ESCI upf. School of International Studies

Bachelor's Degree in Bioinformatics (UPF-UPC-UB)

Final Grade Project

Unraveling the maternal history of Brazilians through the analysis of mitochondrial DNA.

Pol Xavier Pérez Rodríguez

Scientific director: David Comas

Department of Medicine and Life Sciences - UPF, Human Genome Diversity lab. Institute of Evolutionary Biology - CSIC UPF

Abstract

Motivation: Mitochondrial DNA (mtDNA) is widely used in human population genetic studies due to its high variability, high copy number per cell, it is not affected by recombination, it is maternally inherited, and has a refined phylogeography. In this study, we use ancient and present-day mitochondrial samples of Native American lineages from different populations of South America to study the population history of Brazilians.

Results: This paper obtained a total of 1706 mtDNA sequences composed of 13 different subpopulations and 3 general populations in South America. A descriptive analysis of the data was obtained and different population genetic approaches were performed, such as haplotype and haplogroup sharing analyses, haplogroup composition for each population, differentiation between populations based on nucleotide diversity, and haplogroup networks. Our results provide a time-depth description of the maternal lineage composition of the Brazilian population as well as an overview of the relationships between these populations.

1. Introduction

MtDNA has been extremely informative in reconstructing the population history of humans, and over the past decades, it has been widely used in human evolution and population genetic studies due to its properties. It has a limited number of bases (~16.5kb), which makes it easy to analyze; it has high variability due to its higher mutation rates; and high copy number per cell, which makes it easy to sequence. Also, the mtDNA is not affected by recombination and it is maternally inherited; it allows studying genealogical relationships among individuals and exploring the frequency differences of matrilineal clades among human populations through a refined phylogeny.[1] Human mtDNA is a circular double-stranded DNA molecule characterized by high gene density and asymmetry. The two strands are differentiated into heavy (H-strand) and light strands (L-strand). The H-strand contains a higher percentage of guanine bases, while the L-strand has a higher percentage of cytosine bases. The mitochondrial genome in humans contains around 16,569 base pairs that code for 37 genes, 13 of them codify proteins and the rest rRNA and tRNA[2]. The origin of replication of both strands is found in a coding-free region named control region. Inside the control region of the mtDNA, there are three hypervariable (HV) segments (sites with a much faster evolution rate), the HV1, which stands between positions 16024 and 16383 (according to the nomenclature of the mtDNA reference); the HV2, between positions 57 and 372; and the HV3, between 438 and 574, they are indeed mutational hotspots[3].

MtDNA haplogroups (i.e. the combination of variants that define a group of mtDNA lineages) are widely studied and are pivotal in the reconstruction of past population events since they are the product of the common maternal ancestry of the individuals and populations carrying them. Also, the whole mtDNA genome can give us enough molecular resolution to distinguish patterns that have arisen over thousands of years due to mutations, revealing regional patterns of population variation, past human migrations, periods of intermixing, and details of colonization processes.

The nomenclature of mtDNA haplogroups was first introduced in the mid-1990s to describe the different branches or lineages of mitochondrial DNA. This nomenclature system uses a series of letters and numbers to represent the different haplogroups and subhaplogroups, which are based on the specific mutations and variations observed in mitochondrial DNA sequences. The haplogroups A-G were initially assigned to variation observed in Asian and American lineages, H-K were assigned to Europe, whereas only a single letter, L, was assigned to describe the highest level of variation observed in Africa[1]. The current nomenclature used has a robust branch structure which has been expanded based on new research.

Some mtDNA analyses have provided a high-resolution structure of the peopling of the Americas. A study with 92 whole mitochondrial genomes from pre-Columbian South American skeletons dating from 8.6 to 0.5 kya (thousand years ago) suggests that a small population entered the Americas via a coastal route around 16 kya, after previous isolation in eastern Beringia for ~2.4 to 9 thousand years after separation from eastern Siberian populations[4]. This climatic evidence is consistent with a mitogenomic tree showing a sudden burst of lineage diversification starting ~16 to 13kya, followed by an increase in the mean female effective population size and the retreat of coastal glaciers along the northwest Pacific coast associated with a phase of stepwise ocean warming which indicates the entry into the Americas took place via a southward expansion.

The sites of Monte Verde in Chile and Pedra Furada in Brazil have been central to the discussion of early human presence in South America, with evidence implying that peopling in South America was extremely fast, taking into account when the first humans arrived in America and the first evidence of human presence in South America is from ~14.5 kya.

