How Have Humans Populated the Earth?

Out of Africa

An evolutionary quandary in Darwin's writings

Charles Darwin's 1859 *On the Origin of Species by Natural Selection* suffers from a glaring omission. If natural selection is a universal phenomenon, and humans are undoubtedly a part of nature, how then have we evolved? Darwin did opine on the matter, but only sparingly:

> *"In the distant future I see open fields for far more important researches. Psychology will be based on a new foundation, that of the necessary acquirement of each mental power and capacity by gradation. Light will be thrown on the origin of man and his history."*

When Carl Linnaeus published the tenth edition of his famous taxonomy a century earlier, he had been so bold as to combine humans and monkeys into a single order, "Primates". Our old friend Comte de Buffon from Chapter 2 even stated in 1766 that in terms of anatomy, the orangutan

> *"is only an animal, but a very singular animal, which man cannot view without returning to himself".*

Nevertheless, the concept that we share a common ancestor with monkeys was a far more radical one than the idea that natural selection had created giant turtles in the Galapagos. Darwin himself remained silent on the issue, since there were no clear fossil records linking us to our evolutionary predecessors. Little did he know that the answer to the puzzle of our origins was already dawning across the English Channel.

The fossil record elucidates human evolutionary history

Three years earlier, miners working in a German cave had uncovered heavily calcified skeletal remains. The excavators initially thought that the bones belonged to extinct cave bears, which we mentioned in Chapter 11 as being extremely common in Europe until the most recent ice age. Yet some researchers suspected that these fossils belonged to a member of an early human population. The name that they bestowed upon these ancient humans was borrowed from the valley in which the remains had been excavated: **Neanderthals**.

In the ensuing years, the fossil record accumulated evidence in favor of our evolution several million years ago from a chimp-human ancestor. Throughout the 20th Century, scientists discovered fossils from intermediate species between chimps and Neanderthals like australopithecines, *Homo habilis*, and *Homo erectus* (Figure 1).

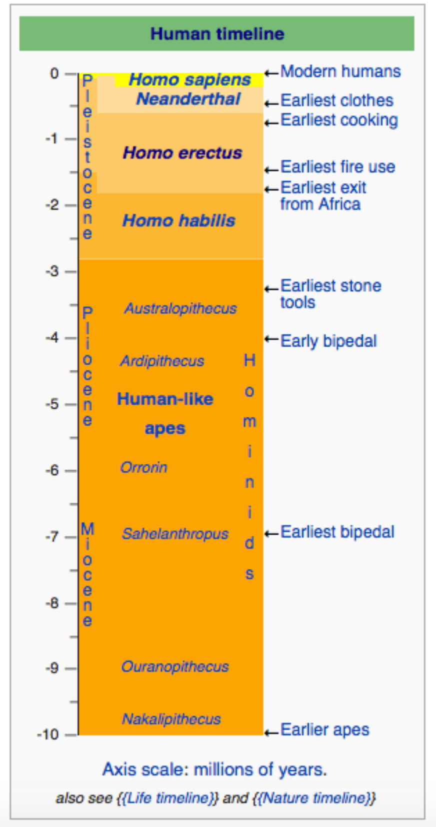


Figure 1: A timeline of recent human ancestors. All these species have only been discovered in Africa until *Homo erectus*, whose fossils have been found across Europe and Asia. Neanderthals arose in Europe and are often considered distinct from *Homo sapiens*, a species that includes Cro-Magnons. Courtesy: Wikipedia.

Until 2 million years ago, all our previous ancestors had been confined to Africa. Yet researchers have discovered *Homo erectus* fossils throughout Eurasia. We can therefore conclude that *Homo erectus* must have been the first human ancestor to emigrate from Africa and survive long enough for us to be able to find their fossils.

Genetic data resolves the origin of modern humans

What happened next became the subject of a much larger debate. **Monogenists** held that despite the far-flung travels of *Homo erectus*, modern humans arose recently in only one location. **Polygenists**, on the other hand, believed that modern humans evolved from *Homo erectus* along distinct paths in different corners of the globe. The polygenist camp would come to house a great number of racist and eugenicist supporters, since their theory allowed for a crystalline view of human race.

The issue of human origins is one more biological question that the advent of genetic data would help resolve. In 1987, Allan Wilson (along with Rebecca Cann and Mark Stoneking) three scientists provided resounding support for monogenism by demonstrating that all modern humans share a common female ancestor who lived approximately 200,000 years ago in Africa. Occam's razor led Wilson to propose what became known as the **Out of Africa hypothesis:** despite the earlier movements of *Homo erectus*, all non-Africans trace their roots to a recent migration of modern humans from Africa. Subsequent, more accurate, research has dated this migration as taking place approximately 70,000 years ago.

The Out of Africa hypothesis helps explain a substantial evolutionary riddle. The fossil record indicates that Neanderthals inhabited Europe from several hundred thousand years ago until approximately 40,000 years ago, at which point they were quickly replaced by the physically weaker **Cro-Magnons**, who resemble modern humans. Polygenists are forced to conclude, illogically, that Neanderthals evolved practically overnight into the very different Cro-Magnons. Yet the Out of Africa hypothesis permits the explanation that Neanderthals were a separate species, *Homo neanderthalensis*, who were less fit than the Cro-Magnons and were displaced over a period of hundreds of generations. This conclusion has been supported by fossil evidence, which shows that Neanderthals and Cro-Magnons coexisted in the same regions before Neanderthals' extinction.

We are left with several questions. First, how exactly does genetic evidence support the Out of Africa hypothesis? During their coexistence in Europe, was any romance kindled between Neanderthals and Cro-Magnons, and if so, how much? What were the paths that modern human populations took as they emigrated from Africa across the globe? And finally, how have these distinct populations contributed to your own genetic identity?

Mitochondrial DNA Confirms the Out of Africa Hypothesis

Mitochondria provide record of female inheritance

Billions of years ago, in the darkness of the primordial ocean, an ancient cell engulfed a bacterium. The bacterium was more efficient at producing energy than the cell, and so the two organisms became symbiotic; as the cell replicated, so did the bacterium within it. Over the eons, the bacterium slowly lost its identity as a distinct organism, and it eventually became an organelle serving as the energy center of the eukaryotic cell, which we now know as the **mitochondrion** (Figure 2)**.**

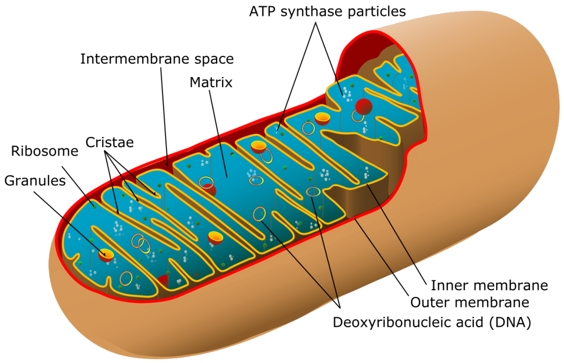


Figure 2: Diagram of an animal mitochondrion. Courtesy Mariana Ruiz Villarreal.

Many researchers support this origin story for the mitochondrion in part because it has retained its own **mitochondrial genome** (**mtDNA**), a short circular chromosome that replicates independently of nuclear DNA. Most mtDNA comprises just 37 genes, all of which are critical to mitochondrial functions. Only 13 of these genes encode proteins – the rest are translated into noncoding RNAs. Furthermore, mtDNA has practically uniform length across species --- throughout the animal kingdom, the mitochondrial genome is approximately 16,000 nucleotides long (in humans, the exact number of nucleotides is typically 16,569). The mitochondrion even has its own genetic code, as four of the 64 RNA codons translate into a different amino acid in the mitochondrion than they do in the nucleus.

