mitoBench Documentation

Release 1.0

Judith Neukamm & Alexander Peltzer

Sep 17, 2021

Contents

This is the main mitoBench documentation, where you can find information about the prerequisites, the installation, and the usage of this workbench.

CHAPTER 1

Prerequisites for the installation of mitoBench

1.1 Operating System Support

mitoBench has been implemented as a platform-independent tool and can thus be installed on Linux, Windows and MacOS. A Java 8+ platform has to be installed on the workstation used for running the tool.

1.1.1 Linux

It has been successfully tested on different flavors of Linux based operating systems, including ArchLinux and Ubuntu 17.04.

1.1.2 Mac OSX

The application works out of the box on macOS X 10.12 Sierra. An up to date Java Runtime Environment (>8) is required to run the application.

1.1.3 Windows

The application can be run on Windows 10. Solely the log file cannot be saved at the end for now.

1.2 Software Requirements for mitoBench

Install a suitable Java 8 runtime environment. We tested both Oracle Java 8 SE and OpenJRE 8 on Linux. The latter requires to install JavaFX in addition, whether the Oracle JRE already ships with the required JavaFX libraries.

CHAPTER 2

Installation Instructions for mitoBench

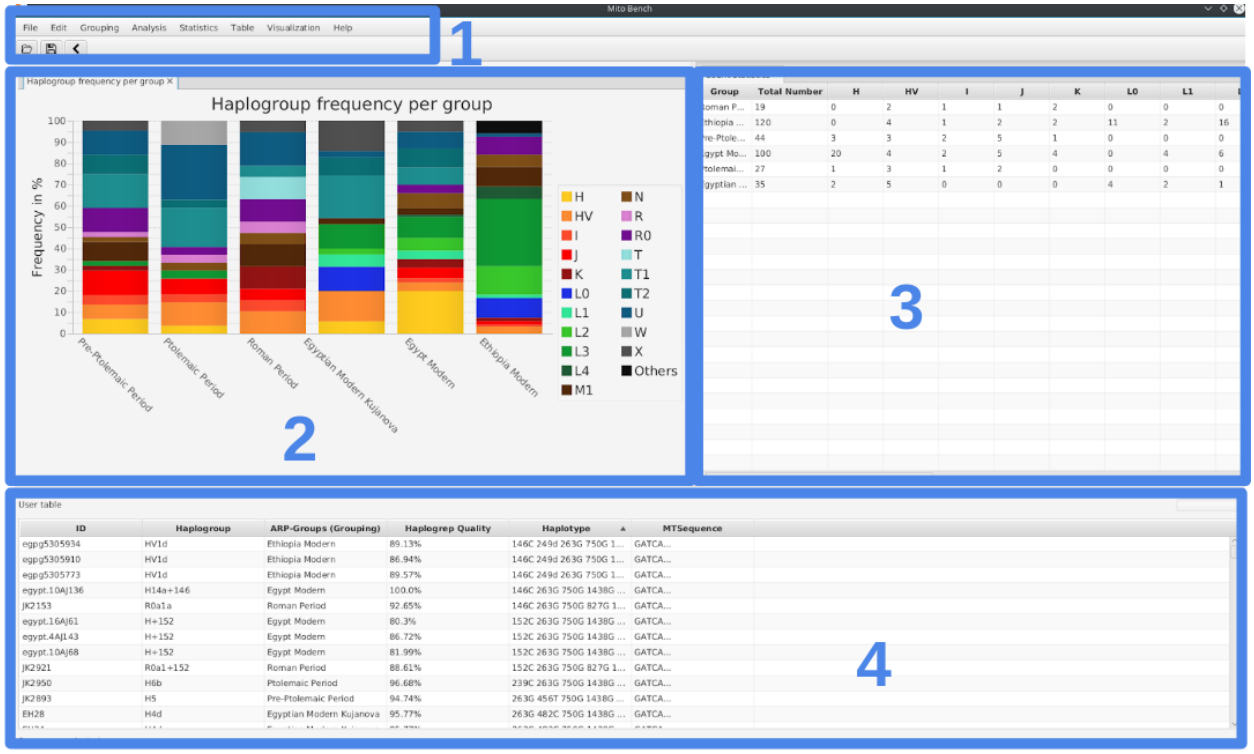
The tool can be downloaded from MitoBench's GitHub page. After downloading the JAR file, you can start the application via double click on most operating systems (OSX, Windows and Linux) via a simple double click. If not, please install Java on your workstation.

CHAPTER 3

General Usage

mitoBench can be used to upload and convert files and to manipulate, analyze and visualize mitochondrial data. It can can be used in offline mode, but some functions (such as map view and DB import) require a working Internet connection.

The main window is divided into three parts. The upper part contains a toolbar (1), on the left a visualization field (2), and a statistic/analysis area (3) on the right side. The following part shows the samples in a table format (4).



3.1 Log file

By closing mitoBench, you will be asked if you want to save a log file. This file contains all steps that were done during your analysis with mitoBench, which helps you to reconstruct your analysis at a later date.

CHAPTER 4

Workbench

4.1 Data Import

mitoBench offers different ways to import data. The imported information is represented in table format. If e.g. samples have been imported from different files, they are merged into one row based on the column *accession id*.

Note: To merge information from different files, make sure that the samples have identical accession ids!

4.1.1 Import via data upload

mitoBench supports different file formats:

- Multi-FastA (.fasta, .fa, .fas, .fna)

  mitoBench supports the upload of fasta and multifasta files. The header of each entry will be set as accession ID without a version information, which is the case e.g. for GenBank entries.

  Example:

  KJ154949.1 Homo sapiens isolate Y5728 mitochondrion, complete genome will be shortened to *KJ154949* and set as accession id.

- Arlequin (.arp)

  Files that were used for analyses in Arlequin can be imported to mitoBench as well. The grouping will be set as a new column.

- Haplogrep (.hsd)

  The hsd file must be tab-separated. Files separated by comma or space cannot be read. HaploGrep2 automatically creates tab-delimited files.

- Excel (.xls, .xlsx)

  - The file needs to have the same format for the first two rows as the generic file. First row will be used as header and has to start with ##, the second row must contain the data types (starting with #).

- Generic file (.tsv, .csv)

  To upload a generic file, the file must have a specific format:

  - The first line starts with ## and contains the column names separated with tabs/commas.

    ##<colname1>\t<colname2>\t. . . .

    ##<colname1>,<colname2>,. . . .

  - The second line starts with # and specifies the data type of the column. You can find a list of all possible data types in the section below.

    #<data type1>\t<data type1>\t. . .

    #<data type1>,<data type1>,. . .

  - Third line to end:

    Contains the actual data. One line per sample, tab-/comma-separated.

    Example:

```
##ID    C14-Date    Sample Country
#String    C14 String
JK2916    cal BC 1111-998 Eygpt
JK2895    cal AD 25-111   Eygpt
JK2907    cal AD 26-84    Eygpt
JK2907    cal AD 26-84    Eygpt
```

- MitoProject (.mitoproj)

  This file contains all information about a previous project, like grouping, filtering, and project-specific haplogroup list. Only one project can be imported per session.

