frontiers in Plant Science

ORIGINAL RESEARCH
published: 22 November 2021
doi: 10.3389/fpls.2021.777028

Genome-Wide Association Studies of Soybean Yield-Related Hyperspectral Reflectance Bands Using Machine Learning-Mediated Data Integration Methods

*Mohsen Yoosefzadeh-Najafabadi[1], Sepideh Torabi[1], Dan Tulpan[2], Istvan Rajcan[1] and Milad Eskandari[1]**

[1] Department of Plant Agriculture, University of Guelph, Guelph, ON, Canada, [2] Department of Animal Biosciences, University of Guelph, Guelph, ON, Canada

OPEN ACCESS

Edited by:
Dick de Ridder,
Wageningen University and Research, Netherlands

Reviewed by:
Jianbo He,
Nanjing Agricultural University, China
Milind B. Ratnaparkhe,
ICAR Indian Institute of Soybean Research, India

***Correspondence:**
Milad Eskandari
meskanda@uoguelph.ca

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: 14 September 2021
Accepted: 18 October 2021
Published: 22 November 2021

Citation:
Yoosefzadeh-Najafabadi M, Torabi S, Tulpan D, Rajcan I and Eskandari M (2021) Genome-Wide Association Studies of Soybean Yield-Related Hyperspectral Reflectance Bands Using Machine Learning-Mediated Data Integration Methods. Front. Plant Sci. 12:777028. doi: 10.3389/fpls.2021.777028

In conjunction with big data analysis methods, plant omics technologies have provided scientists with cost-effective and promising tools for discovering genetic architectures of complex agronomic traits using large breeding populations. In recent years, there has been significant progress in plant phenomics and genomics approaches for generating reliable large datasets. However, selecting an appropriate data integration and analysis method to improve the efficiency of phenome-phenome and phenome-genome association studies is still a bottleneck. This study proposes a hyperspectral wide association study (HypWAS) approach as a phenome-phenome association analysis through a hierarchical data integration strategy to estimate the prediction power of hyperspectral reflectance bands in predicting soybean seed yield. Using HypWAS, five important hyperspectral reflectance bands in visible, red-edge, and near-infrared regions were identified significantly associated with seed yield. The phenome-genome association analysis of each tested hyperspectral reflectance band was performed using two conventional genome-wide association studies (GWAS) methods and a machine learning mediated GWAS based on the support vector regression (SVR) method. Using SVR-mediated GWAS, more relevant QTL with the physiological background of the tested hyperspectral reflectance bands were detected, supported by the functional annotation of candidate gene analyses. The results of this study have indicated the advantages of using hierarchical data integration strategy and advanced mathematical methods coupled with phenome-phenome and phenome-genome association analyses for a better understanding of the biology and genetic backgrounds of hyperspectral reflectance bands affecting soybean yield formation. The identified yield-related hyperspectral reflectance bands using HypWAS can be used as indirect selection criteria for selecting superior genotypes with improved yield genetic gains in large breeding populations.

Keywords: proximal sensing, support vector machine, hierarchical data integration, soybean breeding, recursive feature elimination (RFE), genome-wide association study (GWAS), multi-omics

INTRODUCTION

Soybean (*Glycine max* [L.] Merr.) can be considered one of the super crops that is substantially used for food and feed, green manure, biodiesel, and fiber (Seck et al., 2020). Soybean breeders continually breed soybean genotypes with improved desired traits of interest, such as yield (Yoosefzadeh-Najafabadi et al., 2021a). However, yield is a complex trait affected by intrinsic and extrinsic factors as well as their interactions (Anuarbek et al., 2020; Yoosefzadeh-Najafabadi et al., 2021a). Therefore, a sophisticated understanding of the biological aspects of plant genomes is required for sustainable improvements of yield potential in major crops (Somegowda et al., 2021), such as soybean. Soybean breeding programs are moving to implement new genomics, phenomics, and big data analysis for a deeper understanding of soybean yield formation. Having a phenotypic profile of a large plant population with high-density genetic markers are two of the most important factors for better understanding the phenotype and genotype of complex quantitative traits that are usually controlled by various genes with minor and major effects (Wang et al., 2020).

Proximal/remote sensing can be considered as one of the promising high throughput phenotyping tools that can measure the spectral properties of genotypes in a short time in a large breeding population. Most of the spectral measurements are focused on the visible (400–700 nm), red edge (680–780 nm), and near-infrared (700–1100 nm) spectral regions (Alonzo et al., 2014; Hennessy et al., 2020). The visible range can be dissected into blue/blue-green edge (400–499 nm), the green peak (550 nm), and the red (650–700 nm) regions (Rivard et al., 2008; Hennessy et al., 2020). Most of the reflection in the visible region is regulated by the absorption of different foliar pigments such as chlorophyll a and b, carotenoids, and anthocyanins (Castro-Esau et al., 2006; Pu, 2009; Peerbhay et al., 2013; Alonzo et al., 2014; Hennessy et al., 2020). The red-edge region resides between the red and the near-infrared (NIR) regions, which is correlated with internal leaf structure and chlorophyll absorptions (Clevers et al., 2002; Clark et al., 2005; Liu C. et al., 2021). The NIR plateau (780–1327 nm) is another important hyperspectral reflectance region that is dominated by the amount and interaction of water and air within the intercellular spaces (Hennessy et al., 2020; Paulus and Mahlein, 2020; Okubo, 2021).

Several studies reported the high potential of using spectral reflectance to estimate and classify the yield (Yoosefzadeh-Najafabadi et al., 2021a), leaf area index (Chen et al., 2020), plant stress (Feng et al., 2020), and carbon and nitrogen contents (Omidi et al., 2020). In a study done by Zhang et al. (2019), significant association of red and NIR regions with yield are reported. They also demonstrated R5 as the best growth stage for predicting yield and the efficiency of using regression models in predicting soybean yield from the selected hyperspectral reflectance bands. This potential would allow breeders to accurately predict complex traits such as yield, which are typically controlled by several secondary correlated traits, in a short time at early growth stages (Yoosefzadeh-Najafabadi et al., 2021a). While hyperspectral sensors can measure hundreds of wavebands, most of them are redundant due to their high correlation with the adjacent ones (Omidi et al., 2020). Therefore, there is a dire need to find the redundancy of wavebands not only based on the correlation with adjacent bands but also with the estimation of the interaction with other bands in different regions. Genome-wide association studies (GWAS) can be considered as one of the common genetic approaches used for discovering quantitative trait loci (QTL) that are highly associated with a trait of interest (Eltaher et al., 2021). By using GWAS, a QTL associated with a trait of interest can be detected using linkage disequilibrium (LD), which is the non-random association of alleles at specific loci (Somegowda et al., 2021). The detected QTL can be implemented in marker-assisted selection (MAS) for screening large breeding populations in a time- and cost-effective manner (Dababat et al., 2021; Eltaher et al., 2021). Over the past two decades, several statistical methods were used in GWAS to improve statistical power and computational speed (Brachi et al., 2011; Xu et al., 2018). The mixed linear model (MLM) and the fixed and random model circulating probability unification (FarmCPU) approach are known as two of the most common GWAS methods that are currently used in a wide range of genetic studies (Brachi et al., 2011; Lee et al., 2020; Singh et al., 2020). Also, Bonferroni correction and false discovery rate (FDR) are commonly used to set up a threshold for selecting associated QTL with major effects (Brachi et al., 2011; Xu et al., 2018; Lee et al., 2020; Singh et al., 2020).

The application of GWAS was reported in different plant species such as soybean (Brown et al., 2021), maize (Xu et al., 2018), wheat (Tsai et al., 2020), rice (Zhong et al., 2021), and sorghum (Somegowda et al., 2021). While there is no report on the genetic dissection of soybean yield-related hyperspectral reflectance bands, genetic dissection of vegetation index was previously reported in wheat (Wang, 2019; Galán et al., 2020; Wang et al., 2021). Several detected candidate genes related to NDVI, SPAD, and LR in durum wheat (Wang et al., 2021) overlapped with dry biomass, grain yield, and chlorophyll contents. Although GWAS can be considered as a powerful tool to detect the associated genomic regions with major effects, there are several barriers in applying conventional statistical methods in GWAS for identifying genomic regions associated with complex traits (Szymczak et al., 2009). One of the major challenges in GWAS is the high possibility of a false-positive rate that is due to the stochastic noise arise when the population structure is not well defined (Platt et al., 2010) or a high false-negative rate because of the unappropriated way of selecting the threshold (Kaler and Purcell, 2019). Another challenge associated with using the conventional statistical procedures is the "large markers (p), small samples (n)" problem that habitually happens in GWAS when these methods are applied to datasets where the number of markers (i.e., single nucleotide polymorphisms (SNPs)) is significantly larger ($p \gg n$) than the number of genotypes (Kaler et al., 2020; Mohammadi et al., 2020; Xavier and Rainey, 2020). It is well documented that current conventional GWAS methods are only powerful to detect common SNPs with large effects on the target traits that can reach the minimum level of significance (Lee et al., 2020). Therefore, current conventional GWAS approaches may not be well-suited for discovering minor effect SNPs associated with the target traits, especially in plants

with a significantly narrow genetic background (Zhou et al., 2019). For example, several crops, such as soybean, suffer from narrow genetic diversity mainly due to the genetic bottlenecks associated with their domestications and the lack of introducing new sources of genetic diversity (Mikel et al., 2010). While recent advances in sequencing technologies facilitate the accessibility of high-density genetic markers in a short time at a reduced cost, using sophisticated big data analysis methods combined with accurate and rapid large scale phenotyping methods, especially for complex traits, represent major bottlenecks (Hennessy et al., 2020; Yoosefzadeh-Najafabadi et al., 2021a). Recently, Machine Learning (ML) algorithms were shown to be promising computational strategies when applied to plant sciences because of their potential to analyze complex multivariable and nonlinear biological processes, which are commonly observed in complex traits in plants (Jafari and Shahsavar, 2020; Hesami et al., 2021; Yoosefzadeh-Najafabadi et al., 2021a). In general, ML algorithms can be programmed based on existing patterns present in the dataset. Recent studies showed the effectiveness of using ML algorithms to predict complex traits using secondary traits that are highly correlated with the trait of interest (Pantazi et al., 2016; Liakos et al., 2018; Palanivel and Surianarayanan, 2019; Yoosefzadeh-Najafabadi et al., 2021a). Based on the type of problems they solve, ML algorithms can be characterized in four categories as follows: (i) identification, (ii) classification, (iii) quantification, and (iv) prediction. These four categories could be used to identify important variables from multi-dimensional datasets (Liakos et al., 2018; Hesami et al., 2020; Sharifi, 2021). Variable selection methods are commonly used to improve the prediction performance and avoid overfitting rates for classification and prediction problems in high-dimensional datasets (George, 2000). The variable selection methods are, in general, classified into three distinct groups: wrappers, filters, and embedded methods (George, 2000; Heinze and Dunkler, 2017). In filter methods, subsets of variables are selected based on selection criteria that are independent from those used for the final classifier (Chowdhury and Turin, 2020). However, both wrapper and embedded methods implement variable selection based on individual learners (Albashish et al., 2021). For example, recursive feature elimination (RFE) is a representative wrapper-type variable selection algorithm widely used to extract important features from phenomics and genomics data (Gupta and Gupta, 2020; Albashish et al., 2021; Yoosefzadeh-Najafabadi et al., 2021a). RFE discards in an iterative fashion the weak and unstable variables until a target number of variables is reached and thus retains independent variables from the dataset resulting in significant improvements in performance and reduced overfitting of ML algorithms (Sanz et al., 2018).

In addition to the proper ML algorithm choice, adopting an accurate data integration strategy is required for a better understanding of the structure of complex multidimensional traits at different omics levels (Tarazona et al., 2021). These days, more and more data are generated using different omics such as genomics and phenomics, and several data integration strategies such as early, intermediate, late, mixed, and hierarchical strategies are available (Jamil et al., 2020; Picard et al., 2021). A hierarchical data integration strategy is built upon prior knowledge about the relationship between and among different tested omics layers (Picard et al., 2021). For instance, a hierarchical data integration strategy can be used for a better understanding of soybean yield formation by having prior knowledge about the physiological concept of each hyperspectral reflectance in explaining the overall yield variation. The effectiveness of using RFE was reported previously by Yoosefzadeh-Najafabadi et al. (2021a) to extract the important wavelengths for predicting soybean seed yield. In addition, a few studies used ML algorithms in GWAS for detecting QTL associated with complex traits (Zhou et al., 2019; Xavier and Rainey, 2020; Najafabadi et al., 2021). In a GWAS study, Xavier and Rainey (2020) investigated the potential use of Random Forest (RF) for detecting QTL associated with soybean yield components, such as the number of pods and nodes. In addition, the use of RF for detecting QTL with minor effects was reported in a study by Asif et al. (2020). However, using other promising ML algorithms such as support vector regression (SVR) and implementing a hierarchical data integration strategy for better understanding soybean yield using hyperspectral and genome-wide association studies is long overdue. Therefore, this study aimed to: (1) investigate the use of RFE for selecting hyperspectral reflectance wavelengths influencing soybean yield, (2) evaluate the potential use of SVR- mediated GWAS for genetic dissection of important hyperspectral reflectance bands affecting soybean yield, and (3) discover candidate genes linked to identified hyperspectral reflectance bands associated with yield. To the best of our knowledge, this study is the first report where GWAS was used to discover hyperspectral reflectance bands associated with soybean yield. This study also demonstrates the benefits of using ML algorithms and variable selection methods in phenome-phenome and phenome-genome association studies for discovering yield-related physiological traits and genomic regions associated with these traits in soybean. The results of this study can be useful for selecting high yielding soybean genotypes at early growth stages and, therefore, increasing the rate of genetic gain for yield in cultivar development programs.

MATERIALS AND METHODS

Genome-Wide Association Studies Panel and Experimental Design

The GWAS panel was consisted of 227 diverse soybean genotypes that were grown and evaluated for the target traits in Ridgetown (42°27′14.8″N 81°52′48.0″W, 200 m above sea level) and Palmyra (42°25′50.1″N 81°45′06.9″W, 195 m above sea level), Ontario, Canada, over two consecutive years, 2018 and 2019. The experimental design was conducted based on the randomized complete block design (RCBD) in four environments (two locations × two years) with two replications in each environment. Each phenotypic plot for each genotype was consisted of five rows, each 4.2 m long with a row spacing and seedling rate of 43 cm and 57 per m^2, respectively. Overall, there were 1000 soybean plots per year and 500 soybean plots per environment. Also, nearest-neighbor analysis (NNA), as one of the most common error control methods, was used to estimate the accuracy of the phenotypic evaluations and control the spatial

variability in the field (Stroup and Mulitze, 1991; Bowley, 1999; Katsileros et al., 2015).

Seed Yield and Hyperspectral Reflectance Data Collection

Soybean seed yield (t ha^{-1}) was estimated for each plot after harvesting of three middle rows and adjusting for day to maturity and the seed moisture to 13%.

Previous studies reported that environmental stresses at the seed development growth stage (R5), where seeds are 1/8 inches long in pods at one of the four uppermost nodes (Yoosefzadeh-Najafabadi et al., 2021a), in compared to other growth stages, could have greater damage to the soybean yield. It can be because of the fact that plants have no time to recover the yield before physiological maturity (Zhang et al., 2019; Yoosefzadeh-Najafabadi et al., 2021a). Therefore, the R5 growth stage in soybean can be considered as a reliable growth stage for measuring hyperspectral reflectance, if the goal is to predict the overall seed yield (Zhang et al., 2019; Yoosefzadeh-Najafabadi et al., 2021a). The hyperspectral reflectance was measured at the beginning of the R5 stage. Hyperspectral reflectance bands were measured via UniSpec-DC Spectral Analysis System (PP Systems International, Inc., 110 Haverhill Road, Suite 301 Amesbury, MA, United States), which covers 250 bands from 350 to 1100 nm with a 3 nm bandwidth. In general, the measured hyperspectral reflectance consisted of three main regions: visible, red-edge, and near-infrared (NIR) regions. Spectralon panels and a dark reference background were used to adjust incoming solar radiation and calibrate the dual channels, respectively. For each plot, three measurements were recorded at the same spot in order to reduce the noise, and their average, calculated by the best linear unbiased prediction (BLUP) model, was used as the reflectance band datapoint. All of the measurements were conducted close to solar noon to reduce the signal-to-noise (SNR) ratio.

Hyperspectral Data Pre-processing

The pre-processing step is one of the most important steps in hyperspectral reflectance analysis, which reduces the possible electronic fluctuations and sensor noises in datasets. By checking the quality of hyperspectral reflectance data for the tested panel and the result of sensor-specific artifacts, hyperspectral reflectance data of 1,005–1,100 and 350–395 nm, were removed from the original data. The number of reflectance bands was decreased from 250 bands to 62 by increasing the bandwidth from 3 to 10 nm. In order to improve the signal-to-noise ratio, a Savitzky–Golay filter was applied for each reflectance band and data scaling, centering, and principal component analysis (PCA) was conducted in order to detect potential outliers in the dataset. All the pre-processing steps were performed using R software (version 3.6.1).

Statistical Analyses

In order to estimate the genetic values of each soybean genotype, the BLUP was used as one of the most well-known mixed models (Goldberger, 1962). For this aim, 'environment' and 'genotype' factors were considered as fixed and random effects, respectively. Based on the protocol developed by Bowley (1999), all outliers were detected and treated the same as missing data points in further analysis. Overall, the statistical model used in this study is as follows:

$$Y = X_b + Z_g + W_i + e \quad (1)$$

where Y is the vector of trait of interest (selected hyperspectral reflectance bands), b is the vector of block effects, encompasses all the replications and locations, added to the overall mean (assumed fixed), g in the vector of genotype effects (assumed random), in which $g \sim N(0, \sigma^2_g)$, i is the vector of random GxE interaction effects, in which $i \sim N(0, \sigma^2_{int})$, and e is the vector of residuals, in which $e \sim N(0, \sigma^2_e)$. X, Z, and W represent the incidence matrices of b, g, and i effects, respectively.

