

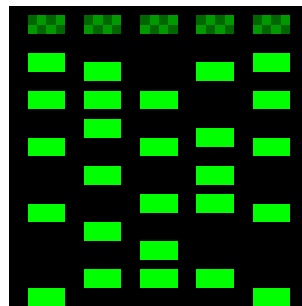
# FAMD - Fingerprint Analysis with Missing Data 1.31

## - Manual -

Philipp M. Schlüter

Institute of Systematic Botany  
University of Zurich  
Zürich, Switzerland

September 2013



# INDEX

## FAMD - Fingerprint Analysis with Missing Data 1.3 Manual

INDEX.....	2
GENERAL.....	5
Purpose.....	5
Author.....	5
Citation.....	5
Licence and Legal Disclaimer.....	5
System requirements.....	6
Linux.....	6
Source code and compilation.....	6
What was new in FAMD 1.3?.....	6
What was new in FAMD 1.2?.....	7
What was new in FAMD 1.1?.....	7
Limitations and known issues.....	7
Opening the input file.....	7
AMOVA.....	8
HOW TO USE.....	9
Installing and uninstalling.....	9
Installing/Uninstalling and using the FAMD 1.1 plug-in for MS Excel.....	9
General directions.....	10
FAMD command line options.....	10
Reporting bugs.....	11
DEFAULTS.....	12
Default settings.....	12
Default input format(s).....	13
PROGRAM FUNCTIONS.....	16
File Menu.....	16
Load.....	16
Load from Clipboard.....	17
Save DataMatrix.....	17
Import.....	18
Import Distance Matrix.....	18
Binary matrix from clusters in trees.....	18
Select File Names.....	19
Select Analysis Output File.....	19
Select Tree Output File.....	19
Select Consensus Tree Input File.....	19
Select Consensus Tree Output File.....	19
Select Log File.....	19
Export.....	20
Nexus.....	20
Arlequin Project.....	20
Tab-delimited text / R.....	20
Genepop.....	20
NTSys-pc.....	21
List of OTUs.....	21
SynTax.....	21
hindex.....	21
Structure.....	22
Hickory (Nexus).....	22
Dfdist.....	22
BayeScan.....	22
AFLPDat.....	23
AFLPop.....	23
Phylip distance matrix.....	23
Turn log on/off.....	23
Exit.....	23

DataMatrix Menu.....	23
Restore Original Matrix.....	24
Matrix Statistics .....	24
Missing Data Statistics .....	24
Count Bands.....	24
Mean Number of Bands Per Individual.....	24
Polymorphic Bands.....	24
Fixed Bands.....	24
Private Bands .....	25
Fixed Private Bands.....	25
Frequency Statistics.....	25
Frequencies per Individuals.....	25
Frequencies per Loci .....	25
Replace Missing Data .....	26
Resample Loci (Bootstrap Replicate).....	26
Remove Individuals.....	26
With missing data... ..	26
With band frequency below... ..	26
With band frequency above... ..	26
Remove Loci... ..	27
With missing data... ..	27
Monomorphic.....	27
Monomorphic Absences.....	27
Monomorphic Presences.....	27
With band frequency below... ..	27
With band frequency above... ..	27
With more missing data than presences .....	28
Pairwise Individual Comparison .....	28
Group Manager .....	28
Defining Groups (groups of individuals) .....	29
Defining and Selecting LocusSets (groups of loci) .....	30
Select Group Only .....	30
Select LocusSet Only .....	30
Group-based Profiles .....	30
Additive Band Profile .....	30
Fixed Band Profile .....	30
Analysis Menu.....	31
Standard Similarity .....	31
Minimum Similarity .....	32
Maximum Similarity .....	33
Average Similarity .....	34
Null Allele Frequencies.....	34
Square root .....	34
Lynch-Milligan.....	35
Bayesian (uniform prior) .....	35
Bayesian (among-population prior) .....	35
Bayesian (among-locus prior).....	36
Shannon's Index.....	37
ML Hybrid Index .....	38
AMOVA.....	39
Pairwise PhiST .....	39
Population Distance .....	40
Trees Menu.....	40
UPGMA .....	40
Neighbour Joining .....	41
Strict Consensus .....	41
Majority Rule Consensus .....	41
Principal Coordinate Analysis .....	41
Replicate Analyses Menu .....	43
Bootstrap Shannon's Variance (SH + RMD).....	43
Bootstrap Shannon's Variance @ NIndiv.....	43
Bootstrap Std Tree (BS + RMD).....	43

Multiple Avg Trees (TR + JC).....	44
Bootstrap Avg Trees (BS + JC).....	44
Bootstrap Population Tree.....	45
Estimate R-support (TR) .....	46
Shannon t-tests .....	46
View Menu .....	48
Input File.....	48
Analysis File .....	48
Tree File .....	48
Consensus Tree File .....	48
PCoA 3D Viewer .....	48
Log File.....	50
Options Menu.....	50
Missing data replacement .....	50
Shannon scaling .....	51
(Dis)Similarity Coefficients .....	52
Distance Transformation .....	52
Character Weights.....	52
Bootstrapping & Replicates .....	52
Trees .....	53
Consensus Trees .....	53
R-Support .....	53
AMOVA.....	54
PCoA .....	54
Allele Frequencies .....	54
Populations distances .....	54
I/O Options .....	55
Project .....	55
Help Menu.....	56
About .....	56
Citation .....	56
Version .....	56
Check for new version.....	56
Help .....	56
REFERENCES .....	57

# GENERAL

## **Purpose**

This program was originally written for analysis of RAPD, AFLP or other **dominant** (binary) fingerprint data, especially for data sets that contain ambiguous or missing data, so that the impact that missing data may have on the analysis can be better evaluated. The second intention was to provide a means to easily calculate the variance associated with Shannon's index. However, as outlined in this manual, since its beginnings FAMD has accumulated a number of additional features.

## **Author**

Philipp M. Schlüter

### Current address:

Institute of Systematic Botany  
University of Zurich  
Zollikerstr. 107  
CH-8008 Zürich  
Switzerland

E-mail: [philipp.schlueter@systbot.uzh.ch](mailto:philipp.schlueter@systbot.uzh.ch)  
Telephone (office): +41 44 63 48328

**For any problems or queries about the program please contact the author.**

## **Citation**

FAMD was originally described in the following publication:

Schlüter, P. M. & Harris, S. A., 2006. Analysis of multilocus fingerprinting data sets containing missing data, *Mol. Ecol. Notes*: **6**: 569-572.

## **Licence and Legal Disclaimer**

By downloading and using FAMD you accept the following.

*FAMD (c) Copyright 2002-2013 Philipp M Schlüter.*

*This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY, whether expressed or implied; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. The author (PMS) will not be liable for any special, incidental, consequential, indirect or similar damages due to loss of data or any other reason, even if he or an agent of his has been advised of the possibility of such damages. In no event shall the author be liable for any damages, regardless of the form of the claim. The person using the software bears all risk as to the quality and performance of the software.*

For some statistical functions, FAMD uses routines from J. Debord's TPMath numerical library.

FAMD may be distributed freely for non-commercial purposes provided that the copyright notice is not removed. If you intend to include it in any commercially distributed package, the author should first be contacted. If you wish to modify the source code and rebuild this application so that it fits your own need, you may do so provided you do not remove the original copyright notice and provided you do not intend to distribute it commercially. For access to the source code please contact the author.

Any source code (e.g. the R script) provided with FAMD is open-source under the *Artistic License 2.0*; a copy of the licence terms can be found at <http://www.opensource.org/licenses/artistic-license-2.0.php>.

## **System requirements**

FAMD was developed for 32-bit (or higher) Windows operating systems and was tested on Windows 98, Windows XP x32/x64, Windows Vista x64 and Windows 7 x64. The program was developed and tested on Intel Pentium M, Intel Core 2 Duo T7300 and higher systems.

FAMD uses floating-point instructions and requires a CPU with a floating-point unit (FPU). FAMD uses instructions such as the CPUID instruction introduced on later 486-processors which therefore represent the minimum processor requirement for FAMD.

For most purposes, FAMD uses FPU-based (80-bit) floating point mathematics for increased numerical precision, but for some purposes, FAMD will try to use SIMD floating point instructions if available on your system for increased speed.

On multithreading-enabled FAMD versions, multithreaded versions of routines can only be used on Windows 2000 and later because the internal threading architecture relies on system functions introduced with that Windows version.

## **Linux**

FAMD is a windows executable and, as such, is not designed to run on other systems. However, I have seen FAMD work nicely on Linux under WINE. Just be aware that FAMD expects to read files that contain DOS/Windows-style line breaks.

## **Source code and compilation**

Source code is available from the author upon request (for licence terms see above). FAMD was developed in Borland Delphi 7, later Delphi 2006 and Embarcadero CodeGear's Delphi 2009/2010, and includes a number of (BASM) assembly-language routines. Some routines are based on older ones (Turbo Pascal and Microsoft Macro Assembler). I admit that the source code documentation is rather poor and therefore cannot guarantee that the code will be easy to follow for anybody but you are free to give it a try.

## **What is new in FAMD 1.3?**

Aside from bug fixes (including a bug with the LocusSet selection), FAMD **1.30** now has

- Maximum likelihood hybrid index calculation
- Centroid calculation for PCoA
- Export filter for AFLPData
- Export filter for AFLPop
- Export filter to tab-delimited text files (.tab/.tsv) for easy loading of data into R or Excel
- Export filter for BayeScan
- Updated manual (this document)
- Shannon's index estimation at specifies sample size.
- Ability to avoid dialogue box when opening files saved with FAMD.
- Improved handling of LocusSets, including quick LocusSet selection.
- Expanded AMOVA functionality (beware, this is in *beta testing* phase !).

Version **1.31** fixes some bugs and adds new features:

- Maximum likelihood population (re)allocation
- Import of count data (e.g. from next-generation sequencing-based tags)
- Plotting of (selected) analysis results via R
- Fixed bug that prevented users to start hybrid index calculation
- Fixed bug that prevented AMOVA calculation on a subset of loci

## ***What was new in FAMD 1.2?***

Version 1.2 has

- an improved and clearer Graphical User Interface, including an improved interface for changing Options and Settings
- a completely rewritten (but thoroughly tested) and more flexible internal computational framework which makes the new version more maintainable and extensible.
- a number of small bug fixes and performance improvements, including a more stable handling of user-defined Groups of individuals.
- support for reading and writing of a small File Specification block to speed up file loading.
- an updated manual (this document)
- new features such as Pairwise Individual Comparison Dialogue, and Data Import features

Version 1.21 updates the Structure export filter (and includes bug-fixes).

Version 1.23 further introduces a data integrity check upon loading of input files and fixes some small graphical user interface-related issues (especially for Windows Vista and later)

Version 1.25 includes minor bug fixes and the option to check whether FAMD is up to date (by connecting to the FAMD web site).

## ***What was new in FAMD 1.1?***

Version 1.1 added a number of features to FAMD, such as...

- read and write of undelimited data files
- export data matrices to a number of other file formats
- provide a graphical interface for easily defining groups within a data set
- calculate a number of different similarity/distance measures (Jaccard, Dice, SMC similarities, and NeiLi, Euclidean and Squared Euclidean distances)
- carry out different distance transformations
- carry out AMOVA analysis with all implemented (dis)similarity measures
- construct strict and majority rule consensus trees
- carry out principal coordinate analysis with all implemented (dis)similarity measures
- estimate  $R_x$ -support of branches based on the missing data present in the data matrix
- estimate null allele frequencies using different methods
- calculate pairwise  $\Phi_{ST}$  distances and the chord distance among populations

## ***Limitations and known issues***

*Opening the input file*

In some rare cases it may be difficult to read in an input file generated by some standard Windows Editors, because it may hide some additional characters in the file that you don't see in the Editor. When saving files to be read by FAMD from Unicode-enabled applications, please make sure that the files to be read by FAMD are saved as ANSI/ASCII-text files and not as Unicode files.

FAMD 1.2 and later should no longer have any problems reading in files that contain remark characters.

### *AMOVA*

The current AMOVA implementation makes certain assumptions about the Groups defined in your data. Please see the *AMOVA* section for further details.



# HOW TO USE

## *Installing and uninstalling*

You need not specifically install FAMD. Just place it anywhere on your computer and start the application. If you wish to remove FAMD, just delete the executable (famd.exe or similar) and/or any files that came along with it (typically, a help file). For the `Help` command to properly display this pdf help file, it should be named famdhelp.pdf and reside in the same directory as the FAMD executable.

## *Installing/Uninstalling and using the FAMD 1.1 plug-in for MS Excel*

FAMD 1.1 contained an optional plug-in for Microsoft Excel, which works on Excel versions up to Excel 2003 (at least on Windows XP systems). **It does not work with Excel 2007 or later.** Because of this and because it does not contain any really critical features, it is not included with FAMD 1.2 and later. However, you can still download FAMD 1.1 and use the Excel plug-in, *famd4xls.dll*, if you wish, and install it manually (see below).

The FAMD plug-in for Microsoft Excel (contained in the file famd4xls.dll) provides an easy way of getting scored dominant fingerprint data into FAMD without the need to create a FAMD input file. If you have installed this add-in, simply select the area of an Excel worksheet with and then press on the “FAMD” button in the FAMD4XLS toolbar. If this toolbar is not visible, right-click any of the visible Excel toolbars, and make sure FAMD4XLS is checked.

Clicking FAMD4XLS’s “FAMD” button copies the selected area into the clipboard, starts FAMD and instructs it to load data from the clipboard. If the FAMD4XLS plug-in should fail to work on your machine for whatever reason, you may still carry out these steps manually, so as to load data into FAMD without the need to create a FAMD input file.

To install the FAMD Excel plug-in manually (here described for WinXP):

- ) Register the Excel Plug-in, which is contained in the dynamic link library famd4xls.dll with the system by calling `regsvr32 famd4xls.dll` from the command line. This will usually display a dialogue box notifying the user of the success (or failure) of the action.
- ) Start `regedit.exe` and set a registry key which tells Microsoft Excel about the existence of the plug-in and stores the path of a current copy of FAMD’s executable in the registry so that the plug-in, when called by Excel, knows where to find FAMD on your computer. The key that to be created should be:  
`HKEY_CURRENT_USER\Software\Microsoft\Office\Excel\Addins\FAMD4XLS.famdXLplugin`  
Within this key, create a the `REG_DWORD` value named `LoadBehavior` and set it to 3 and, in addition, create a `REG_SZ` value named `FAMDpath` and set it to a valid path pointing to a FAMD executable (e.g., `famd.exe`).

To uninstall the Excel plug-in:

- ) Unregister the Excel Plug-in, which is contained in the file famd4xls.dll with the system by calling `regsvr32 /u famd4xls.dll`  
This will usually display a dialogue box notifying the user of the success (or failure) of the action.
- ) Delete the registry key Microsoft Excel needs to know about the FAMD’s plug-in, i.e. delete the following key (and everything contained in it):  
`HKEY_CURRENT_USER\Software\Microsoft\Office\Excel\Addins\FAMD4XLS.famdXLplugin`
- ) You may now delete `famd4xls.dll`

**NB:** If you cannot, for whatever reason, install the Excel plug-in, you can still transfer data from Excel to FAMD easily by selecting and copying on the clipboard in Excel, and loading it from Clipboard in FAMD, using `File → Load from Clipboard`.

## General directions

To use FAMD:

- ) generate an input file that can be read by FAMD
- ) load the input file into FAMD
- ) if desired, specify the file you want your analysis results to be saved to
- ) if desired, specify the file you want your trees to be saved to
- ) if desired, modify your data matrix (`DataMatrix` menu); possibly define groups of individuals or loci using the `DataMatrix` → `Group Manager` function.
- ) to change preset parameters, use the `Options` menu
- ) carry out your analysis using the `Analysis` or the `Replicate Analyses` menu. The `Analysis` menu writes results to the analysis file you selected and, if you generated a similarity/distance matrix expects you to analyse it further yourself (e.g., tree-building, to be found in the `Trees` menu). The `Replicate Analysis` menu performs more complex analyses that run for some time and handle tree-building themselves.
- ) if desired, visualise results by using the options in the `View` menu or third party software.

