SISG 2023 - Module 2

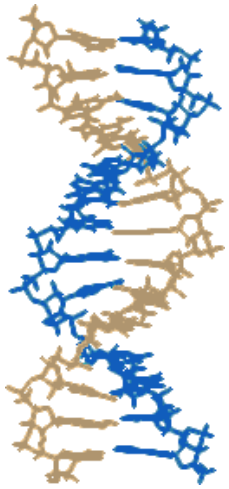
Introduction to Genetics and Genomics

Genetic Ancestry

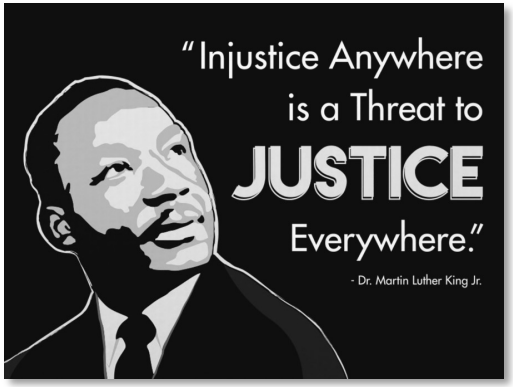
Block #10 – Wednesday, July 12

Joe Lachance and Greg Gibson

joseph.lachance@biology.gatech.edu



Terminology



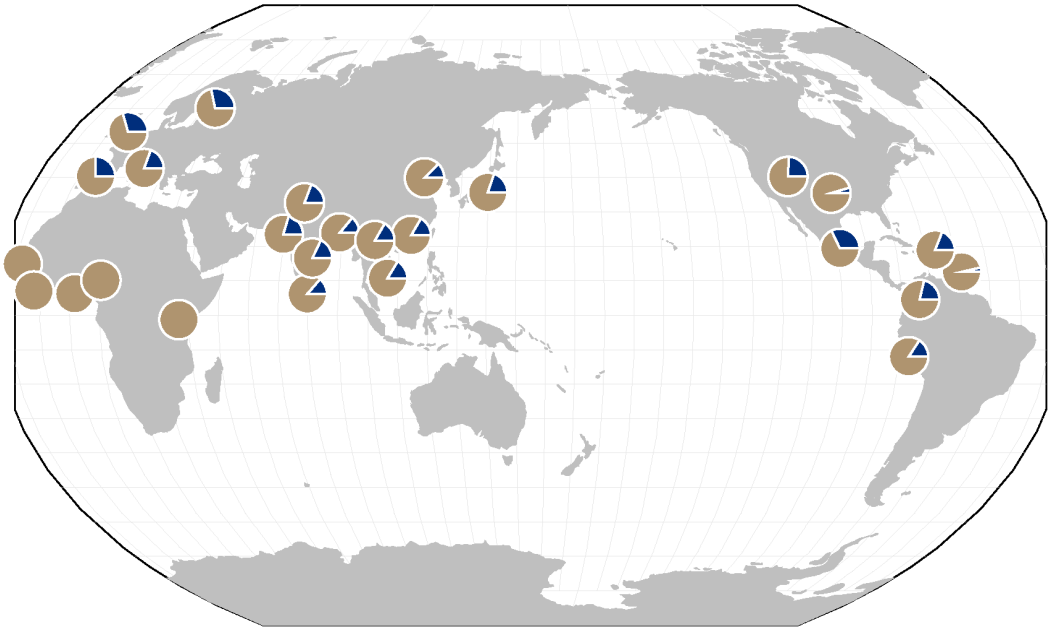
NATIONAL ACADEMIES Sciences Engineering Medicine

- **Race** refers to a socially constructed classification based on perceived biological similarities
- **Ethnicity** refers to a socially constructed classification based on perceived cultural similarities (e.g., language and beliefs)
- **Ancestry** refers to a person's origin or descent, lineage, "roots," or heritage, including kinship (this term focuses on genetics)
- **Populations** are often defined in terms of sampling locations

The changing face of humanity

Photos by Martin Schoeller for National Geographic

AIMs

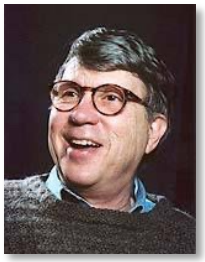


- Ancestry Informative Markers (AIMs) have large allele frequency differences between populations
- Rare alleles are more likely to be population-specific
- No single AIM is a perfect classifier

Image source: GGV

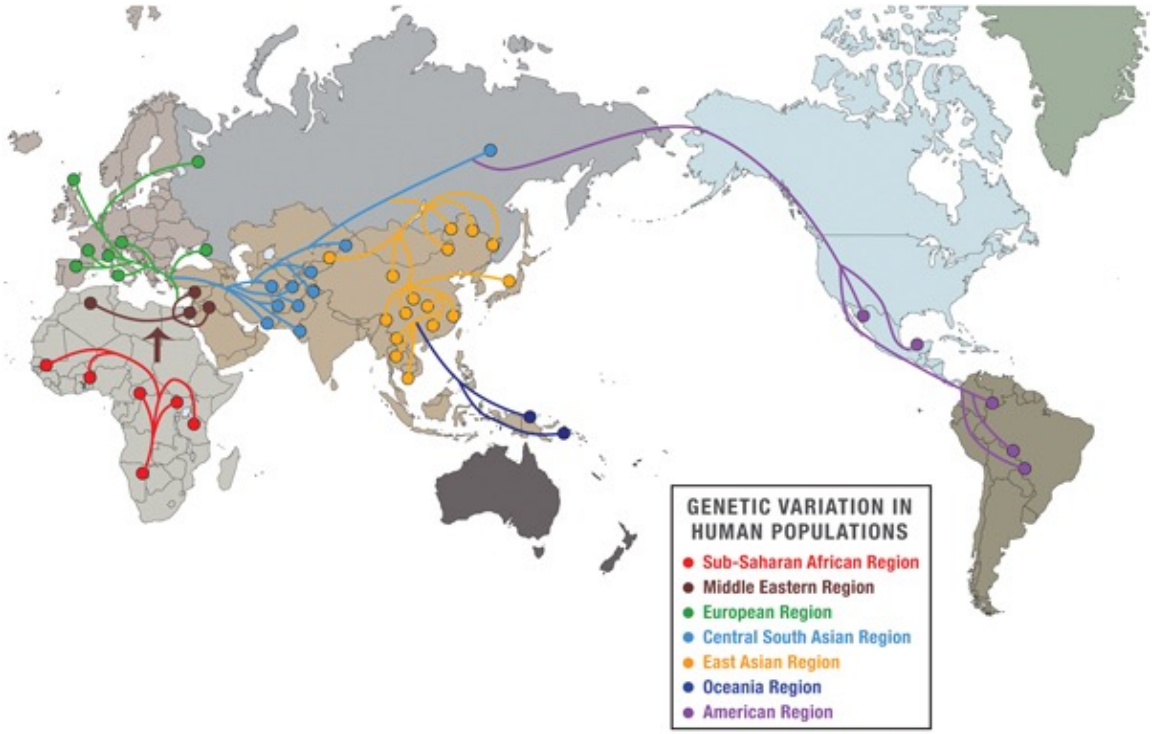
Variance partitioning and Lewontin's Fallacy

- Richard Lewontin (1972)
  - 85% of genetic diversity is found within populations, as opposed to between populations or between continents



- A.W.F. Edwards (2003)
  - Individuals can be assigned to different populations if multilocus data are analyzed ("Lewontin's Fallacy")

HGDP



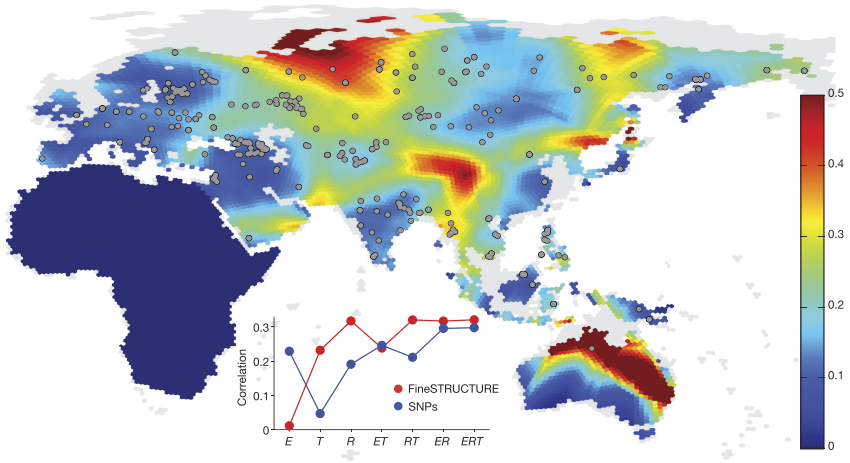
- The Human Genome Diversity Project (HGDP): >50 sampled populations
- Ethical issues:
  - Indigenous groups need not be isolated populations
  - Accusations of “helicopter science”

1000 Genomes Project

- Whole genome sequencing of 2504 samples from 26 global populations

SGDP and EGDP

Simons Genome Diversity Project
Mallick et al. (*Science*, 2106)



Estonian Biocentre Human Genome Diversity Panel
Pagani et al. (*Nature*, 2016)

- More granular sampling, but fewer samples per location

Dangers of limited sampling

- If highly divergent locations are sampled it can lead one to think human diversity falls into distinct categories

- Ideally, each living individual has an equal chance of being sampled in genetic studies

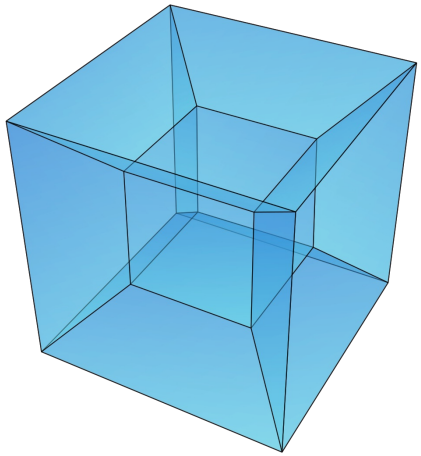
Image: VectorStock

What do population genetic datasets look like?