Also, the heritability (Eq. 2) of each tested trait was calculated based on the following equation:

$$H^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_{int/n} + \sigma^2_{e/nr}} \quad (2)$$

where $\sigma^2 g$ is the genotypic variance; $\sigma^2 int$ is the variance of GxE; $\sigma^2 e$ is error variance; n is the number of locations; r is the number of replications.

Hyperspectral Wide Association Study

With respect to the genome-wide association study, we proposed the term of hyperspectral wide association study (HypWAS) for detecting hyperspectral reflectance bands associated with the trait of interest. In order to detect the important hyperspectral reflectance bands associated with the trait of interest, RFE, as one of the most common variable selection methods (Guyon et al., 2002), was used in this study. The main basis of the RFE is to eliminate variables with low importance scores and select the high importance score variables that explain the trait of interest. In RFE, the first step is to build a model on the complete set of the inputs and computing the importance of each input based on sequential selection strategy (Guyon et al., 2002). The next step is to remove the least important inputs and rebuilding the model to recursively repeat the process. In general, RFE shows how important is a feature for a model with respect to predicting a value and it is strictly describing the prediction power of a feature. In this study, we implemented RFE, considering reflectance bands as input variables and soybean yield as an output variable. All the analyses were done using the *caret* package (Kuhn, 2008) in R software version 3.6.1.

Genotyping

For extracting DNA, young trifoliate leaf tissue was collected from the first soybean phenotyping plot of each genotype at the Ridgetown location and stored after freeze-drying using the Savant ModulyoD Thermoquest (Savant Instruments, Holbrook, NY, United States). DNA was isolated using NucleoSpin Plant II kit (Macherey–Nagel, Duüren, Germany) as per the manufacturer's instructions, and the quality of DNA was checked with Qubit® 2.0 fluorometer (Invitrogen, Carlsbad, CA, United States). The extracted DNA were sent to Genomic

Analysis Platform at Université Laval (Laval, Quebec, Canada) for genotyping-by sequencing (GBS) based on the enzymatic digestion with *ApeKI* (Sonah et al., 2013). GBS for each genotype was done via the Fast-GBS pipeline (Torkamaneh et al., 2020), using Gmax_275_v2 reference genome. After imputing the missing loci by the Markov model using Beagle v5 pipeline and removing markers with a minor allele frequency less than 0.05, a total of 17,958 high-quality single-nucleotide polymorphisms (SNPs) from 227 soybean genotypes used for genomic analysis.

Population Structure Analysis

A total of 17,958 high-quality SNPs were used to conduct the population structure analysis using fastSTRUCTURE (Raj et al., 2014) with K values from 1 to 15. Afterward, the optimum number of subpopulations was calculated using the K tool in the fastSTRUCTURE software.

Association Studies

In this study, MLM and FarmCPU, as the two conventional GWAS methods, were compared with the developed SVR-mediated GWAS method. All the conventional GWAS methods were implemented using the *MVP* (Yin et al., 2021) package in R software version 3.6.1. The popular *Caret* package (Kuhn et al., 2020) in R, was used to develop the SVR-mediated GWAS method.

Mixed Linear Model

One of the most common methods for GWAS is the MLM method developed by Yu et al. (2006). This method has been widely used in GWAS because of its effectiveness in controlling the bias in the population and correcting the inflation from different small genetic effects caused by polygenic background (Bulik-Sullivan et al., 2015; Wang S.-B. et al., 2016; Wen et al., 2018). While the likelihood ratio is not specific to the MLM method, this method is based on the likelihood ratio between the full model (with a marker of interest) and the reduced model (without a marker of interest) (Wen et al., 2018). If we considered Y as the phenotypic value, the MLM equation would be as follows (Eq. 3):

$$Y = X_b + Z_u + e \tag{3}$$

where Y is the vector of phenotypic observations; b is the vector of SNP markers and population structure effects (assumed fixed); u is the vector of additive genetic effects for genotypes (assumed random); e is the vector of residuals (assumed random). X and Z represent the incidence matrices of b and u effects, respectively.

Fixed and Random Model Circulating Probability Unification

This GWAS method was first introduced by Liu et al. (2016) in order to reduce the shortcoming and false discoveries that existed in previously proposed GWAS methods. FarmCPU takes advantage of using the random-effect (REM) and fixed-effect (FEM) models iteratively (Liu et al., 2016). In brief, FEM was used to test the S number of SNPs, simultaneously, based on the following equation (Eq. 4):

$$Y_i = S_{i1}B_1 + S_{i2}B_2 + S_{i3}B_3 + \ldots + S_{it}B_t + M_{ij}K_j + e_i \tag{4}$$

Where Y_i stands for the observation on the ith sample, S_{i1}, S_{i2},..., S_{it} stand for the genotypes of the t pseudo-QTNs, B_1, B_2, B_3, . . ., B_t is the corresponding effect for the pseudo-QTNs, M_{ij} is the genotype of the jth SNPs and ith sample, K_j is known as the corresponding effect of the jth SNPs, and e_i is the residual.

The REM model is used in the FarmCPU method to optimize the selection of the genetic markers based on the *p*-values as follows (Eq. 5):

$$Y_i = U_i + e_i \tag{5}$$

Where Y_i stands for the observation on the ith sample, e_i is the residual, and U_i is the total genetic effect of the ith sample.

Also, false discovery rate (FDR) was used for MLM and FarmCPU to set the significant threshold (Benjamini and Hochberg, 1995).

Support Vector Regression

Support vector regression represents a support vector machine (SVMs) approach used to solve regression problems (Awad and Khanna, 2015). SVR is characterized by the use of the Vapnik-Chervonenkis (VC) theory, sparse solution, and kernels for controlling the number of vectors and margin (Smola and Schölkopf, 2004; Awad and Khanna, 2015). This algorithm is trained by implementing an asymmetrical loss function that equally penalizes low and high misestimates (Vapnik, 1998). The association statistics for SVR can be obtained by evaluating the feature importance, which is previously proposed by Weston et al. (2001). In this study, SNPs were considered as inputs, and the selected hyperspectral reflectance bands were selected as output variables for evaluating the feature importance using the SVR algorithm. In brief, the following equation was used to determine SVR (Eq. 6):

$$Y = W\beta(c) + b \tag{6}$$

Where Y is the output, W stands for the weights for each high dimensional input (β) that is constructed non-linearly on the input space of (c). The upper and lower borderlines are presented as $Y = W\beta\,(c) + b + e+$ and $Y = W\beta\,(c) + b - e$, respectively.

The five-fold cross-validation strategy (Siegmann and Jarmer, 2015) was used to run the variable importance analysis with ten repetitions. The impurity index was selected as the common metric to evaluate the importance of each SNP in explaining the trait of interest. After implementing the variable importance, the achieved scores were scaled to a 0–100% scale. After fitting the algorithm, high variable importance was stored during 1000 times repetitions. Then, all significant SNPs were selected based on a confidence level $\alpha = 0.05$. The global empirical threshold was used for estimating the significant threshold of SNPs associated with selected hyperspectral reflectance bands (Churchill and Doerge, 1994; Doerge and Churchill, 1996).

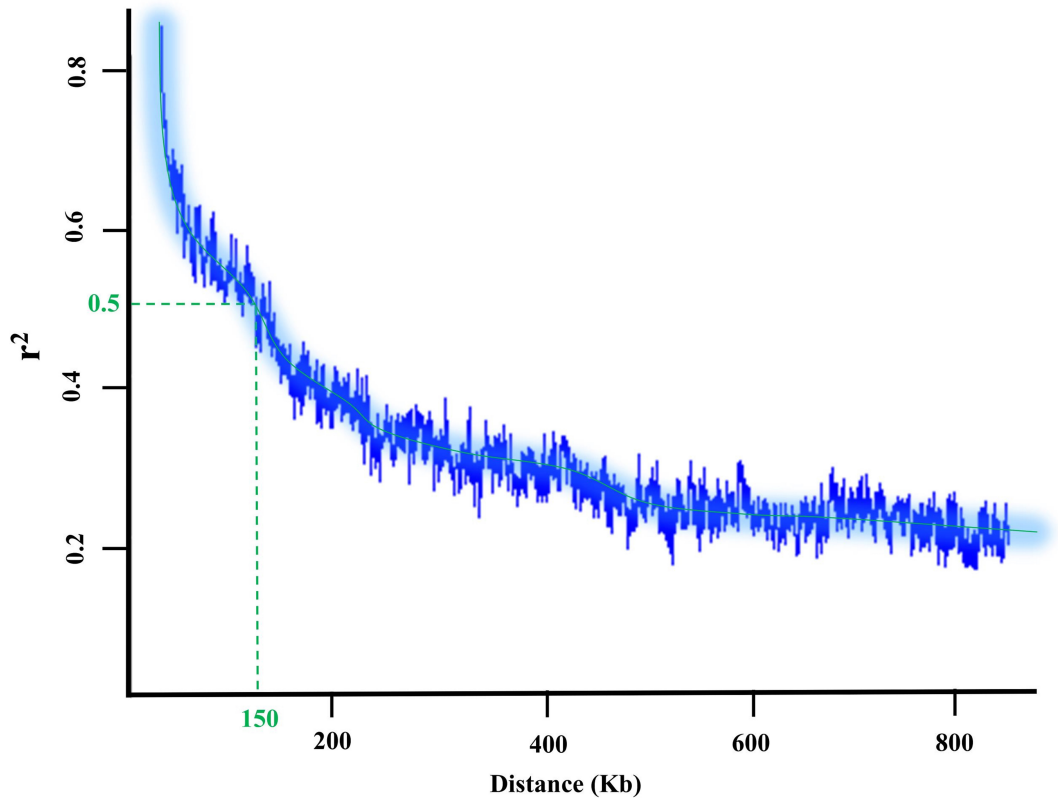

FIGURE 1 | LD decay plot and the flanking regions of each detected SNP in 227 soybean genotypes.

Extracting Candidate Genes Underlying Detected Quantitative Trait Locis

The potential candidate genes for the tested GWAS methods were extracted from the *Glycine max* William 82 reference gene models 2.0 using the SoyBase database[1]. The flanking regions of associated peak SNPs with the trait of interest were obtained based on the LD decay distance (**Figure 1**). Also, Gene Ontology (GO) enrichment analysis (see text footnote 1), and previous studies were used to detect genes associated with the trait of interest and investigate their system biology functions for each trait. Finally, the Electronic Fluorescent Pictograph (eFP) browser for soybean[2] and transcriptomics data from Severin et al. (2010) were included to generate further information about the candidate genes, including developmental- and tissue-stage dependent gene expression levels.

Data Integration Strategy

In this study, the hierarchical data integration strategy was used to accommodate the using the of HypWAS results as prior knowledge for GWAS analyses. By using a hierarchical data integration strategy, genomic regions that are directly associated with the selected hyperspectral reflectance bands and indirectly related to the overall soybean yield can be detected. Afterward, the associated candidate genes with the tested hyperspectral reflectance bands can be identified based on the results of the GWAS analysis (**Figure 2**).

[1]https://www.soybase.org

[2]www.bar.utoronto.ca

RESULTS

Yield Statistics and Hyperspectral Reflectance Profile

The average seed yield for all soybean genotypes evaluated in four environments ranged from 2.6 to 5.7 t ha^{-1} with a standard deviation and mean of 0.57 and 4.22 t ha^{-1}, respectively. The results of analysis of variance for yield as well as heritability are represented in **Supplementary Table 1**. Overall, the heritability of yield in the tested panel was 0.24. The complete hyperspectral reflectance profile measured for the tested soybean panel is presented in **Figure 3**. The visible and NIR regions showed the highest variation among the genotypes with a range of 0.40 and 0.56, respectively, whereas the red-edge region had the lowest variation among all the hyperspectral regions with a range of 0.11 (**Figure 3**). Within the visible region, the highest variations were present in the green, and red regions ranged from 0.12 and 0.13, respectively, while other reflectance bands in the visible region had lower variations (**Figure 3**). The reflectance bands greater than 770 nm showed larger variations among soybean genotypes compared

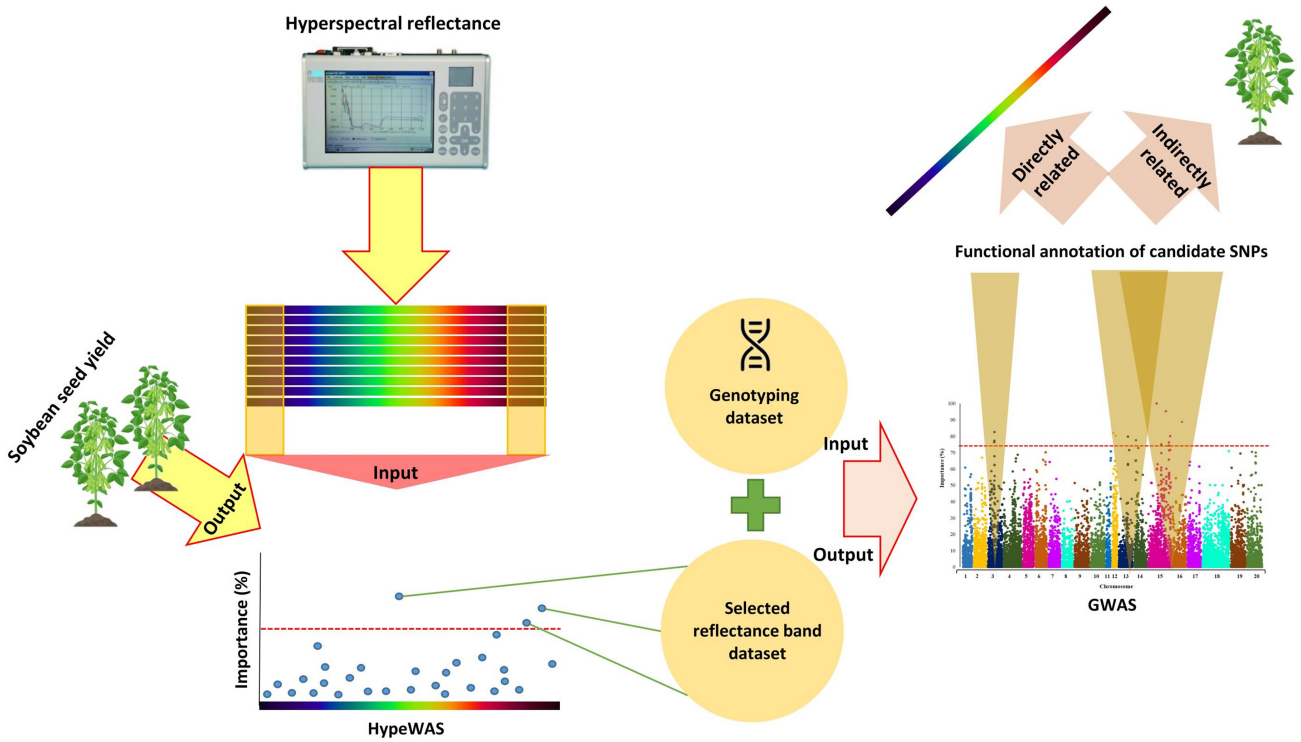

FIGURE 2 | The schematic workflow of the hierarchical data integration strategy of HypWAS and GWAS in Soybean.

to reflectance bands in the red-edge region (**Figure 3**). The results of analysis of variances for each selected hyperspectral reflectance band are presented in **Supplementary Tables 2–6**. Among all the tested hyperspectral reflectance bands, 660 and 730 nm had the highest and lowest heritability with values of 0.85 and 0.28, respectively (**Supplementary Tables 2–6**). All the selected reflectance bands showed a significant difference among genotypes.

Hyperspectral Wide Association Study

The association between reflectance bands and soybean seed yield was assessed using the RFE method. Among all the 62 reflectance bands, five were found to be significantly associated with the soybean yield (**Figure 4**). The selected reflectance bands were 390, 550, 660, 730, and 820 nm with the importance score of 99, 29, 94, 38, and 41%, respectively. Based on the number of important reflectance bands, the visible region was the most informative reflectance region associated with soybean yield by having three out of five important bands, namely, 390, 550, and 660 nm (**Figure 4**). Among all the reflectance bands located in the red-edge and NIR regions, the 730 and 820 nm bands were associated with soybean seed yield in the red-edge and NIR regions, respectively (**Figure 4**). Only the selected reflectance bands were chosen for further analysis. As it can be seen in **Figure 5**, all selected reflectance bands had a normal distribution. The 390 nm, as the hyperspectral reflectance band with the highest importance score, had a mean of 0.1 in the tested panel across different environments. The second hyperspectral reflectance band with high importance score was the 660 nm band with a mean of 0.05 (**Figure 5**). As the lowest importance score among all the selected hyperspectral reflectance bands, the 550 nm band had the mean and standard deviation of 0.24 and 0.02, respectively (**Figure 5**).

Pearson Product-Moment Correlation coefficients of all the selected reflectance bands with soybean yield were estimated and are shown in **Figure 6**. The only positive correlation with yield was found in 820 nm band with the $r = 0.19$ (**Figure 6**). The highest negative correlation with yield was observed in the 660 nm band with a correlation coefficient (r) of −0.80, and the lowest negative correlation was found between the 730 nm band and seed yield with $r = -0.16$ (**Figure 6**).

