Leveraging machine learning to advance genome-wide association studies

Gabrielle Dagasso

Department of Mathematics and Statistics,
Thompson Rivers University,
Kamloops, British Columbia, Canada
Email: gdagasso@gmail.com

Yan Yan

Department of Computing Science,
Thompson Rivers University,
Kamloops, British Columbia, Canada
Email: yyan@tru.ca

Lipu Wang

Department of Plant Sciences,
University of Saskatchewan,
Saskatoon, Saskatchewan, Canada
Email: lipu.wang@usask.ca

Longhai Li

Department of Mathematics and Statistics,
University of Saskatchewan,
Saskatoon, Saskatchewan, Canada
Email: longhai.li@usask.ca

Randy Kutcher

Department of Plant Sciences,
University of Saskatchewan,
Saskatoon, Saskatchewan, Canada
Email: randy.kutcher@usask.ca

Wentao Zhang

National Research Council of Canada,
Saskatoon, Saskatchewan, Canada
Email: Wentao.Zhang@nrc-cnrc.gc.ca

Lingling Jin*

Department of Computer Science,
University of Saskatchewan,
Saskatoon, Saskatchewan, Canada
Email: lingling.jin@cs.usask.ca
*Corresponding author

Abstract: Genome-Wide Association Studies (GWAS) has demonstrated its power in discovering genetic variations to particular traits related to agronomically important features in crops. The typical output of a GWAS program includes a series of Single Nucleotide Polymorphisms (SNPs) and their significance. Currently, there is no standard way to compare results across different programs or to select the most 'significant' results uniformly and consistently. To obtain a comprehensive and accurate set of SNPs associated with a trait of interest, we present a novel automated pipeline that leverages machine learning for GWAS discoveries. The pipeline first performs population structure analysis, then executes multiple GWAS software and combines their results into a single SNP set. After that, it selects SNPs from the set with high individual and/or joint effects with the Least Absolute Shrinkage and Selection Operator analysis. Finally, the predictivity of the model is assessed using cross-validation.

Keywords: genome-wide association studies; machine learning; population structure analysis; cross-validation; LASSO; fusarium head blight.

Reference to this paper should be made as follows: Dagasso, G., Yan, Y., Wang, L., Li, L., Kutcher, R., Zhang, W. and Jin, L. (2021) 'Leveraging machine learning to advance genome-wide association studies', *Int. J. Data Mining and Bioinformatics*, Vol. 25, Nos. 1/2, pp.17–36.

Biographical notes: Gabrielle Dagasso is an undergraduate student at Thompson Rivers University. She is finishing her Bachelor of Science with a major in Mathematics and a minor in Computing Science in 2021. She has been working as an Undergraduate Research Assistant with Lingling Jin since the 3rd year of her program. She is the recipient of several research scholarships, such as NSERC USRA and TRU Undergraduate Research Experience Award Program.

Yan Yan is an Assistant Professor in the Department of Computing Science of Thompson Rivers University (TRU). She received her PhD degree in Biomedical Engineering from the University of Saskatchewan (UofS). Before joining TRU, she worked as a Postdoctoral Fellow at the Department of Computer Science in the University of Western Ontario and the UofS. Her primary research interests include proteomics with a focus on peptide sequencing, population genomics on large scale genome-wide data analysis and machine learning in computational biology.

Lipu Wang received his PhD degree in the Department of Biology from the University of Saskatchewan. Now, he is a Research Officer at the Crop Development Centre, Department of Plant Sciences, University of Saskatchewan. Previously he was a Research Associate in National Research Council of Canada. He has extensive experience and expertise in Plant Molecular Biology and Plant Pathology. Currently, he focuses on characterising the molecular mechanism of Fusarium head blight (FHB) resistance in wheat by using molecular and genomic tools. He developed an

LC-MS/MS-based mycotoxin test methods for FHB breeding programs and investigates FHB resistance components through genome-wide association study (GWAS).

Longhai Li received his PhD degree in Statistics from the University of Toronto in 2007 and BSc degree in Statistics from the University of Science and Technology of China in 2002. Currently, he is a Full Professor at the University of Saskatchewan. His research activities focus on developing and applying statistical machine learning methods for high-throughput data and complex-structured data.

Randy Kutcher is a Professor of Plant Pathology at the Crop Development Centre, Department of Plant Sciences, University of Saskatchewan. Previously, he was a Research Scientist with Agriculture and Agri-Food Canada in Saskatchewan focusing on Canola Pathology and Integrated Pest Management. He received his BSc and MSc degrees from the University of Manitoba and PhD degree from the University of Saskatchewan. He leads the Cereal and Flax Pathology program; his wheat research is focused on applied strategies to mitigate stripe rust, fusarium head blight and leaf spot diseases. He teaches an undergraduate course and conducts extension activities with growers and industry on field crop disease management.

Wentao Zhang is a Quantitative Geneticist from National Research Council of Canada (NRC), is curious about key questions: how does genetic variation give rise to phenotypic variation and how can better predict the performance of these complex traits. His research focuses on understanding the genetic architecture underlying complex traits include disease resistance, high yield, better quality and good agronomics package in crops like wheat, Canola and recently pulses by applying: (1) quantitative genetics, genomics, statistical modelling and computational approaches for complex traits (2) Genomics selection, machine learning and deep learning to predict complex traits.

Lingling Jin received her PhD degree in Computer Science specialised in Bioinformatics from the University of Saskatchewan in 2017. Currently, she is an Assistant Professor of Computer Science at the University of Saskatchewan and an Adjunct Professor at Thompson Rivers University. Previously, she was a Postdoctoral Research Scientist with Agriculture and Agri-Food Canada. Her primary research interest is in computational modelling of polyploid genome evolution and various aspects of plant bioinformatics.

This paper is a revised and expanded version of a paper entitled 'Comprehensive GWAS: a pipeline for genome-wide association studies utilising cross-validation to assess the predictivity of genetic variations' presented at the 'IEEE International Conference on Bioinformatics and Biomedicine 2020 (BIBM 2020)', 16–19 December 2020.

1 Introduction

With the development of next generation sequencing technology and drastically decreasing of sequencing cost, vast amount of whole-genome data becomes available nowadays. Genome-Wide Association Studies (GWAS) becomes powerful tools for identifying associations between genetic variants within a species and phenotype differences among individual samples. Typically, the genetic variants are Single

Nucleotide Polymorphisms, also known as SNPs, which are changes of single DNA base-pairs. The genetic data is referred to as genotypes and the phenotypic data is called phenotypes. In most cases, the number of SNPs is much more than the number of samples as the SNPs are across the whole genome. This makes the genotype "small-*n*-large-*p*" data.

GWAS do not rely on previous knowledge to locate potential causal genes regions; instead, they scan the whole genome to identify causal regions. Therefore, they have the ability to investigate low-frequency and rare variants across the whole genome and identify genes that could be missed by other methods. Since last decade, GWAS have served as primary methods to reveal the relationships between genotypes and phenotypes and made significant advancement in health, agriculture and many other fields. In agriculture, they are widely used to dissect driving genes and biological architecture of complex traits.

