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

April 13, 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.

Availability and implementation: The tool is in R. Source code, examples, documentation and installation instructions are available at https://github.com/agrosavia/multiGWAS-min

Contact: phreyes@agrosavia.co

Keywords: GWAS, tetraploids, SNPs,XXX

1 Introduction

The GWAS (Genome-Wide Association Study) is used to identify which variants through the whole genome of a large number of individuals are associated with a specific trait (CITES). This methodology started with humans and several model plants, such as rice, maize, and Arabidopsis (CITES). Because of the advances in the next-gen sequencing technology and the decreasing of the sequencing cost in recent years, there is an increase in genome sequences in non-model organisms at a faster rate (CITES). Therefore, several research projects are intended for the first time a GWAS analysis for non-model wild plants and crops that often are polyploids (CITES).

One of the main challenges in the GWAS analysis is to identify real associations. A reliable method to validate the results is by replicating the study using different software. This replication process is a challenge if our study organism is polyploid. Currently, the GWAS software to analyze polyploids is limited. Therefore, to confirm the GWAS of polyploids species, it is often necessary to "diploidizing" the data in order to use software designed exclusively for a diploid data matrix. Thus, the replication process is time-consuming. Each software has its characteristics, such as different user interfaces (GUI or command-line based), genotype-phenotype formats, models, algorithm assumptions, and outputs.

To solve this problem, we developed the MultiGWAS tool that performs GWAS analyses for tetraploid species using four software in parallel. Our tool include GWASpoly [11] and the SHEsis tool [13] that accept polyploid genomic data, and PLINK [9] and TASSEL [3] with the use of a "diploidized" genomic matrix. The tool deals with preprocessing data, running four GWAS tools in parallel, and create reports to help the user decide more intuitively the possible true or false associations.

2 Material and Methods

2.1 Tools

We have selected four GWAS software tools to be integrated in our multiGWAS tool, two designed specifically for polyploid species as many important crops are polyploids: GWASpoly [11] and SHEsis [14], and another two designed for diploids species and extensively used in humans and plants: PLINK [9, 4] and TASSEL [3], respectively.

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

2.2 GWASpoly

GWASpoly is a recent R package designed for GWAS in polyploid species that has been used in several studies in plants [2, 5, 12, 15]. It is based on the Q+K linear mixed model with biallelic SNPs that accounts for population structure and relatedness. In addition, to calculate the SNP effect for each genotypic class, GWASpoly provides a general gene action model along with four additional models: additive, simplex dominant, and duplex dominant.

MultiGWAS is using GWASpoly version 1.3. The population structure and relatedness, used in the full model, are estimated using the first five principal components and the kinship matrix, respectively, both calculated with the algorithms built in GWASpoly. For both, naive and full models, all gene action models are tested for detecting associations.

2.3 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 [10, 7].

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).

2.4 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 [8]. 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.

2.5 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, 16], 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.

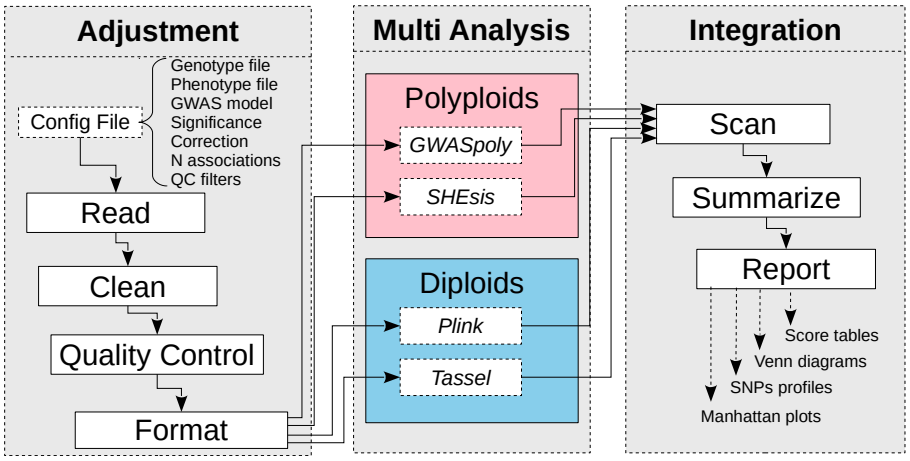


Figure 1: MultiGWAS flowchart has three stages: adjustment, multi analysis, and integration. In the first stage, we process the configuration file. It includes the genotype/phenotype filenames, genome-wide significance threshold, multiple testing correction methods, GWAS model, number of associations to be reported, and TRUE or FALSE whether to use quality control (QC) filters or not. Then the genotype and phenotype are cleaned and filtered using the QC filters. In the second stage, each GWAS tool runs in parallel. In the last stage, after the output files scanning, a summary of results is generated in a report that contains score tables, Venn diagrams, SNP profiles, and Manhattan plots. The QC filters are minor allele frequency, individual missing rate, SNP missing rate, HardyWeinberg threshold.

