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

L. Garreta[1], I. Cerón-Souza[1], M.R. Palacio[2], and P.H. Reyes-Herrera[1]

[1]Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), CI Tibaitatá, Kilómetro 14, Vía a Mosquera, 250047, Colombia

[2]Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), CI El Mira, Kilómetro 38, Vía Tumaco Pasto, Colombia

May 29, 2020

Abstract

Summary: The Genome-Wide Association Studies (GWAS) are essential to determine the association between genetic variants across individuals. One way to support the results is by using different tools to validate the reproducibility of the associations. Currently, software for GWAS in diploids is well-established but for polyploids species 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.

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

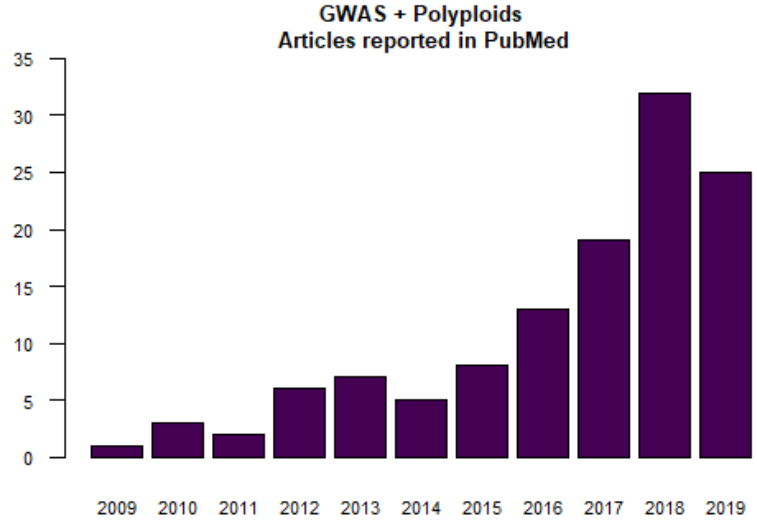


Figure 1: Timeline for articles reported for GWAS studies on polyploid species in PubMed. We present data for completed years.

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. 2). 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.

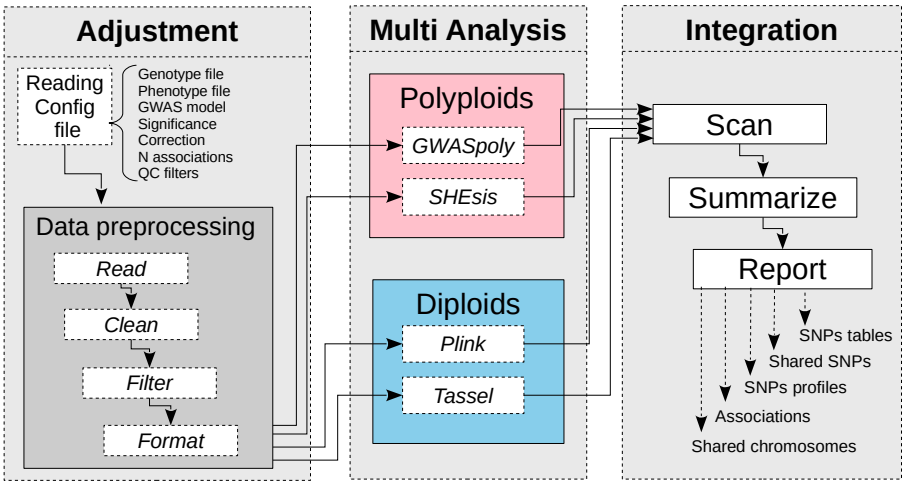


Figure 2: MultiGWAS flowchart has three consecutive steps: adjustment, multi analysis, and integration. The adjustment step manages the input data, reads the configuration file, and preprocessing the input genomic data (genotype and phenotype). The multi analysis step configures and runs the four GWAS 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.

