Association mapping in plants in the post-GWAS genomics era

Pushpendra K. Gupta[a], Pawan L. Kulwal[b,*], Vandana Jaiswal[c]
[a]Department of Genetics and Plant Breeding, Ch. Charan Singh University, Meerut, UP, India
[b]State Level Biotechnology Centre, Mahatma Phule Agricultural University, Rahuri, MS, India
[c]School of Life Sciences, Jawaharlal Nehru University, New Delhi, India
*Corresponding author: e-mail address: pawankulwal@gmail.com

Contents

Advances in Genetics
ISSN 0065-2660
https://doi.org/10.1016/bs.adgen.2018.12.001

Abstract

With the availability of DNA-based molecular markers during early 1980s and that of sophisticated statistical tools in late 1980s and later, it became possible to identify genomic regions that control a quantitative trait. The two methods used for this purpose included quantitative trait loci (QTL) interval mapping and genome-wide association mapping/studies (GWAS). Both these methods have their own merits and demerits, so that newer approaches were developed in order to deal with the demerits. We have now entered a post-GWAS era, where either the original data on individual genotypes are being used again keeping in view the results of GWAS or else summary statistics obtained through GWAS is subjected to further analysis. The first half of this review briefly deals with the approaches that were used for GWAS, the GWAS results obtained in some major crops (maize, wheat, rice, sorghum and soybean), their utilization for crop improvement and the improvements made to address the limitations of original GWA studies (computational demand, multiple testing and false discovery, rare marker alleles, etc.). These improvements included the development of multi-locus and multi-trait analysis, joint linkage association mapping, etc. Since originally GWA studies were used for mere identification of marker-trait association for marker-assisted selection, the second half of the review is devoted to activities in post-GWAS era, which include different methods that are being used for identification of causal variants and their prioritization (meta-analysis, pathway-based analysis, methylation QTL), functional characterization of candidate signals, gene- and gene-set based association mapping, GWAS using high dimensional data through machine learning, etc. The last section deals with popular resources available for GWAS in plants in the post-GWAS era and the implications of the results of post-GWAS for crop improvement.

Abbreviations

DArT	diversity arrays technology
EMMA	efficient mixed model association
Fast-LMM	factored spectrally transformed linear mixed model
FDR	false discovery rate
FWER	family-wise error rate
GBAM	gene based association mapping
GBS	genotyping by sequencing
GEMMA	genome-wide efficient mixed model association
GRAMMAR	genome-wide rapid association using mixed model and regression
GS	genomic selection
GWAS	genome wide association studies
GWER	genome-wide error rate
HT	high throughput
HTS	high-throughput sequencing
JLAM	joint linkage association mapping
LD	linkage disequilibrium
LRT	likelihood ratio test
LSR	localization success rate

MAF	minor allele frequency
ML	machine-learning
MLM	mixed linear model
MLMM	multi-locus mixed model
MTMM	multi-trait mixed model
MTAs	marker-trait associations
NAM	nested association mapping
NGS	next-generation sequencing
PCA	principal component analysis
PheWAS	phenome-wide association studies
PVE	phenotypic variation explained
QTL	quantitative trait loci
SKAT	sequence kernel association test
SNP	single nucleotide polymorphism
SSR	simple sequence repeat
TAD	topologically association domain
TWAS	transcriptome based association studies/transcriptome wide association studies
WGS	whole genome sequencing

1. Introduction

During the last two decades, a large number of studies involving linkage disequilibrium (LD) based association mapping (AM), also called as genome wide association studies (GWAS), have been conducted in humans, livestock and crop plants. Candidate-gene (CG) based association studies have also been conducted in some cases (Cockram et al., 2010; Liu, Xue, Guo, Li, & Tang, 2016; Remington et al., 2001; Thornsberry et al., 2001). A GWA study has twin objectives: *first*, to identify marker-trait associations (MTAs) for one trait at a time, and *second*, to study the genetic architecture of the trait. The latter involves identification of all QTLs/genes (including epialleles) and interactions among QTLs identified through GWAS. Large numbers of such MTAs for a number of traits, identified through GWAS, are now available for all major crops. Genetic architecture for a number of traits has also been worked out in all major crops. The extent and level of this information derived from an individual study has also been improving with continuous increase in the size of the association panel and the number of molecular markers that are used for an individual GWAS. The technique of high-throughput (HT) whole genome sequencing (WGS) is also being increasingly used to improve the power and resolution of GWAS. In some cases, meta-analyses of the results from GWAS,

particularly in humans, have also been undertaken to identify additional newer MTAs and to verify the MTAs identified earlier. Newer bioinformatics pipelines have also been developed to extract more meaningful information from genotyping data obtained using next-generation sequencing (NGS) technology, and the high dimensional phenotypic data obtained using phenomics platforms that have been developed and are being used for GWAS. For instance, NGS Eclipse Plugin (NGSEP) has been developed for accurate, efficient and user-friendly analysis of high-throughput sequencing (HTS) data for genotyping the association panel that is needed for GWAS (Perea et al., 2016). For this purpose, "machine learning" is also being increasing used for GWAS in the post-GWAS era.

Identification of a large number of false positives that appear after the original GWA analysis and false negatives that result after application of Bonferroni or false discovery rate (FDR) corrections has been a problem in GWAS. Also, the phenotypic variation for a trait explained through GWAS still remains to be only a fraction of the total phenotypic variation; this feature has also been described as missing heritability, which has been considered as a serious concern in case of humans, although in case of crops, this has not been such a serious problem. The utilization of MTAs obtained through initial GWAS has apparently been only marginal, so that major efforts are needed to utilize the results of GWAS for MAS.

In the post-GWAS-era, newer approaches are being developed for GWAS. GWAS statistics are also being used for further analysis for deriving useful information. Some of these newer approaches include the following: (i) use of expression profiles, resequencing (using NGS), epigenomics, etc., for GWAS; (ii) use of GWAS summary statistics for conditional GWAS, identification of causal SNPs, prioritization among associated markers and identification of candidate genes (for gene annotation); (iii) identification and further analysis of rare alleles and rare variants. Meaningful information has been generated through these newer approaches in several studies in the post-GWAS era. Bioinformatics and reverse genetics approaches are also being used for identification and functional characterization of candidate genes after identification of MTAs through GWAS. The reverse genetics approaches included retrotransposon-mediated gene disruption, gene silencing through RNAi and site specific genome editing using CRISPR/Cas9 nuclease system (Curtin et al., 2011, 2017) and the modified CSISPR/Cas technology involving development of base editors (Komor, Kim, Packer, Zuris, & Liu, 2016).

In this review, we have two major components. In the first half, we briefly describe the achievements of GWAS in important crops and also about the newer models and software that became available during the past more than a decade; this will also include a brief description of the limitations of these earlier studies, which have already been highlighted in a number of recent reviews (e.g., Gupta, Kulwal, & Jaiswal, 2014; Huang & Han, 2014). In the second part of the review, which is more important, we discuss the current and future possible activities in the post-GWAS era, which overlaps the post-genomics era. While doing this, we also make use of literature on different methods proposed in humans and other organisms for this purpose.

2. Major genes/QTL identified for important traits in major crops following GWAS

In all major crops including maize, wheat, rice, sorghum, etc., thousands of GWA studies have been conducted. These studies involved all kinds of traits including the following: (i) developmental and agronomic traits (e.g., plant height, flowering time, leaf architecture, forage quality, etc.); (ii) yield traits (e.g., grain size, grain number, biomass, etc.), (iii) traits associated with various abiotic and biotic stresses and (iv) biochemical and cellular traits (e.g., oil concentration, tocopherol, carotenoid, lipidome, etc.) (see Table 1 for some details). A number of reviews have also been written on GWAS in plants (Ersoz, Yu, & Buckler, 2007; Gupta et al., 2014; Gupta, Rustgi, & Kulwal, 2005; Huang & Han, 2014; Ingvarsson & Street, 2011; Kulwal, 2016; Sukumaran & Yu, 2014; Xiao, Liu, Wu, Warburton, & Yan, 2017). GWA studies have also been used for validation of QTLs that were earlier identified through interval mapping. It has also been recognized that the GWA studies carried out initially in all these crops were based on limited number of markers (few SSRs) and restricted population size. However, in recent years, with the advances in high-throughput genotyping techniques, large association panels with large numbers of markers (in millions) are being used in such studies (Table 1). However, GWA studies undertaken so far could explain only a limited proportion of the total phenotypic variation for individual traits; this is certainly true for GWAS in humans, but to some extent holds good for plants also. Validation of MTAs identified so far is another serious issue, which limits their use in plant breeding. Salient features of the traits for which GWA studies have been conducted in important crops like maize, wheat, rice, sorghum, and soybean are summarized in Table 1.

Table 1 Summary of list of traits for which genome wide association studies have been conducted in major crop plants.

Crop	Type of traits	Traits studied	Marker types used	Number of markers used (range)
Maize	Morphological traits/ agronomic traits	Plant height; ear length and ear architecture; leaf architecture; inflorescence traits; day-length adaptation; flowering time; stalk strength; internode length; lodging resistance; number of kernel rows; root traits at seedling stage; traits related to nitrogen use efficiency	SSR, SNP	82 SSRs to 26.5 million SNPs
	Nutritional/ biochemical traits	Oil related traits; kernel starch, protein and oil content; tocochromanol levels; α-tocopherol content; stover fodder quality; starch content; carotenoid biofortification; carbon and nitrogen metabolism		
	Biotic stresses	Head smut; aflatoxin accumulation resistance; lethal necrosis disease; Northern leaf blight; Southern leaf blight; gray leaf spot; fusarium ear rot; rough dwarf disease; resistance to the Mediterranean corn borer; hypersensitive defense response		
	Abiotic stresses	Cold tolerance; drought tolerance; water logging tolerance; mercury accumulation; lead accumulation; cadmium accumulation		
Wheat	Morphological traits/ agronomic traits	Plant height; agronomic traits; yield and related traits; morphological traits; grain number; grain weight; earliness	SSR, SNP, DArT, STS	51 SSRs to 90K SNP chip
	Biotic stresses	Resistance to cereal cyst nematode; seedling and leaf rust resistance; Stem rust resistance; fusarium head blight resistance; Stagnospora nodurum glume blotch resistance; Russian wheat aphid		
	Abiotic stresses	Drought stress; drought adaptive traits; Aluminum resistance		
	Quality parameters	Pre-harvest sprouting tolerance; grain color; late maturity alpha amylase; dough mixing properties; grain quality traits; grain protein content; alveograph strength		

Rice	Morphological traits/ agronomic traits	Heading date, plant height and panicle length; grain yield; grain length; mesocotyl length of seedling; agronomic traits; harvest index and related traits; developmental and morphological traits; flowering time; stigma and spikelet characteristics	RFLP, SSR, SNP, InDel	75 SSRs to 3.6 million SNPs
	Biotic stresses	Blast resistance; sheath blight resistance		
	Abiotic stresses	Salinity tolerance; aluminum tolerance; straight head disorder		
	Other	Silica concentration in rice hulls		
Sorghum	Morphological traits/ agronomic traits	Plant height; kernel weight; tiller number; inflorescence architecture; seed size; days to maturity; yield components	SSRs, SNPs	47 SSRs to 404,628 SNPs
	Sugar related traits	Sugar yield related traits; saccharification yield; Brix content		
	Biochemical parameters	Flavonoid pigmentation traits; grain polyphenol concentration; polyphenol concentration		
	Abiotic stress	Drought tolerance		
	Disease resistance	Anthracnose resistance		
Barley	Morphological traits/ agronomic traits	Flowering time; heading date; plant height; agronomic and morphological traits; yield and related traits; photoperiod response	SSR, AFLP, DArT, SSAP; SNP	22 SSRs to 9K SNP chip
	Biotic stresses	Fusarium head blight resistance; spot blotch, stripe rust, and leaf rust resistance; stem rust resistance		
	Abiotic stresses	Winter hardiness; frost tolerance; drought tolerance related traits; salinity tolerance; aluminum tolerance		
	Quality parameters	Amylose, amylopectin and β-glucan concentration; malting quality traits; protein fraction content; tocochromanol concentrations		

Continued

Table 1 Summary of list of traits for which genome wide association studies have been conducted in major crop plants.—cont'd

Crop	Type of traits	Traits studied	Marker types used	Number of markers used (range)
Soybean	Morphological traits/ agronomic traits	Growth period; agronomic and morphological traits; flowering time; days to maturity; plant height; domestication traits; seed weight	SSR, SNP	24 SSR to 50,000 SNPs
	Biotic stresses	Resistance to *Phytophthora sojae*, soybean cyst nematode, white mold, brown stem rot, charcoal rot; sensitivity to Tobacco ringspot virus infection		
	Abiotic stresses	Iron deficiency chlorosis; sudden death syndrome; carbon isotope ratio; salt tolerance; canopy wilting		
	Quality parameters	Fatty acid formation; seed protein content; oil content		
	Other	Shoot ureide concentration		

3. Utilization of genes/QTL identified by GWAS for crop improvement

In general, it is recognized that the results of LD-based GWAS are underutilized for crop improvement (Gupta et al., 2014). In contrast, QTLs identified through linkage-based interval mapping have been successfully introgressed in several cultivars for each of several important crops leading to the development of either pre-bred material or improved varieties (Arruda et al., 2016; Brumlop & Finckh, 2011; Gupta, Langridge, & Mir, 2010). There is also at least one documented example of the utilization of the results of GWAS for crop improvement, which includes improvement of provitamin A in maize (Xiao et al., 2017). This study followed an earlier CG-based association mapping study, which allowed identification of rare favorable alleles of two genes, namely *LcyE* for lycopene epsilon cyclase and *crtRB1* for β-carotene hydroxylase 1 (Harjes et al., 2008; Yan et al., 2010). Improved maize genotypes with higher level of provitamin A developed through MAS are already being used for commercial cultivation, thus addressing the problem of malnutrition of children in Africa. In addition to this, there may be many other undocumented examples of the utilization of the results of GWAS in the private sector.

4. Improvement in GWAS over the years

4.1 Development of newer methods

Development of newer methods mainly involved regular development of improved statistical models and tools (Lipka et al., 2015). This became necessary due to the use of association panels of larger size and due to availability of millions of SNPs for each individual study. Many of the limitations of GWA studies have thus been overcome through newer computational methods involving improved statistical models. These models had major emphasis on reducing the computational demand and were discussed by us in an earlier review (Gupta et al., 2014). These earlier approaches along with newer approaches are listed in Table 2, and will be described briefly in this section.

4.1.1 Single-locus, single trait (SLST) mixed models

Majority of initial GWA studies involved analysis of a single-locus and a single trait at a time; methods for multi-locus and multi-trait were developed

Table 2 Different mixed model approaches proposed over the years for GWAS in crop plants along with their features.

SN	Approach	Features	Reference
1.	Mixed liner model (MLM)	Takes care of multiple levels of relatedness; effectively controls population structure and type I and type II error rates	Yu et al. (2006)
2.	Genome-wide Rapid Association using Mixed Model and Regression (GRAMMAR)	An approximate method which first estimates the residuals adjusted for family effects and then treats these as phenotypes along with genotyping data for analysis using rapid least-squares methods; reduces computation time for each individual SNP	Aulchenko, De Koning, and Haley (2007)
3.	Efficient mixed-model association (EMMA)	An exact method that accounts for population structure and genetic relatedness with substantially increased computational speed and reliability of the results	Kang et al. (2008)
4.	Efficient mixed-model association eXpedited (EMMAX)	An approximate method in which VCA is not repeated for each marker, as each marker is assumed to explain only a small fraction of phenotypic effect; instead, heritability estimated from the null model is used for all markers; can perform AM using vast amount of data in a short time	Kang et al. (2010)
5.	Compressed mixed linear model (CMLM)	Clusters the individuals into fewer groups based on the kinship among the individuals; the kinship between pairs of groups is replaced by the kinship between pairs of individuals; reduces the computation demand substantially	Zhang et al. (2010)
6.	Population parameters previously determined (P3D)	A complementary approach to CMLM; eliminates the need of estimating population parameters (such as VCs); computationally fast	Zhang et al. (2010)
7.	Factored Spectrally Transformed Linear Mixed Models (FaST-LMM)	An exact method with improvement over MLM approach brought out by use of a low-rank relatedness matrix (matrix based on a few thousand markers instead of all the markers); reduces computation time considerably	Lippert et al. (2011)
8.	Multi-locus mixed model (MLMM)	An improvement over MLM; can effectively control for population structure and false discovery rate in GWA studies; takes into account the background genotypes	Segura et al. (2012)

9.	Multi-trait mixed model (MTMM)	Performs GWAS of correlated phenotypes using the principle of MLM; takes into account both, within-trait and between-trait VCs simultaneously for multiple traits	Korte et al. (2012)
10.	GRAMMAR-Gamma	A VC-based two-step approximate method; an improvement over GRAMMAR; reduces computational demand and provides correct estimates of SNP effects; suitable for using genotyping data based on whole-genome resequencing with large sample size	Svishcheva, Axenovich, Belonogova, van Duijn, and Aulchenko (2012)
11.	GEMMA	An efficient-exact method; faster than EMMA; yields accurate *p* values even in the presence of strong population structure, and even when the marker effect is large; suitable for studies with large association panels	Zhou and Stephens (2012)
12.	Linear mixed model-Lasso (LMM-Lasso)	Combines multivariate analysis and corrects for population structure (combination of MLM and Lasso regression); can partition the total phenotypic variance into different components, like the one caused due to individual SNP effects as well as that caused by population structure	Rakitsch, Lippert, Stegle, and Borgwardt (2013)
13.	Selecting CONnected Explanatory SNPs (SConES)	An efficient multi-locus method for discovering sets of loci which are associated with a phenotype while being connected in an underlying network; computationally fast	Azencott, Grimm, Sugiyama, Kawahara, and Borgwardt (2013)
14.	Low rank linear mixed model (LRLMM)	Takes into account the effective degrees of freedom for interpreting model complexity of the LRLMM along with principal components (for controlling population structure) and kinship	Hoffman (2013)
15.	Bayesian sparse linear mixed model (BSLMM)	A combination of MLM and sparse regression models	Zhou, Carbonetto, and Stephens (2013)
16.	Settlement of MLM Under Progressively Exclusive Relationship (SUPER)	An improvement over FaST-LMM; extracts a subset of SNPs and uses them in FaST-LMM; increased statistical power	Wang, Tian, Pan, Buckler, and Zhang (2014)

Continued

Table 2 Different mixed model approaches proposed over the years for GWAS in crop plants along with their features.—cont'd

SN	Approach	Features	Reference
17.	Genetic analysis incorporating Pleiotropy and Annotation (GPA)	Enables joint analysis of multiple GWA data sets and the annotation information	Chung, Yang, Li, Gelernter, and Zhao (2014)
18.	Enriched CMLM (ECMLM)	An improvement over CMLM with increased statistical power; calculates kinship using several different algorithms and uses this information during analysis	Li, Liu, et al. (2014)
19.	Principal components-Select (PC-Select)	A hybrid approach that includes the PCs of the genotype matrix as fixed effects in FaSTLMM Select method	Tucker, Price, and Berger (2014)
20.	Multivariate linear mixed models (mvLMM)	Uses computationally-efficient algorithm for fitting mvLMMs with one covariance component (in addition to the residual error term), and for performing the LR test for GWAS; improvement over GEMMA	Zhou and Stephens (2014)
21.	BOLT-LMM	Based on Bayesian mixed-model association; increased computational power	Loh et al. (2015)
22.	Random-SNP-effect MLM (RMLM)	SNP-effects are treated as random; the threshold p value for significance tests are calculated based on a modified Bonferroni correction	Wang, Feng, et al. (2016)
23.	Multi-locus RMLM (MRMLM)	A multi-locus model that includes markers selected from the RMLM with less stringent selection criterion; multiple test correction is not required	Wang, Feng, et al. (2016)
24.	Fixed and random model Circulating Probability Unification (FarmCPU)	Combines both, the fixed effect and random effect models in analysis and improves statistical power with reduced computing time	Liu, Huang, Fan, Buckler, and Zhang (2016)
25.	Penalized multitrait mixed modeling approach	Accommodate both types of correlations, i.e., between subjects and traits during analysis	Liu, Yang, et al. (2016)
26.	pLARmEB (polygenic-background-control-based least angle regression plus empirical Bayes)	Integrates least angle regression with empirical Bayes and can perform multilocus GWAS; it is more powerful in detection of QTN and its effect; has less false positive rate and require less computing time than Bayesian hierarchical generalized linear model	Zhang et al. (2017)

later (mainly during the last 5–10 years). The first major development in GWAS was the availability of mixed models, which take into account both fixed and random effects and therefore, became popular. The first mixed model for GWAS was proposed by Yu et al. (2006) and has since been undergoing improvements on a regular basis to deal with the problem of computational demand and false positives. These methods have been classified into two major groups: (i) exact methods and (ii) approximate methods. The exact methods like GEMMA and Fast-LMM efficiently refit the model for every marker and provides exact estimates of marker effects; these methods are, however, comparatively slower (Lipka et al., 2015; Zhou & Stephens, 2012), and standard test statistics cannot be used in a user-friendly manner. In contrast, the approximate methods like EMMAX and GRAMMAR are computationally fast, since they eliminate the need of estimating population parameters for every marker. They are based on the score test statistic instead of the likelihood ratio test (LRT) statistic that is considered to be the gold-standard; the two test statistics are, however, asymptotically equivalent. Since dozens of these methods became available, a comparison of different statistical methods and the GWAS programs where these were used has often been made (Eu-ahsunthornwattana et al., 2014; Zhang, Buckler, Casstevens, & Bradbury, 2009; Zhou & Stephens, 2012). However, the choice of method for analysis should be based on the volume of data, desired speed of analysis and the level of user-friendliness (Eu-ahsunthornwattana et al., 2014; Gupta et al., 2014).

4.1.2 Multi-locus and multi-trait mixed models (MLMM and MTMM)

The above SLST analysis carries with it the problem of multiple testing, background genotype effect and inability to identify pleiotropic effects, if any (Buzdugan et al., 2016). In view of this, new approaches and computational tools have been developed, which can analyze multiple loci and multiple traits individually or together in an analysis.

