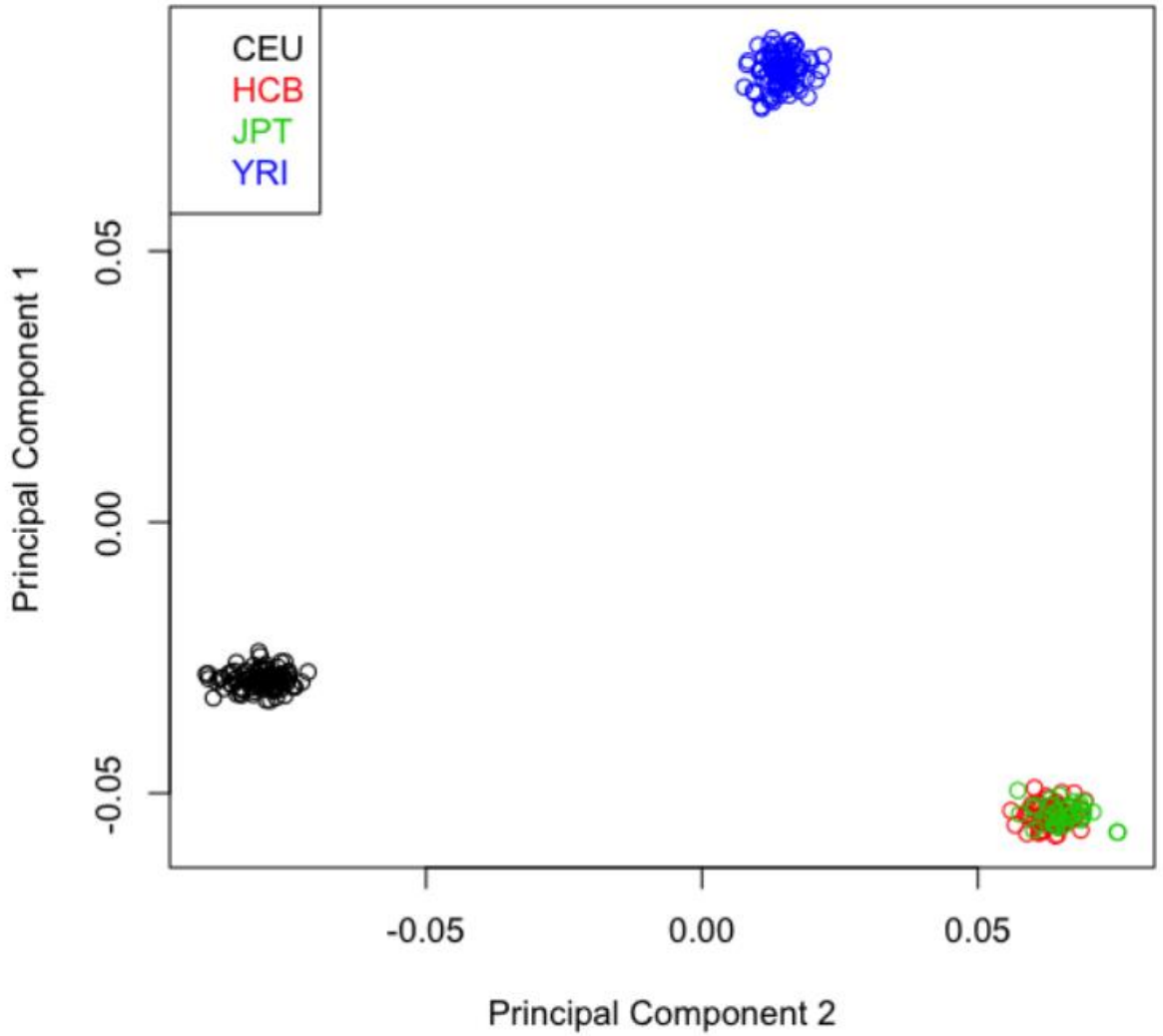
Is it really necessary to consider population differences?

(for the study of complex genetic disorders)

