Machine Learning based Genome-Wide Association Studies for Uncovering QTL Underlying Soybean Yield and its Components

Mohsen Yoosefzadeh-Najafabadi
University of Guelph Department of Plant Agriculture
Sepideh Torabi
University of Guelph Department of Plant Agriculture
Davoud Torkamaneh
Universite Laval
Dan Tulpan
University of Guelph Ontario Agricultural College
Istvan Rajcan
University of Guelph Department of Plant Agriculture
Milad M Eskandari (meskanda@uoguelph.ca)
University of Guelph Department of Plant Agriculture https://orcid.org/0000-0003-0862-4048

Research Article

Keywords: Data-driven Models, FarmCPU, Genome-wide association study, MLM, Soybean Breeding, Support vector machine

Posted Date: October 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-936233/v1

Abstract

Genome-wide association study (GWAS) is currently one of the most recommended approaches for discovering quantitative trait loci (QTL) associated with complex traits in plant species. Insufficient statistical power is a limiting factor that current conventional GWAS methods are suffering from, especially in narrow genetic bases plants such as soybean. In this study, we evaluated the potential use of two machine learning (ML) algorithms, support vector machine (SVR) and random forest (RF), in GWAS and compared them with two conventional methods of mixed linear models (MLM) and fixed and random model circulating probability unification (FarmCPU), for identifying QTL associated with soybean yield and its components. In this study, important soybean yield component traits, including the number of reproductive nodes (RNP), non-reproductive nodes (NRNP), total nodes (NP), and total pods (PP) per plant along with yield and date to maturity were assessed using a panel of 227 soybean genotypes evaluated across four environments. Using the SVR-mediated GWAS method, we were able to discover a greater number and most relevant QTL associated with the target traits, supported by the functional annotation of candidate gene analyses. This study for the first time demonstrated the potential benefit of using sophisticated mathematical approaches, such as ML algorithms, in GWAS for identifying QTL that can improve the efficiency of genomic-based breeding programs.

Key Message

Implementing machine learning algorithms in GWAS can simultaneously consider a wide range of interconnected biological processes and mechanisms that shape the phenotype of complex traits such as yield and its components in soybean.

Introduction

Soybean (*Glycine max* [L.] Merr.) is known as one of the most important legume crops with substantial economic value (Rębilas et al. 2020). Soybean is widely used for food, feed, fiber, biodiesel, and green manure (Carter et al. 2018; Temesgen and Assefa 2020). Despite the importance of genetic improvement in soybean yield, the soybean germplasm has in general a narrow genetic basis, especially within the North American germplasm, which has resulted in limited historical enhancement of the genetic gain (Xavier and Rainey 2020). Therefore, there is a great need for analytical breeding to explore the optimum genetic potential for soybean yield (Mangena 2020; Suhre et al. 2014).

Analytical breeding strategy as an alternate breeding approach requires a better understanding of the factors, or individual traits, responsible for more complex characteristics and phenomenon such as plant growth, development, and yield (Richards 1982). This strategy considers highly correlated secondary traits with the trait of interest as the selection criteria that can make empirical selection more efficient for improving the genetic gain (Reynolds 2001; Richards 1982; Xavier and Rainey 2020). The application of the analytical approaches in plant breeding programs has been limited due mainly to limited resources for evaluating several secondary traits that are mostly time and labor-consuming (Richards 1982; Xavier

et al. 2018). Most of analytical breeding studies were conducted on small populations with limited number of genotypes and, therefore, the results with limited generalization ed in the limitation of the knowledge in the genome-to-phenome analysis process (Kahlon et al. 2011; Nico et al. 2019; Robinson et al. 2009).

Yield potential in soybean is mainly determined by its yield component traits that are the total number of pods, seeds, and nodes, which can be divided into reproductive and non-reproductive nodes, in plants as well as seed size (Pedersen and Lauer 2004; Reynolds 2001; Xavier et al. 2018; Xavier and Rainey 2020; Yoosefzadeh-Najafabadi et al. 2021b). Of these traits, the total number of nodes and pods play more important roles in seed yield production (Robinson et al. 2009; Yoosefzadeh-Najafabadi et al. 2021b). Several studies reported a steady increase in the total number of nodes and the total number of pods in soybean cultivars from 1920 to 2010 (Kahlon et al. 2011; Suhre et al. 2014; Xavier and Rainey 2020). These findings may highlight the importance and potential use of the phenotypic and genotypic information on these traits, along with yield per se, as selection criteria in cultivar development programs (Ma et al. 2001).

Genetic information of soybean yield component traits can accelerate the efficiency of cultivar development programs through selecting genotypes in improved genetic gains (Xavier and Rainey 2020). Genome-Wide Association Studies (GWAS), as one of the most common genetic approaches, can be implemented on genetically diverse populations to detect the quantitative trait loci (QTL) associated with the soybean yield component traits (Kaler et al. 2020). Associated QTL can be used for screening large soybean populations in a more cost-effective and timely manner (Xavier et al. 2018). Up to date, several GWAS approaches such as mixed linear models (MLM), multiple loci linear mixed model (MLMM), and fixed and random model circulating probability unification (FarmCPU) have been developed for genetic studies of complex traits (Kaler et al. 2020). However, due to the narrow genetic base of some plant species, including soybean, these conventional approaches may not have sufficient statistical power to detect reliable QTL (Kaler et al. 2020; Mohammadi et al. 2020; Xavier and Rainey 2020). Therefore, the development of more sophisticated statistical methods can help to establish effective GWAS methods for plant species with narrow genetic bases.

Current GWAS methods are based on the conventional statistical methods that are useful for studying less complex traits in plant species with broader genetic bases (Lipka et al. 2015; Pasaniuc and Price 2017). Machine learning (ML) algorithms as powerful and reliable mathematical methods have been considered as an alternative to conventional statistical methods for dealing with large data set as well as conducting GWAS analyses (Xavier and Rainey 2020). Recently, the use of ML algorithms has been reported in different areas such as plant science (Hesami et al. 2020; Yoosefzadeh-Najafabadi et al. 2021a), animal science (Tulpan 2020), human science (Chen and Verghese 2020), engineering (Kim et al. 2020), and computer science (Jordan and Mitchell 2015). The application of ML algorithms in GWAS was previously investigated in human science by Szymczak et al. (2009). In this study, the potential use of different ML algorithms such as artificial neural networks (ANN), Bayesian network analysis (BNA), and random forests (RF) are explained for using in GWAS studies focused on human disease studies

(Szymczak et al. 2009). One of the most common used ML algorithms is RF, developed by Breiman (2001), which generates a series of trees from the independent samples for a better prediction performance (Meinshausen 2006). The latter algorithm has been widely used in plant genomics (Ogutu et al. 2011), phenomics (Yoosefzadeh-Najafabadi et al. 2021a), proteomics (Jamil et al. 2020), and metabolomics (Sun et al. 2020).

The first and only use of the RF-mediated GWAS in soybean, for detecting the genomic regions association with yield component traits, was reported by Xavier and Rainey (2020). Support vector machine (SVM) is another common algorithm that can detect behavior and patterns of nonlinear relationships (Auria and Moro 2008; Hesami and Jones 2020; Su et al. 2017). Theoretically, SVM should have high performance due to the use of structural risk minimization instead of the empirical risk minimization inductive principles (Belayneh et al. 2014; Yoosefzadeh-Najafabadi et al. 2021a). There is a significant number of reports on the successful using of SVM in prediction problems (Denton and Salleb-Aouissi 2020; Duan et al. 2005; Hesami et al. 2020; Tulpan 2020; Yoosefzadeh-Najafabadi et al. 2021a). Support vector regression (SVR) is known as the regression version of SVM that commonly used for continuous variables. There are also reports on the successful use of SVR for addressing plant prediction problems (Awad and Khanna 2015). However, the potential use of SVR in GWAS is still unexplored remained to be discovered in the area of plant science.

In this study we aimed to: (1) gain a better understanding of the genetic relationships between soybean yield and its component traits, and (2) investigate the potential use of RF and SVM algorithms in GWAS for discovering QTL underlying soybean yield components in compared with the conventional GWAS methods of MLM and FarmCPU. The results of this study will help soybean breeders and geneticists to have a better perspective of exploiting ML algorithms in GWAS studies and may offer new genomic tools for screening high yielding genotypes with improved genetic gain in large breeding populations.

Materials And Methods

Population and experimental design

An GWAS panel of 250 soybean genotypes was grown at the University of Guelph, Ridgetown Campus in two locations, Palmyra (42°25'50.1"N 81°45'06.9"W, 195 m above sea level) and Ridgetown (42°27'14.8"N 81°52'48.0"W, 200m above sea level) in Ontario, Canada, in two consecutive years, 2018 and 2019. The panel used in this study consisted of the main germplasm of the soybean breeding program at the University of Guelph, Ridgetown Campus, that has been established over 35 years for cultivar development and genetic studies. The randomized complete block design (RCBD) with two replications was used for all four environments. In general, there were 500 and 1000 research plots per environment and year, respectively. Each plot consisted of five 4.2 m long rows with 57 seeds per m^2 seeding rate.

Phenotyping

In this experiment, soybean seed yield (t ha^{-1} at 13% moisture) for each plot was estimated by harvesting three middle rows and adjusted based on the maturity data. Soybean seed yield components, including the total number of reproductive nodes per plant (RNP), the total number of non-reproductive nodes per plant (NRNP), the total nodes per plant (NP), and the total number of pods per plant (PP), were measured using 10 randomly selected plants from each plot. The maturity was recorded as the number of days from planting to physiological maturity (R7, (Fehr and Caviness 1971) for each genotype.

Genotyping

Young trifoliate leaf tissue for each soybean genotype from the first replication of the trail at the Ridgetown in 2018, were collected and in a 2 mL screw-cap tube. The leaf samples were freeze-dried for 72 hours, using the Savant ModulyoD Thermoquest (Savant Instruments, Holbrook, NY). By using the DNA Extraction Kit (SIGMA®, Saint Louis, MO), DNA was extracted for soybean genotypes, and the quantity of DNAs was checked via Qubit® 2.0 fluorometer (Invitrogen, Carlsbad, CA). For genotyping-by-sequencing (GBS), DNA samples were sent to Genomic Analysis Platform at Université Laval (Laval, Quebec, Canada). The GWAS panel was genotyped via a GBS protocol based on the enzymatic digestion with *ApeKI* (Sonah et al. 2013). Single-nucleotide polymorphisms (SNPs) were called by the Fast GBS pipeline (Torkamaneh et al. 2020), using Gmax_275_v2 reference genome. Markov model was used to impute the missing loci, and SNPS with a minor allele frequency (MAF) less than 0.05 were removed below the threshold. In total, after checking the quality of reading sequence and removing SNPs with more than 50% heterozygosity, 23 genotypes were eliminated from the experiment and 17,958 high-quality SNPs from 227 soybean genotypes used for genetic analysis.

Statistical analyses

The best linear unbiased prediction (BLUP) as one of the common linear mixed models (Goldberger 1962) was used to estimate the genetic values of each soybean genotype. Also, R package *lme4* (Bates et al. 2014) was used to analyze yield and yield components with 'environment' as a fixed effect and 'genotype' as a random effect. To control for the possible soil heterogeneity among the plots within a given block and reduce the associated experimental errors, nearest-neighbor analyses (NNA) was used as one of the common error control methods (Bowley 1999; Katsileros et al. 2015; Stroup and Mulitze 1991). Outliers were determined in the raw dataset based on the protocols proposed by Bowley (1999) and treated the same as missing data points in the analysis. Overall, the following statistical model (Eq. 1) was used in this study:

$$Y_{ij} = \mu + f(s) + G_i + E_j + GE_{ij} + \varepsilon_{ij}, i = 1, \dots, k; j = 1, \dots, n \quad \text{(Eq. 1)}$$

Where Y_{ij} stands for the trait of interest (soybean seed yield and yield component traits) as a function of an intercept μ, f(s) stands for the spatial covariate, G_i is the random genotype effect, E_j stands for the

fixed environment effect, GE_{ij} is the genotype x environment interaction effect, and ε_{ij} stands for the residual effect.

The heritability (Eq. 2) was calculated for soybean seed yield and yield components using *lme4* open-source R package (Bates et al. 2007) based on the following equation:

$$H^2 = \frac{6_G^2}{6_G^2 + 6_E^2} \quad \text{(Eq. 2)}$$

where 6^2_G stands for the genotypic variance, and 6^2_E is the environmental variance.

Analysis of population structure

A total of 17,958 high-quality SNPs from 227 soybean genotypes were used to conduct population structure analysis using fastSTRUCTURE (Raj et al. 2014). Five runs were conducted for K set from 1 and 15 to estimate the most appropriate number of subpopulations by using the K tool from the fastSTRUCTURE software.

Association studies

Since different GWAS methods may capture different genomic regions (Yang et al. 2018). Therefore, MLM and FarmCPU (two most common GWAS methods) and RF and SVM (two most common machine learning algorithms) were used in this study. MLM and FarmCPU were implemented by using *GAPIT* package (Lipka et al. 2012), and RF, as well as SVM, were conducted through the *Caret* package (Kuhn et al. 2020) in R software version 3.6.1. A brief description of each of the GWAS methods is provided below:

Mixed Linear Model (MLM)

This GWAS is based on the likelihood ratio between the full model, consisting of the marker of interest, and the reduced model, which is known as the model without the marker of interest (Wen et al. 2018).

Fixed and random model circulating probability unification (FarmCPU)

This GWAS takes the advantages of using MLM as the random model, and stepwise regression as the fixed model iteratively (Liu et al. 2016). False discovery rate (FDR) is used for setting the threshold both in the FarmCPU and MLM models (Benjamini and Hochberg 1995).

