SISG 2022 - Module 2

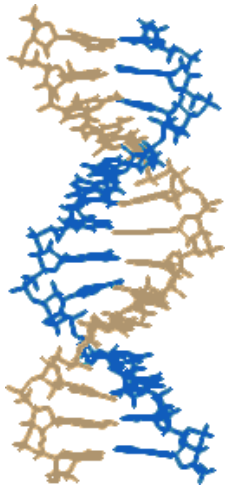
Introduction to Genetics and Genomics

Genetic Ancestry

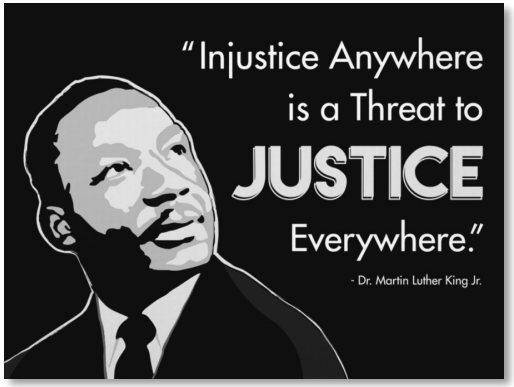
12:15pm EDT, Wednesday, July 13th

Joe Lachance and Greg Gibson

joseph.lachance@biology.gatech.edu



Terminology

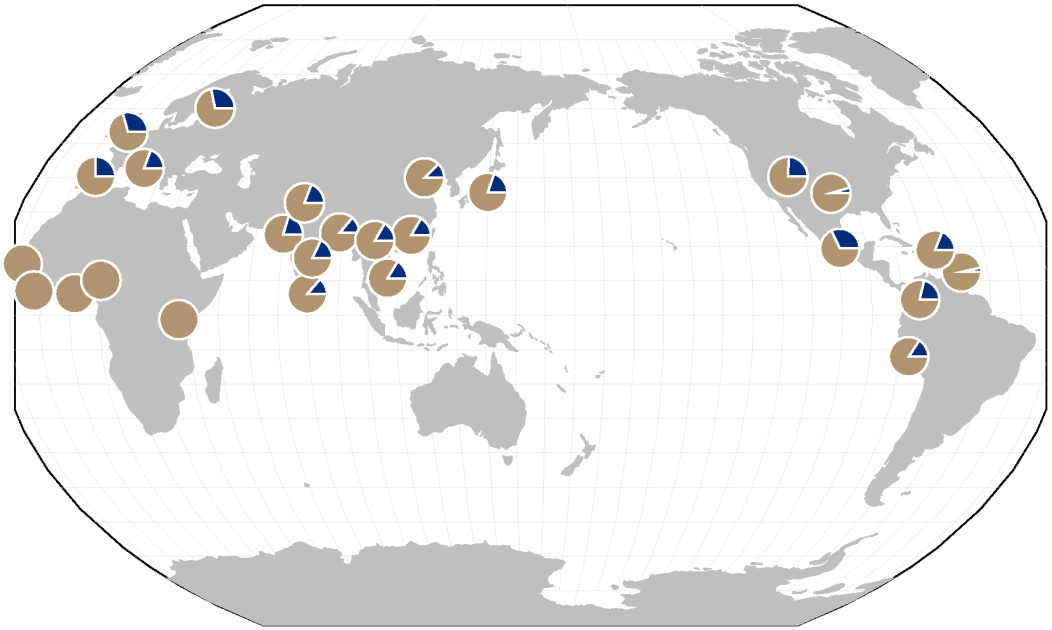


- **Race** refers to an individual's self-identification with one or more social groups (as defined by the US census bureau)
- **Ethnicity** refers to the way in which one identifies learned aspects of themselves (e.g., language and culture)
- **Ancestry** refers to the populations that individuals are descended from (this term is preferred by geneticists)
- **Populations** are often defined in terms of sampling locations

The changing face of humanity

Photos by Martin Schoeller for National Geographic

AIMs

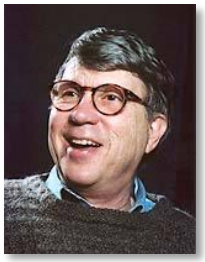


- Ancestry Informative Markers (AIMs) have large allele frequency differences between populations
- Rare alleles are more likely to be population-specific
- No single AIM is a perfect classifier

Image source: GGV

Variance partitioning and Lewontin's Fallacy

- Richard Lewontin (1972)
  - Shannon's diversity index used
  - 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")

- Variance partitioning and classification are separate issues

HGDP

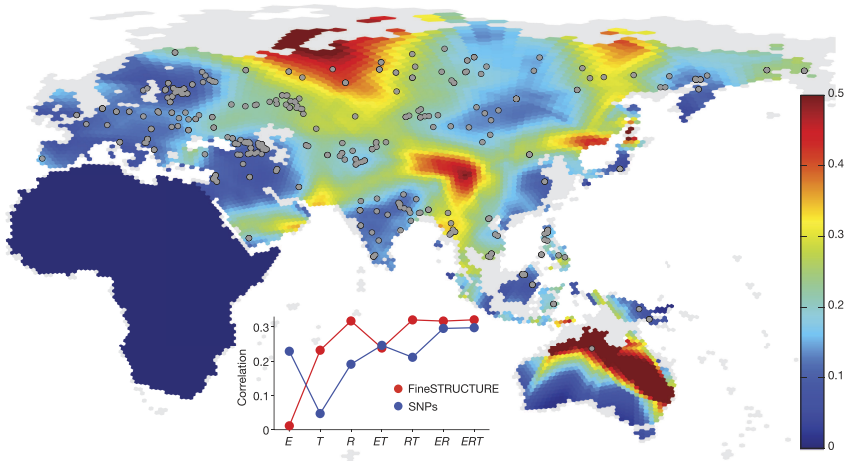
- The Human Genome Diversity Project (HGDP): >50 sampled populations
- Ethical issues:
  - Indigenous groups need not be need not be isolated populations
  - Accusations of “helicopter science”