Data types

- String
- Categorical

  The same as data type *String* yet

- Location

  The location is expected as latitude and longitude. Each value has a separate column.

  Warning: The decimal point has to be a point (.), no comma!

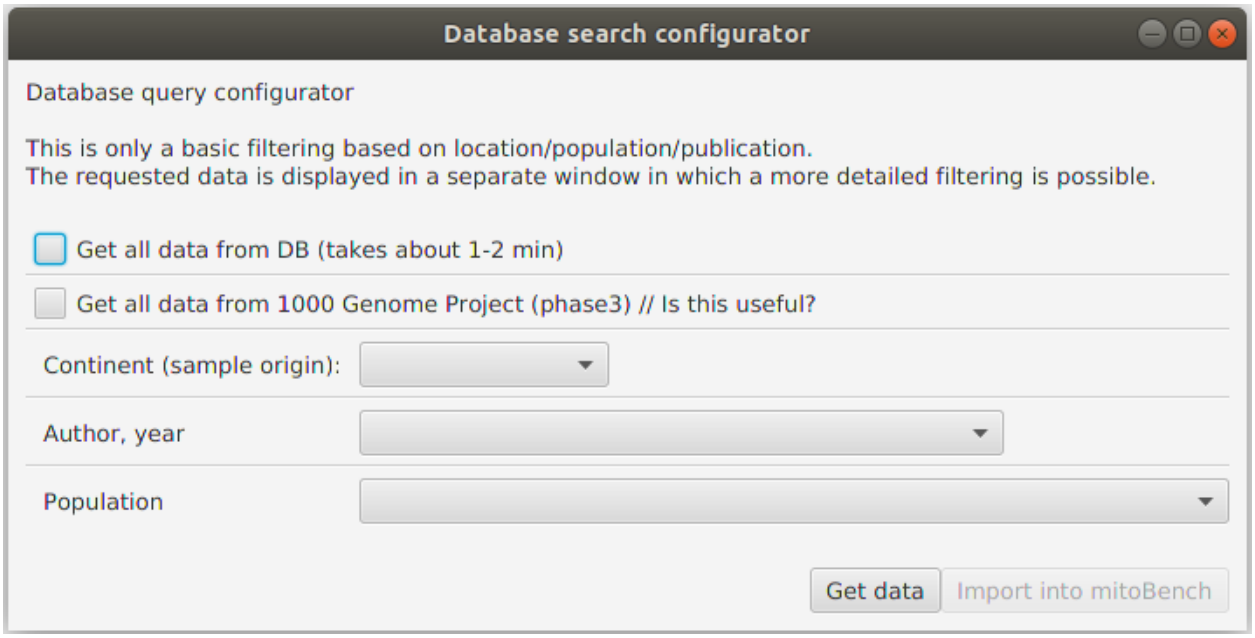
- C14

  We are working on this at the moment.

Note: All files can also be imported into mitoBench via drag & drop.

4.1.2 Import from mitoDB

To import data from mitoDB, select *File –> Import Data from DB*. This opens the database search configurator, where you can do a initial filtering of the data.



Currently, three different filter modes are supported. This is an either-or filtering, so the different options cannot be combined, (which is planned in the future).

- Getting all data

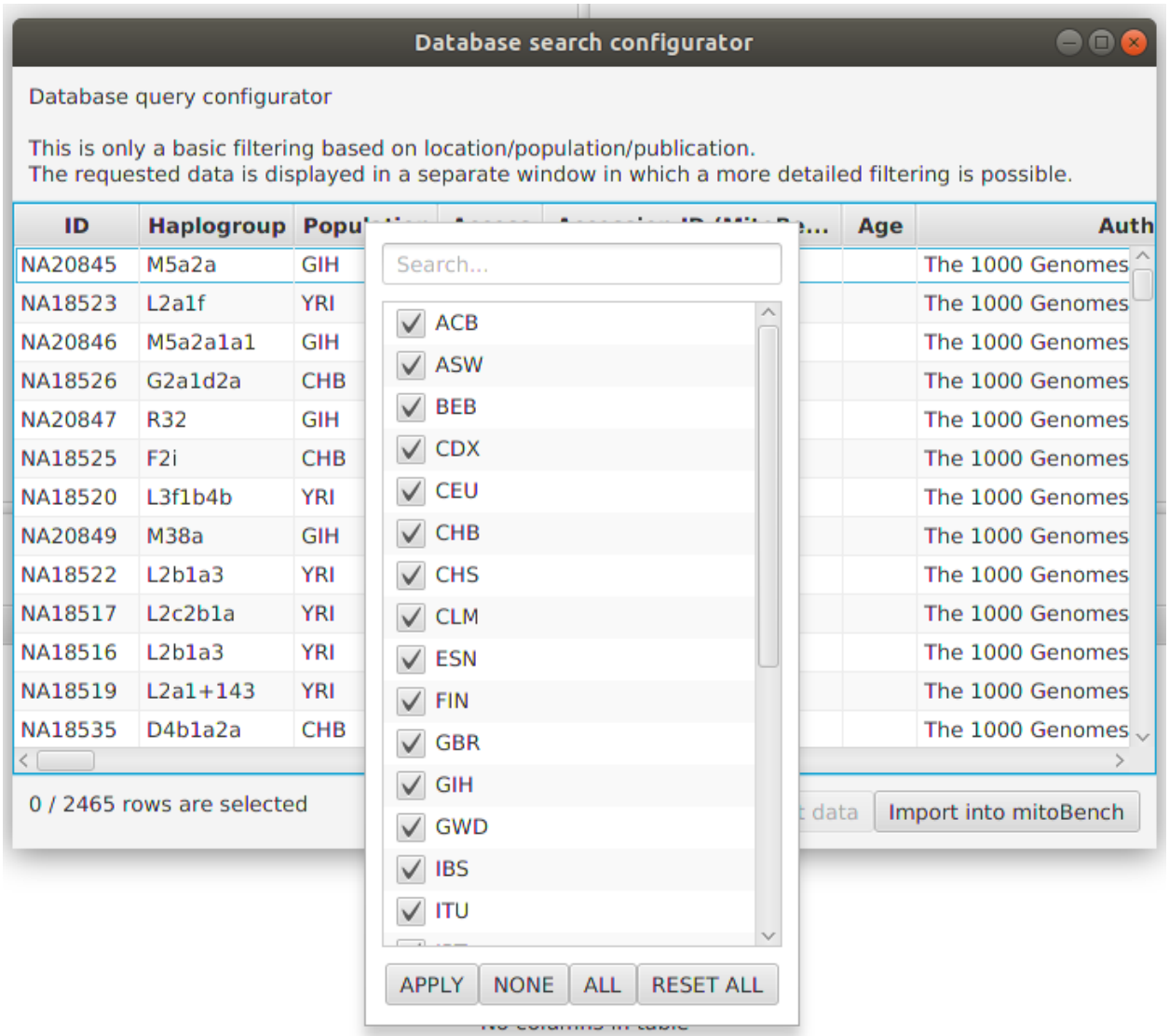
  Returns all data that are contained in the database. Depending on the internet connection, this can take some minutes.

