frontiers in Plant Science

ORIGINAL RESEARCH
published: 25 February 2020
doi: 10.3389/fpls.2019.01794

Comparing Different Statistical Models and Multiple Testing Corrections for Association Mapping in Soybean and Maize

Avjinder S. Kaler[1], *Jason D. Gillman*[2], *Timothy Beissinger*[3] *and Larry C. Purcell*[1]*

[1] Department of Crop, Soil, and Environmental Sciences, University of Arkansas, Fayetteville, AR, United States, [2] Plant Genetic Research Unit, USDA-ARS, Columbia, MO, United States, [3] Division of Plant Breeding Methodology, Center for Integrated Breeding Research, Georg-August-Universität, Göttingen, Germany

Association mapping (AM) is a powerful tool for fine mapping complex trait variation down to nucleotide sequences by exploiting historical recombination events. A major problem in AM is controlling false positives that can arise from population structure and family relatedness. False positives are often controlled by incorporating covariates for structure and kinship in mixed linear models (MLM). These MLM-based methods are single locus models and can introduce false negatives due to over fitting of the model. In this study, eight different statistical models, ranging from single-locus to multilocus, were compared for AM for three traits differing in heritability in two crop species: soybean (*Glycine max* L.) and maize (*Zea mays* L.). Soybean and maize were chosen, in part, due to their highly differentiated rate of linkage disequilibrium (LD) decay, which can influence false positive and false negative rates. The fixed and random model circulating probability unification (FarmCPU) performed better than other models based on an analysis of Q-Q plots and on the identification of the known number of quantitative trait loci (QTLs) in a simulated data set. These results indicate that the FarmCPU controls both false positives and false negatives. Six qualitative traits in soybean with known published genomic positions were also used to compare these models, and results indicated that the FarmCPU consistently identified a single highly significant SNP closest to these known published genes. Multiple comparison adjustments (Bonferroni, false discovery rate, and positive false discovery rate) were compared for these models using a simulated trait having 60% heritability and 20 QTLs. Multiple comparison adjustments were overly conservative for MLM, CMLM, ECMLM, and MLMM and did not find any significant markers; in contrast, ANOVA, GLM, and SUPER models found an excessive number of markers, far more than 20 QTLs. The FarmCPU model, using less conservative methods (false discovery rate, and positive false discovery rate) identified 10 QTLs, which was closer to the simulated number of QTLs than the number found by other models.

Keywords: association mapping, genome-wide association analyses, multiple testing correction, statistical model analysis, quantitative trait loci

OPEN ACCESS

Edited by:
Marco Maccaferri, University of Bologna, Italy

Reviewed by:
Ralf Uptmoor, University of Rostock, Germany
Steven B. Cannon, United States Department of Agriculture, United States

***Correspondence:**
Larry C. Purcell
lpurcell@uark.edu

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

Received: *19 July 2019*
Accepted: *23 December 2019*
Published: *25 February 2020*

Citation:
Kaler AS, Gillman JD, Beissinger T and Purcell LC (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

INTRODUCTION

Connecting genotype to phenotype, known as genetic mapping, is important for modern crop breeding and improvement (Mackay, 2001). Two of the most commonly used approaches for genetic mapping are association mapping (AM) and biparental linkage mapping (LM). AM is an alternative approach to traditional mapping of biparental populations and is currently widely used in plant, animal (Goddard and Hayes, 2009), model species (Brachi et al., 2010), and human genetics (Risch and Merikangas, 1996; Nordborg and Tavaré, 2002). Most important traits in plants are complex and controlled by many genes and influenced by environment. With advancements in high throughput genotyping and sequencing technologies, single nucleotide polymorphisms (SNPs) provide relatively low cost and dense marker coverage across various genomes (Syvänen, 2005). Genotyping diverse lines provides thousands of SNPs across the genome that enables fine mapping complex trait variation down to nucleotide sequences by exploiting historical recombination events (Zhu et al., 2008). AM has lower overall statistical power to detect rare alleles and epistatic interactions than traditional LM, but it has several advantages, which include increased mapping resolution, broader allele coverage, reduced time and cost compared to developing biparental mapping populations, and potentially greater number of alleles evaluated (Yu et al., 2006).

Associating functional variants (alleles, loci) to phenotypes is a fundamental aim of both AM and LM (Botstein and Risch, 2003). The detection of quantitative trait loci (QTLs) through AM depends on the level of linkage disequilibrium (LD) between functional loci and markers. LD refers to a nonrandom association of alleles at different loci. LD is influenced by physical linkage and recombination, but it is a separate phenomenon—unlinked loci can be in a state of LD, and linked loci can be in a state of linkage equilibrium (Terwilliger and Weiss, 1998). The level of LD extent in a specific set of experimental genotypes can be measured statistically and has been widely leveraged in plants and animals to map and clone genes controlling complex genetic traits (Risch and Merikangas, 1996; Dunning et al., 2000; Pritchard and Przeworski, 2001; Nordborg and Tavaré, 2002). LD can be measured based upon the correlation between alleles at pairs of loci as physical distance between the loci increases. Outcrossing crop species, such as maize, have more genetic diversity (Remington et al., 2001; Yan et al., 2009) and also more rapid LD decay than self-pollinated species such as soybean, which has less overall genetic diversity (Gupta et al., 2006; Hyten et al., 2007; Schmutz et al., 2010; Song et al., 2015). Species with faster LD decay over physical distance, as compared to those with slow LD decay, require higher marker density over the genome to capture associations between marker and phenotype (Yu et al., 2006).

Several statistical models are available to identify associations between marker loci and numerous phenotypes that range from simple to increasingly complex. As genotypic data are becoming more readily available, accurately decoding the complexity of traits in a diverse population is only possible if accurate and more comprehensive statistical models can distinguish true biological associations from false positives arising from population structure and family relatedness without overcorrecting and resulting in false negatives. Using covariates for structure and kinship in the statistical model can control these confounding factors. STRUCTURE (Pritchard et al., 2000), principal component analysis (PCA) (Price et al., 2006), and a discriminant analysis of principal components (DAPC) (Jombart et al., 2010) are approaches that use genetic markers to determine population organization. Results from STRUCTURE and PCA are similar, but PCA is generally more commonly used due to lower required computational resources and time to generate covariates. False positives can also arise due to more recent common ancestry and family relatedness, which can be controlled by inclusion of a kinship matrix into the linear model. Identity-by-state is one of the most commonly used approaches to estimate familial relatedness among individuals in a diverse population (Loiselle et al., 1995).

The incorporation of population structure and a kinship matrix as covariates in mixed linear models (MLM) has become a popular approach to control false positives. Since the first MLM of AM was published by Yu et al. (2006), many MLM-based methods have been proposed (Zhang et al., 2010; Wang et al., 2014). All these models are single-locus models, which means that they comprise a one-dimensional genome scan by testing one marker at a time, iteratively for every marker in a dataset. This single-locus approach fails to match the true genetic model of complex traits that are controlled by numerous loci simultaneously. To cope with this problem, multilocus AM models have been recommended because these models consider the information of all loci simultaneously (Wang et al., 2016). MLM-based models can also induce false negatives due to over fitting of the model where some potentially important associations can be missed (Liu et al., 2016).

False negatives in AM can result when multiple comparison adjustments are used to determine statistical significance. Two commonly used multiple comparison methods in AM are Bonferroni correction (Holm, 1979) and false discovery rate (FDR) (Benjamini and Hochberg, 1995), which select the significant threshold. However, overly conservative thresholds can lead to high false negative error rates. Therefore, selection of an appropriate model and threshold are important steps in identifying markers that are truly associated with specific traits and which could be located within or very close to genes that control the trait variation, while controlling both false-positive and false-negative associations.

The objective of this study was to compare eight different AM statistical models, ranging from single to multilocus, for three previously reported traits and six simulated traits in soybean and maize. These crops were selected because of their difference in LD as indicated by the LD decay rate: maize, which is naturally outcrossing, displays much more rapid LD decay than soybean, a self-pollinating species. We also compared these eight statistical models for six qualitative traits in soybean, all of which have known causal genes with published genomic positions. Finally,

we evaluated five multiple comparison methods when used in conjunction with these eight AM models.

