Tropical Plant Biology (2020) 13:212–224
https://doi.org/10.1007/s12042-020-09261-4

Principles and approaches of association mapping in plant breeding

Aminu Kurawa Ibrahim [1,2] · Liwu Zhang [1] · Sylvain Niyitanga [1,2] · Muhammad Zohaib Afzal [1,2] · Yi Xu [1,2] · Lilan Zhang [1,2] · Liemei Zhang [1,2] · Jianmin Qi [1,2]

Received: 21 November 2019 / Accepted: 2 April 2020 / Published online: 14 May 2020

Abstract

Association mapping (AM) is an approach that accounts for thousands of polymorphisms to evaluate the effects of quantitative trait loci (QTL). It is an important instrument for identification of alleles and new genes as well as dissection of complex characters. AM is more advantageous than linkage analysis due to the comparatively high-resolution provided, which is based on the structure of linkage disequilibrium (LD). Marker density, population, sample sizes and population structure are among the critical factors that should be considered when AM is used. It is necessary to note that, the choice of germplasm, genotypic and phenotypic data quality, the use of appropriate statistical analysis for marker-phenotype association detections and verifications are key to association analysis. Great potentials to enhance crop genetic improvement are offered by AM. However, to understand its application, extensive research is needed, such as improvements in computational and statistical methods and its integration with gene annotation data or functional analysis. Statistical apparatuses that are user-friendly and genetic resources are also needed and must be enhanced. Rare allele/variant analysis is an important area to be considered to enhance AM studies. Joint linkage association mapping has now been proposed to improve linkage-based QTL mapping and AM limitations. In the future, new candidate genes and QTL can be easily identified if genome-wide association studies (GWAS) are combined with functional genomics. As such, this review describes association mapping, its utilization in plant breeding, limitations as well as advantages over linkage mapping.

Keywords Association mapping (AM) · Genome-wide association studies (GWAS) · Recombinant inbred lines · Quantitative trait loci · Single nucleotide polymorphisms · Candidate genes

Introduction

Association mapping (AM) as a complement to linkage mapping (Breseghello and Sorrells, 2006a) overcomes the limitations of bi-parental population. Loci controlling the traits of interest in bi-parental population escape detection due to their inability to identify loci from similar parents. This technique requires the utilization of different individuals that raises allele numbers examined as well as multiple historical recombinant events. Rare alleles are difficult to be detected in AM and to get higher alleles frequency that includes the genetic diversity of the crop species. AM panels must be suitably chosen. This will effectively decrease duration and costs while detecting markers connected to quantitative characters. AM considers the use of panels with diverse cultivars for the purpose of recording more recombination events that contribute to a higher resolution to find regions associated with trait than linkage mapping (Zhu et al. 2008; Stich and Melchinger 2010). AM is divided into two components: Genome-wide association studies (GWAS) and candidate gene. Quantitative trait loci (QTL) obtained through interval mapping are validated using GWAS. Markers sharing an association with traits are identified in AM. However, diversity between the sample and the number of chromosomes affects the genotype study. Markers such as Simple Sequence Repeats (SSR), Expressed Sequence Tag (EST), Restriction Fragment Length Polymorphism (RFLP), Random Amplified Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphism (AFLP), Diversity Arrays Technology

Communicated by: Zhi-Liang Zheng

✉ Liwu Zhang
zhang_liwu@hotmail.com; lwzhang@fafu.edu.cn

[1] Key Laboratory of Ministry of Education for Genetics, Breeding and Multiple Utilization of Crops/Fujian Provincial Key Laboratory of Crop Breeding by Design/College of Agriculture, Fujian Agriculture and Forestry University, Fuzhou 350002, China

[2] Experiment Station of Jute and Kenaf in Southeast China of Ministry of Agriculture and Rural Affairs / Public Platform for Germplasm Resources of Bast Fiber Crops of Fujian, Fujian Agriculture and Forestry University, Fuzhou 350002, Fujian, China

(DArT), and Single Nucleotide Polymorphisms (SNP) have contributed to AM. To determine if any relationship exists between phenotypes and markers, GWAS scans the whole genome. Nevertheless, more markers are required to cover the genome, considering the anticipated rate of linkage disequilibrium (LD) decay. Most AM activities are done using GWAS with SNPs. Nevertheless, this method has several restrictions. First, knowledge about the genome is needed for designing SNP arrays and the location of SNPs in the genome. Second, the phenotype might be caused by rare variants that are not on the SNP chip. Another restriction is the existence of structural variations. However, high throughput sequencing data are utilized to overcome some of these limitations. Mapping all the reads to the reference genome, followed by variant calling using mapping by sequencing (MBS) (including RNA-Seq, Whole-genome sequencing WGS and Bulked Segregant Analysis (BSA)) are used to overcome such limitations (Hartwig et al., 2012). These variants are then tested for association. But this needs a reference genome, and the regions not in the reference genome will not be captured in the study and it may induce biases in variant calling. Moreover, genotype calling will be complicated when sequencing depth is low (Nielsen et al. 2012) due to the sequencing errors and repetitive regions. An alternate approach is genotyping using tools such as Cortex and simultaneous de-novo assembly (Iqbal et al. 2012). However, it should be noted that both approaches are computational and costly. Sequence identification which is significantly diverged from the reference genes, like R genes, is complicated in GWAS. But, the use of traits association on sub-sequences (*k*-mers) overcomes such limitations. Needle in the k-stack (NIKS) for mutation identification was introduced by comparison of sequencing data from two strains using k-mers (Nordström et al., 2013). Based on counting and identifying k-mers associated with the phenotype, the overlapped k-mers are then assembled to obtain sequences corresponding to associated regions. *K*-mer-based association genetics was used by Arora et al. (2019) to clone R genes from plant diversity panel. The data were then combined with R gene enrichment sequencing (AgRenSeq) to identify *Sr* genes in the phenotyped group. However, the authors observed that complete Nucleotide-binding Leucine-rich Repeat (NLR) contigs would not be generated if local assembly approaches that use only those *k*-mers that are strongly linked to the trait were used.

Linkage analysis and AM

Recombination number is very few within pedigree and families in linkage mapping (Zhu et al. 2008), which leads to low mapping resolution. While recombination tends to be high and diverse in AM, natural genetic diversity is exploited, leading to higher resolutions. Wu and Zeng (2001) proposed Joint linkage-association mapping (JLAM) to overcome low resolution and power in bi-parental mapping and AM limitations, respectively, to harness their potentials. Equally, Chromosomal crossing over has been used in association and linkage mappings to break up allele associations into new haplotypes that link to phenotypic variations (Myles et al. 2009). The key difference between the two methods lies on the degree (either through mating design or selection of the set of germplasms) at which the researcher has overcome the recombination events. Usually, in linkage mapping, the researcher makes use of biparental populations, thereby making it more feasible to control the possibility for recombination events in the progeny, though having a corresponding loss in mapping resolution related to the AM. Association panels in AM can be regarded as a more natural experiment because there is no control over the number of recombination events that produce the tested genotypes (Álvarez et al., 2015). Diverse panels such as the bi- and multi-parent populations, as well as breeding populations are used in both AM and linkage studies, though they have their own advantages and limitations (Xiao et al. 2017). Thus, AM assesses correlations between phenotypes and genotypes, from which QTL can be detected in traits that show variation. The main advantages of AM over linkage mapping are resolution power and accommodation of multiple alleles to be tested for associations. Moreover, the probability of creating populations with positive versus negative alleles exists in linkage mapping, whereas only phenotypic range values for the alleles are involved in AM present in a population (Álvarez et al., 2015). It should not be assumed that the frequency distribution of alleles at functional loci be the same as that of the distribution of alleles at random loci. Instead, it will be tough to account for most phenotypic differences using AM, because rare alleles usually cause most of it.