## FAMD command line options

FAMD can be started from command line by typing in the name of its executable. FAMD will accept a command line option which specifies an input file to be opened by FAMD, i.e.

```
C:\> famd.exe test.txt
```

will start FAMD and tell it to attempt to open the file `test.txt`, in the same way as starting FAMD from within Windows and opening the file via the `File` → `Load` menu. This also means that you can open a file from within Window's Explorer by right clicking on an input file, selecting the "Open with" option and specifying FAMD as the program with which to open the file.

Alternatively, FAMD can be started with the `/M` command line option, i.e.

```
C:\> famd.exe /M
```

telling FAMD to load input data from the clipboard. This is equivalent to starting FAMD and loading data via the `File` → `Load from Clipboard` menu.

Apart from these, FAMD currently does not accept any command line options.

## FAMD's R plotting script

As of version 1.31, FAMD provides the functionality to run an R script to plot the results provided in its analysis output file. This is meant allow easy and customisable visualisation of results. At the time of writing, the provided script (`plotfamd.R`) only handles a subset of possible analysis results:

- ) AMOVA,
- ) missing data statistics
- ) ML Population Reallocation
- ) ML Hybrid Indices

and even those may still require some further development and aesthetic improvements. I would greatly value any contribution you may make to improve the plotting script. In any event, the script is

provided with open source and I encourage you to change the plotting script as needed for your own purposes.

You can find details on the interoperation of FAMD and its plotting script in the comments provided in the R script. Briefly, FAMD will call R's script interpreter, providing as two additional parameters the name of the analysis output file (to be plotted) and the plotting output file like so (by default):

```
RScript.exe plotfamd.R analysis.txt famdplot.pdf
```

By default, FAMD will try to locate `RScript.exe` by evaluating the Windows Registry for R's entries as recommended by the R development team. Also by default, FAMD will choose the file present in the most recent installation, preferring local user over administrator, and 64-bit over 32-bit installations. Alternatively, you may manually select the R script interpreter to be used.

## ***Reporting bugs***

While a lot of effort goes into testing and validating FAMD, no software is perfect and there will be bugs, and some of these may only manifest themselves under very special circumstances. Therefore, in order to enhance the quality, reliability and usability of FAMD, I rely on users informing me of any bugs they may bump into. Otherwise, I may never see them and hence never get the opportunity to fix them. Please also feel free to write me about feature requests or other things you think may be improved.

In order to fix a bug, it is necessary for me to reproduce it so that I can track it down. Therefore, I would ask you to provide me with all the information and data necessary to reproduce a bug. I would recommend you to start FAMD, go to `File→Turn log on`, and then carry out all the steps that led to the bug. Please send me the log file, along with a description of the error message (if there is one and if it isn't captured in the log file), the FAMD version you're using, and input data file. In many cases, also information on the system under which FAMD is running (see `File→I/O Options`) is useful.

I will treat your input data absolutely **confidential**, but in most cases, access to the same data file you were using is absolutely essential to reproduce a bug!

# DEFAULTS

## *Default settings*

Individual names given in input file:	yes
Locus names given in input file:	no
Input file has individuals in:	columns
End of data character:	*
Open remark character:	{
Close remark character:	}
Presence:	1
Ambiguity/missing data point:	?
Absence:	0
Characters ignored in input file:	Blank (00h), whitespace, tab, CR, LF
Include groups:	yes
Delimited data:	yes
Output negative ambiguities	no
Write FAMD File Specification block	yes
Trust File Specification Block	yes
Overwrite or Append to output files:	Ask User
Display FAMD Warnings:	yes
Log to file:	no
JC:	100
TR:	50
BS:	1000
SH:	10000
MDR:	yes
Missing data replacement by presences:	50%
Data removal threshold: MD (individuals):	18%
Data removal threshold: MD (loci):	25%
Data removal threshold: Below Frq (loci)	5%
Data removal threshold: Above Frq (loci)	95%
Data removal threshold: Below Frq (indiv.)	30%
Data removal threshold: Above Frq (indiv.)	70%
Shannon Log Base:	2
Frequencies = presences/locus per:	total presences
Include Shannon I=0 values in statistics:	no
Save all Shannon Bootstrap replicates to file:	no
Majority Rule consensus percentage	50%
Majority Rule consensus for Rx analysis	100%
Rx Consensus Threshold r=	10000
Values of r to be analysed	1 - 10000
Desired precision for R values	1 digit
(Dis)Similarity Coefficient	Jaccard
Distance Transformation	d = 1-s
Write similarities to analysis file	no
Write distances to analysis file	yes
Nei-Li R-value	6.00
Similarity Mode Preference	Standard
Preferred FPU/SSE rounding mode:	Round-Up
PCoA Precision (Error Tolerance):	0.000000024
Max No Iterations for PCoA:	1000000
Auto-start PCoA Viewer:	yes
Calculate centroids:	no
Default Allele Estimation method:	Bayesian, uniform prior
Bayesian Allele Estimation Correction Factor:	0.01

```

ML Population Reallocation correction method: D&B (2002) correction formula
ML Pop. Realloc. Correction epsilon value: 0.00001
Min. log-likelihood difference for pop alloc.: 2.0

```

```

Analysis output file name: analysis.txt
Tree output file name: outtree.ph
Consensus tree input file: outtree.ph
Consensus tree output file: constree.ph
Log file name: famdlog.txt
R plotting output file: famdplot.pdf
R script to plot results: plotfamd.R

```

## Default input format(s)

FAMD is quite flexible as regards input files. The default input format is a text file containing a simple data matrix whose end is indicated by a single user-defined `EndOfData` character (default: `'*'`). The file can contain labels for individuals and/or loci and the data matrix can be in either orientation, individuals in columns or individuals in rows. Finally, data can be delimited by space, tabs, etc., or undelimited. When opening an input file, FAMD will ask you about these things. FAMD 1.2 can also read/write a small File Specification block (bracketed by `'<'` and `'>'`) at the beginning of the file to speed up the loading of data files by filling in these parameters automatically.

There are, however, some limitations to the input data format: Locus and individual names should be **UNIQUE** names, and should **NOT** contain spaces. The input data file should not contain the characters `'<'` and `'>'` (except for a possible File Specification block generated by FAMD itself). For FAMD 1.1 and before, any names should not exceed 19 characters and the data matrix should not contain empty lines. If you wish to use FAMD for the construction of dendrograms, then please do **NOT** include brackets (i.e., `'('` or `')'`) in your individual names.

As of FAMD version 1.23, the program will check whether the names you provide are unique and tell you if there are any obvious problems. **NB:** FAMD versions before (and including) 1.21 did not check whether the names you provide are unique, but since some of the internal data handling makes the implicit assumption of unique names, old FAMD versions may behave unexpectedly and possibly provide erroneous results if you violate this assumption.

FAMD does allow for remarks in the input file, i.e., information bracketed by the `OpenRemark` and `CloseRemark` characters (defaults are `'{'` and `'}'`) is intended solely for your information and is not read in. In older FAMD versions (1.1 and before), this feature only worked partially and inclusion of remarks could in some cases lead to FAMD being unable to load an input file.

The input format is like this:

```

      [Ind1 Ind2 Ind3 ... IndN]
[Loc1]  1    0    1    ...  0
[Loc2]  ?    0    ?    ...  0
      ..    ..    ..    ...  ..
[LocM]  1    0    0    ...  1
*
```

Items in `[ ]` can be left out and blanks/tabs separating the data points need not be present.

The asterisk (`EndOfData` character) at the end of the data block is compulsory. Blanks or tabs can delimit the entries; the program is fairly flexible regarding that. You can also swap individuals and loci (i.e. rows and columns) - you just need to set the input file parameters accordingly.

You can also choose the characters you use to represent presences, absences and missing data and you can re-save modified data matrices, changing any of those parameters.

Examples for input files are given below:

Example 1 - Delimited data; individuals in columns; individual names present; no locus names present.

(This is the default input file format; File Specification Block: <FAMD:10?>{\* :c+,i+,l-,d+>).

```
IndA IndB IndC IndD IndE IndF
1      0      1      1      ?      0
?      0      ?      1      0      0
1      1      0      0      1      1
1      ?      1      1      1      1
*
```

Example 2 - Delimited data; individuals in columns; individual names present; locus names present.  
(File Specification Block: <FAMD:10?>{\* :c+,i+,l+,d+>).

```
      IndA IndB IndC IndD IndE IndF
AA01  1      0      1      1      ?      0
AA02  ?      0      ?      1      0      0
AA03  1      1      0      0      1      1
AA04  1      ?      1      1      1      1
*
```

Example 3 - Delimited data; individuals in rows; individual names present; no locus names present.  
(File Specification Block: <FAMD:10?>{\* :c-,i+,l-,d+>)

```
IndA 1 ? 1 1
IndB 0 0 1 ?
IndC 1 ? 0 1
IndD 1 1 0 1
IndE ? 0 1 1
IndF 0 0 1 1
*
```

Example 4 - Undelimited data; individuals in rows; individual names present; locus names present.  
(File Specification Block: <FAMD:10?>{\* :c-,i+,l+,d->)

```
AA01 AA02 AA03 AA04
IndA 1?11
IndB 001?
IndC 1?01
IndD 1101
IndE ?011
IndF 0011
*
```

Example 6 - Undelimited data; individuals in rows; individual names present; locus names absent.  
Character 'A' as band presence, 'B' as band absence, 'X' as ambiguity, '#' as EndOfData character.  
(File Specification Block: <FAMD:ABX>{# :c-,i+,l-,d->)

```
IndA AXAA
IndB BBAX
IndC AXBA
IndD AABA
IndE XBAA
IndF BBAA
#
```

You may put anything after the EndOfData character; FAMD doesn't care. However, FAMD does allow you to add additional information blocks in an input file AFTER the data matrix/EndOfData character. For instance, a block containing information about sample grouping (such blocks may be generated using the Group Manager function; see this section for further information). Every block begins with a block name (currently case-sensitive), contained in square brackets and ends with an asterisk (\*). Currently, there are two block types supported, a Groups and a LocusSets block. Briefly, such blocks in the input file would look like the following:

```

[Groups]
AllData/East= A, C, D, E;
AllData/West= F, G, H, I, J;
AllData/East/PopEast1= A, C, D;
*

[LocusSets]
ExcludeLoc3= Loc1, Loc2, Loc4, Loc5, Loc6;
*
```

You will not usually have to write `Groups` and a `LocusSets` block manually, nor will you have to write a File Specification Block (optional) manually. However, for completeness, I include a description of the File Specification block below. Its function is to allow FAMD to enter values contained in it File Parameters dialogue box that is displayed before FAMD tries to load a file, which can save the user time if he/she is unsure about the exact input file format or if the file format deviates from the default format. The format of the File Specification Block, beginning with '<' and ending with '>' is:

<FAMD:#####:\$S[, \$S]>                      where items in square brackets are optional.

The six '#' characters between the two ':' characters are 1-byte (ANSI/ASCII) characters that are interpreted as the characters for band `presence`, `absence`, `ambiguity`, `CloseRemark`, `OpenRemark`, and `EndOfData` character, respectively. After the second ':' character, there is one or more fields of the format '\$\$', delimited by ',' if there are several.

In any '\$\$' field, the second ('\$') character can be either '+' (yes) or '-' (no) indicating the state of the parameter indicated by the first ('\$') character. This character can be: 'c' (individuals in columns), 'i' (labels for individuals present), 'l' (locus names present), 'd' (delimited data), 'a' (read in auxiliary data such as `Groups/LocusSets` blocks).

# PROGRAM FUNCTIONS

## *File Menu*

### Load

Displays a file-open dialogue box which lets you select a file to open. For instructions of the supported file formats, please see the `Default input format(s)` section in this manual. The File Parameter Selection box is displayed which lets you specify which characters are used for which purpose in the file:

Individuals in...  
columns      Data points in the file are read such that each line represents one locus.  
rows         Data points in the file are read such that each line represents one individual.

Header presence for  
individuals    If this option is checked, individual labels are assumed to be present. If this is not set, labels will be given as Ind01, Ind02, etc.  
loci           If this option is checked, locus labels are assumed to be present. If this is not set, labels will be given as Loc001, Loc002, etc.

Delimited data  
If this option is checked, the program assumes that individual data points in your input data matrix are separated by delimiting characters, such as space or tabs. Unchecking this option essentially allows the input of rows of data points (either loci or individuals) without any delimiting characters. Characters specifying data points MUST be single characters. See the examples in the `Default input format(s)` for the difference between delimited and undelimited data.

Include groups  
The default value is yes (checkbox is checked). This tells FAMD to scan the input file for any information on sample of locus groups defined that may be present in the input file AFTER the EndOfData character (default character: '\*'). For more details, please see the `Default input format(s)` and `Group Manager` sections in this manual.

### Characters:

These need to be single characters. Characters '<' and '>' should not be used because they are used internally in the program.

Presence:      Default value='1'.  
                 Specifies a band presence.  
Absence:       Default value='0'.  
                 Specifies a band absence.  
Ambiguity:     Default value='?'.  
                 Specifies a missing data point (e.g., an unclear band).  
OpenRemark:    Default value='{'.  
                 After this character, text or numbers will be treated as a remark.  
CloseRemark:   Default value='}'.  
                 Defines the end of a remark.  
EndOfData:     Default value='\*'.  
                 This character represents the end of data in the file. Any data after this character will be ignored.



## Load from Clipboard

This function is similar to the `File→Load` function (please see there for details), except that FAMD attempts to load data (in the specified input file format) from the clipboard rather than a file. You can use this option to “copy and paste” data from other programs such as office applications.

FAMD’s command line option `/M` also uses this function.

## Save DataMatrix

This lets you re-save your original data matrix. In the process, you can change the characters used to specify e.g. band absences and presences, or you can swap data rows and columns (individuals and loci). Remarks that might have been present in the original data file will be lost, as will be all data after the `EndOfData` character (except for `Groups` and `LocusSets` defined, if there are any). If you have removed individuals or loci from the data set, you can use this option to save a sub-set of your original data set.

Use the File Parameter Selection box to specify which characters are used for which purpose in the file:

Individuals in...  
    columns      Data points in the file are read such that each line represents one locus.  
    rows         Data points in the file are read such that each line represents one individual.

Header presence for  
    individuals    If this option is checked, individual labels are assumed to be present. If this is not set, labels will be given as Ind01, Ind02, etc.  
    loci          If this option is checked, locus labels are assumed to be present. If this is not set, labels will be given as Loc001, Loc002, etc.

Delimited data  
    If this option is checked, the program will save your data matrix as delimited data. For more details, please see the `Default input format(s)`.

Include groups  
    If this option is checked, FAMD will save `Groups` and `LocusSets` blocks to the data file, if groups or `LocusSets` have been defined. For more details, please see the `Default input format(s)` and `Group Manager` sections in this manual.

Ambiguity negative  
    Checking this option will save missing data values preceded by a minus (“-”) sign.

Characters:

These need to be SINGLE characters. Characters ‘<’ and ‘>’ should not be used because they are used internally in the program.

Presence:      Default value='1'.  
                 Specifies a band presence.  
Absence:        Default value='0'.  
                 Specifies a band absence.  
Ambiguity:      Default value='?'.  
                 Specifies an unclear band; missing data point.  
OpenRemark:    Default value='{'.  
                 After this character, text or numbers will be treated as a remark.

CloseRemark: Default value='}'.

Defines the end of a remark.

EndOfData: Default value='\*'.  
This character represents the end of data in the file. Any data after this character will be ignored.

If the option "Write FAMD File Specification Block", to be found under Options→I/O Options, is set (default: yes), then FAMD will encode the File Parameters described above in a FAMD File Specification Block and write it to the first line of the File to be saved

## Import...

The commands included under this option let you import data from 'non-standard' input files into FAMD.