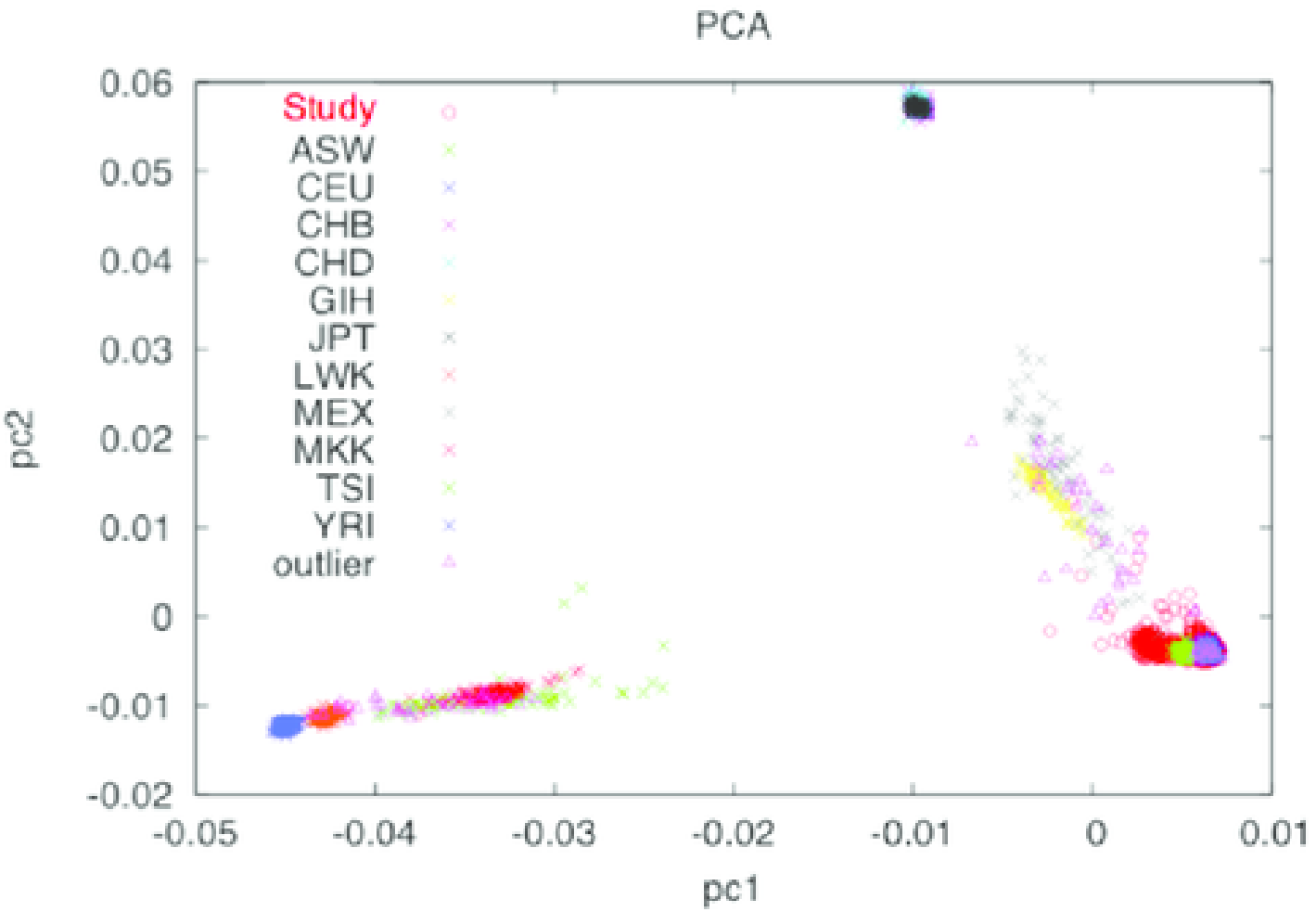
Andrew G. Clark

Cornell University

