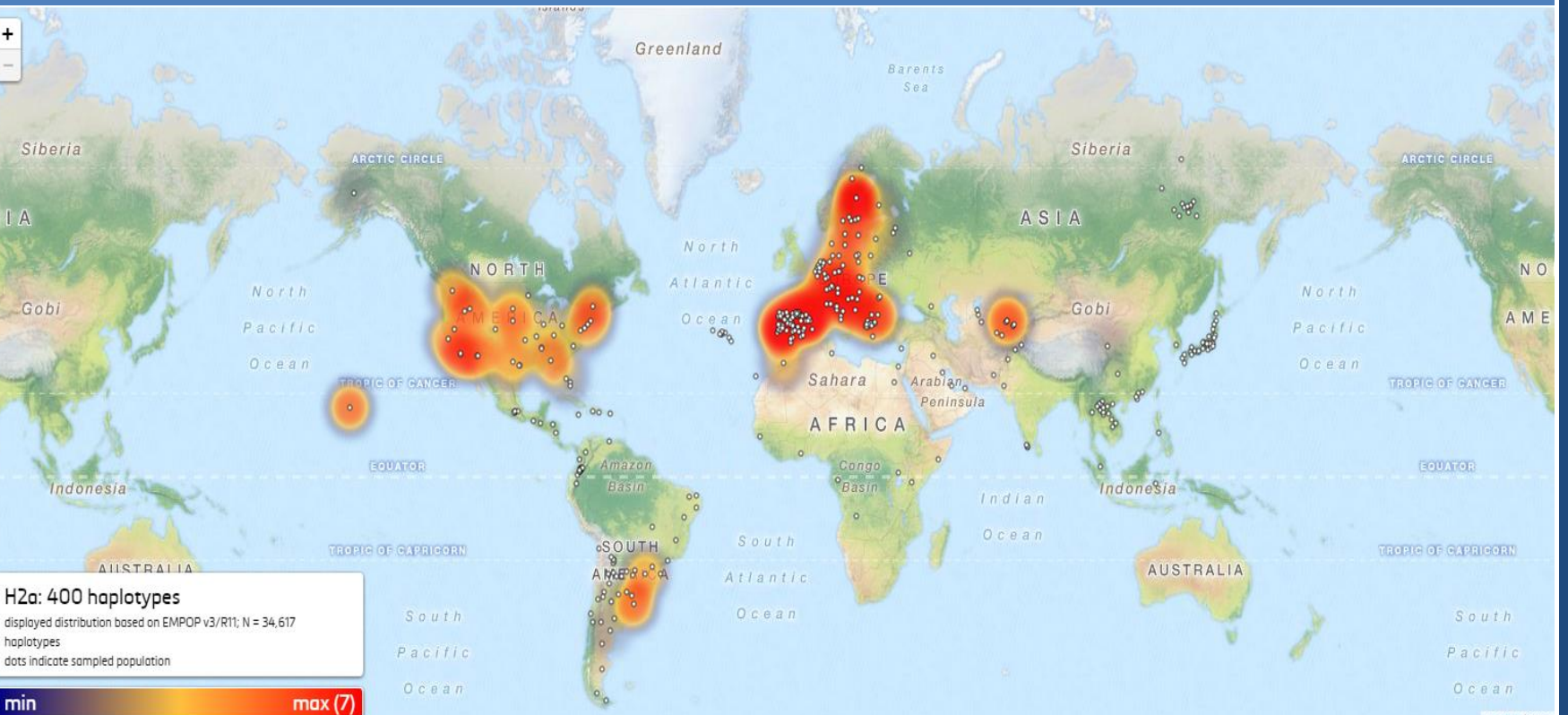
DIRECTIONS FOR USE

Version v4 Sep 2018



EMPOP mtDNA Database

EMPOP mtDNA Database – Directions for Use

Revision Overview

Version V4

- September 2018
  - Release of EMPOP 4
    - The new query engine SAM2 offers enhanced search functions, phylogenetic alignment and haplogrouping (see Huber et al 2018 for details)
    - Revision of several sections of the USE document
- December 2015
  - Section 4.3.2. – When no matches are found was updated
    - Tolerance value of EMMA was changed from „0.3" to „0.1"
- November 2015
  - Section 4.1.8. EMPOP haplogroup estimation – EMMA was added
- October 2015
  - Information about special positions was added in Section 4.1.2. Ranges
- July 2015
  - Revision Overview was added
  - Section 4.3.3. Ambiguous haplogroup estimates was added
- May2015
  - Initial Release of EMPOP mtDNA Database – Directions for Use

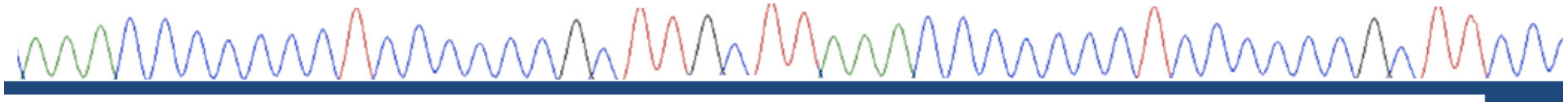
Table of Contents

How to use EMPOP

1. Introduction

The high copy number per cell, the stability against degradation and the maternal mode of inheritance make the mitochondrial (mt) genome particularly suitable for palaeo-, medical- and forensic genetic investigations. Its increased evolutionary rate led to sequence variation that has been generated by sequential accumulation of new mutations along radiating maternal lineages during human dispersal into different parts of the world.

Forensic molecular biology takes advantage of this variation for human identity testing by sequence analysis of the (hypervariable segments within the) mtDNA control region (CR). New developments in Massively Parallel Sequencing (MPS) demonstrated that also full mtGenome information can be obtained from even degraded forensic samples (Parson et al 2013, Eduardoff et al 2017, Strobl et al 2018). MtDNA analysis is a powerful tool to exclude samples as originating from the same individual/matriline. If two samples cannot be excluded the significance of the mtDNA match is assessed by making reference to the abundance of that particular mtDNA sequence (= mitotype) in a relevant population.



2. Concept

The EMPOP database aims at the collection, quality control and searchable presentation of mtDNA mitotypes from all over the world. EMPOP has carefully been envisioned and designed as high quality mtDNA database, where available primary sequence lane data are permanently linked to the database entries. The scientific concept and the quality control measures using logical and phylogenetic tools were found suitable for forensic purposes, e.g. by a declaration of the German Supreme Court of Justice (2010), the SWGDAM mtDNA interpretation guidelines (2013), and the updated ISFG guidelines for mtDNA analysis and interpretation (2014).

The scientific contents presented in EMPOP were developed by the Institute of Legal Medicine (GMI), Medical University of Innsbruck and the Institute of Mathematics, University of Innsbruck.
The mitotypes stored in EMPOP are not considered for partial or full download. The concept of data quality management requires a centralized supervision of the data. Necessary updates (e.g. haplogroup status, Release updates) will be introduced by the database curators to ensure continuous data quality and are made publicly available (see Release history).

3. Register/login

An EMPOP user is identified by the Email address to which account information (voluntary basis) and search history are connected. Follow the instructions for registration. An Email will be sent with a link that completes registration. Note that queries can be deleted in the account history by the user.

