

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

4       L. Garreta<sup>1</sup>, I. Cerón-Souza<sup>1</sup>, M.R. Palacio<sup>2</sup>, and P.H. Reyes-Herrera<sup>1</sup>

5       <sup>1</sup>Corporación Colombiana de Investigación Agropecuaria  
6       (AGROSAVIA), CI Tibaitatá, Kilómetro 14, Vía a Mosquera, 250047,  
7       Colombia

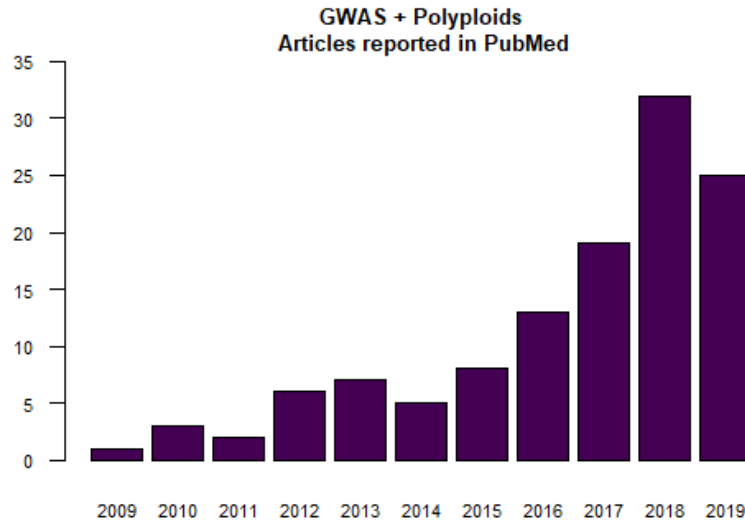
8       <sup>2</sup>Corporación Colombiana de Investigación Agropecuaria  
9       (AGROSAVIA), CI El Mira, Kilómetro 38, Vía Tumaco Pasto,  
10       Colombia

11       May 23, 2020

12       Abstract

13       **Summary:** The Genome-Wide Association Studies (GWAS) are essential to  
14       determine the association between genetic variants across individuals. One way  
15       to support the results is by using different tools to validate the reproducibility of  
16       the associations. Currently, software for GWAS in diploids is well-established  
17       but for polyploids species is scarce. Each GWAS software has its characteris-  
18       tics, which can cost time and effort to use them successfully. Here, we present  
19       MultiGWAS, a tool to do GWAS analysis in tetraploid organisms by executing  
20       in parallel and integrating the results from four existing GWAS software: two  
21       available for polyploids (GWASpoly and SHEsis) and two frequently used for  
22       diploids (PLINK and TASSEL). The tool deals with all the elements of the GWAS  
23       process in the four software, including (1) the use of different control quality  
24       filters for the genomic data, (2) the execution of two GWAS models, the full  
25       model with control for population structure and individual relatedness and the  
26       Naive model without any control. The summary report generated by MultiG-  
27       WAS provides the user with tables and plots describing intuitively the significant  
28       association found by both each one and across four software, which helps users  
29       to check for false-positive or false-negative results.

30  
31       MultiGWAS generates five summary results integrating the four tools. (1)  
32       Score tables with detailed information on the associations for each tool. (2)  
33       Venn diagrams of shared SNPs among the four tools. (3) Heatmaps of signi-  
34       ficative SNP profiles among the four tools. (4) Manhattan and QQ plots for the  
35       association found by each tool. And (5) Chord diagrams for the chromosomes



**Figure 1:** Timeline for articles reported for GWAS studies on polyploid species in PubMed. We present data for completed years.

vs. SNP by each tool. **Contact:** phreyes@agrosavia.co

**Keywords:** GWAS, tetraploids, SNPs,XXX

## 1 Introduction

The Genome-Wide Association Study (GWAS) is used to identify which variants through the whole genome of a large number of individuals are associated with a specific trait [6, 2]. This methodology started with humans and several model plants, such as rice, maize, and *Arabidopsis* [20, 30, 7, 19, 15]. Because of the advances in the next-gen sequencing technology and the decline of the sequencing cost in recent years, there is an increase in the availability of genome sequences of different organisms at a faster rate [10, 11]. Thus, the GWAS is becoming the standard tool to understand the genetic bases of either ecological or economic phenotypic variation for both model and non-model organisms. This increment in GWAS includes complex species such as polyploids (Fig 1) [10, 26].

The GWAS for polyploid species has three related challenges. First, as all GWAS, we should replicate the study as a reliable method to validate the results and recognize real associations. This replication involves finding the same associations either in several replicates from the study population using the same software or testing different GWAS tools among the same study population. This approach involved

the use of different parameters, models, or conditions, to test how consistent the results are [9, 17]. However, the performance of different GWAS software could affect the results. For example, the threshold *pvalue* for SNP significance change through four GWAS software (i.e., PLINK, TASSEL, GAPIT, and FaST-LMM) when sample size varies [31]. It means that well-ranked SNPs from one package can be ranked differently in another.

Second, although there are many GWAS software available to repeat the analysis under different conditions [14], most of them are designed exclusively for the diploid data matrix [4]. Therefore, it is often necessary to "diploidizing" the polyploid genomic data in order to replicate the analysis.

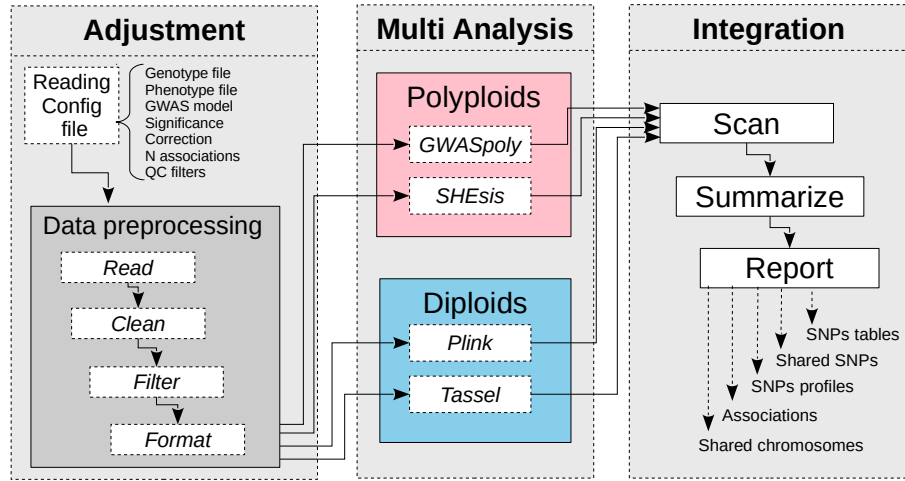
Third, there are very few tools focused on the integration of several GWAS software, to make comparisons under different parameters and conditions across them. As far as we know, there is only two software with this service in mind, such as iPAT and easyGWAS.