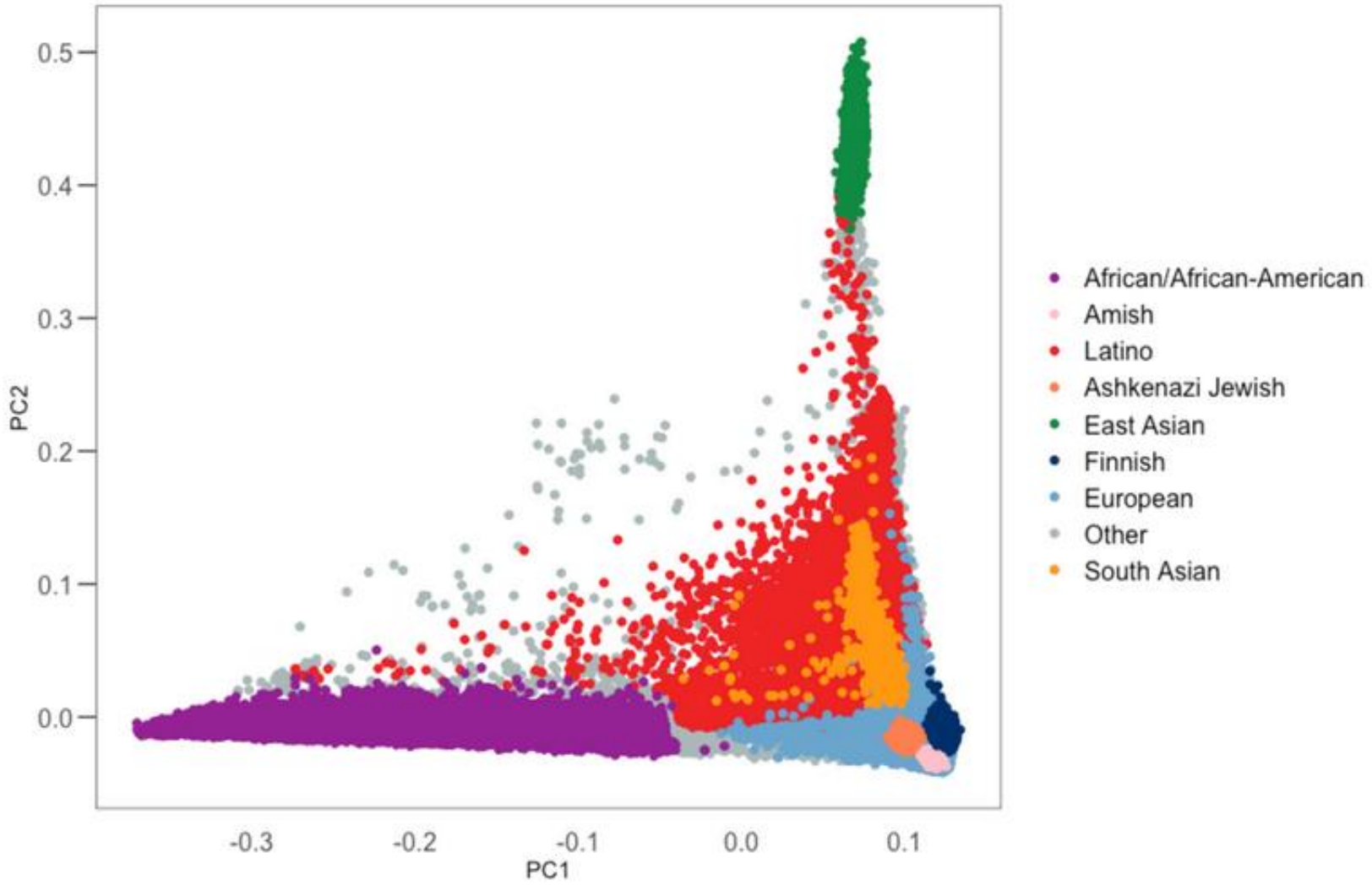
Initial view of human diversity (2000-2003)