The main haplogroups from Native American populations are labeled as A, B, C and D and they exhibit a low genetic variability. Studies have been able to perform a deeper phylogenetic dissection of these four haplogroups which the most relevant are the following ones (A2, B2, C1b, C1c, C1d, C1d1, D1, and D4h3a) which are classified as pan-American because they are found across all America. It has been also described the presence of subhaplogroups originated in situ after a few millennia from the initial peopling, such as B2b, which is shared between North and South America but its origin is best explained in North America; B2a, which is just present in Central and North America and originated in North America; and the subhaplogroups D1g, D1j, C1b13, and B2i2 which are present in South America, originated in the Southern Cone[5]. The shape and geography of the South American continent caused this continent to be inhabited following two main routes, one along the Pacific coastline and the other along the Atlantic one. This phenomenon made the region south of the Amazonas basin the meeting point between these two principal peopling events. This region plays an important role in understanding the demographic processes during the peopling that happened in the Southern Cone[6].

After subsequent lineage diversification within each haplogroup and limited gene flow between populations, the European colonization caused a dramatic bottleneck in Native American genetic diversity due to multiple factors such as war and foreign diseases from the Old World, and many ancient lineages are extinct in contemporary indigenous populations. Nowadays, we found higher Native American mtDNA genetic diversity in admixed populations than in indigenous populations, meaning that admixed populations represent an important genetic reservoir of Native American lineages[7].

Objectives

Using complete mtDNA sequences from different populations of South America, we aim to clarify the genetic matrilineal origin of present-day Brazilians as well as the relationships between present-day populations and ancient samples. This requires an understanding of their past relationships as well as identifying their lineages. To achieve this, we:

- Analyze the haplogroup composition for each of the populations.
- Assess the matrilineal continuity between ancient and present-day populations.
- Explore the relationship between present-day populations and ancient populations with other present-day populations.
- Infer past demographic and historical events from present-day data.

2. Methods

2.1. Sample collection and data preprocessing

MtDNA complete sequence data was previously obtained from native, admix and ancient populations from Brazil, Argentina and surrounding areas. Since the sample size for each population was very small, we grouped the samples into language families. The dataset analyzed consisted exclusively of Native American lineages, i.e. belonging to superhaplogroups A, B, C and D. It also contains samples from ancient Argentinian and ancient Brazilian populations, retrieved from ancient bones and previously sequenced in our group. All samples used in the present analysis were previously filtered after

adequate quality controls, resulting in a reliable dataset for population genetics analysis. This initial classification was performed by our collaborators and it was taken as the preliminary dataset for our project.

2.2. Validation of the dataset and its manipulation

All sequences were formatted in a multifasta file. The haplogroup information for each individual was obtained using HAPLOGREP 2[8]. The aim was to validate if the multifasta was congruent with the previous haplogroup classification provided by our collaborators and correctly aligned. The results of the validation indicated that there were some individuals with discordant haplogroup classifications. In order to overcome these discrepancies, we applied the diff command in Linux, which is used to display differences between files and then these were corrected manually. Also, in some cases, the same sequence belonged to a different population between both files, which is obviously incongruent, and original data was provided to fix the excel file manually. In order to perform the statistical analysis per population, we classified the fasta files according to their corresponding population.

2.3. Statistical descriptive analysis of the data

We created a multifasta file for each population and in order to make use of it, it was necessary to align each of them. The alignments were performed using the online version of MAFFT[9]. Each statistic was calculated separately for each population, including the number of sequences, the number of haplogroups, and haplogroup diversity. For the number of haplotypes and haplotype diversity using the function *haplotype* from R package pegas[10] and gaps and N values in the sequences were ignored. Using the function *nuc.div* from the same package mentioned before we computed the nucleotide diversity.

2.4. Haplogroup composition of each population

Creating a histogram for each population using the haplogroup classification from PhyloTree (HAPLOGREP 2[8]), we analyzed its composition for each population. Counts for each haplogroup were calculated to observe which haplogroup is more common in each population or which haplogroup is present in one population and not in another. For the creation of the plots, we used the R package *ggplot2*[11], and the specification of the haplogroup classification was reduced to visualize a clearer representation.

2.5. Differentiation between populations

We computed a matrix of distances based on the pairwise differences between all the populations using the R function *pairwise_Gst_Nei* from package *mmod*[12], we calculated the average nucleotide diversity between populations. Then we used it to create a multidimensional scaling (MDS), and to visualize it, we used *ggplot*[11] and *ggrepel*[13]. To visualize the matrix of distances, we plotted a heatmap with the function *heatmapSpp* from the R package *spider*[14].

2.6. Haplogroup sharing

Similar to haplogroup composition, but in this scenario, we put into practice the assumption of having a donor population and a receptor one. Basically, what we did was to extract the haplogroups present in Native and Ancient American populations and count their abundance in present day admixed populations for each possible case. The objective is to quantify the genetic impact of the donor populations with respect to the receptor ones. We represented this measurement in plots using the *ggplot* package from R, taking the excel provided as input and the haplogroup classification from PhyloTree (HAPLOGREP 2[8]).

