Complete mitogenome in a population sample from Cameroon

Dear Editor

Mitochondrial DNA (mtDNA) can be a useful tool for forensic applications, particularly when the amount of nuclear DNA is limited and/or is severely degraded [1]. Its maternal transmission, lack of recombination, and high mutation rate have also proved useful in population-genetics studies [2]. In forensics, large amount of data is indeed important to accurately estimate the frequency of a questioned haplotype.

EMPOP (EDNAP mtDNA Population Database, https://empop.online; [3]) has become the reference repository of high-quality mtDNA sequences. The last available version (v4/release 13) features sequences from about 50,000 individuals. However the majority of these sequences (>95%) only cover the control region (CR). As regards the proportion of entries from African individuals (about 5%), this figure shows that Africans are under-represented as ~17% of people in the world are African. Besides, only a small number of these entries in EMPOP correspond to full mitogenomes. This underrepresentation is a matter of particular concern from a population genetic and human evolution point of view, since Africa has the highest mtDNA diversity in the world [4].

To address this lacuna, we present 100 mitogenomes from Cameroonian individuals. Cameroon, a name derived from the colonial Portuguese settler's expression Rio dos Camarões, is often referred to as "Africa-in-miniature" due to its high ecological, linguistic, and ethnic diversity. The Cameroonian people practice a varied set of subsistence strategies, depending on their habitat. Around 250 languages are spoken in Cameroon, mainly belong to three big language families: the Niger-Congo, Nilo-Saharian, and Afro-Asiatic language families. Those local languages are, in general, confined to specific geographic areas in the country [5].

1. Materials and methods

Buccal swabs of 100 Cameroonian individuals of 30 ethnic groups were analyzed (Fig. 1). Samples were collected anonymously and all study participants gave their informed consent. Local permissions were obtained in Cameroon by the Ministry of Higher Education and Scientific Research, permits 0188/MINREST/B00/D00/D10/D12 and 317/MINREST/B00/D00/D10 and University of Yaounde I. Sample collection was approved by the UK ethics committee London Bentham REC (formally the Joint UCL/UCLH Committees on the Ethics of Human Research: Committee A and Alpha, REC reference number 99/0196, Chief Investigator Mark G. Thomas).

Genomic DNA was extracted by a standard phenol-chloroform separation and isopropanol precipitation procedure. All experiments were conducted in accordance with quality control measures following ISFG recommendations on good laboratory practices for databasing.

mtDNA files in FASTQ format were generated from BAM files obtained from whole exome sequencing (WES), previously mapped to the hg19 human reference genome (manuscript in preparation). Briefly, WES was performed using Illumina DNA Prep with Enrichment, following the manufacturer's recommendations. Library preparation began from 50 ng of genomic DNA, generating libraries flanked by dual indexes with mean sizes between 300 and 400 bp.

Each library pool was clustered in two lanes of the flow cell, and sequenced on the Illumina HiSeq 4000 sequencing platform using 2×75 bp paired-end reads. The sequencing experiments were carried out at the Instituto Tecnológico y de Energías Renovables (Santa Cruz de Tenerife, Spain).

To remap mitochondrial reads against rCRS (the Revised Cambridge Reference Sequence, sequence number NC_012920), the Genome Analysis Toolkit (GATK) v4 [6], SAMtools and HTSlib v1.3.1 [7] and BEDtools bamtofastq v2.26.0 [8] were integrated in an in-house bioinformatics pipeline. The rCRS-fasta file was downloaded from MITOMAP (https://www.mitomap.org/foswiki/bin/view/MITOMAP/MitoSeqs). Mapping was done using the Burrows-Wheeler Alignment Tool (BWA-MEM v.0.7.12-r1039) [9]. A fasta file was also generated from rCRS-aligned bam files, using samtools mpileup and bcftools (v1.9) (SAMtools package). The commands used are available in the Supplementary File 1. To look for unexpected or missing mutations, we further manually checked the corresponding BAM files with Integrative Genomics View (IGV v2.8.13) [10].

Mitochondrial haplogroups were assigned using HaploGrep 2, v2.1.21 [11], according to PhyloTree Build 17 (www.phylotree.org; [12]) and confirmed and/or adapted using SAM2 [13] provided by EMPOP (https://empop.online; [3]). Confidence intervals were calculated using the Beta distribution with qbeta function in base R [14]. The Cameroonian mitogenome sequences are available on EMPOP in phylogenetic alignment [13,15] under the accession number EMP00844. The entire dataset underwent EMPOP quality control according to [3].