To measure the LD decay rate in AM, different germplasms should be used. Therefore, the density of the marker is typically higher than that of linkage mapping (Álvarez et al. 2015). LD decay is slightly higher in Recombinant Inbred Lines (RIL) than in F2 populations. Nevertheless, the resolution power is lower than that of the AM population (Álvarez et al. 2015). The order of 5–10 cM is the resolution required to locate the QTL in linkage mapping, from which many genes within each QTL are present (Buckler IV and Thornsberry 2002). In addition, for those germplasm panels with low LD, the diagnostic power of a single marker will only extend a short way. Thus, high number of markers is needed for whole-genome scan. Additionally, the population in AM is obtained either by a strategy that is advantageous for sampling or breeding objectives, whereas in linkage mapping, the population structures are usually constructed and maintained. The breeding lines are challenging to keep in QTL mapping (Myles et al. 2009), but the germplasm accessions are supported adequately in AM due to the excessive number of alleles

contained in them. Generally, AM is an alternate to QTL mapping, that does not need the screening of progeny generation or the development of bi-parental crosses.

Limitations

The detection power of AM relies on the phenotype under study and the association of the marker locus (Álvarez et al. 2015). However, in most germplasm collections, relevant alleles are found and are frequently very significant sources of desirable alleles (Rafalski 2010). Likewise, in some germplasms, the existence of different individuals with diverse growing condition will be a barrier to its usage. As such, in association study, phenotypic evaluation for diverse germplasm must be given due considerations (Myles et al. 2009). Gupta et al. (2019) stated that identification of false positives or negatives and the issue of missing heritability are the main problems in GWAS. Recently, however, some approaches are employed to overcome such limitations. They include epigenetics, the use of expression profiles re-sequencing, identification of candidate genes and functional characterization using reverse genetic approaches such as gene silencing or retrotransposon-mediated gene disruption among others. Furthermore, analysis of rare alleles and rare variants are also among the approaches used to enhance GWAS. The advantage of association over linkage mapping in populations where LD is vast and does not occur rapidly across most of the genome appears to be common in many self-pollinating species. In this case, no dependable relationships can be attained among traits and for specific genes. However, the genome of the entire region is associated due to lack of haplotype chunk disintegration. Alternatively, some replacements that take advantage of AM still exist in such crops where there is low resolution due to restrictions to AM or high LD. Nevertheless, population structure and rare allele limitations can be overcome. This limitation can be achieved by crossing breeding lines to form a multi-parent population, from which functional allele combinations are identified and are used directly to identify marker x trait associations effectively (Kover et al. 2009). AM utilizing Q + K model has been modified to deal with large p, multiple testing and small n limitations in GWAS (Yu and Buckler, 2006). With the development of high throughput technology, haplotypes and SNP-sets (instead of single SNPs) are being used for GWAS, thereby overcoming the limitations of multiple testing and enhancing the identification of candidate genes which in turn facilitate gene-set-based and gene-based association mappings.

Genomic technology

The technology involved in manipulating and analyzing genomic information is referred to as genomic technology. It was initiated following the invention of DNA cloning in the 1970s (Galas and McCormack 2003). The availability of model species and their genome annotations as well as the application of genomic technology provide sequences for various complex traits and candidate genes for further association analysis (Zhu et al. 2008). Genome re-sequencing, reduced representation sequencing and pool-seq are very accessible and inexpensive approaches for population genomic studies (Therkildsen and Palumbi 2017). Targeting Induced Local Lesions in Genomes (TILLING, a method in molecular biology that allows direct identification of mutations in a specific gene) and Ecotype TILLING (EcoTILLING, a modification of TILLING technique that looks for natural mutations in individuals, usually for population genetics analysis) are among the genomic methods used for germplasm collections and screening allelic variant mutants in target genes. Genome re-sequencing is very useful for genome-wide discovery of markers for high-throughput genotyping, such as SNPs and SSRs or for the construction of high-density genetic maps. These, in turn, enhance genetic diversity study, and make identifications of markers linked to genes and QTL achievable via a variety of approaches including fine genetic mapping, bulked segregant analysis (BSA) and association mapping (M Perez-de-Castro et al. 2012). Currently, whole genome sequencing and characterization have been achieved using molecular markers and play a role in marker-assisted breeding (Song et al. 2010). High-density markers are needed to detect alleles that are involved in agronomic traits (Tardivel et al. 2014). Genetic improvement of complex characters (especially drought and salt tolerance) has now been achieved with genomic technology. Today, detection of specific genes is effectively accomplished at faster rate by combining marker-assisted selection (MAS) with genomic technology as compared to classical breeding (Saade et al. 2016).

Natural diversity

Introgression library (IL) and advanced backcross QTL (AB-QTL) techniques are used to explore natural diversity. They are used to remove alleles from germplasm to improve quality, productivity, nutritional value and adaptation of crops (Zamir 2001). Interestingly, large scale functional diversity of a crop species can be evaluated using AM, which makes it different from AB-QTL and IL (Breseghello and Sorrells 2006b). Generally, information derived from association mapping applies to a broader germplasm base while that of bi-parental mapping is specific to the same or genetically similar population (Zhu et al. 2008).

Different phases of association mapping

Five stages are involved in AM as illustrated in (Fig. 1): (i) Individuals for the **population are selected**, (ii) the selected

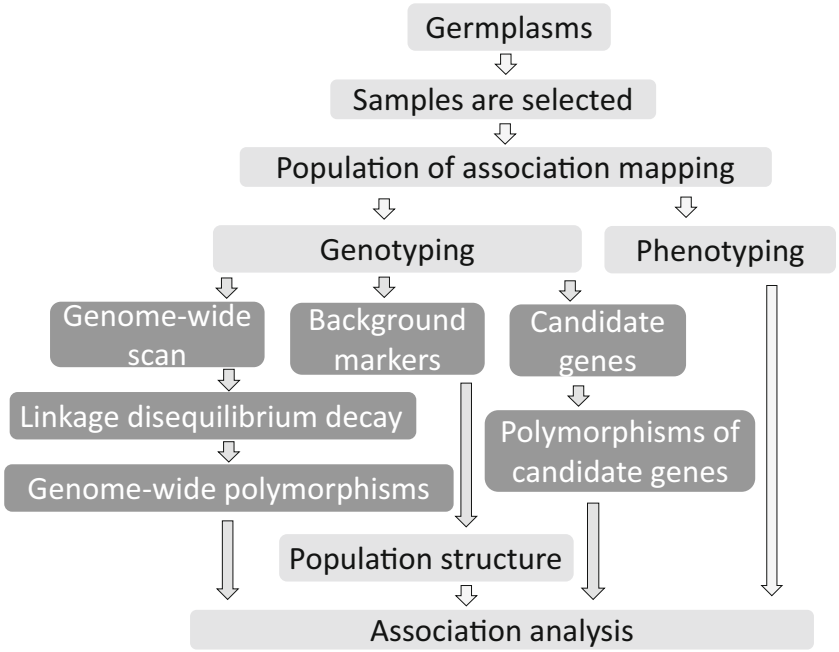

Fig. 1 Simple illustration of association mapping

population are **genotyped,** (iii**) Population structure** based on the genotyping is analyzed, (iv) phenotypic traits of interest among the population are characterized **(phenotyping)**, and lastly, (v) phenotype versus genotype relationships are determined **(Association analysis)**.

