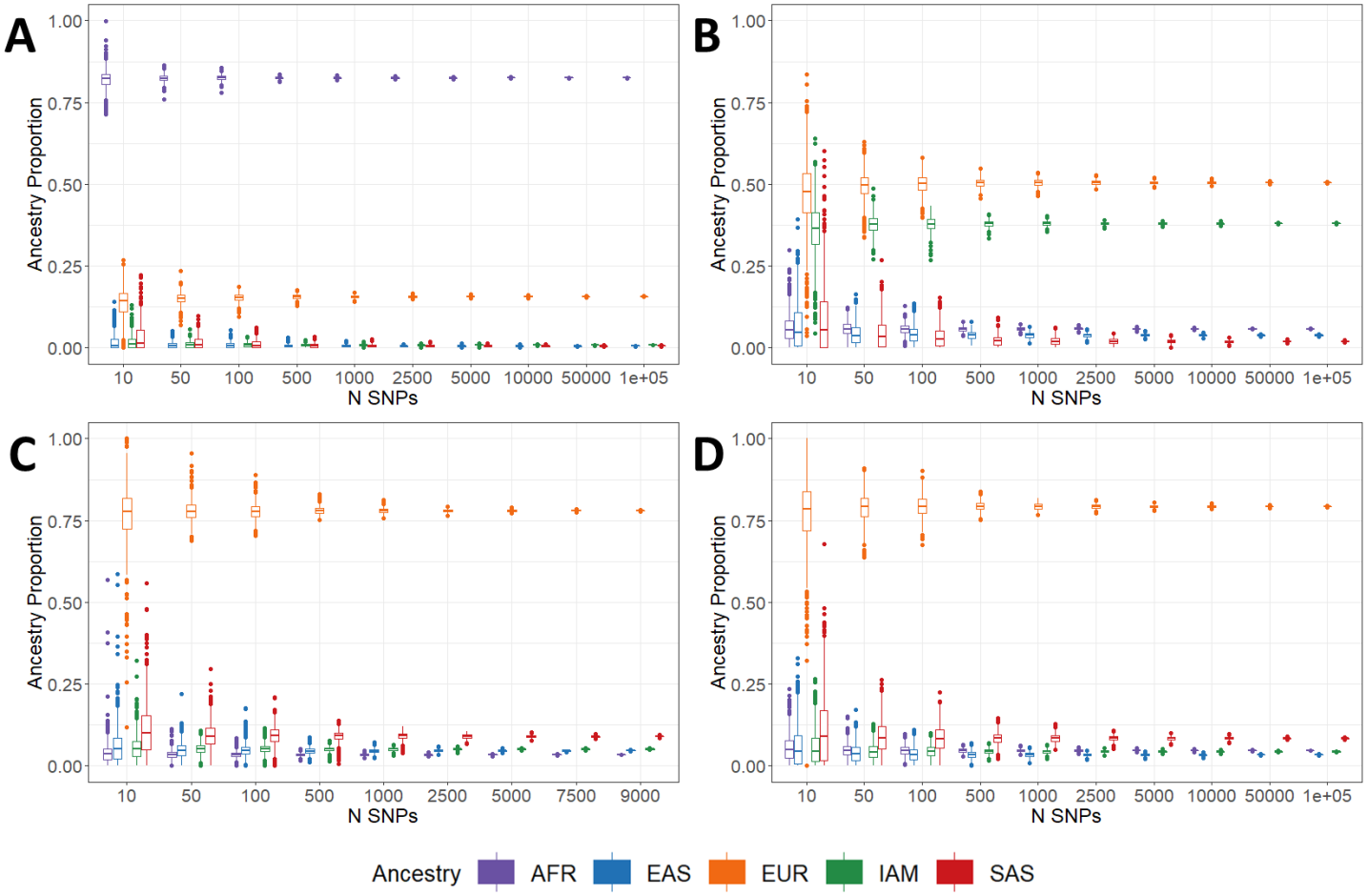

Figure S5. Precision in ancestry estimates for AFR, AMR and OTH gnomad groups by number of SNPs. Number of SNPs (x-axis), estimated ancestry proportion (y-axis) for 1,000 replicates; **A)** AFR genome. **B)** AMR genome. **C)** OTH exome. **D)** OTH genome.