2.7. Haplotype sharing

The original dataset was subjected to data processing using the MEGA[15] software, with the aim of excluding mutational hotspots, poly-c track regions, and all the positions with N values. The presence of hypervariable regions poses a significant drawback in haplotype sharing analysis, as these specific positions in the mitochondrial DNA exhibit a high mutation frequency and low phylogenetic information. Subsequently, the output obtained from MEGA[15] was converted into an arp file format using the PGDSpider[16] software, enabling its utilization in Arlequin[17]. Arlequin is a widely employed software package utilized for the computation of population genetics statistics. Ultimately, the desired outcomes were achieved, wherein two sequences were considered to belong to the same haplotype only if they were entirely identical.

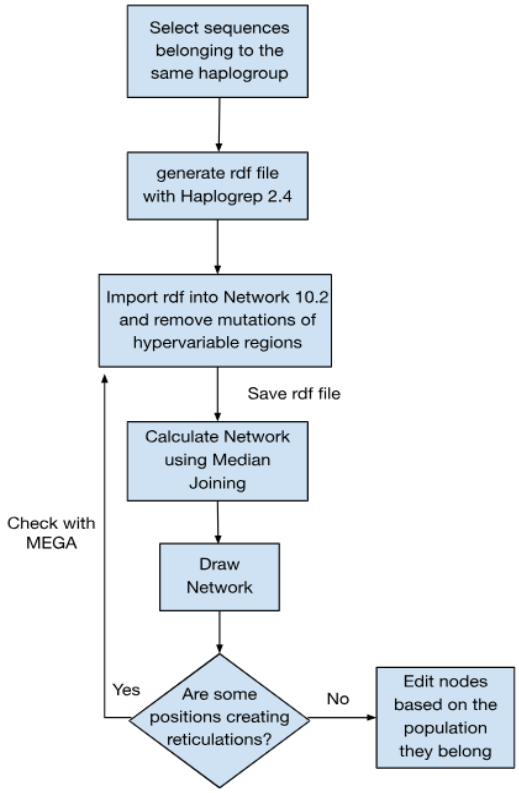
2.8. Haplogroup networks

The use of ancient sequences in our analysis proves particularly intriguing in our present investigation. This choice stems from our objective of assessing the continuity of haplotypes between ancient and contemporary populations. Consequently, we isolated most ancient sequences and determined their respective haplogroups. Subsequently, we gathered all sequences associated with these haplogroups from our dataset. To increase the relevance of our findings, we expanded each network by incorporating additional sequences using Blast[18]. Our aim was to construct distinct networks for each haplogroup. **Figure 1** illustrates the workflow for haplogroup network construction.

To adapt our data for network visualization, we employed HaploGrep 2.4[19] to modify its format, enabling compatibility with the Network 10.2[20] software, our chosen tool for network depiction. To ensure the accuracy of the networks, we took the mutations that deviate from the human mitochondrial reference genome[21] mutational hotspots in human mitochondrial DNA and we eliminated the ones corresponding to hypervariable positions. It is imperative to exclude these positions as they can lead to misleading interpretations when constructing the network since they are not phylogenetically informative. To identify such positions, we relied on a study containing annotated mutational hotspots[22] and utilized MEGA[15] for visualizing poly-c track regions.

Once identified, we computed the network using the Median Joining method[23] and subsequently generated its graphical representation. Given that the removal of hypervariable positions is a manual process, it is crucial to thoroughly examine the resulting network for any reticulations or overlapping connections. In the event of reticulations, we scrutinize the positions that may be responsible for these occurrences and selectively eliminate those associated with hypervariable positions. For our purposes, nodes representing individual sequences were color-coded to correspond with their respective populations.

Figure 1. Workflow of reconstructing a haplogroup network. It is important to note that there are other softwares to perform the process.



2.9. Code and Data Availability

The main data and scripts used for this project are available on **GitHub**.

3. Results and Discussion

3.1 Validation of the dataset

Our final obtained data is composed of a total of 1706 sequences with a total of 13 different subpopulations and 3 general populations. Each sequence identifier is composed of its original id followed by the subpopulation they belong to. Correct haplogroup classification was assigned to all the sequences. Subpopulations Arawak, EasternTukanoan, Ecuador, Guahiban, Je, MakuPuinave, Native2 (also called “rest of natives grouped”) were classified in a single group since they present small sample sizes; Tupi, and WesternTukanoan are engulfed inside Native population. Subpopulations admix.arg (Admixed Argentina) and admix.Bra (Admixed Brazil) represent the admixed population samples. Finally, subpopulations ancient.arg (Ancient Argentina) and ancient.Bra (Ancient Brazil) are the data obtained from ancient populations. Note that, depending on the analysis we perform, it is more useful to use subpopulations or populations. In the next sections, we will refer to these subpopulations as populations as this consideration is just arbitrary and it was just useful for this section to make clear how the data has been manipulated.

