iTaxoTools - tools for integrative taxonomy

Addendum to iTaxoTools manual: Hapsolutely and ConvPhase

Version: 7 November 2023

The code of Hapsolutely and ConvPhase is available at:
https://github.com/iTaxoTools/ConvPhaseGui // https://github.com/iTaxoTools/Hapsolutely

Both the previous and new tools can be downloaded from the project's website http://itaxotools.org which also features a section with useful links, news, and FAQs.

How to cite: When using one of the programs included in the iTaxoTools 0.1.1. release in your study, please cite the main paper as follows.

Vences, M., Miralles, A., Brouillet, S., Ducasse, J., Fedosov, A., Kharchev, V., Kumari, S, Patmanidis, S., Puillandre, N., Scherz, M. D., Kostadinov, I., Renner, S. S. (2021). iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists. *Megataxa* **6**: 77-92.

For ConvPhase and Hapsolutely, we recommend citing also the original paper, and where appropriate the papers that introduced the code of Phase and SeqPhase, as well as the Fitchi approach, and the network reconstruction algorithms from PopArt:

Vences, M., Patmanidis, S., Schmidt, J.-C., Matschiner, M., Miralles, A. & Renner, S.S. (2024). Hapsolutely: a user-friendly tool integrating haplotype phasing, network construction and haploweb calculation. #####

Flot, J.F. (2010). seqphase: a web tool for interconverting phase input/output files and fasta sequence alignments. Mol Ecol Resour 10: 162-166.

Flot J.F., Couloux A. & Tillier S (2010). Haplowebs as a graphical tool for delimiting species: a revival of Doyle's "field for recombination" approach and its application to the coral genus *Pocillopora* in Clipperton. BMC Evol Biol 10: 372.

Leigh J.W. & Bryant D. (2015). popart: full-feature software for haplotype network construction. Method Ecol. Evol., 6, 1110–1116.

Matschiner, M. (2016). Fitchi: haplotype genealogy graphs based on the Fitch algorithm. Bioinformatics 32: 1250–1252.

Miralles, A., Ducasse, J. Brouillet, S., Flouri, T., Fujisawa, T., Kapli, P., Knowles, L. L., Kumari, S., Stamatakis, A., Sukumaran, J., Lutteropp, S., Vences, M. & Puillandre, N. (2021). SPART, a versatile and standardized data exchange format for species partition information. Molecular Ecology Resources 22: 430-438.

Stephens, M., Smith, N.J. & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978–989.

Disclaimer: The programs included in iTaxoTools are free software. All code specifically programmed for iTaxoTools can be redistributed and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.
All tools not specifically programmed for iTaxoTools but constituting a modification or extension of the original code are also free software and most of them licensed under GNU v3, but partly, other licences apply.

We welcome all suggestions, comments, questions and bug reports. Please use the GitHub platform to submit those: for this, you just need to sign up at GitHub (a quick and easy procedure), navigate to the respective repository (see table with links below), and create "New Issue". The team of developers regularly checks all the issues and will deal with them asap (bug reports will be treated as priority).

1. ConvPhase

Example files provided:

ConvPhase_examplefile1.tsv
Convphase_examplefile1_regular_fasta.fas
Convphase_examplefile1_MolDformat_fasta.fas
Convphase_examplefile1_HapViewformat_fasta.fas

ConvPhase_examplefile2.tsv
Convphase_examplefile2_regular_fasta.fas
Convphase_examplefile2_MolDformat_fasta.fas
Convphase_examplefile2_HapViewformat_fasta.fas

All example files are nucleotide sequences. They can be inspected in a text editor. Example file 1 (in its different flavors) is a very simple file as shown in the explanations below. Example file 2 is a more complex and realistic file with a true example of sequences of the RAG-1 gene for 37 individuals of *Gephyromantis* frogs.

The tsv file is in tab-delimited text format and includes sequence identifier and optionally taxon identifier, as well as the (aligned) sequence.

The remaining example files are in fasta format, either in a generic format that includes only the sequence identifier in the sequence description line, or in specific formats where sequence identifier is separated by a pipe or a period character from the taxon identifier.

To test the functions of ConvPhase, import any of these files, and run the program either with default settings or with custom settings.

ConvPhase is a tool to apply the Phase algorithm (Stephens et al. 2001) to separate alleles in DNA sequences of diploid organisms. Sanger sequencing and some assembly strategies of other sequencing techniques produce consensus sequences where heterozygote positions are coded with IUPAC ambiguity codes. As soon as a stretch of nucleotides includes two or more ambiguity codes, the separation (phasing) process is not straightforward and requires specific algorithms such as Phase.

However, in order to use the original Phase code, nucleotide alignments must be reformatted, for which Flot (2010) proposed a specific program, SeqPhase.

ConvPhase combines the code of both Phase and SeqPhase in order to offer a more convenient means for the process of phasing sequences from multi-individual alignments.

Input/output formats

ConvPhase accepts various different input formats, in particular, tab-delimited text (.tsv) files, and different flavors of fasta files.

In each format, all sequences in an input file must be aligned (i.e., all sequences must be of the same length).

For the program to work properly, no missing data, gaps, or ambiguity codes coding for more than two bases should be included (i.e., only Y, R, W, S, K, M are allowed, no D, V, H, or B, and no N, ? or -). If one or several sequences include these not allowed characters, phasing may still proceed but could lead to errors and unreliable output.

When opening an input file, ConvPhase tries to automatically assess its format and parse the respective identifiers and sequences. In case of errors in the input stage, verify the format of the file and its file extension (which should be either .tab / .tsv, or .fas / .fasta).

While the program is of course able to process generic alignments in fasta format with just the sample or specimen numbers in the sequence description line, for subsequent processing e.g. in the construction of haplotype networks or genealogies, ConvPhase can also recognize and transfer to the output files information on species, population or any other grouping category. In this context, the following terminology is used: *sequence identifier* is for instance an isolate or specimen-voucher number, whereas *taxon identifier* is the category that later will be used for grouping sequences into subsets, for instance species, population, or geographical location.

tab-separated value file (.tsv or .tab). The following example shows the structure of this format. The seqid column provides the sequence identifier, the species column provides the taxon identifier. The columns can also be named otherwise, in which case the user must later specify which column is to be used as sequence and taxon identifier, respectively.

To prevent errors, sequence and taxon identifiers must not include spaces or special characters (only use underscores and regular alphanumeric characters)

seqid	species	sequence
sample1	Mantella_aurantiaca	ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
sample2	Mantella_aurantiaca	ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
sample3	Mantella_aurantiaca	ACGTYTTACATRCTTAAGCACGACTTAGCTAGTAATTCCCG
sample4	Mantella_crocea	ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
sample5	Mantella_crocea	ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG

Regular fasta (.fas, .fasta). The following example shows the same data as in the tsv file example above. In this case, only the sequence identifiers (here: sample numbers) are given in the sequence description line.

```
>sample1
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample2
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample3
ACGTYTTACATRCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample4
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample5
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
```

Fasta in MolD format (.fas, .fasta). The program MolD, written by A. Fedosov for molecular diagnosis, introduced a format where a pipe character (|) separates sequence identifier and taxon identifier. Keep in mind that no spaces or special characters are allowed in either identifier (only use underscores).

```
>sample1|Mantella_aurantiaca
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample2|Mantella_aurantiaca
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample3|Mantella_aurantiaca
ACGTYTTACATRCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample4|Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample5|Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
```

Fasta in HapView format (.fas, .fasta). The program HapView, written by G. Ewing for the construction of haplotype genealogies, uses a format where a period (.) separates sequence identifier and taxon identifier. Keep in mind that no spaces or special characters are allowed in either identifier (only use underscores). If you plan to use the phased sequences for HapView (=Haplotype Viewer) rather than visualize haplotype genealogies in Hapsolutely, we recommend keeping both sequence and taxon identifiers relatively short

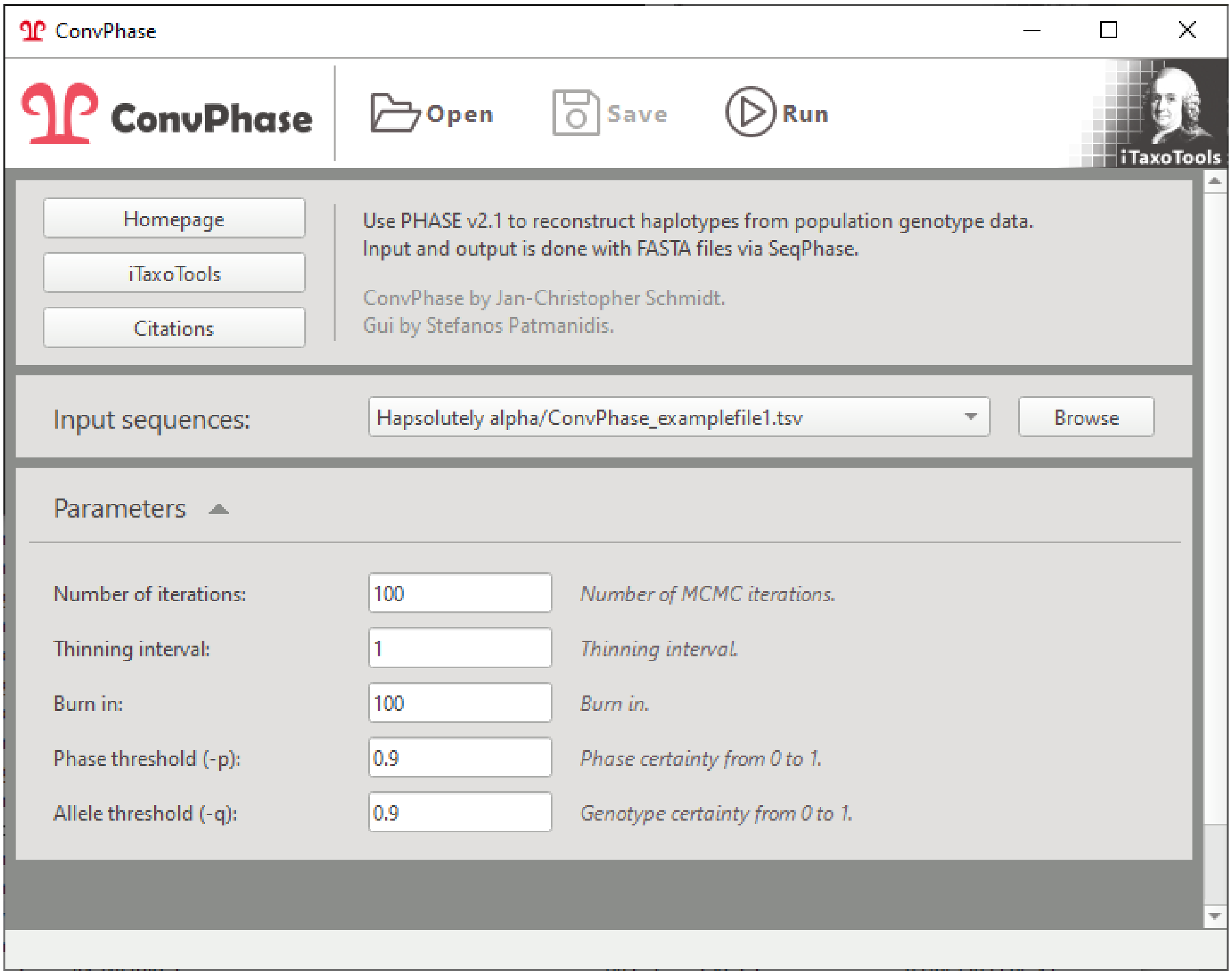
```
>sample1.Mantella_aurantiaca
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample2.Mantella_aurantiaca
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample3.Mantella_aurantiaca
ACGTYTTACATRCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample4.Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample5.Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
```

Running ConvPhase

The program is distributed as standalone executable for Windows and Mac. It should run in different Windows environments, including Windows 10 and Windows 11.

