BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 11 2008, pages 1403–1405
doi:10.1093/bioinformatics/btn129

Genetics and population analysis

adegenet: a R package for the multivariate analysis of genetic markers

Thibaut Jombart*

Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, France

Received on February 12, 2008; revised on April 2, 2008; accepted on April 4, 2008

Advance Access publication April 8, 2008

Associate Editor: Alex Bateman

ABSTRACT

Summary: The package *adegenet* for the R software is dedicated to the multivariate analysis of genetic markers. It extends the *ade4* package of multivariate methods by implementing formal classes and functions to manipulate and analyse genetic markers. Data can be imported from common population genetics software and exported to other software and R packages. *adegenet* also implements standard population genetics tools along with more original approaches for spatial genetics and hybridization.

Availability: Stable version is available from CRAN: http://cran.r-project.org/mirrors.html. Development version is available from *adegenet* website: http://adegenet.r-forge.r-project.org/. Both versions can be installed directly from R. *adegenet* is distributed under the GNU General Public Licence (v.2).

Contact: jombart@biomserv.univ-lyon1.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Genetic markers are now widely used in many fields of population biology, and can be analysed using various approaches. Among these, multivariate methods such as principal component analysis (PCA) are compelled to play an important role because they can summarize the genetic variability without making strong assumptions about an evolution model: they do not rely on Hardy–Weinberg equilibrium, nor do they suppose the absence of linkage disequilibrium. This is especially valuable when no or very little information is known about the system under study, as is frequent in landscape genetics (Manel *et al.*, 2003). Recently, multivariate methods have proven useful to assess the consensus genetic structuring among a set of genetic markers (Laloë *et al.*, 2007), as well as to investigate the spatial pattern of the genetic variability (Jombart *et al.*, in press). However, multivariate methods currently available in population genetics software are very restricted, despite the fairly large number of these programs (Excoffier and Heckel, 2006). An exception to this is the R software (R Development Core Team, 2008) which contains both packages devoted to multivariate methods like *ade4* (Chessel *et al.*, 2004; Dray *et al.*, 2007), and packages dedicated to the analysis of genetic markers (http://cran.r-project.org/web/views/Genetics.html). Currently there are no bridges between multivariate analysis packages and genetic marker packages, and genetic markers data cannot be readily analysed using multivariate approaches. The purpose of *adegenet* is to build this connection. This package aims at extending the *ade4* package so that genetic markers can be analysed using multivariate methods. This is achieved by defining new classes of objects to represent genetic markers, and providing functions to import, export and manipulate these objects. Moreover, *adegenet* also implements some usual population genetics methods, as well as more original tools for spatial genetics and data simulation. This article presents an overview of these functionalities.

2 CONTENT

2.1 Data representation

Basic genetic markers data are genotypes obtained for a set of markers, each allele being coded by a character string (Warnes, 2003). In order to use statistical methods, such information cannot be used directly, and needs to be recoded numerically into a matrix of allelic frequencies. In *adegenet*, allelic frequencies of genotypes are stored inside objects of the class `genind`, which is the basic class of the package. In addition to allelic frequencies stored in a `@tab` component (the '@' designs a slot), a `genind` object stores other useful information whose description is provided by the R command `class?genind`. This class `genind` was designed to allow flexibility (it can virtually store any relevant information about genotypes) but also to be robust. As `genind` is a formal class (or 'S4 class' in R language), it is naturally robust: the content of an object is checked for validity when it is created and each time it is modified, which considerably limits the risks of having wrong or missing items in it. Moreover, `genind` internally uses generic labels for markers, alleles and genotypes, so that missing or redundant user-defined labels cannot originate an error in further analyses. Whenever the study involves groups of genotypes (say, 'populations') rather than genotypes, `genpop` objects are used. This formal class is very similar to `genind`, except that `@tab` contains counts of alleles per population instead of allelic frequencies of genotypes. Objects of both classes can be analysed by multivariate methods using

*To whom correspondence should be addressed.

the @tab slot as input. Main available functions to import to, export from, manipulate and analyse genind and genpop objects are listed in Supplementary Material.