### *Import Distance Matrix*

This lets you import a distance matrix (NB: not a similarity matrix!) generated in another program into FAMD, e.g. for PCoA or dendrogram construction. A distance matrix to be imported must conform to the following format:

```
My Very Special Distance Matrix
Ind01 Ind02 Ind03 Ind04
0.000
0.250 0.000
0.500 0.750 0.000
0.333 0.666 0.375 0.000
```

This means (1) a header line with a name or description of the distance matrix, (2) a line with the names of the entities in the matrix, (3) the lower half of a pairwise distance matrix, including the diagonal, where the order of appearance of elements must match that expected from line 2.

### *Binary matrix from clusters in trees*

This lets you generate a binary matrix from a PHYLIP-format tree file, where every tree will be treated as an OTU, and every cluster found in the tree file as a 'locus'; if a given tree includes a given cluster, the corresponding binary data point will be '1', otherwise it will be '0'. To obtain an explicit listing of clusters in a tree file, please use the Consensus Trees commands available in the **Trees** menu and examine the analysis output file afterwards.

### *Import from count data (tab-separated)*

This option lets you import count data to generate a (dominant) binary data matrix for use with FAMD. This function is useful if you wish to generate a binary data matrix from count data, such as sequence tags derived from next-generation sequencing methods like genotyping by sequencing (GBS) or restriction-site associated DNA tag sequencing (RAD-Seq).

Input data is expected to be in the form of a count table in tab-separated values text file (typically with extension .tsv or .tab), with individuals in columns and tags (=loci) in rows; individual and tag names are expected to be present. An example of such an input file would look like this:

	Ind1	Ind2	Ind3	Ind4	Ind4	Ind5	Ind7
T1	184	116	188	3017	125	115	186
T2	4	1	0	0	0	0	4
T3	0	0	3	4	1	1	2
T4	0	NA	0	17	16	9800	2
T5	8	1	NA	1	7	0	1
T6	0	0	0	1	0	0	0
T7	3	0	4	0	0	2	0
T8	0	0	8	1	1	0	0

T9	1	1	0	1	1	2	1
T10	6	6	2	7	3	3	14

Because people may apply different stringencies, FAMD gives you the options of specifying thresholds for counting data points as presences, absences or ambiguities (missing data), additionally letting you specify an explicit character string that codes for missing data in the input file ("NA" in the above example). You are asked to provide

- threshold for absence ( $T_0$ ): count values will be treated as absent (0) IF  $\leq$  this value
- threshold for presence ( $T_1$ ): count values will be treated as present (1) IF  $\geq$  this value
  - any count values  $x$  in the range  $T_0 < x < T_1$  will be treated as missing data (?)
- missing data character string. Data points matching this are treated as missing data (?)

**Note** that FAMD does not per se impose a limit on the size of the input file or data set it can load. However, it is limited by the system's memory and, since it is a 32-bit program, it cannot use more than ~3Gb of memory. Since FAMD also requires a fair bit of memory during the parsing of input files, it may not be possible to import big population genomic data sets into the current FAMD version. (Certainly, .tsv files > 3Gb will be too large to import.)

## Select File Names

Note that control of input/output file names is also possible via the `Options→I/O Options` command.

### *Select Analysis Output File*

This lets you select a file to which numerical results can be saved. Note that by selecting a file you do not actually save to this file. If you do not select a file here, FAMD will use the default file name `analysis.txt` in the current directory.

### *Select Tree Output File*

This lets you select a file to which trees generated during the analysis will be saved. Note that by selecting a file you do not actually write to this file. If you do not select a file name here, FAMD will use the default file name `outtree.ph` in the current directory. Trees will be saved in PHYLIP format. Tree names will be stored as root labels. Note that not every program may be able to interpret this correctly. You can use programs such as TreeView to open the tree file and convert trees to another format if desired, and in doing so choose to remove branch labels (i.e. tree names).

### *Select Consensus Tree Input File*

This lets you select a file to which contains the trees that will be used as input information for strict and majority rule consensus trees. If you do not select a file name here, FAMD will use the default file name `outtree.ph` (the default tree output file) in the current directory. FAMD expects input trees to be in PHYLIP format.

### *Select Consensus Tree Output File*

This lets you select a file to which consensus trees will be saved. Note that by selecting a file you do not actually write to this file. If you do not select a file name here, FAMD will use the default file name `constree.ph` in the current directory. Consensus trees will be saved in PHYLIP format with the percentages of cluster occurrences saved as branch labels, and not as branch lengths. You can use programs such as TreeView to open the consensus tree file and convert trees to another format if desired.

### *Select Log File*

This lets you select a file to which FAMD - if told to do so - will log all commands it is told to carry out until logging is stopped. If you do not select a file here, FAMD will use the default file name `famdlog.txt` in the current directory.

## Export

This function lets you export your data matrix into a number of file formats that can be read by other analysis program. Please note that support for other file formats cannot be expected to be complete and have only been tested to work with particular versions of targeted programs. However, the different export filters can make your life a lot easier.

### *Nexus*

Exports your data matrix to Nexus format. Programs that have been tested to read exported nexus files are PAUP\* 4.0 beta 10 (Windows), MrBayes 3.1 (Windows) and SplitsTree 4.2 (Windows). The nexus format comes in two flavours ('old' and 'new' nexus format) and some programs may be able to read one, but not the other of these.

FAMD will export your data matrix as space-delimited data (setting the Nexus INTERLEAVE option set), writing a single row of data per individual. The missing data character will be that defined for FAMD. FAMD will write a TAXA block, and will ask you whether to export your data in a CHARACTERS ('new' nexus, recommended for PAUP\*) or DATA block ('old' nexus, recommended for MrBayes). In addition, it will ask you what nexus DATATYPE your data should be assigned to. The available options are: STANDARD and RESTRICTION. STANDARD should be used for use with e.g., PAUP\*. RESTRICTION makes more sense for AFLP data, but is not recognised by PAUP\*. MrBayes does recognise this data type.

If you have already generated a similarity/distance matrix in FAMD, the program will ask you whether to a distance matrix to the nexus file as DISTANCES block. The distance transformation used as defined under FAMD's Options→Distance Transformation option panel.

### *Arlequin Project*

This option lets you save Arlequin Project (ARP) files. This export filter was tested for use with Arlequin 3.0 (Windows). For this option to work, groups of individuals must be defined (see Group Manager section). FAMD does not currently save distance matrices for use with Arlequin.

Depending upon how many hierarchical levels of group you have defined for your data, FAMD may ask you whether to use either the first or the first and the second level of group hierarchy for generating an ARP file. Note that FAMD will export ALL groups at a given level whether or not such groups may 'contradict' each other. FAMD will export all individuals in your current data matrix as haplotypes, even if some of them are not present in any of the exported groups (which should not upset Arlequin). It is the user's responsibility to make sure that the group structure, as exported by FAMD, actually makes sense. FAMD will not check for e.g., individuals which are members of more than one groups (which may be done in FAMD but does not make sense in the context of some of the things you might want to use Arlequin for).

### *Tab-delimited text / R*

This option lets you save the data matrix in tab-delimited text format (typical suffix .tsv or .tab), which makes it easy to load data into R or Excel. The exported data matrix will have loci in rows and individuals in columns, with locus and individual names (row and column names) present. It can be read into R, e.g. by:

```
data <- read.delim ("export_R.tsv")
```

In the data file, band presences and absences are encoded as 1 and 0, respectively, and missing data as NA.

### *Genepop*

This option tells FAMD to generate a genepop-format file. This option was tested using BAPS 3.2 (Windows). Note that FAMD will encode your dominant data as diploid data, the first allele taken from your data matrix, and the second typically encoded as missing data. Since for dominant data, a band absence may be interpreted as a 0/0 genotype, FAMD will ask you whether you wish absences to be

treated thus. Please note, however, that most software manuals suggest simply encoding the second allele as missing data. Using this export filter, presences are encoded as 01, absences as 02 and missing data as 00.

The genepop file generated by FAMD does NOT contain any information about populations, i.e., your data matrix is exported as a single population. If you wish to define additional populations, please edit the FAMD-generated text file, introducing population structures manually using the term 'POP'. An example is provided below

The following is an example of a FAMD-generated genepop-format file with all individuals (IndivA, IndivB, IndivC and IndivD) in one population.

```
FAMD exported data
Locus1, Locus2, Locus3, Locus4, Locus5
POP
IndivA, 0100 0100 0200 0100 0100
IndivB, 0100 0100 0200 0100 0100
IndivC, 0100 0100 0200 0100 0100
IndivD, 0100 0100 0200 0100 0100
```

Suppose you want to have two populations, one consisting of IndivA and Indiv B, and the second one consisting of IndivC and IndivD. To do so, you simply introduce the term 'POP' before the first individual which should belong to population 2 (i.e., IndivC). Your modified genepop file will look as follows:

```
FAMD exported data
Locus1, Locus2, Locus3, Locus4, Locus5
POP
IndivA, 0100 0100 0200 0100 0100
IndivB, 0100 0100 0200 0100 0100
POP
IndivC, 0100 0100 0200 0100 0100
IndivD, 0100 0100 0200 0100 0100
```

### *NTSys-pc*

Using this option, FAMD will write a text file (.NTS) that can be read by NTSYS-pc. This function was tested using NTSYS-pc 2.1 (Windows). FAMD will output labels for loci and individuals. Presences, absences and missing data will be encoded as 1, 0, and -9, respectively.

### *List of OTUs*

This option simply generates a text file containing the names of individuals in your data matrix. The file will not contain the data matrix itself.

### *SynTax*

Use this option to write files that can be read by SYN-TAX. This option was tested using SYN-TAX 2000. Three files will be generated. The first one (default name `syntax.txt`) will contain your data matrix. The other two files, whose name will be based on the name of your syntax file name, will contain the individual and locus label files for use with SYN-TAX. Their default file names are `syntax_indivlabels.txt` and `syntax_rowlabels.txt`, respectively.

### *hindex*

Use this option to generate an input file for hindex. This option was tested using hindex 1.42 (Linux). For this option to work, groups of individuals must be defined (see *Group Manager* section). FAMD will ask you to select two groups that act as parent 1 and parent 2 for the third group, which is the group of individuals that is to be tested using hindex. Two files will be generated, one `filename_indiv.txt` containing the individuals to be tested. Missing data will be encoded 'NA' in this file. The second file, `filename_parents.txt`, will contain information about the parents, i.e. the frequency of a marker at a given locus in parent 1 and parent 2. Marker frequencies are calculated, treating missing data as defined under the *Options* menu. If the *RMD* (replace missing data)

variable under Options→Bootstrapping & Replicates is set, missing data will be replaced according to Options→Missing data replacement before marker frequencies are calculated. If RMD is not set, then missing data will be ignored for parents 1 and 2.

### Structure

This option allows you to generate an input file for Structure, version 2.2 or later. It was tested using Structure 2.3. (This export filter was updated in FAMd 1.21; previous versions of FAMd exported to Structure 2.0 format, which did not really support dominant data.) FAMd will add information to identify band absences as recessive alleles (i.e., use Structure's RECESSIVEALLELES=1 option). You will also be asked about the ploidy level for generating the Structure input file. FAMd's export filter has not been tested for anything higher than diploids, so please do check the file FAMd generates for you. If you have groups defined (see Group Manager section), FAMd will ask you whether this information should be exported as population information for use with Structure (for Structure's POPDATA=1 option etc.). Since FAMd allows an individual to occur in more than one group, an individual will be encoded as belonging to the FIRST group it appears in. Structure populations will be consecutively numbered, starting from 1, depending on the order of appearance of groups in FAMd's Group Manager. If an individual is not assigned to any group, it will be assigned to Structure's 'Population 0'.

To load data into Structure (information here for Structure 2.3), start a new project and follow the wizard and enter number of individuals and loci, ploidy, and as missing data symbol enter "-9". On the next wizard pages, check the options "Row of marker names" and "Row of recessive alleles" (step 3), and then "Individual ID for each individual", and – if you have decided to include population information – "Putative population origin for each individual" (step 4).

### Hickory (Nexus)

To use this option, you need to have groups defined (see Group Manager section). This option will generate a nexus (.NEX) file for use with Hickory. The nexus file generated will contain a TAXA and an ALLELES block.

### Dfdist

This option generates an input file for ddatacal.exe of the dfdist package. If the data matrix contains missing data, then it will be treated differently depending on the current missing data replacement settings. If Options→Bootstrapping & Replicates → RMD (Replace Missing Data) is turned off (checkbox is unchecked), missing data are ignored and the number of band presences ( $P=o$ ) and band absences ( $A=z$ ) will be used to prepare the dfdist input file. Otherwise,  $P$  and  $A$  are calculated as follows:

$$P = o + \rho m \text{ (rounded to the nearest integer)}$$

$$A = o + m + z - P$$

where...

$o$ ... is the number of actual band presences (1)

$z$ ... is the number of actual band absences (0)

$m$ ... is the number of missing data (ambiguities, ?)

$\rho$  ... fraction of missing data to be replaced by band presences (as set in the options)

### BayeScan

This option generates an input file for BayeScan. If the data matrix contains missing data, then it will be treated differently depending on the current missing data replacement settings. If Options→Bootstrapping & Replicates → RMD (Replace Missing Data) is turned off (checkbox is unchecked), missing data are ignored and the number of band presences ( $P=o$ ) and band absences ( $A=z$ ) will be used to prepare the BayeScan input file. Otherwise,  $P$  and  $A$  are calculated as follows:

$$P = o + \rho m \text{ (rounded to the nearest integer)}$$

$$A = o + m + z - P$$

where...

$o...$  is the number of actual band presences (1)

$z...$  is the number of actual band absences (0)

$m...$  is the number of missing data (ambiguities, ?)

$\rho$  ... fraction of missing data to be replaced by band presences (as set in the options)

### *AFLPDat*

This option generates a data file for input into the AFLPDat R package. If groups are defined, then these can be used as population information to be written to the AFLPDat file. Since FAMd allows an individual to occur in more than one group, an individual will be encoded as belonging to the FIRST group it appears in. Since not all functions in AFLPDat are happy about missing data, you may need to replace missing data (if applicable) before saving to AFLPDat format.

### *AFLPop*

This option generates input data files for AFLPop in Excel using the currently defined groups. These files are tab-delimited text files that will be prefixed with the file name stem you select in the export filter, by default `aflpop_*.txt`.

### *Phylip distance matrix*

To use this option, you must already have calculated a (dis)similarity matrix from your input data. This function will save the distance matrix currently in memory in PHYLIP-compatible format. The lower triangle, including the diagonal will be saved. An output file might hence look like the following:

```
3
Indiv1 0.0000000000
Indiv2 0.3500000000 0.0000000000
Indiv3 0.4500000000 0.5500000000 0.0000000000
```

## **Turn log on/off**

This option tells FAMd to turn logging on or off. If enabled, FAMd will log all commands it carries out to a log file (default is `famdlog.txt`). This is useful for purposes of documenting what you are doing. The messages saved to the log files are identical to those displayed on the screen.

## **Exit**

Quits the program. FAMd will not ask you to save data prior to quitting.

## **DataMatrix Menu**

The `DataMatrix` menu lets you view and modify different aspects of your data. Generally, results are displayed on the screen and you are asked whether you would like to save them to the analysis file. If the results of an analysis are too lengthy to be displayed on screen, they will be saved to the analysis output file. If this file already exists you are asked whether data should be appended to it or whether

the file should be overwritten. Alternatively, you can instruct FAMD to always append to or overwrite its output files by setting the respective option under Options→I/O Options ("Overwrite File Prompt").

## Restore Original Matrix

This option lets you restore the original data matrix as first loaded when an input file was opened. It is useful, e.g. after data points have been removed from or replaced in the data matrix.

## Matrix Statistics

This option tells you how many individuals and loci were detected in your input file. It is often useful to use this option to check that the program loads exactly the data set you wanted it to load, or that the data set was loaded completely. In addition to the number of individuals and loci, also the size of the data matrix (i.e. number of data points, given by #individuals × #loci) and the amount of missing data in the data set are displayed.

## Missing Data Statistics

This will write information about missing data to your analysis output file. The output values are the percentages of missing data points in each individual and in each locus.

