Population Structure

Section 5

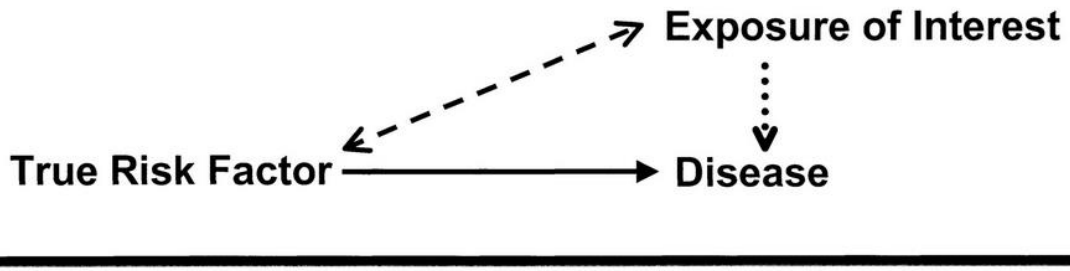
Learning Objectives

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

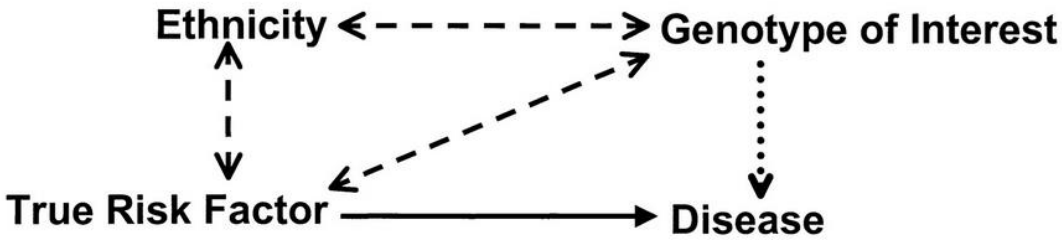
Population Substructure

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

Confounding



Population Stratification

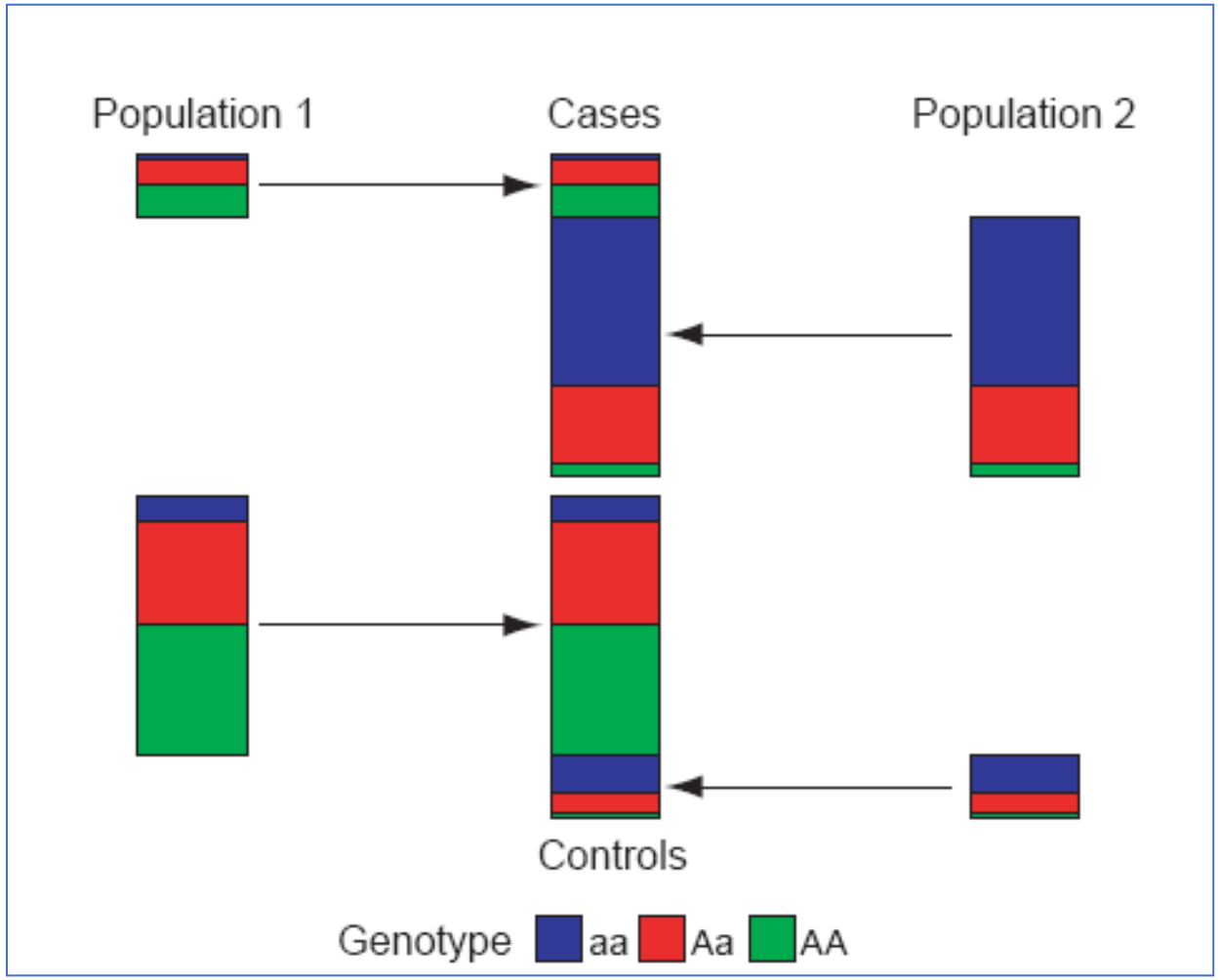


Assume we conduct a case-control GWAS…