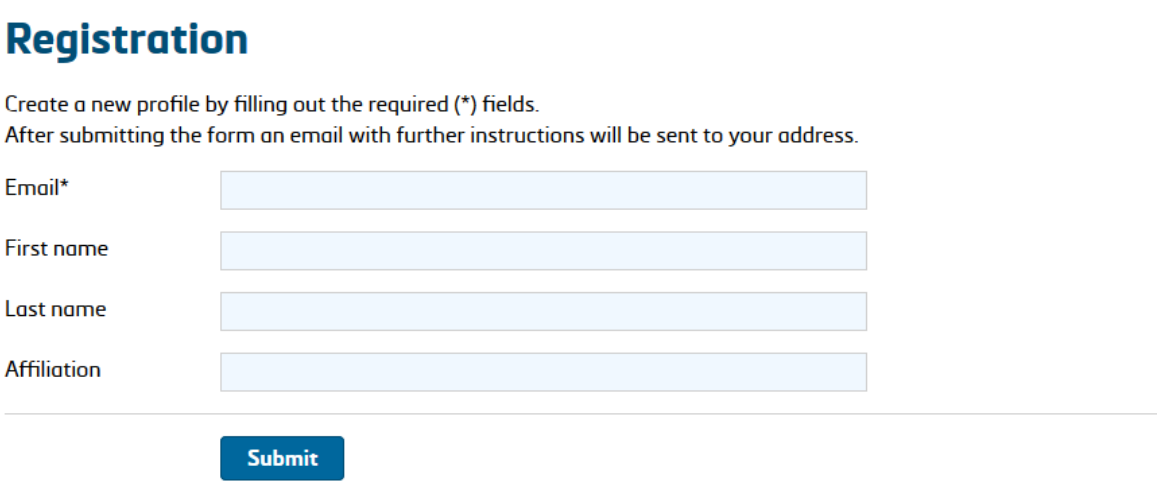


Figure 1 – User registration

4. Using EMPOP for mitotype searches

EMPOP follows the revised and extended guidelines for mitochondrial DNA typing issued by the DNA commission of the ISFG (Parson et al. 2014). See document for further details.

EMPOP's query engine - SAM 2:

EMPOP uses SAM 2, an updated and optimized software based on tests with carefully curated full mitogenome sequences to perform unbiased and conservative database queries to assist statistical evaluation of the evidence in forensic practice. The major changes to the earlier version of the software SAM (Röck et al 2011) include:

i. updated alignment/nomenclature conventions for the phylogenetically instable regions 50-70, 310-316, 455-460, 961-966, 8276-8279, 16180-16193, and 16258-16262

ii. 'count' and 'cost' search modes for neighbours

iii. implementation of 28 block indels containing between 2 and 264 base-pairs (Table 1)

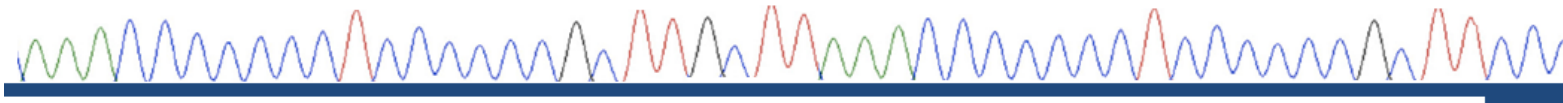
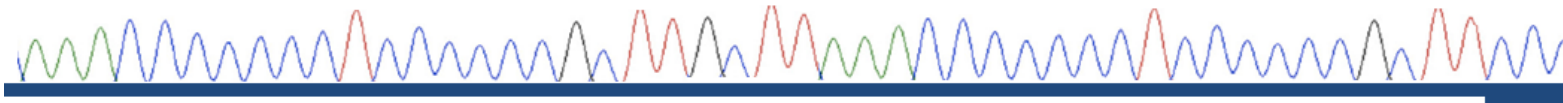


Table 1 - List of implemented block indels:

No.	Insertion position	Deletion positions	Pattern	Length [in bp]
1	16032	16032.1-16032.15	TCTCTGTTCTTTCAT	15
2	16164	16165-16318	AACCCAATCCACA...	154
3	16309	16310-16316	GTACATA	7
4	104	105-110	CGGAGC	6
5	105	106-111	GGAGCA	6
6	209	209.1-209.7	GTGTGTT	7
7	241	241.1-241.3	TAA	3
8	286	286.1-286.5	TAACA	5
9	290	291-294	ATTT	4
10	291	291.1-291.16	ACATCATAACAAA...	16
11	292	292.1-292.2	AT	2
12	292	292.2-292.4	AT	2
13	306	307-356	CCCTCCCCCCGCT...	51
14	309	310-315	TCCCCC	6
15	315	316-319.0	GCTT	4
16	342	342.1-342.2	AT	2
17	343	343.1-343.3	ATC	3
18	368	368.1-368.4	AGAA	4
19	398	398.1-398.14	ACCAGATTTCAAAT	14
20	470	471-478	TACTACTA	8
21	524	524.1-524.2	GC	2



22	563	563.1-563.204	AACAAAGAACCC...	204
23	588	568.1-588.16	CACAGTTTATGTA...	16
24	3326	3327-3590	ACTCCTCATTGTA...	264
25	6019	6020-6024	CGAGC	5
26	9486	9487-9501	TCGCAGGATTTTT...	15
27	14786	14787-14790	TTAA	4
28	16006	16006.1-16006.3	CTA	3

4.1. Query options

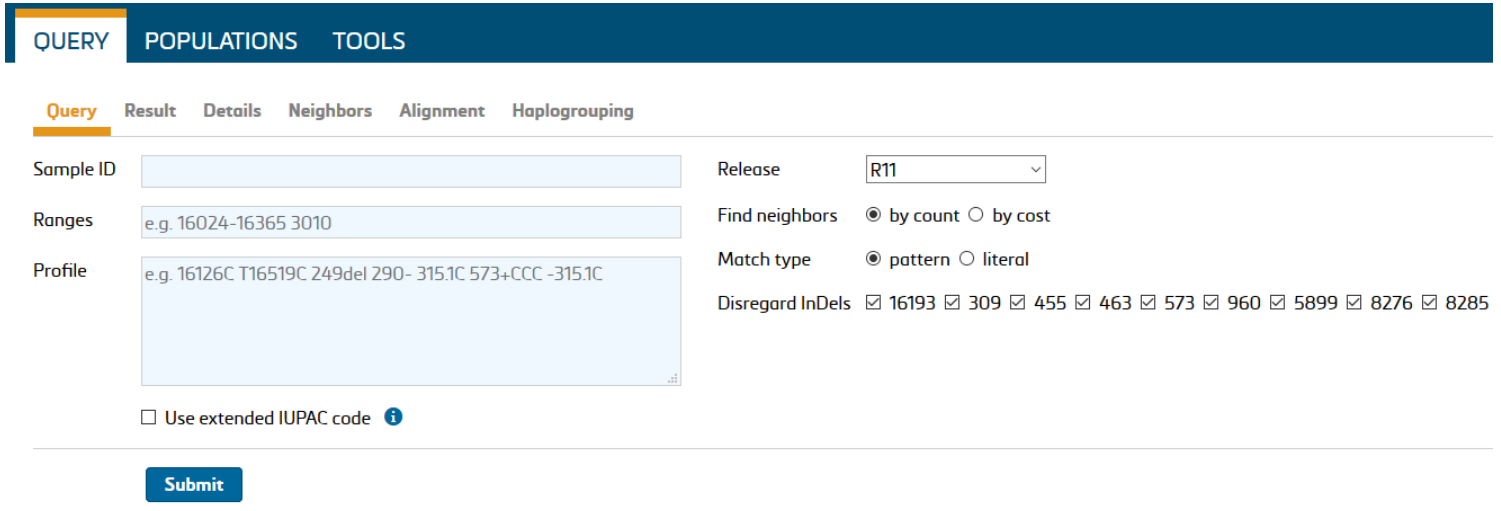


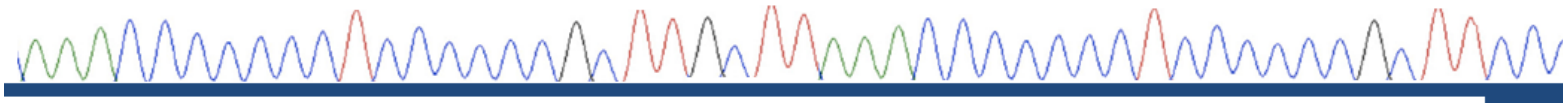
Figure 2 – Query input

4.1.1. Sample ID

Use this field to enter the ID of a mitotype. Search results are linked to this information and also provided on printouts. Sample IDs are used to identify queries in the search history of each individual user.