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

2.6 Method

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

2.6.1 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. It starts by preprocessing the genomic data by selecting individuals present in both genotype and phenotype and excluding individuals and SNPs that have

poor quality. Moreover, the format "ACGT" suitable for the polyploid software GWASpoly and SHEsis, is "diploidized" for PLINK and TASSEL. The homozygous tetraploid genotypes are converted to diploid thus: (e.g.,AAAA→AA, CCCC→CC, GGGG→GG, TTTT→TT). Moreover, for tetraploid heterozygous genotypes, the conversion depends on the reference and alternate alleles calculated for each position (e.g. AAAT→AT, ... ,CCCG→CG). After this process, MultiGWAS transform the genomic data to the formats required for each software.

2.6.2 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) [12]. The Full model (Q+K) controls for both population structure and individual relatedness. For population structure, MultiGWAS uses the Principal Component Analysis (PCA) and takes the top ten PC as covariates. For relatedness, the tool uses kinship matrices that TASSEL and GWASpoly calculated separately, and for PLINK and SHEsis depends on the King software [6].

2.6.3 Integration stage.

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

3 Results

We tested the MultiGWAS tool with the data for the Solanaceae Coordinated Agricultural Project (SolCAP) potato diversity panel implemented in the GWASpoly software [11] and the tuber shape trait. We present the Venn diagram (Fig. 1B) that summarizes the full model results. The two polyploid software, GWASpoly, and SHEsis identify two SNPs: one is the c1_8019, also the most significant association from the original study[11]. Therefore, it could be considered a real association. Also, other five SNPs are predicted simulatenously by at least two tools. For each significant association, MultiGWAS generates a heat map figure to summarize the genotype associated with a trait for each individual (Fig. 1C). The complete report from MultiGWAS for the naive and full model is in the Supplementary information.

TOOL	MODEL	CHR	POS	SNP	P	SCORE	THRESHOLD	SIGNF
GWASpoly	Full	10	48863165	c1_8019	0.000017	4.780000	4.250000	TRUE
GWASpoly	Full	10	48808404	c2_25471	0.000027	4.570000	4.270000	TRUE
GWASpoly	Full	10	48203431	c2_45611	0.000044	4.360000	4.270000	TRUE
GWASpoly	Full	10	48218826	c2_45606	0.000021	4.680000	4.500000	TRUE
PLINK	Full	10	67293176	c1_16001	0.000187	1.769349	3.260071	FALSE
PLINK	Full	10	77351069	c1_329	0.000662	1.179470	3.301030	FALSE
PLINK	Full	11	51404231	c2_29435	0.000845	1.118849	3.255273	FALSE
PLINK	Full	10	69323144	c2_45611	0.001054	1.022917	3.255273	FALSE
PLINK	Full	2	41814861	c2_16350	0.001097	0.959793	3.301030	FALSE
PLINK	Full	10	69311500	c2_45606	0.001445	0.848906	3.292256	FALSE
PLINK	Full	10	69809843	c1_16351	0.002539	0.613066	3.283301	FALSE
SHEsis	Full	2	13697423	c1_8019	0.000000	9.471083	3.301030	TRUE
SHEsis	Full	1	30837971	c1_13526	0.000000	8.450065	3.292256	TRUE
SHEsis	Full	5	46046095	c2_53380	0.000000	8.240929	3.260071	TRUE
SHEsis	Full	3	39255236	c2_25471	0.000000	7.824082	3.292256	TRUE
SHEsis	Full	5	49804489	c2_54811	0.000000	6.963331	3.269513	TRUE
SHEsis	Full	1	69809843	c1_16351	0.000000	6.024734	3.283301	TRUE
SHEsis	Full	4	69311500	c2_45606	0.000000	5.955695	3.292256	TRUE
TASSEL	Full	8	54838024	c2_16999	0.000247	3.607621	3.894316	FALSE
TASSEL	Full	8	54838005	c2_16998	0.000329	3.482989	3.894316	FALSE
TASSEL	Full	1	71450400	c2_12125	0.003287	2.483226	3.894316	FALSE
TASSEL	Full	1	70474651	c2_17191	0.003548	2.449995	3.894316	FALSE
TASSEL	Full	1	70472380	c2_17193	0.005137	2.289293	3.894316	FALSE
TASSEL	Full	10	47539878	c1_16001	0.001230	2.910131	4.551206	FALSE
TASSEL	Full	7	14924207	c2_4342	0.001320	2.879298	4.551206	FALSE

Table 1: Table for Full model results (n=7)....