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

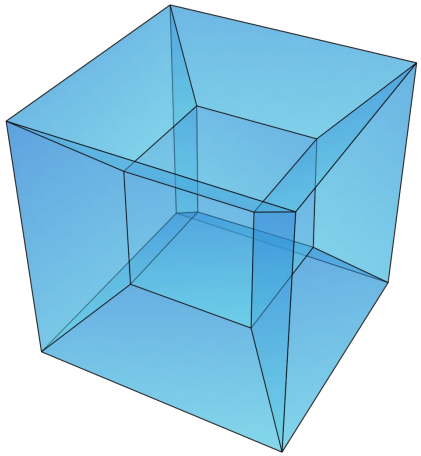
Image: VectorStock

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

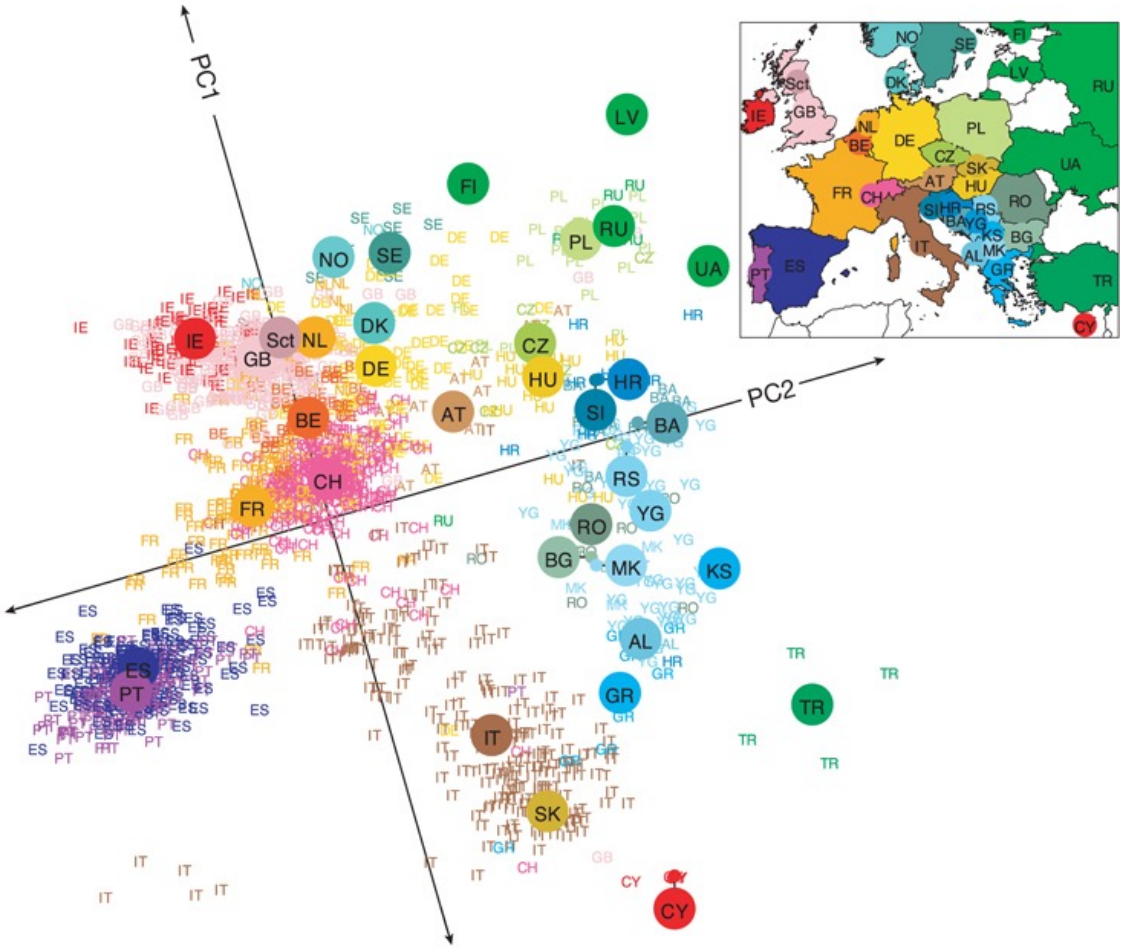
Dimensionality and PCA



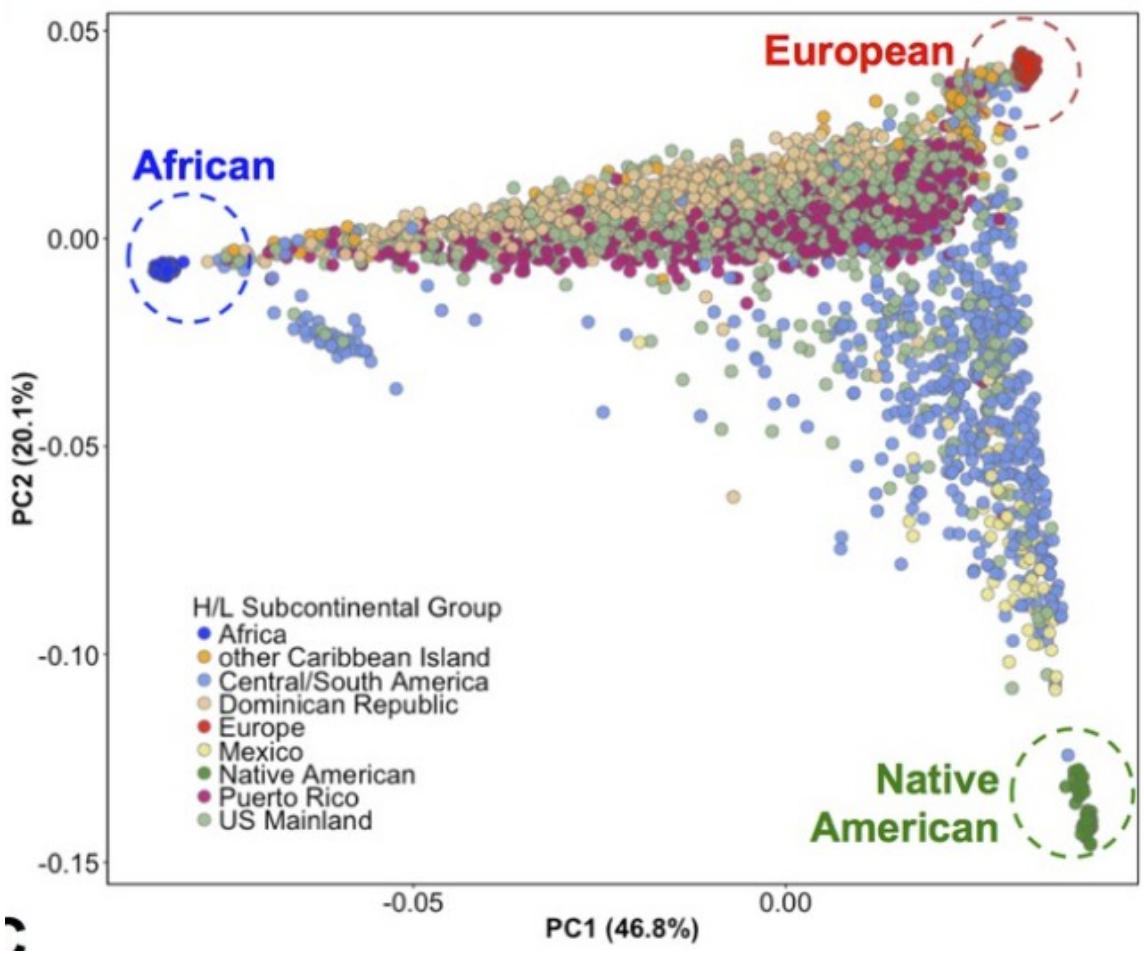
- Principal Component Analysis (PCA) is one way to reduce the dimensionality of genetic datasets
- Each PC refers to an orthogonal (perpendicular) dimension – each PC is an eigenvector and eigenvalues correspond to the % of variance explained by each PC
- PCA can be used to represent samples in a genetic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

Hypercube image from Wikipedia

Genes mirror geography in Europe

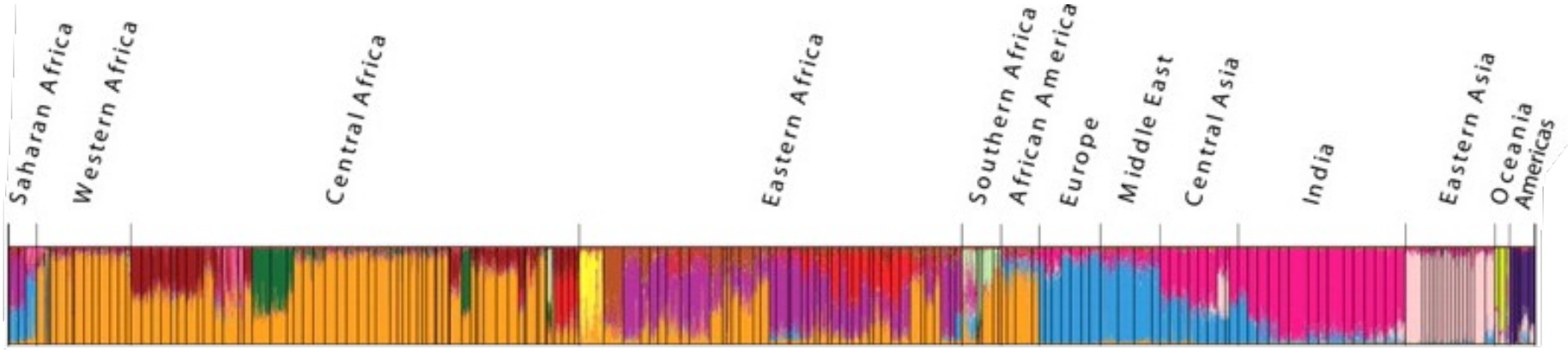


Human diversity exists along a continuum



Belbin et al. (*eLife*, 2017)

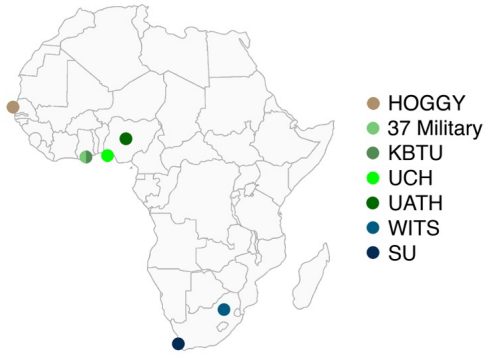
STRUCTURE Plots



- Every individual's genome is a combination of different ancestries
- Each genetic ancestry is represented as a different color
  (K = 14 indicates that there are 14 different colors/ancestries)

Figure from Tishkoff et al. (*Science*, 2009)

