Visualizing genetic data

Alex Diaz-Papkovich

September 2025

How can we visualize genetic diversity?

- *Homo sapiens* have a relatively small pool of DNA compared to other species
- ~99.9% of our DNA is identical; the remaining 0.1% can be interesting and useful
- Visualizations often focus on this remaining 0.1%
- Visualizations can convey interesting aspects of data quickly
  - Can also be poorly made, misused, hard to interpret, etc.

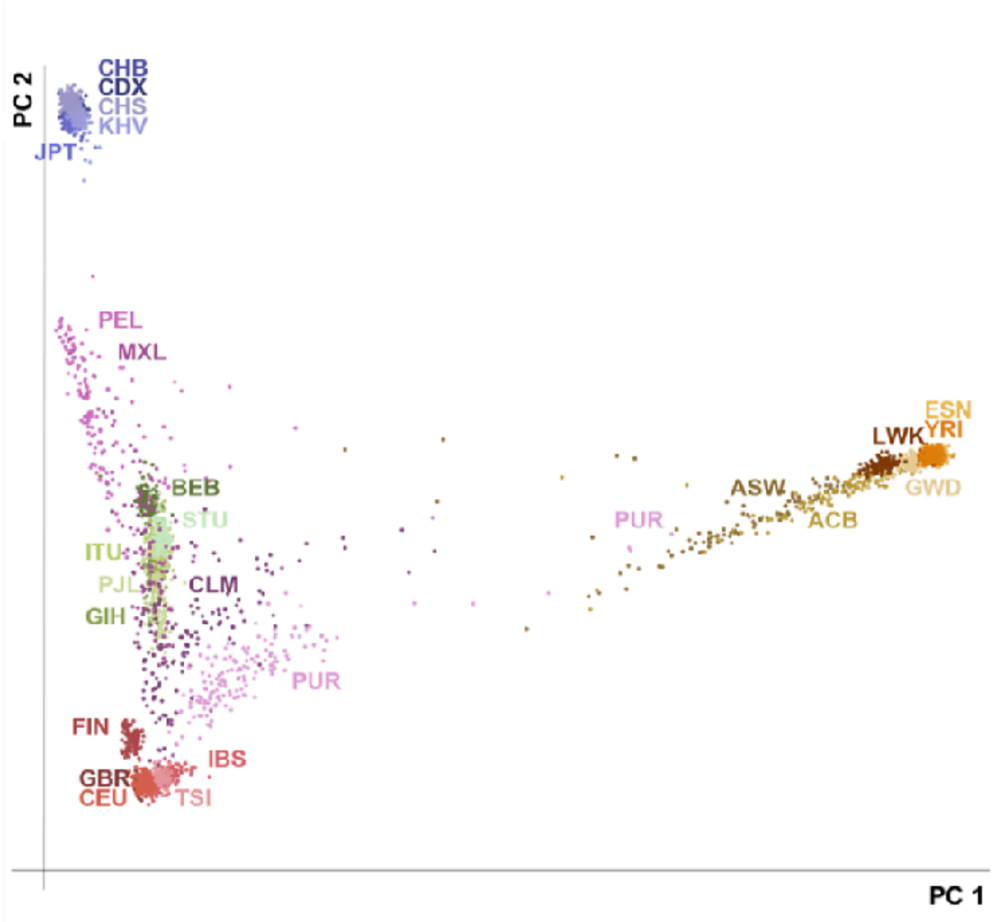
The genotype matrix

```
[0, 1, 0, 0, 0, 1, 0,..., 0, 0, 1, 1, 0, 0, 0, 0]
[2, 2, 2, 2, 2, 2, 2,..., 2, 2, 2, 2, 2, 2, 2, 2]
[0, 0, 0, 0, 0, 0, 0,..., 0, 0, 0, 0, 0, 0, 0, 0]
[1, 0, 0, 0, 0, 0, 0,..., 0, 0, 0, 0, 0, 0, 0, 0]
[2, 2, 2, 2, 2, 2, 2,..., 2, 2, 2, 2, 2, 2, 2, 1]
[0, 0, 0, 0, 0, 0, 0,..., 1, 0, 0, 0, 0, 1, 0, 0]
[1, 0, 0, 1, 2, 0, 1,..., 1, 1, 2, 2, 1, 2, 1, 1]
[2, 2, 2, 2, 2, 2, 2,..., 1, 2, 2, 2, 2, 2, 2, 1]
[1, 0, 0, 0, 0, 0, 0,..., 1, 0, 0, 0, 0, 0, 0, 0]
[0, 0, 0, 0, 0, 0, 0,..., 0, 0, 0, 0, 0, 1, 0, 0]