Random Forest (RF)

This machine-learning algorithm was first implemented by Xavier and Rainey (2020) in a soybean GWAS study. This method is known as the powerful non-parametric regression approach that is derived from aggregating the bootstrapping in various decision trees (Breiman 2001). In this experiment, a 1000-set of decision trees constructed the forest, and the GWAS analysis was done by measuring the importance of each feature (Botta et al. 2014), which was an SNP in this study.

Support vector regression (SVR)

This machine learning algorithm is known as one of the common supervised learning methods in prediction problems (Cortes and Vapnik 1995). This algorithm is based on constructing a set of hyperplanes that can be useful in regression problems (Fletcher 2009). The association statistics in this algorithm can be achieved by estimating the feature importance that was previously proposed by Weston et al. (2001). In this experiment, SNP markers were selected as inputs, and the traits were selected as target variables for estimating the feature importance.

Variable Importance measurement

As one of the common indices for tree-based algorithms, the impurity index was chosen as the metric of the feature importance for the RF algorithm. Regarding the SVR algorithm, the variable importance method for SVR Weston et al. (2001) was implemented in this dataset. For both algorithms, the importance of each SNP was scaled based on 0 to 100 percent scale. Since there is no confirmed way of defining the significant threshold in the tested algorithms, the global empirical threshold that provides the empirical distribution of the null hypothesis (Churchill and Doerge 1994; Doerge and Churchill 1996) was used for establishing threshold in this study. The global empirical threshold was estimated based on fitting the ML algorithm, storing the highest variable importance, repeating 1000 times, and select the SNPs based on α=0.05.

Data-driven model processes

In order to estimate the feature importance in RF and SVR algorithms, a five-fold cross-validation strategy (Siegmann and Jarmer 2015) with ten repetitions was applied on the dataset. All of the tested machine learning algorithms were optimized for their parameters for this dataset accordingly.

Extracting candidate genes undelaying detected QTLs

For each tested GWAS model, the flanking regions of each QTL was determined using LD decay distance (Fig.1), and then potential candidate genes were retrieved using the *G. max* cv. William 82 reference genome, gene models 2.0 in SoyBase (https://www.soybase.org). After listing potential candidate genes in defined windows around each significant SNP, at the peak of each QTL, Gene Ontology annotation, GO

term enrichment (https://www.soybase.org), and the report from previous studies were used as the criteria to select and report the most relevant candidate genes associated with the identified QTL. The Electronic Fluorescent Pictograph (eFP) browser for soybean (www.bar.utoronto.ca) was also used to generate additional information such as tissue- and developmental-stage dependent expression (based on transcriptomic data from Severin et al. (2010)) for the identified candidate genes.

Visualization

All of the visualizations in this study were conducted using the *ggplot2* package (Wickham 2011) in R version 3.6.1 software and Microsoft Excel software (2016).

Results

Phenotyping evaluations

The panel, consisted of 227 soybean genotypes, showed a different level of variations among the genotypes for seed yield, maturity, and yield component traits. The distribution of the phenotypic measures for the target traits across the four environments is presented in Fig. 2. The highest heritability was observed for maturity with an estimate value of 0.78 followed by NP, RNP, NRNP, and PP, with estimate values of 0.34, 0.33, 0.31, and 0.30, respectively (Fig. 2). The lowest heritability was estimated for yield with an estimate value of 0.24 (Fig. 2). Soybean seed yield and PP showed the highest variability across the environments (Fig. 2).

The linear correlations among all the measured traits were estimated using the coefficients of correlation (*r*). Based on the results (Fig. 3), all traits were positively correlated with each other, except NRNP that was negatively associated with yield, maturity, RNP, NP, and PP. NP showed the highest correlation with the RNP (r= 0.97) and NRNP (r= -0.63). RNP had the highest correlation (r =0.86) with yield among all the tested yield components (Fig. 3).

Genotyping evaluations

For the tested GWAS panel, high-quality SNPs were obtained from 210M single-end Ion Torrent reads that were proceeded with Fast-GBS.v2. From 40,712 SNPs, 17,958 SNPs were polymorphic and mapped to 20 soybean chromosomes. The minimum number of 403 SNPs were mapped on 403 and 1780 on chromosomes 11 and the maximum number of 1780 were mapped on chromosome 18. Overall, the average number of SNPs across all the 20 chromosomes was 898, with the mean density of one SNP for every 0.12 cM across the whole genome.

Population structure and kinship

The structure profile for the tested population is presented in Fig. 4. The result of genotypic evaluations suggested that the tested GWAS panel was composed of four to seven subpopulations. Therefore, we chose to conduct the structure analysis using K=7 as the appropriate K for the structure profile of the tested GWAS panel (Fig. 4). In order to reduce the confounding, the kinship was also estimated between genotypes of the GWAS panel.

GWAS analysis

The average value for soybean maturity in the tested GWAS panel was 106 days with a standard deviation of 5 days (Fig. 3). Association analysis by the MLM method identified nine associated SNP markers located on chromosomes 2 and 19 (Fig. 5a). Using FarmCPU, a total of nine associated SNP markers were located on chromosomes 2, 19, and 20 (Fig. 5a). By using the RF method, the total of three SNP markers on chromosomes 3, 7, and 16 were associated with the soybean maturity, whereas SVR-mediated GWAS detected 10 associated SNP markers located on chromosomes 2, 6, 10, 16, 19, and 20 (Fig. 5a). SVR-mediated GWAS detected five QTL directly related to the reproductive period and R8 full maturity (Table 1).

The average soybean seed yield in the GWAS panel was 3.5 t ha^{-1} with a standard deviation of 0.45 t ha^{-1} (Fig. 3). Using MLM, FarmCPU, RF, and SVR approach, we identified two, three, five, and 18 SNP markers associated with the yield, respectively (Fig. 5b). The SNP markers identified by MLM and FarmCPU were located on chromosomes 6 and 8. Using the RF-mediated GWAS method, associated SNP markers were located on chromosomes 4, 7, 12, and 17. By using the SVR-mediated GWAS method, the SNP markers were identified on chromosomes 3, 4, 6, 7, 15, 19, and 20 (Fig. 5b). In SVR-mediated GWAS, the identified QTL were co-localized with eight previously reported yield-related QTL such as seed yield, seed weight, and seed set (Table 2). However, other tested GWAS methods was not able to find QTLs that were co-localized with any previously reported QTL associated with seed yield (Table 2).

The average NP in the tested GWAS panel was 15.21 nodes with a standard deviation of 0.77 nodes (Fig. 3). By using the MLM and FarmCPU methods, one and two associated SNP markers were detected, respectively (Fig. 6a). Four and ten SNP markers associated with NP were detected using RF and SVR methods, respectively. SVR-mediated GWAS was the only method that identified three previously reported NP-related QTL (Table 3). The average NRNP was 3.33 nodes with a standard deviation of 0.28 nodes (Fig. 3). A total of two, three, five, and ten SNP markers were detected associated with NRNP using the MLM, FarmCPU, RF, and SVR methods, respectively (Fig. 6b). The detected SNP markers using the SVR method were located on chromosomes 4, 7, 18, 19, and 20, whereas SNP markers identified through RF were located on chromosomes 1, 4, 7, 18, and 19 (Fig. 6b). Chromosomes number 4, 8, and 15 were identified as carrying SNP markers with NRNP using FarmCPU and the MLM method identified SNP markers located on chromosomes 8 and 15. Most of the detected QTL using all GWAS methods co-localized with previously reported QTL related to seed weight, seed protein, water use efficiency, first flower, and soybean cyst nematode (Table 4).

The average RNP was 11.89 nodes with a standard deviation of 0.98 nodes (Fig. 3). Based on the results of MLM and FarmCPU methods, four associated SNP markers with RNP were located on chromosomes 8 and 19. Using the RF method, four associated SNP markers were identified on chromosomes 8, 9, 15, and 20. Using the SVR method, 11 SNP markers were detected associated with RNP located on chromosomes 4, 7, 8, 15, 18, 19, and 20 (Fig. 7a). Regardless of the type of GWAS methods used in this study, we found SNP markers associated with the trait on chromosome 8. The position of the associated SNP marker on chromosome 8 was identical using all GWAS methods (~ 450 Kbp). The list of detected QTLs for RNP is also presented in Table 5.

The average value for PP in the tested GWAS panel was 45.02 pods with a standard deviation of 8.54 pods. We did not detect any SNP marker associated with PP using the MLM or FarmCPU methods. However, by using the RF method, four SNP markers located on chromosomes 7, 10, 18, and 20 were found to be associated with PP (Fig. 7b). Twelve SNP markers were detected associated with PP using SVR. The markers were located on chromosomes 6, 9, 10, 11, 15, 18, and 19 (Fig. 7b). The associated SNP markers in chromosome 10 were found both in RF and SVR with 4.6 cM distance far from each other. In PP, MLM and FarmCPU did not detect any related QTL for this trait, while SVR-mediated GWAS was identified seven previously reported QTL directly related to the pod number (Table 6).

Extracting candidate genes undelaying detected QTLs

According to the flanking regions of each QTL, which was determined using LD decay distance, 150-kbp upstream and downstream of each SNP's peak were considered to identify potential candidate genes (Fig. 1). Candidate genes were extracted within the flanking region of each QTL with high allelic effects using the gene annotation, enrichment tools and the information from previous studies (Table S1). For date to maturity, three peak SNPs (Chr2_695362, Chr2_720134, and Chr19_47513536) had the highest allelic effects than other detected peak SNPs (Fig. 8a). On the basis of the gene annotation and expression within the QTL, *Glyma.02g006500* (GO:0015996) and *Glyma.19g224200* (GO:0010201), which respectively encode chlorophyll catabolic process and phytochrome A (PHYA) related genes, were identified as the strong candidate genes for maturity. *Glyma.02g006500* (GO:0015996) was exactly detected in the peak SNP position of Chr2_695362, whereas *Glyma.19g224200* (GO:0010201) was 119 Kbp far from the detected peak SNP at Chr19_47513536. The yield-related QTL with the peak SNP positioned on Chr7_1032587 had the highest allelic effect in comparison with other detected peak SNPs (Fig. 8b). Within a 77 Kbp away from the detected peak SNP (Chr7_1032587), *Glyma.07G014100* (GO:0010817), which encodes the regulation of hormone levels, was identified as the strongest candidate genes in yield. For the NP trait, two peak SNPs, Chr7_1032587 and Chr7_1092403, had the highest allelic effects among all the detected peak SNPs (Fig. 8c). In this study, the Chr7_1032587 SNP was found to be associated with yield, NP, and NRNP. The *Glyma.07G205500* (GO:0009693) and *Glyma.08G065300* (GO:0042546) genes, which encode UBP1-associated protein 2C and cell wall biogenesis, respectively, were candidate as plausible genes influencing both NP and NRNP. Both detected gene candidates were exactly collocated at the corresponding peak SNPs at Chr7_1032587

and Chr8_5005929 (Fig. 8d). Regarding peak SNPs associated with RNP, the highest allelic effects were found in peak SNPs of Chr9_40285014 and Chr15_34958361 (Fig. 8e). The *Glyma.15G214600* (GO:0009920) and *Glyma.15G214700* (GO:0009910) genes, which encode cell plate formation involved in plant-type cell wall biogenesis and acetyl-CoA biosynthetic process, respectively, were nominated as strong candidate genes governing NRNP. *Glyma.15G214600* (GO:0009920) and *Glyma.15G214700* (GO:0009910) were 127 and 90 Kbp far from the peak SNP at Chr15_3495836, respectively. For the PP trait, the highest allelic effects were found in peak SNPs at Chr7_15331676, Chr11_5245870, and Chr18_55469601 (Fig. 8f). The *Glyma.07G128100* (GO:0009909) gene, which encodes the regulation of flower development, was the strongest candidate genes that can affect PP. *Glyma.07G128100* (GO:0009909) is located in the peak SNP position, Chr7_15331676.

Discussion

One of the objectives of this study was to attain a better understanding of the roles of soybean yield component traits in the production of total seed yield and how these traits can be used for facilitating the development of high-yielding soybeans with improved genetic gains. The genetic dissection of soybean yield components and establishing genetic and genomics toolkits that can be used for either designing crosses or screening large breeding populations for selecting genotypes with improved yield components will facilitate the improvement of yield genetic gains in new cultivars (Cooper et al. 2009; Hu et al. 2020; Xavier and Rainey 2020). For this aim, a wide range of analyses, including Pearson correlation, normality and distribution plots, GWAS both in combined and separate environments, and functional annotation of candidate genes and QTL, were performed in this study. The collective evaluation of the mentioned analyses contributed to building the wide perspectives of the genetic architecture of the soybean yield component traits. While high phenotypic variations were observed for yield and PP in the panel, date to maturity and NP had the lowest phenotypic variations across the tested environments. These findings are in line with the results of previous research studies on yield component traits (Kahlon and Board 2012; Xavier and Rainey 2020), in which they found higher variation for the total seed yield and total pods per plant. The heritability and correlation analyses showed that NP had the highest heritability and significant linear correlations with RNP and PP. Also, PP had the highest correlation with yield among all the tested soybean yield components. The number of nodes and pods in soybean are known as the two of the key soybean yield components that play important roles in determining the final soybean seed yield (Herbert and Litchfield 1982; Kahlon and Board 2012; Xavier and Rainey 2020). Previous studies reported low heritability rates for soybean yield components, especially NP and PP (KUSWANTORO 2017; Sulistyo and Sari 2018; Xavier et al. 2016a; Xavier and Rainey 2020). These traits are significantly affected by environmental factors (Price and Schluter 1991). Although for a given trait heritability indicates the strength of the relationship between phenotype and genetic variability, it does not necessarily indicate the value of the trait for genetic studies (Cassell 2009). Different low heritable traits are highly correlated with significant economic traits (Cassell 2009). In soybean, for example, yield can be considered as the most important economic trait that is highly determined by its component traits.