3.2 Statistical descriptive analysis of the data

Table 1 summarizes all the statistical descriptive measures applied to our data. We observe that all populations and subpopulations share a similar π (nucleotide diversity). The most contrasting result from this analysis is looking at the aDNA (ancient DNA) population, which has one of the lowest nucleotide diversity values but it has a relatively high haplogroup and haplotype diversity. This can be explained by the fact that probably a lot of their sequences are different but the differences are minimal. We can also see a pattern which is that the higher the sample size of a population, the haplotype and haplogroup diversity result to be lower. It is a matter of fact that the higher the number of sequences, the higher the probability of having a larger number of haplotypes, so it is exactly what we observe, being admix and admix.Bra the populations with more haplotypes. Seems to be that the most informative statistics are the haplogroup diversity and haplotype diversity as they help us compare different populations even though they have very different sample sizes.

Tupi and Je populations result from having a relatively high haplotype diversity compared to the other native

Table 1. Mitochondrial DNA diversity parameters of the populations analyzed

Population	# of sequences	H (# of haplogroups)	Number of haplotypes	Haplotype diversity	Haplogroup diversity	π
admixed	804	60	643	0.8	0.075	0.00219
aDNA	77	18	74	0.961	0.234	0.00152
Native	825	64	623	0,755	0.078	0.00209
admix.arg	93	25	85	0.914	0.269	0.00208
admix.Bra	711	47	558	0.785	0.066	0.00219
ancient.arg	39	11	37	0.949	0.282	0.00187
ancient.Bra	38	12	37	0.974	0.316	0.00145
Arawak	114	21	74	0.649	0.184	0.00216
EasternTukanoan	66	17	51	0.773	0.258	0.00217
Ecuador	208	24	187	0.899	0.115	0.00203
Guahiban	51	4	26	0.51	0.078	0.00193
Je	32	11	28	0.875	0.344	0.0021
MakuPuinave	35	7	16	0.457	0.2	0.00187
Native2	222	46	184	0.829	0.207	0.00217
Tupi	61	21	54	0.885	0.344	0.00212
WesternTukanoan	36	10	22	0.611	0.278	0.00222

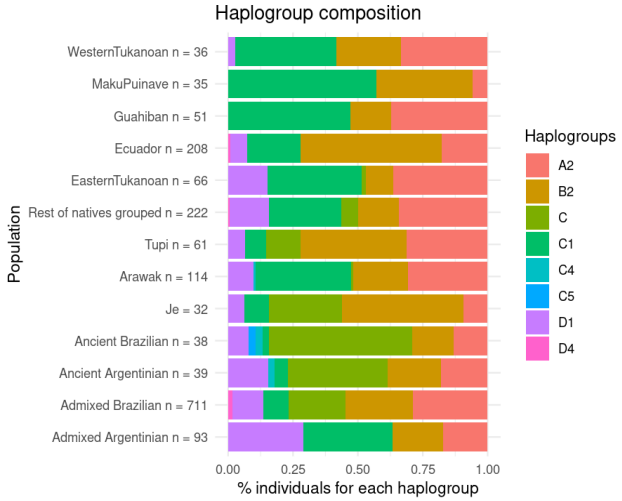
american groups which could be due to the fact that these groups were widely spread around the South American continent. It is found that Admixed Argentina has a very similar haplotype and haplogroup diversity to Ancient Argentina. But this is not the same when comparing Admixed Brazil and Ancient Brazil, this last one having significantly more diversity. The Native group has a similar number of sequences to the Admixed one and they have a similar number of haplogroups, but Admixed has more haplotype diversity, and this must be taken into account in the following sections for further conclusions.

3.3 Haplogroup composition of each population

From **Figure 2** we can see that each population contains a lot of different lineages, which makes it difficult to extract any conclusion at first sight. The frequencies of such lineages vary in each population. Haplogroups C4 is found in Ancient Argentina and Ancient Brazil, and C5 in Ancient Brazil and they are found in low frequencies, being almost nonexistent in present day admixed and native populations, just Arawak having few C4 sequences. With this information, we can discuss the possibility of lineage discontinuity for these two haplogroups. Also, ancient populations have the highest frequencies of haplogroup C, followed by Je and Admixed Brazilian. This might be a sign of lineage continuity between ancient populations, Admixed Brazil and Je but also we should take into account that inside haplogroup C there might be a lot of diversity as these sequences received a very general classification, so it does not necessarily have to imply that. This haplogroup is also present in Tupi. Haplogroup D4 is only present in Ancient Brazil, in Rest of natives grouped, in Admixed Brazil and Ecuador and in all of them in very low

frequencies. Haplogroup C1 is found in a large percentage of samples in Admixed Argentina, as well as in the vast majority of native populations.

Figure 2. MtDNA haplogroup composition by population



Haplogroup A2 and B2 are found in all populations and they do not give us very much information using this kind of approach. The same happens with haplogroup D1, present in almost all populations in not very high frequencies.