[0, 0, 0, 0, 0, 0, 0,..., 0, 0, 1, 0, 0, 0, 0, 0]
[0, 0, 0, 0, 1, 1, 0,..., 0, 0, 1, 1, 0, 1, 1, 0]
[2, 0, 2, 1, 0, 1, 0,..., 0, 1, 0, 0, 0, 1, 0, 0]
[...]
[0, 1, 0, 0, 0, 0, 1,..., 0, 0, 0, 2, 1, 1, 0, 1]
[2, 1, 1, 1, 2, 1, 1,..., 0, 2, 1, 0, 1, 1, 0, 1]
[0, 0, 0, 0, 0, 0, 0,..., 0, 0, 0, 0, 0, 0, 0, 0]
[0, 1, 0, 0, 1, 0, 1,..., 0, 0, 0, 1, 2, 1, 0, 1]
[0, 0, 0, 0, 0, 0, 0,..., 0, 1, 0, 0, 0, 0, 0, 0]
[2, 2, 2, 2, 2, 2, 2,..., 2, 2, 2, 2, 2, 2, 2, 2]
[1, 0, 1, 1, 1, 0, 0,..., 0, 1, 1, 0, 0, 1, 0, 1]
[1, 1, 1, 0, 1, 0, 0,..., 2, 1, 0, 2, 1, 0, 2, 1]
[0, 0, 1, 0, 0, 0, 0,..., 0, 0, 0, 0, 0, 0, 2, 0]
[0, 0, 0, 0, 0, 0, 0,..., 0, 0, 0, 0, 0, 0, 0, 0]
[0, 0, 1, 0, 0, 0, 0,..., 0, 0, 0, 0, 0, 0, 2, 0]
[2, 2, 1, 0, 1, 2, 2,..., 2, 1, 1, 2, 2, 2, 2, 1]
[1, 0, 0, 0, 0, 0, 0,..., 1, 0, 0, 0, 1, 0, 0, 0]
[0, 0, 0, 1, 0, 0, 0,..., 0, 0, 0, 0, 0, 0, 1, 0]
[0, 0, 0, 0, 0, 0, 0,..., 0, 0, 0, 0, 0, 0, 0, 0]
[0, 0, 0, 0, 0, 0, 0,..., 0, 0, 0, 0, 0, 0, 0, 0]
[0, 0, 0, 0, 0, 0, 0,..., 0, 0, 0, 0, 0, 0, 0, 0]
```