- 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

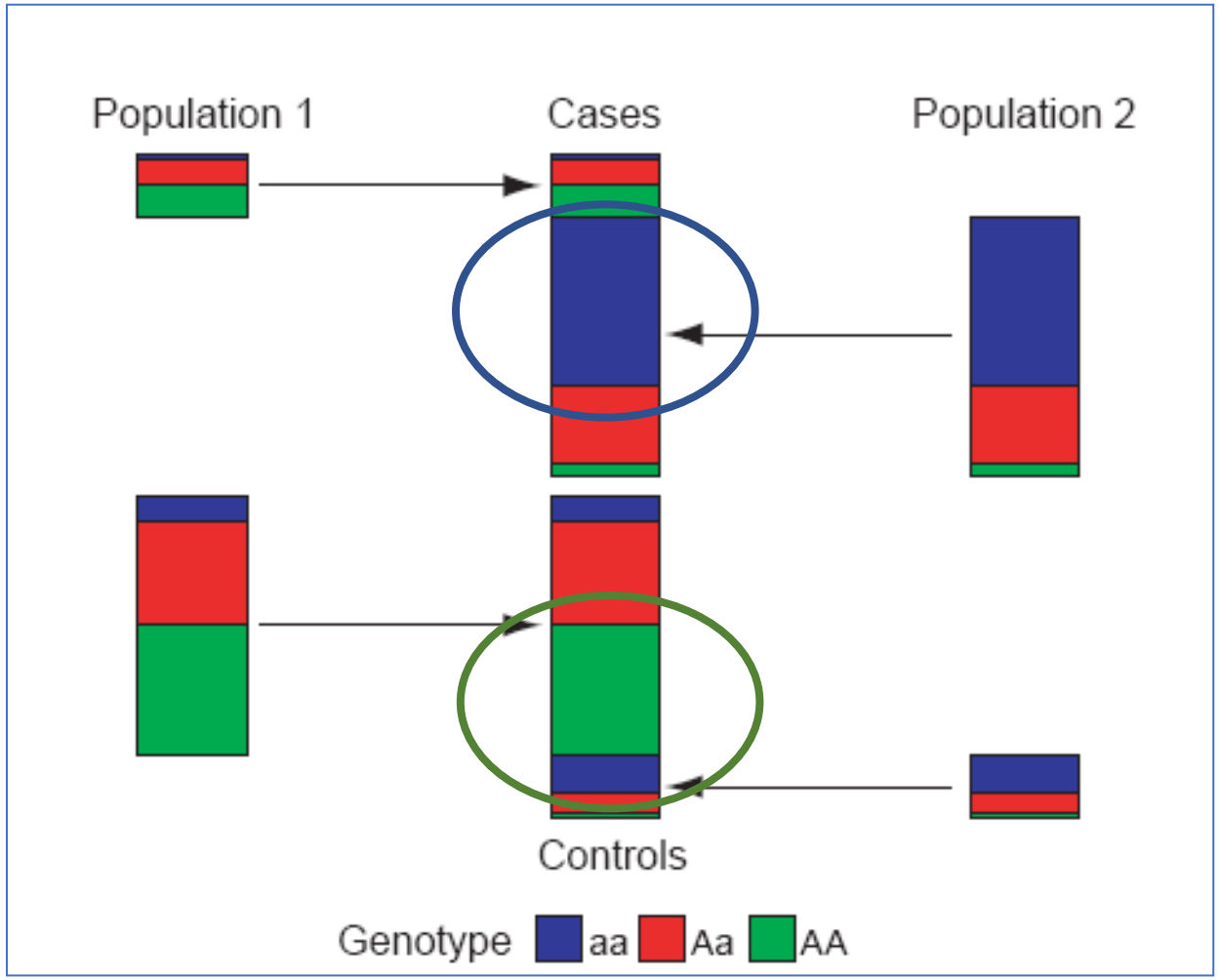
Group differences in ancestry
AND outcome



Marchini, Cardon et al. 2004; Price, Patterson et al. 2006

Population Stratification - Confounding by ancestry

Group differences in ancestry
AND outcome



Marchini, Cardon et al. 2004; Price, Patterson et al. 2006

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?