4.1.2. Ranges

Database queries require specification of the interpretation range(s) for rCRS-coded and FASTA-like string mitotypes. Typical ranges are: HVS-I (16024-16365), HVS-II (73-340), CR (16024-576), mitogenome (ALL). Individual SNPs can be queried by indicating the SNP in the range field, e.g. 3010. Note that EMPOP depends



on the sequence range provided by the submitting laboratory, which is why the ranges can be slightly different for some submissions/populations.

Examples:	
16024-16356 73-340	represents a standard range for a query in HVS-I and HVS-II.
16024-576	represents the control region range
16024-16365 489 3010	represents a query range including HVS-I and the two SNPs 489 and 3010. Note that an insertion between 489 and 490 would not be included in that query range.
ALL	represents the complete mitogenome

Note that there are special positions where a query does not make sense. For example, position 3107: A deletion at this position cannot be queried as 3107 serves as place holder in the rCRS to keep the original numbering system downstream of that position (the CRS mistakenly included that position).

4.1.3. Mitotype

Submit your mitotypes as FASTA-like sequence strings or reported relative to the rCRS.

Query sequence strings:

Copy&paste the sequence string from a text file or a consensus from sequence analysis software. Do not enter header information like in usual FASTA format; enter nucleotides only. For mixtures (e.g point heteroplasmy) use the extended IUPAC code (see below).

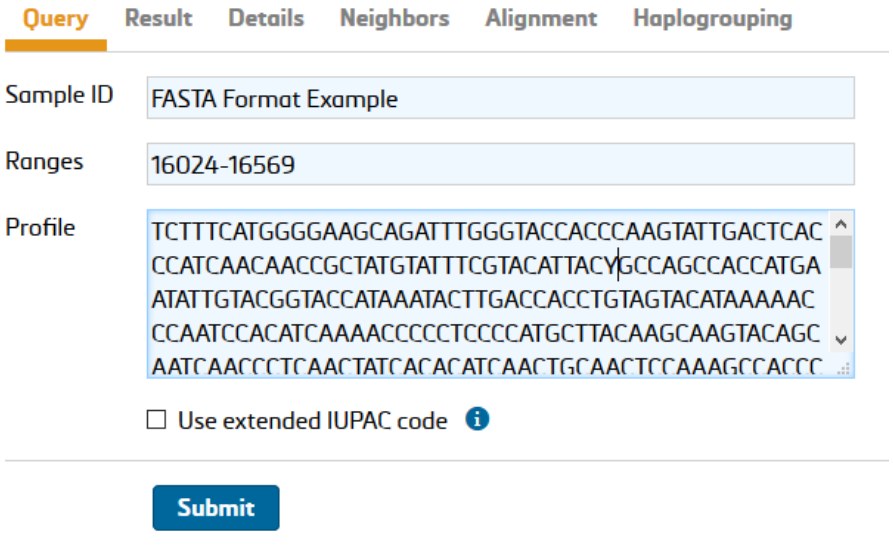


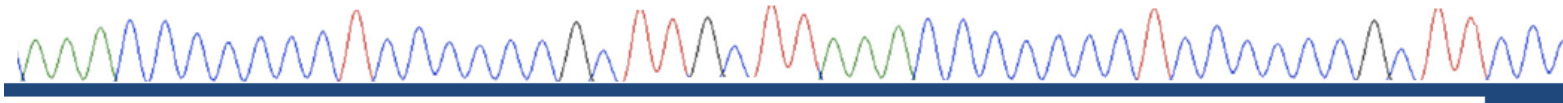
Figure 3 – Query input in FASTA-like format

Query rCRS aligned mitotypes:

Differences to the revised Cambridge Reference Sequence (rCRS, Andrews et al 1999) are entered as mitotypes.

Table 2 – Notation guidelines:

Type	Possible annotations	Comment
Base changes	73G, A73G	If preceding bases are used they must match rCRS base at the given position
Insertions	315.1C -315.1C 315+C 309.1C 309.2C 309+CC	For multiple insertions all preceding insertions need to be stated, i.e. annotating 309.2C is not possible without annotating 309.1C
Deletions	249- A249- 249delA 249del	'del' is treated case insensitive, e.g. Del, DEL, dEL, deL etc is accepted. Please note that the single character 'D' is considered a mixture of A, G, and T (IUB code). The single character 'd' is considered a mixture of A, G, T, and deletion (see Parson et al 2014 for details).



Note that the EMPOP query discerns capital letters (A,G, C, T, Y, ...) from uncapitalized letters (a, g, c, t, y, ...). Uncapizalized letters stand for a mixture of a deletion and a non-deleted variant. E.g. T152c represents two variants, T152C and T152del.

4.1.4. Release

EMPOP 4 offers release-specific queries. The most recent database release is selected by default. Earlier database releases can be selected if available.

4.1.5. Find neighbors

EMPOP offers searching for neighbors by count and by cost. Under current settings EMPOP reports neighbors within a count of 2 differences or costs of 5.34 (see Huber et al 2018 for details).

Finding neighbors by count is the default setting for forensic frequency estimates.

4.1.6. Match type

This is relevant for the consideration of point heteroplasmy in both the query sequence as well as the database sequences.

Pattern match: mixture designations match its individual components (Y={C,T,Y}). Example: 152Y matches 152T and 152C.

Literal match: mixture designations are considered exclusive to all other nucleotide designations (Y={Y}). Example: 152Y matches only 152Y.

Pattern match is the default setting for forensic frequency estimates.

4.1.7. Disregard InDels

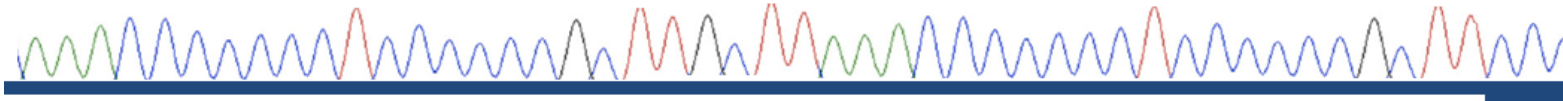
Length variants that are known hotspots for insertion/deletions (indels) should be ignored in a forensic database query. This involves the C-runs around positions 16193, 309, 463 and 573 and the T-run around position 455 relative to the rCRS in the control region. In the coding region length variants around positions 960, 5899, 8276 and 8285 are ignored for a forensic query (Table 2).

Table 3 - List of implemented length hotspots:

Length variant	5' junction	3' junction	Repeating motif
16193	16189	16194	C
309	302	316	C
315*	302	316	C
455	451	456	T
463	460	464	C
573	567	574	C
960	955	961	C
5899	5894	5900	C
8276	8271	8277	C
8285	8280	8286	C

* Note that the C-insertion between 310 and 316 is a stable length variant, not a length hotspot. It is listed here for uninterrupted C-runs as a consequence of T310C.

Standard query settings disregard discrepancies in hotspot length variant regions between query and database sequences.



Note that costs of disregarded InDels do not contribute to the final costs, which influences the ranking of results. See section 4.4. Neighbors.

