Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Genetics and population analysis

MultiGWAS: A software for GWAS analysis on tetraploid organisms by integrating results of four GWAS tools

Luis Garreta [1,*], Paula Reyes [1] and Ivania Cerón [1,*]

[1] Colombian Agricultural Research Corporation (Agrosavia), Kilómetro 14, Vía a Mosquera, 250047, Colombia

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Although the GWAS software for polyploids organisms is scarce, other diploid software has also been use to analyze these organisms. This set of available GWAS software provides the possibility to replicate a study by using multiple tools, which helps to the researches to check for true associations when these ones are found by more than one tool. However, each tool has its own characteristics (interface, inputs, outputs, and arguments) which can may cost time and effort to successfully use them.
Results: We have developed an R tool, MultiGWAS, to do GWAS analysis in polyploids organisms by executing and integrating the analysis results from four existing GWAS software: two for polyploids (GWASpoly and SHEsis) and the other two for diploids (Plink and Tassel). MultiGWAS allows users to run two kinds of GWAS: Full and Naive. The first considering the population structure and cryptic relatedness, and the second without these considerations. MultiGWAS receives a simple configuration file and the tool deals with all matters of the GWAS process to finally summarize and report results in a table and intuitive graphics formats to help researches to check potential true or false associations.
Availability and implementation: Source code is freely available at https://github.com/agrosavia/multiGWAS along with examples, documentation and installation instructions.
Contact: lgarreta@agrosavia.co

1 Introduction

One way to validate GWAS results from a study is to replicate the analysis using a different tool, but using different GWAS software in plants to perform GWAS analyzes to the same genomic data often produce different results. Although the underlying model implemented in this kind of software could be different, each software has its own assumptions which could have a strong influence on the results which produces different associations and confuses the researchers from the true associations. Furthermore, many important crops are polyploids and new genomic data of polyploid plants are available much more frequently, but most of the GWAS software to analyze them have been developed for diploid organisms. Currently, two software tailored for polyploid organisms are the R package GWASpoly (Rosyara *et al.*, 2016) and the SHEsis tool (Shen *et al.*, 2016), and widely used diploid GWAS software, as Plink (Purcell *et al.*, 2007) and Tassel (Bradbury *et al.*, 2007), have been also used to analyze polyploids organisms by «diploidizing» the marker data (Lindqvist-Kreuze *et al.*, 2014; Schulz *et al.*, 2016).

However, each tool has its own way of doing things: different user interfaces (GUI or command line based), different input formats (phenotype and genotype formats), different models and algorithm assumptions, and different outputs. Consequently, in light of all these considerations, researchers must spend great effort when they try to replicate GWAS results by using different tools. With this in mind, we developed the MultiGWAS tool that with a single input it performs GWAS analyses by using four different GWAS software in such a way to help researchers in the selection of true associations across GWAS. The tool scans the different results files, calculates new scores and thresholds from resulting markers, select best ranked and significative makers, and create summary tables and plots that show in an intuitive manner the common and different markers found by the four tools.

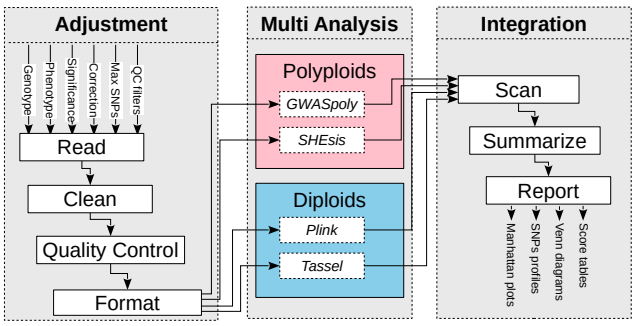

Fig. 1. Central steps in the MultiGWAS tool. In the first stage, inputs are read from a configuration file including: the genotype/phenotype filenames, genome-wide significance threshold α, multiple testing correction method, GWAS model, number of associations to be reported, and TRUE or FALSE whether to use quality control (QC) filters or not. The QC filters are: minor allele frequency, individual missing rate, SNP missing rate, Hardy-Weinberg threshold. Then, genotype and phenotype are cleaned and filtered using the QC filters. In the second stage, the new filtered genotype/phenotype are formated for each GWAS tool which are executed in parallel by MultiGWAS. In the last stage, MultiGWAS scans the output files from each tool and summarizes their results to report them through score tables, Venn diagrams, SNP profiles, and Manhattan plots.

2 Methods and Implementation

MultiGWAS was implemented in R. Genomic data files and parameters are specified in a configuration file (see section 2.1) which is passed as unique argument to MultiGWAS through the command-line interface using a terminal in Linux or Mac systems.

