Population structure, heritability, and polygenic risk

Alicia Martin
Daly Lab
October 18, 2016

armartin@broadinstitute.org

@genetisaur

Project goals

- ✓ Call local ancestry in large case/control PTSD cohort of African Americans

- Estimate heritability using local ancestry tracts. Compare/contrast this estimate with SNP-based heritability in this and European cohort (in progress)

- Perform admixture mapping

- Considerations: transferability of polygenic risk scores, cross-population heritability

(Work with Karestan Koenen, Mark Daly, Laramie Duncan, Caroline Nievergelt)

Data overview

	Study	PI	Analyst	N_{Total}	N_{AA}	Data label
1	GTP (Grady Trauma Project)	Kerry Ressler	Lynn Almli	4752	3492	gt2y
2	Detriot (DNHS)	Monica Uddin	Guia Guffanti	812	650	dnhy
3	Genetics of Substance Dependence	Goel Gelernter	Pingxing Xie	5451	3100	gsdy
4	Marine Resilience Study	Caroline Nievergelt / Dewleen Baker	Adam Maihofer	4036	226	mrsy
5	Family Study of Cocaine Dependence	Laura Bierut	Louis Fox	1271	653	fscy
6	COGEND	Laura Bierut	Louis Fox	2768	711	cogy
7	Nurses Health Study	Karestan Koenen	Andrew Ratanatharathorn	1378		
8	Stein South Africa	Dan Stein / Kerry Ressler	Lynn Almli	434		
9	Ohio National Guard	Israel Liberzon	Tony King	239		
Summary Statistics from imputed data						
10	Duke	J. Beckham / M. Hauser / A. Ashley-Koch	Melanie Garrett	1963		
11	National Center for PTSD (Boston)	Mark Miller / Mark Logue	Mark Logue	652		
	Total			**23,756**	**8,832**	

Local ancestry calling strategy

1. Merge intersecting genotyped SNPs (N=421,607 with MAF > 0.05)
2. Phase aggregated dataset with HAPI-UR 3x and take best combined phase
3. Split jointly phased haplotypes into reference + 50 sets of admixed samples for computational feasibility
4. Aggregate local ancestry calls across all runs
5. Collapse local ancestry output

Heritability estimates

h^2 estimate	Kinship matrix	$\hat{h^2}$	SE	N
h^2_g	REAP	0.018	0.046	7548
h^2_g	GCTA GRM	0.02	0.048	7248
h^2_γ	local ancestry GRM	?	?	

h^2_γ =phenotypic variation described by variation in local ancestry

σ^2_γ =phenotypic variation explained by variation in local ancestry

σ^2_e =residual phenotypic variance

$$h^2_\gamma = \frac{\sigma^2_\gamma}{\sigma^2_\gamma + \sigma^2_e}$$

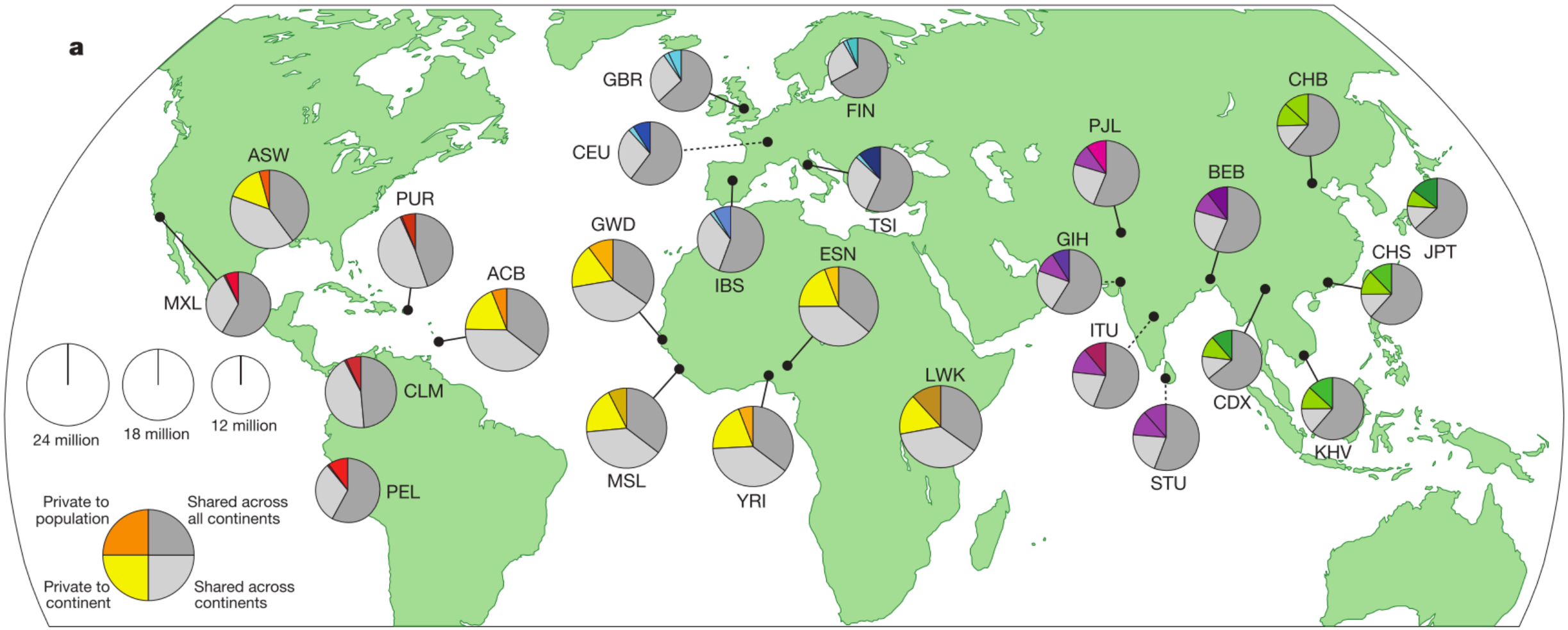
F_{STC} =weighted allele frequency differences between ancestral populations at causal loci

θ =genome-wide ancestry proportions

$$h^2_\gamma = 2F_{STC}\theta(1-\theta)h^2$$

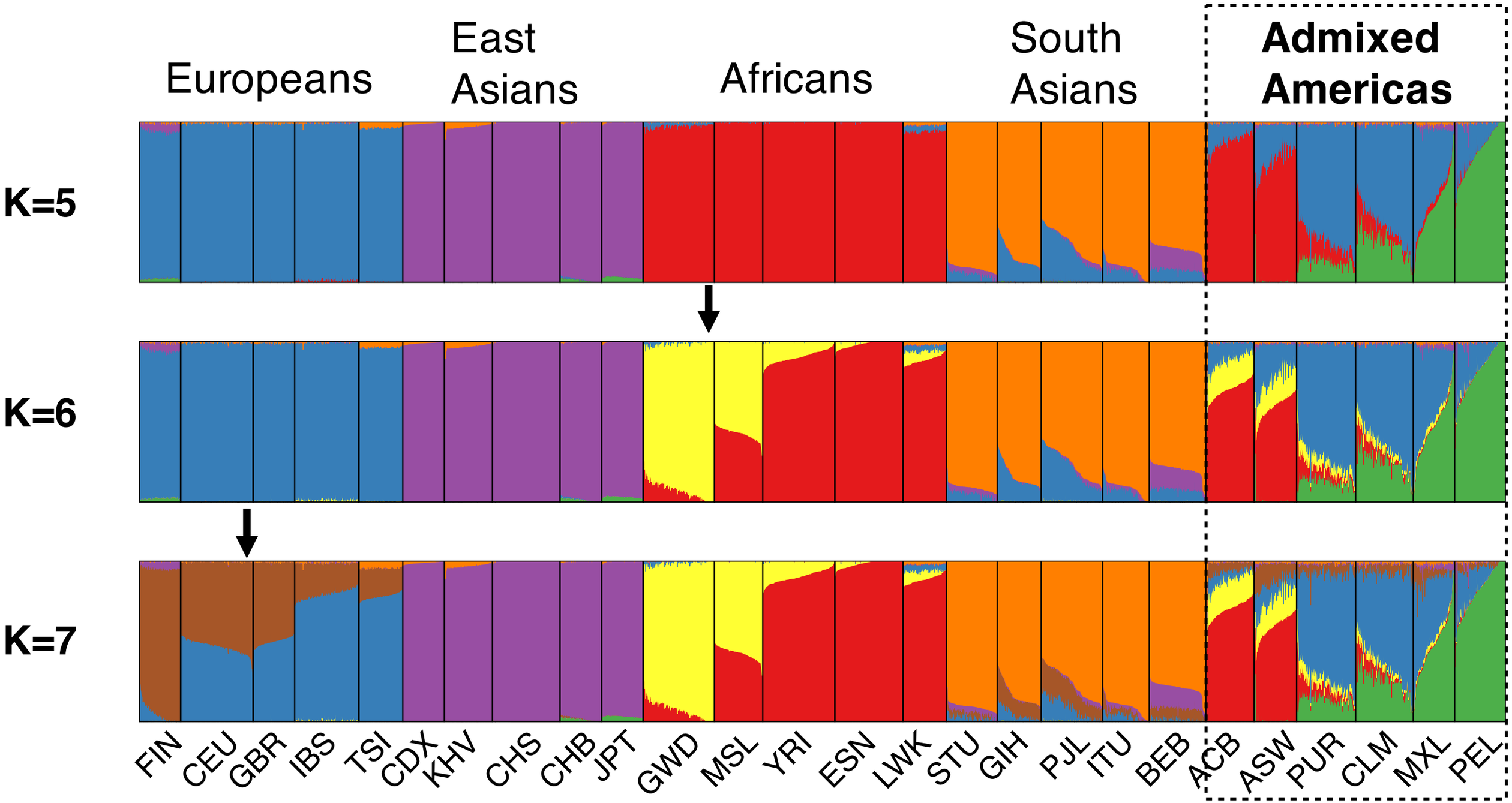
Zaitlen, N., et al. (2014). Nat. Genet. 46, 1356–1362.

