

# *SPAGeDi* 1.5

a program for Spatial *P*attern Analysis of *G*enetic *D*iversity

by Olivier J. HARDY and Xavier VEKEMANS

with the contribution of Reed A. CARTWRIGHT

Copyright (c) 2002-2015 Olivier Hardy and Xavier Vekemans

---

## **User's manual**

---

Address for correspondence:

Evolutionary Biology and Ecology Unit, CP160/12

Université Libre de Bruxelles

50 Av. F. Roosevelt

B-1050 Brussels, Belgium

e-mail: [ohardy@ulb.ac.be](mailto:ohardy@ulb.ac.be)

Last update: 03 March 2015

# Contents

## 1. Note about *SPAGeDi* 1.5 and installation

## 2. What is *SPAGeDi* ?

- 2.1. Purpose
- 2.2. How to use *SPAGeDi* – short overview
- 2.3. Data treated by *SPAGeDi*
- 2.4. Three ways to specify populations
- 2.5. Statistics computed

## 3. Creating a data file

- 3.1. Structure of the data file
- 3.2. How to code genotypes?
- 3.3. Example of data file
- 3.4. Note about distance intervals
- 3.5. Note about spatial groups
- 3.6. Note about microsatellite allele sizes
- 3.7. Using a matrix to define pairwise spatial distances
- 3.8. Defining genetic distances between alleles
- 3.9. Defining reference allele frequencies for relatedness coefficients
- 3.10. Present data size limitations

## 4. Running the program

- 4.1. Launching the program
- 4.2. Specifying the data / results files
- 4.3. Selecting the appropriate options
- 4.4. Information displayed during computations

## 5. Interpreting the results file

- 5.1. Basic information
- 5.2. Allele frequency analysis
- 5.3. Identity disequilibrium and selfing rate estimates
- 5.4. Type of analyses
- 5.5. Distance intervals
- 5.6. Computed statistics
- 5.7. Permutation tests
- 5.8. Matrices of pairwise coefficients/distances

## 6. Technical notes

- 6.1. Statistics for individual level analyses
- 6.2. Statistics for population level analyses
- 6.3. Inference of gene dispersal distances
- 6.4. Estimating the actual variance of pairwise coefficients for marker based heritability and  $Q_{ST}$  estimates
- 6.5. Testing phylogeographic patterns

## 7. References

## 8. Bug reports

# 1. NOTE ABOUT *SPAGeDi* 1.5 AND INSTALLATION

Copyright (c) 2002-2015 Olivier Hardy and Xavier Vekemans

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details (<http://www.gnu.org/licenses/>).

*SPAGeDi* has been tested on several data sets and results were checked for consistency with alternative softwares whenever possible. It may nevertheless still contain bugs (corrected bugs are listed at the end of this manual). Some of these bugs are probably easy to detect by causing the program to crash or leading to obvious erroneous results for particular data sets and analyses. But others, more critical, may just cause biased results that appear plausible. Hence, it is advised to take much care checking the consistency of the information from the results file. The authors would appreciate being informed of any detected bug. The authors claim no responsibility if or whenever a bug causes a misinterpretation of the results given by *SPAGeDi*.

## What's new in *SPAGeDi* ?

Implementations in **version 1.5:**

- 1°) *SPAGeDi* 1.5 can estimate the **standardized identity disequilibrium** between loci, and thereafter the **selfing rate** of hermaphrodites, using a population sample of adults genotyped at codominant markers. It is based on the approach of David et al. (2007), extended to autopolyploids by Hardy (2015).
- 2°) *SPAGeDi* 1.5 can in principle accept datasets **up to 100 000 loci** (1000 loci when importing a dataset from Genpop or FSTAT format), although computer RAM can remain a limitation for performing the computations.

Implementations in **version 1.4:**

- 1°) *SPAGeDi* 1.4 can compute a **new estimator of pairwise relatedness between individuals** using **AFLP** or **RAPD** markers when phenotypes (presence or absence of a band) are **affected by non genetic factors**. To this end, the broad-sense heritability ( $H^2$ ) of each marker must first be estimated using replicated individuals. These  $H^2$  estimates are provided at the end of the data file and used to compute corrected estimates of pairwise relatedness coefficients. Details about the statistics used can be found in Ley & Hardy (2013) who showed that  $H^2$  is an appropriate measure of the reliability of AFLP markers and that  $H^2$  can be much lower than one (the value expected for a “perfect” dominant marker) for the majority of AFLP markers. The way to specify the  $H^2$  of markers in the data file is explained in § 3.1.
- 2°) *SPAGeDi* 1.4 can compute for each population and locus the **effective number of alleles** (Nielsen et al. 2003), the **rarefied allelic richness** for which the subsample size can be specified, and the **inbreeding coefficient with a test significance**.
- 3°) *SPAGeDi* 1.4 can test **differentiation statistics for each pair of populations**.

Implementations in **version 1.3:**

- 1°) *SPAGeDi* 1.3 can be compiled for different platforms including **Windows, Mac OS X, and Linux** (thanks to Reed Cartwright!).
- 2°) The iterative **procedure to estimate gene dispersal parameters** has been improved.
- 3°) A **new statistic** ( $N_{ij}$ ) characterize similarity between individuals using “**ordered alleles**”. It is an analogue of kinship coefficient considering the phylogenetic distance between alleles (or haplotypes). Permutation tests permit to assess whether the allele phylogeny contributes to the genetic structure, providing a **test of phylogeographic patterns at the individual level**.

4°) Spatial coordinates can now be given as latitudes and longitudes in degrees with decimal (using negative numbers for Southern latitudes or Western longitudes). To this end, the number of spatial coordinates (3<sup>rd</sup> number of the first line) must be set to -2.

#### Implementations in **version 1.2:**

- 1°) *SPAGeDi* 1.2 proposes **new statistics** (e.g.  $N_{ST}$ ) to characterize differentiation among populations **using “ordered alleles”**, i.e. considering the phylogenetic distance between alleles (or haplotypes), as proposed by Pons & Petit (1996). Permutation tests permit to assess whether the allele phylogeny contributes to the differentiation pattern, which can be used to **test phylogeographic patterns**.
- 2°) *SPAGeDi* 1.2 proposes an **estimator of the mean kinship coefficient** between populations ( $G_{ij}$ ) closely related to the autocorrelation of population allele frequencies (Barbujani 1987).
- 3°) *SPAGeDi* 1.2 proposes a **new estimator of the relationship coefficient** between individuals (Li et al. 1993).
- 4°) *SPAGeDi* 1.2 can use specific **reference allele frequencies** (to specify in a file) to compute relatedness coefficients between individuals.
- 5°) *SPAGeDi* 1.2 includes an iterative **procedure to estimate gene dispersal parameters** from isolation-by-distance patterns by regressing pairwise kinship coefficients on distance over a restricted distance range (this requires an estimate of the effective population density).
- 6°) *SPAGeDi* 1.2 provides better **error messages**. The most common data file errors are systematically listed in a file called “error.txt” when launching the program. As far as possible, error messages when problems occur were improved. These messages are not yet optimal so that suggestions to improve them are welcome. **Empty lines** in data files are now **allowed**. Problems when entering instructions with the keyboard under Windows 2000 and latter versions have been solved.

#### Implementations in **version 1.1:**

- 1°) *SPAGeDi* 1.1 can treat data from **dominant** genetic markers such as AFLP or RAPD to compute pairwise **relatedness coefficients between individuals**. Details about the statistics used can be found in Hardy (2003). The way to code phenotypes of dominant markers in the data file is explained in § 3.2.2.
- 2°) *SPAGeDi* 1.1 proposes an **allele size permutation test** indicating whether microsatellite allele sizes are informative with respect to genetic differentiation. Details about this test and its applications are given in Hardy et al. (2003).

## Installation

For Microsoft Windows users, an executable (*SPAGeDi.exe*) can be downloaded at <http://ebe.ulb.ac.be/ebe/Software.html> and used directly without installation, as for the previous versions.

Native installers for Windows, Mac OS X, and Linux created by Reed A. Cartwright <reed@scit.us> are also available at <http://scit.us/spagedi/>. These native binaries have not been fully tested and may not work on all versions of their operating systems. If they fail to run properly, you can always compile the source packages. If the Windows version of *SPAGeDi* fails to work, you might also need to install the Microsoft Visual C++ 2008 SP1 Redistributable Package (<http://www.microsoft.com/downloads/details.aspx?familyid=A5C84275-3B97-4AB7-A40D-3802B2AF5FC2&displaylang=en>).

For all users, a **portable source code** (Unix/Windows/OS X port) has been created by Reed A. Cartwright <reed@scit.us>. Compiling *SPAGeDi* from source requires CMake (<http://www.cmake.org/>). Windows, Mac OS X and most Unix-like operating systems can install CMake through their package systems, otherwise it can be compiled from its source code. Once CMake is installed, download *SPAGeDi*'s source code.

SPAGeDi-1.5a.tar.gz: Unix and Mac OS X source (\n linefeeds)

SPAGeDi-1.5a.zip: Windows source (\r\n linefeeds)

Installation from source should be roughly the same on most systems. Detailed instructions for Linux follow. Open a terminal and issue the following commands, once you have downloaded *SPAGeDi*'s source code to your home directory.

```
tar xvzf SPAGeDi-1.5a.tar.gz
cd SPAGeDi-1.5a
cmake .
make
make install
```

A “-g” option can be passed to CMake to designate a different build system. This is most important when compiling *SPAGeDi* on Windows. See CMake’s documentation for more information.

## **How to cite *SPAGeDi*?**

Hardy, O. J. & X. Vekemans (2002). *SPAGeDi*: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* 2: 618-620.

## **Acknowledgments**

We would like to thank the *SPAGeDi* user’s who have identified bugs or have given us other feedback on the program, in particular Dave Coltman, Britta Denise Hardesty, Myriam Heuertz, Xavier Turon, Mine Turktas, Peter Wandeler, Reed Cartwright, Kevin Weitemier, Mariano Giombini, Walter Durka.

## 2. WHAT IS *SPAGeDi* ?

### 2.1. PURPOSE

*SPAGeDi* is primarily designed to characterise the spatial genetic structure of mapped individuals and/or mapped populations using genotype data (e.g. isozymes, RFLP, microsatellites) of any ploidy level. For polyploids, analyses assume polysomic inheritance as in autopolyploids. Polyploids with disomic inheritance (allopolyploids) can be treated correctly only if alleles from different homeologous genomes can be distinguished so that genotypes are treated as diploid data. *SPAGeDi* can compute inbreeding coefficients as well as various statistics describing relatedness or differentiation between individuals or populations by pairwise comparisons. To analyse how values of pairwise comparisons are related to geographical distances, *SPAGeDi* computes 1°) average values for a set of predefined distance intervals, in a way similar to a spatial autocorrelation analysis, 2°) linear regressions of pairwise statistics on geographical distances (or their logarithm). The slopes of these regressions can potentially be used to obtain indirect estimates of gene dispersal distances parameters (e.g. neighbourhood size), and provide a synthetic measure of the strength of spatial structuring. *SPAGeDi* can also treat data without spatial information, providing global estimates of genetic differentiation and/or matrices of pairwise statistics between individuals or populations.

Different permutation procedures allow to test if there is significant inbreeding, identity disequilibrium between loci, population differentiation, spatial structure, or if microsatellite allele size or the phylogenetic distance between alleles carries relevant information about genetic structure.

Analyses can be carried out on data sets containing individuals with different ploidy levels, but not on data sets mixing loci corresponding to different ploidy levels within individuals (e.g. genotypes based nuclear and cytoplasmic DNA can not be analysed simultaneously, except for an haploid organism). Data from dominant markers (RAPD, AFLP) can be used to carry out analyses at the individual level with diploids (relatedness coefficients between individuals), and new estimators are available for markers showing a broad-sense heritability inferior to unity. Presently, there is no statistics adapted to dominant markers for analyses at the population level or for higher ploidy levels. One can always enter such data as haploid data (not mixed with data from codominant markers), but much caution must be taken in the interpretation of the results.

### 2.2. HOW TO USE *SPAGeDi* – SHORT OVERVIEW

*SPAGeDi* has no fancy windowing features. To launch the program just double click on the program icon and type the name of the data file or drag a data file icon into the program window. You can also launch the program with a command line (see §4.1). A **single data file** must contain all individual characteristics (name, category, spatial coordinate(s), genotypes). Details of the analyses to be carried out (individual *versus* population level, population definition, statistics, permutation tests, various options) will be specified after the program has been launched. Results of the analyses are written to a **single results file**. Data and results files are text files with tab delimited pieces of information. Hence they are best opened and edited using a worksheet software such as Excel. Data files can be converted from and into FSTAT and GENEPOP formats. Although **error messages** are displayed when problems occur, typically because the data file is not properly formatted, they may not be sufficient to find out the errors. Therefore we urge users to read carefully the instructions for preparing data files (next chapter).

### 2.3. DATA TREATED BY *SPAGeDi*

*SPAGeDi* requires that the following information is provided for each individual: 1°) one to three spatial coordinates on a Cartesian coordinate system or latitude and longitude in degrees (facultative), 2°) value of a categorical variable (facultative) and 3°) its genotype at each locus (missing data allowed). The categorical variable can be used to define populations or to restrict analyses within or among categories. The spatial coordinate(s) permit(s) *SPAGeDi* to compute pairwise distances between individuals or populations (Euclidian distances). Alternatively, pairwise distances can be defined in a separate matrix. To compute some statistics, a matrix giving phyletic distances between alleles can also be provided.

## 2.4. THREE WAYS TO SPECIFY POPULATIONS

Populations can be defined in three different ways: 1°) as categorical groups, where one population includes all individuals sharing the same categorical variable; 2°) as spatial groups, where a spatial group includes all individuals sharing the same spatial coordinates and following each other in the data file; 3°) as spatio-categorical groups, where a spatio-categorical group includes all individuals belonging to both the same spatial group and categorical group. When populations are defined using the categorical variable, the spatial coordinates of a given population are computed by averaging the coordinates of the individuals it contains.

## 2.5. STATISTICS COMPUTED

Statistics computed for each population and/or the whole sample include diversity metrics computed per locus and a multilocus average (number of alleles per locus, effective number of alleles, rarefied allelic richness, observed and expected heterozygosities, inbreeding coefficient) and a metric for pairs of loci (standardized identity coefficient).

