Forensic Science International: Genetics 61 (2022) 102783

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Sequence diversity of the uniparentally transmitted portions of the genome in the resident population of Catalonia

Neus Font-Porterias [a,1], Carla García-Fernández [a,1], Julen Aizpurua-Iraola [a], David Comas [a], David Torrents [b,c], Rafael de Cid [d,e], Francesc Calafell [a,*]

[a] Institute of Evolutionary Biology (UPF-CSIC), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[b] Life Sciences Dpt, Barcelona Supercomputing Center (BSC), Barcelona, Spain
[c] ICREA, Barcelona, Spain
[d] Genomes for Life-GCAT lab. Germans Trias i Pujol Research Institute (IGTP), Badalona, Spain
[e] In behalf of Genomes for Life GCAT Project, Spain

ARTICLE INFO

Keywords:
Mitochondrial DNA
Y-chromosome
Whole uniparental sequences
Genetic census
Catalonia

ABSTRACT

Genomic reference databases of residing populations are available in different countries and regions. Since they represent the whole genetic diversity of a geographical region, they have wide applications, from biomedical studies to forensic identifications. Uniparentally transmitted portions of the genome specifically are highly suitable for kinship analyses, mixed DNA cases and geographical ancestry inferences. We have sampled 808 individuals currently residing in Catalonia within the GCAT cohort, from which we have generated 808 high-quality whole mitochondrial DNA (mtDNA) genomes and 399 sequences of the male-specific part of the Y chromosome (MSY). We observe higher genetic diversity than in classical population genetics datasets. We test the robustness of whole sequences for unequivocal identifications, and we found that they have higher resolution than mitochondrial control region and Y chromosome short tandem repeats (Y-STRs), and that most of the variants they present are at low frequencies, increasing the discrimination capacity between individuals. These results confirm the forensic applicability of whole uniparental sequences and provide one of the largest high-quality reference datasets ever published.

1. Introduction

Uniparentally transmitted portions of the genome have been extensively used in forensic studies. They are a powerful tool to infer biogeographical origin, kinship or paternity, and even identification. The mitochondrial genome has become of special relevance when dealing with highly-degraded samples (e.g. ancient bones and hair shafts), where the extraction of mitochondrial DNA (mtDNA) is possible even when autosomal DNA is hard to retrieve, because each cell contains a high number of mtDNA copies. Due to the development of new sequencing protocols, complete mtDNA sequences are increasingly available and have proved to outperform control region-based analyses with a higher fine-scale resolution [1–3]. The non-recombining part of the Y chromosome (MSY) is the region that does not recombine with the pseudoautosomal parts of the X chromosome. Within this region, 8.9 Mb of sequence are single-copy and much easier to map and to call variation on than the rest. These unique regions of the MSY have been widely used in population genetic studies [4–6]. The advent of cost-effective techniques to retrieve the MSY, such as target enrichment or flow cytometry, is increasing the availability of reference datasets for different populations [7]. These techniques have also improved the extraction of Y chromosome information from ancient or degraded samples [8], increasing the potential for using whole MSY sequences in forensic studies [9]. Combining the analysis of both uniparentally transmitted portions of the genome could complete the information in complex cases, such as those involving degraded or mixed DNA samples.

Catalonia occupies the NE corner of the Iberian Peninsula, limiting with France and Andorra to the North, the Spanish regions of Aragon and Valencia to the West and South, and the Mediterranean Sea to the East. It occupies 32,108 Km2 and has a population of 7,727,029 (2020). The distinct Catalan culture and identity is mostly based on the Catalan language, which is also spoken, in different varieties, in SW France,

* Corresponding author.
E-mail address: francesc.calafell@upf.edu (F. Calafell).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.fsigen.2022.102783
Received 10 June 2022; Received in revised form 23 September 2022; Accepted 26 September 2022
Available online 28 September 2022

Valencia, and the Balearic Islands. Historically a population crossroads, the local Iron-Age Iberians had contacts with Greek settlers and Phoenician merchants, but were brought into the Roman dominion in the 2nd century BCE. After the Germanic invasions, the region was completely or partly under Muslim, North African control between the 8th and the 12th centuries CE. A dynastic union brought the Catalan counties (and foremost among them, the earldom of Barcelona) under the kingdom of Aragon, which, between the 13th and 15th centuries expanded southwards towards Valencia and eastwards to the Western Mediterranean, including southern Italy, Corsica, Sardinia, Sicily, and briefly Athens. The late 16th and early 17th centuries saw an influx of French migrants fleeing the Religion Wars which assimilated into the local population even if their numbers may have been up to a quarter of the local population [10]. Catalonia industrialized earlier and more intensively than the rest of Spain, which attracted immigration.

The genetic profile of the Catalan population has been previously characterized in autosomal SNP-array population genetic studies of the Iberian Peninsula [11]. As for mitochondrial DNA, 121 mitogenomes from Catalonia were sequenced as part of a large Spanish dataset [12], and showed haplogroup frequencies that were similar to the country-wide averages. In previous publications, the hypervariable region I (and, occasionally, hypervariable region II) were sequenced and showed haplotype and haplogroup distributions within the Western European diversity [13–15]. As for the Y chromosome, a survey of the MSY haplogroups in the Iberian Peninsula showed the North-to-South gradients that were also prominent in autosomal SNPs [11]. A notable feature of haplogroup frequencies in Catalonia is the high frequency of R1b-DF27 [16], and, in particular, of R1b-M167 [17]. Over 60 SNPs and the Yfiler™ Y-STR kit were typed in > 1500 Catalan men in a study of surnames [18]. Other studies of Y-STRs in Catalonia can be found in references [19,20].

However, the studies cited above were carried out on samples of autochthonous Catalans, which represent only a fraction of the actual resident population in the region. Several waves of migration from elsewhere in Spain (particularly in the 1920 s, 1950 s and 1960 s) and from abroad (mostly from Morocco, Romania and Latin America, from ~1995–2008) constitute a non-negligible part of the population currently residing in Catalonia. Today 16.0% of the Catalan residents were born elsewhere in Spain (and probably as many are second- and third-generation immigrants) and 20.4% were born abroad (data from the Catalan Institute of Statistics, www.idescat.cat). For forensic purposes, the creation of reference databases is needed to cover the present genetic diversity of the population. In this study we aim to describe the current genetic diversity of the Catalan resident population, obtaining a comprehensive database for future studies. To do so we analyse the mtDNA and MSY sequences from a cohort of 808 residents in Catalonia, assessing also the quality and suitability of the whole sequences to unequivocally identify individuals.

2. Materials and methods

2.1. Sampling

The samples used in this study are a subset of the GCAT Genomes For Life Cohort [21,22] (see full details in www.genomesforlife.com). The participants were blood donors with access to the public national healthcare system, recruited from the general population (2014–2017) with the only restriction of having lived for at least five years in Catalonia and being aged between 40 and 65 years. All participants who agreed to be part of the study provided informed consent and were asked to sign a consent agreement. This work was approved by Hospital Germans Trias i Pujol IRB, ref. PI-19–081, on April 5th, 2019.

2.2. Sequencing

A random, gender-balanced sample of 808 individuals was selected for whole-genome sequencing. Sequencing data from selected individuals was obtained from parallel-short-reads sequences using HiSeq 4000 sequencer (Illumina, 30X coverage, read length 150 bp, insert size 600 bp) in FASTQ format. FASTQ files are deposited to the European Genome-Phenome Archive (EGA, EGAS00001003018) [21].

