Population structure

Alicia Martin
Postdoctoral Research Fellow
Daly Lab
July 28, 2017
armartin@broadinstitute.org

My background

* Family
* BS at UW in bioengineering, *Drosophila* development
* PhD in Stanford genetics, MS in bioinformatics
* Broad for postdoc (2.5 yrs)




My advice

* **Identify great mentors**
* **Work on a problem that fascinates you**. (What intellectual concepts keep you up at night?)
* **Establish a mentorship committee**. Career goals? Areas of development? Milestones? Funding sources? “Soft” skills. Networking opportunities.
* **Use us!** Reach out! We really want this to be a lasting collaboration and envision you as the next generation of psychiatric genetics leaders in Africa.

European bias leaves vast genetic and phenotypic diversity undiscovered

* Popejoy et al: Non-European study participants increased 4% → 20% between 2009 → 2016. Mostly Asian, US minorities unchanged.

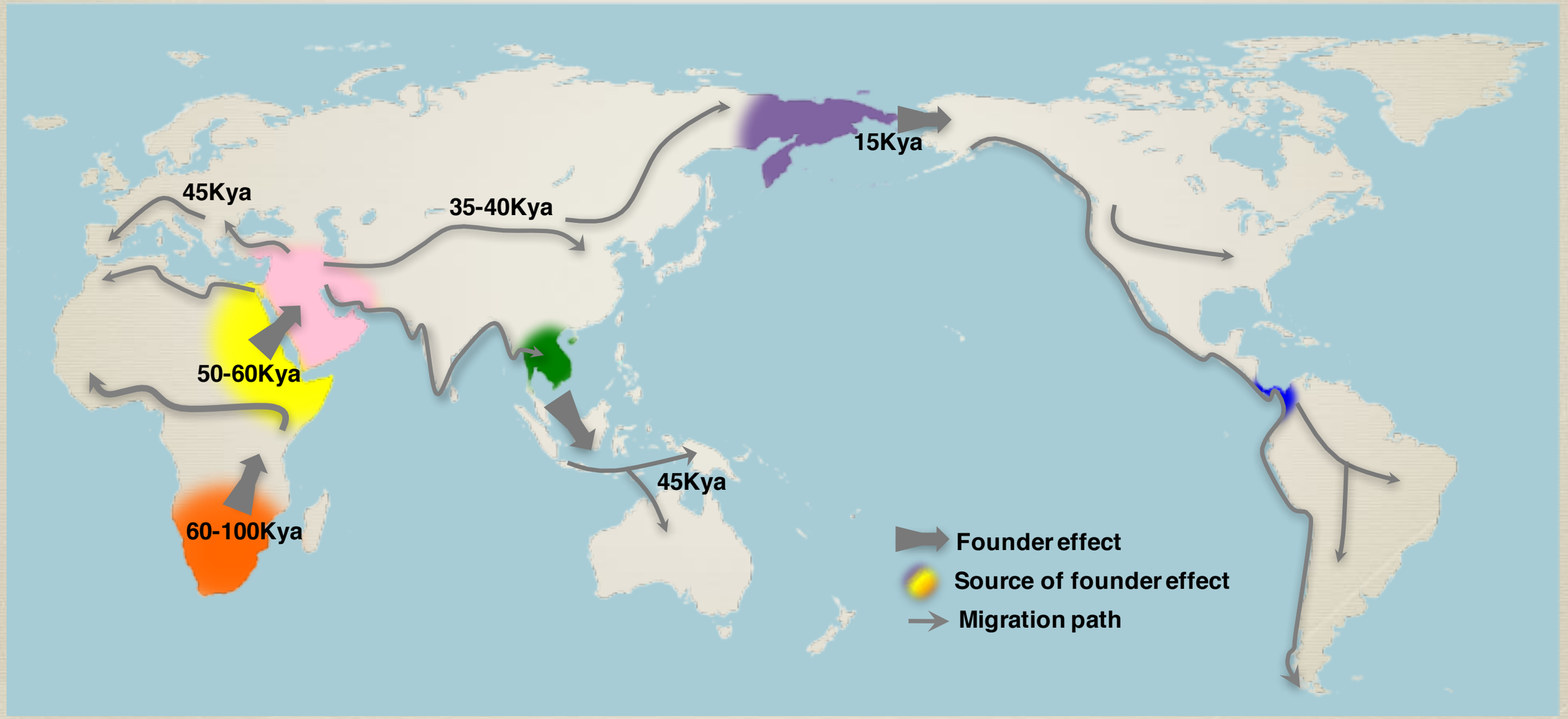
* Manrai et al: Allele frequency differences → genetic misdiagnoses of hypertrophic cardiomyopathy in African Americans

* ExAC: Europeans have the fewest homozygous loss-of-function variants, not helpful for disentangling disease role

Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature 538, 161–164.
Manrai, A.K., et al. (2016). Genetic Misdiagnoses and the Potential for Health Disparities. NEJM 375, 655–665.
Lek, M., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291.

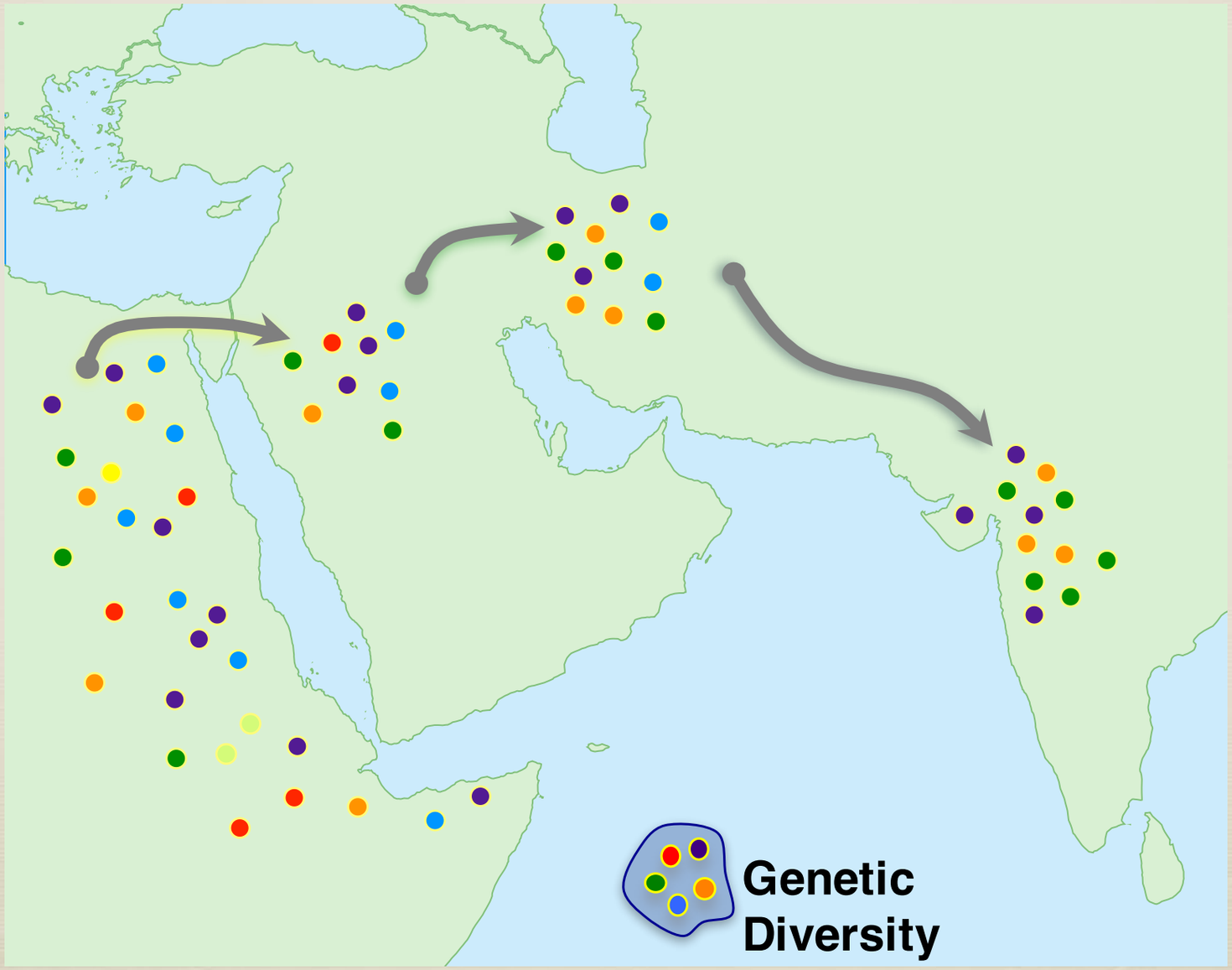
Serial founder effects

Historical human migration routes



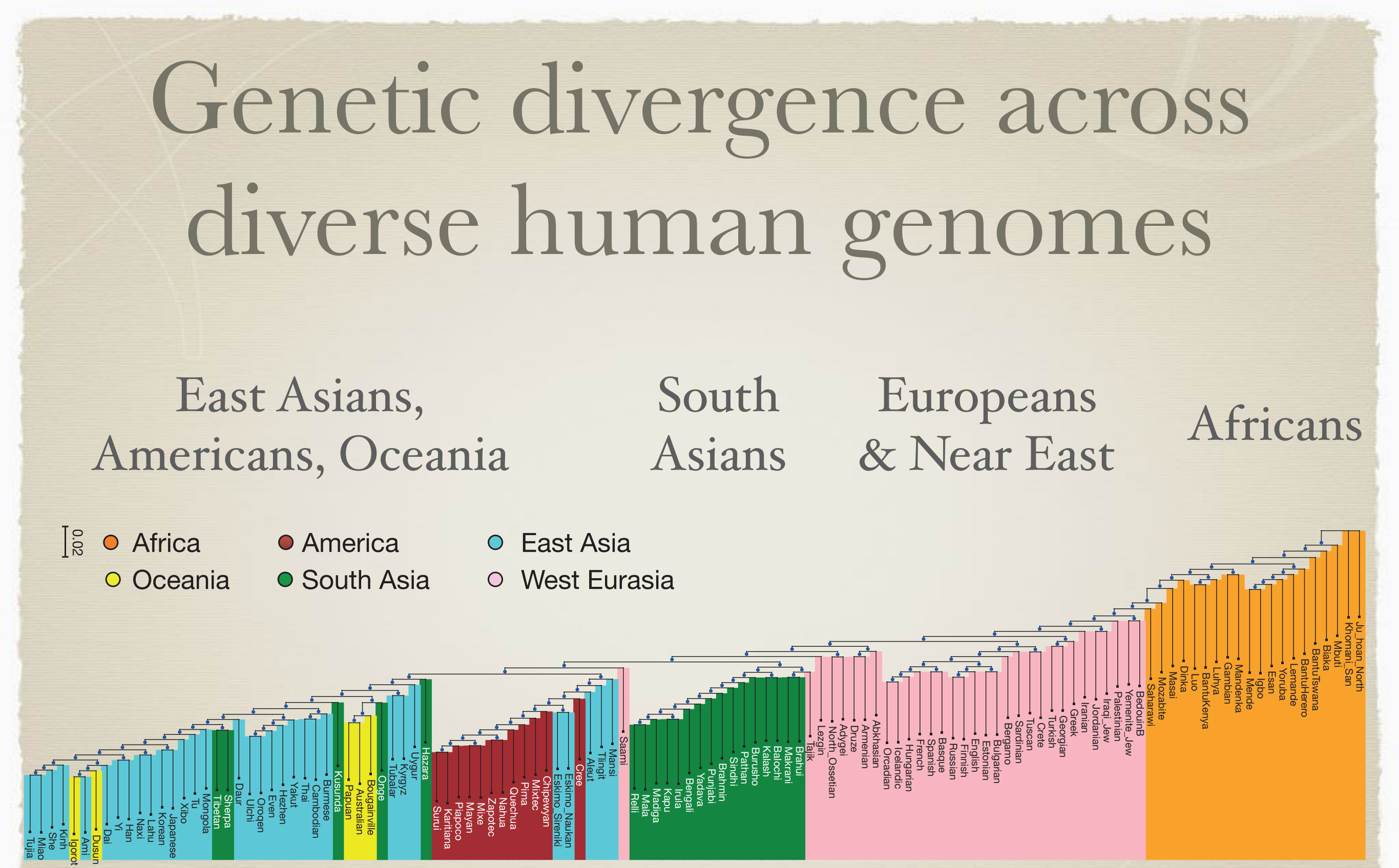
Henn, Cavalli-Sforza, and Feldman (2012) PNAS

Reduction in diversity due to serial founder effects



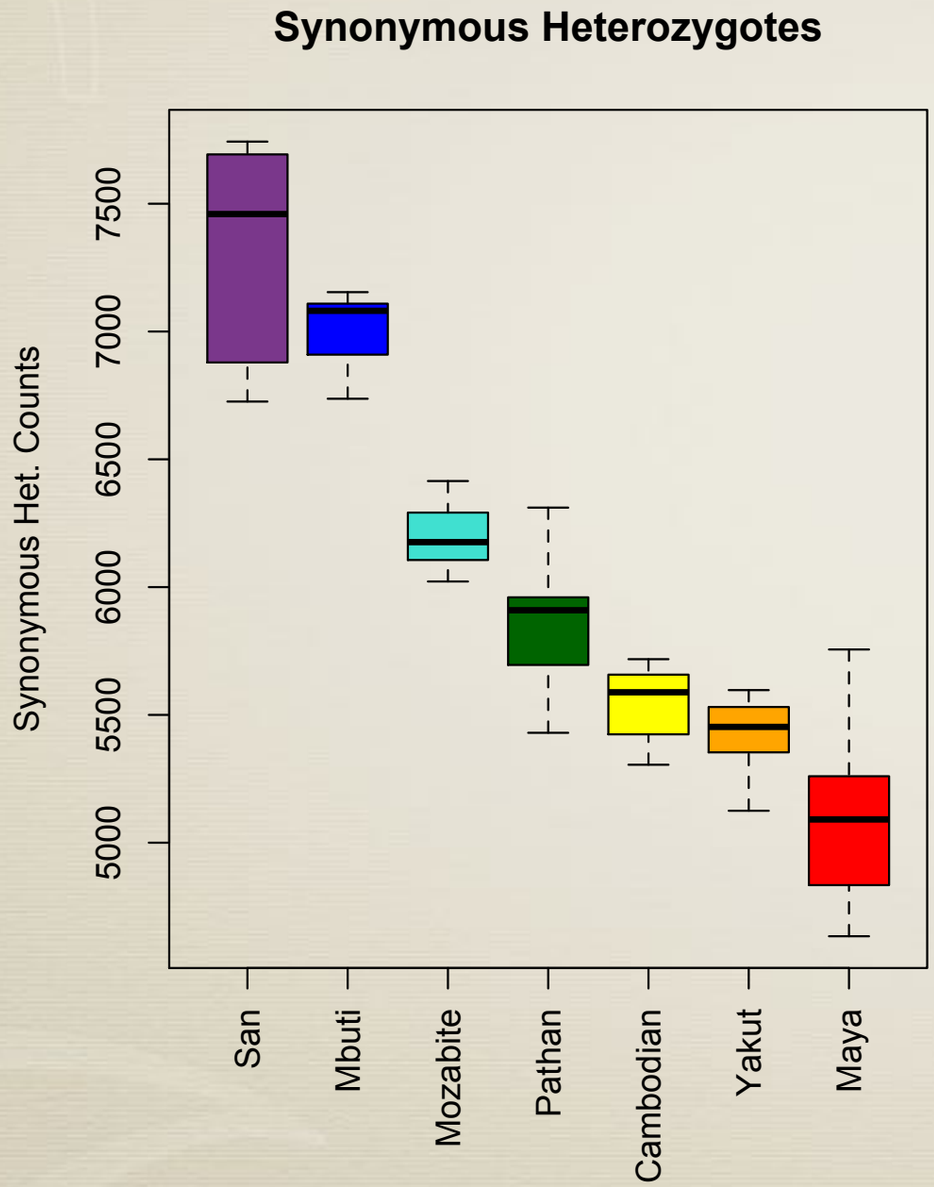
Henn, Cavalli-Sforza, and Feldman (2012) PNAS

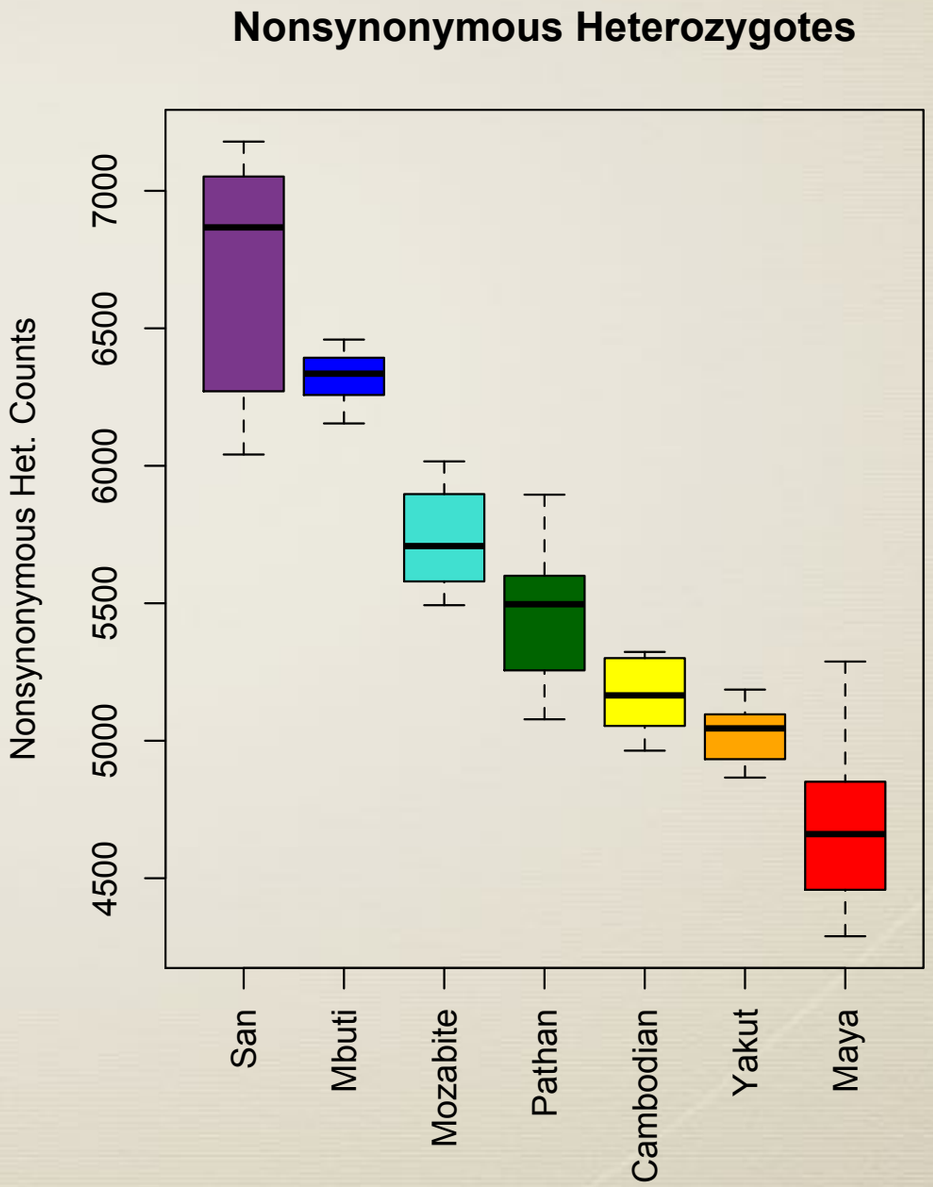
Genetic divergence across diverse human genomes



S Mallick et al. Nature 1–6 (2016) doi:10.1038/nature18964

Decline in heterozygosity out-of-Africa

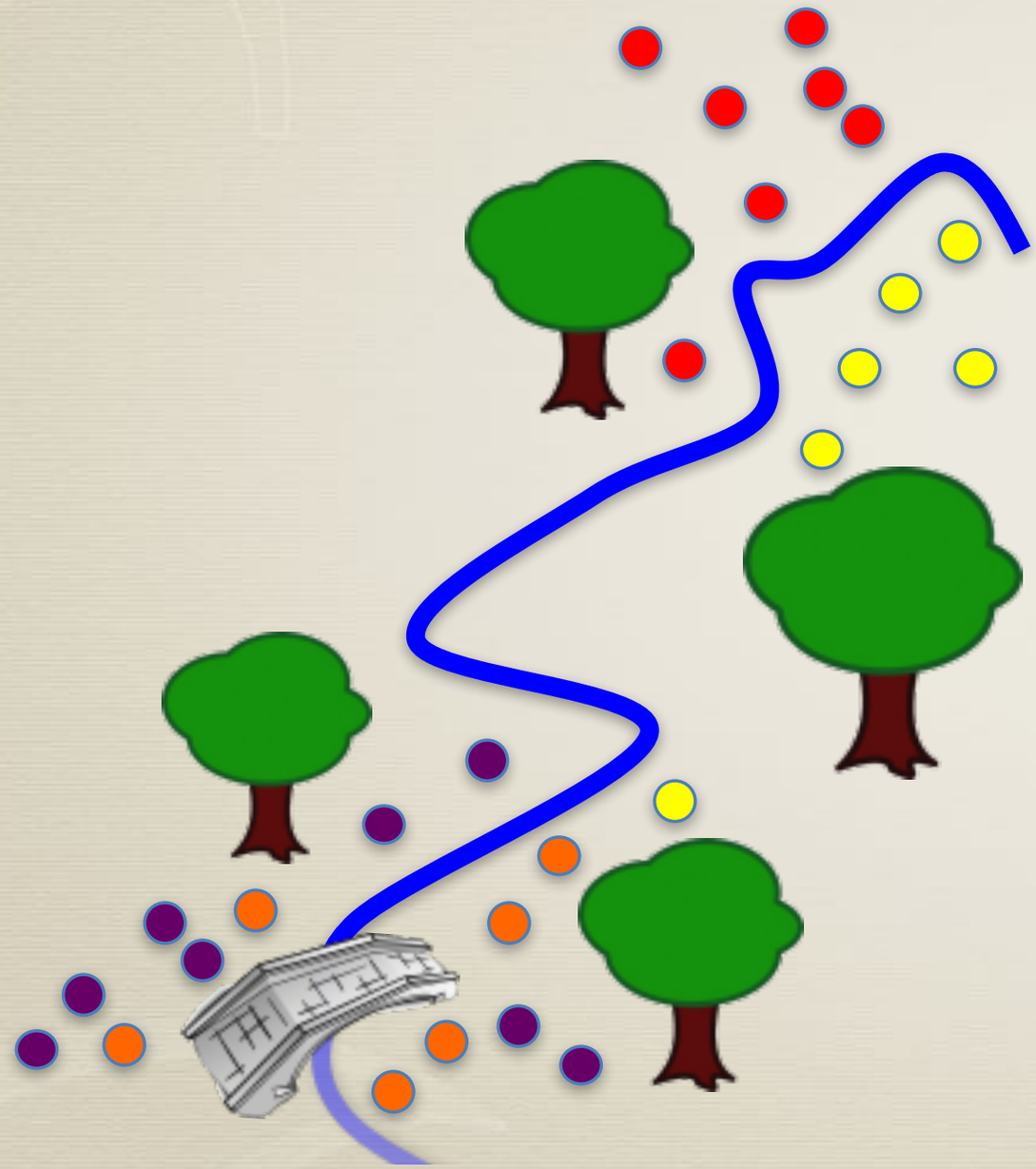




Henn, B.M., et al. (2016). PNAS. 113, E440-9.

Basic population structure

What is population structure?