Selection

In the AM process, an important component to consider is the careful selection of the population, and the resolution power of the study would be better if more recombinations are observed. Generally, a diverse population must be considered rather than classified or structured (Álvarez et al. 2015). Association analysis can be successfully achieved by a careful selection of the population (Breseghello and Sorrells 2006b). Appropriate techniques for association analysis and the power of statistics to detect marker-phenotype association depend on the genetic germplasm diversity, the LD level in genome-wide association, and population structure level and population relatedness under study (Stich and Melchinger 2010).

Plant population can be classified in two dimensions: (a) the extent of population structure and (b) relatedness of the family (Yu and Buckler 2006). Based on these dimensions, the populations are further classified into the following categories: (i) ideal sample with familial relatedness and subtle population structure, (ii) multifamily without considering population structure, (iii) population structured sample that does not consider relatedness of the family, (iv) samples that consider relatedness of the family with that of population structure, and (v) severe population structured sample and relatedness of the family. Therefore, the existence of a population in one of the categories mentioned will determine the kind of statistical methods to be applied for the association analysis. Furthermore, AM populations can also be categorized according to the source of materials (Breseghello and Sorrells 2006b). These sources could be from germplasm collection bank, elite breeding lines, natural population, or synthetic population. These sources of population materials are expected to vary according to the extent of LD, genotypic and phenotypic diversity and the importance of structured population and relatedness of the family.

Genotyping

GWAS and candidate genes analysis are the two approaches used in AM (Fig. 1). However, the selection of each depends upon the amount of marker for the association. GWAS usually tests for an association that represents most of the segments of the genome, and considers genotype of population of individuals that are densely distributed across genetic marker loci covering all the chromosomes (Rafalski, 2010). However, in candidate gene association analysis, markers are chosen based on their location in the genome and based on previous QTL studies/functions of the genes involved that led to the final variation.

Unlinked neutral background markers are selectively mounted for successful coverage and are considered in association studies (Zhu et al. 2008). They have been actively engaged in characterizing the genetic composition of individuals. Additionally, these markers are highly beneficial in conveying individuals to populations (Pritchard and Rosenberg 1999). As such, population structure and relatedness limitations are overcome (Yu and Buckler 2006), and inbreeding and kingship are determined (Lynch and Ritland 1999).

Molecular markers easily trace genetic loci that can be computed in a population and may be related to a specific trait or gene of interest (Hayward et al. 2015). Generally, heritable differences within a population are affected by mutations in the form of translocation, inversion or insertion, which can be noticed and screened using molecular markers (Hayward et al. 2015). Markers can be used to identify the true uniqueness of individual plants. 1 cM distance is an ideal location for an active marker for MAS of the anticipated characteristic and is capable of high throughput and reproducibility genotyping (Mohan et al. 1997). AFLP and RAPD Markers have poor genomic distribution, reproducibility, and low polymorphism and these limit their application in MAS (Vos et al. 1995; Williams et al. 1990). They need unique statistical methods if intended to be used for estimating genetic population parameters. On the other hand, SNPs and SSR are profoundly revealed markers and are used in determining the relative kinship matrix and population structure which make them appear more powerful (Zhu et al. 2008). When calculating genetic parameters using SSR and in the presence of size homoplasy, high mutation rate and size of the alleles may be serious challenges especially if the population is large (Estoup et al. 2002). However, for a valid selection of genotyping technology (Syvänen 2005) SNP markers and scored individuals are required. The rate of mutation per generation in SNPs is shallow compared to that of SSR (Li et al. 2002). Consequently, the biallelic nature of SNPs makes them less informative than multiallelic SSRs. It should be noted that in SNPs, expected heterozygosity is lower (and therefore are required) than SSR background markers for the successful attainment of a practical assessment of population structure and family relatedness of most crops. Additionally, SNPs are distributed widely throughout the genome and inexpensive to score than SSRs. Wessinger et al. (2018) established that the effectiveness of detecting SNPs to explain phenotypic variation depends on some genetic factors of the population such as allele frequency of the population, size effect, sampling effects along with epistasis and genotype uncertainty. SNPs are the most heritable and fine mapping markers (Singh et al. 2001). Polymorphic markers of large sets can be screened through SNPs even in complex polyploid species and large-scale sequencing (Collard and Mackill 2007), and, as such, support genome-wide association studies.

Population structure

Changes in allele frequencies occur due to non-random mating within a species in population structure (Ersoz et al. 2007), which is considered a limiting factor in association mapping. It produces false positives (spurious associations), and it is complicated to follow up on perceived signals through expensive biochemical and independent studies as well as molecular analyses to replicate significant results (Zhu et al. 2008). Recently, approaches such as principal component analysis (PCA), mixed model approach and structured association and genomic control (GC), among others, are used to justify family relatedness and structure of the population (Price et al. 2006; Yu and Buckler 2006). False positives from population structure can be overcome through explicit (e.g., mixed model and SA) or ad hoc adjustment approaches (Zhao et al. 2007). To take care of problems arising from population structure in most association studies, structured association has recently emerged as a method of choice. For this, individuals in population substructures are calculated and assigned through random unlinked markers (Pritchard et al. 2000). Population structure is often calculated using STRUCTURE software (Pritchard et al. 2000), through which the proportion of an individual's genome that initiated from different inferred populations is calculated using Bayesian algorithm. Different groups of individuals are then clustered based on their genome classification. STRUCTURE 20 assumes that all individual population is in Hardy-Weinberg equilibrium and unrelated. The degree of population admixture of each individual is calculated through this program. Additionally, PCA is also used to estimate population structure as reported by Price et al. (2006), which is quicker and more effective than STRUCTURE (Zhao et al. 2007). Generalized linear model (GLM, one of the various structured association models) usually correlates genotypes with phenotypes using subpopulations (Q) as covariates in a regression model (Thornsberry et al. 2001). However, this may not control false positives even when used along with GC model (Yu and Buckler 2006). Subpopulations (Q) are usually assigned as covariates in a unified mixed-model (or Q + K model; K = kinship matrix); nevertheless, they use covariate in the regression as K (Yu and Buckler 2006). It has been concluded that Q + K is more advantageous than Q model (Zhao et al. 2007) in Arabidopsis studies and is therefore recommended in most GWAS. Currently, GWAS analysis can be achieved using Trait Analysis by Association, Evolution and Linkage (TASSEL) Software (Bradbury et al. 2007).

Linkage Disequilibrium (LD) decay

Non-random associations of alleles at diverse loci are referred to as LD (Oraguzie et al. 2007), through which the resolution of AM studies is determined. It should be noted that the resolution is expected to be very high when the LD decays are displayed in a short distance, even though they need many markers. Additionally, mapping resolution will be low when the LD spreads in long-distances, but it requires only a few markers here. Generally, low resolutions reveal high LD, and vice versa. Many factors affect LD. They include population subdivision and population size, genetic isolation among lineage, recombination rates, mutation and amount of inbreeding, among others (Mackay and Powell 2007; Gupta et al.

