Received: 5 August 2020 | Accepted: 18 November 2020

DOI: 10.1002/tpg2.20077

The Plant Genome

REVIEW & INTERPRETATION

Status and prospects of genome-wide association studies in plants

Laura Tibbs Cortes[1] | **Zhiwu Zhang**[2] | **Jianming Yu**[1]

[1] Department of Agronomy, Iowa State University, Ames, IA 50010, USA

[2] Department of Crop and Soil Sciences, Washington State University, Pullman, WA 99164, USA

Correspondence
Jianming Yu, Dep. of Agronomy, Iowa State Univ., Ames, IA, 50010, USA.
Email: jmyu@iastate.edu

Abstract

Genome-wide association studies (GWAS) have developed into a powerful and ubiquitous tool for the investigation of complex traits. In large part, this was fueled by advances in genomic technology, enabling us to examine genome-wide genetic variants across diverse genetic materials. The development of the mixed model framework for GWAS dramatically reduced the number of false positives compared with naïve methods. Building on this foundation, many methods have since been developed to increase computational speed or improve statistical power in GWAS. These methods have allowed the detection of genomic variants associated with either traditional agronomic phenotypes or biochemical and molecular phenotypes. In turn, these associations enable applications in gene cloning and in accelerated crop breeding through marker assisted selection or genetic engineering. Current topics of investigation include rare-variant analysis, synthetic associations, optimizing the choice of GWAS model, and utilizing GWAS results to advance knowledge of biological processes. Ongoing research in these areas will facilitate further advances in GWAS methods and their applications.

1 | CURRENT STATUS OF GWAS

Understanding the genetic architecture of complex traits is fundamental to understanding biology. Most traits of agricultural and evolutionary importance are complex traits that are influenced by many genetic loci and environmental conditions as well as their interaction (Mackay, Stone, & Ayroles, 2009). Advances in genomic technology and methodology development and a desire to examine trait variation across diverse genetic backgrounds were the major driving forces behind the initial wave of association mapping studies in model plant and crop species (Zhu, Gore, Buckler, & Yu, 2008). Continued progress in sequencing technologies and coordinated community effort have made genome-wide association studies (GWAS) a method of choice, particularly when resequencing is conducted after the assemblage of the reference genome or when a high-density genotyping array becomes

Abbreviations: BLINK, Bayesian information and linkage disequilibrium iteratively nested keyway; CMLM, compressed mixed linear model; ECMLM, enriched compressed mixed linear model; EMMA, efficient mixed-model association; EMMAX, efficient mixed-model association expedited; FarmCPU, fixed and random model circulating probability unification; FaST-LMM, factored spectrally transformed linear mixed model; FDR, false discovery rate; GEMMA, genome-wide efficient mixed model analysis; GLM, general linear model; GP, genomic prediction; GS, genomic selection; GWAS, genome-wide association studies; LD, linkage disequilibrium; MLM, mixed linear model; MLMM, multi-locus mixed model; OWAS, omic-wide association studies; P3D, population parameters previously determined; PCA, principal component analysis; QTL, quantitative trait loci; QTN, quantitative trait nucleotide; REML, restricted maximum likelihood; RIL, recombinant inbred line; SLIDE, sliding-window approach for locally intercorrelated markers with asymptotic distribution errors corrected; SNP, single nucleotide polymorphism; SUPER, settlement of MLM under progressively exclusive relationship.

available (Michael & Jackson, 2013). High-throughput phenotyping has also been expanding the trait list for GWAS. For many crops, diversity panels and related genetic populations were built for GWAS, resulting in common resources for the research community including genetic material, phenotyping protocols, genotyping and sequencing platforms, analysis pipelines, and curated data.

In the decade since the previous review (Zhu et al., 2008) and many other earlier reviews (e.g., Sukumaran & Yu, 2014; Visscher, Brown, McCarthy, & Yang, 2012; Xiao, Liu, Wu, Warburton, & Yan, 2017), GWAS have transformed from a promising new tool to a powerful, ubiquitous technique for understanding complex traits in plants. Major highlights in GWAS include the following:

- Genome-wide association studies have investigated agriculturally important traits in many major crop species, including maize (*Zea mays* L.), wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), soybean [*Glycine max* (L.) Merr.], sorghum [*Sorghum bicolor* (L.) Moench], barley (*Hordeum vulgare* L.), cotton (*Gossypium hirsutum* L.), and numerous other crops beyond the model plant species *Arabidopsis* (e.g., Ersoz, Yu, & Buckler, 2007; Liu & Yan, 2019; Sukumaran & Yu, 2014; Varshney et al., 2012).
- Genome-wide association studies have identified genomic regions associated with many agronomic, physiological, and fitness traits including flowering time, plant height, kernel number, stress tolerance, and grain yield (e.g., Ersoz et al., 2007; Gupta, Kulwal, & Jaiswal, 2019; Liu & Yan, 2019; Sukumaran & Yu, 2014).
- Genome-wide association studies have also been used to study other types of phenotypes. Genome-wide association studies in rice have identified genes associated with geographical divergence and adaptation during domestication (Chen, Huang, Tian, Wing, & Han, 2019) as well as with biochemical and molecular phenotypes including flavonoid, fatty acid, amino acid, and nucleic acid metabolites (Chen et al., 2016). Data generated by high-throughput automated phenotyping have also been analyzed by GWAS. For example, GWAS in sorghum have detected significant associations for panicle architecture using automated feature extraction from images (Zhou et al., 2019) and for biomass traits using measurements taken by aerial drones (Spindel et al., 2018).
- Genome-wide association studies are used both to detect novel associations with valuable traits and to validate loci identified by other methods. Genome-wide association studies may be conducted as stand-alone investigations, as a component of gene cloning studies, or as the foundational step in marker-assisted selection, among other uses. In turn, exploiting this information accelerates crop breeding. For example, loci identified by GWAS on provitamin A levels in maize grain were used as the basis of marker-assisted and genomic selection for this important nutritional trait (Owens et al., 2014).
- Genome-wide association studies have also been used to enable genetic engineering, as in the case of transgenic drought-tolerant maize developed after detection of *ZmVPP1* by GWAS (S.B. Wang et al., 2016). As genome-editing technologies continue to improve, particularly those based on CRISPR (Zhang, Massel, Godwin, & Gao, 2018), the use of GWAS is expected to increase to identify target genes for editing.
- Genome-wide association studies were first developed in the context of human disease genetics (Lander, 1996; Lander & Kruglyak, 1995; Lander & Schork, 1994; Risch & Merikangas, 1996) and have led to the detection of thousands of genetic variants significantly associated with these diseases. New understanding gained from these GWAS has been clinically relevant, enabling the development of new therapeutic approaches for diseases ranging from schizophrenia to diabetes (Visscher et al., 2012, 2017).

Core Ideas

- GWAS dissect complex traits by testing genome-wide SNPs across an assembled population.
- Unified mixed-model GWAS control for both population structure and kinship.
- New GWAS methods build on this widely adopted mixed model foundation.
- Ongoing challenges call for the further development of GWAS methods and software.

2 | DEVELOPMENT AND HISTORY OF GWAS

In typical GWAS, phenotype and genotype data are collected for a large sample of assembled individuals such as a diversity panel. The genotype data usually consist of genome-wide single nucleotide polymorphisms (SNPs) identified through resequencing, genotyping-by-sequencing, or array-based genotyping. The genetic markers most associated with the phenotype of interest are found using statistical methods. While it is possible that a genetic marker detected in this way resides within a causative gene for the phenotype of interest, this is often not the case. Instead, GWAS rely on linkage disequilibrium (LD) between markers under testing and functional polymorphisms of causative genes. Loci that are physically near to one another on the chromosome are separated by recombination less often than are loci that are farther from each other. This nonrandom association of alleles at two loci is

called LD or gametic-phase disequilibrium. Those SNPs near the causative locus can be in high LD with the functional polymorphisms and thus associated with the phenotype of interest. Genome-wide association studies detect these associations and mark up the genomic regions harboring these significant SNPs and the implicated genes. If the time elapsed since the last common ancestor in which functional polymorphisms were generated through mutation is considerable in the unrelated populations typically desirable in GWAS, the genomic regions in LD can be narrow and are therefore well suited for mapping of the gene responsible in a high resolution (Lander & Schork, 1994; Lipka et al., 2015; Visscher et al., 2017; Xiao et al., 2017).

Linkage analysis for quantitative trait locus (QTL) mapping was the direct precursor of association studies including GWAS. Instead of assembling individuals into a diverse panel for GWAS, linkage mapping studies individuals with a known relationship. For example, linkage mapping analyses in crop species often use progeny purposefully generated from biparental crosses, either F_2 individuals or recombinant inbred lines (RILs). Genetic markers that are linked to the QTL will cosegregate with the phenotype of interest more often than expected by chance. Because the individuals studied are closely related in pedigree, fewer rounds of recombination have occurred since their most recent common ancestor, and therefore large linkage blocks are present. This means that the genetic markers used do not have to be as dense as those used in GWAS in order to ensure the detection of genomic regions harboring the causative locus. Once a QTL has been found and validated, the area can be targeted for fine mapping and QTL cloning (Korte & Farlow, 2013; Lander & Schork, 1994; Mackay et al., 2009; Miles & Wayne, 2008; Sukumaran & Yu, 2014). Before the advent of next-generation sequencing technologies, this was a considerable advantage. This allowed the first genome-wide QTL analysis to be conducted in tomato (*Solanum lycopersicum* L.) in 1988 (Paterson et al., 1988), 14 years before the publication of the first GWAS (Ozaki et al., 2002). Today, linkage analysis and GWAS are complementary approaches that can be used to understand complex traits in different populations.

