Studies of Diverse Populations

Genetic and Molecular Epidemiology: EPID 217, Winter 2021

Taylor Cavazos

Outline

- What do we mean by diverse populations?
- Genetic admixture, admixture analysis, and admixture mapping
- Why does diversity matter in clinical research or genetic epidemiology?
- Example: Polygenic risk scores in non-European populations

What do we mean by diverse populations?

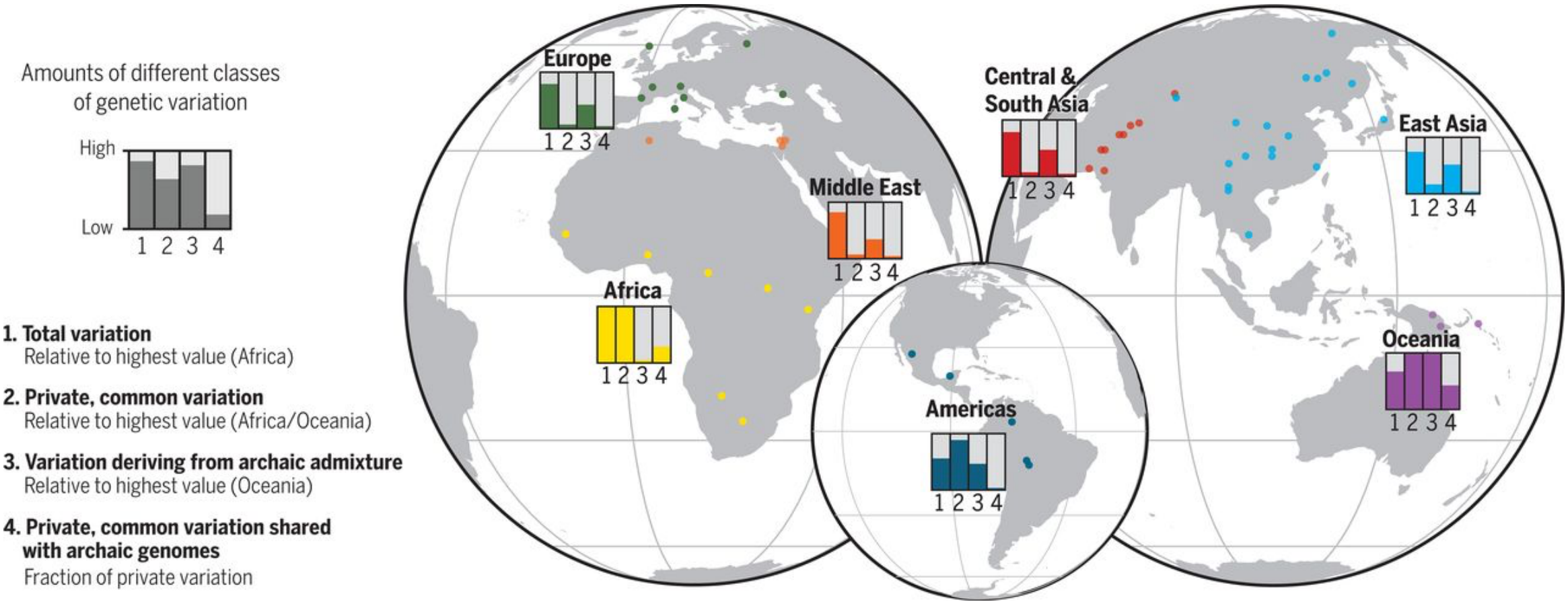
Race vs. Ethnicity vs. Ancestry

- **Race** is a social construct often linked to ancestry

- **Ethnicity**: cultural heritage, ancestry, origin myth, history, homeland, language and/or dialect, symbolic systems such as religion, mythology and ritual, cuisine, dressing style, and art

- **(Genetic) Ancestry** defined in terms of continental (or sub-continental) origin of ancestors

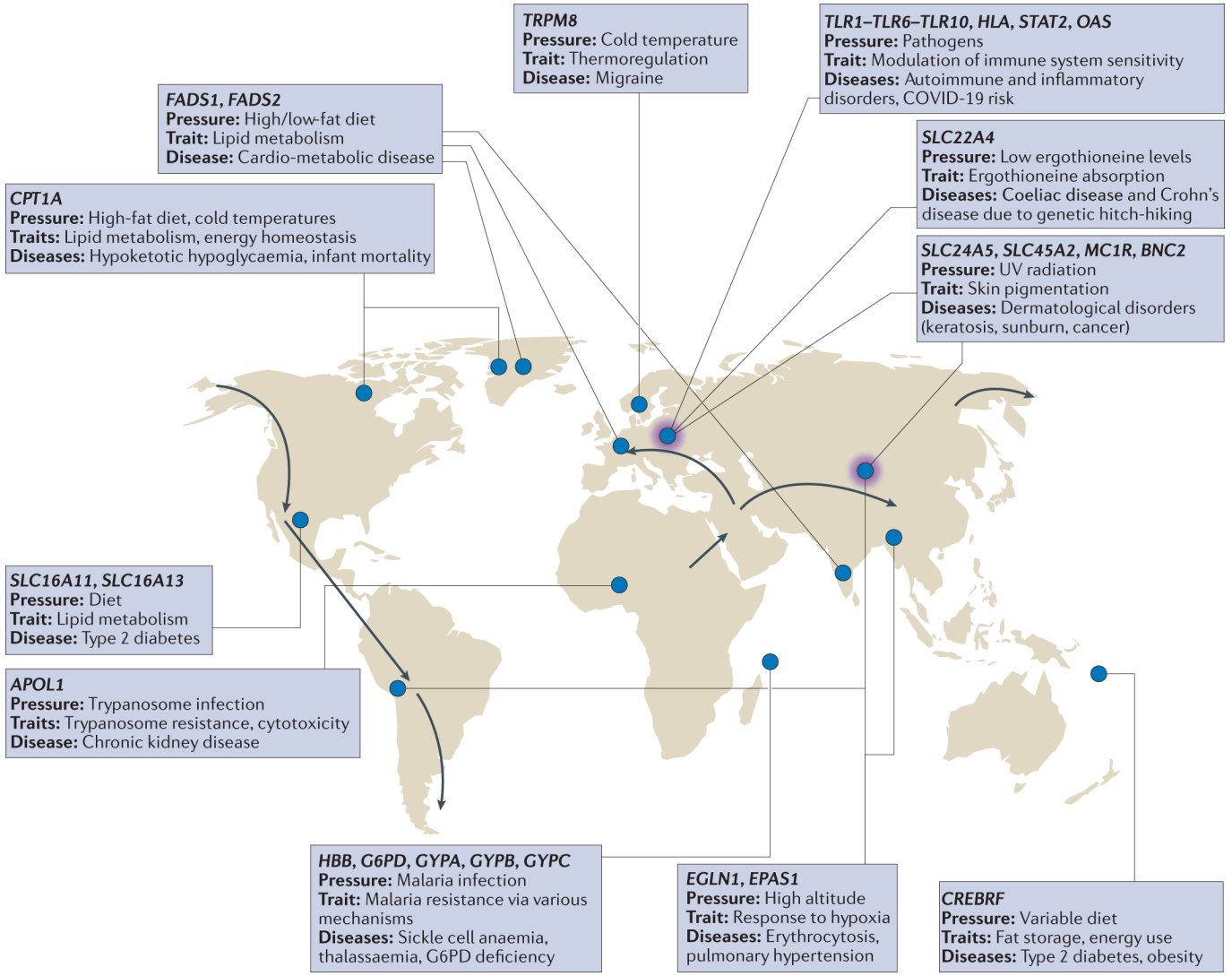
Genomic diversity **across** populations



Bergström et al., Science (2020)

Genomic diversity **across** populations

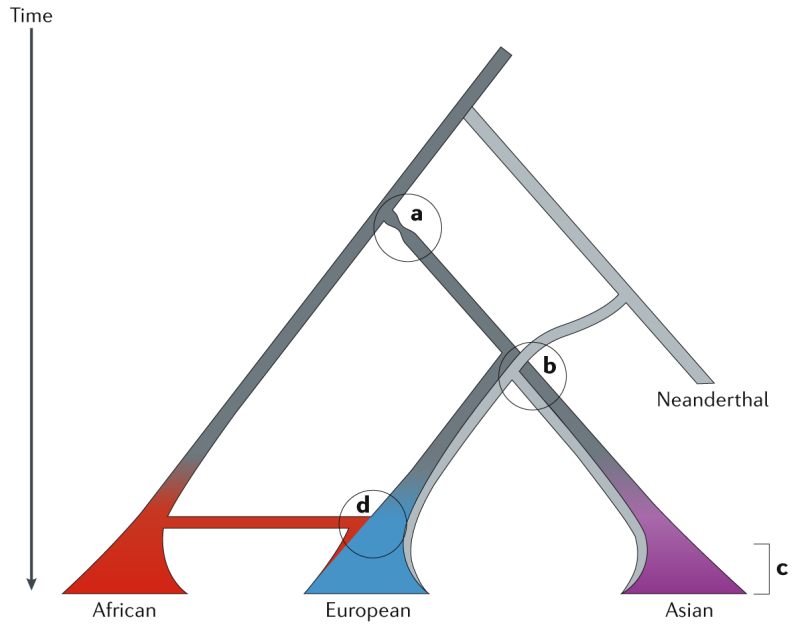
- Most genetic variants in human populations recently evolved (over the past ~100K years)
- Genetic variation arises from evolutionary/ environmental pressures and can have trade-offs that may lead to disease

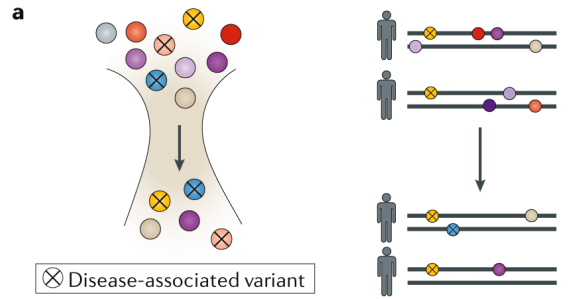


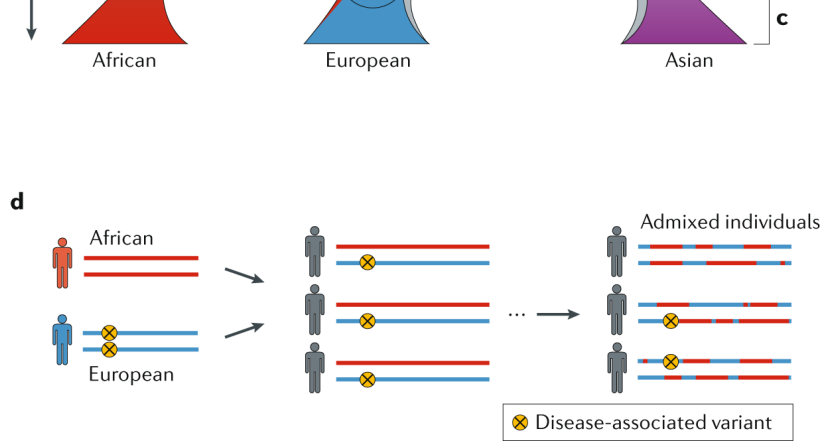
Benton et al., Nature Reviews Genetics (2021)

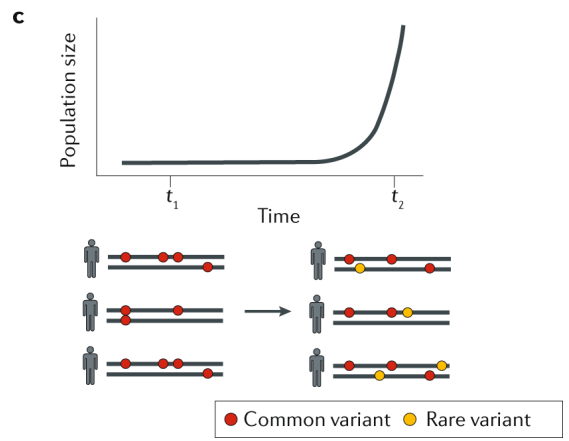
Genomic diversity **across** populations

- The demographic history of modern humans resulted in different genetic architectures among human populations
- Population bottlenecks resulted in a loss of genetic variation, but higher mutation load







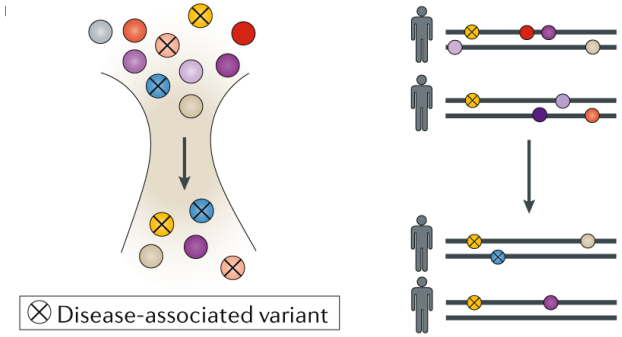


Benton et al., Nature Reviews Genetics (2021)

Factors influencing variant allele frequency changes

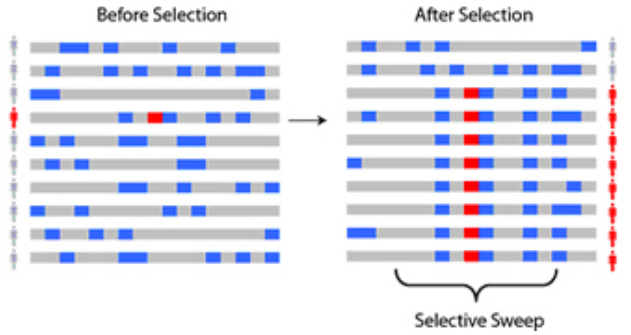
Genetic Drift – a <u>random</u> sampling resulting in AF differences