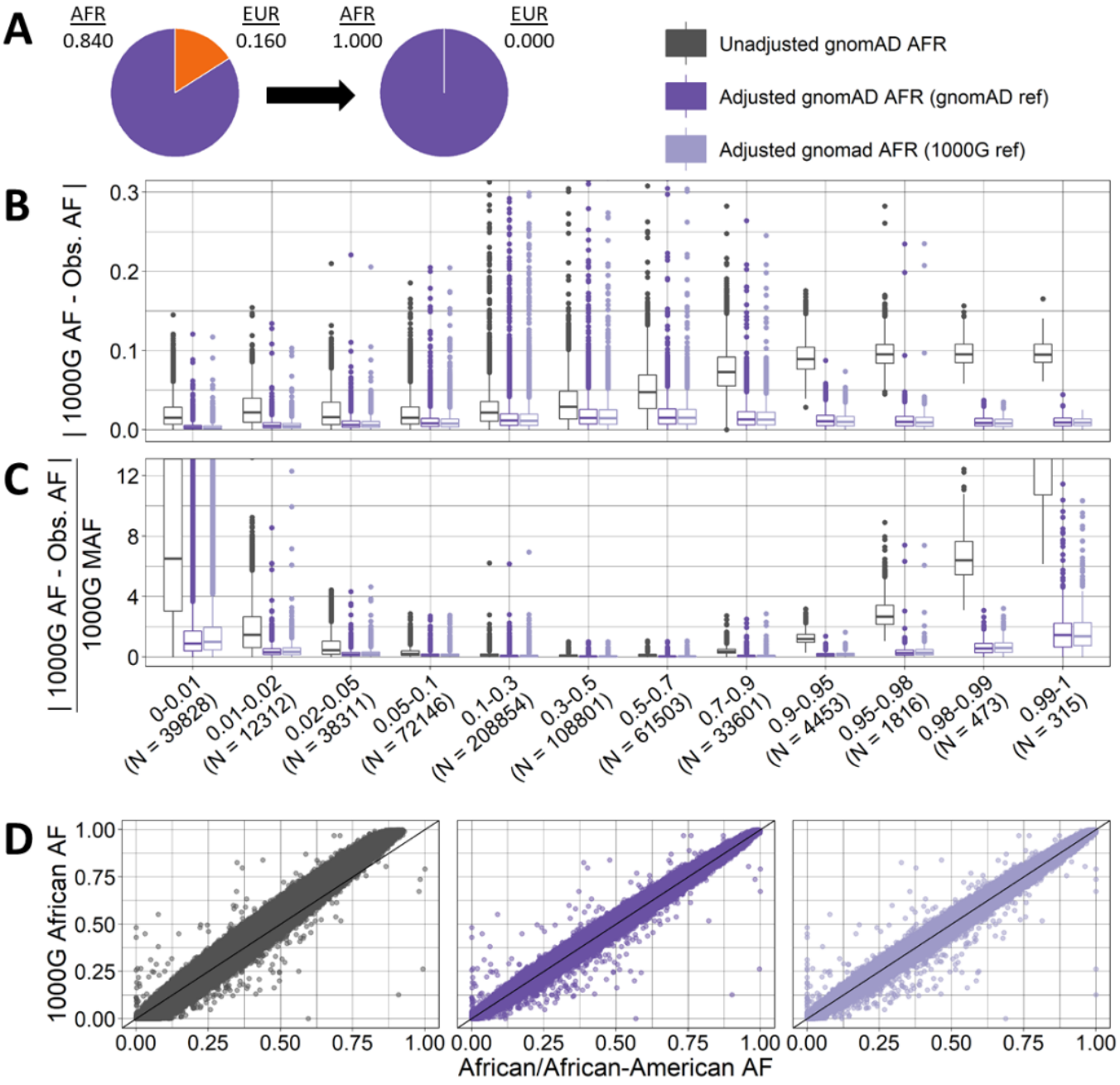

Figure S6. Ancestry-adjusted vs. unadjusted allele frequency for gnomAD African/African American genomes for a target sample with African ancestry. Ancestry-adjusted AF was estimated for a target sample with 100% African ancestry using gnomAD (dark purple) or 1000 Genomes (light purple) Non-Finnish European as reference and compared to unadjusted AF (grey) for 582,413 SNPs. **A)** ancestry proportions for gnomAD African/African American gnomes (AFR = 0.840, EUR=0.160) and target sample (AFR = 1); **B)** absolute difference between target sample AF (1000 Genomes African ancestry) and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category; **C)** relative difference between target 1000 Genomes African ancestry AF and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category; unzoomed figures B and C are available in the supplemental (**Figure S7**). **D)** scatter plot of target sample 1000 Genomes AF (y-axis) and unadjusted (left), ancestry-adjusted with gnomAD reference (center), and ancestry-adjusted with 1000 Genomes reference (right) gnomAD AF (x-axis).