Random match probability (RMP) and power of discrimination (haplotype diversity) were calculated as in [16], disregarding indels in poli-C and AC tracts. Mean pairwise differences (average number of nucleotide site differences between pairs of sequences) were calculated using an in-house Excel worksheet.

To visualize how our Cameroonian sample relates to other African population samples, a multi-dimensional scaling (MDS) plot was performed on haplogroup frequencies using the FinePop package [14,17]. For that, F_{ST} values were obtained by means of the Empirical Bayes estimator of F_{ST}. Publicly available mtDNA haplogroup frequencies from Rwanda ($n = 153$), Ivory Coast ($n = 100$), a mixed West African dataset ($n = 145$) [18], Ghana ($n = 192$) [19], Kenya ($n = 84$) [20], Somalia
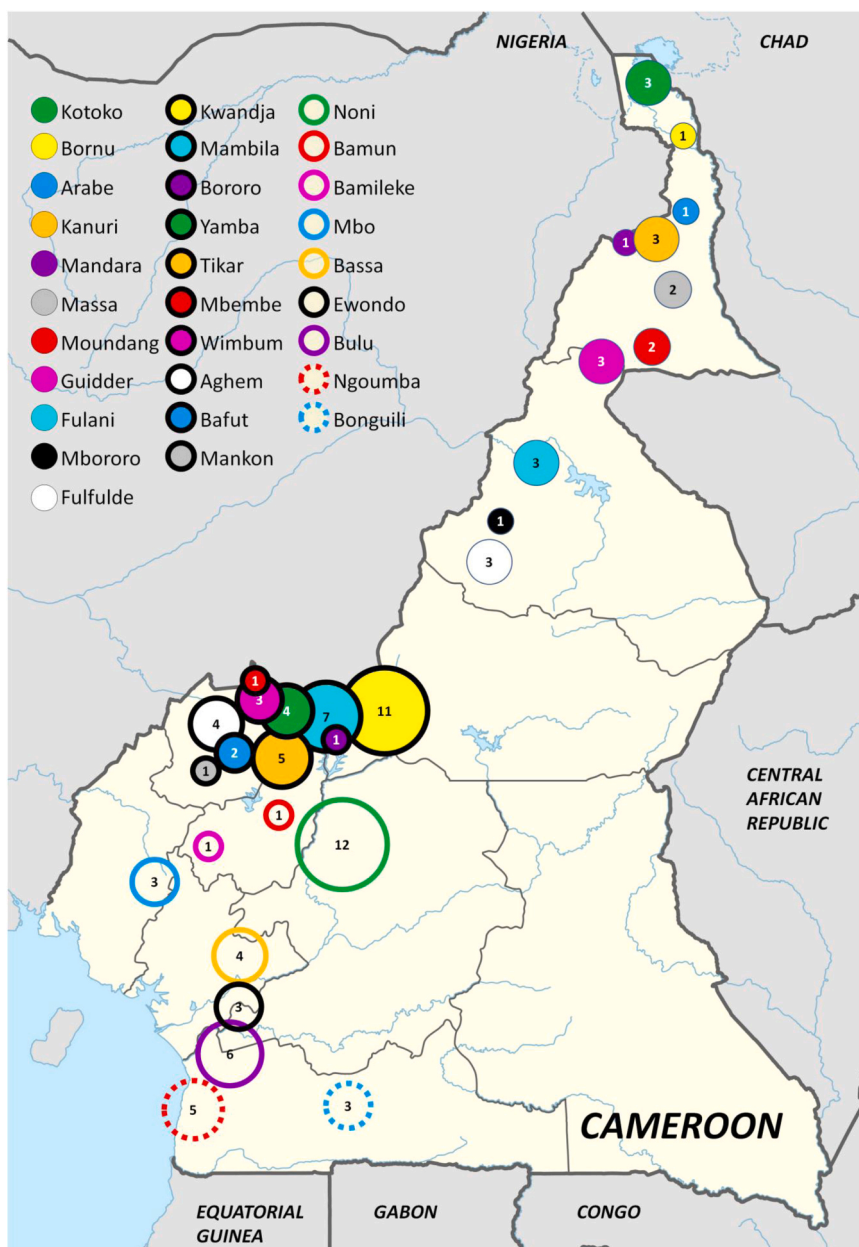
Fig. 1. Sampling locations in Cameroon. Map by: NordNordWest, Creative Commons by-sa-3.0 de license: https://creativecommons.org/licenses/by-sa/3.0/de/legalcode). Dot colors represent ethnic groups and dot sizes represent the number of samples (Map modified from: Cameroon adm location map.svg.

