



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

Anita Brandstätter, Harald Niederstätter, Marion Pavlic, Petra Grubwieser, Walther Parson*

Institute of Legal Medicine, Innsbruck Medical University, Müllerstr. 44, 6020 Innsbruck, Austria

Received 27 February 2006; received in revised form 9 May 2006; accepted 9 May 2006

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. Here we describe a safe laboratory scheme involving electronic data handling and computer-aided data transfer, which help to minimize errors originating from potential sample mix-up, data misinterpretation and incorrect transcription. The procedure is demonstrated by example of an mtDNA control region population study on 273 unrelated individuals from Austria. Our population sample was compared with five other European populations via an analysis of molecular variance (AMOVA). The inclusion of regions outside HVS-I and HVS-II increased the amount of information on the haplogroup diagnostic sites in the control region. Most of the haplotypes in Austrians fell into haplogroups H, J, K, T, and U. The random match probability in Austrians was 1:125; the average number of nucleotide differences between individuals in the Austrian database was 9.32.

© 2006 Published by Elsevier Ireland Ltd.

Keywords: Mitochondrial DNA polymorphisms; Haplogroup; Phylogenetic; Mean pairwise distances; AMOVA; FST values; Random match probabilities

1. Introduction

Forensic molecular biology takes advantage of the high copy number of mitochondrial DNA (mtDNA) molecules in a cell, and meanwhile, mtDNA typing has become routine in its application to analyze samples where the amount of genomic DNA is very small or degraded. Because of the lack of recombination, mtDNA profiles are not unique but regarded as haplotypes. When an mtDNA haplotype derived from an evidence sample cannot be discriminated from the one of a reference sample (e.g. from the suspect), its relative rarity can be estimated by comparing it to a collection of mtDNA sequences – usually assembled in databases – in order to support assessment of the weight to the evidence [1].

Some of the published mtDNA data have faced severe criticism in the past, as an unacceptable high rate of errors was detected therein (e.g. [2–4]). Although good laboratory practice guidelines have been published [5,6], problems still occur in the

course of mtDNA data generation. Sources of these errors are meanwhile well known and have been described in some detail [2,7,8], however, the amplification and sequencing process as well as *a posteriori* data control and transfer of haplotypes still bear shortcomings which result in erroneous data.

We here present in detail a reliable laboratory concept for mtDNA typing based on the generation of redundant sequence information for unequivocal base-assignment, independent double data evaluation, and IT-based (manual-free) comparison and transfer of results. This newly developed strategy is showcased on 273 Austrian mtDNA control region sequences. The thus created population data will be incorporated in the EMPOP database (<http://www.empop.org/>) [7,9].

2. Materials and methods

2.1. DNA samples and extraction

DNA was extracted from blood samples obtained from 273 unrelated West-Euradians from Austria using Chelex 100 as outlined in [10].

* Corresponding author. Tel.: +43 512 5073303; fax: +43 512 5072764.

E-mail address: walther.parson@i-med.ac.at (W. Parson).

2.2. Amplification and sequencing of the entire mtDNA control region

In order to facilitate efficient sequencing of the mitochondrial DNA control region, a 96-well method of processing population samples was developed independently by two laboratories: AFDIL, Rockville, USA [11–13] and GMI Innsbruck, Austria [14]. In the latter aliquots of DNA extracts were dispensed in 96-well plates; four negative control samples were assayed per plate. For amplification, to each well of a MicroAmp Optical 96-well reaction plate (AB), 18 μ l of PCR master mix containing 1.0 unit of AmpliTaq Gold polymerase (AB, Foster City, CA), 1.0 unit of PCR reaction buffer (AB), 200 μ M each dNTP (AB), and 0.5 μ M each primer (L15900 and H00599) were dispensed. Then 2 μ l DNA extract were put into the PCR master mix with a multichannel pipette. The amplification reaction was conducted on a thermal cycler (e.g. 9600/9700 GeneAmp Thermal Cycler; AB). The reaction cocktails were heated to 95 °C (11 min) and then put through 35 reaction cycles: 95 °C for 15 s, 56 °C for 30 s and 72 °C for 90 s, followed by a final extension phase at 72 °C for 10 min. PCR primers and unincorporated dNTPs were removed by adding 8 μ l of ExoSAP-IT (USB, Cleveland, OH) with a repeater pipette (Eppendorf Multipette; Hamburg, Germany) and heating the samples to 37 °C (15 min) for enzyme activation and then to 80 °C (15 min) for enzyme deactivation. Subsequently, 2 μ l of purified PCR product were combined with the sequencing master mix (containing 2 μ l BigDye Terminator v1.1 Cycle Sequencing RR mix (AB), 2 μ l BigDye Terminator v1.1 Sequencing Buffer (AB), 1.6 pmol primer and distilled water up to 8 μ l) with an 8-channel epMotion workstation (Eppendorf, Germany). Cycle sequencing was performed (after a first denaturation step of 95 °C, 1 min) for 25 cycles of 10 s at 95 °C, 5 s at 50 °C, and 4 min at 60 °C. 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 (Table 1(a) and (b)). Sequencing reaction products were purified from residual dye terminators using Sephadex G-50 Fine (Amersham, Buckinghamshire, UK) and Multi-screen filter plates (Millipore) according to the manufacturer's protocol. To this end, the cycle sequencing products were diluted by adding 10 μ l of distilled water and the dilutions were centrifuged through the filter plate into an optical 96-well plate for electrophoretic separation. The entire procedure of diluting cycle sequencing products and transferring the dilutions onto the Sephadex columns in the filter plate was again performed by the epMotion workstation. When spinning cycle sequencing products through the filter plate, unequal amounts of product may be recovered throughout the plate. In order to avoid this, the blocks' orientations in the centrifuge carriage were reversed after 2.5 min and the blocks were spun a second time for 2.5 min to obtain consistent amounts of purified products. Electrophoretic separation was carried out on an ABI3100 capillary sequencer using POP6 and a 36 cm capillary array.

