Chapter 3.1-3.4

Siyi Wang

May 12, 2025

Chapter 3.1

0.1 The Clustering

Clustering is used for grouping genomic data and for identifying groups of individuals with similar genotypes to analyze genetic population structure and ancestry estimation. A 1997 study by Mountain and Cavalli-Sforza analyzed genetic distances using hierarchical clustering, revealing global population structures. They encoded genotypes as follows:

- AA $\rightarrow 0$
- Aa $\rightarrow 1$
- aa $\rightarrow 2$

They computed the simple multi-SNP genotype distance by summing the absolute values of differences across L SNPs between two individuals. The results showed that individuals from the same continent tend to cluster together, while the differences between continents remain small.

0.2 No-Admixture Model

The **no-admixture model** assumes that each individual's genome is derived entirely from a single ancestral population. This simplification does not account for historical population mixing and is thus limited in practical applications. In this model, each individual is assigned to one of K populations, with the ancestry matrix Q containing a single 1 in one column and 0 elsewhere.

While useful for conceptualizing genetic structure, this model fails to capture real-world genetic diversity due to extensive historical admixture among human populations. More robust approaches, such as the admixture model, provide a more accurate representation of human ancestry.

0.3 Admixture Model

Admixture Model acknowledges that individuals may inherit genetic material from multiple ancestral populations. Instead of a discrete assignment, this model allows for fractional ancestry contributions from multiple populations.

0.3.1 Admixture in the Ancestry Matrix, Q

The ancestry proportion matrix Q extends the no-admixture framework by representing continuous values rather than discrete assignments. Each individual's ancestry satisfies:

$$\sum_{k=1}^{K} q_{i,k} = 1 \tag{1}$$

where $q_{i,k}$ represents the fraction of ancestry that individual i derives from population k.

For instance, an individual with **25% ancestry from population 1** and **75% from population 3** is represented as:

$$Q_i = (0.25, 0, 0.75, 0, \dots) \tag{2}$$

0.3.2 Genotype Probability Under the Admixture Model

Under the admixture model, an individual's genotype at SNP l is determined by a weighted sum of allele frequencies across multiple populations:

$$\text{Personal Allele Frequency}(i,l) = \sum_{k=1}^{K} q_{i,k} p_{k,l} = Q_i \cdot P_l \tag{3}$$

where P is the allele frequency matrix, and $p_{k,l}$ represents the derived allele frequency at SNP l in population k.

The genotype probabilities follow:

$$Pr(g_{i,l} = 0) = (1 - Q_i \cdot P_l)^2 \tag{4}$$

$$Pr(g_{i,l} = 1) = 2(Q_i \cdot P_l)(1 - Q_i \cdot P_l) \tag{5}$$

$$Pr(g_{i,l} = 2) = (Q_i \cdot P_l)^2 \tag{6}$$

0.3.3 Ancestry Estimation and Model Optimization

Estimating the parameters Q and P requires computational methods such as:

- **Averaging columns**: Given a labeled population sample, estimate the population frequencies (dividing by 2 for allele frequencies). $P^{1,l}$ is the estimated frequency in population 1 at SNP l.
- **Bayesian Inference**: To maximize the likelihood function:

$$\hat{Q}, \hat{P} = \arg\max_{Q,P} \prod_{i=1}^{N} Pr(G_i | Q_i, P) \tag{7}$$

- **Expectation-Maximization Algorithm**:
  1. **Expectation Step**: Under fixed P, update the ancestral proportions Q of individuals to maximize the posterior probability.
  2. **Maximization Step**: Under fixed Q, calculate the allele frequencies P of the population.
  3. **Repeat iterations** until convergence.

0.4 Applications of the Admixture Model

The admixture model has played a critical role in understanding human genetic variation. Notable applications include:

- **Human Genome Diversity Project**: Demonstrated that most populations exhibit genetic contributions from multiple ancestral sources, reflecting the complexity of migration, structure, and history.
- **Principal Components Analysis (PCA)**: Revealed that linear dimension reduction in high-dimensional data:

$$G = \Lambda F + E \tag{8}$$

  where G represents genotypes, Λ is the individual loadings matrix, F is the factor matrix, and E represents errors. The first two columns of the loading matrix generate the PC1 vs. PC2 plot, which can indicate ancestral populations.
- **Haplotype-based clustering**: Detects recent shared ancestry by analyzing shared haplotype segments rather than just SNP frequencies, offering high geographic resolution.

Chapter 3.2

0.5 Ancestry Block Breakdown

The plot below shows that two populations, after a few generations, with their chromosome segments, break into smaller ancestry blocks at a rate of 1 per Morgan. African Americans, on average, have 20% European ancestry, with genetic block sizes of approximately 10-15 MB.