2005). Under linkage disequilibrium, variation between the observed and expected gamete haplotype frequencies is measured as LD (Soto-Cerda and Cloutier 2012). Graphics' view of LD is used to present r^2 (the Pearson's squared value (product-moment) correlation coefficient) over genetic distances among polymorphic sites (Bradbury et al. 2007) within the loci/gene along a chromosome (Bradbury et al. 2007). r^2 is usually preferred to decay plot D when measuring LD through pair-wise measurements between markers due to fewer biases (Soto-Cerda and Cloutier 2012). The LD decay rate over distances must be understood for the most straightforward determination of the number of markers that would be required for GWAS. The number of marker required to saturate the genome for GWAS should be known prior to measuring the LD in AM. LD is used to identify genomic areas of the candidate related to a specific character or diseases and can offer a more exceptional resolution more than that of linkage-based mapping (Mackay and Powell 2007). Generally, to quantify genetic diversity, LD is used and can easily be explored to make extrapolation about the populations' evolutionary history (Zhu et al. 2015; Slatkin 2008). Genetic drift, population growth, admixture (introduction of genes from a previously distinct population to another) or migration, population structure, natural selection, gene conversion, variable recombination, and rate of mutation are among the factors that influence linkage disequilibrium (Ardlie et al. 2002). Additionally, the level of LD in a varied population will be determined from the species mode of reproduction (Flint-Garcia et al. 2003). Crops, such as self-pollinated ones, have much longer LD distances than cross-pollinated ones (like wheat and maize, respectively). For LD generated by population structure, there should be careful consideration of the sample to avoid faulty analyses of the results (Ersoz et al. 2007). Bilton et al. (2018) reported that evolutionary and genetic forces affect LD, as such, its pattern is utilized in computing genetic diversity and can make inferences about the evolutionary history of natural populations (Bilton et al., 2018). In addition, the association between the map distance and LD level can be used to estimate adequate population size (Sved et al. 2013; Waples 2006). Sequencing data of low coverage that accounts for under-called heterozygous genotypes was used to calculate pairwise disequilibrium by establishing new likelihood methods and Genotyping Uncertainty with Sequencing data - Linkage Disequilibrium (GUS-LD) (Bilton et al. 2018). The authors concluded that using GUS-LD, reliable estimates were obtained whereas the results will be underestimated for linkage disequilibrium if no adjustment is made for the errors. Many authors studied gene controls by one or few loci with significant effects, especially in areas concerning the biochemical basis of essential phenotypes like abiotic and biotic stress tolerance. These phenotypes have greater impacts on enhancing crop production, especially in MAS breeding (Foolad and Panthee 2012). But, complex trait differences have proven to be very difficult to understand, as the genetic architecture of these essential traits (especially salt and drought tolerance) involves many loci with small effects associated with one another and the environment (Buckler et al. 2009; Collard and Mackill 2007). Grouping of statistical tools is now being used to distinguish such small effects. Among them, LD is used to survey genetic variances with a limited resolution to a mapping population rather than the density of the marker. The relationship between polymorphisms in a population is usually stated by LD. Myles et al. (2009) stated that the distance between any two markers functionally relies upon the strength of the relationship between them (Myles et al., 2009). The faster the rate of LD decays over distance signifies how far the resolution from which QTL can be mapped. Thus, the first stage in the design of AM studies could be structural analysis of LD.

Identification of candidate genes

Genes that are indirectly or directly affecting the developmental process of characters with known biological functions that can be valid by assessing the effects of the causative gene differences in association analysis are referred to as candidate genes (Zhu and Zhao 2007). It has been used and applied for genetic association studies, research for biomarkers, gene-disease and drug target selection in many organisms from animals to humans (Tabor et al. 2002). Apart from genome scan, candidate gene analysis is also used for position cloning of QTL regulating main genetic differences of characters of interest. It should be noted that the causative genes are the QTL that show significance in a region of chromosome affecting the genetic variations of characters under study. This region of QTL consists of several genes gathered at about ~20 cM confidence interval (Zhu and Zhao 2007). The highest resolution power for mapping QTL and in LD was offered by SNPs with the causative polymorphism; for this reason, they are usually prepared as a candidate-gene variant to genotype in AM (Rafalski 2002). SNPs within specific genes and between line identifications are necessary for candidate-gene AM. Consequently, candidate gene SNPs identification procedure depends on the amplicons resequencing from numerous individuals that are genetically diverse from a larger association population and within specific Genes (Zhu et al. 2008). Generally, to identify rarer SNPs, individual SNP panel is required while in identifying common SNPs, fewer are needed. For identifying a candidate gene, promoter SNPs, exon, intron and untranslated 5′/3′ regions are all reasonably targeted, with coding regions that have less level of nucleotide diversity than the non-coding part (Zhu et al. 2008). The SNPs number per unit length required to detect significant associations is dictated by a candidate gene locus, which depends on the rate of LD decay (Flint-Garcia et al. 2003). Hence, the Locus of the candidate gene is entirely reliant on SNP

distribution and LD as well as amplicon numbers and the pair-based length needed to sample it adequately. Seven hundred and thirteen upland kinds of cotton (*Gossypium hirsutum* L.) accessions of a natural population were evaluated for salt tolerance-related characters (Sun et al. 2018). From the GWAS result, the authors obtained seven genomic regions that were represented from 23 SNPs. Salt-tolerance and survival rate are among the significantly associated traits. These traits were simultaneously related to two SNP markers on the D09 chromosome (i47388Gh and i46598Gh). Two hundred and eighty possible candidate genes were also screened based on all loci under salt stress (Sun et al. 2018). Genes such as MYB, NAC, WD40, NXH, CDPK, CIPK and LEA are involved in plant salt tolerance and are transporters that participate in numerous enzymatic and transcriptional activities. Because of the limitation of including all causative genes and low repetition of results, the candidate gene approach has been disapproved (Tabor et al. 2002). The digital candidate gene approach (DigiCGA) has been developed to overcome some bottleneck limitations for successful detection of candidate genes in some studies (Zhu and Zhao 2007).

Phenotyping

In association mapping, diverse accessions must be relatively needed in large numbers, thereby making it challenging while taking the phenotypic replicated data across environments and years. However, for the inhomogeneous field, careful consideration of QTL x environment interactions, employing incomplete block design and appropriate statistical methods enhance mapping power (Eskridge 2003). Influences of variabilities within and between years, environments and seasons may complicate trait phenotyping for G x interactions (Atlin et al. 2011). For abiotic stress responses in plants, further improvement is achieved under controlled environment (Negin and Moshelion 2017). Nevertheless, observations of the actual field conditions under this environment, particularly in drought, are challenging (Passioura 2012). Phenotyping in association mapping has not been given much consideration compared to genotyping (Zhu et al. 2008). For large-scale association mapping, obtaining vigorous phenotypic data remains very difficult. Since AM often comprises relatively large and diverse accessions, phenotypic data collection with enough replications across multiple locations and years is equally challenging. Therefore, the experimental area must be effectively laid out with latex design (incomplete block design) due to its potential to increase the mapping power (Piepho et al. 2006). In addition, if unbalanced plant breeding trials are used as sources of phenotypic data, appropriate statistical modeling of the experimental design as well as genotype x environment and marker x environment interactions, must be taken into consideration (Malosetti et al. 2008). As such, mapping power will be increased (Stich et al. 2008). As stated by Cobb et al. (2013) for reliable phenotyping approaches based on quantitative measurements, proper quantitative characterization is needed to dissect genetic differences precisely. Heritability is usually calculated individually to understand the ratio of genetic variances explained by the detected QTL. Some phenomics systems have been established and used for some data like biomass content, photosynthesis, pigment content and attributes of the canopy using rapid and guided-GPS (Simko et al. 2016), responses due to abiotic stress factors (Cobb et al. 2013), flowering (Guo et al. 2015) and pathogenesis (Mahlein 2016). Phenotypic variation relationships under field and control environments must be observed critically so that important information is provided to enhance phenotyping techniques in the control environment.