1000 Genomes phase 3 populations

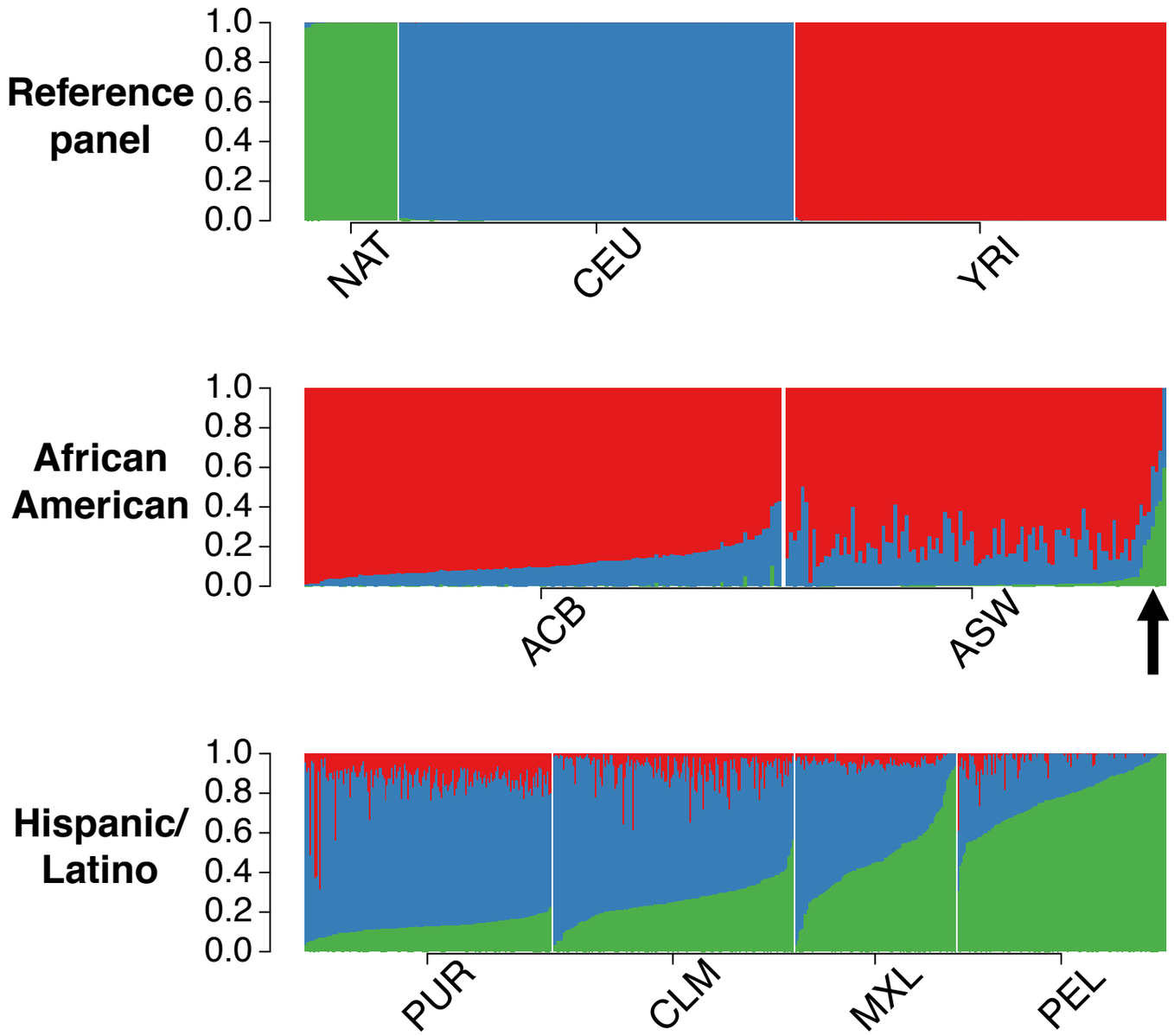


Auton, A., et al. (2015). Nature 526, 68–74.

Substantial global genetic diversity in 1000 Genomes



Varying admixture proportions across populations in the Americas



NAT = Mao et al, (2007). AJHG. *80*, 1171–1178.

