THURSDAY (all times in PT)			
8:30 – 9:15	Alie	HWE/LD	Hardy Weinberg Equilibrium, and Linkage Disequilibrium
		BREAK	
9:30-10:15	Alie	Population Structure	Ancestry and Principal Component Analysis
		BREAK	
10:30-11:15	Sara	Study Designs	Types of genetic epidemiology studies, imputation
		LUNCH BREAK	
11:45 – 12:30	Alie	Association studies	Conducting association studies and calculating odds ratios
		BREAK	
12:45-1:30	Sara	GWAS	Genome wide association studies
		BREAK	
1:45 – 2:30	Sara/Alie	Office Hours	Stop by to ask questions from the day, or schedule time to discuss your own project.

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

SCHOOL OF PUBLIC HEALTH
EPIDEMIOLOGY
UNIVERSITY of WASHINGTON

SISG 2022:
Module 11
Genetic Epidemiology

Population Structure

Podcast on the Human Genome Project

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

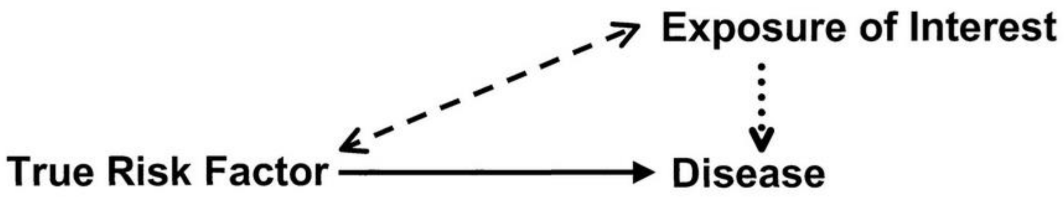
Learning Objectives

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

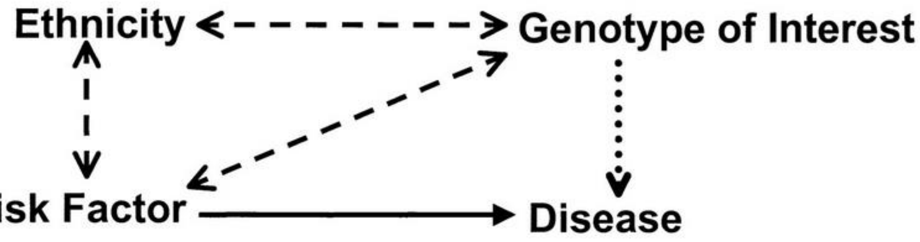
Population Substructure

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

Confounding



Population Stratification

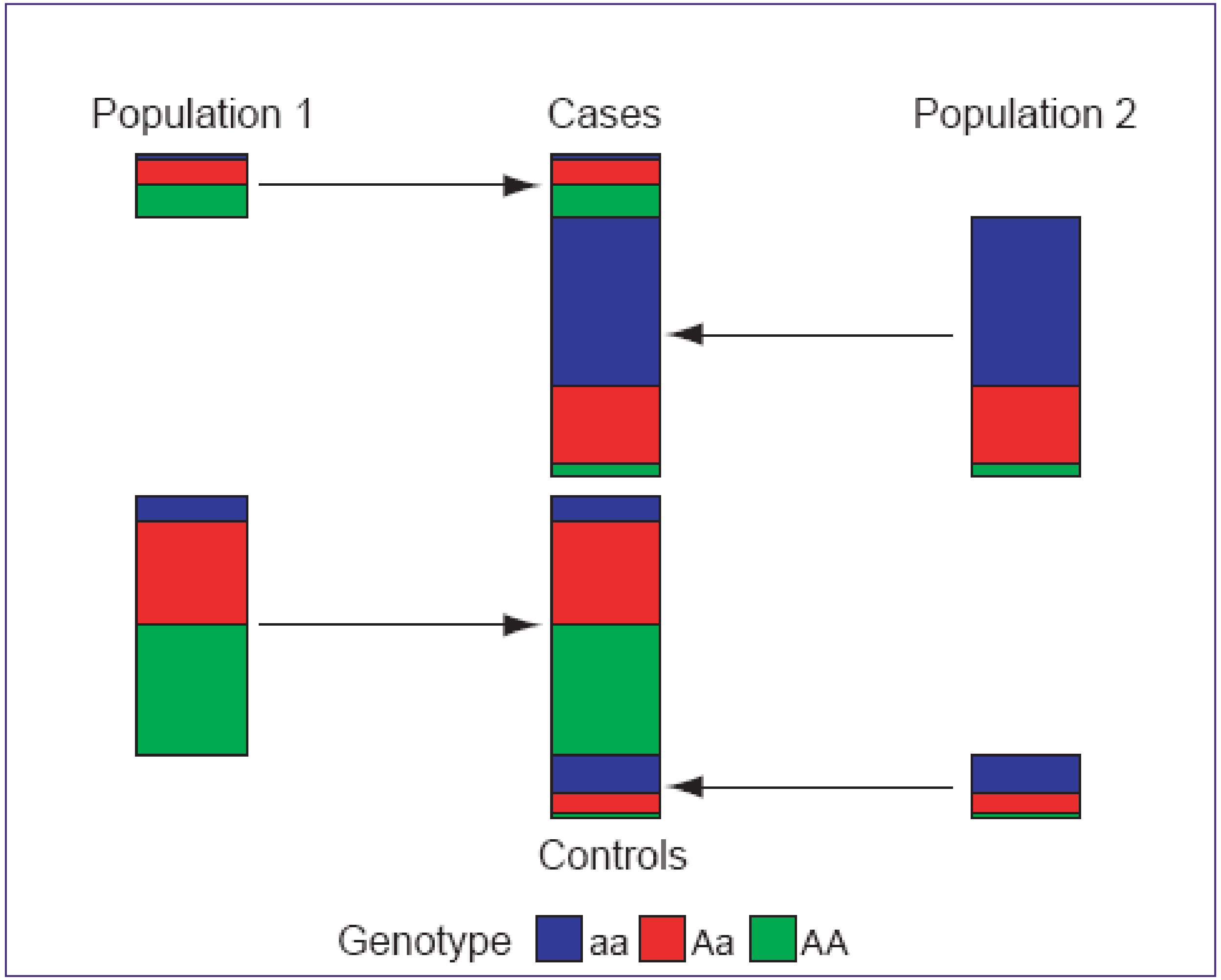


Assume we conduct a case-control GWAS…

- 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

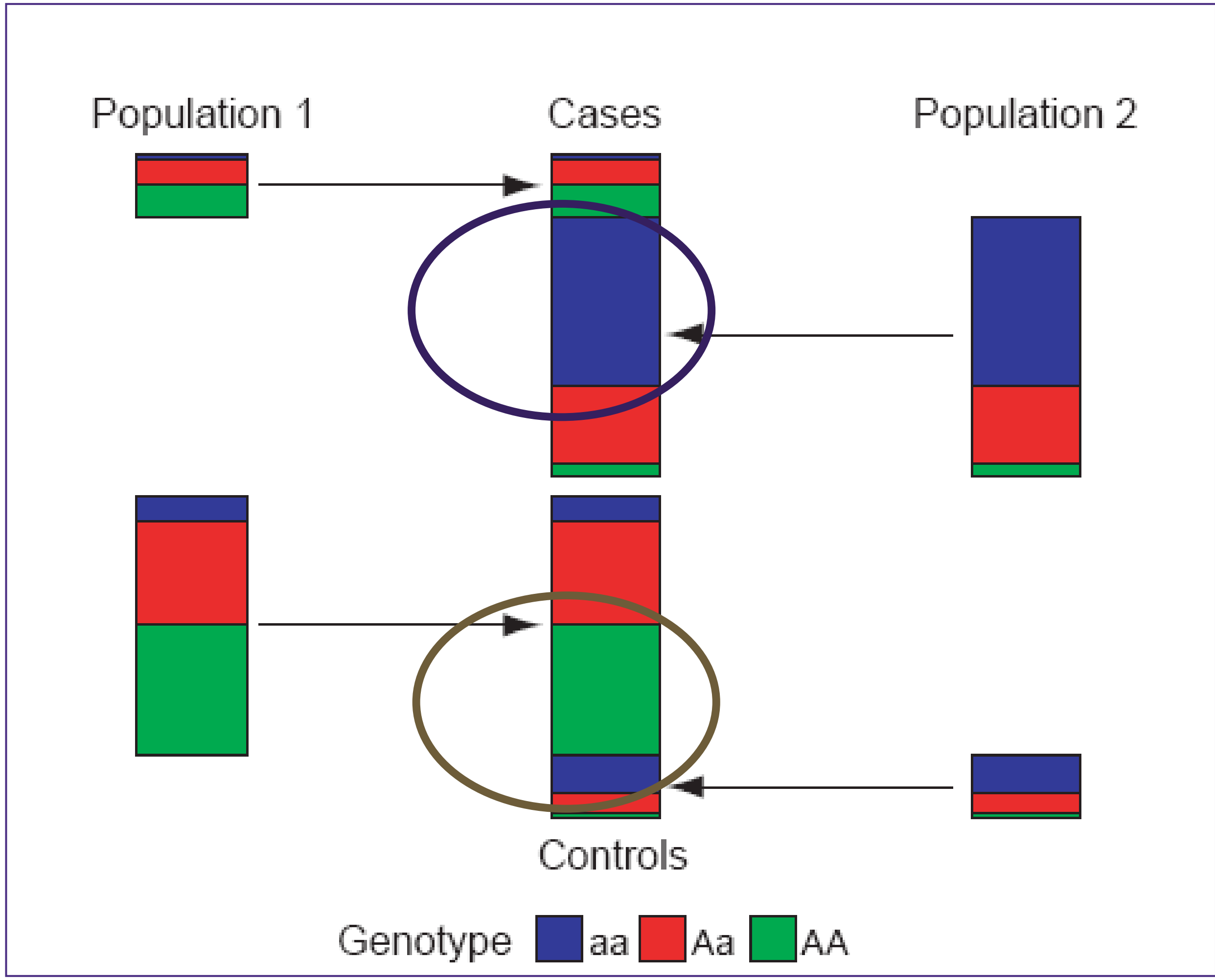
Group differences in ancestry AND outcome



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

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.

Definitions

What is Race?

