Available online at www.sciencedirect.com

ScienceDirect

Journal of Genetics and Genomics 43 (2016) 133−143

JGG

ORIGINAL RESEARCH

Comprehensive Analysis of Pan-African Mitochondrial DNA Variation Provides New Insights into Continental Variation and Demography

María Cerezo [a,b,1], Leonor Gusmão [c,d], Viktor Černý [e], Nabeel Uddin [f], Denise Syndercombe-Court [g], Alberto Gómez-Carballa [a], Tanja Göbel [h], Peter M. Schneider [h], Antonio Salas [a,*,1]

[a] *Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, Galicia 15782, Spain*

[b] *The Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK*

[c] *DNA Diagnostic Laboratory, Institute of Biology, State University of Rio de Janeiro, Rio de Janeiro 20550-900, Brazil*

[d] *IPATIMUP Institute of Molecular Pathology and Immunology of the University of Porto, Porto 4200-465, Portugal*

[e] *Archaeogenetics Laboratory, Institute of Archaeology of the Academy of Sciences of the Czech Republic, Prague 118-01, Czech Republic*

[f] *Faculty of Life Sciences and Medicine, King's College London, London SE1 9NH, UK*

[g] *Barts and The London School of Medicine and Dentistry, Blizard Institute, London E1 2AT, UK*

[h] *Institute of Legal Medicine, Medical Faculty, University of Cologne, Cologne D-50823, Germany*

Received 26 May 2015; revised 30 July 2015; accepted 15 September 2015
Available online 25 September 2015

ABSTRACT

Africa is the cradle of all human beings, and although it has been the focus of a number of genetic studies, there are many questions that remain unresolved. We have performed one of the largest and most comprehensive meta-analyses of mitochondrial DNA (mtDNA) lineages carried out in the African continent to date. We generated high-throughput mtDNA single nucleotide polymorphism (SNP) data (230 SNPs) from 2024 Africans, where more than 500 of them were additionally genotyped for the control region. These data were analyzed together with over 12,700 control region profiles collected from the literature, representing more than 300 population samples from Africa. Insights into the African homeland of humans are discussed. Phylogeographic patterns for the African continent are shown at a high phylogeographic resolution as well as at the population and regional levels. The deepest branch of the mtDNA tree, haplogroup L0, shows the highest sub-haplogroup diversity in Southeast and East Africa, suggesting this region as the homeland for modern humans. Several demographic estimates point to the coast as a facilitator of human migration in Africa, but the data indicate complex patterns, perhaps mirroring the effect of recent continental-scaled demographic events in re-shaping African mtDNA variability.

KEYWORDS: mtDNA; Haplotype; Haplogroup; SNP; MALDI-TOF

INTRODUCTION

Africa has become the focus of population genetic studies for more than two decades, and a large number of these studies are based on the analysis of mitochondrial DNA (mtDNA) variation (Cann et al., 1987; Vigilant, 1990; Chen et al., 1995; Watson et al., 1997). The studies by Pereira et al. (2001) and Salas et al. (2002) were the first attempts aimed at reconstructing the African mtDNA variation globally. These authors performed a phylogeographic analysis of mtDNA control region profiles of samples representing the main continental regions, and the results allowed the first general genetic picture of the Bantu expansion, one of the

* Corresponding author. Tel: +34 64 734 4311, fax: +34 88 181 2459.
E-mail address: antonio.salas@usc.es (A. Salas).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jgg.2015.09.005

most important demographic events that occurred within the African continent, to be outlined. Since then, several studies have been published aimed at reconstructing the southwestern route of the Bantu expansion (Plaza et al., 2004; Beleza et al., 2005; de Filippo et al., 2012), the demographic history of different African regions (Kivisild et al., 2004; Černý et al., 2007; Gonder et al., 2007; Podgorná et al., 2013), and histories of gene flow between Bantu farmers and hunter-gatherer pygmies (Batini et al., 2007; Quintana-Murci et al., 2008; Verdu et al., 2013) or between the Fulani herders and their farmer neighbors (Černý et al., 2006, 2011). Some of these studies have also focused on the reconstruction of the Trans-Atlantic slave trade (TAST) (Pereira et al., 2001; Salas et al., 2004a, 2004b, 2005b; Brucato et al., 2010; Veeramah et al., 2010). The studies by Behar et al. (2008) and Barbieri et al. (2013) represent the major mtDNA genotyping effort aimed at unraveling the phylogenetic history of Africa with main focus on the Khoisan people of South Africa. The results of these studies indicated that Khoisan ancestors diverged from the rest of the human mtDNA pool 90,000–150,000 ka (kilo years ago) and that the early settlement of humans in the continent involved small isolated populations, which are broadly consistent with autosomal resequencing datasets (Veeramah et al., 2012).