Statistics for pairwise comparisons between **populations** include:

$F_{ST}$	a measure of population differentiation (intra-class kinship coefficient) (Weir & Cockerham 1984)
$G_{ST}$	equivalent to $F_{ST}$ but estimator with different statistical properties (Pons & Petit 1996)
$R_{ST}$	$F_{ST}$ analogue based on allele size (Slatkin 1995, estimated as Michalakis & Excoffier 1996)
$N_{ST}$	$F_{ST}$ analogue accounting for the genetic distances between alleles (Pons & Petit 1996)
$Rho$	intra-class relatedness coefficient permitting among ploidy comparisons (Ronfort et al. 1998)
$G_{ij}$	mean kinship coefficient between populations (Barbujani 1987)
$N_{ij}$	$G_{ij}$ analogue accounting for the genetic distances between alleles (OJ Hardy, unpubl)
$D_S$	Nei's 1978 standard genetic distance
$(\delta\mu)^2$	$D_S$ analogue based on allele size (Goldstein and Pollok 1997)

**Global  $F$ - or  $R$ - statistics** (inbreeding coefficients) are also provided.

Statistics designed for pairwise comparisons between **individuals** include

**Kinship coefficients,  $F_{ij}$ :** 3 estimators including one for dominant markers in diploids (Loiselle et al. 1995, Ritland 1996, Hardy 2003, Ley and Hardy 2012)

**Relationship coefficients:** 6 estimators including one for dominant markers in diploids (Hardy & Vekemans 1999, Lynch & Ritland 1999, Queller & Goodnight 1989, Wang 2002, Li et al. 1993, Hardy 2003)

**Fraternity coefficients:** 2 estimators (Lynch & Ritland 1999, Wang 2002)

**Rousset's distance** between individuals,  $A_{ij}$  (Rousset 2000)

**A kinship analogue based on allele size,  $R_{ij}$**  (I' in Streiff et al. 1998)

**A kinship analogue based on the genetic distances between alleles,  $N_{ij}$**  (OJ Hardy, unpubl)

**Inbreeding coefficients** (computed as kinship coefficients between gene copies within individuals)

All statistics are computed for each locus and a multilocus weighted average. Note that an estimate of the inbreeding coefficient must be entered to compute kinship or relationship coefficients with dominant markers in diploids.

The **actual variance** of these coefficients (i.e. the remaining variance when sampling variance has been removed) can be estimated following the method of Ritland (2000) for each distance interval defined. The actual variance of kinship (or relatedness) coefficients and of pairwise  $F_{ST}$  is necessary for *in situ*, genetic markers based inference of, respectively, the heritability and  $Q_{ST}$  of quantitative traits.

For pairwise coefficients, mean values per distance intervals and regression slopes on spatial distance are given (unless spatial information are lacking).

**Jackknifing loci** (i.e. deleting information from one locus at a time) provides approximate standard errors for the multilocus estimates.

**Permutation tests** whereby the statistics are computed again after that locations, individuals or gene copies are permuted provide ad hoc tests for spatial genetic structure, population differentiation or inbreeding coefficients, respectively. Note that permuting locations is equivalent to carrying out a Mantel test. Permutation of microsatellite allele sizes or of the phylogenetic distances between alleles also permit to test if the mutation rate is sufficient to affect the genetic structure (test of phylogeographic patterns) (Hardy et al. 2003; Pons & Petit 1996).

### 3. CREATING A DATA FILE

The data file is a text file. It is advised to create the data file using a worksheet program such as Excel and then save it as a “**tab delimited text file**”. If you do not have this option, try “DOS text” (“Text Unicode” or “ASCII” formats might not work).

#### 3.1 STRUCTURE OF THE DATA FILE

**Comments lines:** they are not read by the program and can be put anywhere in file. Comment lines must begin by the two characters `//`. Empty lines are allowed.

The data file must be in the following format, with each piece of information within a line being separated by a **tab** (i.e. each piece of information put in adjacent columns if using a worksheet program to generate the data file). Hereafter, first, second, third,... line refers to non-comment and non-empty lines.

- **first line:** 6 format numbers separated by a tab in the following order:
  - **number of individuals**
  - **number of categories** (0 if no category defined)
  - **number of spatial coordinates** in a Cartesian coordinate system (0 to 3) or put -2 for latitude + longitude
  - **number of loci** (or the number you wish to use if the data set contains more)
  - **number of digits** used to code one allele (1 to 3); **or** set a **value  $\leq 0$**  (in fact the value given for missing data) in the case of **dominant** markers (restricted to diploid data)
  - **ploidy** (2 = diploid; for data with several ploidy levels, give the largest)
- **second line:** definition of **distance intervals**:
  - number of distance intervals ( $n$ )
  - the  $n$  maximal distances corresponding to each interval

Note 1: alternatively you can enter only the desired number of intervals preceded by a **negative sign**; the program then defines the  $n$  maximal distances in such a way that the number of pairwise comparisons within each distance interval is approximately constant.

Note 2: if you do not wish distance intervals, put 0.

Note 3: if you use latitude + longitude, distance intervals must be given in km .
- **third line:** the names used as column **labels** (up to 15 characters long, without space):
  - a generic name for individuals (e.g. “Ind”)
  - a generic name for categories (e.g. “Cat”), only if categories are defined
  - a generic name for each spatial coordinates (e.g. “X”, “Y”)
  - the name of each locus (e.g. “Pgm”, “Est”, ...)
- **fourth line and next ones: individual data** (each line = 1 individual):
  - name of the individual (up to 15 characters)
  - name of the category (up to 15 characters), only if categories are defined
  - coordinate along each axis (up to 10 digits), or latitude followed by longitude
  - genotype at each locus (also separated by a tab)

Note : if the number of spatial coordinates is set to -2, latitudes and longitudes must be given in degrees with decimal, using negative numbers for Southern latitude or Western longitude.
- **line after the last individual:** the word “END” (in uppercase)
- **next line in the case of dominant markers (optional):** the word “H2” followed by the broad-sense heritability estimates of each dominant marker (separated by tabs; cf. Ley & Hardy 2013).
- **next line in the case of polyploid data (optional):** the word “ALPHA” followed by estimated rates of double reduction for each locus (separated by tabs; cf. Hardy 2015).

## 3.2. HOW TO CODE GENOTYPES ?

### 3.2.1. Codominant data

Single locus genotypes are represented by numbers in either of the following ways:

1°) the **allele** of each homologous gene is up to  $n$  digits long and alleles are **separated by** any number of **non numerical characters** other than a tab ( $n$  is specified in the first line): e.g.,

*12/45*                      *1 12*                      *99, 23*                      *6.6*                      *36--01*

are correct genotypes for a diploid with up to 2 digits per allele.

2°) the allele of each homologous gene is **exactly  $n$  digits long** and alleles are not separated by other characters: e.g.,

*1245*                      *0112*                      *9923*                      *0606*                      *3601*

are the same genotypes as above.

In both cases, non numerical characters cannot follow the righter most digit.

#### Notes:

1°) **missing genotypes** are represented by giving the value **0**: e.g.,

*0*                      *0 0*                      *000,000,000,000*                      *000*

all represent a missing genotype.

2°) **incomplete genotypes** are represented by giving the **value 0** to undetermined alleles **on the right**: e.g.,

*05-00*                      *05,0*                      *500*                      *0500*

all represent the same incomplete genotype of a diploid (2 digits per allele).

3°) the **first 0's** are **optional** so that *0112* and *0606* could also be written as *112* and *606*, respectively

4°) **different ploidy** levels can co-occur within a data set (not within a single individual), therefore alleles are defined only for the necessary number of gene copies, or **0 values** are attributed to "alleles" **on the left**: e.g.,

*123 125 125 121*                      *97 123*                      *0 0 97 123*

are correct genotypes for a tetraploid and two diploids, respectively (3 digits per allele).

5°) do not confound incomplete genotypes with genotypes for a ploidy level lower than announced: e.g.,

*2 3 4 0*                      *4500*                      *0 2 3 4*                      *45*

successively represent 2 tetraploids (with incomplete genotypes), a triploid and a diploid (the two latter with complete genotypes), respectively (1 digit per allele).

### 3.2.2. Dominant data (set the 5<sup>th</sup> format number, "number of digits", $\leq 0$ and the 6<sup>th</sup> format number, ploidy = 2)

Single locus genotypes are represented by numbers in either of the following ways:

1°) if the "number of digit" is set to 0, put

0                      for a missing data  
1                      for a recessive genotype  
2                      for a dominant genotype

2°) if the "number of digit" is set to  $-X$  (i.e. a negative integer), put

$-X$                       for a missing data  
0                      for a recessive genotype  
1                      for a dominant genotype

### 3.3. EXAMPLE OF DATA FILE

```
// this an example (lines beginning by // are comment lines)
// #ind #cat #coord #loci #dig/loc ploidy
5 0 2 4 2 2
4 10 20.5 50 100
Ind Lat Long adh got pgm lap
ind1 7.3 21 0101 0303 0 0101
ind2 8.4 52 101 101 102 103
3 5.11 103 0101 0303 0102 0103
4 1.0 13.2 1,1 3 3 1-2 01 03
lastind 1.94 129 0 1701 0118 1799
END
```

which specifies that 5 diploid individuals, not defined by categories, with location defined by 2 spatial coordinates, are scored at 4 loci where alleles are defined by 2 digits, and that 4 distance intervals will be considered as follow: [0 to 10], ]10 to 20.5], ]20.5 to 50], ]50 to 100]. Note that individuals 3 and 4 share the same genotypes but written in different ways.

### 3.4. NOTE ABOUT DISTANCE INTERVALS

When specific distance intervals are defined in the data file, the program checks that the **maximal distance** between two individuals / populations is not greater than the maximal distance of the last distance interval. Otherwise, an **additional interval is created**. Additional classes are also created for analyses at the individual level: an **intra-individual class** containing inbreeding coefficients (only for kinship statistics), and an **intra-group class** if individuals are organised into spatial groups (see §3.5.).

**Use a point to indicate decimals** (as in American notation), using a coma (as in French notation) would cause distances to be misinterpreted.

### 3.5. NOTE ABOUT SPATIAL GROUPS

If individuals consist of **spatial groups** that should be recognized (e.g. sibs from a given family, individuals from a given population), **individuals** belonging to a same group must **follow each other** in the data file **and** they must be given the **same spatial coordinates**. For analysis carried at the individual level, the program will then add a distance class for the pairwise coefficients between members of the same group (*intra-group* class).

For **analyses at the individual level**, when each individual receives specific spatial coordinates (no spatial groups, i.e. no two adjacent individuals in the data file share the same location), individuals are considered as independent from one another. This is typically the kind of analysis focusing on one continuously distributed population. If instead individuals are organised in spatial groups, **individuals from a same group are treated as dependent**. In such case, regression analyses do not take into account pairwise comparisons between individuals from a same group. The procedures for location permutations is also affected, as spatial group locations rather than individual locations are permuted (see §5.6.). When asking for the matrices of pairwise spatial and genetic distances between individuals, the value of the spatial distance between members of the same group is set conventionally to -1.

### 3.6. NOTE ABOUT MICROSATELLITE ALLELE SIZES

Several statistics are based on microsatellite allele sizes (e.g. R-statistics, Goldstein and Pollok's (1997)  $\delta\mu^2$ , Streiff et al. (1998) kinship analogue) using the size specified in the genotypes of the data file. Ideally, this size should be the number of repeats of the microsatellite motif. The computed statistics will still be valid if the size correspond to a constant plus the number of repeats (but the *mean allele size* information, see § 5.2., will not give the mean number of repeats). Problems may occur if allele sizes are given in terms of number of nucleotides rather than repeats. For the  $\delta\mu^2$  statistic, single locus estimates will be multiplied by the square of the motif size

(the same holds for the *Variance of allele size* information, § 5.2). For R-statistics and Streiff et al. (1998) kinship analogue, single locus estimates will not be affected, but multilocus estimates would be affected if the motif size vary among loci, in which case one should change the data file, dividing allele sizes per locus by the corresponding motif size.

### 3.7. USING A MATRIX TO DEFINE ARBITRARY PAIRWISE SPATIAL DISTANCES

Pairwise spatial distances between individuals or populations are normally computed as Euclidian distances using the spatial coordinates or from latitudes and longitudes. However, you can also specify each pairwise distance in an arbitrary way using a matrix. This can be useful in three cases: 1°) If you wish to consider non Euclidian spatial distances, such as distances more closely related to the probability of gene movements between locations. 2°) If you are not interested in spatial distances but in some other kind of pairwise distances (e.g. a morphological distance between individuals or populations) that you wish to correlate with genetic distance. 3°) If you wish to compute average statistics for particular pairwise comparisons between individuals / populations (for this purpose, you can define “distance” intervals and pairwise “distances” using integers).

The matrix of pairwise distances can be put at the end of the data file (just after the word “END”, see section 3.2.), or at the beginning of another text file. The use of such matrix and its location are specified while running the program (see §4.2.5. and 4.2.9.). The matrix can be written in two formats: a matrix format or a column format.

#### **Matrix format:**

This is a square matrix. The first line must begin with the letter **M**, followed by a number representing the matrix size (# of lines and columns). Then, individual or population names corresponding to each column must be written (separated by tab). Each of the next lines begin by the corresponding individual or population name followed by the pairwise distances attributed. The last line must contain the word “END”. Example:

```
// This is an example of a pairwise distance matrix written in matrix format with 5 rows and columns
M5   pop1   pop2   pop3   pop4   pop5
pop1  0      10.3   12     6      0
pop2  10.3   0      65     18     98
pop3  12     65     0      34     54
pop4  6      18     34     0      15
pop5  0      98     54     15     0
END
```

#### **Column format:**

In column format, each line corresponds to a pairwise comparison. The first line must begin with the letter **C**, followed by the number of lines (# of pairwise distances defined). Each of the next lines begins by the two individual or population names, separated by a tab, followed by the pairwise distance attributed. The last line must contain the word “END”. Example (the following matrix contains the same information as the one above except that self comparisons are left undefined):

```
// This is an example of a pairwise distance matrix written in column format with 15 pairwise distances defined
C15
pop1   pop2   10.3
pop1   pop3   12
pop1   pop4   6
pop1   pop5   0
pop2   pop3   65
pop2   pop4   18
pop2   pop5   98
pop3   pop4   34
pop3   pop5   54
pop4   pop5   15
END
```

#### **Notes:**

- 1°) For both matrix and column formats, the **order** of individuals / populations is **unimportant** (i.e. does not need to follow that of the data file).
- 2°) **Self-comparisons** are **not taken into account**.
- 3°) The **names must match** exactly those of the **data file** (case also matters!). This is straightforward for analyses at the individual level. However, for analyses at population level, population names vary: A) If one population = one **categorical group**, its name is that of the category. B) If one population = one **spatial group**, its name is that of the first individual of the spatial group in the data file. C) If one population = one **spatio-categorical group**, its name is written by joining the name of the first individual of the spatial group (as found in the data file) with the name of the category, the two being separated by the character '-'.  
 In order to create a template of the arbitrary matrix with the correct individual / population names, it can be convenient to run the program a first time without defining a pairwise distance matrix but asking to write pairwise distances and statistics in matrix or column formats (see §4.2.5. and 4.2.7.).
- 4°) Each pairwise comparison does not need to be defined, so that a matrix that does not contain all individuals / populations, or a matrix incompletely filled, are also accepted.
- 5°) Symmetrical comparisons (e.g. i-j and j-i) can not contain different distances (but one can be undefined).