TOOL	MODEL	CHROM	POSITION	SNP	PVALUE	SCORE	THRESHOLD	SIGNIFICANCE
GWASpoly	Naive	10	48863165	c1_8019	0	11.560000	4.500000	TRUE
GWASpoly	Naive	10	48020996	c1_13526	0	10.610000	4.500000	TRUE
GWASpoly	Naive	1	70371898	c2_53380	0	10.370000	4.500000	TRUE
GWASpoly	Naive	10	48808404	c2_25471	0	9.890000	4.500000	TRUE
GWASpoly	Naive	1	46270954	c2_54811	0	8.540000	4.500000	TRUE
GWASpoly	Naive	10	48218826	c2_45606	0	7.970000	4.500000	TRUE
GWASpoly	Naive	10	48761642	c1_16351	0	7.970000	4.500000	TRUE
PLINK	Naive	5	32820618	c2_11977	0	6.848054	3.283301	TRUE
PLINK	Naive	10	13697423	c1_8019	0	5.040577	3.301030	TRUE
PLINK	Naive	1	46046095	c2_53380	0	4.967240	3.260071	TRUE
PLINK	Naive	2	72026885	c2_47760	0	4.771482	3.292256	TRUE
PLINK	Naive	1	49804489	c2_54811	0	4.691867	3.269513	TRUE
PLINK	Naive	10	30837971	c1_13526	0	4.560996	3.292256	TRUE
PLINK	Naive	10	39255236	c2_25471	0	4.536164	3.292256	TRUE
SHEsis	Naive	5	46046095	c2_53380	0	9.497204	3.313867	TRUE
SHEsis	Naive	2	13697423	c1_8019	0	8.031405	3.357935	TRUE
SHEsis	Naive	1	30837971	c1_13526	0	7.933190	3.330414	TRUE
SHEsis	Naive	2	32820618	c2_11977	0	6.565176	3.334454	TRUE
SHEsis	Naive	3	39255236	c2_25471	0	6.264880	3.342423	TRUE
SHEsis	Naive	4	72026885	c2_47760	0	5.941574	3.342423	TRUE
SHEsis	Naive	3	46475259	c2_28012	0	5.781570	3.318063	TRUE
TASSEL	Naive	5	3515956	c2_11977	0	8.830326	4.549984	TRUE
TASSEL	Naive	10	48863165	c1_8019	0	7.040577	4.549984	TRUE
TASSEL	Naive	1	70371898	c2_53380	0	6.926245	4.549984	TRUE
TASSEL	Naive	1	46270954	c2_54811	0	6.859964	4.549984	TRUE
TASSEL	Naive	2	20034673	c2_47760	0	6.762783	4.549984	TRUE
TASSEL	Naive	10	48020996	c1_13526	0	6.552175	4.549984	TRUE
TASSEL	Naive	10	48808404	c2_25471	0	6.527434	4.549984	TRUE

Table 2: Table for Naive model results (n=7) ....