The iPAT allows running in a graphic interface three well-known command-line GWAS software such as GAPIT, PLINK, and FarmCPU (Chen and Zhang, 2018). However, the output from each package is separated. On the other hand, the easyGWAS allows running a GWAS analysis on the web using different algorithms. This analysis could run independently of both the computer capacity and operating system. However, it needs either several datasets available or a dataset with a large number of individuals to make replicates in order to compare among algorithms. Moreover, the output from different algorithms is separated [13]. Thus, for both software iPAT and easyGWAS, the integrative and comparative outputs among software or algorithms are missing.

To solve all the three challenges above, we developed the MultiGWAS tool that performs GWAS analyses for tetraploid species using four software in parallel. Our tool include GWASpoly [25] and the SHEsis tool [28] that accept polyploid genomic data, and PLINK [23] and TASSEL [5] with the use of a "diploidized" genomic matrix. The tool deals with preprocessing data, running four GWAS tools in parallel, and create comparative reports from the output of each software to help the user to decide more intuitively the true or false associations.

## 2 Method

The MultiGWAS tool has three main consecutive steps: the adjustment, the multi analysis, and the integration (Fig. ??). In the adjustment step, MultiGWAS processes the configuration file. Then it cleans and filters the genotype and phenotype, and MultiGWAS "diploidize" the genomic data. Next, during the multi analysis, each GWAS tool runs in parallel. Subsequently, in the integration step, the MultiGWAS tool scans the output files from the four packages (i.e., GWASPoly, SHEsis, PLINK, and TASSEL). Finally, it generates a summary of all results that contains score tables, Venn diagrams, SNP profiles, and Manhattan plots.



**Figure 2:** MultiGWAS flowchart has three consecutive steps: adjustment, multi analysis, and integration. The adjustment step manages the input data, reads the configuration file, and preprocessing the input genomic data (genotype and phenotype). The multi analysis step configures and runs the four GWAS packages in parallel. The integration step summarizes and reports results using different tabular and graphical visualizations.

## 2.1 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. Once the configuration file is processed, MultiGWAS preprocess the data that is cleaning, filtering, and checking data quality. The output of this stage corresponds to the inputs for the four programs at the Multi Analysis stage.

### 2.1.1 Reading configuration file

The configuration file includes the following settings that we briefly describe:

**Input genotype and phenotype files:** Currently, MultiGWAS uses two input files, one for genotype and the other for the phenotype. Both data correspond to data matrices with column and row names (Figure 3). The genotype file uses SNP markers in rows and samples in columns (Figure 3a). The phenotype file uses samples in rows and traits in columns (Figure 3b) with the first column corresponding to the sample name and the second column to trait value.

<pre> Marker,Chrom,Pos,Indiv01,Indiv02,Indiv03,... c2_41437,0,805179,AAAG,AAGG,AAGG,... c2_24258,0,1252430,AAGG,AGGG,GGGG,... c2_21332,0,3499519,TTCC,TTCC,TTCC,... ... </pre>	<pre> Individual,Traitname Indiv01, 3.59 Indiv02, 4.07 Indiv03, 1.05 ... </pre>
a	b

**Figure 3: MultiGWAS genotype and phenotype formats.** Both files are in CSV format (Comma Separated Values) and contain as first row the header labels of the columns. Although the header labels are arbitrary, the column order is obligatory. **a.** Genotype file format, where “Marker”, “Chrom”, and “Pos”, correspond to the names for marker name, chromosome, and position in the first three columns respectively. The next columns names correspond to the individual names and the column content correspond to the genotype of each individual. **b.** Phenotype file format, where “Individual” and “Traitname” are the column for the individual ID and the column for the numerical value of the trait, respectively.

**GWAS model:** MultiGWAS is designed to work with quantitative phenotypes and can run GWAS analysis using two types of statistical models that we have called *full* and *naive* models. The *full model* is known in the literature as the Q+K model [32] and includes a control for structure (Q) and relatedness between samples (K). In contrast, the *naive model* does not include any correction. Both models are linear regression approaches and the four GWAS packages used by MultiGWAS implemented variations of them. The *naive* is modeled with Generalized Linear Models (GLMs, Phenotype + Genotype), and the *full* is modeled with Mixed Linear Models (MLMs, Phenotype + Genotype + Structure + Kinship). The default model used by MultiGWAS is the *full model* (Q+K) [32], following this equation:

$$y = X\beta + S\alpha + Qv + Z\mu + e$$

The vector  $y$  represents the observed phenotypes depends on the following factors: the fixed effect vector  $\beta$ , the SNP effects vector  $\alpha$ , the population effect vector  $v$ , the polygene background effect vector  $\mu$ , and, the residual effect vector  $e$ . The  $Q$ , modeled as a fixed effect, refers to the incidence matrix for subpopulation covariates relating  $y$  to  $v$ . Moreover,  $X$ ,  $S$ , and  $Z$  are incidence matrices of ones and zeros relating  $y$  to  $\beta$ ,  $\alpha$ , and  $\mu$ , respectively.

**Genome-wide significance:** GWAS searches SNPs associated with a phenotype trait in a statistically significant manner. A threshold or significance level  $\alpha$  is specified and compared with the *p-value* derived for each association score. Standard significance levels are 0.01 or 0.05 [14, 25], and MultiGWAS uses an  $\alpha$  of 0.05 for the four GWAS packages. However, the adjustment of the threshold is according to each package. For example, GWASpoly and TASSEL calculate the SNP effect for each genotypic class using different gene action models (see “Multi analysis stage”). Therefore, the number of tested markers may be different in each model (see below) that results in different *p-value* thresholds.