GWAS usually rely on statistical models to evaluate associations and find significant SNPs that are related to the phenotype. Traditional GWAS perform a series of single-SNP statistical hypothesis tests. These tests are independent for each SNP and the null hypothesis is that there is no association between the SNP and the phenotype of interests. If the null hypothesis is rejected, the tested SNP is reported to be associated with the phenotype and a *p*-value is usually output along with the SNP to indicate the significance. Out of all SNPs, the ones correlated with phenotypes are just a small subset, which means that the genotype data is sparse. This property could bring complex challenges to current GWAS methods as many of reported SNPs may not be truly related to the phenotype. Phenotype can be binary (e.g., 1/0 representing having a particular disease or not) or quantitative (e.g., height of individuals). When dealing with quantitative phenotypes, linear regression is usually applied. The most used models include Generalised Linear Models (GLMs) and Mixed Linear Models (MLMs). MLMs are also known as Linear Mixed Models (LMMs). In this study, we will refer to them as LMMs.

GWAS packages commonly used include PLINK (Purcell et al., 2010), BOLT-LMM (Loh et al., 2015), FaST-LMM (Lippert et al., 2011), GCTA (Yang et al., 2011), TASSEL (Bradbury et al., 2007), GAPIT (Lipka et al., 2012; Tang et al., 2016) and others. All of them are based on traditional statistical models and many of them have the options to choose between LMM, GLM and other statistical models. PLINK is a free, open-source toolset for GWAS and population-based linkage analyses. It provides a range of basic yet rapid computation for large biological data sets. It was first developed for human genomes and is one of the earliest GWAS packages. Nowadays, PLINK is considered as a standard GWAS method in many fields of study. BOLT-LMM utilises an efficient Bayesian mixed-model and it is reported to increase the statistical power and computational speed for expanded data sets. FaST-LMM applies a factorised log-likelihood function in the LMM and claims to scale linearly with data size in both run time and memory. It can be ideal for large data to achieve satisfying results. GCTA is a tool for genome-wide complex trait analysis. By utilising LMM, it fits the contribution of all SNPs as random effects and addresses the "missing heritability" problem of human genomes.

TASSEL and GAPIT were developed specifically for GWAS of agricultural plants. TASSEL was first developed and tested for maize with the purpose to handle a wide range of insertions and deletions. Many other existing software did not consider these types of polymorphisms before. GAPIT was developed after TASSEL and used statistical

methods that were similar as TASSEL. It is capable to perform both GWAS and Genomic Selection (GS) (Goddard and Hayes, 2007). Genomic selection is a marker-assisted selection in which genetic markers across the whole genome (typically SNPs) are used and breeding values are predicted. GAPIT has been updated frequently to incorporate the state-of-the-art methods. It is therefore reported to achieve the most accurate and computationally efficient results.

2 Motivations

GWAS typically output a series of SNPs along with *p*-values, which are used to evaluate the significance of SNPs associated with phenotypes of interest. A pre-defined threshold is set to select significant SNPs for further analysis, e.g., 10^{-7} is a commonly used cut-off in the literature. However, output SNPs from these programs could vary largely and their *p*-values may not be comparable to each other. For example, one study investigated different GWAS packages (Yan et al., 2018) reported that for the same input plant data, *p*-values of the same SNPs output by PLINK were dramatically smaller than the ones output by TASSEL and GAPIT. Their differences were by orders of magnitude. Moreover, none of these programs utilises cross-validation which ensures that a model is generalised and measures the accuracy of the models, such as GLM and LMM. Cross validation is an important step in statistical analysis to ensure the results found are validate and not false positives due to over estimations of the model's accuracy (Oetting et al., 2017). A false positive occurs when something assumed to be true is false, likewise a false negative occurs when something assumed to be false is true. For example, when someone tests positive for a disease but do not have it, this would be a false positive. A common fault of some programs when dealing with false positives is to utilise either the Bonferroni or Benjamini-Hochberg methods which may lead to false negatives and loss of true significant SNPs (Waldmann et al., 2013).

In addition, traditional GWAS methods are criticised as causing the "missing heritability". It means that a large number of SNPs identified by GWAS may only account for a proportion of the heritability of complex traits. This is because *p*-values from GWAS only indicate whether SNPs are related to a phenotype or not; they cannot tell how strongly they are related. Therefore, some individual causal SNPs fail to pass the stringent significance thresholds because of their small effects to the phenotype (Manolio et al., 2009; Tam et al., 2019). Thus, we cannot completely rely on GWAS results as predictors for casual SNPs. Moreover, GWAS can be sensitive to the degree of match between the assumed statistical model (such as GLM or LMM) and the real data distribution. Last but not least, the single SNP analysis performed in these methods ignores the potential correlations and joint effects among SNPs.

To overcome these limitations, machine learning can be integrated in the statistical modelling of GWAS. Machine learning algorithms can identify relevant features from a complex data set or make accurate predictions from such data. They have been widely applied to whole-genome data analysis of various bioinformatics problems. When incorporating with GWAS, it could identify high orders of interactions between SNPs based on biological pathways and gene modules (epistasis), identify and prioritise SNPs with small effects by evaluating simultaneously pan-genomic SNPs (Sun et al., 2020), and serve as alternative to find casual SNPs and even predict phenotypes. Instead of

using machine learning or traditional statistical modelling alone, combining the two methods into a two-step model will integrate the strength of the separate methods.

The Least Absolute Shrinkage and Selection Operator (LASSO) is one of such machine learning methods that is suitable for "big-*n*-small-*p*" data just as the genotype. LASSO and its variations have been used to whole-genome data for feature selection and produce a subset of SNPs with significantly large effects to the phenotype (Waldmann et al., 2019). LASSO selects a subset of SNPs with the best joint effect explaining the phenotype of interests. These selected SNPs can also be used in the phenotype prediction and viewed as an alternative for genomic selection. Since LASSO can produce a reduced set of SNPs, it could be extremely useful as a "pre-processing" step for conducting GWAS on large genomes or polyploid crops, or as a "post-processing" step to select relevant SNPs with high individual effects and/or high joint effects from the "significant" SNPs reported by statistical methods like GWAS. GWAS methods often have limitations on the scale of SNPs and could fail to find the correlated SNPs on large complex data. The predictive model in LASSO can be validated using an independent test data set to measure its effectiveness. In order to utilise a machine learning method as LASSO, the whole data set is often arbitrarily split into two parts, a training data set and a test data set. When applying cross-validation, it splits the whole data set into training and testing repeatedly to increase the testing sample size. Cross-validation is thus usually applied to data sets with small sample size just like the genotype data.

In this paper, we present a two-step wrapper model that integrates LASSO into GWAS workflow and develop an automatic pipeline. It first uses traditional GWAS with statistical models to select SNPs with their significant values, and then uses LASSO to further select SNPs with high individual effects and/or high joint effects to enhance the model's results for relating SNPs to the phenotypes. In addition, the pipeline automatically integrates population structure calculation into the model to improve the prediction accuracy. Users do not need to generate the population structure matrix separately and manually feed it into any GWAS program in the pipeline. Since different GWAS programs often produce dissimilar association results, the proposed pipeline includes multiple GWAS packages to mitigate these effects. The pipeline is publicly available online at https://github.com/notTrivial/Comprehensive-GWAS.