Genotyping Evaluations

High-quality SNPs were obtained for the tested GWAS panel from 210M single-end Ion Torrent reads that were proceeded with Fast-GBS.v2. After the filtering process, 17,958 out of 40,712 SNPs were detected as polymorphic and then mapped onto 20 soybean chromosomes. In the tested GWAS panel, the maximum number of SNPs was 1780 on chromosome 18, and the minimum number of SNPs was 403 on chromosome 11. The average number of SNPs was 898 across all the 20 chromosomes, with the mean density of one SNP for every 0.12 cM across the genome. As illustrated in **Figure 7A**, the tested soybean panel was composed of four to seven subpopulations. Therefore, the $K = 7$ was used as the optimum K for the tested association panel. Also, the

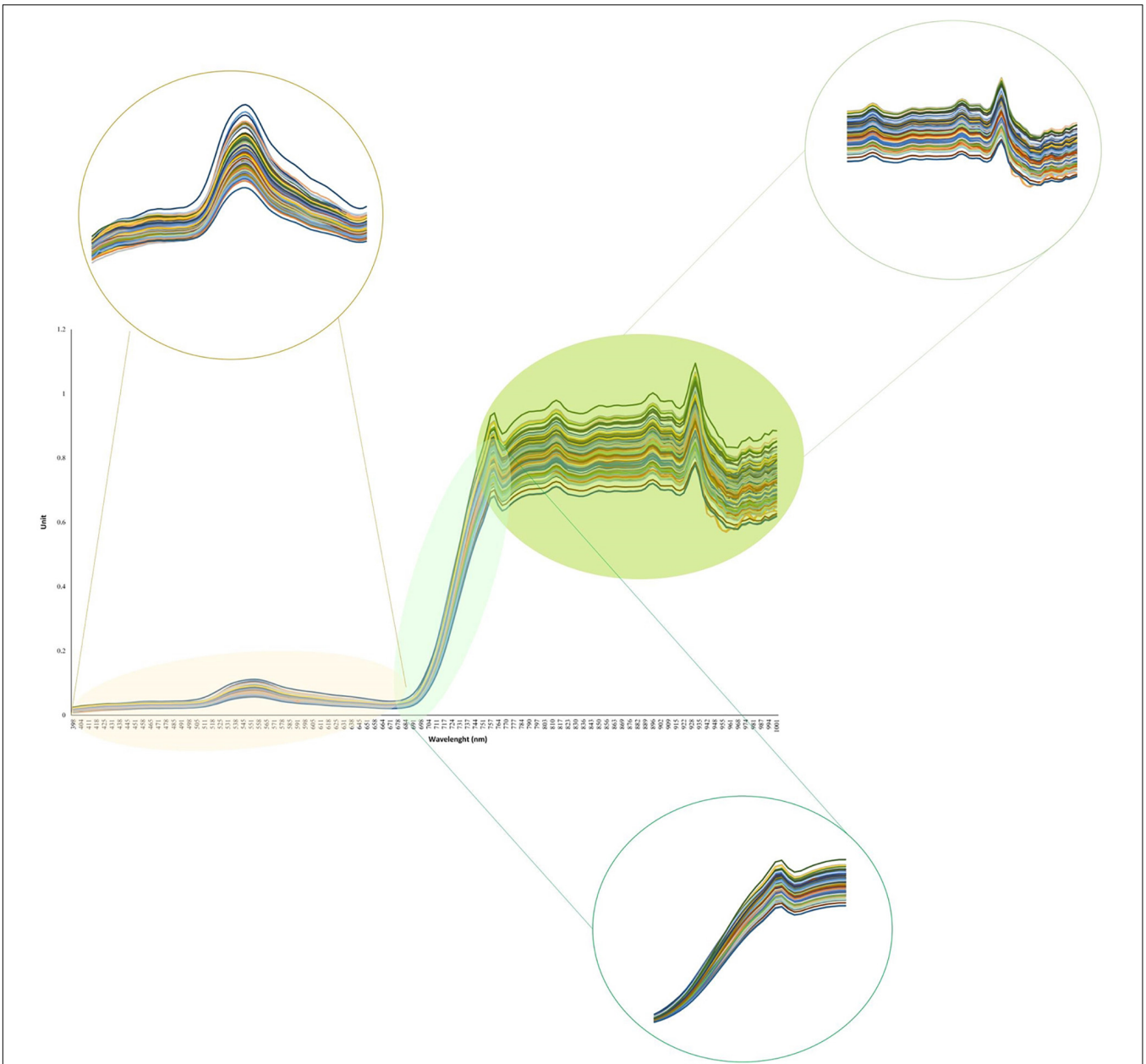

FIGURE 3 | The hyperspectral reflectance profile of the tested 227 soybean genotypes.

kinship was calculated between soybean genotypes to reduce the confounding effect (**Figure 7B**).

Genome-Wide Association Studies Analysis

In this study, we conducted GWAS analysis for dissecting the genetic control of the yield-related reflectance and identified SNPs that are linked to selected bands from HypWAS. According to the GWAS analysis of the 390 nm band, using the MLM method, 10 SNPs located on chromosomes 3, 6, and 10 were found to be associated with this band (**Figure 8**). Using FarmCPU, 13 SNPs on chromosomes 3, 6, 10, and 13 were identified, and exploiting the SVR-mediated GWAS, 12 SNPs on chromosomes 2, 3, 6, 9, 15, 16, and 20 were found to be linked to this band (**Figure 8**). Based on the results, chromosomes 3 and 6 are two chromosomes that were found associated with the 390 nm band using all three GWAS methods. Most of the detected QTL on chromosomes 3 and 6 were co-localized with previously reported QTL such as ureide content, resistance to *Phytophthora sojae*, internode length, and pubescence color (**Table 1**).

Genome-wide association studies analyses on the 550 nm band resulted in discovering 5 and 10 SNPs to be associated with this band using the MLM and the FarmCPU methods, respectively.

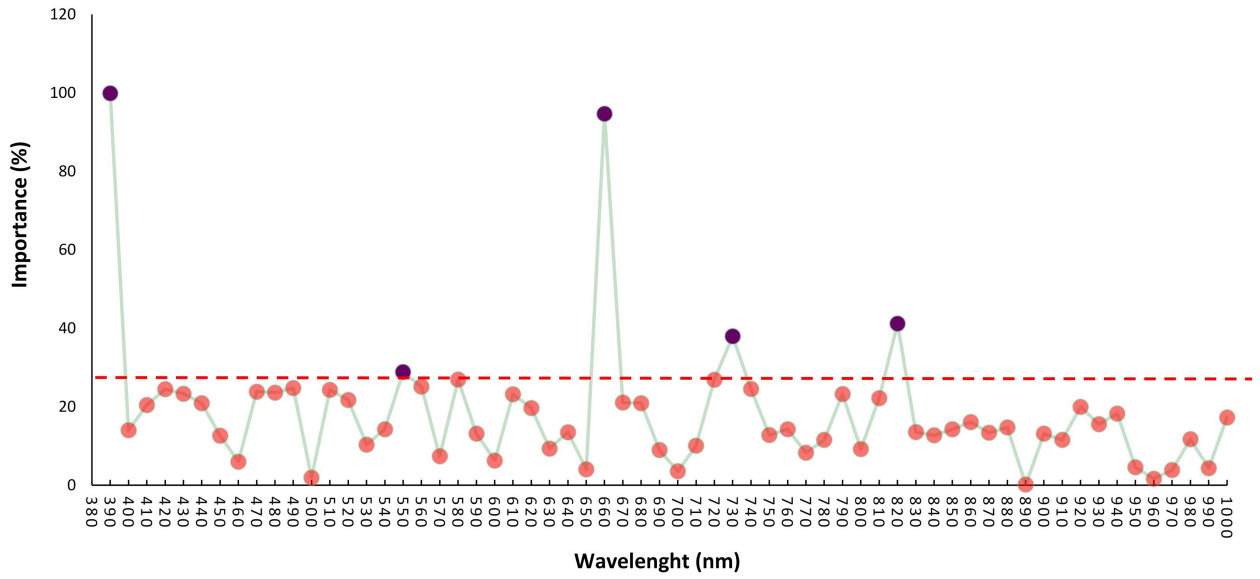

FIGURE 4 | Hyperspectral wide association analysis (HypWAS) of yield in the tested soybean panel. The threshold for significant importance values was indicated as the red dash line.

Among all the associated SNPs using the MLM method, four of them were located on chromosome 5, and one on chromosome 3. The 10 SNPs identified using the FarmCPU method were located on chromosomes 1, 3, 5, and 13 (**Figure 8**). Using SVR-mediated GWAS analysis, we identified nine SNPs linked to the 550 nm band, of which two were located on chromosome 6, three on chromosome 9, and one SNP on each of the chromosomes 5, 15, 18, and 19, respectively (**Figure 8**). Comparing the results of the three GWAS methods, chromosome 5 was consistently found to be associated with the 550 nm band. Most of the detected QTL on chromosome 5 were co-localized with previously reported QTL such as oil-related traits, water use efficiency, and full maturity (**Table 2**).

According to **Figure 8**, a total of 9, 14, and 15 associated SNPs were detected using MLM, FarmCPU, and SVR, respectively. Out of 9 detected SNPs using the MLM method, four SNPs were located on chromosome 3, three SNPs were located on chromosome 7, and two SNPs were located on chromosome 18 (**Figure 8**). All the detected SNPs using FarmCPU were located on chromosomes 3, 6, and 18 (**Figure 8**). GWAS analysis of the 660 nm band using the SVR-mediated GWAS method detected 13 SNPs on chromosome 15 and one SNP on chromosomes 14 and 19 (**Figure 8**). Most of the detect QTL for the 660 nm band were related to first flower, water use efficiency, soybean cyst nematode resistance, seed yield, pod number, and plant height (**Table 3**).

Genome-wide association studies analyses on the 730 nm band resulted in discovering 10, 15, and 10 associated SNPs using MLM, FarmCPU, and SVR, respectively (**Figure 8**). Using MLM, 10 SNPs were located on chromosomes 1, 2, 5, 14, 15, 17, and 18 (**Figure 8**). By using the FarmCPU method, five associated SNPs with the 730 nm band were located on chromosome 5, three SNPs were located on chromosome 2, two SNPs were on chromosomes 1 and 18, and one SNP was located on chromosomes 14, 15, and 17 (**Figure 8**). Using SVR-mediated GWAS, 10 SNPs were located on chromosomes 1, 4, 6, 9, and 10 (**Figure 8**). Based on these results, chromosome 1 was unanimously determined to be associated with the 730 nm band by all three GWAS methods. Most of the detected QTL for the 730 nm band were related to water use efficiency, seed oil and protein-related traits, reproductive stage length, and pod number (**Table 4**).

Using MLM, FarmCPU, and SVR methods for GWAS analysis of the 820 nm band, a total of four, five, and seven SNPs were detected to be associated with this reflectance band. The associated SNPs were located on each of the chromosomes 1, 4, 6, and 16 using MLM (**Figure 8**). Using the FarmCPU method, two associated SNPs were found on chromosome 16, and one was found on chromosomes 1, 5, and 6, respectively (**Figure 8**). Two associated SNPs using the SVR-mediated GWAS method were located on chromosome 1, and one SNP was found on chromosomes 2, 4, 6, 10, and 16 (**Figure 8**). Chromosomes 1, 6, and 16 were selected as the commonly detected chromosomes among all the tested GWAS methods associated with the 820 nm band. Most of the detected QTL for the 820 nm band were related to first flower, soybean cyst nematode resistance, water use efficiency, seed set, seed long-chain fatty acid, and seed width to height (**Table 5**).

Extracting Candidate Genes Underlying Detected Quantitative Trait Locis

The flanking regions of the QTL were determined within the 150-kbp upstream and downstream of each peak SNP for a given QTL, and these regions were searched for identifying potential candidate genes associated with the target bands (**Figure 1**). For the 390 nm band, five

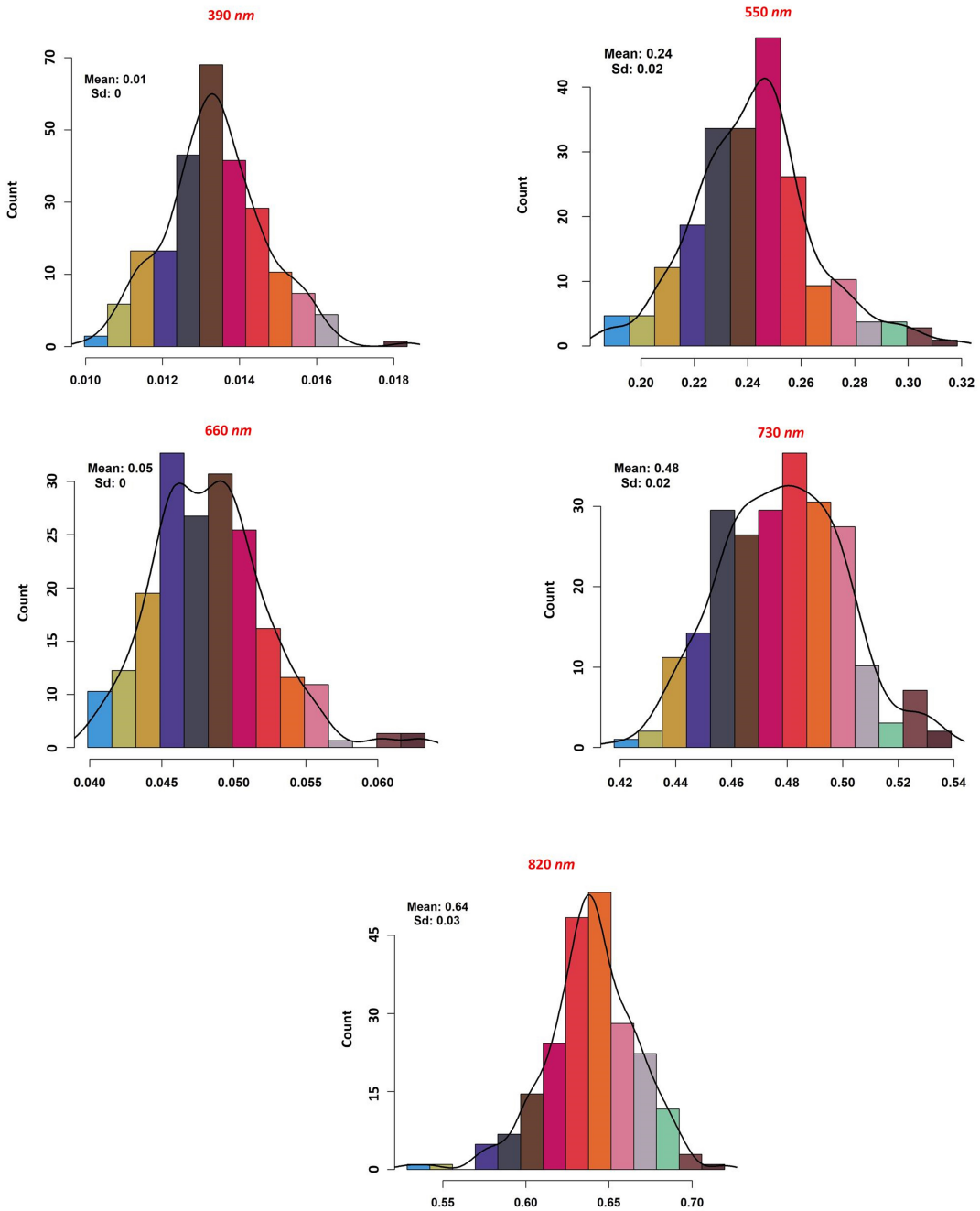

FIGURE 5 | The distribution, mean and standard deviation of the selected hyperspectral reflectance bands in the tested soybean panel.

peak SNPs (Chr2_858458, Chr3_3729245, Chr9_39344365, Chr9_39344439, and Chr16_7313753) had the highest allelic effects (**Figure 9A**). Based on the gene annotation and expression data, the following genes were identified as selected candidates governing the 390 nm band: *Glyma.02G008900* (GO:0006979), *Glyma.02G008700* (GO:0006970), *Glyma.03G033100* (GO:0019748), *Glyma.03G033100* (GO:0031347), *Glyma.09G168700* (GO:0009624 and GO:0042742), and *Glyma.16G073100* (GO:0050660 and GO:0016491). These genes have been annotated for oxidative and osmotic stresses, secondary metabolic process, regulation of defense response, response to nematode, defense response to bacterium, flavin adenine dinucleotide binding, and oxidoreductase activity, respectively. For the 550 nm band, two peak SNPs (Chr3_22283256 and Chr5_40467080) had the highest allelic effects compared to other detected peak SNPs (**Figure 9B**). There was no previously reported QTL linked with Chr3_44326068, while three QTL (water use efficiency and shoot macro- and