**Multiple testing correction:** Due to the massive number of statistical tests performed by GWAS, it is necessary to perform a correction method for multiple hypothesis testing and adjusting the *p-value* threshold accordingly. Two standard methods for multiple hypothesis testing are the false discovery rate (FDR) and the

138 Bonferroni correction. The latter is the default method used by MultiGWAS because  
 139 it is one of the most rigorous. MultiGWAS adjust the threshold below which a *p*-  
 140 value is considered significant, that is  $\alpha/m$ , where  $\alpha$  is the significance level and  $m$   
 141 is the number of tested markers from the genotype matrix.

142 **Number of reported associations:** Criticism has arisen in considering only sta-  
 143 tistically significant associations as the only possible correct associations [29, 18].  
 144 Many low *p*-value associations are closer to being significant, are discarded due to  
 145 the stringent significance levels, and, consequently, increase the number of false  
 146 negatives. To help to analyze both significant and non-significant associations,  
 147 MultiGWAS provides the option to specify the number of best-ranked associations  
 148 (lower *p*-values), adding the corresponding *p*-value to each association found. In this  
 149 way, it is possible to enlarge the number of results, and we can observe replicabil-  
 150 ity in the results for different programs. Nevertheless, MultiGWAS always presents  
 151 each associated SNP with its corresponding *p*-value.

152 **Quality control filters:** A control step is necessary to check the input data for  
 153 genotype or phenotype errors or poor quality that can lead to spurious GWAS re-  
 154 sults. MultiGWAS provides the option to select and define thresholds for the fol-  
 155 lowing filters that control the data quality: Minor Allele Frequency (MAF), individ-  
 156 ual missing rate (MIND), SNP missing rate (GENO), and HardyWeinberg threshold  
 157 (HWE):

- 158 • **MAF of  $x$ :** filters out SNPs with minor allele frequency below  $x$  (default 0.01);
- 159 • **MIND of  $x$ :** filters out all individuals with missing genotypes exceeding  $x*100\%$   
 160 (default 0.1);
- 161 • **GENO of  $x$ :** filters out SNPs with missing values exceeding  $x*100\%$  (default  
 162 0.1);
- 163 • **HWE of  $x$ :** filters out SNPs which have Hardy-Weinberg equilibrium exact test  
 164 *p*-value below the  $x$  threshold.

165 MultiGWAS does the MAF filtering and uses the PLINK package [14] for the other  
 166 three filters: MIND, GENO, and HWE.

### 167 2.1.2 Data preprocessing

168 Once the configuration file is processed, the genomic data is read and cleaned by  
 169 selecting individuals present in both genotype and phenotype. Then, based on pre-  
 170 vious selected quality-control filters and their thresholds, MultiGWAS remove indi-  
 171 viduals and SNPs with poor quality.

172 During this step, the format "ACGT" suitable for the polyploid software GWASpoly  
 173 and SHEsis, is "diploidized" for PLINK and TASSEL. The homozygous tetraploid  
 174 genotypes are converted to diploid thus: AAAA→AA, CCCC→CC, GGGG→GG, TTTT→TT.

175 Moreover, for tetraploid heterozygous genotypes, the conversion depends on the  
176 reference and alternate alleles calculated for each position (e.g., AAAT→AT, ... ,CCCG→CG).  
177 After this process, the genomic data, genotype and phenotype, are converted to  
178 the specific formats required for each of the four GWAS packages.

## 179 2.2 Multi analysis stage

180 MultiGWAS runs in parallel using two types of statistical models specified in the  
181 parameters file, the Full model (Q+K) and Naive (i.e., without any control) [27].  
182 The Full model (Q+K) controls for both population structure and individual related-  
183 ness. For population structure, MultiGWAS uses the Principal Component Analysis  
184 (PCA) and takes the top five PC as covariates. For relatedness, MultiGWAS uses  
185 kinship matrices that TASSEL and GWASpoly calculated separately, and for PLINK  
186 and SHEsis, relatedness depends of kinship coefficients calculated with the PLINK  
187 2.0 built-in algorithm [8].

188 As MultiGWAS implements two types of GWAS analysis, naive and full, each  
189 tool is called in two different ways.

### 190 2.2.1 GWASpoly

191 GWASpoly [25] is an R package designed for GWAS in polyploid species used in  
192 several studies in plants [3, 12, 27, 33]. GWASpoly uses a Q+K linear mixed model  
193 with biallelic SNPs that account for population structure and relatedness. Also, to  
194 calculate the SNP effect for each genotypic class, GWASpoly provides eight gene  
195 action models: general, additive, simplex dominant alternative, simplex dominant  
196 reference, duplex dominant alternative, and duplex dominant. As a consequence,  
197 the number of statistical test performed can be different in each action model and  
198 so thresholds below which the *p-values* are considered significant.

199 MultiGWAS is using GWASpoly version 1.3, employing all gene action models  
200 to find associations and repoting the top *N* best-ranked (the SNPs with lowest *p-*  
201 *values*), where *N* is defined by the user in the input configuration file. The *full*  
202 model used by GWASpoly includes the population structure and relatedness, which  
203 are estimated using the first five principal components and the kinship matrix, re-  
204 spectively, both calculated with the GWASpoly built-in algorithms.

### 205 2.2.2 SHEsis

206 SHEsis is another program designed for polyploid species that includes single locus  
207 association analysis, among others. It is based on a linear regresion model, and it  
208 has been used in some studies of animals and humans [24, 21].

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

### 213 2.2.3 PLINK

214 PLINK is one of the most extensively used programs for GWAS in diploids species. It  
215 was developed for humans but it is applicable to any species [22]. PLINK includes  
216 a range of analysis, including univariate GWAS using two-sample tests and linear  
217 regression models.

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

### 224 2.2.4 TASSEL

225 TASSEL is another common GWAS program based on the Java software. It was de-  
226 veloped for maize and it has been used in several studies in plants [1, 34], but like  
227 PLINK, it is applicable to any species. For association analysis, TASSEL includes the  
228 general lineal model (GLM) and mixed linear model (MLM) that accounts for popu-  
229 lation structure and relatedness. And, in the same manner that GWASPoly, TASSEL  
230 provides three gene action models to calculate the SNP effect of each genotype class:  
231 general, additive, and dominant, and so the significance threshold depends of each  
232 action model.

233 MultiGWAS is using TASSEL 5.0, with all gene action models used to find the  $N$   
234 best-ranked associations and reporting the top  $N$  best-ranked associations (SNPs  
235 with lowest  $p$ -values). Naive GWAS is achieved by the GLM, and full GWAS is  
236 achieved by the MLM with two parameters: population structure that uses the first  
237 five principal components, and relatedness that uses the kinship matrix with centered  
238 IBS method, both calculated with the TASSEL built-in algorithms.