2.3. Sequence preprocessing

We mapped the raw sequencing reads to the human reference genome hs37d5 including rCRS mtDNA reference [23], using the BWA mem algorithm [24]. Only those reads mapping to the mtDNA and the Y chromosome (Ychr) were selected; PCR duplicates were removed and base quality scores were recalibrated using Picard tools and GATK v3.7–0 [25]. Following GATK best practices [26] we then called the sequence variants using HaplotypeCaller and GenotypeGVCFs GATK tools [25].

The mtDNA variant calls were manually curated using the BAM files and the Integrative Genomics Viewer v 2.12.2 [27], with special attention to unexpected variants or missing calls. The final mtDNA dataset underwent EMPOP [28] quality control (EMPOP accession id EMP00860) and contains 808 high quality complete sequences with 1767 polymorphic sites in 16,569 base pairs (Table S1). Analyses performed using the whole mtDNA sequence include the range 1–16,569 and control region analyses used nucleotide positions 1–576 plus 16024–16569.

We filtered the Ychr variants using VariantFiltration according to GATK best practices recommendations [26], applying a coverage filter threshold between half and double of the average coverage [29]. All the analyses were restricted to the 8.97 Mb of high quality regions of the Y chromosome as defined by Wei et al. [30]. The final MSY dataset is composed of 399 male sequences and 23,594 variant positions.

The assignment of haplogroups was performed with yHaplo [31] for the MSY, and EMMA [32] based on phylotree build 17 [33] for the mtDNA. Haplogroup assignment was subsequently manually curated, and, in the case of the MSY, independently verified by YHRD (www.yhrd.org) [34] with accession number YA004753.

2.4. Quality control analyses

The quality of the mapping process and variant calling of the uniparental sequences was controlled on the BAM and VCF files, respectively. As a measure of mapping quality, we estimated the rates of mapping efficiency (number of mapped reads / total number of reads), the mean read length and strand balance (forward strand read depth / total read depth), considering strand bias is present when it falls outside of the 0.3–0.7 interval [35]. We also estimated the proportion of PCR duplicates using SAMtools [36], and the coverage per site using VCFtools [37]. We next analysed the quality of the MSY variant calling by calculating the genotype missingness per site and per sample, the genotype quality (phred-scaled score of the confidence of genotype assignation), and the SNP quality (phred-scaled score of the confidence that the site is a variant) of each variant using VCFtools [37]. mtDNA variant calling was manually curated as explained above.

2.5. Statistical analyses

We constructed two different reference datasets, one for each uniparentally transmitted portion of the genome. For the MSY we used 54 Spanish (IBS, from the 1000 Genomes Project [38]), 15 Spanish Basques [39], 53 Tuscan (TSI) [38] and 27 other Italians [39] (7 Bergamese, 6 Tuscan and 14 Sardinian), 11 French [39], 46 British (BRI) [38], 7 Orcadian [39], 60 European-Americans (CEU) [38], 38 Finns (FIN) [38], and 12 samples from North Africa [40] (2 Algerian non-Imazighen, 2 Tunisian Imazighen, 2 Algerian Imazighen Zenata, 2 Egyptian non-Imazighen, 2 Moroccan non-Imazighen, 1 Saharawi and 1 Tunisian non-Imazighen). Samples from the Human Genome Diversity Project

Table 1

Genetic diversity metrics for the mtDNA GCAT cohort and mtDNA reference dataset 1. N Haps = number of haplotypes. π = nucleotide diversity. S= number of segregating sites. MPD = mean pairwise differences. RMP = Random Match Probability for the whole mtDNA sequence. N shared Hap = Number of shared haplotypes between Catalonia and reference datasets. Freq. shared hap = Frequency of shared haplotypes between Catalonia and reference datasets. Number of haplotypes and RMP calculated considering or not heteroplasmic positions (indicated with [a] and [b] respectively). MPD, S, and π calculated without considering heteroplasmic positions, insertions or deletions. A. Metrics using the whole mtDNA sequence (1–16,569). B. Metrics from the control region (16024–576). See Table S3 for diversity measures of the mtDNA reference dataset 2.

A) Whole mtDNA sequence	N	N Haps[a]	N Haps[b]	π	S	MPD	RMP[a]	RMP[b]	N shared Hap	Freq. shared Hap	EMPOP	Reference
Catalonia	808	777	759	0.0017 ± 0.0008	1685	28.9 ± 12.6	0.0014	0.0014			EMP00860	present study
Basques1	177	149	149	0.0015 ± 0.0007	490	24.1 ± 10.6	0.0095	0.0095	10	0.1525	EMP00756	García O et al. 2020
Zamora	101	98	96	0.0018 ± 0.0009	540	30.0 ± 13.2	0.0105	0.0109	4	0.0495	EMP00555	Ramos A et al. 2013
Sephardic_Portuguese	57	47	47	0.0017 ± 0.0009	279	28.7 ± 12.7	0.0274	0.0274	2	0.0351	EMP00619	Nogueiro I et al. 2015
USEuropeans1	83	83	83	0.0019 ± 0.0009	492	31.3 ± 13.8	0.012	0.012	1	0.0120	EMP00659	King JL et al. 2014
USEuropeans2	263	261	258	0.0019 ± 0.0009	938	30.7 ± 13.4	0.0039	0.0040	0	0.0000	EMP00689	Just RS et al. 2015
Serbian	225	212	211	0.0017 ± 0.0008	773	27.5 ± 12.1	0.0050	0.0050	0	0.0000	EMP00739	Davidovic S et al. 2020
Hungarian	96	91	91	0.0018 ± 0.0009	480	29.1 ± 12.82	0.0115	0.0115	1	0.0044	EMP00735	Malyarchuk B et al. 2018
Armenian	206	187	187	0.0021 ± 0.001	909	34.17 ± 14.9	0.0058	0.0058	0	0.0000	EMP00740	Margaryan A et al. 2017
B) Control Region	**N**	**N Haps[a]**	**N Haps[b]**	**π**	**S**	**MPD**	**RMP[a]**	**RMP[b]**	**N shared Hap**	**Freq. shared Hap**	**EMPOP**	**Reference**
Catalonia	808	662[a]	626	0.0078 ± 0.004	262	8.8 ± 4.1	0.0033	0.0042			EMP00860	present study
Mallorca	79	72	72	0.0077 ± 0.004	98	8.6 ± 4.0	0.0155	0.0155	15	0.2031	EMP00672	Ferragut JC, et al. 2015
Xuetes	104	60	60	0.0085 ± 0.004	93	8.4 ± 3.9	0.0201	0.0201	10	0.194	EMP00672	Ferragut JC, et al. 2015
Basques1	177	109	109	0.0063 ± 0.003	115	7.1 ± 3.4	0.0182	0.0182	26	0.3979	EMP00756	García O et al. 2020
Basques2	106	70	70	0.0075 ± 0.004	93	8.4 ± 3.9	0.0201	0.0201	24	0.4661	EMP00365	Cardoso S et al. 2012
Basques3	158	97	88	0.0071 ± 0.004	101	8.0 ± 3.7	0.0173	0.018	21	0.3478	EMP00668	Palencia-Madrid L et al. 2017
Zamora	101	89	86	0.0084 ± 0.004	124	9.4 ± 4.4	0.0130	0.0136	16	0.2376	EMP00555	Ramos A et al. 2013
PasValley	61	34	34	0.0067 ± 0.003	60	7.4 ± 3.5	0.0523	0.0523	6	0.2334	EMP00400	Cardoso S et al. 2010
Portuguese1	292	248	248	0.0082 ± 0.004	177	9.2 ± 4.2	0.0047	0.0047	30	0.1919	EMP00292 EMP00552–554	Marques SL et al. 2015
Portuguese2	121	75	75	0.0071 ± 0.004	101	8.0 ± 3.7	0.0200	0.0200	13	0.2729	EMP00617	Mairal Q et al. 2013
Sephardic_Portuguese	57	40	40	0.0084 ± 0.004	124	9.4 ± 4.4	0.0130	0.0130	6	0.1577	EMP00619	Nogueiro I et al. 2015
USEuropeans1	83	81	81	0.0089 ± 0.005	118	10.0 ± 4.6	0.0130	0.0130	9	0.1342	EMP00659	King JL et al. 2014
USEuropeans2	263	241	237	0.0086 ± 0.004	176	9.6 ± 4.4	0.0048	0.0051	31	0.1757	EMP00689	Just RS et al. 2015
Dutch	678	504	496	0.0080 ± 0.004	116	9.0 ± 4.2	0.0038	0.0040	40	0.1901	EMP00666	Chaitanya L et al. 2016
Germany	100	91	90	0.0082 ± 0.004	114	9.2 ± 4.3	0.0122	0.0126	15	0.202	EMP00020	Brandstätter A et al. 2006
Austria	273	227	222	0.0078 ± 0.004	167	8.7 ± 4.0	0.0060	0.0063	31	0.1846	EMP00001	Brandstätter A et al. 2007
Croatian	200	182	181	0.0086 ± 0.004	176	9.6 ± 4.4	0.0048	0.0061	17	0.1155	EMP00738	Barbarić L et al. 2020
Serbian	225	191	191	0.0079 ± 0.004	165	8.9 ± 4.1	0.0062	0.0062	23	0.1789	EMP00739	Davidovic S et al. 2020
Hungarian	96	83	83	0.0080 ± 0.004	116	9.0 ± 4.2	0.0148	0.0148	9	0.1263	EMP00735	Malyarchuk B et al. 2018
Greece	319	256	246	0.0084 ± 0.004	124	9.4 ± 4.4	0.0050	0.0061	18	0.1033	EMP00026	Irwin J et al. 2008
Armenian	206	177	177	0.0098 ± 0.005	197	11.0 ± 5.0	0.0066	0.0066	8	0.0636	EMP00740	Margaryan A et al. 2017
Lebanon	195	177	172	0.0091 ± 0.005	175	10.2 ± 4.7	0.0061	0.0064	7	0.0672	EMP00717	Zimmermann B et al. 2019
Cyprus	91	76	76	0.0092 ± 0.005	179	10.3 ± 4.7	0.0168	0.0168	5	0.1444	EMP00016	Irwin J et al. 2008