Upon pressing either the "Open" button in the upper row of icons, or the "Browse" button in the field "Input sequences", you can select your input file. In the following screenshot, the file "ConvPhase_examplefile1.tsv" has been selected.

You can now phase your sequences either with default parameter settings, or adjust the parameters as desired.



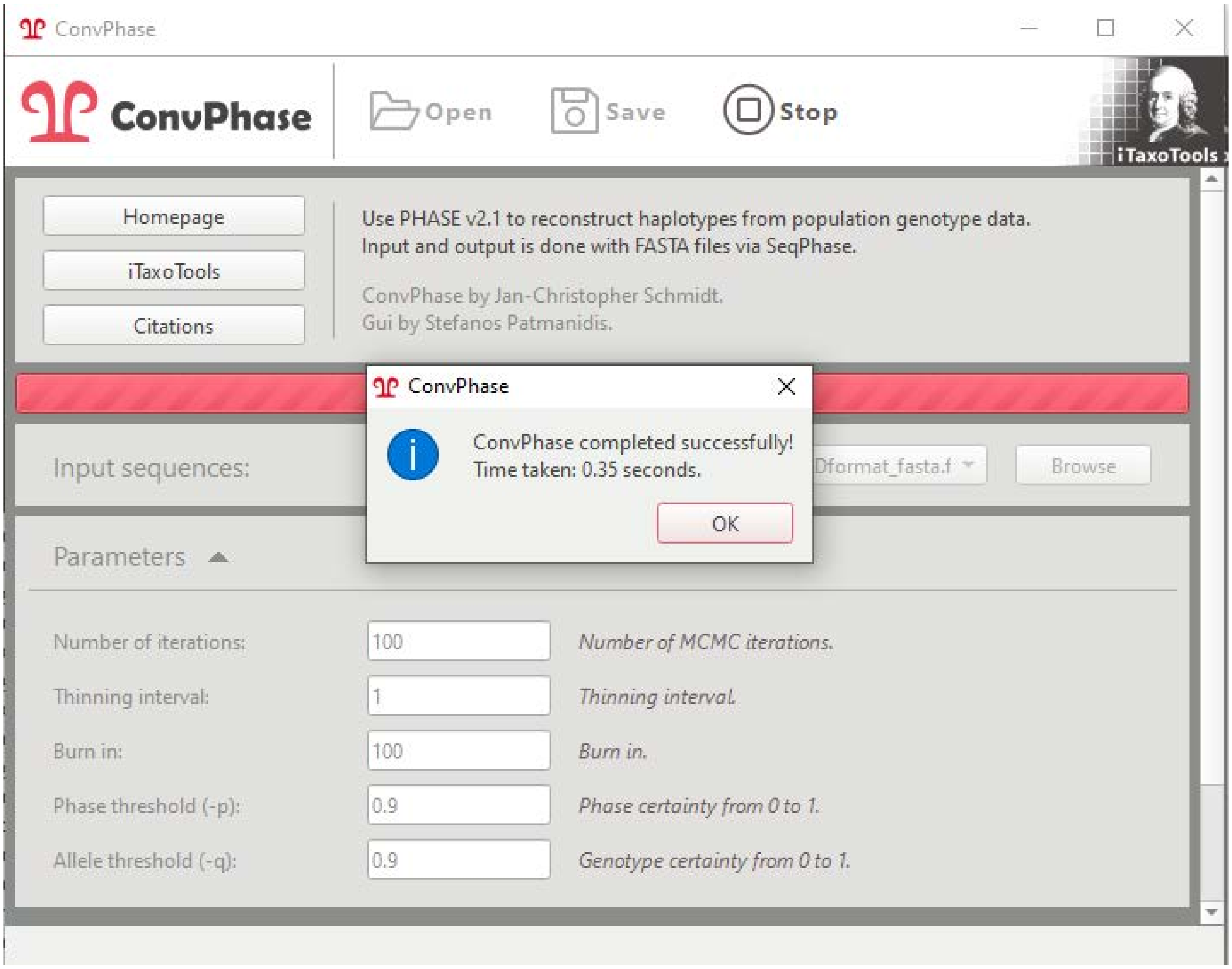
Use PHASE v2.1 to reconstruct haplotypes from population genotype data.
Input and output is done with FASTA files via SeqPhase.

ConvPhase by Jan-Christopher Schmidt.
Gui by Stefanos Patmanidis.

Input sequences: Hapsolutely alpha/ConvPhase_examplefile1.tsv

Number of iterations:	100	*Number of MCMC iterations.*
Thinning interval:	1	*Thinning interval.*
Burn in:	100	*Burn in.*
Phase threshold (-p):	0.9	*Phase certainty from 0 to 1.*
Allele threshold (-q):	0.9	*Genotype certainty from 0 to 1.*

After phasing is completed, an announcement appears. In the case of example file 1, the program performs successfully with default parameter settings.



Use PHASE v2.1 to reconstruct haplotypes from population genotype data.
Input and output is done with FASTA files via SeqPhase.

ConvPhase by Jan-Christopher Schmidt.
Gui by Stefanos Patmanidis.

ConvPhase

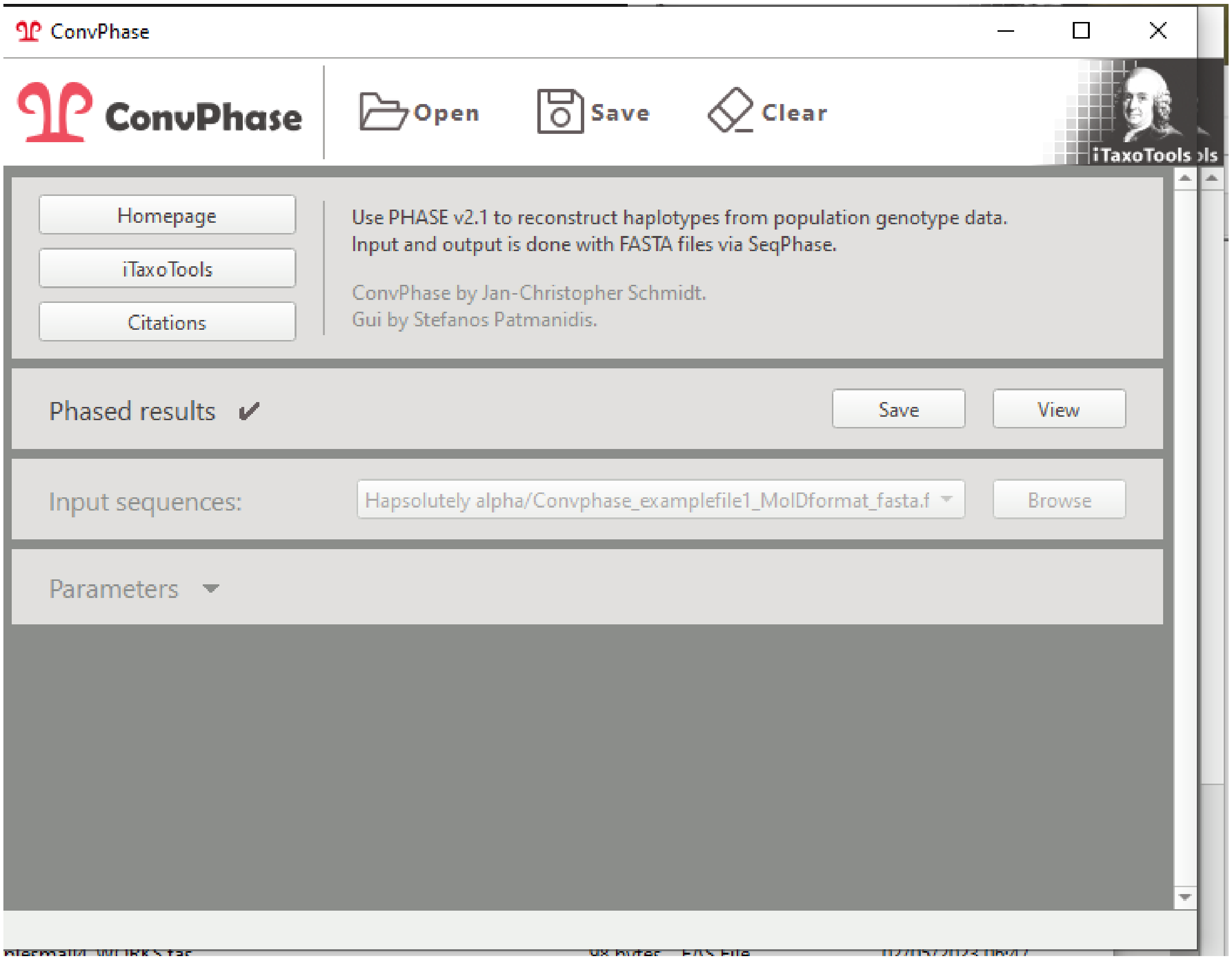
ConvPhase completed successfully!
Time taken: 0.35 seconds.

