mtDNA

Charles Davi

ACGT

“Your genetics loads the gun.” - Mehmet Oz.

What is DNA?

DNA is an abbreviation for Deoxyribonucleic Acid. DNA is a very large molecule that has two main components:

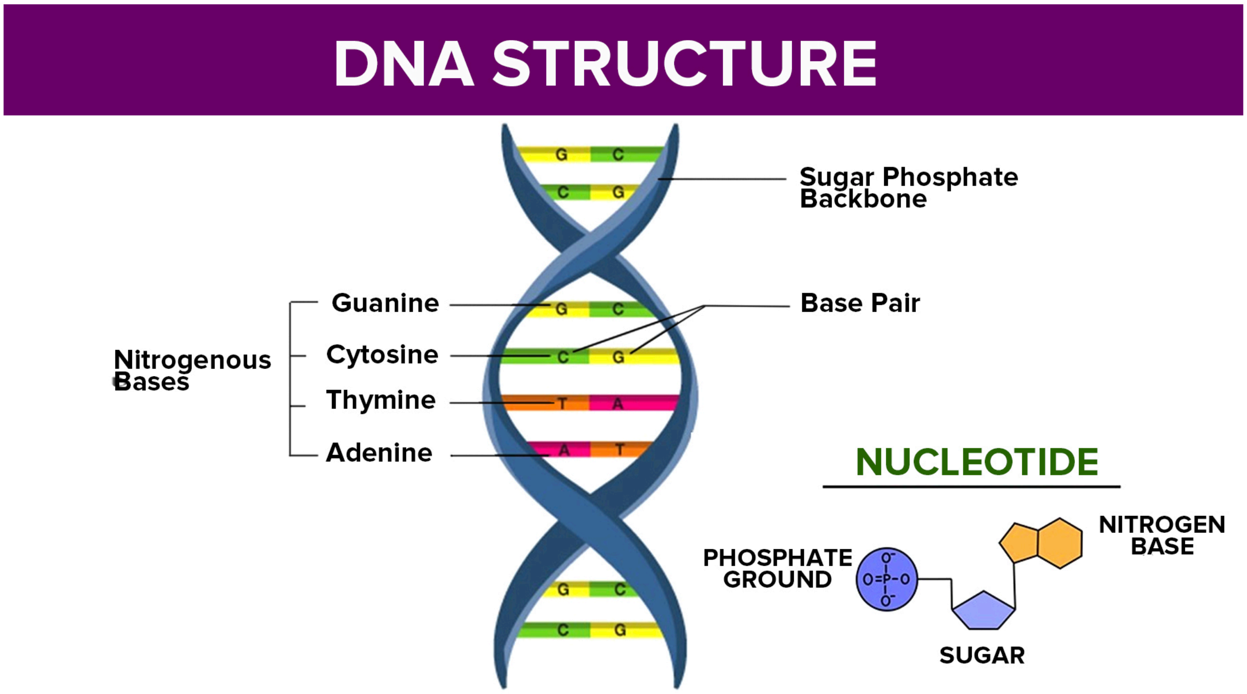


Figure 1: A diagram showing the two main components of DNA, the base pairs that encode information, and the sugar-phosphate backbone to which the base pairs attach.

1. **Base Pairs**: These are the familiar labels ACGT, which stand for the molecules Adenine, Cytosine, Guanine, and Thymine.

2. **Deoxyribose-Phosphate Backbone**: The backbone of a DNA molecule is made of alternating molecules of Phosphate and Deoxyribose (sugar).

DNA is primarily located in the nucleus of a cell, and since DNA is also acidic, it is referred to as Deoxyribonucleic Acid.

Some basic math yields an intuition for the simply gigantic scales at work in the human genome:

On average, a single base molecule contains 14.75 atoms. Therefore, a single "row" of DNA base pairs consists of the following average number of atoms:

2 bases (2 x 14.75 atoms) + 2 sugars (2 x 19 atoms) + 2 phosphates (2 x 5 atoms) = 77.5 atoms.

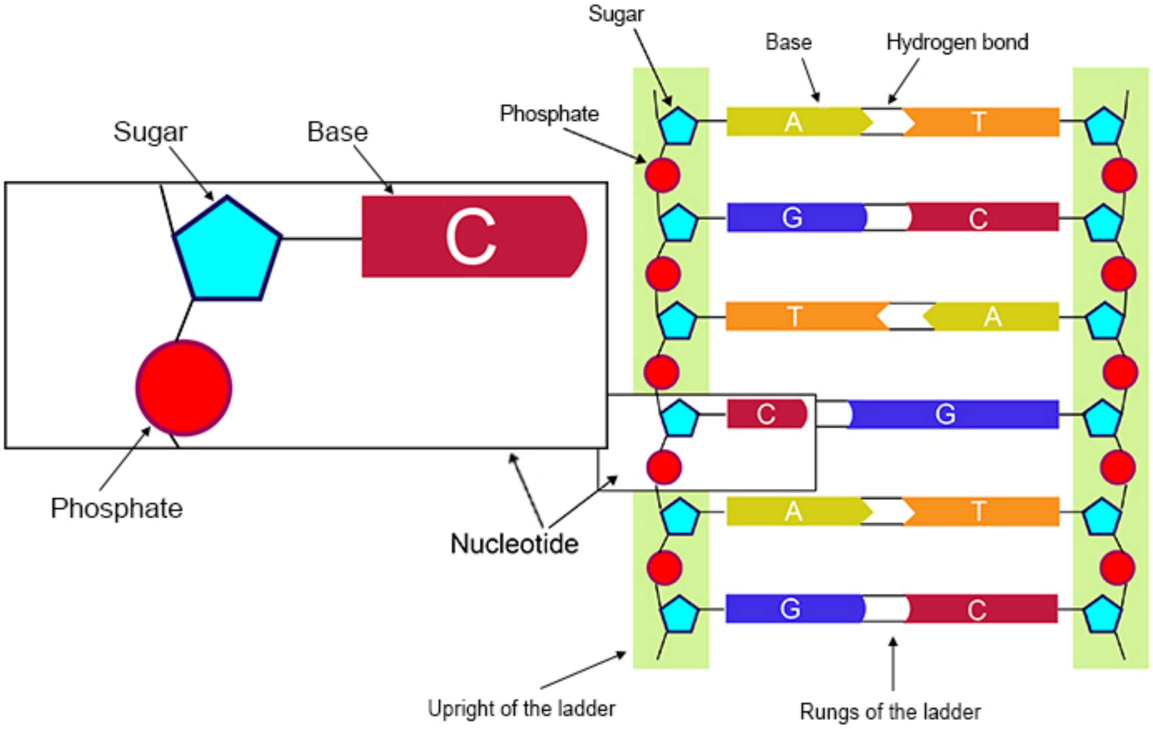


Figure 2: A diagram showing the sugar-phosphate bonds in the backbone of DNA, and the bases bonded to the backbone.

The total number of base pairs in the human genome is approximately 3 billion. This implies that the total number of atoms in the human genome is approximately 232.5 billion. For context, the largest manmade molecule "PG5" consists of approximately 20 million atoms, and has no dynamic function that I'm aware of. In contrast, the human genome is approximately 10,000 times larger, and encodes the most complicated known structures in the Universe, including the human brain.

The obvious conclusion is that DNA is a profoundly complex molecule, with astonishing properties.

Molecular Machines

How does DNA actually function: After studying mtDNA for the last few years, I think the honest answer is that we don't know. That said, there are plenty of things that we do know about how DNA works, and a lot of that knowledge is very impressive, in particular, our growing understanding of Molecular Machines, which are literally everywhere in the human body.

Figure 3: "Becoming", by Jan van Ijken, showing a salamander growing from a single cell.

My assessment is that no one can explain what's happening in this video in mechanical detail. More precisely, I don't think anyone can write a program that is in 1-to-1 correspondence with the processes observed in this video. But we can nonetheless get an intuition by taking an even closer look at cellular function, in particular, Molecular Machines. There are now computer simulations that mimic the behavior of at least some of the Molecular Machines we know to exist in living systems, which is, given their complexity, an unimaginable accomplishment.

The ATP Synthase: The ATP Synthase is a Molecular Machine responsible for energy production in essentially all Eukaryotic species and bacteria.

Eukaryotic species are those that have cells with a well-defined nucleus. Note that even though Bacteria are an example of a Non-Eukaryotic organism, since they don't have a nucleus at all, Bacteria nonetheless contain ATP Synthases, and in fact, human cells contain a bacteria-like organelle called the Mitochondrion, which is responsible for energy production within Eukaryotic cells and Bacteria.

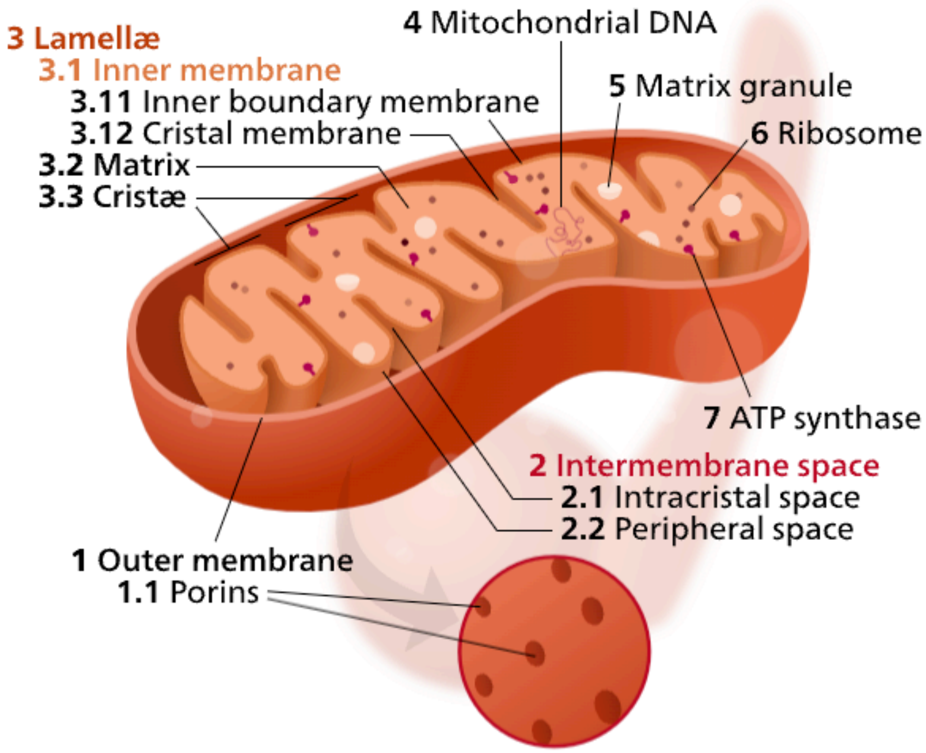


Figure 4: A diagram showing the components of a Mitochondrion, including the location of various ATP Synthases.

Note that there are numerous ATP Synthases throughout the Mitochondrion, providing a sense of scale, given that a Mitochondrion is itself $O(10^{-6})$ meters, which is approximately one-hundredth the width of a human hair. Specifically, there are around 1,000 to 5,000 ATP Synthases per Mitochondrion, around 1,000 Mitochondria per cell, and around 30 trillion cells in the human body. Therefore, the number of ATP Synthases in the human body is $O(10^{19})$. With that incredibly small scale, and incredibly large number, in mind, consider the following video, which is a computer simulation of the ATP Synthase, again responsible for energy production within Eukaryotic cells and Bacteria.

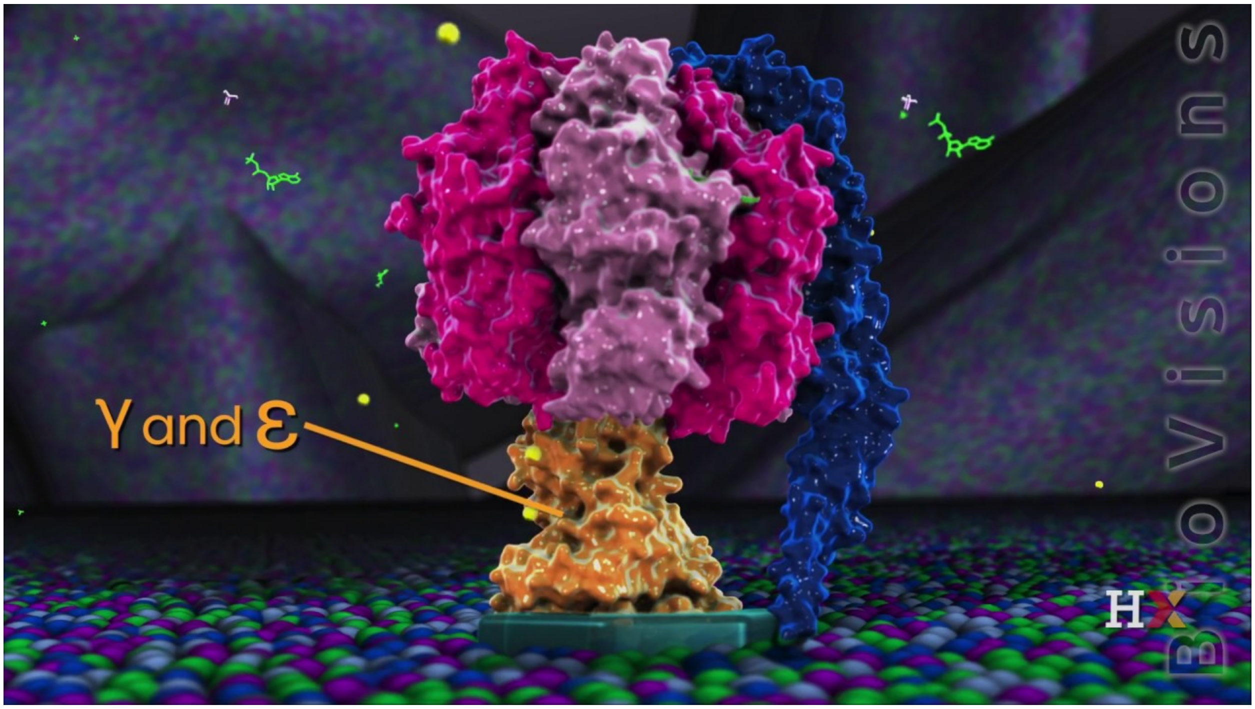


Figure 5: A video demonstrating the mechanics of the ATP Synthase, an example of a Molecular Machine.

To provide an even grittier appreciation of both the scale and complexity of the ATP Synthase, consider the electron microscope images below. Note that they are plainly numerous, and in fact have the mushroom-like shape depicted in the video above. Finally, note that the ATP Synthase is only one example of the many Molecular Machines that exist in the human body, and living systems generally, creating an imposingly complex portrait of life.

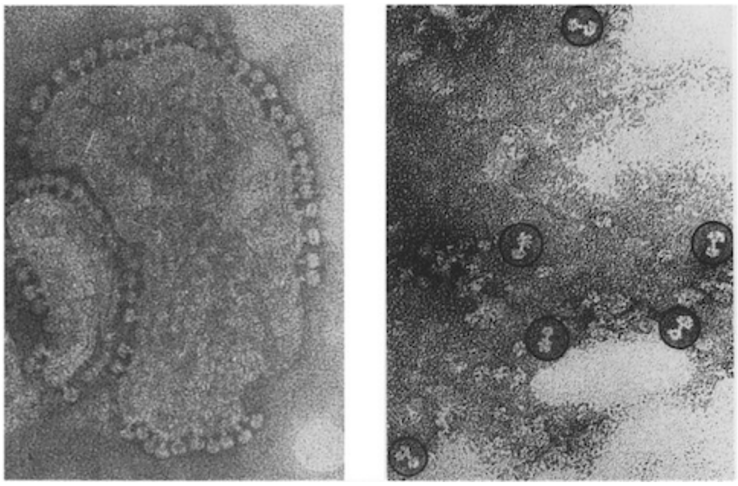


Figure 6: Images of the ATP Synthase produced using an electron microscope.