4.2. Result

The execution of a database query automatically directs the user to the **Results** tab. Sample ID, query range(s) and mitotype are indicated in the top lines. Following information is listed in the results table:

1. number of observed matches in the entire database
2. number of observed matches sorted by geographic origin and
3. number of observed matches by metapopulation affiliation

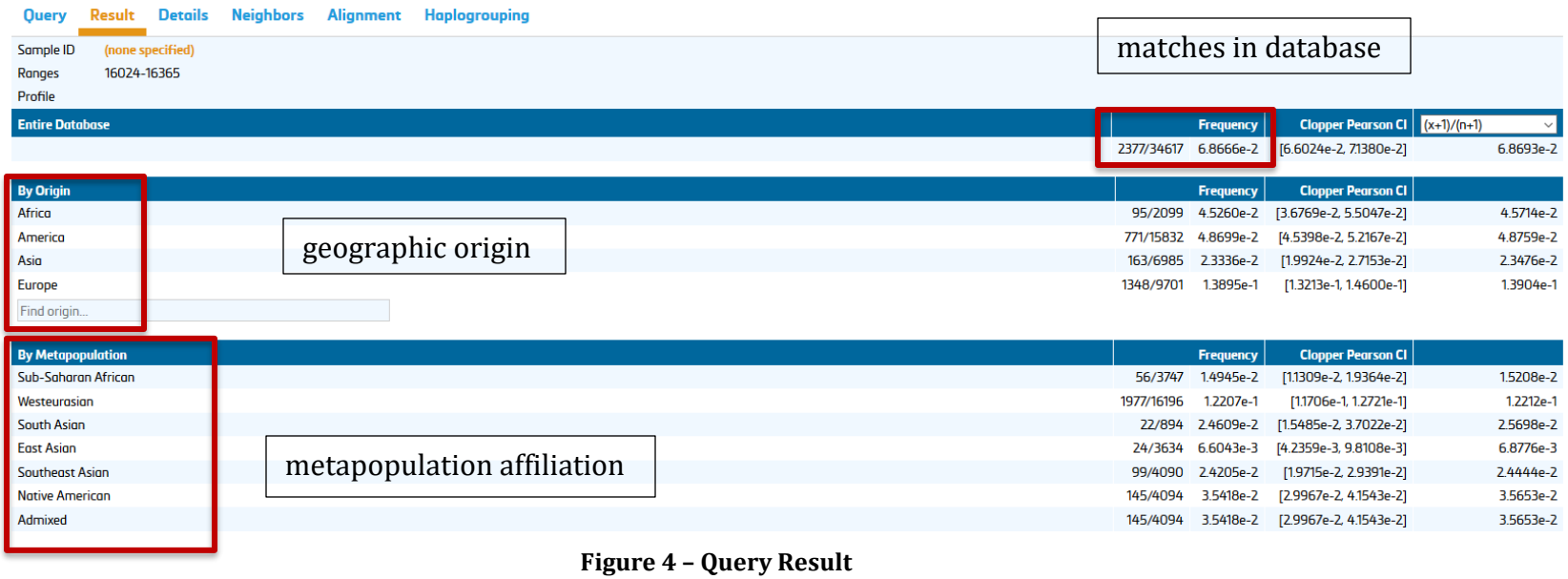


Figure 4 – Query Result

An uncorrected frequency estimate is provided including a two-tailed Clopper Pearson confidence interval. Correction for sampling bias is provided and alternative methods to calculate probabilities are provided in the drop-down box to the right. P values can be estimated based on following formulas:

1. (x+1)/(n+1)
2. (x+2)/(n+2)
3. CI from zero pop

Where x… number of database hits and n… database size

Free text searches are possible for origin and metapopulation to address the relevant subset of the database. This depends on the formulation of the hypothesis, e.g. the reduction of the dataset to the country of Spain.

Note that the number of mitotypes included in a query result depends on the indicated sequence range. Only mitotypes with overlapping sequence ranges to the query sequence are considered. E.g. the query range 16024-576 includes all database sequences that were typed for the entire control region. HVS-I/II data (16024-16365 73-340) are not included in such a query. It may therefore be conservative to also perform a query with standard HVS-I/II sequence ranges.

Below the tabular representation of the database query an interactive map can be found that depicts the sampled populations within the query range (red) and the matches in the sampled populations (green).

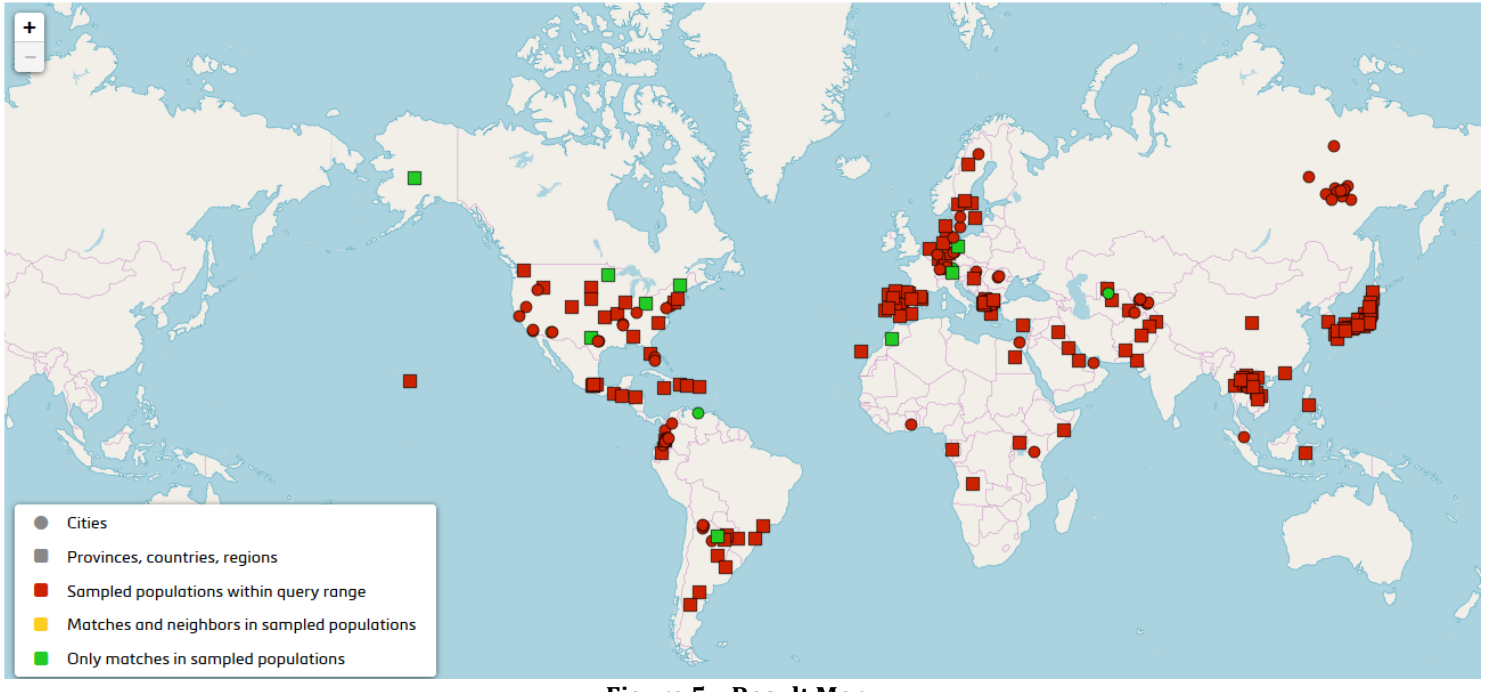


Figure 5 – Result Map

4.3. Details

The Details tab provides a more detailed presentation of the matching mitotypes.

4.3.1. When matches are found

Query | Result | Details | Neighbors | Alignment | Haplogrouping

Sample ID testquery nh
Ranges 16024-576
Profile 16183C 16189C 16191.1C 16519C 263G 309.1C 309.2C 315.1C

