Genepop version 4.8.2

F. Rousset

This documentation: 19 January 2023

Contents

Chapter 1

Introduction

1.1 Purpose

This is a documentation for the Genepop software, distributed both as standalone software and as an R package. 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".

A formal reference for the current version of Genepop is Rousset (2008). Likelihood methods based on coalescent algorithms are being developed in a distinct software, Migraine (Rousset and Leblois 2007, 2012; R. Leblois et al. 2014).

Genepop also converts data from the Genepop input format to formats of some softwares that were around in Genepop's youth (Raymond and Rousset

1995b); there has been little need to update this option as many more recent softwares for population genetic analyses read input files in the Genepop format.

1.2 The two Genepop distributions

Genepop is now distributed both as an R package, and as stand-alone software. See the Genepop distribution page for the latter. This documentation describes the use of the executable. The functionalities it describes are available in an R session, using R functions described only in the package documentation.

1.3 Changes since version 4.0

Changes since version 4.6.9 that also affect the R package are described in the NEWS file of that package. Only changes affecting the stand-alone executable are reported here.

Version 4.8.2

Two additional variants of the non-parametric bootstrap have been implemented in analyses of isolation by distance. The new setting `BootstrapMaethod` can be used to select a non-default one.

Version 4.7.2

A new keyword `intra_all_types` for setting `popTypeSelection` allows one to perform a single spatial regression (but not Mantel tests) for all pairs of individuals or populations within types (e.g., individuals within patches, excluding pairwise statistics for pairs of individuals between patches).

Yet another problem has been fixed for Mantel tests' handling of missing pairwise genetic information (specifically for pairs of "pop" – most likely, individuals – sharing no genotypic information at any locus).

Version 4.6.9

Genepop is now also distributed as an R package. It now uses the implementation of the Mersenne twister pseudo-random number generator found in recent C++ compilers. This has two implications. First, a recent compiler

must be used, as described below. Second, test results of previous versions cannot be exactly replicated.

The format of a few file outputs has been modified (in particular the reporting of extreme values of some global tests).

Version 4.6

A bootstrap analysis of mean differentiation has been introduced, in particular to allow comparison of the mean differentiation observed over a given range of geographical distances, in intra vs. inter-ecotypic analyses. It can be called by the setting `meanDifferentiationTest`.

The Mantel test based on regression slope (not the one on ranks) was not handling appropriately cases where some pairwise data had to be excluded. This is corrected. Such cases concern in particular pairs of samples in the same location (e.g., pairs of individuals), when geographical distance is log-transformed, because the pairwise differentiation between such individuals cannot be used for the computation of the regression. The bootstrap analyses were already handling correctly this case.

Version 4.5

A new keyword `inter_all_types` for setting "popTypeSelection" allows one to perform spatial regressions (but not Mantel tests) between all pairs of individuals or populations belonging to different types (e.g., individuals belonging to different patches, excluding pairwise statistics for pairs of individuals within patches).

Version 4.4

Mantel tests are by default no longer based on rank correlation. The older rank tests can be performed using the new `MantelRankTest` setting. In addition, a `MaximalDistance` setting has been added, affecting the computation of spatial regressions.

Version 4.3

Two new "miscellaneous" conversion options have been added: option 8.5 converts population data to individual data (as 8.4) but keeps the individual names (hence the geographic location of each individual); and option 8.6 randomly samples haploid data at diploid loci.

Version 4.2

One can now perform all isolation-by-distance analyses with a user-provided distance matrix instead of the geographic distance matrix computed from the coordinates of the samples (`geoDistFile` setting).

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

Version 4.0 was a complete rewrite of the fossil version 3.4, with the following changes:

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.

Chapter 2

Installing Genepop and session examples

2.1 Installation

2.1.1 R package

As any R package, it can be installed by `install.packages("genepop")` if on CRAN, and more generally by `install.packages(,type="source",repos=NULL)`. See the R documentation for more information.

2.1.2 Stand-alone executable

Under **Microsoft Windows**, one only needs to unzip/copy the executable on hard disk. Both 32- and 64-bit versions of the executables are distributed. Under **Linux/Mac OSX**, extract all c++ sources from the distributed `sources.tar.gz` (or from the `src/` subdirectory of the R package sources, **except** `RcppExports.cpp`), and compile with a compiler that supports the C++11 standards. For **Windows**, one can use `g++` version 4.9.3 (distributed with recent versions of the R tools) with an ad hoc flag:

```
g++ -std=c++11 -o Genepop *.cpp -O3
```

(`O` in `-O3` is the letter O, not zero). With more recent versions of `g++` (>=6.0) or `clang++`, no such flag is required:

```
g++ -o Genepop *.cpp -O3.
```

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.