Though not relevant to the work discussed in this presentation, it's nonetheless worth considering the example of Molecular Motor Proteins, which literally have legs, and walk inside cells, along microtubules. The motion is disturbingly similar to modern robots, in particular, Boston Dynamic's "Big Dog", though this is of course happening inside your body, and again, at an unbelievably small scale. Further, the locomotion is the result of the exploitation of the laws of Physics that govern the interactions of individual molecules. Contrast this, to a combustion or battery powered engine. Living systems are, in this view, as a matter of engineering, miraculously complicated, whether or not design was involved.

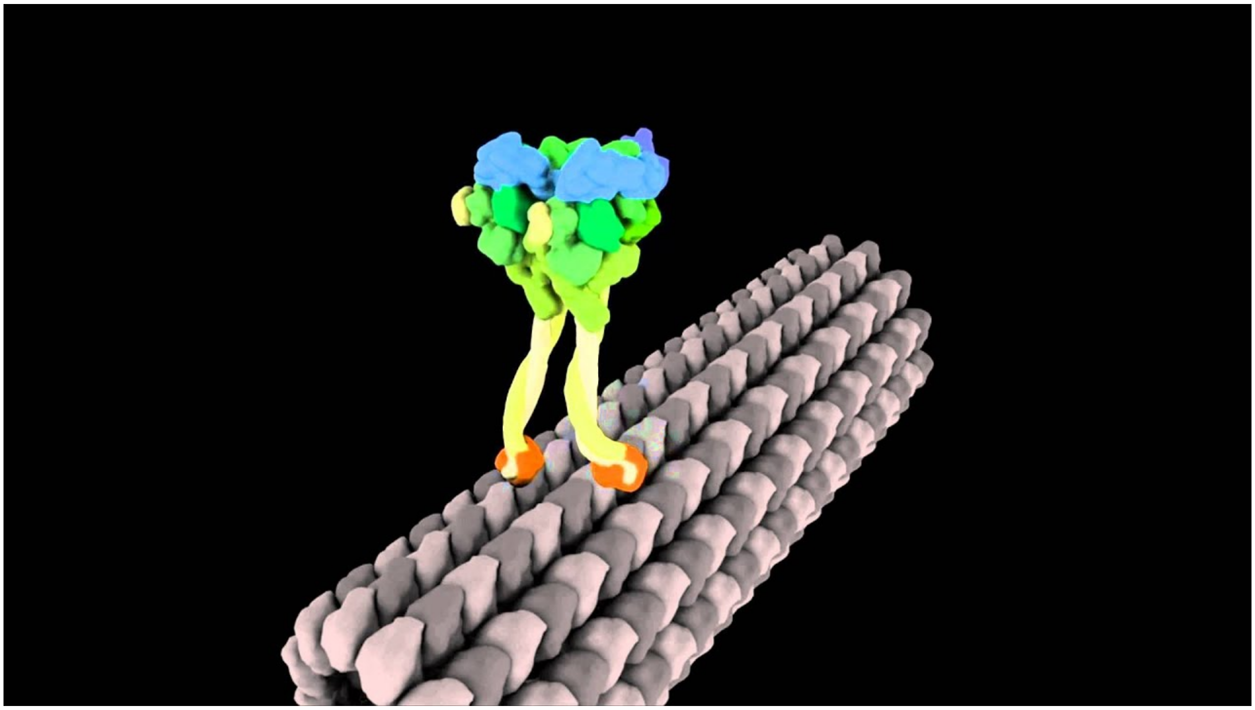


Figure 7: A video demonstrating the motions of Molecular Motor Proteins.

What is mtDNA?

mtDNA: Mitochondrial DNA, known as mtDNA, is a genome found inside the Mitochondrion. As a result, unlike most DNA, mtDNA is not inside the nucleus. Further, mtDNA is inherited directly from the mother to its offspring, generally with no mutations at all. That said, mutations do occur, but they are extremely rare, and as a consequence, we can use mtDNA to uncover deep maternal histories. Though DNA was discovered in 1869 by Friedrich Miescher, and its structure was revealed in 1953 by James Watson and Francis Crick, mtDNA was not discovered until 1963, by Margit M. K. Nass and Sylvan Nass.

The Structure of mtDNA: Just like bacterial DNA, mtDNA is circular, and forms a loop. However, the loop has clearly defined regions, which allow us to impose an order on the genome. That is, even though the genome is physically structured as a loop, we can nonetheless define objective starting points on the genome. In fact, basically all humans have the same sequence of 15 bases, allowing us to define an objective Global Alignment.

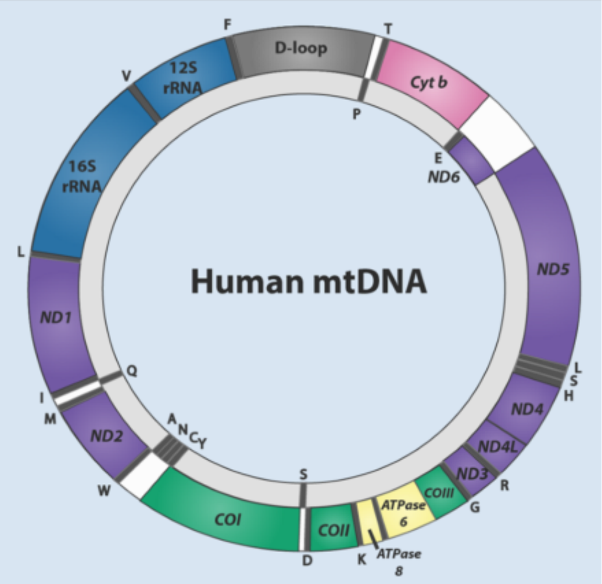


Figure 8: A diagram of human mtDNA, broken down into distinct segments.

Representing Genomes: Mathematically, we can represent a genome as a vector of labels, where each element of the vector is taken from the set

$\{A, C, G, T\}$. For example, the vector (A, C, G, T) in this view represents a genome, even though it probably doesn't code for anything useful.

However, unlike a computer program, which is compiled, a genome is instead physically manipulated using Molecular Machines to do one of two things: (1) replicate or copy the genome and (2) produce proteins encoded for in the genome. As a result, in addition to having function, like a computer program, DNA must also be physically stable as a molecule, and in addition, encode for physically stable proteins that have meaningful physical functions, and in some cases, these proteins spontaneously assemble into structures like microtubules, which in turn spontaneously assemble into organelles like Mitochondria, and into Molecular Machines like the ATP Synthase. In this sense, DNA is not analogous to a computer program, and is instead an incomprehensibly more complex system for the conveyance of information that ultimately has physical function.

Nonetheless, for our purposes, we can reduce a genome to a vector of labels, and use that representation to perform meaningful analysis on computers. But in order to do so, we must first impose an order on the genome, known as an alignment, which are either global or local.

Global Alignment: Assume we are presented with two genomes, $x = (A, C, G, T, A, T)$ and $y = (A, C, G, A, T)$. If these two genomes are complete, i.e., we're not missing any additional bases, then we can conclude that they are not the same length. However, it is not clear how we should read the genomes, since they could e.g., be read backwards or forwards. As such, we will assume that genomes expressed as vectors are read from left to right. However, we still need a method for comparing the two genomes.

An **alignment** assigns an ordinal to each entry in the vector, mapping bases in a given genome to bases in another genome, thereby allowing us to compare two genomes. So in this case, we could e.g., define the alignment $A_x = (1,2,3,4,5,\varnothing)$. Note that we have inserted a single special character in the alignment, to account for the fact that the two genomes are of unequal length, which we will interpret as assigning base six of genome x to nothing.

Because $A_x(1) = 1$, we would compare entry 1 of x to entry 1 of y, to find that they are both Adenine, and we say that the two genomes have matching bases at those indexes. If we count all matching bases using this alignment, we have a total **match count** of 3 bases. We write $|x \cdot y| = 3$ to indicate a match count of 3, with the tacit understanding that the applicable alignment is known given the context, for simplicity. This is called a Global Alignment because there are no gaps, and the mapping is sequential.

Local Alignment: Looking again at genomes $x = (A, C, G, T, A, T)$ and $y = (A, C, G, A, T)$, we note that they are in fact identical, if we remove the Thymine base at index 4 of genome x. We can then represent this using a Local Alignment, which can be expressed as $A_x = (1,2,3,\varnothing,4,5)$. That is, we ignore index 4 of genome x, and shift the remaining segment (A, T) to the left by one index.

x = (A,C,G,~~T~~,A,T)
y = (A,C,G,A,T)

This is referred to as a Local Alignment, because we are comparing less than all of the bases in genome x to genome y. In this case we ignore index 4 of genome x, and as such, this Local Alignment produces a match count of 5, which is higher than the Global Alignment match count of 3. We could infer that the T at index 4 of genome x is what's called an **insertion**, in that genome x can be produced by taking genome y, and inserting a T at index 4. Or, similarly, we could infer that genome y is the result of a **deletion** from genome x. In general, insertions and deletions are called **indels**.

As a general matter, we can use Local Alignments to maximize the similarity between genomes, which could be useful for certain applications, such as locating identical protein producing regions, which are referred to as **genes**. However, you can plainly see that we are disregarding information when making use of Local Alignments. As it turns out, intuition corresponds to outcomes, and I showed unambiguously in my paper, "A New Model of Computational Genomics" [1] that Global Alignments outperform Local Alignments for purposes of imputation when using **whole-genome** mtDNA.

Whole-Genomes: In 1981, Fredrick Sanger was the first person to fully sequence a human mtDNA genome. **Sequencing** a genome means that every base pair in the genome is identified, and listed in some order. As a general matter, because sequencing was difficult, scientists focused on particular regions in a given genome, to conserve time, effort, and money. However, as sequencing techniques improved, it became increasingly faster and cheaper to sequence entire genomes, in particular mtDNA, because it contains only approximately 16,500 base pairs. For context, human Chromosome 1 contains 249 million base pairs, making mtDNA uniquely easy to work with.

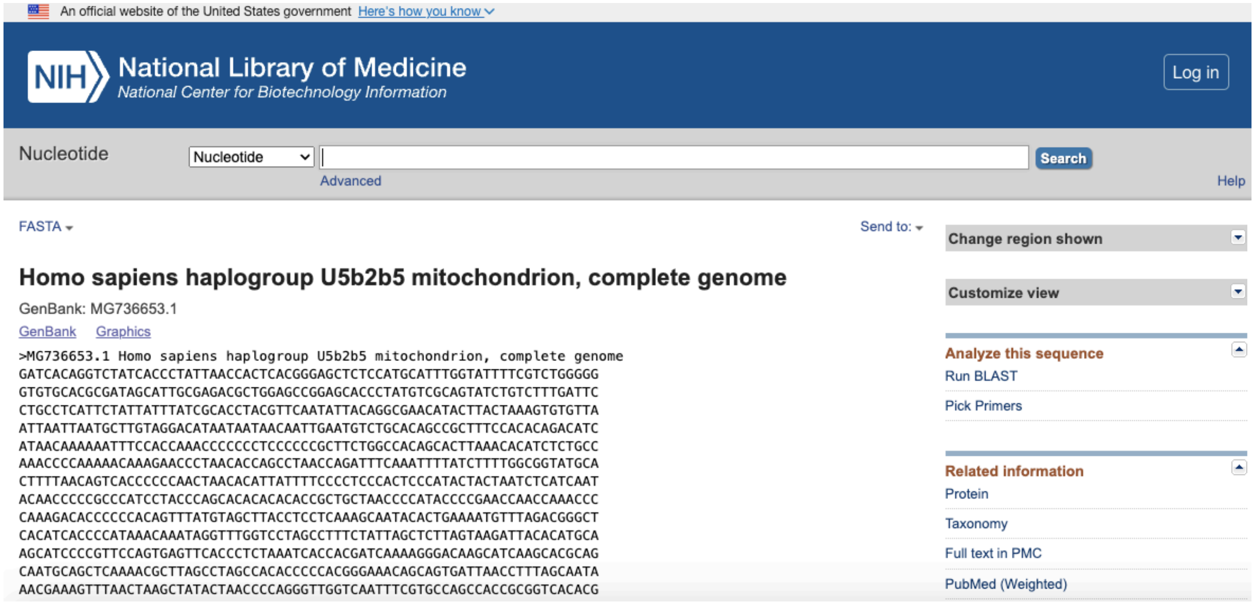
15 Common Bases: Empirically, nearly all human mtDNA contains exactly the same sequence of 15 bases. Specifically, around 99% of the 644 complete mtDNA genomes in the dataset contain exactly these 15 bases. To define the alignment we use in this presentation, we search for those 15 bases in a given genome, and once found, we treat the first base in the sequence of 15 bases as the first base of the entire genome. This also appears to be the default alignment in the National Institutes of Health Nucleotide database.

The 15 bases are: GATCACAGGTCTATC.

We assume the alignment is fixed going forward using these 15 bases, and as a result, for two genomes A and B, we denote the number of matching bases using this alignment as $M = |A \cdot B|$, which we call the **match count**.

FASTA Files: FASTA is a plain text format that contains the actual sequencing data for a genome, or genome segment. All of the FASTA files used to assemble the dataset are taken from the NIH Nucleotide Database, which also features a tool called NIH BLAST, that allows you to compare genomes using a Local Alignment. BLAST is a powerful tool that can quickly compare a given genome to the entire collection of genomes in the NIH Nucleotide Database. However, because BLAST uses a Local Alignment, we will not use it.

The Dataset: The dataset was assembled by simply copy / pasting FASTA files, and reformatting them into an $M \times (N+1)$ matrix, where M is the number of genomes / rows, and $N = 16{,}579$ is the number of bases in the



Homo sapiens haplogroup U5b2b5 mitochondrion, complete genome

GenBank: MG736653.1

GenBank Graphics

```
>MG736653.1 Homo sapiens haplogroup U5b2b5 mitochondrion, complete genome
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGG
GTGTGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTC
CTGCCTCATTCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTA
ATTAATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGCACAGCCGCTTTCCACACAGACATC
ATAACAAAAAATTTCCACCAAACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCC
AAACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCA
CTTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAATCTCATCAAT
ACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACCAACCAAACCC
CAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATACACTGAAAATGTTTAGACGGGCT
CACATCACCCCATAAACAAATAGGTTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCA
AGCATCCCCGTTCCAGTGAGTTCACCCTCTAAATCACCACGATCAAAAGGGACAAGCATCAAGCACGCAG
CAATGCAGCTCAAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATA
AACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGCGGTCACACG
```

Figure 9: A screenshot taken of a FASTA File stored in the NIH Nucleotide Database.

longest genome, stored as columns. Column $N+1$ contains the ethnicity classifier of the genome, where e.g., a classifier of 4 implies the individual in question is Japanese. As such, each row of the dataset consists of $N+1$ columns, where each of the first N entries is one of the bases ACGT, and the last column $N+1$ is an integer classifier, signifying the nationality of the person in question. There's a Genome Population Annex at the end of this presentation, with the ethnicity classifiers and their abbreviations, which is used for a few of the charts.