Even while linkage analysis was the dominant method of understanding complex traits, the potential value of association studies was appreciated. Advantages of association studies compared with linkage studies include eliminating the need to perform experimental crosses (Lander & Schork, 1994), increasing power to detect genes with smaller effect sizes (Risch & Merikangas, 1996), and improving resolution with smaller blocks of LD (Lander & Schork, 1994). However, the requirement for high-density genotypes meant that the first association studies, carried out before the development of next-generation sequencing, could focus only on small subsets of the genome. This meant that an association study could only investigate a region of interest already identified by other methods. For example, an early association study of plant height and flowering time in maize focused on a single candidate gene that had been identified by previous mutagenesis and QTL mapping, *dwarf8*, by examining 123 polymorphisms in and near *dwarf8* along with 141 genome-wide markers (Thornsberry et al., 2001). Researchers at the time recognized, though, that a dense, genome-wide set of genetic markers would alleviate this problem and allow association studies to become genome-wide (Lander, 1996; Lander & Kruglyak, 1995; Lander & Schork, 1994; Risch & Merikangas, 1996).

The ideas behind GWAS were theoretically discussed beginning in the mid-1990s (Lander, 1996; Lander & Kruglyak, 1995; Lander & Schork, 1994; Risch & Merikangas, 1996). These early papers anticipated problems that are still being addressed today, including high rates of false positives as a result of population structure (Lander & Schork, 1994) and multiple testing (Lander & Kruglyak, 1995). However, their ideas had to wait to be put into practice until the publication of the draft human genome in 2001 (International Human Genome Sequencing Consortium, 2001) and the subsequent development of early SNP datasets such as dbSNP (Sherry et al., 2001) and HapMap (The International HapMap Consortium, 2003). In 2002, the first GWAS paper was published. This study, based on 65,000 SNPs and 94 individuals, reported genetic associations for myocardial infarction risk (Ozaki et al., 2002). Three years later, in 2005, the first GWAS in an area outside of human medical genetics was published in *Arabidopsis* (Aranzana et al., 2005). Soon after, GWAS papers based on thousands of SNPs were published for livestock (Abasht & Lamont, 2007) and crop species (Beló et al., 2008).

All association studies, genome-wide or otherwise, use statistical methods to associate genetic markers with phenotypes. Conceptually, GWAS aims to find those SNPs at which variation in genotype is significantly associated with variation in phenotype. At the simplest level, this can be accomplished by performing a statistical test such as an ANOVA on each SNP individually. The null hypothesis that there is no difference between the trait mean for any genotype group (i.e., AA, Aa, and aa) can then be tested for every SNP (Bush & Moore, 2012).

A major problem of this naïve approach is its high rate of false positives, which occur when a finding is declared significant even though it is not actually true. In part, this is because testing each SNP in the dataset results in thousands or even millions of statistical tests. To understand the problem with this, consider that the significance threshold of 0.05 commonly used in statistical tests means that the researcher is accepting a false positive rate of up to 5%. For a single test, this is an acceptable risk. However, as more markers are tested, the probability that at least of these test will be a false positive increases (Brzyski et al., 2017; Bush & Moore,

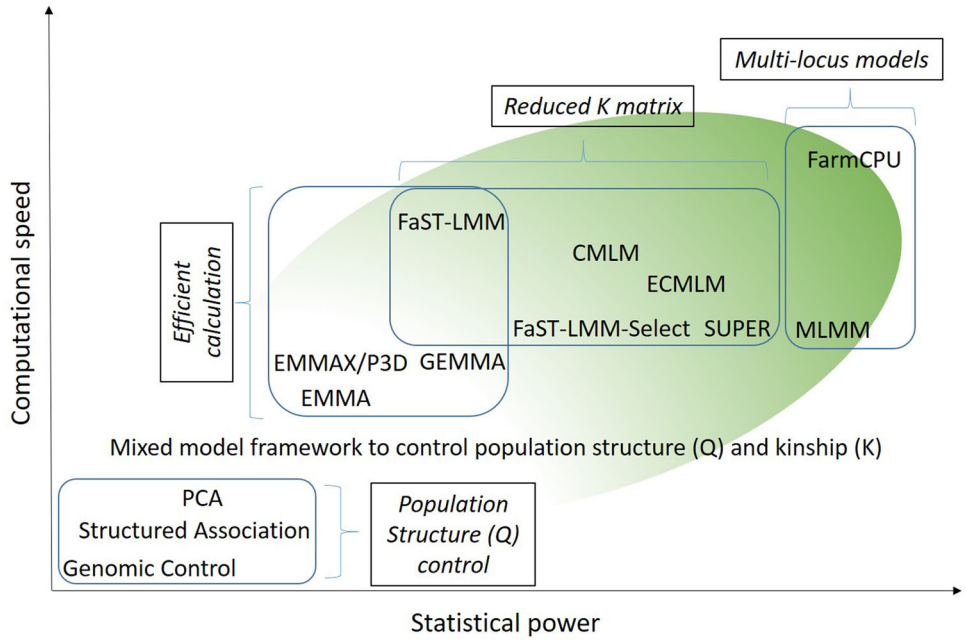

FIGURE 1 Genome-wide association study methods for improving computational speed and statistical power. Different methods are grouped by general strategy, and the position of each method shows the general trend of improved statistical power (shown on the *x* axis) and computational speed (shown on the *y* axis). CMLM, compressed mixed linear model; ECMLM, enriched compressed mixed linear model; EMMA, efficient mixed-model association; EMMAX, efficient mixed-model association expedited; FarmCPU, fixed and random model circulating probability unification; FaST-LMM, factored spectrally transformed linear mixed models; GEMMA, genome-wide efficient mixed model analysis; *K*, kinship; MLMM, multi-locus mixed-model; P3D, population parameters previously determined; PCA, principal component analysis; *Q*, population structure; SUPER, settlement of mixed linear model under progressively exclusive relationship

2012). Common methods of multiple testing correction include limiting the false discovery rate (FDR), which is the proportion of all positive results that are expected to be false positives (Benjamini & Hochberg, 1995; Storey & Tibshirani, 2003) or using the Bonferroni correction, which divides the desired significance threshold by the total number of tests conducted to determine the corrected significance threshold. However, setting an appropriate significance threshold presents additional challenges in the context of GWAS.

Another major contributor to false positives in naïve GWAS is relatedness among individuals. In the diverse populations typically used for GWAS, some individuals are more closely related to one another than they are to are others, forming subpopulations of related individuals within the population. The existence of unequal relationships within an assembled population for GWAS is not easy to minimize or avoid (Yu et al., 2006; Zhu & Yu, 2009). Because of this, SNPs that are more common in a given subpopulation may show spurious associations with the phenotype of interest if the phenotype happens to be present at a higher frequency in that group.

2.1 | Addressing population structure

Different methods of using genome-wide markers to control for population structure have been developed, and this continues to be an important research topic (Figure 1). The first of these methods was genomic control (Devlin & Roeder, 1999), which used null markers—markers unlikely to affect the trait of interest—to estimate the effect of population structure on the test statistic. This information was used to adjust the final *p* value for each marker, reducing false positives (Devlin & Roeder, 1999). Soon after, Pritchard, Stephens, and Donnelly (2000) developed the structured association (or STRUCTURE) method. Structured association also uses null markers but uses them to define a set of subpopulations within the dataset. The individuals are assigned to one or more of these subpopulations, and subpopulation membership is used as a cofactor within the association model (Pritchard et al., 2000). Adding cofactors to correct for population structure in this way is called the general linear model (GLM) (Pritchard et al., 2000).

These older methods have now been widely replaced by the mixed linear model (MLM) (Yu et al., 2006) (Figure 1). The

MLM accounts for relatedness at two levels: population structure (Q) and kinship (K) (Yu et al., 2006). Population structure can be determined from genotype data using STRUCTURE (Pritchard et al., 2000) or principal component analysis (Price et al., 2006). The kinship matrix estimates the relatedness among all individuals in the dataset using their genotype data (Yu et al., 2006). There are several algorithms available to calculate kinship. For example, the methods of Loiselle, Sork, Nason, and Graham (1995) and VanRaden (2008) both use allele frequencies and identity-by-state to estimate identity-by-descent and thereby kinship coefficients (Speed & Balding, 2015). Under the mixed model framework of MLM, control of false positives is realized by having both a fixed effect of population structure and a random effect of polygenic background that is defined by the kinship (Yu et al., 2006).

2.2 | Improving computational efficiency

Several methods have been introduced to increase the efficiency of solving MLM equations. The first of these to be developed was efficient mixed-model association (EMMA), which improved computational speed in part by eliminating redundant matrix operations at each iteration of computation of the likelihood function (Kang et al., 2008). In addition, the EMMA implementation can also be used to calculate the kinship matrix, which it calculates simply by using identity by state to produce a matrix of pairwise genetic similarity among all individuals.

Other methods improve computational speed using approximation. These include population parameters previously determined (P3D) (Zhang et al., 2010) and EMMA expedited (Kang et al., 2010), both of which apply a computational shortcut in the mixed model. Rather than estimating the genetic and residual components repeatedly as each SNP is added to the base model, these methods estimate the variance components only once, using the base model before any SNPs are tested. These estimated variance components are used when calculating all SNP effects (Kang et al., 2010; Zhang et al., 2010). Genome-wide rapid association using mixed model and regression, a residual-based method, also estimates model parameters once before testing SNPs but uses a GLM rather than MLM for SNP testing (Aulchenko, De Koning, & Haley, 2007). However, these approximate solutions may differ from the exact solutions to the MLM, especially in the presence of strong population structure or SNPs with large effect sizes (Zhou & Stephens, 2012).

Methods to improve the speed of exactly solving MLM equations have also been developed. These methods include factored spectrally transformed linear mixed models (FaST-LMM) (Lippert et al., 2011) and genome-wide efficient mixed model analysis (GEMMA) (Zhou & Stephens, 2012). Both methods improve efficiency by rewriting the likelihood function of the MLM in a form that is easier to evaluate. Compared with the approximate methods discussed above, these exact methods do not require the assumption that variance parameters be the same across all SNPs, potentially increasing power (Lippert et al., 2011; Zhou & Stephens, 2012). The two exact methods differ in that FaST-LMM uses only a subset of SNPs to estimate kinship (Lippert et al., 2011), while GEMMA uses all markers and therefore produces results identical to EMMA with increased speed (Zhou & Stephens, 2012).

