3.1 Population structure: I. Ancestry estimation

In which we discuss how genetic data are used to study the population structure, ancestry, and mixing of modern human populations. Specifically we ask: ***How can we use genotype data from a collection of individuals to learn about population structure?*** *We close with a discussion of the relationship between ancestry and cultural concepts of race* [364].

Figure 3.1: **"I come from the Black Lagoon, but it turns out I'm 14 percent Atlantic dolphin."**

Cartoon copyright 2019 Robert Leighton [Link]. Used with permission.

In Chapter 2.4 we showed how barriers to random mating–most notably due to the geography of where we live and who we reproduce with–allow allele frequencies to drift apart. This leads to **population structure**: i.e., differences in allele frequencies or haplotype frequencies among different groups or populations within a species. A closely related concept, **ancestry**, refers to how much of a person's genome comes from each of a set of defined groups or populations.

In this chapter we'll describe how genetic data is used to study population structure, and to estimate the ancestry of individuals. The key analysis methods here will have a different feel from approaches we've used so far. In particular, we'll make heavy use of **clustering techniques**. In statistical data analysis, "clustering" refers to a variety of methods for grouping objects that share similarities – in this case, for grouping individuals with similar genotypes.

While these approaches are extremely useful for understanding human variation, **we should always remember that they are tools to model a much more complex reality** [a]. Human populations are structured at a range of geographic scales from the level of continents down to, in some cases, differences between nearby towns or even extended families. As we'll see, different analyses focus on ancestry at different scales.

Moreover, **population structure is not fixed over time**. We are familiar with the idea that modern societies are often melting pots of peoples from many different places; but as we will discuss, ancient populations were also highly mobile, and population mixing has been a constant force in human genetic history.

Another key point is that each of us has ancestors from many locations and likely even from all continents [365]. As I will explain, estimated genetic ancestry actually corresponds to complicated statements about which groups of individuals have similar mixtures of ancestors.

For all these reasons, there is no single "correct" description of structure. As you read the next three chapters, pay attention to how different methods, and different data sets, tell us about structure and ancestry at a range of geographic and temporal scales; consider also what aspects of history may be entirely invisible to us with any given data set.

Lastly, at the end of the chapter we will touch on another complicated issue: namely, the relationship between ancestry and the socially-defined concepts of race and ethnicity.

[a] *"All models are wrong but some models are useful"–George Box. As we shall see, the twin concepts of population structure and ancestry are powerful tools for studying human history and the genetic basis of disease. But you should remember that what we call 'populations' are models to approximate a complex tangled web of past migrations and mixtures of ancestors, and nonrandom mating at different geographic scales.*

The concept of clustering. In data analysis we often want to sort things into groups of similar objects. This is called **clustering**. For example, consider the following items in a grocery store:

ice cream, wine, tea, apples, oranges, onions, celery, sausages, steaks, frozen peas, beer, milk, chocolates, coffee.

How might we cluster these into groups? Perhaps by food category:
{beer, wine, tea, coffee, milk}, {apples, oranges, onions, celery, frozen peas}, {sausages, steaks}, {ice cream, chocolates}.

Maybe you think some of these clusters are too broad – you might prefer to split the drinks more narrowly:
{beer, wine}, {tea, coffee}, {milk}.

Or maybe you prefer to group by aisle in the store:
{beer, wine}, {tea, coffee}, {apples, oranges, onions, celery}, {milk, sausages, steaks}, {ice cream, frozen peas}, {chocolates}

This example shows how clusters can capture meaningful structure in the data. And yet you may be bothered that there's no obviously "correct" clustering: reasonable observers may differ on how finely to define the clusters (should beer and tea be in the same cluster?) or on which features to emphasize (are frozen peas more like onions or ice cream?) [b].

Our next example shows how we can do a bit better by allowing objects to have membership in multiple clusters.

Suppose we want to cluster Wikipedia pages into topics on the basis of the words they use. How should we do this? To illustrate this I picked out interesting words from 7 different Wikipedia pages. The colors highlight words with similar themes:

Jennifer Doudna	Sewall Wright	CRISPR gene editing	David Beckham	sickle cell disease	Manchester United	population genetics
born	genetics	CRISPR	football	haemoglobin	football	genetics
school	evolution	gene	Manchester	blood	club	selection
patent	born	DNA	born	genetics	England	drift
mother	died	clinical	career	mutation	title	Wright
edit	population	engineering	goal	malaria	Beckham	synthesis
CRISPR	selection	nuclease	wife	selection	goal	evolution
genetic	statistical	therapy	England	allele	record	mutation
disease	quantitative	Nobel	scored	CRISPR	Ferguson	quantitative
Nobel	enzymes	patent	corner	therapy	Premier	statistical

How should we cluster these 7 Wikipedia pages?

Some of the same issues we saw in the grocery store example come up again. Is the Jennifer Doudna page more related to the other biographies (Wright and Beckham) or to CRISPR editing, which she developed?

To solve this, we introduce a concept called a **topic** [366]. Each topic has a characteristic set of word frequencies: for example we could define

[b] *Much the same is true for genetic data: even when there is clear evidence for population structure, the precise clustering output depends on the choice of methods and samples.*

a topic named *biography* (blue), that uses words like *born*, *died*, *school*, *mother*, *career* at higher frequency.

Crucially, **we allow each page to have fractional membership in multiple topics.** We might estimate that the Doudna page is 30% *biography*, 50% *genome editing*, and 20% *genetics*. In a formal model, this would mean that 30% of the words in the Doudna page come from the word distribution for *biography* and similarly for the other topics.

Similarly, each topic contributes to multiple pages: here, *evolution* (olive-color) contributes to at least three of the pages: Sewall Wright, Sickle Cell, and Population Genetics.

In this chapter we use similar ideas to model population structure. *In this analogy, the Wikipedia pages are like people's genomes, and the topics are like different populations. Someone's ancestry is defined by the mixture of different "topics" that are represented in their genome. We'll explore a variety of techniques for understanding ancestry.*

Genotype clustering. Suppose we collect genotype data for a sample of individuals, using SNPs spread across the genome.

First, it's convenient to code the genotype data as integers. As elsewhere in the book, we'll code each possible genotype at a single SNP as the number of alternate or derived alleles, with possible values of 0, 1, or 2:

$$\begin{aligned} AA &\rightarrow 0 \\ Aa &\rightarrow 1 \\ aa &\rightarrow 2 \end{aligned}$$

Then we can record a genotype data set in a matrix G, like this:

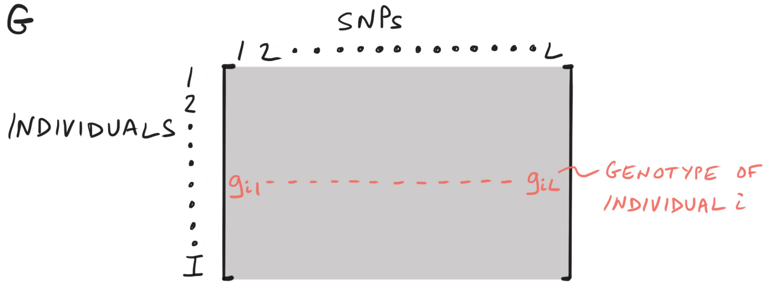


Figure 3.2: **The genotype matrix, G.** *Each entry, $g_{i,l}$, contains the number of derived alleles carried by individual i at SNP l. Each row gives the full genotype for a single individual.*

where the rows (indexed by i) represent individuals and the columns (indexed by l for locus) represent the SNPs. G_i denotes a single row of G: that is, the genotype for individual i.

Our goal in this chapter is to use G to study population structure.

Before we go on, take a moment to look at a hypothetical example of G, below. Can you cluster the individuals into sensible groups by eye? If so, what criteria did you use?

$$\begin{array}{c|cccccccccc}
 & \multicolumn{10}{c}{\text{SNPs}} \\
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
\hline
\text{INDIVIDUALS} \; 1 & 0 & 0 & 2 & 2 & 1 & 0 & 0 & 0 & 2 & 2 \\
2 & 2 & 1 & 1 & 0 & 1 & 2 & 1 & 2 & 1 & 1 \\
3 & 1 & 2 & 0 & 1 & 0 & 2 & 2 & 1 & 0 & 0 \\
4 & 0 & 1 & 2 & 2 & 2 & 0 & 1 & 0 & 2 & 2 \\
5 & 1 & 2 & 0 & 0 & 0 & 1 & 2 & 1 & 1 & 0 \\
6 & 1 & 0 & 2 & 1 & 2 & 0 & 0 & 1 & 1 & 2 \\
7 & 0 & 0 & 1 & 2 & 2 & 1 & 0 & 0 & 2 & 1 \\
8 & 2 & 2 & 0 & 0 & 1 & 2 & 0 & 2 & 0 & 1
\end{array}$$

Figure 3.3: **Can you find structure in this genotype matrix?** *We'll return to this example later.*

Early work on genotype clustering. The first studies of individual ancestry came in the 1990s, with the discovery that a simple measure of genetic distance can cluster individuals by continent [367].

