

SPADS 1.0

>> [manual & tutorials \(22-11-16\)](#)

SPADS 1.0 is available free from
ebe.ulb.ac.be/ebe/Software.html

Contact and bugs report:
Simon.Dellicour@ulb.ac.be

S. Dellicour & P. Mardulyn (2014)

Evolutionary Biology & Ecology, Université Libre de Bruxelles, Av. FD Roosevelt 50, 1050 Brussels, Belgium

Dellicour & Mardulyn, 2014, Evolutionary Biology & Ecology
Manual and example files: ebe.ulb.ac.be/ebe/Software.html

Number of user-defined groups:

Number of loci:

Summary statistics on user-defined populations:

- total number of haplotypes
- G_{st} , N_{st} (Pons & Petit, 1995, 1996)
Number of permutations:
- Φ_{st} estimator for $K=1$ (Excoffier et al, 1992)
Number of permutations:
- IBDSC (isolation by distance slope coefficient)
- $m\Phi_{stdgeo}$ (new statistic for PhyloGeoSim)

Summary statistics on user-defined groups:

- X_h : number of haplotypes ratios
- P_i : nucleotide diversities (Nei & Li, 1979)
- P_{ir} (Mardulyn et al, 2009)
- allelic richness (El Mousadik & Petit, 1996)
- pairwise Φ_{st} 's (Excoffier et al, 1992)
Number of permutations:
- AMOVA Φ -statistics (Excoffier et al, 1992)
Number of permutations:

Input files conversion:


- SPAGeDi (Hardy & Vekemans, 2002)
- Structure (Pritchard et al, 2000)
- BAPS (Corander et al, 2003, 2008)
- Geneland (Guillot et al, 2005)
- GDisPAL and GDivPAL functions

SAMOVA analysis (Dupanloup et al, 2002):
Number of groups (K): to
Number of iterations:
Number of repetitions:

Monmonier algorithm (Manni et al, 2004):
Number of barriers (B): to

Inputs folder:

Inputs name:



SPADS 1.0 (for “Spatial and Population Analysis of DNA Sequences”) is a population genetics toolbox computing several summary statistics from populations or groups of populations, performing several input file conversions for other population genetics programs and implementing two clustering algorithms to study the genetic structure of populations. The toolbox also includes R functions to represent distance and diversity patterns across landscapes. SPADS has been specifically developed for the analysis of multi-locus datasets of DNA sequences.

Table des matières

1. Methods implemented in SPADS.....	3
1.1. <i>Computation of summary statistics.....</i>	3
1.1.1. Summary statistics based on populations	3
1.1.2. Summary statistics based on groups of populations.....	3
1.2. <i>Clustering analysis</i>	4
1.3. <i>Input file conversions</i>	5
1.4. <i>GDisPAL and GDivPAL functions.....</i>	5
2. Input files.....	8
3. How to run the program	11
4. Output files.....	12
5. Tutorials.....	13
5.1. <i>Tutorial 1: analysing population structure on a simulated dataset</i>	13
5.2. <i>Tutorial 2: GDisPAL and GDivPAL functions on a bee (C. hederæ) dataset.....</i>	17
6. SPADZ1 and SPADZ2	19
7. Software limitations.....	20
8. Toolbox availability	21
9. Version history	21
10. References	21

1. Methods implemented in SPADS

1.1. Computation of summary statistics

SPADS computes several summary statistics for each locus, based on user-defined populations or groups of populations.

1.1.1. Summary statistics based on populations

- total number of haplotypes: number of different sequences detected for each locus.
- global G_{ST} estimator (Pons & Petit, 1995) of populations differentiation.
- global N_{ST} estimator (Pons & Petit, 1996) of populations differentiation.
- AMOVA Φ_{ST} estimator for $K=1$ (Excoffier *et al*, 1992).
- *IBDSC*: isolation by distance slope coefficient. This is the slope coefficient of the linear regression estimated from $y = f(\ln(x))$ with $y = (\Phi_{ST}/(1-\Phi_{ST}))$ (Rousset, 1997).
- $m\Phi_{ST}d_{geo}$: the average of ratios between Φ_{ST} estimators and geographical distances between all pairwise populations.

$$m\Phi_{ST}d_{geo} = \frac{2(p-2)!}{p!} \sum_{j_1 \neq j_2} \left(\frac{\Phi_{ST_{j_1 j_2}}}{d_{j_1 j_2}} \right)$$

with:

- p , the number of populations.
- $\Phi_{ST_{j_1 j_2}}$, Φ_{ST} between populations j_1 and j_2 .
- $d_{j_1 j_2}$, geographical distance between populations j_1 and j_2 .

Statistical tests for the significance of three F -statistics (G_{ST} , N_{ST} and AMOVA Φ_{ST} for $K=1$) are based on random permutations of individuals between populations, while the statistical test for the significance of the difference between N_{ST} and G_{ST} (highlighting the extent of the phylogeographical signal) is based on random permutations of haplotypes (Hardy & Vekemans, 2002). Corresponding p-values are the proportions of permuted datasets with a F -statistic value higher or equal to the value estimated for the real dataset.

1.1.2. Summary statistics based on groups of populations

- X_H : ratio between the number of haplotypes in a user-defined group of populations and the total number of haplotypes in the dataset.
- π : nucleotide diversity (Nei & Li, 1979) within each user-defined group of populations.

$$\pi = \frac{2(n-2)!}{n!} \sum_{i=1}^{n-1} \sum_{i'=i+1}^n k_{ii'}$$

with:

- $k_{ii'}$, number of differences between sequences i and i' .
- n , number of sequences in the considered user-defined group.
- π_R : computed for each user-defined group of populations, this is the ratio between the nucleotide diversity within the user-defined group of populations and the nucleotide diversity within the virtual group formed by all other populations (Mardulyn *et al*, 2009).
- A_R : estimator of allelic richness within each user-defined group of populations (El Mousadik & Petit, 1996).
- Φ_{SC} , Φ_{ST} , Φ_{CT} : AMOVA Φ -statistics (Excoffier *et al*, 1992) computed for the population structure linked to the user-defined groups.

- Pairwise Φ_{ST} : pairwise AMOVA Φ_{ST} (Excoffier *et al*, 1992) computed between all the user-defined groups.

Statistical tests for the Φ -statistics (Φ_{SC} , Φ_{ST} and Φ_{CT}) are based on random permutations. The kind of permutations depends on the Φ -statistic tested (Excoffier *et al*, 1992): permutations of sampled sequences across populations but within the same group for Φ_{SC} , permutations of sampled sequences across populations (without regard to their original group) for Φ_{ST} and permutations of whole populations across groups for Φ_{CT} . Corresponding p-values are the proportion of permuted datasets with a Φ -statistic value higher or equal to the value estimated for the real dataset.

Note: for the F -statistics computations (G_{ST} , N_{ST} , Φ_{ST} for $K=1$, and AMOVA Φ -statistics based on user-defined groups), when more than one locus are specified, SPADS also automatically estimates multilocus weighted averages for these statistics. For each locus taken separately, a given F -statistic is always the ratio between inter-population diversity (numerator) and total diversity (denominator). The multilocus estimate is computed as the ratio of the sum of locus-specific numerators to the sum of locus-specific denominators, as suggested by Weir & Cockerham (1984).