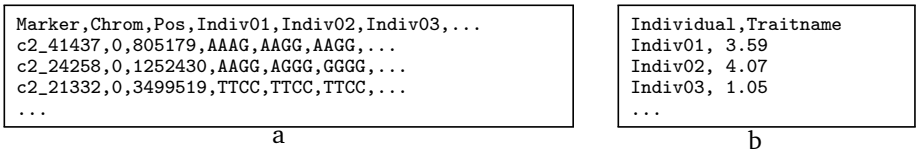


Figure 3: MultiGWAS genotype and phenotype formats. Both files are in CSV format (Comma Separated Values) and contain as first row the header labels of the columns. Although the header labels are arbitrary, the column order is obligatory. **a.** Genotype file format, where "Marker", "Chrom", and "Pos", 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:

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

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

Genome-wide significance: GWAS searches SNPs associated with a phenotype trait in a statistically significant manner. A threshold or significance level α 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 α 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

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

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

Quality control filters: A control step is necessary to check the input data for genotype or phenotype errors or poor quality that can lead to spurious GWAS results. MultiGWAS provides the option to select and define thresholds for the following filters that control the data quality: Minor Allele Frequency (MAF), individual missing rate (MIND), SNP missing rate (GENO), and HardyWeinberg threshold (HWE):

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

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

2.1.2 Data preprocessing

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

During this step, 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: 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, the genomic data, genotype and phenotype, are converted to the specific formats required for each of the four GWAS packages.

2.2 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) [27]. 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 five PC as covariates. For relatedness, MultiGWAS uses kinship matrices that TASSEL and GWASpoly calculated separately, and for PLINK and SHEsis, relatedness depends of kinship coefficients calculated with the PLINK 2.0 built-in algorithm [8].

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

2.2.1 GWASpoly

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

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

2.2.2 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 [24, 21].

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).

2.2.3 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 [22]. 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.

2.2.4 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, 34], 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. And, in the same manner that GWASPoly, TASSEL provides three gene action models to calculate the SNP effect of each genotypc class: general, additive, and dominant, and so the significance threshold depends of each action model.

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

2.3 Integration stage.

The outputs resulting from the four GWAS packages are scanned and processed to identify both significant and best-ranked associations with *p-values* lower than and close to a significance threshold, respectively.

2.3.1 Calculation of *p-values* and significance thresholds

GWAS packages compute *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* drops below a predefined significance threshold. But, most GWAS packages compute differently *p-values* with the possibility to compute them too high or too low. If *p-values* are too high, then it would lead to false negatives or SNPs with true associations with the phenotype but that does not reach the significance threshold. Conversely, if *p-values* are too

low, then it would lead to false positives or SNPs with false associations with the phenotype but that reaches the significance threshold.

To overcome these difficulties, in the case of too high *p-values*, MultiGWAS identifies and reports both significant and best-ranked associations (the ones closer to being statistically significant). Whereas, in the case of too low *p-values*, MultiGWAS provides two methods for adjusting *p-values* and significance threshold: the false discovery rate (FDR) that adjust *p-values,* and the Bonferroni correction, that adjusts the threshold.

By default, MultiGWAS uses the Bonferroni correction in which the significance threshold is adjusted as α/m, where α is the significance level defined by the user in the configuration file, and m is the number of tested markers in the GWAS study. However, the significance threshold can be different for each GWAS package as some of them use several action models to calculate the SNP effect of each genotypic class. For both PLINK and SHEsis packages, which use only one model, m is equal to the total number of SNPs, but for both GWASpoly and TASSEL packages, which use eight and three gene action models, respectively, m is equal to the number of test performed in each model, which is different between models.

2.3.2 Selection of significant and best-ranked associations

After corrections, significant associations are selected as the ones with *p-values* falling below a significant threshold, which is calculated for each GWAS package. But, as described above, it is equally important to know the best-ranked associations, closer to being statistically significant, as they may represent important associations to consider for posterior analysis.