(*continued on next page*)

Table 1 (*continued*)

A) Whole mtDNA sequence	N	N Haps[a]	N Haps[b]	π	S	MPD	RMP[a]	RMP[b]	N shared Hap	Freq. shared Hap	EMPOP	Reference
Jordan	213	185	182	0.0117 ± 0.006	194	13.2 ± 5.9	0.0063	0.0065	1	0.0047	EMP00333	Zimmermann B et al. 2019
Iraqi	203	148	148	0.0095 ± 0.005	134	10.7 ± 4.9	0.0086	0.0086	13	0.0846	EMP00814	Jabbar SM et al. 2021
Bahrain	202	184	182	0.0092 ± 0.005	179	10.3 ± 4.7	0.0063	0.0065	13	0.1197	EMP00012	Zimmermann B et al. 2019

were downloaded in VCF format and lifted to the hs37d5 reference with Picard tools LiftoverVcf [41], and samples from Serra-Vidal et al. were downloaded in fastq format and processed with the same pipeline as the GCAT cohort. For the mtDNA, we included the following datasets as reference: [101] individuals from Zamora (Spain) [42], 57 Sephardic Portuguese [43], 83 US European-descent individuals (USEuropeans1) [44], 263 US European-descent individuals (USEuropeans2) [45], 96 Hungarians [46], 225 Serbians [47], 206 Armenians [48], and 177 Spanish Basques (Basques1) [35]. All these datasets are deposited to EMPOP and underwent the corresponding EMPOP [28] quality controls (mtDNA reference dataset 1) (Table 1). To increase geographic coverage of the reference datasets, we also downloaded from Genbank in FASTA format the following datasets, all of which were published as population studies: 181 French [49–54], 204 North Africans (32 Mozabites, 19 Egyptian non-Imazighen, 48 Moroccan non-Imazighen, 47 Moroccan Imazighen, and 52 Tunisian non-Imazighen [49,52,55–66]), 65 Portuguese [64,67], 352 Italians and 28 Sardinians [52,53,55,62,68–77] (mtDNA reference dataset 2). For the mtDNA control region analyses, we included an additional set of reference datasets with control region sequences deposited at EMPOP (Table 1).

The forensic informativity of the haploid sequences generated was measured with the random match probability (RMP), i.e. the probability that two randomly selected sequences carry identical haplotypes by chance [78,79], for the complete mtDNA and control region using Arlequin 3.5 [80]. As population diversity parameters, we estimated a set of genetic diversity metrics for the Catalan and reference datasets with Arlequin 3.5 [80]: number of haplotypes, nucleotide diversity (π), number of segregating sites (S) and mean pairwise differences (MPD). The same software was used to perform the mismatch distribution in MSY and mtDNA Catalan cohorts. In mtDNA, indels and heteroplasmic sites were excluded from the π and MPD calculations.

We then focused on the frequency distribution of the MSY and mtDNA variants, specifically on the proportion of singletons and the minor allele frequency spectrum in the GCAT cohort using VCFtools [37] and Arlequin 3.5 [80].

Heteroplasmic sites at mtDNA were called by visual inspection of the BAM files with IGV 2.12.2. A site was deemed heteroplasmic if it contained at least two alleles with a quality-weighted frequency above 10%. For calculations such as π and MPD that require a single allele per site, the most frequent allele at heteroplasmic sites was used.

3. Results and discussion

3.1. Quality assessment of the uniparental sequences

We first assessed the quality of the sequencing and mapping processes. The proportion of PCR duplicates present in the whole genome raw reads was 9% (Fig. S1). The mean read length was 150 base pairs for mtDNA and 141 for the Ychr, which corresponds to the expected value range for the sequencing platform used (see Materials and Methods). The mean strand balance per site was 0.498 and 0.499 for mtDNA and Ychr sequences (Fig. S2), and, in both cases, the fraction of sites in which the strand balance score was within 0.3 and 0.7 was > 99%. Strand balance is expected to be around 0.5 when both DNA strands are sequenced equally and it provides a higher degree of confidence in the base calling [81]. The mean numbers of mapped reads per sample were 324,843 and 7,059,683 for mtDNA and the Ychr (Fig. S3) and the mapping efficiency is close to the maximum of 1, with 0.997 and 0.996 respectively. The mean coverage per site was 1937X for mtDNA and 6.2X for the Ychr (Fig. S4), which are in both cases sufficient to perform reliable subsequent analyses. Overall, all the metrics point to a high quality of both sequencing and mapping processes.

Table 2
Genetic diversity metrics for the Catalan and reference datasets in the MSY. N Haplotypes = number of haplotypes. π = nucleotide diversity (x1000). S= number of segregating sites. MPD = mean pairwise differences. The North African group includes the following non-Imazighen populations: Algerian, Tunisian, Egyptian, Libyans, Moroccans, and one Saharawi, together with Zenata and Tunisian Imazighen.