2.3 | Improving power

Correcting for population structure in the MLM can also increase false negatives, particularly when the true biological signal is correlated with population structure. This increase in false negatives also represents a decline in the statistical power of the GWAS, which is defined as the probability that an association between a trait and a given marker will be detected given that the association truly exists (Klasen et al., 2016). Therefore, methods have been developed to improve the power of GWAS, often—though not always—while increasing computational speed.

The compressed MLM (CMLM) and enriched CMLM (ECMLM) methods improve power by using a lower-rank kinship matrix. Both approaches use a clustering algorithm to divide individuals into groups based on similar genotypes. The number of groups used is optimized for each population studied. A summary of kinship within and between groups is then used as a reduced kinship matrix when solving the MLM (Li et al., 2014; Zhang et al., 2010). The CMLM always uses unweighted pair-group method with arithmetic mean clustering and calculates kinship between groups as the mean of all individual pair-wise kinship values between those groups (Zhang et al., 2010). In contrast, ECMLM adds two more parameters to be optimized: the algorithm used to cluster the individuals into groups (chosen from eight hierarchical clustering algorithms) as well as the method used to calculate kinship between groups (mean, maximum, or median) (Li et al., 2014). These additional parameters are optimized by using P3D to maximize model fit before adding marker effects. In effect, MLM and GLM are both extreme types of CMLM run without this optimization step; each individual comprises its own kinship group in MLM, while all individuals are compressed into a single group in GLM. The CMLM and ECMLM methods improve both computational speed and statistical power compared with typical MLM through improved model fit and reduced kinship matrix rank (Li et al., 2014; Zhang et al., 2010). Because of its additional parameters, ECMLM provides the greater increase in power but is somewhat slower than CMLM (Li et al., 2014).

Other methods calculate the kinship matrix more rapidly by using a reduced number of SNPs. The FaST-LMM method

uses this approach simply to improve computational efficiency as described above (Lippert et al., 2011), but careful selection of the SNPs used can also improve power, as implemented in FaST-LMM-Select (Listgarten et al., 2012) and settlement of MLM under progressively exclusive relationship (SUPER) (Wang, Tian, Pan, Buckler, & Zhang, 2014). While FaST-LMM uses SNPs spaced equidistantly throughout the genome (Lippert et al., 2011), the newer methods select a subset of SNPs that are associated with the trait of interest, so that calculated kinship matrices are specific to each trait. These SNPs are expected to be the most informative because this association may be due to confounding by kinship. In both methods, the first step is to perform simple linear regression and then sort SNPs based on the significance of their association with the trait of interest (Listgarten et al., 2012; Q. Wang et al., 2014). In FaST-LMM-Select, the next step is to construct genetic similarity matrices with ever-increasing numbers of these SNPs, beginning with those SNPs with the lowest p values under linear regression. The matrix that minimizes the genomic control factor, a measure used to control inflation or deflation of the test statistic, is used as a reduced-rank kinship matrix in the MLM (Listgarten et al., 2012). In SUPER, after sorting SNPs by association with the trait of interest, the genome is divided into bins. Within each bin, the SNP with the lowest p value is designated the pseudo quantitative trait nucleotide (QTN). Maximum likelihood is used to optimize the size and number of bins. Finally, these QTNs are used to build a reduced kinship matrix (Wang et al., 2014). Both methods were designed to use the FaST-LMM algorithm to solve the MLM, though SUPER's developers suggest using P3D and EMMA expedited instead to improve computational efficiency. In addition, while a given SNP is being tested in the MLM, these methods will exclude this SNP and those in LD with it from the kinship calculation to avoid confounding. Overall, SUPER is somewhat more powerful than FaST-LMM-Select, particularly for traits with higher heritability, but has lower computational efficiency. Both methods could also potentially be combined with CMLM or ECMLM (Listgarten et al., 2012; Wang et al., 2014).

Multi-locus GWAS methods improve power over single-locus methods by incorporating multiple markers in the model simultaneously as covariates. This approach was first implemented in the multi-locus mixed model (MLMM) (Segura et al., 2012). The MLMM is an iterative approach; in each step, the genetic and error variance components are estimated then used to calculate p values for the association of each SNP with the trait of interest. The EMMA method is used to calculate kinship. The most significant SNP found is then added to the model as a fixed cofactor, and the process is repeated. This continues until a user-set threshold or until the genetic variance unaccounted for by covariate SNPs approaches zero. Then, backward stepwise regression is carried out as the least significant cofactor SNP is removed at each iteration. Finally, the optimal number of iterations is determined using extended Bayesian information criterion or multiple Bonferroni criteria, and the SNP effect sizes and p values from that step provide the final results (Segura et al., 2012).

Other multi-locus methods that build upon MLMM include fixed and random model circulating probability unification (FarmCPU) (Liu, Huang, Fan, Buckler, & Zhang, 2016) and Bayesian information and LD iteratively nested keyway (BLINK) (Huang, Liu, Zhou, Summers, & Zhang, 2018). The FarmCPU is a multi-locus method that uses the reduced-rank kinship matrix of SUPER to improve power and efficiency. This method iterates between the fixed-effect model based on MLMM and the random-effect model of SUPER, using restricted maximum likelihood (REML) as the optimization criterion (Liu et al., 2016). It is also implemented more efficiently in FarmCPUpp using C language (Kusmec & Schnable, 2018). The FarmCPU method was also modified by its creators to produce the BLINK method, which enhances power by relaxing the requirement of SUPER that QTNs be evenly distributed in bins throughout the genome, recognizing that true QTNs are often clustered within the genome. This modification also improves speed, as optimization of bin size and number is no longer required. In addition, BLINK increases speed by replacing the computationally expensive random-effect model and associated REML optimization with a fixed-effect model using Bayesian information criterion optimization (Huang et al., 2018). In general, multi-locus approaches are particularly powerful for complex traits controlled by several large-effect loci (Segura et al., 2012), especially when these loci are closely linked (Li, Li, Fridman, Tesso, & Yu, 2015).

The genome-wide complex trait analysis program is a tool developed for human GWAS that performs several functions. This program can estimate values including SNP-based kinship and population structure, variance explained by SNPs, and LD structure, as well as enabling GWAS simulations and data management. The genome-wide complex trait analysis estimate of variance explained by all genotyped SNPs for a given trait, which is typically less than the estimated heritability of that trait, provides an upper bound for the variance expected to be captured in GWAS for a given SNP set and trait (Yang, Lee, Goddard, & Visscher, 2011).

2.4 | Bayesian methods

Although this review focuses on frequentist, linear model-based approaches, GWAS can also be conducted using Bayesian methods originally developed for genomic prediction. While other GWAS methods discussed in this review test only one or a few markers at a time for association with a trait, Bayesian methods leverage prior information about marker effects and phenotypes to estimate all marker effects

simultaneously (Fernando & Garrick, 2013). For example, this prior information may include whether a given SNP is near a gene with known relevant function, the marker's minor allele frequency, or whether the trait is thought to be controlled by additive or nonadditive genetic effects. Bayesian approaches allow this prior knowledge to be incorporated in the GWAS analyses themselves through such choices as the prior probability of association with the trait at each SNP, the genetic model (e.g., additive or dominant), and the expected distributions of SNP effects, although the need to estimate these priors explicitly does add additional modeling challenges and complexity to Bayesian GWAS. This type of information may also be incorporated into frequentist approaches but typically this is done only when evaluating significant associations for follow-up after GWAS have been conducted (Stephens & Balding, 2009).

Markov-chain Monte Carlo sampling is used to obtain results in Bayesian methods (Fernando & Garrick, 2013); these results can be presented as a Bayes factor for each marker, which is the ratio between the probability of the data under the alternative hypothesis of a marker–trait association and its probability under the null hypothesis of no association. In turn, the Bayes factor can be used to calculate the posterior probability of association, which is the probability that a marker is truly associated with the phenotype of interest given the specified model assumptions. This value already takes into account such factors as how many markers were tested and the power of the analysis, enabling the control of the proportion of false positives among all positive results in the analysis, an approach similar to FDR, without the need for an additional multiple-testing correction step (Stephens & Balding, 2009). Rather than searching for individual SNPs that have a significant association with the phenotype as in frequentist GWAS, Bayesian GWAS methods typically aim to detect genomic windows that explain more than a specified proportion of the total genetic variance (Fernando & Garrick, 2013).

While a full description of available Bayesian methods is beyond the scope of this review, some of the common Bayesian methods include Bayesian RR-BLUP (Meuwissen, Hayes, & Goddard, 2001; Whittaker, Thompson, & Denham, 2000), Bayesian LASSO (de los Campos et al., 2009), BayesA (Meuwissen et al., 2001), BayesB (Meuwissen et al., 2001), BayesC (Kizilkaya, Fernando, & Garrick, 2010), BayesCπ (Habier, Fernando, Kizilkaya, & Garrick, 2011), and many others (Gianola, 2013). Because these methods differ in the prior distribution assumed for the marker effects, they differ in relative accuracy and power depending on how well their assumed distributions reflect the genetic architecture of the trait in question, such as its heritability and the number of causal markers (Wang et al., 2018), and the genetic structure of the study population. For example, BayesA assumes that all SNPs have nonzero effects, while BayesB assumes that at least some markers have no effect; therefore, BayesA is expected to outperform BayesB when the trait of interest is controlled by a very large number of genes, while BayesB is expected to outperform BayesA when the trait is controlled by a few large-effect loci (Fernando & Garrick, 2013; Habier et al., 2011).

