Forensic Science International: Genetics 5 (2011) 133–137

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Application of a west Eurasian-specific filter for quasi-median network analysis: Sharpening the blade for mtDNA error detection

Bettina Zimmermann [a], Alexander Röck [b], Gabriela Huber [a], Tanja Krämer [c], Peter M. Schneider [c], Walther Parson [a,*]

[a] Institute of Legal Medicine, Innsbruck Medical University, Austria
[b] Institute of Mathematics, University of Innsbruck, Austria
[c] Institute of Legal Medicine, University Hospital of Cologne, Germany

ARTICLE INFO

Keywords:
mtDNA population data
West Eurasia
Network analysis
Filter analysis
Error detection
EMPOP

ABSTRACT

The application of quasi-median networks provides an effective tool to check the quality of mtDNA data. Filtering of highly recurrent mutations prior to network analysis is required to simplify the data set and reduce the complexity of the network. The phylogenetic background determines those mutations that need to be filtered. While the traditional *EMPOPspeedy* filter was based on the worldwide mtDNA phylogeny, haplogroup-specific filters can more effectively highlight potential errors in data of the respective (sub)-continental region. In this study we demonstrate the performance of a new, west Eurasian filter *EMPOPspeedyWE* for the fine-tuned examination of data sets belonging to macro-haplogroup N that constitutes the main portion of mtDNA lineages in Europe. The effects on the resulting network of different database sizes, high-quality and flawed data, as well as the examination of a phylogenetically distant data set, are presented by examples. The analyses are based on a west Eurasian etalon data set that was carefully compiled from more than 3500 control region sequences for network purposes. Both, etalon data and the new filter file, are provided through the EMPOP database (www.empop.org).

1. Introduction

Mitochondrial (mt) DNA data generation is challenging and prone to error [1]. Despite guidelines for failsafe mtDNA typing [2], serious problems are abundant and can lead to severe misunderstandings and errors in interpretation [1]. Consequently, there is a need for quality control to prevent the release of erroneous mtDNA data.

The application of quasi-median (QM) networks provides an opportunity to examine the quality of mtDNA data by graphically representing the genetic structure of the lineages in a data set [3]. Filtering of highly recurrent (i.e. homoplasic) mutations prior to network analysis is required to reduce the complexity of the resulting network, which then enables the detection of data idiosyncrasies and potential artefacts. In general, the number of filtered mutations is positively correlated with the simplicity of the resulting network torso. However, in order to keep the network analysis powerful for forensic quality control, filtered mutations should be kept at a minimum and only concern sites that display highly recurrent mutations in the given sample. We have previously presented the universal filter *EMPOPspeedy* [4] that was based on hotspot mutations observed in the worldwide phylogeny [5]. However, some of the mutations included there are not relevant for subsets of the west Eurasian phylogeny. Therefore, we here present the new filter *EMPOPspeedyWE* that is specific to the west Eurasian phylogeny. In addition, we have compiled a high-quality set of mtDNA haplotypes that is representative for the west Eurasian mtDNA pool and serves as etalon.

2. Materials and methods

Thirty-one DNA samples from different west Eurasian countries (Germany, France, Poland, Bosnia-Herzegovina, Romania, Switzerland, Italy, Austria, Turkey, Kazakhstan, Denmark, Portugal, Spain, and the Russian Federation) were subjected to entire mtDNA control region sequencing according to high quality amplification and sequencing procedures [6]. Volunteer donors gave written consent. Those samples were used to highlight the effect of applying a small data set to network analysis.

A reference data set of 3673 west Eurasian control region (CR) haplotypes served as basis for selecting the filtered mutations and

* Corresponding author at: Müllerstrasse 44, A-6020 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
doi:10.1016/j.fsigen.2010.10.003

the make-up of the etalon. These data were extracted from the EMPOP database [4], including their corresponding haplogroup affiliation based on the nomenclature updated in [7,Build 7]. Samples outside macrohaplogroup N as well as haplotypes that could not be assigned to a specific haplogroup within R0 by CR polymorphisms were removed (1201 samples affected). The final west Eurasian reference data set comprised 2472 defined mtDNA haplotypes of typically west Eurasian origin. Based on these data, 202 haplotypes were selected to compose the etalon data set (Table S1). This selection is based on the observation of haplotypes/haplogroups frequently present in Europe and was adapted to the requirements of network analysis. This involves a sample size of about 200 distinct haplotypes [4] that builds a reasonable network torso on its own and to which small data sets can be added to allow their depiction in a useful manner.