($n = 190$) [21], and Egypt ($n = 276$) [22] were used in this comparison. While our study is based on complete mitogenomes, those datasets include only the control region (CR). For this reason, our Cameroonian haplogroups were clustered to the closest common haplogroup level with the rest of the datasets.

2. Results and discussion

The hierarchical haplogroup frequencies of 100 Cameroonian mitogenomes are presented in Table 2, according to Phylotree Build 17 (www.phylotree.org; [12]). This sample contained 95 different mitogenomes and 87 different CR sequences of which 91 and 76 were singletons, respectively (Supplementary Table 1). For the mitogenome and the CR, the random match probability (RMP) was estimated at 1.12% and 1.30%, respectively. The power of discrimination was 98.88% and 98.70%, and the mean number of pairwise differences (MNPD) was 27.87 ± 21.16 and 6.59 ± 4.85, respectively. These results confirm

Table 1

Forensic parameters in mitogenomes from Cameroon (disregarding length variants at positions 309, 16093 and point heteroplasmies).

	Mitogenome 1–16569	Control region 16024–576
No. of samples	100	100
No. of haplotypes	95	87
No. of unique haplotypes	91	76
No. of haplogroups	62	54
Power of discrimination (%)	98.88	98.70
Haplotype diversity	0.9988	0.9970
Mean pairwise differences	27.87 ± 21.16	6.59 ± 4.85

marginal gains in discriminatory capacity of mitogenomes when compared to the standard CR for this population (Table 1).

The most common mitogenome, with three observations (3%;

Table 2

Haplogroup frequencies of 100 mitogenomes from Cameroon according to Phylotree build 17 [12].

HV0 (T195C)	1
L0a1a2	1
L0a1b1a	1
L0a1b2	1
L0a1e	3
L0a2	1
L0a2a1	3
L0a2a1b	1
L0a2a2a	2
L1b1a	3
L1b1a15	1
L1b1a3b	1
L1b1a6	2
L1c1'2'4'5'6	1
L1c1d1	2
L1c2a2	1
L1c2a3a	1
L1c2b1a	1
L1c2b1a'b	1
L1c2b1b	3
L1c3a1b	1
L1c3b1a	2
L1c3b1b	2
L2a1(G143A T16189C (C16192T))	2
L2a1a1	2
L2a1a2	1
L2a1a2a1a	1
L2a1b	2
L2a1c5	1
L2a1d1	1
L2a1f	1
L2a1q	1
L2a2a1	2
L2c2b1b	1
L2d1	1
L2d1a	2
L3b1a	2
L3b1a1	1
L3b1a1a	2
L3b1a7a	1
L3b2a	1
L3d1a1a	1
L3d1b	1
L3d1b3	1
L3d1b3a	1
L3d1d	2
L3e1	2
L3e1b2	3
L3e1e	1
L3e2a1b1	1
L3e2a1b2	4
L3e2b	8
L3e2b1a2	1
L3e3b1	1
L3e3b2	2
L3e5c	1
L3f1b1a1	2
L3f1b4a	1
L3f2b	1
L3f3	2
L3h1b1a	2
L4b2a2	1
	100

093, CAM-094, CAM-098, respectively in Supplementary Table 1).

Setting the heteroplasmy threshold at 20% of total coverage, twelve heteroplasmic positions were found (6581 R, 7163Y, 7347 R, 8746Y, 9890 R, 10410Y, 10685 R, 10846Y, 12366 R, 16093Y, 16172Y, 16189Y) in a total of eleven samples (one sample shows two heteroplasmic mutations). The HVI positions were previously described as hotspots [24].

The 95 haplotypes from Cameroon were assigned to 62 distinct haplogroups (Table 2). The majority (99%) belonged to the main common haplogroups in Africa (L1-L4). The sample composition was as follows: 45% (population 95% CI 35.600–54.778%) were L3 haplotypes (in descending frequency L3e, L3b, L3d, L3f, L3h), 22% (population 95% CI 15.017–31.101%) were L1 haplotypes (mainly L1c, L1b), 18% (population 95% CI 11.723–26.696%) were L2 haplotypes (predominantly L2a; also L2d, L2c), 13% (population 95% CI 7.790 21.004%) were L0 haplotypes (all L0a), and two singletons haplotypes belonged to L4b and HV0(T195C) with 1% frequency each. The most frequent of the 62 haplogroups were L3e2b (8%) and L3e2a1b2 (4%).