- Getting only data from 1000 Genome Project (phase 3)

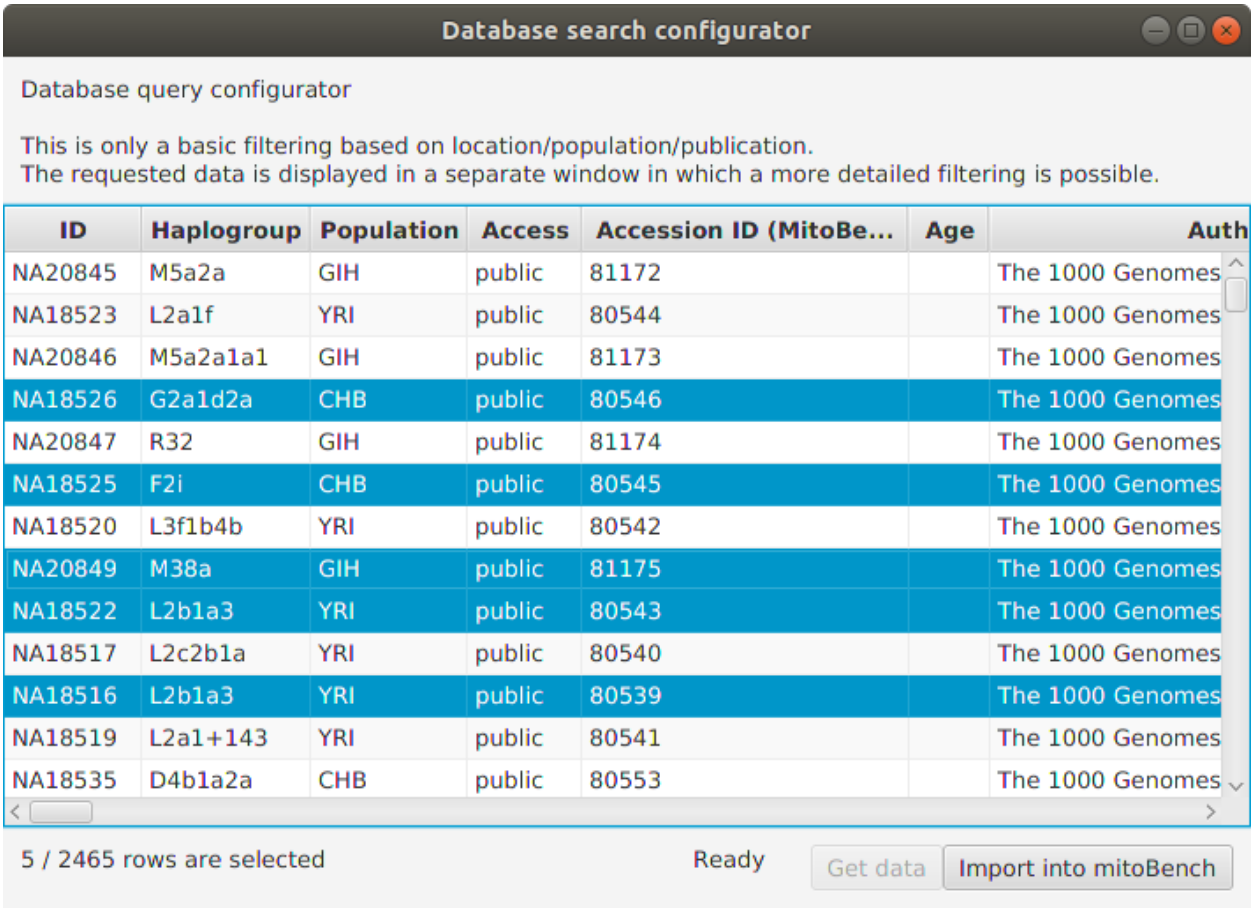
  Returns 2,504 sequences from the 1000 GP Phase 3.

- Filtering data by sample location, publication, or population.

After a successfully getting the data (clicking on 'Get Data'), they will displayed in this window. Now, a more detailed filtering is possible by right-clicking on the corresponding column. This will open a list with all entries contained in the data selection and allows to select and deselect certain values.



It is also possible to select rows and add only the selection to the workbench.



After clicking the 'Import into mitoBench' button, the data can further be explored in the workbench.

4.2 Datatable

The datatable shows all data that are load into mitoBench.

4.2.1 Column order

The first three columns sho the ID, Haplogroup, and Population. All other columns are sorted alphabetically. The column order can be changed via *Table –> Define Column order*.

4.2.2 Manipulating columns

Deleting / Adding columns

It is possible to add and delete columns to the table by right-clicking on the table.

Replacing column content

Values for new created or existing columns can be changed by right-clicking on the table as well. If multiple rows are selected, the value will be set for all selected entries.

Copy/Delete column

Needless columns can be removed from the table. Moreover, single columns also can be copied.

4.2.3 Use selected rows

> Table -> Use selection

In case you want to reduce your dataset further, select the data you need in the data table and use the *Get selected rows* option to use only this data selection for all further analyses. The unselected data will not be saved.

4.2.4 Clear table

> Table -> Clear table

This removes all data from the data table.

4.3 Analyses

4.3.1 Project-based Haplogroup list

> Edit -> Define Haplogroup List

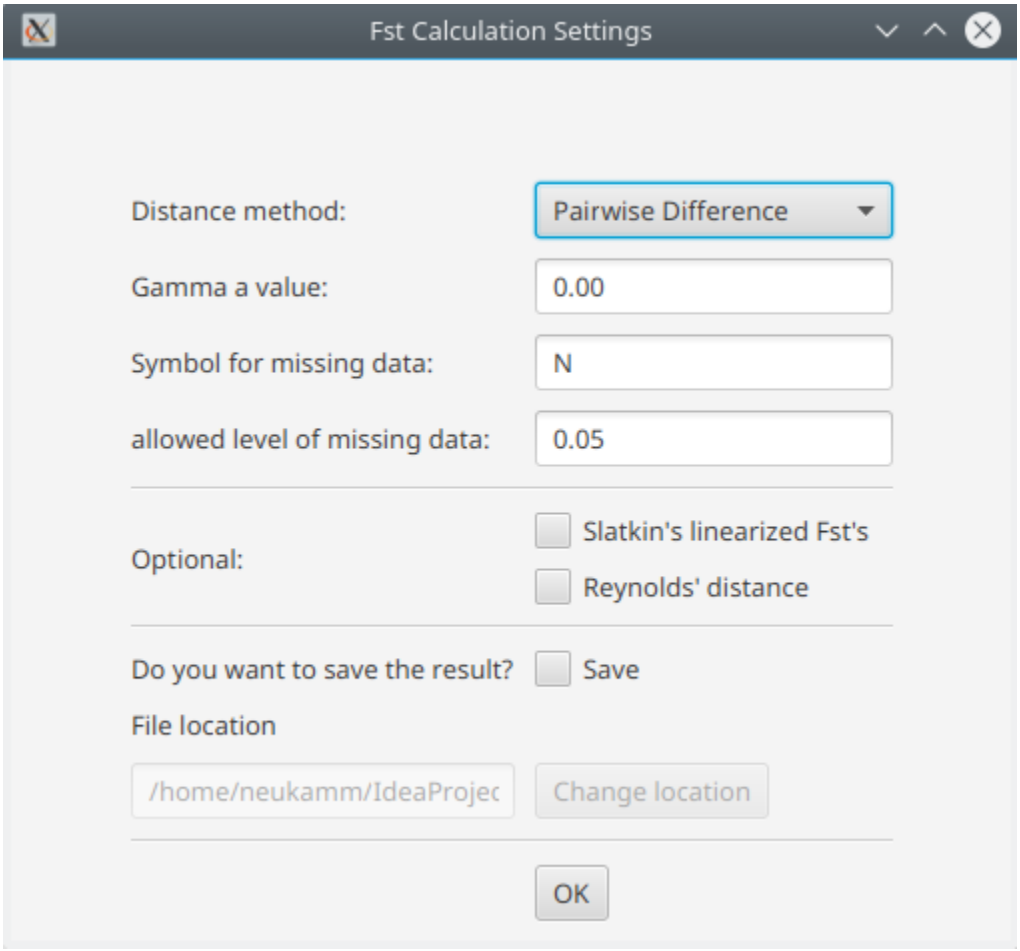
This allows the user to specify a list of Haplogroups for a project. The list will be applied whenever the user is asked for, i.e. some visualization methods, Haplogroup counting and PCA analysis. This list is also stored in the mito project file (.mitoproj) and automatically loaded when opening the project file again.

