

EMPOP—the EDNAP mtDNA population database concept for a new generation, high-quality mtDNA database

Walther Parson*, Anita Brandstätter, Martin Pircher,
Martin Steinlechner, Richard Scheithauer

Institute of Legal Medicine, University of Innsbruck, Muellerstrasse 44, 6020 Innsbruck, Austria

Abstract. The European DNA Profiling Group (EDNAP) MtDNA Population Database (EMPOP) is an international collaborative project between DNA laboratories performing mtDNA analysis and the DNA laboratory of the Institute of Legal Medicine (GMI) in Innsbruck, Austria. The goal is to set up a directly accessible mtDNA population database, which can be used in routine forensic casework for frequency investigations. Most forensic laboratories do not have the capacity to rapidly generate massive amounts of sequence data. Altogether, however, they dispose of a respectable body of data. Apart from a mutual, non-competitive exchange of molecular genetic techniques, every forensic institute can contribute to the EMPOP project with its familiarity with the individual technology and with the polymorphisms typical for the population inhabiting its surroundings. The EMPOP core laboratory in Innsbruck provides the bioinformatic infrastructure with which to analyze, transcribe, store, and share the data. © 2003 Elsevier B.V. All rights reserved.

Keywords: Forensic mtDNA profiling; Control region sequences; Frequency estimations; Internet database

1. Introduction

The analysis of the control region (CR) of the human mitochondrial genome (mtDNA), in particular, the hypervariable segments of the CR, HV1 and HV2, has proven to be a useful tool for forensic identifications and human evolution studies [1,2]. The lack of mitochondrial recombination urges forensic DNA analysts to compare the HV1/HV2 profile in question to a database of mitochondrial CR profiles for biostatistical frequency estimations. At this very moment, mtDNA databases in general, and forensic mtDNA databases in particular, are at the centre of a scientific controversy relating to the detection of large numbers of errors in published databases. Using phylogenetic investigations and

Abbreviations: CR, Control region; mtDNA, mitochondrial DNA.

* Corresponding author. Tel.: +43-512-507-3303; fax: +43-512-507-2764.

E-mail address: Walther.Parson@uibk.ac.at (W. Parson).

referring to known patterns of mutations in mtDNA, rampant errors have recently been reported [3–6]. The combination of biochemical, technical, and informatical difficulties made mtDNA databases extraordinarily prone to error. The actual “crisis of quality” in published forensic mtDNA databases requires forensics to achieve a new, higher level of database quality control and supervision.

2. Structure

The European DNA Profiling Group (EDNAP) MtDNA Population Database (EMPOP) is a Web-based SQL database designed for the management of mtDNA sequence data from various populations worldwide, and in general for the assessment of the rarity of a forensic mtDNA haplotype in a given population. In order to address the high error rate connected with mtDNA typing and databasing, logistical features were installed after assessing potential sources of error by performing collaborative exercises (CEs, see below). As a consequence of the CE results, full electronic transfer of the polymorphisms from the raw data (electropherograms) onto the database without further need of manual interaction (transcription) was identified as the most important feature, as clerical errors proved to be the dominant source of errors. Furthermore, sequence raw data of the haplotypes are directly linked to the entries in the database, which enables immediate control of the sequences if necessary. Before loading population data onto the database, these will undergo phylogenetic analysis in order to test for errors caused by mixing up of samples during the process of assessing a HV1/HV2 haplotype, a phenomenon that was observed in the course of the CE. The program of EMPOP allows for searching any fragment in the CR, independent of length and position in the human CR. Example: If only a 150-bp fragment of HV1 brought successful sequencing results (e.g. from a hair shaft), this particular fragment of the CR is searched in the database. The result, i.e., the number of sequences matching the fragment, will then be given according to the total number of sequences that include this 150-bp fragment. Thus, diverse amplification strategies and different primer pairs developed in the laboratories can be used for searching the result in EMPOP.

