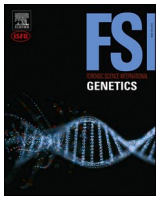
Forensic Science International: Genetics 66 (2023) 102906

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen



Population analysis of complete mitogenomes for 334 samples from El Salvador

Julen Aizpurua-Iraola [a], Raquel Rasal [b], Lourdes Prieto [c,d], David Comas [a], Núria Bonet [b], Ferran Casals [b,e,f], Francesc Calafell [a,*], Patricia Vásquez [g,*]

[a] Institut de Biologia Evolutiva (CSIC-UPF), Universitat Pompeu Fabra, Departament de Medicina i Ciències de la Vida, Barcelona, Spain
[b] Genomics Core Facility, Departament de Medicina i Ciències de la Vida, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, Spain
[c] Instituto de Ciencias Forenses, Universidad de Santiago de Compostela, Santiago de Compostela, Spain
[d] Comisaría General de Policía Científica. DNA Laboratory, Madrid, Spain
[e] Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain
[f] Institut de Biomedicina de la Universitat de Barcelona (IBUB), Universitat de Barcelona, Barcelona, Spain
[g] Asociación Pro-Búsqueda de Niñas y Niños Desaparecidos de El Salvador, San Salvador, El Salvador

ARTICLE INFO

Keywords:
Mitochondrial DNA
El Salvador
Mitogenomes

ABSTRACT

The use of mitochondrial DNA (mtDNA) in the field of forensic genetics is widely spread mainly due to its advantages when identifying highly degraded samples. In this sense, massive parallel sequencing has made the analysis of the whole mitogenome more accessible, noticeably increasing the informativeness of mtDNA haplotypes. The civil war (1980–1992) in El Salvador caused many deaths and disappearances (including children) all across the country and the economic and social instability after the war forced many people to emigration. For this reason, different organizations have collected DNA samples from relatives with the aim of identifying missing people. Thus, we present a dataset containing 334 complete mitogenomes from the Salvadoran general population. To the best of our knowledge, this is the first publication of a nationwide forensic-quality complete mitogenome database of any Latin American country. We found 293 different haplotypes, with a random match probability of 0.0041 and 26.6 mean pairwise differences, which is similar to other Latin American populations, and which represent a marked improvement from the values obtained with just control region sequences. These haplotypes belong to 54 different haplogroups, being 91% of them of Native American origin. Over a third (35.9%) of the individuals carried at least a heteroplasmic site (excluding length heteroplasmies). Ultimately, the present database aims to represent mtDNA haplotype diversity in the general Salvadoran populations as a basis for the identification of people that disappeared during or after the civil war.

1. Introduction

El Salvador is the smallest country in Central America and home to an estimated population of 6,325,827 people, 61.7% of which live in urban areas [1]. According to the last census in 2007, the majority (83%) of the Salvadoran population identifies as mestizo, with 15% identifying as European, 0.23% as Native American, and 0.13% as Afro-descendant [2]. This figure has been disputed by indigenous associations and academics, as some of the Native American groups appear to be significantly underrepresented [3] and some authors estimate the percentage of indigenous population could be as much as 12–17% of the total population[4]. Three primary indigenous groups live in the country, namely the Lencas, Kakawiras, and Nahua or Pipiles; recent migrants may include the Q'eqchi' people from Guatemala and Belize [3].

The history of El Salvador, like for most of the countries in the region, has been complex and often turbulent. The early prehistory of the region is not very well known. The first evidence of human settlement in the area are several possibly pre-Archaic (~10,000 BC) period petroglyphs in different caves along the country [5]. Different Native American groups have inhabited the region such as the Lencas, whose origin is unclear, the Mayan who reached the peak of their civilization around the 6th century CE, or the Nahua-Pipil of Toltec origin, who entered the area in the Classic or post-Classic period (~900–1200 CE) [6], [7]. During the colonial period, the Spanish empire violently conquered the region and

Abbreviations: mtDNA, mitochondrial DNA.

* Corresponding authors.

https://doi.org/10.1016/j.fsigen.2023.102906
Received 22 March 2023; Received in revised form 9 June 2023; Accepted 10 June 2023
Available online 16 June 2023

incorporated it to its territories in 1542[6]. During and after the colonial period, El Salvador was marked by political instability, economic challenges, and social inequality partly caused by the uneven distribution of wealth and political power, and foreign interference supporting local elites and suppressing movements for social change [6]. In 1932, an indigenous peasant uprising sparked a brutal repression and an ethnocide from the government, reducing the indigenous population considerably [8]. The social injustice and political instability continued and erupted in a civil war from 1980 to 1992. The war left thousands of deaths, missing people and forced many others to emigrate outside of the country. In the recent years, despite the attempt of the peace agreements to establish a framework for social and economic development, poverty, corruption and violence are still a challenge in El Salvador and forces many people to leave the country.

During the war, many children of different ages disappeared. After the peace agreements, in 1994, the families of the missing children formed the Pro-Búsqueda Association (http://www.probusqueda.org.sv/) with the objective of finding the missing children. The association manages a genetic database with profiles of both of those seeking their biological families and of those seeking their missing children. Besides this, since 2014 more than 7000 migrants (a fraction of which are from El Salvador) have disappeared in the Mesoamerican migration corridor through Mexico to the United States [9]. In 2009, the Argentine Forensic Anthropology Team (EAAF), together with different governments (including El Salvador) and relatives of deceased and disappeared migrants founded the Border Project (https://eaaf.org/proyecto--frontera/). The Border Project is an association that helps in the exchange of forensic information and remains between different countries and whose one of their main tools is the creation of forensic databanks with genetic information from the unidentified remains of migrants all along the Mesoamerican corridor to the U.S.

Mitochondrial DNA (mtDNA) has had an essential role as a biogeographical marker in different disciplines [10]. From the forensic science point of view, it is pivotal for the identification of highly degraded samples (i.e. hair samples or ancient samples), as the high copy number of molecules per cell enables the retrieval of good quality DNA, even when extraction of autosomal DNA is complicated. Besides, with the improvement of sequencing techniques, complete mtDNA samples are more and more available. These are more informative than partial mtDNA sequences (i.e. hypervariable segments within the control region) in terms of fine-scale resolution because they harbor more variation, and therefore, have gained popularity in the field in the recent years [11–13].

For each of the forensic endeavors mentioned earlier, and as a part of the standard forensic routine procedure, developing and organizing appropriate reference databases representing the target population is key [14]. The reference datasets serve as an approximation to the real population haplotype frequencies, which inform about the strength of the forensic association in the context of that population. Previous studies have provided information on haplotype frequencies of forensic markers, including autosomal STRs [15–21], X-STRs [21], [22], and Y-STRs [21], [23–26], in El Salvador. However, studies on Salvadoran mtDNA haplotypes are scarce, with only one study on control region sequences for non-forensic purposes [27]. This paucity of reference data is not limited to El Salvador but also extends to the entire Latin American region when it comes to forensic-quality full mitogenome data. We present here full mitogenome sequences for 334 individuals in the general population of El Salvador, as a resource in the identification of missing persons in the country.