OK

Input sequences: Dformat_fasta.f

```
Number of iterations:   100   Number of MCMC iterations.
Thinning interval:      1     Thinning interval.
Burn in:                100   Burn in.
Phase threshold (-p):   0.9   Phase certainty from 0 to 1.
Allele threshold (-q):  0.9   Genotype certainty from 0 to 1.
```

Subsequently, you can either view and inspect the phased sequence file using the "View" button, or save it using either of the two "Save" buttons (in the upper row of icons, or in the "Phased Results" field).



Use PHASE v2.1 to reconstruct haplotypes from population genotype data.
Input and output is done with FASTA files via SeqPhase.

ConvPhase by Jan-Christopher Schmidt.
Gui by Stefanos Patmanidis.

Phased results ✔

Input sequences: Hapsolutely alpha/Convphase_examplefile1_MolDformat_fasta.f

After phasing, the output consists of two sequences for each original sequence, corresponding to the two alleles on the two homologous chromosomes. If the original sequence had no heterozygous positions, both output sequences will be identical. As a default, the format of the output file will be in the same format as the input file (but this can be selected before phasing). The two alleles will be named as a and b.

In tsv format, the allele information will be provided as separate column, as follows:

seqid	**species**	**allele**	**sequence**
sample1	Mantella_aurantiaca	a	ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
sample1	Mantella_aurantiaca	b	ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
sample2	Mantella_aurantiaca	a	ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
sample2	Mantella_aurantiaca	b	ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
sample3	Mantella_aurantiaca	a	ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
sample3	Mantella_aurantiaca	b	ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
sample4	Mantella_crocea	a	ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
sample4	Mantella_crocea	b	ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
sample5	Mantella_crocea	a	ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
sample5	Mantella_crocea	b	ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG

In regular fasta format, the allele information will be appended to the sequence identifiers, preceded by an underscore.

```
>sample1_a
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample1_b
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample2_a
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample2_b
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample3_a
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample3_b
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample4_a
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample4_b
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample5_a
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample5_b
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
```

If the input was a fasta file in MolD or HapView format, the allele information will equally be appended to the sequence identifiers, preceded by an underscore, and before the character (pipe or period) separating the sequence and taxon identifier.

```
>sample1_a|Mantella_aurantiaca
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample1_b|Mantella_aurantiaca
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample2_a|Mantella_aurantiaca
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample2_b|Mantella_aurantiaca
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample3_a|Mantella_aurantiaca
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample3_b|Mantella_aurantiaca
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample4_a|Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample4_b|Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample5_a|Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample5_b|Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
```

```
>sample1_a.Mantella_aurantiaca
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample1_b.Mantella_aurantiaca
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample2_a.Mantella_aurantiaca
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample2_b.Mantella_aurantiaca
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample3_a.Mantella_aurantiaca
ACGTCTTACATGCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample3_b.Mantella_aurantiaca
ACGTTTTACATACTTAAGCACGACTTAGCTAGTAATTCCCG
>sample4_a.Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample4_b.Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample5_a.Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
>sample5_b.Mantella_crocea
ACGTATTACATCCTTAAGCACGACTTAGCTAGTAATTCCCG
```

For more detailed information on the Phase algorithm, consult the original paper of Stephens et al. (2001) (see citation on first page of this manual) or the manual of the Phase program: https://stephenslab.uchicago.edu/phase/phasefaq.html

In general, the two threshold values (Phase threshold and Allele threshold) determine the probabilities at which the program phases a sequence. By default, these values are set quite stringent (0.9) which means, only if a heterozygote sequence can be separated into the two corresponding alleles with a probability of 0.9 or higher, phasing will proceed. Otherwise, the program will produce an error message. ConvPhase will still produce an output file where sequences may be partially phased, but this file might lead to erroneous inferences in downstream analyses. This will happen with Example file 2.

In such cases, if you require a fully reliable phasing of every heterozygote position for a fully reliable haplotype genealogy, you need to determine and add more homozygote sequences, or single alleles obtained by cloning or single-molecule sequencing to the dataset before phasing.

If instead your goal is to obtain tentative alleles for visualizing or quantifying major differences and assessing the presence or absence of haplotype sharing between taxa, consider lowering the threshold values to 0.5 or below.You can test this with example file 2.

2. Hapsolutely

Example files provided:

ConvPhase_examplefile1.tsv
Convphase_examplefile1_regular_fasta.fas
Convphase_examplefile1_MolDformat_fasta.fas
Convphase_examplefile1_HapViewformat_fasta.fas

ConvPhase_examplefile2.tsv
Convphase_examplefile2_regular_fasta.fas
Convphase_examplefile2_MolDformat_fasta.fas
Convphase_examplefile2_HapViewformat_fasta.fas

[these are the same example files also used for ConvPhase]

Hapsolutely_examplefile3_regular_fasta.fas
Hapsolutely_examplefile3_regular_fasta_phased.fas
Hapsolutely_examplefile3_spartitions.spart

Hapsolutely_examplefile4_regular_fasta_long.fas
Hapsolutely_examplefile4_regular_fasta_short.fas
Hapsolutely_examplefile4_spartition.tsv
Hapsolutely_examplefile4_spartitions.spart

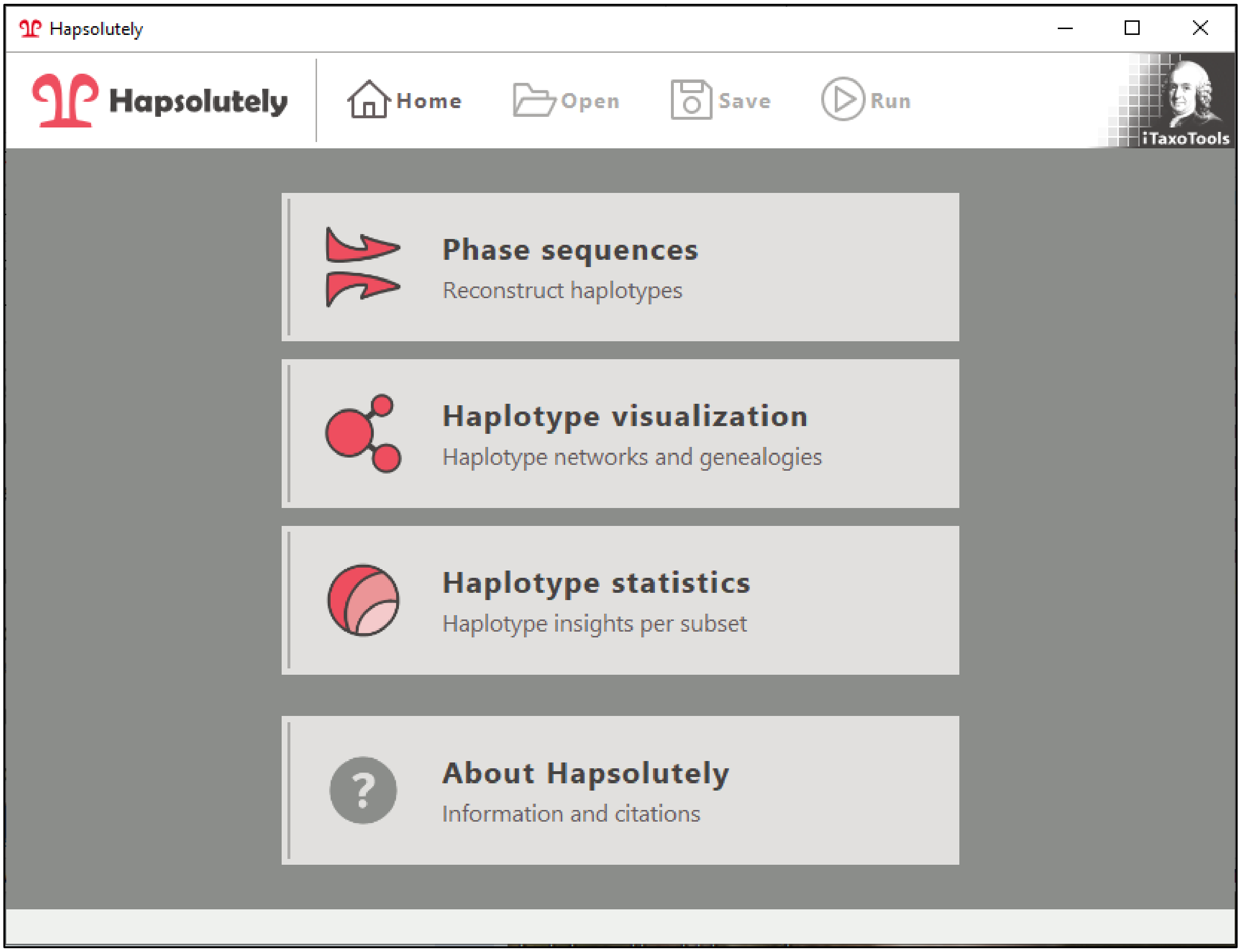
Hapsolutely_examplefile5_haploidseqs_regular_fasta.fas
Hapsolutely_examplefile5_spartition.tsv

Hapsolutely includes an implementation of ConvPhase and therefore, the same example files 1 and 2 introduced for ConvPhase can be used. In addition, example files 3 and 4 are here introduced to illustrate how different species partitions can be coded and visualized by Hapsolutely. Example file 4 contains mitochondrial sequences without heterozygote positions which can be used to directly reconstruct a network or a genealogy without phasing.

