3.3 Inferring human prehistory from genetic data

During the past 30 years population genetics has helped to rewrite our understanding of human prehistory, alongside parallel advances in paleoanthropology and archaeology [449] [a].

This work has shown a fascinating diversity of hominins [b] *that lived simultaneously in Africa, Eurasia, and Oceania during the past 2 million years, until very recently. The main branch of human evolution occurred in Africa, with at least three major migration events out of Africa, the last of which was of modern humans.*

This chapter covers the side-by-side use of fossils and genetics to study the deep history of our species. The next chapter will bring ancient DNA into the story, while focusing on more-recent timescales [450].

[a] *As we shall see, the distribution of genetic variation in modern human populations is a direct result of our species' history, and so these topics influence every other field of human genetics. And, conversely, we can use genetic variation to learn about history.*

[b] *Here the term **hominin** refers to humans and all earlier forms that arose on the human lineage since the common ancestor of humans and chimpanzees.*
***Homo** refers to the genus that includes humans; it appears in the fossil record starting around 2.5 million years ago.*

Bones and stones: the fossil record. In 1856, workers in a limestone quarry in Germany's Neander Valley uncovered 16 bones including a skullcap, buried in layers of clay at a cave entrance [451]. They passed the bones to a local teacher and fossil collector named Johann Fuhlrott. Fuhlrott recognized the bones as something truly extraordinary: human-like, and yet not quite human, with a sloping forehead and prominent brow-ridges. You can see an example in the upper skull on the right.

This chance find was the first recognized discovery of an archaic hominin, and it effectively launched the field of **paleoanthropology**: the study of hominin fossils. Given its distinctiveness, the new fossil was tentatively designated as a new species, *Homo neanderthalensis* – commonly known as **Neanderthals** – within the same genus as modern humans, *Homo sapiens*.

Since 1856, paleoanthropologists have unearthed a stunning array of early hominins. We now know that until just 50 KYA [c], practically yesterday in population genetics terms, there were multiple major hominin lineages in Africa and Eurasia. But within a short timescale these diverse lineages were entirely replaced by modern humans.

[c] *Abbreviations: KYA: thousand years ago; MYA: million years ago.*

Here we ask: How were these early hominins related to each other, and to us? As we shall see, population genetics has played an essential role in tackling these questions.

Figure 3.46: **Contrasting skulls: Neanderthal (top) and early human**. *The two skulls have similar brain capacities but very different shapes.* *Both skulls were found in France, the Neanderthal dates to around 60 KYA [Link], and the human to 30 KYA [Link]. Photo credit Chris Stringer [Link]; fossils at Musée de l'Homme, Paris. Image used with permission.*

How we know what we know. In these two chapters, we'll draw on three main types of evidence, each informative over different timescales, and with different strengths and limitations:
Skeletal remains and other artifacts: These tell us when and where different hominin lineages lived, sometimes allowing insights into their lifestyle and culture. Several key lineages are known only from physical remains. Physical remains can survive for millions of years; however the remains are usually fragmentary, and biological relationships among

groups are heavily debated.

Inference from modern genomes: Population genetics allows us to infer ancestral relationships among modern populations back to ∼300 KYA, and potentially to detect deep population structure as far back as 1–2 MY. But modern genomes don't tell us *where* their ancestors lived, or their physical attributes, so it's difficult to connect genetic signals to specific hominin lineages. And lineages that did not contribute genetic material to modern populations are completely invisible.

Inference from ancient DNA: The study of *ancient* DNA seeks to recover DNA sequences from ancient skeletal remains and sediments. This work has made fantastic contributions to our understanding of recent human prehistory and our relationship with recent archaic hominins including Neanderthals. But DNA degrades over time, thus limiting its use to study older specimens, especially in hot climates. DNA preservation past ∼10,000 years is unusual, and past ∼100,000 years is extraordinary [452].

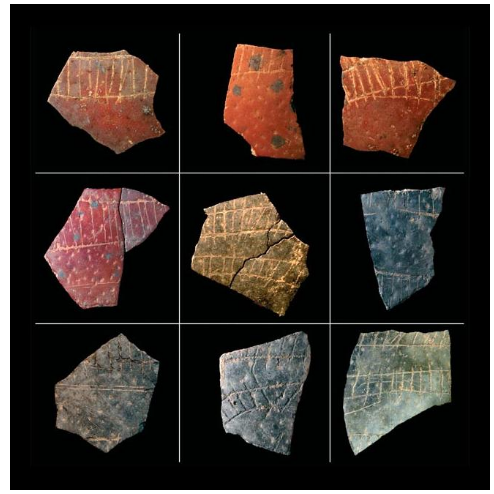

Figure 3.47: **Fragments of engraved ostrich eggshell containers** *from South Africa from 60 KYA. Physical remains can also provide insights into the development of culture; these shells may be an early example of symbolic representations.* *Credit: Pierre-Jean Texier et al (2010) [Link], reproduced from Balter (2010) [Link].*

A crowded family tree. Paleoanthropology has revealed a large cast of characters who lived in Africa, Eurasia, and Oceania during the past 2 million years. In current models, the main trunk of the evolutionary tree was in Africa, but with probably at least three major outward migration events during the past 2 million years.

The figure below shows a model of the major lineages and their relationships [453] [454] [455]. We'll introduce our main characters next:

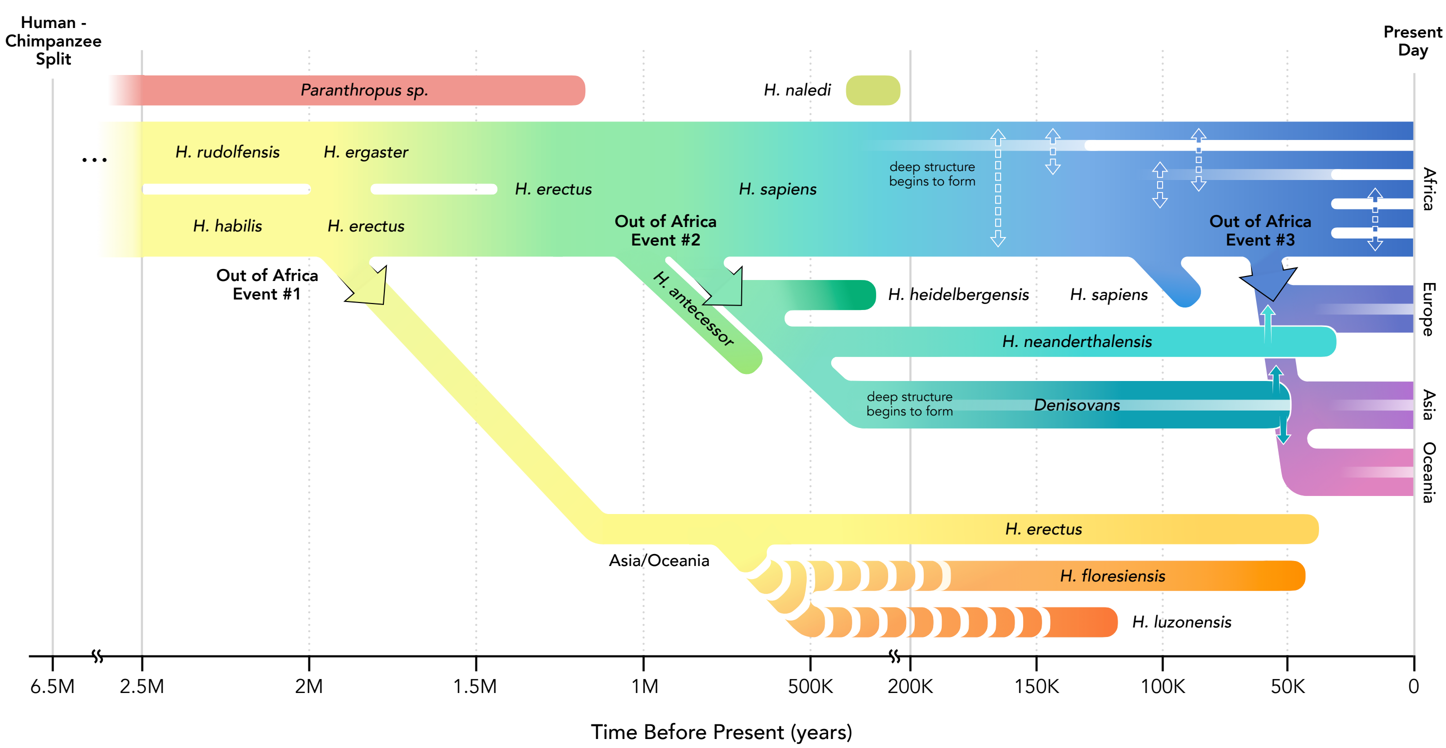

Figure 3.48: A timeline of hominin evolution. *Simplified model showing one interpretation of the relationships among the major hominin lineages during the past 2.5 MY, based on fossil morphology and genetics. Small arrows indicate gene flow between populations. Broken lines leading to H. floresiensis and luzonensis reflect their hypothesized relationship to H. erectus. The direct ancestors of H. naledi are unclear. The Americas and Polynesia are not shown as there is only evidence for recent habitation by modern humans. Credit: Unpublished figure kindly contributed by Alyssa Lyn Fortier. CC BY 4.*

Dramatis personae. We can now set the stage with the major players. The names printed in brown indicate groups that lived within the last 300 KY, and were therefore contemporaries of early *Homo sapiens* [456].

The early hominin lineage (∼6.5–2 MYA), is known from a series of fossils found mainly in Central, Eastern and Southern Africa [457]. The skull shown on the right from Chad, nicknamed *Toumaï*, is dated to ∼7 MYA, near the time of the time of the human-chimpanzee divergence. After this, hominin fossils are mainly classified into two genera: first ***Ardepithecus*** (∼5.8–4.4 MYA), followed by ***Australopithecus*** (∼4.5–2 MYA). *Australopithecus* had a small brain (only slightly larger than modern chimpanzees) but was **the first hominin known to walk truly upright** (see footprints on the right). The late-surviving archaic genus ***Paranthropus*** (∼2.9–1.2 MYA) is a likely descendant of *Australopithecus*, and lived alongside early *Homo* species [458].

Figure 3.49: **The Toumaï skull**, *dated to 7 MYA, from Chad, is from near the time of the human-chimpanzee split. Species name: Sahelanthropus tchadensis; Credit: Didier Descouens [Link] CC BY-SA 4.0*

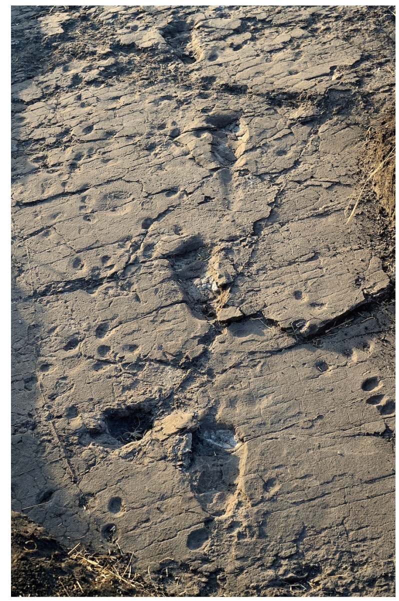

Figure 3.50: **The Laetoli footprints** *dated to 3.6 MYA, from Tanzania, are the first unequivocal evidence for fully bipedal walking. The image shows one of a series of trackways at the Laetoli site. Credit: Fidelis Masao et al [Link] CC BY 4.0*

Early archaic Homo (from ∼2.5 MYA), and archaic descendants, evolved from *Australopithecus*. In Africa, these early Homo are ancestral to the lineage that eventually gives rise to modern humans; meanwhile, *H. erectus* is the first known hominin outside Africa. Notice how late many of these lineages, below, persisted:

• ***Homo habilis***: Africa, ∼2.8–1.6 MYA. *H. habilis*, meaning "handy man" was the first group to make sophisticated **stone tools**, starting ∼2.5 MYA in present-day Ethiopia. Other species that were contemporaries, or slightly later, include early African ***H. erectus*** and (in some classification systems) ***H. rudolfensis*** and ***H. ergaster*** [459].

• ***H. erectus***: Africa, west Asia, east Asia, Indonesia, ∼2.0 MYA–50 KYA. In Africa, *H. erectus* is assumed to be part of the ancestral lineage that gave rise to humans; *erectus* were also **the first hominins to leave Africa**. In 1891, the Dutch scientist Eugène Dubois went to Java in Indonesia on the hunch that humans had evolved from east Asian apes, and that he might be able to find transitional fossils there. His hypothesis was incorrect, but with amazing luck he stumbled onto what were then the oldest-known hominin fossils, now classified as *Homo erectus*.
H. erectus was geographically widespread across Asia, and persisted until around the time that modern humans arrived, but apparently did not admix with them. *H. erectus* is the likely ancestor of two quite archaic groups that persisted in Southeast Asia until relatively recently: *H. floresiensis* and *luzonensis*.

• ***H. antecessor***: Spain, ∼800 KYA. Known from fossils in the Gran Dolina cave site in northern Spain, these are among the oldest known hominin fossils in western Europe [460]. Their precise phylogenetic position is unclear, but they were probably in Europe too early to be directly ancestral to the second-wave archaics including *heidelbergensis*, Neanderthals and Denisovans.

• ***H. naledi***: South Africa, ∼335K–235 KYA. Remarkable, relatively-recent hominins with small brains and a mixture of older Australopithecus-like and more modern Homo-like characters; they overlapped in time with the evolution of early modern humans. Currently known from only a sin-

Figure 3.51: **The first hominins outside Africa: *Homo erectus*** *skull dated to 1.8 MYA, from Georgia (at the border of Eastern Europe/ Western Asia). Dmanisi skull 3. Credit: Rama [Link] Public Domain*

gle cave system near Johannesburg. The main chamber containing fossils from at least 15 *naledi* individuals is so inaccessible that it was excavated by an all-female team who were slender enough to reach it [461].