2. Materials and methods

2.1. Samples

An initial number of 384 samples from either saliva or buccal cells were donated by volunteers from different cities in El Salvador as a part of the initiative carried out by the *Pro-Búsqueda* Association to build a mtDNA haplotype frequency database. All but 32 of these samples overlap with those typed for the Verogen ForenSeq™ Primer Mix A loci [21]. The participants provided informed consent and this study was approved by the National Health Research Ethics Committee of El Salvador Nº CNEIS/2018/020.

2.2. Sequencing

DNA was extracted from a total of 384 samples from buccal cells and saliva in Buccal DNA Collector (Bode Technology, 250 samples) or EasiCollect (Qiagen, 134 samples) collectors. DNA extraction was performed with the Prep Filer BTA kit (Thermofisher Scientific, Walham, MA, USA) and the extracted DNA was quantified with Qubit Broad Range (Thermofisher). Library preparation was carried out with the Kappa HyperPlus kit and Kappa Dual indexed adapters (Roche, Basel, Switzerland) were used. Then mitochondrial DNA fragments were isolated and amplified using a mtDNA capture experiment with the Xgen Hybridization capture kit (IDT, Coralville, IA, USA). The samples were pooled in sets of 24 samples in single tubes, containing 250 ng of each library, with which the capture was performed. Finally, the sequencing was performed using a MiSeq sequencer (Illumina, San Diego, CA, USA).

2.3. Sequence preprocessing

An initial quality check of the raw sequencing reads was performed with FastQC [28]. Then, we mapped the sequencing reads to the revised Cambridge Reference Sequence (rCRS) [29] using the BWA algorithm [30]. PCR duplicates were removed, and the base quality scores were recalibrated with Picard Tools [31] and a final quality report of the alignment (BAM) files was obtained with Qualimap2 [32]. Samples with no coverage at any of the reference positions (12 samples) were directly excluded. Finally, the sequence variants were called with GATK tools HaplotypeCaller and GenotypeGVCFs [33].

After these steps, the dataset underwent the EMPOP quality control [34], after which the unexpected variants and missing calls were manually checked by using Integrative Genomics Viewer v2.12.2 [35]. The final dataset consists of 334 high-quality complete mitogenome sequences, with EMPOP accession number EMP00865.

2.4. Quality assessment

We assessed the quality of the mapping and the variant calling by using the BAM and VCF files. The number of mapped reads per individual, mean coverage depth per site and the coverage across the reference were estimated by using Qualimap [32] and VCFtools [36]. The strand balance and read length were analyzed with Picard Tools CollectAlignmentSummaryMetrics [31]. Finally, as explained above, the variant calling was curated manually after the quality check by EMPOP.

2.5. Assesment of heteroplasmies

The heteroplasmies of each sample were initially assessed with mtDNA-Server [37] with a heteroplasmic detection level of 10%, meaning the minor heteroplasmic level had to be supported by 10% of the read depth for that position. After the quality check by EMPOP, the novel heteroplasmic positions were checked manually, and in the end all individuals with > 3 heteroplasmies (the expected maximum amount of heteroplasmies in an individual with a conservative MAF threshold of 10% [14]) were discarded. With a MAF threshold of 10%, 22 samples contained > 10 heteroplasmies possibly due to contamination (Haplocheck [38] was run and most samples presented contamination levels over the 10% MAF threshold), and 16 samples presenting between 4 and 10 heteroplasmies were also discarded with no clear contamination signs.

2.6. Population diversity statistics

We have computed different diversity statistics in our samples and compared them within three different reference datasets: i) a forensic-quality dataset of geographically relevant complete mitogenomes stored in EMPOP (Table 1), ii) given the scarcity of samples in the former dataset, we collected also from the literature a dataset of different populations, which covers a wider geographical range (Table 2), and iii) a forensic-quality dataset of mainly Latin American control region (HVSI positions: 15996–16401 and HVSII positions: 29–408) sequences stored in EMPOP (Table 3).

For each of the populations, including the population from El Salvador, we computed the number of haplotypes, nucleotide diversity (π), number of segregating sites (S) and the mean pairwise differences (MPD) with the *pegas* and *ape* packages in R [39], [40]. We also assessed the forensic informativity of the dataset by calculating the random match probability (RMP, i.e. the probability that two randomly taken sequences are identical) as the sum of the squared haplotype frequencies. Heteroplasmic positions and indels were not considered when computing π and MPD statistics, and the number of haplotypes and RMP statistics were calculated both with and without considering nucleotide substitution heteroplasmies (length heteroplasmies were not considered in any case). We estimated pairwise ϕ_{st} distances with Arlequin (v.3.5) [41] and the distances were plotted in two dimensions after multidimensional scaling analysis with the R *stats* package [42].

3. Results and discussion

3.1. Quality assessment of the mtDNA sequences

After quality filtering and manual curation of the original dataset, the final dataset contained 334 forensic-quality samples from the general Salvadorian population. We first assessed the quality of the mapping process for the samples. The mean number of mapped reads was 108,064 per sample (Suppl. Fig. 1). The mean read length was 172.57 base pairs, which falls within the expected range for the library preparation kit used (see Materials and Methods). The average strand balance was 49.87%, with all of the samples having between 45% and 55% strand balance. A strand balance of 50% indicates both DNA strands are sequenced equally providing more confidence in the base calling. The coverage distribution across the reference was roughly homogeneous and the mean depth per site was 536.5X (Suppl. Figs. 2 and 3), a reliable value to perform the following analyses confidently. Across individuals, the mean coverage was 984.7X, with a minimum of 16.5X (Suppl. Figure 4).

3.2. Genetic diversity, forensic informativeness and variant distribution

Within our 334 sample dataset, we found 293 different haplotypes, which reduced to 284 if heteroplasmies were not taken into account. Nucleotide diversity (π) and Mean Pairwise Differences (MPD) values for our dataset were 0.0016 and 26.6 respectively, falling within the range of other mitogenome forensic and non-forensic datasets (Tables 1 and 2). However, El Salvador seems to present lower diversity levels in comparison with other Latin American datasets and shows values closer to those of European populations. The Random Match Probability (RMP) of our dataset is 0.0044, which put into context is higher than the RMP of the Catalan dataset (N = 808, RMP=0.0014) but lower than the Alto Paraná samples (N = 105, RMP=0.0117). This agrees with the fact that RMP is inversely correlated with sample size[43].