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?

This is the extreme case, what about more subtle differences?

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

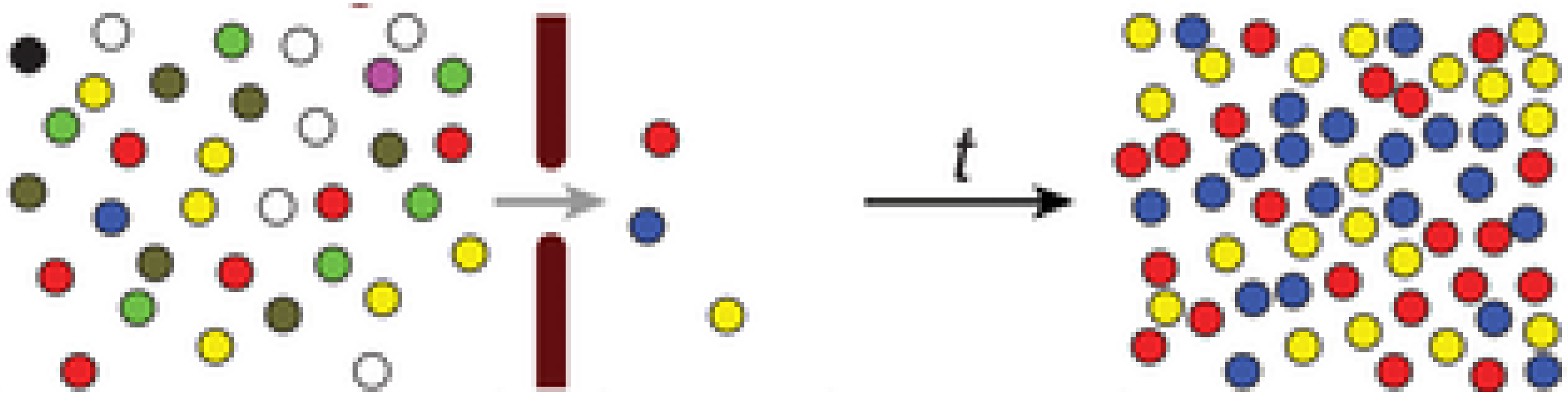
But these are very obviously different populations…

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.

