

Population Structure

Section 5

Podcast on the Human Genome Project

- <https://geneticsunzipped.com/blog/2020/10/22/s322-the-past-present-and-future-of-the-human-genome-project>

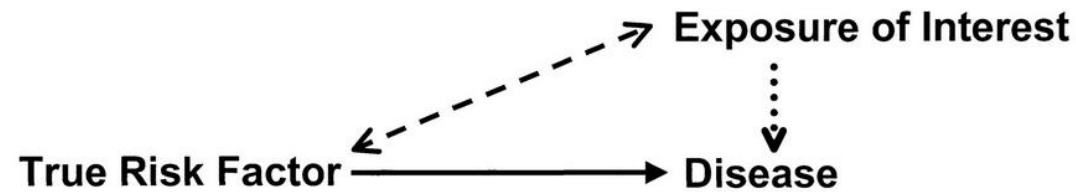
Learning Objectives

- Describe population substructure and how it can confound results.
- Understand methods for accounting for it in analysis.

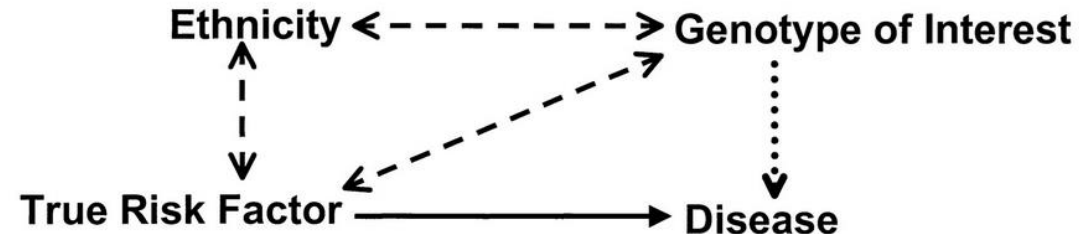
Population Substructure

The presence of a systematic difference in allele frequencies between subpopulations due to different ancestry

Confounding



Population Stratification

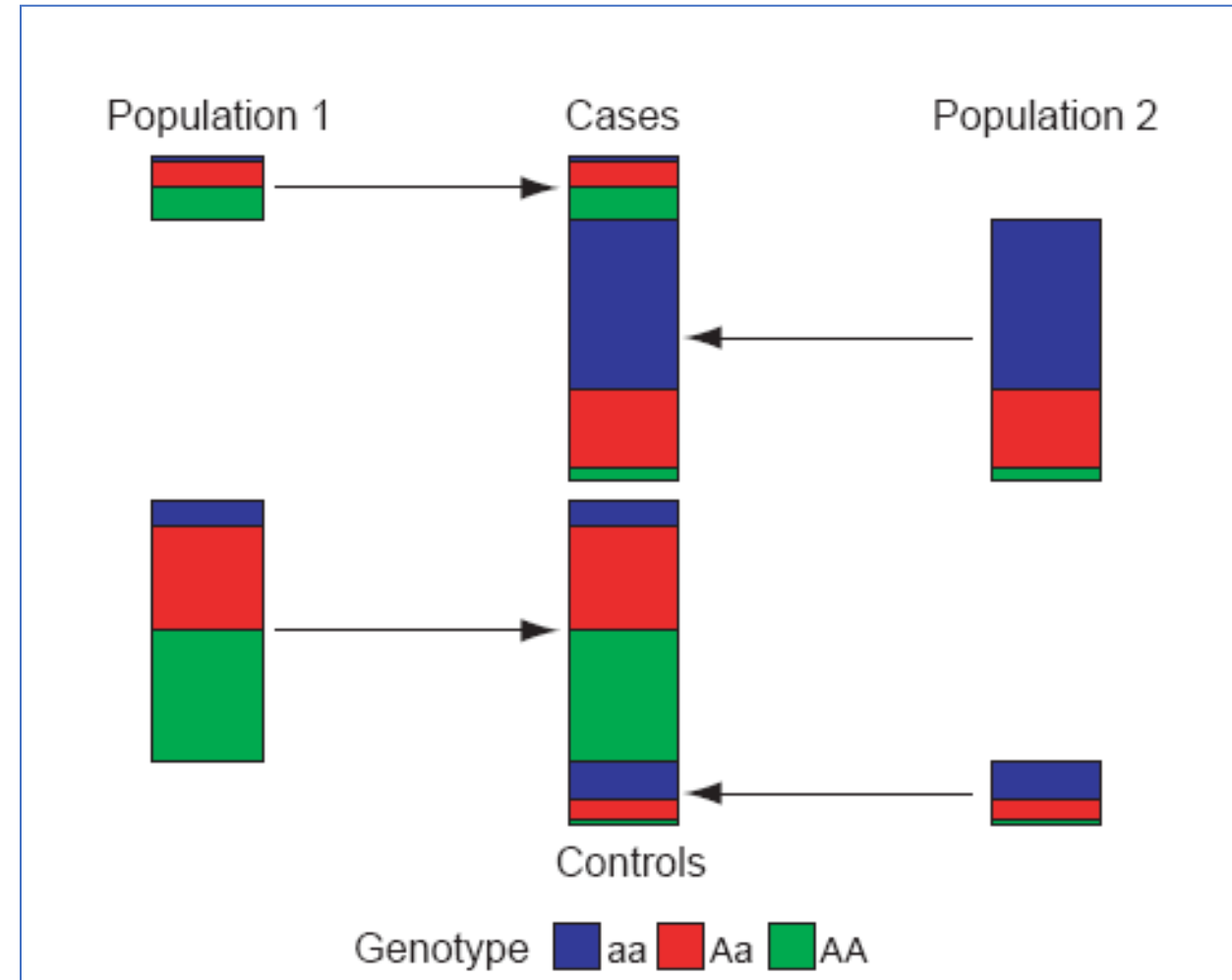


Assume we conduct a case-control GWAS...