Statistical analysis

The most straightforward statistical approach for association analysis of quantitative traits is the analysis of variance (Yu and Buckler 2006). However, restrictions of AM studies, especially arising from population structure quantitative transmission disequilibrium test (QTDT), were modified to apply to inbred populations of plants (Stich et al. 2006). Genomic control and structured association are now in use for both human and plant association studies for population-based samples. Random effects such as multiple background QTL and population membership estimates of Q-matrix are combined in a mixed model for correction of false association at the same vain, considering covariances due to relatedness (Bradbury et al. 2007). Kinship (K) derived from random markers or pedigree can be used to estimate the average relationship between individuals. However, the most effective one is that which combined both Q and K (Yu and Buckler 2006). In population structure diagnosis, PCA is used for genetic diversity study in an association mapping context (Patterson et al. 2006). In structure association analysis, the implemented Q method has been utilized in GLM function in TASSEL software. STRUCTURE program and PCA have been used to derive covariates in the model using population membership estimates (Pritchard et al. 2000; Zhao et al. 2007). To calculate the structure of the population and use the outcome for further analysis, a set of random markers must be initially utilized in structured association (Falush et al. 2003; Pritchard and Rosenberg 1999). Logistic regression has been used for the modified structured association (Thornsberry et al. 2001). Chhatre (2013) reported the use of StrAuto v0.3.1. It is a Python-based structure with an automated procedure software for Linux-based computers, and is recently been utilized for (i) discovery of genetic structure in sample populations for medical purposes (Pritchard and Donnelly 2001); (ii) population structure studies (Randi and Lucchini 2002); and (iii) detection of cryptic genetic structure of natural populations (Caizergues et al. 2003). PCA and Multiple Correspondence

Analysis (MCA) are performed for 3D or 2D space to observe the relative distribution of subpopulation (Rahim et al. 2018). They require less computing time than maximum likelihood estimation. Therefore, Rahim et al. (2018) concluded that PCA and discriminate analysis are the most frequently used analytical techniques in population structure analysis (Rahim et al., 2018). However, STRUCTURE is frequently used in Bayesian clustering method.

Software packages

A number of software/statistical packages have been used in AM studies. Theses include TASSEL, Statistical Analysis System (SAS), R package, STRUCTURE, Spatial Pattern Analysis of Genetic Diversity (SPAGeDi), EINGENSTRAT, Multiple Trait Derivative-Free Restricted Maximum Likelihood (MTDFREML), and Residual Maximum Likelihood (ASREML) (Zhu et al. 2008). Additionally, STRAT, Bimbam and GEN STAT 11 software have been added recently (Álvarez et al. 2015). Summary-data-based Mendelian randomization (SMR) and heterogeneity in dependent instru-ments (HEIDI) tools have been used to test pleiotropic interaction between gene expression level and complex traits using expression quantitative trait loci (eQTL) and GWAS data (Zhu et al. 2016). Moreover, these tools can be employed to assess the size of the effect of SNP on phenotype mediated by the expressed gene.

Application of association mapping in plant breeding

AM sustains breeding practices that capture superior alleles and support their introgression into elite breeding germplasm from diverse individuals. It is noted that most studied characters are abiotic stresses, quality, yield, and morphological parameters (Table 1). Liu et al. (2018) identified 122 and 134 QTL for yield-related traits and fiber quality in cotton, respectively (Liu et al., 2018). The same authors also identified 139 quantitative trait nucleotides (QTNs) for yield components and 209 QTNs for fiber quality among which 74 were observed in two environments using GWAS. Four were possibly "pleiotropic" among the 35 common candidate genes observed. Patishtan et al. (2018) used a panel of 306 diverse rice accession to perform GWAS and identified transcription factors and components of the ubiquitination pathway as an important source of genetic diversity (Patishtan et al. 2018). RD2, HAT22, PIP2 and PP2C genes were proposed to be potentially significant for drought tolerance in cotton using RNA-seq and were verified through a Quantitative reverse transcription-polymerase chain reaction (qRT-PCR) (Hou et al. 2018). Resende et al. (2018) carried out Regional heritability mapping (RHM) and GWAS for lodging, productivity, and plant architecture across two environments using 188 common bean germplasms. The study detected three trait-associated markers using GWAS, whereas 145 markers along chromosomes 5 with eight QTL were identified using RHM. The authors concluded that combining allelic differences of QTL with the large effect could be successfully combined into whole-genome prediction models and can easily be traced using marker-assisted selection. Identification of salt tolerance loci in rice was also carried using GWAS, where Na+/ K+ ratios with the major association were measured at the reproductive stages and were equally detected and found to contain saltol as the major QTL on chromosome I at the seedling stage, regulating salinity tolerance (Kumar et al. 2015). Maulana et al. (2018) Mapped QTL and identified SNP associated with seedling heat tolerance in wheat. Their findings revealed some effective QTL that are tolerant to heat from seedling to reproductive stages. Interestingly, however, new QTL that have never been reported previously at the reproductive stage were found responding to seedling heat stress. Analysis of candidate genes also indicated high sequence resemblances of some loci with candidate genes involved in plant stress responses, such as salt, heat and drought stresses. Su et al. (2018) determined the genetic basis of cotton plant architecture using GWAS, from which 30 significant relations among five-plant architecture and 22 SNP markers were identified. Additionally, more plant architecture component traits concurrently associated with chromosome D03 with four peak SNPs were identified. 37,901 SNP markers in switchgrass were obtained and utilized for GWAS (Taylor et al. 2018). *Arabidopsis* pseudo-response regulator 5 homolog was related to heading date across environments and years on chromosome 8a. The study found that genetic deviations associated with floral enhancement influence the dates of flowering and productivity. Significant quantitative trait SNP markers comprising about 87, 21 and 16 for fatty acid, oil and proteins, respectively, were identified (Du et al. 2018). Protein contents have been controlled by epistasis influence, accounting for a total variation of about 65.18%. However, 16 chromosomes containing 20 QTNs were found to contribute to six-drought tolerance. Moreover, Messenger RNA (mRNA) expression levels of the genes were verified in the target interval through which the potential loci/genes that regulated branch number in *Brassica napus* expression were identified (He et al. 2017). Two SNP markers i47388Gh and i46598Gh on chromosome D09 were found to be associated with salt tolerance level and relative survival rate in cotton, respectively (Sun et al. 2018). Additionally, different expression levels of about 280 candidate genes under salt stress were screened, from which CIPK, NXH, MYB, LEA, WD40 and CDPK genes were responsible for plant salt tolerance. Most of these genes are transcription factors, transporters or enzymes. SNP markers and QTL were identified that could effectively