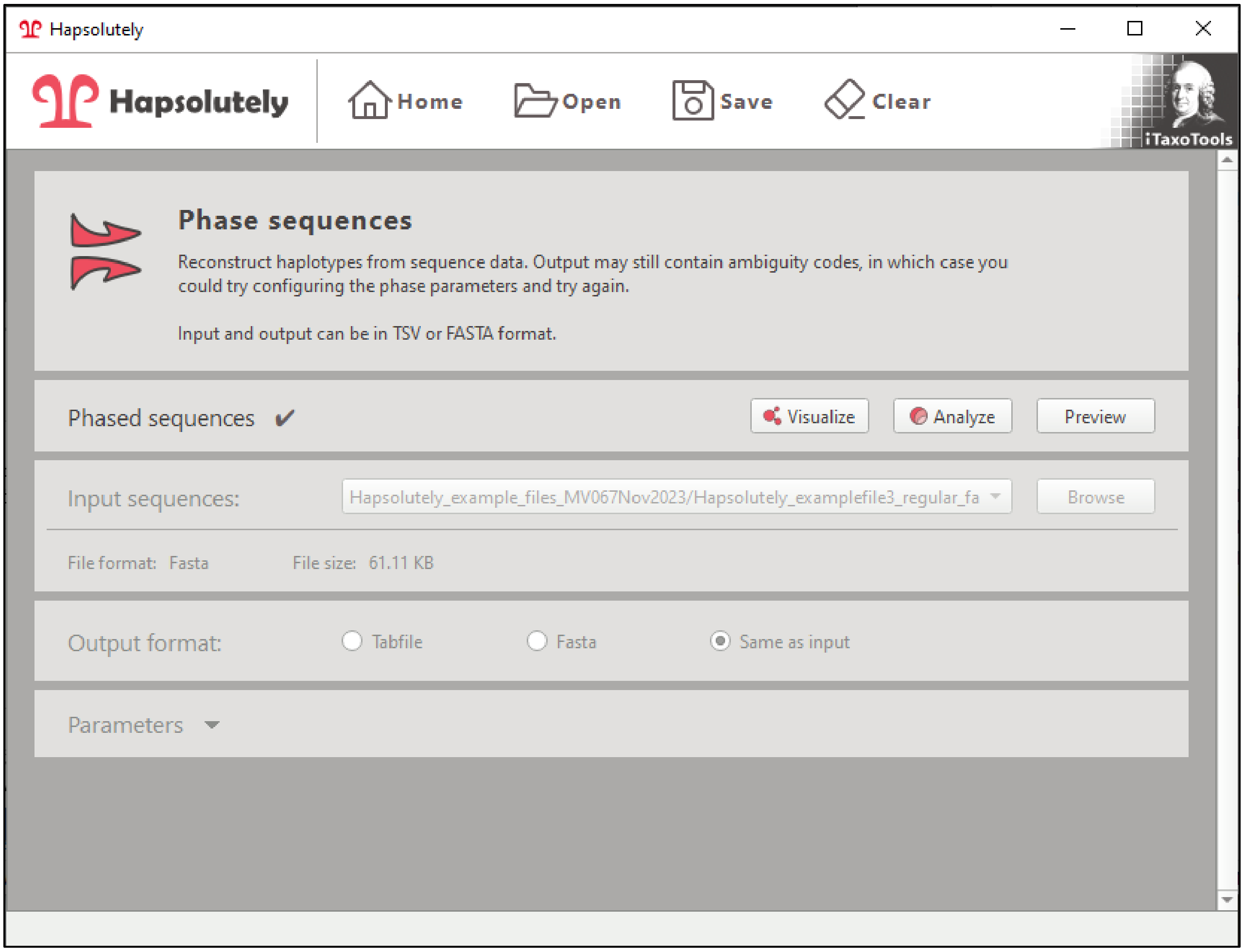
Hapsolutely is a comprehensive program that includes ConvPhase to phase heterozygous sequences, and can directly produce haplotype genealogies and networks, as well as various statistics focused especially on haplotype sharing and fields of recombination between subsets of sequences.

The program integrates original code from Phase and SeqPhase (in the phasing module) (Stephens et al. 2001; Flot 2010), from PopArt (for haplotype network reconstruction using different approaches) (Leigh & Bryant 2015), and from Fitchi (for reconstruction of haplotype genealogies using the Fitch approach) (Matschiner 2016). It also infers fields of recombination (Flot et al. 2010) and is able to use species partitions from SPART files (Miralles et al. 2021).

The starting (home) window of Hapsolutely consists of four main buttons which can be used to start the different analysis modules: (i) phasing, (ii) network/genealogy reconstruction, and (iii) statistics. A fourth button provides general information on the program. Furthermore, the upper row of icons provide access to routine functions such as Open input files, Save results, Clear data for starting a new analysis, and move back to the Home window.



The phasing module is identical to ConvPhase. As an added function, after completion of phasing there are direct links to Visualize or Analyze the phased data.



Phase sequences

Reconstruct haplotypes from sequence data. Output may still contain ambiguity codes, in which case you could try configuring the phase parameters and try again.

Input and output can be in TSV or FASTA format.

Phased sequences ✔

Input sequences: Hapsolutely_example_files_MV067Nov2023/Hapsolutely_examplefile3_regular_fa

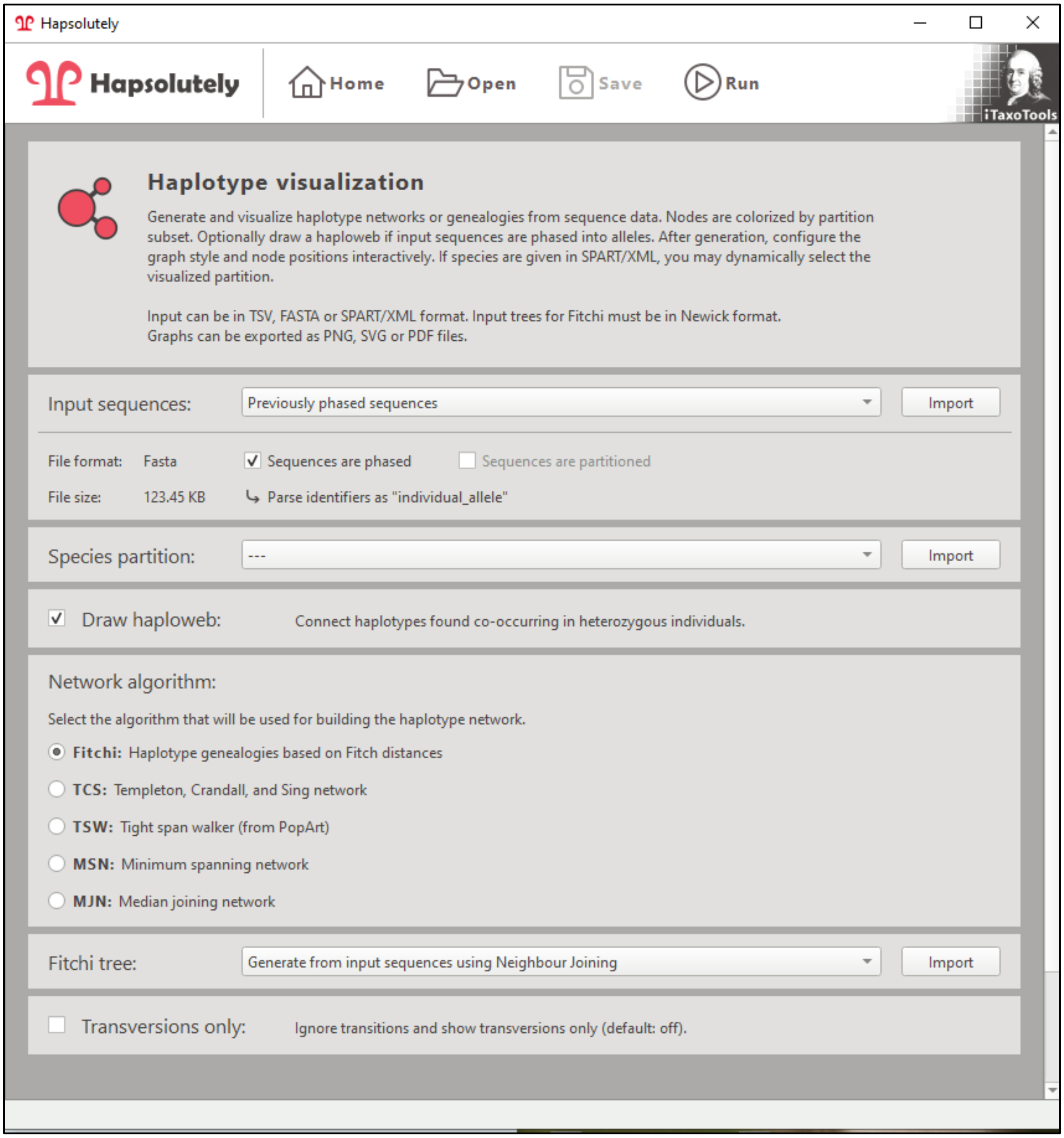
File format: Fasta File size: 61.11 KB

Output format: ○ Tabfile ○ Fasta ● Same as input

Clicking the Visualize button (or the Haplotype networks button on the home window) opens the window with the dialogue for inferring haplotype networks or haplotype genealogies using different approaches.

The field "Species partition" determines which information is used for specimen and taxon identifiers. If the original input file had such information (e.g., as tsv or MolD-fasta format), then the program suggests parsing this information. The "Species" information typically corresponds to the taxon identifier and will be used to color the "bubbles" representing the alleles.

As default, the program will use the Fitch algorithm for reconstructing haplotype genealogies (see Matschiner 2016). For this, typically a user tree inferred under Maximum Likelihood should be imported. Instead, the program also allows generation of the tree under Maximum Parsimony or, for quick data exploration, with Neighbor-joining. It is also possible to specify that only transversions should be represented in the genealogy.



Haplotype visualization

Generate and visualize haplotype networks or genealogies from sequence data. Nodes are colorized by partition subset. Optionally draw a haploweb if input sequences are phased into alleles. After generation, configure the graph style and node positions interactively. If species are given in SPART/XML, you may dynamically select the visualized partition.

Input can be in TSV, FASTA or SPART/XML format. Input trees for Fitchi must be in Newick format.
Graphs can be exported as PNG, SVG or PDF files.

Input sequences: Previously phased sequences ▾ Import

File format: Fasta ☑ Sequences are phased □ Sequences are partitioned

File size: 123.45 KB ↳ Parse identifiers as "individual_allele"

Species partition: --- ▾ Import

☑ Draw haploweb: Connect haplotypes found co-occurring in heterozygous individuals.

Network algorithm:

Select the algorithm that will be used for building the haplotype network.