Table 1
Sequences of primers

Primer	Nucleotide sequence
(a) Sequences of primers used for amplification and sequencing of the entire mitochondrial DNA control region	
H00016	5' TGA TAG ACC TGT GAT CCA TCG TGA 3'
H00159	5' AAA TAA TAG GAT GAG GCA GGA ATC 3'
H00484	5' TGA GAT TAG TAG TAT GGG AG 3'
H00599	5' TTG AGG AGG TAA GCT ACA TA 3'
L00015	5' CAC CCT ATT AAC CAC TCA CG 3'
L00029	5' CTC ACG GGA GCT CTC CAT GC 3'
L00314	5' CCG CTT CTG GCC ACA GCA CT 3'
L00361	5' ACA AAG AAC CCT AAC ACC AGC 3'
L15900	5' TCA AAG CTT ACA CCA GTC TTG TAA ACC 3'
L15971	5' TTA ACT CCA CCA TTA GCA CC 3'
L15989	5' CCC AAA GCT AAG ATT CTA AT 3'
L16268	5' CAC TAG GAT ACC AAC AAA CC 3'
(b) Sequencing primers that were published in [14] and that were replaced in the present study ^a	
L16169 (replaced by L00029)	5' CCC CCC CCC CAT G 3'
L00318 (replaced by L00314)	5' CCC CCC CCC CCC GCT 3'
L00403 (replaced by L00361)	5' TCT TTT GGC GGT ATG CAC TTT 3'
H16410 (replaced by H00016)	5' GAG GAT GGT GGT CAA GGG AC 3'

^a The primers, which they were substituted for, are written in parentheses.

2.3. Data analysis and quality assurance

After the raw data were analyzed with Sequencing Analysis (Version 3.3, Applied Biosystems, AB, Foster City, CA, USA), the sequences were aligned and the basecalls were scrutinized twice by two independent scientists. Consensus sequences were aligned and compared to the revised Cambridge Reference Sequence (rCRS; [15,16]) using Sequencher (Version 4.1.4Fb4, GeneCodes, Ann Arbor, MI, USA), following nomenclature guidelines for mtDNA typing [5,6]. In an independent step the sequences were evaluated by another scientist using the sequence analysis and alignment software SeqScape (Version 2.0, AB). The results of both analyses were exported 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 Laboratory Information Management System (LIMS) ([17]). This software allows for electronic comparison of the exported haplotypes, which are finally reviewed by a third scientist. After evaluation the mtDNA haplotypes were stored and assembled for subsequent phylogenetic analyses and final database export.

2.4. Haplogroup assignment

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 control region, as reported in [18–27]. The haplogroup assignments were confirmed with the results of a previous study involving the analysis of 16 phylogenetic informative single nucleotide polymorphisms (SNPs) from the coding region [28].

2.5. Random match probability

The random match probability was calculated as the sum of the squares of the haplotype frequencies [29]. C-stretch length variants in HVS-I (around position 16189) and in HVS-II (around position 310) were ignored in distinguishing haplotypes for calculation of random match probability.

2.6. Population genetic and molecular evolution analysis

The Austrian data were compared with 93 control region (CR) sequences from the Czech Republic [30], 244 CR sequences from Poland [19], 156 CR sequences from Bosnia-Herzegovina [22], 104 CR sequences from Slovenia [22] and 200 CR sequences from Germany [31]. For all comparisons, C-stretch length variants in HVS-I and HVS-II were ignored. Molecular diversity indices, analysis of molecular variance (AMOVA) and pairwise differences were calculated with the ARLEQUIN software (Version 2.0; [32]). Pairwise F_{ST} values were used to describe the short-term genetic distance between populations. Permutation tests (1000 replicates) were used to evaluate the significance of calculated genetic distances between populations. For this alignment, sequences were trimmed to fit the greatest common range 16024–16368 and 71–340.

2.7. Phylogenetic analyses

Two different kinds of phylogenetic analyses were performed on the data. The first method was aimed at identifying possible artificial HVS-I/HVS-II recombinants (i.e. samples that contain a combination of HVS-I sequences from one person and HVS-II sequences from another person). The occurrence of artificial recombination in mtDNA databases has been described several times [33,3,7], and is difficult to address as this phenomenon cannot be detected by evaluation of the raw sequence data. To address this problem, we have developed an IT-based solution to systematically search for artificial recombinants in population data that consist of HVS-I + HVS-II-sequences. In brief, the computer program identifies the mitochondrial haplogroup of each hypervariable region of a sample separately. This is achieved by finding a haplotype with the minimal genetic distance to the sequence in question from a flexible background dataset containing control region sequences, whose haplogroup-affiliation was confirmed by coding region SNPs. The distance is calculated applying a combination of the simple Hamming-distance for sites with gaps (insertions/deletions) with the more complex GTR (general time reversible; [34]) distance for sites without gaps. Then, the implementation checks whether the hg-assignments of the two hypervariable regions are conclusive—if not, it is indicated to re-examine the mismatching profile. In this study,

the mtDNA haplotypes were generated by sequence analysis of the full control region amplicons. Therefore, the risk of mixing up sample was significantly reduced. Nevertheless, we tested the haplotypes for artificial recombination, which could have potentially been introduced at a later stage of the analysis.