- Our cases were collected in Africa
- Our controls were collected in Asia
- If we find multiple SNPs that are significantly more/less common in cases than controls, **do we believe that these results are due to association with disease or population differences?**

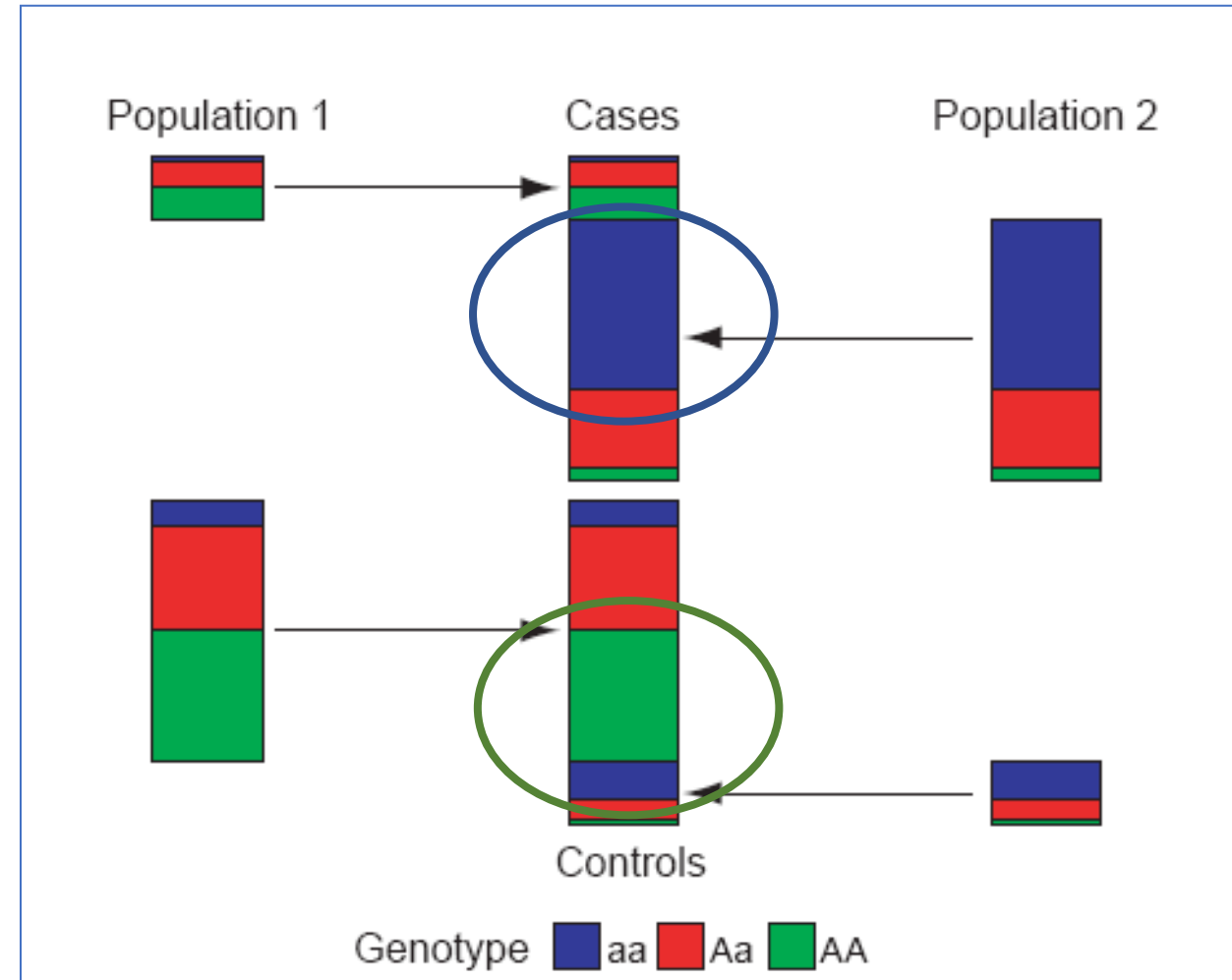
Population Stratification - Confounding by ancestry

Group differences in ancestry
AND outcome



Population Stratification - Confounding by ancestry

Group differences in ancestry
AND outcome



But these are very obviously different populations...

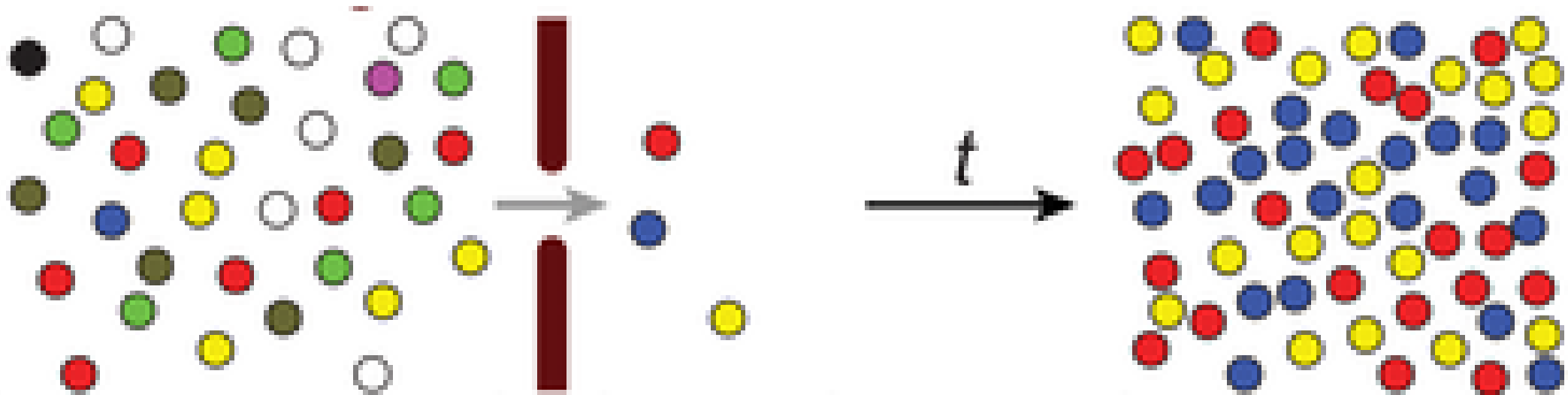
What about more subtle differences?