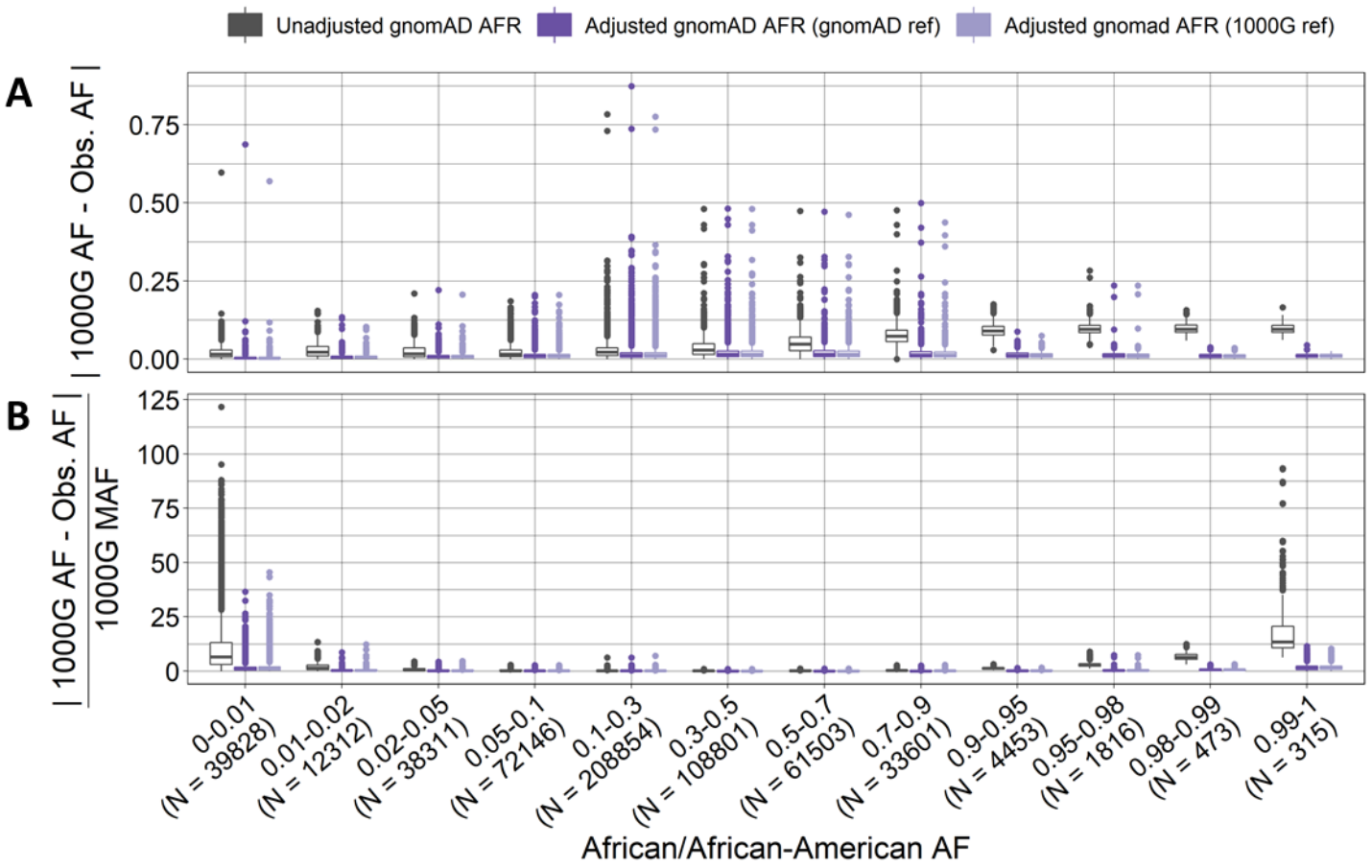

Figure S7. Unzoomed Ancestry-adjusted vs. unadjusted allele frequency for gnomAD African/African American genomes for a target sample with African ancestry (Complimentary to Figure S6). A) absolute difference between target sample AF (1000 Genomes African ancestry) and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category; **B)** relative difference between target 1000 Genomes African ancestry AF and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category.