One of the first examples came in a 1997 paper by Joanna Mountain and Luca Cavalli-Sforza [368]. They analyzed SNP data from 144 humans currently living in different parts of the world, at about 100 SNPs per person [369]. They computed genotype distances between each pair of individuals, like this:

$$\begin{array}{l|cccccccccc}
 & \multicolumn{10}{c}{\text{SNPs}} \\
\text{GENOTYPE 1} & 0 & 0 & 2 & 1 & 1 & 2 & 0 & 1 & 2 & 1 \\
\text{GENOTYPE 2} & 1 & 2 & 2 & 1 & 1 & 1 & 2 & 2 & 2 & 1 \\
\hline
\frac{\sum |\text{DIFFERENCE}|}{L} & \multicolumn{10}{c}{1+2+0+0+0+1+2+1+0+0 \; /_{10} = 0.7}
\end{array}$$

Figure 3.4: **A simple multi-SNP genotype distance.** *Here the two rows show the genotypes for two different individuals at L SNPs. The genotype distance between the two individuals is computed as the sum of the absolute genotype difference at each SNP, divided by L.*

Next, they used a statistical method called **hierarchical clustering** to fit the pairwise distances between the 144 individuals. This results in *a graph that clusters pairs of individuals with shorter genetic distances closer together*, and pairs of individuals with greater genetic distances further apart. The total line distance between any two individuals in the graph is roughly proportional to the genetic distance between them:

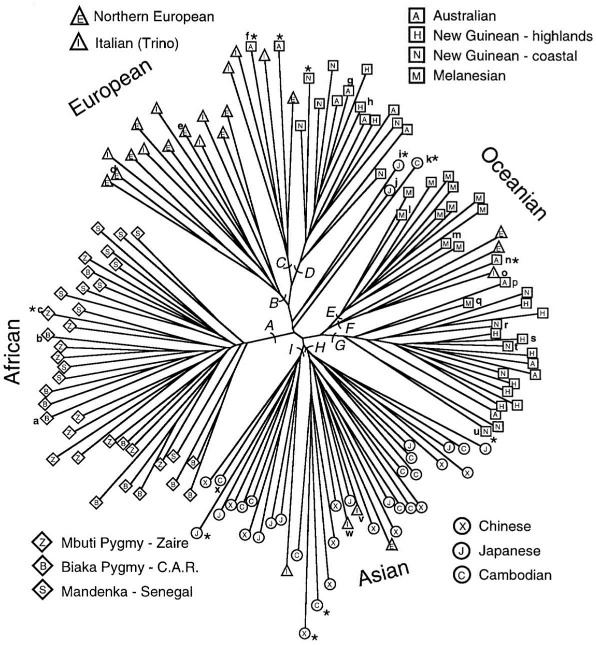


Figure 3.5: **Hierarchical clustering of individuals based on genotype distances (1997).** *Each individual is represented by a tip on the graph. Line distances are roughly proportional to genetic distances. Key:*

- ◇ *African*
- △ *European*
- □ *Oceanian*
- ○ *East Asian*

All of the individuals here are from native populations and not from recent immigrant populations. Credit: Figure 5 from Mountain and Cavalli-Sforza (1997). [Link] Used with permission.

As you see above, this produced the remarkable result that, using a small number of SNPs, individuals cluster according to continent.

This graph represents something quite different from the coalescent trees we talked about in previous chapters: here **the clustering represents the overall *similarity* of multi-SNP genotypes**, and it averages information over many independent SNPs. You should also notice some key points:

- Individuals from the same continent tend to appear together on the tree, reflecting genotype similarity to one another.
- And yet, differences between continents are quantitatively small: **individuals from the same population are only slightly more similar than individuals from different continents** – see figure on the right [370].
- **No one SNP on its own would be enough to distinguish the different continental groupings**; instead the signal comes from averaging weak information across many SNPs.

As you see here, this early type of model was already quite informative, but in the next few sections we'll see how more advanced models, combined with modern data, can provide far more information.

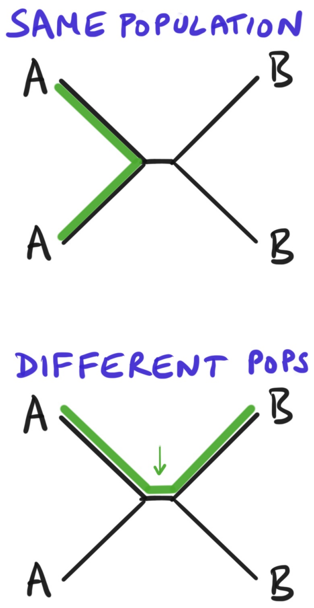


Figure 3.6: **Genetic distances within versus between populations.** *Distances between two individuals from the same population (green path between two individuals from population A) are only slightly shorter than between two individuals from different populations (A to B).* ***The branch with the green arrow shows the small extra distance between populations.***

A genotype model with population structure. As in other parts of the book, we'll find it's helpful to write down a simple model to understand the data.

First we need a concept of **populations**: in our **idealized model** we assume that populations are discrete groups of random-mating individuals. *We assume that, within populations, SNPs are in Hardy Weinberg equilibrium and pairs of SNPs that are far apart in the genome are in linkage equilibrium.* Different populations will generally have different allele frequencies.

We need to define some **notation**.

Assume that there are K different populations, and we measure the genotypes for a panel of individuals at L SNPs. We record the **allele frequencies** in a matrix P, where $p_{k,l}$ to represents the derived allele frequency at the lth SNP in population k. P_l will be a column of P corresponding to the frequencies of SNP l in each of the populations:

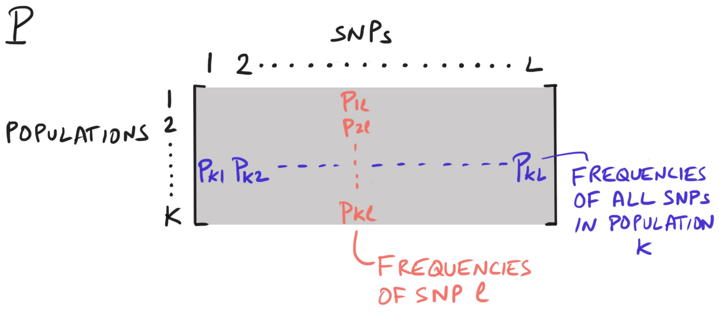


Figure 3.7: **The allele frequency matrix, P.** *Each entry, $p_{k,l}$, contains the allele frequency of SNP l in population k.*

Next, the matrix Q contains the **ancestry** of each individual. Here, $q_{i,k}$ represents the fraction of ancestry that individual i has from population k. By analogy with our Wikipedia example, you can think of i as indexing a single page, and $q_{i,k}$ saying what fraction of that page is generated from topic k. I'll give another example of this shortly.

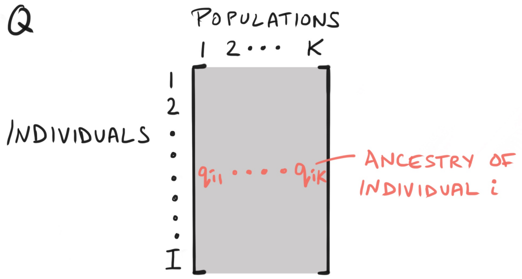


Figure 3.8: **The allele frequency matrix, Q.** *Each entry, $q_{i,k}$, contains the ancestry of individual i in population k. For example if an individual is from Population 2, then their row would read* $(0, 1, 0, \cdots)$.

We'll use Q_i to indicate a single row of Q – the ancestry vector for individual i.

A generative model [371]**.** Now consider an individual i. Suppose I tell you their ancestry, Q_i, and the allele frequencies, P. How do we model their genotype G_i?

To start, we'll assume that an individual's genome comes entirely from a single population – this is known as a **no-admixture** model. Suppose that individual i comes from population k. Then Q_i will have a single entry of 1 in population k, and 0 for the other populations: e.g., $\{0, 0, 0, 1, 0\}$.

The genotype at SNP l results from a random sampling process like this:

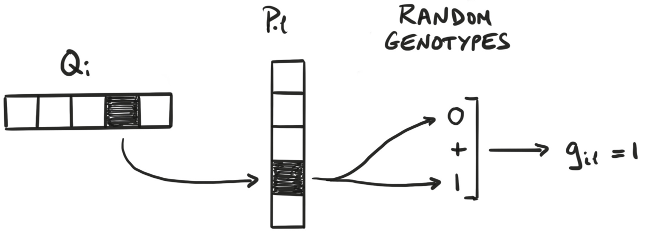


Figure 3.9: **Genotype sampling in the no-admixture model.** *Q_i tells us that individual i is from population k (here $k = 4$). We then look up the relevant allele frequency at SNP l for population k. Then the genotype $g_{i,l}$ is the sum of two random alleles, each sampled given the allele frequency $p_{k,l}$.*

More formally, the genotype probabilities for SNP l are:

$$\begin{aligned} \Pr(g_{i,l} = 0) &= (1 - p_{k,l})^2 \\ \Pr(g_{i,l} = 1) &= 2p_{k,l}(1 - p_{k,l}) \\ \Pr(g_{i,l} = 2) &= p_{k,l}^2 \end{aligned} \tag{3.1}$$

where $p_{k,l}$ is the relevant allele frequency for the population and SNP. These are simply the Hardy Weinberg proportions for population k.

We can compute the probability of individual i's entire genotype, given their ancestry Q_i, by multiplying these probabilities across all L SNPs [c]:

$$\Pr(G_i|Q_i) = \prod_{l=1}^{L} \Pr(g_{i,l}|Q_i) \tag{3.2}$$

This is known as the **likelihood**. It's an important principle of statistical theory that we should generally prefer models that produce the highest likelihoods.

[c] *The vertical bar notation | is math-talk for 'given that'.* $\Pr(G_i|k)$ *can be read as 'the probability of observing genotype G_i given that individual i comes from population k'.*

The admixture model. The no-admixture model is very limiting because it can't handle the fact that many of us have recent ancestry from multiple populations.

As in our Wikipedia example, we can make a much better model by allowing people to have fractional ancestry in each population. For example, suppose that $K = 5$, and a person has 1/4 of their ancestry from population 1 and 3/4 from population 3, then $q_i = (0.25,\ 0,\ 0.75, 0, 0, 0)$.

This is known as an **admixture model**.

The way we interpret this is that each SNP allele now has to make two "decisions": what population does it come from, and then what allele is it?