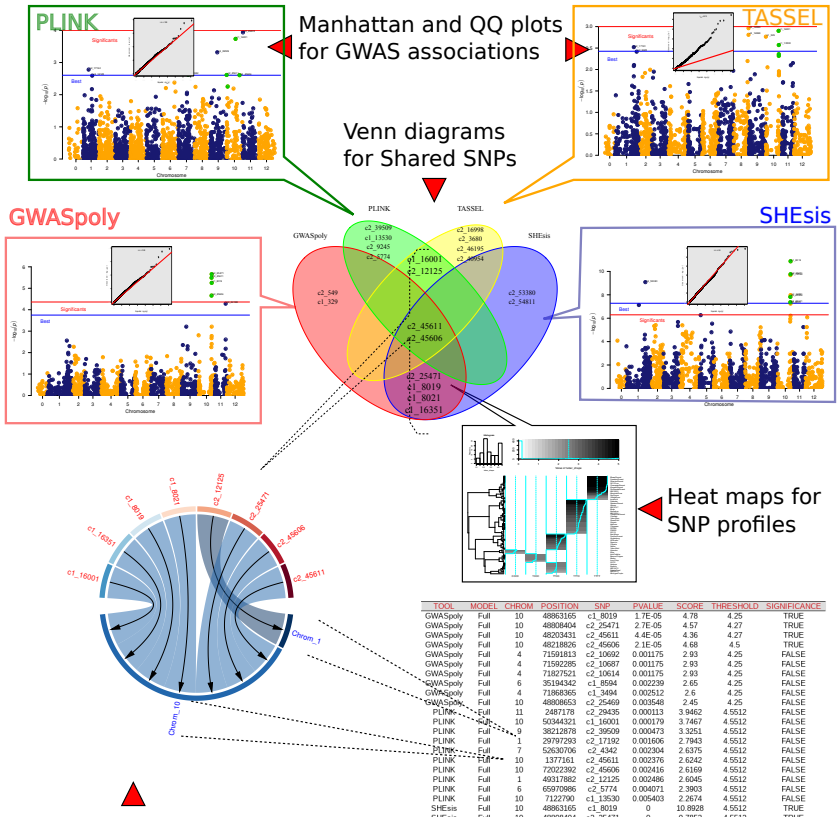
In the case of GWAS packages with only one gene action model (PLINK and SHESIS), the best-ranked associations are selected from the top N identified by the package. But, in the case of GWAS packages with several gene action models (GWASpoly and TASSEL), the best-ranked associations are selected as the top N from the "best action model", the one with more shared SNP associations, in other words, from the action model that identifies more associations that are also identified in the other models.

2.3.3 Integration of results

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

- A Quantile-Quantile (QQ) plots for the resultant *p-values* of each tool and the corresponding inflation factor λ to asses the degree of the test statistic inflation.
- 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 significative SNPs provided by each tool and (2) the set of all the best-ranked SNPs. For each set,



