Forensic Science International: Genetics 7 (2013) 601–609

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA)☆

Alexander W. Röck [a], Arne Dür [b], Mannis van Oven [c], Walther Parson [a,d,*]

[a] Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria
[b] Institute of Mathematics, University of Innsbruck, Innsbruck, Austria
[c] Department of Forensic Molecular Biology, Erasmus MC, University Medical Center Rotterdam, The Netherlands
[d] Penn State Eberly College of Science, University Park, PA, USA

ARTICLE INFO

Article history:
Received 27 February 2013
Received in revised form 1 July 2013
Accepted 8 July 2013

Keywords:
mtDNA
Haplogroup
EMPOP
Fluctuation rates
Phylotree

ABSTRACT

The assignment of haplogroups to mitochondrial DNA haplotypes contributes substantial value for quality control, not only in forensic genetics but also in population and medical genetics. The availability of Phylotree, a widely accepted phylogenetic tree of human mitochondrial DNA lineages, led to the development of several (semi-)automated software solutions for haplogrouping. However, currently existing haplogrouping tools only make use of haplogroup-defining mutations, whereas private mutations (beyond the haplogroup level) can be additionally informative allowing for enhanced haplogroup assignment. This is especially relevant in the case of (partial) control region sequences, which are mainly used in forensics. The present study makes three major contributions toward a more reliable, semi-automated estimation of mitochondrial haplogroups. First, a quality-controlled database consisting of 14,990 full mtGenomes downloaded from GenBank was compiled. Together with Phylotree, these mtGenomes serve as a reference database for haplogroup estimates. Second, the concept of fluctuation rates, i.e. a maximum likelihood estimation of the stability of mutations based on 19,171 full control region haplotypes for which raw lane data is available, is presented. Finally, an algorithm for estimating the haplogroup of an mtDNA sequence based on the combined database of full mtGenomes and Phylotree, which also incorporates the empirically determined fluctuation rates, is brought forward. On the basis of examples from the literature and EMPOP, the algorithm is not only validated, but both the strength of this approach and its utility for quality control of mitochondrial haplotypes is also demonstrated.

1. Introduction

Human mitochondrial (mt)DNA is passed from mother to offspring and therefore inherited along a phylogeny. The first human mitochondrial genome (mtGenome) was sequenced in the early 1980s [1] and revised 18 years later [2], serving as a reference sequence (rCRS) relative to which other mtDNA sequences have been reported in a difference-coded format. A plethora of partial as well as complete mtGenomes has been produced since, permitting an increased understanding of the evolution of this molecule. Its dispersal through human migration left characteristic footprints induced by mutations that have been used to assign sequences to haplogroups [3]. The growing collection of established clades has meanwhile reached 3925 discernible haplogroups based on 16,810 full mtGenomes [Phylotree (www.phylotree.org) Build 15 [4]].

The understanding of sequences from the standpoint of their haplogroup affiliation has become increasingly valuable in studies of human mtDNA. Not only does haplogroup nomenclature and assignment facilitate comparison and communication of genetic variability, but it is also employed to characterize mitochondrial lineages for population [5], medical [6] and forensic genetic [7] purposes. Most importantly for forensics, haplogroup assignment has proven to be an important tool for sequence data quality control [8]. Haplogrouping of mtDNA sequences has been greatly simplified with the provision of Phylotree, which is widely accepted as "mitochondrial haplogroup dictionary" in the scientific community. The haplogroup-defining mutations listed in Phylotree not only facilitate manual haplogroup assignment, but they also serve as the basis for a number of software applications that perform the task (e.g. MitoTool [9,10], HmtDB [11], HaploGrep [12], and mtDNAoffice [13]). To date, however, none of the available automated solutions provide reliable and unbiased

* Corresponding author at: Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria. Tel.: +43 512 9003 70640; fax: +43 512 9003 73640.
E-mail address: walther.parson@i-med.ac.at (W. Parson).

1872-4973/$ – see front matter © 2013 The Authors. Published by Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.07.005

haplogroup estimates, especially in the case of partial mtDNA sequences [7]. The major limitation of existing tools is that they base their haplogroup assignment solely on defined motifs of diagnostic mutations (virtual haplotypes). The remaining "private" mutations of a sequence which can be additionally informative for the haplogroup status are not considered. As a consequence, the haplogroup assignments are therefore often incorrect or too coarse.

Consider for example the following control region haplotype from Argentina, 16189C 16292T 16519C 71A 153G 204C 207A 263G 315.1C 373G, which harbors no characteristic mutation to match a haplogroup in Phylotree's motif list (Build 15) but nevertheless seems to fall within superhaplogroup R0. This sequence was assigned to haplogroup H by mtDNAmanager [14] and to haplogroup H1 + 16,189 by HaploGrep [12]. Additional coding region sequencing of this sample revealed haplogroup H55 status (4769G 10464A). This conclusion could also have been drawn from the control region haplotype alone, if its near match with the complete sequence JQ705203 known to belong to haplogroup H55 had been considered. In this study, we offer new software (EMMA) that bases haplogroup estimation on Phylotree's list of virtual haplotypes and a database of 14,990 quality-controlled full mtGenomes, and that employs a maximum likelihood approach. We demonstrate, by comparative analysis that our tool yields more precise haplogroup assignments than other available software. For the Argentinean control region haplotype, EMMA correctly assigned the proper haplogroup status as H55 even without coding region information.

