MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2014) **14**, 647–651 doi: 10.1111/1755-0998.12200

SPADS 1.0: a toolbox to perform spatial analyses on DNA sequence data sets

SIMON DELLICOUR and PATRICK MARDULYN

Evolutionary Biology and Ecology, Université Libre de Bruxelles, av. FD Roosevelt 50, Brussels 1050, Belgium

Abstract

SPADS 1.0 (for 'Spatial and Population Analysis of DNA Sequences') is a population genetic toolbox for characterizing genetic variability within and among populations from DNA sequences. In view of the drastic increase in genetic information available through sequencing methods, SPADS was specifically designed to deal with multilocus data sets of DNA sequences. It computes several summary statistics from populations or groups of populations, performs input file conversions for other population genetic programs and implements locus-by-locus and multilocus versions of two clustering algorithms to study the genetic structure of populations. The toolbox also includes two MATLAB and R functions, GDISPAL and GDIVPAL, to display differentiation and diversity patterns across landscapes. These functions aim to generate interpolating surfaces based on multilocus distance and diversity indices. In the case of multiple loci, such surfaces can represent a useful alternative to multiple pie charts maps traditionally used in phylogeography to represent the spatial distribution of genetic diversity. These coloured surfaces can also be used to compare different data sets or different diversity and/or distance measures estimated on the same data set.

Keywords: DNA sequences clustering, GDISPAL, GDIVPAL, landscape genetics, SPADS, summary statistics

Received 27 May 2013; revision received 21 October 2013; accepted 7 November 2013

Introduction

Initially, population genetics and phylogeographic studies using DNA sequence markers were based on single-locus data sets. Today, however, most intraspecific DNA sequence data sets include multiple loci. Such data sets are not only used to characterize and compare patterns of genetic diversity and population structure within or between species, but also to study their past history using, for example, methods based on coalescence models (Rosenberg & Nordborg 2002; Marjoram & Tavaré 2006). Here, we present SPADS 1.0, a toolbox to characterize genetic diversity and population structure from multilocus DNA sequences. SPADS computes several summary statistics to summarize population diversity and structure (e.g. G_{ST}, N_{ST}, *IBDSC*, X_{H}, π, π_R, A_{R}, AMOVA Φ-statistics), including standard statistics available in other population genetic programs, but also new statistics. Multilocus versions of two classic clustering methods defining groups of populations *a posteriori* and based on pairwise genetic distances (SAMOVA and Monmonier algorithm) are implemented. It can also create *ad hoc* input files for other popular clustering methods that are based solely on allelic frequencies (i.e. as opposed to using information on genetic distance between pairs of alleles; Pritchard *et al.* 2000; Corander *et al.* 2003, 2004, 2008; Guillot *et al.* 2005a,b, 2008). Finally, the toolbox SPADS includes MATLAB and R functions implementing an extension of a method developed by Miller (2005) to represent patterns of interindividual distances across landscapes. Our extension adds the possibility to also represent patterns of genetic diversity across landscapes and the ability to use any measure of genetic distance or diversity.

Although many methods implemented in SPADS were already available in separate programs (mostly ARLEQUIN for summary statistics computation; Excoffier *et al.* 2005; Excoffier & Lischer 2010), we here propose a friendly user toolbox specifically designed to deal with multilocus data sets and which can be used as a starting point for exploring a DNA sequence data set, from which the user can easily switch to other programs for complementary analyses using the input file conversion option.

Summary statistics

SPADS computes a number of summary statistics separately for each locus: the total number of haplotypes

Correspondence: Simon Dellicour, Fax: +32 2 650 2445;
E-mail: Simon.Dellicour@ulb.ac.be

