1 Background

Y. Yan et al. (2019) evaluated four GWAS software packages for diploid species, PLINK, TASSEL, GAPIT, and FaST-LMM, in the context of plant genomes and phenotypes, specifically they used two datasets of diploid *Arabidopsis thaliana*. Although, most of the evaluated packages are based on linear regression approaches, the four packages produced association results with different number of SNPs passing a predefined *p-value* threshold for a given GWAS package. It means that well-ranked SNPs from one package can be ranked differently in another, causing difficulty to select the most plausible associations when results from each tool are analyzed separately.

Chen and Zhang (2018) developed a software package called iPAT that incorporates three GWAS software packages for diploid organisms, GAPIT, PLINK, and FarmCPU. The main objective of iPat was to facilitate, using a graphical user interface (GUI), the interaction with these command-line packages, including input data, execution, and presentation of output results. Although iPat helps users to work with these packages with a user-friendly interface, results from the execution of each package are shown separately and the problem of selecting the best associations from multiple package results persists.

2 Methods

2.1 Tools

We have selected four GWAS software tools to be integrated in our multiGWAS tool, two designed specifically for polyploid species as many important crops are polyploids: GWASpoly [10] and SHEsis [12], and another two designed for diploids species and extensively used in humans and plants: PLINK [8, 4] and TASSEL [3], respectively.

As MultiGWAS implements two types of GWAS analysis, naive and full, each tool is called in two different ways. The naive without any additional parameter, but the full with two parameters that take into account for population structure (Q) and relatedness (K) to prevent false associations.

2.1.1 GWASpoly

GWASpoly is a recent R package designed for GWAS in polyploid species that has been used in several studies in plants [2, 5, 11, 13]. It is based on the Q+K linear mixed model with biallelic SNPs that accounts for population structure and relatedness. In addition, to calculate the SNP effect for each genotypic class, GWASpoly provides a general gene action model along with four additional models: additive, simplex dominant, and duplex dominant.

MultiGWAS is using GWASpoly version 1.3. The population structure and relatedness, used in the full model, are estimated using the first five principal

components and the kinship matrix, respectively, both calculated with the algorithms built in GWASpoly. For both, naive and full models, all gene action models are tested for detecting associations.

2.1.2 SHEsis

SHEsis is another program designed for polyploid species that includes single locus association analysis, among others. It is based on a linear regresion model, and it has been used in some studies of animals and humans [9, 6].

MultiGWAS is using the version 1.0 which does not take account for population structure or relatedness, however MultiGWAS externally estimates relatedness for SHEsis by excluding individuals with cryptic first-degree relatedness using the algorithm implemented in PLINK 2.0 (see below).

2.1.3 PLINK

PLINK is one of the most extensively used programs for GWAS in diploids species. It was developed for humans but it is applicable to any species [7]. PLINK includes a range of analysis, including univariate GWAS using two-sample tests and linear regression models.

MultiGWAS is using two versions of PLINK: 1.9 and 2.0. Linear regression from PLINK 1.9 is used to achieve both types of analysis, naive and full. For the full analysis, population structure is estimated using the first five principal components calculated with the PLINK 1.9 built in algorithm. But relatedness is estimated from the kinship coefficients calculated with the PLINK 2.0 built in algorithm, removing the close relatives or individuals with first-degree relatedness.

2.1.4 TASSEL

TASSEL is another common GWAS program based on the Java software. It was developed for maize and it has been used in several studies in plants [1, 14], but like PLINK, it is applicable to any species. For association analysis, TASSEL includes the general lineal model (GLM) and mixed linear model (MLM) that accounts for population structure and relatedness.

MultiGWAS is using TASSEL 5.0, with naive GWAS achieved by the GLM, and full GWAS achieved by the MLM with two parameters: one for population structure, using the first five principal components, and another for relatedness, using the kinship matrix with centered IBS method, both calculated with built in the TASSEL built in algorithms.

3 Results

Although most of the GWAS tools used by MultiGWAS are based on a linear regression approaches, they often produce dissimilar association results for the

same input. For example, computed $pvalues$ for the same set of SNPs are different between tools; SNPs with significant *p-values* for one tool are not significant for the others; or well-ranked SNPs in one tool may be ranked differently in another. To alleviate these difficulties, MultiGWAS produces different visual and tabular outputs intended to help users to compare, select, and interpret the set of possible SNPs associated with a trait of interest. These outputs include score tables, Venn diagrams, and SNP profiles, organized into two categories of SNPs: the best-ranked and significative SNPs. Next, we describe these outputs.

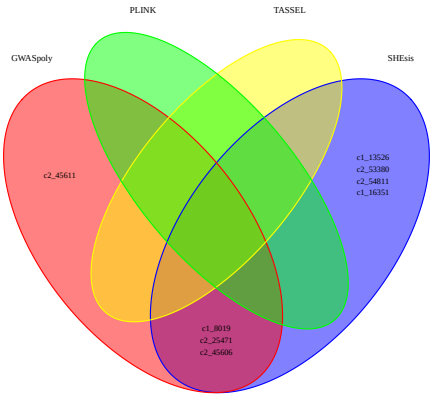
3.1 Significative SNPs

GWAS packages compute $pvalues$ as a measure of association between each individual SNP and the trait of interest. The SNPs are considered statistically significant, and consequently possible true associations, when their $pvalue$ fall below a predefined significance level, usually 0.01 or 0.05.

