GENEPOP 4.1
for Windows/Linux/Mac OS X
This documentation: March 24, 2011

F. Rousset

This is a documentation for the GENEPOP software. This program is a rewrite of the GENEPOP software originally described in Raymond & Rousset (1995b), version 3.4. Additional methods have been implemented has detailed below. A formal reference for the current version of GENEPOP is Rousset (2008). GENEPOP implements a mixture of traditional methods and some more focused developments:

It computes exact tests for Hardy-Weinberg equilibrium, for population differentiation and for genotypic disequilibrium among pairs of loci;

It computes estimates of F-statistics, null allele frequencies, allele size-based statistics for microsatellites, etc., and of number of immigrants by Barton & Slatkin's 1986 private allele method;

It performs analyses of isolation by distance from pairwise comparisons of individuals or population samples, including confidence intervals for "neighborhood size".

Future developments may include a more systematic implementation of bootstrap confidence intervals. Likelihood methods based on coalescent algorithms are being developed in a companion software, MIGRAINE (Rousset & Leblois, 2007).

Recent versions since 4.0 have routinely been checked under both Windows and Linux, and should be easy to compile on Unix-alike operating systems, including Mac OS X.

GENEPOP converts data from the GENEPOP input format to formats of some softwares that were around in GENEPOP's youth; there was little need to update this option as many more recent softwares for population genetic analyses read input files in the GENEPOP format.

What is new to Genepop 4.+?

The following analyses have been added, compared to GENEPOP version 3.4.

Version 4.1

It is possible to test trends in gene diversity among samples.

Analyses of isolation by distance have been strengthened in several ways. Variants of previously described estimators have been implemented for both haploid and diploid data. 0ne can select subsets of the data for analyses of isolation by distance within and between these subsets. Further, analysis of isolation by distance from several one-locus genetic distance matrices is now possible through the `MultiMigFile` option. In contrast to `IsolationFile`, this allows the construction of bootstrap confidence intervals. Finally, it is possible to test specific values of the slope of the spatial regression, using the `testPoint` setting.

The input file reading procedure is better protected against nonstandard file formats (in particular those produced by some Microsoft software under Mac OS X).

The new sub-option 8.4 has been added to convert population-based data to individual-based data (each individual in its own `Pop`).

Version 4.0

Use of the G (log likelihood ratio) statistic has been generalized to all contingency tables (though previous probability tests implemented in GENEPOP are still available). GENEPOP now provides bootstrap confidence intervals for strength of isolation by distance between groups of individuals, an alternative estimator for analyses of "differentiation between individuals", and facilities to evaluate the performance of these methods. The genetic distance matrix produced by these options can also be exported in Phylip (Felsenstein, 2005) format. The option for null allele estimation implements additional estimators with confidence intervals, and its output is better organized.

Some **additional facilities** have been implemented for better ease of use. Earlier versions of GENEPOP required from the user some effort to deal with either 3-digits-coded alleles or with haploid data. GENEPOP is more practical, in that haploid and diploid genotypes in both 2- or 3-digits allele codings are automatically recognized as such by the program and all these different types of data can be mixed in the same input file. The input format is otherwise unchanged so that **input files prepared for earlier versions of Genepop are still read by Genepop** (backward compatibility).

In addition, GENEPOP's behaviour can be controlled using an option file and by inline arguments in a console command line. This allows batch calls to GENEPOP and repetitive use of GENEPOP on simulated data. However, those familiar with the old GENEPOP menus can also use GENEPOP in an almost unchanged way.

Previous GENEPOP distributions included two small utilities, HW.BAT and STRUC.BAT, for testing of single data matrices using a fast ad hoc data input. These facilities are available in GENEPOP 4.0 through the `HWfile` and `StrucFile` options. Previous GENEPOP distributions also included the ISOLDE program for

analysis of isolation by distance between groups of individuals, from one genetic distance and one geographic distance matrices. All such analyses can now be performed through the unique GENEPOP executable (other facilities that were unique to ISOLDE are now accessible through the `IsolationFile` setting).

Other minor, and often trivial, differences with earlier versions of GENEPOP will be pointed out in footnotes.

The remainder of this documentation is as follows:

1 Installing Genepop and session examples **5**
- 1.1 Installation . . . 5
- 1.2 Example sessions . . . 5
  - 1.2.1 Example 1: basic session . . . 5
  - 1.2.2 Example 2: using the settings file . . . 6
  - 1.2.3 Example 3: Batch processing . . . 6

2 The input file **7**

3 The settings file and command line arguments **10**

4 All menu options **13**
- 4.1 Option 1: Hardy-Weinberg (HW) exact tests . . . 13
  - 4.1.1 Sub-options 1–3: Tests for each locus in each population . . . 14
  - 4.1.2 Sub-options 4,5: Global tests across loci or across samples . . . 15
  - 4.1.3 Analyzing a single genotypic matrix . . . 16
- 4.2 Option 2: Tests and tables for linkage disequilibrium . . . 17
  - 4.2.1 Sub-option 1: Tests . . . 18
  - 4.2.2 Sub-option 2: create tables . . . 19
- 4.3 Option 3: population differentiation . . . 19
  - 4.3.1 Sub-options 1 or 2 (genic differentiation) . . . 19
  - 4.3.2 Sub-options 3 or 4 (genotypic differentiation) . . . 20
  - 4.3.3 Output . . . 21
  - 4.3.4 Gene diversity as a test statistic . . . 21
  - 4.3.5 Analyzing a single contingency table . . . 22
- 4.4 Option 4: private alleles . . . 23
- 4.5 Option 5: Basic information, F_{IS}, and gene diversities . . . 23
  - 4.5.1 Sub-option 1: Allele and genotype frequencies . . . 23
  - 4.5.2 Sub-option 2: Identity-based gene diversities and F_{IS} . . . 24
  - 4.5.3 Sub-option 3: Allele size-based gene diversities and ρ_{IS} . . . 24
- 4.6 Option 6: Fst and other correlations, isolation by distance . . . 24
  - 4.6.1 Sub-options 1–4: F-statistics and ρ-statistics . . . 25
  - 4.6.2 Sub-option 5: isolation by distance between individuals . . . 26

4.6.3 Sub-option 6: isolation by distance between groups . . . 30
4.6.4 Former sub-option 5 of GENEPOP: analysis of isolation by distance from a genetic distance matrices . . . 30
4.6.5 Analysis of isolation by distance from multiple genetic distance matrices . . . 32
4.7 Data selection for analyses of isolation by distance . . . 33
4.7.1 Selecting a subset of samples . . . 33
4.8 Option 7: File conversions . . . 33
4.9 Option 8: Null alleles and some input file utilities . . . 34
4.9.1 Sub-option 1: null alleles . . . 35
4.9.2 Sub-option 2: Diploidisation of haploid data . . . 36
4.9.3 Sub-option 3: Relabeling alleles names . . . 36
4.9.4 Sub-option 4: Conversion of population data to individual data . . . 36

5 Evaluating the performance of inferences for Isolation by distance 37

6 Methods 37
6.1 Null alleles . . . 38
6.2 Exact tests . . . 38
6.3 Algorithms for exact tests . . . 39
6.4 Accuracy of P values estimated by the Markov chain algorithms . . . 39
6.5 Test statistics . . . 40
6.6 Estimating F-statistics and related quantities . . . 40
6.6.1 ANOVA estimators: single- and multilocus definitions . . . 41
6.6.2 Microsatellite allele sizes, R_{ST}, and ρ_{ST} . . . 43
6.6.3 Robertson and Hill's estimator of F_{IS} . . . 43
6.7 Bootstraps . . . 43

7 Code history, compilation, credits, contact, etc. 44
7.1 Contact . . . 44

8 Copyright 45

Bibliography 45

Index 49

1 Installing Genepop and session examples

1.1 Installation

Under **Microsoft Windows**, one only needs to unzip/copy the executable on hard disk. Under **Linux**, copy all sources and compile by the simple line
`g++ -DNO_MODULES -o Genepop GenepopS.cpp -O3`.
(note `O` in `-O3` is the letter O, not zero). This should work under **Mac OS X** too.

The data files do not need to be in the same directory as the executable[1]; however, users might find that specifying path names under Windows is not as easy at it should.

You may wish to install the examples, documentation, and source code. All files are available on the Genepop distribution page.

Linkdos, a program described by Garnier-Géré & Dillmann (1992), is distributed with Genepop (but is not part of Genepop). It is originally a DOS program, but the source file distributed with Genepop can be recompiled under Linux using the Free Pascal compiler.

1.2 Example sessions

To reproduce the examples of this session one should download the example package from the Genepop distribution page.

1.2.1 Example 1: basic session

Open a console window in the directory where Genepop has been installed and just execute

```
Genepop
```

If Genepop has never been run before, it will ask for an input file. Otherwise, the main menu should appear, in which case you should use the `C` option to load this input file. For this sample session, the file name to be given is `sample.txt`. Genepop will display some information about the file read, then display the main menu:

```
-------> Change Data ................... C


Testing :
    Hardy-Weinberg exact tests (several options) ..................... 1
    Exact tests for genotypic disequilibrium (several options) ........ 2
    Exact tests for population differentiation (several options) ...... 3

Estimating:
```

[1] ...in contrast to earlier versions of Genepop.

```
    Nm estimates (private allele method) ............................. 4
    Allele frequencies, various Fis and gene diversities ............. 5
    Fst & other correlations, isolation by distance (several options).. 6

Ecumenicism and various utilities:
    Ecumenicism: file conversion (several options) ................... 7
    Null alleles and miscellaneous input file utilities .............. 8

QUIT Genepop ..................................................... 9

Your choice? :
```

Each option will be described later. Let us see some tests for heterozygote deficiency. Reply `1`, next `1`, next `y`(es). As indicated, the results of the analysis are stored in the file `sample.txt.D`.

That was simple enough, even simpler than a first contact with previous versions of GENEPOP. Now exit Genepop (press `(Return)`, next `9`) and discover a new facet of GENEPOP.

1.2.2 Example 2: using the settings file

Execute

```
Genepop settingsFile=SampleSettings.txt
```

Do not add spaces in the arguments. Capitalisation matters for file names (here `SampleSettings.txt`) if it matters for the operating system (i.e. for Linux).

You can see that the previous and additional analyses are performed, and that you just need to hit Return each time GENEPOP stops and waits for feedback. Finally, you are brought back to the main menu. Simple instructions for performing the analyses are contained in the `SampleSettings.txt` file, which you may edit. Section 3 will explain how to use this file. By default, GENEPOP seeks and eventually reads instructions in a `Genepop.txt` file. You can see that one such file is present and was thus read when performing Example 1.

1.2.3 Example 3: Batch processing

Execute the same command as in the previous example but with one more statement:

```
Genepop settingsFile=SampleSettings.txt Mode=Batch
```

GENEPOP should perform the same computations as in the previous example but it will not stop and wait for feedback, and will exit after completion of the computations. Note again that spaces are not allowed within each of the arguments `settingsFile=SampleSettings.txt` and `Mode=Batch`, nor more generally in arguments specified on the command line.

The batch mode makes it easy to analyze multiple files. However, note that concurrent GENEPOP processes should be run in distinct directories. Otherwise, the temporary files of each process might conflict with each other.

2 The input file

As illustrated by the following examples, the input format requested by GENEPOP is:

First line: anything Use this line to store information about your data.

Locus names They may be given one per line, or on the same line but separated by commas.

`Pop` sample indicator (Capitalization does not matter)[2]. Each sample from a different geographical original is declared by a line with a `pop` statement.

Information for first individual. An example is:

`ind#001 fem  ,0101 0202 0000 0410`

Here `ind#001 fem` is an identifier for your personal use. You can use any character (except a comma!). You may leave it blank (at least one space) if you wish. The last identifier of every sub-population is used by GENEPOP as the sample name in output files. The comma between the identifier and the list of genotypes is required. `0101` indicates that this individual is homozygous for the `01` allele at the first locus. The same individual is homozygous for the `02` allele at the second locus (`0202`). Data are missing at the third locus (`0000`). At the fourth locus, the genotype is `0410`, which indicates the presence of alleles `04` and `10`.

More individuals Each individual information starts on a new line, but may extend over several lines (do not start a new line in the middle of a one-locus genotype!).

More samples each declared by a `pop` statement on a new line

Blank lines at the end of the file are removed by GENEPOP.