## Count Bands...

This option provides statistics about your data set and writes these data to the analysis file. The following methods are available:

### *Mean Number of Bands Per Individual*

This option calculates the mean number of bands per individual as an average value for a given defined Group. FAMD outputs the mean number, variance and standard deviation for all Groups that are defined for the current data set. These values are calculated for (a) band presences per individual and (b) band presences and missing data per individual.

### *Polymorphic Bands*

This option outputs the number of polymorphic bands for every Group defined for the current data set. Polymorphic bands (V) are all bands that are not monomorphic:

$$V... \text{ Number of polymorphic bands: } V = N - (P+A)$$

Where...

N... Total number of loci

P... Monomorphic presences: Number of bands that satisfy: IF ( $m \leq o$ ) AND ( $z=0$ ).

A... Monomorphic absences: Number of bands that satisfy: IF ( $m \leq z$ ) AND ( $o=0$ ).

Where  $m = N_z(i)$ , the number of missing/ambiguous data in locus  $i$   
 $z = N_o(i)$ , the number of band absences in locus  $i$   
 $o = N_1(i)$ , the number of band presences in locus  $i$

### *Fixed Bands*

This option outputs the number of fixed bands for every Group defined for the current data set. A fixed band (monomorphic band presence) is defined as any band that satisfies the following:



IF ( $m \leq o$ ) AND ( $z=0$ ).

Where  $m = N_z(i)$ , the number of missing/ambiguous data in locus  $i$   
 $z = N_o(i)$ , the number of band absences in locus  $i$   
 $o = N_1(i)$ , the number of band presences in locus  $i$

### Private Bands

This option outputs the number of private bands for every Group defined for the current data set. A private band is defined as a band which is present only in individuals (OTUs) of the given group but not in any individuals (OTUs) not belonging to the current group. Any band that has at least one band presence in a Group is considered in this method.

Statistics based on private bands are only meaningful if the Groups are defined as mutually exclusive entities.

Example:

In the following data set Group A consists of individuals A1, A2, A3 and Group B of B1, B2, B3.

	A1	A2	A3	B1	B2	B3
Loc01	0	1	1	0	0	1
Loc02	0	1	0	1	0	0
<b>Loc03</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>0</b>	<b>0</b>	<b>0</b>
<b>Loc04</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>1</b>	<b>0</b>
Loc05	0	0	1	0	1	0
*						

Here, Loc03 is a private band for Group A and Loc04 is a private band for Group B.

### Fixed Private Bands

This option outputs the number of fixed private bands for every Group defined for the current data set. A fixed private band is defined as a band that meets the criteria for (a) a fixed band (monomorphic band presence) and (b) a private band at the same time. (Please see above for definitions for these).

In the example above (under *Private Bands*), Loc03 is a fixed private band for Group A, but Loc04 is not fixed private band for Group B, because it satisfies the criteria for a private band but not those for a fixed band.

## Frequency Statistics

### Frequencies per Individuals

This will save the frequency  $p(i)$  of band presences in each individual  $i$  to the analysis output file.

$$p(i) = \frac{N_1(i)}{L} \quad \text{where: } N_1(i) \text{ is the number of band presences (1) in locus } i$$

$L$  is the number of loci

### Frequencies per Loci

This will save the frequency  $p(i)$  of band presences in each locus  $i$  to the analysis output file. You are first asked how you want frequencies to be defined. The first option is:

$$p(i) = \frac{N_1(i)}{n}$$

where:  $N_1(i)$  is the number of band presences (1) in locus  $i$   
 $n$  is the number of individuals

This is intuitive. The second option is:

$$p(i) = \frac{N_1(i)}{\sum_{k=1}^s N_1(k)}$$

where:  $N_1(i)$  is the number of band presences (1) in locus  $i$

$s$  is the number of loci in the data set

$\sum N_1(i)$  is the total number of presences in the data set

The reason for giving this option of calculating frequencies lies in the formula by Bowman *et al.* (1969) for calculating the variance associated with Shannon's index.

## Replace Missing Data

Selecting this option replaces missing data in your data set according to the settings specified under `Options → Missing data replacement`. By doing so, the data matrix is modified and missing data in it replaced by discrete characters. Therefore, if you wish to proceed with the analysis using the original data matrix, you must first restore it using `DataMatrix → Restore Original Matrix`.

## Resample Loci (Bootstrap Replicate)

Selecting this option allows you to manually generate a bootstrap replicate of your current data matrix by resampling its loci. The resulting data matrix will have the same dimensions (individuals x loci) as the one you started out with, and the individuals will be the same. However, the composition of loci will be randomly resampled, meaning that for every locus in the data matrix you started out with, there will be data from one randomly selected locus from the set of available loci, where every source locus is available for every random draw.

Thus, a data matrix of  $n$  individuals x Loc1, Loc2, Loc3, Loc4, Loc5 may end up, for example having a locus composition of Loc2, Loc3, Loc3, Loc3, Loc1.

## Remove Individuals...

This option removes individuals from the data matrix that match the criteria applied. If you wish to undo such a data removal to continue with the original data matrix, you must restore it using `DataMatrix → Restore Original Matrix`.

*With missing data...*

This will remove individuals that have a percentage of missing data that is greater than the specified threshold percentage.

*With band frequency below...*

This will remove individuals whose frequency of band presences is below the specified threshold percentage. This may be useful e.g. to remove individuals whose data stems from poor AFLP reactions which (if they were scored) will often display a strikingly lower number of band presences.

*With band frequency above...*

This will remove individuals whose frequency of band presences is greater than the specified threshold percentage.

## Remove Loci...

This option removes loci from the data matrix according to your choice of options. If you wish to undo such a data removal to continue with the original data matrix, you must restore it using `DataMatrix` → `Restore Original Matrix`.

### *With missing data...*

Removes loci that have a percentage of missing data that is greater than the specified threshold percentage.

### *Monomorphic...*

Removes monomorphic loci from the data set. This should be done for instance prior to calculation of Shannon's index (non-bootstrapping version). A locus with a monomorphic presence or absence is defined as follows:

Monomorphic presence: IF  $(m \leq o)$  AND  $(z=0)$ .

Monomorphic absence: IF  $(m \leq z)$  AND  $(o=0)$ .

Where  $m = N_2(i)$ , the number of missing/ambiguous data in locus  $i$   
 $z = N_0(i)$ , the number of band absences in locus  $i$   
 $o = N_1(i)$ , the number of band presences in locus  $i$

All other bands are considered polymorphic.

### *Monomorphic Absences...*

Removes monomorphic absences from the data set, where a band is considered to be a monomorphic absence

IF  $(m \leq z)$  AND  $(o=0)$ .

Where  $m = N_2(i)$ , the number of missing/ambiguous data in locus  $i$   
 $z = N_0(i)$ , the number of band absences in locus  $i$   
 $o = N_1(i)$ , the number of band presences in locus  $i$

### *Monomorphic Presences...*

Removes monomorphic presences from the data set, where a band is considered to be a monomorphic presence

IF  $(m \leq o)$  AND  $(z=0)$ .

Where  $m = N_2(i)$ , the number of missing/ambiguous data in locus  $i$   
 $z = N_0(i)$ , the number of band absences in locus  $i$   
 $o = N_1(i)$ , the number of band presences in locus  $i$

### *With band frequency below...*

This will remove loci whose frequency of band presences is below the specified threshold percentage.

### *With band frequency above...*

This will remove loci whose frequency of band presences is greater than the specified threshold percentage.

*With more missing data than presences*

This will remove loci if the number of missing data points is greater than the number of band presences.

## Pairwise Individual Comparison

This option brings up a dialogue box that allows the pairwise comparison of two selected individuals from the current data set. For every selected individual, the percentages of band presences (%P), band ambiguities/missing data (%MD) and band absences (%A) will be displayed. The number of band differences excluding and including missing data, respectively, will be displayed: Here, %Differences(0/1) is the percentage of different bands excluding missing data, and %Differences(0/1/?) is the percentage of different bands including missing data.

If  $s$  is the number of loci and  $n_{xy}$  is the number of bands that are of state  $x$  in the first and of state  $y$  in the second individual included in the comparison, then the differences are calculated as follows:

$$\%Differences(0/1) = 100 * (n_{01} + n_{10}) / s$$

Here, a comparison of a band presence (1) or absence (0) with an ambiguity (?) is not counted as a difference.

$$\%Differences(0/1/?) = 100 * (n_{01} + n_{10} + n_{1?} + n_{?1} + n_{0?} + n_{?0}) / s$$

Here, a comparison of a band presence (1) or absence (0) with an ambiguity (?) is counted as a difference.

Furthermore, this dialogue box shows the standard, minimum, maximum and average distance (as currently selected) and a table of all possible pairwise character state combinations among the individuals selected for comparison.

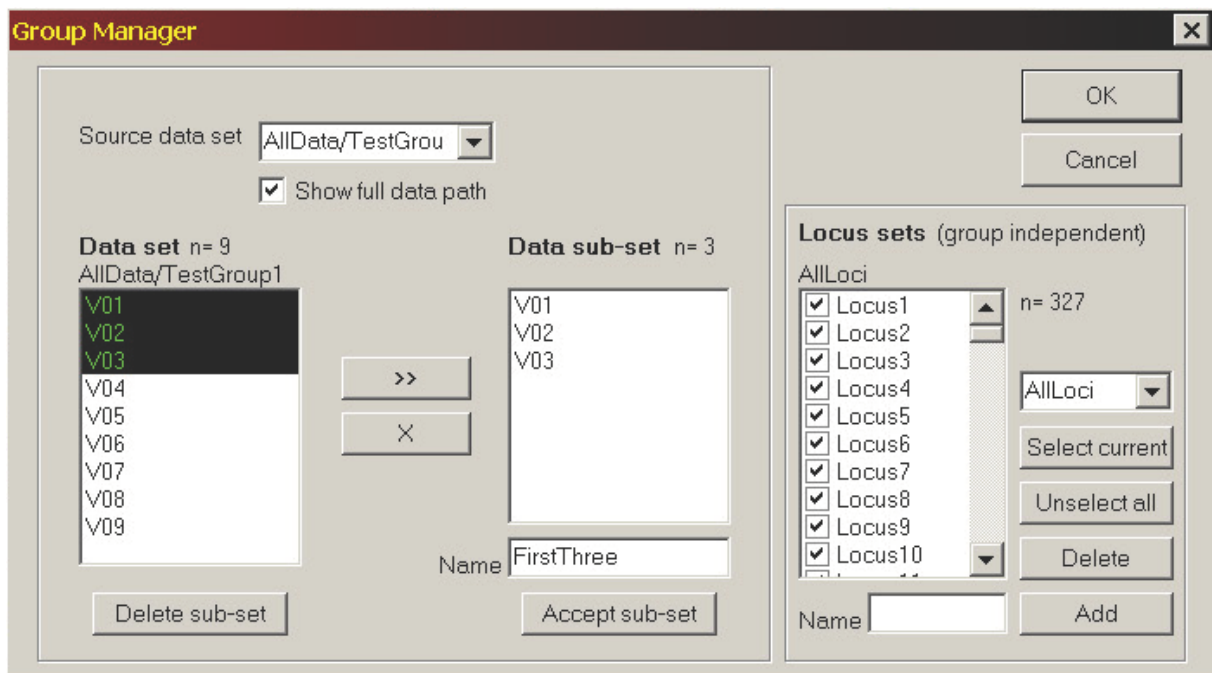
## Group Manager

The group manager serves 3 functions:

- i) to define groups of individuals (herein termed 'Groups')
- ii) to define groups of loci (herein termed 'LocusSets')
- iii) to allow the selection of a LocusSet

If Group or LocusSet information was read in from the input data file, then these groups will appear be available in the Group Manager. Groups of samples can then be selected using the `DataMatrix → Select Group Only → [Group Name]` command. Support for sample or locus groups is in many ways not yet satisfactory and may therefore be subject to change in future versions of FAMd.

To aid description of the Group Manager, a screen-shot is provided below.



### Defining Groups (groups of individuals)

The left side of the Group Manager dialogue box lets you define groups. This allows for selections to be made and enables the comparisons of e.g., different populations. The topmost control lets you select a source data set. If you have no groups defined yet, this source data set will be “AllData”. This is a built-in group that cannot be modified. It is equivalent to all individuals in the current data set. Groups of individuals are hierarchical, with “AllData” acting as the base group. By contrast, none of the selections made using a the Group Manager affect “AllData”.

You can select a group that is a subset of a given group (AllData or any of the groups you have defined) by selecting that group in the topmost control titled “**Source data set**”. All individuals contained in this group are then displayed in the leftmost list, titled “**Data set**”. You can select one or multiple individuals in this list, press the “>>” button, and it/they will appear in the “**Data sub-set**” list. To remove an individual from that list, select it and press the “X” button. If the “**Data sub-set**” list contains those individuals that you wish to be included in your group, enter a name for the group in the empty field titled “**Name**” below the “**Data sub-set**” list and click on the “**Accept sub-set**” button. The “**Data sub-set**” list will then be cleared, and the newly defined group available in the “**Source data set**” control. Groups will appear in that control with their name indicating their placement in the hierarchy in the form of a path name. For instance, if you define a group named, “FirstGroup”, it will have a complete path name of “AllData/FirstGroup”. The same name will also be displayed in other FAMD functions using groups. You can choose to display only the last part of the group name (i.e., the bit you entered) in the Group Manager (and other functions in FAMD) by unchecking the “**Show full data path**” checkbox under the “**Source data set**” control.

To delete a group from the group manager, select the group using the “**Source data set**” control and then press “**Delete sub-set**”.

**NB:** Your selections will only be kept if you press “**OK**” to quit the Group Manager, clicking “**Cancel**” will discard all changes you made. Please also note that FAMD will NOT check whether any of your groups contains one individual several times (although this may result in errors later on).

Note that, “AllData” may or may not be equivalent to the data file initially loaded: Any loci or individuals removed by any functions of the DataMatrix submenu are not present in “AllData”. (If necessary, you can restore the original data matrix via the DataMatrix → Restore Original Matrix command.) Please also note that when saving a file including group information, “AllData” is NOT saved.

### *Defining and Selecting LocusSets (groups of loci)*

The right side of the Group Manager dialogue box lets you define `LocusSets`, that is, groups of loci. This allows the inclusion/exclusion of loci from the data set. In contrast to groups of individuals, `LocusSets` are not hierarchical but simply reflect selections of certain loci that can be stored and loaded. The rightmost list is a list of all loci in the data set. “AllLoci” is a predefined `LocusSet` that cannot be modified; it selects all loci in the data set. You can unselect a single locus by unchecking it, or unselect all loci by clicking on the “**Unselect all**” button. If you are satisfied with your selection, enter a name in the field titled “**Name**” and click on the “**Add**” button. The locus set should now be selectable from the combo box above the “**Select current**” button. To delete a `LocusSet`, select it using that same combo box and click on the “**Delete**” button.

To select a `LocusSet`, select it from the combo box above the “**Select current**” button and then click on the “**Select current**” button. This will exclude unselected loci from any analysis you perform subsequently (until you again select the “AllLoci” `LocusSet`).

## Select Group Only

This option lets you select a group that you have either previously defined using the `Group Manager` or that was read in with the input file with the effect that subsequent analyses will be restricted to the currently selected group of individuals. The `AllData` group refers to all the individuals present in your current data matrix. Note, however, that individuals or loci that have been removed using the `DataMatrix → Remove Loci` or `Remove Individuals` functions have been excluded also from the `AllData` data structure and that once missing data has been replaced using the `DataMatrix → Replace Missing Data` functions is not restored by selecting `AllData`. To restore a data matrix in its entirety, use the `DataMatrix → Restore Original Matrix` function.

## Select LocusSet Only

This option lets you select a `LocusSet` that you have either previously defined using the `Group Manager` or that was read in with the input file with the effect that subsequent analyses will be restricted to the currently selected group of individuals. The `AllLoci` `LocusSet` refers to all the loci present in your current data matrix (similar to `AllData` for Groups of individuals).