Table 1 Examples of AM studies in various plant species

Plant species	Populations	Sample size	Background Markers	Traits	References
Cotton	Diverse germplasm	319	55,060 SNPs	drought tolerant	(Hou et al. 2018)
Sorghum	Diverse lines	648	183,989 genotype by sequence markers	drought tolerant	(Spindel et al. 2018)
common bean	Diverse germplasm	188	17,850 DArTseq	plant architecture, lodging, and productivity	(Resende et al. 2018)
Rice	Diverse germplasm	220	6000 SNPs	Salinity tolerance	(Kumar et al. 2015)
Wheat	Diverse representative lines	200	21,555 SNP	Heat tolerance	(Maulana et al. 2018)
Rice	Diverse accessions	306	700,000 SNPs	Salinity tolerance	(Patishtan et al. 2018)
Cotton	RIL population	231	122 SSR and 4729 SNP	fiber quality traits and yield components	(Liu et al. 2018)
Cotton	Diverse accessions	355	93,250 SNPs	the genetic basis of cotton plant architecture	(Su et al. 2018)
Barley	Diverse accessions	206	408 Diversity arrays technology (DArT)	Salinity tolerance	(Fan et al. 2016)
Brassica napus	Diverse accessions	327	33,186 SNPs	branch number	(He et al. 2017)
Cotton	Diverse germplasm	713	10,511 SNPs	Salinity tolerance	(Sun et al. 2018)
Rice	MAGIC Plus lines	144	14,242 SNP	Agronomic and bio-fortification traits	(Descalsota et al. 2018)
Switchgrass (*Panicum virgatum*)	Four pseudo-F2 populations (two pairs of reciprocal crosses)	588 tetraploid genotypes.	37,901 single nucleotide polymorphisms	heading and anthesis	(Taylor et al. 2018)
Rice	core germplasm collection	419	261,385,070 SLAF-seq	The genetic basis of Gelatinization temperature (GLT), gel consistency and pericarp colour (PC)	(Yang et al. 2018)
Cotton	Accessions	316	390 K SNPs	protein, oil and five fatty acids	(Du et al. 2018)

be used for bio-fortification and breeding disease resistance in rice (Descalsota et al. 2018). Breeding material is used directly in genetic studies, such as recurrent selection or multiple cross pedigree programs. Nevertheless, for greater promising, Marker Assisted Recurrent Selection (MARS) is used. However, AB-QTL methods have been used for introgression and genetic study for commercial production. Geneticist, breeders and statisticians used breeding lines and populations, from which they came-up with models for whole-genome selection that are enhanced at each consecutive generation, season and phenotyping exercise based on whole-genome haplotypes rather than individual gene evaluation, also known as genomic selection (GS). Since GS uses statistical modeling coupled with high-throughput markers, the system has eventually transformed MAS. The GS strategy was recommended in 2001 from several reports of statistical models (Hayes and Goddard 2001). It has been used for enhancing preselection precisions, especially using genomic information for complex agronomic traits. The GS also uses data from genotypic and phenotypic training population (TP), which can be used to calculate genomic estimated breeding values (GEBVs) for accurate selection of each individual from the breeding population that is genotyped without phenotyping (Jonas and de Koning 2013). All marker effects can be directly estimated, and such loci with minor effects for complex characteristics can be easily captured in the whole genome as the main advantage of GS over others (Nakaya and Isobe 2012). Additionally, the rate of annual genetic gain can be significantly enhanced by reducing time, accelerating breeding cycles and cost because selection depends on an individual's genotypes deprived of the required records of the phenotype (Xu et al. 2017). To assess the performance of breeding program in genomic selection, prediction accuracy (rMG) is estimated as Pearson's correlation (r) between the GEBVs of candidate individuals and the true breeding value. The prediction ability of GS is usually affected by many factors that directly influence the accuracy of GEBV. These include population structure, marker density, performances of the model, association between breeding population, target trait heritability, and size of the population of both TP and breeding population (BP). rMG also varies with statistical models of GS (Endelman 2011; Gianola 2013; Juliana et al. 2017; Ornella et al. 2014; VanRaden 2008).

About 94 peach germplasm collections were used by Font et al. (2019) and 347 significant associations were identified between markers and traits, which appeared mapped within the interval where many candidate genes are involved in different pathways. Zhang and Yuan (2019) conducted AM and GP (genomic prediction) analyses using 300 inbred lines of maize from different collection zones. They found out that 1549 SNPs were significantly correlated to 12 trait-environment combinations; the PVE of these significant SNP was about 4.33%, and 541 of them had a phenotypic variance explained (PVE) value greater than 5%. They observed fewer numbers of significant associations and candidate genes with higher PVE values in haplotype-based association mapping than the single SNP-based association mapping. Arab et al. (2019) explored the genomic differences and population structure of *Persian walnut*, from which loci underlying the variation in kernel and nut-related traits were identified using the new Axiom *J. regia* 700 K SNP genotyping arrays. Moreover, they uncovered 55 significant SNPs associated with kernel and nut-related traits. Gao et al. (2019) also identified 17 genes and 4 QTL correlated to 42 significant SNPs associated with thermos tolerance of seed-set by GWAS and linkage mapping, respectively.

Future perspective

Association genetics studies in plants are still in progress; and appropriate phenotyping methods, development of error free statistical software and accessibility to genotyping still remain the major challenges to its effectiveness, despite series of improvements to bridge the AM studies gaps and enhance its efficacy in crop breeding and genetics development. It must be noted that the use of AM in dissecting QTL for evolutionary population studies requires full information about the organism to identify the number markers needed (Álvarez et al. 2015). If the knowledge about recombinational history in breeding populations is known for several population types, The effectiveness of AM studies will be maximized. Additionally, GS and AM phenotyping remain challenging due to the need to capture the right phenotype and differences that occur in different material or breeding programs (Álvarez et al. 2015). Association cannot be found within a single locus particularly when population structure and morphological characteristics are correlated; however, associated prediction with epigenetic interactions and multiple loci can easily be improved due to GS approaches (Jannink et al. 2010). Where variability due to phenotype is established within subpopulations and candidate genes are known, marker density is adequate and AM approaches will be successfully implemented. The low-marker density limitations in GWAS can be overcome by increasing the marker numbers for all crops, although this depends on the types of marker selected in relation to representation and gene distribution in space and LD level. However, large number of markers is not required in AM studies. In the future, AM approaches should look at improvements in computational and statistical methods (such as SNP imputation, Bayesian and haplotypes methods) and their integration with gene annotation data or functional analysis (Zhang et al. 2014). Additionally, advances in crop genome re-sequencing and the expansions of mammalian and other model organisms will influence GWAS (Visscher et al. 2017). Generally, statistical tools that are user-friendly and genomics resources need to be improved. While applying AM, all factors like the population size, the density of marker as well as population structure, should be taken into consideration. For the detecting marker-phenotype relationship, the choice of germplasm, quality of genotypic and phenotypic data, use of the appropriate statistical analysis and verification of the marker-phenotype associations are key to association analysis. To harness the linkage-based QTL mapping and AM, joint linkage association mapping is proposed. Bayesian regression method can be used to overcome the genome-wide error rate (GWER), and it is expected to be used more frequently in GWAS, especially if artificial intelligence networking is involved. It was observed that markers with rare alleles in GWAS are often excluded from the analysis that attributed to missing heritability; as such, rare allele/variant analysis will be an important area to be considered to enhance AM studies.