GWAS is known as one of the most important genetic toolkits for detecting QTL associated with quantitate traits (Kaler et al. 2020). There are several statistical methods implemented in GWAS for improving the detection of associated SNP markers with the trait of interest. While conventional GWAS are appropriate approaches for detecting SNP markers with large effects on complex traits, they are, however, underpowered for the simultaneous consideration of a wide range of interconnected biological processes and mechanisms that shape the phenotype of complex traits (Lee et al. 2020). Therefore, using variable importance values in ML algorithms for identifying SNP-trait associations may improve the power of GWAS for discovering variant-trait association with higher resolution (Szymczak et al. 2009). The variable importance methods based on linear and logistic regressions, support vector machines, and random forests are well established in the literature (Grömping 2009; Williamson et al. 2020; Wu and Liu 2009; Yoosefzadeh-Najafabadi et al. 2021a). Among all the tested GWAS methods in this study, SVR-mediated GWAS was the best method to detect SNP markers with high allelic effects associated with the tested traits. The advantage of SVR-mediated GWAS, over conventional GWAS models, can be explained by the presence of a nonlinear relationship between input and output variables, which is used to build an algorithm with accurate prediction ability (Kaneko 2020). Therefore, reliable genomic regions can be discovered by SVR-mediated GWAS because of its ability to consider the interaction effects between SNPs rather than *p*-values for individual SNP-trait GWAS tests.

In this study, none of the QTL identified using MLM, FarmCPU, and RF was reported to be associated directly with soybean maturity. However, using SVR-mediated GWAS, five QTL were detected on chromosomes 16 and 19 specifically related to the soybean maturity. All the detected QTL were previously reported by Sonah et al. (2015) and Zhang et al. (2015b) in separate studies as it presented in table 1. Also, the peak SNP position of Chr19_47513536 detected by SVR-mediated GWAS had the highest allelic effect among all the detected SNPs in soybean maturity, which is in line with Sonah et al. (2015). For soybean seed yield, five QTL discovered by SVR-mediated GWAS were reported previously (Copley et al. 2018; Hu et al. 2014), while none of the detected QTL from other tested GWAS methods was previously reported for this trait. There was no previous study on the genetic structure of NRNP and RNP, therefore, all the identified QTL in this study are reported for the first time. For PP, conventional GWAS methods were not able to detect any QTL. However, using SVR-mediated GWAS, a total of seven QTL were distinguished to be related to pod numbers aligned with previous studies (Zhang et al. 2015a). It is necessary to be mentioned that the average allelic effects of the QTL presented in this study, Fig. 8, was estimated using the equation develop by Pimentel et al. (2015) but not directly by the GWAS methods. The RF and SVR-mediated GWAS methods do not specifically provide an allele effect therefore, the aim of this study was mostly focused on detecting the associated genes and QTL underlying the soybean yield, maturity, and yield components.

The results of candidate gene identifications within identified QTL using SVR-mediated GWAS analyses reveled important information. For example, from all the detected genes using SVR-mediated GWAS for maturity, candidate gene *Glyma.02g006500* (GO:0015996) is a protein ABC transporter 1, that is annotated as a chlorophyll catabolic process and located exactly in the peak SNP position at Chr02_695362. ATP-binding cassette (ABC) transporter genes play conspicuous roles in different plant

growth and developmental stages by transporting different phytochemicals across endoplasmic reticulum (ER) membranes (Hwang et al. 2016). Because of the central roles of ABC transporters in transporting biomolecules such as phytohormones, metabolites, and lipids, they play important roles in plant growth and development as well as maturity (Block and Jouhet 2015; Hwang et al. 2016). Moreover, recent studies revealed that ER uses fatty acid building blocks made in the chloroplast to synthesize Triacylglycerol (TAG). Therefore, ABC transporter genes are important for the normal accumulation of Triacylglycerol (TAG) during the seed-filling stage and maturity (Block and Jouhet 2015; Kim et al. 2013). Additionally, *Glyma.19g224200* (GO:0010201) in E3 locus, which was previously discovered by Buzzell (1971) and molecularly characterized as a phytochrome A (PHYA) gene (Watanabe et al. 2009), was detected through the SVR-mediated GWAS. Phytochromes, through PHYTOCHROME INTERACTING FACTOR (PIF), regulate the expression of some specific genes encoding rate-limiting catalytic enzymes of different plant growth regulators (e.g., abscisic acid, gibberellins, auxin) and, therefore, play crucial roles in plant maturity (Legris et al. 2019). In addition, PHYB is inactivated after imbibition shade signals, which repress PHYA-dependent signaling in the embryo that results in the maturity of seeds by preventing germination (Casal 2013; De Wit et al. 2016). This is obtained by regulating the balance between abscisic acid and gibberellin. Subsequently, abscisic acid transports from the endosperm to the embryo by ABC transporter (De Wit et al. 2016).

Regarding NRNP, candidate gene *Glyma.07G205500* (GO:0009693- UBP1-associated protein 2C) that annotated as ethylene biosynthetic process was located exactly at the peak SNP position at *Chr7_37469678*, was detected by SVR-mediated GWAS. An interaction screen with the heterogeneous nuclear ribonucleoprotein (hnRNP) results in the production of oligouridylatebinding protein 1 (UBP1)-associated protein (Lambermon et al. 2002). It has been well documented that this protein plays important roles in several physiological processes such as responses to abiotic stresses (Li et al. 2002), leaf senescence(Kim et al. 2008), floral development (Streitner et al. 2008), and chromatin modification (Liu et al. 2007). In addition, previous studies showed that the production of productive or non-reproductive nodes is completely accompanied by the upregulation or downregulation of this protein (Bäurle and Dean 2008; Na et al. 2015). In addition, *Glyma.08G065300* (GO:0042546- MADS-box transcription factor) that is associated with cell wall biogenesis, was located in the SNP position of Chr8_5005929. The genes of the MADS-box family can be considered as the main regulators for cell differentiation and organ determination (Lee et al. 2013). The floral organ recognition MADS-box family has been categorized into A, B, C, D, and E classes. Among these classes, class E was shown to be associated with reproductive organ development (Hussin et al. 2021). Indeed, activation or repression of this transcription factor leads to the development of nodes to productive or non-productive nodes (Ditta et al. 2004; Gao et al. 2010; Liu et al. 2013).

Gene expression data provided by Severin et al. (2010) noted that 20 candidate genes for PP that were detected using the SVR-mediated GWAS were expressed in flowers, 1 cm pod (7 DAF), pod shell (10-13 DAF), pod shell (14-17 DAF) and seeds. In PP, most of the genes detected by SVR-mediated GWAS are associated with auxin influx carrier or auxin response factors (ARFs), gibberellin synthesis, and response to brassinosteroid (Lin et al. 2020; Yin et al. 2018). Song et al. (2020) and Li et al. (2018a) also reported

that some genes related to PP were associated with embryo development, stamen development, ovule development, cytokinin biosynthesis, and response gibberellin that we also identified in our study. Soybean seed yield significantly depends on seed number and seed size (Liu et al. 2010; Rotundo et al. 2009). These two factors are determined from fertilization to seed maturity. Therefore, soybean seed development can be divided into three stages or phases: pre-embryo or seed set, embryo growth or seed growth, and desiccation stages or seed maturation phases (Ruan et al. 2012; Weber et al. 2005). In Arabidopsis, a complex signaling pathway and regulatory networks, including sugar and hormonal signaling, transcription factors, and metabolic pathway, have been reported to be involved in seed development (Le et al. 2010; Orozco-Arroyo et al. 2015). Several key genes and transcription factors (e.g., LEAFY COTYLEDON 1 (LEC1), LEC2, FUSCA3 (FUS3), AGAMOUS-LIKE15 (AGL15), ABSCISIC ACID INSENSITIVE 3 (ABI3), YUCCA10 (YUC10), ARFs) have been determined to control several downstream plant growth regulators pathways to the seed development (Lepiniec et al. 2018; Pelletier et al. 2017; Sun et al. 2010). Indeed, a high ratio of abscisic acid to gibberellic acid can regulate seed development (Figueiredo and Köhler 2018; Wang et al. 2016). The downregulation of FUS3 obtains this through repressing GA3ox1 and GA3ox2 and activating ABA biosynthesis (Weber et al. 2005). In soybean, RNA seq analysis for the seed set, embryo growth, and early maturation stages of developing seeds in two soybeans with contrasting seed size showed cell division and growth genes, hormone regulation, transcription factors, and metabolic pathway are involved in seed size and numbers (Du et al. 2017).

Conclusion

A better understanding of the genetic architecture of the yield component traits in soybean may enable breeders to establish more efficient selection strategies for developing high-yielding cultivars with improved genetic gains. Major yield component traits such as maturity, NP, NRNP, RNP, and PP play important roles in determining the overall yield production in soybean. This study verified the importance of those traits, using correlation and distribution analyses, in determining the total soybean seed yield. Furthermore, by testing different conventional and ML-mediated GWAS methods, this study demonstrated the potential benefit of using ML-mediated methods in GWAS. SVR-mediated GWAS outperformed all the other methods tested in this study, and therefore, it is recommended as an alternative to conventional GWAS methods with a greater power for detecting genomic regions associated with complex traits such as yield and its components in soybean, and possibly other crop species. To the best of our knowledge, this study is the first endeavor in which SVR was used for GWAS analyses in plants. In order to verify the causal relationship between identified QTL and the target phenotypic traits, we identified candidate genes within each QTL region using gene annotation procedures and information. The results demonstrated the efficiency of SVR-mediated GWAS in detecting reliable QTL that can be used for marker-assisted selections. Nevertheless, further investigation is recommended to confirm the efficiency of SVR-mediated GWAS for discovering genomic regions associated with complex traits in other plant species.

Declarations

Acknowledgments

The authors are grateful to the past and current members of Eskandari laboratory at the University of Guelph, Ridgetown, Bryan Stirling, John Kobler, and Robert Brandt for their technical support. We would like to thank Maryam Vazin and Mohsen Hesami for their assistance with the field data collection and reviewing the manuscript, respectively. The preprint of this manuscript previously deposited in bioRxiv as a non-commercial pre-print server with the doi: https://doi.org/10.1101/2021.06.24.449776.

Author Contribution

ME conceptualized, designed and directed the experiments. MY-N performed the experiments, modeled, summed up, and wrote the manuscript. ST participated in candidate gene analyses; DT, DTOR, ST, IR, and ME revised the manuscript and validated the results. All authors have read and approved the final manuscript.

Funding

This project was funded in part by the Grain Farmers of Ontario (GFO) and SeCan. The funding bodies did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Availability Statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

References

Auria L, Moro RA (2008) Support vector machines (SVM) as a technique for solvency analysis

Awad M, Khanna R (2015) Support vector regression. Efficient learning machines. Springer, pp 67-80

Bao Y, Kurle JE, Anderson G, Young ND (2015) Association mapping and genomic prediction for resistance to sudden death syndrome in early maturing soybean germplasm. Molecular Breeding 35:1-14

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:14065823

Bates D, Sarkar D, Bates MD, Matrix L (2007) The lme4 package. R package version 2:74

Bäurle I, Dean C (2008) Differential interactions of the autonomous pathway RRM proteins and chromatin regulators in the silencing of Arabidopsis targets. PLoS One 3:e2733

Belayneh A, Adamowski J, Khalil B, Ozga-Zielinski B (2014) Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. Journal of Hydrology 508:418-429

Benjamini Y, 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) 57:289-300

Block MA, Jouhet J (2015) Lipid trafficking at endoplasmic reticulum–chloroplast membrane contact sites. Current opinion in cell biology 35:21-29

Botta V, Louppe G, Geurts P, Wehenkel L (2014) Exploiting SNP correlations within random forest for genome-wide association studies. PloS one 9:e93379

Bowley S (1999) A hitchhiker's guide to statistics in plant biology. Guelph, Ont.: Any Old Subject Books

Breiman L (2001) Random forests. Machine learning 45:5-32

Buzzell R (1971) Inheritance of a soybean flowering response to fluorescent-daylength conditions. Canadian Journal of Genetics and Cytology 13:703-707

Carter A, Rajcan I, Woodrow L, Navabi A, Eskandari M (2018) Genotype, environment, and genotype by environment interaction for seed isoflavone concentration in soybean grown in soybean cyst nematode infested and non-infested environments. Field Crops Research 216:189-196

Casal JJ (2013) Photoreceptor signaling networks in plant responses to shade. Annual review of plant biology 64:403-427

Cassell BG (2009) Using heritability for genetic improvement

Chang H-X, Hartman GL (2017) Characterization of insect resistance loci in the USDA soybean germplasm collection using genome-wide association studies. Frontiers in plant science 8:670

Che Z, Liu H, Yi F, Cheng H, Yang Y, Wang L, Du J, Zhang P, Wang J, Yu D (2017) Genome-Wide Association Study Reveals Novel Loci for SC7 Resistance in a Soybean Mutant Panel. Frontiers in Plant Science 8

Chen JH, Verghese A (2020) Planning for the Known Unknown: Machine Learning for Human Healthcare Systems. The American Journal of Bioethics 20:1-3

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963-971

Contreras-Soto RI, Mora F, de Oliveira MAR, Higashi W, Scapim CA, Schuster I (2017) A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. PloS one 12:e0171105

Cook DE, Bayless AM, Wang K, Guo X, Song Q, Jiang J, Bent AF (2014) Distinct copy number, coding sequence, and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode. Plant physiology 165:630-647

Cooper M, van Eeuwijk FA, Hammer GL, Podlich DW, Messina C (2009) Modeling QTL for complex traits: detection and context for plant breeding. Current opinion in plant biology 12:231-240

Copley TR, Duceppe M-O, O'Donoughue LS (2018) Identification of novel loci associated with maturity and yield traits in early maturity soybean plant introduction lines. BMC genomics 19:1-12

Cortes C, Vapnik V (1995) Support vector machine. Machine learning 20:273-297

De Wit M, Galvão VC, Fankhauser C (2016) Light-mediated hormonal regulation of plant growth and development. Annual review of plant biology 67:513-537