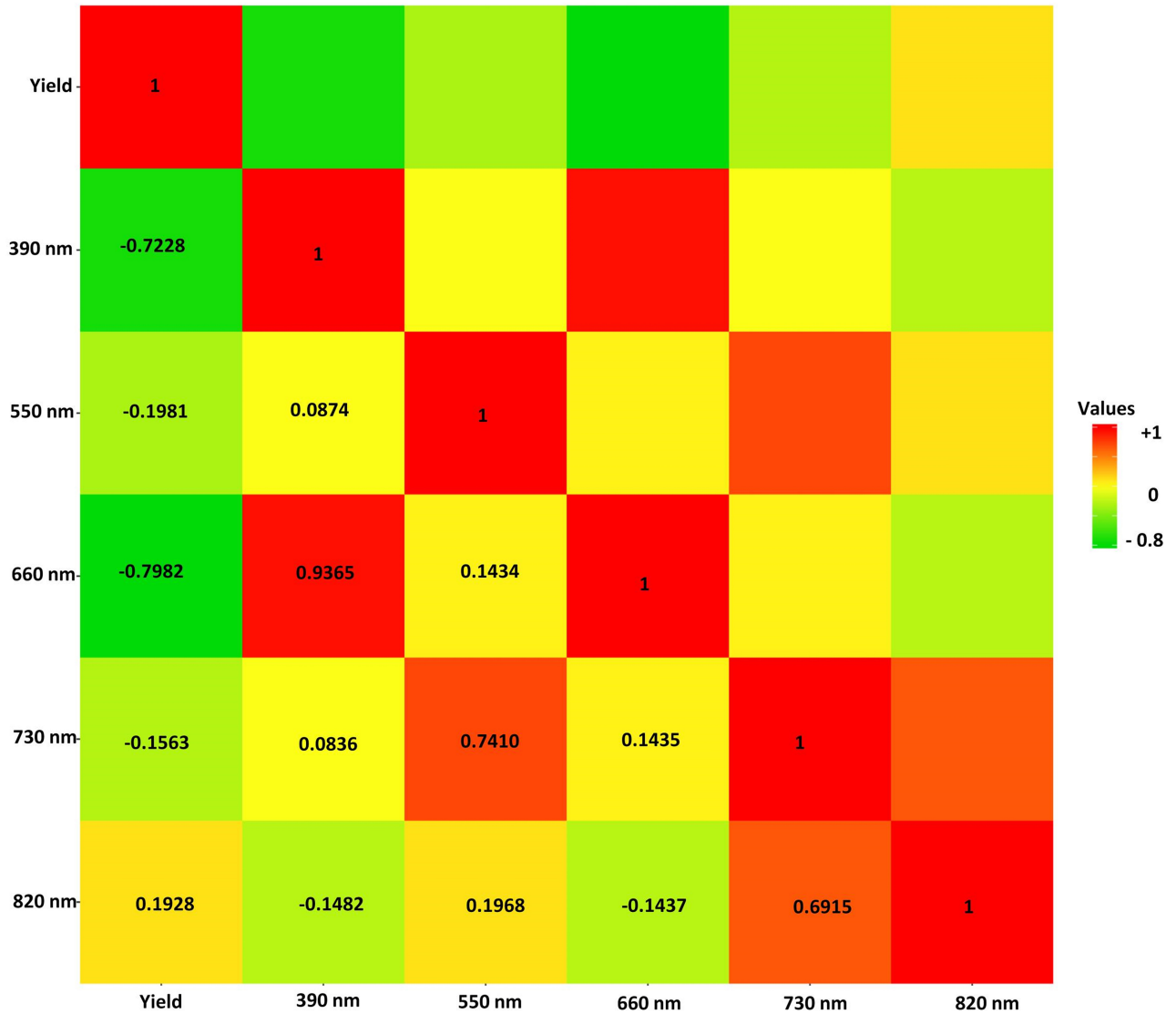

FIGURE 6 | The Pearson correlation of the selected hyperspectral reflectance bands in the tested soybean panel. The heat map scale for values is provided by color for the panel.

micronutrient concentrations) were linked to Chr5_40467080. The candidate genes *Glyma.03G081700* (GO:0010224) and *Glyma.05G226000* (GO:0009411, GO:0009813) were selected candidates for the 550 nm band, which encode response to UV-B and to UV, and flavonoid biosynthetic process, respectively. Based on **Figure 9C**, the highest allelic effect for the 660 nm band was found in one peak SNP (Chr19_36064225) detected by the SVR-mediated GWAS, whereas no previously reported QTL was linked to the detected peak SNP position. Based on the gene ontology analysis, *Glyma.19G108200* (GO:0009911, GO:0048573, and GO:0009909) was the selected candidate gene for the 660 nm band, which encodes positive regulation of flower development, photosynthesis, flowering, light reaction, and regulation of flower development. The allelic effect analysis of the 730 nm band indicated a high allelic effect for two peak SNPs at Chr2_42645195 and Chr18_6728432 (**Figure 9D**). The selected peak SNPs were linked to water use efficiency and seed protein contents (**Figure 9D**). There were two selected candidate genes [*Glyma.02G237700* (GO:0009853) and *Glyma.02G237400* (GO:0031347)] associated with the 730 nm band, which encode photorespiration and regulation of defense response. On the 820 nm band, the highest allelic effect was found in two peak SNPs (Chr4_20000416 and Chr6_11029779) that were detected by the SVR-mediated GWAS method (**Figure 9E**). Gene ontology analysis of the selected peak SNPs identified two candidate genes, Glyma.04G135300 (GO:0009658 and GO:0009055) and Glyma.10G033600 (GO:0015250 and GO:0005215), which encode chloroplast organization, electron carrier activity, water channel activity, and transporter activity, respectively.

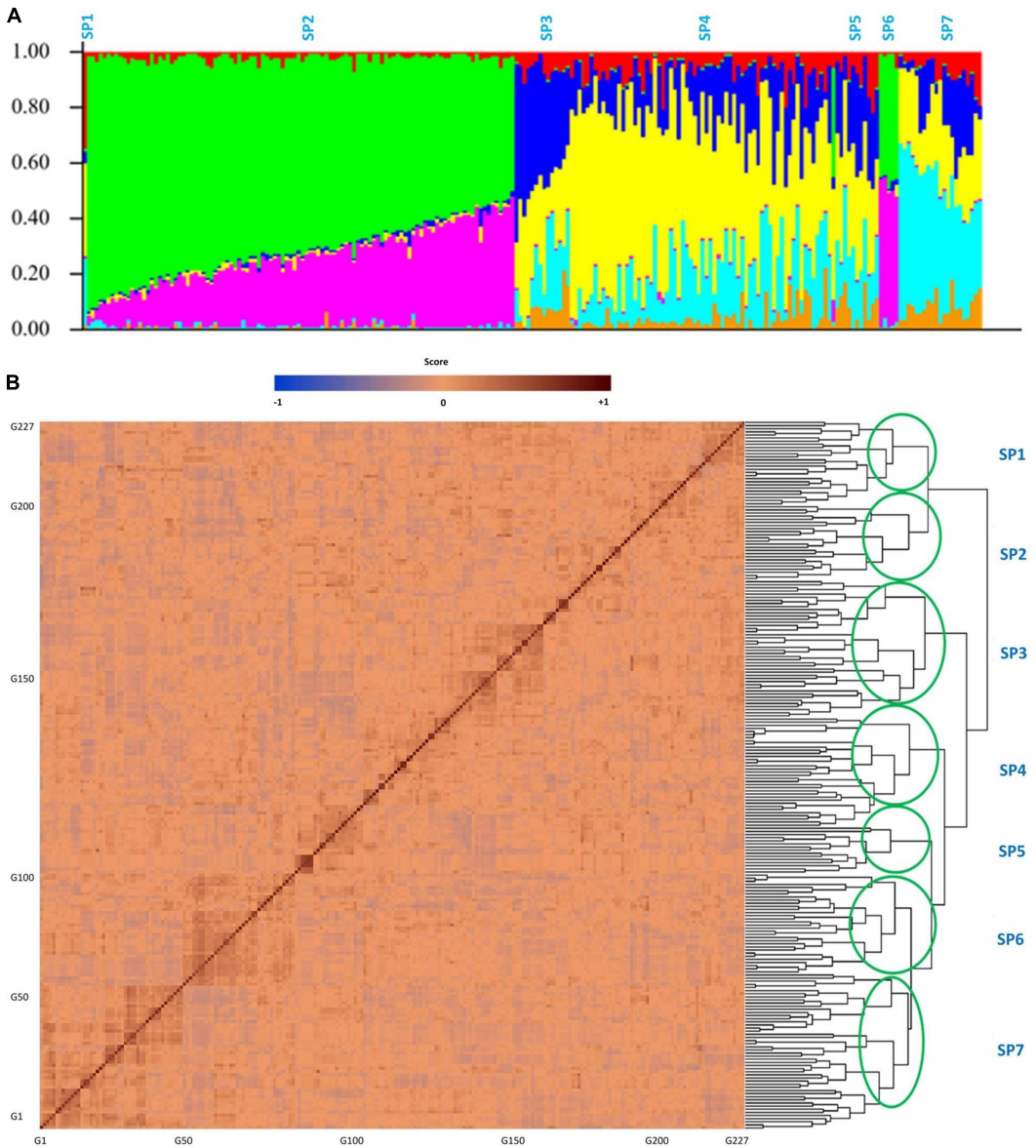

FIGURE 7 | Structure **(A)** and kinship **(B)** plots for the 227 soybean genotypes. The *x*-axis represents the number of genotypes used in this GWAS panel, and the *y* axis represents the membership of each subgroup. SP1–SP7 stands for subpopulations 1 to 7.

DISCUSSION

Accurate predictions of the final performance of germplasm at early growth stages is paramount for breeders to make early selection decisions in germplasm advancement and cultivar development programs (Maimaitijiang et al., 2020). The end-season selection based on yield *per se* may result in overlooking other factors such as selecting for other yield-unrelated traits and environmental factors (Yoosefzadeh Najafabadi, 2021). Yet, trait measurement in large breeding nurseries and working with large phenotypic and genotypic data is still a bottleneck in the genome-to-phenome analysis process (Parmley et al., 2019). In recent years, the combination of high throughput phenotyping and genotyping tools has greatly accelerated the plant breeding progress (Yang et al., 2020). However, merging different omics datasets for better characterization of the complex traits extensively depends on an appropriate selection of a data integration strategy (Picard et al., 2021). Extracting the biological information from the secondary related traits that are in high correlation with the trait of interest, detecting the associated SNPs, and identifying candidate genes for each detected peak SNPs provided a valuable complement for understanding the biological mechanism of a trait of interest. The hierarchical strategy is based on including prior knowledge of relationships between different omics layers (Wang et al., 2013; Picard et al., 2021). In this study, we used this strategy in a soybean

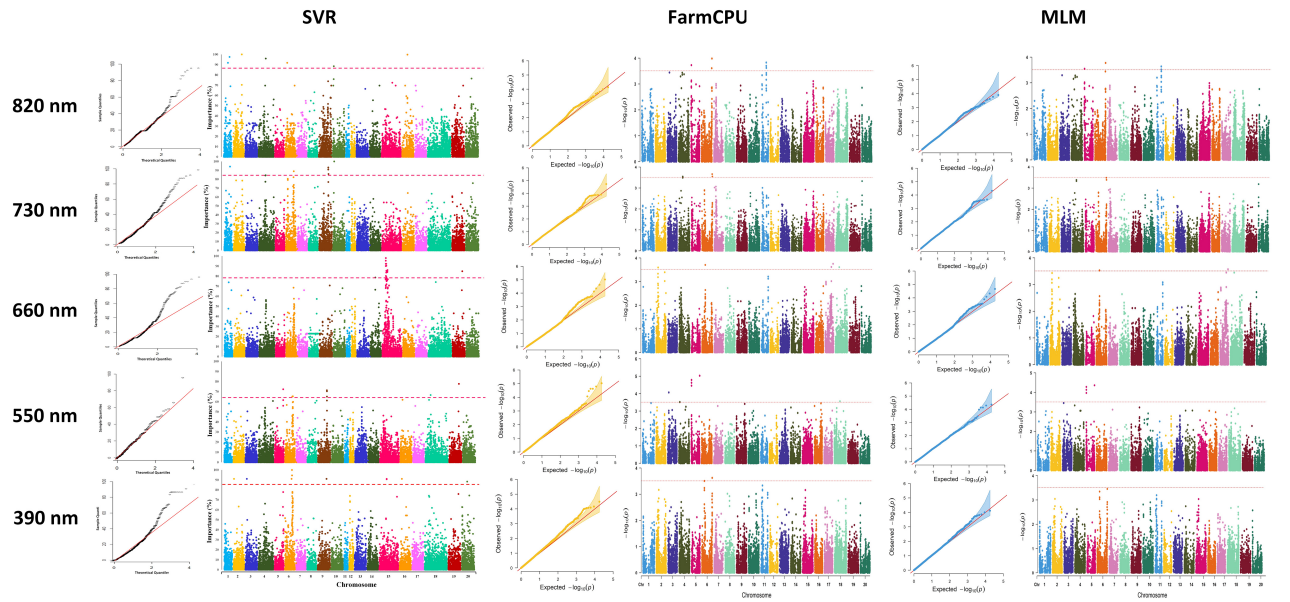

FIGURE 8 | Genome-wide Manhattan plots for GWAS studies of the selected hyperspectral reflectance bands using MLM, FarmCPU, and SVR methods in the tested soybean panel. MLM, mixed linear model; FarmCPU, fixed and random model circulating probability unification; SVR, fixed and random model circulating probability unification.

hyperspectral reflectance-yield association study to acquire a better understanding of the genetic architecture of soybean yield using GWAS.

As one of the important high-throughput phenotyping tools, hyperspectral reflectance has provided plant breeders with an efficient plant evaluation strategy in a large population at early growth stages for important agronomic traits (Yoosefzadeh-Najafabadi et al., 2021a). The use of spectral reflectance for predicting crop yield has been extensively investigated. For example, Qiao et al. (2021) reported the efficiency of using long-time series multi-spectral images for yield mapping of different crop species using an automated spatial-spectral feature extractor. The application of hyperspectral reflectance in predicting the wheat grain yield was studied by Fei et al. (2021), who reported the effectiveness of red and NIR regions in predicting the grain yield in different irrigation regimes. The use of hyperspectral reflectance in predicting yield was not limited to agronomy crops and used for vegetables (Awika et al., 2021), trees (Ali and Imran, 2021), and industrial plants (Holmes et al., 2020). Therefore, breeders can use spectral data to establish phenome-to-genome relationships by performing HypWAS and applying it via phenomics selection. Genomic selection methods suffer from the lack of consideration for environmental effects (Zhong et al., 2009; Tong and Nikoloski, 2021). However, by using HypWAS, both environmental and genetic effects can be considered in the final decision. The proposed idea of HypWAS is novel, and, to our best knowledge, there is no example in the literature.

The rationale behind the HypWAS is to increase the efficiency of indirect selection for breeders and to offer them a strategy to reduce noise and possible errors from the hyperspectral reflectance data that is used in genetic studies. HypWAS makes it possible to select reflectance bands with high importance scores for complex traits such as yield. Therefore, plant breeders and geneticists can investigate more aspects of the selected reflectance bands to find the genomic regions, gene candidates, and physiological processes behind each reflectance band. Previously, we implemented one of the most common variable selection methods, RFE, to select important reflectance bands in association with soybean yield production (Yoosefzadeh-Najafabadi et al., 2021a). However, there was no further information about the genetic background of the selected reflectance bands. In this study, we selected the five most important reflectance bands and we used HypWAS to investigate various aspects of the physiological and genetic background of each reflectance band.

Among the five selected reflectance bands, three of them were located in the visible range of the spectrum. Most of the reflection in the visible region is significantly dominated by the foliar pigments' absorption (Hennessy et al., 2020). Chlorophyll a and b have a stronger reflection in comparison with other foliar pigments in the visible region (Fernandes et al., 2013). Chlorophylls play a major role in photosynthesis due to all the photosynthesis structures, such as the antenna systems of photosystem I and photosystem II (Pettai et al., 2005; Hennessy et al., 2020). Light energy absorbed by antenna systems of photosystem I and II, is then rapidly transferred to the respective reaction centers (Pettai et al., 2005). Several studies reported the strong correlation of the 680–700 nm bands with photosystem I and II (Ke, 2001; Pettai et al., 2005; Hoa et al., 2017; Hennessy et al., 2020). In this study, the 680 nm band was selected as one of the high-importance reflectance bands that explains the total soybean seed yield. This can reveal the importance of

TABLE 1 | The list of detected QTLs for 390 nm using different GWAS methods in the tested soybean panel.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment[a]	References
MLM	3	44121448	Ureide content 1-g13.1	NA	Ray et al., 2015
			Ureide content 1-g13.2	NA	Ray et al., 2015
			Ureide content 1-g13.3	NA	Ray et al., 2015
			Ureide content 1-g13.4	NA	Ray et al., 2015
			Ureide content 1-g13.5	NA	Ray et al., 2015
		44198005		NA	
		44326068		NA	
		44328629		NA	
		45039908		NA	
	6	11953156		NA	
		11953223		NA	
	10	2906452		NA	
		2906459		NA	
		2906496		NA	
FarmCPU	3	44121448	Ureide content 1-g13.1	NA	Ray et al., 2015
			Ureide content 1-g13.2	NA	Ray et al., 2015
			Ureide content 1-g13.3	NA	Ray et al., 2015
		44198005	Ureide content 1-g13.4	NA	Ray et al., 2015
			Ureide content 1-g13.5	NA	Ray et al., 2015
		44326068		NA	
		44328629		NA	
		45039908	Internode length 1-g3	NA	Fang et al., 2017
			Pubescence color 2-g1.1	NA	Fang et al., 2017
	6	11953223		NA	
		11953223		NA	
	10	2906452		NA	
		2906459		NA	
		2906496		NA	
		50548092		NA	
	13	30231980	Seed Yield 3-g5	NA	Contreras-Soto et al., 2017
		30232000	seed weight 6-g2	2	Sonah et al., 2015
			Phytoph 3-g22	2	Chang et al., 2016
			Phytoph 2-g4	2	Qin et al., 2017
			Phytoph 2-g5	2	Qin et al., 2017
			Phytoph 3-g23	2	Chang et al., 2016
			Leaflet width 1-g1	2	Fang et al., 2017
SVR	2	858431		NA	
		858458		NA	
	3	3729245	Phytoph 3-g1	NA	Chang et al., 2016
			Phytoph 3-g2	NA	Chang et al., 2016
			Phytoph 3-g3	NA	Chang et al., 2016
			Phytoph 3-g4	NA	Chang et al., 2016
			Phytoph 3-g5	NA	Chang et al., 2016
			Phytoph 3-g6	NA	Chang et al., 2016
	6	38679976		NA	
		38787915		NA	
		36746367		NA	
	9	39344365	pod number 1-g4.1	NA	Fang et al., 2017
			pod number 1-g4.2	NA	Fang et al., 2017
		39344439	pod number 1-g4.3	NA	Fang et al., 2017
			seed thickness 2-g4	NA	Fang et al., 2017
	15	22232022		NA	
		22231908		NA	
	16	7313753	Sclero 3-g60	NA	Moellers et al., 2017
			ureide content 1-g43	NA	Ray et al., 2015
	20	22656999		NA	