MATERIALS AND METHODS

Data Collection

This study included three datasets collected from previously published or online sources (referred to as "previously reported traits" subsequently). These previously reported datasets were the best linear unbiased predictions (BLUP) values across different environments for each trait. We also simulated six datasets from two crop species: soybean and maize. Previously reported data for soybean included canopy wilting (CW) with a broad sense heritability (H) of 80% (Kaler et al., 2017a), carbon isotope ratio ($\delta^{13}C$, H = 60%, Kaler et al., 2017b), and oxygen isotope ratio ($\delta^{18}O$, H = 20%, Kaler et al., 2017b). For maize, the previously reported data included days to tasseling (DT, H = 85%), ear height (EH, H = 80%), and ear diameter (ED, H = 85%) (Flint-Garcia et al., 2005). For both soybean and maize, six traits were simulated that varied in heritability and the number of QTLs (Q). These simulated traits were generated using the same genotypic markers that were used for AM of previously reported data. These six simulated datasets for each crop had varying heritabilities and genetic architectures. We simulated traits with H = 20% and Q = 20 (H20_Q20), H = 60% and Q = 20 (H60_Q20), H = 80% and Q = 20 (H80_Q20), H = 20% and Q = 40 (H20_Q40), H = 60% and Q = 40 (H60_Q40), and H = 80% and Q = 40 (H80_Q40). The R-script to generate the simulated data sets is provided in a supplement (**Table S1**). These data were simulated to have random QTLs effects. The simulated data for soybean and maize are provided in **Supplementary Data Files 1** and **2**, respectively. Previously reported data of soybean consisted of 346 accessions as described by Kaler et al. (2017b), and previously reported data of maize (Flint-Garcia et al., 2005) consisted of 279 accessions from the Panzea database website (www.panzea.org).

Genotypic Data and LD

Genotypic data for both crops consisted of SNP markers. In soybean, SNP data were obtained from the Illumina Infinium SoySNP50K iSelect SNP BeadChip that provided 42,509 SNP markers for all 346 accessions (Song et al., 2013; Song et al., 2015). In maize, SNP data were obtained from the Illumina MaizeSNP50 BeadChip that provided 50,896 SNP markers for 273 accessions (Flint-Garcia et al., 2005). Quality control checks were performed, which included removing monomorphic markers, markers with minor allele frequency (MAF) ≤ 5%, and markers with a missing rate higher than 10%. An LD-kNNi method, which is based on a k-nearest-neighbor-genotype, was applied to impute the remaining missing marker datasets (Money et al., 2015).

After performing quality controls, 31,260 SNPs for soybean and 48,833 SNPs for maize with MAF > 5% were used for AM. For maize, SNPs were more or less equally distributed across the genome for both euchromatic and heterochromatic regions (**Figures S1**). For soybean, SNPs were not equally distributed across the genome; there was higher marker density in euchromatic than heterochromatic regions (**Figure S2**). All chromosomes of maize had more SNPs than those of soybean (**Table S2**). The decay rate of LD was estimated using the GAPIT R package (Lipka et al., 2012). The decay rate of LD was much greater in maize than soybean with an average LD across all chromosomes decaying to r^2 = 0.25 in less than 1 kb. In comparison, in soybean, an average LD across all chromosomes decayed to r^2 = 0.25 in approximately 2,000 kb (**Figure S3**). In soybean, LD decay rates were different in euchromatic and heterochromatic regions (Hyten et al., 2007; Schmutz et al., 2010; Kaler et al., 2017a; Kaler et al., 2017b). Using both regions together affected the results of LD decay rate.

Broad sense heritability of traits was calculated using the formula: $H = \sigma^2_G \ / \ (\sigma^2_G + (\frac{\sigma^2_\epsilon}{r}))$, where σ^2_G is the genotypic variance, σ^2_ϵ is the residual variance, and r is the number of replications. Marker-based narrow sense heritability (h^2) was estimated to understand the variation and trend of predictive ability across traits (Kruijer et al., 2015) using the GAPIT R package. In the GAPIT package, the MLM model can be described as: $Y=X\beta+Zu+e$, where Y is the vector of observed phenotypes; β is an unknown vector containing fixed effects, including the genetic marker, population structure (Q), and the intercept; u is an unknown vector of random additive genetic effects from multiple background QTL for individuals/lines; X and Z are the known design matrices; and e is the unobserved vector of residuals. The u and e vectors are assumed to be normally distributed with a null mean and a variance of: $Var \begin{pmatrix} u \\ e \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}$, where $G = \sigma^2_a K$ with σ^2_a as the additive genetic variance and K as the kinship matrix. Homogeneous variance is assumed for the residual effect; i.e., $R = \sigma^2_e I$, where σ^2_e is the residual variance. The proportion of the total variance explained by the genetic variance is defined as marker-based heritability.

Description of AM Models

The eight AM models evaluated ranged from simple to complex and included: (i), analysis of variance (ANOVA), (ii) general linear model (GLM) with PCA (principle component analysis) (Price et al., 2006), (iii) MLM with PCA + K (Kinship matrix for family relatedness estimates) (Yu et al., 2006), (iv) compressed MLM (Zhang et al., 2010), (v) enriched compressed MLM (Li et al., 2014), (vi) settlement of MLM under progressively exclusive relationship (SUPER) (Wang et al., 2014), (vii) multiple loci MLM (MLMM) (Segura et al., 2012), and (viii) fixed and random model circulating probability unification (FarmCPU) (Liu et al., 2016). Models from (i) to (vi) are single locus models, and (vii) and (viii) are multilocus models. **Table 1** lists and briefly summarizes keys aspects of models evaluated in the present study.

The GLM with PCA model is expected to reduce the false positives that arise due to only population structure (Price et al., 2006). The MLM with PCA and K model includes the kinship matrix in the model and is expected to reduce the false positives that arise from family relatedness (Yu et al., 2006). Both GLM and MLM are reported to control false positives better than

TABLE 1 | Description of eight genome-wide association mapping models.

Model	Description	References
Analysis of variance (ANOVA)	Single locus analysis, the null hypothesis of an ANOVA using a single SNP is that there is no difference between the trait means of any genotype group.	Lewis, 2002
General Linear Model (GLM)	Single locus analysis, the GLM uses principle components as covariates in the model to reduce the false positives that arise due to only population structure.	Price et al., 2006
Mixed Linear Model (MLM)	Single locus analysis, the MLM uses principle components and kinship matrix in the model to reduce the false positives that arise from the family relatedness and population structure.	Yu et al., 2006
Compressed MLM (CMLM)	Single locus analysis, the CMLM clusters the individuals into groups and fits genetic values of groups as random effects in the model that improves statistical power compared to regular MLM methods.	Zhang et al., 2010
Enriched CMLM (ECMLM)	Single locus analysis, the ECMLM calculates kinship using several different algorithms and then chooses the best combination between kinship algorithms and grouping algorithms.	Li et al., 2014
Settlement of MLM Under Progressively Exclusive Relationship (SUPER)	Single locus analysis, the SUPER model uses the associated genetic markers (pseudo Quantitative Trait Nucleotides), instead of all the markers, to derive kinship. Whenever a pseudo QTN is correlated with the testing marker, it is excluded from those used to derive kinship.	Wang et al., 2014
Multiple Loci Mixed Linear Model (MLMM)	Multi-locus analysis, the MLMM incorporates a kinship matrix and selected cofactors, performed better with regard to the false-discovery rate and the QTL detection power than a model incorporating only a kinship matrix or only cofactors.	Segura et al., 2012
Fixed and random model Circulating Probability Unification (FarmCPU)	Multi-locus analysis, this model uses a modified MLM method, Multiple Loci Linear Mixed Model (MLMM), and incorporates multiple markers simultaneously as covariates in a stepwise MLM to partially remove the confounding between testing markers and kinship. To completely eliminate the confounding, MLMM is divided into two parts: Fixed Effect Model (FEM) and a Random Effect Model (REM) and uses them iteratively. FEM contains testing markers, one at a time, and multiple associated markers as covariates to control false positives. To avoid model over-fitting in FEM, the associated markers are estimated in REM by using them to define kinship. The *P*-values of testing markers and the associated markers are unified at each iteration.	Liu et al., 2016

ANOVA (Price et al., 2006; Yu et al., 2006). The MLM model is reported to perform better than the GLM model alone by controlling false positives (Yu et al., 2006). Advantages of the MLM model to control false positives disappear for complex traits when they are associated with population structure having extensive genetic divergence. The MLM model controls the *P*-value inflation well, but it also leads to false negatives, thereby weakening identification of true associations (Zhang et al., 2010). To deal with this problem, the compressed MLM model (CMLM) was developed, which clusters the individuals into groups and fits genetic values of groups as random effects in the model (Zhang et al., 2010). The CMLM method improves statistical power compared to regular MLM methods (Zhang et al., 2010). Another suggested way to deal with *P*-value deflation due to MLM is to use a SUPER model in which only the associated genetic markers, instead of all the markers, are used as pseudo Quantitative Trait Nucleotides (QTNs) to derive kinship (Wang et al., 2014). Whenever a pseudo QTN is correlated with the testing marker, it is excluded from those used to derive kinship. The SUPER model applies a threshold on LD between the pseudo QTNs and the testing marker. This method improves the statistical power compared to using overall kinship from all markers.