4.3.2 Pairwise Fst values

> Analysis -> Calculate pairwise Fst

This calculates the pairwise Fst values based on the approach of Hudson et al. and implemented according Brid et al. *(Bird, CHRISTOPHER E., et al. "Detecting and measuring genetic differentiation." Phylogeography and population genetics in Crustacea 19.3 (2011): l-55.)* to achieve results comparable to the calculations done in Arlequin *(Excoffier, L. and H.E. L. Lischer (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 10: 564-567.)*.

Advanced settings are provided by the configuration dialogue.



- **Distance method:**
  - Pairwise difference
  - Jukes & Cantor
  - Kimura 2-parameters
- **Gamma a value**
  - This parameter is used to perform gamma correction on the distance measure. The gamma a has no effect on the distance method 'pairwise difference'
- **Symbol for missing data**
  - Specify symbol that is used for missing data in your samples.
- **allowed level of missing data**
  - Specify allowed the percentage of missing data per locus/position. Loci with a value higher than the threshold are excluded from all analyses.
- **Slatkin's linearization (Slatkin 1995)**
  - Linearize the Fst values with Slatkin's linearization (D = Fst / (1-Fst))
- **Reynolds' distance (Reynolds et al. 1983)**
  - Linearize the Fst values with Reynolds' linearization (D = -ln(1-Fst))
- **Number of permutations**

  - The null hypothesis is the assumption that there is no difference between the population. Through permuting haplotypes between pairs of populations, a p-value is calculated to get the significance of the test. The number of permutations can be set by the user, 0 permutations mean that no test will be performed. The p-value is defined as the ratio between the number of permutations that lead to a Fst value higher or equal than the observed one and the total number of calculated p-values.

- **Significance level**
  - Significance level for p-value
- **Save result**
  - The result is displayed in the mitoBench and can be downloaded as text file as well. The file location can be specified here.

Finally, the result is displayed a text format and the Fst values are visualized as a heatmap.

4.3.3 Haplogroups

> Analysis -> Calculate haplogroups

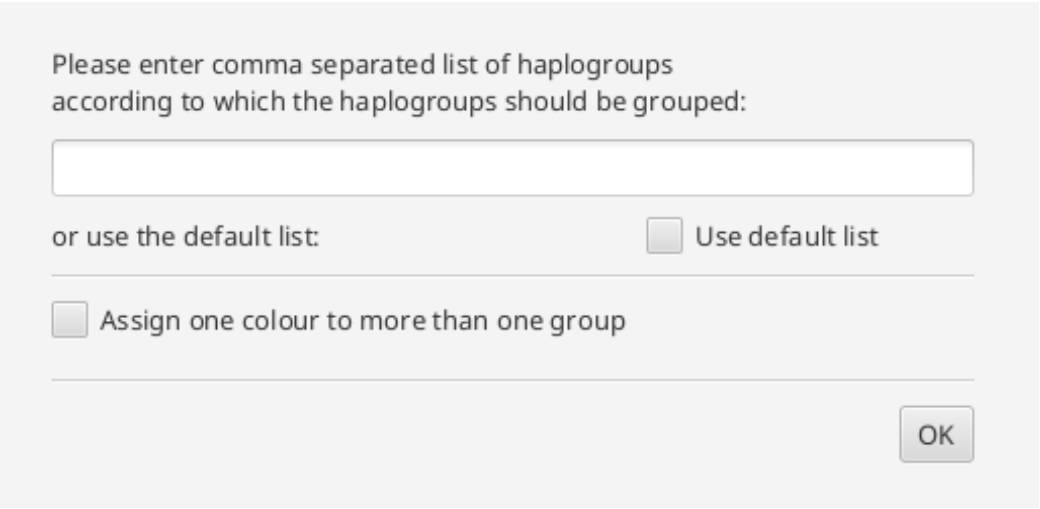
The Haplogroups are determined by the current version of HaploGrep2 (version 2.1.18) based on the FastA sequence. The calculated haplogroups are added as new column in the table view and can be downloaded as hsd file as well.

4.3.4 PCA (still in testing)

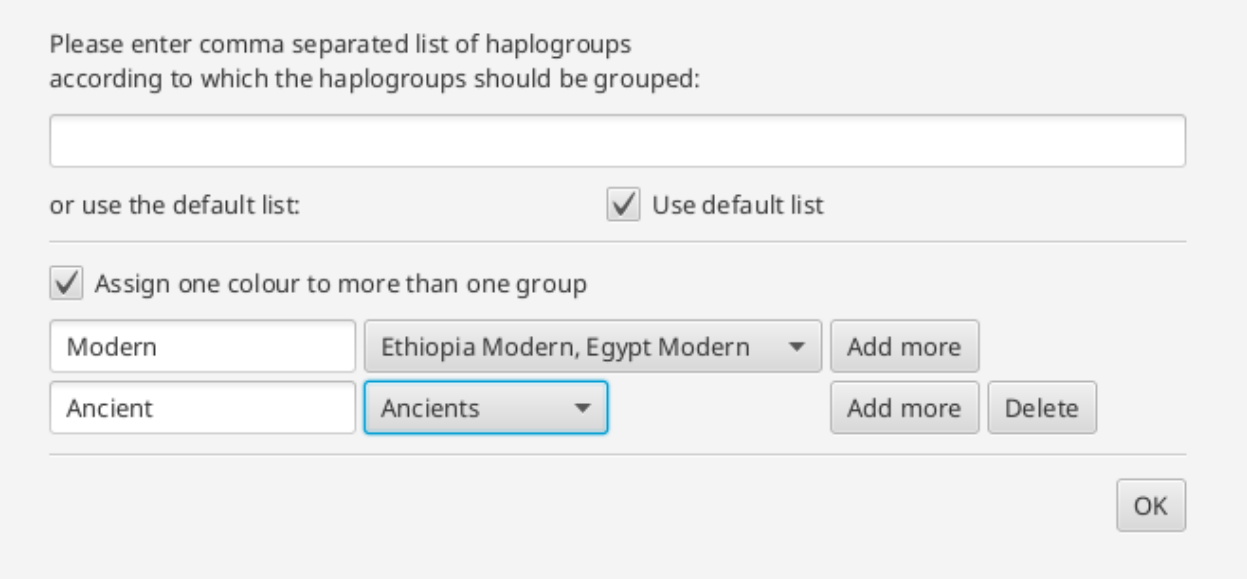
> Analysis -> PCA analysis

Warning: This functionality has to be tested in detail. Please let us know, if you get unexpected or obviously incorrect results.

