Current Plant Biology xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

Identification, deployment, and transferability of quantitative trait loci from genome-wide association studies in plants

Mohsen Mohammadi[a,*], Alencar Xavier[a,b], Travis Beckett[a], Savannah Beyer[a], Liyang Chen[a], Habte Chikssa[c], Valerie Cross[a], Fabiana Freitas Moreira[a], Elizabeth French[c], Rupesh Gaire[a], Stefanie Griebel[a], Miguel Angel Lopez[a], Samuel Prather[a], Blake Russell[a], Weidong Wang[a]

[a] Department of Agronomy, Purdue University, 915 West State Street, West Lafayette, IN, 47907, USA
[b] Whole-genome Analytics, Corteva Agrisciences, 7000 NW 62nd Avenue, Johnston, IA, 50131, USA
[c] Department of Botany and Plant Pathology, Purdue University, 915 West State Street, West Lafayette, IN, 47907, USA

ARTICLE INFO

Keywords:
Linkage disequilibrium
Genetic background effect
Expectation in QTL deployment
Candidate gene identification
QTL introgression

ABSTRACT

Over the past decade, the use of genome-wide association studies (GWAS) in plant breeding for discovery and validation of quantitative trait loci (QTL) has vastly increased. Successful deployment and transferability of these findings, however, have been limited. To increase the value of GWAS for plant breeding, experimental and methodological aspects must be addressed and refined. Population designs and statistical techniques are necessary to properly account for the effect of long-range linkage disequilibrium. Success of current methods has been restricted to the detection of common-variants with moderate additive effects; discovery of rare variants or QTL that depart from additivity has been elusive. Pleiotropy casts doubt on the cause-effect relationships between markers and multiple traits. Major criticisms of association studies center on reproducibility of results and the lack of transferability to other environments and populations. We also discuss population structure, phenotypic plasticity, epistasis, and genotype-by-environment interactions as they apply to GWAS. Perspectives are given for environment-dependent QTL, experimental settings, use of next-generation populations, and deployment of GWAS results for breeding applications and strategic exploitation of genotype-by-environment interactions. In summary, we present an overview of contemporary issues in identification and deployment of marker-trait associations and suggest future avenues of research towards new methods and new sources of data.

1. Introduction

The ultimate goal in genome-wide association studies (GWAS) is to associate Mendelian co-segregation of genomic segments to the phenotypic characteristics of the plant. In 1923, Karl Sax published a groundbreaking study that reported on a correlation between seed color and seed weight in the common bean (*Phaseolus vulgaris*). Co-inheritance of these two phenotypes led Sax to conclude that these two traits must be linked. In the six decades that followed, numerous studies confirmed genetic linkage between monogenic traits [1], such as fruit weight and shape in tomato [2].

In the early years of genetic markers, restriction fragment length polymorphisms (RFLP) allowed the construction of the first large scale high-precision genetic linkage maps [3] and inference of association between a measurable phenotype and allelic state of marker loci [4]. Paterson et al. [3], used a saturated RFLP map to identify six chromosomal regions that harbor genes controlling fruit mass, concentration of soluble solids, and fruit pH in tomato. The chromosomal regions that harbor the genes that control quantitative traits were called quantitative trait loci (QTL).

In 1996, Risch and Merikangas identified a crucial deficiency in linkage mapping by pointing out that method had insufficient power to detect genes of modest effects in human populations. They argued that while basic statistical approaches could successfully identify QTL for a trait which is controlled by a small number of loci each with large effects, when a trait is controlled by many loci each with small effects, the same testing approach would not have sufficient power to detect true QTL. This phenomenon was thoroughly examined by Beavis [5] and is now appropriately called the 'Beavis effect'.

Genome-wide screenings were first applied to humans; a few years passed before the methods would be leveraged for use in plants. In 2006, the first successful genome-wide screening for QTL in the human

* Corresponding author at: Wheat Breeding and Quantitative Genetics, Department of Agronomy, College of Agriculture, Purdue University, 915 West State Street, West Lafayette, IN, 47907, USA.
E-mail address: mohamm20@purdue.edu (M. Mohammadi).

https://doi.org/10.1016/j.cpb.2020.100145
Received 11 October 2019; Received in revised form 4 March 2020; Accepted 22 March 2020