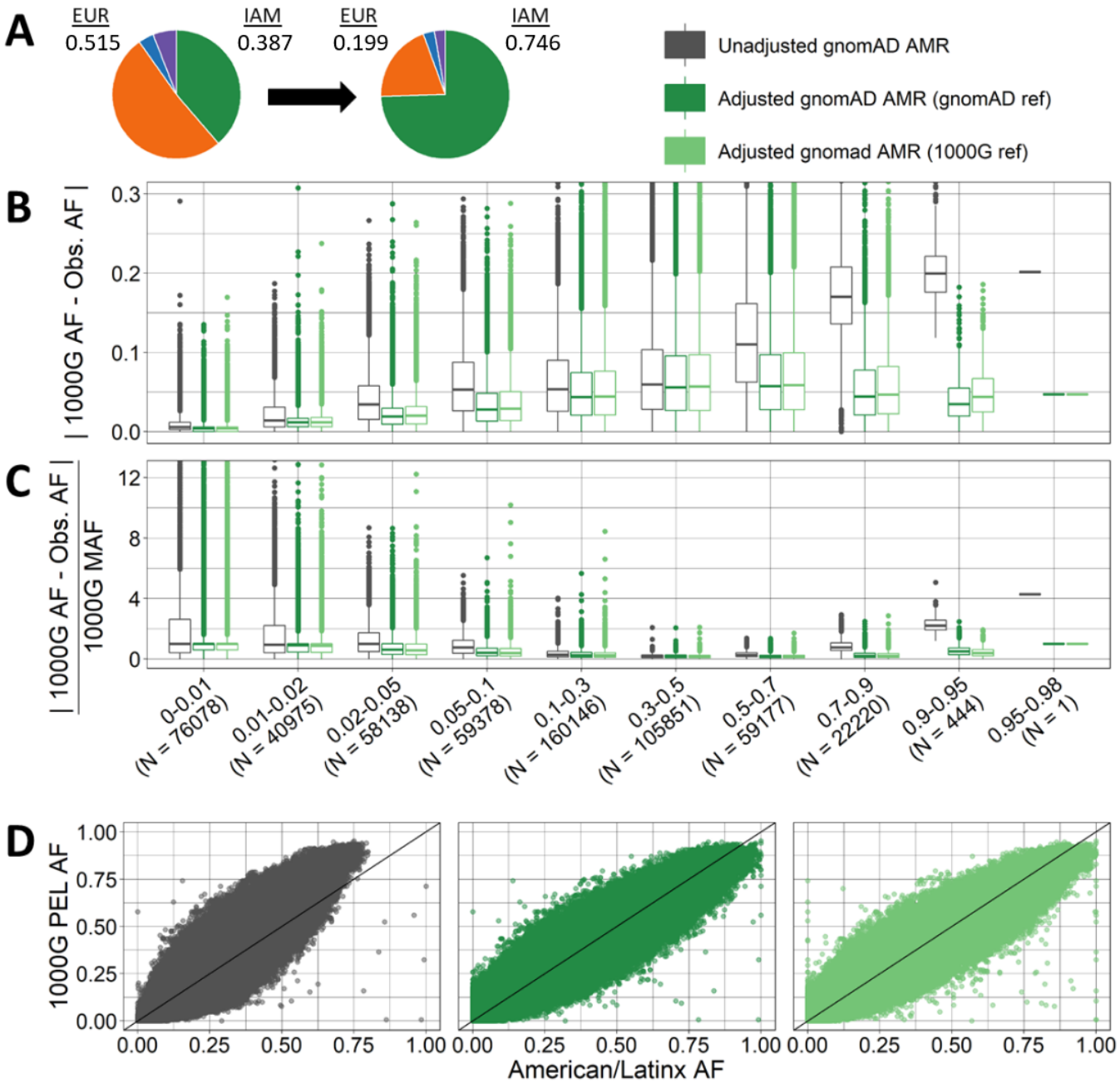

Figure S8. Ancestry-adjusted vs. unadjusted allele frequency for gnomAD American/Latinx genomes for a target sample of Peruvian ancestry. Ancestry-adjusted AF was estimated for a target Peruvian sample using gnomAD (dark green) or 1000 Genomes (light green) East Asian, European, and African as reference ancestral populations and compared to unadjusted AF (grey) for 582,408 SNPs. **A)** normalized ancestry proportions estimated for gnomAD American/Latinx genomes (purple AFR = 0.059, blue EAS=0.039, orange EUR=0.515, green IAM=0.387) and target Peruvian ancestry proportions (purple AFR = 0.028, blue EAS=0.027, orange EUR=0.199, green IAM=0.746); **B)** absolute difference between target 1000 Genomes Peruvian ancestry AF and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category; **C)** relative difference between target 1000 Genomes Peruvian ancestry AF and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category; unzoomed figures B and C are available in the supplemental (**Figure S9**). **D)** scatter plot of target 1000 Genomes AF (y-axis) and unadjusted (left), ancestry-adjusted with gnomAD reference (center), and ancestry-adjusted with 1000 Genomes reference (right) gnomAD AF (x-axis).

