Inferring the Early History of Northern South America through Mitochondrial DNA Analysis

Mateo Rojas, Miguel Vilar

Abstract

Prior to the mass colonization of the Americas, early peopling consisted of Indigenous Americans who, over the course of tens of thousands of years, came to occupy the continents that would later be settled by the ambitious European powers of the time. Through the use of mitochondrial DNA (mtDNA) sequencing throughout numerous samples, as well as temporal analysis of genetic mutations within samples of the same haplogroup, the migrations of different early indigenous populations can be tracked across the geography of the continent, and the ethnographic composition of different regions can be reconstructed with respect to the time period. Given that haplogroups belonging to maternal Amerindian descent are identified by haplogroups A, B, C, and D, and that genetic variations in haplogroups can be tracked over time, it is possible to approximate a model of the early genetic composition of Colombia and other regions of northern South America through the use of samples from model haplogroups like A2al, B2d14, C1d2, and D4h3a. For this study, 94 samples across these haplogroups have been surveyed, and individual phylogenetic trees have been constructed, which each infer a genetic timeline for the mutations found under each haplogroup. With the construction of a phylogenetic tree for each of the four haplogroups, a larger, cumulative phylogenetic tree was also constructed, which in combination with the known presences of haplogroups across the geography of Colombia, can provide insight into the early migration and settling patterns of these early pre-columbian Indigenous Americans.

Introduction

Ever since the first full sequencing of a human genome in 2002, DNA testing kits have become ubiquitous and highly popular for those seeking genetic consultation, forensic analysis, and for the most part, simple curiosity about one's ancestry and origin. Since then, companies such as Ancestry, 23andMe, and FamilyTreeDNA have grown to astonishing proportions, and have compiled equally astonishing amounts of data on the genetic history and composition of their clientele. Utilizing FamilyTreeDNA's online mtDNA database, one is able to not only access anonymized samples, haplogroups, and the individual mutations that comprise them, but also geographic approximations for the dawn of novel mutations and haplogroups across time. This allows researchers to construct migratory patterns for populations of different ethnographic origin, and the ability to simulate the early peopling of different regions across the world.

Over the past few months, we have been utilizing this same mtDNA database to compile the Amerindian genetic composition of present-day Colombia, to simulate the migratory patterns of these early-Amerindian lineages into northern South America.

Mitochondrial DNA

mtDNA is present in almost all eukaryotes and multicellular organisms in the form of haploidic, circular-shaped DNA molecules, called plasmids. mtDNA, as opposed to nuclear DNA, is inherited solely from one's maternal lineage, and thus is not exposed to the nature of genetic recombination. As a result, mtDNA is very stable and consistent across time, while still experiencing mutations around 10-20 times faster than nuclear DNA. These comparatively rapid mutations coupled with their preservation across time makes mtDNA highly desirable for tracking one's maternal ancestry, as the time period and location in which novel mutations were acquired can be approximated.

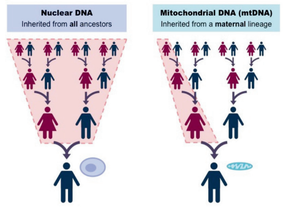

Figure 1. A depiction of the unique inheritance pattern of mitochondrial DNA, wherein only maternal mtDNA is inherited, and only females pass it on to their descendants.

Methodology

While haplogroups A, B, C and D are too large and historically long-lived to acquire specific enough results, especially solely by human effort, using newer variants of these haplogroups that are prominent in samples acquired from Colombia are much easier to analyze for research purposes. To this end, haplogroups A2al, B2d14, C1d2, and D4h3a were identified and served as the primary focus for the study, given their presence in northern South America, and their workable sample sizes. A total of 94 samples were analyzed, comprised of 46 samples from haplogroup A2al, 15 samples from haplogroup B2d14, 10 samples from haplogroup C1d2, and 23 samples from haplogroup D4h3a.