● **Fitchi:** Haplotype genealogies based on Fitch distances

○ **TCS:** Templeton, Crandall, and Sing network

○ **TSW:** Tight span walker (from PopArt)

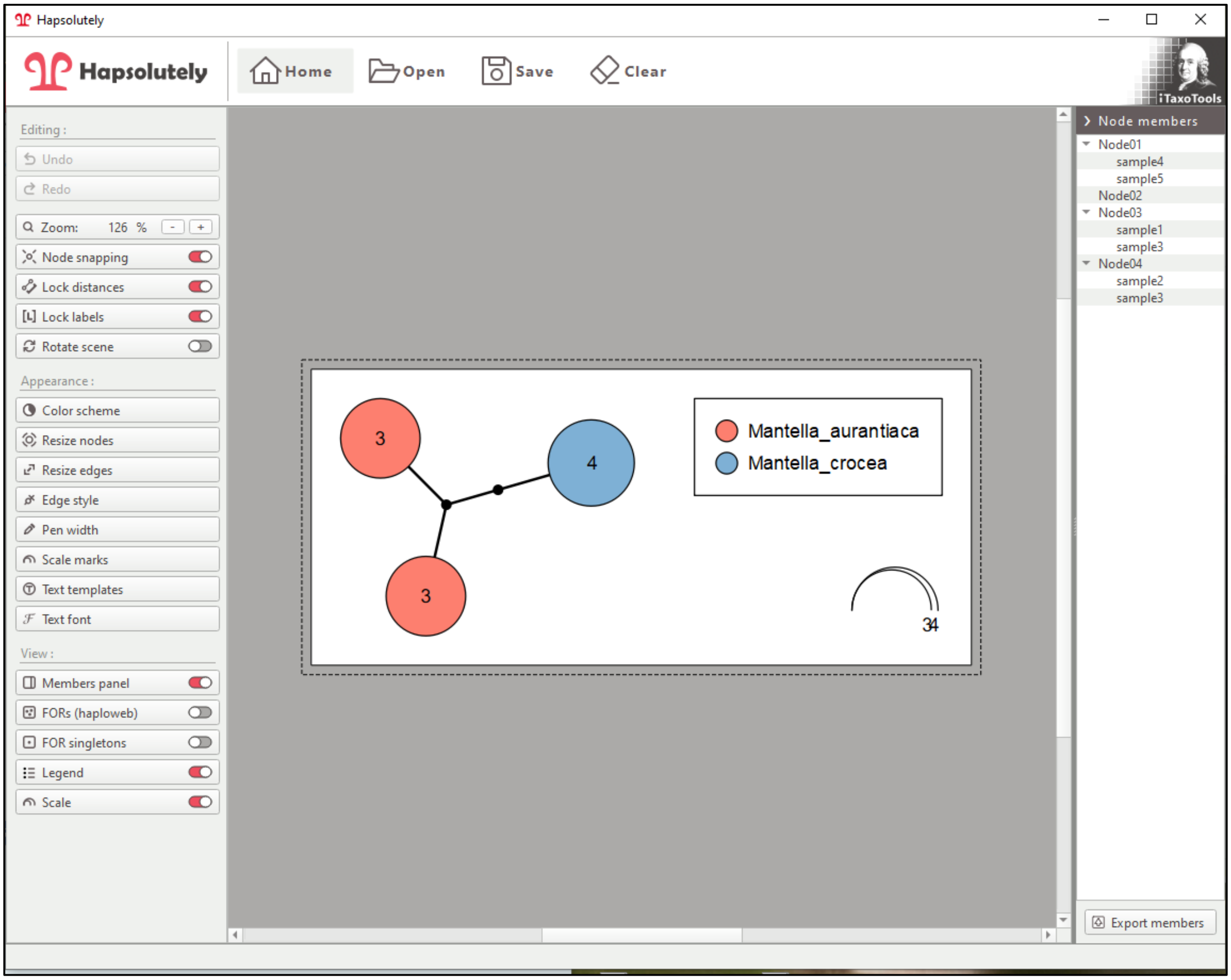
○ **MSN:** Minimum spanning network

○ **MJN:** Median joining network

Fitchi tree: Generate from input sequences using Neighbour Joining ▾ Import

□ Transversions only: Ignore transitions and show transversions only (default: off).

The output will be shown in the network viewer which provides multiple interactive options to adjust details of the visualization. Some of these will be explained in the following pages.



Node01
- sample4
- sample5

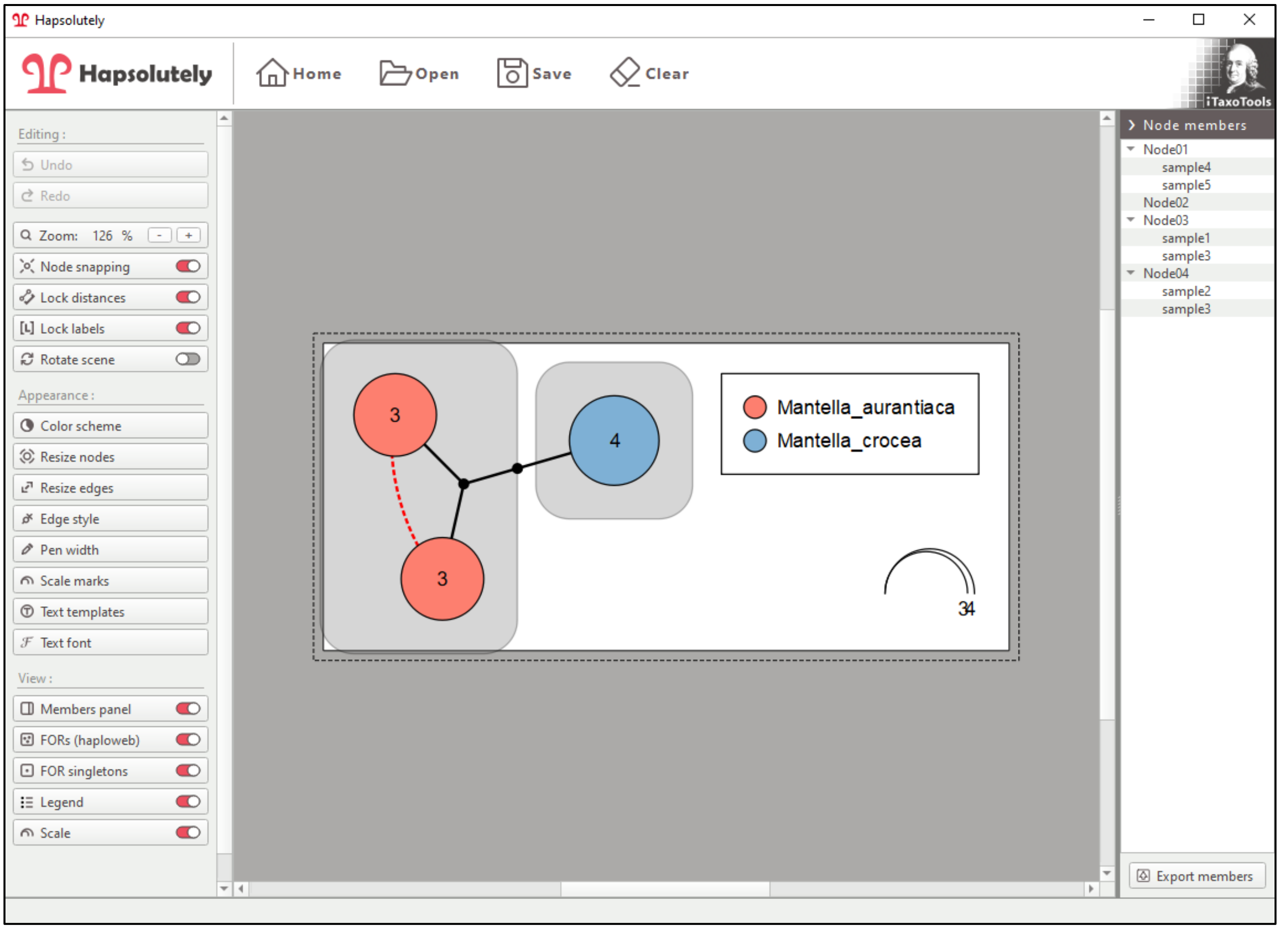
Node02

Node03
- sample1
- sample3

Node04
- sample2
- sample3

A unique option of Hapsolutely is to represent fields of recombination (FORs) and alleles connected by heterozygous sequences/individuals (Flot et al. 2010). The option to compute FORs can be specified before inferring the network / genealogy, but can also be activated/deactivated in the viewer.

The red dotted lines connect alleles that are found in a single heterozygous sequence. In the example file 1, the sequence "sample3" is heterozygous and after phasing, consists of the two alleles represented in red; therefore these two allele "bubbles" are connected. The entire set of alleles connected by heterozygous individuals represents one field of recombination (FOR) and is shown in a gray box. Users can choose not to display boxes for the singleton FORs (i.e., a single allele not connected to any other allele by heterozygous sequences), or switch off FOR visualization altogether.



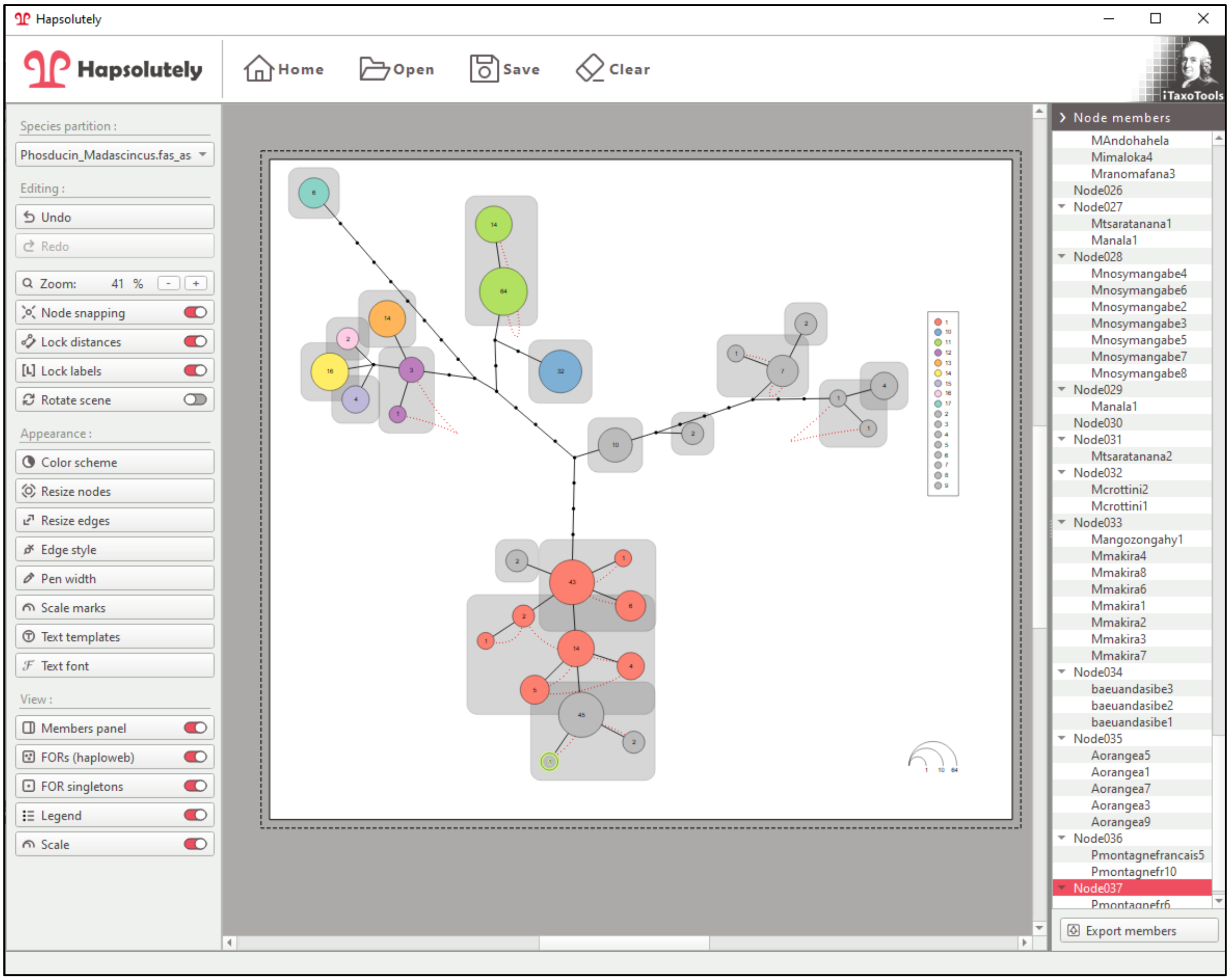
Zoom: 126 %

Node snapping ●
Lock distances ●
Lock labels ●
Rotate scene ○
Members panel ●
FORs (haploweb) ●
FOR singletons ●
Legend ●
Scale ●