The second method applied to the data was a mathematical algorithm developed to aid the determination of the most probable haplogroup(s) of a set of mtDNA control region samples. The algorithm is implemented into a software package (manuscript in preparation) and is based on propositional logic via checking of the presence or absence of haplogroup diagnostic sites.

3. Results and discussion

We here present an optimized laboratory strategy for the sequence analysis of the complete mitochondrial DNA control region, which involves semi-automated 96 well-based pipetting and LIMS- and IT-aided sample processing. A sample of 273 unrelated persons from Austria serves as example (Table 2). This strategy developed for high-throughput typing of mtDNA control region sequences has proven to be a reliable and easily reproducible procedure for generating high-quality population data. The concept of keeping all laboratory stages in the 96-well format facilitates the handling of the samples and minimizes the possibility of artificial recombination (sample mix-up of hypervariable segments from different individuals [35,3,7]) considerably. Another important step in the avoidance of sample mix-up is that the entire control region is amplified in one piece, which reduces the complexity of post-PCR purification treatment and the further transfer of PCR products into the different cycle sequencing cocktails. The choice of the number of PCR cycles (35) is based on the intention to amplify the samples until the plateau phase of PCR is reached, largely independent of their initial DNA concentrations. This normalizes the amount of DNA that is finally added to the cycle sequencing reaction. Also in the post-PCR area, the application of automated pipetting workstations for transferring PCR-products into the different cycle sequencing cocktails, diluting the cycle sequencing products and transferring the diluted cycle sequencing products onto the Sephadex columns in the Multiscreen filter plate, enables a fast and reliable processing of a large number of samples by significantly reducing the number of manual pipetting steps. The sequencing reactions produce widely overlapping DNA sequences, generating overlapping redundant sequence information across the entire mitochondrial control region. This has proven very useful for the determination of point heteroplasmic positions as these were usually confirmed multiple times by independent sequence strands. Even in the case of length heteroplasmy occurring in any of the three C-stretches within the control region, at least full double strand coverage of all nucleotide positions was given with the 10 sequencing primers (Table 1(a)). The major benefit of this strategy is to reduce the need for repeated individual data handling due to length heteroplasmy or artifacts, which challenge sequence interpretation and require additional analyses. Some of the

Table 2
Control region sequences in Austria (haplogroup-assignments were confirmed with coding region SNPs [28])

Hg	Sample	HVS-I (16024–16569)								HVS-II (1–576)						
C	f1G6	16093C	16223T	16234T	16288C	16298C	16327T	16518T	16519C	16527T	73G	249del	263G	309.1C	315.1C	489C
H*	f1A5	16519C									235G	263G	309.1C	309.2C	315.1C	
H*	f1A6	16037G	16188A	16519C							263G	309.1C	309.2C	315.1C	524.1A	524.2C
H*	f1B3										263G	315.1C	340T	523del	524del	
H*	f1B6	16248T	16519C								73G	187T	263G	315.1C	523del	524del
H*	f1C1	16519C									189G	263G	315.1C			
H*	f1D1	16183C	16189C	16519C							263G	309.1C	309.2C	315.1C		
H*	f1E1	16293G	16311C								143A	195C	263G	309.1C	315.1C	
H*	f1E2	16291T	16519C								263G	315.1C				
H*	f1E3	16093C	16271C	16519C							263G	309.1C	315.1C			
H*	f1F1	16234T	16293G	16519C							263G	315.1C	315.2C			
H*	f1G2	16311C									263G	309.1C	315.1C	523del	524del	
H*	f1G3	16325C	16519C								263G	309.1C	315.1C			
H*	f1H2	16519C									153G	263G	309.1C	315.1C		
H*	f1H3	16311C	16519C								263G	309.1C	315.1C			
H*	f1H5	16266T	16311C	16399G	16519C						263G	315.1C	523del	524del		
H*	f2A4	16189C	16519C								263G	315.1C				
H*	f2B1	16189C	16519C								152C	263G	309.1C	315.1C	385G	
H*	f2B5	16189C	16519C								263G	315.1C				
H*	f2D1	16519C									146C	263G	309.1C	315.1C		
H*	f2D3	16072T	16245T	16311C	16519C						263G	315.1C	523del	524del		
H*	f2E4	16519C									263G	309.1C	315.1C			
H*	f2E5	16519C									263G	309.1C	315.1C			
H*	f2F3	16293G	16294T	16311C							143A	195C	263G	309.1C	315.1C	
H*	f2F5	16519C									152C	263G	309.1C	315.1C	334C	524.1A
H*	f2G1	16311C	16519C								152C	263G	315.1C		524.2C	
H*	f2G2										263G	309.1C	309.2C	315.1C	523del	524del
H*	f2H4	16172Y	16327T								263G	309.1C	315.1C	438T		
H*	f3A5	16519C									146C	263G	309.1C	315.1C		
H*	f3A6	16189C	16519C								152C	263G	309.1C	309.2C	315.1C	385G
H*	f3B1	16519C									152C	263G	315.1C	524.1A	524.2C	
H*	f3B4	16519C									259G	263G	315.1C			
H*	f3B5	16248T									146C	263G	309.1C	309.2C	315.1C	
H*	f3B6	16519C									263G	309.1C	309.2C	315.1C		
H*	f3C2	16209C	16519C								263G	315.1C				
H*	f3C3	16183C	16189C	16519C							146C	263G	309.1C	309.2C	315.1C	
H*	f3C4										263G	315.1C	523del	524del		
H*	f3D5										263G	309.1C	309.2C	315.1C		
H*	f3E1										309.1C	315.1C				
H*	f3E4	16519C									263G	309.1C	315.1C			
H*	f3F3	16093Y	16519C								263G	309.1C	315.1C	573.1-6CCCCC		
H*	m1A2	16093C	16519C								263G	315.1C				
H*	m1A4	16299G	16519C								150T	263G	315.1C			
H*	m1A5	16519C									263G	309.1C	309.2C	315.1C		
H*	m1B3	16239T	16355Y	16519C							263G	309.1C	315.1C			
H*	m1B4	16519C	16527T								146C	195C	263G	315.1C		
H*	m1D2	16234T	16293G	16519C							263G	315.1C	315.2C			
H*	m1D5	16129A	16311C	16519C							195C	263G	309.1C	315.1C		
H*	m1E2	16519C									263G	309.1C	315.1C			
H*	m1E3	16519C									263G	309.1C	315.1C			
H*	m1E5	16519C									263G	309.1C	315.1C			
H*	m1F1	16519C									146C	263G	309.1C	309.2C	315.1C	
H*	m1F2	16519C									152C	263G	315.1C			
H*	m1F3	16170G	16390A	16519C							263G	309.1C	315.1C			
H*	m1F5	16327T									263G	309.1C	315.1C	438T		
H*	m1H1	16291T									193G	263G	315.1C			
H*	m2B2	16042A	16288C	16519C							263G	315.1C	477C			
H*	m2B3	16261T	16291T	16311C	16519C						200G	263G	309.1C	315.1C		
H*	m2B6	16519C									263G	309.1C	315.1C			
H*	m2C2	16124C	16519C								263G	309.1C	315.1C	573.1-3CCC		