Predicting Ethnicity: As noted above, mtDNA is inherited directly from the mother, to its offspring, generally without any mutations at all. As a result, it's not obvious that we would be able to predict ethnicity using just mtDNA, since ethnicity is the product of both maternal and paternal lineage. Further, ethnicity is in this case defined at the national level, so e.g., Swedish is distinct from Norwegian. This is far more precise than what is typically done, which is to use what are known as **Haplogroups**, to define categories based on genetic similarity that generally span multiple countries.

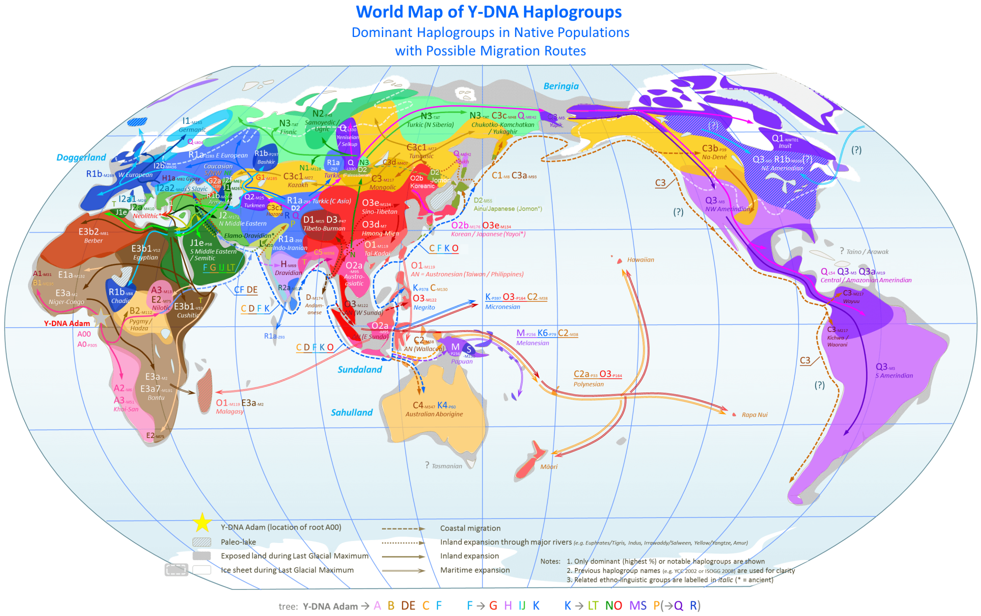


Figure 10: A global map of YDNA haplogroups.

The net takeaway is, Machine Learning is so powerful, you can predict nationality given only mtDNA. The fact that this is possible, implies interesting theories regarding the information conveyed by mtDNA. As a general matter, Machine Learning allows us to not only make predictions, but also develop new theories of genetics, including theories of imputation, ancestry, and selection, all of which are discussed in [1], some of which we'll discuss below.

Nearest Neighbor and Clustering

Nearest Neighbor: The Nearest Neighbor Algorithm was discovered by Evelyn Fix and Joseph Hodges in 1951. It is an incredibly powerful algorithm, and in my opinion began and ended the subject of Machine Learning. Specifically, I proved that the Nearest Neighbor Algorithm will produce literally perfect accuracy given Euclidean data that satisfies certain reasonable assumptions. See, Analyzing Dataset Consistency, generally. This implies a whole host of similar algorithms that run in low-degree polynomial time, that nonetheless produce accuracies comparable to Neural Networks. You can find out more on my website, www.blacktreeautoml.com.

For Euclidean data, the Nearest Neighbor of a given vector x is the vector y, such that the norm of the difference $\|x - y\|$ is minimal over the dataset. That is, given input vector x, the Nearest Neighbor of x is the vector y that is closest to x, hence the name of the algorithm.

DNA is of course not Euclidean, though we can nonetheless construct a rigorous analog. Specifically, given a dataset of genomes A, and a given genome x, we define the Nearest Neighbor of x over A as the genome y such that the match count $|x \cdot y|$ is maximized. In simple terms, the Nearest Neighbor of a given genome is the other genome with the most bases in common with that given genome. This is similar to the Hamming Distance, which would instead count the number of unequal bases, which lacks intuitive appeal in my view, since we're concerned with genetic similarity.

Clustering: A cluster is a collection of similar or otherwise related rows of a dataset. So e.g., if you're analyzing colors, and the dataset is red, pink, blue, and aqua, we could cluster red and pink together, and blue and aqua together, grouping similar colors. In the case of Euclidean data, we can define a distance δ, and then given an input vector x, retrieve all other vectors y such that $\|x - y\| \leq \delta$. That is, we retrieve all vectors that are within δ of x, and that defines a cluster of vectors similar to x.

We can analogously define clusters of genomes by setting a minimum match count M, and given an input genome x, retrieve all other genomes y, such that $|x \cdot y| \geq M$. That is, given x, we retrieve all other genomes that have a match count of at least M with x, producing a cluster of genomes similar to x. If M is too high, the cluster could be empty or too small to be useful. If M is too low, the cluster could be over-inclusive and therefore not useful. This problem is not unique to genomes, and is instead a general problem in Machine Learning, of how to discern between objects, which are typically represented as rows in a dataset. In [1], I present an algorithm that optimizes M using information theory. For this presentation, we will simply fix values of M that work given the context.

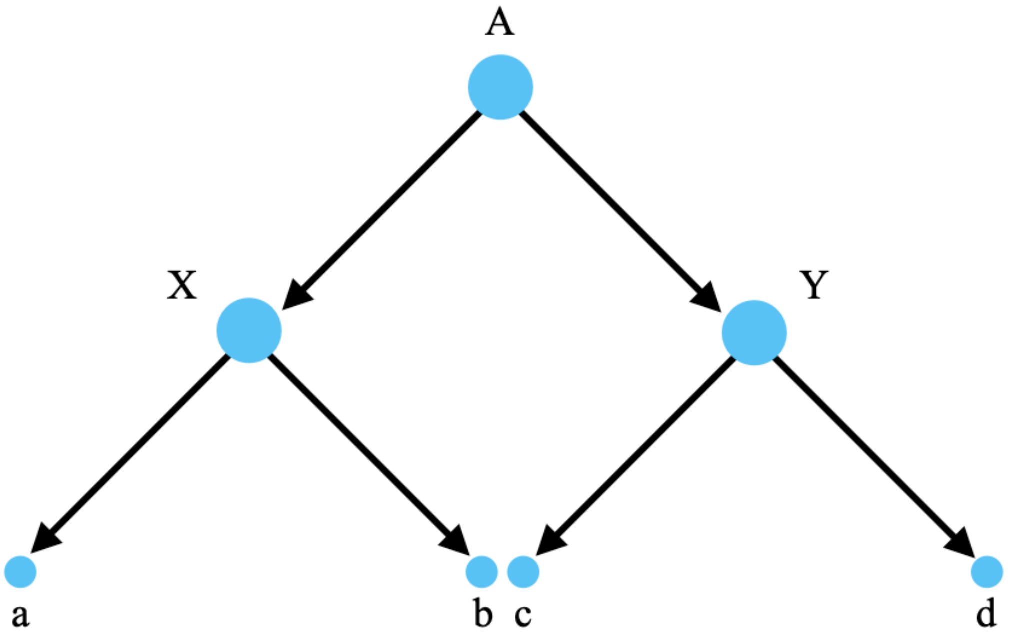


Figure 11: A diagram showing a set of maternal ancestors represented as dots beginning at genome A.

Though mtDNA mutates very slowly as a general matter, if there are any mutations, then in the diagram above, the match counts between e.g., genomes a and b should be higher than the match counts between genomes a and d. This is because genomes a and b will likely share any mutations that occur at genome X. In contrast, they will not contain any mutations that occur at genome Y, and similarly, genomes c and d will not contain any mutations that occur at genome X. Therefore, intuitively, if we increase the minimum match count M for any inquiry, precision should increase, and we will show

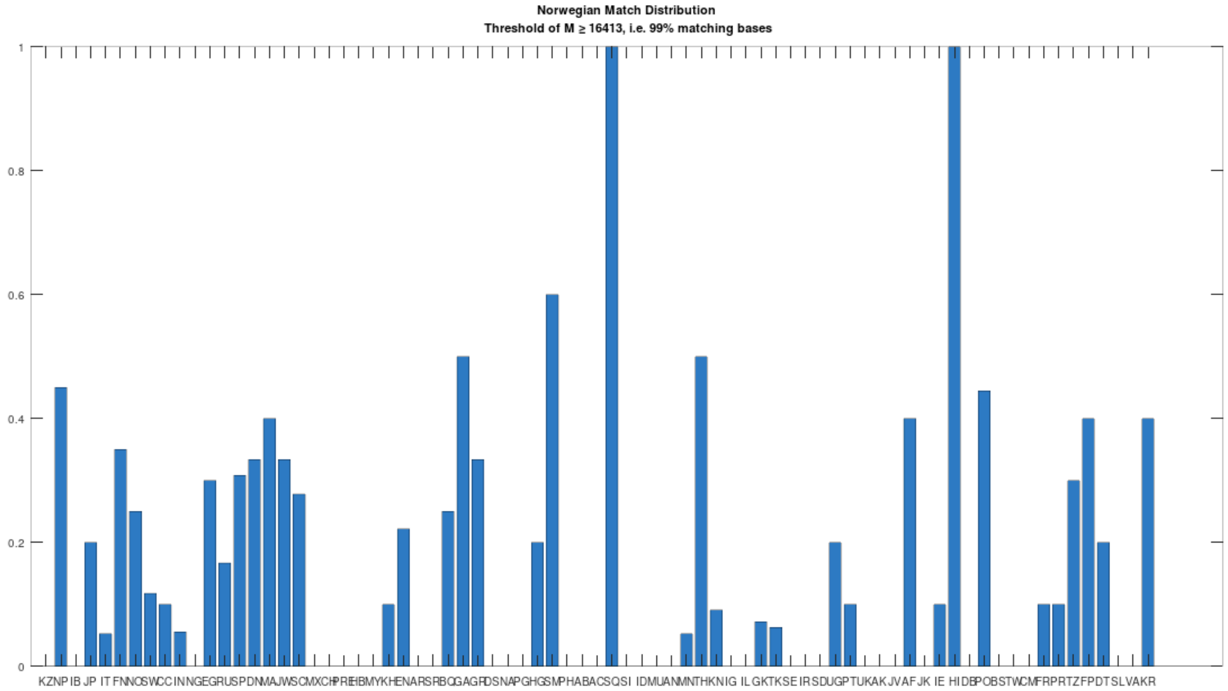


Figure 12: A bar chart showing the distribution of 99% matches by ethnicity for the one input Norwegian genome.

that this is the case below. Though as noted above, if M is too high, the resultant cluster could be completely empty.

Applying Nearest Neighbor and Clustering: Applying Nearest Neighbor to this dataset would entail iterating through each row, and returning the classifier of the row that is the Nearest Neighbor of that row. So if e.g., row 1 is a Kazakh genome, we would find the Nearest Neighbor of row 1, say e.g., row 10, and report the classifier of row 10 as our predicted ethnicity for row 1. That is, our predicted ethnicity for a given row is the known ethnicity of the Nearest Neighbor of that row. This produces an accuracy of 30.87%.

This is of course not great accuracy, but it is significantly higher than chance, which implies that mtDNA alone, can be used to predict ethnicities, which is astonishing. Specifically, there are 75 possible ethnicities, and so chance implies an accuracy of 1.33%. We can therefore conclude, that mtDNA, despite being inherited directly from the mother to its offspring, must include information about ethnicity. This is the type of careful thinking that will allow us to state new theories in genetics, using Machine Learning.

We can improve accuracy by setting a minimum match count M. The intuition is, the higher the match count M between two genomes, the closer the two genomes are to each other, and therefore, the closer they are to their shared maternal ancestor. To gain intuition, assume that in Figure 11 above, A is Norwegian, X is Swedish, and Y is Danish. By increasing M, we're making it more likely that we're comparing two genomes from the same mother that will therefore share all upstream mutations. Further, if you're from the same mother, you must have the same ethnicity, barring truly idiosyncratic outcomes (e.g., two siblings raised in different countries). If this is true, then accuracy should increase as a function of M, and this is in fact the case.

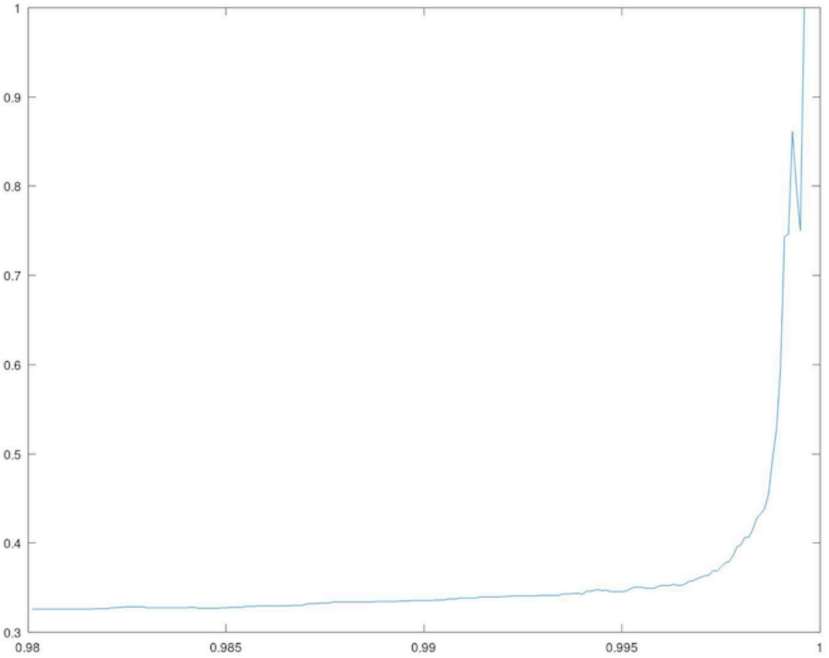


Figure 13: A plot of accuracy as a function of M, expressed as a percentage from $M = .98 \times N$ to $M = 1.0 \times N$.

The plain conclusion is that mtDNA must carry information about ethnicity, despite the fact it is inherited directly from the mother to its offspring. We can therefore, use clusters to meaningfully examine connections between ethnicities. That is, we can take e.g., a Norwegian genome, and ask what other genomes are at least a 99% match (i.e., $|x \cdot y| \geq .99 \times N$). This cluster of 99% matches should provide us with information about what populations are reasonably similar to Norwegians, in terms of maternal ancestry. I say reasonably similar, because you'll note that accuracy at 99% of the genome is still only 33.55%, which is not great, but it is again significantly higher than chance. Further, in the step above, we were trying to predict ethnicity. In this case, we are instead attempting to explore connections between populations.

Again, we are not trying to maximize prediction accuracy, and are instead, using the tools of Machine Learning to engage in scientific inquiry.

Applying this to row 135 of the dataset, which is a Norwegian genome, produces the chart below. The x-axis is labelled by ethnicity, using acronyms that can be found at the end of this presentation. The y-axis gives the percentage of a given population that the input genome matched to. That is, if the input Norwegian genome is at least a 99% match to e.g., 5 out of 10 genomes of a given ethnicity, then the column for that ethnicity will have a height of 0.5 out of 1.0.

