Rayees Rahman

Practical Analysis of a Human Genome

12/18/2017

Individual Capstone Project

Analyzing my mitochondrial heritage using haplogroups and machine learning

In addition to facilitating one of the most well-known reactions in biochemistry: ATP synthesis; mitochondria and mitochondrial DNA have shed light not only onto the processes behind cellular respiration and physiology, but, surprisingly, also onto to the complex field of human evolution and ancestry. Mitochondria, as well as the chloroplasts found in plants, are thought to have originated as organelles in eukaryotic cells hundreds of millions of years ago through an 'endosymbiotic' process. Originally, mitochondria derive from an ancient alphaproteobacterium living independently of the progenitors of modern eukaryotic cells. At some point in time this bacterium was endocytosed, or swallowed, by a proto-eukaryotic cell. This endocytosis resulted in a symbiosis between the a proto-eukaroyotic cell and the bacteria. The bacteria would remove toxic oxygen molecules from within the proto-eukaroyotic cell by generating ATP, and the larger cell would protect the bacterium from the harsh external environment of a much younger Earth. The endosymbiosis ultimately resulted in two distinct DNAs, two distinct evolutionary linages, present in a single cell, both replicating and existing independently of one another.

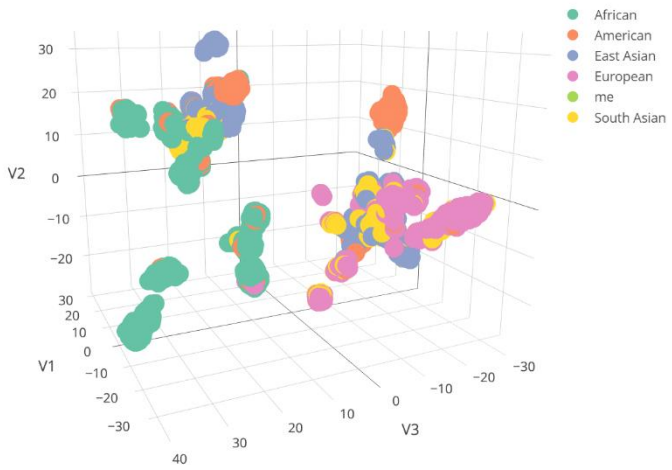
Over vast stretches of time the alphaproteobacterium started to become more and more specialized in ATP production, losing thousands of genes simply because it was too costly, in terms of energy, to keep them. Why express genes for protein transport or structure when your

host can do that for you? Eventually, the once several thousand genome of the ancient alphaproteobacterium shrunk to tens of genes in modern human mitochondria. As a result of this shrinkage in genetic content, polymorphisms in mitochondrial DNA became exceedingly rare over evolutionary time. This was because DNA polymerase, the enzyme that replicates DNA, makes, on average, 1 mistake over 10 billion bases. There are many variations in human genome since there are approximately 3 billion bases, which means that the possibility for error is relatively high. However mitochondrial DNA is approximately only 16,500 bases, this makes variations extremely rare, especially over short amounts of time. This means that during the process of human evolution, which, time-wise, is tiny compared to the age of life on Earth, distinct sets of variations in mitochondrial DNA, called haplogroups, can define certain groups and migrations of human peoples. And because mitochondrial DNA is passed maternally we can identify one's maternal ancestry with high accuracy.

Using current software such as Haplogrep and phylotree we can discover our maternal linages using the most parsimonious reconstruction of our mitochondrial variants based on predefined haplogroups. When I analyzed my own mitochondrial DNA, for example, I discovered that my haplogroup was R11'B6, the R'11 implying that I descended from the R line of matriarchs that is thought to have spread from the Middle East into parts of South Asia and China. B6 is a further sub classification that specifies that I have a 9 base deletion in my mitochondrial DNA that is present in East and South East Asia.