- Bottleneck Effects
- Founder Effects



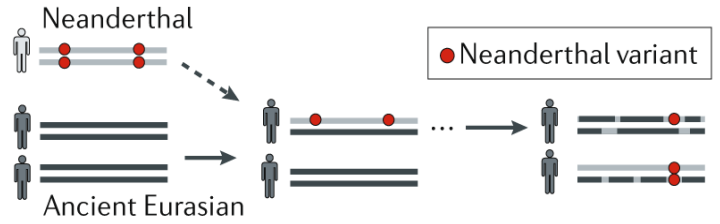
Benton et al., Nature Reviews Genetics (2021)

Natural Selection – acts on heritable variation that impacts fitness



Schaffner et al., Nature Education (2008)

Gene Flow – transfer of alleles from one population to another



Benton et al., Nature Reviews Genetics (2021)

Databases describing genetic variation across human populations

Resource	Individuals	Populations	Variants (million)
gnomAD	141,456	> 60	241
1000 Genomes	2,504	26	84.4
HapMap	1,397	11	1.5

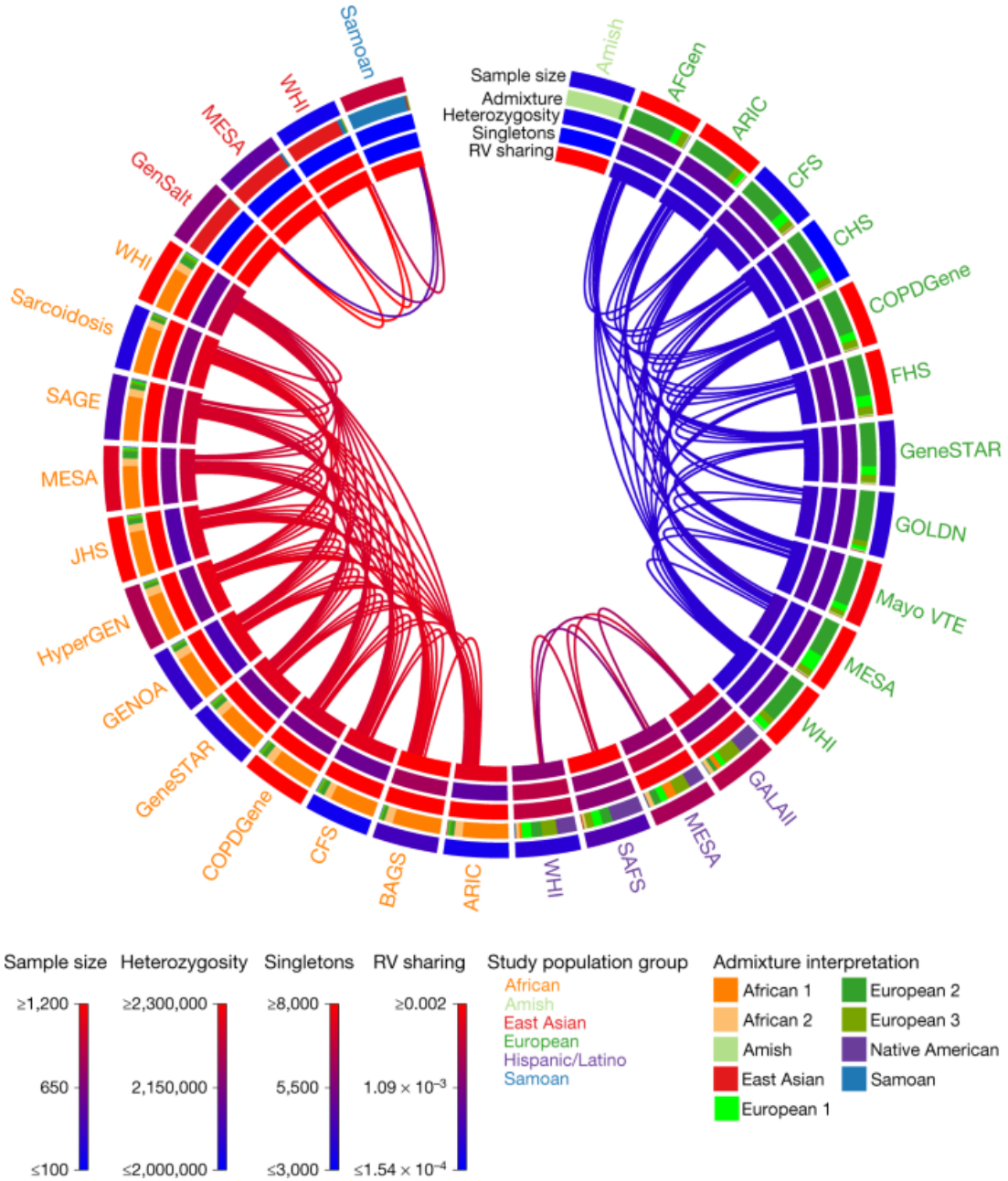
What are some of the factors influencing the number of genetic variants found?

There is even more genetic variation being discovered every day!

WGS of **53,831** diverse participants from **TOPMed**

410 million genetic variants identified

(published Feb 10th)



Taliun et al., Nature (2021)

Summary

- The primary source of genetic structure in the human population is based on continent of origin
- Genetic disease is a by-product of evolution
- With increased sequencing of diverse human populations, we can further catalogue the vast amount of genetic variation present

Methods to find and describe patterns of variation across populations

Kimmel February 14, 2016 at 4:42 pm

I've always seen some Asian in her. I'm leaning towards Southeast Asian mixed with typical 'Hispanic' traits.

LOG IN TO REPLY

Adrian1Cram February 12, 2016 at 10:36 pm

She doesn't look fully white to me, I could tell she's mixed. Also with the natural hair she looks much more mixed then right now.

LOG IN TO REPLY

phaedra January 17, 2016 at 8:07 pm

She's probably like 75% Spanish, 25% Amerindian.

LOG IN TO REPLY

fuzzybear44 January 18, 2016 at 2:53 am

I doubt those amounts. Plus all the DNA results I've seen say PR's are multiethnic(Whte, Black, Amerindian). That's most likely what she is.

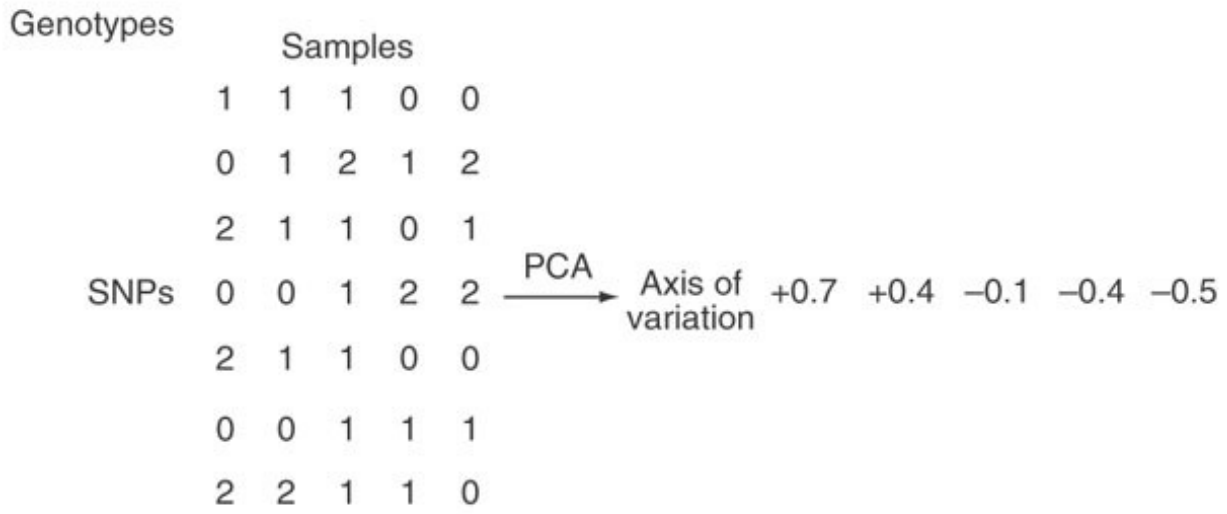
http://ethnicelebs.com/jennifer-lopez

Defining genetic ancestry

Approach: Principal Components Analysis (PCA)

- **PCA** is commonly used for identifying and adjusting for ancestry differences among sample individuals

- **PCA** is applied to genetic data to identify independent **principal components** (PCs) that explain differences among individuals

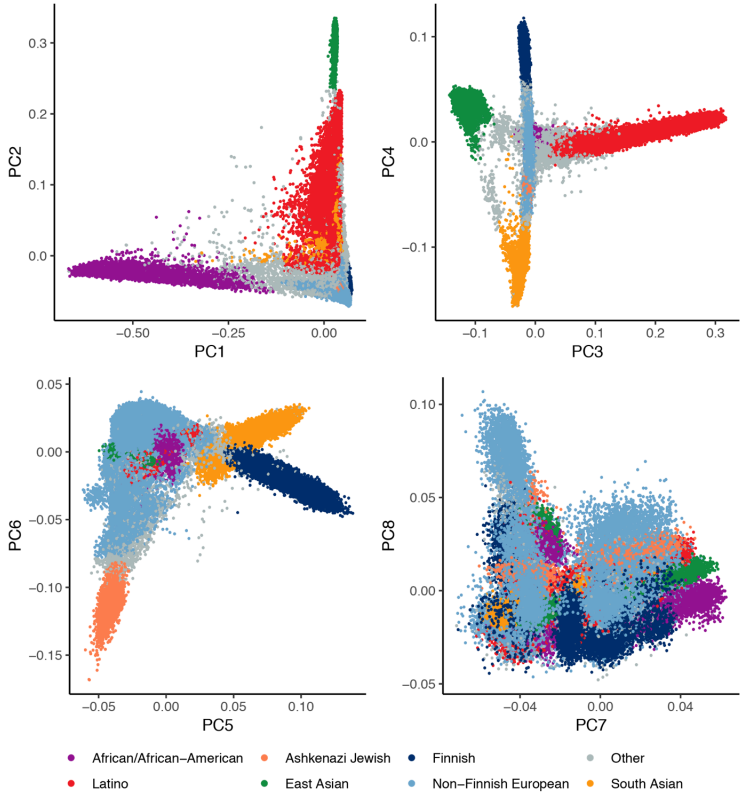


Price et al., Nature Genetics (2006)

Defining genetic ancestry

Example: ~60K WGS samples representing > 60 populations in gnomAD v3.1

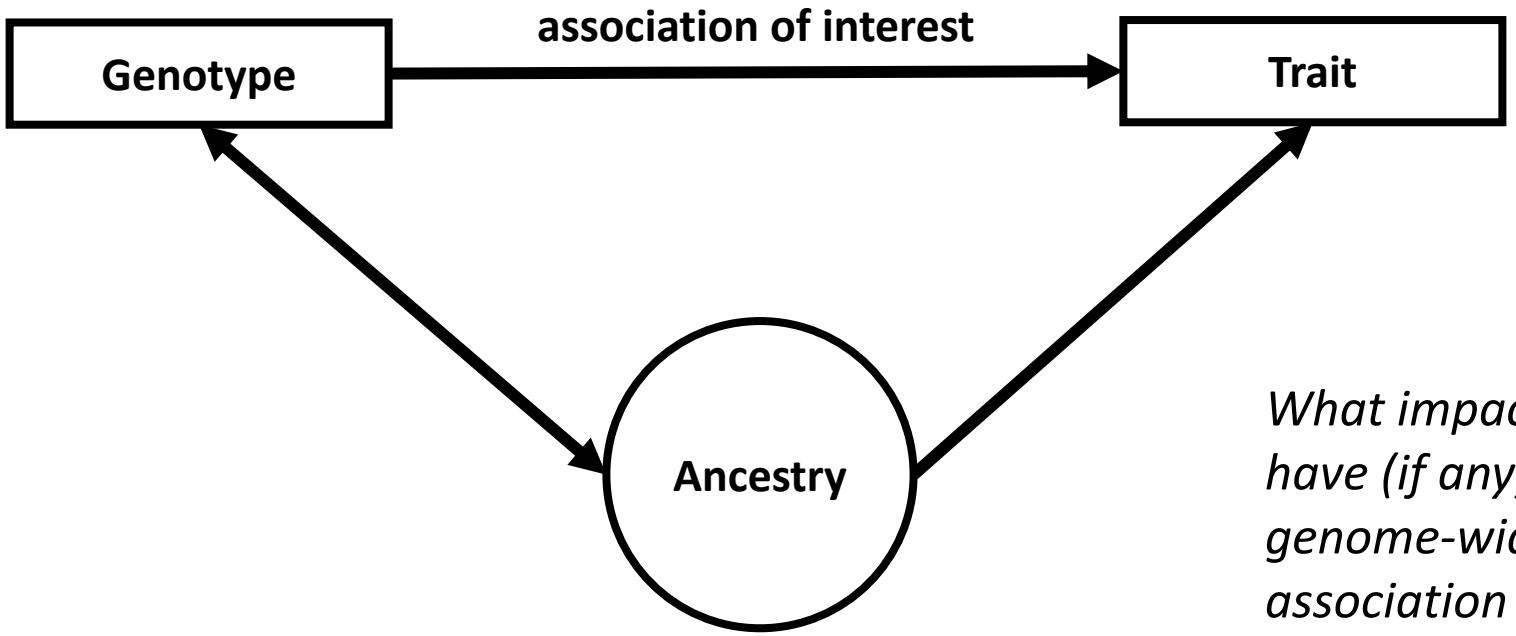
Plot of Top 8 Genetic PCs



UMAP Representation of Top 10 PCs

Karczewski et al., Nature (2020)