The mitochondrial genome's short, uniform length across different species make it a perfect candidate for inexpensive comparative studies (Frederick Sanger sequenced human mtDNA all the way back in 1981). However, the mitochondrion has one additional interesting property: in mammals, sperm mitochondria are usually destroyed during fertilization. This means that you almost certainly inherited all your mitochondrial DNA from your mother, who inherited hers from her mother, and so on, back through the centuries to the most recent common female ancestor of all modern humans, who is commonly called **mitochondrial Eve**.

A human phylogeny from mitochondrial DNA

But how old was Eve, exactly? Since mtDNA acquires mutations over time, we can construct an evolutionary tree from modern mitochondrial genomes. Furthermore, in Chapter 9, we saw that the rooted tree constructed by UPGMA presumes an "evolutionary clock", in which the age of an internal node corresponds to its distance from the leaves beneath it. If we use UPGMA to construct the desired tree, then the age of the root will correspond to the age of mitochondrial Eve.

Exercise Break: Apply UPGMA to construct an evolutionary tree based on human mtDNA data (click here for data). You can use MEGA to generate this tree.

As shown in Figure 3, which contains an evolutionary tree of human mtDNA, all non-Africans clump together on one side of the root, but Africans can be found on both sides of the tree. In other words, Africans must be more genetically diverse than non-Africans, a result that has been confirmed by more advanced subsequent studies.

STOP and Think: What is the Occam's Razor explanation of the fact that Africans are more diverse than non-Africans?

If non-Africans clump together on one side of the tree of modern humans, then it goes to reason that they share a *more recent* common ancestor than all humans do. This deduction led Wilson to propose the out of Africa hypothesis, proposing that this non-African common ancestor must have migrated out of Africa relatively recently. The argument is far from a proof, but it gives damning evidence against polygenism.

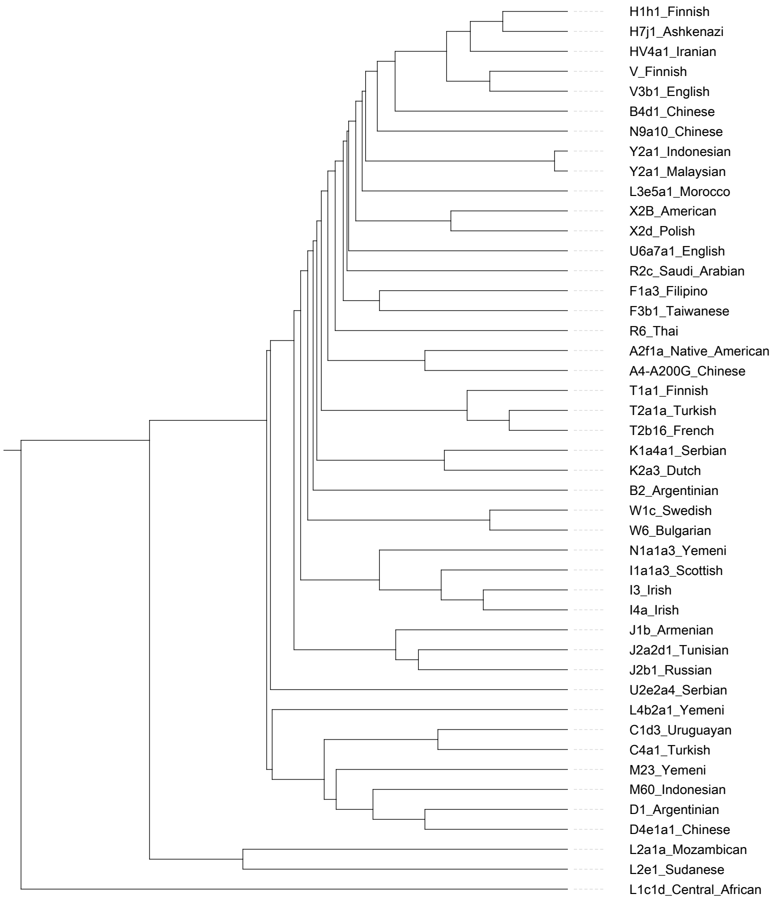


Figure 3: A rooted evolutionary tree constructed by UPGMA from a multiple alignment of complete mitochondrial genomes taken from present-day humans. There are two main subtrees derived from the ancestor. One contains exclusively Africans; the other contains a wide variety of individuals, including Africans.

The Out of Africa hypothesis is only as sound as the dataset used to deduce it, which in 1987 contained only 147 samples. For example, if a 148th sample were discovered to be very different from all existing samples in the tree, it would move the most recent common ancestor of all humans far backward in time. Researchers have not been surprised in the last three decades of mitochondrial studies, but they did receive a shock recently when analyzing Y chromosomes, which are inherited through the male line (see **DETOUR: Aging Y Chromosomal Adam**).

Aging Mitochondrial Eve

Now that we know that mitochondrial Eve was African, our goal is to determine when she lived.

Exercise Break: Use the multiple alignment you constructed to compute the average number of point mutations between mitochondrial Eve and modern humans. Do we now have enough information to age Mitochondrial Eve?

The number of mutations between mitochondrial Eve and modern humans does not help us age Eve without also knowing the **mutation rate** of mtDNA, i.e., the average rate at which mtDNA mutates at the population level over time. But how can we infer this mutation rate?

In practice, researchers often use the fossil record as a proxy for estimating the mutation rate: if fossils indicate that two similar species diverged about a million years ago, and their mitochondrial genomes exhibit x point mutations, then we can conclude that there are approximately $x/2$ mtDNA mutations per million years. Accordingly, Wilson estimated the mutation rate at 2-4% of the entire mitochondrial genome per million years. In the following exercise, we allow you to replicate the Out of Africa hypothesis by estimating the age of mitochondrial Eve yourself.

Exercise Break: Using your estimate from the previous exercise, how old is Eve if the mutation rate is 2% per million years? What if the mutation rate is 4% per million years? How old is the most recent ancestor of all non-Africans?

STOP and Think: The estimate in the preceding exercise relies on the assumption of a constant mutation rate across time and across the length of the mitochondrial genome. Do you think that it is accurate? What are the potential flaws with this assumption?

Unfortunately, estimating the mutation rate in humans can prove difficult. For one, the fossil record has indicated that the size of the human population has fluctuated widely over time. The more individuals there are in a species, the faster that mutations will propagate throughout the species, and a population bottleneck can reduce genetic diversity. As a result, researchers have disagreed about mutation rates.

Further complicating matters, different regions of the mitochondrial genome have different mutation rates. For example, the **control loop** region of mtDNA, which contains the mitochondrion's origin of replication as well as noncoding DNA, is highly variable. Nuclear DNA, which has a different mutation rate than mtDNA, has also been inserted in some parts of the mitochondrial genome.

Nevertheless, multiple studies using differing techniques have all estimated Eve as between 100,000 and 200,000 years old. This estimate seems reasonable in light of the 200,000 year-old **Omo remains** from Kenya, the oldest *Homo sapiens* fossils ever discovered. But where do Neanderthals fit into the picture of human migrations?