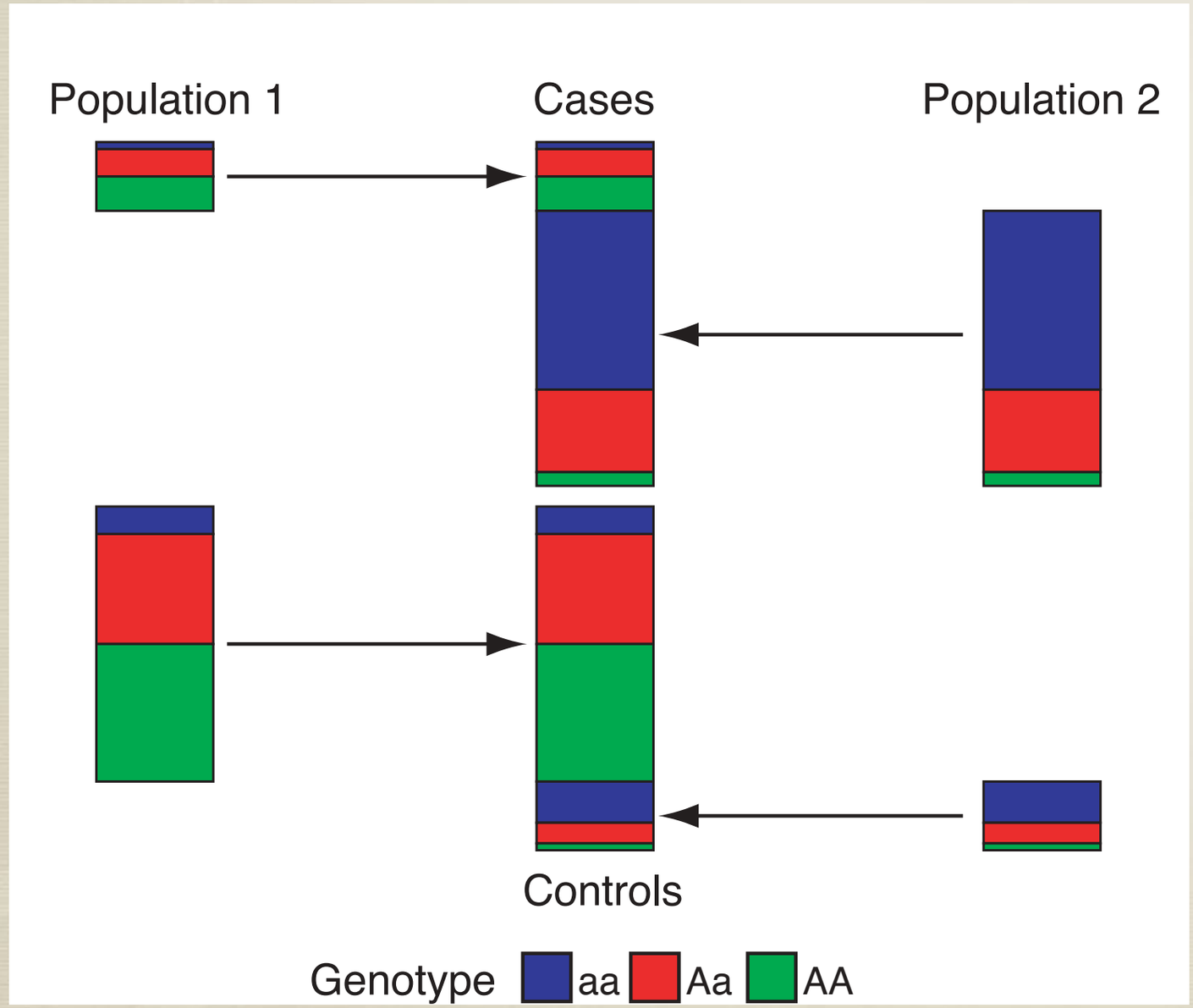


* Can be caused by multiple barriers to random mating: geography, language, ancestry

* Random mating is an important assumption in pop gen and stat gen models, usually assess population structure first

* Two commonly used methods of detecting structure are allele frequency-based clustering algorithms and principle component analysis

How does population stratification affect association analyses?



Disease more common in Population 2

- oversampling cases from this population relative to controls
- any allele that is more common in Pop 2 appears associated with the disease

Marchini et al., Nat Genet 2004

Population structure with clustering algorithms

I'm 80% red and 20% blue!

Each bar represents 1 individual. The number of colors is the number of potential ancestries. Proportion of different colors is the proportion of different ancestries for that individual

Continental ancestry

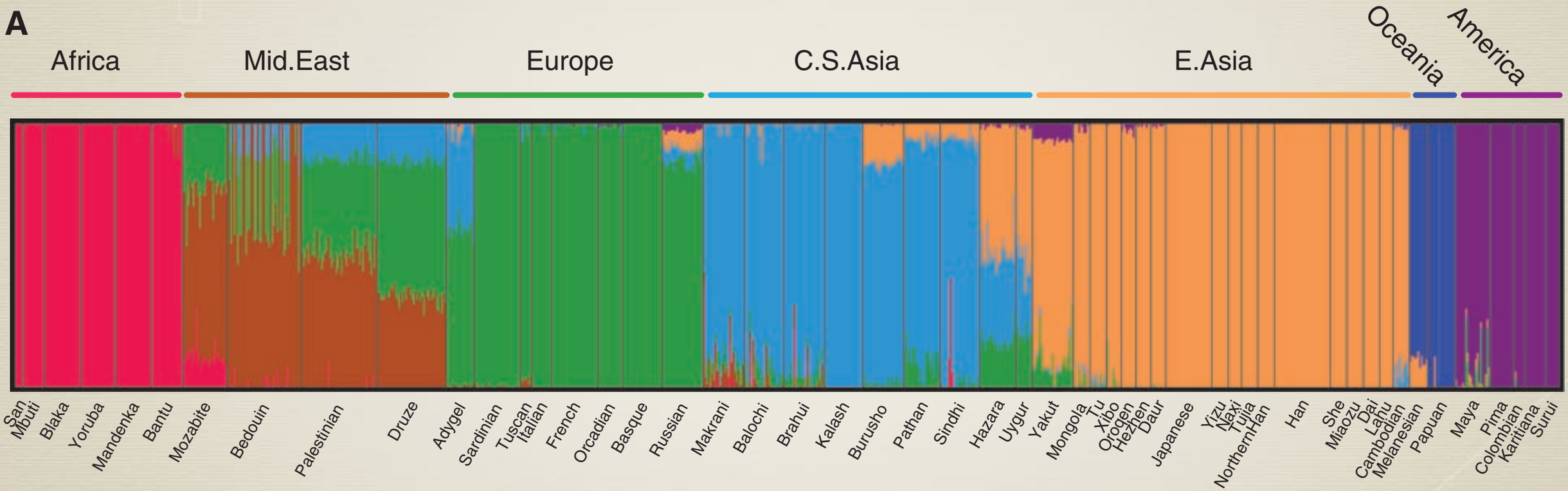
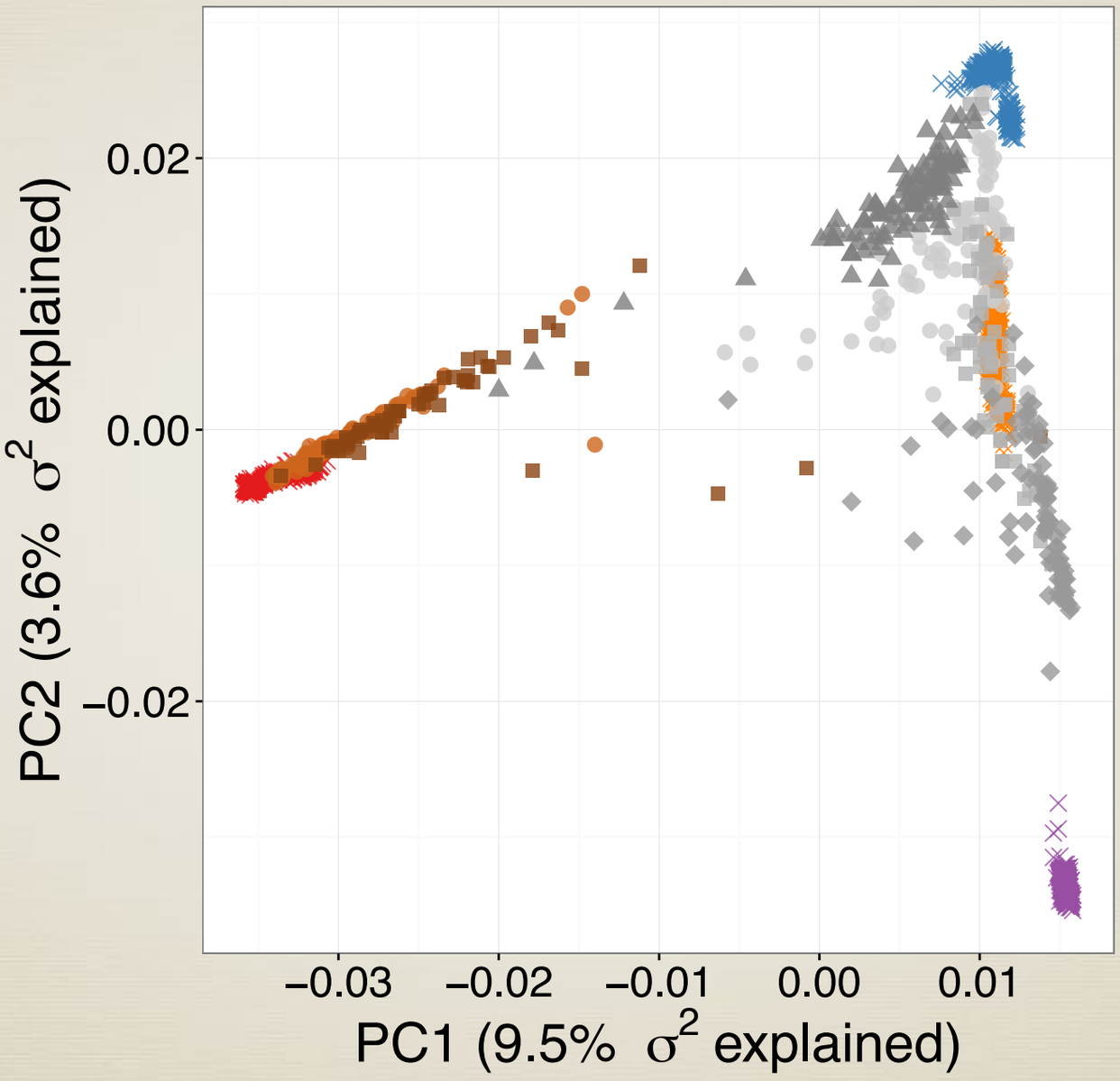


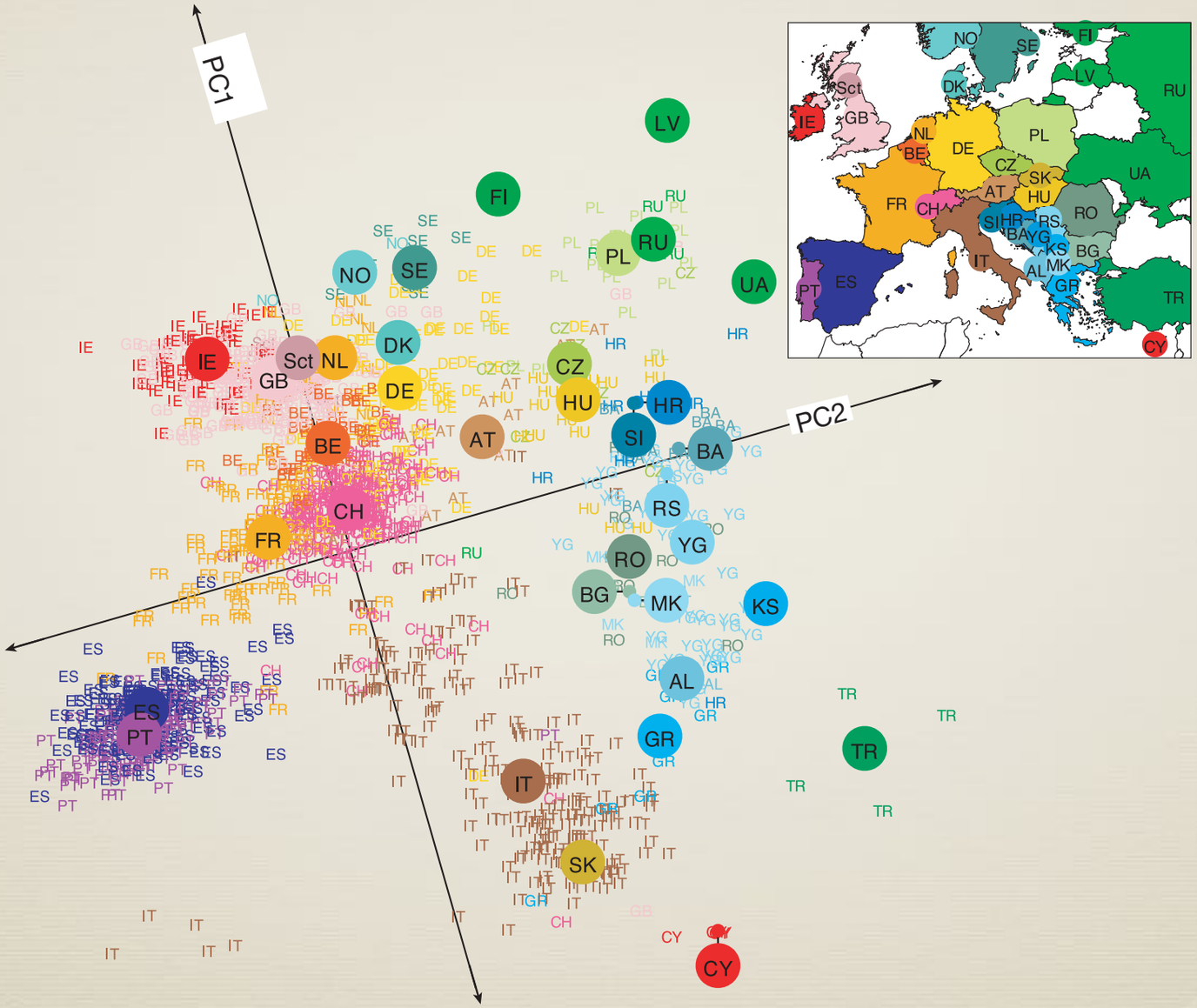
Fig. 1. Individual ancestry and population dendrogram. (**A**) Regional ancestry inferred with the *frappe* program at $K = 7$ (*13*) and plotted with the Distruct program (*31*). Each individual is represented by a vertical line partitioned into colored segments whose lengths correspond to his/her ancestry coefficients in up to seven inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in plotting.

Li, J.Z., et al. (2008). Science 319, 1100–1104.

Global PCA

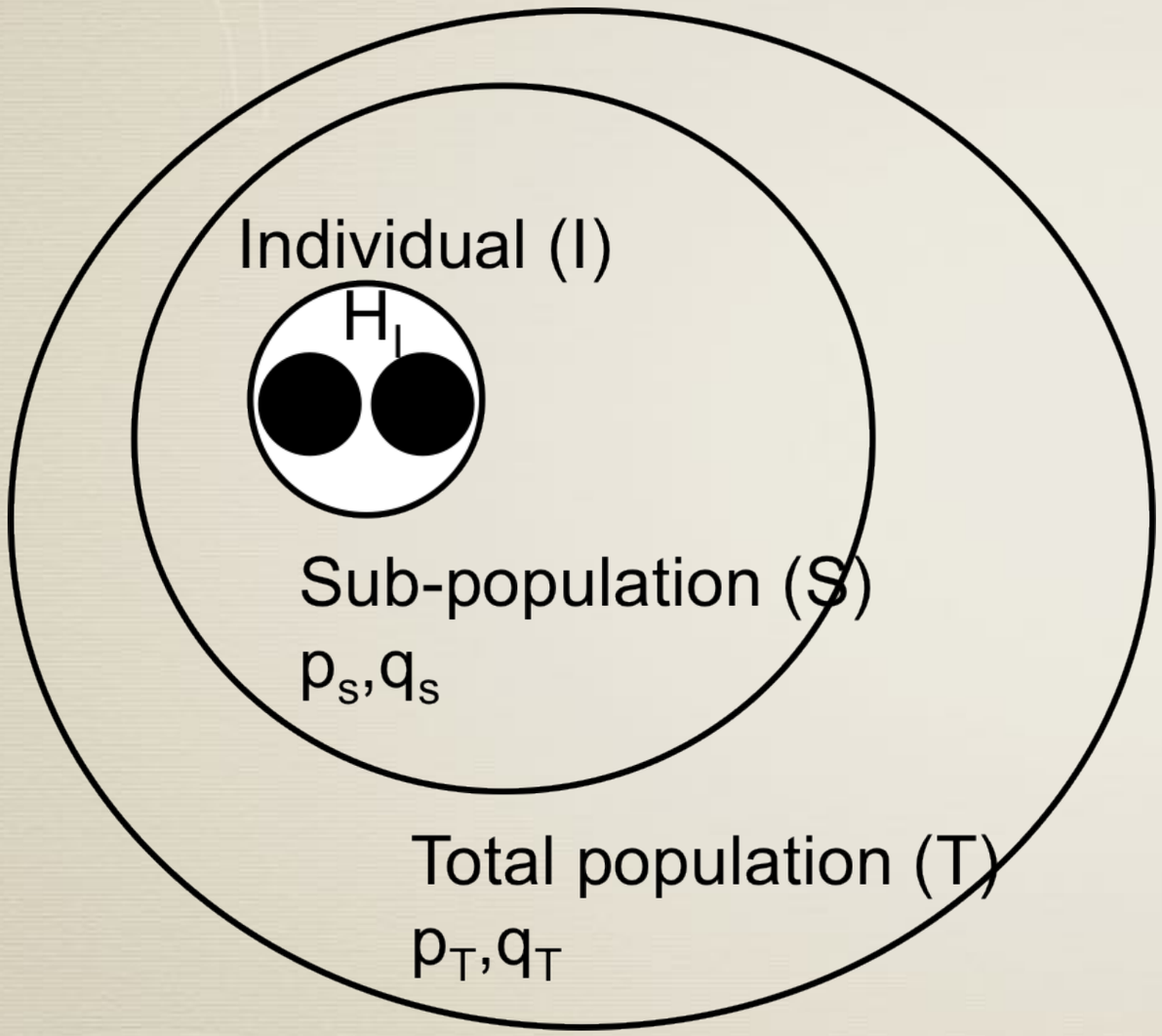

0.02
0.00
−0.02
PC2 (3.6% σ² explained)
−0.03
−0.02
−0.01
0.00
0.01
PC1 (9.5% σ² explained)
Reference panel
AFR
EUR
EAS
SAS
African Americans
ACB
ASW
Hispanic/Latinos
CLM
MXL
PEL
PUR

Genes mirror geography



Novembre, J., et al. (2008). Nature 456, 98–101.

Fixation index (F_{ST})



* Measures divergence across population pairs (S = subpopulations, T = total)

* H = heterozygosity

$$F_{ST} = 1 - \frac{H_S}{H_T}$$

$$= 1 - \frac{2p_S q_S}{2p_T q_T}$$

Graham Coop's pop gen notes:
http://bit.ly/2fEXzUe

How genetic structure changes

How does population structure change?

Changes in allele frequencies through time

* mutation
* migration
* natural selection
* genetic drift
* non-random mating

How does population structure change?

Changes in allele frequencies through time

* mutation
* migration
* natural selection
* genetic drift
* non-random mating

spontaneous change in DNA

Human mutation rate:
~1.2 x 10^{-8} / bp

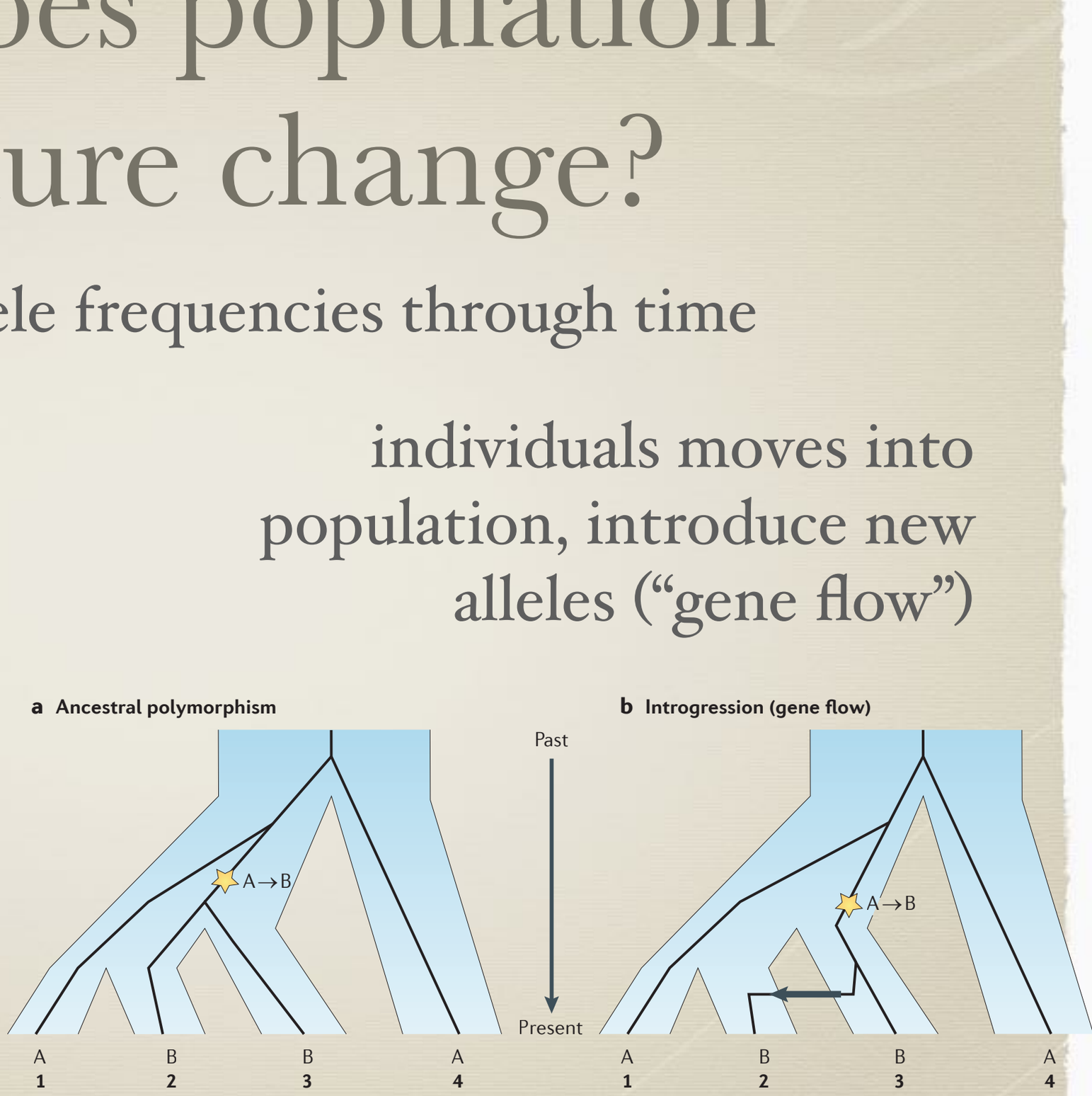
- ~80-100 total *de novo* variants
- <1 *de novo* coding variant

How does population structure change?

Changes in allele frequencies through time

* mutation
* migration
* natural selection
* genetic drift
* non-random mating

individuals moves into population, introduce new alleles (“gene flow”)

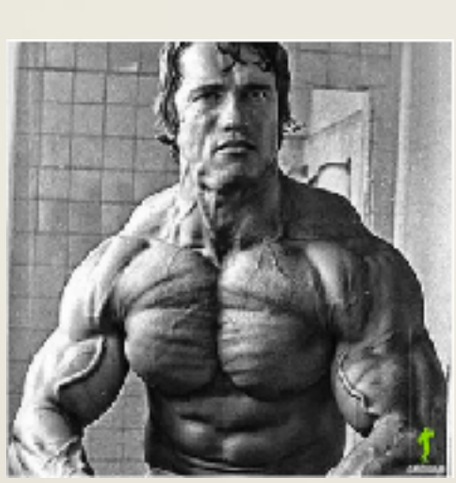


Sousa, V., and Hey, J. (2013). Nat. Rev. Genet. 14, 404–414.

How does population structure change?

