Search, Align and Haplogroup – improved forensic mtDNA analysis via EMPOP

Arne Dür[1], Nicole Huber[2], Walther Parson[2,3]

[1]Institute of Mathematics, University of Innsbruck, Innsbruck, Austria
[2]Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria
[3]Forensic Science Program, The Pennsylvania State University, PA, USA

nicole.huber@i-med.ac.at

Introduction

Mitochondrial DNA (mtDNA) databases continue to grow and simultaneously the value of information that can be derived from an mtDNA profile is increased by different population and dispersal studies of humankind. However, especially for forensic purposes, quality must become more important than quantity. Clerical errors, inconsistent nomenclature or an insufficient sequence length (not covering the complete control region) are observed most frequently in submitted datasets, illustrating the absolute need of analytical software tools to maintain standards that are required for forensic databases [1].

So far, the only legally defensible database is EMPOP which, besides other tools, provides a string-based sequence search algorithm for mitochondrial DNA databases (SAM) [2].

In addition to the profile search, EMPOP now also provides a phylogenetic alignment of the queried sequence that is based upon Phylotree Build 17 and a confidential haplogroup estimation by use of the most recent common ancestor (MRCA).

Here, we describe the extension and refinement of the original algorithm with the aim of improving forensic mtDNA analysis. SAM 2 supports the interpretation of obtained results and promotes inter-laboratory collaboration, not only in the forensic field.

Materials & Methods

Extension of the basic algorithm
The fundamental basis of SAM 2 is formed by the 5435 haplogroup motifs of Phylotree Build 17. In order to evaluate the updated algorithm and to check the plausibility of the phylogenetic alignment as well as the haplogroup assignment, evaluation was carried out on the EMPOP database.

Alignment and Haplogrouping
Under consideration of the fluctuation rate model that was introduced by Röck et al. (2013), multiple realignment cycles followed by manual analysis and adaption of the fluctuation rates for affected positions were undertaken to approximate the human phylogeny. SAM's ability for haplogroup estimation was evaluated by comparison with haplogroup assignment based on EMMA, a concept for estimating mtDNA haplogroups [3].
SAM 2 with additional features will be provided for the scientific community via the EMPOP website (https://empop.online).

Conclusion & Outlook

The steady growth of mtDNA sequencing data emphasizes the importance of an appropriate data processing and management that encompasses quality control of sequences. SAM 2 demonstrated its reliability when it comes to database searches and the phylogenetic alignment allows for harmonization of mtDNA data obtained from different laboratories. The possibility of a confident haplogroup estimation extends the scope of application why the services provided by SAM and EMPOP may also be useful in other scientific areas.

Results & Discussion

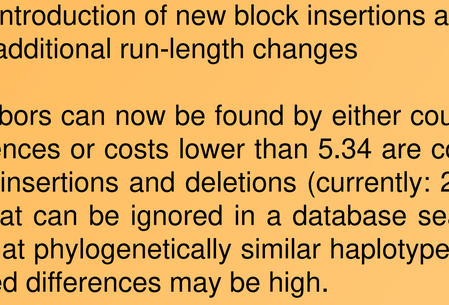
Extension of the basic algorithm
The implemented enhancements of the basic algorithm are:

- selectable mode of finding neighbors
- introduction of new block insertions and deletions
- additional run-length changes

Neighbors can now be found by either counts or costs where database profiles with up to 2 differences or costs lower than 5.34 are considered as neighbors. The introduction of more block insertions and deletions (currently: 26) as well as new run-length changes (currently: 12) that can be ignored in a database search improved the search strategy in view of the fact that phylogenetically similar haplotypes are recognized as such although the number of notated differences may be high.

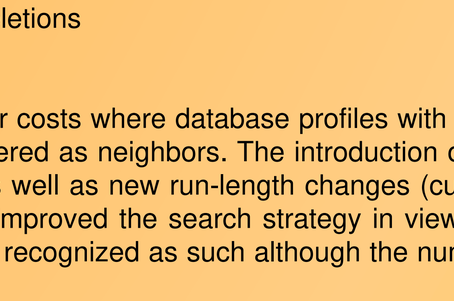
Database search
Regarding matches, an equal or higher number was observed for the default search parameter in EMPOP. In contrast, the number of neighbors was sometimes decreased in case of a higher match number, since some of the neighbors were now considered as a match. In case of equal matches, the number of neighbors was equal or higher.
However, results of database searches are reproducible between SAM and SAM 2 when the mode of finding neighbors was set to count, new introduced block insertions/deletions and run length changes are not considered. That indicates that SAM 2 includes all matches of SAM and a multitude of additional neighbors.

Alignment
The application of the introduced alignment approach on the EMPOP database, that currently holds 34,617 high quality mtDNA profiles, revealed 301 (0.87%) alignment changes in total. The majority of sequence changes was observed around position 310 and was a consequence of the new introduced phylogenetic variant for 310C, which is 309del 310C 315.1C.



Description	# of samples	Percentage
Total number of samples	34617	100%
number of unchanged samples	34316	99,13%
number of changed samples	301	0,87%

Figure 1 – Total Changes of Sequence Notation: Overview of changed samples



Description	# of samples	Percentage
Total number of changed sequences	301	100%
around position 60	23	7,64%
around position 310	198	65,78%
around position 450	21	6,98%
around position 960	3	1,00%
around position 16180	19	6,31%
around position 16190	33	10,96%
around position 16260	3	1,00%
around position 8270	1	0,33%

Figure 2 – Changes grouped by Region: In total, eight different regions were affected by sequence changes

Haplogrouping
The comparison of the haplogroup assignment between EMMA (2013) and SAM (2017) revealed 1,393 changes based on Phylotree Build 17. In 95.12% SAM's haplogroup estimation was finer (e.g. T2c1a instead of T2c1). In 4.67% SAM's haplogroup estimation was coarser and in 0.22% the haplotype was classified into another branch of Phylotree (e.g. B4'5 instead of E1a1a1).

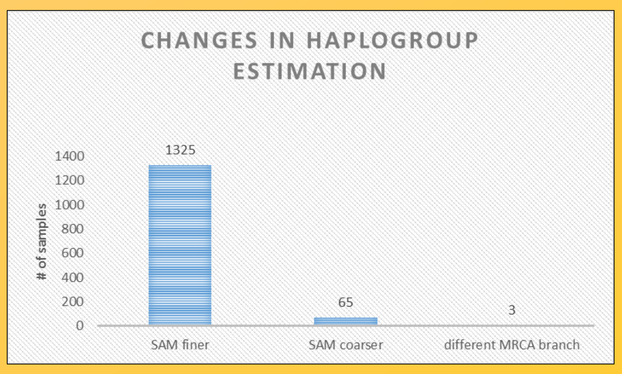


Figure 3 – Changes in haplogroup estimation SAM vs EMMA

References & Acknowledgements

[1] Bandelt, H-J. & Parson, W. (2008) IJLM; 122(1),11-21
[2] Röck, A. et al. (2011) FSIG; 5(2), 126-132
[3] Röck , A. et al. (2013) FSIG; 7(6), 601-609

We would like to thank the EMPOP Team at GMI, in particular Martin Bodner, Antonia Heidegger, Gabriela Huber, Martin Pircher, Lisa Schnaller, Christina Strobl, Stefan Troger, Catarina Xavier, and Bettina Zimmermann. vxWeb is acknowledged for programming and IT support.