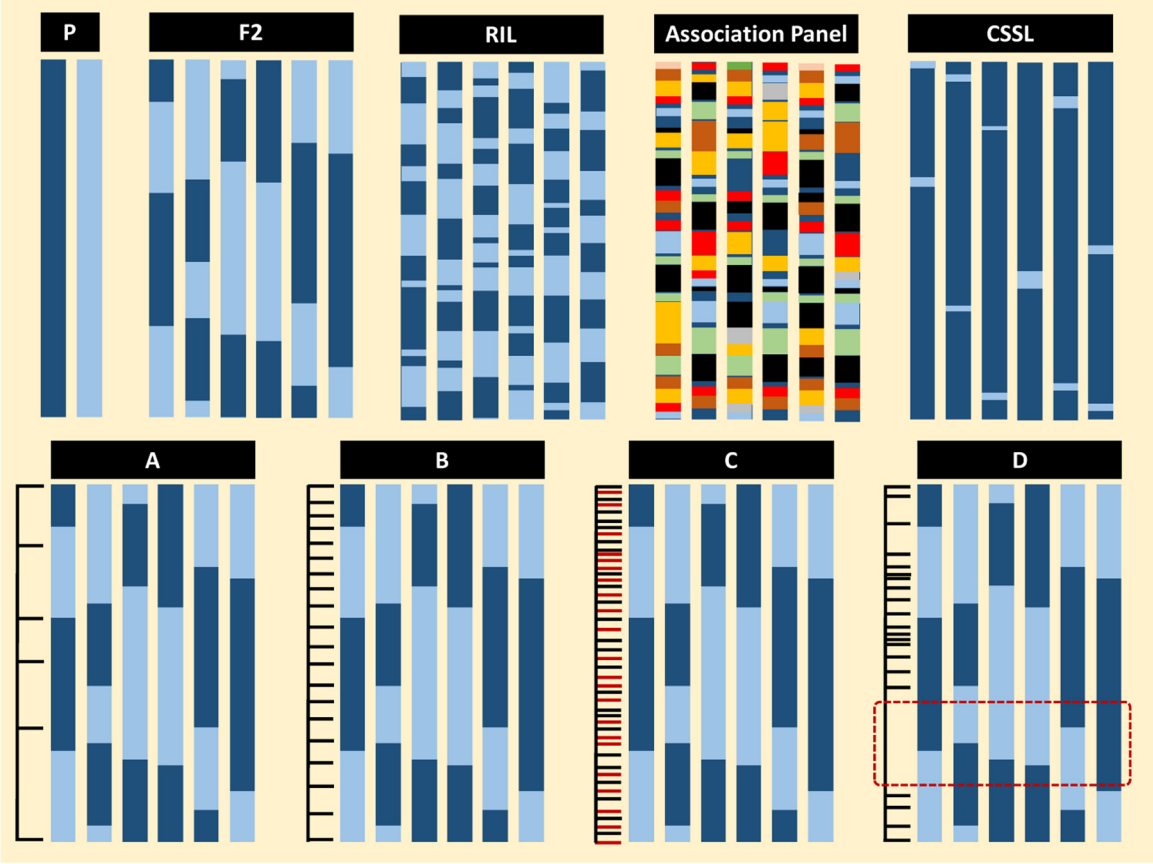

Fig. 1. Nine populations are represented, with each vertical bar representing the same genomic region for each individual within the respective populations. A mapping population is created by crossing two inbreds (**P**). After one generation of meiosis and recombination, the resulting population ($\mathbf{F_2}$) has a genomic resolution of approximately 20 cM. Additional generations of selfing (≥ 7 or 8) results in an immortal bi-parental recombinant inbred line (**RIL**) population with 5x greater genomic resolution than the F_2 population. A set of diverse and unrelated individuals (**Association Panel**) which originated from common ancestors many generations ago incorporates recombination events that are much older. Several generations of backcrossing a RIL individual to one of the original parents (P)—with the aid of marker-assisted selection targeting specific QTL—will create chromosome segment substitution lines (**CSSL**). These **CSSL** populations can be powerful tools for fine mapping, QTL validation, and gene discovery. In populations **A-D**, the side ladder and the horizontal footsteps to the left of each population represent the physical location of molecular markers. More footsteps indicate a denser molecular marker system. Population **A** can still benefit from additional molecular markers, as the six existing markers do not capture all of the genomic differences. Population **B** is covered by a reasonably dense marker set. Population **C** has many more markers than is needed to capture the genomic differences. In such cases, SNP tagging (e.g., removing the red markers and retaining the black markers) will benefit statistical threshold setting. Population **D** has a marker set that covers most of the genomic region but fails to cover the genomic differences in the sub-region identified by red rectangle.

genome identified a link between the Factor H and age-related macular degeneration [6–8]. One year later, Burton et al. [9], using 17,000 samples, 5000 cases, and 500,000 SNPs, reported on associations between seven complex diseases and molecular markers. Genome-wide screenings in humans would go on to identify molecular markers associated with Crohn's Disease, type-1 diabetes, multiple sclerosis, rheumatoid arthritis, colorectal cancer, and lung cancer [10].

In 2004, a publication first appeared that applied GWAS methods to plants. The target was dissection of genetic architecture for yield metrics in barley [11]. Three years later, a study of association mapping in Arabidopsis was published [12], followed by two 2010 studies, one a study in barley [13] and another in Arabidopsis [13]. Then in 2012, three notable GWAS studies were published on rice [14], tomato [15], and barley [16]. To the present day, most GWAS studies in plants have been performed on major crops of economic importance [17]. Through 2017, the GWAS catalog identified at least 3092 publications and 49,769 unique SNP-trait associations generated from analyses with at least 100,000 SNPs [18].

The large number of published GWAS is a testament to its widespread adoption and optimistic valuation by the scientific community. Research programs have invested heavily in design and execution of GWAS projects, with the goal of building powerful fast-track pipelines to identify and deploy QTL or candidate genes for important traits [19,20]. However, several fundamental questions remain, as noted by Yang et al. [21]: What defines a successful GWAS? How can we increase the true discovery rate? How can we validate QTL across populations or environments?

Perhaps a correct identification of true positive alleles is all that ought to be required for a GWAS to be considered successful. Over the past fifteen years, GWAS studies have identified thousands of statistically significant marker-trait associations (MTA) [18]. However, most MTAs show such a small effect that result in little to no improvement in plant and other organisms [22,23]. A successful GWAS also ought to incorporate elements of candidate gene discovery and QTL deployment—two topics which are at the frontlines of scientific pursuit [24]. Therefore, we propose the following criteria for a GWAS to be considered successful: the study needs to identify true MTAs with (1) meaningful effect sizes; (2) a close proximity to underlying genes for the traits of interest; and (3) transferability within a similar population and across a reasonably broad set of environments. This minimal definition still bears considerable consequence and requires the study to include certain validation steps. In the following sections, we suggest elements of experimental design that ought to be given due consideration prior to allocation of resources and execution of GWAS

experiments.

2. Design considerations

Proper experimental design prior to a GWAS will improve the ability of an analysis to identify true MTAs and candidate genes. The basic statistical requirements for the detection of association between markers and phenotypes center on the population, and both the genotypic and phenotypic data [25,26]: (1) The population must display genetic variation for the trait of interest and be sufficiently large for the detection of QTL of modest effect; (2) Each molecular marker included in the genotypic dataset must have sufficient polymorphism to ensure detection as well as sufficient coverage and recombination around the regions of interest to locate the marker closest to the causative polymorphism; (3) The phenotypic data must be accurate and reliable enough to be able to discriminate between genetic signal and noise. These topics are explored in further details in the sections below. First, different types of populations are discussed with respect to their genomic resolutions. Second, kinship among individuals and population structure, marker density, coverage, and the impact of allele frequency are considered. Third, the effects and control of non-genetic factors and trait correlations will be addressed.

3. Genomic resolution and populations

Experimental populations offer improved genomic resolution in the context of GWAS, as recombination events are captured by markers, assuming a portion of the phenotypic variation is due to genetics and underlying loci and candidate genes are captured by the genotyping platform. For instance, an F_2 population (Fig. 1) of reasonable size provides resolution of approximately 20 centi-Morgans (cM) per 95 percent of QTL confidence interval, whereas generations of random mating without selection followed by generations of inbreeding can substantially increase genetic resolution as well as statistical power.