The presence of one HV0 haplotype stands out in our study, as this haplogroup is typically found in West Eurasia, particularly Northwest Europe, and it has been inferred to have originated 20 kya in Anatolia (present day Turkey) [25]. However, this haplogroup has also been reported in Tcheboua Fulani (North of Cameroon) and Kanuri (Nigeria) individuals by [26], consistent with gene flow from western Eurasia to Africa. The HV0 individual reported in this study is also a Kanuri and this could indicate a possible recent migratory event from the Near East to Cameroon.

To visualize haplogroup sharing between our Cameroonian sample and a set of other African population samples, MDS analysis was performed (Supplementary Figure 1). Pairwise F_{ST} distances (Supplementary Table 2) show the highest distance values with Egypt. Egypt is located at the nexus of Africa and Eurasia, and has a haplogroup composition reflecting this, i.e., it shares many haplogroups with other North African populations as well those from southwest Asia [22]. Reflecting this, we can observe how the West African populations (Ghana, West Africa and Ivory Coast) map closely together on one end of the X axis, whereas Egypt maps on the opposite end. Our Cameroon sample tends to cluster with this West African set of samples along the X axis, although it tends to occupy and intermediate position between the West and East (Kenya, Rwanda) samples along the Y axis.

3. Conclusion

This is the first study presenting the complete mitogenomes of an appreciably-sized Cameroonian sample. The Cameroonian sample mtDNA haplogroup distributions shows similarities to other West African populations. In addition, this study contributes to palliating the current underrepresentation of African populations in mtDNA databases, particularly concerning whole mitogenomes.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2021.102597.

population 95% CI 1.089–8.436%), belonged to haplogroup L1b1a (PhyloTree, build 17; [13]) (see samples CAM-002, CAM-005 and CAM-017 in Supplementary Table 1), relative to the rCRS [23], followed by haplotypes belonging to the haplogroups L1b1a6, L3e1b2, as well as the haplogroup L3f1b1a1, with two observations each (2%; population 95% CI 0.617–6.971%).

The two most common CR mitotypes with three observations each (3%; population 95% CI 1.089–8.436%), belonged to the haplogroups L1b and L3e2b (see samples CAM-002, CAM-005, CAM-017 and CAM-

References

[1] T.E. King, G.G. Fortes, P. Balaresque, M.G. Thomas, D. Balding, P. Maisano Delser, R. Neumann, W. Parson, M. Knapp, S. Walsh, L. Tonasso, J. Holt, M. Kayser, J. Appleby, P. Forster, D. Ekserdjian, M. Hofreiter, K. Schurer, Identification of the remains of King Richard III, Nat. Commun. 5 (2014) 5631.

[2] A. Amorim, T. Fernandes, N. Taveira, Mitochondrial DNA in human identification: a review, PeerJ 7 (2019) 7314.

[3] W. Parson, A. Dür, EMPOP-a forensic mtDNA database, Forensic Sci. Int. Genet. 1 (2007) 88–92.

[4] M. Cerezo, L. Gusmão, V. Černý, N. Uddin, D. Syndercombe-Court, A. Gómez-Carballa, T. Göbel, P.M. Schneider, A. Salas, Comprehensive analysis of pan-african mitochondrial DNA variation provides new insights into continental variation and demography, J. Genet. Genom. 43 (2016) 133–143.

[5] J.P. Kouega, The language situation in Cameroon, Curr. Issues Lang. Plan. 8 (2007) 3–93.

[6] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303.

[7] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 genome project data processing subgroup, the sequence alignment/map (SAM) format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[8] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842.

[9] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv:1303.3997v2 (2013).

[10] H. Thorvaldsdóttir, J.T. Robinson, J.P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, Brief. Bioinform. 14 (2013) 178–192.

[11] H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, G. Spetch, H.J. Bandelt, F. Kronenberg, A. Salas, S. Schönherr, HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing, Nucleic Acids Res. 44 (2016) W58–W63.

[12] M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation, Hum. Mutat. 30 (2009) e386–e394.

[13] N. Huber, W. Parson, A. Dür, Next generation database search algorithm for forensic mitogenome analyses, Forensic Sci. Int. Genet. 37 (2018) 204–214.