ADMIXTURE plots of African diversity



Study site	Location	Cases	Controls
Hôpital Général de Grand Yoff (HOGGY)	Dakar, Senegal	56	59
37 Military Hospital (37 Military)	Accra, Ghana	59	59
Korle-Bu Teaching Hospital (KBTH)	Accra, Ghana	53	58
University College Hospital (UCH)	Ibadan, Nigeria	56	56
University of Abuja Teaching Hospital (UATH)	Abuja, Nigeria	56	57
WITS Health Consortium (WITS)	Johannesburg, South Africa	61	61
Stellenbosch University (SU)	Cape Town, South Africa	58	53

MADCaP NETWORK

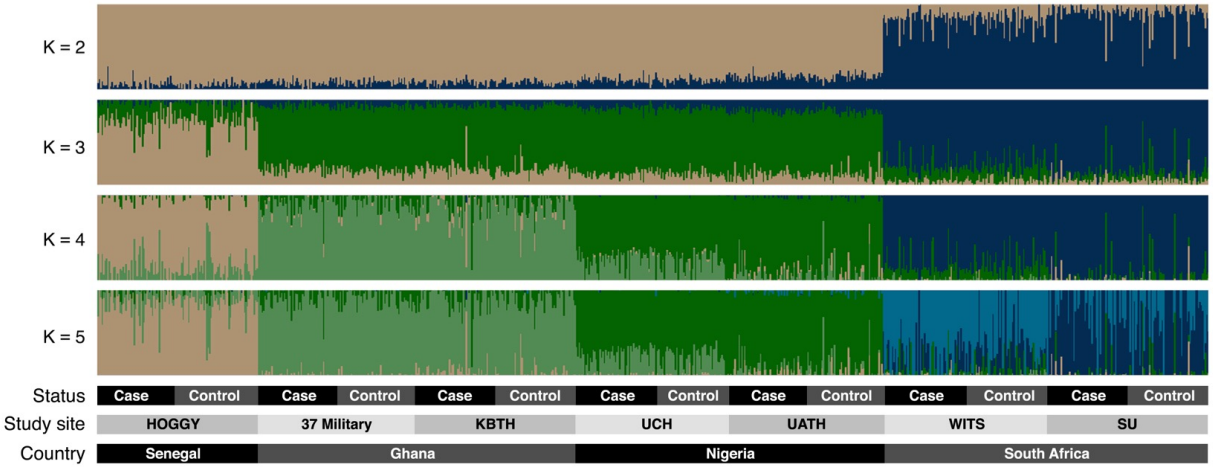
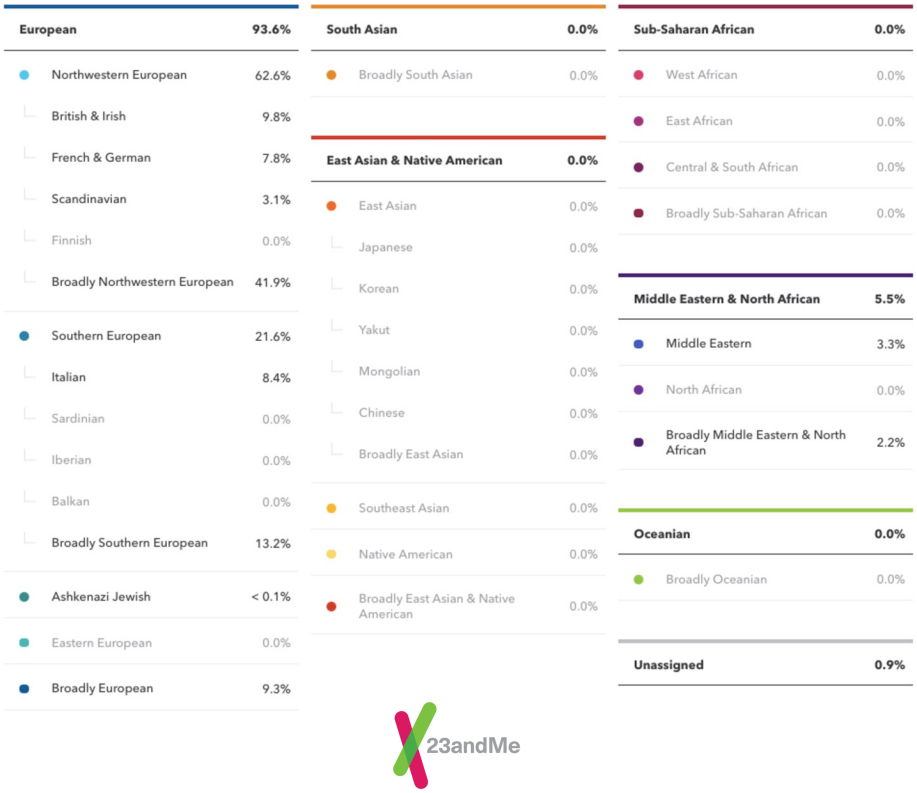
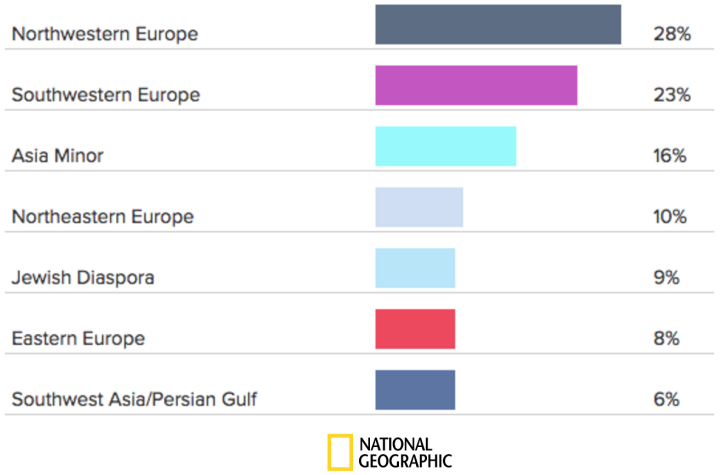
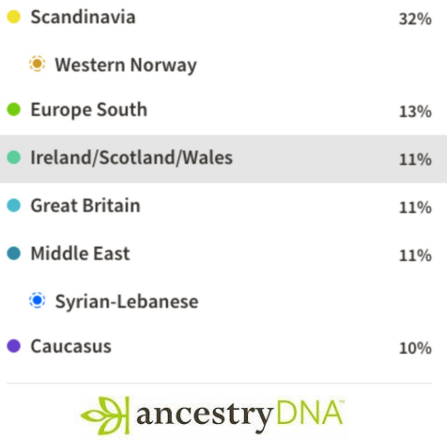


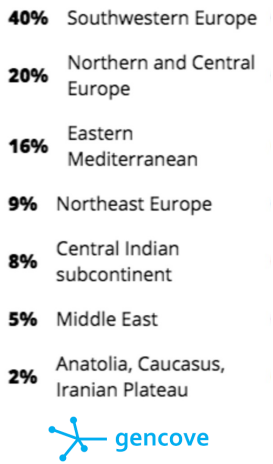
Figure from Harlemon et al. (*Cancer Research*, 2020)

