BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 14 2008, pages 1643–1644
doi:10.1093/bioinformatics/btn241

Genetics and population analysis

SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data

Carlos Morcillo-Suarez[1,2,3], Josep Alegre[1,2,3], Ricardo Sangros[1,2,3], Elodie Gazave[1], Rafael de Cid[2,4], Roger Milne[2,5], Jorge Amigo[2,6], Anna Ferrer-Admetlla[1], Andrés Moreno-Estrada[1], Michelle Gardner[1], Ferran Casals[1], Anna Pérez-Lezaun[1,2], David Comas[1,7], Elena Bosch[1,7], Francesc Calafell[1,7], Jaume Bertranpetit[1,2,7] and Arcadi Navarro[1,2,3,7,8,*]

[1]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Barcelona, [2]National Genotyping Centre (CeGen), [3]Population Genomics Node (GNV8) National Institute for Bioinformatics (INB), [4]Genes and Disease Program, Center for Genomic Regulation (CRG), Barcelona, [5]Human Genetics Group, Human Cancer Genetics Program, Spanish National Cancer Centre, Madrid, [6]Unidade de Xenética, Facultad de Medicina, Santiago de Compostela, [7]CIBER en Epidemiologia y Salud Pública (CIBERESP) and [8]Institució Catalana de Recerca i Estudis Avançats, ICREA and Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Dr Aiguader 88, 08003 Barcelona, Spain

Received on February 21, 2008; revised and accepted on May 16, 2008

Advance Access publication May 30, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: Single nucleotide polymorphisms (SNPs) are the most widely used marker in studies to assess associations between genetic variants and complex traits or diseases. They are also becoming increasingly important in the study of the evolution and history of humans and other species. The analysis and processing of SNPs obtained thanks to high-throughput technologies imply the time consuming and costly use of different, complex and usually format-incompatible software. SNPator is a user-friendly web-based SNP data analysis suite that integrates, among many other algorithms, the most common steps of a SNP association study. It frees the user from the need to have large computer facilities and an in depth knowledge of genetic software installation and management. Genotype data is directly read from the output files of the usual genotyping platforms. Phenotypic data on the samples can also be easily uploaded. Many different quality control and analysis procedures can be performed either by using built-in SNPator algorithms or by calling standard genetic software.

Availability: Access is granted from the SNPator webpage http://www.snpator.org.

Contact: arcadi.navarro@upf.edu; bioinformatica.cegen@upf.edu

Supplementary information: Additional information, including tutorials and example datasets, is available from SNPator's webpage.

1 INTRODUCTION

The vast number of SNPs identified in the last few years and the development of high-throughput genotyping technologies have provided the opportunity for many research groups to undertake association studies of varying scales on a regular basis. SNP association studies have become crucial in the uncovering of genetic correlations of genomic variants with complex diseases, quantitative traits and physiological responses to drugs (e.g. Andrawiss, 2005). SNPs are also increasingly employed to study the history of populations and the evolution of species (e.g. Moreno-Estrada *et al.*, 2008; Tishkoff *et al.*, 2007).

In spite of the increasing popularity of SNP studies, processing and analyzing the huge amounts of data generated by genotyping technologies is still a burdensome and time consuming task. Hundreds of different software packages, most of them free for research purposes, have been developed to deal with particular problems and are available on the Web (http://linkage.rockefeller.edu/soft). Much time and effort is required, not only to identify the most appropriate algorithms and programs for each goal, but also to install them on local computers, to learn how they work or to give the appropriate format to input data. Within many genotyping projects, post-genotyping data management and analysis have become a bottleneck hindering the achievement of results. In order to help tackling these problems we have developed a web-based software solution called SNPator (for SNP analysis to results).

2 IMPLEMENTATION

2.1 Architecture and database features

The basic structure of SNPator consists of a central Linux server with MySQL and the PHP written application. This central node acts as a webserver and database manager. All the tasks and analyses that SNPator performs are coded in the form of WebServices that are executed remotely by computing servers and which can be called by external software other than SNPator.

Users can log into the application via web using a standard browser and introducing the usernames and passwords that can be obtained—without registration—from SNPator's webpage. Users

*To whom correspondence should be addressed.

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

have different levels of privileges and can only access their own studies. A study is a working space—shared by as many users as necessary—where a set of data and all results generated from its analysis are stored. Each study starts with three types of data in highly customizable tables: a set of SNPs with related genomic information, a set of samples with population or phenotypic information and a set of genotypes. SNP and sample information can be easily uploaded using several methods, including, for SNPs, automatic upload from public databases such as dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/) or HapMap (www.hapmap.org). Information on samples can include any sort of numerical, categorical or textual variables and can also be automatically uploaded after customizing the fields in the study. Genotypes can be uploaded directly from the output files generated by the most usual genotyping technologies (Illumina, Sequenom, SNPlex, etc.). All data within SNPator can be uploaded and downloaded in XML format to ease interaction with other software.

