The American Journal of Human Genetics, Volume *100*

Supplemental Data

Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations

Alicia R. Martin, Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny

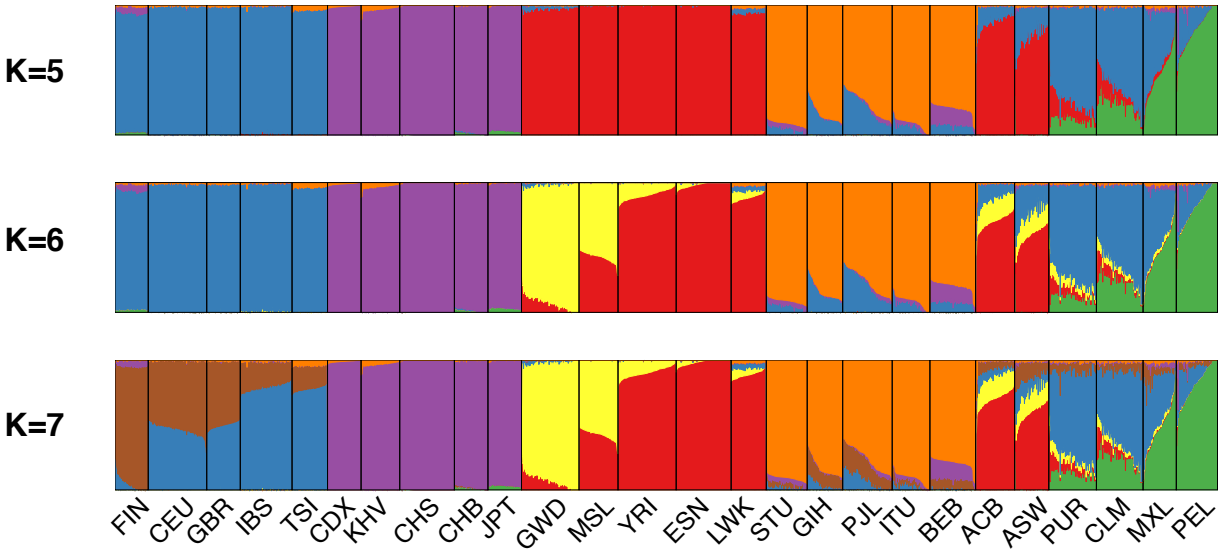


Figure S1 – ADMIXTURE analysis at K=5, K=7, and K=8. K=8 has the lowest 10-fold cross-validation error of K=3-12. At K=5, this analysis separates continental ancestries in the super populations (AFR, AMR, EAS, EUR, and SAS, population abbreviations in Table S1). These results also highlight sub-continental substructure; for example, there is detectable substructure resembling European (EUR) and East Asian (EAS) ancestries in the SAS populations (population means range from 6.1-15.9% and 0.3-12.2%, respectively), with the highest rates of East Asian-like ancestry in the Bengalis from Bangladesh (BEB). In contrast, the greatest quantity of European-like ancestry in the SAS populations is in the Punjabi from Lahore, Pakistan (PJL), who are geographically the closest to Europe. Ancestral clines have been observed along geographical, caste, and linguistic axes in more densely sampled studies of South Asia.[1,2] Increasing the model to K=6 there is also an east-west cline among African populations, while at K=7 we observe the north-south cline of European ancestry.[3] While there is minimal Native American ancestry (<1%) in most African Americans across the United States, there is a substantial enrichment in several ASW individuals from 1000 Genomes (mean of 3.1%, and 9 samples with >5%, including NA19625, NA19921, NA20299, NA20300, NA20314, NA20316, NA20319, NA20414, and NA20274).[4,5] Interestingly, one ASW individual has no African ancestry (NA20314, EUR= 0.40, NAT=0.59) but is the mother of NA20316 in an ASW duo with few Mendelian inconsistencies that suggest that the father mostly likely has ~80% African and ~20% European ancestry, similar to other ASW individuals. We also find evidence of East Asian admixture in several PEL samples (39% in HG01944, 12% in HG02345, 6% in HG0192, 5% in HG01933, and 5% in HG01948). Consistent with the autosomal evidence, the Y chromosome haplogroup for HG01944 (Q1a-M120) clusters most closely with two KHV samples and other East Asians rather than the Q-L54 subgroup expected in samples from South America.[6]

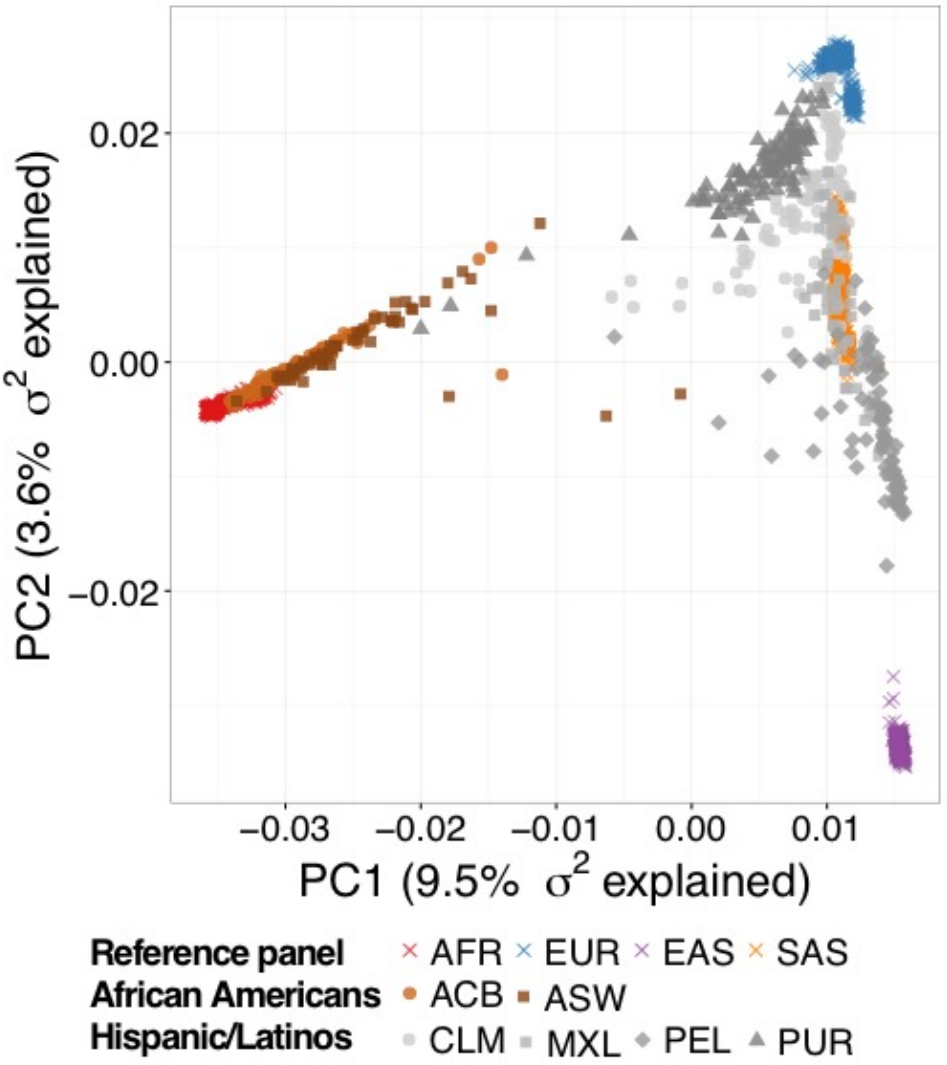


Figure S2 – Principal components analysis of all samples showing the relative homogeneity of AFR, EUR, EAS, and SAS continental groups and continental mixture of admixed samples from the Americas (ACB, ASW, CLM, MXL, PEL, and PUR).