	N	N Haplotypes	π	S	MPD
Catalonia	399	399	0.0622 ± 0.0266	23594	553.9 ± 237.0
Spanish (IBS)	54	54	0.0128 ± 0.0062	23806	115.3 ± 55.4
Basque	15	15	0.0266 ± 0.0121	1443	236.9 ± 107.5
French	11	11	0.0562 ± 0.0261	1979	499.8 ± 231.7
Tuscan	6	6	0.0722 ± 0.0361	1737	642.0 ± 321.3
Tuscan (TSI)	53	53	0.0214 ± 0.0103	26221	191.9 ± 92.3
Bergamese	7	7	0.0396 ± 0.0194	1043	352.3 ± 172.2
Sardinian	14	14	0.0505 ± 0.0230	1985	449.2 ± 204.4
British (GBR)	46	46	0.0092 ± 0.0044	19839	82.6 ± 39.9
Orcadian	7	7	0.0420 ± 0.0205	977	373.2 ± 182.4
European-American (CEU)	60	60	0.0155 ± 0.0074	24093	138.8 ± 66.6
Finnish (FIN)	38	38	0.0112 ± 0.0055	17308	100.9 ± 48.9
North African non-Imazighen	7	7	0.0399 ± 0.0195	1081	355.3 ± 173.6
North African	12	12	0.0251 ± 0.0116	1127	224.0 ± 103.3

We then evaluated the quality of the MSY variant calling process. The mean genotype quality score per sample is above 98 (Fig. S5), within a scale of 0–99 [26], and the mean SNP quality score is 1878.6 (Fig. S6). SNP quality scores tend to scale to high values when quality increases [26]. Given the stringent filtering applied, each sample has a mean of 194.9 missing sites (Fig. S7), and on average each site is missing in 2.19 samples (Fig. S8). However, in relative terms, these are 0.002% of the sites and 0.30% of the samples, which implies that variant calling in our dataset is fairly robust. A distribution of the missing sites along the MSY sequence is shown in Fig. S9. We could call all sites for all samples in mtDNA, with no missing values.

3.2. Genetic diversity and forensic informativity parameters in Catalonia

Within our dataset composed of 808 and 399 mtDNA and MSY sequences, we found 777 (759 without considering heteroplasmies) and 399 unique haplotypes respectively (Tables 1, 2).

For the complete mitochondrial genome, π and MPD are 0.0019 ± 0.001 and 28.9 ± 12.6, respectively (Fig. S10). These values fall within the range for forensic-quality mitogenomes in Europe (Table 1A); for instance, almost all MPD values range between 27 and 31, with the exception of the isolated Basques (24.1), while Armenians reach 34.2. RMP is inversely correlated with sample size (whole mitogenome, Spearman's rho=−0.733, p = 0.025; control region, Spearman's rho=−0.733, p = 0.000021), which is expected, since singleton haplotypes would have a lower frequency in population samples with larger sample sizes and would contribute much less to RMP. Thus, given that our sample is almost four times larger than any EMPOP-curated, whole-mtDNA West Eurasian population sample, its RMP (0.0014) is the lowest, which increases the informativity of this sample in a forensic context. When comparing whole mtDNA haplotypes (Table S2), matches between Catalan residents and non-Iberians are rare (only one each in European-Americans and in Hungarians), while we found 2–10 matches with other Iberian samples.

Comparisons with non-forensic general population samples (reference dataset 2), chosen to improve the geographical coverage of the reference datasets (Table S3), may be harder to interpret. π and MPD values are again similar between Catalan residents and other West Eurasians, with the exception of the French, with lower values, and the North Africans, where sub-Saharan admixture carrying the divergent L-haplogroup haplotypes can explain the higher values [56]. Haplotype matches (Table S4) are rare, from four with an Italian sample and six from a French one, to 24 in a large (N = 1023) Spanish sample. Still, in relative terms, individuals carrying these matching haplotypes have joint frequencies 1–4% in the respective reference populations.

Since EMPOP-curated, whole-mtDNA West Eurasian population samples are still scarce (at the time of writing these lines, only eight population samples match these criteria), we also compared the Catalan mtDNA control region (CR) sequences. In Table 1B, we present the informativity statistics for Catalans and 25 other datasets. At different scales, informativity statistics for the CR replicate the patterns observed for the whole mitogenomes. Thus, π and MPD values are again similar between Catalan residents and other West Eurasians, albeit, for instance, MPD in the CR has a range of 8–10 differences, about three times smaller than in the whole mtDNA. And, at 0.0033, Catalans have the lowest RMP, followed closely by a large Dutch dataset (N = 678, RMP=0.0038). As expected, many more haplotype matches are found for the CR (Table S5) than for the whole mitogenome: haplotypes shared with Catalan residents make up 10–20% of the individuals in non-Iberian European datasets, while in Iberia that frequency rises to 30–40%. Again, these patterns are similar in reference dataset 2 (Table S3B, S6).

The Catalan resident mitogenomes present significant ϕ_{st} values with all reference populations except with Zamora (Spain), probably due to the low sample size of the latter. ϕ_{st} values range between 0.0007 and 0.0046 with European populations, and reach 0.01473 with Armenians (Table S7A). Considering the CR of forensic-grade mtDNA datasets, ϕ_{st} values (Table S7B) tend to be higher. They range from 0 with Mallorca (which was repopulated in the Middle Ages from Catalonia) to 0.0078, although some isolated populations such as the Pas Valley (Spain) and the Xueta crypto-Jews are more differentiated from Catalans. For Middle Eastern populations, ϕ_{st} distances to Catalans are larger and range from 0.012 to 0.040. For reference dataset 2, patterns are similar. For whole mitogenomes, ϕ_{st} values range between 0.0011 and 0.0520 with SW European populations, and they are clearly larger with North African populations (0.0351 – 0.2888) (Table S8A); similar values can be found for the control region (Table S8B).

As for the MSY, reference datasets are generally more scarce than for the mtDNA, and a forensic standard for quality control and for reporting haplotypes has not been developed for MSY whole sequences, as it exists for mtDNA. Finally, batch effects and variations in sequencing coverage may also bias comparisons. Thus, although we report below informativity statistics for some MSY datasets, comparisons should be taken with care. In particular, population samples from the 1000 Genomes Project show a high number of polymorphic sites, but lower nucleotide diversity. As mentioned above, all 399 male individuals in our sample carried different MSY haplotypes, and thus, RMP= 0. Indeed, that was the case for all reference populations (Table 2) and no haplotype matches were observed across populations either. The RMP using 23 Y-STRs at a global scale is around 5.63×10^{-5} [82]; however, it has been shown that this probability is different in each geographical region, probably due to differences in social structure [83]. As seen in Table 2, we were not aware of any West Eurasian MSY sequence datasets that represented random population samples and were as large as ours; in fact, sample sizes were an order of magnitude lower. Catalan residents had one of the largest nucleotide diversity values, with 0.0622×10^{-3}, while the range in West Eurasia was $0.009–0.075 \times 10^{-3}$. This may reflect a sampling bias, since most other datasets were composed of autochthonous rather than resident individuals. Notice that these values are much lower than those observed for mtDNA, a phenomenon that has been previously described and attributed to the smaller effective population size of males compared to females and to various historical bottlenecks in many populations [31]. Still, the size of the target regions sequenced (8.93 Mb) compensates the lower nucleotide diversity, and the MPD between MSY sequences reaches 553.9 in Catalan residents, with a range of 80–650 in the reference datasets (Fig. S10). Finally, ϕ_{st} values between Catalan and other populations (Table S9) range between 0.04 and 0.08 with other West European populations, and are larger with Sardinians (0.33) and North Africans (0.46–0.60), which reflects their different haplogroup composition [84–89].