[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 QTLs for 550 nm using different GWAS methods in the tested soybean panel.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment[a]	References
MLM	3	22283256		NA	
	5	1115619	seed palmitic 4-g1	2	Zhang et al., 2018
			seed palmitic 4-g1.2	2	Zhang et al., 2018
			Seed long chain fatty acid 1-g21.2	2	Fang et al., 2017
			Seed long chain fatty acid 1-g19.2	2	Fang et al., 2017
		1115673	Seed long chain fatty acid 1-g14.1	2	Fang et al., 2017
			Seed long chain fatty acid 1-g14.2	2	Fang et al., 2017
		1115689	Seed long chain fatty acid 1-g1.2	2	Fang et al., 2017
			seed stearic 3-g1	2	Zhang et al., 2018
		41767154	Phytoph 2-g28	2	Qin et al., 2017
			First flower 4-g18	2	Mao et al., 2017
			Seed oil 4-g18	2	Bandillo et al., 2015
			Seed oil 4-g17	2	Bandillo et al., 2015
			Seed oil 4-g16	2	Bandillo et al., 2015
			Seed oil 6-g1	2	Cao et al., 2017
			Seed oil 6-g16	2	Cao et al., 2017
			R8 full maturity 5-g1	2	Fang et al., 2017
			Seed oil 4-g15	2	Bandillo et al., 2015
			Seed oil 4-g14	2	Bandillo et al., 2015
			Seed oil 6-g2	2	Cao et al., 2017
			Seed oil 6-g4	2	Cao et al., 2017
FarmCPU	1	41281098			
	3	22283184			
		22283256			
		44326068		4	
	5	1115619	Seed palmitic 4-g1	2	Zhang et al., 2018
			Seed palmitic 2-g1.2	2	Fang et al., 2017
			seed long chain fatty acid 1-g21.2	2	Fang et al., 2017
		1115673	seed long chain fatty acid 1-g19.2	2	Fang et al., 2017
		1115673	seed long chain fatty acid 1-g14.1	2	Fang et al., 2017
		1115689	seed long chain fatty acid 1-g14.2	2	Fang et al., 2017
		1153958	seed long chain fatty acid 1-g1.2	2	Fang et al., 2017
		1115689	seed stearic 3-g1	2	Zhang et al., 2018
		41767154	First flower 4-g18	2	Mao et al., 2017
			seed oil4-g18	2	Bandillo et al., 2015
			seed oil 4-g17	2	Bandillo et al., 2015
			seed oil 4-g16	2	Bandillo et al., 2015
			seed oil 6-g1	2	Cao et al., 2017
			seed linolenic 4-g5	2	Li et al., 2015
			R8 full maturity 5-g1	2	Fang et al., 2017
			seed oil 4-g15	2	Bandillo et al., 2015
			seed oil 4-g14	2	Bandillo et al., 2015
			seed oil 6-g2	2	Cao et al., 2017
			seed oil 8-g4	2	Zhang et al., 2018
	13	29190399	Phytoph 2-g30	2	Qin et al., 2017
			seed set 1-g29-.2	2	Fang et al., 2017
			seed weight 13-g8	2	Wang J. et al., 2016
SVR	5	40467080	WUE 2-g13	3	Kaler et al., 2017
			Shoot p 1-g10.2	3	Dhanapal et al., 2018
			shoot p 1-g10.1	3	Dhanapal et al., 2018
	6	41477920		4	
		38960003		NA	

(Continued)

TABLE 2 | (Continued)

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment[a]	References
	9	39366957	Pod number 1-g4.1	1	Fang et al., 2017
			Pod number 1-g4.2	1	Fang et al., 2017
			Pod number 1-g4.3	1	Fang et al., 2017
			Seed thickness 2-g4	1	Fang et al., 2017
		39372117	Seed Thr 2-g1	1	Li et al., 2018
			Seed Sar 2-g1	1	Li et al., 2018
			Seed Tyr 2-g1	1	Li et al., 2018
		39659468	Seed Lys 2-g1	NA	Li et al., 2018
			Seed Leu 2-g1	NA	Li et al., 2018
			Seed Llu 2-g1	NA	Li et al., 2018
			Seed Ala 2-g1	NA	Li et al., 2018
			Seed Gly 2-g1	NA	Li et al., 2018
	15	11293240	Seed protein 7-g14	NA	Zhang et al., 2017
	18	11166966		NA	
	19	36064225		NA	

[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.

photosynthesis components in determining the overall soybean yield (Yoosefzadeh-Najafabadi et al., 2021c). The 660 nm band can be used to screen a large population of plants in a short time for selecting genotypes with a high potential for photosynthesis activity. Similarly, another important reflectance band in this study (390 nm) is located in the blue region of the spectrum, which is highly correlated with the chlorophyll content (Richter et al., 2016; Hennessy et al., 2020). The 390 nm band had the highest importance value in predicting the final soybean yield among all the selected reflectance bands, confirming the importance of photosynthesis in the final soybean yield formation. The third detected reflectance band in this study (550 nm), represents the green peak in the visible spectrum (Stommel et al., 2009; Liu et al., 2018). This reflectance band was reported to be correlated with Chlorophyll and Anthocyanin contents in plants (Hennessy et al., 2020). Anthocyanins are known as a diverse class of flavonoid components that play a significant role in protecting plants against abiotic and biotic stresses (Gitelson et al., 2001; Hennessy et al., 2020). Since most of the soybean fields in North America are grown in rainfed areas, water deficit stress would be inevitable during the soybean growing season in these areas.

Many studies reported the strong correlation between red-edge and near-infrared regions and water content as well as spongy mesophyll conditions in plants. Based on the HypWAS analysis, the 730 and 820 nm bands were selected as the important reflectance bands in predicting the overall soybean yield. The 730 nm band is located in the red-edge region and correlated with the leaf water content, chlorophyll concentration, and leaf layering (Horler et al., 1983). The 820 nm band is located in the near-infrared region and reflects the spongy mesophyll condition in leaves (Gao et al., 2014; Salvatori et al., 2015). The major role of spongy mesophyll in plants is to interchange the required CO_2 for photosynthesis (Veromann-Jürgenson et al., 2020). All spongy mesophylls are covered by a thin layer of water, hence environmental stresses can significantly affect mesophylls resulting in a reduced photosynthesis activity level in plants (Veromann-Jürgenson et al., 2020; Liu M. et al., 2021). Therefore, changes in the water and gas level in mesophylls is the first sign of detecting stresses in plants, so the difference between the reflectance of the 730 nm as well as the 820 nm band in normal and stress conditions can be considered as a measurement for abiotic and biotic stresses (Momayyezi et al., 2020). Overall, adjusting breeding selection criteria based on the selected reflectance bands might lead to select a genotype with significant levels of tolerance against stresses and high photosynthesis activity. This can be done by measuring those reflectance bands by remote sensing tools in a short time in a less labor-intensive manner. Also, understanding the genetic background of each selected reflectance band would be helpful to design appropriate genetic markers for the fast screening of genotypes.

Genome-wide association studies is currently considered as an imperative approach for discovering genomic regions associated with complex traits in diverse areas from human genetics to plant and animal breeding (Alqudah et al., 2020; Khanzadeh et al., 2020; Li et al., 2020; Tibbs Cortes et al., 2021). Insufficient statistical power is the most fundamental challenge when it comes to using conventional GWAS for characterizing quantitative traits (Nicholls et al., 2020), especially in plants with narrow genetic bases. In ML-mediated GWAS analyses, the significance levels or thresholds for identifying SNP-trait associations are estimated using variable importance methods, which are different from statistical methods that are used for estimating *p*-value in conventional GWAS (Szymczak et al., 2009). The main advantage of using variable importance, rather than *p*-values, for individual SNP-trait association tests, consists in the ability of these approaches to consider the interaction effects between SNPs (Szymczak et al., 2009; Asif et al., 2020). In other words, ML algorithms are more practical in terms of allowing for high-order interactions that are not pre-specified

TABLE 3 | The list of detected QTLs for 660 nm using different GWAS methods in the tested soybean panel.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment[a]	References
MLM	3	22283184		NA	
		22283218		NA	
		22283256		NA	
		44326068		NA	
	7	20220381		NA	
		20280812		NA	
		23349322		NA	
	18	56952847	pod number 4-g8	NA	Hao et al., 2012
			plant height 3-g14	NA	Contreras-Soto et al., 2017
		56952858	shoot K 1-g39	NA	Dhanapal et al., 2018
			WUE 3-g32	NA	Dhanapal et al., 2018
FarmCPU	3	22283184		NA	
		22283218		NA	
		22283256		NA	
		44326068		NA	
	7	20220330	Shoot Mn 1-g2	NA	Dhanapal et al., 2018
		20220381		NA	
		20280812		NA	
		23317163	Ureide content 1-g9	NA	Ray et al., 2015
		23349322	shoot Mn 1-g3	NA	Dhanapal et al., 2018
		27492738		NA	
	18	56952847	pod number 4-g8	NA	Hao et al., 2012
			plant height 3-g14	NA	Contreras-Soto et al., 2017
		56952858	shoot K 1-g39	NA	Dhanapal et al., 2018
			WUE 3-g32	NA	Dhanapal et al., 2018
		57794992		NA	
SVR	14	16425108		NA	
	15	13118545		NA	
		12895268		NA	
		12894320		NA	
		14174744		NA	
		14378690		1	
		12892003	seed coat color 3-g3	NA	Vuong et al., 2015
			seed yield, soyNAM 7-g14	NA	Diers et al., 2018
		18302021		NA	
		14406716		1	
		17877705		NA	
		17362699		NA	
		14088239		1	
		13163194	SCN 5-g33	NA	Li et al., 2016
			First flower 4-g58	NA	Mao et al., 2017
			First Flower 5-g29.1	NA	Fang et al., 2017
			First flower 5-g29.2	NA	Fang et al., 2017
			First flower 5-g29.3	NA	Fang et al., 2017
		14421366		1	
	19	36064225		1	

[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.

in the model using non-linear kernels (Sun et al., 2021). Conventional statistical methods need to have pre-identified parameters for the analysis of special traits of interest (Sun et al., 2021). Conventional statistical methods are significantly useful in the presence of inherent uncertainty, small signal-to-noise ratio, insufficient training dataset, a small number of variables, predefining the parameters involved in the variance of the trait of interest. Therefore, conventional GWAS are

TABLE 4 | The list of detected QTLs for 730 nm using different GWAS methods in the tested soybean panel.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment[a]	References
MLM	1	41281098		NA	
	2	17694706		NA	
		17694726		NA	
	5	1115689	Seed palmitic 4-g1	NA	Zhang et al., 2018
			Seed palmitic 2-g1.2	NA	Fang et al., 2017
			Seed long-chain fatty acid 1-g21.2	NA	Fang et al., 2017
		1153958	Seed long-chain fatty acid 1-g19.2	NA	Fang et al., 2017
			Seed long-chain fatty acid 1-g14.1	NA	Fang et al., 2017
			Seed long-chain fatty acid 1-g14.2	NA	Fang et al., 2017
			Seed long-chain fatty acid 1-g1.2	NA	Fang et al., 2017
			Seed stearic 3-g1	NA	Zhang et al., 2018
	14	2259506	Reproductive stage length 1-g3.1	NA	Fang et al., 2017
			Reproductive stage length 1-g3.2	NA	Fang et al., 2017
	15	9020829		NA	
	17	31797213		NA	
	18	6686269		NA	
		6728432	Seed protein 7-g27	NA	Zhang et al., 2017
FarmCPU	1	16343505		NA	
		41281098		NA	
	2	17694706		NA	
		17694726		NA	
		42645195	WUE 2-g6	4	Kaler et al., 2017
			WUE 3-g4	4	Dhanapal et al., 2018
			WUE 3-g5	4	Dhanapal et al., 2018
			WUE 3-g6	4	Dhanapal et al., 2018
	5	1115619	Seed palmitic 4-g1	NA	Zhang et al., 2018
			Seed palmitic 2-g1.2	NA	Fang et al., 2017
			Seed long-chain fatty acid 1-g21.2	NA	Fang et al., 2017
		1115673	Seed long-chain fatty acid 1-g19.2	NA	Fang et al., 2017
			Seed long-chain fatty acid 1-g14.1	NA	Fang et al., 2017
		1115689	Seed long-chain fatty acid 1-g14.2	NA	Fang et al., 2017
			Seed long-chain fatty acid 1-g1.2	NA	Fang et al., 2017
		1153958	Seed stearic 3-g1	NA	Zhang et al., 2018
		1467115	Seed palmitic 5-g2	NA	Li et al., 2015
			Seed palmitic 5-g1	NA	Li et al., 2015
			Seed palmitic 2-g1.3	NA	Fang et al., 2017
			Seed long-chain fatty acid 1-g21.3	NA	Fang et al., 2017
			Seed long-chain fatty acid 1-g19.3	NA	Fang et al., 2017
			Seed long-chain fatty acid 1-g1.3	NA	Fang et al., 2017
			Seed width to height ratio 1-g2.1	NA	Fang et al., 2017
			Seed width to height ratio 1-g2.2	NA	Fang et al., 2017
			Seed width to height ratio 1-g2.3	NA	Fang et al., 2017
	14	2259506	Reproductive stage length 1-g3.1	NA	Fang et al., 2017
		2259506	Reproductive stage length 1-g3.2	NA	Fang et al., 2017
	15	9020829	Seed coat luster 1-g1.1	NA	Fang et al., 2017
			WUE 3-g26	NA	Dhanapal et al., 2018
			WUE 3-g27.1	NA	Dhanapal et al., 2018
	17	31797213		NA	
	18	6686269	Seed protein 7-g27	NA	Zhang et al., 2017
		6728432	Seed protein 7-g27	NA	Zhang et al., 2017
SVR	1	47953926		1	
	4	19007585		NA	
		18643026		NA	

(Continued)

TABLE 4 | (Continued)

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment[a]	References
	6	41477920		NA	
	9	39659468	Seed Ser 2-g1	2	Li et al., 2018
			Seed Thr 2-g1	2	Li et al., 2018
		39664525	Seed Tyr 2-g2	NA	Li et al., 2018
			Seed Lys 2-g2	NA	Li et al., 2018
			Seed Leu 2-g2	NA	Li et al., 2018
			Seed Ala 2-g2	NA	Li et al., 2018
			Seed Gly 2-g2	NA	Li et al., 2018
		39366957	Pod number 1-g4.1	NA	Fang et al., 2017
			Pod number 1-g4.2	NA	Fang et al., 2017
		39372117	Pod number 1-g4.3	NA	Fang et al., 2017
			seed thickness 2-g4	2	Fang et al., 2017
		40355403		NA	
	10	3054709		NA	

[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.

appropriate approaches for detecting SNPs with large main effects on complex traits. However, they are underpowered to simultaneously consider a wide range of interconnected biological processes and mechanisms that shape the phenotype of complex traits (Lee et al., 2020). By using ML algorithms in GWAS, the interaction and joint effect of multiple SNPs can be estimated using variable importance methods, and the best set of SNPs will be selected to give the best performance (Pahikkala et al., 2012). Recent studies showed that the SNPs with high importance scores are not necessarily the SNPs with significant *p*-values resulted from single SNP analyses (Arshadi et al., 2009; Grömping, 2009; Szymczak et al., 2009; Ziliak, 2017; Di Leo and Sardanelli, 2020). Therefore, using variable importance values estimated by ML algorithms for identifying SNP-trait associations may improve the power of ML-mediated GWAS for discovering variant-trait associations with higher resolution (Szymczak et al., 2009). The variable importance methods based on linear and logistic regressions, Support vector machines, and random forest algorithms are well established in the literature (Grömping, 2009; Wu and Liu, 2009; Chun and Keleş, 2010; Williamson et al., 2020; Yoosefzadeh-Najafabadi et al., 2021b). In this study, we found that SVR-mediated GWAS had the same performance in detecting numbers of QTL when compared to conventional GWAS methods. However, the detected QTL by SVR-mediated GWAS was more related to the physiological background of each tested hyperspectral reflectance bands. For instance, in the 820 nm band, the SVR-mediated GWAS detected 5 QTL related to water use efficiency, which is clearly in agreement with the physiological background of this trait. The same scenario happened in the 660 nm band, where most of the detected QTL by SVR-mediated GWAS were related to flowing and soybean cyst resistance. In all the tested hyperspectral reflectance bands, several QTL related to the soybean seed protein, oil, pod number, seed yield, and seed thickness were detected in all GWAS methods. Meanwhile, seed protein, oil, pod number, seed yield, and seed thickness can be considered as the yield component traits, which directly and indirectly regulate the final soybean seed yield. Therefore, the detected QTL confirmed the efficiency of HypWAS and GWAS in indirect selection for complex traits such as yield.