- Convert this to something visual
- Lots of methods exist
  - PCA, ADMIXTURE, CHROMOPAINTER, UMAP
  - Each plays to a different strength/aspect
- Interested in visualization that is useful, compelling, and responsible

Principal component analysis

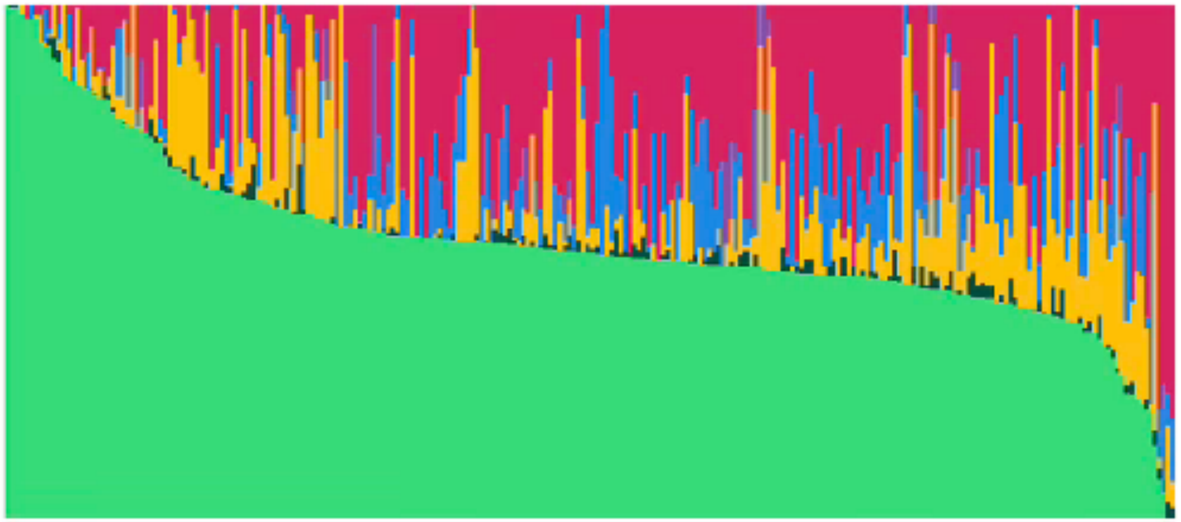
Explaining global variation in data



- Data fall along multiple axes of variation
- Here, reflects geographic distribution
  - Demographic history: Allelic drift over time
- As our species spreads continuously, we have gradual changes in genetic makeup
- Individuals with recent admixture fall “between” populations

Partial cluster membership (ADMIXTURE)

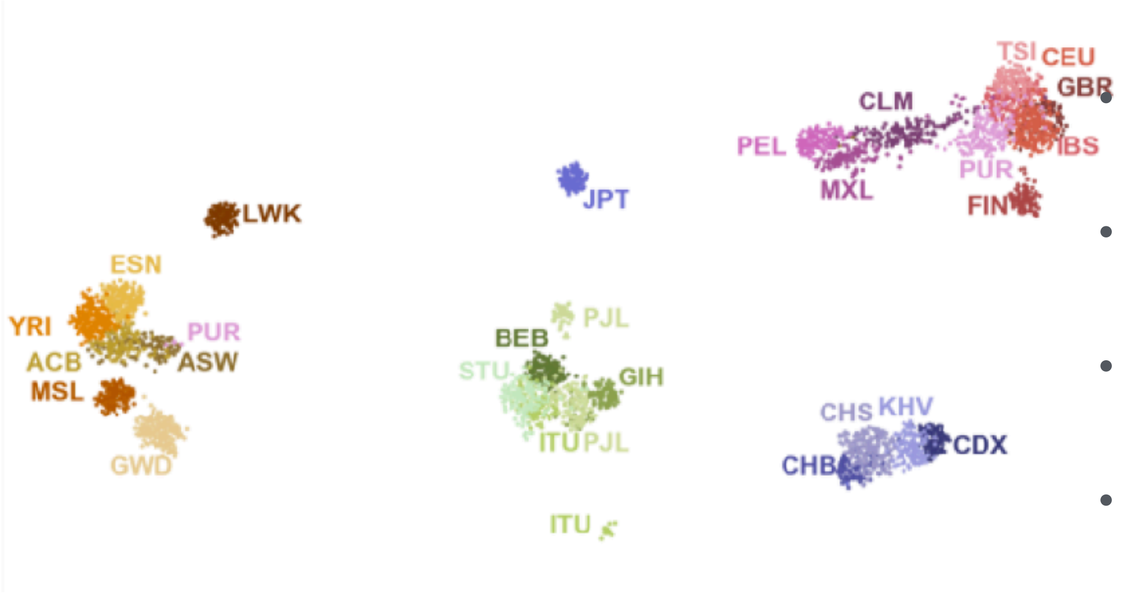
Inferring source population membership for each individual



- Assume *K* source populations
- Determine each individual's ancestry from each of the source populations
- Visualize as a barplot
  - Each column is a person
  - Each bar is a proportion