Chrom	Position	SNP_ID	Ref	Alt	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	Sample_6	Sample_7	Sample_8
6	95632928	rs138026492	C	T	00	00	00	00	00	00	00	00
6	95636167	rs7776290	C	A	11	11	11	11	11	11	11	11
6	95638707	rs9490131	C	T	01	01	11	01	00	00	00	01
6	95639314	rs111993428	G	C	00	00	01	00	00	00	00	00
6	95644518	rs76301071	G	A	00	00	00	00	00	00	00	00
6	95658829	rs9320918	G	T	11	11	11	11	11	11	11	11
6	95676882	rs73546580	G	A	00	00	00	00	01	00	01	00
6	95677999	rs9491308	T	C	01	00	00	00	00	00	00	00
6	95678247	rs117120297	T	C	00	00	00	00	00	00	00	00
6	95689368	rs117996333	G	A	00	00	00	00	00	00	00	00
6	95722603	rs143147841	A	C	00	00	00	00	00	00	00	00
6	95726175	rs116190944	T	C	00	00	00	00	00	00	00	00
6	95747602	rs112599693	G	C	00	00	00	00	00	00	00	00
6	95757249	rs73757480	G	C	00	00	01	00	00	00	00	00
6	95769070	rs62417884	C	T	00	00	00	01	00	01	00	00
6	95788421	rs117816213	T	C	00	00	00	00	00	00	00	00
6	95793344	rs147072022	T	C	00	00	00	01	00	01	00	00
6	95795036	rs77874428	C	A	00	00	00	00	00	00	00	00

- Each row is a different SNP, and each column is a different individual

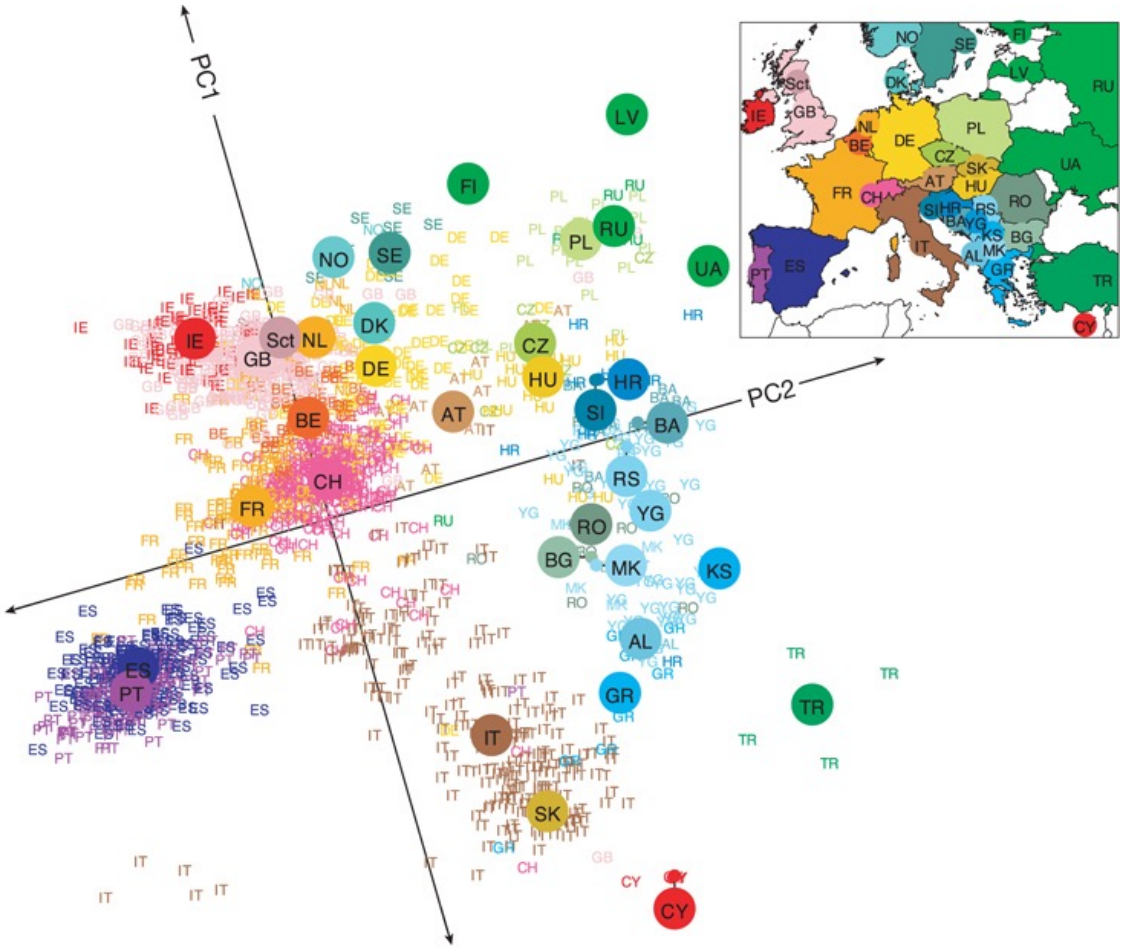
Dimensionality and PCA



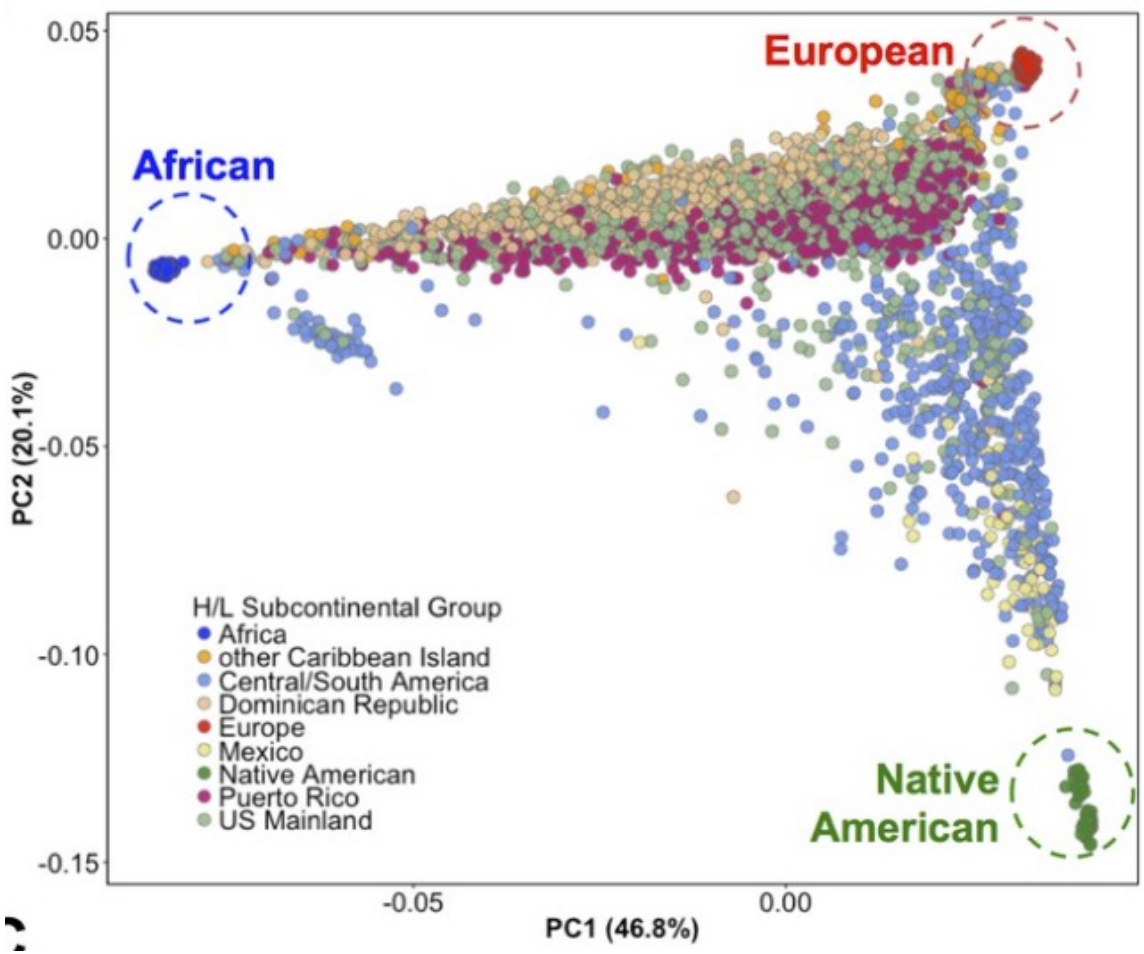
- Principal Component Analysis (PCA) is one way to reduce the dimensionality of genetic datasets
- Each PC refers to an orthogonal (perpendicular) dimension – each PC is an eigenvector and eigenvalues correspond to the % of variance explained by each PC
- PCA can be used to represent samples in a genetic “space” (samples closer together in this space share more alleles)