15 of 15 haplotypes shown

Origin					Metapopulation		Haplogroup (MRCA)		
filter origins					filter metapopulati		filter haplogroup		
Continent	Region	Country	Province	City		Ignored Mutations	Rank 1	Rank 2	Publications
Africa	Northern Africa	Morocco			Westeurasian		HV	HV	Aboukhalid 2013
Asia	Central Asia	Uzbekistan	Karakalpakstan	Halkabad	Westeurasian	M16183C(0.00)	HV	HV	Irwin 2010
Europe	Western Europe	Austria	Tyrol		Westeurasian	16193insC(0.00)	H55b	H55b	Brandstätter 2007
Europe	Western Europe	Austria	Tyrol		Westeurasian	16193insC(0.00)	H55b	H55b	Brandstätter 2007
America	Northern America	United States of America	Minnesota		Westeurasian	309.1insC(0.00)	HV	HV	AFDIL 2011
America	Northern America	United States of America	Ohio		Westeurasian	309.3delC(0.00)	HV	HV	AFDIL 2012
America	South America	Venezuela	Metropolitan District	Caracas	Native American Admixed	309insCC(0.00)	HV	HV	Castro de Guerra 2012
America	Northern America	United States of America	Vermont		Native American	309insCC(0.00)	HV	HV	AFDIL 2011
America	Northern America	United States of America	Arizona	Phoenix	Westeurasian	309insCC(0.00)	HV	HV	AFDIL 2011
America	Northern America	United States of America	Texas		US Hispanics	309insCC(0.00)	HV	HV	AFDIL 2012
Europe	Western Europe	Germany	Berlin-Brandenburg		Westeurasian	16193insC(0.00)	HV	HV	Zander 2012
America	South America	Argentina	Chaco		Native American Admixed	16193insC(0.00)	H	H	Bobillo 2010
Europe	Western Europe	Germany	Bavaria	Munich	Westeurasian	16193insC(0.00)	HV	HV	Eduardoff 2013
America	Northern America	United States of America	Texas		Westeurasian	16193insC(0.00)	HV	HV	AFDIL 2012
America	Northern America	United States of America	Alaska		African	M16183C(0.00)	HV	HV	AFDIL 2012

Figure 6 - Details

EMPOP provides a summary table of all matching mitotypes that meet the queried sequence range. Columns can be sorted by clicking on the column headers.

Geographic and **metapopulation** origins can be filtered using the text boxes.

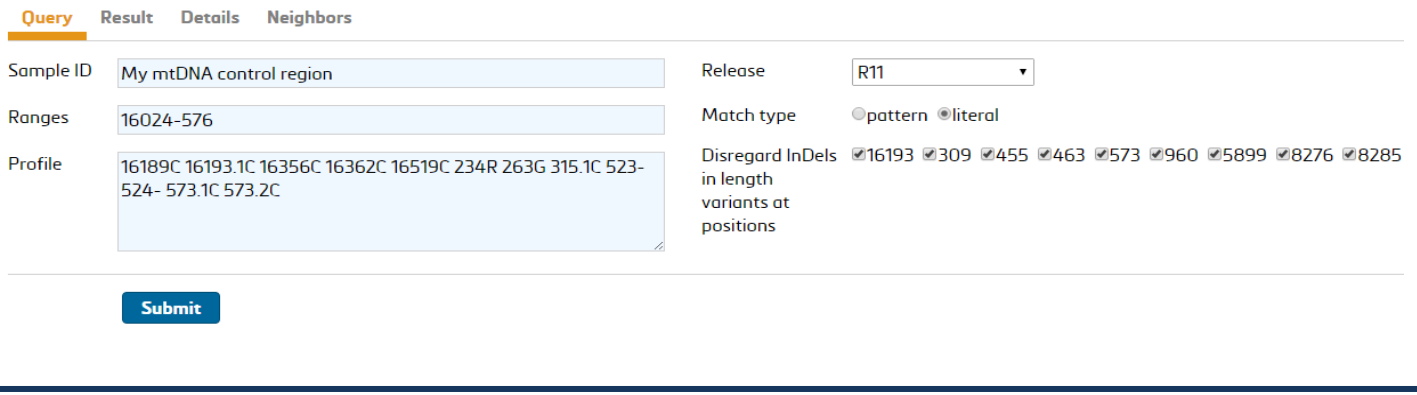
Ignored mutations list the differences between database and query sequences that were disregarded for the search (see 4.1.7. Disregard InDels). The values in brackets display the costs of the listed mutation (details see Huber et al 2018).

Haplogroup indicates the samples' haplogroup assignment. In case of a database match, there is no need to estimate the haplogroup as this column simply indicates the haplogroup of the matching samples. Rank 1 displays the haplogroup estimate with lowest costs (including a tolerance of 0.1) and Rank 2 displays the haplogroup estimate with the next lowest costs (including a tolerance of 0.1).

4.3.2. When no matches are found

MtDNA sequence queries often do not result in database matches. In these cases the Details tab stays empty.

Example:



Unobserved mitotypes are indicated by a frequency value of “0”:

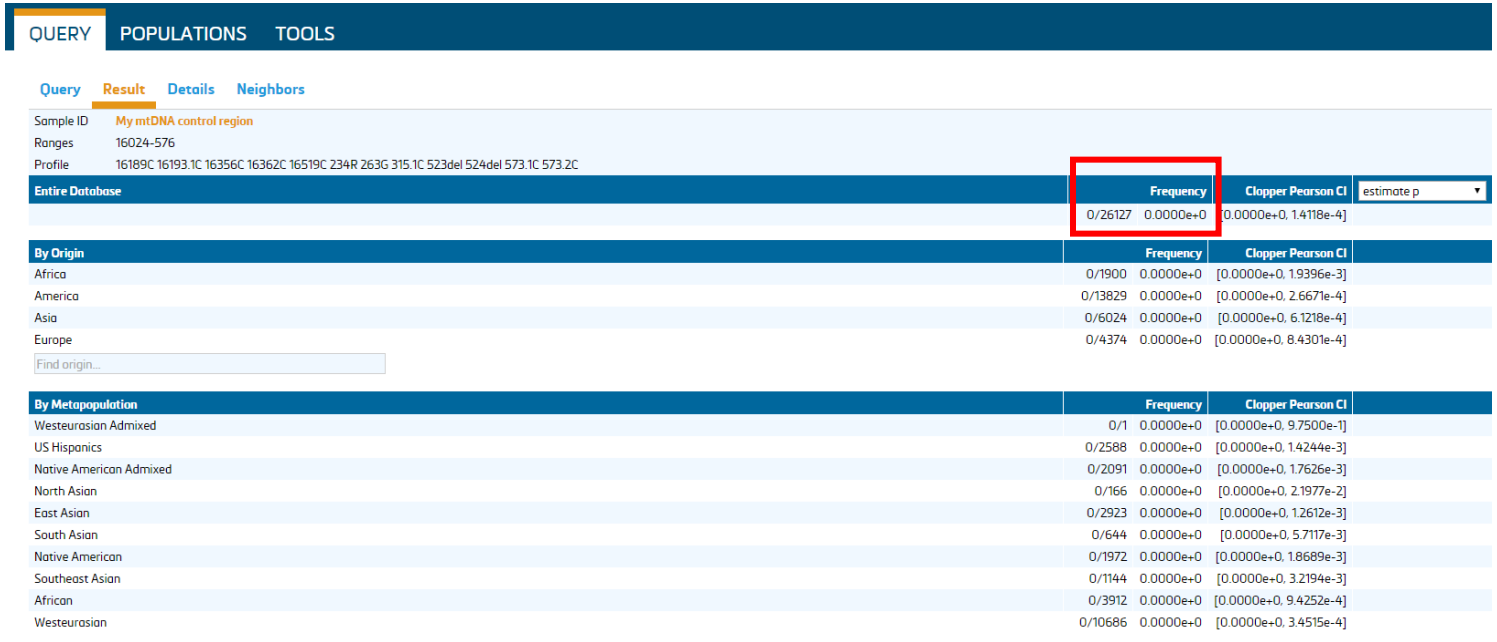


Figure 7 - Results view when no matches were found

In Details no matches are listed:

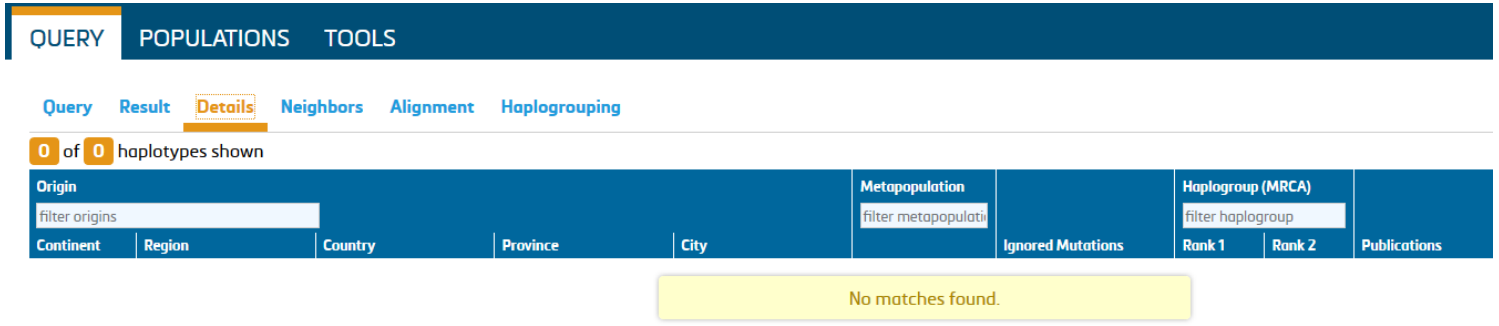


Figure 8 – Details Tab in case of no matches

4.4. Neighbors

Similar sequences with a low number of differences are displayed here.