3 Methods

In this section, the methods and software used in the proposed two-step wrapper model are introduced.

The typical statistical models used in GWAS include the GLM, LMM and the Multiple Locus Mixed Effects Model (MLMM). In this study, we select GLM and LMM. Typically, GLM models assumes the data is independent and LMM can deal with non-independence by adding random effects. The statistical model of a GLM can be described as

$$y = X\beta + e$$

and a LMM model can be described as

$$y = X\beta + Zu + e$$

where y is the vector of observations, β is an unknown vector containing fixed effects including genetic marker and population structure (Q), u is an unknown vector of random additive effects, X and Z are the known design matrices and e is the residuals (Bradbury et al., 2007). When the covariance between individuals is considered in the LMM model, it is viewed as a random additive genetic effect.

Our proposed model combines several open-source software to achieve a clear and simple GWAS workflow for end users. It conducts GWAS analysis utilising multiple packages. The current version includes TASSEL and GAPIT as the GWAS programs. TASSEL utilises GLM and LMM in determining associations; GAPIT uses similar statistical models as TASSEL and claims to be 7-fold faster than it. They are particularly popular in plant genome analysis. Since our study focuses on the plant genomes and agriculture, the two programs are included in the pipeline by default. However, it can be easily extended to include additional GWAS programs such as PLINK.

One important feature of the proposed model is that it integrates the automatic calculation of population structure of the input data. Including population structure in GWAS helps reduce false positive rate of output associations that are not related to the phenotype (Sul et al., 2018). To determine the most likely population structure, a widely used program, Structure, is integrated to calculate population structure matrix *Q*. This enables an accurate determination of the number of sub-populations (denoted as *k*) in a data set (Pritchard et al., 2000). To further verify these results, the Evanno method is included to perform an assessment and visualise the likelihood across all tested sub-population numbers in the data set (Evanno et al., 2005). It is implemented using the method from the pophelper v2.3.0 package (Francis, 2019) in *R*. The optimal population structure is then integrated in the model as a co-variance. Both TASSEL and GAPIT could take into consideration of this and use the population matrix as a co-variance.

After the GWAS analysis, SNPs along with *p*-values reported from each program go through a False Discovery Rate (FDR) adjustment, unless already done. The adjusted *p*-values are also known as *q*-values. Only those SNPs with *q*-values less than the cut-off are viewed as significant SNPs. In this study, the *q*-value=0.05 is set as the default cut-off.

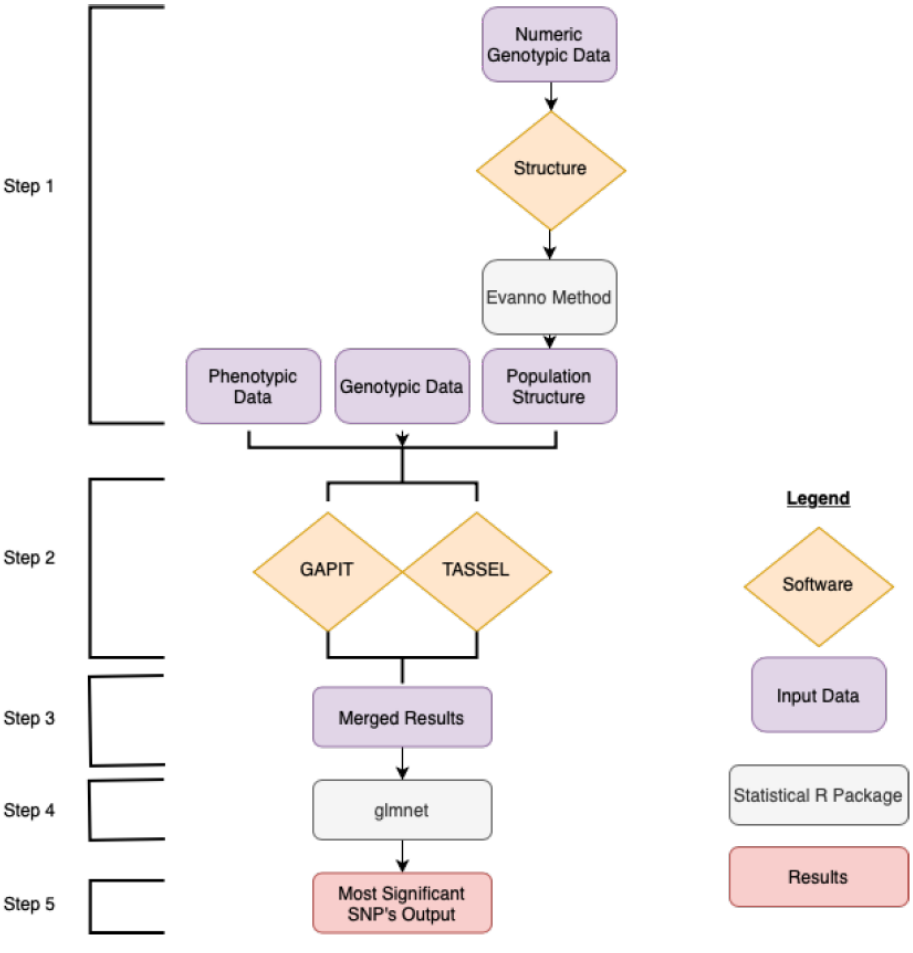
The combined significant SNPs and their *q*-values are then verified by LASSO with the use of glmnet package in *R* (Friedman et al., 2010). This package is responsible for fitting a GLM via penalised maximum likelihood. It is computed for both the elastic-net and LASSO penalty paths (Friedman et al., 2010). LASSO penalises non-zero coefficients by taking the sum of their absolute values, referred to as $L1$ and $L2$ penalties from Ridge Regression, where it penalises the model with the sum of squared individuals. Elastic-net combines both $L1$ and $L2$ penalties. Ridge regression is to shrink regression coefficients for variables with minor contribution to the model (Friedman et al., 2010). When combined with GWAS, the application of LASSO further narrows down the SNPs found by these programs and builds a predictive model for the phenotype. The most significant SNPs passed by LASSO are the final output from the model.

4 Pipeline overview

In this section, the details of the pipeline are introduced with clear explanations about input and output data of each step.

The pipeline takes genotype and phenotype data as input and produces association results along with visualisations and verification of association with the integration of LASSO. The entire workflow of the pipeline is shown as a flowchart in Figure 1.

Figure 1 The flowchart depicts the overall workflow of the pipeline



The pipeline is written in *R* v3.6.1. To execute the pipeline, TASSEL and Structure need to be pre-installed in the appropriate locations as described on https://github.com/notTrivial/Comprehensive-GWAS. The rest will be installed by the pipeline automatically at run time. It is designed to be as seamless as possible to end users to utilise different GWAS software and adjust output from each software to be the input for the next one. The steps taken in the pipeline are further described as follows.