HapMap picture of 10 human groups

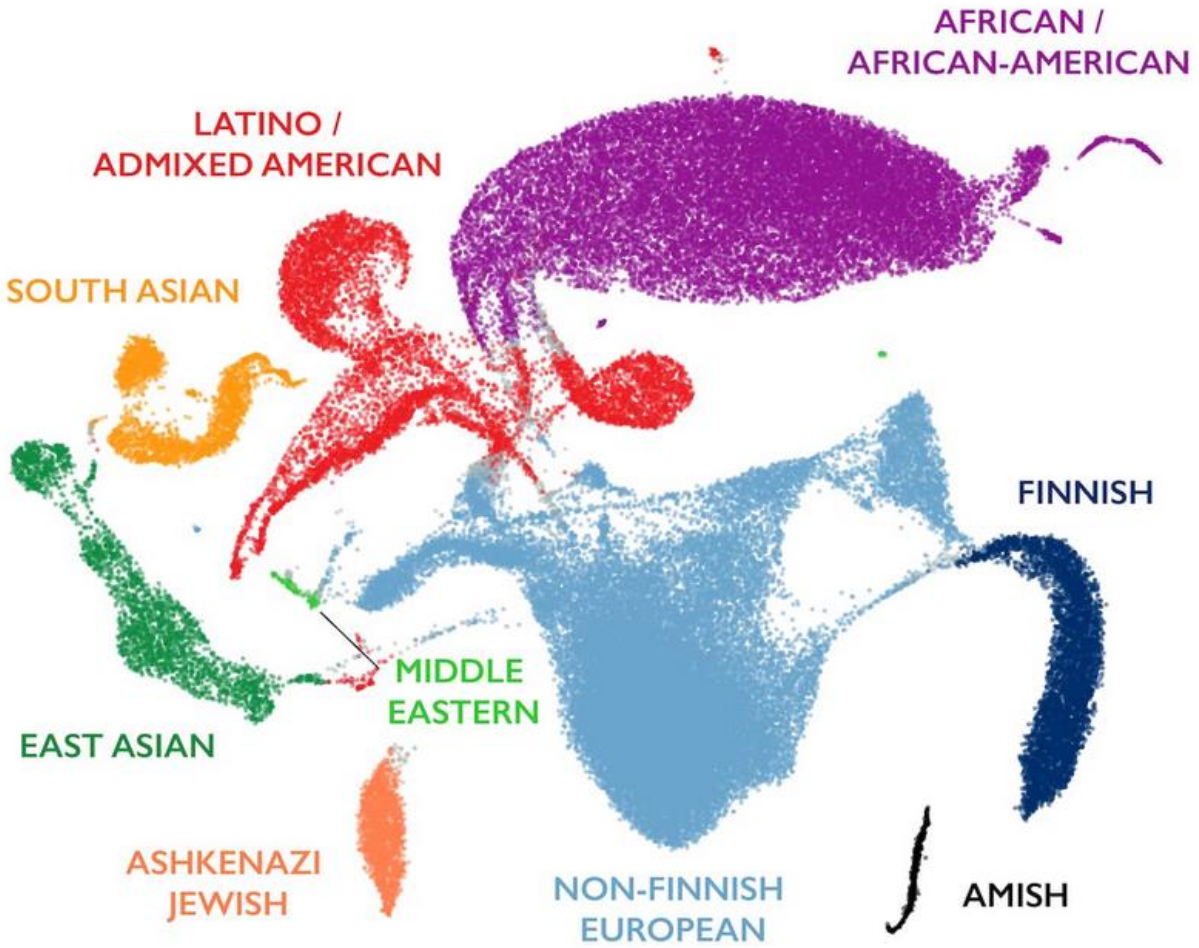


Current gNomad view of human diversity



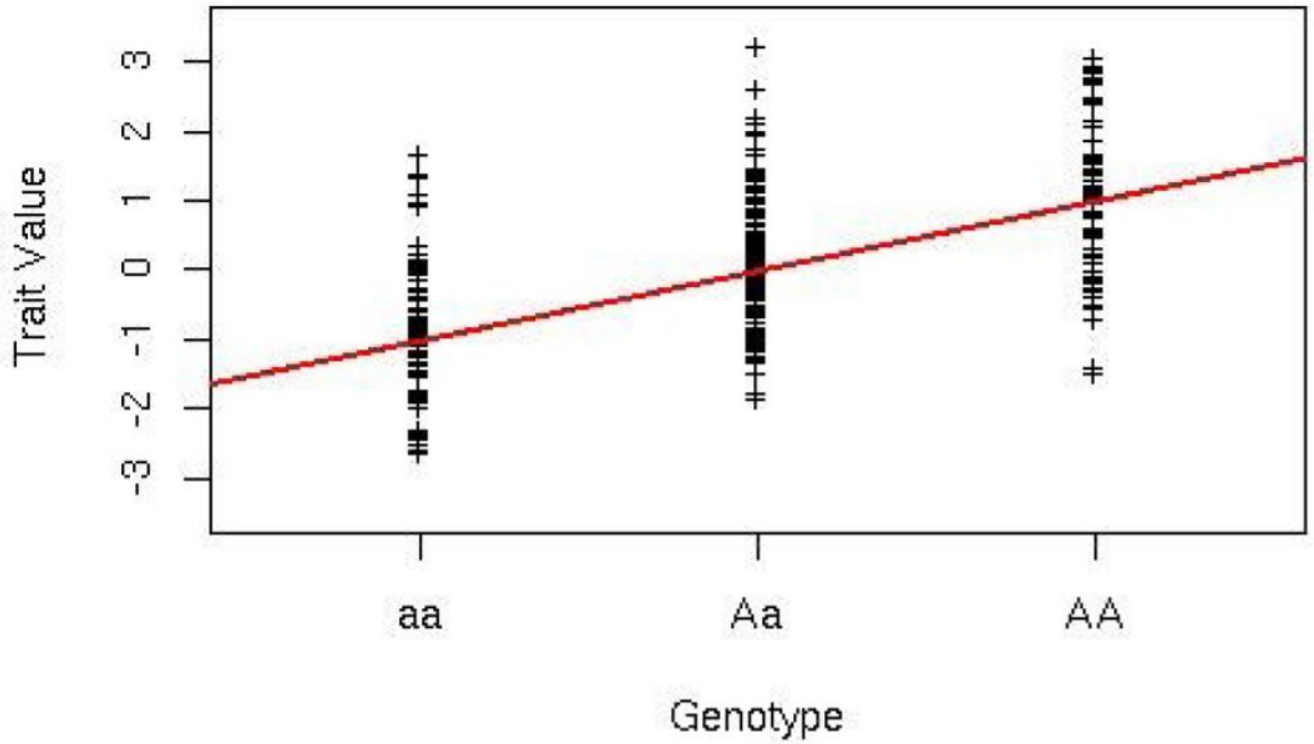
From gNomad online report

The same data can tease groups apart