Furthermore, several candidate genes were detected by SVR-mediated GWAS related to the oxidative and osmotic stresses, regulation of defense response, response to nematode, defense response to bacterium, and oxidoreductase activity. It is well documented that the violet spectrum and UV radiation are key factors in secondary metabolite production (e.g., terpenes, alkaloids, phenolic compounds, glucosinolates, and carotenoids) that can play a pivotal role in a plant's defense systems (Schreiner et al., 2012; Matsuura et al., 2013). It has also been shown that these spectra lead to the activation of several signaling pathways such as defense signaling, reactive oxygen species (ROS), and photomorphogenic signaling (Schreiner et al., 2012; Matsuura et al., 2013). These signaling can stimulate and induce the specific gene expression patterns involved in different secondary metabolism pathways, such as the isoflavonoid biosynthesis pathway (Kim et al., 2014). MYB family is one of the most important transcriptional factors that may interact with light-responsive elements and thereby activate selected genes involved in isoflavonoid biosynthesis (Du et al., 2010). Moreover, a positive correlation was reported between the expression profiles of the selected genes (*Glyma.02G008700* and *Glyma.09G168700*) and the patterns of isoflavonoid accumulation, which shows the biosynthesis of isoflavonoid might be activated by violet spectra through the up-regulation of these genes (Du et al., 2010). The spatiotemporal regulation of chlorophyll metabolism is necessary for various cellular processes such as chloroplast development, photosynthesis, plastid-derived retrograde signaling (Chan et al., 2016), RNA metabolism (Zhang et al., 2014), singlet oxygen-mediated signaling (Shen et al., 2006), abscisic acid signaling, and programmed cell death (Woodson et al., 2015; Dogra et al., 2019). Chlorophyll metabolism can be categorized into four functional classes including (i) Chlorophyll a synthesis through the branched tetrapyrrole biosynthesis pathway (Tanaka and Tanaka, 2007;

TABLE 5 | The list of detected QTLs for 820 nm using different GWAS methods in the tested soybean panel.

GWAS Method	Chromosome	Peak SNP position	Co-located QTL	Environment[a]	References
MLM	1	16343505		NA	
	5	1467115	seed palmitic 5-g2	3	Li et al., 2015
			seed palmitic 5-g1	3	Li et al., 2015
			seed palmitic 2-g1.3	3	Fang et al., 2017
			seed long-chain fatty acid 1-g19.3	3	Fang et al., 2017
			seed long-chain fatty acid 1-g1.3	3	Fang et al., 2017
			seed long-chain fatty acid 1-g21.3	3	Fang et al., 2017
			seed width to height ratio 1-g2.1	3	Fang et al., 2017
			seed width to height ratio 1-g2.2	3	Fang et al., 2017
			seed width to height ratio 1-g2.3	3	Fang et al., 2017
	6	50187091	Sclero 3-g32	3	Moellers et al., 2017
	16	36071599	Seed set 1-g21.2	NA	Fang et al., 2017
			Seed set 1-g21.1	NA	Fang et al., 2017
			First Flower 4-g66	NA	Mao et al., 2017
			SCN 5-g38	NA	Li et al., 2016
FarmCPU	1	16343505		NA	
	5	1467115	seed palmitic 5-g2	3	Li et al., 2015
			seed palmitic 5-g1	3	Li et al., 2015
			seed palmitic 2-g1.3	3	Fang et al., 2017
			seed long-chain fatty acid 1-g19.3	3	Fang et al., 2017
			seed long-chain fatty acid 1-g1.3	3	Fang et al., 2017
			seed long-chain fatty acid 1-g21.3	3	Fang et al., 2017
			seed width to height ratio 1-g2.1	3	Fang et al., 2017
			seed width to height ratio 1-g2.2	3	Fang et al., 2017
			seed width to height ratio 1-g2.3	3	Fang et al., 2017
	6	50187091	Sclero 3-g32	3	Moellers et al., 2017
	16	36071599	Seed set 1-g21.2	NA	Fang et al., 2017
			Seed set 1-g21.1	NA	Fang et al., 2017
		36236254	First Flower 4-g66	NA	Mao et al., 2017
			SCN 5-g38	NA	Li et al., 2016
SVR	1	39990647		NA	
		47953926		NA	
	2	42411031	WUE 3-g-2	4	Dhanapal et al., 2018
			WUE 3-g-3	4	Dhanapal et al., 2018
			WUE 2-g6	4	Kaler et al., 2017
			WUE 3-g4	4	Dhanapal et al., 2018
			WUE 3-g5	4	Dhanapal et al., 2018
	4	20000416		NA	
	6	11029779		NA	
	10	3054709		3,4	
	16	17376063		NA	

[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.

Mochizuki et al., 2010), (ii) the 'Chlorophyll cycle,' which catalyzes the interconversion of Chlorophyll a and Chlorophyll b (Tanaka and Tanaka, 2011), (iii) the degradation of Chlorophyll a to yield colorless through the pheophorbide a oxygenase (PAO)/phyllobilin pathway (Christ and Hörtensteiner, 2014), and (iv) Chlorophyll recycling pathway through dephytylase 1 (CLD1) (Lin et al., 2016). The combination of divinyl reductase (DVR) and light-dependent protochlorophyllide oxidoreductase (POR) produces chlorophyllide (Chlide) a. Subsequently, Chlorophyll synthase (CHLG) catalyzes

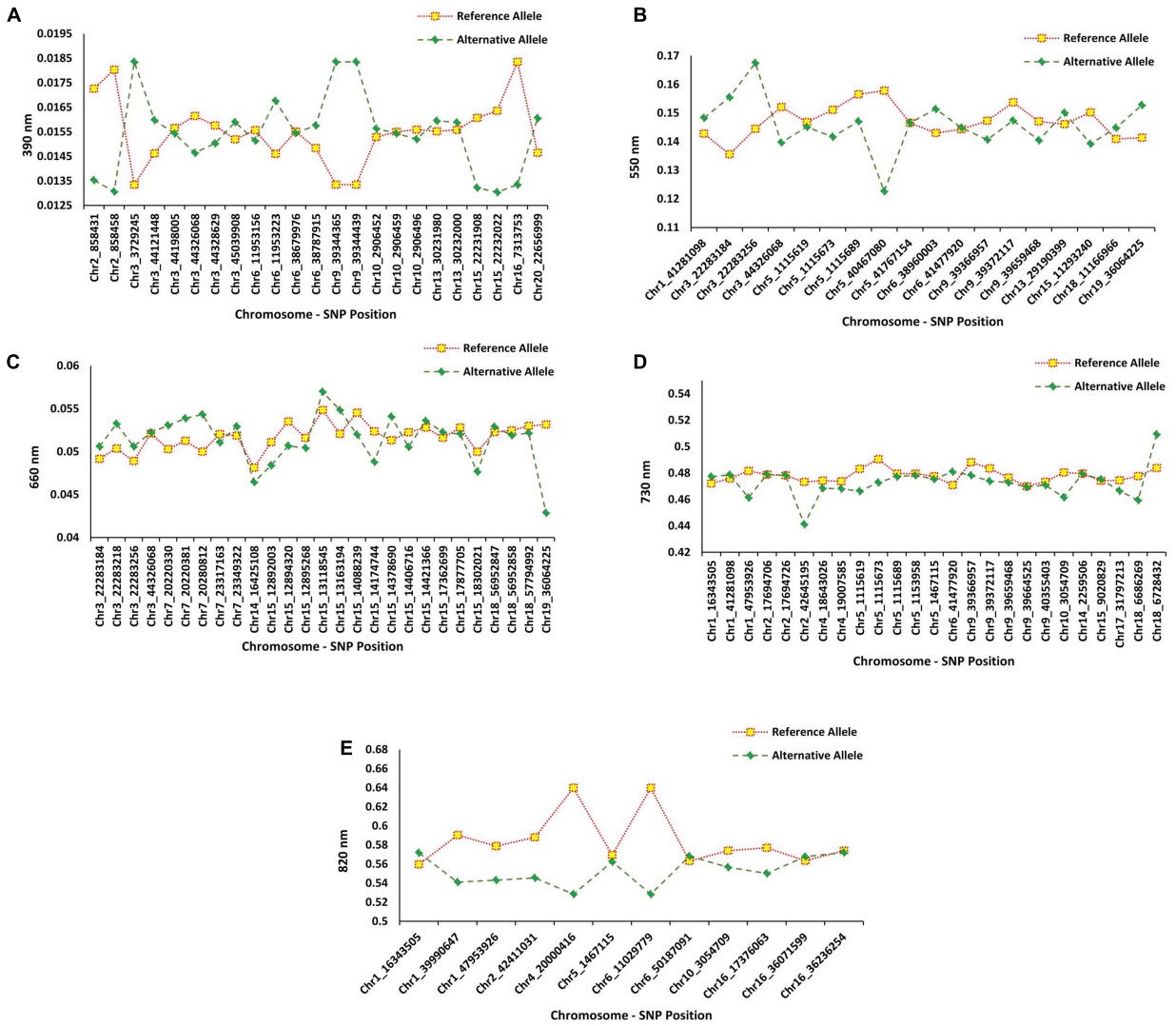

FIGURE 9 | The average effects of reference allele and alternative allele from the detected SNP's peak for **(A)** 390 nm, **(B)** 550 nm, **(C)** 660 nm, **(D)** 730 nm, and **(E)** 820 nm in 227 soybean genotypes across four environments.

Chlorophyll a biosynthesis through the combination of Chlide a and phytyl pyrophosphate (phytyl-PP) (Wang and Grimm, 2021). CLD1 can reversibly convert Chlorophyll a into Chlide a during Chlorophyll recycling. Also, it is well documented that high-light inducible proteins (Hlips) play an important role in binding Chlorophyll a and β-carotene and consequently the photosynthesis capacity (Chidgey et al., 2014; Staleva et al., 2015; Shukla et al., 2018).

It has been well documented that three main families of carbonic anhydrase (CA) genes, including α-CAs, β-CAs, and γ-CAs, play essential roles in the conductance of inorganic carbon through carbon fixation rates and the mesophyll (de Araujo et al., 2014; Cano et al., 2019). They may also be involved in sensing light, CO2, and water availability (Momayyezi et al., 2020). Therefore, CAs can affect photosynthetic efficiency through their impacts on stomatal response to light, CO2-facilitating components (aquaporins), ABA signaling, and other signaling pathways (de Araujo et al., 2014; Cano et al., 2019; Momayyezi et al., 2020). ABA is a key phytohormone associated with stomatal closure. ABA receptors (e.g., PYL, PYR, RCAR proteins) play an important role in executing ABA's function in water relations (Cutler et al., 2010; Kim et al., 2010). ABA regulates the stress-activated kinase signaling network that controls stomatal closure (Mega et al., 2019). In reacting to water deficit, the level of ABA increases, which regulates the ligand-receptor complex formation that represses the clade A protein phosphatase 2Cs (PP2Cs) activity, which is considered negative regulators for ABA signaling (Fujii et al., 2009; Ma et al., 2009; Park et al., 2009). Because of the central role of ABA receptors in transpiration regulation, they can be considered as promising targets for breeding programs in order to manipulate ABA sensitivity and water productivity (Mega et al., 2019).

CONCLUSION

Indirect selection of complex traits would be of paramount importance in analytical breeding strategies. Nowadays, the use of advanced high throughput phenotyping and genotyping combined with big data analysis methods can ease the assessment of large plant breeding populations in a very effective short time. For the first time in this study, we are proposing the HypWAS method for identifying hyperspectral reflectance bands associated with complex traits such as yield. Based on this method, we were able to discover, five hyperspectral reflectance bands significantly associated with the soybean seed yield. The visible region of the spectra was found to be the most informative region related to the seed yield. The GWAS analyses of the selected hyperspectral reflectance bands using MLM, FarmCPU, and a newly developed SVR-mediated GWAS method revealed several QTL revealing the bands that seem to be related to the soybean seed yield, water use efficiency, and soybean cyst nematodes resistance based on previous studies. In general, all of the tested GWAS methods had acceptable performance. However, we were able to detect more relevant QTL using the SVR-mediated GWAS. Regarding the Gene Ontology of the selected traits, most of the detected genes were reported to be related to the water status, photosynthesis, and light intensity. The obtained genetic results confirmed the physiological background of the selected hyperspectral reflectance bands. The result of this study can be used to accelerate the indirect breeding selection strategy for selecting high-yielding genotypes based on specific hyperspectral reflectance bands at early plant growth stages. In addition, the genetic results can be employed to use the detected QTL in each hyperspectral reflectance band for MAS selection in large breeding populations.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/Mohsen1080/Available-Datasets/blob/92f27c80fa3e460900589b42188a943570dee86d/FastGBS.SNPs.232.imputed.Het50.maf0.05.hmp.txt, GitHub.

AUTHOR CONTRIBUTIONS

ME conceptualized, designed and directed the experiments. MY-N conducted the experiments, modeled, summarized the results, and writing the manuscript. ST participated in candidate gene analyses. ST, DT, 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 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.

ACKNOWLEDGMENTS

We 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 Mohsen Hesami for his assistance with reviewing the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.777028/full#supplementary-material

REFERENCES

Albashish, D., Hammouri, A. I., Braik, M., Atwan, J., and Sahran, S. (2021). Binary biogeography-based optimization based SVM-RFE for feature selection. *Appl. Soft Comput.* 101:107026. doi: 10.1016/j.asoc.2020.107026

Ali, A., and Imran, M. (2021). Remotely sensed real-time quantification of biophysical and biochemical traits of *Citrus* (*Citrus sinensis* L.) fruit orchards–A review. *Sci. Hortic.* 282:110024. doi: 10.1016/j.scienta.2021.110024

Alonzo, M., Bookhagen, B., and Roberts, D. A. (2014). Urban tree species mapping using hyperspectral and lidar data fusion. *Rem. Sens. Environ.* 148, 70–83. doi: 10.1016/j.rse.2014.03.018

Alqudah, A. M., Sallam, A., Baenziger, P. S., and Börner, A. (2020). GWAS: fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley–A review. *J. Adv. Res.* 22, 119–135. doi: 10.1016/j.jare.2019.10.013

Anuarbek, S., Abugalieva, S., Pecchioni, N., Laidò, G., Maccaferri, M., Tuberosa, R., et al. (2020). Quantitative trait loci for agronomic traits in tetraploid wheat for enhancing grain yield in Kazakhstan environments. *PLoS One* 15:e0234863. doi: 10.1371/journal.pone.0234863

Arshadi, N., Chang, B., and Kustra, R. (2009). Predictive modeling in case-control single-nucleotide polymorphism studies in the presence of population stratification: a case study using Genetic Analysis Workshop 16 Problem 1 dataset. *BMC Proc.* 3(Suppl. 7):S60. doi: 10.1186/1753-6561-3-s7-s60

Asif, H., Alliey-Rodriguez, N., Keedy, S., Tamminga, C. A., Sweeney, J. A., Pearlson, G., et al. (2020). GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. *Mol. Psychiatry* 26, 2048–2055. doi: 10.1038/s41380-020-0670-3

Awad, M., and Khanna, R. (2015). "Support vector regression," in *Efficient Learning Machines*, eds M. Awad and R. Khanna (Berkeley, CA: Apress), 67–80. doi: 10.1007/978-1-4302-5990-9_4

Awika, H. O., Solorzano, J., Cholula, U., Shi, A., Enciso, J., and Avila, C. A. (2021). Prediction modeling for yield and water-use efficiency in spinach using remote sensing via an unmanned aerial system. *Smart Agric. Technol.* 1:100006. doi: 10.1016/j.atech.2021.100006

Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., et al. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome* 8, 1–13. doi: 10.3835/plantgenome2015.04.0024

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bowley, S. (1999). *A Hitchhiker's Guide to Statistics in Plant Biology*. Guelph, ON: Any Old Subject Books.