1.2. Clustering analysis

SPADS implements two clustering methods to define groups of populations a posteriori from genetic data:

- (1) a locus-by-locus SAMOVA analysis: the algorithm is similar to the one implemented in the software SAMOVA (Dupanloup *et al*, 2002). It analyses one locus at a time. One difference with the software SAMOVA is that users can choose the number of “iterations” performed for each run of the algorithm. In the SAMOVA algorithm, the number of iterations is automatically set to 10,000. The number of “iterations” corresponds to the number of repetitions of steps 5 to 9 of the SAMOVA algorithm (Dupanloup *et al*, 2002) and each SAMOVA run is started with a different initial partition of populations.
- (2) a locus-by-locus Monmonier algorithm: the Monmonier algorithm (Monmonier, 1973) similar to the one implemented in the BARRIER software (Manni *et al*, 2004). This method treats each locus separately.

In addition, SPADS also offers a multi-loci version of these two methods:

- a multi-loci SAMOVA analysis*: instead of performing independent locus-by-locus analysis for each assumption of the number of groups (K), this algorithm uses all the available loci in one analysis. While the locus-by-locus SAMOVA uses the Φ_{CT} estimator (Excoffier *et al*, 1992) to compare two successive iterations, the multi-loci SAMOVA computes a multilocus weighted average Φ_{CT} (computed as the ratio of the sum of locus-specific Φ_{CT} numerators to the sum of locus-specific Φ_{CT} denominators). **Important note:** this method requires that for each locus, there is at least one sequence available in each sampled population.
- a multi-loci Monmonier algorithm*: instead of performing independent locus-by-locus analyses for each assumption of the number of barriers (B), this algorithm uses all the available loci in the same analysis. While the locus-by-locus Monmonier algorithm uses the pairwise Φ_{ST} estimator (Excoffier *et al*, 1992) to choose positions of barriers to construct between sampled populations, the multi-loci version computes multilocus weighted average Φ_{ST} estimators (computed as the ratio of the sum of locus-specific Φ_{ST} numerators to the sum of

locus-specific Φ_{ST} denominators). **Important note:** this method requires that for each locus, there is at least one sequence available in each sampled population.

(*) WARNING: while we have tested, based on computer simulations, that a “multi-loci” analysis with the SAMOVA and Monmonier algorithms were able to identify the clusters or barrier implemented in a few simple models, we have not thoroughly tested the performances of these methods. Users wishing to use these multi-locus versions are thus strongly advised to compare their results with those obtained with a locus-by-locus analysis, to check whether all loci agree or whether they lead to contradictory results. Also, to test the performances of the method in the conditions of a specific study, it could be useful to perform simulations of DNA sequences in a geographic setting similar to the studied area (e.g. using CDPOP, Landguth & Cushman 2010, or PHYLOGEOSIM, Dellicour *et al*).

Note: contrary to the automated version of the Monmonier algorithm implemented in SPADS, the software BARRIER generates helpful graphical outputs, like the Voronoï polygonation used by the algorithm and the different inferred barriers. When selecting this clustering method in SPADS, the program will produce input files based on multi-loci information that can be read by the software BARRIER of Manni *et al* (2004) to generate these graphical outputs.

1.3. Input file conversions

SPADS can be used to construct input files based on multiple DNA sequence alignments for the population genetics programs SPAGeDi (Hardy & Vekemans, 2002), STRUCTURE (Pritchard *et al*, 2000), BAPS (Corander *et al*, 2003, 2004, 2008), GENELAND (Guillot *et al*, 2005a, 2005b, 2008, 2012; Guedj & Guillot, 2011) and BARRIER (Manni *et al*, 2004). Furthermore, SPADS can also create several input files for the GDisPAL and GDivPAL R functions (see below).

STRUCTURE, BAPS and GENELAND all propose clustering methods based on allelic frequencies (Hardy-Weinberg and linkage equilibriums). While STRUCTURE includes an option allowing to define sampling groups *a priori* in the case of weak population structure (Hubisz *et al*, 2009), BAPS and GENELAND implement algorithms that take the geographical information into account to infer population clusters. More recently, Cheng *et al* (2013) extended the spatially explicit BAPS model for clustering DNA sequence data. For BAPS, SPADS can then create two distinct BAPS inputs: (i) a BAPS input for the spatially explicit clustering method based on allelic frequencies (“BAPS_myDataSet_input_genotypes.txt”) and (ii) a BAPS input for the spatially explicit clustering method based on DNA sequences (an MLST¹ Excel file: “BAPS_myDataSet_input_MLST.xls”).

1.4. GDisPAL, GDivPAL functions

In addition to the Java executable SPADS 1.0, we also included in this toolbox two R functions implementing an extension of the method initially developed by Miller (2005) to represent patterns of inter-individual genetic distances across a species distribution. The method of Miller (2005) is based on a connectivity network (e.g. a Delaunay triangulation) built from the sampling localities. In this method, inter-individual genetic distances are then estimated and assigned to landscape coordinates at midpoints of each connectivity network edge. Finally, an interpolation procedure (i.e. an inverse distance-weighted interpolation; Watson & Philips, 1985; Watson, 1992) is used to infer genetic

¹ MLST format: format as applied in multilocus sequence typing (MLST) databases.

distances at locations on a uniformly spaced grid. Here, we proposed an extension of this interpolation method in order to use any different measures of genetic distances and, furthermore, any different measures of genetic diversity. In the case of diversity measures, the interpolation procedure is not based on (distance) values assigned at midpoints of each edge of a connectivity network but on (diversity) values directly estimated at each sampling point here defined as a “sampled population” or “population”. These interpolation methods are implemented in the two R (R Development Core Team, 2016) functions: GDisPAL for “genetic distance patterns across landscapes” and GDivPAL for “genetic diversity patterns across landscapes”.

SPADS can build different input files for this interpolation method:

- “GDisPAL_myDataSet_input_coordinates.txt”;
- “GDisPAL_myDataSet_input_distances_matrix_IID1.txt”;
- “GDisPAL_myDataSet_input_distances_matrix_IID2.txt”;
- “GDisPAL_myDataSet_input_log(10)_pseudoslopes_matrix_IID1.txt”;
- “GDisPAL_myDataSet_input_log(10)_pseudoslopes_matrix_IID2.txt”;
- “GDisPAL_myDataSet_input_regression_residuals_matrix_IID1.txt”;
- “GDisPAL_myDataSet_input_regression_residuals_matrix_IID2.txt”;
- “GDivPAL_myDataSet_input_Ar.txt”;
- “GDivPAL_myDataSet_input_Pi.txt”;
- “GDivPAL_myDataSet_input_Pir.txt”.

IID1 (for “inter-individual distance 1”) refers (i) to the inter-individual distance defined by Miller (2005) for diploid individuals:

$$IID1 = \frac{\sum_{l=1}^L (1 - \sum_{a=1}^{A_l} \sqrt{p_{ia}p_{ja}})}{L}$$

with:

- L , the number of different loci.
- A_l , the number of different alleles at locus l .
- p_{ia} and p_{ja} , the relative frequencies of allele a in individuals i and j .

(ii) or to the inter-individual distance defined by Miller (2005) for haploid individuals:

$$IID1 = \frac{\sum_{l=1}^L d_l}{L}$$

with:

- d_l , a distance value equals to 1 if individuals i and j have different alleles at locus l and equals to 0 if individuals i and j have the same allele at locus l .