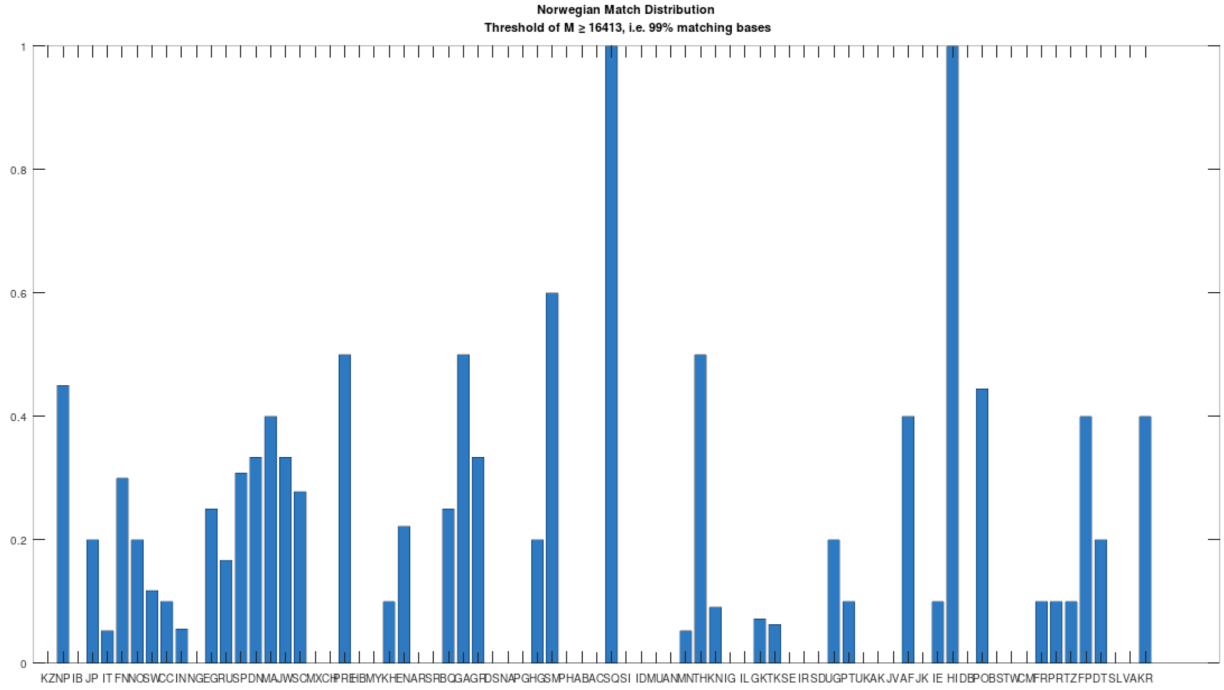


Figure 14: A bar chart showing the distribution of 99% matches by ethnicity for the one input Norwegian genome.

First note that what we have done is to produce a cluster of genomes, by finding all genomes that are at least a 99% match to the input Norwegian genome. This produces a cluster with 105 genomes, which is 15.81% of the 664 genomes / rows in the dataset. Then, we search through that cluster, and identify the ethnicity of each genome in that cluster. If we find e.g., a Finnish genome, we increment the counter for the Finnish column in the chart above, increasing the height of the bar. Then, we divide each column in the chart above by the number of genomes in each ethnicity, normalizing to [0,1].

There are plainly a few surprises, including the fact that 15.81% of all genomes in the dataset, are a 99% match to a Norwegian. Norway is a small country, with less than 6 million people. Though as you'll see, generally speaking, the world is much smaller on the maternal line, though there are nonetheless a handful of distinct categories of maternal lines.

Examining Scandinavian countries first, you can see that the strongest match to any country is to Denmark, despite the fact that Norway shares a giant border with Sweden. Looking closer however, we see that the highest match to a Scandinavian people, is to the Sami people, who are indigenous people in Scandinavia and parts of Russia. The next closest match, is Finland, again despite Norway sharing a giant border with Sweden. Can we formulate a hypothesis that is consistent with even these limited facts, ignoring the bulk of the data in the rest of the chart? This is obviously not Machine Learning, and is instead arguably best described as Genetic Anthropology. However, we are using Machine Learning to inform our research.

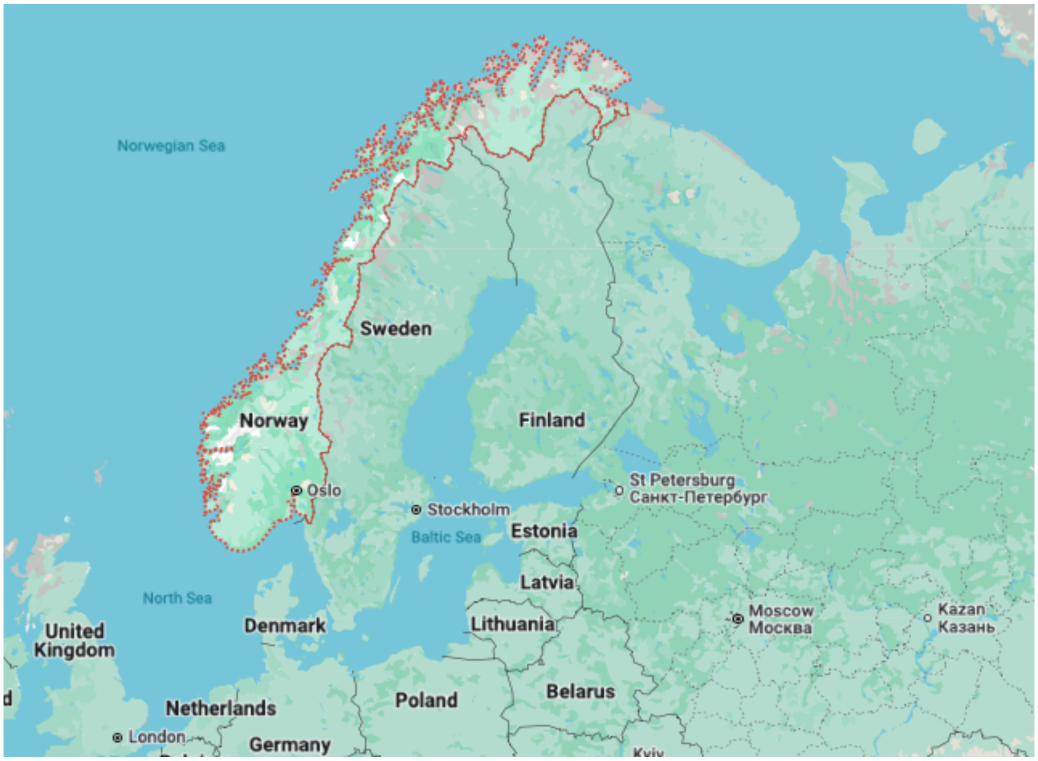


Figure 15: A map of Northern Europe, with Norway highlighted.

Hypothesis formulation is in this case not entirely deductive, and instead requires creativity, and common sense. For example, one possibility is that all of Norway, Sweden, Denmark, and Finland were once homogenous, and similar to the input vector for our cluster, and the Swedes later killed those people off. This is not an absurd hypothesis, given that the highest match among Scandinavians is to the Sami, who are the indigenous people of Scandinavia. But for now, the important point is that formulating hypotheses is non-obvious, and requires some degree of creativity. The net point being, that individuals with subject matter expertise can likely elevate their research to an entirely new level, by using basic Machine Learning techniques like clustering. Note that the kind of thinking we're engaging in can be applied to any subject, where there is a large amount of data, including business applications.

Another surprising observation is that this Norwegian individual has a higher match to the Nepalese, Thai, and Koreans, than to the Germans. The Norwegians are generally considered Germanic people, given their language. However, as we'll see below, the Northern Europeans are, generally speaking, very close to East Asians on the maternal line, though this still leaves open the question of the Nepalese. As you can already tell, hypothesis formation is going to be very difficult. That said, these observations will inform other hypotheses we'll develop regarding overall migration patterns, once we have a few more tools to compare genomes and develop inferences. This will all ultimately culminate into an overall theory of the history of humanity, all the way back to some of the first true humans. Remarkably, the NIH database includes many genomes from early humans, specifically, Denisovan, Heidelbergensis, and Neanderthal. The dataset is, as a whole, courtesy of the National Institutes of Health.

Comparing Populations: In the bar chart above, we compared a single input genome, to every genome in the dataset. Then, we produced a bar chart based upon the normalized number of 99% matches. Because populations are generally heterogenous (e.g., there is more than one maternal line in Norway), if we want to understand the maternal line of a population, we should compare every available genome in that population, to the full dataset. We can of course simply iterate the process above, applying it to every Norwegian genome. However, the normalization step changes slightly. Specifically, there are 20 Norwegian genomes. As a result, when comparing e.g., all Norwegian genomes to all 17 Swedish genomes, there are $340 = 20 \times 17$ opportunities for a successful match count of at least 99%.

Therefore, we divide the actual number of matches by this product, to normalize to [0,1]. Otherwise, the process is no more complicated than what we did above. Below is the bar chart produced comparing all 20 Norwegian genomes to the entire dataset.

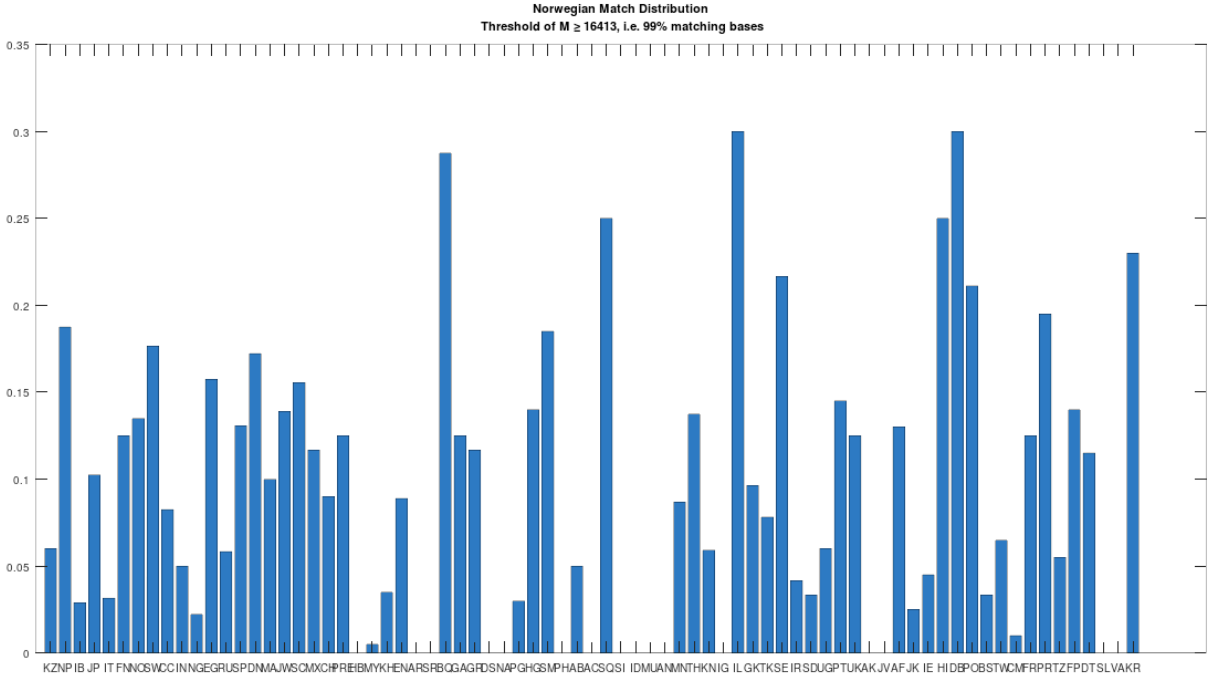


Figure 16: A bar chart showing the results of comparing all 20 Norwegian genomes to the full dataset at $M = .99 \times N$.

We begin by again examining all Scandinavian ethnicities. In this case, note that Denmark is no longer the highest match, confirming that the Norwegian maternal line is heterogenous, and that the Norwegian individual above, just happened to be closer to Danish people. Instead, the highest match across all populations is Icelandic and Dubliner, with about 30% of Norwegians being at least a 99% match to the both. Dublin was settled by Vikings around 800 AD, and so all of this makes intuitive sense. However, we still have a few surprising results, specifically, that the Norwegians are far closer to the Basque people of Spain than they are to the Germans and the Swedes. Further, they are again much closer to the Koreans, Sami, Saqqaq (ancient Greenlanders), and Hawaiians, than they are to the Germans and the Swedes. Moreover, the Norwegians are slightly closer to the Nepalese than the Swedes. Intriguingly, we also see that the Norwegians are closer to the Kenyans, than they are to the Irish and Italians. Note that the Dublin genome is distinct from the Irish set of genomes. We could dismiss all of this as noise, or we can instead consider the possibility that the Norwegian people have a genetic connection to Asian people, and that some indigenous and African people also have a connection to those same populations in Asia. This would imply a migration at some point in time, from Asia to Northern Europe, Greenland, and Africa, creating three distinct populations with common ancestors. It turns out the Swedes do not have as pronounced a connection to indigenous people, but the Norwegians, Finns, and Danes do, leaving us back where we started, in that it seems plausible some form of displacement or even genocide took place in Sweden, creating a population that is decidedly different from the rest of Scandinavia.

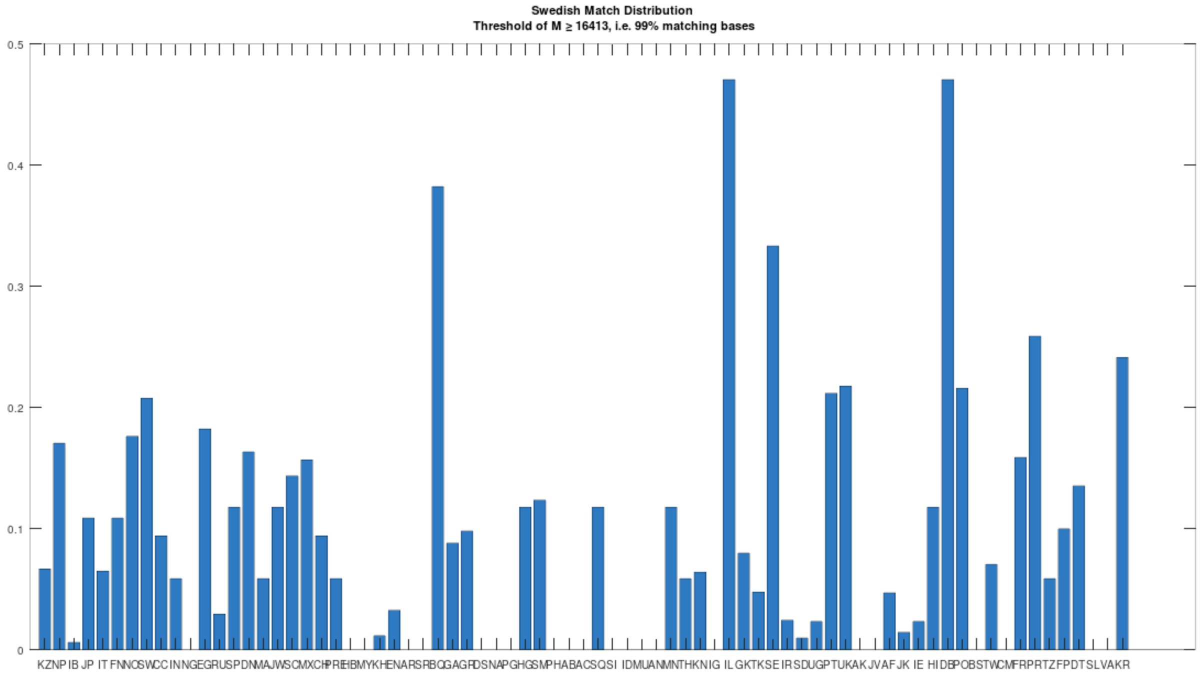


Figure 17: A bar chart showing the results of comparing all 17 Swedish genomes to the full dataset at $M = .99 \times N$.