Relics of human history are present across the genomes, making some genetic variants more/less common in different populations, even if the variants don't have any impact on human traits or health.

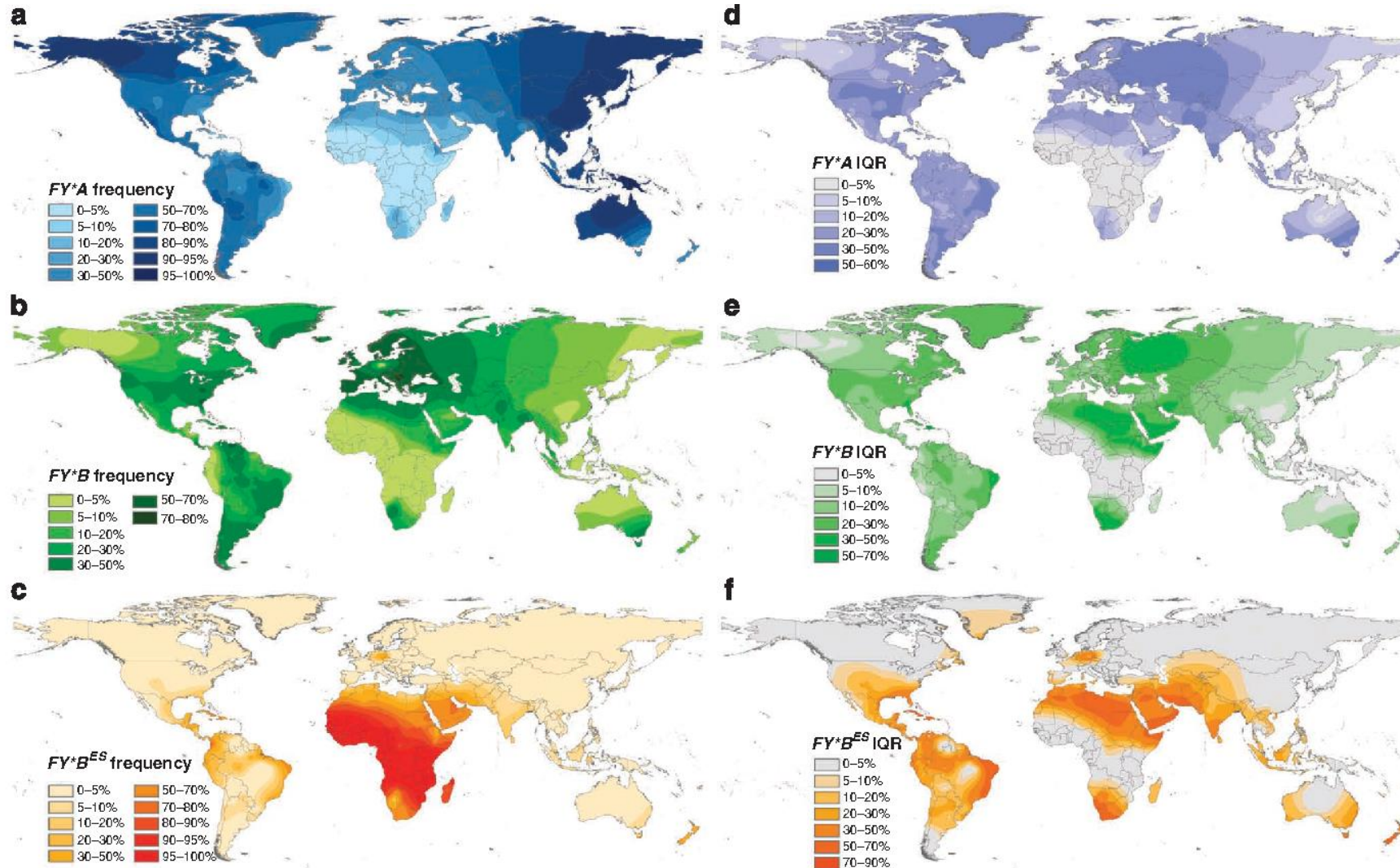
We can use genetic data to determine ancestry and to adjust for ancestry in association studies.



Slight changes in allele frequencies with every population migration/expansion



Global allele frequencies are clinal, with different patterns for every allele...