• ***H. floresiensis***: Flores, Indonesia, ∼190–50 KYA years ago; tentatively linked to remains from 700 KYA; discovered on the Indonesian island of Flores in 2003. Tiny people at just over 1 meter in height, they were dubbed 'hobbits', a name that has stuck in popular usage [462]. Their small size likely reflects a frequent adaptation known as **insular dwarfism** in which island species adapt to low food availability by reduced body size. Although they persisted until relatively recently, it is believed that they are late-surviving descendants of *Homo erectus*, rather than relations of a later archaic such as Denisovans.

• ***H. luzonensis***: Luzon, Philippines, ∼130 KYA. Known from limited material from a single cave on the island of Luzon; unclear how long they inhabited Luzon. Like the Flores hobbits, they were small in stature, and had a mix of archaic features suggesting they may have descended from *Homo erectus*, with some features even reminiscent of *Australopithecus*. They may have been sophisticated enough to reach Luzon by raft or boat across open water.

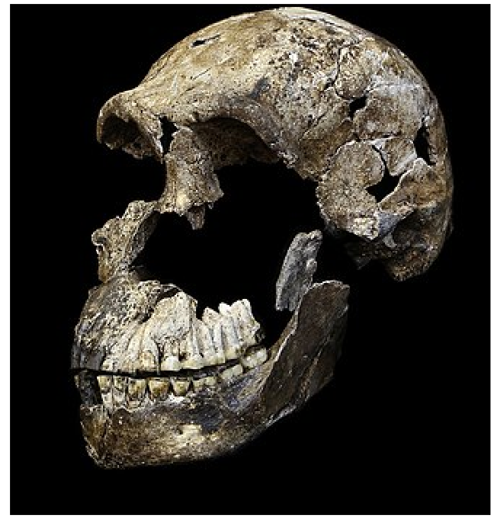

Figure 3.52: ***Homo naledi* skull**, *dated 235–335 KYA, from South Africa. H. naledi, who featured an ambiguous mixture of Australopithecine and Homo features, was contemporaneous with early Homo sapiens.* Credit: from Figure 5 of John Hawks et al (2017) [Link]. CC BY 4.0

Later archaic Homo (the second major wave out of Africa):

• ***H. heidelbergensis***: Africa, Europe, possibly East Asia. ∼600K–300 KYA. *Heidelbergensis* is best known from remains in western Europe. However, the species may have been very widespread as fossils with similar features in the same timeframe have been found in China, as well as south and east Africa. *Heidelbergensis* contains a mixture of early- and late-archaic features, and it's unclear if it's a link between African archaics and either modern humans or Neanderthals, or an evolutionary side branch. The name ***H. rhodesiensis*** is used for African specimens but this is likely the same species as *heidelbergensis*.

• ***H. neanderthalensis***: Europe, Middle East, Central Asia. ∼430–40 KYA. Neanderthals were stockier than modern humans, with large brains. They had sophisticated technology including characteristic stone tools, sea-faring, and potentially early forms of symbolic and ornamental art. Genetic data show that Neanderthals interbred with *Homo sapiens* around 50,000 years ago, likely in the Middle East, and all non-African populations carry around 2% Neanderthal ancestry. 'Neanderthal' is usually pronounced with a silent 'h', i.e., *Neandertal* [463].

• ***Denisovans***. Central Asia, Tibet, Southeast Asia, probably East Asia. ∼400–40 KYA; closest relative of Neanderthals. The discovery of Denisovans was a triumph of ancient DNA: routine genetic screening of bone fragments from Denisova Cave in the Altai Mountains in central Asia revealed a new archaic hominin without a skull or other clear fossil evidence to go with it. Since then, a likely Denisovan jawbone has been identified in the Tibetan plateau of China and several archaic skulls from China are tentatively associated with Denisovans [464]. Genetic data indicate multiple interbreeding events of Denisovans and humans, likely in East Asia and in the islands of Southeast Asia.

Figure 3.53: **Archaic flint handaxe, England.** *This lovely ancient axe is ∼400 KY old and was likely made by heidelbergensis or early neanderthal. It illustrates the craftsmanship of archaic Homo. Nicknamed the "Big Boy" it's over 25 cm long and weighs nearly 1.5 kg.*
Credit: British Museum [Link]. CC BY-NC-SA 4.0

Anatomically modern humans:

• ***H. sapiens***. Modern humans evolved in Africa, where fossils with modern features appear at around 200–300 KYA at sites in Morocco, Ethiopia, and South Africa. Fossils also show evidence for anatomically modern humans into the Arabian Peninsula and Middle East by ∼120 KYA. Further afield, there is evidence for human habitation in Europe, east Asia and Oceania by 50–70 KYA, and perhaps as early as 100 KYA [465] [466]. However, as we shall see, the genetic data suggest that these early dispersers did not contribute significantly to modern human populations.

Now that you've met the major characters, it's helpful to see them plotted on a map: here are the major fossil locations from the last 500 KY. I'm always struck by the remarkable global diversity of hominins that persisted until very recently in our evolution.

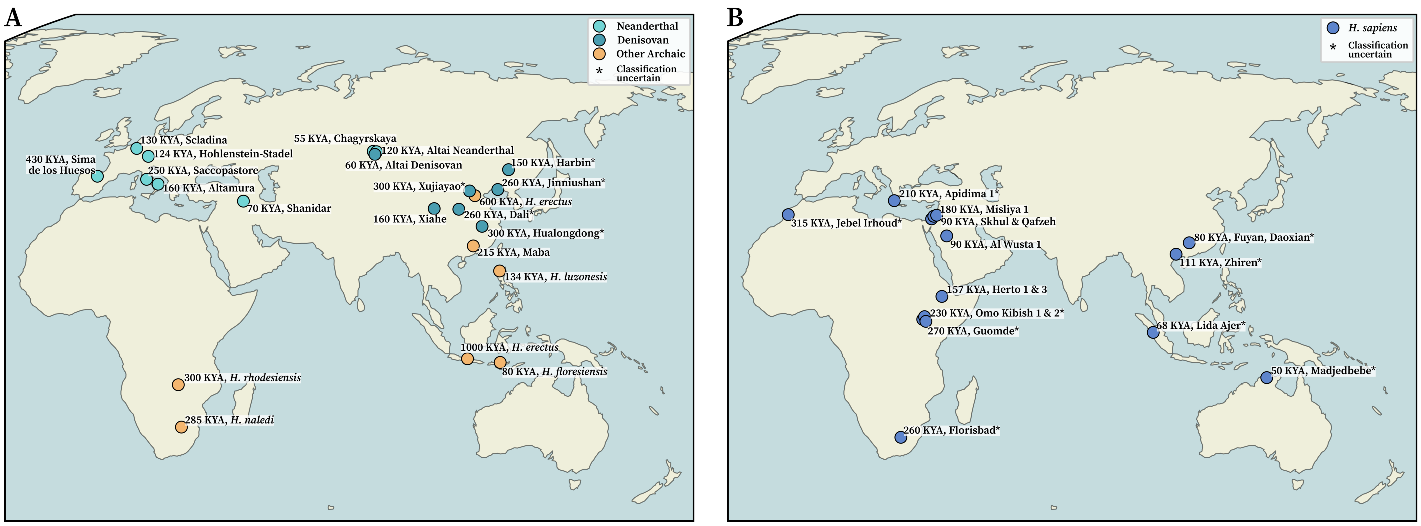

Figure 3.54: Overview of key late-stage hominin fossils: ***(A)*** *Archaic Homo from 0–500 KYA;* ***(B)*** *Homo sapiens prior to 50 KYA. The east Asian fossils marked as Denisovans are tentative assignments at this point. Neanderthals and Denisovans admixed with modern human populations in the past 50 KY, but the more archaic lineages – H. erectus, luzonensis, floresiensis and naledi – likely did not. Early H. sapiens populations outside Africa were largely or entirely replaced by a sapiens expansion after 50 KYA.* ***Indicated ages are very approximate*** *as they may not be known precisely and some sites, such as Denisova Cave, were occupied for long periods.* *Credit: Unpublished figure kindly contributed by Clemens Weiß; CC BY 4. Based in part on figures by Anders Bergström et al (2021) [Link].*

In the remainder of this chapter we'll focus on what we learn about this story from genetic data [d].

[d] *We'll cover ancient DNA in the next chapter. We now have high quality genomes from Neanderthals and Denisovans, but it has not yet been possible to retrieve data from other archaic lineages.*

Genetic perspectives on human origins. The first huge contribution to this story from genetics came in the early 1990s.

As you now know, fossil evidence shows that early hominin evolution was in Africa, but by 1.8 million years ago hominins had spread broadly across Eurasia. **When, and where did modern *Homo sapiens* evolve?**

By the 1980s and 1990s, this question had crystallized into two main alternative hypotheses.

The first theory, argued most prominently by the paleoanthropologist

Milford Wolpoff and colleagues in the 1980s and 90s, was known as the **multiregional hypothesis**. Wolpoff claimed that there was morphological continuity between early archaic fossils such as *Homo erectus*, and modern forms from the same regions, including in China, in Europe, and in Australasia (where Australian fossils were claimed to have similarities to *Homo erectus* from Java). Based on this, Wolpoff wrote that "Multiregional evolution traces all modern populations back to when humans first left Africa at least a million years ago... Modern humanity originated within these widespread populations" [467]. The model allowed for gene flow between populations so that key genetic adaptations shared by all modern humans – for example, enabling human language – could spread between populations.

In contrast, the **Recent African Origin** model proposed that anatomically modern humans evolved entirely within Africa. In this model, modern humans spread out of Africa within the last 100 KY, replacing archaic forms as they spread throughout Eurasia and beyond.

Resolution of this debate began with pioneering studies of mitochondrial DNA (mtDNA) variation from Alan Wilson's lab at Berkeley in 1987 and 1991 [468]. Wilson's lab focused on mtDNA for two reasons: (1) mtDNA is maternally inherited and does not recombine, meaning that the data can be described by a single coalescent tree; (2) mtDNA has a very high SNP density due to the high mitochondrial mutation rate – this was helpful in the 1990s, when sequencing was expensive and laborious [469].

You can see below the crucial figure from the 1991 paper, showing the inferred tree of mtDNA sequence variation in Africans and non-Africans. The analysis suggests two key points: (1) The deepest lineages in the tree, on either side of the MRCA, are African, while most non-African diversity is found in a restricted set of clades; (2) the MRCA time is around 200 KYA.

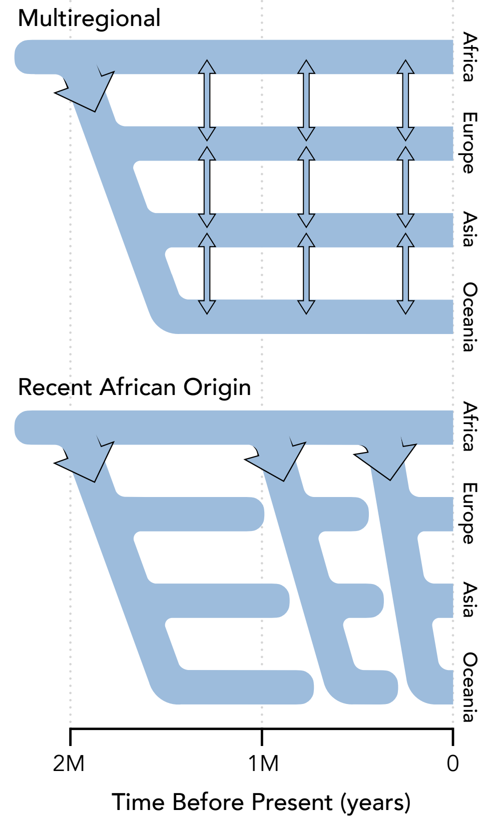

Figure 3.55: **Human origin models**. *The (incorrect)* ***Multiregional Model*** *was characterized by local genetic continuity: e.g., modern Asians would descend from Asian Homo erectus. In the* ***Recent African Origin Model*** *there are 2-3 migrations out of Africa and all modern humans share recent ancestry in Africa.* Credit: *Unpublished figure kindly contributed by Alyssa Lyn Fortier. CC BY 4.*

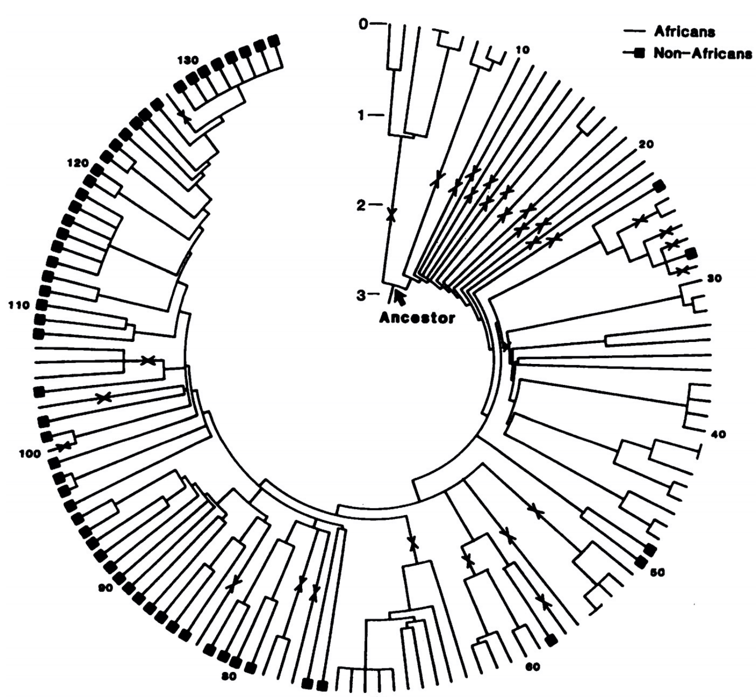

Figure 3.56: **A recent African origin of mtDNA variation (1991).** *Tree of mitochondrial DNA variation for* 135 *distinct mtDNA types sampled from* 189 *individuals. Types from Africans are marked with straight tips, and non-Africans with terminal boxes.* ***The deepest splits in the tree are among Africans, suggesting an African common ancestor.*** Credit: *Figure 3 from Linda Vigilant et al. (1991) [Link]. Used with permission.*

The authors correctly interpreted this as implying a recent African origin of all humans. In particular it is difficult to reconcile the multiregional model, with strong population structure going back more than 1 million years, with such a recent common ancestor; moreover the multiregional model would predict the deepest lineage splits to be between different regions, instead of all being within Africa.