Did Neanderthals Contribute to Your Genome?

The older a fossil, the more difficult it becomes to extract DNA from it, and many scientists believe that DNA cannot be read if it is more than one million years old. Fortunately, mtDNA is relatively less difficult than nuclear DNA to isolate from a sample – even one that is hundreds of thousands of years old – because of its high **copy number** in the cell. That is, there are hundreds of thousands of copies of mitochondria within the typical human cell (but just one nucleus), and all these mitochondria essentially share the same genome.[1]

STOP and Think: Why might contamination be particularly difficult to identify when attempting to isolate Neanderthal DNA?

The first reliable Neanderthal mtDNA sample was sequenced in 1997, after being isolated from the original German Neanderthal fossil. Since that time, researchers have sequenced mtDNA from several more Neanderthals, as well as from **Denisovans**, another population of early Eurasian humans whose fossils were first discovered in Siberia in 2008.

Our goal is to age the most recent common ancestor of modern *Homo sapiens* and Neanderthals. To do so, we will add Neanderthal mtDNA samples to the existing human samples and construct an evolutionary tree for the augmented collection. We will then use the estimated age of mitochondrial Eve to calibrate this tree's molecular

[1] The assumption that every mitochondrion in an organism has the same genome is not a perfect one, since mtDNA can have many variations within the same cell, not to mention within the same tissue, or within the same organism. Variants can arise at any of these levels in an individual, a phenomenon called **heteroplasmy**.

clock and assign a date to the root, which corresponds to the most recent common ancestor of modern humans and Neanderthals.

Exercise Break: Add five Neanderthal samples (click here for data) and three Denisovan samples (click here for data) to the set of *Homo sapiens* samples and construct the resulting evolutionary tree using UPGMA. What is the age of the most recent Neanderthal-modern human ancestor?

The expanded mtDNA phylogeny, shown in Figure 4 (left), indicates that Neanderthals and Denisovans diverged from *Homo sapiens* much earlier than the out of Africa migration. In fact, the shared ancestor of the three groups is YYY years old, vastly predating mitochondrial Eve.

However, just because an evolutionary tree indicates a divergence between modern humans and Neanderthals does not imply that there was no **gene flow**, or exchange of genetic material, among the two populations due to interspecies relationships. In fact, there is some genetic evidence in favor of this hypothesis (see **DETOUR: Gene Flow Between Neanderthals and Cro-Magnons**), which has led researchers to draw the evolutionary tree of humans to indicate this gene flow (Figure 4 (right)).

Now that we have strong evidence in support of the Out of Africa hypothesis, we ask ourselves, "What happened once humans left Africa?" As human populations spread out around the globe, they must have left clues to their movements, especially once these populations became isolated and their genomes started diverging. We will therefore try to reconstruct the paths modern human populations took by unearthing the hidden indicators lurking within our genomes.

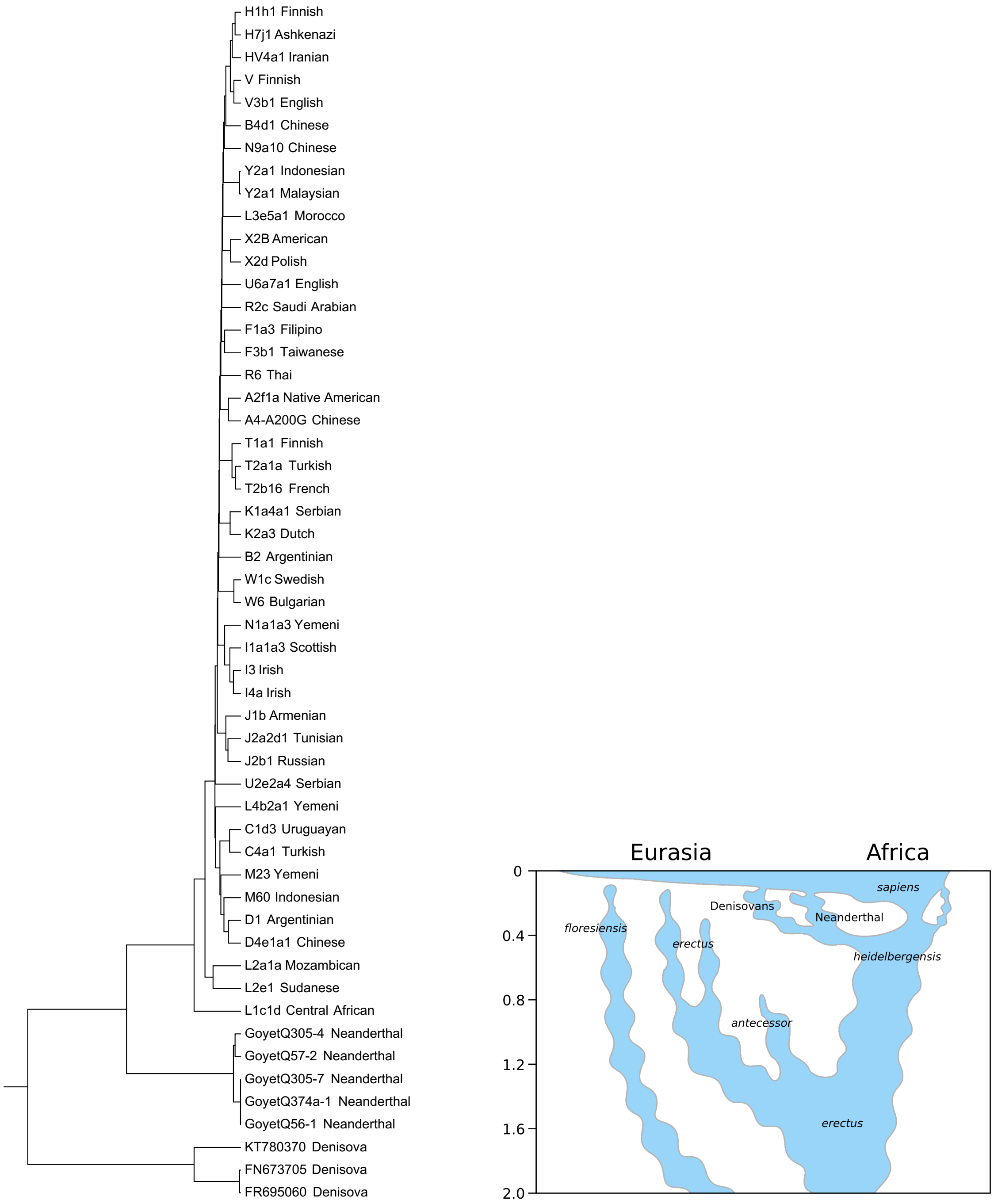


Figure 4: (Left) A UPGMA phylogeny constructed from the modern human mitochondrial genomes in Figure 3, along with five Neanderthal samples and three Denisovan samples. (Right) A rendering of the evolutionary tree for genus *Homo* showing intermingling of Neanderthals and Denisovans with *Homo sapiens* after the second out of Africa migration. Courtesy Chris Stinger.

Identifying Genetic Markers to Compare Humans

Introduction to genotyping

Despite the falling cost of DNA sequencing, it still costs on the order of $200 to sequence a complete mitochondrial genome. Yet any two humans share most of their mtDNA. So rather than resorting to full mitochondrial genome sequencing, it would be much more cost effective to examine the regions of mtDNA in which humans typically differ, i.e., study a library of common variants called **genetic markers**. The use of genetic markers to determine intra-species differences is called **genotyping** and is a practice that predates the modern era (see **DETOUR: The First Genotyping Test**).