Changes in allele frequencies through time

* mutation
* migration
* natural selection
* genetic drift
* non-random mating



certain genotypes produce more/less offspring

differences in survival and reproduction → differences in "fitness"

Many kinds: balancing (e.g. sickle-cell), positive (e.g. height), negative (most common), etc

Positive selection

- **High altitude**: convergent evolution in Tibet, the Andes, and Ethiopian highlands
- **Host-pathogen interactions**: Trypanosomes-African sleeping sickness, malaria-sickle cell
- **Arctic environment/diet**: Greenlandic population, FADS
- **Dairy consumption**: Lactase persistence
- **UV radiation**: skin pigmentation

Yi, X., et al. (2010). Science 329, 75–78.
Zhou, D., et al. (2013). AJHG. 1–11.
Alkorta-Aranburu, G., et al. (2012). PLoS Genet. 8, e1003110.
Genovese, G., et al. (2010). Science 329, 841–845.
McManus, K.F., et al. (2017). PLoS Genet. 13, 48–65.
Moltke, I., et al. (2014). Nature 512, 190–193.
Tishkoff, S. A., et al. (2007). Nat. Genet. 39, 31–40.
Martin, A. R., et al (2017) (submitted)

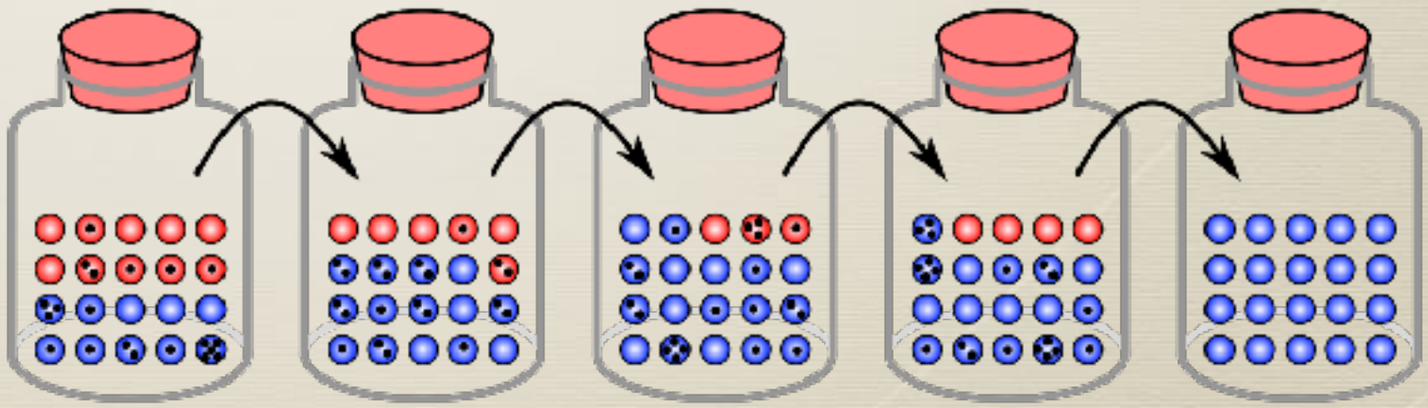
How does population structure change?

Changes in allele frequencies through time

* mutation
* migration
* natural selection
* genetic drift
* non-random mating

genetic change by chance alone

occurs in small populations



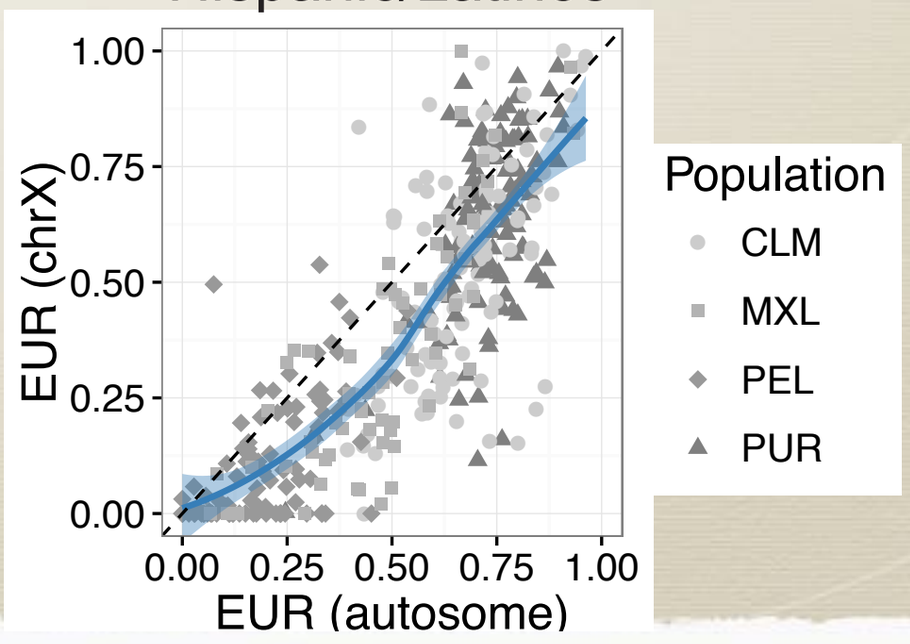
How does population structure change?

Changes in allele frequencies through time

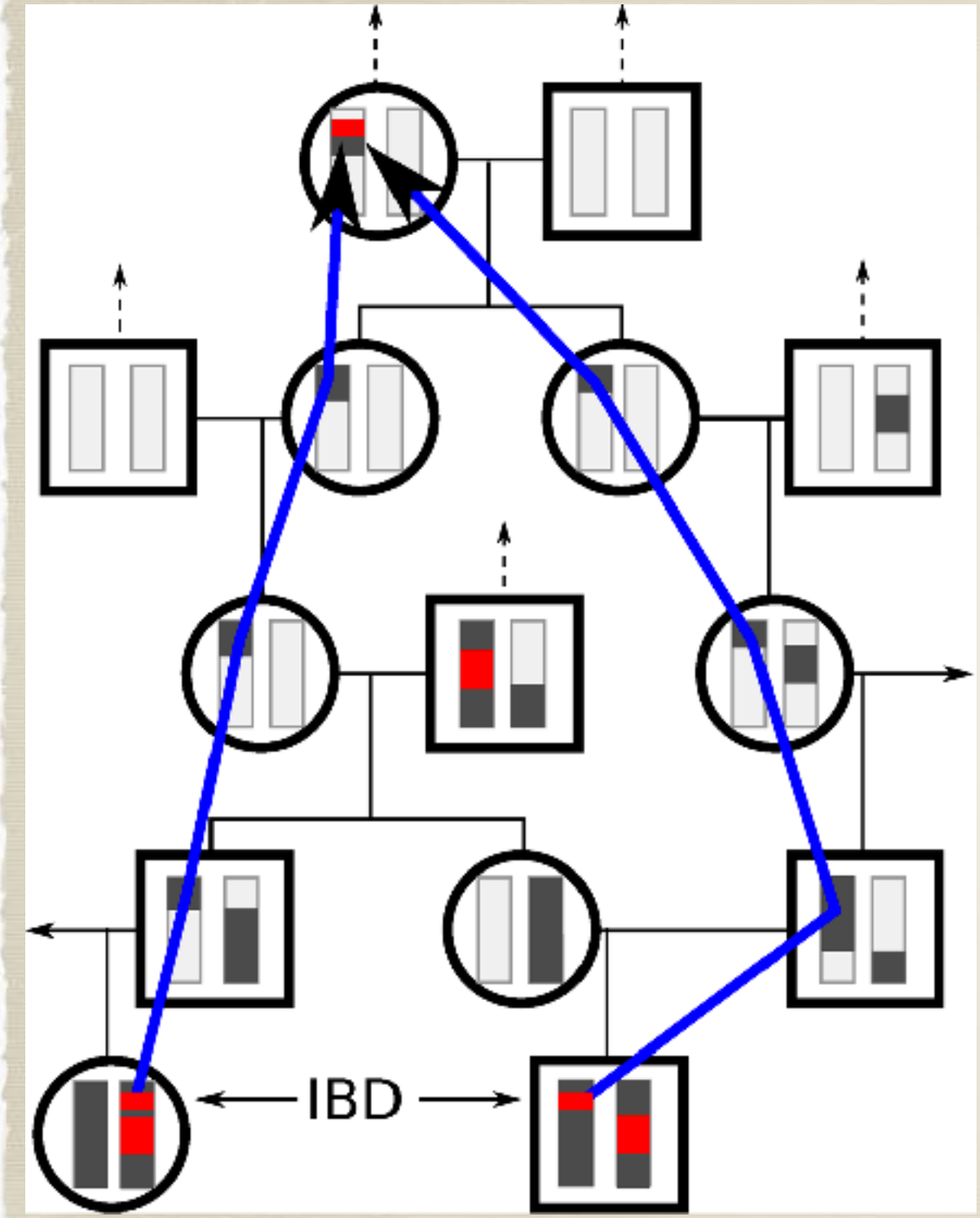
* mutation
* migration
* natural selection
* genetic drift
* non-random mating

assortative mating: mate with more similar type than random

Examples: education, height, skin color



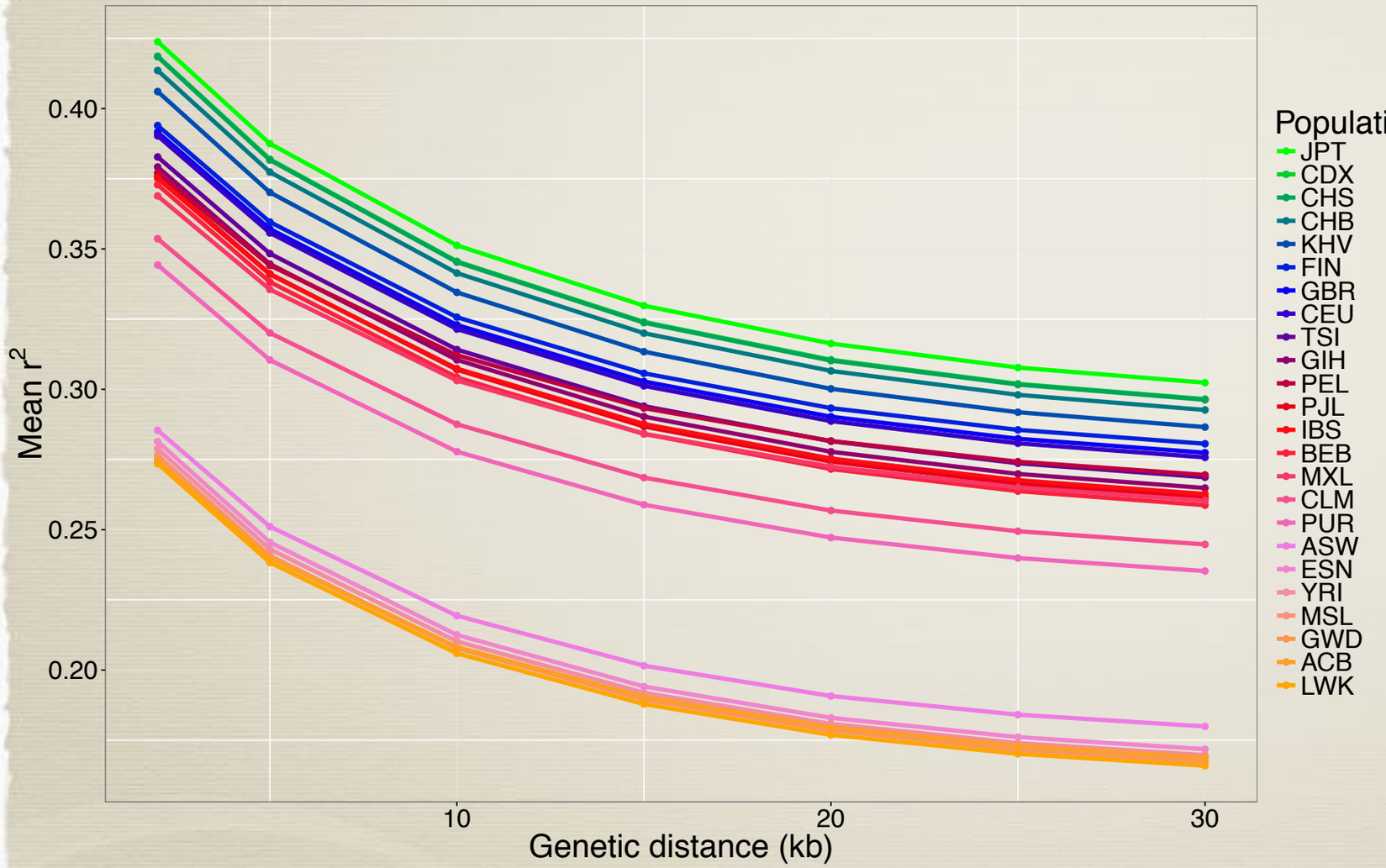
Linkage disequilibrium

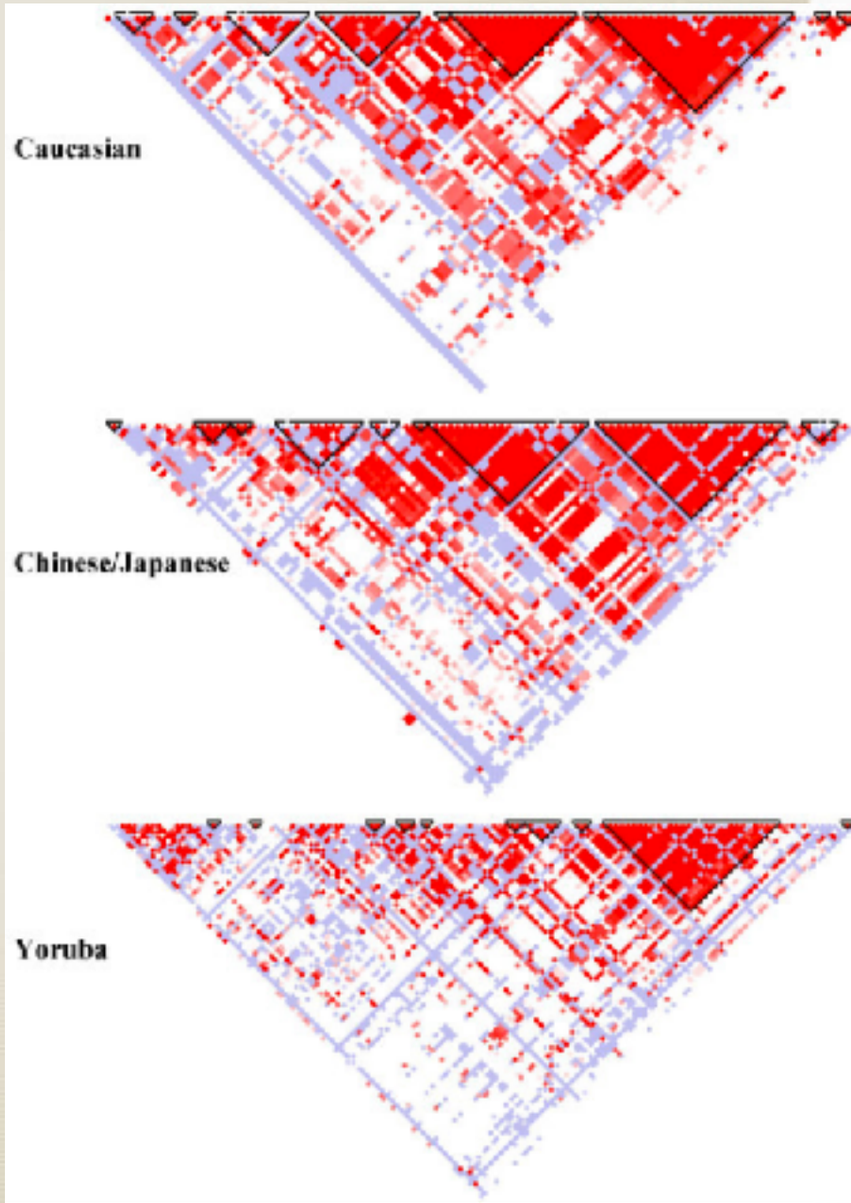


Linkage disequilibrium is the non-random association of alleles at different loci.

Recombination is the exchange of DNA between chromosomes, resulting in a new genetic combination that is different from parents.

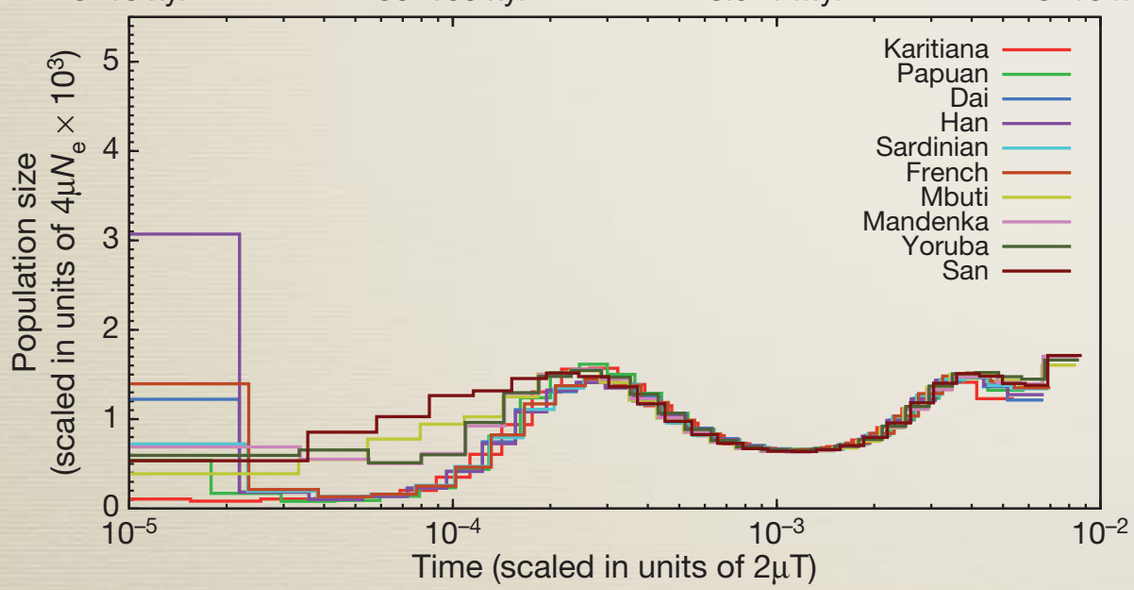
LD decay across 1000 Genomes populations

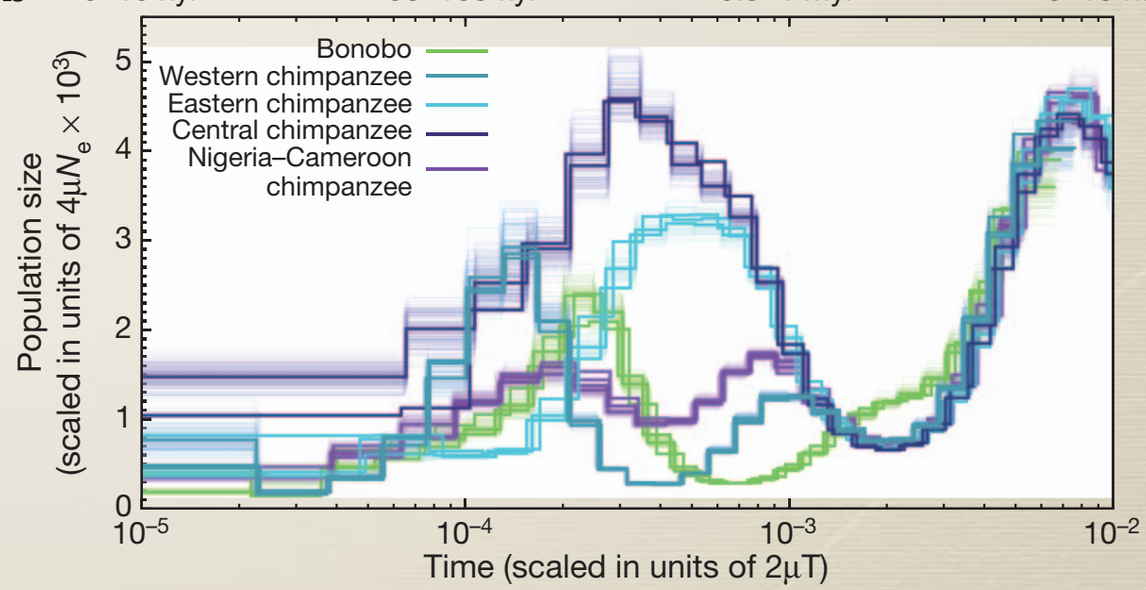




Effective population size

The **effective population size** (N_e) is the population size that would result in the same rate of drift in an idealized constant population size, obeying our modeling assumptions, as that observed in our true population.





Prado-Martinez, J., et al. (2013). Nature 1–5.