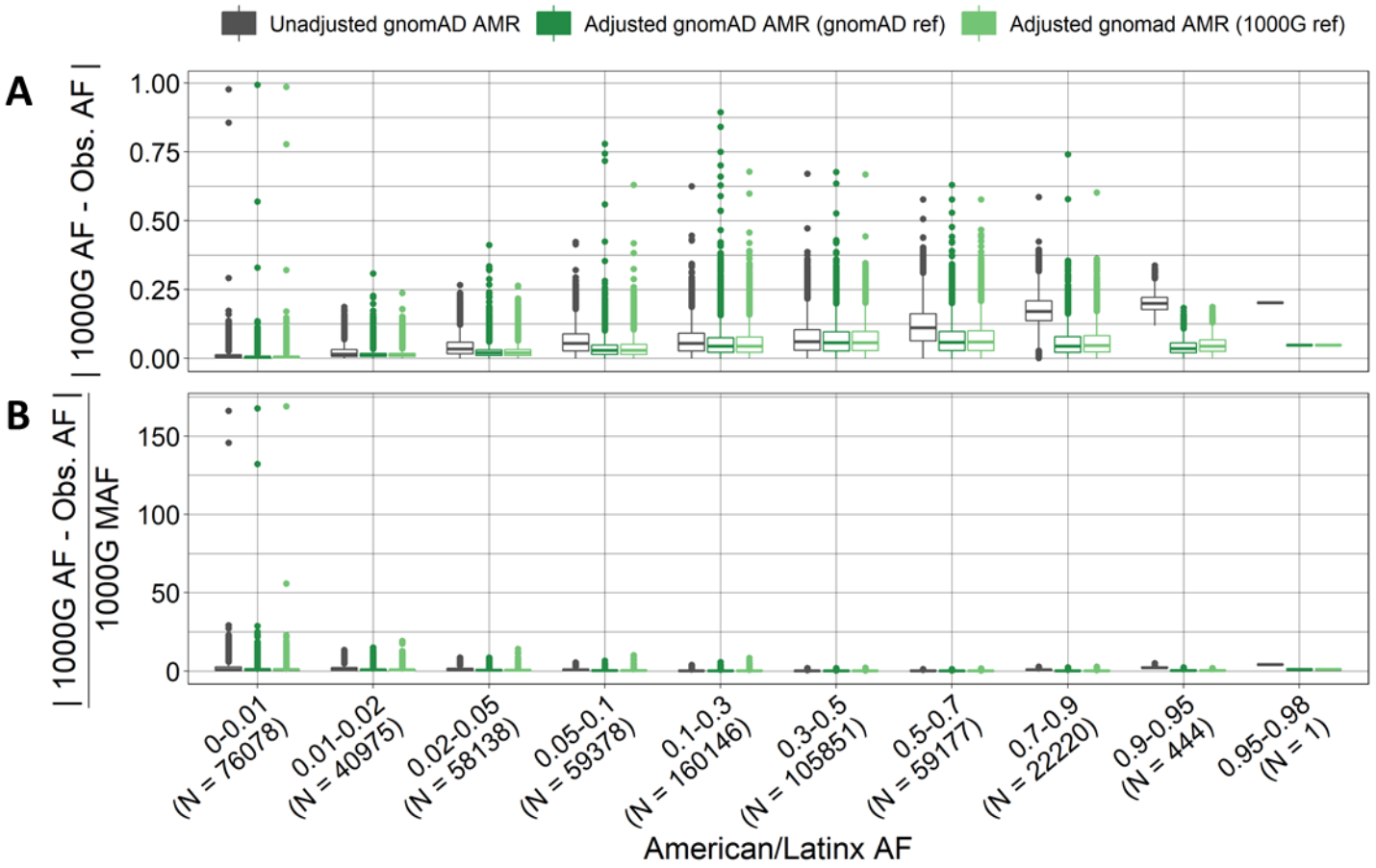

Figure S9. Unzoomed Ancestry-adjusted vs. unadjusted allele frequency for gnomAD American/Latinx genomes for a target sample of Peruvian ancestry (Complimentary to Figure S8). A) absolute difference between target 1000 Genomes Peruvian ancestry AF and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category; **B)** relative difference between target 1000 Genomes Peruvian ancestry AF and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category.