(a) Multi-locus models

Methods proposed for multi-locus analysis include the following: (i) a Bayesian-inspired penalized maximum likelihood approach (Hoggart, Whittaker, De Iorio, & Balding, 2008), (ii) penalized logistic regression approach (Ayers & Cordell, 2010), (iii) elastic-net approach (Cho et al., 2010), (iv) empirical Bayes approach (Lu, Liu, Wei, & Zhang, 2011), (v) multi-locus mixed model (MLMM; Segura et al., 2012) and (vi) random-SNP-effect MLM (RMLM; Wang, Feng, et al., 2016). However, not all these methods are efficient, when the number of markers

exceed the number of genotypes, a problem often described as "*large p, small n*" problem. In order to handle this issue of high dimensionality, Buzdugan et al. (2016) proposed a multivariate approach, which analyzes all the SNPs in a multiple GLM setup and yields *p*-values for assessing significance of single SNP or groups of SNPs (haplotypes), while controlling the effect of all other SNPs. Similarly, Zhou, Hu, Qiao, Cho, and Zhou (2016) proposed several approaches including computationally efficient, exact (nonasymptotic) score (eScore), likelihood ratio test (eLRT) and restricted likelihood ratio test (eRLRT). These tests are described as efficient exact variance component tests (ExactVCTests) and can achieve high power, even when size of samples is small. The underlying idea for these tests is that the SNP-set used examines groups of SNPs (haplotypes) rather than individual SNPs, and can enhance the power of detection. The ExactVCTests are supposed to be superior to the popular sequence kernel association test (SKAT; Wu et al., 2011), particularly when the sample size used in analysis is small. However, it is necessary that the data being studied should be normally distributed, which is not always possible (this condition is not required in case of SKAT). The efficiency of almost all these multi-locus methods (except empirical Bayes) was demonstrated using the data from humans. It will therefore be interesting to see their practical utility using the data from plant systems.

Often, in any GWA study, correction for population structure is undertaken as a measure to reduce confounding, which often leads to the generation of false positives/negatives (Klasen et al., 2016). In order to address this issue, a Quantitative Trait Cluster Association Test (QTCAT) was proposed by Klasen et al. (2016). Taking into account the correlations between the markers, this test can identify multi-locus associations simultaneously. It is also not necessary to correct for population structure and genetic background in this method, which was also found to be better than the other commonly used LMMs.

(b) Multi-trait models

We know that there can be genomic regions (QTLs/genes), each controlling more than one traits (Korte et al., 2012; Zhan et al., 2017). A mixed model approach, called multi-trait MLM (MTMM) accommodating pairs of correlated traits, was proposed by Korte et al. (2012). This model increases the number of tests relative to those conducted in a single-trait analysis and can identify MTAs, each MTA controlling a pair of correlated traits (for a review, see Gupta et al., 2014).

This approach was successfully used in our own GWA studies involving a variety of correlated traits in bread wheat (Jaiswal et al., 2016; Kumar et al., 2018). Unfortunately, in this method, one could use only two correlated traits at a time, so that methods had to be developed, where more than two correlated trait can be used for multi-trait analysis. Furlotte and Eskin (2015) proposed another method called matrix-variate linear mixed model (mvLMM) which can reduce the time required to perform maximum-likelihood inference by using data transformation in a multiple-trait model. Such multi-trait analyses can also be combined now with multi-locus analysis (see step (c) in Section 4.1.2).

The power of multiple trait analysis can also be improved by transforming multiple traits into a group of pseudo-principal components based on residual covariance matrix. Using this principle, Gao et al. (2014) proposed a method which performs PCA for the residual covariance matrix and allows analysis of each pseudo-PC separately. Several methods have been proposed for the analysis of multi-trait data in the GWAS framework in humans; different popular methods have also been compared (Galesloot, Van Steen, Kiemeney, Janss, & Vermeulen, 2014; Porter & O'Reilly, 2017). In one such comparative study, Porter and O'Reilly (2017) made several conclusions: *first*, that the performance of any method is dependent on specific combination of genetic effects and phenotypic correlations; *second*, that most of the available methods have similar statistical power, and *finally*, that these methods can offer a substantial improvement in the discovery of genetic variants over the standard univariate approach (Porter & O'Reilly, 2017). Recently, Thoen et al. (2017) used MTMM approach for a study of the genetics of responses to different stresses in Arabidopsis using 11 individual stresses as well as several combinations of different stresses. Using the principle of MTMM, they identified candidate genes (CGs) for plant responses to multiple stresses. These CGs were also validated by gene expression and mutant analyses. Recently, Turley et al. (2018) proposed another method called Multi-Trait Analysis of GWAS (MTAG), which can analyze multiple traits and has unique features which can be used with GWAS summary statistics (meta-analysis). This can handle the sample overlap (it is not necessary that the summary statistics should come from independent samples) and can quickly generate estimates of trait-specific effect for individual SNP. This method takes advantage of the correlations between the traits and correlations between the estimation errors of SNPs effects across traits.

(c) Multi-locus, multi-trait models

Multi-locus multi-trait analyses are more rewarding than any of the methods involving either multi-locus analysis or multi-trait analysis individually (Lippert, Casale, Rakitsch, & Stegle, 2014). These models also address the concerns for high computational demands. For instance, Kim, Zhang, and Pan (2016) proposed an association test, which involves analysis of multiple genetic variants and multiple traits simultaneously. With this method, they observed gain in power in identification of multiple associated SNPs having weak effects in the gene *AMOTL1* which is involved in the human brain default mode network. These SNPs could not be identified by using single SNP-based tests as well as several other gene-based tests. This method can also be used for analysis of rare variants in sequencing data and can also be used for pathway analysis. The method is yet to be tried in plant systems. Similarly, Zhan et al. (2017) proposed a novel approach called "dual kernel-based association test" (DKAT), which can evaluate association between high-dimensional structured traits and multiple SNPs or rare variants. This method also addresses the "large p small n problem," since variable number of markers can be used for association panels, which differ in size.

4.1.3 Joint-linkage association mapping (JLAM)

Over the years, it has been recognized that linkage-based QTL interval mapping (IM) and LD-based AM, each has its own merits and demerits (low resolution in IM and low power in AM), when performed independently. In order to harness the potential of both these methods simultaneously, integration of these two approaches into one approach was proposed and referred to as joint linkage-association mapping (JLAM) (Wu & Zeng, 2001). In crop plants, initially JLAM was made possible by Nested Association Mapping (NAM) involving multiparental populations in maize (Yu, Holland, McMullen, & Buckler, 2008). Later, many more such multiparental mapping populations (MPP) including Multi-parent Advanced Generation Intercross (MAGIC), Multi-line Cross Inbred Lines (MCILs) and Recombinant Inbred Advanced Intercross Lines (RIAILs) were developed in several important crop plants (Cavanagh, Morell, Mackay, & Powell, 2008; Gupta, Kulwal, & Mir, 2013; Huang et al., 2015) facilitating JLAM. This approach allows not only identification but also validation of QTLs in the same experiment, so that the validated QTLs and the linked markers can then be used in crop improvement programs efficiently.

The utility of JLAM has also been enhanced by incorporating multi-trait data analysis (Meuwissen & Goddard, 2004; Stich et al., 2008; Wu, Ma, &

Casella, 2002). Recently, it was also reported that for JLAM, the model accounting for cofactors and a population effect can effectively control population structure and has a high predictive power than any other linear and linear mixed model (Wurschum et al., 2012). Another advantage of JLAM approach is that the issue of rare alleles can be effectively addressed (Gupta et al., 2014). Notable examples of JLAM studies include identification of QTLs and the CGs for drought tolerance in maize (Lu et al., 2010) and pleiotropic QTLs for seed weight and silique length in rapeseed (Li, Shi, Wang, Liu, & Wang, 2014).

4.1.4 Use of diverse panels for GWAS

Biparental populations, multiparental populations [e.g., NAM, MAGIC and random-open-parent association mapping (ROAM)] and breeding populations have already been used both for linkage studies and GWAS. Each of these population types has their own advantages and limitations, as discussed elsewhere (Gupta et al., 2014; Wang, Xu, et al., 2017; Xiao et al., 2017). The choice of any of these populations also depends upon the nature of breeding system (inbred versus outbred) in the plant species, so that a multi-parental population like NAM may not be as desirable in autogamous crops, as in case of highly cross pollinated-crops like maize. Utilizing the principle of NAM, new designs have also been proposed, which include doubled haploid NAM (DH-NAM) and backcross NAM (BC-NAM) in rapeseed (Li, Bus, Spamer, & Stich, 2016) and advanced backcross NAM (AB-NAM) in barley (Nice et al., 2016). A multiple-hybrid population (MHP) has also been used in maize for identification of key loci for flowering traits (Wang, Xu, et al., 2017). It was shown that for cross-pollinated crops, MHPs are better suited relative to bi-parental and multi-parental populations that are commonly used. This GWAS approach can make use of a panel consisting of a limited number of parental lines and the large numbers of hybrids that can be generated by crossing the parental lines in diallel fashion. It has also been argued that sharing a set of parental lines among collaborators is easier than sharing an association panel with several hundreds to thousands of genotypes (Wang, Xu, et al., 2017). More such studies in cross-pollinated crops are likely to be conducted in future. In the year 2015, the journals Genetics and G3 also started a series on development and use of multiparental populations (MPPs) for linkage and association mapping studies. More than a dozen papers have already been published under this series which provide valuable information on the use of MPPs for QTL analysis involving both linkage based IM and LD-based AM (Bouchet et al., 2017).

4.1.5 Epistasis and Q × E interactions

The detection of markers associated with Q × Q interactions (epistasis) and QTL × environment (Q × E) interactions has not been a regular feature of GWAS. Apparently, this is one of the several reasons why only limited fraction of total genetic variation could be explained in most GWA studies, particularly in humans, but to some extent in plants also (Bubb & Queitsch, 2016; Ritchie, 2015; Upton, Trelles, Cornejo-Garcıa, & Perkins, 2015). Several methods and the associated software are now available for detection of epistasis in GWA studies (for reviews see Upton et al., 2015; Wei, Hemani, & Haley, 2014). Choice of method also depends on the size of the data set, the aim of the user and the availability of processing power (Upton et al., 2015). Bayesian statistical methods which have already shown tremendous promise in genetic analyses is an alternative approach; however, these methods also do not examine all pairwise interactions (Upton et al., 2015).

Often epistasis is discussed in relation to its effect as either only functional or only statistical (Wei et al., 2014). While, functional epistasis suggests that the effect of a variant at one locus depends on the variant at another locus, the statistical epistasis refers to the variance attributed to the interaction between variants that are causal only when occurring together, but not in terms of their independent effects (Wei et al., 2014). If the amount of epistasis is substantial, then the predictability of complex traits based on genotypic data can be improved beyond the theoretical limits of heritability estimates.

Q × E interactions are also sometimes substantial and need attention, particularly for explaining missing-heritability (Thomas, 2010). Several computational methods for GWAS have been proposed to study such interactions in humans (Thomas, 2010). However, they may not always be suitable for studies involving plants. The MTMM approach proposed by Korte et al. (2012), which accounts for the correlations between traits can be useful in finding G × E interactions. In doing so, one can consider a trait measured in two environments as two correlated traits (Korte et al., 2012; Korte & Farlow, 2013).

Reports of GWAS involving epistasis and Q × E interactions in plants are limited. However, in a GWAS involving analysis of G × E interactions, often a small set of genotypes and a limited marker data set have been used. In one such study, Li, Paulo, van Eeuwijk, and Gebhardt (2010) identified two-way epistatic interactions for tuber traits in potato, which were responsible for increased starch content and starch yield. Later, Lu et al. (2011) proposed an epistatic association mapping (EAM) approach in plants using empirical Bayes approach, which includes in its model the main-effect QTL, environmental effects, QTL × QTL interactions and Q × E interactions.

The method was used in soybean for study of the genetics of seed length; three epistatic QTLs were identified in addition to the main effect QTL. Similarly, Saidou, Thuillet, Couderc, Mariac, and Vigouroux (2014) proposed a mixed-model approach, which accounts for SNP × environment as well as Q × E, SNP × structure and three way interactions between SNP, ancestry and the environment; the approach was used for detecting these interactions in pearl millet and maize. Jia et al. (2014) reported significant epistasis and environmental interaction for yield component traits in cotton. In a rather interesting study in Arabidopis, Lachowiec, Shen, Queitsch, and Carlborg (2015) reported that narrow-sense genomic heritability for root length was statistically zero, which resulted in no associations with the root length in the GWA study. However, epistatic GWA analysis identified four significant interacting pairs of loci for root length. It was also reported that epistasis canceled out the additive genetic variance, and was responsible for non-significance of these loci in the additive GWA analysis (Lachowiec et al., 2015). In the post-GWAS era, efficient integration and handling of rare variants in the analysis pipeline involving epistatic interaction will however be difficult. With the advances in the computational techniques capable of analyzing high-dimensional data, the data used in earlier studies can be reanalyzed to find the epistatic interactions in these studies.

4.1.6 Bayesian methods for GWAS

In recent years, Bayesian statistical methods have become an important, rather integral part of studies involving genotype-phenotype associations and genomic prediction. What makes them superior over the usual or frequentist approach is their ability to specifically incorporate background information (prior) into the specification of the model (Stephens & Balding, 2009). Bayesian methods can be effective for fine mapping in candidate regions (Schaid, Chen, & Larson, 2018) as well as in studies involving meta-analyses. These approaches not only increase the computation speed but also deal with the problems of multiple testing and rare marker alleles (Fernando & Garrick, 2013). In GWAS, generally, issue of multiple testing is handled by controlling the genome-wide error rate (GWER), which is the probability of occurence of atleast one false-positive QTL among all tests. However, due to numbers of such tests involved in a GWA study, control for GWER can result in low power (Fernando & Garrick, 2013). Therefore, an alternative approach based on Bayesian regression method has been suggested, which accounts for multiple testing correction by controlling the proportion of false positives among all the positives in the study. Using the same principle of Bayesian regression,

recently, Fernando, Toosi, Wolc, Garrick, and Dekkers (2017) showed that in a GWA study, for controlling false positives, it is more appropriate to control the posterior type I error rate than to control the GWER. Another advantage with Bayesian regression models is that they can simultaneously fit more markers in the analysis as compared to the number of observations. It is because of this growing interest in the Bayesian statistics that newer methods are constantly being developed which can help in the integration of the functional information in GWAS (Yang et al., 2017). It is also important to note that Bayesian statisticians use probability theory to model uncertainty in analysis (Ball, 2013). In the era of machine learning involving use of artificial intelligence, it is expected that Bayesian approaches will be used more frequently in GWAS. Several mixed model approaches based on Bayesian statistics have been proposed for use in GWAS; some of them are listed in Table 2. The differences between frequentist and Bayesian approaches have been discussed elsewhere (Ball, 2013).

4.2 High-throughput (HT) genotyping and phenotyping

4.2.1 HT genotyping

The availability of NGS and GBS for SNP genotyping has greatly accelerated the speed of GWA studies. Consequently, several thousands to millions of markers are now being used in individual GWA studies. The technique of GBS is also rapidly becoming popular due to its cost-effectiveness even in the crops, where a reference genome sequence is not available (Elshire et al., 2011; He et al., 2014). In the post-GWAS framework, the issues which will drive the methodological advances should focus now on efficient handling of rare alleles, missing marker data and appropriate treatment of the multiple testing problems arising due to the progress made in the area of genotyping techniques.

(a) HT marker techniques

Advances in the cost-effective NGS techniques have facilitated whole genome sequencing/resequencing of the association panels in many crops. Resequencing of the entire association panel, albeit at a low depth coverage has helped in the identification of novel MTAs through GWAS in some major crops. Details of some of these studies are summarized in Table 3. These studies have shown that sequencing even at low depth coupled with efficient marker-imputation techniques can be very effective in identifying novel loci in large sample populations using the approach of GWAS. This has offered greater understanding of the genetic basis of agronomically important traits as well as domestication process in several crops. Besides this, RNA sequencing (RNA-seq)

Table 3 Examples of resequencing based GWA studies in important crops.

Crop	Number of accessions	Coverage depth	Number of markers	Important features of the study	Reference
Rice	517 landraces	1×	~3.6 million SNPs	High-density haplotype map was used to identify key loci for 14 different agronomic traits; 6 of the identified MTAs were in close proximity of the previously identified genes	Huang et al. (2010)
	950 worldwide rice cultivars	1×	~4.1 million SNPs	32 new loci associated with 10 grain-related traits and flowering time were identified; CGs for 18 associated loci were also identified	Huang et al. (2012)
	176 japonica rice varieties	5.8×	426,337 SNPs and 67,544 InDels	Important genes for agronomic traits were identified; later CGs were screened and four new genes for these traits were identified; shows that gene-based association analysis can be more-rewarding in dealing with spurious associations	Yano et al. (2016)
Foxtail millet	916 diverse lines	~0.7×	0.8 million SNPs	A haplotype map allowed identification of 512 loci for 47 agronomic traits	Jia et al. (2013)
Soybean	302 wild and cultivated accessions	>11×	>9 million (SNP+InDels)	10 MTAs were identified for 9 domestication or improvement traits; 13 novel loci were identified	Zhou et al. (2015)

Continued

Table 3 Examples of resequencing based GWA studies in important crops.—cont'd

Crop	Number of accessions	Coverage depth	Number of markers	Important features of the study	Reference
Chickpea	69 varieties and advanced breeding lines	3.35×	~0.4 million SNPs	A 100 kb region (AB4.1) from chromosome 4 was found to be associated with resistance to Ascochyta blight	Li, Ruperao, et al. (2017)
Pigeonpea	292 accessions (breeding lines, landraces and wild species)	5–12×	446,568 SNPs	Several CGs associated with agronomically important traits were identified and had sequence similarity with the genes which were functionally characterized in other plants for traits like control of flowering time, seed development and pod dehiscence	Varshney, Saxena, et al. (2017)
Pearl millet	288 test cross hybrids	1.68×	3,117,056 SNPs	1054 MTAs associated with 15 different yield and yield contributing traits were identified	Varshney, Shi, et al. (2017)

has also become common in many crop plants for transcriptome-wide association analysis; useful markers have also been generated from the transcriptome (see Section 7.2.5.1).

Different HT marker genotyping techniques which are now available have their own advantages and limitations (Huang & Han, 2014; Pfeifer, 2017) so that choice of method depends on the objective of the study and the resources available. However, in order to make use of these NGS techniques, one should have thorough understanding of the data so generated, and should also have knowledge of different available methods for analysis of this data (Pfeifer, 2017). The volume of genotypic data obtained using any of these techniques often brings computational and statistical challenges, which demand appropriate analytical tools for effective analysis. Moreover, low sequence coverage generates lot of missing data, which necessitates constant improvements in the marker imputation techniques. Some of these issues including characteristics of popular NGS platforms, tools for pre-processing of NGS data, NGS aligners and SNP callers have been discussed by Pfeifer (2017).

HT marker techniques have also allowed generation of some new marker systems like copy number variations (CNVs) and presence/absence variations (PAVs), InDels, insertion-site-based polymorphisms (ISBPs), epigenetic variations, and transposons. In future, these markers will be increasingly utilized for GWAS (for reviews, see Edwards & Gupta, 2013; Lipka et al., 2015).

(b) Missing data and imputation for genotyping

In recent years, although HT sequencing has accelerated the speed of marker development, it has also created the problem of missing data. This poses some problem in using these markers in analysis and increases the chances of spurious associations. In several studies, data imputation involving use of a reference genome has been used to overcome this limitation. In multiple genetic studies, it has been reported that marker imputation greatly increased the efficiency of GWAS in crop plants (He et al., 2015; Ramstein et al., 2015). It has also been realized that in the absence of imputation, a sufficiently higher sequencing depth would be required for achieving the desired sensitivity of analysis; this may not be possible in a cost-effective and time-effective manner.

Imputation of genotypes not only increases the power of association studies, but also harmonizes data sets for meta-analysis (Porcu, Sanna, Fuchsberger, & Fritsche, 2013; see Section 7.2.1). Several algorithms and computer programs are available for this purpose, but not all of

them may be suitable for plants. Therefore, algorithms have also been developed to handle the data from crop plants including those for which reference genome sequence is not available (Ramstein et al., 2015; Rutkoski, Poland, Jannink, & Sorrells, 2013; Swarts et al., 2014; Ward et al., 2013). In recent years, emphasis is also laid on the use of multiparental populations (MPPs) in GWAS. However, unavailability of good quality data on founder genotypes often creates a problem while conducting imputation of the genotypes. Accordingly, a method has been proposed, which can effectively impute the marker genotypes for the founder genotypes (Huang, Raghavan, Mauleon, Broman, & Leung, 2014). Accuracy resulting from imputation generally depends on the availability of a reference sequence. For example, using a reference sequence, imputation accuracy is high even for low depth GBS data as against the one where no reference sequence is available. However, one of the challenges with imputation techniques is the ability to distinguish correctly between missing data arising due to biological reasons (PAVs) and that arising due to sampling variation (Lipka et al., 2015). Technique of genotype imputation is also useful in dealing with the rare variants. One of the efficient ways to deal with rare variants is to impute genotypes for missing data from corresponding markers into existing genome-wide data (information from a reference panel) (Hoffmann & Witte, 2015). However, this is dependent on the marker array and the number of reference samples used for this purpose. Different strategies have been proposed for genotype imputation in the rare variants (Hoffmann & Witte, 2015).

4.2.2 HT *phenotyping and genotype-phenotyping gap*

HT genotyping platforms are now being extensively used for genotyping at a large scale (millions of markers). However, similar developments did not take place for HT phenotyping. This is often described as genotype–phenotype gap (Gjuvsland, Vik, Beard, Hunter, & Omholt, 2013). In order to bridge this gap, high-throughput phenomics technologies are being regularly developed (Yang, Duan, Chen, Xiong, & Liu, 2013). Among all HT phenotyping tools, image-based phenomics is the most important development (see next section for some details).

(a) Image-based phenomics and software for analysis of phenomics data

During the last one decade different image-based phenomics facilities have been developed in the public domain around the world (Australia, France, China, USA, India and member countries of European Plant Phenomics Network); details are available in Knecht,

Campbell, Caprez, Swanson, and Walia (2016). All these facilities are fully automated and allow accurate and non-destructive recording of data through a series of cameras, thus eliminating errors, which occur in traditional phenotyping methods. This greatly enhances the ability to record data on various quantitative traits on a temporal and spatial manner, and make high-throughput plant phenotyping possible. Various factors associated with image-based phenomics, setting up of such experiments, handling and analysis of data, and many other issues have been described in detail by Fraas and Luthen (2015).

In any HT-phenotyping, often large-scale image data-sets are used, which cannot be easily processed on desk-top computers. Therefore, software has also been developed for processing the data. The number of files generated for a GWAS experiment can easily exceed millions, thus making the task of handling this data computationally challenging. This problem has largely been overcome through several resources including RootNav (Pound, French, Atkinson, Well, & Bennett, 2013), GrainScan (Whan et al., 2014), Integrated analysis platform (Klukas, Chen, & Pape, 2014), TraitCapture (Brown et al., 2014), PlantCV (Fahlgren, Gehan, & Baxter, 2015), leaf angle processing toolbox (Müller-Linow, Pinto-Espinosa, Scharr, & Rascher, 2015), LemnaGrid (www.LemnaTec.com), Image Harvest (Knecht et al., 2016) and others (Fraas & Luthen, 2015).