The principal component analysis requires a grouping of the data and the haplogroups. A basic grouping of the data has to be done previously. The Haplogroups can be set in the configuration pane. The coloring can either be set like the groups (each group gets one color)



or several groups can be assigned to one color. In the text field, the user can specify a name. The color is chosen by the tool.



The result will be shown as a 2-dimensional plot in the visualization pane, and the counts used for the calculation in the statistics pane. Hovering over the data point opens gives information about the represented group.

4.4 Statistics

4.4.1 Haplogroup counts

> Statistics -> Count Haplogroups

This counts the Haplogroups per group - or of the entire data set - if no grouping is defined. The result is represented in a table format and can be exported via *File -> Export statistics*.

4.4.2 Haplotype frequency

> Statistics -> Calculate Haplotype frequency

This calculates the occurrence and frequency of each haplotype based on the current version of PhyloTree. The result also contains the Haplogroups assigned to the Haplotype.

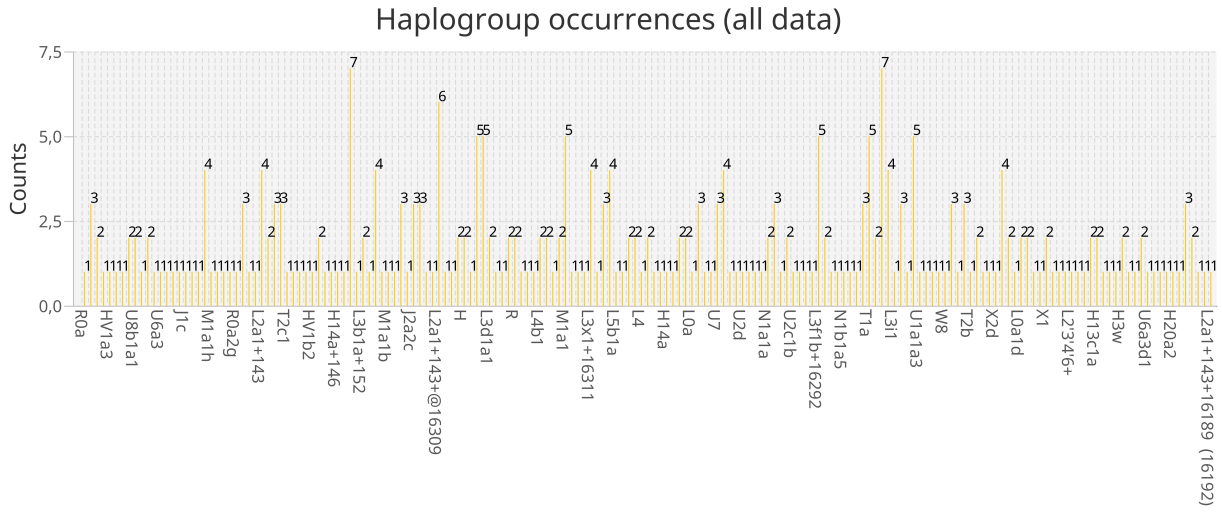
4.5 Visualizations

mitoBench provides a variety of different visualizations for the Haplogroup distribution within a data set.

4.5.1 Bar plot

> Visualization -> Haplogroups -> Create Barchart -> Plot Haplogroup frequency

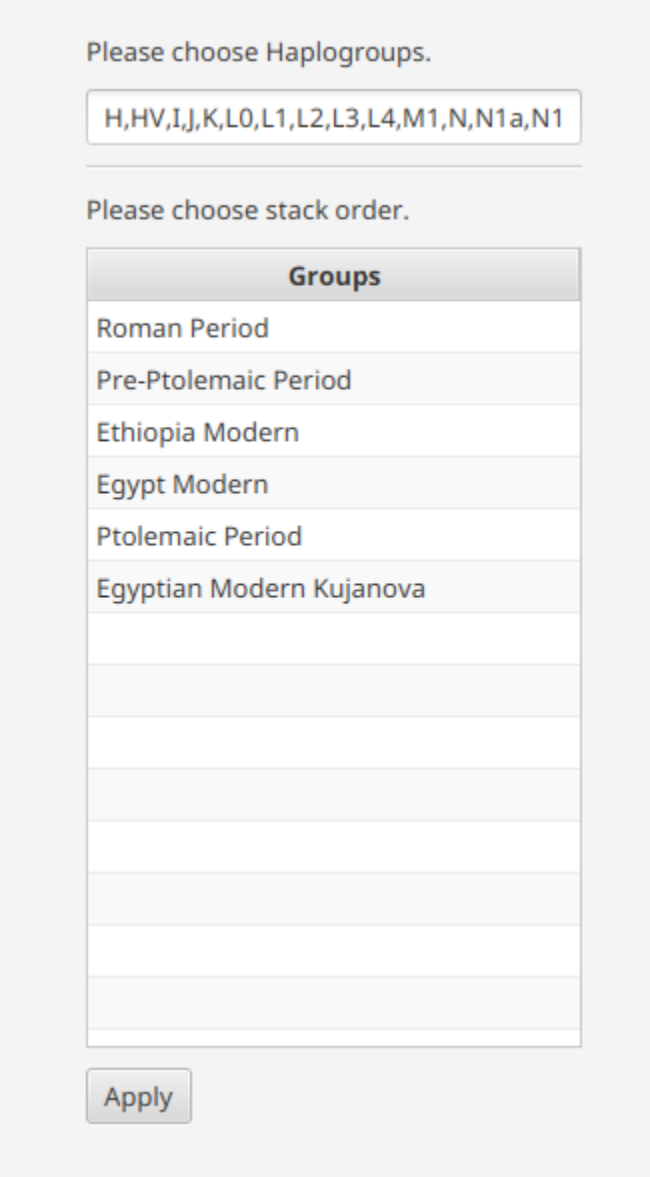
A general visualization of the Haplogroup occurrences in the whole dataset visualized using a simple barplot.



