Genetic characterization of the maternal lineages in Andean Colombian populations

Bibiana Patrícia Augusto Ribeiro

Genética Forense

Departamento de Biologia

2021

Orientador

Maria João Ribeiro, Professor Associado, Faculdade de Ciências da Universidade do Porto (FCUP)

Coorientador

Filipa Simão, PhD, Universidade do Estado do Rio de Janeiro (UERJ)

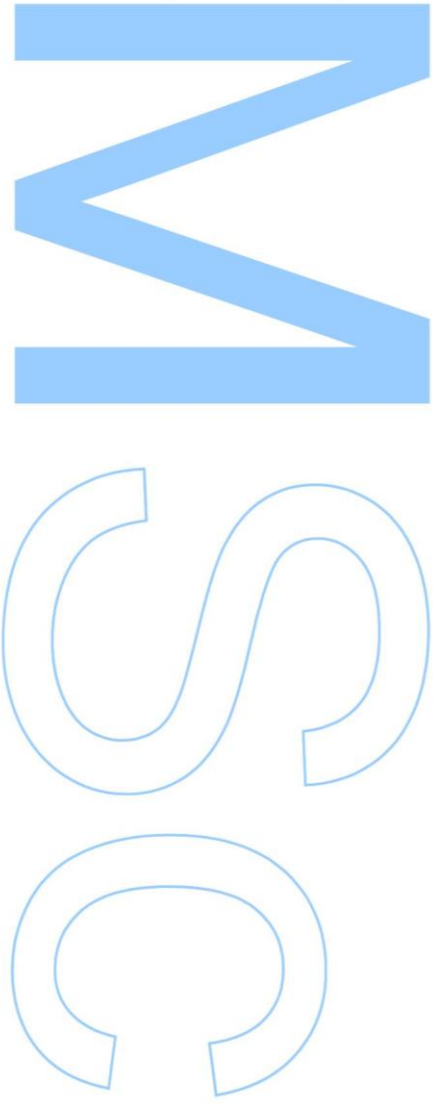
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



Agradecimentos

À Professora Doutora Maria João, minha orientadora, pela orientação exemplar pautada por um elevado e rigoroso nível científico, pelo interesse permanente e fecundo, pela visão crítica e oportuna, pelo empenho inexcedível e saudavelmente exigente, que contribuíram para enriquecer todas as etapas subjacentes ao trabalho realizado;

À Professora Doutora Filipa Simão, minha coorientadora, por todo o acompanhamento, acessibilidade e valiosa ajuda durante a realização deste trabalho;

À Professora Doutora Leonor Gusmão, porque me quis honrar com o seu apoio, e pela confiança que em mim depositou;

À Professora Doutora Verónica Gomes, pela prontidão, paciência e apoio nos procedimentos laboratoriais;

Ao Professor Doutor António Amorim, diretor de curso, que aprovou e permitiu a realização deste projeto, e também pela disponibilidade e simpatia;

A todos os elementos que constituíram o grupo de investigação Genética Populacional e Evolutiva do i3S durante este ano letivo, pela ajuda, boa disposição e por me terem proporcionado um bom ambiente de trabalho;

À minha família por todo o amor, apoio e sacrifícios feitos ao longo destes dois anos;

Muito obrigada.

Resumo

A genética populacional e evolutiva são disciplinas que têm contribuído muito para entender melhor a origem do homem moderno e a estrutura e relações entre populações humanas. Estudos sobre a composição genética de populações da América do Sul têm sido fundamentais não só para recuperar a história da colonização inicial deste continente pelo homem moderno, mas também para documentar a profunda remodelação que as populações nativo-americanas sofreram em consequência das recentes migrações ocorridas desde o início da era colonial. As populações atuais da Colômbia são muito heterogéneas devido à complexidade da história, geografia, demografia e cultura da região em que se encontram. Daí que sejam necessários estudos genéticos abrangendo a diversidade de populações da Colômbia para obter uma visão mais fina sobre os seus padrões de miscigenação. Tal é importante não só do ponto de vista da história das populações, mas também tem implicações importantes em áreas como, por exemplo, genética clínica e forense. Tendo isso em consideração, o objetivo principal deste trabalho foi investigar a diversidade genética materna em populações da Colômbia ainda por analisar, através do estudo do ADN mitocondrial.

Um conjunto de 53 amostras de sangue foram recolhidas sob consentimento informado de indivíduos não relacionados nascidos em dois departamentos da região Andina da Colômbia: Tolima e Cundinamarca. Sequenciou-se e analisou-se a região controlo – entre as posições 16024 e 576 – do ADN mitocondrial. Encontraram-se 43 haplótipos únicos e 5 que eram partilhados entre dois indivíduos. A maioria dos haplótipos encontrados pertencia a haplogrupos nativo-americanos, e os restantes eram haplogrupos euroasiáticos que constituíam, aproximadamente, 7,5% da amostra total. Esta última proporção indica que foi bastante modesta a incorporação de linhagens femininas de origem europeia que ocorreu após a redescoberta quinhentista da América e colonização pelos Europeus.

Foram detetadas diferenças genéticas estatisticamente significativas entre as amostras de Tolima e Cundinamarca. Em termos de linhagens maternas, as populações colombianas aqui estudadas apresentam elevada diversidade genética interna, e também se diferenciam bem entre si. Este é, aliás, um padrão que é comum encontrar em populações da América do Sul.

Análises comparativas integrando populações miscigenadas e nativas da Colômbia bem como de outras regiões da América do Sul, revelaram a ausência de correlação entre perfis genéticos e distribuição geográfica das populações.

Deste trabalho resultaram novos dados de ADNmt, que, no futuro, analisados mais detalhadamente e em conjunto com os já existentes quanto a outras populações americanas contemporâneas e os que começam a surgir através da análise de ADNmt antigo, poderão ajudar a clarificar como surgiu o complexo cenário populacional que caracteriza a América do Sul.

Palavras-chave: genética populacional; ADN mitocondrial; linhagens maternas; haplogrupos do ADNmt; haplótipos; ancestralidade nativo-americana; América do Sul; região Andina da Colômbia.

Abstract

Evolutionary and population genetics are disciplines that are giving extraordinary contributions to understand the origin of modern human populations, their structure, and relationships. The knowledge of the genetic background of South American populations is crucial to a better understanding of the colonization of this continent by modern humans, as well as to obtain insights into the deep remodeling that American populations went through in consequence of the more recent human migrations during and after the colonial Era. The present-day populations of Colombia are heterogeneous duo to the complex history, geography, demography, and culture of the people inhabiting the region. Thus, comprehensive genetic studies involving genetic diversity of Colombian populations are required in order to finely dissect their admixture patterns. This is important not only from the point of view of the history of populations, but it also has important implications in areas such as clinical genetics and forensics. Taking this into account, the main objective of this work was to investigate the maternal genetic diversity in Colombian populations yet to be analyzed, through the study of mitochondrial DNA.

A set of 53 blood samples were collected under informed consent from unrelated individuals born in the Colombian Andean departments of Tolima and Cundinamarca. The entire control region – between positions 16024 and 576 – of the mtDNA was sequenced and analyzed. A total of 43 CR haplotypes were unique and 5 were shared between two individuals. The majority of the haplotypes belonged to Native American haplogroups and the remaining were Eurasian haplogroups encompassing ~7.5% of the entire sample. This last proportion indicates that the incorporation of female lineages of European origin that occurred after the 16th century rediscovery of America and colonization by Europeans was quite modest.

Statistically significant genetic differences were detected between the samples from Tolima and Cundinamarca. In terms of maternal lineages, the Colombian populations studied here show high internal genetic diversity, and also differ well from each other. This is, in fact, a pattern that is common to find in South American populations.

Comparative analyzes integrating admixed and native populations from Colombia as well as from other regions of South America, revealed the absence of correlation between genetic profiles and geographic distribution of populations.

This work resulted in new mtDNA data, which, in the future, analyzed in more detail and in conjunction with those that already exist for other contemporary American populations and those that are beginning to emerge through the analysis of old mtDNA, may help to clarify how the complex population scenario that characterizes South America arose.

Keywords: population genetics; mitochondrial DNA; maternal lineages; mtDNA haplogroups; CR haplotypes; Native-American ancestry; South America; Andean Colombian region.

Table of Contents

Figures and tables

Abbreviations

Table of abbreviations and acronyms and corresponding definition

AMOVA	Analysis of Molecular Variance
ANA	Ancestral Native American
ANS	Ancient North Siberian
ATP & ADP	Adenosine triphosphate & adenosine diphosphate
cal BP	Calibrated years before present
CR	Control Region
D-loop	Displacement loop
DNA	Deoxyribonucleic acid
dNTPs & ddNTPs	Deoxynucleotide triphosphate & dideoxynucleotides triphosphates
HVSI, HVSII and HVSIII	Hypervariable segment I, hypervariable segment II and hypervariable segment III
HVS	Hypervariable segment
IFC	Ice-Free Corridor
Indels	Insertion/deletion
ka	*kilo annum* / thousand years
LGM	Last Glacial Maximum
MDS	Multi-Dimensional Scaling
MgCl2	Magnesium chloride
min & s	minutes & seconds
mix	Mixture
MM	Multiregional Model
mtDNA	Mitochondrial Deoxyribonucleic acid
ng	Nanogram(s)
NNA	Northern Native American
NPC	North Pacific Coast
NRY	Non-recombining region of the Y chromosome
nuDNA	Nuclear Deoxyribonucleic acid
PCR	Polymerase Chain Reaction
rCRS	revised Cambridge Reference Sequence
RNA	Ribonucleic acid
SNA	Southern Native American
U/uL	Units per microliter
uL	Microliter(s)
uM & Mm	Micromolar & millimolar
UOM	Unique Origin Model

1. Introduction

1.1. Population Genetics

Advances in genetics has provided us with new ways of understanding what our ancestors could hardly imagine. Considering all that we have already been able to decipher about our biological information, we can look to the future with great expectations of what will follow - maintaining the humble capacity to understand how much remains to be discovered. Ever since Watson & Crick [1] first interpreted DNA (deoxyribonucleic acid) molecules as carriers of the genetic information, a major endeavor was to gather biological data and elucidate the biological processes guided by the genetic information. By 2001, the Human Genome Project launched in 1990, had drafted the raw information of a typical human genome [2] and many other genome projects were also launched in succession. These collaborative efforts provided an abundance of DNA data available and thus allowed a global perspective on the genome of different species, leading genomic research to an unprecedented level of protagonism [3]. Scientific innovations allied with the progressive advances in artificial intelligence are already showing signs of what it will be possible to do and understand in the next century. It is up to us, as scientists, to take advantage of these scientific and technological advances by developing studies, complementing, or refuting theories that try to explain various biological issues. The origin of humans and the evolution of modern human populations is one of the subjects that has been debated for many years. Genetics, in particular population genetics, has been highly important to clarify unsolved questions since it can bring many hints on evolutionary history. Expectedly, genetic approaches will remain essential tools for elucidating the origin and evolution of mankind, dissecting genetic diversity either in contemporary human populations or even in some already extinct.

Chakraborty, R. in 2001 [4] referred to population genetics as a discipline of scientific inquiry that deals with the extent and pattern of genetic variation among individuals within and between populations. In 2011, Goodwin and colleagues [5] describe it, in a more detailed perspective, as the study of factors affecting the allele and genotype frequencies at different genetic loci in a population; and assuming “population” as referring to a group of individuals that share a common ancestry. Those factors are responsible for changing and diversifying the genetic composition of a population, and consequently for its evolution. Therefore, population geneticists dedicate to describe and analyze the genetic structure of populations in order to draw inferences on the evolutionary events and forces that occur within and between populations, such as natural selection, mutation, genetic drift, admixture, and migration. Many plausible models were developed on human origins and migrations through the interpretation of comparative patterns of genetic diversity within and between populations in light of the principles of population genetics. Key demographic processes were already untangled based on current patterns of genetic diversity.

Darwin's thoughts on biological variation and Mendel's pioneer studies on the rules of transmission of traits from one generation to the next, are considered for many, the foundation stones of population genetics, a sub-field of genetics that soon would gain increasing autonomy [4]. In 1859, in the influential and controversial book the Origin of Species, Darwin defended, in a simplistic way, that modern species were descended from common ancestors and that the process of natural selection was the major mechanism of evolutionary change. It was as well seminal the Mendel's interest in how the distinctive characteristics of living beings are transmitted from parents to offspring, which has led to a theory that marked a turning point in our understanding of inheritance. His work published in *Experiments in Plant Hybridization* in the *Proceedings of the Natural History Society of Brünn* (1866) [6] resulted, several years later, in what we know today as Mendelian theory of inheritance. In the 1920s and early 30s, the mathematical theory developed in the pioneering work of Fisher [7], Haldane [8] and Wright [9] would result in the presentation of formal models to explore how selection, mutation and other evolutionary forces can modify the genetic composition of a population over time [10]. However, the main principle in population genetics is the Hardy-Weinberg theorem, derived independently by G.H. Hardy [11] and W. Weinberg [12] in 1908. This principle allows to make previsions on genotypic frequencies based on estimates of allele frequencies in populations with random mating and where selection, mutation, migration, and genetic drift are absent. Furthermore, the Hardy-Weinberg principles represents the basis that turns possible modelling microevolutionary change.

Contrarily to the deterministic model of Darwinian evolution, stochastic and non-adaptive processes, as contemplated in the neutral theory of evolution, are among the chief pillars of genome evolution. The Japanese geneticist Motoo Kimura and the American geneticist James F. Crow [13] assumed an infinite allele model and proposed that genetic variation in populations arises due to the balance between mutations and genetic drift. Kimura, M. [14] also proposed the "infinite sites model" (ISM), which provided a basis for understanding how mutation generates new alleles in DNA sequences. Both models would allow the refinement of previsions on the heterozygosity or genetic diversity in a finite population using allele frequencies and, ultimately, on the estimates of genetic distances between populations.

The realization that the Darwin's theory of evolution and the Mendel rules of inheritance were reconcilable with the formal models being devised in the field of populations genetics, played a key role in the consolidation of the Neo-Darwinism, a term referring to the modern synthesis of evolution that would prevail for long as the paradigm for evolutionary biology [15, 16].

With the advent of molecular genetics, it became possible to directly analyze deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) sequences, and long-term evolutionary studies became the focus. Then, population genetics took an even bigger leap with the wealth of data coming from the study of genetic similarities between organisms/lineages/populations and the development of the coalescent theory [17]. Modern population genetics has built on the theoretical edifice that already contained the classical models of population genetics, most notably by integrating the theory with data from molecular

biology [10]. Allied to approaches from molecular systematics and more prominently from molecular phylogeny, population genetics rapidly changed to a modern computational population genetics, which has provided essential tools for the interpretation of results that often interest empirical molecular evolutionists [17-19]. Molecular phylogeny applies to nucleotide sequences a series of molecular and statistical methods to infer the evolutionary connections among living beings or genes. On the basis that the closer the relatives, the more genetically similar they will be, genetic differences between species/sequences can be used to infer their evolutionary past. Relying in the concept of molecular clock [20], it is possible to recuperate evolutionary relationships and the timescales of the divergence of organisms/lineages [21].

The more recent advances in the sequencing of genomes and the sophistication of computer-based techniques for storing and analyzing genomic information, has prompted the strengthening of ties between population genetics and many other areas with some shared aims, such as phylogeography, paleoanthropology, archaeogenetics, sociobiology and anthropological linguistics.

While continuing to provide deeper insights into the understanding of the molecular basis of microevolution, population genetics is also a useful discipline in areas concerning public health, where interventions must be modelled integrating multidisciplinary levels and many issues that are societal problems. In the ‘postgenomic’ era, population genetics principles are being applied in diverse fields, including clinical medicine, epidemiology, demography, forensics, and risk analysis [4]. The importance of using those principles in areas such as epidemiology is clear and has been the focus of attention worldwide, particularly in the past two years due to the global pandemic of Covid-19 (Coronavirus Disease 2019), given that previsions of population genetic models can help in tracing the transmission routes of infectious diseases [22]. Genetics has also become the driving force in medical research [23, 24].

Forensic genetics uses molecular biology tools in an attempt to answer problems in the forensic context. Although forensic genetics has been around since the discovery of ABO blood groups [25, 26], it dramatically came into the limelight when Jeffreys and colleagues [27, 28] published the technique of “DNA fingerprinting”. In the following years, the mastering of technologies made it possible to directly disclose differences in DNA sequences, and soon more suitable molecular markers were recruited for the forensic setting. DNA forensics was born as a new area of scientific inquiry, in which population genetics played an integral role in interpreting forensic DNA evidence [4].

According to the definition that appeared in 2007 in the leading “Forensic Science International: Genetics” journal [29], forensic genetics is “the application of genetic to human and non-human material (in the sense of a science with the purpose of studying inherited characteristics for the analysis of inter- and intra-specific variations in populations) for the resolution of legal conflicts”. DNA databases are essential tools, used in a variety of different situations in the forensic context. Whether to estimate the frequency of a genetic profile or to direct search a profile, the quality and quantity of data available in those databases is a key to correctly report results and their statistical weight. This requires the

implementation of DNA databases that accurately describe the composition of reference populations, and genetic population studies contribute to the representativeness of a given population.

1.2. Mitochondrial DNA: a significant marker to assess genetic diversity

1.2.1. Mitochondria

Mitochondria are prominent and vital residents of the cytoplasm of eukaryotic cells. These double-membraned intracellular organelles are intimately involved in cellular homeostasis and in the cellular energy metabolism including the respiratory chain where the ATP, released from the mitochondrion in exchange for cytosolic ADP, is the high-energy source used for most active metabolic processes within the cell [30].

According to the endosymbiotic theory, proposed by Boston University biologist Lynn Margulis [31] in the late 1960s, mitochondria are descended from specialized bacteria that somehow survived endocytosis by another prokaryote or some other cell type becoming incorporated into the cytoplasm. Considerable evolutionary advantage favored the established symbiotic relationship between host cells and symbiont bacteria. Thus cells with mitochondria depend on these organelles to carry out oxidative phosphorylation, while mitochondria in turn depend on the cell to ensure their own existence.

Mitochondrial contain their own genome, mitochondrial DNA (mtDNA), that co-evolved with the nuclear genome (nuDNA) in a way that implied continuous exchange of information between mitochondria and the nucleus. The genome of each eukaryote encompasses, thus, two components: mtDNA and nuDNA.

1.2.2. The human mitogenome

After the discovery that mitochondria from chick embryos contained DNA, made by Margit Nass and Sylvan Nass [32] in the 1960s, around 20 years passed until the first complete sequence of the human mtDNA was reported by Fredrick Sanger's group in Cambridge [33]. The mitochondrial DNA of humans, located in the internal matrix of the mitochondria, consists of a circular, double-stranded molecule, with approximately 16.6 kilobases (kb), which represents only 0.00055% of the entire human genome. The two polynucleotide chains that make up the circular mtDNA molecule are differentiated from each other by the nucleotide composition of guanines. The polynucleotide chain with a higher contribution of guanine purines is called heavy chain (H) because it has a higher molecular weight, whereas the light chain (L) has a higher percentage of pyrimidines and, therefore, a lower molecular weight [34, 35].

The mitochondrial genome is present in a high copy number and, in most human cells, each mitochondria contains 4 or 5 copies of this genome, representing hundreds or even millions of DNA

molecules per cell [34, 36]. The coding region, accounting for 90% of the mitochondrial genome, encodes two ribosomal RNAs, 22 tRNAs, and 13 polypeptides that are components of the respiratory chain located in the inner mitochondrial membrane. The mtDNA displacement loop (D-loop, or control region) is a 1.1-kb non-coding region which is involved in the regulation of transcription and replication of the molecule. The D-loop extends from position 16024 to position 576 of the mtDNA and includes the origin of H strand replication, the promoters for H and L strand transcription, two transcription-factor binding sites, three conserved sequence blocks associated with the initiation of replication, and the D-loop strand-termination-associated sequences [37]. Despite its functional importance this region contains three short regions which, in comparison to the rest of the genome, hold highly variability at individual and population level: hypervariable segment (HVS) HVS-I, HVS-II, and HVS-III (Figure 1).

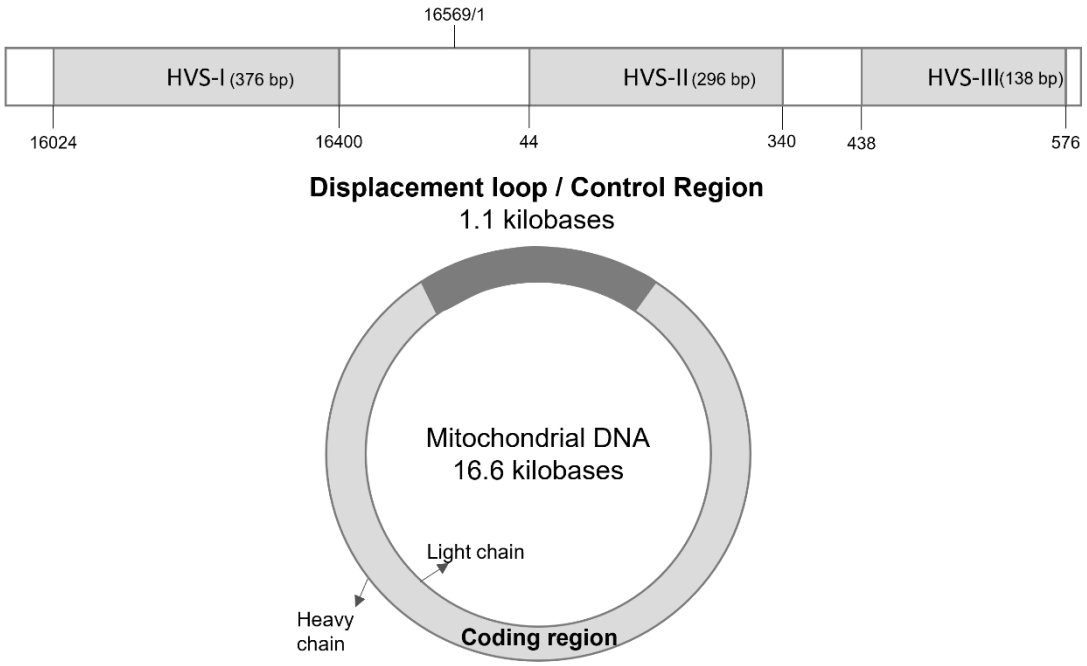


Figure 1 - Mitochondrial DNA: its coding region and control region/D-loop and correspondent length; on top, three hypervariable segments HVS-I, HVS-II, and HVS-III that make-up the control region.

Properties of mitochondrial transmission differ in many ways from Mendelian rules. mtDNA has unique features, such as the haploidy, absence of recombination, small effective population size (fourfold reduced compared to autosomes), that underlines its unique mode of inheritance. Together with the high mutation rate, these features makes mtDNA a focus of interest in various studies from population [38, 39] and forensic genetics [40, 41]. mtDNA is a uniparental or lineage marker because of its haploid genome which is, apart from rare exceptions [42, 43], maternally inherited [44] and does not undergo recombination. The genetic information defined by the combined variation at different non-allelic sites is designated as a haplotype, that behaves as if it were a single mtDNA allele per individual. Since they do not recombine, haplotypes are transmitted without changes to the next generation (except if mutation occurs). This non-recombining uniparental mode of inheritance is extremely useful for phylogenetic studies because enables researchers to trace related lineages back through time,

highlighting the maternal ancestry of a population, without the confounding effects of biparental inheritance and recombination inherent to nuclear DNA (excepting the non-recombining region of the Y chromosome). The high copy number, along with the extranuclear cytoplasmic location of mtDNA, makes it easier in many circumstances to obtain mtDNA for analysis than nuDNA. For this reason, mtDNA is the target of choice for analyzing ancient DNA and for addressing certain forensic questions, such for instance those involving Disaster Victim Identification [45], where usually DNA is found in low quantity. The mitochondrial genome has a very high mutation rate, 10- to 17-fold higher than that observed in nuDNA [46, 47]. The highest degree of variation in the mtDNA is found within the non-coding in hypervariable regions I and II (HVSI and HVSII, respectively). Thus, the mutation rate per site is not constant along the entire mitochondrial genome; being well established that some sites are mutational hotspots while others are prone to much lower rates of change [48-50]. Although the variation rate among sites and other issues, including related to germ line and somatic heteroplasmy, have been major challenges for obtaining accurate estimates of human mtDNA mutation rate, substantial differences have been consistently reported among functional domains of the molecule [51].

With the exponential increase of mtDNA studies, it has been necessary to create high-quality mitochondrial databases and standardize database searches so that reliable comparisons between haplotypes and sequencies can be performed. These databases are constructed using sequence information from which it is possible to assess if a given haplotype was previously identified in a population. This way, estimations on haplotype frequencies and on mtDNA diversity patterns can be obtained. The most worldwide used forensic database is the EMPOP – EDNAP (The European DNA Profiling Group) mitochondrial DNA population database (empop.online) [52], which is managed, maintained and updated by the Innsbruck Institute of Forensic Medicine. In order to meet the highest forensic standards, this database only host sequences that have previously passed through its quality control. Besides containing haplotypes corresponding to the total control region and mitogenomes of individuals from all over the world, the EMPOP database also provides software tools to obtain important parameters in population and forensic studies.

1.2.3. Analysis of the mtDNA

The analysis of human mitochondrial DNA is based on comparison of mtDNA sequences within and between human populations. The initial studies considering mtDNA variation employed restriction fragment length polymorphisms (RFLPs) [53-56] and were instrumental in developing mtDNA as a new molecular tool. When Kocher, et al. [57] reported on highly conserved primers that could be used to amplify by polymerase chain reaction the mtDNA from a wide range of taxa, it opened the door to bring mtDNA to the research limelight, with a huge amount of mtDNA data generated in the 1990s. The first mtDNA sequencing studies were performed by Sanger Sequencing method, and explored only the HVSI [58, 59] or the HVSI plus HVSII [60, 61], since these two segments are the most hypervariable. Later, the study of the mitochondrial genome has been extended to the HVSIII of the control region and next to the

entire control region [62, 63], which significantly increases the discrimination power between samples. In order to augment more that discrimination power, the coding region of mtDNA also began to be analyzed, initially at limited positions, but soon after, with the advances in technologies, the target was extended to the whole mitochondrial genome. The implementation of massively parallel sequencing for DNA analysis, has allowed to obtain more routinely whole mitochondrial sequences in laboratories performing forensic, clinical, or research-based investigations [64].

The first complete sequence of human mitochondrial DNA was published in 1981, and it became known as the Cambridge Reference Sequence (CRS - Cambridge Reference Sequence) [33]. In 1999, R. Andrews noted inconsistencies in the CRS by identifying and correcting 11 errors. Thus, the resequenced and revised sequence was published as revised Cambridge Reference Sequence (rCRS – revised Cambridge Reference Sequence) [65], which has continuously been used for comparative purposes [66]. The rCRS belongs to an individual who fits into a typically European group of lineages. The routine procedure in mtDNA analyses, is to align a given sequence with the rCRS and then the detected polymorphisms (single nucleotide polymorphisms [SNPs] and insertions/deletion polymorphisms [indels]) are scored as differences (“mutations”) to the rCRS, making up the haplotype of the sample. The haplotype is described following specific and universal guidelines established by and for forensic community [67]. Since the differentiation between haplotypes is due to mutations that accumulate along time, mtDNA diversity can be analyzed from a phylogenetic viewpoint. According to the level of molecular similarity, haplotypes are grouped into monophyletic clusters, referred to as haplogroups, each of which has specific defining mutations [68]. Then, haplogroups are used to infer phylogenetic relationships, in an attempt to group lineages by the hierarchic order of their descent [69]. A regularly updated classification of global mtDNA variation is available in a single phylogenetic tree provided on the website Phylotree (http://www.phylotree. org) [70] (Figure 2).

