Bioinformatics, 32(12), 2016, 1901–1902
doi: 10.1093/bioinformatics/btw080
Advance Access Publication Date: 15 February 2016
Applications Note

Genetics and population analysis

solarius: an R interface to SOLAR for variance component analysis in pedigrees

Andrey Ziyatdinov[1,*], Helena Brunel[1], Angel Martinez-Perez[1], Alfonso Buil[2], Alexandre Perera[3,4,†] and Jose Manuel Soria[1,†]

[1]Unitat De Genòmica De Malalties Complexes, Institut D'investigació Biomèdica Sant Pau (IIB-Sant Pau), Barcelona, Spain, [2]Department of Genetics Medicine and Development, University of Geneva Medical School, Geneva, Switzerland, [3]B2SLab, Department ESAII, Universitat Politècnica De Catalunya, Barcelona, Spain and [4]CIBER in Bioengineering Biomaterials and Nanomedicine, Barcelona, Spain

Associate Editor: Janet Kelso

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Received on May 14, 2015; revised on January 19, 2016; accepted on February 5, 2016

Abstract

Summary: The open source environment R is one of the most widely used software for statistical computing. It provides a variety of applications including statistical genetics. Most of the powerful tools for quantitative genetic analyses are stand-alone free programs developed by researchers in academia. SOLAR is one of the standard software programs to perform linkage and association mappings of the quantitative trait loci (QTLs) in pedigrees of arbitrary size and complexity. *solarius* allows the user to exploit the variance component methods implemented in SOLAR. It automates such routine operations as formatting pedigree and phenotype data. It parses also the model output and contains summary and plotting functions for exploration of the results. In addition, *solarius* enables parallel computing of the linkage and association analyses that makes the calculation of genome-wide scans more efficient.
Availability and implementation: *solarius* is available on CRAN and on GitHub https://github.com/ugcd/solarius.
Contact: aziyatdinov@santpau.cat

1 Introduction

Variance component (VC) models or linear mixed models are powerful tools for genetic studies particularly of quantitative traits. These models are attractive because they account for the contribution of individual genetic loci, while efficiently including polygenic and other confounding effects shared among individuals. Implementation of the VC methods has been traditionally a computationally challenging task, and SOLAR is one of the first and well-established VC tools that focuses on the analysis of quantitative trait loci (QTLs) in extended pedigrees (Almasy and Blangero, 1998). *solarius* delivers to the R user three main quantitative genetic models: polygenic, linkage and association.

The motivation to develop the *solarius* software came from the extensive experience of the group that studies the Genetic Analysis of Idiopathic Thrombophilia (GAIT) Project (Soria *et al.*, 2002). The first goal of *solarius* was to provide an effortless data manipulation in a polygenic analysis needed to be explored for such a large number of phenotypes. The second goal was to conduct the genome-wide scans for both linkage and association mappings in an efficient way by means of parallel computing.

2 Approach

2.1 Implementation

The *solarius* package allows the import and export of data, automated manipulation of intermediate directories and configuration of SOLAR commands. The user works with top-level R functions which correspond to low-level SOLAR commands, as summarized

Table 1. Implementation of the three main models in *solarius*

Model	SOLAR command	solarius function	Tables of results
Polygenic	polygenic	solarPolygenic	cf, vcf, lf
Linkage	multipoint	solarMultipoint	lodf, lodf2
Association	mga	solarAssoc	snpf

The high-level functions of the package (column 3) correspond to the low-level SOLAR commands (column 2). The results of an analysis are extracted from SOLAR output files and stored in elements of the returned objects in R (column 4). The main elements contain results for covariates (cf), variance components (vcf), likelihood statistics (lf), SNP associations (snpf) and logarithm of odds (LOD) scores (lodf and lodf2 for the first and the second passes, respectively).

in Table 1. Each function performs the analysis with default SOLAR behavior, but the user can pass a specific configuration, for example, by means of polygenic.settings and polygenic.options arguments in the solarPolygenic function.

The package has a number of benefits as a part of the R environment. The main functions returns output results as objects of S3 classes, for which print, summary and plot methods are defined. Pedigree relationships can be examined by plotPed and plotKinship2 functions based on *kinship2* and *Matrix* R packages, respectively. The large tables of results from the association and linkage analyses are efficiently stored and accessed via the *data.table* R package. The results of the association analysis are explored with quantile–quantile (QQ) and Manhattan plots from the *qqman* R package. In addition, the *rsnps* R package is used to retrieve SNPs information by sending queries to public databases.

Implementation of parallel calculations is straightforward, since the association and linkage analyses are implicitly parallel problems, and the R environment offers a number of packages with parallel interfaces (*parallel*, *iterators* and *doParallel* packages). The user needs to introduce only the parameter cores (the number of cores) to configure parallel computing.