UMAP

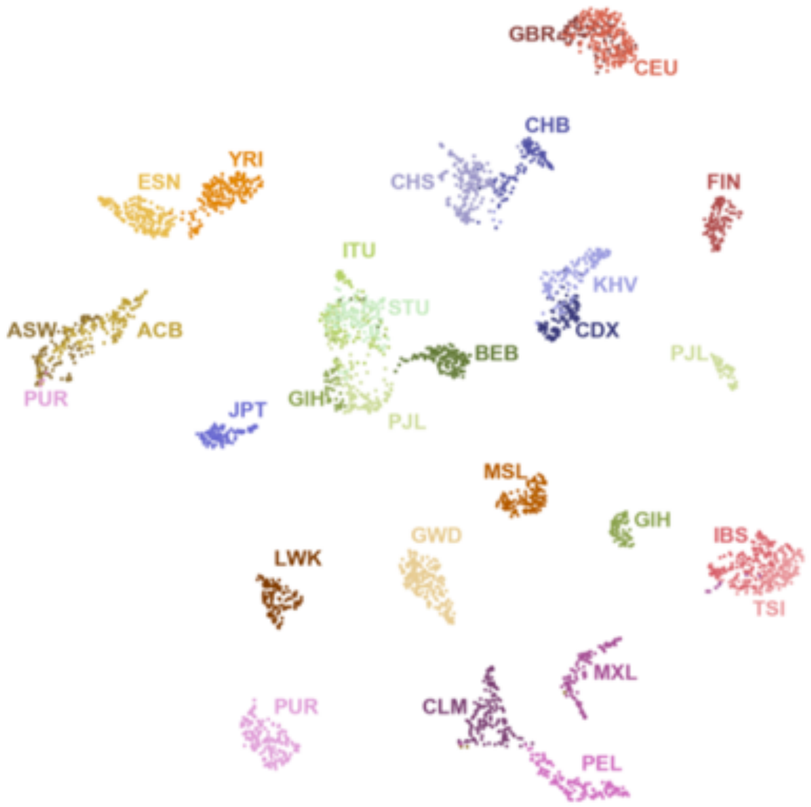
Visualizing local structure in data



- Represent local structure
- Choose output dimensionality (usually 2D)
- Data often forms clusters
- Long-range distances not meaningful
- Clustering implies shared demographic history