Comparison studies using both simulated (Miao, Yang, & Schnable, 2018) and empirical data (Yang et al., 2018) suggest that Bayesian and frequentist mixed model approaches can complement one another when each is used in the appropriate context. For example, Miao et al. (2018) found that FarmCPU outperformed BayesCπ when analyzing moderately complex traits, while the opposite was true in the case of highly complex traits. The authors therefore recommended that researchers first estimate the number of causal variants controlling a trait of interest before model fitting in order to choose the optimal GWAS model for each trait. Where computational resources are a limiting factor, mixed model approaches may be preferred, as they are typically less computationally demanding. However, the constant development of new implementations to improve the computational efficiency of both Bayesian and mixed-model GWAS may alleviate this constraint (Miao et al., 2018).

2.5 | Ongoing method and software development

The methods described above have been included because they provide major improvements in speed and power compared with naïve GWAS models and have also been incorporated into common software packages such as TASSEL (Bradbury et al., 2007), GAPIT (Lipka et al., 2012; Tang et al., 2016), and GEMMA (Zhou & Stephens, 2012). Of course, an exhaustive list of methods used to improve power and efficiency of GWAS is impossible to compile as development of new methods continues. Some of these methods integrate frequentist mixed models with Bayesian concepts. For example, the BOLT-LMM method incorporates both Bayesian priors and a mixed-model framework to achieve improvements (Loh et al., 2015). Several multi-locus methods have been developed that unite the MLM framework with an expectation–maximization empirical Bayes approach. These methods include multi-locus random-SNP-effect MLM (S.B. Wang et al., 2016) and fast multi-locus random-SNP-effect EMMA (Wen et al., 2018). Judging by the many manuscripts describing new methods to conduct GWAS or analyze their results on the biology preprint server bioRxiv at the time of this writing, more methods will certainly be added to this list in the next few years.

In addition, the datasets used for GWAS are constantly increasing in size as high-throughput methods decrease the cost of obtaining both genotypic and phenotypic data. Using more phenotypic and genotypic data in GWAS improves power and resolution but may also make previous methods

computationally intractable. Therefore, any new methods developed must be not only statistically robust but also computationally efficient to address this challenge. For methods already in existence, more efficient implementations must be developed, as has occurred in the case of GEMMA and FarmCPUpp, if they are to keep pace with the ever-increasing availability of data (Kusmec & Schnable, 2018; Zhou & Stephens, 2012).

3 | PERSPECTIVES

3.1 | Challenges and opportunities in further development of GWAS

The challenge posed by loci with low minor allele frequency was known when the GWAS approach was initially proposed. (Risch & Merikangas, 1996). Some researchers remove variants with minor allele frequency below 5% before performing GWAS (e.g., Chen et al., 2016; Kremling, Diepenbrock, Gore, Buckler, & Bandillo, 2019; X. Wang et al., 2016). Their argument is that because statistical power is very low for these rare alleles, preventing identification unless their effect size is extremely large, the large number of these variants only exacerbates the multiple-testing issue. However, the unequal sample size of two alleles of these variants is already considered in the test statistics the same way as other variants and there is no a priori reason why a rare allele should not be biologically important. In fact, because of purifying selection, many deleterious alleles will be present at low frequencies (Sukumaran & Yu, 2014; Visscher et al., 2012; Xiao et al., 2017; Zhu, Li, & Yu, 2011). Many new statistical models have been designed to test the rare variants, often by aggregating nearby rare variants and testing their combined effects (reviewed in the context of human disease genetics in Lee, Abecasis, Boehnke, and Lin [2014]). Many tests designed for rare alleles should be implemented in software packages. Unless including many variants with low minor allele frequency inflates the genome-wide significance threshold as a result of multiple testing, we recommend the testing of variants with low minor allele frequency.

Synthetic associations are misleading associations that occur when GWAS identifies noncausal SNPs as more significant than truly causal variants (Dickson, Wang, Krantz, Hakonarson, & Goldstein, 2010). The most significant peak may actually be located in a different LD block from the true gene, making the gene very difficult to identify. This may happen in the case of allelic heterogeneity, when multiple independent alleles of a given gene are present in a population. If each allele affects the phenotype similarly, none of the alleles responsible will be perfectly correlated with the trait of interest and so their tagging SNPs may not be detected by GWAS. However, there may be SNPs in a different location that are associated with the presence or absence of all alleles responsible. These SNPs may then be detected as synthetic associations. For example, in the case of the *Hd1* gene that controls days to heading in rice, allelic heterogeneity prevented SNPs in the true gene from being significant. However, the adjacent linkage block included SNPs that were well correlated with functional vs. nonfunctional versions of the gene and therefore could be detected (Yano et al., 2016). If the true causal alleles are rare and therefore already difficult to detect, this problem can be exacerbated (Lin et al., 2014; Lin et al., 2012). Rare alleles can also produce synthetic associations even in the absence of allelic heterogeneity. For example, sickle cell anemia is controlled by a single rare allele, but Dickson et al. (2010) showed that common variants in other locations in the genome are significantly associated with this allele by chance. Genome-wide associations studies may then detect these common SNPs as synthetic associations (Dickson et al., 2010). Genome-wide association study approaches based on genes or regions rather than SNPs have been helpful in addressing this problem. However, more work remains to be done in developing these methods, particularly because several of the methods developed have yet to be implemented in freely available software (Yano et al., 2016; Zhu et al., 2011).

Establishing an appropriate significance threshold for GWAS has been a research topic since the beginning (Lander & Kruglyak, 1995; Lander & Schork, 1994). Although Bonferroni correction using the total number of markers tested is known to be overly stringent because it assumes that each marker is independent, which is not the case in GWAS, it has been used in cases where significant GWAS peaks can still be declared with such a threshold (e.g., Li et al., 2015). False discovery rate (Benjamini & Hochberg, 1995), an approach proposed under an expression QTL mapping context (Storey & Tibshirani, 2003), was also used in GWAS (e.g., Owens et al., 2014). However, FDR assumes that test statistics are independent (Benjamini & Hochberg, 1995) and so is not appropriate for GWAS because SNPs in LD within a genomic region yield similar test statistics. Permutation tests are considered the gold-standard method to establish the significance threshold (Gao, Becker, Becker, Starmer, & Province, 2010; Joo, Hormozdiari, Han, & Eskin, 2016) because this method directly samples the test statistic's distribution under the null hypothesis of no association between the markers and the trait of interest. This is accomplished by shuffling the phenotypes while keeping the genotypic data constant and calculating the test statistic for each marker. However, many such permutations are required, often making this method so computationally intensive as to be impractical (Gao et al., 2010; Joo et al., 2016). In the MLM framework specifically, researchers must also be sure to only permute the permutable part in order to avoid disrupting the covariance structure stemming from genetic relatedness (Joo et al., 2016).

Because of these challenges, other methods of setting significance thresholds for GWAS have been developed. In one

example, SimpleM addresses the dependency among markers by calculating the number of effective markers (M_{eff}) and then using $0.05/M_{eff}$ as the genome-wide significant threshold (Gao et al., 2010; Gao, Starmer, & Martin, 2008). The number of effective markers is obtained as the number of principal components that cumulatively capture a high percentage (e.g., 99.5%) of variance in the pairwise correlation matrix for all SNPs, which is derived from the composite LD among SNPs. Another method, a sliding-window approach for locally intercorrelated markers with asymptotic distribution errors corrected (SLIDE), is designed to account for the correlation among SNPs within a sliding window and corrects for the departure of the true null distribution of the statistic from the asymptotic distribution (Han, Kang, & Eskin, 2009). The SLIDE method was shown to have a near identical false positive control to permutation but to be more computationally efficient. In a comparison study with empirical data, Bonferroni correction using the number of LD blocks was found to be inadequate, but SLIDE and SimpleM were recommended (Johnson et al., 2010). More recently, a parametric bootstrapping resampling method, called 'multiple testing in transformed space,' was proposed for GWAS with the linear mixed model (Joo et al., 2016); this method performs a transformation of genotype data to account for genetic relatedness and heritability under linear mixed models and is shown to be computationally efficient by directly sampling statistics instead of sampling phenotypes as in bootstrapping. For studies with local dependency in test statistics, defining discovery at the genomic region level and grouping SNPs within a genomic region together as a single discovery were proposed (Benjamini & Heller, 2007; Sabatti, Service, & Freimer, 2003; Siegmund, Zhang, & Yakir, 2011), and this approach was recently applied to GWAS (Brzyski et al., 2017). Studies to examine these multiple testing correction methods with data in crops need to be carried out.

In addition to conducting GWAS with individual SNPs, genome scans with haplotypes can also be conducted. How to construct haplotypes has been a standing question, and different haplotype construction methods have been proposed (Cardon & Abecasis, 2003). Haplotypes can be built through a sliding-window approach (Guo, Li, Bonham, Wang, & Deng, 2009), based on LD among adjacent SNPs (Gabriel et al., 2002), by selecting informative SNPs (Laramie, Wilk, DeStefano, & Myers, 2007), or inferred by considering different subgroups (Pook et al., 2019). While it is commonly agreed that multiple alleles exist for genetic loci underlying complex traits, it is still debated whether testing individual SNPs or haplotypes better approximates testing of functional alleles. As sequencing technologies continue to improve (Hickey et al., 2019), haplotype-based GWAS in plants is expected to be conducted more often after the sequencing depth is increased to a level where haplotype information is not used extensively in the imputation of missing data.

Ongoing investigation stems from the fact that different GWAS methods often yield similar but nonidentical results. In some cases, significant SNPs detected by one method and experimentally validated as biologically relevant are not detected at all by other methods (Klasen et al., 2016; Li et al., 2015; Yang et al., 2018). These differences in results are expected to occur as a result of differences in the details of the statistical methods used. In many cases, the known strengths of various GWAS methods in traits with disparate genetic architectures and populations with differing structures can explain these differences. For example, MLM, MLMM, and FarmCPU all identify *Dw1* and *Dw2* as top loci affecting sorghum height. However, the dwarfing gene *Dw3* in sorghum was detected by the multi-locus methods MLMM and FarmCPU but not by MLM GWAS because it was in tight repulsion linkage with *qHT7.1*, another locus associated with plant height (Figure 2). Because multi-locus approaches consider multiple SNPs simultaneously as cofactors in the model, they will perform better than single-locus approaches in detecting repulsion-linked loci (Li et al., 2015).