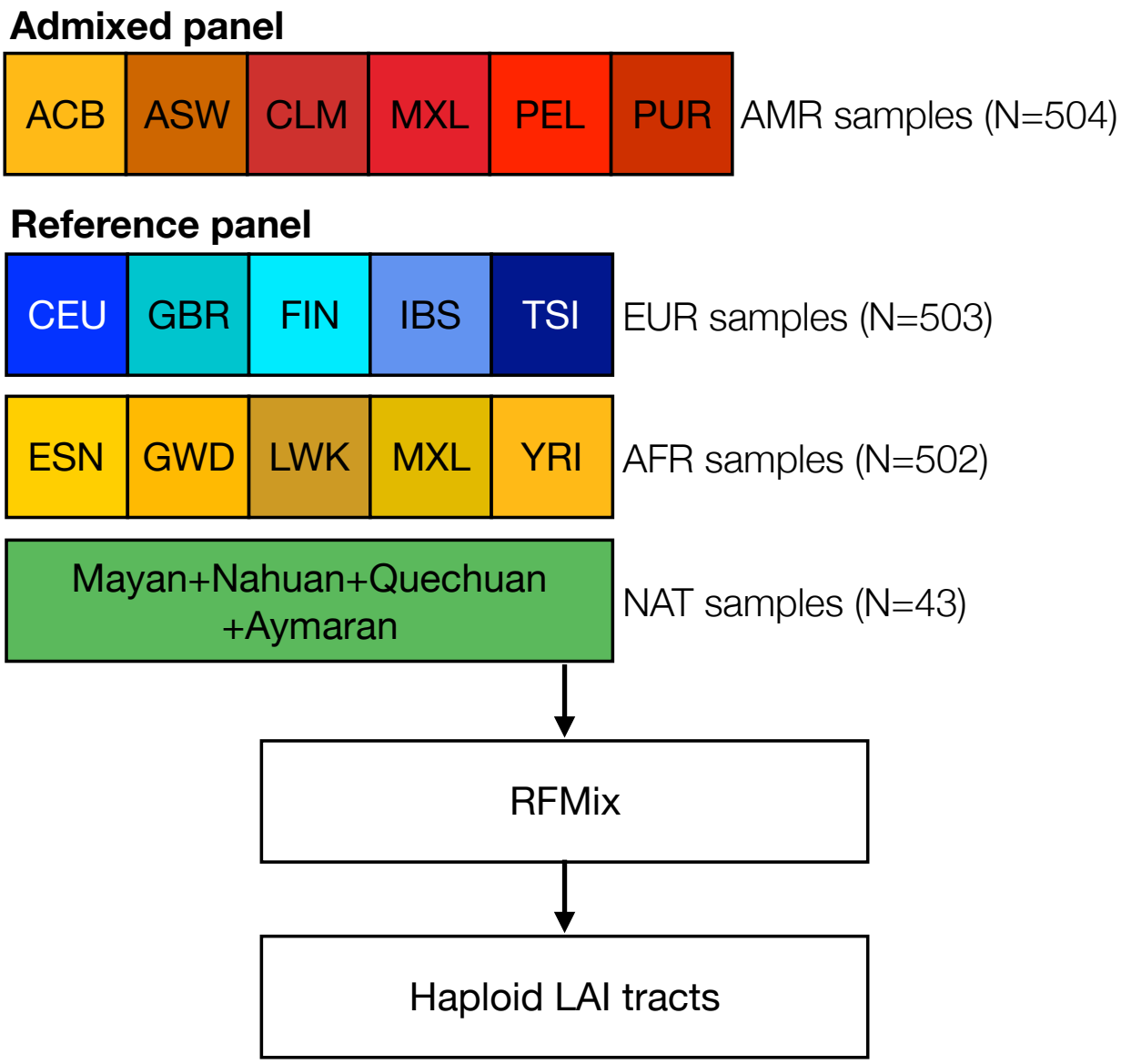


Figure S3 – Schema of local ancestry calling pipeline

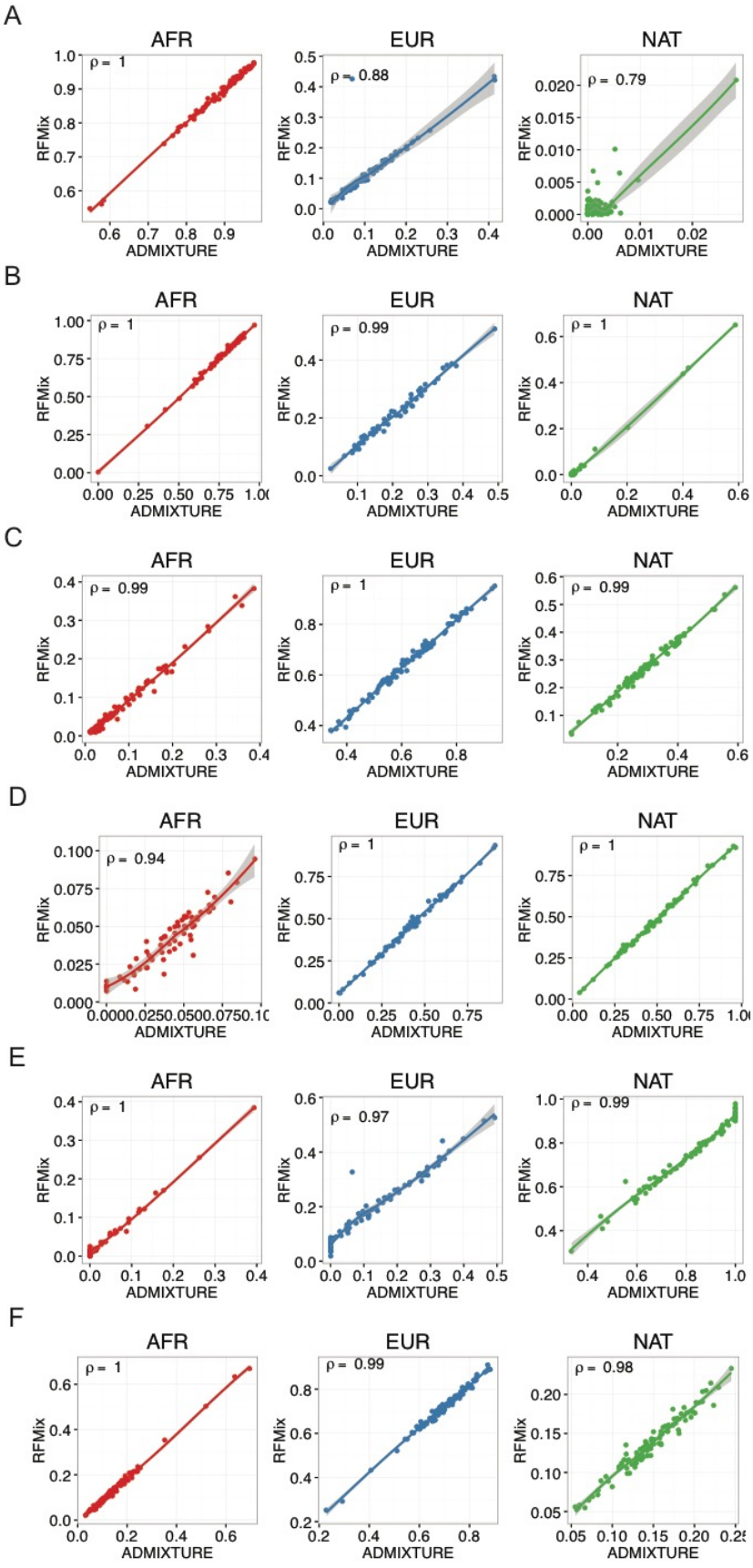

A
AFR
EUR
NAT
ρ = 1
ρ = 0.88
ρ = 0.79
B
ρ = 1
ρ = 0.99
ρ = 1
C
ρ = 0.99
ρ = 1
ρ = 0.99
D
ρ = 0.94
ρ = 1
ρ = 1
E
ρ = 1
ρ = 0.97
ρ = 0.99
F
ρ = 1
ρ = 0.99
ρ = 0.98
RFMix
ADMIXTURE