In early QTL studies, the problem was the scarcity and low coverage of molecular markers. With high-density genotyping platforms, that is not the case anymore. In current GWAS platforms, increasing the number of markers will not increase the genetic resolution—instead, the limiting factors are population characteristics (Fig. 1). While linkage mapping populations (F_2 and backcross) incorporate recent recombination, random-mated populations incorporate historic recombination events (Fig. 1). An example of recent versus old recombination and their consequences is the comparison between European and African human populations [27], where the existence of historical recombination events offer the opportunity of detecting the loci associated to traits in narrow genomic region. In plants, the opportunities to leverage historical recombination events often relies on germplasm collections [28].

4. Kinship

In the ideal mapping population, all individuals are evenly related according to prior expectations, without subpopulations. In contrast, real-life populations display various level of co-ancestry, violating the assumption that any pair of individuals sampled at random will provide the same level of relatedness [29]. Episodes of natural selection imposed by nature, adaptation, and migration as well as artificial selections imposed by breeders, historical ancestry, and islands of local recombination all result in population structure or stratification [30–32]. Stratifications cause a systematic difference in allele frequencies between sub-populations (Fig. 2), creating a confounding effect that leads to identification of spurious associations in GWAS. Individual relatedness due to ancestry or kinship is evidence of genetically distinct subgroups within the population [33]. Such conditions of recent common ancestry produce false linkage disequilibrium (LD) (see Fig. 2) [34,35].

The effect of subpopulations is a major statistical issue, causing an

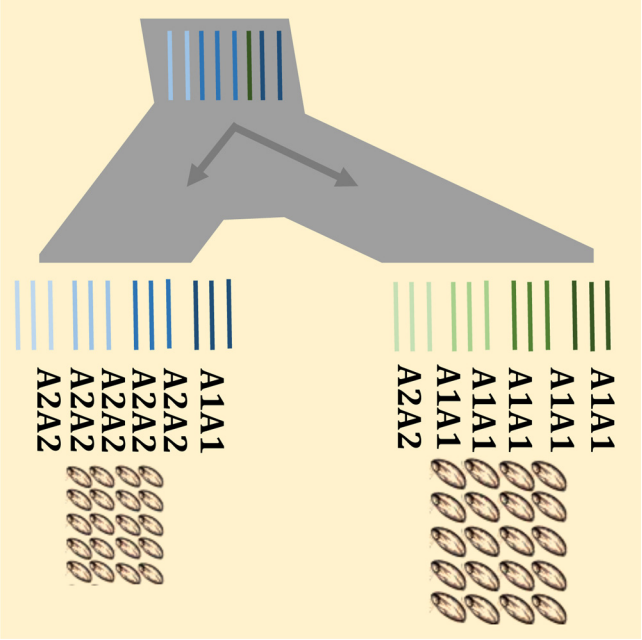

Fig. 2. In ideal unstructured GWAS populations, individuals are evenly related. In reality, some degree of relationship exists due to shared ancestors, selection, or migration. Genealogical descent of populations or the local recombination landscape may result in two independent phenomena: (1) stratification may cause differential expression for the trait of interest e.g., the left and right wheat sub-populations may be different in their kernel weight; or (2) stratification causes systematic changes in allele frequencies, e.g. allele A2 is more prevalent in the sub-population on the left while allele A1 is more prevalent in the sub-population on the right. By ignoring and not accounting for the stratification of sub-populations, the A locus can show spurious (false positive) association with kernel weight.

increased rate of both false positives and false negatives. When MTA signal detection is confounded with these sources of noise, genomic control measures in the form of co-ancestry estimates should be used to mitigate the effect of genetic background [36]. Pedigrees were first used to determine measures of co-ancestry, whereas the genomic similarity among individuals more accurately estimate the relationship among individuals in the post-genomic era [37].

5. Structure

When subsets of an original population are created by drift, migration or selection, allele frequencies do not necessarily reflect those from the source population. When multiple evolutionary events occur, a population tends to stratify and form clusters composed of more closely related individuals. A stratified population will contain two or more subpopulations.

Cluster analyses are commonly employed in statistical genetics to determine the existence of subgroups and to define which individuals belong to which group as the probability of belonging to one group or the other. There is a wide range of machine learning methods broadly used for germplasm classification [26]. Clustering can be performed via unsupervised methods, such as hierarchical clustering [38,39], or supervised methods, such as random forest [40]. It is also possible to frame the subpopulations to identify the general number of sub-groups without explicitly assigning individuals into groups. For instance, subpopulations can be described through applying principal component analysis (PCA) to genotypic data [41], where principal components refer to the most relevant Eigenvectors of the genomic relationship matrix.

When principal components are the method of choice, GWAS include Eigenvectors as fixed-effect covariates, such that the effect

associated to subpopulations is accounted for when testing the statistical association between marker and phenotype [42,43]. Including nuisance parameters that capture some level of population structure during the statistical testing is necessary to reduce the number of spurious MTAs that would have otherwise been identified as real associations [44,45].

When subpopulations are known beforehand and individuals have been assigned to clusters, it is possible to quantify the degree of systematic deviations among subpopulations through the calculation of F_{ST} statistics [46]. The F_{ST} statistic measures the discrepancy in variation of alleles of a given locus between subpopulations (S) and the total population (T). F_{ST} values close to zero indicate that the subpopulations have relatively similar allele frequency for the locus under evaluation, whereas F_{ST} values close to one indicate a large discrepancy in allele frequency. A locus with $F_{ST} = 1$ indicates that subpopulations are entirely homozygous for different alleles. Genome-wide F_{ST} analysis can provide an important insight on genomic regions associated to subpopulation [31]. This information is particularly relevant when a QTL identified through GWAS is situated in a F_{ST} hotspot, in which case the MTA is prone to be a false positive as the association is being confounded by the population structure.

6. Marker coverage and genotyping density

Adequate coverage means a sufficient number of SNP markers are located across the genome to enable detection of the causal variant that controls the trait of interest (Fig. 1). High-density coverage is achieved by genotyping every segment of the genome. LD between the SNP markers can be used as a measure of the effectiveness of genomic coverage [47]. Genome size and mating system both have direct consequences on marker density. The same number of markers will have different density for organisms with different genome sizes (Fig. 1). Genome sizes can vary largely across plant species: wheat has a genome size of 17 Gbp [48]; maize, 2300 Mb [49]; rice, 390 Mb (International Rice Genome Sequence Project); and Arabidopsis, 125 Mb (Arabidopsis Genome Initiative, 2000).