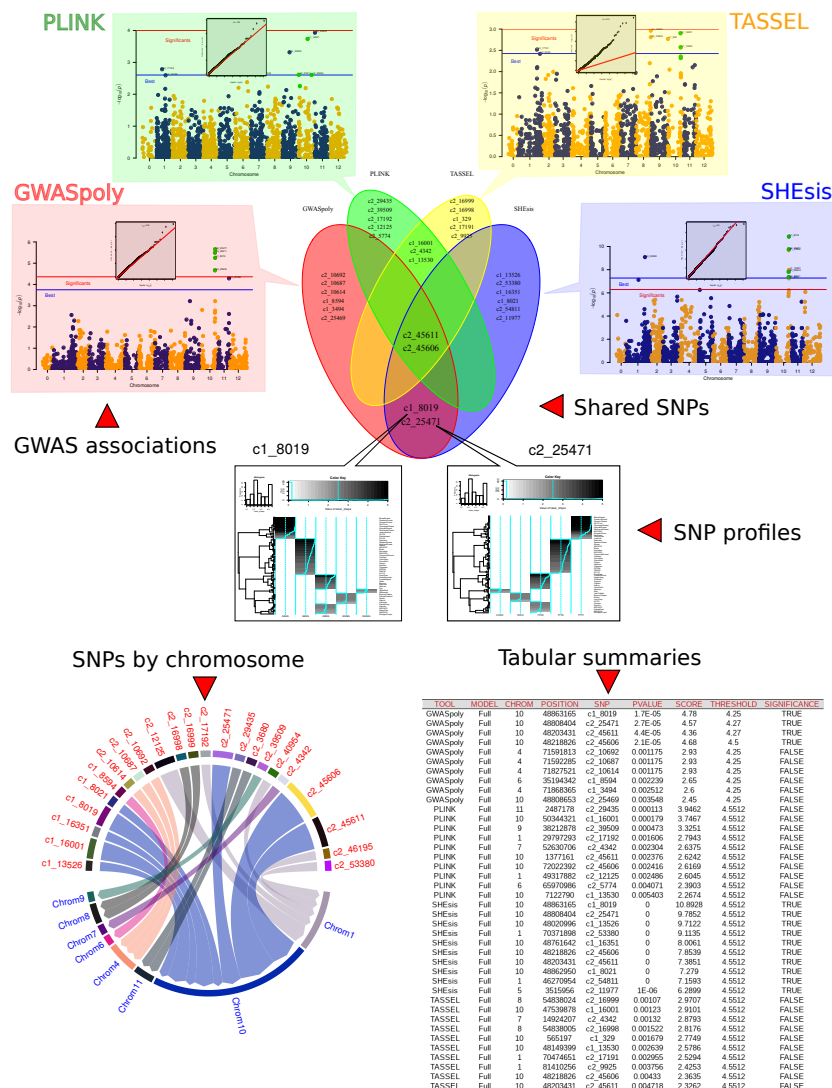
## 239 2.3 Integration stage.

240 The outputs resulting from the four GWAS packages are scanned and processed  
241 to identify both significant and best-ranked associations with  $p$ -values lower than  
242 and close to a significance threshold, respectively. MultiGWAS corrects (correction  
243 method defined in the configuration file) the  $p$ -values and calculates the threshold  
244 value for each associated marker. The calculation uses the number of statistical tests  
245 performed in the GWAS study, but this threshold can be different for each GWAS  
246 package as some of them use more than one gene action model to calculate the SNP  
247 effect of each genotypic class (see “Methods”).

248 At this stage, MultiGWAS integrates the results to evaluate reproducible results  
249 among tools (Fig 4). But, it still reports a summary for the results of each tool:

- 250 • A Quantile-Quantile (QQ) plots for the resultant  $p$ -values of each tool and  
251 the corresponding inflation factor  $\lambda$  to asses the degree of the test statistic  
252 inflation.





**Figure 4: Reports presented by MultiGWAS.** For each tool, first a QQ plot that assesses the resultant p-values. Second, a Manhattan plot for each tool with two lines, blue and red, respectively, is the lower limit for the best ranked and significant SNPs. We present two Venn diagrams, one for the significant SNPs and one for N best-ranked SNPs of each tool. We show the results for GWASpoly, PLINK, TASSEL, and SHEsis in red, green, yellow, and blue, respectively. For each SNP that is in the intersection; thus, that is predicted by more than one tool we provide SNP profile. **SNPs by chromosome** chord diagrams shows how the strongest associations are mostly found on few chromosomes. And we also present tabular summaries with details of significant and best-ranked associations.

- A Manhattan plot of each tool with two lower thresholds, one for the best-ranked SNPs, and another for the significant SNPs.

To present the replicability, we use two sets: (1) the set of all the significant SNPs provided by each tool and (2) the set of all the best-ranked SNPs. For each set, we present a Venn diagram that shows SNPs predicted exclusively by one tool and intersections that help to identify the SNPs predicted by one, two, three, or all the tools. In addition, we provide detailed tables for the two sets.

For each SNP identified more than once, we provide what we call the SNP profile. That is a heat diagram for a specific SNP, where each column is a genotype state AAAA, AAAB, AAB, ABBB, and BBBB. And each row corresponds to a sample. Samples with close genotypes form together clusters. Thus to generate the clusters, we do not use the phenotype information. However, we present the phenotype information in the figure as the color. This figure visually provides information regarding genotype and phenotype information simultaneously for the whole population. We present colors as tones between white and black for color blind people.

MultiGWAS generates a report, one document with the content previously described. Besides, there is a folder with the individual figures just in case the user needs one. In the supplementary information, we include a report and a description of the report content (supplementary information XXX)

In the following section, we present the results applied to a public dataset.