2. Materials and methods

2.1. Virtual Phylotree haplotypes and full mtGenome database

The phylogenetic tree in Phylotree [4] represents known global mtDNA variation by defining haplogroups and their signature mutations. This tree is regularly updated incorporating newly available mtGenomes and is made available to the user in HTML format. All results presented in this study use Phylotree Build 15 (September 30, 2012) as the reference tree. For our purposes, an R [15] script was developed that transforms the tree into a list of hypothetical haplotypes carrying the signature mutations of the respective haplogroups (tree nodes) as differences to the rCRS [2]. As these haplotypes are inferred rather than observed in the real world they are herein referred to as virtual haplotypes.

Recently, a new reference sequence for mtDNA, the so-called Reconstructed Sapiens Reference Sequence (RSRS), has been proposed [16]. Instead of using a contemporary European mtGenome as reference sequence the authors suggest switching to a reconstructed ancestral sequence that is allocated between haplogroups L0 and L1′2′3′4′5′6′. A switch to the RSRS in the forensic field, however, is not expected to occur soon [17]. The software presented here is primarily designed for forensic purposes, thus input of mtDNA data is currently based on the rCRS.

The defined haplogroup motifs in Phylotree are based on a database of published mtGenomes (http://www.phylotree.org/mtDNA_seqs.htm) that were downloaded from GenBank and evaluated for their application within EMMA. Some sequences were incomplete (e.g. [18], accession number EF657231, lack of control region; [19], accession number EF661002, sequencing frame 1–4167 4434–5483 5785–8314 8566–10683 10749–16548) and therefore excluded from this study. MtGenomes highlighted as problematic in Appendix S1 of Ref. [20] have also been removed. Additionally, mtGenomes that have been generated in the course of second generation sequencing attempts in Refs. [21,22], and flawed data published by [[23], I.P. Maksum, unpublished, V.C. Phan, unpublished] have been excluded from the database. Of the remaining mtGenomes, those containing ten or more 'N' designations in the FASTA string were excluded because of insufficient sequence quality.

Subsequent analysis of our quality-controlled mtGenome database revealed that 20 haplogroups of Phylotree 15 were not represented by complete sequences anymore due to the strict policy applied above. To ensure maximum coverage of haplogroups, 29 mtGenomes rejected in the first step were reincluded in the database. With the expected future availability of more reliable mtGenomes those questionable haplotypes will be replaced. A single haplogroup was considered unreliable and therefore not reincluded: according to Phylotree haplogroup M25 is represented by two mtGenomes from [24] (accession numbers DQ246830, DQ246833). Due to the lack of the first 250 bases in these mtGenomes and the presence of several doubtful variants, both sequences were not considered for the database, thus haplogroup M25 is the only haplogroup that is not represented by any mtGenome in the database for EMMA. See ESM1 for the list of genomes added and the reason for initial exclusion.

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.07.005.

Finally, all FASTA strings were translated into rCRS-coded haplotypes using SAM [25] and subsequently checked with in-house software to harmonize alignment. The quality filtering of the sequences finally resulted in a database of 14,990 full mtGenomes stored with their accession numbers and version. In conjunction with the 3925 virtual Phylotree motifs, these 18,915 virtual and real mtGenomes form the basis for haplogroup estimation.

2.2. Fluctuation rates

Haplogrouping of mtDNA data in rCRS-based format requires consistent alignment and notation of sequences following a phylogenetic approach [26] in order to assess the stability of mutations in defined haplogroups. Here, we refer to this mutational (in)stability as a fluctuation rate. The weighting scheme presented for the string-search method in Ref. [25] was updated by assessing the stability of mutations within the mtDNA control region among 19,171 full control region haplotypes for which raw lane data were available. Haplogroups were manually assigned to all sequences in this dataset between November 2011 and September 2012 following the classification outlined in Phylotree Builds 12 through 15. Consequently, the sequences were grouped into discernable control region haplogroup clusters (CR-HGs), i.e. clusters of haplogroups that can be confidently determined based on control region motifs. We set a minimum of four available sequences to define a CR-HG with the exception of CR-HGs L0, L2, L6, U4′9, K3, and P9 for which only one or two sequences were available. In these cases, merging the sequences with their parent haplogroups L and R would have rendered the resolution too coarse. For a list of CR-HGs based on Phylotree Build 15 and the number of samples for each cluster see ESM2. Samples that were assigned to multiple haplogroups due to uncertainty were split equally into the respective CR-HGs.

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.07.005.

Assuming independent positions we estimated the fluctuation rate by

$$r_{\alpha\beta} = \frac{\sum_{\gamma} \min(n(\alpha,\gamma), n(\beta,\gamma))}{\sum_{\gamma} n(\gamma)}$$

where α, β are elements of the set A, C, G, T, – with α not equal to β, γ runs over all CR-HGs where α or β are dominant, $n(x,\gamma)$ denotes