Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics 1 (2007) 88–92
www.elsevier.com/locate/fsig

# EMPOP—A forensic mtDNA database

Walther Parson [a,*], Arne Dür [b]

[a] Institute of Legal Medicine, Innsbruck Medical University Müllerstreet 44, 6020 Innsbruck, Austria
[b] Institute of Mathematics, University of Innsbruck Technikerstreet 45, 6020 Innsbruck, Austria

Received 19 January 2007; accepted 27 January 2007

**Abstract**

Mitochondrial DNA databases stand as the basis for frequency estimations of mtDNA sequences that became relevant in a case. The establishment of mtDNA databases sounds trivial; however, it has been shown in the past that this undertaking is prone to error for several reasons, particularly human error. We have established a concept for mtDNA data generation, analysis, transfer and quality control that meets forensic standards. Due to the complexity of mtDNA population data tables it is often difficult if not impossible to detect errors, especially for the untrained eye. We developed software based on quasi-median network analysis that visualizes mtDNA data tables and thus signposts sequencing, interpretation and transcription errors. The mtDNA data ($N = 5173$; release 1) are stored and made publicly available via the Internet in the form of the EDNAP mtDNA Population Database, short EMPOP. This website also facilitates quasi-median network analysis and provides results that can be used to check the quality of mtDNA sequence data. EMPOP has been launched on 16 October 2006 and is since then available at http://www.empop.org.

*Keywords:* Mitochondrial DNA sequencing; mtDNA population databases; Network analysis; Quality control; Phantom mutations

## 1. EMPOP considerations

The need for a collaborative project to establish a new forensic mtDNA database was raised at the European DNA Profiling (EDNAP) Group meeting at the 18th ISFH (now ISFG) congress in San Francisco, 1999 (http://www.isfg.org). The following terms of references were defined: the new database should include high-quality data and be open and directly accessible for the scientific community. The new concept should meet the forensic requirements in terms of data documentation.

### 1.1. Finding sources of error

Blind tests are a valuable means for demonstrating proficiency and have therefore been widely used as external quality check for forensic laboratories. Regardless of its actual relevance to mtDNA population typing we performed a collaborative exercise to learn potential pitfalls associated with the laboratory process [1]. The results of this experiment confirmed the initial apprehension that mtDNA typing seems to be more prone to human error than other forensic DNA analysis (e.g. STR-typing). Our findings were in large parts confirmed by similar investigations, such as the mtDNA proficiency testing programme of the GEP-ISFG [2].

In parallel, error-reports from the scientific literature put mtDNA analysis at the centre of high-profile discussions [3–6]. By means of *a posteriori* data analysis the authors demonstrated that published mtDNA sequence data are prone to contain errors, mainly due to misinterpretation of sequence raw data (phantom mutations) and due to the introduction of clerical errors during data transcription. A more detailed view on the entire process of data generation revealed a compound picture of causes and phenotypes of error [7–9] that triggered the development of refined laboratory methods and safety steps for the establishment of high-quality mtDNA population data.

### 1.2. Generating high quality mtDNA data

It has more often been stated than actually followed that consensus mtDNA haplotypes were created by full double-strand sequence analysis. The mere application of both forward and reverse PCR primers for cycle sequencing does not suffice to produce full redundant double-strand sequences, as a reliable consensus sequence needs to be inferred from more than these

[*] Corresponding author. 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 © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2007.01.018

two reactions. This applies not only to samples that display length heteroplasmy, which hampers basecalling beyond the variant regions and thus provides partial sequence information only; it is crucial to decipher problematic positions that suffer from elevated background, sequencing artefacts or sub-optimal reaction and electrophoresis conditions in general [10]. New amplification and sequencing strategies that lead to forensically acceptable sequence-quality have been developed recently and are now increasingly applied to generate high-quality population data (e.g. [11–13]).