Ancestry inference and DTC testing







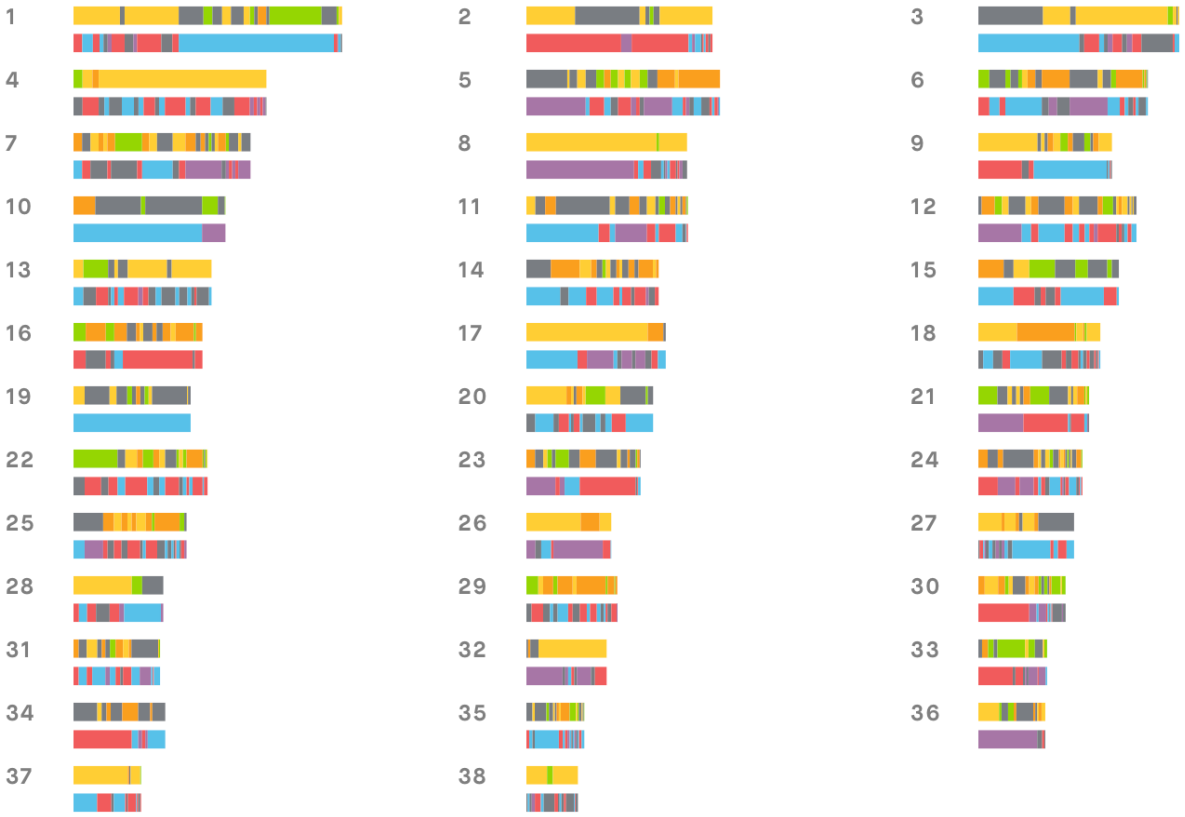


Chromosome painting

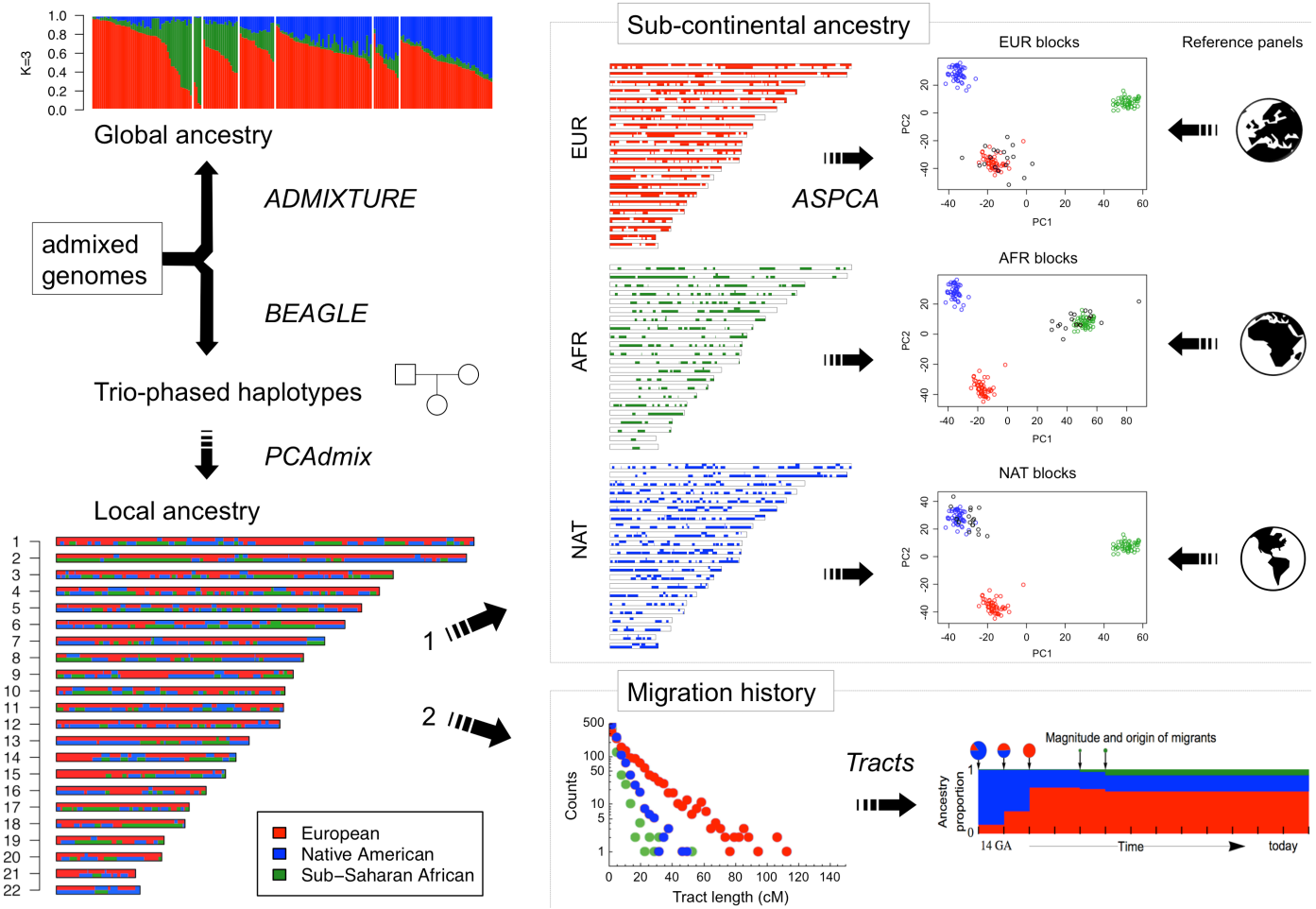
embark

Boxer German Shepherd Dog Golden Retriever Chow Chow

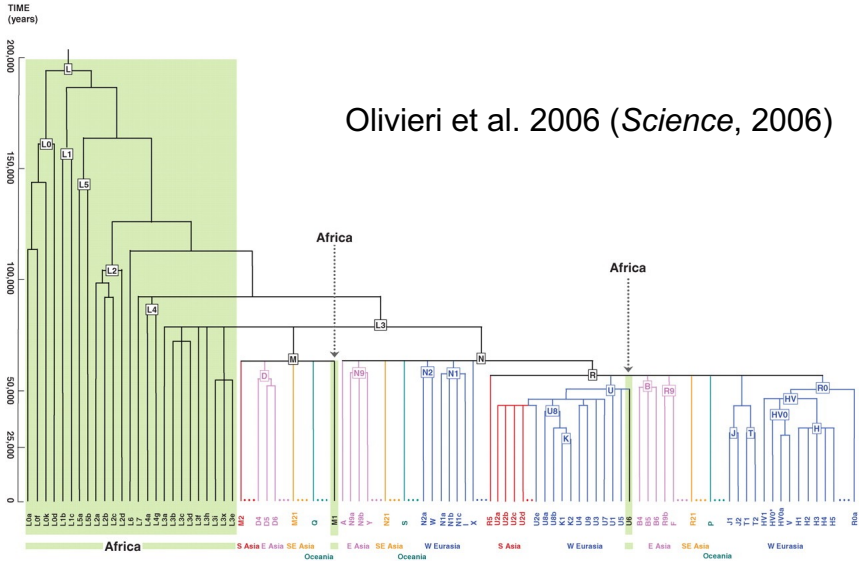
American Staffordshire Terrier Collie Supermutt



