Discerning the "identical": unexpected mitogenome diversity behind the most common European mtDNA control region (D-loop) haplotype

Martin Bodner[1]*, Christina Strobl[1], Simone Nagl[1], Catarina Xavier[1], Gabriela Huber[1], Irene Cardinali[2], Hovirag Lancioni[2], Ornella Semino[3], Anna Olivieri[3], Francesca Gandini[3,4], Alessandro Achilli[3], Antonio Torroni[3], Walther Parson[1,5]

[1]Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria
[2]Dipartimento di Chimica, Biologia e Biotecnologie, University of Perugia, Perugia, Italy
[3]Dipartimento di Biologia e Biotecnologie "L. Spallanzani", University of Pavia, Pavia, Italy
[4]School of Applied Sciences, University of Huddersfield, UK
[5]Forensic Science Program, The Pennsylvania State University, PA, USA
***contact: martin.bodner@i-med.ac.at**

Introduction

The circular human mtDNA molecule, due to its specific characteristics as maternally inherited lineage marker present in a high copy number and stable against degradation, is a vital tool in forensic and population genetics. The utility of this marker depends on the variation present in a population. The output of investigations is often restricted, as most data only include (partial) sequence information from the non-coding ~1.1 kbp control region (CR) that displays highly condensed variation. However, many mtDNA lineages are poorly defined in CR with identical nucleotide variants observed on different haplogroup backgrounds across the mtDNA phylogeny.

In West Eurasia, the most prevalent mitochondrial lineage (haplogroup H, often referred to as "**Helena**") comprises ~40% in many populations [1-3]. Some samples thus share identical haplotypes. However, a matching CR does not imply that the linked remaining mitogenomes (~15.5 kbp) are also identical or belong to the same phylogenetic lineage (Figure 1). This is particularly the case for the **most common West Eurasian CR haplotype (263G 315.1C 16519C)** observed in numerous subclades within haplogroup H and some HV relatives [4,5]. Its frequency of 3-4% in any West Eurasian population and uniform dispersal limit the power of forensic mtDNA analyses [1].

MtDNA variation studies have now entered the final phase of phylogenetic refinement - the analysis of entire mitogenomes. Benchtop high-throughput massively parallel (or next generation) sequencing solutions allow obtaining complete mtDNA sequence information relatively easy. In a recent publication, we were able to realize the full potential of forensic mtDNA testing on 29 Italian "most common CR haplotype" samples previously considered identical - and unexpectedly revealed 28 different mtGenome sequences, making the **forensic geneticist's ultimate desire of discerning the "identical"** come true [6]. We here present results from further investigations on the degree of variation within the most common CR haplotype in an enlarged Italian sample set.

Material and Methods

We collected a pan-Italian sample of almost 300 "most common CR haplotype" individuals (sequenced for at least HVS-I and -II), 80 of which have currently been analyzed (Figure 2). Representing a screening of some 10,000 Italians [1], this study uniquely offers the possibility of investigating the forensically most important haplotype in the currently largest sample set available for a single country.

After verification of integrity and quantitation of mtDNA in a real-time quantitative assay [7], we amplified two overlapping ~8.5 kbp fragments covering the **entire mitogenome** [8] and applied **massively parallel sequencing (MPS)** on the Ion Torrent Personal Genome Machine using Ion PGM Sequencing 200 Kit v2 chemistry on an Ion 316 chip v2. We followed strict quality control according to a validation study of the PGM in forensic mtDNA sequencing [9]. Raw data were inspected twice using independent software, mirroring the gold standard in Sanger-type sequencing.

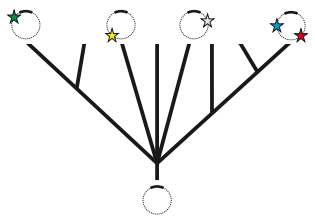

Figure 1: Schematic "phylotree" of haplogroup H from entire mitogenome sequencing. The root haplogroup H haplotype (at the bottom) is the origin of five (for the ~100) radiating lineages that give rise to further sub-branches. Four explicative derived H haplotypes are shown at tips of the tree. If only the mtDNA CR (continuous segments of the circles) were sequenced, all haplotypes in this figure would be identical and would fall into a single H* lineage. Their individual or lineage diagnostic coding region variation, indicated by asterisks in the dotted segments of the circular genomes, would not be detected (modified from [6]).

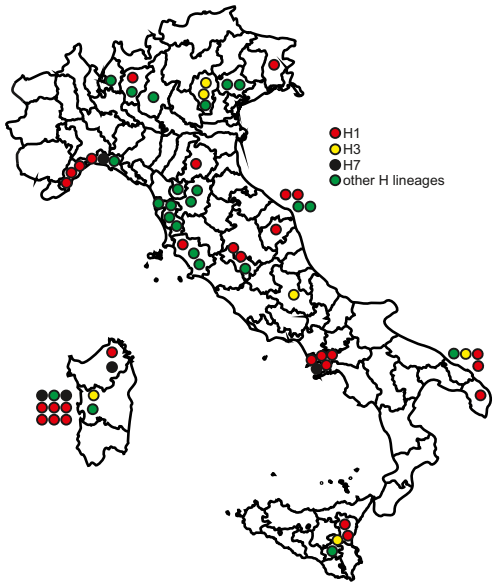

Figure 2: Origin of the 63 Italian samples included in this study. Circles represent single samples and are assigned to their province of origin. When the province was not known, circles were placed in the sea adjacent to their region of origin. The color codes distinguish haplogroup H1, H3, H7, and those falling into other H clades.