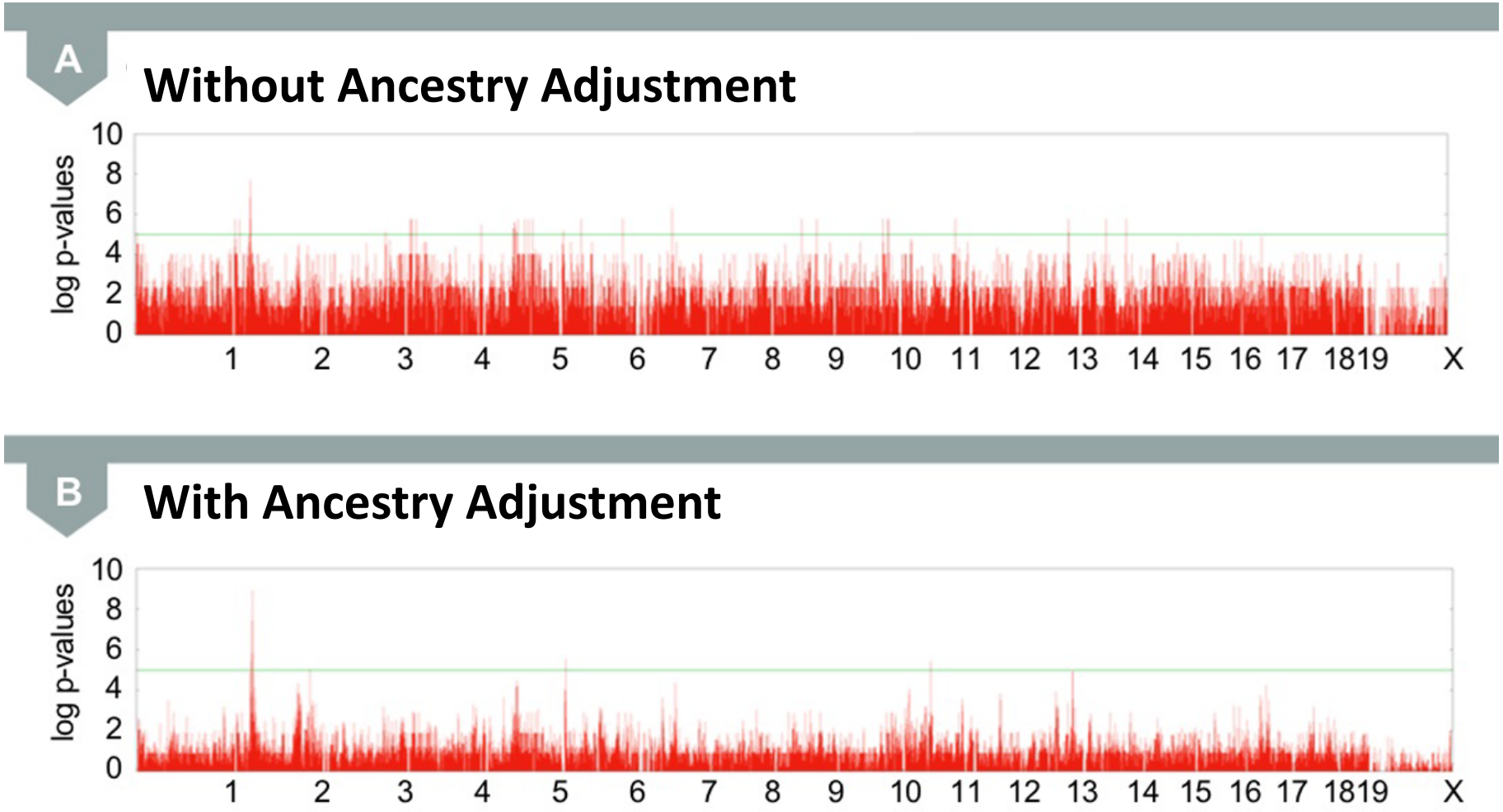
Confounding by ancestry



What impact does this have (if any) on genome-wide association studies?

confounder: a variable in a statistical model that is correlated with the independent and dependent variable

Confounding by ancestry



Hoon Sul et al., PLOS Genetics (2018)

Correcting for ancestry confounding

- PCs model ancestry differences between cases and controls
- It is standard practice to include the top 10 PCs in your model to adjust for ancestry

$$Y = \beta_0 + \beta_1 X + \beta_2 PC_1 + \beta_3 PC_2 + \beta_4 PC_3 + \ldots + \varepsilon$$

- Sometimes additional correction is needed for structured data and mixed models can be used.

Assigning individuals to ancestry groups

Motivations

- Control for confounding by ancestry
- Reconstruct genetic history of recently admixed populations
- Understand heterogeneity within your dataset

Challenges

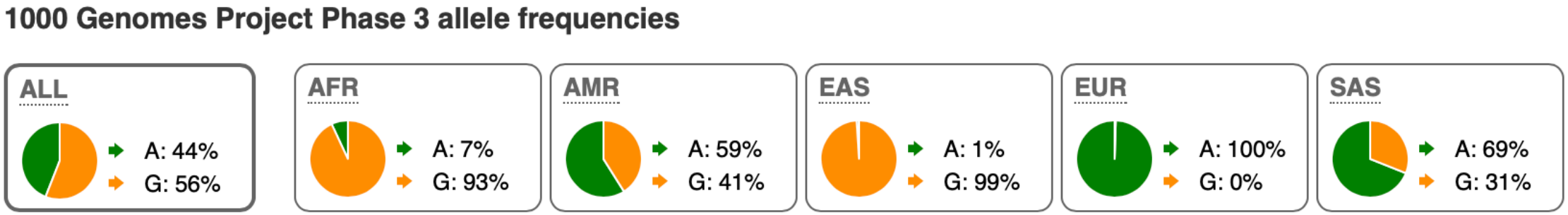
- Determining the number of ancestral populations present
- Assessing whether ancestry assignments are correct

Ancestry Informative Markers (AIMs)

A SNP that exhibits significantly different frequencies between different populations

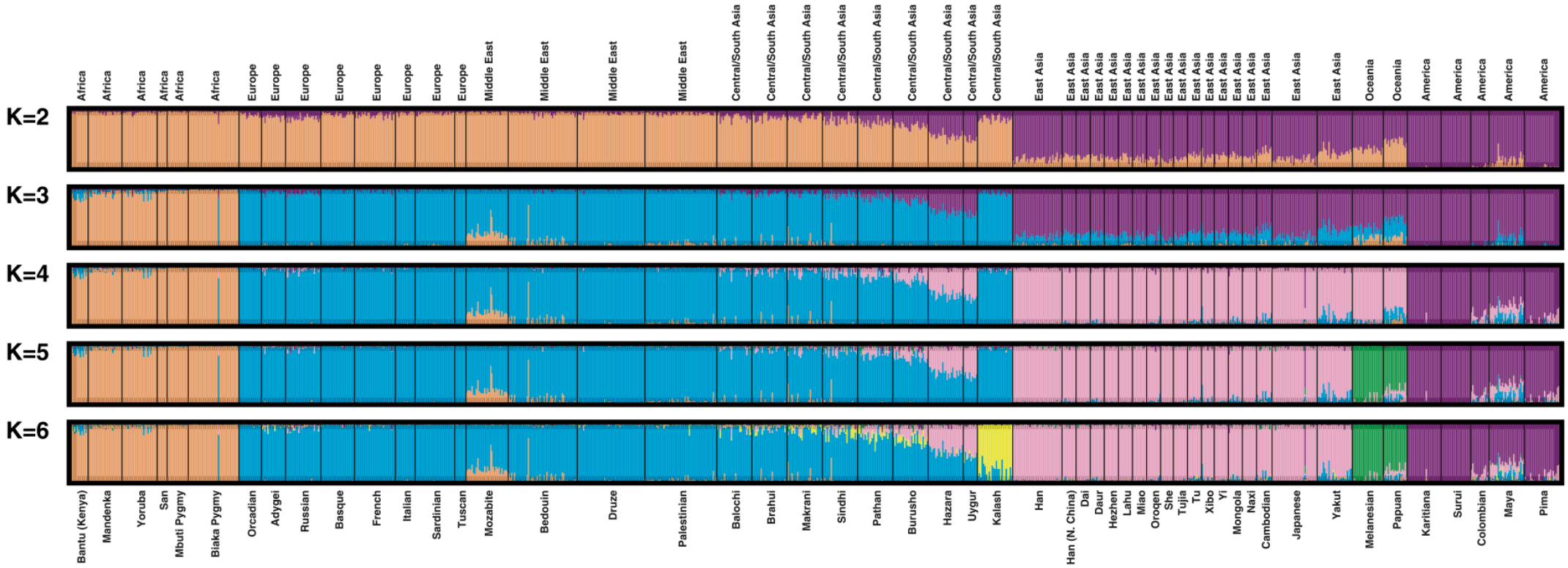
Example:

rs1426654 – A allele is under strong positive selection in EUR



Estimating *STRUCTURE* in human populations

- Assign individuals to **K** populations using a model-based clustering algorithm to group individuals with similar patterns of variation
- **377** markers in **1,056** individuals across **52** populations



Rosenberg et al., Science (2002)

Genetic admixture estimation

$m_k = fraction\ of\ loci\ from\ pop\ k$

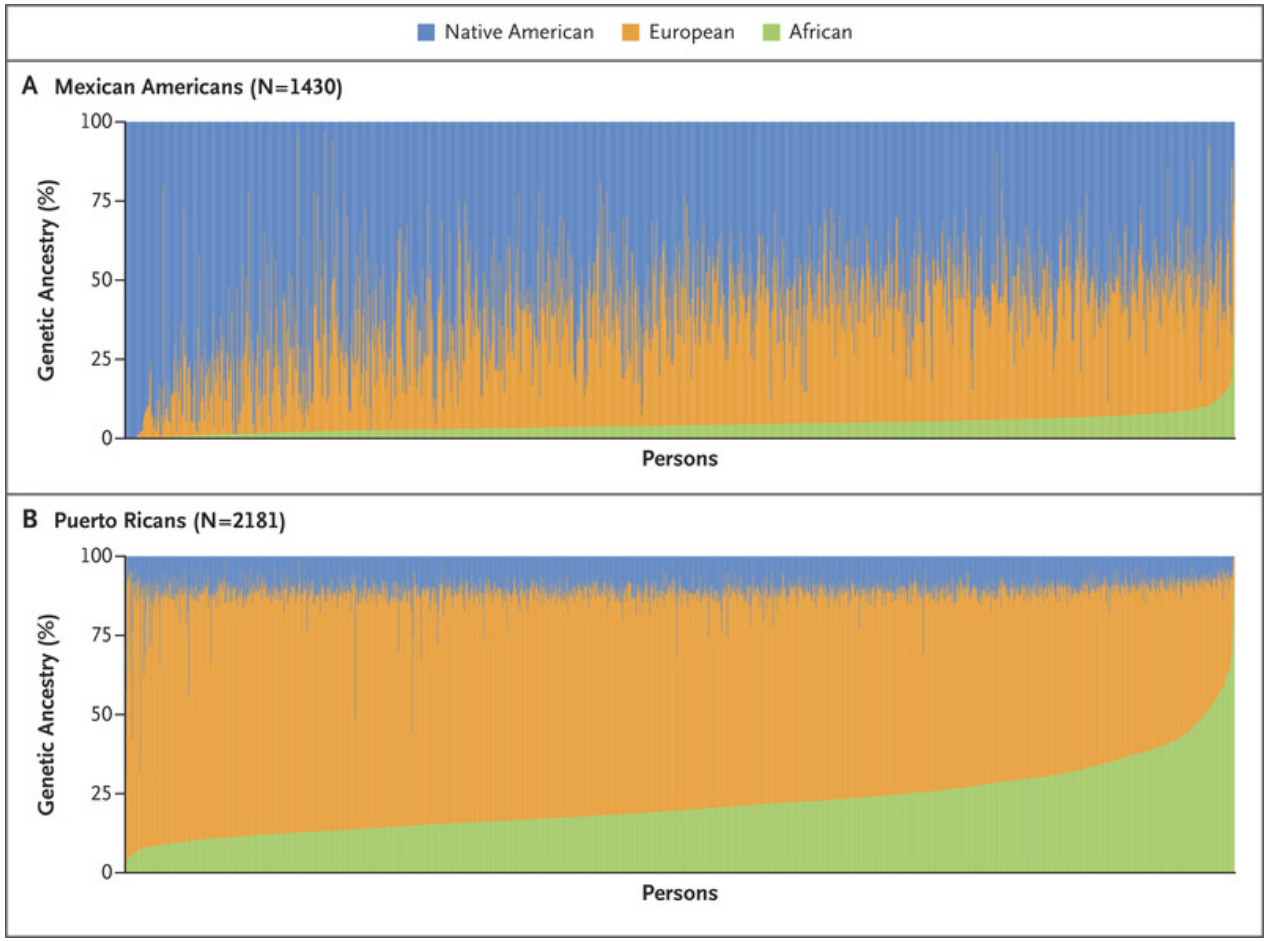
Goal: Identify the $\mathbf{m_k}$ that maximizes the probability of obtaining the observed genotype given the AF in each ancestral population