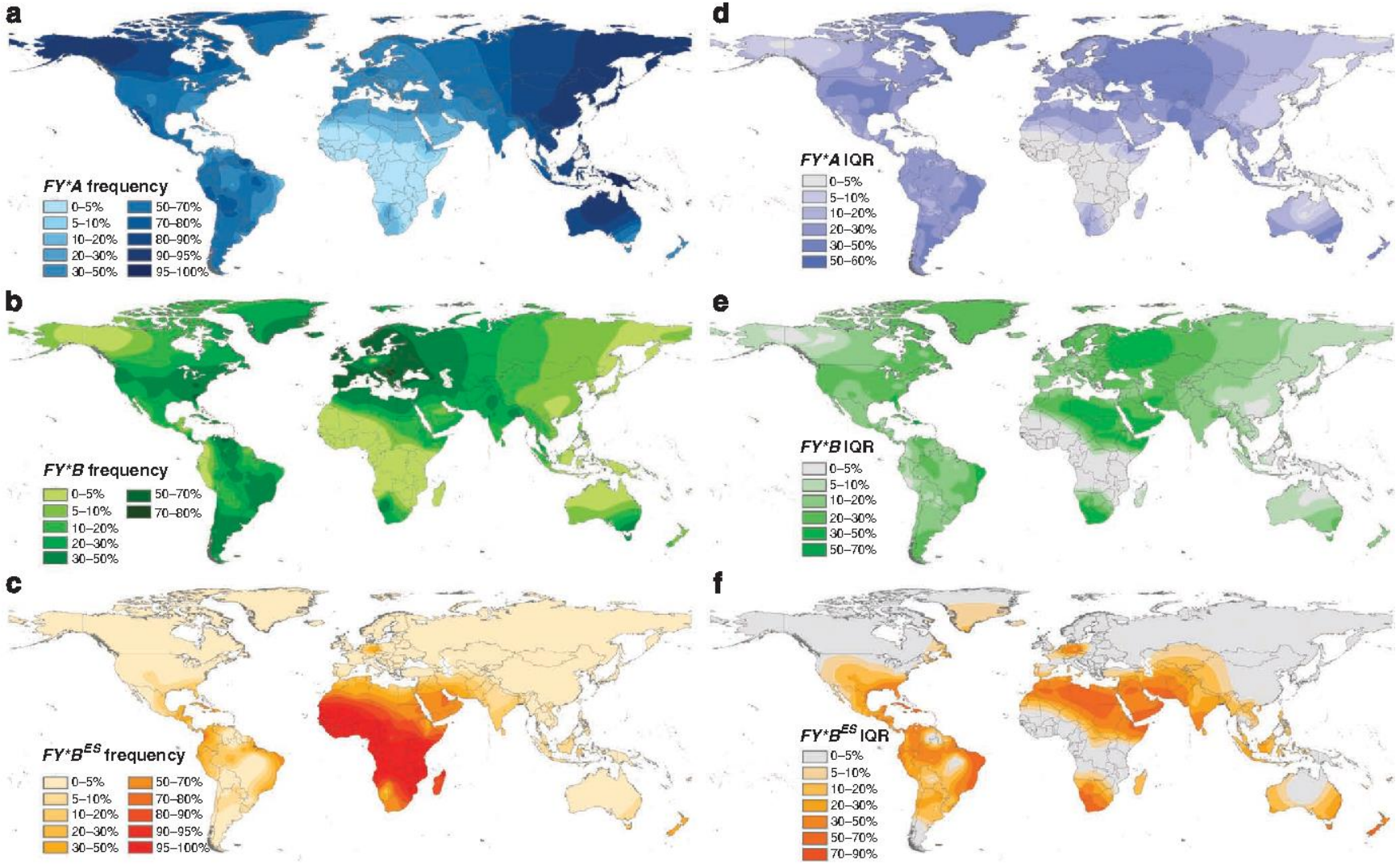
deMenocal & Stringer, Nature 2016

Slight changes in allele frequencies with every population migration/expansion



Abel, PLoS Pathogenics 2015

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

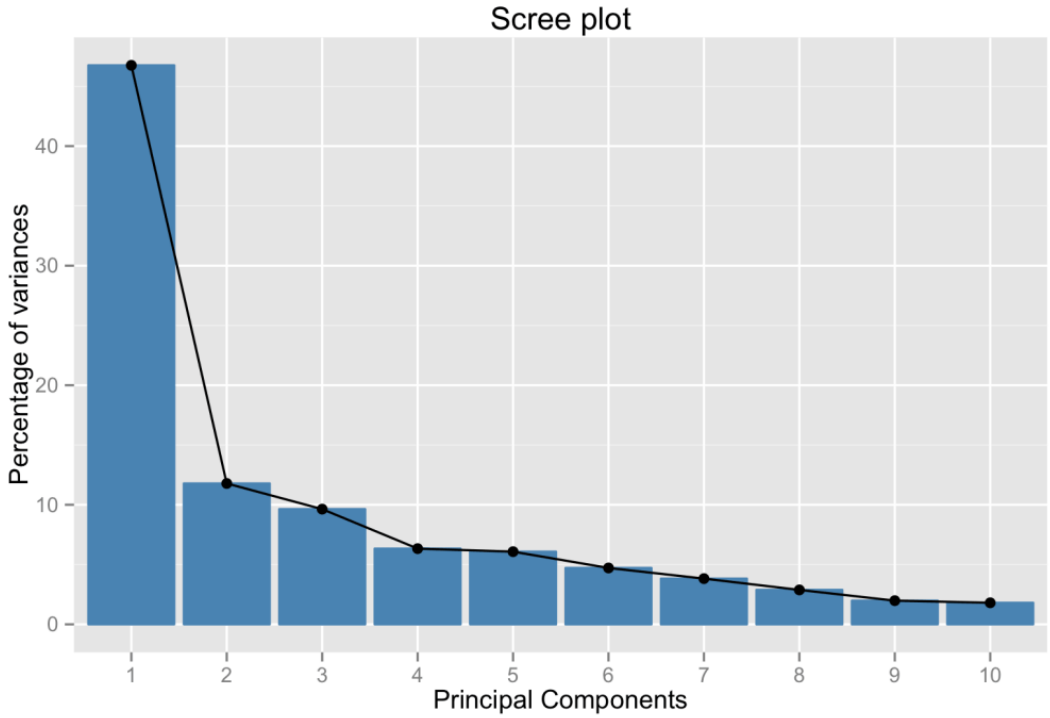


We can use all of these clinal patterns together to “adjust” away the background population substructure effects.