IID2 (for “inter-individual distance 2”) is similar to the inter-individual distance defined by Miller (2005) for DNA sequences but averaged over the different loci:

$$IID2 = \frac{\sum_{l=1}^L (\sum_{m=1}^{M_l} d_{lm} / M)}{L}$$

with:

- M_l , the length (in bp) of the locus l .
- d_{lm} , a distance value equals to 1 if individuals i and j have a different nucleotide at site m of locus l and equals to 0 if individuals i and j have the same nucleotide at site m of locus l .

When there is a significant correlation between genetic and geographical distances, Miller *et al* (2006) advise to follow the recommendation of Manni *et al* (2004) by using residual genetic distances derived from the linear regression of genetic against geographical distances. These regression residuals are computed by SPADS and available in “GDisPAL_myDataSet_input_regression_residuals_matrix_IID1/IID2.txt” files. Another way to deal with a correlation between genetic and geographical distances is to use “pseudoslopes” that Miller (2005) defined as the quotient of congruent elements from the genetic and geographical distance matrices. The logarithms of these “pseudoslopes” distances to base 10 are also computed by SPADS and available in “GDisPAL_myDataSet_input_log(10)_pseudoslopes_matrix_IID1/IID2.txt” files. Beside one of these distance matrices, the GDivPAL always requires the sampling of geographic coordinates of the corresponding individuals. These coordinates are ordered and given in “GDisPAL_myDataSet_input_coordinates.txt”.

SPADS can create three distinct inputs to map the genetic diversity pattern across the landscape using the GDivPAL function. The difference between these inputs simply lies in the summary statistic used to measure diversity. The three different statistics proposed by SPADS are: (i) the estimator of allelic richness A_R calculated within each population (El Mousadik & Petit, 1996), (ii) the nucleotide diversity π (Nei & Li, 1979) of each population and (iii) the relative nucleotide diversity π_R (Mardulyn *et al*, 2009) of each population.

Interpolation surfaces (heat maps or 3-dimensional graphs) can then be generated with the two R functions: “GDisPAL” and “GDivPAL”. With these two functions, the inverse distance interpolation parameter a can be set to different values. We advise users to explore the effect of this parameter on the shape of the interpolations. In GDisPAL, interpolations are all based on a Delaunay triangulation network. See the second tutorial (section 5.2) at the end of this manual for further details over the use of these functions. More recently, we also added the possibility to perform a preliminary “sliding window” approach based on nucleotide diversity values (developed for the study of Lecocq *et al*, *submitted*). This preliminary step can be performed with the R function “slidingWindowPi” (see the third tutorial, section 5.3, for further details).

2. Input files

Running SPADS requires three types of input files:

- (1) the DNA sequence matrices: one sequential Phylip (Felsenstein, 2004) file per locus. These files have to be named as follows:

```
myDataSet_locus1.phy
myDataSet_locus2.phy
myDataSet_locus3.phy
...
```

Only the beginning of the file name (in this example: “myDataSet”) may be modified, and must be defined in the “input file name” field of the SPADS interface. Sequential Phylip format can be created manually or generated by many programs, e.g., DnaSP 5 (Librado & Rozas, 2009) or MEGA 5 (Tamura *et al*, 2011).

```
80 800
1-1 CCGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...
2-1 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGGTTGAA ...
3-1 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...
4-1 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...
5-2 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTACAGGGTTAAGATTGAA ...
6-2 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGAA ...
7-3 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGATTGAA ...
8-3 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGAA ...
9-3 CCGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGAA ...
10-3 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGATTGAA ...
...
```

n Figure 1: example of a sequential Phylip file format.

The Phylip format requires the first line to contain the number of sequences (here “80”) and the sequence length (here “800” b.p.). For SPADS, each sequence name must contain the individual and population Ids, separated by a dash. In this example (Figure 2), the fifth sequence name “5-2” refers to individual n°5 sampled in population 2. Users are free to choose individuals and populations IDs, using letters or numbers (e.g., 05-bru could refer to the fifth individual sampled in Brussels). The names of the sequences can be modified in a text editor, but it is important to note that Phylip format requires exactly 10 characters (including spaces) before the beginning of each sequence. Furthermore, SPADS also requires that the name of the sequence and the sequence itself are separated by at least one space. As a consequence, the name of the sequence cannot be longer than 9 characters.

Notes about ambiguous nucleotides: (1) any character in the sequence that is not “A”, “C”, “G”, “T”, “a”, “c”, “g” or “t” will be considered ambiguous by SPADS. Ambiguous nucleotides are treated as missing data and are thus not taken into account when comparing a given pair of DNA sequences. This means that two sequences differing only by ambiguous nucleotides will be treated as identical haplotypes. (2) SPADS can handle missing sequences at one or more loci. For a given individual, if sequence information is lacking for an entire locus, it should simply be omitted from the corresponding data set.

For the input file conversions, users can specify the level of ploidy and which sequences originated from the same individuals. As displayed in Figure 2, the ploidy level needs to be added as a third parameter at the end of the first line (i.e. after the number of sequences and the length of the locus). In the case of a ploidy level higher than 1, sequences originated from the same individual are required to display the same individual name followed by a dot "." and an integer corresponding to the sequence number. The sequence number itself has little importance but needs to be different than that of other sequences from the same individual. For example in Figure 2, individual "3" has two sequences ("3.1" and "3.2") and was sampled in population "1", which results in sequence names "3.1-1" and "3.2-1".

```

80 800 2
1.1-1 CCGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...
1.2-1 CCGAGCCGATTTGATGACAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...
2.1-1 CTGACCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGGTTGTA ...
2.2-1 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGGTTGAA ...
3.1-1 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...
3.2-1 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...
4.1-2 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...
4.2-2 CTGACCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...
5.1-2 CTGAGCCGATTTGATGACAGGCAAGCATTAGGATTACAGGGTTAAGATTGAA ...
5.2-2 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTACAGGGTTAAGATTGAA ...
6.1-2 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGTA ...
6.2-2 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGAA ...
...

```

Figure 2: example of a Phylip file containing sequences from diploid individuals. The ploidy level (2) is the third parameter of the first line added after the number of sequences (80) and the size of sequences (800 b.p.).

Note about the ploidy level: SPADS only uses the ploidy level and the individual information when generating input files for other programs. For summary statistics computation and clustering analyses, SPADS will not read the third parameter on the first line of the file and will not consider the character "." as a separator between the individual name and the sequence number, thus interpreting for example "3.1-1" as one independent sequence with name "3.1" sampled in population "1".

- (2) the "myDataSet_populations.txt" file: this text file contains the different populations IDs followed by their geographical coordinates (latitude followed by longitude). Only the first part of the name (in this example: "myDataSet") can be customized, and must be defined in the "input file name" field of the SPADS interface (Figure 1). Each line corresponds to a population and contains three data separated by single spaces: the population ID, followed by the latitude and longitude of the population in decimal degrees (Figure 3). Population IDs (number or name) must be identical to those used in the Phylip input file. The creation of this file is compulsory. Here is an example:

```

Pop1      50.457540    4.746462
Pop 2     48.108268   -0.701022
Pop 3     44.686060    3.228962
Pop 4     46.921486    9.618748
Pop 5     41.271108   -6.023386
Pop 6     39.263063    1.197266
Pop 7     52.145201   15.878675
...

```

Figure 3: example of a "myDataSet_populations.txt" input file.