(b) Examples of GWAS involving HT phenotyping

HT phenotyping has already been utilized in different plant systems for a number of quantitative traits; some of these studies include the following: (i) HT phenotyping data through automated confocal microscopy for a study of symbiotic/pathogen interactions, disease resistance and rate of root growth in model legume *Lotus* (Hansen, 2014); (ii) HT phenotyping of 15 agronomic traits in rice using HRPF (Yang et al., 2014); (iii) phenotyping for carotenoid content as well as chlorophyll traits including chlorophyll *a*, chlorophyll *b* and total chlorophyll content estimated through HT canopy spectral reflectance in soybean (Dhanapal et al., 2015, 2016); (iv) phenotyping for salinity tolerance in rice using high-throughput visible and fluorescence imaging (Al-Tamimi et al., 2016; Campbell et al., 2015).

5. Limitations of GWAS

Although the approach of GWAS has proved useful in identifying large numbers of loci for a variety of traits, both in humans and plants, it

has certain limitations which need to be addressed carefully in order to harness the utility of this approach. These limitations include but are not limited to population structure, familial relatedness, computational speed, multiple testing problem, "large *p* small *n* problem" and FDR, markers with rare alleles and rare genetic variants. While these issues have been discussed by us earlier (Gupta et al., 2014) and elsewhere and solutions have been sought to effectively manage these concerns, we will emphasize on other issues which were not discussed in greater details earlier. Identification of false positive (false discovery) associations as well as reproducibility of identified loci and family-wise error rate (FWER) are now considered the major limitations of this approach and are discussed here in brief.

5.1 False discovery, reproducibility and family-wise error rate (FWER) in GWAS

Besides other limitations discussed earlier, following are the two other major limitations of GWAS, which need further discussion: (i) validity of FDR corrections and (ii) lack of reproducibility. As discussed earlier, for FDR, generally Bonferroni's correction and Storey's FDR correction are used, which often represent a trade-off, because when the correction is applied, it leads to excessive false negatives. Similarly, if a particular study is repeated in the same or another laboratory using the same population for the same trait, the reproducibility has been found to be no more than 30%, and sometimes as low as 4% (Hirschhorn, Lohmueller, Byrne, & Hirschhorn, 2002). Therefore, experiments have been conducted to test not only the validity of FDR correction, but also to estimate the reproducibility rate (RR) versus false irreproducibility rate (FIR) (Jiang, Xue, & Yu, 2015). RR is the conditional probability that measures the likelihood of primary association being also positive in the replication study, whereas FIR measures the likelihood of primary positive association being true positive, even if it is not positive in the replication study.

Hirschhorn et al. (2002) reviewed the work involving reproducibility of GWAS results and found that of the 600 associations that were found positive between common gene variants and diseases in humans, majority of them were not robust. Only six associations from among the 166 associations which were studied three or more times were found to be consistently replicated, while >50% of the remaining 160 associations were observed only in one or more studies again. They studied the possible causes for this

irreproducibility and also provided guidelines for conducting and interpreting such studies.

5.2 How to reduce false positives and how to validate FDR corrections

In order to reduce false positives, corrections are generally used but additional evidence is often needed to validate the results of FDR corrections. Commonly, there are two strategies used to verify associations discovered using GWAS: one is joint analysis and the other is replication based analysis. (i) Joint analysis uses all available GWAS data for the same trait (e.g., disease) in the same population to find associated SNPs, either by pooling multiple stage genotyping data (mega-analysis) or by using meta-analysis; additional biological experiments are needed to verify the associations. (ii) Replication-based analysis splits the data into two parts, one for discovery (commonly called primary study) and the other for validation (commonly called replication study). Since only a subset of available data is used each in the primary study and the replication study, the replication-based analysis is less powerful than joint analysis (Skol, Scott, Abecasis, & Boehnke, 2006). However, it gives us an alternative way to examine findings without carrying out additional experiments (Chanock et al., 2007; Kraft, Zeggini, & Ioannidis, 2009). Thus, replication-based analysis is a common method of choice, which is also cost-effective.

5.3 Reproducibility of GWAS results

A number of examples of lack of reproducibility of the results of GWAS have been reported in humans (Hong et al., 2012). Some results of lack of reproducibility are also available in plants. For instance, in maize, limited reproducibility of small effect associations involving the trait southern leaf blight resistance was observed using an improved NAM population. While examining several factors, the effect of marker density was found to have the maximum effect on the accuracy of results of QTL interval mapping and GWAS; the study also facilitated identification of CGs for southern blight resistance (Bian, Yang, Balint-Kurti, Wisser, & Holland, 2014).

5.4 FDR versus FWER and power

We know that multiple testing in GWAS leads to false positives at a rate much higher than in the normal cases. Therefore, often we make estimates of the "probability (P) of having one or more false significant tests." This is

described as family-wise error rate (FWER). Many times FDR is also used as an alternative to the commonly used FWER (Xu & Iglewicz, 2016). Although different approaches of multiple testing have been proposed with the same goal, but each of these methods deal with this problem in different ways. FWER approach may be more appropriate when one wants to have stringency in the results by avoiding any false positives (for example, in case of human genetics). However, in practice using genomics data, one can expect at least certain number of false positives (for example, plant genomics). Under this scenario, approach of FDR is more relevant. Recently, Stevens, Al Masud, and Suyundikov (2017) compared different methods underlying principle of either FDR or FWER. Based on simulation studies, they concluded that different methods differ in power depending on the sample size and the number of tests involved, so that one method cannot be suitable in all scenarios and one should consider the trade-off between specificity and the sensitivity of the method.

6. Rare variants and missing heritability

During GWAS, markers with rare alleles are often excluded from the analysis (Marjoram & Thomas, 2014). This is also considered to be one of the several reasons attributed for the missing heritability (Manolio et al., 2009). It may be possible that at least some of these markers are associated with desirable traits, although it is still not known as to what extent these rare genetic variants actually contribute to the target traits both in humans and plant systems. Therefore, rare alleles/variants analysis has become an important area of research, and will be discussed in relatively greater detail.

6.1 Types of variants and MTAs: Common variants with small effects versus rare variants with large effects

The variants commonly used for GWAS have often been classified into the following three groups: (i) common variants [with minor allele frequency (MAF) >5%], (ii) less common variants (MAF, 1%–5%) and (iii) rare variants (MAF <1%; Hoffmann & Witte, 2015). There are also two views regarding large proportions of variants that account for majority of MTAs: common variants with small effects, or rare variants with large effects. There are arguments for and against each of these two views (Gibson, 2012). However, except for a trait like human height, no rare variants with large

effects have been reported (Marouli et al., 2017). This is sometimes attributed to the low power of the currently available methods, so that efforts are being made to increase the power of detection of these associations in GWA studies (Korte & Farlow, 2013). It has also been shown that some derived traits (e.g., sum of or ratios between absolute levels of different metabolites) can also provide useful information about the genetics of a metabolic network (Angelovici et al., 2017) (see step (b) in Section 7.2.5). Since the hypothesis of "common disease common variant (CDCV)" has now been rejected, at least for humans, three different models which could explain the genetics of complex traits have also been proposed in different studies (Gibson, 2012). These include (i) the infinitesimal model which suggests that there are large number of common variants each having small effect, (ii) the rare allele model which suggests that there are large number of rare variants, each having large effect or (iii) the broad sense heritability model which suggests that there are some combination of genotypic, environmental and epigenetic interactions. All these models have been explained in detail by Gibson (2012). One would like to know which one or more of these models really holds good. In order to study this, it is necessary to find methods with enhanced power for detection of association between rare alleles and rare variants with targeted traits.

6.2 CVAS and RVAS

Since rare variants may sometimes represent the desirable variation, it has been suggested that common variants association study (CVAS) and rare variants association studies (RVAS) be conducted separately, independent of each other (Zuk et al., 2014). Methods have also been suggested, where analysis of rare variants can be combined with that of common variants, by giving weightage to different variants on the basis of frequency of the marker allele. Therefore, the analysis of rare and low frequency variants explaining missing heritability has become an important area of research (Asimit & Zeggini, 2010; Auer & Lettre, 2015). However, in order to explain this missing heritability, fairly large sample size will be required for identification of such variants (Auer & Lettre, 2015; Lee, Abecasis, Boehnke, & Lin, 2014; Zuk et al., 2014). The appropriate sample size for RVAS depends on the mutation rate, selection coefficient, and the size of the effect for null alleles in the gene (Zuk et al., 2014).

Majority of the issues related to rare variants have been discussed using results of studies carried out in humans. More such studies need to be

undertaken in plants, where whole genome sequences are already available. The above findings also suggest the need for improving the existing methods in terms of their power.

6.3 Sequence-based rare variants GWAS

Sequence-based rare variants GWA studies have also been undertaken using NGS technologies, particularly in humans. This allows direct testing of the association between an individual minor/rare allele and the target trait. In a recent study in humans as part of the UK10K project, it was shown that the increase in sequencing depth also increases power to identify rare variants, and that the power to detect association between rare variants and target trait (occurrence of disease) increases with sample size (The UK10K Consortium, 2015). In this study, 10,000 individuals from population-based and disease-based collections were sequenced at the levels of whole genomes (with read depth of 7 ×) and exomes (with read depth of 80 ×). These data were used for a detailed study involving markers with rare alleles.

Rashkin et al. (2017) studied the power of association tests using different study designs which included sequencing at low and high read depth on varying sample sizes, frequencies of singletons, and relative risks and prevalence of disease. They observed that for a fixed cost, power of detecting association is maximum at a read depth of 15–20 ×, while it decreases with increase in coverage of sequencing beyond this threshold.

However, deep WGS of association panels with large samples is still cost-prohibitive. A variety of approaches have thus been suggested to deal with this issue (Auer & Lettre, 2015; Bansal, Libiger, Torkamani, & Schork, 2010; Lee et al., 2014), some of which include the following: (i) low-depth whole genome sequencing (WGS), (ii) sequencing of exome, (iii) sequencing of targeted-regions, (iv) custom genotyping arrays, and (v) sampling of extreme-phenotypes. All these approaches have been used in human genetic studies and have provided useful information (Auer & Lettre, 2015). In addition, different methods have also been proposed and explained in detail for testing of rare-variant associations (Feng et al., 2015; Lee et al., 2014). These include the following: (i) single-variant test, (ii) gene or region-based aggregation tests of multiple variants, (iii) adaptive burden tests, (iv) variant component tests, (v) omnibus test, (vi) The EC-tests and (vii) meta-analysis. Wei et al. (2016) proposed robust adaptive sum of powered score (aSPUr) method for analysis of rare variants for the non-normal distributed traits and found this method to be more effective than the existing methods like SKAT in controlling type-I error rate.

7. Post-GWAS analysis

With the advances in methodology and statistical tools that became available, meaningful further work in the post-GWAS era can be undertaken either using GWAS summary statistics, or by designing new experiments based on the results of earlier studies. Following approaches are already being used in the post-GWAS era. (i) conditional analysis or, joint multiple SNP analysis or conditional and joint multiple SNP analysis (COJO) at a top individual locus or on whole genome level, using Genome-Wide Complex Trait Joint Analysis (GCTA-COJO); (ii) use of several diverse panels for GWAS; (iii) meta-analysis (within an individual study or involving several studies); (iv) rare alleles/variant analysis (discussed in earlier section); (v) use of associated markers in the coding versus non-coding regions; (vi) annotation of candidate genes/alleles, sometimes using TILLING or Eco-TILLING; (vii) use of non-phenotypic variation [RNA-seq and eQTLs (*cis*-eQTLs and *trans*-eQTLs)], DNA methylation and mQTL, metabolite analysis for GWAS. The approaches for Post-GWAS analyses will be discussed in this section under the following six heads: (i) Identification of causal variants among GWAS signals; (ii) prioritization of the identified GWAS signals; (iii) identification and functional characterization of candidate genes; (iv) gene-based and gene-set based association mapping (GBAM, GSBAM); (v) GWAS using machine learning; (vi) use of high-dimensional data for molecular networking.

7.1 Identification of causal variant

In GWAS, often a trait-associated SNP is not causal, but is simply in LD with the causal SNP. Therefore, identification of causal variant among GWAS signals becomes important. For identification of causal variants following approaches can be utilized: (i) fine mapping, (ii) localization success rate approach, and (iii) conditional analysis.

7.1.1 Fine mapping of GWAS signals

For fine mapping of a GWAS signal, dense genotyping or sequencing of associated genomic region is required. High confidence imputation may also help in generating dense genotyping data (Spain & Barrett, 2015). In order to increase the power of fine mapping that can differentiate among several SNPs in LD, a large population is required. In case of humans, large consortia have been developed by combining datasets of custom-designed genotyping arrays with up to ~200,000 variants. For this purpose,

"Cardio-Metabochip Consortium" focused on diabetes and coronary artery disease, and "Immunochip Consortium" utilized variants selected for immune-mediated phenotypes. These consortia enabled genotyping of all samples on a solitary chip and were thus suited for fine mapping of association signals. In case of plants, such platforms (custom chip for particular phenotype) and collaborations are yet to be developed. However, with the advances in HT marker techniques, it is expected that such consortia in plants will also be developed.

Following dense genotyping, the major challenge is to distinguish true causal SNP from other associated SNPs. A simple approach considers all SNPs with a certain cut-off of *p*-value for causality. This approach is not always suitable, since *p*-value is affected by several factors like power, minor allele frequency and the effect size. It is also to be understood that *p*-values from different studies are not always comparable and can have different implications for the possibility of true association (Stephens & Balding, 2009). Graphical tools like LocusZoom which can provide the extent of an association signal and the position relative to nearby genes in a GWA study can also be useful (Pruim et al., 2010). Bayesian approaches have also been used for the identification of causal SNPs, where Bayes factor is used for calculating the posterior-probability for each variant (Spain & Barrett, 2015). In a recent review, Schaid et al. (2018) described in detail various other approaches for fine mapping, which include the following: (i) Heuristic approach, which involves filtering SNPs based to their pairwise correlation (r^2) with the lead SNP or hierarchical clustering of all SNPs in a region based on their pairwise r^2 values or study of pair-wise LD among SNPs within haplotypes. (ii) Penalized regression approach, which involves joint analysis of all the SNPs in a region using regression model. (iii) Bayesian fine mapping approach, which involves incorporation of prior information in the analysis (see Section 4.1.6.). (iv) Multi-region fine mapping, which involves use of multiple loci. (v) *Trans*-ethnic fine-mapping, which involves combining GWAS results of same trait from genetically diverse populations; this will be equivalent to the use of multiple panels for GWAS for fine mapping. Besides this, they have also discussed the factors which can influence the fine-mapping.

7.1.2 Localization success rate (LSR)

LSR is the probability of the causal SNP being top-ranked within an associated region. Often, LSR can be improved through a use of multiple populations for a joint analysis, rather than a single large population.

Zaitlen, Paşaniuc, Gur, Ziv, and Halperin (2010) proposed an approach, where LSR can be improved without using a very large population and very dense genotyping data. This approach considers following two issues, while conducting analysis: (i) structure of the LD in the population being studied, and (ii) identification of the population(s) achieving an increase in LSR for fine mapping. Zaitlen et al. (2010) observed that studies which involve a set of two or three populations give higher average LSR than those which are based on single population.

7.1.3 Conditional analysis

Statistical methods are now available for conditional analysis, which can identify a causal SNP from among many correlated variants within a LD block or a haplotype. Such conditional analysis may be conducted at the level of an individual locus representing a genomic region or on whole genome level. Sometimes, conditional analysis may also involve a network, where a set of other genes (involved in the network) are selected on the basis of prior knowledge about the biological network, and used for conditional analysis.

(a) Conditional analysis (cGWAS) at a locus or in an LD block

In GWAS, if we have a primary lead SNP (say SNP A) associated with the trait, and also find within the same locus/interval another secondary associated SNP (say SNP B) that is correlated with SNP A, one may like to find out whether or not the association of SNP B is independent of correlated SNP A, although both SNPs are within the same genomic region. In such cases, we may conduct GWAS, after adjusting the model for SNP A. Such analysis can be used as a tool to identify secondary association signals, which are independent of the primary signals. This will involve analysis, conditioning on the primary associated SNP (SNP A in this case) at the locus, to test whether within the same region, there are one or more other SNPs significantly and independently associated with the trait. Many times, it may be possible that there are several genes in the interval of one GWAS locus (including humans and maize) and only one of this may contribute to the identified MTA. However, a follow-up analysis of GWAS loci as well as additional experiments will be required to pin-point the causal genes (Huang & Han, 2014).

(b) Conditional analysis (cGWAS) of whole genome

A conditional analysis is often also conducted at the level of whole genome, using the top associated SNPs, which can be followed by a step-wise procedure to select additional SNPs sequentially based on

their conditional p values. This strategy can be useful for the identification of more than two associated SNPs at each individual locus, and may also allow identification of one or more causal SNPs (Huh, Kwon, & Park, 2015; Yang et al., 2012). This will also help in the identification of haplotypes associated with a particular trait.

(c) Network-based conditional analysis (cGWAS)

A network-based cGWAS has also been suggested, where the impact of genetic variants on more than one omics phenotypes is examined. In such a case, for each trait of interest, a set of other traits are selected (based on biological network), and used as covariates in GWAS. The network could be reconstructed either from biological pathway databases or directly from the GWAS data. Such an approach has been used in humans using metabolomics data (151 metabolites), where it was shown that additional loci (not detected by conventional GWAS) can be detected through this approach (Tsepilov et al., 2018).

(d) Conditional analysis using GWAS summary statistics

In addition to the approaches described above, where original individual-level data are used for conditional GWAS, methods have also been proposed, where conditional analysis can be applied to GWAS summary data. These methods include the following: (i) Genome-wide Complex Trait Analysis—Conditional and Joint Effect Analysis (GCTA-COJO) (Yang et al., 2012); (ii) Sequential Sentinel SNP Regional Association Plots (SSS-RAP) (Zheng, Gaunt, & Day, 2013). (iii) "HAPlotype Regional Association-analysis Program" or HAPRAP (Zheng et al., 2017). Although, each of these methods has been designed to suit a specific situation, it was shown that HAPRAP outperforms other two methods and had an increased power for fine mapping (Zheng et al., 2017).

7.2 Prioritization of GWAS signals

Although hundreds of MTAs for different traits have been identified in different crops using GWAS, not all of them can be used in practical plant breeding program. It is therefore imperative to prioritize the most important loci for their functional characterization as well as for possible use in crop improvement program. Sometimes the terms "prioritization" and "identification of causal variant" are used interchangeably, which may not be desirable. Identification of a causal variant refers to distinction between a true causal variant and an association that arises just due to LD

with causal one (for details see above). The term "prioritization," on the other hand, refers to selection of an associated marker that allows identification and selection of a CG for functional characterization. Different approaches have been proposed for prioritizing the most important loci identified through GWAS. Most of these approaches have been suggested for post-GWA studies in humans, and only limited literature is available on prioritization of GWAS signals in plants. However, with suitable modifications, the approaches used in humans can also be used in plants (Table 4).

As can be seen from the data presented in Table 4, the most important prioritization approaches rely on calculating prioritization score based on *p*-values. Prioritization methods, which do not require calculation of prioritization score are also available and include the following: (i) meta-analysis; (ii) analysis of eQTLs and interacting QTLs (Cantor, Lange, & Sinsheimer, 2010); (iii) study of DNA methylation pattern (Heyn et al., 2013); (iv) application of newer omics approaches; (v) haplotype-based analysis (see Section 7.2.6.); (vi) transcription factor binding sites; (vii) DNase hypersensitive sites; (viii) histone modifications (Edwards, Beesley, French, & Dunning, 2013; Hou & Zhao, 2013), etc. Other approaches have also been suggested in case of human beings, which include Network Interface Miner for Multigenic Interactions (NIMMI) (Akula et al., 2011), use of functionally-coherent sub-networks (Taşan et al., 2015), guilt by association (Lee, Blom, Wang, Shim, & Marcotte, 2011) and network-assisted analysis (Jia & Zhao, 2014). Some of these approaches are also available in the form of tools like GWASrap (Li, Sham, & Wang, 2012) and cepip (Li, Li, et al., 2017). Several bioinformatics pipelines have also been proposed for this purpose in humans (Cheng et al., 2015; Hiersche, Rühle, & Stoll, 2013; Uren et al., 2017; Vaez et al., 2016). However, not all approaches proposed in the context of human system are suitable for plants. The approaches, which have been used or are likely to be used in plant system, are described in this section in relatively greater detail.

7.2.1 Meta-analysis

GWAS meta-analysis should not be confused with meta-QTL analysis that is frequently practiced using the results of multiple QTL interval studies for the same trait, using the same markers. This gives us metaQTLs, which are more robust and reliable for MAS. On the other hand, GWAS-meta-analysis is used for the following two purposes, and may involve one study or several studies involving the same trait: (i) discovery analysis for identification of new MTAs, and (ii) replication analysis for confirming MTAs already

Table 4 List of prioritization approaches used for post-GWAS analysis in humans which can potentially be used in plants with suitable modifications.

Approach	Steps involved	Reference
GIN (*genomic information network*)	• Prioritization score (*S*) is calculated for each SNP which is a cumulative measure of scores derived from pathway information, comparative genomics, linkage scan, and results of other independent GWAS studies. The weights are decided by the strength of the linkage between the SNP and the annotations (closer an SNP with the gene, more will be the score) • SNPs are ranked according to their score for further study • The approach has been implemented in SNP prioritization online tool (SPOT)	Saccone et al. (2010, 2008)
Functional priority of the *p*-values (fpp-value) approach	• Functional priority of the *p*-values (fpp-values) of a selected locus is determined from prepared *p*-values of interactions between (i) SNPs and gene expression patterns, and (ii) genetic loci and gene expression for target phenotypes	Paik et al. (2012)
Prioritization using Bayesian probability	• Important signals are prioritized using genome-wide data, SNP information from bioinformatics databases, empirical SNP weights, and the researchers' subjective prior probabilities • Prior probabilities are combined with GWAS data to calculate posterior probabilities	Thompson et al. (2013)
GPA (Genetic analysis incorporating Pleiotropy and Annotation)	• Integration of GWAS data of multiple genetically related phenotypes and incorporation of required biological information in the analysis to prioritize GWAS results • Testing for the presence of pleiotropy • A modified version of this approach is available in the form of graph-GPA • The basic idea is that genetic basis is not shared only within a phenotype group but also between phenotype groups	Chung et al. (2014) and Chung, Kim, and Zhao (2017)

Table 4 List of prioritization approaches used for post-GWAS analysis in humans which can potentially be used in plants with suitable modifications.—cont'd

Approach	Steps involved	Reference
GenoWAP (Genome Wide Association Prioritizer)	• Integrated analysis of GWAS *p*-values and genomic functional annotation as well as pleiotropic effects • Integrates functional prediction and each SNP is assigned a new score that measures its importance • For prediction of the functional potential of each nucleotide of human, GenoCanyon is used which can distinguish true signals from among highly correlated SNPs. • The method can reduce noises caused due to LD and can also identify marginal signals in studies having insufficient sample sizes	Lu, Yao, Hu, and Zhao (2016) and Lu et al. (2017)

identified. Meta-analysis may also involve prioritization through integration of information of multiple complimentary studies from the same or closely related populations (Cantor et al., 2010; Evangelou & Ioannidis, 2013; Lee, Kim, Choi, Huh, & Park, 2015; Magi & Morris, 2010; Tang & Lin, 2015). GWAS meta-analysis have already been conducted in humans (e.g., height and BMI), but its use in plants has yet to realized. A distinction has also been made between GWAS meta-analysis and mega-analysis.