The genotyping density and genomic distribution of marker cannot achieve a mapping resolution greater than that imposed by the genetic limitations of the association population (Fig. 1). Markers in close physical proximity are often in strong LD and offer redundant information [50], thus creating statistical limitations that hinder researchers' efforts to pinpoint exact locations of true QTL.

7. Haplotypes and SNP tagging

Markers in close genomic proximity often display high LD due to limited recombination. Such groups of markers form blocks across a chromosomal region known as haplotype blocks (Fig. 3). When haplotype blocks are represented by a relatively homogeneous number of markers, GWAS have less risk of inflating effect size variance estimates [50] and losing power to detect candidate genes or genomic region associated with the trait [51].

A set of SNP markers that represent a haplotype block is commonly referred as tagging SNPs or haplotype tagging SNPs (htSNPs) [52]. There are several methods to identify htSNPs and can be broadly categorized into block-based and block-free methods [53]. Block-based methods use prior haplotype information and identifies a subset of htSNPs based on how well it can capture the diversity in the full set of SNPs. Block-free methods do not require prior information about haplotype blocks and can be further categorized into LD-based methods and haplotype reconstruction-based methods [54]. LD-base methods rely on the LD (correlation) estimates to partition the chromosomes into haplotype blocks (no prior information required) and identify a set of linked SNPs that represents the haplotypes [55]. SNP tagging efficiency appears to be higher in genomic regions with strong LD than in regions with moderate to weak LD [53,56]. Ke et al. [56] suggested optimizing

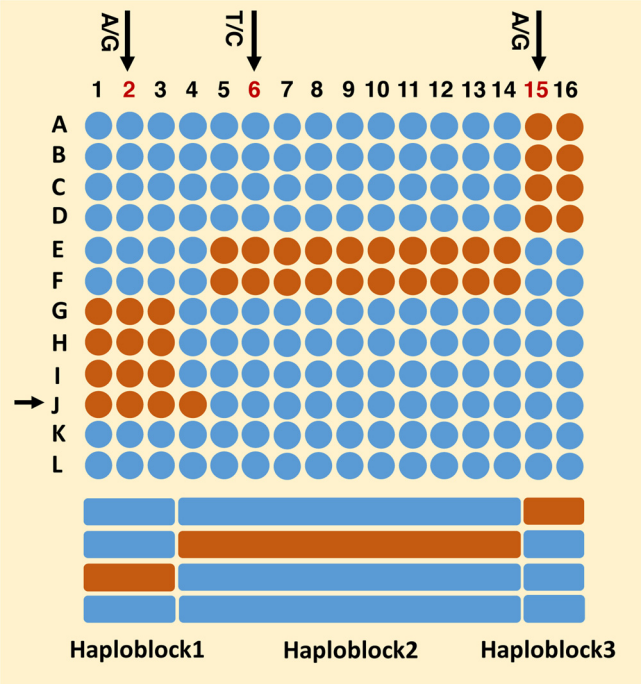

Fig. 3. The procedure for SNP tagging and the risk for missing rare variants are presented. The diversity of 11 individuals A:L is assessed by 16 SNP markers i.e., 1:16. When fewer number of SNPs are selected (2, 6, and 15) the majority of diversity can still be represented without losing much of the information. The three SNPs 2, 6, and 15 together can each mark a haplotype, simplifying the entire diversity to three haplotypes and four genotypic classes in 10 out of 11 individuals. However, one individual (J) cannot be correctly categorized by using only SNPs 2, 6, and 15 because it has a rare variant at SNP4.

tagging through a slight increase of minimum r^2 values for high LD regions and a slight decrease of maximum r^2 values for weak LD regions. Similarly, in haplotype reconstruction methods, a series of post-analyses are performed to reconstruct the haplotype from initial inference such that the informative SNPs (htSNPs) produce a desired level of accuracy in predicting the non-informative SNPs (non-tagged SNPs) (He and Zelikovsky 2006). One major limitation of reconstruction-based method is that it entails considerable computational complexity [54]. Although SNP tagging increases the accuracy of effect size and power to detect QTL linked to common alleles (see Fig. 3), it may also misrepresent rare variants [57].

Alternatively, haplotypes have been used as polymorphisms for association studies and have some advantages over using SNP. Since the SNPs are grouped into haplotypes, haplotype-based association mapping reduces the number of multiple comparisons, reducing false discovery. Furthermore, haplotypes are multi-allelic in nature and thus increase the heterozygosity providing more power. In plants, haplotype-based GWAS has been performed in barley ([58], maize [59], durum wheat [60], soybean [61] and wheat [62].

8. Rare variants

Falconer et al. [63] state that a rare variant is an allele with a frequency less than 1%. Cirulli and Goldstein [64] define four categories of variants in terms of frequency: (1) very common alleles, with frequency between 5% and 50 %; (2) less common variants, with frequency between 1% and 5%; (3) rare variants, with frequency less than 1%; and (4) private variants that are restricted to immediate relatives.

The detection of QTL is a function of many genetic and statistical factors, including the effective population size, genotyping density and the distribution of allele frequencies. It is easier to detect a QTL represented by markers with higher minor allele frequencies than

