BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 3 2010, pages 419–420
doi:10.1093/bioinformatics/btp696

Genetics and population analysis

pegas: an R package for population genetics with an integrated–modular approach

Emmanuel Paradis
Institut de Recherche pour le Développement, UR175 CAVIAR, BP 5095, 361 rue Jean-François Breton, F-34196 Montpellier Cédex 5, France

Received on October 7, 2009; revised on December 14, 2009; accepted on December 16, 2009

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: pegas (Population and Evolutionary Genetics Analysis System) is a new package for the analysis of population genetic data. It is written in R and is integrated with two other existing R packages (ape and adegenet). pegas provides functions for standard population genetic methods, as well as low-level functions for developing new methods. The flexible and efficient graphical capabilities of R are used for plotting haplotype networks as well as for other functionalities. pegas emphasizes the need to further develop an integrated–modular approach for software dedicated to the analysis of population genetic data.

Availability: pegas is distributed through the Comprehensive R Archive Network (CRAN): http://cran.r-project.org/web/packages/pegas/index.html Further information may be found at: http://ape.mpl.ird.fr/pegas/

Contact: emmanuel.paradis@ird.fr

Population genetics has a strong mathematical background, and therefore genetic data analyses heavily relies on computer programs. Currently, no unified framework for these programs exists making the use of the many different population genetics programs a complicated task (Excoffier and Heckel, 2006). On the other hand, R (R Development Core Team, 2009) appeared as a unified framework for analysing data in bioinformatics (Gentleman, 2008) and phylogenetics (Paradis, 2006). In this note, I introduce pegas (Population and Evolutionary Genetics Analysis System) which aims to fill the gap between the existing R packages and the current need in population genetic data analyses.

pegas is entirely written in R to insure maximum portability among operating systems. It requires a standard R installation as well as two packages: ape (Paradis *et al*., 2004) and adegenet (Jombart, 2008). These three packages make an integrated environment for population genetic data analysis.

One of the strengths of R is the flexibility of its data structures, so that it is easy to adapt them for a particular field, resulting in much easier data input/output and manipulation (one of the most time-consuming step in data analysis). Two such structures have been derived for pegas. The main one is the class `'loci'` which is a simple data frame where the rows are individuals and the columns are loci and optional variables. The latter may be of any kind (continuous, integer, logical, etc.). Loci are coded with R factors (i.e. categorical variables) which are vectors of integers. Since computing on integers is fast, counting genotypic frequencies can be done efficiently. The convention is to code a genotype with the alleles separated with a forward slash, so that once genotypes are counted, allelic frequencies are easily obtained by counting the number of alleles in each genotype. Consequently, any level of ploidy may be handled. Since most operations on allelic data require counting allele and genotype frequencies, this operation has been optimized in the function `summary.loci` which is used by other functions in pegas.

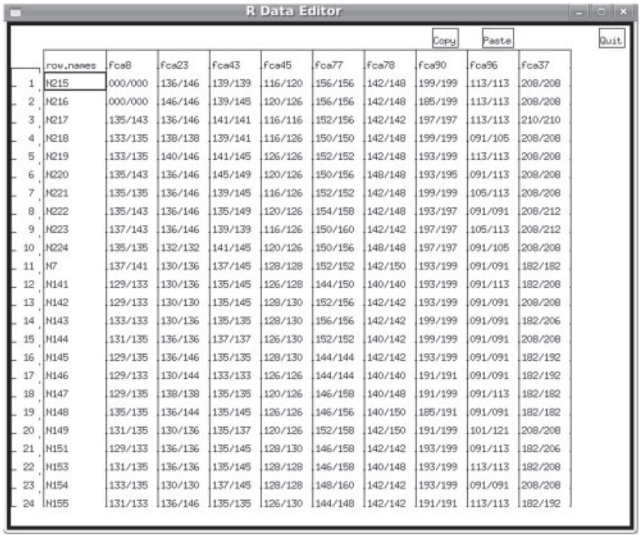
The class `'loci'` inherits the class `'data.frame'` (a standard class in R) so all functionalities available for the latter (manipulation, subsetting, etc.) can be used for the former (Chambers, 2008). The typical example is in the use of the indexing operator `[, ]`. A special version of this operator has been written for the class `'loci'` so that the loci columns are correctly identified even after deletion of some rows and/or columns. Another advantage of the data frame structure is that each row is identified by a unique label (the `rownames`) so that these data may easily be matched with other data that have similar labels (e.g. a distance matrix, a DNA sequence alignment or a genealogy).