Unknowns:

Admixture proportions, AF in population k

Approaches:

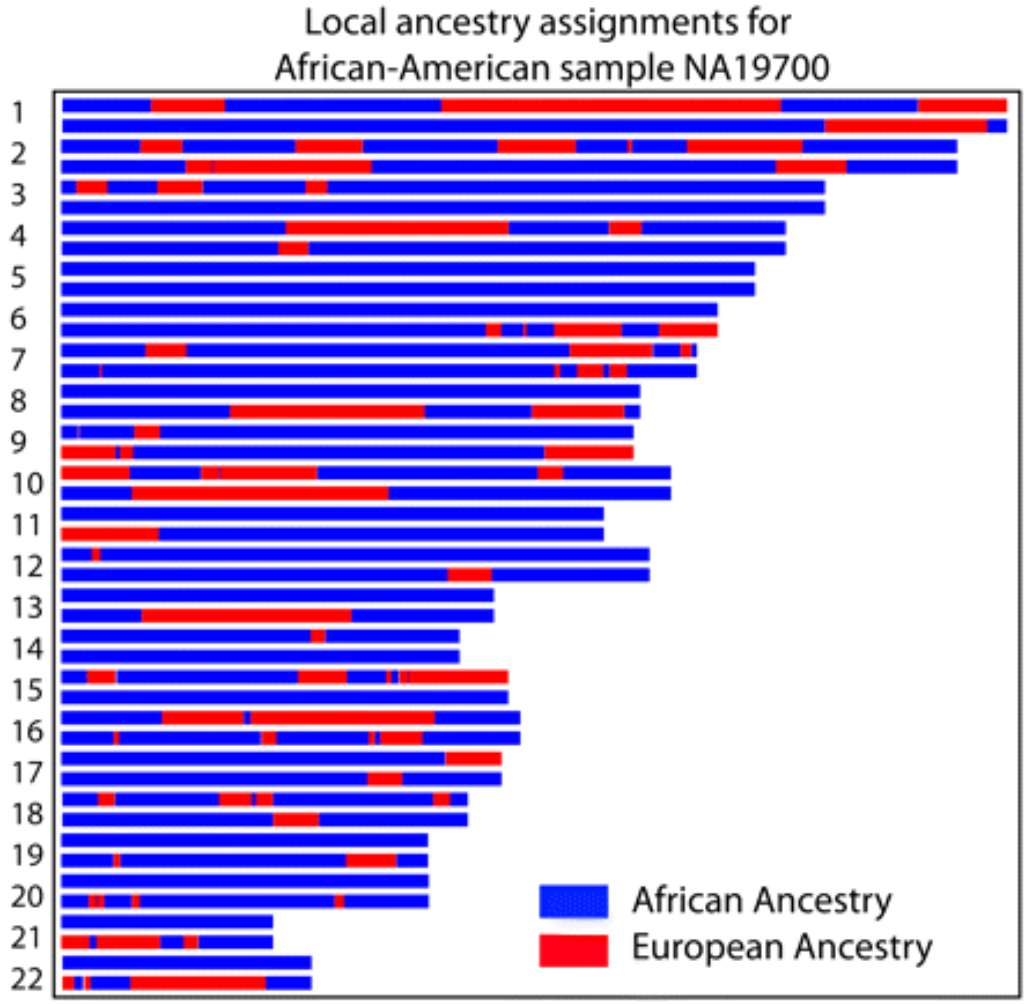
- Maximum Likelihood Estimation
- Bayesian Inference via MCMC



Burchard et al., Am J Public Health (2005)

Locus-specific ancestry

- Identifies ancestry of each segment in the genome
- An individual can have 0, 1, or 2 copies of an allele derived from each ancestral population
- Requires reference populations



Gravel, GENETICS (2012)

Methods for admixture estimation and ancestry assignment

Genetic Admixture Estimation Methods

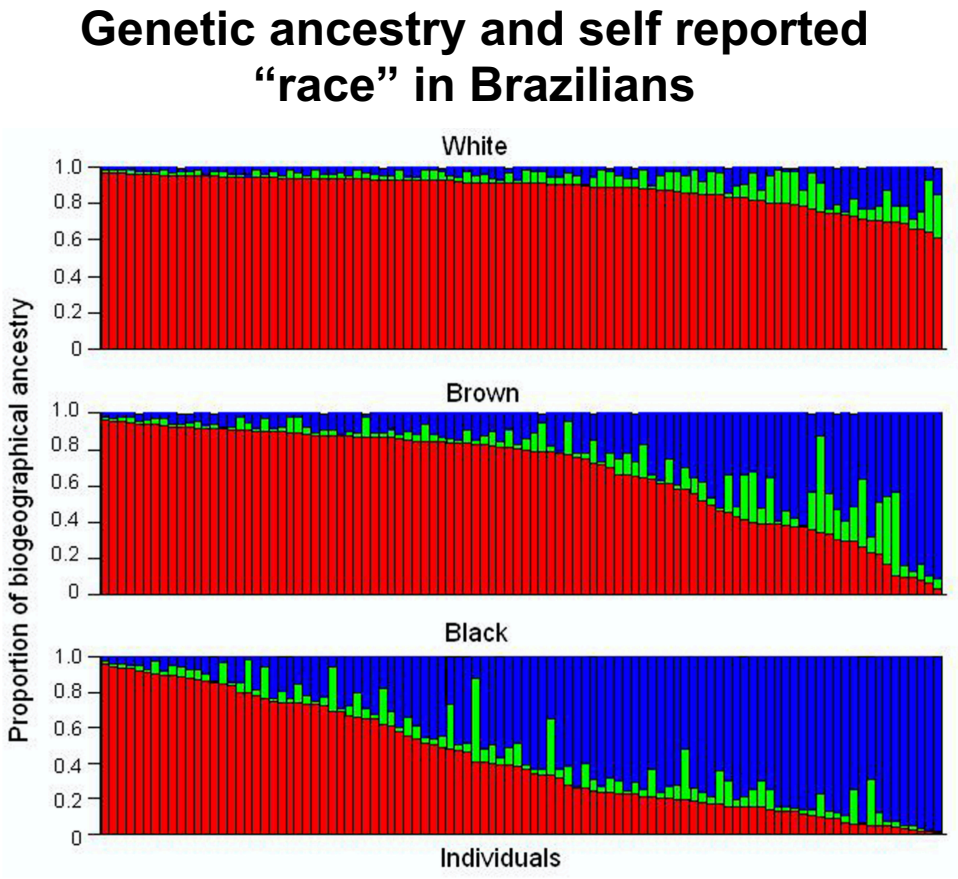
- STRUCTURE
  - http://pritch.bsd.uchicago.edu/structure.html
- ADMIXTURE
  - http://www.genetics.ucla.edu/software/admixture/index.html

Local Ancestry Inference Methods

- LAMP-LD
  - http://lamp.icsi.berkeley.edu/lamp/
- RFMix
  - http://med.stanford.edu/bustamantelab/

Summary

- Race/ethnicity and genetic ancestry are different concepts
- Adjusting for ancestry can prevent confounding in genome-wide association studies
- Structure exists across and within populations



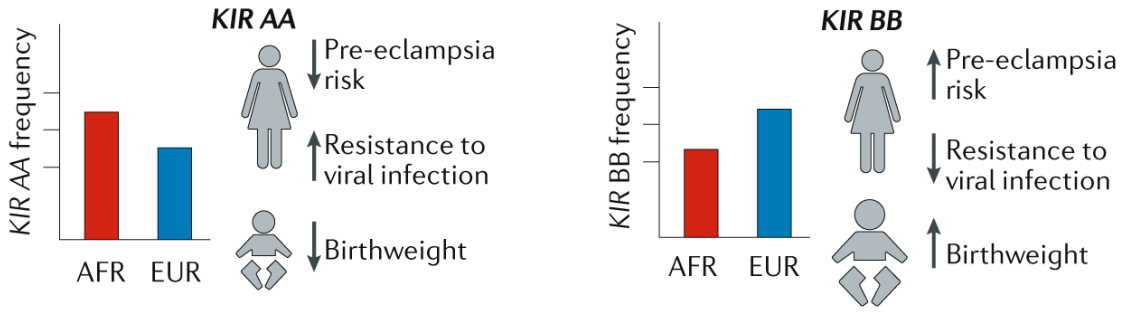
Suarez-Kurtz Front. Pharmacol. (2010)

Why diversity matters

The importance of considering diverse human populations in genetic analyses

Example 1: Trade-offs in pregnancy influenced by maternal KIR genotype

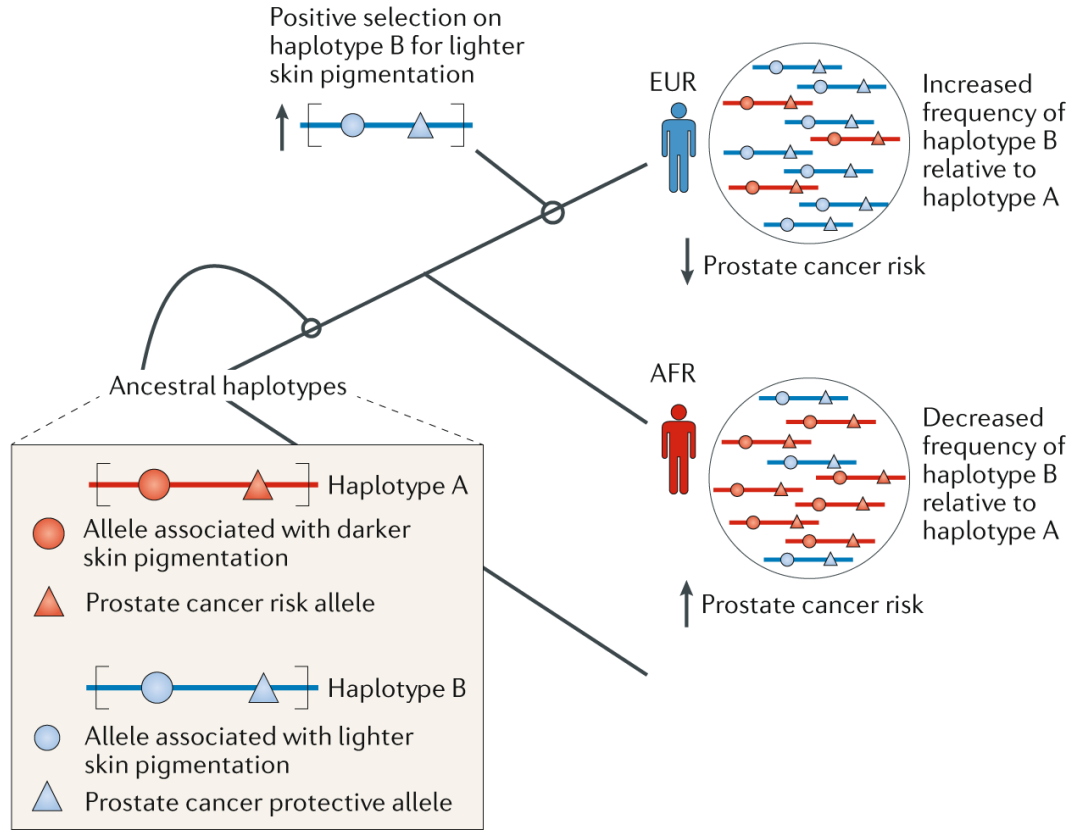
- KIR-AA genotype, present at higher proportions in African ancestry populations, results in improved maternal immune response, but lower birthweight



Benton et al., Nature Reviews Genetics (2021)

Example 2: Prostate cancer risk

- Selection of skin pigmentation variant results in decreased prostate cancer risk in Europeans



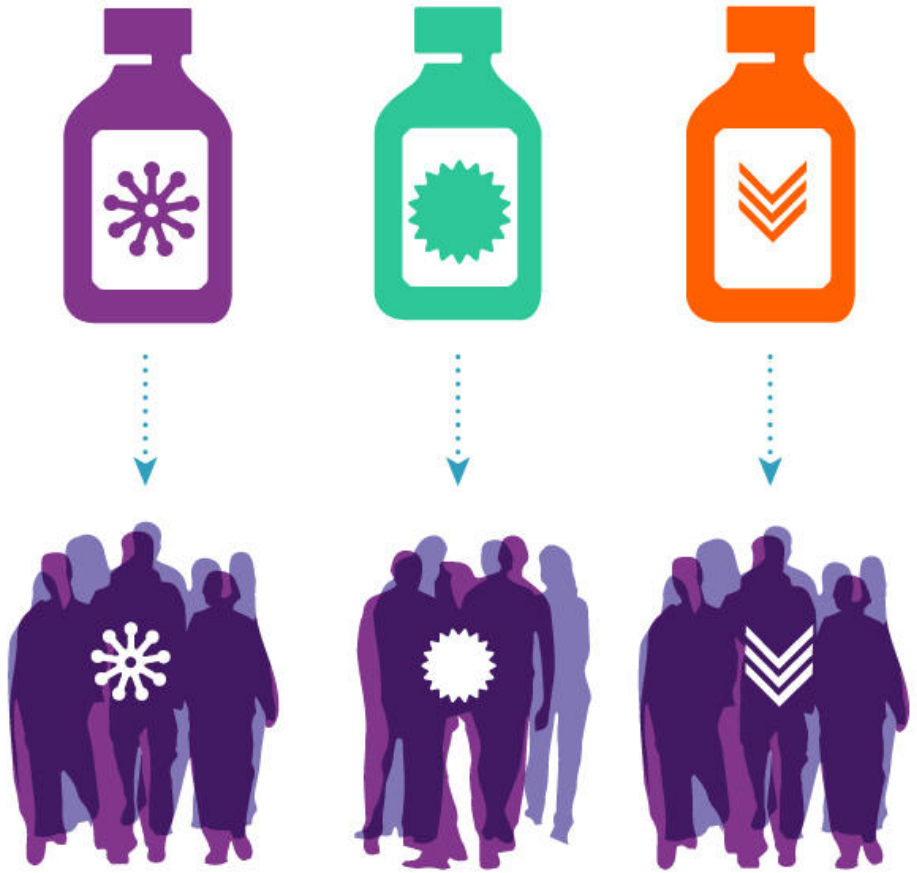
The importance of considering diverse human populations in genetic analyses