To make a more general analysis it is recommended to observe *Supplementary Figure 1*, here we can see the differences between ancient, natives and admixed populations, and what we extract from this, is a clearer difference in the frequency of haplogroup C between ancient and Native populations. In contrast, haplogroup C1 is very common in Native populations but found in very low frequencies in ancient populations. Admixed populations do not present a significant difference between these two haplogroups.

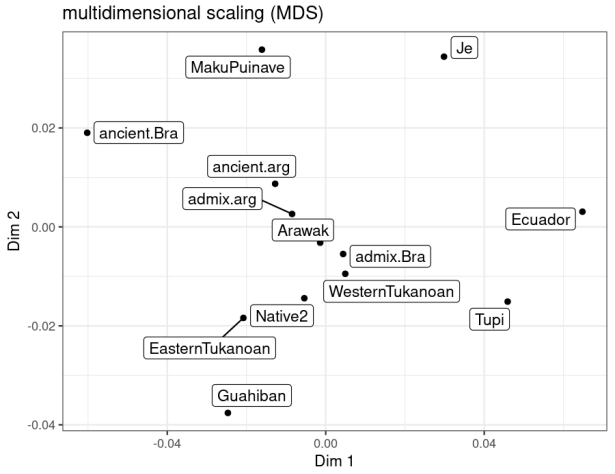
Supplementary Figures 2-3 are provided to observe how the real numbers of the data look without the normalization process being applied. So basically, it is useful to make clear that maybe, in some cases, there could be a little bit of sample bias, even though in the normalized figures commented before, the sample size of each population was indicated.

3.4 Differentiation between populations

In the MDS plot (see **Figure 3**) we observe the high similarity between Ancient Argentina, Admixed Argentina, and Arawak. Ancient Brazil and Admixed Brazil do not cluster very closely, being Admixed Brazil closer Admixed and Ancient Argentina. Je, MakuPuinave, and Ecuador do not cluster with any population, being very distant from each other. Admixed Brazil cluster with WesternTukanoan.

In *Supplementary Figure 4* we see a heatmap of distances and we can observe the differentiation values between populations. We see that Ancient Brazil and Ecuador have a high differentiation value, which is why they are so distant in the MDS plot. It is totally the opposite for Admixed Argentinian and Ancient Argentinian, clustering together due to their low value in the matrix of distances as represented in this figure.

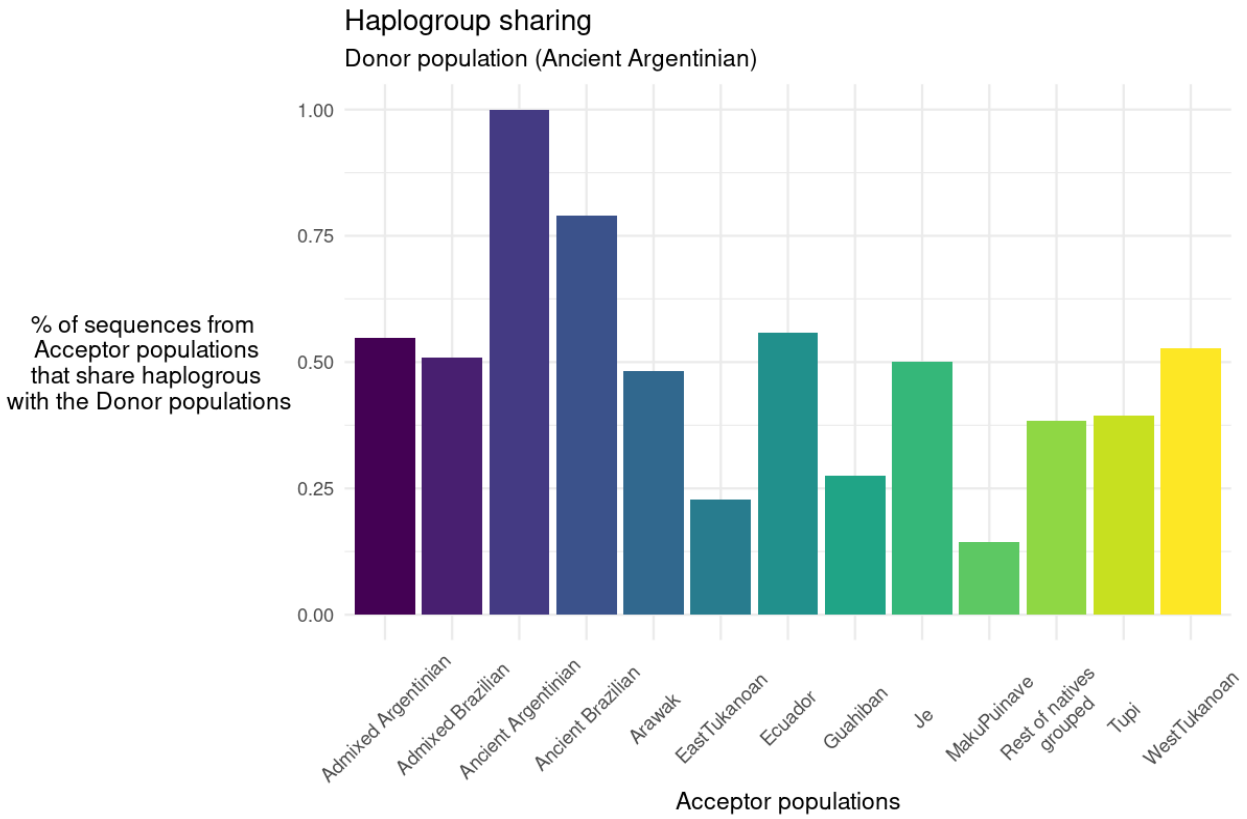
Figure 3. Multidimensional scaling (MDS) plot based on genetic distances