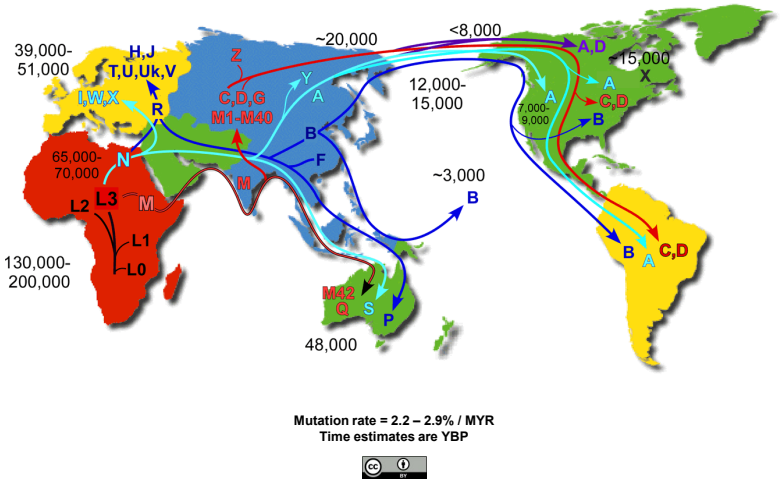
Inferring history from ancestry bocks



Maternal (mtDNA) lineages



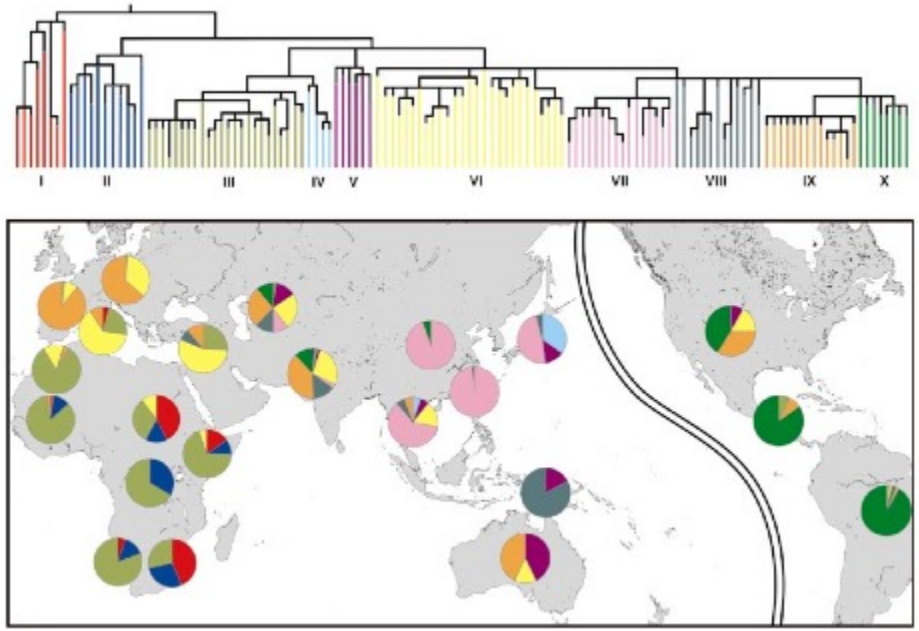
Olivieri et al. 2006 (*Science*, 2006)

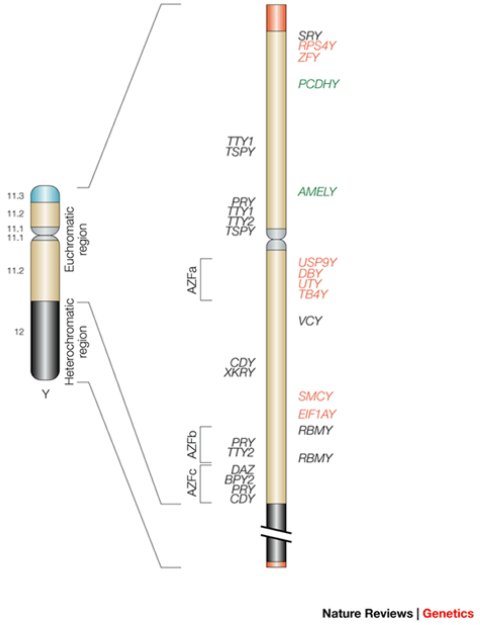


Scally and Durbin
(*Nature Reviews Genetics*, 2012)

Paternal (Y chromosome) lineages

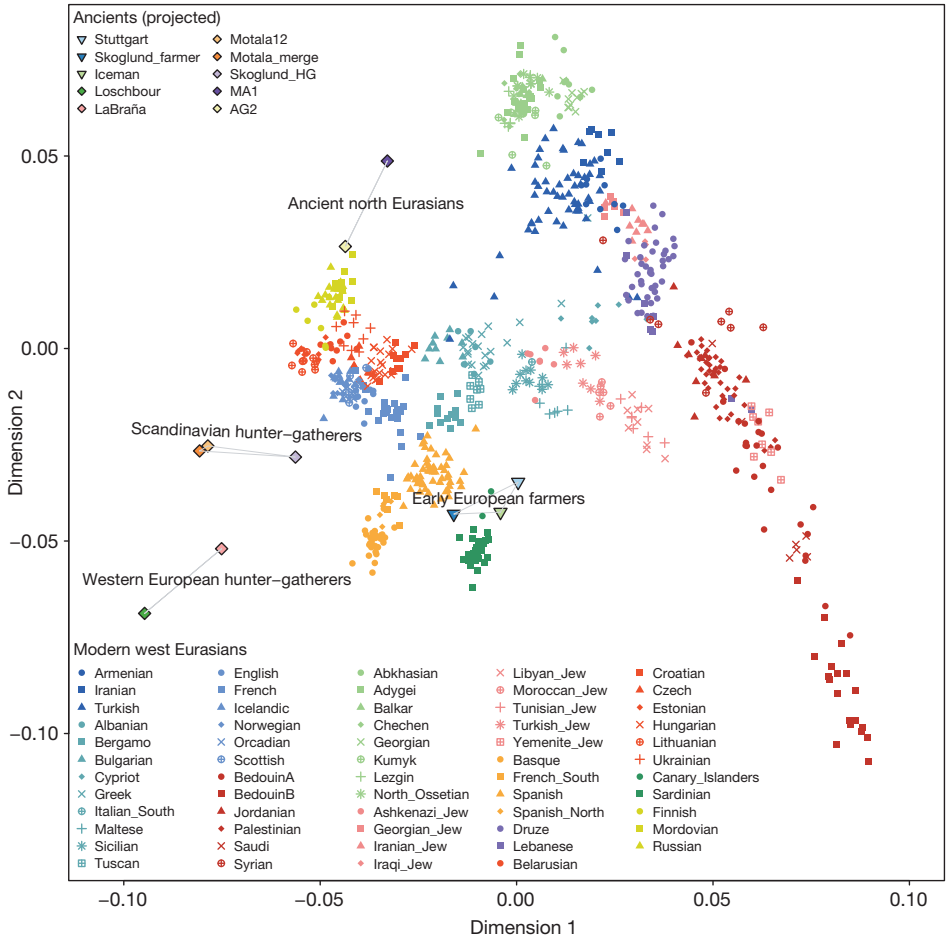
Underhill et al. 2001 (*Ann Hum Genet*, 2000)





- 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)

Archaic introgression

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

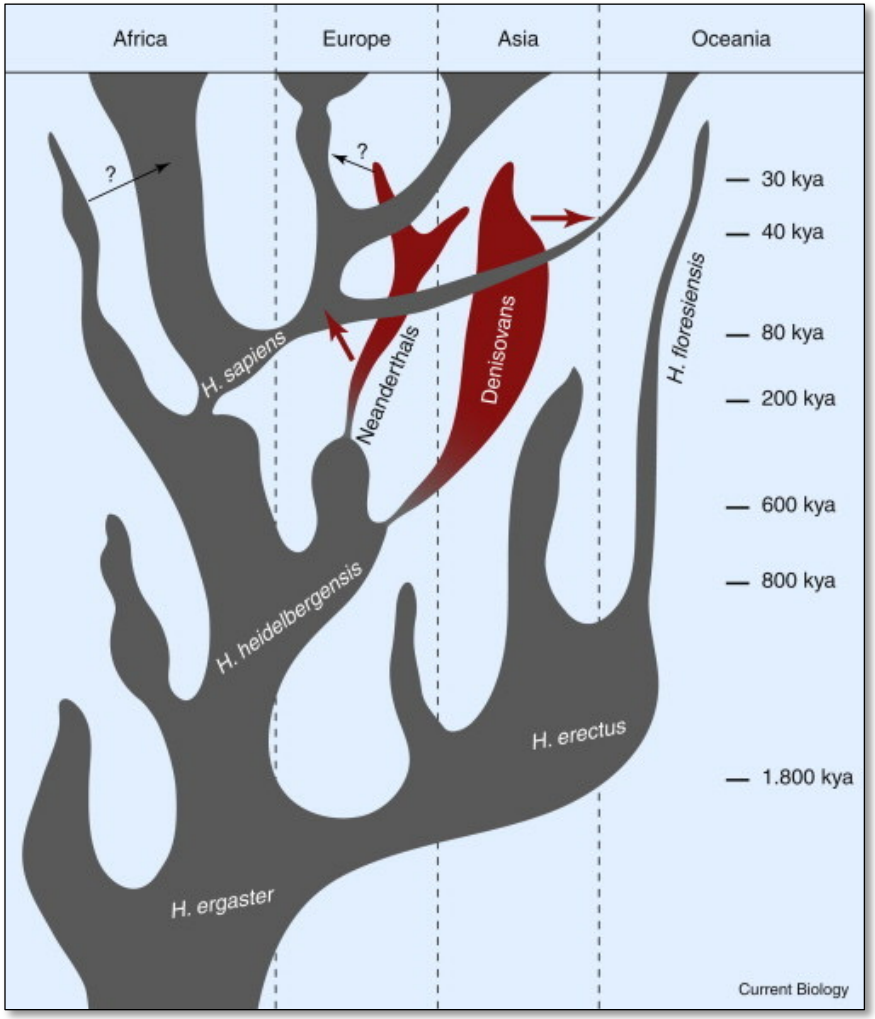
Sebastien Chabal
(Rugby player or Neanderthal?)