Hypercube image from Wikipedia

Genes mirror geography in Europe



Human diversity exists along a continuum



Belbin et al. (*eLife*, 2017)

ADMIXTURE plots

- Human variation exists along a continuum
- Every individual's genome contains a mix of different ancestries
- Each genetic ancestry can be represented by a different color

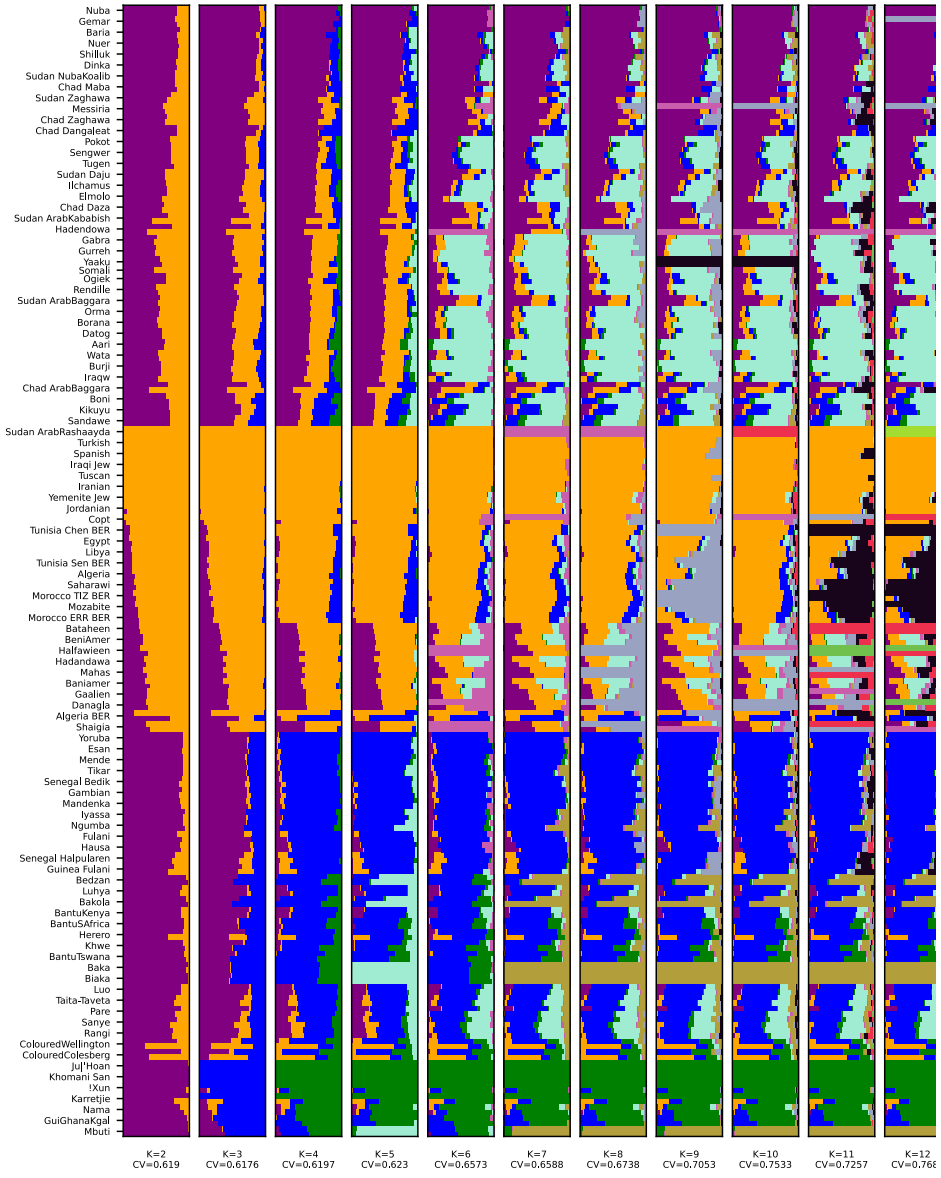
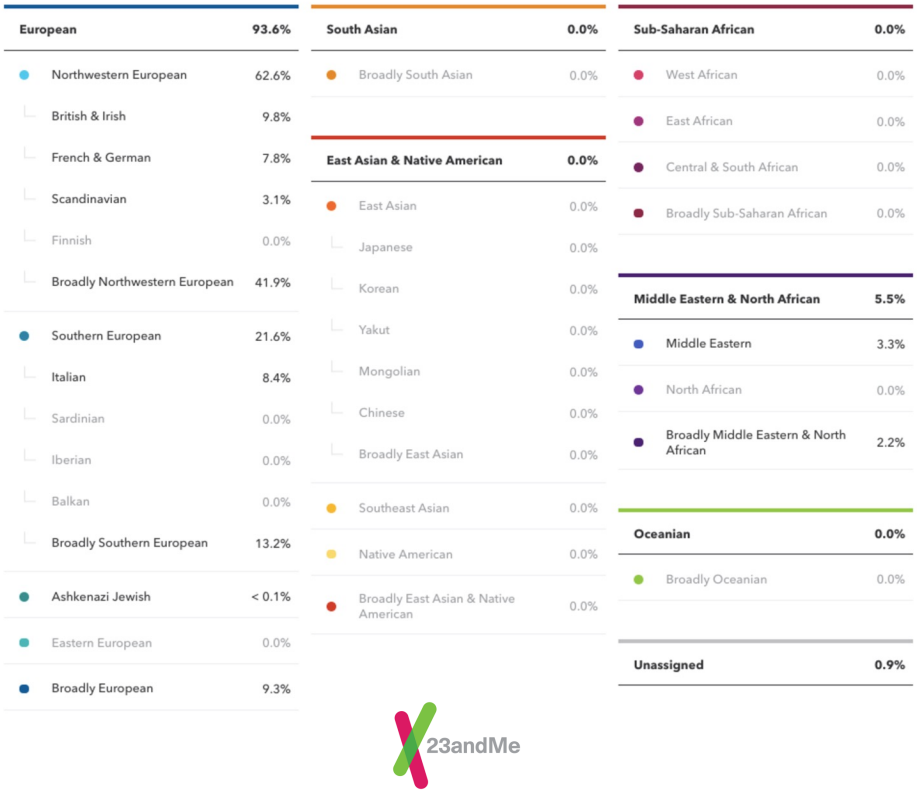
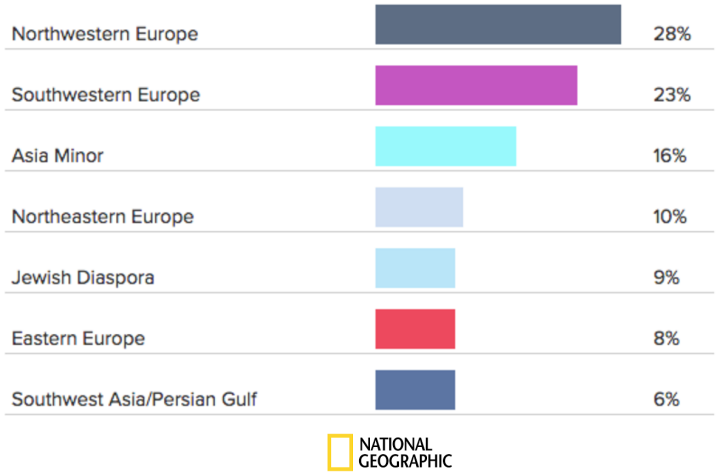
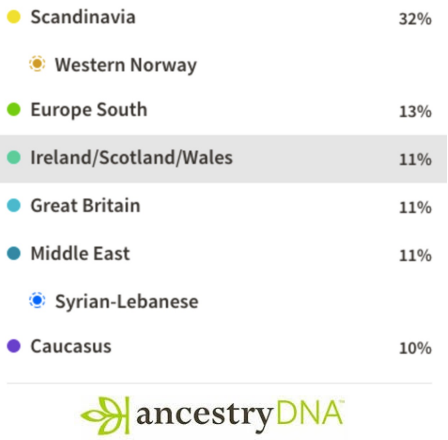