R has many tools to analyse data frames and these may be used directly for the class `'loci'` and functions in pegas. A common task in population genetics is to perform some analyses for different subsets of the data (populations, breeds, plots, etc.). For instance, if a data set `x` has a variable `population`, the R function `by` may be used to compute some statistics for each level of this variable, e.g. `by(x, x$population, hw.test)` to perform Hardy–Weinberg test. The second argument may be a list of variables in which case the analyses will be done for each combination of them, e.g. `by(x, list(x$plot, x$treatment), summary)` will compute allele and genotype frequencies for every combination of `plot` and `treatment`.

The second main special data structure of pegas is the class `'haplotype'` which inherits the class `'DNAbin'` of ape. In addition to the set of unique DNA sequences, this class includes a vector of indices identifying the individuals belonging to each haplotype. This makes possible to link a set of haplotypes with individual data (phenotypes, geographic locations, etc.).

Missing data are coded explicitly in pegas. In the class `'loci'`, the standard `NA` is used and these are handled in the R standard way. In the class `'haplotype'`, the IUPAC code for ambiguous nucleotides is used and this is taken into account during calculations.

The current version of pegas includes standard tools for population genetic analyses: genotypic and allelic frequencies,



	row.names	fca8	fca23	fca43	fca45	fca77	fca78	fca90	fca96	fca37
1	N215	000/000	136/146	139/139	116/120	156/156	142/148	199/199	113/113	208/208
2	N216	000/000	146/146	139/145	120/126	156/156	142/148	185/199	113/113	208/208
3	N217	135/143	136/146	141/141	116/116	152/156	142/142	197/197	113/113	210/210
4	N218	133/135	138/138	139/141	116/126	150/150	142/148	199/199	091/105	208/208
5	N219	133/135	140/146	141/145	126/126	152/152	142/148	193/199	113/113	208/208
6	N220	135/143	136/146	145/149	120/126	150/156	148/148	193/195	091/113	208/208
7	N221	135/135	136/146	139/145	116/126	152/152	142/148	199/199	105/113	208/208
8	N222	135/143	136/146	135/149	120/126	154/158	142/148	193/197	091/091	208/212
9	N223	137/143	136/146	139/139	116/126	150/160	142/142	197/197	105/113	208/212
10	N224	135/135	132/132	141/145	120/126	150/156	148/148	197/197	091/105	208/208
11	N7	137/141	130/136	137/145	128/128	152/152	142/150	193/199	091/091	182/182
12	N141	129/133	130/136	135/145	126/128	144/150	140/140	193/199	091/113	182/208
13	N142	129/133	130/130	135/145	128/130	152/156	142/142	193/199	091/091	208/208
14	N143	133/133	130/136	135/135	128/130	156/156	142/142	199/199	091/091	182/206
15	N144	131/135	136/136	137/137	126/130	152/152	140/142	199/199	091/091	208/208
16	N145	129/135	136/146	135/135	128/130	144/144	142/142	193/199	091/091	182/192
17	N146	129/133	130/144	133/133	126/126	144/144	140/140	191/191	091/091	182/192
18	N147	129/135	138/138	135/135	120/126	146/158	140/148	191/199	091/113	182/182
19	N148	135/135	136/144	135/145	126/126	146/156	140/150	185/191	091/091	182/182
20	N149	131/135	130/136	135/137	120/126	152/158	142/150	191/199	101/121	208/208
21	N151	129/133	136/136	135/145	128/130	146/158	142/142	193/199	091/113	182/206
22	N153	131/135	136/136	135/145	128/128	146/158	140/148	193/199	113/113	182/208
23	N154	133/135	130/130	137/145	128/128	148/160	142/142	193/199	091/091	208/208
24	N155	131/133	136/146	135/135	126/130	144/148	142/142	191/191	113/113	182/192

Fig. 1. The R data editor showing microsatellite data. Note the first column allowing to edit the individual (row) labels.