Examples and documentation files are included in the R package and are available on the Genepop distribution page.

Linkdos, a program described by Garnier-Géré and Dillmann (1992), is distributed with (but is not part of) Genepop. It is originally a DOS program, but the source file distributed can be recompiled under Linux using the Free Pascal compiler (or at least "could'', since this is no longer maintained/checked).

2.2 Example sessions

To reproduce the examples of this session one should copy in a personal directory the examples files found in the `extdata/` subdirectory of the packageor on the Genepop distribution page.

2.2.1 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
```

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

```
    Exact tests for population differentiation (several options) ..... 3

Estimating:
    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`.

The next example illustrates a slightly more elaborate use of Genepop.

2.2.2 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 4 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.

2.2.3 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. `Mode=Batch2File` is a variant of the batch mode that also removes some console outputs. It is suitable for use in running environments where the console output is redirected to a file.

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.

Chapter 3

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)[1]. 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

[1]Earlier versions of Genepop only accepted `Pop`, `POP` and `pop`...

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
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!).[2]
- To designate alleles, consecutive numbers are not required.
- haploid and diploid data can be intermixed.[3] 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 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

[2] New to Genepop 4.0
[3] Also new to Genepop 4.0

- 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`[4] 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 a text editor such as Notepad++.[5] It has also appeared that certain Microsoft products under **Mac OS X** still produced 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.

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.

There are some limits to the number of samples and individuals imposed by the compiler. These values, and a few other ones, are shown by running "`Genepop Maxima=`" (see the `Maxima` setting). However, these built-in maxima are so large[6] as to be practically infinite even in the era of whole-genome sequencing. Computer memory, or user patience, are more likely limits.

[4] New to Genepop 4.0

[5] Other text editors including the Windows basic text editor may not show all end-of-line characters correctly.

[6] in constrast to earlier versions of Genepop

Chapter 4

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.[1]

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:

```
// sample Genepop settings file, showing all options.
/********** Syntax of this file:
```

[1]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`.

```
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
//MantelRankTest=
/PopTypes= 1 2 1 2 3
/PopTypeSelection= all
//PhylipMatrix=
/           Emulating ISOLDE
//IsolationFile=Isoldetest
/           Extending ISOLDE to multiple matrices
//MultiMigFile=perlocusStuff
/ Isolation by distance with user-provided geographic distances
//geoDistFile=someFile
/********** 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 7.3:

the dememorisation number The default is 10000;[2] 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;[3] values below 400 are not allowed.

The maximum allowed value of these parameters will depend on the C++ compiler (it is its maximum `size_t`, that is at least 65535, and typically much more on recent compilers). See the setting `Maxima` if you really need more information about this value.

`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 change some settings in the course of a session). For example it will ask for confirmation of the MC parameters. `Mode=Batch`

[2] increased from Genepop 3.4's default
[3] increased from Genepop 3.4's default

will not wait for feedback: 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`).

Chapter 5

All menu options

5.1 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
```

5.1.1 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), Bruce S. Weir (1996), Guo and 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 and 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 and 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 and Thompson 1992), and three parameters are needed to control this algorithm (see Section 7.3). 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`.

5.1.2 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 "-" if 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}, B. S. Weir and Cockerham (1984)'s (1984) estimate (W&C), and Robertson and Hill (1984)'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.[1]

5.1.3 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).

[1] New to Genepop 4.0.

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 multisample score test of Rousset and 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.41).[2]

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).

5.1.4 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.[3] 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
```

[2] Again new to Genepop 4.0.

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

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).

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.

5.1.5 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 and Selander 1989) for two allele cases, and with data published in Louis and Dempster (1987) and Guo and Thompson (1992) for more alleles. The sample files `LouisD87.txt` and `GuoT92.txt` contain two such test samples, in single-matrix format.

5.2 Option 2: Tests and tables for linkage disequilibrium

The following menu appears:[4]

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

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

5.2.1 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 (Bruce S. Weir 1996, 126–28). 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
```

[4]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.

```
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 and Rousset (1995a). The number of switches of the algorithm is given for each table analyzed.[5]

5.2.2 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.

5.2.3 Sub-option 2: create tables

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

5.2.4 Code checks

See code checks for Option 3.

5.3 Option 3: population differentiation

The following menu appears:

```
Testing population differentiation :
```

[5]This was not the case in earlier versions of Genepop

```
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.

5.3.1 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 and Rousset (1995a). A simple modification of this algorithm is used for the exact G test.[6] Genepop's default is the G test. You can revert to Fisher's test by using the `DifferentiationTest=Proba` setting. Finally,

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

the level of gene diversity can be used as a test statistic when coupled with the `GeneDivRanks` setting (this was new to version 4.1; see Section 5.3.4).

