Int J Legal Med (2006) 120:310–314
DOI 10.1007/s00414-006-0114-x

SHORT COMMUNICATION

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)

Anita Brandstätter · Rachel Klein · Nina Duftner · Peter Wiegand · Walther Parson

Received: 26 January 2006 / Accepted: 12 June 2006 / Published online: 27 July 2006
© Springer-Verlag 2006

Abstract Entire mtDNA control region sequences from 100 individuals in a west Eurasian population sample from southern Germany (around the city of Ulm) were generated and analyzed. The control region was amplified in one piece and sequenced with ten different sequencing primers. Sequence evaluation was performed independently. Phylogenetic analyses were used for quality assurance purposes and for the determination of the haplogroup affiliation of the samples. The sequences were scrutinized performing a quasi-median network analysis. To visualize character conflicts, frequent mutations were filtered, and the reduced data were represented by the torso of their quasi-median network. Character incompatibilities were found to be based on real biological patterns of homoplasy. The population data will be incorporated in the EMPOP database (http://www.empop.org).

Keywords West Eurasian population · mtDNA polymorphisms · Phylogenetic analyses · Haplogroups assignment · Forensic

Electronic supplementary material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00414-006-0114-x and is accessible for authorized users.

A. Brandstätter · N. Duftner · W. Parson (✉)
Institute of Legal Medicine, Innsbruck Medical University,
Müllerstrasse 44,
Innsbruck 6020, Austria
e-mail: walther.parson@i-med.ac.at

R. Klein · P. Wiegand
Institute of Legal Medicine, University Clinics Ulm,
Ulm, Germany

N. Duftner
Bauer Center for Genomics Research, Harvard University,
Cambridge, MA, USA

Introduction

The compilation of mtDNA population databases through sequencing of the hypervariable segments (HVS-I and HVS-II) of the mtDNA control region (CR) has become a routine matter in the past decade in many anthropological and forensic laboratories [1–3]. The discovery and description of highly polymorphic sites outside the commonly tested HVS-I/HVS-II ranges have triggered a trend to target the entire mtDNA CR for the assessment of intraspecific mtDNA variation [4–7]. Although the general sequencing technology (including polymerases, sequencing chemistries, and sequencing platforms) has improved greatly in the past few years, the results of sequencing efforts are still not free of artifacts. Collaborative exercises performed by different forensic working groups have shown that clerical errors (including misreading of sequencing outputs and errors in data transcription and preparation of population data tables) are mainly responsible for flawed mtDNA population data compilations [8, 9]. In addition, phantom mutations, which are generated in the course of the sequencing process, have been creeping in mtDNA datasets [10, 11].

In the course of the establishment of the EMPOP mtDNA population database [8, 12], a strategy for the generation of high-quality mtDNA population data has been developed [7]. The high quality of the data was assured by contiguous double-stranded sequence coverage (also in the case of length heteroplasmy), independent double evaluation of the data, electronic export of the CR profiles from the sequence analysis software to the Laboratory Information Management System, comparison and validation of the profiles by a third senior mtDNA scientist, and a phylogenetic analysis for haplogroup assignment.

As an additional tool for quality assurance, the dataset was inspected for unusual variants by application of a quasi-median network analysis. The applied median-joining algorithm [13] visualizes character incompatibilities, which are displayed as hypercubes [14] in the network. Thus, these reticulations pinpoint suspicious sites, which deserve detailed attention and confirmation by the raw data.

Materials and methods

DNA samples, extraction, amplification, and sequencing

One hundred samples were collected from native German speakers in the middle of southern Germany, a region that is situated between Lake Constance and the Swabian Alb. Selection criteria involved a German surname and a birth place in the above-mentioned area. The selected samples stem from unrelated persons as far as this is possible to determine with the above-mentioned selection criteria. DNA was extracted from blood samples and buccal swabs using Chelex 100 as outlined [15]. Amplification and sequencing has been described in detail [7]. In short, aliquots of DNA extracts were dispensed in 96-well plates (including no-template controls), and 2 μl of the DNA extracts was amplified in an 18-μl PCR master mix containing 1.0 unit of AmpliTAQ Gold polymerase (AB, Foster City, CA, USA), 1.0 unit of PCR reaction buffer (AB), 200 μM each dNTP (AB), and 0.5 μM each primer (L15900 and H00599). Cycling conditions were initial denaturation at 95°C (11 min) followed by 35 reaction cycles at 95°C (15 s), 56°C (30 s), and 72°C (90 s) with a final extension at 72°C (10 min). ExoSAP-IT (USB, Cleveland, OH, USA) was used for post-PCR treatment, and the BigDye Terminator v1.1 Cycle Sequencing RR mix (AB) was used for cycle sequencing, both according to the manufacturers' protocols. Each template was sequenced in the forward direction with primers L15971, L15989, L16268, L00015, L00029, L00314, and L00361 and in the reverse direction with primers H00016, H00159, and H00484. Sequencing reaction products were purified using Sephadex G-50 Fine (Amersham, Buckinghamshire, UK) and Multiscreen filter plates (Millipore). Electrophoretic separation was carried out on an ABI3100 capillary sequencer using POP6 and a 36-cm capillary array.