Denton SM, Salleb-Aouissi A (2020) A Weighted Solution to SVM Actionability and Interpretability. arXiv preprint arXiv:201203372

Dhanapal AP, Ray JD, Singh SK, Hoyos-Villegas V, Smith JR, Purcell LC, King CA, Cregan PB, Song Q, Fritschi FB (2015a) Genome-wide association study (GWAS) of carbon isotope ratio (δ 13 C) in diverse soybean [Glycine max (L.) Merr.] genotypes. Theoretical and Applied Genetics 128:73-91

Dhanapal AP, Ray JD, Singh SK, Hoyos-Villegas V, Smith JR, Purcell LC, King CA, Fritschi FB (2015b) Association mapping of total carotenoids in diverse soybean genotypes based on leaf extracts and high-throughput canopy spectral reflectance measurements. PloS one 10:e0137213

Dhanapal AP, Ray JD, Smith JR, Purcell LC, Fritschi FB (2018) Identification of Novel Genomic Loci Associated with Soybean Shoot Tissue Macro and Micronutrient Concentrations. The plant genome 11:170066

Ditta G, Pinyopich A, Robles P, Pelaz S, Yanofsky MF (2004) The SEP4 gene of Arabidopsis thaliana functions in floral organ and meristem identity. Current biology 14:1935-1940

Doerge RW, Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. Genetics 142:285-294

Du J, Wang S, He C, Zhou B, Ruan Y-L, Shou H (2017) Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. Journal of Experimental Botany 68:1955-1972

Duan K-B, Rajapakse JC, Wang H, Azuaje F (2005) Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE transactions on nanobioscience 4:228-234

Fang C, Ma Y, Wu S, Liu Z, Wang Z, Yang R, Hu G, Zhou Z, Yu H, Zhang M (2017) Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome biology 18:1-14

Fehr W, Caviness C (1971) Burmood DT, Penington J S. Development description of soybean, Glycine max (L) Mer. Crop Science 11:929-931

Figueiredo DD, Köhler C (2018) Auxin: a molecular trigger of seed development. Genes & development 32:479-490

Fletcher T (2009) Support vector machines explained. Tutorial paper, Mar:28

Gao X, Liang W, Yin C, Ji S, Wang H, Su X, Guo C, Kong H, Xue H, Zhang D (2010) The SEPALLATA-like gene OsMADS34 is required for rice inflorescence and spikelet development. Plant Physiology 153:728-740

Goldberger AS (1962) Best linear unbiased prediction in the generalized linear regression model. Journal of the American Statistical Association 57:369-375

Grömping U (2009) Variable importance assessment in regression: linear regression versus random forest. The American Statistician 63:308-319

Herbert S, Litchfield G (1982) Partitioning Soybean Seed Yield Components 1. Crop science 22:1074-1079

Hesami M, Jones AMP (2020) Application of artificial intelligence models and optimization algorithms in plant cell and tissue culture. Applied Microbiology and Biotechnology:1-37

Hesami M, Naderi R, Tohidfar M, Yoosefzadeh-Najafabadi M (2020) Development of support vector machine-based model and comparative analysis with artificial neural network for modeling the plant tissue culture procedures: effect of plant growth regulators on somatic embryogenesis of chrysanthemum, as a case study. Plant Methods 16:1-15

Hu D, Zhang H, Du Q, Hu Z, Yang Z, Li X, Wang J, Huang F, Yu D, Wang H (2020) Genetic dissection of yield-related traits via genome-wide association analysis across multiple environments in wild soybean (Glycine soja Sieb. and Zucc.). Planta 251:39

Hu Z, Zhang D, Zhang G, Kan G, Hong D, Yu D (2014) Association mapping of yield-related traits and SSR markers in wild soybean (Glycine soja Sieb. and Zucc.). Breeding science 63:441-449

Hussin SH, Wang H, Tang S, Zhi H, Tang C, Zhang W, Jia G, Diao X (2021) SiMADS34, an E-class MADS-box transcription factor, regulates inflorescence architecture and grain yield in Setaria italica. Plant

Molecular Biology 105:419-434

Hwang J-U, Song W-Y, Hong D, Ko D, Yamaoka Y, Jang S, Yim S, Lee E, Khare D, Kim K (2016) Plant ABC transporters enable many unique aspects of a terrestrial plant's lifestyle. Molecular plant 9:338-355

Jamil IN, Remali J, Azizan KA, Muhammad NAN, Arita M, Goh H-H, Aizat WM (2020) Systematic Multi-Omics Integration (MOI) Approach in Plant Systems Biology. Frontiers in Plant Science 11

Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. Science 349:255-260

Kahlon CS, Board JE (2012) Growth dynamic factors explaining yield improvement in new versus old soybean cultivars. Journal of crop improvement 26:282-299

Kahlon CS, Board JE, Kang MS (2011) An analysis of yield component changes for new vs. old soybean cultivars. Agronomy Journal 103:13-22

Kaler AS, Dhanapal AP, Ray JD, King CA, Fritschi FB, Purcell LC (2017) Genome-wide association mapping of carbon isotope and oxygen isotope ratios in diverse soybean genotypes. Crop Science 57:3085-3100

Kaler AS, Gillman JD, Beissinger T, Purcell LC (2020) Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. Frontiers in plant science 10:1794

Kan G, Zhang W, Yang W, Ma D, Zhang D, Hao D, Hu Z, Yu D (2015) Association mapping of soybean seed germination under salt stress. Molecular Genetics and Genomics 290:2147-2162

Kaneko H (2020) Support vector regression that takes into consideration the importance of explanatory variables. Journal of Chemometrics:e3327

Katsileros A, Drosou K, Koukouvinos C (2015) Evaluation of nearest neighbor methods in wheat genotype experiments. Communications in Biometry and Crop Science 10:115-123

Kim CY, Bove J, Assmann SM (2008) Overexpression of wound-responsive RNA-binding proteins induces leaf senescence and hypersensitive-like cell death. New Phytologist 180:57-70

Kim GB, Kim WJ, Kim HU, Lee SY (2020) Machine learning applications in systems metabolic engineering. Current opinion in biotechnology 64:1-9

Kim S, Yamaoka Y, Ono H, Kim H, Shim D, Maeshima M, Martinoia E, Cahoon EB, Nishida I, Lee Y (2013) AtABCA9 transporter supplies fatty acids for lipid synthesis to the endoplasmic reticulum. Proceedings of the national academy of sciences 110:773-778

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Team RC (2020) Package 'caret'. The R Journal

KUSWANTORO H (2017) Genetic variability and heritability of acid-adaptive soybean promising lines. Biodiversitas Journal of Biological Diversity 18

Lambermon MH, Fu Y, Kirk DAW, Dupasquier M, Filipowicz W, Lorković ZJ (2002) UBA1 and UBA2, two proteins that interact with UBP1, a multifunctional effector of pre-mRNA maturation in plants. Molecular and Cellular Biology 22:4346-4357

Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S (2010) Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proceedings of the National Academy of Sciences 107:8063-8070

Leamy LJ, Zhang H, Li C, Chen CY, Song B-H (2017) A genome-wide association study of seed composition traits in wild soybean (Glycine soja). BMC genomics 18:1-15

Lee JH, Ryu H-S, Chung KS, Posé D, Kim S, Schmid M, Ahn JH (2013) Regulation of temperature-responsive flowering by MADS-box transcription factor repressors. Science 342:628-632

Lee S, Liang X, Woods M, Reiner AS, Concannon P, Bernstein L, Lynch CF, Boice JD, Deasy JO, Bernstein JL (2020) Machine learning on genome-wide association studies to predict the risk of radiation-associated contralateral breast cancer in the WECARE Study. PloS one 15:e0226157

Legris M, Ince YÇ, Fankhauser C (2019) Molecular mechanisms underlying phytochrome-controlled morphogenesis in plants. Nature communications 10:1-15

Lepiniec L, Devic M, Roscoe T, Bouyer D, Zhou D-X, Boulard C, Baud S, Dubreucq B (2018) Molecular and epigenetic regulations and functions of the LAFL transcriptional regulators that control seed development. Plant reproduction 31:291-307

Li C, Zou J, Jiang H, Yu J, Huang S, Wang X, Liu C, Guo T, Zhu R, Wu X (2018a) Identification and validation of number of pod-and seed-related traits QTL s in soybean. Plant Breeding 137:730-745

Li J, Kinoshita T, Pandey S, Ng CK-Y, Gygi SP, Shimazaki K-i, Assmann SM (2002) Modulation of an RNA-binding protein by abscisic-acid-activated protein kinase. Nature 418:793-797

Li X, Tian R, Kamala S, Du H, Li W, Kong Y, Zhang C (2018b) Identification and verification of pleiotropic QTL controlling multiple amino acid contents in soybean seed. Euphytica 214:1-14

Li Y-h, Reif JC, Ma Y-s, Hong H-l, Liu Z-x, Chang R-z, Qiu L-j (2015) Targeted association mapping demonstrating the complex molecular genetics of fatty acid formation in soybean. BMC genomics 16:1-13

Li Yh, Shi Xh, Li Hh, Reif JC, Wang Jj, Liu Zx, He S, Yu Bs, Qiu Lj (2016) Dissecting the genetic basis of resistance to soybean cyst nematode combining linkage and association mapping. The plant genome 9:plantgenome2015.2004.0020

Lin F, Wani SH, Collins PJ, Wen Z, Li W, Zhang N, McCoy AG, Bi Y, Tan R, Zhang S (2020) QTL mapping and GWAS for identification of loci conferring partial resistance to Pythium sylvaticum in soybean (Glycine max (L.) Merr). Molecular Breeding 40:1-11

Lipka AE, Kandianis CB, Hudson ME, Yu J, Drnevich J, Bradbury PJ, Gore MA (2015) From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Current Opinion in Plant Biology 24:110-118

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397-2399

Liu B, Liu X, Wang C, Li Y, Jin J, Herbert S (2010) Soybean yield and yield component distribution across the main axis in response to light enrichment and shading under different densities. Plant, Soil and Environment 56:384-392

Liu C, Teo ZWN, Bi Y, Song S, Xi W, Yang X, Yin Z, Yu H (2013) A conserved genetic pathway determines inflorescence architecture in Arabidopsis and rice. Developmental cell 24:612-622

Liu F, Quesada V, Crevillén P, Bäurle I, Swiezewski S, Dean C (2007) The Arabidopsis RNA-binding protein FCA requires a lysine-specific demethylase 1 homolog to downregulate FLC. Molecular cell 28:398-407

Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS genetics 12:e1005767

Ma B, Dwyer LM, Costa C, Cober ER, Morrison MJ (2001) Early prediction of soybean yield from canopy reflectance measurements. Agronomy Journal 93:1227-1234

Mangena P (2020) Phytocystatins and their Potential Application in the Development of Drought Tolerance Plants in Soybeans (Glycine max L.). Protein and Peptide Letters 27:135-144

Mao T, Li J, Wen Z, Wu T, Wu C, Sun S, Jiang B, Hou W, Li W, Song Q (2017) Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. BMC genomics 18:1-17

Meinshausen N (2006) Quantile regression forests. Journal of Machine Learning Research 7:983-999

Moellers TC, Singh A, Zhang J, Brungardt J, Kabbage M, Mueller DS, Grau CR, Ranjan A, Smith DL, Chowda-Reddy R (2017) Main and epistatic loci studies in soybean for Sclerotinia sclerotiorum resistance reveal multiple modes of resistance in multi-environments. Scientific reports 7:1-13

Mohammadi M, Xavier A, Beckett T, Beyer S, Chen L, Chikssa H, Cross V, Moreira FF, French E, Gaire R (2020) Identification, Deployment, and Transferability of Quantitative Trait Loci from Genome-Wide Association Studies in Plants. Current Plant Biology:100145

Na J-K, Kim J-K, Kim D-Y, Assmann SM (2015) Expression of potato RNA-binding proteins StUBA2a/b and StUBA2c induces hypersensitive-like cell death and early leaf senescence in Arabidopsis. Journal of experimental botany 66:4023-4033

Nico M, Miralles DJ, Kantolic AG (2019) Natural post-flowering photoperiod and photoperiod sensitivity: Roles in yield-determining processes in soybean. Field Crops Research 231:141-152

Ogutu JO, Piepho H-P, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. BMC proceedings. Springer, p S11

Orozco-Arroyo G, Paolo D, Ezquer I, Colombo L (2015) Networks controlling seed size in Arabidopsis. Plant reproduction 28:17-32

Pasaniuc B, Price AL (2017) Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics 18:117-127

Pedersen P, Lauer JG (2004) Response of soybean yield components to management system and planting date. Agronomy Journal 96:1372-1381

Pelletier JM, Kwong RW, Park S, Le BH, Baden R, Cagliari A, Hashimoto M, Munoz MD, Fischer RL, Goldberg RB (2017) LEC1 sequentially regulates the transcription of genes involved in diverse developmental processes during seed development. Proceedings of the National Academy of Sciences 114:E6710-E6719

Pimentel E, Edel C, Emmerling R, Götz K-U (2015) How imputation errors bias genomic predictions. Journal of dairy science 98:4131-4138

Price T, Schluter D (1991) On the low heritability of life-history traits. Evolution 45:853-861

Priolli RHG, Campos J, Stabellini N, Pinheiro J, Vello N (2015) Association mapping of oil content and fatty acid components in soybean. Euphytica 203:83-96

Qin J, Song Q, Shi A, Li S, Zhang M, Zhang B (2017) Genome-wide association mapping of resistance to Phytophthora sojae in a soybean [Glycine max (L.) Merr.] germplasm panel from maturity groups IV and V. PLOS ONE 12:e0184613

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197:573-589

Ray JD, Dhanapal AP, Singh SK, Hoyos-Villegas V, Smith JR, Purcell LC, King CA, Boykin D, Cregan PB, Song Q (2015) Genome-wide association study of ureide concentration in diverse maturity group IV soybean [Glycine max (L.) Merr.] accessions. G3: Genes, Genomes, Genetics 5:2391-2403

