3.2 Population structure: II. More about admixture

People often conceptualize human populations as static, homogeneous groups, perhaps with recent mixing in the age of modern travel.

*But genetics teaches us that population mixing – known as **admixture** – is ubiquitous. Evidence of population mixing, both ancient and recent, has been found practically everywhere that scientists have looked for it. Here we explore theory and methods for studying admixture.*

Figure 3.23: **Moai sculpture from Rapa Nui (Easter Island)** *in eastern Polynesia. Later in this chapter we discuss how admixture analysis sheds light on the peopling of Rapa Nui. Credit: Aurbina, [Link] Public Domain.*

Recent population admixture. I've emphasized that populations have been continually splitting and merging throughout human evolution. These population mergers are known as **admixture** events – referring to the sudden mixing of distinct genetic groups.

The concept of admixture is closely related to *migration*, but in population genetics 'migration' usually refers to low levels of gene flow continuously over many generations as opposed to abrupt mixing events [407].

Many populations in the world are **recently admixed** (within the last ∼100–1000 years, say). For example, in the US, many individuals who identify as Native American have recent ancestry from both Native American and European groups; Hispanics often have recent Native American, European, and occasionally African ancestry; and most African Americans have both recent African and European ancestry [408].

Meanwhile, most populations are also products of deeper admixture. For example, ancient DNA tells us that what we might think of as a relatively homogeneous group, modern Europeans, are actually a mixture of at least three highly divergent human groups that no longer exist, plus admixture from Neanderthals.

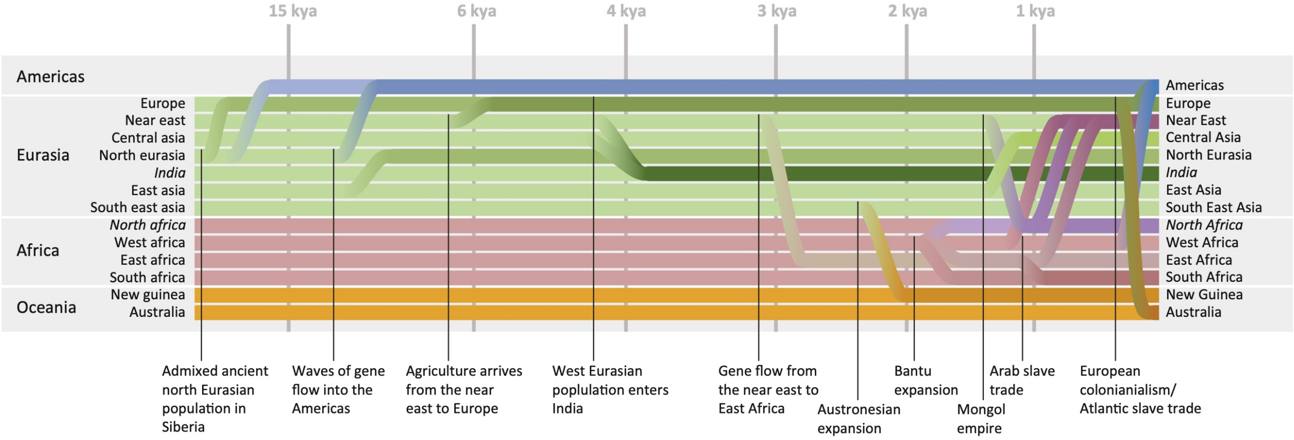


Figure 3.24: Examples of genetically-documented admixture events. The flow chart shows an assortment of major admixture events from the past 20 KY of human history. *Credit: Figure 2 from Joseph Pickrell and David Reich (2014) [Link].*

Models of Admixture. To understand admixture, let's start with a simple scenario: Two populations mix in a single mixing event t generations in the past. We write the initial mixing proportions as w from population 1 and $1 - w$ from population 2. We assume that mating occurs at random with respect to ancestry, and that chromosomal blocks are selectively neutral [409] [a].

In the first generation after the mixing, each chromosome comes entirely from one population or the other; but over time recombination breaks the chromosome segments into smaller and smaller ancestry blocks:

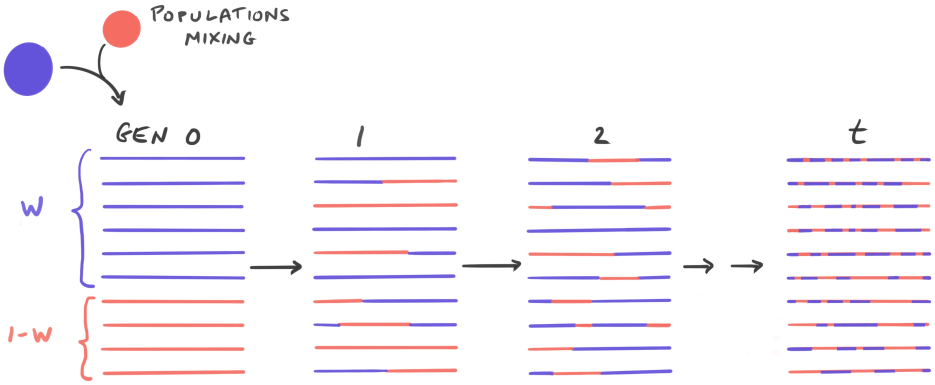


Figure 3.25: **A basic admixture model.** *The cartoon shows ancestry patterns in a population after admixture, for a single chromosome in many individuals. As the chromosomes recombine with one another, the ancestral blocks get smaller over time.*

You can think of the genome of a person in generation t as stitching together blocks of chromosomes from different ancestors who lived at the time of the initial mixture event. Whenever successive blocks come from ancestors with different ancestries (one red, and one blue), this results in an ancestry switch [410].

How large are these unrecombined blocks at generation t? In each generation, recombination events occur at a rate of 1 per Morgan, by definition [b]. So after t generations, we get on average t recombination events per Morgan, and the average size of the blocks is $1/t$ Morgans (or equivalently $100/t$ cM). For example, after $t = 10$ generations, the average block size is 10 cM, a bit less than 10 megabases.

Lastly, how does this relate to the Structure/Admixture model of the last chapter, in which we treated each SNP independently? It turns out that if we mainly care about estimating genome-wide ancestry fractions (Q), it's ok to ignore the details of the ancestry blocks. This is because in a genome-wide SNP set, we'll find about the right fraction of SNPs in blue blocks versus red blocks so we still get an unbiased estimate of Q [411]. But as we'll see, by using a more detailed model, we can learn a great deal about the admixture process itself.

Admixture in African Americans. One of the most studied examples of recent admixture is among African Americans. During the 18th and 19th Centuries, as a result of the transatlantic slave trade, there was extensive gene flow from European slave owners into the African American population.

Typical African American individuals carry around 20% European ancestry in a series of blocks of ~10–15 MB. You can see this illustrated in the

[a] *Notice the analogy between w and q: we use w for mixing proportions of populations and, as before, q for the ancestry proportions of individuals.*

[b] *Recall that Morgans are a measure of recombination distance (Chapter 2.3). Recombination distances are usually reported in centiMorgans (cM); 1 cM contains an expected 0.01 recombination events per generation.*

left-hand panel below, where the blue blocks represent chunks of European ancestry within a single African American genome:

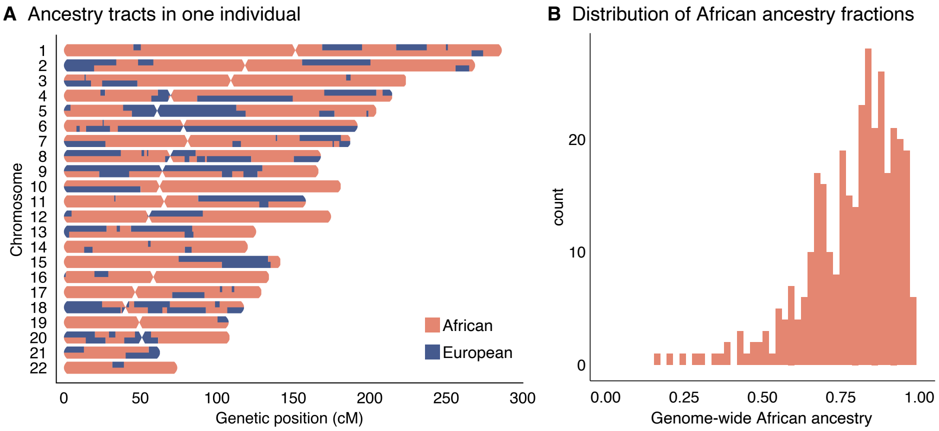


Figure 3.26: **Admixture in African American individuals. A.** *Inferred ancestry blocks in a single African American individual. The colored segments on each chromosome show estimated ancestry of each of the two homologs. Phasing between blocks is generally unknown.* **B.** *Distribution of African ancestry in a sample of individuals from the MESA cohort. Average African ancestry in this cohort is 79%.* *Credit: Unpublished figure kindly contributed by Roshni Patel, CC BY 4*

The highest rates of admixture into the African American population occurred in the early-to-mid 1800s, roughly 6 generations before the present [412]. This means that the average size of European blocks is about $100/6 \approx 17$cM (or roughly 13 MB) [c].