Methodological timeline for human N_e inference

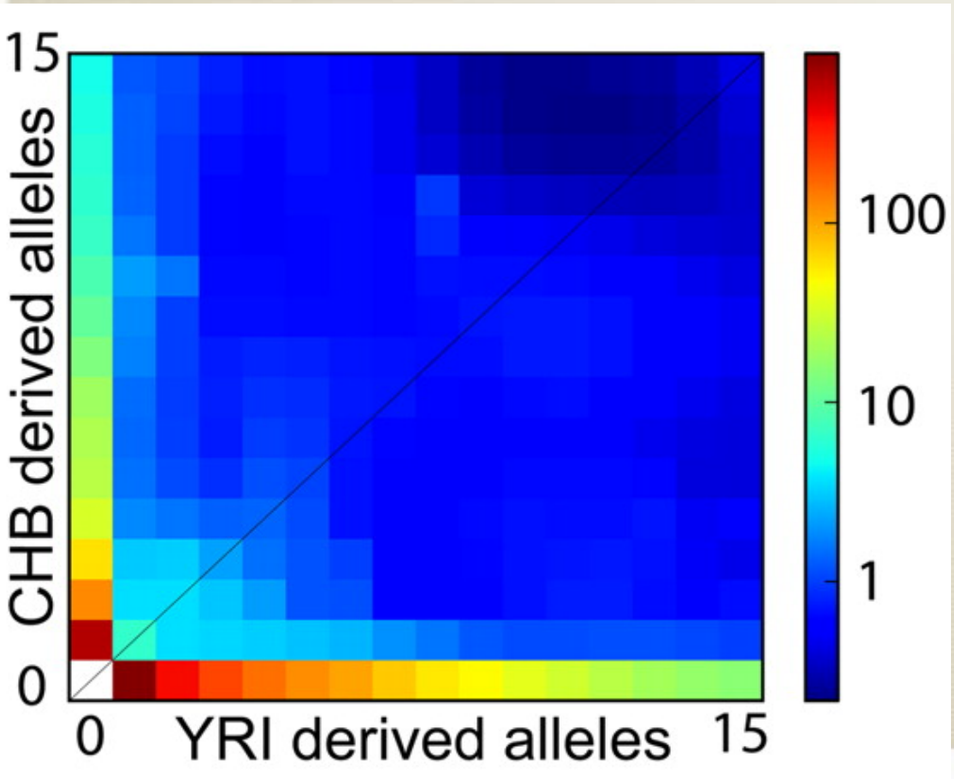
SFS

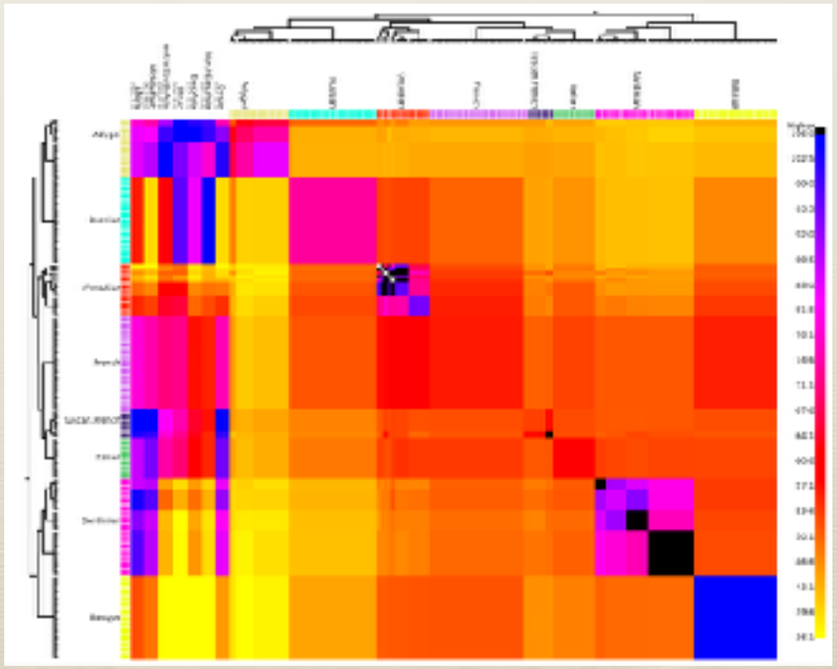
Pedigrees

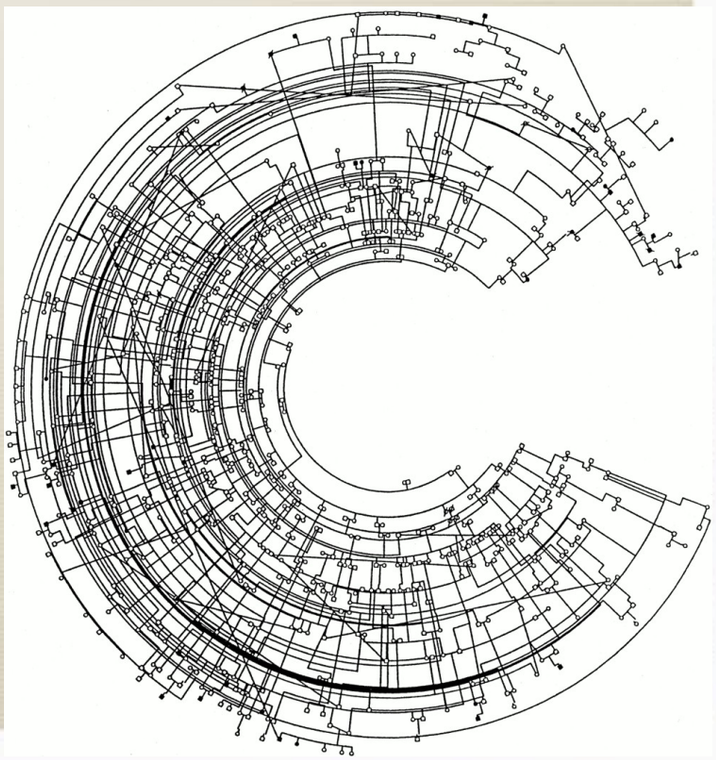
Haplotypes

1000 100 10 Present

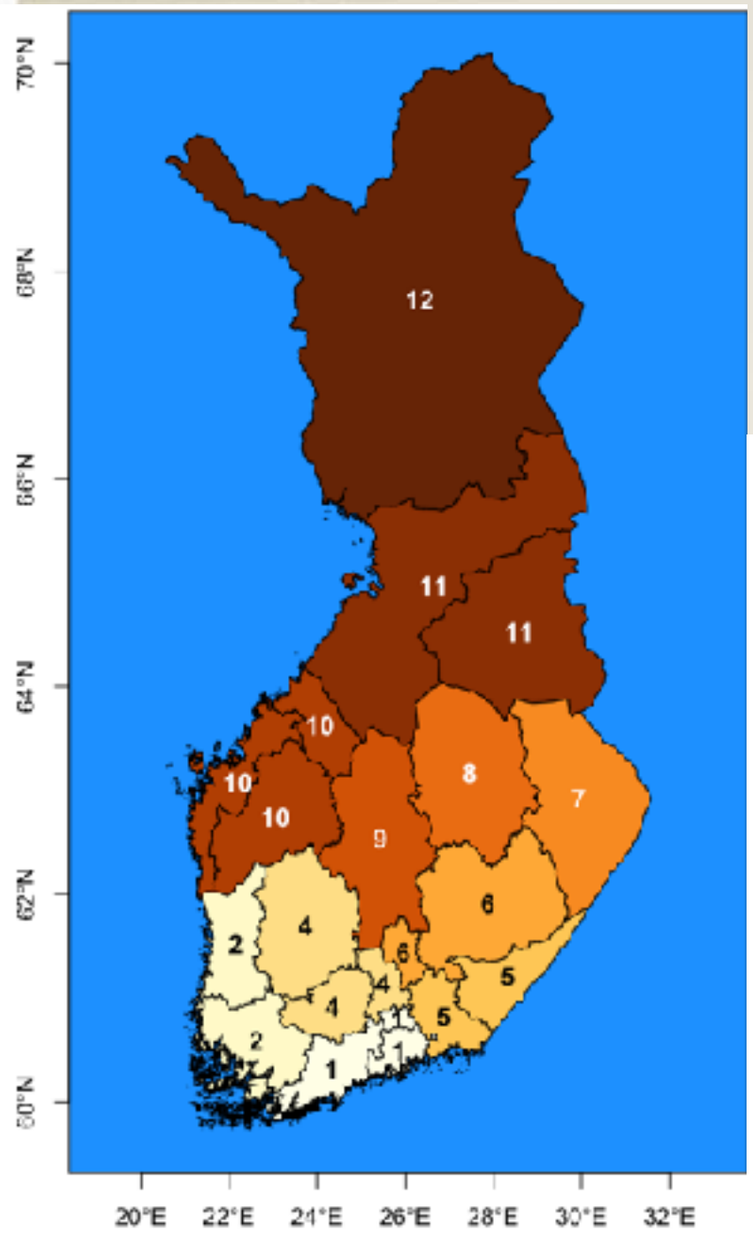
Generations





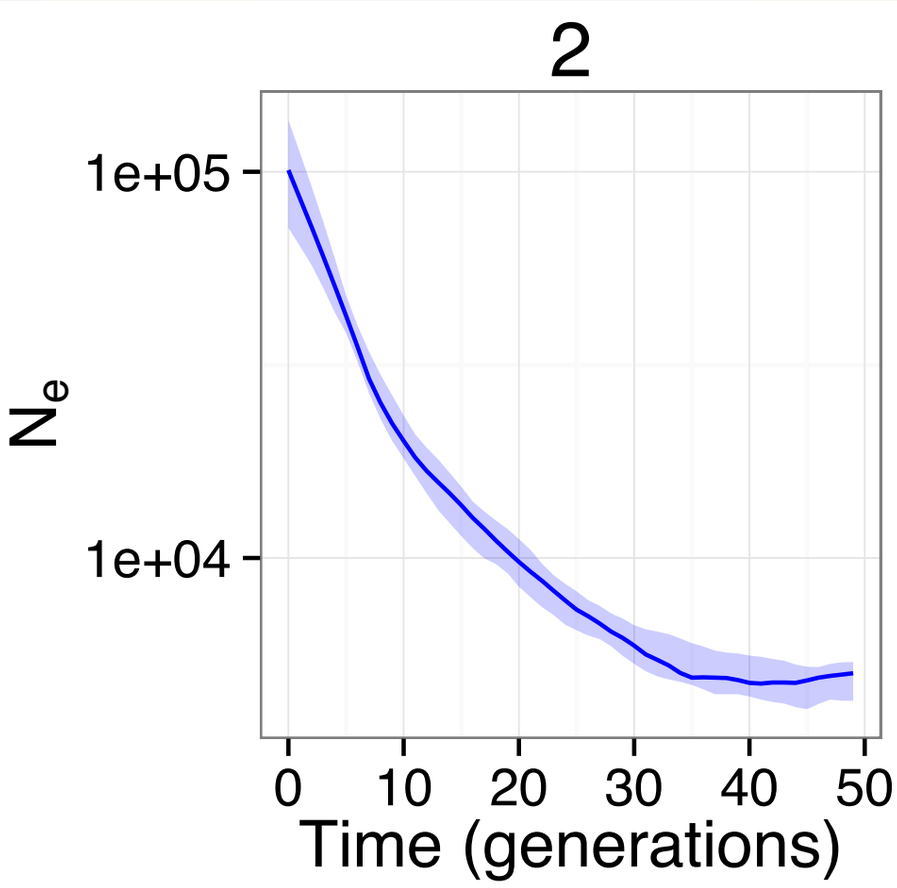


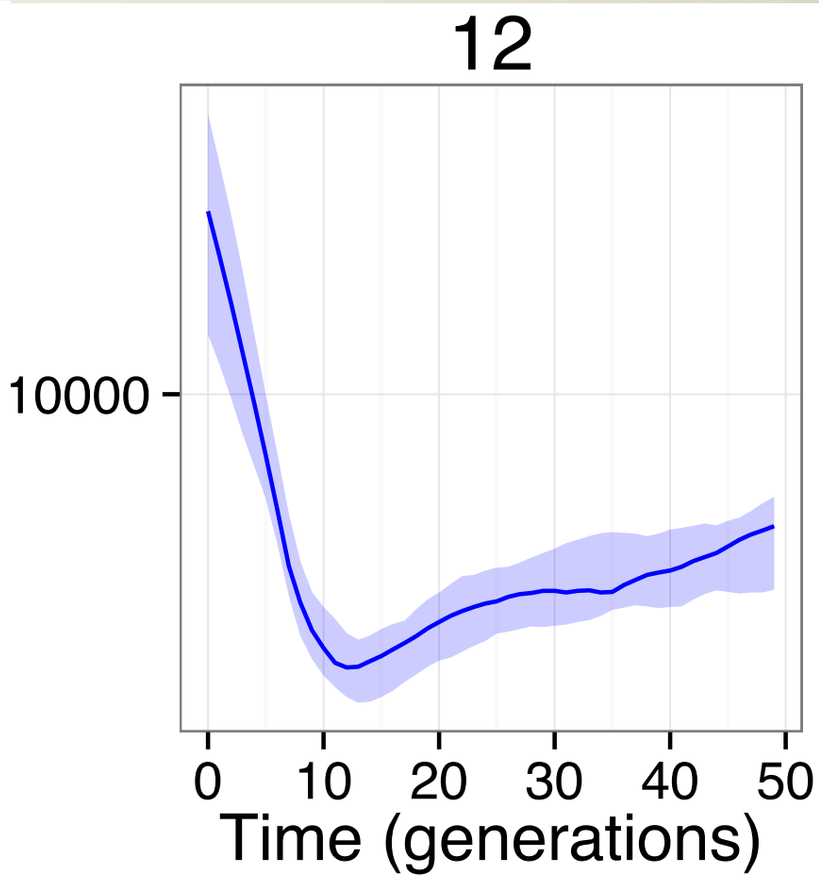
Haplotype sharing provides insights into fine-scale population history



2: Southwest coastal region started growing longer ago

12: Lapland maintained very little growth for extended period





African origins and population structure

What do we know about African population history?

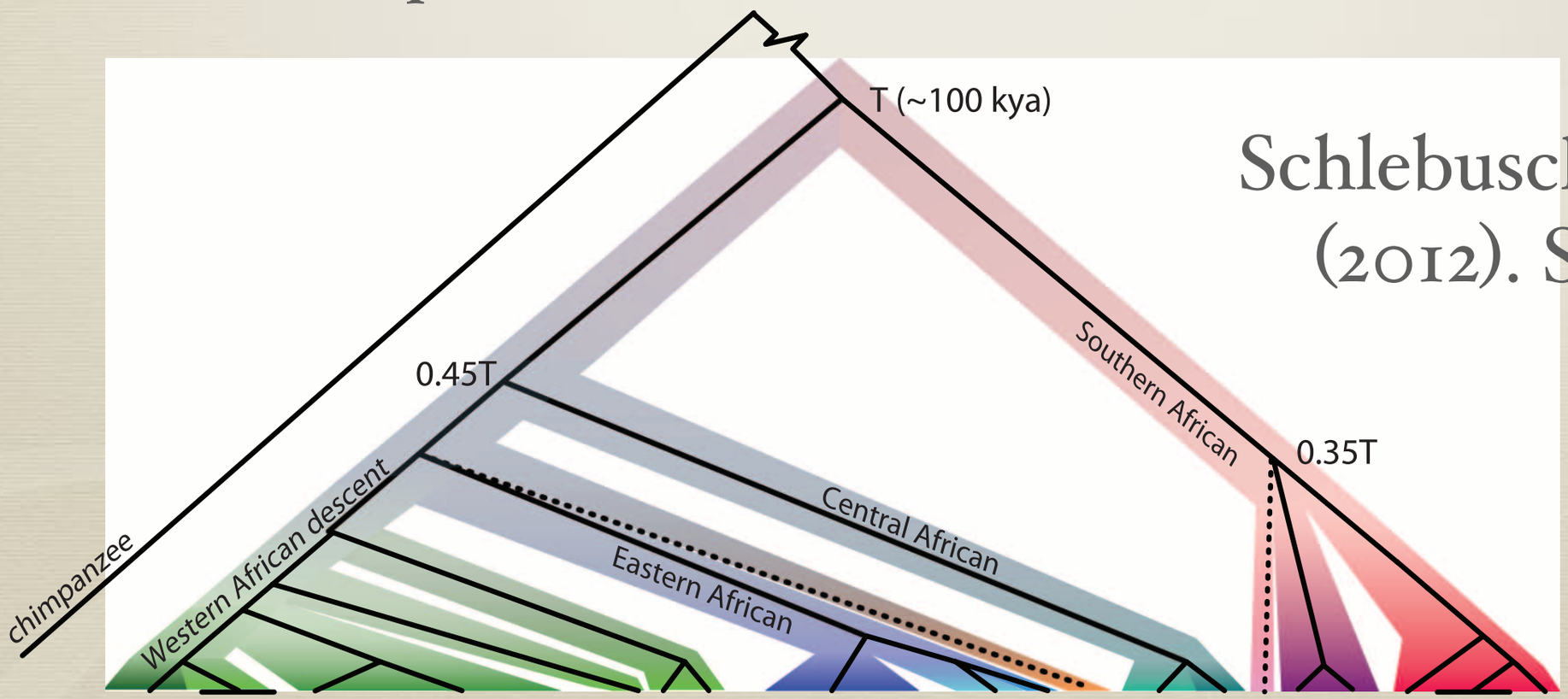
Anatomically modern humans originated in Africa

Klein 1999

White 2003

Timing of population divergence within Africa

- Oldest divergence is between KhoeSan populations and everyone else (120-90 kya)
- Divergence between Central and Eastern Africans: 70-45 kya
- Eurasians split from Eastern African common ancestor



Schlebusch, C.M., et al. (2012). Science 374.

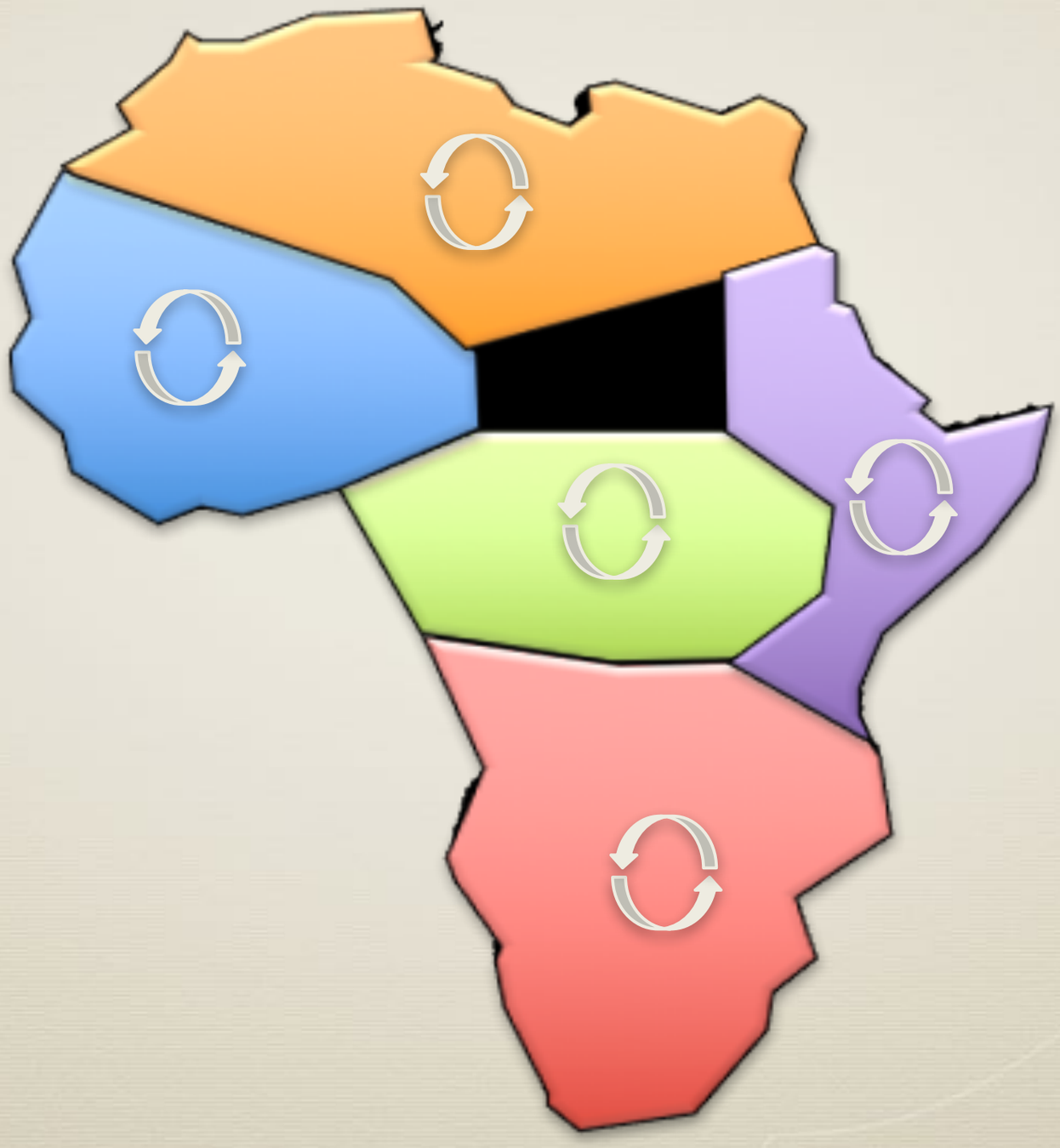
Linguistic structure

* 5 major language families in Africa

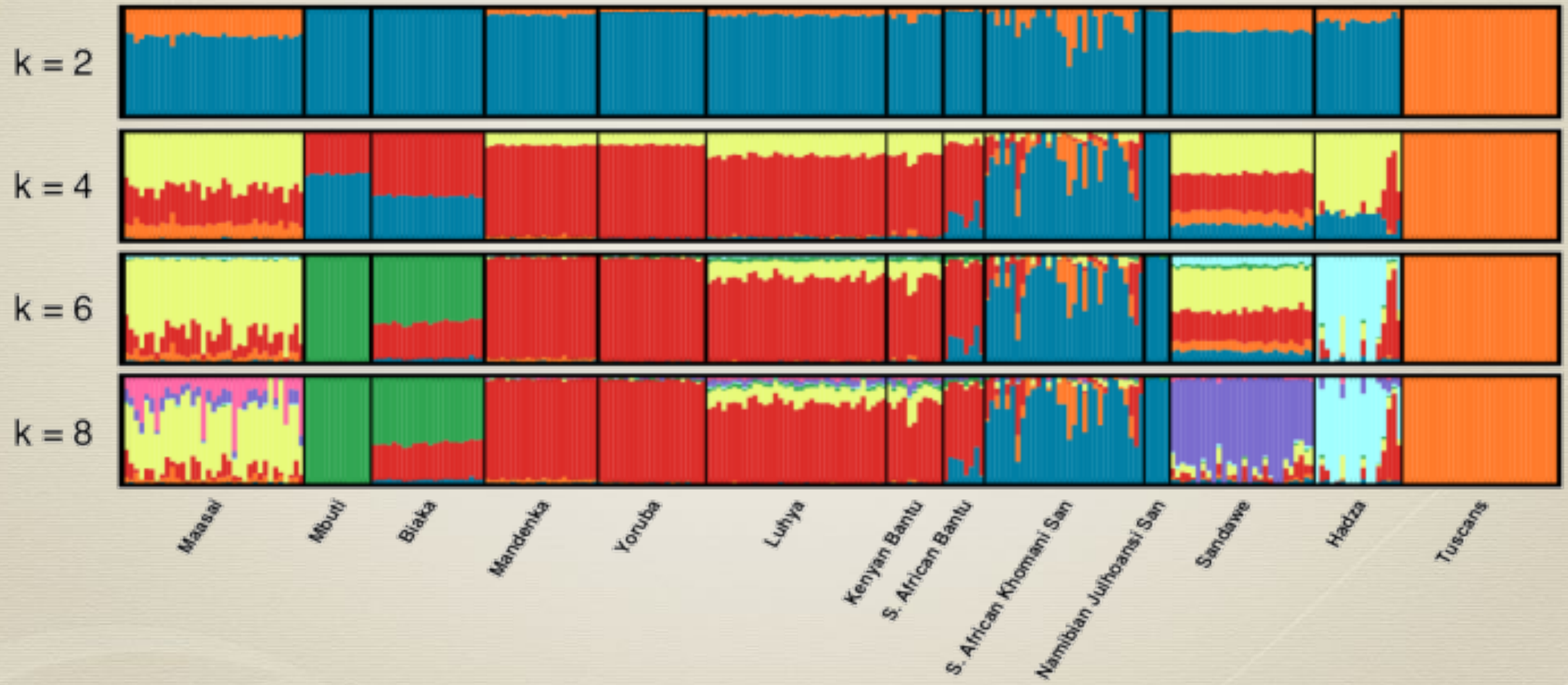
* Expansion of Niger-Congo language 4,000 years ago

* Most isolated and most controversial language family is Khoisan

Population structure



Structure within Africa



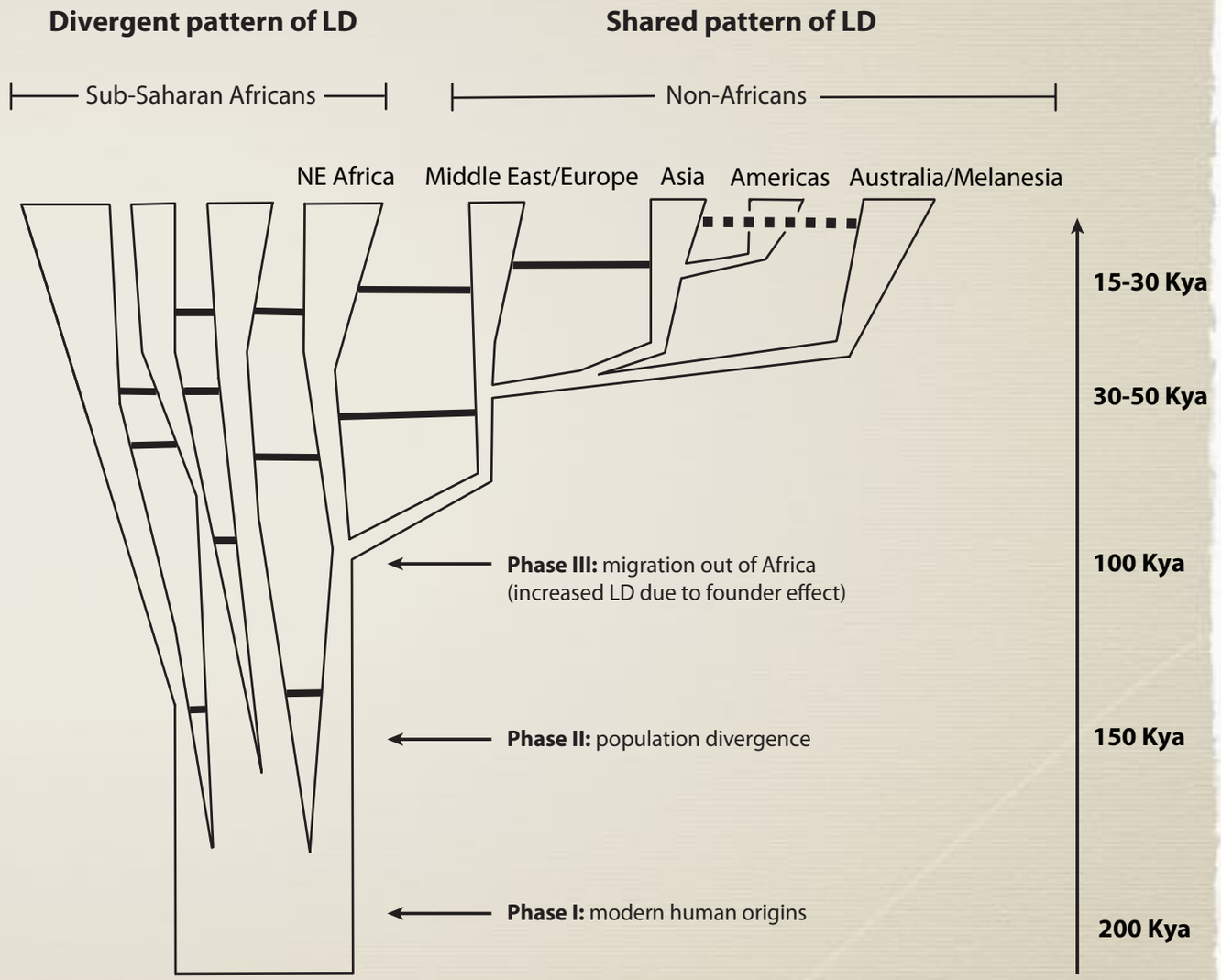
Henn, B.M., et al. (2011). PNAS. 108, 5154–5162.