(total number of different sequences detected for each locus), estimators of population differentiation G_{ST} (Pons & Petit 1995) and N_{ST} (Pons & Petit 1996), the AMOVA Φ_{ST} estimator when considering only one group of populations (Excoffier *et al.* 1992), the isolation by distance slope coefficient *IBDSC* defined as the slope coefficient of the linear regression estimated from $(\Phi_{ST}/(1\text{-}\Phi_{ST})) = f(\ln(x))$ (Rousset 1997), and $m\Phi_{ST}dgeo$, the average ratio between Φ_{ST} and geographical distance over all pairs of populations (a similar statistic was already computed by Mulcahy *et al.* 2006). As $m\Phi_{ST}dgeo$ will highlight geographically close populations that are genetically distant, this statistic can be used to detect departure from a pattern of isolation by distance. Statistical tests for the three *F*-statistics (G_{ST}, N_{ST} and Φ_{ST}) are based on random permutations of individuals between populations, while the statistical test for the difference between N_{ST} and G_{ST} (highlighting the extent of the phylogeographic signal) is based on random permutations of haplotypes (Hardy & Vekemans 2002). Corresponding *P*-values are the proportions of permutated data sets with a *F*-statistic higher or equal to the one estimated from the real data. The software also estimates several statistics on groups of populations defined *a priori* by the user: the ratio X_H between the number of haplotypes in each group and the total number of haplotypes, nucleotide diversity π (Nei & Li 1979), the ratio π_R between nucleotide diversity within each group and nucleotide diversity within the virtual group formed by all other populations (Mardulyn *et al.* 2009), allelic richness A_R (El Mousadik & Petit 1996) and AMOVA Φ-statistics (Excoffier *et al.* 1992). Statistical tests for the three AMOVA Φ-statistics (Φ_{SC}, Φ_{ST} and Φ_{CT}) are based on random permutations, the kind of permutation implemented depending on the Φ-statistic tested (Excoffier *et al.* 1992). All these summary statistics are briefly described in Table S1 (Supporting information; see the toolbox manual for further details). To verify the accuracy of the computations provided by the program, standard statistics computed by SPADS were compared with values obtained with other available programs (FSTAT, Goudet 1995; SAMOVA, Dupanloup *et al.* 2002; SPAGEDI, Hardy & Vekemans 2002; ARLEQUIN, Excoffier *et al.* 2005; Excoffier & Lischer 2010) .

Clustering analyses

SPADS implements two clustering methods based on pairwise distances among DNA sequences to define groups of populations *a posteriori*: a SAMOVA (i.e. spatial analysis of molecular variance, Dupanloup *et al.* 2002) and a Monmonier algorithm (Dupanloup *et al.* 2002; Manni *et al.* 2004). Compared with the original implementation of the SAMOVA (Dupanloup *et al.* 2002), users can specify the number of search iterations to perform for each run (independent repeats with different initial partitions of populations) of the algorithm. The second method is a Monmonier algorithm as implemented in the software BARRIER (Manni *et al.* 2004) but here based on a matrix of pairwise Φ_{ST} among populations (Excoffier *et al.* 1992) directly computed on the DNA sequence data sets. SPADS includes multilocus versions of these two methods. For each assumption of the number of groups (*K*), the multilocus version of the SAMOVA algorithm analyzes all loci simultaneously, using a multilocus weighted average Φ_{CT} (instead of the Φ_{CT} calculated for one locus; Excoffier *et al.* 1992) to compare two successive iterations. Similarly to the multilocus version of the SAMOVA, the multilocus version of the Monmonier algorithm computes multilocus weighted average Φ_{ST} estimators (instead of pairwise Φ_{ST} for one locus; Excoffier *et al.* 1992) to position barriers between populations on the map. Note that, contrary to the automated version implemented in SPADS, the software BARRIER does not allow to start from DNA sequences, but, on the other hand, generates graphical outputs that are very helpful to interpret the results. When selecting this clustering method in SPADS, the program will automatically generate input files based on multilocus information that can be read by the software BARRIER of Manni *et al.* (2004) to produce these graphical outputs. For generating this input file and for the inference of barrier(s) with the Monmonier algorithm, users can provide their own matrix of pairwise distances among sampled populations. Otherwise, pairwise Φ_{ST} computed from DNA sequences is used as default.