(a) Meta-analysis and mega-analysis

Meta-analysis is generally conducted using summary-statistics using the data from one study or several studies. If the individual-level data from different studies are also available, one can perform the mega-analysis. In meta-analysis, the results reported in a number of earlier studies, or those obtained by several collaborators are utilized. Sometimes, all results are not reported in a publication, so that mega-analysis involving original data of individual studies may be more rewarding.

(b) Discovery analysis and replication analysis

Based on the purpose of the study, meta-analysis that is conducted following GWAS can be classified in the following two groups: (i) discovery analysis for the discovery of new variants, and (ii) replication analysis for the replication of earlier findings. While the purpose of discovery analyses is to look for new variants across the whole genome, replication analyses usually focuses on a limited number of pre-specified variants. However,

depending on the situation, both these methods can be combined as part of hybrid design wherein, one would first replicate the previous findings and then the results are used to identify new candidate regions. Accordingly, a typical meta-analysis involves the following different stages: (i) processing of the preliminary data to understand between-study heterogeneity; (ii) replication of previous discoveries; (iii) actual discovery of new variants, and (iv) replicating these new discoveries.

(c) Heterogeneity in component studies

Meta-analysis also involves identification of heterogeneity in the size of the effects across different studies that are used for meta-analysis (Lee et al., 2015; Panagiotou, Willer, Hirschhorn, & Ioannidis, 2013). Different factors causing heterogeneity have been discussed (Bush & Moore, 2012; Gogele et al., 2012; Han & Eskin, 2012; Lee et al., 2015). Some of the factors which are relevant in the context of plant systems include the following: (i) use of different study designs in different studies leading to variation in the genetic effects across populations; (ii) differences in the genotyping platforms used and different thresholds for genotyping quality control procedures; (iii) quality of imputation, particularly for the data involving low-frequency variants; (iv) population structure in association panels that are used and (v) publication-bias caused due to incomplete reporting of quality criteria in the individual analysis. An important heterogeneity test is the Cochran's Q test, which can be used to identify if the differences between the primary studies really exist or are due to chance (Whitehead & Whitehead, 1991). Heterogeneity can also be tested using I^2 value, which is independent of the number of studies or the type of output data (Higgins & Thompson, 2002).

(d) Methods of GWAS meta-analysis

The methods that are used for GWAS meta-analysis were reviewed by Evangelou and Ioannidis (2013) and include the following: (i) use of *P*-values accompanied either with or without weights, and (ii) use of fixed/random effects model for combining the size of effects. The *P*-values can be used either by Fisher's *P*-method, where all studies are weighed equally or by *Z*-score method (*Z*-scores are calculated by transforming *P*-values); this will allow attaching different weights to different studies (Cantor et al., 2010; De Bakker et al., 2008). Also, the choice of appropriate model for meta-analysis is dependent upon the presence/absence of heterogeneity, so that a fixed effects model is suited in the absence of heterogeneity, while a random effects model is used in the presence of heterogeneity (Lee et al., 2015). It is known

that in the presence of heterogeneity, interpretation of results is difficult. A new statistic called *M*-value has been suggested for this purpose (Han & Eskin, 2012).

(e) Crop-wise consortia for meta-analysis in crops

Studies involving meta-analysis of GWA data in humans have mainly been benefited by powerful consortia between different research groups as well as advances in the technique of data imputation. Utility of such studies is also seen from the fact that several studies in humans have identified associations that could not be identified in any individual study. However, despite its demonstrated success in human GWAS, such studies have not been conducted in plants. One of the major reasons for this can be the non-availability of common sets of markers across different GWA studies in plants; another reason can be heterogeneity in the results caused due to various reasons mentioned above.

The success of meta-analysis in humans suggests that it will be desirable to have strong consortia for individual crops. This will help in prioritization of associations reported in earlier studies, and can be followed by mega-analyses. Important crop plants like maize, rice and wheat, where major international consortia for genome sequencing already exist can be the starting point for such studies. The advances made in techniques involving data imputation and variety of computer programs enabling meta-analysis in humans can facilitate such studies in plants. Excellent reviews describing different methods of meta-analysis along with issues related to it are already available (Begum, Ghosh, Tseng, & Feingold, 2012; Bush & Moore, 2012; Evangelou & Ioannidis, 2013; Han & Eskin, 2012; Panagiotou et al., 2013). In order to handle the computational challenge involving meta-analysis, large numbers of software packages are also available, which have been discussed and compared by Evangelou and Ioannidis (2013). Grimm et al. (2017) developed a cloud-based platform called easyGWAS, which can be used for meta-analysis, and also for comparison of results from different GWA studies. It can also compute, store, share and annotate GWAS results across different experiments and different species. This is probably the first such interactive resource for GWA studies in plants and will facilitate more such studies in future.

7.2.2 Pathway-based analysis

Pathway-based analysis is a promising post-GWAS approach to understand genetic basis of the trait of interest (Wang, Li, & Bucan, 2007). The approach

focuses on the study of the combined effect of several genes which are grouped based on their shared biological function. Following steps are involved in this analysis: (i) information is collected for all genomic regions and genes associated with the trait of interest, directly or indirectly; (ii) screening of available transcriptome data (this may sometimes involve identification of eQTLs); (iii) grouping of genes on the basis of gene ontology; (iv) analysis of groups of genes to identify the possible pathway that may be associated with the trait of interest. Initially, the approach was based on single or few genes, which were differentially expressed at the highest level; the approach, therefore, was then called gene-based analysis (Liu et al., 2010; Tang, Perkins, Williams, & Warburton, 2015). Later, modifications were made to extract more information regarding genetic architecture of trait of interest, so that detailed pathways involved in the expression of a specific trait could be worked out in some cases.

Pathway-based analyses are routinely used for human genetic studies involving complex diseases (Carlson, Eberle, Kruglyak, & Nickerson, 2004; Kwak & Pan, 2017; Li, Jiao, et al., 2015; Torkamani, Topol, & Schork, 2008; Weng et al., 2011). However, in case of plants, only few such studies have been conducted (Lu, Liu, et al., 2015; Tang et al., 2015). The first such study in plants was conducted in maize and involved identification of defense mechanism against aflatoxin accumulation which is caused due to infection of fungus *Aspergillus flavus* in maize kernel (Tang et al., 2015). It was found that jasmonic acid pathway is the most important pathway associated with aflatoxin resistance. It was also observed that the inbred lines of GWAS panel with desirable alleles of genes involved in jasmonic acid pathway had reduced level of aflatoxin. However, other genes that were not involved in this pathway were also identified, which reduced aflatoxin content and coded for the different gene products.

(a) Phenome-based GWAS (PheGWAS)

Pathway-based approach may sometimes also involve the so-called phenome-based GWAS (PheGWAS; Denny et al., 2010). The approach facilitated identification of causal SNPs involved in regulatory functions, since large number of SNPs identified through GWAS are present in either intronic regions or intergenic regions; this aspect is poorly understood and cannot be studied through conventional GWAS involving genomic data (van der Sijde, Ng, & Fu, 2014).

PheWAS was used in humans to interpret the results of GWA studies for several diseases, based on International Classification of Disease (ICD9) clinic codes (Denny et al., 2010). PheWAS was also used for identification of enzymes and metabolites involved in specific pathways

(Denny et al., 2013). R package is also available for automatic run of PheWAS (Carroll, Bastarache, & Denny, 2014).

A metabolic pathway-based PheGWAS (M-PheGWAS) was also proposed and utilized in rice (Lu, Liu, et al., 2015). In this approach, GWAS is first conducted involving metabolites to identify SNPs associated with metabolites. Identified SNPs were then used to identify corresponding eQTL (using expression QTL analysis and genetical genomics approach). The eQTL analysis was followed by pathway analysis, leading to identification of a specific pathway involved in regulation of trait variation. In this study, results of two earlier studies were utilized, one involving metabolite dataset with 840 distinct metabolites from leaves (Chen et al., 2014) and the other involving eQTL dataset with more than 13,000 eQTLs for over 10,000 e-traits (Wang et al., 2014). The study successfully demonstrated the functional relationships among metabolites, which play an important role in downstream regulation (flavonoids and enzymes, which regulate transcript level).

(b) Topologically association domain (TAD) pathway-based approach

TADs are defined as regions with DNA sequences, which physically contact and interact with each other more frequently than with sequences outside the TAD. Thus, TADs define the boundaries of an interactome and can aid in defining the limits within which an association impacts gene function. These TADs can range in size from thousands to millions of base pairs.

In "TAD pathway-based approach," TAD boundaries are first identified and then gene ontology analysis is performed using all the genes present within TAD boundaries. In many cases, this method identified a gene other than the gene that is nearest to the associated marker identified through GWAS, thus demonstrating its utility. While conducting GWAS for "bone mineral density (BMD)" in humans, Way, Youngstrom, Hankenson, Greene, and Grant (2017) successfully identified "skeletal system development" as the top ranked pathway. This approach helps in developing an understanding of the inter-relationship between causal genes and associated signals present in coding as well as non coding regions.

7.2.3 Methylation QTL (meQTL) in the post-GWAS era

It is widely known that in genomic DNA, 20–40% of cytosine residues in CG islands in humans and CG, CHG and CHH (H = any base other than G) islands in plants are methylated (Gruenbaum, Naveh-Many, Cedar, & Razin, 1981; Messeguer, Ganal, Steffens, & Tanksley, 1991). This methylation

controls the expression of gene, thus leading to changes in phenotypes, both in humans and plant systems (for details, see Kalisz & Purugganan, 2004; King, Amoah, & Kurup, 2010; Long et al., 2011; Suzuki & Bird, 2008). In humans, correlations were also reported between SNPs identified through GWAS and differential DNA methylation, suggesting that methylated SNPs identified through GWAS may regulate gene expression and control phenotype (Heyn et al., 2013). In two recent studies in humans, methylation QTL (meQTL) was identified; in one of these two studies, one-third of the loci associated with schizophrenia identified through GWAS were found to be methylation QTL (meQTL; Jaffe et al., 2016). In the other study, involving asthma, most methylated CpG sites were found to be associated with SNPs manifesting *cis*-effects (Kumar et al., 2016).

In higher plants, changes in DNA methylation due to vernalization and due to treatment with azacytidine (a DNA demethylating agent) was shown to alter phenotype (for a review, see Horvath et al., 2002). For instance, in *Perilla frutescens*, flowering was induced by azacytidine (Kondo, Ozaki, Itoh, Kato, & Takeno, 2006). In model plant Arabidopsis, epiRILs were also developed using mutants for *met1* (DNA methyltranferase) and *ddm1* (decrease in DNA methylation) genes (Johannes et al., 2009; Reinders et al., 2009). *Trans*-generational epigenetic variation involving DNA methylation has also been reported in many plant systems including model plant Arabidopsis (Amoah et al., 2012; Garg, Chevla, Shanker, & Jain, 2015; Johannes et al., 2009; Reinders et al., 2009). A hypomethylated population was also developed in *Brassica* using azacytidine, where mutants for methylation were not available (Amoah et al., 2012).

Whole genome methylation studies have also been conducted in several plant systems including Arabidopsis, wheat, brassica, cotton, rice, etc., where epigenetic variants were shown to affect phenotype (Chen et al., 2015; Gardiner et al., 2015; Hu, Chen, Zhang, & Ding, 2015; Lu, Liu, et al., 2015; Lu, Zhao, et al., 2015; Zhang et al., 2006). However, in case of plants, no study is available where epigenetic analysis is utilized in order to interpret the GWAS results. Integration of GWAS and methylome may provide better insights into genetic architecture of traits.

7.2.4 Prioritization of variants in the non-coding regions

The markers associated with traits of interest may often lie in the non-coding region of the genome, so that such variants in the non-coding region need special treatments. The examples of such variants include e-QTL, miRNAs and lncRNAs.

(a) Expression QTL (eQTL)

It is known that most associated SNPs (identified using GWAS) for the trait of interest are often present in the non-coding region of the genome (Atanasovska, Kumar, Fu, Wijmenga, & Hofker, 2015; Schaid et al., 2018). Therefore, it is possible that some of these SNPs are actually in LD with regulatory regions of structural genes. However, in many cases the gene nearest to the associated SNP is not the causal gene (French et al., 2013; Zhou et al., 2012). For instance, in a study on obesity in humans, it was shown that a causal gene is not the nearest gene (Claussnitzer et al., 2015; Smemo et al., 2014). These associated SNPs often represent the so-called expression QTL or eQTL (Jansen & Nap, 2001), which can be *cis*-QTL or *trans*-eQTL. In actual practice, it has been shown that eQTL analysis is a powerful tool for identification of causal genes (Nica & Dermitzakis, 2013; Zhu et al., 2016). Although most of these studies on eQTL analysis involved use of mapping populations (DH or RIL populations) for interval mapping, there are also reports of GWAS involving eQTL analysis both in human beings (Westra & Franke, 2014) and in plants (Bajaj et al., 2015; Cubillos, Coustham, & Loudet, 2012; Kliebenstein, 2009; West et al., 2007).

GWA studies have also been conducted using transcriptome data leading to the identification of eQTL carrying the associated SNPs. Since most traits are complex in nature, very large populations are needed to detect genetic variants/gene expression. This makes it difficult to phenotype and to perform eQTL GWAS (genetical genomics) in a population with large numbers of genotypes. To overcome this limitation, Zhu et al. (2016) proposed "Summary Data-Based Mendelian Randomization (SMR)" approach to identify causal genes using summarized data of GWAS and eQTL studies, which are available in public domain. This is a useful tool to prioritize genes with known MTAs for functional studies.

While studying the genetics of a complex trait involving expression data along with genotyping and phenotyping data that is routinely used, we may have two possible scenarios: (i) a genetic variant may be an eQTL and controls the expression of another gene [adjoining gene (eQTL) or a distant gene (*trans*-eQTL)], thus indirectly influencing the trait of interest; this is described as causal relationship between the gene and the phenotype (even though it is through expression); (ii) a genetic variant may have direct effect on both, the expression

and the phenotype (this is pleiotropy); alternatively this genetic variant controlling expression may be in close linkage with another gene controlling the phenotype. "Heterogeneity independent instruments" (HEIDI) test is generally used for discriminating between pleiotropy and linkage.

In higher plants, thousands of eQTL have been identified that regulate expression of genes associated with complex traits (Cubillos et al., 2012; Kliebenstein, 2009; West et al., 2007). However, information from only few of these eQTL has been utilized to interpret GWAS (Chan, Rowe, Corwin, Joseph, & Kliebenstein, 2011; Yang et al., 2014). In Arabidopsis, through combination of GWAS and eQTL analysis, Chan et al. (2011) cloned some novel genes involved in glucosinolate (GSL) synthesis. For this purpose, they used 96 Arabidopsis accessions and 230,000 SNPs to conduct GWAS for more than 40 GSL traits. Through co-expression network approach using eQTL, candidate genes (identified through GWAS) were prioritized and then several genes associated with GSL synthesis were characterized and cloned. In another study in maize, out of six CGs associated with agronomic traits identified through GWAS, one was found in the region of eQTL and was prioritized for further study (Yang, Lu, et al., 2014). Such studies need to be extended to other important crops like wheat, rice, etc., to prioritize MTAs identified through GWAS for further functional validation and for their utilization in crop improvement programs.

(b) Integration of GWAS and eQTLs

In order to overcome some of the limitations of traditional GWAS, integration of GWAS with eQTL has been recommended using following two methods: (i) PrediXcan [for individual level data (Gamazon et al., 2015) and for summary data (Torres et al., 2017)] and (ii) transcriptome-wide association study (TWAS) (for individual level and summary statistics data + eQTL; Gusev et al., 2016). By incorporating information on gene regulation from a set of markers, PrediXcan increases power of association analysis over that of traditional GWAS and gene-based tests. On the other hand, TWAS identifies genes whose *cis*-regulated expression is associated with complex traits by integrating gene expression data with that of the GWAS summary statistics. However, under certain common situations, both these methods suffer from loss of power.

Recently, Xu, Wu, Wei, and Pan (2017) has proposed a new test called TWAS-aSPU, which is based on the reformulation of PrediXcan

and TWAS and can integrate single or multiple sets of eQTL data with GWAS individual-level data or summary statistics and has been proven to be more powerful than the original methods. Using lipid GWAS summary data from large number of samples, this new method identified novel trait-associated genes and also showed much improved performance (Xu et al., 2017). R package for this method is also freely available.

(c) Micro-RNA (miRNA)

Since presence of trait-associated SNPs in non-coding region is a rule rather than exception, one would expect the presence of GWAS SNPs in miRNAs regulome also. During the last 5 years, several strategies, bioinformatic tools, online databases have been developed that enabled identification of GWAS signals, which alter miRNA regulome. For instance, in humans, target site of miR196 was shown to be associated with Crohn's disease through GWAS (Georges, 2011). In several other studies in humans, SNPs detected through GWAS were found to be present within the regions of miRNA regulome (Goulart et al., 2015). This allows one to better understand and prioritize the GWAS signals and genetic architecture of complex traits (Bulik-Sullivan, Selitsky, & Sethupathy, 2013; Thomas, Saito, & Saetrom, 2011; Ziebarth, Bhattacharya, Chen, & Cui, 2012). A recent development is the integration of expression data of miRNAs and their targets with data on mRNAs and GWAS in a database to enable user to prioritize and select functional SNPs (Gong et al., 2015). However, only limited efforts have been made in plants to study the involvement of GWAS SNPs in miRNA regulome. Efforts, therefore, are needed to collect all the information related to miRNA and its targets and expression analysis that may be relevant to examine further the results of GWAS in post-GWAS era for a number of traits in a variety of crops. For this purpose, statistical and bioinformatics tools as well as databases are also need to be developed.

(d) Long noncoding RNA (lncRNA)

Since many GWAS SNPs are present in non-coding regions, lncRNAs may also be involved in associations identified through GWAS. In humans GWA studies, association of lncRNA was reported with several diseases including the following: intellectual disorder (D'haene et al., 2016), lung cancer (Yuan et al., 2016) and cardiometabolic diseases (Dechamethakun & Muramastu, 2017). However, in case of plants, no such association of lncRNA has been reported in post-GWAS studies, although there is no reason why such an association of lncRNAs should not be available.

7.2.5 Application of omics approaches

In the post-GWAS era, omics approaches (including genomics, transcriptomics, proteomics and metabolomics) will also be increasingly used in association mapping. Some progress has been made in this area, which will be covered briefly in this section.

(a) Associative transcriptomics and transcriptome based GWAS (TWAS)

RNA-Seq is often used for a study of transcript abundance and for providing functional information such as quantitative variation in expression and indications of epigenetic silencing (Bancroft, 2013). With the advances in NGS techniques, RNA-Seq is becoming increasingly popular. The advantage with this technique is that the complexity associated with sequencing large/polyploid genomes is reduced by sequencing only the transcribed region of the genome (Harper et al., 2012). In addition, it can provide a means to conduct GWAS with a single sequence data set to analyze variation in gene sequences (as SNP markers) and regulatory sequences (as gene expression markers = GEMs) (Bancroft, 2013). This technique of rapid identification of molecular markers associated with trait variation due to gene sequences and regulatory sequences has been described as associative transcriptomics (AT) (Harper et al., 2012). The approach can be used even in species where limited genomic resources are available.

The technique of AT has already been utilized in *Brassica napus* which led to identification of genomic deletions associated with two QTLs for seed glucosinolate content (Lu et al., 2014) and anion homeostasis (Koprivova, Harper, Trick, Bancroft, & Kopriva, 2014). It has also been successfully utilized in wheat for identification of the novel causative genes contributing to stem strength and plant height (Miller et al., 2016) and in European ash (*Fraxinus excelsior*, a tree) for identification of genes providing tolerance against dieback disease (Harper et al., 2016). The technique allows identification of important associations even in a smaller set of genotypes, as against a GWA study, where a larger sample size is needed. The SNPs and GEMs which have been found to be associated with the traits using AT can be converted into user-friendly markers for use in future breeding programs (Miller et al., 2016). The limitation of this technique, however, is that one should have a draft genome sequence scaffolds from a related species in order to establish a hypothetical marker order (Bancroft, 2013).

(b) Metabolite based GWAS (mGWAS)

In recent years, the focus of genetic studies has shifted toward identifying associations with the intermediate or end products of enzymatic

reactions which have a direct effect on the phenotype. These products are called as metabolites and the plant metabolome serves as the link between the "genome and its phenome" (Luo, 2015). The abundance of a specific metabolite in the genotypes of a population can be treated as a trait and can be used to perform metabolite-based GWAS (mGWAS).

This new approach of mGWAS is emerging as one of the powerful alternative genetic strategies to elucidate the genetic and biochemical bases of metabolism in not only model plants but also in important crop plants. The significance of metabolites in plants is also evident from their diversity which is far greater than that found in other organisms (Riedelsheimer et al., 2012).

Metabolite based GWA studies have been carried out using natural populations leading to identification of significant MTAs. Metabolome profiling of plants accompanied with advances in quantitative genetics and genomics can significantly aid in identification of the causal genes controlling natural variation in metabolome profile and the variation in abundance of individual metabolites. This is evident from such studies carried out in maize (Riedelsheimer et al., 2012; Wen et al., 2014), rice (Chen et al., 2014; Matsuda et al., 2015) and tomato (Sauvage et al., 2014; Zhao et al., 2016), which suggested that metabolites serve as the vital links between the genotype and the phenotype. Wen et al. (2014) combined the mGWAS approach with the expression profiling (RNA-seq) data for 983 metabolite features and identified important MTAs for kernel weight in maize. Following the re-sequencing and CG based analysis, they identified potential causal variants for five CGs involved in metabolic traits. Similarly, in a recent study, Zhang, Warburton, et al. (2016) identified, genetic determinants of metabolic response to drought stress using a high density SNP set in maize. The important candidate loci identified in such studies, if validated functionally, can yield important insights about the genetics of these traits. More such studies are needed to be carried out on large scale in other crop plants to harness the potential of this approach.