## Group-based Profiles

To use this function, groups must be defined (see `Group Manager` section). Band profiles will be generated as detailed below and every defined group (irrespective of its hierarchical level) will be represented as one ‘pooled’ individual whose genotype is the band profile calculated from the information of all individuals in the group. Note that by using this feature, your data matrix (`AllData`) containing ‘real’ individuals will be overwritten. The original data matrix can be reloaded using the `DataMatrix → Restore Original Matrix` function.

### *Additive Band Profile*

This is (at least theoretically) the *in silico* equivalent of pooling DNA of a group of individuals and subjecting the pooled DNA to a dominant fingerprinting method. The resulting band profile will have a band presence at a given marker locus if at least one individual of the group has a band presence at this marker locus. The resulting band profile will NOT contain any missing data.

### *Fixed Band Profile*

The resulting band profile will only contain band presences at those loci that are fixed in the group. A band presence will be considered fixed if there are no band absences at the locus and the number of actual band presences is greater than the number of missing data points. The resulting band profile will NOT contain any missing data.

## Analysis Menu

The **Analysis** menu lets you perform different calculations on your current data set, such as different calculation of similarity/distance matrices which can then be subjected, e.g. to UPGMA tree reconstruction in the **Trees → UPGMA** menu. To control which (dis)similarity measures and, if applicable, distance transformations are being used, use **Options→(Dis) Similarity Coefficients** and **Options→Distance Transformation** commands.

## Standard Similarity

This calculates a similarity (or distance) matrix from your current data matrix, based on the coefficient selected under **Options→(Dis) Similarity Coefficients**. The same dialogue lets you check options to save the similarity matrix - and/or the matrix after the distance transformation has been applied (**Options→Distance Transformation**) - to the analysis output file.

The default coefficient selected is Jaccard's similarity coefficient.

Jaccard's similarity coefficient for a pair of individuals  $i$  and  $j$  is defined as:

Standard Jaccard:

$$S_{ij,Jaccard} = \frac{n_{11}}{n_{11} + n_{01} + n_{10}}$$

where  $n_{xy}$  is the number of characters that have state  $x$  in individual  $i$  and state  $y$  in individual  $j$ . Possible character states are band presence (1), band absence (0) and missing data (?).

Similarly, other the similarity coefficients are defined as follows.

Standard Dice/Sørensen:

$$S_{ij,Dice} = \frac{2n_{11}}{2n_{11} + n_{01} + n_{10}}$$

Standard SMC (Simple Matching Coefficient):

$$S_{ij,SMC} = \frac{n_{11} + n_{00}}{n_{11} + n_{01} + n_{10} + n_{00}}$$

The following are distance measures rather than similarity measures:

Nei-Li distance (following Nei & Li, 1979) for restriction-site data:

$$d_{ij,NeiLi} = -\frac{3}{2} \ln \frac{4\left(\frac{2n_{11}}{2n_{11} + n_{10} + n_{01}}\right)^{1/2r} - 1}{3} = -\frac{3}{2} \ln \frac{4(s_{ij,Dice})^{1/2r} - 1}{3}$$

where  $r$  is the length of the restriction enzyme's recognition sequence, e.g. for *EcoRI*, recognising GAATTC,  $r$  would be 6.

NB: The implementation in PAUP\* 4.0 beta 10 assumes that  $r = 6$  (J. Wilgenbusch, pers. comm.). For comparability, this is also the default value for  $r$  in FAMD, but you can change it using the Similarity Coefficient Selection dialogue box.

Standard Euclidean distance:

$$d_{ij, Euclid} = \sqrt[n]{n_{10} + n_{01}}$$

Standard Squared Euclidean distance:

$$d_{ij, SqEuclid} = d_{ij, Euclid}^2 = n_{10} + n_{01}$$

## Minimum Similarity

This calculates a minimum similarity (or maximum distance) matrix from your current data matrix (Schl ter & Harris, 2006), based on the coefficient selected under Options→(Dis) Similarity Coefficients. The same dialogue lets you check options to save the similarity matrix - and/or the matrix after the distance transformation has been applied (Options→Distance Transformation) - to the analysis output file.

The default coefficient used the minimum Jaccard's coefficient, taking into account missing data; for a pair of individuals  $i$  and  $j$ , it is defined as:

Minimum Jaccard:

$$S_{ij, \min} = \frac{n_{11}}{n_{11} + n_{01} + n_{10} + n_{?1} + n_{1?} + n_{?0} + n_{0?}}$$

where  $n_{xy}$  is the number of characters that have state  $x$  in individual  $i$  and state  $y$  in individual  $j$ . Possible character states are band presence (1), band absence (0) and missing data (?).

If the data set does not contain missing data, the minimum Jaccard value will be identical to the standard Jaccard's coefficient. (Note that in the above case, the value  $n_{??}$  is not used because for it could serve both to increase and to decrease the measure under consideration, and as such is not overly useful for estimating a minimum or maximum.)

Similarly, other the minimum similarity coefficients are defined as follows:

Minimum Dice:

$$S_{ij, \min} = \frac{2n_{11}}{2n_{11} + n_{01} + n_{10} + n_{1?} + n_{?1} + n_{0?} + n_{?0}}$$

Minimum SMC:

$$S_{ij, \min} = \frac{n_{11} + n_{00}}{n_{11} + n_{01} + n_{10} + n_{1?} + n_{?1} + n_{0?} + n_{?0} + n_{00}}$$



The maximum distance coefficients are defined as follows:

Maximum Nei-Li distance:

$$d_{ij,\max} = -\frac{3}{2} \ln \frac{4(s_{ij,Dice,\min})^{1/2r} - 1}{3}$$

where  $r$  is the restriction enzyme's recognition site length

Maximum Euclidean distance:

$$d_{ij,\max} = \sqrt[2]{n_{10} + n_{01} + n_{1?} + n_{?1} + n_{0?} + n_{?0}}$$

Maximum squared Euclidean distance:

$$d_{ij,\max}^2 = n_{10} + n_{01} + n_{1?} + n_{?1} + n_{0?} + n_{?0}$$

## Maximum Similarity

This calculates a maximum similarity (or minimum distance) matrix from your current data matrix (Schlüter & Harris, 2006), based on the coefficient selected under Options→(Dis)Similarity Coefficients. That same dialogue lets you check options to save the similarity matrix - and/or the matrix after the distance transformation has been applied (Options→Distance Transformation) - to the analysis output file.

The default coefficient used the maximum Jaccard's coefficient, taking into account missing data; for a pair of individuals  $i$  and  $j$ , it is defined as:

Maximum Jaccard:

$$s_{ij,\max} = \frac{n_{11} + n_{?1} + n_{1?} + n_{??}}{n_{11} + n_{01} + n_{10} + n_{?1} + n_{1?} + n_{??}}$$

where  $n_{xy}$  is the number of characters that have state  $x$  in individual  $i$  and state  $y$  in individual  $j$ . Possible character states are band presence (1), band absence (0) and missing data (?).

If the data set does not contain missing data, the maximum Jaccard value will be identical to the standard Jaccard's coefficient.

Similarly, other the maximum similarity coefficients are defined as follows:

Maximum Dice:

$$s_{ij,\max} = \frac{2(n_{11} + n_{1?} + n_{?1} + n_{??})}{2(n_{11} + n_{1?} + n_{?1} + n_{??}) + n_{01} + n_{10}}$$

Maximum SMC:

$$s_{ij,max} = \frac{n_{11} + n_{1?} + n_{?1} + n_{0?} + n_{?0} + n_{??} + n_{00}}{n_{11} + n_{01} + n_{10} + n_{1?} + n_{?1} + n_{0?} + n_{?0} + n_{??} + n_{00}}$$

The minimum distance coefficients are defined as follows:

Minimum Nei-Li distance:

$$d_{ij,min} = -\frac{3}{2} \ln \frac{4(s_{ij,Dice,max})^{1/2r} - 1}{3}$$

where  $r$  is the restriction enzyme's recognition site length

The standard Euclidean and squared Euclidean distances already represent the minimum possible value for these distances, so the minimum [squared] Euclidean measure is equivalent to the standard [squared] Euclidean distance measure.

## Average Similarity

This function calculates a similarity matrix from your current data matrix based on the average similarity  $s_{ij}^*$  (or distance  $d_{ij}^*$ ) coefficient (Schlüter & Harris, 2006) you have selected under Options → (Dis) Similarity Coefficients. For this, minimum ( $s_{ij,min}$ ) and maximum ( $s_{ij,max}$ ) similarity coefficients are calculated. The average similarity coefficient is defined as the arithmetic average of values drawn randomly (uniformly) from the interval  $[s_{ij,min}; s_{ij,max}]$ . The number of random draws, JC can be defined in the same options page (Options → (Dis) Similarity Coefficients) dialogue box. In addition to the average similarity value, the associated variance and standard deviation (square root of variance) is calculated and output to the analysis file (provided the respective option is turned on). If the data set does not contain missing data, the average Jaccard value will be identical to the standard similarity coefficient and variance and standard deviation will be zero. The same procedure is followed also if the selected coefficient is a distance rather than a similarity.

## Null Allele Frequencies

This function estimates the null allele frequency in your data set and outputs it to your analysis file. It assumes that the data are from dominant fingerprints in a diploid organism. The recommended method is the Bayesian method (Zhivotovsky, 1999) with a prior from among-population information if groups are defined, or with another prior if there are no groups defined. For a comparison of these methods, see Zhivotovsky (1999). In all cases, the number of band absences,  $Z$ , are used to estimate the null allele frequency,  $q$ , at a locus. If the RMD option (Options → Bootstrapping & Replicatea → RMD ["Replace missing data"]) is turned on, then  $Z$  is calculated as

$$Z = z + (1 - \rho) m$$

where  $z$  .. number of band absences  
 $\rho$  .. fraction of missing data to be replaced by band presences  
 $m$  .. number of ambiguous bands

If RMD is turned off,  $Z = z$ . The parameter  $\rho$  (default 50%) can be modified under Options → Missing data replacement.

### Square root

This estimates the frequency of null alleles at a given locus as

$$q = \sqrt[3]{Z}$$

### *Lynch-Milligan*

This estimates the frequencies of null alleles and the associated variance at a locus using the method of Lynch & Milligan (1994), ignoring all loci with  $Z < 3$  in a given population.

$$q = \frac{\sqrt[2]{x}}{\left(1 - \frac{\text{var}(x)}{8x^2}\right)} = \frac{\sqrt[2]{x}}{\left(1 - \frac{1-x}{8Nx}\right)}$$

and

$$\text{var}(q) = \frac{1-x}{4N}$$

where  $N$ ... population size  
 $x$ ...  $Z/N$  = frequency of band absences

### *Bayesian (uniform prior)*

This estimates the frequencies of null alleles,  $q$ , at a given locus in a population and the associated squared standard error,  $s_q^2$ , as described by Zhivotovsky (1999), using a uniform prior.

$$q = \frac{B(Z+1, N-Z+1)}{B(Z+0.5, N-Z+1)}$$

and  $s_q^2 = \frac{B(Z+1.5, N-Z+1)}{B(Z+0.5, N-Z+1)} - q^2$

where  $B(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$  is the Beta-distribution. FAMD uses the beta (a,b) function implemented in J. Debord's TPMath library.

$Z$  = number of band absences (as defined above)

$N$  = population size

### *Bayesian (among-population prior)*

This estimates the frequencies of null alleles,  $q$ , and the associated squared standard error,  $s_q^2$ , at a given locus in a population as described by Zhivotovsky (1999), using a non-uniform prior derived from among-population information. This function is therefore only available, if groups (populations) are defined. `AllData` will not be used as a valid population for this. If necessary, FAMD will perform a correction of values as suggested in Zhivotovsky's (1999) note 2, using a correction constant (default 0.01, can be changed under `Options→Allele Frequencies`).

$$q = \frac{B(Z+a+0.5, N-Z+b)}{B(Z+a, N-Z+b)}$$

and  $s_q^2 = \frac{B(Z+a+1, N-Z+b)}{B(Z+a, N-Z+b)} - q^2$

where  $B(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$  is the Beta-distribution. FAMD uses the beta (a,b) function implemented in J. Debord's TPMath library.

$Z$  = number of band absences (as defined above)

$N$  = population size (of the current population  $i$ , the same as  $n_i$  below)

and 
$$a = \bar{R} \left( \frac{\bar{R}(1 - \bar{R})}{\sigma_R^2} - 1 \right)$$

$$b = (1 - \bar{R}) \left( \frac{\bar{R}(1 - \bar{R})}{\sigma_R^2} - 1 \right)$$

where  $R_i = \frac{Z_i}{n_i}$  = fraction of band absences in population  $i$

$$\bar{R} = \sum_{i=1}^k \frac{n_i}{n} R_i = \sum_{i=1}^k \frac{Z_i}{n} \quad \text{FAMD actually calculates the latter.}$$

$$\sigma_R^2 = \left( \sum_{i=1}^k \frac{n_i}{n} R_i^2 \right) - \bar{R}^2 = \left( \sum_{i=1}^k \frac{Z_i^2}{nn_i} \right) - \bar{R}^2$$

FAMD actually calculated the latter. FAMD will issue a warning if  $\sigma^2 < 0$ , set  $\sigma$  to 0 and attempt to re-calculate  $\sigma$  with a correction constant. If this also fails, FAMD will issue an error message.

$$n = \sum_{i=1}^k n_i = \text{sum of population sizes, } n_i, \text{ of the } k \text{ populations}$$

#### *Bayesian (among-locus prior)*

This estimates the frequency of null alleles,  $q$ , and the associated squared standard error,  $s_q^2$ , at a given locus as described by Zhivotovsky (1999), using a non-uniform prior derived from among-locus information. If necessary, FAMD will perform a correction of values as suggested in Zhivotovsky's (1999) note 2, using a correction constant (default 0.01; can be changed under `Options→Allele Frequencies`).

$$q = \frac{B(Z + a + 0.5, N - Z + b)}{B(Z + a, N - Z + b)}$$

and 
$$s_q^2 = \frac{B(Z + a + 1, N - Z + b)}{B(Z + a, N - Z + b)} - q^2$$

where  $B(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$  is the Beta-distribution. FAMD uses the beta (a,b) function implemented in J. Debord's TPMath library.

$Z$  = number of band absences (as defined above)

$N$  = population size (of the current population  $i$ , or of the entire data set if no Groups are defined; the same as  $n_i$  below)

and 
$$a = \bar{R} \left( \frac{\bar{R}(1 - \bar{R})}{\sigma_R^2} - 1 \right)$$

$$b = (1 - \bar{R}) \left( \frac{\bar{R}(1 - \bar{R})}{\sigma_R^2} - 1 \right)$$

The remaining parameters are similar to the among-population prior (see above), but summing over loci instead of populations:

$$R_j = \frac{Z_j}{n_j} = \text{fraction of band absences in locus } j.$$

Note here that  $n_j$  is actually constant, and that  $n_j = n_i$  (the size of the population  $i$  under consideration).

$$\bar{R} = \sum_{j=1}^s \frac{n_j}{n} R_j = \sum_{j=1}^s \frac{Z_j}{n}$$

FAMD actually calculated the latter.

$$\sigma_R^2 = \left( \sum_{j=1}^s \frac{n_j}{n} R_j^2 \right) - \bar{R}^2 = \left( \sum_{j=1}^s \frac{Z_j^2}{nn_i} \right) - \bar{R}^2$$

FAMD actually calculated the latter. FAMD will issue a warning if  $\sigma^2 < 0$ , set  $\sigma$  to 0 and attempt to re-calculate  $\sigma$  with a correction constant. If this also fails, FAMD will issue an error message.

$$n = \sum_{j=1}^s n_j = \sum_{j=1}^s n_i = sn_i = \text{sum of population sizes (since } n_j = n_i, \text{ the size of the population under consideration, is the same for all loci) across all loci, } s.$$

## Shannon's Index

This function calculates Shannon's index and its variance from your current data set. You should always first remove monomorphic loci manually from the data matrix, using the respective function in the `DataMatrix` menu. This is because depending on your definition of band frequencies, monomorphic loci may still contribute to the sum calculated (although they aren't supposed to), which will artificially change Shannon's index. FAMD 1.1 will, however, warn you if your data matrix still contains monomorphic bands. Note that monomorphic bands, if removed will not be present in the data matrix after this function has completed the calculation of Shannon's index.