On the basis of the west Eurasian reference data set, the fluctuation rate of the observed mutations was determined. For this purpose, the haplotypes were clustered according to their major CR haplogroups and the relative frequency of the mutations with respect to these clusters was estimated. These ranged between 0.00% and 40.00% with an average positional log-likelihood ratio of 2.81. For deciding which mutations to include in the filter file, a fluctuation rate threshold of 0.85% was applied. Furthermore, additional mutations with lower thresholds that increased the complexity of the network were identified empirically by analyzing example data together with the etalon.

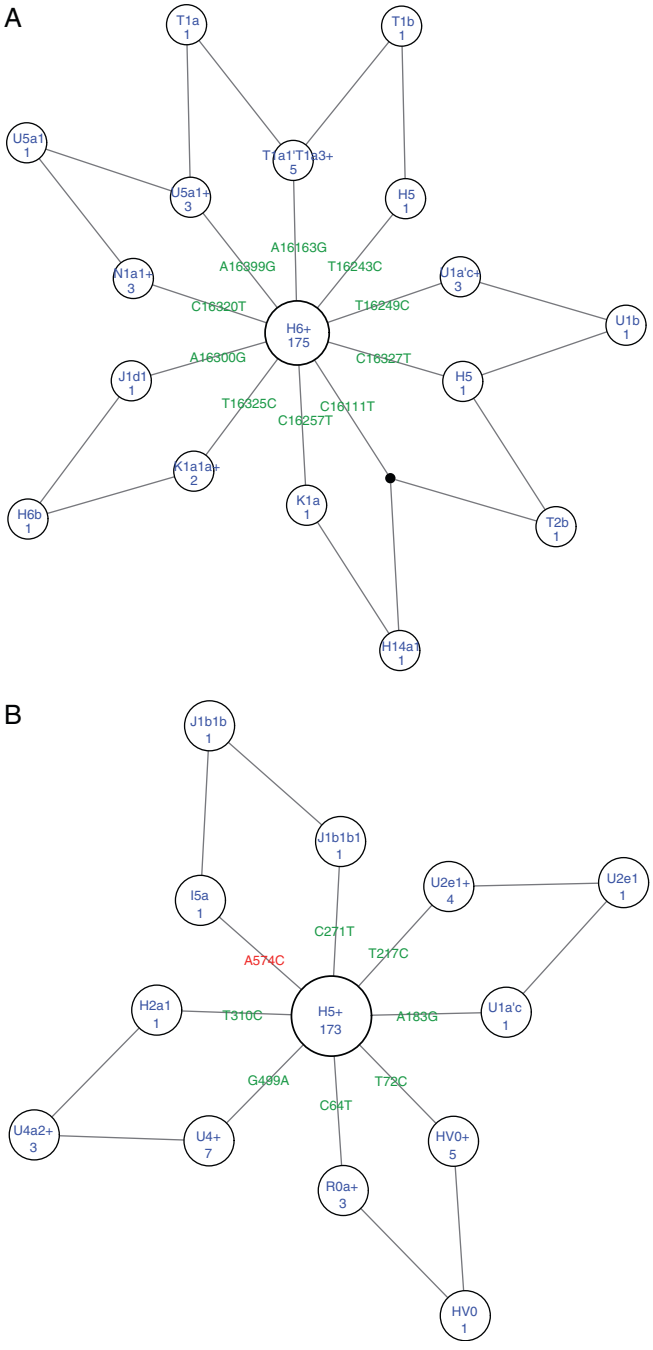

Fig. 1. QM network torsi of the etalon data set. (A) HVS-I: nps 16024–16569; (B) HVS-II: nps 1–576. The nodes correspond to reduced and condensed haplotypes. The most frequent haplogroup and a "+" are given if more than one haplogroups are involved. The number of condensed haplotypes is indicated below. The branches represent mutational events; transitions are shown in green, transversions in red. Small black nodes represent a quasi-median indicating that this haplotype was not observed in the data set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

