

Drawing and interpreting quasi-median networks with EMPOP

A short introduction

1. Quasi-median network basics
2. How to perform a QM network analysis?
3. How to interpret the QM network?
4. Networking via EMPPOP means that...

Bettina Zimmermann, Alexander W. Röck, Walther Parson
Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria

<http://empop.org>

info@empop.org

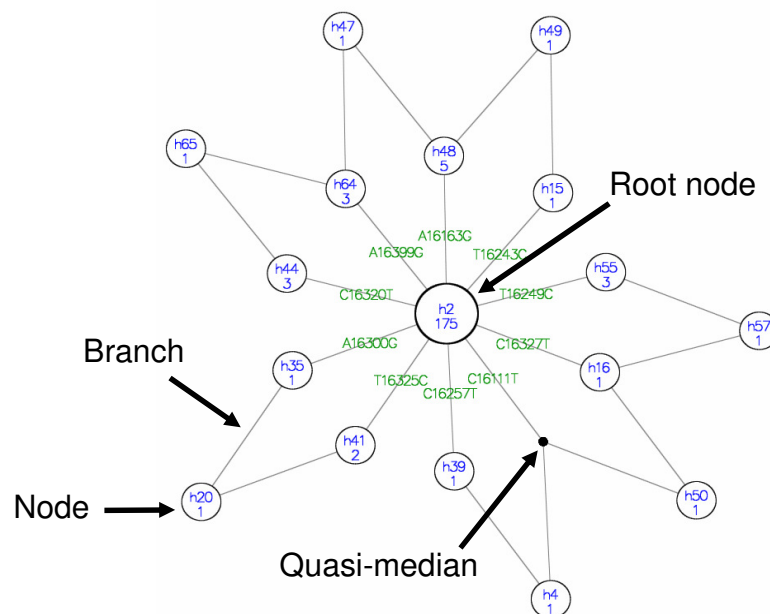
October 20, 2011

1. Quasi-median (QM) network basics

Quasi-median networks are an effective tool to check the quality of mtDNA data.

The application of quasi-median networks enables mtDNA examination by graphically representing the genetic structure of the lineages in the data set. Hence, data idiosyncrasies in regard to homoplasy and potential artifacts which are caused by sequencing or data interpretation problems can be detected. Filtering of highly recurrent mutations prior to network analysis is required to reduce the complexity of the resulting network. The quasi-median network displays the filtered, condensed and reduced haplotypes of the analyzed data set. For the torso of the network all pendant subtrees are collapsed into their base nodes. Use the **torso** for a **quick data check**, the **network** is useful when a **closer look** on the data is preferred.

Here the quasi-median network torso of 202 mtDNA haplotypes from the example data set WE_Etalon with the analysis range set to 16024-16569 and filtered with the WE filter is displayed.



- The **nodes** correspond to filtered, reduced and condensed haplotypes.
- The **root node** is displayed with a thick border; in most cases it corresponds to the rCRS haplotype; if this is not the case this is annotated in the accompanying report file.
- The **branches** represent mutational events and are to be read starting at the root node; mutational events given on the branches are colored either in green for transitions or in red for all other mutations (insertions, deletions and transversions); parallel branches carry the same mutation and are labeled only once.
- The small black node is a **quasi-median** which represents a virtual haplotype; quasi-medians are required to link haplotypes within a QM network and are calculated by the network software.

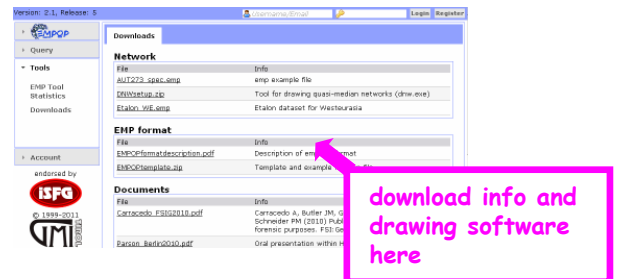
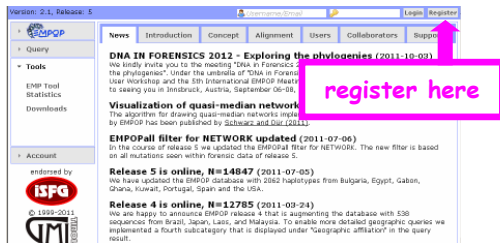
2. How to perform a QM network analysis?