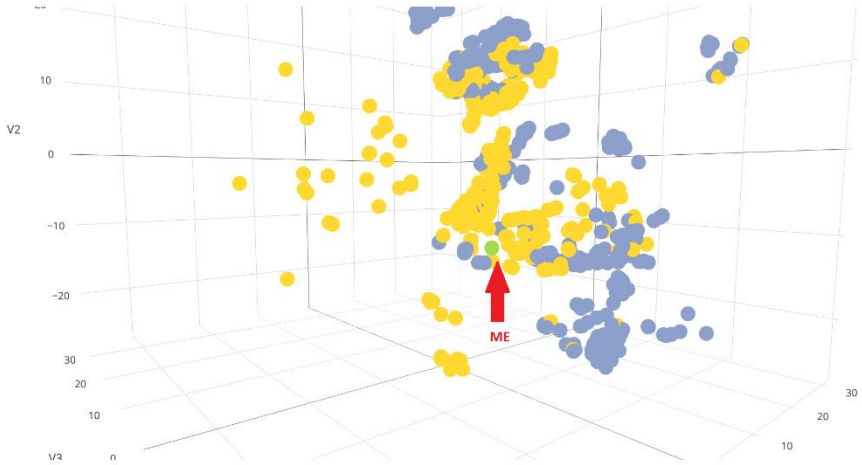
This result was puzzling to me. I know for certain that my origins are from Bangladesh, so the R lineage made sense, however the B6 haplotype was odd. There were a few possible explanations for the B6 haplotype: my ancestors may have come from China and into Bangladesh several hundred years ago, the sample size for the B6 haplotype was small and was

biased for East and South East Asian populations, or, as unlikely as it sounds, I had a spontaneous 9 base pair deletions in my mitochondrial DNA that gave the incorrect Haplotype. Additionally, I found that after discovering my haplotype that the information present about it was minuscule at best, and that most websites and articles had a significant bias towards European populations. Using machine learning, I was wondering if I could define my own haplogroups that were both consistent with known human migration patterns, potentially create an algorithm to predict a person's ethnicity based on their mitochondrial variants and perhaps find out more about my mitochondrial lineage.