Since whole mitogenome Latin American datasets are scarce in EMPOP, with only US Hispanics [13] and Alto Paraná in Paraguay [44], we also compared the control region (CR) from our dataset with other forensic-quality control region datasets. We found 231 different haplotypes in our CR dataset, and π and MPD values of 0.011 and 8.9 respectively. Again, these values fall within the distribution of other worldwide populations (Table 3). The RMP was 0.0064, which is close to that of other Latin American populations such as Argentina (N = 209, RMP=0.0072) or the Brazilian population in Rio de Janeiro (N = 205, RMP=0.0058), again considering that the RMP is dependent on sample size. Using the whole mitogenome rather than the control region results in a 31% reduction in RMP (from 0.0064 to 0.0044), but the improvement is particularly remarkable for MPD, which almost triples (from 8.9 to 26.6).

We found 295 singleton variants in our dataset (0.88 singletons per sample). The control region shows 0.16 singletons per site while the coding region contains 0.037 singletons per site. 56.4% and 34.1% of variants in our dataset have a MAF $\leq$ 0.5% and a MAF between 0.5% and 5% respectively. The remaining 9.4% falls in the MAF > 5% frequency bin (Suppl. Figure 5). Most variants are very rare (MAF $\leq$0.5%) or rare (0.5% <MAF$\leq$ 5%), which, when combined into haplotypes, evidence the potential of mtDNA as a forensic tool.

3.3. Heteroplasmies

The sequencing technical advances during the past years have enabled us to reach sufficient power to detect with high precision both the presence of alternate nucleotides at heteroplasmic sites and their respective frequencies. Here, we describe the point heteroplasmies detected within our samples; length heteroplasmies are not assessed due to the high occurrence of certain heteroplasmies at poly-C tracts such as 303–315. We detected 158 heteroplasmies in 120 individuals, in other words, 35.9% of the samples carried at least one heteroplasmy. Besides, 32 individuals harbored two heteroplasmies, and three samples contained three heteroplasmies each. The distribution of the number of heteroplasmies carried by each individual (Suppl. Figure 6) follows a Poisson distribution (χ^2 =5.86, p-value=0.191) which indicates that

Table 1
Genetic diversity metrics for the general population from El Salvador and reference forensic-quality whole mtDNA dataset. Nhaps: number of different haplotypes; p: nucleotide diversity; S: number of polymorphic sites; MPD: mean pairwise differences; RMP: random match probability; EMPOP: EMPOP accession id. (a) including or (b) excluding heteroplasmic sites from the calculations.

Forensic whole mtDNA sequences	N	N Haps[a]	N Haps[b]	π	S	MPD	RMP[a]	RMP[b]	EMPOP	Reference
El Salvador	334	293	284	0.0016 ± 0.0007	788	26.6 ± 11.1	0.0041	0.0044	EMP00865	Present study
Catalonia	808	777	759	0.0017 ± 0.0008	1685	28.9 ± 12.6	0.0014	0.0014	EMP00860	Font-Porterias N et al. 2022
USA African	170	169	168	0.0034 ± 0.0015	996	57.1 ± 25.6	0.006	0.006	EMP00690	Just RS et al. 2014
USA Hispanic	155	150	150	0.0024 ± 0.0011	910	40.2 ± 18.6	0.007	0.007	EMP00690	Just RS et al. 2014
USA European	263	261	260	0.0018 ± 0.0008	1008	30.7 ± 14.0	0.0039	0.0039	EMP00690	Just RS et al. 2014
Alto Paraná	105	94	89	0.0025 ± 0.001	626	41.9 ± 20.1	0.0117	0.0126	EMP00728	Simão F et al. 2019[44]

Table 2
Genetic diversity metrics for the general population from the reference non-forensic quality whole mtDNA dataset. Nhaps: number of different haplotypes; p: nucleotide diversity; S: number of polymorphic sites; MPD: mean pairwise differences; RMP: random match probability; EMPOP: EMPOP accession id. (a) including or (b) excluding heteroplasmic sites from the calculations.

Non-forensic whole mtDNA sequences	**N**	N Haps[a]	N Haps[b]	**π**	**S**	**MPD**	RMP[a]	RMP[b]	**EMPOP**	**Reference**
CEU	99	-	98	0.0017 ± 0.0008	495	27.8 ± 13.5	-	0.0103	-	1KGP[48]
FIN	99	-	80	0.0016 ± 0.0008	356	26.3 ± 12.9	-	0.016	-	1KGP
GBR	92	-	89	0.0017 ± 0.0009	434	29.0 ± 14.2	-	0.0116	-	1KGP
IBS	107	-	106	0.0016 ± 0.0008	566	26.3 ± 12.8	-	0.0095	-	1KGP
TSI	107	-	105	0.0017 ± 0.0008	567	28.5 ± 13.9	-	0.0097	-	1KGP
Iberian	1142	-	1091	0.0017 ± 0.0006	2025	27.9 ± 10.3	-	0.001	-	Silva et al. 2021[49]
Colombia	94	-	58	0.0022 ± 0.0011	469	36.3 ± 17.6	-	0.0349	-	1KGP
Mexico	67	-	64	0.0021 ± 0.001	375	34.2 ± 16.7	-	0.0163	-	1KGP
Peru	86	-	86	0.0021 ± 0.001	507	35.7 ± 17.4	-	0.0116	-	1KGP
Puerto Rico	105	-	69	0.0027 ± 0.0013	540	45.3 ± 21.9	-	0.0304	-	1KGP
NW Amazonia	432	-	267	0.0022 ± 0.001	732	37.1 ± 17.8	-	0.0069	-	Arias L et al. 2017[50]
Ecuador/Peru	223	-	201	0.0020 ± 0.001	628	33.6 ± 16.0	-	0.0063	-	Brandini S et al. 2018[51]
YRI	108	-	106	0.003 ± 0.002	651	56.6 ± 27.3	-	0.0096	-	1KGP

Table 3
Genetic diversity metrics for the general population from El Salvador and reference forensic-quality control region mtDNA dataset. Nhaps: number of different haplotypes; p: nucleotide diversity; S: number of polymorphic sites; MPD: mean pairwise differences; RMP: random match probability; EMPOP: EMPOP accession id. (a) including or (b) excluding heteroplasmic sites from the calculations.