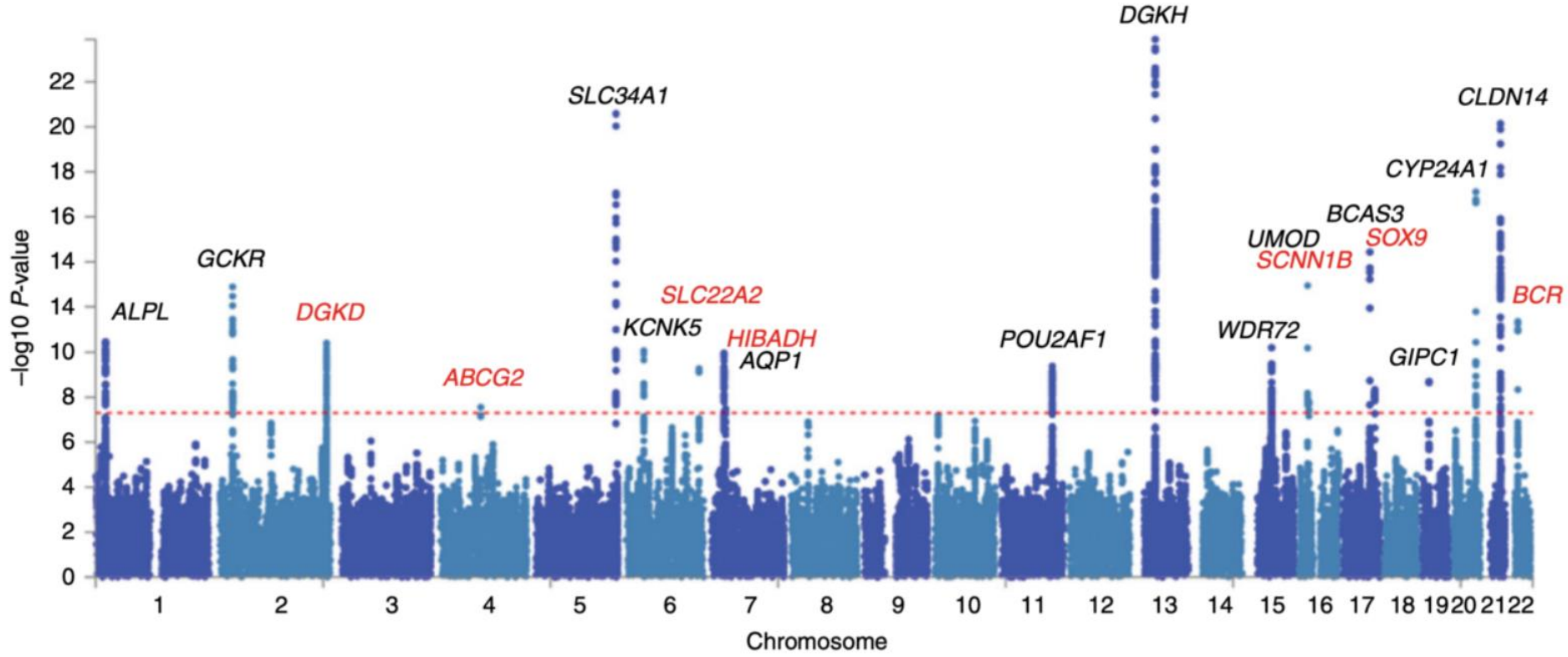
From gNomad online report

How GWAS is done



If the slope is different from zero, there is AN ASSOCIATION between the SNP and the phenotype.

GWAS results are displayed with a “Manhattan plot”