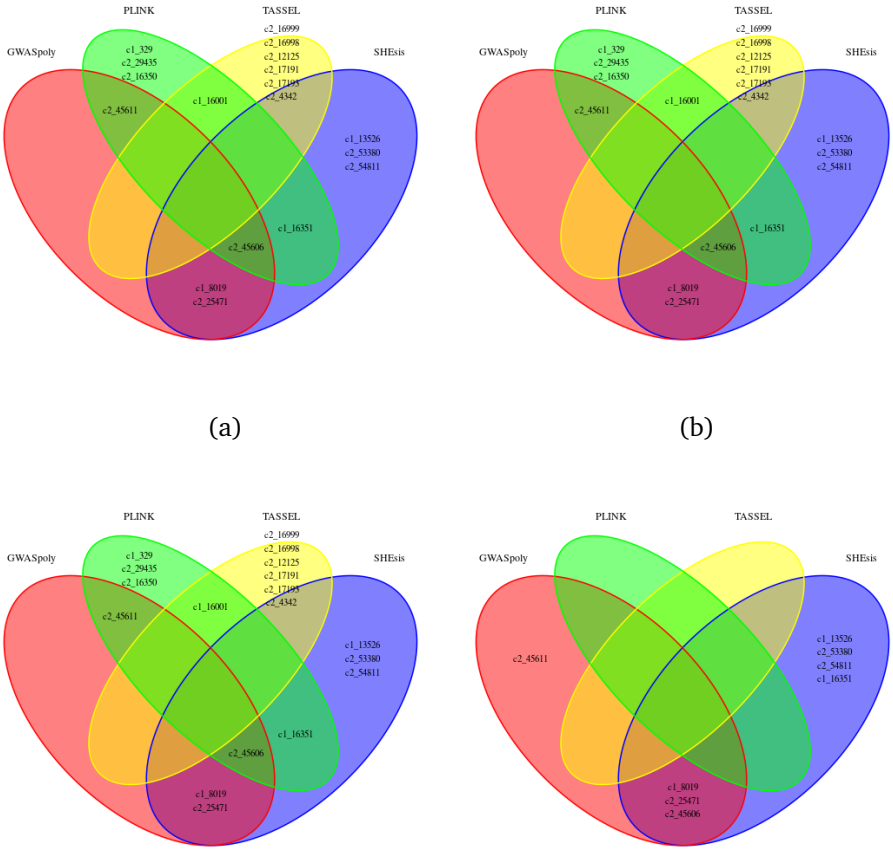


Figure 2: Venn diagrams generated by the MultiGWAS tool for the SolCAP potato panel GWAS for tuber shape. The results are for the Full model in which both diploid software found one marker in common (red text), but the other two polyploid tools found a different marker in common (yellow text). (a)Full model: n Best Ranked, (b) Full model: Significative, (c) Naive model: n Best Ranked, (d)Naive model: Significative