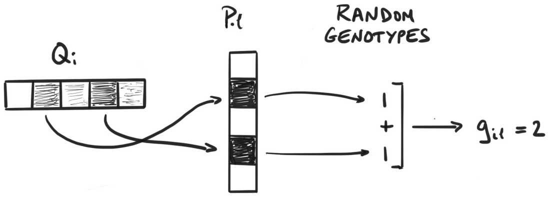


Figure 3.10: **Genotype sampling in the admixture model.** *In this model, an individual has fractional ancestry in each population. To simulate a genotype, we would first pick the population-of-origin for each allele at random according to the ancestry probabilities in Q_i. These tell us where to look in $P_{\cdot,l}$ for the relevant frequencies for each allele (populations 2 and 4 in this example). Genotype $g_{i,l}$ is then the sum of two random alleles, each sampled according to the appropriate population frequency.*

We can think of individual i as having a "personal" allele frequency for SNP l that depends on their ancestry proportions. This personal allele frequency is a weighted average of the allele frequencies across populations. In vector notation it's the dot product of Q_i and P_l:

$$\begin{aligned}\text{Personal Allele Freq}(i, l) &= q_{i,1}p_{l,1} + q_{i,2}p_{l,2} + \cdots \\ &= Q_i \cdot P_l \end{aligned} \tag{3.3}$$

Then the genotype probabilities for the admixture model are like those in Equation 3.1 but using the personal allele frequency instead of $p_{k,l}$ [372]. In the basic model we assume that we can treat the information from each SNP independently [373].

Estimating P. So far we've been assuming that P and Q are known. How can we estimate these?

One standard approach is to start with labeled samples – for example, samples of individuals like the YRI (Yoruba) or BEB (Bengali) from the 1000 Genomes Project [Link]. If we are willing to assume these are representative of "true" populations then we can use them to estimate population-specific allele frequencies:

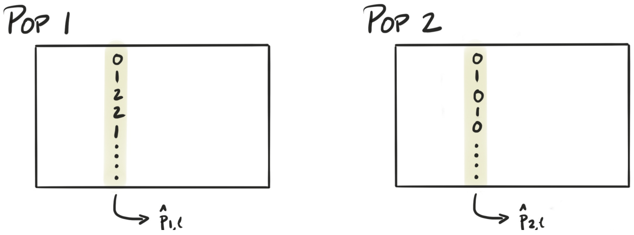


Figure 3.11: **Estimating allele frequencies.** *If we have labeled population samples then we can estimate population frequencies simply from the column averages of the genotype data (dividing by 2 for allele frequencies rather than genotypes). Here $\hat{p}_{1,l}$ is the estimated frequency in population 1 at SNP l.*

As a technical point, people often add so-called pseudocounts to the allele counts data to avoid frequency estimates of zero as these can cause problems in the downstream analysis; see [374].

Estimating Q: Ancestry Estimation. Next, if we have allele frequency estimates P for reference populations, we can estimate ancestry of new unknown individuals.

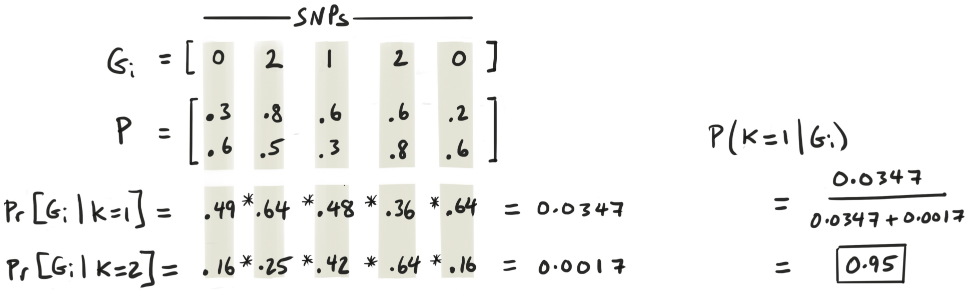
Suppose that I tell you somebody's genotype G_i. How can you infer their ancestry (starting with the **no-admixture model**)? We can estimate their ancestry using a method called Bayes' Rule of conditional probability [375].

If all possible source populations are equally likely, then the probability that the individual comes from population k is simply proportional to the probability of observing their genotype in population k: [d]

$$\Pr(\text{individual } i \text{ from population } k) = \frac{\Pr(G_i|k)}{\sum_{j=1}^{K} Pr(G_i|j)} \tag{3.4}$$

The numerator is the likelihood of observing their genotype in population k, while the denominator is a sum of the likelihoods over all populations 1 to K. Remember that Equation 3.2 shows how to compute this likelihood.

Here's how this works in a toy example assuming we know P:



$$G_i = \begin{bmatrix} 0 & 2 & 1 & 2 & 0 \end{bmatrix}$$

$$P = \begin{bmatrix} .3 & .8 & .6 & .6 & .2 \\ .6 & .5 & .3 & .8 & .6 \end{bmatrix}$$

$$\Pr[G_i|k=1] = .49 * .64 * .48 * .36 * .64 = 0.0347$$

$$\Pr[G_i|k=2] = .16 * .25 * .42 * .64 * .16 = 0.0017$$

$$P(k=1|G_i) = \frac{0.0347}{0.0347+0.0017} = \boxed{0.95}$$

Figure 3.12: **Population assignment of a single individual.** *Here* $\mathbf{G_i}$ *shows the genotype for an individual at 5 SNPs and* **P** *shows the allele frequencies for each SNP in two different populations.*
The rows at the bottom show likelihood calculations assuming either k=1 or 2 (Eqs 3.1, 3.2). The likelihood of any specific genotype is always small, but the key point is that Population 1 has much higher likelihood than Population 2.
The calculation to the right gives the posterior probability this individual is from Population 1 (Eq. 3.4).

The **admixture model** is conceptually similar: we simply find the ancestry vector Q_i that maximizes the individual's genotype data G_i given the population allele frequencies (but the math is a bit more complicated) [376].

Genetic clustering: the Structure/Admixture model.

Assignment tests like this are fine if you're willing to assume that you know how to define the relevant populations in advance: for example based on sampling location, self-defined group, or language.

But how would we know that pre-defined populations actually reflect underlying structure in a sensible way? To address this limitation, a paper from 2000 by Pritchard, Stephens, and Donnelly asked if we can infer structure directly from the genotype matrix G [e] [377]

Given the genotype matrix G we want to estimate both the allele frequencies, P and ancestry of individuals, Q.

Some intuition. Where does the information come from to allow us to identify populations? *One way to think about this is that within populations*

[d] *The most probable source for the individual is whichever population makes their observed genotype most likely.*

[e] *This type of approach has been used widely for studying the population structure of humans and other species. The original algorithm, named* **Structure**, *is too slow for genome-scale data and modern applications use a faster method called* **Admixture**.

we can expect alleles to be independent: that is, individual SNPs should be in Hardy-Weinberg proportions, and unlinked SNPs should not be in LD.

But if we have a collection of individuals with different ancestries, this creates correlations between the genotypes even at unlinked SNPs. Here's the hypothetical example I showed earlier. Notice how the coloring helps us to explain the correlation structure across SNPs:

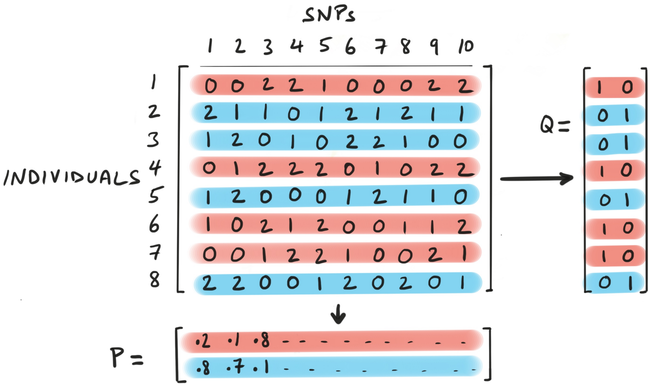


Figure 3.13: **Inference of structure in a toy example with K = 2 and no admixture.** *Populations 1 and 2 are colored in red and blue respectively.*
The **Q** *matrix shows population assignments for each individual (the columns show fraction of ancestry in each population).*
The **P** *matrix shows possible (true) allele frequencies that could produce the observed data (the rows show allele frequencies for each population). You may like to suggest allele frequencies for SNPs 4–10 as an exercise.*

Of course in practice, we can't infer the relevant populations by eye. We won't go into the algorithms in full detail, but the overall goal is to find estimates of P and Q that maximize the likelihood of the genotype data G. Roughly speaking, **you can think of the Structure algorithm as searching for a partition of individuals into K populations (with admixture) that minimize Hardy-Weinberg and linkage disequilibrium within populations.**

In short, the approach is to start from a random guess about Q and then iterate toward the true estimates roughly as follows:

- *Initialization:* Set the ancestry for each individual, Q_i, at random (typically from a uniform distribution).
- *repeat*
  {
  - *Update P given Q:* For each SNP, update the allele frequency estimates according to the allele counts in each population.
  - *Update Q given P:* For each individual, update their ancestry estimate Q_i given the allele frequencies.

  } until converged.

Somewhat magically, this type of algorithm converges to sensible solutions for P and Q. In *Structure* this is implemented as a fully Bayesian model using Markov chain Monte Carlo; *Admixture* frames it as a maximum likelihood problem and performs optimization using an Expectation-Maximization algorithm.

One last challenging question is how to choose the number of clusters (K). As in the grocery store example at the beginning of the chapter, increasing K usually corresponds to adding increasingly finer subdivisions, and there is usually not a single "best" choice of K [378]. Instead, different values of K reveal different aspects of the structure, as we'll see below.

Structure in practice. The first large-scale application of the Structure algorithm was by Rosenberg, Pritchard, and others in 2002 [379], using a worldwide sample of 51 populations from the Human Genome Diversity Panel (HGDP) [380]. The plot below is from an updated version of this data set published by Jun Li and colleagues in 2008 [381].