TOOL	MODEL	CHROM	POSITION	SNP	PVALUE	SCORE	THRESHOLD	SIGNIFICANCE
GWASpoly	Full	10	48863165	c1_8019	1.7E-05	4.78	4.25	TRUE
GWASpoly	Full	10	48808404	c2_25471	2.7E-05	4.57	4.27	TRUE
GWASpoly	Full	10	48203431	c2_45611	4.4E-05	4.36	4.27	TRUE
GWASpoly	Full	10	48218826	c2_45606	2.1E-05	4.68	4.5	TRUE
GWASpoly	Full	4	71591813	c2_10692	0.001175	2.93	4.25	FALSE
GWASpoly	Full	4	71592285	c2_10687	0.001175	2.93	4.25	FALSE
GWASpoly	Full	4	71827521	c2_10614	0.001175	2.93	4.25	FALSE
GWASpoly	Full	6	35194342	c1_8594	0.002239	2.65	4.25	FALSE
GWASpoly	Full	4	71868365	c1_3494	0.002512	2.6	4.25	FALSE
GWASpoly	Full	10	48808653	c2_25469	0.003548	2.45	4.25	FALSE
PLINK	Full	11	2487178	c2_29435	0.000113	3.9462	4.5512	FALSE
PLINK	Full	10	50344321	c1_16001	0.000179	3.7467	4.5512	FALSE
PLINK	Full	9	38212878	c2_39509	0.000473	3.3251	4.5512	FALSE
PLINK	Full	1	29797293	c2_17192	0.001606	2.7943	4.5512	FALSE
PLINK	Full	7	52630706	c2_4342	0.002304	2.6375	4.5512	FALSE
PLINK	Full	10	1377161	c2_45611	0.002376	2.6242	4.5512	FALSE
PLINK	Full	10	72022392	c2_45606	0.002416	2.6169	4.5512	FALSE
PLINK	Full	1	49317882	c2_12125	0.002486	2.6045	4.5512	FALSE
PLINK	Full	6	65970986	c2_5774	0.004071	2.3903	4.5512	FALSE
PLINK	Full	10	7122790	c1_13530	0.005403	2.2674	4.5512	FALSE
SHEsis	Full	10	48863165	c1_8019	0	10.8928	4.5512	TRUE
SHEsis	Full	10	48808404	c2_25471	0	9.7852	4.5512	TRUE
SHEsis	Full	10	48020996	c1_13526	0	9.7122	4.5512	TRUE
SHEsis	Full	1	70371898	c2_53380	0	9.1135	4.5512	TRUE
SHEsis	Full	10	48761642	c1_16351	0	8.0061	4.5512	TRUE
SHEsis	Full	10	48218826	c2_45606	0	7.8539	4.5512	TRUE
SHEsis	Full	10	48203431	c2_45611	0	7.3851	4.5512	TRUE
SHEsis	Full	10	48862950	c1_8021	0	7.279	4.5512	TRUE
SHEsis	Full	1	46270954	c2_54811	0	7.1593	4.5512	TRUE
SHEsis	Full	5	3515956	c2_11977	1E-06	6.2899	4.5512	TRUE
TASSEL	Full	8	54838024	c2_16999	0.00107	2.9707	4.5512	FALSE
TASSEL	Full	10	47539878	c1_16001	0.00123	2.9101	4.5512	FALSE
TASSEL	Full	7	14924207	c2_4342	0.00132	2.8793	4.5512	FALSE
TASSEL	Full	8	54838005	c2_16998	0.001522	2.8176	4.5512	FALSE
TASSEL	Full	10	565197	c1_329	0.001679	2.7749	4.5512	FALSE
TASSEL	Full	10	48148399	c1_13530	0.002639	2.5786	4.5512	FALSE
TASSEL	Full	1	70474651	c2_17191	0.002955	2.5294	4.5512	FALSE
TASSEL	Full	1	81410256	c2_9925	0.003756	2.4253	4.5512	FALSE
TASSEL	Full	10	48218826	c2_45606	0.00433	2.3635	4.5512	FALSE
TASSEL	Full	10	48203431	c2_45611	0.004718	2.3262	4.5512	FALSE

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 significative SNPs. We present two Venn diagrams, one for the significative 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.

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, AABB, 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