Preparations

Register and login at <http://empop.org>

Download, unzip and install the drawing program DNWsetup.zip from the section „Tools → Downloads“

Prepare your data in emp file format (see „EMPOPtemplate.zip in the „Downloads“-section)



Perform the network analysis

Open the Network link in the „Tools“-section

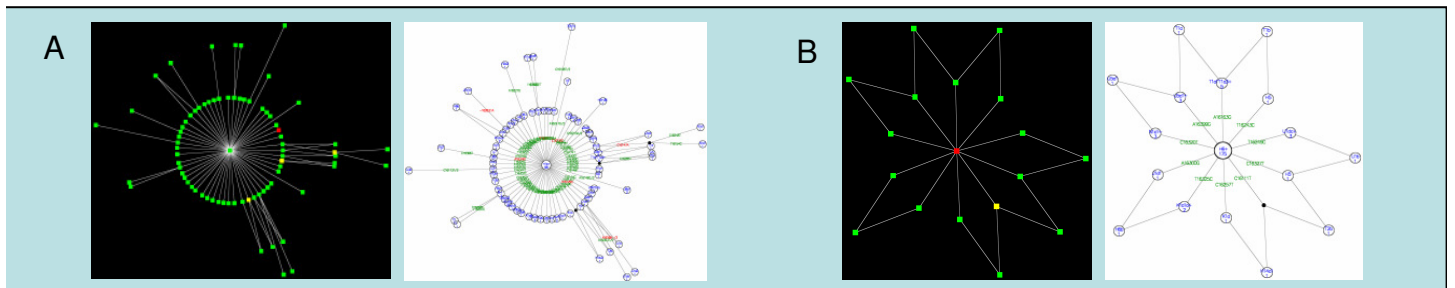
Input your emp file and the required analysis settings (i.e. filter, analysis range) and proceed

Download and unzip the resulting zip file that includes all of the analysis results

Draw the network

By double-clicking the „*_network.dnw“ file the image of the network will automatically be drawn by the previously installed drawing software. The same for „*_torso.dnw“.

The quasi-median network (A) displays the filtered, condensed and reduced haplotypes of the analyzed data set, for the torso of the network (B) all pendant subtrees have been collapsed into their base nodes which facilitates identification of complex structures



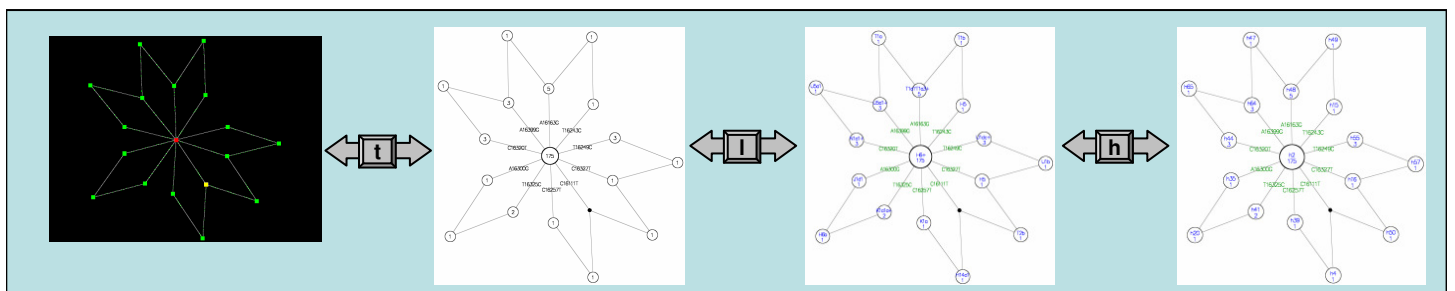
Configure the network

Change the appearance of the network/torso

Pressing „t“ changes the draft version to the drawing version

Pressing „l“ changes the labeling of the nodes and coloring of the mutations

Pressing „h“ changes the labels from haplogroup to haplotype designation (see *_report.txt file)



Change the font size using „f“ and „F“, respectively, the circle size using „c“ and „C“

Move nodes (left mouse button) and branches (right mouse button) by clicking and holding the center of the nodes and the mutation labels, respectively

3. How to interpret the QM network?

Example "Etalon_WE"

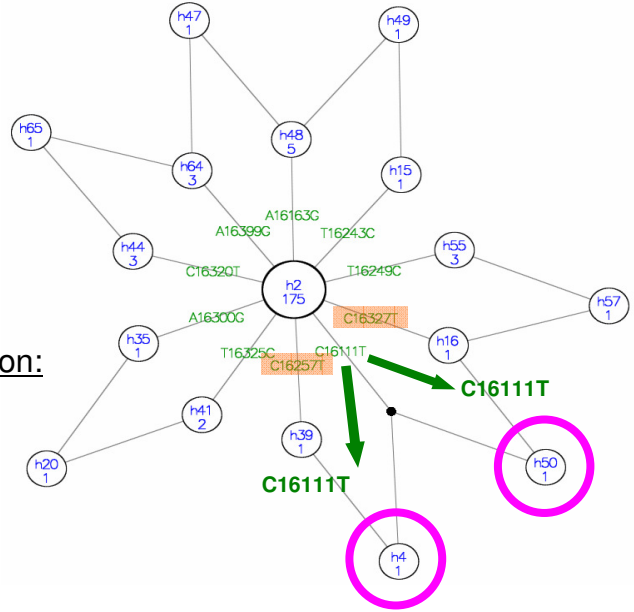
Two samples of the West Eurasian Etalon data set contain the transition **C16111T**. In the network torso the two samples are named **h4** and **h50**. The two samples are separated by the mutations **C16257T** (h4) and **C16327T** (h50). To connect these two samples in this network the quasi-median is required.