Riffing off the fact that mtDNA is maternally inherited, the science writer Roger Lewin dubbed the common ancestor as the **Mitochondrial Eve**, after the Biblical Eve who is the mother of all humanity in the Old Testament creation story. This caused no small amount of confusion among lay-people: it seemed to imply, incorrectly, that there was only a single female alive at that time [e], and it incorrectly linked the analysis to the Biblical creation story. But, like an unfortunate childhood nickname, the name has stuck.

[e] *As we discussed in Chapter 2.2, every region in the genome has a most recent common ancestor; this corresponds to a random person and timepoint at which all lineages coalesced. It does not imply that the population was particularly small at the time of the MRCA (and certainly not a single person!).*

These results have held up with modern data: the mitochondrial MRCA time was recently estimated at 160 KYA, not so far off the 1991 estimate based on very limited data [470]. Similar results have been found for the other large, non-recombining region of the genome – the **Y chromosome** – where the deepest lineages are also African, and the MRCA time has been estimated at $\sim$140 KYA [471]. *These very recent MRCA dates with deep lineages in Africa are inconsistent with expectations under the multiregional model.*

Genome-wide data also support the Recent African Origin model. The mtDNA and Y chromosome data were major landmarks in the history of the field, but early on there was no way of knowing if these loci might be unusual in some way – for example if they happened to have been targets of selective sweeps. Indeed, we now know that the MRCA times in these regions are very recent compared to the autosomal genome [472] [473].

Genetic diversity declines with distance from Africa. Following the mtDNA work, another key early observation was that genetic diversity is highest in Africa, and declines smoothly with distance from Africa [f].

[f] *The gradient of genetic diversity with distance from Africa can also be seen in Table 1.2.*

A striking illustration was shown in a 2005 paper by Sohini Ramachandran and colleagues, who computed the average heterozygosity at short tandem repeat loci in indigenous populations from the Human Genome Diversity Panel (HGDP) (panel A [474]):

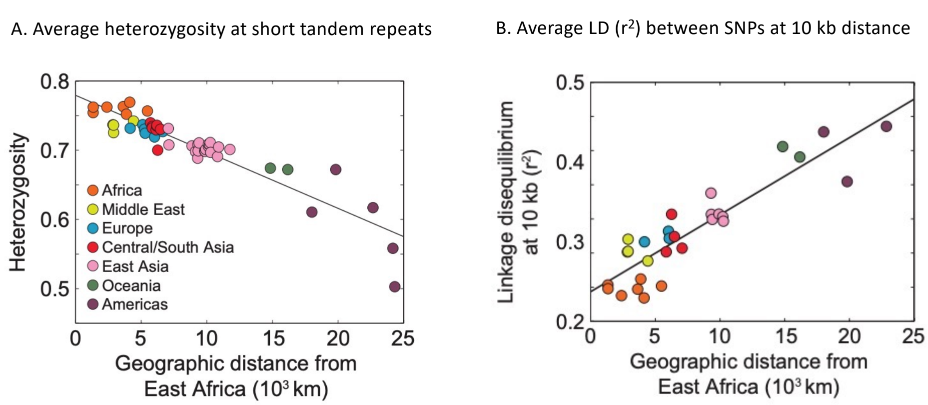

Figure 3.57: **Genetic diversity decreases and LD increases with distance from Africa.** **A.** *Heterozygosity at STR loci for populations in the HGDP. Distance from east Africa is measured along likely dispersal routes: notably, the distance for native American populations is measured via the Bering Strait.* **B.** *Average LD in the same HGDP samples, measured using r^2 between pairs of SNPs at 10 kb spacing.* *Credit: Modified from Figure 1, Michael DeGiorgio et al (2009) [Link].*

In the plot above, 85% of the variation in genetic diversity among 53 human populations is explained simply by dispersal distance from east Africa.

Ramachandran *et al.* explained this pattern using a **serial founder** model. The central idea of the model is to imagine human populations spreading out across the globe in a series of discrete steps, starting from sub-Saharan Africa. At each step moving away from Africa, a new population is generated by subsampling variation from the previous step. Since the dispersal out of Africa was relatively recent on a coalescent timescale, there has not been long enough to regenerate diversity within populations since the major dispersal event [g].

[g] *In contrast, under a multiregional model there is no clear reason why N_e should track closely with distance from Africa–e.g., if there had been genetic continuity in East Asia since Homo erectus, we might have expected high diversity there.*

Although we now know that human population history was much more complicated than this, including migration, admixture and population replacements, this simple model captures the essential pattern of reduced genetic diversity as modern humans spread out across the landscape [475].

Similarly, the scale of **linkage disequilibrium (LD)** tends to increase with distance from Africa [h], as shown in **Panel B, above**. This should come as no surprise given the results on heterozygosity. As we discussed in Chapter 2.3, the scale of LD is proportional to effective population size: recombination is less effective at breaking down LD in small populations than in large.

[h] *Recall that LD refers to the correlations between genotypes at different variants close together in the genome.*

An example of this pattern is shown below for a single genomic region (panel A, below). Notice that blocks of haplotype sharing are generally longer in populations located further from Africa (shown toward the right on the plot):

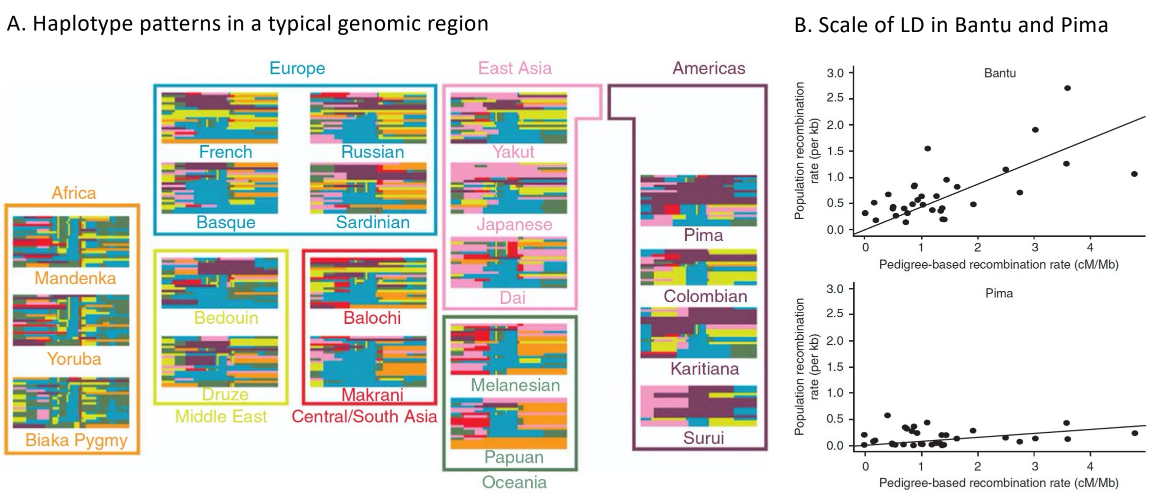

Figure 3.58: LD increases with distance from Africa. A. *Visualization of haplotypes in different populations, for a typical region of* 330 *Kb. Haplotypes are plotted in rows and the horizontal axes reflect genomic position. Colors indicate haplotype sharing and the haplotypes are sorted so that similar haplotypes are adjacent.* **B.** *The decay-rate of LD, measured by* $4Nr$, *is much higher in the Bantu (African) than in Pima (native American), shown here for* 32 *genomic regions.* *Credit: Figs 1a, 4b from Don Conrad et al (2006) [Link]. JKP is a copyright holder.*

Later in the book when we get to genome-wide association studies, we'll see that these varying levels of LD have important consequences for our ability to detect and localize causal variants in different populations.

Non-African populations share the same Neanderthal introgression event [i]. Another critical piece of evidence about the timing and spread of human populations comes from Neanderthal ancestry in non-Africans. As modern humans spread out of Africa, they encountered and interbred with Neanderthals. As a result of this, *all non-African populations carry* 1.5%–2.0% *Neanderthal ancestry*. As you might expect, African populations – who are not descended from this key non-African source population – carry only trace amounts of Neanderthal ancestry [476].

[i] *Spoiler Alert! We'll talk about archaic introgression in detail in the next chapter.*

Genetic dating indicates that contact between Neanderthals and humans took place around 45–50 KYA, most likely in the Middle East – *and that all non-African genomes are descended from the same mixture event* [477] [j]. I mentioned above the fossil evidence for Homo sapiens in Eurasia before 50 KYA. These results imply that the expansion after 50 KYA must have largely, or entirely, replaced these earlier populations [478]. Together, these observations are entirely inconsistent with a long-term multiregional model.

[j] *The Neanderthal admixture at* $\sim$50 *KYA, inherited by all non-Africans, places an important upper bound on the time that modern non-Africans expanded from a single source population.*

In summary: Fossils show modern human features appearing in Africa around 200–300 KYA. Genetic data show deep structure within Africa over the past 200 KYA, and a single Recent African Origin that largely replaced the diversity of archaic forms that were previously spread across most of Eurasia. As modern humans spread, they encountered late-stage archaics – Neanderthals and Denisovans – who contributed 1.5–5% to the genomes of modern non-Africans [479].

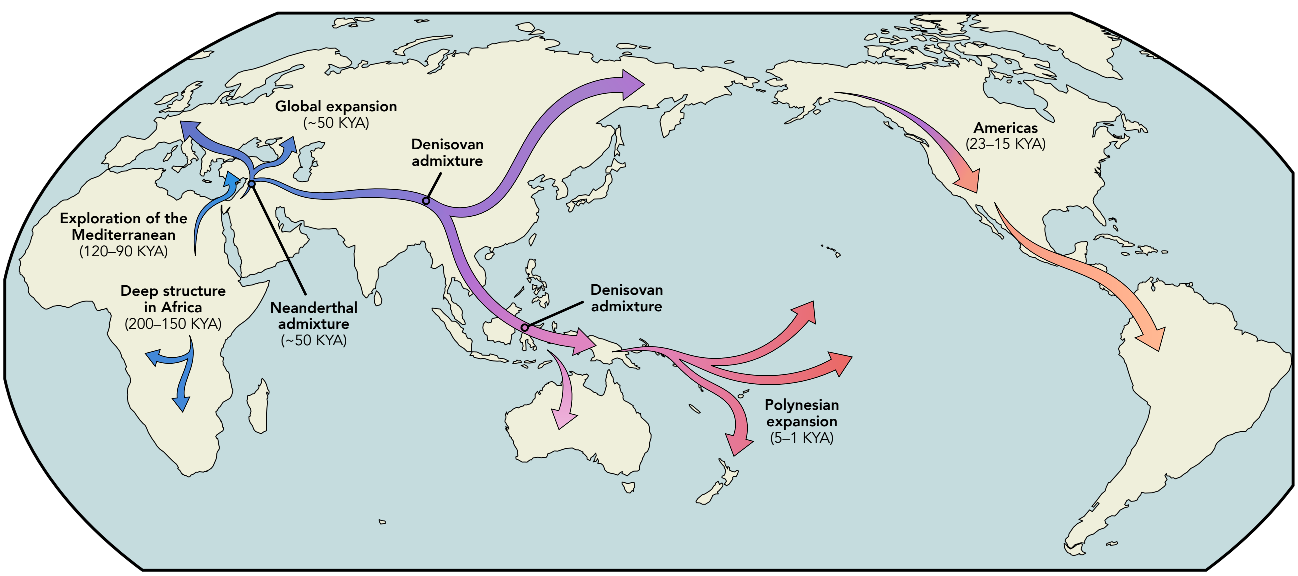

Figure 3.59: Global spread of modern humans. *The earliest fossils with modern features appear in Africa* 200–300 *KYA. Early dispersals of modern humans into Eurasia (*120–50 *KYA) were likely replaced by a single major dispersal after* 50 *KYA. The precise locations of arrows and admixture events are generally not known in detail.* ***All times are approximate.***

We haven't said much yet about what's going on **inside Africa**. This is another fascinating story. But first we need to say a bit more about tools for interpreting genetic data.

Genome-wide inference of population history. So far we have discussed inferences that can be drawn from relatively simple aspects of the genetic data: for example a single locus such as mtDNA, or a simple measure such as heterozygosity or extent of LD [k].

[k] *The term **inference** refers to statistical approaches to estimate parameters or choose among models.*

Moreover, so far the analysis that I have described has been mainly qualitative – e.g., a recent MRCA argues against deep multiregional population structure, and the decreased variability outside Africa reflects the late dispersal out of Africa.

This kind of qualitative reasoning – based on "vibes" – is often very helpful, but sometimes one's intuition can be wrong [480]. In any event it doesn't make full use of the immensely rich data that we now have available: tens of thousands of whole genomes from modern and ancient humans.