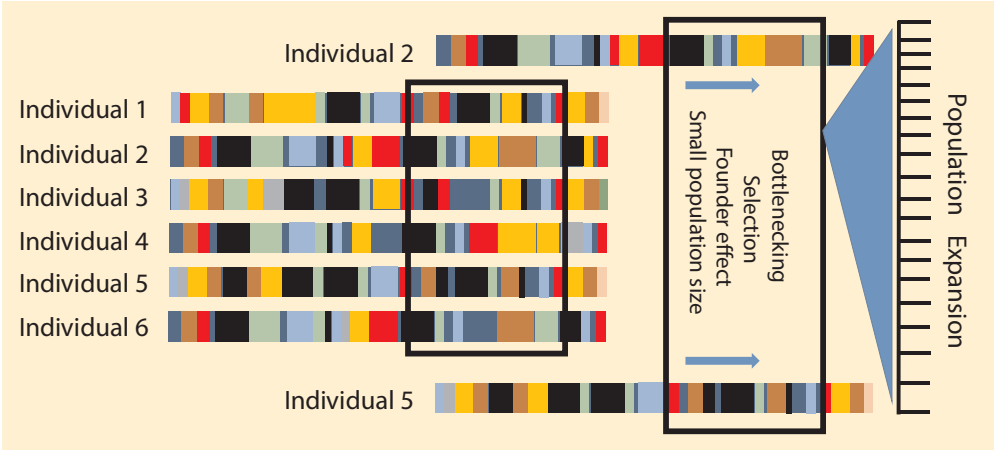

Fig. 4. The effect of any episode of bottlenecking on shrinking diversity and lengthening LD in immediately expanding population is shown. Two individuals (2 and 5) are selected out of a population of size n = 6 to become parents of a new population. With minimal recombination events, individuals in the immediate new generation are likely to have the same haplotype blocks as the founding parents i.e., 2 and 5. To generate diversity in the genomic region within the box, several generations of recombination are needed to re-shuffle the long range stretche into several short-range segments. If a GWAS is performed on a population that recently underwent selection or migration, any QTL identified on the recently selected region can be associated with a large number of linked candidate genes. In contrast, when a GWAS is performed on a population with genomic regions that have been subject to many generations of recombination, the QTL identified will be associated with shorter genomic regions that contain a manageable number of linked candidate genes.

markers with lower minor alleles that rarely appear in the population [65]. That is due to the higher variance of markers and, consequently, higher statistical power. Filtering the SNP panel by removing loci with low minor allele frequency ensures evaluation of variants shared by a relatively sizeable proportion of the populations. The maize causative mutation F469 deletion allele that controls oil content in maize kernels is a prime example of a common variant detected by multiple studies in different populations [66–69].

GWAS performed on complex traits generally identify loci that partially explain only a fraction of the estimated heritability, as a significant proportion of the inherited complex trait may be caused by the cumulative effects of low frequency genes [70]. Much of the speculation about the 'missing heritability' from GWAS focuses on the possible contributions of variants with low minor allele frequency, i.e. < 0.05 [71], as the effect of such variants is usually not sufficient to be captured by current GWAS [72,73].

9. Trait correlations

Phenotypic correlations among traits create ambiguous GWAS results. Pleiotropy occurs when one gene influences two or more traits, and can result in the association between traits. Another reason for association of traits is genetic linkage. These genomic regions cast doubt on the interpretation of MTAs, since most GWAS methods are not designed for multi-trait associations [74]. GWAS will capture epistatic QTL in all traits, unless the trait of interest is conditioned to the associated traits [75]. Zhang et al. [76] performed associations on a primary trait (kernel weight) while conditioning on a secondary trait (spike number). Also, it is important to note that multi-trait association methods that handle correlated traits can be more effective at detecting QTL, especially those with small effects [77,78].

10. Candidate genes and regulatory elements

The overall goal of QTL mapping and GWAS is the identification of genetic variants associated with the phenotype of interest. The resulting QTL often translates into a lengthy genomic region that likely harbors several polymorphic genes or regulatory elements. Identification of the precise genetic factor often requires further investigation of the QTL region—a follow-up study called 'fine mapping'.

Fine mapping relies on producing more recombination resolution in the coarsely mapped region. It is necessary for cloning, gene editing, and precise introgression of causal variants into the plant breeding pipeline. Researchers create near isogenic lines with identical genetic background, varying only for the polymorphism within the QTL region, for the functional assessment of the effect of the allele on the trait of interest. Uga et al. [79] present the fine mapping of deep rooting in rice (gene *DRO1*), where authors cloned and backcrossed *DRO1* into a shallow-rooting rice cultivar to enhance drought avoidance [79].

A limiting factor for fine mapping is creating experimental populations with high enough recombination resolution in the target region. The ability to create the necessary experimental populations depends on the mode of reproduction and mating system of the target species. Open-pollinated populations, such as maize, exhibit a high rate of LD decay that can be leveraged to narrow the QTL region, effectively decreasing the number of potential candidate genes. Self-pollinated crops, on the other hand, have much slower rates of LD decay; consequentially, MTAs are often located on large haplotype blocks [80]. The purpose of creating a new population for fine mapping is to promote crossovers in the coarsely mapped QTL target region, thus breaking up the linkage blocks and reducing the width of the QTL region.

11. Recombination

We have established that recombination is a key factor for successful fine mapping and GWAS by increasing both statistical power and genetic mapping resolution. To disrupt the LD within QTL regions, researchers can either increase the number of generations per year or increase the density of crossing over in each generation to create more opportunities for genomic regions to recombine (Fig. 4).

Crossover control is biologically achieved by three known mechanisms: crossover homeostasis, interference, and assurance. Crossover homeostasis maintains the number of crossovers at a relatively constant number. This phenomenon has been previously shown in populations of mice, yeast and *C. elegans* [81–83]. Interference, on the other hand, is the mechanism by which crossovers do not form in close proximity to one another. The result is non-random occurrence of crossovers along the length of each chromosome. Assurance refers to the observation of at least one crossover between a pair of homologs. In maize, no evidence was found for crossover control except assurance that at least one crossover remains in each homolog pair [84].

The number of crossovers per generation can be manipulated through external agents or natural genetic processes. Hanneman et al. [85] showed that cisplastin, an anticancer drug, increased the frequency of recombination in mice and Arabidopsis. Jackson et al., (2015) reported that exposure to acute thermal stress was associated with elevated recombination rates in Arabidopsis, *Drosophila*, and barley. However, the increased recombination rate also led to decreased fitness of the population (Jackson et al., 2015). In plants, pathogen-

induced stress can also lead to an increase in somatic cell recombination rates. Arabidopsis inoculated with *Peronospora parasitica* and tobacco inoculated with mosaic virus both showed increased recombination rates in somatic cells [86,87]. Similarly, X-rays and ultraviolet C (UVC 100–280 nm) can increase recombination rate [88].