In this plot, each cluster (population) is given a different color. Each individual is represented as a thin vertical line, and the colors indicate proportional membership in each of the 5 clusters:

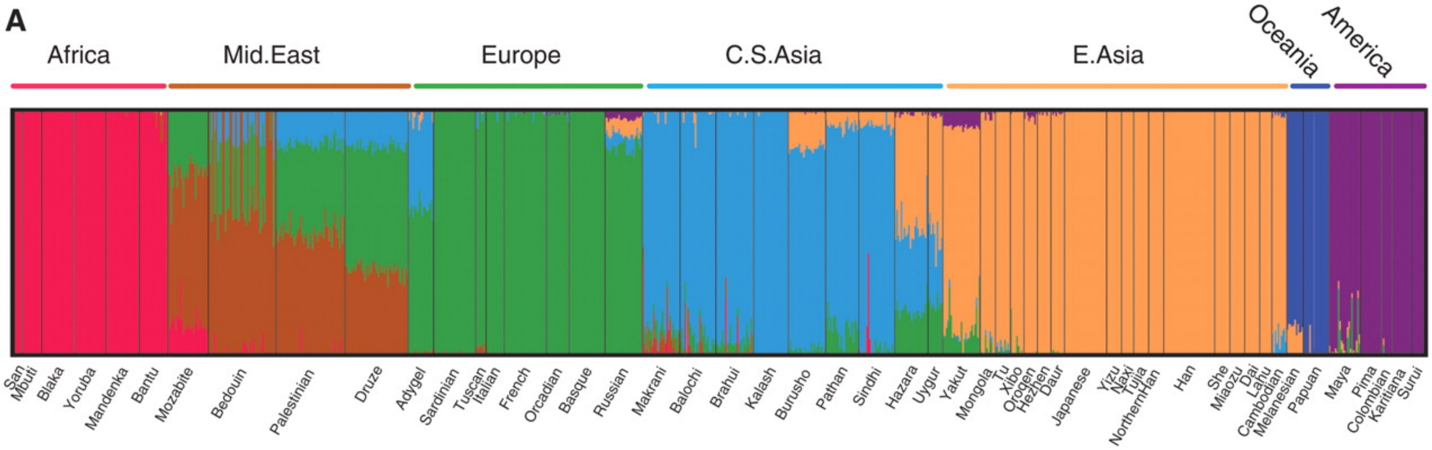


Figure 3.14: Structure/Admixture plot of human populations. *This study collected genome-wide SNP data from* 931 *individuals representing* 51 *indigenous human groups, listed at bottom. The data were fit under a version of the Structure/Admixture model, assuming* $K = 5$. *The sampling labels were not used in the analysis.* Credit: Figure 1A from Jun Li et al (2008). [Link] Used with permission.

Perhaps the most striking feature of the clustering results is that the main genetic clusters in this analysis correspond roughly to continental groupings. This occurs despite the fact that the clustering algorithm did not use population labels or geographic information.

Secondly, you'll notice that *some groups are mixtures of different clusters*: especially in the Middle East populations where there is a Middle Eastern component (brown), but also variable amounts of ancestry from Europe (green), Central/South Asia (blue) and Sub-Saharan Africa (red) [382]. These signals reflect the location of the Middle East as a geographic crossroads that has experienced extensive immigration from each of these other regions during the past few thousand years [383].

Similarly, the Uygurs, a Turkic population of northwest China, appear as a mixture of multiple components including East Asian, Central Asian, and European, reflecting their geographic and historical position as a crossroads between these regions.

In certain other cases, there is highly *variable ancestry within a population sample*: for example individuals in the Maya group (from Mexico) carry varying amounts of admixture, mainly from Europeans, likely reflecting historical Spanish migration into Central America.

Finer-scale structure: Northeast Africa. *This type of analysis also reveals much finer-scale structure within regions* of the world. For example, the analysis below by Nina Hollfelder and colleagues [384] focuses on structure within Northeast Africa, and how it relates to geographically nearby groups. As hinted at by the global analysis above, we see that this region shows highly complex patterns of structure and mixing:

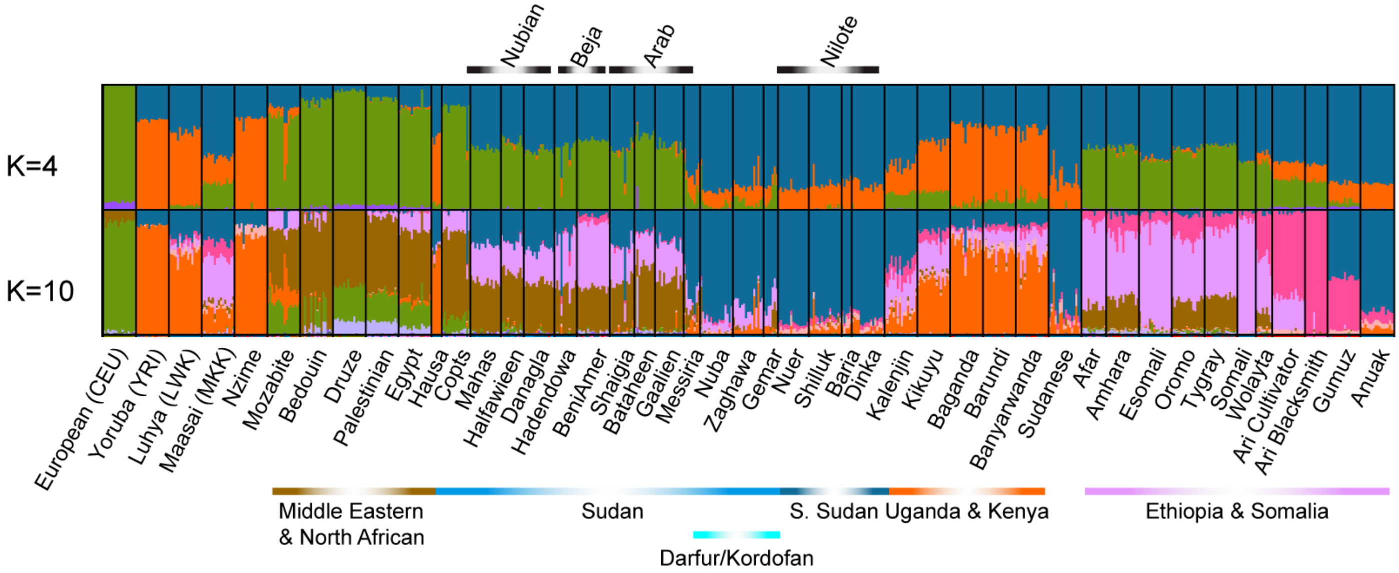


Figure 3.15: Admixture analysis of northeast Africa and nearby populations. *Populations are indicated along the bottom, grouped by region. The rows show clustering results for different values of K, the number of assumed clusters. Notice that different values of K emphasize different aspects of the structure in the data visualization.*

Credit: Modified from Figure 2 of Nina Hollfelder et al (2017) [Link] CC-BY-4. See the original paper for additional values of K, and more interpretation.

At $K = 4$ the genotypes are modeled as varying mixtures of three main components: a European/Middle Eastern cluster (green); a West African Bantu cluster (orange) and a North African Nilotic cluster (blue) [385].

But at $K = 10$ the seemingly simple structure from $K = 4$ is broken down into a variety of subgroups: for example splitting the green cluster into European (green) and Middle Eastern clusters (brown); adding a light purple cluster to describe East African ancestry found in the Ethiopian populations as well as to lesser extents in the Sudanese and the Maasai (Kenya); and a pink cluster most associated with the Ari and Gumuz populations of Ethiopia.

I can't do justice to the complex history and structure of North African populations here, but you can read more about this topic here: [386] [f].

Comments. As you can see from these examples, *structure is usually both highly complex, and hierarchically nested*. Structure that may seem relatively simple in the global analysis is revealed to contain additional structuring within groups with more extensive local sampling of diverse ethnic

[f] *We'll see another application of Admixture in Chapter 3.3, applied to populations in sub-Saharan Africa.*

groups.

Moreover, in this example we see *almost every population shows extensive admixture between clusters*, reflecting the dynamic history of the region. Admixture is a ubiquitous process in human populations, and we'll talk more about it in the next chapter.

Genetic clustering: PCA. A second major approach uses **Principal Components Analysis (PCA)**. PCA is widely used in data science as a technique for linear dimension reduction in high-dimensional data [387].

Most relevant to us, PCA has become another hugely important technique for studying population structure in genetic data, sometimes revealing different aspects of the data than Structure/Admixture [g].

As above, we start with a genotype matrix G, with I rows and L columns (individuals x loci). One way to think of G is that it represents each individual's genotype as a point in L-dimensional space. Since L, the number of SNPs, is extremely large, this is not directly useful for visualizing data.

PCA is an example of a technique for *dimensionality reduction*, which means that it "projects" the data into a "low-dimensional" approximation of the data that is far more useful for interpretation. In short, PCA tries to predict the genotype matrix G as a product of two much smaller matrices [h]: the "loadings" matrix Λ and the "factors" matrix F.

In this model the expected value for a single genotype (the red dot in G, below) is predicted by the product of a row in Λ times a column in F. We can visualize this here:

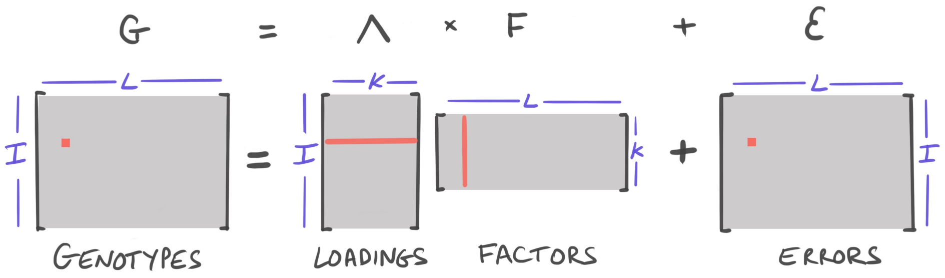


Figure 3.16: **PCA is a low-dimensional matrix factorization of the genotype data:** *it estimates two low-rank matrices, Λ and F to predict G as accurately as possible. The errors in predicting G are contained in the matrix $\mathcal{E}$; these are similar to residuals in regression. The loadings matrix contains the ancestry of each individual; this is what is shown in PCA plots.*

Credit: Redrawn from Figure 1 of Barbara Engelhardt and Matthew Stephens (2010) [Link]

Mathematically, the first PC is chosen to explain as much of the variance in G as possible; each subsequent PC is chosen to explain as much of the remaining variance that has not already been captured by earlier PCs. As we'll discuss in the upcoming box, *there's a strong mathematical connection between PCA and the Structure model, if you think of Λ as serving the role of Q, and F being like P*; meanwhile the PCA representation is mathematically more flexible (for better or worse, depending on context).