Most of the previous studies carried out on African populations were based on control region data and/or a few selected Restriction Fragment Length Polymorphisms (RFLPs) or mtSNPs (Beleza et al., 2005; Černý et al., 2006, 2007; Pereira et al., 2010). Only a few of the studies, which focused on selected African haplogroups (Černý et al., 2006; Behar et al., 2008; Pereira et al., 2010; Soares et al., 2012) or on particular regions or countries (Kivisild et al., 2004; Gonder et al., 2007; Kujanová et al., 2009), targeted complete mtDNA sequences. The present study aims to investigate African variability based on two approaches: (1) use of the large amount of data that has been accumulated in the literature during the past decade, which allow the inference of demographic and diversity parameters that was not possible a few years ago (Salas et al., 2002); (2) use of mtSNP data generated in the present study to a high haplogroup resolution aiming to improve our knowledge on phylogeographic patterns across the African continent.

RESULTS

Genetic diversity

North Africa shows the lowest values for molecular diversity (with the exception of South Africa for the haplotype diversity in control region sequences), mirroring the Mediterranean nature of the region (Table 1) compared to sub-Saharan Africa. Thus, only 17% of the lineages belong to typical L sub-Saharan haplogroups (Fig. S1). Diversity values observed for mtSNPs and HVS-I sequences are consistent for North Africa.

Within sub-Saharan Africa, the Southeast shows very low values of haplotype diversity (in particular for mtSNPs). This pattern is compatible with Southeast Africa being a *cul-de-sac* of the Bantu expansion (Salas et al., 2002). Nucleotide diversity, which correlates with the average number of nucleotide differences, does not show the same signal, which

Table 1
Diversity indices in the main African regions

mtDNA marker	Continental region	*n*	*K*	*S*	*H*	π	*M*
mtSNP	North	107	52	79	0.938 (0.017)	0.032 (0.003)	7.5
	West-Central	745	224	168	0.984 (0.001)	0.049 (0.001)	11.3
	Southwest	106	68	106	0.986 (0.004)	0.055 (0.002)	12.7
	Southeast	835	148	142	0.935 (0.004)	0.055 (0.001)	12.6
	East	231	111	130	0.982 (0.003)	0.050 (0.001)	11.7
	Total	2024	471	205	0.982 (0.001)	0.054 (0.000)	12.6
HVS-I	Northeast	600	250	102	0.963 (0.005)	0.020 (0.001)	5.2
	Northwest	2619	764	156	0.970 (0.002)	0.018 (0.000)	4.9
	North	3219	876	151	0.966 (0.002)	0.018 (0.000)	4.7
	Senegambia	1394	397	116	0.983 (0.001)	0.020 (0.000)	5.4
	Bight of Biafra	4063	950	155	0.989 (0.001)	0.033 (0.000)	8.8
	Gold Coast	713	258	100	0.984 (0.002)	0.025 (0.000)	6.9
	West-Central	6170	1093	157	0.984 (0.001)	0.025 (0.000)	6.3
	Southwest	157	96	77	0.991 (0.002)	0.032 (0.001)	8.9
	Southeast	587	190	95	0.971 (0.003)	0.030 (0.000)	8.2
	East	627	343	123	0.995 (0.001)	0.030 (0.000)	8.1
	South	303	12	34	0.898 (0.027)	0.031 (0.002)	8.7
	Total	10793	1706	170	0.984 (0.000)	0.021 (0.000)	5.1

The data are based on mtSNPs (genotyped in the present study) and the HVS-I (compiled from the literature; sequence range, 16090–16365). *n*, sample size; *K*, number of different sequences; *S*, number of segregating sites; *H*, haplotype diversity; π, nucleotide diversity; *M*, average number of pairwise differences (mismatch observed mean). The numbers in brackets are the standard deviations.