African Americans

ACB = African Caribbean in Barbados

ASW = African Ancestry in SW US

Hispanic/Latinos

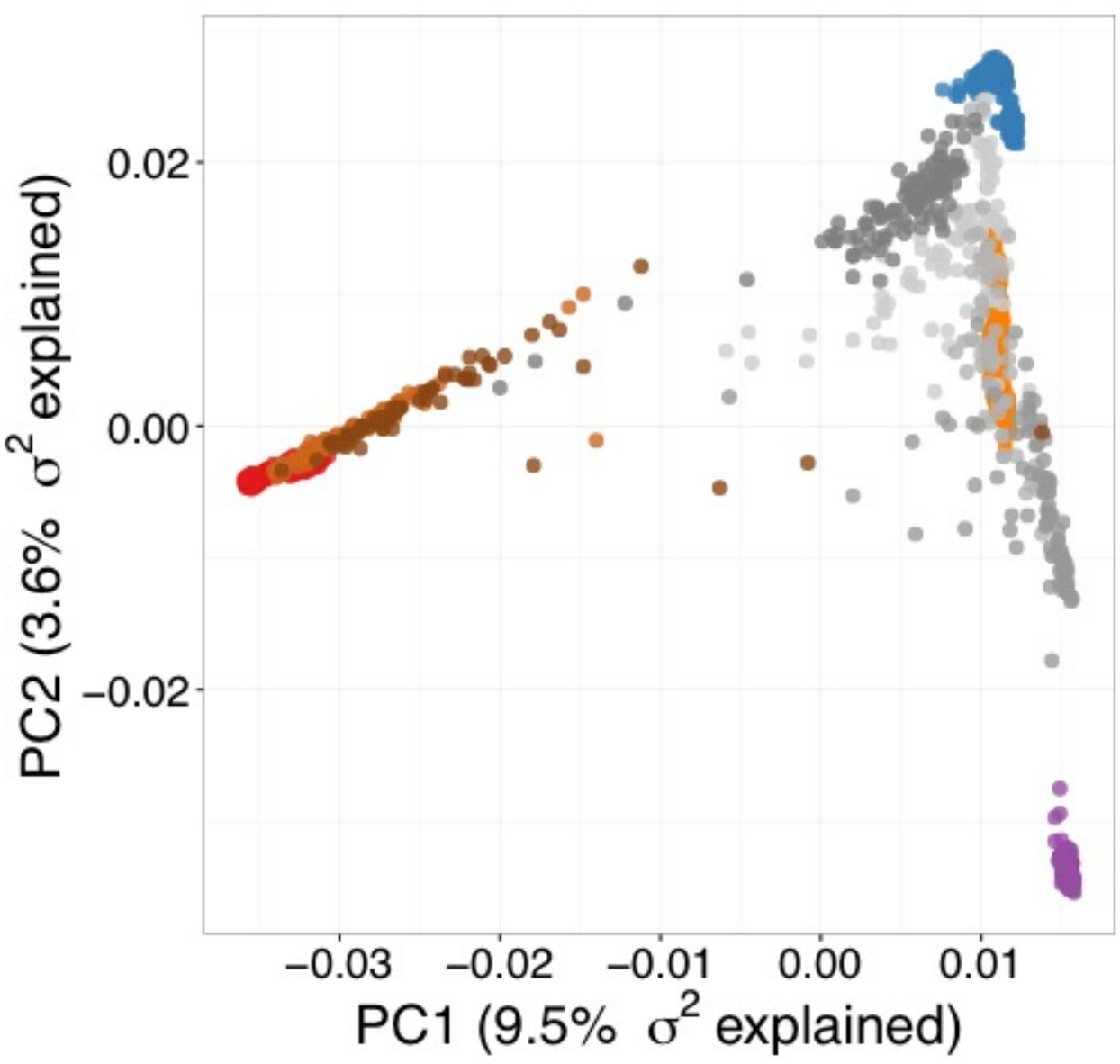
CLM = Colombians

MXL= Mexicans

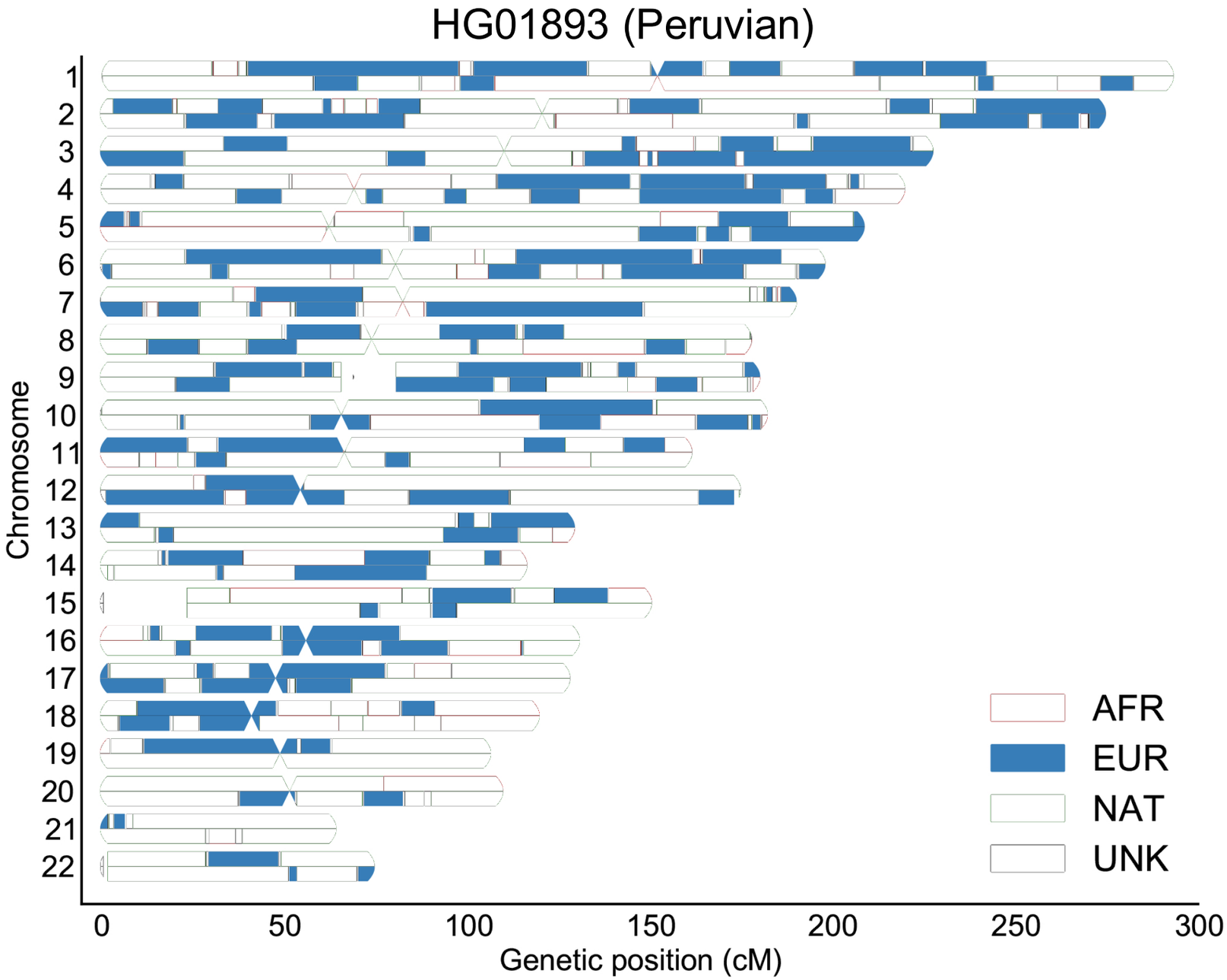
PUR = Puerto Ricans

PEL = Peruvians

Admixed samples in the Americas

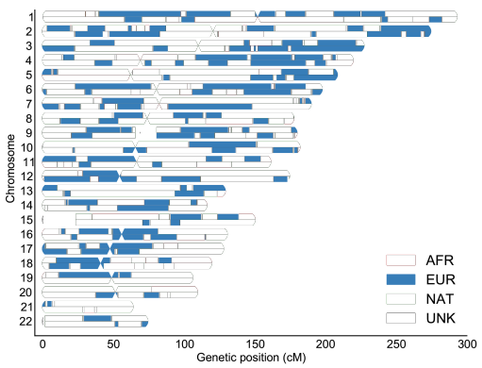


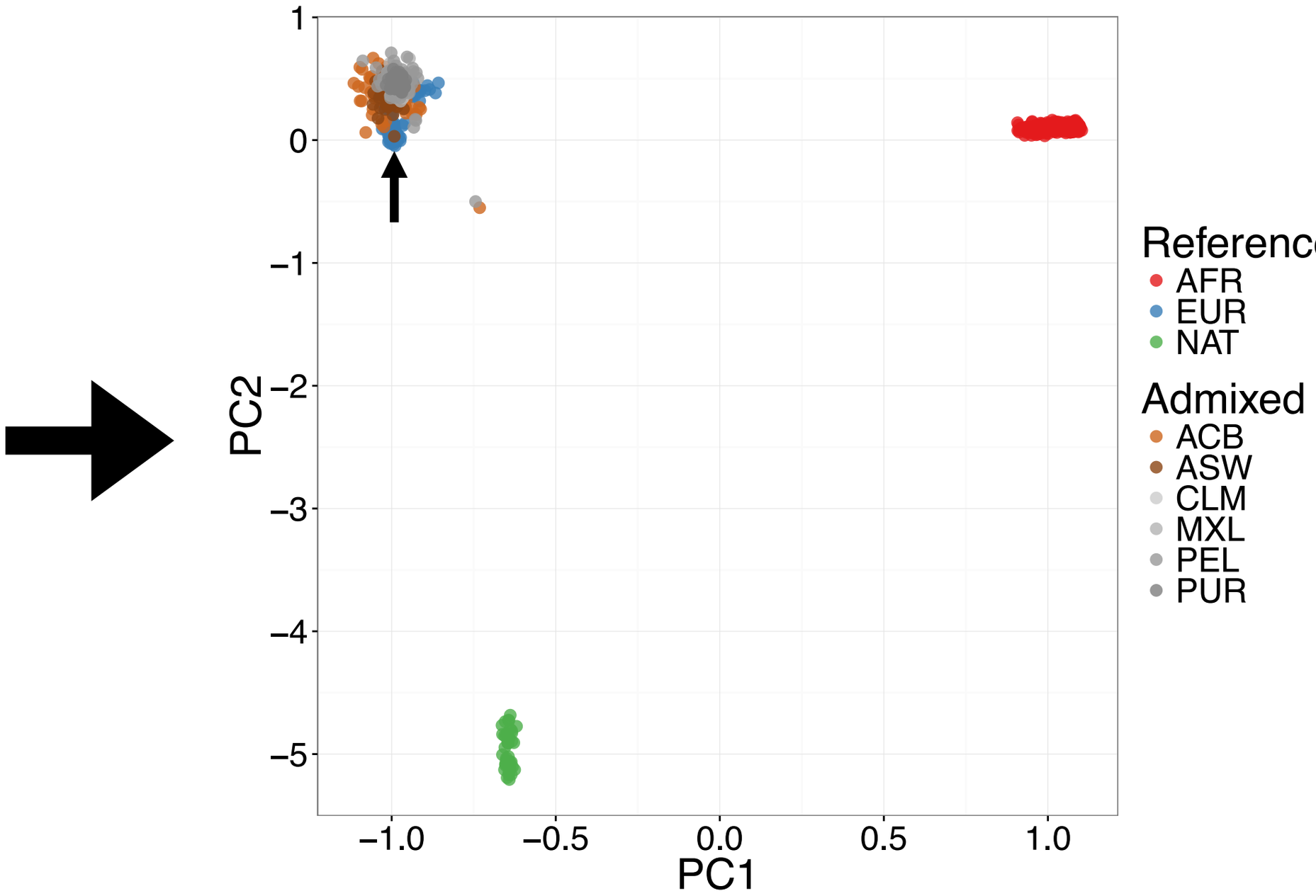
Admixture tracts inform subcontinental-level ancestral populations



RFMix: Maples, B.K., et al (2013). AJHG. 93, 278–288.

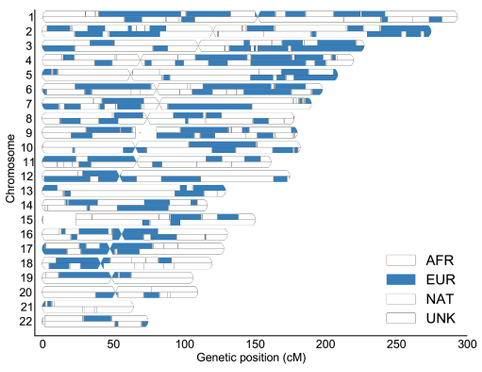
Ancestry-specific PCA provides insight into subcontinental admixture origins

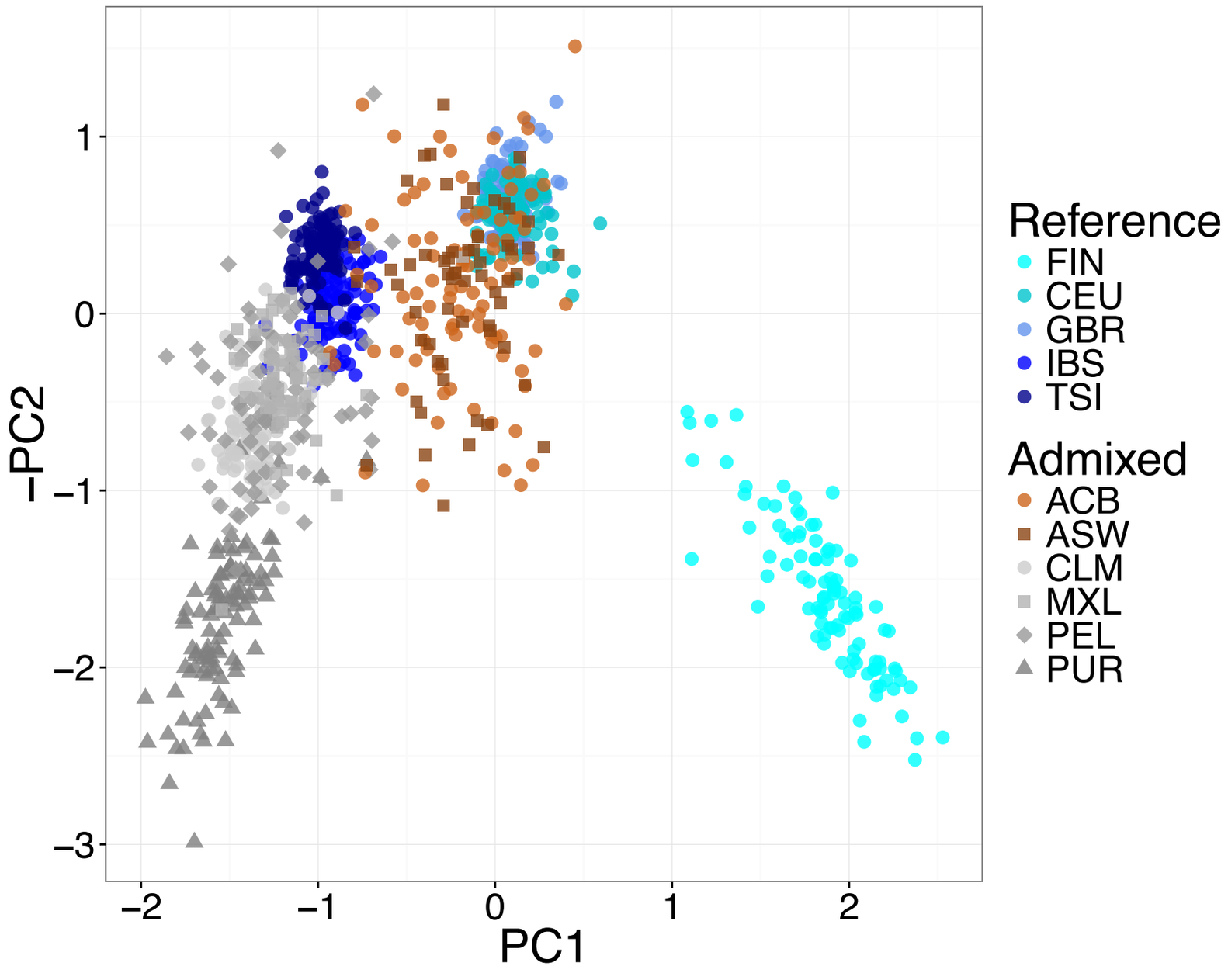




ASPCA: Moreno-Estrada, A., et al. (2013). PLoS Genetics. *9*, e1003925.