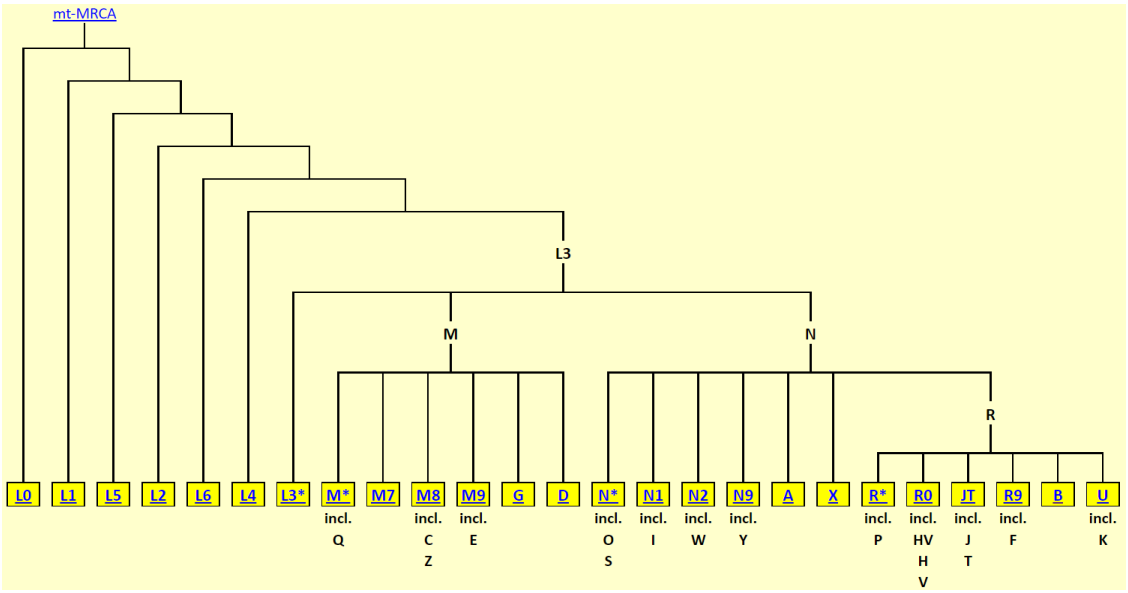


Figure 2 - Phylotree, the updated classification tree of global mtDNA variation. (http://www.phylotree.org - accessed in 06/2021)

The human mtDNA phylogeny associated to the fine knowledge of the geographic distribution of haplogroups has permitted a very plausible reconstruction of the human origins and the ancient migrations of women.

The deepest and oldest branches of the phylogenetic tree include haplogroups characteristic of Sub-Saharan Africa: they comprise, the macrohaplogroup L, which include haplogroups L0, L1, L2, L3, L4, L5 and L6 [59, 71]. It was from haplogroup L3 that the remaining haplogroups diverged. Haplogroup L3 branches into haplogroups M and N originated in eastern Africa and dispersed, first, into Eurasia [72]. Sequences typically found in European populations are derived from the haplogroup N, which includes the subclades H, I, J, K, T, U, V, W and X. Asian lineages are derived from both haplogroup M and haplogroup N. The first includes the C, D, E and Z and the second the A, B, F and J. In Native American populations, the haplogroups typically found are A, B, C, and D, which can also be found in Asian populations. The four basal Native-American clusters were described for the first time in alphabetical order – haplogroups A, B, C and D – by Torroni, A. and colleagues [73] when studying mtDNA of Native Americans. The presence of these lineages in America has long preceded the European colonization of the continent. Because haplogroups might inform on maternal ancestry, they can be helpful for estimating admixture proportions in populations and for bringing light on migration routes that still remain uncertain. Furthermore, analysis of mtDNA diversity can led to detect other important events, such as founder and bottleneck effects. As a matter of fact, these two genetic drift effects occurred often in the process of peopling of Americas, either during the migration(s) that culminated with the arrival of the first inhabitants in North America or later when populations expanded towards South of the continent, so that the multiple founder and bottleneck effects were key players in modelling the degree of diversity in pre-Columbian Native-American populations [74-76]. In the 16th century, America entered a new level of demographic and population complexity, with successive non-American waves of migration and admixture processes that would result in a dramatic reconfiguration of its population scenario.

mtDNA is only a single maternally-inherited locus that cannot reflect the whole complexity of the past demographic processes that America experienced. However, it allows an extremely detailed reconstruction of the nesting relationships within a phylogeny, a feature that can be extremely informative for recuperating and dating migration and population separation events, explaining why studies of mtDNA have extensively contributed to the current view we have on American populations [77].

1.3. The peopling of America

1.3.1. Origin of Modern Humans

The origin and demographic history of modern humans is a topic that continues to be hotly debated. Broadly speaking, there are two main theories on the subject, the Multiregional Model (MM)

and the Unique Origin Model (UOM). The first model states that all human populations living today were originated from multiple archaic human populations that evolved in different parts and local contexts of the world, following the earlier migrations of the ancestral Homo erectus populations within and out of Africa. The model further assumes intense gene flow between all archaic human populations [78]. The other model, also referred to as the “Out-of-Africa” or “Recent African Origins” model [54, 79], is nowadays more widely accepted, and holds that anatomically modern humans arose in Africa ~200 ka (thousand years) ago as a new species [39, 54]. Approximately 100 ka to 60 ka ago, small population(s) from Africa started expanding and eventually spread across the world, replacing all non-African archaic humans [80]. Recent genetic and archaeological evidence are sustaining a less simplistic scenario then before assumed, and revised versions of UOM have emerged, such as the assimilation model [81], which considers an initial African origin, spread of modern humans out of Africa to regions already inhabited by archaic humans with which the new comers admixed in some extent. This might explain the signs for low, but not insignificant, levels of archaic human gene flow into modern populations [82].

Fossil records are consistent with the first advent of modern humans in Africa, much earlier than in other geographic regions [83, 84]. Furthermore, phylogeographic analyses based on the genetics of present-day populations point most probably to an African origin, combined with a series of dispersions out of Africa. This is corroborated by the levels of genetic diversity, which are typically higher in Sub-Saharan Africa when compared to other parts of the world. Furthermore, genetic diversity outside Africa often represents a subset of the diversity within Africa, and the level of genetic diversity within populations decreases as the geographic distance from Africa increases [85-87]. These observations can be explained by the effects of genetic drift, including successive bottleneck and founder effects, following the exit from Africa [88, 89]. mtDNA diversity, for instance, captures well these signatures, since among the sequences currently available, the first five branches in the mtDNA tree encompass exclusively sequences from people living in sub-Saharan Africa (or with ancestry there). The first scientific study based on mtDNA that provided strong evidence on the time and place of the origin of modern humans was developed in 1987, by Allan Wilson and his collaborators [54] who constructed a phylogenetic tree with mtDNA haplotypes from 147 individuals worldwide, from which it was extrapolated that the most recent maternal common ancestor was African, present in a hypothetical woman that become commonly known as “the mitochondrial Eve”. In the next decades, the result was systematically replicated, strengthening the UOM. However, some data became to appear not supporting a model of complete replacement. Unprecedented insights emerged when it was possible to obtain partial or complete genomic sequences from archaic hominins fossils. Effectively, ancient DNA from Neanderthal fossils showed that roughly 1–4% of the ancestry of present-day Eurasians is Neanderthal [90, 91], likely deriving from interbreeding in the Middle East following the initial dispersion of modern humans out of Africa. In 2010, another group of extinct archaic humans was identified on the basis of genomic data, the Denisovans which appear to be a sister group of the Neanderthals, but different genetically from both modern humans and Neanderthals. Denisovan ancestry has been detected in present-day Melanesians, Australian aborigines, and Polynesians, and in some parts of Indonesia [92, 93]. Very recent analyses

corroborate that the genetic variation humans carry today can be traced back to multiple ancient populations, and further suggest that the current-day human genome retains ancestry components that can be attributed not only to Neandertal and Denisovan introgression but also to admixture with other yet-to-be-discovered extinct hominins populations [89]. The genetic data obtained so far support an origin of modern humans largely influenced by the out of Africa event(s) but allowing for some level of gene flow with local archaic forms following the dispersal out of Africa, as it is sustained in the assimilation model, a model that recently was also called “out of Africa with introgression” [94].

1.3.2. Exploration and settlement of the Americas

North and South America were the last continents to be explored and settled, representing the culmination of a Late Pleistocene expansion of anatomically modern humans out of Africa [95]. Evidence suggests that the arrival and spread of *Homo sapiens* across the Americas involved initial migrations from a structured Northeast Asian source population with differential relatedness to present-day Australasian populations, followed by a divergence into northern and southern Native American lineages [96]. Solutions to many unanswered or controversial questions and new highlights into this subject have been the aim of recent archaeological and genetic studies.

1.3.2.1. Entrance in the Americas

During the Late Pleistocene period – the ice age that lasted from about 126 ka to 11 ka ago – the northeast of Siberia was connected to North America through a land bridge that existed from about 34 ka to 11.7 ka ago and that is thought to have been the passage of people that entered in the Americas, sometime between 30 ka and 15 ka ago. Considering today’s biogeographical borders, this region called Beringia, extends from the Verkhoiansk Range in eastern Siberia east to the Mackenzie River in northwestern Canada and includes Kamchatka, Chukotka, and the Bering Sea area. As early as 1590, the Spanish missionary Fray Jose de Acosta had already suggested that the ancestors of Native Americans might have walked over a yet-to-be-discovered land bridge between Asia and North America [97, 98]. Although the Acosta’s view meant to address political and religious issues (and not properly the origin of the Native-Americans), his intuition would later be incorporated in most of the theories on the peopling of the Americas that emerged based on the growing evidence coming from disparate fields as linguistics, archeology, paleoclimatology and genetics.

In the late Pleistocene, most of Siberia was occupied by individuals of a population designated Ancient North Siberian (ANS), which had diverged from West Eurasian populations shortly after their split from East Eurasian populations [99]. Current evidence also suggests that around 23–20 ka, there was gene flow between an ANS group and an East Asian group. Gene flow between these populations ultimately gave rise to at least two distinct lineages, the Ancient Paleo-Siberians forming the ancestral

population of present-day groups of northeast Siberia, and the basal American branch, whose descendants ultimately crossed to the Americas [100, 101].

Archaeological evidence of humans in eastern Beringia at Swan Point in central Alaska, where distinctive technological artifacts dates from 18 to 12.6 ka ago, appear to document the dispersal of microblade-producing humans from Siberia to Beringia during the late glacial. Genetic studies convincingly demonstrate that Asia gave rise to the first Beringians and Americans. Genetic data reveals that the ancestors of all contemporary Indigenous people had descended from only five maternal lineages (haplogroups A, B, C, D, and X) and two paternal lineages (haplogroups C and Q) [102, 103]. These lineages also indicated that the founding population came from Asia and experienced a severe genetic bottleneck, in which a small number of people with limited genetic diversity gave rise to all Indigenous people who occupied the continent before European arrival [104].

Precisely where and when the basal American branch emerged remains uncertain, nevertheless its emergence must have been before approximately 21–20 ka, as by then the basal American branch had begun to diverge into separate lineages, and none shows evidence of subsequent gene flow from Ancient Paleo-Siberian or other northeast Asian populations [100]. Gene flow to and from east Siberia certainly appears to have ceased by the height of the Last Glacial Maximum (LGM). Current coalescent estimates based on variation in extant mtDNA lineages, set the event at 25 to 20 ka [105] or less than 20 ka [98, 106], after the LGM, and estimates based on Y chromosome variability suggest that divergence occurred after 22.5 ka, possibly as late as 20 to 15 ka [107, 108].

1.3.2.2. Beringian Standstill

The first founding population of the Americas was isolated from Eurasian populations before its radiation into a multitude of sub-populations in America [96, 109]. mtDNA data suggested that, before spreading across the Americas, the ancestral population paused in Beringia – “Beringian Standstill” – long enough for specific mutations to accumulate, that separate the New World founder lineages from their Asian sister-clades. This incubation or “standstill” was originally hypothesized to have taken 15 ka [110], although new studies argue it began later and lasted less than 8 ka [105, 111].

From that isolated population, several lineages emerged: unsampled population A, a ‘genetic ghost’ of which little is currently known, ‘Ancient Beringian’ individuals, and ‘Ancestral Native American’ (ANA) individuals. This last three populations ultimately crossed into North America, but the deep divergence and limited gene flow between them indicate that they probably did so in separate movements [112].

Evidence suggests that ANA individuals crossed Beringia and reached North America south of the continental ice sheets ahead of Ancient Beringian individuals. A recent analysis of the 11.5 ka old Upward Sun River 1 genome supports the idea of a single population source separated from East Asians

by 26.1 to 23.9 ka ago, with two deep branches: an Ancient Beringian population that split off ~22 to 18 ka ago and a second branch that split into northern and southern lineages ~17.5 to 14.6 ka ago, originating the Northern Native American (NNA) and Southern Native American (SNA) populations, which are genetically equidistant to Ancient Beringian individuals [113].

NNAs appear to have remained in northern North America and, perhaps after the disappearance of Ancient Beringian peoples, shifted further northward, as they are presently in Alaska and the Yukon [114]. Members of the SNA branch ultimately reached southern South America, and on the basis of mtDNA, Y chromosome, and genome-wide evidence, this likely occurred quickly and involving complex admixture events between earlier established populations [112, 115].

1.3.2.3. South from Beringia

The route(s) that people used to travel southward from Alaska and the timing of the first entry of humans into North America are still hotly debated within the scientific community. Two main models try to elucidate the first migratory paths taken into the American continent: Ice-Free Corridor hypothesis (or IFC) and the North Pacific Coast hypothesis (or NPC).

Canada and lower North America are thought to have been inaccessible to the occupants of Beringia during the LGM due to the coalescence of both the Laurentide glacier, which extended between the Atlantic and the Rocky Mountains, and the Cordilleran ice sheet, which extended from the Pacific Coast to the Rocky Mountains. The IFC hypothesis has been a reasonable theory for decades, stating that the spread to the American continent occurred after an ice-free corridor opened in postglacial times along the eastern flank of the Rocky Mountains, allowing populations trapped in Beringia to continue exploring towards the south. According to this traditional notion, Clovis culture hunters – then considered the earliest inhabitants in North America – arrived by chasing now-extinct large-bodied mammals (assumedly elephant and buffalo-like) following an open corridor between the ice slabs [100].

The Ice-Free Corridor's existence and usefulness for human colonization are not questioned, but the latest theories about the timing of its viability and of human colonization have seemingly ruled it out as the first pathway taken by people arriving from Beringia and northeastern Siberia.

Geological evidence shows that the corridor was not fully ice-free until around 15–14 ka, and ancient DNA from both fossil bison and lake sediments, indicate that the plants and animals that hunter-gatherers would have needed for food along the roughly 1,500-kilometers route were not available in the corridor region until about 13 ka [116, 117]. Thus, this route would not have been viable early enough for the first peoples' travels.

Analysis of mitogenomes places the arrival of humans into unglaciated America at ~16 ka ago [105], and Y-chromosome estimates place their arrival sometime between ~19.5 and 15.2 ka ago [118], predating the opening of the inland ice-free corridor and pointing to a southward expansion along

recently emerged northwest Pacific coastal land. Additionally, a late-entry and rapid dispersal model of humans across the New World is inconsistent with the distribution of genetic variation observed in Native American populations today [104].

Several archaeological sites date to well before the supposed ice-free corridor was open — including Manis Mastodon Site in Washington (dated to 13.8 ka ago) [119], Paisley Caves in Oregon (dated to 14.5 ka ago) [120, 121], Monte Verde in Chile (dated to 14.8 ka ago) [122], and, known as the oldest confirmed human occupation site within the main route of the corridor, Charlie Lake Cave (dated to 12.4 ka ago) [123], where the recovery of both southern bison bone and Clovis-like projectile points also suggest that the corridor was not used just for southbound movement and that people were moving north as well. Clearly, people who lived in far southern Chile by 15 ka ago could not have used the ice-free corridor to get there.

An alternative route for the first settlers has been proposed along the Pacific coast – North Pacific coast hypothesis – which would have been ice-free and available for migration for pre-Clovis explorers in boats or along the shoreline. Glacial ice blocked the interior route as early as around 23 ka, but with the post-LGM retreat long reaches were ice-free after 17 ka and, by 16–15 ka, the coast was largely clear and supported the resources necessary for human travelers [124-126]. A coastal route would have enabled people to reach the Americas south of the continental ice sheets well before the earliest currently accepted archaeological presence. The presence of human remains dating to 13.1 to 13 ka at Arlington Springs, on Santa Rosa Island off the coast of California, indicates that the first Americans used watercraft [127].

The genetic, archeological, and paleoecological evidence are largely consistent with either an inland (IFC) or Pacific coastal (NPC) routes, but neither can be rejected at present [128]. It is also not possible to exclude that, the entrance in the double continent might have occurred through both the Pacific coastal corridor, which apparently brought mtDNA haplogroups A–D often defined as “pan-American” (A2, B2, C1b, C1c, C1d, C1d1, D1, particularly marked by D4h3a sub-haplogroup [129]) and NRY haplogroups P-M45a and Q-242/Q-M3 haplotypes with them to the Americas, with these being dispersed throughout all continental areas of the New World and leaving the greatest genetic mark; and through the subsequent expansion coincident with the opening of the ice-free corridor that probably brought mtDNA haplogroup X (marked by X2a and C4c sub-haplogroup [129, 130]) and NRY haplogroups P-M45b, C-M130, andR1a1-M17, with these being disseminated in only North and Central America [106, 131]. Much later arrivals are thought to have brought other sub-clades such as A2a, A2b, D2a, and D3, reshaping the original maternal makeup of NNA by additional streams of gene flow and local population dynamics [132].

1.3.2.4. Dispersals into South America

South America was the last major area to be occupied by modern humans, but the process and subsequent population movements that led to the peopling of South America remain poorly known. The west and east sides of South America have different archaeological records, different chronologies, and different human histories, reflecting distinctive patterns of genetic, morphological, and cultural variation [133]. The current most consensual view on is peopling is that the first Paleoindian settlers reached the northern edge of South America from a single wave of migration around 16–15 ka ago and split into two groups that rapidly expanded southwards along two main routes: throughout the Pacific coastline, and eastwards crossing the Amazon basin and proceeding South along the Atlantic coastline [112, 134-136] (Figure 3).

Based on archaeological evidence of an early occupation of the Andean highlands, some scholars hypothesized that an additional inland Andean route could had been important in the spread of people in South America. However, other archaeological record and ancient DNA studies showed an interconnection between highlands and coastal groups, suggesting that the western Andean occupation occurred through the Pacific route [136], reinforcing the model assuming the Pacific and the Atlantic as the two mains routes of migration in the initial settlement of South America.

Figure 3 - Demographic scenario for the peopling of South America suggested by Gómez-Carballa, A. and colleagues [135], where mitochondrial DNA variation at different Andean locations and >360,000 autosomal SNPs from 28 Native American ethnic groups were analyzed.

The Pacific coast was the main route for South America in prehistoric time and was characterized by mostly north to south direction population movements and rapid population growth marked by a series of bottlenecks. Contemporary and ancient mitogenomes from the western Andes revealed sub-haplogroups that originated in South America between 15.7 and 13.5 ka ago after initial entry into this region [104]. A small number of individuals succeeded to access and adapt to the high-altitude Andean regions and were then subjected to very limited genetic exchange with the East of the subcontinent, where the Amazon Forest is also thought to have acted as a geographical barrier to gene flow between these two main expansion waves. This model of southward migration along the western side of the Andes is consistent with an interpretation that modern speakers of Andean languages may represent descendants of the first occupiers of the region [137, 138]. Native South American variation based on Y-chromosome microsatellites documented a west-to-east genetic discontinuity between Andean and eastern Brazilian tribes, representing one of the strongest signals of sub-continental genetic differentiation [139-141].

The Atlantic migration wave played a major role in the colonization of the non-Andean regions. Movements along the Atlantic coast could have been important not only in a north to south direction but also with an opposite course that might have even involved movements toward the Caribbean. Both waves of migration have been further supported by analysis of genetic variation across South America, particularly by the analysis of geographically restricted mtDNA sub-haplogroups. The continental Caribbean connection is, for instance, reflected by the presence of B2b3a1a in Puerto Rico at relatively high frequencies, almost 10 ka after the occurrence of the ancestral B2b3 in the Atlantic coast of South America [77, 135].

A signal of assumed Australasian ancestry was recently detected in present day Native American populations and was proposed as another contributor to the South American gene pool. Ancient human remains of individuals from Lagoa Santa (Brazil) and some present-day Amazonian tribes were found to share intriguing genetic similarities with native people from New Guinea, Australia, and Andaman Islands. According to Skoglund and colleagues [142], this Australasian signal derived from an extinct ancient ancestor that contributed for both Australasian and Native American ancestors (Population Y) and was anciently introduced in South America through the Pacific coastal route before the formation of the Amazonian branch, yet leaving no apparent traces in North America [104, 143].

The archaeological and genetic evidence show that the colonization of the Americas was a complex process that we are only beginning to understand.

The current territory of Colombia played an important role on these migratory routes as crossing passage between North and South America. A comprehensive study on the Colombian gene pool will help not only to characterize the admixture structure of the country but also to clarify pre-Columbian migrations by dissecting information on native lineages.

1.3.3. Colombia

1.3.3.1. Geography and demographics

Colombia is divided into six natural regions (Figure 4) constituted by differences in topography, weather, vegetation and types of soil: the Andean Region, covering the three branches of the Andes mountains; the Caribbean Region, covering the area adjacent to the Caribbean Sea; the Pacific Region adjacent to the Pacific Ocean; the Orinoquía Region, part of the Llanos plains mainly in the Orinoco river basin along the border with Venezuela; the Amazon Region, part of the Amazon rainforest; and finally the Insular Region, comprising islands in both the Atlantic and Pacific oceans. The Andes region and the Caribbean region were and remain the most densely populated and economically active in the country.

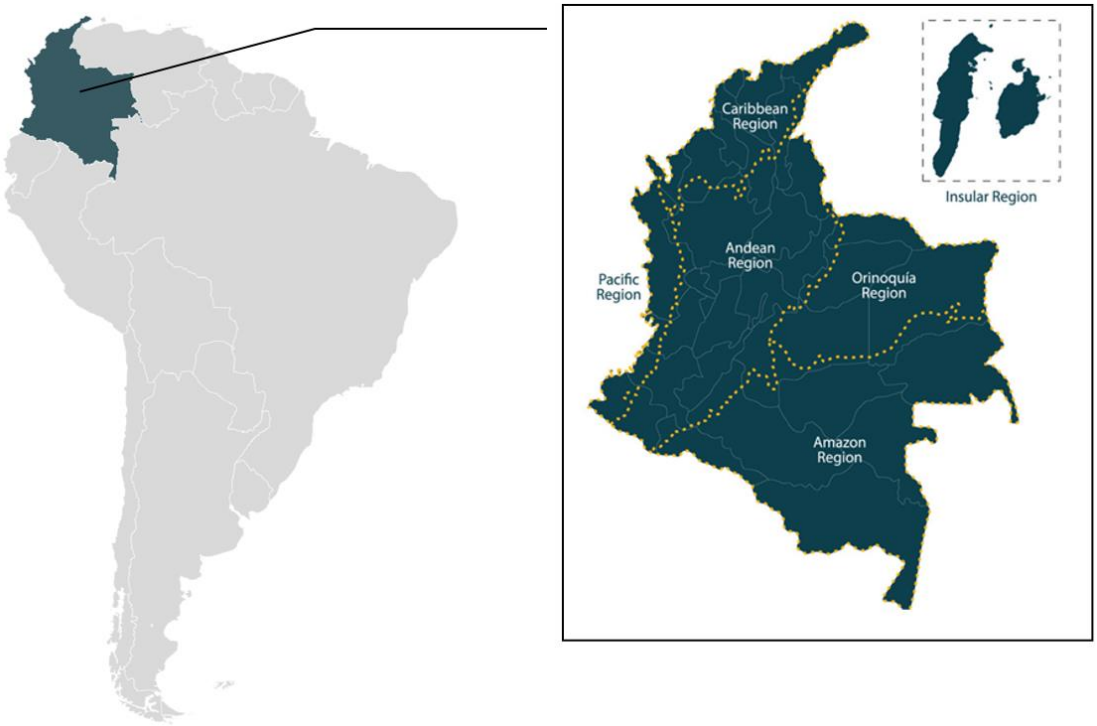


Figure 4 - Geographical illustration of Colombia in the South American continent and, on the right, Colombia divided into six natural regions. Adapted from dreamstime.com

With an estimated 50 million people in 2020, Colombia is the third-most populous country in Latin America, after Brazil and Mexico [144]. Concerning political subdivisions, Colombia is a unitary republic made up of a total of thirty-two departments. Traditionally a rural society, movement to urban areas was very heavy in the mid-20th century, and Colombia is now one of the most urbanized countries in Latin America. The Andean Region in Colombia correspond to 24.5% of Colombia's territory and is the most populated natural region of Colombia. Its impressive ecosystem, species and landscapes diversity are the result of the complex topography and of the periodic isolation determined by the Ice Ages during the Pleistocene, which propitiated high variation in climate and soil conditions [145].

The Colombian Andes can be divided into three branches known as "cordilleras": the West Andes run adjacent to the Pacific coast, the Central Andes run up the center of the country between the Cauca

and Magdalena River valleys (to the west and east respectively), and the East Andes extend northeast towards the Guajira Peninsula.

The Indigenous and admixed populations from Latin America have unique patterns of structure, to which accounted their more or less remote past demography, linguistic features and degree of geographic isolation. Colombia is characterized by a great heterogeneity of ethnic groups/populations inhabiting amply different geographic regions. Their patterns of diversity were modeled by the vaguely known ancient history of the people that lived in what is now known as Colombia until the arrival of the European colonists, and then by the events that followed the European colonization, including forced migration, admixture, society segmentation, isolation and decimation of native groups with consequent drift effects [146]. Besides the early Spanish settlers and African slaves, the 20th-century migrants from Europe and the Middle East, have also contributed to the current cultural and genetic heritage. Several linguistic groups coexist in Colombia since before the Spanish colonization, some of the most relevant being the Chibchan, Carib and Arawakan, in the Atlantic coast, Chocoan in the Pacific coast and Paezan, Barbacoan and Quechua in the Southern Andean region. Nowadays Colombia presents 87 recognized ethnic groups and even though Castilian is the official language in Colombia, there are still 64 indigenous languages spoken throughout the country [147]. According to the 2018 census, “Native-American” only accounts to 3.4% of the total population of Colombia, “African Colombian+Mulatto+Palenquero+Raiza” to 10.5%, “European Colombian” to 30.7% and “Mestizo” to 53.5% [148]. The large “Mestizo” population includes people living in rural areas but mainly in the cities where they have played a major role in the urban expansion of recent decades.

1.3.3.2. First settlers in Colombia

The so-called Isthmo-Colombian area, tucked between Mesoamerica and the Central Andes, forms the northwestern corner of South America and represents a probable gateway to migrations of Paleoindians from North to South America. Hunters and gatherers who migrated to South America via the Isthmus of Panama could have entered the Andean highlands by way of the Cauca and Magdalena River valleys which flow from south to north in Colombia [149]. Thus, Colombia represents this pivotal juncture between Central and South America, where peoples from the Mesoamerican, Caribbean, Amazonian, and Andean regions have all interacted.

Researchers point the time for the first human arrival into northern South America to have occurred sometime between 16,600–15,100 cal BP (calibrated years before the present), and probably at 15,500 cal BP [134, 150, 151]. There is an increase of archaeological sites in Colombia suggesting that this was a period of human expansion, adaptive adjustments, and population growth in the region [152, 153].

The Sabana de Bogotá (SB) in the eastern highlands of Colombia, Northern South America, is a well-known archaeological region that played an important role in the initial human expansion into South

America [154]. The earliest human contexts in Colombia are presently from Andean sites on the Bogota plateau, such those from the Tibito site, dating to ~13,600 cal BP (calibrated years before the present), and the Tequendama rock shelter, dating to ~12,850 cal BP [155-157]. Archaeological research over the past two decades in the sub-Andean Forest of the Middle Cauca region of central Colombia has documented sites dating from the terminal Pleistocene to middle Holocene, starting at ~12,600 cal BP [158, 159]. In the Colombian Amazon, along the Caquetá River, lies the Peña Roja site, dated between ~11,069 and 9,168 cal BP [160]. Recent excavations of three rock shelters in the Serranía La Lindosa (on the northern edge of the Colombian Amazon) document the earliest occupation of the Colombian Amazon, starting ~12,600 cal BP; and provide evidence of these earliest inhabitants and their environmental interactions [155].

1.3.3.3. European arrival, colonial exchange, and recent migratory events

In 1492, a Spanish expedition headed by Christopher Columbus landed in the Bahamas and set in motion the European settlement of the American continent dominated by immigrants from Spain and Portugal. The first century of the Spanish occupation was marked by a predominance of men as immigrants. Following their arrival, the annihilation of natives, the implementation of systems of forced labor and the spread of European diseases, resulted in a dramatic reduction of the population size of native Americans. Early in the colonial period, the Spanish and the Portuguese also initiated the slave trade process of Africans to the Americas which gained strong impetus with the reduction of the native population [161].