Above is the bar chart produced by comparing all 17 Swedish genomes to the dataset. We begin by again examining all Scandinavian ethnicities. In this case, the Swedes are again closest to the Icelandic and Dublin genomes, with about 45% of Swedes at least a 99% match to both genomes. Further, we again see that the Swedes are close to the Basque people of Spain, though they are nowhere near as close to the Spanish generally. Further still, the Swedes are relatively close to Nepalese and Korean people. Putting it all together, it seems plausible that Scandinavians were once fairly homogenous, and possibly of Asian origin, and that the Swedes are a distinct and therefore later people, that displaced and possibly killed the indigenous population. As for the Basque, it could be that they are the source population for Northern Europeans, or in my opinion, that Northern Europeans are instead the source population for the Basque, which makes geographical sense, given that the Basque are concentrated in the Basque Region of Spain, in the North.

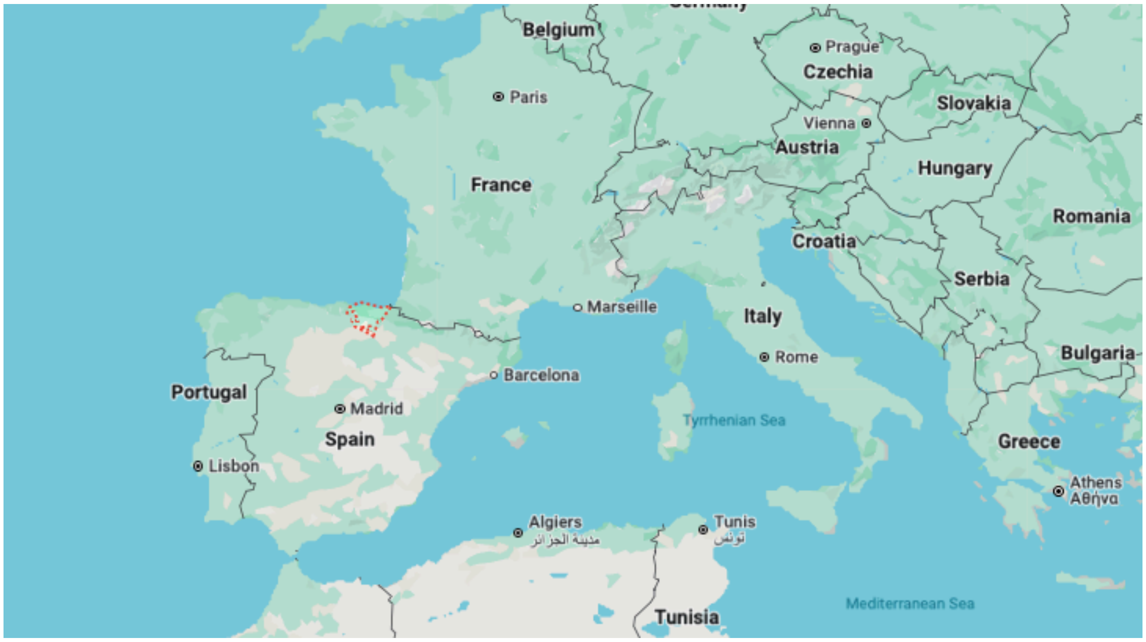


Figure 18: A map of Europe, with the Basque Region of Spain highlighted.

All of this would be a lot easier to consider if there were a method for testing ancestry using mtDNA, and as it turns out, basic combinatorics and probability imply a method that allows us to test for ancestry, and that test is so simple, it can implemented at scale algorithmically.

Testing for Ancestry

Algorithmic Testing: Assume two groups of people originally from location X migrated to distinct locations Y and Z. Further, assume it is known that genome A is the ancestor of both genomes B and C, and that genomes B and C evolved separately, in those two distinct locations Y and Z, respectively.

Although mtDNA does not mutate much if at all from one generation to the next, it does mutate over longer periods of time. We can therefore posit a point in time where there were three basically identical genomes in location X, one of which remained and is still basically identical to genome A, and two of which migrated, eventually producing genomes B and C. Whatever mutations occur in B and C should generally speaking be independent of one another, and therefore, B and C should have increasingly fewer bases in common as a function of time. That is, B and C started out identical to A, and each mutated independently, which should cause the match count between B and C to decrease faster than the match counts between A and B, and A and C. As a result, we have the following fundamental inequalities that are consistent with ancestry from genome A, to genomes B and C:

$$|A \cdot B| > |B \cdot C| \text{ and } |A \cdot C| > |B \cdot C|.$$

That is, if A is truly the ancestor of B and C, and B and C evolved independently, then the above inequality almost certainly holds. You can read footnote 16 of [1] for details on the underlying probability theory, but the intuition is that random, independent mutations shouldn't overlap significantly, causing genomes B and C to evolve away from A. However, one obvious limitation of this test is the fact that it requires 3 genomes, and the test is sensitive to the selection of those 3 genomes. Again, as a general matter, we are using Mathematics and Computer Science to create measurable answers to questions, as opposed to reaching absolute conclusions about a particular topic. Said otherwise, we are asking questions that can be answered using Machine Learning, and then using the answers produced to inspire hypotheses. The results we reach will have to be combined, and we'll eventually supplement our results with archeological observations that are consistent with those results.

The origins of mankind: We'll begin with a big question, specifically, what species of archaic human is the most ancient? We have three "ethnicities", or perhaps better stated, "classes", of archaic humans in the dataset, specifically, Denisovan, Heidelbergensis, and Neanderthal. It turns out, a simply enormous number of people have mtDNA that is plainly related to the mtDNA of all three classes of archaic humans. This is demonstrated in the chart below, where we have set $M = .95 \times N$ as the minimum match count for inclusion in the cluster generated for the single Heidelbergensis genome.

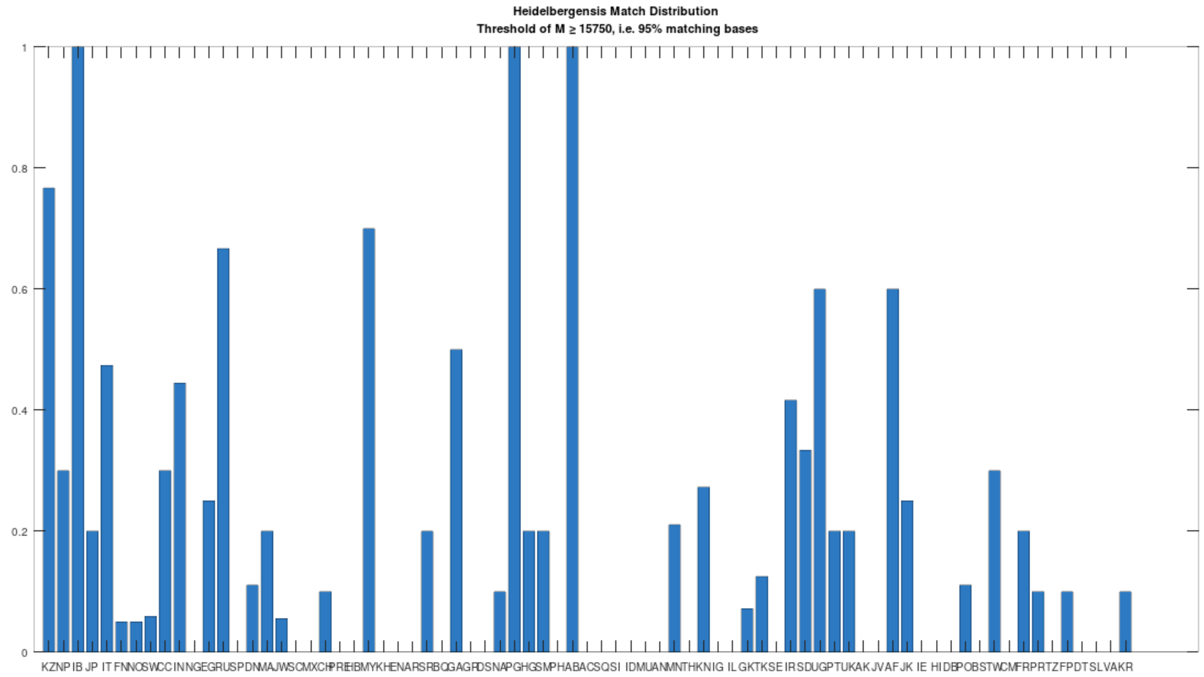


Figure 19: A chart showing the distribution of ethnicities that are at least a 95% match to Heidelbergensis.

If we instead set $M = .97 \times N$, the chart will be completely empty, as no living person in the dataset is a 97% match to Heidelbergensis, but as you can plainly see, many living people are a 95% match. In fact, 100% of the Iberian Roma and Papuans are at least a 95% match to Heidelbergensis. There is no obvious connection between the Spanish Romani and the Papuans, however, it turns out the Roma are originally from India. And as you can see, a little over 40% of Indians are a match to Heidelbergensis. Further, over 60% of Russians are a match to Heidelbergensis. One sensible hypothesis, is that the root population is somewhere in Russia, and they migrated South to India, East to Papua, and West to Europe.

This might seem shocking, but it is well known that modern living people carry at times significant quantities of archaic DNA. Because mtDNA is so stable, this result makes perfect sense. However, as we will see, not all descendants of Heidelbergensis have the same rate of mutation, and in fact, we can plot a clear course of evolution from Heidelbergensis, to the Phoenicians, to the Ancient Egyptians, plainly demonstrating that mtDNA has a heterogenous rate of mutation, which we will discuss below.

Specifically, the rate of mutation for modern day people with Heidelbergensis mtDNA is around 0.0099% per 1,000 years, whereas the Ancient Egyptian line mutates around 0.935% every 1,000 years. The net conclusion is that many modern people carry archaic mtDNA that has barely changed in hundreds of thousands of years, consistent with the slow rate of mutation of mtDNA generally, whereas other people carry mtDNA that is mutating relatively rapidly.

Back to the original question, we can perform the algorithmic ancestry test above, by treating each of Denisovan, Heidelbergensis, and Neanderthal as the root population A. However, because there are a different number of genomes in each archaic population, we must normalize the results to $[0,1]$, to ensure we are testing fairly. In this case it's simple, we divide by the number of times we perform the test for a given root population. There are 8 Denisovan genomes, 1 Heidelbergensis genome, and 10 Neanderthal genomes. Therefore, e.g., treating Denisovans as the root population A will generate $8 \times 1 \times 10 = 80$ tests. We simply count how many times the test is satisfied, and divide that number by 80. Doing so, we find that Denisovans pass the ancestry test 17.5% of the time, whereas Heidelbergensis and Neanderthals pass the test 0% of the time. This is consistent with the hypothesis that the Denisovans are our true ancestors.

It turns out there are a significant number of people with Denisovan mtDNA in Cameroon. If we substitute Denisovan, with Cameroon in the ancestry test, it turns out that the Cameroon pass the test 41% of the time, which is consistent with the hypothesis that the people of Cameroon are the root population of humanity. How could it be that living people in Cameroon have mtDNA that tests as more ancient, than genomes that are hundreds of thousands of years old? It’s possible they literally never left Cameroon, and as a consequence, their mtDNA has not changed much if at all since the dawn of humanity. If instead you look at neighboring Nigeria, you see a very different picture, again with a population that has plain connections to Northern Europeans and Asians. All of this suggesting, the Cameroon simply got lucky, and were not displaced or killed by more modern neighboring populations like the Nigerians, that include many genomes common in Asia, the Middle East, and Europe.

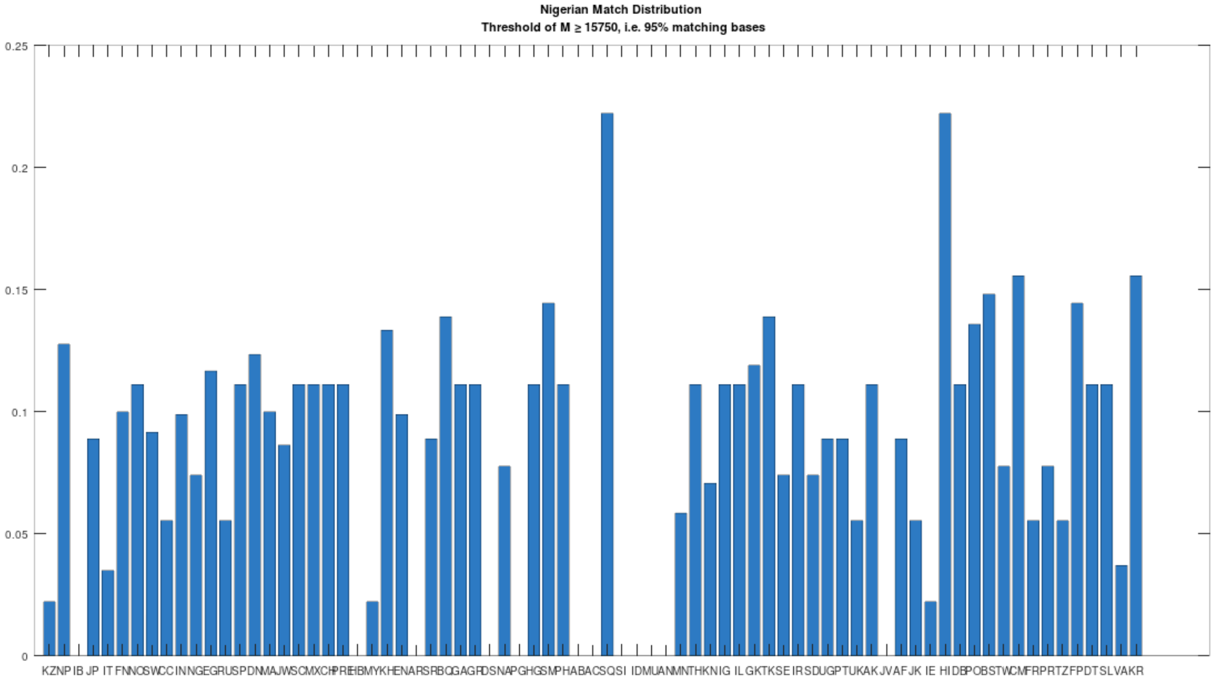


Figure 20: A chart showing the distribution of ethnicities that are at least a 95% match to the Nigerians.

The chart above for Nigerians is plainly similar to the chart for the Norwegians in Figure 14, though note that $M = .95 \times N$, whereas at $M = .99 \times N$, you see a very different chart below that is somewhat more consistent with intuition for an African population. However, you can’t ignore the chart above, and putting it all together, it is consistent with the hypothesis that some Nigerians arrived relatively recently, and share a common ancestor with the Northern Europeans, but have mutated independently by approximately 5% of the total genome.