Another source of error that is repeatedly found in mtDNA data is the mix-up of hypervariable segments (HVS-I/HVS-II) between individuals especially when separate amplifications of the hypervariable regions are performed. This error that is also known as 'artificial recombination' cannot be detected by use of the raw data, but with the aid of phylogenetic analysis when the individual mutation patterns are compared between haplogroups. This however is only a limited tool for quality control as a number of haplogroups harbour HVS-II motifs that cannot be reliably discriminated between even distant haplogroups, such as the HVS-II sequence pattern 73G, 263G 315.1C that constitutes the basal motif in hgs H(1a), K, U and T in the West Eurasian population between positions 73 and 340. This motif is also found in some lineages of super-haplogroups B, D, E and G in East Asians and Native Americans. Therefore, strategies that use a single large amplification product are advantageous as artificial recombinants can almost completely be avoided.

### 1.3. Data transfer

The EMPOP collaborative exercise on mtDNA typing [1] revealed that 62% of the errors were clerical errors that arose during the manual transcription of mutations relative to the reference sequence. This value is surprisingly high given the reduced number and complexity of the experimental data set. In that respect it is beyond doubt that larger sample sizes harbour an increased risk of wrong transcriptions, which makes IT-based solutions for safe data transfer indispensable. All mtDNA data included in EMPOP is handled with a modified version of a self-developed in-house LIM system [14]. This software monitors the analysis and evaluation of the consensus sequences, archives the history of data generation and permanently links the profiles to their raw data for any later inspection. This kind of documentation is a valued forensic principle that is widely used in routine casework and intelligence databasing [15]. The evaluation process of local EMPOP data is carried out in two IT-aided analysis steps involving quasi-median network analysis (online) and phylogenetic analysis (local development, presented elsewhere) of the haplotypes.

### 1.4. Quasi-median network analysis

MtDNA data tables can be visualized as quasi-median networks which represent a helpful tool to enhance our understanding of the data in regard to homoplasy and potential artefacts. Network analysis has proven a probative means to detect data idiosyncrasies that pinpoint sequencing and data interpretation problems [16]. Each mtDNA data set should undergo routine *a posteriori* data analysis regardless of the sequence strategy or quality. This evaluation is facilitated by the software package NETWORK that is made available via the EMPOP website (http://www.empop.org). NETWORK accepts mtDNA control region data compiled as motif lists in so-called 'emp' format. An example file describing a population data set of 273 Austrian control region sequences [13] is accessible for download at the website.

An important feature of the network analysis is the filtering option, which highlights mutations that should be reviewed by inspection of the raw lane data. Currently NETWORK employs three different filters that can be selected depending on the application:

- *EMPOPspeedy*. This filter removes highly recurrent mutations based on the lists provided in Refs. [3,17]. In addition we added more mutations to this filter that were homoplasic in Release 1 of EMPOP ($N = 3830$ west Eurasians). The individual filtered mutations can be viewed in the NETWORK section of EMPOP (http://www.empop.org). This filter is typically used for the analysis of mtDNA data including the hypervariable segments—HVS-I (16024–16569) and HVS-II (1–576) of medium sized datasets ($N = 50$–300).
- *EMPOPall*. EMPOPall disregards all mutations observed in the database (currently from Release 1; $N = 5173$). This filter produces networks that highlight only unobserved mutations and thus provides a quick and effective check on new data. It can be applied to large datasets (>300 haplotypes) encompassing the above mentioned hypervariable segments.
- *Unfiltered*. None of the mutations are removed from the tested dataset. This blank filter is used for network analyses that should display the full variation in a given dataset. This filter can only be meaningfully applied to short sequence segments of the control region. The complexity of the network could increase rapidly if no filter is applied to the analysis of larger sequence regions.

Table 1
Excerpt of the report.txt file displaying general analysis settings and a tabular summary of the network analysis

| Filter analysis | | | | | | |
|---|---|---|---|---|---|---|
| $n$ | $p$ | $p'$ | $h$ | $q$ | $t$ | $t'$ |
| 273 | 38 | 37 | 40 | 43 | 15 | 1 |

EMPOP Network Analysis Report Tuesday 21. 11. 2006, 13:01:14 UTC. Input data set: AUT273 spec.emp (273 samples), filter: EMPOPspeedy (Version 1: Region: 16024–16569). $n$: number of samples; $p$: number of polymorphic positions; $p'$: number of partitions (condensed characters); $h$: number of haplotypes; $q$: number of nodes of the network; $t$: number of nodes of the torso (network without periphery); $t'$: number of nodes of the peeled torso (network without extrusions).