### 3.8. DEFINING GENETIC DISTANCES BETWEEN ALLELES

When a statistic based on the genetic distances between alleles is request (e.g.  $N_{ST}$ ,  $N_{ij}$ ), the program asks to specify the file containing the **distance matrix between alleles**. The latter can be put at the end of the data file or in another file, and must be a symmetric square matrix with the following format:

**First line:** name of the locus followed by the allele names (numbers)

**Next lines:** allele name followed by the genetic distance between alleles

Example:

```
// This is an example of a distance matrix between alleles for a locus called "Hapl"
Hapl 1 2 3 4 15 26 7 18
1 0 6 5 2 4 4 3 3
2 6 0 1 4 2 2 3 3
3 5 1 0 3 1 1 2 2
4 2 4 3 0 2 2 1 1
15 4 2 1 2 0 2 1 3
26 4 2 1 2 2 0 3 1
7 0 2
18 0
END
```

#### Notes:

- 1°) **Locus names** must match exactly those of the data file (**case matters**).
- 2°) The **order** of alleles must be the same along rows and columns.
- 3°) Each allele found in the data file must occur in the matrix but the latter can contain additional alleles.
- 3°) **Self-comparisons** are **not taken into account**.
- 4°) The distance between each allelic pair must be defined but it can be so only one time in the matrix (i.e. a half matrix is also accepted).

### 3.9. DEFINING REFERENCE ALLELE FREQUENCIES FOR RELATEDNESS COEFFICIENTS

Most statistics available for analyses at the individual level (coefficients of kinship, relationship,...) provide measures of genetic similarity between individuals that are **relative** to a sample of individuals (usually all individuals in the data set, but other options are available with multiple categories, see §4.3.3 5°), which defines the “**reference allele frequencies**”. However, specific reference allele frequencies **can be given in a distinct file** (see option § 4.3.3. - 6bis) with the following format:

**First line:** for consecutive loci, name of each locus followed by the total number of alleles

**Next lines** (one per allele): for consecutive loci, allele name followed by the allele frequency

Example:

// This is an example of a matrix with reference allele frequencies for 3 loci called “Loc1”, “Loc2”, “Loc3”.

Loc1	5	Loc2	8	Loc3	3
1	0.3	120	0.01	2	0.67
2	0.1	122	0.04	43	0.32
3	0.05	124	0.35	3	0.01
4	0.15	130	0.13		
15	0.4	132	0.10		
		140	0.05		
		142	0.07		
		144	0.25		

#### Notes:

- 1°) These allele frequencies must be in a **distinct file** (the default name is “freq.txt”), not in the data file.
- 2°) **Locus names** must match exactly those of the data file (**case matters**).
- 3°) All loci in the data file must occur but additional loci may also occur (they will not be read).
- 4°) The order of alleles is unimportant.
- 5°) **All alleles found in the data file** must occur and be given a **non-null frequency**. Other alleles can also be present.
- 6°) The sum of allele frequencies at each locus must be one (sum between 0.999 and 1.001 accepted).

### 3.10. PRESENT DATA SIZE LIMITATIONS

max. 100 000 individuals

max. 100 000 loci

max. 999 alleles per locus (i.e. max 3 digits per allele)

max. 30 characters for the individual, category and locus names

max. 20000 random permutations

max. ploidy = 20 (note that all analyses on polyploids assume polysomic inheritance)

max. 100 distance intervals

max. length of any line in the data file: 1000 000 characters

Please contact us if these limitations are a problem for you, we may be able to send you a recompiled version with other specifications.

## 4. RUNNING THE PROGRAM

The program is written in C language. It has no fancy windowing features.

### 4.1. LAUNCHING THE PROGRAM

Launch the program by double-clicking on its icon and type the name of the data file (which should reside in the same folder as the program file, otherwise you must type its full path) or drag the icon of the data file into the *SPAGeDi* windows (the data file can then reside anywhere and the result file will be written in the directory of the data file). (This is also the procedure to follow if you need to import a data file.) For Windows version, you can also directly drag the data file on the *SPAGeDi* program icon or a shortcut to it. The program can also be launched using a **command line** or via another application, which can be useful to analyse numerous data sets obtained for instance by simulations. A command file that contains the keystrokes used to run an analysis can be associated to automatize the runs (e.g. “spagedi < cmds.txt” where cmds.txt is the file with the keystrokes commands). On Unix-derived systems, the program “tee” can be used to record keystrokes for playback later (e.g. “tee cmds.txt | spagedi” to record and “spagedi < cmds.txt” to repeat).

**Error messages** are given when files cannot be opened, data files are not well formatted or contain inconsistent information. These messages are not yet optimal and you may have difficulties finding out what is wrong in your data file (suggestions to improve this are welcome). When launching *SPAGeDi*, an **error file** “*error.txt*” is opened (and its previous content erased) and common errors made when preparing data files are listed. Additional information is added in this file whenever a problem occurs. *SPAGeDi* checks that the number of individuals and the number of categories found are the one specified in the data file, but there is no check for the number of loci (analyses considering only the first loci listed can thus be done by adjusting the number of loci given in line containing the format numbers).

### 4.2. SPECIFYING THE DATA / RESULTS FILES

Once the program is launched, you are requested to enter the name of the data file (unless you dragged the data file icon on the program) and the name of the results file.

If you just press *RETURN* to these questions, the default names “in.txt” and/or “out.txt” will be considered as data and results files, respectively (this can be useful if you wish to carry out many different analyses on the same data set without having to enter the file names each time).

You can also **import data** from a file in FSTAT (Goudet 1995) or GENEPOP (Raymont and Rousset 1995) format. Therefore, press *SPACE* and then *RETURN* when asked to enter the data file name, and select the format of the data file (FSTAT or GENEPOP). A new data file in *SPAGeDi* format will then be created, but it will not contain spatial information, so that you need to add them (as spatial coordinates per individual or as a matrix of pairwise distances), unless you do not need spatial analyses.

If a file with the same name as the results file already exists in the folder, the program will ask if you wish to: erase the existing file first (enter ‘e’), add results to the end of this file (enter ‘a’ or simply press *RETURN*), or change the name of the output file (enter the new name).

Once the data and result files are specified, the program first displays the basic information from the data file on the screen and waits for user to hit the *RETURN* key. The first set of information displayed is: the number of individuals, the number of categories and their names, the number of spatial coordinates and their names, the number of loci and their names, the number of digits used to specify alleles, the specified ploidy of the data, and the number of individuals of each ploidy. At this stage, if some individuals have missing genotypes at all loci, a warning message is addressed (**but the analysis can go on anyway**), and if different loci suggest different ploidy levels within some individuals, a warning message is addressed and the data file must be modified (the program stops here). The second set of information displayed is the groups recognised (categorical, spatial and spatio-categorical ones) with the minimal and maximal numbers of individuals per group.

### 4.3. SELECTING THE APPROPRIATE OPTIONS

You define the analyses to carry out and the results to write down by selecting options in 4 successive panels: 1°) *Level of analyses*, 2°) *Statistics*, 3°) *Computational options*, 4°) *Output options*. Some of the options will not be available depending on the structure of the data. You can come back to the beginning at different stages if you made an error of selection.

#### 4.3.1. Level of analyses: individual vs population

Analyses are carried out at the individual level or population level. When both categorical and spatial groups occur, you have also the choice among three different ways to define populations: as categorical, spatial, or spatio-categorical groups. If there are no categorical nor spatial groups in the data set, analyses are restricted to the individual level.

#### 4.3.2. Statistics

You must select the statistics to be computed (you can select several simultaneously). These statistics are computed for each pair of individuals or populations and the average values per distance interval as well as the regression statistics are given in the results file. More details about those statistics are given in § 6.1 and 6.2.

For analyses at the **individual level** with codominant markers, 12 statistics for pairwise comparisons between individuals are available:

- 1°) A kinship coefficient estimated according to J. Nason (described in Loiselle *et al.* 1995).
- 2°) A kinship coefficient estimated according to Ritland (1996).
- 3°) A relationship coefficient computed as Moran's *I* statistic (Hardy and Vekemans 1999).
- 4°) A relationship coefficient estimated according to Queller and Goodnight (1989).
- 5°) A relationship coefficient estimated according to Lynch and Ritland (1999) (*r* coef).
- 6°) A relationship coefficient estimated according to Wang (2002) (*r* coef).
- 7°) A relationship coefficient estimated according to Li *et al.* (1993).
- 8°) A fraternity coefficient (4-genes coefficient) estimated according to Lynch and Ritland (1999) ( $\Delta$  coef).
- 9°) A fraternity coefficient (4-genes coefficient) estimated according to Wang (2002) ( $\Delta$  coef).
- 10°)  $A_{ij}$ : a distance measure described in Rousset (2000) (the one called  $\hat{a}$  by Rousset).
- 11°)  $R_{ij}$ : a kinship coefficient analogue based on allele sizes (for microsatellites) ( $I'$  in Streiff *et al.* 1998).
- 12°)  $N_{ij}$ : a kinship coefficient analogue based on distances between alleles (OJ Hardy, unpublished).

Note: statistic 10° can not be computed for haploid data, and statistics 4°, 5°, 6°, 7°, 8° and 9° can presently be computed only for diploid data (5°, 6°, 8° and 9° also assume a population with Hardy-Weinberg genotypic proportions).

For the kinship coefficients, intra-individual values are also computed (as kinship between gene copies within individuals), providing estimates of an inbreeding coefficient.

For analyses at the **individual level** with dominant markers in diploids (see §3.2.2), 2 statistics are available:

- 1°) A kinship coefficient (Hardy 2003, Ley & Hardy 2012).
  - 2°) A relationship coefficient (Hardy 2003, Ley & Hardy 2012).
- In both cases you must specify the assumed inbreeding coefficient.

If estimates of broad-sense heritability are provided for each marker in the data file (line starting with H2 followed by each  $H^2$  estimate, see §3.1), they will be accounted for following Ley & Hardy (2012), otherwise  $H^2$  will be assumed equal to 1 and the estimation follows Hardy (2003).

For analyses at the **population level** with codominant markers, there are 9 choices for global and pairwise statistics between populations:

Statistics based on allele identity / non-identity

- 1°) Global F-statistics and pairwise  $F_{ST}$  (Weir & Cockerham 1984)
- 2°) Global F-statistics and pairwise Rho (Streiff *et al.* 1998)
- 3°) Global  $G_{ST}$  and pairwise  $G_{ST}$  (Pons & Petit 1996)
- 4°) Global  $G_{ST}$  and pairwise  $G_{ij}$  (Barbujani 1987)
- 5°) Global F-statistics and pairwise  $D_s$  (Nei's standard genetic distance, Nei 1978)

Statistics based on allele size for microsatellites

- 6°) Global R-statistics and pairwise  $R_{ST}$  (

7°) Global R-statistics and pairwise  $d_{m2}$  (Goldstein's ( $\delta\mu$ )<sup>2</sup> distance, Goldstein and Pollock 1997)

Statistics based on distances between alleles

8°) Global  $N_{ST}$  and pairwise  $N_{ST}$  (Pons & Petit 1996)

9°) Global  $N_{ST}$  and pairwise  $N_{ij}$  (OJ Hardy, unpublished)

When a statistic based on distance between alleles is asked, the program will ask to specify the file containing the matrix of distances between alleles.

#### **4.3.3. Computational options**

Once the statistics are chosen, you can select among different options regarding computations (several options can be selected simultaneously):

##### **1°) Use a matrix to define pairwise spatial distances.**

This option allows to define pairwise spatial distances between individuals / populations in an arbitrary way (otherwise, Euclidian distances are computed from the spatial coordinates given in the data file). Therefore, you must enter the name of the file containing the matrix (if the matrix follows the genotype information in the data file, just press *Return*). Details of the format of the matrix are given in § 3.6.

##### **2°) Make partial regression analyses (i.e. over restricted distance range).**

This option allows to define a distance range within which the spatial regression is computed, a useful option for gene dispersal parameter estimations (§ 6.3.). If this option is not selected, the regressions are carried out using all pairwise comparisons, except those with a distance of zero for the regressions on  $\ln(\text{distance})$ . Otherwise, minimal and maximal distances defining the range must be given. Entering no values (i.e. just pressing *RETURN*) means that the minimal or maximal distance is not bounded.

##### **3°) Make permutation tests.**

This option allows to test the significance of different statistics by random permutations of gene copies, individuals, locations, or allele sizes. More details in § 4.3.5.

##### **4°) Jackknife over loci.**

With this option, mean jackknifed estimators and jackknife standard errors (SE) are computed for multilocus average statistics. They can be used to derive approximate confidence intervals using the mean  $\pm 2SE$ . Jackknifing necessitates at least 2 polymorphic loci, but many polymorphic loci estimates are necessary to obtain reliable SE. Because SE are approximate, they should not be used for formal tests.

##### **5°) Restrict pairwise comparisons within or among (selected) categories.**

If the data are organised in categorical groups and analyses are carried out at the level of individuals or populations defined as spatio-categorical groups, you can select the type of pairwise comparisons for which the pairwise statistics are to be computed:

- 1°) All pairs (i.e. irrespective of categorical groups) = default option
- 2°) Only pairs within categories
- 3°) Only pairs among categories
- 4°) Only pairs within a specified category
- 5°) Only pairs between two specified categories

When 4° or 5° is selected, the name(s) of the category(ies) is(are) to be given.

When analyses are carried out at the **individual level**, you can select between **three options regarding reference allele frequencies** to compute the statistics (see § 6.1.1. for explanations):

- 1°) whole sample (i.e. reference = allele frequencies found in the whole sample)
- 2°) unweighted average allele frequencies over categories (i.e. reference = mean of allele frequencies found in each category)
- 3°) sample within category (i.e. reference = sample to which the pair of individuals belongs); option available only for pairs within categories or within a specified category.

6°) ***Pairwise Fst (or Rst, or Rho) provided as Fst/(1-Fst) ratio.***

When this option is selected, pairwise differentiation between population will be estimated using  $F_{ST}/(1-F_{ST})$  ratios. This is useful to analyse isolation-by-distance patterns because  $F_{ST}/(1-F_{ST})$  is expected to vary linearly with the distance or its logarithm. See §6.3.