Figure S4 – Concordance between global ancestry estimates across individuals via Pearson's correlation from ADMIXTURE at K=5 as in Figure S1 versus 3-way RFMix inferences for AFR, EUR, and NAT ancestries. The correlation between ADMIXTURE and global ancestry estimates from RFMix was lower when there was minimal ancestry from a given source population and/or tracts were very short (<5 cM), e.g. NAT ancestry in the ACB (ρ=0.79) and AFR ancestry in the MXL (ρ=0.94). A) ACB. Substantial differences occurred in 1 ACB individual, HG01880, where considerable South Asian ancestry (31.8%) was classified as European ancestry due to limitations of the 3-way local ancestry reference panel. B) ASW. C) CLM. D) MXL. E) PEL. Substantial differences occurred in 2 PEL individuals, HG01944 and HG02345, where considerable East Asian ancestry (38.2% and 12.3%, respectively) was classified in RFMix as EUR and NAT ancestry due to limitations of the 3-way local ancestry reference panel. F) PUR.

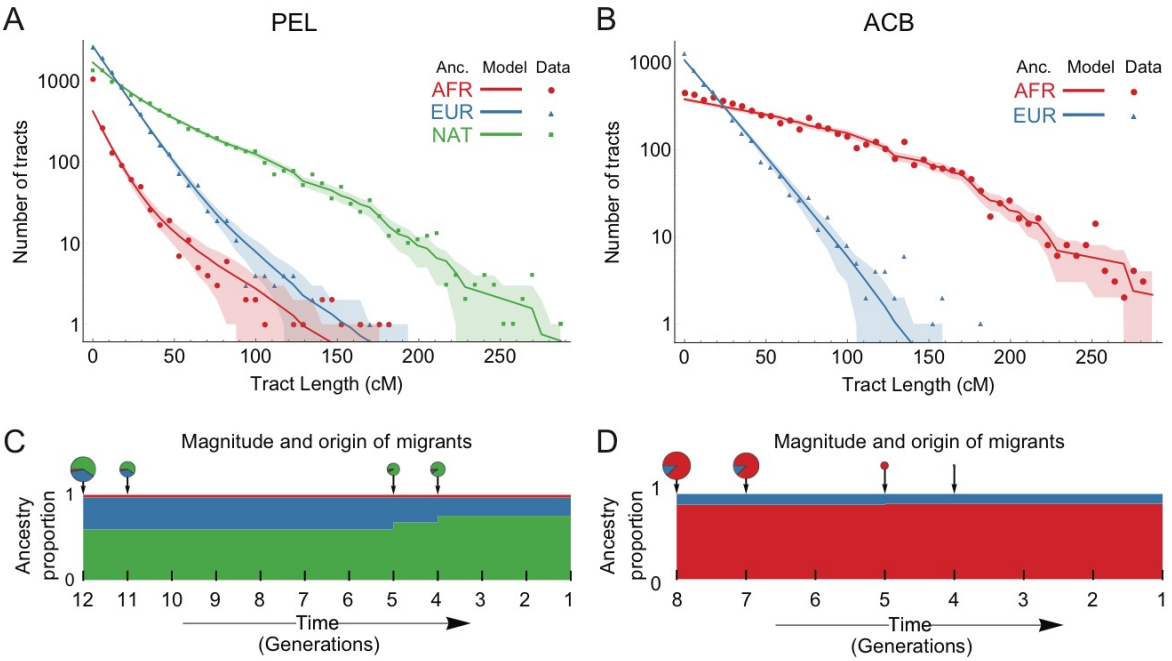


Figure S5 – Demographic reconstruction through genetically dated recent admixture events in the Americas. A-B) Local ancestry tract length decay of AFR, EUR, and NAT continental ancestry tracts for the A) PEL and B) ACB. Points represent the observed distribution of ancestry tracts, and solid lines represent the distribution of the best-fit Markov model inferred using *Tracts*, with the shaded areas indicating one standard deviation confidence intervals. C-D) Admixture time estimates in number of generations ago, relative quantity of migrants, and ancestry proportions over time under the best-fitting model for the C) PEL and D) ACB. C) The best-fit model for the PEL begins ~12 generations ago, which is slightly more recent than for insular and Caribbean mainland populations. For example, admixture in Colombian and Honduran mainland populations was previously inferred to have begun 14 generations ago, whereas admixture in Cuban, Puerto Rican, Dominican, and Haitian populations began 16-17 generations ago.[7] There is minimal African ancestry (2.9%), some European ancestry (37.6%) and primarily Native ancestry (59.4%) in the first pulse of admixture, followed by a later pulse (~5 generations ago) of primarily Native ancestry (91.1%). This later pulse of primarily Native ancestry is unique to the PEL compared to other admixed populations of the Americas.[7] D) The best-fit model for the ACB was an initial pulse of admixture between Europeans and Africans followed by a later pulse of African ancestry. The best model indicates that admixture in the ACB began ~8 generations ago with the initial pulse containing 87.4% African ancestry and 12.6% European ancestry. The second pulse of African ancestry began ~5 generations ago and had only a minor overall contribution (4.4% of total pulse ancestry), which is consistent with either a later small pulse of African ancestry or movement of populations within the Caribbean. The admixture events we infer in the ACB are more recent than previous ASW and African American two-pulse models, which estimated that admixture began ~10-11 generations ago.[4,8] Potential explanations for this small difference include differences in the ages of individual between the two cohorts and the fact that pulse timings indicate the generations that admixture most likely spanned rather than the exact generation during which admixture began.[7]