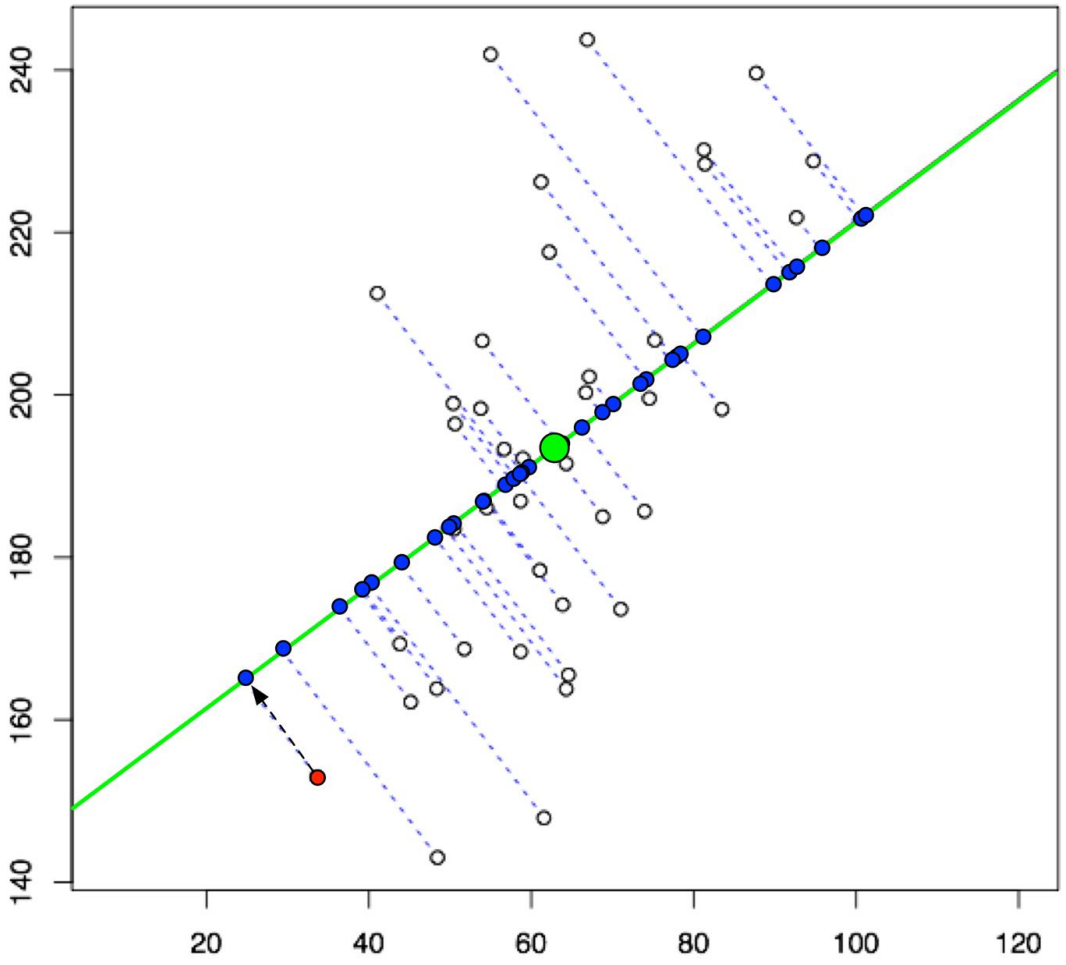
= Principal Component Analysis

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?

