

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>.

© 2007 Elsevier Ireland Ltd. All rights reserved.

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).

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 homoplasmy 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 homoplastic 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						
<i>n</i>	<i>p</i>	<i>p'</i>	<i>h</i>	<i>q</i>	<i>t</i>	<i>t'</i>
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).

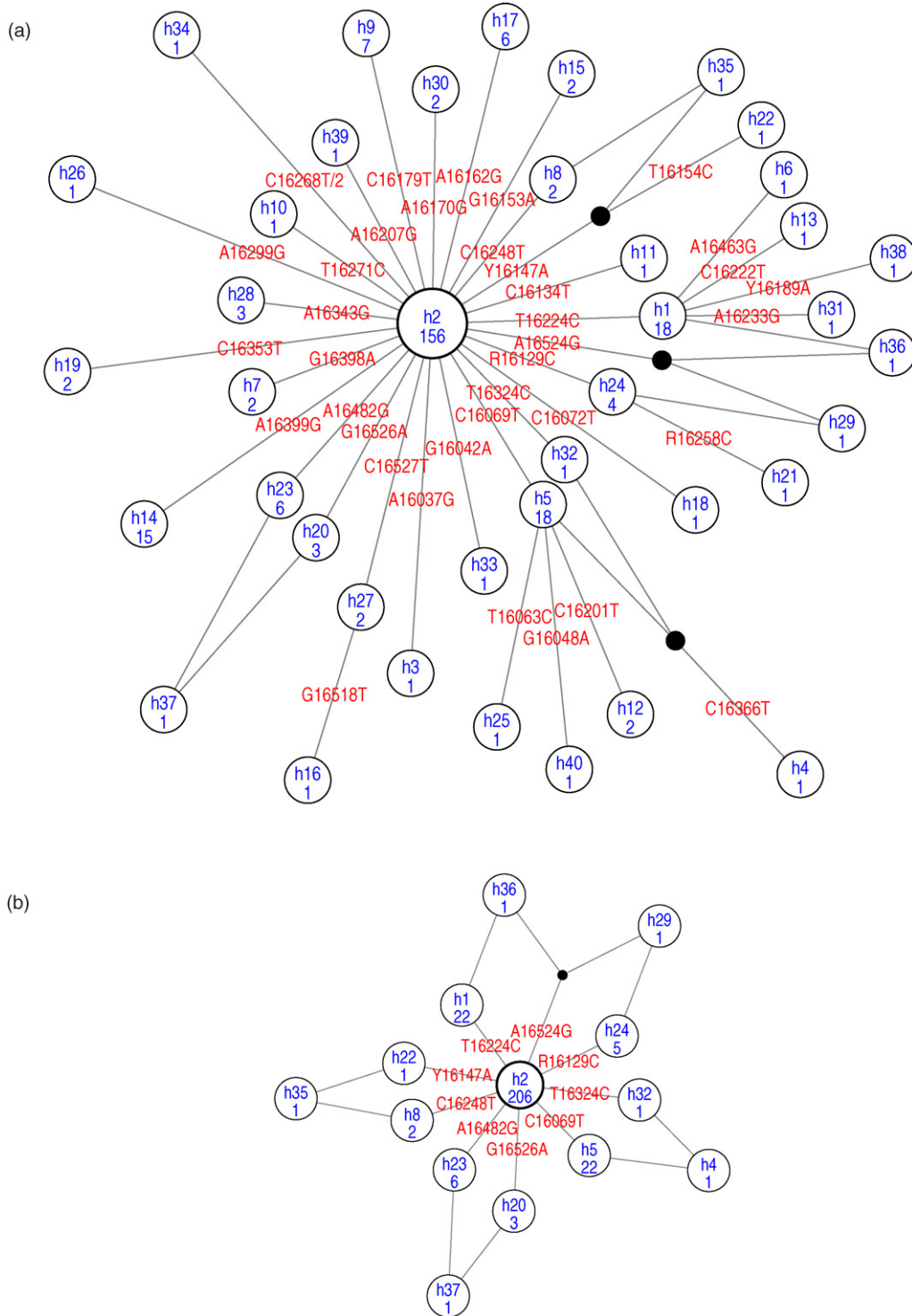


Fig. 1. The complete network (a) and the torso (b) of 273 mtDNA sequences (npt 16024–16569) from Austria [13] is presented (EMPOSpeedy filter). The software for drawing networks (dnw.exe) is made available at the EMPOP-Downloads-page. Circles represent filtered and reduced haplotypes, numbers indicate the frequency of identical haplotypes. The full dot represents a quasi-median, a logical mtDNA haplotype that has not been observed in the sample. Nucleotide positions on the links describe observed mutations, whereas preceding symbols characterise the observed bases in the data set and the following symbol describes the variant that is separated by the link. Combined mutations, i.e. multiple mutations that always occur jointly in the data set, are specified by the first (lowest) mutation and their number is indicated following the slash sign (/). This applies to ht34 in the network, in which the transition at position 16268 is coupled with the transition at position 16391 (details in report, not shown here). (a) Quasi-median network (273 Austrian HVS-I sequences, EMPOSpeedy) and (b) torso of the quasi-median network (273 Austrian HVS-I sequences, EMPOSpeedy).

1.5. Network results

The interpretation of the network results is demonstrated by the above mentioned example file, that represents 273 control region sequences from Austria, and that can be downloaded to evaluate the network function of the database.

A results file is created (report.txt) that reviews relevant information of the quasi-median network analysis (Table 1). The values which are given in the tabular summary are indicative for the quality of the resulting network. However, they depend on the size and composition of the population data set in question. Generally, small t' -values (ideally 1) describe a star-like structure of the network, which is in agreement with the expected evolutionary pattern. More suggestive however, is the graph of the quasi-median network (Fig. 1a and b), which can be created with the downloadable software dnw.exe. The drawing algorithm implemented in NETWORK constructs Eulerian tours in the non-strong-compatibility graph of the DNA data table [16] to lay out the network in the plane. The nodes of the network represent the (reduced and filtered) haplotypes (circles, indicating the haplotype frequency) and the quasi-medians (full dots) that were generated from the data. The root node is drawn with a bold circle and contains the filtered and reduced rCRS. In case that the filtered and reduced rCRS haplotype is not included in the dataset, the first haplotype is chosen instead and a respective warning indicates this fact in the report. The links represent single or combined mutations specified by the syntax <old base><position><new base> for single mutations or <old base><position><new base>/<number of mutations> for combined mutations. ‘Combined’ in this context refers to multiple mutations that always occur jointly in the data set. For display purposes not all but only the first (lowest in position in 16024–576) mutation is specified on the link and the total sum of (combined) mutations is indicated by its number after the slash (see Fig. 1). The other (combined) mutations are listed in the report. The reading orientation is chosen from the root node outwards. Ambiguous IUPAC designation refers to the filtering approach, e.g. Y16147A in Fig. 1b indicates that the transition C16147T is disregarded by the filter (here EMPOPspeedy), and that the variant 16147A was observed in the two haplotypes h22 and h35. Parallel links in the graph indicate identical mutations and are only labelled once.

The full quasi-median network (Fig. 1a) displays the observed variation in the data set, i.e. all mutations that are not removed by the filter are included in the graph or indicated by numbers for combined mutations, respectively. This view helps to pinpoint unusual mutations that should be checked for their plausibility. Here transversions are of particular interest as they are not expected at high frequency in the mtDNA control region.

The torso (Fig. 1b) is obtained from the quasi-median network by collapsing all pendant sub-trees into their base nodes. Thus the analysis of homoplasmy can be restricted to the torso which contains all the reticulation of the network. For each base node the coinciding haplotypes are listed in the report to easily trace back corresponding samples (data not shown