Shannon's index is defined as:

$$I \approx - \sum_{i=1}^s p_i \log_2 p_i$$

where  $I$  is Shannon's index (often also referred to as  $H_{Sh}$ )  
 $p_i$  is the frequency of band presences in locus  $i$ , as defined as in the `Shannon Scaling...` dialogue box  
 $s$  is the number of loci  
 $\log_2 x = \text{ld } x$  is the logarithm to base 2.

The associated variance is calculated using the formula of Bowman *et al.* (1969):

$$\text{var}(I) \approx \frac{\sum_{i=1}^s p_i \log_2^2 p_i - \left( \sum_{i=1}^s p_i \log_2 p_i \right)^2}{n} + \frac{s-1}{2n^2} = \frac{\sum_{i=1}^s p_i \log_2^2 p_i - I^2}{n} + \frac{s-1}{2n^2}$$

where  $I$  is Shannon's index  
 $p_i$  is the frequency of band presences in locus  $i$ , as defined under  
 Options → Shannon Scaling  
 $s$  is the number of loci  
 $n$  is the number of individuals  
 $\log_2 x = \lg x$  is the logarithm to base 2.

The calculation of this variance will only work if  $p_i$  is defined as band presences in a locus relative to all band presences in the data set. Otherwise, doing the calculation may result in a negative value. The standard deviation is calculated as the square root of the variance.

## ML Hybrid Index

This function calculates a maximum likelihood-based hybrid index for dominant data similar to the hindex software (Buerkle, 2005). (If there are no missing data, and accounting for possible rounding/precision differences, the values obtained from FAMD and hindex should be identical.) You will be asked to specify two Groups that identify the putative parental populations and a Group that is to be tested (i.e. the Group for which hybrid index values will be calculated). The results will be written to the analysis output file.

This function uses the parameter `RMD` set in the Options → Bootstrapping & Replicate options dialogue box. If `RMD` is set, missing data will be replaced internally according to the current settings, otherwise missing data are ignored in the same fashion as in the hindex software.

The results contain lists of poor and excluded loci for hybrid index calculation (as in hindex). The actual data table consist of a hybrid index estimate, a lower and upper bound, a  $\ln$  Likelihood value, along with Error columns which indicate whether calculations converged onto a result ('OK'), did not converge, were obtained by a correction, or whether a more serious error occurred.

*Implementation details:* FAMD internally uses the `Bisect2` and `GoldSearch2` functions in `TPMath` by Jean Debord, in order to find maximum likelihood values and confidence bounds.

## ML Population Reallocation

This function performs maximum likelihood (re)allocation of individuals to a set of possible source populations, using the same methods as in AFLPOP software (Duchesne & Bernatchez, 2002). During this process, any tested individual is excluded from all populations for the purpose of allele frequency calculation. The user is asked to select a set of defined Groups (here assumed to be populations) to individuals may be assigned to. (*Note:* each individual in the data set will be allocated to these target populations; if you are only interested in the (re)allocation of a subset of individuals, simply disregard those which are unimportant to you.)

This function uses options set in the Options → Population Reallocation tab. This includes a choice of the allele frequency correction method: either the formula provided in the Duchesne & Bernatchez (2002) paper [ $1/(1+N)$ , where  $N$  is population size], or a user defined  $\epsilon$  correction value. The user can also determine the minimum difference in  $\log_{10}$ -likelihood value between best and second most likely populations for an allocation to be performed (default: 2). If missing data are

present, missing data points are evaluated as being equal to the user-defined missing data replacement probability (set under `Options→Missing Data Replacement`).

FAMD outputs the results of this analysis to the analysis output file, providing a table of  $\log_{10}$ -likelihood values for the allocation likelihood of each individual to each target population. It also lists the log-likelihood difference among best and second-best populations for the allocation, along with the null hypothesis of population membership ("Group\_H0") and the likeliest allocation. Note that if the log-likelihood difference among the best two populations is less than the specified threshold value, an individual will not be allocated to any population ("None" in the output).

Small differences in numeric output between FAMD and AFLPOP may be observed for the results of this procedure; they are most likely due to differences in the internal precision of computation and rounding.

## AMOVA

This function carries out an AMOVA (Analysis of Molecular Variance) analysis as described in Excoffier *et al.* (1992). For this to work you must have groups defined (see `Group Manager` section). Note that for AMOVA results to be meaningful, group must be defined that correctly reflect the population structure you wish to be tested. This means that all individuals that are present in the data set (in the `AllData` structure) must be partitioned into one and only one group. If there are more levels of group hierarchy, again all individuals in all groups must be assigned to one and only one sub-group. FAMD will not check whether these criteria are met - this is your responsibility!

The AMOVA calculation does not require you to previously calculate a distance matrix since the function calculates its own distances as defined under `Options→Similarity Coefficients`. The similarity preference selection under `Options→AMOVA` defines whether an AMOVA will be conducted on standard, minimum, maximum or average similarities. Note that a minimum similarity naturally goes together with a maximum distance if your selected coefficient is a distance measure.

FAMD will ask you whether or not to save AMOVA results to your analysis file.

The AMOVA SSD terms (see Excoffier *et al.*, 1992) operate on squared distances ( $\delta_{ij}^2$ ). FAMD will square any input distance value, so there is no need for you to manually select the squared distance transformation mode. In other words, the relationship is as follows: (i) FAMD calculates a similarity ( $s_{ij}$ ) or distance ( $d_{ij}$ ) matrix. (ii) FAMD carries out a distance transformation (a function of the similarity or primary distance) as  $\delta_{ij}=d'_{ij}=f(s_{ij})$  or  $\delta_{ij}=d'_{ij}=f(d_{ij})$  and (iii) FAMD performs the AMOVA on  $\delta_{ij}^2$ .

You should be aware that (a) the calculated  $\phi_{ST}$  etc. values (genotypic variation) cannot be directly compared with  $F_{ST}$  etc. values estimated by different procedures, (b) different distance measures will obviously lead to different AMOVA values (although values based on different input distances will be correlated), and (c) that depending on your combination of parameters,  $\delta_{ij}$  may not be a Euclidean metric which could potentially impact on the AMOVA.

Please note also that AMOVA values can differ between FAMD and, e.g. Arlequin, if a data set contains missing data because programs such as Arlequin may treat missing data differently. If a data set does not contain missing data and all groups are set up correctly, then AMOVA results between FAMD and Arlequin should be identical. Arlequin uses the standard Euclidean distance with no distance transformation ( $d'_{ij}=d_{ij}$ ).

## Pairwise PhiST

This function calculates the pairwise  $\phi_{ST}$  values between populations using an AMOVA (see the AMOVA section for details) and the current distance as set for AMOVA under `Options → (Dis)Similarity Coefficients.`, and the (dis)similarity mode preference as set under `Options → AMOVA`. The current version of FAMD will calculate pairwise values for all groups defined in for your

data (see `Group Manager` section), except `AllData`. It is the user's responsibility to make sure that groups are defined in a meaningful way. Please note that using this function will overwrite the data matrix (and individual names) in memory. Therefore, if you need these later on, you will have to restore your original data matrix using the `DataMatrix → Restore Original Matrix` command.

Pairwise  $\Phi_{ST}$  values are calculated as the proportion of population variance due to among population variation [i.e.  $\Phi_{ST} = V_a/(V_a+V_b)$ ]. The total population that is used to calculate  $\Phi_{ST}$  among Pop1 and Pop2 conforms to Pop1 + Pop2 (instead of `AllData`).

## Population Distance

FAMD uses all defined groups (but not `AllData`; see `Group Manager` section for information about groups) to calculate a population distance matrix from allele frequency data. The default method for estimation of allele frequencies is the Bayesian method with a nonuniform prior from among-population information (for more details see `Null allele frequencies`). Currently, only one population distance is implemented. It is a chord distance based upon the single-locus chord distance by Cavalli-Sforza & Edwards (1967).

There are different versions for the chord distance across many loci. FAMD implements the distance as described e.g. by Takezaki & Nei (1996) which is an arithmetic average over the individual locus chord distances.

The formula employed is:

$$d_{xy} = \frac{2}{\pi L} \sum_{j=1}^L \sqrt[2]{2 \left( 1 - \sum_{i=1}^{a_j} \sqrt[2]{p_{xi} p_{yi}} \right)} = \frac{2}{\pi L} \sum_{j=1}^L \sqrt[2]{2 \left( 1 - \sqrt[2]{q_{xj} q_{yj}} - \sqrt[2]{(1 - q_{xj})(1 - q_{yj})} \right)}$$

where..

$d_{xy}$ ...	distance between populations x and y
$L$ ...	number of loci
$a_j$	number of alleles at locus j
$p_{xi}$	frequency of allele i (at locus j) in population x
$q_{xj}$	null allele frequency at locus j in population x

## Trees Menu

This menu lets you carry out tree-based analyses (UPGMA and consensus methods) as well as principal coordinate analysis which, admittedly, has very little to do with trees.

### UPGMA

This generates a UPGMA (unweighted pair group method using arithmetic averages) tree from a distance [default distance= 1-similarity(Jaccard)] matrix that you have generated previously and stores the tree in the tree output file you have defined (or else to the default tree output file). The distance transformation applied can be changed under `Options → Distance Transformation`.

The algorithm implemented is a modified UPGMA algorithm that can generate multifurcating trees, in contrast to a strictly bifurcating implementation. This means that if there are two or more equally good choices for clustering groups of individuals, this will be realised as a multifurcation in the tree, rather than randomly choosing one of the possible choices to generate a strictly bifurcating tree. However, this seldom occurs in real data sets.

FAMD produces trees in the PHYLIP format. You can use other software, such as TreeView to convert between formats and to view/print trees.



## Neighbour Joining

This generates a Neighbour Joining (NJ) tree from a distance matrix [default distance= 1-similarity (Jaccard)] you have generated previously and stores the tree in the tree output file you have defined (or else the default tree output file). The distance transformation applied can be changed under `Options → Distance Transformation`.

The NJ algorithm implemented in FAMD follows the original one (Saitou & Nei, 1987).

FAMD produces trees in the PHYLIP format. You can use other software, such as TreeView to convert between formats and to view/print trees.

## Strict Consensus

This function calculates a strict consensus tree from the trees present in the consensus input trees file and saves the consensus to the consensus output trees file. The input file is expected to contain trees in PHYLIP format. The output file likewise will contain PHYLIP format trees. You can use other software, such as TreeView to convert between formats and to view/print trees.

## Majority Rule Consensus

This function calculates a majority rule consensus tree from the trees present in the consensus input trees file and saves the consensus to the consensus output trees file. You can set the majority rule consensus threshold (default 50%) under `Options → Consensus Trees`. If a consensus threshold <50% is selected, a 50% threshold will be used. The input file is expected to contain trees in PHYLIP format. The output file likewise will contain PHYLIP format trees. Output trees will contain the percentage of occurrence of nodes saved as node labels (not as branch lengths). You can use other software, such as TreeView to convert between formats and to view/print trees.

## Principal Coordinate Analysis

This function carries out a principal coordinate analysis (PCoA or PCO; also referred to as metric multidimensional scaling or MDS) (Gower, 1966) on the current (dis)similarity matrix. The distance transformation used to generate the distance matrix can be changed under `Options → Distance Transformation`.

As of version 1.3, FAMD can calculate centroids for each defined Group after PCoA to display in the plot. You can enable this option via `Options → PCoA`.

PCoA is calculated from this distance ( $d_{ij}$ ) matrix as follows:

- a new matrix **A** is calculated whose elements ( $a_{ij}$ ) are given by  $a_{ij} = -0.5 d_{ij}^2$
- from this, the centred matrix  $\Delta$  is derived whose elements  $\delta_{ij}$  are calculated as  $\delta_{ij} = a_{ij} - \bar{a}_i - \bar{a}_j - \bar{a}$ , where the last 3 terms are the row, column and total means of all  $a_{ij}$  values in the matrix **A**, respectively.
- eigenvalues  $\lambda_i$  and normalised eigenvectors  $\mathbf{V}_i$  of the centred matrix  $\Delta$  are computed using the iterative method of Jacobi implemented in TPMath by Jean Debord. (FAMD calls the Jacobi routine from J. Debord's TPMath library to do this.)
- normalised eigenvectors are then scaled so that  $|\vec{V}_k| = \sqrt{\lambda_k}$ . This is done by multiplying all vector components with the square root of their associated eigenvalue.
- these scaled eigenvector components are the final PCoA coordinates.

Please note that there is no correction for negative eigenvalues, they (and their associated eigenvectors) are simply ignored.

The first (largest) eigenvalue is output to the analysis file, as is the number and sum of all positive eigenvalues and the PCoA coordinates (scaled eigenvector components). The eigenvalue percentages given in the output file are calculated relative to the sum of all positive eigenvalues found.

Apparently, the output for PCoA analyses are not exactly identical between different pieces of software (i.e., there seem to be some alternative implementations around). For instance, PCoA implemented in the R-package 4.0d9 (Mac) yields different coordinates than SynTax 2000 (Windows). During test runs, FAMD consistently gave results similar to the R-package, but different from SynTax.

## Replicate Analyses Menu

The functions of the `Replicate Analyses` menu work with replicates of your current data set and modify these replicate data sets automatically. Before carrying out any replicate analysis, you should define the parameters for analysis using the `Options → Bootstrapping & Replicate` options page. The parameters used by the individual functions are indicated in brackets after their name.

### Bootstrap Shannon's Variance (SH + RMD)

This option estimates Shannon's index and its variance by data resampling. Unlike the function `Shannon's Index` that operates on your current data set and for which you should first remove monomorphic loci from the data set, this option generates resampled data sets from your current data set and automatically removes monomorphic loci from it before calculating Shannon's index. Since every resampled data set may contain a different configuration of loci from the original data matrix and since missing data may be replaced randomly, it may be that loci that are treated as monomorphic are not monomorphic in another data set replicate. Therefore, you **SHOULD NOT** remove monomorphic bands manually from your data set (or replace missing data manually) for carrying out this function, since thereby you will limit the variation generated during bootstrapping.

This function uses the parameters `SH` and `RMD` set in the `Options → Bootstrapping & Replicate` options dialogue box. It is implemented as follows:

Repeat the following `SH` times:

- ) Generate a resampled data set by randomly choosing `s` loci from your current data set (every locus can be picked in every random draw), where `s` is the number of loci present in your current data set.
- ) If `RMD` is set, replace missing data so that monomorphic loci are removed from the newly generated replicate data set (according to the definitions given under `Remove Loci...`)
- ) From this data matrix, Shannon's index is calculated as described in the respective section

The series of values for Shannon's index are averaged and the variance calculated.

Please be aware that for high `SH` values, this procedure may take a considerable time.

### Bootstrap Shannon's Variance @ NIndiv

This function works exactly like the "`Bootstrap Shannon's Variance (SH + RMD)`", except that the sample size in terms of number of individuals per group (`NIndiv`) is fixed to the number specified by the user (as carried out e.g. by Schlüter *et al.*, 2011). During each bootstrap replicate, a different set of `NIndiv` individuals is randomly chosen.

The rationale for this analysis is that, although Shannon's index does not directly depend on the number of samples per group, it does depend on the number on markers observed in that group. Since this number of markers may indirectly depend on the number of individuals sampled (usually, you get a roughly logarithmic curve of markers as a function of number individuals sampled), for Groups consisting of only few samples (say,  $< 10$ ), it may be better to explicitly account for sampling size differences – as is done in the function outlined here – so as to ensure the Shannon index values remain comparable among Groups. For larger individual sample sizes, this should not usually be an issue, because the marker/individual curve would already have plateaued.

### Bootstrap Std Tree (BS + RMD)

This function can be used to generate multiple UPGMA or NJ trees (as set under `Options→Trees`) based upon the standard selected similarity of distance coefficient from resampled data matrices and stores them in the tree output file. You can then use the `Trees → Strict Consensus` or `Trees → Majority Rule Consensus` functions, or alternatively use other software to analyse the trees.