Node members
- Node01
  - sample4
  - sample5
- Node02
- Node03
  - sample1
  - sample3
- Node04
  - sample2
  - sample3

The species partition information can also be provided from an additional file. For this, under "Species partition", either a tsv or a SPART file can be imported. The tsv file simply contains two columns containing the sequence identifier (which must coincide with that in the sequence file) and the name of the subset (= species, taxon). See examplefile 3. Additional columns may be present in the tsv, but then users must make sure to select the right ones under "Species" and "Individuals.

The following screenshot shows the haplotype genealogy of example file 3. These are sequences of a group of lizards of the genus *Madascincus*.



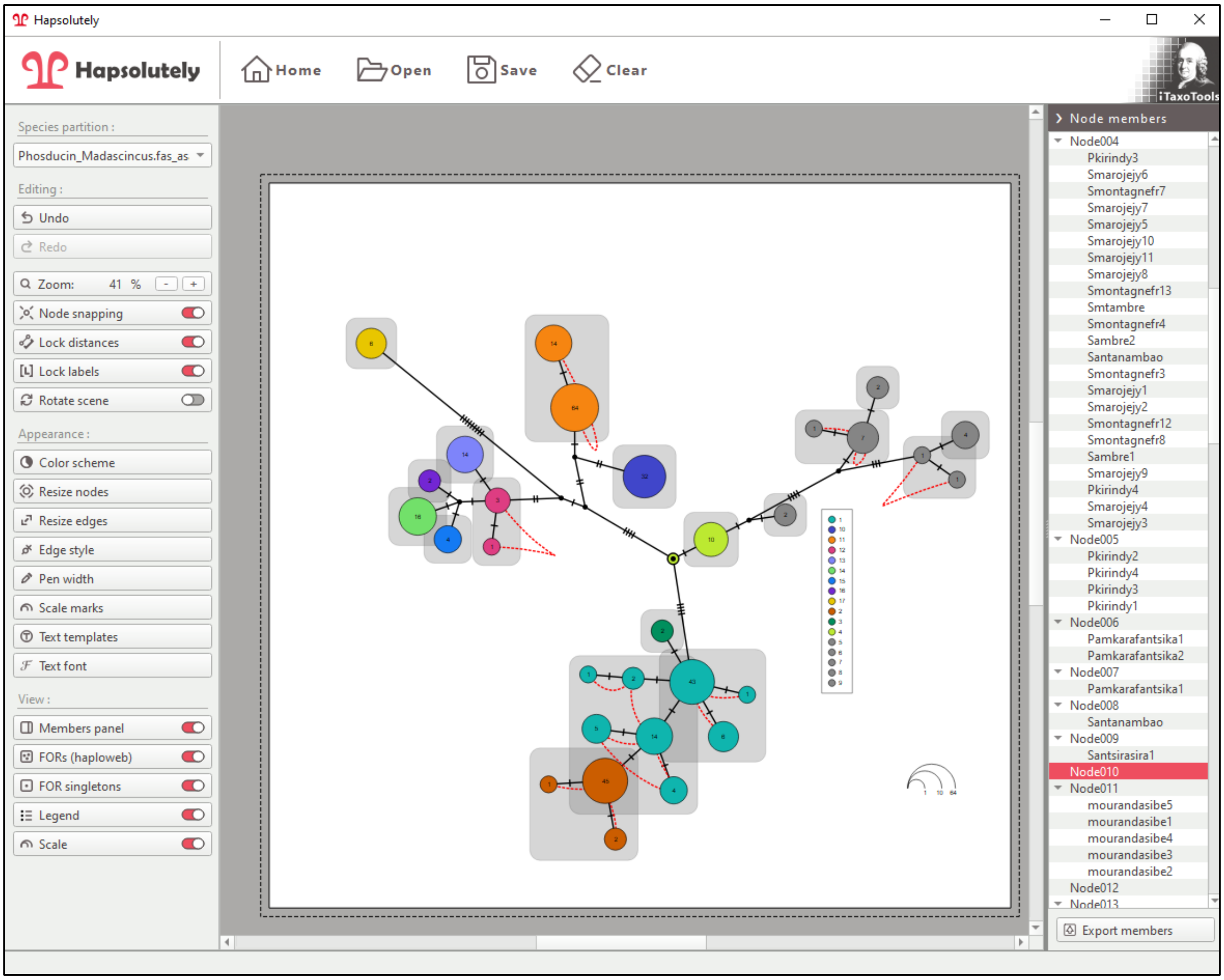
Species partition : Phosducin_Madascincus.fas_as

Zoom: 41 %

Node members

MAndohahela
Mimaloka4
Mranomafana3
Node026
Node027
Mtsaratanana1
Manala1
Node028
Mnosymangabe4
Mnosymangabe6
Mnosymangabe2
Mnosymangabe3
Mnosymangabe5
Mnosymangabe7
Mnosymangabe8
Node029
Manala1
Node030
Node031
Mtsaratanana2
Node032
Mcrottini2
Mcrottini1
Node033
Mangozongahy1
Mmakira4
Mmakira8
Mmakira6
Mmakira1
Mmakira2
Mmakira3
Mmakira7
Node034
baeuandasibe3
baeuandasibe2
baeuandasibe1
Node035
Aorangea5
Aorangea1
Aorangea7
Aorangea3
Aorangea9
Node036
Pmontagnefrancais5
Pmontagnefr10
Node037
Pmontagnefr6

Hapsolutely allows for extensive editing of genealogies/networks. Here, the genealogy has been rotated, the color palette has been changed, and the little dots representing additional mutational steps (or unsampled alleles) have been changed to crossbars. Note that the crossbars count the numbers of steps (i.e., one crossbar = 1 mutational step) whereas the dots represent the unsampled haplotypes (i.e., no dot = 1 mutational step; 2 dots = 3 steps).



Species partition :
Phosducin_Madascincus.fas_as

Editing :
Zoom: 41 %
Node snapping ●
Lock distances ●
Lock labels ●
Rotate scene ○

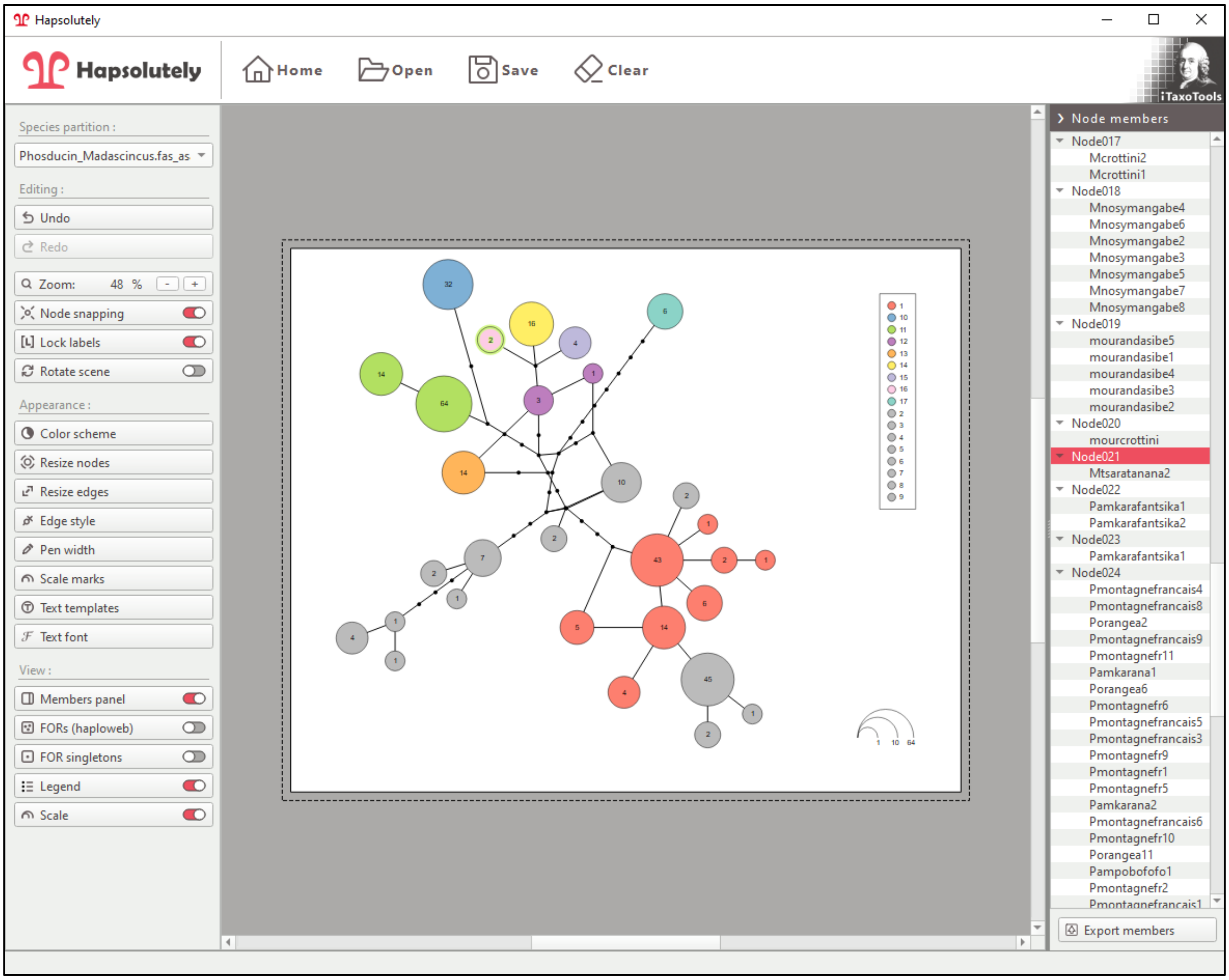
Appearance :