An example of a short input file is given below:

```
Title line: "Grape populations in southern France"
ADH Locus 1
ADH #2
ADH three
ADH-4
ADH-5
mtDNA
Pop
```

[2]Earlier versions of GENEPOP only accepted `Pop`, `POP` and `pop`...

```
Grange des Peres  ,  0201 003003 0102 0302 1011 01
Grange des Peres  ,  0202 003001 0102 0303 1111 01
Grange des Peres  ,  0102 004001 0202 0102 1010 01
Grange des Peres  ,  0103 002002 0101 0202 1011 01
Grange des Peres  ,  0203 002004 0101 0102 1010 01
POP
Tertre Roteboeuf ,      0102 002002 0201 0405 0807 01
Tertre Roteboeuf ,      0102 002001 0201 0405 0307 01
Tertre Roteboeuf ,      0201 002003 0101 0505 0402 01
Tertre Roteboeuf ,      0201 003003 0301 0303 0603 01
Tertre Roteboeuf ,      0101 002001 0301 0505 0807 01
pop
Bonneau 01   , 0101    002002 0304 0805 0304 01
Bonneau 02   , 0201    002002 0404 0505 0304 01
Bonneau 03   , 0101    002100 0304 0505 0101 01
Bonneau 04 , 0101    100100 0204 0805 0304 01
Bonneau 05   , 0101    100002 0104 0808 0304 01
Pop
 ,            0000 002001 0202 0402 0007 01
 ,            0200 002001 0202 0205 0707 01
 ,            0010 002001 0101
0105 0807 01
last pop,      0101 002001 0101 0401 0807 02
```

This example shows some useful features of the input file:

- There is no constraint on the number of blanks separating the various fields.
- The individual identifier has a free format.
- Alleles are numbered from 01 to 99 or 001 to 999 if needed. In 3-digits coding, (say) homozygotes for the `90` allele are noted `090090`, not `9090` as in the 2-digits format. 2-digits and 3-digits coding of alleles can be intermixed (among loci, not within loci!).[3]
- To designate alleles, consecutive numbers are not required.
- haploid and diploid data can be intermixed.[4] 6-digits genotypes are recognized as 3-digits diploid genotypes; 4-digits genotypes are recognized as 2-digits diploid genotypes; 2- and 3-digits genotypes are recognized as haploid genotypes. The same coding should be used consistently within each locus. See the `EstimationPloidy` setting for more information about analyzing haploid data. For haplo-diploid data at a given locus, the haploid genotypes should be coded as diploid genotypes with

[3] New to GENEPOP 4.0
[4] Also new to GENEPOP 4.0

one unknown allele; note however that the information from haploid genotypes at haplo-diploid loci will be used only for genic contingency table tests, and will be ignored in estimation of genetic structure.

- Genotypes can extend on more than one line (see penultimate individual)
- To group various samples, just remove each relevant `Pop` separator.

It is possible to write all the locus names on one line, provided that a comma is used as separator. This could be useful to clearly label each column. Thus the above input file could have started as

```
Title line: "Grape populations in southern France"
                     Loc1,Loc2,  ADH3,ADH4,ADH5,mtDNA
Pop
Grange des Peres  ,  0201 003003 0102 0302 1011 01
...
```

Note the absence of comma after the last locus name.

There are however constraints to be obeyed

- Missing data should be indicated with `00` (or `000` for 3-digits coding) and not with blanks. The first locus in the last sample illustrates the various possibilities of missing data: no information (first individual coded `0000`) or partial information (only one allele is determined: allele `02` for the second individual coded `0200` and allele `10` for the third individual coded `0010`).
- The number of locus names should correspond to the number of genotypes in each individual. If you remove one or several loci from your input file, you should remove both their names and the corresponding genotypes.
- No empty line should be present in the data file.
- GENEPOP accepts input file names either with the extension `.txt`[5] or without any extension.
- GENEPOP input files are ASCII text files.

The last point implies that under **Windows**, you should avoid using Microsoft Word to edit input files (and settings files as well). Rather use the MFC Wordpad program found in recent distributions of Windows (The Windows more basic text editor may not show all end-of-line characters correctly, which may cause trouble).
It also appeared that certain Microsoft products under **Mac OS X** still produces files formatted according to the older Mac format. GENEPOP now catches and corrects this miserable feature.

One can also find some conversion tools (e.g. from EXCEL) on the web.

[5] New to GENEPOP 4.0

If the input file is correctly read, the name of the larger allele number is indicated for each locus. The number of distinct alleles for each locus is provided upon request. If alleles have been labeled with consecutive numbers from `01` onwards, then the name of the larger allele will correspond to the number of distinct alleles for each locus.

In principle, there are no built-in effective maxima on sample size in GENEPOP.[6] There are some limits to the number of samples and individuals imposed by the compiler. These unrealistically large values, and a few other ones, are shown by running "`Genepop Maxima=`" (see the `Maxima` setting).

3 The settings file and command line arguments

The settings file allows finer control of GENEPOP and/or batch processing. Further control is possible by using optional arguments when launching GENEPOP through the operating system command line, following the general syntax explained below for the settings file, e.g.

```
Genepop EstimationPloidy=Haploid DifferentiationTest=Proba
```

Indeed, command line arguments are written in the file `cmdline.txt`, then this file is read much as the settings file.[7]

Henceforth, menu options are called *options* and batch file/command line options are called *settings*.

Running `Genepop help` will display the help information, which so far is no more than a list of available settings, loosely grouped semantically. A file showing all possible settings is the following:

[6] in constrast to earlier versions of GENEPOP

[7] Long command lines: under some old versions of Windows, the command line had a fairly limited maximum length, so it should have been used with moderation. This should no longer be a problem with recent versions of Windows, but who knows with Microsoft... one may try to find more information about command-line string limitation on `support.microsoft.com`.

```
// sample Genepop settings file, showing all options.
/********** Syntax of this file:
lines without 'equal' symbol are ignored (hence this one is).
Lines beginning with a '/', /a '#' or a '%' are also ignored,
even if they contain '=' (hence this one is).
/********** General options **********
Mode=Ask
GenepopInputFile=sample.txt
Dememorisation=10000
BatchLength=5000
BatchNumber=100
//EstimationPloidy=Haploid
//RandomSeed=12345678
//MantelSeed=87654321
/***     allele sizes stuff
//AllelicDistance=Size
AlleleSizes=1:5,2:10,3:15,10:50
/*** selecting menu options
MenuOptions=8
/********** Option 1 (HW tests) **********
HWtests=Enumeration
/           Emulating HW.BAT
//HWFile=HWtest
//HWfileOptions=4,3
/********** Option 2 ("linkage" disequilibrium) **********
//          old Genepop behaviour
/GameticDiseqTest=Proba
/********** Option 3 (differentiation) **********
//          old Genepop behaviour
/DifferentiationTest=Proba
/           Emulating STRUC.BAT
//strucFile=structest
/********** Option 4 (private alleles) **********
//no specific setting, but may be affected
//by the estimationPloidy setting
/** Option 5 (basic information, Fis, gene diversities... )
//no specific setting, but may be affected
// by the AlleleSizes setting
/***** Option 6 (F-statistics, isolation by distance) *****
IsolationStatistic=e
GeographicScale=Linear
MinimalDistance=1
CIcoverage=0.9
testPoint=0.00123
/PopTypes= 1 2 1 2 3
/PopTypeSelection= all
//PhylipMatrix=
/           Emulating ISOLDE
//IsolationFile=Isoldetest
/           Extending ISOLDE to multiple matrices
//MultiMigFile=perlocusStuff
/********** Option 7 (file conversions) **********
//no specific setting
/********** Option 8 (Various utilities) **********
NullAlleleMethod=ApparentNulls
CIcoverage=0.9
/******** Testing performance of some options *********
// Option 6.x: options as above plus
//Performance=aLinear
//GenepopRootFile=file
//JobMin=1
//JobMax=100
/********* Checking some limits of Genepop **********
//Maxima=
```

Each setting is specified following a *Keyword=value* syntax. Capitalisation is not important (it is here only to ease reading) *except* for file names if the operating system cares about it (as Linux does).

By default, GENEPOP seeks settings in the file `Genepop.txt`, but one can specify another settings file through the command line, as was shown in the session examples:

```
Genepop settingsFile=SampleSettings.txt
```

The `SettingsFile` setting must be the first argument on the command line.

Settings specific to each menu option will be explained along with the description of each option. Settings affecting several menu options are the following:

`GenepopInputFile` (or simply `InputFile`)

which is the name of the input file in GENEPOP format

`Dememorisation`, `BatchLength` and `BatchNumber`

which are Markov Chain parameters, which meaning is explained in Section 6.3:

the dememorisation number The default is 10000;[8] values below 100 are not allowed.

the number of batches The default is 20 for sub-options 1.4 and 1.5 (multisample HW tests), and 100 otherwise; values below 10 are not allowed.

the number of iterations per batch The default is 5000;[9] values below 400 are not allowed.

The maximum allowed values will depend on the compiler, being the much-more-than-needed 2,147,483,647 for all three parameters for the distributed Windows executable (see the setting `Maxima` if you really need more information about this).

`EstimationPloidy`

In multilocus estimates only diploid data are taken into account, unless the setting `EstimationPloidy=Haploid` is given, in which case only haploid data are taken into account. This setting applies to options 4 (private allele method), 5.2 and 5.3 (for multilocus estimates of gene diversities), and 6 (F-statistics and isolation by distance).

`Mode`

GENEPOP has three modes: `Mode=Ask` will ask for some feedback even in cases where the answer has been prespecified (e.g. through some setting; this may be useful when one wishes to chance some settings in the course of a GENEPOP session). For example it will ask for confirmation of the MC parameters. `Mode=Batch` will not wait for feedback:

[8] increased from GENEPOP 3.4's default
[9] increased from GENEPOP 3.4's default

execution of GENEPOP should complete without any user intervention. The third mode, `Mode=Default` (which in most cases does not need to be explicitly specified) will ask for unspecified settings but not request confirmation of prespecified ones, and will also pause and wait for feedback when some notable information is displayed.

`MenuOptions`

This tells GENEPOP to run the analyses as given through the menus: `MenuOptions=1.1` will run option 1 sub-option 1 (test for heterozygote deficit), `MenuOptions=1.1,2.2` will run option 1.1 then 2.2, and so on.

`AllelicDistance=Size` (or `=AlleleSize`)

This tells GENEPOP to use allele size-based statistics (where meaningful). Allele sizes are allele names unless specified by the next setting:

`AlleleSizes`

In the above example, the first such line `AlleleSizes=1:5,2:10,3:15,10:50` says that at the first locus, allele 1 has size 5, allele 2 has size 10... 0 cannot be given a size since it means missing information. Any unlisted allele retain its name as its size. The second line specifies allele size at the second locus. The third line `AlleleSizes=` implies that at the third locus, all alleles retain their name as their size (don't forget the '='). It is needed only so that the next line `AlleleSizes=1:5,2:10,3:15,10:50` refers to the fourth locus. As there are four `AlleleSizes` declarations, alleles retain their name as their size for any locus beyond the fourth one.

`RandomSeed` and `MantelSeed`

One may change the seed of the pseudo-random number generator by the setting `RandomSeed=`*value*, except for the Mantel test for which the seed is given by the setting `MantelSeed=`*value*. The default value for both seeds is 67144630.

`Maxima`

With this setting, GENEPOP will only display some maximal values, including the maximum `int` and `long int` values for the compiler (the Markov chain dememorization and batch length are `long int` and the number of batches is `int`).

4 All menu options

4.1 Option 1: Hardy-Weinberg (HW) exact tests

The following menu appears:

```
Hardy Weinberg tests:
```

```
HW test for each locus in each population:
   H1 = Heterozygote deficiency.......1
   H1 = Heterozygote excess...........2
   Probability test...................3

Global test:
   H1 = Heterozygote deficiency.......4
   H1 = Heterozygote excess...........5

Main menu..............................6
```

4.1.1 Sub-options 1–3: Tests for each locus in each population

Three distinct tests are available, all concerned with the same null hypothesis (random union of gametes). The difference between them is the construction of the rejection zone. For the Probability test (sub-option 3), the probability of the observed sample is used to define the rejection zone, and the P-value of the test corresponds to the sum of the probabilities of all tables (with the same allelic counts) with the same or lower probability. This is the "exact HW test" of Haldane (1954), Weir (1996), Guo & Thompson (1992) and others. When the alternative hypothesis of interest is heterozygote excess or deficiency, more powerful tests than the probability test can be used (Rousset & Raymond, 1995). One of them, the score test or U test, is available here, either for heterozygote deficiency (sub-option 1) or heterozygote excess (sub-option 2). The multi-samples versions of these two tests are accessible through sub-options 4 or 5.

Two distinct algorithms are available: first, the complete enumeration method, as described by Louis & Dempster (1987). This algorithm works for less than five alleles. As an exact P-value is calculated by complete enumeration, no standard error is computed. Second, a Markov chain (MC) algorithm to estimate without bias the exact P-value of this test (Guo & Thompson, 1992), and three parameters are needed to control this algorithm (see Section 6.3). These different values may be provided either at Genepop's request, or through the `Dememorisation`, `BatchLength` and `BatchNumber` settings. Two results are provided for each test by the MC algorithm: the estimated P-value associated with the null hypothesis of HW equilibrium, and the standard error (S.E.) of this estimate.