One last point is that the human genetics literature usually discusses admixture with a certain detachment, but it's important to consider that most of the European admixture component in African Americans reflects the genetic legacy of slave ownership [413] [414]. It's been estimated that the pedigree of a typical African American individual contains around 50 distinct European ancestors [415].

Methods for detecting and measuring admixture. In the remainder of this chapter, we'll discuss three main types of methods for studying admixture. We'll end with two examples involving admixture in Native Americans and in Polynesia, respectively, and revisit these ideas with human archaic admixture in Chapter 3.4 [d]:

- **Chromosome painting:** detection of admixture blocks, as shown in the African American example above. Works best when the blocks are very large and/or the population allele frequencies are quite diverged;
- **Decay of admixture LD:** measurement of the genome-wide effect of admixture on LD. Powerful for dating admixture events;
- **Covariance of allele frequencies:** robust detection of admixture events can be powerful even for subtle signals.

Chromosome Painting. How can we infer ancestry blocks in admixed individuals, as in Panel A above? This process is known as **chromosome painting** and is used for many applications in recently admixed populations [416].

[c] *In terms of our model, w, representing African ancestry, is 0.8, and t is 6.*

[d] *Some of the upcoming methods are quite specialized for admixture studies and if you prefer you can skip to the examples.*

To formalize this, consider the admixture process for one chromosome in a single admixed individual. We denote this individual's genome-wide ancestry as q_1 from population 1, and q_2 from population 2 (where $q_1 + q_2 = 1$). I'm going to assume that we can treat each of the two homologs in an individual one-at-a-time. (In practice we usually don't know the haplotype phase of SNPs – i.e., which SNP alleles come from which homolog at large chromosomal distances – so in data analysis we need to extend the algorithm to deal with this phase uncertainty.)

We introduce a vector, Z, where each element z_l is either 1 or 2, to indicate which population this homolog is from at SNP l. As we scan along the chromosome, Z forms a sequence with occasional switches corresponding to ancestral recombination events: for example, the z values might be

$$1\ 1\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2. \tag{3.11}$$

We model Z using a probability model known as a Markov chain that tells us what to expect for z at SNP $l+1$, given z at SNP l. Specifically, we assume that the ancestry of the first SNP on the chromosome, z_1, is a random draw from this individual's ancestry proportions:

$$\Pr(z_1 = 1) = q_1 \tag{3.12}$$

$$\Pr(z_1 = 2) = q_2 \tag{3.13}$$

We then define the transition probabilities. Let r_l be the recombination distance in Morgans between SNP l and $l+1$. Then the expected number of recombination events between the two SNPs in the last t generations is $r_l t$. The probability of having zero recombination events in this interval –i.e., that the two SNPs come from the same ancestral block – is $e^{-r_l t}$. We compute the transition probabilities as follows:

$$\text{Probability of no recombination} = e^{-r_l t}; \tag{3.14}$$

$$\text{in that case, } z_{l+1} = z_l$$

$$\text{Probability of at least 1 recombination} = 1 - e^{-r_l t}; \tag{3.15}$$

$$\text{then with probability } q_1,\ z_{l+1} = 1, \text{ and otherwise 2.}$$

Importantly, notice that when there *is* a recombination event (Equation 3.15), this does not mean that the ancestry necessarily switches. Instead, the next block is a random draw depending on the individual's ancestry. For an individual with high ancestry in one population (like most African Americans), most block changes do not result in an ancestry switch.

Of course we can't observe Z directly, but the genotype data depend on Z. Whenever a chromosome chunk comes from population 1 the allele frequencies in that block reflect frequencies in population 1, and otherwise they reflect population 2. This means that we can infer which parts of the chromosome come from each population, but of course we cannot detect block changes that did not also result in an ancestry switch.

This model lends itself to a statistical algorithm for chromosome painting called a Hidden Markov Model (HMM), that decodes the ancestry along each chromosome [417] [418].

Admixture creates LD at different scales. The presence of these ancestry blocks has important implications for one of our old friends: **linkage disequilibrium** (LD). Recall from Chapter 2.3 that LD refers to correlations between the genotypes at different SNPs.

When I introduced LD, I discussed how it can be generated by the fact that nearby sites tend to share much of the same coalescent genealogy (i.e., the ancestral recombination graph), going back over hundreds of thousands of years. It's important to be aware that population structure also generates LD, but at much larger genetic distances, and it decays much faster with time.

It's helpful to conceptualize LD in admixed populations at three different length scales [419]:

- **Background LD** (up to $\sim$100 Kb): This is the type of LD that is found within populations due to the structure of coalescent genealogies (Chapter 2.3). The length-scale depends on the relative rate of recombination to coalescence ($4Nr$).
- **Admixture LD** (up to multi-Mb): As we discussed above, genomes of admixed individuals consist of blocks of ancestry from one parent population or the other. When the parent populations have different allele frequencies, this creates correlations between SNPs in the same block. This type of LD depends on the size of ancestry blocks (which have a mean size of $100/t$ cM) and can be detectable over background LD for tens of generations.
- **Mixture LD** (Genomewide): When different individuals in the population have different ancestry (i.e., different q) this creates LD even between unlinked SNPs. This is the kind of signal used by Structure and PCA, and it decays very rapidly – at rate of $(1/2)^t$ – becoming undetectable within a few generations of random mating.

The decay of admixture LD provides an important quantitative signal that we can use to date admixture events, as follows.

Optional details: The decay of Admixture LD. Recall that the most basic measure of LD is denoted D. If we have two SNPs with alleles A, a at the first SNP, and B, b at the second SNP, then D is computed $p_{AB} - p_A p_B$, where p_{AB} is the frequency of the AB haplotype, and p_A and p_B are the frequencies of A, and of B, separately. If genotypes at the two SNPs are independent, then $D = 0$.

We extend the notation to look at what happens to D following an admixture event. As before, populations 1 and 2 mix together in proportions w and $1 - w$. We denote D in the two parental populations as D_1 and D_2, respectively, and $D_m^{(0)}$ denotes D in generation 0 in the mixed population. D_1 and D_2 reflect any background LD prior to population mixing.

The first key result [420] is that, immediately following admixture, initial LD in the admixed population is:

$$D_m^{(0)} = \underbrace{wD_1 + (1-w)D_2}_{\textit{Background LD}} \quad + \quad \underbrace{w(1-w)\delta_A\delta_B}_{\textit{Mixture/Admixture LD}} \tag{3.16}$$

where δ_A is the allele frequency difference for allele A between the two populations: $\delta_A = p_{1,A} - p_{2,A}$; and similarly δ_B measures the frequency differences at the other SNP. The background LD term represents LD that is present in the initial populations, and is only relevant for SNPs that are extremely close. The mixture/admixture LD term measures LD that is created by mixing together individuals with different allele frequencies genomewide.