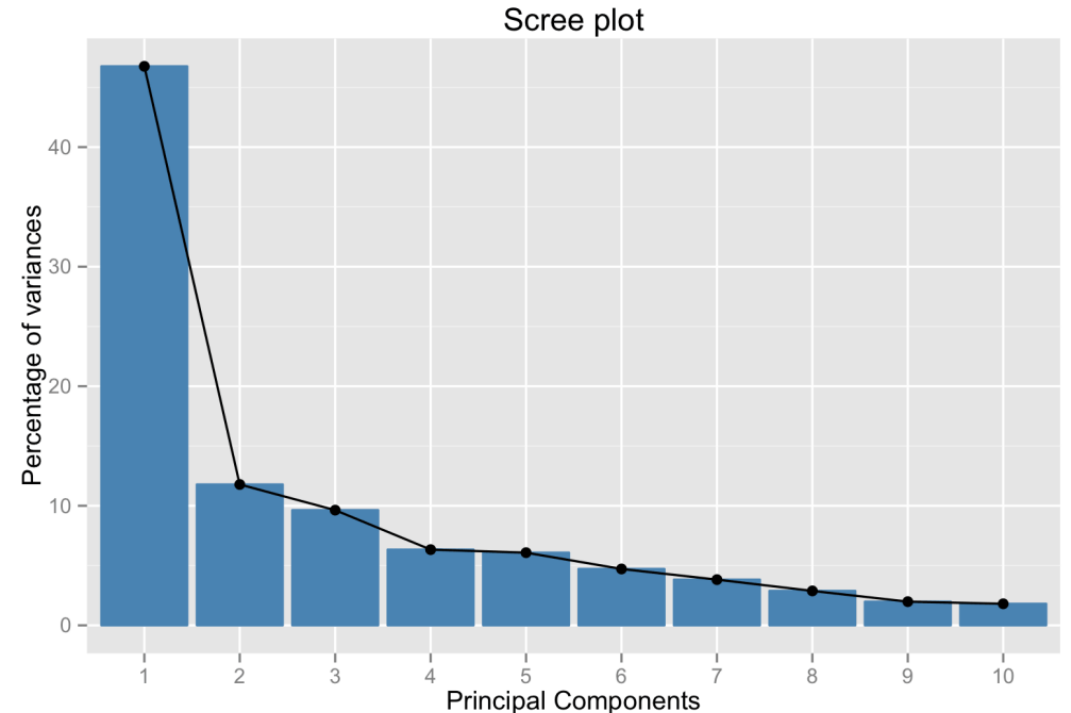


We can use all of these clinal patterns together to “adjust” away the background population substructure effects.

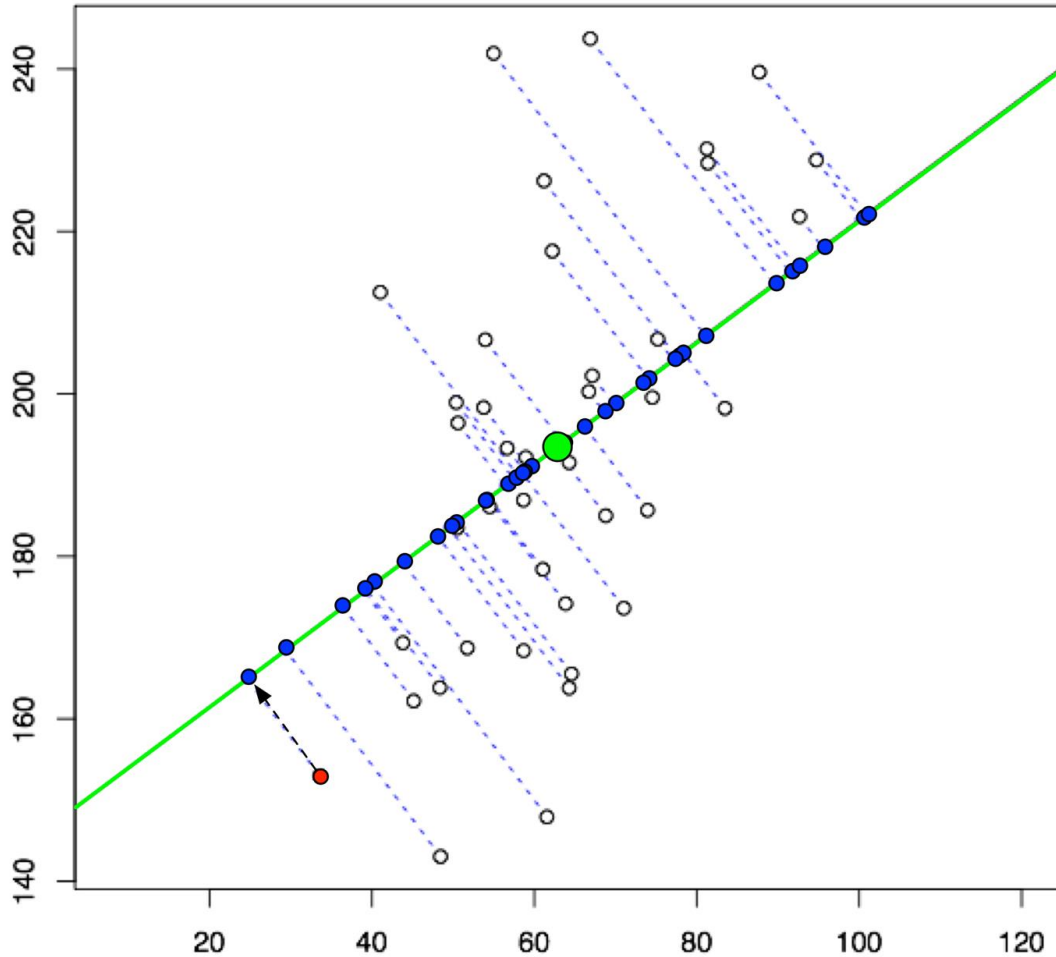
=> Principal Component Analysis

Principal Component Analysis (PCA)

- Reduces the dimension of the data from many, many variables to a small set (“principal components ” or “PCs”– eigenvectors) that still explain the majority of variation seen in the data.
- The first PC (PC1) is constructed to explain as much of the variation as possible, the second (PC2) is constructed to explain as much of the remaining variation as possible....
- The more correlation in the data (i.e. between SNPs), the fewer PCs are needed to explain most of the variation.
- Each PC is a linear combination of the original variables (SNPs)
- PCs are independent of each other.



How does PCA work?