Each of these samples is identifiable via a sample ID and the respective haplotype of said sample. In order to further ensure anonymity, only the haplotypes were used to identify the samples in the study. Each sample for each haplogroup tends to be unique in mutations, denoted by the Novel Variants, which are differences between the sample's sequence and the sequenced haplogroup. For example, mutation C3516T, which is typical for A2al, describes a transition at position 35116 in the mtDNA genome, from a cytosine to a thymine, and 16519T, which is very common for haplogroup A2al, describes an insertion of a thymine at position 16519.

Using the software Network, designed and made free to access by Fluxus Technology, samples and their mutations were able to be plotted in a comprehensive phylogenetic tree, with four trees being constructed for each of the four model haplogroups. These trees not only depict the relationship and ancestry of samples in the same haplogroup, and even samples in different haplogroups, but can also give an estimate of time during which the haplogroup gave rise to the variants found in the tree. For the sake of the study, we assumed an average mutation rate of one mutations every 3600 years, in accordance with previous studies.

Haplotype	Haplogroup	Novel Variants
M8036650	A2al	16519T 9926G
M1509327	A2al	1530R 16519T
M2153603	A2al	16519T 6782Y
M7456588	A2al	16519T 9926G

Table 1. An example of how each sample for each haplogroup was organized, and their mutations read as variants.

Results

Using the Time Estimates feature on Network once the trees had been constructed, and assuming an average mutation rate of one mutation every 3600 years, we approximated the age of each haplogroup (see **Table 2**).

And in combination with the migration database on FamilyTreeDNA, we were able to conclude, based on the data available, how long it would have taken each haplogroup to have migrated to their present locations in Colombia from their point of origin.

Haplogroup A2al, which differentiated from haplogroup A in present-day Alaska, would have taken around 13,000 years to migrate down to South America.

Haplogroup B2d14, which differentiated from haplogroup B in present-day Baranof Island, would have taken around 7,000 years to migrate down.

Haplogroup C1d2, which differentiated from haplogroup C in the present-day Amur Oblast of Russia, would have taken around 11,000 years to migrate across to the Americas and down.

And haplogroup D4h3a, which differentiated from haplogroup D in the present-day Northwestern Heilongjiang Province of China, would have taken around 21,000 years to migrate across and down.

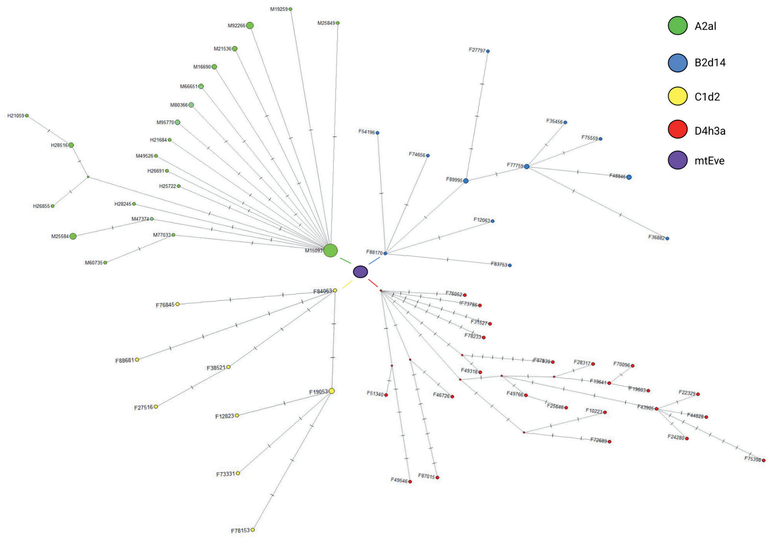

Figure 2. The composite phylogenetic tree constructed by analyzing the samples of each haplogroup, and combining each individual tree together, joined at the oldest maternal ancestor. The center represents the hypothetical mitochondrial Eve, from which each maternal lineage, including haplogroups A, B, C, and D, first diverged.