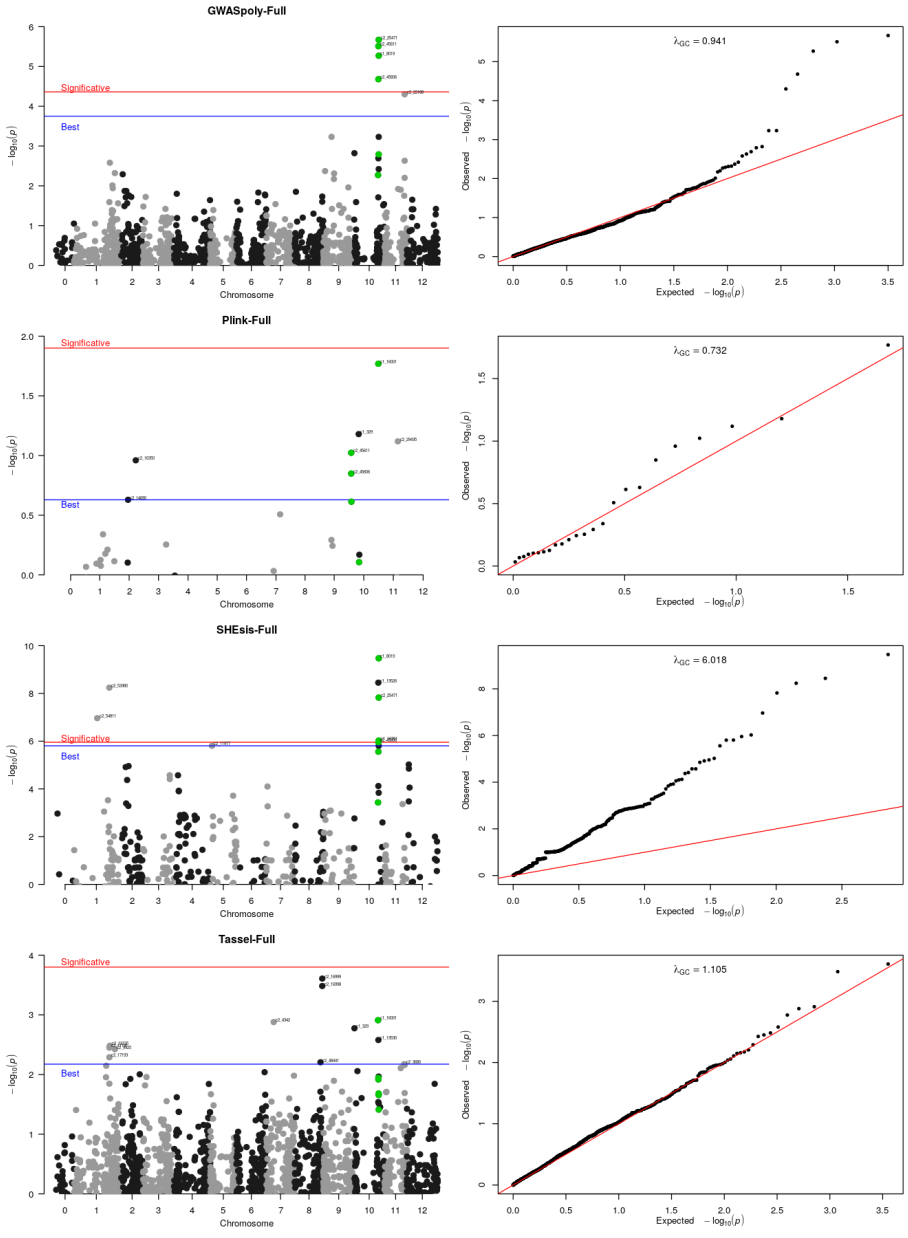
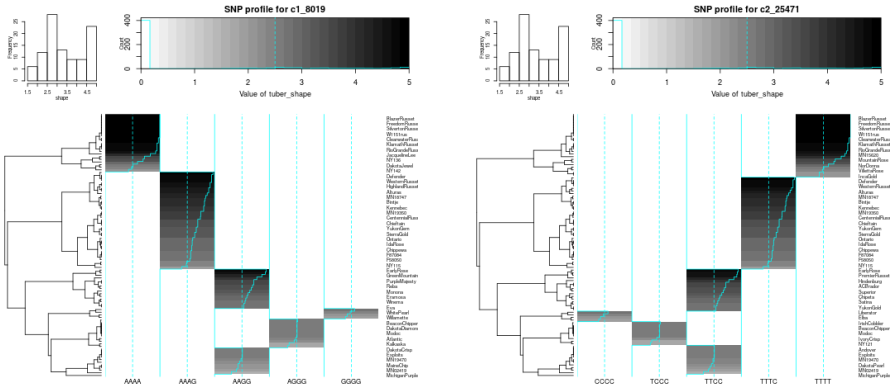
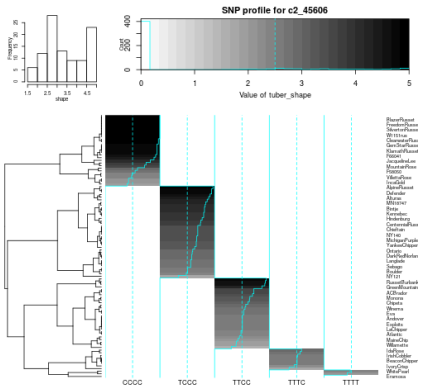


Figure 3: Full model report section Manhattan and QQ plot

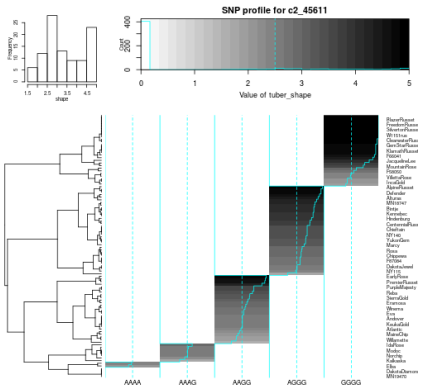


(a) c1_8019

(b) c2_25471



(c) c2_45606



(a) c2_45611

Figure 4: Full model section SNP profile is a heatmap for best-ranked SNPs (in this case, c1_8019). The clusters group the samples(rows) based only on the genotype. The color represents the phenotype value for the trait of interest. On top of the figure, there is the frequency histogram and the color representation for the numeric phenotype.

4 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Ani Manichaikul et al. “Robust relationship inference in genome-wide association studies”. In: *Bioinformatics* 26.22 (2010), pp. 2867–2873. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq559.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Yong Yong Shi and Lin He. “SHEsis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci (Cell Research (2005) 15, (97-98) DOI: 10.1038/sj.cr.7290272)”. In: *Cell Research* 16.10 (2006), p. 851. ISSN: 10010602. DOI: 10.1038/sj.cr.7310101.

[15] 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.

[16] 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.