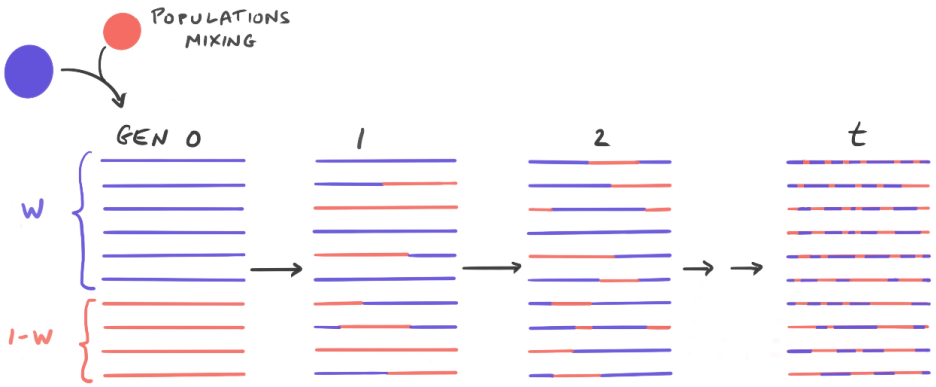

Figure 1: Basic Admixture Model

Admixture Analysis: Core Methods and Concepts

Understanding population admixture involves three main analytical approaches, each suited to different timescales and data types.

1. Chromosome Painting

Chromosome painting is used to trace ancestry along the genome by identifying chromosomal segments derived from distinct ancestral populations. It is especially effective when the source populations are genetically divergent.

Key Principle:

- The genome of an admixed individual consists of contiguous blocks inherited from multiple ancestral sources.
- A **Hidden Markov Model (HMM)** is used to infer these ancestry blocks:
  1. Let $Z_l = 1$ if SNP l is from population 1, and $Z_l = 2$ if from population 2.
  2. Transition probabilities depend on recombination rates and admixture time:
  $$\Pr(Z_{l+1} = Z_l) = e^{-r_l t} \tag{9}$$
  where r_l is the genetic distance in Morgans and t is the number of generations since admixture.
  3. Genotype data and known allele frequencies in reference populations inform the inference of local ancestry.

2. Admixture Linkage Disequilibrium

This method uses the decay of linkage disequilibrium (LD) over time to date admixture events.

Key Equations:

Initial LD after admixture:

$$D_m^{(0)} = wD_1 + (1-w)D_2 + w(1-w)\delta_A\delta_B \tag{10}$$

LD decay over generations:

$$D_m^{(t)} = (1-r)^t \cdot w(1-w)\delta_A\delta_B \tag{11}$$

Here:

- w and $1-w$: admixture proportions from two source populations.
- δ_A, δ_B: allele frequency differences.
- r: recombination rate.

By fitting decay curves of LD across loci with known recombination distances, one can estimate the time t since admixture occurred.

3. Covariance of Allele Frequencies

This method infers ancient admixture events using allele frequency correlations across populations.

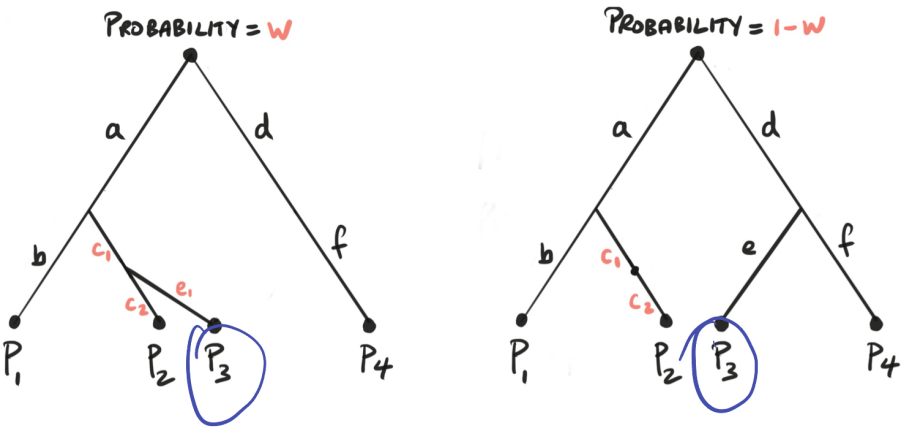

Figure 2: Admixture Graph

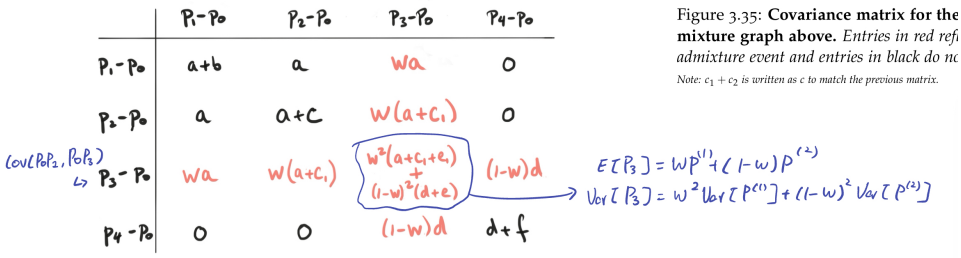

Figure 3.35: **Covariance matrix for the mixture graph above.** *Entries in red refl admixture event and entries in black do nc*