Next, consider pairs of SNPs that are further apart than the typical scale of background LD: let's say more than about 100 kb. Then D_1=D_2= 0. Recall from Chapter 2.3 that if the two SNPs are separated by a recombination distance r, then recombination will reduce the admixture LD D_m at a rate $(1-r)$ per generation. If the SNPs are very far apart, or even unlinked, then $r \sim 0.5$, and $D_m^{(t)}$ decays to zero within a few generations of random mating.

But if the SNPs are within a few centiMorgans in the genome, LD is maintained for tens of generations – hundreds or even thousands of years. We can predict the decay of admixture LD as follows. After t generations we have:

$$D_m^{(t)} = \underbrace{(1-r)^t \Big[w(1-w)\delta_A\delta_B \Big]}_{\textit{Admixture LD}}. \tag{3.17}$$

Rearranging, we see that

$$\frac{D_m^{(t)}}{\delta_A\delta_B} = (1-r)^t \times w(1-w) \tag{3.18}$$

To fit this model to data, we can consider all pairs of SNPs in the genome that are separated by some recombination distance r. The intercept at $r = 0$ is an estimate of $w(1-w)$ and allows us to estimate the initial mixing proportions, and the decay rate is a function of the mixture time t. For more on these methods see [421]. We'll show an application next.

This theory of admixture LD can be used to estimate the date of admixture events, provided that we know the allele frequencies in the parental populations.

One example of this considers the genetic impact of the so-called Bantu expansion in Africa. It's long been hypothesized, based on linguistic data, that a group of populations known as Bantu Speaking Peoples expanded south and west out of Cameroon starting around 3000 years ago. Genetic evidence confirms this expansion, showing that it created a wave of admixture with local populations, as Bantu-ancestry populations spread south and east across most of sub-Saharan Africa [422] [e].

The example below comes from work by Joe Pickrell and colleagues [423] studying the genetic structure of the distinctive Khoisan populations of southern and eastern Africa. Pickrell et al showed that all of these popu-

[e] *You can read more about African genetic history in the next chapter.*

lations carry Bantu admixture, to varying degrees.

This plot shows the decay of admixture LD as a function of recombination distance in a Khoisan population, the Jul'hoan (black dots).

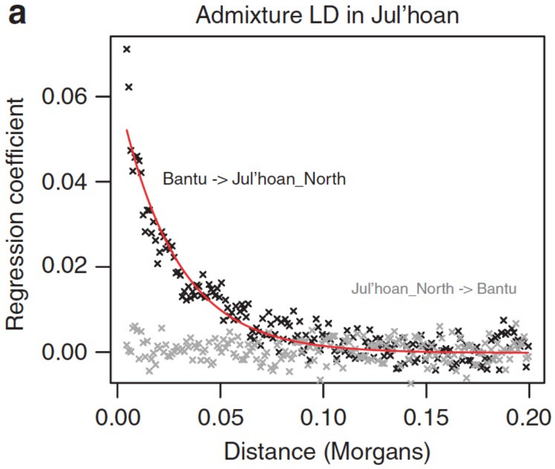


Figure 3.27: **Admixture LD in a southern African population.** *The plot shows the decay of admixture LD in Jul'hoan as a function of recombination distance (black dots). The grey data show a negative control: testing for admixture from Jul'hoan into Bantu. The plotted values are computed using a function similar to the left-hand side of Equation 3.18, averaged over many pairs of SNPs at similar recombination distances; the data use allele frequencies from Khoisan and Bantu as the two donor populations. Credit: Figure 2a from Joseph Pickrell et al 2012. CC BY-NC-SA 3.0.*

As you can see, admixture LD in the Jul'hoan extends to about 5 cM (i.e, around 6 MB) – far beyond the typical range of background LD [f]. The rate of decay of admixture LD reflects the timing of admixture: the data show that w=6% of Jul'hoan ancestry comes from Bantu admixture about 35 generations (or 1000 years) ago.

We'll see another example of this technique in Chapter 3.4 where it's used to date the admixture of humans and Neanderthals (Figure 3.86).

Ancient admixture, allele frequency covariances, and F statistics. The last important class of methods focuses on allele frequencies instead of LD. The key idea is that if a population carries ancestry from multiple sources then, immediately after admixture, its allele frequencies are a weighted average of the source populations:

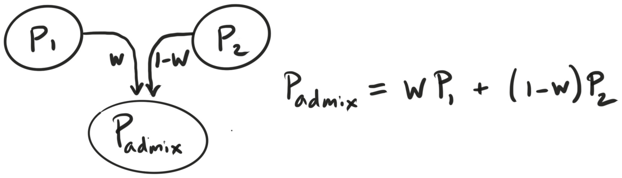


Figure 3.28: **Expected allele frequency in an admixed population** *is simply a weighted average of the frequencies in the source populations. Here, w and $1 - w$ are the mixing proportions.*

For a recent admixture event, we can estimate the allele frequencies in the source populations and use those to estimate w. For example we could use this to estimate the amount of Bantu admixture in the example above.

But if admixture took place further back in time, then it gets more complicated because the population frequencies are changing through drift and we don't know the allele frequencies of the parental populations at the time of admixture. **As usual, we'll need a model to understand this**[g] [424].

As the first building block for a model, we consider a variant with

[f] *One way to think about this signal is that if an individual carries an allele that is more common in Bantu at one SNP, then they are more likely to carry a Bantu allele at a nearby SNP; this correlation decays over about 5 cM, a lengthscale that reflects the date of admixture.*

[g] *This model extends the Nicholson-Donnelly model of drift from Chapter 2.4 to trees of populations.*

frequency p_0 in an ancestral population, and frequency p_1 in a modern population. Recall that for a neutral allele, the expected change in frequency down each branch is zero:

$$\mathrm{E}[p_1 - p_0] = 0 \qquad (3.19)$$

and the expected squared change in frequency (i.e., the variance) is

$$\mathrm{E}[(p_1 - p_0)^2] \approx \frac{T}{2N_e} p_0(1 - p_0) \qquad (3.20)$$

where T is the elapsed time and N_e is the effective population size [425].

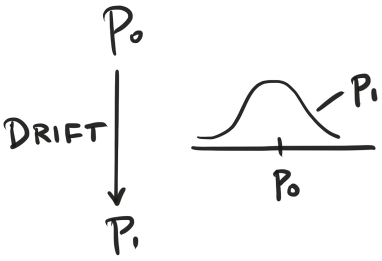


Figure 3.29: **Genetic drift from an ancestral frequency $\mathbf{p_0}$.** *The cartoon shows the distribution of p_1 given p_0. The variance of p_1 is proportional to the elapsed time divided by population size; we use the variance as a measure of branch length.*

Noting that the mean squared change in frequency (left-hand side of Equation 3.20) down a branch is proportional to the elapsed time T, we'll refer to the mean squared change as a **branch length** [426].

Next, how does this look if we have multiple populations related by a population tree? A key implication of our model is that **drift along different branches is independent** [427] [428]:

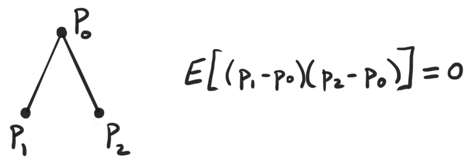


Figure 3.30: **Drift is independent along different branches,** *implying that the covariance of p_1 and p_2 is 0.*

Notice that the quantity $\mathrm{E}[(p_1 - p_0)(p_2 - p_0]$ measures the covariance of frequencies between populations 1 and 2, so in this tree we expect zero covariance between these two populations.

What about for populations that share a branch? Using the assumption that drift is independent along each branch we can show the following:

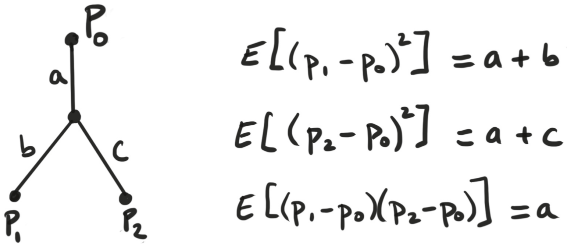
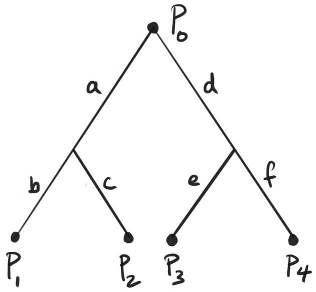


Figure 3.31: **Covariance along shared branches.**

As you can see, the drift from root to tip is given by the sum of the branch lengths: for example $a + b$ for population 1. And the covariance for two populations is given by the sum of the shared branch-length: namely a in the case of populations 1 and 2 together.