PCA provides a powerful tool for visualizing and modeling data. We're usually most interested in the individual loadings, Λ, which reflect the ancestries of individuals. People often plot just the first two dimensions of this, known as **principal components** or **PCs**.

As one example, in what may be the single most famous data visualiza-

[g] *PCA is widely used in association studies (GWAS) as a tool for removing confounding effects of population structure (Chapter 4.5).*

[h] *A note on Greek letters: Λ is the upper case Greek letter lambda; $\mathcal{E}$ is the upper-case epsilon.*

tion in human genetics, John Novembre and colleagues showed a remarkable correspondence between geography and the PCA projection of European individuals. Novembre et al performed PCA on genome-wide SNP data for 1387 European individuals.

They plotted each individual into a 2-dimensional scatterplot based on their loadings in the first two PCs (i.e., the first two columns of the loadings matrix Λ). When each point was colored according to the country of origin of their grandparents, the data showed a remarkable correspondence between the position of individuals in the PC coordinate space, and the geography of the populations in Europe:

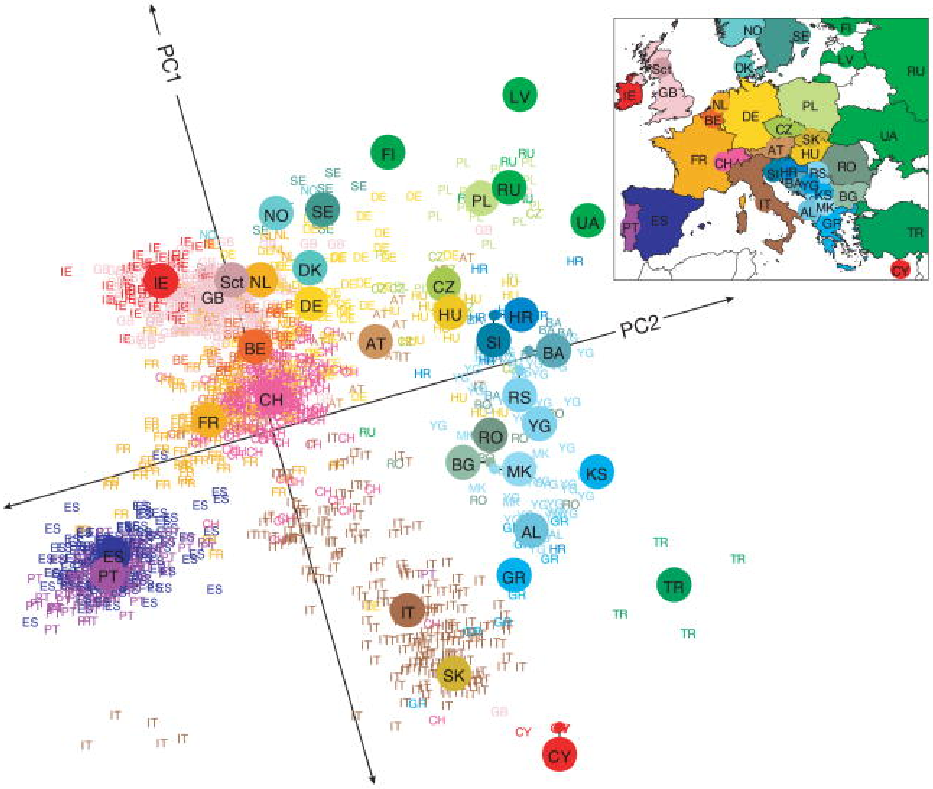


Figure 3.17: **Genes mirror geography within Europe (2008).** *The plot shows the PCA projection of 1387 European individuals. Each two-letter code shows the projection of a single individual, and the circled letters show the average projections of individuals from the corresponding country. The map in the upper-right gives the two-letter country codes. After slightly rotating the PCA axes, the PCA projection reflects the geographic positions of countries to a remarkable degree, with only minor distortions.* *Credit: Figure 1A from John Novembre et al (2008). [Link] Used with permission.*

You may wonder *why* the population structure of Europe is so regular. In theory, this kind of PCA pattern can result from stable population structure with short-range migration and drift [388]. But in Europe, the lead PCs arise because modern Europeans are derived from varying mixtures of three different ancestral populations who spread into Europe from different directions: Western Hunter Gatherers, who dominated Europe 10,000 years ago; Anatolian farmers who spread into Europe from modern-day Turkey around 8000 years ago; and herders who spread from the Russian steppes around 5000 years ago [389]. We'll revisit this in Chapter 3.4.

Optional details: PCA, and its relationship to Structure.

As above, with Structure, we start with a genotype matrix G, with I rows and L columns (individuals x loci). For PCA it's conventional to *center* and *rescale* the data to make a new genotype matrix G^* where each entry $g^*_{i,l}$ is calculated as

$$g^*_{i,l} = \frac{g_{il} - 2p_l}{\sqrt{2p_l(1-p_l)}}, \tag{3.5}$$

where p_l is the allele frequency for SNP l in the sample. The numerator of this expression $(g_{il} - 2p_l)$ centers the genotypes so that every SNP has a mean of 0. The denominator $(\sqrt{2p_l(1-p_l)})$ rescales the data so that every SNP has variance ≈ 1. This effectively upweights rare SNPs so that every SNP contributes equally to the PCA [390].

PCA provides a *low-dimensional projection* of the individuals into K dimensions. For data visualization, K is usually chosen as 2, but people regularly use 10 or more PCs as covariates to control for structure in GWAS analyses (Chapter 4.5). In matrix notation, we can write the PCA projection as predicting the genotype matrix G^* as a linear combination of two much smaller matrices [391]:

$$\mathrm{E}(G^*) = \Lambda F \tag{3.6}$$

where the loadings matrix Λ is an $n \times K$ matrix, and the factors matrix F is $K \times L$. Equivalently, we can write this in terms of the expected genotype for a single SNP in a single individual:

$$\mathrm{E}(g^*_{i,l}) = \sum_{k=1}^{K} \lambda_{i,k} f_{k,l} \tag{3.7}$$

where $\lambda_{i,k}$ and $f_{k,l}$ are elements of the Λ and F matrices. The meaning of the expectation here is that we predict the *expected* genotype matrix as well as possible given the individual ancestries (i.e., Λ) and the allele frequency shifts for each population (i.e., F).

How are Λ and F chosen? I won't go into the linear algebra here, but the key idea is that to get the first principal component, we solve for the first column of Λ and row of F that minimize the squared error in predicting G^*. The second principal component then fits the "residual" error that is left behind after removing the first principal component: i.e., it minimizes the squared error in $G^* - \Lambda_{.,1}F_{1,.}$. And so on, for subsequent PCs [392].

Relationship between PCA and Structure. At first glance, PCA and Structure seem like very different models [393]. However, both algorithms essentially aim to estimate the allele frequencies in an individual as a linear combination of K factors that reflect *individual ancestry × population allele frequencies*. In the case of PCA we can rearrange Equation 3.7 to show that each individual genotype g_{il} is predicted as

$$\mathrm{E}(g_{i,l}) = 2p_l + c \sum_{k=1}^{K} \lambda_{i,k} f_{k,l} \tag{3.8}$$

where c is $\sqrt{2p_l(1-p_l)}$.

Meanwhile, in the Structure model, we can think of each individual i as having a "personal" expected allele frequency r_{il} at SNP l that depends on their ancestry vector Q_i and the population allele frequencies P_l:

$$r_{i,l} = \sum_{k=1}^{K} q_{i,k} p_{k,l}. \tag{3.9}$$

Here, $q_{i,k}$ takes the place of the loading $\lambda_{i,k}$ while $p_{k,l}$ takes the role of the factor $f_{k,l}$. In Structure, the individual's genotype at this SNP is a binomial sample of two alleles given $r_{i,l}$:

$$g_{i,l} = \text{Binomial}(2, r_{i,l}) \tag{3.10}$$

However, there are important differences between the two models. Compared to PCA, Structure is motivated by an explicit genetic model, and hence the ancestry components q are non-negative and add to 1 for each individual, while the allele frequencies p are required to be in $[0,1]$. In contrast, in PCA, both the λs and the fs are typically centered around 0 and can take on any real value. Since PCA minimizes total squared error, it also makes an implicit assumption that the squared errors $(g_{i,l} - \text{E}[g_{i,l}])^2$ should be weighted equally across all SNPs. Unlike in Structure, the PCA factors are constructed to be orthogonal to one another.

These specific differences between the algorithms, including the difference in how λ versus q are represented, often give rise to visualizations that may be more or less appealing for one or other method. I find that PCA is usually easier to interpret for geographic structure at local scales, especially when it is fairly continuous, as in the European example. However, for more complex structure, with meaningful loadings on more than two dimensions, the standard PCA plots quickly become overplotted and hard to interpret. For this reason, Structure/Admixture plots provide better visualization for complicated structure with many distinct clusters as in the examples above.

So far we have been treating SNPs as independent. But we can potentially gain a different type of information by looking at haplotype sharing between individuals.

Haplotype-based clustering. When we use SNPs for clustering, we rely on genetic drift to produce allele frequency differences between groups. However, drift occurs very slowly in large populations, and allele frequencies in closely related populations are extremely similar. But with haplotype sharing, we may be able to identify pairs of individuals with recent shared ancestors. Those pairs of individuals are more likely to come from the same, or closely related, populations. *Clustering approaches that use this kind of information can often identify structure at remarkably recent, and local, timescales.*

One version of this, known as fineSTRUCTURE [394], starts from the Li and Stephens copying algorithm that we covered in Chapter 2.3, in the context of phasing and imputation. Other recent approaches have used sharing of IBD (identity-by-descent) segments [395] and sharing of recent coalescent events in an ancestral recombination graph [396] [i].