3. Results and discussion

The presented west Eurasian-specific filter *EMPOPspeedyWE* contained a total of 111 mutational positions (Table S2), 29 less than the general *EMPOPspeedy* filter ([4]; see www.empop.org for details). When applied to the west Eurasian etalon data set of 202 haplotypes (Table S1) the network torsi of both hypervariable segments HVS-I and HVS-II displayed typical star-like structures (Fig. 1). This combination (etalon and filter) lends itself to the addition of small sample size data sets to enhance the demonstration of the included sequence information in the network. The

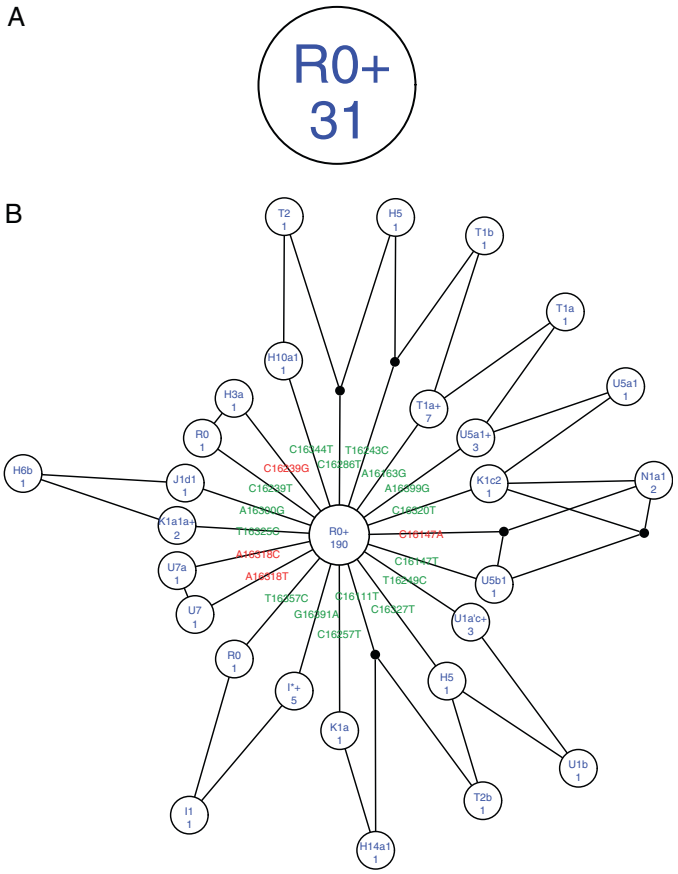

Fig. 2. QM network torsi (nps 16024–16569) of a small high-quality data set ($N = 31$) from west Eurasia. (A) The torso of the 31 samples. (B) The resulting torso of the 31 samples combined with the etalon ($N = 233$).

following examples illustrate the application of the west Eurasian etalon in combination with the new filter to mtDNA data sets of different sizes and qualities.

3.1. QM network analysis of a high-quality but small data set (N = 31)

The QM network torsi of a small high-quality data set ($N = 31$, Table S3) are shown in Fig. 2. The small sample size and the high quality of the data led to a very simple torso with all haplotypes being condensed into one node (Fig. 2A). The addition of these 31 haplotypes to the west Eurasian etalon resulted in a more complex but still star-like torso (Fig. 2B). The three-dimensional reticulation on the right side of the HVS-I torso can be explained by the known parallel occurrence of a transition and a transversion at position 16147 in two samples (C16147T: hg U5b1; C16147A: hg N1a1).

3.2. Effect of a small ambiguous data set on QM torsi (N = 29)

Another west Eurasian data set of similar size ($N = 29$; from Dagestan [8]) resulted in a different picture (Fig. 3). The torso of the 29 Darginian HVS-I haplotypes alone (Fig. 3A) already shows two three-dimensional reticulations, suggesting phantom mutations at positions 16280, 16281, 16384 and 16391. When combined with the etalon 13 reticulations were observed in the resulting network torso (Fig. 3B). These data (among others from [8]) were in the centre of a debate [9]. Finally, the authors presented their raw data and thus provided evidence for sequence misinterpretation due to low quality of the raw data [10].