Conclusion

AM is a tool used in plant breeding and genetics to comprehend QTL location and ascertain and monitor essential characters. It provides a vast prospect to assess and discover diversity of plant species for modern agricultural production. Many loci controlling the traits of interest escape detection and failure to identify the loci from similar parents are among the limitations of linkage based mapping. However, integrating it with AM produces high-resolution power and multiple alleles can be tested easily in the same experiment. Additionally, PCA and discriminate analysis are suitable for population structurer while STRUCTURE is recommended for Bayesian clustering method.

Acknowledgments National Natural Science Foundation of China (31771369) and the China Agriculture Research System (CARS-19-E06) supported this work.

Authors' Contributions AKI was responsible for literature search and writing the draft; LZ conceived the idea for the work and revised the draft; SN, MZ, YX, LZ, LMZ, and JQ did critical revision of the manuscript.

Compliance with Ethical Standards

Conflict of Interest The authors declare no conflict of interest.

References

Álvarez MF, Mosquera T, Blair MW (2015) The use of association genetics approaches in plant breeding Reviews 38:17–68

Arab MM, Marrano A, Abdollahi-Arpanahi R, Leslie CA, Askari H, Neale DB, Vahdati K (2019) Genome-wide patterns of population structure and association mapping of nut-related traits in Persian walnut populations from Iran using the axiom J. Regia 700K SNP array. *Scientific reports 9*(1):6376

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3:299

Arora S et al (2019) Resistance gene cloning from a wild crop relative by sequence capture and association genetics. Nature biotechnology 37:139

Atlin G, Kleinknecht K, Singh K, Piepho H (2011) Managing genotype x environment interaction in plant breeding programs: a selection theory approach. Journal of the Indian Society of Agricultural Statistics 65:237–247

Bilton TP, McEwan JC, Clarke SM, Brauning R, van Stijn TC, Rowe SJ, Dodds KG (2018) Linkage disequilibrium estimation in low coverage high-throughput sequencing data. Genetics 209:389–400

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635

Breseghello F, Sorrells ME (2006a) Association analysis as a strategy for improvement of quantitative traits in plants. Crop Science 46:1323–1330

Breseghello F, Sorrells ME (2006b) Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. Genetics 172:1165–1177

Buckler ES et al (2009) The genetic architecture of maize flowering time. Science 325:714–718

Buckler ES IV, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. Current opinion in plant biology 5:107–111

Caizergues A, Bernard-Laurent A, Brenot JF, Ellison L, Rasplus J (2003) Population genetic structure of rock ptarmigan *Lagopus mutus* in Northern and Western Europe. Molecular Ecology 12:2267–2274

Chhatre VE (2013) Population structure, association mapping of economic traits and landscape genomics of East Texas loblolly pine (Pinus taeda L.). Texas a&M University,

Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement Theoretical and Applied Genetics 126: 867–887

Collard BC, Mackill DJ (2007) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences 363:557–572

Descalsota GIL et al. (2018) Genome-wide association mapping in a rice MAGIC Plus population detects QTL and genes useful for biofortification Frontiers in plant science 9

Du X et al (2018) Dissection of complicate genetic architecture and breeding perspective of cottonseed traits by genome-wide association study. BMC genomics 19:451

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP the plant. Genome 4:250–255

Ersoz ES, Yu J, Buckler ES (2007) Applications of linkage disequilibrium and association mapping in crop plants. In: Genomics-assisted crop improvement. Springer, pp 97–119

Eskridge K (2003) Field design and the search for quantitative trait loci in plants. Available,

Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Molecular ecology 11:1591–1604

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Fan Y, Zhou G, Shabala S, Chen Z-H, Cai S, Li C, Zhou M (2016) Genome-wide association study reveals a new QTL for salinity tolerance in barley (Hordeum vulgare L.). Frontiers in plant science 7: 946

Font I Forcada C, Guajardo V, Reyes Chin Wo S, Moreno Sánchez MÁ (2019) Association mapping analysis for fruit quality traits in *Prunus persica* using SNP markers. Front Plant Sci 9:2005

Flint-Garcia SA, Thornsberry JM, Buckler ES IV (2003) Structure of linkage disequilibrium in plants. Annual review of plant biology 54:357–374

Foolad MR, Panthee DR (2012) Marker-assisted selection in tomato breeding. Critical reviews in plant sciences 31:93–123

Galas DJ, McCormack SJ (2003) An historical perspective on genomic technologies

Gao J, Wang S, Zhou Z, Wang S, Dong C, Mu C et al (2019) Linkage mapping and GWAS reveal candidate genes conferring thermotolerance of seed-set in maize. J Exp Bot 70:4849–4864

Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194:573–596

Guo W, Fukatsu T, Ninomiya S (2015) Automated characterization of flowering dynamics in rice using field-acquired time-series RGB images. Plant methods 11:7

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant molecular biology 57:461–485

Gupta PK, Kulwal PL, Jaiswal V (2019) Association mapping in plants in the post-GWAS genomics era. Advances in genetics 104:75-154

Hartwig B, James GV, Konrad K, Schneeberger K, Turck F (2012) Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. Plant physiology 160:591–600

Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Hayward AC, Tollenaere R, Dalton-Morgan J, Batley J (2015) Molecular marker applications in plants. In: Plant Genotyping. Springer, pp 13–27

He Y et al (2017) GWAS, QTL mapping and gene expression analyses in Brassica napus reveal genetic control of branching morphogenesis. Scientific reports 7:15971

Hou S et al. (2018) Genome-wide association studies reveal genetic variation and candidate genes of drought stress-related traits in cotton (*Gossypium hirsutum* L) Frontiers in plant science 9

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature genetics 44:226

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics 9:166–177

Jonas E, de Koning D-J (2013) Does genomic selection have a future in plant breeding? Trends Biotechnol 31:497–504

Juliana P et al. (2017) Comparison of models and whole-genome profiling approaches for genomic-enabled prediction of Septoria tritici blotch, Stagonospora nodorum blotch, and tan spot resistance in wheat the plant genome

Kover PX et al (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana* PLoS genetics 5: e1000551

Kumar V, Singh A, Mithra SVA, Krishnamurthy SL, Parida SK, Jain S, Tiwari KK, Kumar P, Rao AR, Sharma SK, Khurana JP, Singh NK, Mohapatra T (2015) Genome-wide association mapping of salinity tolerance in rice (Oryza sativa). DNA Res 22:133–145

Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions, and mutational mechanisms: a review. Mol Ecol 11:2453–2465

Liu R et al. (2018) GWAS analysis and QTL identification of fiber quality traits and yield components in upland cotton using enriched high-density SNP markers Frontiers in Plant Science 9

Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. Genetics 152:1753–1766

Perez-de-Castro AM et al (2012) Application of genomic tools in plant breeding. Current genomics 13:179–195

Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends in plant science 12:57–63