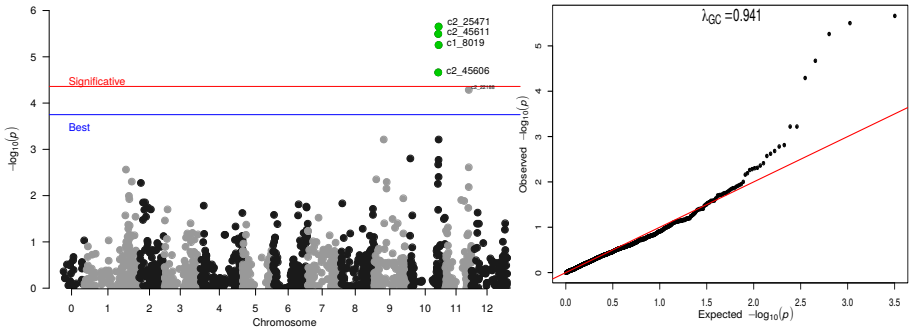
Although most of the GWAS packages used by MultiGWAS use linear regression approaches, they often produce different association results for the same input. For example, computed *p-values* for the same set of SNPs are different between packages; SNPs with significant *p-values* for one package maybe 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 intersections 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 significative, best-ranked, and shared SNPs (found by more than one tool).



(a) (b)

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 assesed 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.

PKG	MDL	CHR	POS	SNP	SCR	THR	SGN
GWASpoly	Full	10	48863165	c1_8019	4.78	4.25	TRUE
GWASpoly	Full	10	48808404	c2_25471	4.57	4.27	TRUE
GWASpoly	Full	10	48203431	c2_45611	4.36	4.27	TRUE
GWASpoly	Full	10	48218826	c2_45606	4.68	4.5	TRUE
GWASpoly	Full	4	71591813	c2_10692	2.93	4.25	FALSE
GWASpoly	Full	4	71592285	c2_10687	2.93	4.25	FALSE
GWASpoly	Full	4	71827521	c2_10614	2.93	4.25	FALSE
PLINK	Full	10	67293176	c1_16001	1.7693	3.2601	FALSE
PLINK	Full	10	77351069	c1_329	1.1795	3.301	FALSE
PLINK	Full	11	51404231	c2_29435	1.1188	3.2553	FALSE
PLINK	Full	10	69323144	c2_45611	1.0229	3.2553	FALSE
PLINK	Full	2	41814861	c2_16350	0.9598	3.301	FALSE
PLINK	Full	10	69311500	c2_45606	0.8489	3.2923	FALSE
PLINK	Full	10	69809843	c1_16351	0.6131	3.2833	FALSE
SHEsis	Full	2	13697423	c1_8019	9.4711	3.301	TRUE
SHEsis	Full	1	30837971	c1_13526	8.4501	3.2923	TRUE
SHEsis	Full	5	46046095	c2_53380	8.2409	3.2601	TRUE
SHEsis	Full	3	39255236	c2_25471	7.8241	3.2923	TRUE
SHEsis	Full	5	49804489	c2_54811	6.9633	3.2695	TRUE
SHEsis	Full	1	69809843	c1_16351	6.0247	3.2833	TRUE
SHEsis	Full	4	69311500	c2_45606	5.9557	3.2923	TRUE
TASSEL	Full	8	54838024	c2_16999	3.6076	3.8943	FALSE
TASSEL	Full	8	54838005	c2_16998	3.483	3.8943	FALSE
TASSEL	Full	1	71450400	c2_12125	2.4832	3.8943	FALSE
TASSEL	Full	1	70474651	c2_17191	2.45	3.8943	FALSE
TASSEL	Full	1	70472380	c2_17193	2.2893	3.8943	FALSE
TASSEL	Full	10	47539878	c1_16001	2.9101	4.5512	FALSE
TASSEL	Full	7	14924207	c2_4342	2.8793	4.5512	FALSE

(a)