[14] R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL ⟨https://www.R-project.org/⟩.

[15] W. Parson, L. Gusmão, D.L. Hares, J.A. Irwin, W.R. Mayr, N. Morling, E. Pokorak, M. Prinz, A. Salas, P.M. Schneider, T.J. Parsons, 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.

[16] 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.

[17] S. Kitada, T. Kitakado, H. Kishino, The empirical Bayes inference of pairwise FST and its distribution in the genome, Genetics 177 (2007) 861–873.

[18] T.M.K. Göbel, M. Bodner, C. Robino, C. Augustin, G.E. Huber, M. Marra, L. Mutesa, S. Pasino, A. Santovito, B. Zimmermann, P.M. Schneider, W. Parson, Mitochondrial DNA variation in Sub-Saharan Africa: forensic data from a mixed West African sample, Côte d'Ivoire (Ivory Coast), and Rwanda, Forensic Sci. Int. Genet. 44 (2020), 102202.

[19] L. Fendt, A. Röck, B. Zimmermann, M. Bodner, T. Thye, F. Tschentscher, E. Owusu-Dabo, T.M. Göbel, P.M. Schneider, W. Parson, MtDNA diversity of Ghana: a forensic and phylogeographic view, Forensic Sci. Int. Genet. 6 (2012) 244–249.

[20] A. Brandstätter, C.T. Peterson, J.A. Irwin, S. Mpoke, D.K. Koech, W. Parson, T.J. Parsons, Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database, Int. J. Leg. Med. 118 (2004) 294–306.

[21] M. Mikkelsen, L. Fendt, A.W. Röck, B. Zimmermann, E. Rockenbauer, A.J. Hansen, W. Parson, N. Morling, Forensic and phylogeographic characterisation of mtDNA lineages from Somalia, Int. J. Leg. Med. 126 (2012) 573–579.

[22] J.L. Saunier, J.A. Irwin, K.M. Strouss, H. Ragab, K.A. Sturk, T.J. Parsons, Mitochondrial control region sequences from an Egyptian population sample, Forensic Sci. Int. Genet. 3 (2009) e97–e103.

[23] 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 (1999) 147.

[24] J.A. Irwin, J.L. Saunier, H. Niederstätter, K.M. Strouss, K.A. Sturk, T.M. Diegoli, A. Brandstätter, W. Parson, T.J. Parsons, Investigation of heteroplasmy in the human mitochondrial DNA Control region: a synthesis of observations from more than 5000 global population samples, J. Mol. Evol. 68 (2009) 516–527.

[25] D.M. Behar, M. van Oven, S. Rosset, M. Metspalu, E.L. Loogväli, N.M. Silva, T. Kivisild, A. Torroni, R. Villems, A "Copernican" reassessment of the human mitochondrial DNA tree from its root, Am. J. Hum. Genet. 90 (2012) 675–684.

[26] M. Cerezo, V. Černý, A. Carracedo, A. Salas, New Insights into the lake chad basin population structure revealed by high-throughput genotyping of mitochondrial DNA coding SNPs, PLoS One 6 (2011) 18682.

Sonia Olaechea-Lázaro[a,1], Óscar García[b,1], Rafaela González-Montelongo[c,d], José Miguel Lorenzo-Salazar[c], Carlos Flores[c,d,e,f], Saioa López[g], Krishna R. Veeramah[h], Garrett Hellenthal[i,j], Mark G. Thomas[i,j], Santos Alonso[a,*]

[a] *Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country (UPV/EHU), Barrio Sarriena s/n, 48940 Leioa, Spain*

[b] *Forensic Science Unit, Forensic Genetics Section, Basque Country Police, Erandio, Bizkaia, Spain*

[c] *Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain*

[d] *Instituto de Tecnologías Biomédicas (ITB), Universidad de La Laguna, Santa Cruz de Tenerife, Spain*

[e] *Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain*

[f] *CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain*

[g] *Department of Strategy and Data Insights, Wellcome Trust, 215 Euston Road, Bloomsbury, London NW1 2BE, United Kingdom*

[h] *Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, United States*

[i] *UCL Genetics Institute, University College London (UCL), Gower Street, London WC1E 6BT, United Kingdom*

[j] *Research Department of Genetics, Evolution & Environment, University College London (UCL), Darwin Building, Gower Street, London WC1E 6BT, United Kingdom*

[*] Corresponding author.
E-mail address: santos.alonso@ehu.eus (S. Alonso).

---

[1] Equal contribution