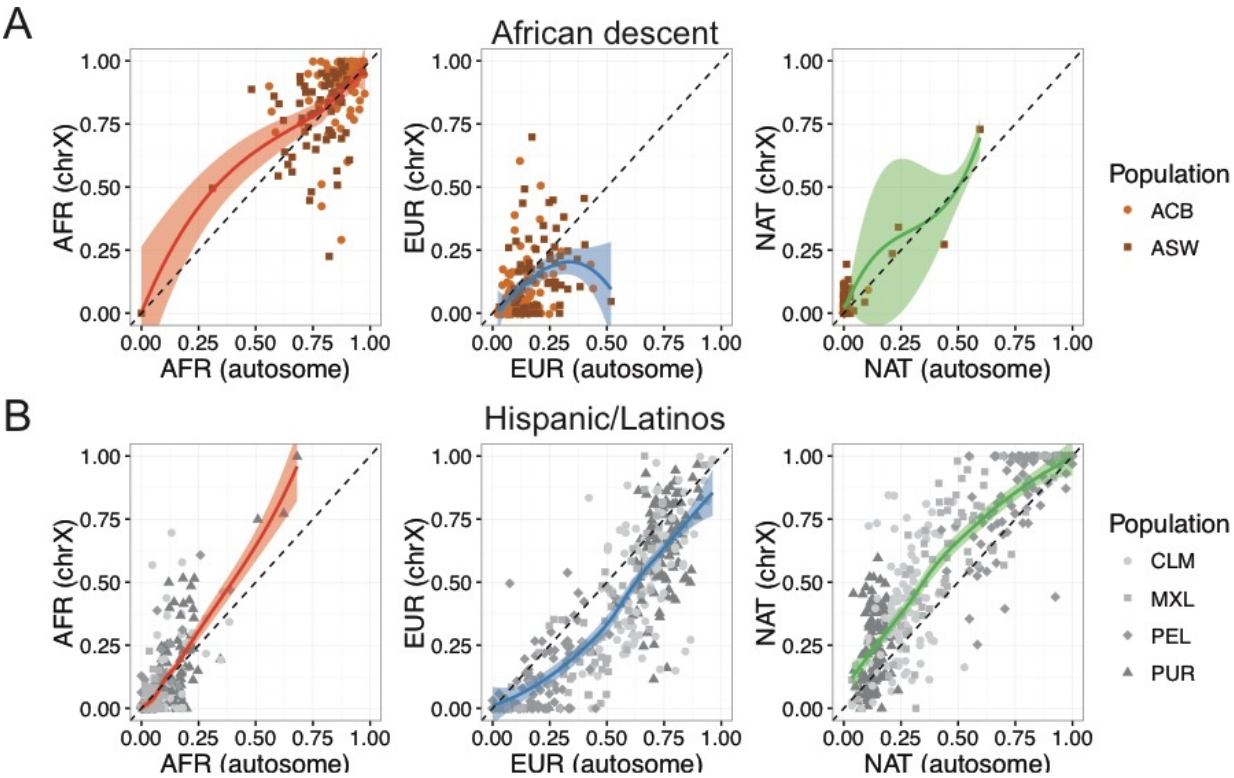


Figure S6 – Comparison of ploidy-adjusted ADMIXTURE ancestry estimates obtained on the autosomes and X chromosome at K=3 with CEU, YRI, and NAT[9] reference samples. 700,093 SNPs on the autosomes and 10,503 SNPs on the X chromosome were used to infer ancestry proportions. A) African descent and B) Hispanic/Latino samples. Sex-biased admixture has previously been shown to be ubiquitous in the Americas, impacting phenotypes strongly correlated with ancestry, such as pigmentation.[7,10-14] We inferred sex-biases in admixture events by separately querying ploidy-adjusted admixture proportions on the X chromosome versus the autosomes, as previously described[10]. We computed 3-way admixture proportions for AMR and AFR/AMR via ADMIXTURE[15] and consistently find across all six admixed AMR populations that the ratio of European ancestry is significantly depleted on the X chromosome compared to the autosomes, indicating a ubiquitous excess of breeding European males in the Americas, as seen previously[4,12,16]; there is also a significant excess of Native American ancestry ($p<1e-2$, Table S3) on the X chromosome in each of the AMR populations ($p < 1e-4$).

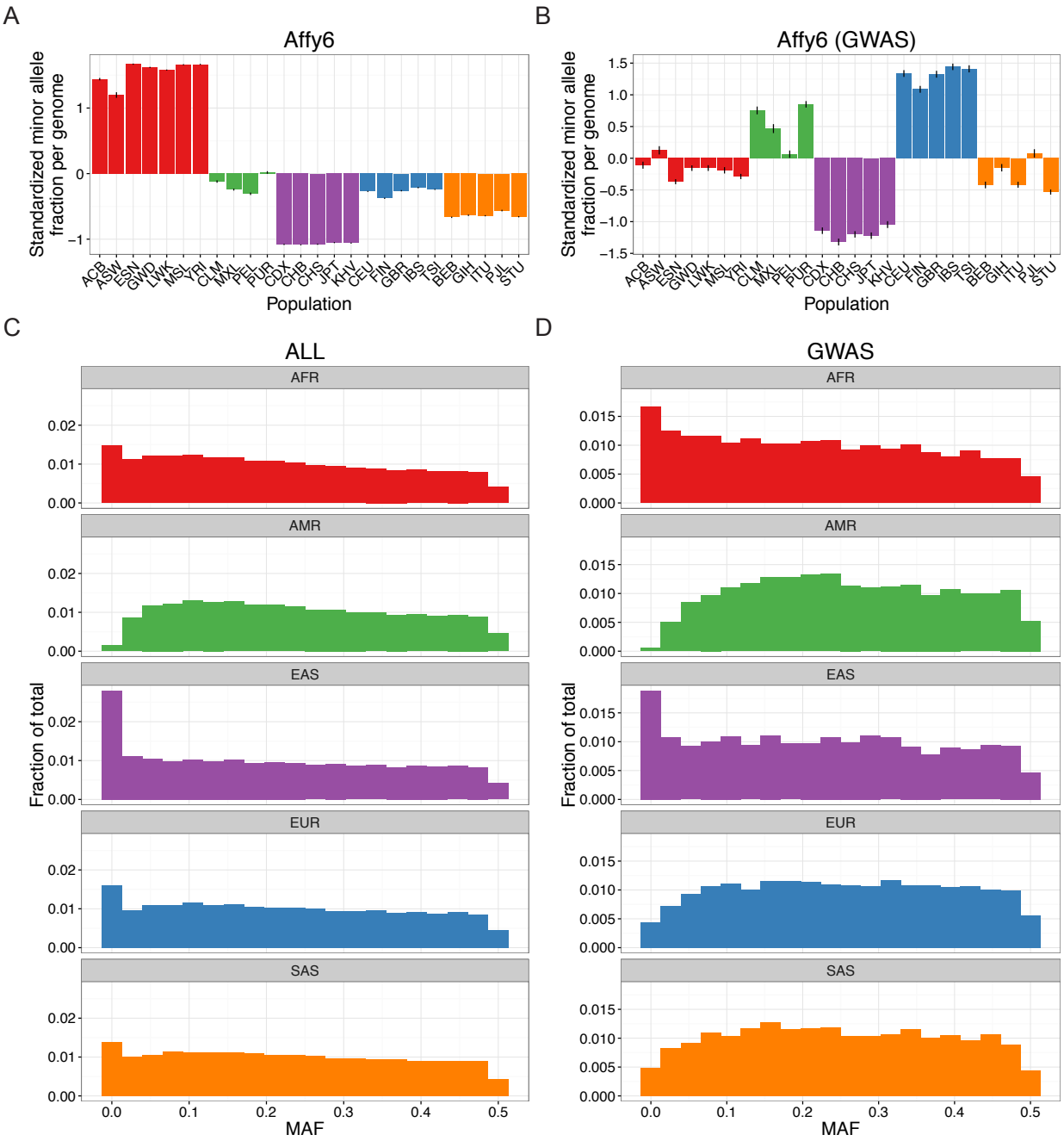


Figure S7 – Genetic variation and allele frequencies in global populations across all sites and at GWAS sites. A-B) GWAS study bias in European and American samples compared at all Affy6 sites from which local ancestry calls were made. All standardizations are computed as the ratio of minor alleles to total alleles per population minus the mean ratio across all individuals, then all divided by the standard deviation of this ratio. Error bars shows the standard error of the mean. A) Standardized across all Affy6 sites. B) Standardized across the intersection of Affy6 sites and the GWAS catalog. C-D) Allele frequencies within all super populations. Minor allele frequency fraction across C) all sites Affy6 sites, and D) the intersection of all Affy6 and GWAS catalog sites.