Summary

* African populations are highly structured (pre-Bantu expansion)

* Time depth of structure is unresolved (~120-40 kya)

* Despite recent gene flow, underlying structure and diversity is detectable

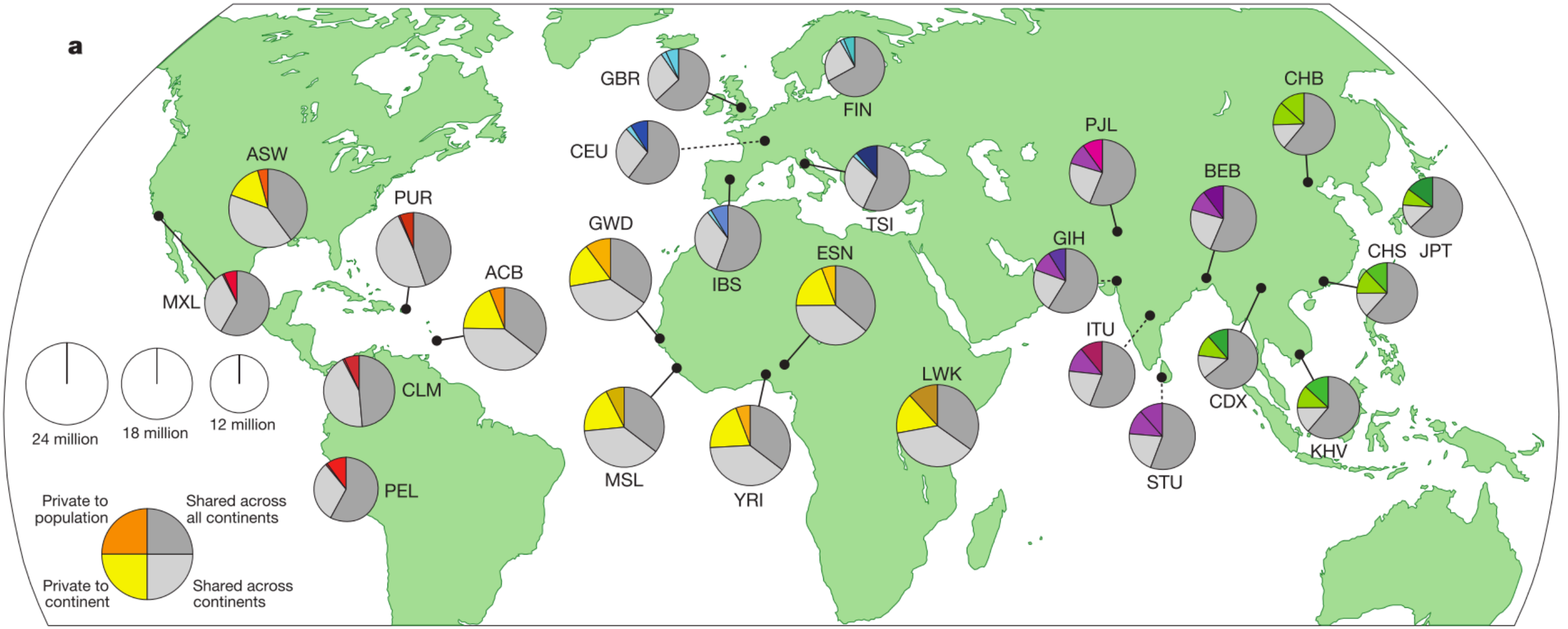


Campbell, M.C., & Tishkoff, S.A. (2008). Annu Rev Genomics Hum Genet 9, 403–433.

Transferability of Euro-centric genetic studies to diverse populations

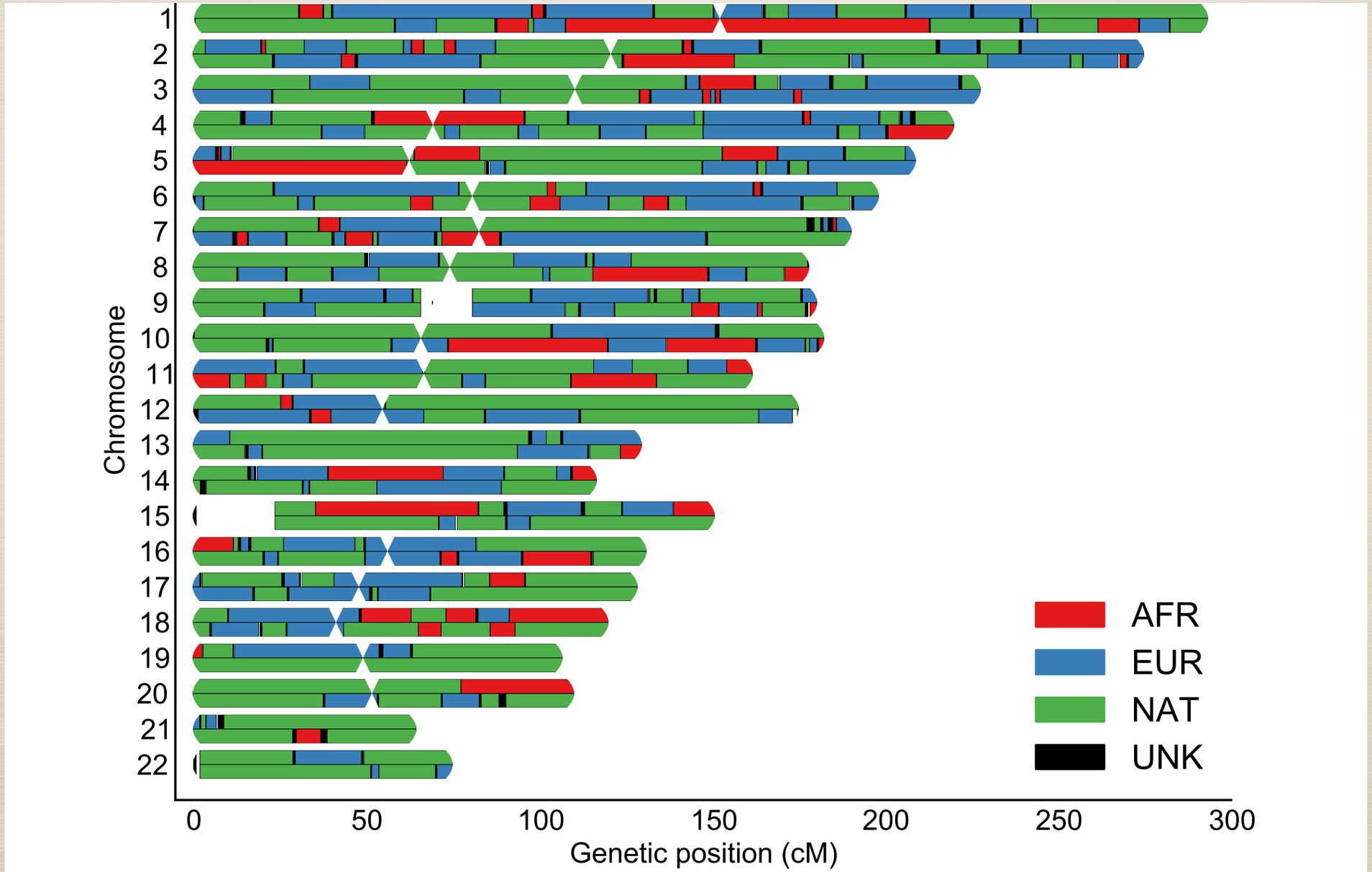
Martin, A.R., et al. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. AJHG. 100, 635–649.

1000 Genomes Project



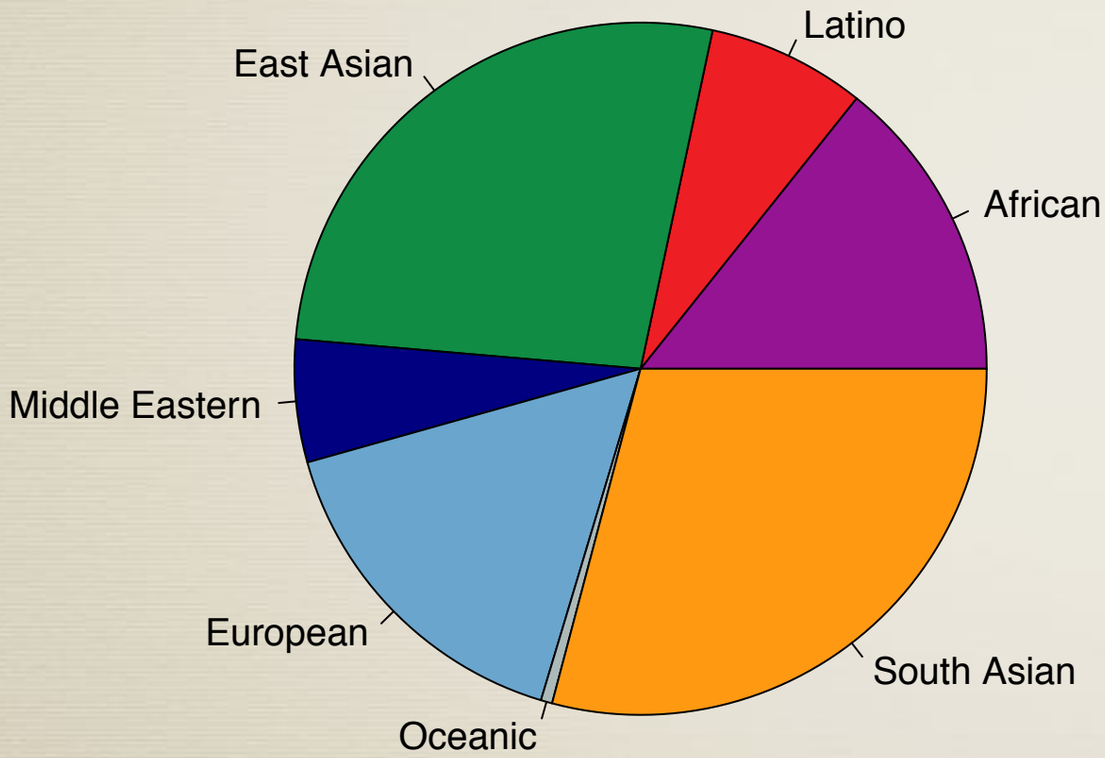
The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature 526, 68–74.

Local ancestry inference in recently admixed genomes

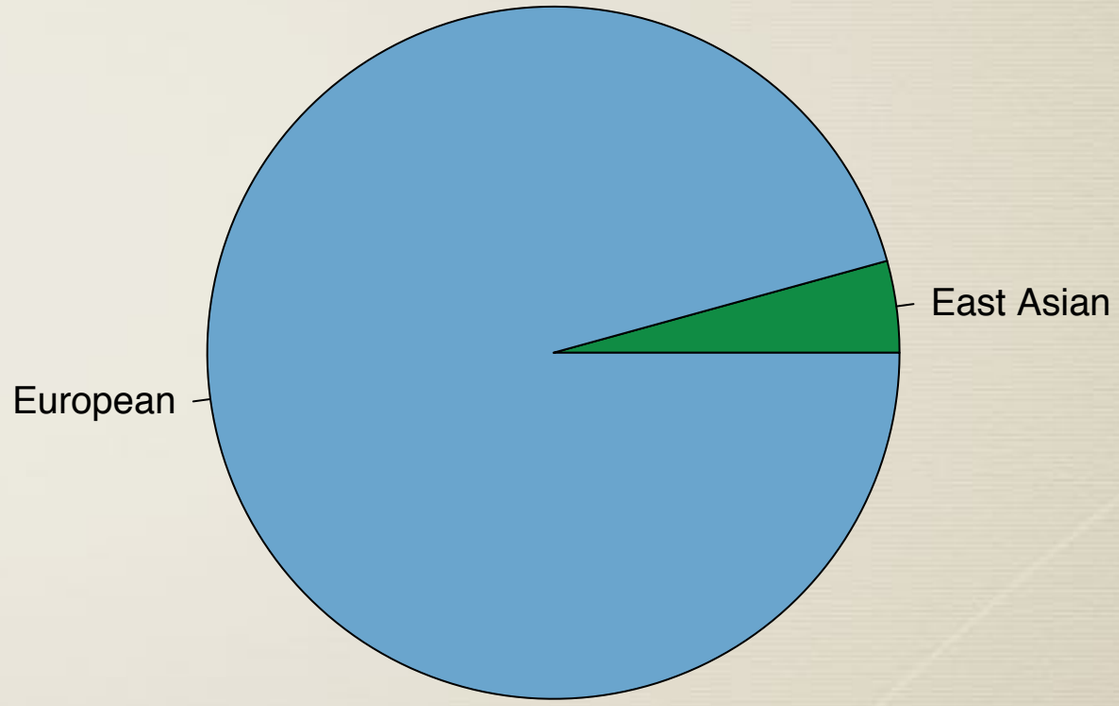


Biased genetic discoveries

Global population

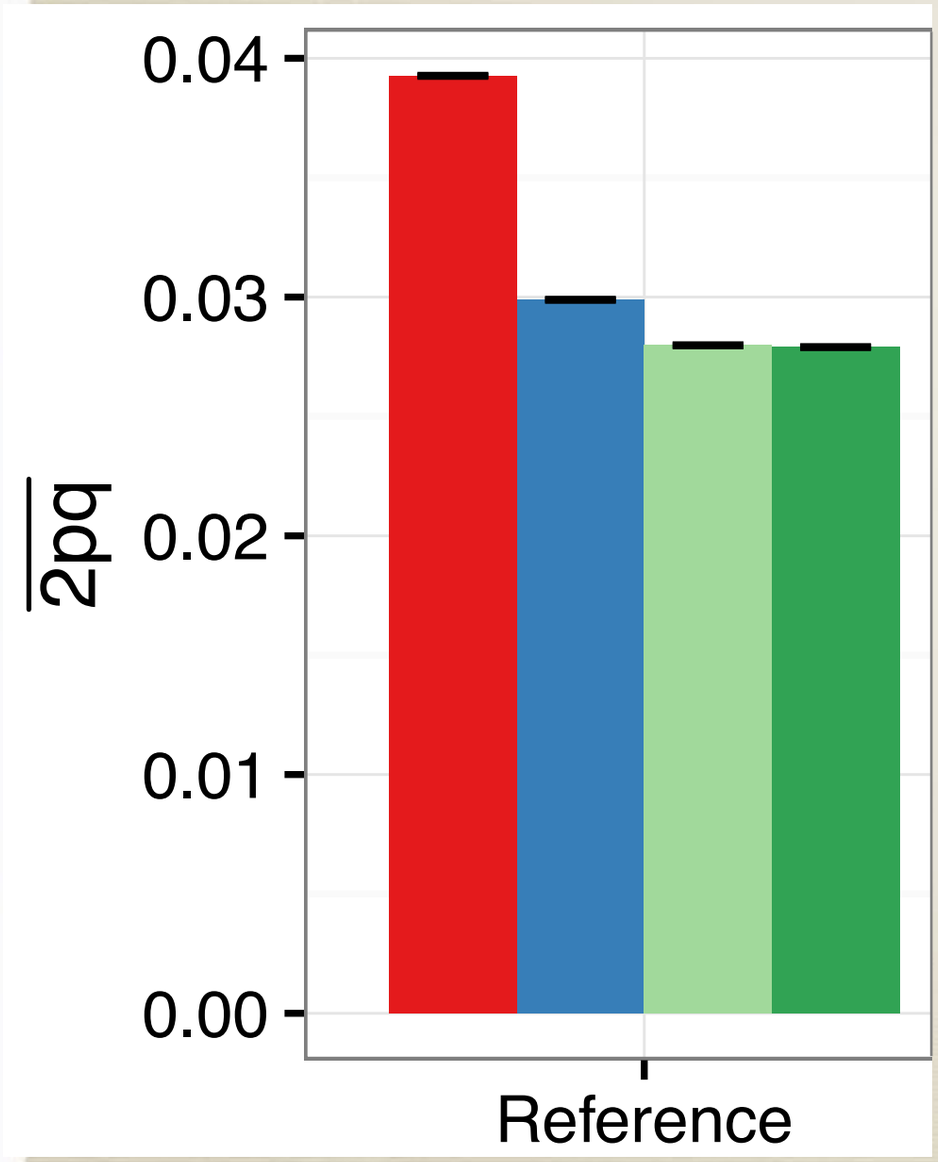


Psychiatric Genetics Consortium GWAS



Biased genetic discoveries

Whole genome



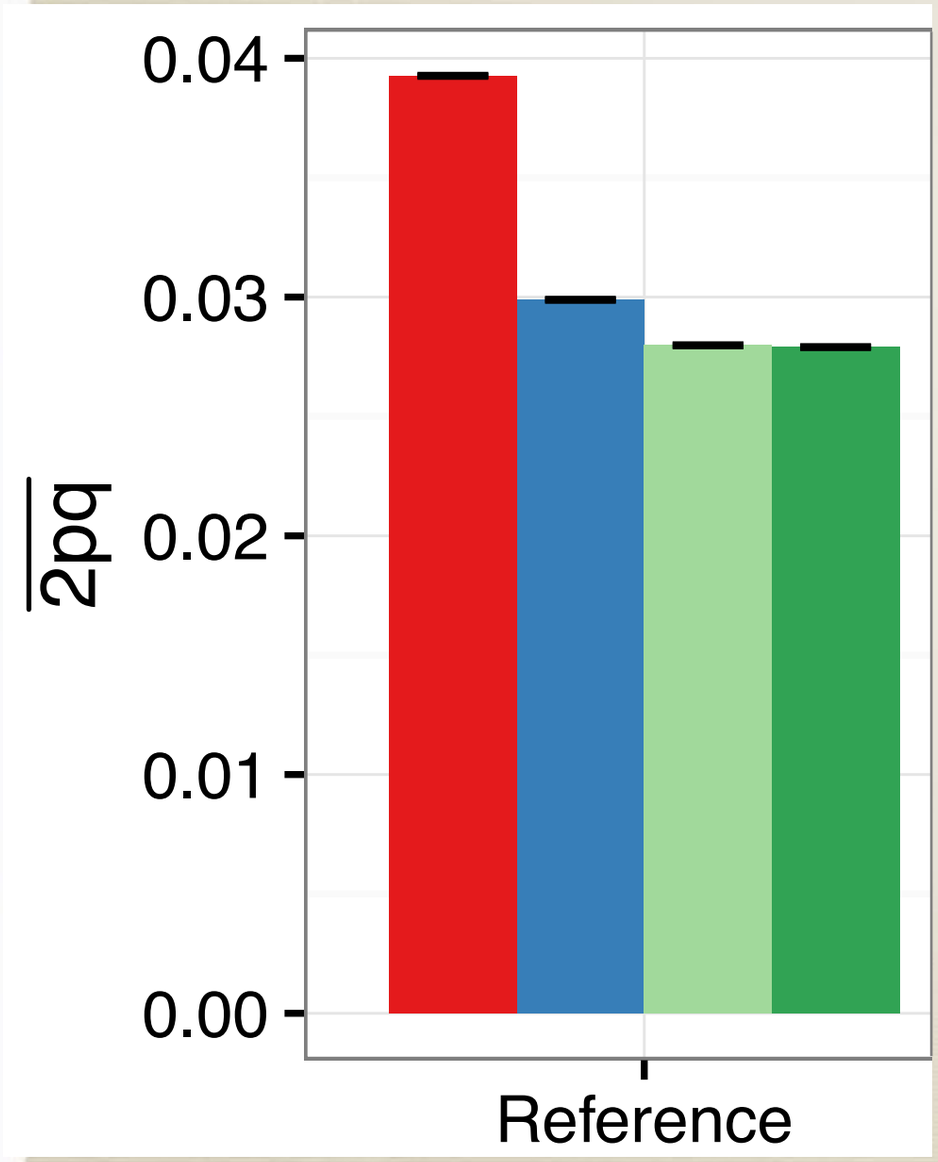
African

European

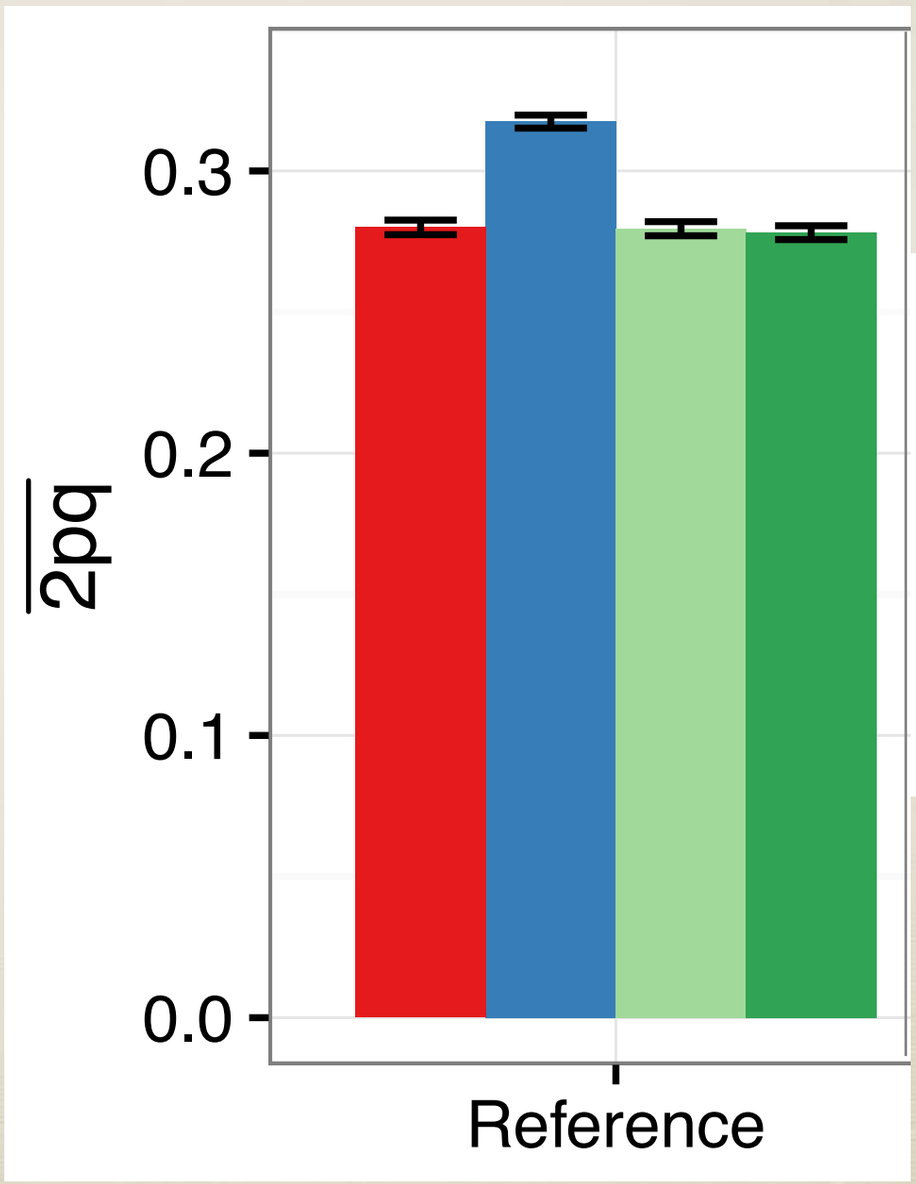
East Asian

Biased genetic discoveries

Whole genome



GWAS catalog



African

European

East Asian

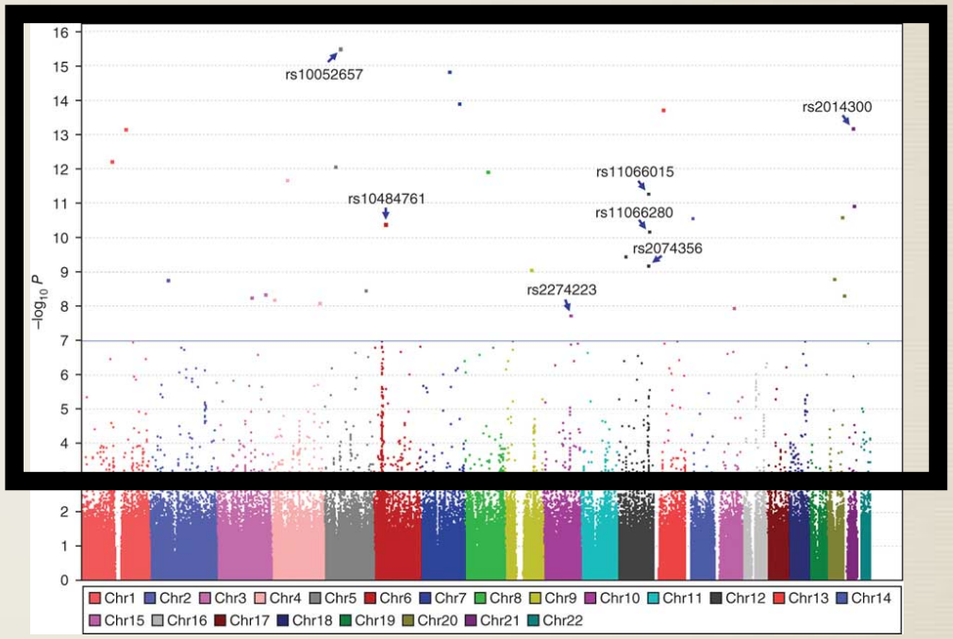
How do biased genetic studies impact the transferability of GWAS findings?

?