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

5.3.2 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 on the genic table derived from the genotypic one (see Jérome 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.

5.3.3 Output

For the four sub-options, results are stored in a file named as follows:[7]

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.

5.3.4 Gene diversity as a test statistic

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

`DifferentiationTest=GeneDiv` makes Genepop use gene diversity as test

[7] slightly modified in comparison to earlier versions of Genepop

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

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

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

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.

5.3.5 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.[8] 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:

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

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`.

5.3.6 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 and Patel 1983) or by hand calculations. The example file `MehtaP83.txt` contains one such test sample.

5.4 Option 4: private alleles

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

5.5 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
```

5.5.1 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 B. S. Weir and Cockerham (1984).
- global estimate of F_{IS} over alleles according to B. S. Weir and Cockerham (1984) (W&C) and Robertson and Hill (1984) (R&H).
- observed and "expected" numbers 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.

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

This option takes the observed frequencies of identical pairs of genes as estimates (Q) of corresponding probabilities of identity (Q) and then simply computes diversities as $1 - Q$: 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 B. S. Weir and 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`.

5.5.3 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 7.6.2) 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`.

5.6 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
               For all population pairs ....... 4

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

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

Table 5.3: Genetic distance statistics available in options 6.5 and 6.6

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)	$F_{\mathrm{ST}}/(1\text{-}F_{\mathrm{ST}})$
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)	$F_{\mathrm{ST}}/(1\text{-}F_{\mathrm{ST}})$
Haploid	No (option 6.6)	`=singleGeneDiv`	$F/(1-F)$ variant with denominator common to all pairs

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. Table 5.3 summarizes the choice of methods, each of which will now be detailed.

5.6.1 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; B. S. Weir and

Cockerham 1984), and the analogous measure of correlation in allele size (ρ_{ST}) is estimated by the same technique (see Section 7.6.2). Multilocus estimates are computed as detailed in Section 7.6.1). 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.

5.6.2 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), Raphael Leblois, Estoup, and Rousset (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.

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 $x = 0.0$, $y = 15.0$ (showing that the decimal separator is a period), the second at the point $x = 0$, $y = 30$, etc. This example also shows that *individual identifiers can be added after these coordinates.*

```
Title line: A really too small data set
ADH Locus 1
ADH #2
ADH three
ADH-4
ADH-5
Pop
0.0 15.0,  0201 0303 0102 0302 1011
Pop
0 30 Second indiv,  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 (Raphael Leblois, Estoup, and Rousset 2003; Rousset 2007), so such an analysis has been purposely disabled.

Pairwise statistics for diploid data: Watts et al. (2007) contrasted two pairwise genetic distance statistics, $\hat{a}$ and $\hat{e}$. Using $\hat{e}$ is practically equivalent to using Loiselle's statistic (Loiselle et al. 1995), which has previously been advocated by e.g. Vekemans & Hardy (2004). Genepop actually uses a statistic e_r that handles missing data differently from $\hat{e}$ (see Methods) but the following discussion holds for both.

The pairwise statistic is selected by the setting `IsolationStatistic=a` or `=e`, or at runtime (in batch mode, the default is $\hat{a}$). $\hat{e}$ 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, Estoup, and Rousset 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 $\hat{a}$.

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`.

Nonparametric bootstrap is used in particular to obtain confidence intervals (DiCiccio and Efron 1996). The default method is the `ABC` bootstrap, but this can be changed by the setting `bootstrapMethod` to `BC` or `BCa` method. The number of computations of regression estimates scales as the number of loci for the ABC method, and as a chosen number of bootstrap resamples for the BC method (which is controlled by the `BootstrapNsim` setting, with default 999). The latter may thus be useful when the data include thousands of loci. The BCa method differs from the BC one by an additional step that scales as the number of loci.

Coverage probability of confidence interval This is the target probability that the confidence interval contains the parameter value. The usage is to compute intervals with 95% coverage and equal 2.5% tails, 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 and maximal geographic distances: 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 strictly 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 of the data. Thus, the suggested minimal distance here is any distance large enough to exclude only pairs at zero geographical distance. Negative values are thus not recommended (and rejected in 2D), and the default in batch mode is 0.0001.

There is also a setting `MaximalDistance=`*value*. This should not be abused, and is (therefore) available only through the settings file, not as a runtime option.

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

non-parametric bootstrap. 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. See Section 7.8 for limitations of this test. In the present context this is an exact test of the null hypothesis that there in no spatial correlation between genetic samples.