Rębilas K, Klimek-Kopyra A, Bacior M, Zając T (2020) A model for the yield losses estimation in an early soybean (Glycine max (L.) Merr.) cultivar depending on the cutting height at harvest. Field Crops Research 254:107846

Reynolds M (2001) Application of physiology in wheat breeding. Cimmyt

Richards R (1982) Breeding and selecting for drought resistant wheat. p. 303–316. Drought resistance in crops with emphasis on rice. IRRI, Manila, Philippines. Breeding and selecting for drought resistant wheat p 303–316 In Drought resistance in crops with emphasis on rice IRRI, Manila, Philippines:-

Robinson AP, Conley SP, Volenec JJ, Santini JB (2009) Analysis of high yielding, early-planted soybean in Indiana. Agronomy Journal 101:131-139

Rotundo JL, Borrás L, Westgate ME, Orf JH (2009) Relationship between assimilate supply per seed during seed filling and soybean seed composition. Field crops research 112:90-96

Ruan Y-L, Patrick JW, Bouzayen M, Osorio S, Fernie AR (2012) Molecular regulation of seed and fruit set. Trends in plant science 17:656-665

Severin AJ, Woody JL, Bolon Y-T, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE (2010) RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. BMC plant biology 10:1-16

Siegmann B, Jarmer T (2015) Comparison of different regression models and validation techniques for the assessment of wheat leaf area index from hyperspectral data. International journal of remote sensing 36:4519-4534

Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, Normandeau É, Laroche J, Larose S, Jean M, Belzile F (2013) An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. PLOS ONE 8:e54603

Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F (2015) Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant biotechnology journal 13:211-221

Song J, Sun X, Zhang K, Liu S, Wang J, Yang C, Jiang S, Siyal M, Li X, Qi Z (2020) Identification of QTL and genes for pod number in soybean by linkage analysis and genome-wide association studies. Molecular Breeding 40:1-14

Streitner C, Danisman S, Wehrle F, Schöning JC, Alfano JR, Staiger D (2008) The small glycine-rich RNA binding protein AtGRP7 promotes floral transition in Arabidopsis thaliana. The Plant Journal 56:239-250

Stroup W, Mulitze D (1991) Nearest neighbor adjusted best linear unbiased prediction. The American Statistician 45:194-200

Su Q, Lu W, Du D, Chen F, Niu B, Chou K-C (2017) Prediction of the aquatic toxicity of aromatic compounds to tetrahymena pyriformis through support vector regression. Oncotarget 8:49359

Suhre JJ, Weidenbenner NH, Rowntree SC, Wilson EW, Naeve SL, Conley SP, Casteel SN, Diers BW, Esker PD, Specht JE (2014) Soybean yield partitioning changes revealed by genetic gain and seeding rate interactions. Agronomy Journal 106:1631-1642

Sulistyo A, Sari K (2018) Correlation, path analysis and heritability estimation for agronomic traits contribute to yield on soybean. IOP Conference Series: Earth and Environmental Science, p 012034

Sun S, Wang C, Ding H, Zou Q (2020) Machine learning and its applications in plant molecular studies. Briefings in Functional Genomics 19:40-48

Sun X, Shantharaj D, Kang X, Ni M (2010) Transcriptional and hormonal signaling control of Arabidopsis seed development. Current opinion in plant biology 13:611-620

Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, Sun YV (2009) Machine learning in genome-wide association studies. Genetic epidemiology 33:S51-S57

Temesgen D, Assefa F (2020) Inoculation of native symbiotic effective Sinorhizobium spp. enhanced soybean [Glycine max (L.) Merr.] grain yield in Ethiopia. Environmental Systems Research 9:1-19

Torkamaneh D, Laroche J, Belzile F (2020) Fast-GBS v2.0: an analysis toolkit for genotyping-by-sequencing data. Genome 63:577-581

Tulpan D (2020) 311 A brief overview, comparison and practical applications of machine learning models. Journal of Animal Science 98:44-45

Vuong T, Sonah H, Meinhardt C, Deshmukh R, Kadam S, Nelson R, Shannon J, Nguyen H (2015) Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. BMC genomics 16:1-13

Wang L, Hu X, Jiao C, Li Z, Fei Z, Yan X, Liu C, Wang Y, Wang X (2016) Transcriptome analyses of seed development in grape hybrids reveals a possible mechanism influencing seed size. BMC genomics 17:1-15

Watanabe S, Hideshima R, Xia Z, Tsubokura Y, Sato S, Nakamoto Y, Yamanaka N, Takahashi R, Ishimoto M, Anai T (2009) Map-based cloning of the gene associated with the soybean maturity locus E3. Genetics 182:1251-1262

Weber H, Borisjuk L, Wobus U (2005) Molecular physiology of legume seed development. Annu Rev Plant Biol 56:253-279

Wen Y-J, Zhang H, Ni Y-L, Huang B, Zhang J, Feng J-Y, Wang S-B, Dunwell JM, Zhang Y-M, Wu R (2018) Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Briefings in bioinformatics 19:700-712

Wen Z, Tan R, Yuan J, Bales C, Du W, Zhang S, Chilvers MI, Schmidt C, Song Q, Cregan PB (2014) Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean. BMC genomics 15:1-11

Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V (2001) Feature selection for SVMs. Advances in neural information processing systems, pp 668-674

Wickham H (2011) ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics 3:180-185

Williamson BD, Gilbert PB, Simon NR, Carone M (2020) A unified approach for inference on algorithm-agnostic variable importance. arXiv preprint arXiv:200403683

Wu Y, Liu Y (2009) Variable selection in quantile regression. Statistica Sinica:801-817

Xavier A, Jarquin D, Howard R, Ramasubramanian V, Specht JE, Graef GL, Beavis WD, Diers BW, Song Q, Cregan PB (2018) Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. G3: Genes, Genomes, Genetics 8:519-529

Xavier A, Muir WM, Rainey KM (2016a) Assessing predictive properties of genome-wide selection in soybeans. G3: Genes, Genomes, Genetics 6:2611-2616

Xavier A, Muir WM, Rainey KM (2016b) Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans. BMC bioinformatics 17:1-9

Xavier A, Rainey KM (2020) Quantitative Genomic Dissection of Soybean Yield Components. G3: Genes, Genomes, Genetics 10:665-675

Yang J, Yeh C-TE, Ramamurthy RK, Qi X, Fernando RL, Dekkers JC, Garrick DJ, Nettleton D, Schnable PS (2018) Empirical comparisons of different statistical models to identify and validate kernel row number-associated variants from structured multi-parent mapping populations of maize. G3: Genes, Genomes, Genetics 8:3567-3575

Yin Z, Qi H, Mao X, Wang J, Hu Z, Wu X, Liu C, Xin D, Zuo X, Chen Q (2018) QTL mapping of soybean node numbers on the main stem and meta-analysis for mining candidate genes. Biotechnology & Biotechnological Equipment 32:915-922

Yoosefzadeh-Najafabadi M, Earl HJ, Tulpan D, Sulik J, Eskandari M (2021a) Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean. Frontiers in Plant Science 11

Yoosefzadeh-Najafabadi M, Tulpan D, Eskandari M (2021b) Application of machine learning and genetic optimization algorithms for modeling and optimizing soybean yield using its component traits. Plos one 16:e0250665

Zhang H, Hao D, Sitoe HM, Yin Z, Hu Z, Zhang G, Yu D (2015a) Genetic dissection of the relationship between plant architecture and yield component traits in soybean (Glycine max) by association analysis across multiple environments. Plant Breeding 134:564-572

Zhang J, Song Q, Cregan PB, Jiang G-L (2016) Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycinemax). Theoretical and Applied Genetics 129:117-130

Zhang J, Song Q, Cregan PB, Nelson RL, Wang X, Wu J, Jiang G-L (2015b) Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. BMC genomics 16:1-11

Zhang J, Wang X, Lu Y, Bhusal SJ, Song Q, Cregan PB, Yen Y, Brown M, Jiang G-L (2018) Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. Molecular Plant 11:460-472

Tables

Table 1. The list of detected QTL for soybean maturity using different GWAS methods in the tested soybean population.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment [a]	Reference
MLM	2	2212910	Sclero 3-g31	NA	(Moellers et al. 2017)
		8233782	Seed Weight 6-g1	NA	(Sonah et al. 2015)
FarmCPU	2	2212910	Sclero 3-g31	NA	(Moellers et al. 2017)
		8233766	Seed Weight 6-g1	NA	(Sonah et al. 2015)
	20	37765851	WUE 2-g53	NA	(Kaler et al. 2017)
RF	3	2978272	Leaflet area 1-g2.1	NA	(Fang et al. 2017)
			Leaflet width 1-g4.1	NA	(Fang et al. 2017)
			Leaflet area 1-g2.2	NA	(Fang et al. 2017)
			Leaflet width 1-g4.2	NA	(Fang et al. 2017)
			Salt tolerance 1-g12	NA	(Kan et al. 2015)
	16	5730281	Plant height 6-g17	NA	(Zhang et al. 2015b)
			Plant height 1-g17	NA	(Zhang et al. 2015b)
			First flower 4-g63	NA	(Mao et al. 2017)
	17	34757372	SDS root retention 1-g6	NA	(Bao et al. 2015)
SVR	2	695362	Seed linolenic 2-g1	NA	(Leamy et al. 2017)
			Seed linolenic 2-g2	NA	(Leamy et al. 2017)
		720134	SDS 1-g12.1	2	(Wen et al. 2014)
			SDS 1-g12.2	2	(Wen et al.

				2014)
		Ureide content 1-g2	2	(Ray et al. 2015)
	827374	SDS 1-g12.3	NA	(Wen et al. 2014)
10	1595239	Shoot Cu 1-g8	NA	(Dhanapal et al. 2018)
	1689395	Seed oil 5-g3	NA	(Sonah et al. 2015)
16	2438652	Reproductive period 4-g16	NA	(Zhang et al. 2015b)
		R8 full maturity 9-g2	NA	(Zhang et al. 2015b)
	2460921	Reproductive period 2-g16	NA	(Zhang et al. 2015b)
		R8 full maturity 2-g2	NA	(Zhang et al. 2015b)
19	47513536	R8 full maturity 4-g1	NA	(Sonah et al. 2015)
	47513572	First flower 4-g81	NA	(Mao et al. 2017)

[a] Detected in separate environments in addition to the combined environment. 1) 2018Ridgetown, 2) 2019Ridgetwon, 3) 2018Palmyra, 4) 2019Palmyra, NA) Not found in any separate environment.

MLM: Mixed Linear Model, FarmCPU: Fixed and random model circulating probability unification, RF: Random Forest, SVR: Support Vector Regression

Table 2. The list of detected QTL for soybean yield using different GWAS methods in the tested soybean population.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment [a]	Reference
MLM	5	34391386	Ureide content 1-g16.1	NA	(Ray et al. 2015)
			Ureide content 1-g16.2	NA	(Ray et al. 2015)
FarmCPU	5	34391386	Ureide content 1-g16.1	NA	(Ray et al. 2015)
			Ureide content 1-g16.2	NA	(Ray et al. 2015)
RF	7	1032587	WUE 2-g18	NA	(Kaler et al. 2017)
SVR	3	36309302	First flower 4-g10	NA	(Mao et al. 2017)
			First flower 3-g2	NA	(Hu et al. 2014)
			Seed weight 4-g3	NA	(Hu et al. 2014)
			Seed yield 4-g2	NA	(Hu et al. 2014)
			R8 full maturity 3-g3	NA	(Hu et al. 2014)
		37617293	Plant height 3-g17	NA	(Contreras-Soto et al. 2017)
			Leaflet shape 1-g1.1	NA	(Fang et al. 2017)
			Leaflet shape 1-g1.2	NA	(Fang et al. 2017)
			Leaflet shape 1-g1.3	NA	(Fang et al. 2017)
			Seed set 1-	NA	(Fang et al.

		g32.1		2017)
		Seed set 1-g32.2	NA	(Fang et al. 2017)
7	44488152	Seed yield 4-g4	NA	(Hu et al. 2014)
	1032587	WUE 2-g18	NA	(Kaler et al. 2017)
15	34958361	SCN 5-g35	NA	(Li et al. 2016)
19	41385139	Seed weight 5-g20	NA	(Zhang et al. 2016)
		Seed weight 4-g18	NA	(Hu et al. 2014)
		Seed yield 4-g5	NA	(Hu et al. 2014)
		Shoot Zn 1-g28.1	NA	(Dhanapal et al. 2018)
		Shoot Zn 1-g28.2	NA	(Dhanapal et al. 2018)
		Shoot Zn 1-g29.1	NA	(Dhanapal et al. 2018)
		Shoot Zn 1-g29.2	NA	(Dhanapal et al. 2018)
		Shoot Zn 1-g29.3	NA	(Dhanapal et al. 2018)

[a] Detected in separate environments in addition to the combined environment. 1) 2018Ridgetown, 2) 2019Ridgetwon, 3) 2018Palmyra, 4) 2019Palmyra, NA) Not found in any separate environment.