WARNING: if some adjacent populations have coordinates forming a perfect square on the map, the Delaunay triangulation algorithm will have to arbitrary choose between the two possible triangulations inside the square. Users also need to avoid populations with exactly the same geographical coordinates.

- (3) the “myDataSet_groups.txt” file: this text file contains the different groups IDs followed by the IDs of populations included in each group. Only the first part of the name (in this example: “myDataSet”) may be changed, and must be defined in the “input file name” field of the SPADS interface (Figure 1). These groups are thus defined by the user and used for the computation of some summary statistics (X_H , π , π_R , A_R and AMOVA Φ -statistics). Here is an example:

GroupA	Pop3 Pop4 Pop5 Pop6 Pop7 Pop10 Pop11
GroupB	Pop1 Pop2 Pop12 Pop13 Pop14
GroupC	Pop15
GroupD	Pop16 Pop17 Pop18
...	

Figure 4: example of a "myDataSet_groups.txt" input file.

Each line must correspond to a group and begin with the group ID (name or number), followed by some populations IDs, separated by single spaces. For example, the second line in Figure 4 refers to group “B” which contains populations 1, 2, 12, 13 and 14.

Note about overlapping groups: except for the AMOVA analysis, SPADS allows the user to specify overlapping groups (i.e. groups having at least one population in common). As a result, if one or more populations are assigned to more than one group, all statistics based on user-defined groups will be estimated but the AMOVA analysis will be cancelled (if selected) and SPADS will display a warning message.

In addition to these three compulsory input files, two optional input files can be analysed by SPADS:

- (4) “myDataSet_monmonier_distances.txt”: a text file containing a matrix of pairwise genetic (or other type of) distances among all sampled populations, with columns and lines in the same order than in the “populations” file. This matrix is only for the barrier inference with the Monmonier algorithm and thus allows users to specify their own kind of distances between populations. Otherwise, the default distances used by the Monmonier algorithms are pairwise Φ_{ST} measures computed from DNA sequences.
- (5) “myDataset_population_distances.txt”: matrix of pairwise geographic distances among all sampled populations, with columns and lines in the same order than in the “populations” file. By default (in the absence of this external geographical distance matrix), SPADS computes its own matrix of spatial distances based on Euclidian distance among populations. However, Euclidian distances are not suitable at all when considering distant points on the globe. In such cases, corrected distance matrices are sometimes necessary. Such corrected matrices can easily be computed with software like Geographic Distance Matrix Generator (Ersts 2012).

3. How to run the program

A double click on the program file will prompt the program interface. Input files (the Phylip files, the “populations.txt” file and the “groups.txt” file) must be located in the same folder. If this folder is not the one containing the “SPADS.jar” program file, users have to click on the “open” button to select the directory where input files are located. Before starting the run, the following information needs to be specified from the user interface:

- the summary statistics the user wants to estimate. For the F -statistics (global G_{ST} , N_{ST} , Φ_{ST} and AMOVA Φ -statistics), the user must define the number of permutations to perform for the computation of the p-value associated with statistical tests.
- the number of groups of populations. This number can be set to “0” if the user does not wish to estimate the summary statistics based on these user-defined groups. If bigger than “0”, this number must correspond to the number of lines in the “groups.txt” file.
- The number of loci. This number must correspond to the number of Phylip files.
- The range of K (number of SAMOVA groups) values to explore. The minimum ($Kmin$) and maximum ($Kmax$) K values must be specified. The program will perform a locus-by-locus and a multi-loci SAMOVA algorithm for each value of K contained in this interval. For example, if the user sets $Kmin = 3$ and $Kmax = 6$, the program will perform four locus-by-locus SAMOVA and four multi-loci SAMOVA analyses: for $K = 3$, $K = 4$, $K = 5$ and $K = 6$. If the user does not wish to perform SAMOVA analyses, $Kmin$ and $Kmax$ must be set to 0. If the number of loci is set to “1”, SPADS will only perform a locus-by-locus SAMOVA analysis.
- The number of SAMOVA iterations to perform (cfr. 2. Methods implemented in SPADS). This field will not be taken into account if $Kmin = 0$ and $Kmax = 0$ for the SAMOVA analysis.
- The number of SAMOVA runs to perform for each value of K explored (cfr. 2. Methods implemented in SPADS). This field will not be taken into account if $Kmin = 0$ and $Kmax = 0$ for the SAMOVA analysis.
- The number of barriers (B) to construct with the locus-by-locus and multi-loci Monmonier algorithms. Users must give the minimum ($Bmin$) and maximum ($Bmax$) numbers of barriers to construct. The program will perform a Monmonier algorithm for each value of B contained in this interval. If users do not wish to perform Monmonier algorithm analyses, $Bmin$ and $Bmax$ should both be set to 0.
- The user-defined part of the input file name, which corresponds to the beginning of all Phylip files (cfr. 2. Input files).

4. Output files

SPADS creates at least two output files:

- (1) A text file containing the results of all analyses. The name of this file is "SPADS_myDataSet_results.txt".
- (2) A messages file ("SPADS_myDataSet_messages_file.txt") containing the following additional information:
 - Parameter values defined by the user.
 - The user-defined groups pairwise Φ_{ST} matrices (one per locus).
 - The mismatch distributions (one per locus).
 - The Delaunay triangulation used to perform the SAMOVA and Monmonier algorithms.
 - The positions of the barriers constructed by the different SAMOVA and Monmonier algorithms (locus-by-locus and/or multi-loci) for each K - or B -assumptions.
 - The populations pairwise Φ_{ST} matrices used by the locus-by-locus Monmonier algorithm.

In addition SPADS optionally creates the input files for SPAGeDi, STRUCTURE, BAPS, GENELAND and/or GDisPAL-GDivPAL functions. The text blocks corresponding to these input files (e.g. list of individuals with genotypes and/or populations coordinates) are also added in the messages file.

WARNING: SPADS automatically overwrites output files with the same name. Users must change the names of their output files if they want to keep them before launching a second run of the program.

5. Tutorials

5.1. Tutorial 1: analysing population structure on a simulated dataset

Dataset: “Example 1 (simulated)”, available on the toolbox website (see below).

Description of the dataset: includes sequences for three loci (800, 350 and 1000 pb) simulated with the software PHYLOGEOSIM 1.0 (software freely available, along with a detailed manual, at <http://ebe.ulb.ac.be/ebe/Software.html>), that implements a spatially explicit model of coalescence. The sequences were simulated under a history of geographic fragmentation, separating two groups of populations: a “North” and a “South” group. Four populations were sampled (10 sequences/population) for each group. Northern populations (Pop1, 2, 3 and 4) were separated (no migration) from southern populations (Pop5, 6, 7 and 8) for 100,000 generations. We analyse this dataset to show an example of structure/fragmentation analyses that can be conducted with SPADS.