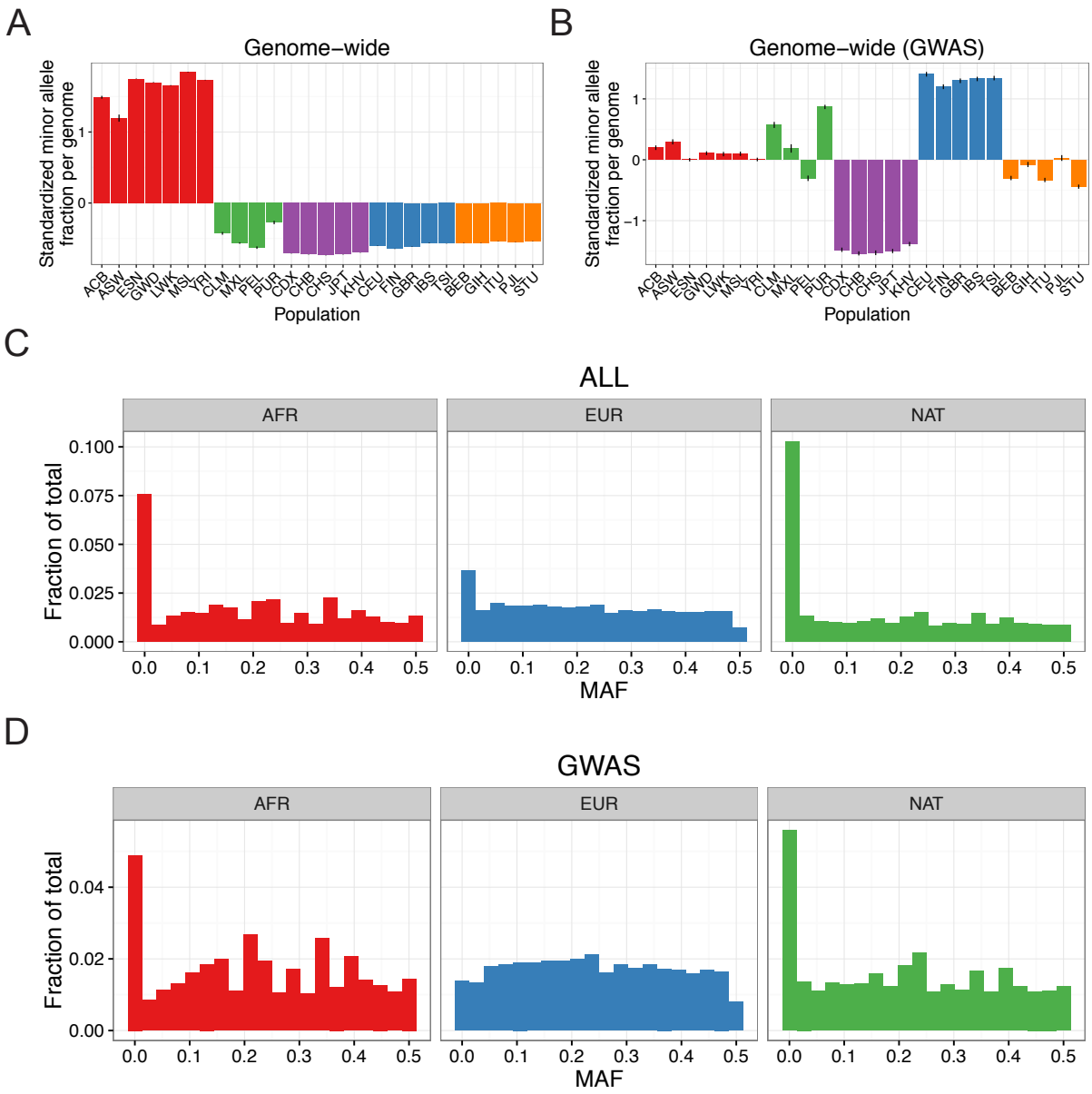


Figure S8 – Genetic variation in global and admixed populations across all sites and at GWAS sites. A-B) GWAS study bias in European and American samples compared to genomic background. All standardizations are computed as the ratio of minor alleles to total alleles per population minus the mean ratio across all individuals from all populations, then all divided by the standard deviation of this ratio. Error bars shows the standard error of the mean. A) Standardized across the whole genome. B) Standardized across all sites from the GWAS catalog. C-D) Allele frequencies in local ancestry calls from admixed AMR and AFR/AMR samples are specifically enriched on European tracts and depleted on African and Native American tracts across all genotyped sites and specifically at GWAS sites. Minor allele frequency fraction across C) all sites in admixed AFR/AMR and AMR populations stratified by local ancestry tracts, and D) sites from the GWAS catalog in admixed AFR/AMR and AMR populations stratified by local ancestry tracts.

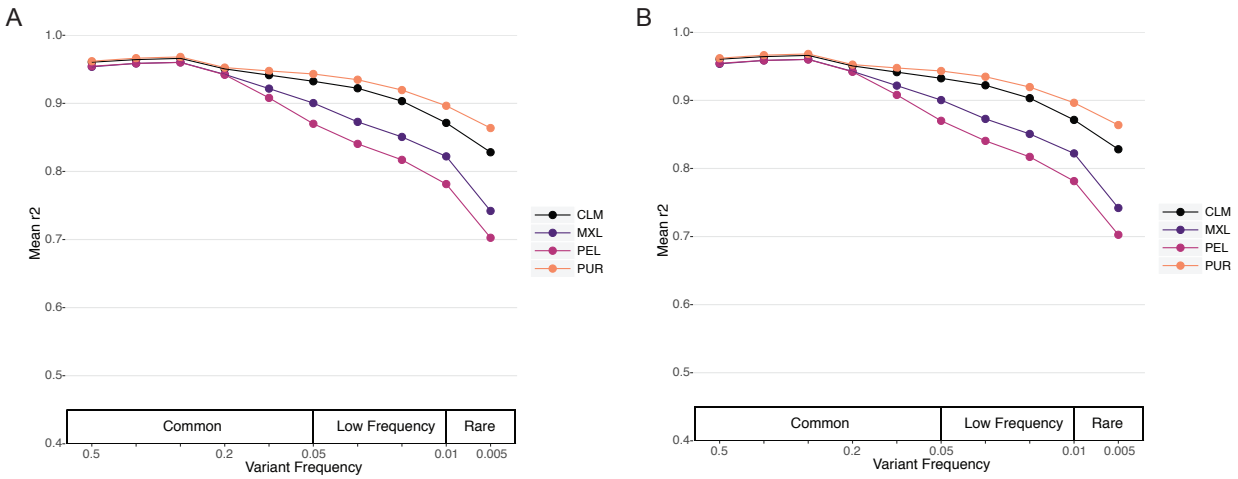


Figure S9 – Imputation accuracy by population for chromosome 9. A) Illumina OmniExpress. B) Affymetrix Axiom World Array LAT