Recall that the Li and Stephens algorithm models each genome as a copying path from a panel of other individuals. The key concept for fineSTRUCTURE is that an individual should be more likely to copy from individuals in the same, or closely related populations, than from distantly related populations:

[i] *For an ARG-based approach to this see Chapter 3.3.*

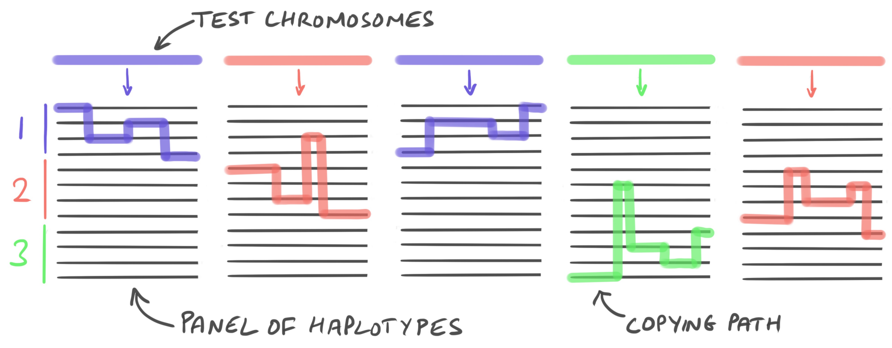


Figure 3.18: **The fineSTRUCTURE copying process.** *In each example, a test chromosome is modeled as a copying path through a panel of other haplotypes. Individuals from the same population (red, blue, and green) tend to copy from the same rows (as indicated at left). fineSTRUCTURE clusters individuals based on whom they copy. (In practice, individuals are diploid, so each individual actually makes two copying paths.)*

fineSTRUCTURE starts by estimating a matrix of copying rates between all pairs of individuals in the sample. It uses that matrix to cluster individuals into groups that share similar copying rates. Critically, **in a large sample, copying often takes place from individuals who are genetically related within the last few tens of generations, and this provides power to detect clusters of individuals with very recent shared ancestry**.

Stephen Leslie and colleagues used this approach to study population structure in Britain and Northern Ireland [397]. The authors genotyped 2,039 British individuals, sampled on the basis that all four grandparents were born close together. PCA and conventional Structure provide only limited resolution in this data set.

FineSTRUCTURE clustered the individuals into 17 groups. These show a remarkable agreement with geography, often identifying structure at extraordinarily fine geographic scales:

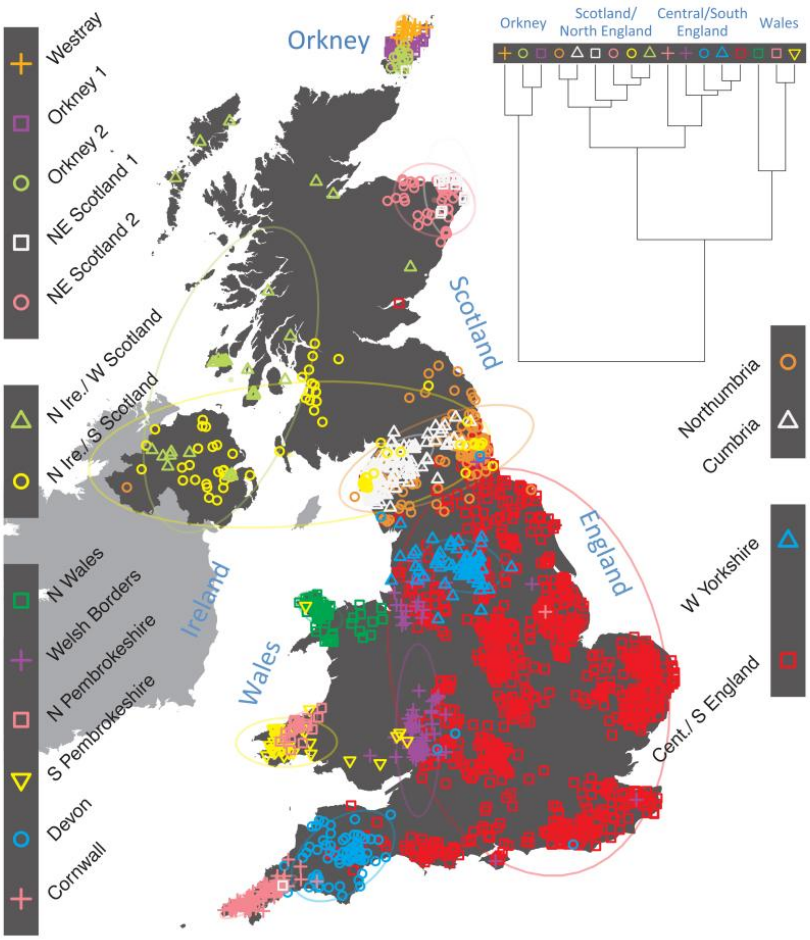


Figure 3.19: **Fine structure of Britain and Northern Ireland.** 2,039 *individuals were clustered into 17 groups using fineSTRUCTURE. Each data point indicates the geographic sampling location, and the symbols indicate the assigned cluster. The bars at left and right indicate labels assigned to each based on their geographic distribution. Credit: Figure 1 from Stephen Leslie et al (2015). [Link] Used with permission.*

For example in the southwest corner of England, the algorithm identifies distinct clusters from the neighboring counties of Cornwall and Devon, each of which is only about 50 miles across.

Even the clusters that are geographically dispersed in the plot above are likely informative. For example, two clusters that are shared between Northern Ireland and Scotland (in yellow and lime) likely reflect extensive historical migration in both directions, including the Ulster Scots migration to Ireland in the 17th Century, and the Great Famine migration from Ireland back to Scotland in the 19th Century.

We can expect that similarly high resolution estimates of genetic ancestry will be attainable in many parts of the world, as high-density data are increasingly available.

In summary, it's interesting to contemplate the advances in resolution in studies of population genetic structure since the early work from the 1990s, using $\sim$*100 SNPs to identify continent-level clustering, to present day studies at or below the scale of countries or even counties.*

In the last sections of this chapter we step back a little to consider what clustering does, and doesn't, tell us about ancestry and race.

Population clusters and their interpretation. Clustering techniques are important because they play a central role in how we understand human variation and human history, and are an essential tool for data analysis in genome-wide association studies. And yet it's also important to understand what these models *don't* teach us [398].

• First: What do the population clusters represent? One way to think about this is in terms of an individual's ancestors. Remember that each of our genomes is made up of many segments of DNA, inherited from many different ancestors. Going backward in time, you can think about these ancestors as occupying a distribution across geographic space or populations.

Perhaps counterintuitively, there's a strong mathematical argument that we all share pedigree ancestors within the past few thousand years [399]. Even when we look at individuals from different clusters – even from different continents – they all share many ancestors.

Instead clustering works on something much more subtle: *individuals who are genetically similar – for example, the two blue individuals below – have similar* ***distributions*** *of ancestors across space and time*, while both differ from the red individual. But if we go back a bit further in time, to the two upper time-slices, all three individuals have very similar distributions of ancestors in space:

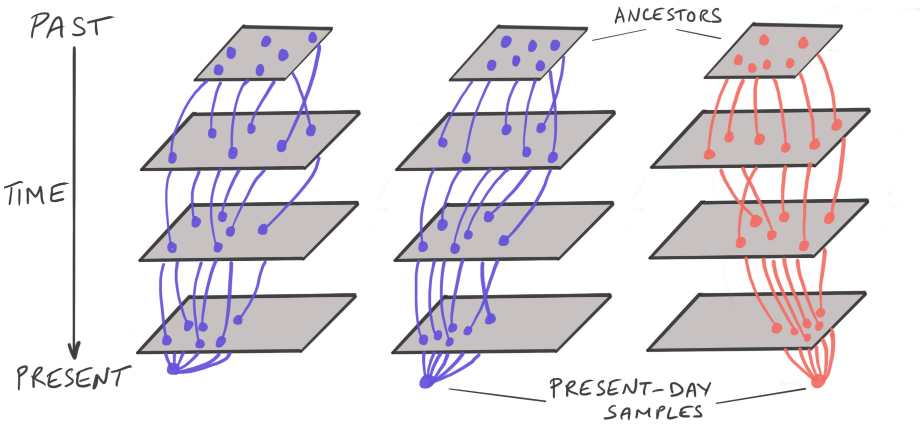


Figure 3.20: **You can think of genetically similar individuals as having similar geographic distributions of ancestors.** *The cartoon shows the geographic distributions of ancestors for three individuals. Each slice shows the locations of ancestors at a single point in time. Individuals who cluster together have similar geographic distributions of ancestors.*

So, even though we all share many ancestors, *a helpful way to think about clustering is that individuals who cluster together tend to have similar distributions of ancestors at each point in time.*

Different methods implicitly reflect distributions of ancestors at different time-scales. Notably, fineSTRUCTURE and other haplotype-based methods can achieve high geographic resolution because they are designed to detect similarity of ancestors at the most recent timescales.

• Second, while clustering techniques do clearly identify real signals of structure in the data – e.g., consider the PCA of Europeans – it's important to be aware that the precise results are often sensitive to the numbers of samples, precisely which samples are included, and the details of the analysis [400].