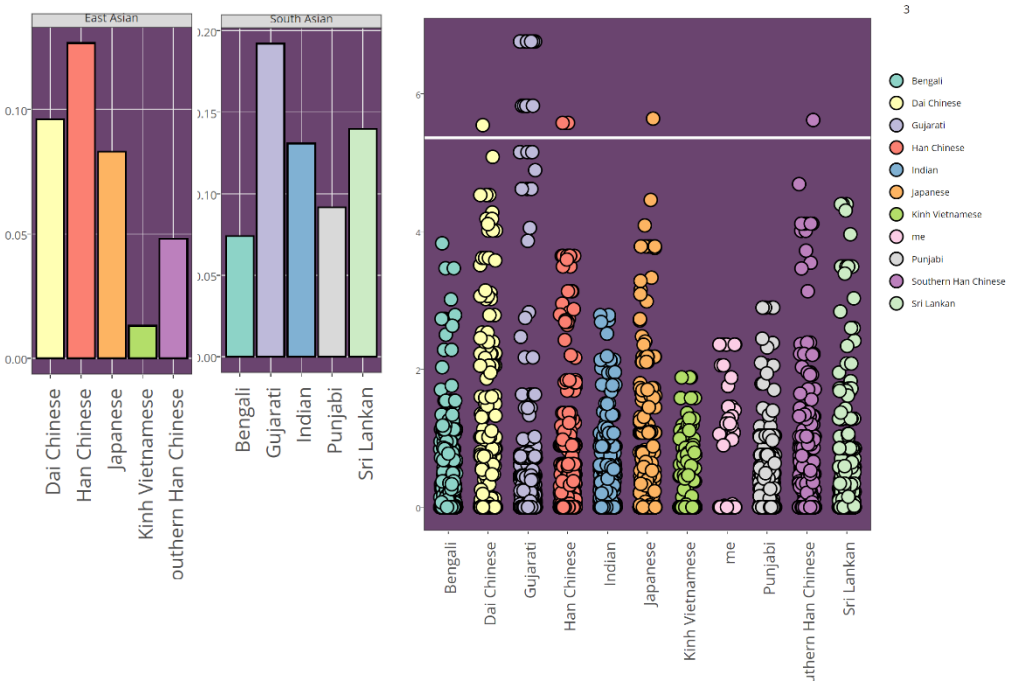


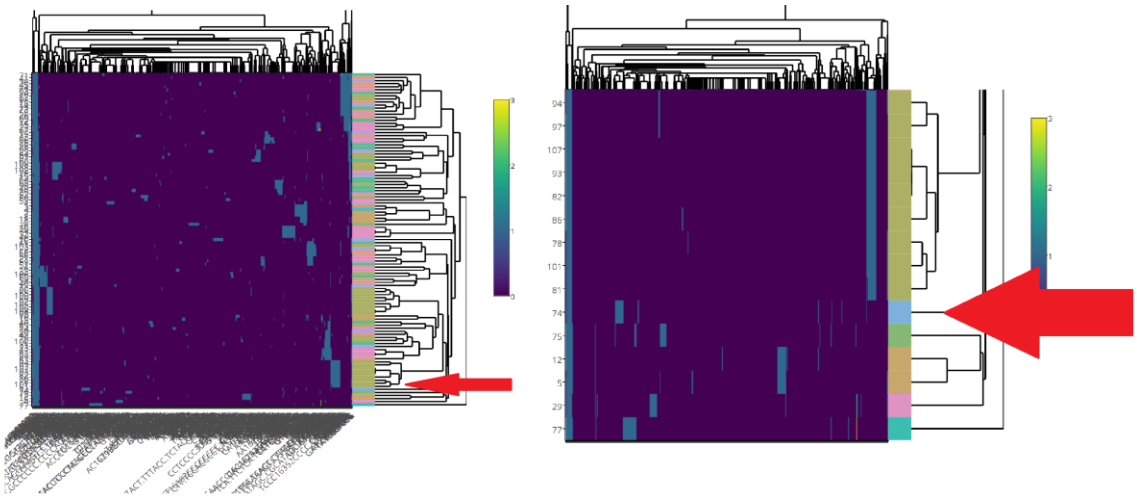
Using t-distributed stochastic neighbor embedding (t-SNE), I clustered mitochondrial presence or absence haplotypes data from the 1000 genomes consortium. I plotted this result in 3 dimensions using the R programming language, right. Here we can see that certain population groups are distinctly separated using this method, specifically European and African populations have a plane of separation between them, when East Asian and South Asian populations are combined. This mirrors the migration of human populations well, since south Asian and east Asian populations do not have as much divergence times as Europeans and Africans and as a result, it seems that all Asians have some admixture in their mitochondrial DNA, irrelevant from geographic location.

Looking into my own location on this plot, I noticed that my mtDNA variants cluster quite well with South Asian and East Asian populations, this is consistent the R haplogroup which defines the Middle East, East Asian and South Asian populations.



To further investigate what variants define my lineage and of people that cluster with me, I plotted a histogram of the ethnicities that clustered with me, along with a heatmap and a Manhattan plot to discover what variants are present or absent in our mitochondrial DNA.





After plotting the heatmap I discovered that while I clustered with Gujarati, a South Asian subpopulation, I lacked several key variants that defined that subpopulation within my cluster, and that I have several unique variants that are not seen in other groups. For example I had C8275T, A302C variant that uniquely identified me compared to the other members of my cluster.

Following this I wanted to see if I could use a machine learning algorithm to predict what ethnicities people are based on their mitochondrial DNA haplotypes. I used a Random Forest algorithm to define superpopulations (i.e. South Asian, African, etc.). After training I observed that my classifier had 89% accuracy, or an 11% error rate for superpopulation classifications. Looking at the error rates for specific populations, my algorithm had particular difficulties in differentiating East Asian (7% error rate) from South Asian (11% error rate) populations, often confusing one for the other. African, American and European populations had under 3% error rates, I had noticed that my model also had some difficulties separating African populations from South American populations, though not to the extent of Asian populations.

Ultimately what I learned from this experience is that while mitochondrial DNA is highly definitive of one matriarchically lineage, there is much work that need to be done in terms of assigning and describing haplogroups well. My method described results that was consistent with what I expected for what my lineage should be, but it was still not definitive. *De novo* mutations, especially ones that are convergent with other mitochondrial mutations highly bias my results, while haplogroups, themselves give information about population movements, may be inaccurate due to lack of samples from diverse areas. Both of these problems, as well as latent admixture in the variants, affected my classifier as well, which wasn't extremely accurate in determining superpopulations just from mitochondrial DNA. Going forward I feel that in order to develop a good classifier for ethnicity, we may need to include autosomal DNA in the process.