Data analysis and phylogenetic analyses

All sequences were evaluated twice using the sequence analysis software SeqScape (Version 2.0, AB) by two independent scientists. The sequences were aligned and compared to the revised Cambridge Reference Sequence (rCRS [16, 17]) following the nomenclature guidelines for mtDNA typing [18, 19]. The sequence evaluations were exported from SeqScape as mutation reports, which were directly imported into a data storage and data evaluation software, specifically designed for mtDNA analysis and implemented into the previously described in-house LIMS [20]. This software allows electronic comparison of the exported haplotypes to be finally reviewed by a third scientist. After evaluation, the mtDNA haplotypes were stored and assembled for subsequent phylogenetic analyses and final database export. Phylogenetic analyses were performed as described in the paper of Brandstätter et al. [7] with in-house computer programs. In brief, the data were compared to assembled mtDNA CR sequences with known haplogroup affiliations, and the haplogroup status suggested by a nearest neighbor search was confirmed by checking the presence or absence of haplogroup-diagnostic sites. The mtDNA haplotypes from this study were affiliated to (sub)haplogroups based on the patterns of shared haplogroup-specific or haplogroup-associated polymorphisms in the CR as previously reported [21–32].

Quasi-median network analysis

The consensus sequences were automatically converted into a table using an in-house computer program. Transitions at the following sites, which are known to show a higher phylogenetic variability [10, 33–36], were filtered from the dataset, whereas deletions and transversions at these positions remained: 93; 146; 150; 151; 152; 182; 183; 185; 189; 194; 195; 198; 199; 200; 204; 207; 228; 499; 513; 16051; 16069; 16093; 16111; 16114; 16124; 16126; 16129; 16145; 16147; 16148; 16153; 16163; 16166; 16167; 16168; 16169; 16172; 16173; 16176; 16184; 16186; 16187; 16188; 16189; 16192; 16207; 16209; 16214; 16218; 16223; 16224; 16230; 16234; 16243; 16245; 16248; 16249; 16256; 16260; 16261; 16264; 16265; 16266; 16270; 16271; 16278; 16287; 16290; 16292; 16293; 16294; 16295; 16298; 16300; 16309; 16311; 16316; 16318; 16319; 16320; 16325; 16327; 16344; 16354; 16355; 16360; 16362; 16390; and 16519. Furthermore, C-stretch length variation in HVS-I (around np 16189), HVS-II (around np 310), and HVS-III (around np 573) as well as variations in the dinucleotide repeat in HVS-III [37] was not taken into consideration.

To obtain networks feasible to display, the dataset was partitioned into four smaller datasets comprising segments 16024–16365, 16366–72, 73–340, and 341–576, respectively. Then, the quasi-median networks of these subsets were constructed with the computer program Network 4.1 (Shareware Phylogenetic Network Software Web site) by applying the median joining algorithm [13] with the tolerance threshold epsilon set to 99. Incompatibilities between characters are manifest in squares and higher

dimensional cubes in the associated median network; these reticulations (i.e., cycles) constitute the torso of the network. For the visualization of the sites that cause reticulations, the resulting graphs were drawn with Network 4.1 by selecting the "display only torso" option.

Results and discussion

Haplotype frequencies

We determined the complete mitochondrial DNA CR nucleotide sequences of 100 individuals from the middle of southern Germany around the city of Ulm (Table 1). We observed 91 distinct haplotypes defined by 137 variable sites (taking length variants of the C tracts in HVS-I, HVS-II, and HVS-III into consideration). Among these, 84 sequences were observed only once, five sequences were found twice, and two sequences were detected three times. The most frequently observed haplotypes were 73G-150T-263G-309.1C-315.1C-16343G (hg U3) and 73G-263G-315.1C-16126C-16294T-16296T-16304C-16519C (hg T2). Ignoring length variants of the major C-tracts for distinguishing haplotypes, 87 different sequences were observed. From these, 79 haplotypes were observed once, five sequences were found twice, two sequences were detected three times, and one haplotype was observed five times. The most abundant sequence was the T2-haplotype 73G-263G-16126C-16294T-16296T-16304C-16519C.