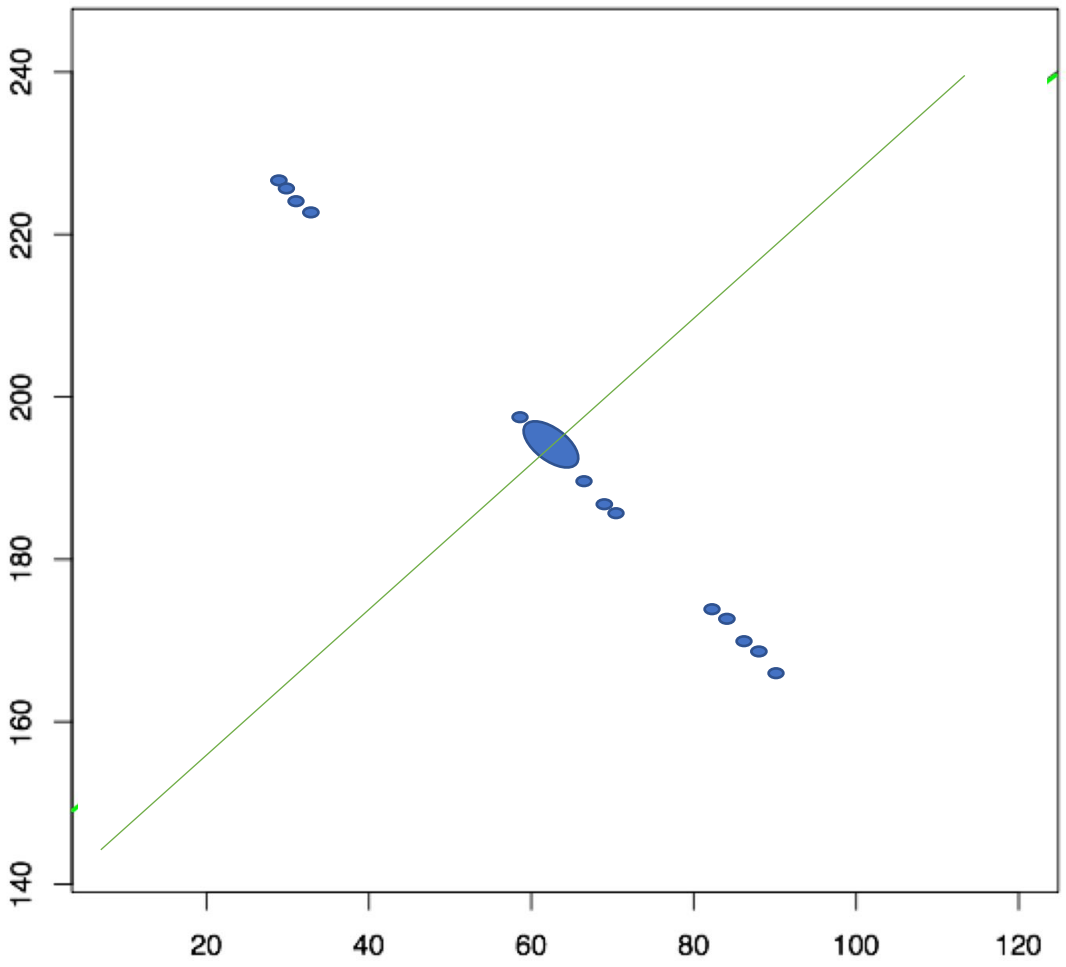


Each PCA maximizes variance and minimizes error

Basically to “absorb any systematic differences.”

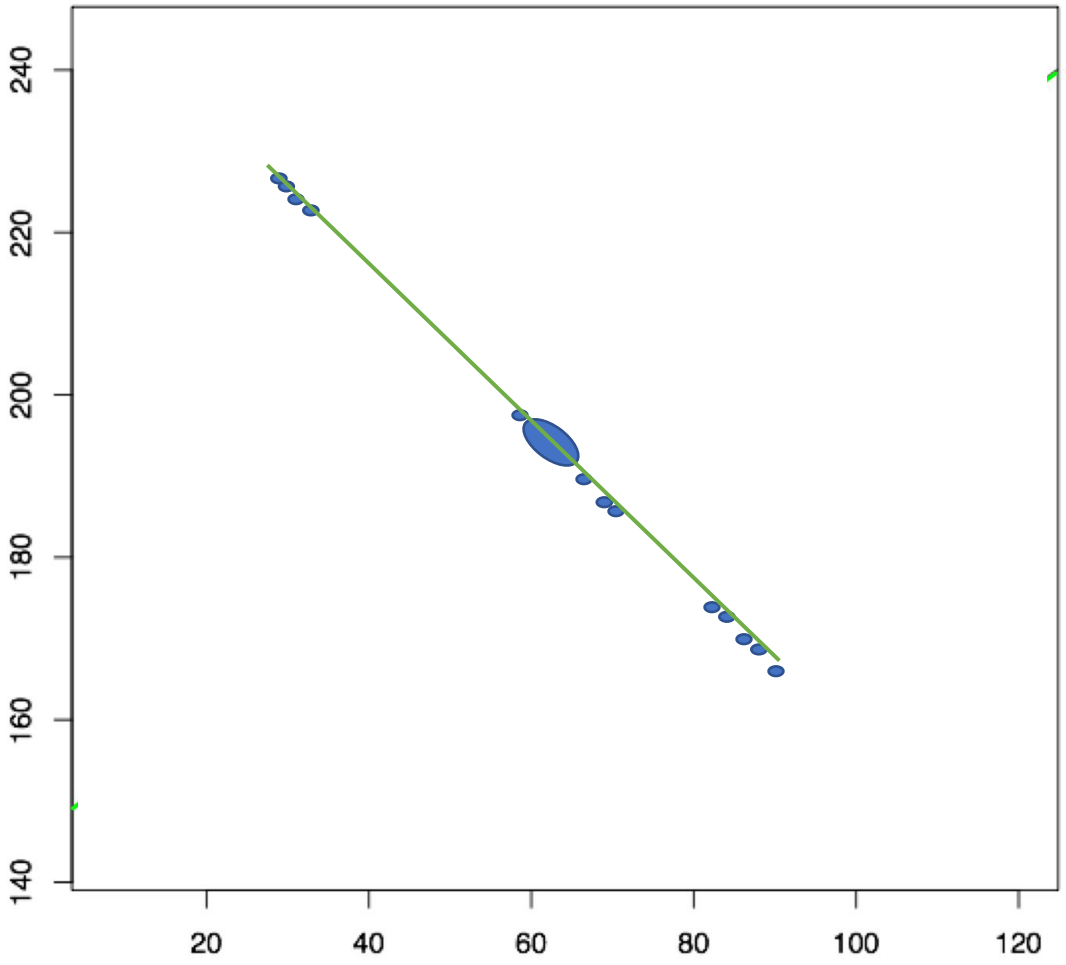
Reduces data dimensions.