In practice, biologists use several types of genetic markers. Before the sequencing era, DNA was treated with a **restriction enzyme**, which cleaves DNA whenever it encounters some specific short DNA “keyword”, resulting in a collection of fragments. Individuals possessing a single-nucleotide variant within an occurrence of the keyword will not have their DNA cleaved by the restriction enzyme at this location, resulting in a longer DNA fragment called a **restriction fragment length polymorphism** (**RFLP**). RFLP analysis rose to prominence in the 1980s, when it led to highly accurate paternity testing and forensics analysis.

Another commonly used genetic marker is a **short tandem repeat**, a *short* DNA string of at most a few nucleotides that *repeats* dozens of times in *tandem*, i.e., consecutively. For example, the ends of chromosomes, called **telomeres**, consist of a short tandem repeat that occurs thousands of times (in vertebrates, the repeated string is `TTAGGG`). And Huntington’s disease is caused by a short tandem repeat on chromosome 4 in which the triplet `CAG` occurs too often; if an individual possesses over 40 copies of this triplet, then they will develop the disease.

Yet the genetic marker most frequently used in genotyping is the **single-nucleotide polymorphism** (**SNP**, pronounced “snip”), a variation at a single nucleotide position that occurs in at least 1% of the human population. Biologists have developed a variety of relatively cheap (i.e., under $100) lab methods to test an individual’s DNA sample against a library of common SNPs without needing to resort to sequencing.

Isolation produces mitochondrial haplotypes

Imagine a small population of individuals. If we assume a simple population model in which every woman in this population has exactly two children, then about a quarter of the mitochondrial genomes will be lost every generation: those corresponding to women who had two sons. As a result, the mitochondrial diversity of this population will steadily decay until everyone in the population possesses a mitochondrial genome derived from a single female ancestor.

This toy example helps us understand what we observe when analyzing mitochondrial genomes from differing human populations. Because many human populations have been isolated for significant time periods, mitochondrial genomes sampled from the same population tend to be very similar.

Another way of viewing the statement that individuals from the same population have similar mitochondrial genomes is that mitochondrial SNPs are often very positively correlated; that is, possessing one SNP makes you far more likely to possess a collection of many other SNPs. As a result, multiple SNPs can be grouped together and treated as a single genetic marker called a **haplotype**; all individuals possessing a haplotype form a **haplogroup**.

In the widest view, SNP correlation means that all humans can be divided into a relatively small family of mitochondrial haplogroups. However, we can only divide humans into these groups if we know how to form haplotypes from a collection of SNPs.

Selecting informative SNPs

To move toward a computational problem modeling haplotype selection, we will first (randomly) select a collection of SNPs T to serve as a benchmarking set. Given a (larger) collection of SNPs S, we would like to select an "informative" subset of SNPs S' from S that do the best job of "explaining" the variance that we observe in T for a population.

STOP and Think: How might we formulate a computational problem for determining the best choice of S' given T?

We will assume that all SNPs correspond to one of two possibilities, so that we can represent whether a given individual possesses a SNP as a binary value. Accordingly, given a collection of m individuals, we view a single SNP for these individuals as a binary vector s such that $s_i = 1$ if individual i possesses the SNP and $s_i = 0$ otherwise. Given a collection of n SNPs, we can consolidate the information stored in these vectors into an m x n **SNP matrix** A; that is, we set $A_{i,s} = 1$ if individual i possesses SNP s and $A_{i,s} = 0$ if individual i does not possess SNP s. Note that if we think of each SNP as a character, then an SNP matrix is just a character table from Chapter 7.

Our first question is how well a single SNP s explains another SNP t; that is, how much information does s provide regarding whether an individual will possess t? In particular, if individuals i and j have $t_i \neq t_j$, then in a simple sense s explains t for these two individuals if it is also the case that $s_i \neq s_j$. To be more precise, we will define a function $Diff(s, t)$ that averages this behavior over all pairs of individuals,

$$Diff(s,t) = \frac{\text{\# of pairs of individuals } (i,j) \text{such that } s_i \neq s_j \text{ and } t_i \neq t_j}{\text{\# of pairs of individuals } (i,j) \text{ such that } t_i \neq t_j}.$$

The larger the value of *Diff*(*s*,*t*), the more that any variance exhibited in *t* can be explained by *s*; if *Diff*(*s*,*t*) = 1, then we say that *s* is **fully informative** for *t*.

For example, say that we have five individuals along with the binary vectors *s* = (1, 1, 0, 0, 0) and *t* = (0, 0, 1, 1, 0) representing two SNPs. Of the ten possible pairs of individuals, six pairs have differing values for *t*, and four of these pairs have differing values for *s*. As a result,

$$\mathit{Diff}(s,\, t) = 4/6.$$

Exercise Break: Compute *Diff*(*s, t*) for the SNP vectors *s* = (1, 0, 0, 0, 0) and *t* = (0, 1, 0, 1, 1).

We can extend the definition of our function *Diff*() to quantify how well a collection of SNPs *S′* explains a single SNP *t*. To do so, we ask that for every pair of individuals that differ according to *t*, there is *some* SNP in *S′* that explains this difference in *t*,

$$\mathit{Diff}(S', t) = \frac{\#\text{ of pairs of individuals } (i,j) \text{such that } s_i \neq s_j \text{ for some } s \in S' \text{ and } t_i \neq t_j}{\#\text{ of pairs of individuals } (i,j) \text{ such that } t_i \neq t_j}\,.$$

Exercise Break: Compute *Diff*(*S′, t*) if *S′* consists of the vectors (1, 0, 0, 0, 0) and (1, 1, 1, 0, 0), and *t* = (0, 1, 0, 1, 1).

Finally, we return to our original problem: given SNP sets *S* and *T*, identify a collection of SNPs *S*′ that best explain the variance present in the entire collection *T*. We can quantify the informativeness of a subset *S*′ with respect to *T* by simply summing *Diff*(*S*′, *t*) over all SNPs *t* in *T*,

$$\mathit{Diff}(S', T) = \sum_{t \in T} \mathit{Diff}(S', t)\,.$$

Our goal, then, is to find a collection of SNPs maximizing this sum.

***k*-Most Informative SNP Problem:** *Identify a subset of most informative SNPs with respect to another collection of other SNPs.*

Input: SNP matrices corresponding to two collections of SNPs *S* and *T*, along with an integer *k*.

Output: A subset *S*′ of *S* containing *k* SNPs maximizing *Diff*(*S*, *T*) over all possible choices of *S* with *k* SNPs.

The *k*-Most Informative SNP Problem is *NP*-Hard, but it is still possible to define effective heuristics for it. For example, inspired by **RandomizedMotifSearch** from

Chapter 2, we can implement an algorithm called **RandomizedHaplotypeSearch** by starting with a random collection of *k* SNPs in *S*. At each step of the algorithm, we try every possible replacement of one SNP in our current collection with some SNP not in the collection and update *S′* to be the set that maximizes *Diff*(*S′*, *T*) among all those considered. We continue iterating until we can obtain a set such that no SNP replacement can increase *Diff*(*S′*, *T*). Like **RandomizedMotifSearch**, to avoid getting stuck in a local optimum, **RandomizedHaplotypeSearch** should be run many times over different collections of randomly chosen initial collections of SNPs.

```
RandomizedHaplotypeSearch(S, T, k)
    bestSNPs ← random collection of k SNPs from S
    while forever
        currentSNPs ← bestSNPs
        for each element s in currentSNPs
            for each element s’ in S - currentSNPs
                S’ ← currentSNPs with s replaced by s’
                if Diff(S’, T) < Diff(bestSNPs, T)
                    bestSNPs ← S’
        if bestSNPs = currentSNPs
            return bestSNPs
```

Code Challenge: Implement **RandomizedHaplotypeSearch**.