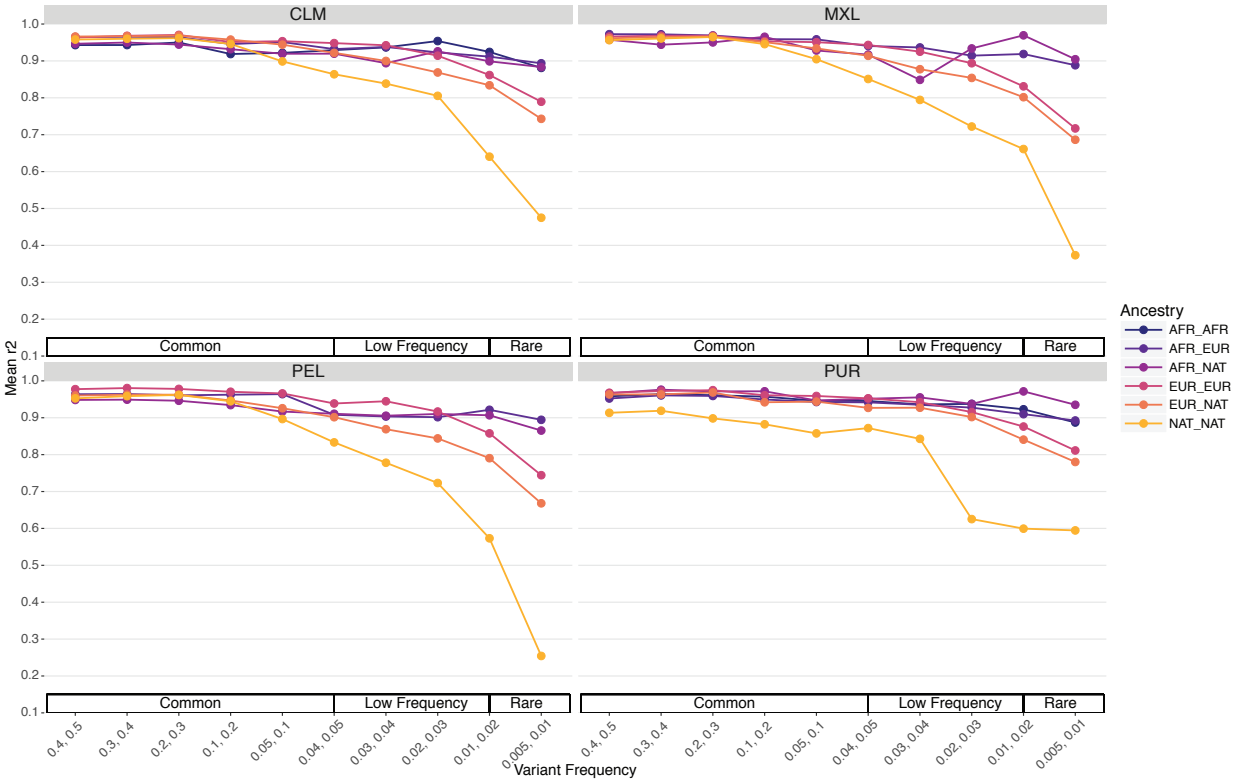


Figure S10 – Imputation accuracy by population assessed using a leave-one-out strategy, stratified by diploid local ancestry on chromosome 9 for the Affymetrix Axiom World Array LAT genotyping array.

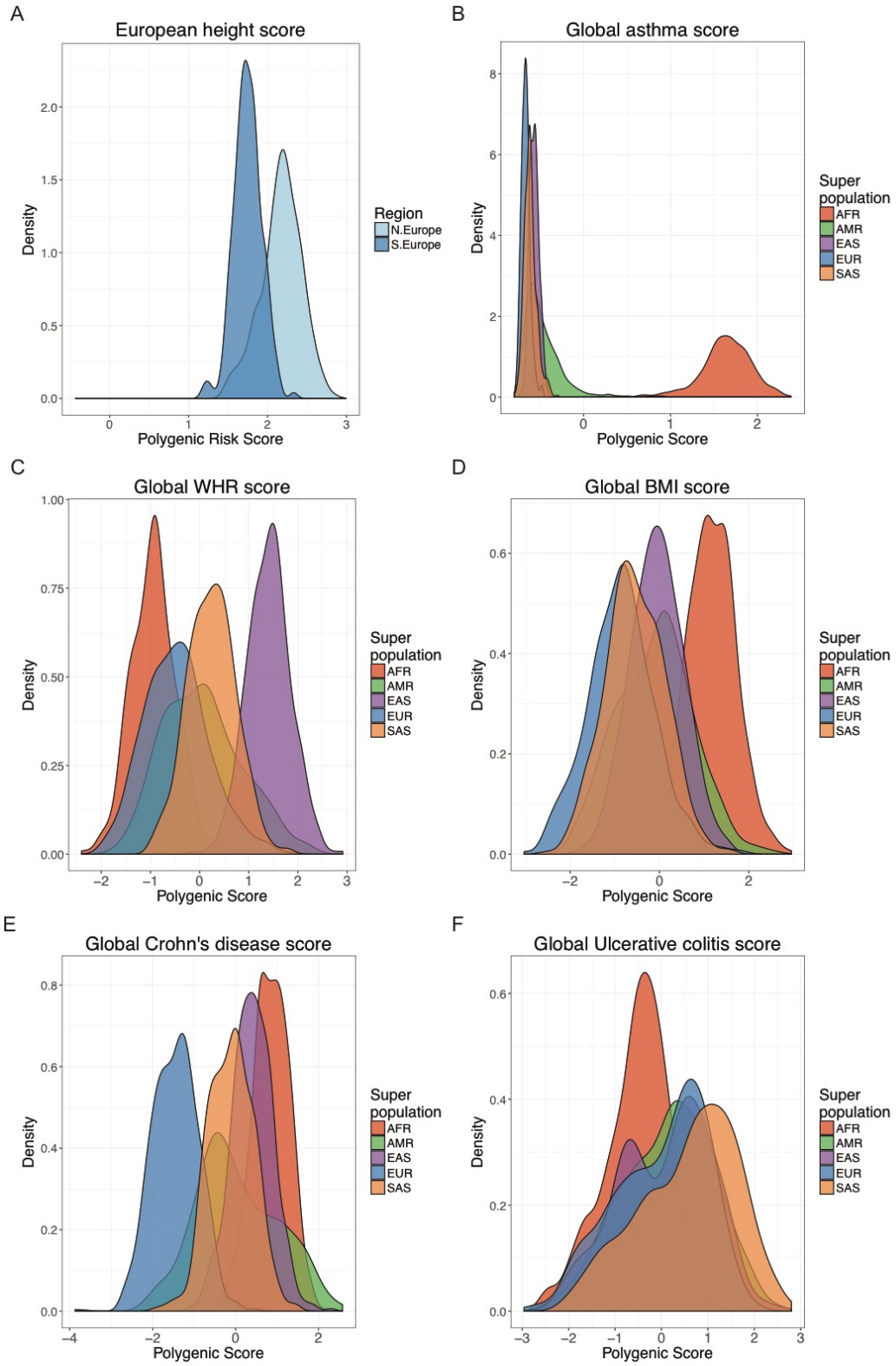


Figure S11 – Standardized polygenic risk score distributions for: A) northern/southern European height, B) asthma, C) waist-hip ratio, D) body mass index, E) Crohn's disease, and F) ulcerative colitis.