2.2 Functionalities

Great attention was devoted to developing input/output functions, because interoperability of data is crucial to facilitate data analysis. Until now, data could only be imported into R from FSTAT (Goudet, 2002) using the *hierfstat* package (Goudet, 2005). Currently, *adegenet* can read files from the software GENETIX (Belkhir *et al*., 1996–2004), STRUCTURE (Pritchard *et al*., 2000), FSTAT (Goudet, 2002), and Genepop (Raymond and Rousset, 1995), which are among the most common data formats in population genetics software (Excoffier and Heckel, 2006). Data can also be read inside R from a data.frame of genotypes coded by character strings (using df2genind), and exported back (genind2df). Outputs are possible from genind to the R packages *genetics* (Warnes, 2003) and *hierfstat* (Goudet, 2005), using genind2genotype and genind2hierfstat, respectively. Note that the data representation in the *genetics* package was intended to be consensual, and is used by many other R packages. Moreover, the output of genind2df is customisable and can be designed to fit usual formats like those of GENETIX or STRUCTURE.

To perform analyses at a population level, a genind object can be translated into a genpop object using genind2genpop. Other data manipulations include splitting information by marker (seploc) or by population (seppop), computing allelic frequencies for populations (makefreq), or subsetting genotypes, populations or alleles according to a given criterion. Basic methods are implemented such as the Hardy–Weinberg equilibrium test for all combinations of populations and markers (HWE.test.genind), a matrix of *P*-values allowing a quick overview of the results. Observed and expected heterozygosity, number of alleles by marker or population, sample sizes and other miscellaneous information are provided by summary functions. Missing data can be replaced in different ways—which is required by most statistical methods—using na.replace. Several genetic distances among populations can be computed using dist.genpop. Goudet's *G* statistic (Goudet *et al*., 1996) can be tested by a Monte–Carlo procedure to assess the hierarchical structuring of a set of genotypes (gstat.randtest).

The last goal of *adegenet* is to implement more original methods, either by extending existing ones, or by proposing new methods. Hybridization between individuals from two genind objects can be simulated using hybridize, which can be useful to evaluate the power of methods based on genetic differentiation. Monmonier's algorithm (Monmonier, 1973), which is used to infer genetic boundaries among geo-referenced genotypes (Manni *et al*., 2004), has been extended to include different degrees of connectivity among genotypes (monmonier) and implemented with an optimization function (optimize.monmonier). Finally, recently developed methods to investigate spatial patterns of the genetic variability (Jombart *et al*., in press) are also part of the package (functions spca, global.rtest and local.rtest).

3 EXAMPLE

This example illustrates how a theoretical hybrid population would appear on a typology provided by a multivariate method. First, we load the required packages, and the dataset microbov containing 30 microsatellite markers for 704 genotypes of 15 cattle breeds, described in Laloë *et al*. (2007).

```
> library(adegenet)
> library(ade4)
> data(microbov)
```

To simulate a hybrid population, two parent breeds (Salers and Zebu) are isolated:

```
> temp <- seppop(microbov)
> salers <- temp$Salers
> zebu <- temp$Zebu
```

The hybrid population ('Zebler') is obtained using the hybridize function (with n = 40 simulated genotypes). All data are pooled in a new object newbov:

```
> zebler <- hybridize(salers, zebu,
+ pop = "Zebler", n = 40)
> newbov <- repool(microbov, zebler)
```

Now we seek a typology displaying the diversity between breeds. For this, the inter-class PCA (Dolédec *et al*., 1987) is appropriate: this modification of PCA maximizes the variance between populations (here, breeds), instead of the total variance. Missing data are replaced (na.replace) before performing a centred PCA (dudi.pca) and an inter-class PCA (between):

```
> newbov <- na.replace(newbov, method =
"mean")
> pca1 <- dudi.pca(newbov$tab, center = TRUE,
+ scale = FALSE, scannf = FALSE)
> bet1 <- between(pca1, fac = newbov$pop,
+ scannf = FALSE, nf = 3)
```

The resulting typology (Fig. 1) is obtained by:

```
> s.class(bet1$ls, fac = newbov$pop,
+ clab = 1.2, lab = newbov$pop.names)
> add.scatter.eig(bet1$eig, nf = 2, xax = 1,
+ yax = 2, pos = "bottomright", csub = 1.2)
```

The first principal axis of the analysis (Fig. 1) differentiates African and French breeds, while the second axis expresses the genetic variability between African breeds. Interestingly enough, the simulated hybrid population (Zebler) appears between its parent populations (Salers and Zebu).