For all tests concerned with sub-options 1-3, there are three possible cases. The number of distinct alleles at each locus in each sample is

no more than 4: Genepop will give you the choice between the complete enumeration and the MC method. If you have less than 1000 individuals per sample, the complete enumeration is recommended. Otherwise, the MC method could be much faster. But there are no general rules, results are highly variable, depending also on allele frequencies.

always 5 or more: GENEPOP will automatically perform only the MC method.

sometimes higher than 4, sometimes not: For cases where the number of alleles is 4 or lower, GENEPOP will give you the choice between both methods. For the other situations (5 alleles or more in some samples), the MC method will be automatically performed.

Whether one wants enumeration or MC methods to be performed can be specified at runtime, or otherwise by the `HWtests` setting, with options `HWtests=enumeration` and `HWtests=MCMC`. The default in the batch mode is `enumeration`.

Output

Results are stored in a file named as follows

sub-option	Extension
1	*yourdata*`.D`
2	*yourdata*`.E`
3	*yourdata*`.P`
4	*yourdata*`.DG`
5	*yourdata*`.EG`

where *yourdata* is (throughout this document) the name of the input file.

For each test, several values are indicated on the same line: (i) the P-value of the test (or "-" is no data were available, or only one allele was present, or two alleles were detected but one was represented by only one copy); (ii) the standard error (only if a MC method was used); (iii) two estimates of F_{IS}, Weir & Cockerham's (1984) estimate (W&C), and Robertson & Hill's (1984) estimate (R&H). The latter has a lower variance under the null hypothesis. Finally, the number of "steps" is given: for the complete enumeration algorithm this is the number of different genotypic matrices considered, and for the Markov chain algorithm the number of switches (change of genotypic matrice) performed.[10]

4.1.2 Sub-options 4,5: Global tests across loci or across samples

For sub-option 3, a global test across loci or across sample is constructed using Fisher's method. This method (sometimes conservative because discrete probabilities are analyzed), is only performed for convenience and its relevance should be first established (e.g. statistical independence of loci).

General statistical theory shows that there is no uniformly better way to combine P-values of different tests. When an alternative model is specified, it is possible to find a better way of combining results from different data sets than Fisher's method, and usually not by combining P-values. In the present context one such method is the

[10] New to GENEPOP 4.0.

multisample score test of Rousset & Raymond (1995), which defines a global test across loci and/or across samples generalizing the tests of sub-options 1 and 2. The global tests are performed by sub-options 4 and 5, only by the MC algorithm. Independence of loci is also assumed for these global tests.

The output file reports global P value estimates and standard errors per population, per locus, and over all loci and populations. For each global P value, the average number of switches per test combined is also reported. Since it is tempting to reduce the chain length parameters in this option, special care is needed in checking this accuracy diagnostic (see p. 39).[11]

This option generates several large temporary files. The space used temporarily by GENEPOP can be estimated as: (# of Loci+# of pop+1)*batches*(iterations per batch)*8 octets. For example it will require about 240 Mo of temporary hard disk space if you have 10 loci, 50 samples and if you use a chain of 500,000 steps (100 batches of 5000 iterations).

4.1.3 Analyzing a single genotypic matrix

It is possible to perform a single HW test independently of the GENEPOP input file. This option is not presented in the GENEPOP menu. You should have an input file with a genotypic matrix (which can be taken from the output file of option 5 and edited), and use the `HWfile` setting.[12] When GENEPOP is launched in this way, the following menu will appear:

```
HW test for each locus in each population:
   H1 = Heterozygote deficiency .................1
   H1 = Heterozygote excess .....................2
   Probability test .............................3

Allele frequencies, expected genotypes, Fis .... 4
Quit ........................................... 5
```

All HW tests corresponding to options 1.1–3 of "regular" GENEPOP are available through options 1–3, and basic information similar to that given by regular option 5.1 is available through the present option 4. Results are stored at the end of your input file. The exact format of the input file is:

First line: anything. Use this line to store information about your data.

Second line: The number of alleles n.

Line three through $n+2$: the genotypic matrix (see example).

[11] Again new to GENEPOP 4.0.

[12] In earlier versions of GENEPOP, this analysis was done through the HW.BAT batch file.

Beyond line $n+2$: anything (this is not read by the program).

An example with four alleles is:

```
Human Monoamine Oxidase (MOAO) Data
4
2
12 24
30 34 54
22 21 20 10
```

If this file is named `MOAO`, you can analyze it by setting `HWfile=MOAO` in the settings; you can also set `HWfileOptions=1` to run option 1 without making your way through the menus. All this can be done through the console command line. For example

```
Genepop HWFile=MOAO HWfileOptions=1,2,3,4
```

will perform all four analyses available through the above menu. General settings `Dememorisation`, `BatchLength`, `BatchNumber`, and `Mode` all affect these analyses in the same way as they affect analyses of regular input files.

Code checks

Code for HW tests has a now venerable history of testing. Early versions of GENEPOP were compared with the EXACTP step in BIOSYS (Swofford & Selander, 1989) for two allele cases, and with data published in Louis & Dempster (1987) and Guo & Thompson (1992) for more alleles. The sample files `LouisD87.txt` and `GuoT92.txt` contain two such test samples, in single-matrix format.

4.2 Option 2: Tests and tables for linkage disequilibrium

The following menu appears:[13]

```
Pairwise associations (haploid and genotypic disequilibrium):
      Test for each pair of loci in each population ......... 1
      Only create genotypic contingency tables ............... 2

Menu  ..................................................... 3
```

[13]The distinct option 2.3 of GENEPOP 3.4 is no longer necessary as option 2.1 of GENEPOP 4.0 more gracefully handles haploid data.

4.2.1 Sub-option 1: Tests

For this option the null hypothesis is: "Genotypes at one locus are independent from genotypes at the other locus". For a pair of diploid loci, no assumption is made about the gametic phase in double heterozygotes. In particular, it is not inferred assuming one-locus HW equilibrium, as such equilibrium is not assumed anywhere in the formulation of the test. The test is thus one of association between diploid genotypes at both loci, sometimes described as a test of the composite linkage disequilibrium (Weir, 1996, p. 126–128). For a haploid locus and a diploid one, a test of association between the haploid and diploid genotypes is computed (there is no concern about gametic phase in this case). This makes it easy to test for cyto-nuclear disequilibria. For a pair of loci with haploid information, a straightforward test of association of alleles at the two loci is computed.

The default test statistic is now the log likelihood ratio statistic (G-test). However one can still perform probability tests (as implemented in earlier versions of GENEPOP) by using the `GameticDiseqTest=Proba` setting.

For a given pair of loci within one sample, the relevant information is represented by a contingency table looking e.g. like

```
       GOT2
       1.1  1.3  3.3  1.7  3.7
EST    -------------------------
 1.1   1    1    0    0    1      3
 1.2   16   6    1    3    2      28
       -------------------------
       17   7    1    3    3      31
```

for two diploid loci (`1.1`, etc., are the diploid genotypes at each locus). Contingency tables are created for all pairs of loci in each sample, then a G test or a probability test for each table is computed for each table using the Markov chain algorithm of Raymond & Rousset (1995a). The number of switches of the algorithm is given for each table analyzed.[14]

Output

Results are stored in the file *yourdata*`.DIS`. Three intractable situations are indicated: empty tables ("No data"), table with one row or one column only ("No contingency table"), and tables for which all rows or all columns marginal sums are 1 ("No information"). For each locus pair within each sample, the unbiased estimate of the P-value is indicated, as well as the standard error. Next, a global test (Fisher's method) for each pair of loci is performed across samples.

See also the next section for analysis of a single table.

[14] This was not the case in earlier versions of GENEPOP

4.2.2 Sub-option 2: create tables

Suboption 2 only generates the above contingency tables and stores them in the file *yourdata*`.TAB`

Code checks

See code checks for Option 3.

4.3 Option 3: population differentiation

The following menu appears:

```
 Testing population differentiation :

      Genic differentiation:
           for all populations ..................... 1
           for all pairs of populations ............. 2

      Genotypic differentiation:
           for all populations ..................... 3
           for all pairs of populations ............. 4

      Main menu  ................................... 5
```

All tests are based on Markov chain algorithms. The Markov chain parameters are controlled exactly as in option 1.

4.3.1 Sub-options 1 or 2 (genic differentiation)

They are concerned with the distribution of alleles is the various samples. The null hypothesis tested is "alleles are drawn from the same distribution in all populations". For each locus, the test is performed on a contingency table like this one:

```
Sub-Pop.  Alleles
          1    2   Total

          -------
 1        14   46   60
 2        6    76   82
 3        10   74   84
 4        4    58   62

          -------
Total     34   254  288
```

For each locus, an unbiased estimate of the P-value is computed. The test statistic is either the probability of the sample conditional on marginal values, the G log likelihood ratio, or the level of gene diversity. In the first case, the test is Fisher's exact probability test, and the algorithm is described in Raymond & Rousset (1995a). A simple modification of this algorithm is used for the exact G test.[15] GENEPOP's default is the G test. You can revert to Fisher's test by using the `DifferentiationTest=Proba` setting. Finally, the level of gene diversity can be used as a test statistic when coupled with the `GeneDivRanks` setting (this new to version 4.1; see Section 4.3.4).

For sub-option 2, the tests are the same, but they are performed for all pairs of samples for all loci.

4.3.2 Sub-options 3 or 4 (genotypic differentiation)

are concerned with the distribution of diploid genotypes in the various populations. The null hypothesis tested is "genotypes are drawn from the same distribution in all populations". For each locus, the test is performed on a contingency table like this one:

	Genotypes:						
	1	1	2	1	2	3	
Pop:	1	2	2	3	3	3	All
Pop1	142	27	0	13	1	0	183
Pop2	149	20	0	11	0	4	184
Pop3	131	12	0	9	0	1	153
Pop4	119	22	1	10	0	0	152
Pop5	120	17	1	10	1	0	149
Pop6	134	18	2	15	0	0	169
Pop7	116	15	1	10	1	1	144
Pop8	214	41	3	14	2	1	275
Pop9	84	17	0	7	2	0	110
Pop10	107	18	0	15	3	0	143
Pop11	134	32	1	21	4	0	192
Pop12	105	26	1	11	1	4	148
Pop13	97	19	2	23	4	0	145
Pop14	95	28	3	19	3	1	149
All:	1747	312	15	188	22	12	2296

An unbiased estimate of the P-value of a log-likelihood ratio (G) based exact test is performed. For this test, the statistics defining the rejection zone is the G value computed

[15] Up to version 3.4, GENEPOP only computed Fisher's exact test in these sub-options.

on the genic table derived from the genotypic one (see Goudet *et al.*, 1996 for the choice of this statistic), so that the rejection zone is defined as the sum of the probabilities of all tables (with the same marginal genotypic values as the observed one) having a G value computed on the derived genic table higher than or equal to the observed G value.

For sub-option 4, the test is the same but is performed for all pairs of samples for all loci.

4.3.3 Output

For the four sub-options, results are stored in a file named as follows:[16]

sub-option	test	output file name
1	Probability test	*yourdata*`.PR`
1	G	*yourdata*`.GE`
2	Probability test	*yourdata*`.PR2`
2	G	*yourdata*`.GE2`
3	G	*yourdata*`.G`
4	G	*yourdata*`.2G2`

All contingency tables are saved in the output file. Two intractable situations are indicated: empty tables or tables with one row or one column only ("No table"), and tables for which all rows or all columns marginal sums are 1 ("No information"). Estimates of P-values are given, as well as (for sub-options 1 and 3) a combination of all test results (Fisher's method), which assumes a statistical independence across loci. For sub-options 2 and 4, this combination of all tests across loci (Fisher's method) is performed for each sample pair. The result `Highly sign.`[ificant] is reported when at least one of the individual tests being combined yielded a zero P-value estimate.

4.3.4 Gene diversity as a test statistic

New to version 4.1

```
DifferentiationTest=GeneDiv
GeneDivRanks=2,1,3,3,3
```

`DifferentiationTest=GeneDiv` makes GENEPOP use gene diversity as test statistic in tests of genetic differentiation (option 3). The test will look for a decrease in gene diversity from populations ranked first (value `1` in `GeneDivRanks`) to populations ranked last. This should work for both genic and genotypic tables, and for pairwise comparisons as well as for all populations, i.e. for all sub-options 3.1 to 3.4. The test statistic is

$$\sum_{\text{all subsamples } i} \sum_{j>i} (Q_j - Q_i)(R_j - R_i) \qquad (1)$$

where Q_i is gene identity in subsample i and R_i is the `GeneDivRanks` value for this subsample.

[16] slightly modified in comparison to earlier versions of GENEPOP

This option also works on input files in contingency table format (`strucfile` setting). In that case each *row* of the table is interpreted as a new population.