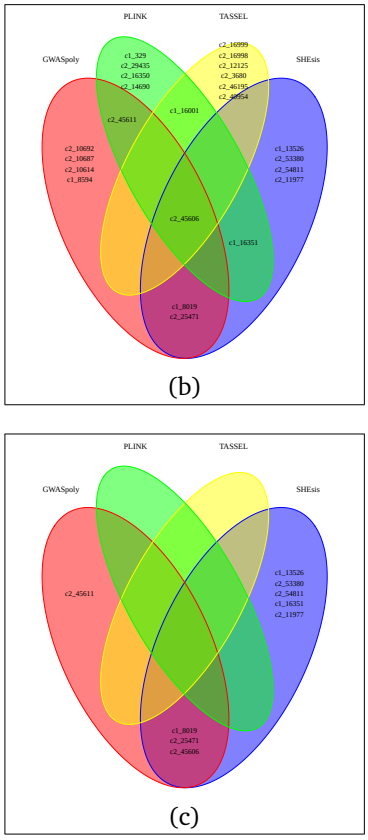
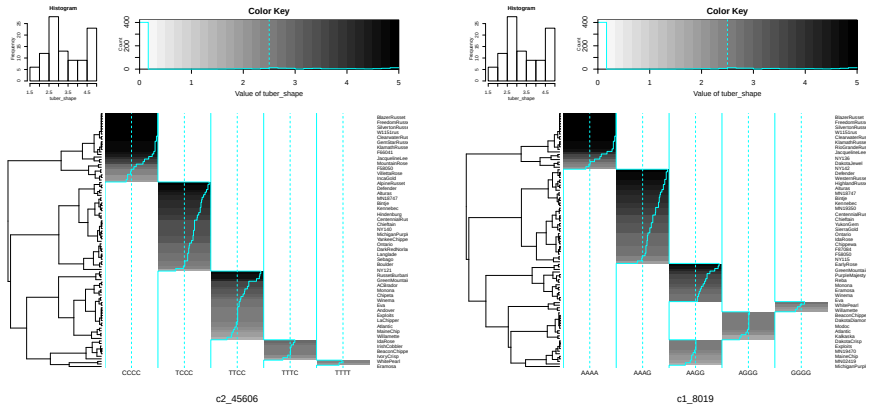


Figure 6: Visualization of shared SNPs. Tabular and graphical views of the best-ranked and significant SNPs identified by the four packages. **(a)** Tabular view with detailed information of each SNPs, including: package name (PKG), GWAS model used (MDL), chromosome (CHR), position in the genome (POS), ID (SNP), score (SCR), threshold (THR), and significance flag (SGN), wether the SNP was evaluated statistically significant or not (score > theshold). **(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 representation, what we called SNP profile. This profiles enables to visualize each trait by individuals and genotypes as rows and columns, respectively (Figure 7). At the left, the individuals are grouped in a dendrogram by their genotype. At the right, there is the name or ID of each individual. At the bottom, the genotypes are ordered from left to right, starting from the major to the minor allele (i.e., AAAA,

AAAB, AABB, ABBB, BBBB). At the top, there is a description of the trait based on a histogram of frequency (top left) and by an assigned color for each numerical phenotype value using a grayscale (top right). Thus, each individual appears as a colored line by its phenotype value on its genotype column. For each column, there is a solid cyan line with the mean of each column and a broken cyan line that indicates how far the cell deviates from the mean.

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



(a) (b)

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).

4 Availability and implementation:

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

4.1 Input parameters

MutiGWAS uses as the only input a simple configuration text file where users set the values for the main parameters that drives the GWAS process. The input parameters include: the output folder where results will be written, input genotype/phenotype filenames, genome-wide significance threshold, method for multiple testing correction, GWAS model, number of associations to be reported, and TRUE or FALSE whether to use quality control filters or not. The filters are: minor allele frequency, individual missing rate, SNP missing rate, and Hardy-Weinberg threshold.