View :
Members panel ●
FORs (haploweb) ●
FOR singletons ●
Legend ●
Scale ●

Node members

- Node004
  - Pkirindy3
  - Smarojejy6
  - Smontagnefr7
  - Smarojejy7
  - Smarojejy5
  - Smarojejy10
  - Smarojejy11
  - Smarojejy8
  - Smontagnefr13
  - Smtambre
  - Smontagnefr4
  - Sambre2
  - Santanambao
  - Smontagnefr3
  - Smarojejy1
  - Smarojejy2
  - Smontagnefr12
  - Smontagnefr8
  - Sambre1
  - Smarojejy9
  - Pkirindy4
  - Smarojejy4
  - Smarojejy3
- Node005
  - Pkirindy2
  - Pkirindy4
  - Pkirindy3
  - Pkirindy1
- Node006
  - Pamkarafantsika1
  - Pamkarafantsika2
- Node007
  - Pamkarafantsika1
- Node008
  - Santanambao
- Node009
  - Santsirasira1
- Node010
- Node011
  - mourandasibe5
  - mourandasibe1
  - mourandasibe4
  - mourandasibe3
  - mourandasibe2
- Node012
- Node013

Hapsolutely also allows computing haplotype networks using the TCS (statistical parsimony), median joining network, minimum spanning network, and tight span walker approaches. The respective algorithms were coded by Leigh & Bryant (2015) for PopArt. Different from Fitchi haplotype genealogies, circular connections between alleles are possible in these networks.



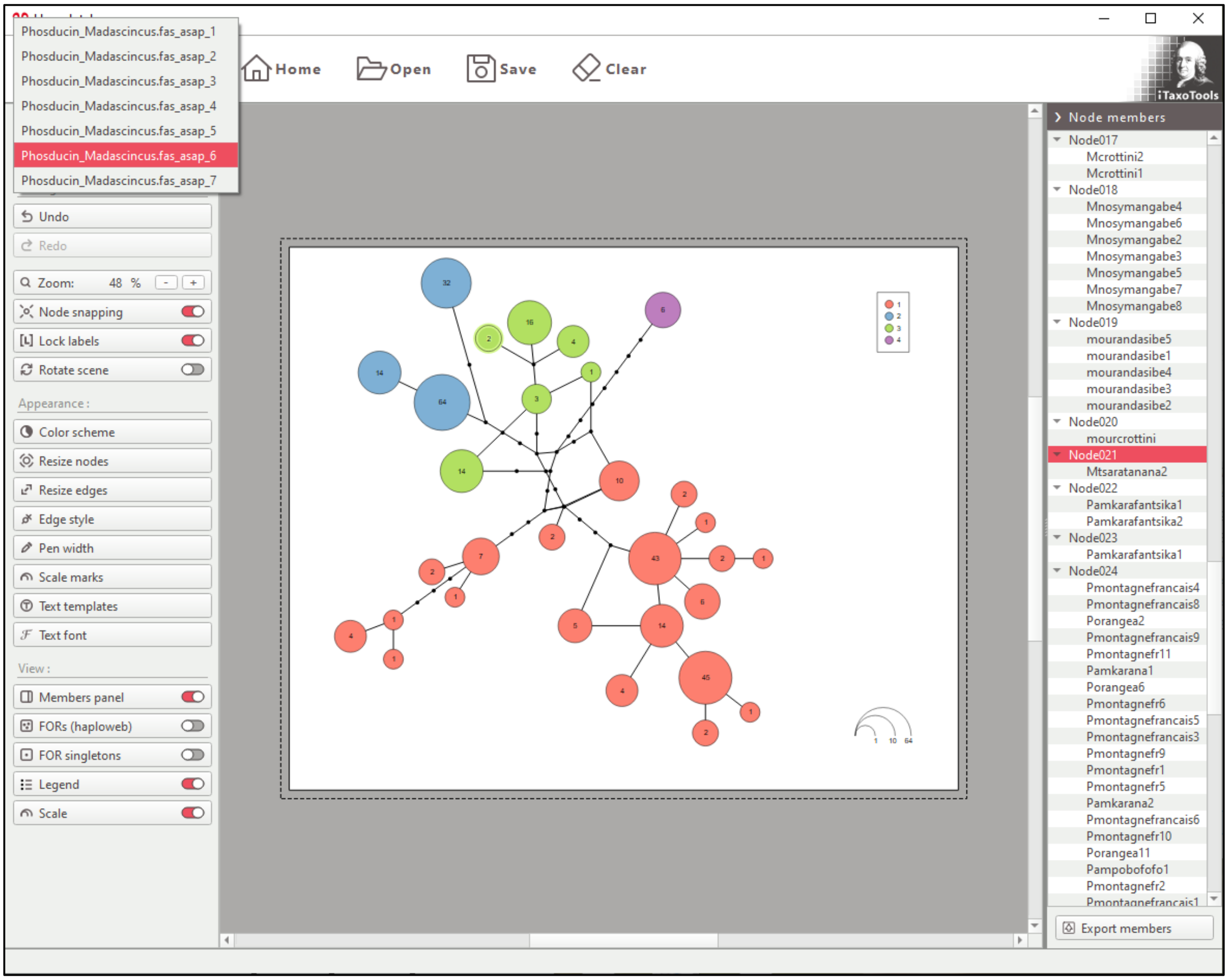
Species partition : Phosducin_Madascincus.fas_as

Zoom: 48 %

Node017
Mcrottini2
Mcrottini1
Node018
Mnosymangabe4
Mnosymangabe6
Mnosymangabe2
Mnosymangabe3
Mnosymangabe5
Mnosymangabe7
Mnosymangabe8
Node019
mourandasibe5
mourandasibe1
mourandasibe4
mourandasibe3
mourandasibe2
Node020
mourcrottini
Node021
Mtsaratanana2
Node022
Pamkarafantsika1
Pamkarafantsika2
Node023
Pamkarafantsika1
Node024
Pmontagnefrancais4
Pmontagnefrancais8
Porangea2
Pmontagnefrancais9
Pmontagnefr11
Pamkarana1
Porangea6
Pmontagnefr6
Pmontagnefrancais5
Pmontagnefrancais3
Pmontagnefr9
Pmontagnefr1
Pmontagnefr5
Pamkarana2
Pmontagnefrancais6
Pmontagnefr10
Porangea11
Pampobofofo1
Pmontagnefr2
Pmontagnefrancais1

Press the Save button in the upper row of icons to save the graph in PNG, SVG or PDF format.

Hapsolutely also supports the SPART format. See the original publication by Miralles et al. (2021) for specifications of this format. A SPART file is provided along with example file 3. The SPART file can include multiple species partitions (i.e., multiple combinations of individuals into subsets, which, in this context, may represent multiple combinations of sequences into species). The sequence identifiers in the SPART file must coincide with those in the sequence file. Hapsolutely then allows selection of one of the species partitions and assigns colors in the network / genealogy accordingly. The program also allows toggling between species partitions in the viewer.



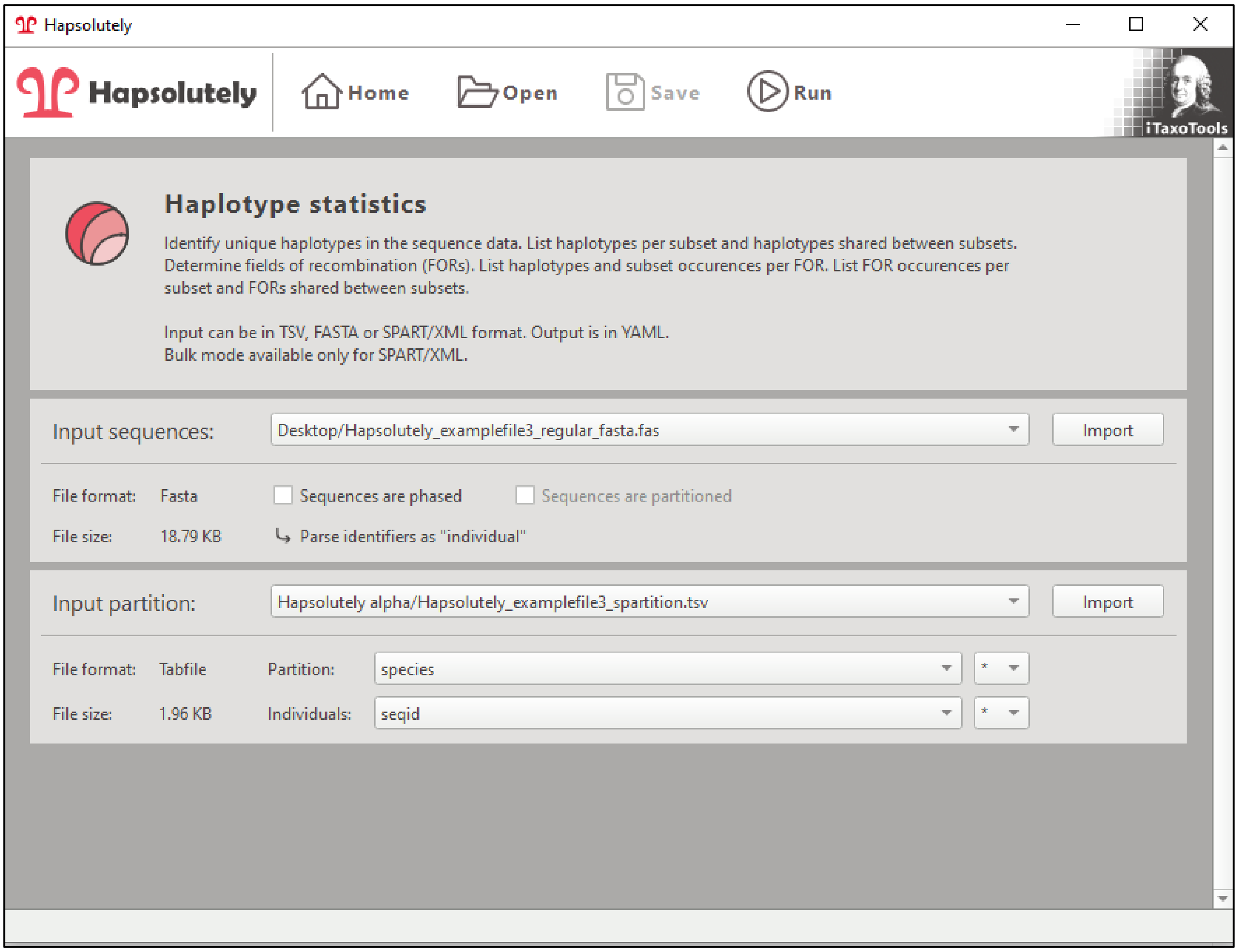
Phosducin_Madascincus.fas_asap_1
Phosducin_Madascincus.fas_asap_2
Phosducin_Madascincus.fas_asap_3
Phosducin_Madascincus.fas_asap_4
Phosducin_Madascincus.fas_asap_5
Phosducin_Madascincus.fas_asap_6
Phosducin_Madascincus.fas_asap_7
Home
Open
Save
Clear
iTaxoTools
Undo
Redo
Appearance :
Color scheme
Resize nodes
Resize edges
Edge style
Pen width
Scale marks
Text templates
Text font
View :
Export members
Node members
32
16
6
2
4
14
64
1
3
14
10
2
1
2
7
2
1
43
2
1
1
6
4
1
1
5
14
45
4
1
2
1
2
3
4
1
10
64