3. Collaborative exercises

CEs, which have been proven to be a valuable tool for inter-laboratory quality assurance [7], were conducted to ensure analytical reliability and revealed a high correspondence rate of the analytical results obtained by the individual member institutes. A detailed description of the organization and the results of the collaborative exercises will be published soon (Parson et al. submitted). In the EMPOP CE, the mtDNA sequencing results were evaluated by means of inspecting both the raw data and the tabular results submitted by each participating laboratory. Consequently, two different types of results were obtained from the study. First, a measure of the quality of the sequence data reflecting the entire laboratory process involving DNA extraction, PCR setup, and sequence analysis was obtained. Second, the quality of data interpretation and transcription, the latter being a crucial step when reporting mtDNA evidence, was assessed. Since the major portions of all discrepancies observed were clerical errors (the raw data were in agreement with the expected results), the data would meet the requirements in terms of the

raw data but at the same time, the data would be considered erroneous with respect to the tabular result. Consequently, such data would display the errors when regarded as results listed in a report. In the context of the laboratory process, however, the samples were typed correctly. Considering the EMPOP CE as means of monitoring the quality of the mtDNA typing process for database purposes (the raw data would be transferred electronically omitting the occurrence of clerical errors), the results demonstrated that the laboratories had accomplished all necessary requirements in order to arrive at the correct result (furthermore, the individual laboratories immediately discovered the clerical errors when being contacted). Following this, a laboratory demonstrating good laboratory practice would still meet the necessary requirements for submitting population to EMPOP even when a clerical error was introduced in the tabular report.

Clerical errors were the major source of error in our collaborative exercises (62.5% of all errors). The errors were introduced in the course of transcription of the variant positions that differed from the CRS sequence from the raw data to the result tables. In two instances, mix-up of the sequenced regions of two samples was observed: This phenomenon, which can be described as a common source of error [3,6], caused 12.5% of all observed errors. Whereas this kind of error can easily be identified in a collaborative exercise, it is difficult to reveal this error in a set of population data. It has been described that phylogenetic methods as prefigured by Bandelt et al. [3] could help to identify erroneous haplotypes, although the methods may be more focussed on the identification of a wrong combination of mtDNA regions from different individuals. In conclusion, clerical errors and sample mix-up errors made up 75% of all errors observed in the EMPOP CEs. This fact strongly emphasizes the need for appropriate safety regulations when mtDNA profiles are compiled for database purposes in order to accomplish the high standard required for forensic mtDNA databases that are used for forensic purposes.

References

- [1] W. Parson, T.J. Parsons, R. Scheithauer, M.M. Holland, Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: application of mtDNA sequence analysis to a forensic case, *Int. J. Legal Med.* 111 (3) (1998) 124–132.
- [2] E. Watson, P. Forster, M. Richards, H.J. Bandelt, Mitochondrial footprints of human expansions in Africa, *Am. J. Hum. Genet.* 61 (3) (1997) 691–704.
- [3] H.J. Bandelt, P. Lahermo, M. Richards, V. Macaulay, Detecting errors in mtDNA data by phylogenetic analysis, *Int. J. Legal Med.* 115 (2) (2001) 64–69.
- [4] H.J. Bandelt, L. Quintana-Murci, A. Salas, V. Macaulay, The fingerprint of phantom mutations in mitochondrial DNA data, *Am. J. Hum. Genet.* 71 (5) (2002) 1150–1160.
- [5] P. Forster, To err is human, *Ann. Hum. Genet.* 67 (Pt. 1) (2003) 2–4.
- [6] A. Rohl, B. Brinkmann, L. Forster, P. Forster, An annotated mtDNA database, *Int. J. Legal Med.* 115 (1) (2001) 29–39.
- [7] S. Rand, M. Schurenkamp, B. Brinkmann, The GEDNAP (German DNA profiling group) blind trial concept, *Int. J. Legal Med.* 116 (4) (2002) 199–206.