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

Figure 3: Admixture Graph

Concepts and Calculations:

- Immediately after admixture, allele frequencies are a weighted average:

$$p_{\text{admix}} = wp_1 + (1-w)p_2 \tag{12}$$

- Over time, genetic drift alters allele frequencies. The expected change is:

$$E[p_1 - p_0] = 0 \tag{13}$$

$$E[(p_1 - p_0)^2] \approx \frac{T}{2N_e} p_0(1-p_0) \tag{14}$$

This defines the **branch length** (genetic drift distance) between ancestral population and the current one.

- The covariance between allele frequencies of two populations (i and j) sharing ancestry is:

$$\text{Cov}(p_i, p_j) = E[(p_i - p_0)(p_j - p_0)] \tag{15}$$

This equals the sum of shared drift segments (branches) on the population tree.

F4 Statistic:
A formal test for admixture involving four populations:

$$F4(\text{Pop1}, \text{Pop2}; \text{Pop3}, \text{Pop4}) = E[(p_1 - p_2)(p_3 - p_4)] \tag{16}$$

If $F4 \neq 0$, it indicates that gene flow occurred and that the populations cannot be explained by a simple unadmixed tree.

Chapter 3.3

Skeletal remains and artifacts can be used in human prehistory. Examples include footprints in Laetoli, Tanzania, and Neanderthal skulls in Europe. However, these remains are often fragmentary, and reconstructing their evolutionary relationships is difficult.

Recently, by comparing genomes from modern populations, population genetics can reconstruct relationships going back 1–2 million years. However, this approach cannot provide detailed information about physical appearance. Fossil and genetic evidence both indicate Africa as the origin point of our species. There were at least three *Out-of-Africa* events, but only the last one, around 50 KYA, led to the global spread of modern humans. During this expansion, *Homo sapiens* replaced other archaic hominins like *Homo erectus*, Neanderthals, and Denisovans.

Multiregional Hypothesis vs. Recent African Origin

Multiregional Hypothesis: Suggests modern humans evolved in parallel in multiple regions from local archaic groups like *Homo erectus*, with some gene flow maintaining species unity.

Recent African Origin (RAO): Proposes that modern humans evolved in Africa and spread globally around 50 KYA, largely replacing local archaic populations.

Strong genetic evidence supports RAO:

- mtDNA and Y-chromosome data show the most recent common ancestor (MRCA) lived in Africa around 160 KYA and 140 KYA respectively.
- The deepest branches in the genetic tree are exclusively African.
- Genetic diversity is highest in Africa.

Genetic Diversity and Expansion Pattern

Genetic diversity decreases with increasing distance from Africa, while linkage disequilibrium (LD) increases. This supports the Serial Founder Model, which posits that new populations were founded by small groups moving out from Africa, each carrying only a subset of the previous population's genetic diversity.

As a result, Native American populations today have the lowest genetic diversity, while Central African hunter-gatherers have the highest.

Modeling Methods:

- For simple models: $F_{ST} \approx \frac{T}{2N_e}$ to estimate population sizes and allele frequency divergence times.
- For complex models: Simulate datasets and compare summary statistics to observed data (flexible but high-dimensional).
- Use summary statistics like the Site Frequency Spectrum (SFS) to approximate likelihood. It ignores linkage but captures substantial information.
- Methods like `ARGweaver` and `Relate` reconstruct genome-wide ancestral trees, including recombination, to model population history (computationally intensive for large datasets).

Pairwise Sequentially Markovian Coalescent (PSMC) Model: Infers historical effective population size (N_e) using only a single diploid genome. Particularly useful for detecting population bottlenecks and expansions.

Deep Structure in African Populations

African populations show deeper splits than any between non-African populations. Four major lineages have been identified:

- Southern African Khoisan (known for click languages)
- Central African rainforest hunter-gatherers (notably short stature)
- Eastern African pastoralists and hunter-gatherers
- Western African Bantu-speaking agriculturalists

PSMC and cross-coalescence analysis suggest these groups diverged around 200–300 KYA. Some models infer the existence of *ghost lineages* (e.g., Stem 1 and Stem 2), with origins 1–2 million years ago. These ancient branches may no longer exist today but contributed genetically to current populations.

This detailed structure within Africa emphasizes the continent's complexity and richness in human evolutionary history. It also underscores that modern humans emerged from a diverse and structured set of populations within Africa.