A flowchart of the main central steps involved in the three stages of the MultiGWAS tool is outlined in figure 1. The first stage preprocesses the inputs; the second stage performs GWAS on each tool; and the third stage postprocesses the outputs to summarize the results. These stages are detailed below.

2.1 Adjustment stage

The allowed format for the marker data is the "ACGT" configuration which is suitable for the polyploid tools GWASpoly and SHEsis, but not for the diploid tools Plink and Tassel. So, our MultiGWAS tool "diploidizes" the tetraploid marker data for these two tools by coding each marker in two ways: all possible homozygous genotypes (AAAA, CCCC, GGGG, TTTT), are coded with two nucleotides (AA, CC, GG, and TT); and all possible heterozygous genotypes (e.g. AAAT, ... ,CCCG) are coded with the combination of their reference and alternate alleles calculated from the tetraploid marker (e.g. AT, ... ,CG).

MultiGWAS takes as input a configuration file where the user specify the genomic data for the GWAS analysis (genotype and phenotype files), along with the arguments used by the four tools to perform GWAS quality control (QC) on this data (see below). It starts with a cleaning and filtering steps performed on the genomic data. The first to get the common individuals to work with, and the second, to exclude from the analysis individuals and SNPs that are likely to be of poor-quality. Then, the filtered genotype and phenotype are transformed to the specific formats required for each tool.

The used format of the phenotype and genotype are the same as the the GWASpoly software (Rosyara *et al.*, 2016), except that the phenotype file must include only one trait without covariates. For the genotype file, the first three column correspond to the marker name, chromosome number, and position in the chromosome; while the next columns contain the marker data for each individual in the population codified in the "ACGT" format (e.g. AATT, CCGG, AAAT, GGCG). For the phenotype file, the first column contains the individual name, and the second contains the trait value.

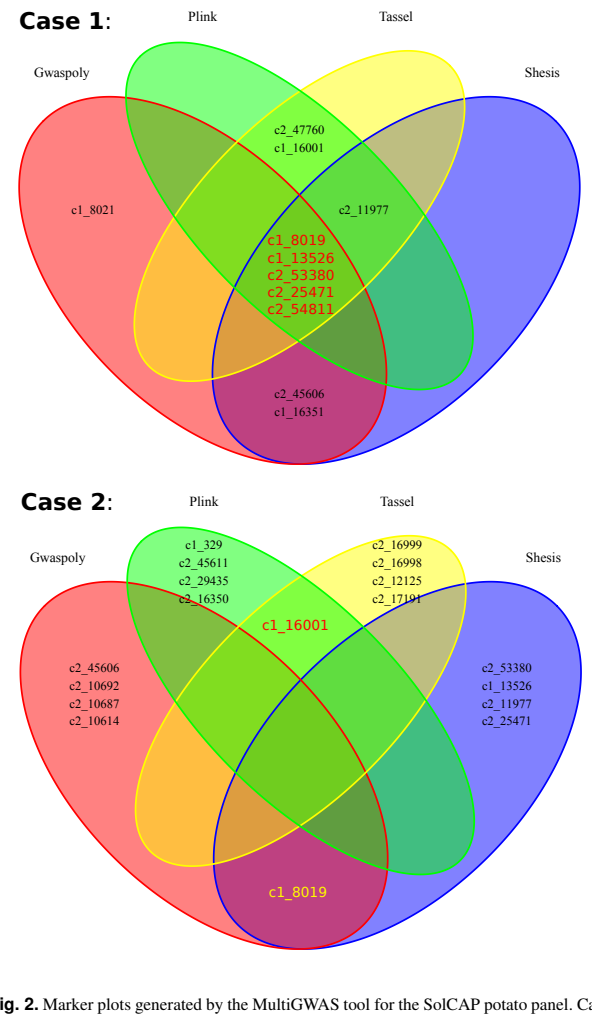

Fig. 2. Marker plots generated by the MultiGWAS tool for the SolCAP potato panel. Case 1: Naive GWAS analysis, the four tools found the same set of five markers (center area, red marker names). Case 2, Full GWAS analysis, the two diploid tools found one common marker (upper-central area, red marker name), but the other two polyploid tools found a different common marker (lower-central area, yellow marker name). In both cases, other marker were found either by two tools or by one tool.

The current arguments included in the configuration file are: (1) The genotype/phenotype filenames, (2) The genome-wide significance threshold α (commonly 0.01 or 0.05), (3) The method for multiple testing correction ("Bonferroni" or "FDR"), (4) The GWAS model ("Full" or "Naive"), (6) The maximum number of best ranked association to be reported, and (6) TRUE or FALSE whether to use quality control (QC) filters or not. The QC filters used by MultiGWAS are: (a) MAF (minor allele frequency), (b) MIND (individual missing rate), (c) GENO (SNP missing rate), and (d) HWE (Hardy-Weinberg threshold).