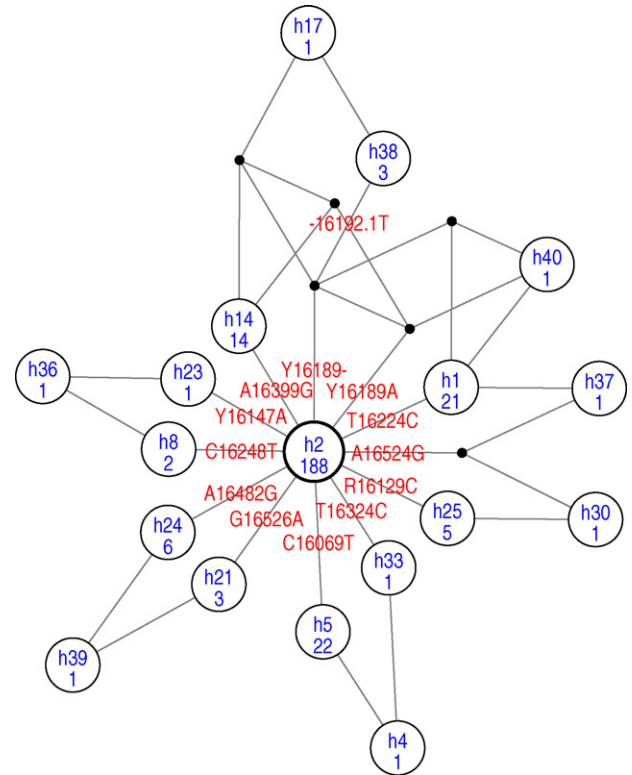


Fig. 2. Torso of the quasi-median network (273 Austrian HVS-I sequences, EMPOPspeedy), including the modification 16189del 16192.1T; original notation: 16189C 16192T.

here but available in the report). Reticulations can be attributed to homoplasmy mutations that have not been filtered from the data set; however, these can also constitute data idiosyncrasies and therefore need to be checked by controlling the raw data.

Fig. 2 shows the torso of a data set that has been modified for all four samples with the control region sequence motif 16189C 16192T. This has intentionally been changed to 16189del 16192.1T following earlier expressed guidelines for the notation of length variants [18] that used a binary (non-phylogenetic) alignment biased towards the reference sequence. The resulting twin-prism in the torso describes the inconsistency (by introduction of five additional quasi-medians) that arises in comparison to other haplotypes that either share the 16399G variant (h14) or a transversion at position 16189 to A (h40). This example demonstrates how NETWORK signposts “unusual” mutations, be they sequencing artefacts, transcriptional errors or notational differences. A reappraisal of the guidelines for sequence notation in length variants is presented in Ref. [19].

1.6. Samples included in Release 1

EMPOP (Release 1) holds 5173 mtDNA haplotypes which are distinguished at the most basal level of classification, i.e. sub-Saharan African, West Eurasian, East Asian and Southeast Asian metapopulations. The vast majority of haplotypes in Release 1 stem from west Eurasia, while the other metapopulations are underrepresented. Ongoing and future sampling and

sequence analysis will add more data which will be made available in subsequent releases. Further, the haplotypes stored in EMPOP are distinguished by their documentation with ‘forensic data’ ($N = 4527$) being linked to high-quality sequences, and ‘literature data’ ($N = 646$) where no appropriate raw lane sequence information is available. The latter are however checked with great scrutiny using the above-mentioned NETWORK tool and other methods of phylogenetic evaluation.

Acknowledgements

The mtDNA staff at the Institute of Legal Medicine, Innsbruck Medical University, Anita Brandstätter, Nina Duftner (currently University of Texas at Austin), Cordula Eichmann, Anna König, Roswitha Mühlmann (now Department for Internal Medicine), Daniela Niederwieser and Bettina Zimmermann, is acknowledged for excellent technical work in generating and analyzing thousands of mtDNA sequences. We thank Martin Pircher, Stefan Troger, Alexander Röck and Konrad Schwarz for software development and programming of the local and Internet databases. We would like to thank Hans-Jürgen Bandelt for discussion. All EMPOP collaborators are greatly acknowledged for their supportive collaboration; a list is provided at the contributors site of EMPOP (<http://www.empop.org>).