We can use these simple rules to compute variance-covariance matrices for any bifurcating tree. For example:



	P_1-P_0	P_2-P_0	P_3-P_0	P_4-P_0
P_1-P_0	$a+b$	a	0	0
P_2-P_0	a	$a+c$	0	0
P_3-P_0	0	0	$d+e$	d
P_4-P_0	0	0	d	$d+f$

Figure 3.32: **Variance-covariance matrix for a simple tree.** *The entries in the matrix show the expected values of the product of the row labels times column labels: e.g.,* $\mathrm{E}[(p_1 - p_0)(p_2 - p_0)] = a$.

In short, we calculate expected values of the form $\mathrm{E}[(p_i - p_0)(p_j - p_0)]$ *as the sum of the branches shared between i and j.* This simple calculation rule extends directly to larger trees as well.

What about admixture? The tree below shows a similar relationship among the populations, but with an admixture event contributing a fraction w of the ancestry to Population 3:

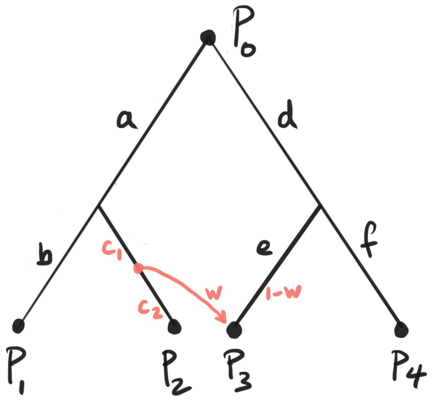


Figure 3.33: **A simple admixture graph.** *The way to interpret this is that a fraction of w of the ancestry in population 3 comes via the left-hand branch, and* $1 - w$ *from the right-hand branch.*

The red admixture arrow effectively creates alternative paths that alleles can take: an allele observed in Population 3 may have come down the left-hand branch (with probability w) or down the right-hand branch (with probability $1 - w$). If we consider someone from Population 3, different parts of their genome come from a mixture of two different trees, like this:

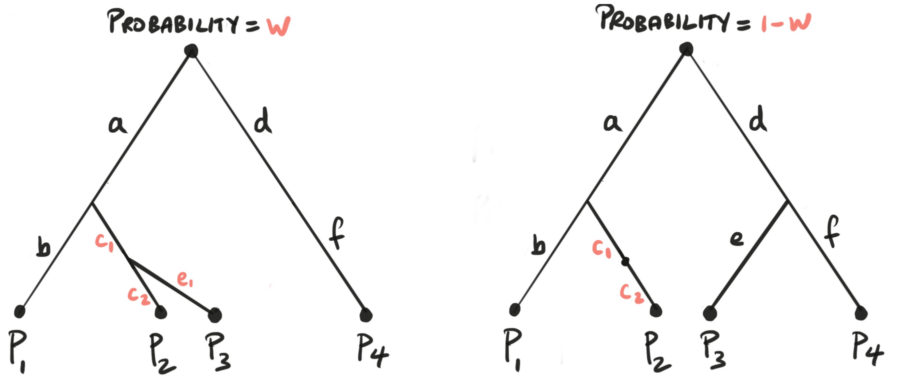
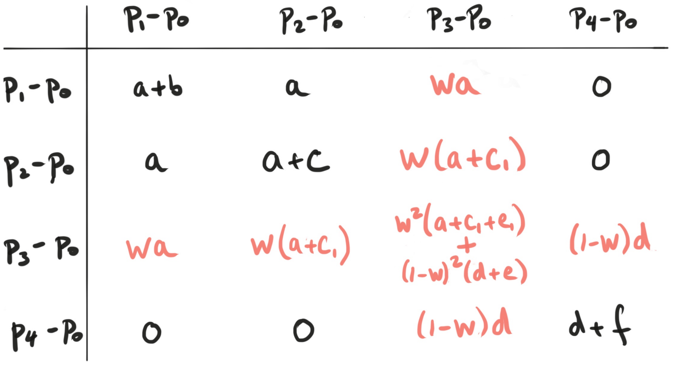


Figure 3.34: **The admixture graph above can be viewed as a weighted mixture of these two trees.** *Different parts of an individual's genome can follow one or other graph, with probabilities w and* $1 - w$*, respectively.*

Now the covariances are a weighted sum of the covariances in these two different trees:



	P_1-P_0	P_2-P_0	P_3-P_0	P_4-P_0
P_1-P_0	$a+b$	a	wa	0
P_2-P_0	a	$a+c$	$w(a+c_1)$	0
P_3-P_0	wa	$w(a+c_1)$	$w^2(a+c_1+e_1) + (1-w)^2(d+e)$	$(1-w)d$
P_4-P_0	0	0	$(1-w)d$	$d+f$

Figure 3.35: **Covariance matrix for the admixture graph above.** *Entries in red reflect the admixture event and entries in black do not.*

Note: $c_1 + c_2$ is written as c to match the previous matrix.

and the variances are weighted sums of variances [429]. Notice how the admixture graph results in a more complicated covariance structure than the tree without admixture. *In particular, there is no model without admixture that could provide a good fit to these data.*

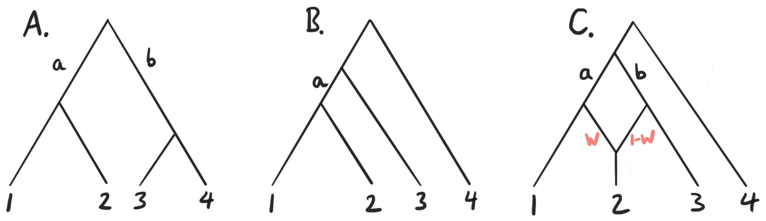
Now that we have developed some theory, how can we use it to detect admixture in data?

One important approach known as the **F statistics** was developed by Nick Patterson, David Reich, and colleagues [430] [h]. We focus here on one of the F statistics, F_4, which is defined in terms of the allele frequencies in four populations at a time:

$$F_4(Pop_1, Pop_2;\ Pop_3, Pop_4) = \mathrm{E}[(p_1 - p_2)(p_3 - p_4)] \qquad (3.21)$$

where $Pop_1...Pop_4$ represent four population samples, $p_1...p_4$ are the corresponding sample allele frequencies, and where the expectation is taken over all SNPs [431] [432].

The key idea here is that if $Pop_1 + Pop_2$ form an independent clade from $Pop_3 + Pop_4$ then the drift is independent in the two clades (Panels A and B below). This means that F_4 should be zero. But if there's gene flow between clades then $F_4 \neq 0$ regardless of how we label the populations (Panel C) [433]:



	A.	B.	C.
$F_4(1,2;3,4)$	0	0	$(1-w)b$
$F_4(1,3;2,4)$	$a+b$	a	$wa+(1-w)b$
$F_4(1,4;2,3)$	$a+b$	a	wa

Figure 3.36: **Examples of F_4 tests.** *The expected value of $F_4(1,2;3,4)$ is given by the overlap of the path from $Pop_1 \rightarrow Pop_2$ with the path from $Pop_3 \rightarrow Pop_4$. Panels* **A** *and* **B** *show examples for trees without admixture, for different permutations of the populations in the F test. In both cases, one of the permutations results in $F_4 = 0$. In Panel* **C***, Pop_2 is formed by admixture of ancestors of Pop_1 and Pop_3. No F_4 test equals zero.*

We may not know in advance which populations might be most closely related, in which case we can compute F_4 while permuting the labels of Pop_1, Pop_2, Pop_3, and Pop_4. *If none of the permutations have $F_4 = 0$ this implies that there is no simple unadmixed tree for the four populations. Hence, F_4 provides a formal test for whether four populations can be fit by a simple tree.*

[h] *We'll cover a related concept, D statistics, when we get to archaic admixture in Chapter 3.4.*

F-tests have been used widely to test for admixture among populations, and show that past admixture among human populations is virtually ubiquitous: in well-powered data sets it is usually difficult to find populations that do *not* show evidence for past admixture [434].