3.5 Haplogroup sharing

As a reminder, the haplogroup composition was computed taking into account a directional relationship between populations, having a Donor population that donates haplogroups to the rest of the populations. We consider that Donor populations are just all the ones belonging to native and ancient populations, and we performed this measurement for each of these populations. It consists in extracting all the haplogroups present in the Donor population and counting how many of them are present in each of the rest of the populations, and then dividing it by the sample size of each population. From this, we will obtain the percentage of sequences that share haplogroup with the Donor population for each Acceptor population. It is surprising that in the case of Ancient Argentina as Donor (see **Figure 4**), the second population that shares more percentage of sequences from haplogroups present in the Ancient Argentina population are the Admixed Argentina, and with the distance matrix it results that these two populations are the most similar between them (as seen in section **3.4**), which at first sight it does not seem very congruent. In the rest of *Supplementary*

Figure 4. MtDNA haplogroup sharing sequences between populations



Figures 5-12 it is also interesting to see that it seems that Je haplogroups are very present in these two ancient populations, being Admixed Argentinian the total opposite. In the case of Tupi as a donor, Guahiban shares the highest amount and Ecuador the lower amount, and they share a similar amount of haplogroups with the ancient populations. WesternTukanoan does not share a large amount of sequences with other populations. Je shares a high percentage of haplogroups with Ancient Argentina and Ancient Brazil. This last one as a Donor shares a high percentage with Ancient Argentina, Je, and Admixed Brazil.

3.6 Haplotype sharing

Ancient DNA contains a lot of missing values, which makes comparisons challenging; therefore, after dealing with this problem, previously explained in section **2.7**, we obtained *Supplementary File 1*. Focusing on haplotypes shared between ancient and modern populations we just found five sequences shared between Ancient Argentina and Admixed Brazil, all of them belonging to haplogroup C. Ancient Brazil does not share any identical sequence with other populations. This is interesting because it is congruent with section **3.3** when talking about lineage continuity of this haplogroup between ancient populations and Admixed Brazil but we note the fact that C haplogroup can be very diverse so it does not necessarily imply lineage continuity.

Nonetheless, with this new result, we can affirm that there is probably lineage continuity between Ancient Argentina and Admixed Brazil. However, finding only five identical sequences between these two populations is not enough evidence, since we do not know the amount of similarity between the rest of the sequences belonging to haplogroup C are between them. To see how similar sequences are between these two populations, we have to look at section **3.4** and, as explained, it is extremely low. This is just more evidence to ensure that, indeed, lineage continuity between these populations exists.

In the case of Admixed Argentina, we did not find shared haplotypes with other populations, which is surprising but it is probably due to its low sample size and that if just one position between sequences differ they are not counted as shared haplotypes. In contrast, due to the high number of sequences that our data has of Admixed Brazilians, we can find that it shares haplotypes with Tupi, Rest of natives grouped,

Arawak, Je, EasternTukanoan, Ecuador, and of course, Ancient Argentina as commented before.