Second PCA “soaks up” leftover variance



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

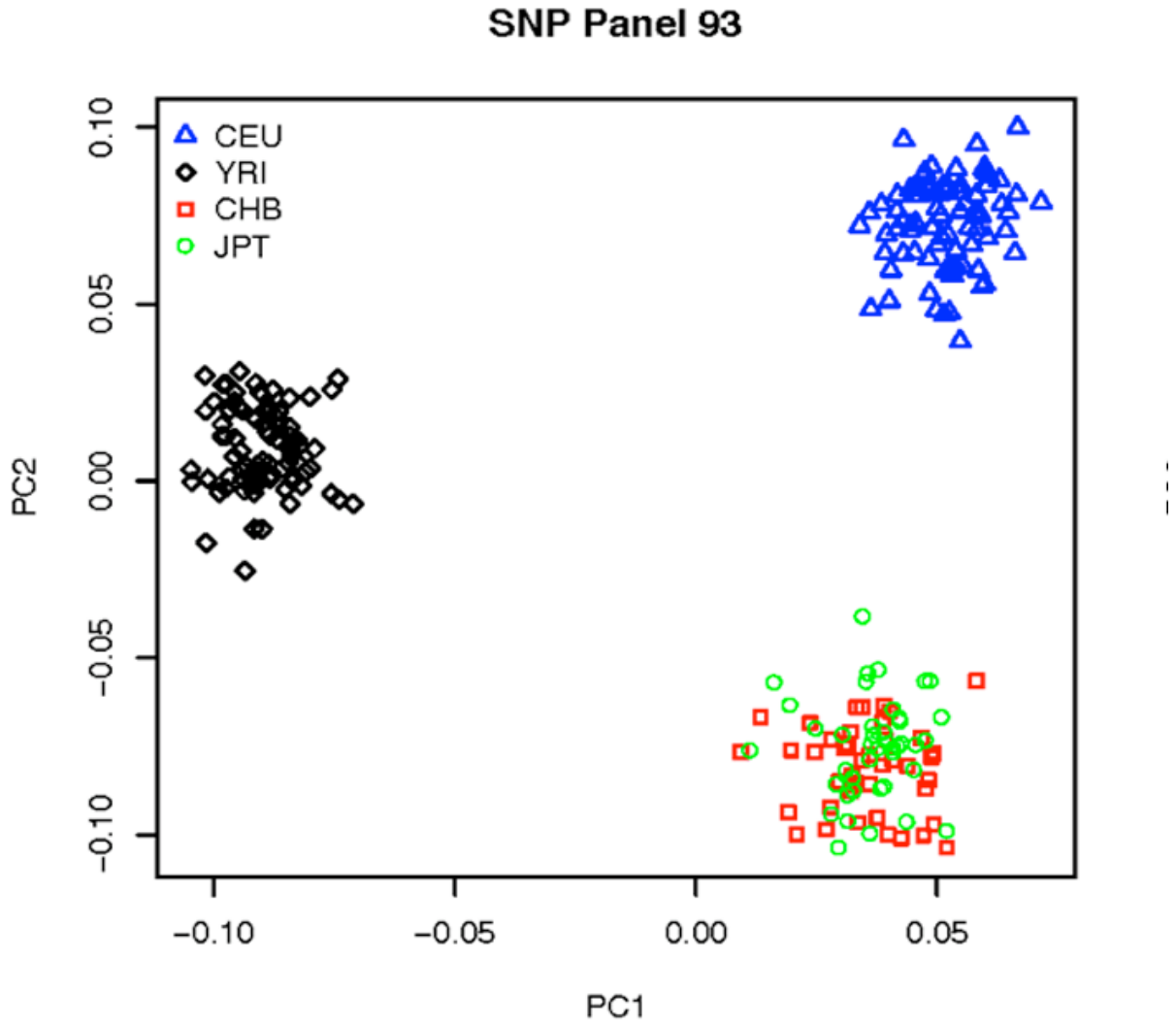
Second PCA “soaks up” leftover variance



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



Hou, PLoS One 2011

Zoom breakout

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

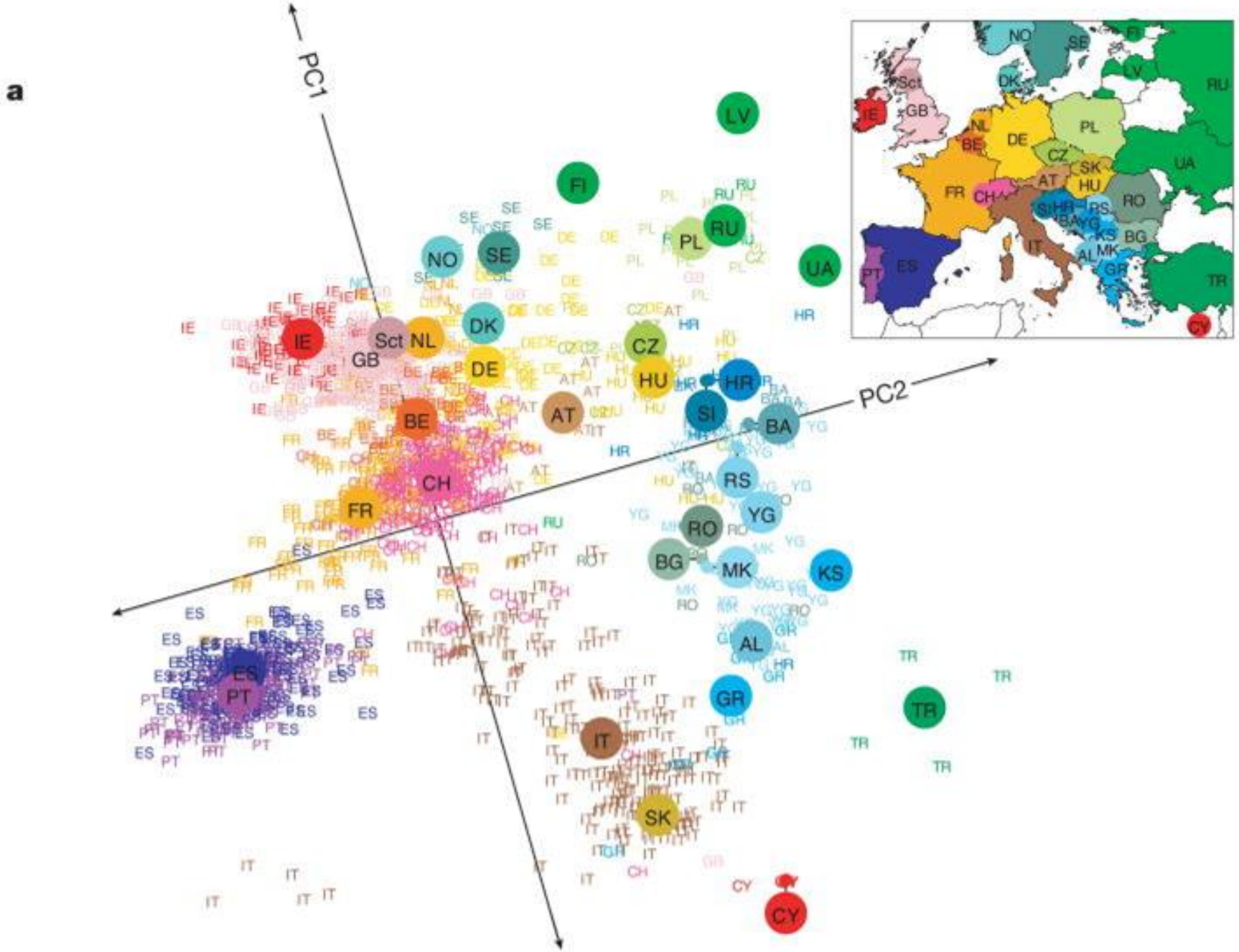
- Translate each genotype into 0, 1, 2 (Additive form), with homozygous reference allele = 0, heterozygous = 1, homozygous variant = 2.
- Multiply that value * loading value for every SNP location
- Add all of those values together.

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



Novembre, 2008 Nature

Include top PCs in genetic association study

phenotype = m*genotype + aPC1 +bPC2 + cPC3 + dPC4 + ePC5

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

Summary

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