Computing polygenic risk scores from summary statistics

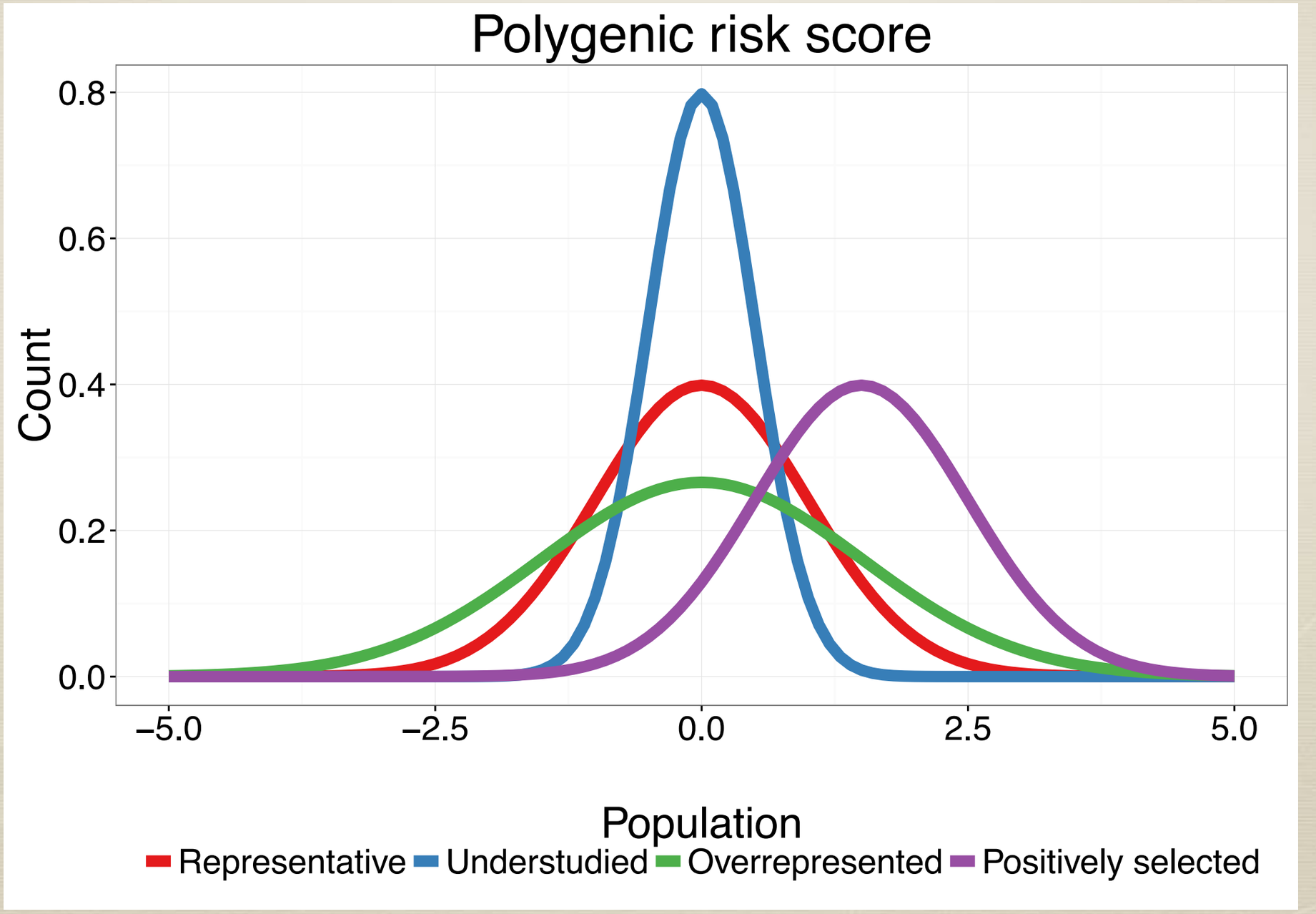
- LD clumping
- P-value thresholds

Σ

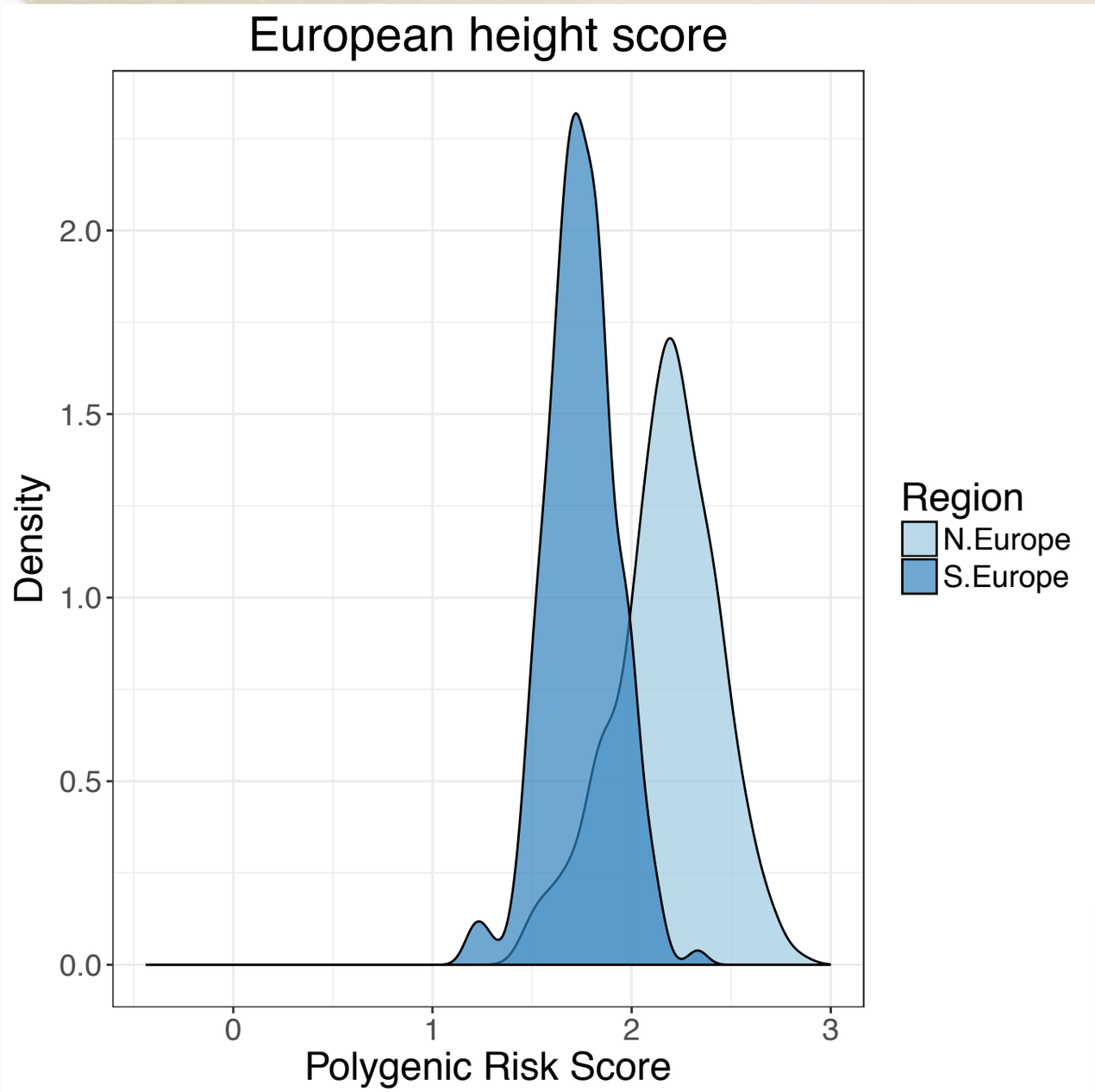


$$X = \sum_{i=1}^{m} g_i \beta_i$$

Interpreting polygenic risk scores

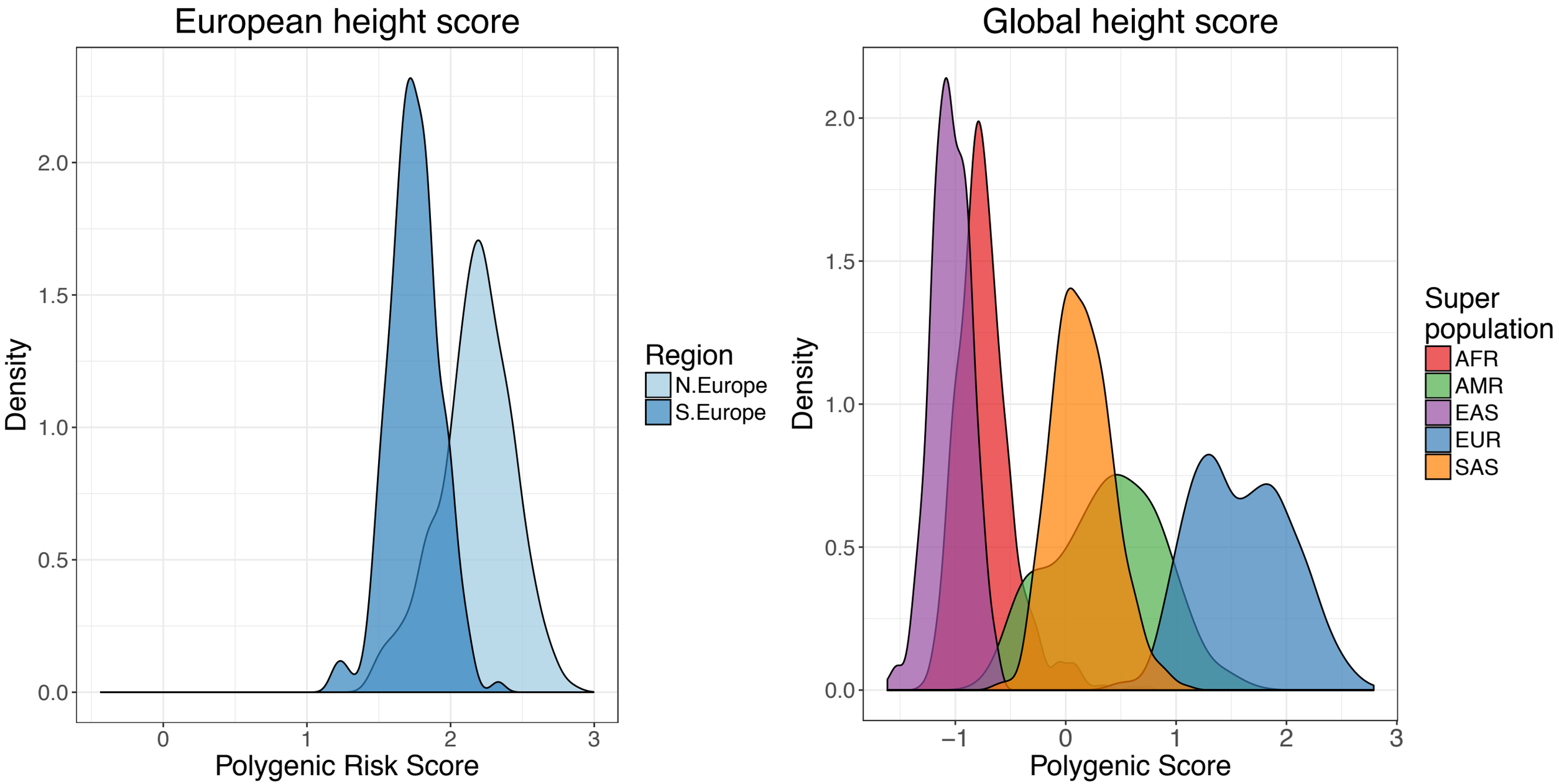


Polygenic height score appears to reflect adaptive event in Europeans



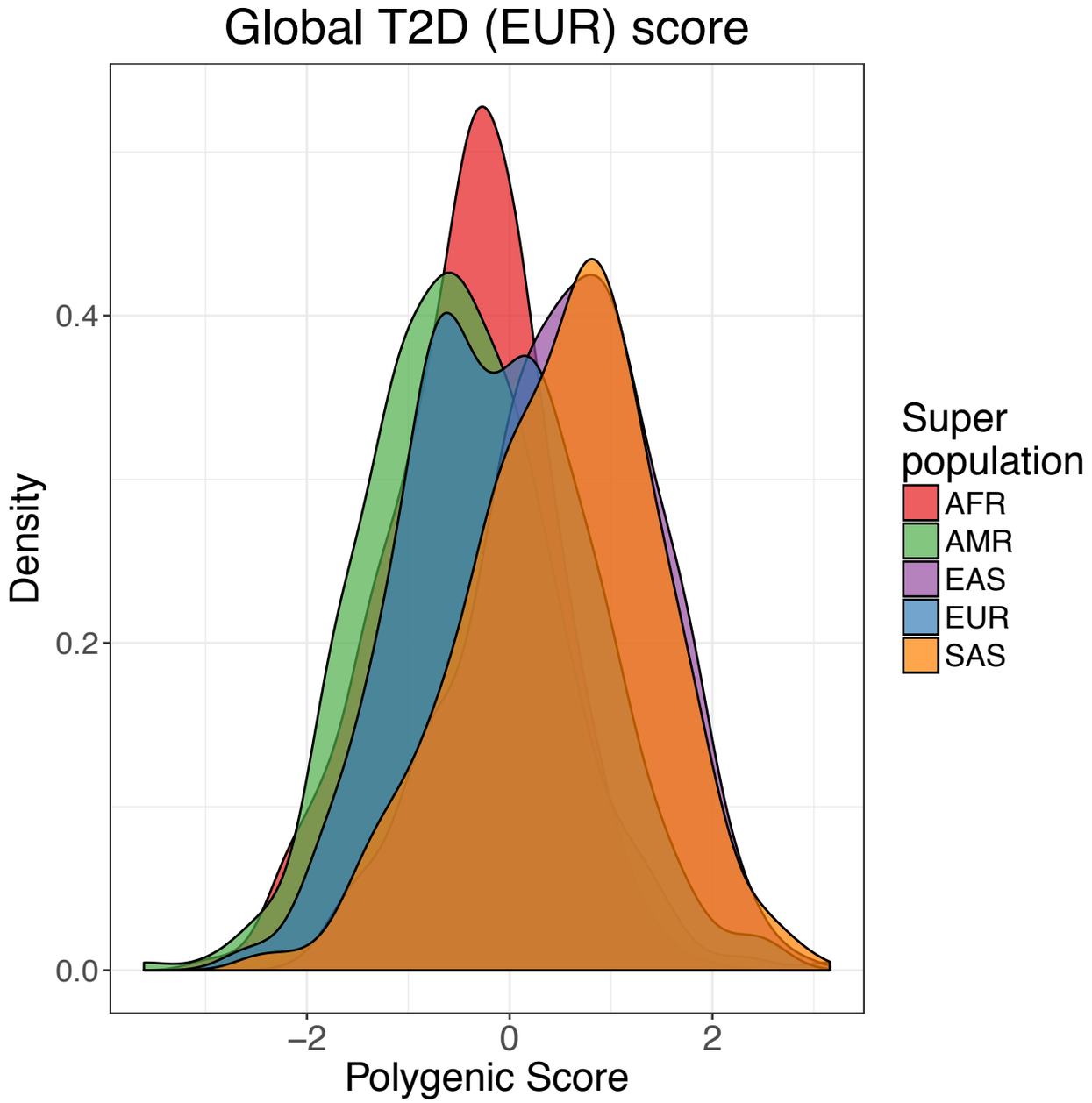
Wood, A.R., et al. (2014). Nature Genetics 46, 1173–1186.

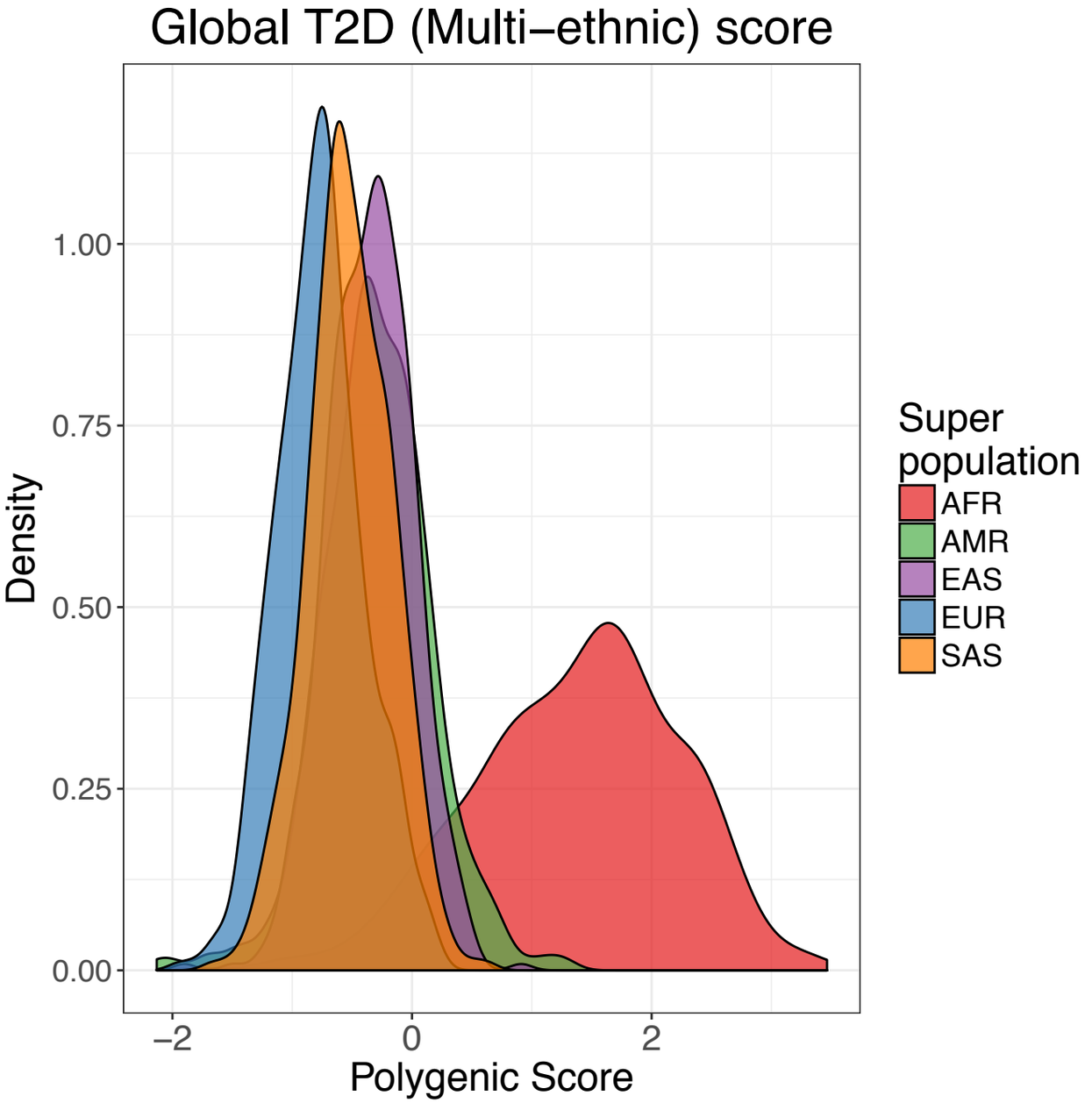
Polygenic height score appears to reflect adaptive event in Europeans... and bias



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

Polygenic risk of Type II diabetes highlights role of demography

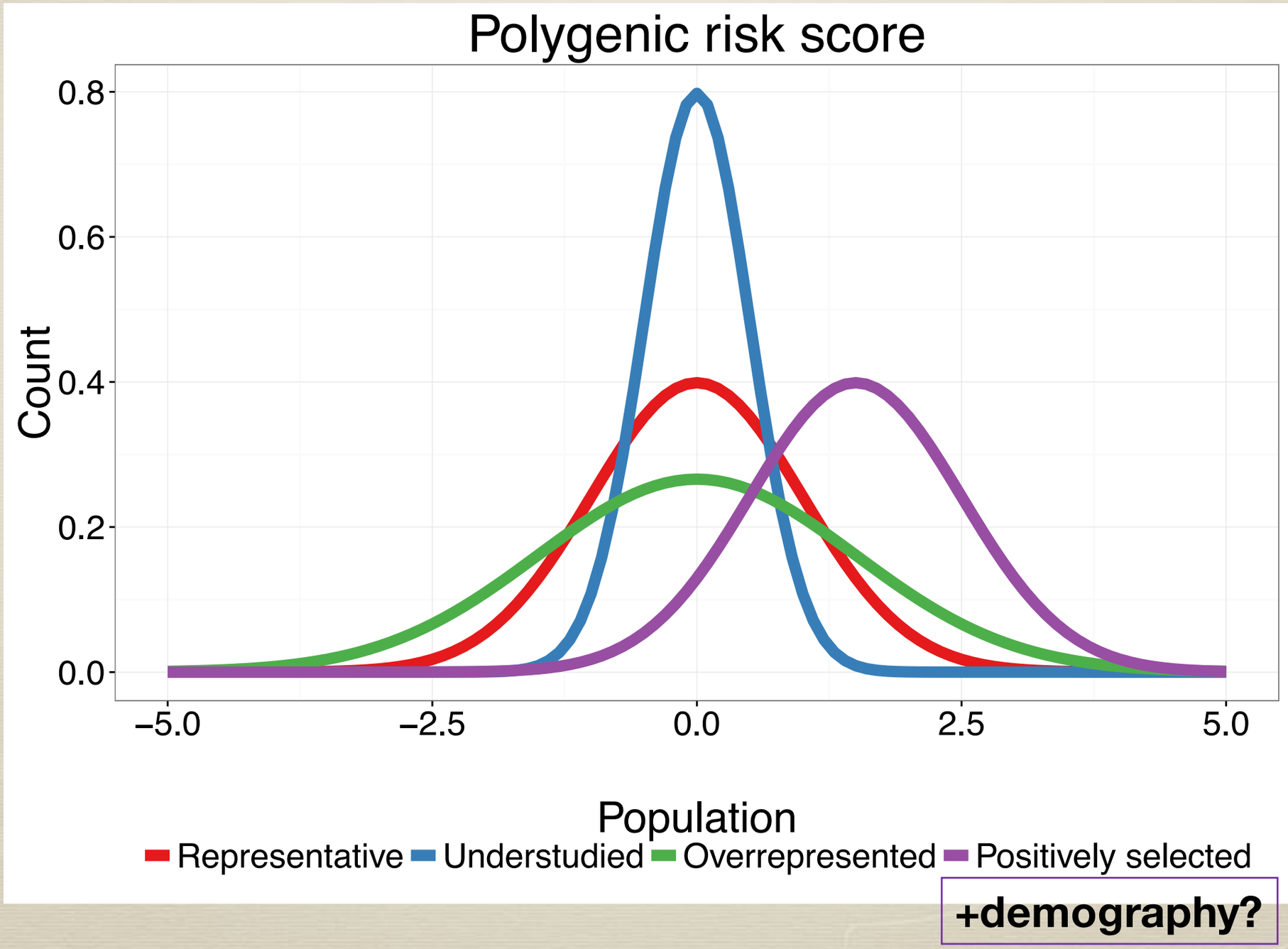




European: Gaulton, K.J., et al. (2015). Nat. Genet. *47*, 1415–1425.

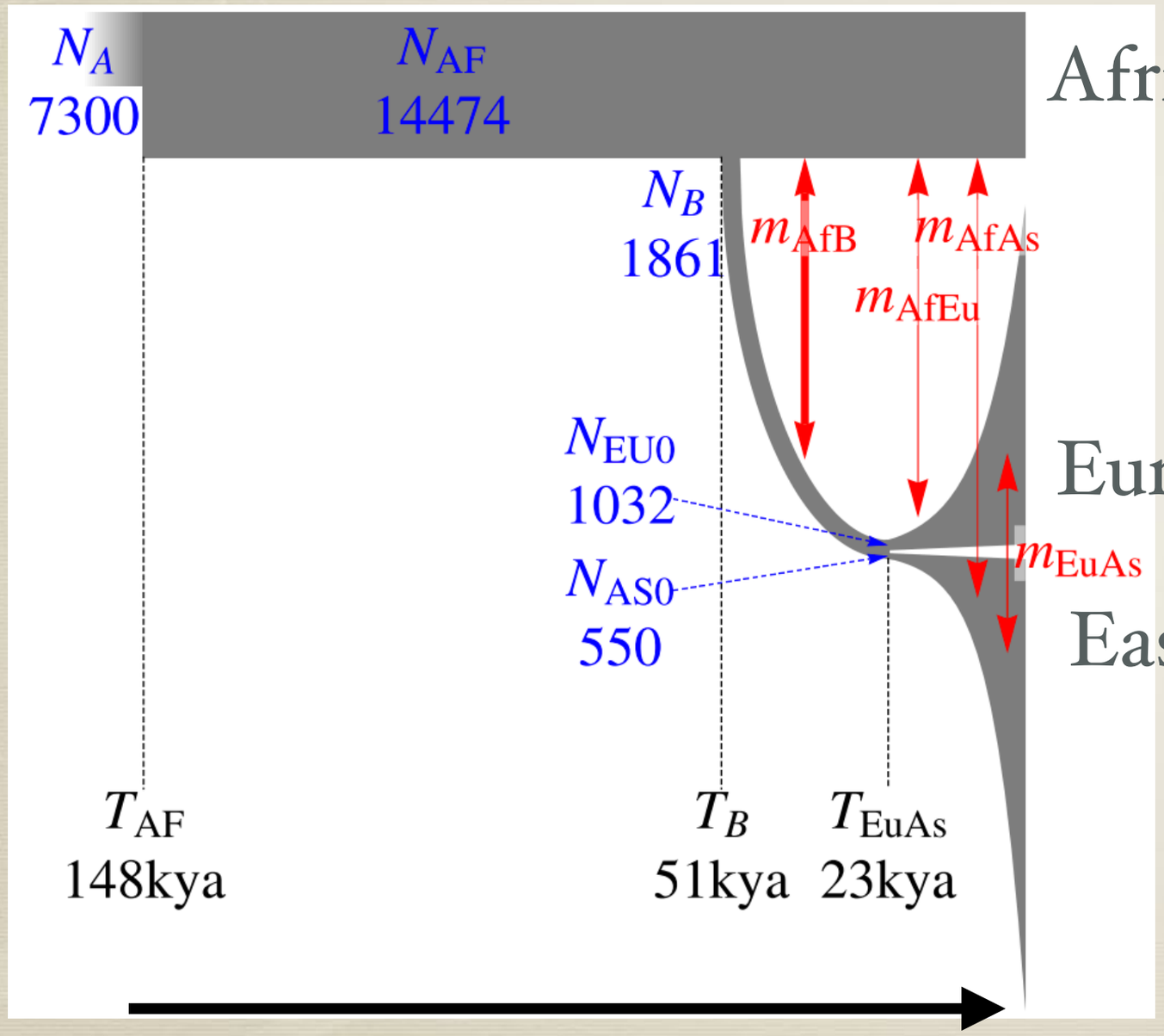
Multi-ethnic: Mahajan, A., et al. (2014). Nat. Genet. *46*, 234–244.