Here, the MultiGWAS reports the SNPs considered statistically significant by each GWAS package. For that purpose, it provides two views: a tabular and Venn diagram. The table shows detailed information of the SNPs, where both $pvalues$ and significance levels have been scaled as $-log_{10}(pvalue)$. Whereas, Venn diagram visually shows the same SNPs but emphasizing if these were significant either a single package or for more than one. As an example, Figure 1 shows the significative SNPs resulting from running MultiGWAS on a tetraploid potato dataset.

PKG	MDL	CHR	POS	SNP	SCR	THR	SGN
GWASpoly	Full	10	488631	c1_8019	4.78	4.25	TRUE
GWASpoly	Full	10	488084	c2_25471	4.57	4.27	TRUE
GWASpoly	Full	10	482034	c2_45611	4.36	4.27	TRUE
GWASpoly	Full	10	482188	c2_45606	4.68	4.50	TRUE
SHEsis	Full	2	136974	c1_8019	9.47	3.30	TRUE
SHEsis	Full	1	308379	c1_13526	8.45	3.29	TRUE
SHEsis	Full	5	460460	c2_53380	8.24	3.26	TRUE
SHEsis	Full	3	392552	c2_25471	7.82	3.29	TRUE
SHEsis	Full	5	498044	c2_54811	6.96	3.26	TRUE
SHEsis	Full	1	698098	c1_16351	6.02	3.28	TRUE
SHEsis	Full	4	693115	c2_45606	5.95	3.29	TRUE

(a)



(b)

Figure 1: **MultiGWAS views for significant SNPs.** (a) Table with detailed information of significant SNPs found by tool and sorted by decreasing score computed as $-log_{10}(pvalue)$. The information includes: the package reporting the SNP, the GWAS model used by the package, the chromosome, the position in the genome, the name, and the score and threshold to evaluate the SNP as significant. (b) Venn diagram with the SNPs found either for one package or for more than one. It shows one ellipse by package. At the top are the SNPs found by only one package, while at the intersections are those found (shared) by more than one package. For example, the SNP c2_45611 at the top left was found significant only by one tool: GWASpoly, but the three SNPs c1_8019, c2_25471, and c2_45606, were found significative by both packages GWASpoly and SHEsis. However, the other packages, PLINK and TASSEL, did not report any significant SNP.

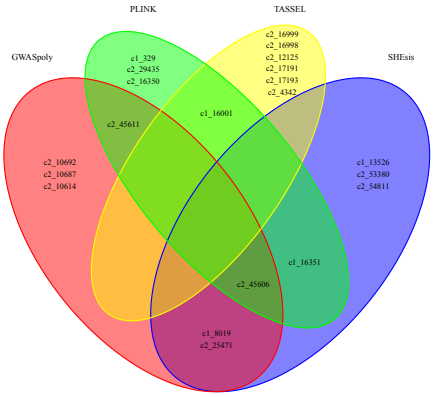
3.2 Best-ranked SNPs

GWAS packages compute $pvalues$ as a measure of association between each individual SNP and the trait of interest. The SNPs are considered statistically significant, and consequently possible true associations, when their $pvalue$ fall below a predefined significance level, usually 0.01 or 0.05. However, these $pvalues$ and significance levels are computed differently by most of the GWAS packages, and SNPs with low $pvalues$ but not so much to be considered significant may represent true associations, the so-called false negatives.

The MultiGWAS tool provides both a table and a Venn diagram to report the best-ranked SNPs for each GWAS package wheter they were assesed significant or not. The table shows detailed information of the SNPs, where both $pvalue$ and significance level are scaled as $-log_{10}(pvalue)$, so that best-ranked SNPs are listed in order their decreasing score. Whereas, Venn diagram shows the same SNPs but emphasizing if they were well ranked either by a single package or by more than one. Figure 1 shows the best-ranked SNPs resulting from running MultiGWAS on a tetraploid potato dataset.

PKG	MDL	CHR	POS	SNP	SCR	THR
GWASpoly	Full	10	488631	c1_8019	4.78	4.25
GWASpoly	Full	10	488084	c2_25471	4.57	4.27
GWASpoly	Full	10	482034	c2_45611	4.36	4.27
...						
PLINK	Full	10	672931	c1_16001	1.76	3.26
PLINK	Full	10	773510	c1_329	1.17	3.30
PLINK	Full	11	514042	c2_29435	1.11	3.25
...						
SHEsis	Full	2	136974	c1_8019	9.47	3.30
SHEsis	Full	1	308379	c1_13526	8.45	3.29
SHEsis	Full	5	460460	c2_53380	8.24	3.26
...						
TASSEL	Full	8	548380	c2_16999	3.60	3.89
TASSEL	Full	8	548380	c2_16998	3.48	3.89
TASSEL	Full	1	714504	c2_12125	2.48	3.89
...						