The Spanish explored Colombia in 1500 and began to colonize it soon after, with Santa Marta founded in 1525, followed by Cartagena in 1533, and by mid-century the occupation was complete. During these periods, a large number of African slaves from the west-central coast of Africa were taken to Colombia, to work in agricultural, livestock and mining exploitation. The coast of Cartagena de Indias in Colombia was the main port for slaves arriving in America.

In the beginning of the nineteenth century, comparatively to what was observed in other parts of the subcontinent, the immigration of Europeans to Colombia was demographically less important. However, the European presence in Colombia increased substantially between the mid-19th and mid-20th centuries. Two waves of Jewish immigrants settled in Colombia between 1830 and 1938, contributing significantly to regional financial sectors through trade and business. Germans arrived in the mid-1800s, and by the early 1900s were leading revitalization of the coffee, tobacco, and banking industries. After World War II, many European technicians and agricultural experts migrated to Colombia. Other immigrant communities that developed successful enterprises were the Spanish, French, Italians, and Americans, who settled in the capital of Bogotá for the most part, and the Japanese who settled in the Cauca Valley region [162].

The encounter of Native Americans with large numbers of Europeans and Africans resulted in extensive admixture in the continent. The extent to which this admixture has taken place has been influenced by geography, the timing and magnitude of population migration, and a range of social factors. All of these factors ultimately affected the patterns of genetic diversity across the American continent, resulting in what we detect today, that is presence of Native American, African, and European genetic ancestries varying throughout the country.

1.4. mtDNA studies in Colombia: state of the art

Colombia has been the target of several forensic and anthropological genetic studies [149, 163-180]. Many Colombian populations were already characterized for mtDNA aiming, through the comparison of the haplotype and haplogroup composition of different populations helps to clarify migration events including the peopling of the South America and the more recent demographic transformations that arose with the European colonization. Other studies were conducted with more strict forensic purposes, which in Colombia have an added value since it is a country with serious social and violent conflicts, requiring a well-established genetic population and forensic anthropological frame in the post-conflict eras, for the correct statistical evaluation of the evidence in forensic cases.

Native Americans exhibit a low haplogroup variability when compared with other continental contexts, with only four haplogroups, initially named A, B, C, and D, later relabeled as A2, B2, C1, and D1, encompassing the vast majority of native mtDNA haplogroups in the entire double continent. Native American mtDNA haplogroups, according to the current classification, are A2, B2, C1b, C1c, C1d, D1, and the less frequent C4c, D4h3a, and X2a [181]. If most South American populations reveal relatively high frequencies of these haplogroups, many also show clear signatures of the significant inflow of European and African lineages over the last 500 years, which naturally complicates to make inferences on the genetic diversity patterns in pre-Columbian South Americans using data from contemporary populations. Importantly, however, the recent studies on ancient mitogenomes hold the hope that in the near future much more will be discovered about the demographic past of human populations, including from America.

In respect to contemporary Colombian populations, globally their genetic composition share many features with those seen in other South American populations, including the clear admixed profiles [164, 176, 181, 182].

Also in concordance with the generality observed in the South American populations studied to date, the mtDNA lineages found in Colombian populations are predominantly of Native-American descent [183-185], associated with small inputs of European and African ancestries.

A differential male/female admixture ratio is quite visible in several Colombian regions. This is the common pattern across America, that was established since the initial stages of the colonization

process. It is well documented that only a minority of European settlers and African slaves brought to America were women. Analysis of uniparental genetic markers in non-Amerind populations from Colombia has shown that native male lineages were almost entirely washed away, but at the mtDNA level, native lineages were by far predominant, which is fully consistent with an asymmetric pattern of mating, involving mostly European men and native women to establish neo-American populations [146, 163, 164, 184].

A remarkable aspect of Colombian Amerindian groups is their high level of genetic heterogeneity when compared to Amerindian groups from other regions from South America [186-188].

In short, the pools of mtDNA lineages nowadays found in Colombian populations were determined by the poorly known pre-Columbian historical events and migrations that occurred during more than ~15000 years since the arrival of the first settlers to South America, and, later, during the scarce last five centuries by the disturbing population contacts initiated with the European colonization.

2. Aims

The knowledge of the genetic background of South American populations is crucial to a better understanding of the colonization of this continent by modern humans, as well as to obtain insights into the deep population(s) remodeling that America went through in consequence of the more recent human migrations during the colonial Era. There are already some genetic studies focused on South American populations. However, due to the high heterogeneity of these populations, more genetic data is still required for a comprehensive view of the genetic diversity and its stratification in this sub-continent.

The main objective of this work is to obtain a better landscape of the maternal genetic diversity in Colombia, assessing its distribution across different population groups, through the study of mitochondrial DNA (mtDNA).

To achieve this main aim, the following goals were proposed:

- Genotype and characterize the maternal lineages in two admixed populations living in the Andean Colombian region, through the analysis of the entire control region of the mitochondrial DNA
- Interpret the mtDNA profiles obtained in light of the current mitochondrial phylogeny
- Accomplish statistical processing and data analysis
- Perform comparative analysis with data available from published studies
- Contribute to a deeper understanding of the colonization of the New World and the history of its current-day populations

3. Material and methods

In order to carry out the objectives listed above, the experimental procedure carried out is briefly described below in flowchart format, and in more detail in the following points.

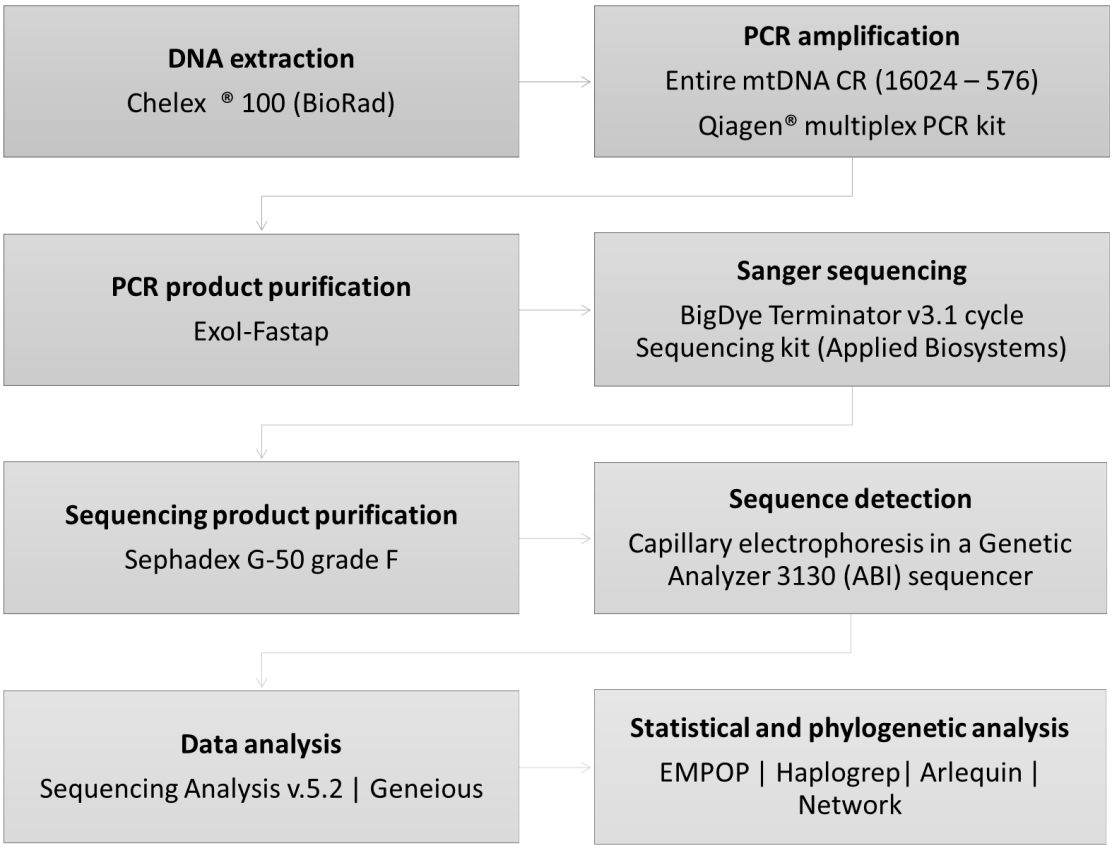


Figure 5 - Experimental procedure overview.

3.1. Population samples and DNA extraction

A set of 53 blood samples were collected under informed consent from unrelated individuals born in the urban areas of Colombian Andean departments of Tolima and Cundinamarca, thus belonging to admixed populations. DNA was extracted using Chelex® 100 (BioRad), technique developed by Walsh and collaborators in 1991 [189, 190].

3.2. mtDNA typing

The mtDNA typing was carried out following the recommendations of the DNA Commission of the International Society for Forensic Genetics [40, 67, 191]. We pursued an approach able to type the control region (CR) variation. Amplification of the entire control region was performed in a single PCR assay and independent sequence analysis (with internal primers) was followed to generate a consensus

sequence. Therefore, the entire control region – between positions 16024 and 576 – of the mitochondrial DNA was sequenced and analyzed.

Primers are named according to the mtDNA chain and position they interact with. Therefore, L correspond to the Light chain and H to the Heavy chain and are followed by the nucleotide position in which the 3' end of each primer binds. All primers used during the practical work are presented in Table 1.

Primer	Sequence (5'— 3')
L15978	CACCATTAGCACCCAAAGCT
L16268	CACTAGGATACCAACAAACC
L16536	CCCACACGTTCCCCTTAAAT
H016	CCCGTGAGTGGTTAATAGGGT
H036	CCCGTGAGTGGTTAATAGGGT
H159	AAATAATAGGATGAGGCAGGAATC
H649	TTT GTT TAT GGG GTG ATG TGA

Table 1 - Name and sequence of each primer used in the present work.

Amplification

The PCR is an *in vitro* method for the enzymatic synthesis of specific DNA sequences, using two oligonucleotide primers that hybridize to opposite strands and flank the region of interest in the target DNA [192]. These oligonucleotides anneal to the separated template DNA strands such that the primer 3' OH ends, which are the "growing ends" of a newly synthesized strand, face each other. When the new strand that results from the extension by DNA polymerase of one primer extends past the other primer site, that strand becomes a new template. Because the primer extension products synthesized in one cycle can serve as a template in the next, the number of target DNA copies doubles approximately at every cycle. The process developed by Kary Mullis in the 1980s [193] was named Polymerase Chain Reaction (PCR) and it is characterized by repetitive series of cycles that ultimately results in the exponential accumulation of a specific fragment whose termini are defined by the 5' ends of the primers. Generally speaking, each cycle consists of three main steps which demand an optimal operating temperature, (i) a heating step to separate the two DNA strands (template denaturation), (ii) a primer annealing step, and (iii) a primer extension step at the optimum temperature for the DNA polymerase in question.

The CR of mitochondrial DNA was amplified in a final volume of 10 µl, using 5 µl of Qiagen® multiplex PCR kit, 1 uL of primer mix – L15978 and H649 (Figure 6) at a concentration of 2,5 µM each – 3 uL of deionized water and 1 uL of DNA (1-5 ng). The Qiagen® multiplex PCR kit contains HotStarTaq Plus DNA Polymerase, PCR Buffer with $MgCl_2$ at a concentration of 3mM, and dNTP's (triphosphate

deoxynucleotides) at a 400 µM concentration each. A negative control (using deionized water instead of DNA product) was used in each PCR reaction to ascertain the presence of contamination. The pre-PCR reaction mix was prepared in a separated room from where the DNA extraction product was added to each pre-identified and prepared microtubes to avoid contamination.

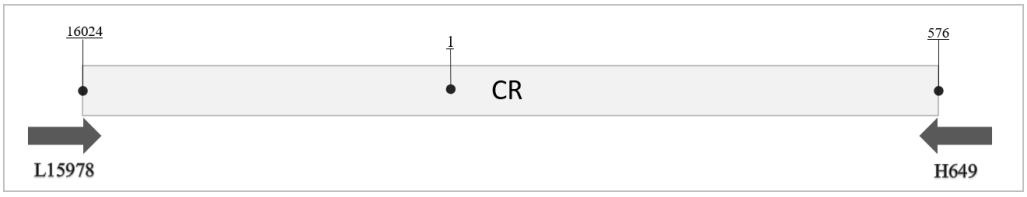


Figure 6 - Control region amplified in this work. Nucleotide positions and primers used are presented. CR meaning control region.

The PCR conditions were an initial denaturation at 95ºC for 15 min, followed by 35 cycles of 94ºC for 30 s, 58ºC for 90 s and 72ºC for 90 s, ending with a final extension at 72ºC for 10 min. Graphic illustration of these temperature cycles are in Figure 7. Minor changes to the described protocol were made to optimize the results in some samples that in a first assay provided unsatisfactory sequences.

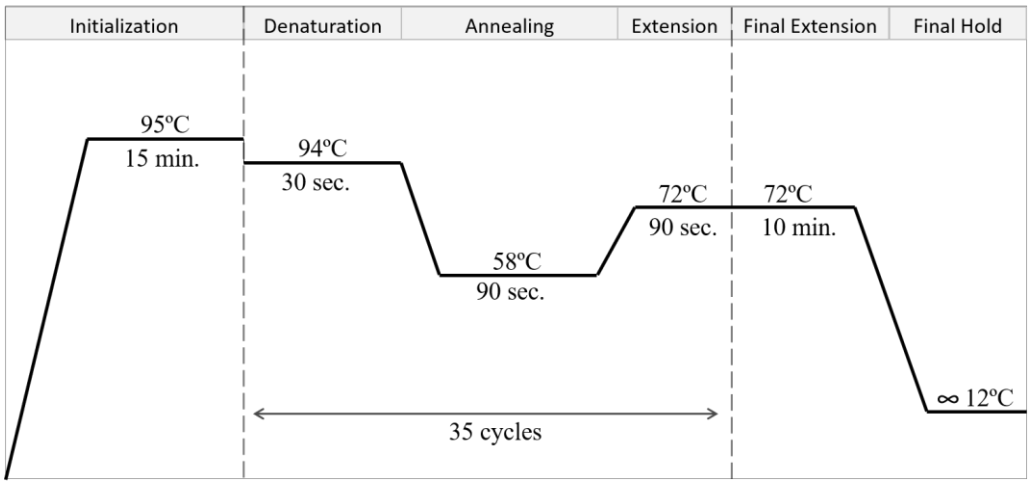


Figure 7 - PCR conditions for DNA amplification using Qiagen Multiplex PCR kit.

The PCR thermocycle used throughout the entire work were the Applied Biosystems' (ABI) GeneAmp™ PCR System 2700 or the Bio-Rad MyCycler Thermal Cycler. After amplification, samples were restored at -20ºC to prevent further evaporation.

The amplification products were confirmed on polyacrylamide gel (T9%, C5%) by silver staining method. To identify the size of each amplified fragment, the GRS Ladder 100bp (GRiSP) was used as pattern. This step was important not only to confirm if amplification occurred but also to adapt the next steps according to the appearance of the DNA bands, which were indicative of the amount of DNA amplified. In order to obtain sequences with quality it was important to take into account this concentration, avoiding either high or low concentrations in further steps.

The PCR product was purified with an enzyme mixture of Exonuclease I (20U/uL) and FastAP Thermosensitive Alkaline Phosphatase (1U/uL) made at a 1:2 proportion (ExoI:FastAP), which was responsible for the removal of residual/unincorporated primers and unused dNTP's in the amplification process. Thus, 2 uL of ExoI-Fastap were added to 5 uL of PCR product. Two incubations were followed: the first incubation (37ºC for 15 minutes) activates the enzymes, allowing them to digest excess primer and dephosphorylates nucleotides; the second, high-temperature incubation (85ºC for 15 min) inactivates the enzymes. Samples were then restored at -20ºC.

Sequencing

Sequencing is used to determine the exact position of each nucleotide of a DNA segment previously amplified. This method is based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication, and was first developed by Frederick Sanger and colleagues in 1977 [194]. The chain-terminating dideoxynucleotides (ddNTPs) differ from triphosphate deoxynucleotides (dNTPs) by the absence of a hydroxyl group at the end 3', which makes phosphodiester attachment of a new nucleotide impossible. Thus, when the polymerase introduces a new ddNTP in the chain to be synthesized, its growth ends. Each of the ddNTPs is labeled with a fluorescent dye that allows its detection and identification during the analysis: ddTTP (thymine) is marked in red, ddCTP (cytosine) in blue, ddATP (adenine) in green and ddGTP (guanine) in yellow, although it is displayed in black, in order to facilitate the visualization of the results obtained.

The forward primers of choice were L15978 and L16536. Reverse primers were used in cases which some nucleotide position in the sequence needed clarification or in cases witch changes in sequence cause length heteroplasmy: in these cases, the reading was not possible (or partially not possible) due to overlap of nucleotides that follow such changes. Long tracts of (more than seven) cytosines ('long C-stretches') regularly cause slippage in amplification, so beyond such a long C-stretch reading is inhibited by an overlay of shifted sequences.

Sequencing was achieved with the BigDye Terminator v3.1 cycle Sequencing kit (Applied Biosystems), which has dNTP's, ddNTP's (marked with Dye Terminator) and the DNA polymerase. For a final volume of 5 uL, it was used 0,8 uL of BigDye, 0,6 uL of the Sequencing Buffer 5x, 0,5 uL of the primer in question (2,5uM), 2,4 uL of deionized water and 0,5 uL of the amplificated DNA. The sequencing process was performed in a thermocycler with prior programming of the temperature and time ranges – an initial denaturation at 96 ºC for 2 min, followed by 35 cycles of 96 ºC for 15 s, 55 ºC for 9 s and 60 ºC for 2 min, ending with a final extension at 60 ºC for 10 min (Figure 8). Some changes to the described protocol were necessary to optimize the results, namely for the different primers.

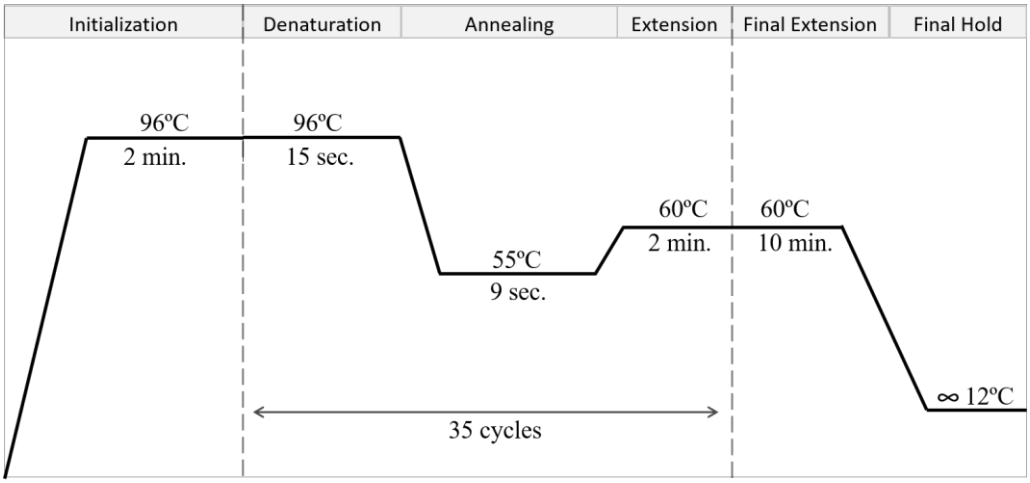


Figure 8 - PCR conditions for DNA sequencing using the BigDye Terminator v3.1 cycle Sequencing kit.

Final purification was completed using the cross-linked dextran gel Sephadex G-50 Fine DNA Grade (GE Healthcare) by gel filtration in spin columns, to purify DNA from small molecules by size exclusion. To the final purified DNA, formamide was added before running samples on the Genetic Analyzer.

Sequence detection

Size separation and detection of the fragments was performed by capillary electrophoresis in a Genetic Analyzer 3130 (AB) sequencer at the I3S facility – The Genomics platform (GenCore). The

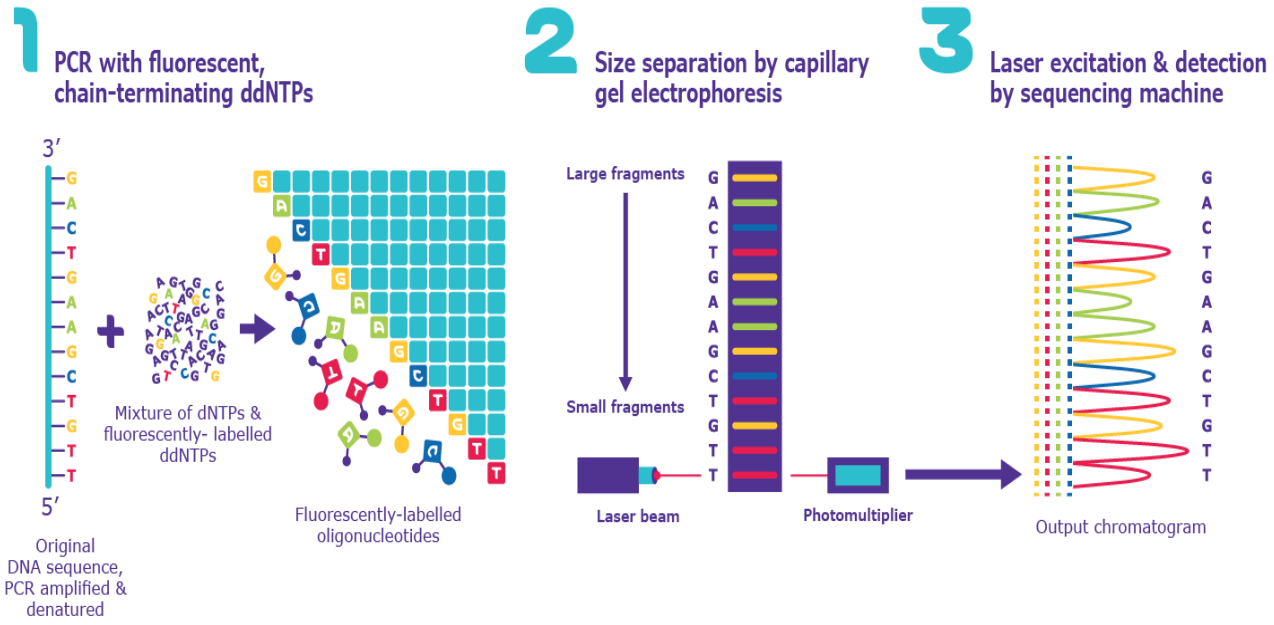


Figure 9 - Sanger Sequencing, capillary gel electrophoresis and sequence detection methods used in the present work. Source: www.sigmaaldrich.com

different fragments migrate according to their molecular weight, with those with a lower molecular weight going through the capillaries at a higher speed than those with a higher molecular weight. Detection was made by the presence of ddNTPs labeled with fluorochromes at the end of the chain, which, when excited by a laser beam incident on the capillary, emitted fluorescence. This was represented by a peak, with a color corresponding to the excited ddNTP, in the electropherogram (Figure 9).

3.3. Data analysis

The sequencing output – electropherogram – was first opened with the Sequencing Analysis v.5.4. software, which allowed to assess the quality of sequences by identifying the nucleotides in each position. This software was an important tool as a preliminary approach because it helped understand which sequences and specific regions were correctly/undoubtedly identified and which sequences needed clarification, so that further/additional steps could be performed.

After obtaining clear and complete CR sequences, the software Geneious v.5.5.8. was used to perform the alignment and comparison of those sequences with the rCRS (rCRS oriented version of Phylotree). The Geneious alignment between each sequence and the rCRS allowed the identification of substitutions, insertions, and deletions, that were then annotated for further classification. The set of the characteristic polymorphisms of a mtDNA sequence belonging to a specific individual gives rise to the haplotype. Its characterization must be carried out based on internationally standardized rules in order to ensure uniformity in the study of mtDNA. In this sense, the guidelines described by the European DNA Profiling Group (EDNAP) [195] and the nomenclature rules were followed according to the standards of the International Union of Pure and Applied Chemistry (IUPAC). Samples and their polymorphisms are presented in Appendix 1, divided into the two Andean populations.

Annotated polymorphisms for each sample were then searched on EMPOP, a (semi-)automated software solution for haplogrouping based on Phylotree. EMPOP worked as quality-control tool and indicated the samples' haplogroup assignment or estimation [52, 70].

Phylogenetic and statistical analysis

The haplogroup and haplotype frequencies, the number of shared and unique haplotypes, as well as the genetic diversities of the studied populations were calculated using Arlequin v. 3.5.2.2 Software [196]. Native and Non-Native mtDNA haplogroups proportions were calculated by direct counting.

Fixation index (F_{ST}) values, which evaluates the genetic distances between populations, and the p-value, which indicates the significance level (0.05), were obtained using Arlequin v. 3.5.2.2 Software. To estimate the interpopulation genetic distances, the samples from the current study as well as samples from other Andean Colombian departments and from other Latin American countries selected from the

literature, were used. The results of the pairwise genetic distances analyses were visualized in a two-dimensional graphic through the Multi-Dimensional Scaling (MDS) analysis included in the software STATISTICA 10 (www.statsoft.com) [197].

Genetic structure was calculated with AMOVA (analysis of molecular variance) [198] using Arlequin v. 3.5.2.2, considering geographic classification to group populations.

The Network 10.2 software (Fluxus Technology Ltd.) [199] was used to investigate genetic relationships within specific haplogroups, and median-joining networks were constructed.

4. Results and Discussion

4.1. mtDNA diversity in Tolima and Cundinamarca

The whole mtDNA control region haplotypes observed in the studied population groups, together with their classification in haplogroups, are described in detail in Appendix 1.

In a total of 53 Colombians analyzed in this study, 48 CR different haplotypes were found. They encompassed 43 haplotypes that were unique and 5 that were shared between two individuals, resulting in a haplotype diversity of 0.9964 +/- 0.0045 and a mean number of pairwise differences among sequences of 14.229318 +/- 6.480534.

Assuming a territorial division of the samples based on the respective departments of Colombia, 31 samples belong to the Cundinamarca department and 22 belong to the Tolima department, resulting in a haplotype diversity, for each department separately only slightly lower than in the pooled sample (Cundinamarca: 0.9957 +/- 0.0095; Tolima: 0.9957 +/- 0.0153). These high levels of haplotype diversity suggest that no important founder effects appear to have been experienced by any of the Columbian populations under study.

Within Cundinamarca, two individuals shared a A2ac haplotype and other two shared a A2+(64)+@16111 haplotype, whereas within Tolima two individuals had the same C1d+194 haplotype. Additionally, two different sequences belonging to haplogroup A2+(64), were shared between one individual from Tolima and other from Cundinamarca. It is remarkable this residual level of haplotype sharing between two populations from so close geographic proximity as Tolima and Cundinamarca are. Absence of shared haplotypes between Colombian populations from Antioquia and Cauca had been previously reported by Xavier, C. and colleagues [184]. However, in the later study, only Native-American populations were sampled, and knowing that native populations usually have reduced population sizes being so more prone to drift effects inducing substantial genetic differentiation, the absence of shared haplotypes is understandable. Since Tolima and Cundinamarca are both admixed populations, the scarcity of identical mtDNA haplotypes might still reflect substantial genetic differentiation between the ancestral groups of Native Americans from which derived the current-day admixed populations from Colombia. Several lines of evidence indicate that, in diverse regions from America, geographic structure was rapidly established after the initial human occupation and was thereafter followed by limited gene flow between populations [105]. Furthermore, data provided by mtDNA indicate that the European colonization was followed by local mass, rapid and intense decline of Native-American populations, leading to extinction of lineages especially in the major population centers of the pre-Columbian past [105], in a time-horizon concomitant with the fast growing of admixed populations. Thus, the consequent drift effects might have accounted to the differentiation of current-day admixed populations.

Comparing the levels of haplotype diversity estimated in the current study with those obtained in previous works relying in identical resolution power, that is based on the variation of the entire CR, the observed values are higher when compared to other mixed and native populations from the Andean region of Colombia [183, 184]. Furthermore, some South American countries such as Peru [186], Argentina [188] and Paraguay [187] have presented lower diversity indices, while Ecuador [200] and other three central east cities of Brazil [201-203] show an even higher haplotype diversity (Table 2).