Step 0: data pre-processing

The input data required for this pipeline is numeric genotype, non-numeric genotype, phenotype data and a kinship matrix. The numeric genotype is used for determining the population structure by Structure and validation by LASSO. To be specific, numeric genotype is a matrix with rows as individuals and columns as the SNP loci. When used in Structure, marker names and headings are removed as per its requirement. When used in LASSO, marker names and headings are present in the file. The non-numeric genotype is used for GWAS along with phenotypes and the kinship matrix. The phenotype data and kinship are in TASSEL required format and both need to have two files, one with the header and one without.

Step 1: determine optimal population structure

The Structure program (Pritchard et al., 2000) is a Bayesian clustering program that characterises population clusters to discriminate populations, to determine population structure and to reveal the genetic composition of individuals using molecular markers. It uses multilocus allele frequencies to assign individuals to a pre-defined number of populations (*k*). The assignment is usually run for a range of k such as 2 to 10. Multiple repeats are carried out for each *k*. Each output file for each repeat of k showing the assignment probabilities of all individuals is referred to as the *Q*-matrix.

The Evanno method (Evanno et al., 2005) is used to estimate the optimal number of populations, *k*, based on the results generated from the Structure program (Pritchard et al., 2000). The pophelper package (Francis, 2019) in *R* is the analysis tool of the Evanno method to analyse and visualise population structure. First, Structure examines possible numbers of sub-populations from 2 to *k*, where k is specified by the user, i.e. $k = 10$ is specified in the results section. This process is repeated E times for each possible number of sub-populations to ensure that the Evanno method is able to run. In this study, $E = 10$ is used as default. Results are then presented to the users and the optimal number of populations (*k*) is chosen. The resultant Q-matrix is the output of this step and then will be automatically fed into the follow-up GWAS analysis.

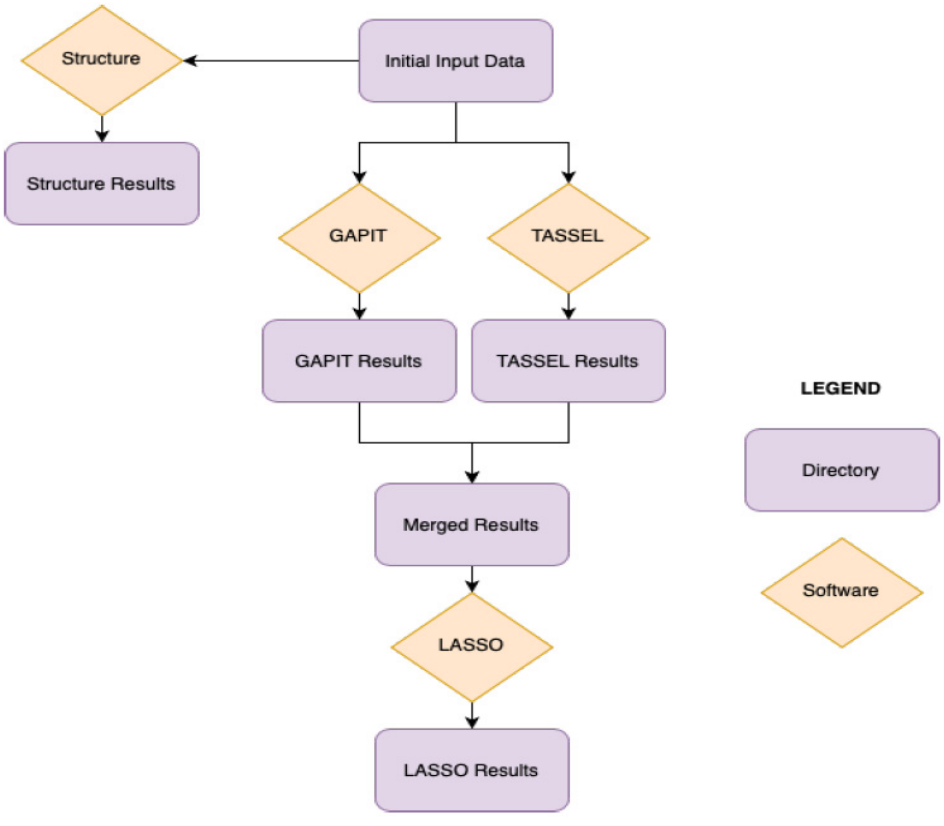
Step 2: GWAS with traditional statistical methods

The non-numeric genotype data, phenotype data, kinship matrix *K*, and population matrix *Q* are used as input to run GAPIT and TASSEL. The pipeline has the flexibility to choose between GLM and LMM models as both programs support them (Bradbury et al., 2007; Tang et al., 2016). Output of this step includes Manhattan Plots, QQ-Plots and statistical association results from TASSL and GAPIT. They are then output to their respective directories as shown in Figure 2.

Step 3: merge data set from statistical methods

Statistical association results from GAPIT and TASSEL are read back into the pipeline as the input of this step. Reported *p*-values of each program are adjusted using FDR. Since GAPIT can report both adjusted and un-adjusted *p*-values, in the current version, only TASSEL needs this adjustment. They are then merged together and the appropriate Manhattan plots of each statistical model (GLM or LMM) per phenotype are generated. Only the SNPs that meet the cut-off and are found by both programs will be kept as the output of this step and used in further verification steps.

Figure 2 The file structure chart depicts the overall data flow of the pipeline



Step 4: LASSO and cross-validation on merged data set

The significant SNPs output from above Step are verified using the LASSO method from the glmnet package in *R* to identify joint correlations. Cross-validation with 10 folds are applied to the significant SNPs of each statistical method per phenotype. The phenotype data is converted to discrete variables from numerical, anything over 50% is converted to 1 and less than 50% is converted to 0. The binomial method is applied in the validation. The coefficient plot and cross-validation plot are the output of this step and stored under the LASSO directory.

Step 5: generate final results

The final significant SNPs passed from above step of each statistical method per phenotype are output to respective directories. Combined with other output from previous steps, all the results are stored under the directories as described in Figure 2.

5 Experiments and results

In this section, we describe the experiments and report the results of using the proposed pipeline on a plant genomeic and phenotypic data set.

5.1 Experimental data

Fusarium Head Blight (FHB) is a fungal disease that affects both wheat and barley worldwide. It is a serious disease that causes reduced crop yields and produces mycotoxins which pose serious health threats to both humans and livestock (Walter et al., 2010; Hilton et al., 1999). Therefore, detecting FHB resistance has significant meaning in plant breeding. GWAS has been widely used to provide important insights into the linkages of complex traits to causal SNPs and its related genes in plant breeding (Yan et al., 2019).

In this study, 199 bread wheat varieties were selected with the lowest FHB severity rate among a poll of 4000 varieties. 2235 SNPs were detected using wheat 90 k SNP array (Wang et al., 2014). Three phenotypes including flowering date, plant height and disease severity have been screened in a FHB disease greenhouse in Saskatchewan, Canada in 2019.

5.2 Population structure determination

The pipeline's built-in function identifies the most likely population structure of the experimental data. Figures 3 and 4 show the number of population verified by Evanno method. As noted in both figures, the graph plateau's at $k = 2$, which supports that $k = 2$ is the optimal number. The Q-matrix produced by Structure with $k = 2$ is used in the subsequent GWAS runs (Evanno et al., 2005). Figure 5 is a stacked barplot that shows the population group assignment of each individual in 100 repetitions for k ranging from 2 to 7 groups of populations.