Query | Result | Details | Neighbors | Alignment | Haplogrouping

Sample ID testquery nh
Ranges 16024-576
Profile 16183C 16189C 16191.1C 16519C 263G 309.1C 309.2C 315.1C

22 of 737 haplotypes shown

Origin					Metapopulation					Haplogroup (MRCA)		
filter origins					of		filter			filter haplogroup		
Continent	Region	Country	Province	City		Cost	Count	Mutations	Ignored Mutations	Rank 1	Rank 2	Publications
America	Northern America	United States of America	Idaho		African	0.50	1	A16183C (0.50)	16193insC(0.00)	HV	HV	AFDIL 2012
America	Northern America	United States of America	California	Orange County	African	0.50	1	A16183C (0.50)	16193insC(0.00)	HV	HV	AFDIL 2013
America	Northern America	United States of America	Missouri		African	0.50	1	A16183C (0.50)	16193insC(0.00) 309.1insC(0.00)	HV	HV	AFDIL 2012
America	Northern America	United States of America	Washington		African	0.50	1	A16183C (0.50)	16193insC(0.00) R189A(0.00) 309insCC(0.00)	HV	HV	AFDIL 2012
America	Northern America	United States of America	Connecticut		African	0.96	2	C16093T (0.46) A16183C (0.50)	16193.2delC(0.00) 309.1insC(0.00)	H1f+16093	H1f+16093	AFDIL 2013
America	Northern America	United States of America	Colorado		African	0.98	2	A16183C (0.50) T16189C (0.48)	16193insC(0.00)	RO	RO	AFDIL 2011
America	Northern America	United States of America	Alaska		African	0.98	2	A16183C (0.50) T16189C (0.48)	16193insC(0.00)	RO	RO	AFDIL 2012

Figure 9 - Neighbors

The display of neighbors follows the same concept as the summary of matches (see 4.3. Details) and includes all mitotypes that are at a distance to the query sequence of **one and two differences ("events")**.

An "Event" refers to the biological meaning of any difference but not the absolute number of differing nucleotides. As such, a tandem deletion (or insertion) in the AC-repeat region between 514 and 524 is regarded as one event, and therefore one difference between otherwise matching mitotypes. The same

rationale applies to the 6 bp Chibcha deletion between 105 and 110 or 106 and 111, the 9 bp deletion between 8281 and 8290, as well as other (less abundant) block indels in the mitochondrial genome.

Additional information is provided with regard to differences between query mitotype and neighbors. These are listed in the columns **cost**, **count** and **mutations**.

Costs are determined by the change from the base profile symbol to the test profile symbol (approximately 1.0 for an average mutation). See Huber et al 2018 for further details.

Count lists the number of mutational events between query and database mitotypes.

Note that some combined mutations are single events, e.g. 523del 524del or 106-111del and treated as such in EMPOP.

Mutations specifies differences between query and database mitotypes, which are listed with the individual costs. Disregarded indels do not contribute to the final costs.

4.5. Alignment

Since EMPOP 4 the phylogenetic alignment is displayed as illustrated in Figure 10 below:

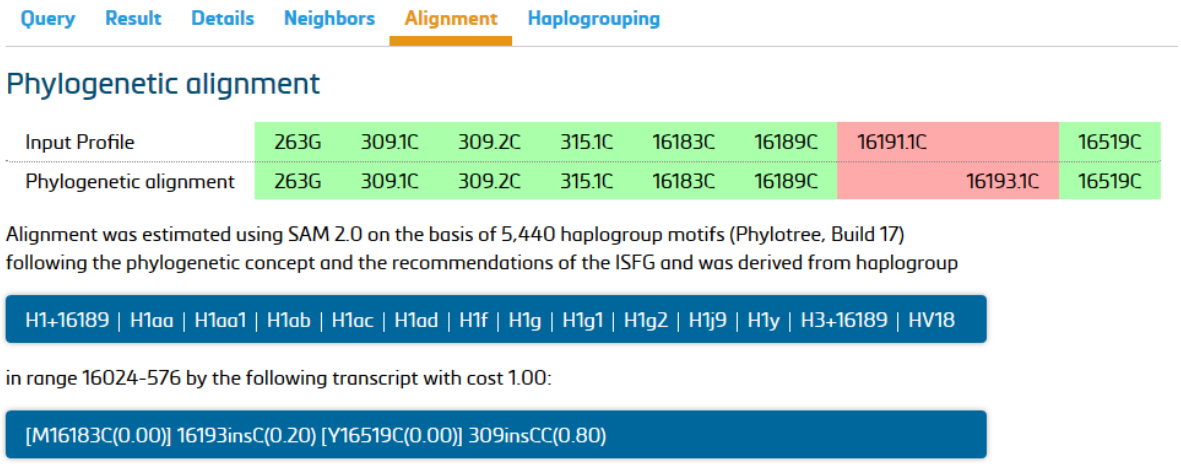
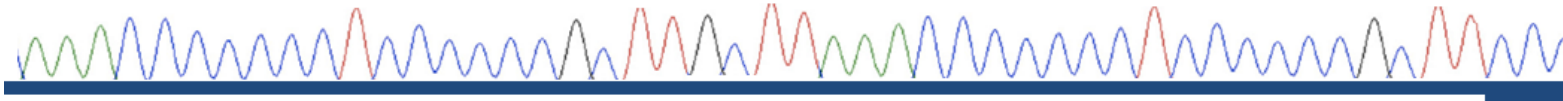


Figure 10 - Phylogenetic Alignment

The scientific publication is currently under preparation, details will follow later.