A social construct that artificially divides people into distinct groups based on characteristics such as physical appearance (particularly skin color), cultural affiliation, cultural history, ethnic classification, and the social, economic and political needs of a society at a given period of time. There are no distinctive genetic characteristics that truly distinguish between groups of people.

What is Ethnicity?

A social construct that divides people into smaller social groups based on characteristics such as shared sense of group membership, cultural heritage, values, behavioral patterns, language, political and economic interests, history and ancestral geographical base.

University of Washington Department of Epidemiology Equity, Diversity, and Inclusion Committee. Glossary of Equity, Diversity, and Inclusion Terms. 2019

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Ancestry

Genealogical ancestry

Genealogical ancestry describes information about your ancestors from whom you are biologically descended. If one of your ancestors belonged to a particular group X, you might say that you have some "X" ancestry. For example, if one of your four grandparents belonged to X you might describe yourself as "one fourth X".

Genetic ancestry

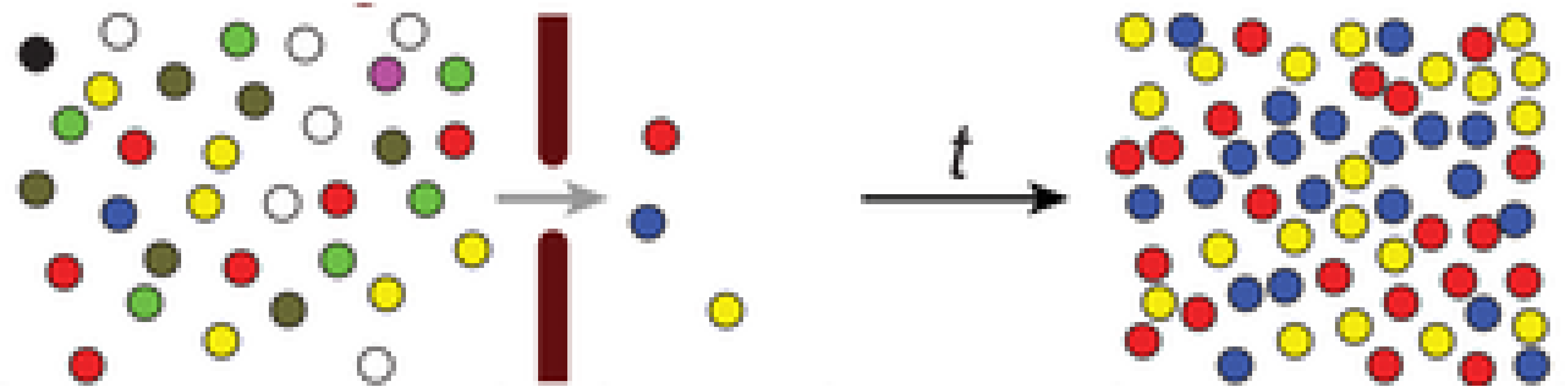
The subset of paths by which the material in your genome has been inherited. Because parents transmit only half their DNA to offspring each generation, your genetic ancestry involves only a small proportion of all your genealogical ancestors.