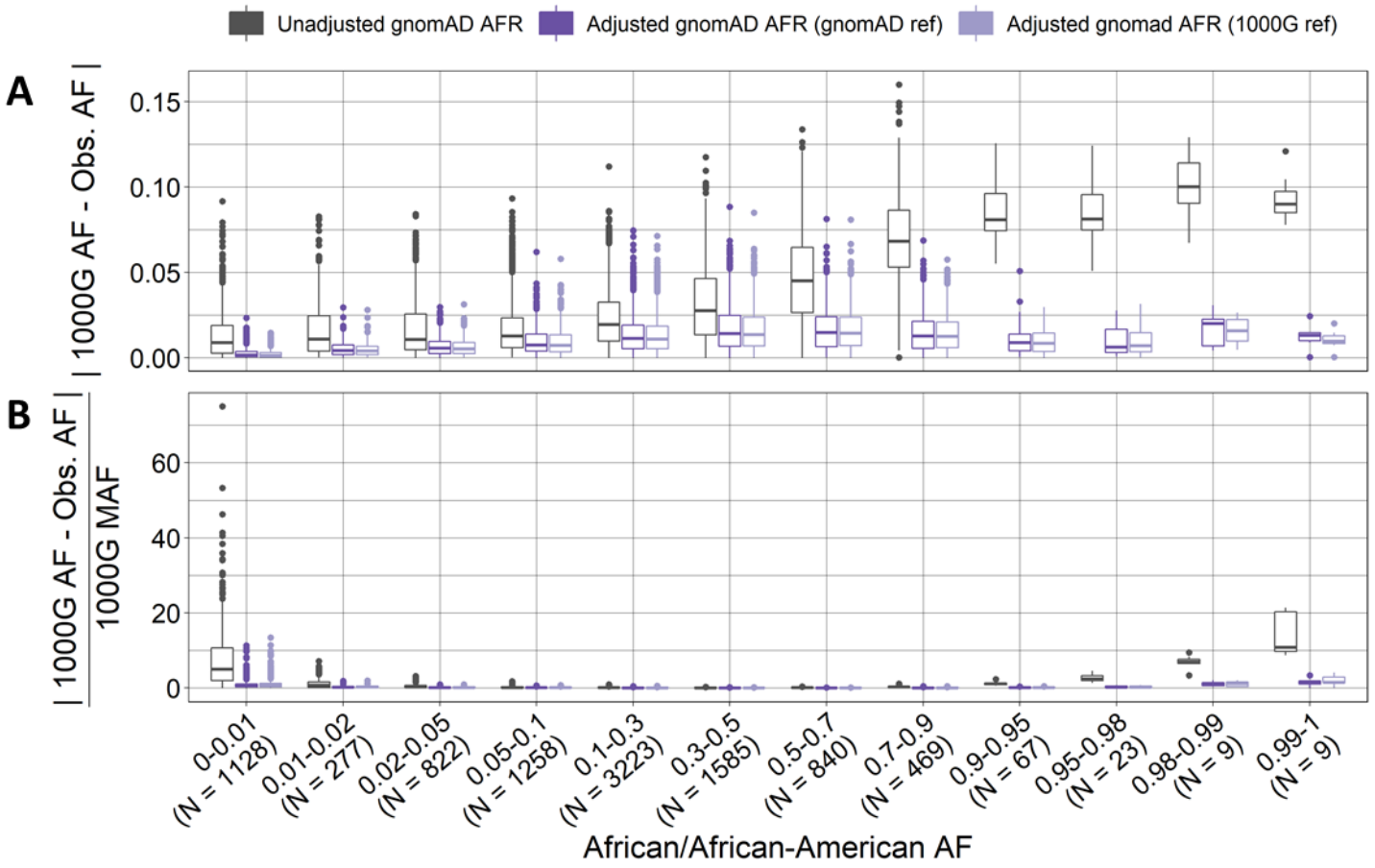

Figure S10. Unzoomed Ancestry-adjusted vs. unadjusted allele frequency for gnomAD African/African American exomes for a target sample with African ancestry (Complimentary to Figure 3). A) absolute difference between target sample AF (1000 Genomes African ancestry) and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category; **B)** relative difference between target 1000 Genomes African ancestry AF and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category.

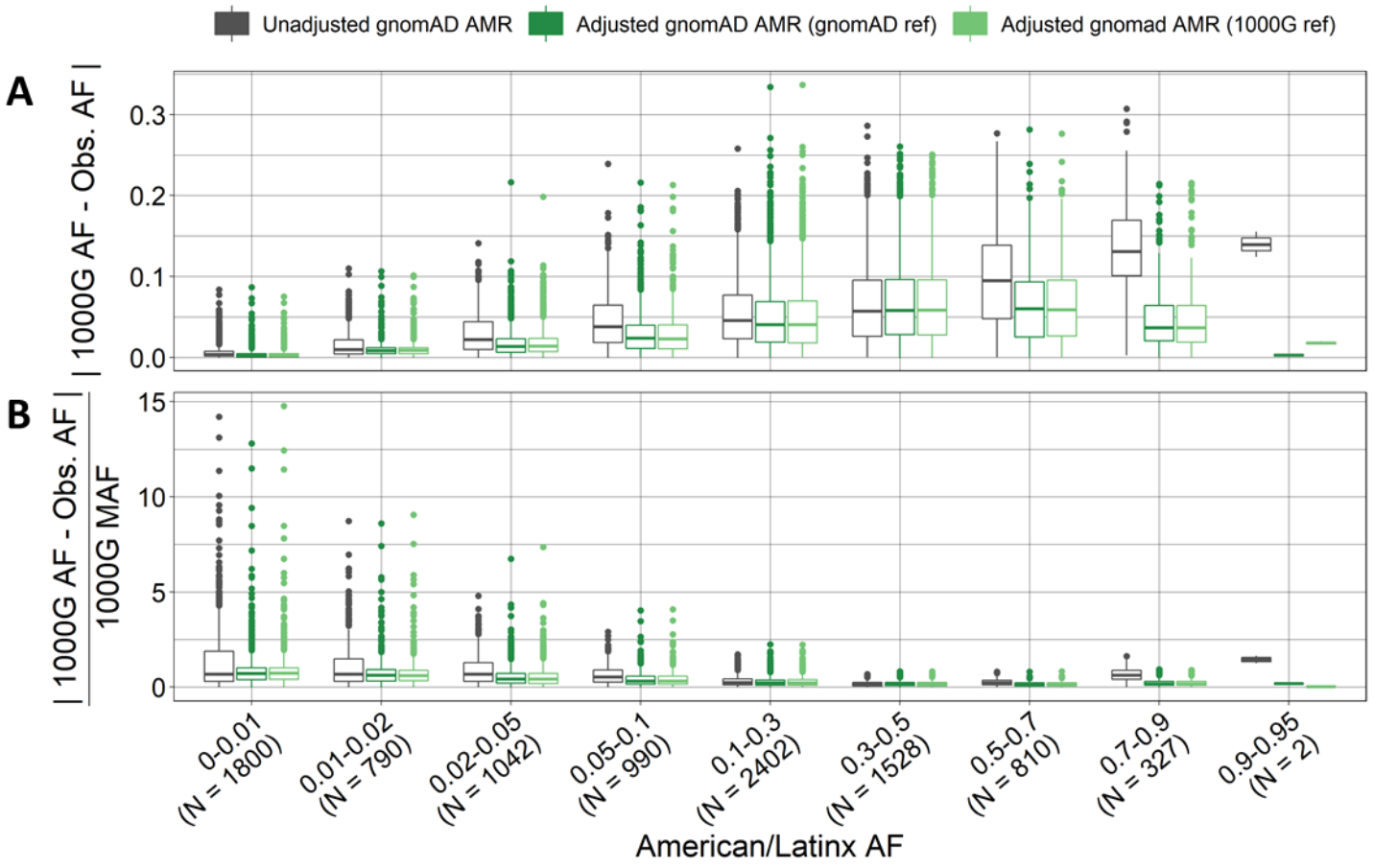

Figure S11. Unzoomed Ancestry-adjusted vs. unadjusted allele frequency for gnomAD American/Latinx exomes for a target sample of Peruvian ancestry (Complimentary to Figure 4). A) absolute difference between target 1000 Genomes Peruvian ancestry AF and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category; **B)** relative difference between target 1000 Genomes Peruvian ancestry AF and unadjusted or ancestry-adjusted gnomAD AF by 1000 Genomes AF category.