UMAP

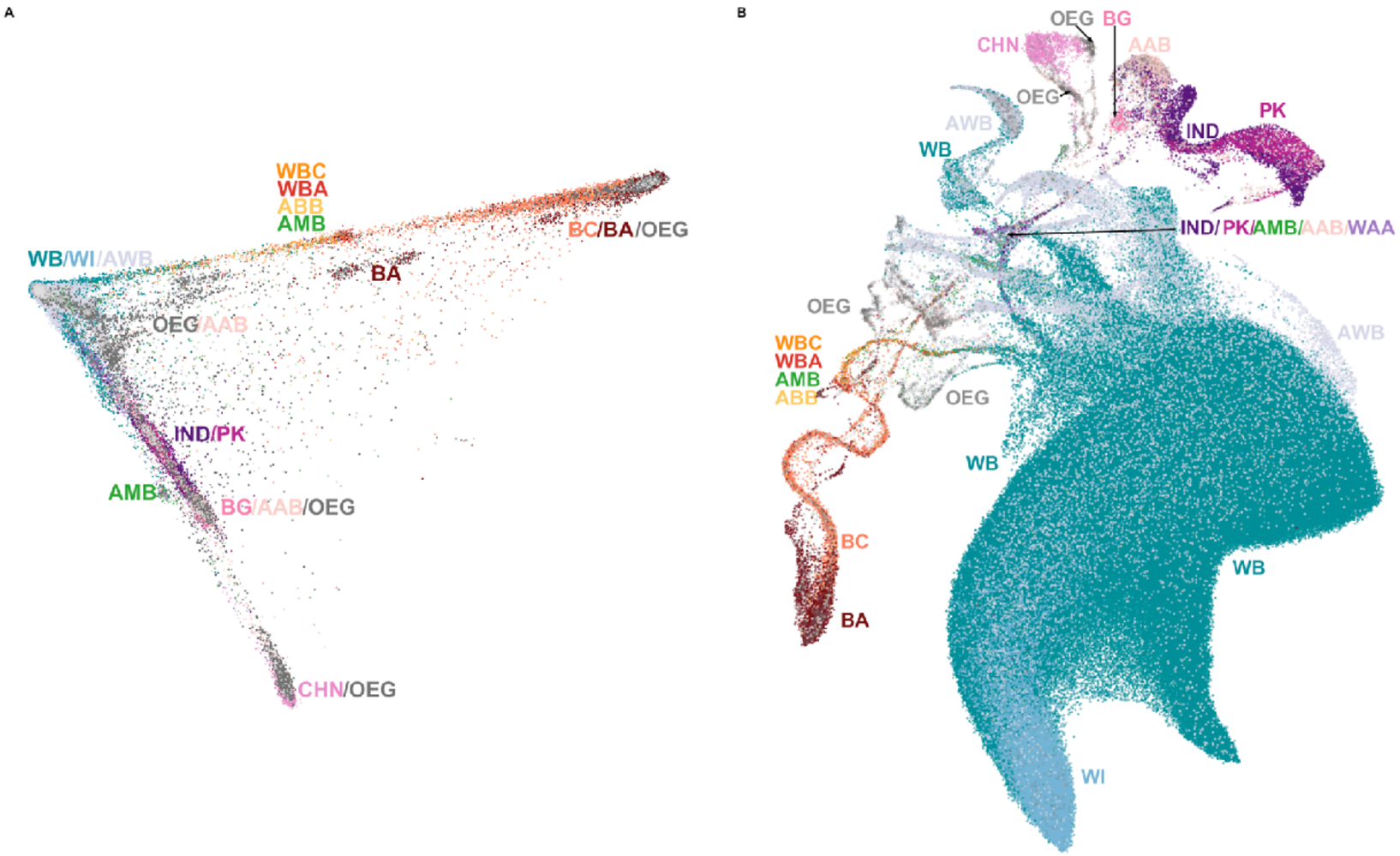
Visualizing local structure in data



- Can be pre-processed with PCA
- De-noise, create finer clusters
- Can select how many PCs to run through UMAP