Up to version 4.3 Genepop implemented only a Mantel test based on the rank correlation. It now also implements, and performs by default, Mantel tests based on the regression coefficient for the "genetic distance" statistic used to quantify isolation by distance. The latter tests should generally be more congruent with the confidence intervals based on the same distances than the rank-based tests are. The rank test can now be performed by using the setting `MantelRankTest=` (no *value* needed).

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 bootstrap method is only approximate, but such exceptions appear to be rare. Exceptions may more commonly occur when the bootstrap is based on the regression of genetic "distance" and geographic distance over a selected range of the latter.

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; (v) a bootstrap interval for the intercept. 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.

5.6.3 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 $F_{\mathrm{ST}}/(1\text{-}F_{\mathrm{ST}})^*$ 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 5.3). 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 $F/(1-F) = (Q_0 - Q_r)/(1-Q_0)$ 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.

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.

5.6.4 Former sub-option 5 of Genepop: analysis of isolation by distance from a genetic distance matrix

That option (using the Isolde program) allowed one 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[9] (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

[9] more extensively than in earlier versions of Genepop.

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
```

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.

Genepop IsolationFile=*input file name* MantelRankTest= will further replicate the rank test of the old Isolde program.

5.6.5 User-provided geographic distance matrices

The setting `geoDistFile=`*file name*[10] can be used to provide a geographic distance matrix. Its format is that of other geographic distances matrices, with one required line of comment:

```
Geographic distances:                    <---anything (comments)
 21
 31 32
 41 42 43
 ...
```

The number of samples does not need to be given.

5.6.6 Analysis of isolation by distance from multiple genetic distance matrices

If another program has 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 show
...
 locus 16:                               <---anything (comments)
...
 Geographic distances:                   <---anything (comments)
  158.0
  158.0 1215.0
```

[10] New to Genepop 4.2

```
 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.

5.6.7 Analysis of mean differentiation

It is possible to perform a bootstrap analysis of the mean pairwise differentiation, through all menu options that lead to bootstrap analyses of isolation by distance, when additionally using the setting `MeanDifferentiationTest=TRUE`. It takes into account selection of data by both `PopTypes` and range of geographical distances.

5.7 Data selection for analyses of isolation by distance

5.7.1 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. `inter_all_types` excludes all pairs within types; no Mantel test is performed in that case. `intra_all_types` keeps all pairs within types, and performs a single regression for all types; again, no Mantel test is performed in that case.

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.

5.8 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 J. 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 and 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 and Dillmann (1992) and based on Black and Krafsur (1985). This program performs pairwise linkage disequilibria analyses in subdivided populations and Ohta (1982)'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 and Dillmann (1992) should be cited whenever this program is used.

5.9 Option 8: Null alleles and some input file utilities

The following menu appears[11]

```
Miscellaneous :
   Null allele: estimates of allele frequencies ..... 1
   Diploidisation of haploid data ..................... 2
   Relabeling alleles ................................. 3
   Conversion to individual data with population names ... 4
   Conversion to individual data with individual names ... 5
   Random sampling of haploid genotypes from diploid ones  6