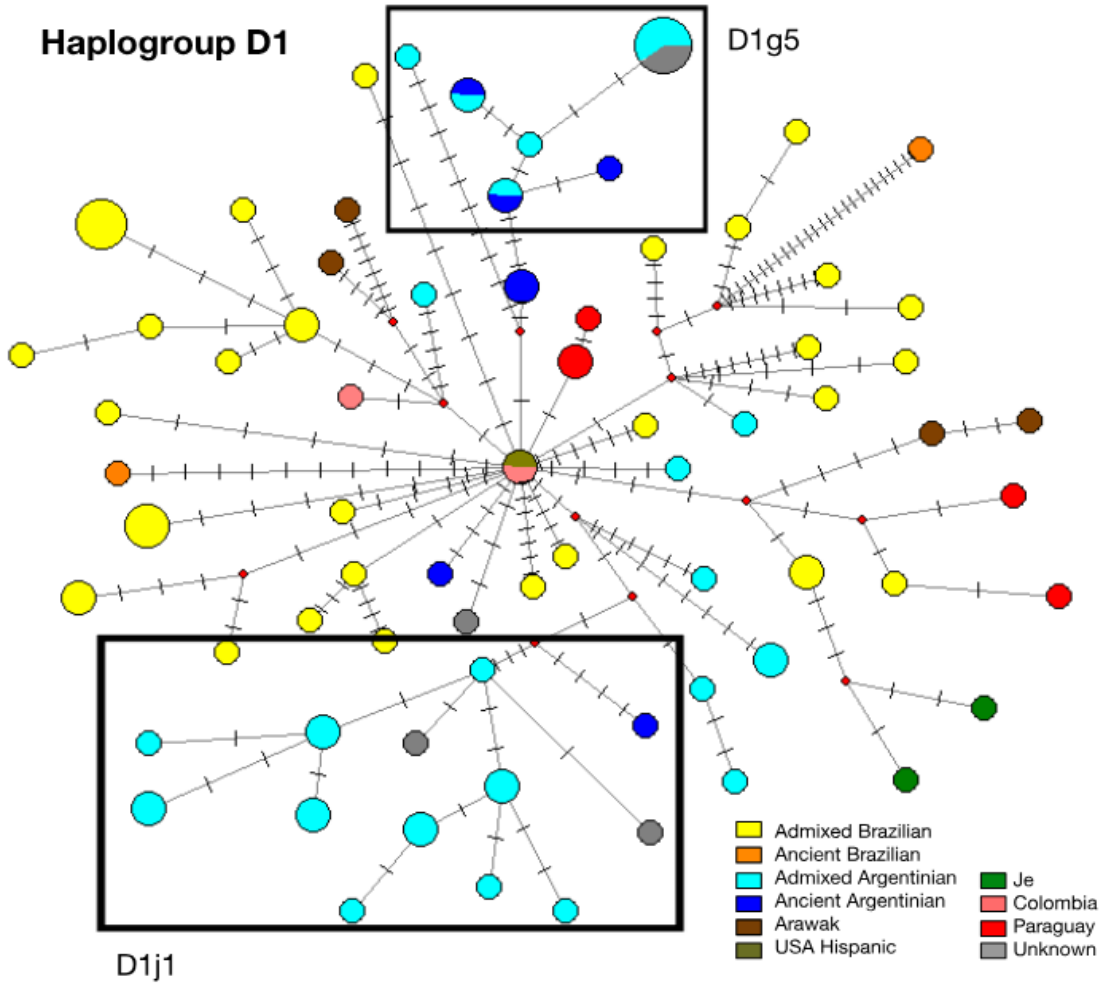
3.7 Haplogroup networks

A total of five networks were reconstructed, each of them belonging to a different haplogroup. In this section, we will discuss the five of them separately. Note that lines represent mutations in a specific position and it is important to understand that ancient samples usually have more because of missing values interpreted as deletions.

Starting with **Figure 5**, we could conclude that, in this case, there is a very clear lineage continuity between Ancient Argentina and Admixed Argentina in two ramifications of the graph. It is interesting to note that these two ramifications belong to specific subhaplogroups which appear to be specific for Ancient and Admixed Argentinian populations. Admixed Brazil is all spread around the network, however, we can observe that there is a connection between the Ancient Brazilian sequences and some other Admixed Brazilian ones, suggesting a possible continuity between the ancestral sequence and the others that are in the ramification.

Supplementary Figure 13 corresponds to the haplogroup C4c. We obtained two different ramifications belonging to two different subhaplogroups, C4b which correspond to sequences from Russia (some of them ancient), and C4c1 which all their sequences belong to North America, being all of them except for three, ancient samples. All these sequences were obtained from doing Blast. And finally, we have three more sequences belonging to our original data, one Ancient Brazilian, another Ancient Argentinian, and one Arawak all of them belonging to C4c. As a conclusion for this network, this haplogroup seems to be almost extinct in present-day populations from America, as, in our network, it is just present in three non-ancient

Figure 5. Network plot of haplogroup D1. The size of the circles represents the frequency of each sequence, different colors represent populations (see inserted legend), and the crosses in the branches represent mutations between sequences. Subhaplogroups D1g5 and D1j1 are highlighted in the network.



samples from North American individuals and one Arawak. The network for C1d1 is found in *Supplementary Figure 14* and it shows two clear cases with lineage continuity. In the case of Ancient Brazil with Admixed Brazil, Paraguay, and Admixed Argentina, and the other case with that Argentinian ramification, with ancient and modern samples. Samples from Mexico and USA can also be considered for such lineage continuity of the ancient Brazilians.