(c) Network-guided GWAS

Although, in many studies mentioned above, absolute values of metabolites have been used as traits in GWAS (mGWAS), it is now known that derived traits which are generated from absolute values of metabolites can offer unique understanding about the metabolic network and can be used as a trait for genetic dissection of a trait like free amino acids (FAAs) (Angelovici et al., 2017). These derived traits can either be the ratio of two related metabolites or the sum of related

metabolites. In many GWA studies, it has also been shown that these derived traits have exhibited more significant associations over those obtained using absolute levels of metabolites (Angelovici et al., 2013; Gonzalez-Jorge et al., 2013; Lipka et al., 2013; Owens et al., 2014). One such study in Arabidopsis involved use of traits like branched-chain amino acid (BCAA) and other related seed traits, including ratio of isoleucine (Ile) either with the amino acids of the BCAA family or with the total amino acids (Angelovici et al., 2013). This study identified *BCAT2* (*BRANCHED-CHAIN AMINO ACID TRANSFERASES*) as the key-locus associated with variation in seed BCCA. Similarly, Lipka et al. (2013) observed highly significant association with a known tocochromanol biosynthesis gene for the ratio of δ-tocotrienol to the sum of γ- and α-tocotrienols in maize grain as compared to that of the absolute levels of tocotrienol. In another study in maize, utilizing a NAM panel, Richter et al. (2016) identified a new cytochrome P450 gene for the ratio of two homoterpenes using a joint-linkage-assisted GWAS. Using this network-guided approach involving a panel of 313-ecotypes of Arabidopsis, Angelovici et al. (2017) performed GWAS for 98 traits which were derived from known metabolic pathways of amino acids. The results were compared with those obtained from 92 traits which were generated from the analysis of an unbiased correlation-based metabolic network. It was found that the latter approach was superior over the former approach as additional novel metabolic interactions as well as SNP-trait associations for FAAs were identified with this approach. This shows the potential application of network-based approach in elucidating the genetic basis of a complex metabolic network.

7.2.6 Haplotype-based analysis

Prioritization of GWAS signals can also be achieved through the use of haplotype-based analysis. It is well known that the patterns of variation present in the genomes are inherited as blocks. Therefore, it will be more appropriate to cluster the markers into haplotypes for analysis. This ultimately can improve statistical power and can identify novel associations (N'Diaye et al., 2017). There are several advantages with this approach. For instance, one can use information on multiple markers simultaneously, which leads to increased power of the analysis. It has also been observed that the haplotype approach leads to an increase in the PVE (up to 50%) as well as the allelic effect (up to 34%). Both, simulation based as well as empirical studies

have shown that the power of QTL detection and the mapping accuracy can be improved by grouping of markers into haplotype blocks (Barrero, Bellgard, & Zhang, 2011; Calus et al., 2009; Hamblin & Jannink, 2011; Hao et al., 2012; Lipka et al., 2013; Lorenz, Hamblin, & Jannink, 2010; Lu et al., 2012; N'Diaye et al., 2017).

By haplotyping the entire genome, tag-SNPs can also be identified representing haplotype blocks that are used in genetic studies, because the haplotype blocks can describe common patterns of genetic variation. This reduces the expense and time spent on GWAS, since it eliminates the need to study every individual SNP. Several GWA studies have demonstrated the importance of this approach in identification of associations with the traits. Recently, this approach has been used for grain quality traits in wheat and it was observed that by combining multiple SNPs into haplotype blocks the average PIC increased from 0.27 per SNP to 0.50 per haplotype. In rice also, using a set of 258 genotypes, when haplotype-based GWAS was conducted, this allowed identification of a number of important QTL and CGs for cooking characteristics and protein content (Wang, Pang, et al., 2017). In future, with the advances in the HTP techniques and availability of large numbers of SNP markers, haplotype-based GWAS will be the promising technique (for a review, see Gupta et al., 2014).

7.3 Functional characterization of candidate genes (CGs)

Identification of causal markers and prioritization of associated markers should generally be followed by identification and functional characterization of candidate genes through bioinformatics analysis. But for validation and functional characterization, reverse genetics technologies are often used, where the effect of variations/alterations in a gene on phenotype is examined (Heikoff, Till, & Comai, 2004).

7.3.1 Targeting induced local lesions in genomes (TILLING)

Validation and functional characterization of CGs identified using GWAS is often achieved using TILLING, where a mutagenized population is examined for the presence of a rare allele of a CG and its effect on the phenotype (Heikoff et al., 2004; Stemple, 2004). A modified approach is called Eco-TILLING, where a germplasm collection (e.g., association panel) or a natural population is used for allele mining. This will allow validation of the function of a CG without the need for producing transgenic plants for this purpose.

7.3.2 Insertional mutagenesis, VIGS and RNAi

Insertional mutagenesis involving transposon tagging is another approach to validate gene function of a CG identified using GWAS (Kim et al., 2004; Kuromori, Takahashi, Kondou, Shinozaki, & Matsui, 2009). In one such study in Arabidopsis, GWAS was used to identify new effector genes responsible for accumulation of proline under drought condition which were subsequently validated using reverse genetic approach with the help of T-DNA insertion mutants (Verslues, Lasky, Juenger, Liu, & Kumar, 2014).

Post-transcriptional gene silencing involving RNAi and virus-induced gene silencing (VIGS) and promoter-mediated overexpression/misexpression approaches are also widely used for functional characterization of genes in plant systems like wheat, cotton, etc. (Barro et al., 2016; Czarnecki et al., 2016; Lu et al., 2003; Waterhouse, Wang, & Lough, 2001; Younis, Siddique, Kim, & Lim, 2014). VIGS has been used in plant systems including tomato (Fantini, Falcone, Frusciante, Giliberto, & Giuliano, 2013), tobacco (Senthil-Kumar & Mysore, 2014), Arabidopsis (Manhaes, de Oliveira, & Shan, 2015), soybean (Zhang, Whitham, & Hill, 2013), barley (Yuan et al., 2011), wheat (Scofield & Brandt, 2012), maize (Mei, Zhang, Kernodle, Hill, & Whitham, 2016), rice (Kant, Sharma, & Dasgupta, 2015), etc.

7.3.3 Genome editing and base editing

In recent years, the genome editing techniques (also called targeted mutagenesis), including those involving clustered regulatory interspaced short palindromic repeats (CRISPR/Cas), zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs), have been developed for functional characterization of genes, and have been widely discussed in published literature. CRISPR/Cas has become the most popular approach for gene targeting, and has already been utilized in a number of plant systems including Arabidopsis (Feng et al., 2013; Mao et al., 2013; Osakabe et al., 2016), tobacco (Gao et al., 2015; Nekrasov, Staskawicz, Weigel, Jones, & Kamoun, 2013), rice (Mao et al., 2013; Miao et al., 2013; Shan et al., 2013; Wang, Wang, et al., 2016), wheat (Wang, Cheng et al., 2014; Zhang, Lian, et al., 2016) and tomato (Pan et al., 2016). A modified CRISPR/Cas approach termed base editing has also been developed recently (Komor et al., 2016), and will be increasingly used in future, where it will be possible to validate even an individual causal SNP. Appropriate bioinformatic tools have also been developed to select for the best possible CRISPR/Cas target sites (Belhaj, Chaparro-Garcia, Kamoun, Parton, & Nekrasov, 2015).

In all the above approaches, the first generation plants (M_1 or T_0) are usually heterozygous for the modified genomic region. Therefore, to validate their phenotypic effect, one or more generations will be required for obtaining homozygous recessives. In order to overcome this limitation, Shen, Pan, and Lubberstedt (2015) proposed haploid scheme of mutagenesis for functional validation of a single mutation. This approach has the advantage of reducing cycle time either through pollen mutagenesis or through pollen culture.

7.4 Gene-based and gene-set based association mapping (GBAM, GSBAM) for quantitative traits

With the availability of NGS technology in recent years, millions of SNPs with large populations could be used for GWAS, so that in each study, one had to face "multiple testing" and "large *p* small *n*" problems. Haplotype-based association mapping was proposed and successfully used to only partially overcome these problems. Therefore, gene-based association mapping (GBAM) and gene-set based association mapping (GSBAM) have been recently proposed and used to overcome the limitations of single SNP-based and haplotype-based analysis.

GBAM or GSBAM makes use of variants within each CG or gene-set (gene or gene-set is used as a unit of test), which is generally identified using the results of GWAS based on SNPs or haplotypes. These approaches have increased power, and can also be used for the study of the biological processes involved in any complex trait. GBAM was first proposed by Neale and Sham (2004) and has the following advantages, which are also shared by GSBAM. *First*, it increases the power by using all the associated SNPs within a gene together, because a gene for a trait may contain multiple independent causal variants. *Second*, it may increase the power of GWAS by minimizing the multiple testing problem, since instead of testing several million SNPs only thousands of genes or hundreds of gene sets are tested. *Third*, it can also deal with the problem of allelic heterogeneity by using multiple SNPs within a gene together; this results in more consistent results across studies. *Fourth*, it can provide greater insight into the biology of the trait, since the genes are basic functional units of the genome, and *finally*, it can be readily extended to pathway or network-based analysis of GWAS (see Section 7.2.2). In view of these merits, GBAM and GSBAM will be increasingly utilized in future, when GWAS summary results become publicly available for a number of traits in each important crop.

Because of the apparent similarity between GBAM and GSBAM, similar chi-square tests are used for both. These tests can be classified into two major types (univariate and multivariate tests): (i) The univariate tests involve chi-square tests, each involving a set of single SNPs which are then combined. The best approach for combining the results of individual tests is to first obtain the minimum *p*-values as measures of significance and utilize these values in Fisher's method for combining chi-square-based *p*-values. (ii) The multivariate chi-square tests involve testing all the variants in a gene together, which increases the power for prediction of an unobserved causal variant. A number of methods have been suggested and reviewed for this purpose (Wang, Huang, et al., 2017) and are given in Table 5. Some of these methods have already been referred in earlier sections. A comparison between different such methods proposed for analysis of rare variants has been made by Moutsianas et al. (2015). The steps involved in performing gene-set analysis using GWAS data have also been described in published literature (Mooney & Wilmot, 2015).

Entropy-based statistics have also been designed for GBAM and GSBAM, and have been found to have higher power than a chi-square based statistic. The basic purpose of using entropy is to amplify the difference in allele frequencies between contrasting phenotypes, so that even if this difference is minute and difficult to detect, entropy will allow its detection by amplification of this difference. Non-linear transformations are used to amplify the difference in allele (or haplotype) frequencies between contrasting phenotypes.

7.5 GWAS using machine learning

The most common methods of GWAS discussed above either deal with analysis of single-locus at a time or multiple-loci at a time. However, in both these cases, the methods involved in analysis rely on making certain assumptions and building a mathematical model for analysis. This in turn increases the computational demand. Moreover, the predictive power of the analysis is inversely proportional to the number of assumptions made.

The recent progress in genotyping techniques, however, enabled collection of enormous and high dimensional data. These data are being regularly used for GWAS. The methods to deal with this volume of data are also being developed regularly. In this wave of big data, methods which involve minimal human efforts in making assumptions and which can learn from data without relying on rules-based programming are therefore needed in the post-GWAS era for rapid analysis of the data and for increasing the predictive power of the analysis.

Table 5 Different tests available for gene-based association mapping and gene-set based ssociation mapping.

SN	Test	Features	Reference
1	Smallest *p*-value method	The smallest *P*-value over all SNPs within a gene is used as an overall gene-based *P*-value	Wang et al. (2007)
2	Weighted Fourier transform Test	Multivariate test; reduces degrees of freedom by transforming genotyping by weighted Fourier transformation	Wang and Elston (2007)
3	Fisher's combination of *p*-values test	Combines results from several independent tests using the same null hypothesis; particularly suitable, when there are multiple independent causal SNPs within a gene	Curtis, Vine, and Knight (2008)
4	Versatile Gene-based Association Study (VEGAS)	Based on a multivariate distribution and computes a gene-level *p*-value without requiring raw genotype data; does not require permutation analysis	Liu et al. (2010)
5	Gene-based Association Test using Extended Simes (GATES)	Calculates gene-level *p*-value without using permutation or simulation; not suitable when multiple independent causal variants are present in a gene	Li, Gui, Kwan, and Sham (2011)
6	PrediXcan	Incorporates information on regulation of gene using a set of markers; has increased power than SNP-based GWAS and known gene-based tests; suitable for analysis of transcriptome data	Gamazon et al. (2015)
7	Multivariate Gene-based Association test by extended Simes (MGAS)	Allows efficient testing of multivariate phenotypes in unrelated individuals; available as KGG v3.0	Van der Sluis et al. (2015)
8	Flexible and Adaptive test for Gene Sets (FLAGS)	Takes into account the unique association patterns of gene sets; Suitable for GSBAM	Huang et al. (2016)
9	Combined gene-Based Association Test (COMBAT)	Requires only SNP-level *p*-values and correlations between SNPs from ancestry-matched samples for analysis; makes use of the strengths of several gene-based tests listed above; superior performance over many other tests	Wang, Huang, et al. (2017)

Machine-learning (ML) methods make use of iterations, where computer tries to find out the patterns hidden in data which is then subsequently used to predict future data (Murphy, 2012). ML algorithms provide several alternatives to perform multi-SNP analyses. One such method called "penalized regression method" extends technique of standard regression for analysis of correlated variables (Szymczak et al., 2009). ML approaches are also shown to be useful in handling small n, large p problem, as well as LD structure between the SNPs resulting in correlated variables (Szymczak et al., 2009).

Machine learning can be divided into two types: supervised learning and unsupervised learning (Murphy, 2012; Tarca, Carey, Chen, Romero, & Draghici, 2007). In supervised learning, the aim is to establish a link from input to output given a dataset consisting of labeled pairs of input and output. In unsupervised learning there are no pairs of data, but only have data in which we attempt to find some kind of structure; this would, however, not be applicable to this setting. Both these approaches can be used in the Post-GWAS setup in either protein function prediction or constructing gene networks from gene expression data (Caragea & Honavar, 2009). ML models can also deal with genetic interactions, which will not be possible in single-locus association studies (Okser et al., 2014).

In recent years, the use of ML approaches is becoming common in the analysis involving GWAS and GS. In one such application, a method called COMBI, combining ML and statistical testing was proposed which takes into account the correlations within the set of SNPs used in the study (Mieth et al., 2016). It is a two-step process in which a support vector machine is trained first and a subset of candidate SNPs is identified followed by a hypothesis test along with appropriate threshold correction. Similarly, a software GenAMap has been developed making use of ML approach and can detect different associations among genotypes, gene expression data, and clinical or other macroscopic traits in addition to providing many other features (Xing et al., 2014). PILGRM (Platform for Interactive Learning by Genomics Results Mining) is another platform which uses ML methods and is being utilized for genetic analysis and allows its users to utilize their own knowledge and genome wide data for designing the future experiments (Greene & Troyanskaya, 2011). It contains a compendium of gene expression data which can be used by the researcher based on their research needs. Recently, this platform was successfully used in *Drosophila* and genes involved in learning and memory were identified (Kacsoh, Greene, & Bosco, 2017). There is no doubt

that ML approaches are becoming very popular for efficient analysis and making sense of the ever-growing databases of biomedical nature (Szymczak et al., 2009) and will prove important in the post-GWAS era. Sufficient literature is now available on this aspect, but a more detailed treatment of this subject is beyond the scope of this review. However, as with any other approach, caution should be exercised when judging the superiority of some ML approaches over other methods (Tarca et al., 2007). Several methods employing ML are available, which have been described and compared (Szymczak et al., 2009).

7.6 Use of high-dimensional data for molecular networking in the post-GWAS era

A complex trait is often a manifestation of a large number of related phenotypes, which are not independent of each other. Thus, identification of causal genetic variations and understanding the mechanisms underlying such complex traits requires a joint analysis of different interactions (including pleiotropic, epistatic and plastic) and integration of information of different omic data (Xing et al., 2014). This will require use of high-dimensional data, which is being generated and used for molecular networking in the post-GWAS era.

7.6.1 High-dimensional multi-omics data

It is also known now that a large number of SNPs associated with a complex trait are likely to be eQTLs. Therefore, it is necessary to include gene expressions and/or phenotypic traits as association responses. A major approach, which will receive major attention in the post-GWAS era, includes the analysis of complex traits using high dimensional structured phenotypes including multi-omics data to construct the molecular networks (Runcie & Mukherjee, 2013). For this purpose, in the post-GWAS era, following different data are now often available for more detailed association mapping: (i) genome-wide SNP genotyping data; (ii) NGS based whole-genome transcription data,; (iii) expression data for individual genes using microarrays (used for eQTL analysis); (iv) chromatin immunoprecipitation (ChIP) sequencing data (used for epigenetics) and (v) imaging data involving phenomics. Among these, the use of data from different areas including genotyping, transcriptomics (including gene expression data and eQTL analysis), metabolomics and phenomics has already been discussed earlier. However, the success of the studies focusing on the individual omic level data will remain restricted if one set of data is used in isolation. Moreover, these single omic layer analyses do

not directly explain interaction across multiple omic layers. It has now become possible to integrate data at different molecular levels so that the biological processes associated with them can be described in detail (Carreno-Quintero, Bouwmeester, & Keurentjes, 2013).

Several methods have been proposed to conduct GWAS using high dimensional molecular data like transcriptome, metabolome, etc. The simplest approach is to conduct association between each pair and then apply a stringent cut-off, but this approach requires very large population size to capture minor effects. Sparse regression model (Kim & Xing, 2009) and latent variable regression model have been proposed for GWAS using high dimensional phenotypes and successfully utilized in case of humans (Fusi, Stegle, & Lawrence, 2012; Stegle, Parts, Durbin, & Winn, 2010). Canonical correlation analysis (CCA) is another statistical technique to deal with multivariate data in GWAS (Ferreira & Purcell, 2009). In CCA, association is tested between two groups of variables rather than testing each pair of variable. In case, where data set is large enough and cannot fit in CCA model, sparse CCA can be utilized (Parkhomenko, Tritchler, & Beyene, 2009; Witten & Tibshirani, 2009). These different methods have also been discussed and compared (Marttinen, Gillberg, Havulinna, Corander, & Kaski, 2013).

7.6.2 Networking of genes

Earlier in this review, we described the identification of causal variants and prioritization as essential steps in the analysis of the results of GWAS in the post-GWAS era. However, the knowledge of one or more causal variants and their prioritization is not enough, because it does not tell us about the mechanism involved, which connects the gene(s) associated with the causal variants for the trait. A complex trait is always controlled by intricate interactions between large numbers of genes, which constitute a network. Therefore, following conventional GWAS, one can't stop at the identification of a causal variant and its prioritization, and will have to understand the intricate interactions between different genes involved. The causal variants for a trait need to be related to other molecular states in a cell in terms of RNA transcripts, proteome and metabolome data, to construct a network, which determines the status of the trait. Construction of such a network that is involved in the expression of the trait is thus becoming an important area of research in the post-GWAS era. This amounts to using systems biology approach to define physiological state of the system, which seems to be possible in view of the availability of complete information about variation in the DNA, RNA, metabolite and protein. This network provides a direct

link between causal variants and crop improvement programs by connecting molecular biology to physiology at the cellular and organismal levels. Such large-scale data is also becoming available in many cases, which will facilitate not only the construction of above molecular networks, but also for causal association of such networks with the physiological states for the trait of interest. Several studies and reviews are available, which address this subject.

7.6.3 Trans-*ome wide association study* (trans-*OWAS*)

Recently, a combination of omics approaches has been proposed and termed as *trans*-omics to utilize the benefits of all the omic approaches simultaneously and can lead to *trans*-ome-wide association study (*trans*-OWAS) (Yugi, Kubota, Hatano, & Kuroda, 2016). Such multi-omic approaches make sense as they will allow us to understand the molecular mechanisms that underlie genotype-phenotype relationships (Li, Pearl, & Jackson, 2015). The *trans*-OWAS approach proposed by Yugi et al. (2016) includes information from all the omic layers and aims at identification of the molecular mechanism involved in multifactorial diseases. This approach involves reconstruction of individual networks from the multiple omic data which is then used to characterize the phenotypes (Yugi et al., 2016). The important advantage of this approach over GWAS is that it can associate phenotypes with genetic as well as environmental factors (Yugi et al., 2016). Although at the conceptual stage, a few *trans*-OWAS studies have already been carried out in humans, but such studies are yet to be carried out in plants. However, the available individual omics data in crop plants if integrated/networked carefully, it is expected that in future, such studies will become common in plants. Model plants like Arabidopsis and maize can be the starting point, where extensive omic resources are already available.

7.6.4 *Structured association mapping (SAM)*

While using the multivariate omic data, one of the limitations, that is often encountered during traditional GWA studies, is that they ignore the structural information contained in each set of data; another limitation is imposed due to multiple testing (Marttinen et al., 2013; Xing et al., 2014). This information involving multivariate omic data is valuable in boosting the statistical power of GWA mapping and should not be ignored. Analysis of such data requires development of new algorithms, such as structured association mapping (SAM) algorithms. In the post-GWAS era, one will have to conduct GWAS with modern statistical and ML technologies and will therefore require software and algorithms for this purpose which would also allow SAM.

Recently, a suit of program called GenAMap (discussed earlier) has been described which can facilitate SAM (Xing et al., 2014). The utility of this software has also been demonstrated using yeast and mice data (Xing et al., 2014). More such software need to be developed which can address the issue of high-dimensional data in case of plants.

8. Popular resources available for GWAS in plants

Although large numbers of GWA studies have been carried out for a variety of traits in different plant species and are continuously increasing, there is no common platform where one can access all these studies. This is in contrast to the studies carried out in humans where almost every study is well cataloged in the National Human Genome Research Institute catalog. Cataloging of studies is important because using such a catalog, one can get the information about the most significant associations for different traits and plan future studies. In plants, although, such comprehensive catalog is not available, but in model plant species *Arabidopsis thaliana* and in important crop plants maize and rice, efforts have been made to develop user-friendly platforms, which can facilitate GWA studies in these species. Details of these resources are given in Table 6.

9. Post-GWAS results for crop improvement

Interval mapping and GWAS are two major approaches, which have been extensively used to determine MTAs for a variety of traits in all major crops. Millions of SNPs have been utilized in such studies and thousands of MTAs have already been identified. These MTAs are being utilized for marker-assisted selection (MAS), without having any knowledge about the associated gene(s) that may be involved, and without any causal relationship between the associated marker and the trait or the gene controlling the trait. In this connection, interval mapping (IM) proved to be more useful than GWAS, although the latter provides higher resolution, albeit with no knowledge about linkage relationships. However, recently, in the post-GWAS era, efforts are being made to find out the causal SNPs from among a number of correlated MTAs, which are detected in GWAS. Using these causal SNPs, the CGs and the pathways leading to the phenotype may be worked out (Fig. 1). Such studies will prove useful not only in

Table 6 Popular resources available for GWAS in plants.