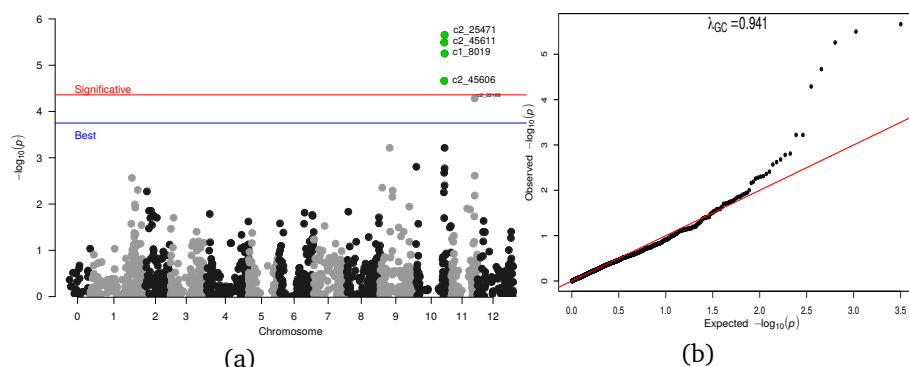
### 3 Results

Although most of the GWAS packages used by MultiGWAS use linear regression approaches, they often produce different association results for the same input. For example, computed *p-values* for the same set of SNPs are different between packages; SNPs with significant *p-values* for one package maybe not significant for the others, or well-ranked SNPs in one package may be ranked differently in another. To alleviate these difficulties, MultiGWAS produces four types of outputs using different graphics and tabular views, including score tables, Venn diagrams, Manhattan and Q-Q plots, and SNP profiles. We designed these outputs to help users visually to compare, select, and interpret the set of possible SNPs associated with a trait of interest.

As an example of the functionality of the tool, here we show the results of running MultiGWAS tool in the genomic data from a tetraploid potato diversity panel, genotyped and phenotyped as part of the USDA-NIFA Solanaceae Coordinated Agricultural Project (SolCAP) [16]. The reports include: significant SNPs, best-ranked SNPs, profile SNPs, and visualization of associations. First, the best-ranked SNPs (Figure 6.b), where the SNP c2\_45606 was evaluated with a high score by the four packages, but other four SNPs were also ranked with high scores by two packages simultaneously. Second, the significant SNPs (Figure 6.c), where the two polyploid software, GWASpoly and SHEsis, found as significant three SNPs, c1\_8019, c2\_25471, and c2\_45606. In particular, the c1\_8019 was also the most significant association found in the same potato dataset analyzed by Rosyara et al. (2016).

For each SNPs identified by more than one package, we provide the SNP profile (Figure 7), where for each significant association, a heat map figure is generated to summarize the genotype associated with a trait for each individual. Here we present, the SNP profile for the SNPs ranked among the best for the four tools c2\_45606. We also include the SNP profile for the c1\_8019 the most significant SNP for both polyploid tools. The SNP profiles for the markers present in the inter-

sections are in the (supplementary information XXX) And fourth, the visualization of associations (Figure 5), where for each package, a Manhattan and QQ plots are generated using special marks to help to identify significant, best-ranked, and shared SNPs (found by more than one tool).



**Figure 5: MultiGWAS visualization of associations.** MultiGWAS adds special marks to the Manhattan and QQ plots to help identify different types of SNPs: (a) In Manhattan plots, significant SNPs are above a red line, best-ranked SNPs are above a blue line, and shared SNPs (See Figure 6.b) are colored in green (b) In QQ plots, a red diagonal line indicates the expectation, so potential associations can be observed when the number of SNPs deviating from the diagonal is small, as in the case of monogenic traits, or when this number is somewhat higher, as in the case of truly polygenic traits. However, deviations for a high number of SNPs could reflect inflated  $p$ -values owing to population structure or cryptic relatedness.

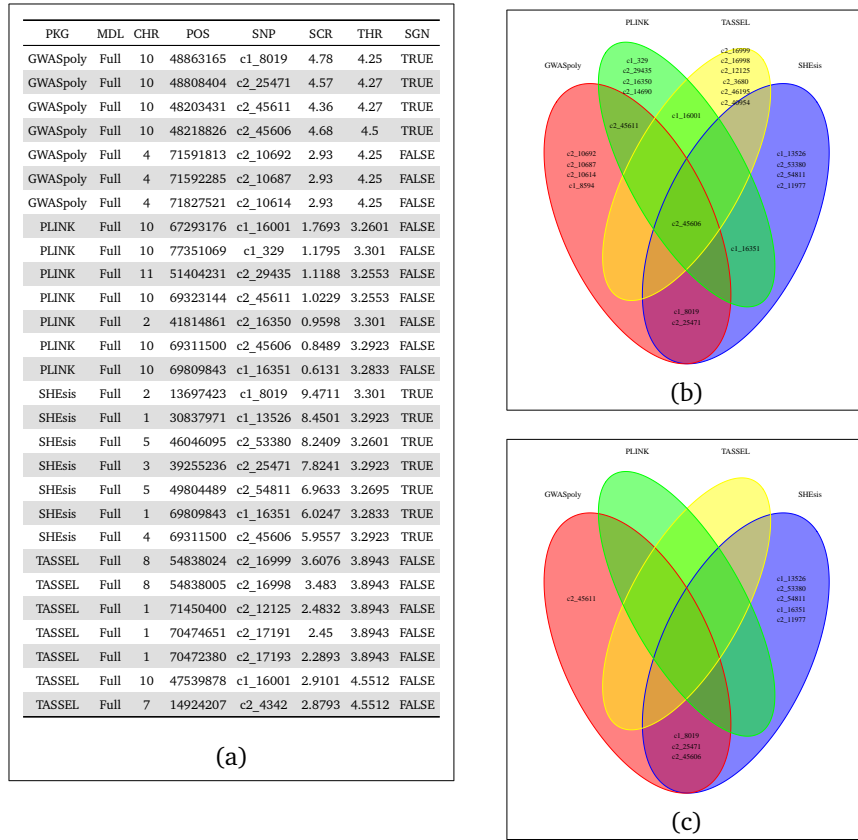
The complete report from MultiGWAS for the naive and full model is in the Supplementary information (<https://github.com/agrosavia-bioinformatics/multiGWAS>)

### 3.1 Visualization of shared SNPs

GWAS packages rely on  $p$ -value as a measure of association between each individual SNP and the trait of interest. The SNPs are considered statistically significant, and so possible true associations, when their  $p$ -value falls below a predefined significance level, usually 0.01 or 0.05. But, most GWAS packages compute differently both  $p$ -values and significance levels, it could result in non-significant SNPs. Consequently, it is important to know the significant SNPs. It is equally important to know the best-ranked SNPs closer to being statistically significant, as they may represent important associations to consider for posterior analysis (e.g. false negatives).