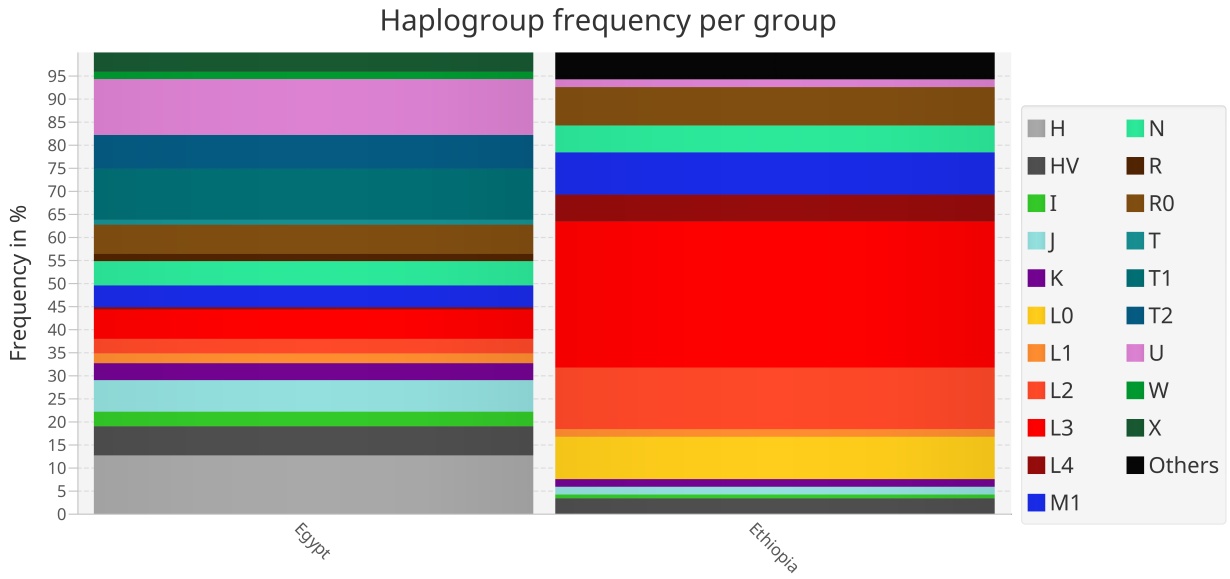
4.5.2 Stacked Barplot

> Visualization -> Haplogroups -> Create Barchart -> Plot Haplogroup frequency per group

The stacked bar plot visualizes the Haplogroup frequency per group. The order of the stacks can be defined by drag-and-dropping the groups in the desired order, e.g. chronological order.



The stacks, which represent the groups, are placed next to each other. This offers a direct comparison of the Haplogroup frequency between the different groups.



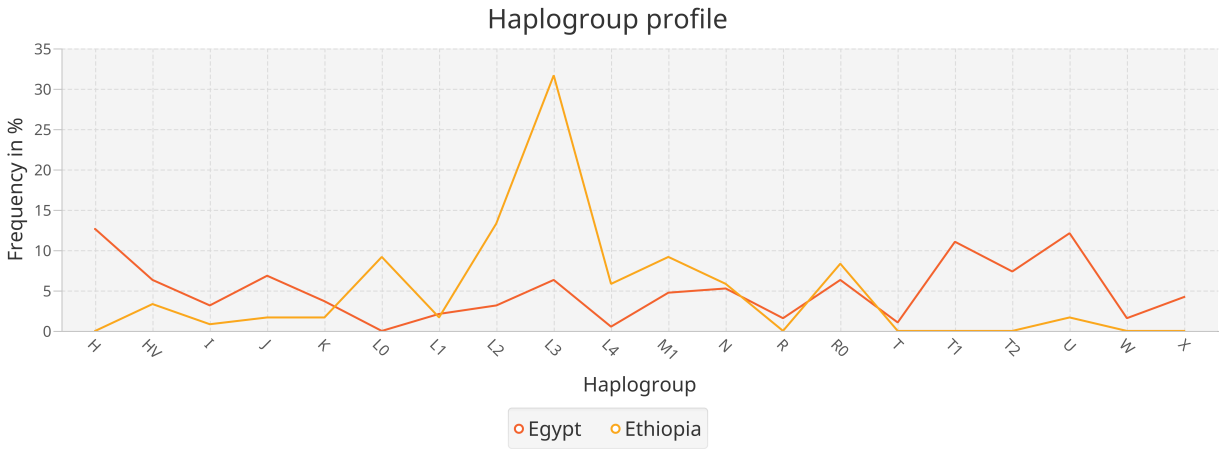
4.5.3 Profile plot

> Visualization -> Haplogroups -> Create Profile plot

The profile plot can visualize the haplogroup profile per group. The x-axis represents the haplogroups, the y-axis the frequency. This enables the comparison of the frequency of one HG in different groups.

In addition, the represented data is provided in a table format next to the visualization panel. Hovering over the rows highlights the corresponding line in the profile plot.

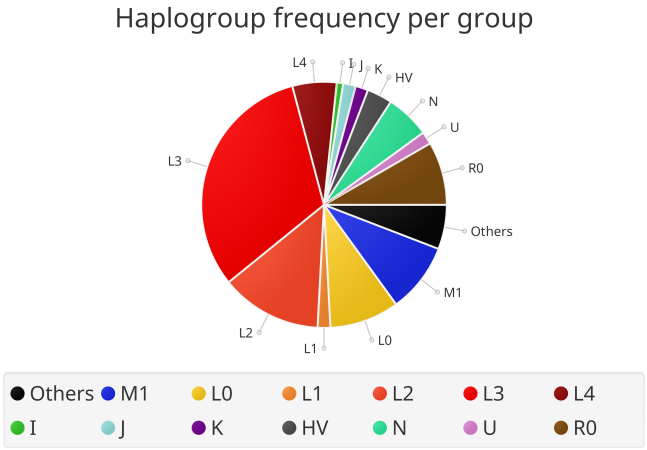
Tabs relying on each other are marked with the same automatically increasing number.



4.5.4 Pie chart

> Visualization -> Haplogroups -> Create Pie Chart

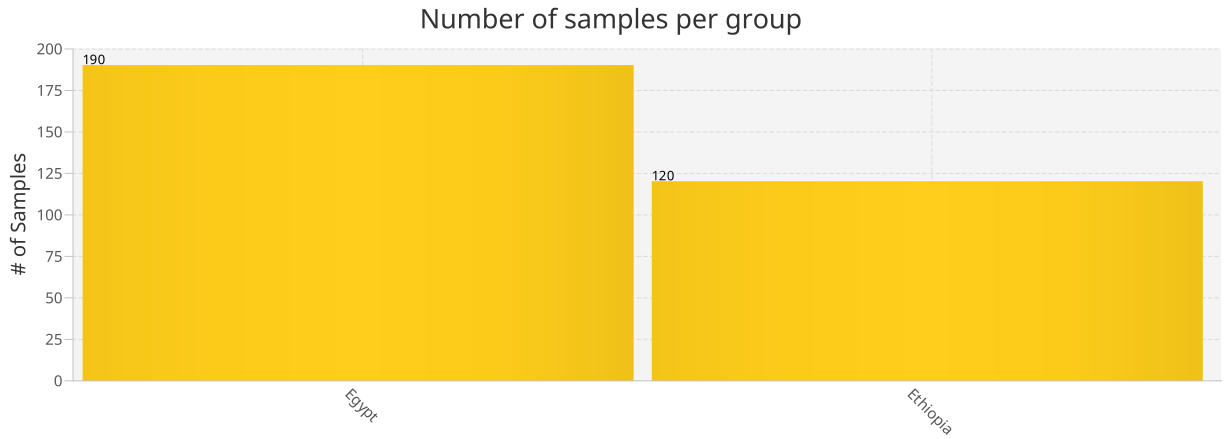
The Haplogroup distribution can also be visualized as a pie chart. This plot can be done on groups and ungrouped data. In case of multiple groups, one pie chart per group is created.