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* 12:232. doi: 10.1186/gb-2011-12-10-232

Brown, A. V., Conners, S. I., Huang, W., Wilkey, A. P., Grant, D., Weeks, N. T., et al. (2021). A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 49, D1496–D1501. doi: 10.1093/nar/gkaa1107

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295. doi: 10.1038/ng.3211

Cano, F. J., Sharwood, R. E., Cousins, A. B., and Ghannoum, O. (2019). The role of leaf width and conductances to CO 2 in determining water use efficiency in C4 grasses. *N. Phytol.* 223, 1280–1295. doi: 10.1111/nph.15920

Cao, Y., Li, S., Wang, Z., Chang, F., Kong, J., Gai, J., et al. (2017). Identification of major quantitative trait loci for seed oil content in soybeans by combining linkage and genome-wide association mapping. *Front. Plant Science* 8:1222. doi: 10.3389/fpls.2017.01222

Castro-Esau, K. L., Sánchez-Azofeifa, G. A., Rivard, B., Wright, S. J., and Quesada, M. (2006). Variability in leaf optical properties of Mesoamerican trees and the potential for species classification. *Am. J. Bot.* 93, 517–530. doi: 10.3732/ajb.93.4.517

Chan, K. X., Phua, S. Y., Crisp, P., Mcquinn, R., and Pogson, B. J. (2016). Learning the languages of the chloroplast: retrograde signaling and beyond. *Annu. Rev. Plant Biol.* 67, 25–53. doi: 10.1146/annurev-arplant-043015-111854

Chang, H.-X., Lipka, A. E., Domier, L. L., and Hartman, G. L. (2016). Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Phytopathology* 106, 1139–1151. doi: 10.1094/PHYTO-01-16-0042-FI

Chen, Z., Jia, K., Xiao, C., Wei, D., Zhao, X., Lan, J., et al. (2020). Leaf area index estimation algorithm for GF-5 hyperspectral data based on different feature selection and machine learning methods. *Rem. Sens.* 12:2110. doi: 10.3390/rs12132110

Chidgey, J. W., Linhartová, M., Komenda, J., Jackson, P. J., Dickman, M. J., Canniffe, D. P., et al. (2014). A cyanobacterial chlorophyll synthase-HliD complex associates with the Ycf39 protein and the YidC/Alb3 insertase. *Plant Cell* 26, 1267–1279. doi: 10.1105/tpc.114.124495

Chowdhury, M. Z. I., and Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Fam. Med. Community Health* 8:e000262. doi: 10.1136/fmch-2019-000262

Christ, B., and Hörtensteiner, S. (2014). Mechanism and significance of chlorophyll breakdown. *J. Plant Growth Regul.* 33, 4–20. doi: 10.1007/s00344-013-9392-y

Chun, H., and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 72, 3–25. doi: 10.1111/j.1467-9868.2009.00723.x

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963–971. doi: 10.1093/genetics/138.3.963

Clark, M. L., Roberts, D. A., and Clark, D. B. (2005). Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Rem. Sens. Environ.* 96, 375–398. doi: 10.1016/j.rse.2005.03.009

Clevers, J., De Jong, S., Epema, G., Van Der Meer, F., Bakker, W., Skidmore, A., et al. (2002). Derivation of the red edge index using the MERIS standard band setting. *Int. J. Rem. Sens.* 23, 3169–3184. doi: 10.1080/01431160110104647

Contreras-Soto, R. I, Mora, F., De Oliveira, M. A. R., Higashi, W., Scapim, C. A., and 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. doi: 10.1371/journal.pone.0171105

Cutler, S. R., Rodriguez, P. L., Finkelstein, R. R., and Abrams, S. R. (2010). Abscisic acid: emergence of a core signaling network. *Annu. Rev. Plant Biol.* 61, 651–679. doi: 10.1146/annurev-arplant-042809-112122

Dababat, A., Arif, M. A. R., Toktay, H., Atiya, O., Shokat, S., Gul, E., et al. (2021). A GWAS to identify the cereal cyst nematode (*Heterodera filipjevi*) resistance loci in diverse wheat prebreeding lines. *J. Appl. Genet.* 62, 93–98. doi: 10.1007/s13353-020-00607-y

de Araujo, C., Arefeen, D., Tadesse, Y., Long, B. M., Price, G. D., Rowlett, R. S., et al. (2014). Identification and characterization of a carboxysomal γ-carbonic anhydrase from the cyanobacterium *Nostoc* sp. PCC 7120. *Photosynth. Res.* 121, 135–150. doi: 10.1007/s11120-014-0018-4

Dhanapal, A. P., Ray, J. D., Smith, J. R., Purcell, L. C., and Fritschi, F. B. (2018). Identification of novel genomic loci associated with soybean shoot tissue macro and micronutrient concentrations. *Plant Genome* 11:170066. doi: 10.3835/plantgenome2017.07.0066

Di Leo, G., and Sardanelli, F. (2020). Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *Eur. Radiol. Exp.* 4, 1–8. doi: 10.1186/s41747-020-0145-y

Diers, B. W., Specht, J., Rainey, K. M., Cregan, P., Song, Q., Ramasubramanian, V., et al. (2018). Genetic architecture of soybean yield and agronomic traits. *G3 Genes Genomes Genetics* 8, 3367–3375. doi: 10.1534/g3.118.200332

Doerge, R. W., and Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* 142, 285–294. doi: 10.1093/genetics/142.1.285

Dogra, V., Li, M., Singh, S., Li, M., and Kim, C. (2019). Oxidative post-translational modification of EXECUTER1 is required for singlet oxygen sensing in plastids. *Nat. Commun.* 10:2834. doi: 10.1038/s41467-019-10760-6

Du, H., Huang, Y., and Tang, Y. (2010). Genetic and metabolic engineering of isoflavonoid biosynthesis. *Appl. Microbiol. Biotechnol.* 86, 1293–1312. doi: 10.1007/s00253-010-2512-8

Eltaher, S., Baenziger, P. S., Belamkar, V., Emara, H. A., Nower, A. A., Salem, K. F., et al. (2021). GWAS revealed effect of genotype× environment interactions for grain yield of Nebraska winter wheat. *BMC Genomics* 22:2. doi: 10.1186/s12864-020-07308-0

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 18, 1–14. doi: 10.1186/s13059-017-1289-9

Fei, S., Hassan, M. A., He, Z., Chen, Z., Shu, M., Wang, J., et al. (2021). Assessment of ensemble learning to predict wheat grain yield based on UAV-multispectral reflectance. *Rem. Sens.* 13:2338. doi: 10.3390/rs13122338

Feng, X., Zhan, Y., Wang, Q., Yang, X., Yu, C., Wang, H., et al. (2020). Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping. *Plant J.* 101, 1448–1461. doi: 10.1111/tpj.14597

Fernandes, M. R., Aguiar, F. C., Silva, J. M., Ferreira, M. T., and Pereira, J. M. (2013). Spectral discrimination of giant reed (*Arundo donax* L.): a seasonal study in riparian areas. *ISPRS J. Photogramm. Rem. Sens.* 80, 80–90. doi: 10.1016/j.isprsjprs.2013.03.007

Fujii, H., Chinnusamy, V., Rodrigues, A., Rubio, S., Antoni, R., Park, S.-Y., et al. (2009). In vitro reconstitution of an abscisic acid signalling pathway. *Nature* 462, 660–664. doi: 10.1038/nature08599

Galán, R. J., Bernal-Vasquez, A.-M., Jebsen, C., Piepho, H.-P., Thorwarth, P., Steffan, P., et al. (2020). Hyperspectral reflectance data and agronomic traits can predict biomass yield in winter rye hybrids. *BioEnergy Res.* 13, 168–182. doi: 10.1007/s12155-019-10080-z

Gao, J., Li, P., Ma, F., and Goltsev, V. (2014). Photosynthetic performance during leaf expansion in *Malus micromalus* probed by chlorophyll a fluorescence and modulated 820 nm reflection. *J. Photochem. Photobiol. B Biol.* 137, 144–150. doi: 10.1016/j.jphotobiol.2013.12.005

George, E. I. (2000). The variable selection problem. *J. Am. Stat. Assoc.* 95, 1304–1308. doi: 10.1080/01621459.2000.10474336

Gitelson, A. A., Merzlyak, M. N., and Chivkunova, O. B. (2001). Optical properties and nondestructive estimation of anthocyanin content in plant leaves. *Photochem. Photobiol.* 74, 38–45. doi: 10.1562/0031-8655(2001)074<0038:OPANEO>2.0.CO;2

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *J. Am. Stat. Assoc.* 57, 369–375. doi: 10.1080/01621459.1962.10480665

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *Am. Stat.* 63, 308–319. doi: 10.1198/tast.2009.08199

Gupta, M., and Gupta, B. (2020). A novel gene expression test method of minimizing breast cancer risk in reduced cost and time by improving SVM-RFE

gene selection method combined with LASSO. *J. Integr. Bioinform.* 18, 139–153. doi: 10.1515/jib-2019-0110

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422. doi: 10.1023/A:1012487302797

Hao, D., Cheng, H., Yin, Z., Cui, S., Zhang, D., Wang, H., et al. (2012). Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. *Theor. Appl. Genet.* 124, 447–458. doi: 10.1007/s00122-011-1719-0

Heinze, G., and Dunkler, D. (2017). Five myths about variable selection. *Transpl. Int.* 30, 6–10. doi: 10.1111/tri.12895

Hennessy, A., Clarke, K., and Lewis, M. (2020). Hyperspectral classification of plants: a review of waveband selection generalisability. *Rem. Sens.* 12:113. doi: 10.3390/rs12010113

Hesami, M., Alizadeh, M., Naderi, R., and Tohidfar, M. (2020). Forecasting and optimizing Agrobacterium-mediated genetic transformation via ensemble model-fruit fly optimization algorithm: a data mining approach using chrysanthemum databases. *PLoS One* 15:e0239901. doi: 10.1371/journal.pone.0239901

Hesami, M., Yoosefzadeh Najafabadi, M., Adamek, K., Torkamaneh, D., and Jones, A. M. P. (2021). Synergizing off-target predictions for in silico insights of CENH3 knockout in cannabis through CRISPR/CAS. *Molecules* 26:2053. doi: 10.3390/molecules26072053

Hoa, P., Giang, N., Binh, N., Hieu, N., Trang, N., Toan, L., et al. (2017). Mangrove species discrimination in Southern Vietnam based on in-situ measured hyperspectral reflectance. *Int. J. Geoinform.* 13, 25–35.

Holmes, W. S., Ooi, M. P.-L., Kuang, Y. C., Simpkin, R., Lopez-Ubiria, I., Vidiella, A., et al. (2020). "Classifying *Cannabis sativa* flowers, stems and leaves using statistical machine learning with near-infrared hyperspectral reflectance imaging," in *Proceedings of the 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, (Piscataway, NJ: IEEE), 1–6. doi: 10.1109/I2MTC43012.2020.9129531

Horler, D., Dockray, M., and Barber, J. (1983). The red edge of plant leaf reflectance. *Int. J. Rem. Sens.* 4, 273–288. doi: 10.1080/01431168308948546

Jafari, M., and Shahsavar, A. (2020). The application of artificial neural networks in modeling and predicting the effects of melatonin on morphological responses of citrus to drought stress. *PLoS One* 15:e0240427. doi: 10.1371/journal.pone.0240427

Jamil, I. N., Remali, J., Azizan, K. A., Nor Muhammad, N. A., Arita, M., Goh, H.-H., et al. (2020). Systematic multi-omics integration (MOI) approach in plant systems biology. *Front. Plant Sci.* 11:944. doi: 10.3389/fpls.2020.00944

Kaler, A. S., and Purcell, L. C. (2019). Estimation of a significance threshold for genome-wide association studies. *BMC Genomics* 20:618. doi: 10.1186/s12864-019-5992-7

Kaler, A. S., Dhanapal, A. P., Ray, J. D., King, C. A., Fritschi, F. B., and Purcell, L. C. (2017). Genome-wide association mapping of carbon isotope and oxygen isotope ratios in diverse soybean genotypes. *Crop Sci.* 57, 3085–3100. doi: 10.2135/cropsci2017.03.0160

Kaler, A. S., Gillman, J. D., Beissinger, T., and Purcell, L. C. (2020). Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. *Front. Plant Sci.* 10:1794. doi: 10.3389/fpls.2019.01794

Katsileros, A., Drosou, K., and Koukouvinos, C. (2015). Evaluation of nearest neighbor methods in wheat genotype experiments. *Commun. Biometry Crop Sci.* 10, 115–123.

Ke, B. (2001). *Photosynthesis Photobiochemistry and Photobiophysics*. Dordrecht: Springer Science & Business Media.

Khanzadeh, H., Ghavi Hossein-Zadeh, N., and Ghovvati, S. (2020). Genome wide association studies, next generation sequencing and their application in animal breeding and genetics: a review. *Iran. J. Appl. Anim. Sci.* 10, 395–404.

Kim, T.-H., Böhmer, M., Hu, H., Nishimura, N., and Schroeder, J. I. (2010). Guard cell signal transduction network: advances in understanding abscisic acid, CO2, and Ca2+ signaling. *Annu/ Rev. Plant Biol.* 61, 561–591. doi: 10.1146/annurev-arplant-042809-112226

Kim, Y. B., Thwe, A. A., Li, X., Tuan, P. A., Zhao, S., Park, C. G., et al. (2014). Accumulation of flavonoids and related gene expressions in different organs of *Astragalus membranaceus* Bge. *Appl. Biochem. Biotechnol.* 173, 2076–2085. doi: 10.1007/s12010-014-1004-1

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26. doi: 10.18637/jss.v028.i05

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., et al. (2020). Package 'caret'. *R J.*

Lee, S., Liang, X., Woods, M., Reiner, A. S., Concannon, P., Bernstein, L., et al. (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. doi: 10.1371/journal.pone.0226157

Li, X., Tian, R., Kamala, S., Du, H., Li, W., Kong, Y., et al. (2018). Identification and verification of pleiotropic QTL controlling multiple amino acid contents in soybean seed. *Euphytica* 214, 1–14. doi: 10.1007/s10681-018-2170-y

Li, Y. H., Shi, X. H., Li, H. H., Reif, J. C., Wang, J. J., Liu, Z. X., et al. (2016). Dissecting the genetic basis of resistance to soybean cyst nematode combining linkage and association mapping. *Plant Genome* 9:150020. doi: 10.3835/plantgenome2015.04.0020

Li, Y.-H., Reif, J. C., Ma, Y.-S., Hong, H.-L., Liu, Z.-X., Chang, R.-Z., et al. (2015). Targeted association mapping demonstrating the complex molecular genetics of fatty acid formation in soybean. *BMC Genomics* 16:841. doi: 10.1186/s12864-015-2049-4

Li, Z., Votava, J. A., Zajac, G. J., Nguyen, J. N., Jaimes, F. B. L., Ly, S. M., et al. (2020). Integrating mouse and human genetic data to move beyond GWAS and identify causal genes in cholesterol metabolism. *Cell Metab.* 31, 741–754.e745. doi: 10.1016/j.cmet.2020.02.015

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Machine learning in agriculture: a review. *Sensors* 18:2674. doi: 10.3390/s18082674

Lin, Y.-P., Wu, M.-C., and Charng, Y.-Y. (2016). Identification of a chlorophyll dephytylase involved in chlorophyll turnover in *Arabidopsis*. *Plant Cell* 28, 2974–2990. doi: 10.1105/tpc.16.00478

Liu, C., Hu, Z., Islam, A. T., Kong, R., Yu, L., Wang, Y., et al. (2021). Hyperspectral characteristics and inversion model estimation of winter wheat under different elevated CO2 concentrations. *Int. J. Rem. Sens.* 42, 1035–1053. doi: 10.1080/01431161.2020.1823038

Liu, M., Liu, X., Du, X., Korpelainen, H., Niinemets, Ü, and Li, C. (2021). Anatomical variation of mesophyll conductance due to salt stress in *Populus cathayana* females and males growing under different inorganic nitrogen sources. *Tree Physiol.* 41, 1462–1478. doi: 10.1093/treephys/tpab017

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12:e1005767. doi: 10.1371/journal.pgen.1005767

Liu, Y., Tikunov, Y., Schouten, R. E., Marcelis, L. F., Visser, R. G., and Bovy, A. (2018). Anthocyanin biosynthesis and degradation mechanisms in *Solanaceous vegetables*: a review. *Front. Chem.* 6:52. doi: 10.3389/fchem.2018.00052

Ma, Y., Szostkiewicz, I., Korte, A., Moes, D., Yang, Y., Christmann, A., et al. (2009). Regulators of PP2C phosphatase activity function as abscisic acid sensors. *Science* 324, 1064–1068. doi: 10.1126/science.1172408

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., and Fritschi, F. B. (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Rem. Sens. Environ.* 237:111599. doi: 10.1016/j.rse.2019.111599

Mao, T., Li, J., Wen, Z., Wu, T., Wu, C., Sun, S., et al. (2017). Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. *BMC Genomics* 18:415. doi: 10.1186/s12864-017-3778-3

Matsuura, H. N., De Costa, F., Yendo, A. C. A., and Fett-Neto, A. G. (2013). "Photoelicitation of bioactive secondary metabolites by ultraviolet radiation: mechanisms, strategies, and applications," in *Biotechnology for Medicinal Plants*, eds S. Chandra, H. Lata, and A. Varma (Berlin: Springer), 171–190. doi: 10.1007/978-3-642-29974-2_7

Mega, R., Abe, F., Kim, J.-S., Tsuboi, Y., Tanaka, K., Kobayashi, H., et al. (2019). Tuning water-use efficiency and drought tolerance in wheat using abscisic acid receptors. *Nat. Plants* 5, 153–159. doi: 10.1038/s41477-019-0361-8

Mikel, M. A., Diers, B. W., Nelson, R. L., and Smith, H. H. (2010). Genetic diversity and agronomic improvement of North American soybean germplasm. *Crop Sci.* 50, 1219–1229. doi: 10.2135/cropsci2009.08.0456

Mochizuki, N., Tanaka, R., Grimm, B., Masuda, T., Moulin, M., Smith, A. G., et al. (2010). The cell biology of tetrapyrroles: a life and death struggle. *Trends Plant Sci.* 15, 488–498. doi: 10.1016/j.tplants.2010.05.012

Moellers, T. C., Singh, A., Zhang, J., Brungardt, J., Kabbage, M., Mueller, D. S., et al. (2017). Main and epistatic loci studies in soybean for *Sclerotinia sclerotiorum* resistance reveal multiple modes of resistance in multi-environments. *Sci. Rep.* 7:3554. doi: 10.1038/s41598-017-03695-9

Mohammadi, M., Xavier, A., Beckett, T., Beyer, S., Chen, L., Chikssa, H., et al. (2020). Identification, deployment, and transferability of quantitative trait loci from genome-wide association studies in plants. *Curr. Plant Biol.* 24:100145. doi: 10.1016/j.cpb.2020.100145

Momayyezi, M., Mckown, A. D., Bell, S. C., and Guy, R. D. (2020). Emerging roles for carbonic anhydrase in mesophyll conductance and photosynthesis. *Plant J.* 101, 831–844. doi: 10.1111/tpj.14638

Najafabadi, M. Y., Torabi, S., Torkamaneh, D., Tulpan, D., Rajcan, I., and Eskandari, M. (2021). Machine learning based genome-wide association studies for uncovering QTL underlying soybean yield and its components. *bioRxiv* [preprint] doi: 10.1101/2021.06.24.449776