MultiGWAS provides tabular and graphic views to report in an integrated way both the best-ranked and significant SNPs identified by the four GWAS packages (see Figure 6). Both  $p$ -values and significance levels have been scaled as  $-\log_{10}(p\text{-value})$  to give high scores to the best statistically evaluated SNPs.

First, the best-ranked SNPs correspond to the top-scored  $N$  SNPs that are the ones corresponding to the  $N$  lower p-values. 6.a) and in a Venn diagram (Figure 6.b). The table lists them by package and sorts by decreasing score, whereas the Venn diagram shows them emphasizing if these ones were best-ranked either in a single package or in several at once (shared). And second, the significant SNPs correspond to the ones assessed statistically significant by each package (depending on the threshold), they are shown in a Venn diagram (Figure 6.b), and they are also shown in the SNPs table, marked with significance TRUE and score greater than threshold, columns SGN, SCR, and THR, respectively in the table of the Figure6.a.

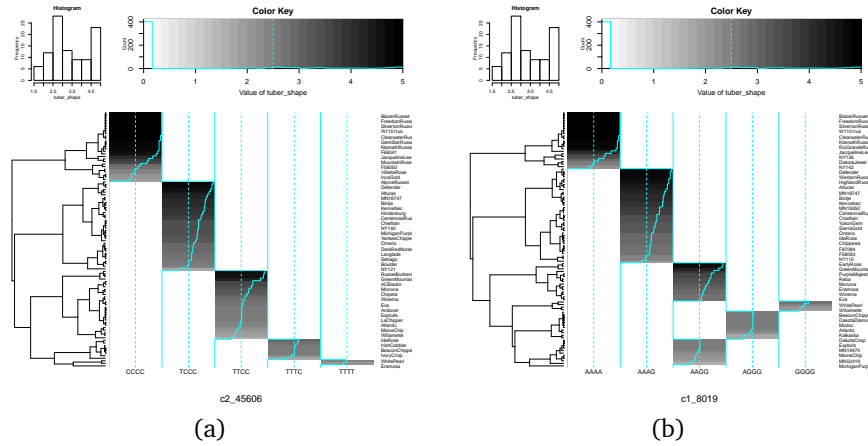


**Figure 6: Visualization of shared SNPs.** Tabular and graphical views of the best-ranked and significant SNPs identified by the four packages. (a) Tabular view with detailed information of each SNPs, including: package name (PKG), GWAS model used (MDL), chromosome (CHR), position in the genome (POS), ID (SNP), score (SCR), threshold (THR), and significance flag (SGN), whether the SNP was evaluated statistically significant or not (score > threshold). (b) Venn diagram with the best-ranked SNPs, showing that one SNP was shared by the four packages (c2\_45606), other two only by the two polyploid packages GWASpoly and SHESis (c1\_8019 and c2\_25471), and other one only by the two diploid packages PLINK and TASSEL (c1\_16001). (c) Venn diagram with the significant SNPs, showing that only three SNPs (c1\_8019, c2\_25471, and c2\_45606) were evaluated as significant by the two polyploid packages GWASpoly and SHESis.

For each SNP predicted by more than one too, MultiGWAS creates a two-dimensional

330 representation, what we called SNP profile. This profiles enables to visualize each  
 331 trait by individuals and genotypes as rows and columns, respectively (Figure 7).  
 332 At the left, the individuals are grouped in a dendrogram by their genotype. At the  
 333 right, there is the name or ID of each individual. At the bottom, the genotypes are  
 334 ordered from left to right, starting from the major to the minor allele (i.e., AAAA,  
 335 AAAB, AABB, ABBB, BBBB). At the top, there is a description of the trait based on  
 336 a histogram of frequency (top left) and by an assigned color for each numerical  
 337 phenotype value using a grayscale (top right). Thus, each individual appears as  
 338 a colored line by its phenotype value on its genotype column. For each column,  
 339 there is a solid cyan line with the mean of each column and a broken cyan line that  
 340 indicates how far the cell deviates from the mean.

341 Because each multiGWAS report shows one specific trait at a time, the histogram  
 342 and color key will remain the same for all the best-ranked SNPs.



**Figure 7: SNP profiles.** SNP profiles for two of the best-ranked significant SNPs shown in the figure 6.b. (a) SNP c2\_45606 best-ranked by the four packages (central intersection of the Venn diagram Figure 6.b) (b) SNP c1\_8019 best-ranked by the two tetraploid packages (Figure 6.b), and also identified as significant by the same packages (at the bottom of the Figure 6.a).

## 343 4 Availability and implementation:

344 The core of the MultiGWAS tool was developed in R and users can interact with the  
 345 tool by either a command line interface (CLI) developed in R or a graphical user  
 346 interface (GUI) developed in Java (Figure 8). Source code, examples, documenta-  
 347 tion and installation instructions are available at <https://github.com/agrosavia-bioinformatics/multiGWAS>.

### 349 4.1 Input parameters

350 MutiGWAS uses as the only input a simple configuration text file where users set the  
 351 values for the main parameters that drives the GWAS process. The input parameters

352 include: the output folder where results will be written, input genotype/phenotype  
353 filenames, genome-wide significance threshold, method for multiple testing cor-  
354 rection, GWAS model, number of associations to be reported, and TRUE or FALSE  
355 whether to use quality control filters or not. The filters are: minor allele frequency,  
356 individual missing rate, SNP missing rate, and Hardy-Weinberg threshold.

357 The configuration file can be created either using a general text editor or using  
358 the GUI application. In the first case, the file must have the structure shown in the  
359 Figure 8.a, where parameter names and values are separated by colon, filenames  
360 are enclosed in quotation marks, and TRUE or FALSE indicates whether filters are ap-  
361 plied or not. Moreover examples for the config file [https://github.com/agrosavia-](https://github.com/agrosavia-bioinformatics/MultiGWAS/tree/master/examples)  
362 [bioinformatics/MultiGWAS/tree/master/examples](https://github.com/agrosavia-bioinformatics/MultiGWAS/tree/master/examples)

363 In the second case, the user creates the config file in a simple and straightforward  
364 way using the input parameter view from the GUI application (Figure 8.b) and  
365 clicking the “Save” button.