For example, the vibes-based approach does not really help us answer key questions such as: When did modern African and non-African populations separate? What are the deepest splits among extant human populations? Do all non-African populations descend from the same out-of-Africa migration? When did human populations undergo bottlenecks, or population expansions?

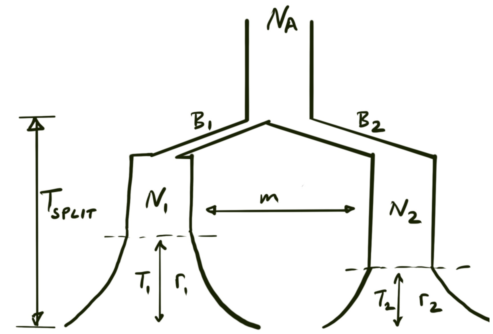

Figure 3.60: **Example of a model** *for two populations shows how complicated such models can be – including parameters for population sizes, population split time, bottleneck sizes, migration rate m, and an exponential growth phase for each population specifying start times and growth rates – while noting that even this model is surely simpler than reality.*

These kinds of questions are complicated by the fact that the underlying history is inherently complex, including population splits, migrations, and ancient population structure and population size changes. Hence, a simple measure such as F_{ST} (which quantifies allele frequency divergence between populations as a single number) could be produced by many different population histories.

Suppose that we want to infer population histories, using genome sequences from one or more populations. How should we analyze the data?

In population genetics, we have the luxury of well-defined models that describe how the data – i.e., genome sequences – are produced by the standard processes of population genetics: mutation, drift, recombination; as well as by the specific details of population histories: ancient population sizes, population split times, and migration rates [481]. This means that it is relatively straightforward to *simulate data* under complex models of population history [482].

However, it is much harder to perform **inference** in population genetic models. Statistical theory tells us that when we have well-defined models as we do here, all of the available information about the parameters is present in what is known as the **likelihood**: i.e., the probability of generating the observed data for a given set of parameter values.

To compute the full likelihood we would need to be able to infer the unknown ancestral recombination graph for the samples (ARG) [483]. Remember that the ARG records the coalescent genealogy at every position along the genome, as well as how the different genealogies are glued together by linkage and separated by recombination. Crucially, the ARG tells us about the rates of coalescence within and between populations at different times in the past, and these rates reflect ancient population sizes, population split times, and migration rates. The cartoon below shows how the coalescent genealogy at a single position in the genome might vary depending on population history:

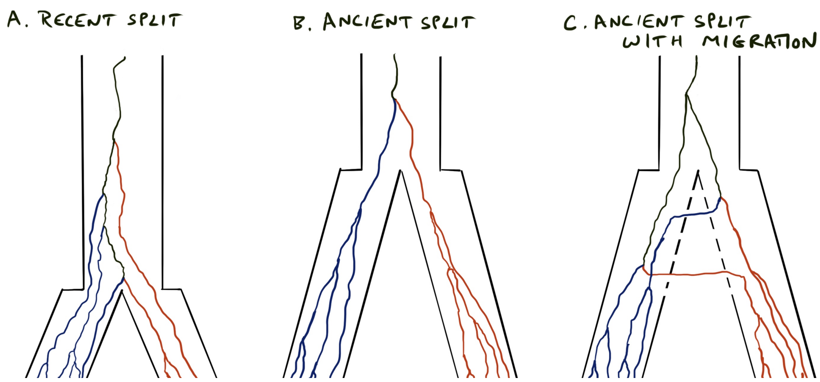

Figure 3.61: **Ancestral genealogies reflect population history.** *For example, a recent split allows recent coalescence between populations (compare A vs B); migration allows recent coalescence but at lower rates (C). The information provided by a single locus is quite limited, but genome-wide data can be very informative about history. We can also infer how effective population sizes changed over time using rates of coalescence as a function of time.* *For visualization lineages are colored blue if they are ancestral only to tips in the left population; red for right; black for both.*

Unfortunately, although it's relatively straightforward to simulate coalescent models, it turns out to be extremely difficult to perform likelihood comparisons under the ARG, especially in large-scale data.

For this reason, there has been a lot of work to develop methods that can extract the most relevant information from the data, while still being computationally practical. As we discuss in the Box below, these methods include various approximations of the likelihood that do not require getting the full ARG [1], simulation-based approaches, and recent advances toward ARG estimation.

[1] *We have already discussed one important example of this type of approach: the* ***Li and Stephens*** *algorithm for haplotype inference, which simplifies the coalescent model to make the likelihood computation practical (Chapter 2.3).*

The box is a bit specialized, and you can skip over it if you like. After the box we'll cover one important approach, PSMC, in some detail.

Optional Box: Approaches to historical inference. *This box is more specialized and not required.*

Moment-matching methods. Some of the earliest approaches to estimating population parameters simply used the basic formulas of population genetics to estimate parameters like N_e and split times [484].

In statistics, the **moments** of a distribution are quantities relating to the mean, variance, and other higher-order statistics [485]. There are often simple formulas to predict these, and then we can estimate parameters by plugging in values observed in the data. As a simplified example, recall that the expected heterozygosity H is $4N_e\mu$, and that expected F_{ST} is approximately $T/2N_e$. So if we had sequence data from two populations we could estimate H and F_{ST}. Then, since we have an estimate of the mutation rate μ, we could solve for N_e within each population, and next solve for the split time T. A more sophisticated version of this approach used counts of the numbers of shared SNPs, private SNPs, and fixed differences within and between populations [486].

But this kind of approach is not very satisfying. First, it's a bit annoying to start with huge amounts of data, collected at great expense, and then just compute two or three simple summary measures for which we have analytical formulas. Second, this approach immediately breaks down if we want to consider anything at all complicated in the history – for example population splits *and* migration; population growth or bottlenecks; or ancestral population structure – as it becomes increasingly difficult to estimate multiple parameters, and there are not always simple formulas for more complicated scenarios.

Simulation-Based Inference. Starting in the late 1990s, there was a move toward using simulation approaches to allow a broader range of summary statistics and much more complicated evolutionary models. (**Summary statistics** are low-dimensional quantities that we can compute from the data, sometimes as simple as H or F_{ST}, or sometimes requiring more complicated algorithms such as estimates of ρ.) The basic process is to:

1. *Choose a series of summary statistics* that are expected to be informative about population history;

2. *Simulate a large number of data sets*, across a range of models of population histories. The simulations are set up to match the experimental design of the real data for basic features such as sample size and amount of genotype/sequence data;

3. *Record population histories for which the simulated summary statistics are close to the observed values in the data*; these are considered to be part of the acceptable parameter space. Meanwhile, parameter values that consistently produce poor matches to the observed data are rejected.

This is formalized in an approach known as **Approximate Bayesian Computation (ABC)** [487] [488]. The most basic version of this uses a simulation technique known as **rejection sampling**. The user specifies plausible ranges (technically, prior distributions) for the parameters of interest. Next, each replicate simulation starts by sampling randomly from the possible parameter values. If the simulated data in a replicate are deemed to be "close enough" to the true data, then the parameter values are recorded as part of the distribution of parameter values that are consistent with the data (technically, the posterior distribution).

These kinds of approaches are extremely flexible in terms of the range of population histories that can be modeled – limited only by the flexibility of the simulation software. They are also very good at measuring model uncertainty: that is, to quantify the range of parameter values and models consistent with the data. These approaches can also easily account for uncertainty in external parameters such as mutation rates or generation times.

The main downsides of ABC methods are that: (1) The choice of summary statistics is rather arbitrary, and depend on the judgment and creativity of the data analyst. For example Voight et al. (2005) used the number of segregating sites; the mean and variance of Tajima's D (a measure of the allele frequency spectrum), and a measure of LD [489]. There is no strong theoretical reason to choose these specific summaries, and there may well be other choices that would perform better. (2) We cannot use more than a handful of summary statistics simultaneously – if we do, then the rate at which simulated data sets are tolerably close to the real data becomes vanishingly small. The limitations of ABC methods are an example of the 'curse of dimensionality': i.e., that ABC becomes rapidly less efficient as we increase the number of summary statistics. This in turn means that we cannot perform inference for models with large numbers of parameters, because they become under-constrained when there are more parameters than independent summary statistics.

A new alternative that may help for these issues combines simulation with methods from **machine learning** [490]. The key idea is to simulate data under a variety of models and then develop **deep learning** classifiers by training on labeled simulated data. The classifier can then be applied to the real data. This can be viewed as a form of Simulation-Based Inference, where the deep learning classifier takes the place of user-defined summary statistics [491]. The advantages are that the classifier bypasses the need for investigators to choose their own summary statistics, and it can potentially use higher dimensional summaries of the data to achieve higher accuracy. However, there are domain-specific challenges in how to apply machine learning to population genetic data, and it's still unclear for which problems these methods will ultimately prove most useful.

Composite likelihood Methods. The next important category of methods steps closer to computing likelihoods. *We saw an important example of this in Chapter 2.2 (Figure 2.32).*

Composite likelihood methods simplify the likelihood calculation by breaking the data down into small pieces for which it's easier to compute the likelihood. We compute the **composite likelihood** by multiplying each of these individual likelihoods together as if they were independent. For example, suppose that we sequence $m/2$ genomes from a population. Instead of trying to compute a likelihood for the full data, we could simplify the data to just look at the **site frequency spectrum (SFS)**. At each position in the genome we simply count the number of derived alleles: 0, 1, 2, ... m. Then, let n_i be the number of sites in the genome where there are i derived alleles. The SFS simply lists the numbers of sites with each possible number of derived alleles: n_0, n_1, n_2, ... , n_m [492].

We could also tabulate a **joint SFS** for two or more populations: $n_{i,j}$ is the number of sites where there are i derived alleles in Population 1, and j in Population 2.

Notice that in both cases, we're taking the highly complex whole-genome data, and compressing it down to a much lower-dimensional summary. When we do this, we effectively treat every site as independent of every other site – that is, we ignore all the information that is present in haplotypes.

Next, let's suppose that we can compute the probability π that a particular site will be in state i, j given some population history model h; denote this $\pi_{i,j}(h)$. In practice, $\pi_{i,j}$ can be computed by simulation or, for some models, with diffusion theory. In effect, we want to find model parameters that maximize the fit of the predicted joint frequency spectrum $\pi_{i,j}$ to the observed data $n_{i,j}$.

Formally we do this by computing the composite likelihood for the data, assuming some history h:

$$\Pr(Data|h) \approx c \prod_{i=0}^{m} \prod_{j=0}^{m} \{\pi_{i,j}(h)\}^{n_{i,j}}. \tag{3.22}$$

In words, the composite likelihood is obtained by multiplying, over all combinations of i and j, the expected frequency of sites with variant counts i and j (written $\pi_{i,j}(h)$), raised to the power of the actual number of sites with counts i and j (written $n_{i,j}$). You can derive this composite likelihood using the standard formula for multinomial sampling – after making the simplifying assumption that every site in the genome is independent. (The constant c contains a bunch of factorials and cancels out when we compare different models for the same data.) To perform inference, we search for the parameter values in h that maximize this likelihood.

Composite likelihoods are extremely useful for a variety of problems in population genetics because they can often capture a huge amount of information in the data while still being computationally practical. Applications include estimating population histories, detecting selective sweeps, and estimating local recombination rates using pairs of variants [493]. One key weakness is that these methods usually ignore information contained in haplotypes between pairs of sites.

ARG approximation. We started this section by discussing that the true underlying structure of population genetics data is given by an ancestral recombination graph; however full ARG prediction has long been recognized as an extraordinarily challenging problem [494].

But in the last few years there has been exciting progress on this problem. One pioneering approach, ARGweaver, samples from the probability distribution of ARGs conditional on an explicit historical model with splits and migration. ARGweaver can estimate parameters of the model and identify interesting features such as blocks of archaic introgression. Given the complexity of the problem, ARGweaver is limited to small numbers of samples (tens of samples at most) [495].

Recently, a new generation of methods have implemented extremely efficient heuristic methods that can estimate the ARG, including tsinfer, Relate, and SINGER [496]. These methods are very fast and can be applied to genome-wide data for samples on the order of 10^4 or even 10^5 individuals.

These new techniques don't directly solve all the major inference problems as we still cannot compare likelihoods under different models, but they offer an extremely promising road toward tackling a range of problems in population genetics, including studies of past population size changes, relationships among populations, relationships between modern and ancient genomes, and for studying selection.

We close this section with more detail about a technique that is particularly useful, and builds nicely on concepts you learned in Part 2 of the book, about the coalescent.

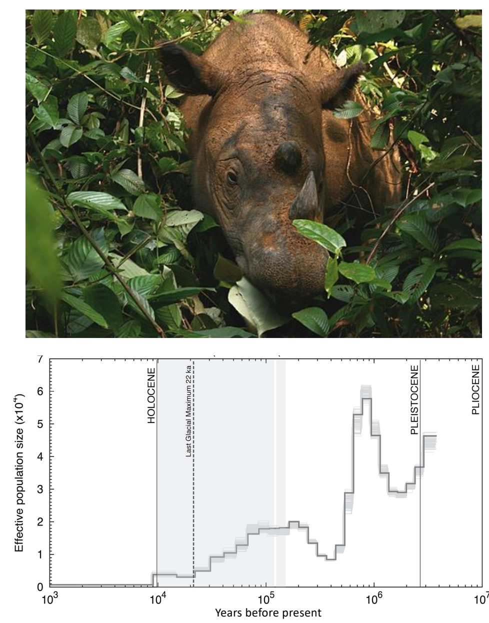

Figure 3.62: **PSMC of the critically endangered Sumatran Rhino.** *PSMC has become an important analysis tool for many species. This analysis shows that the Sumatran Rhino has had a tiny population size since the end of the last ice age* 10 *KYA.* ***Typical of PSMC plots, time is shown on a log scale with the most recent times to the left.*** *Credit: William Strien [Link] CC-BY-2.0; Modified Fig 1 from Herman Mays Jr et al (2017) [Link] CC-BY-NC-ND.*

Population history from one individual!? All of the methods described so far can use samples from many individuals, but then collapse the data from many sites into a much smaller number of summary values.