The difference between ***genealogical*** and ***genetic*** ancestry can be illustrated by full siblings. Full siblings have identical genealogical ancestry but differ in their genetic ancestry, due to differences in transmission of chromosomal segments from their parents.

Mathieson I and Scally A. PLoS Genet, 2020

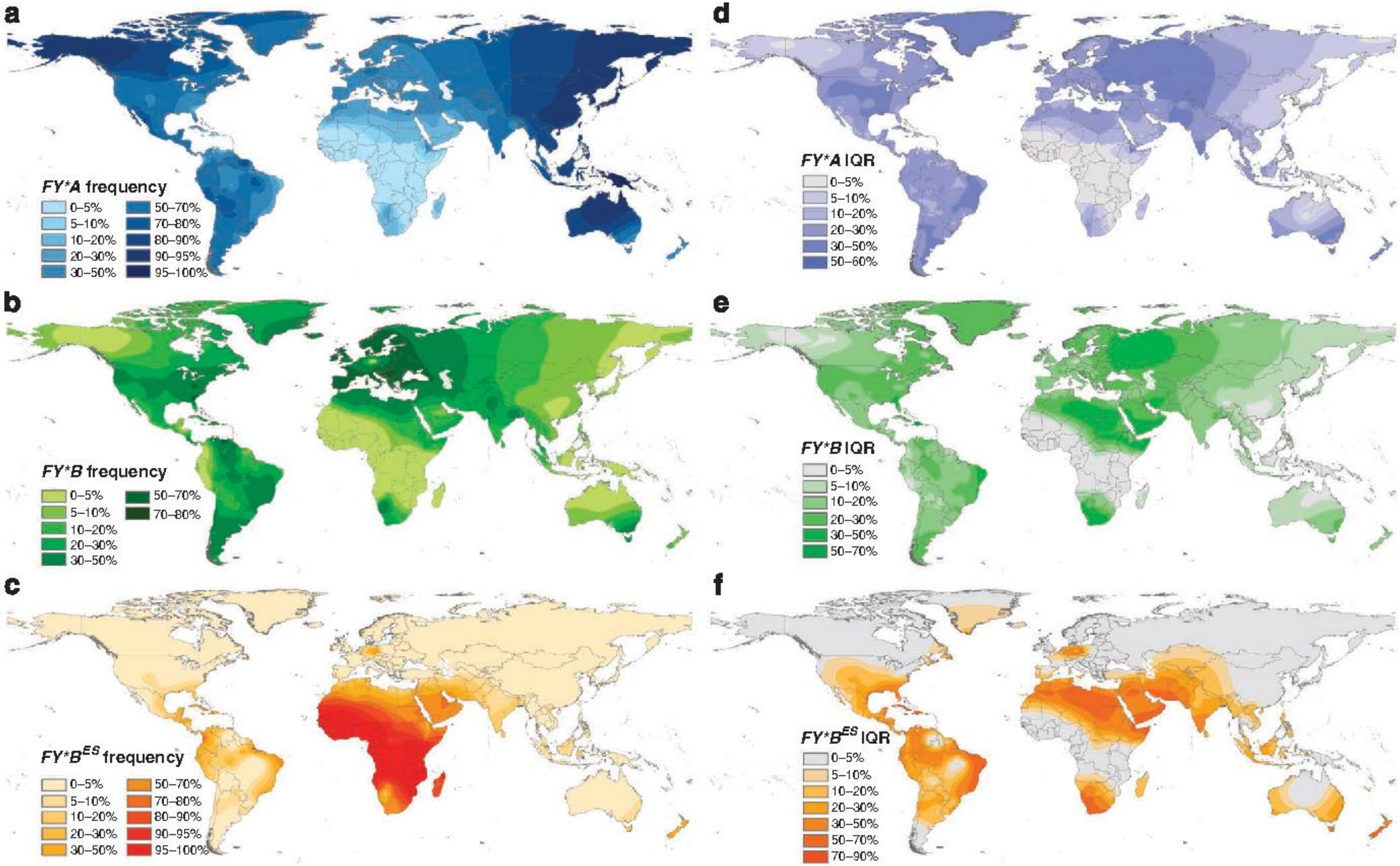
deMenocal & Stringer, Nature 2016

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…

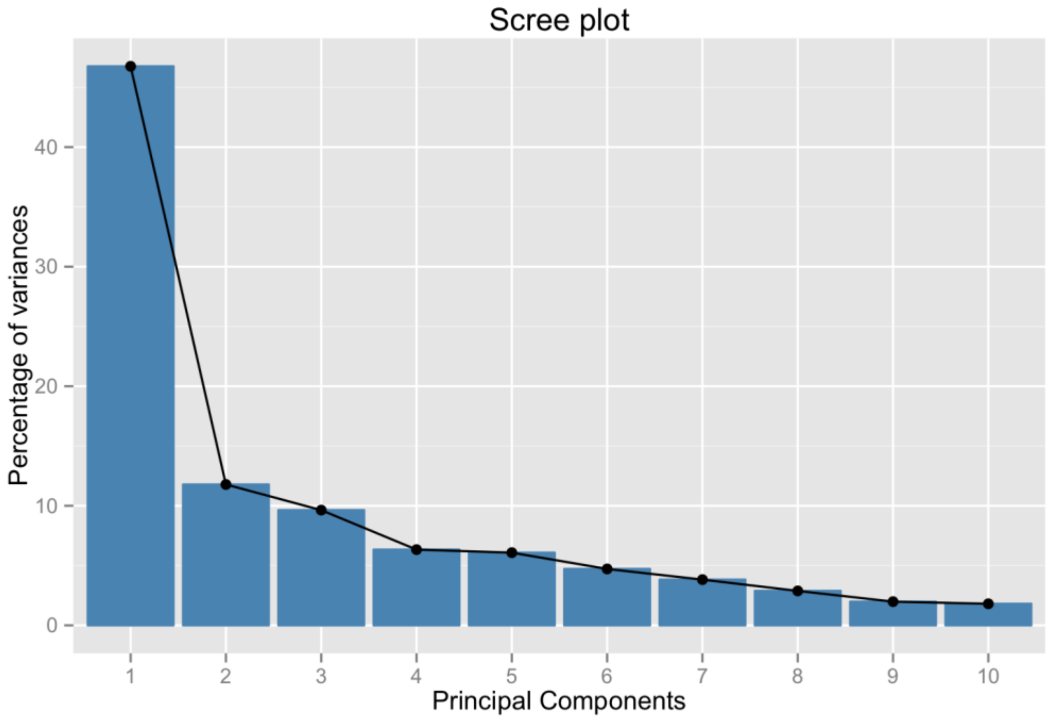


We can use *all* of these clinal patterns together to “adjust” away the background population substructure effects.