6bis°) ***Define reference allele frequencies to compute relatedness coefficients.***

When this option is selected, pairwise relatedness coefficients will be computed relative to reference allele frequencies given in a separate file (see §3.9. for the format). *SPAGeDi* will ask the name of this file. This option cannot be applied for the statistics developed for dominant markers in diploids, the relationship coefficient computed as a Moran's *I* statistic, and Rousset's (2000) *a* coefficient.

#### **4.3.4. Output options**

A second set of options concerns the information given in the results file:

1°) ***Report allele frequencies for each population / category (otherwise only averages reported).***

In the results file, global allele frequencies and gene diversities are reported. Activating this option means that this information will also be given for each population (or for each categorical group in the case of analyses at the individual level including categories).

2°) ***Set subsample size to compute allelic richness (or use minimal sample size).***

To compare the genetic diversity of populations with different sample sizes, one can estimate the expected allelic richness for a random subsample of gene copies of constant size *k*. Select this option to define *k*, otherwise *k* will be the smallest sample found.

3°) ***Report all stat of regression analyses (otherwise only slopes reported).***

When this option is activated, the following statistics of the regressions of pairwise statistics on spatial distances are provided: slope, intercept, determination coefficient, number of pairs, mean and variance of values of (log) distance and statistics.

4°) ***Report matrices with pairwise spatial distances and genetic coefficients.***

With this option, pairwise spatial distances and pairwise statistics are given at the end of the results file. You must also specify whether the pairwise statistics are to be given for each locus or only the multilocus estimates, and whether pairwise values are to be written only in columnar form or also in matrix form. You can also select Phylip format which gives a square matrix of genetic distances that can be copied directly to a text file for further analyses (there is no tab delimitations). Note that in Phylip format, negative genetic distances are given the value -0.0000. Estimates of the inbreeding coefficient for each individual are given in the columnar format if you asked to compute a kinship coefficient between individuals (the inbreeding coefficients given are computed as kinship coefficient between homologous gene copies within individual).

5°) ***Convert data file into GENEPOP or FSTAT format.***

This option allows to create a data file that can be used by the software FSTAT (Goudet 1995) or GENEPOP (Raymond and Rousset 1995), and it is available only with diploid data. If analyses were asked at the population level, the GENEPOP or FSTAT file codes data for the same populations as selected. For analyses selected at the individual level, the FSTAT file code data as a single population, whereas the GENEPOP file code data as if each individual constituted a single population (this is the necessary format to use Rousset's pairwise distance between individuals in GENEPOP).

6°) ***Estimate gene dispersal sigma.***

For analyses at the individual level, this option can be used to estimate the gene dispersal distance parameter  $\sigma$  from the regression of pairwise kinship coefficients (or Rousset distance) on the logarithm of the distance (Rousset 2000; Hardy et al. 2006). **You must assume that genotypes come from a two-dimensional population at drift-dispersal equilibrium** so that theoretical expectations of isolation-by-distance models hold (§ 6.3.). You will be asked to enter the effective population density. Be consistent with units: the density must be given in number of individuals per square distance unit where the latter is the same as the one used for spatial coordinates or for the spatial distance matrix. You must also enter *X* defining the distance range ( $\sigma$  to  $X \cdot \sigma$ ) over which regression is applied (*X* should be between 10 and 50, the default value=20). *SPAGeDi* will then apply an iterative procedure to estimate the  $\sigma$  from the genetic

structure on the restricted distance range (see § 6.3.). The iterative procedure might not converge, indicating that the data are not powerful enough to get reliable estimates (see § 6.3. for additional advices).

7°) ***Estimate selfing rate from standardized identity disequilibrium.***

This option is available for codominant markers and multiple loci. Standardized identity disequilibrium is then computed for each pair of loci in each population or category, and a multilocus estimator is provided, together with estimates of the selfing rate assuming a mixed mating model (i.e. a proportion  $S$  of selfing and  $1-S$  or random outcrossing) at equilibrium (David et al. 2007; Hardy 2015). Standard errors are provided by jackknifing over loci and, optionally, the significance of the multilocus standardized identity disequilibrium coefficient is tested unilaterally by a permutation test whereby single-locus data are randomly permuted among individuals within each population.

8°) ***Report actual variance of pairwise genetic coefficients (Ritland 2000).***

With this option activated, the actual variance (i.e. excluding sampling variance) of pairwise statistics is given for each distance class following the approach described in Ritland (2000), which requires independent loci (at least two). An estimate of the standard error by jackknifing over loci is also given with at least 3 loci. This variance is useful to compute marker based estimates of the heritability ( $h^2$ ) or population differentiation ( $Q_{st}$ ) at quantitative traits (Ritland 1996, 2000).

#### **4.3.5. Permutation tests**

If permutation tests are selected, you have two sets of additional options (you can select several at once):

Firstly (only if statistics based on allele size or distance between alleles have been selected),

1°) ***Test of genetic structuring (permuting gene copies, individuals and/or locations)***

To test individual inbreeding, population differentiation, and/or spatial structure.

2°) ***Test of mutation effect on genetic structure (permuting alleles)***

To test if the allele size (microsatellites) or the phylogenetic distance between alleles contributes to explain the genetic structure (e.g. test of phylogeographic signal).

Secondly,

1°) ***Report only P-values (otherwise details of permutation tests are reported)***

If this option is selected, only P-values for 2-sided tests are reported. Otherwise, the following details are given: object permuted, # permutations, # of different values of the statistic after permutation, observed values before permutation, mean values after permutation, standard errors of mean values after permutation, 95% confidence intervals, P-values of 1- and 2-sided tests.

2°) ***Define # of permutations for each randomised unit (otherwise same #)***

Allows defining a high number of permutations for the statistics that most interest you, and no or few permutations for the ones that are not of interest for you or that would take a lot of computation time.

3°) ***Initialise random number generator (otherwise initialisation on clock)***

Define initial seed for random number generator, otherwise the latter is defined according to the computer's internal clock (this option is useful for debugging).

4°) ***Perform tests of genetic differentiation on each population pair***

Select this option to test genetic differentiation or the impact of mutations on genetic differentiation for each pair of populations.

You must then enter the number(s) of permutations you wish. On large data sets, resampling can be time consuming; hence there is a compromise between computation time and precision of the probability ( $P$ -values). It is advisable to enter at least 199 if you are satisfied with a 5% significance level, 999 for a 1% level, 9999 for a 0.1% level. Enter "0" if you do not need tests.

## **4.4. INFORMATION DISPLAYED DURING COMPUTATIONS**

Once the program proceeds to the calculations, it displays the computational stage: computation of allele frequencies, of distance intervals, of pairwise statistics, permutation tests. When the computations are finished, a message will appear on the screen and pressing any key will close the window. You can proceed to examination

of the results file. If the program crashed, do not forget to open the file "**error.txt**", because this may give you some information on the origin of the problem.

Details relative to distance intervals are displayed once computed and computations proceed.

Each interval (class) is characterised by

- 1°) *max d* its maximal distance (the minimal distance is the maximal distance of the preceding interval)
- 2°) *mean d* the average distance between individuals / populations for the pairs belonging to the interval
- 3°) *mean ln(d)* idem but using the  $\ln(\text{distance})$  between individuals / populations
- 4°) *# pairs* the number of pairwise comparisons belonging to the interval
- 5°) *% partic* the proportion (%) of all individuals / populations represented at least once in the interval
- 6°) *CV partic* the coefficient of variation of the number of times each individual / population is represented

**Notes:**

- 1°) If analyses are restricted to pairwise comparisons within or among (specified) category(ies), the information per distance intervals considers only pairs satisfying these conditions.
- 2°) Information on distance intervals can be useful for fine-tuning them. For example, low *% partic* and/or high *CV partic* means that the statistics computed for the corresponding interval involve data from only a fraction of the individuals / populations. Hence, as a **rule of thumb**, we advise that for each distance interval: *% partic* > 50%, and *CV partic* <= 1. For **analyses at the individual level** we also advise that *# pairs* > 100, given the large standard errors typically observed for pairwise coefficients between individuals (with many loci or highly polymorphic loci this number could be reduced).

## 5. INTERPRET THE RESULTS FILE

All the results are found in a single results file. The results file can be read as a text file but it is best to open it with a worksheet program (e.g. Excel), tabs being used to delimit columns. The results appear in the following order.

### 5.1. BASIC INFORMATION

First, the basic information as it appeared on the screen when running the program is written: names of data and results files, numbers of individuals, categories, spatial coordinates and loci, names of categories, spatial coordinates and loci, ploidy, numbers of individuals for each ploidy level, number of categorical, spatial and spatio-categorical groups (see § 4.2.).

### 5.2. ALLELE FREQUENCY ANALYSIS

Second, for each locus are written: the number of missing genotypes (*# missing genotypes*), the number of incomplete genotypes (*# incomplete genotypes*), the total number of defined gene copies (*# of defined gene copies*), the number of alleles with non zero frequency (*NA*), the effective number of alleles estimated following Nielsen et al. (2003) (*NAe*), the allelic richness expressed as the expected number of alleles among *k* gene copies (*AR(k=...)*), the gene diversity corrected for sample size (*He*), the inbreeding coefficient (*Fi*) followed by the P-value of a permutation test if asked (*Pval(Fi<>0)*), the name (or size) of each allele (*allele names* or *allele size*) (i.e. the number given in the data file), and the allele frequencies (*allele frequencies*). When a statistic based on allele size (e.g. R-statistics) has been selected, the mean (*Mean allele size*) and variance (*Variance of allele size*) of allele sizes are also given. When a statistic based on distances between alleles (e.g. N-statistics) has been selected, the genetic diversity or ordered (*v*) and non-ordered (*h*) alleles following Pons and Petit (1996) are also given. This information is given for the whole sample and, if asked, for each population (analysis at population level) or each category (analysis at individual level).

If relatedness coefficients were computed using specified reference allele frequencies (individual level analyses), the latter are written.

### 5.3. IDENTITY DISEQUILIBRIUM AND SELFING RATE ESTIMATES

If the option for estimating identity disequilibrium and selfing rate was selected, results appear just after the allele frequency tables and are given for each population or category, or for the whole sample if no categories or population was defined. For diploids, results consist in a multilocus estimate of the standardized identity disequilibrium (*g2*) followed by its standard error (SE) and a unilateral test of significance (Pval), then an estimate of the selfing rate based on *g2* with SE, and finally the *g2* values for each pair of loci. For polyploids, *g2* is estimated considering (1) the genotypes as given in the dataset (*g2g*) and (2) the phenotypes where allele dosage is not used (*g2p*) and, in each case, the selfing rate is estimated assuming either that chromosome segregation occurred or considering the rates of double reduction (ALPHA) optionally given in the dataset (cf §3.1; Hardy 2015).

### 5.4. TYPE OF ANALYSES

After the allele frequencies information, it is specified whether the analyses are carried out at the individual or population level, if pairwise comparisons are restricted to pairs within or among category(ies), and, for analyses at the individual level, if statistics are computed on basis of the global (whole sample) or local (within category) allele frequencies (for comparisons within (a) category) or relative to given reference allele frequencies.

### 5.5. DISTANCE INTERVALS

Next, for each distance interval corresponding to a column are written:

- *Dist classes*: the names of the distance classes (1, 2,...)
- *Max distance*: the maximum distance defining the interval: distance interval  $c = ] \text{Max dist } (c-1), \text{Max dist } (c) ]$
- *Number of pairs*: the number of pairs of individuals separated by the given distance interval

- *% partic*: the percentage of individuals participating at least once in a pairwise comparison within the interval
- *CV partic*: the coefficient of variation (i.e. the ratio of the standard deviation over the average) of the number of times each individual participate in pairwise comparisons within the interval
- *Mean distance*: the average distance separating pairs of individuals within the interval
- *Mean ln(distance)*: the average natural logarithm of the distance separating pairs of individuals within the interval

**Note:** For analyses at the individual level, an *intra individual* class is added for comparison of gene copies within individual (only defined for kinship statistics when ploidy is larger than one), and this class actually corresponds to an inbreeding coefficient. When individuals consist of groups, the distance class “1” corresponds to *intra group* comparisons.

## 5.6. COMPUTED STATISTICS

For each selected statistic, the following results are given for the multilocus estimate and each locus:

- in columns labelled  $F_{IT}$ ,  $F_{IS}$ ,  $F_{ST}$  or  $R_{IT}$ ,  $R_{IS}$ ,  $R_{ST}$  or  $G_{ST}$  or  $N_{ST}$  (for analyses at population level only): the global statistics. When analyses are restricted to comparisons within a given category or between two given categories, global statistics are computed considering only the populations included in the concerned category(ies).
- in columns corresponding to each distance class: the average value of the pairwise coefficients computed over all pairs of individuals or populations within the distance interval (all pairs of gene copies within individuals in the case of the “*intra individual*” class, for analyses at the individual level).
- under the column “average”: the average value of the coefficients computed over all pairs of individuals or populations, whatever the distance (for analyses at individual level, it includes *intra group* class but not *intra individual* class).
- under “distance range for regression analyses”: the distance range used to compute regressions of pairwise statistics on spatial distance or  $\ln(\text{distance})$ .

The next columns report the results of the regression analyses, first with the linear distance, then with the  $\ln(\text{distance})$ . If the option “*Report details of regression analyses*” has not been selected (see §4.2.5), only the slopes (*b-lin* and *b-log*) are given; otherwise the following statistics are reported for each regression analysis:

- the slope *b*
- the intercept *a*
- the coefficient of determination  $r^2$  (i.e. squared correlation coefficient)
- the number of pairwise comparisons *N* (taking account of missing data)
- the mean (*Md*) and variance (*Vd*) of pairwise distances or  $\ln(\text{distances})$
- the mean (*Mv*) and variance (*Vv*) of pairwise statistics

If the option ***Jackknifing over loci*** has been selected (see §4.2.5), results of a jackknife procedure deleting each locus at a time are given on the two lines following the information of the last locus: the first line gives the jackknifed estimates, the second one gives their standard errors. Calculations follow Sokal and Rohlf (1995, p.821).

If the option ***Report actual variance of pairwise genetic coefficients*** has been selected (see §4.3.4), estimations of the actual variance of pairwise genetic coefficients within each distance interval is provided following the method of Ritland (2000).

If the option ***Estimate gene dispersal sigma*** has been selected (see §4.3.4), estimates of the neighbourhood size (*Nb*) and sigma (the square root of half the mean square parent-offspring distance) are given (see §6.3 for technical notes). If the iterative procedure used to get these estimates did not converge, the message “no convergence” appears. If successive estimates cycled among a set of values, the average and the range of values obtained over a cycle are given. When the option ***Jackknifing over loci*** was also selected, the iterative procedure is repeated again after removing one locus at a time, and standard errors on estimates are given.

