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

OXFORD

Genetics and population analysis

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

L. Garreta [1], I. Cerón-Souza [1], M.R. Palacio [1] and P.H. Reyes-Herrera [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

Summary: The Genome-Wide Association Studies (GWAS) are essential to determine the association between genetic variants across individuals and traits of agronomic interest in non-model tetraploid crops. One way to support the results is by using different tools to validate the reproducibility of the associations. Currently, software for GWAS in polyploids is scarce. Each GWAS software has its characteristics, which can cost time and effort to use them successfully. Here, we present MultiGWAS, a tool to do GWAS analysis in tetraploid organisms by executing in parallel and integrating the results from four existing GWAS software: two available for polyploids (GWASpoly and SHEsis) and two frequently used for diploids (PLINK and TASSEL). The tool deals with all the elements of the GWAS process in the four software, including (1) the use of different control quality filters for the genomic data, (2) the execution of two GWAS models, the full model with control for population structure and individual relatedness and the Naive model without any control. The summary report generated by MultiGWAS provides the user with tables and plots describing intuitively the significant association found by both each one and across four software, which helps users to check for false-positive or false-negative results.
Availability and implementation: The tool is in R. Source code, examples, documentation and installation instructions are available at https://github.com/agrosavia/multiGWAS
Contact: phreyes@agrosavia.co

1 Introduction

GWAS allows analyzing genomics data to identify the set of variants across different individuals of a species that are associated with a specific trait. Due to the advances in the next-gen sequencing technology, currently, the GWAS analysis is extended to non-model tetraploid crops important for agriculture.

One of the main challenges in the GWAS analysis is to identify real from false associations. A reliable method to validate the results is by replicating the study using different software. Currently, the GWAS software to analyze tetraploid species is limited. Therefore, to confirm the GWAS of tetraploid species, it is necessary to use software designed exclusively to a diploid organism. Besides the conversion from the tetraploid to a diploid matrix, this is a time-consuming process. Each software has its characteristics, such as different user interfaces (GUI or command-line based), genotype-phenotype formats, models, algorithm assumptions, and outputs.

To solve this problem, we developed the MultiGWAS tool that performs GWAS analyses for tetraploid species using four software in parallel. Our tool include GWASpoly (Rosyara *et al*., 2016) and the SHEsis tool (Shen *et al*., 2016) that accept polyploid genomic data, and PLINK (Purcell *et al*., 2007) and TASSEL (Bradbury *et al*., 2007) with the use of a "diploidized" genomic matrix. The tool deals with preprocessing data, running four GWAS tools in parallel, and create reports to help the user decide more intuitively the possible true or false associations.