In general, it is probably not possible to identify a single best GWAS method for all situations given the biological complexity inherent among GWAS samples, but each GWAS method does provide a tool to uncover associations that may be missed by other methods depending on the unique genetic architecture and population structure in the study in question (Figure 3). We recommend that researchers conduct both a genome scan of individual SNPs with the MLM methods and genome scans with other multi-locus methods. For additional genomic regions identified by the multi-locus methods, researchers need to evaluate whether the genome-wide marker coverage was adequate so that the multi-locus method's attempt to dissect the polygenic effect of population structure and kinship and attribute it to covariate markers is justified. Developing criteria for selecting the optimal method or methods in any given experiment is therefore an important topic for further investigation.

3.2 | Beyond GWAS

While GWAS are powerful, they are only one step in the process of understanding the underlying genetics of a trait. Significant associations may be prioritized for further investigation based on criteria including p value, replication of significance across multiple locations and related traits, and degree of confounding with population structure (Spindel et al., 2018). After promising SNPs have been prioritized, the genes underlying those peaks remain to be identified. Significant SNPs may be near to and in strong LD with true causal variants, but this is not always the case. Even when a significant SNP is near a true causal gene, other genes may be within the same LD block, making it difficult to determine

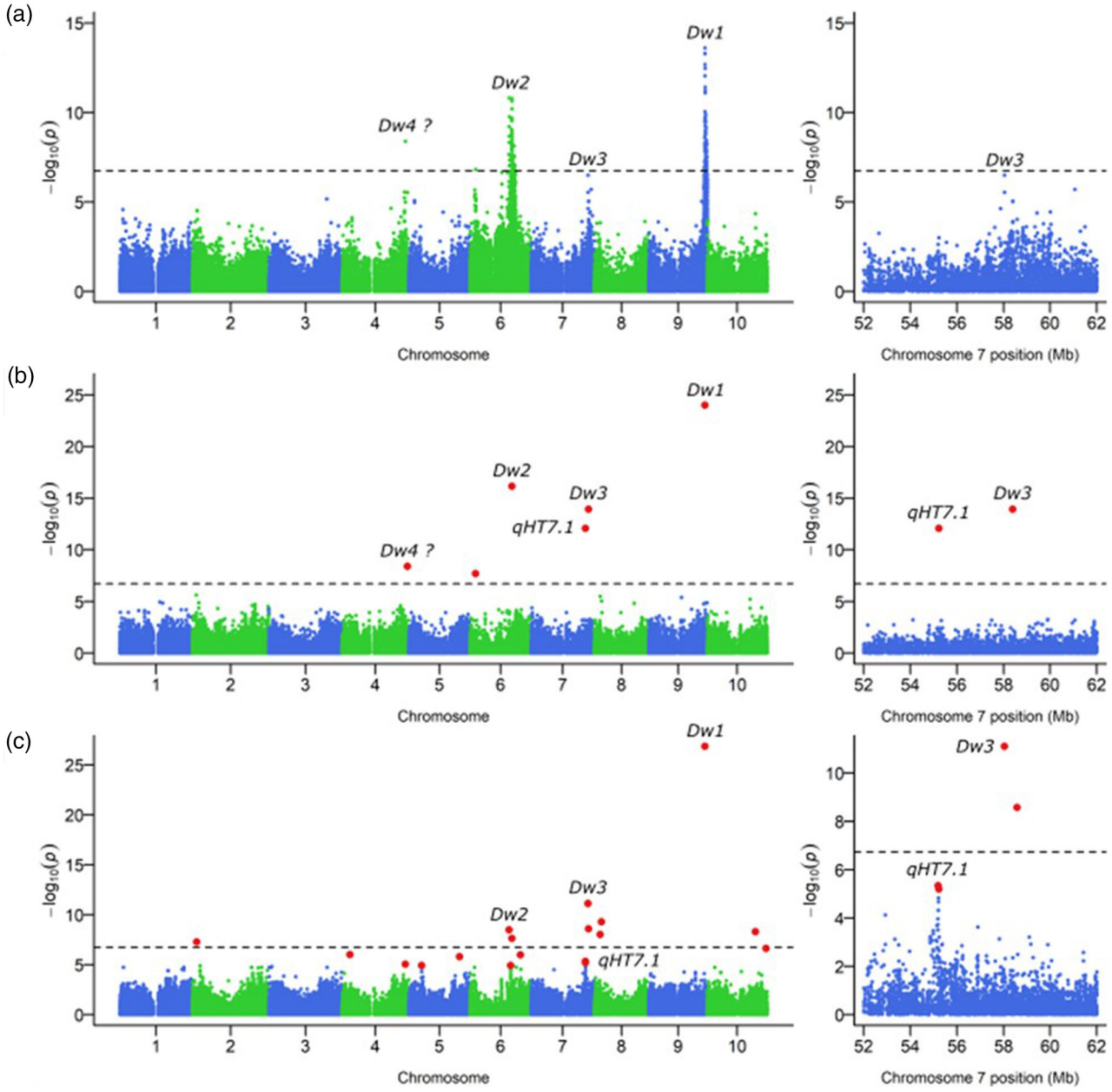

FIGURE 2 Genome-wide association studies of plant height in sorghum. (a) Mixed linear model (MLM), (b) multi-locus mixed-model (MLMM), and (c) Fixed and random model circulating probability unification (FarmCPU). All three methods detected the same strong signals at *Dw1* and *Dw2*. (Left) GWAS plots across all chromosomes. (Right) Regional plots on chromosome 7 near *Dw3*. (a) Because *Dw3* and *qHT7.1* are in tight repulsion linkage, MLM could not detect these loci. The capacity of (b) MLMM and (c) FarmCPU to consider multiple SNPs as cofactors revealed that *Dw3* and *qHT7.1* both affect plant height in sorghum. The horizontal line in each plot represents the significance threshold used in the original publication. The red points in (b) are the SNPs identified as covariates in the optimal model by MLMM; the red points in (c) are the SNPs identified as covariates by FarmCPU

which gene is responsible for the signal (Yano et al., 2016). One of the most basic ways to prioritize genes for further validation is to manually examine nearby genes for relevant functions as predicted by homology, gene networks, or other methods. Gene network tools including RiceNet (Lee et al., 2015a), AraNet (Lee et al., 2015b), and PlaNet (Mutwil et al., 2011) expedite this process in many plant species, especially when a priori candidate genes for the trait are known. Other tools such as RafSee and RAP combine these network results with evolutionary, epigenetic, or other information to rank candidate genes (Zhai et al., 2016). Another method is to prioritize SNPs based on the predicted effect of the polymorphism, that is, whether the mutation is synonymous or nonsynonymous, or in a coding or noncoding region, for example. (Yano et al., 2016). These and other techniques can be combined, as in composite resequencing-based

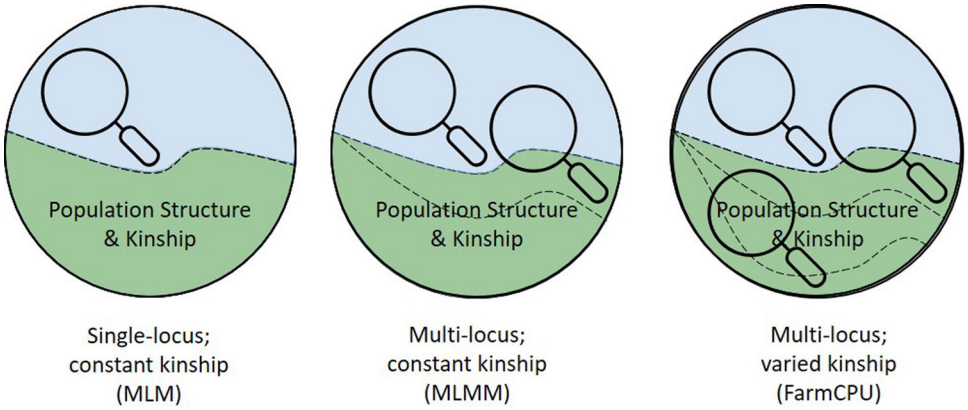

FIGURE 3 Different genome-wide association study methods ask different questions. For the single-locus mixed linear model (MLM) method, acknowledging the embedded unequal genetic relationship among the assembled samples, we ask the question: "After controlling for the relatedness among individuals, where are the association signals?" For the multi-locus methods, multi-locus mixed-model (MLMM) and fixed and random model circulating probability unification (FarmCPU), attempting to identify variants to explain the observed phenotypic variation, we ask the question, "Within the current data, where are the association signals?" The multi-locus methods should only be used as a stand-alone analysis when there is an adequate marker coverage and with the assumption that a genomic region harboring the causal variant can be generally identified consistently even if the chance of tagging causal variants in other genomic regions may vary

GWAS, which prioritizes candidate genes by integrating conventional GWAS with rare allele testing as well as functional prediction, prior biological knowledge, and gene networks (Zhu et al., 2011). Because these methods require well-annotated genomes and fine-scale genome sequencing, their utility in facilitating the move from GWAS to biology will increase as sequencing costs decrease and annotations improve.

As data collection accelerates in all levels of biology, GWAS approaches will expand to omic-wide association studies (OWAS) (Xiao et al., 2017). In a typical GWAS, genotype explains a single terminal phenotype. However, GWAS can also predict other omic endophenotypes, which include epigenomic, transcriptomic, proteomic, and metabolomic data. For example, expression QTL studies explain transcript number variation (Brem, Yvert, Clinton, & Kruglyak, 2002), while other analyses predict epigenetic markers using genotype data (Schmitz et al., 2013). Endophenotypes may also act as the predictors rather than as the predicted trait. A veritable alphabet of association studies have been developed in this area; EWAS, TWAS, and MWAS explain traits using epigenomic (Teschendorff et al., 2009), transcriptomic (Gusev et al., 2016), and metabolomic data (Holmes et al., 2008), respectively. These can even be combined, such as by integrating TWAS and GWAS into a single analysis using Fisher's combined test (Kremling, Diepenbrock, Gore, Buckler, & Bandillo, 2019). In addition, GWAS can expand to analyze multiple traits simultaneously. Multi-trait GWAS models can be used on either individual-level phenotype and genotype data or as a metaanalysis method combining summary statistics from previous GWAS of single traits (methods reviewed and compared in Porter & O'Reilly, 2017). By applying GWAS to multiple levels and multiple traits simultaneously in a multi-OWAS approach, we can expect to achieve a better understanding of the intricate biological systems that produce terminal phenotypes.