## Notes:

- 1°) For analyses at the individual level, **the *intra individual* kinship coefficient is an inbreeding coefficient expressing the departure from Hardy-Weinberg genotypic proportions (cf.  $F_{IS}$ )**. When individuals consist of spatial groups corresponding to different populations, this is equivalent to  $F_{IT}$  (not  $F_{IS}$ ). Kinship statistics for the *intra group* class provides an estimator similar to  $F_{ST}$  if groups correspond to different populations.
- 2°) For analyses at the individual level, the slopes of the regressions do not include the pairs of individuals within spatial groups (*intra group* class). As slopes do not depend on an arbitrary choice of distance, they offer a convenient measure of the degree of spatial genetic structuring. Moreover, under some conditions, these slopes can be related to population genetic parameters like neighbourhood size (see §6.3.).

## 5.7. PERMUTATION TESTS

If permutation tests are selected as option (see §4.2.5), results of these tests are written after the pairwise coefficients. These tests are based on the comparison of the observed values with the corresponding frequency distributions when random permutations of the data are performed. For each locus and the multilocus estimates, tests are given for global statistics (population level analyses), each distance class, and the slopes of the regressions analyses. When asked in the options, differentiation coefficients are also given for each pair of populations.

The following information is reported (unless the option “*Report only P-values*” has not been selected - see §4.3.5 – in which case only P-values for the two-sided tests are given):

- the object (gene copies, individuals or location) permuted (and how):	<i>Object permuted</i>
- the number of valid permutations (i.e. for which the statistic was computable):	<i>N valid permut</i>
- the number of different values obtained for the different permutations:	<i>N different permut val</i>
- the observed value (i.e. before permutation):	<i>Obs val</i>
- the average value after permutation:	<i>Mean permut val</i>
- the standard error of the distribution of values after permutation:	<i>SD permut val</i>
- the lower 95% confidence interval value:	<i>95%CI-inf</i>
- the upper 95% confidence interval value:	<i>95%CI-sup</i>
- the P-value for the 1-sided test observed value < permuted value:	<i>P(1-sided test, H1: obs&lt;exp)</i>
- the P-value for the 1-sided test observed value > permuted value:	<i>P(1-sided test, H1: obs&gt;exp)</i>
- the P-value for the 2-sided test observed value different from permuted value:	<i>P(2-sided test, H1: obs!=exp)</i>

The following **code** is used to designate the object permuted and how it is permuted (*Objected permuted*):

GaI	permutation of <u>G</u> ene copies <u>a</u> mong all <u>I</u> ndividuals
GaIwC	permutation of <u>G</u> ene copies <u>a</u> mong <u>I</u> ndividuals <u>w</u> ithin <u>C</u> ategory
GaIwP	permutation of <u>G</u> ene copies <u>a</u> mong <u>I</u> ndividuals <u>w</u> ithin <u>P</u> opulation
IaSG	permutation of <u>I</u> ndividuals <u>a</u> mong <u>S</u> patial <u>G</u> roups
IaSGwC	permutation of <u>I</u> ndividuals <u>a</u> mong <u>S</u> patial <u>G</u> roups <u>w</u> ithin <u>C</u> ategory
IaP	permutation of <u>I</u> ndividuals <u>a</u> mong all <u>P</u> opulations
IaPwC	permutation of <u>I</u> ndividuals <u>a</u> mong <u>P</u> opulations <u>w</u> ithin <u>C</u> ategory
IbPP	permutation of <u>I</u> ndividuals <u>b</u> etween a <u>P</u> air of <u>P</u> opulations
ILaI	permutation of <u>I</u> ndividual <u>L</u> ocations <u>a</u> mong all <u>I</u> ndividuals
ILaIwC	permutation of <u>I</u> ndividual <u>L</u> ocations <u>a</u> mong <u>I</u> ndividuals <u>w</u> ithin <u>C</u> ategory
SGLaSG	permutation of <u>S</u> patial <u>G</u> roup <u>L</u> ocations <u>a</u> mong all <u>S</u> patial <u>G</u> roups
SGLaSGwC	permutation of <u>S</u> patial <u>G</u> roup <u>L</u> ocations <u>a</u> mong <u>S</u> patial <u>G</u> roups <u>w</u> ithin <u>C</u> ategory
PLaP	permutation of <u>P</u> opulation <u>L</u> ocations <u>a</u> mong all <u>P</u> opulations
PLaPwC	permutation of <u>P</u> opulation <u>L</u> ocations <u>a</u> mong <u>P</u> opulation <u>w</u> ithin <u>C</u> ategory
ASaAwL	permutation of <u>A</u> llele <u>S</u> izes <u>a</u> mong <u>A</u> lleles <u>w</u> ithin <u>L</u> ocus
RCoDMbA	permutation of <u>R</u> ows and <u>C</u> olumns <u>o</u> f <u>D</u> istance <u>M</u> atrices <u>b</u> etween <u>A</u> lleles

When permutation of an object is done *within category*, it means that the permuted objects remain in their original categorical group after permutation. This is done when pairwise comparisons are restricted to within category(ies) (see §4.2.3.).

As the preceding code shows, the **object permuted** varies:

- Gene copies are permuted among individuals, each locus independently, for tests on  $F_{IS}$ ,  $F_{IT}$ ,  $R_{IS}$ ,  $R_{IT}$  and *intra individual* coefficients (individual inbreeding coefficient). Missing data are not permuted (i.e. permutation concerns only defined gene copies). For  $F_{IS}$  and  $R_{IS}$ , gene copies are permuted only within population.
- Individuals (i.e. whole genotypes) are permuted among populations or spatial groups for tests on global  $F_{ST}$ ,  $R_{ST}$ ,  $Rho$ ,  $G_{ST}$ ,  $N_{ST}$  and *intra group* coefficients.
- Individual Locations (for analyses at the individual level without spatial groups), Spatial Group Locations (for analyses at the individual level with spatial groups), or Population Locations (for analyses at the population level) are permuted among the available locations for tests on each distance class (except the *intra individual* and *intra group* ones), and tests on the regression slopes. This is equivalent to a Mantel test between a matrix of genetic distances and a matrix of geographic distances.
- Allele Sizes represented within each locus are permuted among allelic states to test if allele sizes are informative (assuming stepwise mutations) on global R-statistics ( $R_{IS}$ ,  $R_{IT}$ ,  $R_{ST}$ ), pairwise  $R_{ST}$  (and regression slope), or the correlation coefficient between allele sizes (Streiff et al. 1998) for individual level analyses (cf. Hardy et al. 2003).
- Rows and Columns of Distances Matrices between Alleles are permuted to test if the genetic distances between alleles are informative on global or pairwise  $N_{ST}$  or  $N_{ij}$  (and regression slope) (cf. Pons & Petit 1996; Burban et al. 1999).

#### Notes:

- 1°) Tests based on individual permutations indicate whether population or spatial groups are genetically differentiated, whereas tests based on location permutations indicate whether the degree of differentiation or relatedness between individuals, spatial groups or populations depends on the geographical distance.
- 2°) When using an **arbitrary matrix** to define **pairwise distances** (§3.6.), location permutations correspond to permutations of the rows and columns of this matrix (as in a Mantel test).
- 3°) For tests based on individual or location permutations, the presence of missing data is not taken into account, i.e. even if there is no gene copies defined at some loci for a given individual (location), this individual (location) will be permuted with the other ones. Hence, these **tests** can be **biased** for loci with a substantial proportion of **missing data**. In such case, it might be preferable to make the tests on each locus separately using single locus data files in which missing data are removed.
- 4°) When analyses are restricted to **comparisons among categories** (see §4.2.3.), care must be taken in the **interpretation of the tests** based on location permutations (i.e. tests on pairwise coefficients and regression slopes). It can be tempting to interpret these tests as indicating whether the spatial structures within each category have developed independently or not, because if gene flow occurs or had occurred recently among categories, one would indeed expect a spatial correlation between the patterns of genetic variation of the different categories. However, these tests are **biased** for such purpose because they are based on random permutations that not only make the spatial structures of the different categories independent from one another, but also break down the structure within each category (ideally the level of structuring within category should be kept intact). Therefore, a test may be significant because the patterns of spatial genetic variation within category match for different categories just by chance, whereas these structures developed truly independently. Nevertheless, a test of the independence of the spatial structures of the different categories can be done if different independent loci (i.e. in linkage phase equilibrium within category) are available, using conventional non parametric methods (e.g. sign tests) on the regression slopes per locus.
- 5°) For identity disequilibrium (§5.3), there is also a specific permutation test whereby single-locus genotypes are permuted randomly among individuals from the same population/category, independently for each locus, except for missing genotypes (no permutation, so that the distribution of missing data is kept intact after randomization). The Pval is the proportion of identity disequilibrium values after randomization that are larger than the observed identity disequilibrium value.

## **5.8. MATRICES OF PAIRWISE DISTANCES AND STATISTICS**

When asked as option (see §4.2.5), matrices of pairwise geographical distances between individuals and pairwise coefficients (for the multilocus estimates, optionally for every locus) are provided in two possible formats (defined as options, see §4.2.7): as square matrices (lines correspond to one individual / population, column to another), or in columns (first and second individuals / populations are figured in two columns, spatial distances, multilocus estimates and/or per locus estimates are given in the next columns). You can also select Phylip format which gives a square matrix of genetic distances that can be copied directly to a text file for further analyses (there is no tab delimitations). Note that in Phylip format, negative genetic distances are given the value  $-0.0000$ . Estimates of the inbreeding coefficient for each individual are given in the columnar format if you asked to compute a kinship coefficient between individuals. Note that the value reported for the geographical distance between individuals belonging to the same spatial group is  $-1$ .

## 6. TECHNICAL NOTES

### 6.1. STATISTICS FOR INDIVIDUAL LEVEL ANALYSES

Analyses at the individual level are carried out by computing measures of genetic relatedness or genetic distance between individuals for each possible pair (unless stated differently, see §4.3.3). These pairwise coefficients are computed for each locus and a multilocus weighted average. They are regressed on pairwise spatial distances and they are averaged to compute mean values per distance interval. Hence, a multilocus estimate for a distance interval is computed by first averaging pairwise coefficients over loci (weighted average), then averaging multilocus pairwise coefficients over all pairs included in the distance interval.

For codominant data, *SPAGeDi* allows the user to compute five types of “relatedness” coefficients between individuals: “kinship”, “relationship” and “fraternity” coefficients, plus a distance measure based on allele identity, and a kinship analogue based on allele size. For some of these coefficients, several estimators are available, so that a total of 13 different statistics can be estimated. Comparisons of the statistical properties of different estimators can be found in Lynch & Ritland (1999), Van de Casteel et al. (2001), Wang (2002), Vekemans & Hardy (2004). The fraternity coefficient is a “4-genes” coefficient, in the sense that it is based on the simultaneous comparison of all of the 4 homologous gene copies of two diploid individuals. The other coefficients are “2-genes” coefficients, because they are ultimately based on comparisons between 2 homologous gene copies. For dominant data, *SPAGeDi* allows to compute two types of “relatedness” coefficients between individuals: “kinship” and “relationship” coefficients. There is no unified terminology for these different coefficients so that we attempt to define them below.

Most statistics available are relative measures of genetic similarity that depend on the definition of a reference sample or reference allele frequencies (see below). Specific reference allele frequencies can be defined in a distinct file (§3.9, §4.3.3.) but they will not be taken into account for the relationship estimator based on Moran's *I* statistic (§6.1.2), Rousset's distance measure (§6.1.4), and statistics developed for dominant markers in diploids. Note that sampling bias corrections that are normally applied for some statistics (see below) are not applied when specific reference allele frequencies are defined.

#### Synthetic table of the statistics proposed by *SPAGeDi* for individual level analyses and their properties.

Coefficient Estimator (ref)	Intra-indiv. estimate (inbreeding)	Assumptions		Statistical properties <sup>3</sup>	
		Ploidy	Inbreeding <sup>2</sup>	Accuracy (low bias)	Precision (low variance)
<b>Kinship</b>					
Loiselle et al. 1995	+	1 to 8		+++	++
Ritland 1996	+	1 to 8		++	+++
Hardy 2003 (for <b>dominant marker</b> )		2	$F_I$ to give	++	++
<b>Relationship</b>					
Hardy & Vekemans 1999 (Moran's I)		1 to 8		+++	++
Lynch & Ritland 1999		2	H-W	+++	++
Queller & Goodnight 1989		2 <sup>1</sup>		+++	++
Wang 2002		2	H-W	+++	++
Li et al. 1993		2	H-W	+++	+++
Hardy 2003 (for <b>dominant marker</b> )		2	$F_I$ to give	++	++
<b>Fraternity</b>					
Lynch & Ritland 1999		2	H-W	++	++
Wang 2002		2	H-W	+++	+++
<b>Rousset's 'a'</b>					
Rousset 2000		2 to 8	no selfing	+++	+
<b>Kinship analogue based on allele size <math>R_{ij}</math></b>					
Streiff et al. 1998	+	1 to 8		+++	+
<b>Kinship analogue based on allele distances <math>N_{ij}</math></b>					
OJ Hardy, unpublished	+	1 to 8		+++	+

<sup>1</sup> Queller & Goodnight's estimator is defined for any ploidy level but *SPAGeDi* computes it only for diploids.

<sup>2</sup> *H-W* means that the estimator was derived assuming Hardy-Weinberg proportions and may be biased under inbreeding.  $F_I$  to give means that an independent estimate of the individual inbreeding coefficient must be provided.

<sup>3</sup> Statistical properties are based on literature results (Lynch & Ritland 1999, Van de Casteel et al. 2001, Wang 2002, Vekemans & Hardy 2004) and personal experience. These indications must be considered with caution because the actual ranking of the performances of the statistics depend on the data set.

### 6.1.1. Kinship coefficient

#### Definition and interpretation

In a generic way, kinship coefficients, also called coancestry coefficients, are based on the probability of identity of alleles for two homologous gene copies sampled in some particular way. In the case of a kinship coefficient between two individuals, the two gene copies are randomly sampled within each of the two individuals. F-statistics are also kinship coefficients but for gene copies sampled in different ways (see §6.2.).

A kinship coefficient ( $F$ ) is often defined as the probability of identity by descent of the gene copies compared (e.g. Ritland 1996) but estimators based on genetic markers actually estimate a “relative kinship”, that can be defined as ratios of differences of probabilities of identity in state (Rousset 2002; Vekemans & Hardy 2004). Thus, equating these kinship coefficients with probability of identity by descent is not true in general (Rousset 2002). In the case of two individuals  $i$  and  $j$ , the kinship coefficient between them can be **defined as  $F_{ij}=(Q_{ij}-Q_m)/(1-Q_m)$** , where  $Q_{ij}$  is the probability of identity in state for random gene copies from  $i$  and  $j$ , and  $Q_m$  is the average probability of identity by state for gene copies coming from random individuals from the sample (i.e. “reference population” = sample). As defined here, kinship is not really a population genetics parameter as it depends on an arbitrary sample. Note also that with this definition, **negative relative kinship coefficients** naturally occur between some individuals, it simply means that these are less related than random individuals (a definition equating kinship and probability of identity by descent would not allow negative values).