Geographic location	Haplotype diversity	References
Cundinamarca	0.9957 +/- 0.0095	Current study
Tolima	0.9957 +/- 0.0153	Current study
Norte Santander + Santander + Cundinamarca	0.9932	Castillo, A. et al [182]
Cauca*	0.952	Xavier, C. et al [183]
Antioquia*	0.751	Xavier, C. et al [183]
Peru	0.9134	Simão, F. et al [185]
Argentina north / central / south	0.906 / 0.937 / 0.878	Bobillo, M. C. et al [187]
Paraguay	0.9937	Simão, F. et al [186]
Ecuador	0.9986	Burgos, G. et al [199]
Rio de Janeiro (Brazil)	0.9994	Simão, F. et al [200]
Espírito santo (Brazil)	0.999	Dos Reis, R. S. et al [201]
Brasília (Brazil)	0.9988	Freitas, J. M. et al [202]

Table 2 - Haplotype diversity values from different geographic locations in South America. In bold letters are the samples from this study; shaded in blue are the Colombian departments; *native populations.

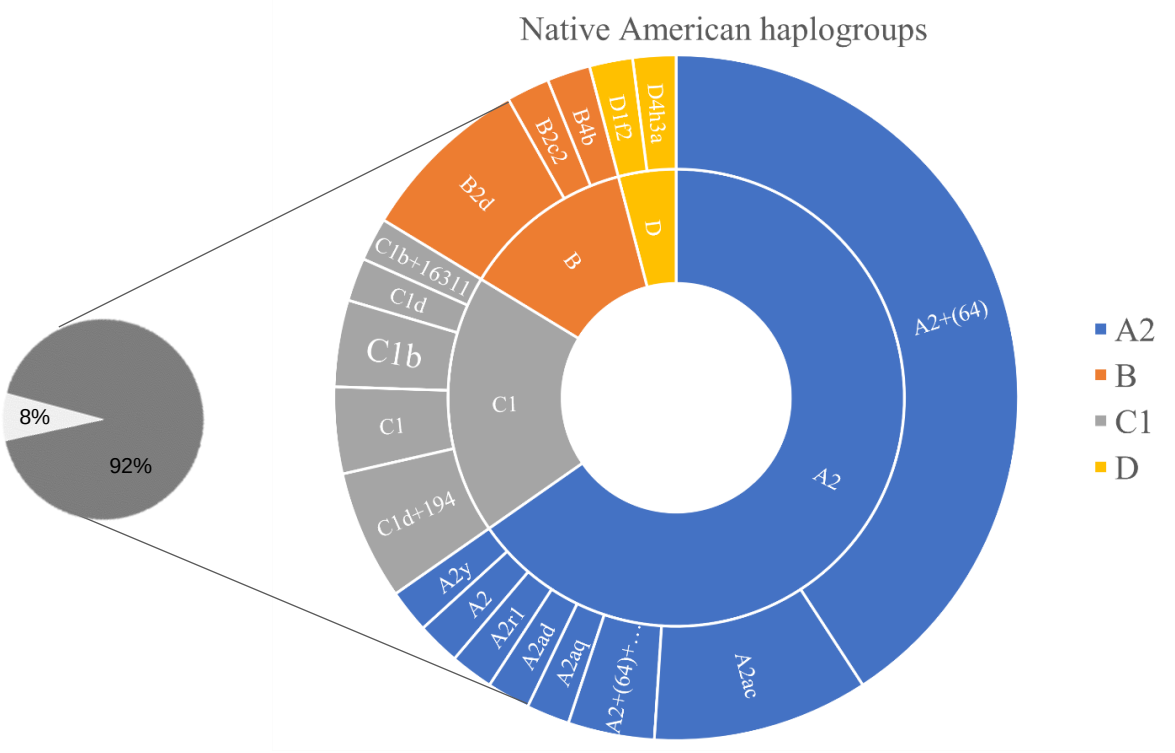


Figure 10 – Pie-chart representation of the Native (92%) and Non-Native (8%) proportions obtained for the studied samples. Donut-chart representation of the Native American haplogroups distribution from the studied samples.

The majority of the detected haplotypes (92.4%) belong to Native American haplogroups, specifically to the branches A2, B2, B4, C1, D1 and D4. The remaining (7.6%) belong to the Eurasian haplogroups J2a2b1, M74a, U5b2b3 and R0. No haplogroups of African origin were detected in the studied individuals (Figure 10).

Assuming the whole component of Native American ancestry, the observed results are in consonance with those, for instance, reported by Castillo, A. and colleagues [183] for a Colombian Andean sample containing mostly individuals from Santander (North-East Andean region), among whom 91% of mtDNA haplogroups were Native American.

In Figure 11 is presented the haplogroup distribution detected in the subsamples of Tolima (Center-West Andean region) and Cundinamarca (Center Andean region).

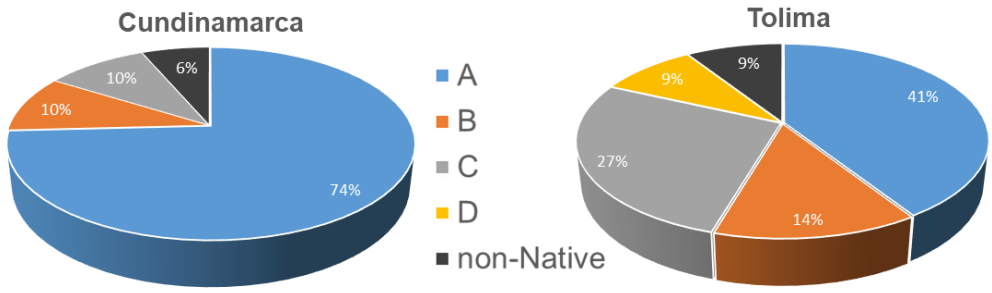


Figure 11 - Donut-chart representing the Amerindian haplogroups A, B, C and D distribution for each studied department, as well as the non-Native haplogroup "slice" present in each department.

Concerning the Native-American haplogroups, there are clear differences in the frequencies found in Tolima and Cundinamarca. While in Cundinamarca, D lineages are absent and the repertoire of Native-American lineages is overtly dominated by those belonging to haplogroup A (74%), in Tolima the four main representatives of Native-American haplogroups (A, B, C and D) are present, and despite the most frequent being also the A, it only reaches 41%, followed by C (27%), B (14%) and finally D (9%). Accordingly, the genetic distance between Cundinamarca and Tolima was statistically significant (F_{ST} =0.05042; p=0.009).

Again, the high differentiation between the two subsamples could be seen as uncommon given their geographical proximity: Tolima is bordered on the West by Cundinamarca and both departments are located in the valley of the Magdalena River, which is the main hydrographic system and the most important waterway in Colombia that represented a major development axis of the population scenario in the region since the establishment of the first human settlements until present.

Yet, such differentiation fits well the pattern of strong substructure that nowadays is widely documented for Native-American or admixed populations from South America. In fact, global patterns of variation in the Americas reveal that most of the genetic variation occurs within populations rather than

among linguistic or ecoregional groups, and that isolation by distance is barely detectable in most population sets, even in those involving populations located in close geographical proximity [181].

Most native American haplogroups here detected are included in what are assumed to be the maternal founding lineages of Asian/Beringian origin, among which eight (A2, B2, C1b, C1c, C1d, C1d1, D1, and D4h3a) are often defined as “pan-American,” as they are found across populations all over the continent [77]. However, there are also sub-branches within the pan-American haplogroups that instead have arose later, after the front of the expansion wave had already passed through, and thus remained mostly confined to the geographic area where each arose, and that might have been associated, at least in part, to the early demographic dynamics of the Paleo-Indian settlers [77].

The haplogroup A2 represents the majority of the native dataset of the present study (60.4%) being the most diverse concerning its sub-clades. Represents almost 75% of the dataset from Cundinamarca and 41% from Tolima. A2+(64) is the main sub-clade follow by A2ac, the later with a much less strong representativeness and restricted to Cundinamarca. A search in Empop ([52] accessed 20/09/2021) reveals that A2+(64) is dispersed within the entire American continent especially Central America; the sub-haplogroup A2ad is found in Central America; A2y is restricted to the northern part of South America, whereas the A2ac and A2aq are only found in Colombia so far. In Central America, the haplogroup A2 is indeed the major haplogroup: in populations from El Salvador [204], Nicaragua [205] and Guatemala [206] the notably high frequencies for haplogroup A2 of 91%, 74% and 75% respectively, have been reported. Thus, the high frequencies of haplogroup A2 observed in Colombia might still reflect the result of prehistorical immigrations from Central America that contributed to the initial peopling of South America.

C1 is the second major haplogroup (17%) followed by B (11.3%) and, in a less extent, D (3.8%). Within C1, all of the sub-clades are virtually dispersed through the entire American continent with the exception of the haplotype belonging to C1b+16311, here found in one individual from Cundinamarca, which up to now had been restrictively detected in North America [52]. Nevertheless, its presence in Colombia is not hard to explain given the geographic role of Colombia as a gateway to north-south movements of people.

A search in Empop ([52] accessed 20/09/2021) also revealed that, within haplogroup B, the sub-haplogroup B2d, appears to be virtually dispersed through the American continent but with a more prominent presence in Central America and in the northern part of South America. Contrarily, the sub-haplogroup B4b, found here in one individual from Cundinamarca, is widely dispersed throughout the continent with additional sporadic presence in Asia. The sub-haplogroup B2c2 is restricted to the American continent with a major presence in the North region, namely in the United States. The sub-haplogroups within haplogroup D found in this work were D4h3a and D1f2, both in Tolima. Based on the Empop search, while the first one is found dispersed through the American continent especially South region, the second seems to be restricted to Colombia.

In respect to the frequency distribution of the native American haplogroups here obtained, the proportions are close to those found in Colombia in previous studies, such as in Rojas, W. et al [176], where haplogroup A had the highest frequency (37%), followed by haplogroup B (33%) and haplogroup C (15%) in the combined sample and where A (65%) and B (35%) were found specifically in Cundinamarca. Criollo-Rayo, A. et al [182], obtained similar results, with a high frequency of both A and C (> 60%) in the indigenous groups and A and B (> 60%) in the admixed groups. Yet, a slightly different distribution was reported by Yunis, J. J., & Yunis, E. J. [185] revealing higher frequencies of haplogroups D and C and a lower frequency of haplogroup A in the southeast region of Colombia, although haplogroup B is indeed less frequent in the eastern region of Colombia, similarly to what was found in the current study.

Concerning non-Native lineages, J2a2b1 and U5b2b3 were found in two individuals from Cundinamarca. A query in the EMPOP database ([52] accessed 22/09/2021) of the haplotype J2a2b1 revealed no match but very close neighboring haplotypes were present in populations from Southern and Western Europe, as well as in admixed populations from eastern regions of South America (namely Brazil and Argentina); whereas for the U5b2b3, a full match was observed in Southern Europe, and another in a population from southern Brazil. The presence of both lineages in Colombia are likely the result of the major migratory flows to the South American East coast in the nineteenth century onwards, from Europe and particularly from the Iberian Peninsula, following the Spanish conquest. Given the differential male/female admixture ratio as the consequence of the asymmetric pattern of mating involving mostly interbred between immigrant men and native women to establish neo-American populations, this admixture dynamics has largely erased the Amerind male lineages, contrarily to the fate of the mtDNA lineages, which, as we can see in this study, were much more slightly disturbed [163, 164, 207].

In Tolima, R0 and M74a were the unique non-native haplotypes found in the sample. R0 shows a current distribution very wide both in Eurasia and America, but has the higher predominance in Europe and is thought to be of Middle East/Western Asia origin. Its presence in Colombia is certainly due to the massive impact of post-Columbian influx of people. Concerning the M74a lineage, a query in the EMPOP database ([52] accessed 22/09/2021) revealed no match nor neighbor haplotypes. However the 22 M74a haplotypes available in this database were mainly from Southeast Asia, although having sporadic representatives in North America (North-Central United States and South-West United States). So, its presence in Tolima result from different major waves of recent migration from Asia to the Americas, the first of which had as main destiny continental United States, as occurred, for instance, during the California Gold Rush that started in the 1850s. The proliferation in the United States of exclusionary immigration laws in the late 19th, changed the migration flow and entire Latin-America, including Columbia, received many Asian people. Nowadays in Colombia, there are many Asian communities that have experienced considerable population grow in the last decades.

A last note on the absence of mtDNA lineages attributed to be of clear African descent. The finding is in keeping with previous data for American populations, including from Colombia, which have clearly documented the highly sex-imbalanced admixture of different ancestries that emerged with the initial colonial expansion that afforded the genetic contacts between Native American, European, and African individuals. As a consequence, nowadays Colombian populations harbor a pool of maternal lineages that are mainly Native-American followed by those of European origin. African mtDNA haplogroups are very scarcely represented, except in a few small populations, usually referred to as African-Colombian. Comparatively, in the repertoire of male lineages, both European and African haplogroups are much well represented, especially the first [146, 181, 184]. Moreover, the recent analysis of ancestry informative markers in Colombia revealed that the lowest proportion of African ancestry was found in populations from the Andean region [146].

Two networks of mtDNA haplotypes were constructed, one containing the entire set of lineages identified in Cundinamarca and Tolima (Figure 12.1) and the other restrictively with lineages falling in haplogroup A2+(64) (Figure 12.2).

In the first network, haplotypes are very coherently clustered according to their haplogroup class, being possible to sharply discriminate those belonging to haplogroup A, B and C, whereas the remaining haplogroups, all of non-Native-American ancestry excepting D1f2 and D4h3a representative, and only sporadically present in the dataset, are quite divergent in the network and show high differentiation between each other.

In the 3 clusters encompassing lineages belonging to the most common Native-American haplogroups here found (A, B and C), haplotypes from Tolima and Cundinamarca are randomly interspersed, revealing close molecular relationships. The same feature emerges in the network with the A2+(64) haplotypes.

The part of the network encompassing the A haplotypes has a clear star-like shape, with a central haplotype from which radiate a number of one mutational step haplotypes. This network structure seems to indicate that the A lineages underwent considerable expansion in loco, probably in a broad region where Tolima and Cundinamarca are located, as the Colombian Andean region, since there is no sign of evolutionary differentiation between the A lineages identified in the two populations. So, it appears that both Tolima and Cundinamarca are subsamples of a precedent pool that accumulated differentiation along the time.

Some reticulation seen in the part of network with the A haplotypes, most probably would be resolved if the sample size of A lineages was increased.

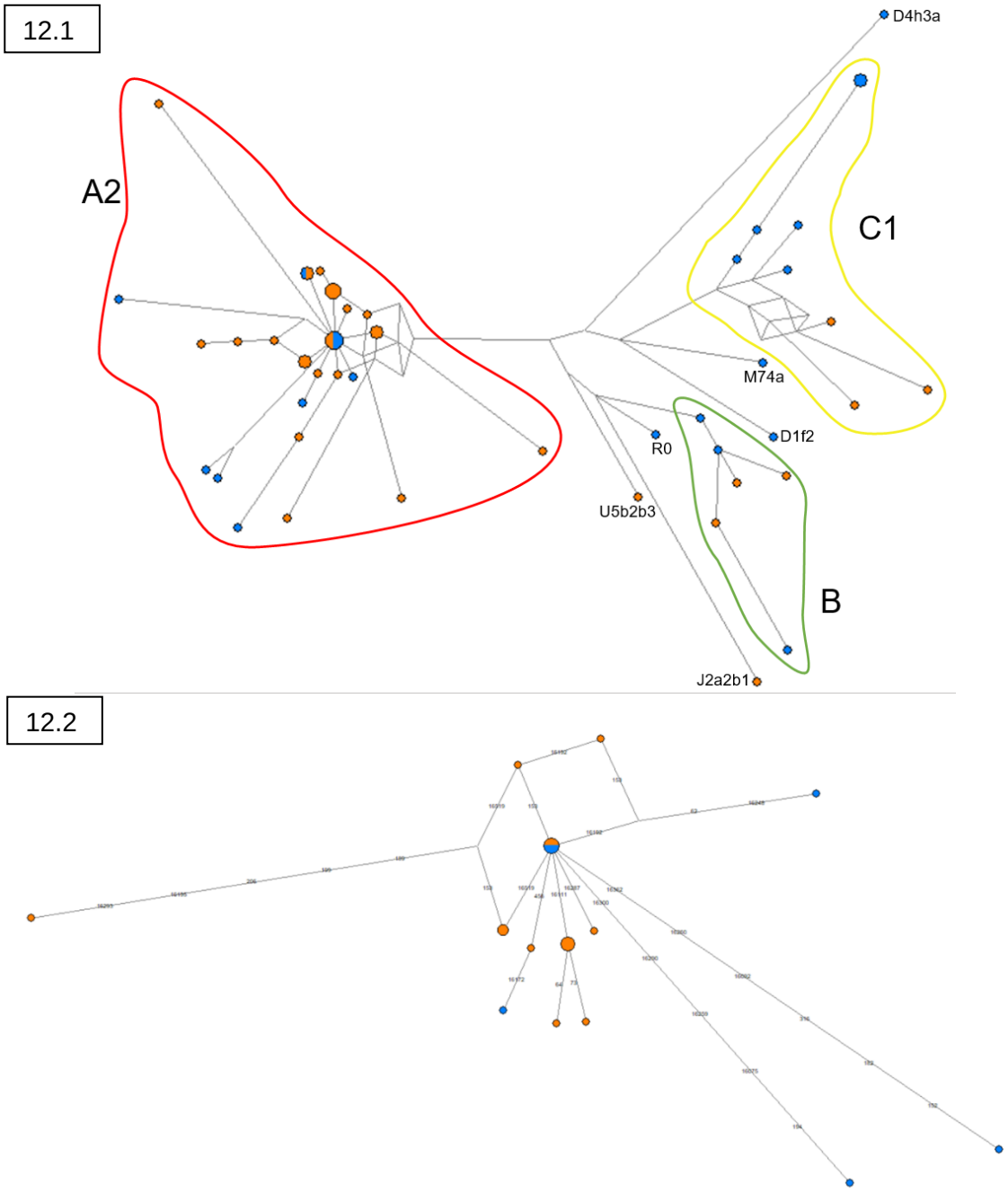


Figure 12 - Median-joining networks of haplotypes present in studied samples from Cundinamarca (in orange) and Tolima (in blue). (12.1) From the entire dataset (53 sequences), 44 haplotypes were found (disregarding indels); (12.2) 20 sequences were used and 14 haplotypes were found (disregarding indels); diameter size is proportional to haplotype frequency.

4.2. Comparative Analysis

4.2.1. Colombian populations

In order to make comparisons between populations belonging to different departments of the Andean region of Colombia and between different Andean sub-regions we have compiled, from different sources, a total of 356 samples with data of the entire CR – disregarding indels in the following positions: 309, 315, 523, 524, 573 and 16193. From the 356 total, 53 samples were the ones obtained in the

current study, while 96 samples were found available on the published work of Xavier, C. et al [184], 94 samples belong to the data available from the 1000 Genomes Project [208], 67 of them were published in Castillo, A. et al (2019) [183] and the last 47 were produced by Simão, F. (non-published data).

Samples were distributed along 7 different departments of Colombia, all from the Andean region: Norte de Santander, Santander, Boyacá, Antioquia, Cauca, Cundinamarca and Tolima (Figure 13). Sequences from Norte de Santander were included in the Santander population group because Norte de Santander was only represented by 7 sequences that, after pairwise genetic distances analysis, were found to be genetically close to those from the neighboring department.

The haplogroup distribution among the considered departments, represented in Table 3 and illustrated in Figure 12, is very differentiated. Haplogroup A predominates in Cundinamarca (71%) and Tolima (43%) as we have already mentioned, but also in Antioquia (46%) and Boyacá (45%). Santander is represented mainly by A (35%) and B (37%) sub-haplogroups as opposed to Cauca that is mostly represented by sub-haplogroups within C (64%). Boyacá has the highest proportion of non-Amerindian haplogroups (17%) follow by Antioquia (11%).

			Haplogroups				
Sources	Population	N	A	B	C	D	non-Native
Current study; Castilho, A. et al [182]; Simão, F.*	Cundinamarca	35	25	3	3	1	3
Current study; Castilho, A. et al [182]; Simão, F.*	Tolima	23	10	3	6	2	2
Castilho, A. et al [182]; Simão, F.*	Santander	57	20	21	7	6	3
Xavier, C. et al [183]	**Cauca**	58	11	4	37	5	1
Castilho, A. et al [182]; Simão, F.*	Boyacá	41	18	13	0	2	7
Castilho, A. et al [182]; Simão, F.*; Xavier, C. et al [183]; 1000 Genomes Project [207]	Antioquia	142	65	44	6	12	15

Table 3 - Haplogroup distribution among the selected Colombian samples for comparative purposes and respective sources. N=number of samples; In bold: native population sample; *Simão, F. (non-published).

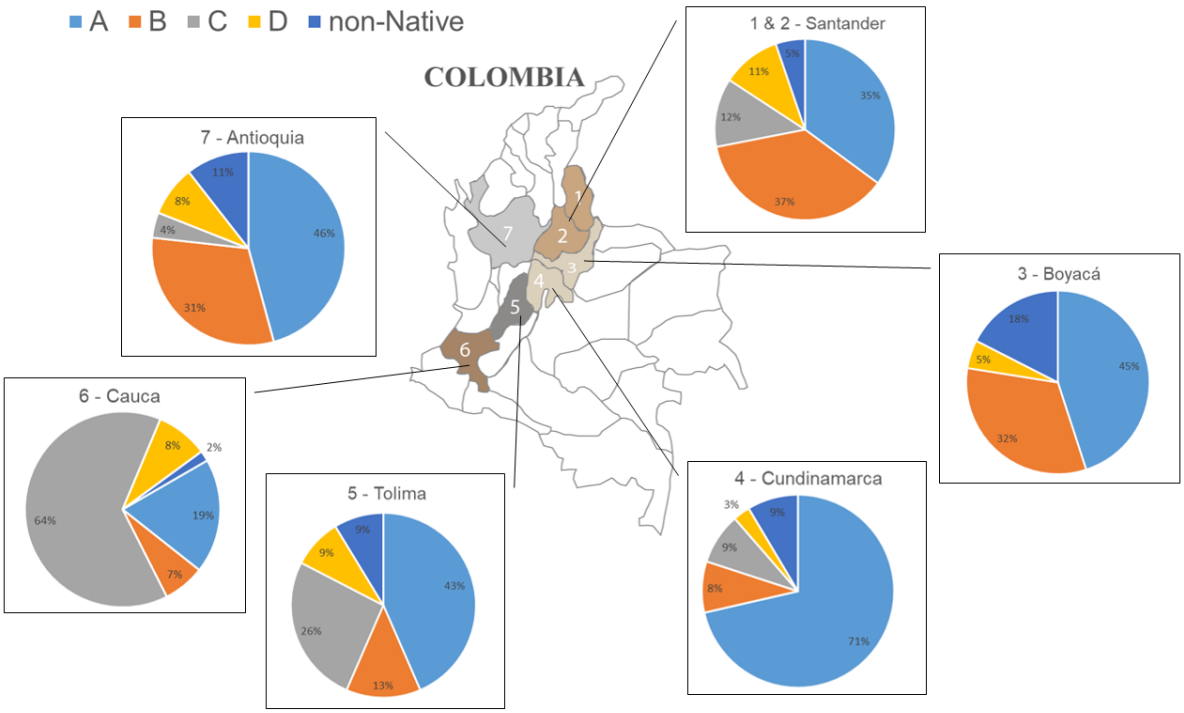


Figure 13 - Haplogroup distribution among the selected Colombian samples for comparative purposes.

Pairwise genetic distances (F_{ST}) and corresponding differentiation p-values, calculated based on CR haplotypes with the maximum common resolution between the considered Colombian Andean departments (see Appendix 4), show that Antioquia and Boyacá (the most Western and Eastern Andean regions, respectivelly) are the ones that present the lowest genetic distances between each other (p>5% associated to the F_{ST} values). Cauca stands out due to the high genetic distances with all populations, which can derive from the fact that the sample only contains Native-Americans, contrarily to the samples from the remaining regions that comprehend admixted populations. Cundinamarca is also genetically quite well differentiated from all samples. Both Cundinamarca and Cauca present statistically significant distances in the pairwise comparisons with all the other departments. Santander, the northern Andean region, shows scarce differentiation from Boyacá and Antioquia and a more modest affinity with Tolima.

The values from the pairwise genetic distances were recruited to construct a two-dimensional plot using MDS analysis (Figure 14). The plot clearly reveals the overall high differentiation between Colombian populations, and further shows that Cundinamarca, Tolima and Cauca which are the most southern departments analyzed, are positioned oppositely to the northern samples from Antioquia, Boyacá and Santander, which integrate the cluster of populations with the highest genetic affinities.

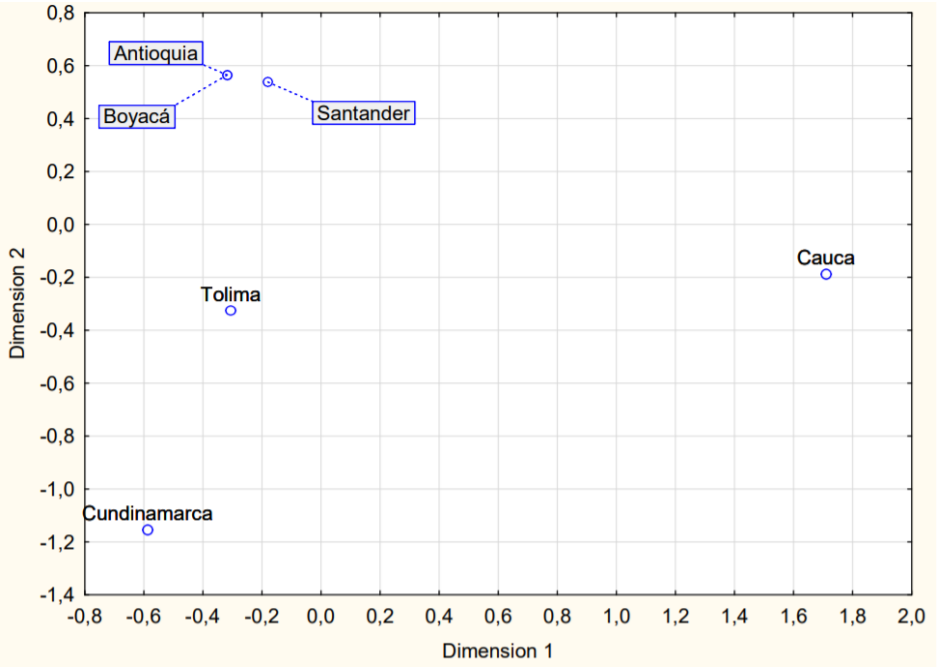
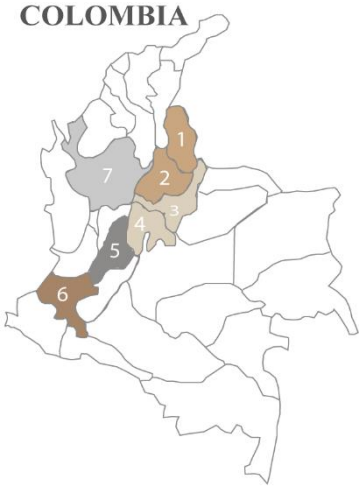


Figure 14 - Multi-Dimensional Scaling (MDS) plot made from the pairwise genetic distances between the Andean departments. *stress*=0,0000046

Next, different geographic classifications were assayed to better understand the genetic patterns present in the Andean region. We started by grouping the populations into 5 Andean Sub-regions (defined elsewhere [146]), as described in Table 4. Then, an East/West sub-division based on the Magdalena River bed, was considered: East including (1) Norte Santander, (2) Santander, (3) Boyacá and (4) Cundinamarca; and West including (5) Tolima, (6) Cauca and (7) Antioquia. As well as a North/South sub-division with (1) Norte Santander, (2) Santander, (3) Boyacá and (7) Antioquia forming the Northern region; and (4) Cundinamarca, (5) Tolima and (6) Cauca forming the Southern region.



Andean Sub-region	Colombian Department	N
North-East	1 - Norte de Santander	7
North-East	2 - Santander	50
Central-East	3 - Boyacá	41
Central-East	4 - Cundinamarca	35
South-East	5 - Tolima	23
South-West	6 - Cauca	58
Central-West	7 - Antioquia	142

Table 4 - Colombian samples belonging from seven different Colombian departments grouped in five Andean Sub-regions. N = number of samples.

The percentage of Native and non-Native haplogroups from the subset are illustrated in the Appendix 5. Non-Native haplogroups are low frequent, ranging from 2% in the South-West population to 13% in the Central-East population. Considering an East/West sub-division, non-Native haplogroups represents 10% of the East group and 7% of the West group. Similarly, low levels of non-Native haplogroups have been systematically reported for populations from Colombia, which actually might reflect some bias in the sampling process that is difficult to scrutinize. But, for instance, in the 102 mtDNA lineages belonging to Colombians collected by Bisso-Machado, R., & Fagundes, N. J. (2021) [181] from 14 studies, all haplotypes were included in the typically Native American haplogroups (A, B, C and D). Even considering the entire dataset examined by the authors (American continent), the presence of mtDNA genomes of probable non-Amerindian origin was in general rare.