   Main Menu   ........................................ 7
```

5.9.1 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 7.1.[12]

Genepop takes the allele with the highest number for a given locus **across all populations** as the null allele.[13] 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, Laird, and Rubin (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

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

[12] The last two methods are new to Genepop 4.0.

[13] This is a notable difference from Genepop 3.4, 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.

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.

- 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 bootstrap 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).

The confidence intervals for null allele frequencies are obtained by a bootstrap method, and are **not suitable** for testing for the presence of null alleles, because the null hypothesis is at the boundary of the parameter space (Andrews 2000). Instead, the exact score test for Hardy-Weinberg proportions can be used.

5.9.2 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*.[14]

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).

5.9.3 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.

5.9.4 Sub-options 4 and 5: Conversion of population data to individual data

These sub-options convert "population" data (with several individuals per `Pop` to "individual" data where each individual is put in a distinct `Pop`. This

[14] No longer truncated to 8 letters as it was in earlier versions of Genepop

is useful for individual-based analyses of isolation by distance and, in this perspective, the name of each individual is replaced by what should be its coordinates, that is, either the name of the last individual in the original population (sub-option 4), or the name of each individual if their locations are distinguished (sub-option 5)[15].

5.9.5 Sub-option 6: Random sampling of haploid genotypes from diploid ones

This sub-option randomly samples haploid genotypes at diploid loci.[16] This may be useful for external analyses that require haploid data or that would be biased by Hardy-Weinberg disequilibria.

[15]New to Genepop 4.3
[16]New to Genepop 4.3

Chapter 6

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`

analysis	*value*
$\hat{e}$, 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, in which case Genepop will carry all further computations in `Default` mode.

Chapter 7

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.

7.1 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, Laird, and Rubin 1977; Hartl and Clark 1989; Kalinowski and 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 and Taper 2006). This analysis is possible when at least three alleles are obviously present.

7.2 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 and Hinkley (1974) and Lehmann (1994) for the underlying theory; a much more elementary introduction to the tests implemented in Genepop is Rousset and Raymond (1997).

7.3 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 and 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 and 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.

7.4 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 and 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).

7.5 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 Lehmann (1994). Efficient tests for detecting heterozygote excesses and deficits (Rousset and 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 (Jérome 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.

Global tests are performed either using methods tuned to specific alternative hypotheses (for heterozygote excess or deficiency) or using Fisher's combination of probabilities technique. While the latter has been criticized (Whitlock 2005), the recommended alternative can fail spectacularly on discrete data.

7.6 Estimating F-statistics and related quantities

The definition of F-statistics used here is

$$F_{\mathrm{IS}} \equiv \frac{Q_1 - Q_2}{1 - Q_2}$$
$$F_{\mathrm{ST}} \equiv \frac{Q_2 - Q_3}{1 - Q_3}$$
$$F_{\mathrm{IT}} \equiv \frac{Q_1 - Q_3}{1 - Q_3}$$

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 and 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. Raphael Leblois, Estoup, and Rousset 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.

7.6.1 ANOVA estimators: single- and multilocus definitions

Well-known work by Cockerham (e.g. Cockerham 1973; B. S. Weir and 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. Bruce S. 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 $\hat{Q}_2 - \hat{Q}_3 =$

$(MSP - MSI)/(2n) \equiv \hat{\sigma}_P^2$ where n is group size. Hence the single-group (single-population) F_{IS} estimator is

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

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 $MSG \equiv \hat{\sigma}_G^2$ and $(MSI - MSG)/2 \equiv \hat{\sigma}_I^2$, and $(MSP - MSI)/(2n_c) \equiv \hat{\sigma}_P^2$ 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

$$\begin{aligned}
\hat{F}_{\mathrm{IS}} &= \frac{MSI - MSG}{MSI + MSG} = \frac{\hat{\sigma}_I^2}{\hat{\sigma}_I^2 + \hat{\sigma}_G^2}, \\
\hat{F}_{\mathrm{ST}} &= \frac{MSP - MSI}{MSP + (n_c - 1)MSI + n_c MSG} = \frac{\hat{\sigma}_P^2}{\hat{\sigma}_P^2 + \hat{\sigma}_I^2 + \hat{\sigma}_G^2}, \\