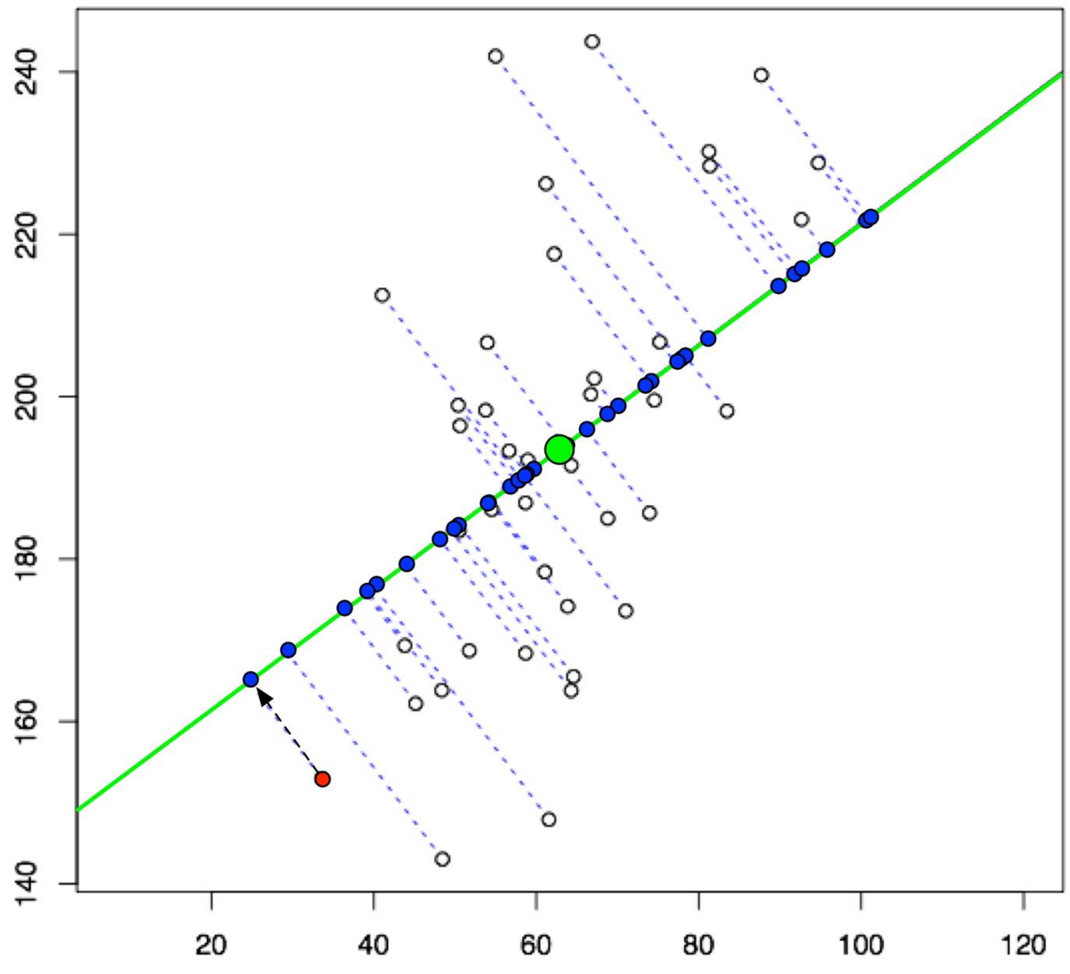
=> Principal Component Analysis

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?

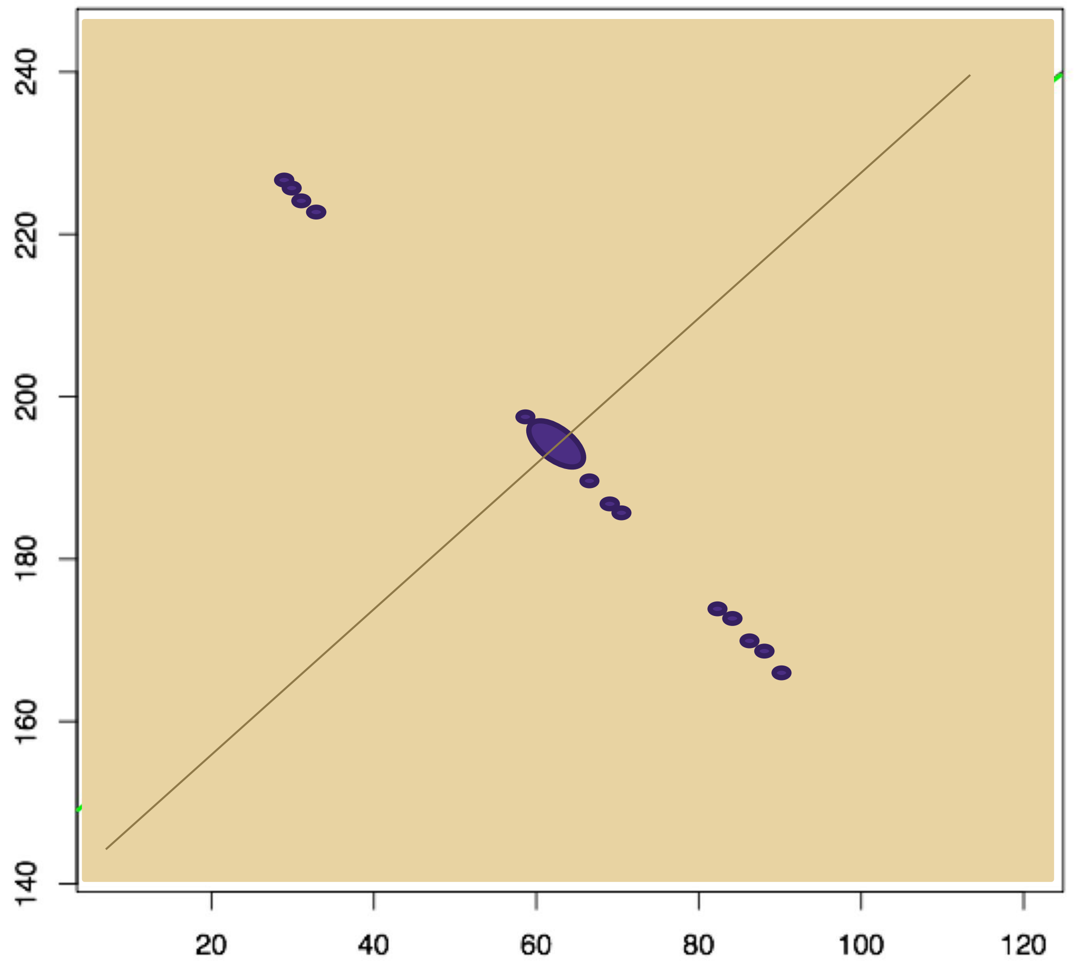


Each PCA maximizes variance and minimizes error

Basically to "absorb any any systematic differences."