To explore how geography influenced the genetic structure of the Colombian population groups, AMOVA was performed and the results are shown in Table 5. AMOVA results show that 87,06-90,90% of the differences observed are explained by variation within populations but variation among populations within groups (4,69-13.30%) is also statistically significant. Contrarily, when grouping was made based on the different geographic clustering (Sub-Andean regions, East/West-Andes and North/South-Andes) none of the three F_{cT} values were statistically significant. This means, that the geographic criteria used did not captured any significant structure in the degree of diferenciation, i.e, it does not represent a factor with weight in determining the general structure of the populations.

Perhaps considering ethno-linguistic afilliations to classify groups of populations, the AMOVA results would reveal to be different, but that could not be confirmed because information on ethnic group and language family was not available for all the samples used in this study. Anyway, in the recent work conducted in Native American populations by Bisso-Machado, R., & Fagundes, N. J. [181], both language and geography were used to assess structure in the Continental or sub-continental patterns of variation, leading to conclude that variation among linguistic or ecoregional groups had a very modest role (although some tests reached statistical significance) in the total diversity of Native-Americans.

Geographic criteria	Number of groups	Among groups (F_{CT})	Among populations/within groups (F_{SC})	Within populations (F_{ST})
Sub-Andean regions	5	6.01 (0.060)*	4.69 (0.049)	89.30 (0.107)
East/West	2	0 (0)*	13.30 (0.127)	90.90 (0.091)
North/South	2	6.32 (0.0631)*	6.63 (0.070)	87.06 (0.129)

Table 5 - AMOVA analysis results considering Colombian populations.
All p-values were statistically significant except when noted with an asterisk (p-value > 0.05).

In this study, no evidence emerged of strong correlation between genetics and geographic affiliation, or, by other words, geography did not appear to constitute a barrier to gene flow.

There are studies reporting that populations living in the same river basin or along closely connected rivers tend to be genetically more similar than those living on different rivers [38]. However, we did not observe any significant trend of differentiation between populations from departments bordered and separated by the Magdalena River.

Regardless of the factors underling the high values of differentiation within the studied Colombian populations/regions, the finding highlights the importance of the elaboration and implementation of detailed genetic forensic databases, representative of diverse geographic regions, linguistic families and ethnic groups. Additional information, for instance from the coding region of the mtDNA, would allow a more detailed analysis of the genetic composition of these samples and consequently of Colombian populations.

4.2.2. South American populations

In order to make comparisons between populations belonging to different South American regions we have compiled, from different sources, a total of 1810 samples with data of the entire CR – disregarding indels in the following positions: 309, 315, 523, 524, 573 and 16193. Besides the 356 Colombian samples; 83 samples are from Lima, Peru; 496 samples belong the Brazilian cities of Rio de Janeiro [201] (n= 205) and Espírito Santo [202] (n=291); another 107 samples belong to Ecuador [209]; Argentina [188] is represented by a total of 338 samples where 98 come from the North-East, 193 from Central-West and 47 from the South region; 325 samples belong to Chile [210], namely North (n=148) and Southern (n=177) regions; and finally, another 105 samples that belong to the Paraguayans of Alto Paraná [211] (see Appendix 6).

The haplogroup distribution among the considered Latin American countries is represented in Table 6 and illustrated in Figure 15. Besides Colombia, the haplogroup A is also dominant among the samples from Ecuador (43%), although the second major haplogroup in this case is D (34%) as opposing to Colombia where D has a minor presence. Alto Paraná in Paraguay has the most homogeneous distribution with both B and C representing 25% each. The samples from Lima in Peru belong mostly to haplogroup B (47%) and the ones from Chile belong mainly to haplogroup C (35%). Notably, the eastern countries of South America are the ones with the major prevalence of non-Native haplogroups: almost 80% of the Brazilian samples and 44% of the Argentines. Additionally, the absence of non-Native haplogroups in the represented population of Ecuador might be to the fact that only ~40% of this sampled population is admixed.

			Haplogroups				
Sources	Population	N	A	B	C	D	non-Native
Current study; Castilho, A. et al [182]; Simão, F.*; Xavier, C. et al [183]; 1000 Genomes Project [207]	Colombia	356	149	88	59	29	31
1000 Genomes Project [207]	Peru (Lima)	83	12	39	15	13	4
Simão, F., et al [200]; Dos Reis, R. S. et al [201]	Brazil	496	35	34	28	5	394
Baeta, M. et al [208]	Ecuador	107	46	16	9	36	0
Bobillo, M. C. et al [187]	Argentina	338	52	33	52	52	149
Gómez-Carballa, A. et al [209]	Chile	325	18	74	113	90	30
Simão, F. et al [210]	Paraguay	105	22	26	26	16	15

Table 6 - Haplogroup distribution among the selected South American samples for comparative purposes and respective sources. N=number of samples; *Simão, F. (non-published)

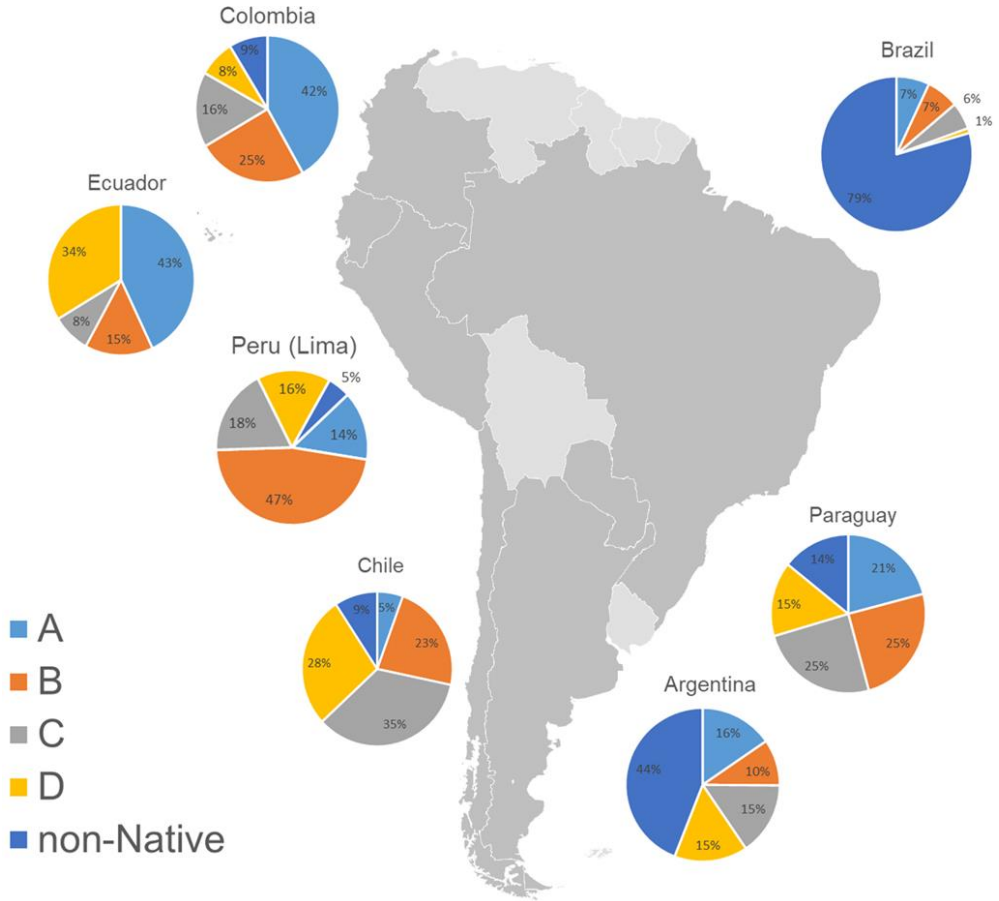


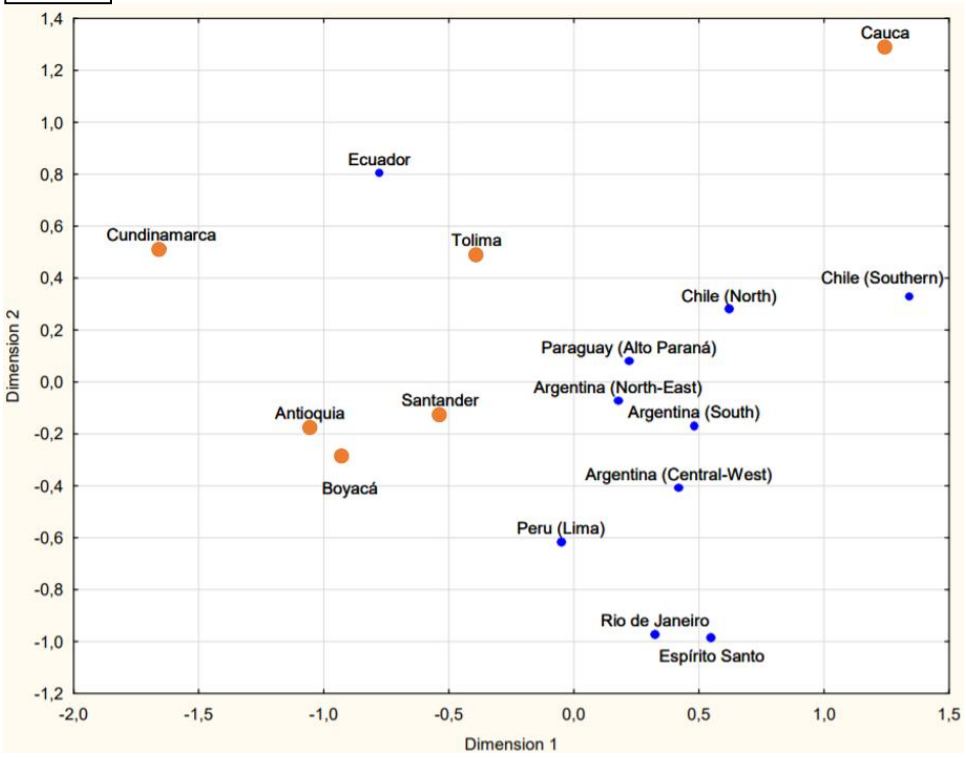
Figure 15 - Haplogroup distribution among the selected South American samples for comparative purposes.

4.2.2.1. Comparisons between admixed/entire dataset

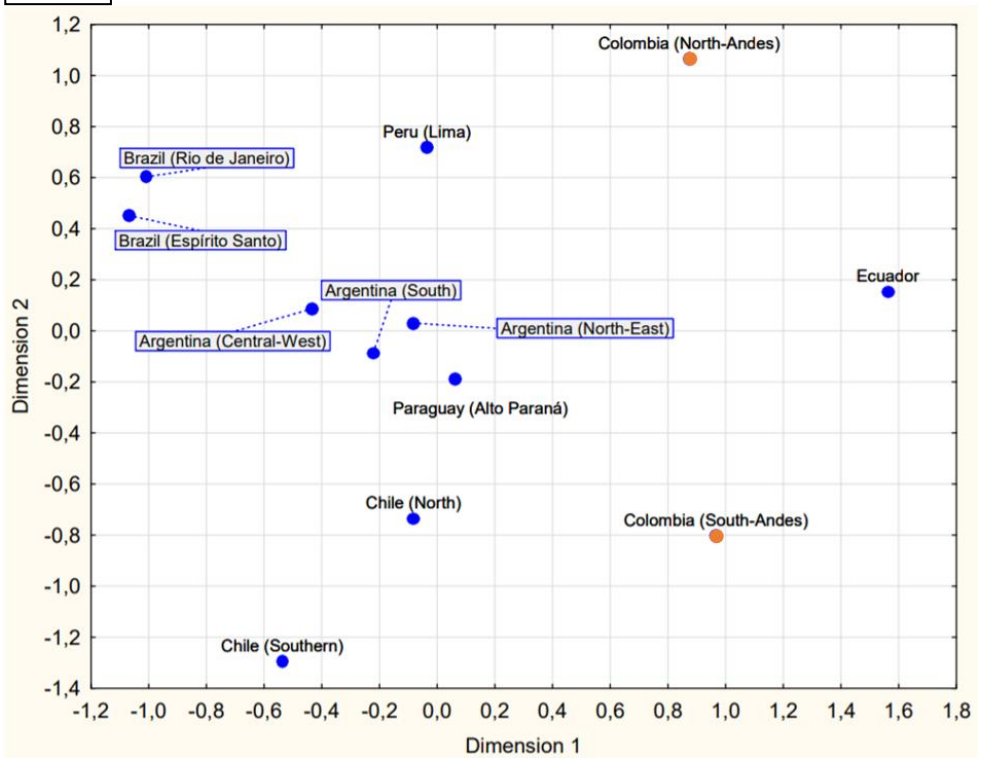
Pairwise genetic distances were calculated based on CR haplotypes identified in our dataset and other Latin American populations (Appendix 7). Most of the populations presented statistically significant F_{ST} values (P<0.05) with all Colombian populations, except the population from Paraguay (Alto Paraná) that do not differ significantly from the population of Tolima as well from the population of North-East Argentina which is less unexpected given their geographic proximity. It is of note that populations belonging to the same country present statistically significant values between each other, as for instance the East-Brazilian populations considered. Belonging to a country, does not says much about the genetics affinities between populations from South America.

Two-dimensional plots (Figure 16) were constructed through MDS using the pairwise F_{ST} distances present in Appendix 7. The first aspect that leaps out in the first MDS plot (Figure 16.1) is the fact that, with the exception of Ecuador, the populations that most differentiate from one another are the ones belonging to Colombia. Taking into account this global South American context, is remarkable the level of heterogeneity among population within Colombia, as we have commented before. Also, when departments of Colombia are grouped in a “North/South” sub-division (Figure 16.2), they appear to be genetically more distant between one another, whereas the “East/West” groups (Figure 16.3) emerge as less differentiated genetically.

16.1



16.2



16.3

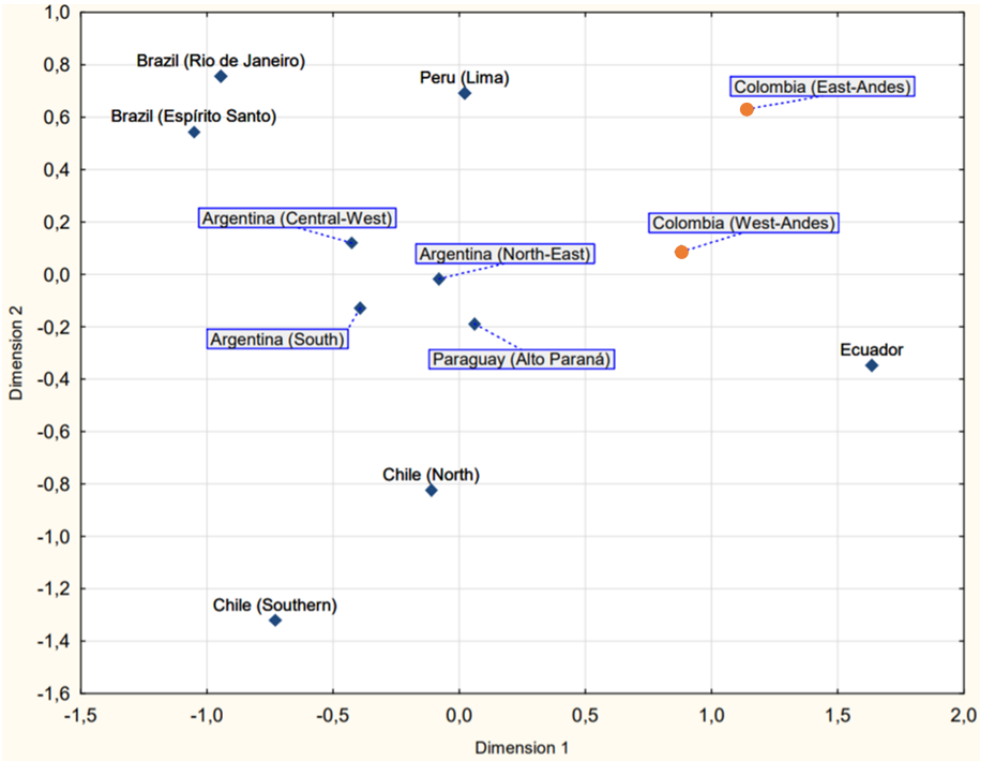


Figure 16 - Two-dimensional plots with MDS analysis. Orange dots represent Colombian populations and blue dots represent the others South American populations. (16.1) Considering each Colombian department as individual variables; *stress*=0,0906167; (16.2) considering the North/South sub-divisions of the Colombian samples; *stress*=0,0905513; (16.3) considering the East/West sub-division; *stress*=0,101473.

To evaluate whether genetic structure within South American populations was influenced by geography, three different geographical criteria were attempted to group populations viewing to perform AMOVA (see Appendix 8). We started by grouping regions belonging to the "West of South America" which included populations from Colombia, Ecuador, Peru (Lima) and Chile; and a second group "East of South America" comprising Brazil, Paraguay (Alto Paraná) and Argentina. Then, we conducted AMOVA with other three major groups: one comprising the countries that the Andes runs through (which extend through the western part of South America) including Colombia, Ecuador, Peru (Lima), Chile and Argentina; and then Paraguay and Brazil as independent groups. We also evaluated another geographic classification assuming three distinctive groups: the Northern-West region of South America (Colombia + Ecuador + Peru) which is also the "Andean Community" (based on a geopolitical sense), the Southern Cone (Chile + Paraguay + Argentina) and then, the Brazilian samples as the eastern group of South America (Table 7).

According to the 3 grouping criteria, the AMOVA results show that 92,46-92,87% of the differences observed are explained by variation within populations but variation among populations within groups (4,07-6,54%) is also statistically significant. Yet, variation among groups only reached a value statistically significant (F_{CT} =3,46%; p-value=0.02737+/-0.00452) when the sub-division "Northern-West / Southern Cone / Brazil" was made.

Geographic criteria	Number of groups	Among groups (F_{CT})	Among populations/within groups (F_{SC})	Within populations (F_{ST})
East / West of South America	2	1.80 (0.0179)*	5.75 (0.0585)	92.46 (0.0754)
Andean States / Brazil / Paraguay	3	0.58 (0.0058)*	6.54 (0.0658)	92.87 (0.0712)
Northern-West / Southern Cone / Brazil	3	3.46 (0.0345)	4.07 (0.0421)	92.48 (0.0752)

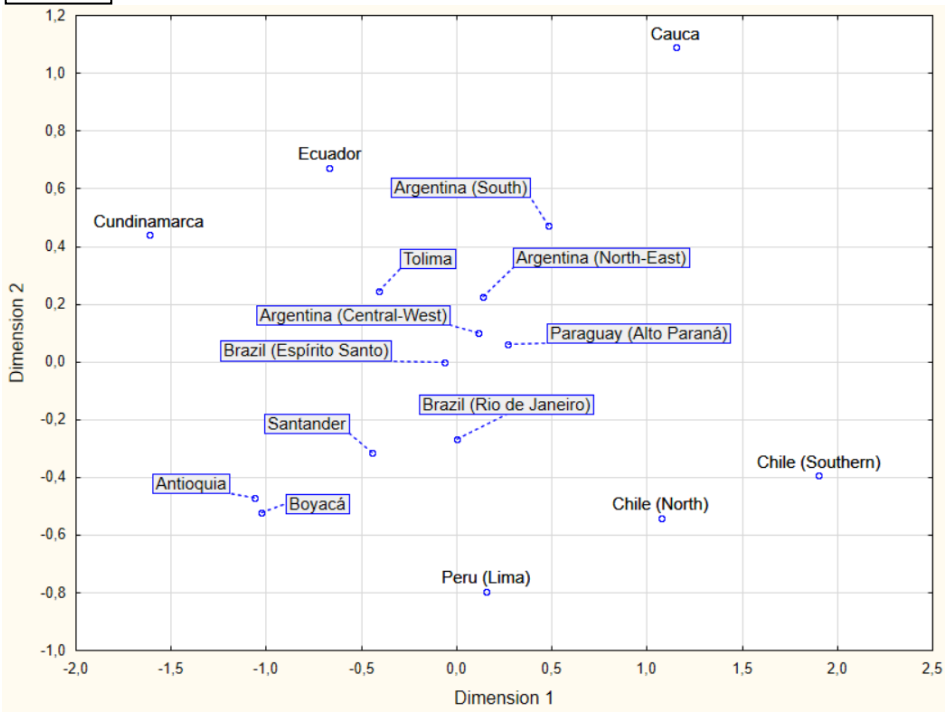
Table 7 - AMOVA analysis results considering South American populations.
All p-values were statistically significant except when noted with an asterisk (p-value > 0.05).

However, this pattern of sub-structuring detected in South American populations, seemed to mirror considerably the differential proportions of non-Native mtDNA lineages, which indeed were quite disparate among the populations, and peaked by far in the Brazilian samples. For this reason, all analyses were redone taking only in account Native-American haplogroups.

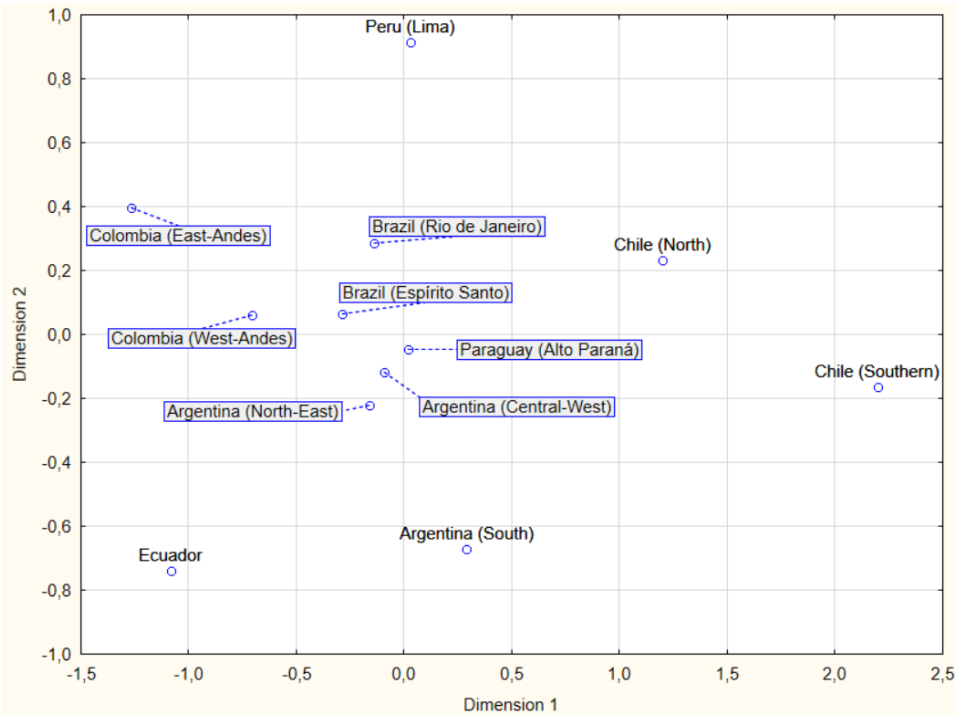
The new pairwise genetic distances based on CR Native haplotypes (Appendix 9) were used to construct the two-dimensional MDS plots shown in Figures 17. AMOVA was again performed considering the same three geographical criteria used in the previous analysis (Table 8; Appendix 10).

According to the 3 grouping criteria, the AMOVA show that 90,55-94,66% of the differences observed are explained by variation within populations, which are results that basically replicate those obatined with the whole set of lineages. However, the proportion of variation among populations within groups increased substantiaily reaching almost double the previous values (F_{Sc} values: 6,93-11,13%). Contrarily the F_{cT} values decreased, and none was statistically significant, meaning that the geographic criteria assumed to define the large clusters of populations, do not contribute to explain the pattern of genetic structure across populations from South America.

17.1



17.2



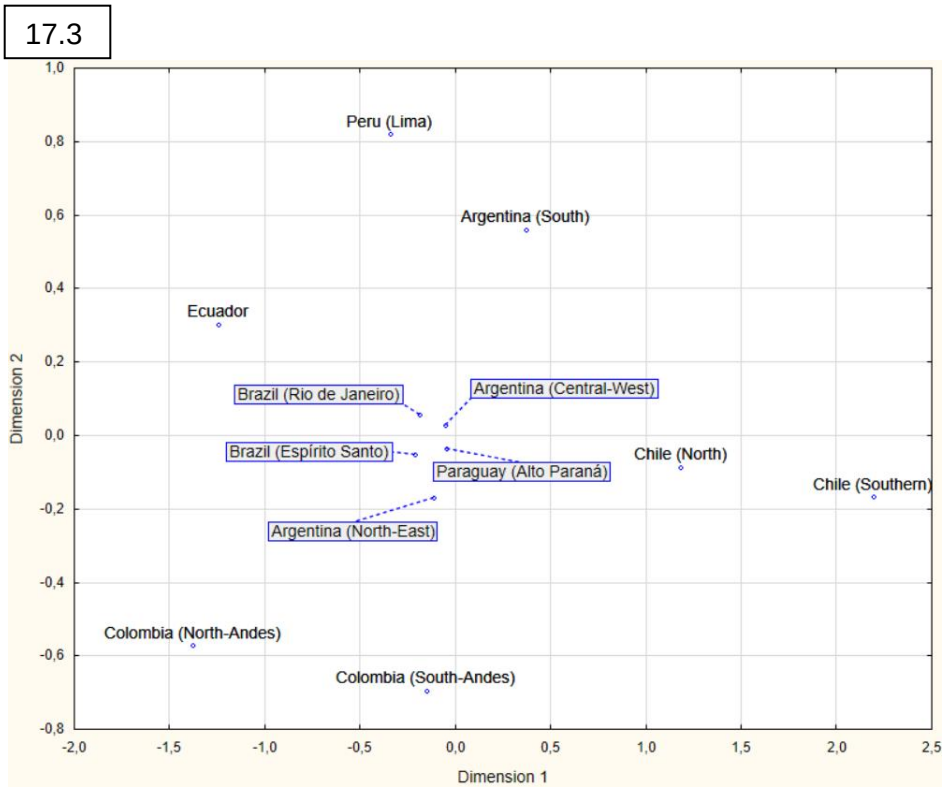


Figure 17 - Two-dimensional plots with MDS analysis. (17.1) Considering each Colombian department as individual variables, *stress*=0,0732014; (17.2) considering the East/West sub-division, *stress*=0,0376780; and (17.3) considering the North/South sub-divisions of the Colombian samples, *stress*=0,0928497.

Geographic criteria	Number of groups	Among groups (F_{CT})	Among populations/within groups (F_{SC})	Within populations (F_{ST})
East / West of South America	2	0 (0)*	10.09 (0.0986)	92.19 (0.0780)
Andean States / Brazil / Paraguay	3	0 (0)*	11.13 (0.1051)	94.66 (0.0533)
Northern-West / Southern Cone / Brazil	3	2.52 (0.0252)*	6.93 (0.0711)	90.55 (0.0945)

Table 8 - AMOVA analysis results considering Native haplogroups from the South American populations considered.
All p-values were statistically significant except when noted with an asterisk (p-value > 0.05).

Two interesting results that became more clear when the Native-American lineages were addressed, are on the one hand the tight affinities between the populations from Brazil, Paraguay and North/Central Argentina (from the Atlantic South America), and on the other the positions that populations from Colombia, Ecuador and Peru (from the Pacific South America), which despite very distant from each other, are arrayed in such way that seems to surround externally the populations from Brazil, Paraguay and North/Central Argentina.

Although it might be hastily to integrate these findings in the context of the models about the peopling of South America, in a certain sense they seem to support the scenario of an initial split of the

first Paleoindian settlers of South America into two main groups whose expansion routes followed different paths – along the Pacific and the Atlantic coastlines. In addition, South Argentina is often placed in the midway between Pacific and the Atlantic populations, suggesting that it might have received multiple influences from quite different and distant populations. Remarkably, a recent study based on ancient mitochondrial genomes from the Argentinian Pampas led to conclude that likely both Atlantic and inland Andean routes have played the major roles for the initial peopling of the region [212].

5. Conclusion

The analysis of the entire control region of mtDNA in two admixed populations from Cundinamarca and Tolima, both located in the Andean region of Colombia, allowed to detect very high values of haplotype diversity within each population, and in addition, in spite of the great geographic proximity, a remarkable level of differentiation between the two samples based on the mtDNA haplogroup distribution harbored by each of them.