Mahlein A-K (2016) Plant disease detection by imaging sensors–parallels and specific demands for precision agriculture and plant phenotyping. Plant Dis 100:241–251

Malosetti M, Ribaut JM, Vargas M, Crossa J, Van Eeuwijk FA (2008) A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L). Euphytica 161:241–257

Maulana F, Ayalew H, Anderson JD, Kumssa TT, Huang W, Ma X-F (2018) Genome-Wide Association Mapping of Seedling Heat Tolerance in Winter Wheat Frontiers in plant science 9

Mohan M, Nair S, Bhagwat A, Krishna T, Yano M, Bhatia C, Sasaki T (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. Molecular breeding 3:87–103

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell 21:2194–2202

Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? Ann Bot 110:1303–1316

Negin B, Moshelion M (2017) The advantages of functional phenotyping in pre-field screening for drought-tolerant crops. Functional Plant Biology 44:107–118

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS One 7:e37558

Nordström KJ et al (2013) Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. Nature biotechnology 31:325

Oraguzie NC, Wilcox PL, Rikkerink EH, De Silva HN (2007) Linkage disequilibrium. In: Association mapping in plants. Springer, pp 11–39

Ornella L et al (2014) Genomic-enabled prediction with classification algorithms. Heredity 112:616

Passioura J (2012) Phenotyping for drought tolerance in grain crops: when is it useful to breeders? Funct Plant Biol 39:851–859

Patishtan J, Hartley TN, Fonseca de Carvalho R, Maathuis FJ (2018) Genome-wide association studies to identify rice salt-tolerance markers. Plant, cell & environment 41:970–982

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS genetics 2:e190

Piepho H, Büchse A, Truberg B (2006) On the use of multiple lattice designs and α-designs in plant breeding trials. Plant breeding 125: 523–528

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38:904

Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. Theoretical population biology 60:227–237

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. The American Journal of Human Genetics 65:220–228

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. The American Journal of Human Genetics 67:170–181

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Rafalski JA (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13:174–180

Rahim MS, Sharma H, Parveen A, Roy JK (2018) Trait mapping approaches through association analysis in plants. In: Plant Genetics and Molecular Biology. Springer, pp 83–108

Randi E, Lucchini V (2002) Detecting rare introgression of domestic dog genes into wild wolf (Canis lupus) populations by Bayesian admixture analyses of microsatellite variation. Conserv Genet 3:29–43

Resende RT et al (2018) Genome-wide association and regional heritability mapping of plant architecture, lodging and productivity in Phaseolus vulgaris G3: genes. Genomes, Genetics 8:2841–2854

Saade S et al (2016) Yield-related salinity tolerance traits identified in a nested association mapping (NAM) population of wild barley. Scientific reports 6:32586

Simko I, Hayes RJ, Furbank RT (2016) Non-destructive phenotyping of lettuce plants in early stages of development with optical sensors. Frontiers in plant science 7:1985

Singh S et al (2001) Pyramiding three bacterial blight resistance genes (xa5, xa13, and Xa21) using marker-assisted selection into indica rice cultivar PR106. Theoretical and Applied Genetics 102:1011–1015

Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485

Song Q, Jia G, Zhu Y, Grant D, Nelson RT, Hwang EY, Hyten DL, Cregan PB (2010) Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCSOYSSR_1. 0) in soybean. Crop Sci 50:1950–1960

Soto-Cerda BJ, Cloutier S (2012) Association mapping in plant genomes Genetic Diversity in Plants:29–54

Spindel JE, Dahlberg J, Colgan M, Hollingsworth J, Sievert J, Staggenborg SH, Hutmacher R, Jansson C, Vogel JP (2018) Association mapping by aerial drone reveals 213 genetic associations for Sorghum bicolor biomass traits under drought. BMC Genomics 19:679

Stich B, Melchinger AE (2010) An introduction to association mapping in plants. CAB Rev 5:1–9

Stich B, Melchinger AE, Piepho H-P, Heckenberger M, Maurer HP, Reif JC (2006) A new test for family-based association mapping with inbred lines from plant breeding programs. Theoretical and applied genetics 113:1121–1130

Stich B, Möhring J, Piepho H-P, Heckenberger M, Buckler ES, Melchinger AE (2008) Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754

Su J et al (2018) Genome-wide association study identified genetic variations and candidate genes for plant architecture component traits in Chinese upland cotton. Theoretical and applied genetics 131:1299–1314

Sun Z et al. (2018) Identification of SNPs and candidate genes associated with salt tolerance at the seedling stage in cotton (*Gossypium hirsutum* L) Frontiers in plant science 9

Sved JA, Cameron EC, Gilchrist AS (2013) Estimating effective population size from linkage disequilibrium between unlinked loci: theory and application to fruit fly outbreak populations. PLoS One 8: e69078

Syvänen A-C (2005) Toward genome-wide SNP genotyping. Nature genetics 37:S5

Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Reviews Genetics 3:391

Tardivel A, Sonah H, Belzile F, O'Donoughue LS (2014) Rapid identification of alleles at the soybean maturity gene E3 using genotyping by sequencing and a haplotype-based approach the plant genome 7

Taylor MS et al. (2018) Genome-Wide Association Study in Pseudo-F2 Populations of Switchgrass Identifies Genetic Loci Affecting Heading and Anthesis Dates Frontiers in plant science 9:1250

Therkildsen NO, Palumbi SR (2017) Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Molecular ecology resources 17:194–208

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES IV (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28:286–289

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 101:5–22

Vos P, Hogers R, Bleeker M, Reijans M, Lee T, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conservation Genetics 7:167

Wessinger CA, Kelly JK, Jiang P, Rausher MD, Hileman LC (2018) SNP-skimming: A fast approach to map loci generating quantitative variation in natural populations. Molecular ecology resources 18:1402–1414

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic acids research 18:6531–6535

Wu R, Zeng ZB (2001) Joint linkage and linkage disequilibrium mapping in natural populations. Genetics 157(2):899–909

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

Xu Y, Li P, Zou C, Lu Y, Xie C, Zhang X, Prasanna BM, Olsen MS (2017) Enhancing genetic gain in the era of molecular breeding. J Exp Bot 68:2641–2666

Yang X et al (2018) Identification of candidate genes for gelatinization temperature, gel consistency and pericarp color by GWAS in rice based on SLAF-sequencing. PloS one 13:e0196690

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Current opinion in biotechnology 17:155–160

Zamir D (2001) Improving plant breeding with exotic genetic libraries. Nature reviews genetics 2:983

Zhang X, Bailey SD, Lupien M (2014) Laying a solid foundation for Manhattan–'setting the functional basis for the post-GWAS era'. Trends in Genetics 30:140–149

Zhang X, Yuan Y (2019) Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and flowering time under drought and heat stress conditions in maize. Front Plant Sci 9:1919

Zhao K et al (2007) An Arabidopsis example of association mapping in structured samples. PLoS genetics 3:e4

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. The plant genome 1:5–20

Zhu M, Zhao S (2007) Candidate gene identification approach: progress and challenges. an International journal of biological sciences 3:420

Zhu X, Xu F, Zhao S, Bo W, Jiang L, Pang X, Wu R (2015) Inferring the evolutionary history of outcrossing populations through computing a multiallelic linkage–linkage disequilibrium map. Methods in Ecology and Evolution 6:1259–1269

Zhu Z et al (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics 48:481