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

Ars Technica

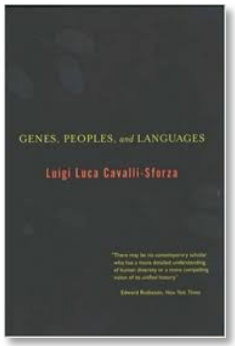
Ancient population structure

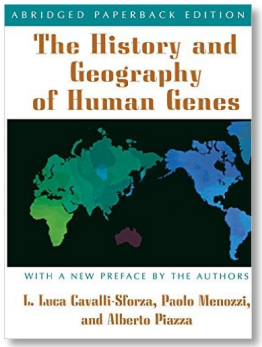
- Complex historical patterns
- Many archaic lineages died out (including *H. florensiensis*)
- Some archaic populations may have mated with our ancestors



Genetics and language

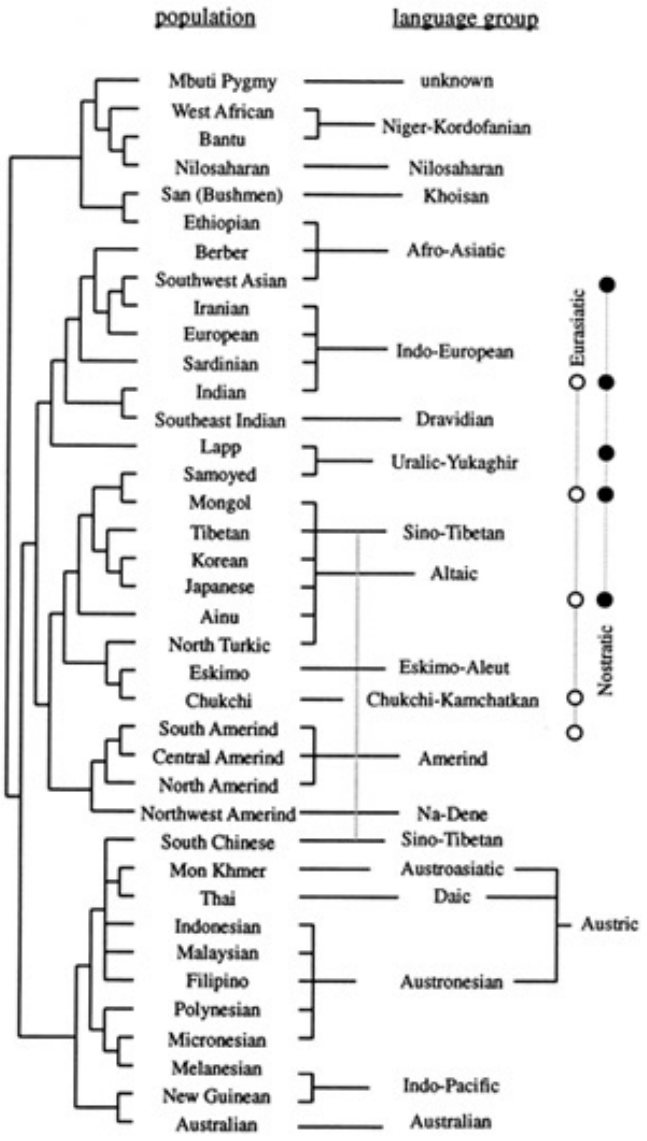
- Luca Cavalli-Sforza



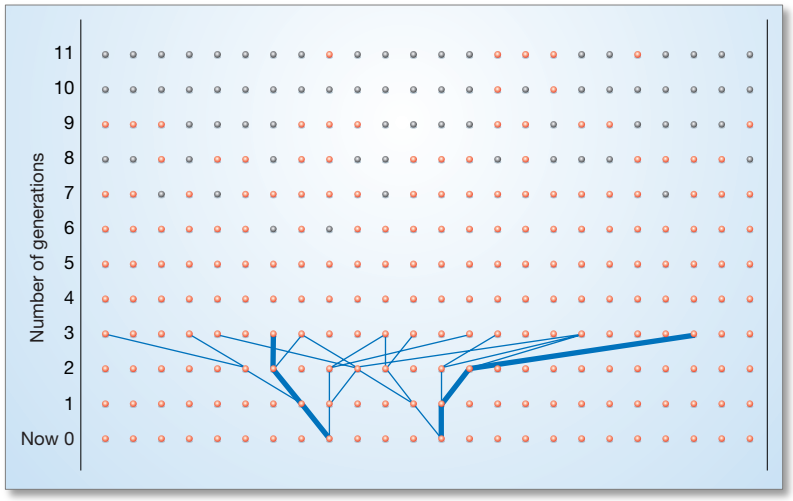


- Pioneering work using blood groups

- Populations with similar languages tend to have similar genetics



Biparental inheritance and shared ancestry

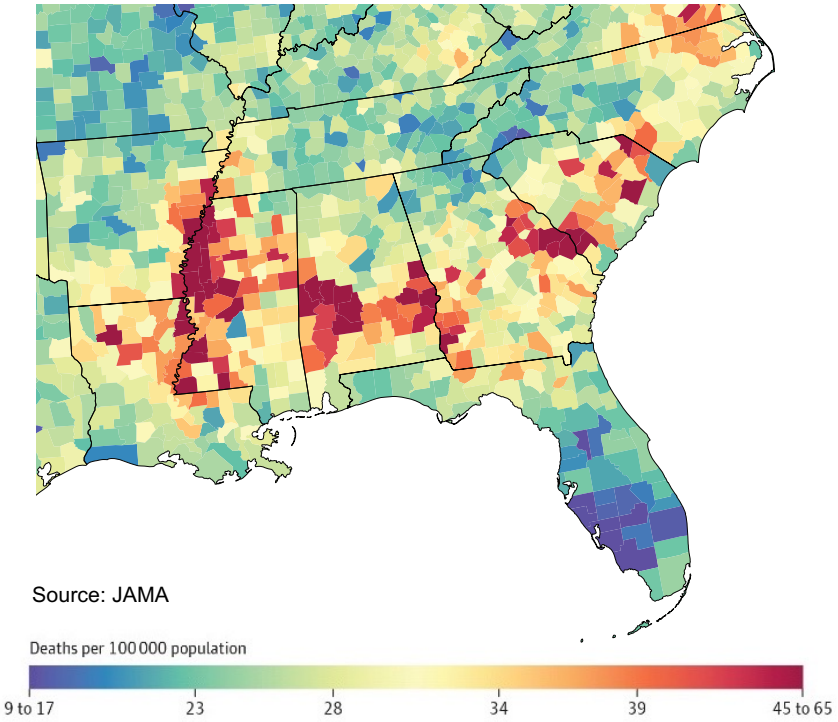


- Ancestry involves more than just DNA
- Number of ancestors t generations ago $\approx 2^t$
  - Chang (*Adv. Appl. Prob.*, 1999)
- Spain and Jewish ancestry
  - Weitz (*PLoS One*, 2014)
- Shared biparental ancestry as recent as 2500 years ago?
  - Rohde et al. (*Nature*, 2004)
  - Lachance (*Theo. Pop.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