Figure 3 The likelihood distribution of k results using the pophelper package in *R* (Step 1 in Figure 1). The graph shows the mean of $L(k)(\pm SD)$ over 10 runs for each k value. It's plateau at $k = 2$, which supports the usage of its corresponding Q-matrix

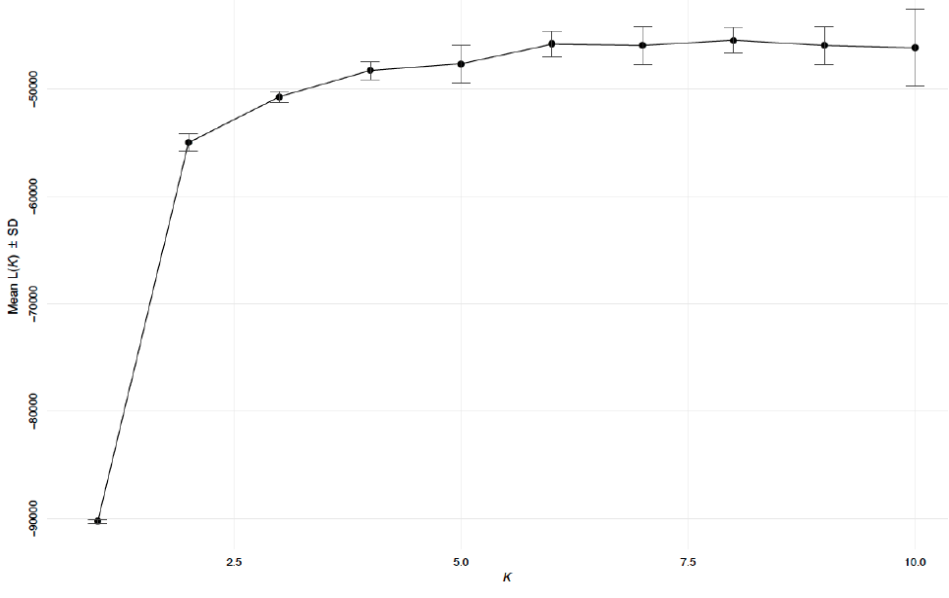
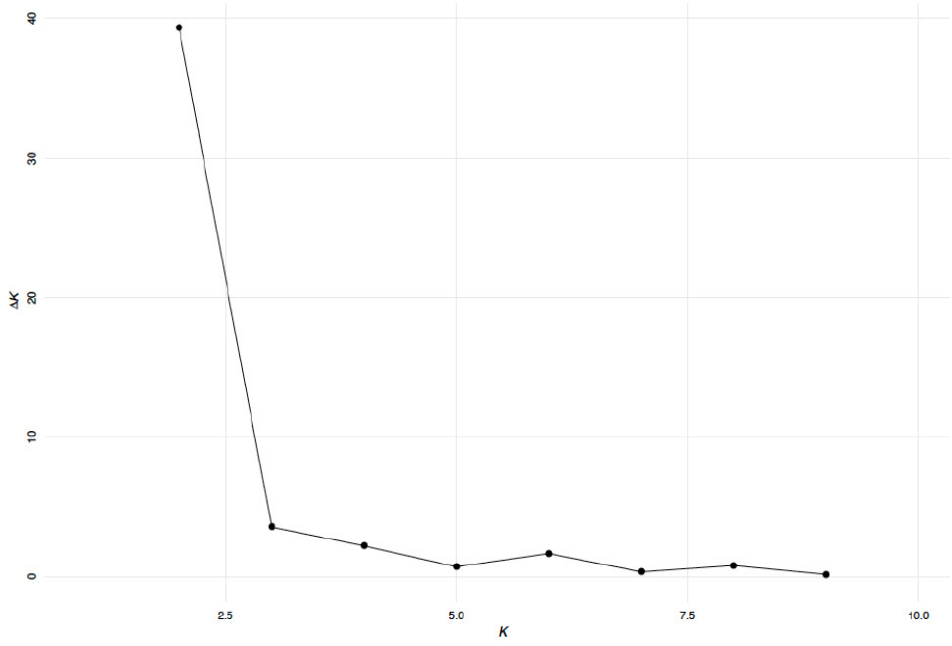


Figure 4 Δk plot results using the Evanno method in the pophelper package in *R*. Δk is calculated from the absolute value of the second order rate of change of the likelihood distribution divided by the standard deviation of the likelihood distribution (Step 1 in Figure 1). The modal value of this distribution is the true k or the uppermost level of structure. Here $k = 2$ clusters because the graph plateau's at $k = 2$. This supports the usage of the Q-matrix of $k = 2$



5.3 GWAS results

From all three tested phenotypes, two of them, plant height and disease severity, did not get significant results by TASSEL and GAPIT. The resultant QQ plots from them showed that the model did not fit the data well. However, this is also likely due to the fact that the 90 k SNP array only detects SNPs in "preselected" genomic regions that have a higher likelihood to be associated with traits of interest without including any irrelevant positions (Tang et al., 2016; Bradbury et al., 2007). In the following, results on flowering date are shown and explained in details.

Figures 6 and 7 show the Manhattan plots of TASSEL and GAPIT's GWAS results. The SNP's *p*-values are with a scale of $-\log10(p-value)$. They are both produced by the GLM of the two programs as The LMM found no significant results for either of them. Both programs identified the most significant SNPs located on chromosomes 2B and 6B.

Figure 5 Multiline barplots from Q-matrices from $k = 2$ to $k = 7$ for step 1 in Figure 1. Each group is individually labelled