3.3. Effect of a high-quality but oversized data set on QM network torsi (N = 786)

Oversized data sets produce crowded network torsi as enormous complexity arises from the high number of condensed haplotypes. Fig. 4A demonstrates an oversized data set by joining four (high-quality) data sets retrieved from the EMPOP database [6,11,12, Lutz-Bonengel et al., unpublished], comprising a total of 786 haplotypes. The resulting torso is still star-like but overloaded and impossible to read. Data sets of such size should be partitioned and then tested independently. Hence, we recommend the application of sample sizes between 200 and 300 haplotypes per query to get useful graphical representations as shown in Fig. 4B (273 Austrian Europeans from [6]). This high-quality Austrian data set exhibits one phylogenetically distant sample, a C5b1 haplotype (nested in macrohaplogroup M) that can be easily identified by the three-dimensional reticulations on the left bottom of the network torso.

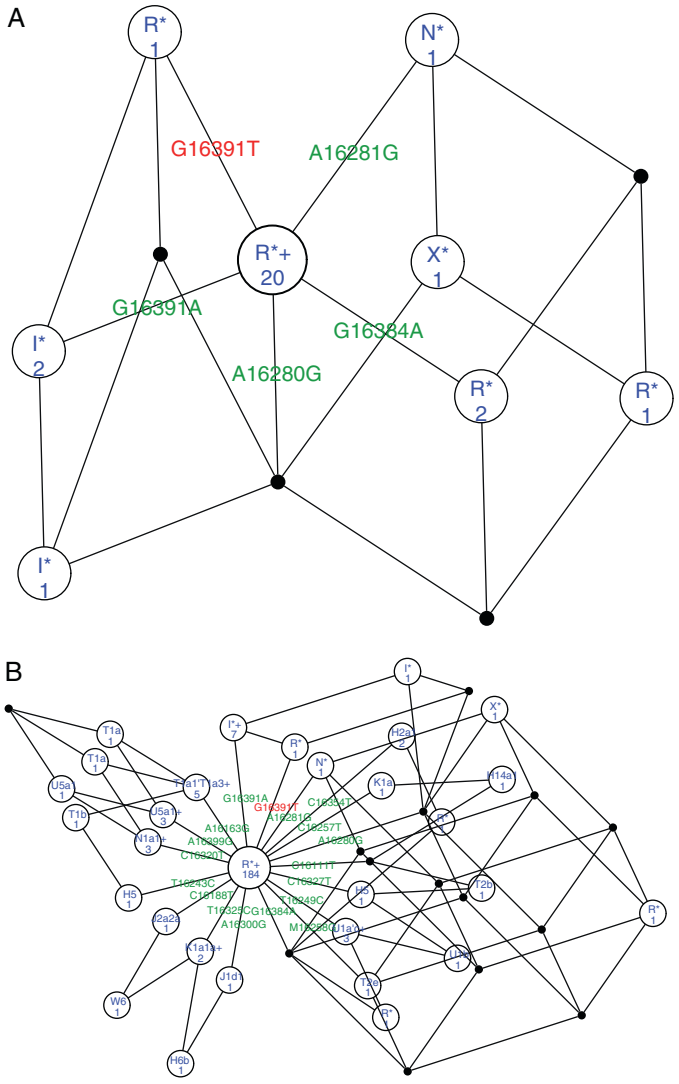

Fig. 3. QM network torsi (nps 16024–16400) of a Darginian data set ($N = 29$). (A) The torso of the 29 haplotypes. (B) The resulting torso of the 29 samples combined with the etalon ($N = 231$).