3.3. Frequency distribution of variants

Given the lack of recombination, the basic unit of inheritance in mtDNA and the MSY is the haplotype; for instance, the relevant figure needed when assessing a match is the haplotype frequency. Still, mtDNA and MSY haplotypes are (mostly) composed of SNPs, and, ultimately, the power of mtDNA and MSY to discriminate is based on the variants they contain. The number of singletons in the mtDNA and MSY GCAT dataset is 864 (1.07 singletons per sample) and 16,352 (40.98 singletons per sample) respectively (Fig. S11). mtDNA values are in agreement with previous estimates using complete mitogenomes (1.09 singletons per sample) [90]. The mtDNA control region contains more singletons per site (0.106) than in the coding region (0.048), since the former is the most polymorphic region in the mtDNA [79]. The mean number of singletons per individual in the MSY present in our dataset is much higher than the one found in previous studies in the Spanish population (15 singletons per sample) [91], probably because this study, besides the Spanish sampled, contained populations from across Eurasia.

We next examined the minor-allele frequency (MAF) spectrum or proportion of variants in frequency bins in mtDNA and MSY sequences. There are 77.96% and 68.9% of variants with MAF $\leq$ 0.5%, 18.19% and 24.9% with MAF between 0.5% and 5%, and the remaining 3.85% and 6.14% of variants have a MAF $\geq$ 5%. Most variants are rare in the mtDNA and MSY sequences as previously suggested [92], which further exemplifies the ability to differentiate two samples using complete uniparental sequences. Previous studies have already shown the potential of complete mitogenomes for forensics [1,2,35], while YSTRs are the most analysed markers for the MSY [82,93,94]. However, MSY sequences present high levels of rare variants and singletons and can be a potential tool with forensic applications to infer paternal biogeographic ancestry, and in cases in which close male-line relatives are involved.

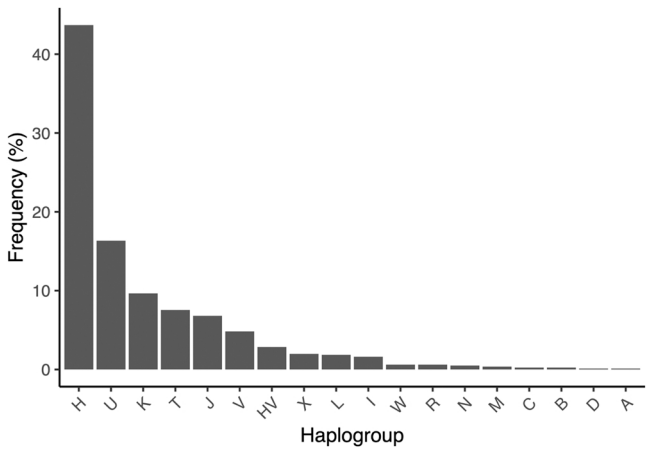

Fig. 1. Mitochondrial haplogroup (superhaplogroup) relative frequencies for the GCAT cohort.

3.4. Heteroplasmy

Massive parallel sequencing affords unprecedented power not only to detect heteroplasmic sites at mtDNA, but also to produce precise estimates of the frequency of each allele at each site. Next, we report point heteroplasmies rather than the length heteroplasmies that are exceedingly common in poly-C tracts such as 303–309 or 311–315. 207 heteroplasmic sites were detected in 181 individuals; that is, 22.4% of the individuals in our sample carried at least one heteroplasmic site. Twenty-two individuals carried two heteroplasmies, and three heteroplasmies were detected in each of two individuals. The number of individuals carrying different numbers of heteroplasmies followed a random Poisson distribution (χ^2 =0.211, p = 0.646), which means that the accumulation of heteroplasmies seemed to be independent from any individual background factors. Heteroplasmies were found in 163 different nucleotide positions (Table S10), and tended to accumulate in known sites such as 16183 M (12 samples), 16182 M (9 samples) and 16192Y (6 samples). 43% of the heteroplasmies happened in the control region, which spans only 6.8% of the mtDNA sequence. Still, 30 out of 89 heteroplasmies in the control region were found between positions 16182 and 16192, but even if those were discounted, the control region would still account for one third of the total heteroplasmies. Most (86.5%) of the heteroplasmies are transitions, and this figure reaches 95.7% if 16183 M and 16182 M are omitted.

The alleles found at a heteroplasmic site can be classified as identical to the reference (in this case, the revised Cambridge Reference Sequence [23]) or alternate. The distribution of the frequency of the alternate allele is shown in Fig. S12A; it can be seen that it is skewed towards lower values. This is expected, since the random change in frequencies at a heteroplasmic site from generation to generation may cause a heteroplasmy to revert to its original, homoplasmic state, and, thus, it is likelier that a low frequency of the derived allele be observed, either because it is recent or because it has reverted to low values from higher frequencies. Obviously, the distinction we made between reference and alternate alleles cannot be equated with ancestral and derived alleles in this mutation process. Alternatively, we can note simply the frequency of the rarer allele at a heteroplasmic site; the distribution of the minor allele frequency (Fig. S12B) also skews towards low values.

3.5. Haplogroup description

The mtDNA GCAT dataset contains 378 different haplogroups (Table S1, S11), being superhaplogroup H the most common (43.7%), followed by U (16.3%), and K (9.7%) (Fig. 1). In West Eurasia, superhaplogroup H is the most frequent [95], and within it, H1 and H3 have the highest frequencies in Catalonia (16.9% and 7.9%) and in the Iberian Peninsula [95]. The U superhaplogroup is highly common in Europe and Southwest Asia, with U5 (9.7%), U2e (2.3%) and U4 (1.6%) the most common U haplogroups in our dataset and in Europe [96–98]. Superhaplogroup K is the third most common GCAT lineage and it is mainly distributed in Europe [99]. When grouping samples according to place of birth, a similar haplogroup distribution is found among autonomous communities in Spain, suggesting that GCAT genetic diversity is not structured by geography (Table S11). As mentioned above, the GCAT cohort includes present-day residents in Catalonia (without considering genealogical background), which may explain the presence of 3.9% of the samples with a mtDNA haplogroup with an origin outside Europe. Native American haplogroups A2 + 64, B2b, B2c1, C1b, and C1c [100] are identified in 5 participants born in Ecuador, Mexico and Colombia. North African haplogroups are found in 11 individuals born in Spain: U6 is mainly distributed in West North Africa with U6b1a being a Canary