(a)



(b)

Figure 2: **MultiGWAS best-ranked SNPs from GWAS on tetraploid potato dataset.** Table and Venn diagram of Best-ranked SNPs. (a) Table with detailed information of high-score SNPs by package sorted in decreasing score. The information includes for each SNP, package reporting the SNP, GWAS model used by the package, chromosome, position in the genome, name or ID, score computed as $-log_{10}(pvalue)$, and threshold to define whether the SNP is significant or not. (b) The Venn diagram shows four ellipses, one for each package. At the top of each ellipse are the SNPs found by only one package, while at the intersections are those found (shared) by more than one package. For example, the two SNPs at the botom, c1_8019 and 25471, were ranked with high scores by two packages: GWASpoly and SHEsis.

3.3 Significative SNPs

Here, MultiGWAS reports the SNPs considered statistically significant by each GWAS package. The SNPs are considered statistically significant, and consequently possible true associations, when their $pvalue$ fall below a predefined significance level, usually 0.01 or 0.05.

.

.

Each GWAS tool computes a threshold that is compa

This section shows a table and a Venn diagram for the signifcative SNPs (score is above the significance threshold for each tool)

References

[1] María F. Álvarez, Myrian Angarita, María C. Delgado, Celsa García, José Jiménez-Gomez, Christiane Gebhardt, and Teresa Mosquera. Identification of Novel Associations of Candidate Genes with Resistance to Late Blight in Solanum tuberosum Group Phureja. *Frontiers in Plant Science*, 8:1040, 2017.

[2] Jhon Berdugo-Cely, Raúl Iván Valbuena, Erika Sánchez-Betancourt, Luz Stella Barrero, and Roxana Yockteng. Genetic diversity and association mapping in the colombian central collection of solanum tuberosum L. Andigenum group using SNPs markers. *PLoS ONE*, 12(3), 2017.

[3] Peter J Bradbury, Zhiwu Zhang, Dallas E Kroon, Terry M Casstevens, Yogesh Ramdoss, and Edward S Buckler. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19):2633–2635, 2007.

[4] Christopher C. Chang, Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):1–16, 2015.

[5] Luís Felipe V. Ferrão, Juliana Benevenuto, Ivone de Bem Oliveira, Catherine Cellon, James Olmstead, Matias Kirst, Marcio F. R. Resende, and Patricio Munoz. Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a GWAS Context. *Frontiers in Ecology and Evolution*, 6:107, 2018.

[6] Jie Meng, Kai Song, Chunyan Li, Sheng Liu, Ruihui Shi, Busu Li, Ting Wang, Ao Li, Huayong Que, Li Li, and Guofan Zhang. Genome-wide association analysis of nutrient traits in the oyster Crassostrea gigas: Genetic effect and interaction network. *BMC Genomics*, 20(1):1–14, 2019.

[7] Robert A. Power, Julian Parkhill, and Tulio De Oliveira. Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics*, 18(1):41–50, 2016.

[8] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. De Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575, 2007.

[9] Hui Ping Qiao, Chun Yang Zhang, Zhi Long Yu, Qi Min Li, Yang Jiao, and Jian Ping Cao. Genetic variants identified by GWAS was associated with colorectal cancer in the Han Chinese population. *Journal of Cancer Research and Therapeutics*, 11(2):468–470, 2015.

[10] Umesh R. Rosyara, Walter S. De Jong, David S. Douches, and Jeffrey B. Endelman. Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *The Plant Genome*, 9(2):1–10, 2016.

[11] Sanjeev Kumar Sharma, Katrin MacKenzie, Karen McLean, Finlay Dale, Steve Daniels, and Glenn J. Bryan. Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. *G3: Genes, Genomes, Genetics*, 8(10):3185–3202, 2018.

[12] Yong Yong Shi and Lin He. SHEsis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci (Cell Research (2005) 15, (97-98) DOI: 10.1038/sj.cr.7290272). *Cell Research*, 16(10):851, 2006.

[13] Jiazheng Yuan, Benoît Bizimungu, David De Koeyer, Umesh Rosyara, Zixiang Wen, and Martin Lagüe. Genome-Wide Association Study of Resistance to Potato Common Scab. *Potato Research*, 2019.

[14] Shengkui Zhang, Xin Chen, Cheng Lu, Jianqiu Ye, Meiling Zou, Kundian Lu, Subin Feng, Jinli Pei, Chen Liu, Xincheng Zhou, Ping'an Ma, Zhaogui Li, Cuijuan Liu, Qi Liao, Zhiqiang Xia, and Wenquan Wang. Genome-wide association studies of 11 agronomic traits in cassava (Manihot esculenta crantz). *Frontiers in Plant Science*, 9(April):1–15, 2018.

Threshold:

- **Rosyara2016**:
- **Gumpinger2018**:

Manhattan plots:

- **Powel2016:**

QQ-plots:

- **Powel2016:**

False_negatives:

- **Gumpinger2018:**

Logarithm_scale:

- **Powel2016:**