\hat{F}_{\mathrm{IT}} &= \frac{MSP + (n_c - 1)MSI - n_c MSG}{MSP + (n_c - 1)MSI + n_c MSG} = \frac{\hat{\sigma}_P^2 + \hat{\sigma}_I^2}{\hat{\sigma}_P^2 + \hat{\sigma}_I^2 + \hat{\sigma}_G^2}.
\end{aligned}$$

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. B. S. Weir and Cockerham (1984)'s multilocus estimators are defined in terms 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 (Bruce S. Weir 1996's) 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 $F_{\rm ST}$ 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:

$$\hat{F}_{\rm IS}=\frac{\sum_i[n_c(MSI-MSG)]_i}{\sum_i[n_c(MSI+MSG)]_i}=\frac{\sum_i[n_c\hat{\sigma}_I^2]_i}{\sum_i[n_c\hat{\sigma}_I^2+n_c\hat{\sigma}_G^2]_i},$$
$$\hat{F}_{\rm ST}=\frac{\sum_i[MSP-MSI]_i}{\sum_i[MSP+(n_c-1)MSI+n_cMSG]_i}=\frac{\sum_i[n_c\hat{\sigma}_P^2]_i}{\sum_i[n_c\hat{\sigma}_P^2+n_c\hat{\sigma}_I^2+n_c\hat{\sigma}_G^2]_i},$$
$$\hat{F}_{\rm IT}=\frac{\sum_i[MSP+(n_c-1)MSI-n_cMSG]_i}{\sum_i[MSP+(n_c-1)MSI+n_cMSG]_i}=\frac{\sum_i[n_c\hat{\sigma}_P^2+n_c\hat{\sigma}_I^2]_i}{\sum_i[n_c\hat{\sigma}_P^2+n_c\hat{\sigma}_I^2+n_c\hat{\sigma}_G^2]_i}.$$

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

Estimate	$F_{\rm IS}$	$F_{\rm ST}$	$F_{\rm IT}$
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 $F_{\rm ST}$ estimates under weak differentiation.

7.6.2 Microsatellite allele sizes, $R_{\rm ST}$, and $\rho_{\rm ST}$

Following Slatkin (1995), statistics based on allele size have been widely used. The parameters $\rho_{\rm IS}$, $\rho_{\rm ST}$ and $\rho_{\rm IT}$ 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

and Excoffier 1996). Then the estimators become plain ANOVA estimators of intraclass correlation for allele size; if there are only two alleles, $\hat{\rho}_{\mathrm{ST}} = \hat{F}_{\mathrm{ST}}$, but Slatkin's $R_{\mathrm{ST}} \neq \hat{F}_{\mathrm{ST}}$.

7.6.3 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 that deviations from Hardy-Weinberg proportions are characterized by the same F_{IS} for all pairs of alleles (Robertson and Hill 1984). The score test computed in heterozygote excess and deficiency sub-options of option 1 is equivalent to this estimator for testing purposes.

7.7 Bootstraps

Option 6 constructs approximate bootstrap confidence intervals (DiCiccio and Efron 1996), assuming that each locus is an independent realization of genealogical and mutation processes. The bootstrap is a general methodology with different incarnations: ABC, BC and BCa variants are implemented for this option. The default bootstrap method, ABC, was chosen for typical microsatellite data sets because it balances moderate computation needs (for small number of loci) 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 Raphael Leblois, Estoup, and Rousset (2003) and Watts et al. (2007).

For SNP data sets of thousands of loci, the ABC method can become very slow and the alternative BC bootstrap method may be useful. BC is the bias-corrected percentile method discussed in the early bootstrap literature (Efron 1987) and superseded by the BCa method which is more accurate for small samples. However the BCa method (also implemented) will again be slow for large number of loci, while the BC may be both reasonably accurate and reasonably fast in that case.

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

7.8 Mantel test

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 no such 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 correlation is the sum of these two P-values minus 1.

7.8.1 Misuse 1: tests of correlation at different distance

Genetic processes of isolation by distance generate asymptotically decreasing variation in genetic differentiation with increasing geographic distances, and there is some temptation to use the Mantel test to test for the presence of correlation at specific distances. However, Genepop prevents this as this is logically unsound, and the more quantitative methods it provides are better suited to address variation of patterns with distance.

As soon as a process generates data with an expected non-zero correlation at some distance, it contradicts the null hypothesis under which the Mantel test is an exact test. Thus it may not make sense to use a Mantel test for testing correlation at some distance if there is correlation at another distance.

One can still wonder whether a permutation-based test could have some approximate validity for testing absence of correlation at some distance. However, the bootstrap procedure already addresses this case. Alternative procedures would require further definition on an ad-hoc basis to be operational (e.g., the idea of eliminating *samples* that form *pairs* below or above a given distance may not unambiguously define a *sample* selection procedure that will retain power) and would be likely to generate some confusion.

For these reasons, in the present implementation the Mantel tests are always based on all pairs, ignoring all selection of data according to distance.

7.8.2 Misuse 2: partial Mantel tests

