Bioinformatics, 31(14), 2015, 2380–2381
doi: 10.1093/bioinformatics/btv144
Advance Access Publication Date: 11 March 2015
Applications Note

OXFORD

Genome analysis

MUGBAS: a species free gene-based programme suite for post-GWAS analysis

S. Capomaccio[*,†], **M. Milanesi**[†], **L. Bomba**[†], **E. Vajana and P. Ajmone-Marsan**

Istituto di Zootecnica, Università Cattolica del Sacro Cuore, 29122, Piacenza, Italy

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: John Hancock

Received on November 12, 2014; revised on March 5, 2015; accepted on March 6, 2015

Abstract

Genome Wide Association Studies between molecular markers and phenotypes are now routinely run in model and non-model species. However, tools to estimate the probability of association of functional units (e.g. genes) containing multiple markers are not developed for species other than humans. Here we introduce MUGBAS (MUlti species Gene-Based Association Suite), software that estimates the *P*-value of a gene using information on annotation, single marker GWA results and genotype. The software is species and annotation independent, fast, highly parallelized and ready for high-density marker studies.

Availability and implementation: https://bitbucket.org/capemaster/mugbas

Contact: capemaster@gmail.com

1 Introduction

With the availability of high-density SNP panels, Genome Wide Association Studies (GWAS) have become a standard for dissecting the architecture of complex traits (McCarthy *et al*., 2008). This is true for humans, and other species. The underlying idea of GWAS is simple: find marker-trait associations exploiting the linkage disequilibrium (LD) that exists between causative mutations—which we ignore—and flanking markers of known chromosomal position.

A number of well established approaches can be used in a standard GWAS (Aulchenko *et al*., 2007; Kang *et al*., 2010; Purcell *et al*., 2007). These generally use many thousand and sometimes millions of markers. Statistical thresholds are set considering multiple comparisons and may exclude a number of moderate or small effects actually associated with the traits investigated.

Reducing the number of tests while maintaining all the information provided by high-density panels would be a clear advantage.

In addition, once significant signals are detected, different strategies are used to tag the candidate causative genes. Sometimes, all genes lying in the significant SNP(s) flanking region are included in further analyses, sometimes only the gene closest to the signal is further investigated. These choices are arbitrary, may end up in choosing the wrong genes or in including a large number of false candidate genes.

Methods have recently been proposed to overcome these limitations using a priori knowledge. The 'candidate pathway' analysis (Raven *et al*., 2013), and 'gene-based' association strategies (Akula *et al*., 2011; Cantor *et al*., 2010; Liu *et al*., 2010) restrict GWA studies to known genes or gene subsets, resizing the multiple testing problem and increasing the power of the analyses.

Bioinformatics pipelines are generally developed for the analysis of human and model organisms, and are often not suited to the analysis of other species. This is the case of the VEGAS software (Liu et al. 2010), developed for gene-based analyses in humans. In fact, this PLINK-based software expects data from 22 autosomes and pre-calculated LD values available for human HapMap phase 2 data.

Here, we introduce MUGBAS, a software based on VEGAS that uses GWAS data and a given annotation (typically genes but any region is suitable) to estimate gene-based and region-based association *P*-values. The suite is species and annotation free, ready to analyze high-density data from any diploid species.

2 Methods

MUGBAS is a suite of scripts built in Bash, R and Python. The suite requires several pre-requisites (such as GNU Parallel, bedtools and a few R libraries) that can be easily fulfilled in a Linux/Unix/Mac environment with a few command lines. Further information on how to meet these requirements is detailed in the online repository.

MUGBAS can be invoked launching the Bash wrapper that checks dependencies and stores user choices.

Required input files are: MAP/PED files in PLINK format with genotype data from at least 200 individuals for a proper LD calculation, an annotation file in BED format (chromosome, start position, end position, name and user custom fields) of the desired genome feature (e.g. a gene or a genomic region) and the GWAS results file with mandatory headers 'SNPNAME' and 'PVALUE'. Multiple association results can be tested at the same time, if provided in separate files in the same directory. A trial dataset is downloadable along with the software release. Please note that for simplicity we provide a gene annotation, but any BED file with any desired feature can be used.

The program starts the analysis assigning SNPs to the feature of the given annotation using bedtools (Quinlan and Hall, 2010). This step is customizable by user defined boundaries (up- and downstream) to catch regulatory regions in LD with the current gene/region. To speed up the analysis, the program subsamples 200 individuals from the input dataset. Two levels of parallelization are implemented: one for the LD and one for the 'gene-wise' or 'region-wise' *P*-value calculation. The latter is particularly useful when more than one GWA results is analysed from the same dataset.