Hardy–Weinberg equilibrium, F_{ST}, analysis of molecular variance, haplotype network, mismatch distribution, Tajima's D and R_2 tests for population stability, nucleotide diversity (π), the population parameter θ ($=4N_e\nu$), the site frequency spectrum, as well as several functions for reading and writing data files. pegas has also several low-level functions for manipulating its data structures (extracting observed alleled and genotypes, their frequencies, ploidy levels, building tables of all possible genotypes), making possible to extend its functionalities in a straightforward way. pegas includes basic tools for the coalescent such a computing likelihoods for a given tree, and estimating θ by maximum likelihood. All functions are accompanied with a help page describing how the calculations are done and giving the relevant literature references. These pages are compiled in a reference manual in PDF available from the above web site.

There are already a large number of computer programs for handling genetic data, so it is crucial for R users to be able to analyse the data files from the most widely used ones. pegas provides a function `read.loci` to read allelic data in tabular form from a text file. This function has several options to specify the loci and allele separator as well as which columns should be considered as loci or as additional variables. pegas has several functions to convert the data structures from adegenet (class `'genind'`) to the class `'loci'`. Since adegenet has functions to read files created for the programs STRUCTURE, FSTAT, GENETIX and GENEPOP (`read.structure`, `read.fstat`, `read.genetix`, `read.genepop`), these file formats may also be readily used into pegas as well. Once read into R, and possibly converted into the class `'loci'`, data can be edited by hand with the R spreadsheet data editor (Fig. 1). pegas is distributed with a tutorial explaining, step-by-step, how to input data from different file formats (in R type `vignette('ReadingFiles')`).

pegas uses the efficient and flexible graphical capabilities of R (Murrell, 2006). For instance, the options for plotting a haplotype network (Templeton *et al*., 1992) are:

```
plot.haploNet(x, size = 1, col = "black",
    bg = "white", col.link = "black",
    lwd = 1, lty = 1, pie = NULL,
    labels = TRUE, scale.ratio = 1,
    legend = FALSE, fast = FALSE, ...)
```

These options make possible to plot symbols of different sizes, colours (contour and background) and the links between them may be of different colours, widths or line types (solid, dotted, dashed, etc.). These may be controlled by variables computed from the original dataset with R's standard statistical and computing tools.

pegas illustrates the use of an integrated–modular approach for the development of software data analysis. The three packages adegenet, ape and pegas, complement each other for population genetics as they provide functions for spatial and multivariate analyses (adegenet), trees structures and DNA sequences manipulation (ape) and basic population genetic analyses (pegas). Furthermore, each package has enough functionalities to be used on its own, independently of the others.

ACKNOWLEDGEMENTS

I am grateful to two anonymous referees for their comments on a previous version of this article and to Thibaut Jombart for useful discussions.

Funding: IRD program 'Spirales', and European Science Foundation (ESF) through the MolArch (Molecular Archives of Past Climatic Changes) project.

Conflict of Interest: none declared.

REFERENCES

Chambers,J.M. (2008) *Software for Data Analysis: Programming with R*. Springer, New York.

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

Gentleman,R. (2008) *R Programming for Bioinformatics*. Chapman & Hall/CRC, Boca Raton.

Jombart,T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Murrell,P. (2006) *R Graphics*. Chapman & Hall/CRC, Boca Raton, FL.

Paradis,E. (2006) *Analysis of Phylogenetics and Evolution with R*. Springer, New York.

Paradis,E. *et al*. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

Templeton,A.R. *et al*. (1992) A cladistic analysis of phenotypic association with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, **132**, 619–635.