Tracing the Footsteps of Ancient Human Populations

The infinite sites model and perfect phylogeny

Researchers have found that a small handful of haplotypes can prove very informative for human mitochondrial studies. Furthermore, the resulting haplogroups are *hierarchical*; that is, they derive from an underlying evolutionary tree as populations branched into subpopulations. The question, then, is how to reconstruct this evolutionary tree to infer the evolution of human populations from an *m* x *n* SNP matrix *A* corresponding to *n* binary haplotypes over a collection of *m* individuals.

We will make a simplifying assumption that because we have comparatively few haplotypes compared to the length of the mitochondrial genome, the probability of the same site mutating more than once given a reasonably small collection of mitochondrial genomes is close to zero. This assumption is a variant of the **infinite sites model**, which was introduced by Motoo Kimura in 1969. It states that because a mutation at a given location is relatively rare during species evolution, we can assume that a genome is a continuous strand of infinite mutation sites rather than a finite sequence of nucleotides, so that the probability of mutations occurring more than once at a given site is approximately zero. (Note that we have already made this assumption when inferring the age of mitochondrial Eve from the mutation rate combined with the number of mutations between Eve and present-day humans.)

The infinite sites assumption implies that there has been no **convergence** of characters during the evolutionary process, meaning that each character arose exactly once. In other words, given an m x n SNP matrix A, we are seeking a rooted tree T in which each node is labeled by a vector of length n, called the node's **SNP vector**, satisfying the following properties:

1. The SNP vector at the root consists of only zeros (as it corresponds to an ancestor possessing none of the characters corresponding to the SNPs).
2. Each row of A is the SNP vector of exactly one leaf.
3. For any column j of A, there is a single node v such that every node in T_v, the subtree of T rooted at v, contains a 1 at the j-th position, and every other node in T contains a 0 at the j-th position.

This framework is collectively called the **perfect phylogeny** assumption. If a tree satisfying the perfect phylogeny assumption exists for a given matrix A, then we say that the tree **perfectly fits** A, and we call A **phylogenetic**. We can now state the following computational problem.

Perfect Phylogeny Problem: *Reconstruct a perfect phylogeny from a binary character matrix.*
Input: A binary character matrix.
Output: A tree that fits the matrix perfectly, if such a tree exists.

STOP and Think: Is perfect phylogeny a reasonable assumption for the character of wings during insect evolution?

Identifying compatible SNPs

Examine the 6 x 5 SNP matrix shown in Figure A (top). In Figure A (bottom), we present a rooted tree fitting this matrix. If we know how to assign each row of the matrix to a leaf in this tree, then we can easily infer the SNP vector at internal nodes from the leaves upwards --- an internal node possesses a 1 in the i-th element of its SNP vector if all its children do, and it possesses a 0 otherwise. This assignment ensures point 3 in the perfect phylogeny assumptions is satisfied; furthermore, any SNP can be assigned to some node v in T such that v and all its descendants possess the SNP (but no other nodes in the tree do).

STOP and Think: How does this inference of internal states in Figure A differ from how the **SmallParsimony** algorithm presented in Chapter 7 assigned ancestral states?

a	1	1	1	0	1
b	0	1	1	0	1
c	0	0	0	0	0
d	0	0	1	0	0
e	0	0	1	1	1
f	0	0	1	1	1

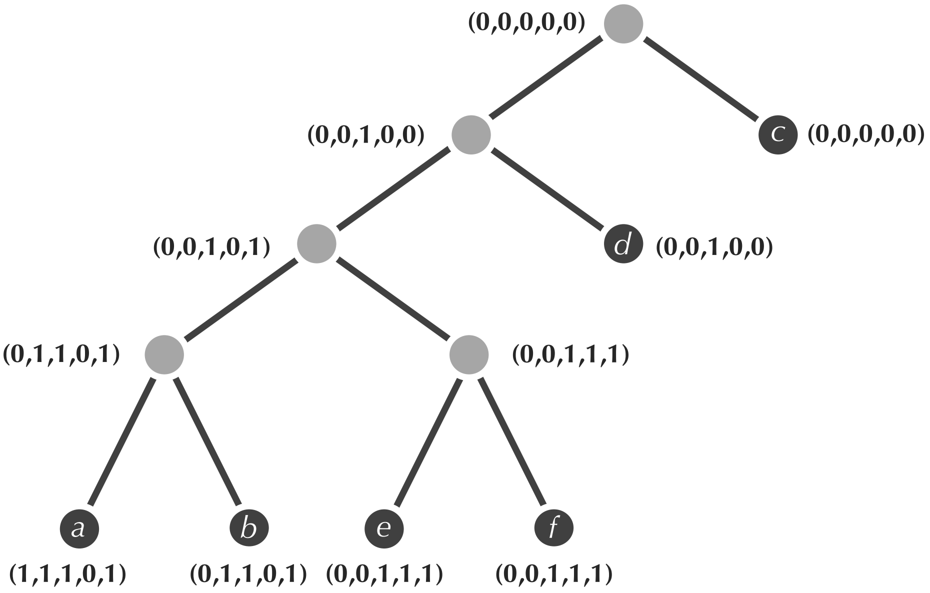


Figure A: (Top) A 6 x 5 SNP matrix. (Right) The phylogenetic tree that perfectly fits this matrix. We have labeled every leaf with a binary vector corresponding to the row of the matrix representing its character states; we can then infer the SNP vectors for internal nodes of the tree by working our way from the leaves upward, assigning a 1 at position *i* of an internal node's SNP vector if its children all have a 1 at this position.

Say that we examine two different characters s and t that are associated with nodes v and w. Then assumption 3 also guarantees that the subtrees T_v and T_w are either disjoint (i.e., contain none of the same nodes), or one is a subtree of the other.

However, we are getting ahead of ourselves, since we do not know the correct assignment of rows of A to leaves of a tree in advance (or, for that matter, whether A is phylogenetic to begin with). Our goal is therefore to find a direct test for whether A is phylogenetic; like the four-point condition from Chapter 7, this test should rely solely on properties of A.

Fortunately, the information about the two characters *s* and *t* is represented within *A* as two columns, corresponding to some indices *i* and *j*. Using the notation O_i to denote the collection of rows possessing a 1 in their *i*-th element, we conclude that one of three possibilities must be true: $O_i \subseteq O_j$, $O_j \subseteq O_i$, or O_i and O_j are disjoint (the first two cases include the possibility that $O_i = O_j$). If columns *i* and *j* satisfy this condition, then we call them **compatible**.

Exercise Break: Verify that all pairs of columns in the matrix in Figure A (top) are compatible.

Solving the Perfect Phylogeny Problem

We have just demonstrated that if a matrix is phylogenetic, then its columns are all pairwise compatible. It turns out that this simple condition is enough to *guarantee* that the matrix is phylogenetic, which we will prove next.