2.2 Quality control and analysis features

Once data have been uploaded, SNPator offers many quality control and analysis possibilities. Quality control options range from the detection of contradictory genotypes to the generation of graphical reports of uploaded plates. As to analysis, the simplest way in which SNPator can be used is to generate formatted data files ready to be used by other programs. Data can be downloaded into different formats ranging from ordered lists and matrices (to be imported into Excel or SPSS, for example) to input files for standard genetic software. Other analysis possibilities range from the simplest tests (such as Hardy Weinberg) to genomic overview measures (linkage disequilibrium, haplotype inference, population differentiation statistics, and others), disease-oriented analyses (allele or haplotype association tests, TDT and others) or multiple test corrections. Some analysis algorithms have been implemented as PHP scripts in SNPator, while others use standard external software that has been wrapped into WebServices.

Any action demanded by the user generates a job that will be sent to a queue and performed when resources become available. Most jobs will be performed immediately but those requiring more computational resources (haplotype estimations, for instance) will be put on hold while other such jobs are running. The appropriate screen provides users with information about the generation, execution, completion time and current status of their jobs. All the actions performed in SNPator generate results which are stored in a section called User Results. Results remain there as long as the user wants them and can be read, downloaded and even reused for further analysis in SNPator (in the case of workflows with more than one step). When launching an action, the user can ask to be sent an e-mail when this action is finished.

2.3 Filters and batch mode

SNPator implements several features that ease complex analysis procedures. First, users may define a set of criteria (filters) to select a subset of SNPs and samples from a study by means of easily created Boolean statements. The fraction of genotyping success of a SNP or sample can also be used as a criterion to set up a filter. When one of the filters is activated, all operations performed with SNPator will affect only the SNPs and samples selected in the filter and its genotypes.

Another feature which facilitates analysis is the Batch Mode in which several jobs can be simultaneously generated using as inputs different values in a field. If, for example, 'Sample Batch mode' for the field 'Population' is selected when running an allele frequency job, SNPator examines the 'Population' field, determines how many different values are there and runs as many 'allele frequency' tests as populations in that field.

2.4 System management

A web-based administration application has also been developed. Using it, it is possible to perform tasks such as managing the set of extant studies and the user privileges. It is also possible to obtain usage statistics by means of text or a graphical output. Such statistics include summaries of user logins and their actions, lists of currently running and waiting jobs, memory usage parameters and many others. This feature will be made generally available in future 'pre-packaged' versions of SNPator that users will be able to install on their own servers

3 APPLICATIONS AND USE TO DATE

SNPator is open to all users and it is currently the core application in the Spanish National Genotyping Center (www.cegen.org). CeGen is a nodal network of different genotyping facilities distributed in three different cities and created to allow scientists access to distinct genotyping technologies. Once samples are genotyped, data are uploaded into SNPator from the different platforms so that users can access them at a single point, add their own data and perform any analysis. External users can upload their own data by themselves. Over the last two years, SNPator has been used to perform, either in part or completely, more than 200 studies, ranging from association studies (Goertsches *et al.*, 2008) to population genetic analysis of genes or genome regions in different populations (Gardner *et al.*, 2007). SNPator differs from other packages in both its wide and ever-growing spectrum of possibilities and its extremely easy usage.

ACKNOWLEDGEMENTS

We are grateful to the CeGen coordination team for continuing support. We are indebted to the many users that have provided us with feedback about features to improve.

Funding: This work is funded by the National Institute for Bioinformatics and the National Genotyping Center, two platforms of Genoma España, and projects BFV2005 – 00243 to EB and BFU2006-15413-C02-01 to A.N.

Conflict of Interest: none declared.

REFERENCES

Andrawiss,M. (2005) First phase of HapMap project already helping drug discovery. *Nat. Rev. Drug Discov.*, **4**, 947.

Gardner,M. *et al.* (2007) Extreme individual marker F(ST)values do not imply population-specific selection in humans: the NRG1 example. *Hum. Genet.*, **121**, 759–762.

Goertsches,R. *et al.* (2008) Evidence for association of chromosome 10 open reading frame (C10orf27) gene polymorphisms and multiple sclerosis. *Mult. Scler.*, **14**, 412–414.

Moreno-Estrada,A. *et al.* (2008) Signatures of selection in the human olfactory receptor OR5I1 gene. *Mol. Biol. Evol.*, **25**, 144–154.

Tishkoff,S.A. *et al.* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.*, **39**, 31–40.