Zoom: 48 %

- Node snapping: on
- Lock labels: on
- Rotate scene: off
- Members panel: on
- FORs (haploweb): off
- FOR singletons: off
- Legend: on
- Scale: on

```
Node017
    Mcrottini2
    Mcrottini1
Node018
    Mnosymangabe4
    Mnosymangabe6
    Mnosymangabe2
    Mnosymangabe3
    Mnosymangabe5
    Mnosymangabe7
    Mnosymangabe8
Node019
    mourandasibe5
    mourandasibe1
    mourandasibe4
    mourandasibe3
    mourandasibe2
Node020
    mourcrottini
Node021
    Mtsaratanana2
Node022
    Pamkarafantsika1
    Pamkarafantsika2
Node023
    Pamkarafantsika1
Node024
    Pmontagnefrancais4
    Pmontagnefrancais8
    Porangea2
    Pmontagnefrancais9
    Pmontagnefr11
    Pamkarana1
    Porangea6
    Pmontagnefr6
    Pmontagnefrancais5
    Pmontagnefrancais3
    Pmontagnefr9
    Pmontagnefr1
    Pmontagnefr5
    Pamkarana2
    Pmontagnefrancais6
    Pmontagnefr10
    Porangea11
    Pampobofofo1
    Pmontagnefr2
    Pmontagnefrancais1
```

The “Haplotype statistics” module of Hapsolutely allows to compute a series of basic descriptive statistics. The start window is similar to that of the network builder. Also here, the sequence file can be specified independently from the file which contains information on the species partition to be used.



Haplotype statistics

Identify unique haplotypes in the sequence data. List haplotypes per subset and haplotypes shared between subsets. Determine fields of recombination (FORs). List haplotypes and subset occurences per FOR. List FOR occurences per subset and FORs shared between subsets.

Input can be in TSV, FASTA or SPART/XML format. Output is in YAML.
Bulk mode available only for SPART/XML.

Input sequences: Desktop/Hapsolutely_examplefile3_regular_fasta.fas

File format: Fasta
File size: 18.79 KB

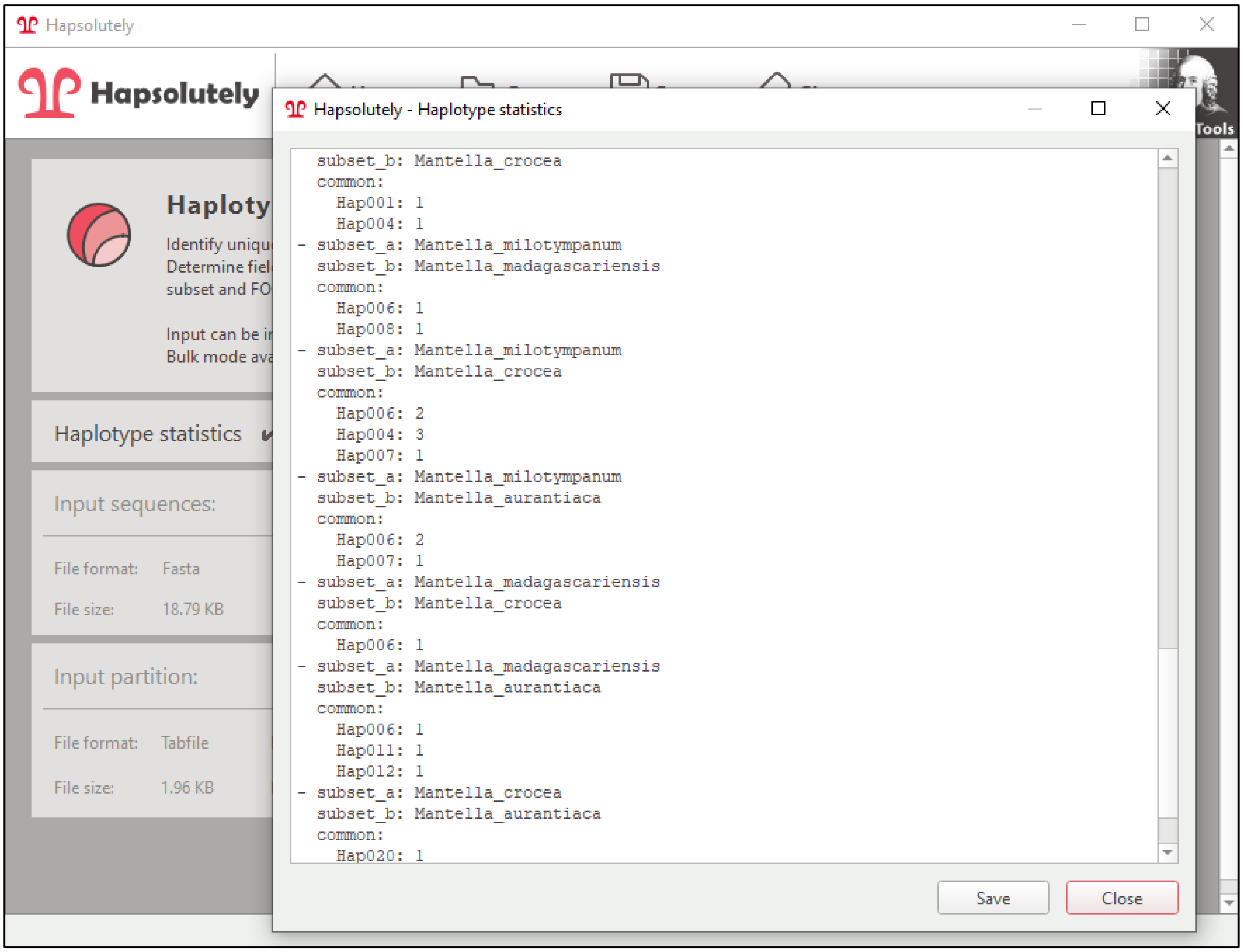
□ Sequences are phased
□ Sequences are partitioned
Parse identifiers as "individual"

Input partition: Hapsolutely alpha/Hapsolutely_examplefile3_spartition.tsv

File format: Tabfile
File size: 1.96 KB

Partition: species *
Individuals: seqid *

The descriptive statistics computed are focused on information relevant for species delimitation, i.e, haplotypes shared between subsets (species), individuals included in each field of recombination, or fields of recombinations shared between pairs of species.



File format: Fasta
File size: 18.79 KB

File format: Tabfile
File size: 1.96 KB

```
  subset_b: Mantella_crocea
  common:
    Hap001: 1
    Hap004: 1
- subset_a: Mantella_milotympanum
  subset_b: Mantella_madagascariensis
  common:
    Hap006: 1
    Hap008: 1
- subset_a: Mantella_milotympanum
  subset_b: Mantella_crocea
  common:
    Hap006: 2
    Hap004: 3
    Hap007: 1
- subset_a: Mantella_milotympanum
  subset_b: Mantella_aurantiaca
  common:
    Hap006: 2
    Hap007: 1
- subset_a: Mantella_madagascariensis
  subset_b: Mantella_crocea
  common:
    Hap006: 1
- subset_a: Mantella_madagascariensis
  subset_b: Mantella_aurantiaca
  common:
    Hap006: 1
    Hap011: 1
    Hap012: 1
- subset_a: Mantella_crocea
  subset_b: Mantella_aurantiaca
  common:
    Hap020: 1
```

Hapsolutely contains numerous other options that are best explored by running the program with the example files. The following list hints at several options and issues to keep in mind.

►In Hapsolutely each allele "bubble" is called a "node" and the branches connecting them are "edges".

►The viewer includes an "Undo" button which allows undoing almost all commands executed to edit the network / genealogy.

►Each node by default shows the number of sequences by which the respective allele is represented. To remove this number, use the option "Text templates" and delete the term "Weight" from node label.

►Under edge styles, you can choose between bubbles =dots) or bars to represent additional mutational steps (or unsampled alleles). Note that the crossbars count the numbers of steps (i.e., one crossbar = 1 mutational step) whereas the bubbles represent the unsampled haplotypes (i.e., no bubble = 1 mutational step; 2 bubbles = 3 steps). The "cutoff" value indicates the maximum number of steps that are shown as single bars or bubbles; connections of more steps are collapsed and the number of steps is shown instead.

►The "Members panel" button opens a list where the sequence identifier is shown for all sequences represented in each node. The list can be exported and saved using the button below the list.

►In Fitch genealogies (without circular connections among haplotypes), edge lengths are strictly proportional to the number of mutational steps, and the length is locked. Nodes can therefore be moved only while maintaining the distance. This can be changed by unselecting the "Lock distances" button. In haplotype networks inferred by other algorithms, distances are by default unlocked (and indeed cannot be locked).

►The "Scale marks" button allows adjusting the scale that indicates the number of sequences per node.

►Node size can be adjusted either using a linear factor, an area factor or a logarithmic factor. The "base radius" specifies the minimal size of a node containing one sequence. In order to have node radius fully proportional to the number of sequences per node, set the linear factor to 1. In order to have node area fully proportional to number of sequences per node, set the area factor to 1.