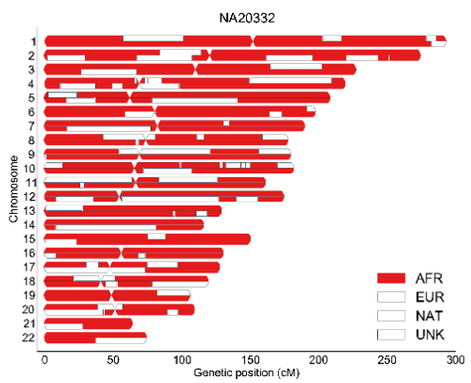
African Americans have northern European tracts, Hispanics have southern European tracts

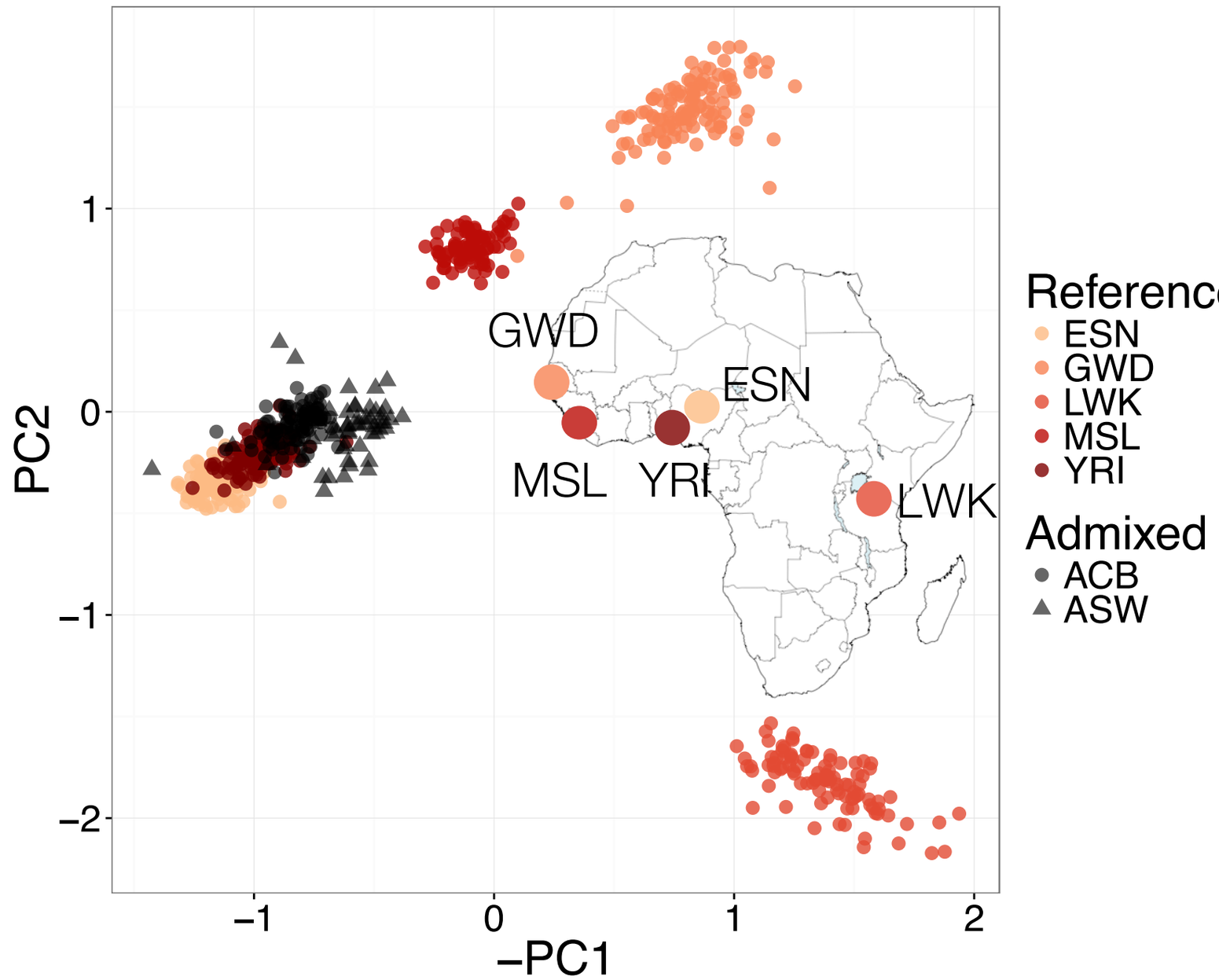




ASPCA: Moreno-Estrada, A., et al. (2013). PLoS Genetics. *9*, e1003925.

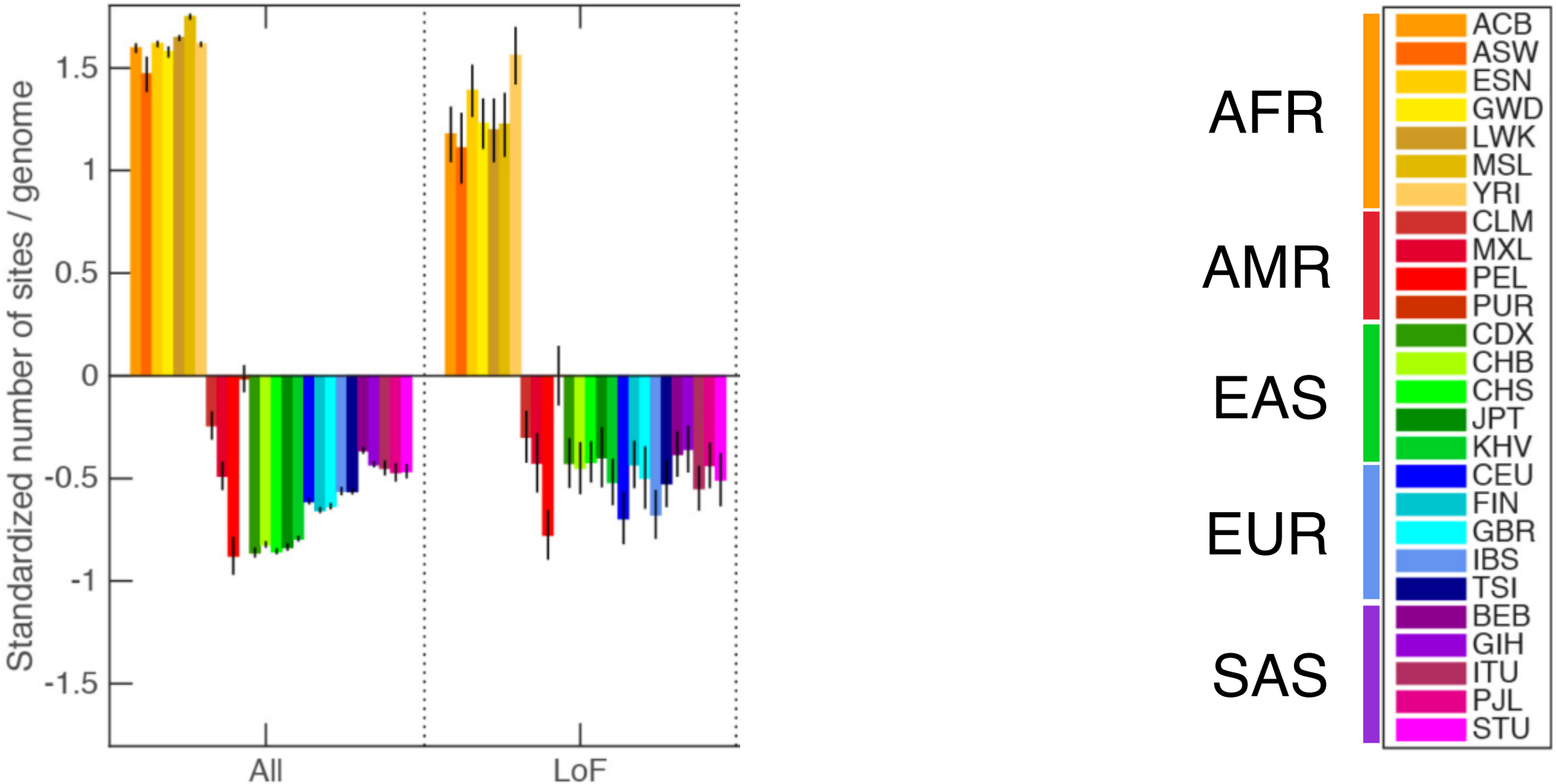
African Americans have African tracts closest to Nigerian reference panel





ASPCA: Moreno-Estrada, A., et al. (2013). PLoS Genetics. *9*, e1003925.

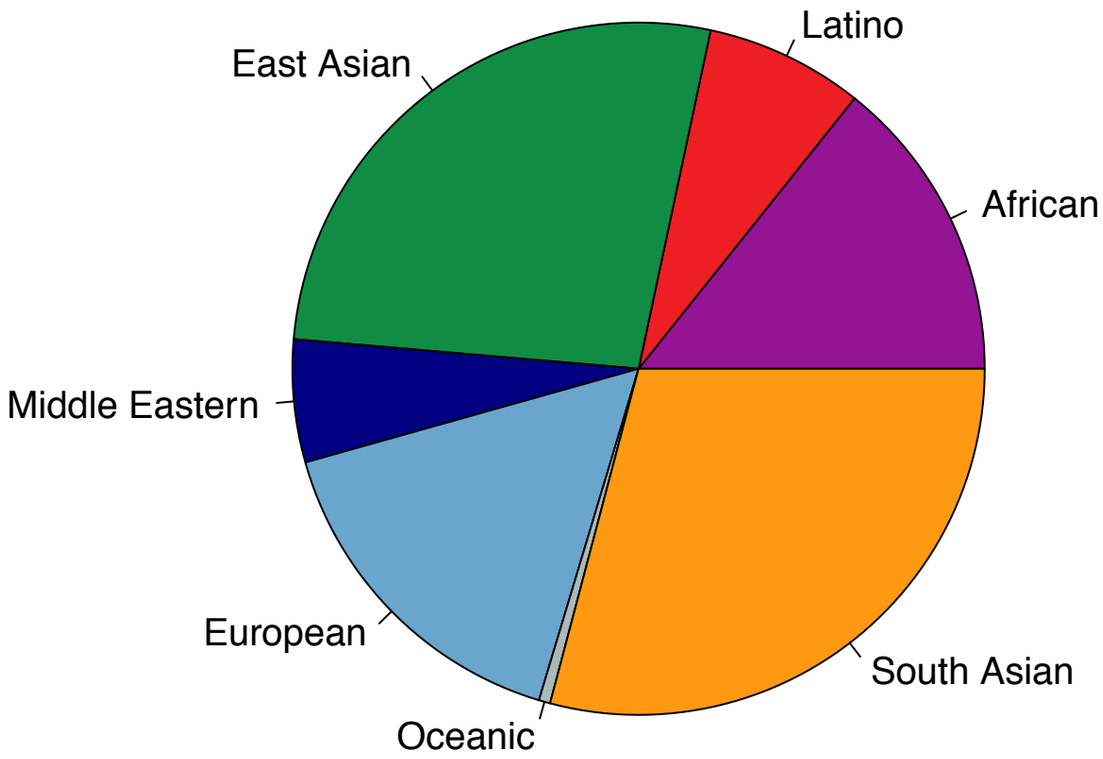
Africans have more genetic variation than out-of-Africa populations