Recombination rates are also driven by genetic factors with quantitative inheritance. Natural variation in recombination rates can be found at the individual, population and species levels. Response to selection has been shown to increase recombination rates in *Drosophila* [89,90]. GWAS within human populations, cattle, and wild sheep have consistently identified two genes, *RNF212* and *CPLX1*, associated with increased recombination rates in females [91–93]. In Arabidopsis, two genes, *AtMSH4* and *AtMER3*, were shown to control formation of interfering crossovers. *The* number of crossovers was reduced by 75 % for *AtMSH mutants* and 85 % for mutants at *AtMER3* [94,95].

12. Candidate gene strategies

A QTL containing a small number of genes is considered fine-mapped, and the genes within this short region are named candidate genes. Researchers assume that there exists a statistical correlation between the DNA polymorphism in or near the candidate gene and the phenotypic trait of interest [96]. Several strategies exist that help identify a manageable list of candidate genes and prioritize genes for functional validations. One is the position-dependent strategy Fig. 5, where gene search is based on the decay of LD in the region. A physical region within which the population reaches a pre-defined level of LD becomes the search window where candidate genes are identified [97]. Although LD and physical distance can provide strong evidential criteria, there are circumstances where the true QTL region includes several genes.

Candidate gene identification methods may combine position-dependent strategy with annotation and ontology, comparative genomics, and gene expression strategies. Annotation and ontology strategies refer to using previously known gene annotations and ontologies to pinpoint genes that are most likely to be associated with a trait. For example, in a QTL analysis of plant height, candidate genes tagged with the gene ontology (GO) biological process category “gibberellic acid mediated signaling pathway” (GO:0009740) are likely to be a good candidate for further testing. Comparative genomics strategies use cross-species fast-track approach to identify and characterize the effect of candidate genes. Gene expression-based strategies employ differential gene expression information between lines containing different alleles in the defined QTL region to define candidate genes [97]. It may be the case, however, that differential expression might occur in other tissues than the organ or tissues under study. For example, in root studies, it may be the differential expression in leaf tissue that derives the signals that promote differential growth in roots.

13. Functional validation

The confirmation of a gene functionality is performed upon genes that have been fine mapped and cloned. Strategies to evaluate gene functionality include genetic complementation and rescuing known mutants, overexpression, and silencing. Genetic resources for functional genomic studies in diploid model species (Arabidopsis and rice) consist of comprehensive reverse-genetics resources, such as genome-wide mutant populations, genome-wide T-DNA insertional mutants in Arabidopsis [98], T-DNA insertional mutants in rice [99], and genome-wide RNAi-silenced lines in Arabidopsis [100]. In polyploid wheat, most genes are present in multiple functional copies known as homologous genes. Multiple copies have a buffering effect, where loss-of-function mutations are masked by the presence of other homologous genes [101].

The evolution of engineered nuclease techniques has led to the development of clustered regularly interspaces short palindromic repeats (CRISPR). Currently, gene and genome editing is mainly accomplished using the CRISPR-Cas9 technique, which offers a wide range of potential uses for candidate gene and regulatory element functional analysis and validation. For example, gene editing has been employed to examine the genes underlying disease resistance in rice and Arabidopsis [102].

14. Validation in other genetic backgrounds

Early optimism about QTL deployment in populations via marker-assisted selection may have mislead researchers [103]. QTL identified in one population may not show the same magnitude of effectin other populations [104,105], as favorable alleles often exhibit population-specific effects. This could be due to epistatic interactions of the QTL within the overall genetic background [106], leading to penetrance and variable degrees of expression.

Under strong epistasis, the effect of a major single-locus QTL is influenced by interacting loci spread across the genome. The epistatic

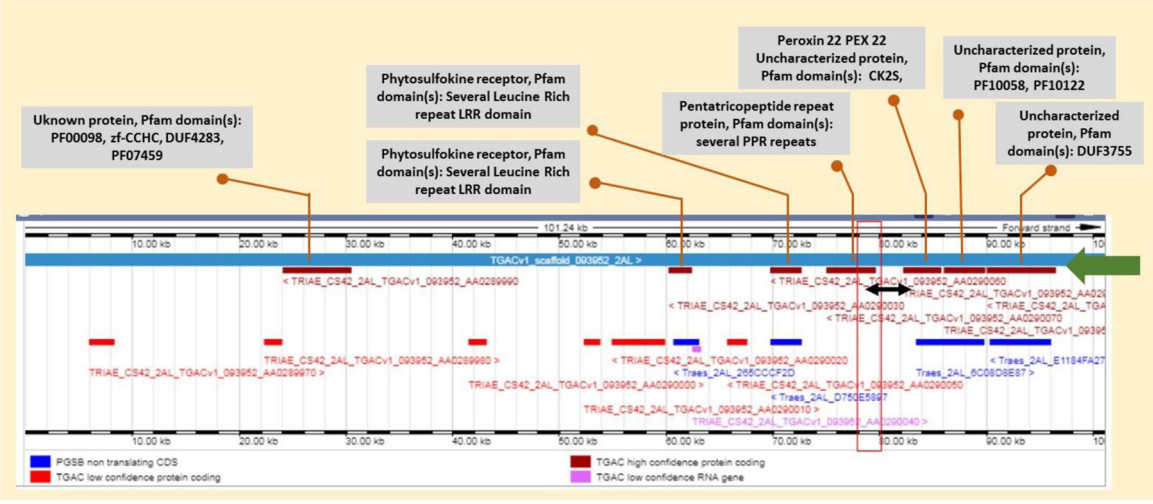

Fig. 5. The position-dependent strategy to identify candidate genes is presented. A genotyping-by-sequencing (GBS) marker in hexaploid wheat was found to be associated with a trait of interest. This marker is located on chromosome 2A. A 2000 bp anchor sequence (the red vertical box) surrounding this SNP was used to blast against Ensembl TGACv1 wheat genome. The blast returned the genomic scaffold 093952, which contains 7 high confidence protein coding genes (dark red horizontal rectangles aligned with the green arrow), which vary in their physical distance from the polymorphic marker (center of the red vertical box).

background effect limits the effectiveness of deploying the QTL in other populations [107–109]. The QTL and the interacting loci act as a package within the specific genetic background of the discovery population [109]; special statistical techniques may be necessary to identify and mitigate effects of background epistasis in these situations. Xavier et al. [110] proposed testing the marker effects within multiple populations simultaneously, thus avoiding inconsistencies in QTL phasing, genetic background, and effect sizes from one population to another [19].