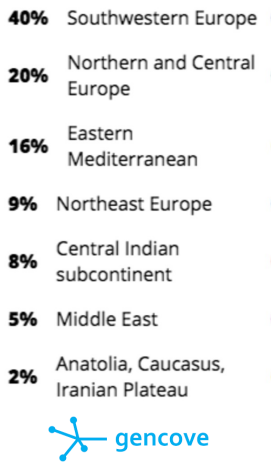
Figure from Pfennig et al. (*GBE*, 2023)

Ancestry inference and DTC testing







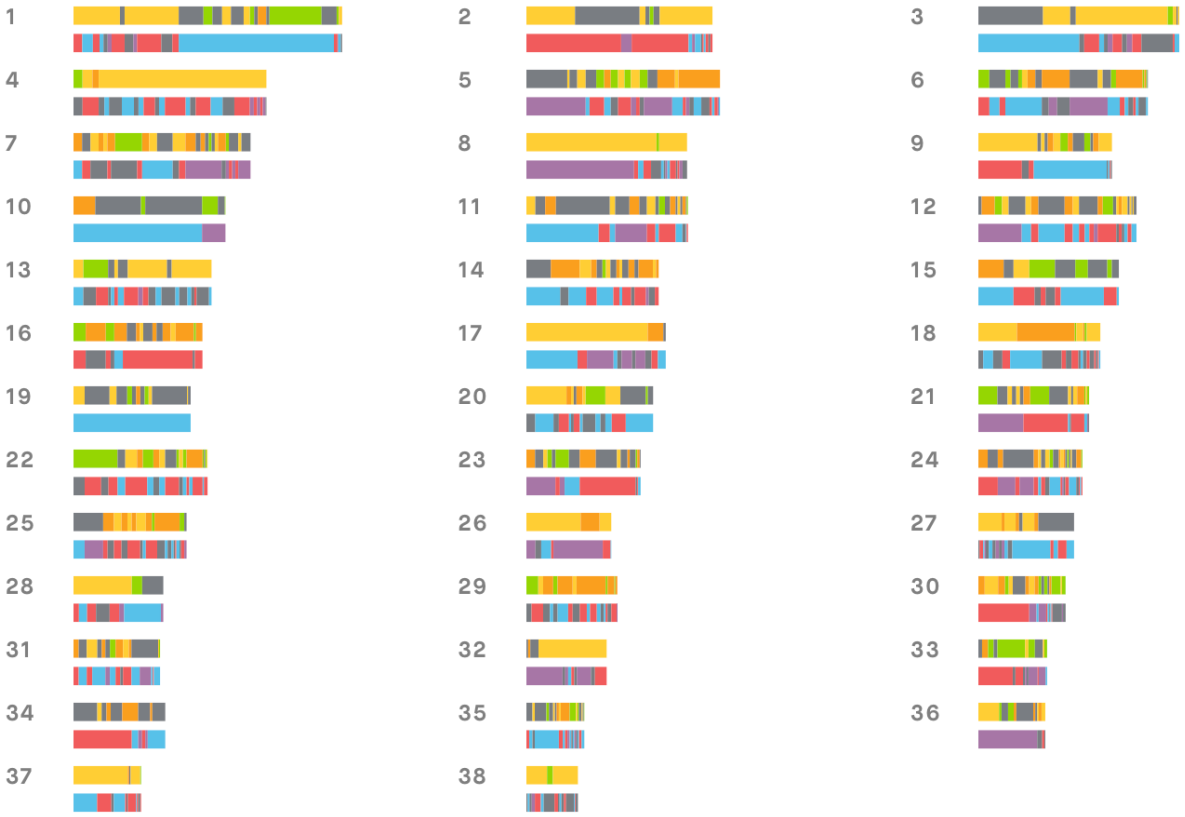


Chromosome painting

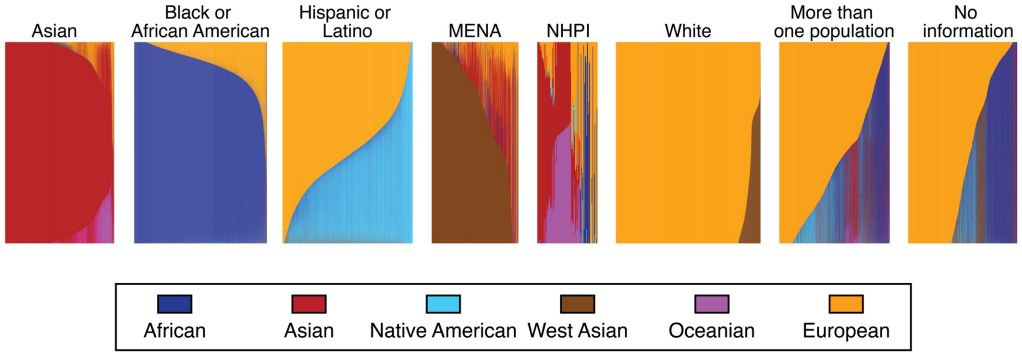
embark

Boxer German Shepherd Dog Golden Retriever Chow Chow

American Staffordshire Terrier Collie Supermutt

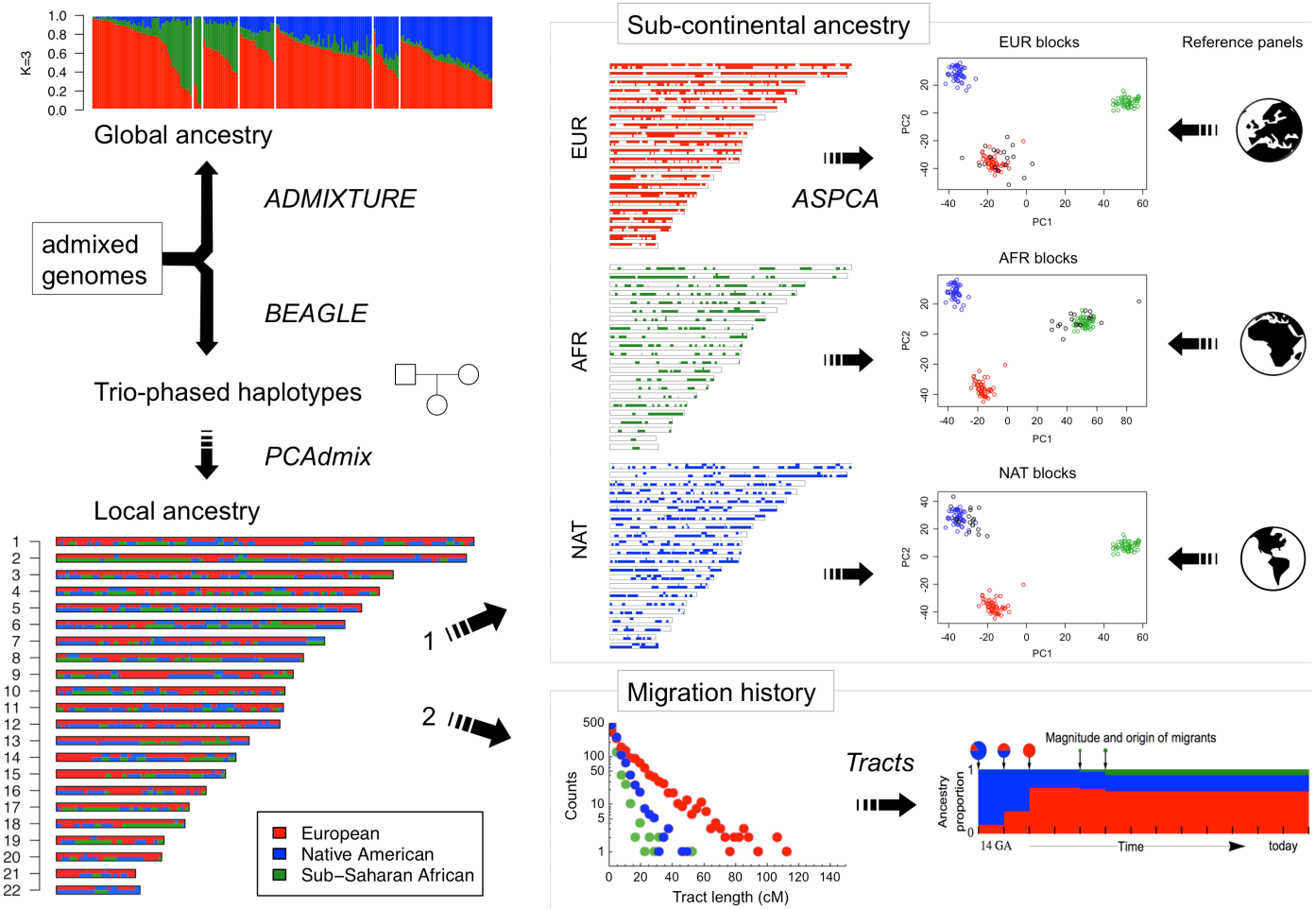


Comparisons between SIRE and ancestry

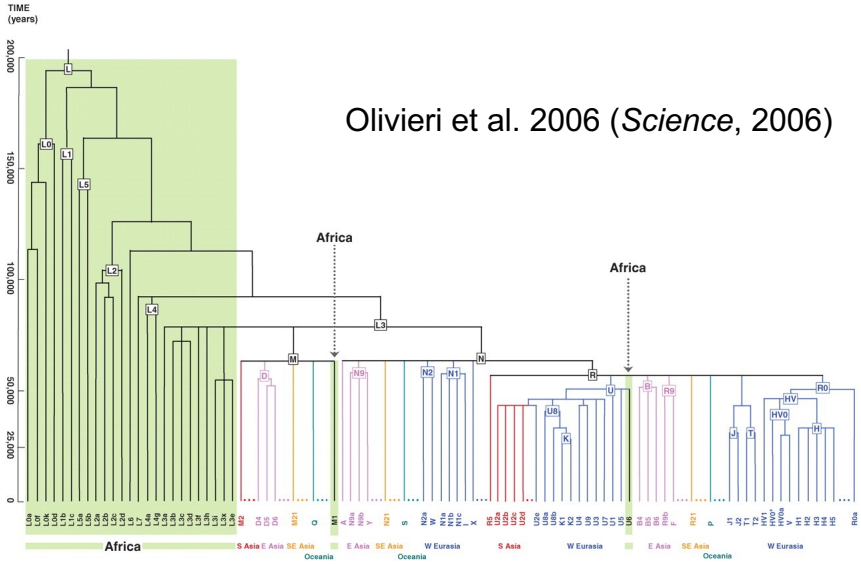