FarmCPU is a multilocus model that was developed to control false positives without comprising false negatives (Liu et al., 2016). This model is not used extensively for AM of complex traits in crops because it has not been compared with other models for previously reported and simulated data. The FarmCPU model uses a modified MLM method, multiple loci linear mixed model (MLMM), and incorporates multiple markers simultaneously as covariates in a stepwise MLM to partially remove the confounding between testing markers and kinship. To completely eliminate the confounding, MLMM is divided into two parts: fixed effect model (FEM) and a random effect model (REM) and uses them iteratively. FEM contains testing markers, one at a time, and multiple associated markers as covariates to control false positives. To avoid model overfitting in FEM, the associated markers are estimated in REM by using them to define kinship. The *P*-values of testing markers and the associated markers are unified at each iteration. This model reportedly improves statistical power, increases computational efficiency, and the ability to control false positives and false negatives as compared to other models (Liu et al., 2016).

Interpretation of Q-Q Plots and Model Evaluation

Examining quantile-quantile (Q-Q) plots is one of the most common ways of determining if models control false positives and false negatives (Stich et al., 2008; Stich and Melchinger, 2009; Würschum et al., 2012; Riedelsheimer et al., 2012; Kristensen et al., 2018). The Q-Q plot shows the expected negative-log of association probability (X-axis) across all markers versus the observed negative-log of association probability values (Y-axis). If a Q-Q plot has a straight line close to the 1:1 line without any tail, then it follows a uniform distribution, which means the null hypothesis is true and that there is no significant association or causal polymorphism. Any deviation of this straight line would indicate that the null hypothesis was not true and there were significant associations present. If the Q-Q plot does not have a straight line and tail, it indicates that there are false positives when a line inflates upward and there are false negatives when line deflates downward. If a Q-Q plot has a straight line, close to the 1:1 line, with a sharp upward deviated tail, it indicates that both false positives and false negatives were controlled, and that there are true associations and causal polymorphisms. This happens because most of the *P*-values observed follow a

uniform distribution (i.e., they are not in LD with a causal polymorphism, so the null hypothesis is true) but the few that are in LD with a causal polymorphism will produce significant *P*-values [extremely low = extremely high -log (*P*-values)] and these are in the "tail".

We evaluated these eight models for false positives and false negatives based on the Q-Q plots. A sharp deviation from the expected *P*-value distribution in the tail area would indicate that a model appropriately controlled both false positives and false negatives. Models were also compared using qualitative traits in soybean, which have known published genes for flower color (Takahashi et al., 2010), stem termination (Bernard, 1972), seed-coat luster (Gijzen et al., 2003), seed-coat color (Clough et al., 2004), hilum color (Carpentieri-Pipolo et al., 2015), and pubescence color (Toda et al., 2002; Zabala and Vodkin, 2003). Models were also compared using simulated data in which there were a known number of QTLs in the simulated data. The accuracy of a model was evaluated by identifying the number of QTLs in the simulated data.

Evaluation of Multiple Comparisons Methods for AM

Three common multiple comparison methods were compared for determining statistical significance with a cutoff of $P = 0.05$. These methods included Bonferroni, false discovery rate, and positive false discovery rate. These comparisons were made using the PROC MULTTEST procedure of SAS version 9.4 (SAS Institute, 2013). The models were also compared to no multiple comparison adjustment at a *P*-value of 0.0003.

RESULTS

Phenotype Descriptions

There were broad phenotypic ranges for all the traits evaluated in both soybean and maize (**Table 2**), which is required for dissecting complex traits through association analysis (McCarthy et al., 2008). Among the three traits in maize, broad and marker-based narrow sense heritability ranged between 80% to 85% and 70% to 80%, respectively. Among the three traits in soybean, broad and marker-based narrow sense heritability ranged between 20% to 80% and 3% to 71%, respectively (**Table 2**).

TABLE 2 | Descriptive statistics of days to tasseling (DT), ear height (EH), and ear diameter (ED) in maize, and canopy wilting (CW), carbon isotope ratio ($\delta^{13}C$), and oxygen isotope ratio ($\delta^{18}O$) in soybean.

	Maize			Soybean		
	DT	EH	ED	CW	$\delta^{13}C$	$\delta^{18}O$
Mean	67.58	61.38	36.74	16.99	−29.06	20.87
Standard Deviation	5.75	20.27	4.05	6.46	0.27	0.43
Minimum	54.50	8	23.72	7.50	−29.81	19.20
Maximum	85.00	136	46.35	45.63	−28.37	22.29
Skewness	0.41	0.64	−0.29	1.39	−0.12	−0.11
Range	30.50	128.00	22.63	38.13	1.46	3.09
Count	279	279	279	346	346	346
Broad sense heritability (%)	85	80	85	80	60	20
Narrow sense heritability (%)	70	72	80	71	29	3

Model Comparison With Soybean Data

Eight different AM models that ranged from simple to complex were compared using three previously reported traits and six simulated traits for soybean and maize (**Figures 1** and **2**). These eight AM models identified different numbers of significant markers associated with the previously reported and simulated traits for soybean when we consider the same significance threshold (**Table 3**). For example, if we consider the significance threshold as −Log10 (P) > 3.5 to declare a significant association for a simulated trait with 20 QTLs, we identified 2465 SNPs from ANOVA, 520 from GLM, 24 from MLM, 24 from CMLM, 16 from ECMLM, 229 from SUPER, 26 from MLMM, and 19 from FarmCPU (**Table 3**). All models, except the FarmCPU and MLMM, identified multiple significant SNP marker associations in close physical distance on the chromosome. These large peaks were generated because one SNP from these peaks had the highest significant association with traits, but the other markers at a given peak were in high LD with this most significant marker.

For the CW trait (H = 80%), the ANOVA, GLM, and SUPER models had a large number of false positives as indicated by a substantial inflation of *P*-values (**Figure 1A**). Q-Q plots of complex models including MLM, CMLM, and ECMLM had a straight line with a slightly deviated tail, which indicated that these models reduced the false positives. However, most markers were close to the straight line of 1:1, indicating that they may have been reported as false negatives (**Figure 1A**). In contrast, the FarmCPU model followed a straight line close to 1:1, with a sharp upward deviated tail, indicating that this model controlled both false positives and false negatives (**Figure 1A**). For $\delta^{13}C$ (moderate H = 60%), results of all models were similar to the CW trait, indicating that the FarmCPU model controlled both false positives and false negatives more effectively than other models (**Figure 1B**). For a low heritability trait, $\delta^{18}O$, the Q-Q plot for all models, except FarmCPU, deflated downward, indicating that these models increased false negatives. In contrast, Q-Q plots of the FarmCPU model for $\delta^{18}O$ had a straight line close to the 1:1 with a slightly deviated tail, indicating that FarmCPU controlled both false positives and false negatives (**Figure 1C**).