In a remarkable 2011 paper, Heng Li and Richard Durbin flipped this logic on its side by focusing on the information that we could get out of a single diploid genome, in a method known as **PSMC (Pairwise Sequentially Markovian Coalescent)** [497]. Their key insight was that a single genome contains a great deal of information about the distribution of coalescence times, and that this, in turn, can tell us a lot about population history.

Li and Durbin drew on the idea that we can infer a great deal about population history if we know the distribution of pairwise coalescence times. When the population size is small we tend to get high rates of coalescence, and when the population size is large, coalescence is rare. Specifically, *we define the coalescence rate at time t as the instantaneous rate of coalescence at time t, provided that they have not coalesced prior to t* (as usual, measuring time backwards from the present).

Recall from Chapter 2.2 that the coalescence rate at time t is $1/2N_t$, where N_t is the effective population size at time t. Thus, if we can estimate the coalescence rates for all times t, then by computing the inverse of the coalescence rate we get a trajectory of population sizes over time [498].

The next insight is that this type of information is recorded in a single diploid genome. At each point in the genome there is a random coalescence time between the maternal and paternal copies – these coalescence times are drawn from the overall distribution of coalescence times. As we look along a chromosome, there is a sequence of (unknown) coalescence times: we can imagine a contiguous block of chromosome that shares the same common ancestor, and then another block with a different common ancestor at a different time. The contiguous blocks are separated by historical recombination events. Blocks with a recent common ancestor will tend to be long (little time for recombination) and have low variation (less time for mutation), compared to blocks with more ancient common ancestors [499].

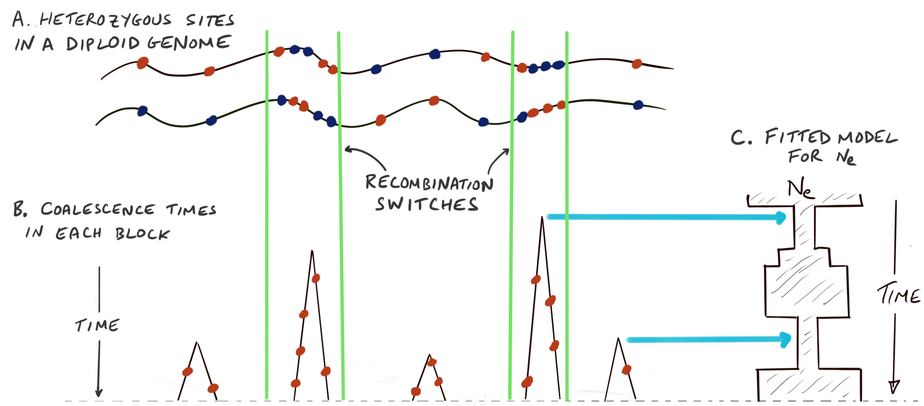

Figure 3.63: **Schematic overview of PSMC.** **A.** *Two haplotypes of a diploid genome. Heterozygous sites are indicated: ancestral in blue, derived in red.* **B.** *Some segments of the genome have deep coalescence time: these tend to be short, with high density of heterozygous sites.Segments with recent coalescence tend to be long with low density of heterozygous sites.* **C.** *PSMC fits the data using a model of discrete time epochs with an N_e parameter estimated for each epoch. Epochs with more coalescence events are inferred to have low N_e.*

To simplify computation, the PSMC algorithm divides time into a series of discrete time epochs. Then, based on the intuition above, PSMC estimates the coalescence times along the genome using a hidden Markov model approach. The input parameters to the hidden Markov model include the coalescence rate in each time epoch (t_i, t_{i+1}). These parameters are estimated from the data by maximizing the overall likelihood of the PSMC model: intuitively, you can imagine that if many regions coalesce in epoch (t_i, t_{i+1}), then PSMC infers the coalescence rate to be high (and N_e to be low) in that epoch.

This approach actually works astonishingly well. Since there are $\sim 10^6$ block switches in a single diploid genome [500], PSMC can easily handle models with tens of epochs. And because PSMC uses haplotype information, it picks up on a fundamentally different aspect of the data than SFS-based approaches. In extensions that we will discuss shortly, PSMC is very useful for complicated population-split models that are difficult to approach with the SFS. PSMC also has weaknesses, including that specific historical events are often smeared across adjacent time bins, and it can perform poorly at recent timescales, such as detecting the recent growth that has occurred in many populations.

The next plot shows a typical application of PSMC to human genomes from a variety of populations:

These plots illustrate several important features:
• The African populations have varied patterns of population sizes until quite far back (around 250 KYA). As we'll see shortly, this reflects deep African structure.
• There is a strong bottleneck in all non-African populations at around 50 KYA prior to their global spread. The indigenous Mexican group has an

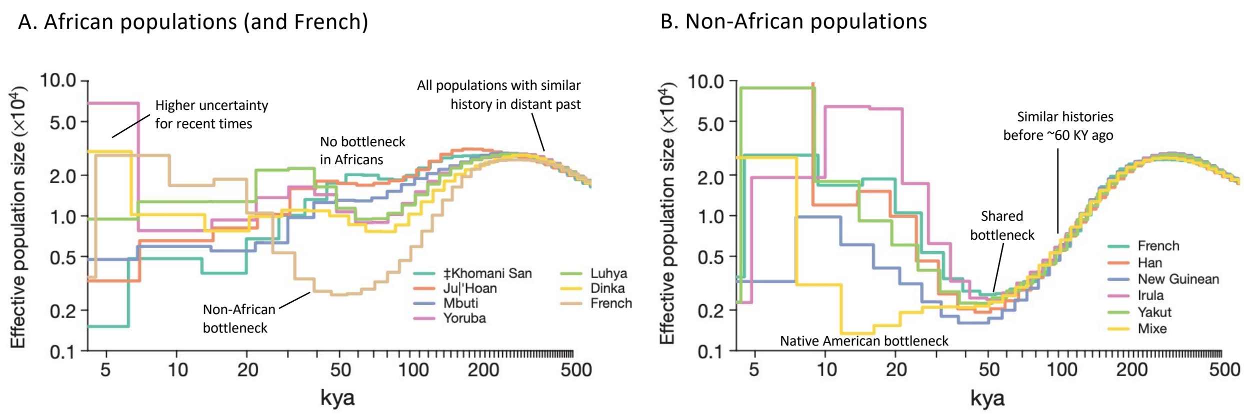

Figure 3.64: PSMC in human populations. A. *Effective population size over time, estimated for a variety of African populations, and French for comparison. Notice the out-of-Africa bottleneck that is apparent in French but absent in the Africans, as well as shared history before* $\sim$ 250*K years ago.* **B.** *Population sizes for non-Africans. Notice the shared bottleneck around* 60*KY ago. The Mixe (yellow), an indigenous Mexican population, show an additional bottleneck at around* 15*KY ago, likely corresponding to low population size during colonization of the Americas through Beringia. The variability at very recent timescales (left of each plot) reflects both population-specific growth or contractions, as well as higher estimation error at recent times.* *Credit: Modified from Figs 2d, 2f from Swapan Mallick et al (2016) [Link]. Used with permission.*

additional bottleneck around 15 KYA, probably reflecting the colonization of the Americans through Beringia (the region around the Bering Strait).
• The non-African populations have similar histories before about 60 KYA, reflecting that they are largely descended from a single source population at that time.

However, you might have noticed that the signal of recent population growth is much weaker in the PSMC analysis than what we saw in the SFS-based analysis of large samples. This highlights a weakness of PSMC, namely that it tends to produce very noisy estimates in the most recent time epoch(s). This is because when we are looking at the coalescence of a pair of genomes there are very few coalescent events in the most recent time frame, and more in earlier time periods.

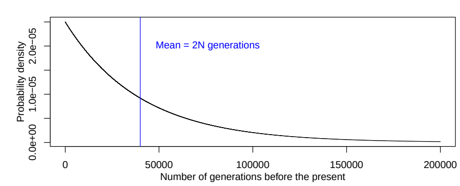

Figure 3.65: **Theoretical probability density of coalescence times** *for* $n = 2$*, and constant* $N_e = 20,000$*.*

But you, Dear Reader, exclaim – "*Wait! Isn't the coalescence probability highest at recent times, and steadily declining over time? Shouldn't PSMC actually perform better at recent times?*" It's true that the coalescence rate is highest at recent times (Figure 3.65), but the key point is that recent coalescence events span much larger chromosomal regions than do older events. So while a *larger fraction of the genome* is involved in very recent coalescence events per unit time, *there are fewer independent events*, and this increases noise in the estimation. Furthermore, if there has been recent population growth, as in many human populations, this makes the problem worse by decreasing the coalescence rate in recent times. Since the original PSMC paper, methods have been developed to improve performance at recent times by increasing the sample to more than just two haplotypes [501].

So far we have been talking about PSMC as a way of measuring the coalescent intensity for a pair of haplotypes from the same population. But we could also apply the same concept for *a pair of haplotypes from different* *populations*. In 2014 Stephan Schiffels and Richard Durbin introduced a concept called the **relative cross coalescence rate, which measures the coalescence rate for two lineages from different populations relative to the average of the within-population rates** (and defined to be ≤ 1) [502].

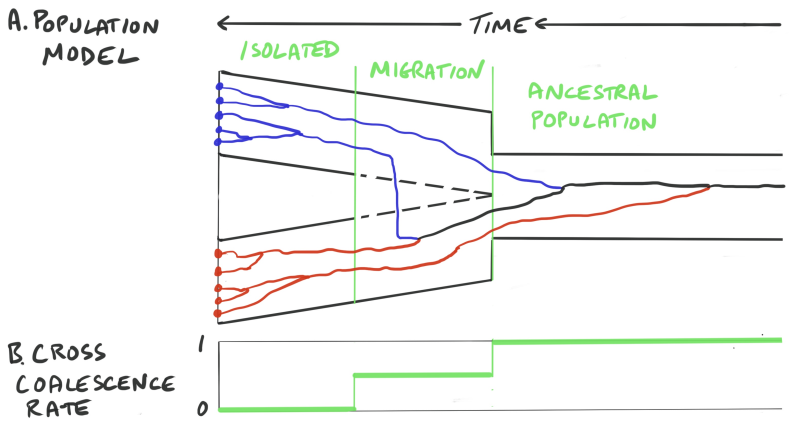

Figure 3.66: **Cross-coalescence during a gradual split**. **A.** *Hypothetical population model: going backward in time, two populations are initially isolated; then they start to exchange migrants (allowing ancestral lineages to move between populations); then they merge into a single ancestral population.*
B. *Under this model the cross-coalescence rate is initially 0 (when the populations are isolated); higher during the migration phase; and 1 when the populations merge.*
Note: *Here the present-day is placed on the left to match the standard orientation of PSMC plots.*

The cross coalescence rate allows us to visualize the process of population mergers in the past: **two populations that do not exchange migrants should have a cross coalescence rate of zero, while two populations that have fully merged into a single ancestral population have a cross coalescence rate of one.**

You can see this in the next plot, which shows cross-population coalescence rates among seven pairs of non-African populations (heavy lines) along with simulated rates assuming instantaneous population splits (light lines):

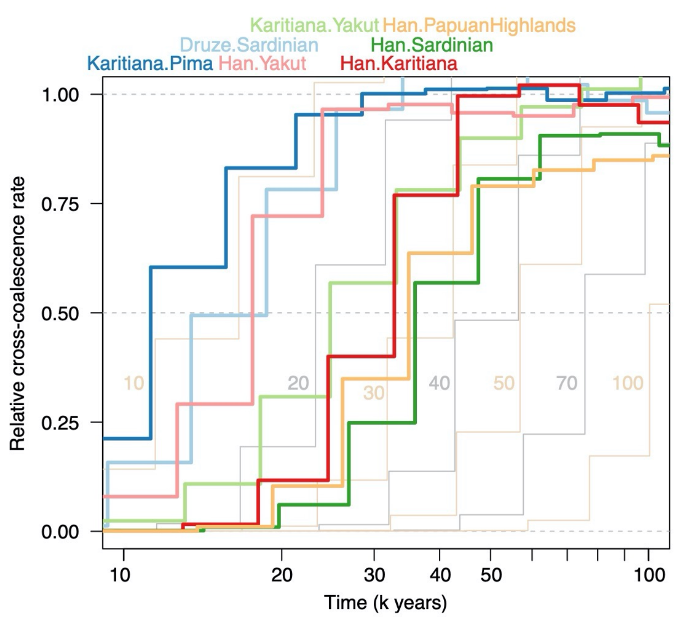

Figure 3.67: **Cross-coalescence of non-African populations**. *The solid lines show cross-coalescence rate estimates as a function of time before present, for pairs of populations. The color key is shown at top. The faint background lines show the estimated cross-coalescence rates obtained from data simulated with instantaneous splits at the times indicated in KYA.*
Populations: ***Karitiana***, ***Pima***: *South/North America (natives);* ***Druze***: *Middle East;* ***Sardinian***: *Europe;* ***Han***: *China;* ***Yakut***: *Siberia;* ***Papuans***: *Papua New Guinea.*
Credit: Fig 5b Anders Bergström et al (2019) [Link] CC-BY-NC-ND 4.0.

As you can see above, in the simulated data, the cross-coalescence rates climb from 0 (distinct populations) to near 1 (single ancestral population), centered around the correct timing of the simulated splits, although the actual signal is somewhat smeared across adjacent time bins both before and after the correct time.