Chapter 3.4

1. Ancient DNA (aDNA)

Ancient DNA refers to DNA retrieved from biological remains (bones, teeth, sediments) long after death. Due to degradation over time, it is often fragmented and chemically damaged. The study of aDNA allows scientists to reconstruct population history, migration, and interbreeding events.

Fragmented and chemically modified DNA, easily contaminated by modern DNA, and difficult to obtain from warm, humid environments.

2. Timeline of Key Discoveries and Events

Year / Period	Event
1984	First aDNA was sequenced by Allan Wilson's lab.
1997	Pääbo's team sequenced Neanderthal mtDNA.
2006	First Neanderthal nuclear DNA fragments were sequenced.
2010	First Neanderthal genome: Non-Africans have ∼1.5–2% Neanderthal DNA.
2008–2014	Discovery & genome of Denisovans from a finger bone in Siberia.
2020	Sediment DNA confirms Denisovans in the Tibetan Plateau.

3. The Story of Neanderthals and Humans

mtDNA studies (1997–2007): Neanderthal mtDNA lies outside modern variation.

Genome sequencing in 2010 confirmed gene flow between species (introgression): non-Africans have 1.5–2% Neanderthal ancestry. Note: Gene flow between populations is admixture.

Gene flow likely occurred in the Middle East shortly after humans left Africa.

4. Ancient DNA Technologies

Challenges: DNA decay, contamination, single nucleotide errors, and microbial DNA (Bacteria/Fungi).

Solutions:

- Cold/dry extraction to prevent contamination
- Clean rooms, and UV treatment to avoid contamination.
- Petrous bone (inner ear) is best for DNA yield.
- Sediment DNA enables detection without fossils.

5. Denisovans

First discovered via mtDNA from a finger bone (Denisova Cave, Altai).

Diverged from Neanderthals ∼400 KYA; all three (humans, Neanderthals, Denisovans) share an African ancestor, and mtDNA shows differences from all around 1 MYA.

D-statistic (ABBA-BABA Test):

Measures allele sharing across genomes. 'B' matches Neanderthal, and 'A' is an outgroup chimpanzee.

Ideally, D should equal 0. Since the null hypothesis states no introgression between Neanderthals and Europeans.

$$D = \frac{\#ABBA - \#BABA}{\#ABBA + \#BABA}$$

It doesn't depend on demography since it cannot produce ABBA or BABA.

Strong signal of Neanderthal introgression into non-Africans, not Africans.

6. Selection and Introgression

Neanderthal and Denisovan ancestry is systematically depleted near genes. Most archaic DNA was removed by natural selection, especially near functional regions shown by B statistic plot.

Two hypotheses model:

- Genetic incompatibilities due to divergence from two distinct species (Neanderthals and Denisovans).
- Higher genetic load in Neanderthals/Denisovans due to small effective population sizes (Low Ne). A bunch of deleterious alleles after admixture to remove these.

Adaptive introgression examples: EPAS1, immunity-related genes

7. The Complex Ancestry of Modern Europeans

Ancient DNA data reveal that modern Europeans are the product of multiple ancient population mixtures and replacement events. Rather than being descended in a straight line from the first *Homo sapiens* who arrived in Europe, modern Europeans derive ancestry from at least three major ancestral sources:

The Three Ancestral Components:

- **Western Hunter-Gatherers (WHG)** — Dominant in Europe around 14,000 years ago, after the last Ice Age. Replaced earlier populations like Oase 1.
- **Neolithic Anatolian Farmers** — Migrated from the Middle East ∼9,000 years ago, bringing agriculture. They largely replaced WHG in many regions of Europe.
- **Yamnaya Steppe Pastoralists** — Originating in the Russian steppe, they expanded ∼5,000 years ago, introducing horses, wheeled transport, and Indo-European languages. They mixed with existing populations instead of fully replacing them.

Key Evidence and Insights:

- Ötzi the Iceman (5.3 KYA) was genetically closer to modern Sardinians than central Europeans, because Sardinia remained less affected by the later Steppe expansions.
- Ancient individuals lie outside the modern European genetic variation, showing that modern genomes are blends of now-extinct populations.
- PCA plots of modern Europeans reflect admixture from these three groups.

Process:

- After the Ice Age: WHG dominates.
- ∼9 KYA: Anatolian Farmers expand across Europe, replacing much of WHG.
- ∼5 KYA: Yamnaya Steppe people spread across Europe, adding a third layer of ancestry.

The genetic continuity is the exception, not the rule. Ancient DNA reveals repeated waves of migration, replacement, and admixture — far more dynamic than previously assumed.

Questions

How to choose the "Best" number of clusters K? (To avoid oversimplified or overly complex interpretations of population structure.)–Comparing AIC/BIC(model)?