Forensic control region sequences	**N**	N Haps[a]	N Haps[b]	**π**	**S**	**MPD**	RMP[a]	RMP[b]	**EMPOP**	**Reference**
El Salvador	334	233	231	0.011 ± 0.004	182	8.9 ± 3.4	0.0063	0.0064	EMP00865	Present study
Argentina	209	176	176	0.015 ± 0.008	198	12.1 ± 6.2	0.0072	0.0072	EMP00008	Bobillo M.C. 2010[52]
Argentina Jujuy	180	134	133	0.014 ± 0.006	171	10.6 ± 5.0	0.0101	0.010	EMP00512	Cardoso S et al. 2013[53]
Brazil	214	200	200	0.017 ± 0.008	194	13.1 ± 6.4	0.0054	0.0054	EMP00748	Dos Reis RS et al. 2019[54]
Ecuador all	107	40	40	0.012 ± 0.006	75	9.7 ± 5.0	0.116	0.116	EMP00421-EMP00422	Baeta M et al. 2012[55]
Ecuador Kichiwa	65	13	13	0.011 ± 0.006	42	8.5 ± 4.5	0.216	0.216	EMP00421	Baeta M et al. 2012
Ecuador Mestizo	42	31	31	0.014 ± 0.007	71	10.6 ± 5.5	0.044	0.044	EMP00422	Baeta M et al. 2012
Catalonia	808	609	608	0.01 ± 0.004	257	7.8 ± 3.1	0.0051	0.0051	EMP00860	Font-Porterias N et al. 2022
Nicaragua	163	105	105	0.012 ± 0.006	132	9.1 ± 4.6	0.021	0.021	EMP00515	Nuñez C et al. 2010[56]
Rio de Janeiro	205	190	190	0.019 ± 0.009	201	14.9 ± 6.9	0.0058	0.0058	EMP00697	Simão F et al. 2018[57]
Santa Catarina	80	64	63	0.015 ± 0.008	130	11.7 ± 6	0.0188	0.0191	EMP00084	Palencia L et al. 2010[58]
USA Hispanic	128	116	116	0.017 ± 0.009	174	13.3 ± 6.7	0.01	0.01	EMP00051	Saunier JL et al. 2008[59]
Venezuela	101	78	78	0.017 ± 0.008	130	13.3 ± 6.3	0.0185	0.0185	EMP00297	Castro de Guerra D et al. 2012 [60]
NE Mexico	179	115	113	0.015 ± 0.007	155	11.5 ± 5.7	0.0177	0.0178	EMP000849–852	Bodner M et al. 2021[61]
NW Mexico	388	168	167	0.013 ± 0.006	169	10.0 ± 2.5	0.0195	0.0196	EMP000849–852	Bodner M et al. 2021
Center E Mexico	265	190	188	0.014 ± 0.007	194	10.7 ± 5.2	0.0076	0.0078	EMP000849–852	Bodner M et al. 2021
Center W Mexico	357	229	226	0.014 ± 0.006	210	11.1 ± 5.0	0.0073	0.0076	EMP000849–852	Bodner M et al. 2021
SE Mexico	80	57	56	0.012 ± 0.006	104	9.5 ± 4.7	0.0244	0.0253	EMP000849–852	Bodner M et al. 2021
SW Mexico	629	264	261	0.012 ± 0.005	202	9.7 ± 4.0	0.0153	0.0155	EMP000849–852	Bodner M et al. 2021
African Americans	170	151	151	0.018 ± 0.007	147	14.3 ± 5.8	0.0074	0.0074	EMP00690	Just RS et al. 2014
European Americans	263	223	223	0.011 ± 0.005	171	8.4 ± 3.6	0.006	0.006	EMP00690	Just RS et al. 2014

heteroplasmies accumulate randomly in the population. We found 98 heteroplasmies in the control region (0.087 heteroplasmies per base) and 60 in the coding region (0.0039 heteroplasmies per base) (Suppl. Table 1). The control region harbored 62% of the total heteroplasmies as a consequence of its higher mutation rate, in comparison with the coding region, where the heteroplasmies were randomly distributed with no position containing more than one heteroplasmy (Suppl. Figure 7). In the control region, however, the known sites 16182 M (17 individuals) and 16183 M (40 individuals) accumulated most heteroplasmies (Suppl. Figure 8), although these point heteroplasmies are really a consequence

of length heteroplasmy when 16189 shows a C residue. Most of the heteroplasmies were transitions (61.4%) and the number reaches 96.0% if 16182 M and 16183 M are excluded.

The distribution of the minor allele frequency shown in Suppl. Figure 9 is skewed towards lower values, as one would expect. Random changes in the frequencies from one generation to the next can lead to a heteroplasmy reverting to its homoplasmic state, resulting in a greater probability of the derived allele having a low frequency due to its recent origin or a decrease in its frequency.

3.4. Haplogroup description and population relationships

Of the 334 sequences, 53 haplogroups were detected in the general population of El Salvador. Whole genome sequencing allowed us to reach the most precise haplogroup classification; out of the 53 detected haplogroups, the terminal branches of 23 of them (43%) were defined solely by positions in the coding region. Most mitogenomes (91%) belonged to Native American superhaplogroups. The A2 haplogroup was the most common at 72.7%, followed by B2 at 9.6% and C1 at 5.99%. The H super haplogroup was the most common non-native haplogroup, comprising 3.9% of the frequency. In addition, some African origin sequences were also detected belonging to super haplogroups L2 (1.5%) and L3 (2.1%) (Fig. 1). These figures are comparable to those in Nicaragua or South East Mexico where A2 is the most common haplogroup (73% and 68% respectively), B2 is lower in Nicaragua (1%) than in S-E Mexico (6%) and this is reversed for the C1 haplogroup which reaches 13% in S-E Mexico but is absent in the Nicaraguan dataset (Suppl. Table 2). Most Latin American populations result from an admixture of Native American, European and African ancestries. This admixture process was sex-biased, with Native American ancestry being more predominant in the matrilineal mtDNA than in the Y chromosome or the autosomal genome [45–47]. In the populations of our reference datasets, and except for Brazil, the average frequency of Native American haplogroups in Latin America is 87%, with a range 57–100%, with the highest values in Central America and in some Mexican regions (Suppl. Table 2). The most abundant haplotype in El Salvador was an A2 haplotype, found in 6 individuals within our dataset. Haplogroups like A2 + (64) or A2w1 contained 41 and 28 different haplotypes respectively evidencing a lack of internal haplogroup refinement within these clades (Suppl. Table 3). Since the geographic origin of the samples was recorded, we were able to verify whether any geographical substructure existed between the main geographic subdivisions of the country: west, center and east. The AMOVA result yielded a − 0.38% of variation between the three regions (Monte Carlo test of significance with 1000 replicates, p-value=0.94), evidencing a lack of genetic substructure that is to be expected by inspecting the haplogroup composition of each region (Suppl. Figure 10), and in agreement with autosomal STRs [21].

In addition, ϕ_{st} values from forensic-quality mitogenomes between the general population from El Salvador and the rest of populations ranged between 0.10 with USA Hispanics to 0.28 with USA Afrodescendants (Suppl. Table 4. A). Fig. 2. A shows El Salvador is closer to the other Latin American populations (USA Hispanic and Paraná) as they all contain Native American sequences (91% in El Salvador, 61% in USA Hispanic and 86% in Paraná), but appear quite separated because they may differ in the frequencies of different Native American haplogroups (see further discussion below) (Suppl. Table 2).