In large part this is because true population structure is usually complex, including both continuous variation and hierarchical levels of structure at different geographic scales, and potentially non-random mating according to ethnicity, religion, or social group. Clustering methods generally capture the major axes of variation within this complex reality, so *small changes in the sample composition can change which aspects of the structure are emphasized by the main PCs or population clusters.*

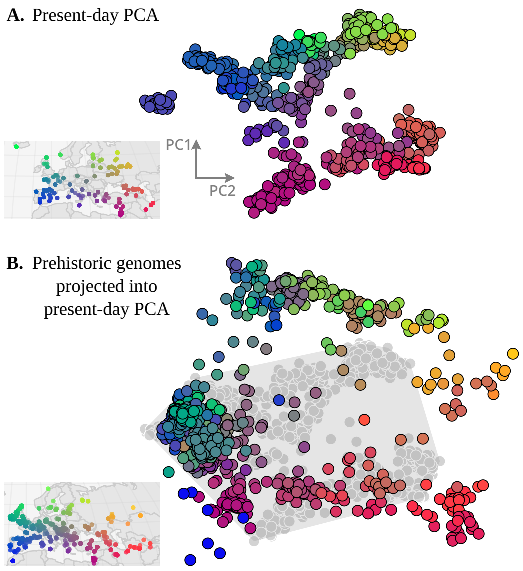


Figure 3.21: **PCA of genomes from Europe and the Middle East. A.** *PCA projection of present-day samples; color code in inset map.* **B.** *Prehistoric samples (5,000-12,000 years ago) projected into the same PCA space (shown in gray). Present-day genomes are mixtures of ancient populations that no longer exist.* Credit: *Unpublished figure kindly contributed by Clemens Weiß, CC BY 4.*

• Third, clusters are not necessarily stable over historical time as population movements, splitting, and merging are constant forces in human evolution. For example, I mentioned above that the genomes of modern Europeans are mixtures of at least 3 distinct ancestral groups (western hunter gatherers, Anatolian farmers, and herders from the Russian steppes) [401]. For this reason, many inhabitants of Europe prior to the Steppe admixture do not cluster especially closely to any modern day populations. What we might describe in modern samples as "European ancestry" did not exist in anything like the current form until just 5,000 years ago.

This complex genetic history of Europe was largely invisible to us until the application of ancient DNA techniques starting around 2010. *Similar complexities have emerged almost everywhere that ancient DNA is available.* Thus, we should think of population clusters as relating only to the population structure detectable in the available samples.

• Last, human genetics doesn't occur in isolation from the rest of society, and there's an important concern that our concepts can be misinterpreted by nonscientists. In particular, the use of clusters may encourage patterns of **typological thinking**: i.e., the incorrect notion that humans fall into a limited set of distinct forms or types [j] [402].

Moreover, the clustering techniques we use may sometimes exaggerate the homogeneity within groups, and the differences between them. A compelling essay by Graham Coop argues that we should move away from a focus on *ancestry groups* to focusing on *genetic similarity* [403].

Coop proposes that one should say, for example: *' "Graham is genetically similar to the GBR 1000 genome samples (on the first 10 principal components)" rather than "Graham has Northwestern European genetic ancestry" '* [404]. He notes that while statements about similarity may seem awkward, they side-step the limitations of clumping all humans into a finite number of discrete groups (and their mixtures). For additional nuanced discussion about when 'similarity' or 'ancestry' may be most appropriate see [405].

Race and ancestry. So far we have talked at length about population structure and genetic ancestry. We close this chapter by touching on the complicated relationship between ancestry and popular conceptions of **race** as used in cultural settings. Race categories are socially constructed labels that group people based on features such as ancestral origins, culture, and often physical characteristics such as skin pigmentation and hair type. Conceptions of race and ethnicity vary among countries, and can change through time, but generally have some imprecise relationship with the concept of ancestry.

The term *race* is not easily defined, but 'race' is usually thought of as relating to ancestral origins and physical characteristics. In the US, current conceptions of race include categories such as *Black/ African American*, *White* and *Asian* [k]. The term *ethnicity* is also used in the US as another categorization alongside race: these categories usually include a cultural component: for example, *Latino/ Hispanic*, which identify people from Latin America or Spanish-speaking countries, respectively. Brazil – another country with diverse ancestry – classifies people somewhat differently, and the official census includes categories of *branco* (white), *pardo* (brown or mixed), *preto* (black), *amarelo* (yellow/ East Asian), and *indigenous*. (Brazilians living in the US might self-identify for census purposes as both Latino and a race category.)

A common source of confusion is that race/ethnicity categories usually overlap–though imperfectly–with genetically-inferred ancestry categories. For example, as we will discuss in the next chapter, most African Americans have mixed west African and European ancestry, albeit in varying proportions. Latinos often have mixed ancestry including potentially various native American groups, European, and west African, though the proportions vary greatly across individuals.

[j] *This topic relates to cultural concepts of race, which we discuss next.*

[k] *A sobering look at how the US Census has categorized Americans since 1790 can be found at the Pew Research Center: [Link].*

Which is more relevant: race, or ancestry? [1] In genetics research and clinical genetics (and for the purpose of this book), the concept of ancestry is more precise and, usually, more relevant than race. I'll give three examples. First, in genome-wide association mapping, genetic ancestry can act as a confounding variable, and may lead to biased effect-size estimates and spurious signals if not controlled for within the analysis (we'll cover this topic in a later chapter). Second, a completely different issue arises in clinical genetics, where it is common to screen patients for rare variants in known disease-causing genes, as such variants may underlie the disease. A standard filter would be to ignore variants that are found at appreciable frequencies in healthy control samples; this kind of filter is less useful if we don't have enough control samples of similar ancestry to the patient. Third, for studies of human population genetics, it's usually more useful to focus on genetic ancestry.

But for many other situations in research and healthcare, race/ ethnicity may actually be *more* relevant than ancestry. For example, social categories of race/ethnicity often correlate with many aspects of lived experience such as culture, the neighborhoods people live in, resourcing of public schools, health-care access, and experiences of discrimination. All of these factors affect health outcomes, and all are likely more related to social perceptions of race than to genetically-measured ancestry. Thus, race can serve as a label that correlates with important, but hard to measure, environmental aspects of environment. When using measures of race in research, however, is important to avoid the facile assumption that racial differences – for example in health and disease – are mainly due to genetic differences between ancestries. There is a common tendency in society, and even among some geneticists, to over-emphasize the role of genetics, and under-emphasize the role of systemic environmental effects [406].

In this chapter we have discussed methods for studying contemporary population structure and recent admixture. In the next chapter we'll expand on the theme of admixture between populations.

[1] *For a deeper dive into this topic I recommend the National Academies Report on Population Descriptors [Link], including especially Chapter 5 and Table 5.1.*

Notes and References.

[364] Thanks again to the fantastic generosity of people who commented on earlier drafts of this chapter and the next (some of whom even commented on multiple versions!): Molly Przeworski, Doc Edge and his lab, Roshni Patel, Aylwyn Scally, and Leo Speidel; thanks also to Roshni Patel and Clemens Weiß for kindly contributing original figures. As always, any errors are my own.

[365] Rohde DL, Olson S, Chang JT. Modelling the recent common ancestry of all living humans. Nature. 2004;431(7008):562-6

[366] This problem is known as **topic modeling** and has wide applications in computer science but has also been used in other areas including single cell RNA-seq modeling. A classical approach to this treats each document as an unstructured bag of words, and each topic is defined by a probability distribution over all words. The most widely-used approach for this problem is known as Latent Dirichlet Allocation [Link], and is closely related to models for genetic structure (see Pritchard et al 2000, cited below). The classic LDA citation is:

Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3(Jan):993-1022

[367] Aspects of this work go back to the 1970s and 1980s, however the data were very limited at that time:

Mitton JB. Genetic differentiation of races of man as judged by single-locus and multilocus analyses. The American Naturalist. 1977;111(978):203-12

Smouse PE, Spielman RS, Park MH. Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. The American Naturalist. 1982;119(4):445-63

[368] The plot shown below is from

Mountain JL, Cavalli-Sforza LL. Multilocus genotypes, a tree of individuals, and human evolutionary history. The American Journal of Human Genetics. 1997;61(3):705-18,

building on ideas from an earlier paper using STRs

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. High resolution of human evolutionary trees with polymorphic microsatellites. Nature. 1994;368(6470):455-7.

This work was led by Luca Cavalli-Sforza, who was a giant of 20th Century human population genetics – pioneering the use of genetic data to study human history, and posing some of the key questions in the field. He also served on my PhD committee, and I like to think of him at age 76, peddling his ancient bike to my PhD defense, his long white hair streaming behind him.

For early methods work using the Mountain and Cavalli data for population assignment see

Rannala B, Mountain JL. Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences. 1997;94(17):9197-201.

[369] The data were actually RFLPs which are an old-fashioned technique that approximates SNP typing.

[370] This fact is very closely related to the observation that F_{ST} is generally low between human populations, topping out at about 0.15 (Chapter 2.4).

[371] As described here, this generative model treats Q and P as fixed parameters. If you prefer, you could consider the allele frequencies at each site as the random outcome of mutation and drift in a population genetic model as outlined in Chapter 2.4. This could be straightforwardly using SLiM or tskit. Similarly you could model Q as the result of a random process: for example as the result of some specified sampling and/or admixture model, or as a sample from a generic probability distribution such as a Dirichlet (see Pritchard et al 2000, cited below).

[372] This standard set of assumptions implies that the two alleles are inherited independently given Q. This isn't precisely correct for very recent pedigree ancestors: for example you can't get two alleles from the same great grandparent!) but it works well in practice and it's a useful simplification for the much more complicated patterns of ancestry that usually crop up in real life.

[373] In other words, the most basic model ignores fine-scale LD between SNPs. This is works well for the basic model because any given LD block only contains a tiny fraction of the genome so the error of ignoring LD is negligible. However, as we'll discuss later in this chapter and in the next chapter, in some cases we can learn more by making specific use of LD and haplotype patterns.

[374] The basic problem is this. Suppose that a particular allele is rare in population k, and not observed in our initial sample. The simple point estimate assigns this allele a frequency of zero, even though it may not be truly zero. Suppose we later sample an individual from population k who *does* carry this allele. Now their genotype likelihood in population k is zero, and we could never assign them to k, even though it is the correct population. If we look at a large enough number of SNPs this will almost certainly occur somewhere in the genome. To solve this, people often add α to the numerator and 2α to the denominator for each population to avoid zeroes in the frequency data. This procedure corresponds

to computing the posterior mean under a beta-binomial sampling model, with a prior of Beta(α, α) (Section 3.4 [Link]). A typical choice would be $\alpha = 1$.