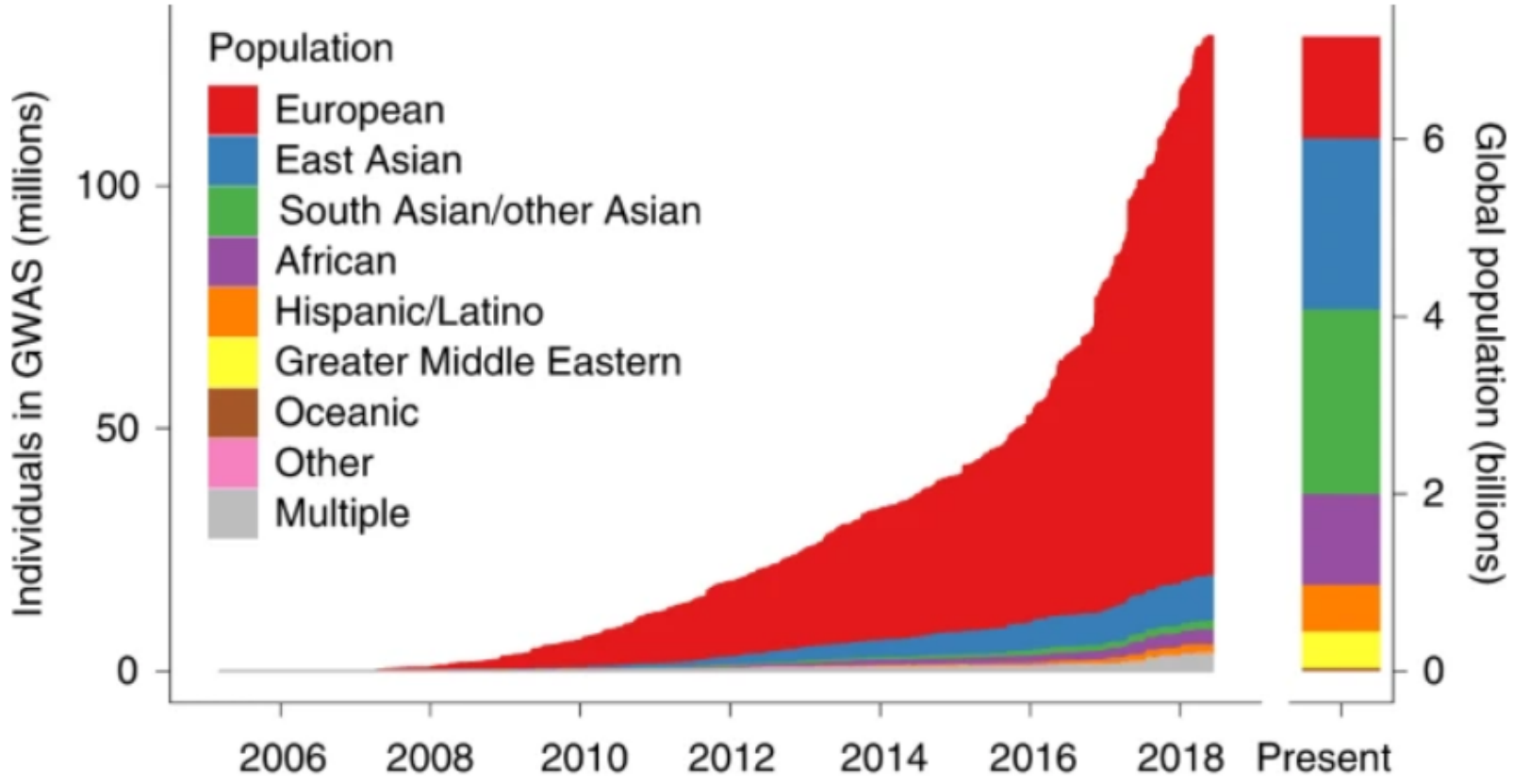
Population subdivision can cause a problem

- Suppose a disease is common in one population, and rare in another.
- If the data pool the sample into one mixed group, any SNP that is common in the population where the disease is common (and rare where the disease is rare) will appear to be associated.
- We call this a FALSE POSITIVE.

Solution to the problem of population substructure?

- Sample only from one homogeneous population.
- Include factors in the model that explain differences between the populations (e.g. Principal Components)
- Repeat the GWAS in each population!