The first input files (Phylip format) contain the DNA sequence alignments, one for each locus. It begins as follows (locus 1, with 80 sequences of 1000 pb):

```
80 800 1
80-pop8 CCATTGGCTTCTGACTCGGTGTGGCGTTTACTACAATT ...
79-pop8 CCATTGGCTTCTGACTCGGTGTGGCGTTTACTACAATT ...
78-pop8 CCATTGGCTTCTGACTCGGTGTGGCGTTTACTACAATT ...
77-pop8 CCATTGGCTTCTGACTCGGTGTGGCGTTTACTACAATT ...
76-pop8 CCATTGGCTTCTGACTCGGTGTGGCGTTTACTACAATT ...
75-pop8 CCATTGGCTTCTGACTCGGTGTGGCGTTTACTACAATT ...
74-pop8 CCATTGGCTTCTGACTCGGTGTGGCGTTTACTACAATT ...
...
```

The first line indicates the number of sequences (80), the number of nucleotides per sequence (800), and the ploidy level (1). Each sequence ID contains a number specific to the sequence and a population ID, separated by the symbol «-». The “populations” file contains the population IDs found in the previous input file, followed by the geographic coordinates of each population:

```
pop1 5 19
pop2 20 19
pop3 3 17
pop4 12 14
pop5 2 7
pop6 3 6
pop7 19 6
pop8 11 3
```

The “groups” input file partitions sampled populations *a priori* in two separate groups. This will allow the program to estimate population structure *a posteriori*. A group named “North” and another named “South” are here defined:

```
North pop1 pop2 pop3 pop4
South pop5 pop6 pop7 pop8
```

To investigate population differentiation between these two groups, we will use SPADS to (i) estimate population and phylogeographic structure statistics, and to perform (ii) an AMOVA analysis, (iii) a locus-by-locus and multi-loci SAMOVA analyses and (iv) a barrier

construction with the Monmonier algorithm. As displayed on Figure 5, these different options are selected on the toolbox interface. During the run, progress is displayed in the message area.

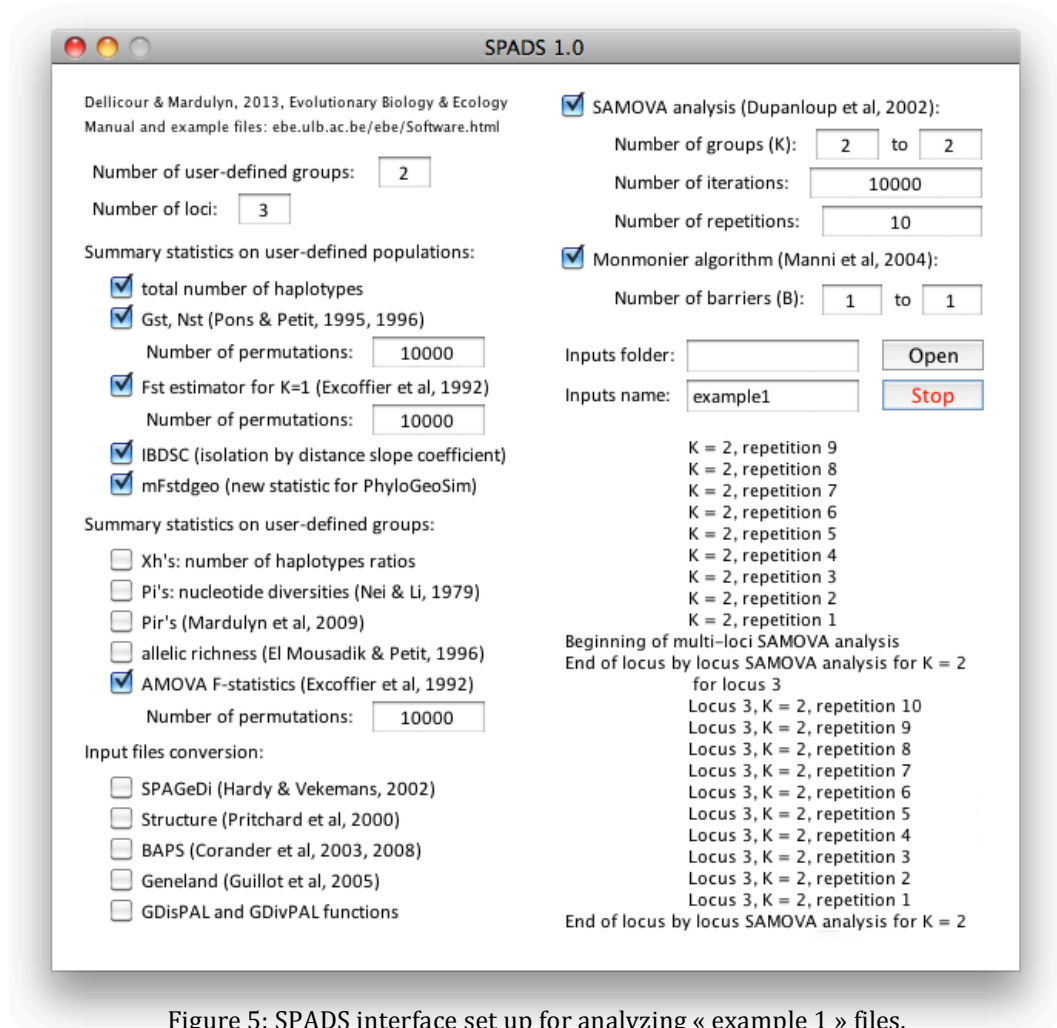


Figure 5: SPADS interface set up for analyzing « example 1 » files.

The “results” file produced by the program should look like this:

Summary statistics computations:

Gst (Pons & Petit, 1995) on combined loci = 0.05350347 (p-val = 0.0)
 Nst (Pons & Petit, 1996) on combined loci = 0.08234435 (p-val = 0.0)
 Nst-Gst on combined loci = 0.028840885 (p-val = 0.0411)
 PhiST (Excoffier et al, 1992) on combined loci = 0.08298171 (p-val = 0.0)
 AMOVA Phi-statistics on combined loci:
 PhiSC: -0.011166253 (p-val = 0.7414)
 PhiST: 0.14283836 (p-val = 0.0)
 PhiCT: 0.15230395 (p-val = 0.0)

LOCUS 1

Total number of haplotypes: 22
 Gst (Pons & Petit, 1995) = 0.079868555 (p-val = 0.0027)
 Nst (Pons & Petit, 1996) = 0.043715846 (p-val = 0.0265)
 Nst-Gst = -0.036152706 (p-val = 0.9273)
 Global PhiST for K=1 (Excoffier et al, 1992) = 0.043715846 (p-val = 0.0273)
 mPhiSTdgeo = 0.0026483338
 IBD regression slope coefficient = 0.0566677

User-defined groups:	North	South
Nseq:	40	40
Nh:	13	10
Xh:	0.59090906	0.45454547
Pi:	0.0025432692	0.0013477564
Pir:	1.8870392	0.5299307
Ar:	13.0	10.0

AMOVA Phi-statistics:
PhiSC: -0.022604952 (p-val = 0.788)
PhiST: 0.088072956 (p-val = 0.0011)
PhiCT: 0.10823134 (p-val = 0.0287)

LOCUS 2

Total number of haplotypes: 21
Gst (Pons & Petit, 1995) = 0.04231139 (p-val = 0.0108)
Nst (Pons & Petit, 1996) = 0.085339405 (p-val = 0.0008)
Nst-Gst = 0.043028016 (p-val = 0.0588)
Global PhiST for K=1 (Excoffier et al, 1992) = 0.085339405 (p-val = 0.0008)
mPhiSTdgeo = 0.0026814754
IBD regression slope coefficient = 0.10084494

User-defined groups:	North	South
Nseq:	40	40
Nh:	11	11
Xh:	0.52380955	0.52380955
Pi:	0.0053113555	0.005923077
Pir:	0.8967223	1.1151724
Ar:	11.0	11.0