[375] Bayes' Rule, named after the 18th century mathematician Thomas Bayes provides a mathematical rule for calculating conditional probabilities. We compute the probability that some outcome A is true given observed data B as follows:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\sum_{A' \in S} \Pr(B|A')\Pr(A')}. \tag{3.11}$$

where A' is a possible outcome and S is the set of all possible outcomes. In this expression $\Pr(B|A)$ is known as the *likelihood*: it is the probability of the data assuming outcome A is true. $\Pr(A)$ is known as the *prior*: it contains our subjective probability that outcome A is true *before* collecting data.

[376] We apply Bayes' Rule to estimate the source of an individual given the data (the genotypes G and allele frequencies P) and any prior information π_j about where the individual is from as follows (**no-admixture**):

$$\Pr(\text{individual } i \text{ from population } k) = \frac{\Pr(G_i|q_{i,k} = 1)\pi_k}{\sum_{j=1}^{K} Pr(G_i|q_{i,j} = 1)\pi_j} \tag{3.12}$$

Here, π_j is known as the *prior probability* that this individual is from population j. The prior probability reflects any information we may have in advance about where someone is from, for example based on sampling location (Hubisz et al 2009). However in most applications it's difficult to quantify the prior information and in any event this information tends to be weak compared to the genotype likelihood. For this reason it's most common to ignore prior information altogether and simply set all $\pi_j = 1/K$. In that case, the priors cancel, and we get simply

$$\Pr(i \text{ from population } k) = \frac{\Pr(G_i|q_{i,k} = 1)}{\sum_{j=1}^{K} Pr(G_i|q_{i,j} = 1)} \tag{3.13}$$

In the **admixture** case, Q_i is a vector of K non-negative real numbers that sum to 1, so we want to estimate a posterior density. The form of this is similar, except that the denominator is an integral over all possible values of Q_i:

$$\Pr(Q_i = \hat{Q}_i) = \frac{\Pr(G_i|\hat{Q}_i)\pi_{\hat{Q}}}{\int_{Q^*} \Pr(G_i|Q^*)\pi_{Q^*} \mathrm{dQ}^*}. \tag{3.14}$$

Here, π_Q is the prior, and again it's convenient to set this to be uniform so that it cancels out. The integral in the denominator is hard to compute so in practice most people just use the maximum likelihood estimate of $\hat{Q}_i$, i.e., simply finding $\hat{Q}_i$ that maximizes $\Pr(G_i|\hat{Q}_i)$.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Molecular ecology resources. 2009;9(5):1322-32

[377] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945-59

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164(4):1567-87

Novembre J. Pritchard, Stephens, and Donnelly on population structure. Genetics. 2016;204(2):391-3

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009;19(9):1655-64

[378] I've been skating over a rather thorny issue: everything above assumes that we have already chosen a particular value of K, the number of clusters. But usually we do not know in advance what K should be and we would like to choose K based on the data. Both the Structure and Admixture software packages recommend criteria for selecting K, but in practice it's often difficult to choose a single best value–indeed this makes sense as identifying the "correct" number of clusters is frequently not a well-defined problem in practical settings where the data may not completely fit the idealized Structure notion of discrete ancestral populations. Structure is usually hierarchical, including additional local structure within the larger populations, and often varying continuously across geographic space rather than being entirely discrete. For these reasons, there is not usually a single "true" value of K. I find that Structure/Admixture are best viewed as tools for visualizing population structure and generally prefer to choose values of K that are useful for understanding the data, and many papers choose to present results for multiple values of K that are useful for visualization and interpretation.

[379] Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science. 2002;298(5602):2381-5

[380] You can see a map of the HGDP populations in Chapter 1.2. The HGDP is a panel of DNA and cell lines from around 1000 unrelated individuals, sampled from 51 populations from around the world. The HGDP was collected by a network

of scientists in the 1980s and 1990s, and aimed to capture the genetic diversity of indigenous populations around the world. The HGDP is complementary to the 1000 Genomes data set: it is about one third as large, but has much better sampling of diverse human populations albeit with important gaps, including among Native North Americans, Native Australians, and from India

[381] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319(5866):1100-4;

The data analysis applied a maximum likelihood algorithm called Frappe that is similar to Structure/Admixture:

Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genetic Epidemiology. 2005;28(4):289-301.

[382] In this plot, the Mozabite have been labeled with the Middle Eastern populations, although in fact they are located in the northern Sahara region of Algeria. In the ancestry analysis they have a small Sub-Saharan component (red), as well as larger Middle Eastern and European components.

[383] Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics. 2012;8:e1002967

Moots HM, Antonio M, Sawyer S, Spence JP, Oberreiter V, Weiß CL, et al. A genetic history of continuity and mobility in the Iron age Central Mediterranean. Nature Ecology & Evolution. 2023;7(9):1515-24

[384] Hollfelder N, Schlebusch CM, Günther T, Babiker H, Hassan HY, Jakobsson M. Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations. PLoS Genetics. 2017;13(8):e1006976

[385] The fourth cluster in purple is mainly present in a single admixed individual from the Bataheen, an Arab group from Sudan. This might be some quite divergent ancestry but it's unclear to me what this represents.

[386] In addition to Hollfelder et al, and Moots et al, see:

Lucas-Sánchez M, Serradell JM, Comas D. Population history of North Africa based on modern and ancient genomes. Human Molecular Genetics. 2021;30(R1):R17-23.

[387] For a somewhat mathy tutorial see

Shlens J. A tutorial on principal component analysis. arXiv preprint arXiv:14041100. 2014

[388] Novembre and Stephens explored how PCA represents the data in isolation-by-distance models

Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nature genetics. 2008;40(5):646-9.

[389] We'll cover this topic more thoroughly in Chapter 3.4, but for a look at spatial patterns of the main ancestral groups in Europe see

Racimo F, Woodbridge J, Fyfe RM, Sikora M, Sjögren KG, Kristiansen K, et al. The spatiotemporal spread of human migrations during the European Holocene. Proceedings of the National Academy of Sciences. 2020;117(16):8989-9000

[390] The mean adjustment is used because otherwise the first principal component reflects the mean allele frequencies. The variance adjustment in the denominator (which is optional) is sometimes motivated by noting that drift occurs at a rate proportional to $\sqrt{2p_l(1-p_l)}$; so with this rescaling the amount of drift at each SNP is approximately independent of p. Patterson et al (2006) rationalized use of this denominator by noting that it gives each SNP approximately equal weighting, and also with reference to the Nicholson and Donnelly drift model (2002). However in certain nonstandard contexts use of the denominator may complicate analysis (McVean 2009; Pickrell and Pritchard 2012). While the variance reweighting is conventional it usually doesn't impact the results much.

[391] There are several different ways of writing the mathematics of PCA, as well as motivating it; my notation here follows an unusually clear description by Barbara Engelhardt and Matthew Stephens:

Engelhardt BE, Stephens M. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS genetics. 2010;6(9):e1001117.

[392] It's beyond our scope here, but there's also **an important connection between PCA and pairwise coalescent times**: specifically, the individual loadings onto PC axes can be interpreted as functions of pairwise coalescent times, such that pairs of individuals separated by greater average coalescent times will project apart than individuals with smaller coalescent times:

McVean G. A genealogical interpretation of principal components analysis. PLoS genetics. 2009;5(10):e1000686.

[393] This section follows the exposition in Engelhardt and Stephens (2010)

[394] Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS genetics. 2012;8(1):e1002453

[395] Ralph P, Coop G. The geography of recent genetic ancestry across Europe. PLoS biology. 2013;11(5):e1001555

Nait Saada J, Kalantzis G, Shyr D, Cooper F, Robinson M, Gusev A, et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. Nature communications. 2020;11(1):6130

[396] Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. Nature Genetics. 2019;51(9):1330-8

[397] Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. Nature. 2015;519(7543):309-14

Another example of fineSTRUCTURE examines genetic structure of the Iberian Peninsula, showing a striking concordance with history:

Bycroft C, Fernandez-Rozadilla C, Ruiz-Ponte C, Quintela I, Carracedo Á, Donnelly P, et al. Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. Nature Communications. 2019;10(1):551

[398] For an excellent and in-depth consideration of these topics see a 2023 report from the National Academy of Sciences [Link]:

committee on Population & National Academies of Sciences Engineering & Medicine. Using Population Descriptors in Genetics and Genomics Research. National Academies; 2023.

See also

Lawson DJ, Van Dorp L, Falush D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nature Communications. 2018;9(1):3258

[399]

Rohde DL, Olson S, Chang JT. Modelling the recent common ancestry of all living humans. Nature. 2004;431(7008):562-6

[400] The results are heavily influenced by the numbers and distributions of samples. It's been argued that the particular locations of the HGDP samples tend to accentuate the gaps between continental clusters: for example with better sampling across northeast Africa we would see a more gradual shift of ancestry components from African to Middle Eastern ancestry, as evident in the Hollfelder et al analysis

[401] Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513(7518):409-13

[402] As human geneticists we have a responsibility to be aware that our work can be misinterpreted by nonscientists, or weaponized by extremists who promote a pseudoscientific form of racism:

Bird KA, Carlson J. Typological thinking in human genomics research contributes to the production and prominence of scientific racism. Frontiers in Genetics. 2024;15:1345631.

[403] Coop G. Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics. arXiv preprint arXiv:220711595. 2022

[404] Coop 2022, p11. Italics are mine.

[405] National Academies Report, including Chapter 5 and especially Table 5.1 [Link].

[406] For a range of views on this topic see the NAS review as well as

Rosenberg NA, Edge MD, Pritchard JK, Feldman MW. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. Evolution, medicine, and public health. 2019;2019(1):26-34

Borrell LN, Elhawary JR, Fuentes-Afflick E, Witonsky J, Bhakta N, Wu AH, et al. Race and genetic ancestry in medicine–a time for reckoning with racism. New England Journal of Medicine. 2021;384(5):474-80

Coop G, Przeworski M. Lottery, luck, or legacy. A review of "The Genetic Lottery: Why DNA matters for social equality". Evolution; International Journal of Organic Evolution. 2022;76(4):846