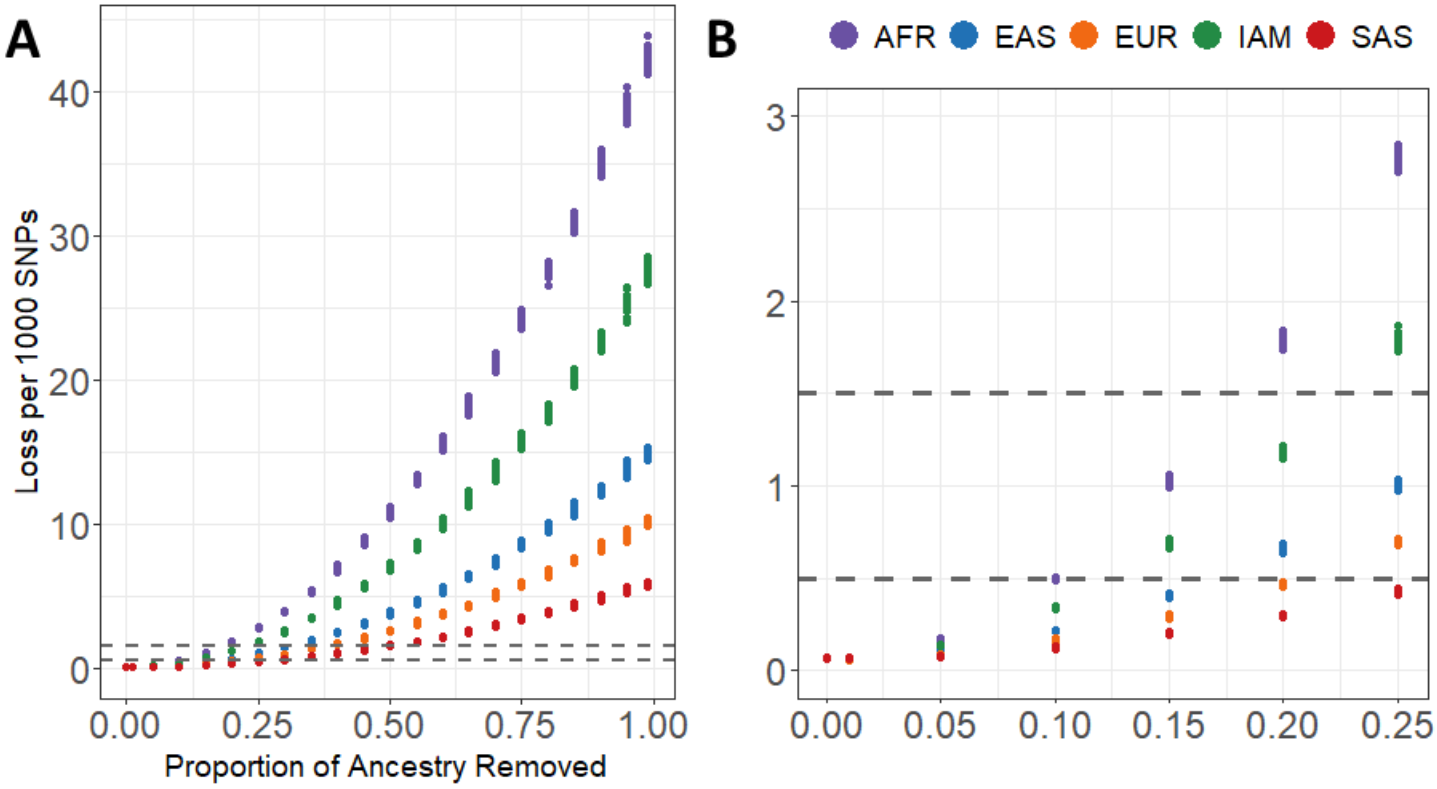

Figure S12. Reference panel sensitivity results for simulated admixed populations. 5-way admixed populations were simulated with a target ancestry held in constant proportion, and other 4 ancestries allowed to be random with constraint that all 5 sum to one. Target ancestry was removed from reference panel, which was used for ancestry estimation and least squares loss. **A)** Least-squares loss per 1000 SNPs for all simulation parameters. **B)** Zoomed plot showing proportion of ancestry removed between 0 and 0.25.

Table S12. Number of variants and proportion rounded to 0 and 1 after adjusting the allele frequency for ancestry.