#### Changing the reference population

In some contexts, one wishes to compare estimates of kinship coefficients with some expected values derived from pedigree information. For example, in the case of sib families coming from non-inbred diploid parents, the kinship between sibs is expected to be 0.125 for half-sibs and 0.25 for full-sibs according to standard computations (e.g. Lynch and Walsh 1998). These are actually the expected values of a kinship coefficient relative to the parental generation (i.e. where  $Q_m$  is for random gene copies from the parental generation, which is here the “reference population”). Thus, kinship is not relative to the same “reference population” when computing it from a data set containing some sib-families (“reference population” = sample) and when considering expected values from pedigree information (“reference population” = ancestors of the genealogy, which are assumed to be “unrelated”). One can however switch between these different references if one can find pairs of individuals in the data set that are expected to be “unrelated” in the sense of the putative pedigree (cf. Hardy 2003). For sib-families, this would be the case of pairs of individuals belonging to different families. Let  $F^\circ$  be the kinship between individuals from different sib families as computed from the sample reference, you can then compute kinship coefficients relative to the pedigree reference,  $F_{ij}'$ , as  $F_{ij}'=(F_{ij}-F^\circ)/(1-F^\circ)$ . These  $F_{ij}'$  are expected to be 0.125 or 0.25 in case of half and full sibs, respectively. When allele frequencies of a reference population (e.g. the parental population) can be assessed precisely, an alternative approach consists in estimating kinship (or other relatedness) coefficients using specified reference allele frequencies corresponding to this reference population (see §3.9, §4.3.3). All alleles found in the individuals being compared must then have a non-null frequency (*SPAGeDi* tests this).

#### Estimators

For codominant markers, *SPAGeDi* proposes two estimators of kinship (coefficients relative to the sample): 1°) a kinship coefficient computed as a correlation coefficient between allelic states proposed by J. Nason (Loiselle *et al.* 1995), 2°) a kinship coefficient estimated according to Ritland (1996).

1°) is computed as  $F_{ij} = \sum_l [ \sum_a (\sum_{ci} \sum_{cj} (x_{l_{cia}} - p_{la})(x_{l_{cja}} - p_{la}) / \sum_{ci} \sum_{cj} 1) + \sum_a (p_{la}(1 - p_{la}) / (n_l - 1)) ] / \sum_l \sum_a (p_{la}(1 - p_{la}))$

where  $x_{l_{cia}}$  is an indicator variable ( $x_{l_{cia}} = 1$  if the allele on chromosome  $c$  at locus  $l$  for individual  $i$  is  $a$ , otherwise  $x_{l_{cia}} = 0$ ),  $p_{la}$  is the frequency of allele  $a$  at locus  $l$  in the reference sample,  $n_l$  is the number of gene copies defined in the sample at locus  $l$  (the number of individuals times the ploidy level minus the number missing alleles), and  $\sum_{ci}$  stands for the sum over the homologous chromosomes of individual  $i$ . Here, the term involving  $(n_l - 1)$  is a sampling bias correction. This estimator should be equivalent to the one computed by John Nason's *FijAnal* software, except that the bias correction might differ slightly.

Note that this formula is identical to:

$$F_{ij} = \sum_l [ \sum_a (p_{ila} - p_{la})(p_{jla} - p_{la}) + \sum_a (p_{la}(1 - p_{la}) / (n_l - 1)) ] / \sum_l \sum_a (p_{la}(1 - p_{la}))$$

where  $p_{ila}$  is the frequency of allele  $a$  at locus  $l$  in individual  $i$ .

2°) is computed as  $F_{ij} = \sum_l ((\sum_a \sum_{ci} \sum_{cj} (x_{l_{cia}} x_{l_{cja}} / p_{la}) / \sum_{ci} \sum_{cj} 1) - 1) / \sum_l (n_l - 1)$

where  $m_l$  is the number of different alleles found in the sample at locus  $l$ . Note that the bias correction consisting in removing the compared individuals when computing  $p_{la}$ , as suggested by Ritland (1996), is not applied.

The two estimators differ mainly by the way information from the different alleles and different loci are combined to provide average estimates per locus or multilocus estimates. Basically, Ritland's estimator weights allele contributions by  $1/p_{la}$ , giving more weight to rare alleles, and this estimator usually shows lower sampling variance, especially for unrelated individuals (Vekemans & Hardy 2004). Hence it is more powerful to detect genetic structure. However, it suffers downward bias as soon as one allele in the data set occurs at a low frequency (e.g. <5%). The estimator described in Loiselle *et al.* (1995) weights allele contribution by  $p_{la}(1 - p_{la})$  and does not suffer particular bias in the presence of low frequency alleles.

For dominant markers, *SPAGeDi* proposes an estimator of kinship defined in Hardy (2003) as well as an extension of this estimator for less reliable markers presenting a broad-sense heritability ( $H^2$ ) statistically significant but inferior to one (see Ley & Hardy 2013). To compute these estimators, the inbreeding coefficient must be given (the estimator is robust to moderate errors made on the assumed inbreeding coefficient), as well as estimates of  $H^2$  per marker when  $H^2 < 1$ . See Ley & Hardy (2013) for the procedure to estimate  $H^2$  for each marker and see §3.1 for specifying  $H^2$  estimates in the data file.

### 6.1.2. Relationship coefficient

#### Definition

Relationship coefficients can be defined as the proportion of gene copies in one individual with alleles identical to those of a reference individual (in several papers (e.g. Queller and Goodnight 1989), the so called "relatedness" coefficient is what is here called "relationship" coefficient). As for kinship coefficients, relationship coefficients depend on a reference population or on reference allele frequencies that can be specified (except for estimator 1° based on Moran's  $I$  statistic). Relationship coefficient is the  $r$  in Hamilton's (1964) famous rule for altruistic behaviour:  $rb > c$  ( $b$  = fitness benefit,  $c$  = fitness cost).

The expected value of the relationship coefficient ( $r_{ij}$ ) between two  $k$ -ploid individuals ( $i$  and  $j$ ) with inbreeding coefficient  $F$  can be expressed in term of the kinship coefficient ( $F_{ij}$ ):  $r_{ij} = F_{ij} k / (1 + (k-1)F)$ , reducing to  $r_{ij} = 2F_{ij}$  for two non-inbred diploids. However, contrary to the kinship coefficient, the relatedness coefficient is not always symmetric (i.e.  $r_{ij}$  and  $r_{ji}$  have not necessarily the same expectations), in particular when comparing individuals with different ploidy levels as in haplo-diploid organisms. Presently, *SPAGeDi* considers only symmetrical relatedness coefficients (for asymmetric coefficients, see the program *Relatedness* by Goodnight and Queller, at <http://www.bioc.rice.edu/~kfg/GSoft.html>).

One advantage of the relationship coefficient when investigating the genetic structure due to gene flow and drift, is that, at constant gene flow parameters, it is not influenced by the ploidy level or the selfing rate (Hardy and Vekemans 1999). Hence, it is useful to compare the level of genetic structuring among ploidy levels (Hardy and Vekemans 2001).

#### Estimators

For codominant markers, *SPAGeDi* proposes 5 estimators.

1°) A first estimator of the relationship coefficient is computed as the correlation between individual allele frequencies (e.g. for a diploid, frequencies can take the following discrete values: 0, 1/2, 1):

$$r_{ij} = \frac{\sum_a [ \sum_a (p_{ila} - p_{la})(p_{jla} - p_{la}) + \sum_a \text{Var}(p_{ila}) / (n_l - 1) ]}{\sum_a \text{Var}(p_{ila})}$$

with  $\text{Var}(p_{ila})$ , the variance of individual allele frequencies. The term  $\sum_a \text{Var}(p_{ila}) / (n_l - 1)$  is a sampling bias correction.

Averaging this estimator over distance classes give mean values of Moran's  $I$  statistic computed in the way proposed by Dewey and Heywood (1988) (Hardy and Vekemans 1999), except for the bias correction.

2°) A second estimator is defined in Queller and Goodnight (1989):

$$r_{ij} = \frac{\sum_l \sum_a \sum_c x_{lca} (p_{jla} - p_{la})}{\sum_l \sum_a \sum_c x_{lca} (p_{ila} - p_{la})}$$

*SPAGeDi* actually computes the average  $(r_{ij} + r_{ji})/2$ . Note that the estimator currently computed in *SPAGeDi* does not exclude related individuals to calculate  $p_{la}$  (a bias correction suggested by Queller and Goodnight 1989).

3°) Two additional estimators are defined in Lynch and Ritland (1999) and Wang (2002), respectively. These estimators can only be computed for diploids without inbreeding (genotypes in Hardy-Weinberg proportions). See these references for definitions and statistical properties in regard to other estimators. There is a sampling bias correction in Wang (2002) estimator which is not applied when reference allele frequencies are given.

4°) A fifth estimator is derived from Li et al. (1993) with a sample size correction by Wang (2002):

$$r_{ij} = \Sigma_l [\omega_l (S_{ijl} - S_{0l}) / (1 - S_{0l})] / \Sigma_l \omega_l$$

where  $S_{ijl}$  is the average proportion of alleles in  $i$  found in  $j$  and vice versa at locus  $l$  [ $S_{ijl} = 1$  (for  $i = aa$ ,  $j = aa$  or  $i = ab$ ,  $j = ab$ ),  $S_{ijl} = 0.75$  (for  $i = aa$ ,  $j = ab$ ),  $S_{ijl} = 0.5$  (for  $i = ab$ ,  $j = ac$ ),  $S_{ijl} = 0$  (for  $i = ab$ ,  $j = cd$ ), where  $a, b, c, d$  indicate alleles]

$$S_{0l} = 2[n_l[(\Sigma_a p_{la}^2) - 1] / (n_l - 1)] - [n_l^2(\Sigma_a p_{la}^3) - 3[n_l(\Sigma_a p_{la}^2) - 1] - 1] / [(n_l - 1)(n_l - 2)]$$

or  $S_{0l} = \Sigma_a [2p_{la}^2 - p_{la}^3]$  when reference allele frequencies are given (no sample size correction).

$\omega_l$  is the empirically determined locus weight defined as the inverse of the variance of single locus  $r_{ij}$  estimates over all  $i$ - $j$  pairs (Van de Casteel et al. 2001).

This estimator often shows a low variance (high precision) compared to other ones.

For dominant markers, *SPAGeDi* proposes an estimator defined in Hardy (2003) as well as an extension of this estimator for less reliable markers presenting a broad-sense heritability ( $H^2$ ) statistically significant but inferior to one (see Ley & Hardy 2013). To compute these estimators, the inbreeding coefficient must be given (the estimator is robust to moderate errors made on the assumed inbreeding coefficient), as well as estimates of  $H^2$  per marker when  $H^2 < 1$ . See Ley & Hardy (2013) for the procedure to estimate  $H^2$  for each marker and see §3.1 for specifying  $H^2$  estimates in the data file.

### 6.1.3. Fraternity coefficient

#### Definition

The fraternity coefficient,  $\Delta_{ij}$ , defined for two diploids ( $i$  and  $j$ ), is a function of the probability that the two gene copies of  $i$  are identical by descent to each of the gene copies of  $j$  (Lynch and Walsh 1998, Lynch and Ritland 1999), hence it depends on the states of all four gene copies. It can be expressed as a function of the kinship coefficients between the parents of  $i$  and  $j$ :  $\Delta_{ij} = F_{mi,mj} \cdot F_{fi,fj} + F_{mi,fj} \cdot F_{fi,mj}$ , where the subscripts  $mi, mj$  and  $fi, fj$  refer to the mother and father of  $i$  and  $j$ , respectively (Lynch and Walsh 1998). Hence, a positive  $\Delta_{ij}$  coefficient means there is a double genetic link between  $i$  and  $j$ .

The use of both 2-genes and 4-genes coefficients can help assessing the type of parentage relationship linking two individuals; for example, in a random mating population,  $\Delta_{ij} = 0, 0.25, 0$ , and  $F_{ij} = 0, 0.25, 0$ , for  $i$  and  $j$  being parent-offspring, full sibs, or half sibs, respectively (when parents constitute the “reference population”).

#### Estimators

Two estimators available are described in Lynch and Ritland (1999) and Wang (2002). These estimators can only be computed for diploids without inbreeding (genotypes in Hardy-Weinberg proportions). Practically, they perform well only with highly polymorphic loci (with at least 4 or 5 alleles). There is a sampling bias correction in Wang (2002) estimator which is not applied when reference allele frequencies are given.

### 6.1.4. Kinship type coefficient based on allele size $R_{ij}$

For microsatellite loci undergoing stepwise mutations, difference of allele sizes (not just the alleles identity vs non-identity information) contain information on coalescence time (Slatkin 1995). This information is taken into account in the  $R_{ij}$  coefficient, computed as an average correlation coefficient between allele sizes for homologous gene copies from two individuals (Streff et al. 1998). This coefficient is to kinship coefficient, what R-statistics are to F-statistics.

$$R_{ij} = \Sigma_l [ (\Sigma_{ci} \Sigma_{cj} (s_{lci} - s_l)(s_{lci} - s_l) / \Sigma_{ci} \Sigma_{cj} 1) + \text{Var}(s_l) / (n_l - 1) ] / \Sigma_l \text{Var}(s_l)$$

where  $s_{lci}$  is the size of the allele at locus  $l$  on chromosome  $c$  from individual  $i$ ,  $s_l$  is the mean allele size at locus  $l$  in the sample, and  $\text{Var}(s_l)$  is the variance of allele size in the sample. The term  $\text{Var}(s_l) / (n_l - 1)$  is a sampling bias correction (removed when reference allele frequencies are defined).

### 6.1.5. Kinship type coefficient based on distances between alleles $N_{ij}$

When alleles can be ordered so that a distance can be attributed for each pair of alleles (e.g. number of mutations separating two sequences), this phyletic distance contains information on coalescence time. This information is taken into account in the  $N_{ij}$  coefficient. This coefficient is to kinship coefficient what N-statistics are to G-statistics.  $N_{ij}$  is positive when gene copies of  $i$  and  $j$  show lower average phyletic distances than two gene copies sampled at random.