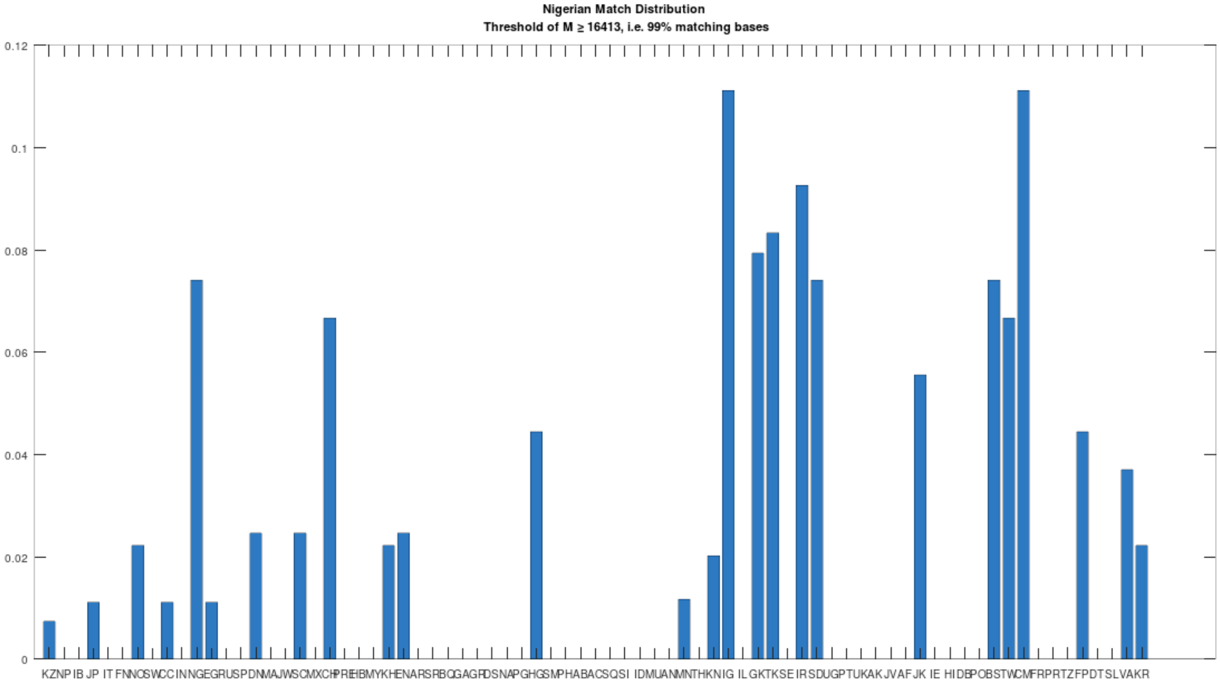


Figure 21: A chart showing the distribution of ethnicities that are at least a 99% match to the Nigerians.

Even at 99%, we see surprising connections to Europeans and Asians, specifically the Greeks and Belarusians, and the Taiwanese, Filipino, Jharkhand of India, Vedic Aboriginals of Sri Lanka, and Koreans. However, note that the y-axis in this case peaks at 12%. If we perform the same test for the Cameroon at 99%, we see they are related basically only to themselves, and the Nigerians, with whom they share a border.

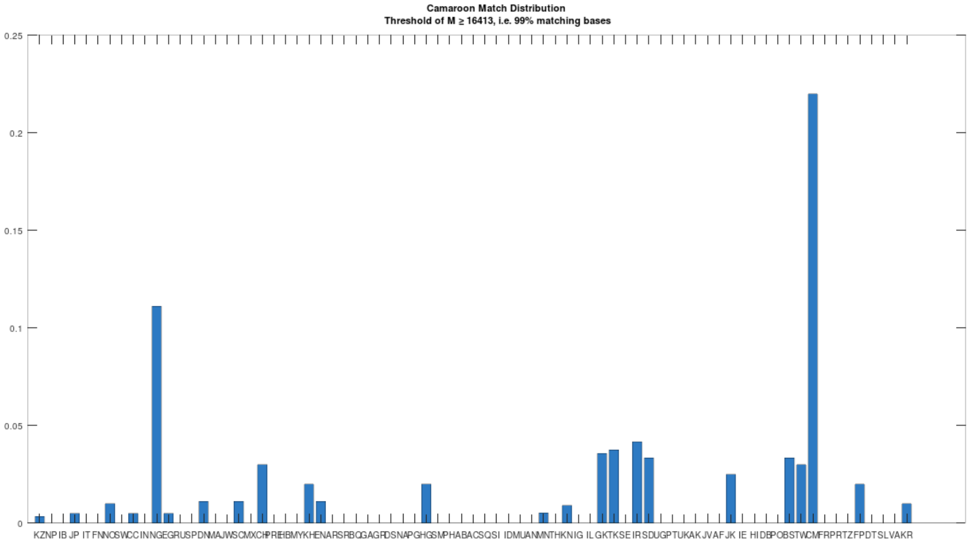


Figure 22: A chart showing the distribution of ethnicities that are at least a 99% match to the Cameroon.

Even though we've only considered Scandinavia, West Africa, and archaic humans, we can already see that humanity presumably originated in Africa, migrated to Asia, and then migrated back to both Europe and Africa, creating the surprising relationships we see above. This is also consistent with the distribution generated by the East African Tanzanian people below, which is plainly similar to the distribution generated by the Swedes, who are again, presumably more modern people than their Scandinavian neighbors.

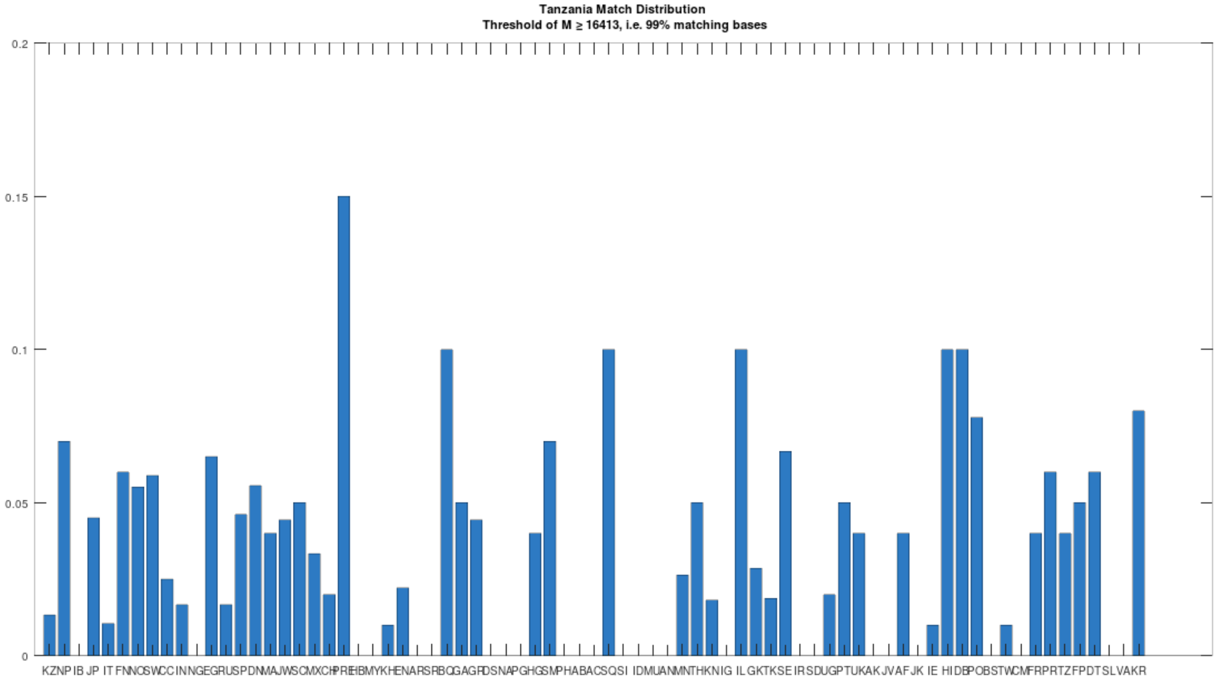


Figure 23: A chart showing the distribution of ethnicities that are at least a 99% match to the Tanzanians.

We can be more precise in our comparisons, by literally taking the difference between the Tanzanian bar chart above, and the corresponding 99% match bar chart generated by the Swedes, and examining the differences. The chart below is produced by taking a column from the Tanzanian chart, and subtracting it from the corresponding column of the Swedish chart. Therefore, a positive number indicates that the Tanzanians have a higher match to the population in question, and a negative number indicates that the Swedes have a higher match to the population in question.

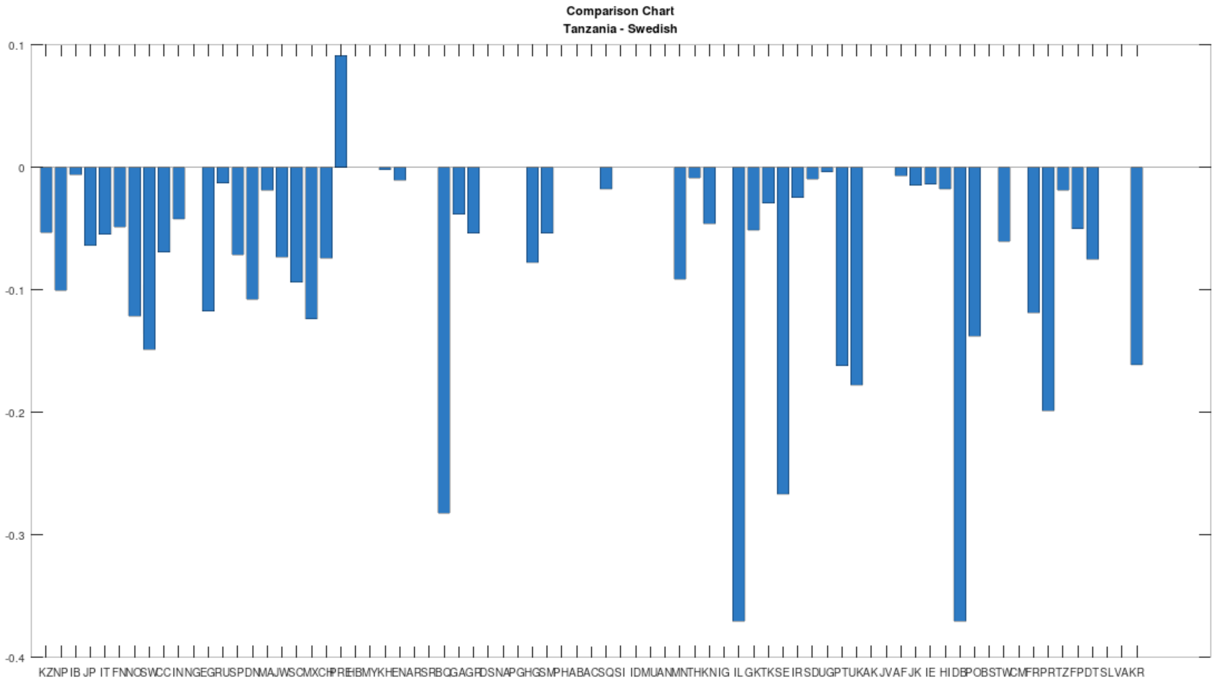


Figure 24: A chart showing the difference between the distributions of ethnicities that are at least a 99% match to the Tanzanians and Swedes.

As you can see, there are plainly significant differences between the two populations, in particular, the Swedes are significantly closer to the Basque, the Icelandic, and the Dubliners, which is not surprising. However, the rest of the differences are relatively small, and in fact, 40% of the columns differ by less than 10%. Because people generally accept the Out of Africa Theory, intuition suggests that the Tanzanians are simply early ancestors of mankind generally, but this is just not true. Instead, the people of Cameroon, who cary deeply archaic mtDNA are far better candidates, and therefore, it seems the people of Tanzania are modern people, that are in particular related to the Ancient Egyptians, which you can see above.

Ancient Egypt

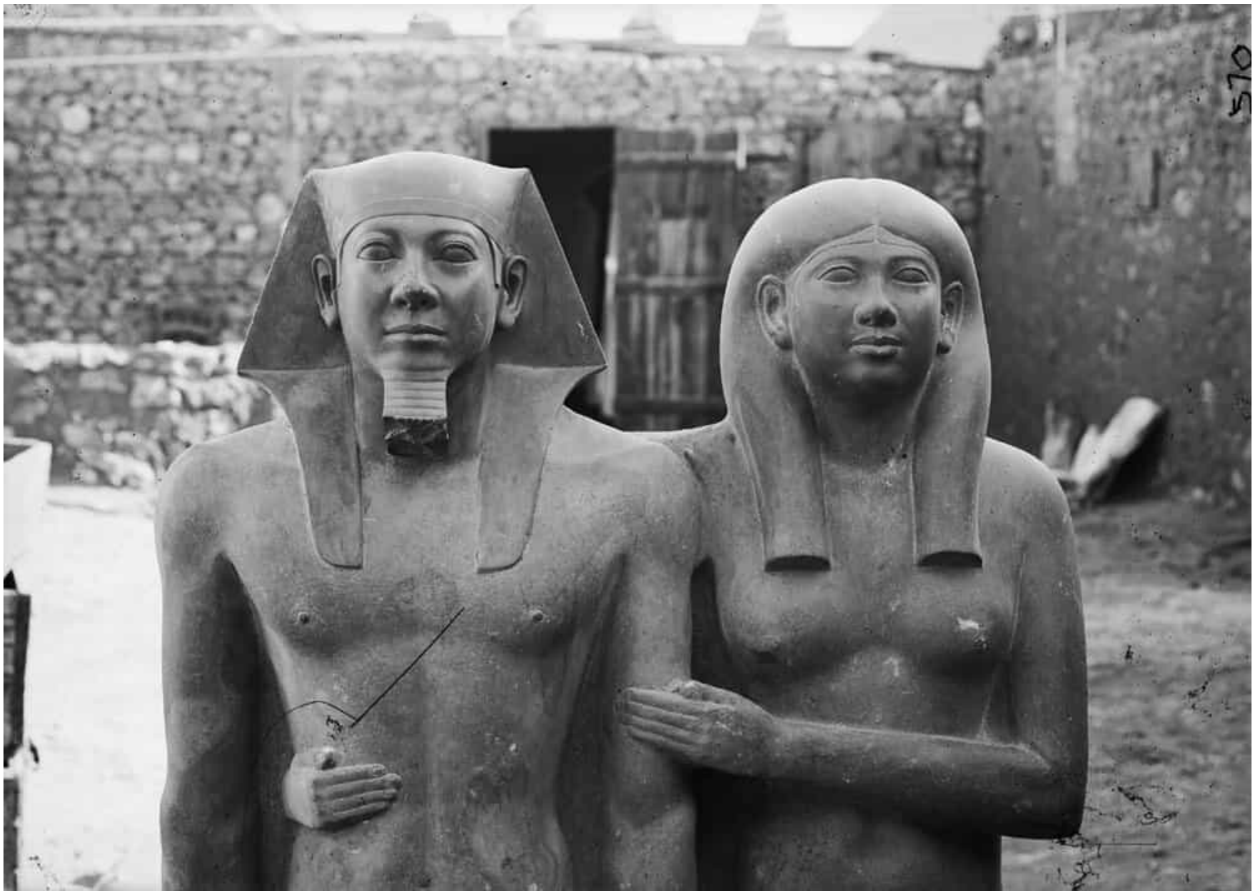


Figure 25: A statue of King Menkaura and Queen Khamerernebty (c. 2,500 BC).

Ancient Egypt is generally divided into time periods, e.g., the Old Kingdom (c. 2,700 to 2,200 BC) versus the New Kingdom (c. 1,600 BC to 1,100 BC). For this presentation, the precise historical labels are less relevant than the fact that Ancient Egypt, like all populations, underwent significant change, in particular the Late Bronze Age Collapse (c. 1,200 BC), which involved wide spread destruction in the Mediterranean generally, purportedly carried out by a mysterious group of people known as the “Sea Peoples”.