Haplogroup structure

The most common haplogroup, cluster HV (including samples belonging to haplogroups HV0, H*, and known subhaplogroups of H), was observed in 40% of the southern Germany sample set. A refined classification within haplogroup H led to the assignment of samples to haplogroups H1a (3%), H1b (2%), H5 (7%), H6 (2%), H8 (1%), and H11 (2%). Samples belonging to haplogroup H, which could not be resolved genealogically in subclades [27, 28, 38], were assigned to haplogroup H* (20%) whereby one sample was assigned to haplogroup HV* and haplogroup HV0 was found in two individuals. Haplogroup I was represented as subhaplogroups I1, I1a, and I3 (combined 4%). Notably, all haplogroup I samples in this population (and also all I samples from the work of Brandstätter et al. [7]) showed a differing number of C-insertions in the C-stretch at the 3′ end of the CR between positions 568 and 573 resulting in length heteroplasmy. The incidence of length heteroplasmy in this segment and the associated nomenclature suggestions have been described before [5] and have been observed occasionally in various haplogroups, e.g., in K and T in this population sample, in L3e3 and L0f in a population from Kenya [5], and in H and U5 in an Austrian population sample [7]; however, in haplogroup I, the occurrence of this special length heteroplasmy seems obligatory [29, 31]. The 573insCC polymorphism might however not only be associated to the I branch of the macrohaplogroup N phylogeny as proposed [29] but might be located on a deeper rooting branch before the N1a/I split, as also the two N1a sequences in the work of Brandstätter et al. [7] show additional C-insertions between positions 568 and 573. Two sequences in the study of Palanichamy et al. [29] (C124 and DMe22i) showed the 455.1T insertion polymorphism, but were not associated to a new haplogroup inside haplogroup I, although clearly distinguished from haplogroup I1a. Also, the 455.1T insertion polymorphism was found in two haplogroup I sequences from this southern Germany population sample, one in an I1a sequence and the other in an I1 sequence. Therefore, 455.1T is either characteristic of a new subhaplogroup of I1 or paraphyletic within I1.

Of the samples, 12% accounted for haplogroup J (consisting of subhaplogroups J1c, J1c1, J2a, and J2b), eight individuals belonged to haplogroup K (here found as K1a and K1c), and 14 individuals were among the haplogroup T lineages (comprising T1a, T2, and T2b). Haplogroup U constituted 18% of the samples, with classifications to U2e, U3, U4, U5a, U5a1, U5b, U5b1, and U7. The two haplogroup X samples were assigned to subhaplogroups X2b and X2c. One sample belonged to the Asian haplogroup M6b, and one sample could not be assigned to any haplogroup described so far. It is interesting to note that neither haplogroup N1a nor haplogroup W was found in the southern Germany population sample; however, the basic phylogenetic structure of this west Eurasian population as derived from its haplogroup distribution is in agreement with other west Eurasian populations [7, 22, 25, 33, 39, 40].

Quasi-median network analysis

For a weighty view [10] of continental west Eurasian mtDNA variation, the dataset from southern Germany was partitioned into four smaller subsets (in fact the three hypervariable segments plus the segment between HVS-I and HVS-II). Then, following the guidelines of Bandelt et al. [10], hypervariable polymorphisms ("speedy transitions") were filtered to condense the data to the weighty variation only. Figure 1 displays the torsos of the resulting weighty networks, in each of which the most frequent haplotype corresponds to the rCRS: the torso of the HVS-I network (16024–16365) displayed a single four-cycle (Fig. 1a), indicating character incompatibility between positions 16296 and 16304, which was caused by the