This function uses the parameters `BS` and `RMD` set in the `Options → Bootstrapping & Replicate` options dialogue box. It is implemented as follows:

Repeat the following `BS` times:

- ) Generate a resampled data set by randomly choosing `s` loci from your current data set (every locus can be picked in every random draw), where `s` is the number of loci present in your current data set.
- ) If `RMD` is set replace missing data so that monomorphic loci are removed from the newly generated replicate data set (according to the definitions given under `Remove Loci...`)
- ) From this data matrix, calculate similarity matrix based upon the selected coefficient of similarity or distance (standard definition)
- ) Perform the selected distance transformation
- ) Perform the UPGMA or NJ clustering algorithm on the distance matrix
- ) Write a tree to the tree output file.

## Multiple Avg Trees (TR + JC)

This function can be used to generate multiple UPGMA or NJ trees (as set under `Options→Trees`) based upon the average selected similarity (or distance) coefficient and stores them in the tree output file. You can then use the `Trees → Strict Consensus` or `Trees → Majority Rule Consensus` functions, or alternatively use other software to analyse the trees.

This function uses the parameters `TR` and `JC` set in the `Options → Bootstrapping & Replicate` and `Options → (Dis) Similarity Coefficients` options pages, respectively. It is implemented as follows:

Repeat the following `TR` times:

- ) Randomly draw `JC` values from the interval  $[s_{j,min}; s_{j,max}]$  (minimum to maximum possible similarity value) and calculate average similarity values from this series of numbers. The same is done if the selected coefficient directly produces a distance measure
- ) Generate a (dis)similarity matrix from the average Jaccard values
- ) Perform the selected distance transformation
- ) Perform the UPGMA or NJ clustering algorithm on the distance matrix
- ) Write a tree to the tree output file

## Bootstrap Avg Trees (BS + JC)

This function can be used to generate multiple UPGMA or NJ trees (as set under `Options→Trees`) based upon the average selected similarity (or distance) coefficient from resampled data matrices and stores them in the tree output file. You can then use the `Trees → Strict Consensus` or `Trees → Majority Rule Consensus` functions, or alternatively use other software to analyse the trees.

This function combines the data resampling of loci used for bootstrapping values and the calculation of average Jaccard coefficients by sampling from the interval of possible values.

This function uses the parameters **BS** and **JC** set in the **Options → Bootstrapping & Replicate** and **Options → (Dis) Similarity Coefficients options** pages, respectively. It is implemented as follows:

Repeat the following **BS** times:

- ) Generate a resampled data set by randomly choosing *s* loci from your current data set (every locus can be picked in every random draw), where *s* is the number of loci present in your current data set.
- ) Randomly draw **JC** values from the interval [*s<sub>ij,min</sub>*; *s<sub>ij,max</sub>*] (minimum to maximum possible similarity value) and calculate average similarity values from this series of numbers. The same is done if the selected coefficient directly produces a distance measure.
- ) Generate a similarity matrix from the average [dis]similarity values
- ) Perform the selected distance transformation
- ) Perform the UPGMA or NJ clustering algorithm on the distance matrix
- ) Write a tree to the tree output file

## Bootstrap Population Tree

This function can be used to generate multiple UPGMA or NJ trees (as set under **Options→Trees**) based upon the currently selected population-population distance (set under **Options→Population distances**) from allele frequency estimates. Groups must be defined to use this functions. All groups, except **AllData**, will be used in this function.

FAMD will first ask you to specify the method for allele frequency estimation to be used for this function (default: Bayesian from a non-uniform among –population prior, for more details see **Null allele frequencies**). Trees generated by this function are stored the tree output file. You can then use the **Trees → Strict Consensus** or **Trees → Majority Rule Consensus** functions, or alternatively use other software to analyse the trees.

The number of bootstrap replicates to be used in this function is specified by the parameter **BS** set in the **Options → Bootstrapping & Replicate**.

The function is implemented as follows:

Repeat the following **BS** times:

- ) Generate a resampled data set by randomly choosing *s* loci from your current data set (every locus can be picked in every random draw), where *s* is the number of loci present in your current data set.
- ) Estimate allele frequencies for all Groups of data using the selected allele frequency estimation method.
- ) Calculate a pairwise group-group distance matrix from allele frequency estimates using the currently selected population distance.
- ) Perform the UPGMA or NJ clustering algorithm on the distance matrix
- ) Write a tree to the tree output file

## Estimate R-support (TR)

This function attempts to estimate the  $R_x$ -support for clusters and generates a  $R_x$ -consensus tree (written to the consensus tree output file) according to the options set under `Options→R-Support Settings`. The output tree will be in PHYLIP format with node labels (e.g., “R1.8”) indicating estimated  $R_x$  support for a given cluster. Clusters appearing only at  $r$  values higher than the specified threshold value (or the maximum  $r$  value analysed, whichever is the smaller value) will be unresolved and branches collapsed.

### NB:

**A)** Even in relatively small data sets, this function can be very resource-intensive. It requires a lot of computing (CPU time = user waiting time), and memory. In some instances, FAMD may ask Windows for more memory than Windows is willing to give it. In this case, there will be an “Out of memory” error. This is not a program error in FAMD, but simply says that the computer system’s resources are insufficient to carry out the analysis given the selected parameters. Performing the analysis on a newer computer system may resolve this problem, if it occurs.

**B)** FAMD is a single-threaded program and will NOT respond to user input until it has finished its calculations (or encounters an error).

This analysis is designed to estimate  $R_x$ -support for clusters found during UPGMA clustering. Average similarity (or distance) coefficients are calculated by averaging  $r$  values drawn from the interval  $[s_{ij,min}; s_{ij,max}]$ . Consensus trees are then constructed from TR trees calculated from average similarity matrices generated using different values of  $r$ . Given the consensus threshold of  $x\%$ , the value  $r_x$  for a cluster considered will be the smallest value of  $r$  in that a cluster appears in the consensus tree. Since these values  $r_x$  may be large numbers, it is convenient to define the replicate support number  $R_x$  at the given consensus threshold percentage of  $x\%$  as

$$\text{Replicate support:} \quad R_x = \log_{10} r_x$$

The smallest possible values are  $r_x = 1$  and  $R_x = 0$ . The smaller the  $R_x$  value obtained for a given cluster, the more highly supported this cluster is based upon data and missing data in the data matrix.

When  $r$  becomes very large, sampling from  $[s_{ij,min}; s_{ij,max}]$  is expected eventually lead to convergence of average similarity,  $s_{ij}^*$ , on arithmetic mean of minimum and maximum similarity values. Since replicate support is estimated essentially using a stochastic process, slight differences between runs may be expected. It is also possible that, very rarely, a tree is found which is apparently inconsistent with previous trees, i.e., does not contain a cluster the present  $r$  value which was found in a tree with a smaller  $r$  value. The current version of FAMD will report such trees - should they be found - but essentially ignores them.

Note that the replicate number,  $r$ , is the same as the variable JC defined for use in other functions in e.g. the `Replicate Analyses` menu with the only difference that JC remains constant during these analyses but different values of  $r$  are considered during the `Estimate R-support` routine.

## Shannon t-tests

This option lets you carry out t-tests comparing different Shannon values. In order to use this function, groups must be defined in your data matrix (see `Group Manager` section). FAMD will display a dialogue box asking you to check/uncheck the groups whose Shannon indices you wish to compare. The dialogue box also asks you what data you wish to see written to your analysis file. You can make the following choices:

- ) Should Shannon's measure and variance will be calculated using the Bowman *et al.* (1969) formula?

- ) Should Shannon's measure and variance will be estimated by bootstrapping?
- ) Should FAMD output the p-value or simply tell you significant vs. insignificant at a given p-value?
- ) Should FAMD output the p-value or simply tell you significant vs. insignificant at a given p-value?
- ) Should FAMD output t(df) values?
- ) Should FAMD output the actual Shannon values and variances and group sizes?

Shannon's index will be calculated according as specified under `Options` → `Shannon Scaling`, and, if the bootstrapping method is selected, the additional settings under `Options` → `Bootstrapping & Replicates`.

NB: If you use the Shannon/Bowman variance option and band frequencies  $p(i)$  are defined on a per-locus basis, the variances may become negative which precludes t-testing. FAMD will not stop you to use such settings, but will display error messages if variances do get negative.

This routine operates on transient copies of the input data matrix and takes care of the removal of monomorphic bands in these replicate data sets.

The implementation for Shannon/Bowman variance is as follows:

For every selected group, do the following:

- ) If `RMD` is set replace missing data so that monomorphic loci are removed from the newly generated replicate data set (according to the definitions given under `Remove Loci...`)
- ) From this data matrix, calculate Shannon's index is calculated using the scaling options selected and the Bowman *et al.* (1969) variance

The implementation for Shannon/bootstrapped variance is as follows:

For every selected group, repeat the following `SH` times:

- ) Generate a resampled data set by randomly choosing  $s$  loci from your current data set (every locus can be picked in every random draw), where  $s$  is the number of loci present in your current data set.
- ) If `RMD` is set replace missing data so that monomorphic loci are removed from the newly generated replicate data set (according to the definitions given under `Remove Loci...`)
- ) From this data matrix, Shannon's index is calculated using the scaling options selected

The series of values for Shannon's index are averaged and the variance calculated.

T-Tests are carried out as follows:

t-values are calculated as

$$t = \frac{I_1 - I_2}{\sqrt{\text{var}(I_1) + \text{var}(I_2)}}$$

where...

$I_X$  is Shannon's index for group X.  
 $\text{var}(I_X)$  is the variance of  $I_X$ .

the degrees of freedom are calculated as

$$df = \frac{(\text{var } I_1 + \text{var } I_2)^2}{\frac{\text{var}^2(I_1)}{n_1} + \frac{\text{var}^2(I_2)}{n_2}}$$

where...

$I_X$  is Shannon's index for group X.

$\text{var}(I_X)$  is the variance of  $I_X$ .

$n_X$  is the sample size (number of individuals) of group X.

p-values are calculated as implemented in TPMath by Jean Debord. [FAMD calls the PStudent (df, tvalue) function.]

## **View Menu**

This menu provides a few shortcuts to easily opening/viewing some of FAMD's files in the default programs that are associated with the relevant file extensions, or in the case of the PCoA viewer, a built-in viewer. The default programs are those that would start if you double-clicked on the file in questions in the Windows Explorer.

## **Input File**

This will open FAMD's input file for viewing in the standard program associated with the file's suffix, i.e., typically a text editor such as Window's Notepad.

## **Analysis File**

This will open FAMD's analysis file (default: `analysis.txt`) for viewing in the standard program associated with the file's suffix, i.e., typically a text editor such as Window's Notepad.

## **Tree File**

This will open FAMD's tree output file (default: `outtree.ph`) for viewing in the standard program associated with the file's suffix. For instance, you might use TreeView or MEGA.

## **Consensus Tree File**

This will open FAMD's tree output file (default: `contree.ph`) for viewing in the standard program associated with the file's suffix. For instance, you might use TreeView or MEGA.

## **PCoA 3D Viewer**

This will open FAMD's built-in 3D PCoA coordinate viewer (see screenshot below), which is also started automatically after a PCoA has been performed (unless this feature is turned off by the user). The Viewer can be used to rotate data points in 3D space, and associate symbol and colours with data points.

The "Export coordinates" button can be used to save a set of PCoA coordinates (symbols associated with data points) to a text file, which can be loaded back into the Viewer using the "Import coordinates" button.

The plot, as displayed on screen, can also be copied to the clipboard for copying to another application ("Copy to clipboard" button) or saved to file ("Save to file" button) in the selected graphics



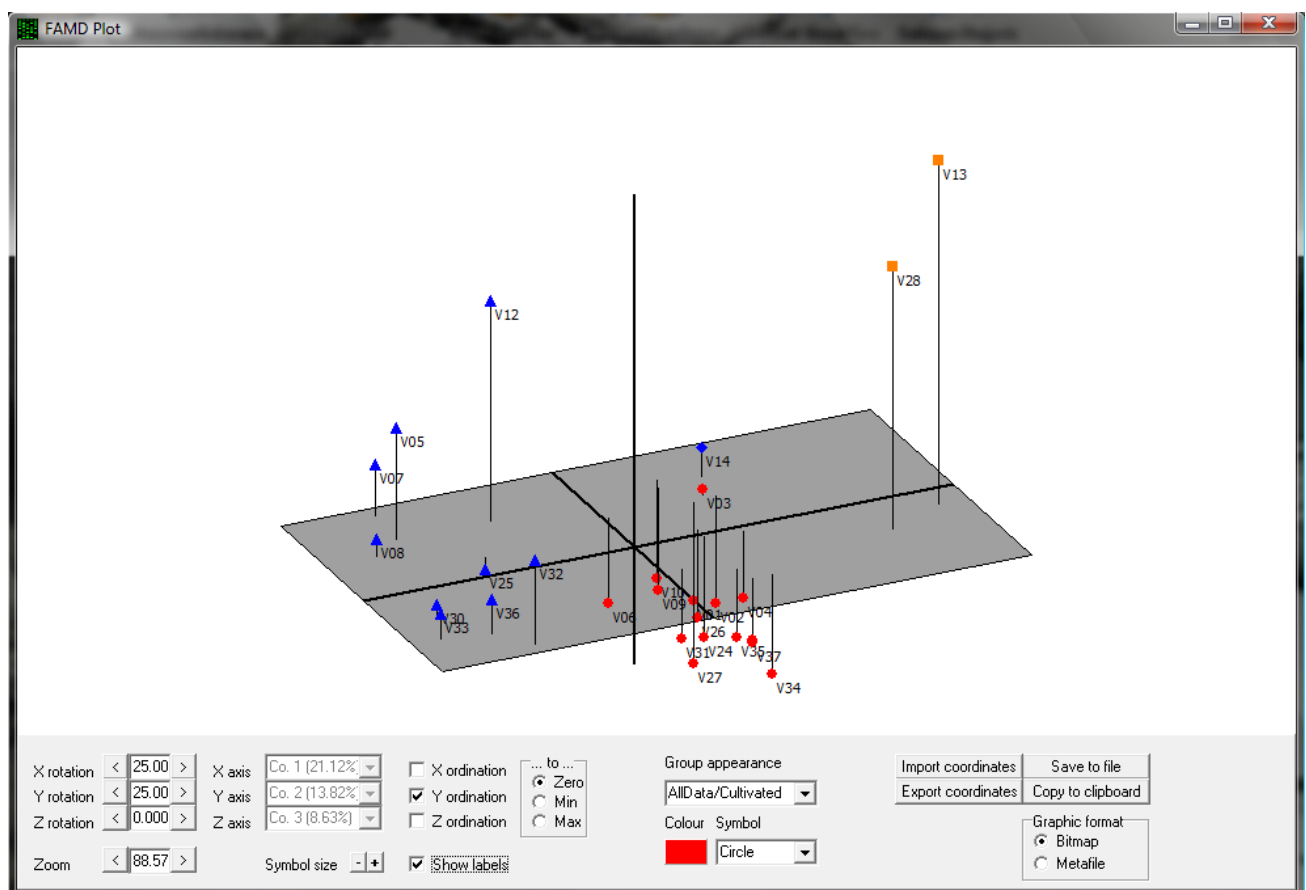
format. This can either be a bitmap (\*.bmp) file or a Windows Enhanced Metafile (\*.emf), as controlled by the respective radio buttons. Since a metafile stores graphics in a vector-oriented format, graphics stored as metafiles can later be scaled to a different size without loss of resolution, or edited in a graphics or office program.

NB: Note however, that some versions of some programs will unfortunately use bitmap data associated with a metafile even if it contains vector-oriented graphics. When using a metafile, the background colour for the entire plot may not get saved.

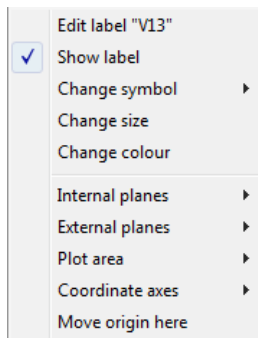
The plot can be rotated and zoomed by clicking on the appropriate buttons, by mouse control (left mouse button with or without Shift key) or by setting in rotation angles and zoom manually.