Haplogroup	Age in Years	std dev (Years)
A2al	13355	± 5037
B2d14	7200	± 3073
C1d2	10800	± 3138
D4h3a	21287	± 3576

Table 2. Each haplogroup's age up until the differentiation of modern variants in Colombia, with some ages more precise than others.

Discussion

This study serves as an example as to how mtDNA data can be utilized to reconstruct the history of early peopling in different regions of the world, based on analysis of present-day inhabitants and their mtDNA lineages, and ancient DNA isolated from remains. While we were limited to only 94 samples total for the chosen haplogroups, further studies can and should expand upon our limited data in order to create a more accurate and representative timeline.

In our study, we were somewhat limited in the necessary workforce to fully derive the conclusions we intended. While Network as a software is extremely potent in its ability to process numerous samples and mutations, it is still an older piece of software that, as opposed to other spreadsheet software (i.e., Microsoft Excel) requires much more manual input, and as such is significantly more time consuming to operate. While 94 samples are hardly representative in nature, it still cost numerous hours just to input their data into Network, thus limiting not only our time for continuing the study, but for increasing the sample size as desired. It is also important to mention that there have likely been numerous migrations from North America to South America across different lineages and time periods, and that examining haplogroups of different ages allows us to better piece together the true nature of these different waves of migrations, whereas we have only just scratched the surface.

Looking at the phylogenetic trees for each haplogroup (see **Figure 2**), less-centralized haplogroups with much more variation like D4h3a tend to be older, while haplogroups that display a clear center and less bifurcations like A2al and the others tend to be younger by comparison. This is a common trend in other mtDNA phylogenetic trees, and corroborates the data observed in **Table 2**.

For the sake of the research, certain samples and mutations were excluded from the study. Samples that were acquired from 'Haplocall' studies instead of other 'Tree' studies were excluded, for the sake of consistency, and mutations 16519T, 64T, and 93G from haplogroup A2al, and C166111 and T152C in haplogroup B2d14 were ignored, due to their lack of precise plotting in their respective phylogenies.

Finally, moving forward, we are hoping to compile and analyze the presence of these haplogroups in other countries beyond Colombia, in order to create a more holistic representation for the migratory patterns of these maternal lineages in other parts of the world.

Figure 3. A map of North and Central America depicting the presumed migration patterns of the analyzed haplogroups, with the divergence of A2al and B2d14 from A and B respectively being visible on the map, signified by hollow circles. Unlike the former two, haplogroups C1d2 and D4h3a diverged in East Asia before these lineages reached the Americas, hence why their migration stems from beyond Alaska. In addition, haplogroup C1d2's approximate migration in the FamilyTreeDNA database ends in Mexico, even though modern-day samples are found not uncommonly in Colombia. Thus, the rest of C1d2's migration pattern was extrapolated based on its last known location in the database, with the cross marking where the known migration data ends.

Conclusion

From our study, we managed to construct a unified phylogenetic tree and a migration pattern for four unique Amerindian haplogroups, A2al, B2d14, C1d2, and D4h3a. We are looking forward to expanding on this research in the future, as exploring not only other samples, but the frequency of these haplogroups in regions of Central and South America will allow us to determine more directionality in early migration patterns. And while our current sample size of 94 is relatively small, it sets a precedent for how we and other researchers can continue to fully apply the analysis of mtDNA in order to recreate the early settlements and peopling across history.

During the colonization of the Americas, indigenous populations were displaced, oppressed, and exterminated, with the Native American populations during the expansion of the United States struggling to integrate themselves into the new English-American status quo. Similarly, indigenous populations in Central and South America were also uprooted to varying degrees. In Central and Northern South America, people of Amerindian descent were either killed, or absorbed into the Spanish and Portuguese populations, living on in their ethnic descendants, but with their history completely obfuscated. Very few purely-indigenous populations remain to this day, and are slowly dwindling. In Southern South America, settlers of present-day Argentina, Chile, and Uruguay were known to be even more ruthless in their massacres of indigenous populations, leaving little room even for ethnographic integration. Across the Americas, there is very little indigenous history before colonization, at least compared to the societies that once blossomed when left to their own devices.