Table 2 (Continued)

Hg	HVS-I (16024–16569)		HVS-II (1–576)	
	Sample	16024–16569	1–576	16024–16569
U5a1	m1D6	16192T 16249C 16256T	73G 263G 315.1C	16270T 16399G 16519C
U5a1	m2G2	16192T 16249C 16256T	73G 263G 315.1C	16270T 16399G 16519C
U5a1	m2G5	16256T 16270T 16399G	73G 263G 309.1C 315.1C	
U5a1	m3C3	16256T 16270T 16399G	73G 263G 315.1C 524.1A	524.2C
U5b	f1B5	16189C 16192T 16270T	73G 150T 263G 315.1C	
U5b	f1D3	16189C 16192T 16270T	73G 150T 263G 315.1C	
U5b	f2D2	16261T 16270T	73G 150T 189G 263G	
U5b	f3A4	16189C 16209C 16270T	73G 150T 185A 263G	524.1A 524.2C
U5b	m2D5	16189C 16270T 16278T	73G 150T 185A 263G	524.1A 524.2C
U5b	m2H5	16189C 16220R 16270T	73G 150T 185A 263G	524.1A 524.2C
U5b	m3D5	16183C 16189C 16270T	73G 150T 263G 309.1C	
U5b	m3D6	16189C 16192T 16270T	73G 150T 263G 309.1C	
U5b	m3F3	16129A 16189C 16270T	73G 150T 263G 309.1C	524.2C
W	f1F2	16223T 16292T 16295T	73G 119C 189G 195C	315.1C 263G
W	f2F1	16223T 16292T 16362C	73G 150T 189G 195C	315.1C 263G
W	m2C1	16093C 16192T 16223T	73G 143A 152C 189G	207A 263G
W	m2H1	16223T 16292T 16519C	73G 152Y 189G 194T	195C 204C
X2	m1C3	16092C 16189C 16223T	73G 153G 195C 263G	315.1C 263G
X2b	f1D2	16182C 16183C 16189C	73G 153G 195C 225A	207A 263G
X2c	m3D2	16182C 16183C 16189C	73G 153G 195C 225A	207A 263G
X2d	f1E5	16134T 16183C 16189C	73G 195C 263G 309.1C	315.1C 226C 315.1C

Variant positions from the rCRS are shown between 16024 and 16569 in HVS-I and 1 and 576 in HVS-II.

sequencing primers that we used in former mtDNA population studies [14] turned out to provide insufficient sequence quality and were replaced (L16196, L00318, L00403 and H16410; Table 1(b)).

To obtain a consistent and comprehensible nomenclature for all population samples generated for the EMPOP database, a sample naming procedure was conceived that takes the geographical origin and the position of the DNA extract in the 96-well DNA masterplate into account. This nomenclature enables an easy and straight-forward post-laboratory handling and organization of population data. Beginning with the creation of sample sheets for electrophoresis, where for new population data only the city and plate identifiers have to be replaced (which can automatically be done by e.g. MS Excel), also the further processing of the data (i.e. import into the sequence alignment software and the final import into the EMPOP database) can be organized consistently and unambiguously. So, for generating further population data, we recommend the following nomenclature to be applied:

- The first two letters identify the geographical origin. For example, if population data from Vienna were generated, the first two letters would, e.g. be “Vi”.
- Following the city identifier, the plate identifier is added (1, 2, ...).
- Lastly, the 96-well plate position of the DNA extract is appended (e.g. “A1”).

The sequencing reactions deriving from different sequencing primers are attached after the sample name and separated by an underscore. The full name of a sequence electropherogram obtained from e.g. sample Vi1A1 generated with the sequencing primer L15971 would thus be: “Vi1A1_L15971.ab1”. The file extension “.ab1” is automatically appended to file names by AB capillary electrophoresis instruments.

3.1. Polymorphisms in the entire mtDNA control region

When taking dominant length variants in C-stretches into consideration, sequence comparisons led to the identification of 222 mitochondrial lineages as defined by 189 variable sites. On average, the samples showed 8.9 (95% CI 8.46–9.35) differences to the rCRS [15,16]. The mean pairwise difference between individuals was 9.32 (95% CI 9.28–9.36). Within the Austrian sequences, the ratio of sites manifesting transition mutations to those manifesting transversion mutations is 13.3.