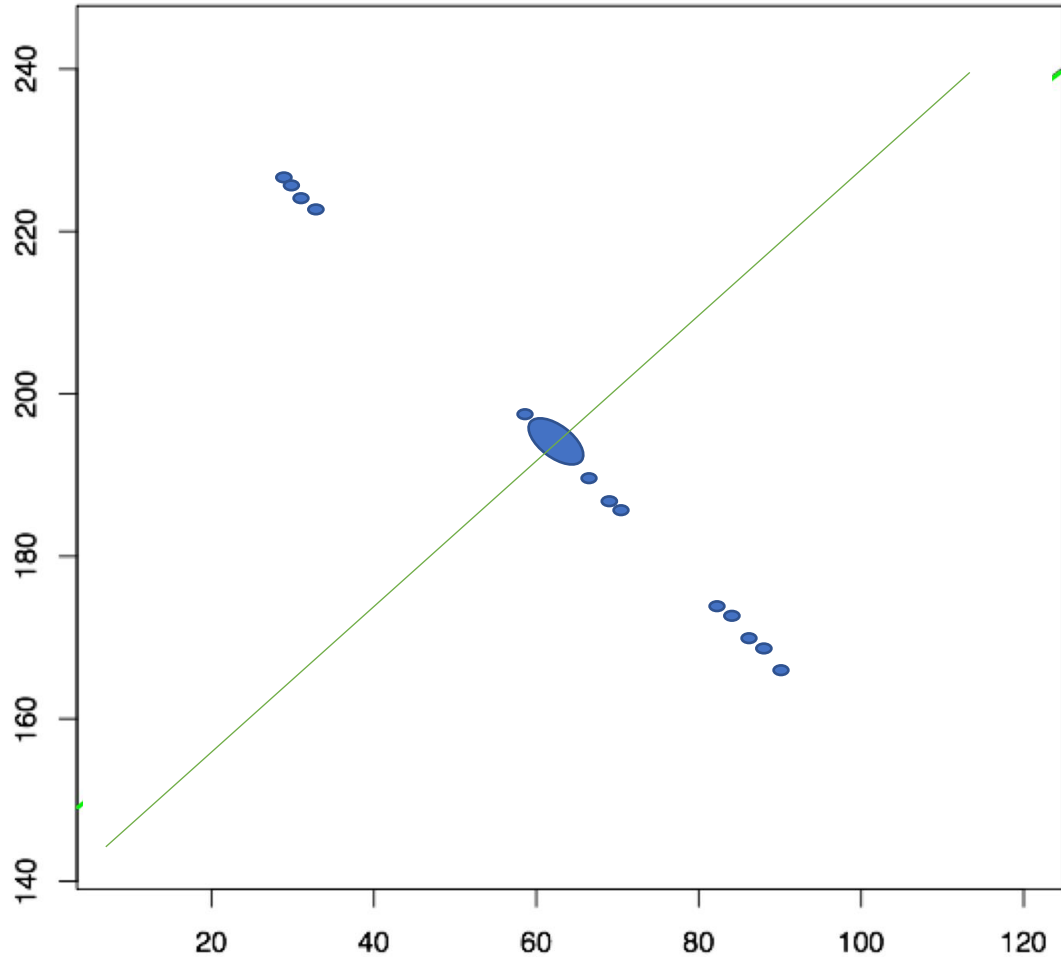


Each PCA maximizes variance and minimizes error

Basically to “absorb any systematic differences.”

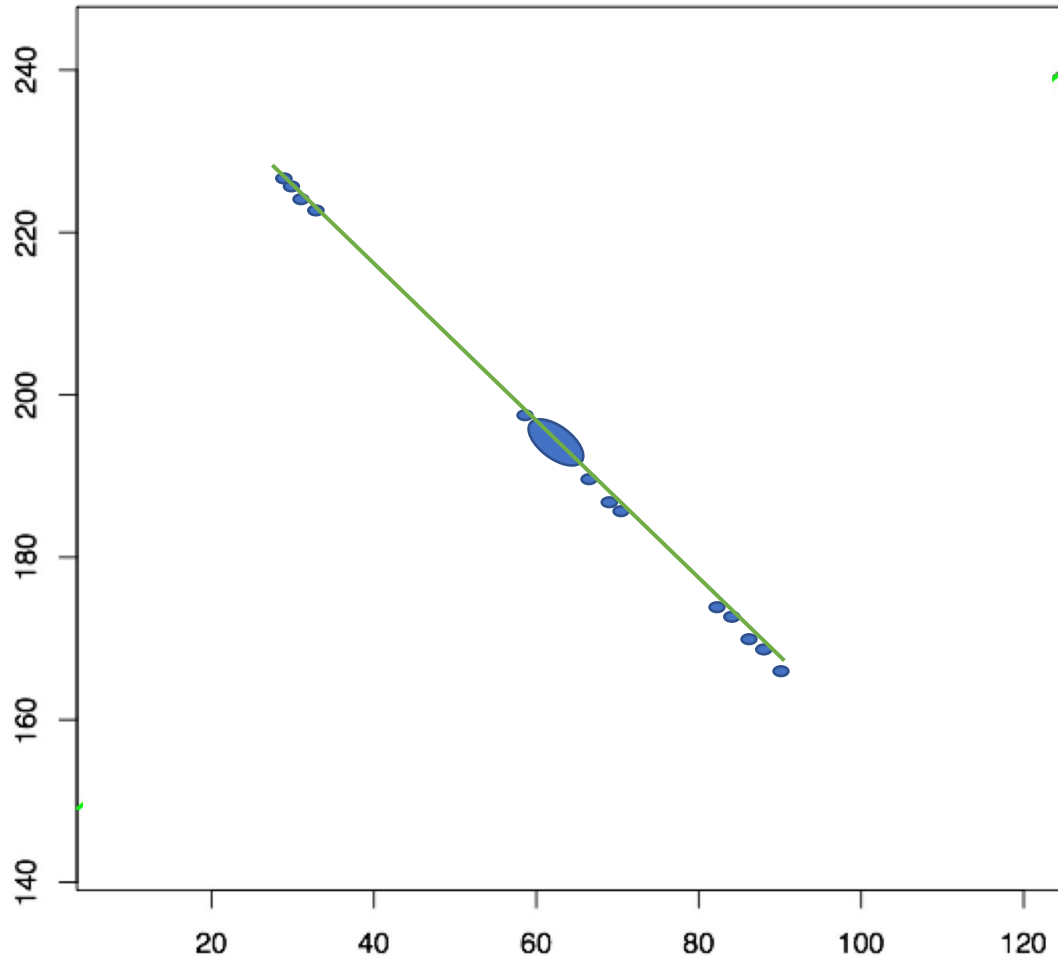
Reduces data dimensions.