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

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

4.2 Using the command line interface

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

```
multiGWAS full.config
```

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

4.3 Using the graphical user interface

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

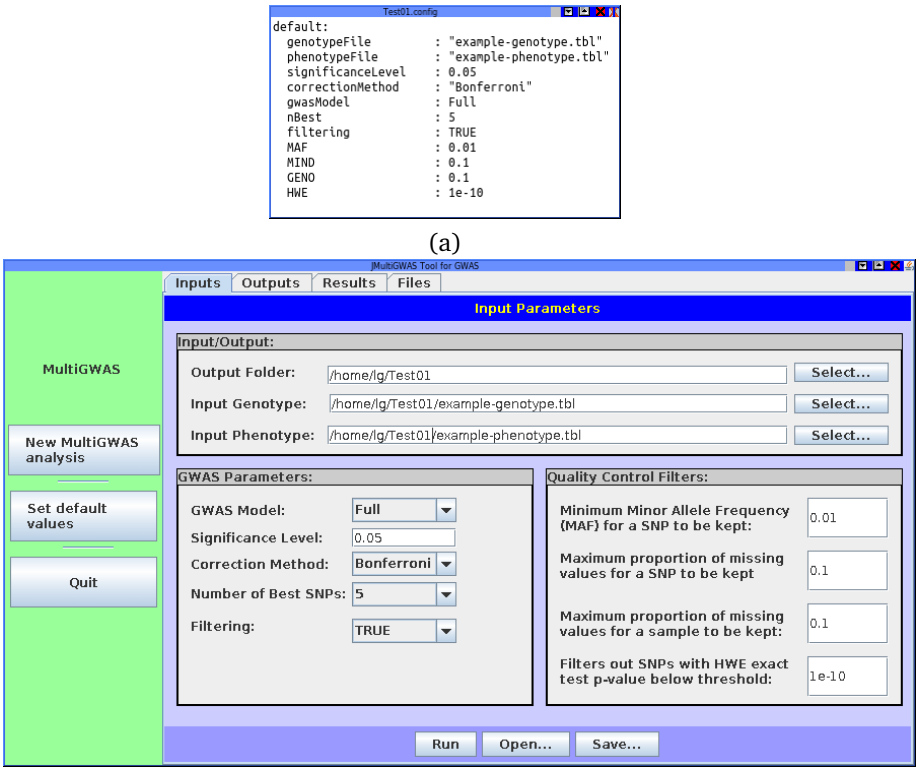


Figure 8: MultiGWAS inputs and interaction. MultiGWAS uses as input a simple configuration text file and can be executed using either a command line interface script in R (CLI) or a graphic user interface application in Java (GUI). **(a)** An example of a configuration text file named *"Test01.config"* 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

XXXXXXXXXXXXXXXXXXXXXXX

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%7B%5C_%7D5.

[10] Robert Ekblom 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.

[13] Dominik G Grimm et al. “easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies”. In: *The Plant Cell* 29.1 (2017), pp. 5–19. ISSN: 1040-4651. DOI: 10.1105/tpc.16.00551. URL: http://www.plantcell.org/content/29/1/5.

[14] Anja C Gumpinger et al. *Methods and Tools in Genome-wide Association Studies*. Vol. 1819. 2018. ISBN: 9781493986187.

[15] Bin Han and Xuehui Huang. “Sequencing-based genome-wide association study in rice”. In: *Current opinion in plant biology* 16.2 (2013), pp. 133–138.

[16] Candice N. Hirsch et al. “Retrospective view of North American potato (Solanum tuberosum L.) breeding in the 20th and 21st centuries”. In: *G3: Genes, Genomes, Genetics* 3.6 (2013), pp. 1003–1013. ISSN: 21601836. DOI: 10.1534/g3.113.005595.

[17] “How to interpret a genome-wide association study”. In: *JAMA - Journal of the American Medical Association* 299.11 (2008), pp. 1335–1344. ISSN: 00987484. DOI: 10.1001/jama.299.11.1335.

[18] Avjinder S Kaler and Larry C Purcell. “Estimation of a significance threshold for genome-wide association studies”. In: *BMC Genomics* 20.1 (2019), p. 618. ISSN: 1471-2164. DOI: 10.1186/s12864-019-5992-7. URL: https://doi.org/10.1186/s12864-019-5992-7.

[19] Arthur Korte and Ashley Farlow. “The advantages and limitations of trait analysis with GWAS: a review”. In: *Plant methods* 9.1 (2013), p. 29.