Results from Q-Q plots of the six simulated traits in soybean were consistent with results from the previously reported data (**Figures 1D–I**). That is, ANOVA, GLM, and SUPER models had an inflation of *P*-values indicating there were a large number of false positives whereas MLM, CMLM, ECMLM, and MLMM controlled false positives but not false negatives. The Q-Q plots for FarmCPU indicated control of both false negatives and false positives. For all simulated traits, the ANOVA model had a large number of false positives because it inflated the *P*-value in the Q-Q plots (**Figures 1D–I**). When a simulated trait had a low heritability (H = 20) and a large QTL number (40), all complex models that incorporated the PCs and kinship matrix increased the number of false negatives, except the FarmCPU model (**Figure 1G**). When a simulated trait had a high heritability with 20 or 40 QTLs, complex models that included

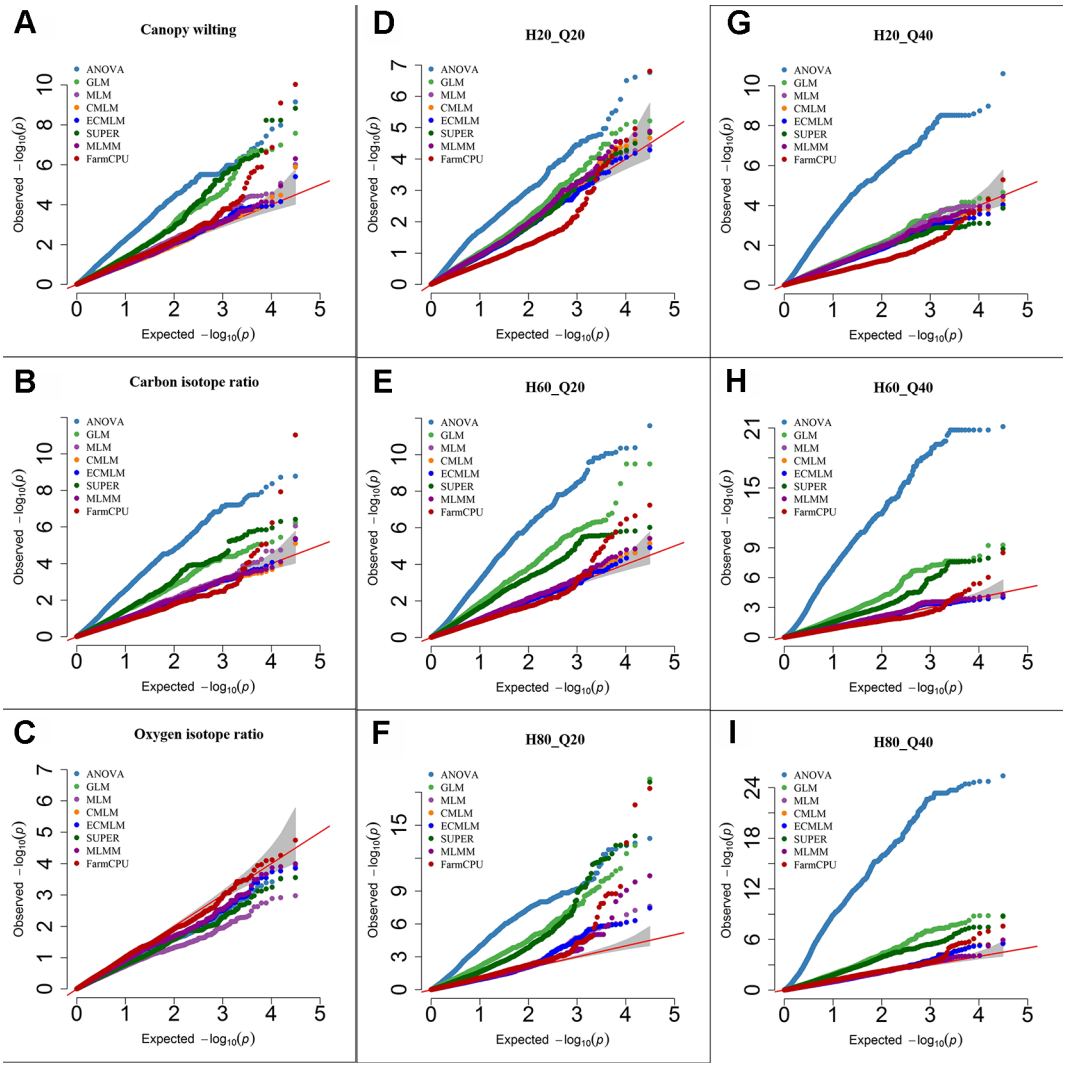

FIGURE 1 | Quantile-quantile (QQ) plots of the eight models including Analysis of Variance (ANOVA), General Linear Model (GLM), Mixed Linear Model (MLM), Compressed MLM (CMLM), Enriched Compressed MLM (ECMLM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), Multiple Loci Mixed linear Model (MLMM), and Fixed and random model Circulating Probability Unification (FarmCPU) for three real traits including canopy wilting **(A)**, carbon isotope ratio **(B)**, and oxygen isotope ratio **(C)**, and six simulated traits that varied in heritability (H) and quantitative trait loci (Q) including H = 20% and Q = 20 (H20_Q20) **(D)**, H = 60% and Q = 20 (H60_Q20) **(E)**, H = 80% and Q = 20 (H80_Q20) **(F)**, H = 20% and Q = 40 (H20_Q40) **(G)**, H = 60% and Q = 40 (H60_Q40) **(H)**, and H = 80% and Q = 40 (H80_Q40) **(I)** in Soybean. The grey area represents the 95% concentration band.

both PCs and kinship matrix reduced the false positives (**Figures 1F, I**), but still the FarmCPU model had a straight line that followed the 1:1 line with a sharp deviated tail compared to other models.

Maize

In maize, we observed large effects of population structure and family relatedness. Similar to soybean (**Table 3**), the models that included the PCs and kinship matrix for maize identified a smaller number of markers than models that did not (data not shown). Likewise, the models that had no adjustment (ANOVA) or included only PCs (GLM) increased the number of significant markers for both previously reported and simulated traits when a specific threshold level was used compared with other complex models (**Table 3**). All single-locus models gave a peak of multiple, significant SNPs, which may result in missing the identification of other important genomic regions that may not have that high level of significance (*P*-value) as the markers in the peak region that are in high LD with the most significant marker. However, the multilocus model, FarmCPU and MLMM, did not show any clusters of significant markers in maize; instead they provided the highest significant marker at a specific genomic location, which led to identification of more markers at different locations (data not shown). Based on the Q-Q plots for all