AMOVA Phi-statistics:
PhiSC: -0.028156996 (p-val = 0.8262)
PhiST: 0.15527515 (p-val = 0.0)
PhiCT: 0.1784087 (p-val = 0.0293)

LOCUS 3

Total number of haplotypes: 26
Gst (Pons & Petit, 1995) = 0.040639862 (p-val = 0.0142)
Nst (Pons & Petit, 1996) = 0.111838326 (p-val = 0.0)
Nst-Gst = 0.07119846 (p-val = 2.0E-4)
Global PhiST for K=1 (Excoffier et al, 1992) = 0.111838326 (p-val = 0.0)
mPhiSTdgeo = 0.008023103
IBD regression slope coefficient = 0.06898533

User-defined groups:	North	South
Nseq:	40	40
Nh:	14	13
Xh:	0.53846157	0.5
Pi:	0.0020833334	0.0016461539
Pir:	1.2655764	0.79015386
Ar:	14.0	13.0

AMOVA Phi-statistics:
PhiSC: 0.016028496 (p-val = 0.2409)
PhiST: 0.17228465 (p-val = 0.0)
PhiCT: 0.1588015 (p-val = 0.0215)

SAMOVA RESULTS (locus by locus):

LOCUS 1
 Clusters ID pop1 pop2 pop3 pop4 pop5 pop6 pop7 pop8
 best partition for K = 2: 1 1 1 1 2 2 2 2
 PhiCT: 0.10823134
 PhiST: 0.088072956
 PhiSC: -0.022604952

LOCUS 2
 Clusters ID pop1 pop2 pop3 pop4 pop5 pop6 pop7 pop8
 best partition for K = 2: 1 1 1 1 2 2 2 2
 PhiCT: 0.1784087
 PhiST: 0.15527515
 PhiSC: -0.028156996

LOCUS 3
 Clusters ID pop1 pop2 pop3 pop4 pop5 pop6 pop7 pop8
 best partition for K = 2: 1 1 1 1 2 2 2 2
 PhiCT: 0.1588015
 PhiST: 0.17228465
 PhiSC: 0.016028496

SAMOVA RESULTS (multi-loci):

Clusters ID pop1 pop2 pop3 pop4 pop5 pop6 pop7 pop8
 best partition for K = 2 : 1 1 1 1 2 2 2 2
 multilocus weighted average PhiCT: 0.15230395
 multilocus weighted average PhiST: 0.14283836
 multilocus weighted average PhiSC: -0.011166253

Monmonier algorithm RESULTS (locus by locus):

the different inferred barriers defined groups of populations.

The corresponding group ID for each population are reported below.

LOCUS 1
 Clusters ID pop1 pop2 pop3 pop4 pop5 pop6 pop7 pop8
 for B = 1: 1 1 1 1 2 2 2 2

LOCUS 2
 Clusters ID pop1 pop2 pop3 pop4 pop5 pop6 pop7 pop8
 for B = 1: 1 1 1 1 2 2 2 2

LOCUS 3
 Clusters ID pop1 pop2 pop3 pop4 pop5 pop6 pop7 pop8
 for B = 1: 1 1 1 1 2 2 2 2

Monmonier algorithm RESULTS (multi-loci):

Clusters ID pop1 pop2 pop3 pop4 pop5 pop6 pop7 pop8
 for B = 1: 1 1 1 1 2 2 2 2

All three loci are associated with low but significant global G_{ST} , N_{ST} and Φ_{ST} statistics (with the exception of Φ_{ST} calculated for locus 1). A significant phylogeographic signal (as measured by $N_{ST}-G_{ST}$), was highlighted for loci 2 and 3. The AMOVA analysis estimated similar values for the Φ_{ST} and Φ_{CT} statistics at all three loci. Since Φ_{CT} measures the proportion of variation among groups, the analysis did identify that a significant part of the genetic variation is associated to the group partition. These results are further supported by the SAMOVA and the Monmonier algorithm: all the locus-by-locus and multilocus analyzes identified our two groups of populations correctly.

5.2. Tutorial 2: GDisPAL and GDivPAL functions on a bee (*C. hederæ*) dataset

Dataset: "Example 2 (*C. hederæ*)", available on the toolbox website (see below).

Description of the dataset: *Colletes hederæ* is a solitary bee currently studied for its recent range expansion in Western Europe, possibly a result of current global warming (Dellicour *et al*, *in press*). In this context, it is interesting to analyse the distribution of genetic diversity across the species range, especially for comparing old and newly colonized areas. One hundred haploid males sampled across the western portion of its range (i.e. France, Belgium, Germany and Switzerland) were sequenced at three nuclear loci. We used this dataset to test the accuracy of SPADS when computing standard summary statistics, by comparing its output with those of other softwares like SPAGeDi (Hardy & Vekemans, 2002), ARLEQUIN (Excoffier *et al*, 2005, 2010), FSTAT (Goudet *et al*, 1995) and SAMOVA (Dupanloup *et al*, 2002). In this tutorial, we show how to analyse inter-individual distances and population diversity using the GDisPAL and GDivPAL functions in R.

SPADS is used to generate the different input files needed for using these two functions, by selecting the appropriate input file conversion in the SPADS main window. It is important to specify on the software interface that there are three DNA sequence alignments to read and convert (field "number of loci" in the main window of the program). SPADS will then return the following outputs:

```
Chederæ_GDisPAL_input_coordinates.txt
Chederæ_GDisPAL_input_distances_matrix_IID1.txt*
Chederæ_GDisPAL_input_distances_matrix_IID2.txt*
Chederæ_GDisPAL_input_log(10)_pseudoslopes_matrix_IID1.txt*
Chederæ_GDisPAL_input_log(10)_pseudoslopes_matrix_IID2.txt*
Chederæ_GDisPAL_input_regression_residuals_matrix_IID1.txt
Chederæ_GDisPAL_input_regression_residuals_matrix_IID2.txt
Chederæ_GDivPAL_input_Ar.txt
Chederæ_GDivPAL_input_Pi.txt
Chederæ_GDivPAL_input_Pir.txt
```

Note that, as advised by Manni *et al* (2004), we will use inter-individual distances computed using residual distances derived from the linear regression of genetic vs. geographical distances. (*) refers to input files that we will not use here. We can now use R to produce interpolating surfaces using the two functions. The following instructions are entered in a R console. All input files have to be located in the same folder that contains the GDisPAL and GDivPAL function files. This folder is set as the working directory in R. If they are NOT installed yet, the following packages need to be installed and loaded:

```
> install.packages("rgl",dependencies=T); require(rgl)
> install.packages("fields",dependencies=T); require(fields)
> install.packages("raster",dependencies=T); require(raster)
> install.packages("geometry",dependencies=T); require(geometry);
```

If they are installed already, these packages just need to be loaded:

```
> require(rgl)
> require(fields)
> require(raster)
> require(geometry)
```

The next step is to load the GDisPAL and GDivPAL functions:

```
> source(file= "GDisPAL.r")
> source(file="GDivPAL.r")
```

Then, the interpolation parameter “a” and the “template” raster have to be specified:

```
> a = 5
> template = raster("template_Chederae.asc")
```

Note that the template raster will specify the area on which the interpolation will be performed. This can be any raster files and it will be use to delimitate the area of interest (the values it contained will not be used). Regarding the inverse distance interpolation algorithm “a”, this is important to test different values in order to investigate its impact on the interpolation result.