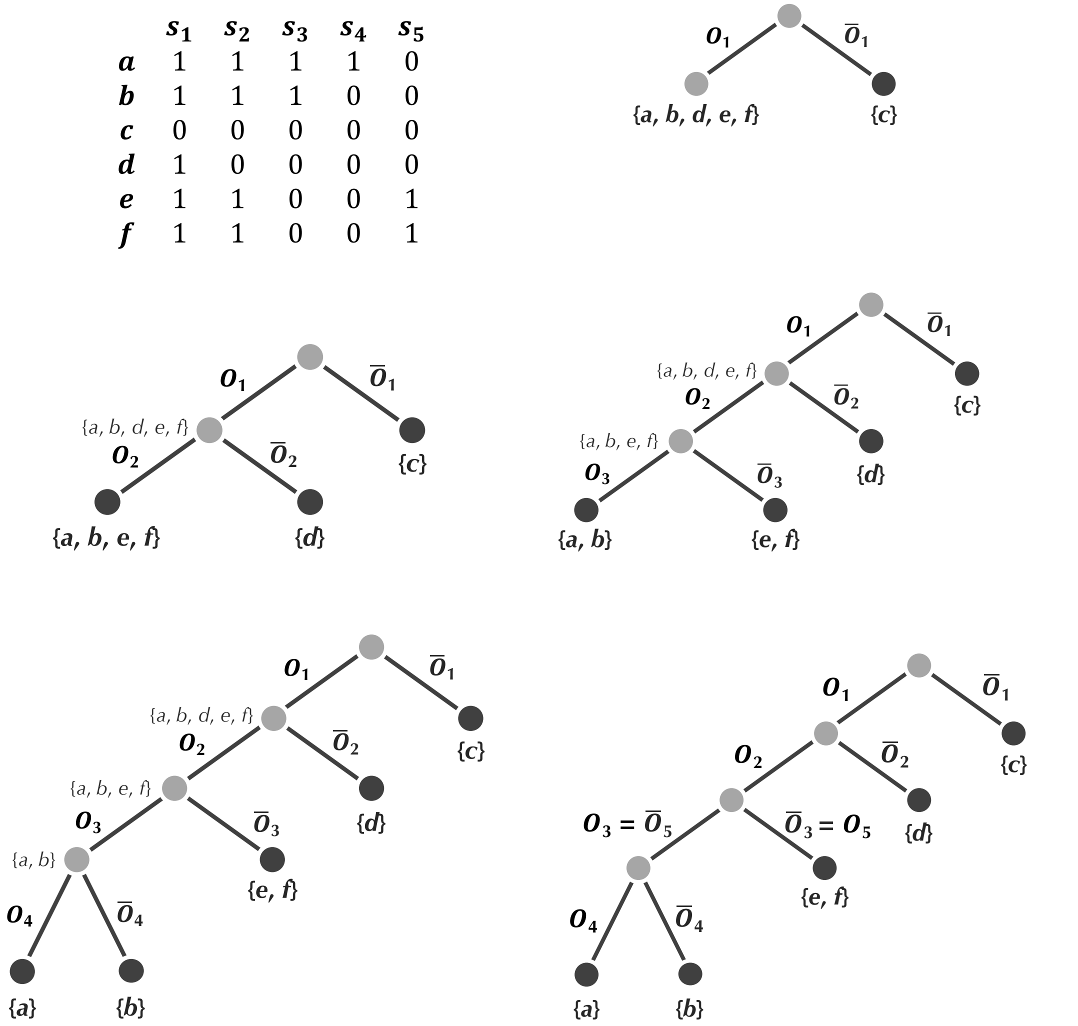
Theorem: A SNP matrix *A* is phylogenetic if and only if every pair of columns in *A* are compatible.

To prove this theorem, assume that *A* is an arbitrary SNP matrix with pairwise compatible columns; we will also assume without loss of any generality that no two columns of *A* are identical. We will provide a *constructive proof* that will lead us to an algorithm for constructing a tree fitting *A*. Our idea for this algorithm is to start at a tree consisting of a single root, identifying the characters that are possessed by the most individuals, and iteratively moving downward to the leaves.

If we treat the columns of *A* as binary vectors, then sorting the columns into descending lexicographic order from left to right will guarantee that if $j > i$, then either $O_j \subseteq O_i$ or O_i and O_j are disjoint (Figure B (top left)). We can therefore form two children of the root and assign all members of O_1 to one child and all remaining members (which we denote as $\overline{O_1}$) to the other child (Figure B (top right)).

We then iterate this process, moving left to right within the columns of *A*. When considering the *i*-th column, we move downward in the tree, at each step choosing a child *v* if O_i is a subset of the individuals contained in T_v. (This process will never stop at an internal node because the columns were previously sorted.) When we reach a leaf *v*, there are two possibilities: if O_i is equal to the individuals assigned to *v*, then we stop; if there are individuals in $\overline{O_i}$ that are present at *v*, then we create two children of *v* corresponding to O_i and $\overline{O_i}$ and subdivide the individuals accordingly. This process is illustrated in the middle and bottom panels of Figure B.

After considering all columns of *A*, if there is any leaf node containing more than one individual, then we form as many children of this leaf as needed so that each leaf contains a single individual; for example, we subdivide the leaf {*e*, *f*} in Figure B (bottom right) into two leaf nodes to yield the original tree shown in Figure A (bottom). The resulting algorithm is called **PerfectPhylogeny**.



	s_1	s_2	s_3	s_4	s_5
a	1	1	1	1	0
b	1	1	1	0	0
c	0	0	0	0	0
d	1	0	0	0	0
e	1	1	0	0	1
f	1	1	0	0	1

Figure B: (Top left) The matrix *A* from Figure A (top) with columns sorted into decreasing lexicographic order. (Top right) Dividing all individuals according to whether they have a 1 or a 0 for the first column in *A*. (Middle, bottom panels) Subdividing the tree at the appropriate node for each subsequent column in *A*, moving left to right. Note that when we insert the fifth character (bottom right), it coincides perfectly with the set of individuals not containing the third character, so there is no need to draw a new node. In the final panel, subdividing the node containing *e* and *f* into two leaves will result in the rooted tree from Figure A (bottom) as desired.

Code Challenge: Implement **PerfectPhylogeny** to solve the Perfect Phylogeny Problem.

Practical considerations in perfect phylogeny

Once we have constructed an evolutionary tree from a SNP matrix, we would like to augment this tree with the data from a newly genotyped individual, which we can state as the following problem.

Augmented Perfect Phylogeny Problem: *Add a new individual to an existing perfect phylogeny.*

Input: A rooted tree T along with an SNP matrix A that T perfectly fits and a SNP vector c for an additional individual.

Output: A rooted tree T perfectly fitting the matrix A augmented by an additional row corresponding to c (if the augmented matrix is still phylogenetic).

We could solve this problem by simply reconstructing the entire evolutionary tree from the augmented matrix. However, this approach is overkill, since we can use the structure of the tree that we already know fitting the original matrix. Instead, we can determine where the individual belongs in this tree by starting at the root and iteratively walking downward, determining the child to which the individual belongs at each step.

Code Challenge: Solve the Augmented Perfect Phylogeny Problem.

In the preceding discussion, we have assumed that the ancestral state of a given character is 0 for each character. But in practice, when we observe the two states of a haplotype, we won't know in advance which of the two states is ancestral (i.e., which individuals should receive a 0 for this haplotype). Fortunately, if there exists an assignment of each character to 0 and 1 yielding a phylogenetic matrix, then this assignment is easy to find (see **DETOUR: Perfect Phylogeny for Undirected Characters**).