2.2 A practical example

Here, we show an application of *solarius* package and place emphasis on its features by performing genome-wide QTL mapping for coagulation Factor XII (FXII) levels in the GAIT1 Project. The original work (Soria *et al.*, 2002) showed unequivocally that a locus in the *F12* gene influences both FXII activity and susceptibility to thrombosis.

The recruitment, phenotyping and genotyping methods used in the GAIT1 Project have been described extensively elsewhere (Soria *et al.*, 2002). High-quality genotypes at 363 DNA microsatellite markers spaced at a density of 9.5 cM and 299 695 SNPs were available for linkage and association mappings.

The following three lines of R code performs polygenic, linkage and association analyses for FXII phenotype by calling the three main functions of *solarius* package.

```
solarPolygenic (FXII ~ AGE + SEX, dat, covtest = TRUE)
solarMultipoint (FXII ~ 1, dat, household = FALSE,
mibddir = gait1.mibddir, cores = 2)
solarAssoc (FXII ~ 1, dat, household = FALSE,
mga.files = gait1.files, cores = 2)
```

All the three functions support the formula interface for fixed effects, similar to that of the standard linear regression lm function. The second argument of the functions is a data frame (dat) that contains not only phenotypic variables given in the formula and pedigree-specific identificators, but also optional proband and household variables for ascertainment and shared-environment corrections. The matchIdNames function defines the naming controls.

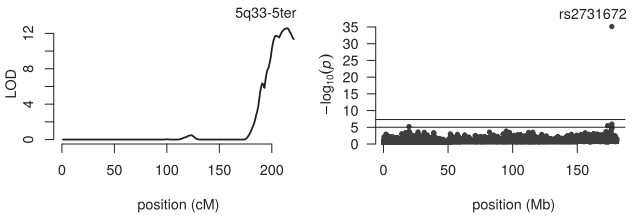

Fig. 1. Results of (**a**) linkage and (**b**) association mappings on Chromosome 5 for Factor FXII in the GAIT1 sample. The identified locus is the *F12* gene

Two mapping functions, solarMultipoint and solarAssoc, take input genetic data in a plain-text format of SOLAR. The user needs to prepare these data in advance likely using external tools and custom quality control pipelines. The mibddir argument specifies a directory with identity by descent (IBD) matrices. The mga.files argument takes a list of files with allele-dosage SNP data, optionally split into batches. solarAssoc function accepts also both genotype and allele-dosage data in PLINK and R data frame formats.

The initial polygenic model included two fixed AGE and SEX effects and two random polygenic and household effects. The covariates were tested for statistical significance, as indicated by the covtest argument. Neither a covariate or a household effect was statistically significant at the 0.05 level and, thus, were excluded from the following models. The heritability in the final polygenic model was 0.64 ± 0.08 with P-value 1.21×10^{-16}.

Both linkage and association scans identified the *F12* gene locus at the genome-wide significant level. The mapping results on Chromosome 5 in Figure 1 were produced by the default plot methods. Annotation of the association results based on *rsnps* package (annotate function) showed that the only significant SNP rs2731672 belongs to the *F12* gene and tags the untyped causal 46C/T polymorphism rs180102 reported in (Soria *et al.*, 2002) (linkage disequilibrium measures are $D' = 1$ and $R^2 = 1$).

A considerable speed-up of mapping scans can be achieved by parallel computing. Our computation time in minutes of FXII association mapping (average over 5 runs, standard error) was 264.1 (0.2), 132.2 (0.1) and 34.5 (0.1) on 2, 4 and 16 cores, respectively. The gains are closely proportional to the relative difference in the number of cores, as it is expected for implicitly parallel problems.

More information on polygenic, linkage and association studies with *solarius* is on http://ugcd.github.io/solarius/vignettes/tutorial.html.

Acknowledgment

The authors thank Professor W.H. Stone for revising the manuscript.

Funding

This research was funded by the TEC2013-44666-R grant. This work was partially funded by the 2014SGR-2016 consolidated research group of the Generalitat de Catalunya, Spain. CIBER-BBN is an initiative of the Spanish ISCIII. This research was supported partially by grants PI-11/0184, PI-14/0582 and UIN2013-50833 from the Instituto Carlos III (Fondo de Investigacin Sanitaria FIS).

Conflict of Interest: none declared.

References

Almasy,L. and Blangero,J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.*, **62**, 1198–1211.

Soria,J.M. *et al.* (2002) A quantitative-trait locus in the human factor xii gene influences both plasma factor xii levels and susceptibility to thrombotic disease. *Am. J. Hum. Genet.*, **70**, 567–574.