In regards to our study, mtDNA analysis poses an invaluable avenue by which to study the forgotten history of early Native American migrations and geographical diversity. It is especially important to research the forgotten history of America's early indigenous populations, as well as other marginalized groups across history, as we are piecing back together the lineages and stories of populations that still exist and flourish today.

References

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet. 2009 Jun;84(6):740-59. doi: 10.1016/j.ajhg.2009.05.001. Epub 2009 Jun 4. PMID: 19500773; PMCID: PMC2694979.
Yunis JJ, Yunis EJ. Mitochondrial DNA (mtDNA) haplogroups in 1526 unrelated individuals from 11 Departments of Colombia. Genet Mol Biol. 2013 Sep;36(3):329-35. doi: 10.1590/S1415-47572013000300005. Epub 2013 Aug 30. PMID: 24130438; PMCID: PMC3795164.

Luis Roniger (1997) Human Rights Violations and the Reshaping of Collective Identities in Argentina, Chile and Uruguay, Social Identities, 3:2, 221-246, DOI: 10.1080/13504639752078

Uricoechea Patiño, D., Collins, A., García, O. J. R., Santos Vecino, G., Cuenca, J. V. R., Bernal, J. E., Benavides Benítez, E., Vergara Muñoz, S., & Briceño Balcázar, I. (2023). High Mitochondrial Haplotype Diversity Found in Three Pre-Hispanic Groups from Colombia. Genes, 14(10), 1853. https://doi.org/10.3390/genes14101853

Ribeiro, B. P. A. (2021). Genetic characterization of the maternal lineages in andean colombian populations (Order No. 29140543). Available from ProQuest Dissertations & Theses Global. (2689299298). Retrieved from https://www.proquest.com/dissertations-theses/genetic-characterization-maternal-lineages-andean/docview/2689299298/se-2

Uricoechea Patiño, D., Collins, A., Romero García, O. J., Santos Vecino, G., Aristizábal Espinosa, P., Bernal Villegas, J. E., Benavides Benitez, E., Vergara Muñoz, S., & Briceño Balcázar, I. (2023). Unraveling the Genetic Threads of History: mtDNA HVS-I Analysis Reveals the Ancient Past of the Aburra Valley. *Genes*, *14*(11), 2036. https://doi.org/10.3390/genes14112036

Adriana Castillo, Fernando Rondón, Gerardo Mantilla, Leonor Gusmão, Filipa Simão, Maternal ancestry and lineages diversity of the Santander population from Colombia, *Forensic Sciences Research*, Volume 8, Issue 3, September 2023, Pages 241–248, https://doi.org/10.1093/fsr/owad032

de Saint Pierre M, Bravi CM, Motti JMB, Fuku N, Tanaka M, et al. (2012) An Alternative Model for the Early Peopling of Southern South America Revealed by Analyses of Three Mitochondrial DNA Haplogroups. PLOS ONE 7(9): e43486. https://doi.org/10.1371/journal.pone.0043486

Melton, P. E. (2008). Genetic history and pre -columbian diaspora of chibchan speaking populations: Molecular genetic evidence (Order No. 3349820). Available from ProQuest Dissertations & Theses Global. (304617164). Retrieved from https://www.proquest.com/dissertations-theses/genetic-history-pre-columbian-diaspora-chibchan/docview/304617164/se-2

Díaz-Matallana, Marcela, Gómez, Alberto, Briceño, Ignacio, & Rodríguez, José Vicente. (2016). Genetic analysis of Paleo-Colombians from Nemocón, Cundinamarca provides insights on the early peopling of northwestern South America. Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales, 40(156), 461-483. https://doi.org/10.18257/raccefyn.328

Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, et al. (2015) Genomic Insights into the Ancestry and Demographic History of South America. PLOS Genetics 11(12): e1005602. https://doi.org/10.1371/journal.pgen.1005602