- Self-identified race and ethnicity (SIRE) is positively correlated with genetic ancestry

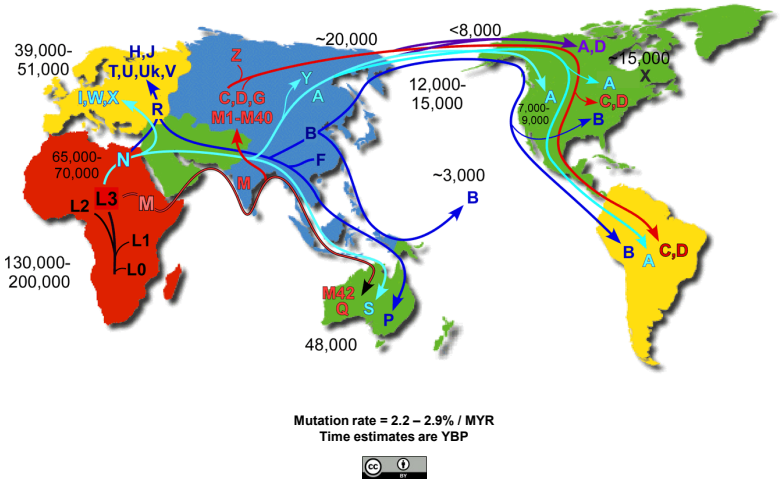
Inferring history from ancestry bocks



Maternal (mtDNA) lineages



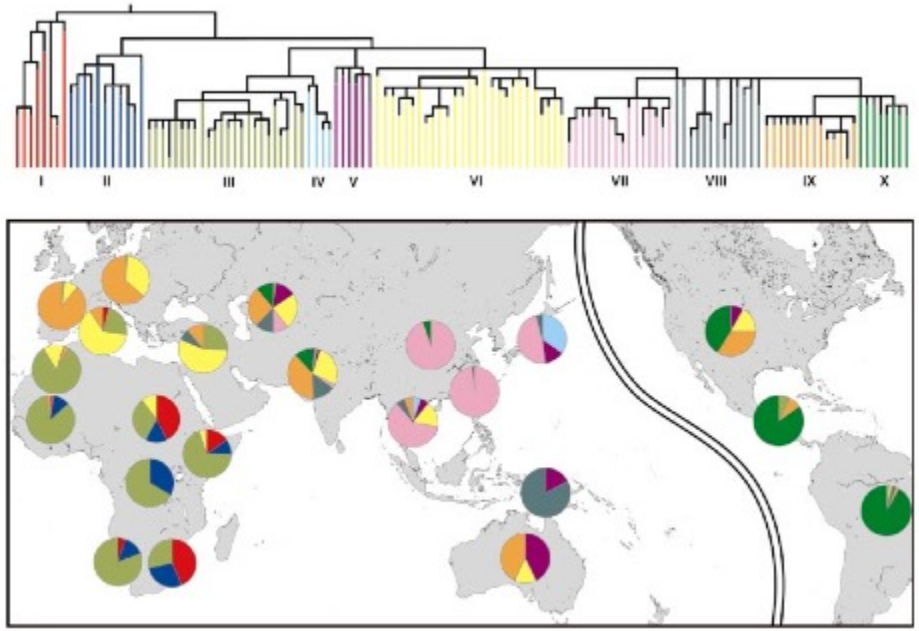
Olivieri et al. 2006 (*Science*, 2006)

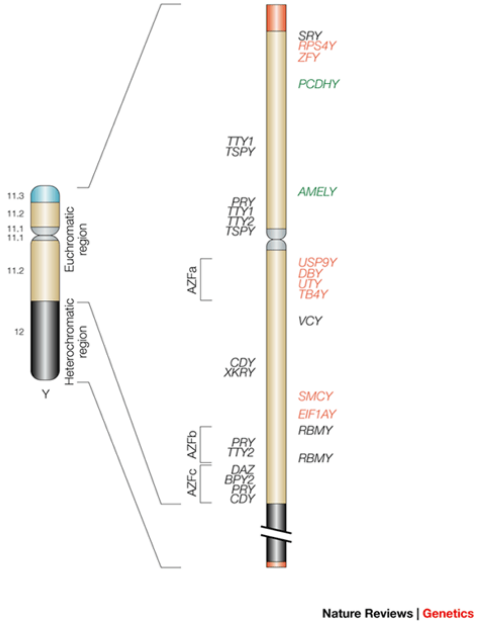


Scally and Durbin
(*Nature Reviews Genetics*, 2012)

Paternal (Y chromosome) lineages

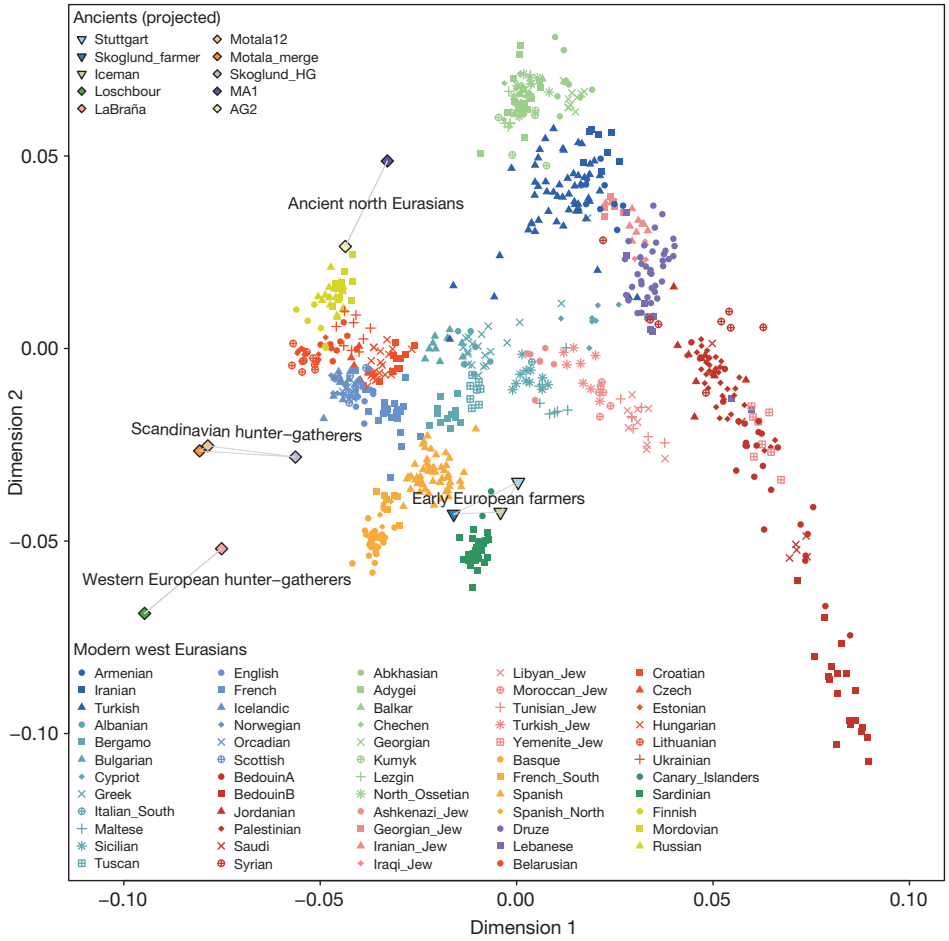
Underhill et al. 2001 (*Ann Hum Genet*, 2000)