Example: The use of warfarin as an anticoagulant

Incorporation of genetic information meant to improve dosing

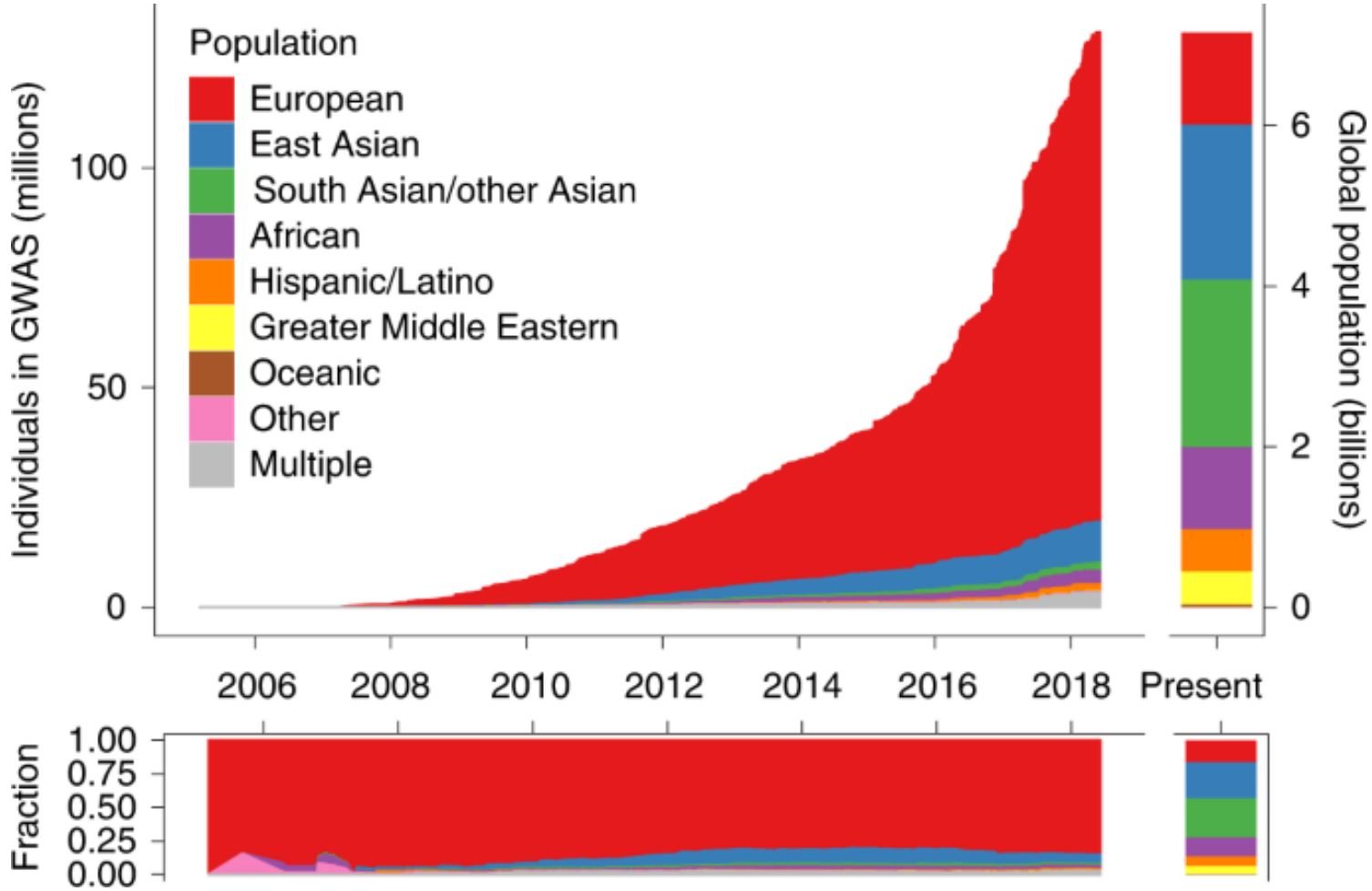
Result:

- Increased accuracy in predicting dose in European ancestry participants
- Worse coagulation in African ancestry participants



Oni-Orisan et al., NEJM (2021)

Representation of diverse populations in GWAS

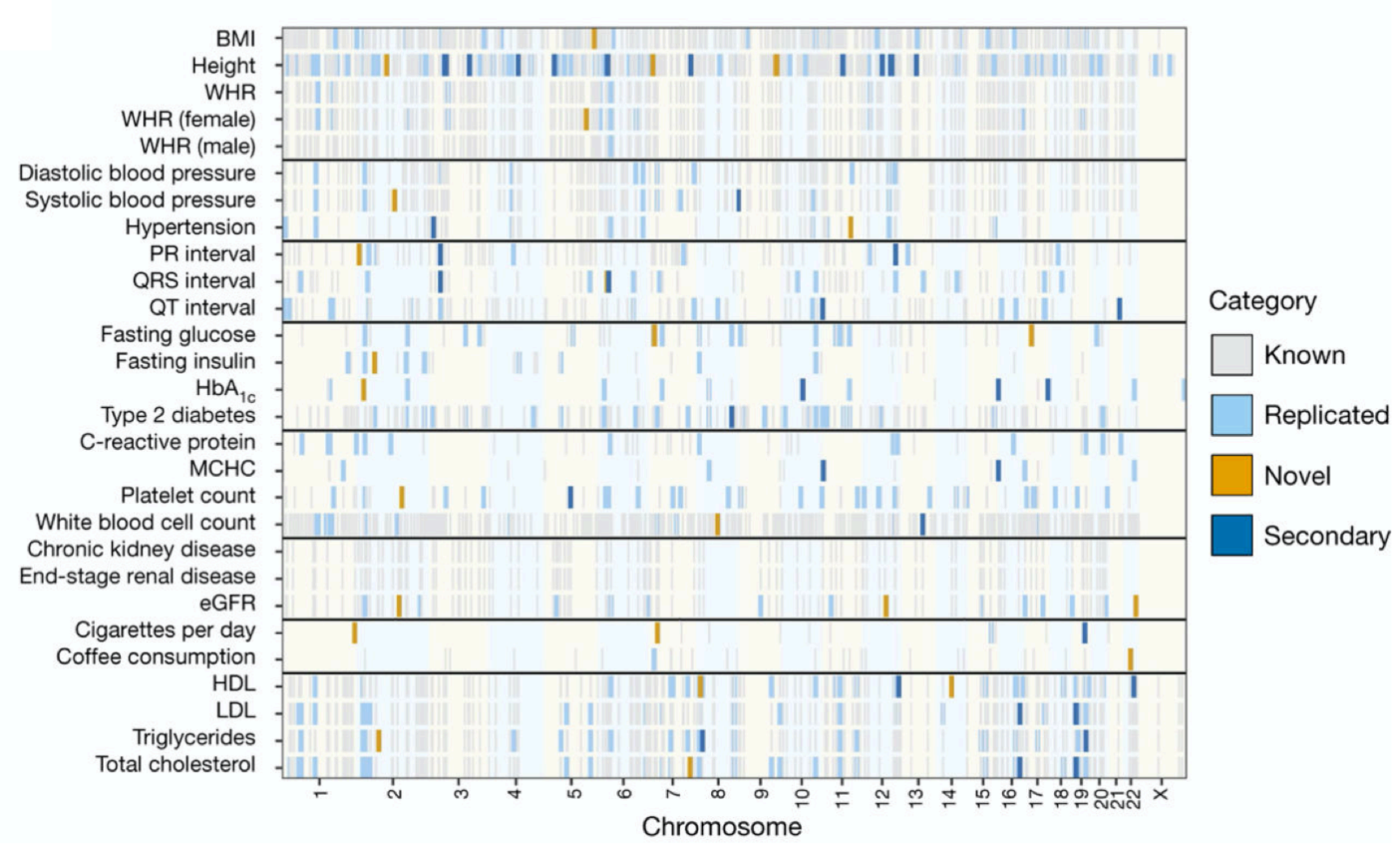


Martin et al., Nat Genet (2019)

From diversity to discovery

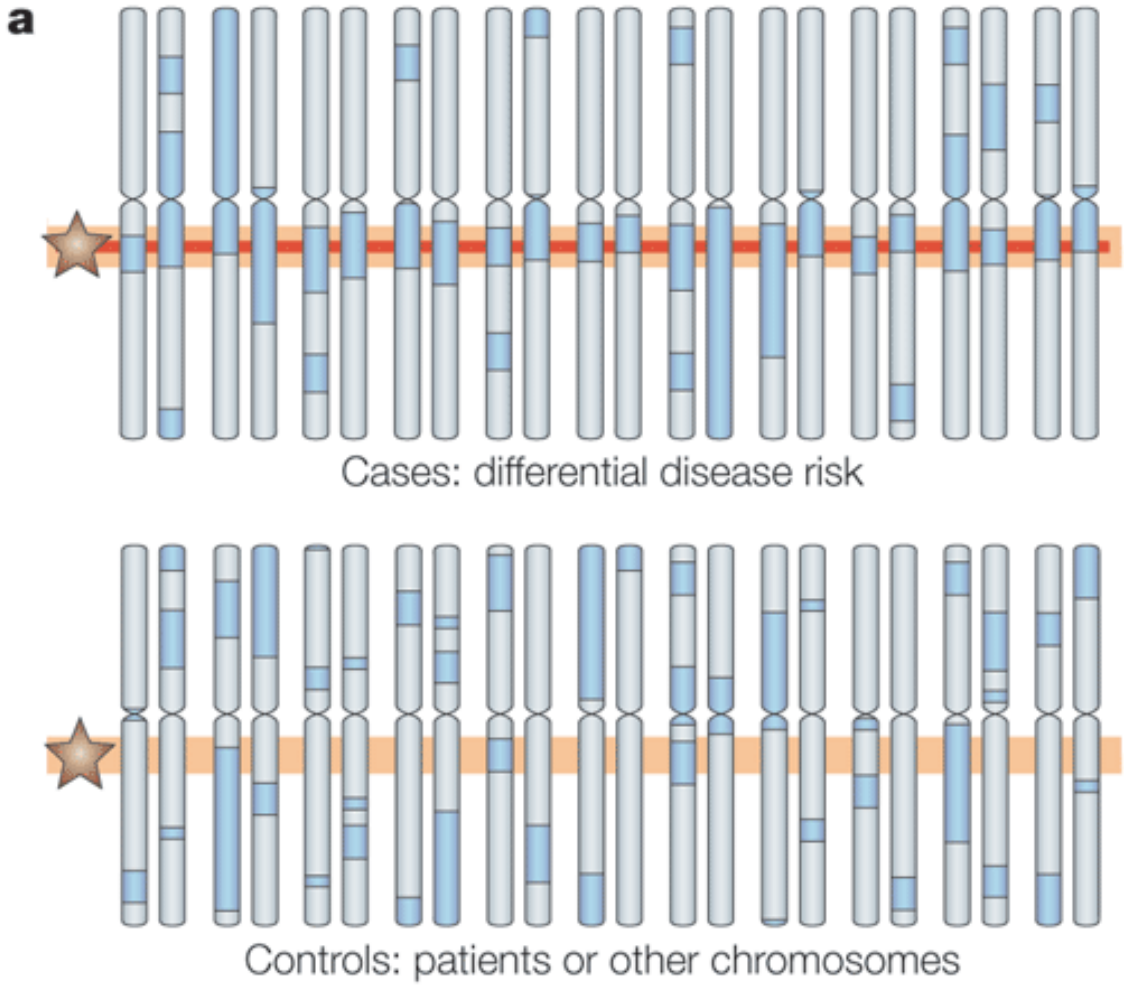
PAGE Study

- **49,839** non-European individuals
- **1,444** (of 8,979 total) variant-trait associations were replicated
- **27** novel loci
- **38** secondary signals at known loci