MLM: Mixed Linear Model, FarmCPU: Fixed and random model circulating probability unification, RF: Random Forest, SVR: Support Vector Regression

Table 3. The list of detected QTL for soybean total number of nodes per plant (NP) using different GWAS methods in the tested soybean population.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment [a]	Reference
FarmCPU	19	40131952	Pubescence density 1-g17	NA	(Chang and Hartman 2017)
			Seed weight 9-g5.1	NA	(Copley et al. 2018)
RF	4	1205787	Shoot Ca 1-g10	NA	(Dhanapal et al. 2018)
	6	50570624	Seed set 1-g51.1	NA	(Fang et al. 2017)
			Seed set 1-g43.1	NA	(Fang et al. 2017)
			Seed set 1-g25.1	NA	(Fang et al. 2017)
			Seed set 1-g43.2	NA	(Fang et al. 2017)
			Seed set 1-g25.2	NA	(Fang et al. 2017)
			Seed set 1-g51.2	NA	(Fang et al. 2017)
		50570473	Seed set 1-g43.3	NA	(Fang et al. 2017)
			Seed set 1-g51.3	NA	(Fang et al. 2017)
			Seed set 1-g25.3	NA	(Fang et al. 2017)
			Pod number 1-g3	NA	(Fang et al. 2017)
			Seed palmitic 2-g2	NA	(Fang et al. 2017)
			Seed long-chain faty acid 1-g22	NA	(Fang et al. 2017)
SVR	6	50570624	Seed set 1-g51.1	NA	(Fang et al. 2017)
			Seed set 1-g43.1	NA	(Fang et al. 2017)

		Seed set 1-g25.1	NA	(Fang et al. 2017)
		Seed set 1-g43.2	NA	(Fang et al. 2017)
		Seed set 1-g25.2	NA	(Fang et al. 2017)
		Seed set 1-g51.2	NA	(Fang et al. 2017)
	50570473	Seed set 1-g43.3	NA	(Fang et al. 2017)
		Seed set 1-g51.3	NA	(Fang et al. 2017)
		Seed set 1-g25.3	NA	(Fang et al. 2017)
		Pod number 1-g3	NA	(Fang et al. 2017)
		Seed palmitic 2-g2	NA	(Fang et al. 2017)
		Seed long-chain faty acid 1-g22	NA	(Fang et al. 2017)
7	1032587	WUE 2-g18	NA	(Kaler et al. 2017)
	1092403	WUE 2-g18	NA	(Kaler et al. 2017)
		First flower 3-g4	NA	(Fang et al. 2017)
18	55645699	Leaflet shape 1-g4.1	NA	(Fang et al. 2017)
		Leaflet shape 1-g4.2	NA	(Fang et al. 2017)
		Leaflet shape 1-g4.3	NA	(Fang et al. 2017)
		Seed stearic 4-g5	NA	(Li et al. 2015)
		Node number 1-g6.1	NA	(Fang et al. 2017)
		Node number 1-g6.2	NA	(Fang et al. 2017)
		Pod number 1-g1.1	NA	(Fang et al. 2017)

		Pod number 1-g1.2	NA	(Fang et al. 2017)
		Pode number 1-g1.3	NA	(Fang et al. 2017)
		WUE 3-g31	NA	(Kaler et al. 2017)
		Seed weight, SoyNAM 14-g28	NA	(Xavier et al. 2016b)
		Lodging, SoyNAM 4-g15	NA	(Cook et al. 2014)
		Branching 1-g1.1	NA	(Fang et al. 2017)
		Plant height 5-g4.2	NA	(Fang et al. 2017)
		Plant height 5-g4.3	NA	(Fang et al. 2017)
		Shoot p 1-g30	NA	(Dhanapal et al. 2018)
19	47350110	Node number 1-g2.3	NA	(Fang et al. 2017)

[a] Detected in separate environments in addition to the combined environment. 1) 2018Ridgetown, 2) 2019Ridgetwon, 3) 2018Palmyra, 4) 2019Palmyra, NA) Not found in any separate environment.

MLM: Mixed Linear Model, FarmCPU: Fixed and random model circulating probability unification, RF: Random Forest, SVR: Support Vector Regression

Table 4. The list of detected QTL for soybean total number of non-reproductive nodes per plant (NRNP) using different GWAS methods in the tested soybean population.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment [a]	Reference
MLM	15	10193796	Seed protein 6-g2	NA	(Zhang et al. 2018)
			Seed Arg 1-g4	NA	(Zhang et al. 2018)
			Seed coat luster 1-g1.3	NA	(Fang et al. 2017)
FarmCPU	15	10193796	Seed protein 6-g2	NA	(Zhang et al. 2018)
			Seed Arg 1-g4	NA	(Zhang et al. 2018)
			Seed coat luster 1-g1.3	NA	(Fang et al. 2017)
RF	1	54647498	First flower 4-g2	NA	(Mao et al. 2017)
	7	329800	Phytoph 2-g32	NA	(Qin et al. 2017)
			Phytoph 2-g7	NA	(Qin et al. 2017)
	18	12945778	SCN 4-g14	NA	(Vuong et al. 2015)
	19	40218800	Seed weight 9-g5.1	NA	(Copley et al. 2018)
SVR	7	1032587 [2]	WUE 2-g18	2	(Kaler et al. 2017)
	19	40218800	Seed weight 9-g5.1	NA	(Copley et al. 2018)

[a] Detected in separate environments in addition to the combined environment. 1) 2018Ridgetown, 2) 2019Ridgetwon, 3) 2018Palmyra, 4) 2019Palmyra, NA) Not found in any separate environment.

MLM: Mixed Linear Model, FarmCPU: Fixed and random model circulating probability unification, RF: Random Forest, SVR: Support Vector Regression

Table 5. The list of detected QTL for soybean total number of reproductive nodes per plant (RNP) using different GWAS methods in the tested soybean population.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment [a]	Reference
RF	9	40285014	Shoot Fe 1-g8.1	NA	(Dhanapal et al. 2018)
			Shoot Fe 1-g8.2	NA	(Dhanapal et al. 2018)
			Shoot Fe 1-g8.3	NA	(Dhanapal et al. 2018)
			Shoot Fe 1-g9	NA	(Dhanapal et al. 2018)
			Shoot Fe 1-g10	NA	(Dhanapal et al. 2018)
			Shoot Fe 1-g11	NA	(Dhanapal et al. 2018)
			Soybean mosaic virus 2-g5	NA	(Che et al. 2017)
	15	34958361	SCN 5-g35	NA	(Li et al. 2016)
SVR	7	1032587	WUE 2-g18	NA	(Kaler et al. 2017)
	15	34958361 [1]	SCN 5-g35	1	(Li et al. 2016)

[a] Detected in separate environments in addition to the combined environment. 1) 2018Ridgetown, 2) 2019Ridgetwon, 3) 2018Palmyra, 4) 2019Palmyra, NA) Not found in any separate environment.

MLM: Mixed Linear Model, FarmCPU: Fixed and random model circulating probability unification, RF: Random Forest, SVR: Support Vector Regression

Table 6. The list of detected QTL for soybean total number of pods per plant (PP) using different GWAS methods in the tested soybean population.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment [a]	Reference
RF	7	15331676	Seed weight, SoyNAM 14-g11	NA	(Xavier et al. 2016b)
	19	42300695	First flower 4-g77	NA	(Mao et al. 2017)
			Lodging, SoyNAM 4-g17	NA	(Cook et al. 2014)
SVR	9	39366957	Pod number 1-g4.1	NA	(Fang et al. 2017)
			Pod number 1-g4.2	NA	(Fang et al. 2017)
			Pod number 1-g4.3	NA	(Fang et al. 2017)
			Seed thickness 2-g4	NA	(Fang et al. 2017)
	9	39372117	Seed Thr 2-g1	NA	(Li et al. 2018b)
			Seed Ser 2-g1	NA	(Li et al. 2018b)
			Seed Tyr 2-g2	NA	(Li et al. 2018b)
			Seed Lys 2-g2	NA	(Li et al. 2018b)
			Seed leu 2-g2	NA	(Li et al. 2018b)
			Seed ile 2-g2	NA	(Li et al. 2018b)
			Seed Ala 2-g2	NA	(Li et al. 2018b)
			Seed Gly 2-g2	NA	(Li et al. 2018b)
	11	5245870	Ureide content 1-g29	NA	(Ray et al. 2015)
			Pod number 1-g6	NA	(Fang et al. 2017)
	18	55645699	Leaflet shape 1-g4.1	NA	(Fang et al.

				2017)
	55469601	Leaflet shape 1-g4.2	NA	(Fang et al. 2017)
		Leaflet shape 1-g4.3	NA	(Fang et al. 2017)
		Seed stearic 4-g5	NA	(Li et al. 2015)
		Node number 1-g6.1	NA	(Fang et al. 2017)
		Node number 1-g6.2	NA	(Fang et al. 2017)
		Pode number 1-g1.1	NA	(Fang et al. 2017)
		Pode number 1-g1.2	NA	(Fang et al. 2017)
		Pode number 1-g1.3	NA	(Fang et al. 2017)
		WUE 3-g31	NA	(Dhanapal et al. 2015a)
		Seed weight, SoyNAM 14-g28	NA	(Xavier et al. 2016b)
		Lodging, SoyNAM 4-g15	NA	(Cook et al. 2014)
		Branching 1-g1.1	NA	(Fang et al. 2017)
		Plant height 5-g4.2	NA	(Fang et al. 2017)
		Plant height 5-g4.3	NA	(Fang et al. 2017)
		Shoot p 1-g30	NA	(Dhanapal et al. 2018)
		Seed yield, SoyNAM 7-g19	NA	(Cook et al. 2014)
		R8 full maturity, SoyNAM 13-g19	NA	(Cook et al. 2014)
		Plant height 5-g4.3	NA	(Fang et al. 2017)
19	43077182	Seed weight 9-g5.2	NA	(Copley et al. 2018)

	Seed weight 5-g21	NA	(Copley et al. 2018)
	First flower 5-g3	NA	(Fang et al. 2017)
	First flower 5-g17	NA	(Fang et al. 2017)
47235604	First flower 4-g77	NA	(Mao et al. 2017)
	Seed palmitic 1-g19	NA	(Priolli et al. 2015)
47350110	Leaf carotenoid content 1-g14	NA	(Dhanapal et al. 2015b)
	Ureide content 1-g50.3	NA	(Ray et al. 2015)
	Ureide content 1-g50.4	NA	(Ray et al. 2015)
47224293	Node number 1-g2.3	NA	(Fang et al. 2017)

[a] detected in separate environments in addition to the combined environment. 1) 2018Ridgetown, 2) 2019Ridgetwon, 3) 2018Palmyra, 4) 2019Palmyra, NA) Not found in any separate environment.

MLM: Mixed Linear Model, FarmCPU: Fixed and random model circulating probability unification, RF: Random Forest, SVR: Support Vector Regression

Figures

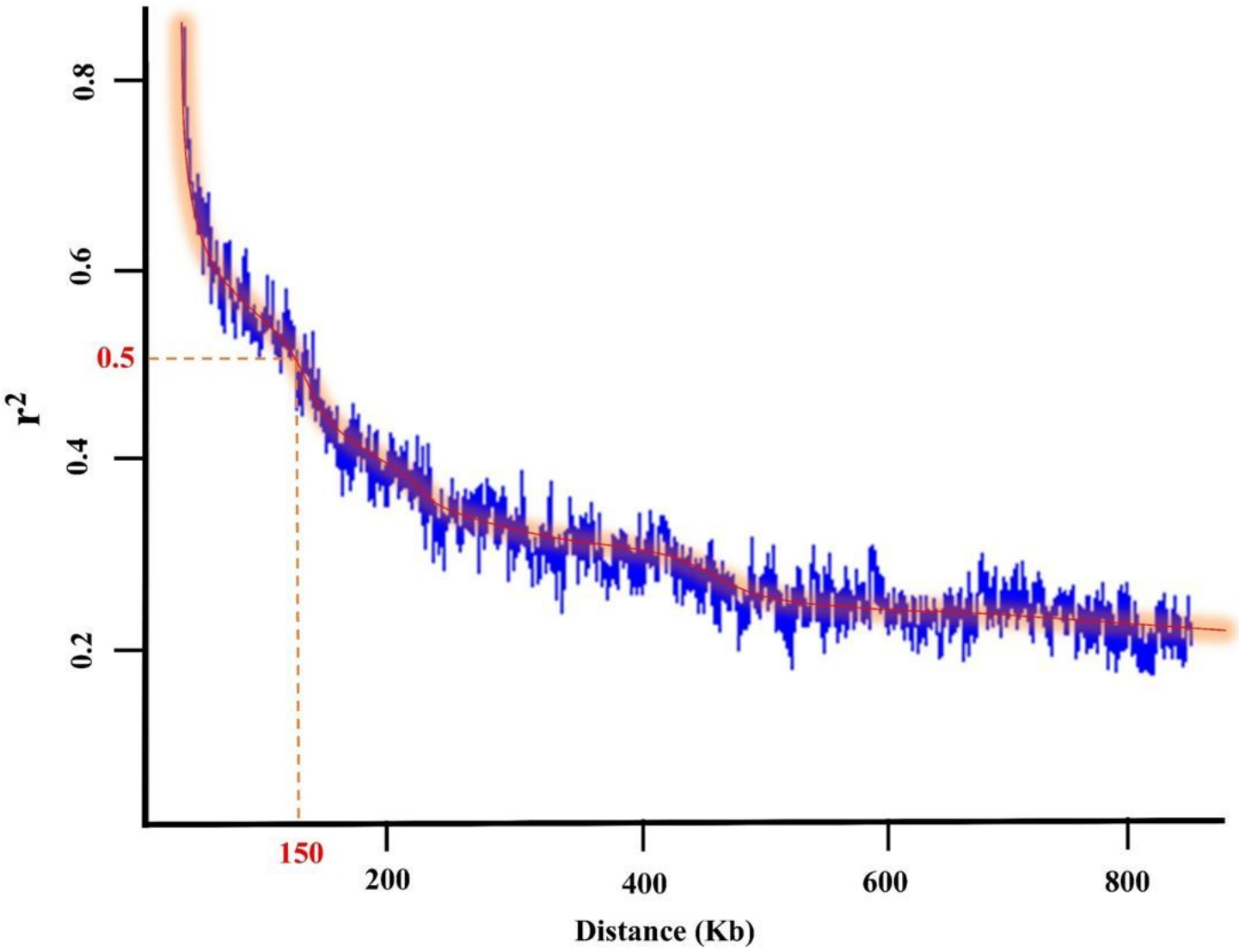

Figure 1

LD decay distance in the tested 227 soybean genotypes.