Next, looking at the real data, we see that – as expected – the closely-related population pairs join first, starting with the two Native American populations, Karitiana and Pima. All seven population pairs reach 80%–100% cross-coalescence by 50 KYA; this is broadly consistent with the analysis of Neanderthal introgression that I described above, showing a common source for non-Africans around 50 KYA [503].

In all cases, we see that cross-coalescence accumulates more slowly than in the simulations with instantaneous splits. This suggests that the relationships among populations are more complicated than a simple split, likely including both gradual splitting and subsequent gene flow.

What happens when we look at cross-coalescence in African populations? You'll have to wait a couple of pages for the answer. In this last section of the chapter, we'll explore the population structure and history of sub-Saharan Africa.

Case study: Deep population structure in sub-Saharan Africa. We've just seen that all non-African populations are fairly closely related. But the story of African population history, which we are just starting to understand now, is fascinatingly different!

African population genetics have been studied through a combination of modern population sampling and limited ancient DNA [504]. For in-depth reviews of African population genetics see [505].

Recent work identifies at least four major ancestry groups in sub-Saharan Africa, with additional structure within each major group [506] [m]. **Divergence times among all four groups are more ancient than between any non-African populations!** We describe these groups briefly, below.

[m] *For more about structure in North Africa see Figure 3.15.*

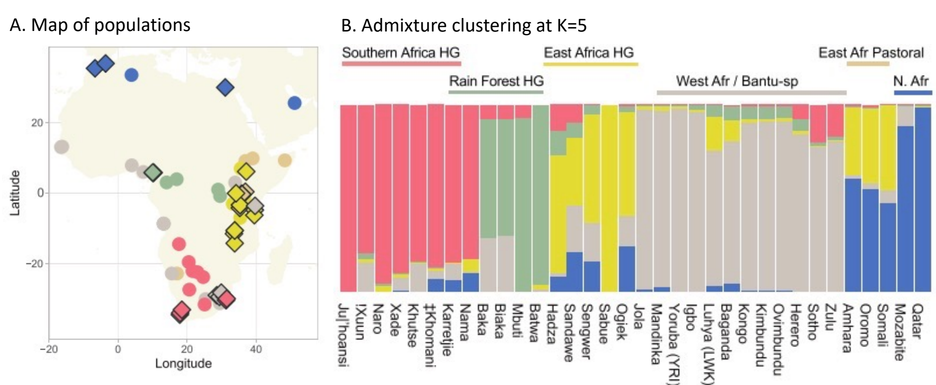

Figure 3.68: **African population structure inferred by Admixture. A.** *Map of population samples. Color labels match those in the right-hand panel. Diamonds indicate ancient DNA.* **B.** *Admixture analysis of modern populations. The labels at the top indicate the major geographic or ethnographic groups; with some exceptions these correspond roughly to Admixture clusters. Credit: Modified from Fig 1 of Mário Vicente and Carina Schlebusch (2020) [Link]. CC BY 4.0.*

• **Khoisan hunter gatherers of Southern Africa** [shown above in red]. The Khoisan (or Khoe-San) refer to a number of groups in the Kalahari region of southern Africa who, as we shall see, have the deepest splitting date of any human population. The Khoisan languages make heavy use of "click" consonants in their speech [n].

[n] *The click languages are quite remarkable and I recommend searching for an online video if you are unfamiliar with them.*

• **Central African rainforest hunter gatherers** [green]. These populations are skilled at foraging in dense tropical rainforests, with specialized techniques for hunting and honey gathering. They are known for their very short stature, which is likely an adaptation to an rainforest environment although the specific reason why this is advantageous is unresolved [507]
• **East Africans** (hunter gatherers and pastoralists [yellow]. Consistent with their geographic location, the East African groups are likely most closely related to the out-of-Africa source population [508]. They include hunter gatherer populations such as the Hadza click speakers [509]
• **West African Bantu farmers** [gray]. The Bantu peoples are traditionally grouped together as speakers of any one of several hundred languages within the Bantu language family. During the Bantu Expansion in the last 5,000 years, they expanded from their ancestral range in Cameroon and Nigeria to colonize most of sub-Saharan Africa.

The origins of **North Africans** [blue] are usually considered separately from sub-Saharan Africans because much of their ancestry is from non-African sources. Aside from occasional greening periods, the Sahara desert has long been a major barrier to gene flow between sub-Saharan regions and North Africa. Ancient DNA work in Morocco indicates that most early north African ancestry came from a non-African source in the eastern Mediterranean; additionally there were smaller sub-Saharan contributions related to East and West Africans [510].

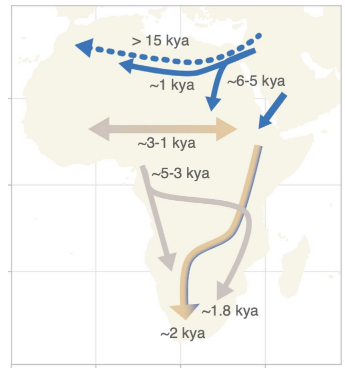

Figure 3.69: **Recent African migration events.** *Blue arrows show back-migrations from the middle east into Africa; gray/brown is migration along the Sahel; gray arrows show the Bantu expansion south and east; brown/blue is the spread of East African pastoralists carrying middle east admixture.* *Credit: Fig 2b of Mário Vicente and Carina Schlebusch (2020) [Link]. CC BY 4.0.*

Recent population movements and admixture. You can see from the Admixture plot above that most of the populations carry ancestry from more than one cluster [511]. One major driver of admixture was the Bantu expansion into regions historically occupied by the Central African, and Southern African hunter gatherer populations (Figure 3.69). We saw an example of this in the last chapter, in the case of the Khoisan group Jul'hoan who carry about 6% Bantu ancestry as the result of an admixture event $\sim 1,000$ years ago (Figure 3.27).

Secondly, there has been a major back-flow of Eurasian genotypes, mainly from the Middle East, back into East African populations during the past $\sim 3,000$ years. Most Kenyan, Tanzanian and Ethiopian populations have from 5%–50% Eurasian ancestry. More recently some of this Eurasian ancestry reached the southern Khoisan populations, likely carried by southward dispersal of East Africans [512].

Deep African population structure. Now that we know a bit about contemporary African population structure, let's explore the deeper historical relations among the major population groups.

The next plot shows the **cross-coalescence** analysis for several African and non-African populations:

As you can see here, for most population pairs, coalescence starts by about 50 KYA, but continues well past 200 KYA. This indicates the presence of long-standing population structure in Africa [o].

[o] *Contrast the timescale here to the plot for non-African populations (Figure 3.67), where all pairs were already above 80% by 50 KYA.*

We also see on the right-hand side of the plot a comparison between

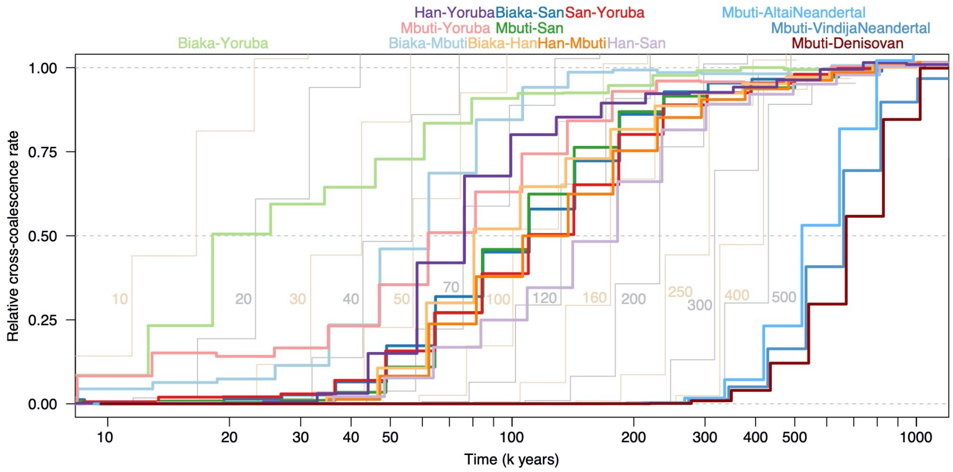

Figure 3.70: **Cross-coalescence of African populations and selected others.** *Pairs indicated by solid colors. The faint background lines show cross-coalescence rates estimated from simulated data with splits at the indicated times (in KY).*
Populations: **Biaka**, **Mbuti**: *Central African hunter gatherers;* **Yoruba**: *West Africans;* **San**: *Southern African Khoisan. From outside Africa:* **Han**: *Chinese;* **Neanderthals**, **Denisovans**: *archaic Eurasians.*
Credit: Fig 5a Anders Bergström et al (2019) [Link] CC-BY-NC-ND 4.0.

Mbuti pygmies and ancient DNA genomes from Neanderthals and Denisovans. These show that the ancestors of Neanderthals and Denisovans diverged from the modern human lineage around 700 KYA.

These plots do show that there was ancient structure at least 200 KYA, that persists to the present day. But this doesn't tell us much about what the population looked like prior to this: in particular, it doesn't show that there was a single ancestral source population at that time. Instead these plots may reflect convergence to a point where the ancestors of modern populations are identically distributed among a set of ancient structured populations.

There has been a great deal of work to estimate detailed, quantitative models of deep human history in Africa, using a variety of methods including composite likelihoods and extensions of PSMC [513]. While there is still a great deal of uncertainty in the details, most models include the following features:

- Europeans are most closely related to East Africans, with a divergence time of 50–100 KYA;
- Four major lineages in Africa (East Africans, West African Bantu, Khoisan, and Central African hunter gatherers) are more divergent than the deepest divergence among non-African populations;
- East and West Africans are most closely related, with deepest divergences to Central African hunter gatherers and Khoisan;
- There is a great deal of disagreement about the deepest divergence times, but these are 200–300 KYA in many models [514];
- Many papers infer some form of super-deep structure that is shared among groups going back 1–2 MY.

One specific model, by Aaron Ragsdale and colleagues, is shown here, though I should remind you that there is still a great deal of uncertainty in the details [515]:

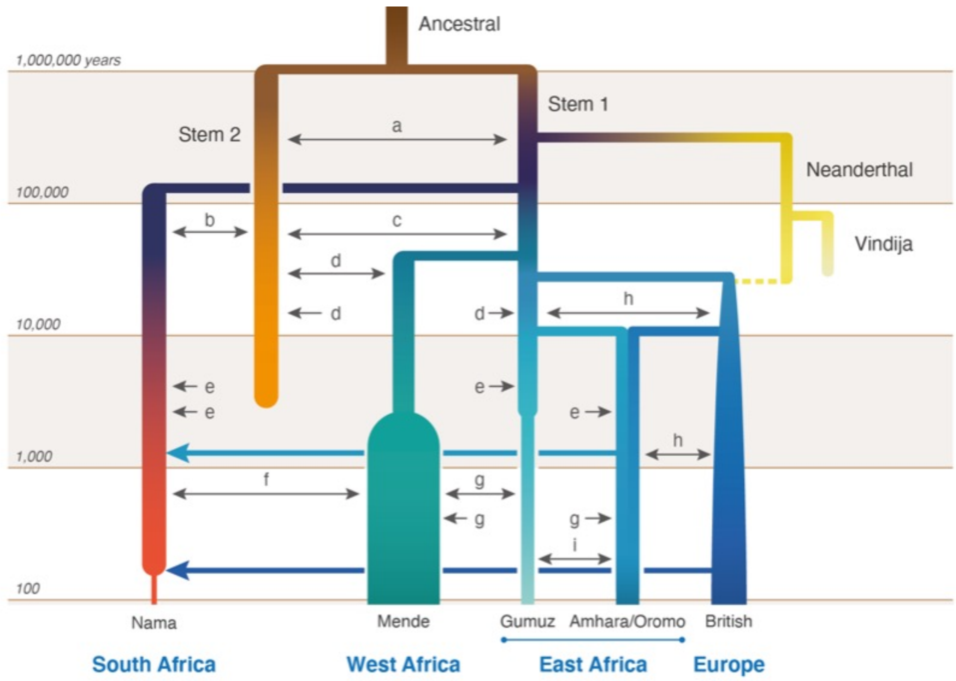

Figure 3.71: **A model for deep African population structure** *based on a composite likelihood approach. Notice that time on the y-axis is shown on a log scale; this model infers deep structure ("Stem 1" and "Stem 2") extending back to 1.1 MYA. While Stem 1 is the majority ancestor of all populations, Stem 2 makes its largest contribution to the Mende branch.*

Notes: Arrows indicate migration rates that are model parameters. This analysis includes three of the main African lineages (Central African hunter gatherers are not shown); Nama are a Khoisan population; Vindija is a high-quality Neanderthal genome. The original paper preferred a more complex model that is not shown here, for clarity. Credit: Fig 3A Aaron Ragsdale et al (2022) [Link] CC-BY-NC-ND 4.0.

This model illustrates the key features that I described above: the non-Africans (here represented by British) are a side branch from east Africans; the model illustrates older structure between South African Khoisan, West and East Africans (Central Africans are not shown but would be a separate deep branch).

One intriguing feature of this model is the inference of super-deep structure back to 1.1 MYA. One population–which the authors refer to as Stem 1–is the major ancestor of all modern populations as well as Neanderthals; Stem 2 is an example of a so-called **ghost population** that no longer exists and can be inferred only through its genetic contributions to other lineages [p] [516]. It's tantalizing to try to connect this type of deep structure to the great diversity in hominin fossils that we discussed at the start of the chapter, although at this time it's extremely difficult to link inferred genetic lineages to specific fossils.

[p] *Several papers agree there is signal for deep population structure back to 1–2 MYA, although the inferred details vary considerably across studies.*

Summary. The combination of paleoanthropology and genetics have combined to sketch out the deep history of the human species.