4.3.5 Analyzing a single contingency table

It is possible to analyse any contingency table independently of the GENEPOP input file. You should have an input file with a contingency table, and use the `strucFile` setting.[17] This option is not presented in the GENEPOP menu. Both the G and probability tests are available and performed as in option 3.1. Results are stored at the end of your input file. An example of input file is:

```
Dull example
6 5
1   2  5 10 11
2   0  8 11 15
0   0  1  5  6
10 15 20 51 55
0   0  0  2  1
4   5  6 11 10
```

If this file is named `structest`, you can analyze it by writing `StrucFile=structest` in the settings file, or by the console command line

```
Genepop StrucFile=structest
```

The exact format of the input file is:

First line: anything. Use this line to store information about your data.

Second line: The numbers of rows (n) and columns.

Line three through $n+2$: the contingency table (see example).

Beyond line $n+2$: anything (this is not read by the program).

The default is to perform a G test, but as in options 3.1 and 3.2 you can revert to Fisher's exact test by the setting `DifferentiationTest=Proba`.

Code checks

Code for contingency tables also has a venerable history of testing. Early versions of GENEPOP were tested by comparison with published data (e.g. Mehta & Patel, 1983) or by hand calculations. The example file `MehtaP83.txt` contains one such test sample.

[17] In previous versions of GENEPOP, this analysis was done by the STRUC program called through the `Struc.BAT` batch file.

4.4 Option 4: private alleles

This option provides a multilocus estimate of the effective number of migrants (Nm). Three estimates of Nm are provided, using the three regression lines published in Barton & Slatkin (1986), and a corrected estimate is provided using the values from the closest regression line (see Barton & Slatkin, 1986). Results are stored in the file *yourdata*`.PRI`.

4.5 Option 5: Basic information, F_{IS}, and gene diversities

The following menu appears:

```
Allele and genotype frequencies per locus and per sample .. 1

Gene diversities & Fis :
                            Using allele identity ......... 2
                            Using allele size ............. 3

Main menu ..................................................... 4
```

4.5.1 Sub-option 1: Allele and genotype frequencies

This option provides basic information on the data set. The output file is saved in the file *yourdata*`.INF`. For each locus in each sample, several variables are calculated:

- allele frequencies.
- observed and expected genotype proportions.
- F_{IS} estimates for each allele following Weir & Cockerham (1984).
- global estimate of F_{IS} over alleles according to Weir & Cockerham (1984) (W&C) and Robertson & Hill (1984) (R&H).
- observed and "expected" number of homozygotes and heterozygotes. "Expected" here means the expected numbers, conditional on observed allelic counts, under HW equilibrium; the difference from naive products of observed allele frequencies is sometimes called Levene's correction, after Levene (1949).
- the genotypic matrix.

A table of allele frequencies for each locus and for each sample is also computed.

4.5.2 Sub-option 2: Identity-based gene diversities and F_{IS}

This option takes the observed frequencies of identical pairs of genes as estimates ($\hat{Q}$) of corresponding probabilities of identity (Q) and then simply computes diversities as $1-\hat{Q}$: gene diversity within individuals (`1-Qintra`), and among individuals within samples (`1-Qinter`), per locus per sample, and averaged over samples or over loci. One-locus F_{IS} estimates are also computed in a way consistent with Weir & Cockerham (1984). No estimate is given when no information is available (e.g. no estimate of diversity between individuals within a sample when only one individual has been genotyped).

For haploid data, only the gene diversity among individuals is computed. Multilocus estimates ignore haploid loci, or on the contrary ignore diploid loci if the setting `EstimationPloidy=Haploid` is used. Single-locus estimates are computed for both haploid and diploid loci irrespective of this setting.

The output is saved in the file *yourdata*`.DIV`.

4.5.3 Sub-option 3: Allele size-based gene diversities and ρ_{IS}

Option 5.3 is analogous to option 5.2. It computes measures of diversity based on allele size, namely mean squared allele size differences within individuals (`MSDintra`), and among individuals within samples (`MSDinter`), per locus per sample, and averaged over samples or over loci. Corresponding ρ_{IS} (the F_{IS} analogue, see Section 6.6.2) estimates are also computed. Allele size is the allele name unless it has been given through the `AlleleSizes` setting.

For haploid data, only the mean squared difference `MSDinter` among individuals is computed. Multilocus estimates ignore haploid loci, or on the contrary ignore diploid loci if the setting `EstimationPloidy=Haploid` is used. Single-locus estimates are computed for both haploid and diploid loci irrespective of this setting.

The output is saved in the file *yourdata*`.MSD`.

4.6 Option 6: Fst and other correlations, isolation by distance

The following menu appears:

```
Estimating spatial structure:

The information considered is :
     --> Allele identity (F-statistics)
               For all populations ............ 1
               For all population pairs ....... 2
     --> Allele size (Rho-statistics)
               For all populations ............ 3
```

Data ploidy	pop=individual?	`isolationStatistic` setting	Estimator used
Diploid	Yes (option 6.5)	`=a`	$\hat{a}$
Diploid	Yes (option 6.5)	`=e`	$\hat{e}$
Diploid	No (option 6.6)	none (default)	$F_{ST}/(1-F_{ST})$
Diploid	No (option 6.6)	`=singleGeneDiv`	$F/(1-F)$ variant with denominator common to all pairs
Haploid	Yes (option 6.5)	none (default)	$\hat{a}$-like statistic with stand-in for within-deme gene diversity
Haploid	No (option 6.6)	none (default)	$F_{ST}/(1-F_{ST})$
Haploid	No (option 6.6)	`=singleGeneDiv`	$F/(1-F)$ variant with denominator common to all pairs

Table 1: Genetic distance statistics available in options 6.5 and 6.6

```
              For all population pairs ..... 4

Isolation by distance
              between individuals ........... 5
              between groups................. 6

   Main menu  ............................... 7
```

Suboptions 5 and 6 provide a variety of analyses of isolation by distance patterns, including bootstrap confidence intervals of the slope of spatial regression (or equivalently, for "neighborhood" size estimates). Starting with version 4.1, it is even possible to test given values of the slope, through the `testPoint` setting; and additional estimators (merely minor variation on a common logic) have been implemented, in particular for haploid data. The following Table summarizes the choice of methods, each of which is further explained in Table 1

New to version 4.1

4.6.1 Sub-options 1–4: F-statistics and ρ-statistics

These options compute estimates of F_{IS}, F_{IT} and F_{ST} or analogous correlations for allele size, either for each pair of population (sub-options 2 and 4) or a single measure for all populations (sub-options 1 and 3). F_{ST} is estimated by a "weighted" analysis of variance Cockerham (1973); Weir & Cockerham (1984), and the analogous measure of correlation in allele size (ρ_{ST}) is estimated by the same technique (see Section 6.6.2). Multilocus

estimates are computed as detailed in Section 6.6). For haploid data, remember to use the `EstimationPloidy=Haploid` setting.

In sub-option 1, the output is saved in the file *yourdata*`.FST`. Beyond F_{IS}, F_{IT} and F_{ST} estimates, estimation of within-individual gene diversity and within-population among-individual gene diversity are reported as in option 5.2.

In sub-option 2 (pairs of populations), single locus and multilocus estimates are written in the *yourdata*`.ST2` file and multilocus estimates are also written in the *yourdata*`.MIG` file in a format suitable for analysis of isolation by distance (see option 6.6 for further details).

Sub-option 3 is analogous to sub-option 1, but for allele-size based estimates. the output is saved in the file *yourdata*`.RHO`. Beyond ρ_{IS}, ρ_{IT} and ρ_{ST} estimates, estimation of within-individual gene diversity and within-population among-individual gene diversity are reported as in option 5.3.

Sub-option 4 is analogous to sub-option 2, but for allele-size based estimates. Output file names are as in sub-option 2.

4.6.2 Sub-option 5: isolation by distance between individuals

This option allows analysis of isolation by distance between pairs of individuals. It provides estimates of "neighborhood size", or more precisely of $D\sigma^2$, the product of population density and axial mean square parent-offspring distance, derived from the slope of the regression of pairwise genetic statistics against geographical distance or log(distance) in linear or two-dimensional habitats, respectively. More details are described in Rousset (2000) ($\hat{a}$ statistic), Leblois *et al.* (2003) (bootstrap confidence intervals) and Watts *et al.* (2007) ($\hat{e}$ statistic). For haploid data, a proxy for the $\hat{a}$ statistic has been introduced in version 4.1.

New to version 4.1

The position of individuals must be specified as two coordinates standing for their name (i.e. before the comma on the line for each individual), and since each individual is considered as a sample, it must be separated by a `Pop`. An example of such input file is given below: The first individual is located at the point $x = 0$, $y = 15$, the second at the point $x = 0$, $y = 30$, etc.

```
Title line: A really too small data set
ADH Locus 1
ADH #2
ADH three
ADH-4
ADH-5
Pop
0 15,  0201 0303 0102 0302 1011
Pop
0 30,  0202 0301 0102 0303 1111
```

```
Pop
0 45,  0102 0401 0202 0102 1010
Pop
0 60,  0103 0202 0101 0202 1011
Pop
0 75,  0203 0204 0101 0102 1010
POP
15 15,      0102 0202 0201 0405 0807
Pop
15 30,      0102 0201 0201 0405 0307
Pop
15 45,      0201 0203 0101 0505 0402
Pop
15 60,      0201 0303 0301 0303 0603
Pop
15 75,      0101 0201 0301 0505 0807
```

Missing information arises when there is no genetic estimate (if a pair of individuals has no genotypes for the same locus, for example), or when geographic distance is zero and log(distance) is used. GENEPOP will correctly handle such missing information until it comes to the point where regression cannot be computed or there are not several loci to bootstrap over.

Options to be described within option 6.5 are: $\hat{a}$ or $\hat{e}$ pairwise statistics (for diploid data); log transformation for geographic distances; minimal geographic distance; coverage probability of confidence interval; testing a given value of the slope; Mantel test settings; conversion to genetic distance matrix in Phylip format. Allele-size based analogues of $\hat{a}$ or $\hat{e}$ can be defined, but they should perform very poorly (Leblois *et al.*, 2003; Rousset, 2007), so such an analysis has been purposely disabled.

Pairwise statistics for diploid data: They are selected by the setting `IsolationStatistic=a` or `=e`, or at runtime (in batch mode, the default is $\hat{a}$). The $\hat{e}$ statistic is asymptotically biased in contrast to $\hat{a}$, but has lower variance. The bias of the $\hat{e}$-based slope is higher the more limited dispersal is, so it performs less well in the lower range of observed dispersal among various species. Confidence intervals are also biased (Leblois *et al.*, 2003; Watts *et al.*, 2007), being too short in the direction of low $D\sigma^2$ values, and on the contrary conservative in the direction of low $D\sigma^2$ values. Based on the simulation results of Watts *et al.* (2007), a provisional advice is to run analyses with both statistics, and to derive an upper bound for the $D\sigma^2$ confidence interval (CI), hence the lower bound for the regression slope, from $\hat{e}$ (which has CI shorter than $\hat{a}$, though still conservative) and the other $D\sigma^2$ bound, hence the upper bound for the regression slope, from $\hat{a}$ (which has too short CI, but less biased than the $\hat{e}$ CI). When the $\hat{e}$-based $D\sigma^2$ estimate is below 2500 (linear habitat) or 4 (two-dimensional habitat) it is suggested to

derive both bounds from $\hat{a}$.

Note that $\hat{e}$ is essentially Loiselle's statistic (Loiselle *et al.*, 1995), which use in this context has previously been advocated by e.g. Vekemans & Hardy (2004).

For **haploid data** (i.e. `EstimationPloidy=Haploid`) the denominators of the $\hat{a}$ and $\hat{e}$ statistics cannot be computed. Ideally the denominator should be the gene diversity among individuals that would compete for the same position, as could be estimated from "group" data. As a reasonable first substitute, GENEPOP uses a single estimate of gene diversity (from the total sample and for each locus) to compute the denominators for all pairs of individuals. This amount to assume that overall differentiation in the population is weak.

Log transformation for geographic distances: This transformation is required for estimation of $D\sigma^2$ when dispersal occurs over a surface rather than over a linear habitat. It is the default option in batch mode. It can be turned on and off by the setting `GeographicScale=Log` or `=Linear` or equivalently by `Geometry=2D` or `=1D`

Coverage probability of confidence interval This is the target probability that the confidence interval contains the parameter value. The usage is to interval with 95% coverage and 2.5% probabilities that either the confidence interval is too low or too high, and this is the default coverage in GENEPOP. This can be changed by the setting `CIcoverage`, e.g. `CIcoverage=0.99` will compute interval with target probabilities 0.5% that either the confidence interval is too low or too high (an unrealistically large number of loci may be necessary to achieve the latter precision)..