Input file conversions

SPADS can construct input files from DNA sequence alignments for several population genetic programs: SPAGEDI (Hardy & Vekemans 2002), STRUCTURE (Pritchard *et al.* 2000), BAPS (Corander *et al.* 2003, 2004, 2008) and GENELAND (Guillot *et al.* 2005a,b, 2008, 2012; Guedj & Guillot 2011). The interest of using SPADS to generate such input files is that, to the best of our knowledge, these other programs cannot directly analyse matrices of DNA sequences alignment. SPADS can also create input files for the GDISPAL and GDIVPAL MATLAB or R functions (see below). STRUCTURE, BAPS and GENELAND all implement clustering methods based solely on allelic frequencies (i.e. that do not take genetic distance among alleles into account). Recently, Cheng *et al.* (2013) extended the spatially explicit BAPS model for clustering DNA sequence data as well. For BAPS, SPADS creates two distinct BAPS input files: (i) one for the method based on allelic frequencies and (ii) a second for the method dealing with DNA sequences. Note that when choosing the clustering analysis based on the Monmonier algorithm, SPADS will also generate input files for the corresponding analysis implemented in the software BARRIER (Manni *et al.* 2004).

Mapping genetic diversity and differentiation: GDISPAL and GDIVPAL

In addition to the Java executable SPADS 1.0, we also included in this toolbox MATLAB and R functions implementing an extension of a method initially developed by Miller (2005, see also Miller *et al.* 2006) to display patterns of interindividual genetic distance 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, interindividual genetic distances are estimated and assigned to landscape coordinates at midpoints of each connectivity network edge. An interpolation procedure (i.e. an inverse distance-weighted interpolation; Watson & Philips 1985; Watson 1992) is used to infer genetic distances at different locations uniformly spaced on a grid. Here, we propose 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, instead of basing the interpolation procedure on distance values assigned at midpoints of each edge of a connectivity network, it uses diversity values directly estimated at each sampling point, hereafter simply designated as a 'population'. These interpolation methods are implemented in the two MATLAB (The MathWorks, Inc) or R (R Development Core Team 2008) functions: GDISPAL for 'genetic distance patterns across landscapes' and GDIVPAL for 'genetic diversity patterns across landscapes'. In these two functions, the inverse distance interpolation parameter *a* can be set to different values. We advise users to explore the influence of this parameter on the shape and smoothness of the interpolations. In GDISPAL, interpolations are all based on a Delaunay triangulation network. Surfaces created with these two functions can be seen in three dimensions or as heat maps to facilitate the comparison between several data sets. In this case, the colour scale can be standardized to allow direct visual comparison between several heat maps based on the same distance or diversity measure.

SPADS can use different distance measures to create GDISPAL function input files: (i) *IID*1 (for 'interindividual distance 1'), an interindividual distance based on allelic frequencies as defined by Miller (2005) and (ii) *IID*2 (for 'interindividual distance 2'), an interindividual distance based on pairwise nucleotide differences between DNA sequences (p-distance averaged across loci). Further details on these interindividual distances, including related formulas, are available in the manual. When there is a significant correlation between genetic and geographical distances, Manni *et al.* (2004) recommend using residual genetic distances derived from the linear regression of genetic against geographical distances. 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. SPADS can then also create regression residual and pseudoslope distance matrices computed from *IID*1 and *IID*2 measures.

Similarly, SPADS uses different diversity measures to create input files for the GDIVPAL function: (i) the allelic richness A_R (El Mousadik & Petit 1996) estimated within each population, (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.

A practical example of using the GDISPAL and GDIVPAL functions

Colletes hederae 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. Figure 1 presents the results of analysing interindividual distances and population diversity with the MATLAB functions GDISPAL and GDIVPAL. Note that, as advised by Manni *et al.* (2004), interindividual distances were computed using residual distances derived from the linear regression of genetic vs. geographical distances. The analysis of this data set is described in detail in a tutorial available in the software manual focusing on (i) the use of SPADS for computing several statistics and converting input files and (ii) the use of the R versions of the GDISPAL and GDIVPAL functions. The four interpolating surfaces shown in Fig. 1 can be used to highlight areas with higher population differentiation (i.e. red areas in *IID*1 and *IID*2 surfaces) and areas with higher genetic diversity (i.e. red areas in A_R and π surfaces). Also, this example illustrates that indices taking the genetic distance between DNA sequences into account (i.e. *IID*2 and π) can display notably different patterns than those based on indices only based on allelic frequencies (i.e. *IID*1 and A_R). These interpolating surfaces can be used to explore data, help to visualize complex multilocus information and also to visually compare spatial patterns of genetic variability among different species analysed with the same DNA sequence markers.