[20] Gordan Lauc et al. “Genomics meets glycomics—the first GWAS study of human N-glycome identifies HNF1α as a master regulator of plasma protein fucosylation”. In: *PLoS genetics* 6.12 (2010).

[21] Jie Meng et al. “Genome-wide association analysis of nutrient traits in the oyster Crassostrea gigas: Genetic effect and interaction network”. In: *BMC Genomics* 20.1 (2019), pp. 1–14. ISSN: 14712164. DOI: 10.1186/s12864-019-5971-z.

[22] Robert A. Power, Julian Parkhill, and Tulio De Oliveira. “Microbial genome-wide association studies: lessons from human GWAS”. In: *Nature Reviews Genetics* 18.1 (2016), pp. 41–50. ISSN: 14710064. DOI: 10.1038/nrg.2016.132.

[23] Shaun Purcell et al. “PLINK: A tool set for whole-genome association and population-based linkage analyses”. In: *American Journal of Human Genetics* 81.3 (2007), pp. 559–575. ISSN: 00029297. DOI: 10.1086/519795.

[24] Hui Ping Qiao et al. “Genetic variants identified by GWAS was associated with colorectal cancer in the Han Chinese population”. In: *Journal of Cancer Research and Therapeutics* 11.2 (2015), pp. 468–470. ISSN: 19984138. DOI: 10.4103/0973-1482.150346.

[25] Umesh R. Rosyara et al. “Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato”. In: *The Plant Genome* 9.2 (2016), pp. 1–10. ISSN: 1940-3372. DOI: 10.3835/plantgenome2015.08.0073. URL: https://dl.sciencesocieties.org/publications/tpg/abstracts/9/2/plantgenome2015.08.0073.

[26] Anna W Santure and Dany Garant. “Wild GWAS—association mapping in natural populations”. In: *Molecular ecology resources* 18.4 (2018), pp. 729–738.

[27] Sanjeev Kumar Sharma et al. “Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato”. In: *G3: Genes, Genomes, Genetics* 8.10 (2018), pp. 3185–3202. ISSN: 21601836. DOI: 10.1534/g3.118.200377.

[28] Jiawei Shen et al. “SHEsisPlus, a toolset for genetic studies on polyploid species”. In: *Scientific Reports* 6 (2016), pp. 1–10. ISSN: 20452322. DOI: 10.1038/srep24095. URL: http://dx.doi.org/10.1038/srep24095.

[29] John R Thompson, John Attia, and Cosetta Minelli. “The meta-analysis of genome-wide association studies”. In: *Briefings in Bioinformatics* 12.3 (2011), pp. 259–269. ISSN: 1467-5463. DOI: 10.1093/bib/bbr020. URL: https://doi.org/10.1093/bib/bbr020.

[30] Feng Tian et al. “Genome-wide association study of leaf architecture in the maize nested association mapping population”. In: *Nature genetics* 43.2 (2011), pp. 159–162.

[31] Yan Y. Yan et al. “Effects of input data quantity on genome-wide association studies (GWAS)”. In: *International Journal of Data Mining and Bioinformatics* 22.1 (2019), pp. 19–43. ISSN: 17485681. DOI: 10.1504/IJDMB.2019.099286.

[32] J Yu et al. “A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.” In: *Nature genetics* 38.2 (2006), pp. 203–208.

[33] Jiazheng Yuan et al. “Genome-Wide Association Study of Resistance to Potato Common Scab”. In: *Potato Research* (2019). ISSN: 18714528. DOI: 10.1007/s11540-019-09437-w.

[34] Shengkui Zhang et al. “Genome-wide association studies of 11 agronomic traits in cassava (Manihot esculenta crantz)”. In: *Frontiers in Plant Science* 9.April (2018), pp. 1–15. ISSN: 1664462X. DOI: 10.3389/fpls.2018.00503.