Ordinations for X, Y and Z coordinates can be displayed to either the zero, minimum, or maximum height of the coordinate system. Display of labels (OTU/individual names) can be turned on or off with the "Show labels" check-box, and the size of symbols increased or decreased using the "+" and "-" buttons beside the label "Symbol size".

If Groups are defined and you performed a PCoA, you can select a Group from the drop-down box and then set the symbol (drop-down box) and colour associated with members of this group (clicking on the "colour" panel).



Right-clicking on a data point brings up the following menu, in which the top functions can be used to change parameters for individual data points, and the bottom functions control the overall appearance of the plot:



Using “Edit label...” you can edit and change the label associated with a data point, and decide whether or not it should be shown manually using “Show label”. Furthermore, symbol shape, size and colour can be changed individually for each data point.

Clicking on “Internal planes” and “External planes”, you can determine which planes you want to be shown and set the colours for internal (zero-level) and external planes (boundaries to the coordinate system displayed). The background and foreground colours for the entire plot can be changed under “Plot area” and presence and width of positive and negative coordinate axes under “Coordinate axes”. “Move origin here” positions the origin ( $x=0$ ,  $y=0$ ,  $z=0$ ) of the coordinate system wherever you last right-clicked to bring up the menu displayed.

## Log File

This will open FAMD’s log file (default: `famdlog.txt`), if it exists, for viewing in the standard program associated with the file’s suffix, i.e., typically a text editor such as Window’s Notepad.

## Plot with R and View

This start FAMD’s R plotting script (default: `plotfamd.R`) in R’s script interpreter (`RScript.exe`), wait for the script to finish and then display the generated plot output file (default: `famdplot.pdf`), if it exists, for viewing in the standard program associated with the file’s suffix, (e.g. a pdf reader).

*Notes:*

- ) This can of course only work if you have R installed on your machine.
- ) You are free to provide your personal plotting scripts, and to enhance the script currently provided with FAMD. It is not very sophisticated yet, so I encourage you to improve on it.

## Last R Plot

This will open the most recently produced plotting result file (default: `famdplot.pdf`), if it exists, for viewing in the standard program associated with the file’s suffix, (e.g. a pdf reader).

## Options Menu

Functions in this menu allow you to set parameters that are use by different routines in the program. Changing any of these parameters does not actually result in any calculations or modifications done to your data matrix.

## Missing data replacement

This options page lets you select how missing data should be treated by those routines that deal with missing data. No action is performed on the data matrix. The options are to replace all missing data

points by presences, by absences or to randomly replace missing data by  $x$  % of presences and  $(100-x)$  % of absences, where  $x$  is the value entered by you. The default is missing data replacement by 50 % presences.

## Shannon scaling

Here you can define which frequency definition and which logarithm base should be used for calculation of Shannon's index.

The frequency of band presences  $p(i)=p_i$  can be defined:

- ) relative to all presences (frequency per data set):

$$p(i) = p_i = \frac{N_1(i)}{\sum_{k=1}^s N_1(k)}$$

where  $p(i)$  is the band frequency in locus  $i$   
 $N_1(i)$  is the number of band presences in locus  $i$   
 $s$  is the number of loci in the data set  
 $\sum N_1$  is the total number of band presences in the data set.

- ) divided by the number of individuals (frequency per locus):

$$p(i) = p_i = \frac{N_1(i)}{n}$$

where  $p(i)$  is the band frequency in locus  $i$   
 $N_1(i)$  is the number of band presences in locus  $i$   
 $n$  is the number of individuals in the data set

Internally, the program will always calculate Shannon's index using  $\log_2 x = \text{ld } x$  (the dual/binary logarithm), because (i) FAMD/Shannon's index deals with binary data and (ii) it is the native logarithm for the computer's processor (FPU):

$$I_2 = I \approx -\sum_{i=1}^s p_i \log_2 p_i$$

where  $I_2$  represents the fact that logarithm with base 2 was used

However, since in principle, any logarithm can be used, the program can re-scale Shannon's index accordingly by multiplying with a correction factor:

$$I_A = I_2 \log_A 2$$

where  $I_2$  represents Shannon's index based on  $\log_2 x$   
 $I_A$  represents Shannon's index based on  $\log_A x$

You can select the following options:

- ) use  $\log_2 x = \text{ld } x$
- ) use  $\log_e x = \ln x$
- ) use  $\log_{10} x = \lg x$

- ) use  $\log_A x$ , where A is user-defined

## (Dis)Similarity Coefficients

Using this options page, you can select the similarity or dissimilarity coefficient to be used by FAMd. Please note that, when using a distance measure, a minimum distance would correspond to a maximum similarity.

Jaccard, Dice and SMC coefficients produce similarities, NeiLi, Euclidean, and squared Euclidean produce distances. Distance measures are marked with an asterisk (\*) in the dialogue box. For details of the different similarity coefficients, please see the `Standard`, `Minimum`, `Maximum`, and `Average Similarity` sections in this manual.

For Nei & Li's distance, the parameter  $r$  can also be set here.

For average similarity/dissimilarity coefficients, the parameter `JC` (Average similarity from how many random draws) can also be set here. This parameter determines the number of random draws ( $r$ ) from the interval  $[s_{j,\min}; s_{j,\max}]$  that is used for calculating an average (dis)similarity value and its variance.

Ticking the appropriate check-boxes tells FAMd to write a similarity matrix (where available) or distance matrix to the analysis output file when the calculation of these is requested by the user.

## Distance Transformation

This options page allows you to select which distance transformation is applied to an input (dis)similarity matrix before use by a downstream method such as dendrogram construction.

The distance transformations available are the following.

- )  $d_{ij} = 1 - s_{ij}$  for similarities or  $d'_{ij} = d_{ij}$  for distances (i.e., distance measures are not modified further)
- )  $d_{ij} = \sqrt[2]{1 - s_{ij}}$  for similarities or  $d'_{ij} = \sqrt[2]{d_{ij}}$  for distances
- )  $d_{ij} = (1 - s_{ij})^2$  for similarities or  $d'_{ij} = d_{ij}^2$  for distances

## Character Weights

The Weights tab is currently only a placeholder, as no character weighting schemes are implemented in FAMd 1.2. Such features are, however, planned for a future release.

## Bootstrapping & Replicates

Here, you can define parameters for different analysis functions that operate on multiple data set replicates. It is advisable if you do this before you start your analyses. The available parameters are:

- SH: Resampled data matrix replicates for Shannon  
This number defines how many replicates of your current data set should be generated for estimating Shannon's index by data resampling (i.e. the number of bootstrap replicates).
- Write all Shannon replicates to file:  
If this option is enabled, all individual bootstrap replicates will be written to the analysis output file when Shannon's variance is estimated by bootstrapping.
- Include Shannon I=0 values in stats:  
If your input data matrix is VERY small, it is possible for data resampling to generate a bootstrap replicate in which all loci are uninformative for calculation of Shannon's index  $I$  and the resulting value is  $I=0$ . Ticking this option will exclude such values for the generation of mean and variance of Shannon's index by bootstrapping; such unsuccessful replicates will count towards the desired number of bootstrap replicates requested by the user.
- RMD: Replace missing data (as set in respective dialogue)  
If checked, missing data in the data matrix will be replaced according to the parameters defined in the Options→Missing Data Replacement options page.
- BS: Resampled data matrix replicates for similarity analyses  
Defines from how many data set replicates UPGMA or NJ trees should be generated.
- TR: Number of average similarity trees to generate  
Defines how many UPGMA or NJ trees based upon average (dis)similarity-derived distance matrices should be generated.

## Trees

This options page lets you select the default tree type (either UPGMA or NJ) to be used for bootstrap analyses etc.

## Consensus Trees

This options page lets you set the threshold the threshold percentage used for the majority rule consensus function (must be 50% or higher). Clusters occurring at a frequency less than the threshold frequency will be unresolved in the resulting consensus tree. (Changing this parameter does not affect  $R_x$ -analysis.)

## R-Support

Here you can define parameters that are used by those routines that require consensus tree methods, especially the Replicate Analyses → Estimate R-support function.

You can define

- the threshold percentage,  $x$ , for the majority rule consensus routine used by FAMD used internally during  $R_x$ -support analysis. 100% (the default) means that a strict consensus is used (and  $R_{100}$  calculated).

- ) the  $R_x$  threshold r-value (corresponding to variable `JC` for average similarity calculations) for the majority rule consensus routine used by FAMD used internally during R-support analysis. Clusters will be unresolved on the  $R_x$ -consensus tree if they occur only in consensus trees generated at r-values greater than or equal to the specified values.
- ) the range of values that should be considered for R-support analysis, i.e. the minimum and maximum r-values to be considered.
- ) the desired precision to be obtained for  $R_x$  values. A precision of 1 digit here would specify an output of values with one digit after the decimal point, e.g.  $R_x=1.3$ , where the analysis should be sufficiently precise that run-to-run variation between  $R_x$ -values should be limited to the last digit displayed. Note that increasing the precision of the analysis may drastically increase computing time and memory used by FAMD.

## AMOVA

This options page lets you select the default mode of (dis)similarity index calculation used by FAMD's AMOVA and Pairwise `PhiST` routines. The choices are: Standard, Minimum, Maximum, and Average Similarity (where a minimum similarity will correspond to a maximum distance).

As of version 1.3, FAMD allows you to estimate significance in AMOVA by a randomisation procedure as also performed in Arlequin. **Caution:** this functionality is currently in beta-testing phase! To enable this functionality, select "Estimate AMOVA p-values by randomisations:" and provide a number after "Number of randomisations:" (The number 0 is equivalent to turning off this option.) The locus-by-locus AMOVA functionality is currently unavailable.

## PCoA

This options page lets you select parameters for FAMD's principal coordinate analysis routine:

If `Automatically start PCoA viewer` is turned on, then FAMD will open the 3D coordinate viewer after a PCoA has been performed.

`Max # iterations`: This defines the maximum number of iterations that FAMD will carry out for PCoA analysis. If the calculations have not converged on a result by after the indicated number of iterations, PCoA will be deemed to have failed.

`Precision for PCoA`: This value defines the error tolerance for convergence of the PCoA calculations (used in the Jacobi routine).

`Calculate centroids`: Ticking this box will enable the calculation of centroids for each defined Group after PCoA.

## Allele Frequencies

This options page lets you select the allele estimation method to be used e.g. for calculation of inter-population distances. The options are: square root, Lynch-Milligan, and the Bayesian method with, uniform prior, or non-uniform prior from among-population or among-locus information.

## Populations distances

Currently, only there is only one inter-population distance implemented, and therefore there's currently no choice: it's the Cavalli-Sforza & Edwards chord distance (Takezaki & Nei formula).

## I/O Options

This options page lets you set input/output and user-interface related parameters, as well as some miscellaneous things:

If the option `"Display program warnings"` is set, then FAMD will display Warning messages. Since these may in some cases be expected by and/or annoying to the user, you can disable this option to suppress such warnings. Note that actual error messages will still be displayed.

If the option `"Write FAMD File Specification Block"` is enabled, FAMD will include such a block in a file being saved by the user.

If option `"Trust File Specification Block (where available)"`, FAMD will assume that the information in an input file's FAMD File Specification Block (if it exists) is correct and attempt to load the file without displaying the data layout dialogue box.

FAMD's behaviour regarding existing output files can be changed: The option `"Always Ask (Show Dialogue Box)"` will ask you whether to overwrite or append to an output file that already exists. Alternatively, you can tell FAMD to Always Overwrite Files or Always Append To Files.

Also, the file names for analysis output, tree output, consensus tree input and output as well as the log file can be changed here.

Additionally, the names of files necessary for FAMD's R plotting functionality can be specified here: the R plotting output file (default: `famdplot.pdf`), the R script to plot the results (default: `plotfamd.R`) and R's script interpreter executable (this is usually `RScript.exe`). FAMD can automatically detect the path of the latter, but may for various reasons fail to do so, in which case you have to select this file manually. Also, if you have several versions of R installed, you might wish to manually select a different version of the executable.

The option `"Preferred FPU/SSE round mode"` lets you select the hardware rounding mode that some FAMD routines set. Note that this setting will not impact on all of FAMD's calculations but only to a subset of them. This option was included because I empirically found that to replicate the output generated by some other programs, it is sometimes necessary to change the processor's rounding mode. The options for this are: Round to nearest, Round up, Round down, Round truncate. I will not explain what they do exactly, but the interested reader may consult the *Intel® 64 and IA-32 Architectures Software Developer's Manual* available as PDF on Intel's web site for technical details.

This options page also contains some information on the machine on which FAMD is running, as determined by the program. This information includes Windows version and bitness, the processor, and number of cores, as well as the instruction sets available on the machine. The maximum number of parallel computations is currently restricted to 1, but you will likely be able to change this value in future versions of FAMD will likely be able to take advantage of multicore technology by running several parallel threads.

## Project

This options page is just a placeholder. There are currently no project options implemented.

## Population reallocation

This options page lets you select parameters for FAMD's population reallocation calculations. You can specify

- ) Allele frequency correction:
  - 1) Duchesne & Bernatchez (2002) correction formula, i.e.  $1/(1+N)$
  - 2) User specified epsilon correction value (default: 0.00001)
- ) Minimum log-likelihood difference for population allocation (default: 2.0)

## Help Menu

### About

Displays a message box with the authors's contact address, a copyright notice and the legal disclaimer.

### Citation

Displays a message box with the authors's contact address, a copyright notice and the legal disclaimer.

### Version

Displays the current version of the FAMD executable.

### Check for new version

Queries the FAMD web page ([www.famd.me.uk](http://www.famd.me.uk)) for the latest available FAMD version to determine whether a newer FAMD version than the one you are running is available. This obviously requires an internet connection. It also requires the presence of the Windows file `wininet.dll`, which should be present on the vast majority of Windows systems. (Since FAMD does not presuppose its existence, FAMD will run just fine on systems on which this file is missing - except for this very program function.)

### Help

Displays this help file (`famdhelp.pdf`) by starting the default program associated with pdf files, e.g. Acrobat Reader.



## REFERENCES

- Bowman, K. O., Hutcheson, K., Odum, E. P. & Shenton, L. R. 1969. Comments on the distribution of indices of diversity. – *Proc. Intl. Symp. Stat. Ecol.* **3**: 315-359.
- Buerkle, C. A. 2005. Maximum-likelihood estimation of a hybrid index based on molecular markers. – *Mol. Ecol. Notes* **5**: 684-687.
- Cavalli-Sforza, L. L. & Edwards, A. W. F. 1967. Phylogenetic analysis: models and estimation procedures. – *Evolution* **21**: 550-570.
- Duchesne, P. & Bernatchez, L. 2002. AFLPOP: a computer program for simulated and real population allocation, based on AFLP data. – *Mol. Ecol. Notes* **2**: 380-383.
- Excoffier, L., Smouse, P. E. & Quattro, J. M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. – *Genetics* **131**: 479-491.
- Gower, J. C. 1966. Some distance properties of latent root and vector methods in multivariate analysis. – *Biometrika* **53**: 325-338.
- Lynch, M. & Milligan, B. G. 1994. Analysis of population genetic structure with RAPD markers. – *Mol. Ecol.* **3**: 91-99.
- Nei, M. & Li, W.-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. – *Proc. Natl. Acad. Sci. USA* **76**: 5269-5273.
- Saitou, N. & Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. – *Mol. Biol. Evol.* **4**: 406-425.
- Schlüter, P. M. & Harris, S. A. 2006. Analysis of multilocus fingerprinting data sets containing missing data. – *Mol. Ecol. Notes* **6**: 569-572.
- Schlüter, P. M., Ruas, P. M., Kohl, G., Ruas, C. F., Stuessy, T. F. & Paulus, H. F. 2011. Evidence for progenitor-derivative speciation in sexually deceptive orchids. – *Ann. Bot.* **108**: 895-906.
- Takezaki, N. & Nei, M. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. – *Genetics* **144**: 389-399.
- Zhivotovsky, L. A. 1999. Estimating population structure in diploids with multilocus dominant DNA markers. – *Mol. Ecol.* **8**: 903-913.