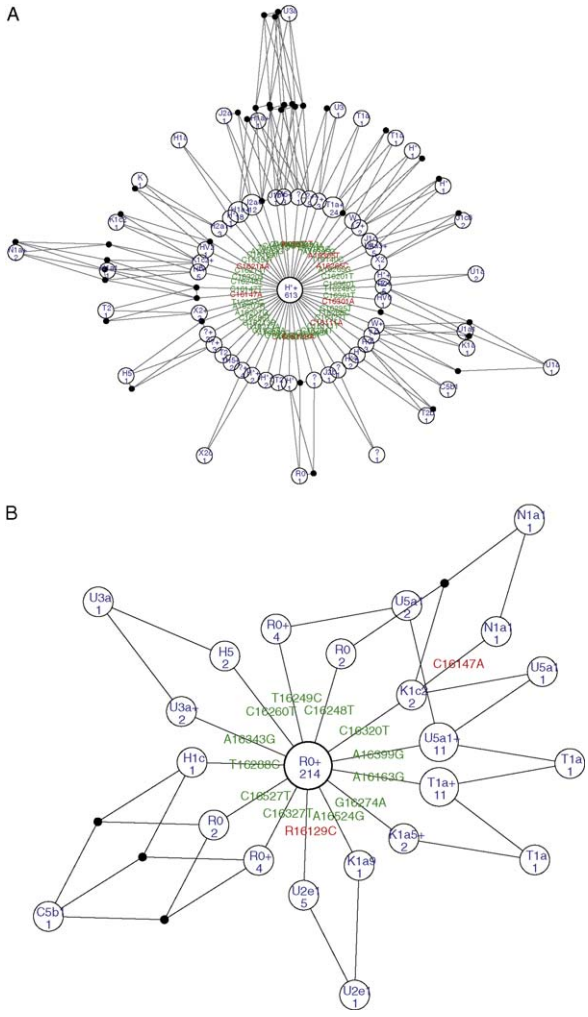

Fig. 4. QM network torsi of high-quality west Eurasian samples. (A) The torso (nps 16024–16365) of an oversized sample set ($N = 786$). (B) The torso (nps 16024–16569) of 273 Austrians.