- Y chromosome lineages are more diverse in Africa
- Mendez et al. (*AJHG*, 2013)
  - Highly divergent Y lineage (A00)… 388kya ← exact date is under contention
  - Found in African American and Central African samples

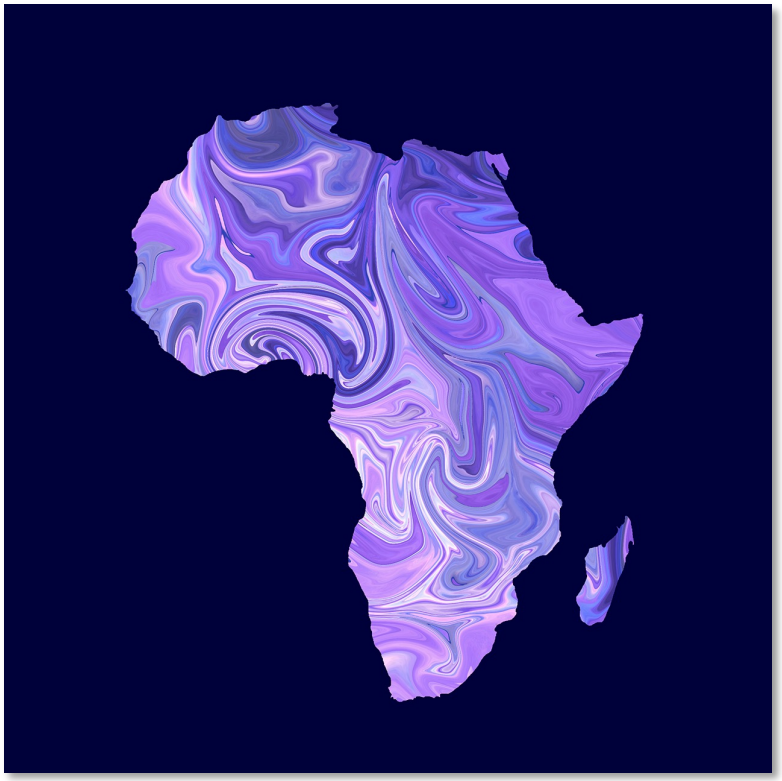
Movement into Europe



- The spread of agriculture was due to the spread of farmers, not the spread of technology

Lazaridis et al. (*Nature*, 2016)

Admixture



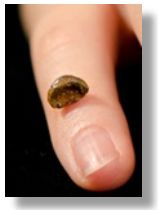
- **Admixture** refers to the mixing of divergent evolutionary lineages

Archaic introgression

- Non-African genomes contain Neanderthal DNA
  Green et al. (*Science*, 2010)

Sebastien Chabal
(Rugby player or Neanderthal?)

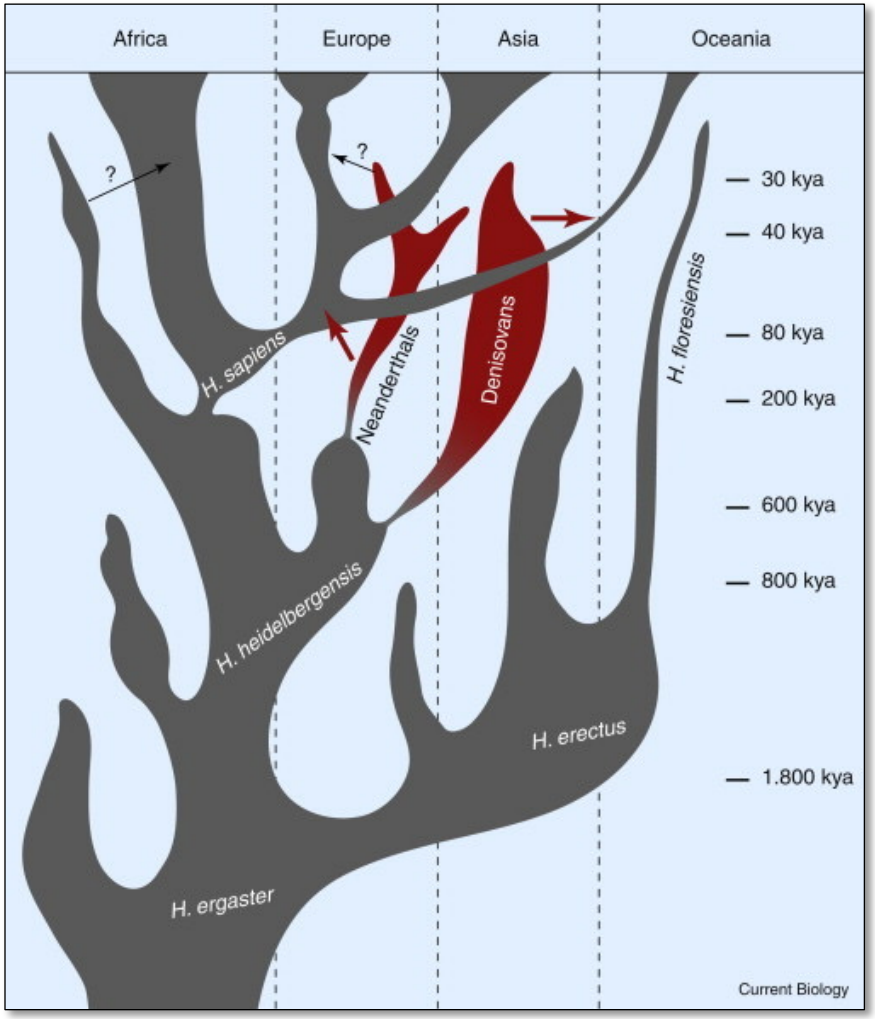
- Some modern humans also have Denisovan DNA
  Reich et al. (*Nature*, 2010)



Ars Technica

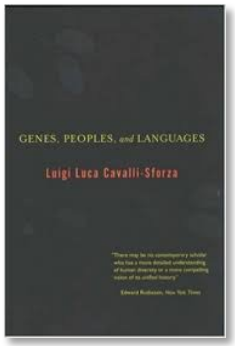
Ancient population structure

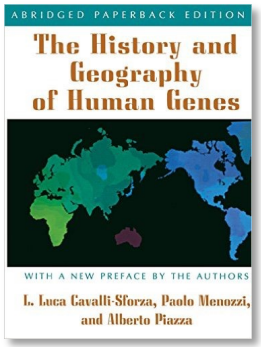
- Complex historical patterns
- Many archaic lineages died out (including *H. florensiensis*)
- Some archaic populations may have mated with our ancestors