SN	Plant species	Resource	Important features	Reference
1	*Arabidopsis thaliana*	GWAPP	An interactive web-based application to perform GWAS in *A. thaliana*; performs GWAS using linear mixed-model	Seren et al. (2012)
2	*Oryza sativa*	Rice Diversity	A high-resolution, open-access platform for facilitating GWAS in rice; provides collection of diverse germplasm, a high-density SNP data set, well-documented analytical strategies, and a suite of bioinformatics resources	McCouch et al. (2016)
3	*Zea mays*	MODEM	A comprehensive database that contains multidimensional omics data for maize; data on 508 diverse inbred lines are available which can facilitate genetic mapping; it can be linked with other databases	Liu, Wang, et al. (2016)

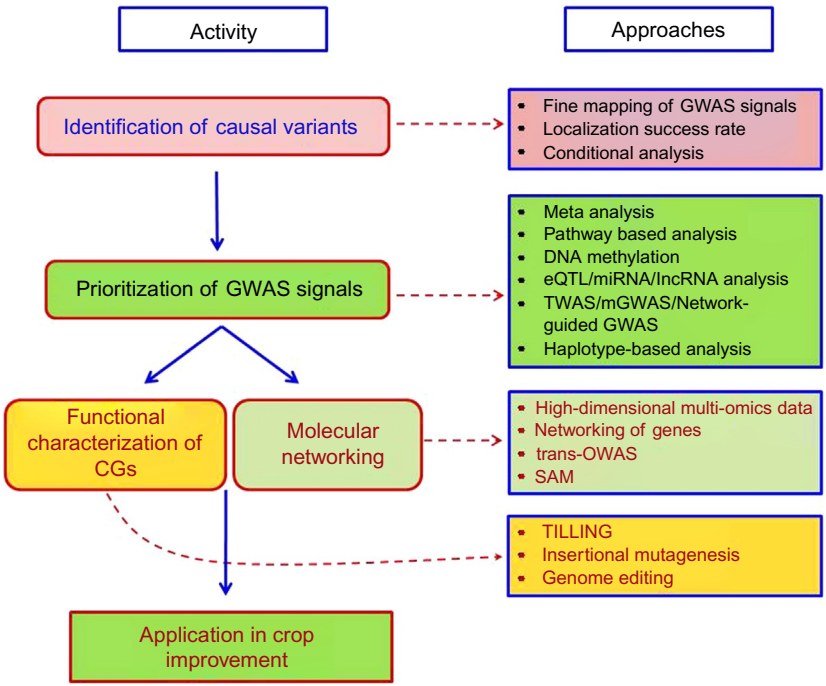

Fig. 1 A flow chart showing the activities, which can be carried out in the post-GWAS era (for details, see text).

understanding the genetic architecture of the trait, but also for manipulation of the genes for improvement of the target trait in a specified crop.

In the post-GWAS era, an effort is also being made to provide interpretations for the results of GWA studies that are now being conducted. The first step in such post-GWA study is the determination of the causal SNPs and the associated CGs including genes encoding transcription factors. Also, often these SNPs may be present in the non-coding regions of the genome, often described as regulome that includes not only the promoter sequences, but also the sequences that are transcribed into miRNAs, lncRNAs and other ncRNAs.

The results of GWAS are also being utilized for Phenome-Wide Association Study (PheWAS). For instance, taking metabolite profiles (involving 840 metabolites) as phenotypic traits, and 6.4 million SNPs for genotyping, a metabolic genome wide association study was conducted in rice (Chen et al., 2014). A modified approach called metabolite pathway-based Phe-WAS (M-PheWAS) was also conducted in rice (Lu, Liu, et al., 2015). A similar study for oil biosynthesis in maize was conducted by Li et al. (2013).

10. Conclusions and perspective

During the last decade, significant advances have been made in the approaches for QTL interval mapping (IM) and GWAS/association mapping for identification of marker-trait associations (MTAs). These MTAs have also been effectively used for marker-assisted selection (MAS) to supplement conventional plant breeding for improvement of simple as well as complex quantitative traits. Association mapping was initially improved through linear mixed model (Q + K model; Yu et al., 2006), which was variously modified not only to reduce the computation demand, but also to deal with the problems of "multiple testing" and "large p small n" that continue to plague GWA studies (Widmer et al., 2014).

It is widely known that GWA studies identify only genomic regions associated with the target trait, and do not discover genes. It is also known that there would be many markers (e.g., SNPs) within a trait-associated genomic region; a number of these markers show association with the trait due to LD with one or few of markers that are causative; one would like to identify these causal SNPs. Several methods are now available for identification of these causal SNPs. Approaches like conditional analysis, meta-analysis and several other approaches discussed in this review can be used in utilizing the results of earlier GWAS and help for the identification of causal SNPs.

Approaches also became available for prioritization of MTAs, so that better MTAs become available for marker assisted selection (MAS) and marker assisted recurrent selection (MARS).

With an enormous increase in the number of markers (millions of SNPs for an individual study) that are now available for each GWA study due to NGS technology, SNP-sets and haplotypes instead of individual SNPs are being used for GWAS. This partly facilitated in overcoming the problem of multiple testing. The identification of causal SNPs and SNP-set/haplotype-based studies also allowed identification of candidate genes (CGs), thus facilitating gene-based association mapping (GBAM) and gene-set based association mapping (GSBAM). A variety of methods were developed for conducting these GBAM and GSBAM studies, thus increasing the power of GWAS. Availability of multi-omics data and their integrated use is also being recommended for understanding the molecular mechanisms that underlie genotype-phenotype relationships. Moreover, with the growing interest in the machine learning techniques, the analysis of multi-dimensional data will be easier in the post-GWAS era. With these advances and the knowledge generated by these newer approaches, it has also become possible to develop the networks, which may be involved in the expression of phenotypes of individual traits.

The major advances in GWAS in the post-GWAS era became possible due to constant development of newer approaches to deal with the limitations of GWAS both in terms of increasing the power of GWAS and also to conduct basic studies to understand the molecular mechanism underlying a trait of interest. Thousands of MTAs have already been identified involving different traits of interest in each of a number of crops. In the post-GWAS era, these results of GWAS are already being utilized for further analysis to provide meaningful results about the key elements governing these traits. In this review, we have tried to discuss different approaches along with their potential applications for analysis of the results of earlier GWAS in the post-GWAS era. Majority of these advances have been possible while working with human system, and their use in plant system is yet to be fully realized.

Acknowledgments

The manuscript was written when P.K.G. was holding the positions of INSA Senior Scientist (2016), and INSA Hony Scientist (2017–2018). P.K.G. would also like to thank Head, Department of Genetics and Plant Breeding, CCS University Meerut for providing the facilities. P.L.K. would like to thank Ministry of Agriculture and Farmers Welfare, Government of India for research grant during this period. V.J. would like to acknowledge DST, New Delhi for INSPIRE faculty award.

References

Akula, N., Baranova, A., Seto, D., Solka, J., Nalls, M. A., Singleton, A., et al. (2011). A network-based approach to prioritize results from genome-wide association studies. *PLoS One*, *6*, e24220.

Al-Tamimi, N., Brien, C., Oakey, H., Berger, B., Saade, S., Ho, Y. S., et al. (2016). Salinity tolerance loci revealed in rice using high-throughput non-invasive phenotyping. *Nature Communications*, *7*, 13342.

Amoah, S., Kurup, S., Lopez, C. M. R., Welham, S. J., Powers, S. J., Hopkins, C. J., et al. (2012). A hypomethylated population of *Brassica rapa* for forward and reverse epigenetics. *BMC Plant Biology*, *12*, 193.

Angelovici, R., Batushansky, A., Deason, N., Gonzalez-Jorge, S., Gore, M. A., Fait, A., et al. (2017). Network-guided GWAS improves identification of genes affecting free amino acids. *Plant Physiology*, *21*, 01287.

Angelovici, R., Lipka, A. E., Deason, N., Gonzalez-Jorge, S., Lin, H., Cepela, J., et al. (2013). Genome-wide analysis of branched-chain amino acid levels in *Arabidopsis* seeds. *Plant Cell*, *25*, 4827–4843.

Arruda, M. P., Lipka, A. E., Brown, P. J., Krill, A. M., Thurber, C., Brown-Guedira, G., et al. (2016). Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum* L.). *Molecular Breeding*, *36*, 84.

Asimit, J., & Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annual Review of Genetics*, *44*, 293–308.

Atanasovska, B., Kumar, V., Fu, J., Wijmenga, C., & Hofker, M. H. (2015). GWAS as a driver of gene discovery in cardiometabolic diseases. *Trends in Endocrinology and Metabolism: TEM*, *26*, 722–732.

Auer, P. L., & Lettre, G. (2015). Rare variant association studies: Considerations, challenges and opportunities. *Genome Medicine*, *7*, 16.

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.

Ayers, K. L., & Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, *34*, 879–891.

Azencott, C. A., Grimm, D., Sugiyama, M., Kawahara, Y., & Borgwardt, K. M. (2013). Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, *29*, i171–i179.

Bajaj, D., Saxena, M. S., Kujur, A., Das, S., Badoni, S., Tripathi, S., et al. (2015). Genome-wide conserved non-coding microsatellite (CNMS) marker-based integrative genetical genomics for quantitative dissection of seed weight in chickpea. *Journal of Experimental Botany*, *66*, 1271–1290.

Ball, R. D. (2013). Designing a GWAS: Power, sample size, and data structure. In *Genome-wide association studies and genomic prediction* (pp. 37–98). Totowa, NJ: Humana Press.

Bancroft, I. (2013). *Association genetics and more from crop transcriptome sequences*. Information Systems for Biotechnology News Report, January1–4.

Bansal, V., Libiger, O., Torkamani, A., & Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, *11*, 773–785.

Barrero, R. A., Bellgard, M., & Zhang, X. (2011). Diverse approaches to achieving grain yield in wheat. *Functional & Integrative Genomics*, *11*, 37–48.

Barro, F., Iehisa, J., Giménez, M. J., García-Molina, M. D., Ozuna, C. V., Comino, I., et al. (2016). Targeting of prolamins by RNAi in bread wheat: Effectiveness of seven silencing-fragment combinations for obtaining lines devoid of coeliac disease epitopes from highly immunogenic gliadins. *Plant Biotechnology Journal*, *14*, 986–996.

Begum, F., Ghosh, D., Tseng, G. C., & Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Research*, *40*, 3777–3784.

Belhaj, K., Chaparro-Garcia, A., Kamoun, S., Parton, N. J., & Nekrasov, V. (2015). Editing plant genomes with CRISPR/Cas. *Current Opinion in Biotechnology*, *32*, 76–84.

Bian, Y., Yang, Q., Balint-Kurti, P. J., Wisser, R. J., & Holland, J. B. (2014). Limits on the reproducibility of marker associations with southern leaf blight resistance in the maize nested association mapping population. *BMC Genomics*, *15*, 1068.

Bouchet, S., Olatoye, M. O., Marla, S. R., Perumal, R., Tesso, T., Yu, J., et al. (2017). Increased power to dissect adaptive traits in global sorghum diversity using a nested association mapping population. *Genetics*, *206*(2), 573–585.

Brown, T. B., Cheng, R., Sirault, X. R., Rungrat, T., Murray, K. D., Trtilek, M., et al. (2014). TraitCapture: Genomic and environment modelling of plant phenomic data. *Current Opinion in Plant Biology*, *18*, 73–79.

Brumlop, S., & Finckh, M. R. (2011). *Applications and potentials of marker assisted selection (MAS) in plant breeding*. Federal Agency for Nature Conservation Konstantinstrasse Bonn, Germany Bundesamt für Naturschutz (BfN)31.

Bubb, K. L., & Queitsch, C. (2016). A two-state epistasis model reduces missing heritability of complex traits. *bioRxiv*. https://doi.org/10.1101/017491.

Bulik-Sullivan, B., Selitsky, S., & Sethupathy, P. (2013). Prioritization of genetic variants in the microRNA regulome as functional candidates in genome-wide association studies. *Human Mutation*, *34*, 1049–1056.

Bush, W. S., & Moore, J. H. (2012). Genome-wide association studies. *PLoS Computational Biology*, *8*, e1002822.

Buzdugan, L., Kalisch, M., Navarro, A., Schunk, D., Fehr, E., & Bühlmann, P. (2016). Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics*, *32*, 1990–2000.

Calus, M. P., Meuwissen, T. H., Windig, J. J., Knol, E. F., Schrooten, C., Vereijken, A. L., et al. (2009). Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genetics Selection Evolution*, *41*, 11.

Campbell, M. T., Knecht, A. C., Berger, B., Chris, J. B., Wang, D., & Walia, H. (2015). Integrating image-based phenomics and association analysis to dissect the genetic architecture of temporal salinity responses in rice. *Plant Physiology*, *168*, 1476–1489.

Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American Journal of Human Genetics*, *86*, 6–22.

Caragea, C., & Honavar, V. (2009). Machine learning in computational biology. In *Encyclopedia of database systems* (pp. 1663–1667). USA: Springer.

Carlson, C. S., Eberle, M. A., Kruglyak, L., & Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. *Nature*, *429*, 446–452.

Carreno-Quintero, N., Bouwmeester, H. J., & Keurentjes, J. J. (2013). Genetic analysis of metabolome–phenotype interactions: From model to crop species. *Trends in Genetics*, *29*, 41–50.

Carroll, R. J., Bastarache, L., & Denny, J. C. (2014). R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*, *31*, 14.

Cavanagh, C., Morell, M., Mackay, I., & Powell, W. (2008). From mutations to MAGIC: Resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology*, *11*, 215–221.

Chan, E. K. F., Rowe, H. C., Corwin, J. A., Joseph, B., & Kliebenstein, D. J. (2011). Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. *PLoS Biology*, *9*, e1001125.

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., et al. (2007). Replicating genotype–phenotype associations. *Nature*, *447*, 655.

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature Genetics*, *46*, 714–721.

Chen, X., Ge, X., Wang, J., Tan, C., King, G. J., & Liu, K. (2015). Genome-wide DNA methylation profiling by modified reduced representation bisulfite sequencing in *Brassica rapa* suggests that epigenetic modifications play a key role in polyploid genome evolution. *Frontiers in Plant Science*, *6*, 836.

Cheng, Z., Chu, H., Fan, Y., Li, C., Song, Y. Q., Zhou, J., et al. (2015). PExFInS: An integrative post-GWAS explorer for functional INDELs and SNPs. *Scientific Reports*, *5*, 17302.

Cho, S., Kim, K., Kim, Y. J., Lee, J. K., Cho, Y. S., Lee, J. Y., et al. (2010). Joint identification of multiple genetic variants via Elastic-Net variable selection in a genome-wide association analysis. *Annals of Human Genetics*, *74*, 416–428.

Chung, D., Kim, H. J., & Zhao, H. (2017). graphGPA: A graphical model for prioritizing GWAS results and investigating pleiotropic architecture. *PLoS Computational Biology*, *13*, e1005388.

Chung, D., Yang, C., Li, C., Gelernter, J., & Zhao, H. (2014). GPA: A statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genetics*, *10*, e1004787.

Claussnitzer, M., Dankel, S. N., Kim, K. H., Quon, G., Meuleman, W., Haugen, C., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. *The New England Journal of Medicine*, *373*, 895–907.

Cockram, J., White, J., Zuluaga, D. L., Smith, D., Comadran, J., Macaulay, M., et al. (2010). Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 21611–21616.

Cubillos, F. A., Coustham, V., & Loudet, O. (2012). Lessons from eQTL mapping studies: Non-coding regions and their role behind natural phenotypic variation in plants. *Current Opinion in Plant Biology*, *15*, 192–198.

Curtin, S. J., Tiffin, P., Guhlin, J., Trujillo, D. I., Burghardt, L. T., Atkins, P., et al. (2017). Validating genome-wide association candidates controlling quantitative variation in nodulation. *Plant Physiology*, *173*, 921–931.

Curtin, S. J., Zhang, F., Sander, J. D., Haun, W. J., Starker, C., Baltes, N. J., et al. (2011). Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases. *Plant Physiology*, *156*, 466–473.

Curtis, D., Vine, A. E., & Knight, J. (2008). A simple method for assessing the strength of evidence for association at the level of the whole gene. *Advances and Applications in Bioinformatics and Chemistry*, *1*, 115–120.

Czarnecki, O., Bryan, A. C., Jawdy, S. S., Yang, X., Cheng, Z. M., Chen, J. G., et al. (2016). Simultaneous knockdown of six non-family genes using a single synthetic RNAi fragment in *Arabidopsis thaliana*. *Plant Methods*, *12*, 16.

D'haene, E., Jacobs, E. Z., Volders, P. J., De Meyer, T., Menten, B., & Vergult, S. (2016). Identification of long non-coding RNAs involved in neuronal development and intellectual disability. *Scientific Reports*, *6*, 28396.

De Bakker, P. I., Ferreira, M. A., Jia, X., Neale, B. M., Raychaudhuri, S., & Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics*, *17*, R122–R128.

Dechamethakun, S., & Muramastu, M. (2017). Long noncoding RNA variations in cardiometabolic diseases. *Journal of Human Genetics*, *62*, 97–104.

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, *31*, 1102–1110.

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., BrownGentry, K., et al. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, *26*, 1205–1210.

Dhanapal, A. P., Ray, J. D., Singh, S. K., Hoyos-Villegas, V., Smith, J. R., Purcell, L. C., et al. (2016). Genome-wide association mapping of soybean chlorophyll traits based on canopy spectral reflectance and leaf extracts. *BMC Plant Biology*, *16*, 174.

Dhanapal, A. P., Ray, J. D., Singh, S. K., Hoyos-Villegas, V., Smith, J. R., Purcell, L. C., et al. (2015). Association mapping of total carotenoids in diverse soybean genotypes based on leaf extracts and high-throughput canopy spectral reflectance measurements. *PLoS One*, *10*, e0137213.

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: Illuminating the dark road from association to function. *American Journal of Human Genetics*, *93*, 779–797.

Edwards, D., & Gupta, P. K. (2013). Sequence based DNA markers and genotyping for cereal genomics and breeding. In P. K. Gupta & R. K. Varshney (Eds.), *Cereal genomics II* (pp. 57–76): Springer.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, *6*, e19379.

Ersoz, E. S., Yu, J., & Buckler, E. S. (2007). Applications of linkage disequilibrium and association mapping in crop plants. In R. Varshney & R. Tuberosa (Eds.), *Genomics approaches and platforms*. : *Vol. 1. Genomic assisted crop improvement* (pp. 97–120).

Eu-ahsunthornwattana, J., Miller, E. N., Fakiola, M., Wellcome Trust Case Control Consortium 2, Jeronimo, S. M. B., Blackwell, J. M., et al. (2014). Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genetics*, *10*, e1004445.

Evangelou, E., & Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, *14*, 379–389.

Fahlgren, N., Gehan, M. A., & Baxter, I. (2015). Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. *Current Opinion in Plant Biology*, *24*, 93–99.

Fantini, E., Falcone, G., Frusciante, S., Giliberto, L., & Giuliano, G. (2013). Dissection of tomato lycopene biosynthesis through virus-induced gene silencing. *Plant Physiology*, *163*, 986–998.

Feng, S., Pistis, G., Zhang, H., Zawistowski, M., Mulas, A., Zoledziewska, M., et al. (2015). Methods for association analysis and meta-analysis of rare variants in families. *Genetic Epidemiology*, *39*, 227–238.

Feng, Z., Zhang, B., Ding, W., Liu, X., Yang, D. L., Wei, P., et al. (2013). Efficient genome editing in plants using a CRISPR/Cas system. *Cell Research*, *23*, 1229–1232.

Fernando, R. L., & Garrick, D. J. (2013). Bayesian methods applied to GWAS. In C. Gondro, J. H. J. van der Werf, & B. Hayes (Eds.), *Genome-wide association studies and genomic prediction* (pp. 237–274). Berlin: Springer Series.

Fernando, R., Toosi, A., Wolc, A., Garrick, D., & Dekkers, J. (2017). Application of whole-genome prediction methods for genome-wide association studies: A Bayesian approach. *Journal of Agricultural, Biological and Environmental Statistics*, *22*, 172–193.

Ferreira, M. A., & Purcell, S. M. (2009). A multivariate test of association. *Bioinformatics*, *25*, 132–133.

Fraas, S., & Luthen, H. (2015). Novel imaging-based phenotyping strategies for dissecting crosstalk in plant development. *Journal of Experimental Botany*, *66*, 4947–4955.

French, J. D., Ghoussaini, M., Edwards, S. L., Meyer, K. B., Michailidou, K., Ahmed, S., et al. (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *American Journal of Human Genetics*, *92*, 489–503.

Furlotte, N. A., & Eskin, E. (2015). Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics*, *200*, 59–68.

Fusi, N., Stegle, O., & Lawrence, N. (2012). Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computational Biology*, *8*, e1002330.

Galesloot, T. E., Van Steen, K., Kiemeney, L. A., Janss, L. L., & Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PLoS One*, *9*, e95923.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*, 1091–1098.

Gao, J., Wang, G., Ma, S., Xie, X., Wu, X., Zhang, X., et al. (2015). CRISPR/Cas9-mediated targeted mutagenesis in *Nicotiana tabacum*. *Plant Molecular Biology*, *87*, 99–110.

Gao, H., Zhang, T., Wu, Y., Jiang, L., Zhan, J., Li, J., et al. (2014). Multiple-trait genome-wide association study based on principal component analysis for residual covariance matrix. *Heredity*, *113*, 526–532.

Gardiner, L. J., Quinton-Tulloch, M., Olohan, L., Price, J., Hall, N., & Hall, A. (2015). A genome-wide survey of DNA methylation in hexaploid wheat. *Genome Biology*, *16*, 273.

Garg, R., Chevla, V. V. S., Shanker, R., & Jain, M. (2015). Divergent DNA methylation patterns associated with gene expression in rice cultivars with contrasting drought and salinity stress response. *Scientific Reports*, *5*, 14922.

Georges, M. (2011). The long and winding road from correlation to causation. *Nature Genetics*, *43*, 180–181.

Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews. Genetics*, *13*, 135–145.

Gjuvsland, A. B., Vik, J. O., Beard, D. A., Hunter, P. J., & Omholt, S. W. (2013). Bridging the genotype–phenotype gap: What does it take? *The Journal of Physiology*, *591*, 2055–2066.

Gogele, M., Minelli, C., Thakkinstian, A., Yurkiewich, A., Pattaro, C., Pramstaller, P. P., et al. (2012). Methods for meta-analyses of genome-wide association studies: Critical assessment of empirical evidence. *American Journal of Epidemiology*, *175*, 739–749.