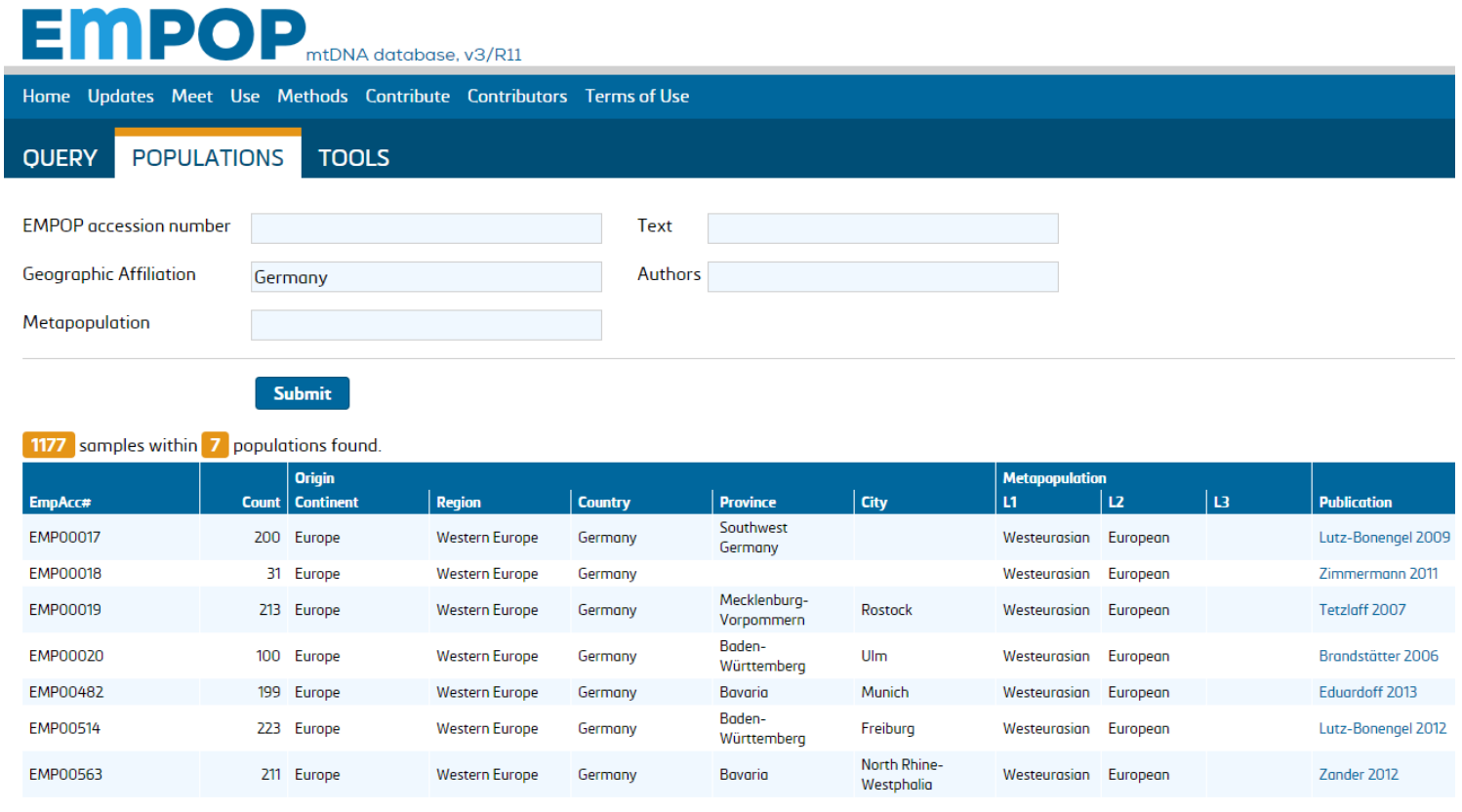


4.6. Haplogrouping

The assignment of haplogroups to mitotypes contributes substantial value for quality control, not only in forensic genetics but also in population and medical genetics. The availability of Phylotree, a widely accepted phylogenetic tree of human mitochondrial DNA lineages, led to the development of several (semi-)automated software solutions for haplogrouping. However, the currently existing tools only make use of haplogroup-defining mutations, whereas private mutations (beyond the haplogroup level) can be additionally informative allowing for enhanced haplogroup assignment.

The scientific publication is currently under preparation, details will follow later.

5. Browsing EMPOP for populations



EmpAcc#	Count	Origin					Metapopulation			Publication
		Continent	Region	Country	Province	City	L1	L2	L3	
EMP00017	200	Europe	Western Europe	Germany	Southwest Germany		Westeurasian	European		Lutz-Bonengel 2009
EMP00018	31	Europe	Western Europe	Germany			Westeurasian	European		Zimmermann 2011
EMP00019	213	Europe	Western Europe	Germany	Mecklenburg-Vorpommern	Rostock	Westeurasian	European		Tetzlaff 2007
EMP00020	100	Europe	Western Europe	Germany	Baden-Württemberg	Ulm	Westeurasian	European		Brandstätter 2006
EMP00482	199	Europe	Western Europe	Germany	Bavaria	Munich	Westeurasian	European		Eduardoff 2013
EMP00514	223	Europe	Western Europe	Germany	Baden-Württemberg	Freiburg	Westeurasian	European		Lutz-Bonengel 2012
EMP00563	211	Europe	Western Europe	Germany	Bavaria	North Rhine-Westphalia	Westeurasian	European		Zander 2012

Figure 11 - Populations

Under the tab POPULATIONS the individual datasets contained in EMPOP can be found by using the accession number (if known), geographic or metapopulation affiliations. Published datasets can be searched by Text (Title) and Authors.

6. EMPOP Tools

The EMPOP tools section provides a suite of software to support the analysis and interpretation of mitochondrial DNA sequence variation.

6.1. Haplogroup Browser

This tool represents the established most recent Phylotree haplogroups in convenient searchable format and provides the number of EMPOP sequences assigned to the respective haplogroups by SAM2. Note that EMPOP provides the MRCA haplogroup if multiple haplogroup assignments are feasible.

Individual haplogroups can also be found by querying differences to the rCRS in a database of > 20.000 mtGenome sequences.

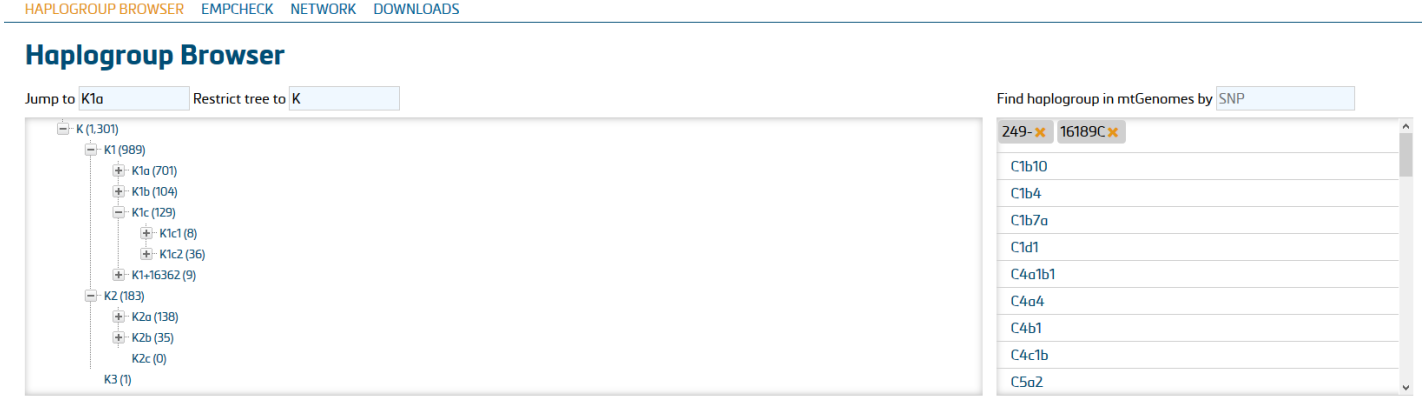


Figure 12 - Haplogroup Browser

6.2. EMPcheck

EMPcheck is a tool to perform plausibility checks on an rCRS-coded data table.

The file format must meet the requirements described below and in Carracedo et al 2014.

6.2.1. Structure of the emp-file

Lines starting with "#!" indicate the sequence range of the mitotype. Note that a given sequence range is applied to all mtDNA mitotypes following this range until a new range is defined. Thus, multiple mitotypes with different sequence ranges can be handled in one file.
The file lists the mitotypes in columns with the following contents.

Column A: Sequence name: don't use blank space or special characters (allowed characters are letters (except umlauts ä, ö, ü), numbers, "-", "_", "/")

Column B: Haplogroup (hg) status: indicate hg, if unknown, use "?"

Column C: Frequency of mitotype (0 - 9999). Typically, this value is “1”, as individual mitotypes should be presented. If it is set to 0 the sample is not considered for the analysis.

Column D: Annotation of the mitotype relative to the rCRS. Separate differences by tabs (or use individual cells in MS Excel). Use forensic notation of sequences as outlined in the revised and updated ISFG recommendations for mtDNA typing (Parson et al (2014)).

Text lines can be included everywhere in the file for comments or description. They need to be marked with "#".
Avoid blank lines (except when marked with "#").

The structure of an EMP file is illustrated below and the file can also be downloaded from: Downloads section in EMPOP.