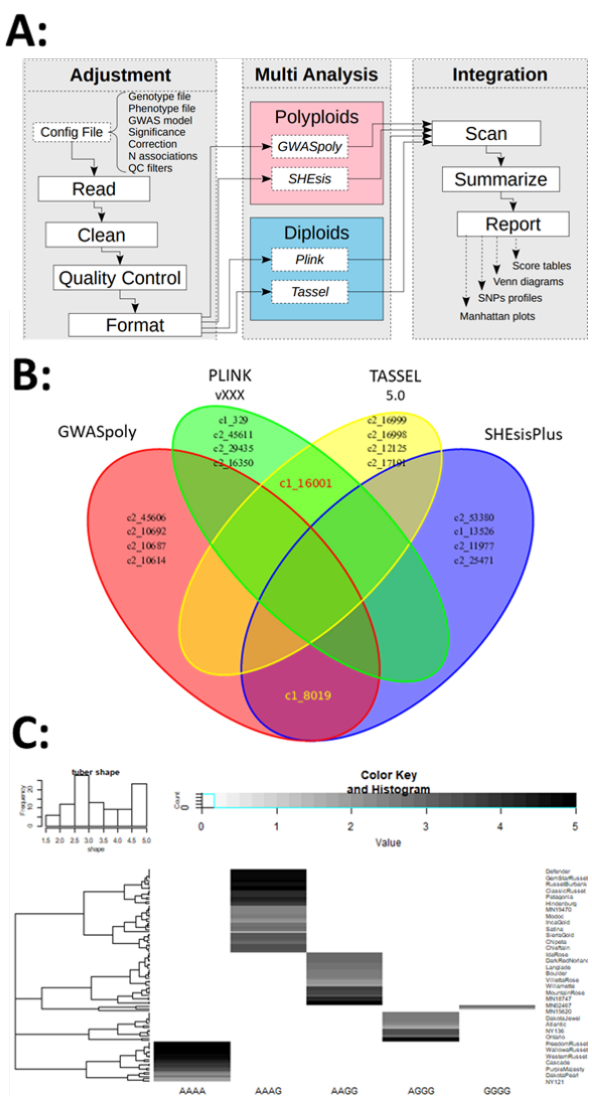

Fig. 1. (A) MultiGWAS flowchart has three stages: adjustment, multi analysis, and integration. In the first stage, we process the configuration file. It includes the genotype/phenotype filenames, genome-wide significance threshold, multiple testing correction methods, GWAS model, number of associations to be reported, and TRUE or FALSE whether to use quality control (QC) filters or not. Then the genotype and phenotype are cleaned and filtered using the QC filters. In the second stage, each GWAS tool runs in parallel. In the last stage, after the output files scanning, a summary of results is generated in a report that contains score tables, Venn diagrams, SNP profiles, and Manhattan plots. The QC filters are minor allele frequency, individual missing rate, SNP missing rate, Hardy-Weinberg threshold. (B) Venn diagrams generated by the MultiGWAS tool for the SolCAP potato panel GWAS for tuber shape. The results are for the Full model in which both diploid software found one marker in common (red text), but the other two polyploid tools found a different marker in common (yellow text). (C) SNP profile is a heatmap for best-ranked SNPs (in this case, c1_8019). The clusters group the samples(rows) and are based only on the genotype. The color represents the phenotype for the trait of interest. On top of the figure, there is the frequency histogram and the color representation for the numeric phenotype.

2 Methods and Implementation

The MultiGWAS tool has three main steps, the adjustment state, the multi analysis stage, and the integration step (Fig. 1A).

Adjustment stage. MultiGWAS takes as input a configuration file where the user specifies the genomics data along with the parameters that will be used by the four tools. It starts by preprocessing the genomic data by selecting individuals present in both genotype and phenotype and excluding individuals and SNPs that have poor quality. Moreover, the format "ACGT" suitable for the polyploid software GWASpoly and SHEsis, is "diploidized" for PLINK and TASSEL. The homozygous tetraploid genotypes are converted to diploid thus: (e.g.,AAAA→AA, CCCC→CC, GGGG→GG, TTTT→TT). Moreover, for tetraploid heterozygous genotypes, the conversion depends on the reference and alternate alleles calculated for each position (e.g. AAAT→AT, ... ,CCCG→CG). After this process, MultiGWAS transform the formats required for each software.

Multi analysis stage. MultiGWAS runs in parallel using two types of statistical models specified in the parameters file, the Full model (Q+K) and Naive (i.e., without any control) (Sharma *et al*., 2018). The Full model (Q+K) controls for both population structure and individual relatedness. For population structure, MultiGWAS uses the Principal Component Analysis (PCA) and takes the top ten PC as a covariate. For relatedness, the tool uses kinship matrices that TASSEL and GWASpoly calculated separately, and for PLINK and SHEsis depends on the King software (Manichaikul *et al*., 2010).

Integration stage. The outputs resulting from the four software are scanned and processed to identify both significant and best-ranked associations. Based on the specification of the configuration file about the correction method and the significant level, MultiGWAS corrects the p-values and calculates the threshold value for each associated marker. The calculation uses the number of valid genotype calls (i.e., non-missing genotypes, phenotypes, and covariates). Then, MultiGWAS summarize the results in tables, Venn diagrams, SNP profiles, and Manhattan plots.

3 Results and Discussion

We tested the MultiGWAS tool with the data for the Solanaceae Coordinated Agricultural Project (SolCAP) potato diversity panel implemented in the GWASpoly software (Rosyara *et al*., 2016) and the tuber shape trait. We present the Venn diagram (Fig. 1B) that summarizes the full model results. The two polyploid software, GWASpoly, and SHEsis identify the c1_8019, also the most significative association from the original study(Rosyara *et al*., 2016). Therefore, it could be considered a real association. Also, the two diploid software, PLINK, and TASSEL, identify a new marker, the c1_16001;. Finally, for each significant association, MultiGWAS generates a heat map figure to summarize the genotype associated with a trait for each individual (Fig. 1C). The complete report for the naive and full model is in the Supplementary information.

References

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

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., 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., 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., Endelman, J. B. (2016). Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *The Plant Genome*, **9**(2).

Sharma, S. K., MacKenzie, K., McLean, K., Dale, F., Daniels, S., 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.