The current knowledge on the mtDNA diversity in Colombia Native-American or admixed populations, points to a population heterogeneity in Colombia that is quite unusual, even in the context of South America, where strong population structure has been widely reported.

The significant differentiation between Cundinamarca and Tolima highlights the importance of considering the haplotype distribution profile in each population when evaluating the efficiency of mtDNA marker analysis in forensic case studies.

Native-American haplogroups comprises most of the obtained high diversity, contrarily to what we could have predicted taking into account the mass decimation of most native populations after the arrival of Europeans. This event resulted in massive extinction of native lineages, and so it would not be surprising that the recovered populations would have low levels of diversity. That was not observed, raising the question on the factors (demographic, cultural, other) that permitted to preserve such rich reservoir of native maternal lineages.

All the native lineages belonged to pan-America native haplogroups. The most well represented sub-haplogroup was A2+(64). Given its elevated frequency, it seems a good candidate to obtain more insights concerning its phylogeography. Eventually, a more refined picture on its wider geographic distribution and on its level of internal molecular heterogeneity, can bring new clues on the lineage diversification in pre-Columbus native populations and in those that persisted or arose by admixture after the arrival of Europeans and Africans to America.

Two much less frequent haplogroups were detected that also deserves special attention: A2ac (attaining 16,1% only in Cundinamarca) and A2aq (one individual in Tolima). Since up to now, the 2 sub-haplogroups were uniquely found in Colombia, it would be interesting to explore better both, to realize

weather they represent relic lineages that uniquely persisted in that region and whether they have diversified in loco.

The non-Native component was exclusively composed by lineages of European descent, which only reached a modest frequency in both populations. As a future perspective, the analyses of Y-chromosome diversity in the same two population samples would allow to derive the admixture pattern in male lineages, which then could be contrasted with that here described for the maternal lineages. Doing that, it would be possible to make direct inferences on the level of sex-bias that occurred in the admixture process.

Comparative analysis extended to population across South America, permitted to put the results here obtained for Cundinamarca and Tolima in the broad diversity frame of the continent, and further allowed to demonstrate that geography was a poor predictor of the genetic structure in South-American populations.

In the future, the new mtDNA data provided in this work, still need to be more deeply dissected, and combining this dataset with other already available for contemporary American populations and those that are being rapidly produced through the analysis of ancient mtDNA, all together will help to clarify how the complex population scenario that characterizes nowadays South America arose.

Among the difficulties to achieve that goal, is the reduced sample sizes often used to characterize South America populations. This was also a limitation in this study, that should be overcome in order to obtain a better description of the maternal diversity in Tolima and Cundinamarca, especially regarding the Native-American lineages. The high diversity here detected might be only a rough proxy of the native heterogeneity still carried by people from both regions.

6. References

1. Jd, W. and C. Fh, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-738.
2. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* 2001.
3. Yue, T. and H. Wang, *Deep learning for genomics: A concise overview.* arXiv preprint arXiv:1802.00810, 2018.
4. Chakraborty, R., *Population Genetics: Historical Aspects.* e LS, 2001.
5. Goodwin, W., A. Linacre, and S. Hadi, *An introduction to forensic genetics*. Vol. 2. 2011: John Wiley & Sons.
6. Mendel, J., *Experiments in Plant Hybridization: The Proceedings of the Natural History Society in Brünn*, 1866.
7. Fisher, R.A., *XV.—The correlation between relatives on the supposition of Mendelian inheritance.* Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 1919. **52**(2): p. 399-433.
8. Haldane, J.B.S. *A mathematical theory of natural and artificial selection*. in *Mathematical Proceedings of the Cambridge Philosophical Society*. 1926. Cambridge University Press.
9. Wright, S., *Evolution in Mendelian populations.* Genetics, 1931. **16**(2): p. 97.
10. Okasha, S., *Population genetics.* 2006.
11. Hardy, G.H., *Mendelian proportions in a mixed population.* Science, 1908. **28**(706): p. 49-50.
12. Weinberg, W., *ber den Nachweis der Vererbung beim Menschen.* Jahres. Wiertt. Ver. Vaterl. Natkd., 1908. **64**: p. 369-382.
13. Kimura, M. and J.F. Crow, *The number of alleles that can be maintained in a finite population.* Genetics, 1964. **49**(4): p. 725.
14. Kimura, M., *The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations.* Genetics, 1969. **61**(4): p. 893.
15. Huneman, P., *Special issue editor's introduction:"revisiting the modern synthesis"*, 2019, Springer.
16. Müller, G.B., *Why an extended evolutionary synthesis is necessary.* Interface focus, 2017. **7**(5): p. 20170015.
17. Yang, Z. and B. Rannala, *Molecular phylogenetics: principles and practice.* Nature reviews genetics, 2012. **13**(5): p. 303-314.
18. Baldauf, S.L., *Phylogeny for the faint of heart: a tutorial.* TRENDS in Genetics, 2003. **19**(6): p. 345-351.
19. Gregory, T.R., *Understanding evolutionary trees.* Evolution: Education and Outreach, 2008. **1**(2): p. 121-137.
20. Van Der Wal, C. and S.Y. Ho, *Molecular Clock.* 2019.
21. Hasegawa, M., H. Kishino, and T.-a. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.* Journal of molecular evolution, 1985. **22**(2): p. 160-174.
22. Vasylyeva, T.I., et al., *Integrating molecular epidemiology and social network analysis to study infectious diseases: towards a socio-molecular era for public health.* Infection, Genetics and Evolution, 2016. **46**: p. 248-255.
23. Korf, B.R., *Genetics in medical practice.* Genetics In Medicine, 2002. **4**(6): p. 10-14.
24. Ginsburg, G.S. and H.F. Willard, *Essentials of genomic and personalized medicine*. 2009: Academic Press.
25. Landsteiner, K., *Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe.* Zentralbl. Bakteriol., 1900. **27**: p. 357-362.
26. Ottenberg, R., *Medicolegal application of human blood grouping.* JAMA, 1983. **250**(18): p. 2525-2527.
27. Jeffreys, A.J., V. Wilson, and S.L. Thein, *Individual-specific 'fingerprints' of human DNA.* Nature, 1985. **316**(6023): p. 76-79.
28. Gill, P., A.J. Jeffreys, and D.J. Werrett, *Forensic application of DNA 'fingerprints'.* Nature, 1985. **318**(6046): p. 577-579.
29. Anonymous, *Forensic Science International: Genetics.* 2007. **1**: p. 1-2.

30. Andersson, G., et al., *On the origin of mitochondria: a genomics perspective.* Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 2003. **358**(1429): p. 165-179.
31. Sagan, L., *On the origin of mitosing cells.* Journal of theoretical biology, 1967. **14**(3): p. 225-IN6.
32. Nass, M.M. and S. Nass, *Intramitochondrial fibers with DNA characteristics I. Fixation and electron staining reactions.* Journal of Cell Biology, 1963. **19**(3): p. 593-611.
33. Anderson, S., et al., *Sequence and organization of the human mitochondrial genome.* Nature, 1981. **290**(5806): p. 457-465.
34. Iborra, F.J., H. Kimura, and P.R. Cook, *The functional organization of mitochondrial genomes in human cells.* BMC biology, 2004. **2**(1): p. 1-14.
35. Macaulay, V. and D.M. Richards, *Human mitochondrial DNA and the evolution of Homo sapiens*. 2006: Springer.
36. Robin, E.D. and R. Wong, *Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells.* Journal of cellular physiology, 1988. **136**(3): p. 507-513.
37. Malyarchuk, B.A., et al., *Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region.* Human genetics, 2002. **111**(1): p. 46-53.
38. Arias, L., et al., *High-resolution mitochondrial DNA analysis sheds light on human diversity, cultural interactions, and population mobility in Northwestern Amazonia.* American journal of physical anthropology, 2018. **165**(2): p. 238-255.
39. Ingman, M., et al., *Mitochondrial genome variation and the origin of modern humans.* Nature, 2000. **408**(6813): p. 708-713.
40. Parson, W., et al., *DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing.* Forensic Science International: Genetics, 2014. **13**: p. 134-142.
41. Bandelt, H.-J., M. van Oven, and A. Salas, *Haplogrouping mitochondrial DNA sequences in legal medicine/forensic genetics.* International journal of legal medicine, 2012. **126**(6): p. 901-916.
42. Schwartz, M. and J. Vissing, *Paternal inheritance of mitochondrial DNA.* New England Journal of Medicine, 2002. **347**(8): p. 576-580.
43. Kraytsberg, Y., et al., *Recombination of human mitochondrial DNA.* Science, 2004. **304**(5673): p. 981-981.
44. Giles, R.E., et al., *Maternal inheritance of human mitochondrial DNA.* Proceedings of the National academy of Sciences, 1980. **77**(11): p. 6715-6719.
45. Holland, M.M., et al., *Mitochondrial DNA sequence analysis of human skeletal remains: identification of remains from the Vietnam War.* Journal of Forensic Science, 1993. **38**(3): p. 542-553.
46. Parsons, T.J., et al., *A high observed substitution rate in the human mitochondrial DNA control region.* Nature genetics, 1997. **15**(4): p. 363-368.
47. Brown, W.M., M. George, and A.C. Wilson, *Rapid evolution of animal mitochondrial DNA.* Proceedings of the National Academy of Sciences, 1979. **76**(4): p. 1967-1971.
48. Stoneking, M., *Hypervariable sites in the mtDNA control region are mutational hotspots.* The American Journal of Human Genetics, 2000. **67**(4): p. 1029-1032.
49. Hagelberg, E., *Recombination or mutation rate heterogeneity? Implications for Mitochondrial Eve.* TRENDS in Genetics, 2003. **19**(2): p. 84-90.
50. Meyer, S., G. Weiss, and A. von Haeseler, *Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA.* Genetics, 1999. **152**(3): p. 1103-1110.
51. Kivisild, T., *Maternal ancestry and population history from whole mitochondrial genomes.* Investigative genetics, 2015. **6**(1): p. 1-10.
52. Parson, W. and A. Dür, *EMPOP—a forensic mtDNA database.* Forensic Science International: Genetics, 2007. **1**(2): p. 88-92.
53. Brown, W.M., *Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis.* Proceedings of the National Academy of Sciences, 1980. **77**(6): p. 3605-3609.
54. Cann, R.L., M. Stoneking, and A.C. Wilson, *Mitochondrial DNA and human evolution.* Nature, 1987. **325**(6099): p. 31-36.
55. Johnson, M.J., et al., *Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns.* Journal of Molecular Evolution, 1983. **19**(3): p. 255-271.

56. Avise, J.C., R.A. Lansman, and R.O. Shade, *The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus Peromyscus.* Genetics, 1979. **92**(1): p. 279-295.
57. Kocher, T.D., et al., *Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers.* Proceedings of the National Academy of Sciences, 1989. **86**(16): p. 6196-6200.
58. Nagai, A., et al., *Sequence polymorphism of mitochondrial DNA in Japanese individuals from Gifu Prefecture.* Legal Medicine, 2003. **5**: p. S210-S213.
59. Salas, A., et al., *The making of the African mtDNA landscape.* The American Journal of Human Genetics, 2002. **71**(5): p. 1082-1111.
60. Pereira, L., et al., *Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade.* Annals of human genetics, 2001. **65**(5): p. 439-458.
61. Costa, H.A., et al., *Mitochondrial DNA sequence analysis of a native Bolivian population.* Journal of forensic and legal medicine, 2010. **17**(5): p. 247-253.
62. De Guerra, D.C., et al., *Sequence variation of mitochondrial DNA control region in North Central Venezuela.* Forensic Science International: Genetics, 2012. **6**(5): p. e131-e133.
63. Mairal, Q., et al., *Linguistic isolates in Portugal: insights from the mitochondrial DNA pattern.* Forensic Science International: Genetics, 2013. **7**(6): p. 618-623.
64. Canale, L.C., W. Parson, and M.M. Holland, *The time is now for ubiquitous forensic mtMPS analysis.* Wiley Interdisciplinary Reviews: Forensic Science: p. e1431.
65. Andrews, R.M., et al., *Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA.* Nature genetics, 1999. **23**(2): p. 147-147.
66. Bandelt, H.-J., et al., *The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies.* Journal of human genetics, 2014. **59**(2): p. 66-77.
67. Parson, W. and H.-J. Bandelt, *Extended guidelines for mtDNA typing of population data in forensic science.* Forensic Science International: Genetics, 2007. **1**(1): p. 13-19.
68. Kloss-Brandstätter, A., et al., *HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups.* Human mutation, 2011. **32**(1): p. 25-32.
69. Richards, M. and R. Villems, *The World mtDNA Phylogeny.* Nucleic Acids and Molecular Biology, 2006.
70. Van Oven, M. and M. Kayser, *Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation.* Human mutation, 2009. **30**(2): p. E386-E394.
71. Rosa, A. and A. Brehem, *African human mtDNA phylogeography at-a-glance.* Journal of anthropological sciences= Rivista di antropologia: JASS, 2011. **89**: p. 25-58.
72. Quintana-Murci, L., et al., *Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa.* Nature genetics, 1999. **23**(4): p. 437-441.
73. Torroni, A., et al., *Asian affinities and continental radiation of the four founding Native American mtDNAs.* American journal of human genetics, 1993. **53**(3): p. 563.
74. Stone, A.C. and M. Stoneking, *mtDNA analysis of a prehistoric Oneota population: implications for the peopling of the New World.* The American Journal of Human Genetics, 1998. **62**(5): p. 1153-1170.
75. Schurr, T.G., et al., *Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages.* American journal of human genetics, 1990. **46**(3): p. 613.
76. Stone, A.C. and M. Stoneking, *Ancient DNA from a pre-Columbian Amerindian population.* American journal of physical anthropology, 1993. **92**(4): p. 463-471.
77. Brandini, S., et al., *The Paleo-Indian entry into South America according to mitogenomes.* Molecular biology and evolution, 2018. **35**(2): p. 299-311.
78. Wolpoff, M.H., *Multiregional evolution: the fossil alternative to Eden.* The Human Revolution-Behavioral and Biological Perspective on the Origin of Modern Humans, 1989: p. 62-108.
79. Vigilant, L., et al., *African populations and the evolution of human mitochondrial DNA.* Science, 1991. **253**(5027): p. 1503-1507.
80. Mellars, P., *Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia.* Science, 2006. **313**(5788): p. 796-800.
81. Smith, F.H., *Assimilation model of modern human origins.* The International Encyclopedia of Biological Anthropology, 2018: p. 1-4.

82. Racimo, F., D. Marnetto, and E. Huerta-Sánchez, *Signatures of archaic adaptive introgression in present-day human populations.* Molecular biology and evolution, 2017. **34**(2): p. 296-317.
83. Bräuer, G., *A craniological approach to the origin of anatomically modern Homo sapiens in Africa and implications for the appearance of modern Europeans.* The origins of modern humans: a world survey of the fossil evidence, 1984. **327**: p. 410.
84. Stringer, C.B. and P. Andrews, *Genetic and fossil evidence for the origin of modern humans.* Science, 1988. **239**(4845): p. 1263-1268.
85. Bergström, A., et al., *Origins of modern human ancestry.* Nature, 2021. **590**(7845): p. 229-237.
86. Macaulay, V. and M.B. Richards, *Mitochondrial genome sequences and their phylogeographic interpretation.* eLS, 2013.
87. Maca-Meyer, N., et al., *Major genomic mitochondrial lineages delineate early human expansions.* BMC genetics, 2001. **2**(1): p. 1-8.
88. Tishkoff, S.A. and B.C. Verrelli, *Patterns of human genetic diversity: implications for human evolutionary history and disease.* Annual review of genomics and human genetics, 2003. **4**(1): p. 293-340.
89. Relethford, J.H., *9 Human Population.* A Companion to Anthropological Genetics, 2019: p. 123.
90. Green, R.E., et al., *A draft sequence of the Neandertal genome.* science, 2010. **328**(5979): p. 710-722.
91. Villanea, F.A. and J.G. Schraiber, *Multiple episodes of interbreeding between Neanderthal and modern humans.* Nature ecology & evolution, 2019. **3**(1): p. 39-44.
92. Reich, D., et al., *Genetic history of an archaic hominin group from Denisova Cave in Siberia.* Nature, 2010. **468**(7327): p. 1053-1060.
93. Reich, D., et al., *Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania.* The American Journal of Human Genetics, 2011. **89**(4): p. 516-528.
94. Gokcumen, O., *Archaic hominin introgression into modern human genomes.* American journal of physical anthropology, 2020. **171**: p. 60-73.
95. Meltzer, D.J., *First peoples in a new world: colonizing ice age America*. 2009: Univ of California Press.
96. Skoglund, P. and D. Reich, *A genomic view of the peopling of the Americas.* Current opinion in genetics & development, 2016. **41**: p. 27-35.
97. De Acosta, J., *Natural and moral history of the Indies*. 2002: Duke University Press.
98. Forster, P., *Ice Ages and the mitochondrial DNA chronology of human dispersals: a review.* Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 2004. **359**(1442): p. 255-264.
99. Pitulko, V.V., et al., *The Yana RHS site: humans in the Arctic before the last glacial maximum.* Science, 2004. **303**(5654): p. 52-56.
100. Willerslev, E. and D.J. Meltzer, *Peopling of the Americas as inferred from ancient genomics.* Nature, 2021. **594**(7863): p. 356-364.
101. Sikora, M., et al., *The population history of northeastern Siberia since the Pleistocene.* Nature, 2019. **570**(7760): p. 182-188.
102. Forster, P., et al., *Origin and evolution of Native American mtDNA variation: a reappraisal.* American journal of human genetics, 1996. **59**(4): p. 935.
103. Brown, M.D., et al., *mtDNA haplogroup X: an ancient link between Europe/Western Asia and North America?* The American Journal of Human Genetics, 1998. **63**(6): p. 1852-1861.
104. Waters, M.R., *Late Pleistocene exploration and settlement of the Americas by modern humans.* Science, 2019. **365**(6449).
105. Llamas, B., et al., *Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas.* Science advances, 2016. **2**(4): p. e1501385.
106. Schurr, T.G. and S.T. Sherry, *Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: evolutionary and demographic evidence.* American Journal of Human Biology, 2004. **16**(4): p. 420-439.
107. Bortolini, M.-C., et al., *Y-chromosome evidence for differing ancient demographic histories in the Americas.* The American Journal of Human Genetics, 2003. **73**(3): p. 524-539.
108. Seielstad, M., et al., *A novel Y-chromosome variant puts an upper limit on the timing of first entry into the Americas.* American journal of human genetics, 2003. **73**(3): p. 700.
109. Tamm, E., et al., *Beringian standstill and spread of Native American founders.* PloS one, 2007. **2**(9): p. e829.

110. Kitchen, A., M.M. Miyamoto, and C.J. Mulligan, *A three-stage colonization model for the peopling of the Americas.* PloS one, 2008. **3**(2): p. e1596.
111. Raghavan, M., et al., *Genomic evidence for the Pleistocene and recent population history of Native Americans.* Science, 2015. **349**(6250).
112. Moreno-Mayar, J.V., et al., *Early human dispersals within the Americas.* Science, 2018. **362**(6419).
113. Moreno-Mayar, J.V., et al., *Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans.* Nature, 2018. **553**(7687): p. 203-207.
114. Verdu, P., et al., *Patterns of admixture and population structure in native populations of Northwest North America.* PLoS genetics, 2014. **10**(8): p. e1004530.
115. Reich, D., et al., *Reconstructing native American population history.* Nature, 2012. **488**(7411): p. 370-374.
116. Heintzman, P.D., et al., *Bison phylogeography constrains dispersal and viability of the Ice Free Corridor in western Canada.* Proceedings of the National Academy of Sciences, 2016. **113**(29): p. 8057-8063.
117. Pedersen, M.W., et al., *Postglacial viability and colonization in North America's ice-free corridor.* Nature, 2016. **537**(7618): p. 45-49.
118. Pinotti, T., et al., *Y chromosome sequences reveal a short Beringian Standstill, rapid expansion, and early population structure of Native American founders.* Current Biology, 2019. **29**(1): p. 149-157. e3.
119. Waters, M.R., et al., *Pre-Clovis mastodon hunting 13,800 years ago at the Manis site, Washington.* Science, 2011. **334**(6054): p. 351-353.
120. Gilbert, M.T.P., et al., *DNA from pre-Clovis human coprolites in Oregon, North America.* Science, 2008. **320**(5877): p. 786-789.
121. Jenkins, D.L., et al., *Clovis age Western Stemmed projectile points and human coprolites at the Paisley Caves.* Science, 2012. **337**(6091): p. 223-228.
122. Dillehay, T.D., *Monte Verde, a Late Pleistocene settlement in Chile: the archaeological context and interpretation*. Vol. 2. 1989: Smithsonian Institution Press.
123. Driver, J.C., et al., *Stratigraphy, radiocarbon dating, and culture history of Charlie Lake Cave, British Columbia.* Arctic, 1996: p. 265-277.
124. Darvill, C., et al., *Retreat of the western Cordilleran Ice Sheet margin during the last deglaciation.* Geophysical Research Letters, 2018. **45**(18): p. 9710-9720.
125. Lesnek, A.J., et al., *Deglaciation of the Pacific coastal corridor directly preceded the human colonization of the Americas.* Science Advances, 2018. **4**(5): p. eaar5040.
126. Menounos, B., et al., *Cordilleran Ice Sheet mass loss preceded climate reversals near the Pleistocene Termination.* Science, 2017. **358**(6364): p. 781-784.
127. Goebel, T., M.R. Waters, and D.H. O'Rourke, *The late Pleistocene dispersal of modern humans in the Americas.* science, 2008. **319**(5869): p. 1497-1502.
128. Potter, B.A., et al., *Current evidence allows multiple models for the peopling of the Americas.* Science Advances, 2018. **4**(8): p. eaat5473.
129. Perego, U.A., et al., *Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups.* Current biology, 2009. **19**(1): p. 1-8.
130. Kashani, B.H., et al., *Mitochondrial haplogroup C4c: A rare lineage entering America through the ice-free corridor?* American journal of physical anthropology, 2012. **147**(1): p. 35-39.
131. Perego, U.A., et al., *The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia.* Genome Research, 2010. **20**(9): p. 1174-1179.
132. Achilli, A., et al., *Reconciling migration models to the Americas with the variation of North American native mitogenomes.* Proceedings of the National Academy of Sciences, 2013. **110**(35): p. 14308-14313.
133. Rothhammer, F. and T.D. Dillehay, *The late Pleistocene colonization of South America: an interdisciplinary perspective.* Annals of human genetics, 2009. **73**(5): p. 540-549.
134. Prates, L., G.G. Politis, and S.I. Perez, *Rapid radiation of humans in South America after the last glacial maximum: A radiocarbon-based study.* PLoS one, 2020. **15**(7): p. e0236023.
135. Gómez-Carballa, A., et al., *The peopling of South America and the trans-Andean gene flow of the first settlers.* Genome research, 2018. **28**(6): p. 767-779.

136. García, A., et al., *Ancient and modern mitogenomes from Central Argentina: new insights into population continuity, temporal depth and migration in South America.* Human Molecular Genetics, 2021. **30**(13): p. 1200-1217.
137. Fix, A.G., *Rapid deployment of the five founding Amerind mtDNA haplogroups via coastal and riverine colonization.* American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists, 2005. **128**(2): p. 430-436.
138. Wang, S., et al., *Genetic variation and population structure in Native Americans.* PLoS genetics, 2007. **3**(11): p. e185.
139. Fuselli, S., et al., *Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders.* Molecular biology and evolution, 2003. **20**(10): p. 1682-1691.
140. Wang, S., et al., *Genetic variation and population structure in Native Americans.* PLoS Genet, 2007. **3**(11): p. e185.
141. Tarazona-Santos, E., et al., *Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome.* The American Journal of Human Genetics, 2001. **68**(6): p. 1485-1496.
142. Skoglund, P., et al., *Genetic evidence for two founding populations of the Americas.* Nature, 2015. **525**(7567): p. 104-108.
143. Castro, M.A., et al., *Deep genetic affinity between coastal Pacific and Amazonian natives evidenced by Australasian ancestry.* Proceedings of the National Academy of Sciences, 2021. **118**(14).
144. LA, P.A.D.E.P., *Departamento Administrativo Nacional de Estadística DANE.* 2020.
145. Safford, F. and M. Palacios, *Colombia: Fragmented land, divided society*. 2002: Oxford University Press New York.
146. Ossa, H., et al., *Outlining the ancestry landscape of Colombian admixed populations.* PLoS One, 2016. **11**(10): p. e0164414.
147. DANE., *Colombia una nación multicultural. Su diversidad étnica.* Departamento Administrativo Nacional de Estadística, 2007.
148. *"Censo Nacional de Población y Vivienda 2018".* www.dane.gov.co. **Retrieved 2021-10-22.**
149. Keyeux, G., et al., *Possible migration routes into South America deduced from mitochondrial DNA studies in Colombian Amerindian populations.* Human Biology, 2002: p. 211-233.
150. Delgado, M., *EARLY NEOTROPICAL HUNTER-GATHERERS AND THE DYNAMICS OF THE INITIAL PEOPLING OF NORTHERN SOUTH AMERICA EARLY NEOTROPICAL HUNTER-GATHERERS AND THE DYNAMICS OF THE INITIAL PEOPLING OF NORTHERN SOUTH AMERICA.* Quaternary International, 2021. **578**: p. 20.
151. Aceituno, F.J., et al., *The initial human settlement of Northwest South America during the Pleistocene/Holocene transition: Synthesis and perspectives.* Quaternary International, 2013. **301**: p. 23-33.
152. Delgado, M., F.J. Aceituno, and G. Barrientos, *14C data and the early colonization of northwest South America: a critical assessment.* Quaternary International, 2015. **363**: p. 55-64.
153. Archila, S., et al., *Dwelling the hill: traces of increasing sedentism in hunter-gatherers societies at Checua site, Colombia (9500-5052 cal BP).* Quaternary International, 2021. **578**: p. 102-119.
154. Delgado, M., et al., *A paleogenetic perspective of the Sabana de Bogotá (Northern South America) population history over the Holocene (9000–550 cal BP).* Quaternary International, 2021. **578**: p. 73-86.
155. Morcote-Ríos, G., et al., *Colonisation and early peopling of the Colombian Amazon during the Late Pleistocene and the Early Holocene: new evidence from La Serranía La Lindosa.* Quaternary International, 2021. **578**: p. 5-19.
156. Correal Urrego, G., *Evidencias culturales y megafauna pleistocénica en Colombia*, 1981.
157. Correal, G. and T. Van der Hammen, *Investigaciones arqueológicas en los abrigos rocosos del Tequendama.* Biblioteca Banco Popular, Bogotá, 1977.
158. Dickau, R., et al., *Radiocarbon chronology of terminal Pleistocene to middle Holocene human occupation in the Middle Cauca Valley, Colombia.* Quaternary International, 2015. **363**: p. 43-54.
159. López, C.E., *Landscapes variability and the early peopling of the inter-Andean Magdalena Valley, Colombia (South America).* Quaternary International, 2021. **578**: p. 139-154.
160. Mora, S. and C. Gnecco, *Archaeological hunter-gatherers in tropical forests: a view from Colombia.* Under the canopy: The archaeology of tropical rain forests, 2002: p. 271-290.