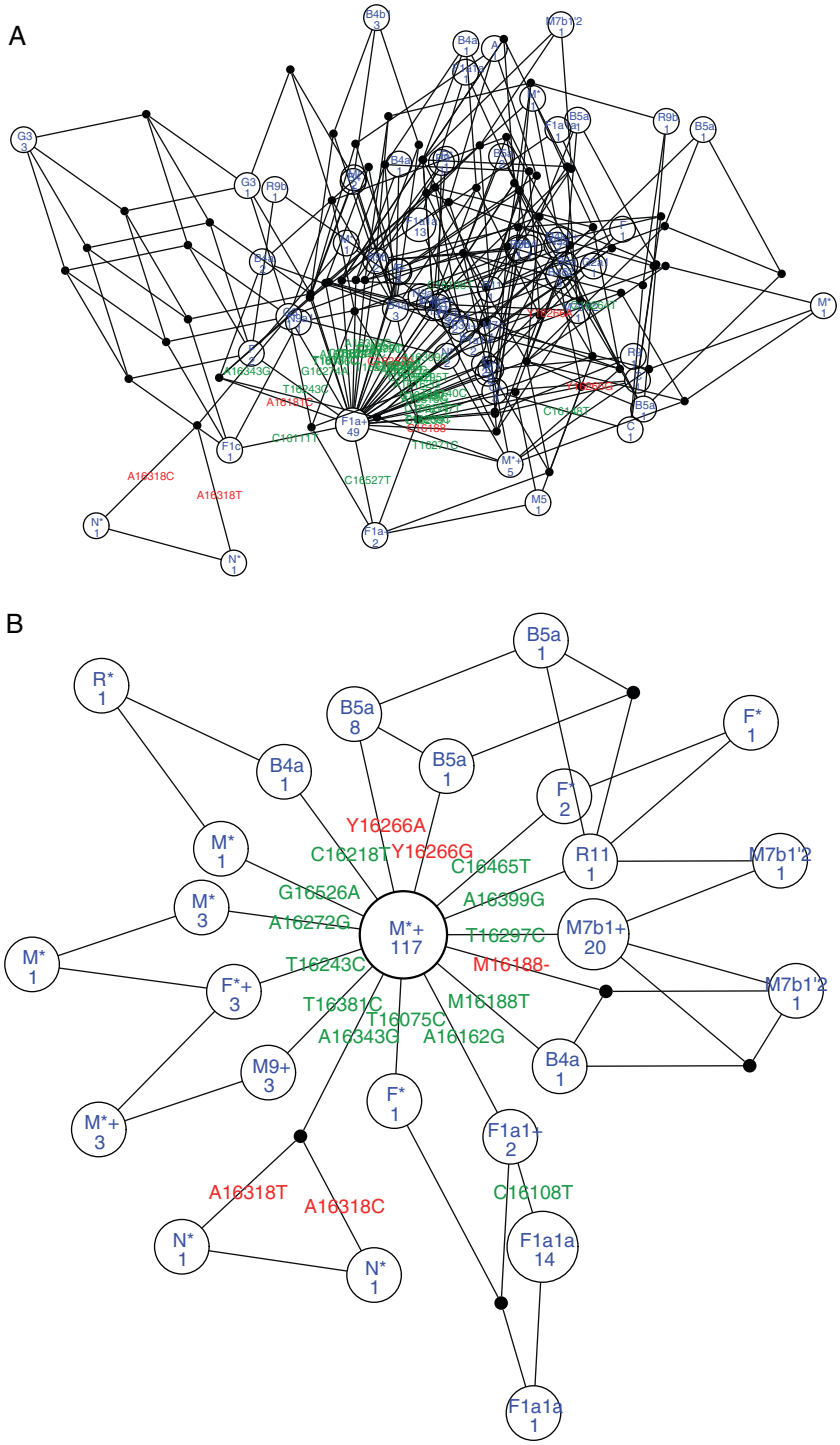

Fig. 5. QM network torsi of 190 haplotypes from Thailand (nps 16024–16569), constructed with the new west Eurasian-specific filter *EMPOPSpeedyWE* (A) and the *EMPOPspeedy* filter (B).

3.4. Application of a Southeast Asian data set to the west Eurasian filter (N = 190)

An apparent misapplication of the west Eurasian-specific filter is shown in Fig. 5. The investigated haplotypes originate from the Southeast Asian phylogeny (Thailand [13]). The complexity of the network torso in Fig. 5A points at the presence of mutational hotspots in the data that are not removed by the west Eurasian filter. For comparison, the torso of the same data after passage through the general *EMPOPspeedy* filter is shown in Fig. 5B, revealing (less specific) reduced complexity.

4. Conclusions

QM network analysis is a useful tool for the quality control of mtDNA sequences as data idiosyncrasies can be unmasked. The complexity of mtDNA data needs to be reduced to simplify the graphical representation of the network and to make it more powerful for the detection of errors. This is achieved by introducing a filter that targets highly recurrent mutations that would otherwise distort the network. The new filter *EMPOPspeedyWE* is specific to sieve homoplasic mutations typical to the west Eurasian mtDNA phylogeny. In combination with the presented etalon data set it is

powerful to examine even small sized sample sets of west Eurasian origin. Since homoplasic mutations from other parts of the phylogeny are not filtered, this approach also allows the detection of phylogenetically distant lineages that may be present in the data.

Acknowledgement

This study was supported by the FWF Austrian Science Fund (TR397).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2010.10.003.

References

[1] A. Brandstätter, T. Sänger, 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.
[2] A. Salas, A. Carracedo, V. Macaulay, M. Richards, H.J. Bandelt, A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics, Biochem. Biophys. Res. Commun. 335 (2005) 891–899.
[3] 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.
[4] W. Parson, A. Dür, EMPOP—a forensic mtDNA database, Forensic Sci. Int. Genet. 1 (2007) 88–92.
[5] 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).
[6] 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.
[7] M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation, Hum. Mutat. 30 (2009) E386–E394.
[8] I. Nasidze, M. Stoneking, Mitochondrial DNA variation and language replacements in the Caucasus, Proc. R. Soc. Lond. B: Biol. Sci. 268 (2001) 1197–1206.
[9] H.-J. Bandelt, T. Kivisild, Quality assessment of DNA sequence data: autopsy of a mis-sequenced mtDNA population sample, Ann. Hum. Genet. 70 (2006) 314–326.
[10] W. Parson, The art of reading sequence electropherograms, Ann. Hum. Genet. 71 (2007) 276–278.
[11] 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.
[12] S. Tetzlaff, A. Brandstätter, R. Wegener, W. Parson, V. Weirich, Mitochondrial DNA population data of HVS-I and HVS-II sequences from a northeast German sample, Forensic Sci. Int. 172 (2007) 218–224.
[13] B. Zimmermann, M. Bodner, S. Amory, L. Fendt, A. Röck, D. Horst, B. Horst, T. Sanguansermsri, W. Parson, A. Brandstätter, Forensic and phylogeographic characterization of mtDNA lineages from northern Thailand (Chiang Mai), Int. J. Legal Med. 123 (2009) 495–501.