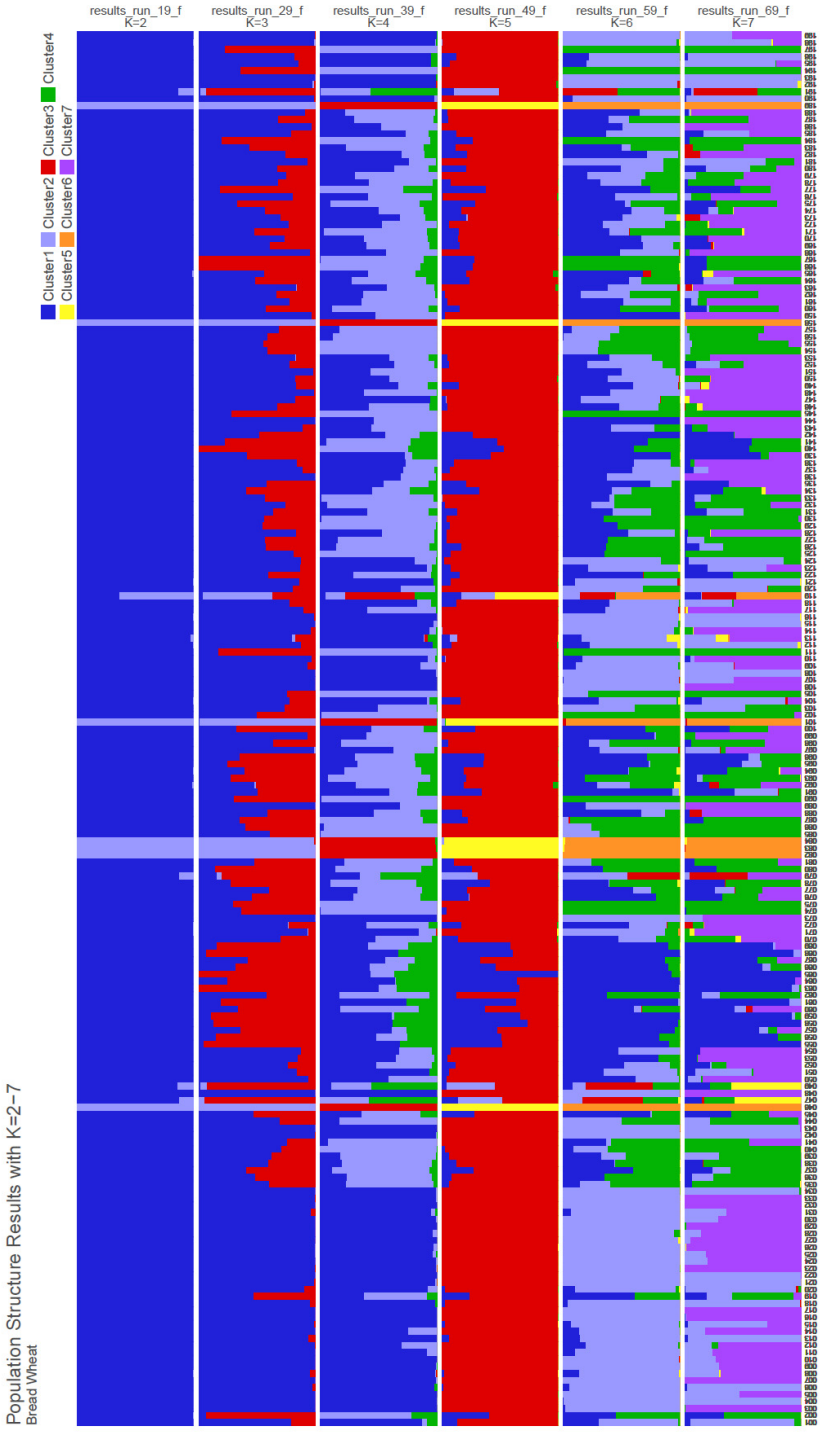


Figure 6 Manhattan plot for flowering date by GLM of GAPIT. The most significant SNPs are above the green line and located on chromosomes 2B and 6B (Step 2 in Figure 1)

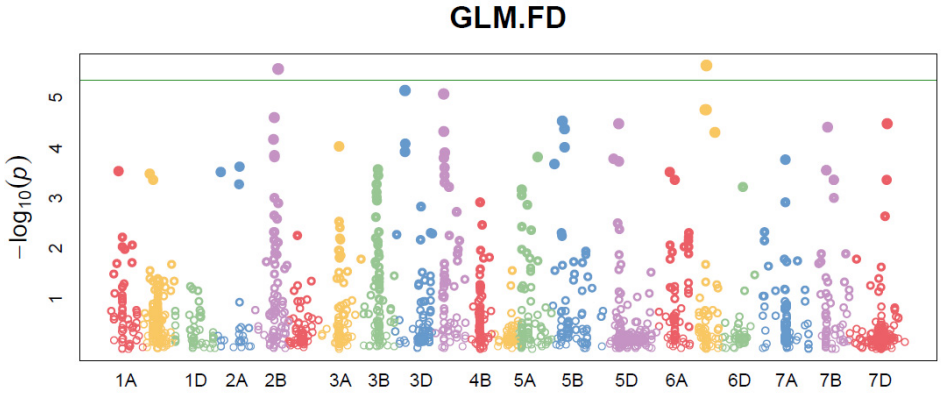
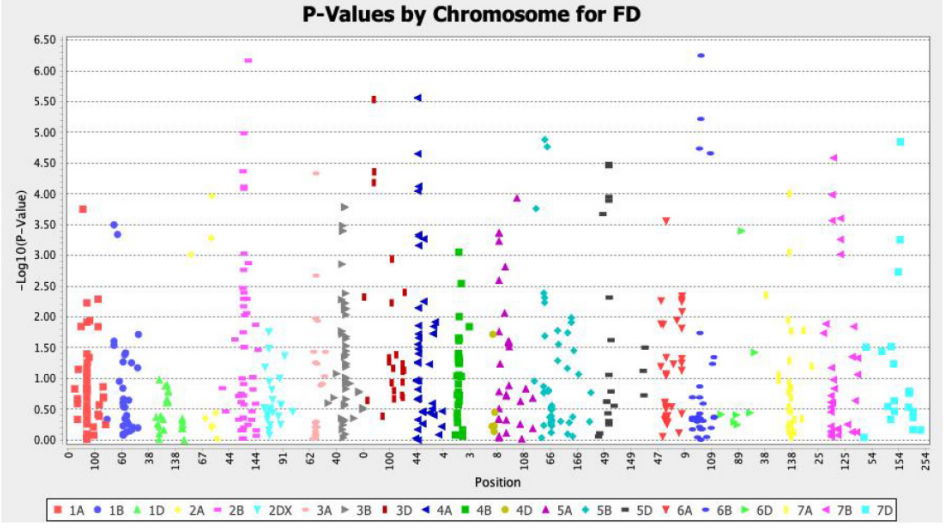
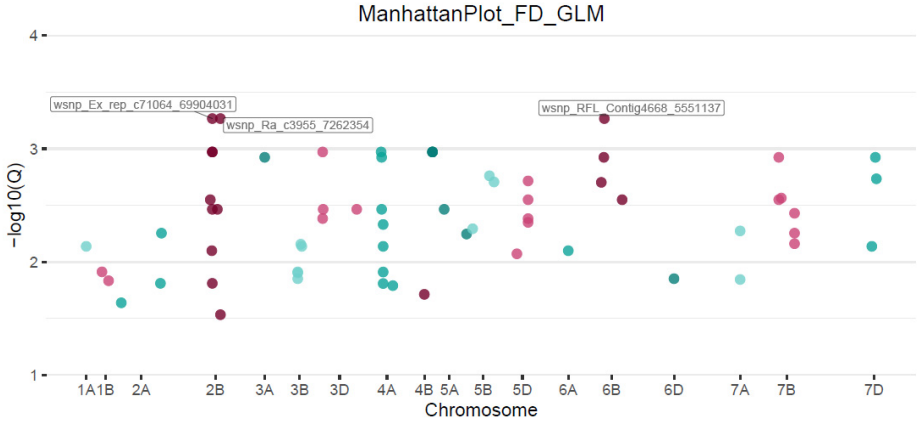


Figure 7 Manhattan plot for flowering date by GLM of TASSEL. The most significant SNPs are located on chromosomes 2B and 6B (Step 2 in Figure 1)



There were a total of 65 significant (q-values $\leq$ 0.05) SNPs predicted by the pipeline using GLM of TASSEL and GAPIT. The q-value is adjusted by the FDR (Benjamini and Hochberg, 1995) of the p-value. They are the merged results from the two programs. Among the whole genome, chromosome 2B and chromosome 6B have the largest number of SNPs which the two programs agree on. This indicates that the two regions may have significant influence to the flowering time phenotype. The pipeline output is summarised in the merged Manhattan plot in Figure 8. It shows the merged significant SNPs Manhattan Plot with a scale of $-\log 10\,(q-value)$. The coloured columns help distinguish different chromosomes. Each dot on this Manhattan plot corresponds to one significant SNP (q-values $\leq$ 0.05), though some overlap to appear as a larger circle.