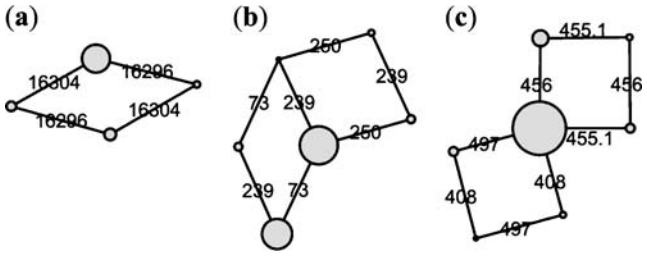

Fig. 1 Quasi-median portraits of the torsos compiled from the 100 CR sequences from southern Germany in the sequence segments **a** 16024–16365, **b** 70–340, and **c** 452–459

double association of the 16304 polymorphism to the two haplogroups H5 and T2b. The network of the segment between HVS-I and HVS-II (16366-72) was a perfect tree (i.e., no homoplasy manifestation in the weighty data), thus the torso was empty; the torsos of the HVS-II network (73–340) as well as the HVS-III-network (341–576) contained two four-cycles (Fig. 1b,c, respectively). The two four-cycles in the HVS-II-torso, which shared one edge, were the result of the occurrence of the haplogroup H6 diagnostic site 239 [27, 28] as a private polymorphism in a haplogroup I sequence. In the HVS-III-torso, one four-cycle was caused by the occurrence of a T-insertion at position 455 (noted as 455.1T) once on a haplogroup H5, and twice on a haplogroup I1-background, while the other four-cycle was generated by a transition polymorphism on position 408 in two sequences, one belonging to haplogroup H* and the other one belonging to haplogroup K1a.

To draw a conclusion on the implementation of a quasi-median network analysis to an mtDNA population sample of CR sequences, it can be said that the computer program Network is easy to use once the data are in the required format. The conversion of a data table into the requested alignment can be tricky, but with the aid of programs available from V. Macaulay's homepage (http://www.stats.gla.ac.uk/~vincent/fingerprint/index.html), or with in-house conversion software, or also with the data entry wizard of the program Network itself, the conversion process can quickly become a routine matter. The occurrence of one or two four-cycles in a network based on an ~350-bp mtDNA CR segment from a west Eurasian population sample is not alarming and will in most cases refer to low levels of homoplasy. Higher dimensional cubes in the network however hint to character incompatibilities that cannot be explained by natural levels of homoplasy only and should thus trigger a closer inspection of the data.

The application of a systematic quality check is a very useful tool for forensic science and human evolution studies, and the results of the quasi-median network analysis on entire CR sequences (partitioned in four segments) from southern Germany can be used to compare the results of similar analyses on other west Eurasian populations, thus serving as a benchmark dataset.

Conclusions

The mtDNA CR data presented in this article were generated according to high-quality laboratory standards, and the frequencies of the different haplogroups are characteristic for west Eurasian populations. The results of the quasi-median network analysis (i.e., the torsos of the networks of the segments 16024–16365, 16366–72, 73–340, and 341–576) can be used as benchmarks for similar analyses on other west Eurasian mtDNA population samples. Character incompatibilities were based on low levels of homoplasy. The population data will be incorporated into the EMPOP mtDNA database.

References

1. Parson W, Parsons TJ, Scheithauer R, Holland MM (1998) Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: application of mtDNA sequence analysis to a forensic case. Int J Legal Med 111:124–132
2. Imaizumi K, Parsons TJ, Yoshino M, Holland MM (2002) A new database of mitochondrial DNA hypervariable regions I and II sequences from 162 Japanese individuals. Int J Legal Med 116:68–73
3. Imaizumi K, Parsons TJ (2005) A new database of mitochondrial DNA hypervariable regions I and II sequences from 162 Japanese individuals (vol 116, p 68, 2001). Int J Legal Med 119:183
4. Lutz S, Weisser H-J, Heizmann J, Pollak S (1998) Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany. Int J Legal Med 111:67–77. Errata in Int J Legal Med 111 (1998) 286 and Int J Legal Med 112 (1999) 145–150
5. Brandstätter A, Peterson CT, Irwin JA et al (2004) Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. Int J Legal Med 118:294–306
6. Vanecek T, Vorel F, Sip M (2004) Mitochondrial DNA D-loop hypervariable regions: Czech population data. Int J Legal Med 118:14–18
7. Brandstätter A, Niederstätter H, Pavlic M et al (2006) 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 (in press)
8. Parson W, Brandstätter A, Alonso A et al (2004) The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation, results and perspectives. Forensic Sci Int 139:215–226
9. Salas A, Prieto L, Montesino M et al (2005) Mitochondrial DNA error prophylaxis: assessing the causes of errors in the GEP'02-03 proficiency testing trial. Forensic Sci Int 148:191–198
10. Bandelt H-J, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71:1150–1160