4.5.5 Grouping bar plot

> Visualization -> Grouping -> Grouping bar chart

The grouping bar plot visualizes the sizes of the different groups.



4.5.6 Map view

> Visualization -> Map view -> Visualize data on map

If the samples have some geographic information, they can be visualized on a map. To add all samples to the map, click on the *Add data* button. In case of grouped data, the grouping is represented with different colors, but only up to 8 different colors are supported by now.

4.6 Filtering

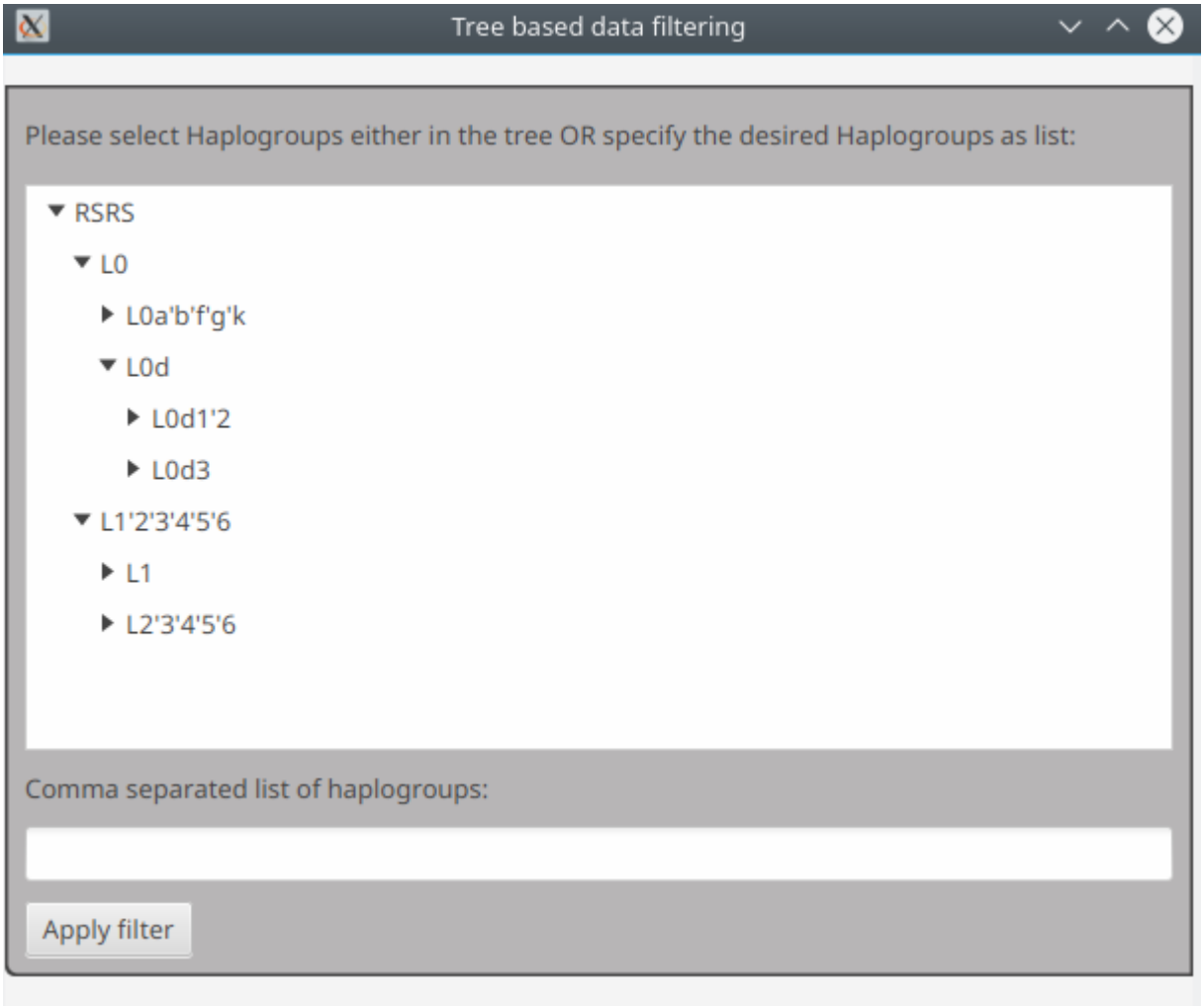
4.6.1 Haplogroup based filtering

> Edit -> Filter data . . . -> . . . based on PhyloTree

To filter out haplogroups that you don't want to include into your analysis, the haplogroup based filtering provides two different ways:

1. The tree-based filtering offers the possibility to select one or multiple Haplogroups via a tree representation of the current version of PhyloTree.

2. Instead of selecting Haplogroups in the tree, you can also enter a comma-separated list of Haplogroups in the text field below the tree representation.



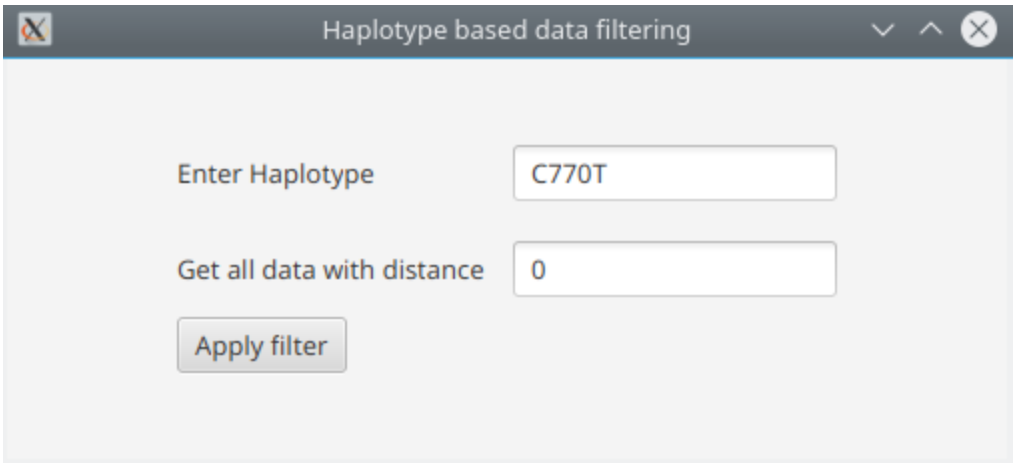
Note: The last filtering step can be redone via *Table -> Reset table*.