Second PCA “soaks up” leftover variance



Remove the dimension from PC1 so that every point is squished together with zero variance along PC1 axis

Second PCA “soaks up” leftover variance

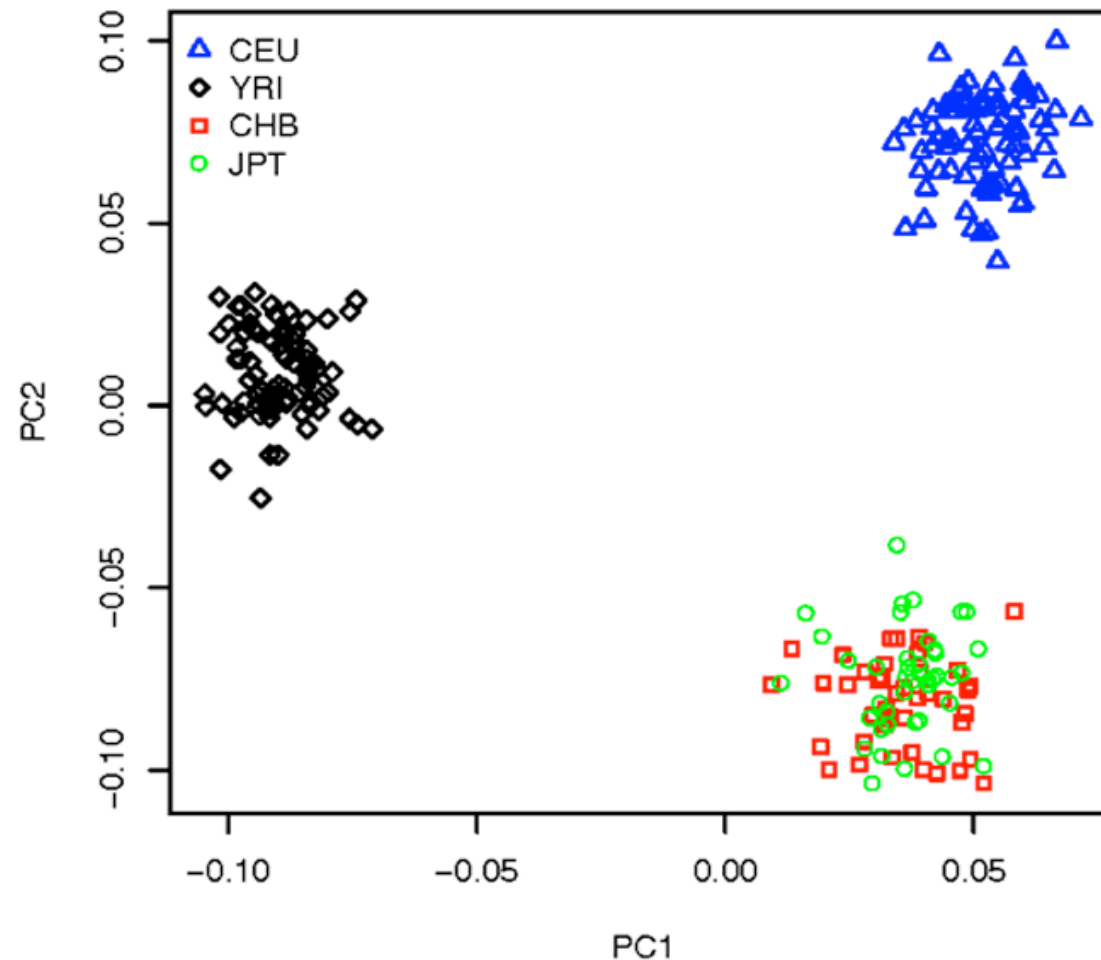


Remove the dimension from PC1 so that every point is squished together with zero variance along PC1 axis

Now, PC2 absorbs the most variance from whatever is left after PC1 dimension is removed

The first two PCs can help distinguish ancestral populations

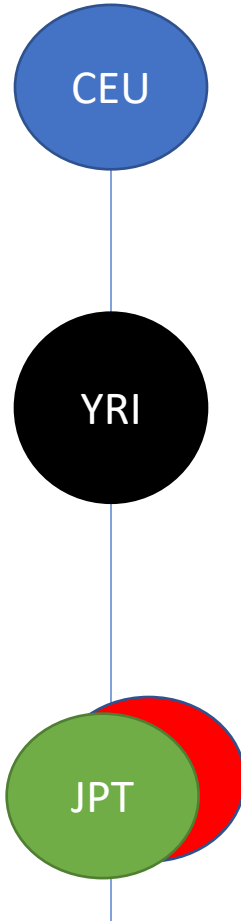
SNP Panel 93



PC1



PC2



Zoom breakout

- Interpret the plot of principal component 1 vs principal component 2 in a sample of people from 4 populations.

Calculate each PC for each individual

- Translate each genotype into 0, 1, 2 (Additive form), with homozygous reference allele = 0, heterozygous = 1, homozygous variant = 2.
- Multiply that value * loading value for every SNP location
- Add all of those values together.

Individual	SNP1	SNP2	SNP3	SNP4	PC1 total
A	2	1	1	0	
B	1	0	2	1	

Calculate each PC for each individual

- Translate each genotype into 0, 1, 2 (Additive form), with homozygous reference allele = 0, heterozygous = 1, homozygous variant = 2.
- Multiply that value * loading value for every SNP location
- Add all of those values together.

Individual	SNP1 Loading = 4	SNP2 Loading = 0.3	SNP3 Loading = -2	SNP4 Loading = 1	PC1 total
A	$2*4 = 8$	$1*0.3 = 0.3$	$1*-2 = -2$	$0*1 = 0$	
B	$1*4 = 4$	0	$2*-2 = -4$	$1*1=1$	

Calculate each PC for each individual

- Translate each genotype into 0, 1, 2 (Additive form), with homozygous reference allele = 0, heterozygous = 1, homozygous variant = 2.
- Multiply that value * loading value for every SNP location
- Add all of those values together.