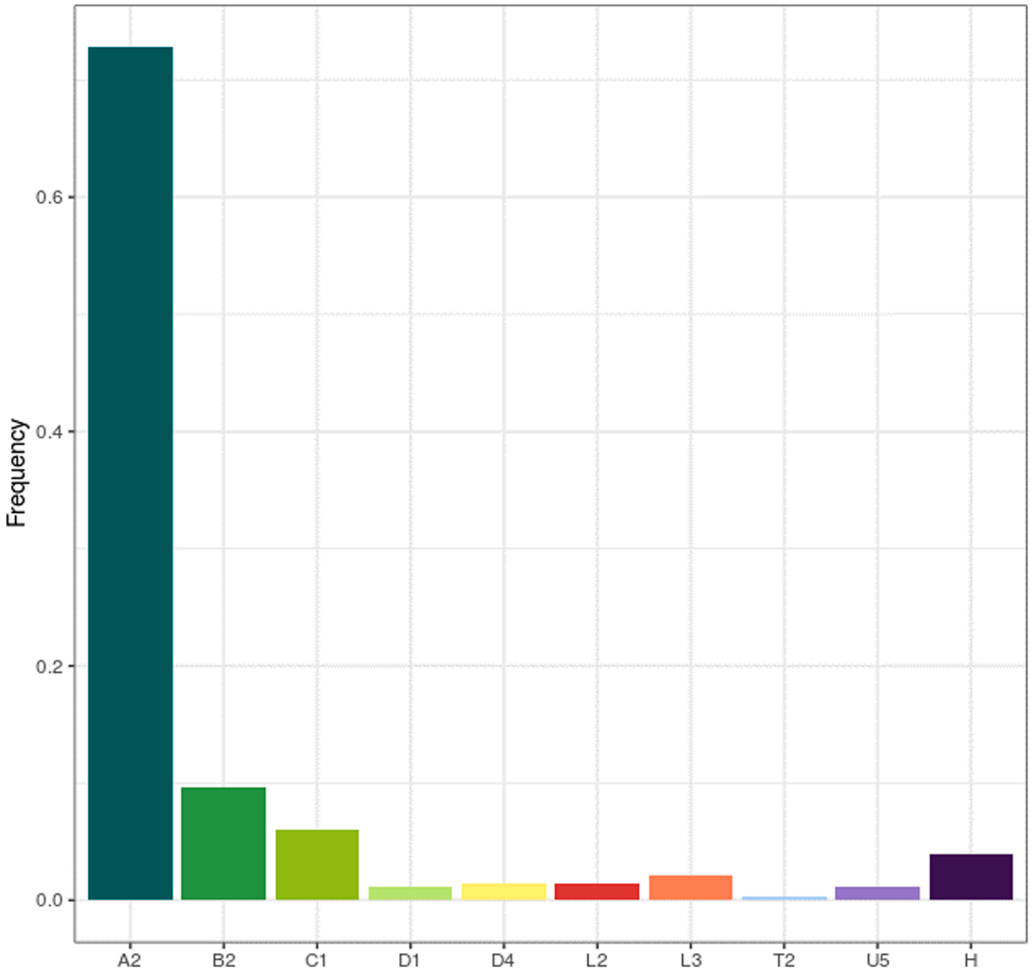


Fig. 1. Haplogroup composition for the general population from El Salvador.

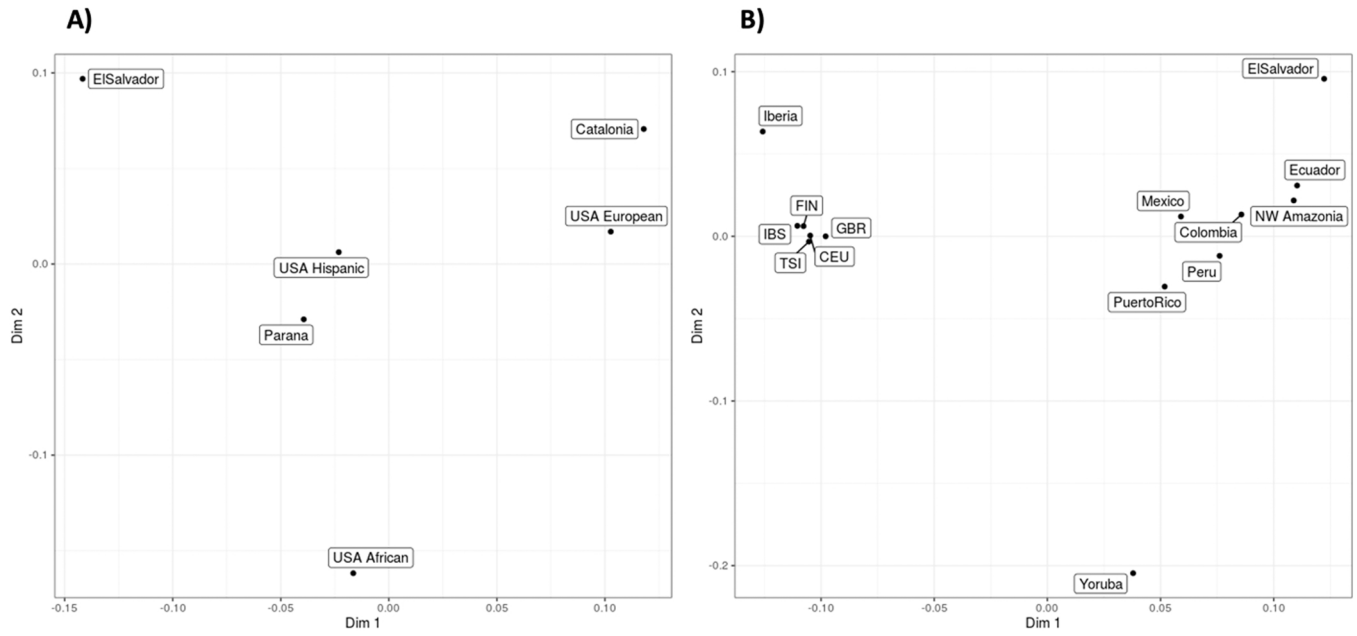


Fig. 2. Multidimensional Scaling plot of population pairwise Φst distances for (A) the dataset of forensic-quality mitogenomes and (B) the dataset of non-forensic quality mitogenomes.