The Ancient Egyptians existed for thousands of years, and so change is to be expected. However, there is a sharp break in the physical appearance of their leadership as you get closer to the Roman era. While the people themselves might not have changed much, you can’t ignore the fact that the early Egyptian leaders appear to be Asian people, and mtDNA suggests that they are closely related to the South East Asians in particular. As a general matter, the earlier Ancient Egyptians, both in terms of their appearance and their mtDNA, make it completely obvious that populations migrated from Asia to the West.

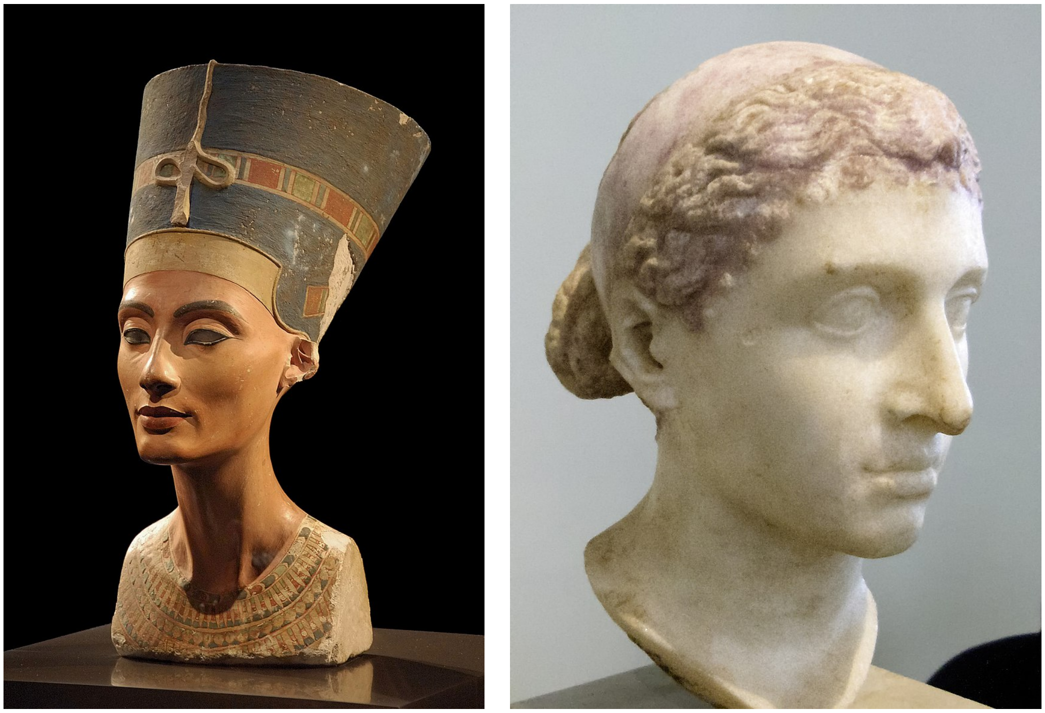


Figure 24: A bust of Nefertiti (left, c. 1,345 BC) and a bust of Cleopatra (right, c. 50 AD).

Assuming the above are accurate representations of both Nefertiti and Cleopatra, it's clear that Nefertiti is either literally Asian, or of Asian descent, whereas Cleopatra is European. That Cleopatra is European, is a surprise to no one, as she is of Macedonian ancestry. That the earlier Ancient Egyptians were Asian, is in contrast probably not generally accepted just yet, but the mtDNA and the archeological evidence make it completely obvious this is the case. You can see this for yourself by simply visiting the Metropolitan Museum of Art, which has an enormous collection of similar works.

Below is a chart showing the distribution of ethnicities that are at least a 99% match to an Ancient Egyptian genome from 4,000 years ago (row 320), which is even earlier than Nefertiti. As you can see, much like the Norwegian distribution, we have matches to the Saqqaq, who were an ancient population from Greenland, and a modern Hawaiian person. Looking closer, we see that 60% of the Sami population (indigenous Scandinavians), are a 99% match to the Ancient Egyptian genome. Further, over 40% of Thai and Nepalese people are a 99% match as well. Further still, you'll note that Denmark, Finland, and Norway are a much higher match than the Swedes. Finally, there's hardly any noticeable representation in Africa, other than modern Egypt itself. Putting it all together, it seems the earlier Ancient Egyptians have the same maternal line as the indigenous Northern Europeans, and that both migrated from Asia, in one case to Northern Europe, and in the other, to North East Africa.

There's some noise to this however, since e.g., Germany, Poland, and Spain are roughly equally significant matches, whereas France and Portugal are not significant matches. But the bottom line is, the Ancient Egyptians are plainly closely related to indigenous Scandinavians, all of whom are presumably of Asian origin, which is astonishing. Note that in Europe, the match is lowest in Italy, Greece, and Turkey, suggesting again, these people really travelled North, generally speaking.

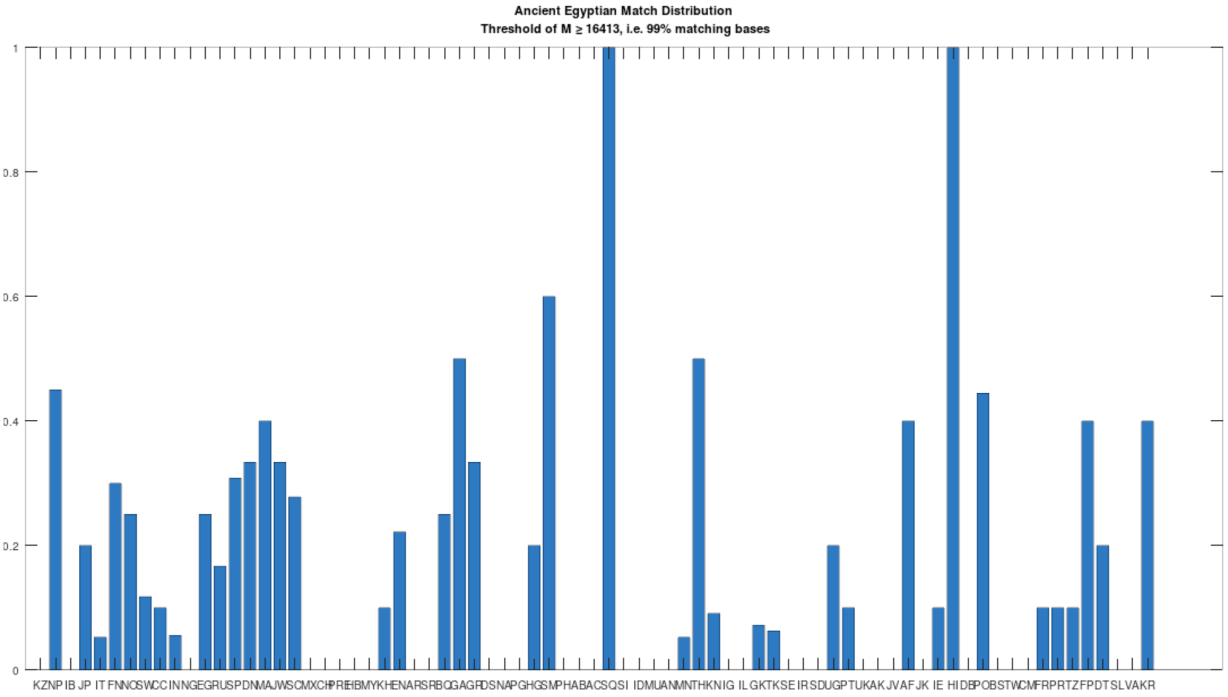


Figure 25: A chart showing the distribution of ethnicities that are at least a 99% match to a single 4,000 year old Ancient Egyptian genome.

Among the Asian ethnicities, we can see China, Mongolia, and India are low matches, whereas Nepal, Korea, Thailand, and Japan are significant matches.

Note that Korea, Thailand, and Japan, are plainly East Asian, whereas Nepal is more central. While anything is possible, I think it’s sensible that the root population of the Ancient Egyptians is somewhere more central in Asia, and that some stayed, whereas others migrated North West, becoming Sami and Saqqaq people and later the Scandinavians, others migrated South East, becoming South East Asians, and a very successful subset migrated to North East Africa, becoming what we refer to as the Ancient Egyptians.

Phoenicia

Figure 26: An Assyrian ivory sculpture, generally attributed to Phoenicians (c. 900 to 700 BC).

Phoenicia was a civilization in the Middle East that lasted from around 2,500 BC to 64 BC. The Phoenician genomes in the dataset were collected from Puig de Molins, Ibiza. Unfortunately, the provenance file for the genomes do not provide a date for the genomes, but the Phoenicians settled Ibiza around the 7th century BC. The Phoenicians are an important civilization historically, that lasted for thousands of years, but they are also important to the evolution of mankind. Specifically, the Phoenician maternal line from Puig de Molins is found in Asia, and in particular, in Sri Lanka, where the population contains nearly identical mtDNA. This suggests the astonishing possibility that the Phoenicians travelled to Asia.

Further, when examining the genomes of the Phoenician people, it is plain that they are the bridge between Heidelbergensis, and the Ancient Egyptian maternal line. That is, as we will see below, the evolutionary history of mankind flows from Heidelbergensis, to Phoenician, to Ancient Egyptian. Further, because it seems the people of Cameroon are the ancestors of all archaic humans, the Phoenicians allows us to complete this portrait of mankind, which begins in Africa, proceeds to the Middle East, then Asia, and then returning West, to Egypt, while also spreading further East, to South East Asia.

Below is a chart showing the distribution of ethnicities that are at least a 99% match to the Phoenicians. As you can see, Phoenician mtDNA is rare, and found only in Japan, Finland, China, Nigeria, Sardinia, Mongolia, Saudi Arabia, and Sri Lanka. Frankly, I have no explanation for the fact that Phoenician mtDNA is found in Nigeria, Mongolia, and China. In contrast, Sardinia and Saudi Arabia make perfect sense give that the Phoenicians were a Middle Eastern and Mediterranean power. As for the Japanese and Sri Lankans, my view is that the Phoenicians, who were certainly a seafaring people, traveled to Asia.

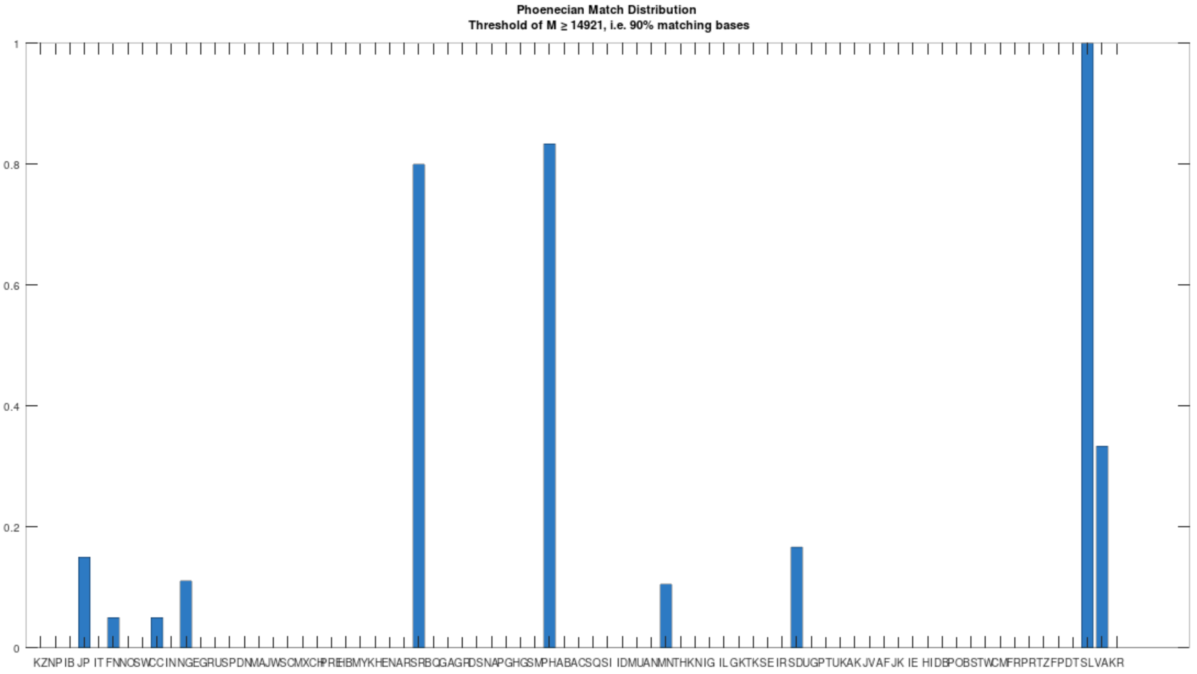


Figure 27: A chart showing the distribution of ethnicities that are at least a 99% match to the Phoenicians.

Because we have an ancestry test, we can test whether or not the Sri Lankans are the descendants of the Phoenicians, or vice versa. Specifically, we can apply the ancestry test above to the Ancient Egyptians, Phoenicians, and Sri Lankans, and just as before, determine which population is the best fit as the common ancestor of the remaining two populations.

Unfortunately, in this case, we have a low confidence answer in that the spreads $|A \cdot B| - |B \cdot C|$ and $|A \cdot C| - |B \cdot C|$ are generally a single digit or low double digit number of bases. In contrast, in the example above, the code set a threshold of 300 bases in order to say that the test was passed. That said, assuming the Phoenicians are the common ancestor satisfies the test 66.67% of the time, the Sri Lankans 25.00%, and the Ancient Egyptians 0.0%. Similar results are produced by substituting the Ancient Egyptians with the Thai and Korean, and in all cases, the Phoenicians are the best fit ancestor. As such, even though the spreads are very low, the pass rate for the Phoenicians is significantly higher than for the Sri Lankans. I would interpret this as implying that even though we can't be confident in the significance of any individual test, we can be confident the Phoenicians are considerably more likely to pass any given test for ancestry than the Sri Lankans. Therefore, if we have to pick, the Phoenicians are the best fit ancestor of the Sri Lankans and Ancient Egyptians, consistent with the hypothesis that the Phoenicians travelled to Asia. Note that the distance from Lebanon (the Phoenician homeland) to Sardinia is comparable to the distance from the

Horn of Africa to Sri Lanka. Finally, note that some credible historical sources claim that the Phoenicians completely circumnavigated all of Africa, making a journey to Sri Lanka plausible.

Evolution from Heidelbergensis

Figure 28: A Heidelbergensis skull (c. 600,000 to 200,000 YBP).

Homo Heidelbergensis was discovered recently, in 1907 in Heidelberg Germany, by Otto Schoentensack, an industrialist turned anthropologist. The genome in the dataset was instead found in Sima de los Huesos, Spain, and is believed to be around 400,000 years old. As noted as above, we've recently developed the ability to sequence entire genomes, in this case the entire mtDNA genome of a Heidelbergensis fossil, extracted from the femur (per the provenance file).

Below we've counted the number of matching bases between the Heidelbergensis and the Phoenician genomes (expressed as a percentage of the maximum), after first breaking up the genome into 100 segments. This is done because a base-by-base comparison would plot roughly 16,000 binary "yes / no" outcomes (i.e., do the bases match), which is very difficult to look at, and doesn't convey much information. In contrast, by breaking up the genome into 100 sequential segments, we can count e.g., how many matching bases there are in the first $\frac{N}{100} \approx 165$ bases. This will allow us to visualize any sharp changes in the match count that could be due to insertions, deletions, or mutations, which as you'll see, really do have location, and are not random. This will become evident when we compare Phoenicians to Ancient Egyptians, and Heidelbergensis to Ancient Egyptians below.