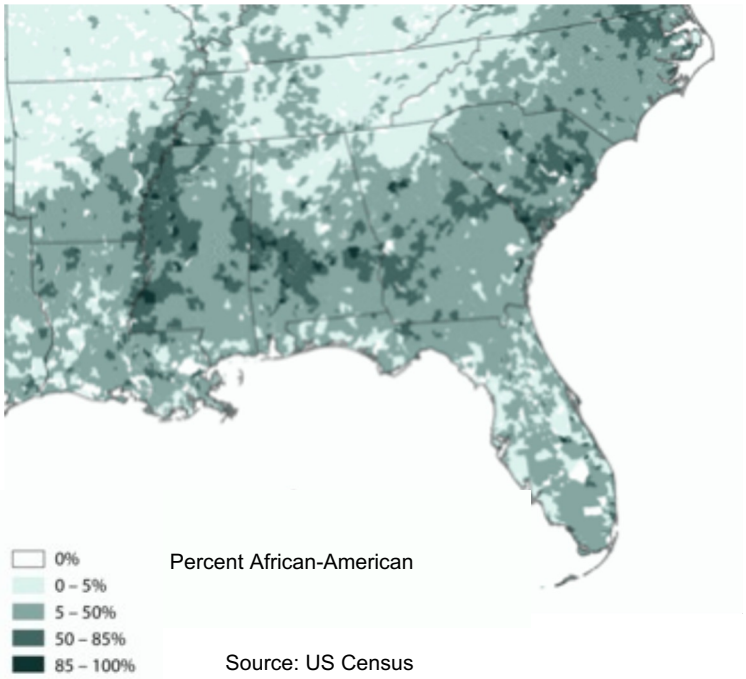
Image from Hein (*Nature*, 2004)

Ancestry and health disparities

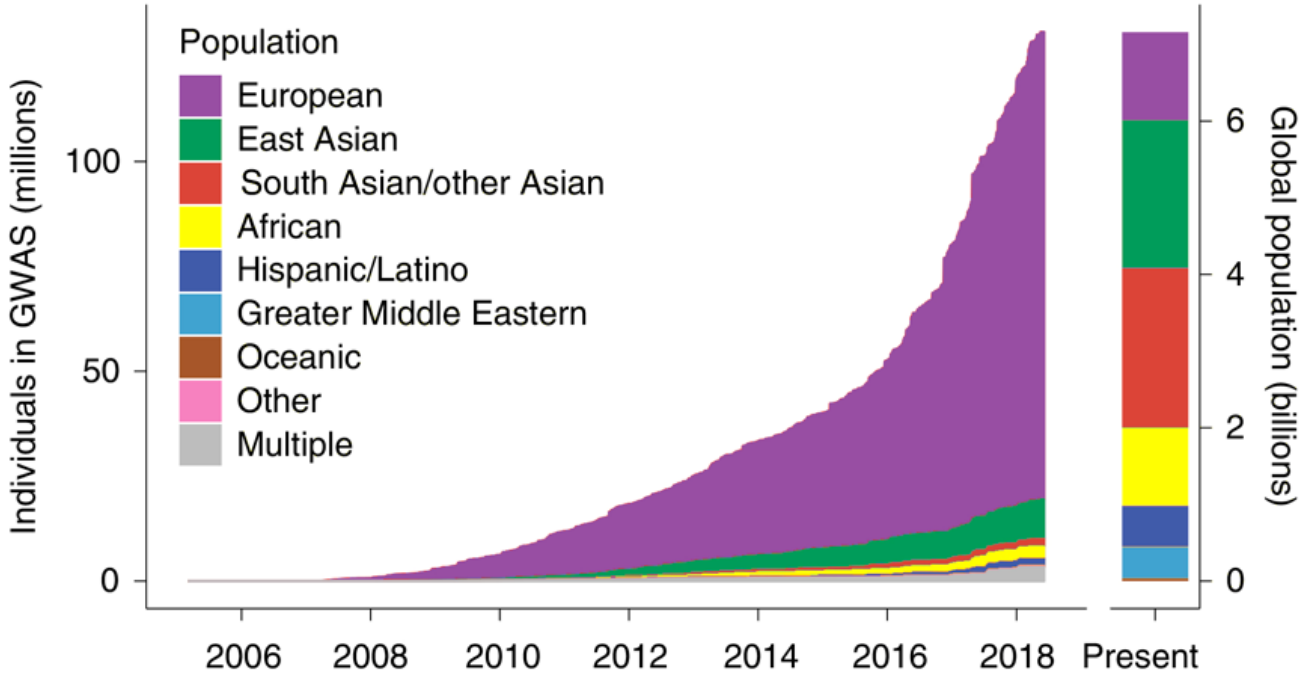
Prostate cancer mortality rate



Proportion African-American



Most GWAS have used European samples



- This sampling exacerbates existing health disparities

Martin et al. (*Nature Genetics*, 2019)

SNP ascertainment bias

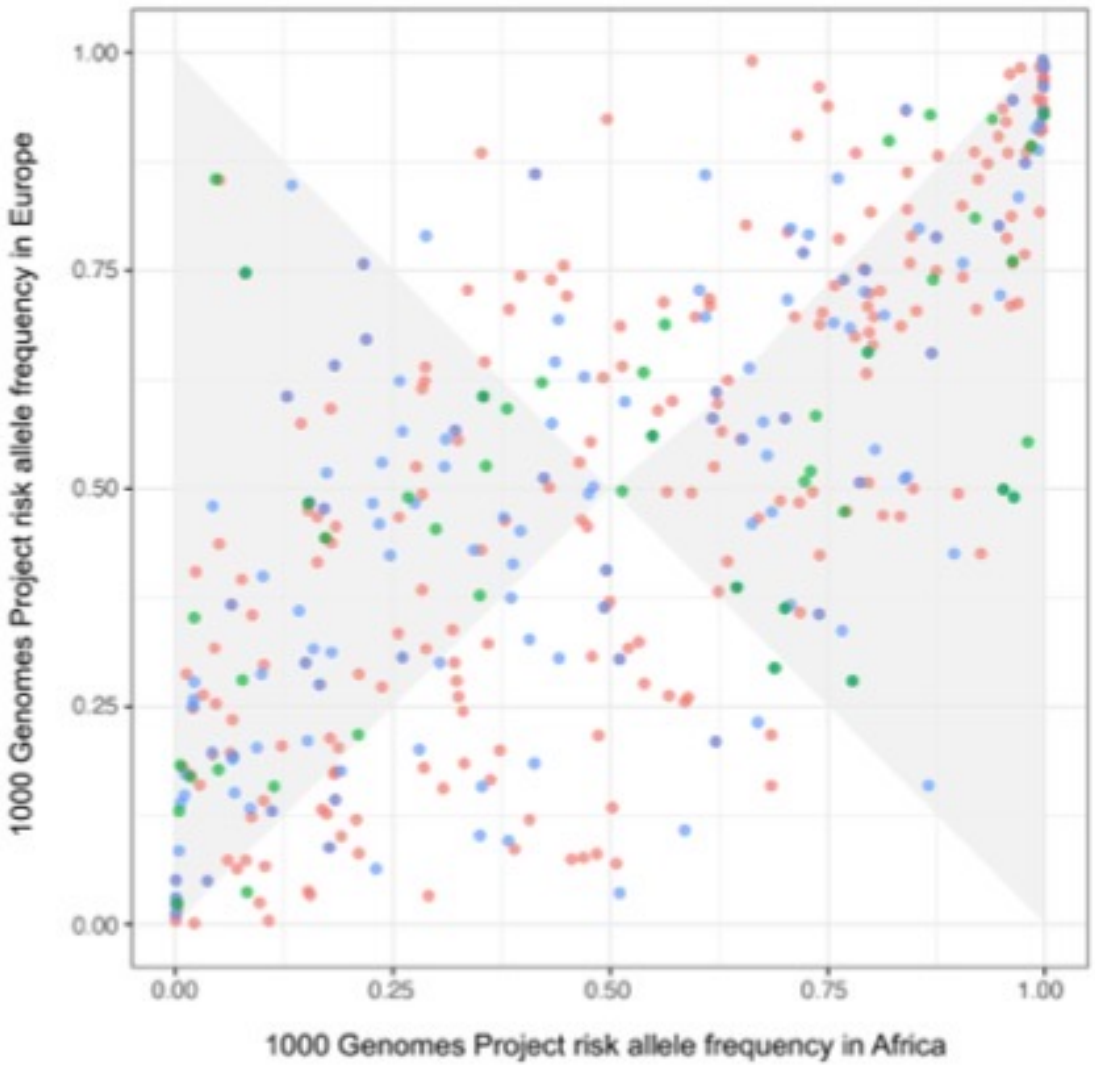
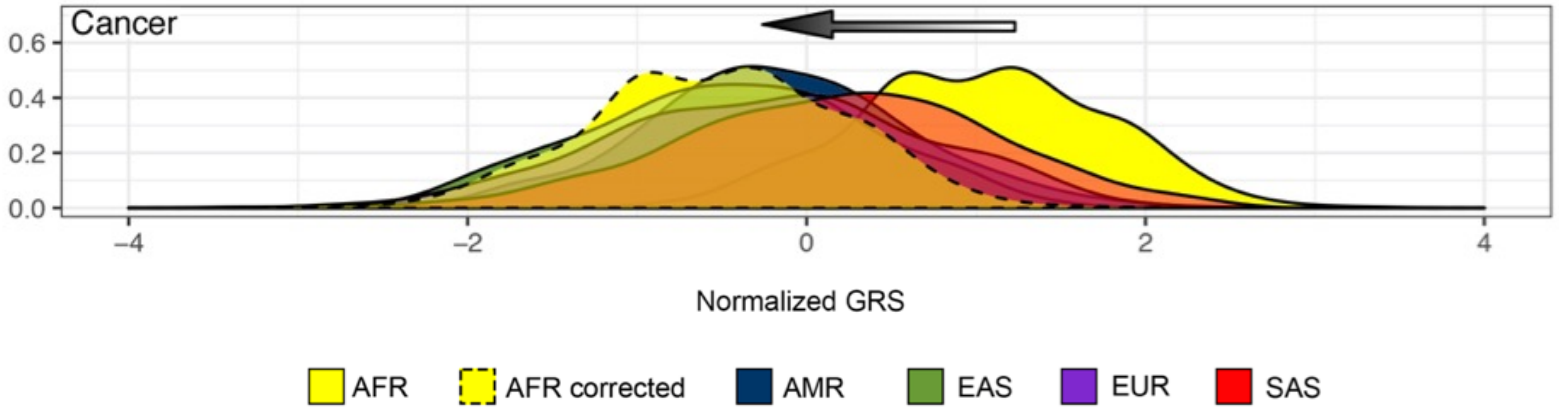