Paleoanthropology has revealed that during the last 2 million years, the hominins were represented by a striking diversity of forms, originating in Africa, but spreading all the way across Eurasia and even into the islands of southeast Asia. Some archaic forms survived until comparatively recently: Neanderthals and Denisovans, persisted until $\sim$40 KYA, and even the ultra-archaic Flores "hobbits" survived until at least 50 KYA. Fossils with features of modern humans start to appear by around 150–250 KYA ago, though it's often unclear which fossils are, or are not, on the human lineage.

At this point, genetics starts to pick up the story. Genetic analysis has revealed deep, and complex structure of African populations, reaching back 200 KYA or more, and with contributions from unknown, highly-diverged ghost populations. Meanwhile, all non-Africans are (mostly) descended from a single recent out-of-Africa event, spreading rapidly from the Middle East in the last 50 KYA. Within the next 10 to 20 thousand years, all the archaic hominins were extinct, and their only surviving traces are through small genetic contributions to modern populations: an unknown archaic in Africa, and (as we'll see next) Neanderthals and Denisovans in non-Africans.

Our next chapter brings ancient DNA into the story.

Notes and References.

[449]Thanks again to the fantastic generosity of people who commented on earlier drafts of this chapter and the upcoming aDNA chapter (some of whom even commented on multiple versions!): Molly Przeworski, Doc Edge, Chris Stringer, Leo Speidel. Also, huge thanks to Alyssa Lyn Fortier and Clemens Weiß for kindly contributing original figures for these chapters. As always, any errors, omissions or over-simplifications are entirely my own fault.

[450]Reviews in **Paleoanthropology**: Some of the best accessible resources in this space are nonacademic online summaries including decent Wikipedia pages for the various hominin species. This Wikipedia page provides an excellent overview of the major fossils: [Link]. For an overview of early hominid evolution from a fossil perspective see Almécija et al (2021). However as far as I am aware there is not a broad entry-level overview of *Homo* fossils available online. There's a highly accessible book by paleoanthropologists Louise Humphrey and Chris Stringer. It's targeted towards lay readers but it provides a great overview, and also includes beautiful images of many of the key fossils. A moderately advanced overview is provided by Chris Stringer (2016).

Almécija S, Hammond AS, Thompson NE, Pugh KD, Moyà-Solà S, Alba DM. Fossil apes and human evolution. Science. 2021;372(6542):eabb4363

Humphrey L, Stringer C. Our Human Story. Natural History Museum; 2018

Stringer C. The origin and evolution of Homo sapiens. Philosophical Transactions of the Royal Society B: Biological Sciences. 2016;371(1698):20150237

Reviews of **Genetic approaches to human history**: Here there is a lot to choose from, including Nielsen et al (2017); Skoglund and Mathieson (2018); Bergstrom et al (2021) provides probably the best integration of fossil and genetic evidence; and a 2018 book by David Reich on aDNA, which provides a long-form overview of many topics.

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017;541(7637):302-10

Skoglund P, Mathieson I. Ancient genomics of modern humans: the first decade. Annual review of genomics and human genetics. 2018;19(1):381-404

Bergström A, Stringer C, Hajdinjak M, Scerri EM, Skoglund P. Origins of modern human ancestry. Nature. 2021;590(7845):229-37

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

[451]Schmitz et al 2002 provides a nice description of the original discovery, as well as carbon dating results (the specimen is a late-stage Neanderthal, dating to 40 KYA):

Schmitz RW, Serre D, Bonani G, Feine S, Hillgruber F, Krainitzki H, et al. The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. Proceedings of the National Academy of Sciences. 2002;99(20):13342-7

[452]DNA preservation varies greatly depending on temperature and humidity (cold and dry are better). The oldest reported human aDNA is from 430 KYA

Meyer M, Arsuaga JL, De Filippo C, Nagel S, Aximu-Petri A, Nickel B, et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature. 2016;531(7595):504-7.

[453]This unpublished figure by Alyssa Lyn Fortier draws on a number of key sources, including a wonderful unpublished figure by James Cole [Link]; Refined dating estimates for key fossils from Figure 85 of Gruen and Stringer (2023) [Link]; Nielsen et al (2017); Denisovan models from Jacobs et al (2019), African models from Ragsdale et al (2023); Fan et al (2023)

Gruen R, Stringer C. Direct dating of human fossils and the ever-changing story of human evolution. Quaternary Science Reviews. 2023;322:108379

Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple deeply divergent Denisovan ancestries in Papuans. Cell. 2019;177(4):1010-21

Ragsdale AP, Weaver TD, Atkinson EG, Hoal EG, Möller M, Henn BM, et al. A weakly structured stem for human origins in Africa. Nature. 2023;617(7962):755-63

Fan S, Spence JP, Feng Y, Hansen ME, Terhorst J, Beltrame MH, et al. Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. Cell. 2023;186(5):923-39

[454]This model provides a minimal simplification of key events: for example, genetic evidence indicates additional gene flow in and out of Africa at different timepoints, and highly complex patterns of human migration.

[455]I'm presenting a version of the current consensus view of deep human evolution. But it's important to note that the evolutionary connections between the major fossils remain uncertain, and it's still possible that major aspects of the consensus view could change in upcoming years as we get more fossil evidence, more clarity from population genetics, and hopefully additional ancient DNA of key early samples.

As one example of the types of alternatives that might be consistent with present data, David Reich suggests a highly provocative model in his 2018 book (Figure 11 in that book and related text) in which *the main trunk of human evolution* *is actually outside Africa* from 2 MYA - 300 KYA. In that scenario, non-African *H. erectus* is ancestral to Neanderthals and Denisovans, and gives rise to early modern humans via *migration back into Africa*. Results showing deep population structure in human ancestors could be consistent with this model if we interpret the two main compartments as corresponding to populations inside and outside Africa, respectively: Fan et al (2023); Cousins et al (2024).

Cousins T, Scally A, Durbin R. A structured coalescent model reveals deep ancestral structure shared by all modern humans. bioRxiv. 2024:2024-03

[456]You'll see below that most of the different fossil forms are named as distinct species based on their morphology. The species naming is quite subjective (and often debated) and there's no guarantee that named species conform to the Biological Species Concept (i.e., populations that cannot interbreed). In fact, we do know now that humans, Neanderthals and Denisovans could interbreed, and I'd guess many other combinations of the named species could too.

[457]For a pair of excellent blog posts by John Hawks about the earliest hominins see [Link] and [Link].

[458]In an exciting development, an international team recently retrieved a small amount of protein sequence from *Paranthropus* teeth. Proteins can survive much longer than DNA, and these experiments were an intriguing indication of the possibility of retrieving genetic information from very old samples. In this case the limited data were enough to confirm the clustering of Paranthropus outside the human-Neanderthal-Denisovan clade, but did not provide specific novel insights into history.

Madupe PP, Koenig C, Patramanis I, Rüther PL, Hlazo N, Mackie M, et al. Enamel proteins reveal biological sex and genetic variability within southern African Paranthropus. bioRxiv. 2023:2023-07

[459]The species naming in this time period is quite uncertain and controversial. Some researchers prefer to group *H. rudolfensis* with *H. habilis* and *H. ergaster* with *H. erectus*. The African *H. erectus* is sometimes referred to as *H. erectus s.l.* (s.l. standing for "sensu lato" which means "broadly defined") – i.e., grouping the African and non-African populations under the same species name.

[460]Older human remains including jawbones and a phalanx have been found nearby at the Sima del Elefante site and dated to $\sim 1.2 - -1.4$ MYA. Different researchers have hypothesized that these may be *H. erectus* or perhaps very early *H. antecessor*.

[461]P101, Our Human Story.

[462]According to Henry Gee, editor at Nature who handled the paper "I know where I was (when he first heard the news). I was at my desk at @Nature and opened a manuscript on a weird hominin originally called Sundanthropus florianus; a referee noted that the sp name meant 'flowery anus'. After review the authors changed it to Homo floresiensis." [Link]

[463]The species name comes from the old-German spelling 'Neander Thal', referring to the valley ('thal') in which it was found. The 'h' is silent, and spelled 'tal' in modern German. Hence 'Neanderthal' is sometimes spelled 'Neandertal', and pronounced *Neandertal* regardless of spelling. In the science literature the species is also occasionally referred to as "Neandertals", e.g., in the 2010 draft genome paper:

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710-22

[464]The skulls have been assigned species names *H. longi* and *H. daliensis* but it seems likely that these may be physical remains of Denisovans.

Ni X, Ji Q, Wu W, Shao Q, Ji Y, Zhang C, et al. Massive cranium from Harbin in northeastern China establishes a new Middle Pleistocene human lineage. The Innovation. 2021;2(3)

[465]For an excellent synthesis of these topics see Bergström et al (2022).

[466]There's also evidence for recurrent gene flow *from early humans into eastern Neanderthals* in the past $\sim$200 KYA. This would also suggest early habitation of Eurasia by humans.

Kuhlwilm M, Gronau I, Hubisz MJ, De Filippo C, Prado-Martinez J, Kircher M, et al. Ancient gene flow from early modern humans into Eastern Neanderthals. Nature. 2016;530(7591):429-33

Li L, Comi TJ, Bierman RF, Akey JM. Recurrent gene flow between Neanderthals and modern humans over the past 200,000 years. Science. 2024;385(6705):eadi1768

[467]Wolpoff et al (1984). In 1992, Thorne and Wolpoff wrote that "Multiregional evolution traces all modern populations back to when humans first left Africa at least a million years ago, through an interconnected web of ancient lineages... Today distinctive populations maintain physical differences despite inter-breeding and population movements; this situation has existed ever since humans first colonized Europe and Asia. Modern humanity originated within these widespread populations"

Wolpoff MH, Wu X, Thorne AG. Modern Homo sapiens origins: a general theory of hominid evolution involving the fossil evidence from East Asia. The origins of modern humans: a world survey of the fossil evidence. 1984;6:411-

83

Thorne AG, Wolpoff MH. The multiregional evolution of humans. Scientific American. 1992;266(4):76-83

[468]Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. Nature. 1987;325(6099):31-6

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. African populations and the evolution of human mitochondrial DNA. Science. 1991;253(5027):1503-7

Tragically, Alan Wilson died of leukemia at the age of 56, two months before publication of the 1991 paper.

An early criticism of the paper was that other tree topologies that included deep non-African lineages were nearly-as-good according to the parsimony criterion used to build the tree. However, subsequent work using entire mtDNA genomes confirmed Vigilant's main result that the deepest lineages of the tree are entirely among African samples:

Maddison DR, Ruvolovo M, Swofford DL. Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. Systematic Biology. 1992;41(1):111-24

Ingman M, Kaessmann H, Pääbo S, Gyllensten U. Mitochondrial genome variation and the origin of modern humans. Nature. 2000;408(6813):708-13

[469] The mtDNA mutation rate of 10^{-6}/bp/generation is about 100-times higher than in the nuclear genome. The authors would have had to do 100-fold more sequencing to get similar amounts of information from a nuclear locus.

[470]Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, Bollongino R, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. Current Biology. 2013;23(7):553-9

[471]Poznik GD, Henn BM, Yee MC, Sliwerska E, Euskirchen GM, Lin AA, et al. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science. 2013;341(6145):562-5

[472]Aspects of both the mitochondrial and Y chromosome data are highly unusual. In particular, the MRCA times in these regions are extraordinarily recent; one study has estimated that no autosomal region has an MRCA time less than about 300 KYA (see the discussion starting on page 77 of Mallick et al (2016), although this estimate may be biased toward neutrality by the PSMC prior. The aberrantly short MRCA times of the mtDNA and Y in part reflect that these maternally and paternally inherited regions are expected to have 1/4 the effective population size of autosomal loci. The effective population size of the Y chromosome may be further reduced by the higher variance in reproductive success experienced by males relative to females, and N_e is likely reduced even further by background selection against deleterious variants, as background selection is particularly strong in non-recombining regions such as these. Nonetheless, it's striking that these MRCA times are actually younger than deep population structure in Africa (as we shall see). Perhaps this points to some additional role of positive selection in the mtDNA and Y chromosome.

[473]While the major conclusions from mtDNA about Recent African Origins were correct, mtDNA has sometimes been actively misleading in other contexts, including ancient DNA studies of Neanderthals and Denisovans. The earliest work on Neanderthal ancient DNA found that Neanderthals are completely outside the modern range of variation. This was taken as evidence against gene flow between humans and Neanderthals (Krings et al 1997), though we now know that this did in fact take place. Even at the time, it was pointed out that data from a single locus cannot completely resolve this question (Nordborg 1998). In a similar vein, the first Denisova mtDNA sequence suggested that Denisovans were much more diverged from Neanderthals than implied by the rest of the genome (Krause et al 2010).

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S. Neandertal DNA sequences and the origin of modern humans. cell. 1997;90(1):19-30

Nordborg M. On the probability of Neanderthal ancestry. The American Journal of Human Genetics. 1998;63(4):1237-40

Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, et al. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature. 2010;464(7290):894-7

[474]Original heterozygosity analysis was performed by Sohini Ramachandran et al (2005). Original LD analysis was by Mattias Jakobsson et al (2008). This version of the figure is from a redrawing by Michael DeGiorgio et al (2009).

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences. 2005;102(44):15942-7

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008;451(7181):998-1003

DeGiorgio M, Jakobsson M, Rosenberg NA. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proceedings of the National Academy of Sciences. 2009;106(38):16057-62

[475]As pointed out by Pickrell and Reich (2014), other models with varying combinations of bottlenecks and admixture can also produce similar patterns of diversity.

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

[476]Neanderthal ancestry in African populations is due to reverse gene flow from Middle Eastern populations back into Africa

Chen L, Wolf AB, Fu W, Li L, Akey JM. Identifying and interpreting apparent Neanderthal ancestry in African individuals. Cell. 2020;180(4):677-87

Harris DN, Platt A, Hansen ME, Fan S, McQuillan MA, Nyambo T, et al. Diverse African genomes reveal selection on ancient modern human introgressions in Neanderthals. Current Biology. 2023;33(22):4905-16.