Point heteroplasmy was reproducibly detected in 18 out of 273 samples (6.6%)—f1C4: 16266Y; f1C6: 16278Y; f2B6: 152Y; f2C3: 16519Y; f2H4: 16172Y; f3F3: 16093Y; f3G1: 16311Y; f3G6: 16086Y and 16168Y; m1B3: 16355Y; m1B5: 152Y; m1C6: 16093Y; m2E5: 16232S; m2H1: 152Y; m2H5: 16220R; m3B2: 234R; m3B5: 16301Y; m3G6: 16216R; m3H3: 16362Y.

We targeted the entire control region for this database to permit access to additional discriminatory variation that resides outside of HVS-I/HVS-II in the control region [31,36,37]. In the presented Austrian database, the entire control region discriminates 15 additional haplotypes compared to HVS-I/

HVS-II sequencing (206 vs 191 different haplotypes) disregarding cytosine insertions in the C-stretches (in HVS-I, HVS-II and HVS-III). The additional power of discrimination is not the only reason for targeting the mini-variable regions outside the control region, many phylogenetically informative sites also lie outside the commonly typed range 16024–16365 and 73–340. Haplogroup HV0 (former pre-V), for example, is characterized by the polymorphisms T16298C and T72C [38]. Many laboratories that amplify HVS-I and HVS-II separately start with the analysis of HVS-II at position 73, although position 72 has been sequenced and is clearly readable. Thus, when only HVS-II is typed from a sample sequence evaluation should start at least at position 72. Similar observations were made with other haplogroups: In this study, we confirm that U2e is among other diagnostic sites characterized by A508G, U3a by G16390A, U4 by G499A, U5a1 by A16399G, H1c by 477C, H5 by C456T, K1a by 497T, J by T489C, J1 by C462T and J1c1 by T482C (Table 2) [25,26].

Several years before, we [39] published an mtDNA population study on 101 Austrian Caucasians. In order to avoid redundancy we randomly selected from an independent set of samples. However, as identified during data analysis, we unintentionally included a sample in the new set that has also been typed in the first study (AUT69 from [39] is identical to m3G3). This is the mtDNA haplotype of a woman, who has accidentally been sampled twice within the past 5 years. Nevertheless, we decided to maintain this sample in the set of this study.

3.2. Random match probability (RMP)

We have targeted the entire control region for this dataset to access additional discriminatory variation that resides outside of HVS-I/HVS-II [36,37]. The probability of a random match between two unrelated individuals from this Austrian dataset ($n = 273$) was calculated 1:89 for HVS-I + HVS-II and 1:125 for the entire CR (consistently disregarding C-insertions in HVS-I, HVS-II and HVS-III). The latter is comparable to RMPs computed for the entire control region databases from the Czech Republic (1:83; value from [30]) and Germany (1:96; [31]; RMP calculated from the data). The slight observed differences may confirm the assumption that the number of haplotypes increases with the sample size [40]. In addition, the observed decrease of discrimination power when only HVS-I + HVS-II is analyzed was confirmed by the results from RMP

analysis of hypervariable regions I and II of the Czech (RMP = 1:67) and German (RMP = 1:66) population samples. The probability of a chance match (HVS-I + HVS-II) was calculated 1:59 for Bosnia-Herzegovina, 1:76 for Poland and 1:56 for Slovenia.

3.3. Comparison with other West-Eurasian populations

In order to evaluate the diversity observed in Austria in relation to other West-Eurasian mtDNA control region variation, we compared the present Austrian sample set to databases of other West-Eurasian populations (Bosnia-Herzegovina, Czech Republic, Germany, Poland, and Slovenia). Only HVS-I and HVS-II data were considered, since not all databases included entire control region sequences. Pairwise comparisons showed a considerable number of matches between Austria and other European populations: 121 individuals (54 haplotypes) from Austria were also found in the databases from Bosnia-Herzegovina, the Czech Republic, Germany, Poland, and Slovenia. In particular, the most common haplotype in the Austrian population sample (263G, 315.1C) was also the most frequent profile in the other European populations (Austria: 7.7%, Bosnia-Herzegovina: 8.3%, Czech Republic: 8.6%, Germany: 9.5%, Poland: 9.0%, Slovenia: 7.8% of the population sample, respectively). On the other hand, the majority of sequences in the particular databases have not been observed in other databases (Austria: 55.7%, Bosnia-Herzegovina: 55.1%, Czech Republic: 76.3%, Germany: 57.5%, Poland: 62.3%, Slovenia: 52.4% of the population sample, respectively).

Pairwise differences between and within populations were calculated with ARLEQUIN (Table 3). Whereas the populations from Austria, Bosnia-Herzegovina, Germany, Poland and Slovenia show on average 7.64 pairwise differences within their populations and 7.66 pairwise differences between the populations, the population sample from the Czech Republic shows 8.89 pairwise differences within its population and 8.29 differences to the other European populations.

AMOVA was used to test for significant variation in the mtDNA distributions among the various populations (Table 4(a)). 99.7% of the variance observed among the six populations is attributable to differences within populations, and 0.3% ($p_{\alpha} < 0.05$) represents differences among populations. The comparison of F_{ST} values from pairs of population samples (Table 4(b) and (c)) revealed that the Austrian

Table 3
Population average pairwise differences (16024–16368 and 71–340)

	Austria	Bosnia-Herzegovina	Czech Republic	Germany	Poland	Slovenia
Austria	7.81507	7.59969	8.36260	7.84875	7.78091	7.69467
Bosnia-Herzegovina	0.02895	7.32642	8.14344	7.59919	7.52743	7.41603
Czech Republic	0.00976	0.03493	8.89060	8.40284	8.29975	8.22483
Germany	0.01856	0.01334	0.03489	7.84530	7.77762	7.71257
Poland	0.03558	0.02643	0.01665	0.01718	7.67559	7.60502
Slovenia	0.02732	−0.00699	0.01971	0.03010	0.00742	7.51963

Above diagonal: average number of pairwise differences between populations ($PiXY$); diagonal elements: average number of pairwise differences within population (PiX); below diagonal: corrected average pairwise difference ($PiXY - (PiX + PiY)/2$).