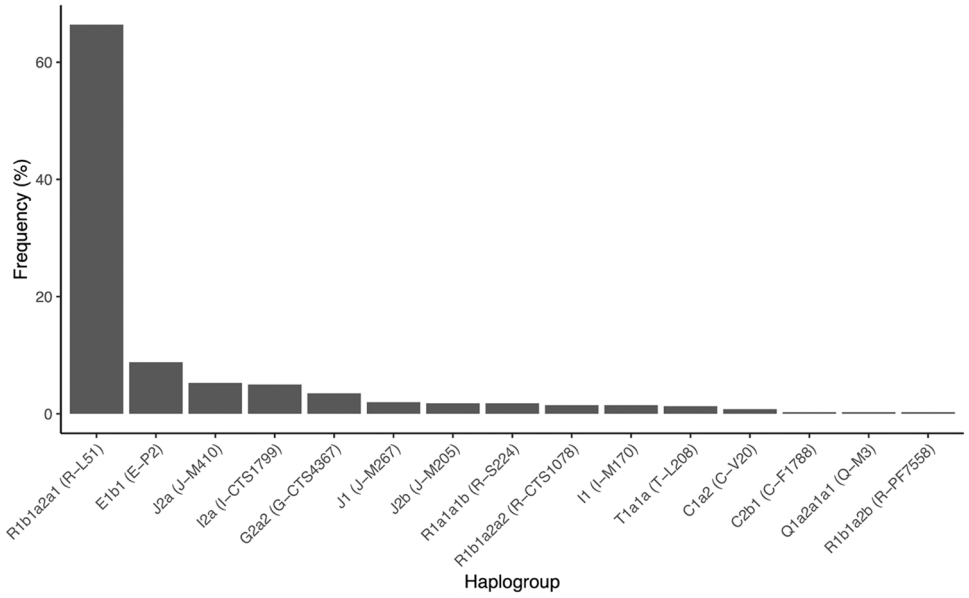

Fig. 2. Y-chromosome haplogroup relative frequencies for the GCAT cohort. Haplogroup longform follows the Y Chromosome Consortium [107] and ISOGG 2016 [108] nomenclature. A representative SNP for each haplogroup is shown between brackets.

Islands-specific lineage [60]; M1a is mostly represented in East and the Mediterranean region of North Africa [64]. L1, L2 and L3 lineages are found in 15 GCAT samples (14 individuals born in Spain and 1 in Argentina) and are widespread in Africa [101]. The presence of haplogroups with a non-European origin can be the result of recent people movement (e.g. Native American haplogroups) or more ancient gene flow (e.g. North African lineages).

A total of 121 different haplogroups were assigned in the MSY GCAT dataset (Table S12). The haplogroup with the highest frequency is R1b1a2a1 (R-L51) and its derivatives (66.4%), of which R-DF27 represents 64.1% (or 42.6% of the total sample) (Fig. 2). R-DF27 is highly prevalent in Western Europe, and especially in the Iberian Peninsula [16], where its parental haplogroup arrived at the beginning of the Bronze Age as a result of the Bell Beaker expansion [102]. The second lineage in frequency is E1b1 (E-P2, 8.8%), of which E-V13 (31.4%) and E-M183 (25.7%) are the most abundant (Fig. 2, Table S12). These lineages are very common in the Mediterranean; E-V13 specifically in Greece and the Balkans [103] and E-M183 in North African populations [6]. Similar patterns of haplogroup frequencies are found when dividing samples by the autonomous community of birth (Table S12). Six individuals (1.5%) carried MSY lineages that are not common in Europe, a proportion that is lower than that in mtDNA (3.9%; $p = 0.0238$, χ^2 test). Five out of these six individuals were born in Catalonia and carried R1a1a1b2-Z93 haplotypes, a haplogroup that has its origin in India and has been related to the Roma diaspora [104]. The remaining individual was born in Mexico and carried a Q1a2a1a1-M3 lineage, which is frequent among Native Americans [105] (his mtDNA sequence also carried a Native American lineage, B2c1). The fact that we observe just one Native American Y haplogroup in the GCAT dataset does not imply a higher female Latin American migration rate into Catalonia. It instead confirms the previously described sex-biased process during the colonization of the Americas, where gene flow was mainly driven by European men and Native American women [106].

4. Conclusion

We report here for the first time 808 mtDNA and 399 MSY complete sequences in individuals from the GCAT cohort. This dataset aims to increase the representation of the population residing in Catalonia and it could be useful as a forensic reference database given the high quality of the sequences and the high number of rare variants. In addition, this dataset presents the potential of whole mtDNA and MSY sequences, which have lower random match probabilities than mtDNA control regions and Y-STRs. We caution however that it might not include all the genetic variability in the resident population, since particular self-reported ethnic minorities (e.g. Asian populations, Romani people), age ranges and socioeconomic groups are underrepresented in the dataset.

Acknowledgments

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), PID2019–106485 GB-I00/AEI/10.13039/501100011033 (MINEICO), and "Unidad María de Maeztu" (CEX2018–000792-M) to FC and DC; and Agència de Gestió d'Ajuts Universitaris i de la Recerca (Generalitat de Catalunya, grant 2017SGR00702). Computing time at Barcelona Supercomputing Centre was granted by Red Española de Supercomputación (BCV-2019–3-00002). NF-P was supported by a FPU17/03501 fellowship. This study makes use of data generated by the GCAT=Genomes for Life. Cohort study of the Genomes of Catalonia, Fundacio IGTP with registration number PI-2018–03. IGTP is part of the CERCA Program / Generalitat de Catalunya. GCAT is supported by Acción de Dinamización del ISCIII-MINECO and the Ministry of Health of the Generalitat of Catalunya (ADE 10/00026) and; the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) (2017-SGR 529). We thank IGTP's scientific director Dr.Jordi Barretina for his support. www.genomesforlife.com/ This study was carried out using anonymized data provided by the Catalan Agency for Quality and Health Assessment, within the framework of the PADRIS Program. The authors of this study would like to acknowledge all GCAT project investigators who contributed to the generation of the GCAT data. A full list of the investigators is available from www.genomesforlife.com. We thank the Blood and Tissue Bank from Catalonia (BST) and all the GCAT volunteers that participated in the study.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2022.102783.

References

[1] M. Nilsson, H. Andréasson-Jansson, M. Ingman, M. Allen, Evaluation of mitochondrial DNA coding region assays for increased discrimination in forensic analysis, Forensic Sci. Int. Genet. 2 (2008) 1–8.
[2] R.S. Just, et al., Development of forensic-quality full mtGenome haplotypes: success rates with low template specimens, Forensic Sci. Int. Genet. 10 (2014) 73–79.
[3] 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.
[4] C. García-Fernández, et al., Sex-biased patterns shaped the genetic history of Roma, Sci. Rep. 10 (2020) 14464.
[5] T. Pinotti, et al., Y Chromosome sequences reveal a short beringian standstill, rapid expansion, and early population structure of native American Founders, Curr. Biol. 29 (2019) 149–157.e3.
[6] N. Solé-Morata, et al., Whole Y-chromosome sequences reveal an extremely recent origin of the most common North African paternal lineage E-M183 (M81), Sci. Rep. 7 (2017) 15941.
[7] W. Kutanan, et al., Contrasting maternal and paternal genetic variation of hunter-gatherer groups in Thailand, Sci. Rep. 8 (2018) 1536.
[8] M. Petr, et al., The evolutionary history of neanderthal and denisovan Y chromosomes, Science 369 (2020) 1653–1656.
[9] P. de Knijff, On the forensic use of Y-chromosome polymorphisms, Genes 13 (2022) 898.
[10] J. Nadal, E. Giralt, F. Braudel, La population catalane de 1553 à 1717. L'immigration française et les autres facteurs de son développement. VI e Section, Centre de Recherches historiques, coll. " Démographie et Sociétés ", III, Année Sociol. 19401948- 12 (1960) 266–269.
[11] C. Bycroft, et al., Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula, Nat. Commun. 10 (2019) 551.
[12] M. Silva, et al., Biomolecular insights into North African-related ancestry, mobility and diet in eleventh-century Al-Andalus, Sci. Rep. 11 (2021) 18121.
[13] S. Plaza, et al., Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean, Ann. Hum. Genet. 67 (2003) 312–328.
[14] M. Crespillo, et al., Mitochondrial DNA sequences for 118 individuals from northeastern Spain, Int. J. Leg. Med. 114 (2000) 130–132.
[15] H.B. Côrte-Real, et al., Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis, Ann. Hum. Genet. 60 (1996) 331–350.
[16] N. Solé-Morata, et al., Analysis of the R1b-DF27 haplogroup shows that a large fraction of Iberian Y-chromosome lineages originated recently in situ, Sci. Rep. 7 (2017) 7341.
[17] M.E. Hurles, et al., Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism, Am. J. Hum. Genet. 65 (1999) 1437–1448.
[18] N. Solé-Morata, J. Bertranpetit, D. Comas, F. Calafell, Y-chromosome diversity in Catalan surname samples: insights into surname origin and frequency, Eur. J. Hum. Genet. 23 (2015) 1549–1557.
[19] M. Gené, et al., Haplotype frequencies of eight Y-chromosome STR loci in Barcelona (North-East Spain), Int. J. Leg. Med. 112 (1999) 403–405.
[20] A. Pérez-Lezaun, et al., Population genetics of Y-chromosome short tandem repeats in humans, J. Mol. Evol. 45 (1997) 265–270.
[21] M. Obón-Santacana, et al., GCAT=Genomes for life: a prospective cohort study of the genomes of Catalonia, BMJ Open 8 (2018), e018324.
[22] I. Galván-Femenía, et al., Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort, J. Med. Genet. 55 (2018) 765–778.
[23] R.M. Andrews, et al., Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet. 23 (1999) 147.
[24] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinforma. Oxf. Engl. 25 (2009) 1754–1760.
[25] A. McKenna, et al., The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data, Genome Res 20 (2010) 1297–1303.