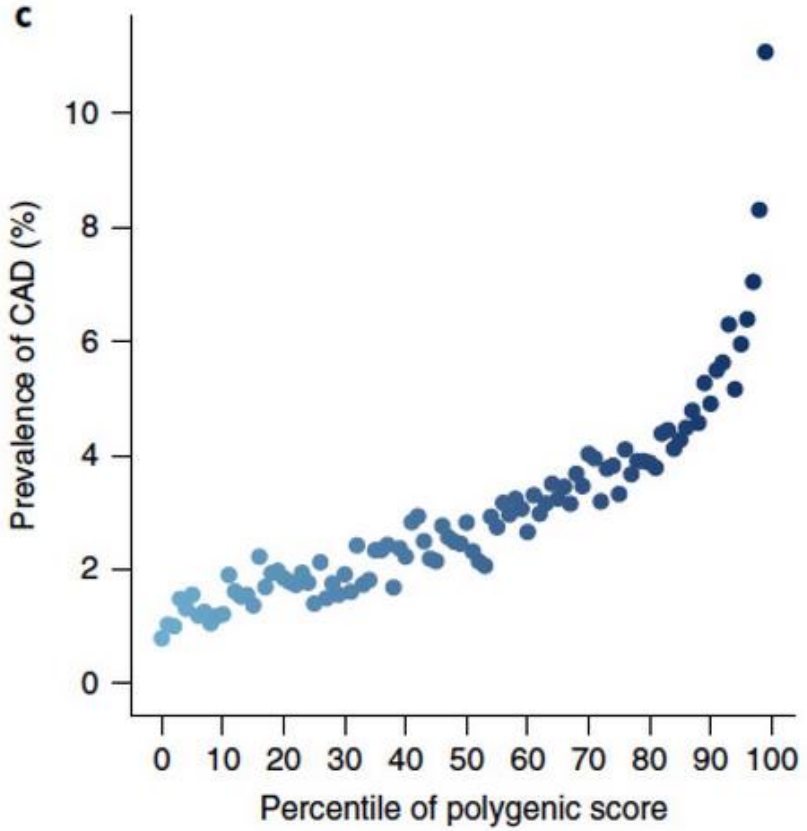
GWAS sampling has been highly biased



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

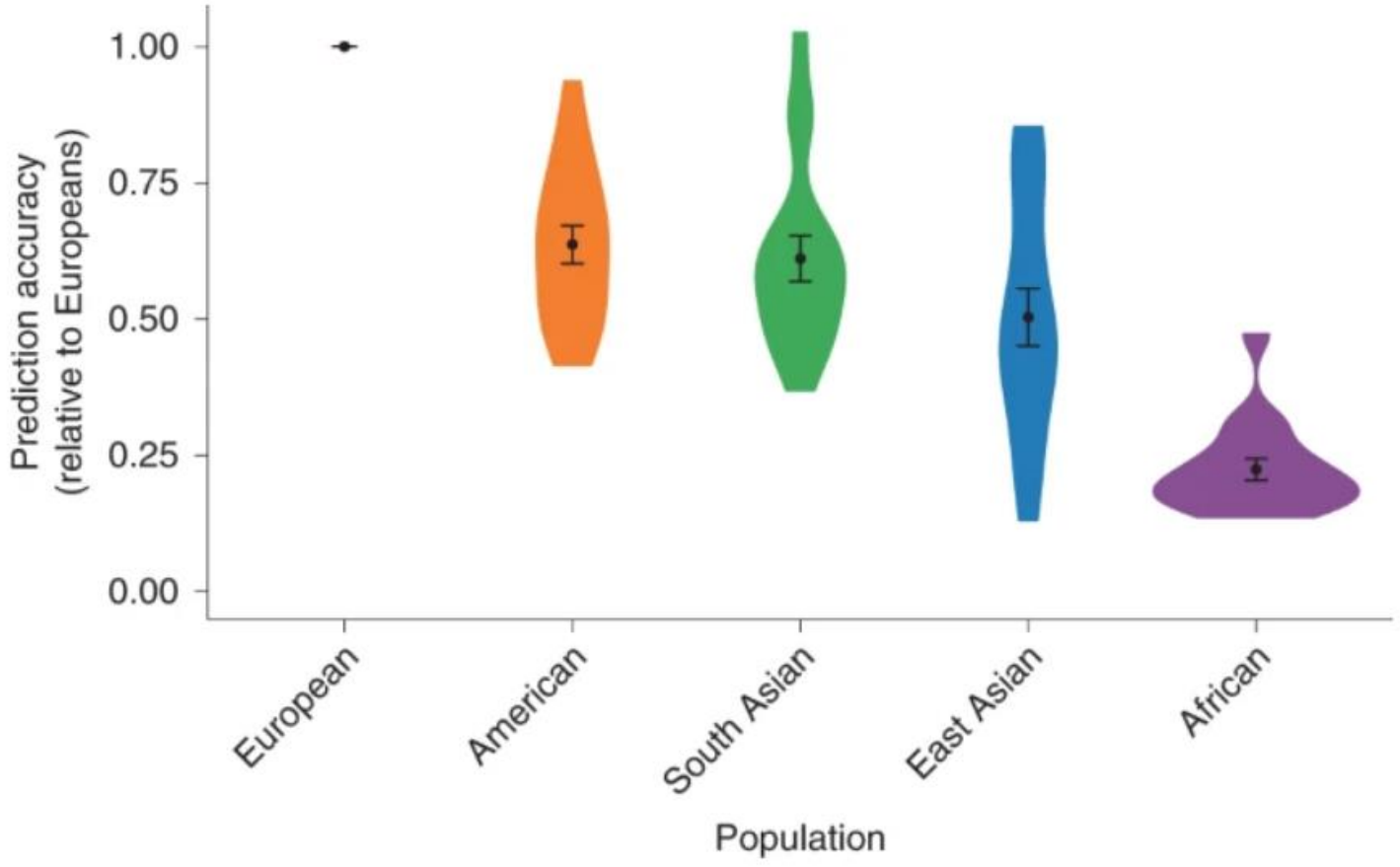
Polygenic scoring promises to be one means of genetic risk prediction

- Start with a very large, well-executed GWA study
- Determine which SNPs are most likely to be causal in their association.
- Use the slopes of their effects and each individual's genotype, and add up the effects over all SNPs – this is the PGS.



Poor transferability of Polygenic Risk prediction

- PGS is based on GWAS – depends on LD and so is population-specific



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

WHY not use statistical tricks to correct?