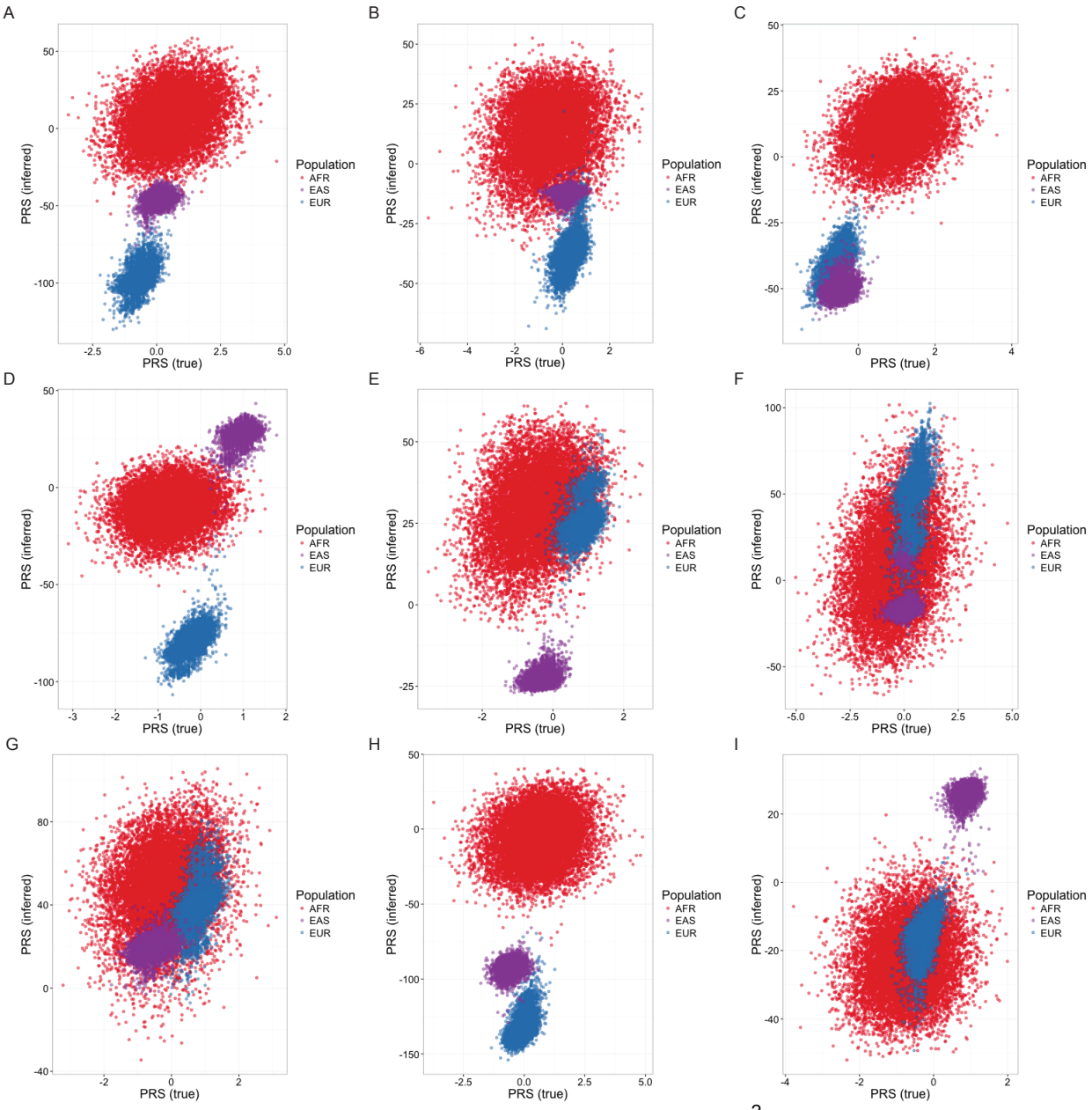
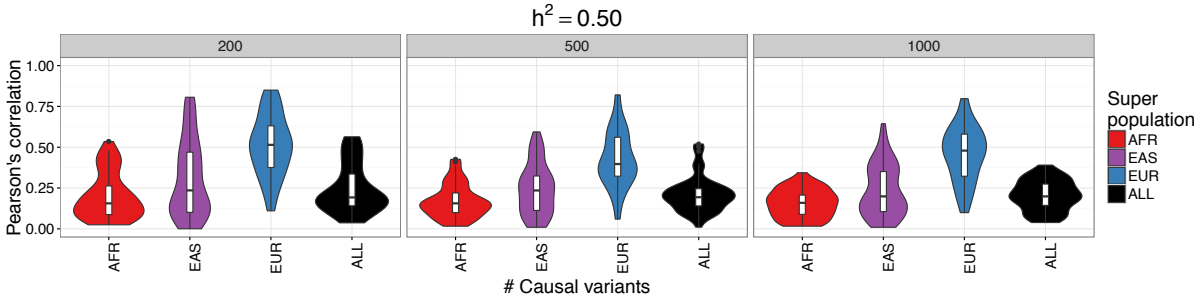


Figure S12 – Simulation runs for the same parameter set (h^2=0.67, m=1000) and same causal variants with varying effect sizes resulting in a wide range of possible biases in inferred polygenic risk scores across populations.

A



B

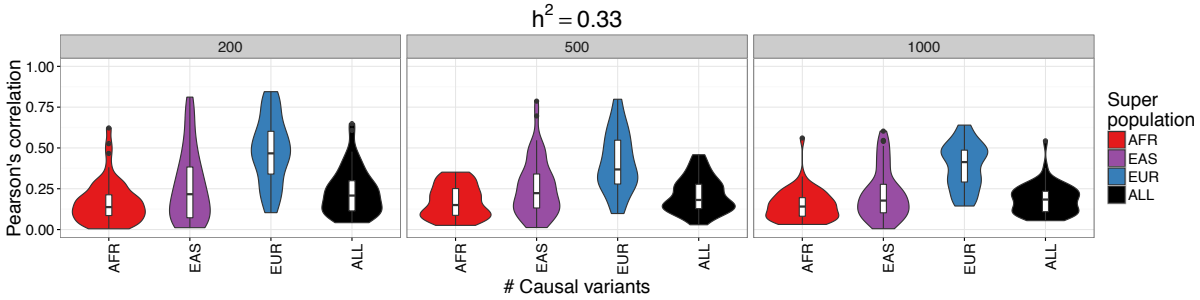


Figure S13 - Violin plots show Pearson's correlation across 50 iterations per parameter set between true and inferred polygenic risk scores across differing genetic architectures, including *m*=200, 500, and 1,000 causal variants and h^2=0.67, as in Figure 5. The "ALL" population correlations were performed on population mean-centered true and inferred polygenic risk scores.