Xu [111,112] proposed a mixed-models-based GWAS using multiple kinship matrices, explicitly accounting for various levels of polygenic background and suitable when interactions are of major concern. Other possiblilities are the GWAS methods that search for additive and epistatic associations simultaneaously [113]. Such methods were deployed by Mathew et al. [114] to seach for multi-loci associations that captured epistatic and additive effects on flowering time in barley.

15. Transferability to other environments

GWAS results can be robust and reproducible across environments environment as long as the level or phenotypic plasticity (PP) is low and genotype-by-environment interactions (GEI) are minimal or non-existent and play no role in the expression of the trait of interest (Fig. 6). Such assumptions are not necessarily valid because the phenotype of a quantitative trait is the result of genotype, environment, and GEI [63,115]. For example, in the study of yield in a population of soft red winter wheat, we observed considerable and significant genotype by year interaction even in single location. Differences between PP and GEI are subtle and frequently confounded [116], as PP refers to the ability of one genotype to react to a range of environments [117] and GEI denotes variations in several genotypes in response to a range of environments [63].

Approaches to study environmental responses include reaction normal plots, GEI analysis, and between-environment correlation [118]. Reaction norms relate to assess the phenotype of a given genotype over a range of environments [119], GEI analysis show the variation in reaction norms of multiple genotypes, and correlation between environments is a measure of consistency of genotypes under different environments [120,63]. Guntrip and Sibly [118] presented graphic illustrations of PP, GEI, and GCE in the context of evolution of specialization (see Fig. 6). In traditional approaches the whole genome is considered in GEI analysis [121,122].

When genotyping information is available, GEI can be further analyzed to the level of QTL-by-environment interaction (QEI) resolution [123–126]. On top of QEI interaction, Sul et al. [127] reported that population structure and polygenic background affect GWAS and increase the number of false positives unless GEI is accounted for. For genome-wide screening of grain yield genes in soybeans, Xavier et al. [128] addressed GEI through within-environment GWAS followed by meta-analysis to combine results into a single association. Deriving full benefit from QEI-GWAS results requires complex interpretation and must be evaluated with caution, as the nature of QEI association may vary significantly across environments.

16. Controlled environments

The transferability of a QTL across environments includes ecological boundaries. GWAS from phenotype collected under controlled environment (Fig. 7) often have QTL validated on distinct environments. This phenomenon has been reported in canola [129], wheat [130], maize [131], and rice [132]. The progression from the single-plant to the field plots requires factoring in plant-plant interactions and mitigating environmental effects. For instance, interactions among plants cannot be ignored, as they compete for different resources [133]. Mutic and Wolf [134] studied an Arabidopsis population grown with Landsberg neighbors, where QTL were evaluated for direct and indirect effects on size, developmental, and fitness related traits. The authors concluded that neighboring plants introduce interactions into the ecosystem. Ignoring indirect genetics effects may over- or underestimate the QTL effect [135].

17. QTL deployment

Crop performance is highly dependent on the growing environment. With the mapping objective of exploiting GEI, association analyses are performed within target environments. The inconsistency in detection of signals across environments is widely attributed to the differential allele effect associated to QEI [136].

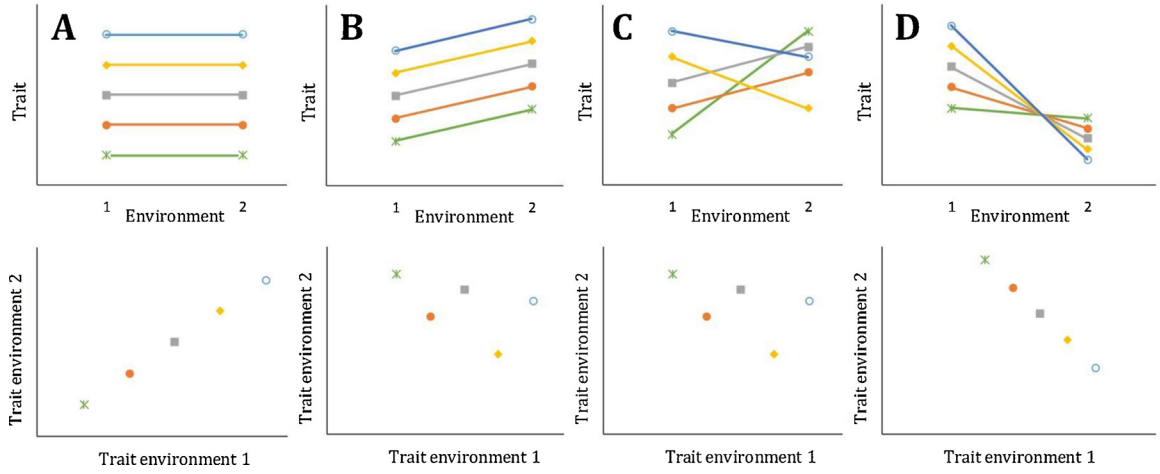

Fig. 6. Reaction norms (top panel) and between-environment correlations (bottom panel) for five genotypes evaluated in two environments. Parallel horizontal lines (top A) demonstrate a lack of phenotypic plasticity as well as absence of GEI. In this case, the between-environment genetic correlation (GCE) equals +1 (bottom A). Changes in the magnitude of the expression of traits (increases from left to right) over a range of environments are suggestive of strong phenotypic plasticity (top B) and indicative of no GEI because such plasticity occurs in all genotypes at an equal rate. The GCE in this case is again +1 because the performance of genotypes in one environment is predictable from the performance of genotypes in the other environment (bottom B). Changes in the magnitude of the expression of traits in a manner that their ranking from one environment to another environment also changes, is a clear representation of both PP and GEI (top C). The GCE for this case tends to be negligible due to lack of credibility in across environment predictions (bottom C). Changes in the magnitude of the expression of traits such that not only their ranking but also their directions are affected from one environment to another (top D) is a typical case for both PP and GEI, where GCE tends to be close to -1 (bottom D). The concepts depicted in this illustration are adapted from Guntrip and Sibly [118].