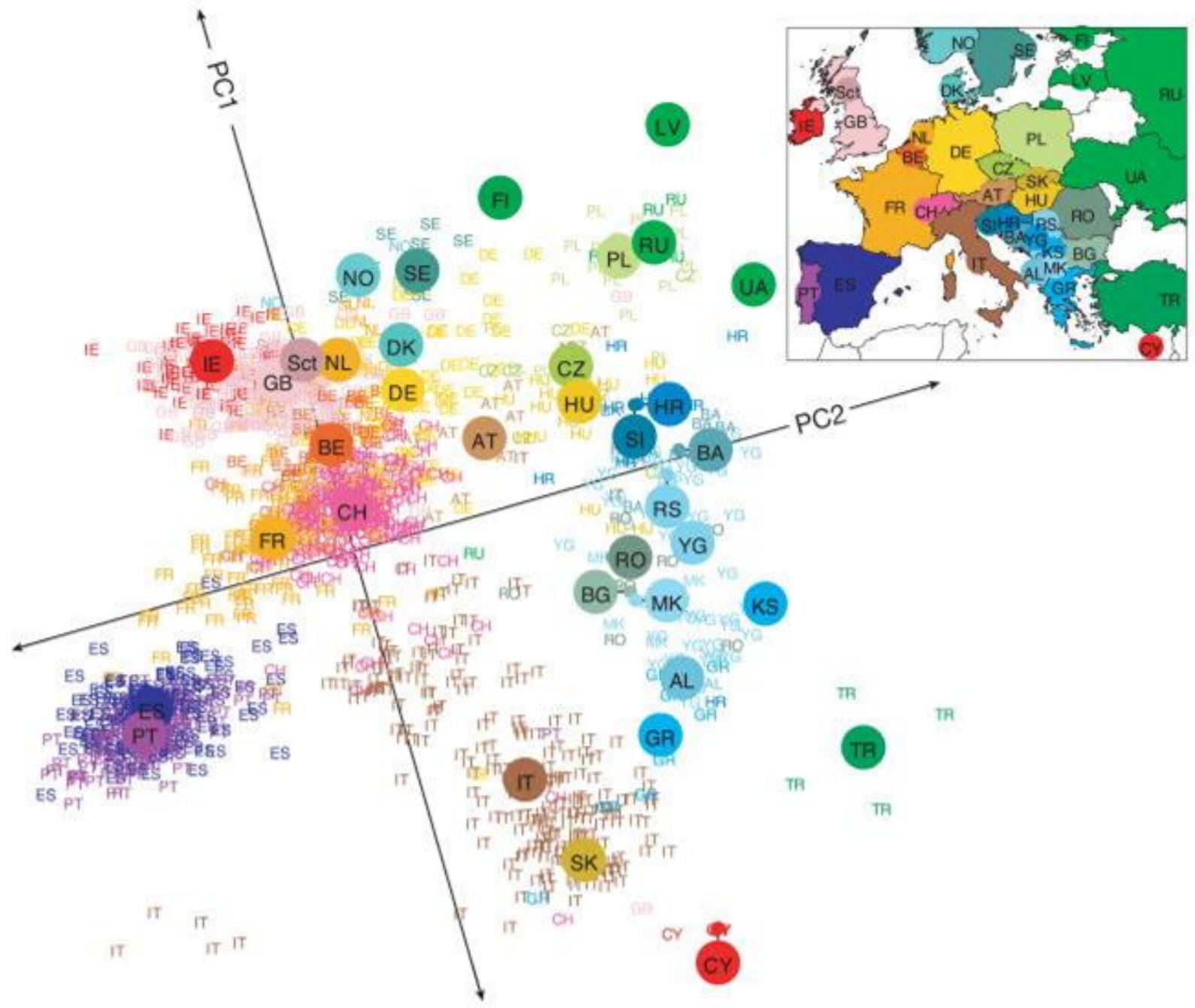
Individual	SNP1 Loading = 4	SNP2 Loading = 0.3	SNP3 Loading = -2	SNP4 Loading = 1	PC1 total
A	$2*4 = 8$	$1*0.3 = 0.3$	$1*-2 = -2$	$0*1 = 0$	6.3
B	$1*4 = 4$	0	$2*-2 = -4$	$1*1=1$	1

Include top PCs in genetic association study

$$\text{phenotype} = m * \text{genotype} + a\text{PC1} + b\text{PC2} + c\text{PC3} + d\text{PC4} + e\text{PC5}$$

Accounts for underlying gradient patterns that aren't truly associated with a phenotype, but may appear so due to allele frequency differences.