## 366 4.2 Using the command line interface

367 The execution of the tool in command line is simple, it only needs to open a linux  
368 console, change to the folder where the configuration file was created, and type the  
369 name of the executable tool followed by the filename of the configuration file, like  
370 this:

```
371 multiGWAS full.config
```

372 Then, the tool starts the execution, showing information of the process in the  
373 console window, and when it finishes the results are saved to a new subfolder called  
374 “outgwas/reports”. Results include a full html report containing the different views  
375 described in the results section, along with the original graphics and summary tables  
376 created by MultiGWAS and used to create the html report. Additionally, results  
377 include the preprocessed tables of the main outputs generated by the four GWAS  
378 packages used by MultiGWAS.

## 379 4.3 Using the graphical user interface

380 The MultiGWAS GUI application can be executed either by running from a Linux  
381 console the *jmultiGWAS* command or by clicking on the Java application file *JMulti-*  
382 *GWAS.jar* located in the “multiGWAS/sources” subfolder. After it opens, it shows a  
383 main frame with four tabs at the top (Figure 8b): “Inputs”, “Outputs”, “Results”, and  
384 “Files”. The “Inputs” tab shows the form to create the configuration file and run the  
385 application. The “Outputs” tab shows the messages from the running process after  
386 it starts the execution. The “Results” tab shows the full html report described above.  
387 And the “Files” tab shows an embedded file browser pointing to the subfolder that  
388 contains the original files used in the html report and described above.

```

Test01.config
default:
  genotypeFile      : "example-genotype.tbl"
  phenotypeFile     : "example-phenotype.tbl"
  significanceLevel : 0.05
  correctionMethod  : "Bonferroni"
  gwasModel         : Full
  nBest             : 5
  filtering         : TRUE
  MAF               : 0.01
  MIND              : 0.1
  GENO              : 0.1
  HWE               : 1e-10

```

(a)

(b)

**Figure 8: MultiGWAS inputs and interaction.** MultiGWAS uses as input a simple configuration text file and can be executed using either a command line interface script in R (CLI) or a graphic user interface application in Java (GUI). (a) An example of a configuration text file named “Test01.config” including the parameters that drive the GWAS process. It can be created using a general text editor or using the GUI application (see below) (b) Main view of the MultiGWAS GUI application (“Inputs” view) where users can create the configuration file by setting values for input parameters. The GUI contains other three views: “Outputs” view shows the logs of the running process. “Results” view shows a report in html format with the tabular and graphics described in the results section. And, the “Files” view shows an embedded file manager pointing to the subfolder that contains the files created by MultiGWAS and used to create the report.

## 5 Discussion

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

## References

- [1] María F. Álvarez et al. “Identification of Novel Associations of Candidate Genes with Resistance to Late Blight in Solanum tuberosum Group Phureja”. In: *Frontiers in Plant Science* 8 (2017), p. 1040. ISSN: 1664-462X. DOI: 10.

- 3389/fpls.2017.01040. URL: <http://journal.frontiersin.org/article/10.3389/fpls.2017.01040/full>.
- [2] Ferdouse Begum et al. "Comprehensive literature review and statistical considerations for GWAS meta-analysis". In: *Nucleic acids research* 40.9 (2012), pp. 3777–3784.
- [3] Jhon Berdugo-Cely et al. "Genetic diversity and association mapping in the colombian central collection of solanum tuberosum L. Andigenum group using SNPs markers". In: *PLoS ONE* 12.3 (2017). ISSN: 19326203. DOI: 10.1371/journal.pone.0173039.
- [4] Peter M. Bourke et al. "Tools for Genetic Studies in Experimental Populations of Polyploids". In: *Frontiers in Plant Science* 9 (2018), p. 513. ISSN: 1664-462X. DOI: 10.3389/fpls.2018.00513. URL: <http://journal.frontiersin.org/article/10.3389/fpls.2018.00513/full>.
- [5] Peter J Bradbury et al. "TASSEL: software for association mapping of complex traits in diverse samples". In: *Bioinformatics* 23.19 (2007), pp. 2633–2635. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm308. URL: <https://doi.org/10.1093/bioinformatics/btm308>.
- [6] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. "Prioritizing GWAS results: a review of statistical methods and recommendations for their application". In: *The American Journal of Human Genetics* 86.1 (2010), pp. 6–22.
- [7] Jun Cao et al. "Whole-genome sequencing of multiple Arabidopsis thaliana populations". In: *Nature genetics* 43.10 (2011), p. 956.
- [8] Christopher C. Chang et al. "Second-generation PLINK: Rising to the challenge of larger and richer datasets". In: *GigaScience* 4.1 (2015), pp. 1–16. ISSN: 2047217X. DOI: 10.1186/s13742-015-0047-8. arXiv: 1410.4803.
- [9] Rishika De, William S Bush, and Jason H Moore. "Bioinformatics Challenges in Genome-Wide Association Studies (GWAS)". In: *Clinical Bioinformatics*. Ed. by Ronald Trent. New York, NY: Springer New York, 2014, pp. 63–81. ISBN: 978-1-4939-0847-9. DOI: 10.1007/978-1-4939-0847-9\_5. URL: [https://doi.org/10.1007/978-1-4939-0847-9\\_5](https://doi.org/10.1007/978-1-4939-0847-9_5).
- [10] Robert Eklom and Juan Galindo. "Applications of next generation sequencing in molecular ecology of non-model organisms". In: *Heredity* 107.1 (2011), pp. 1–15.
- [11] Hans Ellegren. "Genome sequencing and population genomics in non-model organisms". In: *Trends in ecology & evolution* 29.1 (2014), pp. 51–63.
- [12] Luís Felipe V. Ferrão et al. "Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a GWAS Context". In: *Frontiers in Ecology and Evolution* 6 (2018), p. 107. ISSN: 2296-701X. DOI: 10.3389/fevo.2018.00107. URL: <https://www.frontiersin.org/articles/10.3389/fevo.2018.00107/full>.