Gong, J., Liu, C., Liu, W., Wu, Y., Ma, Z., Chen, H., et al. (2015). An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools. *Database*, *2015*, bav029.

Gonzalez-Jorge, S., Ha, S. H., Magallanes-Lundback, M., Gilliland, L. U., Zhou, A., Lipka, A. E., et al. (2013). Carotenoid cleavage dioxygenase4 is a negative regulator of β-carotene content in Arabidopsis seeds. *Plant Cell*, *25*, 4812–4826.

Goulart, L. F., Bettella, F., Sonderby, I. E., Schork, A. J., Thompson, W. K., Mattingsdal, M., et al. (2015). MicroRNAs enrichment in GWAS of complex human phenotypes. *BMC Genomics*, *16*, 304.

Greene, C. S., & Troyanskaya, O. G. (2011). PILGRM: An interactive data-driven discovery platform for expert biologists. *Nucleic Acids Research*, *39*, W368–W374.

Grimm, D. G., Roqueiro, D., Salome, P., Kleeberger, S., Greshake, B., Zhu, W., et al. (2017). easyGWAS: A Cloud-based platform for comparing the results of genome-wide association studies. *Plant Cell*, *29*, 5–19.

Gruenbaum, Y., Naveh-Many, T., Cedar, H., & Razin, H. (1981). Sequence specificity of methylation in higher plant DNA. *Nature*, *292*, 860–862.

Gupta, P. K., Kulwal, P. L., & Jaiswal, V. (2014). Association mapping in crop plants: Opportunities and challenges. *Advances in Genetics*, *38*, 109–147.

Gupta, P. K., Kulwal, P. L., & Mir, R. R. (2013). QTL mapping: Methodology and applications in cereal breeding. In P. K. Gupta & R. K. Varshney (Eds.), *Cereal genomics II* (pp. 275–318). The Netherlands: Springer.

Gupta, P. K., Langridge, P., & Mir, R. R. (2010). Marker-assisted wheat breeding: Present status and future possibilities. *Molecular Breeding*, *26*, 145–161.

Gupta, P. K., Rustgi, S., & Kulwal, P. L. (2005). Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant Molecular Biology*, *57*, 461–485.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*, 245–252.

Hamblin, M. T., & Jannink, J. L. (2011). Factors affecting the power of haplotype markers in association studies. *The Plant Genome*, *4*, 145–153.

Han, B., & Eskin, E. (2012). Interpreting meta-analyses of genome-wide association studies. *PLoS Genetics*, *8*, e1002555.

Hansen, N. E. (2014). *High-throughput phenotyping allows for QTL analysis of defense, symbiosis and development-related traits*. Ph.D. Thesis Denmark: Aarhus University, Faculty of Science and Technology.

Hao, D., Cheng, H., Yin, Z., Cui, S., Zhang, D., Wang, H., et al. (2012). Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. *Theoretical and Applied Genetics*, *124*, 447–458.

Harjes, C. E., Rocheford, T. R., Bai, L., Brutnell, T. P., Kandianis, C. B., Sowinski, S. G., et al. (2008). Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. *Science*, *319*, 330–333.

Harper, A. L., McKinney, L. V., Nielsen, L. R., Havlickova, L., Li, Y., Trick, M., et al. (2016). Molecular markers for tolerance of European ash (*Fraxinus excelsior*) to dieback disease identified using associative transcriptomics. *Scientific Reports*, *6*, 19335.

Harper, A. L., Trick, M., Higgins, J., Fraser, F., Clissold, L., Wells, R., et al. (2012). Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. *Nature Biotechnology*, *30*, 798–802.

He, J., Zhao, X., Laroche, A., Lu, Z. X., Liu, H., & Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science*, *5*, 484.

He, S., Zhao, Y., Mette, M. F., Bothe, R., Ebmeyer, E., Sharbel, T. F., et al. (2015). Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (*Triticum aestivum* L.). *BMC Genomics*, *16*, 168.

Heikoff, S., Till, B. J., & Comai, L. (2004). TILLING. Traditional mutagenesis meets. *Plant Physiology*, *135*, 630–636.

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., et al. (2013). DNA methylation contributes to natural human variation. *Genome Research*, *23*, 1363–1372.

Hiersche, M., Rühle, F., & Stoll, M. (2013). PostGWAS: Advanced GWAS interpretation in R. *PLoS One*, *8*, e71775.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics *in Medicine*, *21*, 1539–1558.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine*, *4*, 45–61.

Hoffman, G. E. (2013). Correcting for population structure and kinship using the linear mixed model: Theory and extensions. *PLoS One*, *8*, e75707.

Hoffmann, T. J., & Witte, J. S. (2015). Strategies for imputing and analyzing rare variants in association studies. *Trends in Genetics*, *31*, 556–563.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, *4*, e1000130.

Hong, H., Xu, L., Liu, J., Jones, W. D., Su, Z., Ning, B., et al. (2012). Technical reproducibility of genotyping SNP arrays used in genome-wide association studies. *PLoS One*, 7, e44483.

Horvath, E., Szalai, G., Janda, T., Páldi, E., Racz, I., & Lasztity, D. (2002). Effect of vernalisation and azacytidine on the DNA methylation level in wheat (*Triticum aestivum* L. cv. Mv 15). In *Proceedings of the Seventh Hungarian Congress on Plant Physiology, 46* (pp. 35–36).

Hou, L., & Zhao, H. (2013). A review of post-GWAS prioritization approaches. *Frontiers in Genetics*, *4*, 280.

Hu, J., Chen, X., Zhang, H., & Ding, Y. (2015). Genome-wide analysis of DNA methylation in photoperiod- and thermo-sensitive male sterile rice Peiai 64S. *BMC Genomics*, *16*, 102.

Huang, X., & Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annual Review of Plant Biology*, *65*, 531–551.

Huang, B. E., Raghavan, C., Mauleon, R., Broman, K. W., & Leung, H. (2014). Efficient imputation of missing markers in low-coverage genotyping-by-sequencing data from multiparental crosses. *Genetics*, *197*, 401–404.

Huang, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, *42*, 961–967.

Huang, B. E., Verbyla, K. L., Verbyla, A. P., Raghavan, C., Singh, V. K., Gaur, P., et al. (2015). MAGIC populations in crops: Current status and future prospects. *Theoretical and Applied Genetics*, *128*, 999–1017.

Huang, J., Wang, K., Wei, P., Liu, X., Liu, X., Tan, K., et al. (2016). FLAGS: A flexible and adaptive association test for gene sets using summary statistics. *Genetics*, *202*, 919–929.

Huang, X., Zhao, Y., Li, C., Wang, A., Zhao, Q., Li, W., et al. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics*, *44*, 32–39.

Huh, I., Kwon, M. S., & Park, T. (2015). An Efficient stepwise statistical test to identify multiple linked human genetic variants associated with specific phenotypic traits. *PLoS One*, *10*, e0138700.

Ingvarsson, P. K., & Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytologist*, *189*, 909–922.

Jaffe, A. E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T. M., Weinberger, D. R., et al. (2016). Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nature Neuroscience*, *19*, 40–47.

Jaiswal, V., Gahlaut, V., Meher, P. K., Mir, R. R., Jaiswal, J. P., Rao, A. R., et al. (2016). Genome wide single locus single trait, multi-locus and multi-trait association mapping for some important agronomic traits in common wheat (*T. aestivum* L.). *PLoS One*, *11*, e0159343.

Jansen, R. C., & Nap, J. P. (2001). Genetical genomics: The added value from segregation. *Trends in Genetics*, *17*, 388–391.

Jia, G., Huang, X., Zhi, H., Zhao, Y., Zhao, Q., Li, W., et al. (2013). A haplotype map of genomic variations and genomewide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nature Genetics*, *45*, 957–961.

Jia, Y., Sun, X., Sun, J., Pan, Z., Wang, X., He, S., et al. (2014). Association mapping for epistasis and environmental interaction of yield traits in 323 cotton cultivars under 9 different environments. *PLoS One*, *9*, e95882.

Jia, P., & Zhao, Z. (2014). Network-assisted analysis to prioritize GWAS results: Principles, methods and perspectives. *Human Genetics*, *133*, 125–138.

Jiang, W., Xue, J. H., & Yu, W. (2015). Estimating reproducibility in genome-wide association studies. *arXiv*. preprint arXiv, 1508.06715.

Johannes, F., Porcher, E., Teixeira, F. K., Saliba-Colombani, V., Simon, M., Agier, N., et al. (2009). Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genetics*, *5*, e1000530.

Kacsoh, B. Z., Greene, C. S., & Bosco, G. (2017). Machine learning analysis identifies Drosophila Grunge/Atrophin as an important learning and memory gene required for memory retention and social learning. *G3: Genes Genomes Genetics*, 7, 3705–3718.

Kalisz, S., & Purugganan, M. D. (2004). Epialleles via DNA methylation: Consequences for plant evolution. *Trends in Ecology & Evolution*, *19*, 309–314.

Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*, 348–354.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*, 1709–1723.

Kant, R., Sharma, S., & Dasgupta, I. (2015). Virus-induced gene silencing (VIGS) for functional genomics in rice using Rice tungro bacilliform virus (RTBV) as a vector. *Methods in Molecular Biology*, *1287*, 201–217.

Kim, C. M., Piao, H. L., Park, S. J., Chon, N. S., Je, B. I., Sun, B., et al. (2004). Rapid, large-scale generation of Ds transposant lines and analysis of the Ds insertion sites in rice. *Plant Journal*, *39*, 252–263.

Kim, S., & Xing, E. P. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genetics*, *5*, e1000587.

Kim, J., Zhang, Y., & Pan, W. (2016). Powerful and adaptive testing for multi-trait and multi-SNP associations with GWAS and sequencing data. *Genetics*, *203*, 715–731.

King, G. J., Amoah, S., & Kurup, S. (2010). Exploring and exploiting epigenetic variation in crops. *Genome*, *53*, 856–868.

Klasen, J. R., Barbez, E., Meier, L., Meinshausen, N., Bühlmann, P., Koornneef, M., et al. (2016). A multi-marker association method for genome-wide association studies without the need for population structure correction. *Nature Communications*, 7, 13299.

Kliebenstein, D. (2009). Quantitative genomics: Analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annual Review of Plant Biology*, *60*, 93–114.

Klukas, C., Chen, D., & Pape, J. M. (2014). Integrated analysis platform: An open-source information system for high-throughput plant phenotyping. *Plant Physiology*, *165*, 506–518.

Knecht, A. C., Campbell, M. T., Caprez, A., Swanson, D. R., & Walia, H. (2016). Image Harvest: An open-source platform for high-throughput plant image processing and analysis. *Journal of Experimental Botany*, *67*, 3587–3599.

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, *533*, 420–424.

Kondo, H., Ozaki, H., Itoh, K., Kato, A., & Takeno, K. (2006). Flowering induced by 5-azacytidine, a DNA demethylating reagent in a short-day plant, *Perilla frutescens* var. crispa. *Physiologia Plantarum*, *127*, 130–137.

Koprivova, A., Harper, A. L., Trick, M., Bancroft, I., & Kopriva, S. (2014). Dissection of the control of anion homeostasis by associative transcriptomics in *Brassica napus*. *Plant Physiology*, *166*, 442–450.

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods*, *9*, 29.

Korte, A., Vilhjalmsson, B. J., Segura, V., Platt, A., Long, Q., & Nordborg, M. (2012). A mixed model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, *44*, 1066–1071.

Kraft, P., Zeggini, E., & Ioannidis, J. P. (2009). Replication in genome-wide association studies. *Statistical Science*, *24*, 561–573.

Kulwal, P. L. (2016). Association mapping and genomic selection—Where does sorghum stand? In *The sorghum genome* (pp. 137–148). Springer.

Kumar, J., Saripalli, G., Gahlaut, V., Goel, N., Meher, P. K., Mishra, K. K., et al. (2018). Genetics of Fe, Zn, β-carotene, GPC and yield traits in bread wheat (*Triticum aestivum* L.) using multi-locus and multi-traits GWAS. *Euphytica*, *214*, 219.

Kumar, A., Soderhall, C., Merid, S., Xu, C., Gruzieva, O., Koppelman, G., et al. (2016). meQTL analysis of asthma GWAS loci and DNA methylation. *European Respiratory Journal*, *48*, PA1209.

Kuromori, T., Takahashi, S., Kondou, Y., Shinozaki, K., & Matsui, M. (2009). Phenome analysis in plant species using loss-of-function and gain-of-function mutants. *Plant Cell Physiology*, *50*, 1215–1231.

Kwak, I. Y., & Pan, W. (2017). Gene- and pathway-based association tests for multiple traits with GWAS summary statistics. *Bioinformatics*, *33*, 64–71.

Lachowiec, J., Shen, X., Queitsch, C., & Carlborg, O. (2015). A Genome-wide association analysis reveals epistatic cancellation of additive genetic variance for root length in *Arabidopsis thaliana*. *PLoS Genetics*, *11*, e1005541.

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

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., & Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, *21*, 1109–1121.

Lee, J., Kim, K. W., Choi, S. H., Huh, J., & Park, S. H. (2015). Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: A practical review for clinical researchers part II. Statistical methods of meta-analysis. *Korean Journal of Radiology*, *16*, 1188–1196.

Li, J., Bus, A., Spamer, V., & Stich, B. (2016). Comparison of statistical models for nested association mapping in rapeseed (*Brassica napus* L.) through computer simulations. *BMC Plant Biology*, *16*, 26.

Li, M. X., Gui, H. S., Kwan, J. S., & Sham, P. C. (2011). GATES: A rapid and powerful gene-based association test using extended Simes procedure. *American Journal of Human Genetics*, *88*, 283–293.

Li, W. D., Jiao, H., Wang, K., Yang, F., Grant, S. F., Hakonarson, H., et al. (2015). Pathway-based genome-wide association studies reveal that the Rac1 pathway is associated with plasma adiponectin levels. *Scientific Reports*, *5*, 13422.

Li, M. J., Li, M., Liu, Z., Yan, B., Pan, Z., Huang, D., et al. (2017). CEPIP: Context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes. *Genome Biology*, *18*, 52.

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 Biology*, *12*, 73.

Li, L., Paulo, M. J., van Eeuwijk, F., & Gebhardt, C. (2010). Statistical epistasis between candidate gene alleles for complex tuber traits in an association mapping population of tetraploid potato. *Theoretical and Applied Genetics*, *121*, 1303–1310.

Li, Y., Pearl, S. A., & Jackson, S. A. (2015). Gene networks in plant biology: Approaches in reconstruction and analysis. *Trends in Plant Science*, *20*, 664–675.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nature Genetics*, *45*, 43–50.

Li, Y., Ruperao, P., Batley, J., Edwards, D., Davidson, J., Hobson, K., et al. (2017). Genome analysis identified novel candidate genes for ascochyta blight resistance in chickpea using whole genome re-sequencing data. *Frontiers in Plant Science*, *8*, 359.

Li, M. J., Sham, P. C., & Wang, J. (2012). Genetic variant representation, annotation and prioritization in the post-GWAS era. *Cell Research*, *22*, 1505–1508.

Li, N., Shi, J., Wang, X., Liu, G., & Wang, H. (2014). A combined linkage and regional association mapping validation and fine mapping of two major pleiotropic QTLs for seed weight and silique length in rapeseed (*Brassica napus* L.). *BMC Plant Biology*, *14*, 114.

Lipka, A. E., Gore, M. A., Magallanes-Lundback, M., Mesberg, A., Lin, H., Tiede, T., et al. (2013). Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. *G3: Genes Genomes Genetics*, *3*, 1287–1299.

Lipka, A. E., Kandianis, C. B., Hudson, M. E., Yu, J., Drnevich, J., Bradbury, P. J., et al. (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.

Lippert, C., Casale, F. P., Rakitsch, B., & Stegle, O. (2014). LIMIX: Genetic analysis of multiple traits. *bioRxiv*. https://doi.org/10.1101/003905.

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.

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.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., et al. (2010). A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics*, *87*, 139–145.

Liu, H., Wang, F., Xiao, Y., Tian, Z., Wen, W., Zhang, X., et al. (2016). MODEM: Multi-omics data envelopment and mining in maize. *Database*, *2016*, 1–9. https://doi.org/10.1093/database/baw117.

Liu, N., Xue, Y., Guo, Z., Li, W., & Tang, J. (2016). Genome-wide association study identifies candidate genes for starch content regulation in maize kernels. *Frontiers in Plant Science*, *7*, 1046.

Liu, J., Yang, C., Shi, X., Li, C., Huang, J., Zhao, H., et al. (2016). Analyzing association mapping in pedigree-based GWAS using a penalized multitrait mixed model. *Genetic Epidemiology*, *40*, 382–393.

Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*, 284–290.

Long, Y., Xia, W., Li, R., Wang, J., Shao, M., Feng, J., et al. (2011). Epigenetic QTL mapping in *Brassica napus*. *Genetics*, *189*, 1093–1102.

Lorenz, A. J., Hamblin, M. T., & Jannink, J. L. (2010). Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PLoS One*, *5*, e14079.

Lu, G., Harper, A. L., Trick, M., Morgan, C., Fraser, F., O'Neill, C., et al. (2014). Associative transcriptomics study dissects the genetic architecture of seed glucosinolate content in *Brassica napus*. *DNA Research*, *21*, 613–625.

Lu, Q., Jin, C., Sun, J., Bowler, R., Kechris, K., Kaminski, N., et al. (2017). Post-GWAS prioritization through data integration provides novel insights on chronic obstructive pulmonary disease. *Statistics in Biosciences*, *9*, 605–621.

Lu, Y., Liu, Y., Niu, X., Yang, Q., Hu, X., Zhang, H. Y., et al. (2015). Systems genetic validation of the SNP-metabolite association in rice via metabolite-pathway-based phenome-wide association scans. *Frontiers in Plant Science*, *6*, 1027.

Lu, H. Y., Liu, X. F., Wei, S. P., & Zhang, Y. M. (2011). Epistatic association mapping in homozygous crop cultivars. *PLoS One*, *6*, e17773.

Lu, R., Malcuit, I., Moffett, P., Ruiz, M. T., Peart, J., Wu, A. J., et al. (2003). High throughput virus-induced gene silencing implicates heat shock protein 90 in plant disease resistance. *EMBO Journal*, *22*, 5690–5699.

Lu, Y., Xu, J., Yuan, Z., Hao, Z., Xie, C., Li, X., et al. (2012). Comparative LD mapping using single SNPs and haplotypes identifies QTL for plant height and biomass as secondary traits of drought tolerance in maize. *Molecular Breeding*, *30*, 407–418.

Lu, Q., Yao, X., Hu, Y., & Zhao, H. (2016). GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. *Bioinformatics*, *32*, 542–548.

Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., et al. (2010). Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 19585–19590.

Lu, X., Zhao, X., Wang, D., Yin, Z., Wang, J., Fan, W., et al. (2015). Whole-genome DNA methylation analysis in cotton (*Gossypium hirsutum* L.) under different salt stresses. *Turkish Journal of Biology*, *39*, 396–406.

Luo, J. (2015). Metabolite-based genome-wide association studies in plants. *Current Opinion in Plant Biology*, *24*, 31–38.

Magi, R., & Morris, A. P. (2010). GWAMA: Software for genome-wide association meta-analysis. *BMC Bioinformatics*, *11*, 288.

Manhaes, A. M., de Oliveira, M. V., & Shan, L. (2015). Establishment of an efficient virus-induced gene silencing (VIGS) assay in *Arabidopsis* by *Agrobacterium*-mediated rubbing infection. *Methods in Molecular Biology*, *1287*, 235–241.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747–753.

Mao, Y., Zhang, H., Xu, N., Zhang, B., Gou, F., & Zhu, J. K. (2013). Application of the CRISPR-Cas system for efficient genome engineering in plants. *Molecular Plant*, *6*, 2008–2011.

Marjoram, P., & Thomas, D. C. (2014). Next-generation sequencing studies: Optimal design and analysis, missing heritability and rare variants. *Current Epidemiology Reports*, *1*, 213–219.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., et al. (2017). Rare and low-frequency coding variants alter human adult height. *Nature*, *542*, 186–190.

Marttinen, P., Gillberg, J., Havulinna, A., Corander, J., & Kaski, S. (2013). Genome-wide association studies with high-dimensional phenotypes. *Statistical Applications in Genetics and Molecular Biology*, *12*, 413–431.

Matsuda, F., Nakabayashi, R., Yang, Z., Okazaki, Y., Yonemaru, J. I., Ebana, K., et al. (2015). Metabolome genome wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. *Plant Journal*, *81*, 13–23.

McCouch, S. R., Wright, M. H., Tung, C. W., Maron, L. G., McNally, K. L., Fitzgerald, M., et al. (2016). Open access resources for genome-wide association mapping in rice. *Nature Communications*, *7*, 10532.

Mei, Y., Zhang, C., Kernodle, B. M., Hill, J. H., & Whitham, S. A. (2016). A foxtail mosaic virus vector for virus-induced gene silencing in maize. *Plant Physiology*, *171*, 760–772.

Messeguer, R., Ganal, M. W., Steffens, J. C., & Tanksley, S. D. (1991). Characterization of the level, target sites and inheritance of cytosine methylation in tomato nuclear DNA. *Plant Molecular Biology*, *16*, 753–770.

Meuwissen, T. H., & Goddard, M. E. (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution*, *36*, 261–279.

Miao, J., Guo, D., Zhang, J., Huang, Q., Qin, G., Zhang, X., et al. (2013). Targeted mutagenesis in rice using CRISPR-Cas system. *Cell Research*, *23*, 1233–1236.

Mieth, B., Kloft, M., Rodríguez, J. A., Sonnenburg, S., Vobruba, R., Morcillo-Suarez, C., et al. (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Scientific Reports*, *6*, 36671.

Miller, C. N., Harper, A. L., Trick, M., Werner, P., Waldron, K., & Bancroft, I. (2016). Elucidation of the genetic basis of variation for stem strength characteristics in bread wheat by associative transcriptomics. *BMC Genomics*, *17*, 500.

Mooney, M. A., & Wilmot, B. (2015). Gene set analysis: A step-by-step guide. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics*, *168*, 517–527.

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., et al. (2015). The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genetics*, *11*, e1005165.

Müller-Linow, M., Pinto-Espinosa, F., Scharr, H., & Rascher, U. (2015). The leaf angle distribution of natural plant populations: Assessing the canopy with a novel software tool. *Plant Methods*, *11*, 11.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: The MIT Press.

N'Diaye, A., Haile, J. K., Cory, A. T., Clarke, F. R., Clarke, J. M., Knox, R. E., et al. (2017). Single marker and haplotype-based association analysis of semolina and pasta colour in elite durum wheat breeding lines using a high-density consensus map. *PLoS One*, *12*, e0170941.