It is worthwhile to point out that GWAS is closely related to genomic prediction (GP) and genomic selection (GS). Both GWAS and GP use large genomic and phenotypic data sets (in GP, called the training set) to estimate marker effects, but rather than simply identifying the most significant loci as in GWAS, GP uses these marker effects to predict the phenotypes of unobserved individuals (the testing set) based on their genotype data (Bernardo & Yu, 2007; Heffner, Sorrells, & Jannink, 2009; Meuwissen et al., 2001). Genomic selection applies the results of GP, using the predicted phenotypes to aid the selection decisions in a breeding program (Crossa et al., 2017; Xu et al., 2020). Because of these data and procedural overlaps, methods developed for GWAS may be applied to GS and vice versa (Fernando & Garrick, 2013; Tang et al., 2016), and many common software programs can use these methods to perform both GWAS and GS (Bradbury et al., 2007; Tang et al., 2016; Yang et al., 2011). Results of previous GWAS, whether significant markers or validated genes, can be incorporated into GS models, an approach that has been particularly useful in complex, low-heritability traits in livestock but has yet to be extensively applied in crops (Xu et al., 2020). Genome-wide association studies can also be incorporated directly into GS through GS + de novo GWAS, in which GWAS is conducted on the training set to identify markers that are then used as fixed effects in genomic prediction of the testing set (Spindel et al., 2016).

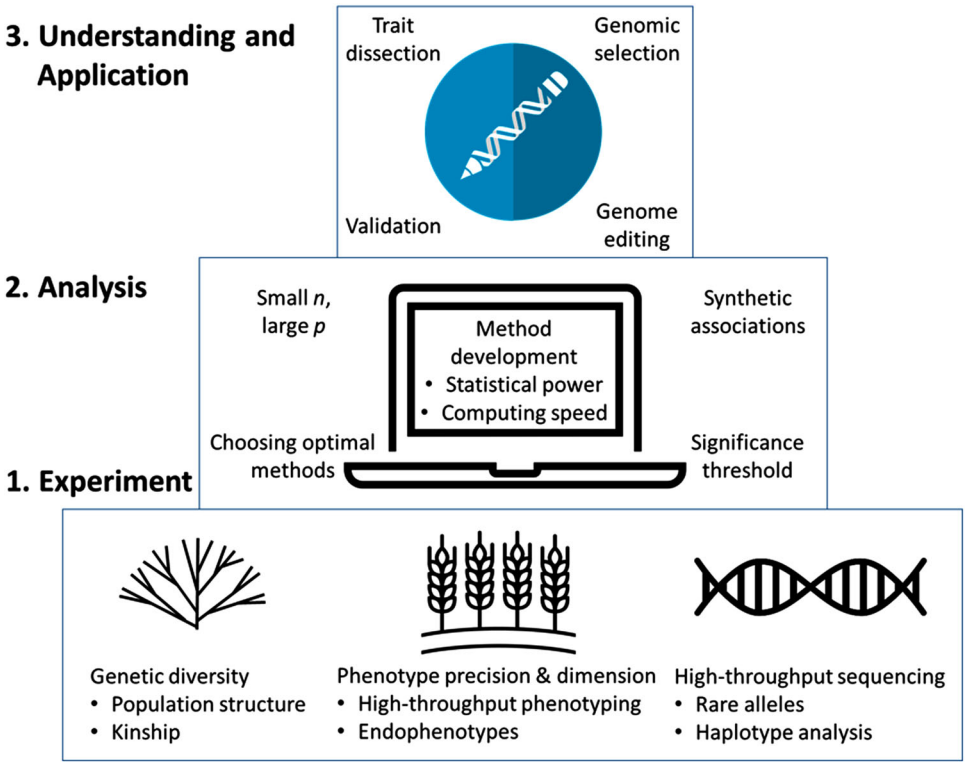

FIGURE 4 Challenges and opportunities in genome-wide association studies (GWAS). Challenges arise in each step of GWAS through the complex interplay of both biology and statistics. Surmounting these challenges will provide new opportunities for understanding and application

3.3 | Methodology development for GWAS in plants

There is no doubt that complex trait dissection requires long-term efforts to address long-standing and newly emerging issues as different technologies advance and understanding improves (Figure 4) (Boyle, Li, & Pritchard, 2017; Mackay et al., 2009; Zhu et al., 2008). Genome-wide association studies present a unique opportunity to study the genotype–phenotype relationship across diverse genetic backgrounds. Concerted efforts are needed in association panel assembly, experimental design, genotyping and sequencing, phenotyping, comprehensive analysis, and postanalysis interpretation and validation. Many biological and statistical perspectives deserve careful consideration. While researchers may think newer statistical methods would solve some long-standing issues, the biological complexity should be regarded as the root cause of these challenges. While many research groups may be devoting efforts to understanding different complex traits through GWAS and other approaches, very few groups have been working on methodology development. Funding support to methodology development and software implementation has been scarce, even though there is a consensus that developing new methods and implementing them in user-friendly software packages have a broad impact.

Genome-wide association studies in plants have several unique aspects different from studies in human genetics. Association mapping panels assembled in plants have complex genetic relatedness. The sizes of these panels with diverse inbreds are typically on the order of hundreds because of the challenges associated with extensively phenotyping a panel with a large number across multiple environments. Rather than a single species as in human genetics, our scope is many different plant species. While the sequencing cost has dramatically reduced, obtaining genomic data through resequencing with a high genome coverage or de novo assembly in crops with complex genomes for hundreds of individuals is still beyond the reach of individual research groups. The relatively constant sample size (n) but increased marker (i.e., SNPs, structural variations) number (p) exacerbate many issues including low minor frequency and rare alleles, synthetic association, multiple testing, haplotype construction, and GWAS model comparison.

As new GWAS methods are created, they should be validated in order to assess their performance. In many cases, the newly developed method and one or more older methods are used to analyze simulated data so that authors can accurately assess their statistical power (Li et al., 2014; Segura et al., 2012; Yu et al., 2006). In addition to statistical validation, biological validation of candidates identified by GWAS is also important. For well-studied traits, peaks detected by

a new GWAS method can be compared with known associated genes or genomic regions for validation (Yang et al., 2018). New causal genes underlying GWAS peaks have been identified through transgenic and other approaches, including RNAi, mutant rescue, and CRISPR/Cas9-mediated gene silencing, knockout, and overexpression (Li et al., 2017; Si et al., 2016; Sun et al., 2018; Yano et al., 2016). In some exemplary cases, a single paper may perform both biological and statistical validation of GWAS methods (Liu et al., 2016; Miao et al., 2018; Wen et al., 2018). Especially in the case of newly identified causal genes, the combination of computational, laboratory, and other resources required means that validation of new GWAS methods will remain challenging.

4 | CONCLUSION

Genome-wide association studies have successfully identified thousands of loci associated with agronomic and other traits in crop species, and several methods have been developed to improve power and computational speed. As the development of GWAS in crops continues, it may emulate the recent progress of GWAS in human diseases. Human disease GWAS have been coupled with tools including in silico models, tissue-specific resources, and wet-lab experiments to understand the biological functions of significant loci and to determine the causal and protective alleles. This deeper understanding has led to tangible results including new drug targets and therapeutic optimization for individual patients depending on genotype (reviewed in Visscher et al., 2017). Today, especially in the major species of crops, similar validation tools are being developed or are already available. It is worthwhile to point out the synergistic relationship between GWAS and genome editing. While GWAS is a major tool to identify genes underlying complex traits, providing targets for genome editing to generate engineered alleles, improved genome editing enables the validation of gene function under different genetic backgrounds, which can inform the method research of GWAS. Ultimately, the biological understanding gained from complementing GWAS with many genomic, phenomic, biotechnological, and data analytical tools in crop species may in turn be used to produce genetically modified plants containing specific alleles known to influence desirable traits including drought resistance, increased yield, and improved nutritional quality.

To help readers grasp the essence of this GWAS review, we summarize the main points as follows:

- Genome-wide association studies techniques were developed beginning in the 1990s, and the first GWAS were published in the early 2000s.
- Genome-wide association studies dissect complex traits by associating genotypic variants with phenotypic variation in a large panel. Although unrelated individuals are preferred, this is often not the case when we assemble existing samples to form the study panel.
- Because of the existence of population structure and kinship among samples within the panel, naïve GWAS analysis with simple linear regression results in many false positives.
- The unified mixed model was developed for GWAS to control for both population structure and kinship. This framework has been widely adopted.
- Genome-wide association studies methods have been developed to improve computational speed by improving the efficiency of solving the MLM equations. Other methods improve power by alternative kinship calculation methods and by including multiple markers as covariates; these methods often improve efficiency as well.
- Ongoing challenges include analyzing variants with low minor allele frequency, avoiding synthetic associations, and understanding differences among results generated by various GWAS methods.
- Candidate gene prioritization methods help in moving from GWAS results to biological understanding.
- Continued methodology development in GWAS is needed and funding support for methodology development and software implementation benefits a wide range of research disciplines.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (IOS-1546657, DBI-1661348, and ISO-2029933), the Iowa State University Plant Sciences Institute, and the Iowa State University Raymond F. Baker Center for Plant Breeding, and the Washington Grain Commission (Endowment and Award 126593 and 134574). Laura Tibbs Cortes was supported by the National Science Foundation Graduate Research Fellowship Program (Grant No. 1744592).