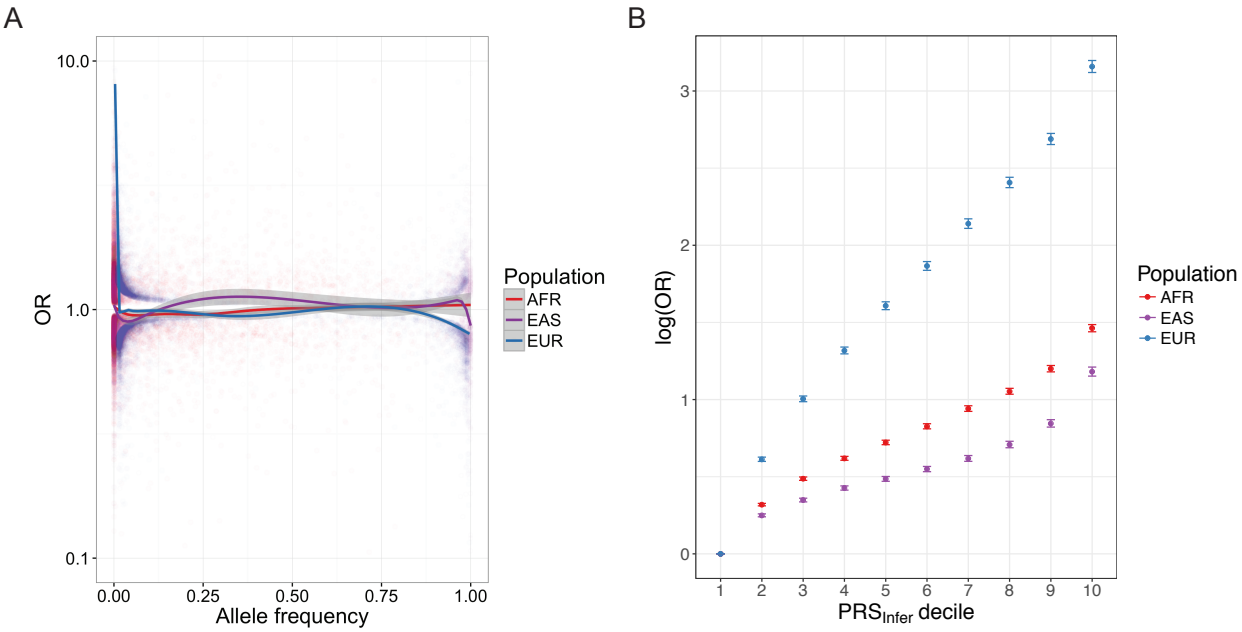


Figure S14 – **Genetic risk prediction differences across populations.** A) Allele frequency versus inferred odds ratio for sites included in inferred polygenic risk scores for each population across 500 simulations, as in Figure 5A-B. B) Log odds ratio by inferred polygenic risk score. The 10,000 individuals with the highest total liability per population were designated as cases, and 10,000 random other individuals in the population were designated as controls. The polygenic risk scores were converted to ordinal deciles, and contrasted with the 1st decile through logistic regression, as: case/control status predicted by polygenic risk decile and population label. Error bars indicate the standard error of the mean across 500 replicates with h^2=0.67 and m=1000 causal variants, as in Figure 5A-B. Small divergence from the population trend of prediction accuracy across the AFR and EAS are driven by differences in population-specific heritabilities arising from different numbers of population-private causal alleles (i.e. more AFR variants in general gives rise to more AFR-specific causal variants).

Table S1 – Population names and abbreviations

Population	Code	Super population	N
Esan in Nigeria	ESN	AFR	99
Gambian in Western Division, Mandinka	GWD	AFR	113
Luhya in Webuye, Kenya	LWK	AFR	99
Mende in Sierra Leone	MSL	AFR	85
Yoruba in Ibadan, Nigeria	YRI	AFR	108
African Caribbean in Barbados	ACB	AFR/AMR	96
People with African Ancestry in Southwest USA	ASW	AFR/AMR	61
Colombians in Medellin, Colombia	CLM	AMR	94
People with Mexican Ancestry in Los Angeles, CA, USA	MXL	AMR	64
Peruvians in Lima, Peru	PEL	AMR	85
Puerto Ricans in Puerto Rico	PUR	AMR	104
Chinese Dai in Xishuangbanna, China	CDX	EAS	93
Han Chinese in Beijing, China	CDX	EAS	103
Southern Han Chinese	CHS	EAS	105
Japanese in Tokyo, Japan	JPT	EAS	104
Kinh in Ho Chi Minh City, Vietnam	KHV	EAS	99
Utah residents (CEPH) with Northern and Western European ancestry	CEU	EUR	99
British in England and Scotland	GBR	EUR	91
Finnish in Finland	FIN	EUR	99
Iberian Populations in Spain	IBS	EUR	107
Toscani in Italia	TSI	EUR	107
Bengali in Bangladesh	BEB	SAS	86
Gujarati Indians in Houston, TX, USA	GIH	SAS	103
Indian Telugu in the UK	ITU	SAS	102
Punjabi in Lahore, Pakistan	PJL	SAS	96
Sri Lankan Tamil in the UK	STU	SAS	102

Table S2 – Three-way admixture proportions between recently admixed populations in the Americas. Values are computed at K=3 on common autosomal SNPs using ADMIXTURE with mean percentages ± standard deviations.

	AFR	EUR	NAT
ACB	88.0% (7.7%)	11.7% (7.3%)	0.3% (1.1%)
ASW	75.6% (13.8%)	21.3% (9.1%)	3.1% (9.2%)
CLM	7.8% (13.8%)	66.6% (12.8%)	25.7% (9.3%)
MXL	4.3% (2.2%)	48.7% (18.6%)	47.0% (19.1%)
PEL	2.5% (5.4%)	20.2% (12.0%)	77.3% (14.2%)
PUR	13.9% (5.4%)	73.2% (10.0%)	12.9% (3.6%)

Table S3 – Comparison of mean ancestry proportions and ratio on chromosome X versus autosomes across populations. Per Lind et al[10], proportion X in a population = (fraction male + 2*fraction female) / 1.5, and proportion autosome in a population = fraction male + fraction female. P-values are from two-sided t-tests on individual ancestries (comparisons are not independent as ancestry proportions must sum to one).

	Ancestry	ACB	ASW	CLM	MXL	PEL	PUR
Relative X/autosome % change	AFR	4.01	0.83	-2.02	-20.32	50.75	12.69
	EUR	-41.73	-17.41	-20.20	-26.60	-41.51	-14.51
	NAT	558.04	87.41	52.70	28.49	9.37	66.89
p-value	AFR	8.9e-2	7.7e-1	9.8e-1	6.8e-2	3.5e-1	4.1e-1
	EUR	1.0e-3	8.9e-2	1.4e-7	7.9e-4	4.5e-6	1.5e-7
	NAT	7.2e-9	1.1e-1	4.0e-9	3.9e-4	1.3e-3	1.4e-10

Table S4 – Empirical polygenic risk score details. OR = odds ratio

Trait	Reference	Effect size	Number of clumps with p≤1e-2
Height	Wood et al, 2014	Beta	35,194
Female WHR	Shungin et al, 2015	Beta	7,351
T2D, EUR	Gaulton et al, 2015	Log(OR)	515
T2D, Multi-ethnic	Mahajan et al, 2014	Log(OR)	11,577
Asthma	Moffatt et al, 2010	Log(OR)	4,786
Schizophrenia	Ripke et al, 2014	Log(OR)	22,047
BMI	Locke et al, 2015	Beta	9,445
Crohn's disease	Jostins et al, 2012	Log(OR)	19,637
Ulcerative colitis	Jostins et al, 2012	Log(OR)	19,078

References

1. Basu A, Sarkar-Roy N, Majumder PP (2016) Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proc Natl Acad Sci U S A 113:1594-1599
2. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. Nature 461:489-494
3. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al (2008) Genes mirror geography within Europe. Nature 456:98-101
4. Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, et al (2016) The Great Migration and African-American Genomic Diversity. PLoS genetics 12:e1006059
5. Mimno D, Blei DM, Engelhardt BE (2015) Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. Proc Natl Acad Sci U S A 112:E3441-E3450
6. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, et al (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nature Genetics
7. Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, et al (2013) Reconstructing the Population Genetic History of the Caribbean. PLoS Genetics 9:e1003925
8. Gravel S (2012) Population genetics models of local ancestry. Genetics 191:607-619
9. Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, et al (2007) A genomewide admixture mapping panel for Hispanic/Latino populations. American journal of human genetics 80:1171-1178
10. Lind JM, Hutcheson-Dilks HB, Williams SM, Moore JH, Essex M, Ruiz-Pesini E, et al (2007) Elevated male European and female African contributions to the genomes of African American individuals. Human Genetics 120:713-722
11. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, et al (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proceedings of the National Academy of Sciences of the United States of America 107:786-791
12. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL (2015) The genetic ancestry of African Americans, Latinos, and European Americans across the United States. American Journal of Human Genetics 96:37-53
13. Beleza S, Campos J, Lopes J, Araújo II, Hoppfer Almada A, Correia e Silva A, et al (2012) The Admixture Structure and Genetic Variation of the Archipelago of Cape Verde and Its Implications for Admixture Mapping Studies. PLoS ONE 7:1-12
14. Marcheco-Teruel B, Parra EJ, Fuentes-Smith E, Salas A, Buttenschøn HN, Demontis D, et al (2014) Cuba: exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers. PLoS genetics 10:e1004488
15. Shringarpure SS, Bustamante CD, Lange KL, Alexander DH (2016) Efficient analysis of large datasets and sex bias with ADMIXTURE. BMC Bioinformatics 17(1):218
16. McHugh C, Thornton TA, Brown L (2015) Detecting Heterogeneity in Population Structure Across the Genome in Admixed Populations. Genetics 204(1):43-56