A more pressing concern is that the perfect phylogeny assumption is an extremely rigid one. As we increase the number of individuals, we will need more haplotypes to differentiate them, and it becomes increasingly likely that random mutations will have caused convergence of characters with respect to some haplotypes.

STOP and Think: How would you address the fact that **PerfectPhylogeny** can only reconstruct a tree from ideal data?

Constructing an evolutionary tree of human mitochondrial haplotypes

From the growing wealth of genetic data, researchers have constructed a phylogeny using the most informative human mitochondrial haplotypes (Figure C (top)). And by gathering information about the heritage of individuals genotyped to form this evolutionary tree, we can simply connect the dots to *visualize* the rough migration patterns of human populations as they spread out from the root in east Africa around the world (Figure C (bottom)).

For example, mitochondrial haplotype A is often found in individuals of east Asian and native American descent, providing genetic evidence for an eastward migration across the land bridge crossing what is now the Bering Strait approximately 15,000 years ago. Furthermore, haplotype L is located at the root of the human mitochondrial phylogeny and corresponds to mitochondrial Eve. All non-Africans belong to the L3 haplogroup, in addition to some Africans, giving us yet another argument in favor of the Out of Africa hypothesis, a once-volatile proposal that has quickly become an accepted scientific fact in the modern era of genetic data.

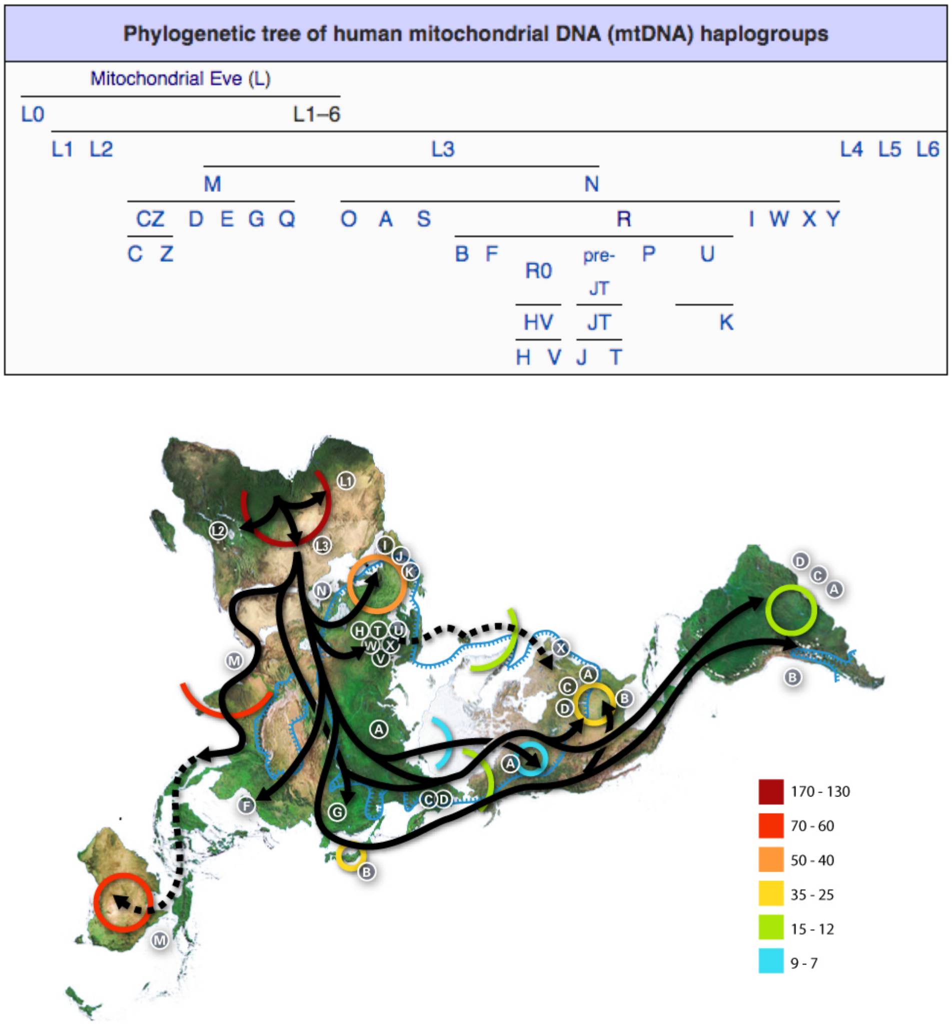


Figure C: (Top) A phylogenetic tree of human mtDNA haplogroups. Courtesy: https://en.wikipedia.org/wiki/Human_mitochondrial_DNA_haplogroup. (Bottom) A map of human haplotypes. Courtesy: Alexandre Van de Sande.

Bibliography Notes

The Out of Africa hypothesis was proposed by Cann, Stoneking, and Wilson, 1981. Neanderthal mtDNA was first isolated in Krings et al., 1997; a draft sequence of the Neanderthal genome was published by Green et al., 2010; a more reliable Neanderthal genome with significantly higher coverage was published by Prüfer et al., 2014. The *k*-Most Informative SNP Problem was adapted from Bafna et al., 2003. The first theorem for proving compatibility of a matrix appeared in Estabrook, Johnson, and McMorris, 1975. The proof of the theorem for perfect phylogeny of undirected characters was taken from McMorris, 1977.

Detours

Aging Y Chromosomal Adam

The Y chromosome offers an analogue to mtDNA in that it serves as a "mini-genome" that can only be passed via the paternal line. Accordingly, we can construct an evolutionary tree from Y chromosome data gathered from human males; such a tree is often called a **surname phylogeny** since many civilizations pass surnames down paternally. The most recent ancestor of all human males is called, predictably, **Y chromosomal Adam**.

We can use evolutionary trees to date the appearance of Y chromosomal Adam similarly to how we constructed an evolutionary tree for mitochondrial genomes. However, research repeatedly indicated that mitochondrial Eve is substantially older than Y chromosomal Adam, who would appear to be only about 60,000 years old!

The fact that Adam is much younger than Eve is not a paradox; all other male lineages may have simply died out before reaching the present. Researchers proposed a host of other hypotheses for why mitochondrial Eve might be older than Y chromosomal Adam, but none gained widespread traction.

Then, in 2012, something amazing happened. As part of National Geographic's Genographic Project, researchers found that an African American man whose Y chromosome was startlingly distinct from the many thousands of Y chromosomes tested up to that point. Subsequent genetic testing found 11 members of the Mbo people of Cameroon who possessed a very similar Y chromosome.

Exercise Break: Construct an updated surname phylogeny after adding Mbo Y chromosome data. Still assuming that the mutation rate on the Y chromosome is XXX, how old is Y chromosomal Adam?

When researchers redrew the human male phylogeny to account for the Mbo outliers, Y chromosomal Adam's age was moved back to approximately 200,000 years ago following some debate.

Gene Flow Between Neanderthals and Modern Humans

It seems highly likely that Neanderthals and Cro-Magnons interacted, and difficult to imagine that there were no sexual encounters during the thousands of years that the two groups inhabited the same parts of Europe. However, it is also possible that Neanderthals had become genetically incompatible with Cro-Magnons by the time the latter arrived in Europe, making gene flow impossible.