Interpreting polygenic risk scores



+demography?

Coalescent model for simulation framework



Model <u>parameters</u>

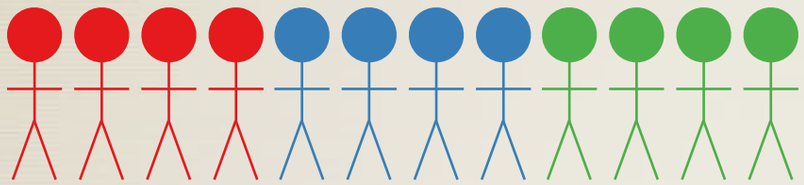
- N_e: population size
- m: migration rates
- T: time
- r: growth

Demographic model: Gravel, S., et al. (2011). Proc. Natl. Acad. Sci. U. S. A. *108*, 11983–11988.

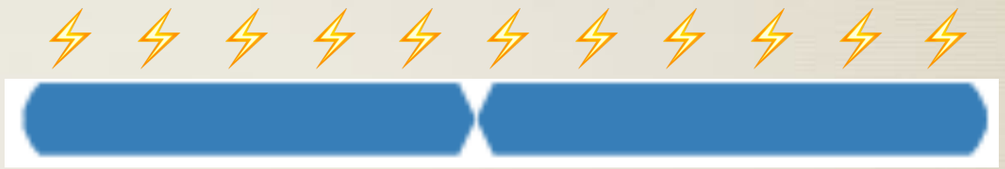
msprime: Kelleher, J., Etheridge, A.M., and Mcvean, G. (2016). PLoS Comput Biol 1–22.

Simulation overview

1. Simulate genotypes (AFR, EUR, EAS)



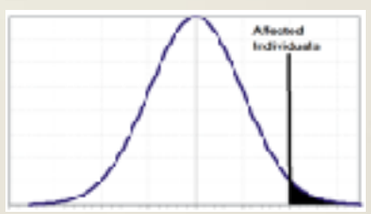
2. Assign evenly spaced causal variants



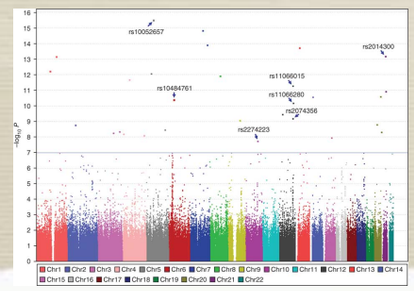
3. Compute PRS_{TRUE}

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

4. Define EUR cases, controls (10k each)



5. Run a EUR GWAS

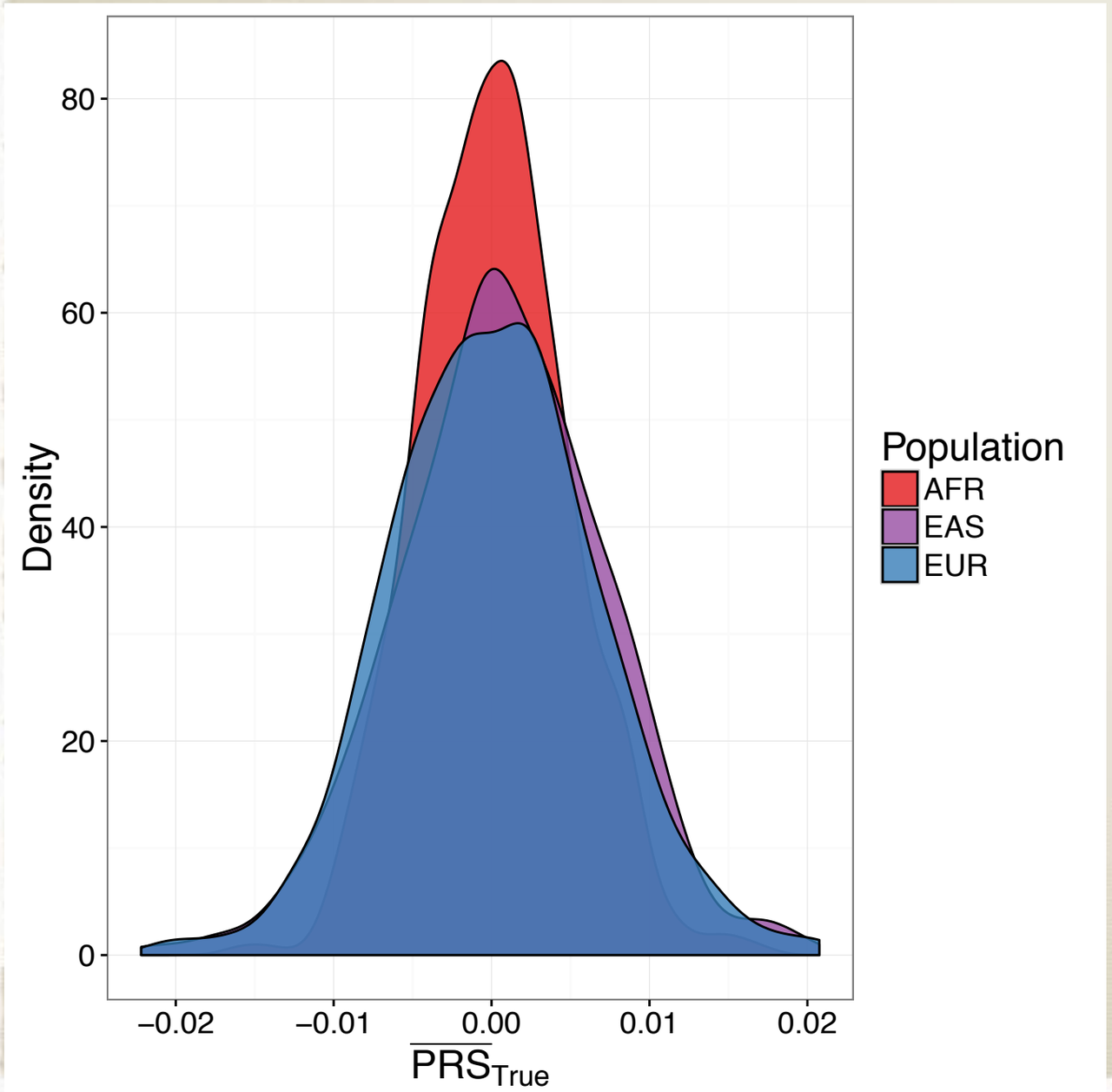


6. Compute PRS_{INFER} across populations

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

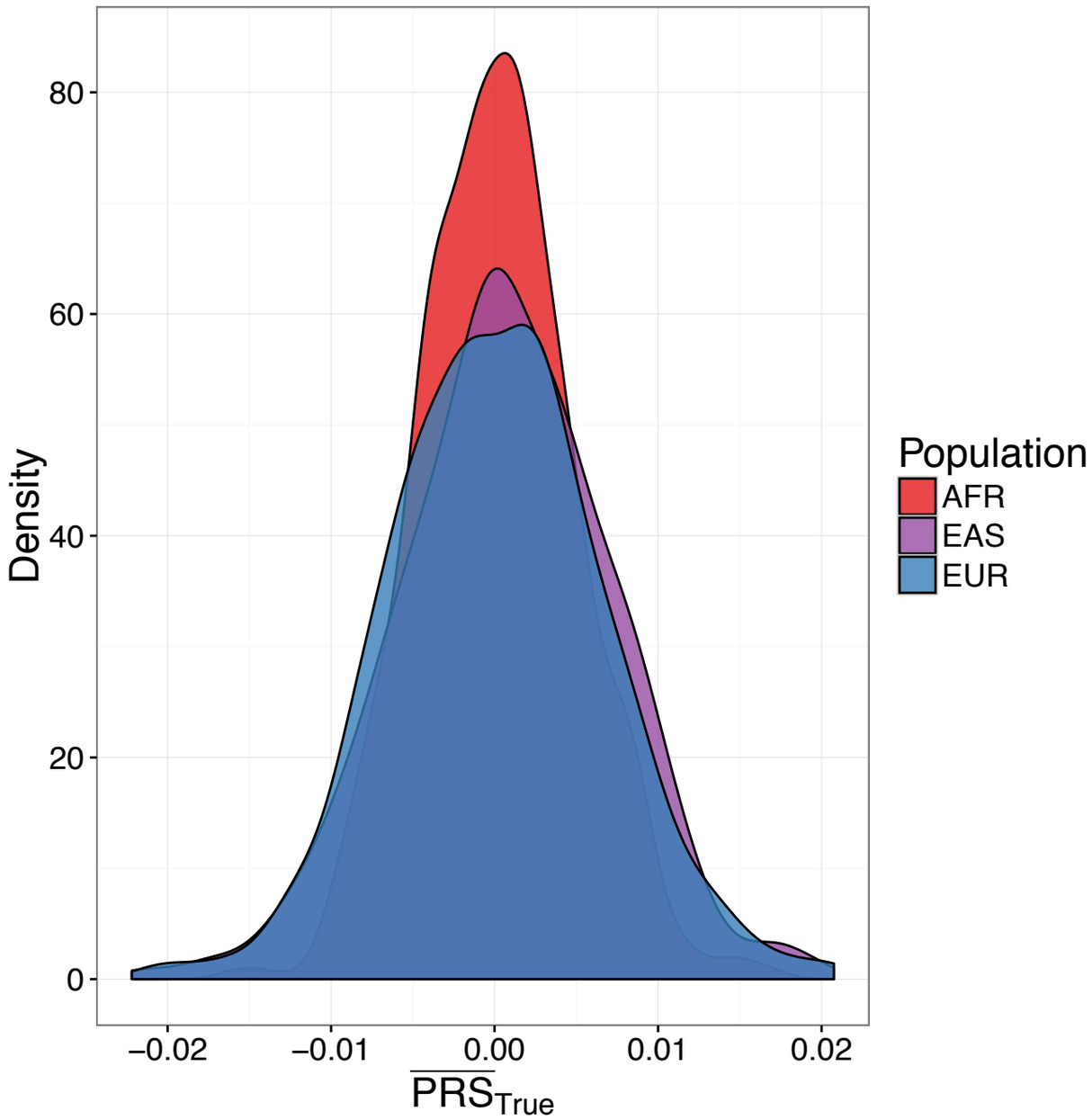
PRS_{TRUE} is not significantly different across populations

True causal variants

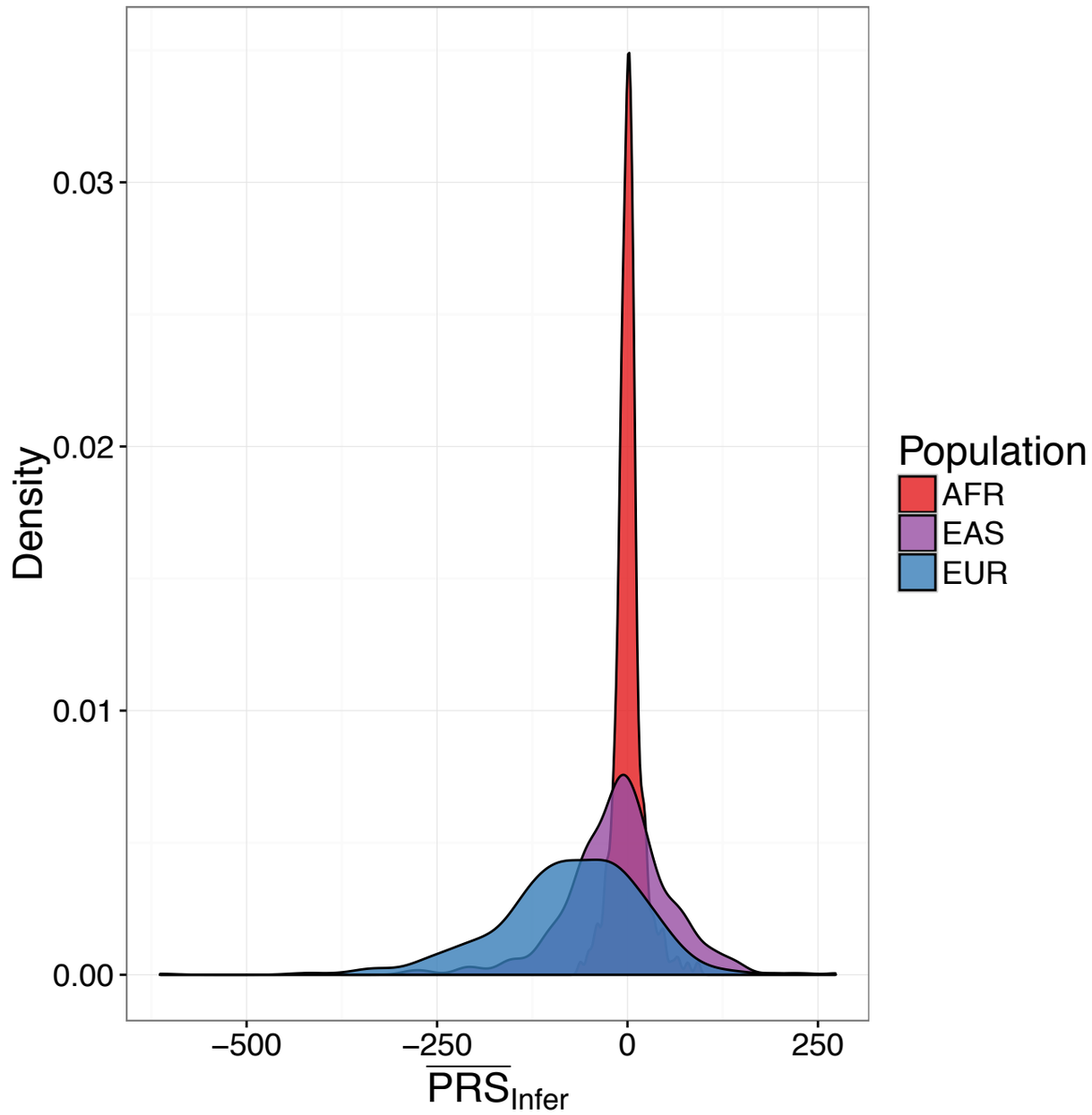


PRS_{INFER} is highly stratified across populations

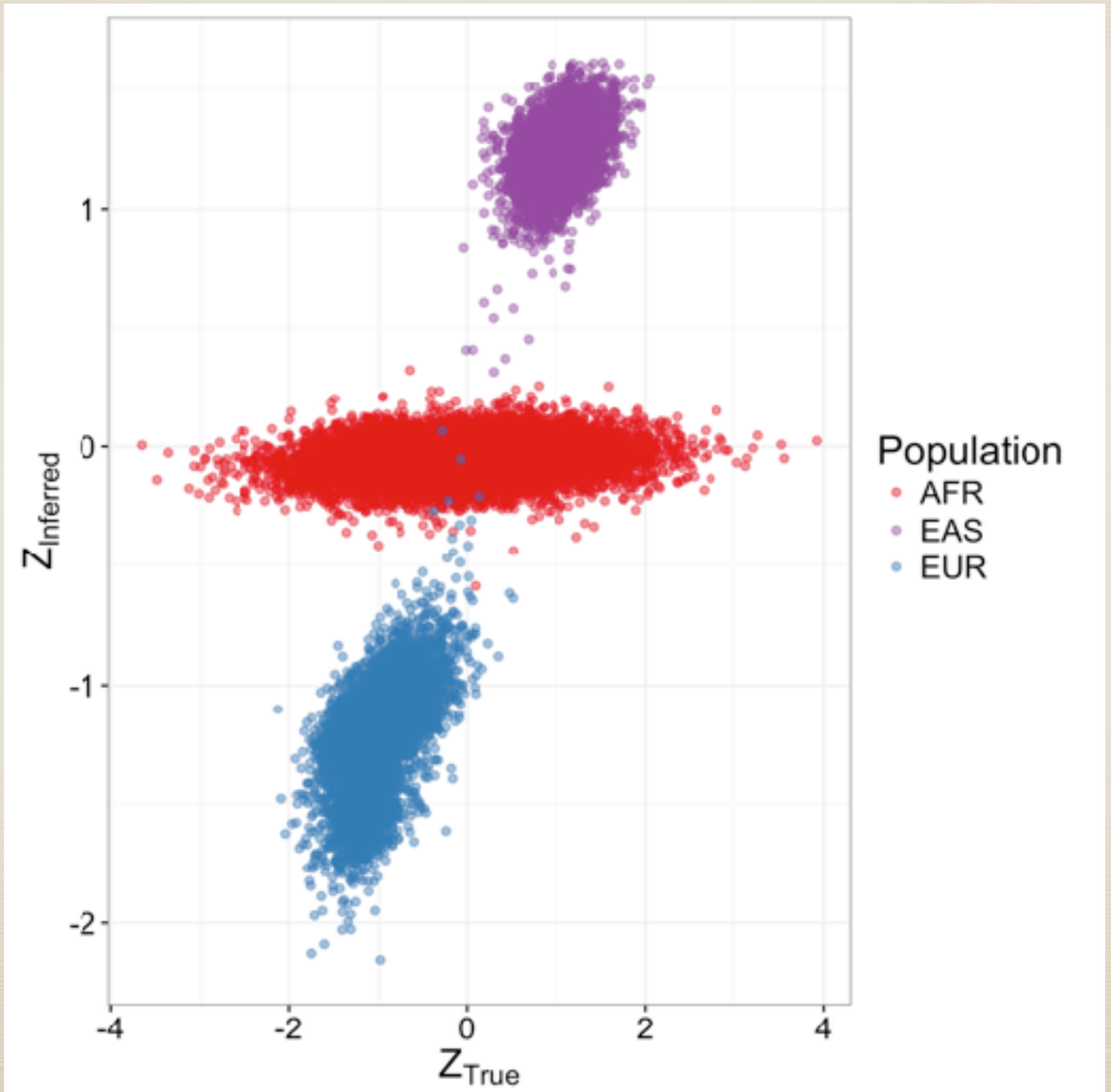
True causal variants



GWAS inferred variants



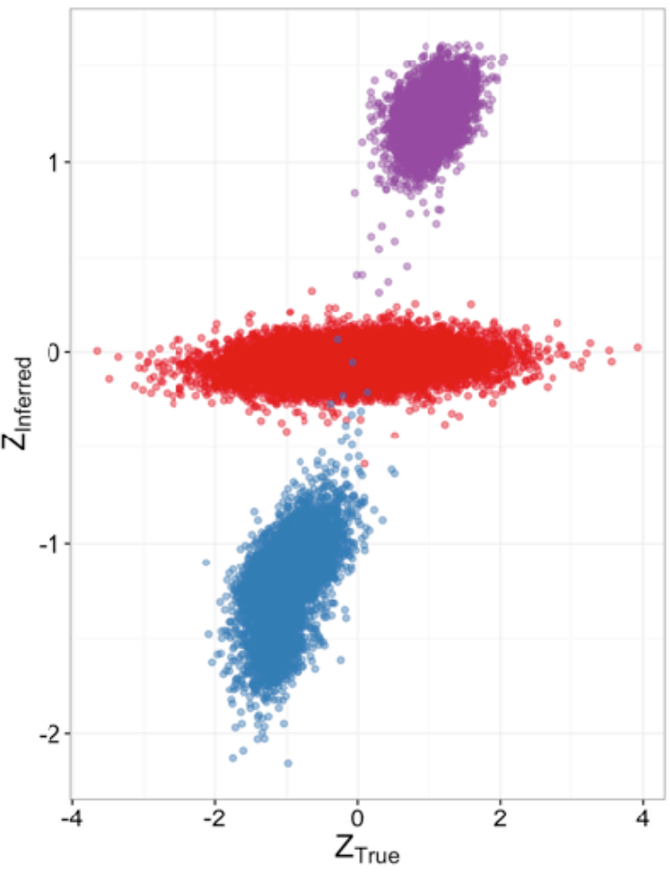
Simulations demonstrate inconsistent, unpredictable biases across populations



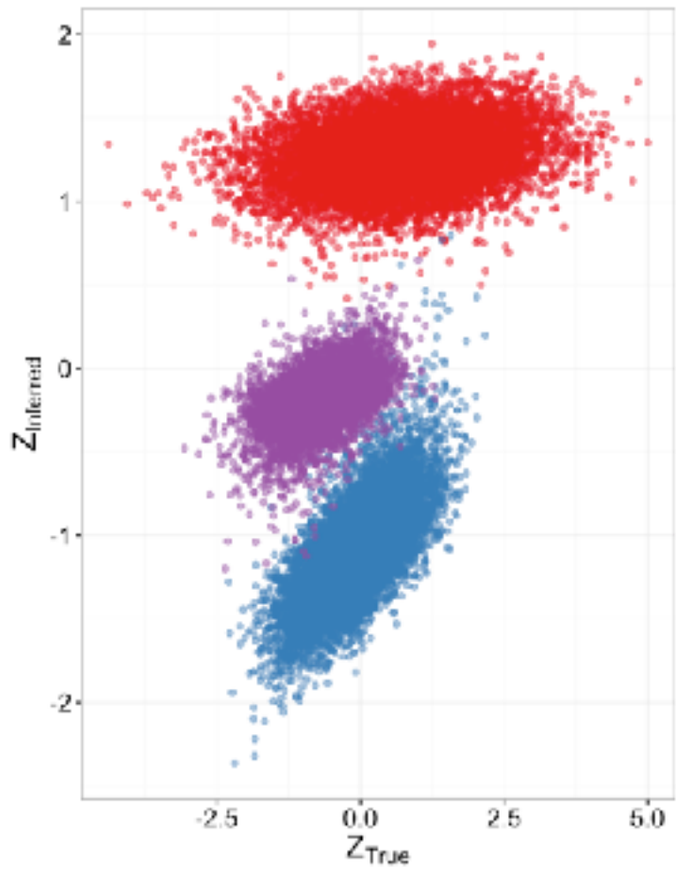
Simulations demonstrate inconsistent, unpredictable biases across populations

Analogous to different traits:

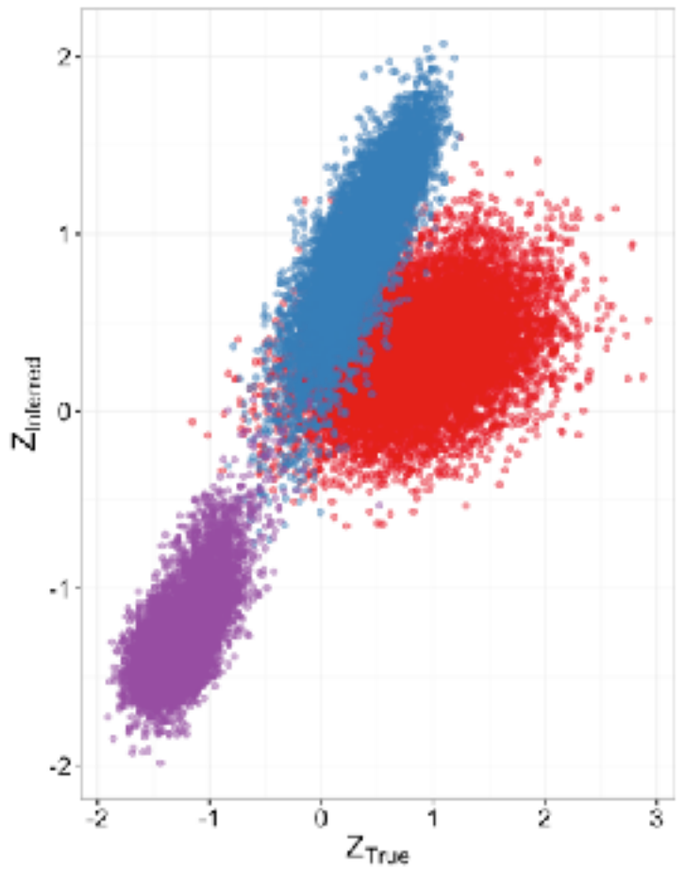
Height



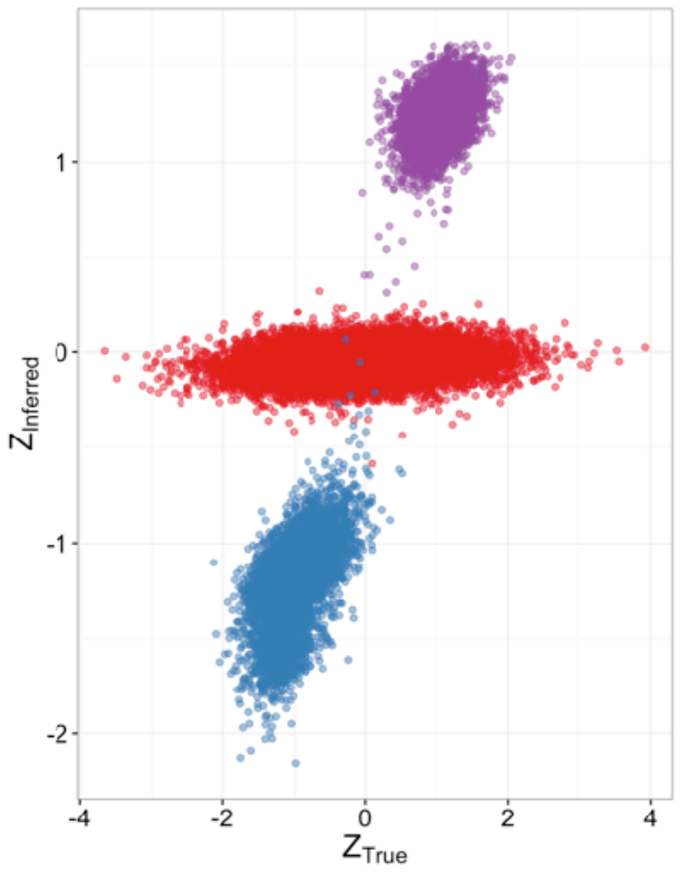
Schizophrenia

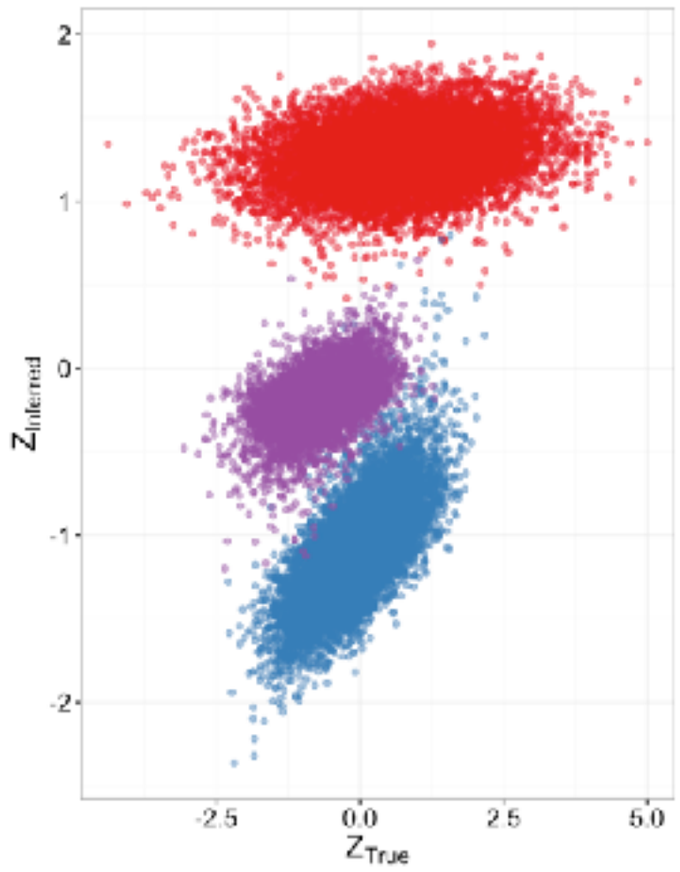


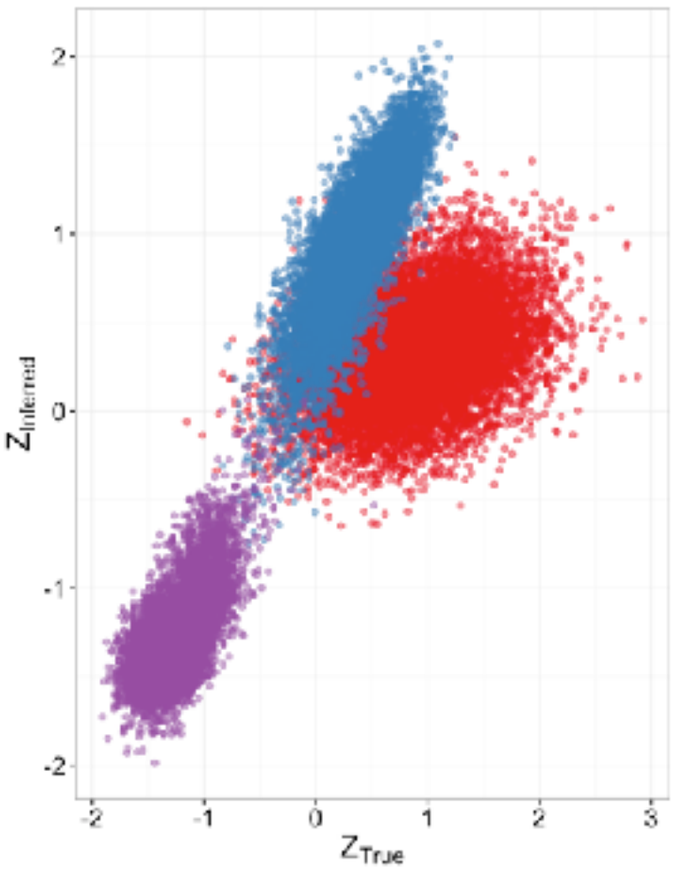
T2D



Simulations demonstrate inconsistent, unpredictable biases across populations

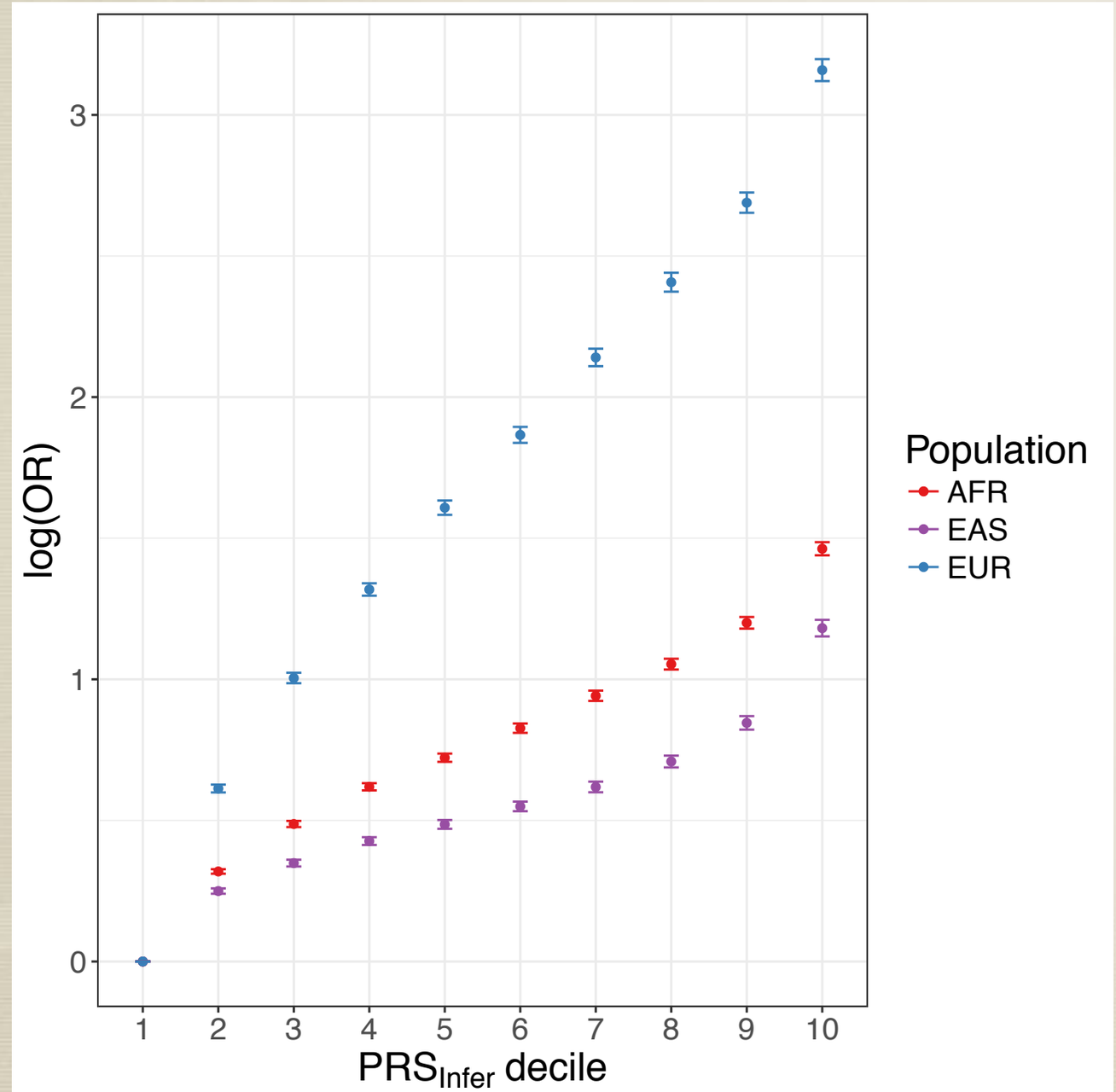






For a given trait, impossible to predict *a priori* which population will have highest inferred risk!

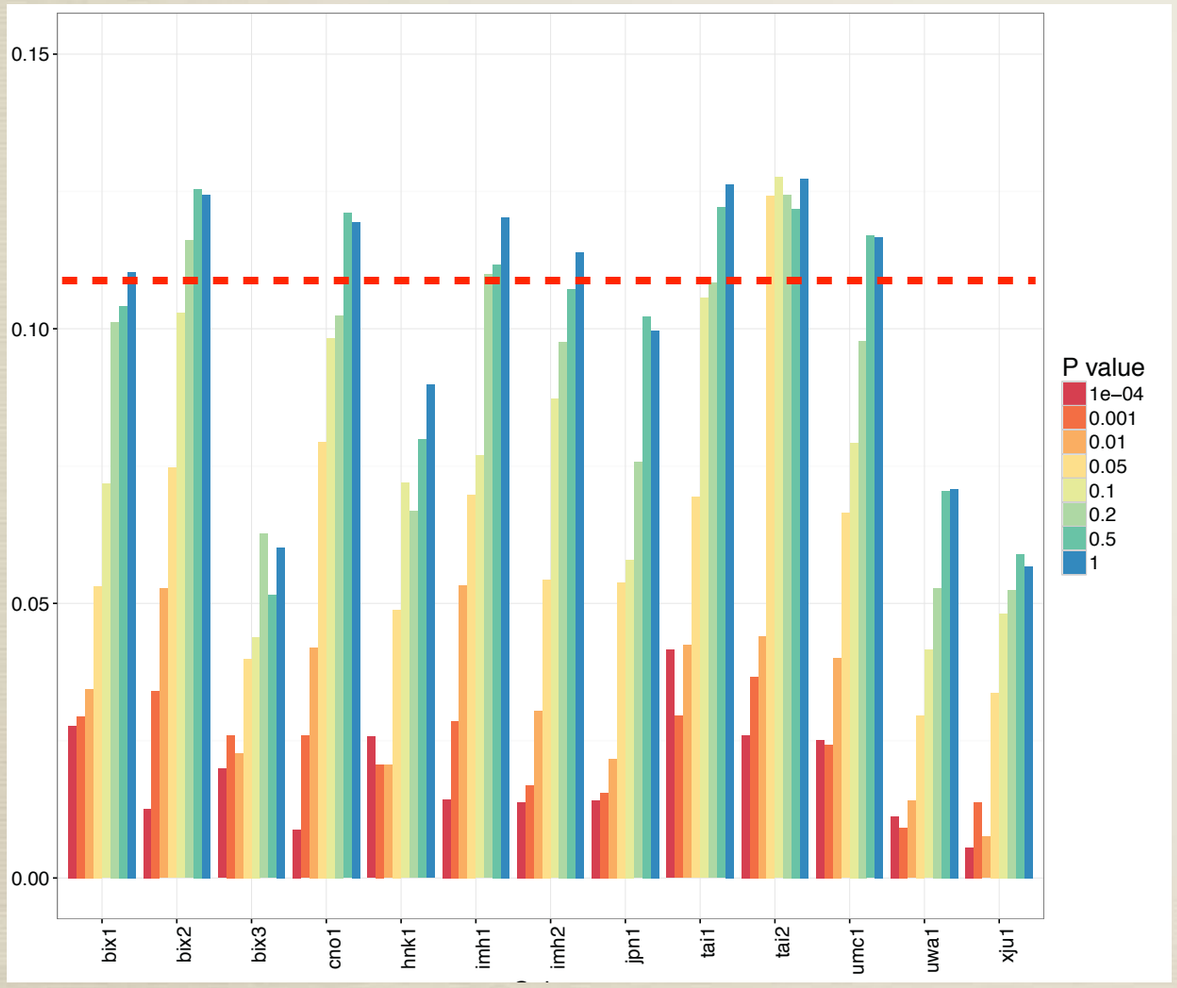
Prediction accuracy decays with genetic divergence



- Prediction accuracy is highest in the European discovery cohort
- The European bias diminishes the potential for clinical viability

Schizophrenia prediction accuracy recapitulates transferability issue

EAS training, EAS testing

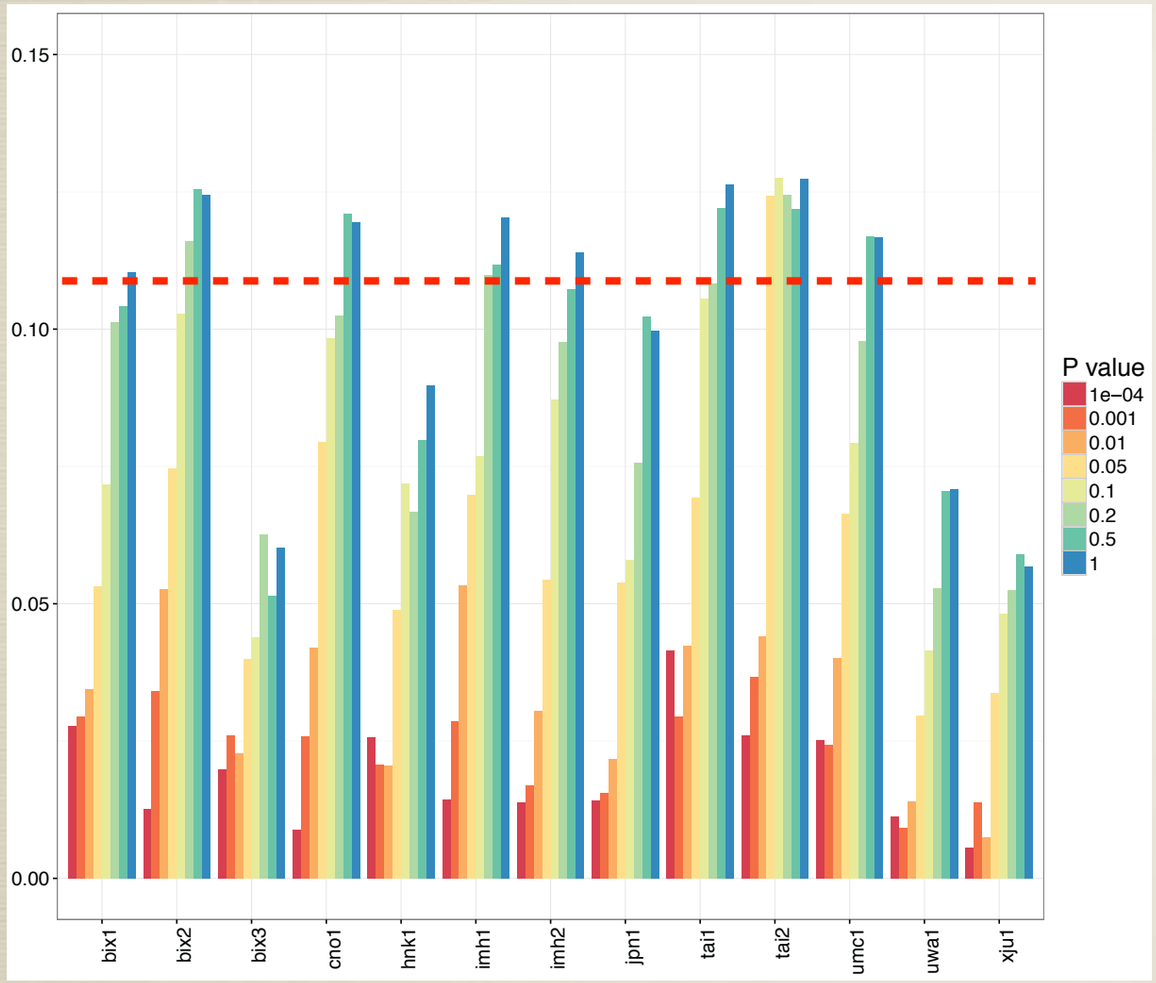


Multi-ethnic schizophrenia GWAS

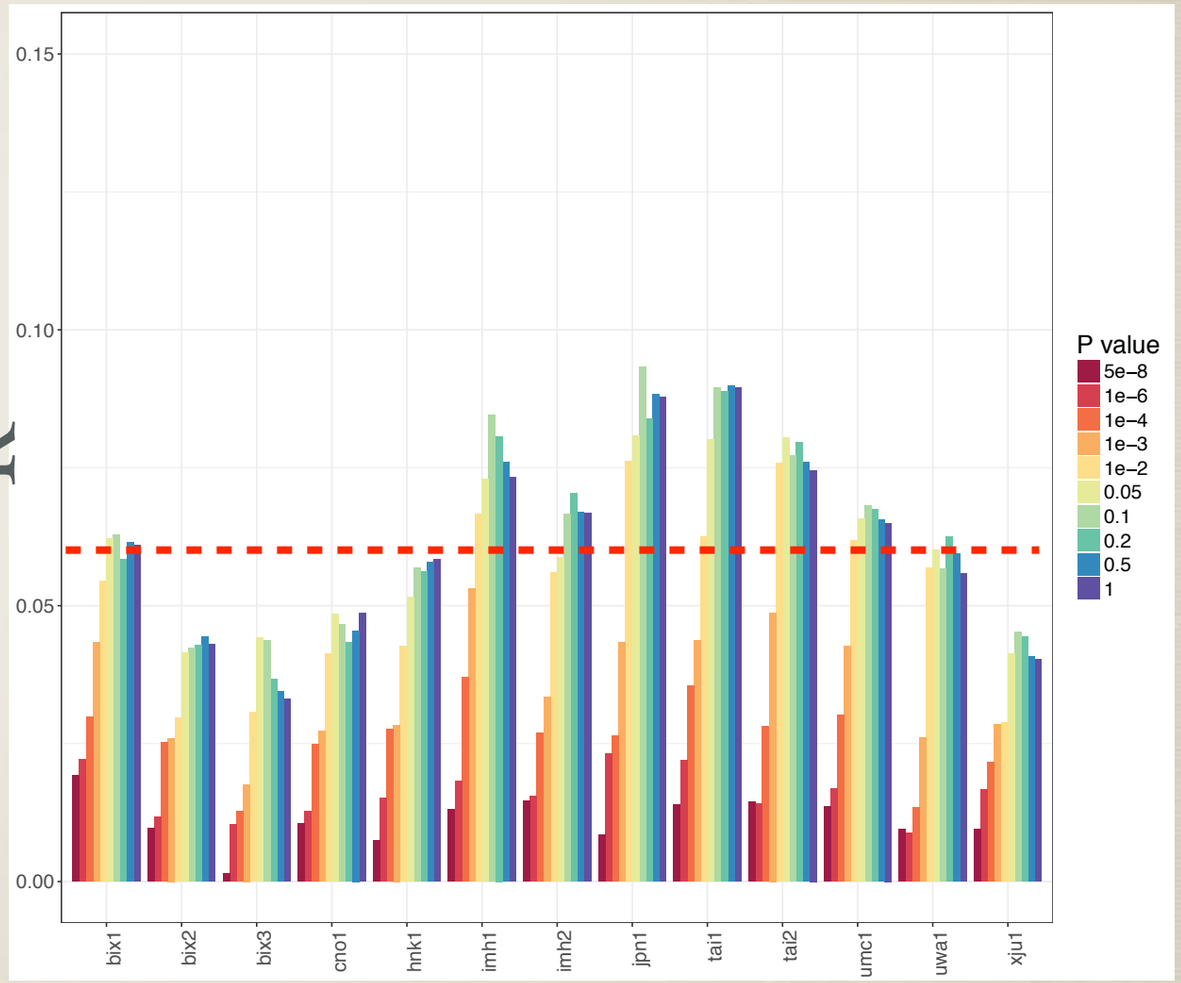
- ~37k and ~113k European cases and controls
- ~13k and ~16k East Asian cases and controls
- Prediction in East Asians from both populations

Schizophrenia prediction accuracy recapitulates transferability issue

EAS training, EAS testing



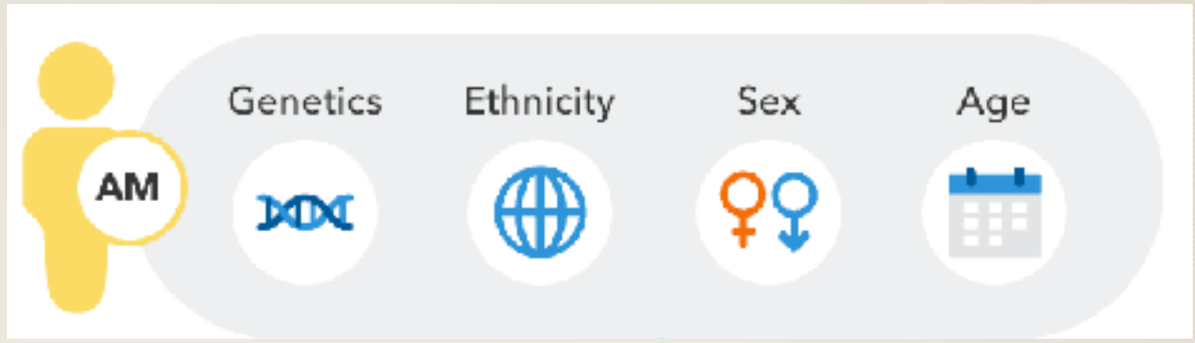
EUR training, EAS testing



Despite ~3X larger European sample size, prediction is 37% worse with European training data

Genetic risk scores are becoming widespread and translational

23andMe



Incorporating a Genetic Risk Score Into Coronary Heart Disease Risk Estimates

Effect on Low-Density Lipoprotein Cholesterol Levels (the MI-GENES Clinical Trial)

Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease

Polygenic risk score predicts prevalence of cardiovascular disease in patients with familial hypercholesterolemia

Conclusions

* Using large-scale genomics, we can learn about population history information modern structure
* GWAS studies and tools (e.g. imputation, arrays, statistical methods) are biased towards Europeans
* Polygenic risk scores are unpredictably biased across populations (not straightforward to correct with PCs alone)
* Clinical challenges of interpretability across populations cautions genomic health disparities

Future directions

* **<u>As a field: Increase diversity in genetic studies</u>**

* Developing better polygenic risk methods: use LD from both populations to correct effect size estimates

* Longer term: incorporate local ancestry in prediction

* Extending simulations: multiple populations are available

* Extending simulations: couple effect size and allele frequency (i.e. invoke selection)

$$\beta \sim N(0, f_i(1-f_i)^{\alpha}c)$$

$$\alpha = -0.35 \pm 0.05$$

Interested in African pop gen in NeuroDev/NeuroGAP?

Let's work together!

armartin@broadinstitute.org