161. Adhikari, K., et al., *The genetic diversity of the Americas.* Annual review of genomics and human genetics, 2017. **18**: p. 277-296.
162. Carvajal, D., *As Colombia emerges from decades of war, migration challenges mount.* Migration Information Source, 2017. **13**.
163. Carvajal-Carmona, L.G., et al., *Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia.* The American Journal of Human Genetics, 2000. **67**(5): p. 1287-1295.
164. Mesa, N.R., et al., *Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: pre-and post-Columbian patterns of gene flow in South America.* The American Journal of Human Genetics, 2000. **67**(5): p. 1277-1286.
165. Yunis, J.J., E.J. Yunis, and E. Yunis, *Genetic relationship of the Guambino, Paez, and Ingano Amerindians of southwest Colombia using major histocompatibility complex class II haplotypes and blood groups.* Human immunology, 2001. **62**(9): p. 970-978.
166. Gaviria, A.b., et al., *Nineteen autosomal microsatellite data from Antioquia (Colombia).* Forensic science international, 2004. **143**(1): p. 69-71.
167. Gaviria, A.A., et al., *Y-chromosome haplotype analysis in Antioquia (Colombia).* Forensic science international, 2005. **151**(1): p. 85-91.
168. Yunis, J.J., O. Garcia, and E.J. Yunis, *Population frequencies for CSF1PO, TPOX, TH01, F13A01, FES/FPS and VWA in seven Amerindian populations from Colombia*. 2005: ASTM International.
169. Bedoya, G., et al., *Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate.* Proceedings of the National Academy of Sciences, 2006. **103**(19): p. 7234-7239.
170. Builes, J.J., et al., *Y-chromosome STRs in an Antioquian (Colombia) population sample.* Forensic science international, 2006. **164**(1): p. 79-86.
171. Palacio, O.D., et al., *Autosomal microsatellite data from Northwestern Colombia.* Forensic science international, 2006. **160**(2-3): p. 217-220.
172. Torres, M.M., et al., *A revertant of the major founder Native American haplogroup C common in populations from northern South America.* American Journal of Human Biology, 2006. **18**(1): p. 59-65.
173. Builes, J.J., et al., *Y chromosome STR haplotypes in the Caribbean city of Cartagena (Colombia).* Forensic science international, 2007. **167**(1): p. 62-69.
174. Melton, P.E., et al., *Biological relationship between Central and South American Chibchan speaking populations: evidence from mtDNA.* American Journal of Physical Anthropology, 2007. **133**(1): p. 753-770.
175. Salas, A., et al., *The mtDNA ancestry of admixed Colombian populations.* American Journal of Human Biology: The Official Journal of the Human Biology Association, 2008. **20**(5): p. 584-591.
176. Rojas, W., et al., *Genetic make up and structure of Colombian populations by means of uniparental and biparental DNA markers.* American Journal of Physical Anthropology, 2010. **143**(1): p. 13-20.
177. Yang, N.N., et al., *Contrasting patterns of nuclear and mtDNA diversity in Native American populations.* Annals of human genetics, 2010. **74**(6): p. 525-538.
178. Casas-Vargas, A., et al., *High genetic diversity on a sample of pre-Columbian bone remains from Guane territories in northwestern Colombia.* American journal of physical anthropology, 2011. **146**(4): p. 637-649.
179. Usme-Romero, S., et al., *Genetic differences between Chibcha and Non-Chibcha speaking tribes based on mitochondrial DNA (mtDNA) haplogroups from 21 Amerindian tribes from Colombia.* Genetics and molecular biology, 2013. **36**: p. 149-157.
180. Ibarra, A., et al., *Comparison of the genetic background of different Colombian populations using the SNP for ID 52plex identification panel.* International journal of legal medicine, 2014. **128**(1): p. 19-25.
181. Bisso-Machado, R. and N.J. Fagundes, *Uniparental genetic markers in Native Americans: A summary of all available data from ancient and contemporary populations.* American Journal of Physical Anthropology, 2021. **176**(3): p. 445-458.
182. Criollo-Rayo, A.A., et al., *Native American gene continuity to the modern admixed population from the Colombian Andes: implication for biomedical, population and forensic studies.* Forensic Science International: Genetics, 2018. **36**: p. e1-e7.

183. Castillo, A., et al., *Maternal genetic characterization of a Colombian Andean population.* Forensic Science International: Genetics Supplement Series, 2019. **7**(1): p. 342-344.
184. Xavier, C., et al., *Admixture and genetic diversity distribution patterns of non-recombining lineages of Native American ancestry in Colombian populations.* PloS one, 2015. **10**(3): p. e0120155.
185. Yunis, J.J. and E.J. Yunis, *Mitochondrial DNA (mtDNA) haplogroups in 1526 unrelated individuals from 11 Departments of Colombia.* Genetics and molecular biology, 2013. **36**(3): p. 329-335.
186. Simão, F., et al., *The maternal inheritance of the Ashaninka native group from Peru.* Forensic Science International: Genetics Supplement Series, 2019. **7**(1): p. 135-137.
187. Simão, F., et al., *Paraguay: Unveiling migration patterns with ancestry genetic markers.* Forensic Science International: Genetics Supplement Series, 2017. **6**: p. e226-e228.
188. Bobillo, M.C., et al., *Amerindian mitochondrial DNA haplogroups predominate in the population of Argentina: towards a first nationwide forensic mitochondrial DNA sequence database.* International Journal of Legal Medicine, 2010. **124**(4): p. 263-268.
189. Bio-Rad, L., *Chelex®-100 and Chelex®-20 Chelating Ion Exchange Resin Instruction Manual.* Bio-Rad Laboratories, 2000.
190. Walsh, P.S., D.A. Metzger, and R. Higuchi, *Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material.* Biotechniques, 1991. **10**(4): p. 506-513.
191. Carracedo, A., et al., *DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing.* Forensic Science International, 2000. **110**(2): p. 79-85.
192. Erlich, H.A., *Polymerase chain reaction.* Journal of clinical immunology, 1989. **9**(6): p. 437-447.
193. Saiki, R.K., et al., *Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase.* Science, 1988. **239**(4839): p. 487-491.
194. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proceedings of the national academy of sciences, 1977. **74**(12): p. 5463-5467.
195. Tully, G., et al., *Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles.* Forensic Science International, 2001. **124**(1): p. 83-91.
196. Excoffier, L. and H.E. Lischer, *Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows.* Molecular ecology resources, 2010. **10**(3): p. 564-567.
197. Statsoft, I., *STATISTICA (data analysis software system), version 10.0*, 2011, StatSoft Tulsa^ eOklahoma Oklahoma.
198. Excoffier, L., P.E. Smouse, and J.M. Quattro, *Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data.* Genetics, 1992. **131**(2): p. 479-491.
199. Bandelt, H.-J., P. Forster, and A. Röhl, *Median-joining networks for inferring intraspecific phylogenies.* Molecular biology and evolution, 1999. **16**(1): p. 37-48.
200. Burgos, G., et al., *An approach to maternal ancestry in a sample of Ecuadorian "mestizo" population by sequencing the control region of mtDNA.* Forensic Science International: Genetics Supplement Series, 2019. **7**(1): p. 537-538.
201. Simão, F., et al., *Defining mtDNA origins and population stratification in Rio de Janeiro.* Forensic Science International: Genetics, 2018. **34**: p. 97-104.
202. Dos Reis, R.S., et al., *A view of the maternal inheritance of Espírito Santo populations: The contrast between the admixed and Pomeranian descent groups.* Forensic Science International: Genetics, 2019. **40**: p. 175-181.
203. Freitas, J.M., et al., *Mitochondrial DNA control region haplotypes and haplogroup diversity in a sample from Brasília, Federal District, Brazil.* Forensic Science International: Genetics, 2019. **40**: p. e228-e230.
204. Salas, A., et al., *Mitochondrial echoes of first settlement and genetic continuity in El Salvador.* PLoS One, 2009. **4**(9): p. e6882.
205. Nuñez, C., et al., *Reconstructing the population history of Nicaragua by means of mtDNA, Y-chromosome STRs, and autosomal STR markers.* American journal of physical anthropology, 2010. **143**(4): p. 591-600.
206. Söchtig, J., et al., *Genomic insights on the ethno-history of the Maya and the 'Ladinos' from Guatemala.* BMC genomics, 2015. **16**(1): p. 1-18.

207. Alonso Morales, L.A., et al., *Paternal portrait of populations of the middle Magdalena River region (Tolima and Huila, Colombia): New insights on the peopling of Central America and northernmost South America.* PloS one, 2018. **13**(11): p. e0207130.
208. Consortium, G.P., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68.
209. Baeta, M., et al., *Mitochondrial diversity in Amerindian Kichwa and Mestizo populations from Ecuador.* International journal of legal medicine, 2012. **126**(2): p. 299-302.
210. Gómez-Carballa, A., et al., *Revealing latitudinal patterns of mitochondrial DNA diversity in Chileans.* Forensic Science International: Genetics, 2016. **20**: p. 81-88.
211. Simão, F., et al., *The maternal inheritance of Alto Paraná revealed by full mitogenome sequences.* Forensic Science International: Genetics, 2019. **39**: p. 66-72.
212. Roca-Rada, X., et al., *Ancient mitochondrial genomes from the Argentinian Pampas inform the early peopling of the Southern Cone of South America.* Iscience, 2021. **24**(6): p. 102553.

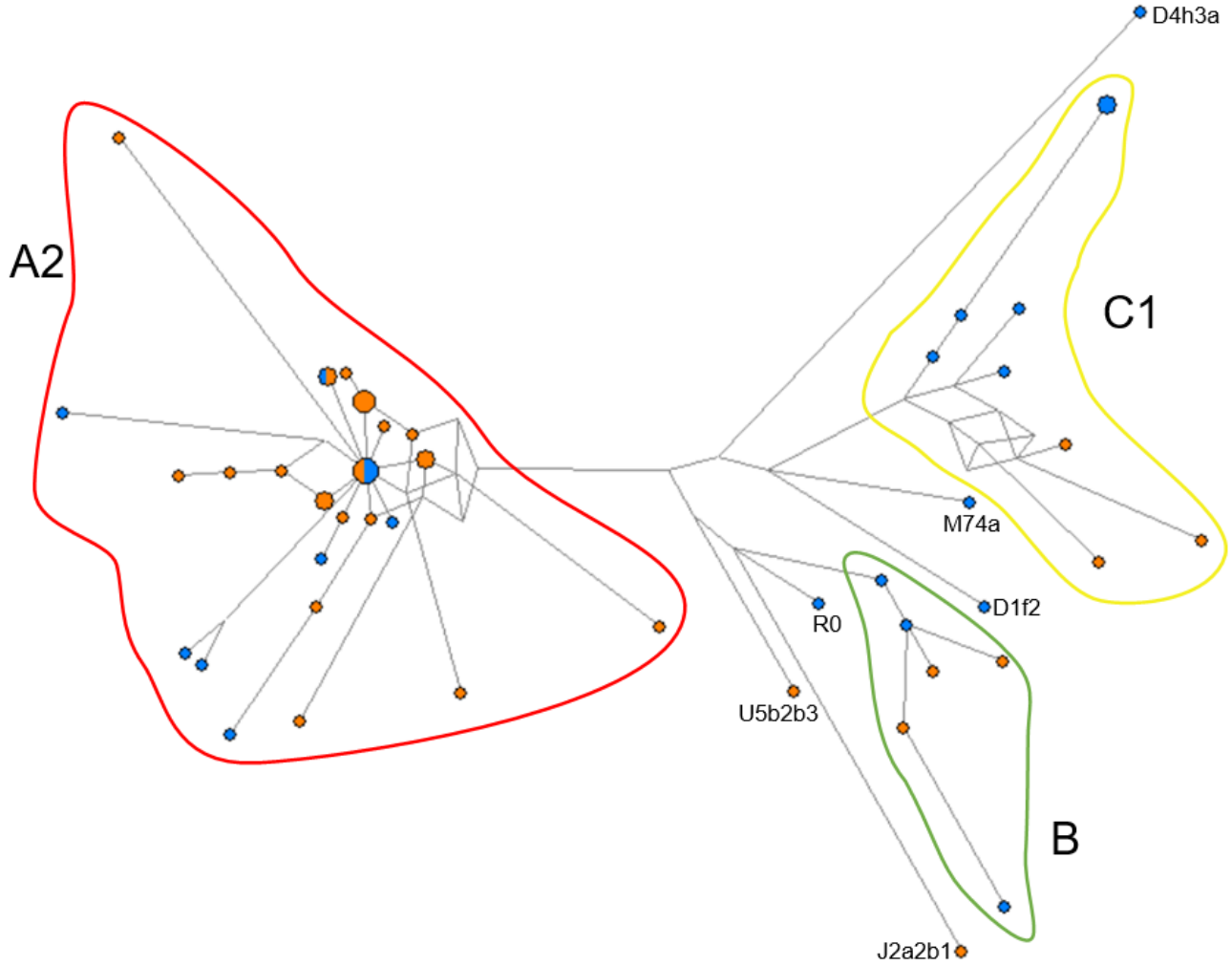
7. Appendix

Appendix 1. mtDNA control region haplotypes observed in Cundinamarca (n=31) and Tolima (n=22). For each sample, polymorphic sites (range: 16024-576) and correspondent haplogroup are presented.

Colombian department	Sample ID	Freq.	Haplogroup	Polymorphic sites (range: 0-576; 16024-16569)																				
Cundinamarca	>92	1	B2d	73G	263G	309.1C	315.1C	498d	499A	16183C	16189C	16217C	16324C	16519C										
Cundinamarca	>93	1	A2 + (64)	64T	73G	146C	153G	235G	263G	291d	309.1C	315.1C	456T	523d	524d	16111T	16223T	16290T	16319A	16362C				
Cundinamarca	>95	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16111T	16223T	16290T	16319A	16362C	16519C					
Cundinamarca	>96	1	A2y	64T	73G	146C	152C	153G	189G	204C	207A	235G	263G	309.1C	315.1C	523d	524d	16111T	16258G	16290T	16293G	16304C	16319A	16362C
Cundinamarca	>97	1	C1b+16311	73G	249d	263G	290d	291d	309.1C	309.2C	315.1C	489C	493G	523d	524d	16189C	16223T	16298C	16311C	16325C	16327T			
Cundinamarca	>98	1	A2ac	64T	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16111T	16213A	16223T	16290T	16319A	16362C					
Cundinamarca	>100	1	A2+(64)	64T	73G	146C	235G	263G	309.1C	315.1C	523d	524d	16111T	16223T	16290T	16319A	16362C							
Cundinamarca	>103	1	C1b	73G	204C	249d	263G	282C	290d	291d	315.1C	489C	493G	523d	524d	16183C	16185T	16189C	16193d	16223T	16298C	16325C	16327T	
Cundinamarca	>105	1	A2+(64)+16129	64T	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16111T	16129A	16223T	16256T	16290T	16319A	16362C				
Cundinamarca	>106	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	309.2C	315.1C	523d	524d	16223T	16290T	16319A	16362C						
Cundinamarca	>107	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	309.2C	315.1C	523d	524d	16111T	16223T	16290T	16319A	16362C					
Cundinamarca	>108	1	A2+(64)	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16223T	16290T	16319A	16362C								
Cundinamarca	>109	1	A2ac	64T	146C	153G	182T	235G	263G	309.1C	315.1C	523d	524d	16111T	16213A	16223T	16256T	16290T	16319A	16362C				
Cundinamarca	>110	1	A2ac	64T	73G	146C	153G	182T	235G	263G	309.1C	309.2C		315.1C	523d	524d	16111T	16213A	16223T	16290T	16319A	16362C		
Cundinamarca	>111	1	A2+(64)	64T	146C	153G	235G	263G	309.1C	309.2C	315.1C	523d	524d	16223T	16290T	16319A	16362C							
Cundinamarca	>113	1	A2+(64)	64T	73G	146C	189G	199C	206C	235G	263G	309.1C	315.1C	523d	524d	16111T	16195C	16223T	16290T	16293G	16319A	16362C	16519C	
Cundinamarca	>114	1	A2ac	64T	146C	153G	182T	235G	263G	309.1C	315.1C	523d	524d	16111T	16213A	16223T	16290T	16319A	16362C					
Cundinamarca	>116	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16223T	16290T	16319A	16362C							
Cundinamarca	>117	1	B4b	73G	263G	309.1C	315.1C	499A	16086C	16182C	16183C	16189C	16193.1C	16217C	16519C									
Cundinamarca	>119	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16111T	16223T	16287T	16290T	16319A	16362C					
Cundinamarca	>121	1	A2ac	64T	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16111T	16213A	16223T	16290T	16319A	16362C					
Cundinamarca	>126	1	A2+(64)	62T	64T	73G	146C	235G	263G	309.1C	315.1C	523d	524d	16111T	16192T	16223T	16290T	16319A	16362C					
Cundinamarca	>127	1	U5b2b3	73G	150T	263G	309.1C	309.2C	315.1C	517T	16224C	16270T	16362C	16519C										
Cundinamarca	>128	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16111T	16223T	16290T	16319A	16362C						
Cundinamarca	>129	1	A2	73G	146C	153G	195C	235G	263G	315.1C	523d	524d	16085G	16111T	16189C	16223T	16266T	16290T	16319A	16356C	16362C	16519C		
Cundinamarca	>130	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	309.2C	315.1C	523d	524d	16111T	16223T	16290T	16319A	16362C	16519C				
Cundinamarca	>131	1	B2d	73G	195C	263G	315.1C	498d	499A	16129A	16183C	16189C	16217C	16519C										
Cundinamarca	>132	1	A2r1	73G	103A	114T	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16111T	16223T	16261T	16290T	16311C	16319A	16362C			
Cundinamarca	>141	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16223T	16290T	16319A	16362C							
Cundinamarca	>142	1	C1b	73G	152C	234G	249d	263G	290d	291d	315.1C	489C	493G	523d	524d	16189A	16223T	16291T	16298C	16325C	16327T	16519C		
Cundinamarca	>150	1	J2a2b1	73G	150T	185A	189G	195C	263G	295T	315.1C	489C	16069T	16126C	16241G	16519C								
Tolima	>633	1	C1d+194	73G	153G	194T	249d	263G	290d	291d	309.1C	315.1C	489C	523d	524d	16051G	16094C	16223T	16261T	16298C	16325C	16327T	16518A	
Tolima	>634	1	C1d+194	16051G	16223T	16298C	16325C	16327T	16519C	73G	194T	249d	263G	290d	291d	315.1C	489C	523d	524d					
Tolima	>636	1	R0	315.1C	523d	524d	16042A	16519C	263G															
Tolima	>639	1	A2+(64)	59C	62T	64T	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16111T	16192T	16223T	16248T	16290T	16319A	16362C		
Tolima	>642	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	315.1C	456T	523d	524d	16111T	16172C	16223T	16290T	16319A	16362C				
Tolima	>643	1	C1d+194	73G	153G	194T	249d	263G	290d	291d	309.1C	315.1C	489C	523d	524d	16051G	16094C	16223T	16261T	16298C	16325C	16327T	16518A	
Tolima	>644	1	B2d	73G	263G	309.1C	309.2C	315.1C	498d	499A	16183C	16189C	16217C	16519C										
Tolima	>645	1	D4h3a	73G	152C	263G	309.1C	315.1C	489C	16187T	16189C	16203G	16223T	16241G	16245T	16301T	16319A	16342C	16362C	16468C	16533C			
Tolima	>650	1	A2ad	64T	73G	146C	153G	194T	235G	263G	309.1C	309.2C	315.1C	523d	524d	16075C	16111T	16175G	16223T	16259T	16290T	16300G	16319A	16362C
Tolima	>653	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	309.2C	315.1C	523d	524d	16111T	16223T	16290T	16319A	16362C					
Tolima	>654	1	B2d	73G	263G	309.1C	315.1C	498d	499A	16189C	16217C	16342.1T	16519C											
Tolima	>700	1	C1d	73G	249d	263G	290d	291d	309.1C	315.1C	489C	523d	524d	16051G	16223T	16298C	16325C	16327T	16519C					
Tolima	>701	1	A2aq	64T	73G	146C	153G	235G	263G	309.1C	309.2C	315.1C	523d	524d	16111T	16223T	16290T	16319A	16355T	16362C				
Tolima	>702	1	B2c2	25d	73G	114T	146C	152C	263G	309.1C	315.1C	499A	16086C	16182C	16183C	16189C	16217C	16474C	16519C					
Tolima	>703	1	C1	73G	249d	263G	290d	291d	315.1C	489C	16188T	16223T	16298C	16325C	16327T									
Tolima	>706	1	A2+(64)	25d	64T	73G	146C	153G	194T	235G	263G	309.1C	315.1C	523d	524d	16075C	16111T	16223T	16259T	16300G	16319A	16362C		
Tolima	>709	1	C1	73G	185A	249d	263G	290d	291d	309.1C	315.1C	489C	16223T	16274A	16298C	16325C	16327T							
Tolima	>710	1	A2+(64)	64T	73G	146C	152C	153G	182T	235G	263G	309.1C	315.1C	316A	523d	524d	16092C	16111T	16223T	16260T	16290T	16319A		
Tolima	>711	1	M74a	63C	64T	66A	73G	263G	309.1C	315.1C	489C	16222T	16223T	16325C	16362C	16519C								
Tolima	>712	1	D1f2	73G	152C	200G	263G	315.1C	489C	16092C	16142T	16223T	16311C	16325C	16362C	16519C								
Tolima	>718	1	A2+(64)	64T	73G	146C	153G	235G	263G	309.1C	315.1C	523d	524d	16111T	16223T	16290T	16319A	16362C						
Tolima	>719	1	A2+(64)+16129	64T	73G	146C	153G	235G	263G	309.1C	309.2C	315.1C	523d	524d	16111T	16129A	16223T	16256T	16290T	16319A	16362C			

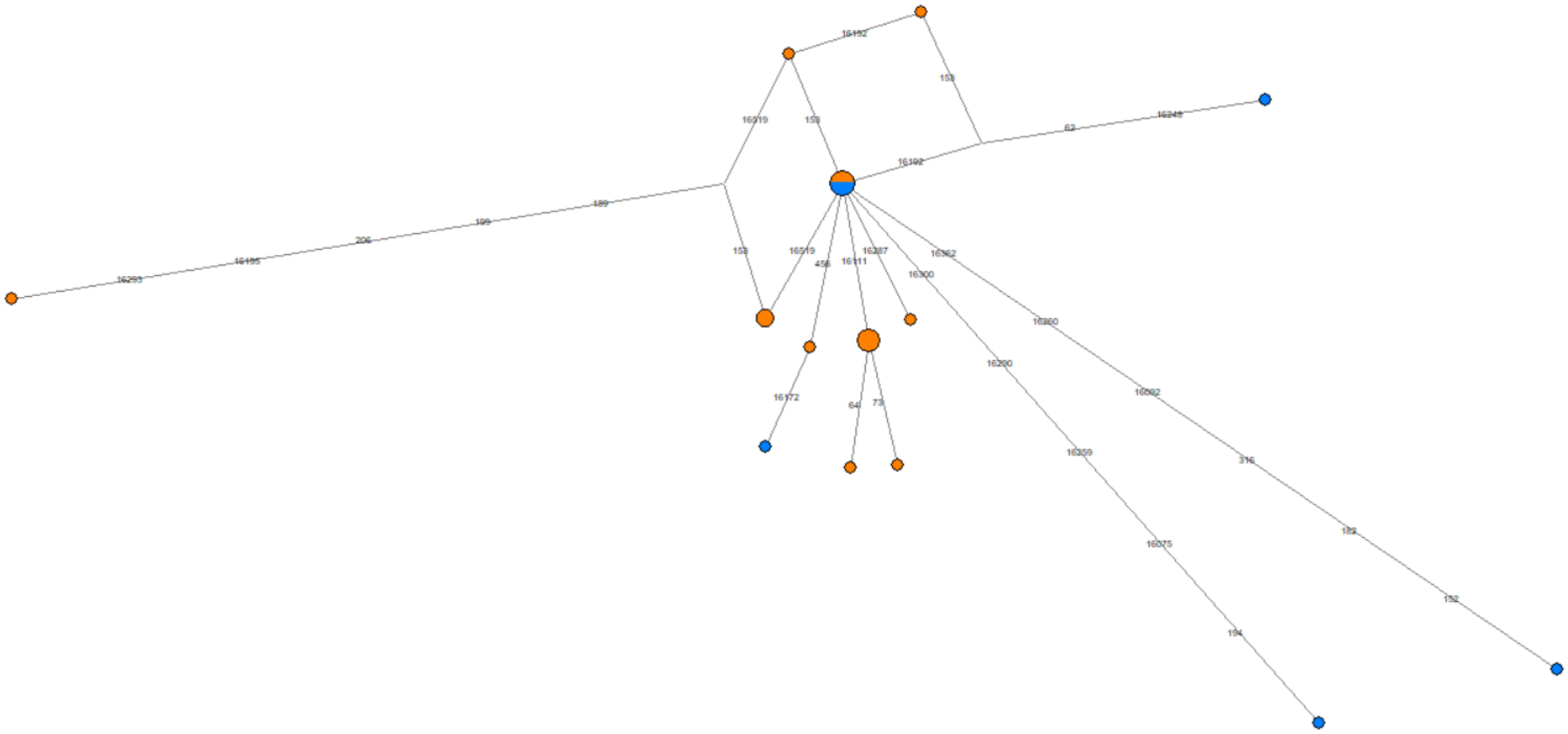
Appendix 2. Median-joining network of haplotypes present in 53 samples from Cundinamarca (in orange) and Tolima (in blue).

From a total of 53 sequences, 44 haplotypes were found (disregarding indels); diameter size is proportional to haplotype frequency.



Appendix 3. Median-joining network of haplogroup A2+(64) in 20 samples from Cundinamarca (orange) and Tolima (blue).

From a total of 20 sequences, 14 haplotypes were found (disregarding indels): 11 haplotypes were unique, and 3 haplotypes were shared between 4,2 and 3 sequences. Diameter size is proportional to haplotype frequency.



Appendix 4. Pairwise genetic distances based on CR haplotypes shared between the Andean Colombian populations considered for comparative purposes. In red are the F_{ST} values and in parentheses the p-values; highlighted in grey are the results statistically significant (p-value < 0.05).

4.1. Considering seven Andean departments

	Santander	Norte de Santander	Cundinamarca	Boyacá	Tolima	Cauca	Antioquia
Santander							
Norte de Santander	0,00001 (0,31532+-0,0365)	0					
Cundinamarca	0,08031 (0)	0,15445 (0,00901+-0,0091)	0				
Boyacá	0,00445 (0,26126+-0,0344)	0,03351 (0,11712+-0,0360)	0,06367 (0)	0			
Tolima	0,02952 (0,08108+-0,0286)	0,00993 (0,36036+-0,0470)	0,04215 (0,01802+-0,0121)	0,04974 (0,01802+-0,0121)	0		
Cauca	0,18076 (0)	0,08684 (0,06306+-0,0194)	0,26019 (0)	0,23984 (0)	0,10474 (0,00901+-0,0091)	0	
Antioquia	0,00555 (0,14414+-0,0454)	0,04563 (0,19820+-0,0227)	0,05725 (0)	0,00068 (0,31532+-0,0311)	0,04977 (0,01802+-0,0121)	0,23804 (0)	0

4.2. Considering six Andean departments (where samples from Norte de Santander were included in the Santander population)

	Santander	Cundinamarca	Boyacá	Tolima	Cauca	Antioquia
Santander	0					
Cundinamarca	0,08647 (0)	0				
Boyacá	0,00806 (0,19820+-0,0503)	0,06367 (0,00901+-0,0091)	0			
Tolima	0,02821 (0,06306+-0,0237)	0,04215 (0,02703+-0,0194)	0,04974 (0,03604+-0,0148)	0		
Cauca	0,16946 (0)	0,26019 (0)	0,23984 (0)	0,10474 (0)	0	
Antioquia	0,00915 (0,14414+-0,0337)	0,05725 (0)	0,00068 (0,29730+-0,0305)	0,04977 (0,02703+-0,0139)	0,23804 (0)	0

Appendix 5. Percentage of Native and non-Native haplogroups from the seven considered Colombian departments, which include samples obtained from the current study and samples from previously published works (as described in Table 3 and Table 4).