Admixture graphs. While *F* tests provide a powerful framework for detecting admixture in up to four populations at a time, we often want to understand the relationships among larger groups of populations. We can represent these relationships using so-called *admixture graphs*. These graphs represent the process of both splits and admixture in the history of a set of populations.

One main approach to estimating admixture graphs is to search for graphs that satisfy the *F* test results for all combinations of populations (implemented by Admixtools and Mixmapper [435]). A second approach is to estimate graphs directly from the sample covariance matrix of population allele frequencies (implemented by TreeMix and AdmixtureBayes [436] [437]).

One example using a global set of populations is shown below. The left-hand panel shows a traditional branching tree of population relationships; however this is a poor fit to the data as many population sets fail *F* tests. The right-hand-panel adds additional admixture events, shown by dashed lines. For example the graph shows North Asian populations as a mixture of East Asian ancestry and a component related to the ancestors of Native South Americans (Surui and Karitiana). Middle Eastern/North African populations are modeled as a mixture of West African (Mandenka and Yoruba) and European ancestry [438].

Figure 3.37: **Gratuitous image of a knight fighting a snail.** *A great mystery of medieval books is that the margins were often decorated with unrelated images of knights battling snails! Nobody knows why – but I like to think they were simply to add humor to dense sections of text.* *Credit: [Link], from the Gorleston Psalter, England, 1310-1324, British Library MS 49622, f. 193v.*

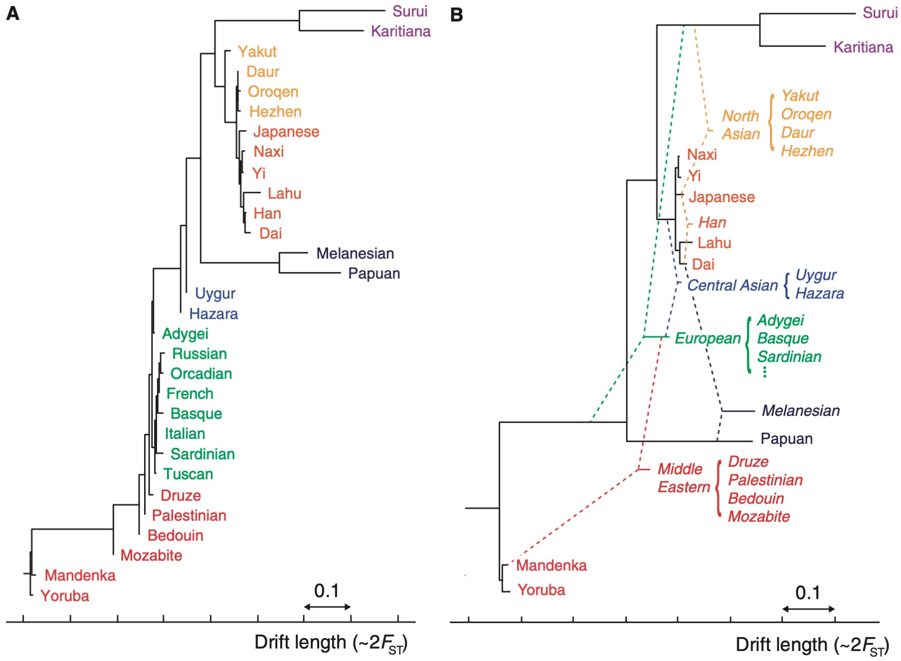


Figure 3.38: **Estimated admixture graph of human populations. A.** *Traditional neighbor-joining tree of 30 human populations.* **B.** *Admixture graph estimated using the tool Mixmapper. Dashed lines represent admixture events.* *Figure 4 from Mark Lipson et al (2013) [Link] CC-BY-NC*

***F* tests as a 'tracer dye' for ancient population movements.** F tests are a powerful tool for studying ancient admixture and gene flow. Even small amounts of ancestry inherited from a divergent population can leave a signal in the allele frequency data tens of thousands of years later. David Reich has likened the signals of divergent alleles to *tracer dyes*, explaining that they are *"like the heavy metals injected into patient's veins in hospitals to track the paths of their blood vessels in an MRI scan"* [439].

One striking application where these have been used is to understand the **peopling of the Americas** [440]. Native Americans descend from one or more migration events through a region called Beringia that connects Eastern Siberia to present-day Alaska. Siberia and Alaska are now separated by a narrow strip of water called the Bering Straits, but during the last Ice Age, when sea levels were lower, Beringia formed a land bridge between the two continents.

When the first Americans reached Beringia, they were initially blocked from spreading southward by large ice sheets. As the ice receded, around 16,000 years ago, humans spread rapidly to colonize the north and south American continents [441].

While many details of this story remain hotly debated, we'll address one question here: *Who were the source population(s) for the first Americans?*

It was long assumed that native Americans descend from the ancestors of modern Siberians and other east Asians. This seems intuitive based on geography. It's also supported by the traditional tree of populations shown above in Panel A of Figure 3.38 – the two south American populations (Surui and Karitiana at the top) are closest to the northeast Asian populations, including the Yakut from Siberia.

But Panel B of Figure 3.38 shows something very surprising: an additional admixture branch from the ancestors of Europeans. This connection between Native Americans and Europeans has shown up in several different analyses. How should we understand this?

The observation was largely enigmatic until the 2014 sequencing of **ancient DNA** from remains of a boy who lived 24,000 years ago in south-central Siberia [442]. This boy's genome (known as Mal'ta-1 or MA-1) is a representative of an early Siberian population that we now call the **Ancient North Eurasians**. Remarkably, genetic data show that Ancient North Eurasians (or related populations) contributed ancestry to both modern Europeans and native Americans!

But that wasn't the end of the surprises. The next year, in 2015, Pontus Skoglund and colleagues [443] reported a result that is perhaps even harder to make sense of: *some indigenous populations from the Amazon share a small but statistically significant amount of allele frequency covariance with present-day Australasians*, including native Australians, Papuans, and Andaman Islanders!

Here's a simplified admixture graph that represents our current understanding of the sources of native American populations:

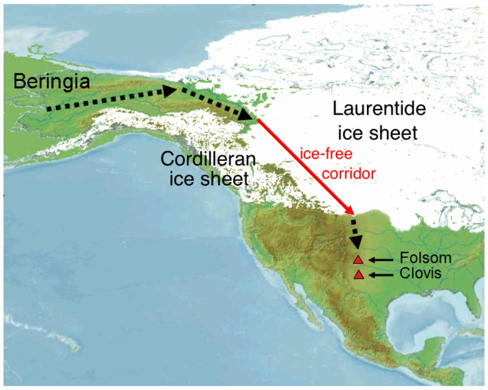


Figure 3.39: **Migration path of early native Americans through Beringia.** *The ice sheets blocked southward migrations until ∼16 KYA. Paleo-American sites at Clovis and Folsom date to ∼13 KYA.* Credit: Roblespepe [Link] CC BY-SA 3.0.

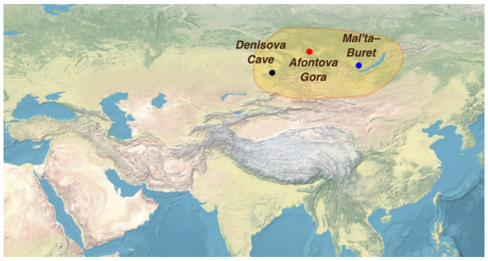


Figure 3.40: **Known range of Ancient North Eurasians,** *who lived ∼20–30 KYA, and contributed ancestry to both native Americans and Europeans.* Credit: [Link] [Link] CC BY SA 4.0.