References

- [1] W. Parson, A. Brandstätter, A. Alonso, N. Brandt, B. Brinkmann, Á. Carracedo, D. Corach, O. Froment, I. Furaç, T. Grzybowski, K. Hedberg, C. Keyser-Tracqui, T. Kupiec, S. Lutz-Bonengel, B. Mevag, R. Ploski, H. Schmitter, P. Schneider, E. Syndercombe-Court, H. Sorensen, G. Thew, R. Tully, Scheithauer, The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation, results and perspectives, *Forensic Sci. Int.* 139 (2004) 215–226.
- [2] A. Salas, L. Prieto, M. Montesino, C. Albarrán, E. Arroyo, M.R. Paredes-Herrera, A.M. Di Lonardo, C. Doutremepuich, I. Fernandez-Fernandez, A.G. de la Vega, C. Alves, C.M. Lopez, M. Lopez-Soto, J.A. Lorente, A. Picornell, R.M. Espinheira, A. Hernandez, A.M. Palacio, M. Espinoza, J.J. Yunis, A. Perez-Lezaun, J.J. Pestano, J.C. Carril, D. Corach, M.C. Vide, V. Alvarez-Iglesias, M.F. Pinheiro, M.R. Whittle, A. Brehm, J. Gomez, Mitochondrial DNA error prophylaxis: assessing the causes of errors in the GEP’02–03 proficiency testing trial, *Forensic Sci. Int.* 148 (2005) 191–198.
- [3] H.-J. Bandelt, L. Quintana-Murci, A. Salas, V. Macaulay, The fingerprint of phantom mutations in mitochondrial DNA data, *Am. J. Hum. Genet.* 71 (2002) 1150–1160.
- [4] H.-J. Bandelt, A. Salas, C. Bravi, Problems in FBI mtDNA database, *Science* 305 (2004) 1402–1404.
- [5] C. Dennis, Error reports threaten to unravel databases of mitochondrial DNA, *Nature* 421 (2003) 773–774.
- [6] P. Forster, To Err is Human, *Ann. Hum. Genet.* 67 (2003) 2–4.
- [7] H.-J. Bandelt, W. Parson, Fehlerquellen mitochondrialer DNA-Datensätze und Evaluation der mtDNA-Datenbank “D-Loop-BASE” [Sources of errors in mitochondrial DNA datasets and evaluation of the mtDNA database “D-Loop-BASE”], *Rechtsmedizin* 14 (2004) 251–255.
- [8] H.-J. Bandelt, T. Kivisild, J. Parik, R. Villems, C. Bravi, Y.-G. Yao, A. Brandstätter, W. Parson, Lab-specific mutation processes, in: H.-J. Bandelt, M. Richards, V. Macaulay (Eds.), *Human Mitochondrial DNA and the Evolution of *Homo sapiens**, Springer-Verlag, Berlin/Heidelberg/New York, 2006, Chapter 6.
- [9] A. Brandstätter, T. Sängler, S. Lutz-Bonengel, W. Parson, E. Béraud-Colomb, B. Wen, Q.-P. Kong, C.M. Bravi, H.-J. Bandelt, Phantom mutation hotspots in human mitochondrial DNA, *Electrophoresis* 26 (2005) 3414–3429.
- [10] W. Parson, The art of reading sequence electropherograms, *Ann. Hum. Genet.* 71 (2007) 276–278.
- [11] 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. Legal Med.* 118 (2004) 294–306.
- [12] A. Brandstätter, R. Klein, N. Duftner, P. Wiegand, W. Parson, Application of a quasi-median network analysis for the visualization of character conflicts to a population sample, of mitochondrial DNA control region sequences from southern Germany (Ulm), *Int. J. Legal Med.* 120 (2006) 310–314.
- [13] A. Brandstätter, H. Niederstätter, M. Pavlic, P. Grubwieser, W. Parson, Generating population data for the EMPOP database—an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example, *Forensic Sci. Int.* 166 (2007) 164–175.
- [14] M. Steinlechner, W. Parson, Automation and high through-put for a DNA database laboratory: development of a laboratory information management system, *Croat. Med. J.* 42 (2001) 252–255.
- [15] W. Parson, M. Steinlechner, Efficient DNA database laboratory strategy for high through-put STR typing of reference samples, *Forensic Sci. Int.* 122 (2001) 1–6.
- [16] H.J. Bandelt, A. Dür, Translating DNA data tables into quasi-median networks for parsimony analysis and error detection, *Mol. Phylogenet. Evol.* 42 (2007) 256–271.
- [17] H.-J. Bandelt, Q.-P. Kong, M. Richards, V. Macaulay, Estimation of mutation rates and coalescence times: some caveats, in: H.-J. Bandelt, M. Richards, V. Macaulay (Eds.), *Human Mitochondrial DNA and the Evolution of *Homo sapiens**, Springer-Verlag, Berlin/Heidelberg/New York, 2006, Chapter 4.
- [18] M.R. Wilson, M.W. Allard, K.L. Monson, K.W. Miller, B. Budowle, Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region, *Forensic Sci. Int.* 129 (2002) 35–42.
- [19] H.-J. Bandelt, W. Parson, Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *Int. J. Legal Med.*, in press.