BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 5 2007, pages 644–645
doi:10.1093/bioinformatics/btm025

Genetics and population analysis

SNPassoc: an R package to perform whole genome association studies

Juan R. González[1,*], Lluís Armengol[1], Xavier Solé[2], Elisabet Guinó[2], Josep M. Mercader[1], Xavier Estivill[1] and Víctor Moreno[2,*]

[1]Genes and Disease Program, Centre for Genomic Regulation and [2]Unit of Biostatistics and Bioinformatics, Epidemiology Service, IDIBELL, Catalan Institute of Oncology, Barcelona, Spain

Received on September 27, 2006; revised and accepted January 22, 2007

Advance Access publication January 31, 2007

Associate Editor: Keith Crandall

ABSTRACT

Summary: The popularization of large-scale genotyping projects has led to the widespread adoption of genetic association studies as the tool of choice in the search for single nucleotide polymorphisms (SNPs) underlying susceptibility to complex diseases. Although the analysis of individual SNPs is a relatively trivial task, when the number is large and multiple genetic models need to be explored it becomes necessary a tool to automate the analyses. In order to address this issue, we developed SNPassoc, an R package to carry out most common analyses in whole genome association studies. These analyses include descriptive statistics and exploratory analysis of missing values, calculation of Hardy–Weinberg equilibrium, analysis of association based on generalized linear models (either for quantitative or binary traits), and analysis of multiple SNPs (haplotype and epistasis analysis).
Availability: Package SNPassoc is available at CRAN from http://cran.r-project.org
Contact: juanramon.gonzalez@crg.es or v.moreno@iconcologia.net
Supplementary information: A tutorial is available on *Bioinformatics* online and in http://davinci.crg.es/estivill_lab/snpassoc

1 INTRODUCTION

Whole genome association studies, in which a dense set of SNPs across the genome is genotyped, are a novel approach to assess the role of genetic variation in disease. To increase the efficiency of this approach, multistage designs have been proposed (Hirschhorn and Daly, 2005). In the first step, thousands of SNPs are tested for association with the disease. In a second and possibly third step, additional detailed studies are performed, in which only a few hundred SNPs, those with a putative association found in the first step, are genotyped.

Although analysis of a single or a small number of SNPs is a relatively simple task to conduct (Solé *et al.*, 2006), the statistical analysis of large-scale studies is challenging. In this article, we present *SNPassoc*, an R package (http://www.r-project.org) designed to analyze genome-wide association studies. *SNPassoc* contains tools for data manipulation, exploratory data analysis with graphics, and assessment of genetic association for both quantitative and binary traits. For the analysis of a small selection of SNPs, the package also provides tools to analyze interactions between SNPs or haplotypes and other SNPs or environmental variables. This note presents an overview of the package but a detailed tutorial is provided in the Supplementary Material.

*To whom correspondence should be addressed.

1.1 Data manipulation and descriptive analysis

SNPassoc uses the object-oriented features of R ('classes and methods') to ease data manipulation, analysis and plots. Variables coding for SNP genotypes are defined with the function `snp`, which takes care of formatting and assigns class 'snp'. The recommended format delimits each allele with a character (i.e. '/'), but two-letter formats or any three codes are also allowed. Objects of class 'snp' can be explored using the generic R functions `print`, `summary` and `plot`. The summary of a 'snp' object shows genotype and allele frequencies, missing values and a test for compliance with Hardy–Weinberg equilibrium. By default, the reference category is the genotype homozygous for the most frequent allele. This may be changed using the method `reorder`.

If the user has a large collection of SNPs coded similarly, the function `setupSNP` prepares the data automatically. Information about chromosome and genomic positions, if given, is used later to classify or sort SNPs in tables and plots. The function `setupSNP` returns a packed object that can be explored and analyzed with a series of functions. For example, the generic function `summary` provides a table with a systematic descriptive analysis, including allele frequencies, percentage of missing values and the test for Hardy–Weinberg equilibrium. This test may also be obtained using the function `tableHWE`, which uses an exact test of Hardy–Weinberg equilibrium as described in Wigginton *et al.* (2005). The function `plotMissing` provides a visual representation of missing values in samples and SNPs (Fig. 2 in the Supplementary Material). Objects with class 'setupSNP' can be manipulated after their creation: variables can be added or deleted and subsets of SNPs can be selected for specific analysis.

1.2 Whole genome association studies

After initial inspection of the data, analysis of association can be performed using the function `WGassociation`, which requires an object of class 'setupSNP'. To demonstrate how to perform this analysis using a real dataset, we have downloaded individual genotypes from the HapMap project (http://www.hapmap.org) and randomly selected close to 10 000 SNPs distributed across the 22 autosomes. We compare the genotype frequencies for all SNPs from this dataset between the European (CEU) and African (YRI) populations. The dataset and the genomic information are loaded typing `data(HapMap)`. The required object of class 'setupSNP' is created executing:

```
myDat<-setupSNP(HapMap, colSNPs=3:9809,
      sort=TRUE, info=SNPs.pos,
      sep="")
```