- Populations may differ in both allele frequencies and their correlations (Linkage disequilibrium).
- The relationship between a SNP and disease depends on both.
- Because we are trying to predict disease based on SNPs that are correlated with risk (but usually not directly causal), we need to have accurate estimates of these correlations.
- Mixed populations (and admixed individuals) pose a serious challenge to these methods.

Inference of segmental ancestry

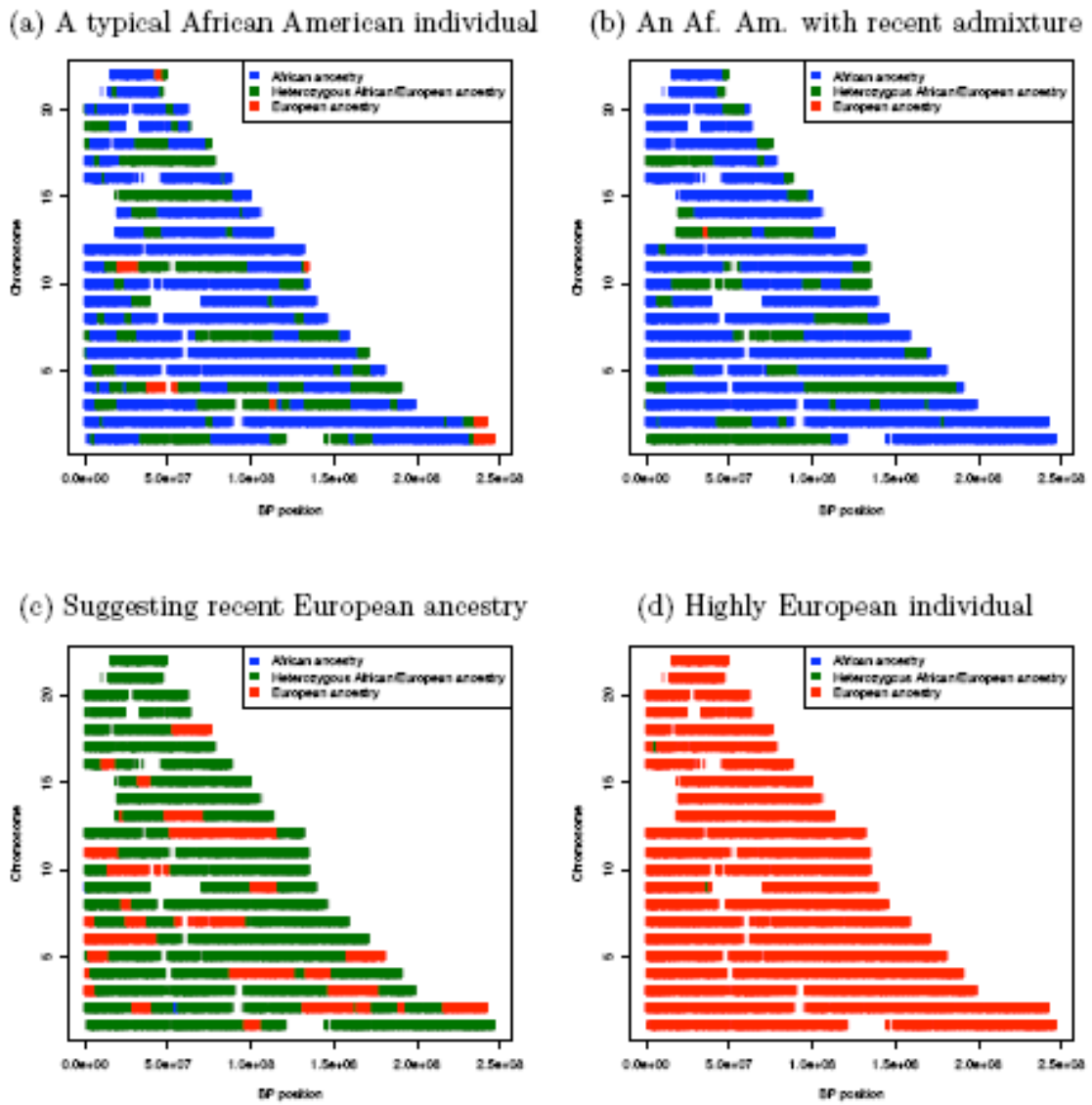


Figure 11: Individual ancestry estimates of several African American individuals.

Bryc et al. (2012)

Take-home conclusions

- Human populations do display differences in allele frequencies.
- Most of these differences are small, but in aggregate allow one to see differences even among closely related populations.
- There are also differences in the correlations between SNPs, including those that are associated with disease.
- Ideally GWAS would be done in each population group.
- GWAS and Polygenic Scoring cannot ignore population differences.
- Purely statistical correction is not yet working, but perhaps analysis of segmental ancestry of genomes will improve.