Figure 3.41: **Paleolithic engraving of mammoth on ivory**, *by Ancient North Eurasians, Mal'ta, Siberia.* Credit: José-Manuel Benito [Link] Public Domain.

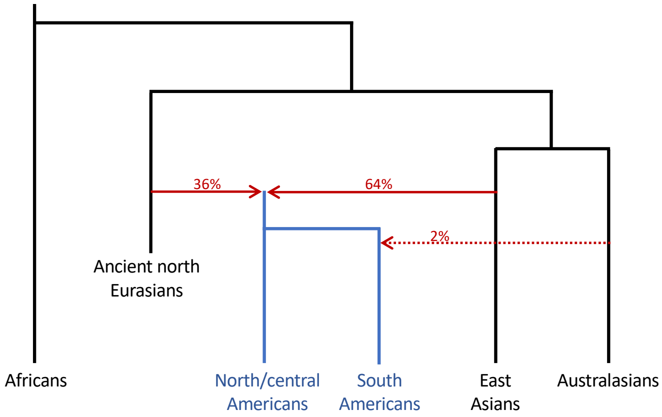


Figure 3.42: **Admixture model for the origin of native American populations.** *Red lines indicate admixture edges, with weights given as percentages.* *Redrawn based on Figure 2 of Skoglund et al (2015) [Link] and Figure 3 of Araújo Castro e Silva et al (2021) [Link].*

Note: *The model was fitted using specific populations or samples including Mbuti (central African), MA1 (ANE), Pima (north American), Mixe (central American), Surui and Karitiana (south American), Han (east Asian), and Onge Andaman Islanders (Australasian).*

The affinity with Australasians remains highly enigmatic. The shared ancestry is unrelated to Polynesians, arguing against a migration event across the Pacific to south America. The signal is not found in other native American populations. The signal shows a rapid decay of LD, indicating that it is an ancient event (though it has not been dated precisely).

In thinking how to interpret this, it's helpful to remember the lesson of the European affinity to native Americans: it's not that Europeans themselves are ancestors of native Americans, but that both Europeans and native Americans carry a tracer dye of Ancestral North Eurasian ancestry.

Together these points suggest the existence of an unknown source population that contributed both to gene flow through Beringia, as well as dispersing southward to contribute to Australasian populations. The migration of this unknown ancestry through Beringia may have been mixed with other first Americans, but it must have been a separate event from the main dispersal that gave rise to north and central American populations, given that this signal is only found in the Amazon.

As you can see, the "tracer dye" of DNA has presented us with both insight and puzzles about the deep relationships among human groups. Hopefully these puzzles will be resolved in future with additional ancient DNA samples!

We close the chapter with an example where admixture analysis has resolved another long-standing question in anthropology.

Case study of admixture: Native American and European admixture in Polynesia. Our last example comes from a study of more-recent admixture, in Pacific Islanders.

Polynesia is a vast region of the central and south Pacific comprising more than a thousand remote islands. Polynesia was settled by seafarers who spread eastward across the sea from south-east Asia and Melanesia over a period of several thousand years, reaching eastern Polynesia by about 1000 years ago. This general model is supported both by archaeological and linguistic evidence as well as genetic similarity between the Polynesians and populations in Oceania [444].

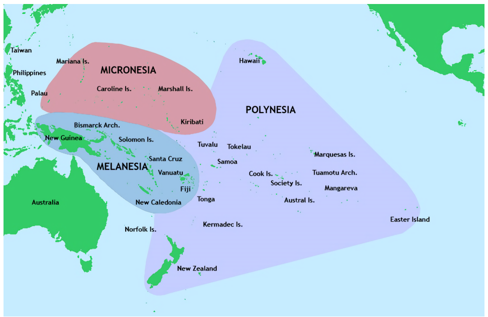


Figure 3.43: **Islands of the central and south Pacific.** *Easter Island (Rapa Nui) is located in eastern Polynesia, 4,300 miles east of New Zealand and 2,100 miles west of Chile.* *Credit: Kahuroa [Link]. Public Domain.*

Although the main ancestry in Polynesia is from Oceania, there has long been debate about whether there may also have been prehistoric contact between eastern Polynesia and native central or south Americans. One line of evidence for early contact is the presence of south American crops including sweet potato in Polynesia, but it has also been argued that sweet potato may have spread independently of humans [445].

Moreover, the ocean distance between Polynesia and south America would have been extremely daunting for early sea travelers: the eastern-most island in Polynesia is Rapa Nui, also known as Easter Island, located some 2,100 miles from the west coast of South America. In 1947, the Norwegian explorer Thor Heyerdahl famously navigated a hand-built balsa wood raft from Peru to eastern Polynesia to argue for the plausibility of south American migration into Polynesia, but claims of a link between prehistoric Americans and Polynesia have remained highly controversial.

Figure 3.44: **Ancestor figure from Rapa Nui.** *Crafted in wood, bird bone, obsidian. Credit: [Link] Public Domain.*

But this is a question that should be resolvable using the techniques of this chapter. In 2020, a group led by Alex Ioannidis and colleagues collected genome-wide SNP data from 807 Polynesians representing 17 island populations, and compared these to data from Pacific coast native Americans and Europeans [446].

As expected, many of the islands had signals of European admixture following post-colonial contact. But, much more remarkably, the study also found compelling evidence for a small but clear genetic contribution from native Americans! This is illustrated here:

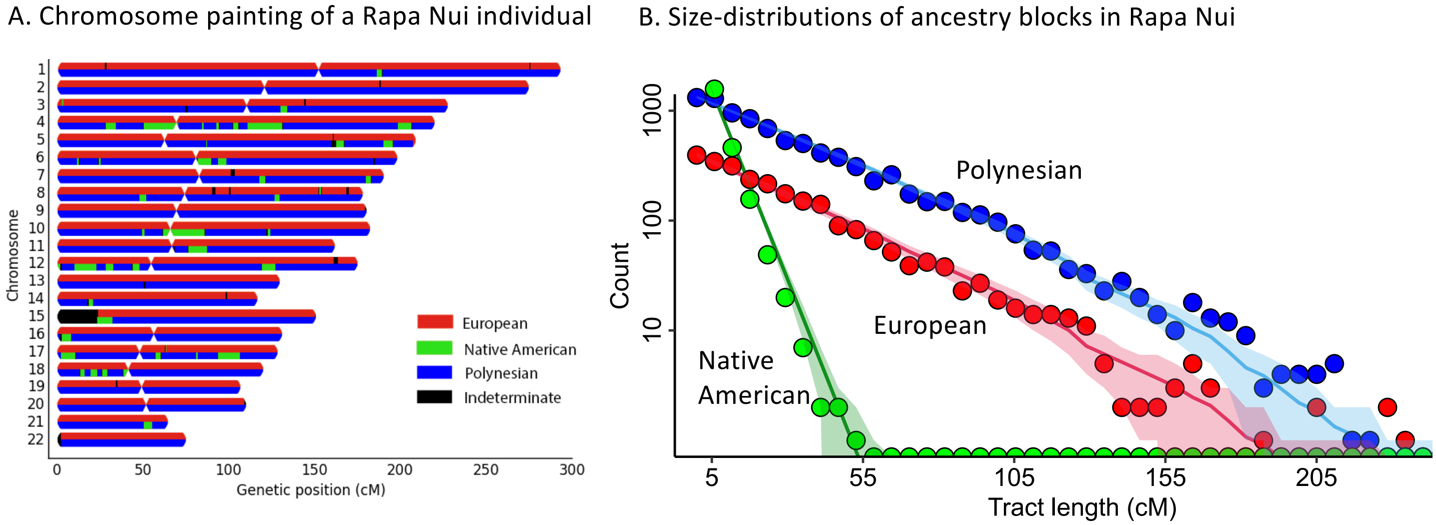


Figure 3.45: Admixture blocks in Rapa Nui (Easter Island). A. *Chromosome painting for a single Rapa Nui individual who is approximately 50% Polynesian (blue) and 50% European (red). Short blocks of Native American ancestry (green) are embedded within the Polynesian segments.* **B.** *Ancestry block (tract)-length distributions. The plot shows total numbers of ancestry blocks at each size, summed across* 64 *Rapa Nui individuals.*

Credit: Unpublished images kindly provided by Alex Ioannidis, CC BY 4. Original study: [Link].