Figure 8 Manhattan plot for flowering date using the merged SNPs from both TASSEL and GAPIT. All 65 significant SNPs (whose q-values ≤ 0.05) are illustrated in this graph. The colours indicate different chromosomes from its neighbouring chromosomes (Step 3 in Figure 1)



The Manhattan plot also highlights the top 3 significant merged SNPs with their names displayed. These top 3 significant SNPs are also listed in Table 1 with their chromosome and location information. The results show that our proposed pipeline captured significant SNPs from both GWAS software and they could have joint effects to the phenotype.

Table 1 Top 3 Significant SNPs after merging (Step 5 in Figure 1)

Name of SNP	*Chromosome*	*Position*
wsnp_RFL_Contig4668_5551137	6B	71
wsnp_Ex_rep_c71064_69904031	2B	100
wsnp_Ra_c3955_7262354	2B	116

5.4 LASSO and cross-validation

Top significant SNPs found by the glmnet of *R* package are shown in Table 2. They played an important role in fitting the model and therefore were concluded to be the top significant SNPs from the glmnet analysis. SNPs used in glmnet are only the 65 significant SNPs from the merged results of GAPIT and TASSEL. We notice that these top SNPs are not the same as the SNPs with the lowest q-values. But they could pinpoint new regions that researchers could focus on to find potential casual relationships to the phenotype. As we stated in Section 2, GWAS results only reflect associations but not how strong they are. SNPs reported by LASSO may influence the phenotype but failed to pass the q-value threshold.

Table 2 Top 5 significant SNPs found by glmnet in step 5 of Figure 1. These SNPs contributed more to the development of the model fitting than the other 65 significant SNPs from the merged results

Name of SNP	*Chromosome*	*Position*
Excalibur_rep_c114833_645	3D	41
D_contig25474_188	7D	161
RAC875_c5243_479	1B	31
BS00065005_51	1A	55
wsnp_Ex_c56629_58677561	5B	12

The coefficients plot from the model fitting of SNPs is shown in Figure 9, which includes all 65 significant SNPs. Each curved line in Figure 9 corresponds to one SNP and it shows the path of each SNPs' coefficient against the $L1$-norm. The number of non-zero coefficients is indicated as the numbers in the x-axis on the top.

Figure 10 shows the corresponding cross-validation result. It shows the validation of the model, the choice of two λs, and the LASSO regularisation parameter, as indicated by the vertical dotted lines. The x-axis on the top is the number of non-zero coefficients just the same as the ones in Figure 9. The red dotted line is the cross-validation curve and the error bars indicate the upper and lower standard deviation of the λ. One λ gives the minimum mean cross-validated error and the other gives the most regularised model such that the error is within one standard error of the minimum (Friedman et al., 2010). Figure 10 also shows that the misclassification error rate is about 0.20, which indicates the predictivity of the selected SNPs. This highlights the significance of the SNPs from the merged results.

Figure 9 Coefficients from LASSO for the 65 merged significant SNPs (Step 4 in Figure 1)

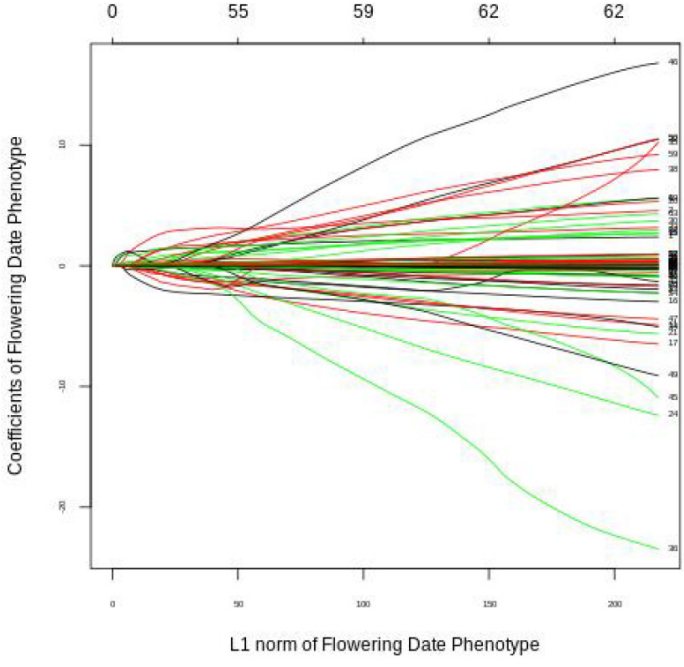
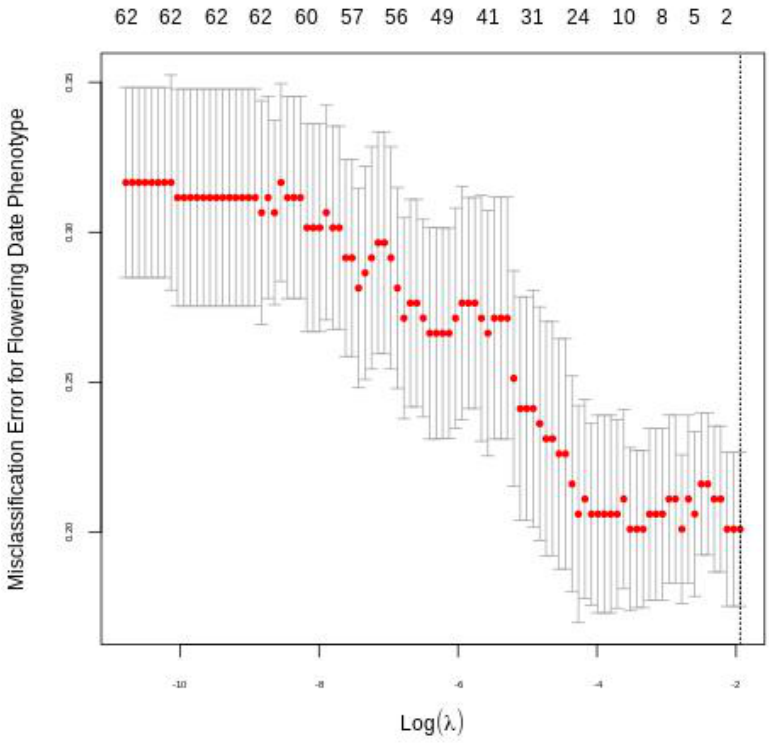


Figure 10 Cross-validation from LASSO for the 65 merged significant SNPs (Step 4 in Figure 1)



6 Conclusions and discussions

In this study, we proposed a computational method for seamless GWAS analyses with statistically modelling and machine learning. It uses an ensemble approach that allows easy analysis of associations between genotype and phenotype data to narrow down areas and SNPs of "true" significance while being efficient and easy to handle. Combining results from two well-known GWAS programs yields more promising results, which allows overlapping analysis of multiple programs across multiple statistical models. Automatically including the population structure analysis in these GWAS programs helps to reduce a high rate of false positive associations for SNPs that are not linked with phenotypes. The final step of the pipeline further verifies the significant SNPs by using LASSO to provide better accuracy in detecting "true" significant SNPs to help narrow down the number of resultant SNPs (Friedman et al., 2010).