From *Supplementary Figure 15* we see that Haplogroup B2b3a is mainly composed of Admixed Brazilian sequences and there are no Argentinian populations involved. It is interesting to note that from Blast we retrieved an ancient Colombian sample and another ancient sample from Perú. This network shows how the continuity of Ancient Brazil is all spread around different latin american and modern native populations.

Haplogroup A2+(64) was just found in one Ancient Brazilian sample from our dataset. After doing Blast, we found an ancient sample from Canada and another one from Colombia. *Supplementary Figure 16* shows the network for this haplogroup. Interestingly, we did not find any Ancient Argentinian sample but we found samples from Admixed Argentina. This may imply that they received this haplogroup from other populations or we were not lucky enough to get samples from Ancient Argentina from this haplogroup.

4. Conclusions

Our results have shown that Ancient Argentina and Ancient Brazilian samples have been the major contributors of mtDNA lineages to the nowadays Admixed Brazilian population. Regarding the nowadays Admixed Argentina population, the Ancient Argentina samples are the main contributors to this group. We also observed that Natives have the lowest genetic diversity, being in ancient populations very high. It is not observed a clear relationship between the lineages of native and ancient populations but in some cases, we found a link between native and admix groups. Taking into account all our results, we conclude that modern admixed populations from Brazil and Argentina are, in most cases, an important repository for those lineages that belonged to those ancient populations that got extinct during the European colonization. In addition, this study corroborates that contemporary indigenous populations suffered a bottleneck event that made them lose a significant amount of their lineages. And finally, it seems that the Brazilian population is composed of a mix of ancient and contemporary native populations.

Acknowledgments

I would like to extend my sincerest appreciation to Julen Aizpurua Iraola for his invaluable guidance in the development of this process and his unwavering personal support, assisting me with all my needs.

Supplementary Material

The supplementary material can be found **here**. The files can be found in folders with the section titles.
File 1. It contains the results of shared haplotypes between populations in Arlequin format. **Figure 1.** Normalized haplogroup composition for each population. **Figure 2.** Haplogroup composition for each subpopulation. **Figure 3.** Haplogroup composition for each population. **Figure 4.** Heatmap represents the matrix of distances between populations. **Figure 5-12**. Each of these figures represents the percentage of shared haplogroups between a donor population and the rest. **Figure 13-16.** Haplogroup networks for each of the haplogroups of interest.

References

1. Kivisild, T. Maternal ancestry and population history from whole mitochondrial genomes. *Investig. Genet.* **6**, 3 (2015).
2. Garcia, I., Jones, E., Ramos, M., Innis-Whitehouse, W. & Gilkerson, R. The little big genome: the organization of mitochondrial DNA. *Front. Biosci. Landmark Ed.* **22**, 710–721 (2017).
3. Stoneking, M. Hypervariable Sites in the mtDNA Control Region Are Mutational Hotspots. *Am. J. Hum. Genet.* **67**, 1029–1032 (2000).
4. Llamas, B. *et al.* Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci. Adv.* **2**, e1501385 (2016).
5. Brandini, S. *et al.* The Paleo-Indian Entry into South America According to Mitogenomes. *Mol. Biol. Evol.* **35**, 299–311 (2018).
6. García, A. *et al.* Ancient and modern mitogenomes from Central Argentina: new insights into population continuity, temporal depth and migration in South America. *Hum. Mol. Genet.* **30**, 1200–1217 (2021).
7. Tavares, G. M., Reales, G., Bortolini, M. C. & Fagundes, N. J. R. Measuring the impact of

European colonization on Native American populations in Southern Brazil and Uruguay: Evidence from mtDNA. *Am. J. Hum. Biol.* **31**, e23243 (2019).
8. Weissensteiner, H. *et al.* HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58-63 (2016).
9. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
10. Paradis, E. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26**, 419–420 (2010).
11. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2009). doi:10.1007/978-0-387-98141-3.
12. Winter, D. J. MMOD: an R library for the calculation of population differentiation statistics. *Mol. Ecol. Resour.* **12**, 1158–1160 (2012).
13. Slowikowski, K. *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. (2022).
14. Brown, S. D. J. *et al.* Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Mol. Ecol. Resour.* **12**, 562–565 (2012).
15. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).
16. Lischer, H. E. L. & Excoffier, L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299 (2012).
17. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
19. Schönherr, S., Weissensteiner, H., Kronenberg, F. & Forer, L. Haplogrep 3 - an interactive haplogroup classification and analysis platform. *Nucleic Acids Res.* gkad284 (2023) doi:10.1093/nar/gkad284.
20. fluxus-engineering.com. https://www.fluxus-engineering.com/.
21. Homo sapiens mitochondrion, complete genome. (2023).
22. Soares, P. *et al.* Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. *Am. J. Hum. Genet.* **84**, 740–759 (2009).
23. Bandelt, H. J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).