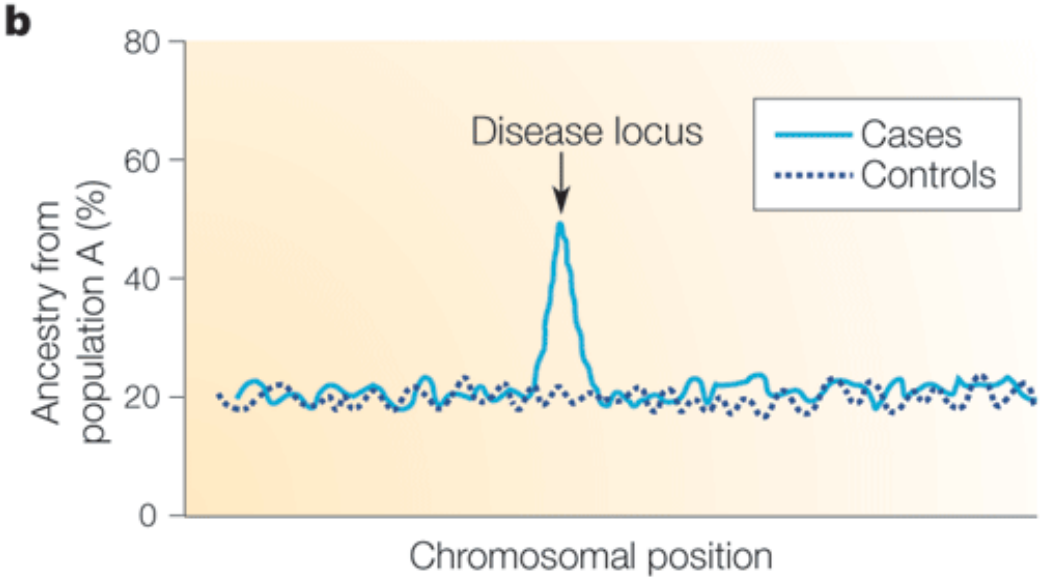


Wojcik et al., Nature (2019)

Fine mapping



- Relies on allele-frequency differences between ancestries
- Leverages admixed genomes to identify regions containing the causal disease variant



Smith and O'Brien, Nature Reviews Genetics (2005)

Challenges in studying diverse populations

- Limited statistical power due to small sample sizes non-European populations
- Lower imputation accuracy due to lack of large diverse reference populations (this is starting to change!)
- Greater heterogeneity and potential for ancestry confounding

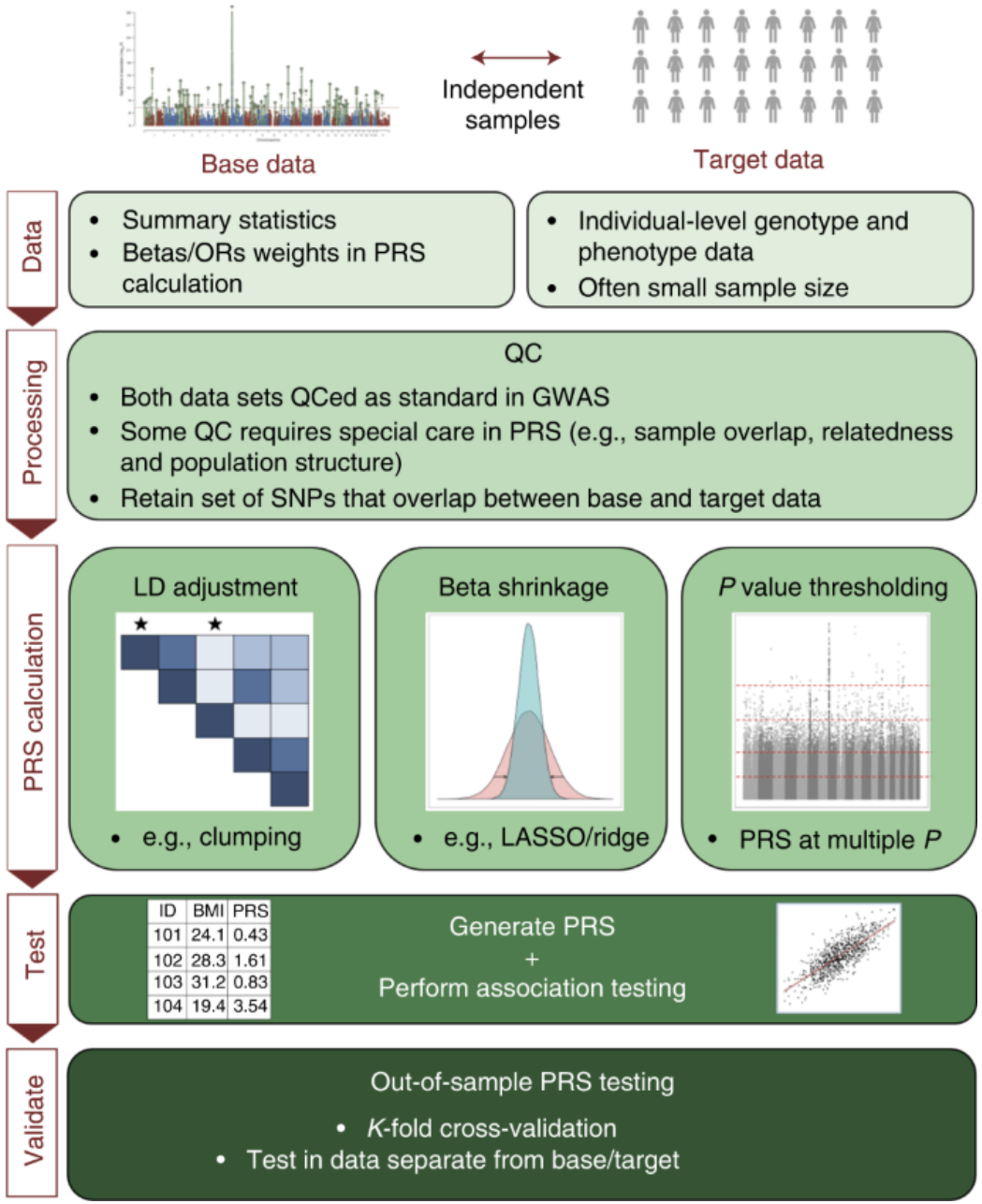
Example: Polygenic risk scores in diverse populations

PRS is a simple additive model for combining many GWAS discovered risk variants into a single score

$$PRS = \sum_{m=1}^{M} \beta_m X_m$$

Genotype

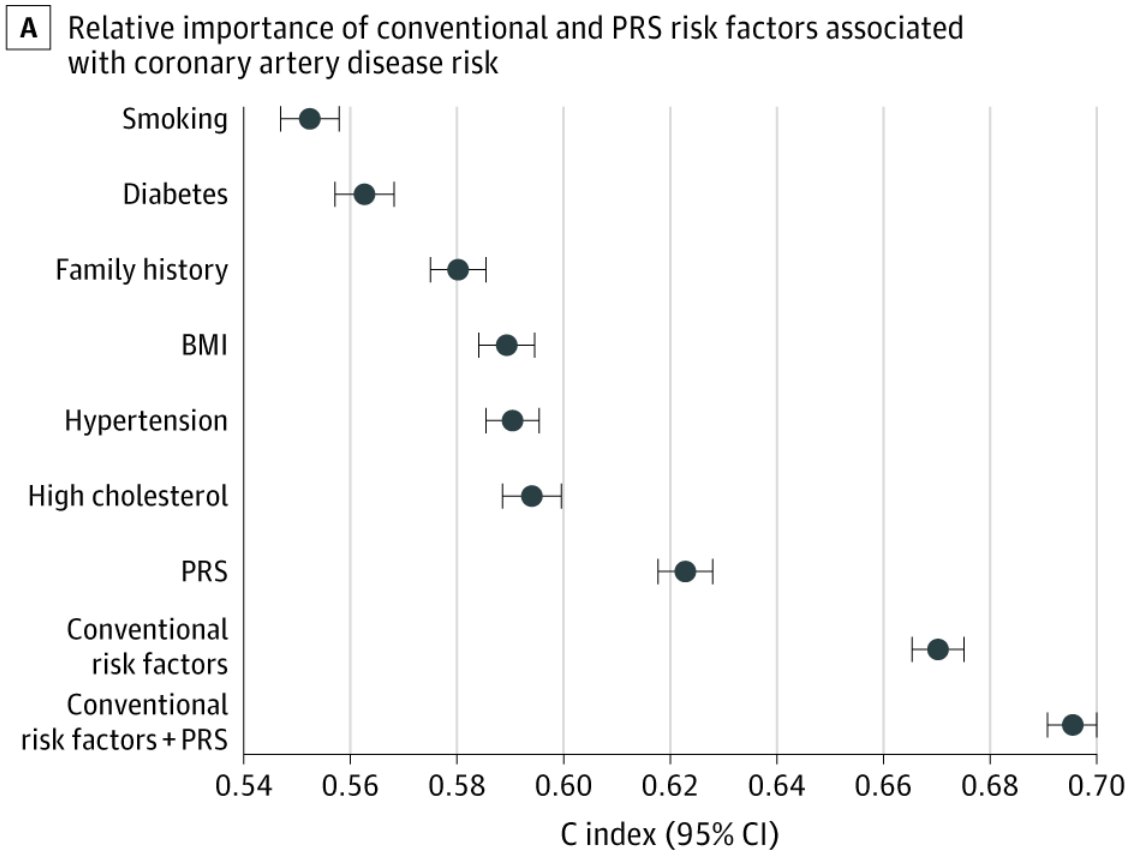
Estimated effect size

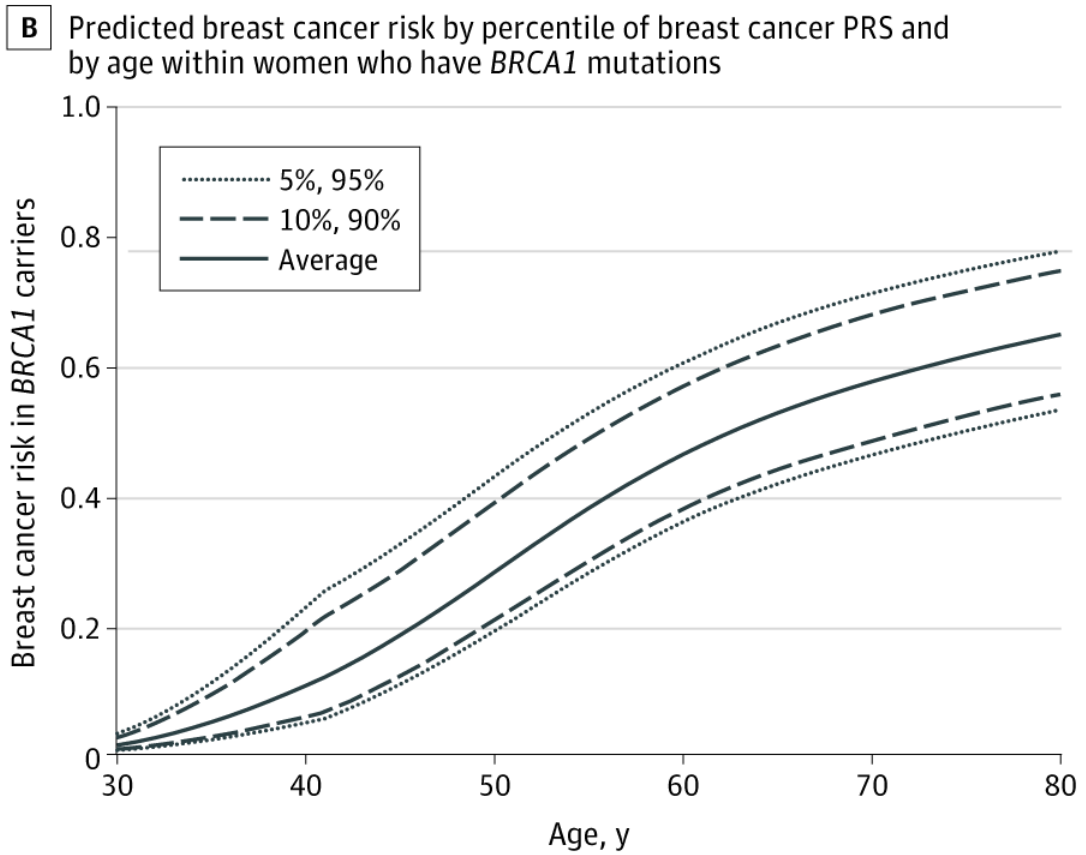


Choi et al., Nature Protocols (2020)

https://choishingwan.github.io/PRS-Tutorial/

Potential clinical utility of PRS





Wray et al., JAMA Psychiatry (2020)

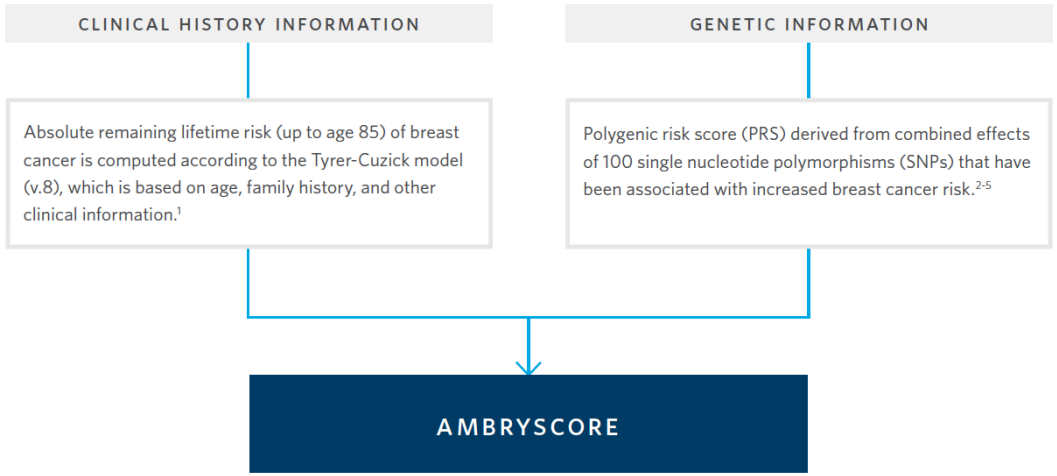
PRS are currently being implemented by genetic testing companies

Myriad: PRS for Breast Cancer

Eligibility:

1. Woman under age 85
2. **European and Ashkenazi Jewish ancestry**
3. No personal history of breast cancer
4. The woman does not have a mutation in a breast cancer gene (excluding monoallelic *CHEK2*)
5. The woman's relatives have not been found to have a mutation in a high-penetrance breast cancer risk gene.