AUTHOR CONTRIBUTIONS

Laura Tibbs Cortes: Data curation; Formal analysis; Funding acquisition; Investigation; Resources; Software; Visualization; Writing-original draft; Writing-review & editing. Zhiwu Zhang: Funding acquisition; Methodology; Resources; Software; Writing-review & editing. Jianming Yu: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Visualization; Writing-review & editing

ORCID

Laura Tibbs Cortes https://orcid.org/0000-0003-3188-6820
Zhiwu Zhang https://orcid.org/0000-0002-5784-9684
Jianming Yu https://orcid.org/0000-0001-5326-3099

REFERENCES

Abasht, B., & Lamont, S. J. (2007). Genome-wide association analysis reveals cryptic alleles as an important factor in heterosis for fatness in chicken F_2 population. *Animal Genetics*, *38*, 491–498. https://doi.org/10.1111/j.1365-2052.2007.01642.x

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., … Nordborg, M. (2005). Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genetics*, *1*, e60. https://doi.org/10.1371/journal.pgen.0010060

Aulchenko, Y. S., De Koning, D.-J., & Haley, C. (2007). Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, *177*, 577–585. https://doi.org/10.1534/genetics.107.075614

Beló, A., Zheng, P., Luck, S., Shen, B., Meyer, D. J., Li, B., … Rafalski, A. (2008). Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. *Molecular Genetics and Genomics*, *279*, 1–10. https://doi.org/10.1007/s00438-007-0289-y

Benjamini, Y., & Heller, R. (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association*, *102*, 1272–1281. https://doi.org/10.1198/016214507000000941

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289–300. Retrieved from https://www.jstor.com/stable/2346101

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, *47*, 1082–1090. https://doi.org/10.2135/cropsci2006.11.0690

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, *169*, 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*, 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, *296*, 752–755. https://doi.org/10.1126/science.1069516

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., & Sabatti, C. (2017). Controlling the rate of GWAS false discoveries. *Genetics*, *205*, 61–75. https://doi.org/10.1534/genetics.116.193987

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PL0S Computational Biology*, *8*, e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Cardon, L. R., & Abecasis, G. R. (2003). Using haplotype blocks to map human complex trait loci. *Trends in Genetics*, *19*, 135–140. https://doi.org/10.1016/S0168-9525(03)00022-2

Chen, E., Huang, X., Tian, Z., Wing, R. A., & Han, B. (2019). The genomics of *Oryza* species provides insights into rice domestication and heterosis. *Annual Review of Plant Biology*, *70*, 639–665. https://doi.org/10.1146/annurev-arplant-050718-100320

Chen, W., Wang, W., Peng, M., Gong, L., Gao, Y., Wan, J., … Luo, J. (2016). Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nature Communications*, *7*, 12767. https://doi.org/10.1038/ncomms12767

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., … Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*, 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., … Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, *182*, 375–385. https://doi.org/10.1534/genetics.109.101501

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*, 997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., & Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biology*, *8*, e1000294. https://doi.org/10.1371/journal.pbio.1000294

Ersoz, E. S., Yu, J., & Buckler, E. S. (2007). Applications of linkage disequilibrium and association mapping in crop plants. In R. K. Varshney, & R. Tuberosa (Eds.), *Genomics-assisted crop improvement* (pp. 97–119). Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-1-4020-6295-7_5

Fernando, R. L., & Garrick, D. (2013). Bayesian methods applied to GWAS. In C. Gondro, J. van der Werf, & B. Hayes (Eds.), *Genome-wide association studies and genomic prediction* (Vol. *1019*, pp. 237–274). Totowa, NJ; Humana Press. https://doi.org/10.1007/978-1-62703-447-0_10

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., … Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, *296*, 2225–2229. https://doi.org/10.1126/science.1069424

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., & Province, M. A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic Epidemiology*, *34*, 100–105. https://doi.org/10.1002/gepi.20430

Gao, X., Starmer, J., & Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, *32*, 361–369. https://doi.org/10.1002/gepi.20310

Gianola, D. (2013). Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics*, *194*, 573–596. https://doi.org/10.1534/genetics.113.151753

Guo, Y., Li, J., Bonham, A. J., Wang, Y., & Deng, H. (2009). Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: A comparison of association-mapping strategies. *European Journal of Human Genetics*, *17*, 785–792. https://doi.org/10.1038/ejhg.2008.244

Gupta, P. K., Kulwal, P. L., & Jaiswal, V. (2019). Chapter Two—Association mapping in plants in the post-GWAS genomics era. In D. Kumar (Ed.), *Advances in Genetics* (pp. 75–154). Cambridge, MA: Academic Press–Elsevier Inc. https://doi.org/10.1016/bs.adgen.2018.12.001

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., … Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*, 245–252. https://doi.org/10.1038/ng.3506

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*, 186. https://doi.org/10.1186/1471-2105-12-186

Han, B., Kang, H. M., & Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of

correlated markers. *PLoS Genetics*, *5*, e1000456. https://doi.org/10.1371/journal.pgen.1000456

Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science*, *49*, 1–12. https://doi.org/10.2135/cropsci2008.08.0512

Hickey, L. T., Hafeez, A. N., Robinson, H., Jackson, S. A., Leal-Bertioli, S. C. M., Tester, M., … Wulff, B. B. H. (2019). Breeding crops to feed 10 billion. *Nature Biotechnology*, *37*, 744–754. https://doi.org/10.1038/s41587-019-0152-9

Holmes, E., Loo, R. L., Stamler, J., Bictash, M., Yap, I. K. S., Chan, Q., … Elliott, P. (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature*, *453*, 396–400. https://doi.org/10.1038/nature06882

Huang, M., Liu, X., Zhou, Y., Summers, R. M., & Zhang, Z. (2018). BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. *GigaScience*, *8*, 1–12. https://doi.org/10.1093/gigascience/giy154

Consortium, International Human Genome Sequencing (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921. https://doi.org/10.1038/35057062

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., & O'Brien, S. J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics*, *11*, 724. https://doi.org/10.1186/1471-2164-11-724

Joo, J. W. J., Hormozdiari, F., Han, B., & Eskin, E. (2016). Multiple testing correction in linear mixed models. *Genome Biology*, *17*, 62. https://doi.org/10.1186/s13059-016-0903-6

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., … Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*, 348–354. https://doi.org/10.1038/ng.548

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*, 1709–1723. https://doi.org/10.1534/genetics.107.080101

Kizilkaya, K., Fernando, R. L., & Garrick, D. J. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *Journal of Animal Science*, *88*, 544–551. https://doi.org/10.2527/jas.2009-2064

Klasen, J. R., Barbez, E., Meier, L., Meinshausen, N., Bühlmann, P., Koornneef, M., … Schneeberger, K. (2016). A multi-marker association method for genome-wide association studies without the need for population structure correction. *Nature Communications*, *7*, 13299. https://doi.org/10.1038/ncomms13299

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods*, *9*, 29. https://doi.org/10.1186/1746-4811-9-29

Kremling, K. A. G., Diepenbrock, C. H., Gore, M. A., Buckler, E. S., & Bandillo, N. B. (2019). Transcriptome-wide association supplements genome-wide association in *Zea mays*. *G3: Genes, Genomes, Genetics*, *9*, 3023–3033. https://doi.org/10.1534/g3.119.400549

Kusmec, A., & Schnable, P. S. (2018). FarmCPUpp: Efficient large-scale genomewide association studies. *Plant Direct*, *2*, e00053. https://doi.org/10.1002/pld3.53

Lander, E. S. (1996). The new genomics: Global views of biology. *Science*, *274*, 536–539. https://doi.org/10.1126/science.274.5287.536

Lander, E. S., & Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, *11*, 241–247. https://doi.org/10.1038/ng1195-241

Lander, E. S., & Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, *265*, 2037–2048. https://doi.org/10.1126/science.8091226

Laramie, J. M., Wilk, J. B., DeStefano, A. L., & Myers, R. H. (2007). HaploBuild: An algorithm to construct non-contiguous associated haplotypes in family based genetic studies. *Bioinformatics*, *23*, 2190–2192. https://doi.org/10.1093/bioinformatics/btm316

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *The American Journal of Human Genetics*, *95*, 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009

Lee, T., Oh, T., Yang, S., Shin, J., Hwang, S., Kim, C. Y., … Lee, I. (2015a). RiceNet v2: An improved network prioritization server for rice genes. *Nucleic Acids Research*, *43*, W122–W127. https://doi.org/10.1093/nar/gkv253

Lee, T., Yang, S., Kim, E., Ko, Y., Hwang, S., Shin, J., … Lee, I. (2015b). AraNet v2: An improved database of co-functional gene networks for the study of *Arabidopsis thaliana* and 27 other nonmodel plant species. *Nucleic Acids Research*, *43*, D996–D1002. https://doi.org/10.1093/nar/gku1053

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

Li, W., Zhu, Z., Chern, M., Yin, J., Yang, C., Ran, L., … Chen, X. (2017). A natural allele of a transcription factor in rice confers broad-spectrum blast resistance. *Cell*, *170*, 114–126. https://doi.org/10.1016/j.cell.2017.06.008

Li, X., Li, X., Fridman, E., Tesso, T. T., & Yu, J. (2015). Dissecting repulsion linkage in the dwarfing gene *Dw3* region for sorghum plant height provides insights into heterosis. *Proceedings of the National Academy of Sciences*, *112*, 11823–11828. https://doi.org/10.1073/pnas.1509229112

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., … Huang, S. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics*, *46*, 1220–1226. https://doi.org/10.1038/ng.3117

Lin, Z., Li, X., Shannon, L. M., Yeh, C. T., Wang, M. L., Bai, G., … Yu, J. (2012). Parallel domestication of the *Shattering1* genes in cereals. *Nature Genetics*, *44*, 720–724. https://doi.org/10.1038/ng.2281

Lipka, A. E., Kandianis, C. B., Hudson, M. E., Yu, J., Drnevich, J., Bradbury, P. J., & Gore, M. A. (2015). From association to prediction: Statistical methods for the dissection and selection of complex traits in plants. *Current Opinion in Plant Biology*, *24*, 110–118. https://doi.org/10.1016/j.pbi.2015.02.010

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

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*, 833–835. https://doi.org/10.1038/nmeth.1681

Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., & Heckerman, D. (2012). Improved linear mixed models for

genome-wide association studies. *Nature Methods*, *9*, 525–526. https://doi.org/10.1038/nmeth.2037

Liu, H. J., & Yan, J. (2019). Crop genome-wide association study: A harvest of biological relevance. *The Plant Journal*, *97*, 8–18. https://doi.org/10.1111/tpj.14139

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

Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., … Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*, 284–290. https://doi.org/10.1038/ng.3190

Loiselle, B. A., Sork, V. L., Nason, J., & Graham, C. (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany*, *82*, 1420–1425. https://doi.org/10.2307/2445869

Mackay, T. F. C., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: Challenges and prospects. *Nature Reviews Genetics*, *10*, 565–577. https://doi.org/10.1038/nrg2612

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*, 1819–1829.