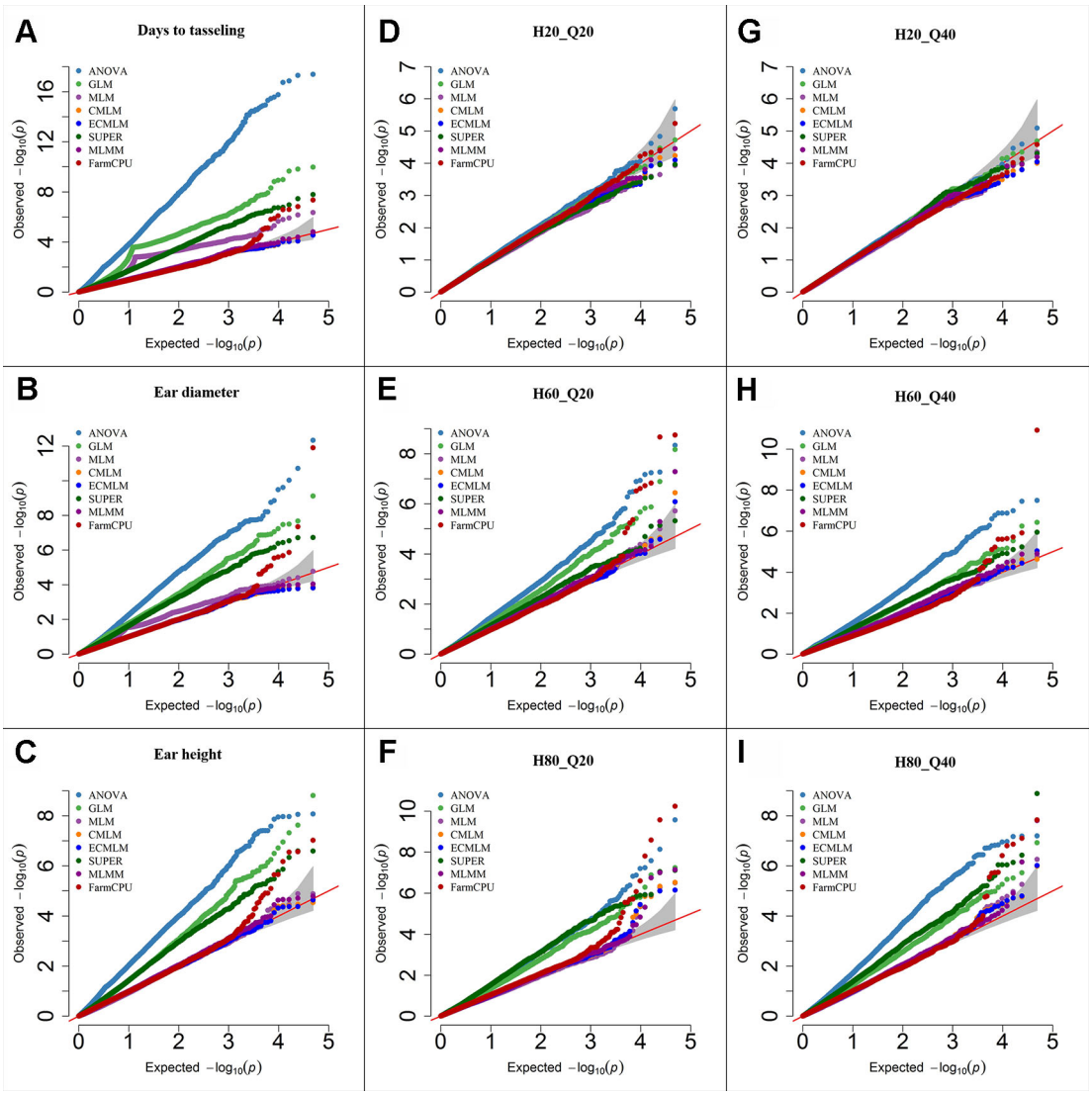

FIGURE 2 | Quantile-quantile (QQ) plots of the eight models including Analysis of Variance (ANOVA), General Linear Model (GLM), Mixed Linear Model (MLM), Compressed MLM (CMLM), Enriched Compressed MLM (ECMLM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), Multiple Loci Mixed linear Model (MLMM), and Fixed and random model Circulating Probability Unification (FarmCPU) for three real traits including days to tasseling **(A)**, ear diameter **(B)**, and ear height **(C)**, and six simulated traits that varied in heritability (H) and quantitative trait loci (Q) including H = 20% and Q = 20 (H20_Q20) **(D)**, H = 60% and Q = 20 (H60_Q20) **(E)**, H = 80% and Q = 20 (H80_Q20) **(F)**, 2 = 20% and Q = 40 (H20_Q40) **(G)**, H = 60% and Q = 40 (H60_Q40) **(H)**, and H = 80% and Q = 40 (H80_Q40) **(I)** in Maize. The grey area represents the 95% concentration band.

previously reported and simulated traits, the FarmCPU model performed much better than other models as indicated by the Q-Q plots with a straight line close to the 1:1 line with most sharply deviated tail (**Figure 2**).

Qualitative Traits of Soybean

Flower color in soybean is a qualitative trait that is conferred by the *W1* gene. A small (65 bp) insertion of tandem repeats in exon 3 that truncates the translation product prematurely, resulting in a white flower instead of the wild-type purple flower (Zabala and Vodkin, 2005; Zabala and Vodkin, 2007). The *W1* locus is located on Gm13 at 4552540-4557331 base pairs in the Wm82.a1.v1.1 genomic assembly (Schmutz et al., 2010). As both alleles are widespread in soybean germplasm, this trait is ideal to determine which model would best identify markers closely linked to the known causative allele. The FarmCPU, GLM, and ANOVA models identified the most significant SNP associated with flower color on Gm13 (**Figure 3**). Other models, except the MLMM, identified most significant markers at different positions on Gm13 that were further away from the published gene on the same chromosome. For example, MLM identified the highest significant SNP at 3,822,639 base pairs.

TABLE 3 | Comparison of the number of significant markers ($P \leq 0.05$) identified by multiple comparison methods including Bonferroni (Bon), false discovery rate (FDR), and positive false discovery rate (PFDR) using a simulated trait for soybean that had a heritability 60% and 20 QTLs (H60_Q20) in eight different association models eight including analysis of variance (ANOVA), general linear model (GLM), mixed linear model (MLM), compressed MLM (CMLM), enriched compressed MLM (ECMLM), settlement of MLM under progressively exclusive relationship (SUPER), multiple loci mixed linear model (MLMM), and fixed and random model circulating probability unification (FarmCPU).

	$-Log_{10}$ $P \geq 3.5$	Bon	FDR	PFDR
ANOVA	2,465	411	7,204	9,760
GLM	520	38	1,336	1,966
MLM	24	0	0	0
CMLM	24	0	0	0
ECMLM	16	0	0	0
SUPER	229	5	327	521
MLMM	26	0	0	0
FarmCPU	19	4	10	10

For comparative purposes, a P-value threshold ($-Log_{10}$ P ≥ 3.5) without any adjustment is included.

MLMM identified the highest significant SNP on Gm19. Unlike the other models, FarmCPU identified only the single SNP on Gm13 at the position 4,559,799 bp, closest to the position of *W1*gene, in the Wm82.a1.v1.1 genomic assembly (Schmutz et al., 2010) (**Figure 3**).

Similar results were observed when models were compared using five other qualitative traits in soybean including hilum color, pubescence color, seed-coat color, seed-coat luster, and stem termination (determinacy) (data not shown). **Figure 4** shows the comparison of FarmCPU models with MLM for different qualitative traits. We chose a comparison of FarmCPU with MLM because it is a commonly used model for AM. The FarmCPU model identified a single significant SNP close to the genes associated with qualitative traits, instead of identifying a large peak of SNPs with MLM (**Figure 4**). For example, the FarmCPU model identified the single most significant SNP associated with stem termination on Gm19 at the position 45,000,827 base pairs closest to the position of the *Dt1* gene (45,183,357– 45,185,175 base pairs), instead of a large peak of SNPs with MLM. For three qualitative traits including, hilum color, pubescence color, and stem termination, the identified significant SNPs with the largest $-Log_{10}$ *P*-value from the peak were similar to the position of the peak identified by the FarmCPU model (**Figure 4**). The most significant marker for hilum color and pubescence color was on Gm06 at the position 18,766,611 base pairs which was 28,586

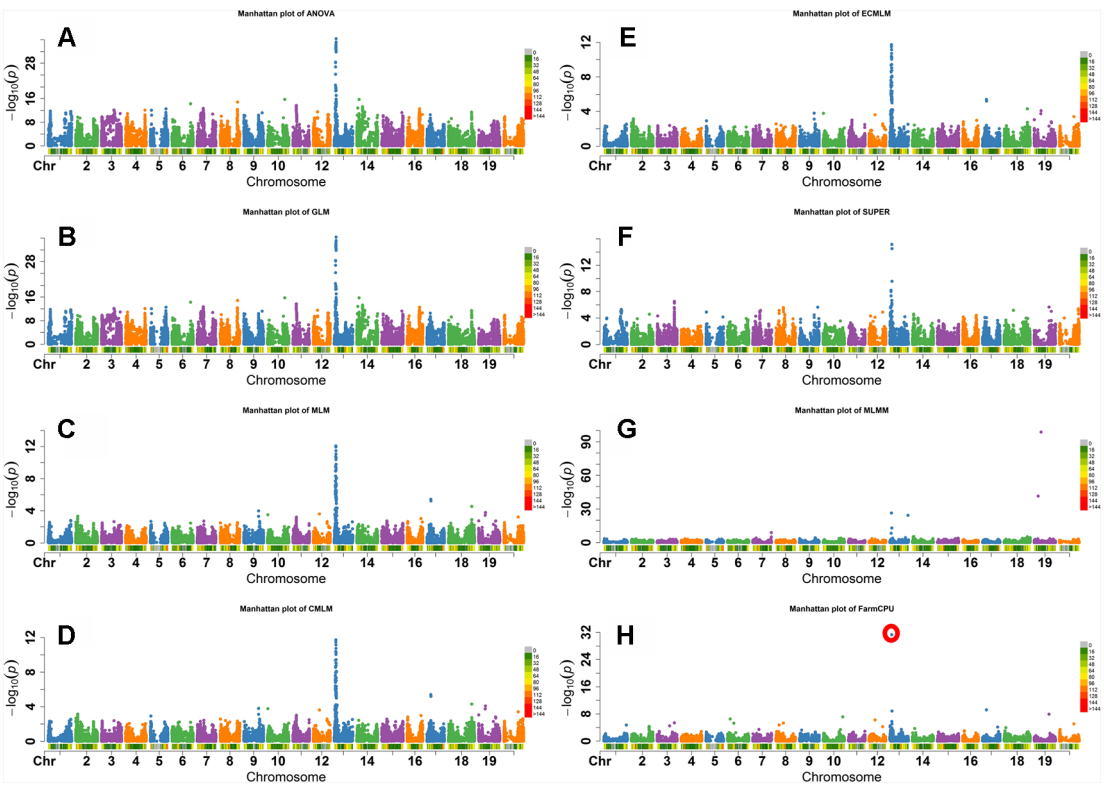

FIGURE 3 | Manhattan plots of -Log10 (P) vs. chromosomal position of SNP markers associated with flower color in soybean from eight models including Analysis of Variance (ANOVA) **(A)**, General Linear Model (GLM) **(B)**, Mixed Linear Model (MLM) **(C)**, Compressed MLM (CMLM) **(D)**, Enriched Compressed MLM (ECMLM) **(E)**, Settlement of MLM Under Progressively Exclusive Relationship (SUPER) **(F)**, Multiple Loci Mixed linear Model (MLMM) **(G)**, and Fixed and random model Circulating Probability Unification (FarmCPU) **(H)**.