Minimal geographic distance: As discussed in Rousset (1997), samples at small geographic distances are not expected to follow the simple theory of the regression method, so the program asks for a minimum geographical distance. Only pairwise comparisons of samples at larger distances are used to estimate the regression coefficient (all pairs are used for the Mantel test). The minimal distance may be specified by the setting `MinimalDistance=`*value* or at runtime. This being said, it is wise to include all pairs in the estimation as no substantial bias is expected, and this avoids uncontrolled hacking the data. Thus, the suggested minimal distance here is any distance large enough to exclude only pairs at zero geographical distance. Only non-negative values are accepted, and the default in batch mode is 0.0001.

Testing a given value of the slope The setting `testPoint=0.00123` (say) returns the unidirectional P-value for a specific value of the slope, using the ABC bootstrap method. This is the reciprocal of a confidence interval computation: confidence intervals evaluate parameter values corresponding to given error levels, say the 0.025 and 0.975 unidirectional levels for a 95% bidirectional CI, while this option evaluates the unidirectional P-value associated with a given parameter value.

Mantel test: The Mantel test is implemented. In the present context this is an exact test of the null hypothesis that the regression slope is zero. The principle of the Mantel permutation procedure is to permute samples between geographical locations,

so it generates a distribution conditional on having n given sets of genotypic data in n different samples. The permutations provide the distribution of any statistic under the null hypothesis of independence between the two variables (here, genotype counts and geographic location). Mantel (1967) considered a particular statistics and approximations for its distribution. Instead, GENEPOP uses a rank correlation coefficient and no approximation. Isolation by distance will generate positive correlations between geographic distance and genetic distance estimates, and this is best tested using one-tailed P-values. The program provides both one-tailed P-values. The probability of observing the sample rank correlation is the sum of these two P-values minus 1.

Ideally the confidence interval for the slope should contain zero if and only if the Mantel test is non-significant. Some exceptions may occur as the two analyses use different statistics and the bootstrap method is only approximate, but such exceptions appear to be rare. Some exceptions may also occur when some geographically close pairs are excluded from the regression, as these pairs are never excluded from the Mantel test.

The number of permutations may be specified by the setting `MantelPermutations=`*value*, or else at runtime. In batch mode, if no such value has been given the default behaviour is not to perform the test.

Export genetic distance matrix in Phylip format. This option is activated by the setting `PhylipMatrix=` (no *value* needed). It may be useful if you wish to use Phylip to draw a tree based on genetic distances. A constant is added to all values if necessary so that all resulting distances are positive. Output is written in the file *yourdata*`.PMA`. No further estimation or testing is done, so the name of the groups/individuals does not need to be their spatial coordinates.

Except for this export option, output files are:

- the *yourdata*`.ISO` output file, containing (i) a genetic distance ($\hat{a}$ or $\hat{e}$) half-matrix and a geographic (log-)distance half-matrix; missing information is reported as '`-`'; (ii) regression estimates and bootstrap confidence intervals; (iii) the result of testing a slope value (using `testPoint`); (iv) results of a Mantel test for evidence of isolation by distance, if requested. The order of elements in the half-matrices is:

```
      1     2     3
2     x
3     x     x
4     x     x     x
```

- a *yourdata*`.MIG` output file, containing the same genetic and geographic distances as in the `ISO` file, but with more digits, and without estimation or test results.

This file was formerly useful as input for the Isolde program (see "Former option 5 of Genepop", below), and is a bit redundant now.

- a *yourdata*`.GRA` output file, where again the genetic and geographic distances are reported, now as (x, y) coordinates for each pair of individuals (one per line). This is useful e.g. for importing the output into programs with good graphics. Pairs with missing values (either x or y) are not reported in this file.

4.6.3 Sub-option 6: isolation by distance between groups

This option is analogous to the previous one, but derives $D\sigma^2$ estimates from a regression of $F_{\mathrm{ST}}/(1 - F_{\mathrm{ST}})$ estimates to geographic distance in a linear habitat, or log(distance) in a two-dimensional habitat (Rousset, 1997).

Both diploid and haploid data (through `EstimationPloidy=Haploid`) are handled. Missing information is handled as in option 6.5. Input format is the same, except that some samples must contain several individuals. The coordinates of each sample are still contained in the name of each sample, that is in the name of the last individual in each sample.

In addition some allele-size based analyses are possible (by the setting `AllelicDistance=Size`) but again they are not advised in general. Further options within option 6.6 are: `isolationStatistic`; `SingleGeneDiv`; minimal geographic distance; log transformation for geographic distances; testing a given value of the slope; Mantel test settings; conversion to genetic distance matrix in Phylip format. They operate as described above for analyses between individuals, the only difference being the genetic distance used (see Table 1). In particular, a minor variant of the $F/(1 - F)$ estimator is introduced in version 4.1, by analogy to the "between individuals" estimators. Recall that $F/(1 - F) = (Q_r - Q_0)/(1 - Q_0)$ where $1 - Q_0$ is the within-deme gene diversity. The $F/(1-F)$ method uses per-pair estimates of this within-deme gene diversity, which may not be best. With `IsolationStatistic=SingleGeneDiv` a single estimate is used for all pairwise statistics. In principe this should be better when small per-group samples are considered, but the generic $F/(1-F)$ method is still available as the default method. Limited testing so far suggests little effect of the choice of the statistic on inferences from samples with 10 haploid individuals per group and high overall diversity.

New to version 4.1

Output is written in three files *yourdata*`.ISO`, *yourdata*`.MIG`, and *yourdata*`.GRA` with the same contents as in option 6.5, except for the nature of the genetic distances.

4.6.4 Former sub-option 5 of Genepop: analysis of isolation by distance from a genetic distance matrices

That option (using the Isolde program) allowed to perform the analyses of sub-options 5 and 6 from a file with two semi-matrices, one for genetic "distances" (F_{ST} or whatever),

the other for Euclidian distances. These analyses are now available through the `IsolationFile` setting. Most choices within options 6.5 and 6.6 are available through this option, and missing data are handled[18] (see example below). However, it is not possible to compute nonparametric confidence intervals for the regression slope since per-locus information is not provided (remarkably, some software pretends to compute nonparametric intervals in this case). This option may serve as a general purpose program for Mantel tests. Of course, some settings (minimal geographic distance, the $F/(1-F)$ transformation, and the interpretation of one one-tailed P value as a test of isolation by distance) make sense in the narrower inference context of options 6.5 and 6.6.

The option is called by `IsolationFile=`*input file name* where the input file follows the format of the *yourdata*`.MIG` file written by options 6.5 and 6.6, which may be used as models. An example is

```
Lousy data                    <------anything (comments)
8 (an example)                       <---# of samples (comments ignored)
Fst estimates:                               <---anything (comments)
 0.003
 0.18 0.107
 0.19 0.068  0.011
 0.20 0.664  0.665 0.009
 0.21 0.098    -   0.673  0.675
 0.22 0.048  0.682  0.683  0.017  0.001
 0.23 0.715  0.721  0.666  0.666  0.037 0.006
distances:                             <---anything (comments)
 158.0
 158.0 1215.0
 158.1 1213.0 2300.0
 158.2 2300.0    2.0 1057.0
 158.3 1055.0 2525.0 2525.0 1000.0
 158.4 1057.0 1055.0 2525.0 2525.0 1000.0
  - 3582.0 3582.0 3582.0 3582.0    1.0 2.222
Anything after the second half matrix       <----as it says
is ignored
```

The order of elements in the half-matrices is again

```
      1     2      3
2     x
3     x     x
4     x     x     x
```

[18] more extensively than in earlier versions of GENEPOP.

Again as in options 6.5 and 6.6, both missing genetic and geographic information ('-') are handled.

Output is written at the end of the input file, and as in options 6.5 and 6.6, (x, y) data points are also written in the file *yourdata*`.GRA`.

4.6.5 Analysis of isolation by distance from multiple genetic distance matrices

New to version 4.1

If another program as generated F_{ST} or $F_{ST}/(1 - F_{ST})$ matrices for a number of loci, the computation of bootstrap confidence intervals is possible. Analysis of such data sets is allowed by the `MultiMigFile=`*input file name* setting. The format of the input file is the same as for a single genetic matrix, except that it contains multiple matrices and that the number of genetic matrices must be given (third line of input):

```
 More lousy data
 8
 16 loci (for example)                <---# of samples (comments ignored)
 locus 1:                             <---anything (comments)
...                                   <-half matrix (not shown here)
 locus 2:                             <---anything (comments)
...
...                                   <-more loci and half matrices (not shown here)
...
 locus 16:                            <---anything (comments)
...
 Geographic distances:                <---anything (comments)
  158.0
  158.0 1215.0
  158.1 1213.0 2300.0
  158.2 2300.0    2.0 1057.0
  158.3 1055.0 2525.0 2525.0 1000.0
  158.4 1057.0 1055.0 2525.0 2525.0 1000.0
   - 3582.0 3582.0 3582.0 3582.0    1.0 2.222
 Anything after the second half matrix       <----as it says
 is ignored
```

The main use of this option is to allow analyses based on genetic distances not considered in Genepop. If the same estimates are input as would be computed by Genepop, the results should be similar to those from options 6.5 and 6.6, but not identical in general, because Genepop's bootstrap estimates are computed as ratio of weighted average numerators and denominators of genetic estimates, while `MultiMigFile` can only use weighted averages of the ratios, i.e. of the input genetic values.

4.7 Data selection for analyses of isolation by distance

New to version 4.1

4.7.1 Selecting a subset of samples

The settings `PopTypes` and `PopTypeSelection` have been developed to facilitate comparison of differentiation patterns within and among different ecotypes or host races. They are used as follows:

```
PopTypes= 1 1 2 1 2 1 1 2 3 4
PopTypeSelection=only 1
// PopTypeSelection=inter 1 2
// PopTypeSelection=all
```

`PopTypes` allows to distinguish different types of samples (e.g. different ecotypes) by integer indices. The number of indices must match the number of samples in the data file.

`PopTypeSelection` allows performing analyses (genetic distance regressions, confidence intervals, Mantel tests) only on pairs of populations belonging to the types specified. That is, the genetic differentiation statistic among excluded pairs is not used in any of these analyses. The different choices are shown above: `all` excludes no pairs (this is the default value); `inter` *a b* will exclude all pairs that do not involve both types *a* and *b* (only two types can be specified); and `only` *a* will exclude all pairs that involve a type different from *a* (only one type can be specified). For the latter two choices, permutations are made only among samples from a given type.

You have to perform the "`only`" and "`inter`" analyses in distinct Genepop runs if you wish to compare their results. Rousset (1999) explains how inferences can be made from such comparisons. Note that in this perspective, some comparison of the intercept may be useful and that Genepop also provides confidence intervals on the intercept at zero distance [or log(distance)].

The inter-type Mantel test may be misleading. The null hypothesis implied by the permutation procedure is that there is no isolation by distance among populations within each type, rather than the often more relevant hypothesis that spatial processes within each type of populations are independent from each other. For this reason, a more appropriate test of the latter hypothesis is whether the bootstrap confidence interval for the inter-types regression slope includes zero or not.

4.8 Option 7: File conversions

This option allows the conversion of the Genepop input file toward other formats required by some other programs (the "ecumenical" function of Genepop). Given the limited interest in some of these conversions, little effort has been made to update them. In particular, data including haploid loci or in three-digits format may not be converted into valid input for the other programs.

The following menu appears:

```
File conversion (diploid data, 2-digits coding only):

     GENEPOP --> FSTAT (F statistics) ..................... 1
     GENEPOP --> BIOSYS (letter code) ..................... 2
     GENEPOP --> BIOSYS (number code) ..................... 3
     GENEPOP --> LINKDOS (D statistics) ................... 4

     Main menu  ........................................... 5
```

Sub-option 1 converts the GENEPOP input file into the format required by the FSTAT program of Goudet (1995). The new format is saved in the file *yourdata*`.DAT`.

Sub-options 2 and 3 converts the GENEPOP input file into the format required by BIOSYS (Swofford & Selander, 1989), either the letter or the number code. The new format is saved in the file *yourdata*`.BIO`. You should add the STEP procedures at the end of this new file before running BIOSYS. Refer to the BIOSYS manual for details.

Sub-option 4 converts the GENEPOP input file into the format required by LINKDOS, a program described by Garnier-Géré & Dillmann (1992) and based on Black & Krafsur (1985). This program performs pairwise linkage disequilibria analyses in subdivided populations and Ohta's (1982) D statistics. The new format is saved in the file *yourdata*`.LKD`. The source LINKDOS program (LINKDOS.PAS) and an executable (LINKDOS.EXE) have been distributed with previous versions of GENEPOP with permission of their authors, and are still available on the Genepop distribution page. The executable distributed with GENEPOP has been compiled for 40 samples, 20 loci and 99 alleles per locus. It may be wise to relabel alleles (option 8.3) before the conversion. Garnier-Géré & Dillmann (1992) should be cited whenever this program is used.