Neale, B. M., & Sham, P. C. (2004). The future of association studies: Gene-based analysis and replication. *American Journal of Human Genetics*, *75*, 353–362.

Nekrasov, V., Staskawicz, B., Weigel, D., Jones, J. D. G., & Kamoun, S. (2013). Targeted mutagenesis in the model plant *Nicotiana benthamiana* using Cas9 RNA-guided. *Nature Biotechnology*, *31*, 691–693.

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*, 20120362.

Nice, L. M., Steffenson, B. J., Brown Guedira, G. L., Akhunov, E. D., Liu, C., Kono, T. J., et al. (2016). Development and genetic characterization of an advanced backcross–nested association mapping (AB-NAM) population of wild × cultivated barley. *Genetics*, *203*, 1453–1467.

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., & Aittokallio, T. (2014). Regularized machine learning in the genetic prediction of complex traits. *PLoS Genetics*, *10*, e1004754.

Osakabe, Y., Watanabe, T., Sugano, S. S., Ueta, R., Ishihara, R., Shinozaki, K., et al. (2016). Optimization of CRISPR/Cas9 genome editing to modify abiotic stress responses in plants. *Scientific Reports*, *6*, 26685.

Owens, B. F., Lipka, A. E., Magallanes-Lundback, M., Tiede, T., Diepenbrock, C. H., Kandianis, C. B., et al. (2014). A foundation for provitamin A biofortification of maize: Genome-wide association and genomic prediction models of carotenoid levels. *Genetics*, *198*, 1699–1716.

Paik, H., Kim, J., Lee, S., Heo, H. S., Hur, C. G., & Lee, D. (2012). Prioritization of SNPs for genome-wide association studies using an interaction model of genetic variation, gene expression, and trait variation. *Molecules and Cells*, *33*, 351–361.

Pan, C., Ye, L., Li, Q., Liu, X., He, Y., Wang, J., et al. (2016). CRISPR/Cas9-mediated efficient and heritable targeted mutagenesis in tomato plants in the first and later generations. *Scientific Reports*, *6*, 24765.

Panagiotou, O. A., Willer, C. J., Hirschhorn, J. N., & Ioannidis, J. P. (2013). The power of meta-analysis in genome Wide Association Studies. *Annual Review of Genomics and Human Genetics*, *14*, 441.

Parkhomenko, E., Tritchler, D., & Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, *8*, 1–34.

Perea, C., De La Hoz, J. F., Cruz, D. F., Lobaton, J. D., Izquierdo, P., Quintero, J. C., et al. (2016). Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. *BMC Genomics*, *17*, 498.

Pfeifer, S. P. (2017). From next-generation resequencing reads to a high-quality variant data set. *Heredity*, *118*, 111–124.

Porcu, E., Sanna, S., Fuchsberger, C., & Fritsche, L. G. (2013). Genotype imputation in genome-wide association studies. *Current Protocols in Human Genetics*, *78*, 1.25.

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

Pound, M. P., French, A. P., Atkinson, J. A., Well, D. M., & Bennett, M. J. (2013). RootNav: Navigating images of complex root architectures. *Plant Physiology*, *162*, 1802–1814.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., et al. (2010). LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics*, *26*, 2336–2337.

Rakitsch, B., Lippert, C., Stegle, O., & Borgwardt, K. (2013). A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, *29*, 206–214.

Ramstein, G. P., Lipka, A. E., Lu, F., Costich, D. E., Cherney, J. H., Buckler, E. S., et al. (2015). Genome-wide association study based on multiple imputation with low-depth sequencing data: Application to biofuel traits in reed canarygrass. *G3: Genes Genome Genetics*, *5*, 891–909.

Rashkin, S., Jun, G., Chen, S., Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), & Abecasis, G. R. (2017). Optimal sequencing strategies for identifying disease-associated singletons. *PLoS Genetics*, *13*, e1006811.

Reinders, J., Wulff, B. B., Mirouze, M., Mari-Ordonez, A., Dapp, M., Rozhon, W., et al. (2009). Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes and Development*, *23*, 939–950.

Remington, D. L., Thornsberry, J., Matsuoka, Y., Wilson, L., Rinehart-Whitt, S., Doebley, J., et al. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 11479–11484.

Richter, A., Schaff, C., Zhang, Z., Lipka, A. E., Tian, F., Köllner, T. G., et al. (2016). Characterization of biosynthetic pathways for the production of the volatile homoterpenes DMNT and TMTT in *Zea mays*. *The Plant Cell*, *28*, 2651–2665.

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. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 8872–8877.

Ritchie, M. D. (2015). Finding the epistasis needles in the genome-wide haystack. *Methods in Molecular Biology*, *1253*, 19–33.

Runcie, D. E., & Mukherjee, S. (2013). Dissecting high-dimensional phenotypes with Bayesian sparse factor analysis of genetic covariance matrices. *Genetics*, *194*, 753–767.

Rutkoski, J. E., Poland, J., Jannink, J. L., & Sorrells, M. E. (2013). Imputation of unordered markers and the impact on genomic selection accuracy. *G3: Genes Genomes Genetics*, *3*, 427–439.

Saccone, S. F., Bolze, R., Thomas, P., Quan, J., Mehta, G., Deelman, E., et al. (2010). SPOT: A web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic Acids Research*, *38*, W201–W209.

Saccone, S. F., Saccone, N. L., Swan, G. E., Madden, P. A., Goate, A. M., Rice, J. P., et al. (2008). Systematic biological prioritization after a genome-wide association study: An application to nicotine dependence. *Bioinformatics*, *24*, 1805–1811.

Saidou, A. A., Thuillet, A. C., Couderc, M., Mariac, C., & Vigouroux, Y. (2014). Association studies including genotype by environment interactions: Prospects and limits. *BMC Genetics*, *15*, 3.

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., et al. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiology*, *165*, 1120–1132.

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews. Genetics*, *19*, 491–504.

Scofield, S. R., & Brandt, A. S. (2012). Virus-induced gene silencing in hexaploid wheat using barley stripe mosaic virus vectors. *Methods in Molecular Biology*, *894*, 93–112.

Segura, V., Vilhjalmsson, B. J., Platt, A., Korte, A., Seren, U., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, *44*, 825–830.

Senthil-Kumar, M., & Mysore, K. S. (2014). Tobacco rattle virus-based virus-induced gene silencing in *Nicotiana benthamiana*. *Nature Protocols*, *9*, 1549–1562.

Seren, U., Vilhjalmsson, B. J., Horton, M. W., Meng, D., Forai, P., Huang, Y. S., et al. (2012). GWAPP: A web application for genome-wide association mapping in Arabidopsis. *Plant Cell*, *24*, 4793–4805.

Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., et al. (2013). Targeted genome modification of crop plants using a CRISPR-Cas system. *Nature Biotechnology*, *31*, 686–688.

Shen, Y., Pan, G., & Lubberstedt, T. (2015). Haploid strategies for functional validation of plant genes. *Trends in Biotechnology*, *33*, 611–620.

Skol, A. D., Scott, L. J., Abecasis, G. R., & Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics*, *38*, 209–213.

Smemo, S., Tena, J. J., Kim, K. H., Gamazon, E. R., Sakabe, N. J., Gomez-Marin, C., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, *507*, 331–375.

Spain, S. L., & Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human Molecular Genetics*, *24*, R111–R119.

Stegle, O., Parts, L., Durbin, R., & Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology*, *6*, e1000770.

Stemple, D. L. (2004). TILLING—A high-throughput harvest for functional genomics. *Nature Reviews. Genetics*, *5*, 145–150.

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews. Genetics*, *10*, 681–690.

Stevens, J. R., Al Masud, A., & Suyundikov, A. (2017). A comparison of multiple testing adjustment methods with block-correlation positively-dependent tests. *PLoS One*, *12*, e0176124.

Stich, B., Mohring, J., Piepho, H. P., Heckenberger, M., Buckler, E. S., & Melchinger, A. E. (2008). Comparison of mixed-model approaches for association mapping. *Genetics*, *178*, 1745–1754.

Sukumaran, S., & Yu, J. (2014). Association mapping of genetic resources: Achievements and future perspectives. In *Genomics of plant genetic resources* (pp. 207–235). The Netherlands: Springer.

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: Provocative insights from epigenomics. *Nature Reviews. Genetics*, *9*, 465–476.

Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M., & Aulchenko, Y. S. (2012). Rapid variance components-based method for whole-genome association analysis. *Nature Genetics*, *44*, 1166–1170.

Swarts, K., Li, H., Romero Navarro, J. A., An, D., Romay, M. C., Hearne, S., et al. (2014). Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. *The Plant Genome*, *7*, 1–12.

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H., et al. (2009). Machine learning in genome-wide association studies. *Genetic Epidemiology*, *33*(S1), S51–S57.

Tang, Z. Z., & Lin, D. Y. (2015). Meta-analysis for discovering rare-variant associations: Statistical methods and software programs. *American Journal of Human Genetics*, *97*, 35–53.

Tang, J. D., Perkins, A., Williams, W. P., & Warburton, M. L. (2015). Using genome-wide associations to identify metabolic pathways involved in maize aflatoxin accumulation resistance. *BMC Genomics*, *16*, 673.

Tarca, A. L., Carey, V. J., Chen, X. W., Romero, R., & Draghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational Biology*, *3*, e116.

Taşan, M., Musso, G., Hao, T., Vidal, M., MacRae, C. A., & Roth, F. P. (2015). Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nature Methods*, *12*, 154–159.

The UK10K Consortium. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*, 82–90.

Thoen, M. P., Davila Olivas, N. H., Kloth, K. J., Coolen, S., Huang, P. P., Aarts, M. G., et al. (2017). Genetic architecture of plant stress resistance: Multi-trait genome-wide association mapping. *New Phytologist*, *213*, 1346–1362.

Thomas, D. (2010). Gene–environment-wide association studies: Emerging approaches. *Nature Reviews. Genetics*, *11*, 259–272.

Thomas, L. F., Saito, T., & Saetrom, P. (2011). Inferring causative variants in microRNA target sites. *Nucleic Acids Research*, *39*, e109.

Thompson, J. R., Gögele, M., Weichenberger, C. X., Modenese, M., Attia, J., Barrett, J. H., et al. (2013). SNP prioritization using a bayesian probability of association. *Genetic Epidemiology*, *37*, 214–221.

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.

Torkamani, A., Topol, E. J., & Schork, N. J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, *92*, 265–272.

Torres, J. M., Barbeira, A. N., Bonazzola, R., Morris, A. P., Shah, K. P., Wheeler, H. E., et al. (2017). Integrative cross tissue analysis of gene expression identifies novel type 2 diabetes genes. *bioRxiv* https://doi.org/10.1101/108134.

Tsepilov, Y. A., Sharapov, S. Z., Zaytseva, O. O., Krumsiek, J., Prehn, C., Adamski, J., et al. (2018). Network based conditional genome wide association analysis of human metabolomics. *Giga Science*, *7*, 1–11.

Tucker, G., Price, A. L., & Berger, B. (2014). Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. *Genetics*, *197*, 1045–1049.

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, *50*, 229–237.

Upton, A., Trelles, O., Cornejo-Garcıa, J. A., & Perkins, J. R. (2015). Review: High-performance computing to detect epistasis in genome scale data sets. *Briefings in Bioinformatics*, *17*, 368–379.

Uren, C., Henn, B. M., Franke, A., Wittig, M., van Helden, P. D., Hoal, E. G., et al. (2017). A post-GWAS analysis of predicted regulatory variants and tuberculosis susceptibility. *PLoS One*, *12*, e0174738.

Vaez, A., van der Most, P. J., Prins, B. P., Snieder, H., van den Heuvel, E., Alizadeh, B. Z., et al. (2016). lodGWAS: A software package for genome-wide association analysis of biomarkers with a limit of detection. *Bioinformatics*, *32*, 1552–1554.

van der Sijde, M. R., Ng, A., & Fu, J. (2014). Systems genetics: From GWAS to disease pathways. *Biochimica et Biophysica Acta*, *1842*, 1903–1909.

Van der Sluis, S., Dolan, C. V., Li, J., Song, Y., Sham, P., Posthuma, D., et al. (2015). MGAS: A powerful tool for multivariate gene-based genome-wide association analysis. *Bioinformatics*, *31*, 1007–1015.

Varshney, R. K., Saxena, R. K., Upadhyaya, H. D., Khan, A. W., Yu, Y., Kim, C., et al. (2017). Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. *Nature Genetics*, *49*, 1082–1088.

Varshney, R. K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., et al. (2017). Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nature Biotechnology*, *35*, 969–976.

Verslues, P. E., Lasky, J. R., Juenger, T. E., Liu, T. W., & Kumar, M. N. (2014). Genome-wide association mapping combined with reverse genetics identifies new effectors of low water potential-induced proline accumulation in *Arabidopsis*. *Plant Physiology*, *164*, 144–159.

Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., Gao, C., et al. (2014). Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nature Biotechnology*, *32*, 947–951.

Wang, T., & Elston, R. C. (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. *American Journal of Human Genetics*, *80*, 353–360.

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. *Scientific Reports*, *6*, 19444.

Wang, M., Huang, J., Liu, Y., Ma, L., Potash, J. B., & Han, S. (2017). COMBAT: A combined association test for genes using summary statistics. *Genetics*, *207*, 883–891.

Wang, K., Li, M., & Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *American Journal of Human Genetics*, *81*, 1278–1283.

Wang, X., Pang, Y., Zhang, J., Wu, Z., Chen, K., Ali, J., et al. (2017). Genome-wide and gene-based association mapping for rice eating and cooking characteristics and protein content. *Scientific Reports*, 7, 17203.

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.

Wang, F., Wang, C., Liu, P., Lei, C., Hao, W., Gao, Y., et al. (2016). Enhanced rice blast resistance by CRISPR/Cas9-targeted mutagenesis of the ERF transcription factor gene *OsERF922*. *PLoS One*, *11*, e0154027.

Wang, H., Xu, C., Liu, X., Guo, Z., Xu, X., Wang, S., et al. (2017). Development of a multiple-hybrid population for genome-wide association studies: Theoretical consideration and genetic mapping of flowering traits in maize. *Scientific Reports*, 7, 40239.

Wang, J., Yu, H., Weng, X., Xie, W., Xu, C., Li, X., et al. (2014). An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. *Journal of Experimental Botany*, *65*, 1069–1079.

Ward, J. A., Bhangoo, J., Fernandez-Fernandez, F., Moore, P., Swanson, J. D., Viola, R., et al. (2013). Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome independent imputation. *BMC Genomics*, *14*, 2.

Waterhouse, P. M., Wang, M. B., & Lough, T. (2001). Gene silencing as an adaptive defence against viruses. *Nature*, *411*, 834–842.

Way, G. P., Youngstrom, D. W., Hankenson, K. D., Greene, C. S., & Grant, S. F. (2017). Implicating candidate genes at GWAS signals by leveraging topologically associating domains. *European Journal of Human Genetics*, *25*, 1286–1289.

Wei, P., Cao, Y., Zhang, Y., Xu, Z., Kwak, I. Y., Boerwinkle, E., et al. (2016). On robust association testing for quantitative traits and rare variants. *G3: Genes Genomes Genetics*, *6*, 3941–3950.

Wei, W. H., Hemani, G., & Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews. Genetics*, *15*, 722–733.

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., et al. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nature Communications*, *5*, 3438.

Weng, L., Macciardi, F., Subramanian, A., Guffanti, G., Potkin, S. G., Yu, Z., et al. (2011). SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*, *12*, 99.

West, M. A., Kim, K., Kliebenstein, D. J., van Leeuwen, H., Michelmore, R. W., Doerge, R. W., et al. (2007). Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics*, *175*, 1441–1450.

Westra, H. J., & Franke, L. (2014). From genome to function by studying eQTLs. *Biochimica et Biophysica Acta*, *1842*, 1896–1902.

Whan, A. P., Smith, A. B., Cavanagh, C. R., Ral, J. P. F., Shaw, L. M., Howitt, C. A., et al. (2014). GrainScan: A low cost, fast method for grain size and colour measurements. *Plant Methods*, *10*, 23.

Whitehead, A., & Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, *10*, 1665–1677.

Widmer, C., Lippert, C., Weissbrod, O., Fusi, N., Kadie, C., Davidson, R., et al. (2014). Further improvement to linear mixed models for genome-wide association studies. *Scientific Reports*, *4*, 6874.

Witten, D. M., & Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, *8*, 1–27.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, *89*, 82–93.

Wu, R., Ma, C. X., & Casella, G. (2002). Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations. *Genetics*, *160*, 779–792.

Wu, R., & Zeng, Z. B. (2001). Joint linkage and linkage disequilibrium mapping in natural populations. *Genetics*, *157*, 899–909.

Wurschum, T., Liu, W., Gowda, M., Maurer, H. P., Fischer, S., Schechert, A., et al. (2012). Comparison of biometrical models for joint linkage association mapping. *Heredity*, *108*, 332–340.

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.

Xing, E. P., Curtis, R. E., Schoenherr, G., Lee, S., Yin, J., Puniyani, K., et al. (2014). GWAS in a box: Statistical and visual analytics of structured associations via GenAMap. *PLoS One*, *9*, e97524.

Xu, Y., & Iglewicz, B. (2016). False discovery versus familywise error rate approaches to outlier detection. *Statistics in Biopharmaceutical Research*, *8*, 143–150.

Xu, Z., Wu, C., Wei, P., & Pan, W. (2017). A powerful framework for integrating eQTL and GWAS summary data. *Genetics*, *207*, 893–902.

Yan, J., Kandianis, C., Harjes, C., Bai, L., Kim, E., Yang, X., et al. (2010). Rare genetic variation at *Zea mays crtRB*1 increases b-carotene in maize grain. *Nature Genetics*, *42*, 322–327.

Yang, W., Duan, L., Chen, G., Xiong, L., & Liu, Q. (2013). Plant phenomics and high-throughput phenotyping: Accelerating rice functional genomics using multidisciplinary technologies. *Current Opinion in Plant Biology*, *16*, 180–187.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, *44*, 369–375.

Yang, J., Fritsche, L. G., Zhou, X., Abecasis, G., & International Age-Related Macular Degeneration Genomics Consortium (2017). A Scalable Bayesian method for integrating functional information in genome-wide association studies. *American Journal of Human Genetics*, *101*, 404–416.

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., et al. (2014). Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nature Communications*, *5*, 5087.

Yang, N., Lu, Y., Yang, X., Huang, J., Zhou, Y., Ali, F., et al. (2014). Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. *PLoS Genetics*, *10*, e1004573.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., et al. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics*, *48*, 927–934.

Younis, A., Siddique, M. I., Kim, C. K., & Lim, K. B. (2014). RNA interference (RNAi) induced gene silencing: A promising approach of hi-tech plant breeding. *International Journal of Biological Sciences*, *10*, 1150–1158.

Yu, J., Holland, J. B., McMullen, M. D., & Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, *178*, 539–551.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*, 203–208.

Yuan, C., Li, C., Yan, L., Jackson, A. O., Liu, Z., Han, C., et al. (2011). A high throughput barley stripe mosaic virus vector for virus induced gene silencing in monocots and dicots. *PLoS One*, *6*, e26468.

Yuan, H., Liu, H., Liu, Z., Owzar, K., Han, Y., Su, L., et al. (2016). A novel genetic variant in long non-coding RNA gene NEXN-AS1 is associated with risk of lung cancer. *Scientific Reports*, *6*, 34234.

Yugi, K., Kubota, H., Hatano, A., & Kuroda, S. (2016). Trans-Omics: How to reconstruct biochemical networks across multiple 'omic' layers. *Trends in Biotechnology*, *34*, 276–290.

Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E., & Halperin, E. (2010). Leveraging genetic variability across populations for the identification of causal variants. *American Journal of Human Genetics*, *86*, 23–33.

Zhan, X., Zhao, N., Plantinga, A., Thornton, T., Conneely, K. N., Epstein, M. P., et al. (2017). Powerful genetic association analysis for common or rare variants with high dimensional structured traits. *Genetics*, *206*, 1779–1790.

Zhang, Z., Buckler, E. S., Casstevens, T. M., & Bradbury, P. J. (2009). Software engineering the mixed model for genome-wide association studies on large samples. *Briefings in Bioinformatics*, *10*, 664–675.

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. *Nature Genetics*, *42*, 355–360.

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: Integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity*, *118*, 517–524.

Zhang, Y., Lian, Z., Zong, Y., Wang, Y., Liu, J., Chen, K., et al. (2016). Efficient and transgene-free genome editing in wheat through transient expression of CRISPR/Cas9 DNA or RNA. *Nature Communications*, *7*, 12617.

Zhang, X., Warburton, M. L., Setter, T., Liu, H., Xue, Y., Yang, N., et al. (2016). Genome-wide association studies of drought-related metabolic changes in maize using an enlarged SNP panel. *Theoretical and Applied Genetics*, *129*, 1449–1463.

Zhang, C., Whitham, S. A., & Hill, J. H. (2013). Virus-induced gene silencing in soybean and common bean. *Methods in Molecular Biology*, *975*, 149–156.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W. L., Chen, H., et al. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell*, *126*, 1189–1201.

Zhao, J., Xu, Y., Ding, Q., Huang, X., Zhang, Y., Zou, Z., et al. (2016). Association mapping of main tomato fruit sugars and organic acids. *Frontiers in Plant Science*, *7*, 1286.

Zheng, J., Gaunt, T. R., & Day, I. N. (2013). Sequential sentinel SNP regional association plots (SSS-RAP): An approach for testing independence of snp association signals using meta-analysis data. *Annals of Human Genetics*, *77*, 67–79.

Zheng, J., Rodriguez, S., Laurin, C., Baird, D., Trela-Larsen, L., Erzurumluoglu, M. A., et al. (2017). HAPRAP: A haplotype-based iterative method for statistical fine mapping using GWAS summary statistics. *Bioinformatics*, *33*, 79–86.

Zhou, X., Baron, R. M., Hardin, M., Cho, M. H., Zielinski, J., Hawrylkiewicz, I., et al. (2012). Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. *Human Molecular Genetics*, *21*, 1325–1335.

Zhou, X., Carbonetto, P., & Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, *9*, e1003264.

Zhou, J. J., Hu, T., Qiao, D., Cho, M. H., & Zhou, H. (2016). Boosting gene mapping power and efficiency with efficient exact variance component tests of SNP sets. *Genetics*, *204*, 921–931.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, *33*, 408–414.

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*, 821–824.

Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, *11*, 407–409.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*, 481–487.

Ziebarth, J. D., Bhattacharya, A., Chen, A., & Cui, Y. (2012). PolymiRTS Database 2.0: Linking polymorphisms in microRNA target sites with human diseases and complex traits. *Nucleic Acids Research*, *40*, D216–D221.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., et al. (2014). Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, E455–E464.