Considering the non-forensic-quality datasets, El Salvador clusters together with Latin American populations and ϕ_{st} values range from 0.11 with Mexico to 0.32 with Yoruba (Fig. 2. A and 2. B, Suppl. Tables 4. B). Overall, Fig. 2. B shows El Salvador clusters together with other Latin American populations but due to the differences in frequencies in Native American haplogroups (namely A2, $p < 0.0001, \chi^2$ test), appears somewhat separated. Finally, the control region ϕ_{st} values ranged from 0.015 ϕ_{st} with southeastern Mexicans to 0.33 with Catalonia (Suppl. Table 5). In any of the datasets, all ϕ_{st} values between El Salvador and any other population were statistically significantly

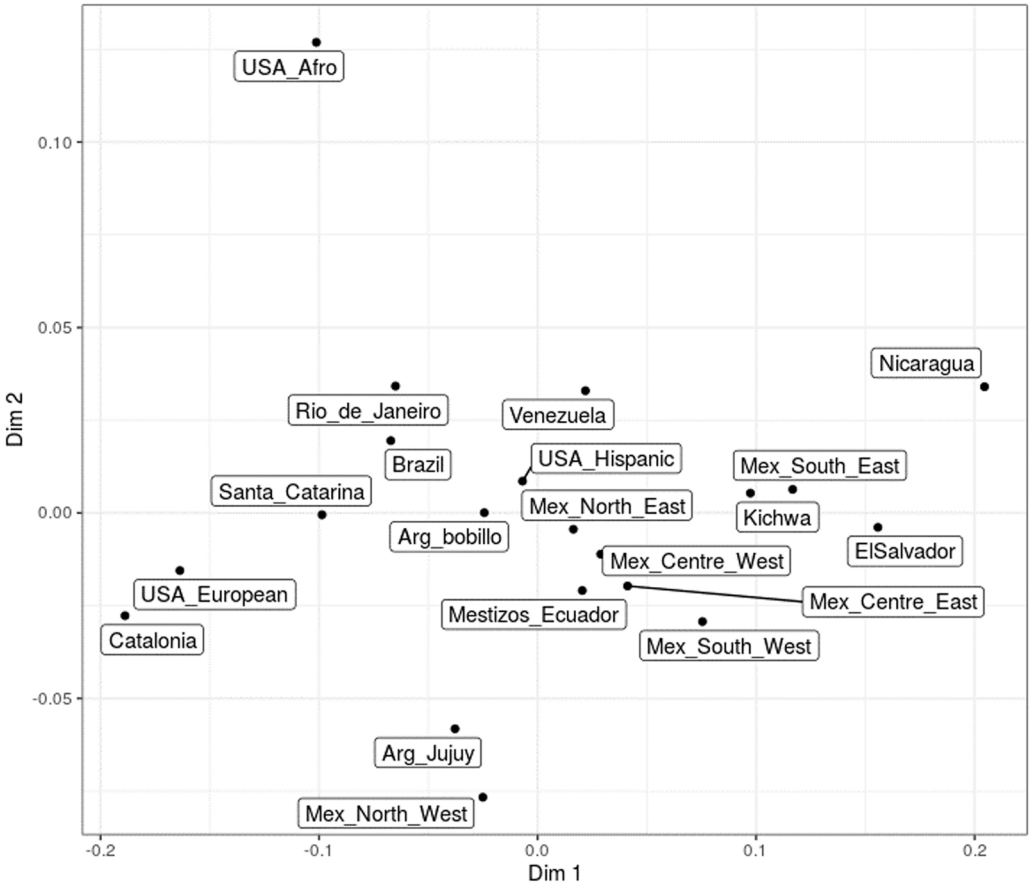


Fig. 3. Multidimensional Scaling plot of population pairwise Φst distances for the control region dataset.

different from zero ($p < 0.001$) (Suppl. Tables 4 A, 4B and 5). These significant differences were also present when only Native American sequences were considered (Suppl. Tables 6 A, 6B and 7). In Fig. 3, the Salvadoran population clusters together next to other central American populations like Southeastern Mexicans and Nicaraguans, who share high frequencies of A2 ($p > 0.05$, χ^2 test) but differ slightly in the prevalence of B2 ($p = 0.348$ with Nicaragua and $p < 0.001$ with Southeastern Mexicans), and C1 ($p = 0.001$ and $p = 0.044$) (Suppl. Table 2). Multidimensional scaling plots based on ϕ_{st} matrices computed only with Native American sequences for the non-forensic-quality mitogenomes (Suppl. Fig. 11) and control-region sequences (Suppl. Fig. 12) show again El Salvador in a peripheral position, which implies that the patterns seen when considering all sequences (Figs. 2B, 3) may be mostly due to the specificity of the Native American sequences in El Salvador.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to particularly thank all the volunteers participating in this study. We are particularly grateful to the late Cristián Orrego Benavente for this initiative and for his general contribution to the defense of human rights in El Salvador. This work was supported by the Spanish Ministry of Economy and Competitiveness and Agencia Estatal de Investigación (grant numbers CGL2016-75389-P (MINEICO/FEDER, UE), PID2019-106485GB-I00/AEI/10.13039/501100011033 (MINEICO), and "Unidad María de Maeztu" (MDM-2014-0370) to FCal and DC; Agència de Gestió d'Ajuts Universitaris i de la Recerca (Generalitat de Catalunya, grant 2017SGR00702); Agència Catalana de Cooperació al Desenvolupament (ACCD004/17/00019 and ACCD016/18/00031); Fundación Panamericana para el Desarrollo (PADF, No. PRDHD-RFA-R-2017-009). We thank also the Ministry of Health of El Salvador, which, in 2018, allowed us to take samples at their facilities.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2023.102906.