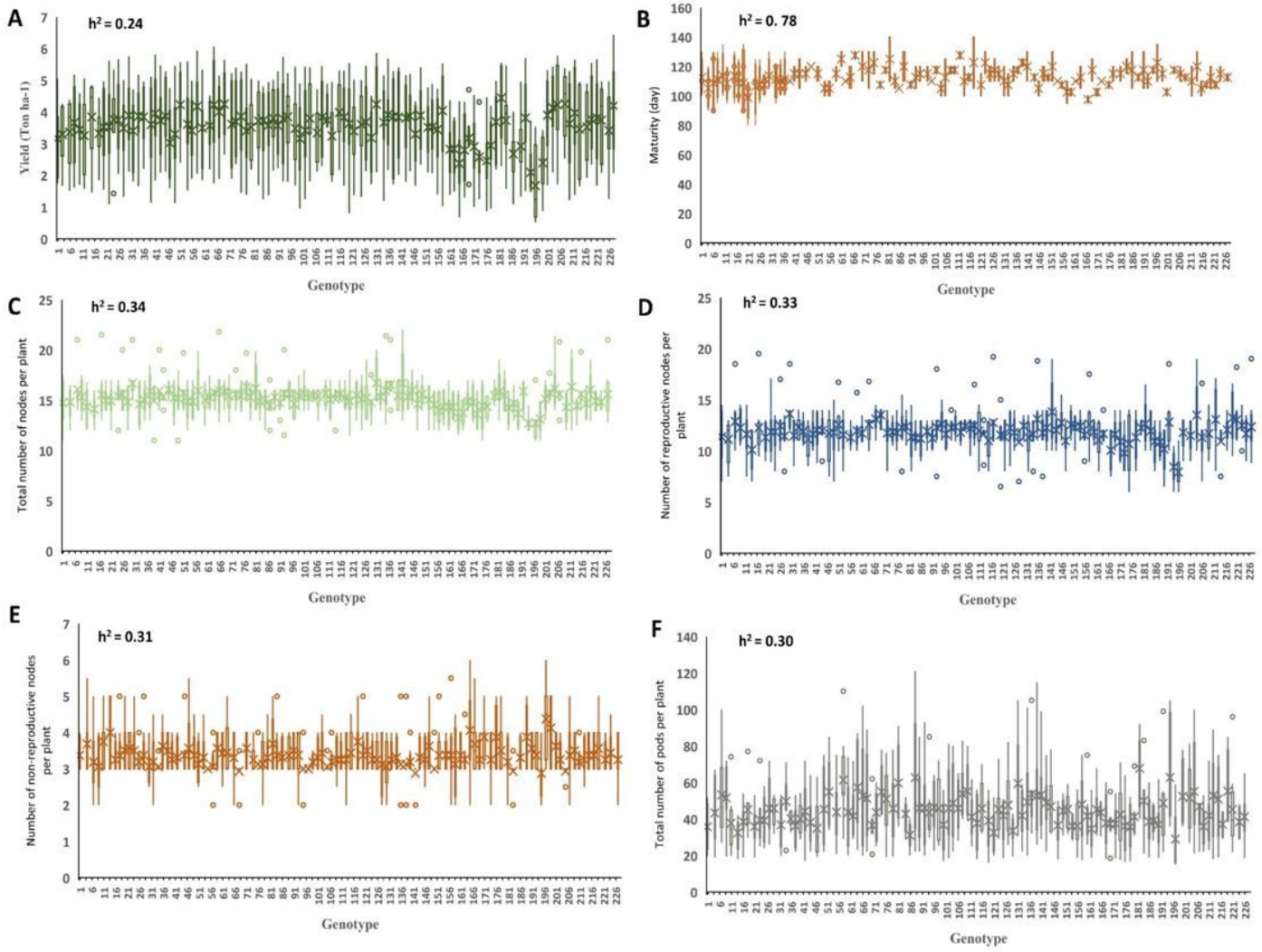

Figure 2

The distribution of seed yield (A), maturity (B), NP (C), NRNP (D), RNP (E), and PP (F) in 227 soybean genotypes across four environments. The estimated heritability is provided for each of the six traits. RNP: Total number of reproductive nodes per plant, NRNP: The total number of non-reproductive nodes per plant, NP: The total nodes per plant, PP: The total number of pods per plant.

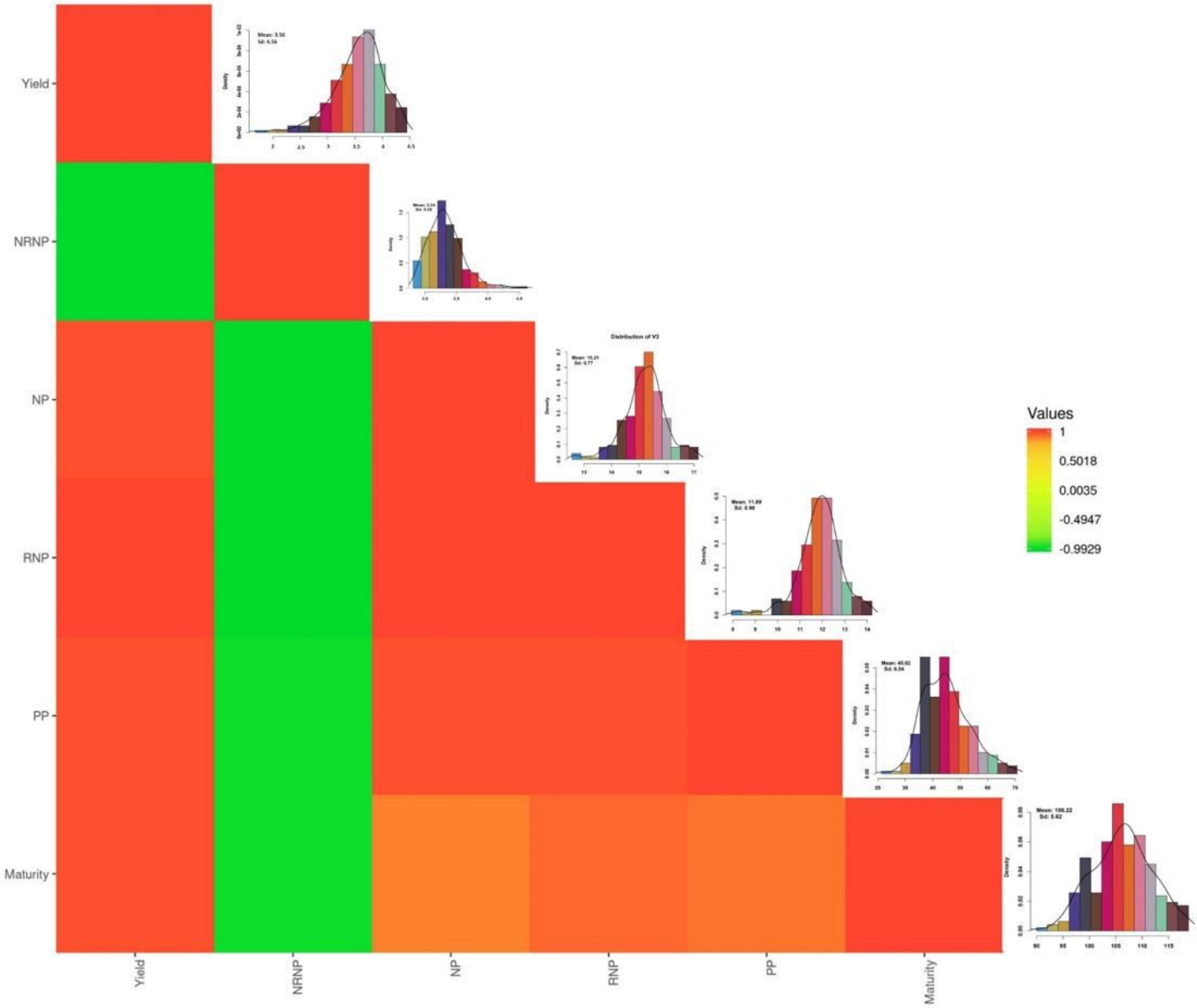

Figure 3

The distributions and Pearson correlations among the soybean seed yield, maturity, and yield component traits. RNP: Total number of reproductive nodes per plant, NRNP: The total number of non-reproductive nodes per plant, NP: The total nodes per plant, PP: The total number of pods per plant. The heat map scale for values is provided by colour for the panel.

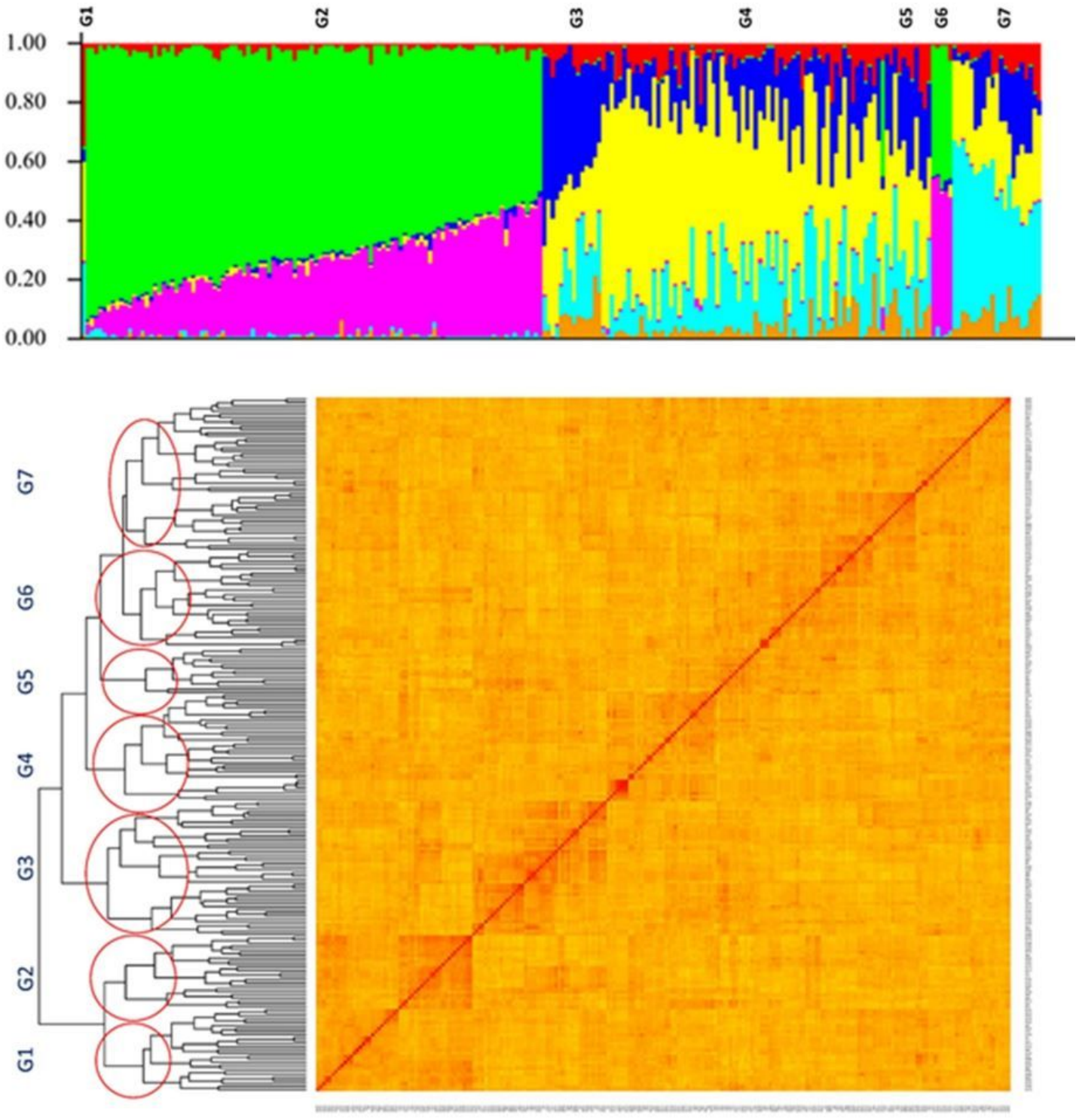

Figure 4

Structure and kinship plots for the 227 soybean genotypes. The x-axis is the number of genotypes used in this GWAS panel, and the y axis is the membership of each subgroup. G1-G7 stands for the subpopulation.