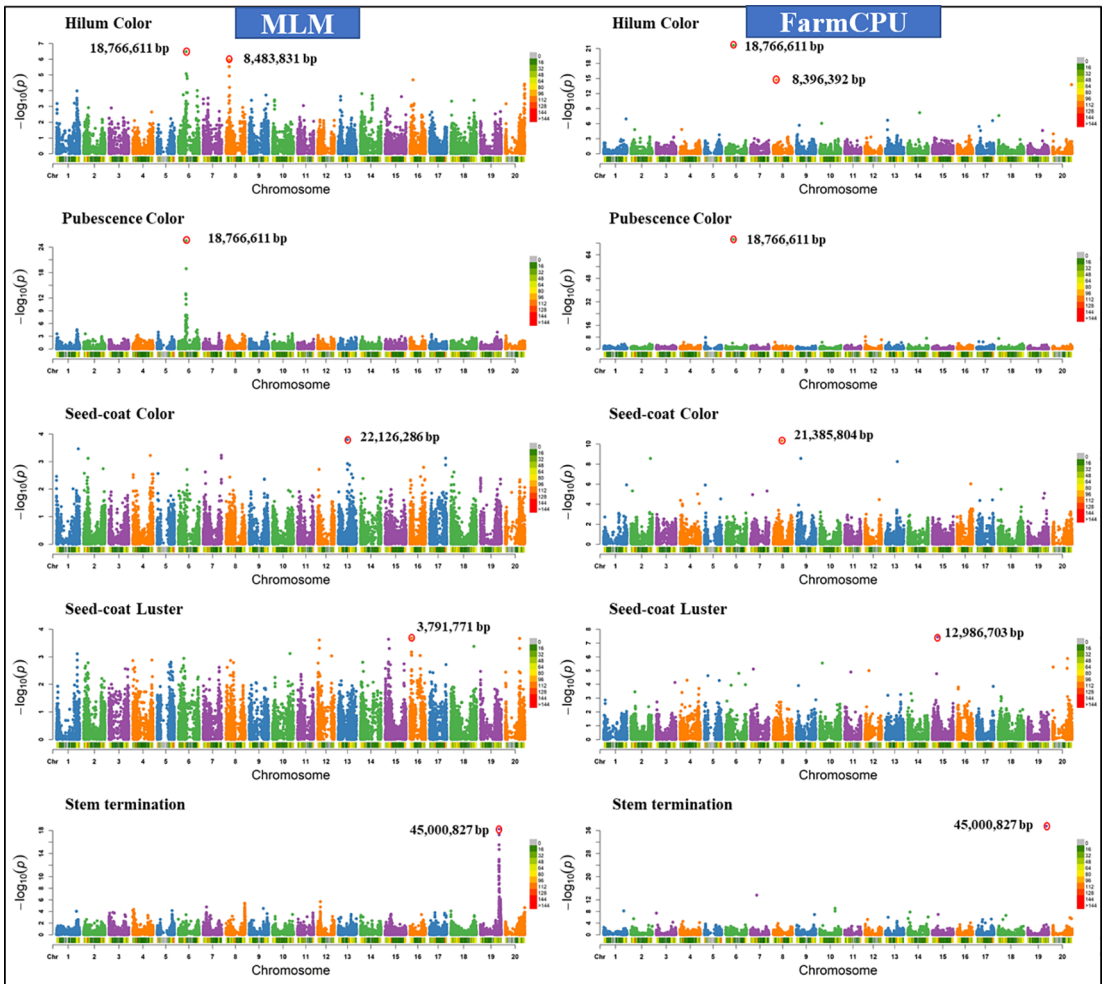

FIGURE 4 | Manhattan plots of –Log10 (P) vs. chromosomal position of SNP markers associated with five qualitative traits in soybean from two models including, Mixed Linear Model (MLM) and Fixed and random model Circulating Probability Unification (FarmCPU).

base pairs distant from the T locus. Pubescence coloration and hilum coloration are in part determined by loss of function mutation affecting the Glyma06g21920 gene which results in grey pubescence at plant maturity and, in the right genetic background, can result in buff or imperfect hila (Zabala and Vodkin, 2003).

For seed-coat color and luster, the MLM identified the most significant SNPs on different chromosomes compared to the FarmCPU model (**Figure 4**). For seed-coat color, the FarmCPU model identified the most significant SNP on Gm8 at 21,385,804 base pairs close to the *Glyma08g27050* gene (21392963-21395430 bp), which is involved in flavonol biosynthesis (Palmer et al., 2004; Zabala and Vodkin, 2007; Yang et al., 2010). This marker association is most likely reporting the natural gene silencing cluster which results in yellow seed-coats (Clough et al., 2004; Tuteja et al., 2004) However, MLM identified the most significant SNP on Gm13 at the position 22,126,286 base pairs, where there was no gene present associated with seed coat color.

For seed-coat luster, the FarmCPU model identified the most significant SNP on Gm15 at 12,986,703 base pairs within a gene *Glyma15g16670* (12,982,823-12,987,622), which is involved in a function of epidermis development. In contrast, MLM identified the SNP on Gm16 at 3,791,771 base pairs, which was not located close to any known gene for seed coat luster. Seed coat color and luster are controlled by more than one gene, hence, the FarmCPU model identified additional significant SNPs on other chromosomes, and all those regions are located close to previously reported genes for these traits (Palmer et al., 2004; Yang et al., 2010).

Multiple Comparisons Methods for AM

Different multiple comparison methods were compared for determining statistical significance with a cutoff of $P = 0.05$. These comparison methods included: Bonferroni, false discovery rate, and positive false discovery rate (Holm, 1979; Hommel, 1988; Hochberg, 1988; Benjamini and Hochberg, 1995). The

significance level of –Log10 ($P > 1.3$, which is an equivalent to the $P < 0.05$) was used as a threshold before and after performing multiple comparison adjustments. We compared these methods for all traits in maize and soybean, but for brevity, we only show results of the simulated trait, H60_Q20 in soybean (**Table 3**). Results of all other traits were consistent with this trait (data not shown). These results indicated that these multiple comparison methods for AM were very conservative and only depended on the *P*-value of the association test. These results were evaluated based on the number of markers identified after adjustments. If the number of markers identified was more than 20, it indicates that there were false positives. If the number of markers identified was less than 20, it indicates that there were false negatives. In this study, ANOVA, GLM, and SUPER models had very large –Log10 *P*-values for significant associations; when multiple comparisons adjustments were performed, all 20 QTLs were above the significance level of –Log10 $P > 1.3$ (**Table 3**). Checking the Q-Q plots for this trait (**Figure 1E**), indicated that the ANOVA, GLM, and SUPER models did not control false positives well in a diverse population. The FDR and PFDR methods gave more false positives than the Bonferroni method in the ANOVA, GLM, and SUPER models (**Table 3**). Complex models (MLM, CMLM, ECMLM, and MLMM), which were expected to control false positives arising from population structure and family relatedness, did not identify any significant associations after performing multiple comparisons adjustments. These complex models reduced the *P*-value inflations (**Table 3**), which led to an increase in the false negative error rates. For the FarmCPU model, the Bonferroni adjustment identified 4 out of 20 highly significant associations, which means that these methods gave 16 false negative associations; the false discovery rate and positive false discovery rate adjustments with the FarmCPU model, identified 10 out of 20 highly significant associations, resulting in 10 false negatives above the selected cutoff value of –Log10 $P > 1.3$ (which is an equivalent to the $P < 0.05$) after adjustments (**Table 3**). Without any multiple comparison adjustments, FarmCPU identified 19 out 20 associations at a cutoff of 3.5 (–Log10 (P) ≥ 3.5; $P \leq 0.0003$).