Partial Mantel tests have been used to test for effects of a variable Y on a response variable Z, while supposedly removing spatial autocorrelation effects on Z. Both standard theory of exact tests (as used by Raufaste and Rousset 2001) and simulation (Oden and Sokal 1992; Raufaste and Rousset 2001; Rousset 2002b; Guillot and Rousset 2013) show that the permutation procedure of the Mantel test is not appropriate for the partial Mantel test when the Y variable itself presents spatial correlations. Asymptotic arguments have also been proposed to support the use of such permutation tests (e.g. Anderson 2001) but they fail in the same conditions. As shown by Raufaste and Rousset (2001), the problem is inherent to the permutation procedure, not to a specific test statistic. Unfortunately, some papers maintain confusion about these different aspects of "partial Mantel tests". Legendre and Fortin (2010) argued how miserable the papers by Raufaste and Rousset (2001) and Rousset (2002b) were, and claimed that some versions of the tests should be preferred because they used pivotal statistics (without evidence that the statistics were indeed pivotal, a property that depends on the statistical model). Guillot and Rousset (2013) reviewed old and more recent literature demonstrating issues with the partial Mantel test, provided new simulations showing that the different tests discussed by Legendre and Fortin (2010) failed, and criticized their verbal arguments. Despite this, Legendre, Fortin, and Borcard (2015) criticized this more recent paper again for ignoring the old literature, and repeated the same kind of verbal explanations that have previously failed.

Chapter 8

Code maintenance, credits, contact, etc.

8.1 Code maintenance

Distribution of Genepop as an R package means that the code is portable to the major operating systems supported by R. New version are checked using a variety of tools available in the R environment (including valgrind and so-called sanitizers). Tests against more or less standard examples from the literature are also applied. These tests can be found in the `tests/testthat` directory of the distributed archive.

8.2 Credits for the current version

The R package and the R markdown version of the documentation were originally developed by Jimmy Lopez (Labex Cemeb) and Khalid Belkhir (Institut des Sciences de l'Évolution) from the C++ sources and LaTeX documentation of the Genepop executable version 4.6, and further modified by F. Rousset.

8.3 Previous history

Version 4.0 of Genepop was a C++ rewrite of Genepop 3.4 (Raymond and 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.

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.

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.

T. Antão, E. Archer, R.I. Bailey, J.S.F. Barker, D. Bourguet, T. Devitt, É. Imbert, R. Leblois, T. de Meeüs, P. Morin, S. Ponsard, V. Ravigné, E. Taschen, and Y. Zimmermann have pointed issues or have stimulated additional developments of more recent versions.

8.4 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 this 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 this documentation. Please read this documentation.

8.4.1 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.

See the distribution page for more recent bug fixes.

Chapter 9

Copyright

Index

Bibliography

Anderson, Marti J. 2001. "Permutation Tests for Univariate or Multivariate Analysis of Variance and Regression." *Can. J. Fish Aquatic Scis.* 58: 626–39.

Andrews, Donald W. K. 2000. "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space." *Econometrica* 68 (2): 399–405.

Black, W. C., IV, and E. S. Krafsur. 1985. "A FORTRAN Program for the Calculation and Analysis of Two-Locus Linkage Disequilibrium Coefficients." *Theor. Appl. Genetics* 70: 491–96.

Brookfield, J. F. Y. 1996. "A Simple New Method for Estimating Null Allele Frequency from Heterozygote Deficiency." *Mol. Ecol* 5: 453–55.

Brooks, Stephen, and Andrew Gelman. 1998. "General Methods for Monitoring Convergence of Iterative Simulations." *J. Comput. Graphical Statistics* 7: 434–55.

Chakraborty, R., M. de Andrade, S. P. Daiger, and B. Budowle. 1992. "Apparent Heterozygote Deficiencies Observed in DNA Typing Data and Their Implications in Forensic Applications." *Ann. Hum. Genetics* 56: 45–57.

Cockerham, C. Clark. 1973. "Analyses of Gene Frequencies." *Genetics* 74: 679–700.

Cockerham, C. Clark, and Bruce S. Weir. 1987. "Correlations, Descent Measures: Drift with Migration and Mutation." *PNAS* 84: 8512–14.

Cox, D. R., and D. V. Hinkley. 1974. *Theoretical Statistics*. London: Chapman & Hall.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the *EM* Algorithm (with Discussion)." *JRSSB* 39: 1–38.

DiCiccio, Thomas J., and Bradley Efron. 1996. "Bootstrap Confidence

Intervals (with Discussion)." *Stat. Sci.* 11: 189–228.
Efron, Bradley. 1987. "Better Bootstrap Confidence Intervals." *Journal of the American Statistical Association* 82: 171–85.
Felsenstein, J. 2005. "PHYLIP (Phylogeny Inference Package) Version 3.6."
Fisher, Ronald Aylmer. 1935. "The Logic of Inductive Inference (with Discussion)." *JRSS* 98: 39–82.
Garnier-Géré, P., and C. Dillmann. 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: 485–86.
Goudet, Jérome, Michel Raymond, Thierry de Meeüs, and François Rousset. 1996. "Testing Differentiation in Diploid Populations." *Genetics* 144: 1931–38.
Guillot, G., and François Rousset. 2013. "Dismantling the Mantel Tests." *Methods Ecol. Evol.* 4: 336–44.