4 CONCLUSION

The first contribution of the R package *adegenet* is to implement classes and functions to facilitate the multivariate analysis of genetic markers. This led to define new formal classes for genotypes (genind) or groups of genotypes (genpop), which can be used as input to multivariate methods proposed in the R software. Several functions are also implemented to manipulate and analyse these objects, including recent development in spatial genetics and data simulation. By assuring a good interoperability of data, *adegenet* contributes to making the R software a unifying platform for the analysis of genetic markers.

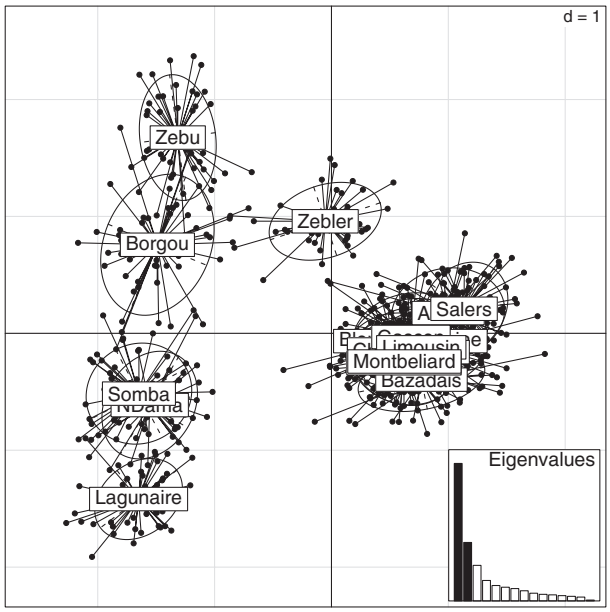

Fig. 1. Typology of cattle breeds (object `newbov`) obtained by inter-class PCA. Eigenvalues corresponding to the represented components are filled in black. Points represent genotypes; breeds are labelled inside their 95% inertia ellipses.

ACKNOWLEDGEMENTS

The author is grateful to R-Forge for hosting *adegenet*, to P. Sólymos for his contribution and to A.-B. Dufour, S. Devillard, D. Laloë and D. Pontier for their constructive comments.

Conflict of Interest: none declared.

REFERENCES

Belkhir,K. *et al*. (1996–2004) GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Genome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier, France. URL: http://www.genetix.univmontp2.fr/genetix/intro.htm.

Chessel,D. *et al*. (2004) The ade4 package-I-one-table methods. *R News*, **4**, 5–10.

Dolédec,S. and Chessel (1987) Rythmes saisonniers et composantes stationnelles en milieu aquatique. *Acta Oecologica, Oecologia Generalis*, **8**, 403–426.

Dray,S. *et al*. (2007) The ade4 package – II: two-table and *K*-table methods. *R News*, **7**, 47–54.

Excoffier,L. and Heckel,G. (2006) Computer programs for population genetics data analysis: a survival guide. *Nat. Rev. Genet*., **7**, 745–758.

Goudet,J. (2002) Fstat 2.9.3.2. URL: http://www2.unil.ch/popgen/softwares/fstat.htm.

Goudet,J. (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes*, **5**, 184–186.

Goudet,J. *et al*. (1996) Testing differentiation in diploid populations. *Genetics*, **144**, 1933–1940.

Jombart, T. *et al*. (in press) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*.

Laloë,D. *et al*. (2007) Consensus genetic structuring and typological value of markers using multiple co-inertia analysis. *Genet. Sel. Evol*., **39**, 545–567.

Manel,S. *et al*. (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecol. Evol*., **18**, 189–197.

Manni,F. *et al*. (2004) Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by ''Monmonier's algorithm''. *Hum. Biol*., **76**, 173–190.

Monmonier,M. (1973) Maximum-difference barriers: an alternative numerical regionalization method. *Geogr. Anal*., **3**, 245–261.

Pritchard,J. *et al*. (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*. ISBN 3-900051-07-0. R Foundation for Statistical Computing, Vienna, Austria.

Raymond,M. and Rousset,F. (1995) Genepop (version 1.2): population genetics software for exact tests and ecumenicism. Journal of *Heredity*, **86**, 248–249.

Warnes, G. (2003) The genetics package. *R News*, **3**, 9–13.