		Target Sample	
		Number of SNPs (Proportion Corrected)	
	Reference data used	**African**	**Peruvian**
Exome	**gnomAD**	152 (0.015)	427 (0.049)
	1000 Genomes	184 (0.018)	473 (0.055)
Genome	**gnomAD**	4,937 (0.008)	49,115 (0.084)
	1000 Genomes	5,448 (0.009)	50,542 (0.087)

Table S13. Ancestry proportion estimates for 1000 Genomes Peruvian sample using Summix and ADMIXTURE (95% CI)

Ancestry	Summix	ADMIXTURE		
		Supervised Estimates	**Unsupervised Estimates**	**Projected Estimates from Unsupervised AF**
Indigenous American	0.723	0.768 (0.736, 0.799)	0.763 (0.732, 0.795)	0.724
Non-Finnish European	0.209	0.196 (0.166, 0.226)	0.199 (0.169, 0.229)	0.209
African	0.033	0.027 (0.018, 0.035)	0.027 (0.018, 0.035)	0.033
East Asian	0.035	0.010 (0.003, 0.017)	0.011 (0.003, 0.020)	0.033
Run Time*	3s	24m 55s	147m 58s	Unsupervised: 97m 51s Projection: 1m 51s

*Dual Intel Xeon E5-2670v2 2.5Ghz (10 core/20 thread) with 192GB DDR3-1600 ECC Registered Memory

Table S14. Unadjusted and adjusted values for *PADI3* variants from Malki et al. 2019.

rsID **Location** **DNA Sequence Variant** **AA Sequence Change**	**AN**	**unadjusted**			**100% AFR adjusted**		
		AF	**AC**	**Number of Homozygotes**	**AF**	**AC***	**Number of Homozygotes*** **
rs139426141 1-17597398-A-G c.856A→G p.Thr286Ala	24952	0.0365	910	23	0.0434	1082	47
rs34097903 1-17607274-G-A c.1744G→A p.Ala582Thr	24964	0.0227	566	6	0.0270	673	18
rs140482516 1-17607199-C-T c.1669C→T p.Arg557Trp	24968	0.0075	188	0	0.0089	223	2
rs1557508308 1-17597372-A-G c.832-2A→G splicing	0	0	0	0	0	0	0
rs1437225536 1-17609534-G-A c.1955G→A p.Arg652Lys	8716	0.0001	1	0	0.0001	1	0
rs139876092 1-17594433-C-T c.628C→T p.Arg210Trp	24962	0.00164	41	0	0.0019	48	0
Total	--	--	1706	29	--	2027	67

* $AC_{adj} = round(AF_{adj} * AN)$

**Adjusted number of homozygotes was estimated assuming Hardy-Weinberg equilibrium, $N_{homozygotes} = round(AF_{adj}^2 * AN)$

Table S15. Number of cases and African/African American gnomAD v2.1 controls with minor alleles in *PADI3* variants reported from Malki et al.

	Cases	gnomAD v2.1 African	
		unadjusted	100% AFR adjusted***
Minor allele	14	1677*	1960
No minor allele	44	10810**	10527
Chi Square p-value		0.029	0.114
Fisher's exact test p-value		0.031	0.101

* number of individuals with at least one minor allele was estimated by removing the number of homozygotes from the total minor allele count

** number of individuals with no minor allele was estimated as the total number of gnomAD v2.1 African/African American individuals (N=12,487) minus the number of estimates individuals with at least one minor allele

*** adjusted number of individuals was estimated by $\sum_{j=1}^{K} round\left(AN_j * AF_{adj_j} - AN_j * AF_{adj_j}^2\right)$ where AN_j and AF_{adj_j} are the observed allele number and adjusted allele frequency for variant j respectively. The values are rounded to the nearest integer before summing.

Table S16. Unadjusted and adjusted allele frequency for F508del in *CFTR*

gnomAD v2.1 group	AC	AN	Number of Homozygotes	Unadjusted AF	Adjusted AF*
Exomes					
Non-Finnish European	1394	113626	1	0.0123	NA
African/ African American	48	16248	0	0.0030	0.0014
Genomes					
Non-Finnish European	204	15408	0	0.0132	NA
African/ African American	17	8710	0	0.0020	0**

* Adjusted AF for 100% African using gnomAD Non-Finish European AF and assuming only European admixture

** We recommend caution when interpreting adjusted AF at or close to 0.

TABLE S19. Least-squares loss per 1000 SNPs across exome and genome gnomAD groups.

Ancestry	Exome			Genome		
	SNPs	**Iterations**	**Loss/1000**	**SNPs**	**Iterations**	**Loss/1000**
African/African-American	9750	24	0.234	582156	32	0.221
American/Latinx	9722	38	1.080	582155	45	0.824
Other	9749	25	0.451	582156	42	0.680
Non-Finnish European	9763	29	0.374	582156	24	0.500
East Asian	9732	30	0.346	582155	20	0.433
South Asian	9719	44	0.337	--	--	--
Finnish	9728	49	2.617	582155	85	2.618
Ashkenazi Jewish	9749	121	2.047	582156	68	2.462