4.9 Option 8: Null alleles and some input file utilities

The following menu appears[19]

```
Miscellaneous :
   Null allele: estimates of allele frequencies ..... 1
   Diploidisation of haploid data ................... 2
   Relabeling alleles ............................... 3
   Conversion of population data to individual data.. 4
   Main Menu   ...................................... 5
```

[19]Former sub-option 3 (erasing all temporary files) has been discarded.

4.9.1 Sub-option 1: null alleles

This sub-option allows estimation of gene frequencies when a null allele is present. Different methods are available: maximum likelihood, maximum likelihood with genotyping failure, and Brookfield's (1996) estimator, which differences are explained in Section 6.1.[20]

GENEPOP takes the allele with the highest number for a given locus **across all populations** as the null allele.[21] For example, if you have 4 alleles plus a null allele, a null homozygote individual should be indicated as e.g. `0505` or `9999` in the input file.

The default estimation method is maximum likelihood, using the EM algorithm of Dempster *et al.* (1977). Apparent null genotypes may also be due to nonspecific genotyping failures. Joint maximum likelihood estimation of such failure rate ("β") and of allele frequencies is available through the setting `NullAlleleMethod=ApparentNulls`. Finally, the estimator of Brookfield (1996) is also available through the setting `NullAlleleMethod=B96`.. Confidence intervals for null allele frequencies are computed for each locus in each population. Their coverage probability can be modified by the same setting `CIcoverage` as in options 6.5 and 6.6.

The output file is saved in the file *yourdata*`.NUL`. This file may contain

- For the maximum likelihood methods, estimated allelic frequencies and predicted numbers of homozygotes and of heterozygotes with a null allele. For example, in an output such as

```
Allele   EM freq.  Homoz.    Null Heter.
 1       0.2762    2.7046    4.2954
 2       0.2576    1.8500    3.1500
 3       0.2251    1.3567    2.6433
 4       0.0217    0.0000    0.0000
Null     0.2193
```

  of the seven (2.7046+4.2954) apparent homozygotes for allele 1, it is predicted that 4.2954 are actually heterozygotes for allele 1 and for the null allele. This predicted value is the expected, or average, number of such heterozygotes over different samples with the same number of apparent genotypes, under the assumptions of the model.

[20] The last two methods are new to GENEPOP 4.0.

[21] This is a notable difference from GENEPOP, where the allele with the highest number in each population was taken as the null allele in this population. Consequently, null allele estimation is now meaningful even if no null homozygote is observed in a given population. The output format has also been improved, compared to earlier versions of GENEPOP, with a more logical ordering of results (samples within loci) and a final locus by population table of estimated null allele frequencies.

- a summary locus-by-population table of estimates of null allele frequencies.
- a summary locus-by-population table of estimates of genotyping failure frequencies ("`beta`"), if applicable.
- A table of confidence intervals for estimates of null allele frequencies.

Note that there may be insufficient information to compute estimates and/or confidence intervals: not enough alleles in the sample, for example. These are indicated by the message `No information`. Sometimes the point estimate can formally be computed but the computed CI is not meaningful. This happens for example in case of heterozygote excess, and generates a `(No info for CI)` warning (if all pseudo-samples generated by some resampling technique show an heterozygote excess, all pseudo-estimates of null allele frequency will be zero and there is no information to construct a non-null CI from this distribution).

4.9.2 Sub-option 2: Diploidisation of haploid data

This sub-option "diploidizes" haploid loci. For example, the line
`popul 1, 01 02 10 00`
of an haploid dataset with 4 loci, will become
`popul 1, 0101 0202 1010 0000`.
Only haploid data are thus modified in a mixed haploid/diploid data file. The new file is named D*yourdata*.[22]

Note that there may no longer be any need for this option for further analyses with GENEPOP (except perhaps as a preliminary to file conversions, option 7), since GENEPOP 4.0 now perform analyses on haploid data without such prior "diploidization" (don't forget the `EstimationPloidy=Haploid` setting).

4.9.3 Sub-option 3: Relabeling alleles names

This sub-option relabels all alleles starting from 1 up to x, x being the true number of distinct alleles for each locus. The new file is named N*yourdata*. The correspondence between the old and the new numbering is indicated in the file *new_file_name*.NUM. This option was originally introduced in GENEPOP because for some options, the memory space required depends on the highest allele number. I don't expect this to be a cause of concern now.

4.9.4 Sub-option 4: Conversion of population data to individual data

New to version 4.1

This sub-option converts "population" data (with several individuals per `Pop` to "individual" data where each individual is put in a distinct `Pop`. This is useful for individual-based

[22] No longer truncated to 8 letters as it was in earlier versions of GENEPOP

analyses of isolation by distance and, in this perspective, the name of each individual is replaced by what should be its coordinates, that is, the name of the last individual in the original population.

5 Evaluating the performance of inferences for Isolation by distance

GENEPOP can analyze multiple files, using the settings settings

```
GenepopRootFile=file                    <-- or GenepopRootFileName...
JobMin=1
JobMax=100
```

This will perform analysis of data in files `file`1 to `file`100. Default values of these three settings are `GP`, 1, and 1. Users need to assemble results from the multiple output files. A more integrated output is provided for analyses of isolation by distance. For the regression estimators of $D\sigma^2$ (menu options 6.5 and 6.6), the `result.CI` file will contain a Table of point estimates, bootstrap confidence intervals, and (if requested using the `testPoint` setting) the bootstrap P-value for a given tested neighborhood value. including the performance of the bootstrap confidence intervals.

The `Performance=`*value* setting provides a convenient (if somewhat ad hoc) shortcut for selecting the following analyses:

analysis	*value*
$\hat{a}$, 1-dim.	`aLinear` or equivalently `a1D`
$\hat{e}$, 2-dim.	`aPlanar` or `a2D`
$\hat{a}$, 1-dim.	`eLinear` or `e1D`
$\hat{e}$, 2-dim.	`ePlanar` or `e2D`
$F/(1-F)$, 1-dim.	`FLinear` or `F2D`
$F/(1-F)$, 2-dim.	`FPlanar` or `F2D`

`Performance` sets GENEPOP in batch mode. Then, the `GenepopRootFile`, `JobMin`, and `JobMax` values must be given in the settings file. Alternatively, these values can be given interactively if the `Ask` or `Default` mode has been specified *after* the `Performance` setting, but GENEPOP will return in batch mode for all further computation.

6 Methods

This section is only intended as a quick reference guide. The primary literature should be consulted for further information about the methods implemented in GENEPOP.

6.1 Null alleles

When apparent null homozygotes are observed, one may wonder whether these are truly null homozygotes, or whether some technical failure independent of genotype has occurred. Maximum likelihood estimates of null allele frequency, or of this frequency jointly with the failure rate, can be obtained by the EM algorithm (Dempster *et al.*, 1977; Hartl & Clark, 1989; Kalinowski & Taper, 2006), which is one of the methods implemented in GENEPOP (menu option 8.1).

Also implemented is a simpler estimator defined by Brookfield (1996) for the case where apparent null homozygotes are true null homozygotes. He also described this as a maximum likelihood estimator, but there are some (often small) differences with the ML estimates derived by the EM algorithm as implemented in this and previous versions of GENEPOP, which may to be due to the fact that Brookfield wrote a likelihood formula for the number of apparent homozygotes and heterozygotes, while the EM implementation is based on a likelihood formula where apparent homozygotes and heterozygotes for different alleles are distinguished.

For the case where one is unsure whether apparent null homozygotes are true null homozygotes, Chakraborty *et al.* (1992) described a method to estimate the null allele frequency from the other data, excluding any apparent null homozygote. The estimator is not implemented in GENEPOP because, beyond its relatively low efficiency, its behavior is sometimes puzzling (for example, where there is no obvious heterozygote in a sample, the estimated null allele frequency is always 1, whatever the number of alleles obviously present and even if only non-null genotypes are present). Actually, even if apparent null homozygotes are not true null homozygotes, their number bring some information, and it is more logical to estimate the null allele frequency jointly with the nonspecific genotyping failure rate by maximum likelihood (Kalinowski & Taper, 2006). This analysis is possible when at least three alleles are obviously present.

6.2 Exact tests

The probability of a sample of genotypes depends on allele frequencies at one or more loci. In the tests of Hardy Weinberg equilibrium, population differentiation and pairwise independence between loci ("linkage equilibrium") implemented in GENEPOP, one is not interested in the allele frequencies themselves and, given they are unknown, the aim is to derive valid conclusions whatever their values. In these different cases, this can be achieved by considering only the probability of samples conditional on observed allelic (e.g. for HW tests) or genotypic counts (e.g. for tests of population differentiation not assuming HW equilibrium). Because exact probabilities are computed, these conditional tests are also known as exact tests. See Cox & Hinkley (1974) and Lehmann (1994) for the underlying theory; a much more elementary introduction to the tests implemented in GENEPOP is Rousset & Raymond (1997).

The Mantel test is one of the exact tests implemented in GENEPOP, but partial Mantel tests are not implemented. The latter have been used to test for effects of a variable Y on a response variable Z, while removing spatial autocorrelation effects on Z. Both standard theory of exact tests and simulation show that the permutation procedure of the Mantel test is not appropriate for the partial Mantel test when the Y variable itself presents spatial gradients (Oden & Sokal, 1992; Raufaste & Rousset, 2001; Rousset, 2002b). Asymptotic arguments have also been proposed to support the use of permutation tests (e.g. Anderson, 2001) but they fail in the same conditions. More recent papers advocating partial Mantel tests tend not to show false positive rates in scenarios with spatial gradients, or to assert that such cases are "special", without further biologically-based arguments. This may say something about the weakness of the case for partial Mantel tests.

6.3 Algorithms for exact tests

Conditional tests require in principle the complete enumeration of all possible samples satisfying the given condition. In many cases this is not practical, and the P-value may be computed by simple permutation algorithms or by more elaborate Markov chain algorithms, in particular the Metropolis-Hastings algorithm (Hastings, 1970). The latter algorithm explores the universe of samples satisfying the given condition in a "random walk" fashion. For HW testing Guo & Thompson (1992) found a Metropolis-Hastings algorithm to be efficient compared to permutations. A slight modification of their algorithm is implemented in GENEPOP. Guo and Thompson also considered tests for contingency tables (Technical report No. 187, Department of Statistics, University of Washington, Seattle, USA, 1989) and again a slightly modified algorithm is implemented in GENEPOP (Raymond & Rousset, 1995a). A run of the Markov chain (MC) algorithms starts with a dememorization step; if this step is long enough, the state of the chain at the end of the dememorization is independent of the initial state. Then, further simulation of the MC is divided in batches. In each batch a P-value estimate is derived by counting the proportion of time the MC spends visiting sample configurations more extreme (according to the given test statistic) than the observed sample. If the batches are long enough, the P-value estimates from successive batches are essentially independent from each other and a standard error for the P-value can be derived from the variance of per-batch P-values (Hastings, 1970). As could be expected, the longer the runs, the lower the standard error.

6.4 Accuracy of P values estimated by the Markov chain algorithms

For most data sets the MC "mixes well" so that the default values of the dememorization length and batch length implemented in GENEPOP appear quite sufficient (in many other

applications of MC algorithms, things are not so simple; e.g. Brooks & Gelman, 1998). Nevertheless, inaccurate P-values can be detected when the standard error is large or, else if the number of switches (the number of times the sample configuration changes in the MC run) is low (this may occur when the P-value estimate is close to 0 or 1). Therefore, it is wise to increase the number of batches if the standard error is too large, in particular if it is of the order of P (the P-value) for small P or of the order of $1 - P$ for large P, or else if the number of switches is low (< 1000).

6.5 Test statistics

The Markov chain algorithms were first implemented for probability tests, i.e. tests where the rejection zone is defined out of the least likely samples under the null hypothesis. Such tests also had Fisher's preference (e.g. Fisher, 1935); in particular the probability test for independence in contingency tables is known as Fisher's exact test. However, probability tests are not necessarily the most powerful. Depending on the alternative hypothesis of importance, other test statistics are often preferable (see again Cox & Hinkley, 1974 or Lehmann, 1994 for textbook accounts). Efficient tests for detecting heterozygote excesses and deficits (Rousset & Raymond, 1995) were introduced in GENEPOP from the start (see option 1), and log likelihood ratio (G) tests were introduced with the implementation of the genotypic tests for population differentiation (Goudet *et al.*, 1996). The allelic weighting implicit in the G statistic is indeed optimal for detecting differentiation under an island model (Rousset, 2007) and use of the G statistic has been generalized to all contingency table tests in GENEPOP 4.0, though probability tests performed in earlier versions of GENEPOP are still available.