PCA vs UMAP in biobanks

~500k individuals, comparing visual structure



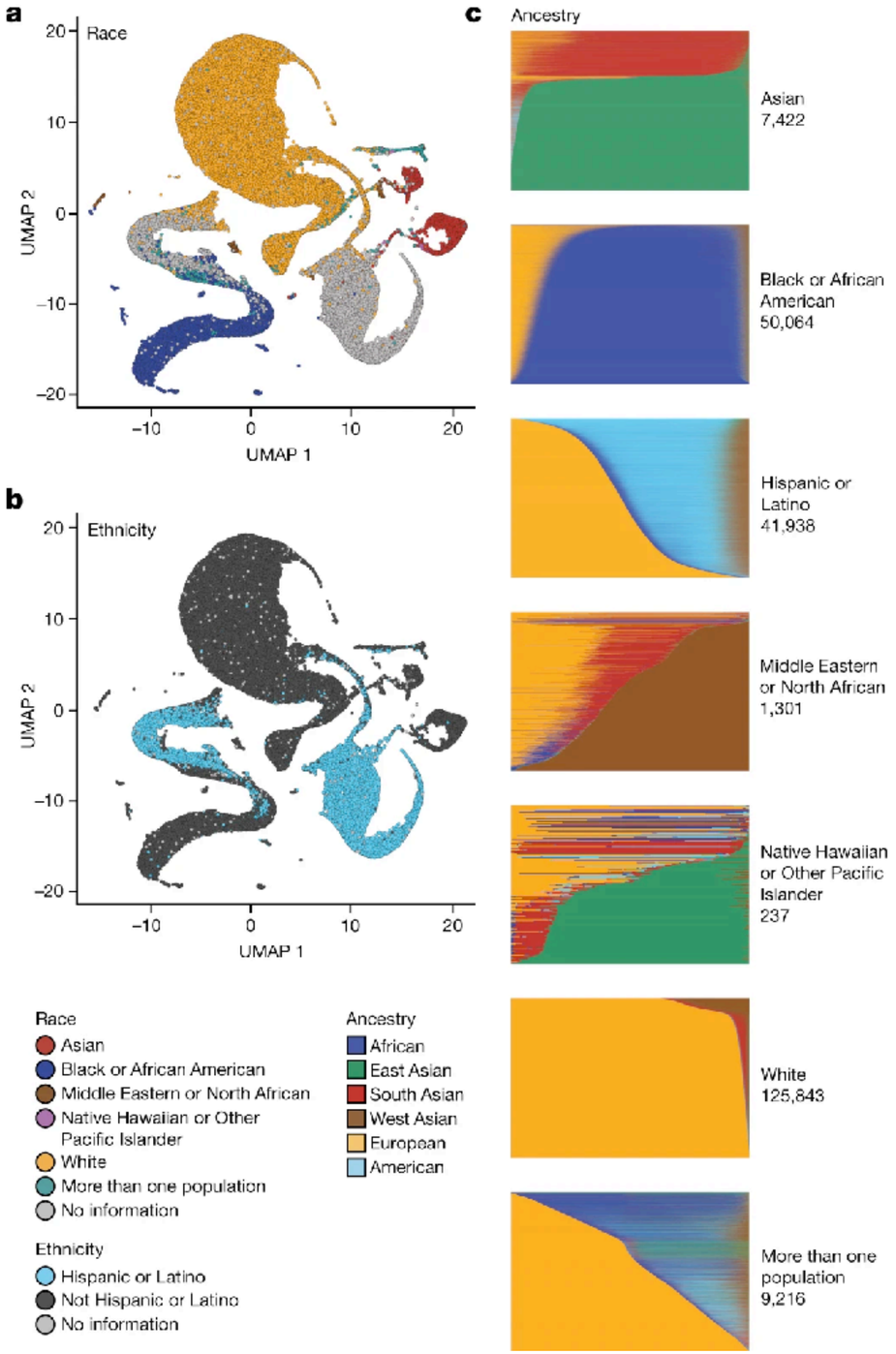
Problems

- Human DNA is almost fully identical, but focus is on differences
  - Emphasizes differences rather than (e.g.) shared underlying humanity
- Can be taken out-of-context, weaponized, misused, poorly-executed, etc.
- Long shadow of eugenics, race science, and other abuses

"Genomic data in the all of us research program." Nature 627.8003 (2024): 340-346.

- Figure criticized for implying race, ethnicity, ancestry, genetics are all one thing