[26] M.A. DePristo, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (2011) 491–498.
[27] J.T. Robinson, et al., Integrative genomics viewer, Nat. Biotechnol. 29 (2011) 24–26.
[28] W. Parson, A. Dür, EMPOP—a forensic mtDNA database, Forensic Sci. Int. Genet. 1 (2007) 88–92.
[29] M. Mondal, et al., Y-chromosomal sequences of diverse Indian populations and the ancestry of the Andamanese, Hum. Genet. 136 (2017) 499–510.
[30] W. Wei, et al., A calibrated human Y-chromosomal phylogeny based on resequencing, Genome Res. 23 (2013) 388–395.
[31] G.D. Poznik, Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men, bioRxiv (2016), https://doi.org/10.1101/088716.
[32] A.W. Röck, A. Dür, M. van Oven, W. Parson, Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA), Forensic Sci. Int. Genet. 7 (2013) 601–609.
[33] M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation, Hum. Mutat. 30 (2009) E386–394.
[34] S. Willuweit, L. Roewer, The new Y chromosome haplotype reference database, Forensic Sci. Int. Genet. 15 (2015) 43–48.
[35] Ó. García, S. Alonso, N. Huber, M. Bodner, W. Parson, Forensically relevant phylogeographic evaluation of mitogenome variation in the Basque Country, Forensic Sci. Int. Genet. 46 (2020), 102260.
[36] H. Li, et al., The sequence alignment/map format and SAMtools, Bioinforma. Oxf. Engl. 25 (2009) 2078–2079.
[37] P. Danecek, et al., The variant call format and VCFtools, Bioinforma. Oxf. Engl. 27 (2011) 2156–2158.
[38] M. Byrska-Bishop, et al., High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios, Cell 185 (2022) 3426–3440.e19.
[39] A. Bergström, et al., Insights into human genetic variation and population history from 929 diverse genomes, Science 367 (2020) eaay5012.
[40] G. Serra-Vidal, et al., Heterogeneity in palaeolithic population continuity and neolithic expansion in North Africa, Curr. Biol. 29 (2019) 3953–3959.e4.
[41] Picard Tools. Broad Institute: ⟨http://broadinstitute.github.io/picard/⟩.
[42] A. Ramos, et al., Frequency and pattern of heteroplasmy in the complete human mitochondrial genome, PloS One 8 (2013), e74636.
[43] I. Nogueiro, J. Teixeira, A. Amorim, L. Gusmão, L. Alvarez, Echoes from Sepharad: signatures on the maternal gene pool of crypto-Jewish descendants, Eur. J. Hum. Genet. 23 (2015) 693–699.
[44] J.L. King, et al., High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq, Forensic Sci. Int. Genet. 12 (2014) 128–135.
[45] R.S. Just, et al., Full mtGenome reference data: development and characterization of 588 forensic-quality haplotypes representing three U.S. populations, Forensic Sci. Int. Genet. 14 (2015) 141–155.
[46] B. Malyarchuk, et al., Whole mitochondrial genome diversity in two Hungarian populations, Mol. Genet. Genom. 293 (2018) 1255–1263.
[47] S. Davidovic, et al., Complete mitogenome data for the Serbian population: the contribution to high-quality forensic databases, Int. J. Leg. Med. 134 (2020) 1581–1590.
[48] A. Margaryan, et al., Eight millennia of matrilineal genetic continuity in the South Caucasus, Curr. Biol. 27 (2017) 2023–2028.e7.
[49] A. Hartmann, et al., Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes, Hum. Mutat. 30 (2009) 115–122.
[50] D.M. Behar, et al., The Basque paradigm: genetic evidence of a maternal continuity in the Franco-Cantabrian region since pre-Neolithic times, Am. J. Hum. Genet. 90 (2012) 486–493.
[51] M. Ingman, H. Kaessmann, S. Pääbo, U. Gyllensten, Mitochondrial genome variation and the origin of modern humans, Nature 408 (2000) 708–713.
[52] S. Lippold, et al., Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences, Investig. Genet. 5 (2014) 13.
[53] A. Gómez-Carballa, et al., Genetic continuity in the Franco-Cantabrian region: new clues from autochthonous mitogenomes, PloS One 7 (2012), e32851.
[54] V. Guillet, et al., Adenine nucleotide translocase is involved in a mitochondrial coupling defect in MFN2-related Charcot-Marie-Tooth type 2A disease, Neurogenetics 11 (2010) 127–133.
[55] A. Olivieri, et al., The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa, Science 314 (2006) 1767–1770.
[56] D.M. Behar, et al., The dawn of human matrilineal diversity, Am. J. Hum. Genet. 82 (2008) 1130–1140.
[57] E. Musilová, et al., Population history of the Red Sea–genetic exchanges between the Arabian Peninsula and East Africa signaled in the mitochondrial DNA HV1 haplogroup, Am. J. Phys. Anthropol. 145 (2011) 592–598.
[58] H. Ennafaa, et al., Mitochondrial DNA haplogroup H structure in North Africa, BMC Genet 10 (2009) 8.
[59] N. Maca-Meyer, A.M. González, J.M. Larruga, C. Flores, V.M. Cabrera, Major genomic mitochondrial lineages delineate early human expansions, BMC Genet 2 (2001) 13.
[60] N. Maca-Meyer, et al., Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography, BMC Genet 4 (2003) 15.
[61] V. Cerný, et al., Internal diversification of mitochondrial haplogroup R0a reveals post-last glacial maximum demographic expansions in South Arabia, Mol. Biol. Evol. 28 (2011) 71–78.
[62] M. Pala, et al., Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians, Am. J. Hum. Genet. 84 (2009) 814–821.
[63] E. Pennarun, et al., Divorcing the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa, BMC Evol. Biol. 12 (2012) 234.
[64] L. Pereira, et al., Population expansion in the North African late Pleistocene signalled by mitochondrial DNA haplogroup U6, BMC Evol. Biol. 10 (2010) 390.
[65] A. Achilli, et al., Saami and Berbers–an unexpected mitochondrial DNA link, Am. J. Hum. Genet. 76 (2005) 883–886.
[66] M.D. Costa, et al., Data from complete mtDNA sequencing of Tunisian centenarians: testing haplogroup association and the 'golden mean' to longevity, Mech. Ageing Dev. 130 (2009) 222–226.
[67] L. Pereira, J. Gonçalves, H.-J. Bandelt, Mutation C11994T in the mitochondrial ND4 gene is not a cause of low sperm motility in Portugal, Fertil. Steril. 89 (2008) 738–741.
[68] F. Brisighelli, et al., The Etruscan timeline: a recent Anatolian connection, Eur. J. Hum. Genet. EJHG 17 (2009) 693–696.
[69] L. Ermini, et al., Complete mitochondrial genome sequence of the Tyrolean Iceman, Curr. Biol. 18 (2008) 1687–1693.
[70] I. Pichler, et al., Genetic structure in contemporary South Tyrolean isolated populations revealed by analysis of Y-chromosome, mtDNA, and Alu polymorphisms. 2006, Hum. Biol. 81 (2009) 875–898.
[71] A. Santoro, et al., Evidence for sub-haplogroup h5 of mitochondrial DNA as a risk factor for late onset Alzheimer's disease, PloS One 5 (2010), e12037.
[72] M.V. Zaragoza, M.C. Brandon, M. Diegoli, E. Arbustini, D.C. Wallace, Mitochondrial cardiomyopathies: how to identify candidate pathogenic mutations by mitochondrial DNA sequencing, MITOMASTER and phylogeny, Eur. J. Hum. Genet. 19 (2011) 200–207.
[73] A. Achilli, et al., Mitochondrial DNA backgrounds might modulate diabetes complications rather than T2DM as a whole, PloS One 6 (2011), e21029.
[74] N. Raule, et al., The co-occurrence of mtDNA mutations on different oxidative phosphorylation subunits, not detected by haplogroup analysis, affects human longevity and is population specific, Aging Cell 13 (2014) 401–407.
[75] M. Bodner, et al., Helena, the hidden beauty: resolving the most common West Eurasian mtDNA control region haplotype by massively parallel sequencing an Italian population sample, Forensic Sci. Int. Genet. 15 (2015) 21–26.
[76] NCBI Resource Coordinators, Database resources of the national center for biotechnology information, Nucleic Acids Res 46 (2018) D8–D13.
[77] K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, E.W. Sayers, GenBank, Nucleic Acids Res 44 (2016) D67–D72.
[78] L. Prieto, et al., The GHEP-EMPOP collaboration on mtDNA population data–A new resource for forensic casework, Forensic Sci. Int. Genet. 5 (2011) 146–151.
[79] M. Stoneking, D. Hedgecock, R.G. Higuchi, L. Vigilant, H.A. Erlich, Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes, Am. J. Hum. Genet. 48 (1991) 370–382.
[80] 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 (2010) 564–567.
[81] S. Seo, et al., Underlying data for sequencing the mitochondrial genome with the massively parallel sequencing platform ion torrent™ PGM™, BMC Genom. 16 (2015) S4.
[82] J. Purps, et al., A global analysis of Y-chromosomal haplotype diversity for 23 STR loci, Forensic Sci. Int. Genet. 12 (2014) 12–23.
[83] G. Iacovacci, et al., Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four Eastern African countries, Forensic Sci. Int. Genet. 27 (2017) 123–131.
[84] P. Francalacci, et al., Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability, Am. J. Phys. Anthropol. 121 (2003) 270–279.
[85] D. Contu, et al., Y-chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: inference for association scans, PloS One 3 (2008), e1430.
[86] S. Triki-Fendri, et al., Paternal lineages in Libya inferred from Y-chromosome haplogroups, Am. J. Phys. Anthropol. 157 (2015) 242–251.
[87] K. Fadhlaoui-Zid, et al., Sousse: extreme genetic heterogeneity in North Africa, J. Hum. Genet. 60 (2015) 41–49.
[88] A. Bekada, et al., Introducing the Algerian mitochondrial DNA and Y-chromosome profiles into the North African landscape, PloS One 8 (2013), e56775.
[89] K. Fadhlaoui-Zid, et al., Genetic structure of Tunisian ethnic groups revealed by paternal lineages, Am. J. Phys. Anthropol. 146 (2011) 271–280.
[90] M.M. Andersen, D.J. Balding, How many individuals share a mitochondrial genome? PLOS Genet 14 (2018), e1007774.
[91] C. Batini, et al., Large-scale recent expansion of European patrilineages shown by population resequencing, Nat. Commun. 6 (2015) 7152.
[92] K. Yamamoto, et al., Genetic and phenotypic landscape of the mitochondrial genome in the Japanese population, Commun. Biol. 3 (2020) 104.
[93] M. Wróbel, A. Parys-Proszek, M. Marcińska, T. Kupiec, Y chromosome sequence variation of common forensic STR markers and their flanking regions among Polish population, Forensic Sci. Int. Genet. Suppl. Ser. 7 (2019) 557–560.
[94] Â. Dente, et al., Study of Y chromosome markers with forensic relevance in Lisbon immigrants from African countries – Allelic variants study, Forensic Sci. Int. Genet. Suppl. Ser. 7 (2019) 906–907.