6.6 Estimating F-statistics and related quantities

The definition of F-statistics used here is

$$F_{\mathrm{IS}} \equiv \frac{Q_1 - Q_2}{1 - Q_2} \tag{2}$$

$$F_{\mathrm{ST}} \equiv \frac{Q_2 - Q_3}{1 - Q_3} \tag{3}$$

$$F_{\mathrm{IT}} \equiv \frac{Q_1 - Q_3}{1 - Q_3} \tag{4}$$

where the Q are probabilities of identity in state, Q_1 among genes (gametes) within individuals, Q_2 among genes in different individuals within groups (populations), and Q_3 among groups (populations). Such formulas appear in Cockerham & Weir (1987); see Rousset (2002a) for an account of most implications of such definitions, except estimation.

The commonly held idea that it is more difficult to estimate F-statistics when there are more alleles is generally incorrect; actually many inferences may be more accurate

when more alleles are present (e.g. Leblois *et al.*, 2003, at least as long as gene diversity is less than 0.8). The issue is not to estimate the frequencies of all alleles, but only to estimate the above ratios. Any expression of the form $(Q_i - Q_j)/(1 - Q_j)$ can be estimated as $(\hat{Q}_i - \hat{Q}_j)/(1 - \hat{Q}_j)$ where any $\hat{Q}_k$ is the observed frequency of identical pairs of genes in the sample, among pairs satisfying the condition designated by the k index. This is only slightly different (see Rousset, 2007) from what the following estimators achieve.

6.6.1 ANOVA estimators: single- and multilocus definitions

Well-known work by Cockerham (e.g. Cockerham, 1973; Weir & Cockerham, 1984) has used the formalism of analysis of variance (ANOVA) to define estimators of F-statistics. These estimators may be expressed in terms of the mean sums of squares MSG, MSI, MSP (for Gametes, Individuals, and Populations) computed by an analysis of variance (see e.g. Weir, 1996). Equivalently, they can be expressed in terms of "components of variances" $\hat{\sigma}^2_G$, $\hat{\sigma}^2_I$, $\hat{\sigma}^2_P$ which are unbiased estimates of the corresponding parametric "components of variances" σ^2_G, σ^2_I, σ^2_P in an ANOVA model. The snag is, in general (and in some notable applications), these parametric "components of variance" are not variances but rather differences between variances and can be negative. The σ^2 notation is misleading in this respect; this is a lasting source of confusion, explained in Rousset (2007). Of course, the $\hat{\sigma}^2$ estimators can be negative even if the σ^2 parameters are positive, but this is a distinct issue.

The mean squares can themselves be interpreted in terms of observed frequencies $\hat{Q}$ of identical pairs of genes in the sample. For balanced samples, the relationships are simple: $1 - \hat{Q}_1 = MSG \equiv \hat{\sigma}^2_G$, $\hat{Q}_1 - \hat{Q}_2 = (MSI - MSG)/2 \equiv \hat{\sigma}^2_I$ and $\hat{Q}_2 - \hat{Q}_3 = (MSP - MSI)/(2n) \equiv \hat{\sigma}^2_P$ where n is group size. Hence the single-group (single-population) F_{IS} estimator is

$$\frac{\hat{Q}_1 - \hat{Q}_2}{1 - \hat{Q}_2} = \frac{MSI - MSG}{MSI + MSG} = \frac{\hat{\sigma}^2_I}{\hat{\sigma}^2_I + \hat{\sigma}^2_G}. \tag{5}$$

For unbalanced groups ("populations" of unequal size), estimates over several groups are complex weighted averages of observed frequencies of identical pairs of genes within groups, not detailed here (see Rousset, 2007). However, ANOVA expressions still satisfy $MSG \equiv \hat{\sigma}^2_G$ and $(MSI - MSG)/2 \equiv \hat{\sigma}^2_I$, and $(MSP - MSI)/(2n_c) \equiv \hat{\sigma}^2_P$ where n_c is a function of the size of each group ($n_c \equiv [S_1 - S_2/S_1]/(n-1)$, where S_1 is the total sample size, S_2 is the sum of squared group sizes, and n is the number of non-empty

groups). Then

$$\hat{F}_{\mathrm{IS}} = \frac{MSI - MSG}{MSI + MSG} = \frac{\hat{\sigma}_I^2}{\hat{\sigma}_I^2 + \hat{\sigma}_G^2}, \tag{6}$$

$$\hat{F}_{\mathrm{ST}} = \frac{MSP - MSI}{MSP + (n_c - 1)MSI + n_c MSG} = \frac{\hat{\sigma}_P^2}{\hat{\sigma}_P^2 + \hat{\sigma}_I^2 + \hat{\sigma}_G^2}, \tag{7}$$

$$\hat{F}_{\mathrm{IT}} = \frac{MSP + (n_c - 1)MSI - n_c MSG}{MSP + (n_c - 1)MSI + n_c MSG} = \frac{\hat{\sigma}_P^2 + \hat{\sigma}_I^2}{\hat{\sigma}_P^2 + \hat{\sigma}_I^2 + \hat{\sigma}_G^2}. \tag{8}$$

With several loci, such an analysis is performed for each locus i and the multilocus estimate is the ratio of a weighted sum of the above locus-specific numerators over locus-specific denominators. However, there is no single consistent way to compute the weighted sums. Weir & Cockerham's (1984) multilocus estimators are defined from sums of intermediate statistics a, b, and c for each locus, which appear to be the $\hat{\sigma}^2$'s. The numerator of the multilocus estimator of F_{ST} is thus $\sum_{\mathrm{loci}\ i} a_i = \sum_i [(MSP - MSI)/(2n_c)]_i$. On the other hand Weir's (1996) multilocus estimators are defined from distinct intermediate statistics S_1, S_2, and S_3 for each locus, where for locus i, $S_{1i} = [(MSP - MSI)]_i/(2\bar{n})$ for an average sample size across loci $\bar{n}$, and the numerator of the multilocus estimate is $\sum_{\mathrm{loci}\ i} S_i = \sum_i [an_c]_i/\bar{n}$. Hence the 1984 and 1996 estimators slightly differ.

However, both give the same weight to the estimates of the Q's for a locus typed at 5 individuals in each subpopulation as for a locus typed at 50 individuals in each subpopulation. GENEPOP follows another logic. The multilocus estimator of F_{ST} has numerator $\sum_i [n_c(MSP - MSI)]_i$, which will give 10 time more weight to the Q estimates for the more intensively typed locus. 'Explicit' formulas for the estimators are:

$$\hat{F}_{\mathrm{IS}} = \frac{\sum_i [n_c(MSI - MSG)]_i}{\sum_i [n_c(MSI + MSG)]_i} = \frac{\sum_i [n_c \hat{\sigma}_I^2]_i}{\sum_i [n_c \hat{\sigma}_I^2 + n_c \hat{\sigma}_G^2]_i}, \tag{9}$$

$$\hat{F}_{\mathrm{ST}} = \frac{\sum_i [MSP - MSI]_i}{\sum_i [MSP + (n_c - 1)MSI + n_c MSG]_i} = \frac{\sum_i [n_c \hat{\sigma}_P^2]_i}{\sum_i [n_c \hat{\sigma}_P^2 + n_c \hat{\sigma}_I^2 + n_c \hat{\sigma}_G^2]_i}, \tag{10}$$

$$\hat{F}_{\mathrm{IT}} = \frac{\sum_i [MSP + (n_c - 1)MSI - n_c MSG]_i}{\sum_i [MSP + (n_c - 1)MSI + n_c MSG]_i} = \frac{\sum_i [n_c \hat{\sigma}_P^2 + n_c \hat{\sigma}_I^2]_i}{\sum_i [n_c \hat{\sigma}_P^2 + n_c \hat{\sigma}_I^2 + n_c \hat{\sigma}_G^2]_i}. \tag{11}$$

Data from the example file `Fmulti.txt` (3 samples, 3 loci) illustrate the difference between results obtained by the different methods:

Estimate	F_{IS}	F_{ST}	F_{IT}
locus 1	-0.0483	0.5712	0.5505
locus 2	-0.1161	0.8560	0.8393
locus 3	0.0051	-0.0023	0.0028
Multilocus (1984 a,b,c method)	-0.0286	0.5606	0.5480
Multilocus (1996 S1,S2,S3 method)	-0.0286	0.5633	0.5508
Multilocus (GENEPOP v3.3 and later)	-0.0275	0.5436	0.5310

Most of the time the different estimators yield close values; I expect the GENEPOP method to provide better F_{ST} estimates under weak differentiation.

6.6.2 Microsatellite allele sizes, R_{ST}, and ρ_{ST}

Following Slatkin (1995), statistics based on allele size have been widely used. The parameters ρ_{IS}, ρ_{ST} and ρ_{IT} and their estimators are defined by replacing any $1-Q_k$ by the expected square difference in allele size between the genes compared (Rousset, 1996) in all formulas above, and any $1-\hat{Q}_k$ by the observed mean square difference (more formulas are given in Michalakis & Excoffier, 1996). Then the estimators become plain ANOVA estimators of intraclass correlation for allele size; if there are only two alleles, $\hat{\rho}_{ST} = \hat{F}_{ST}$, but Slatkin's $R_{ST} \neq \hat{F}_{ST}$.

6.6.3 Robertson and Hill's estimator of F_{IS}

This estimator, reported in options 1 and 5, was designed to have lower variance than the ANOVA estimator and no small-sample bias when F_{IS} is low, assuming a probability model for sample probabilities (Robertson & Hill, 1984). The score test computed in heterozygote excess and deficiency sub-options of option 1 is equivalent to this estimator for testing purposes.

6.7 Bootstraps

Option 6 constructs approximate bootstrap confidence (ABC) intervals (DiCiccio & Efron, 1996), assuming that each locus is an independent realization of genealogical and mutation processes. The bootstrap is a general methodology with different incarnations. The ABC methods were chosen because they balance moderate computation needs with good accuracy compared to alternatives. Bootstrap methods are approximate, and simulation tests of their performance (a too rare deed in statistical population genetics) for the present application are reported in Leblois *et al.* (2003) and Watts *et al.* (2007).

The ABC method is also applied over individuals in option 8 to compute confidence intervals for null allele frequency estimates.

7 Code history, compilation, credits, contact, etc.

The present version of GENEPOP is a C++ rewrite of GENEPOP 3.4 (Raymond & Rousset, 1995b) by F.R., using draft C translations of many GENEPOP modules by O. Guillaume, N. Benhamou and A. André, and some draft C++ classes by R. Leblois. GENEPOP uses R. J. Wagner's implementation of the Mersenne Twister random number generator (Matsumoto & Nishimura, 1998). The GENEPOP Windows executable has been compiled with GNU's C++ compiler from the "Minimalist GNU for Windows" (Mingw) distribution (`g++` version 4.2.1-sjlj).

Beyond M. Raymond and F.R., credit for previous GENEPOP code is as follows. The complete enumeration procedure for HW tests was derived from Fortran code provided by E. J. Louis (Inst. Mol. Med., Oxford, UK). Some of the procedures for isolation by distance "between individuals" were first written by R. Leblois with help from S. Piry (INRA-CBGP, Montpellier). P. David, É. Imbert and S. Samadi wrote some early code in 1993.

J.S.F. Barker, R. Leblois, T. de Meeus, P. Morin, V. Ravigné and Y. Zimmermann have pointed issues with Genepop 4.0 before or after release. É. Imbert, S. Ponsard and D. Bourguet stimulated additional developments.

B. Anderson, M.A. Beaumont, A. Becher, T.J.C. Beebee, S. Bellman, L. Bernatchez, D. Bourguet, J. Britton-Davidian, E. Bucheli, J. Carlier, G. Carmody, R. Castilho, F. Catzeflis, C. Chevillon, J. Clayton, J. Dallas, P. David, P. Dias, B. Dodd, R. Eritja, A. Estoup, A.-B. Failloux, E. Fjerdingstad, R.C. Fleischer, A.J. Gharrett, S. T. Glenn, S.(?) Goodman, J. Goudet, L. Henke, D. Innes, P. Jarne, L. Jermiin, J. Kelso, N. Khromov-Borissov, J. Lagnel, M. Lascoux, L.S. Magnussen, J. Mallet, D., (?) McDonald, C. Moran, F. Nicholas, I. Olivieri, M. van Oppen, N. Pasteur, R. Paxton, F. Renaud, H. Rosa, L., P. W. Shaw, Shapiro, J. Shykoff, D. Sicard, J. Slate, M. Slatkin, M. Small, T. Staedler, F. Thomas, F. Viard, P. Waldmann, K. J. Wetherall, (?) Winker, Z. Xu, made suggestions or tests on the various states of GENEPOP until version 3.4.

7.1 Contact

If you think you have found a bug, you can contact me. Requests which do not meet the following requirements are likely to meet poor response. Please provide a minimal input file illustrating the suspected problem, whenever relevant. Please use the latest version of GENEPOP taken from a web page I maintain. **Note that I do not maintain the "Genepop on the web" port of Genepop: any question related to this port should be addressed to Eleanor Morgan.** Please specify the version of GENEPOP you are using. Please do not ask whether GENEPOP is commercial software. Please read the documentation.