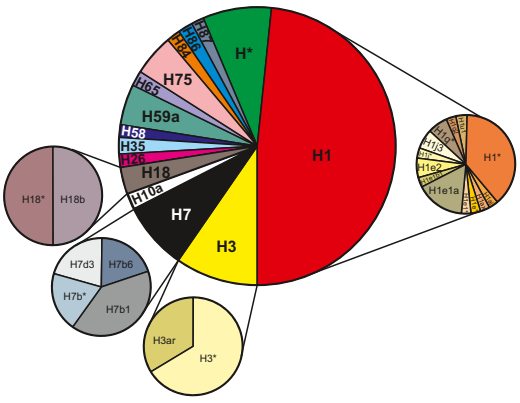

Figure 3: MtDNA diversity within the "most common CR haplotype" revealed by complete mitogenome sequencing. The pie chart indicates the proportions of the 31 described mtDNA clades (and unassigned H*) [4] found in the Italian sample (n=63). The peripheral pie charts depict the proportions of subclades found within the segment they are assigned to.

Results

From the 80 samples currently analyzed in this ongoing project, 17 revealed additional CR polymorphism outside HVS-I and II and were thus excluded. Complete mitogenome sequencing revealed 57 different haplotypes (thereof 52 unique) in the 63 remaining samples considered identical from CR (including those previously published [6]). One trio and four pairs exhibited identical mitogenomes. **The power of discrimination increased from 0.0%** (CR) **to 99.6%** (entire mitogenomes) (Table 1). The 63 samples classified into 31 described subclades of haplogroup H [4] (build 16), 21 of which were unique in the dataset. Five samples could not be assigned to a distinct H lineage and remained H* (Table 1; Figure 3). They possibly highlight yet undescribed lineages.

Table 1. Diversity parameters of the 63 Italian samples

n = 63	mtDNA range	
	CR	mitogenome
haplotypes	1	57
unique haplotypes	0	52
haplogroups[4]	1	32
unique haplogroups[4]	0	22
random match probability	1.000	0.019
haplotype diversity	0.0%	99.6%

Discussion and Outlook

This study clearly demonstrates the **benefit of complete mitogenome sequencing for forensic genetics**. An unexpectedly high coding region diversity rendered nearly every haplotype, previously considered identical from CR data, unique. Resolution using the entire mtDNA molecule only slightly decreased even though the number of samples was more than doubled compared to our initial study (29 samples, power of discrimination 99.8%, [6]) (Table 1).

The geographic distribution of mtDNA lineages necessarily appears rather random (Figure 2). Ongoing sequencing of the complete sample set (n~300) might reveal particular phylogeographic patterns. Thereby, the project will gain **significance beyond forensics** by shedding light on human migration and population genetic questions.

Ultimately, the knowledge on the distribution and proportion of "Helena's daughters" could result in a specific SNP panel for forensic mtDNA testing of Italian samples exhibiting the most common CR haplotype for use when complete mtGenome sequencing is not feasible or affordable, but increased resolution is needed for investigative purposes.

We used the PGM on an mtDNA population sample in a forensic environment as stand-alone approach. This study therefore contributes towards the **implementation of MPS** into high quality mtDNA typing routine on the way to the future of forensic mtDNA databases [10].

References

[1] **EMPOP mtDNA database (https://empop.online)**
[2] Brandstätter A et al. 2008. BMC Evol Biol. 8:19.
[3] Turchi C et al. 2008. Int J Legal Med. 122:199-204.
[4] van Oven M, Kayser M. 2009. Hum Mutat. 30:E386-94; www.phylotree.org
[5] Brandstätter A et al. 2006. Electrophoresis. 27:2541-50.
[6] Bodner M et al. 2015. Forensic Sci Int Genet. 15:21-26.
[7] Niederstätter H et al. 2007. Forensic Sci Int Genet. 1:29-34.
[8] Fendt L et al. 2009. BMC Genomics. 10:139.
[9] Parson W et al. 2013. Forensic Sci Int Genet. 7:543-549.
[10] Irwin JA et al. 2011. Forensic Sci Int Genet. 5:222-225.

Disclosure and Acknowledgements
The authors report no conflicts of interest. This work was supported by the intramural funding program of the Medical University Innsbruck for young scientists MUI-START, Project 2013042025, the *Theodor Körner Fonds zur Förderung von Wissenschaft und Kunst*, the *D. Swarovski Förderungsfonds* (DSF 2015-1-1) and the *Tiroler Wissenschaftsfonds* (TWF) (UNI-404/1998). Gefördert aus Mitteln des vom Land Tirol eingerichteten Wissenschaftsfonds.