Table 4
AMOVA results

Source of variation	d.f.	Sum of squares	Variance components			Percent of variation
(a) Design and results (d.f. stands for degrees of freedom)						
Among populations	5	29.289	0.01136 Va			0.29
Within populations	1064	4136.307	3.89116 Vb			99.71
Total	1069	4165.595	3.90253			
	Austria	Bosnia-Herzegovina	Czech Republic	Germany	Poland	Slovenia
(b) Population pairwise FSTs						
Austria	0.00000					
Bosnia-Herzegovina	0.00369	0.00000				
Czech Republic	0.00168	0.00482	0.00000			
Germany	0.00237	0.00170	0.00462	0.00000		
Poland	0.00457	0.00344	0.00258	0.00222	0.00000	
Slovenia	0.00340	−0.00090	0.00249	0.00378	0.00091	0.00000
	Austria	Bosnia-Herzegovina	Czech Republic	Germany	Poland	Slovenia
(c) FST <i>P</i> values ^a						
Austria	*					
Bosnia-Herzegovina	0.02441	*				
Czech Republic	0.17773	0.02344	*			
Germany	0.05176	0.11328	0.02441	*		
Poland	0.00391	0.02734	0.08789	0.06348	*	
Slovenia	0.06445	0.60645	0.14844	0.04492	0.25391	*

^a Significant FST *P* values are depicted in bold.

population sample displayed significant differences in mtDNA distributions compared to the population databases from Bosnia-Herzegovina and Poland. Additionally, the Bosnian database showed statistically significant differences to the population samples from Poland and the Czech Republic, which also displayed differences to the German database. Lastly, the German sequences also showed differences in population substructure to sequences from Slovenia. The slight differences in mtDNA substructure between the different populations might be explained by the relatively small sizes of the samples compared to the sizes of the populations they were drawn from. The small sample sizes might thus represent faintly different cutouts of the distribution of the mitochondrial haplogroups. However, the inter-population variability calculated for these European populations is in agreement with former estimates from other European populations [41] and is low compared to sub-Saharan African populations [42,13].

3.4. Haplogroup assignment

Much of the evolutionary change in the mtDNA pools of West-Eurasian populations during the past centuries has been introduced by lineage redistribution—both within populations, caused by drift, and between populations, caused by migration [43]. The basic phylogenetic structure of West-Eurasian mtDNA lineages has been revealed by a number of recent studies [44,45,18,46].

The analysis of 16 diagnostic mtDNA coding region SNPs allowed us to assign the major haplogroup status of the CR sequences [28], which were in agreement with the expected haplogroup distribution from West-Eurasian populations (H, I, J, K, N1a, T, U, V, W, and X) [46–50,28,23–25].

The most common haplogroup, cluster HV (including samples belonging to haplogroups HV0 (former pre-V), H* and sub-haplogroups of H), was observed in 45.4% of the Austrian sample set. Samples belonging to haplogroup H, which could not be resolved genealogically in subclades [23,24,51], were assigned to haplogroup H*. Haplogroups observed at intermediate levels included clusters U (19.8%), T (13.2%), J (8.4%), and K (8.4%). The haplogroups observed less frequently included I (0.7%), N1a (0.7%), W (1.5%), and X (1.5%). One sample (f1G6) that could not be unambiguously assigned by the 16 coding region SNPs was identified by control region polymorphisms to belong to the Asian haplogroup C (0.4%; Table 2). Two samples that were assigned to haplogroup I by the 16 coding region SNPs were identified to belong to haplogroup N1a by control region markers. This initial misclassification can be explained by the fact that by the time of the SNP study [28], the marker G1719A was used to classify haplogroups I and X [47,45,50]. According to a later study however [25], the marker G1719A turned out as a deep-rooting marker defining the two branches N1 and N5, where haplogroup I is a only sub-lineage of haplogroup N1. A refined classification into nested sub-haplogroups could be reached for haplogroups I, J, K, T, U and V. Due to the lack of a detailed description of haplogroups W and X, samples belonging to these haplogroups could not be further subdivided.

3.5. Phylogenetic analyses

The phylogenetic check for artificial recombination applied here provides a fast method for identifying samples containing DNA-stretches from two different persons. Such a quality check is especially valuable for mtDNA databases, where the

two hypervariable regions have been amplified and sequenced separately. Our approach to sequence generation involves the amplification of the entire control region in one step and the generation of redundant sequence information by sequencing overlapping DNA fragments. Thus, the occurrence of artificial recombination of DNA stretches deriving from different persons can be excluded. However, a byproduct of this phylogenetic check for artificial recombination is that all samples are assigned with a mitochondrial haplogroup. This phylogenetic haplogroup determination procedure again is approved and completed by the results of the haplogroup determination algorithm based on propositional logic. The combination of the two programs compensates for the weakness of each single program: the phylogenetic inference fails if the background dataset does not contain a sequence of similar ethnic origin to the sample in question; the logic approach does not take the different mutation rates of the individual diagnostic positions into full account and might thus lead the user to the wrong tip of the mitochondrial tree. Thus, both approaches can facilitate the assignment of samples to mitochondrial haplogroups; however, the final decision needs to be based on a profound understanding of the human mitochondrial genome, its mutation rates and evolutionary patterns.