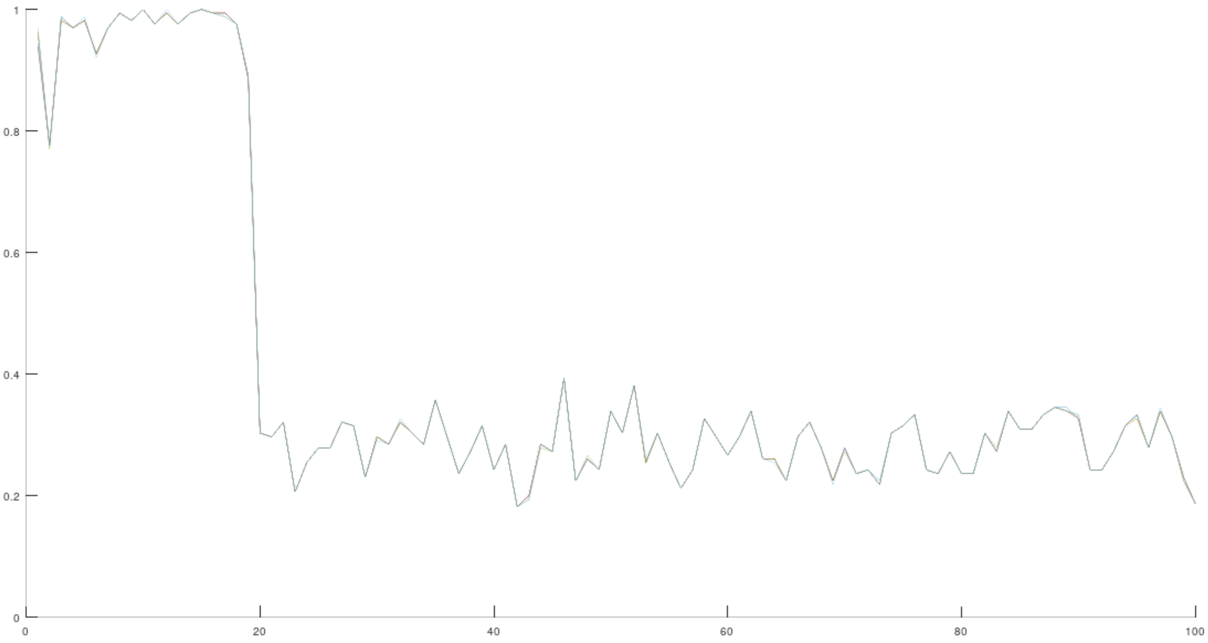


Figure 29: A graph showing the match percentages comparing Heidelbergensis and Phoenician genomes in 100 sequential segments of the genome. The x-axis shows the segment index, and the y-axis shows the match percentage.

As you can see, the match count is very high for the first 20 segments, which contains $165 \times 20 = 3{,}300$ bases. In contrast, it drops down to somewhere around chance (i.e., 25%) to 40%, for the remainder of the genome. Now consider the graph below, which compares Phoenicians to Ancient Egyptians, and you can plainly see, that the Phoenicians evolved away from Heidelbergensis in segments 20 to 100, and the Egyptians complete this process, evolving away from both Heidelbergensis and the Phoenicians in segments 1 to 20, producing a global maternal line that is heavily represented in Northern Europe and East Asia.

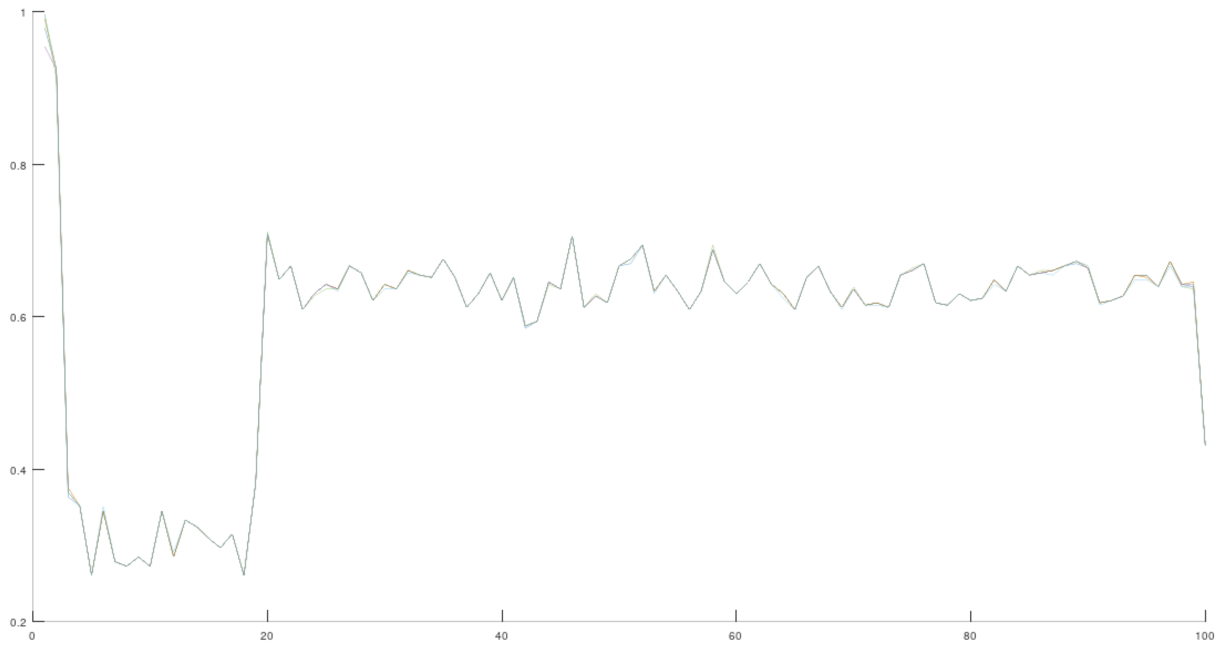


Figure 30: A graph showing the match percentages comparing Phoenician and Ancient Egyptian genomes in 100 sequential segments of the genome. The x-axis shows the segment index, and the y-axis shows the match percentage.

Note that segments 1 through 20 show that the Phoenician and Ancient Egyptians have little more than chance in common in that section. Further, the match count beyond segment 20 is less than 100%, and is instead around 70%, suggesting that in addition to shedding the first 20 segments inherited from Heidelbergensis, the Ancient Egyptian maternal line also evolved away from the Phoenicians to a significant degree on the remainder of the genome.

Finally, consider the graph below, which compares Heidelbergensis to the Ancient Egyptians. As you can see, there's little more than chance in common, except in the opening handful of segments. As noted above, basically all of the genomes in the dataset, including the archaic genomes, have exactly the same 15 bases in common. More generally, the opening 150 bases are very consistent across all genomes, and are therefore not changing much over time. This again suggests that even though mtDNA is a loop, it has objective location and therefore alignment, consistent with the use of the Global Alignment utilized throughout this presentation.

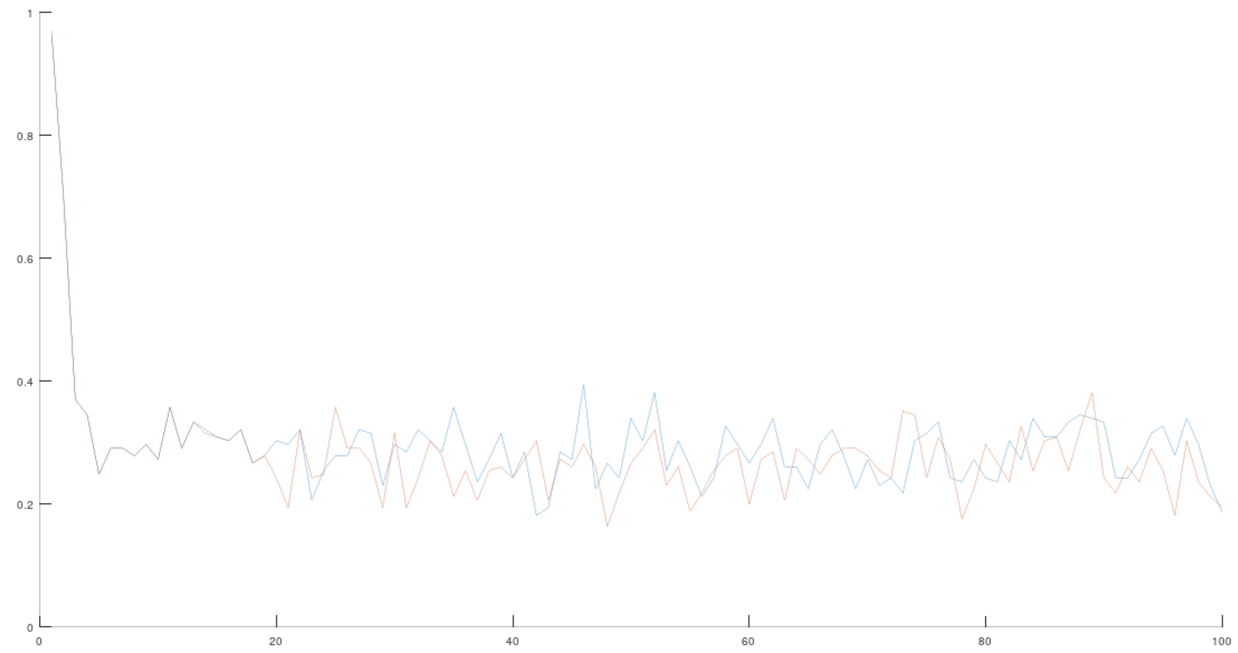


Figure 31: A graph showing the match percentages comparing Heidelbergensis to the Ancient Egyptians in 100 sequential segments of the genome. The x-axis shows the segment index and the y-axis shows the match percentage.

Because the Cameroon test as the ancestor of both Heidelbergensis and the Neanderthals, and Heidelbergensis is plainly the ancestor of the Phoenicians, who are in turn the ancestors of the Ancient Egyptians, who appear to be Asian people, we have therefore reconstructed what could be the entirety of human history, from Africa, to the Middle East, to Asia, and back. Not bad.

As a general matter, this work demonstrates the power of Machine Learning techniques, when applied carefully in a manner that is not necessarily seeking to predict outcomes with high accuracy, but is instead used to create measurable hypotheses, in cases where it was arguably impossible before the advent of computing and routine accumulation of enormous amounts of data. In particular, note that this dataset consists of 664 complete mtDNA genomes, each containing 16,579 bases. This works out to around 1.1 million bases in total, yet these algorithms run in seconds. This kind of work was literally impossible, only decades ago, yet here we are.

Genome Population Annex

Population Class / Name	Abbreviation	No. Of Rows / Genomes
1 = Kazakh	KZ	30
2 = Nepalese	NP	20
3 = Iberian Roma	IB	19
4 = Japanese	JP	20
5 = Italian	IT	19
6 = Finnish	FN	20
7 = Norwegian	NO	20
8 = Swedish	SW	17
9 = Chinese	CC	20
10 = Indian	IN	18
11 = Nigerian	NG	9
12 = Egyptian	EG	20
13 = Russian	RU	6
14 = Spanish	SP	13
15 = Danish	DN	9
16 = Maritime Archaic	MA	10
17 = Ashkenazi	JW	18
18 = Scottish	SC	18
19 = Mexican	MX	3
20 = Chachapoya	CH	10
21 = Ancient Egyptian	PRE	2
22 = Heidelbergensis	HB	1
23 = Mayan	MY	10
24 = Khoisan	KH	10
25 = English	EN	9

Population Class / Name	Abbreviation	No. Of Rows / Genomes
26 = Ancient Roman	AR	5
27 = Sardinian	SR	5
28 = Basque	BQ	4
29 = Georgian	GA	2
30 = German	GR	9
31 = Denisovan	DS	8
32 = Neanderthal	NA	10
33 = Papau New Guinea	PG	5
34 = Hungarian	HG	5
35 = Sami	SM	10
36 = Phoenecian	PH	6
37 = Ancient Bulgarian	AB	1
38 = Ancient Chinese	AC	1
39 = Saqqaq	SQ	1
40 = Sol. Islands	SI	1
41 = Indonesian	ID	1
42 = Munda	MU	2
43 = Andamanese	AN	1
44 = Mongolian	MN	19
45 = Thai	TH	4
46 = Kenyan	KN	11
47 = Igbo	IG	1
48 = Icelandic	IL	1
49 = Greek	GK	14
50 = Turkish	TK	16
51 = Sephardic	SE	3
52 = Iranian	IR	12

Population Class / Name	Abbreviation	No. Of Rows / Genomes
53 = Saudi	SD	6
54 = Uyghur	UG	5
55 = Pashtun	PT	10
56 = Ukrainian	UK	10
57 = Ancient Khoisan	AK	1
58 = Javanese	JV	1
59 = Ancient Finnish	AF	10
60 = Jarkhand	JK	4
61 = Irish	IE	10
62 = Hawaiin	HI	1
63 = Dublin	DB	1
64 = Polish	PO	9
65 = Belarus	BS	3
66 = Taiwanese	TW	10
67 = Camaroon	CM	10
68 = French	FR	10
69 = Portuguese	PR	10
70 = Tanzania	TZ	10
71 = Filipino	FP	10
72 = Dutch	DT	10
73 = Sri Lanka	SL	1
74 = Vedda Abor.	VA	3
75 = Korean	KR	10

Installing the BlackTree Genetics Library

The BlackTree Genetics Library is completely free for non-commercial purposes. **You may not republish any of this material or the Library**. If you believe your use of the Library could be for a commercial purpose (i.e., any purpose for which you are paid either directly or indirectly), contact me and we can discuss if your use is truly non-commercial. Note the Library is setup for Mac, but can be modified to work in Windows. Finally, note that all of these downloads are free.

1. Download and install Octave from GitHub.
2. Download and unzip the dataset from DropBox, and identify the dataset file "mtDNA_dataset.txt".
3. Download and unzip the BlackTree Genetics Library from DropBox.
4. Add the BlackTree Genetics Library by typing the following code into the Octave GUI command line and pressing ENTER, using the folder location of the unzipped file:

```
addpath("/Users/[yourname]/[folder location]")
savepath
```

5. Run the following script from the BlackTree Genetics Library by Copy / Pasting the entire file into the Octave GUI command line and pressing ENTER:

```
UPDATED Genetic Preprocessing.m
```

IMPORTANT: You must update the variable "file = "/Users/charlesdavi/Desktop/Datasets/mtDNA/mtDNA_Dataset.txt" to reflect the location of the dataset file on your computer.

Using the BlackTree Genetics Library

To produce clusters, Copy / Paste "SECTION 4 Fixed Percentage Clustering_CMNDLINE.m" into the Octave GUI command line and press ENTER.

To test for ancestry, Copy / Paste "three_way_test.m" into the Octave GUI command line and press ENTER.

To run Nearest Neighbor on a row, set the variable "index" to the desired row number, and run the function "Genetic_Nearest_Neighbor_Single_Row" by Copy / Pasting the following code into the Octave GUI command line and pressing ENTER:

```
[nearest_neighbor, match_count, match_vector, match_matrix] = Genetic_Nearest_Neighbor_Single_Row(index, dataset, N);
```

The Library includes a total of 130 source code files which you are free to explore and use for non-commercial purposes.