- 437 [13] Dominik G Grimm et al. “easyGWAS: A Cloud-Based Platform for Comparing  
438 the Results of Genome-Wide Association Studies”. In: *The Plant Cell* 29.1  
439 (2017), pp. 5–19. ISSN: 1040-4651. DOI: 10.1105/tpc.16.00551. URL:  
440 <http://www.plantcell.org/content/29/1/5>.
- 441 [14] Anja C Gumpinger et al. *Methods and Tools in Genome-wide Association Stud-*  
442 *ies*. Vol. 1819. 2018. ISBN: 9781493986187.
- 443 [15] Bin Han and Xuehui Huang. “Sequencing-based genome-wide association  
444 study in rice”. In: *Current opinion in plant biology* 16.2 (2013), pp. 133–138.
- 445 [16] Candice N. Hirsch et al. “Retrospective view of North American potato (*Solanum*  
446 *tuberosum* L.) breeding in the 20th and 21st centuries”. In: *G3: Genes, Genomes,*  
447 *Genetics* 3.6 (2013), pp. 1003–1013. ISSN: 21601836. DOI: 10.1534/g3.  
448 113.005595.
- 449 [17] “How to interpret a genome-wide association study”. In: *JAMA - Journal*  
450 *of the American Medical Association* 299.11 (2008), pp. 1335–1344. ISSN:  
451 00987484. DOI: 10.1001/jama.299.11.1335.
- 452 [18] Avjinder S Kaler and Larry C Purcell. “Estimation of a significance threshold  
453 for genome-wide association studies”. In: *BMC Genomics* 20.1 (2019), p. 618.  
454 ISSN: 1471-2164. DOI: 10.1186/s12864-019-5992-7. URL: [https :](https://doi.org/10.1186/s12864-019-5992-7)  
455 [//doi.org/10.1186/s12864-019-5992-7](https://doi.org/10.1186/s12864-019-5992-7).
- 456 [19] Arthur Korte and Ashley Farlow. “The advantages and limitations of trait  
457 analysis with GWAS: a review”. In: *Plant methods* 9.1 (2013), p. 29.
- 458 [20] Gordan Lauc et al. “Genomics meets glycomics—the first GWAS study of hu-  
459 man N-glycome identifies HNF1 $\alpha$  as a master regulator of plasma protein  
460 fucosylation”. In: *PLoS genetics* 6.12 (2010).
- 461 [21] Jie Meng et al. “Genome-wide association analysis of nutrient traits in the  
462 oyster *Crassostrea gigas*: Genetic effect and interaction network”. In: *BMC*  
463 *Genomics* 20.1 (2019), pp. 1–14. ISSN: 14712164. DOI: 10.1186/s12864-  
464 019-5971-z.
- 465 [22] Robert A. Power, Julian Parkhill, and Tulio De Oliveira. “Microbial genome-  
466 wide association studies: lessons from human GWAS”. In: *Nature Reviews*  
467 *Genetics* 18.1 (2016), pp. 41–50. ISSN: 14710064. DOI: 10.1038/nrg.  
468 2016.132.
- 469 [23] Shaun Purcell et al. “PLINK: A tool set for whole-genome association and  
470 population-based linkage analyses”. In: *American Journal of Human Genetics*  
471 81.3 (2007), pp. 559–575. ISSN: 00029297. DOI: 10.1086/519795.
- 472 [24] Hui Ping Qiao et al. “Genetic variants identified by GWAS was associated  
473 with colorectal cancer in the Han Chinese population”. In: *Journal of Cancer*  
474 *Research and Therapeutics* 11.2 (2015), pp. 468–470. ISSN: 19984138. DOI:  
475 10.4103/0973-1482.150346.

- 476 [25] Umesh R. Rosyara et al. “Software for Genome-Wide Association Studies  
477 in Autopolyploids and Its Application to Potato”. In: *The Plant Genome* 9.2  
478 (2016), pp. 1–10. ISSN: 1940-3372. DOI: 10.3835/plantgenome2015.  
479 08.0073. URL: [https://dl.sciencesocieties.org/publications/](https://dl.sciencesocieties.org/publications/tpg/abstracts/9/2/plantgenome2015.08.0073)  
480 [tpg/abstracts/9/2/plantgenome2015.08.0073](https://dl.sciencesocieties.org/publications/tpg/abstracts/9/2/plantgenome2015.08.0073).
- 481 [26] Anna W Santure and Dany Garant. “Wild GWAS—association mapping in  
482 natural populations”. In: *Molecular ecology resources* 18.4 (2018), pp. 729–  
483 738.
- 484 [27] Sanjeev Kumar Sharma et al. “Linkage disequilibrium and evaluation of genome-  
485 wide association mapping models in tetraploid potato”. In: *G3: Genes, Genomes,*  
486 *Genetics* 8.10 (2018), pp. 3185–3202. ISSN: 21601836. DOI: 10.1534/g3.  
487 118.200377.
- 488 [28] Jiawei Shen et al. “SHEsisPlus, a toolset for genetic studies on polyploid  
489 species”. In: *Scientific Reports* 6 (2016), pp. 1–10. ISSN: 20452322. DOI: 10.  
490 1038/srep24095. URL: <http://dx.doi.org/10.1038/srep24095>.
- 491 [29] John R Thompson, John Attia, and Cosetta Minelli. “The meta-analysis of  
492 genome-wide association studies”. In: *Briefings in Bioinformatics* 12.3 (2011),  
493 pp. 259–269. ISSN: 1467-5463. DOI: 10.1093/bib/bbr020. URL: <https://doi.org/10.1093/bib/bbr020>.
- 495 [30] Feng Tian et al. “Genome-wide association study of leaf architecture in the  
496 maize nested association mapping population”. In: *Nature genetics* 43.2 (2011),  
497 pp. 159–162.
- 498 [31] Yan Y. Yan et al. “Effects of input data quantity on genome-wide association  
499 studies (GWAS)”. In: *International Journal of Data Mining and Bioinformatics*  
500 22.1 (2019), pp. 19–43. ISSN: 17485681. DOI: 10.1504/IJDMB.2019.  
501 099286.
- 502 [32] J Yu et al. “A unified mixed-model method for association mapping that ac-  
503 counts for multiple levels of relatedness.” In: *Nature genetics* 38.2 (2006),  
504 pp. 203–208.
- 505 [33] Jiazheng Yuan et al. “Genome-Wide Association Study of Resistance to Potato  
506 Common Scab”. In: *Potato Research* (2019). ISSN: 18714528. DOI: 10.1007/  
507 s11540-019-09437-w.
- 508 [34] Shengkui Zhang et al. “Genome-wide association studies of 11 agronomic  
509 traits in cassava (*Manihot esculenta* crantz)”. In: *Frontiers in Plant Science*  
510 9.April (2018), pp. 1–15. ISSN: 1664462X. DOI: 10.3389/fpls.2018.  
511 00503.