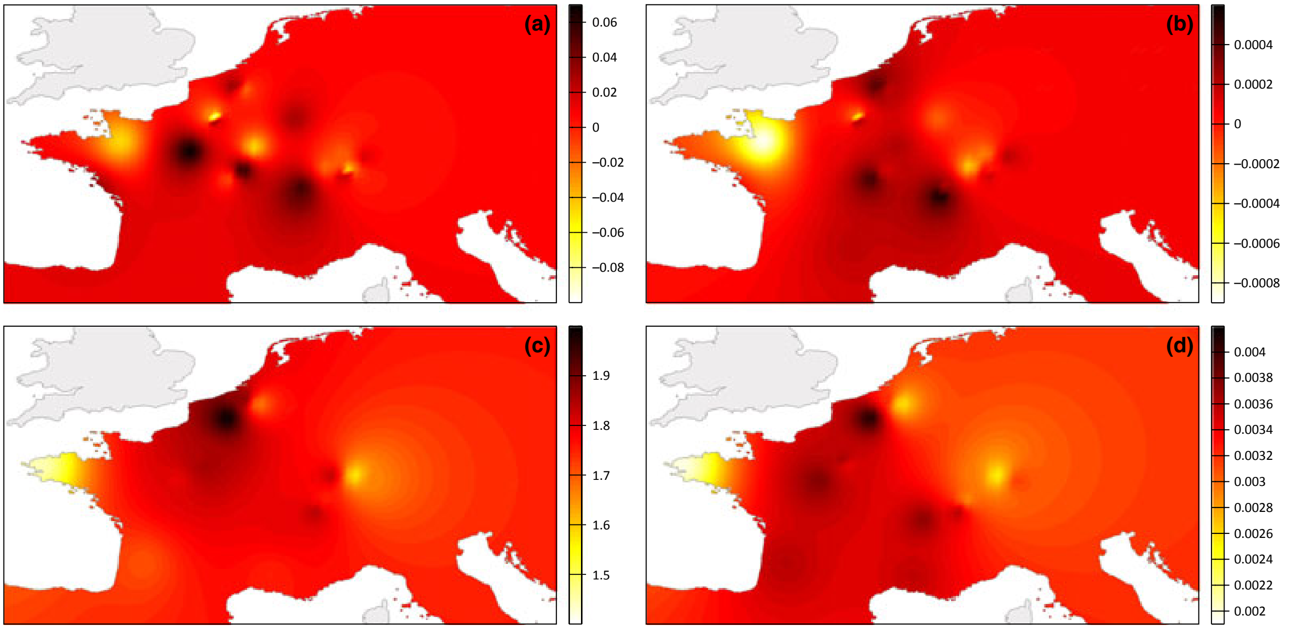

Fig. 1 Examples of inter-individual distance and population diversity interpolating surfaces in the case of *Colletes hederae*. These graphs are based on input files generated by SPADS. (a) graph generated with a matrix of average inter-individual distances based on allelic frequency (*IID*1). (b) graph generated with a matrix of average inter-individual distances based on allelic distances (*IID*2). (c) graph based on the allelic richness (A_R) estimated for each sampled population. (d) graph based on the nucleotide diversity (π) estimated for each sampled population.