Briefly, MUGBAS first splits the LD estimation in *n* cores (as defined by the user) using the R packages *foreach* / *doParallel* (Analytics and Weston, 2014a, b) and then distributes traits across cores using GNU Parallel (Tange, 2011) and *foreach* / *doParallel*.

Since LD calculation is a slow process, VEGAS benefits of HapMap phase 2 data to speed up the process, while preserving the option for a user defined (human) population using PLINK. To expand these analyses to any species, in MUGBAS LD is estimated with the *r2fast()* function of the *GenABEL* package (Aulchenko *et al.*, 2007), while the gene-wise *P*-value is calculated as in VEGAS. Briefly, the gene-wise test statistic is the sum of the *n* upper tail one degree of freedom chi-square values of a SNP subset (where *n* is the number of SNP mapping in the given gene). In the case of perfect linkage equilibrium, the gene-wise *P*-value would be the one-tailed *P*-value of a chi-square distribution with *n* degrees of freedom. Otherwise *P*-values have to be calculated by simulation and weighted by LD values (Liu *et al.*, 2010).

Following gene/region *P*-value estimation, a False Discovery Rate (FDR) *q*-value statistics is calculated with the base R function *p.adjust()*. Results are stored and enriched with 'Manhattan' and 'underground' plots of the 'gene-wise' *P*-value and the *q*-value, respectively, integrating in the suite publicly available code (http://gettinggeneticsdone.blogspot.com.br/2011/04/annotated-manhattan-plots-and-qq-plots.html).

3 Results and discussion

Gene- and region-based GWAS are now widely used and accepted in genomic research (Mooney *et al.*, 2014). However, non model-species suffer for the lack of specific tools to permit these approaches and facilitate post GWAS investigation. MUGBAS provides researchers dealing with species other than human a set of robust and widely used statistics to conveniently jump from associated markers to the desired functional units.

MUGBAS output is gene/region-oriented and provides useful information on: (i) the location of the analyzed feature, the 'Best SNP' genomic coordinates and original *P*-value; (ii) the gene or region statistics (*P*-value, number of simulations and *q*-value); (iii) the custom annotation of that particular entry; (iv) the position of the SNP with respect to the feature and (v) the chosen boundaries. All these information are useful to track the effect of highly significant markers that map into the region of interest and effectively pinpoint the most probable location to focus on with data mining approaches (i.e.pathway analysis).

Benchmarks with medium-density (50K markers) and high-density (800K markers) panels showed good scalability of MUGBAS distributed in 10 mid-performance cores (Intel Xeon X5675 @ 3.07 GHz). Computational performances for a complete analysis ranged from 8 minutes (medium-density with one trait) to 280 min (high density with three traits). RAM usage peak (16 Gb) was reached for the three traits/high density analysis, while for low density panels RAM had its maximum at 2.5 Gb. Further details are available in the repository.

Acknowledgements

We Dr Jimmy Liu for his advices in building the suite.

Funding

MM and EV were supported by the Doctoral School on the Agro-Food System (Agrisystem) of the Università Cattolica del Sacro Cuore (Italy).

Conflict of Interest: none declared.

References

Akula,N. *et al.* (2011) A network-based approach to prioritize results from genome-wide association studies. *PLoS ONE*, **6**, e24220.

Analytics,R. and Weston,S. (2014a) doParallel: Foreach parallel adaptor for the parallel package.

Analytics,R. and Weston,S. (2014b) foreach: Foreach looping construct for R.

Aulchenko,Y.S. *et al.* (2007) Genomewide rapid association using mixed model and regression: a fast and simple method For genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, **177**, 577–585.

Cantor,R.M. *et al.* (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their APPLICATION. *Am. J. Hum. Genet.*, **86**, 6–22.

Kang,H.M. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.

Liu,J.Z. *et al.* (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.*, **87**, 139.

McCarthy,M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.

Mooney,M.A. *et al.* (2014) Functional and genomic context in pathway analysis of GWAS data. *Trends Genet. TIG*, **30**, 390–400.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Raven,L.-A. *et al.* (2013) Genes of the RNASE5 pathway contain SNP associated with milk production traits in dairy cattle. *Genet. Sel. Evol.*, **45**, 25.

Tange,O. *et al.* (2011) GNU Parallel - The Command-Line Power Tool. *Login USENIX Mag.*, **36**, 42–47.