Miao, C., Yang, J., & Schnable, J. C. (2018). Optimising the identification of causal variants across varying genetic architectures in crops. *Plant Biotechnology Journal*, *17*, 893–905. https://doi.org/10.1111/pbi.13023

Michael, T. P., & Jackson, S. (2013). The first 50 plant genomes. *The Plant Genome*, *6*, 1–7. https://doi.org/10.3835/plantgenome2013.03.0001in

Miles, C. M., & Wayne, M. (2008). Quantitative trait locus (QTL) analysis. *Nature Education*, *1*, 208. Retrieved from https://www.nature.com/scitable/topicpage/quantitative-trait-locus-qtl-analysis-53904

Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., … Persson, S. (2011). PlaNet: Combined sequence and expression comparisons across plant networks derived from seven species. *The Plant Cell*, *23*, 895–910. https://doi.org/10.1105/tpc.111.083667

Owens, B. F., Lipka, A. E., Magallanes-Lundback, M., Tiede, T., Diepenbrock, C. H., Kandianis, C. B., … Rocheford, T. (2014). A foundation for provitamin A biofortification of maize: Genome-wide association and genomic prediction models of carotenoid levels. *Genetics*, *198*, 1699–1716. https://doi.org/10.1534/genetics.114.169979

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., … Tanaka, T. (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, *32*, 650–654. https://doi.org/10.1038/ng1047

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., & Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, *335*, 721–726. https://doi.org/10.1038/335721a0

Pook, T., Schlather, M., de los Campos, G., Mayer, M., Schoen, C. C., & Simianer, H. (2019). HaploBlocker: Creation of subgroup-specific haplotype blocks and libraries. *Genetics*, *212*, 1045–1061. https://doi.org/10.1534/genetics.119.302283

Porter, H. F., & O'Reilly, P. F. (2017). Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Scientific Reports*, *7*, 38837. https://doi.org/10.1038/srep38837

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909. https://doi.org/10.1038/ng1847

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959. Retrieved from https://www.genetics.org/content/155/2/945

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

Sabatti, C., Service, S., & Freimer, N. (2003). False discovery rate in linkage and association genome screens for complex disorders. *Genetics*, *164*, 829–833. Retrieved from https://www.genetics.org/content/164/2/829

Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., … Ecker, J. R. (2013). Patterns of population epigenomic diversity. *Nature*, *495*, 193–198. https://doi.org/10.1038/nature11968

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

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, *29*, 308–311. https://doi.org/10.1093/nar/29.1.308

Si, L., Chen, J., Huang, X., Gong, H., Luo, J., Hou, Q., … Han, B. (2016). *OsSPL13* controls grain size in cultivated rice. *Nature Genetics*, *48*, 447–457. https://doi.org/10.1038/ng.3518

Siegmund, D. O., Zhang, N. R., & Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika*, *98*, 979–985. https://doi.org/10.1093/biomet/asr057

Speed, D., & Balding, D. J. (2015). Relatedness in the post-genomic era: Is it still useful? *Nature Reviews Genetics*, *16*, 33–44. https://doi.org/10.1038/nrg3821

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J. L., & McCouch, S. (2016). Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity*, *116*, 395–408. https://doi.org/10.1038/hdy.2015.113

Spindel, J. E., Dahlberg, J., Colgan, M., Hollingsworth, J., Sievert, J., Staggenborg, S. H., … Vogel, J. P. (2018). Association mapping by aerial drone reveals 213 genetic associations for *Sorghum bicolor* biomass traits under drought. *BMC Genomics*, *19*, 679. https://doi.org/10.1186/s12864-018-5055-5

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, *10*, 681–690. https://doi.org/10.1038/nrg2615

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*, 9440–9445. https://doi.org/10.1073/pnas.1530509100

Sukumaran, S., & Yu, J. (2014). Association mapping of genetic resources: Achievements and future perspectives. In R. Tuberosa, A. Graner, & E. Frison (Eds.), *Genomics of plant genetic resources* (pp. 207–235). Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-94-007-7572-5_9

Sun, S., Wang, T., Wang, L., Li, X., Jia, Y., Liu, C., … Wang, X. (2018). Natural selection of a *GSK3* determines rice mesocotyl domestication by coordinating strigolactone and brassinosteroid signaling.

Nature Communications, *9*, 2523. https://doi.org/10.1038/s41467-018-04952-9

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., … Zhang, Z. (2016). GAPIT version 2: An enhanced integrated tool for genomic association and prediction. *The Plant Genome*, *9*. https://doi.org/10.3835/plantgenome2015.11.0120

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Gayther, S. A., Apostolidou, S., … Widschwendter, M. (2009). An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS ONE*, *4*, e8274. https://doi.org/10.1371/journal.pone.0008274

The International HapMap Consortium. (2003). The International HapMap Project. *Nature*, *426*, 789–796. https://doi.org/10.1038/nature02168

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., & Buckler, E. S. (2001). *Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genetics*, *28*, 286–289. https://doi.org/10.1038/90135

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423. https://doi.org/10.3168/JDS.2007-0980

Varshney, R. K., Ribaut, J. M., Buckler, E. S., Tuberosa, R., Rafalski, J. A., & Langridge, P. (2012). Can genomics boost productivity of orphan crops? *Nature Biotechnology*, *30*, 1172–1176. https://doi.org/10.1038/nbt.2440

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, *90*, 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, *101*, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wang, J., Zhou, Z., Zhang, Z., Li, H., Liu, D., Zhang, Q., … Zhang, Z. (2018). Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity*, *121*, 648–662. https://doi.org/10.1038/s41437-018-0075-0

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

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., … Zhang, Y. M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific Reports*, *6*, 19444. https://doi.org/10.1038/srep19444

Wang, X., Wang, H., Liu, S., Ferjani, A., Li, J., Yan, J., … Qin, F. (2016). Genetic variation in *ZmVPP1* contributes to drought tolerance in maize seedlings. *Nature Genetics*, *48*, 1233–1241. https://doi.org/10.1038/ng.3636

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

Whittaker, J. C., Thompson, R., & Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetics Research*, *75*, 249–252. https://doi.org/10.1017/S0016672399004462

Xiao, Y., Liu, H., Wu, L., Warburton, M., & Yan, J. (2017). Genome-wide association studies in maize: Praise and stargaze. *Molecular Plant*, *10*, 359–374. https://doi.org/10.1016/j.molp.2016.12.008

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., … Zhang, A. (2020). Enhancing genetic gain through genomic selection: From livestock to plants. *Plant Communications*, *1*, 100005. https://doi.org/10.1016/j.xplc.2019.100005

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*, 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yang, J., Yeh, C. T. E., Ramamurthy, R. K., Qi, X., Fernando, R. L., Dekkers, J. C. M., … Schnable, P. S. (2018). Empirical comparisons of different statistical models to identify and validate kernel row number-associated variants from structured multi-parent mapping populations of maize. *G3: Genes, Genomes, Genetics*, *8*, 3567–3575. https://doi.org/10.1534/g3.118.200636

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P., Hu, L., … Matsuoka, M. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics*, *48*, 927–934. https://doi.org/10.1038/ng.3596

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., … Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*, 203–208. https://doi.org/10.1038/ng1702

Zhai, J., Tang, Y., Yuan, H., Wang, L., Shang, H., & Ma, C. (2016). A meta-analysis based method for prioritizing candidate genes involved in a pre-specific function. *Frontiers in Plant Science*, *7*, 1914. https://doi.org/10.3389/fpls.2016.01914

Zhang, Y., Massel, K., Godwin, I. D., & Gao, C. (2018). Applications and potential of genome editing in crop improvement. *Genome Biology*, *19*, 210. https://doi.org/10.1186/s13059-018-1586-y

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., … Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*, 355–360. https://doi.org/10.1038/ng.546

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed model analysis for association studies. *Nature Genetics*, *44*, 821–824. https://doi.org/10.1038/ng.2310

Zhou, Y., Srinivasan, S., Mirnezami, S. V., Kusmec, A., Fu, Q., Attigala, L., … Schnable, P. S. (2019). Semiautomated feature extraction from RGB images for sorghum panicle architecture GWAS. *Plant Physiology*, *179*, 24–37. https://doi.org/10.1104/pp.18.00974

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

Zhu, C., Li, X., & Yu, J. (2011). Integrating rare-variant testing, function prediction, and gene network in composite resequencing-based genome-wide association studies (CR-GWAS). *G3: Genes, Genomes, Genetics*, *1*, 233–243. https://doi.org/10.1534/g3.111.000364

Zhu, C., & Yu, J. (2009). Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics*, *182*, 875–888. https://doi.org/10.1534/genetics.108.098863

How to cite this article: Tibbs Cortes L, Zhang Z, Yu J. Status and prospects of genome-wide association studies in plants. *Plant Genome*. 2021;14:e20077. https://doi.org/10.1002/tpg2.20077