The pipeline is created for GWAS analysis in one place while utilising various software. It also has the capacity to extend and include more GWAS programs and statistical models to be utilised in other analyses. We demonstrated the utility of this

pipeline in associating FHB resistance related phenotypes with genomic regions in bread wheat varieties. Significant SNPs were found in the flowering date phenotype by GLM, whereas the other studied phenotypes had no significant SNPs found with the *q*-value threshold of 0.05 using TASSEL and GAPIT. The significant SNPs in the flowering date phenotype was verified by FDR adjusted *p*-values and were further verified by cross-validation using LASSO (Friedman et al., 2010). Having this result verified by both GAPIT and TASSEL lends credibility to the finding of these significant SNPs.

There are a few directions where the pipeline will be extended to obtain better and more comprehensive results. For example, instead of detecting associated SNPs for each phenotype independently, the pipeline can be extended to combine results from multiple related phenotypes. It is known that flowering date, plant height and disease severity rates are phenotypes related to FHB disease; therefore, genetic variants that are common across multiple related phenotypes or unique to any specific phenotype can be aggregated and associated with the FHB disease. Further, given that results from different GWAS programs could vary largely in terms of output SNPs as well as their *p*-values, in addition to the current *q*-value threshold, the pipeline can include other selection criterion such as choosing the top significant SNPs without meeting the *q*-value threshold. Finally, more GWAS packages such as PLINK can be included in the pipeline to complement with current two software, GAPIT and TASSEL. In this case, additional file format conversion could also be added into the pipeline, for example, to convert PLINK ped/bed files to the VCF files. This will enable the pipeline to be further extended to synthesise and report significant SNP sets in a customised but uniform way.

References

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 57, No. 1, pp.289–300.

Bradbury, P.J. and Zhang, Z., Kroon, D.E., Casstevens, T.M. (2007) 'TASSEL: software for association mapping of complex traits in diverse samples', *Bioinformatics*, Vol. 23, No. 19, pp.2633–2635.

Evanno, G., Regnaut, S. and Goudet, J. (2005) 'Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study', *Mol. Ecol.*, Vol. 14, No. 8, pp.2611–2620.

Francis, R.M. (2019) *Tabulate, Analyse and Visualise Admixture Proportions from STRUCTURE, TESS, BAPS, ADMIXTURE and Tab-Delimited q-matrices Files*, R package version 2.3.0.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) 'Regularization paths for generalized linear models via coordinate descent', *Journal of statistical software*, Vol. 33, No. 1, pp.1–22.

Goddard, M.E. and Hayes, B.J. (2007) 'Genomic selection', *Journal of Animal breeding and Genetics*, Vol. 124, No. 6, pp.323–330.

Hilton, A.J., Jenkinson, P., Hollins, T.W. and Parry, D.W. (1999) 'Relationship between cultivar height and severity of fusarium ear blight in wheat', *Plant Pathology*, Vol. 48, No. 2, pp.202–208.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S. and Zhang, Z. (2012) 'GAPIT: genome association and prediction integrated tool', *Bioinformatics*, Vol. 28, No. 18, pp.2397–2399.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I. and Heckerman, D. (2011) 'Fast linear mixed models for genome-wide association studies', *Nature Methods*, Vol. 8, No. 10, pp.833–835.

Loh, P-R., Tucker, G., Bulik-Sullivan, B-K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M. and Berger, B. et al. (2015) 'Efficient Bayesian mixed-model analysis increases association power in large cohorts', *Nature Genetics*, Vol. 47, No. 3, pp.284–290.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R. and Chakravarti, A. et al. (2009) 'Finding the missing heritability of complex diseases', *Nature*, pp.747–753.

Oetting, W.S., Jacobson, P.A. and Israni, A.K. (2017) 'Validation is critical for genome-wide association study-based associations', *American Journal of Transplantation*, Vol. 17, No. 2, pp.318–319.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) 'Inference of population structure using multilocus genotype data', *Genetics*, Vol. 155, No. 2, pp.945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W. and Daly, M.J. et al. (2007) 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *The American Journal of Human Genetics*, Vol. 81, pp.559–575.

Sul, J.H., Martin, L.S. and Eskin, E. (2018) 'Population structure in genetic studies: confounding factors and mixed models', *PLoS Genetics*, Vol. 14, No. 12. Doi: 10.1371/journal.pgen.1007309.

Sun, S., Dong, B. and Zou, Q. (2020) (2020) 'Revisiting genome-wide association studies from statistical modelling to machine learning', *Briefings in Bioinformatics*. Doi: 10.1093/bib/bbaa263.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. and Meyre, D. (2019) 'Benefits and limitations of genome-wide association studies', *Nature Reviews Genetics*, Vol. 20, No. 8, pp.467–484.

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D. and Lipka, A.E. et al. (2016) 'GAPIT version 2: an enhanced integrated tool for genomic association and prediction', *The plant genome*, Vol. 9, No. 2, pp.1–9.

Waldmann, P., Ferenčaković, M., Mészáros, G., Khayatzadeh, N., Curik, I. and Sölkner, J. (2019) 'Autalasso: an automatic adaptive lasso for genome-wide prediction', *BMC Bioinformatics*, Vol. 20, No. 1, pp.1–10.

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C. and Sölkner, J. (2013) 'Evaluation of the lasso and the elastic net in genome-wide association studies', *Frontiers in genetics*, Vol. 4. Doi: 10.3389/fgene.2013.00270.

Walter, S., Nicholson, P. and Doohan, F.M. (2010) 'Action and reaction of host and pathogen during Fusarium head blight disease', *New Phytologist*, Vol. 185, No. 1, pp.54–66.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B.E., Maccaferri, M., Salvi, S., Milner, S.G. and Cattivelli, L. et al. (2014) 'Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array', *Plant Biotechnology Journal*, Vol. 12, No. 6, pp.787–796.

Yan, Y., Burbridge, C., Shi, J., Liu, J. and Kusalik, A. (2018) 'Comparing four genome-wide association study (GWAS) programs with varied input data quantity', *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp.1802–1809.

Yan, Y., Burbridge, C., Shi, J., Liu, J. and Kusalik, A. (2019) 'Effects of input data quantity on genome-wide association studies (GWAS)', *International Journal of Data Mining and Bioinformatics*, Vol. 22, No. 1, pp.19–43.

Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) 'GCTA: a tool for genome-wide complex trait analysis', *The American Journal of Human Genetics*, Vol. 88, No. 1, pp.76–82.