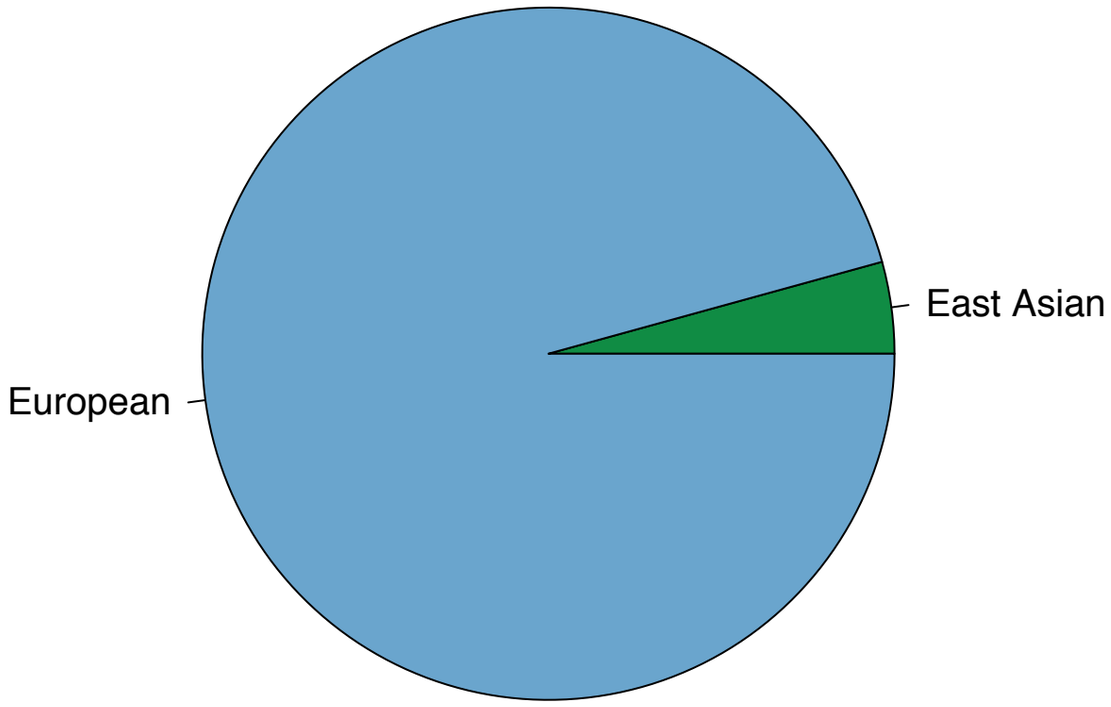
1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature 526, 68–74.

Biased genetic discoveries

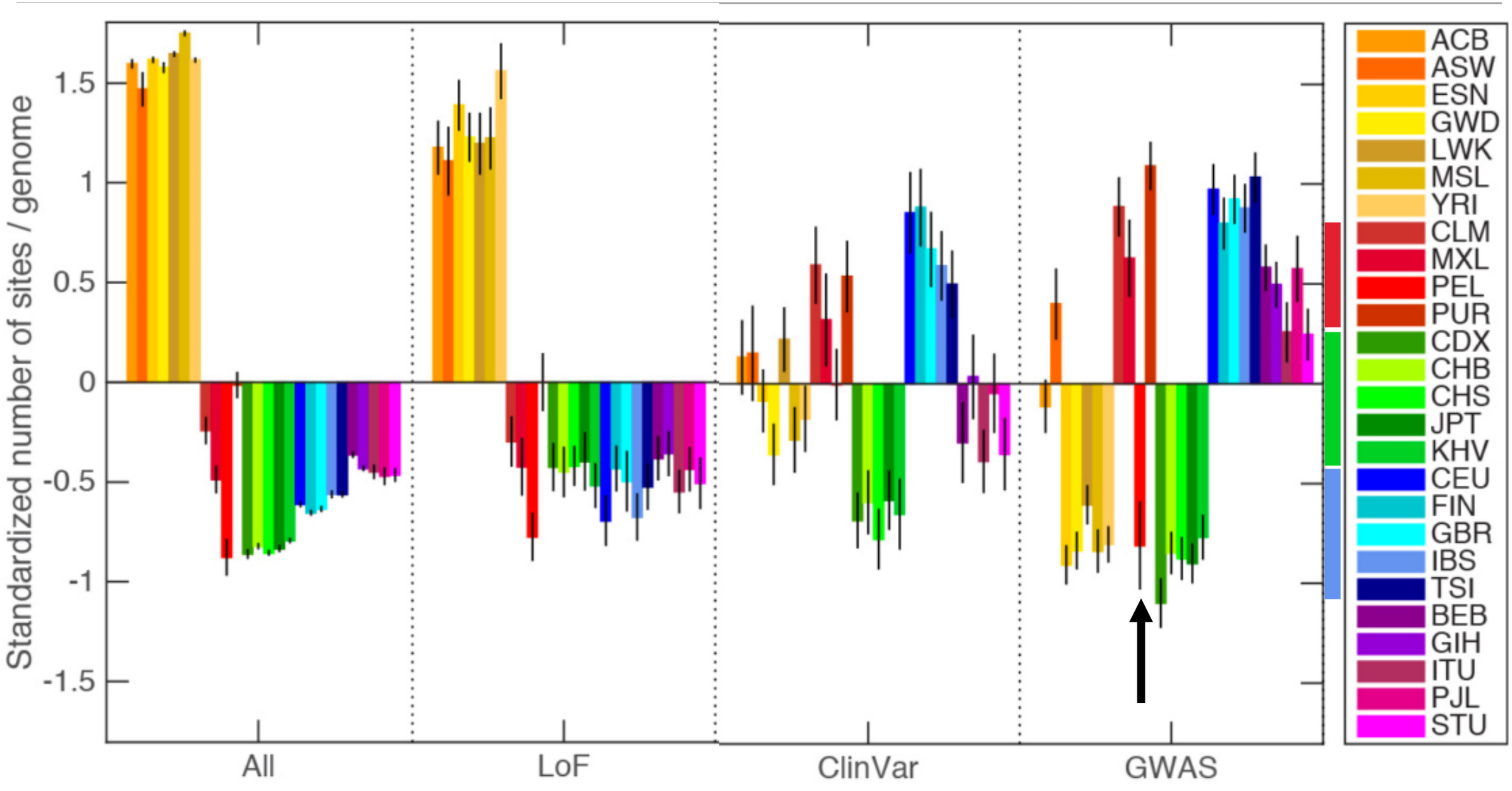
Global population



PGC GWAS
(SCZ, BIP, MDD, ADHD)



Europeans (~~and Hispanic/Latinos~~) are overrepresented in disease databases



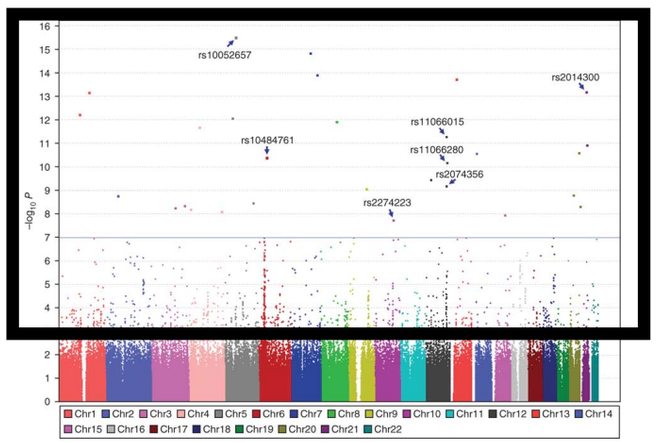
1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature 526, 68–74.