Ambry Genetics: PRS for Prostate Cancer



Eligibility:

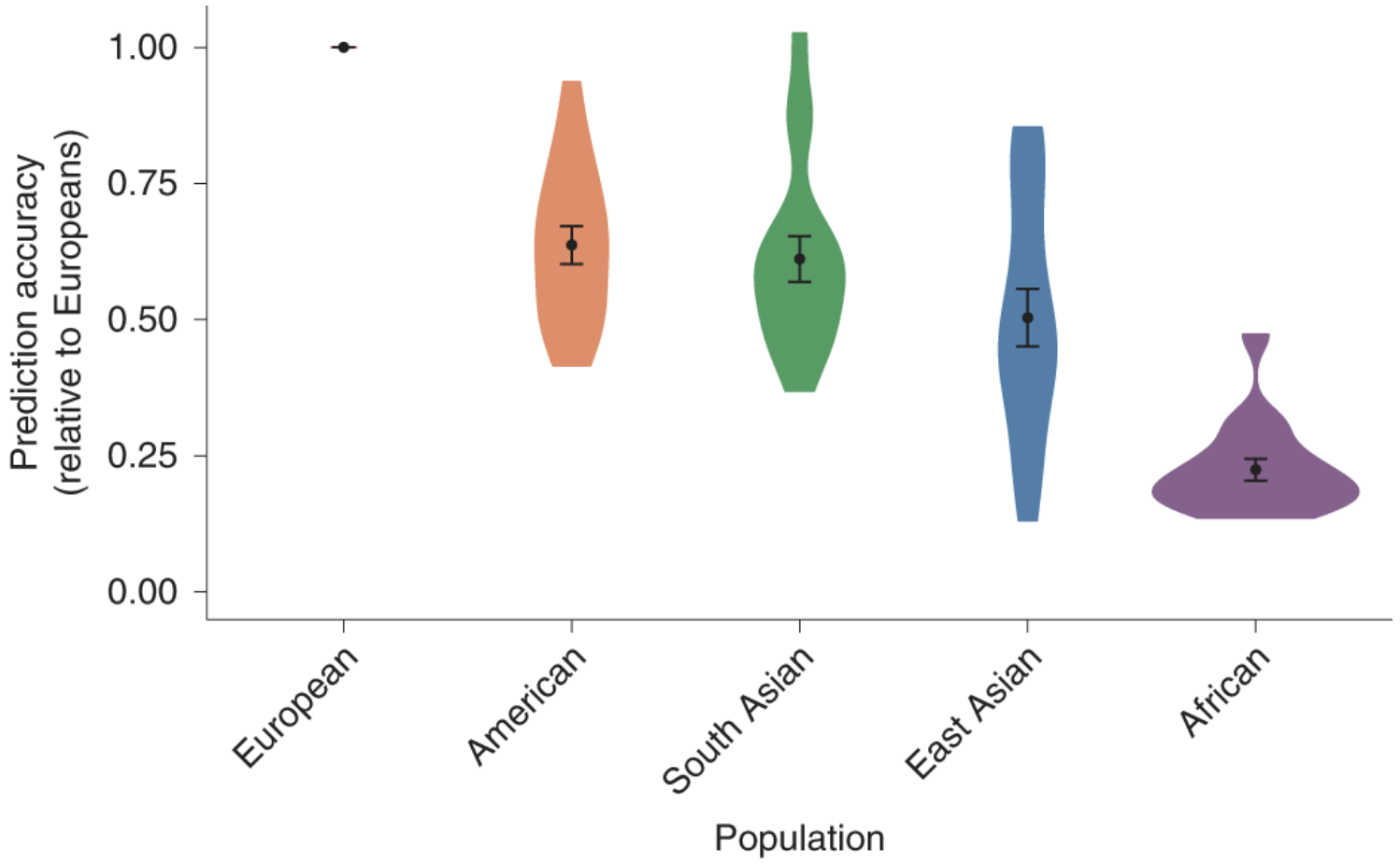
1. Male biological sex
2. 18-84 years old
3. **Northern European ancestry**
4. No personal or family history of a mutation in a prostate cancer susceptibility gene

PRS developed in Europeans have decreased accuracy when applied to non-European populations

GWAS representation compared to percentage of world population

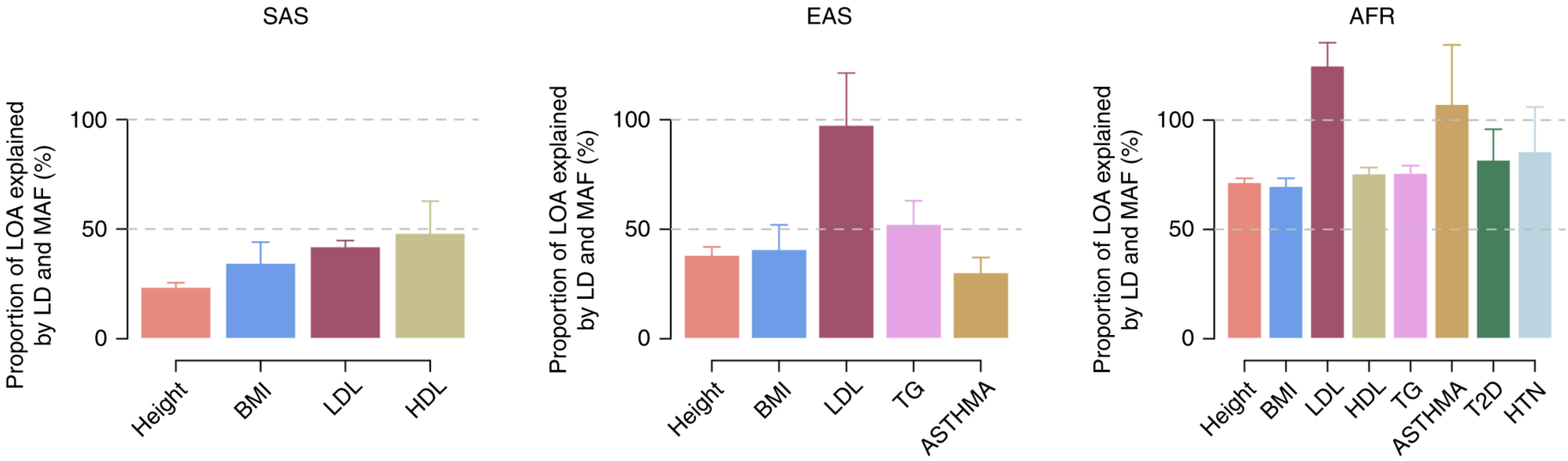
European:	460%
Asian:	40%
Latino:	19%
African:	17%

Duncan et al., Nat. Comm. (2019), PMID: 31346163



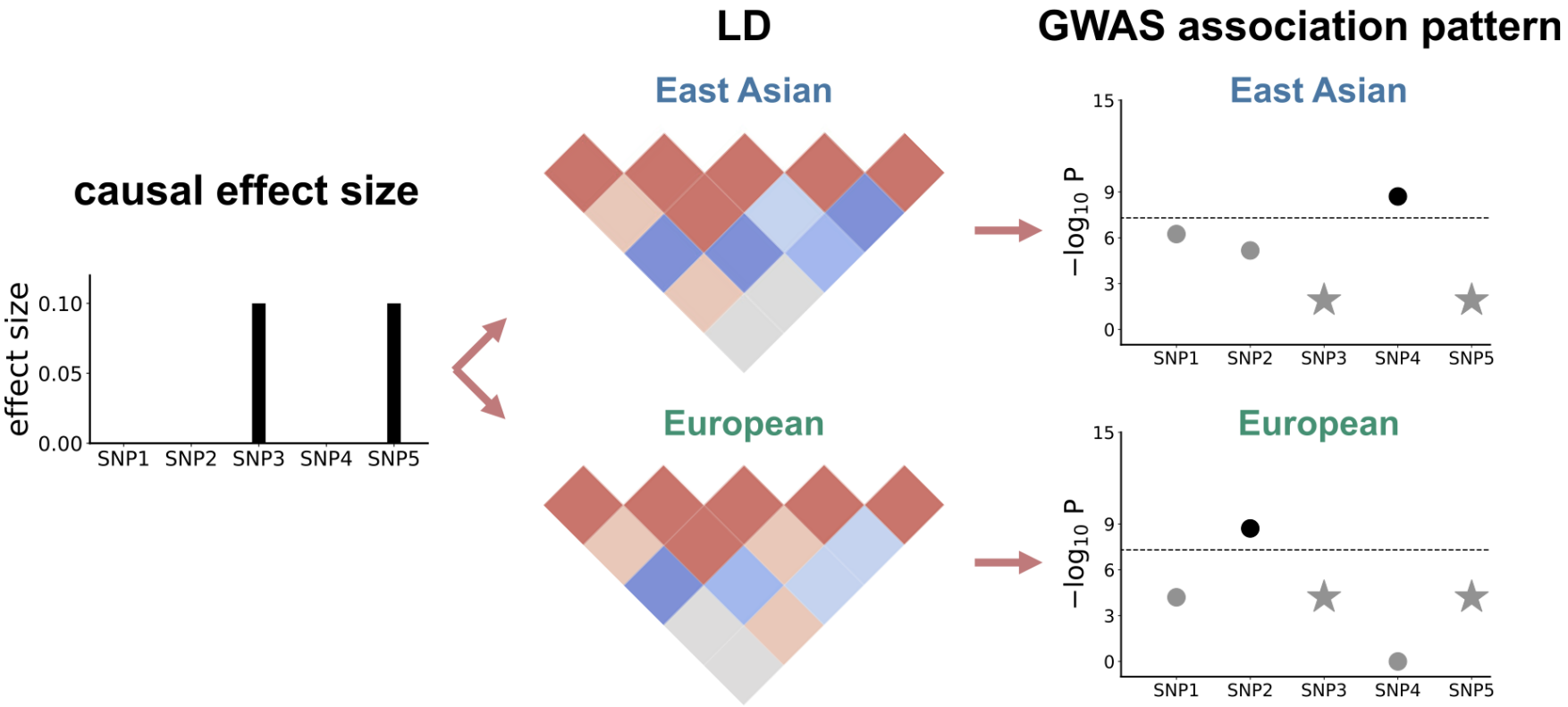
Martin et al., Nat. Genet. (2019)

Loss of PRS accuracy in non-Europeans can be explained by population differences in LD and MAF



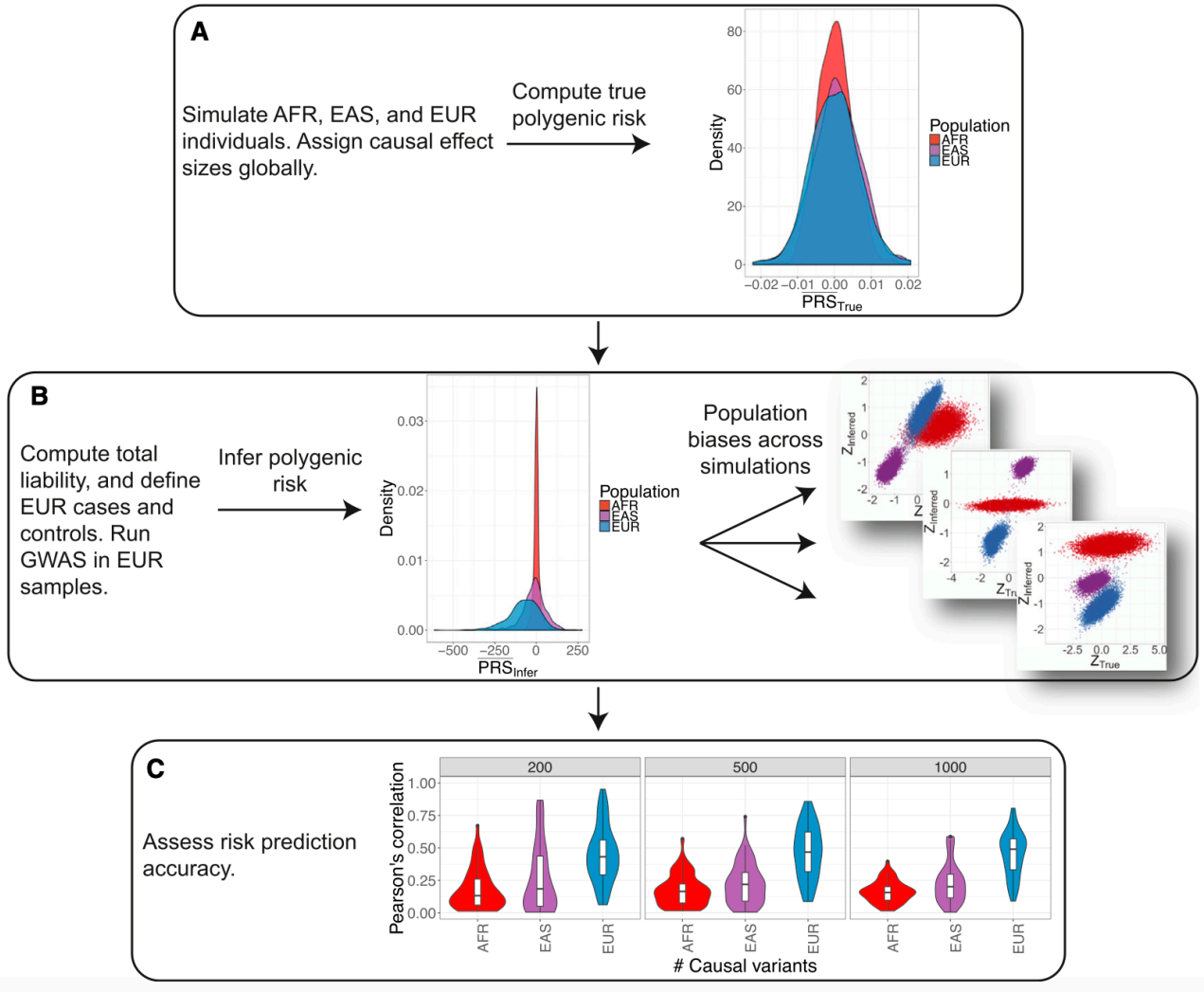
Wang et al., Nat Comm (2020)