[477]Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505(7481):43-9

Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. The date of interbreeding between Neandertals and modern humans. PLOS Genetics. 2012

Moorjani P, Sankararaman S, Fu Q, Przeworski M, Patterson N, Reich D. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. Proceedings of the National Academy of Sciences. 2016;113(20):5652-7

Iasi LN, Chintalapati M, Skov L, Mesa AB, Hajdinjak M, Peter BM, et al. Neandertal ancestry through time: Insights from genomes of ancient and present-day humans. bioRxiv. 2024

[478]There is evidence for small genetic contributions from a "basal Eurasian" population that preceded Neanderthal admixture. This has also been suggested in Oceania but the evidence is controversial. For more on this see the section on "Early Expansion Hypotheses" in Bergstrom et al 2021.

[479]While there was in fact interbreeding between non-African modern *H sapiens* and the second-wave archaic groups, Neanderthals and Denisovans, the fact of this gene flow does not rescue the multiregional model as there is no continuity with earlier groups such as *Homo erectus*. Moreover, the degree of Neanderthal+Denisovan ancestry is very low, between 1.5–5% in most non-African populations, it generally avoids regions of the genome of functional importance, and it did not contribute to morphological continuity. For more on this see

Stringer C. Why we are not all multiregionalists now. Trends in ecology & evolution. 2014;29(5):248-51.

[480]One famous example where "vibes" were misleading was when the first studies of Neanderthal mtDNA showed Neanderthals entirely outside the human phylogeny. The authors concluded that Neanderthals did not interbreed with humans; more detailed modeling showed that this conclusion was not strongly justified (and we now know that in fact they did interbreed): Krings et al (1997); Nordborg (1998), cited above.

[481]It's conventional to ignore positive selection in these models, under the rationale that strong positive selection only impacts a small fraction of the genome; besides, inference is hard enough even with neutral models. Background selection is also usually ignored; however it has meaningful and varying effects on N_e across the genome and probably deserves more attention in demographic inference.

[482]For example msprime by Jerome Kelleher and colleagues, which implements ultra-fast coalescent models [Link]; or SLiM by Philip Messer and colleagues, which uses forward simulations in a very flexible framework [Link].

[483] Statistical theory tells us that when we have well-defined models like in population genetics, that we should compute what is known as the **likelihood** of the data: i.e., the probability of generating the observed data for a given set of parameter values. (Here the parameters are features of the model such as population split times.) Loosely speaking, our estimates of the parameters should be centered around the values that produce the highest likelihoods. Since genome data sets are large and complex, the likelihood of observing any given data set is astronomically tiny, but all we care about is whether the likelihood is higher for some parameters than for others: e.g., do the data become more or less likely if we assume an earlier population split time?

In principle the likelihood can be computed by integrating over all possible ARGs: the probability of each ARG given the history times the probability of the genotype data given that ARG. Unfortunately it is extremely difficult to compute, or even approximate this.

[484]For example we could use F_{ST} to estimate population divergence, or heterozgyosity to estimate the mutation-drift parameter $4N\mu$.

[485]The nth moment of a random distribution is defined as $\mathrm{E}[X^n]$ where X is a random draw from that distribution. The term *moment estimator* refers to methods that use theory to predict moments of a distribution as a function of relevant parameters, and then match those against the corresponding values in the data.

[486]Wakeley J, Hey J. Estimating ancestral population parameters. Genetics. 1997;145(3):847-55

See also an interesting recent approach by

Sjödin P, McKenna J, Jakobsson M. Estimating divergence times from DNA sequences. Genetics. 2021;217(4):iyab008

[487]The core ABC concepts were developed in a theory paper by Tavare et al (1996) [REF]. The first implementation of ABC methods was by Pritchard et al (1999). The methods were formalized in a landmark paper by Beaumont et al (2003), which coined the term ABC, and really helped to promulgate these techniques within population genetics and beyond. In another important paper, Marjoram et al (2003) showed how this could be incorporated into MCMC approaches. A popular implementation is provided by ABCtoolbox (Wegmann et al 2010).

Tavaré S, Balding DJ, Griffiths RC, Donnelly P. Inferring coalescence times from DNA sequence data. Genetics. 1997;145(2):505-18

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular biology and evolution. 1999;16(12):1791-8

Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162(4):2025-35

Marjoram P, Molitor J, Plagnol V, Tavaré S. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences. 2003;100(26):15324-8

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC bioinformatics. 2010;11:1-7

Sisson SA, Fan Y, Beaumont M. Handbook of approximate Bayesian computation. CRC press; 2018

[488]In parallel, several papers have used simulation-based "likelihood" approaches: including

Weiss G, von Haeseler A. Inference of population history using a likelihood approach. Genetics. 1998;149(3):1539-46

Wall JD, Lohmueller KE, Plagnol V. Detecting ancient admixture and estimating demographic parameters in multiple human populations. Molecular biology and evolution. 2009;26(8):1823-7

[489]Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proceedings of the National Academy of Sciences. 2005;102(51):18508-13

[490]Applications of deep learning in population genetics include:

Sheehan S, Song YS. Deep learning for population genetic inference. PLoS computational biology. 2016;12(3):e1004845

Schrider DR, Kern AD. S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS genetics. 2016;12(3):e1005928

Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. Trends in Genetics. 2018;34(4):301-12

[491]Here, the training data are simulated under a variety of parameter values and/or distinct models, and the classifier is designed to minimize error under this input distribution. In this way, the input distribution forms an implicit prior on the parameters.

[492]Minor note about notation: in the one-population case we would usually drop n_m because this corresponds to a state with no variation and is wrapped into n_0. But we do need this for the multi-population case as a site may have variation in one population but be fixed for the derived allele in another. So I'm leaving n_m in the one-population case to avoid complicating the notation.

[493] Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS genetics. 2009;5(10):e1000695

Kamm J, Terhorst J, Durbin R, Song YS. Efficiently inferring the demographic history of many populations with allele count data. Journal of the American Statistical Association. 2020;115(531):1472-87

Ragsdale AP, Gravel S. Models of archaic admixture and recent history from two-locus statistics. PLoS genetics. 2019;15(6):e1008204

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome research. 2005;15(11):1566-75

Hudson RR. Two-locus sampling distributions and their application. Genetics. 2001;159(4):1805-17

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304(5670):581-4

[494]Kuhner MK, Yamato J, Felsenstein J. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics. 1998;149(1):429-34

Fearnhead P, Donnelly P. Estimating recombination rates from population genetic data. Genetics. 2001;159(3):1299-318

[495]Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-wide inference of ancestral recombination graphs. PLoS genetics. 2014;10(5):e1004342

Hubisz MJ, Williams AL, Siepel A. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. PLoS genetics. 2020;16(8):e1008895

[496] Minichiello MJ, Durbin R. Mapping trait loci by use of inferred ancestral recombination graphs. The American Journal of Human Genetics. 2006;79(5):910-22

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

Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. Nature genetics. 2019;51(9):1321-9

Deng Y, Nielsen R, Song YS. Robust and accurate bayesian inference of genome-wide genealogies for large samples. bioRxiv. 2024:2024-03

[497]Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475(7357):493-6

[498]This is related to an earlier concept called Bayesian Skyline Plots, although those were generally focused on individual loci:

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular biology and evolution. 2005;22(5):1185-92.

[499]A very early version of this idea for $m = 2$ was developed by

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, et al. Human genome sequence variation and the influence of gene history, mutation and recombination. Nature genetics. 2002;32(1):135-42.

[500]To get an order of magnitude estimate of the number of recombination switches in a single diploid genome, we can note that the average size of an unrecombined block is about $(2 \times 2N_e r)^{-1}$ where $2N_e$ is the average coalescence time for a pair of sequences, and r is the average recombination rate per base pair per generation, and recombination can occur on either of the two lineages. Then taking the genome size as G, the predicted number of blocks is about $4N_e rG$ or roughly 2×10^6 for humans.

[501]The PSMC extension, MSMC, can consider more than 2 haplotypes simultaneously, and focuses on the *first coalescence* in the sample (the first M in SMC is for 'Multiple'). Another important advance, SMC++ from Yun Song's lab, incorporates SNP data into MSMC-style modeling. For a good example of SMC++ in action see Figure 4 of Bergström et al (2020). Notice that SMC++ also provides smoothed estimates of the N_e trajectory.

Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nature genetics. 2014;46(8):919-25

Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nature genetics. 2017;49(2):303-9

Wang K, Mathieson I, O'Connell J, Schiffels S. Tracking human population structure through time from whole genome sequences. PLoS genetics. 2020;16(3):e1008552

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367(6484):eaay5012

[502]Schiffels and Durbin (2014), cited above.

[503]The lowest cross coalescent rate at the right-hand edge of the plot is for the Han-Papuan comparison. This is likely because Papuans have an unusually high fraction of Denisovan ancestry. It's not clear (to me) why Han-Sardinian also converges slowly to 1, or if this is perhaps a technical artifact.

[504]Ancient DNA degrades quickly in hot climates and so far successful aDNA work in sub-Saharan Africa has been more limited than in temperate regions of the globe. At present most successful aDNA work has been done on samples from the last 10 KYA. These are useful for identifying the major strands of population structure prior to recent population movements and expansions, but so far there's no analog of the wildly successful work on Neanderthals and Denisovans from Eurasia.

Lipson M, Sawchuk EA, Thompson JC, Oppenheimer J, Tryon CA, Ranhorn KL, et al. Ancient DNA and deep population structure in sub-Saharan African foragers. Nature. 2022;603(7900):290-6

[505]There are many good reviews in this area. These include an excellent special issue in Human Molecular Genetics: [Link], with reviews on deep structure in Africa; genetic histories of the Bantu, Ethiopians, Khoisan, and north Africans. I also recommend Vicente and Schlebusch (2020). David Reich's book has a detailed chapter on Africa. There's also a short but helpful section in the Bergstrom et al (2020) review cited above.

Vicente M, Schlebusch CM. African population history: an ancient DNA perspective. Current Opinion in Genetics & Development. 2020;62:8-15

[506]The figure is taken from Vicente and Schlebusch (2020), cited above. The classic study of African population structure is by Tishkoff et al (2009). For other relevant work see Fan et al (2019), cited above.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. science. 2009;324(5930):1035-44

[507]Perry GH, Foll M, Grenier JC, Patin E, Nédélec Y, Pacis A, et al. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. Proceedings of the National Academy of Sciences. 2014;111(35):E3596-603

Perry et al (2014) reported that Batwa have an average height of 152.9cm in male, and 145.7cm in females, and showed evidence that their short stature has a genetic basis

Perry GH, Foll M, Grenier JC, Patin E, Nédélec Y, Pacis A, et al. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. Proceedings of the National Academy of Sciences. 2014;111(35):E3596-603

[508]This conclusion is somewhat complicated by the presence of migration backflow *from* Eurasia in modern East Africans. However, analysis of the 4500 year old Mota sample from Ethiopia finds that this is closest to Eurasians.

Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, et al. Reconstructing prehistoric African population structure. Cell. 2017;171(1):59-71

Lipson M, Sawchuk EA, Thompson JC, Oppenheimer J, Tryon CA, Ranhorn KL, et al. Ancient DNA and deep population structure in sub-Saharan African foragers. Nature. 2022;603(7900):290-6

[509]The Hadza are notable as East African click speakers; intriguingly there is evidence for gene flow from Khoisan suggesting a possible genetic link to the south African click speakers: Fan et al (2019), as well as cattle-herders (whom we met previously in the context of lactase adaptation, Figure 2.100).

Fan S, Kelly DE, Beltrame MH, Hansen ME, Mallick S, Ranciaro A, et al. African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. Genome Biology. 2019;20:1-14

[510]Van de Loosdrecht M, Bouzouggar A, Humphrey L, Posth C, Barton N, Aximu-Petri A, et al. Pleistocene North African genomes link near Eastern and sub-Saharan African human populations. Science. 2018;360(6388):548-52

[511]Even though African ancient DNA work has not yet been able to reach very far back into the past (at the time of writing the oldest genomes are 18,000 years old), these fairly recent large-scale movements and admixture events highlight the value of African ancient DNA for clarifying population structure prior to the major movements, including the Skoglund and Lipson papers cited above.

[512]Pickrell JK, Patterson N, Loh PR, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa. Proceedings of the National Academy of Sciences. 2014;111(7):2632-7

[513]Much of this work is summarized in an excellent review by Hollfelder et al. Key papers that have appeared since that review include Fan et al (2022); Ragsdale et al (2023); Cousins et al (2024), cited above.

Hollfelder N, Breton G, Sjödin P, Jakobsson M. The deep population history in Africa. Human Molecular Genetics. 2021;30(R1):R2-R10

[514]Hollfelder et al summarize the estimates as "The divergence between the ancestors of the Khoe-San and the ancestors of the rest of modern humans is estimated to between 340,000 and 200,000 ya, and with younger estimates (160,000–90,000) based on the MSMC cross-coalescence approach. The next event assuming a simplified bifurcating tree is a divergence between the RHG ancestors and the ancestors of the rest of modern humans (minus the Khoe-San); the estimates vary from 350,000 to 70,000 ya but are generally more recent than the Khoe-San divergence. Eastern African groups, including hunter-gatherers, such as the Hadza and the Sandawe, point to divergences from all other African groups, including western Africans, at $\sim$140,000–70,000 ya." [lightly edited for clarity]

[515]Ragsdale et al (2023), cited above.

[516]Many papers agree on the presence of some kind of deep structure – though the time estimates vary considerable so it's not clear that they are all detecting the same structure.