Computing polygenic risk scores from summary statistics

$$X = \sum_{i=1}^{m} g_i \beta_i$$

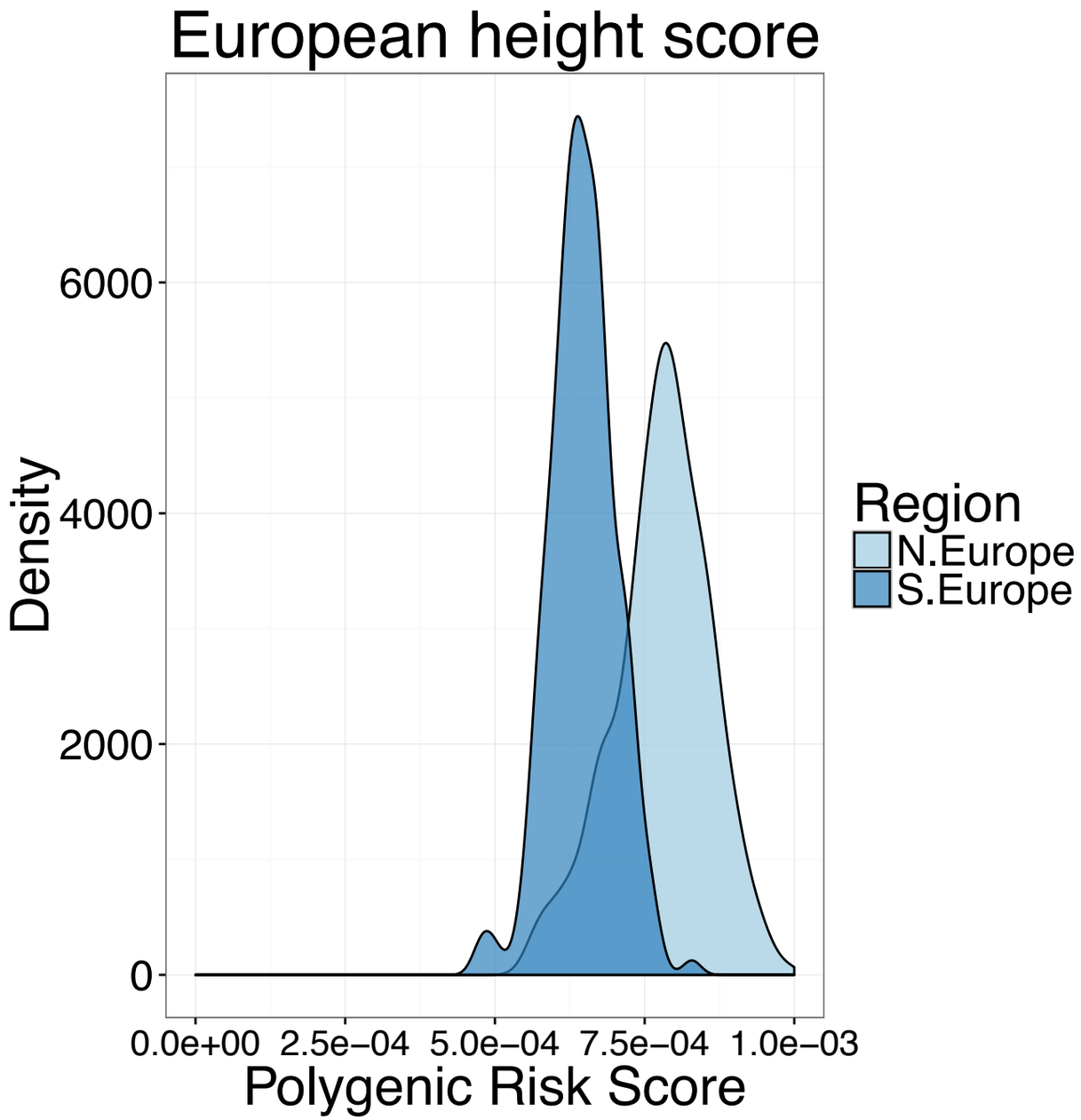
- LD clumping for all variants with MAF ≥ 0.01:
- Apply p-value threshold (p=0.01)
- Thin for LD within window (R^2=0.5, window=250kb)

Σ



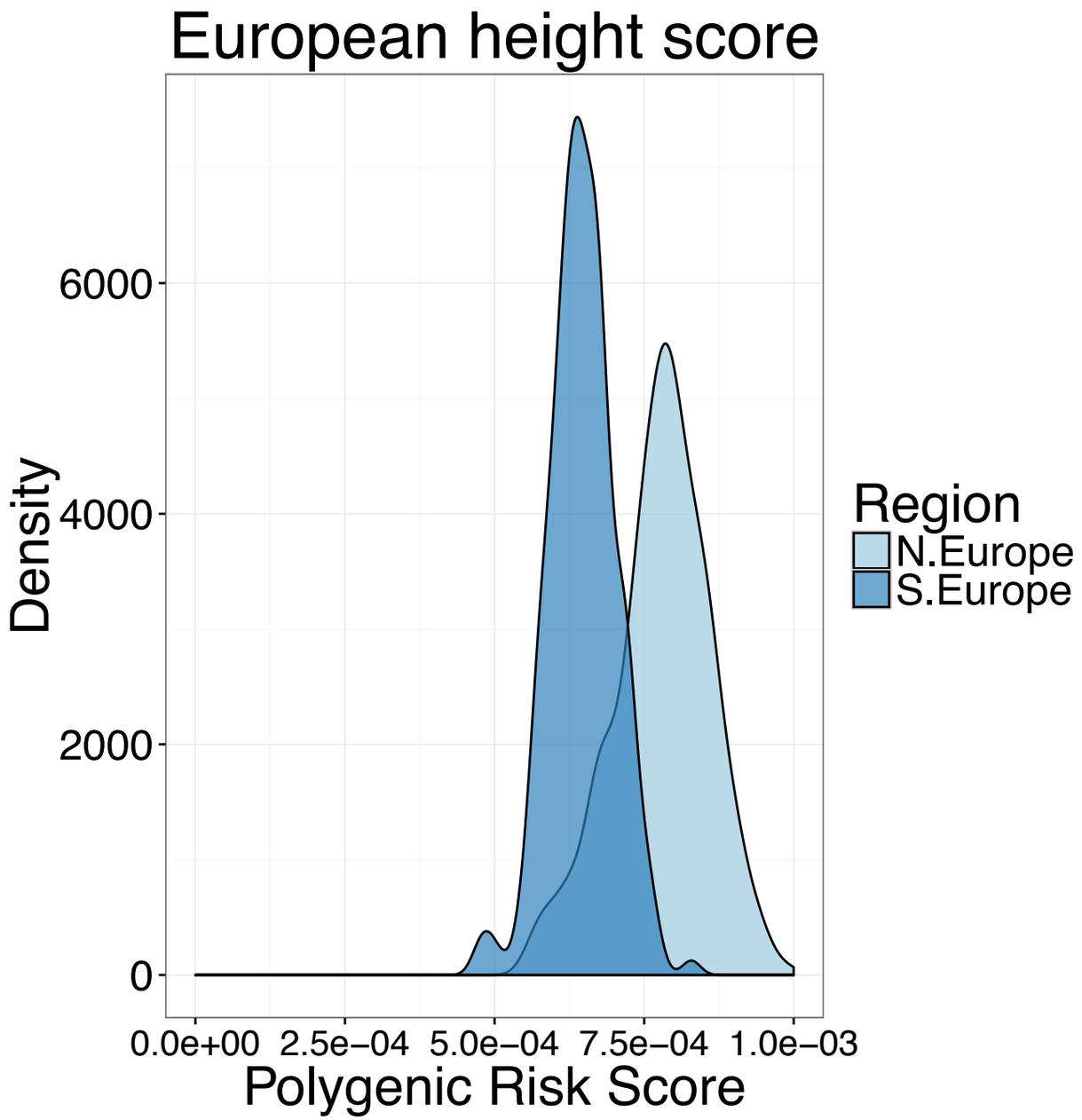
(P+T in LDpred paper)

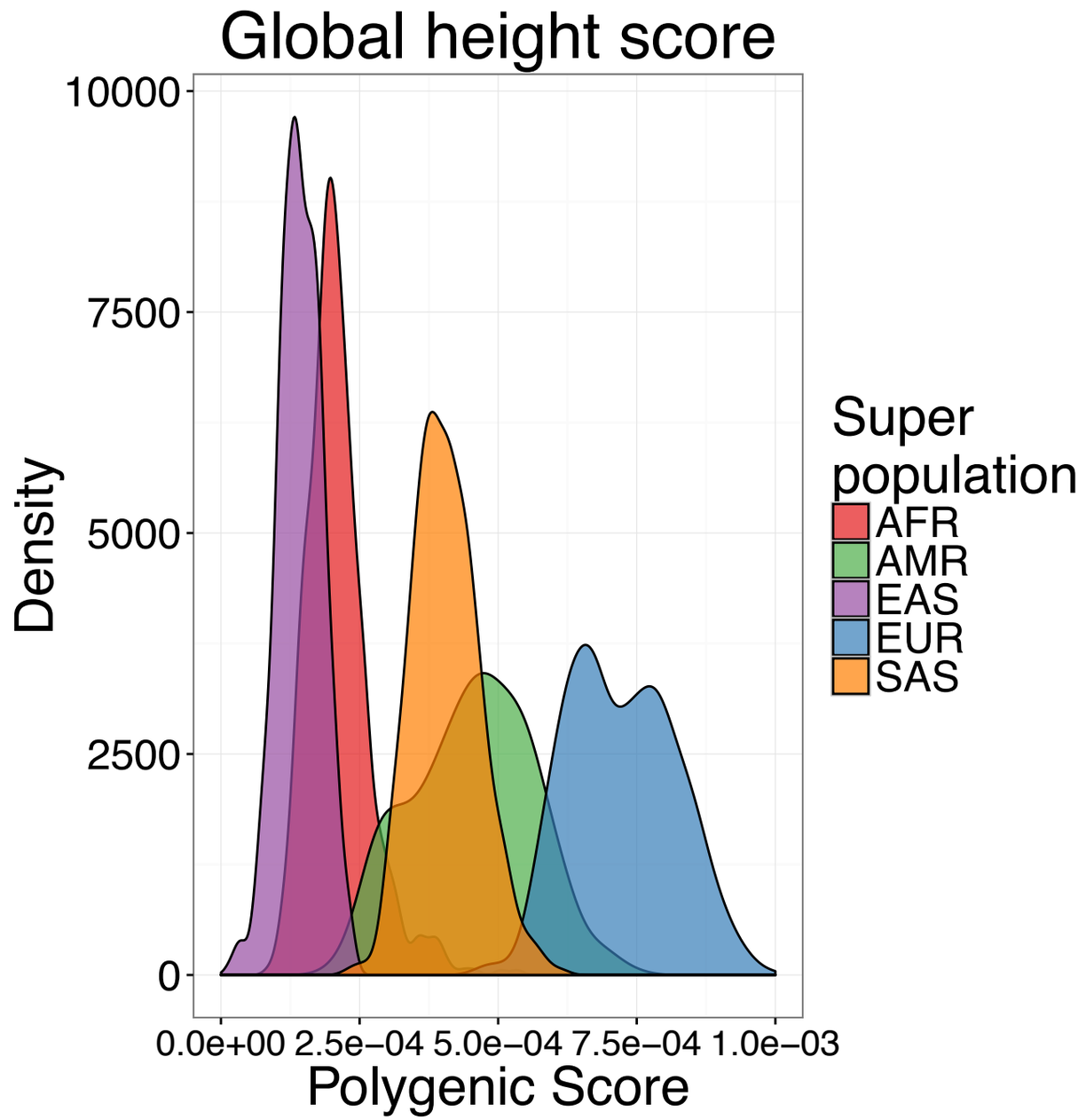
Polygenic risk score for height reflects adaptive event in Europeans



Wood, A.R., et al. (2014). Nature Genetics 46, 1173–1186.