The next step is to load the different SPADS outputs:

```
> coordinates = read.table(file="GDisPAL_Chederae_input_coordinates.txt", h=F)
> distances_iid1 =
read.table(file="GDisPAL_Chederae_input_regression_residuals_matrix_IID1.txt",
h=F)
> distances_iid2 =
read.table(file="GDisPAL_Chederae_input_regression_residuals_matrix_IID2.txt",
h=F)
> diversities_Ar = read.table(file="GDivPAL_Chederae_input_Ar.txt", h=F)
> diversities_Pi = read.table(file="GDivPAL_Chederae_input_Pi.txt", h=F)
```

Once the matrices are loaded, we can call the GDisPAL and GDivPAL functions to build the interpolations surfaces:

```
> iid1_Chederae =
GDisPAL(template, coordinates_Chederae, distances_Chederae_iid1, a)
> iid2_Chederae =
GDisPAL(template, coordinates_Chederae, distances_Chederae_iid2, a)
> ar_Chederae = GDivPAL(template, diversities_Chederae_Ar, a)
> pi_Chederae = GDivPAL(template, diversities_Chederae_Pi,a)
```

These interpolation surfaces can be saved as raster files using the following commands:

```
> writeRaster(iid1_Chederae, file="Chederae_surface_IID1_a5.asc")
> writeRaster(iid2_Chederae, file="Chederae_surface_IID2_a5.asc")
> writeRaster(ar_Chederae, file="Chederae_surface_Ar_a5.asc")
> writeRaster(pi_Chederae, file="Chederae_surface_Pi_a5.asc")
```

Finally, the interpolation surfaces can be displayed using the function “plot” as followed:

```
> plot(iid1_Chederae)
> plot (iid2_Chederae)
> plot (ar_Chederae)
> plot (pi_Chederae)
```

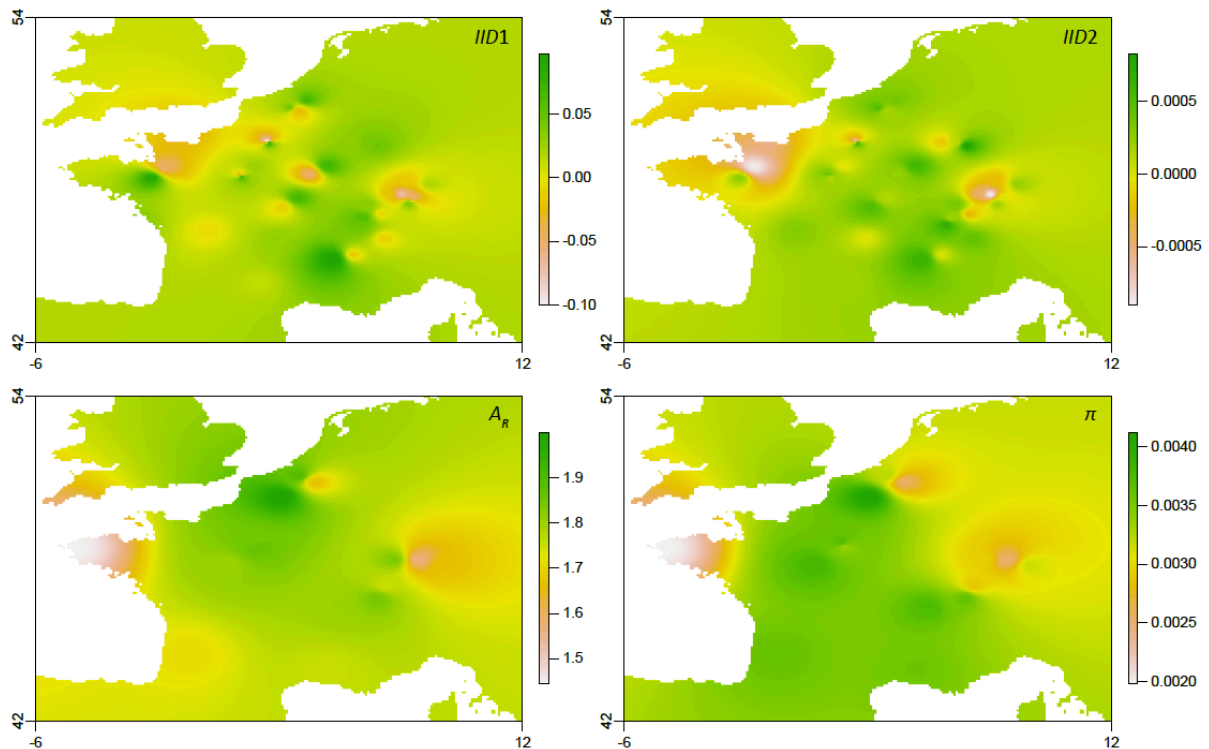


Figure 6: interpolation surfaces produced by GDivPAL and GDivPAL R functions for *C. hederae*.

5.3. Tutorial 3: GDivPAL function with a preliminary sliding window step

Dataset: “Example 2 (*C. hederae*)”, the same dataset as for tutorial 2.

This third tutorial is very similar to tutorial 2 but, instead of basing the interpolation on a SPADS output, the GDivPAL function is rather applied on the output of a “sliding window” function. Here, we use a sliding window approach to estimate the nucleotide diversity π associated with the cells of a template raster covering the study area. In practice, the value assigned to a given raster cell was the nucleotide diversity π estimated for the group of sequences sampled within a circle centred on this cell. The sliding window was implemented in the R function “slidingWindowPi”. As stated above, the GDivPAL function is then applied on the outcome of the sliding window application. See the related R script for further details as well as Lecocq *et al* (*submitted*) for an application.

In summary, the first step of tutorial 2 is replaced by the sliding window application:

```
> source(file="slidingWindowPi.r") # to load the slidingWindowPi function
> source(file="GDivPAL.r") # to load the GDivPAL function
> template = raster("Chederae_template_raster.asc") # to load the template raster
> inputName = "Chederae" # to specify the input name prefix
> numberOfLoci = 3 # to specify the number of loci
> radius = 50 # to define the radius of the slidingwindow (in km)
> a = 5 # interpolation parameter

> slidingWindowPi_Chederae = slidingWindowPi(template, inputName,
numberOfLoci, radius)
> diversities_Chederae_pi = rasterToPoints(slidingWindowPi_Chederae)
> pi_Chederae = GDivPAL(template, diversities_Chederae_pi, a, nberOfCores)
```

6. SPADZ1 and SPADZ2

Locus by locus and multi-loci SAMOVA algorithms can potentially run for a long time. To facilitate the use of these methods on shared computer clusters, we compiled two command line programs for the SAMOVA: SPADZ1 and SPADZ2. SPADZ1 implements a locus-by-locus SAMOVA analysis and SPADZ2 a multi-loci SAMOVA analysis. To use these command line versions, users need to add the following parameters to the first line of the “locus1” Phylip input file: *Kmin*, *Kmax*, the number of iterations, the number of independent runs, and the number of loci (i.e. the number of distinct Phylip files to read). Note that for SPADZ1 and SPADZ2, users do not have to specify the ploidy level. Examples of input files for SPADZ1 and SPADZ2 are given in Figure 7. The commands to launch SPADZ1 and SPADZ2 from a command-line window are:

```
java -jar SPADZ1.jar  
java -jar SPADZ2.jar
```

```
80 800 2 10 10000 3 5  
1-1 CCGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...  
2-1 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGGTTGAA ...  
3-1 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...  
4-1 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGGTTGAA ...  
5-2 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTACAGGGTTAAGATTGAA ...  
6-2 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGAA ...  
7-3 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGATTGAA ...  
8-3 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGAA ...  
9-3 CCGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGAA ...  
10-3 CTGAGCCGATTTGATGATAGGCAAGCACTAGGATTAGAGGGTTAAGATTGAA ...  
11-3 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGAA ...  
12-4 CTGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGAA ...  
13-4 CCGAGCCGATTTGATGATAGGCAAGCATTAGGATTAGAGGGTTAAGATTGAA ...  
...
```

Figure 7: example of input file for SPADZ1 and SPADZ2. In addition to the number of sequences (80) and the size of sequences (800 bp), the first line also contains *Kmin* (2), *Kmax* (10), the number of iterations (10,000), the number of independent runs (3) and the number of loci (5, i.e. the number of input files to read). These additional parameters only have to be added in the first input file (i.e. the input file for “locus1”).