2.2 Multi analysis stage

Two statistical GWAS models can be conducted by MultiGWAS: Full (Q+K) and Naive (Sharma *et al.*, 2018). The first with control for population structure and individual relatedness; and the second without any of these controls. Controlling of population structure is based on the principal components (PCs). Each software calculates the PCs and uses the top ten as covariates in the association analysis, while controlling of individual relatedness is based on kinship matrices calculated by each software, in the case of Tassel and GWASpoly, and calculated externally

by the king software (Manichaikul *et al.*, 2010), in the case of Plink and SHEsis.

When the filtered genotype and phenotype are ready for the four tools, each tool is executed in parallel by MultiGWAS using a parameterized script for each tool. Each script is called using as arguments the filtered genotype and phenotype along with the set of specified values for the arguments described in section 2.1. If the user has had experience with one of the four GWAS tools, these scripts can be modified to improve the analysis according to the specific tool syntax and arguments.

2.3 Integration stage

After the execution of the four tools by MultiGWAS, their outputs are processed to create tables, marker plots, and Manhattan plots that summarizes their resulting associations. Scores and thresholds for the markers are calculated taking into account: the p-value of each marker, the user specified correction method and significance level α. These values are calculated taking into account only the number of valid genotype calls (nonmissing genotype, phenotype, and covariates).

Two tables and marker plots are created: one for the significative associations which score is greater than the threshold, and other for the best *N* ranked associations, with *N* defined by the user in the configuration file. The Manhattan and QQ plots correspond to the associations found by the GWASpoly tool.

3 Results and Discussion

Figure 2 presents two marker plots generated by MultiGWAS in the analysis of the genomic data for the Solanaceae Coordinated Agricultural Project (SolCAP) potato diversity panel, used to test the GWASpoly software (Rosyara *et al.*, 2016). In the first case (upper plot on the figure 2), it shows that the four tools found a large set of common markers as the GWAS type used was Naive, without any controls for population structure or individual relatedness. However, two markers: the c1_8019 and the c2_25471 are reported by Rosayra et al., the first as the most significative association, and second as the second best ranked association.

In the second case (lower plot on the figure 2), it shows that the set of common markers is reduced, as the GWAS type used was Full GWAS with controls for population structure (10 PCs) and individual relatedness (kinship matrix). Now, the two polyploid tools GWASpoly and SHEsis found the common marker c1_8019, as in the Naive GWAS. But, the two diploid tools found a different common marker, the c1_16001.

4 Conclusion

The MultiGWAS tool allows users to perform a GWAS analysis for tetraploid organism using four GWAS software at the same time: GWASpoly, SHEsis, Plink and Tassel. The tool deals with all the matters of the GWAS process in the four software, and only needs as input a simple configuration file with genomic filenames, values for different quality control filters, and the type of analysis to be performed. Analysis include both Full GWAS with control for population structure and individual relatedness, and Naive GWAS without any control.

The summary reports generated by the MultiGWAS tool provide the user with tables and plots that show the significative and best ranked markers found by each tool. And, the marker plots is a powerful visualization that describes graphically the markers found by each tool and the common marker found by more than one tool, which help users to decide in a more intuitively way the possible true or false associations.

References

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**(19), 2633–2635.

Lindqvist-Kreuze, H., Gastelo, M., Perez, W., Forbes, G. A., De Koeyer, D., and Bonierbale, M. (2014). Phenotypic stability and genome-wide association study of late blight resistance in potato genotypes adapted to the tropical highlands. *Phytopathology*, **104**(6), 624–633.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**(22), 2867–2873.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**(3), 559–575.

Rosyara, U. R., De Jong, W. S., Douches, D. S., and Endelman, J. B. (2016). Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *The Plant Genome*, **9**(2), 0.

Schulz, D. F., Schott, R. T., Voorrips, R. E., Smulders, M. J., Linde, M., and Debener, T. (2016). Genome-wide association analysis of the anthocyanin and carotenoid contents of rose petals. *Frontiers in Plant Science*, **7**(DECEMBER2016), 1–15.

Sharma, S. K., MacKenzie, K., McLean, K., Dale, F., Daniels, S., and Bryan, G. J. (2018). Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. *G3: Genes, Genomes, Genetics*, **8**(10), 3185–3202.

Shen, J., Li, Z., Chen, J., Song, Z., Zhou, Z., and Shi, Y. (2016). SHEsisPlus, a toolset for genetic studies on polyploid species. *Scientific Reports*, **6**, 1–10.