Example

```
# Population data of 250 individuals from Austria; Walther Parson (walther.parson@i-med.ac.at)
# 100 samples from Innsbruck, 100 samples from Salzburg, 50 samples from Vienna
#! 16024-576
mitotype1          H1c    1    16519C 263G 523DEL 524DEL 477C
mitotype2          R0     2
#! 16024-16365 73-340
mitotype4          T2b    1    16126C 16294T 16296T 16304C 73G 263G 315.1C
mitotype5          ?      1    16223T 73G 263G 315.1C
```

6.3. NETWORK

This tool can be used to calculate and draw quasi-median networks. They are useful to examine the quality of an mtDNA dataset.

MtDNA data tables can be depicted as quasi-median networks to enhance the understanding of the data in regard to homoplasmy and potential artifacts. Highly recurrent mutations are removed from the dataset (filtering) to help detect data idiosyncrasies that pinpoint sequencing and data interpretation problems. A detailed discussion of the method can be found in Bandelt and Dür (2007) and its application in Parson and Dür (2007).

The following section leads you through

- the input and parameter selection of a network analysis
- the output generated by NETWORK
- network drawing and
- interpretation of the results.

6.3.1. Input

Sample Info

The sample-specific information identifies a search. This is also the reference under which the query is reported. The history of NETWORK searches can be found under YOUR ACCOUNT.

Input file (=emp file)

The input file contains the annotated population data. The emp-file is a tab delimited text file that can be created using standard text software or MS Excel (then, safe file under .txt format and rename "txt" by "emp"). Its format needs to meet the following criteria:

Ambiguous symbols

The software accepts the IUB-code. However, ambiguous symbols (e.g. sequence heteroplasmy Y ~ C/T) can cause artificial nodes and links in the network. Therefore it may be necessary to specify a non-ambiguous symbol either by calling the dominant type or by using the phylogenetic background of the sample. You will be notified on the presence of ambiguous symbols on the screen and in the network analysis report. For your information you also get a list of new insertions in your data set that are not known to the current EMPOP database.

New insertions

We collect positions with observed insertions in an EMPOP datafile to which new data are compared. New insertions that have not been recorded in EMPOP yet are displayed to draw the attention on them. This however does not impact the performance of NETWORK.

Filtering

Highly recurrent mutations are removed from the data set (filtering) that would otherwise increase the complexity of the network. You can choose between different filters depending on the application. The contents of the filters can be viewed by clicking on the symbol next to the dropdown box.

Available filters:

- EMPOPspeedy: This filter removes highly recurrent mutations based on the lists provided in Bandelt et al (2002 and 2006). This filter is typically used for the analysis of mtDNA population data within the hypervariable segments - HVS-I (16024 - 16569) and HVS-II (1 - 576).
- EMPOPspeedyWE: This filter removes highly recurrent mutations as presented in Zimmermann et al (2010). This filter is typically used for the analysis of west Eurasian mtDNA population data within the hypervariable segments - HVS-I (16024 - 16569) and HVS-II (1 - 576).
- EMPOPall_R11: This is a superfine filter that contains all mutations observed in EMPOP. This filter provides a very quick check on the data by highlighting only yet unobserved mutations. We update the EMPOPall filter periodically.
- Unfiltered: None of the mutations are removed from your dataset. This is useful for the analysis of very short sequence stretches in the mtDNA CR (see below). The complexity of the network will increase rapidly if no filter is applied to the analysis of larger sequence regions.

Range

The range determines the region for which the network is computed. Any range within 16024-16569 and 1-576 can be queried. In some data very small regions may be interesting for detailed network analysis (e.g. 450-460).

Submit starts the execution.

6.3.2. Output

After clicking on the **Submit** Button, the network calculation is initiated. Depending on the size of the file and the used filter options this process may take some time. When finished, result files will be listed in "Network Result Files" on the "your account" page.

Figure 13 - Download a created Network Result File

Download the file and unzip it to obtain the folder *[RID_FILTERNAME_REGION]*, which contains the following files:

Results file [FILENAME_report.txt]: This file summarizes the settings and the results of the network analysis - for details see chapter Interpretation.

File for drawing the network [FILENAME_network.dnw]: This file can be used to draw the entire network of the mtDNA datafile by dnw.exe.

File for drawing the torso [FILENAME_torso.dnw]: This file can be used to draw the torso of the network of the mtDNA datafile by dnw.exe.

Difference table of the network [FILENAME_network.txt]: This file contains the filtered and reduced mitotypes of the entire network, displayed in dot table format.

Difference table of the torso [FILENAME_torso.txt]: This file contains the filtered and reduced mitotypes of the torso of the network, displayed in dot table format.

EMP-File [EMPFileName.emp]: The emp file which was uploaded

Info-file [FILENAME_info.txt]: Contains the sample identification (which was defined in “Sample info”, see 6.3.1 Input) and the title of the emp file.

Drawing

1. Download the software for drawing the network (DrawNetWorkSetup.exe) from the EMPOP download page.
2. Execute the file and follow the instructions given by the software. Choose a destination folder where the software is to be installed.
3. Once the installation is finished you can find a folder called DrawNetWork containing the software and an uninstaller in the start menu. Files having ".dnw" as file ending are automatically linked to the software. Double-clicking a dnw file opens the network in a separate window. The help menu contains a legend of keys to edit the network (e.g. t ... for drawing a draft of the network, l ... for adding

labels, etc.). During execution the current drawing can be exported in SVG (Scalable Vector Graphic), EPS (Encapsulated PostScript) or GIF format for printing or editing.

Interpretation

The Report.txt file summarizes relevant information of the network analysis. The network is described in a table by the number of samples (n), the number of polymorphic positions (p), the number of partitions or condensed characters (p'), the number of mitotypes (h), the number of nodes in the network (q), the number of nodes in the torso (t) and the number of nodes of the peeled torso (t'). These values are indicative for the quality of a network. However, they depend on the size and composition of the population data set in question. Generally, small t'-values (ideally 1) describe a star-like structure of the network, which is in agreement with the expected evolutionary pattern.

A more suggestive representation of the data is the graph of the quasi-median network. The nodes of this graph are given by the mitotypes or the quasi-medians generated from the mitotypes. In the drawing the frequencies of the mitotypes or quasi-medians are also shown. The root node is drawn with a bold circle and contains the filtered and reduced Anderson sequence (In the rare case that no mitotype contains the filtered and reduced Anderson sequence, the first mitotype is chosen instead and a warning is included in the report). The links are single or combined mutations specified by the syntax for single mutations or / for combined mutations, where the orientation is from the root node outwards. Links with the same mutation are drawn parallel and are labeled only once. The torso is obtained from the quasi-median network by collapsing all pendant subtrees into their base nodes. Thus the analysis of homoplasmy can be restricted to the torso which contains all the reticulation of the network. For each base node the coinciding mitotypes are listed in the report to make it easy to find all corresponding samples.

References & Further Reading

- Bandelt HJ et al (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71:1150-1160
- Bandelt HJ et al (2006) Estimation of mutation rates and coalescence times: some caveats. In: Human mitochondrial DNA and the evolution of Homo sapiens. Springer-Verlag eds. Hans-Jürgen Bandelt, Vincent Macaulay, Martin Richards
- Bandelt and Dür (2007) Translating DNA data tables into quasi-median networks for parsimony analysis and error detection. Mol Phylogenet Evol 42:256-271
- Parson and Dür (2007) EMPOP - A forensic mtDNA database. FSI:Genetics 1:88-92
- Schwarz and Dür (2011) Visualization of quasi-median networks. Discrete Applied Mathematics 159(15):1608-1616
- Zimmermann et al (2014) Improved visibility of character conflicts in quasi-median networks with the EMPOP NETWORK software. Croat Med J 55(2): 115-120.
- Parson et al (2013) Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). FSI: Genetics 7(5): 543-549
- Eduardoff et al (2017) Optimized mtDNA Control Region Primer Extension Capture Analysis for Forensically Relevant Samples and Highly Compromised mtDNA of Different Age and Origin. Genes 8(10)
- Strobl et al (2018) Evaluation of the precision ID whole MtDNA genome panel for forensic analyses. FSI: Genetics 35: 21-25.