Using SPADS

Running SPADS requires three types of input files: (i) the DNA sequence matrices in sequential Phylip format (Felsenstein 2004), (ii) a 'populations' file containing the geographical coordinates of each population and (iii) a 'groups' file defining the different groups of populations to analyse. A double click on the program file will prompt the program interface. The user is asked to complete different fields (e.g. number of loci, number of user-defined groups), to choose the input files directory and to specify the input files name before launching the analysis. As output, SPADS creates at least a 'results' and a 'messages' file, the second containing possible errors as well as additional information regarding the analyses performed (e.g. populations Φ_{ST} matrices used by the locus-by-locus Monmonier algorithm). In addition, SPADS will also generate input files for SPAGEDI, STRUCTURE, BAPS, GENELAND and/or GDISPAL-GDIVPAL functions, when required by the user.

Software availability

SPADS 1.0, GDISPAL and GDIVPAL R and MATLAB functions are freely available from *ebe.ulb.ac.be/ebe/Software.html*. Java source code, example files and a detailed manual involving tutorials on how to use SPADS and GDISPAL/GDIVPAL functions are also available from this website.

Acknowledgements

We wish to thank three anonymous referees for their constructive and useful comments. We are grateful to François Mayer, Gilles Dauby, Chedly Kastally and Thomas Lecocq for help in testing previous versions of the software. We also thank Vincent Bracke for help in the software interface development and Olivier Hardy for his useful advices. This research project was funded by the Belgian *Fonds National pour la Recherche Scientifique* (FNRS) and the *Fonds pour la recherche dans l'industrie et l'agriculture* (FRIA).

References

Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J (2013) Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution*, **30**, 1224–1228.

Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Corander J, Waldmann P, Marttinen P, Sillanpää MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**, 2363–2369.

Corander J, Sirén J, Arjas E (2008) Bayesian spatial modeling of genetic population structure. *Computational Statistics*, **23**, 111–129.

Dellicour S, Mardulyn P, Hardy OJ, Hardy C, Roberts SPM, Vereecken NJ (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**, 2571–2581.

El Mousadik A, Petit RJ (1996) Chloroplast DNA phylogeography of the argan tree of Morocco. *Molecular Ecology*, **5**, 547–555.

Excoffier L, Lischer HEL (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.

Excoffier L, Smouse PE, Quattro JM (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 Online*, **1**, 47–50.

Felsenstein J (2004). *PHYLIP (PHYLogeny Inference Package) version 3.6a2*, Department of Genome Sciences, University of Washington, Seattle.

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 JF (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, 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.

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.

Hardy OJ, 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.

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 JM (2009) Testing phylogeographic hypotheses in a Euro-Siberian cold-adapted leaf beetle with coalescent simulations. *Evolution*, **63**, 2717–2729.

Marjoram P, Tavaré S (2006) Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, **7**, 759–770.

Miller MP (2005) Alleles In Space (AIS): computer software for the joint analysis of interindividual spatial and genetic information. *Journal of Heredity*, **96**, 722–724.

Miller MP, Bellinger MR, Forsman ED, Haig SM (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.

Mulcahy DG, Spaulding AW, Mendelson JR, Brodie ED (2006) Phylogeography of the flat-tailed horned lizard (*Phrynosoma mcallii*) and systematics of the *P. mcallii*–platyrhinos mtDNA complex. *Molecular Ecology*, **15**, 1807–1826.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, **76**, 5269–5273.

Pons O, Petit RJ (1995) Estimation, variance and optimal sampling of genetic diversity. I. Haploid locus. *TAG. Theoretical and Applied Genetics. Theoretische und Angewandte Genetik*, **90**, 462–470.

Pons O, Petit RJ (1996) Measuring and testing genetic differentiation with ordered versus unordered alleles. *Genetics*, **144**, 1237–1245.

Pritchard JK, 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.

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, **3**, 380–390.

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

Watson DF (1992) *Contouring: A Guide to the Analysis and Display of Spatial Data*. Pergamon Press, New York, NY.

Watson DF, Philips GM (1985) A refinement of inverse distance weighted interpolation. *Geo-processing*, **2**, 315–327.

S.D., and P.M. designed the software; S.D. developed the software; S.D., and P.M. wrote the manuscript.

Data Accessibility

The toolbox, source code, user manual (including tutorials) and example data sets are available from ebe.ulb.ac.be/ebe/Software.html.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 List, brief description and reference of the different summary statistics computed by SPADS 1.0.