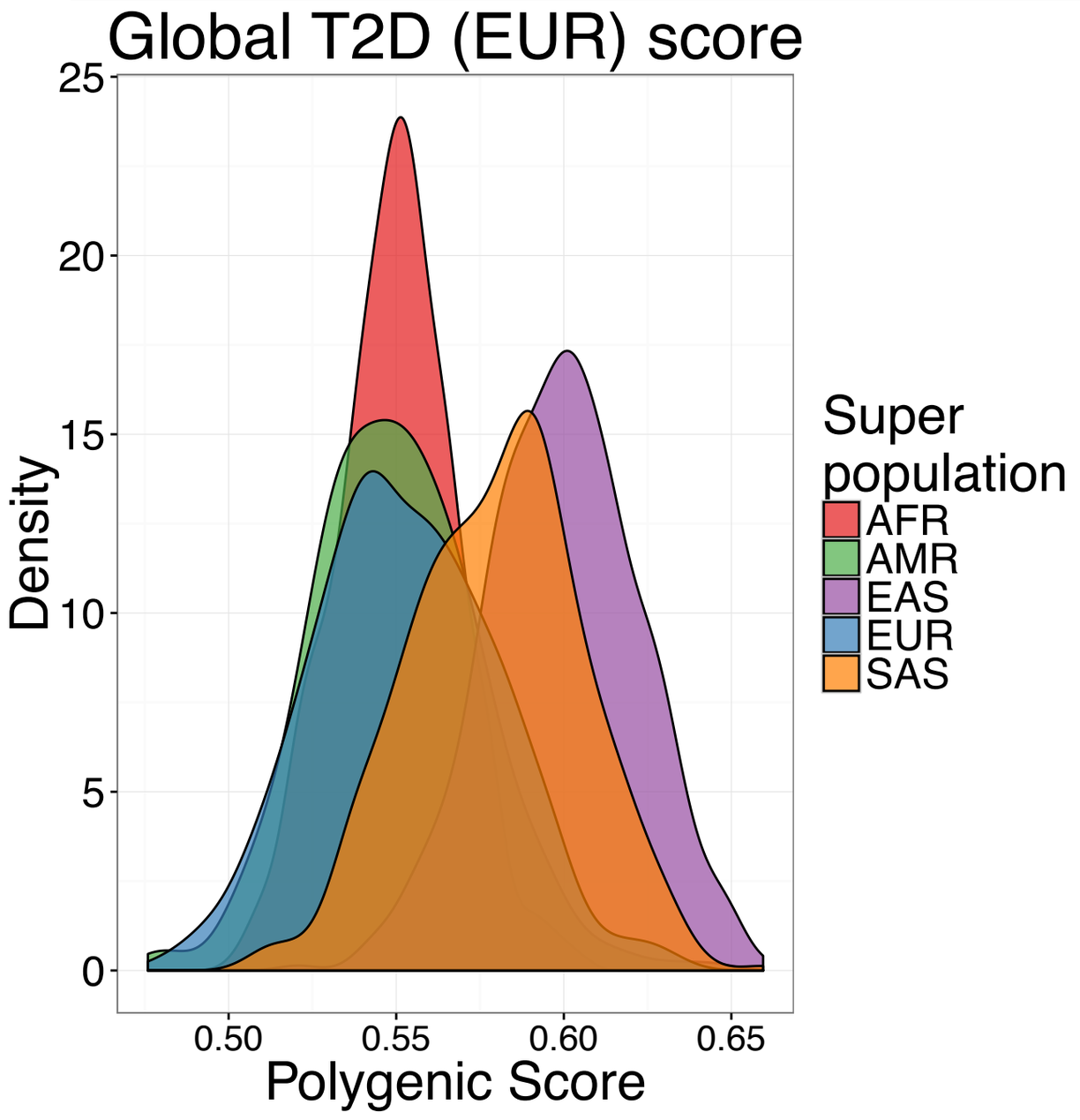
Polygenic risk score for height reflects adaptive event in Europeans… and bias

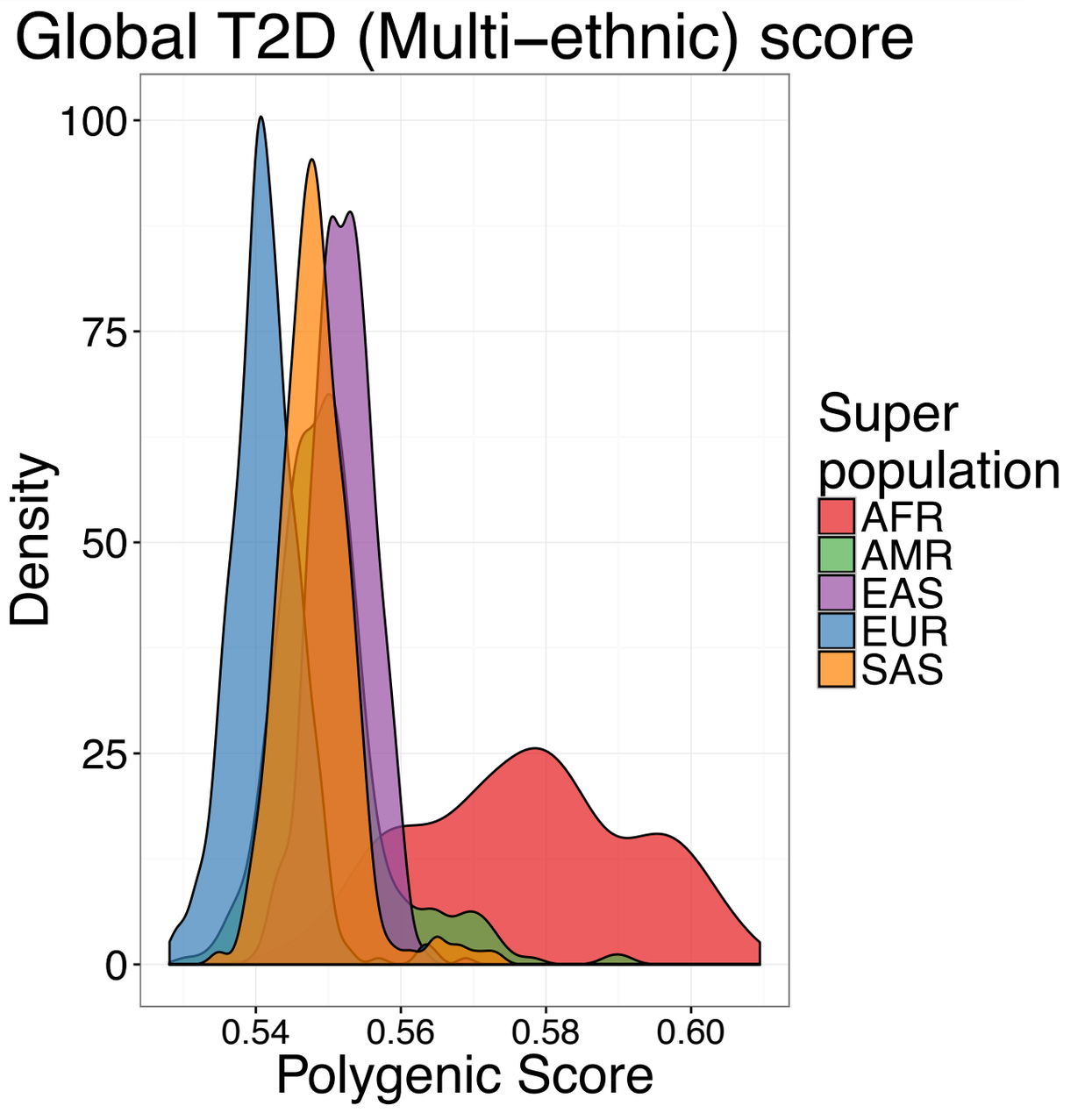




Wood, A.R., et al. (2014). Nature Genetics 46, 1173–1186.

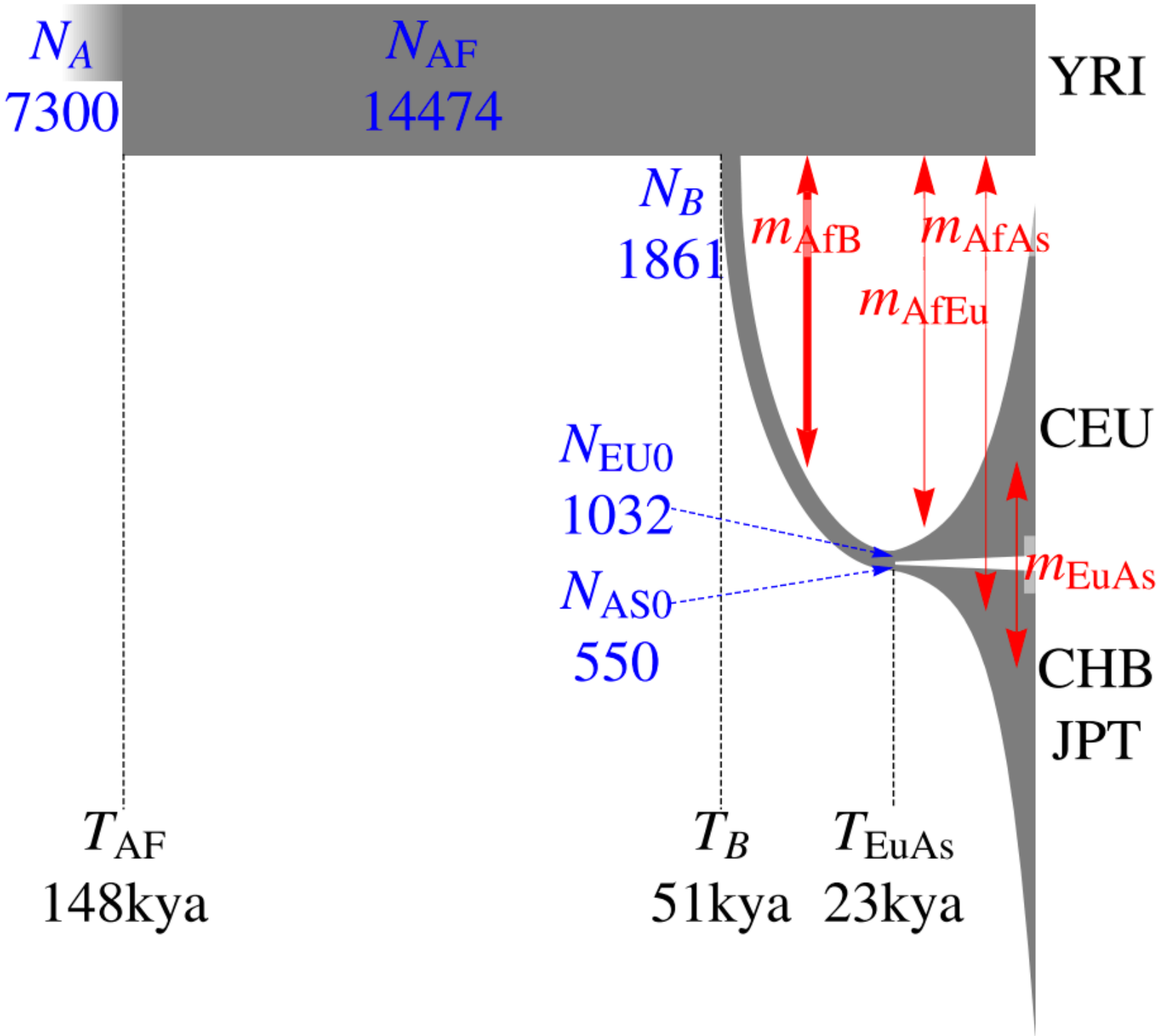
Polygenic risk score for Type II diabetes highlights role of demography