4.6.2 Haplotype-based filtering

> Edit -> Filter data . . . -> . . . haplotype filtering

Another way to filter data is the haplotype-based filtering. You can enter one or multiple Haplotypes (as a comma-separated list). This filters out all samples that do not have this haplotype.

You can also specify a distance value d. This will include all Haplotypes with the distance d to the specified ones.



4.7 Grouping

Many analysis tools require grouping of the data. Moreover, grouping allows comparing e.g. different time periods or different locations. mitoBench is flexible in terms of grouping - you can basically define any criterion in the table for grouping.

- **Grouping by column** Select *Grouping -> Group by column* and choose the column that defines the grouping.

Note: After grouping the data, the column name of the column that defines the grouping is extended by the word *(Grouping)*.

4.8 File conversion

mitoBech provides different file conversions via the data *export function*.

For file formats not supported by mitoBench, PDGSpider (for more information see here) is accessible to converted from one file format to another. This tools is only called via mitobench but works completely independent.

4.9 Export

4.9.1 Export data

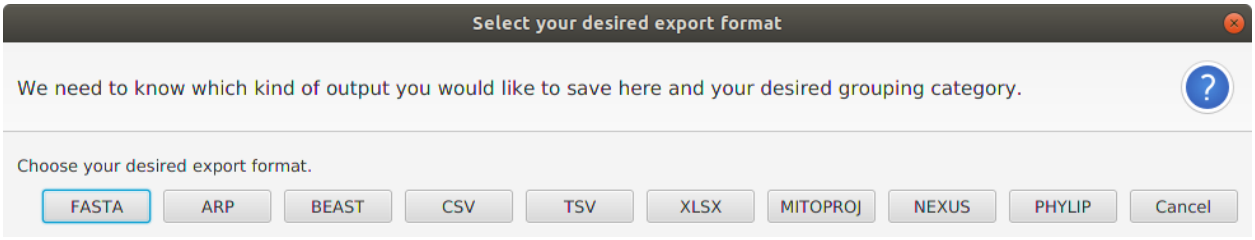
It is possible to export all data shown in the data table via

> File -> Export all Data

Moreover, only a selection can be stored via

> File -> Export selected Data

Both options open an export dialogue where you can choose between different file formats:



The supported file formats are:

- MultiFastA (.fasta)

  Writes all complete MT sequences into a multiFastA file.

- Arlequin format (.arp)

  To export your data to arp format, please specify a column that should be used for grouping. This output file can be used as input file for your analyses with Arlequin.

- BEAST format (.beast)

  This output file can be used as input fasta file for BEAUti / BEAST (*Drummond, Alexei J., et al. "Bayesian phylogenetics with BEAUti and the BEAST 1.7." Molecular biology and evolution 29.8 (2012): 1969-1973.*). The C14 dating will be added to the header line. If your data does not have a C14 dating, the header of the FastA entry will only contain the sample name.

- Generic format (.csv)

  Creates a csv file containing all data displayed in the table.

- Excel format (.xlsx)

  Writes all data as Excel file.

- Nexus format (.nex)

  Writes all data into NEXUS format. This requires aligned sequences. mitoBench only tests whether all sequences have the same length. It's in the user's responsibility to make sure that the sequences are aligned.

- Mitoproject format (.mitoproj)

  Writes all (eventually filtered) data plus grouping information into a text file. This file can later be used to restore the current session/project.

- File conversions via PGDSpider

  For more advanced file conversions, *PGDSpider* can be run directly from mitoBench via *File -> Convert files with PGDSpider*. However, mitoBench currently does not support any file preparing to ease the usage of PGDSpider.

4.9.2 Export images

> File -> Export chart

Each image can be exported individually, either by right-clicking on the chart, which will open a 'Save as png' context menu, or via the File menu (select *File -> Export chart*). This will save the currently displayed visualization as png with a good resolution. Unfortunately, it is not possible to save the figure as a vector graphic so far. The export to pdf also does not result in a higher resolution than the png file.

Note: The chart will be saved with the same aspect ratio displayed in the mitoBench.

4.9.3 Export statistics

> File -> Export statistics

To export calculated statistics, select *File -> Export statistics*. This will write the currently displayed statistics to a comma-separated file.

CHAPTER 5

Database

This database contains currently 22,880 complete modern and ancient published mitochondrial genomes from the 1000 Genome Project and other published sources with the aim of continuous expansion (see all publications here).

5.1 Access

The database can be *accessed* via mitoBench.

5.2 Glossary

In addition to the accession id and the actual sequence, the database contains 72 attributes describing the data such as geographic location, sequencing technologies, and sequence quality information. A complete list describing each attribute in detail including examples can be found here.

5.3 DataValidator

This tool validates the data intended to be uploaded to the database. See DataValidator for more details.

5.4 DataCompleter

DataCompleter is calculating several attributes automatically, such as numbers of Ns in the sequence. Moreover, it completes geographic information. For example, if latitude and longitude of the sampling location is given, it determines the city, country, region, and continent. For more information, see DataCompleter.

Note: DataValidator and DataCompleter can also called directly within mitoBench.

5.5 Future plans:

The user will be able to upload own data to the database via mitoBench. We will provide a template where the data can be filled in (`Download template`) and imported into mitoBench.

CHAPTER 6

How to cite

If you use mitoBench, please cite

Alexander Peltzer, & Judith Neukamm. (2019, March 9). mitobench/MitoBench: mitoBench version 1.1-beta (Version v1.1-beta). Zenodo. http://doi.org/10.5281/zenodo.2588120

Older releases:

Judith Neukamm, & Alexander Peltzer. (2017, November 18). apeltzer/MitoBench: MitoBench Version 0.12.11 (Version v0.12.11). Zenodo. http://doi.org/10.5281/zenodo.1058980

The project URL is:

https://github.com/mitobench/MitoBench

6.1 Tools & Methods

- 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature, 526(7571), p.68.
- Excoffier, L. and Lischer, H.E., 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular ecology resources, 10(3), pp.564-567.
- Hudson, R.R., Slatkin, M. and Maddison, W.P., 1992. Estimation of levels of gene flow from DNA sequence data. Genetics, 132(2), pp.583-589.
- Lischer, H.E. and Excoffier, L., 2011. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics, 28(2), pp.298-299.
- Reynolds, John, Bruce S. Weir, and C. Clark Cockerham. "Estimation of the coancestry coefficient: basis for a short-term genetic distance." Genetics 105.3 (1983): 767-779.
- Slatkin, Montgomery. "A measure of population subdivision based on microsatellite allele frequencies." Genetics 139.1 (1995): 457-462.
- van Oven, M., 2015. PhyloTree Build 17: Growing the human mitochondrial DNA tree. Forensic Science International: Genetics Supplement Series, 5, pp.e392-e394.

- Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.J., Kronenberg, F., Salas, A. and Schönherr, S., 2016. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic acids research, 44(W1), pp.W58-W63.

CHAPTER 7

Indices and tables

- genindex
- modindex
- search