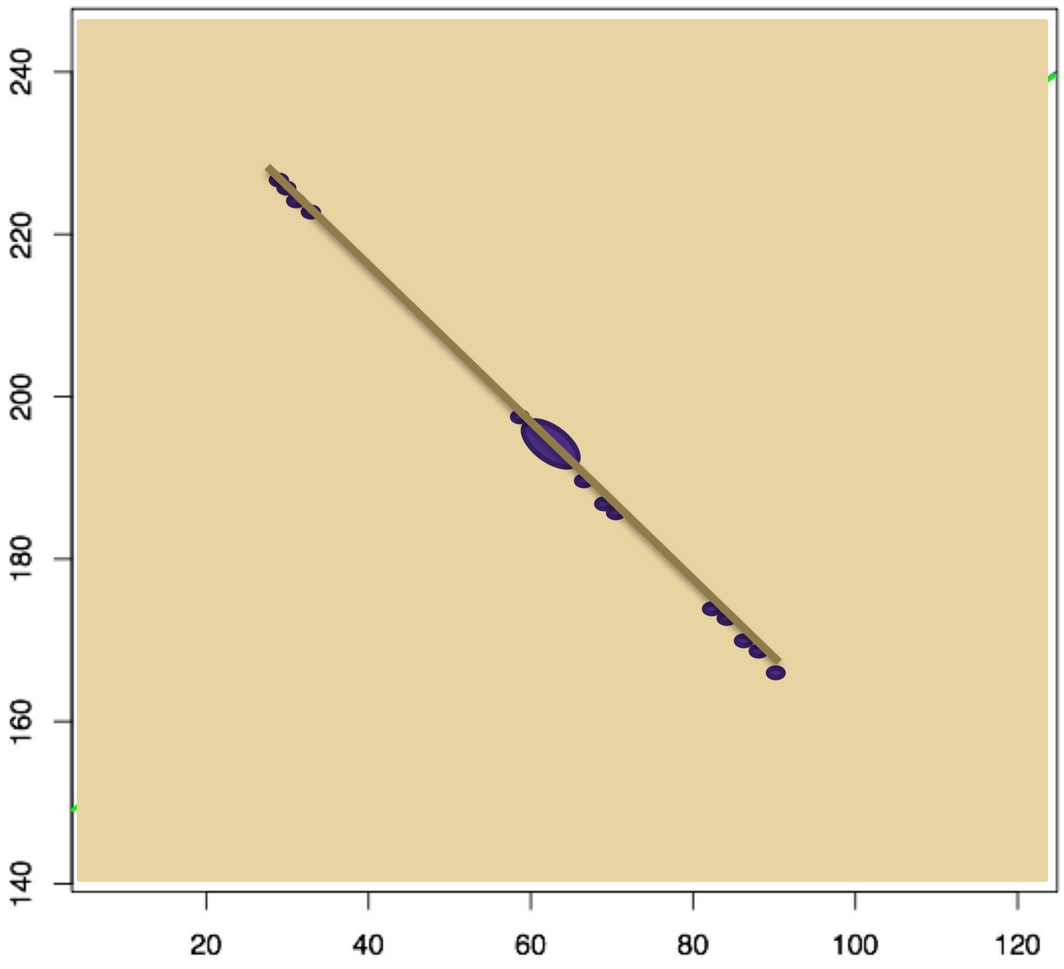
Reduces data dimensions.