Nicholls, H. L., John, C. R., Watson, D. S., Munroe, P. B., Barnes, M. R., and Cabrera, C. P. (2020). Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. *Front. Genet.* 11:350. doi: 10.3389/fgene.2020.00350

Okubo, K. (2021). "NIR hyperspectral imaging," in *Transparency in Biology*, eds K. Soga, M. Umezawa, and K. Okubo (Singapore: Springer), 203–222. doi: 10.1007/978-981-15-9627-8_10

Omidi, R., Moghimi, A., Pourreza, A., El-Hadedy, M., and Eddin, A. S. (2020). Ensemble hyperspectral band selection for detecting nitrogen status in grape leaves. *arXiv* [preprint] arXiv:2010.04225 doi: 10.1109/ICMLA51294.2020.00054

Pahikkala, T., Okser, S., Airola, A., Salakoski, T., and Aittokallio, T. (2012). Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms Mol. Biol.* 7, 1–15. doi: 10.1186/1748-7188-7-11

Palanivel, K., and Surianarayanan, C. (2019). An approach for prediction of crop yield using machine learning and big data techniques. *Int. J. Comput. Eng. Technol.* 10, 110–118. doi: 10.34218/IJCET.10.3.2019.013

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., and Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Comput. Electron. Agric.* 121, 57–65. doi: 10.1016/j.compag.2015.11.018

Park, S.-Y., Fung, P., Nishimura, N., Jensen, D. R., Fujii, H., Zhao, Y., et al. (2009). Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. *Science* 324, 1068–1071. doi: 10.1126/science.1173041

Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S., and Singh, A. K. (2019). Machine learning approach for prescriptive plant breeding. *Sci. Rep.* 9:17132. doi: 10.1038/s41598-019-53451-4

Paulus, S., and Mahlein, A.-K. (2020). Technical workflows for hyperspectral plant image assessment and processing on the greenhouse and laboratory scale. *GigaScience* 9:giaa090. doi: 10.1093/gigascience/giaa090

Peerbhay, K. Y., Mutanga, O., and Ismail, R. (2013). Commercial tree species discrimination using airborne AISA Eagle hyperspectral imagery and partial least squares discriminant analysis (PLS-DA) in KwaZulu–Natal, South Africa. *ISPRS J. Photogramm. Rem. Sens.* 79, 19–28. doi: 10.1016/j.isprsjprs.2013.01.013

Pettai, H., Oja, V., Freiberg, A., and Laisk, A. (2005). The long-wavelength limit of plant photosynthesis. *FEBS Lett.* 579, 4017–4019. doi: 10.1016/j.febslet.2005.04.088

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* 19, 3735–3746. doi: 10.1016/j.csbj.2021.06.030

Platt, A., Vilhjálmsson, B. J., and Nordborg, M. (2010). Conditions under which genome-wide association studies will be positively misleading. *Genetics* 186, 1045–1052. doi: 10.1534/genetics.110.121665

Pu, R. (2009). Broadleaf species recognition with in situ hyperspectral data. *Int. J. Rem. Sens.* 30, 2759–2779. doi: 10.1080/01431160802555820

Qiao, M., He, X., Cheng, X., Li, P., Luo, H., Zhang, L., et al. (2021). Crop yield prediction from multi-spectral, multi-temporal remotely sensed imagery using recurrent 3D convolutional neural networks. *Int. J. Appl. Earth Observ. Geoinf.* 102:102436. doi: 10.1016/j.jag.2021.102436

Qin, J., Song, Q., Shi, A., Li, S., Zhang, M., and 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. doi: 10.1371/journal.pone.0184613

Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197, 573–589. doi: 10.1534/genetics.114.164350

Ray, J. D., Dhanapal, A. P., Singh, S. K., Hoyos-Villegas, V., Smith, J. R., Purcell, L. C., et al. (2015). Genome-wide association study of ureide concentration in diverse maturity group IV soybean [*Glycine max* (L.) Merr.] accessions. *G3 Genes Genomes Genet.* 5, 2391–2403. doi: 10.1534/g3.115.021774

Richter, R., Reu, B., Wirth, C., Doktor, D., and Vohland, M. (2016). The use of airborne hyperspectral data for tree species classification in a species-rich Central European forest area. *Int. J. Appl. Earth Observ. Geoinf.* 52, 464–474. doi: 10.1016/j.jag.2016.07.018

Rivard, B., Sanchez-Azofeifa, G. A., Foley, S., and Calvo-Alvarado, J. C. (2008). "Species classification of tropical tree leaf reflectance and dependence on selection of spectral bands," in *Hyperspectral Remote Sensing of Tropical and Sub-Tropical Forests*, eds M. Kalascka and A. Sanchez-Azofeifa (Boca Raton, FL: CRC Press), 141–159. doi: 10.1201/9781420053432.ch6

Salvatori, E., Fusaro, L., Strasser, R. J., Bussotti, F., and Manes, F. (2015). Effects of acute O3 stress on PSII and PSI photochemistry of sensitive and resistant snap bean genotypes (*Phaseolus vulgaris* L.), probed by prompt chlorophyll "a" fluorescence and 820 nm modulated reflectance. *Plant Physiol. Biochem.* 97, 368–377. doi: 10.1016/j.plaphy.2015.10.027

Sanz, H., Valim, C., Vegas, E., Oller, J. M., and Reverter, F. (2018). SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics* 19:432. doi: 10.1186/s12859-018-2451-4

Schreiner, M., Mewis, I., Huyskens-Keil, S., Jansen, M., Zrenner, R., Winkler, J., et al. (2012). UV-B-induced secondary plant metabolites-potential benefits for plant and human health. *Crit. Rev. Plant Sci.* 31, 229–240. doi: 10.1080/07352689.2012.664979

Seck, W., Torkamaneh, D., and Belzile, F. (2020). Comprehensive genome-wide association analysis reveals the genetic basis of root system architecture in soybean. *Front. Plant Sci.* 11:590740. doi: 10.3389/fpls.2020.590740

Severin, A. J., Woody, J. L., Bolon, Y.-T., Joseph, B., Diers, B. W., Farmer, A. D., et al. (2010). RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biol.* 10:160. doi: 10.1186/1471-2229-10-160

Sharifi, A. (2021). Yield prediction with machine learning algorithms and satellite images. *J. Sci. Food Agric.* 101, 891–896. doi: 10.1002/jsfa.10696

Shen, Y.-Y., Wang, X.-F., Wu, F.-Q., Du, S.-Y., Cao, Z., Shang, Y., et al. (2006). The Mg-chelatase H subunit is an abscisic acid receptor. *Nature* 443, 823–826. doi: 10.1038/nature05176

Shukla, M. K., Llansola-Portoles, M. J., Tichý, M., Pascal, A. A., Robert, B., and Sobotka, R. (2018). Binding of pigments to the cyanobacterial high-light-inducible protein HliC. *Photosynth. Res.* 137, 29–39. doi: 10.1007/s11120-017-0475-7

Siegmann, B., and Jarmer, T. (2015). Comparison of different regression models and validation techniques for the assessment of wheat leaf area index from hyperspectral data. *Int. J. Rem. Sens.* 36, 4519–4534. doi: 10.1080/01431161.2015.1084438

Singh, S., Sehgal, D., Kumar, S., Arif, M., Vikram, P., Sansaloni, C., et al. (2020). GWAS revealed a novel resistance locus on chromosome 4D for the quarantine disease Karnal bunt in diverse wheat pre-breeding germplasm. *Sci. Rep.* 10:5999. doi: 10.1038/s41598-020-62711-7

Smola, A. J., and Schölkopf, B. (2004). A tutorial on support vector regression. *Stat. Comput.* 14, 199–222. doi: 10.1023/B:STCO.0000035301.49549.88

Somegowda, V. K., Rayaprolu, L., Rathore, A., Deshpande, S. P., and Gupta, R. (2021). Genome-Wide association studies (GWAS) for traits related to fodder quality and biofuel in sorghum: progress and prospects. *Protein Peptide Lett.* 28, 843–854. doi: 10.2174/0929866528666210127153103

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., et al. (2013). An Improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* 8:e54603. doi: 10.1371/journal.pone.0054603

Sonah, H., O'donoughue, L., Cober, E., Rajcan, I., and Belzile, F. (2015). Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol. J.* 13, 211–221. doi: 10.1111/pbi.12249

Staleva, H., Komenda, J., Shukla, M. K., Šlouf, V., Kaňa, R., Polívka, T., et al. (2015). Mechanism of photoprotection in the cyanobacterial ancestor of plant antenna proteins. *Nat. Chem. Biol.* 11, 287–291. doi: 10.1038/nchembio.1755

Stommel, J. R., Lightbourn, G. J., Winkel, B. S., and Griesbach, R. J. (2009). Transcription factor families regulate the anthocyanin biosynthetic pathway in *Capsicum annuum*. *J. Am. Soc. Hortic. Sci.* 134, 244–251. doi: 10.21273/JASHS.134.2.244

Stroup, W., and Mulitze, D. (1991). Nearest neighbor adjusted best linear unbiased prediction. *Am. Stat.* 45, 194–200. doi: 10.1080/00031305.1991.10475801

Sun, S., Dong, B., and Zou, Q. (2021). Revisiting genome-wide association studies from statistical modelling to machine learning. *Brief. Bioinformatics* 22:bbaa263. doi: 10.1093/bib/bbaa263

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H., et al. (2009). Machine learning in genome-wide association studies. *Genet. Epidemiol.* 33, S51–S57. doi: 10.1002/gepi.20473

Tanaka, R., and Tanaka, A. (2007). Tetrapyrrole biosynthesis in higher plants. *Annu. Rev. Plant Biol.* 58, 321–346. doi: 10.1146/annurev.arplant.57.032905.105448

Tanaka, R., and Tanaka, A. (2011). Chlorophyll cycle regulates the construction and destruction of the light-harvesting complexes. *Biochim. Biophys. Acta (BBA) Bioenerget.* 1807, 968–976. doi: 10.1016/j.bbabio.2011.01.002

Tarazona, S., Arzalluz-Luque, A., and Conesa, A. (2021). Undisclosed, unmet and neglected challenges in multi-omics studies. *Nat. Comput. Sci.* 1, 395–402. doi: 10.1038/s43588-021-00086-z

Tibbs Cortes, L., Zhang, Z., and Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *Plant Genome* 14:e20077. doi: 10.1002/tpg2.20077

Tong, H., and Nikoloski, Z. (2021). Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data. *J. Plant Physiol.* 257:153354. doi: 10.1016/j.jplph.2020.153354

Torkamaneh, D., Laroche, J., and Belzile, F. (2020). Fast-GBS v2.0: an analysis toolkit for genotyping-by-sequencing data. *Genome* 63, 577–581. doi: 10.1139/gen-2020-0077

Tsai, H.-Y., Janss, L. L., Andersen, J. R., Orabi, J., Jensen, J. D., Jahoor, A., et al. (2020). Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. *Sci. Rep.* 10:3347. doi: 10.1038/s41598-020-60203-2

Vapnik, V. (1998). *Statistical Learning Theory*. New York, NY: Wiley.

Veromann-Jürgenson, L.-L., Brodribb, T. J., Niinemets, Ü, and Tosens, T. (2020). Pivotal role of mesophyll conductance in shaping photosynthetic performance across 67 structurally diverse Gymnosperm species. *Int. J. Plant Sci.* 181, 116–128. doi: 10.1086/706089

Vuong, T., Sonah, H., Meinhardt, C., Deshmukh, R., Kadam, S., Nelson, R., et al. (2015). Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. *BMC Genomics* 16:593. doi: 10.1186/s12864-015-1811-y

Wang, J. (2019). *Pattern Discovery for Genome-wide Base Composition Evolution and Genetic Dissection of NDVI with UAV-based Remote Sensing in Crops.* Ph.D. dissertation. Ames, IA: Iowa State University.

Wang, J., Chu, S., Zhang, H., Zhu, Y., Cheng, H., and Yu, D. (2016). Development and application of a novel genome-wide SNP array reveals domestication history in soybean. *Sci. Rep.* 6:20728. doi: 10.1038/srep20728

Wang, L., Liu, F., Hao, X., Wang, W., Xing, G., Luo, J., et al. (2021). Identification of the QTL-allele system underlying two high-throughput physiological traits in the Chinese soybean germplasm population. *Front. Genet.* 12:600444. doi: 10.3389/fgene.2021.600444

Wang, L., Yang, Y., Zhang, S., Che, Z., Yuan, W., and Yu, D. (2020). GWAS reveals two novel loci for photosynthesis-related traits in soybean. *Mol. Genet. Genomics* 295, 705–716. doi: 10.1007/s00438-020-01661-1

Wang, P., and Grimm, B. (2021). Connecting chlorophyll metabolism with accumulation of the photosynthetic *Apparatus*. *Trends Plant Sci.* 26, 484–495. doi: 10.1016/j.tplants.2020.12.005

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2013). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* 29, 149–159. doi: 10.1093/bioinformatics/bts655

Wen, Y.-J., Zhang, H., Ni, Y.-L., Huang, B., Zhang, J., Feng, J.-Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinformatics* 19, 700–712. doi: 10.1093/bib/bbw145

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. (2001). Feature selection for SVMs. *Adv. Neural Inf. Process. Syst.* 13, 668–674.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2020). A unified approach for inference on algorithm-agnostic variable importance. *arXiv* [preprint] arXiv:2004.03683

Woodson, J. D., Joens, M. S., Sinson, A. B., Gilkerson, J., Salomé, P. A., Weigel, D., et al. (2015). Ubiquitin facilitates a quality-control pathway that removes damaged chloroplasts. *Science* 350, 450–454. doi: 10.1126/science.aac7444

Wu, Y., and Liu, Y. (2009). Variable selection in quantile regression. *Stat. Sin.* 19:801.

Xavier, A., and Rainey, K. M. (2020). Quantitative genomic dissection of soybean yield components. *G3 Genes Genomes Genet.* 10, 665–675. doi: 10.1534/g3.119.400896

Xu, Y., Yang, T., Zhou, Y., Yin, S., Li, P., Liu, J., et al. (2018). Genome-Wide association mapping of starch pasting properties in maize using single-locus and multi-locus models. *Front. Plant Sci.* 9:1311. doi: 10.3389/fpls.2018.01311

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., et al. (2020). Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant* 13, 187–214. doi: 10.1016/j.molp.2020.01.008

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., et al. (2021). rmvp: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics Proteomics Bioinformatics* (in press). doi: 10.1016/j.gpb.2020.10.007

Yoosefzadeh Najafabadi, M. (2021). *Using Advanced Proximal Sensing and Genotyping Tools Combined with Bigdata Analysis Methods to Improve Soybean Yield.* Ph.D. thesis. Guelph, ON: University of Guelph.

Yoosefzadeh-Najafabadi, M., Earl, H. J., Tulpan, D., Sulik, J., and Eskandari, M. (2021a). Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. *Front. Plant Sci.* 11:624273. doi: 10.3389/fpls.2020.624273

Yoosefzadeh-Najafabadi, M., Tulpan, D., and Eskandari, M. (2021b). Using hybrid artificial intelligence and evolutionary optimization algorithms for estimating soybean yield and fresh biomass using hyperspectral vegetation indices. *Rem. Sens.* 13:2555. doi: 10.3390/rs13132555

Yoosefzadeh-Najafabadi, M., Tulpan, D., and Eskandari, M. (2021c). Application of machine learning and genetic optimization algorithms for modeling and optimizing soybean yield using its component traits. *PLoS One* 16:e0250665. doi: 10.1371/journal.pone.0250665

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhang, D., Lü, H., Chu, S., Zhang, H., Zhang, H., Yang, Y., et al. (2017). The genetic architecture of water-soluble protein content and its genetic relationship to total protein content in soybean. *Sci. Rep.* 7:5053. doi: 10.1038/s41598-017-04685-7

Zhang, F., Tang, W., Hedtke, B., Zhong, L., Liu, L., Peng, L., et al. (2014). Tetrapyrrole biosynthetic enzyme protoporphyrinogen IX oxidase 1 is required for plastid RNA editing. *Proc. Natl. Acad. Sci. U.S.A.* 111, 2023–2028. doi: 10.1073/pnas.1316183111

Zhang, J., Wang, X., Lu, Y., Bhusal, S. J., Song, Q., Cregan, P. B., et al. (2018). Genome-wide scan for seed composition provides insights into soybean quality improvement and the impacts of domestication and breeding. *Mol. Plant* 11, 460–472. doi: 10.1016/j.molp.2017.12.016

Zhang, X., Zhao, J., Yang, G., Liu, J., Cao, J., Li, C., et al. (2019). Establishment of plot-yield prediction models in soybean breeding programs using UAV-Based hyperspectral remote sensing. *Rem. Sens.* 11:2752. doi: 10.3390/rs11232752

Zhong, H., Liu, S., Meng, X., Sun, T., Deng, Y., Kong, W., et al. (2021). Uncovering the genetic mechanisms regulating panicle architecture in rice with GPWAS and GWAS. *BMC Genomics* 22:86. doi: 10.1186/s12864-021-07391-x

Zhong, S., Dekkers, J. C., Fernando, R. L., and Jannink, J.-L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* 182, 355–364. doi: 10.1534/genetics.108.098277

Zhou, W., Bellis, E. S., Stubblefield, J., Causey, J., Qualls, J., Walker, K., et al. (2019). Minor QTLs mining through the combination of GWAS and machine learning feature selection. *bioRxiv* [preprint] doi: 10.1101/702761

Ziliak, S. (2017). P values and the search for significance. *Nat. Methods* 14, 3–4. doi: 10.1038/nmeth.4120

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.