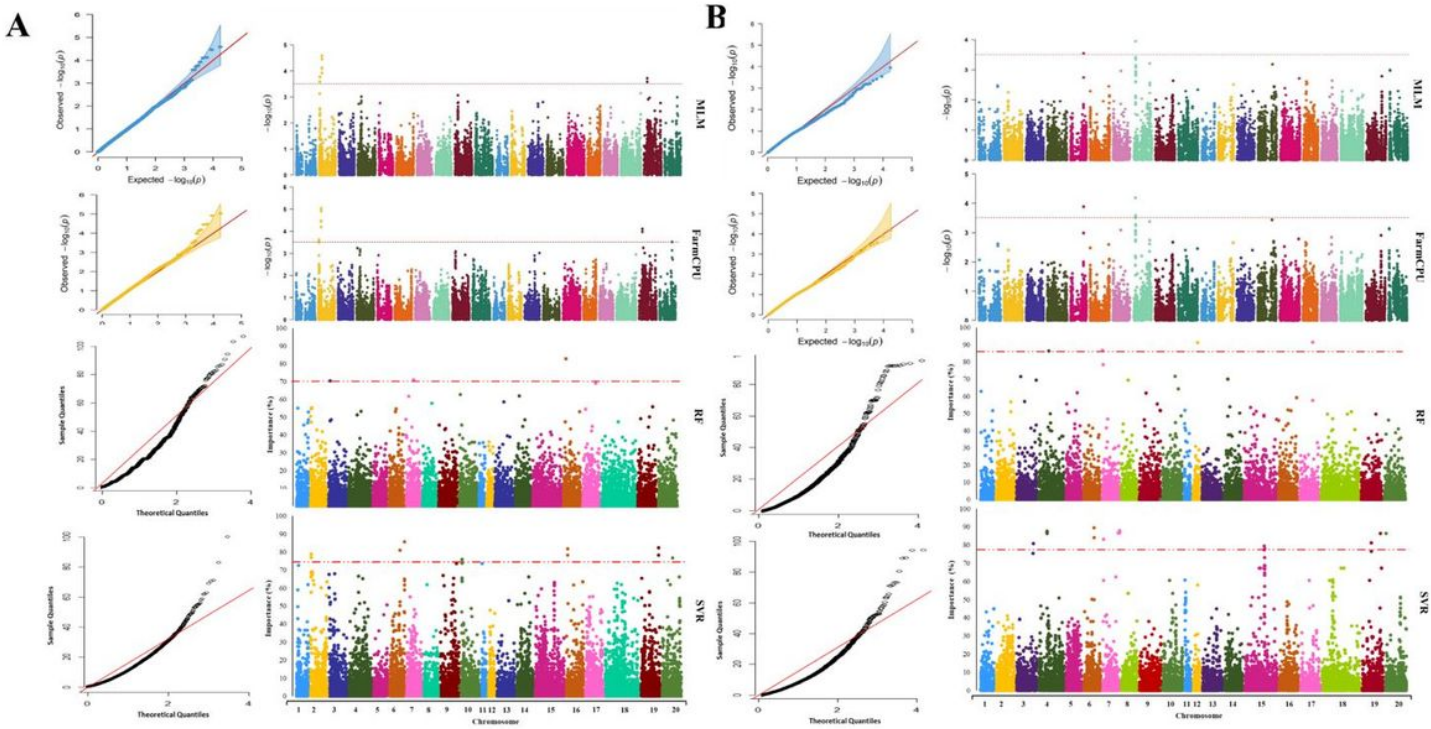

Figure 5

Genome-wide Manhattan and quantile-quantile plots for GWAS studies of A) maturity and B) seed yield in soybean using MLM, FarmCPU, RF, and SVR methods, from top to bottom, respectively.

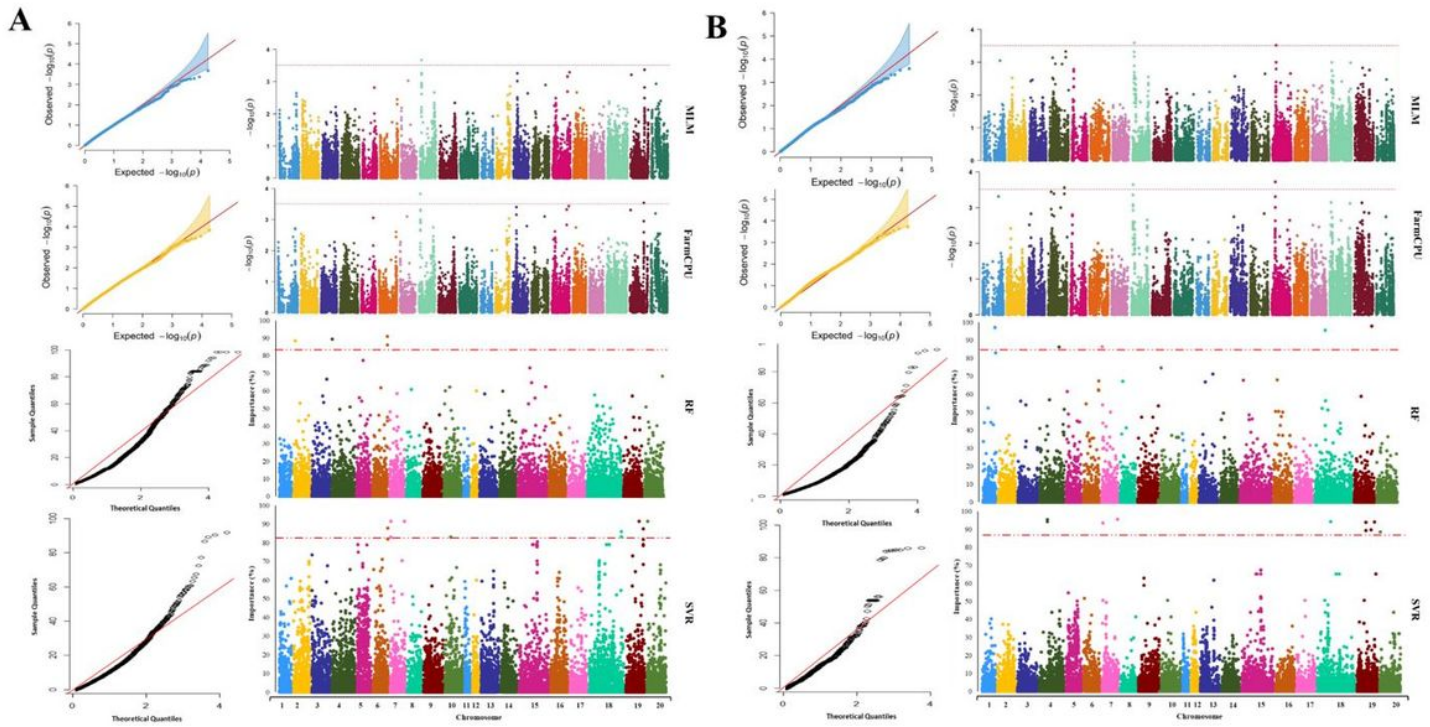

Figure 6

Genome-wide Manhattan and quantile-quantile plots for GWAS studies of A) the total number of nodes (NP) and B) the total number of non-reproductive nodes (NRNP) in soybean using MLM, FarmCPU, RF, and SVR methods, from top to bottom, respectively.

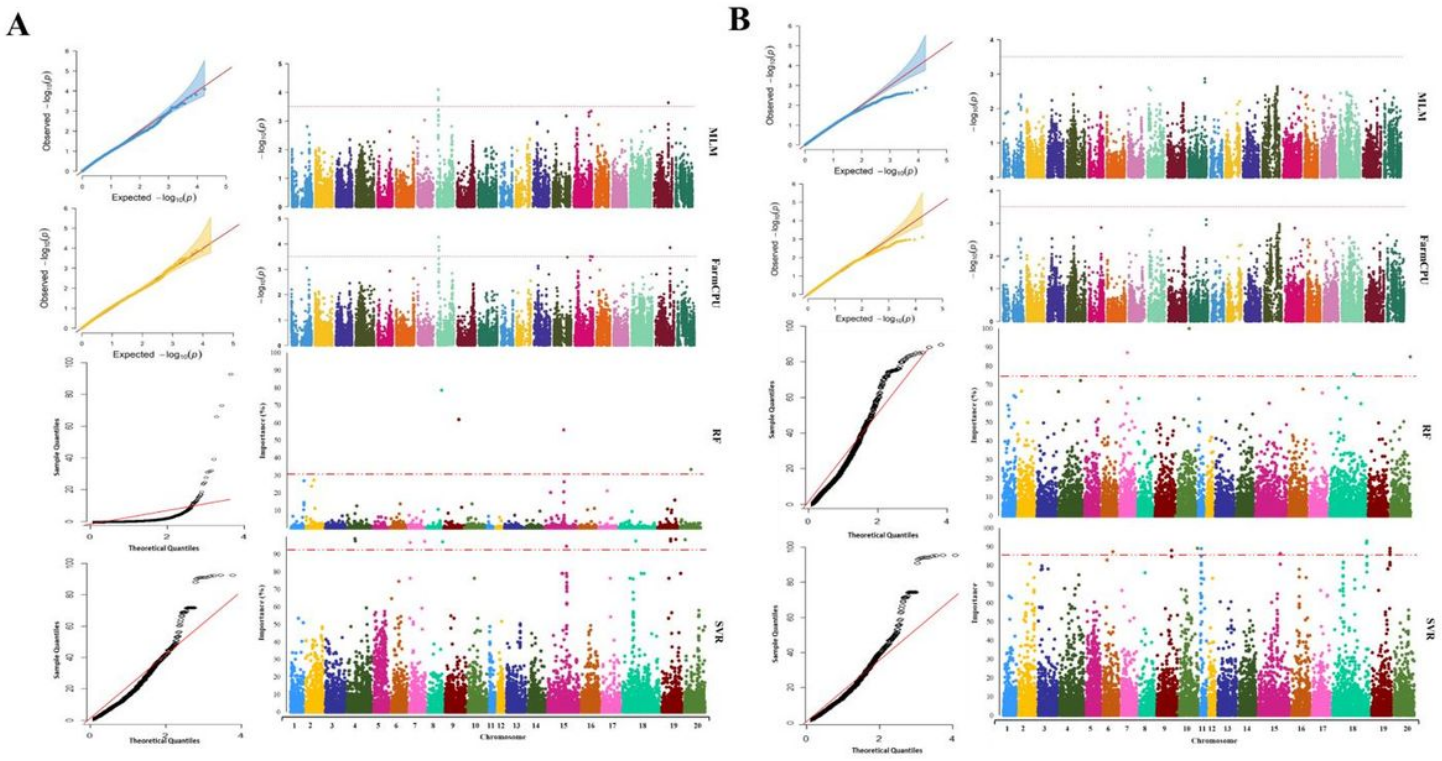

Figure 7

Genome-wide Manhattan and quantile-quantile plots for GWAS studies of A) The total number of reproductive nodes (RNP) and B) the total number of pods (PP) in soybean using MLM, FarmCPU, RF, and SVR methods, from top to bottom, respectively.

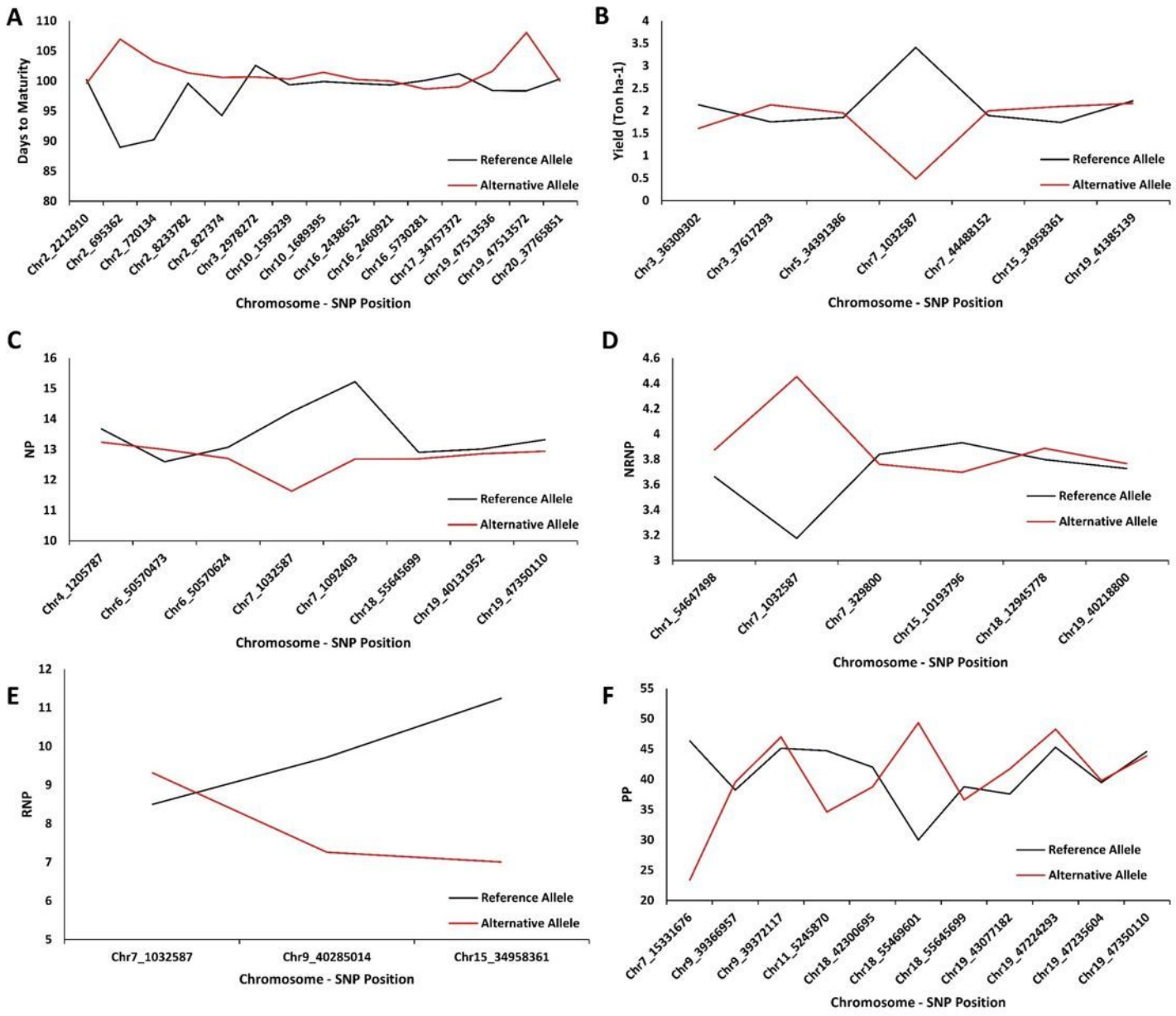

Figure 8

The average effects of reference allele and alternative allele from the detected SNP's peak for maturity (A), seed yield (B), NP (C), NRNP (D), RNP (E), and PP (F) in 227 soybean genotypes across four environments. RNP: Total number of reproductive nodes per plant, NRNP: The total number of non-reproductive nodes per plant, NP: The total nodes per plant, PP: The total number of pods per plant.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.pdf