If you find quasi-medians or complex reticulations in your network check the raw lane data of the corresponding samples!

Where do I find information about the haplotypes shown in the network?

The download zip file includes all relevant information:

- the input *.emp-file
- the two *.dnw-files for drawing the network and the torso
- a report.txt-file
- and three *.txt-files including further information

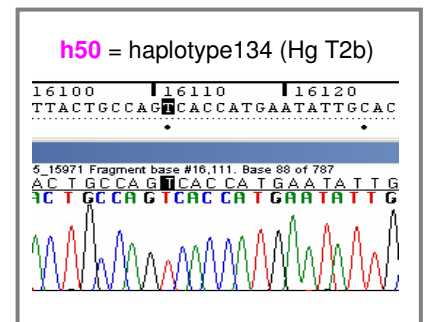
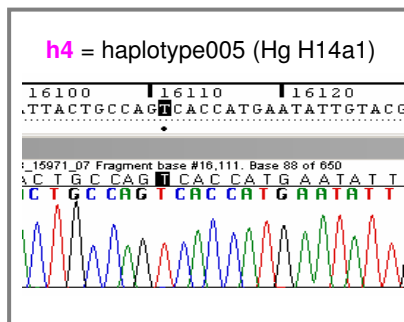


In the *.report.txt-file you find the two haplotypes h4 and h50 and their corresponding sample names „haplotype005“ and „haplotype134“. The .emp-file includes the corresponding mtDNA profile by means of differences with respect to the rCRS.

*_report.txt-File	*_emp-File
<pre> 45 h2: haplotype002 (H11) haploty 46 h3: haplotype004 (H11a2) 47 h4: haplotype005 (H14a1) 48 h5: haplotype008 (H1a) haploty 49 h6: haplotype016 (H1k) haploty ... 90 h47: haplotype124 (T1a) 91 h48: haplotype125 (T1a) haploty 92 h49: haplotype130 (T1b) 93 h50: haplotype134 (T2b) 94 h51: haplotype135 (T2b) 95 h52: haplotype139 (T2b6) 96 h53: haplotype144 (T2e) </pre>	<pre> haplotype001 H10a1 1 16114T 16344T 16519C haplotype002 H11 1 16293G 16311C 195C 263C haplotype003 H11a1 1 16278T 16311C 195C 263C haplotype004 H11a2 1 16092C 16140C 16265G haplotype005 H14a1 1 16111T 16256T 16257T haplotype006 H15 1 55C 57G 185A 263G 309.1C haplotype007 H15a1 1 55C 57G 93G 263G 309.1C ... haplotype132 T2a1a1 1 16126C 16294T 16296T haplotype133 T2b 1 16126C 16294T 16296T 16304 haplotype134 T2b 1 16111T 16126C 16294T 16296 haplotype135 T2b 1 16051G 16126C 16294T 16296 haplotype136 T2b 1 16126C 16294T 16296T 16304 haplotype137 T2b3a 1 16126C 16292T 16294T 16296 </pre>

Check the raw lane data of samples „haplotype005“ and „haplotype134“ and verify the transition C16111T in both samples.

→ the virtual haplotype (quasi-median) is verified



4. Networking via EMPOP means that...

- Registration and software download/installation is required only once
- The calculation and drawing of a network can be repeated any time directly via EMPOP
- A network analysis on EMPOP is stored in your account history for easy future recognition using an arbitrary identifier chosen by you
- You should not apply more than 300 haplotypes per analysis since the resulting network will be too complex in most cases
- You can use the high quality Etalon data set available in the download section to embed your data (every difference to the Etalon network should be evaluated)
- You should play around with different analysis ranges up to the entire control region (depending on sample size)
- Complex reticulations in the network can be caused by
 - errors** → check mutations that connect quasi-medians with nodes
 - sample size** → reduce sample size and try again
 - phylogeny** → remove phylogenetically distant haplotypes and/or use the appropriate filter and try again
- You should filter your data according to their phylogenetic background (e.g. Westeurasian data needs the Westeurasian filter); try the “EMPOPall” filter to pinpoint unobserved mutations; additional filters are added regularly
- You should edit your network with the drawing software trying to find those samples within the network that are responsible for reticulations, quasi-medians and unknown mutations
- There is also a more detailed manual available via EMPOP in case you are curious now that you have read this short introduction! And last but not least, contact us via info@empop.org if you have detailed questions concerning your network analysis.