Guo, Sun Wei, and Elizabeth A. Thompson. 1992. "Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles." *Biometrics* 48: 361–72.
Haldane, John Burdon Sanderson. 1954. "An Exact Test for Randomness of Mating." *Journal of Genetics* 52: 631–35.
Hartl, Daniel L., and Andrew G. Clark. 1989. *Principles of Population Genetics*. Second. Sunderland, Mass.: Sinauer.
Hastings, W. K. 1970. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika* 57: 97–109.
Kalinowski, Steven T., and Mark L. Taper. 2006. "Maximum Likelihood Estimation of the Frequency of Null Alleles at Microsatellite Loci." *Conserv. Genetics* 7: 991–95.
Leblois, Raphael, Arnaud Estoup, and François Rousset. 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.
Leblois, R., P. Pudlo, J. Néron, F. Bertaux, C. R. Beeravolu, R. Vitalis, and F Rousset. 2014. "Maximum Likelihood Inference of Population Size Contractions from Microsatellite Data." *Mol. Biol. Evol.* 31: 2805–23.
Legendre, Pierre, and Marie-Josée Fortin. 2010. "Comparison of the Mantel Test and Alternative Approaches for Detecting Complex Multivariate Relationships in the Spatial Analysis of Genetic Data." *Mol. Ecol. Resources*

10 (5): 831–44.

Legendre, Pierre, Marie-Josée Fortin, and Daniel Borcard. 2015. "Should the Mantel Test Be Used in Spatial Analysis?" *Methods Ecol. Evol.* 6: 1239–47.

Lehmann, E. L. 1994. *Testing Statistical Hypotheses*. Second. New York: Chapman & Hall.

Levene, Howard. 1949. "On a Matching Problem Arising in Genetics." *Annals of Mathematical Statistics* 20: 91–94.

Loiselle, Bette A., Victoria L. Sork, John Nason, and Catherine Graham. 1995. "Spatial Genetic Structure of a Tropical Understory Shrub *Psychotria Officinalis* (Rubiaceae)." *Am. J. Bot.* 82: 1420–25.

Louis, Edward J., and Everett R. Dempster. 1987. "An Exact Test for Hardy-Weinberg and Multiple Alleles." *Biometrics* 43: 805–11.

Mantel, Nathan. 1967. "The Detection of Disease Clustering and a Generalized Regression Approach." *Cancer Research* 27: 209–20.

Mehta, Cyrus R., and Nitin R. Patel. 1983. "A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables." *JASA* 78: 427–34.

Michalakis, Yannis, and Laurent Excoffier. 1996. "A Generic Estimation of Population Subdivision Using Distances Between Alleles with Special Interest to Microsatellite Loci." *Genetics* 142: 1061–64.

Oden, Neal L., and Robert R. Sokal. 1992. "An Investigation of Three-Matrix Permutation Tests." *J. Classif.* 9: 275–90.

Ohta, Tomoko. 1982. "Linkage Disequilibrium Due to Random Genetic Drift in Finite Subdivided Populations." *PNAS* 79: 1940–44.

Raufaste, Nathalie, and François Rousset. 2001. "Are Partial Mantel Tests Adequate?" *Evolution* 55: 1703–5.

Raymond, Michel, and François Rousset. 1995a. "An Exact Test for Population Differentiation." *Evolution* 49: 1283–86.

———. 1995b. "GENEPOP Version 1.2: Population Genetics Software for Exact Tests and Ecumenicism." *J. Hered.* 86: 248–49.

Robertson, Alan, and William G. Hill. 1984. "Deviations from Hardy-Weinberg Proportions: Sampling Variances and Use in Estimation of Inbreeding Coefficients." *Genetics* 107: 703–18.

Rousset, François. 1996. "Equilibrium Values of Measures of Population Subdivision for Stepwise Mutation Processes." *Genetics* 142: 1357–62.

———. 1997. "Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance." *Genetics* 145: 1219–28.

———. 1999. "Genetic Differentiation Within and Between Two Habitats." *Genetics* 151: 397–407.

———. 2000. "Genetic Differentiation Between Individuals." *J. Evol. Biol.* 13: 58–62.

———. 2002a. "Inbreeding and Relatedness Coefficients: What Do They Measure?" *Heredity* 88: 371–80.

———. 2002b. "Partial Mantel Tests: Reply to Castellano and Balletto." *Evolution* 56: 1874–75.

———. 2007. "Inferences from Spatial Population Genetics." In *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop, and C. Cannings, third, 945–79. Chichester, U.K.: Wiley.

Rousset, François, and Raphaël Leblois. 2007. "Likelihood and Approximate Likelihood Analyses of Genetic Structure in a Linear Habitat: Performance and Robustness to Model Mis-Specification." *Mol. Biol. Evol.* 24: 2730–45.

———. 2012. "Likelihood-Based Inferences Under Isolation by Distance: Two-Dimensional Habitats and Confidence Intervals." *Mol. Biol. Evol.* 29: 957–73.

Rousset, François, and Michel Raymond. 1995. "Testing Heterozygote Excess and Deficiency." *Genetics* 140: 1413–19.

———. 1997. "Statistical Analyses of Population Genetic Data: Old Tools, New Concepts." *Tr. Ecol. Evol.* 12: 313–17.

Slatkin, Montgomery. 1995. "A Measure of Population Subdivision Based on Microsatellite Allele Frequencies." *Genetics* 139: 457–62.

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

Watts, Phillip C., François Rousset, Ilik J. Saccheri, Raphaël Leblois, Stephen J. Kemp, and David J. Thompson. 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: 737–51.

Weir, B. S., and C. Clark Cockerham. 1984. "Estimating F-Statistics for the Analysis of Population Structure." *Evolution* 38: 1358–70.

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

Whitlock, M. C. 2005. "Combining Probability from Independent Tests: The Weighted z-Method Is Superior to Fisher's Approach." *J. Evol. Biol.* 18

(5): 1368–73.