Legenda:

1 - Norte de Santander

2 - Santander

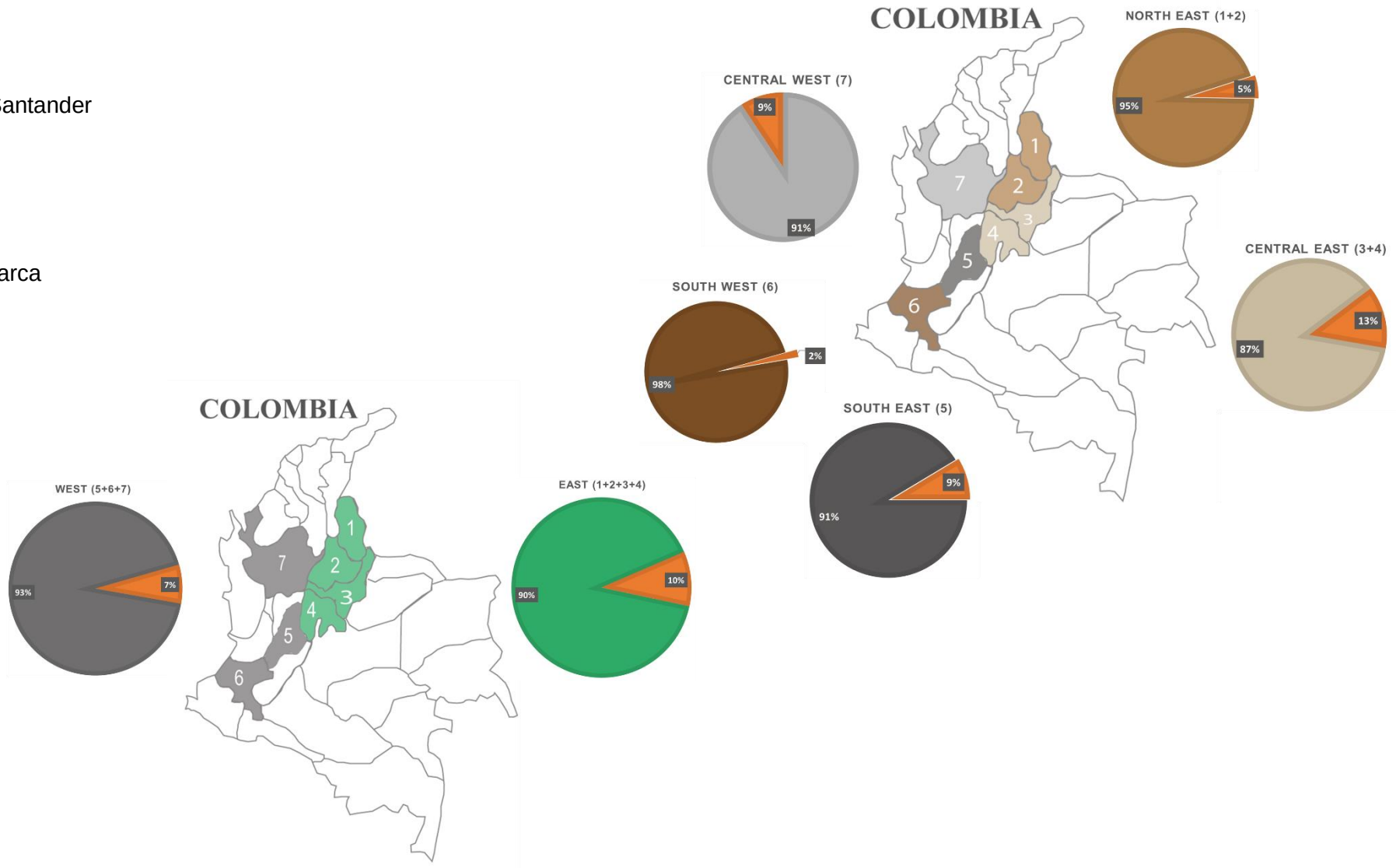
3 - Boyacá

4 - Cundinamarca

5 - Tolima

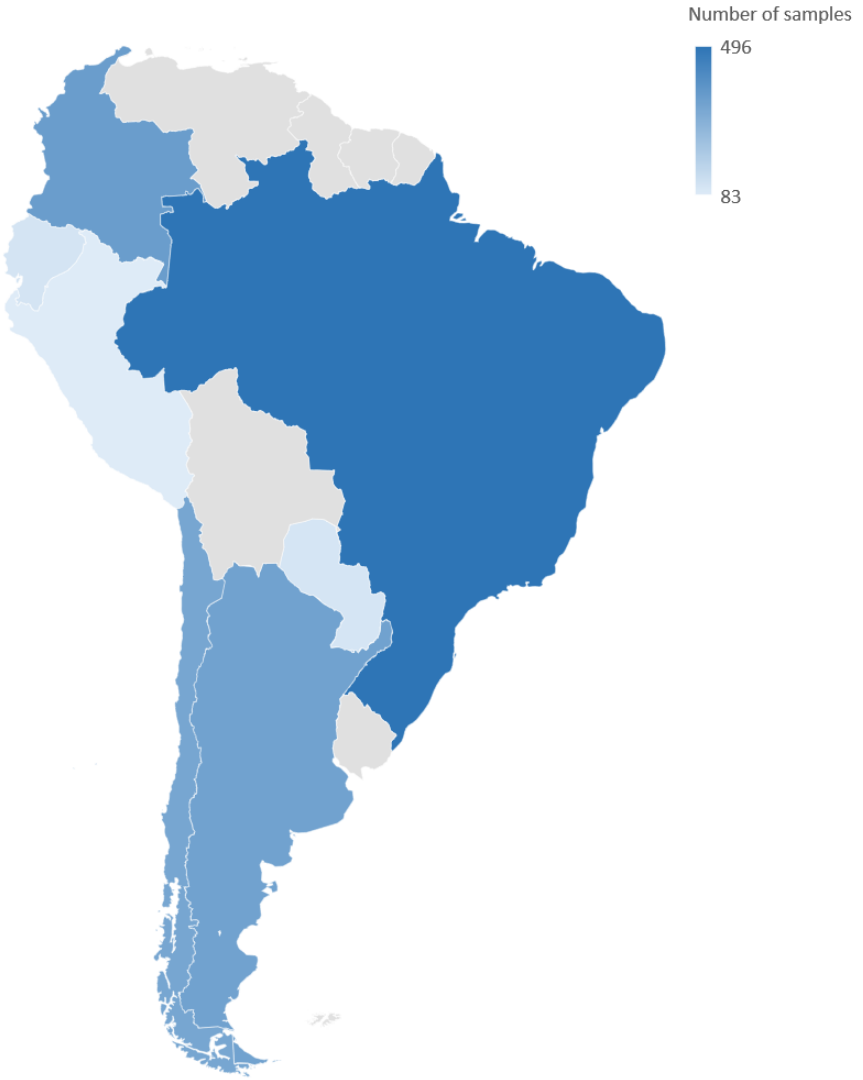
6 - Cauca

7 - Antioquia



Appendix 6. A total of 1810 samples collected from South American populations to perform comparative analysis concerning their CR haplotypes. Geographic distribution is presented as well as the number of samples for each region.

Country	Region	Number of samples
Colombia	Andes	356
Peru	Lima	83
Brazil	Rio de Janeiro	205
Brazil	Espírito Santo	291
Ecuador	-	107
Argentina	North-East	98
Argentina	Central-West	193
Argentina	South	47
Chile	North	148
Chile	Southern	177
Paraguay	Alto Paraná	105



Appendix 7. Pairwise genetic distances based on CR haplotypes shared between the South American populations considered for comparative purposes. In red are the F_{ST} values and in parentheses the p-values; highlighted in grey are the statistically significant values (p-value < 0.05).

7.1. Considering departments within Colombia as independent variables.

	Cundinamarca	Tolima	Santander	Boyacá	Antioquia	Cauca	Peru (Lima)	Brazil (Rio de Janeiro)	Brazil (Espírito Santo)	Ecuador	Argentina (North-East)	Argentina (Central-West)	Argentina (South)	Chile (North)	Chile (Southern)	Paraguay (Alto Paraná)
Cundinamarca	0															
Tolima	0,04215 (0,01802+-0,0121)	0														
Santander	0,08618 (0)	0,02802 (0,08108+-0,0212)	0													
Boyacá	0,06349 (0)	0,04955 (0,05405+-0,0201)	0,00813 (0,20721+-0,0305)	0												
Antioquia	0,05721 (0)	0,04968 (0)	0,00915 (0,10811+-0,0227)	0,0008 (0,31532+-0,0512)	0											
Cauca	0,26439 (0)	0,10766 (0,00901+-0,0091)	0,1727 (0)	0,24311 (0)	0,24157 (0)	0										
Peru (Lima)	0,19226 (0)	0,09348 (0)	0,03485 (0,00901+-0,0091)	0,07451 (0)	0,08082 (0)	0,17396 (0)	0									
Brazil (Rio de Janeiro)	0,17299 (0)	0,10633 (0)	0,07929 (0)	0,07783 (0)	0,11057 (0)	0,18127 (0)	0,0661 (0)	0								
Brazil (Espírito Santo)	0,19806 (0)	0,1261 (0)	0,0886 (0)	0,09762 (0)	0,12498 (0)	0,2029 (0)	0,06761 (0)	0,0153 (0)	0							
Ecuador	0,07551 (0)	0,04269 (0,01802+-0,0121)	0,0598 (0)	0,07243 (0)	0,06815 (0)	0,1841 (0)	0,10881 (0)	0,13243 (0)	0,14361 (0)	0						
Argentina (North-East)	0,12751 (0)	0,03737 (0,00901+-0,0091)	0,03667 (0)	0,06209 (0)	0,07683 (0)	0,10569 (0)	0,03777 (0)	0,04399 (0)	0,03997 (0)	0,07079 (0)	0					
Argentina (Central-West)	0,15959 (0)	0,06931 (0)	0,0521 (0)	0,07813 (0)	0,09476 (0)	0,143 (0)	0,03998 (0)	0,0336 (0)	0,01572 (0)	0,09819 (0)	0,00795 (0,00901+-0,0091)	0				
Argentina (South)	0,16175 (0)	0,0551 (0)	0,0544 (0)	0,08593 (0)	0,10231 (0)	0,12711 (0)	0,04476 (0)	0,04336 (0)	0,03983 (0)	0,08648 (0)	0,01383 (0,03604+-0,0201)	0,00904 (0,06306+-0,0237)	0			
Chile (North9	0,1908 (0)	0,07318 (0)	0,05699 (0)	0,11083 (0)	0,11595 (0)	0,08321 (0)	0,03013 (0)	0,07841 (0)	0,08782 (0)	0,11186 (0)	0,03125 (0)	0,04532 (0)	0,03095 (0,00901+-0,0091)	0		
Chile (Southern)	0,26646 (0)	0,12604 (0)	0,13704 (0)	0,19914 (0)	0,20256 (0)	0,0669 (0)	0,09723 (0)	0,11591 (0)	0,124 (0)	0,15763 (0)	0,06223 (0)	0,07716 (0)	0,04838 (0)	0,02876 (0)	0	
Paraguay (Alto Paraná)	0,13315 (0)	0,0311 (0,08108+-0,0163)	0,03693 (0)	0,07117 (0)	0,08146 (0)	0,08417 (0)	0,02946 (0)	0,05425 (0)	0,05996 (0)	0,07096 (0)	0,00361 (0,13514+-0,0244)	0,02044 (0)	0,01459 (0,04505+-0,0152)	0,02219 (0)	0,04981 (0)	0

7.2. Considering regions North and South to group the considered Colombian departments

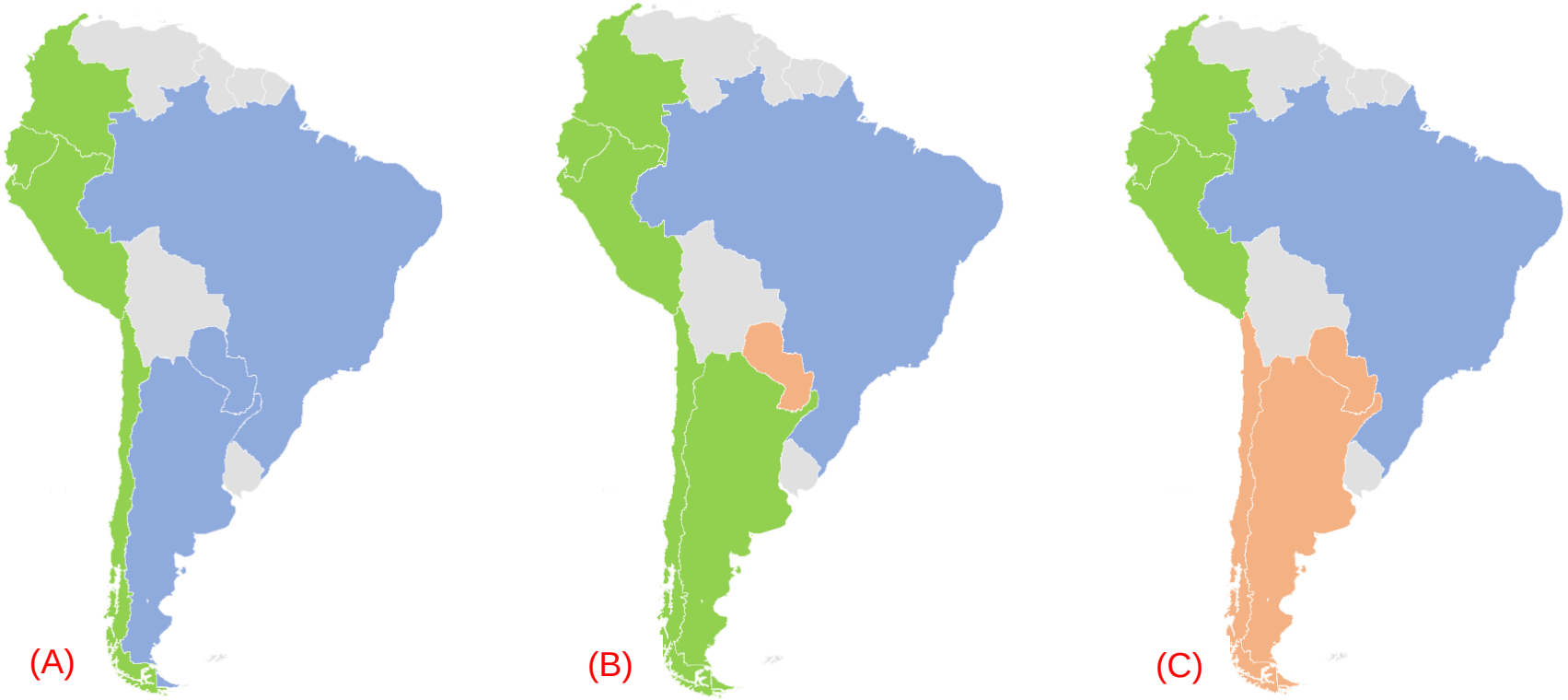
	Peru (Lima)	Ecuador	Colombia (South-Andes)	Colombia (North-Andes)	Argentina (Central West)	Chile (North)	Chile (Southern)	Paraguay (Alto Paraná)	Brazil (Rio de Janeiro)	Brazil (Espírito Santo)	Argentina (North-East)	Argentina (South)
Peru (Lima)	0											
Ecuador	0,10881 (0)	0										
Colombia (South-Andes)	0,10672 (0)	0,06714 (0)	0									
Colombia (North-Andes)	0,0664 (0)	0,06415 (0)	0,09171 (0)	0								
Argentina (Central-West)	0,03998 (0)	0,09819 (0)	0,08368 (0)	0,0814 (0)	0							
Chile (North)	0,03013 (0)	0,11186 (0)	0,06598 (0)	0,10198 (0)	0,04532 (0)	0						
Chile (Southern)	0,09723 (0)	0,15763 (0)	0,09089 (0)	0,1837 (0)	0,07716 (0)	0,02876 (0)	0					
Paraguay (Alto Paraná)	0,02946 (0)	0,07096 (0)	0,03724 (0)	0,06968 (0)	0,02044 (0)	0,02219 (0)	0,04981 (0)	0				
Brazil (Rio de Janeiro)	0,0661 (0)	0,13243 (0)	0,12976 (0)	0,10178 (0)	0,0336 (0)	0,07841 (0)	0,11591 (0)	0,05425 (0)	0			
Brazil (Espírito Santo)	0,06761 (0)	0,14361 (0)	0,1431 (0)	0,11187 (0)	0,01572 (0)	0,08782 (0)	0,124 (0)	0,05996 (0)	0,0153 (0)	0		
Argentina (North-East)	0,03777 (0)	0,07079 (0)	0,04625 (0)	0,06464 (0)	0,00795 (0,00901+-0,0091)	0,03125 (0)	0,06223 (0)	0,00361 (0,18919+-0,0394)	0,04399 (0)	0,03997 (0)	0	
Argentina (South)	0,04476 (0)	0,08648 (0)	0,06745 (0)	0,0872 (0)	0,00904 (0,02703+-0,0139)	0,03095 (0)	0,04838 (0)	0,01459 (0,04505+-0,0203)	0,04336 (0)	0,03983 (0)	0,01383 (0,04505+-0,0279)	0

7.3. Considering regions East and West to group the considered Colombian departments

	Colombia (East-Andes)	Colombia (West-Andes)	Peru (Lima)	Brazil (Rio de Janeiro)	Brazil (Espírito Santo)	Ecuador	Argentina (North-East)	Argentina (Central West)	Argentina (South)	Chile (North)	Chile (Southern)	Paraguay (Alto Paraná)
Colombia (East-Andes)	0											
Colombia (West-Andes)	0,01379 (0,00901+-0,0091)	0										
Peru (Lima)	0,07464 (0)	0,05469 (0)	0									
Brazil (Rio de Janeiro)	0,09953 (0)	0,09472 (0)	0,0661 (0)	0								
Brazil (Espírito Santo)	0,11077 (0)	0,10405 (0)	0,06761 (0)	0,0153 (0)	0							
Ecuador	0,05024 (0)	0,0444 (0)	0,10881 (0)	0,13243 (0)	0,14361 (0)	0						
Argentina (North-East)	0,05658 (0)	0,03101 (0)	0,03777 (0)	0,04399 (0)	0,03997 (0)	0,07079 (0)	0					
Argentina (Central West)	0,07703 (0)	0,05767 (0)	0,03998 (0)	0,0336 (0)	0,01572 (0)	0,09819 (0)	0,00795 (0,06306+-0,0305)	0				
Argentina (South)	0,07935 (0)	0,05184 (0)	0,04476 (0)	0,04336 (0)	0,03983 (0)	0,08648 (0)	0,01383 (0,03604+-0,0148)	0,00904 (0,05405+-0,0201)	0			
Chile (North)	0,10043 (0)	0,05583 (0)	0,03013 (0)	0,07841 (0)	0,08782 (0)	0,11186 (0)	0,03125 (0)	0,04532 (0)	0,03095 (0)	0		
Chile (Southern)	0,17769 (0)	0,11262 (0)	0,09723 (0)	0,11591 (0)	0,124 (0)	0,15763 (0)	0,06223 (0)	0,07716 (0)	0,04838 (0)	0,02876 (0)	0	
Paraguay (Alto Paraná)	0,06232 (0)	0,02831 (0)	0,02946 (0)	0,05425 (0)	0,05996 (0)	0,07096 (0)	0,00361 (0,09910+-0,0252)	0,02044 (0)	0,01459 (0,06306+-0,0194)	0,02219 (0)	0,04981 (0)	0

Appendix 8. AMOVA analysis results considering South American populations; all p-values were statistically significant except when noted with an asterisk (p-value > 0.05).

Geographic criteria	Number of groups	Among groups (F_{CT})	Among populations/within groups (F_{SC})	Within populations (F_{ST})
(A) East / West of South America	2	1.80 (0.0179)*	5.75 (0.0585)	92.46 (0.0754)
(B) Andean States / Brazil / Paraguay	3	0.58 (0.0058)*	6.54 (0.0658)	92.87 (0.0712)
(C) Northern-West / Southern Cone / Brazil	3	3.46 (0.0345)	4.07 (0.0421)	92.48 (0.0752)



Appendix 9. Pairwise genetic distances based on CR Native haplotypes shared between the South American populations considered for comparative purposes. In red are the F_{ST} values and in parentheses the p-values; highlighted in grey are the statistically significant values (p-value < 0.05).

9.1. Considering departments within Colombia as independent variables.

	Cundinamarca	Tolima	Cauca	Boyacá	Antioquia	Santander	Peru (Lima)	Brazil (Rio de Janeiro)	Brazil (Espírito Santo)	Ecuador	Argentina (North-East)	Argentina (Central-West)	Argentina (South)	Chile (North)	Chile (Southern)	Paraguay (Alto Paraná)
Cundinamarca	0															
Tolima	0,05236 (0,02703+-0,0194)	0														
Cauca	0,2967 (0)	0,11778 (0)	0													
Boyacá	0,07232 (0,00901+-0,0091)	0,06182 (0,04505+-0,0244)	0,28604 (0)	0												
Antioquia	0,07255 (0)	0,06494 (0,01802+-0,0121)	0,27386 (0)	0 (0,47748+-0,0360)	0											
Santander	0,1083 (0)	0,03469 (0,03604+-0,0201)	0,18363 (0)	0,01269 (0,13514+-0,0412)	0,01488 (0,04505+-0,0152)	0										
Peru (Lima)	0,22453 (0)	0,10867 (0)	0,1802 (0)	0,10683 (0)	0,10481 (0)	0,03959 (0,00901+-0,0091)	0									
Brazil (Rio de Janeiro)	0,14719 (0)	0,03638 (0,09009+-0,0192)	0,1186 (0)	0,06964 (0)	0,06981 (0)	0,01423 (0,08108+-0,0286)	0,02046 (0,00901+-0,0091)	0								
Brazil (Espírito Santo)	0,10531 (0)	0,0113 (0,20721+-0,0508)	0,10232 (0)	0,06571 (0.00901+-0.0091)	0,06512 (0)	0,02016 (0,07207+-0,0353)	0,04786 (0)	0 (0,56757+-0,0526)	0							
Ecuador	0,09796 (0)	0,05011 (0)	0,18925 (0)	0,08171 (0)	0,07876 (0)	0,06064 (0)	0,11023 (0)	0,07546 (0)	0,05829 (0)	0						
Argentina (North-East)	0,13279 (0)	0,01972 (0,18018+-0,0449)	0,0699 (0)	0,10313 (0)	0,10334 (0)	0,04797 (0)	0,06647 (0)	0,01486 (0,12613+-0,0149)	0,00088 (0,35135+-0,0317)	0,05681 (0)	0					
Argentina (Central-West)	0,13063 (0)	0,01889 (0,12613+-0,0388)	0,07856 (0)	0,09504 (0)	0,0927 (0)	0,03733 (0)	0,05124 (0)	0,01203 (0,11712+-0,0273)	0,00423 (0,19820+-0,0379)	0,05384 (0)	0 (0,45045+-0,0525)	0				
Argentina (South)	0,19235 (0)	0,05493 (0,01802+-0,0121)	0,11613 (0)	0,14912 (0)	0,14829 (0)	0,07704 (0)	0,07905 (0)	0,05312 (0,00901+-0,0091)	0,04559 (0,01802+-0,0121)	0,08514 (0)	0,03778 (0,01802+-0,0121)	0,01517 (0,10811+-0,0353)	0			
Chile (North)	0,2514 (0)	0,12965 (0)	0,12565 (0)	0,18677 (0)	0,18709 (0)	0,10553 (0)	0,07947 (0)	0,06338 (0)	0,08051 (0)	0,1572 (0)	0,0825 (0)	0,07076 (0)	0,08325 (0)	0		
Chile (Southern)	0,35061 (0)	0,21149 (0)	0,13098 (0)	0,31322 (0)	0,30123 (0)	0,21571 (0)	0,18016 (0)	0,15769 (0)	0,16303 (0)	0,22896 (0)	0,13583 (0)	0,12393 (0)	0,12115 (0)	0,03843 (0)	0	
Paraguay (Alto Paraná)	0,16616 (0)	0,03636 (0,02703+-0,0194)	0,0787 (0)	0,1134 (0)	0,10832 (0)	0,04519 (0,00901+-0,0091)	0,03833 (0)	0,01301 (0,07207+-0,0297)	0,01345 (0,05405+-0,0201)	0,07062 (0)	0,00586 (0,16216+-0,0227)	0,00633 (0,09009+-0,0271)	0,03633 (0,01802+-0,0121)	0,07178 (0)	0,13029 (0)	0

9.2. Considering regions North and South to group the considered Colombian departments.

	Colombia (South-Andes)	Colombia (North-Andes)	Peru (Lima)	Brazil (Rio de Janeiro)	Brazil (Espírito Santo)	Ecuador	Argentina (North-East)	Argentina (Central-West)	Argentina (South)	Chile (North)	Chile (Southern)	Paraguay (Alto Paraná)
Colombia (South-Andes)	0											
Colombia (North-Andes)	0,10883 (0)	0										
Peru (Lima)	0,11759 (0)	0,08746 (0)	0									
Brazil (Rio de Janeiro)	0,05177 (0)	0,05622 (0)	0,02046 (0,02703+-0,0194)	0								
Brazil (Espírito Santo)	0,02407 (0,04505+-0,0152)	0,0544 (0)	0,04786 (0)	0 (0,61261+-0,0446)	0							
Ecuador	0,07351 (0)	0,07292 (0)	0,11023 (0)	0,07546 (0)	0,05829 (0)	0						
Argentina (North-East)	0,01656 (0,04505+-0,0279)	0,09098 (0)	0,06647 (0)	0,01486 (0,09910+-0,0316)	0,00088 (0,37838+-0,0504)	0,05681 (0)	0					
Argentina (Central-West)	0,02416 (0,00901+-0,0091)	0,08152 (0)	0,05124 (0)	0,01203 (0,11712+-0,0273)	0,00423 (0,14414+-0,0337)	0,05384 (0)	0 (0,42342+-0,0430)	0				
Argentina (South)	0,06163 (0)	0,13088 (0)	0,07905 (0)	0,05312 (0)	0,04559 (0,00901+-0,0091)	0,08514 (0)	0,03778 (0,01802+-0,0121)	0,01517 (0,08108+-0,0212)	0			
Chile (North)	0,1141 (0)	0,17112 (0)	0,07947 (0)	0,06338 (0)	0,08051 (0)	0,1572 (0)	0,0825 (0)	0,07076 (0)	0,08325 (0)	0		
Chile (Southern)	0,15913 (0)	0,27985 (0)	0,18016 (0)	0,15769 (0)	0,16303 (0)	0,22896 (0)	0,13583 (0)	0,12393 (0)	0,12115 (0)	0,03843 (0)	0	
Paraguay (Alto Paraná)	0,03378 (0)	0,09383 (0)	0,03833 (0)	0,01301 (0,06306+-0,0237)	0,01345 (0,08108+-0,0212)	0,07062 (0)	0,00586 (0,15315+-0,0273)	0,00633 (0,18919+-0,0459)	0,03633 (0,02703+-0,0139)	0,07178 (0)	0,13029 (0)	0

9.3. Considering regions East and West to group the considered Colombian departments.

	Colombia (East-Andes)	Colombia (West-Andes)	Peru (Lima)	Brazil (Rio de Janeiro)	Brazil (Espírito Santo)	Ecuador	Argentina (North-East)	Argentina (Central-West)	Argentina (South)	Chile (North)	Chile (Southern)	Paraguay (Alto Paraná)
Colombia (East-Andes)	0											
Colombia (West-Andes)	0,01785 (0)	0										
Peru (Lima)	0,09501 (0)	0,06622 (0)	0									
Brazil (Rio de Janeiro)	0,05292 (0)	0,02111 (0,01802+-0,0121)	0,02046 (0,02703+-0,0139)	0								
Brazil (Espírito Santo)	0,04137 (0)	0,0098 (0,09009+-0,0332)	0,04786 (0)	0 (0,50450+-0,0388)	0							
Ecuador	0,05729 (0)	0,04728 (0)	0,11023 (0)	0,07546 (0)	0,05829 (0)	0						
Argentina (North-East)	0,07349 (0)	0,02496 (0)	0,06647 (0)	0,01486 (0,09910+-0,0316)	0,00088 (0,36937+-0,0344)	0,05681 (0)	0					
Argentina (Central-West)	0,06746 (0)	0,02388 (0)	0,05124 (0)	0,01203 (0,07207+-0,0297)	0,00423 (0,27928+-0,0344)	0,05384 (0)	0 (0,50450+-0,0411)	0				
Argentina (South)	0,11414 (0)	0,06593 (0)	0,07905 (0)	0,05312 (0,00901+-0,0091)	0,04559 (0,01802+-0,0121)	0,08514 (0)	0,03778 (0,02703+-0,0139)	0,01517 (0,05405+-0,0201)	0			
Chile (North)	0,162 (0)	0,11297 (0)	0,07947 (0)	0,06338 (0)	0,08051 (0)	0,1572 (0)	0,0825 (0)	0,07076 (0)	0,08325 (0)	0		
Chile (Southern)	0,26687 (0)	0,193 (0)	0,18016 (0)	0,15769 (0)	0,16303 (0)	0,22896 (0)	0,13583 (0)	0,12393 (0)	0,12115 (0)	0,03843 (0)	0	
Paraguay (Alto Paraná)	0,08352 (0)	0,03162 (0,00901+-0,0091)	0,03833 (0)	0,01301 (0,08108+-0,0252)	0,01345 (0,07207+-0,0227)	0,07062 (0)	0,00586 (0,12613+-0,0337)	0,00633 (0,10811+-0,0402)	0,03633 (0,00901+-0,0091)	0,07178 (0)	0,13029 (0)	0

Appendix 10. AMOVA analysis results considering Native haplogroups from the South American populations. All p-values were statistically significant except when noted with an asterisk (p-value=0,08504+-0,00824).

Geographic criteria	Number of groups	Among groups (F_{CT})	Among populations/within groups (F_{SC})	Within populations (F_{ST})
(A) East / West of South America	2	0 (0)*	10.09 (0.0986)	92.19 (0.0780)
(B) Andean States / Brazil / Paraguay	3	0 (0)*	11.13 (0.1051)	94.66 (0.0533)
(C) Northern-West / Southern Cone / Brazil	3	2.52 (0.0252)*	6.93 (0.0711)	90.55 (0.0945)