Single-locus estimator:

$$N_{ij}(l) = [\sum_{ci} \sum_{cj} (1 - \delta_{cij(l)}) / D(l)] / \sum_{ci} \sum_{cj} 1 \quad \text{where}$$

$\delta_{cij(l)}$  is the phyletic distance between the alleles found on chromosomes  $ci$  and  $cj$  at locus  $l$ ;

$D(l) = [n_l / (n_l - 1)] \sum_{a(l)} \sum_{a'(l)} p_{la} p_{la'} \delta_{aa'(l)}$  is the mean phyletic distance between two gene copies at locus  $l$

$\delta_{aa'(l)}$  is the phyletic distance between alleles  $a$  and  $a'$  at locus  $l$ ,  $p_{la}$  being the frequency of allele  $a$  at locus  $l$ .

The  $[n_l / (n_l - 1)]$  factor in  $D(l)$  is a sampling bias correction (removed when reference allele frequencies are defined).

Multi-locus estimator:

$$N_{ij} = \sum_l [D(l) N_{ij}(l)] / \sum_l D(l)$$

### 6.1.6. Rousset's distance measure

#### Definition

Rousset (2000) proposed a genetic distance measure between individuals ( $a$ ) analogous of the  $F_{ST} / (1 - F_{ST})$  ratio using pairs individuals instead of populations. In terms of probabilities of identity by state of gene copies (see §6.1.1.), this coefficient can be defined as  $a_{ij} = (Q_0 - Q_{ij}) / (1 - Q_0)$ , where  $Q_0$  refers to gene copies within individuals. The advantage of this distance measure over kinship coefficient is that it is not relative to a "reference" population (the distance is calibrated on the distance between gene copies within individuals). However, this measure is undefined for haploid organisms, and it is much dependent on the selfing rate. It also suffer higher sampling variance than kinship coefficients (Vekemans & Hardy 2004). Rousset (2000) showed that the slope of the regression of this estimator with the distance can be used to provide an estimate of gene dispersal distances (see §6.3.).

#### Estimator

The estimator is the one called  $\hat{a}$  in Rousset (2000).

## 6.2. STATISTICS FOR POPULATION LEVEL ANALYSES

For analyses at population level, **global** and **pairwise statistics** are computed. Global statistics are based on allele identity (F-statistics and  $G_{ST}$ ), microsatellite allele size (R-statistics), or the phylogenetic distances between alleles ( $N_{ST}$ ). Similarly, pairwise statistics are based on allele identity ( $F_{ST}$ ,  $Rho$ ,  $G_{ST}$ ,  $G_{ij}$  or  $D_s$ ), microsatellite allele size ( $R_{ST}$  or  $\delta\mu^2$ ), or the phylogenetic distances between alleles ( $N_{ST}$ ). Pairwise statistics are first computed for each pair of populations. Then they are regressed on pairwise spatial distances (regression analyses), and they are averaged over all pairs belonging to each predefined distance interval.

### 6.2.1. F-statistics and $G_{ST}$

#### Definition

F-statistics are based on allele identity and are types of kinship coefficients. In terms of probabilities of identity by state, they can be defined as  $F_{IT} = (Q_0 - Q_2) / (1 - Q_2)$ ,  $F_{IS} = (Q_0 - Q_1) / (1 - Q_1)$ , and  $F_{ST} = (Q_1 - Q_2) / (1 - Q_2)$ , where  $Q_0$ ,  $Q_1$ ,  $Q_2$ , refer to probabilities of identity of homologous gene copies within individuals, among individuals within population, and among individuals among populations, respectively. For global F-statistics,  $Q_2$  refers to all populations, whereas for pairwise  $F_{ST}$ ,  $Q_2$  refers only to the two populations being compared. Equivalently, these statistics can be defined as intra-class correlation coefficients of allelic states for gene copies within individuals

relative to all populations ( $F_{IT}$ ), gene copies within individuals relative to a population ( $F_{IS}$ ), and gene copies within populations relative to all populations ( $F_{ST}$ ).

### **Estimators**

For F-statistics, the estimation procedure is based on a nested ANOVA following Weir and Cockerham (1984), where populations are weighted according to their sample size.

$G_{ST}$  is an alternative estimator of  $F_{ST}$ , based on a decomposition of diversity indices following Pons and Petit (1996), where populations have equal weight, irrespective of the sample size.

Note that both  $F_{ST}$  and  $G_{ST}$  assume “random population effects” in statistical terms, contrary to Nei’s  $G_{ST}$  (not available in *SPAGeDi*) which assumes “fixed population effects”.

### **6.2.2. Rho statistic**

#### **Definition**

The *Rho* statistic is defined by Ronfort et al. (1998). It is to  $F_{ST}$  what the relationship coefficient is to the kinship coefficient (§6.1.2.), as it can be interpreted as an average relationship coefficient between individuals within population. *Rho* is equivalent to *Relat* in FSTAT software (Goudet 1995). This is a convenient statistic to compare the level of genetic structuring among ploidy levels (Ronfort et al. 1998). It relates to F-statistics in the following way: for a  $k$ -ploid,  $Rho = k \cdot F_{ST} / (1 + (k-1) F_{IT})$ , reducing to  $Rho = 2 F_{ST} / (1 + F_{IT})$  for a diploid.

#### **Estimator**

The estimator is computed as an intra-class correlation coefficient of individual allele frequencies, using an ANOVA framework (Ronfort et al. 1998) equivalent to that of Weir and Cockerham (1984) for F-statistics.

### **6.2.3. G<sub>ij</sub> statistic**

#### **Definition**

$G_{ij}$  is an average kinship coefficient between the individuals of two populations ( $i, j$ ), relative to a sample of populations (it expresses thus the genetic similarity rather than the distance between populations). It is equivalent the mean correlation coefficient between the allele frequencies of the two populations multiplied by the global  $F_{ST}$  among populations (Barbujani 1987).

#### **Estimator**

$$G_{ij} = 1 - h_{ij}/h_T$$

with  $h_{ij} = \sum_l [\sum_{a \neq a'} (p_{ila} p_{jla'})] / L$ , where  $p_{ila}$  is the frequency of allele  $a$  at locus  $l$  in population  $i$ ,  $L$  is the number of loci, and  $h_T$  is the average  $h_{ij}$  over all population pairs ( $i \neq j$ ).

### **6.2.4. D<sub>s</sub> – Nei’s standard genetic distance**

#### **Definition**

Nei’s (1972) standard genetic distance is a measure of genetic differentiation between two populations often used in phylogenetic reconstruction. Under an infinite allele model (IAM), its expected value is approximately  $D_s = 2\mu t$ , where  $\mu$  is the mutation rate, and  $t$  is the number of generations since population divergence.

#### **Estimator**

Nei’s unbiased estimate is computed according to Nei (1978).

### **6.2.5. R-statistics**

#### **Definition**

R-statistics are equivalent to F-statistics but based on allele sizes rather than allele identity (Slatkin 1995, Rousset 1996). They can be defined as intra class correlation coefficients of allelic sizes for gene copies within individuals relative to all populations ( $R_{IT}$ ), gene copies within individuals relative to a population ( $R_{IS}$ ), and gene copies within populations relative to all populations ( $R_{ST}$ ). They were developed for loci undergoing a

stepwise mutation process. Under a random mutation process (IAM, KAM), expectations for R-statistics are equivalent to corresponding F-statistics, but they suffer higher sampling variances (Balloux and Goudet 2002). To test for the impact of stepwise mutations on genetic structuring, R-statistics can be compared to the corresponding F-statistics estimated following Weir and Cockerham (1984) (see §4.3.5. and §5.6. for permutation tests).

### **Estimator**

R-statistics are estimated using a nested ANOVA (Michalakis and Excoffier 1996).

## **6.2.6. dm2 – Goldstein’s genetic distance**

### **Definition**

Goldstein et al. (1995) defined a distance  $\delta\mu^2$  comparable to Nei’s (1972) standard genetic distance but adapted for loci undergoing stepwise mutations (microsatellites). Under the stepwise mutation model, its expected value is approximately  $dm2 = 2\mu t$ , where  $\mu$  is the mutation rate, and  $t$  is the number of generations since population divergence.

### **Estimator**

The unbiased  $\delta\mu^2$  estimator is defined in Goldstein and Pollok (1997).

## **6.2.7. N<sub>ST</sub> statistic**

### **Definition**

$N_{ST}$  is an equivalent to  $F_{ST}$  (or  $G_{ST}$ ) but accounting for the phylogenetic distances between alleles (“ordered alleles”). To test for the impact of the allele phylogeny on genetic structuring,  $N_{ST}$  should be compared with  $G_{ST}$  (see §4.3.5. and §5.6. for permutation tests).

### **Estimators**

The  $N_{ST}$  estimator is described in Pons and Petit (1996).

## **6.2.8. N<sub>ij</sub> statistic**

### **Definition**

$N_{ij}$  is an equivalent to  $G_{ij}$  but accounting for the phylogenetic distances between alleles (“ordered alleles”). It can be useful to test for the impact of the allele phylogeny on genetic structuring, in which case  $N_{ij}$  should be compared to  $G_{ij}$ . An advantage of using  $N_{ij}$  and  $G_{ij}$  instead of  $G_{ST}$  and  $N_{ST}$  occurs when population show no or little genetic variation because  $G_{ST}$  and  $N_{ST}$  then reach the value of 1, irrespective of the phyletic relationships between populations.

### **Estimator**

$$N_{ij} = 1 - v_{ij}/v_T$$

with  $v_{ij} = \sum_l [\sum_{a \neq a'} (p_{ila} p_{jla} \delta_{aa'(l)})] / L$ , where  $p_{ila}$  is the frequency of allele  $a$  at locus  $l$  in population  $i$ ,  $\delta_{aa'(l)}$  is the phyletic distance between alleles  $a$  and  $a'$  at locus  $l$ ,  $L$  is the number of loci, and  $v_T$  is the average  $v_{ij}$  over all population pairs ( $i \neq j$ ).

### 6.3. INFERENCE OF GENE DISPERSAL DISTANCES

Theoretical models of isolation by distance show that, if some conditions are met, the kinship and relationship coefficients between individuals and the pairwise  $F_{ST}$ ,  $Rho$  and  $R_{ST}$  coefficients between populations are expected to vary approximately linearly (at least within some distance range) with the logarithm of the distance in a two-dimensional space, and with the linear distance in a one-dimensional space (Rousset 1997, 2000; Hardy and Vekemans 1999; Hardy 2003; Vekemans & Hardy 2004; for an application see e.g. Fenster et al. 2003). The slope of the corresponding regressions can be used to estimate gene dispersal distances in terms of a product between population density and mean squared distance of gene movements:

In a **two-dimensional space**, defining the *neighbourhood size* as  $Nb = 4\pi D\sigma^2$ , where  $D$  is the “effective” population density (i.e. taking into account the variance of reproductive success among individuals),  $\sigma^2$  is  $\frac{1}{2}$  the mean squared parent-offspring distance, and  $\pi = 3.1415$ ,  $Nb$  can be inferred in the following way for **diploids** using

- |                                    |                            |
|------------------------------------|----------------------------|
| (1) pairwise $F_{ST}/(1-F_{ST})$ : | $Nb \approx 1/blog$        |
| (2) Rousset’s $a$ coefficient:     | $Nb \approx 1/blog$        |
| (3) kinship coefficient:           | $Nb \approx -(1-F_N)/blog$ |

where  $blog$  is the regression slope of pairwise values on the logarithm of spatial distance, and  $F_N$  is the inbreeding coefficient. (2) and (3) are correct in the absence of selfing. With selfing, (2) is biased, but good estimates can be obtained from (3) if  $F$  is replaced by the kinship coefficient between adjacent individuals.

These relationships hold best when the regression ( $blog$ ) is computed within a restricted distance range which is approximately  $\sigma$  to  $20\sigma$ . At shorter distances, the details of the gene dispersal distribution (not just  $\sigma^2$ ) matter (Rousset 2001; Heuertz et al. 2003). At large distances, mutation rate can also matter. Hence an adequate geographic sampling scale is very important to achieve reliable estimates. *SPAGeDi* allows to **define a restricted distance range** to compute the regression slope (§ 4.3.3.).

For haploids or autopolyploids, the relationships can be generalized by defining the *neighbourhood size* as  $Nb = 2k\pi D\sigma^2$  where  $k$  is the ploidy level (Hardy & Vekemans 1999). Using maternally inherited cytoplasmic haplotypes as genetic marker, estimate of female (or seeds) dispersal can be obtained.

For analyses at the individual level and assuming a two-dimensional population at drift-dispersal equilibrium, *SPAGeDi* can use an **iterative procedure to determine  $\sigma$  and  $Nb$**  by regressing pairwise kinship coefficients on  $\ln(\text{distance})$  over a restricted distance range (§ 4.3.4.,  $\sigma$  is denote *sigma* in *SPAGeDi*). The procedure requires an estimate of the effective population density,  $D$ , as well as  $X$ , the width of the distance range  $\sigma$  to  $X\sigma$  over which the regression is applied. Starting from a global regression slope, the procedure consists in estimating  $Nb$  as  $Nb = -(1-F_{(1)})/blog$ , where  $F_{(1)}$  is the kinship coefficient between individuals for the first distance class (assumed to correspond to pairs of neighbours), and  $\sigma$  is estimated as  $\sigma = [Nb/(2\pi.k.D)]^{1/2}$ . Then, restricting the regression ( $blog$ ) to distances between  $\sigma$  and  $X\sigma$ ,  $Nb$  and  $\sigma$  are estimated again. This step is repeated until  $\sigma$  converges, with up to 100 iterations (Fenster et al. 2003; Vekemans & Hardy 2004; Hardy et al. 2006). Successive  $\sigma$  estimates are displayed on the screen. Convergence is not ensured, in which case no estimate is provided. The procedure can also cycle periodically around a set of values, in which case the average, as well as the minimal and maximal values over a cycle is reported. Approximate standard errors (SE) of  $\sigma$  and  $Nb$  estimates can also be obtained by jackknifing over loci (SE will be reported only if the iterative procedure always converge or cycle when a locus is removed). It is advise to select  $X$  between 10 and 100.  $X=10$  should result is less biased but potentially more imprecise estimates because regression might be based on a small number of pairs of individuals. Increasing  $X$  should improve precision but may result in more bias, notably under high mutation rates or when the structure is not yet at drift-dispersal equilibrium.

In a **one-dimensional space**, relationships (1), (2) and (3) also hold but with the regression slope based on linear distance instead of logarithmic distance, and with  $Nb$  defined as  $2.k.D\sigma^2$  where  $\sigma^2$  is the mean squared distance of gene dispersal (Rousset 1997). The iterative procedure to determine  $\sigma$  and  $Nb$  should not be used for one-dimensional populations.