Figure from Kim et al. (*Genome Biology*, 2022)

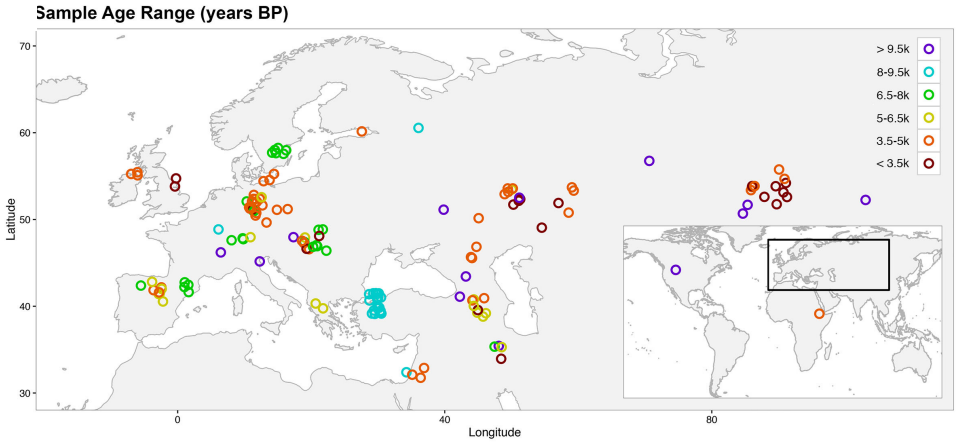
Bias in polygenic risk scores



- Polygenetic risk scores do not always generalize well across populations
- Ascertainment bias can create the illusion of genetic health disparities
- Evolutionary information can yield improved risk scores

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

Polygenic risk scores can be applied to ancient DNA



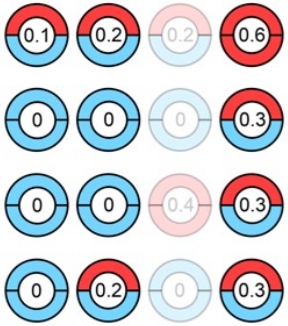
GWAS Catalog

The NHGRI-EBI Catalog of published genome-wide association studies

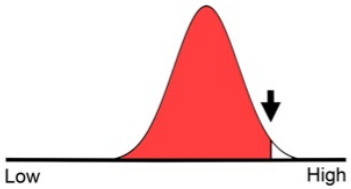
Curated set of ~3000 LD-pruned disease associations

Ancient genomes

Modern genomes

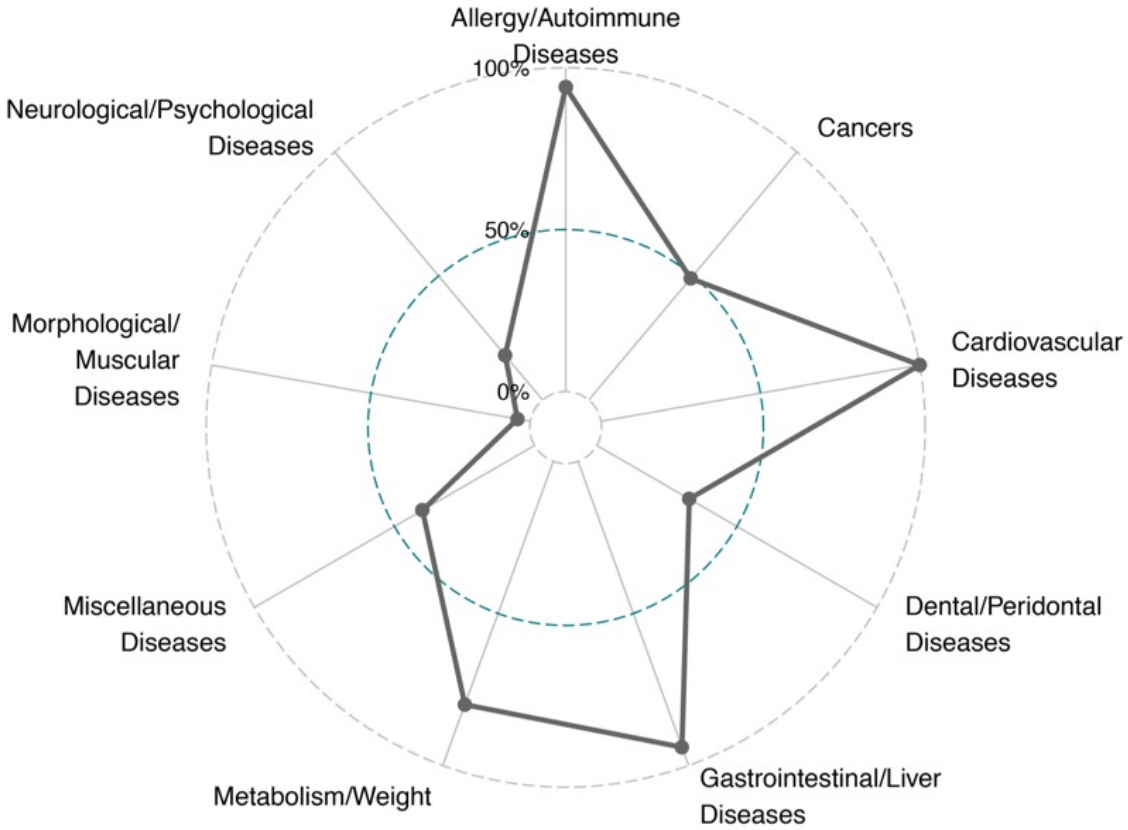


Modern PRS distribution



Polygenic risk score (PRS)

Ötzi the Tyrolean Iceman's genetic risk profile

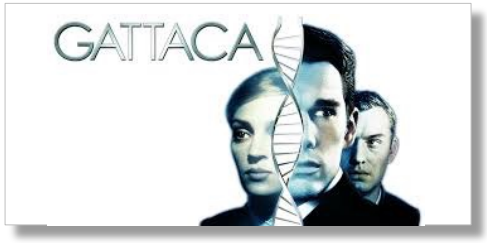


Kennis & Kennis Reconstructions

Berens, Cooper, and Lachance (*Human Biology*, 2017)

What will our genomes look like in the future?

- Genetic engineering



Columbia Pictures

- Changes in selection pressures

Disney · PIXAR

- Population admixture