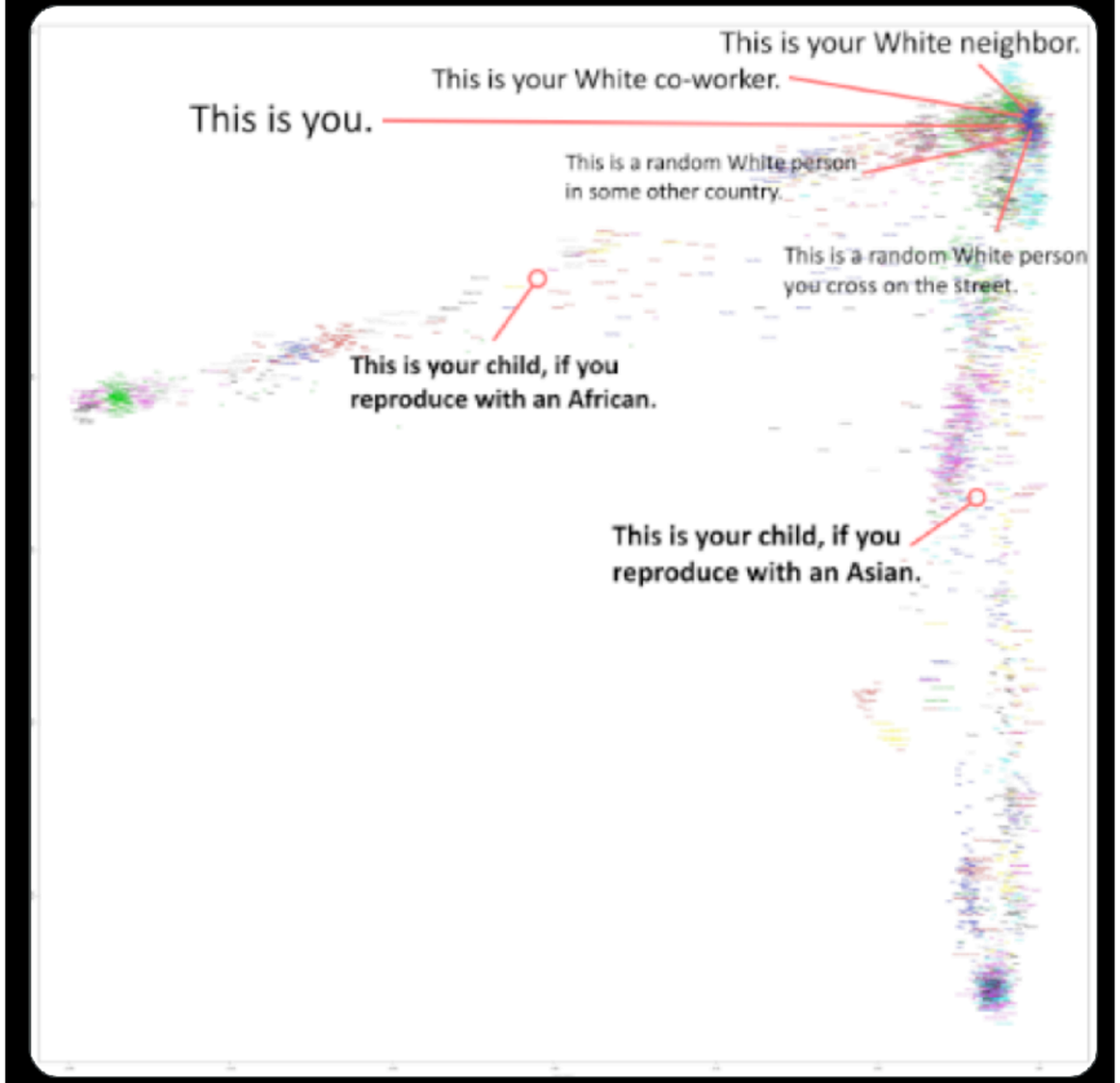
- https://www.science.org/content/article/huge-genome-study-confronted-concerns-over-race-analysis