To test the hypothesis that there was gene flow between the populations, we would ideally have access to a complete Neanderthal genome rather than just mtDNA. Yet when Neanderthal mtDNA was first extracted in 1997, many researchers did not believe that we would ever be able to reliably sequence a full Neanderthal genome because of the effects of degradation and contamination. Nevertheless, in 2010, the advent of more powerful laboratory sequencing methods allowed researchers to publish a low-coverage draft of the Neanderthal genome based on a read dataset containing 4 billion nucleotides. Three years later, whole genome sequencing with much higher coverage was completed using DNA extracted from a Neanderthal toe bone.

STOP and Think: Once we have a Neanderthal genome, how can we test the hypothesis of gene flow between Neanderthals and modern humans?

In the absence of gene flow, all human populations should be approximately equally diverged from Neanderthals, since all modern humans share an ancestor who lived more recently than the first appearance of Neanderthals in Europe. In the 2010 publication of the draft Neanderthal genome, researchers observed a greater discrepancy when comparing African genomes against the Neanderthal genome than when comparing European genomes against the Neanderthal genome. As a result, they concluded that gene flow had been present between Neanderthals and Cro-Magnons.

Nevertheless, the regions of the genome exhibiting significant similarity between Eurasians and Neanderthals were relatively sparse (early estimates indicated just 1-4% of the genome). Recent results showing that many modern Eurasians share parts of the genes related to pigment and keratin might indicate that the Neanderthals provided modern humans with adaptations to living in a colder environment that persisted because of natural selection. Researchers have also found variants lurking within genes powering the immune system, which may imply that Neanderthals also equipped Cro-Magnon settlers with genes helping them resist infections in their new landscape. Both arguments are plausible, but some researchers are still not convinced that significant gene flow between the populations occurred.

The first genotyping test

Blood transfusion has become such a common medical practice that you may be surprised to learn that it has only been widely applied for the last hundred years. 17th century techniques attempted to supplement human blood with that of animals, causing disastrous consequences when the recipient's immune system reacted violently, and scaring away any serious physician who would consider the practice. By the 19th Century, person-to-person transfusions had gained popularity, but only as a last resort; these transfusions were sometimes successful but often produced reactions like those observed with animal blood. Scientists wondered: why would your body sometimes reject the blood from another human, and sometimes accept it?

The breakthrough occurred in 1900, when Karl Landsteiner noticed that after mixing blood from two different individuals, the red blood cells clumped together. He deduced that this clumping process (called "agglutination") must be the physical manifestation of the body's immune response to foreign blood. Moreover, Landsteiner noticed that sometimes the clumping happened, and sometimes it didn't; in fact, he could cluster subjects into disjoint groups for which subjects from the same group were compatible, and subjects from different groups were incompatible.

Landsteiner had discovered the **ABO blood group**, which we now know is encoded by a single locus on chromosome 9 consisting of three alleles: A, B, and O (which differs from A by a single nucleotide deletion). If an individual possesses at least one copy of the A allele, then they produce a specific enzyme that bonds to the surface of red blood cells. An individual who has no A alleles produces an antibody in the blood serum that attacks the enzyme, causing agglutination in incompatible donors. The B allele represents a different surface enzyme and antibody, but the result is the same. The O allele is simply a placeholder for "neither A nor B." Thus, OO individuals can only receive red blood cells from other OO individuals because OO blood serum contains antibodies to attack both A and B surface enzymes; yet these individuals can donate blood cells to anyone. AB individuals are the opposite, since they can only donate blood cells to other AB individuals, but can receive blood from any donor. The donor relationship between different blood groups is summarized in Figure E.

STOP and Think: How does Figure E differ when considering the donation/acceptance of blood *serum*? Recall that if an individual does not possess an A (respectively, B) allele, then they will produce an antibody in the blood serum attacking the A enzyme.

		Donor			
		O	A	B	AB
Recipient	O	yes	no	no	no
	A	yes	yes	no	no
	B	yes	yes	yes	no
	AB	yes	yes	yes	yes
		45%	34%	16%	5%

Figure E: The standard ABO blood group table (with percentages of humans falling into each category), indicating whether a given individual from one group can donate or receive from each other group. "O" indicates an individual with two "O" alleles; "A" indicates an individual who is either "OA" or "AA"; "B" indicates an individual who is "OB" or "BB"; and "AB" indicates an individual who possesses an A allele and a B allele.

A quick test for ABO blood group was developed after about a decade of research following Landsteiner, who would later discover the Rh blood group as well allowing transfusions to be carried out in hospitals around the world. Yet it is important to note that in addition to saving lives, the ABO blood test represents an early example of genotyping, one that predates any direct knowledge of the molecular basis for inheritance.

Perfect phylogeny for undirected characters

In the main text, we assumed that for a binary character, 0 always corresponded to an ancestral state and 1 always corresponded to the state acquired by the character as a mutation. In practice, we may have **undirected characters**, in which case we do not know which state is ancestral. In this case, a binary matrix will be phylogenetic if it is possible to assign a 0 and 1 to each state of each character in such a way as to obtain a matrix whose columns are pairwise compatible. Unfortunately, there are 2^n different assignments of 0 and 1 to the n characters making up a binary matrix. However, we will be able to assign a 0 to a "majority state", i.e., the one held by the majority of individuals (with ties broken arbitrarily).

Theorem: If a compatible assignment of 0 and 1 to the states of two characters exists, then the assignment of the majority state to the two characters is compatible.

To prove this statement, assume that two binary characters s and t have the respective states $\{a, b\}$ and $\{c, d\}$. Define $P(s, t)$ as the collection of different pairs of these states possessed by individuals; that is, $P(s, t)$ consists of at least one of (a, c), (a, d), (b, c), and (b, d). We subdivide our proof based on how many elements $P(s, t)$ contains.

If $P(s, t)$ contains only one pair, then s and t are both uniform and we can assign them both all zeroes. If $P(s, t)$ contains two pairs, say (a, c) and (a, d), then one of the characters is uniform (we set $a = 0$ in this case). We can then set either $c = 0$ or $d = 0$ and still retain compatible characters, since O_s will be a subset of O_t regardless.

On the other hand, if $P(s, t)$ contains all four pairs, then any assignment of 0 and 1 to the values of s and t will mean that each of O_s and O_t will contain elements not found in the other, and that these characters cannot be compatible to begin with.

The only tricky case, then, is if $P(s, t)$ contains exactly three pairs. We will show that s and t are always compatible. Without loss of generality, assume that the pairs are (a, c), (a, d), and (b, c). First, if there are at least as many occurrences of a than b, then we can assign $a = 0$ and $b = 1$. If there are at least as many occurrences of c as d, then we can assign $c = 0$ and $d = 1$, meaning $P(s, t)$ becomes the three pairs (0, 0), (0, 1), and (1, 0), and so O_s and O_t are disjoint. If there are at least as many occurrences of d as c, then we can assign $d = 0$ and $c = 1$, meaning $P(s, t)$ becomes the three pairs (0, 1), (0, 0), and (1, 1); thus, O_s is contained within O_t.

If there are at least as many occurrences of b than a, then we can assign $a = 1$ and $b = 0$. Because (b, d) does not occur in $P(s, t)$, we automatically conclude that there are more occurrences of c than d, and assign $c = 0$ and $d = 1$. This makes $P(s, t)$ becomes the three pairs (1, 0), (1, 1), and (0, 0), and so O_t is contained within O_s.