References

[1] Dirección General de Estadística y Censos (DIGESTYC), Encuesta de Hogares de Propósitos Múltiples, San. Salvador (2021).
[2] Dirección General de Estadística y Censos (DIGESTYC), Censo de Población y Vivienda, San. Salvador (2007).
[3] J. Lemus, Baja Centroamérica: El Salvador. Sociolinguistic Atlas of Indigenous Peoples in Latin America, vol. Tomo 2º, UNICEF and FUNPROEIB,, Cochabamba, 2009, pp. 789–800.
[4] Centre for the Autonomy and Development of Indigenous Peoples updated by IFAD, Country technical note on indigenous peoples' issues, San Salvador, 2017.
[5] R.D. Rivas, Investigaciones recientes en la 'Gruta del Espíritu Santo' en Corinto, Morazán, Apr. 2011, Accessed: Feb. 22, 2023. [Online]. Available: ⟨http://revistas.ues.edu.sv/index.php/launiversidad/issue/view/31⟩.
[6] Ministerio de Educación, Historia de El Salvador, Second Edition., vol. I. San Salvador: Gobierno de El Salvador, 2009.
[7] J.F. Olguín Martínez, Lemus, Jorge E. (2015): El Pueblo Pipil y su lengua: de vuelta a la vida, UniverSOS: revista de lenguas indígenas y universos culturales, ISSN 1698–6083, No. 13, 2016, págs. 265–268, no. 13, pp. 265–268, 2016, Accessed: Feb. 22, 2023. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=5815841.
[8] R.M. DeLugan, Commemorating from the Margins of the Nation: El Salvador 1932, Indigeneity, and Transnational Belonging, Anthr. Q 86 (4) (2013) 965–994 ([Online]. Available), ⟨http://www.jstor.org/stable/43652892⟩.
[9] "The Americas | Missing Migrants Project." ⟨https://missingmigrants.iom.int/region/americas⟩ (accessed Feb. 22, 2023).
[10] T. Kivisild, Maternal ancestry and population history from whole mitochondrial genomes, Invest. Genet 6 (1) (2015) 1–10, https://doi.org/10.1186/S13323-015-0022-2/FIGURES/2.
[11] L.C. Canale, W. Parson, M.M. Holland, The time is now for ubiquitous forensic mtMPS analysis, Wiley Interdiscip. Rev. Forensic Sci. 4 (1) (2022), e1431, https://doi.org/10.1002/WFS2.1431.
[12] W. Parson, et al., DNA Commission of the International Society for Forensic Genetics: Revised and extended guidelines for mitochondrial DNA typing, Forensic Sci. Int Genet 13 (2014) 134–142, https://doi.org/10.1016/J.FSIGEN.2014.07.010.
[13] R.S. Just, et al., Development of forensic-quality full mtGenome haplotypes: success rates with low template specimens, Forensic Sci. Int Genet 10 (1) (2014) 73–79, https://doi.org/10.1016/J.FSIGEN.2014.01.010.
[14] R.S. Just, J.A. Irwin, W. Parson, Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing, Forensic Sci. Int Genet 18 (2015) 131–139, https://doi.org/10.1016/J.FSIGEN.2015.05.003.
[15] J. Lovo-Gómez, A. Salas, Á. Carracedo, Microsatellite autosomal genotyping data in four indigenous populations from El Salvador, Forensic Sci. Int 170 (1) (2007) 86–91, https://doi.org/10.1016/J.FORSCIINT.2006.05.031.
[16] B. Martínez-Jarreta, P. Vásquez, E. Abecia, M. Garde, I. de Blás, B. Budowle, Autosomic STR Loci (HUMTPOX, HUMTH01, HUMVWA, D18S535, D1S1656 and D12S391) in San Salvador (El Salvador, Central America), JFS2003395-2, J. Forensic Sci. 49 (3) (2004), https://doi.org/10.1520/JFS2003395.
[17] J.C. Monterrosa, J.A. Morales, O. García, Genetic variation for 15 short tandem repeat Loci in an El Salvadoran (Central America) population, J. Forensic Sci. 51 (2) (2006) 451–452, https://doi.org/10.1111/J.1556-4029.2006.00097.X.
[18] J.C. Monterrosa, J. Morales, I. Yurrebaso, O. García, Population genetic data for 16 STR loci (PowerPlex ESX-17 kit) in El Salvador, Forensic Sci. Int Genet 6 (5) (2012), e134, https://doi.org/10.1016/j.fsigen.2011.12.004.
[19] P. Muñoz, E.L. Pinto de Erazo, C. Baeza, E. Arroyo-Pardo, A.M. López-Parra, Genetic polymorphism of 15 STR loci in El Salvador, Int J. Leg. Med 129 (5) (2015) 991–993, https://doi.org/10.1007/S00414-015-1148-8/METRICS.
[20] J.A. Morales, et al., Population Data on Nine STR Loci in an El Salvadoran (Central American) Sample Population, J. Forensic Sci. 47 (4) (2002) 1–2, https://doi.org/10.1520/JFS15461J.
[21] F. Casals, et al., A forensic population database in El Salvador: 58 STRs and 94 SNPs, Forensic Sci. Int Genet 57 (2022), 102646, https://doi.org/10.1016/J.FSIGEN.2021.102646.
[22] M. Baeta, et al., Study of 17 X-STRs in Native American and Mestizo populations of Central America for forensic and population purposes, Int J. Leg. Med 135 (5) (2021) 1773–1776, https://doi.org/10.1007/S00414-021-02536-9/FIGURES/2.
[23] J.C. Monterrosa, J.A. Morales, I. Yurrebaso, L. Gusmão, O. García, Population data for 12 Y-chromosome STR loci in a sample from El Salvador, Leg. Med 12 (1) (2010) 46–51, https://doi.org/10.1016/J.LEGALMED.2009.10.003.
[24] B. Martínez-Jarreta, P. Vásquez, E. Abecia, B. Budowle, A. Luna, F. Peiró, Characterization of 17 Y-STR Loci in a Population from El Salvador (San Salvador, Central America) and Their Potential for DNA Profiling, JFS2005173-JFS2005174, J. Forensic Sci. 50 (5) (2005), https://doi.org/10.1520/JFS2005173.
[25] J. Lovo-Gómez, A. Blanco-Verea, M. v Lareu, M. Brión, A. Carracedo, The genetic male legacy from El Salvador, Forensic Sci. Int 171 (2–3) (2007) 198–203, https://doi.org/10.1016/J.FORSCIINT.2006.07.005.
[26] J. Saul, M. Fondevila, A. Salas, M. Brión, M.V. Lareu, Á. Carracedo, Y-chromosome STR-haplotype typing in El Salvador, Forensic Sci. Int 142 (1) (2004) 45–49, https://doi.org/10.1016/J.FORSCIINT.2004.02.004.
[27] A. Salas, et al., Mitochondrial echoes of first settlement and genetic continuity in El salvador, PLoS One 4 (9) (2009), e6882, https://doi.org/10.1371/JOURNAL.PONE.0006882.
[28] S. Andrews, FastQC a quality control tool for high throughput sequence data, Babraham Bioinforma. (2010). ⟨https://www.bioinformatics.babraham.ac.uk/projects/fastqc/⟩. accessed Jan. 26, 2022.
[29] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet 23 (2) (1999) 147, https://doi.org/10.1038/13779.
[30] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, Mar. 2013, Accessed: Nov. 11, 2020. [Online]. Available: http://arxiv.org/abs/1303.3997.
[31] Picard Toolkit, Picard Toolkit, Broad Institute, GitHub repository, 2019, [Online]. Available: http://broadinstitute.github.io/picard/.
[32] K. Okonechnikov, A. Conesa, F. García-Alcalde, Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data, Bioinformatics 32 (2) (2016) 292–294, https://doi.org/10.1093/bioinformatics/btv566.
[33] A. McKenna, et al., The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res 20 (9) (2010) 1297, https://doi.org/10.1101/GR.107524.110.
[34] W. Parson, A. Dür, EMPOP-A forensic mtDNA database, Forensic Sci. Int Genet 1 (2) (2007) 88–92, https://doi.org/10.1016/j.fsigen.2007.01.018.
[35] J.T. Robinson, H. Thorvaldsdóttir, A.M. Wenger, A. Zehir, J.P. Mesirov, Variant review with the integrative genomics viewer, Cancer Res 77 (21) (2017) e31–e34, https://doi.org/10.1158/0008-5472.CAN-17-0337/SUPPLEMENTARY-VIDEO-S1.
[36] P. Danecek, et al., The variant call format and VCFtools, Bioinformatics 27 (15) (2011) 2156–2158, https://doi.org/10.1093/bioinformatics/btr330.
[37] H. Weissensteiner, et al., mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud, Nucleic Acids Res 44 (W1) (2016) W64–W69, https://doi.org/10.1093/NAR/GKW247.