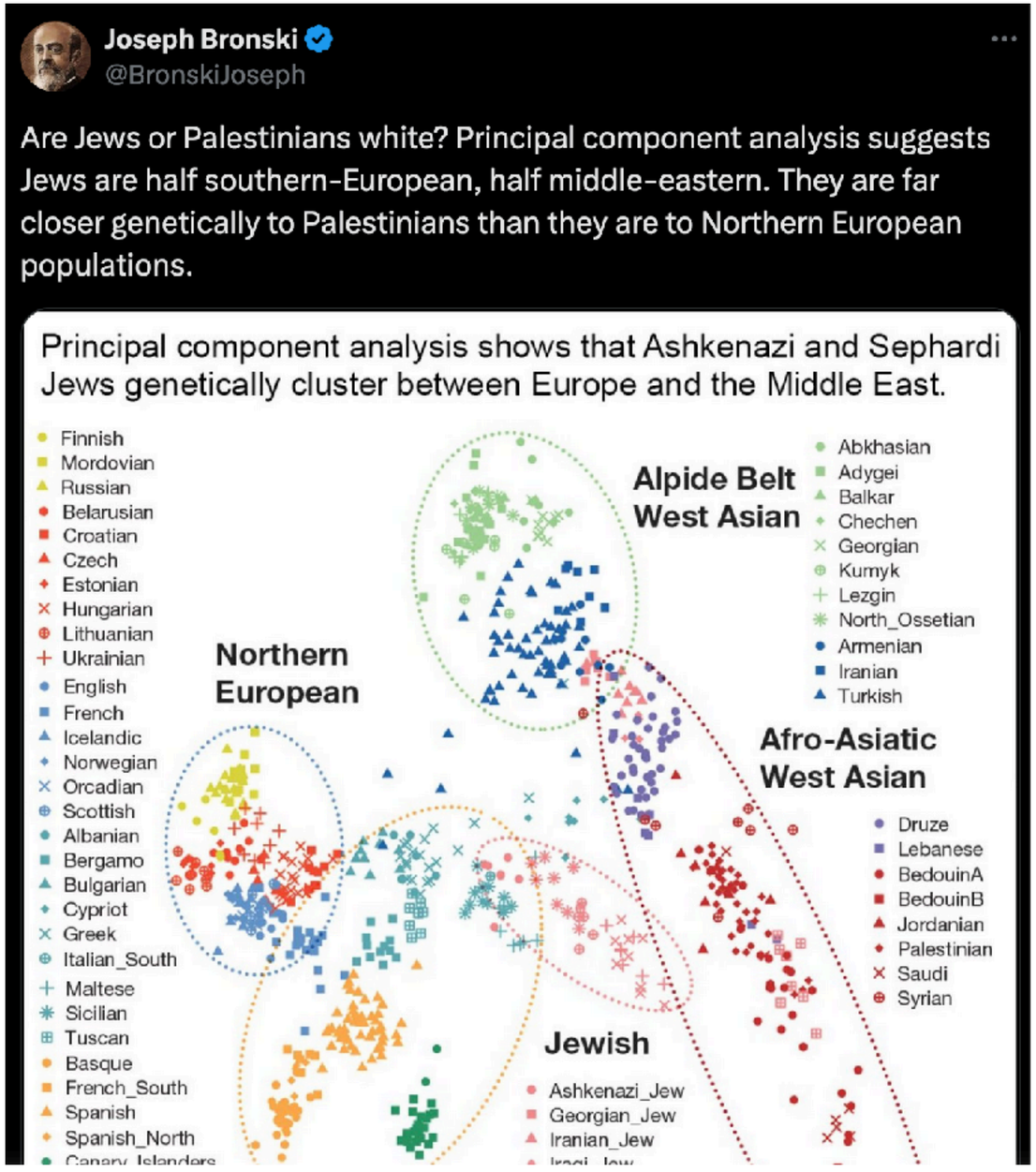


Jean-François Gariépy
@JFGariepy

Most people don't realize that interbreeding with other races means that strangers on the street from your own race are more related to you than your own child. Amplify this, @elonmusk and spread the knowledge!

Traduire le post




Joseph Bronski
@BronskiJoseph
Are Jews or Palestinians white? Principal component analysis suggests Jews are half southern-European, half middle-eastern. They are far closer genetically to Palestinians than they are to Northern European populations.
Principal component analysis shows that Ashkenazi and Sephardi Jews genetically cluster between Europe and the Middle East.
Finnish
Mordovian
Russian
Belarusian
Croatian
Czech
Estonian
Hungarian
Lithuanian
Ukrainian
English
French
Icelandic
Norwegian
Orcadian
Scottish
Albanian
Bergamo
Bulgarian
Cypriot
Greek
Italian_South
Maltese
Sicilian
Tuscan
Basque
French_South
Spanish
Spanish_North
Canary_Islanders
Northern European
Alpide Belt West Asian
Abkhasian
Adygei
Balkar
Chechen
Georgian
Kumyk
Lezgin
North_Ossetian
Armenian
Iranian
Turkish
Afro-Asiatic West Asian
Druze
Lebanese
BedouinA
BedouinB
Jordanian
Palestinian
Saudi
Syrian
Jewish
Ashkenazi_Jew
Georgian_Jew
Iranian_Jew

Further reading

- My papers (on UMAP/density clustering in population genetics):
  - Diaz-Papkovich 2019 (PLoS Genetics), 2020 (Journal of Human Genetics), 2023 (bioRxiv)
- Carlson, Jedidiah, et al. "Counter the weaponization of genetics research by extremists." Nature 610.7932 (2022): 444-447.
- Lewis, Anna CF, et al. "Getting genetic ancestry right for science and society." Science 376.6590 (2022): 250-252.