Acknowledgements

We thank Verena Lubei, Anna König and Dr. Cordula Eichmann (Institute of Legal Medicine, Innsbruck Medical University, Austria) for help with sequence evaluation.

References

- [1] M.M. Holland, T.J. Parsons, Mitochondrial DNA sequences analysis—validation and use for forensic casework, *Forensic Sci. Rev.* 11 (1) (1999) 21–49.
- [2] 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.
- [3] H.-J. Bandelt, A. Salas, S. Lutz-Bonengel, Artificial recombination in forensic mtDNA population databases, *Int. J. Legal Med.* 118 (5) (2004) 267–273.
- [4] Y.-G. Yao, C.M. Bravi, H.-J. Bandelt, A call for mtDNA data quality control in forensic science, *Forensic Sci. Int.* 141 (1) (2004) 1–6.
- [5] A. Carracedo, W. Bär, P. Lincoln, W. Mayr, N. Morling, B. Olaisen, et al., DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing, *Forensic Sci. Int.* 110 (2) (2000) 79–85.
- [6] G. Tully, W. Bär, B. Brinkmann, A. Carracedo, P. Gill, N. Morling, et al., Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles, *Forensic Sci. Int.* 124 (1) (2001) 83–91.
- [7] W. Parson, A. Brandstätter, A. Alonso, N. Brandt, B. Brinkmann, A. Carracedo, et al., The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation, results and perspectives, *Forensic Sci. Int.* 139 (2–3) (2004) 215–226.
- [8] A. Salas, L. Prieto, M. Montesino, C. Albarrán, E. Arroyo, M.R. Paredes-Herrera, et al., Mitochondrial DNA error prophylaxis: assessing the causes of errors in the GEP’02-03 proficiency testing trial, *Forensic Sci. Int.* 148 (2–3) (2005) 191–198.
- [9] W. Parson, A. Brandstätter, M. Pircher, M. Steinlechner, R. Scheithauer, EMPOP—the EDNAP mtDNA population database concept for a new generation, high-quality mtDNA database, *Int. Congr. Ser.* 1261 (2004) 106–108.
- [10] E. Ambach, W. Parson, H. Niederstätter, B. Budowle, Austrian Caucasian population data for the quadruplex plus amelogenin: refined mutation rate for HumvWFA31/A, *J. Forensic Sci.* 42 (6) (1997) 1136–1139.
- [11] W.J. Jones, J.A. Irwin-Ross, F.A. Love, A.B. Welsh, M.M. Holland, T.J. Parsons, Development of an efficient, high-throughput strategy for sequence analysis of the entire human mitochondrial DNA control region. in: Poster, 10th International Symposium on Human Identification, Promega Corp., Orlando, FL, September 27–October 1, 1999.
- [12] E.T. Richon, N. Abassi, A. Coute, T.P. McMahon, J.A. Irwin, S.M. Barritt, et al., Validation of the Tecan Genesis robotic sample processor for automated cycle sequencing of mtDNA database samples, in: Poster, 14th International Symposium on Human Identification, Promega Corp., Phoenix, AZ, September 28–October 3, 2003.
- [13] 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 (5) (2004) 294–306.
- [14] A. Brandstätter, H. Niederstätter, W. Parson, Monitoring the inheritance of heteroplasmy by computer-assisted detection of mixed basecalls in the entire human mitochondrial DNA control region, *Int. J. Legal Med.* 118 (1) (2004) 47–54.
- [15] S. Anderson, A.T. Bankier, B.G. Barrell, M.H. de Bruijn, A.R. Coulson, J. Drouin, et al., Sequence and organization of the human mitochondrial genome, *Nature* 290 (5806) (1981) 457–465.
- [16] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, *Nat. Genet.* 23 (2) (1999) 147.
- [17] 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 (3) (2001) 252–255.
- [18] A. Helgason, a.S. Sigureth, J.R. Gulcher, R. Ward, K. Stefansson, mtDNA and the origin of the Icelanders: deciphering signals of recent population history, *Am. J. Hum. Genet.* 66 (3) (2000) 999–1016.
- [19] B.A. Malyarchuk, T. Grzybowski, M.V. Derenko, J. Czarny, M. Wozniak, D. Miscicka-Sliwka, Mitochondrial DNA variability in Poles and Russians, *Ann. Hum. Genet.* 66 (Pt. 4) (2002) 261–283.
- [20] M.V. Derenko, T. Grzybowski, B.A. Malyarchuk, I.K. Dambueva, G.A. Denisova, J. Czarny, et al., Diversity of mitochondrial DNA lineages in South Siberia, *Ann. Hum. Genet.* 67 (5) (2003) 391–411.
- [21] M.V. Derenko, B.A. Malyarchuk, I.K. Dambueva, I.A. Zakharov, Structure and diversity of the mitochondrial gene pools of south Siberians, *Dokl. Biol. Sci.* 393 (2003) 557–561.
- [22] B.A. Malyarchuk, T. Grzybowski, M.V. Derenko, J. Czarny, K. Drobniec, D. Miscicka-Sliwka, Mitochondrial DNA variability in Bosnians and Slovenians, *Ann. Hum. Genet.* 67 (5) (2003) 412–425.
- [23] A. Achilli, C. Rengo, C. Magri, V. Battaglia, A. Olivieri, R. Scozzari, et al., 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 (5) (2004) 910–918.
- [24] E.-L. Loogväli, U. Roostalu, B.A. Malyarchuk, M.V. Derenko, T. Kivisild, E. Metspalu, et al., Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia, *Mol. Biol. Evol.* 21 (11) (2004) 2012–2021.
- [25] M.G. Palanichamy, C. Sun, S. Agrawal, H.-J. Bandelt, Q.-P. Kong, F. Khan, et al., Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia, *Am. J. Hum. Genet.* 75 (6) (2004) 966–978.
- [26] A. Achilli, C. Rengo, V. Battaglia, M. Pala, A. Olivieri, S. Fornarino, et al., Saami and berbers—an unexpected mitochondrial DNA link, *Am. J. Hum. Genet.* 76 (5) (2005) 883–886.
- [27] H.-J. Bandelt, A. Achilli, Q.-P. Kong, A. Salas, S. Lutz-Bonengel, C. Sun, et al., Low “penetrance” of phylogenetic knowledge in mitochondrial disease studies, *Biochem. Biophys. Res. Commun.* 333 (1) (2005) 122–130.