Genetics and language

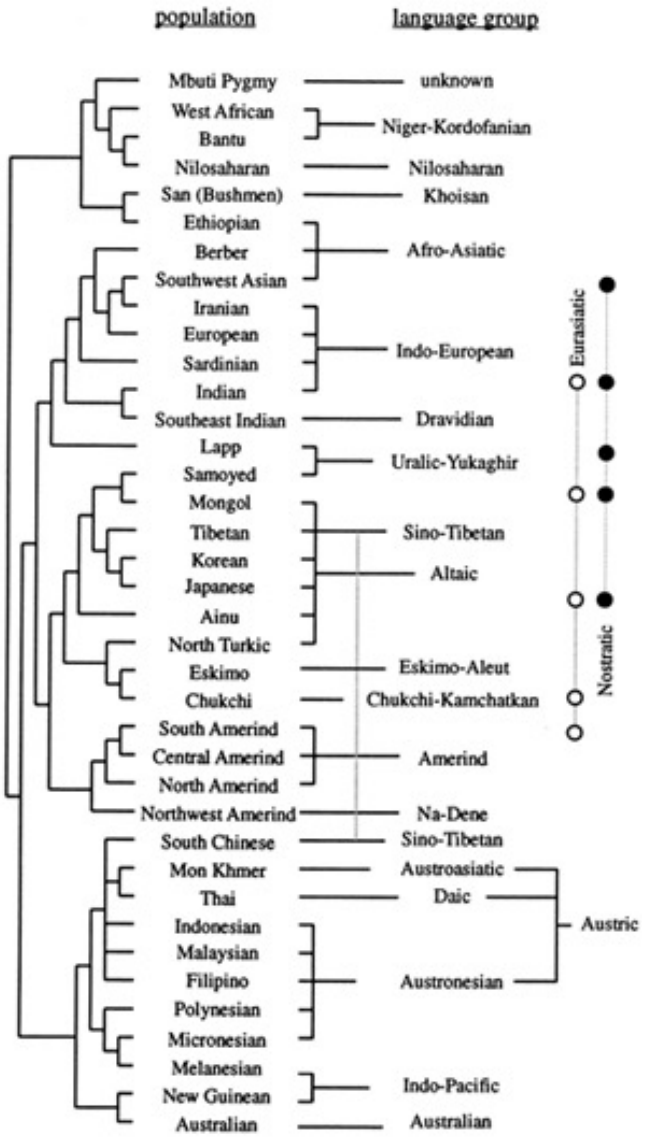
- Luca Cavalli-Sforza



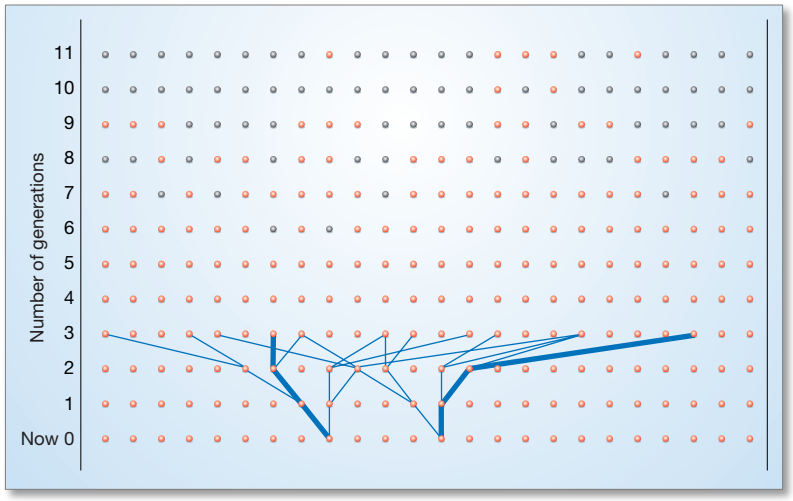


- Pioneering work using blood groups

- Populations with similar languages tend to have similar genetics



Biparental inheritance and shared ancestry

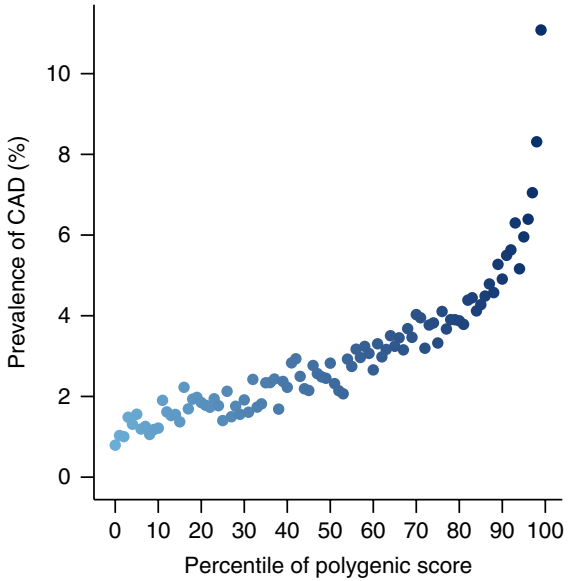


- Ancestry involves more than just DNA
- Number of ancestors t generations ago $\approx 2^t$
  - Chang (*Adv. Appl. Prob.*, 1999)
- Spain and Jewish ancestry
  - Weitz (*PLoS One*, 2014)
- Shared biparental ancestry as recent as 2500 years ago?
  - Rohde et al. (*Nature*, 2004)
  - Lachance (*Theo. Pop. Biol.*, 2009)

Image from Hein (*Nature*, 2004)

The generalizability problem

Polygenic risk scores (PRS)



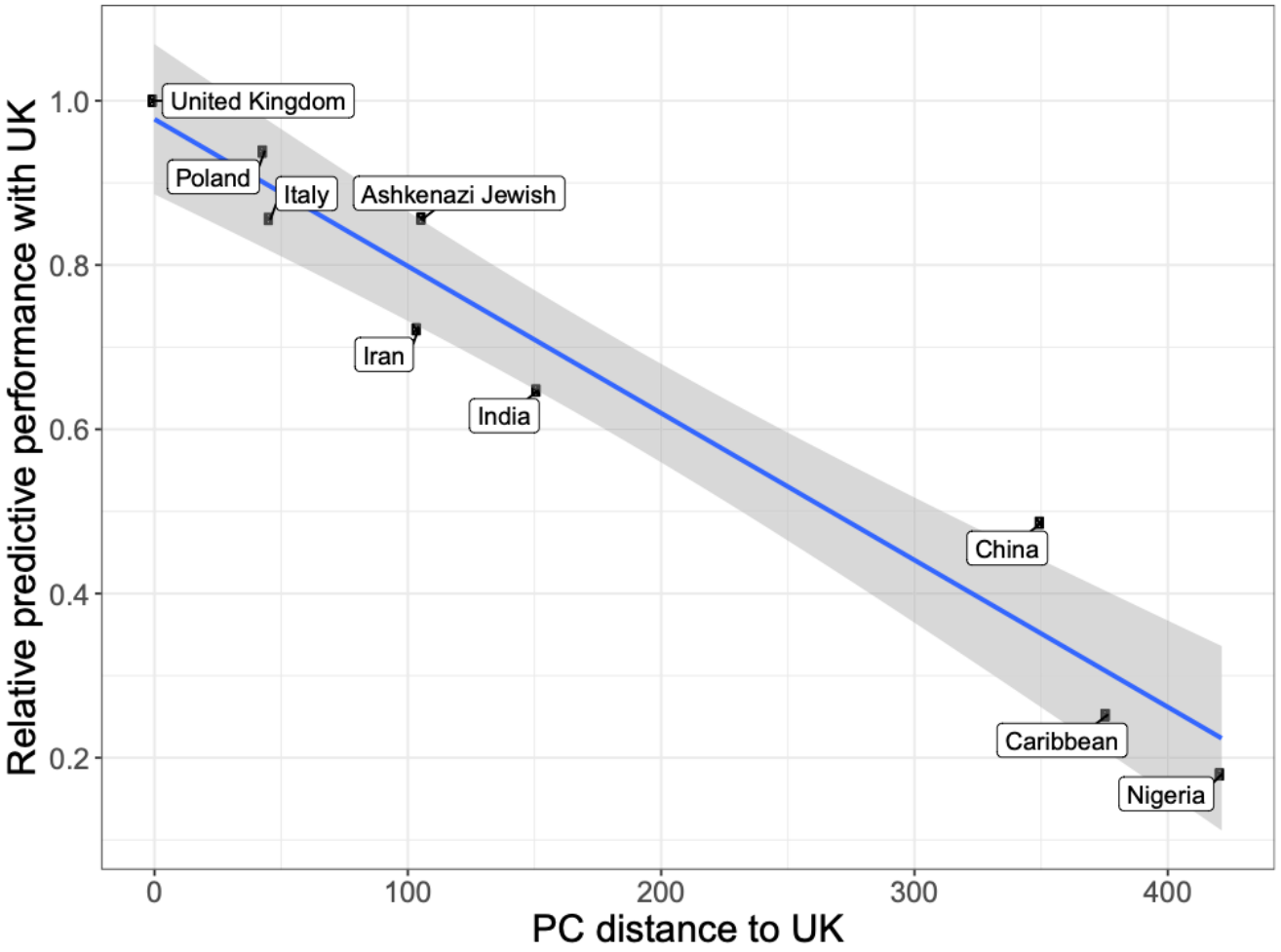
$$PRS_j = \sum_{i=1}^{L} d_{ij} \times \beta_i$$

i : SNP #
j : individual #
d : allele dose
β : effect size
L : total # of SNPs

- Counts of risk-increasing alleles can be used to generate individualize predictions of disease risk
- Importantly, frequencies of risk-increasing alleles differ across populations
- Effect sizes and the ability of PRS variants to tag causal alleles can also differ across populations

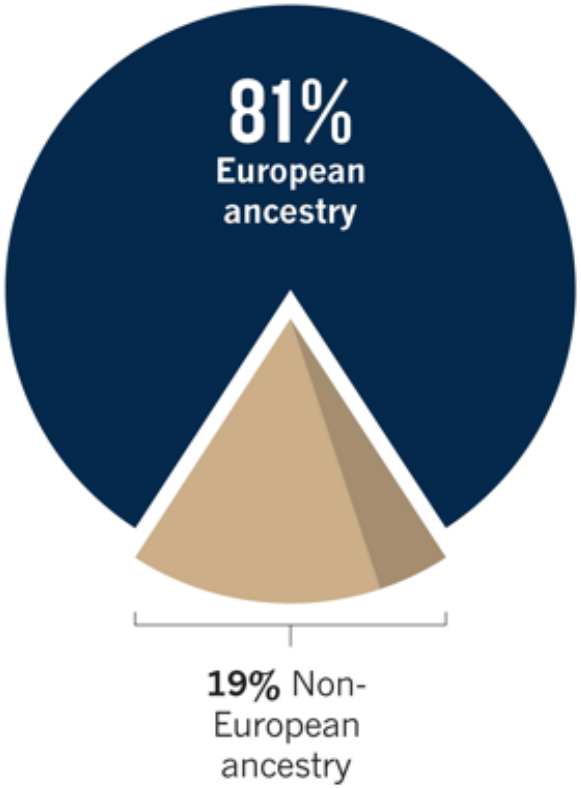
Khera et al. (*Nature Genetics*, 2018)

Polygenic predictions do not generalize well



Privé et al. (*AJHG*, 2022)