Second PCA “soaks up” leftover variance



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

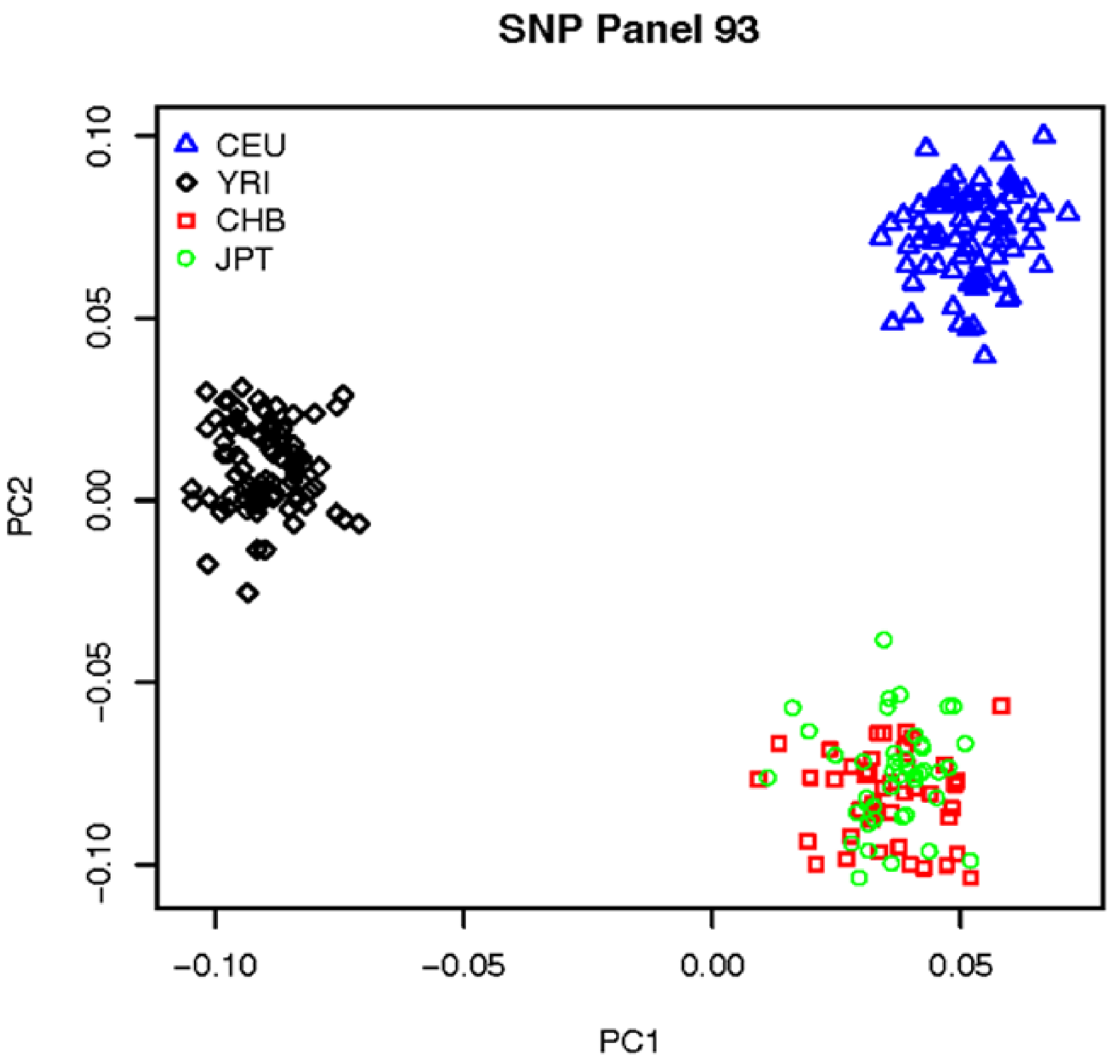
Second PCA “soaks up” leftover variance



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

PC1


PC2
CEU
YRI
JPT

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

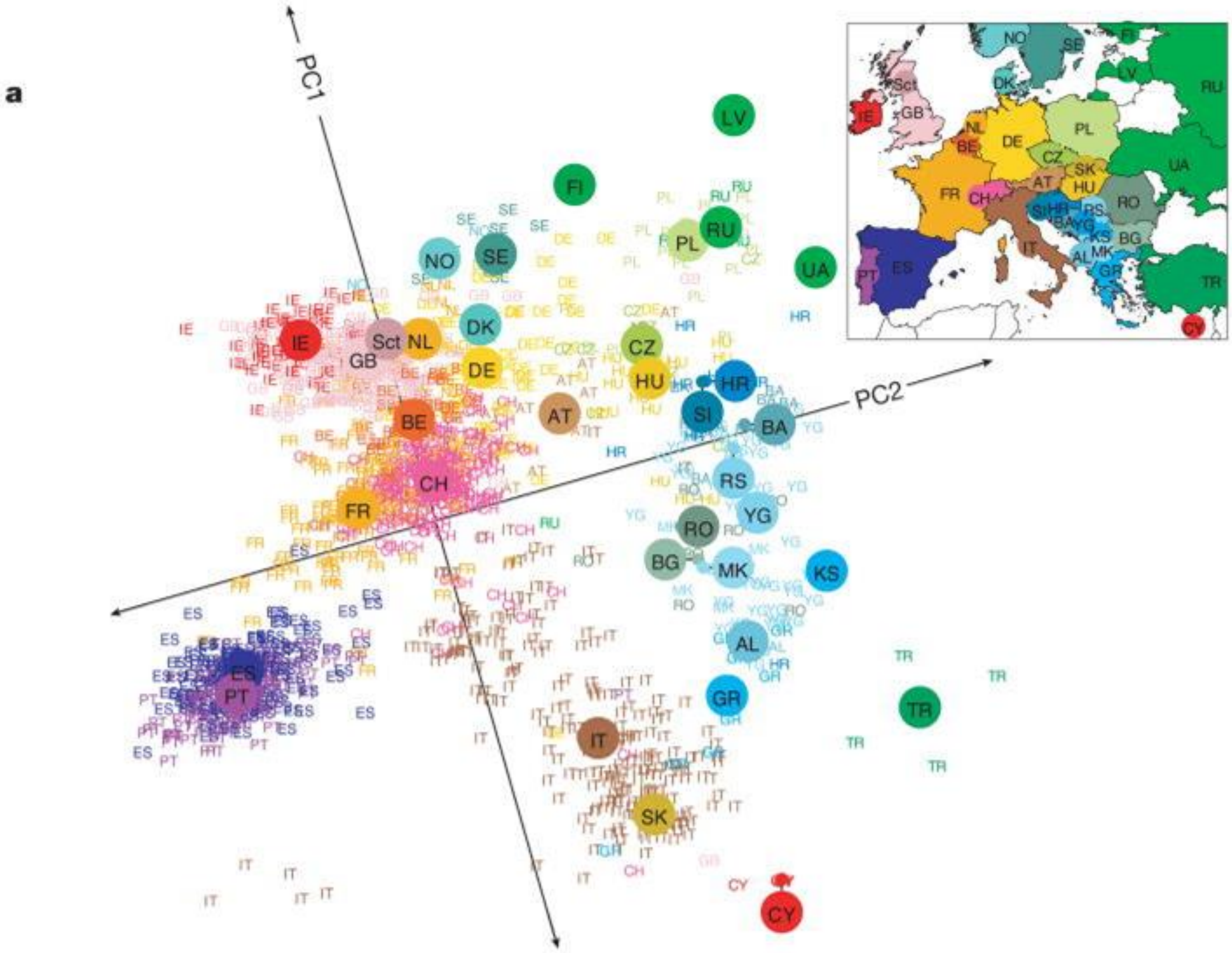
- Translate each genotype into 0, 1, 2 (Additive form), with homozygous reference allele = 0, heterozygous = 1, homozygous variant = 2.
- Multiply that value * loading value for every SNP location
- Add all of those values together.