- [28] A. Brandstätter, T.J. Parsons, W. Parson, Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups, *Int. J. Legal Med.* 117 (5) (2003) 291–298.
- [29] M. Stoneking, D. Hedgecock, R.G. Higuchi, L. Vigilant, H.A. Erlich, Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes, *Am. J. Hum. Genet.* 48 (2) (1991) 370–382.
- [30] T. Vanecek, F. Vorel, M. Sip, Mitochondrial DNA D-loop hypervariable regions: Czech population data, *Int. J. Legal Med.* 118 (1) (2004) 14–18.
- [31] S. Lutz, H.-J. Weisser, J. Heizmann, S. Pollak, Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany, *Int. J. Legal Med.* 111 (2) (1998) 67–77 (Errata in *Int. J. Legal Med.* 111 (1998) 286 and *Int. J. Legal Med.* 112 (1999) 145–150).
- [32] S. Schneider, D. Roessler, L. Excoffier, Arlequin ver. 2.0: a software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland, 2000.
- [33] H.-J. Bandelt, A. Salas, C. Bravi, Problems in FBI mtDNA database, *Science* 305 (5689) (2004) 1402–1404.
- [34] P.J. Waddell, M.A. Steel, General time-reversible distances with unequal rates across sites: mixing gamma and inverse Gaussian distributions with invariant sites, *Mol. Phylogenet. Evol.* 8 (3) (1997) 398–414.
- [35] 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.
- [36] S. Lutz, H.-J. Weisser, J. Heizmann, S. Pollak, A third hypervariable region in the human mitochondrial D-loop, *Hum. Genet.* 101 (3) (1997) 384.
- [37] M.D. Coble, R.S. Just, J.E. O’Callaghan, I.H. Letmanyi, C.T. Peterson, J.A. Irwin, T.J. Parsons, Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians, *Int. J. Legal Med.* 118 (3) (2004) 137–146.
- [38] A. Torroni, H.-J. Bandelt, V. Macaulay, M. Richards, F. Cruciani, C. Rengo, et al., A signal, from human mtDNA, of postglacial recolonization in Europe, *Am. J. Hum. Genet.* 69 (4) (2001) 844–852.
- [39] 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.
- [40] L. Pereira, C. Cunha, A. Amorim, Predicting sampling saturation of mtDNA haplotypes: an application to an enlarged Portuguese database, *Int. J. Legal Med.* 118 (3) (2004) 132–136.
- [41] T. Melton, M. Wilson, M. Batzer, M. Stoneking, Extent of heterogeneity in mitochondrial DNA of European populations, *J. Forensic Sci.* 42 (3) (1997) 437–446.
- [42] T. Melton, C. Ginther, G. Sensabaugh, H. Soodyall, M. Stoneking, Extent of heterogeneity in mitochondrial DNA of sub-Saharan African populations, *J. Forensic Sci.* 42 (4) (1997) 582–592.
- [43] A. Helgason, E. Hickey, S. Goodacre, V. Bosnes, K. Stefansson, R. Ward, B. Sykes, mtDNA and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry, *Am. J. Hum. Genet.* 68 (3) (2001) 723–737.
- [44] M. Richards, V. Macaulay, H.-J. Bandelt, B.C. Sykes, Phylogeography of mitochondrial DNA in western Europe, *Ann. Hum. Genet.* 62 (3) (1998) 241–260.
- [45] V. Macaulay, M. Richards, E. Hickey, E. Vega, F. Cruciani, V. Guida, et al., The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs, *Am. J. Hum. Genet.* 64 (1) (1999) 232–249.
- [46] S. Finnilä, M.S. Lehtonen, K. Majamaa, Phylogenetic network for European mtDNA, *Am. J. Hum. Genet.* 68 (6) (2001) 1475–1484.
- [47] A. Torroni, K. Huoponen, P. Francalacci, M. Petrozzi, L. Morelli, R. Scozzari, et al., Classification of European mtDNAs from an analysis of three European populations, *Genetics* 144 (4) (1996) 1835–1850.
- [48] D.C. Wallace, M.D. Brown, M.T. Lott, Mitochondrial DNA variation in human evolution and disease, *Gene* 238 (1) (1999) 211–230.
- [49] M.W. Allard, K. Miller, M. Wilson, K.L. Monson, B. Budowle, 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 (6) (2002) 1215–1223.
- [50] C. Herrnstadt, J.L. Elson, E. Fahy, G. Preston, D.M. Turnbull, C. Anderson, et al., Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups, *Am. J. Hum. Genet.* 70 (5) (2002) 1152–1171.
- [51] K. Tambets, S. Rootsi, T. Kivisild, H. Help, P. Serk, E.-L. Loogväli, et al., 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 (4) (2004) 661–682.