I may answer queries about methods implemented in GENEPOP, and the more so when they are specific to GENEPOP. But in most cases there are published references

describing the methods, cited in the documentation. Please read the documentation.

Bug fixes since release of Genepop version 3.4 in May 2003 until first release of Genepop 4.0:

The sign of the lower confidence interval bound for regression slope in ISOLDE did not appear on output file when it was negative.

For computation of allele size-based statistics (Option 6.2 and 6.4) with the option "allele name = allele size", the allele '99' was interpreted as having size zero.

8 Copyright

The sources of earlier versions of GENEPOP were distributed as "public domain". The GENEPOP 4.+ code is © F. Rousset, and distributed under the GPL-compatible CeCill licence (see http://www.cecill.info/index.en.html). The Mersenne Twister code is © R. J. Wagner, and open source code under the BSD Licence.

Bibliography

Anderson, M. J., 2001. Permutation tests for univariate or multivariate analysis of variance and regression. *Can. J. Fish. Aquatic Sci.* **58**: 626–639.

Barton, N. H. & Slatkin, M., 1986. A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. *Heredity* **56**: 409–415.

Black, IV, W. C. & Krafsur, E. S., 1985. A FORTRAN program for the calculation and analysis of two-locus linkage disequilibrium coefficients. *Theor. Appl. Genet.* **70**: 491–496.

Brookfield, J. F. Y., 1996. A simple new method for estimating null allele frequency from heterozygote deficiency. *Mol. Ecol.* **5**: 453–455.

Brooks, S. & Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. *J. Computational Graphical Stat.* **7**: 434–455.

Chakraborty, R., de Andrade, M., Daiger, S. & Budowle, B., 1992. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Ann. Hum. Genetics* **56**: 45–57.

Cockerham, C. C., 1973. Analyses of gene frequencies. *Genetics* **74**: 679–700.

Cockerham, C. C. & Weir, B. S., 1987. Correlations, descent measures: drift with migration and mutation. *Proc. Natl. Acad. Sci. U. S. A.* **84**: 8512–8514.

Cox, D. R. & Hinkley, D. V., 1974. *Theoretical statistics*. Chapman & Hall, London.

Dempster, A. P., Laird, N. M. & Rubin, D. B., 1977. Maximum Likelihood from incomplete data via the *EM* algorithm (with discussion). *J. R. Stat. Soc. B* **39**: 1–38.

DiCiccio, T. J. & Efron, B., 1996. Bootstrap confidence intervals (with discussion). *Stat. Sci.* **11**: 189–228.

Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fisher, R. A., 1935. The logic of inductive inference (with discussion). *J. R. Stat. Soc.* **98**: 39–82.

Garnier-Géré, P. & Dillmann, C., 1992. A computer program for testing pairwise linkage disequilibria in subdivided populations. *J. Hered.* **83**: 239.

Goudet, J., 1995. FSTAT (Version 1.2): A computer program to calculate F-statistics. *J. Hered.* **86**: 485–486.

Goudet, J., Raymond, M., de Meeüs, T. & Rousset, F., 1996. Testing differentiation in diploid populations. *Genetics* **144**: 1931–1938.

Guo, S. W. & Thompson, E. A., 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**: 361–372.

Haldane, J. B. S., 1954. An exact test for randomness of mating. *Journal of Genetics* **52**: 631–635.

Hartl, D. L. & Clark, A. G., 1989. *Principles of population genetics*. Sinauer, Sunderland, Mass., 2nd edn.

Hastings, W. K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.

Kalinowski, S. T. & Taper, M. L., 2006. Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conserv. Genetics* **7**: 991–995.

Leblois, R., Estoup, A. & Rousset, F., 2003. Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance. *Mol. Biol. Evol.* **20**: 491–502.

Lehmann, E. L., 1994. *Testing statistical hypotheses*. Chapman & Hall, New York, 2nd edn.

Levene, H., 1949. On a matching problem arising in genetics. *Annals of Mathematical Statistics* **20**: 91–94.

Loiselle, B. A., Sork, V. L., Nason, J. & Graham, C., 1995. Spatial genetic structure of a tropical understory shrub *Psychotria officinalis* (Rubiaceae). *Am. J. Bot.* **82**: 1420–1425.

Louis, E. J. & Dempster, E. R., 1987. An exact test for Hardy-Weinberg and multiple alleles. *Biometrics* **43**: 805–811.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**: 209–220.

Matsumoto, M. & Nishimura, T., 1998. Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. on Modeling and Computer Simulation* **8**: 3–30.

Mehta, C. R. & Patel, N. R., 1983. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Am. Stat. Soc.* **78**: 427–434.

Michalakis, Y. & Excoffier, L., 1996. A generic estimation of population subdivision using distances between alleles with special interest to microsatellite loci. *Genetics* **142**: 1061–1064.

Oden, N. L. & Sokal, R. R., 1992. An investigation of three-matrix permutation tests. *J. Classif.* **9**: 275–290.

Ohta, T., 1982. Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc. Natl. Acad. Sci. U. S. A.* **79**: 1940–1944.

Raufaste, N. & Rousset, F., 2001. Are partial Mantel tests adequate? *Evolution* **55**: 1703–1705.

Raymond, M. & Rousset, F., 1995a. An exact test for population differentiation. *Evolution* **49**: 1283–1286.

Raymond, M. & Rousset, F., 1995b. GENEPOP Version 1.2: population genetics software for exact tests and ecumenicism. *J. Hered.* **86**: 248–249.

Robertson, A. & Hill, W. G., 1984. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics* **107**: 703–718.

Rousset, F., 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics* **142**: 1357–1362.

Rousset, F., 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**: 1219–1228.

Rousset, F., 1999. Genetic differentiation within and between two habitats. *Genetics* **151**: 397–407.

Rousset, F., 2000. Genetic differentiation between individuals. *J. Evol. Biol.* **13**: 58–62.

Rousset, F., 2002a. Inbreeding and relatedness coefficients: what do they measure? *Heredity* **88**: 371–380.

Rousset, F., 2002b. Partial Mantel tests: reply to Castellano and Balletto. *Evolution* **56**: 1874–1875.

Rousset, F., 2007. Inferences from spatial population genetics. In: *Handbook of statistical genetics* (D. J. Balding, M. Bishop & C. Cannings, eds.), pp. 945–979. Wiley, Chichester, U.K., 3rd edn.

Rousset, F. & Leblois, R., 2007. Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model misspecification. *Mol. Biol. Evol.* **24**: 2730–2745.

Rousset, F. & Raymond, M., 1995. Testing heterozygote excess and deficiency. *Genetics* **140**: 1413–1419.

Rousset, F. & Raymond, M., 1997. Statistical analyses of population genetic data: old tools, new concepts. *Trends Ecol. Evol.* **12**: 313–317.

Slatkin, M., 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**: 457–462.

Swofford, D. L. & Selander, R. B., 1989. *BIOSYS-1. A Computer Program for the Analysis of Allelic Variation in Population Genetics and Biochemical Systematics. Release 1.7.* Illinois Natural History Survey, Champaign.

Vekemans, X. & Hardy, O. J., 2004. New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol. Ecol.* **13**: 921–934.

Watts, P. C., Rousset, F., Saccheri, I. J., Leblois, R., Kemp, S. J. & Thompson, D. J., 2007. Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator. *Mol. Ecol.* **16**: 737–751.

Weir, B. S., 1996. *Genetic Data Analysis II.* Sinauer, Sunderland, Mass.

Weir, B. S. & Cockerham, C. C., 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Index

Allele coding, 8
  2-digits, 8
  3-digits, 2, 8
Allele size-based statistics, 13, *Options* 6.3 & 6.4
  ρ_{ST}, 43
  R_{ST}, 43
`AlleleSizes` setting, 13
`AllelicDistance` setting, 13

Batch mode, 6, 37
`BatchLength` setting, 12
`BatchNumber` setting, 12
BIOSYS program, 34
Bootstrap, *see* Confidence intervals
Bug reports, 44
Bugs, 45

`CIcoverage` setting, 28, 35
Code checks, 17, 22
Command line, 10
Concurrent processes, 7
Confidence intervals, 28
  bootstrap, 43

Data selection
  by ploidy, see `estimationPloidy`, 48
  subset of samples, see `popTypeSelection`, 48
`Dememorisation` setting, 12
Differentiation, *Option* 3
  gene diversity, 21
  genic, 19
  genic-genotypic test, 21
  genotypic, 20
`DifferentiationTest` setting, 20, 21
$D\sigma^2$ estimation
  $\hat{a}$ statistic, 27
  $\hat{e}$ statistic, 27
  $F_{ST}/(1-F_{ST})$ statistic, 30
  Loiselle's statistic, 28

`EstimationPloidy` setting, 12
Exact tests, *see also* Differentiation; Linkage disequilibrium; Hardy-Weinberg tests; Mantel test
  conditional tests, 38
  Fisher's, 40
  Metropolis-Hastings algorithm, 39
  permutation algorithms, 39
  probability test, 40

F-statistics, *see also* F_{IS}
  definition, 40
  estimation formulas, 41
  F_{ST}, *Options* 6.1 & 6.2
File conversions, *Option* 7
F_{IS}
  multisample multilocus, *Option* 6.1
  multisample per locus, *Options* 5.2 & 6.1
  per sample multilocus, *Option* 5.2
  per sample per locus, *Options* 5.1 & 5.2
  Robertson & Hill's estimator of F_{IS}, 43
FSTAT program, 34

`GameticDiseqTest` setting, 18
Gene diversities, *Options* 5.2 & 5.3
`GeneDivRanks` setting, 21
GENEPOP, differences from previous versions, 1, *see also footnotes throughout this document*
`GenepopInputFile` setting, 12
`GenepopRootFile` setting, 37
`GeographicScale` setting, 28
`Geometry` setting, 28

Haplo-diploid genotypes, 8
Haploid data, 2, 8, 12, 24, 26, 33, 36
Hardy-Weinberg tests, 15, *Option* 1

multisample score test, 16, *Options* 1.4 & 1.5
score test, 14
`help`, 10
Heterozygosities, *see* Gene diversities
HW program, 2, 16
`HWfile` setting, 2, 16
`HWfileOptions` setting, 17
`HWtests` setting, 15

Individual data from population data, *Option* 8.4
Individual-based analysis
conversion of data for, 36
Input file, *see* `GenepopInputFile`
Input format, 7
for Mantel test, 31
for single contingency table, 22
for single HW test, 16
`InputFile` setting, 12
Isolation by distance
between groups, *Option* 6.6
between individuals, *Option* 6.5
`IsolationFile` setting, 31
`IsolationStatistic` setting, 27
ISOLDE program, 2, 30

`JobMax`, 37
`JobMin`, 37

Levene's correction, 23
Linkage disequilibrium, *Option* 2
composite, 18
cyto-nuclear, 18
Ohta's statistics, 34
LINKDOS program, 5, 34
Linux, 5
installation on, 5

Mac OS X
file format issues, 9
installation on, 5
Mantel test, 13, 28, *Options* 6.5 & 6.6
inter-type, 33
partial, 39
`MantelPermutations` setting, 29
`MantelSeed` setting, 13
Markov chain algorithms
accuracy, 39
parameters, 12
switches, 15, 18, 40
`Maxima` setting, 13
Maximum sample size, 10, *see* `Maxima`
`MenuOptions` setting, 13
Microsoft Windows
file format issues, 9
installation on, 5
`MinimalDistance` setting, 28
Missing data, 9
`Mode` setting, 6, 12, 37
`MultiMigFile` setting, 32

Neighborhood size, *see* $D\sigma^2$ estimation
Null alleles, 35, 38, *Option* 8.1
`NullAlleleMethod` setting, 35

`Performance` setting, 37
PHYLIP package, *see* `PhylipMatrix`
`PhylipMatrix` setting, 29
`PopTypes` setting, 33
`PopTypeSelection` setting, 33
Population differentiation, *see* Differentiation
Population type selection, 33
Private allele method, *Option* 4
Pseudo-random numbers, 13, 44

`RandomSeed` setting, 13
Relabeling alleles, 36, *Option* 8.3
ρ_{IS}
multisample multilocus, *Option* 6.3
multisample per locus, *Options* 5.3 & 6.3
per sample multilocus, *Option* 5.3

per sample per locus, *Option* 5.3
ρ_{ST}, 43, *Options* 6.3 & 6.4
R_{ST}, *see* Allele size-based statistics

Sample size
limitations, 10
Selecting subset of samples, *see* Population type selection
Settings file, 10
`SettingsFile` setting, 6, 12
STRUC program, 2, 22
`StrucFile` setting, 2, 22

`testPoint` setting, 28