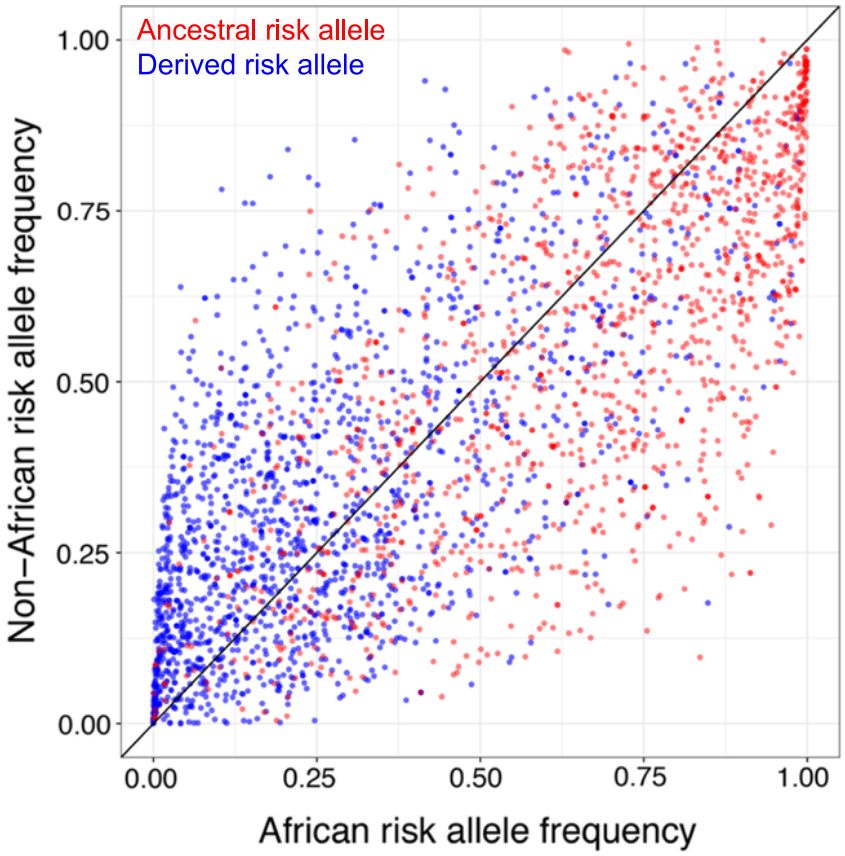
Most GWAS have used European samples



- This sampling exacerbates existing health disparities

Popejoy and Fullerton (*Nature*, 2016)

SNP ascertainment bias



NHGRI-EBI GWAS Catalog

- The set of known disease-associations is biased (there is enrichment for SNPs with intermediate allele frequencies in Europe)

Kim et al. (*Genome Biology*, 2018)

Genotyping technologies contribute to bias

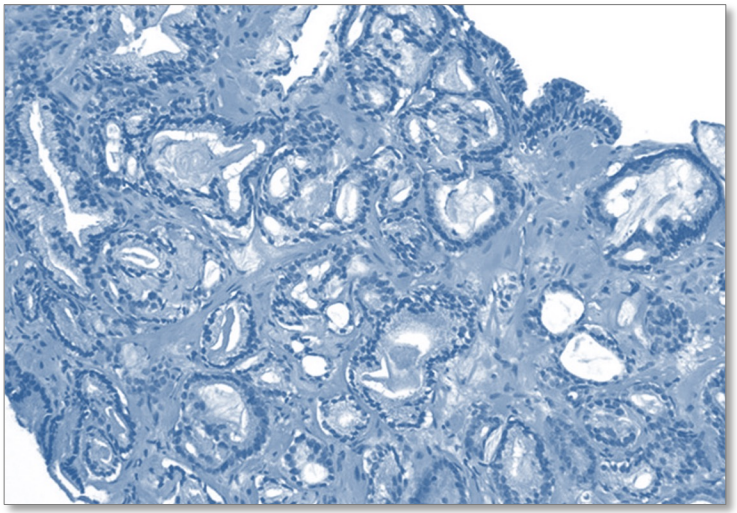
Why might GWAS findings replicate poorly across populations?

- Allele frequency differences
  (including private alleles)

- Linkage disequilibrium varies across populations
  (tag SNPs need not be causal)

- Effect size differences
  (including genotype-by-environment interactions)

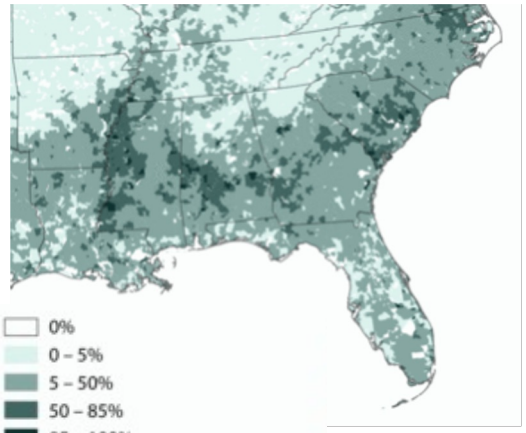
Case study: Prostate cancer genetics



- Prostate cancer has a high heritability (h^2 = 58%)
- The relative risk of men with affected fathers is 2.1-fold higher compared to men without a family history
- However, much of what we know about this disease comes from studies of individuals of European descent

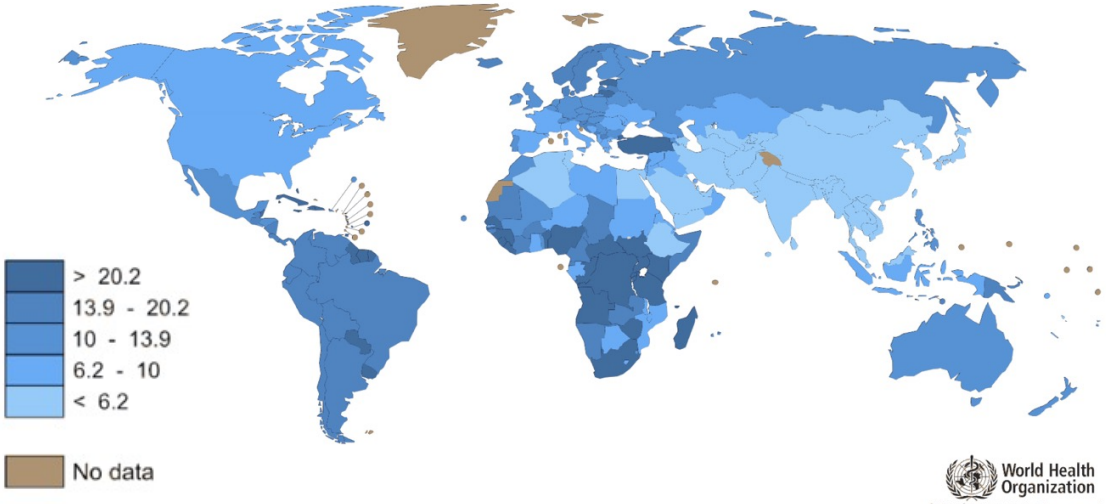
Men of African decent have higher risks of prostate cancer

African-American %

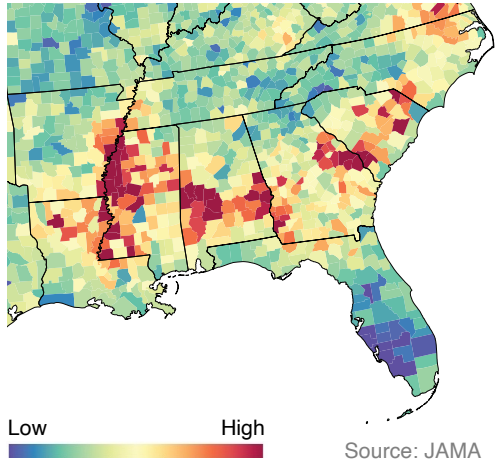


Source: US census

Global prostate cancer mortality rates



Prostate cancer mortality



Source: JAMA

Ancestry-matched risk scores perform better

Source of polygenic risk score	AUC_{UKBB}	AUC_{MADCaP}
Conti (Multi-ancestry)	0.703	0.579
Conti (European)	**0.707**	0.541
Conti (African)	0.671	**0.585**
Conti (Asian)	0.662	0.533
Conti (Hispanic)	0.678	0.527
Karunamuni (European)	**0.612**	0.502
Karunamuni (European + African)	0.608	**0.547**

- Polygenic risk scores perform better if they are ancestry-matched

Going forward…

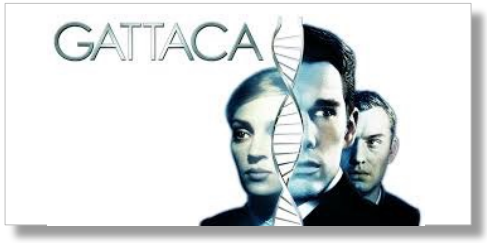
Image from HealthToday

- Evolutionary genomics and functional genomics can be leveraged to better understand human health and disease

- There is a need to conduct genetic studies in diverge populations

What will our genomes look like in the future?

- Genetic engineering



Columbia Pictures

- Changes in selection pressures

Disney · PIXAR

- Population admixture