DISCUSSION

AM is based on the LD of marker with a QTL and is a popular approach for fine mapping traits of interest. LD in an AM population can also result from population structure, family relatedness, selection, and genetic drift (Flint-Garcia et al., 2003; Yu et al., 2006), which are the major reasons of false positive associations. The success of AM to identify true associations depends on the ability to separate LD of the marker with a QTL from LD due to other causes. There is a need for an appropriate model that can correctly identify LD caused by population structure and family relatedness.

In this study, eight different statistical models, ranging from single to multilocus, were compared for AM of three empirical phenotype traits differing in heritability in two crop species, soybean and maize, that vary in LD decay rates. The power of SNP identification is determined by several factors including the size of the population, the population structure, the extent of LD in the population, the heritability and underlying genetic architecture of the trait (Yu et al., 2006). For all previously reported traits, several SNPs were identified in this study, which indicated that all these traits were complex, quantitative traits, controlled by a large number of genes with small effects. The power of detecting SNP and mapping resolution for complex traits depend on the LD exploited in the population by the statistical model (Yu et al., 2006). As expected, faster LD decay over physical distance was observed in maize compared to soybean because maize is a cross-pollinated with a higher recombination rate and soybean is a self-pollinated with a lower recombination rate.

Based on the Q-Q plots, we observed a nonuniform distribution of *P* values in the ANOVA, GLM, and SUPER models of all empirical traits (**Figures 1** and **2**). These results are similar to previous studies (Yu et al., 2006; Stich et al., 2008; Zhao et al., 2011) indicating that these models are inappropriate for AM of complex traits in plants because they generate spurious marker-trait associations. Complex models including MLM, CMLM, and ECMLM were proposed to correct population structure and family relatedness (Yu et al., 2006). We observed a straight line close to the 1:1 line with slightly deviated tail in the Q-Q plots of MLM, CMLM, and ECMLM, indicating that these models reduced the false positives, but increased false negatives because most significant markers were present close to the 1:1 line. These false negatives were generated due to the overfitting of these complex models. Similar results were observed in other studies (Wen et al., 2015; Tamba et al., 2017; Li et al., 2018; Wen et al., 2018) where these complex models generated more false negatives. In contrast, the Q-Q plot of the FarmCPU model, a multilocus model, controlled both false positives and false negatives as indicated by a straight line (close to the 1:1 line) with a sharp deviated tail for all empirical traits in both crops.

Some studies, where multilocus models, including mrMLM (Wang et al., 2016), FASTmrEMMA (Wen et al., 2018), and LASSO (ISIS EM-BLASSO) (Tamba et al., 2017), were used, performed better than MLM-based models. Liu et al. (2016) reported that the FarmCPU model avoids overfitting by using two types of adjustments for testing markers. The first type of adjustment was fitting covariates of population structure, family relatedness, and pseudo-quantitative trait nucleotides; the second type of adjustment either refines how family relatedness is derived from all the markers, or selectively includes or excludes pseudo-quantitative trait nucleotides based on their relationship with the testing markers.

These eight AM models were also compared based on simulated traits in which a known number of QTLs were simulated. Among these models, the FarmCPU model identified the number of QTLs close to the number of simulated QTLs for all traits in both crops. Comparison of Q-Q plots of different models for all simulated traits indicated that the FarmCPU controlled better the false positives and false negatives. Additionally, FarmCPU identified markers of qualitative traits closer to the published location of genes

controlling these traits compared to the other models. Instead of providing a large peak as in other models, the FarmCPU model provided a single most significant marker, which was always present closest to the published genes.

For determining statistical significance in AM, different multiple comparison methods are used with a cutoff of $P = 0.05$, and several of these methods were compared when used in combination with the eight AM models. Complex models (MLM, CMLM, ECMLM, and MLMM) were particularly conservative and did not find any markers after adjustment; these complex models and multiple comparison methods are apparently increasing the number of false negatives. In contrast, ANOVA, GLM, and SUPER models identified more than 20 QTLs after multiple comparison adjustments, indicating that these models increased the false positives. In contrast, the FarmCPU model performed better than other models for these multicomparison adjustments by identifying 10 QTLs with less conservative methods, FDR and PFDR. Based on the Q-Q plots and the number of known simulated QTLs, the FarmCPU was an appropriate model for controlling false positives and false negatives compared to other models. Other multiple comparison methods were overly conservative for selection of significant threshold for AM. Determination of the correct significant threshold for AM can be determined by an empirical relationship based upon marker-based heritability (Kaler and Purcell, 2019).

CONCLUSIONS

This study compared eight statistical models for AM of three empirical phenotypic traits differing in heritability and six simulated traits in two crop species, soybean, and maize, varying in LD decays rates. Based on the Q-Q plots and the number of known simulated QTLs, the FarmCPU was an appropriate model for controlling false positives and false negatives compared to other models. These finding were also supported by the AM of six qualitative traits, which identified a single most significant SNP closest to the known published genes. The FarmCPU model performed better for multiple comparison adjustments compared to other models because adjustments were overly conservative for MLM, CMLM, ECMLM, and MLMM and did not find any QTL. In contrast, for ANOVA, GLM, and SUPER models, these adjustments found more than 20 QTLs. From this study, we conclude that FarmCPU provides a robust model for AM of complex traits in plants, which effectively controls both false positives and false negatives.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

AK and LP conceived of the idea and wrote the manuscript. AK performed simulations and data analysis. JG and TB provided theoretical insights and valuable edits. All authors read and approved the final version.

FUNDING

Partial funding from this project was from the United Soybean Board (project #1920-172-0116-A). Additional funds were provided by the University of Arkansas System, Division of Agriculture and the USDA-ARS.

ACKNOWLEDGMENTS

The authors gratefully acknowledge partial funding of this research from the United Soybean Board. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. The USDA is an equal opportunity provider and employer.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01794/full#supplementary-material

REFERENCES

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

Bernard, R. L. (1972). Two genes affecting stem termination in soybean. *Crop. Sci.* 12, 235–239. doi: 10.2135/cropsci1972.0011183X001200020028x

Botstein, D., and Risch, N. (2003). Discovering genotypes underlying human phenotypes, past successes for Mendelian disease, future approaches for complex disease. *Nat. Genet.* 33, 228–237. doi: 10.1038/ng1090

Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., et al. (2010). Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLOS Genet.* 6 (5), e1000940. doi: 10.1371/journal.pgen.1000940

Carpentieri-Pipolo, V., Almeida, L. A., and Kiihl, R. A. S. (2015). Inheritance of R locus expressing brown hilum on black seed coat in soybean. *Am. J. Plant Sci.* 06, 1857–1861. doi: 10.4236/ajps.2015.611186

Clough, S. J., Tuteja, J. H., Li, M., Marek, L. F., Shoemaker, R. C., and Vodkin, L. O. (2004). Features of a 103-kb gene-rich region in soybean include an inverted perfect repeat cluster of CHS genes comprising the I locus. *Genome* 47, 819–831. doi: 10.1139/g04-049

Dunning, A. M., Durocher, F., Healey, C. S., Teare, M. D., McBride, S. E., Carlomagno, F., et al. (2000). The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* 67, 1544–1554. doi: 10.1086/316906

Flint-Garcia, S. A., Thornsberry, J., and Buckler, E. (2003). Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54, 357–374. doi: 10.1146/annurev.arplant.54.031902.134907

Flint-Garcia, S. A., Thuillet, A. C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., et al. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* 44, 1054–1064. doi: 10.1111/j.1365-313X.2005.02591.x

Gijzen, M., Weng, C., Kuflu, K., Woodrow, L., Yu, K., and Poysa, V. (2003). Soybean seed lustre phenotype and surface protein co-segregate and map to linkage group E. *Genome Nat. Res. Council Canada* 46, 659–664. doi: 10.1139/g03-047

Goddard, M. E., and Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* 10 (6), 381–391. doi: 10.1038/nrg2575

Gupta, P. K., Rustgi, S., and Kumar, N. (2006). Genetic and molecular basis of grain size and grain number and its relevance to grain productivity in higher plants. *Genome* 49, 565–571. doi: 10.1139/g06-063