[95] A. Achilli, et al., The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool, Am. J. Hum. Genet. 75 (2004) 910–918.
[96] B. Malyarchuk, et al., The Peopling of Europe from the Mitochondrial Haplogroup U5 Perspective, PLOS ONE 5 (2010), e10285.
[97] M. Richards, et al., Tracing European founder lineages in the Near Eastern mtDNA pool, Am. J. Hum. Genet. 67 (2000) 1251–1276.
[98] L. Quintana-Murci, et al., Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor, Am. J. Hum. Genet. 74 (2004) 827–845.
[99] M.D. Costa, et al., A substantial prehistoric European ancestry amongst Ashkenazi maternal lineages, Nat. Commun. 4 (2013) 2543.
[100] A. Achilli, et al., The Phylogeny of the Four Pan-American MtDNA Haplogroups: Implications for Evolutionary and Disease Studies, PLOS ONE 3 (2008), e1764.
[101] A. Salas, et al., The making of the African mtDNA landscape, Am. J. Hum. Genet. 71 (2002) 1082–1111.
[102] I. Olalde, et al., The genomic history of the Iberian Peninsula over the past 8000 years, Science 363 (2019) 1230–1234.
[103] F. Cruciani, et al., Tracing past human male movements in Northern/Eastern Africa and Western Eurasia: New Clues from Y-Chromosomal Haplogroups E-M78 and J-M12, Mol. Biol. Evol. 24 (2007) 1300–1311.
[104] P.A. Underhill, et al., The phylogenetic and geographic structure of Y-chromosome haplogroup R1a, Eur. J. Hum. Genet. 23 (2015) 124–131.
[105] V. Grugni, Analysis of the human Y-chromosome haplogroup Q characterizes ancient population movements in Eurasia and the Americas, BMC Biol. 17 (2019) 3.
[106] E. D'Atanasio, et al., Y haplogroup diversity of the Dominican Republic: reconstructing the effect of the European colonisation and the trans-Atlantic slave trades, Genome Biol. Evol. 12 (2020) 1579–1590.
[107] T.Y.C. Consortium, A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups, Genome Res 12 (2002) 339–348.
[108] ISOGG. International Society of Genetic Genealogy (2016): ⟨http://www.isogg.org/⟩.