11. Brandstätter A, Sänger T, Lutz-Bonengel S et al (2005) Phantom mutation hotspots in human mitochondrial DNA. Electrophoresis 26:3414–3429
12. Parson W, Brandstätter A, Pircher M et al (2004) EMPOP—the EDNAP mtDNA population database concept for a new generation, high-quality mtDNA database. Int Congress Series 1261:106–108
13. Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48
14. Bandelt H-J, Macaulay V, Richards M (2000) Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. Mol Phylogenet Evol 16:8–28
15. Ambach E, Parson W, Niederstätter H, Budowle B (1997) Austrian Caucasian population data for the quadruplex plus amelogenin: refined mutation rate for HumvWFA31/A. J Forensic Sci 42:1136–1139
16. Anderson S, Bankier AT, Barrell BG et al (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465
17. Andrews RM, Kubacka I, Chinnery PF et al (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147
18. Bär W, Brinkmann B, Budowle B et al (2000) DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Int J Legal Med 113:193–196
19. Carracedo A, Bär W, Lincoln P et al (2000) DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Forensic Sci Int 110:79–85
20. Steinlechner M, Parson W (2001) Automation and high throughput for a DNA database laboratory: development of a laboratory information management system. Croat Med J 42:252–255
21. Helgason A, Sigureth aS, Gulcher JR et al (2000) mtDNA and the origin of the Icelanders: deciphering signals of recent population history. Am J Hum Genet 66:999–1016
22. Malyarchuk BA, Grzybowski T, Derenko MV et al (2002) Mitochondrial DNA variability in Poles and Russians. Ann Hum Genet 66:261–283
23. Derenko MV, Malyarchuk BA, Dambueva IK, Zakharov IA (2003) Structure and diversity of the mitochondrial gene pools of south Siberians. Dokl Biol Sci 393:557–561
24. Derenko MV, Grzybowski T, Malyarchuk BA et al (2003) Diversity of mitochondrial DNA lineages in south Siberia. Ann Hum Genet 67:391–411
25. Malyarchuk BA, Grzybowski T, Derenko MV et al (2003) Mitochondrial DNA variability in Bosnians and Slovenians. Ann Hum Genet 67:412–425
26. Reidla M, Kivisild T, Metspalu E et al (2003) Origin and diffusion of mtDNA haplogroup X. Am J Hum Genet 73:1178–1190
27. Achilli A, Rengo C, Magri C et al (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75:910–918
28. Loogväli E-L, Roostalu U, Malyarchuk BA et al (2004) Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. Mol Biol Evol 21:2012–2021
29. Palanichamy MG, Sun C, Agrawal S et al (2004) Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet 75:966–978
30. Achilli A, Rengo C, Battaglia V et al (2005) Saami and Berbers—an unexpected mitochondrial DNA link. Am J Hum Genet 76:883–886
31. Bandelt H-J, Achilli A, Kong Q-P et al (2005) Low "penetrance" of phylogenetic knowledge in mitochondrial disease studies. Biochem Biophys Res Commun 333:122–130
32. Torroni A, Achilli A, Macaulay V et al (2006) Harvesting the fruit of the human mtDNA tree. Trends Genet 22:339–345
33. Allard MW, Miller K, Wilson M et al (2002) Characterization of the Caucasian haplogroups present in the SWGDAM forensic mtDNA dataset for 1771 human control region sequences. Scientific Working Group on DNA Analysis Methods. J Forensic Sci 47:1215–1223
34. Malyarchuk BA, Rogozin IB, Berikov VB, Derenko MV (2002) Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region. Hum Genet 111: 46–53
35. Allard MW, Wilson MR, Monson KL, Budowle B (2004) Control region sequences for East Asian individuals in the Scientific Working Group on DNA Analysis Methods forensic mtDNA data set. Leg Med 6:11–24
36. Allard MW, Polanskey D, Miller K et al (2005) Characterization of human control region sequences of the African American SWGDAM forensic mtDNA data set. Forensic Sci Int 148: 169–179
37. Lutz S, Weisser H-J, Heizmann J, Pollak S (1997) A third hypervariable region in the human mitochondrial D-loop. Hum Genet 101:384
38. Tambets K, Rootsi S, Kivisild T et al (2004) The western and eastern roots of the Saami—the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. Am J Hum Genet 74:661–682
39. Brandstätter A, Parsons TJ, Parson W (2003) Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups. Int J Legal Med 117:291–298
40. Poetsch M, Wittig H, Krause D, Lignitz E (2003) Mitochondrial diversity of a northeast German population sample. Forensic Sci Int 137:125–132