European: Gaulton, K.J., et al. (2015). Nat. Genet. *47*, 1415–1425.
Multi-ethnic: Mahajan, A., et al. (2014). Nat. Genet. *46*, 234–244.

Coalescent model for simulation framework



Demographic model: Gravel, S., et al. (2011). Proc. Natl. Acad. Sci. U. S. A. *108*, 11983–11988.
msprime: Kelleher, J., Etheridge, A.M., and Mcvean, G. (2015). PLoS Comput Biol 1–22.

Simulation steps

- Simulate for chr20 (μ=2e-8 mutations/(bp*generation)) genotypes with HapMap recombination map for 200k each: Africans, East Asians, Europeans
- Assign "true" causal effect sizes to *m* evenly spaced variants as:

$$\beta \sim N(0, \frac{h^2}{m})$$

- As before, define *X* as:

$$X = \sum_{i=1}^{m} g_i \beta_i$$

- Normalize:

$$Z_X = \frac{X - \mu_X}{\sigma_X}$$

- Compute true PRS as (such that total variance is h^2):

$$G = \sqrt{h^2} * Z_X$$

Simulation steps

- Compute the total liability for each individual (epsilon is standard normal noise), such that:

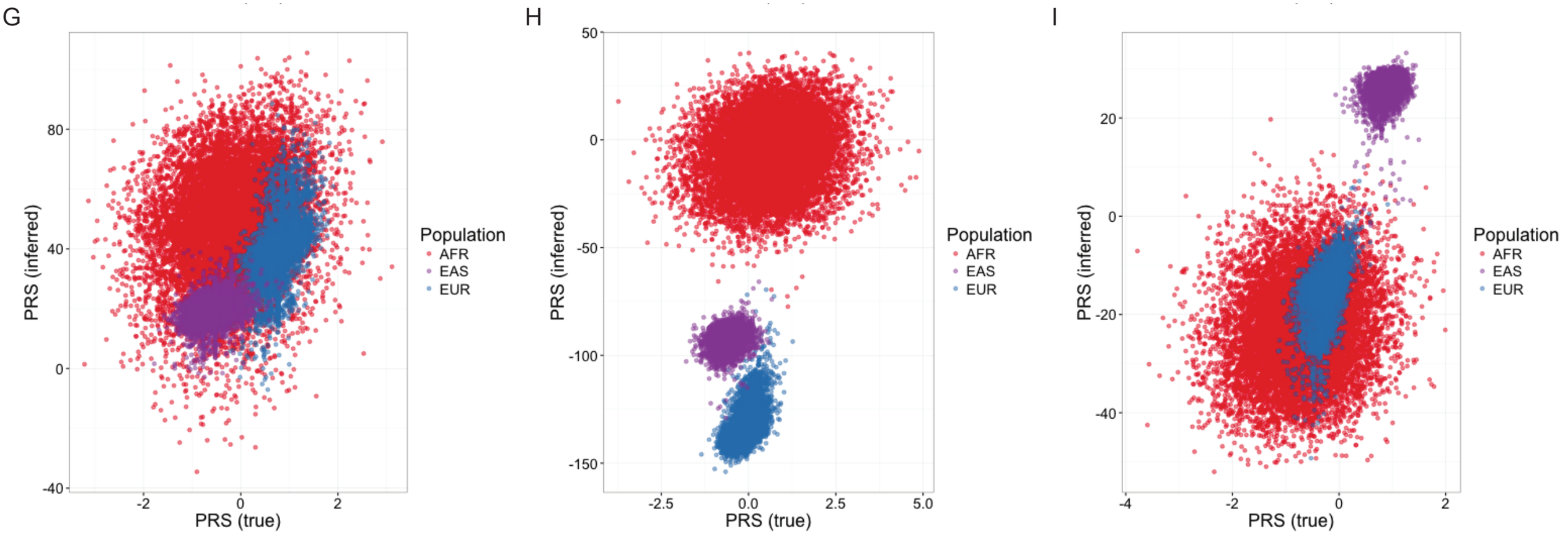
$$T = \sqrt{h^2} * Z_X + \sqrt{1 - h^2} * Z_\epsilon$$

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2}$$

- Assuming a 5% prevalence, assign 10,000 European individuals at the most extreme end of the liability threshold “case” status. Randomly assign different 10,000 European individuals “control” status.
- Run a simulated GWAS, computing Fisher’s exact test for all sites with MAF ≥ 0.01.
- Clump SNPs into LD blocks for all sites with p≤1e-2, R^2≥0.5 in Europeans, and window size of 250kb.
- Compute inferred PRS from summary stats and ρ with true PRS
- Evaluate over 50 simulations for m = 200,500,1000 and h^2=0.33,0.50,0.67

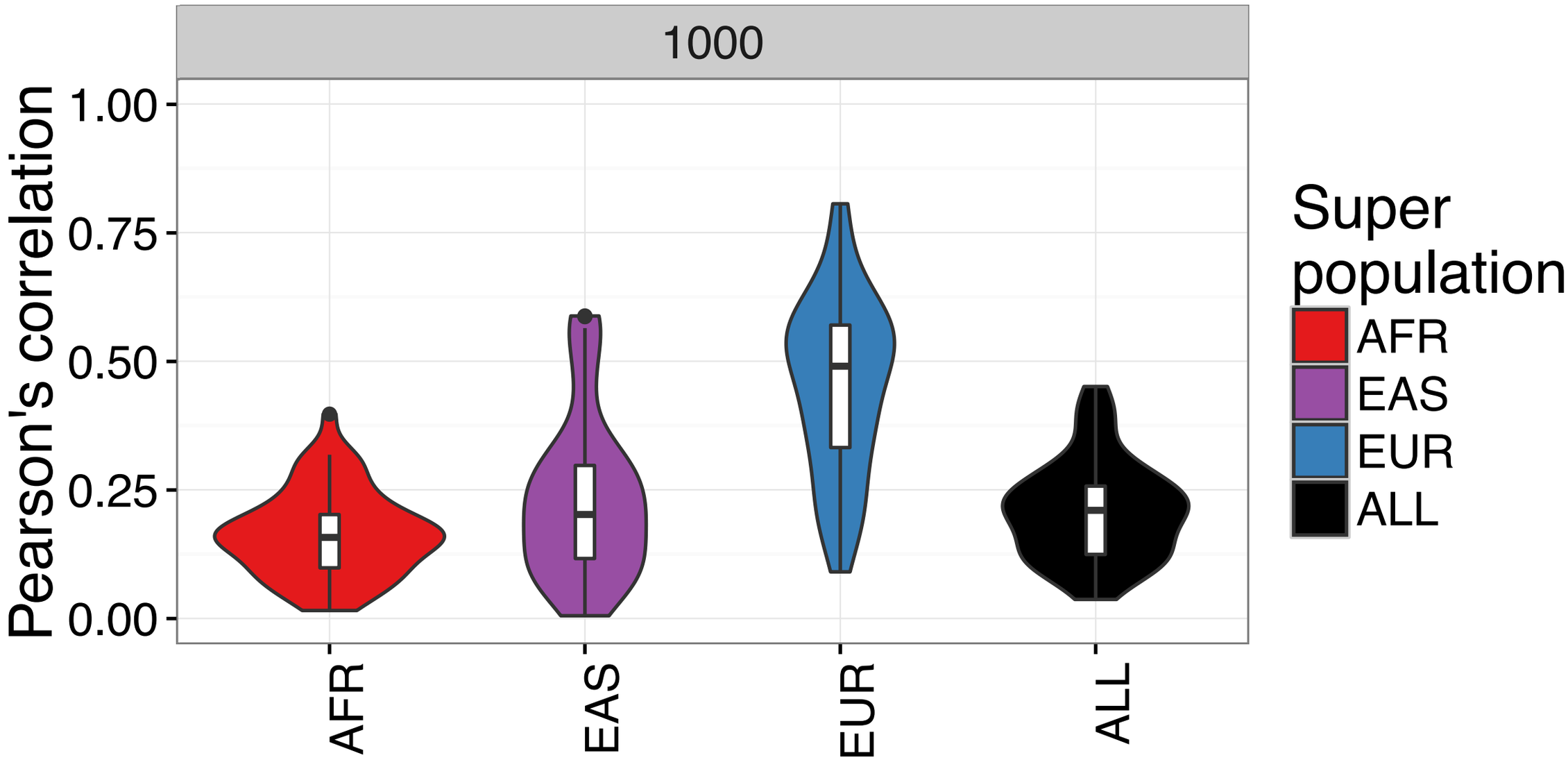
True vs inferred PRS with same causal variants, different effect sizes are inconsistent

h^2=0.67, m=1000



Best performance in European study population

h^2=0.67, m=1000, 50 replicates



New Results

Population genetic history and polygenic risk biases in 1000 Genomes populations

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

doi: http://dx.doi.org/10.1101/070797

http://biorxiv.org/content/early/2016/08/23/070797