Panel A illustrates the results of chromosome painting [447] in a typical individual from Rapa Nui, where they had the largest sample size. Notice the Native American ancestry blocks in green, embedded in a genome that is otherwise mainly of mixed Polynesian and European ancestry.

Panel B shows the overall size distribution of blocks within the population. As you can see, the Native American ancestry blocks are generally much shorter than the European and Polynesian ancestry blocks, indicating that they date to an earlier admixture event: the authors estimate 20 generations for Native American, compared to 6 generations for Europeans. The authors found their best-fitting source for the Native American ancestry was in northern South America–most likely a population related to modern-day Colombians [i].

The study found evidence for earlier native American admixture into other Polynesian populations; the earliest date estimate is 1150 AD on the island of South Marquesas. These earlier dates roughly coincide with the time that Polynesians were completing their spread to the farthest reaches of Polynesia, making it possible that some islands may even have had American habitation before Polynesians arrived [448].

This story shows another example of the fantastic power of DNA to act as a tracer dye for past events in human history, both recent and ancient – in this case resolving a question that anthropologists had argued about for nearly a century.

In the last two chapters we have discussed methods for studying contemporary population structure and admixture. In the next two chapters we'll outline the huge advances in using population genetics to study the deeper relationships among human populations.

[i] *Notice that the large ancestry blocks in Rapa Nui (∼10 cM) show that Native American admixture was quite recent, in contrast to the Australasian gene-flow into South Americans where admixture LD is very short (∼0.2 cM), showing that admixture event was ancient.*

Notes and References.

[407] Another distinct usage that you may come across is that specific individuals with recent ancestors in different populations are also referred to in the genetics literature as *admixed*.

[408] Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. The American Journal of Human Genetics. 2015;96(1):37-53

Verdu P, Pemberton TJ, Laurent R, Kemp BM, Gonzalez-Oliver A, Gorodezky C, et al. Patterns of admixture and population structure in native populations of Northwest North America. PLoS Genetics. 2014;10(8):e1004530

Jordan IK, Rishishwar L, Conley AB. Native American admixture recapitulates population-specific migration and settlement of the continental United States. PLoS Genetics. 2019;15(9):e1008225

[409] The assumption that there is no selective advantage to blocks of one ancestry or the other works well in modern human populations; however as we'll discuss in the ancient DNA chapter, Neanderthal admixture blocks were deleterious on average in anatomically modern humans.

[410] We define a block here as a contiguous chromosomal region that is inherited from the same ancestor at the time of admixture. We assume that the ancestry of successive blocks is independent. This means that successive blocks can potentially come from the same ancestry, in which case we cannot easily detect block switches. But for a minority ancestry (e.g., European ancestry in African Americans, or Neanderthals in non-Africans) it's unlikely that successive blocks both come from the minority ancestry, in which case the block size is approximately the same size as the typical length of ancestry tracts.

[411] Structure/Admixture produces unbiased estimates of Q with admixture, provided that it can estimate the population frequencies accurately. It can usually do ok at this if there are at least some non-admixed representatives of each source population, but it has a hard time if there are only admixed individuals.

[412] A paper by Soheil Baharian et al (2016) provides detailed models of admixture in African Americans. They fit a model with 2 separate pulses of admixture, estimated at 1740 and 1863, and being of roughly equal size. Of course in reality, admixture would be more continuous and variable.

Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, et al. The great migration and African-American genomic diversity. PLoS Genetics. 2016;12(5):e1006059

Mooney JA, Agranat-Tamir L, Pritchard JK, Rosenberg NA. On the number of genealogical ancestors tracing to the source groups of an admixed population. Genetics. 2023;224(3):iyad079

[413] This article about Michelle Obama's roots puts a human face on one person's ancestors and relatives: [Link]. You can also get a numerical perspective on this from Mooney et al (2023) paper cited above.

[414] As you might expect given this history, most of the European ancestry carried by African Americans comes from European males. This can be inferred from the genetic data by comparing the fraction of European ancestry on the autosomes versus European ancestry on the X chromosome. We inherit 1/2 of our autosomal ancestry from male ancestors, but only about 1/3 of our X chromosome ancestry from male ancestors (The precise calculation is a bit complicated:

Goldberg A, Rosenberg NA. Beyond 2/3 and 1/3: the complex signatures of sex-biased admixture on the X chromosome. Genetics. 2015;201(1):263-79

This means that if European admixture was male-biased, then we should see higher European ancestry on the autosomes than on the X. This is, in fact, the case: for example, Bryc et al (2015) estimated that for African Americans in their sample about 37% of the male ancestry is from Europeans, compared with only 10% of the female ancestry

[415] Mooney et al (2023) wrote *"we infer that if all genealogical lines of a random African-American born during 1960–1965 are traced back until they reach members of source populations, the mean over parameter sets of the expected number of genealogical lines terminating with African individuals is 314 (interquartile range 240–376), and the mean of the expected number terminating in Europeans is 51 (interquartile range 32–69)."*

[416] For examples of chromosome painting see the African American plot above, and an example from Polynesia in Figure 3.45A. For a global application see

Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. science. 2014;343(6172):747-51

[417] 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

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, et al. Methods for high-density admixture mapping of disease genes. The American Journal of Human Genetics. 2004;74(5):979-1000

[418] Note that with standard genotyping data, successive SNPs are usually within a few Kb. Hence this basic model can be confused by background LD among nearby SNPs (see below). More complicated methods that deal with background LD perform better than the simplest HMM:

Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics. 2013;93(2):278-88

Hilmarsson H, Kumar AS, Rastogi R, Bustamante CD, Montserrat DM, Ioannidis AG. High resolution ancestry deconvolution for next generation genomic data. bioRxiv. 2021:2021-09

[419] Falush et al (2003)

[420] Chakraborty R, Weiss KM. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proceedings of the National Academy of Sciences. 1988;85(23):9119-23

[421] Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genetics. 2011;7(4):e1001373

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192(3):1065-93

Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring admixture histories of human populations using linkage disequilibrium. Genetics. 2013;193(4):1233-54

See also supplement p33 of

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, et al. The genetic prehistory of southern Africa. Nature Communications. 2012;3(1):1143

[422] Fortes-Lima CA, Burgarella C, Hammarén R, Eriksson A, Vicente M, Jolly C, et al. The genetic legacy of the expansion of Bantu-speaking peoples in Africa. Nature. 2024;625(7995):540-7

[423] Pickrell, Patterson et al (2012), cited above.

[424] There's a long history of these interest in graph models of drift, going back to work such as Cavalli-Sforza and Edwards (1967) and Felsenstein (1982). In the modern era, work by David Reich and collaborators starting around 2008 re-initiated interest in this topic, with particular focus on studying admixture. The presentation here relies heavily on Pickrell and Pritchard (2012) which merged some of the classic graph-based work with the approaches to admixture pioneered by the Reich group (cited below).

Cavalli-Sforza LL, Edwards AW. Phylogenetic analysis. Models and estimation procedures. American journal of human genetics. 1967;19(3 Pt 1):233

Felsenstein J. How can we infer geography and history from gene frequencies? Journal of Theoretical Biology. 1982;96(1):9-20

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

[425] In what follows we'll use the variance assumption to estimate branch lengths on trees, while recognizing that it's not strictly accurate because Nicholson-Donnelly is only an approximation, and only really accurate for small $T/2N_e$. This means that the branch lengths for long branches are biased downward, but we wouldn't expect this bias to produce false signals of admixture.

Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, Donnelly P. Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2002;64(4):695-715

[426] That is, if the frequency for SNP l the top of a branch is $p_{i,l}$ and is $p_{j,l}$ at the bottom of the branch, then the branch length is defined as the expected value of

$$\frac{1}{L}\sum_{l=1}^{L}(p_{i,l}-p_{j,l})^2. \tag{3.22}$$

Under the Nicholson-Donnelly approximation this is approximately $(T/2N_e)\cdot\overline{p_{0,l}(1-p_{0,l})}$, where $\overline{p_{0,l}(1-p_{0,l})}$ is the average computed over all SNPs in the data set. It's important to note that, defined in this way, the numerical value of the branch length depends on the choice of SNPs – for example it would probably be longer using SNPs from a genotyping array (which are ascertained to be common variants) than computed for sequencing data. This definition is convenient, but for settings where we want to interpret the branch lengths as a measure of drift we can rescale the branch lengths at the end of the analysis by dividing by $\overline{p_{0,l}(1-p_{0,l})}$; then the branch lengths are estimators of $T/2N_e$.

[427] By definition, if random variables X and Y are independent then $\mathrm{E}[XY]=0$.

[428] This assumption of independence is central to admixture tests by this method, so you should wonder if it's a good assumption. One thing you might worry about is whether selection might cause frequency changes to be correlated between particular pairs of populations. This is probably not a major problem in practice because selection would need to affect a large number of sites in a coordinated direction, and in a specific subset of populations. In principle that might happen via polygenic adaptation, or if selection against deleterious sites became much stronger (eg due to increasing pop-

ulation size in a few populations). But it seems unlikely that these would be quantitatively strong effects in practice. Another plausible artifact would be if SNP ascertainment is biased in specific ways: e.g., by identifying SNPs for genotyping from two distantly related populations. This could plausibly bias toward intermediate frequencies in both populations, but again this is probably not a major concern in real settings.

[429] The weights for variances are of the form w^2.

[430] F statistics are introduced in the supplement of Reich et al (2009), and described comprehensively by Patterson et al (2012). Patterson et al is comprehensive but challenging to read, so I also recommend Peter (2016), Lipson (2020), and the documentation for AdmixTools 2 [Link]

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461(7263):489-94

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192(3):1065-93

Peter BM. Admixture, population structure, and F-statistics. Genetics. 2016;202(4):1485-501

Lipson M. Applying f4-statistics and admixture graphs: Theory and examples. Molecular Ecology Resources. 2020;20(6):1658-67.

[431] Reich and Patterson define three different F statistics:

$$F_2(Pop_1;\ Pop_2) = \mathrm{E}[(p_1 - p_2)^2] \tag{3.23}$$

$$F_3(Pop_1;\ Pop_2, Pop_3) = \mathrm{E}[(p_1 - p_2)(p_1 - p_3)] \tag{3.24}$$

$$F_4(Pop_1, Pop_2;\ Pop_3, Pop_4) = \mathrm{E}[(p_1 - p_2)(p_3 - p_4)] \tag{3.25}$$

Here F_2 measures the total branch length between populations 1 and 2. It is mathematically related to F_{ST}, thus motivating use of the F nomenclature here. F_3 is a test of whether Pop_1 is an admixture of Pop_2 and Pop_3: specifically, if F_3 is significantly negative this implies that Pop_1 is admixed. F_4 tests for independence between the clade of Pop_1 and Pop_2 versus Pop_3 and Pop_4. If no permutation of the four populations produces $F_4 = 0$ (up to sampling error) then this implies that there is no unadmixed tree that can explain the data. See Peter (2016) for details.

[432] You might reasonably worry about sampling error for all the analyses in this section. For example, we never observe p_1 directly, but instead we collect a smaller sample of m diploid individuals, say, and from that we estimate $\hat{p_1}$. Happily, this winds up not being a major problem. You can think of the sampling process as like adding a one-generation bottleneck down to m individuals at the end of each terminal branch on the tree. This will increase the estimated branch lengths (by $1/2m$ for each tip). Luckily, the F_3 and F_4 statistics, and the covariances, are still unbiased (albeit noisier) because the binomial sampling error is uncorrelated between population samples. If we want to, we can correct the branch length biases according to sample size.

[433] We can relate F_4 to the covariance statistics in a tree with ancestral allele frequency p_0, as follows:

$$\begin{aligned} F_4(Pop_1, Pop_2;\ Pop_3, Pop_4) &= \mathrm{E}[(p_1 - p_2)(p_3 - p_4)] && (3.26)\\ &= \mathrm{E}[((p_1 - p_0) - (p_2 - p_0))((p_3 - p_0) - (p_4 - p_0))] && (3.27)\\ &= V_{1,3} - V_{2,3} - V_{1,4} + V_{2,4} && (3.28) \end{aligned}$$

where $V_{i,j} = \mathrm{E}[(p_i - p_0)(p_j - p_0)]$. Recall from the main text that $V_{i,j}$ can be computed as the sum of branches *shared* by i and j. Then you can show with some simple examples that if the pair Pop_1 and Pop_2 form a distinct clade with respect to Pop_3 and Pop_4 then $F_4 = 0$ (Pritchard and Pickrell 2012; Supplementary Eq 18).

[434] For an early example see Figure 3

Pickrell JK, Reich D. Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics. 2014;30(9):377-89.

[435] These tools have been developed mainly by the Reich lab across a series of papers (Patterson et al 2012, Lipson et al 2013). Their tree construction process is semi-supervised by the user.

Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B. Efficient moment-based inference of admixture parameters and sources of gene flow. Molecular Biology and Evolution. 2013;30(8):1788-802

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

Nielsen SV, Vaughn AH, Leppälä K, Landis MJ, Mailund T, Nielsen R. Bayesian inference of admixture graphs on Native American and Arctic populations. PLoS genetics. 2023;19(2):e1010410

[437] Something you might wonder about for these methods is that the theory was presented in terms of the ancestral frequencies p_0, which is unknown in practice. Pickrell and Pritchard (2012) showed how to write the observed sample covariance matrix in terms of tree parameters.

It's also worth noting that admixture graphs are generally under-constrained, such that only a subset of the gene flow events can be detected, and many details may not be identifiable.

[438] The TreeMix version of this graph is broadly similar despite using a different algorithm: see Figures 3 and 4 of Pickrell and Pritchard (2012).

[439] p181 of David Reich's book:

Reich D. Who we are and how we got here: Ancient DNA and the new science of the human past. Oxford University Press; 2018

[440] For reviews of this topic with citations to the key literature see

Skoglund P, Reich D. A genomic view of the peopling of the Americas. Current opinion in genetics & development. 2016;41:27-35

Willerslev E, Meltzer DJ. Peopling of the Americas as inferred from ancient genomics. Nature. 2021;594(7863):356-64

[441] Skoglund and Reich (2016), cited above

[442] Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014;505(7481):87-91

[443] Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, et al. Genetic evidence for two founding populations of the Americas. Nature. 2015;525(7567):104-8

[444] Kayser M, Brauer S, Cordaux R, Casto A, Lao O, Zhivotovsky LA, et al. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. Molecular biology and evolution. 2006;23(11):2234-44

Hudjashov G, Endicott P, Post H, Nagle N, Ho SY, Lawson DJ, et al. Investigating the origins of eastern Polynesians using genome-wide data from the Leeward Society Isles. Scientific reports. 2018;8(1):1823

For a nice overview of the work highlighted here see

Wallin P. Native South Americans were early inhabitants of Polynesia. Nature. 2020;583:524-5

[445] Muñoz-Rodríguez P, Carruthers T, Wood JR, Williams BR, Weitemier K, Kronmiller B, et al. Reconciling conflicting phylogenies in the origin of sweet potato and dispersal to Polynesia. Current Biology. 2018;28(8):1246-56

[446] Ioannidis AG, Blanco-Portillo J, Sandoval K, Hagelberg E, Miquel-Poblete JF, Moreno-Mayar JV, et al. Native American gene flow into Polynesia predating Easter Island settlement. Nature. 2020;583(7817):572-7

See also

Ioannidis AG, Blanco-Portillo J, Sandoval K, Hagelberg E, Barberena-Jonas C, Hill AV, et al. Paths and timings of the peopling of Polynesia inferred from genomic networks. Nature. 2021;597(7877):522-6

[447] Chromosome painting was performed using RFMix

Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. The American Journal of Human Genetics. 2013;93(2):278-88.

[448] Ioannidis et al (2020) and Wallin (2020), cited above.