Hochberg, Y. (1988). A sharper Bonferonni procedure for multiple tests of significance. *Biometrika* 75, 800–803. doi: 10.1093/biomet/75.4.800

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75, 383–386. doi: 10.1093/biomet/75.2.383

Hyten, D. L., Choi, I.-Y., Song, Q., Shoemaker, R. C., Nelson, R. L., et al. (2007). Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 175, 1937–1944. doi: 10.1534/genetics.106.069740

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components, a new method for the analysis of genetically structured populations. *BMC Genet.* 11, 94. doi: 10.1186/1471-2156-11-94

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

Kaler, A. S., Ray, J. D., King, C. A., Schapaugh, W. T., and Purcell, L. C. (2017a). Genome-wide association mapping of canopy wilting in diverse soybean genotypes. *Theor. Appl. Genet.* 130, 2203–2221. doi: 10.1007/s00122-017-2951-z

Kaler, A. S., Dhanapal, A. P., Ray, J. D., King, C. A., Fritsch, F. B., and Purcell, L. C. (2017b). 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

Kristensen, P. S., Jahoor, A., Andersen, J. R., Cericola, F., Orabi, J., Janss, L. L., et al. (2018). Genome-wide association studies and comparison of models and cross-validation strategies for genomic prediction of quality traits in advanced winter wheat breeding lines. *Front. Plant Sci.* 9(69), 69. doi: 10.3389/fpls.2018.00069

Kruijer, W., Boer, M. P., Malosetti, M., Flood, P. J., Engel, B., et al. (2015). Marker-based estimation of heritability in immortal populations. *Genetics* 199, 379–398. doi: 10.1534/genetics.114.167916

Lewis, C. M. (2002). Genetic association studies: design, analysis and interpretation. *Brief. Bioinform.* 3, 146–153.

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y.-M., Todhunter, R. J., et al. (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol.* 12, 73. doi: 10.1186/s12915-014-0073-5

Li, C., Huang, Y., Huang, R., Wu, Y., and Wang, W. (2018). The genetic architecture of amylose biosynthesis in maize kernel. *Plant Biotechnol. J.* 16, 688–695. doi: 10.1111/pbi.12821

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT, genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399. doi: 10.1093/bioinformatics/bts444

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 (2), e1005767. doi: 10.1371/journal.pgen.1005767

Loiselle, B. A., Sork, V. L., Nason, J., and Graham, C. (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am. J. Bot.* 82, 1420–1425. doi: 10.1002/j.1537-2197.1995.tb12679.x

Mackay, T. F. (2001). The genetic architecture of quantitative traits. *Annu. Rev. Genet.* 35, 303–339. doi: 10.1146/annurev.genet.35.102401.090633

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., et al (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9 (5), 356–369.

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G. Y., and Myles, S. (2015). LinkImpute fast and accurate genotype imputation for non-model organisms. *G3* 5 (11), 2383–2390. doi: 10.1534/g3.115.021667

Nordborg, M., and Tavaré, S. (2002). Linkage disequilibrium, what history has to tell us. *Trends Genet.* 18 (2), 83–90. doi: 10.1016/S0168-9525(02)02557-X

Palmer, R. G., Pfeiffer, T. W., Buss, G. R., Kilen, T. C., Boerma, H. R., and Specht, J. E. (2004). "Qualitative genetics, Soybeans, improvement, production, and uses," in *Madison (WI) ASA, CSSA, and SSSA*, *3rd edn*, 137–214.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38 (8), 904–909. doi: 10.1038/ng1847

Pritchard, J. K., and Przeworski, M. (2001). Linkage disequilibrium in humans, models and data. *Am. J. Hum. Genet.* 69 (1), 1–14. doi: 10.1086/321275

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., et al. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci.* 98, 11479–11484. doi: 10.1073/pnas.201394398

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., et al. (2012). Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proc. Natl. Acad. Sci.* 109(23), 8872–8877. doi: 10.1073/pnas.1120813109

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* 273 (5281), 1516–1517. doi: 10.1126/science.273.5281.1516

SAS Institute (2013). *The SAS System for Windows. Version 9.3* (Cary, NC: SAS Inst. Inc.).

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463 (7278), 178–183. doi: 10.1038/nature08670

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE* 8 (1), e54985. doi: 10.1371/journal.pone.0054985

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2015). Fingerprinting soybean germplasm and its utility in genomic research. *Genes* 50 (10), 1999–2006. *3*. doi: 10.1534/g3.115.019000

Stich, B., and Melchinger, A. E. (2009). Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and arabidopsis. *BMC Genom.* 2710, 94. doi: 10.1186/1471-2164-10-94

Stich, B., Mohring, J., Piepho, H. P., Heckenberger, M., Buckler, E. S., et al. (2008). Comparison of mixed-model approaches for association mapping. *Genetics* 178, 1745–1754. doi: 10.1534/genetics.107.079707

Syvänen, A. C. (2005). Toward genome-wide SNP genotyping. *Nat. Genet.* 37, S5–10. doi: 10.1038/ng1558

Takahashi, R., Dubouzet, J. G., Matsumura, H., Yasuda, K., and Iwashina, T. (2010). A new allele of flower color gene W1 encoding flavonoid 3'5'-hydroxylase is responsible for light purple flowers in wild soybean *Glycine soja*. *BMC Plant Biol.* 10, 155. doi: 10.1186/1471-2229-10-155

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13, e1005357. doi: 10.1371/journal.pcbi.1005357

Terwilliger, J., and Weiss, K. (1998). Linkage disequilibrium mapping of complex diseases, fantasy or reality? *Curr. Opin. Biotechnol.* 1998 9, 578–594. doi: 10.1016/S0958-1669(98)80135-3

Toda, K., Yang, D., Yamanaka, N., Watanabe, S., Harada, K., and Takahashi, R. (2002). A single-base deletion in soybean flavonoid 3'-hydroxylase gene is associated with gray pubescence color. *Plant Mol. Biol.* 50, 187–196. doi: 10.1023/A:1016087221334

Tuteja, J. H., Clough, S. J., Chan, W. C., and Vodkin, L. O. (2004). Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in *Glycine max*. *Plant Cell* 16 (4), 819–835. doi: 10.1105/tpc.021352

Würschum, T., Tucker, M. R., Reif, J. C., and Maurer, H. P. (2012). Improved efficiency of doubled haploid generation in hexaploid triticale by *in vitro* chromosome doubling. *BMC Plant Biol.* 12, 109. doi: 10.1186/1471-2229-12-109

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., and Zhang, Z. (2014). A SUPER powerful method for genome wide association study. *PLoS ONE* 9, e107684. doi: 10.1371/journal.pone.0107684

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

Wen, Z., Boyse, J. F., Song, Q., Cregan, P. B., and Wang, D. (2015). Genomic consequences of selection and genome-wide association mapping in soybean. *BMC Genom.* 16 (1), 671. doi: 10.1186/s12864-015-1872-y

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. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Yan, J., Shah, T., Warburton, M. L., Buckler, E. S., McMullen, M. D., and Crouch, J. (2009). Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS ONE* 4 (12), e8451. doi: 10.1371/journal.pone.0008451

Yang, K., Jeong, N., Moon, J. K., Lee, Y. H., Lee, S. H., Kim, H. M., et al. (2010). Genetic analysis of genes controlling natural variation of seed coat and flower colors in soybean. *J. Hered.* 101, 757–768. doi: 10.1093/jhered/esq078

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

Zabala, G., and Vodkin, L. O. (2003). Cloning of the pleiotropic T locus in soybean and two recessive alleles that differentially affect structure and expression of the encoded flavonoid 3' hydroxylase. *Genetics* 163, 295–309.

Zabala, G., and Vodkin, L. O. (2005). The wp mutation of *Glycine max* carries a gene-fragment-rich transposon of the CACTA superfamily. *Plant Cell* 17 (10), 2619–2632. doi: 10.1105/tpc.105.033506

Zabala, G., and Vodkin, L. O. (2007). A rearrangement resulting in small tandem repeats in the F3′5′H gene of white flower genotypes is associated with the soybean W1 locus. *Crop Sci.* 47, S2. doi: 10.2135/cropsci2006.12.0838tpg

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhao, K., Tung, C. –W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* 2 (1), 467. doi: 10.1038/ncomms1467

Zhu, C., Gore, M. A., Buckler, E. S., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genom.* 1, 5–20. doi: 10.3835/plantgenome2008.02.0089

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.