Note that the conditions necessary for valid inferences might not be met in general within natural populations although simulation studies suggest that the method is fairly robust (Leblois 2003, 2004; see also Rousset 1997, 2000, 2001). In any case, when a relatively linear relationship is observed, the slope expresses the degree of genetic structuring and contains most of the information regarding intra-locus structure (Vekemans & Hardy 2004). *GENEPOP* (Rousset 2008) is an alternative software to conduct such inferences using Rousset’s  $a$  coefficient. It implements bootstrap procedure to get approximate confidence intervals.

#### 6.4. ESTIMATING THE ACTUAL VARIANCE OF PAIRWISE COEFFICIENTS FOR MARKER-BASED HERITABILITY AND $Q_{ST}$ ESTIMATES

Ritland (1996, 2000) proposed methods to estimate heritability and  $Q_{ST}$  using genetic markers. These methods require an estimate of the “actual” variance  $V$  (i.e. excluding sampling variance) of pairwise kinship coefficients between individuals (inference of heritability) or pairwise  $F_{ST}$  between populations (inference of  $Q_{ST}$ ):

$$V = \Sigma_p((\Sigma_l W_l R_{pl})^2 - \Sigma_l (W_l R_{pl})^2) / (1 - \Sigma_l W_l^2) / N - (\Sigma_p \Sigma_l W_l R_{pl} / N)^2$$

Where  $\Sigma_p$  stands for the sum over the considered pairs (i.e. within a distance class) of individuals or populations ( $N$  being the number of pairs),  $R_{pl}$  is the value of the pairwise statistic (kinship, relatedness,  $F_{ST}$ , ...) for pair  $p$  at locus  $l$ , and  $W_l$  is the locus  $l$  specific weight when computing multilocus  $R_p$  averages.

For heritability inference, the advantage of the approach is that quantitative characters can be measured in situ, avoiding the problem of having an heritability measure valid only for some experimental conditions. For  $Q_{ST}$  inference, the advantage of the Ritland’s approach is that there is no need to estimate the heritability of the characters.

The method to obtain the actual variance of pairwise coefficients follows Ritland (2000) and requires at least two loci. A jackknife procedure over loci (at least 3 loci necessary) provides approximate standard errors of the variance estimate. The estimates are given under request (§ 4.3.4.) for pairs of individuals or populations belonging to each distance interval as well as for all pairs (under “average”). Note that a large sample size and a high number of loci and/or very polymorphic loci are required for reliable heritability inference.

#### 6.5. TESTING PHYLOGEOGRAPHIC PATTERNS

A phylogeographic pattern occurs when gene copies sampled at nearby locations (e.g. within the same population) carry alleles that are more related on average than for gene copies sampled further apart. Under neutrality, such pattern is expected when the mutation rate is non negligible compared to the migration rate (Hardy et al. 2003).

Phylogeographic patterns can be tested only for ‘ordered’ alleles, such as microsatellites where differences in allele sizes informs on genetic distance (if stepwise mutations occur), or sequence data (or other multiple site polymorphisms at non recombinant DNA) where genetic distances between alleles can be attributed, for example as the number of mutations differentiating two alleles. *SPAGeDi* proposes several statistics that account for ‘ordered’ alleles, such as  $R_{ST}$  and  $(\delta\mu)^2$  for microsatellites, and  $N_{ST}$  for alleles for which a matrix of genetic distances is provided (cf. § 3.8.). A phylogeographic structure appears when  $R_{ST}$  or  $N_{ST}$  is significantly larger than  $F_{ST}$ .

Testing for a phylogeographic pattern can be done using  $R_{ST}$  by permuting allele sizes among alleles (Hardy et al. 2003), or using  $N_{ST}$  by permuting genetic distances among alleles (Burban et al. 1999; cf. § 4.3.5., § 5.6.). The expected value after such permutation is equal to an  $F_{ST}$  because only the structure due to allele identity remains (note that considering the statistical properties of the different estimators computed by *SPAGeDi*, the  $R_{ST}$  statistic should be compared with the  $F_{ST}$  statistic, whereas the  $N_{ST}$  statistic should be compared with the  $G_{ST}$  statistic). The permutation procedures permit to assess the distribution of  $R_{ST}$  or  $N_{ST}$  under the null hypothesis that there is no phylogeographic pattern. Therefore the unilateral test corresponding to the alternative hypothesis that the observed  $R_{ST}$  or  $N_{ST}$  is superior to the corresponding value after permutation should be considered.

Testing the global  $R_{ST}$  or  $N_{ST}$  (or the average pairwise values) tell us whether there is a **phylogeographic signal within populations**, answering the question: Are distinct alleles more related within populations than among populations?

Testing the slope (*b-lin* or *b-log* values) of pairwise  $R_{ST}$  or  $N_{ST}$  tell us whether there is a **phylogeographic signal among populations**, answering the question: Are distinct alleles more related between nearby populations than between distant populations?

## 7. CITED REFERENCES

- Balloux F, Goudet J (2002) Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Molecular Ecology* 11, 771-783.
- Barbujani G (1987) Autocorrelation of gene frequencies under isolation by distance. *Genetics* 117, 777-782.
- Burban C, Petit RJ, Carcreff E, Jactel H (1999) Rangewise variation of the maritime pine bast scale *Matsucoccus feytaudi* Duc. (Homoptera: Matsucoccidae) in relation to the genetic structure of its host. *Molecular Ecology* 8: 1593-1602.
- David P, Pujol B, Viard F et al. (2007) Reliable selfing rate estimates from imperfect population genetic data. *Molecular Ecology* 16: 2474-2487.
- Dewey, S. E., and J. S. Heywood, 1988. Spatial genetic structure in a population of *Psychotria nervosa*. I. Distribution of genotypes. *Evolution* 42: 834-838.
- Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.
- Fenster, C. B., X. Vekemans, and O. J. Hardy, 2003. Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (leguminosae). *Evolution* 57: 995-1007.
- Goldstein, D. B., A. R. Linares, M. W. Feldman and L. L. Cavalli-Sforza, 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics* 139: 463-471.
- Goldstein, D. B., and D. D. Pollock, 1997. Launching microsatellites: a review of mutation processes and method for phylogenetic inference. *Journal of Heredity* 88: 335-342.
- Goodnight, K. F., and D. C. Queller, . Relatedness ver 5.0: program available for downloading at <http://www.bioc.rice.edu/~kfg/GSoft.html> .
- Goudet, J., 1995. FSTAT (version 1.2): a computer program to calculate F-statistics. *Journal of Heredity* 86: 485-486.
- Hamilton, W. D., 1964. The genetical evolution of social behaviour. *J. Theor. Biol.* 7: 1-16.
- Hardy, O. J., 2003. Estimation of pairwise relatedness between individuals and characterisation of isolation by distance processes using dominant genetic markers. *Molecular Ecology* 12: 1577-1588.
- Hardy, O. J., 2015. Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Molecular Ecology Resources* (in press).
- Hardy, O. J., and X. Vekemans, 1999. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* 83: 145-154.
- Hardy, O. J., and X. Vekemans, 2001. Patterns of allozymic variation in diploid and tetraploid *Centaurea jacea* at different spatial scales. *Evolution* 55: 943-954.
- Hardy, O. J., and X. Vekemans, 2002. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* 2: 618-620.
- Hardy, O. J., N. Charbonnel, H. Fréville and M. Heuertz, 2003. Microsatellite allele sizes: a simple test to assess their significance on genetic differentiation. *Genetics* 163: 1467-1482.
- Hardy, O. J., L. Maggia, E. Bandou, P. Breyne, H. Caron, M-H. Chevallier, A. Doligez, C. Dutech, A. Kremer, C. Latouche-Hallé, V. Troispoux, V. Veron, B. Degen. 2006. Fine-scale genetic structure and gene dispersal inferences in ten neotropical tree species. *Molecular Ecology* 15: 559-571.
- Hardy, O. J., S. Vanderhoeven, M. De Loose and P. Meerts, 2000. Ecological, morphological and allozymic differentiation between diploid and tetraploid knapweeds (*Centaurea jacea* s.l.) from a contact zone in the Belgian Ardennes. *New Phytologist* 146: 281-290.
- Heuertz M, Vekemans X, Hausman J-F, Palada M, Hardy OJ (2003) Estimating seed versus pollen dispersal from spatial genetic structure in the common ash. *Molecular Ecology* 12, 2483-2495.
- Li CC, Weeks DE, Chakravarti A (1993) Similarity of DNA finger-prints due to chance and relatedness. *Human Heredity* 43, 45-52.
- Leblois R, Estoup A, and Rousset F. 2003. Influence of Mutational and Sampling Factors on the Estimation of Demographic Parameters in a "Continuous" Population Under Isolation by Distance. *Mol Biol Evol* 20: 491-502.
- Leblois R, Rousset F and Estoup A. 2004. Influence of Spatial and Temporal Heterogeneities on the Estimation of Demographic Parameters in a Continuous Population Using Individual Microsatellite Data. *Genetics* 166: 1081-1092.

- Ley A, and Hardy OJ. 2013. Improving AFLP analysis of large-scale patterns of genetic variation – a case study with the Central African lianas *Haumania* spp (Marantaceae) showing interspecific gene flow. *Molecular Ecology* 22:1984-1997.
- Loiselle, B. A., V. L. Sork, J. Nason and C. Graham, 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany* 82: 1420-1425.
- Lynch, M., and K. Ritland, 1999. Estimation of pairwise relatedness with molecular markers. *Genetics* 152: 1753-1766.
- Lynch, M., and B. Walsh, 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland.
- Michalakis, Y., and L. Excoffier, 1996. A genetic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* 142: 1061-1064.
- Nei, M., 1972. Genetic distance between populations. *American Naturalist* 106: 283-292.
- Nei, M., 1978. Estimation of average heterozygosity and genetic distance for small number of individuals. *Genetics* 89: 583-590.
- Nielsen R, Tarpay DR, Kern Reeve H. 2003. Estimating effective paternity number in social insects and the effective number of alleles in a population. *Molecular Ecology* 12: 3157–3164.
- Pons O, Petit RJ (1996) Measuring and testing genetic differentiation with ordered versus unordered alleles. *Genetics* 144, 1237-1245.
- Queller, D. C., and K. F. Goodnight, 1989. Estimating relatedness using genetic markers. *Evolution* 43: 258-275.
- Raymond, M., and F. Rousset, 1995. GENEPOP, ver. 1.2. A population genetic software for exact tests and eucumenism. *Journal of Heredity* 86: 248-249.
- Ritland, K., 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet. Res., Camb.* 67: 175-185.
- Ritland, K., 1996. A marker-based method for inferences about quantitative inheritance in natural populations. *Evolution* 50: 1062-1073.
- Ritland, K., 2000. Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology* 9: 1195-1204.
- Ritland, K., and C. Ritland, 1996. Inferences about quantitative inheritance based on natural population structure in the yellow monkeyflower, *Mimulus guttatus*. *Evolution* 50: 1074-1082.
- Ronfort, J., E. Jenczewski, T. Bataillon and F. Rousset, 1998. Analysis of population structure in autotetraploid species. *Genetics* 150: 921-930.
- Rousset, F., 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics* 142: 1357-1362.
- Rousset, F., 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics* 145: 1219-1228.
- Rousset, F., 2000. Genetic differentiation between individuals. *J. Evol. Biol.* 13: 58-62.
- Rousset F. 2001. Genetic approaches to the estimation of dispersal rates. In: *Dispersal* (eds. Clobert J, Danchin E, Dhondt AA, Nichols JD), pp. 18-28. Oxford University Press, Oxford.
- Rousset, F., 2002. Inbreeding and relatedness coefficients: what do they measure? *Heredity* 88, 371-380.
- Rousset, F. 2008. GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* 8, 103–106.
- Slatkin, M., 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139 1463-1463.
- Sokal, R. R., and F. J. Rohlf, 1995. *Biometry*. W. H. Freeman and Company, New York.
- Streiff, R., T. Labbe, R. Bacilieri, H. Steinkellner, J. Glössl *et al.*, 1998. Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Molecular Ecology* 7: 317-328.
- Van de Castelee, T., P. Galbusera, and E. Matthysen. 2001. A comparison of microsatellite-based pairwise relatedness estimators. *Molecular Ecology* 10:1539-1549.
- Vekemans, X, and O. J. Hardy. 2004. New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* 13: 921-935.
- Wang, J., 2002. An estimator for pairwise relatedness using molecular markers. *Genetics* 160: 1203-1215.

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

## 8. BUG REPORTS

### Ver. 1.1

- Two computational bugs occurred in version 1.1 and earlier and were corrected in version 1.1b :
  - 1° Bug leading to erroneous jackknife estimates (mean and standard error) for Wang(2002) relationship estimator when missing data occur.
  - 2° Bug leading occasionally to erroneous estimates for the Rousset's a distance between individuals for the last locus, and consequently for the multilocus and jackknife estimators.

### Ver. 1.2

- A bug causing occasional crash when reading the file with reference allele frequencies has been corrected in version 1.2b (released on 11 Nov 2005).
- When selecting both options "Test of mutation effect on genetic differentiation for each population pair" and "Report only P-values", a bug caused inconsistent results regarding these tests in the output file. It has been corrected in version 1.2c (released on 23 Dec 2005).
- WARNING: For analyses at the individual level, when selecting the "allele size correlation coefficient" (Streiff et al. 1998), a problematic bug caused erroneous multilocus and jackknife estimates for this statistic (single-locus estimates were correctly computed). The bug has been corrected in version 1.2d (released on 17 Jan 2006) but all previous versions are affected !
- WARNING: For analyses at the individual level using dominant markers (kinship or relationship coefficient), a bug could cause the program to crash (it may also have lead to inconsistent results). The problem was resolved in version 1.2e.
- Two bugs causing occasional crash when reading the file with reference allele frequencies and when writing down results have been corrected in version 1.2f (released on 11 Apr 2007).
- Bugs occurring when importing large data sets in Genepop format have been corrected in version 1.2g (released on 24 Apr 2007).

### Ver. 1.3

- WARNING: For analyses at the individual when asking to restrict pairwise comparison within categories (option 5 followed by option 2), a bug could cause the program to crash or, on some computers, it may lead to biased mean kinship. The problem was resolved in version 1.4.

### Ver. 1.4

- In version 1.4, a bug occurred when the number of loci was larger than the number of populations. The problem was resolved in version 1.4b.
- For dominant markers, bugs could occur in versions 1.4 and 1.4b when trying to read the last line with H2 estimates. The problem was resolved in version 1.4c.

## FINAL NOTE

The program is regularly modified for further improvements. Any suggestion for improvement is welcome. Also if you have trouble with some data sets you can send us the data set by e mail and we could try to fix the problem.

Good luck!