7. Software limitations

SPADS has no practical limitations, but the bigger the datasets (number of loci, sequence lengths, number of individuals), the slower the analyses will be. All analyses are quite fast (in most case instantaneous), but a SAMOVA run can potentially take a long time. Note that for a classical SAMOVA analysis (10,000 iterations and 10 repeats) of the “example 1” data set presented in the tutorial and available on the software website, SPADS takes around one minute for the multi-loci and the three locus-by-locus SAMOVA’s for $K = 2$, using Java version 1.6.0_43-b01-447 on a 2.4 GHz Intel Core i5 machine running Mac OS X. Much larger datasets and/or a much larger number of runs to test for convergence could be needed in complex cases, however. In such situations, command line versions SPADZ1 and SPADZ2 represent useful alternatives to launch the algorithm on a computer cluster.

8. Toolbox availability

SPADS 1.0, GDisPAL-GDivPAL R functions, SPADZ1 and SPADZ2 are available from ebe.ulb.ac.be/ebe/Software.html. Java source code, example files and software manual are also available at this address.

9. Version history

- SPADS_1.0_261113.jar: first version.
- SPADS_1.0_240414.jar: minor bug fixed for the Delaunay triangulation (used in SAMOVA and Monmonier algorithms).
- 29th October 2014: R versions of the GDisPAL and GDivPAL functions have been updated and now works with the R package “raster”. The corresponding tutorial has also been changed accordingly.

10. References

- Cheng L., Connor T.R., Sirén J., Aanensen D.M., Corander J. (2013). Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution*, *in press*.
- Corander J., Sirén J., Arjas E. (2008). Bayesian spatial modeling of genetic population structure. *Computational Statistics* **23**: 111-129.
- Corander J., Waldmann P., Marttinen P., Sillanpää M.J. (2004). BAPS 2: Enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**: 2363-2369.
- Corander J., Waldmann P., Sillanpää M.J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**: 367-374.
- Dellicour S., Mardulyn P., Hardy O.J., Hardy C., Roberts S.P.M., Vereecken, N.J. (*in press*). Inferring the mode of colonisation of a rapid range expansion from multi-locus DNA sequence variation. *Journal of Evolutionary Biology*.
- Dupanloup I., Schneider S., Excoffier L. (2002). A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology* **11**: 2571-81.
- El Mousadik A., Petit R.J. (1996). Chloroplast DNA phylogeography of the argan tree of Morocco. *Molecular Ecology* **5**: 547-555.
- Ersts P.J. (2012) Geographic Distance Matrix Generator (version 1.2.3). American Museum of Natural History, Center for Biodiversity and Conservation. Available from http://biodiversityinformatics.amnh.org/open_source/gdmg.
- Excoffier L., Smouse P.E., Quattro J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**: 479-491.
- Excoffier L., Laval G., Schneider S. (2005). Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics* **1**: 47-50.
- Excoffier L., Lischer H.E.L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**: 564-567.
- Felsenstein J. (2004). PHYLIP (PHYLogeny Inference Package) version 3.6a2, Department of Genome Sciences, University of Washington, Seattle (2004).
- Goudet J. (1995). FSTAT Version 1.2: a computer program to calculate F-statistics. *Journal of Heredity* **86**: 485-486.

- Guedj B., Guillot G. (2011). Estimating the location and shape of hybrid zones. *Molecular Ecology Resources* **11**: 1119-1123.
- Guillot G., Estoup A., Mortier F., Cosson J.F. (2005a). A spatial statistical model for landscape genetics. *Genetics* **170**: 1261-1280.
- Guillot G., Mortier F., Estoup A. (2005b). GENELAND: A computer package for landscape genetics. *Molecular Ecology Notes* **5**: 712-715.
- Guillot G., Renaud S., Ledevin R., Michaux J., Claude J. (2012). A unifying model for the analysis of phenotypic, genetic, and geographic data. *Systematic Biology* **61**: 897-911.
- Guillot G., Santos F., Estoup A. (2008). Analysing georeferenced population genetics data with Geneland: A new algorithm to deal with null alleles and a friendly graphical user interface. *Bioinformatics* **24**: 1406-1407.
- Hardy O.J., Vekemans X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* **2**: 618-620.
- Landguth E.L., Cushman S.A. (2010). CDPOP: A spatially-explicit cost distance population genetics program. *Molecular Ecology Resources* **10**: 156-161.
- Lecocq T., Michez D., Gérard M., Vereecken N.J., Delangre J., Rasmont P., Vray S., Dufrêne M., Mardulyn P., Dellicour S. (*submitted*). Divergent geographic patterns of genetic diversity among wild bees: conservation implications.
- Librado P., Rozas J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451-1452.
- Manni F., Guérard E., Heyer E. (2004). Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by "Monmonier's algorithm". *Human Biology* **76**(2): 173-190.
- Mardulyn P., Mikhailov Y., Pasteels J.-M. (2009). Testing phylogeographic hypotheses in a Euro-Siberian cold-adapted leaf beetle with coalescent simulations. *Evolution* **63**: 2717-2729.
- Miller M.P. (2005). Alleles In Space (AIS): Computer software for the joint analysis of interindividual spatial and genetic information. *Journal of Heredity* **96**: 722-724.
- Miller M.P., Bellinger M.R., Forsman E.D., Haig S.M. (2006). Effects of historical climate change, habitat connectivity, and vicariance on genetic structure and diversity across the range of the red tree vole (*Phenacomys longicaudus*) in the Pacific Northwestern United States. *Molecular Ecology* **15**: 145-159.
- Nei M., Li W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76**: 5269-5273.
- Pons O., Petit R.J. (1995). Estimation, variance and optimal sampling of genetic diversity. I. Haploid locus. *Theor. Appl. Genet.* **90**: 462-470.
- Pons O., Petit R.J. (1996). Measuring and testing genetic differentiation with ordered versus unordered alleles. *Genetics* **144**: 1237-1245.
- Pritchard J.K., Stephens M., Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945-959.
- R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Rousset F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**: 1219-1228.
- Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**: 2731-2739.
- Watson D.F. (1992). Contouring: a guide to the analysis and display of spatial data Pergamon Press, New York, NY.

- Watson D.F., Philips G.M. (1985). A refinement of inverse distance weighted interpolation. *Geo-processing* **2**: 315-327.
- Weir B.S., Cockerham C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**: 1358-1370.