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

Include top PCs in genetic association study

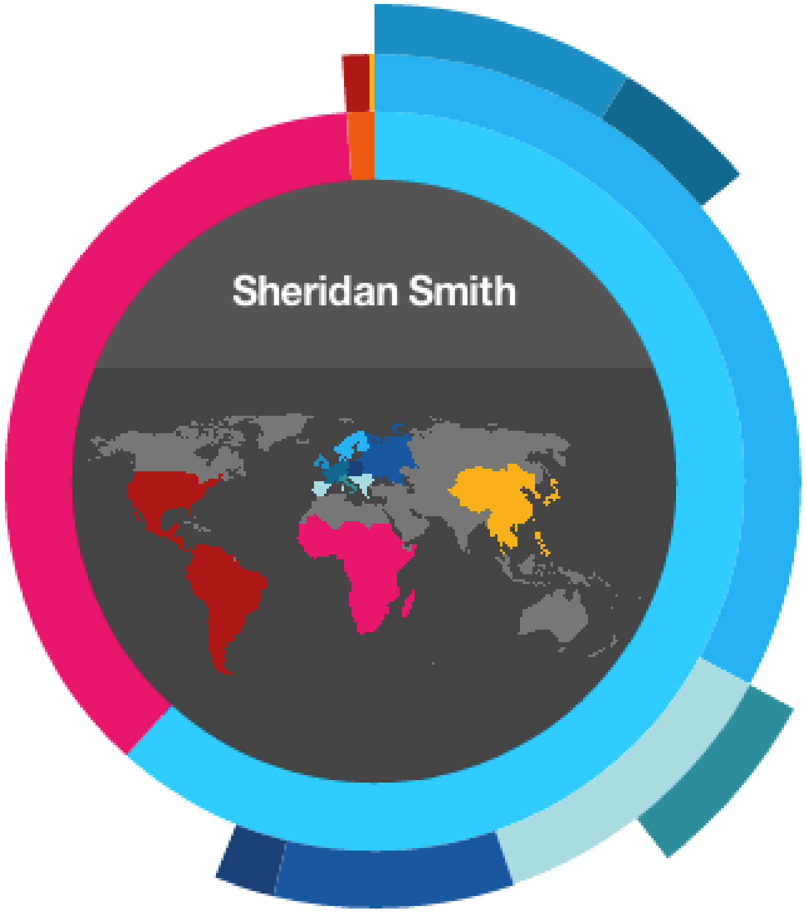
phenotype = m*genotype + aPC1 +bPC2 + cPC3 + dPC4 + ePC5