[38] H. Weissensteiner, et al., Contamination detection in sequencing studies using the mitochondrial phylogeny, Genome Res 31 (2) (2021) 309–316, https://doi.org/10.1101/GR.256545.119.
[39] E. Paradis, J. Barrett, pegas: an R package for population genetics with an integrated–modular approach, Bioinformatics 26 (3) (2010) 419–420, https://doi.org/10.1093/BIOINFORMATICS/BTP696.
[40] E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R, Bioinformatics 35 (3) (2019) 526–528, https://doi.org/10.1093/BIOINFORMATICS/BTY633.
[41] L. Excoffier, H.E.L. Lischer, Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (3) (2010) 564–567, https://doi.org/10.1111/j.1755-0998.2010.02847.x.
[42] R.Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria, 2013. [Online]. Available: ⟨http://www.R-project.org/⟩.
[43] N. Font-Porterias, et al., Sequence diversity of the uniparentally transmitted portions of the genome in the resident population of Catalonia, Forensic Sci. Int Genet 61 (2022), 102783, https://doi.org/10.1016/J.FSIGEN.2022.102783/ATTACHMENT/616D2CD4-9278-47E7-9A08-4C55D2D4C018/MMC2.DOCX.
[44] F. Simão, et al., The maternal inheritance of Alto Paraná revealed by full mitogenome sequences, Forensic Sci. Int Genet 39 (2019) 66–72, https://doi.org/10.1016/J.FSIGEN.2018.12.007.
[45] I. Mendizabal, et al., Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba, BMC Evol. Biol. 8 (1) (2008) 1–10, https://doi.org/10.1186/1471-2148-8-213/TABLES/3.
[46] L. Ongaro, et al., The genomic impact of european colonization of the americas, e4, Curr. Biol. 29 (23) (2019) 3974–3986, https://doi.org/10.1016/J.CUB.2019.09.076.
[47] A. Moreno-Estrada, et al., Reconstructing the population genetic history of the caribbean, PLoS Genet 9 (11) (2013), e1003925, https://doi.org/10.1371/JOURNAL.PGEN.1003925.
[48] S. Fairley, E. Lowy-Gallego, E. Perry, P. Flicek, The International Genome Sample Resource (IGSR) collection of open human genomic variation resources, Nucleic Acids Res 48 (D1) (2020) D941–D947, https://doi.org/10.1093/NAR/GKZ836.
[49] M. Silva, et al., Biomolecular insights into North African-related ancestry, mobility and diet in eleventh-century Al-Andalus, Sci. Rep. 11 (1) (2021) 1–13, https://doi.org/10.1038/s41598-021-95996-3.
[50] L. Arias, C. Barbieri, G. Barreto, M. Stoneking, B. Pakendorf, High-resolution mitochondrial DNA analysis sheds light on human diversity, cultural interactions, and population mobility in Northwestern Amazonia, Am. J. Phys. Anthr. 165 (2) (2018) 238–255, https://doi.org/10.1002/AJPA.23345.
[51] S. Brandini, et al., The Paleo-Indian entry into south america according to mitogenomes, Mol. Biol. Evol. 35 (2) (2018) 299–311, https://doi.org/10.1093/MOLBEV/MSX267.
[52] M.C. Bobillo, et al., Amerindian mitochondrial DNA haplogroups predominate in the population of Argentina: towards a first nationwide forensic mitochondrial DNA sequence database, Int J. Leg. Med 124 (4) (2010) 263–268, https://doi.org/10.1007/S00414-009-0366-3.
[53] S. Cardoso, et al., Mitochondrial DNA control region data reveal high prevalence of Native American lineages in Jujuy province, NW Argentina, Forensic Sci. Int Genet 7 (3) (2013) e52–e55, https://doi.org/10.1016/J.FSIGEN.2013.01.007.
[54] R.S. dos Reis, et al., A view of the maternal inheritance of Espírito Santo populations: The contrast between the admixed and Pomeranian descent groups, Forensic Sci. Int Genet 40 (2019) 175–181, https://doi.org/10.1016/J.FSIGEN.2019.03.007.
[55] M. Baeta, et al., Mitochondrial diversity in Amerindian Kichwa and Mestizo populations from Ecuador, Int J. Leg. Med 126 (2) (2012) 299–302, https://doi.org/10.1007/S00414-011-0656-4/TABLES/2.
[56] C. Nuñez, et al., Reconstructing the population history of Nicaragua by means of mtDNA, Y-chromosome STRs, and autosomal STR markers, Am. J. Phys. Anthr. vol. 143 (4) (. 2010) 591–600, https://doi.org/10.1002/AJPA.21355.
[57] F. Simão, A.P. Ferreira, E.F. de Carvalho, W. Parson, L. Gusmão, Defining mtDNA origins and population stratification in Rio de Janeiro, Forensic Sci. Int Genet 34 (2018) 97–104, https://doi.org/10.1016/J.FSIGEN.2018.02.003.
[58] L. Palencia, et al., Mitochondrial DNA diversity in a population from Santa Catarina (Brazil): predominance of the European input, Int J. Leg. Med 124 (4) (2010) 331–336, https://doi.org/10.1007/S00414-010-0464-2.
[59] J.L. Saunier, J.A. Irwin, R.S. Just, J. O'Callaghan, T.J. Parsons, Mitochondrial control region sequences from a U.S. 'Hispanic' population sample, Forensic Sci. Int Genet 2 (2) (2008), https://doi.org/10.1016/J.FSIGEN.2007.11.004.
[60] D. Castro De Guerra, et al., Sequence variation of mitochondrial DNA control region in North Central Venezuela, Forensic Sci. Int Genet 6 (5) (2012) e131–e133, https://doi.org/10.1016/j.fsigen.2011.11.004.
[61] M. Bodner, et al., The mitochondrial dna landscape of modern mexico, Genes 12 (9) (2021) 1453, https://doi.org/10.3390/GENES12091453/S1.