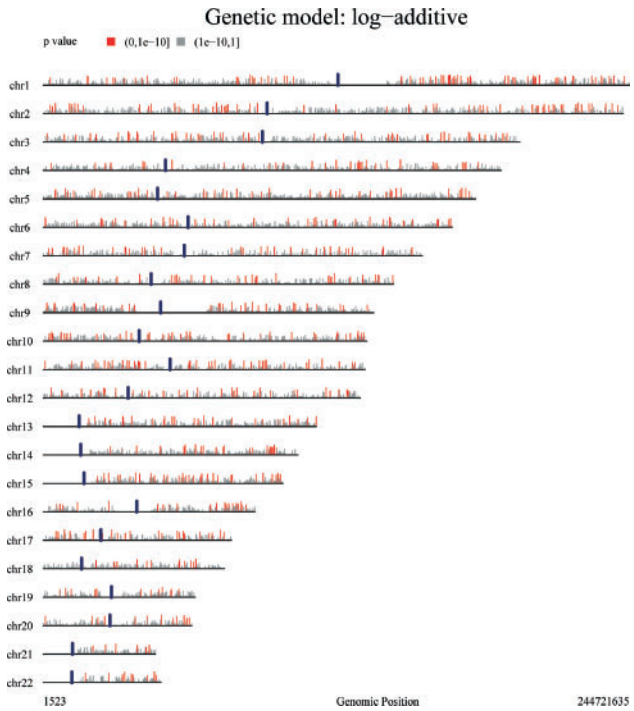

Fig. 1. Results of `WGassociation` for the `HapMap` dataset. The $-\log P$ values for a whole genome analysis assuming a log-additive genetic model are shown for each chromosome. The statistically significant associations at level 10^{-10} are plotted in red, while the other associations are in gray. Blue lines indicate the centromeres.

The association analysis is then performed by typing:

```
resHapMap<-WGassociation
           (group, data=myDat,
           model=''log-additive'')
```

The first argument, group, is the response variable that can be either binary or continuous (quantitative trait). In this example, `WGassociation` fits individual logistic regression models to each of the variables class 'snp' provided in the 'setupSNP' data object `myDat`. If we need an analysis adjusted for covariates, these can be indicated using a model formula. For example, to adjust the association of each SNP for age and gender the formula would be: `group~age+gender`. The argument `model` allows for the selection of different genetic models: codominant, dominant, recessive, overdominant, log-additive (or all). Since a genome-wide association analysis may be computationally intensive, we recommend exploring only one or two models in a preliminary step (in our example `''log-additive''`). We provide a faster alternative function, `scanWGassociation`, but it only reports *P*-values.

The function `WGassociation` returns a special object that can be inspected using generic R methods. Figure 1 shows the results of `plot` for a genome-wide analysis assuming a log-additive model of inheritance. The function `summary` shows additional global information in tabular format. For each SNP and genetic model, the function `WGstats` provides a cross tabulation with numbers and percentages, odds ratios (or mean differences for quantitative traits), 95% confidence intervals, the *P*-value for the likelihood ratio test of association (or Fisher's exact test if some cell is empty), and the Akaike information criteria.

Computing time and memory requirements are critical issues for the practical usefulness of this package. The analyses of this package use R code, but the computing times are reasonable. The exception is `scanWGassociation` for binary traits, that has been programmed in Fortran (including permutation test). For example, a typical desktop computer needs 5 min to import the data from a study with 30 000 SNPs in 1000 subjects, 40 min to prepare `setupSNP` and 30 s to calculate the *P*-values (using `scanWGassociation`). More examples can be found in the Supplementary Material. These analyses can be performed in parallel on several computers and the results can be combined. The peak memory requirements are 750 MB since R stores everything in memory.

1.3 Association studies with reduced number of SNPs

After identifying a subset of SNPs with putative associations, they are normally analyzed in more detail. The function `getSignificantSNPs` helps to select these SNPs (see Section 3 from Supplementary Material). This function currently uses a Bonferroni correction for multiple comparisons, but other methods based on *P*-values as implemented in the R package `multtest` (Pollard *et al.*, 2006) could be used. The function `association` provides a detailed analysis of the association between a given SNP and the response, analogous to the previously described output of `WGstats`. This function performs crude, adjusted, stratified, subset and interaction analyses in a similar way as the web application *SNPstats* (Solé *et al.*, 2006).

1.4 Analysis of multiple SNPs

Two different analyses can be performed when multiple SNPs are considered: epistasis (SNP–SNP interaction) and analysis of haplotypes. The function `interactionPval` performs an epistasis analysis between all pairs of SNPs. This analysis probably should be restricted to a reduced subset of selected SNPs. The output matrix of *P*-values can be visualized with `plot` (Supplementary Fig. 6). Finally, the R package `haplo.stats` (Sinnwell and Schaid, 2005) implements the analysis of association of haplotypes with a response via generalized linear models (`haplo.glm`). We provide auxiliary functions to prepare the data (`make.geno`) and to obtain a more detailed and complete output for the analysis of interactions between haplotypes and covariates (`haplo.interaction`). This function generates three tables of interactions, the cross-tabulated and two conditional tables of one SNP nested within the other.

ACKNOWLEDGEMENTS

Authors thank M. Gratacòs, R. de Cid and M. Cáceres for testing the functions. This work has been supported by CEGEN (Spanish Genotyping Center) funded by Genoma España, Danone Institute and grants FIS03/0114 (Instituto de Salud Carlos III) and SAF2005-01005 (Spanish Ministry of Science and Education). Funding to pay the Open Access publication charges was provided by Genetic Causes of Disease Lab, Center for Genomic Regulation.

Conflict of Interest: none declared.

REFERENCES

Hirschhorn,J.N. and Daly,M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev.*, **6**, 95–108.

Pollard,K.S. *et al.* (2006). multtest: Resampling-based multiple hypothesis testing. R package version 1.10.2 http://cran.r-project.org.

Sinnwell,J.P. and Schaid,D.J. (2005) haplo.stats: Statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. R package version 1.2.2. http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm.

Solé,X. *et al.* (2006) SNPstats: a web tool for the analysis of association studies. *Bioinformatics*, **22**, 1928–1929.

Wigginton,J.E. *et al.* (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.*, **76**, 887–893.