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



Novembre, 2008 Nature

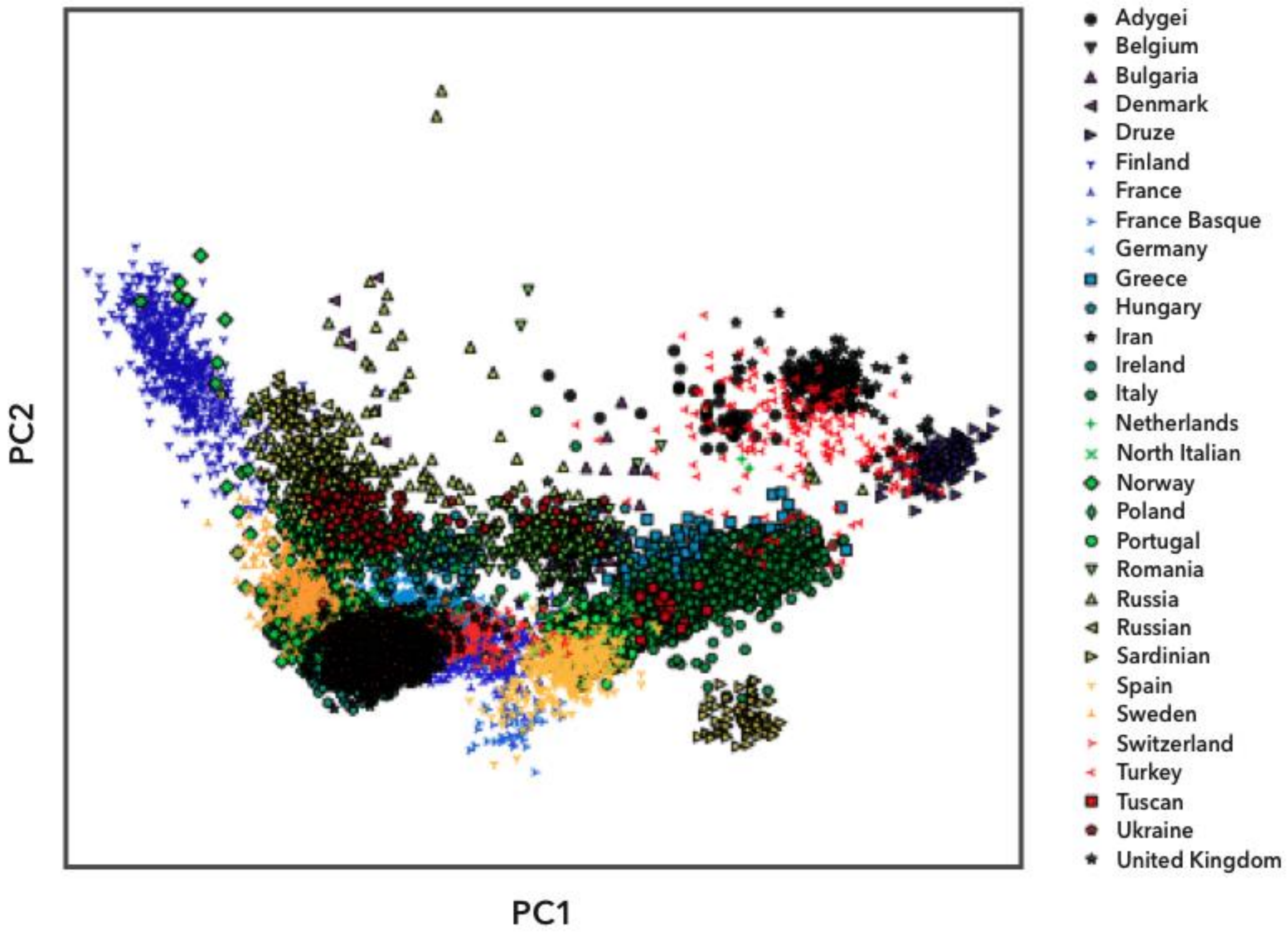
Map View | − Sub-regional Resolution +



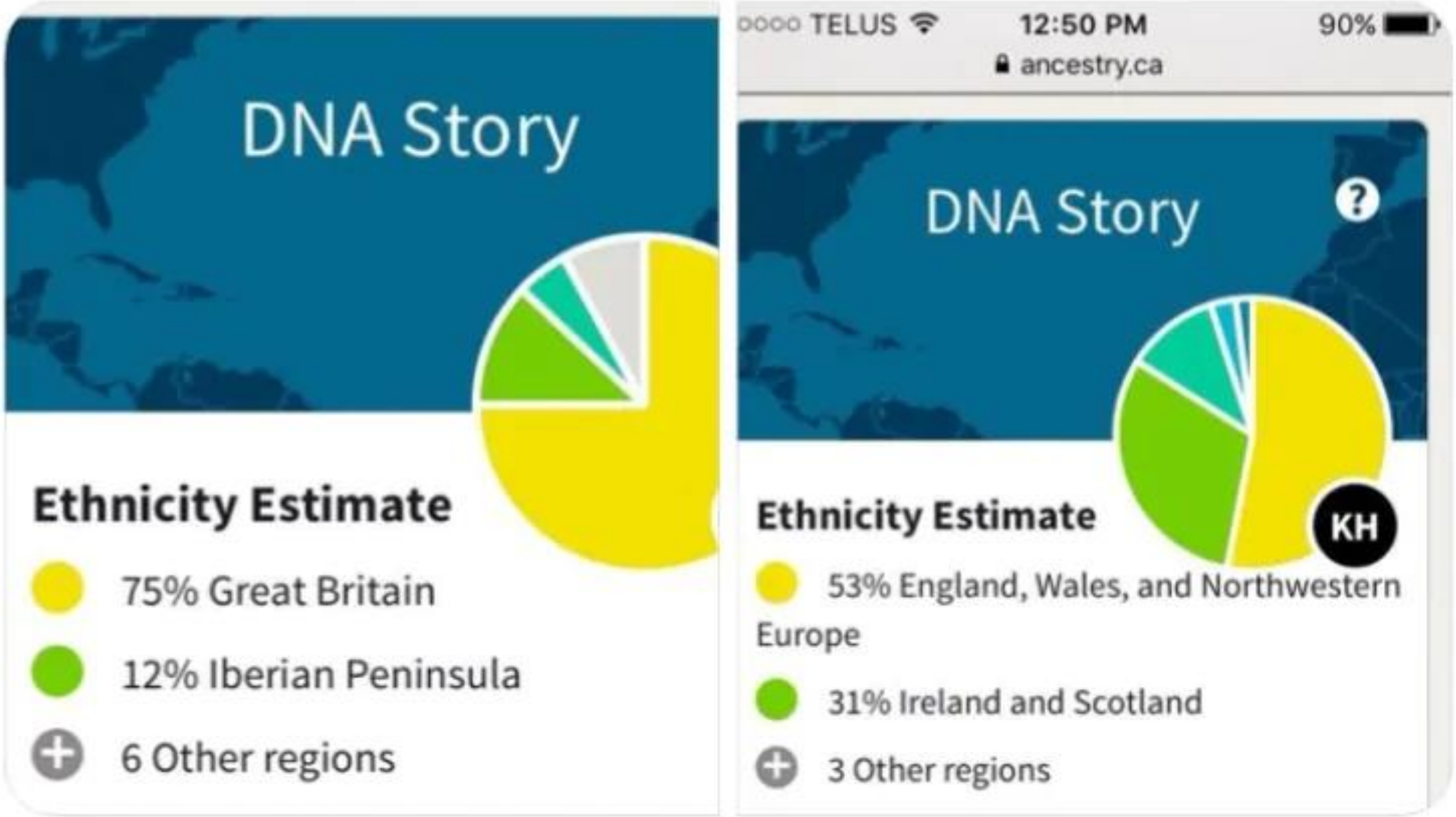
Ancestry Composition tells you what percent of your DNA comes from each of 22 populations worldwide. The analysis includes DNA you received from all of your ancestors, on both sides of your family. The results reflect where your ancestors lived 500 years ago, before ocean-crossing ships and airplanes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

61.7%	European
	Northern European
8.9%	British and Irish
5.0%	French and German
19.0%	Nonspecific Northern Eur…
	Southern European
6.6%	Italian
5.2%	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**

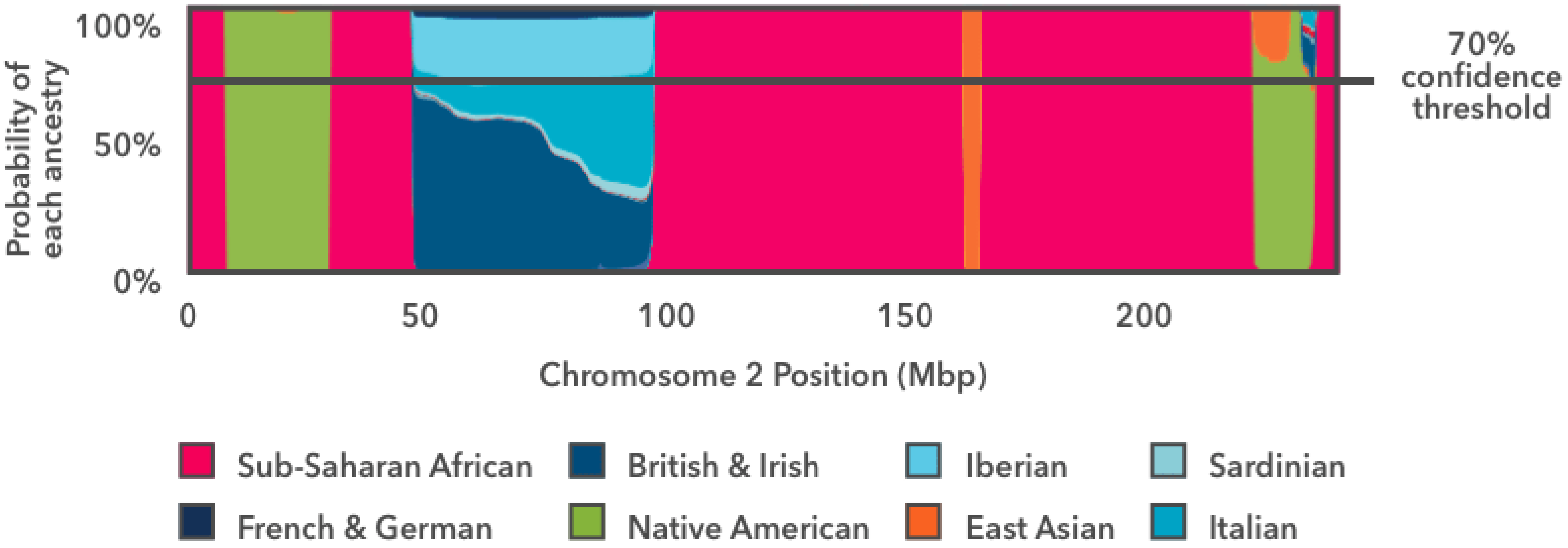
PCA matching at 300bp genomic windows - 23andMe



Sep 22, 2018


DNA Story
Ethnicity Estimate
75% Great Britain
12% Iberian Peninsula
6 Other regions
TELUS
12:50 PM
90%
ancestry.ca
DNA Story
Ethnicity Estimate
KH
53% England, Wales, and Northwestern Europe
31% Ireland and Scotland
3 Other regions

DNA segment and ancestry probability



Summary

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