a



Map View



Sub-regional Resolution

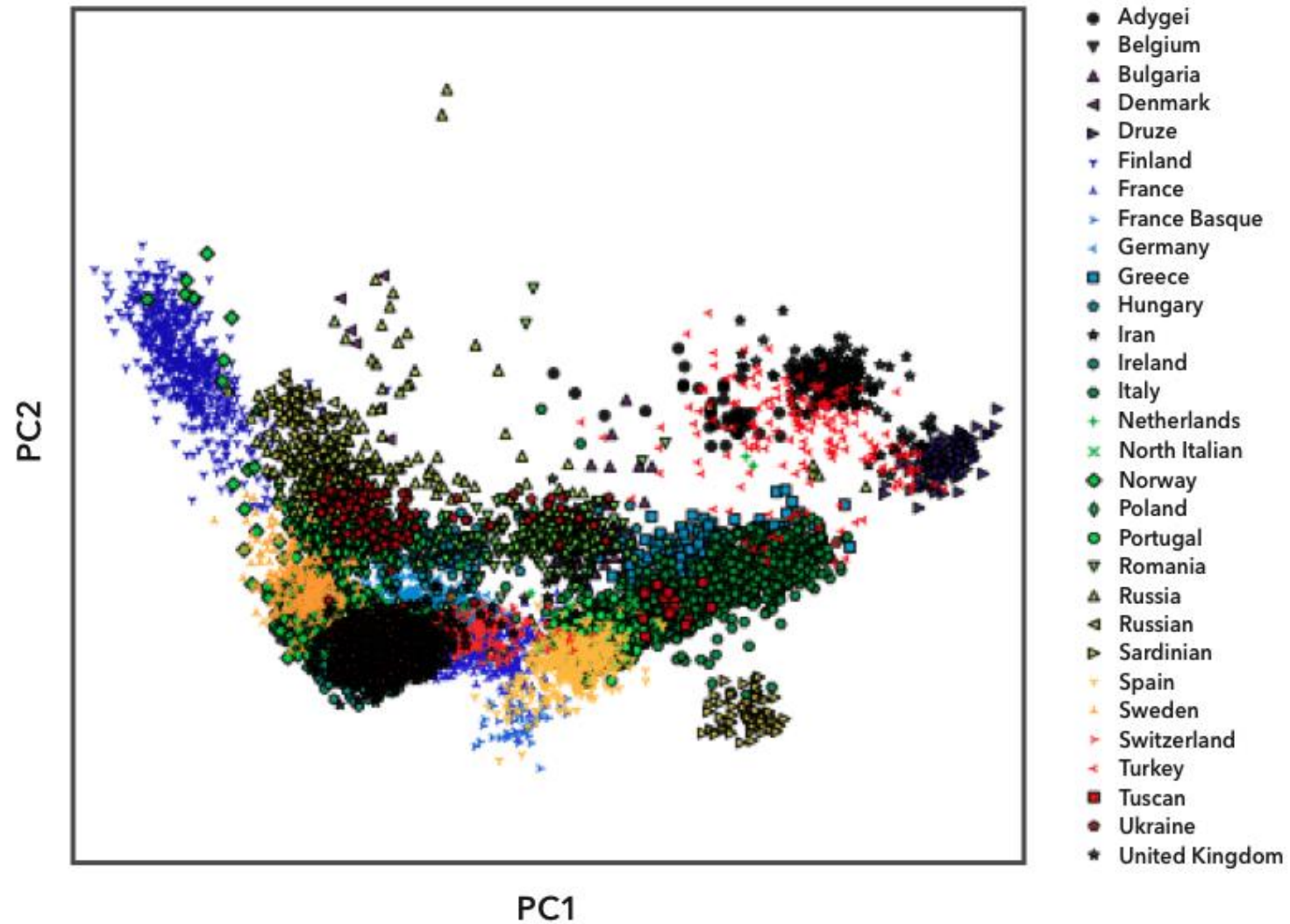


Ancestry Composition tells you what percent of your DNA comes from each of 22 populations worldwide. The analysis includes DNA you received from all of your ancestors, on both sides of your family. The results reflect where your ancestors lived 500 years ago, before ocean-crossing ships and airplanes came on the scene.



61.7%	European
8.9%	Northern European
5.0%	British and Irish
19.0%	French and German
	Nonspecific Northern Eur...
6.6%	Southern European
5.2%	Italian
	Nonspecific Southern Eur...
9.1%	Eastern European
2.3%	Ashkenazi
5.6%	Nonspecific European
37.1%	Sub-Saharan African
1.2%	East Asian & Native American
1.0%	Native American
0.2%	East Asian
< 0.1%	Unassigned
100.0%	Sheridan Smith

PCA matching at 300bp genomic windows - 23andMe



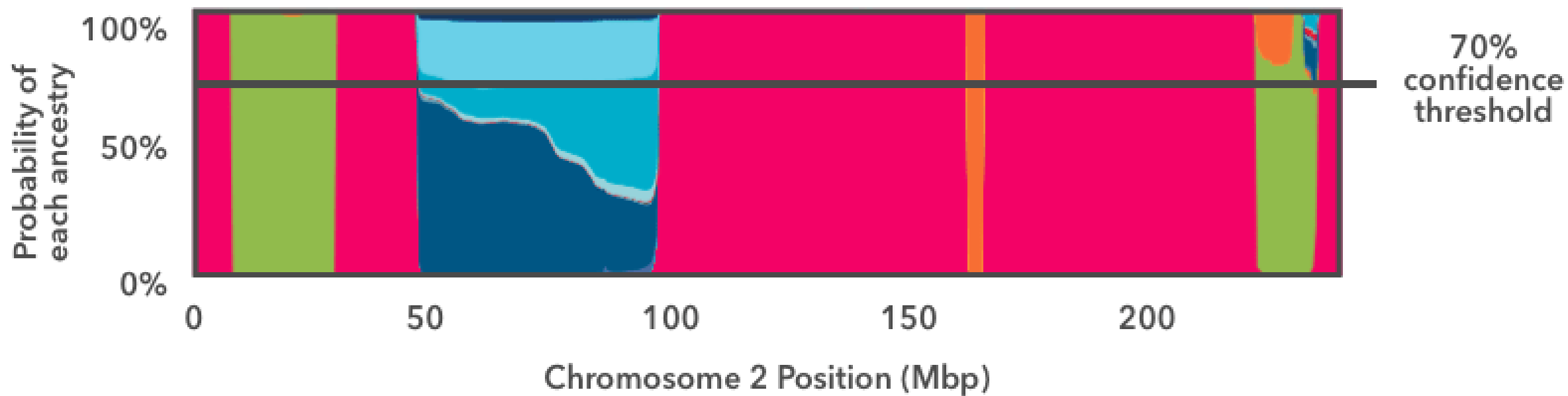


For matching ancestry - 23andMe

Reference data sets!!

Reflecting populations that existed before transcontinental travel and migration were common (at least 500 years ago). People who report four grandparents all born in the same country are included in the reference data.

DNA segment and ancestry probability



- Sub-Saharan African
- British & Irish
- Iberian
- Sardinian
- French & German
- Native American
- East Asian
- Italian

Summary

- Population structure can confound genetic association studies, but using principal component analysis can reveal and adjust.