Most causal variants are expected to be shared across populations



Shi et al., Nat Comm (2020), PMID: 32442408

Simulation framework to understand limited PRS portability across populations

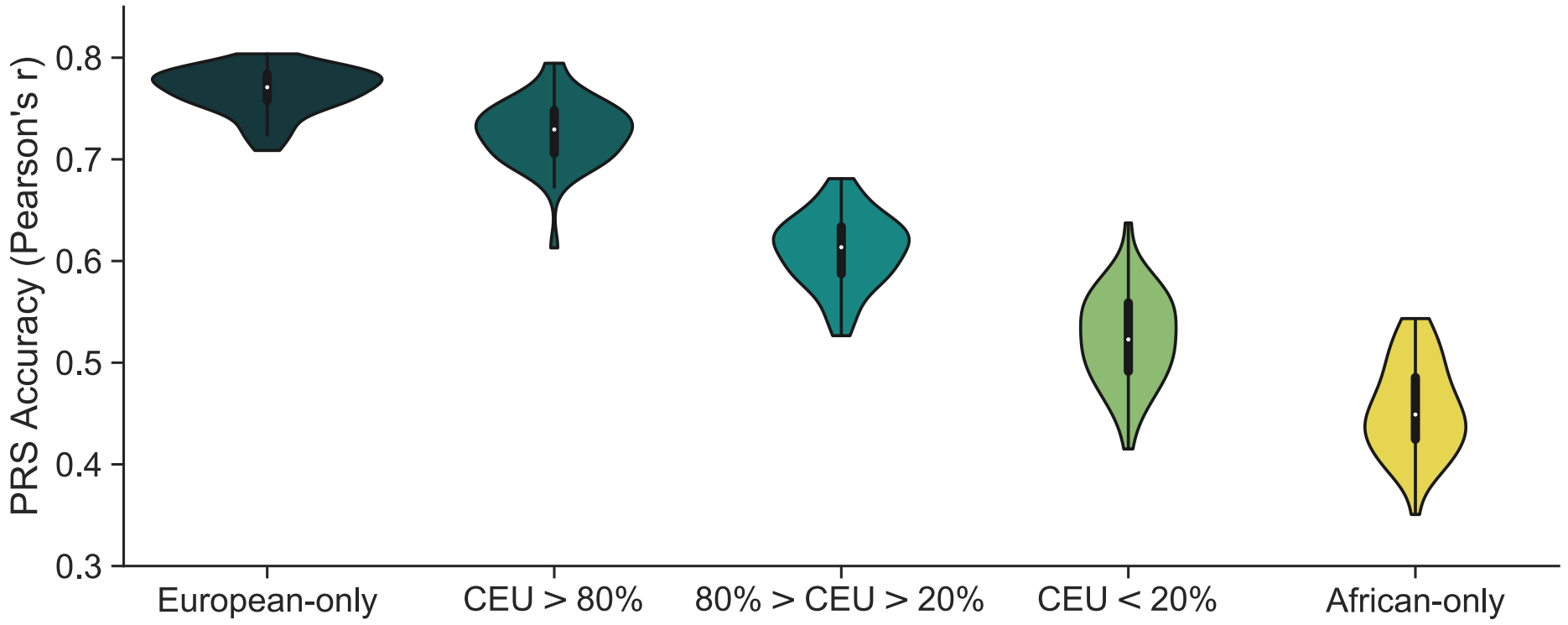


Main Differences:

1. We incorporate admixture.
2. Test multiple PRS methods.
3. Explore the impact of sample size.

Martin et al. AJHG (2017)

European-derived PRS performance decreases with increasing proportion of African ancestry



Cavazos and Witte, HGG Adv (2020)

Population-specific weighting of European GWAS variants has limited improvement in PRS accuracy

PRS Weighting Strategies:

(1) European ancestry GWAS

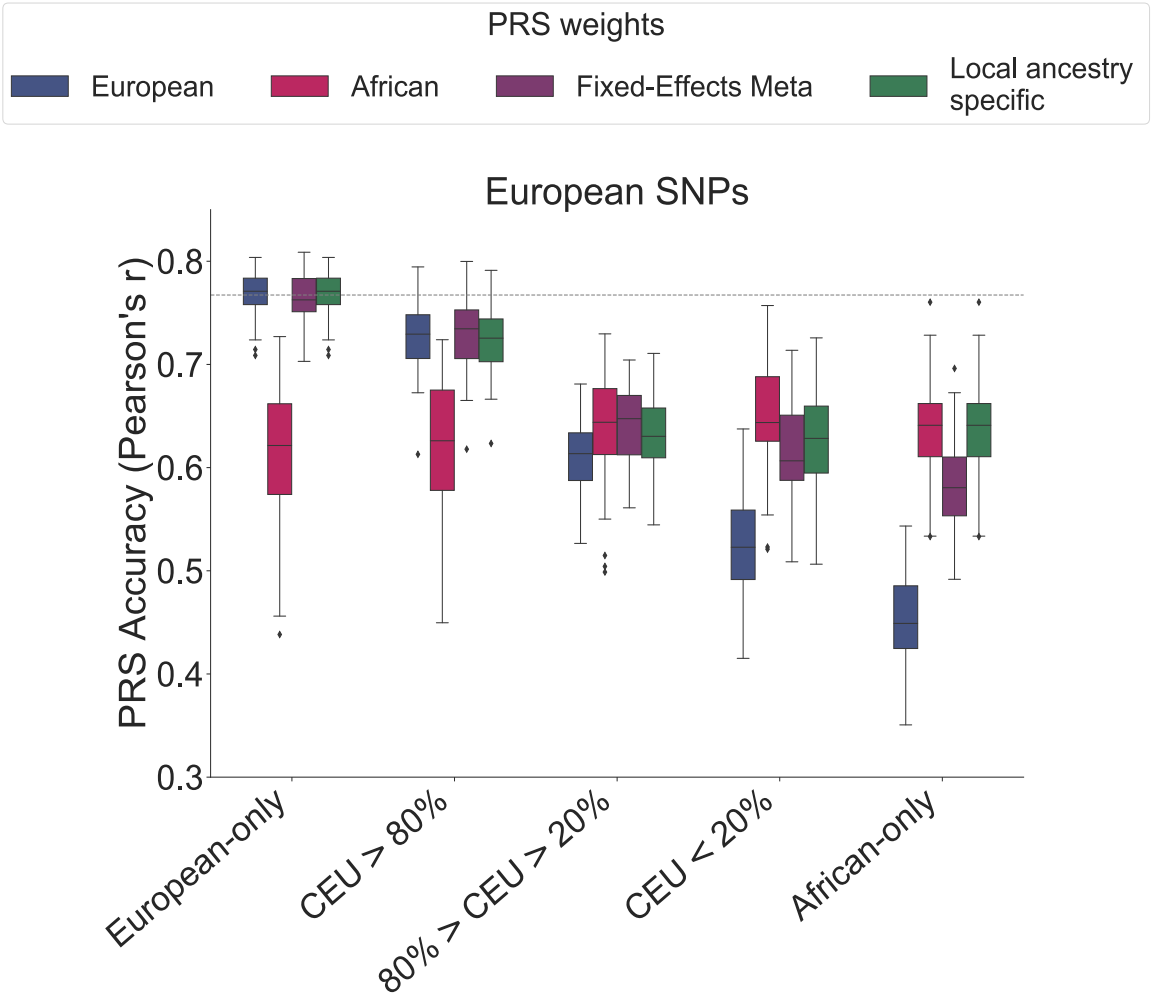
(2) African ancestry GWAS

(3) Fixed-effects meta

(4) Local ancestry specific weighting

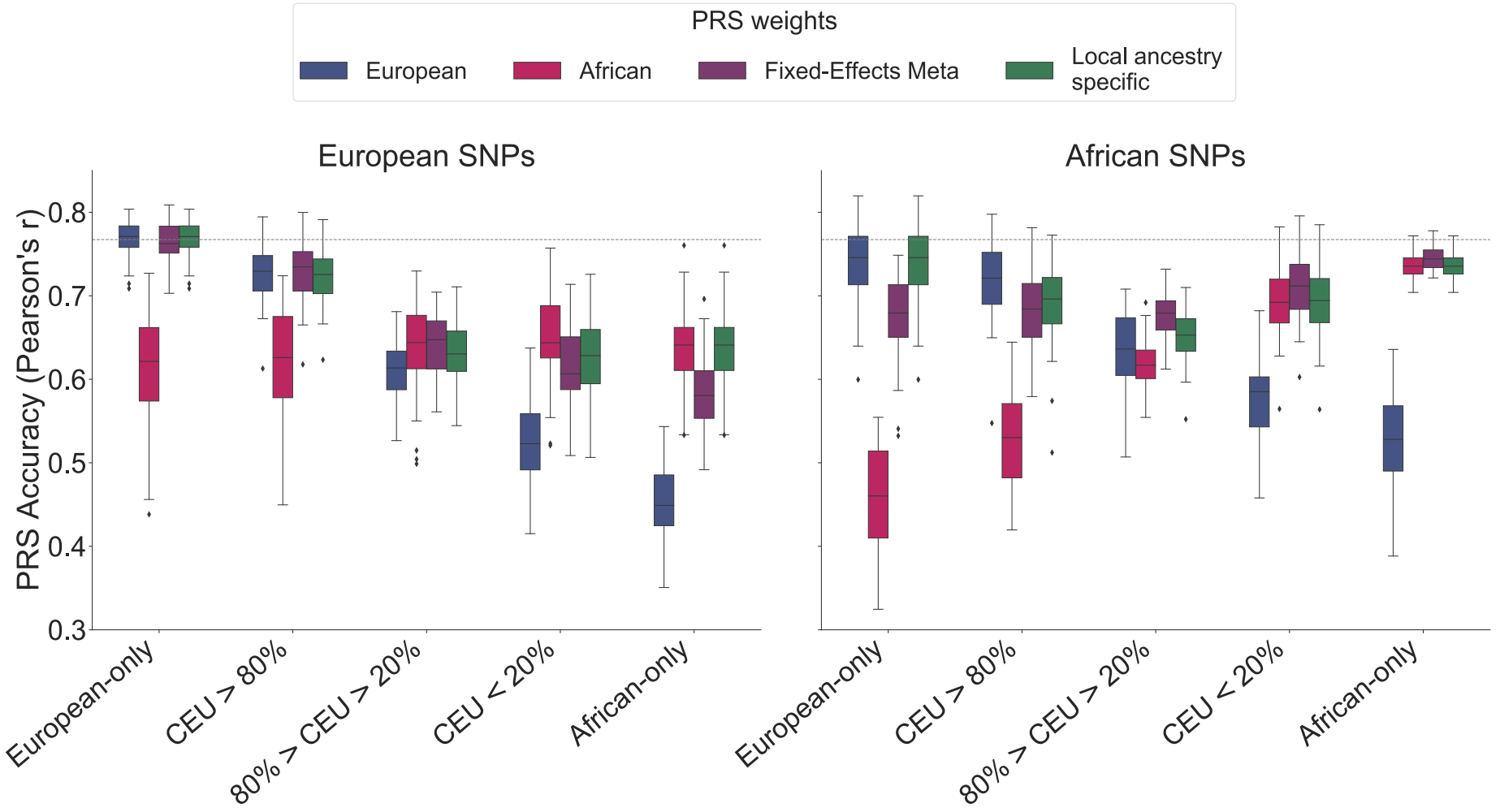
$$PRS = \rho_{EUR} \sum_{i \in EUR} \beta_{i,EUR} G_i + (1 - \rho_{EUR}) \sum_{i \in AFR} \beta_{i,AFR} G_i$$

Marnetto et al. Nat Comm (2020)



Cavazos and Witte, HGG Adv (2020)

PRS with African GWAS variants and population-specific weights performs accurately across ancestries



Cavazos and Witte, HGG Adv (2020)

Summary

- Studying the genetic diversity in African ancestry populations has the potential to advance scientific discovery of disease associated variants
- Risk prediction in diverse populations should not be based on European findings only
- Future data collection initiatives should prioritize the inclusion of diverse populations to ensure genetic findings do not further contribute the disparities in health care

Additional Resources

Race, Ethnicity, Ancestry and Medicine

Oni-Orisan et al., NEJM, 2021

Borrell et al., NEJM, 2021

Cooper et al., JAMA, 2018

Fine Mapping

Schaid et al., Nat Rev Genet, 2018

Atkinson et al., Nat Rev Genet, 2021

PRS Diverse Populations

Duncan et al., Nat Comm, 2019

Weissbrod et al., medRxiv, 2021

Questions?
taylor.cavazos@ucsf.edu