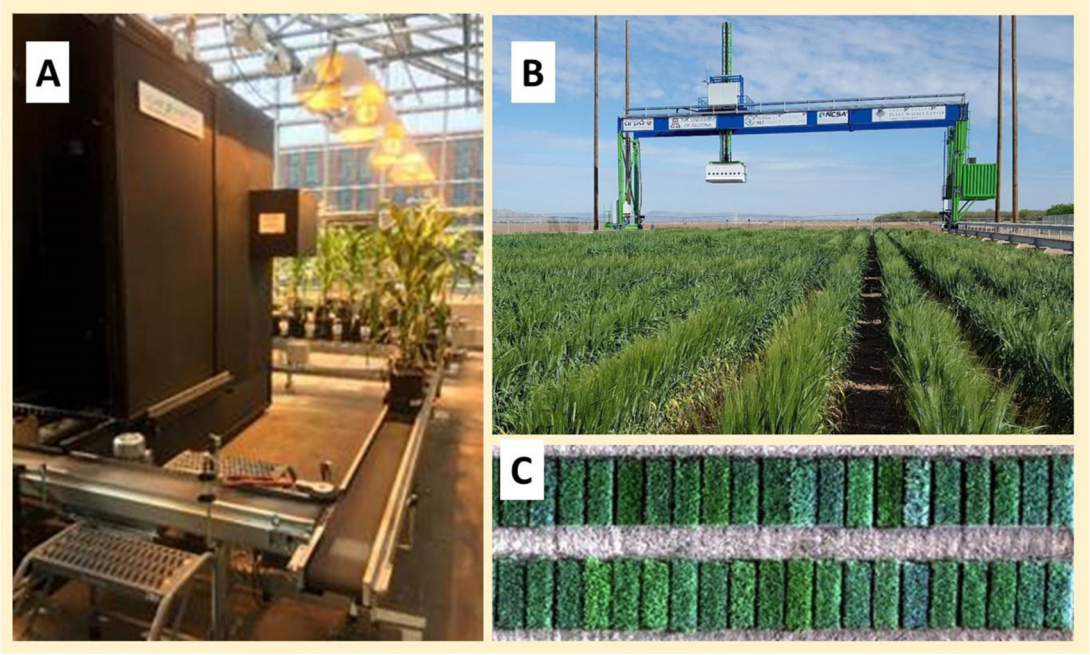

Fig. 7. A. Maize plants in automated hyperspectral imaging room. **B.** The durum wheat diversity panel planted under Scanalyzer. **C.** A part of the preliminary wheat yield trial, 1m × 3m plots imaged at 40 m altitude and speed of 7 m/s using hyperspectral sensors mounted on Unmanned Aircraft Vehicles. **A** is an example of a single plant phenotypic analysis for discovery of mechanisms (e.g., development, stress tolerance, nutrient uptake efficiency). **B** and **C** are examples of analysis of plants in their community, which entails plant-plant interactions, and results in identification of traits that are relevant to production scale.

A considerable fraction of GWAS results are performed under QEI influence, supporting the need for phenotyping in multiple locations to accurately characterize the QTL effect. The levels of QEI was estimated to be 27–52 % for the inflorescence development traits in Arabidopsis [137]; 48 % for water-soluble carbohydrate accumulation in drought studies in wheat [138]; and 87.4 % in grain yield and its components in rice [139]. The success of a variety in environment 'A' yet developed in environment 'B' depends on the level of similarities between the two environments. Similarly, a compilation of germplasm from multiple locations is unlikely to reveal truly advantageous germplasm, which is referred to as genotype-environment covariation.

Next generation populations such as Nested Association Mapping (NAM) panels and Multi-parent Advanced Generation Inter-Cross (MAGIC) populations are powerful resources to identify QTL with high power and resolution [140]. However, these populations have limited environmental scope and are restricted to detect environment-dependent QTL. For example, while soybean varieties in North America are developed and tested for performance in major maturity zones (Fig. 8), the results present a high level of correlation between varietal performance and the environments in which the varieties are bred [141]. Likewise, maize hybrids are developed and tested for performance within six general zones of comparative relative maturity. NAM populations were developed in maize, with 25 diverse varieties crossed to B73 with 200 individuals per offspring [142], and in soybeans, with 40 diverse varieties crossed to IA3023 with 140 individuals per offspring [128]. When these NAM panels are grown in a narrow range of environments, some of the varieties will not perform optimally. These individuals will not express their true genetic potential, and the true genetic effect will be confounded with GEI (Fig. 8).

18. Structural variants

Research efforts in the past decade concentrated on identifying genetic variants at the SNP level, the causal quantitative trait nucleotide (QTN). While SNPs are created by DNA replication errors, replicate-independent variants result from the ability of cells to repair damaged DNA [143]. The resulting structural variants include insertion-deletion (indel) polymorphisms, block substitutions, inversions and copy-number variants [144]. Structural variants tend to account for a greater proportion of the total nucleotide differences as compared to nucleotide variants.

Torkamaneh et al. [144] summarized nucleotide diversity and structural variants using sequencing of a representative set of 102 short-season soybeans genotypes. The authors demonstrated that only 0.071 % of all genetic variants were predicted to have a disruptive impact on the protein. These variants correspond to 4113 markers in 3064 genes, of which 2279 were single nucleotide polymorphisms, 230 were multiple nucleotide polymorphism variants, and 1604 markers were indel variants. Indels therefore represented 39 % of the 4113 functionally high impact variants. Despite the progress in sequence variant detection, structural variants come with challenging assessment of reproducibility [145]. It is expected that structural variants will be further explored with long-read sequencing technologies [146].

19. Capturing time-dependent QTL

The growth cycle for most field crops follows a continuous progression beginning at germination, proceeding through vegetative growth, flowering, grain fill, and terminating at maturity. Many traits are measured at one particular point in time (e.g. grain yield), whereas other traits can have multiple measures across the season (e.g. biomass). Traits collected with repeated measures throughout the season are referred to as longitudinal traits. In longitudinal traits, genotypes may display higher values of a time-dependent trait in one stage, but lower values in another stage. For those traits, studies can either be conducted by (strategy 1) focusing on the phenotypic measure of each stage, or (strategy 2) index or non-linear functions that compress the various time points into a single proxy phenotype.

Growth dynamics in soybeans [147,148] and maize [149] have been performed using GWAS adopting strategy 1, where each time point is treated as a different trait, thus supporting that different QTL were acting in distinct time points. Sikorska et al. [150] preferred treating the measurements from different time points as various correlated traits, conditioning each time to each other time point.

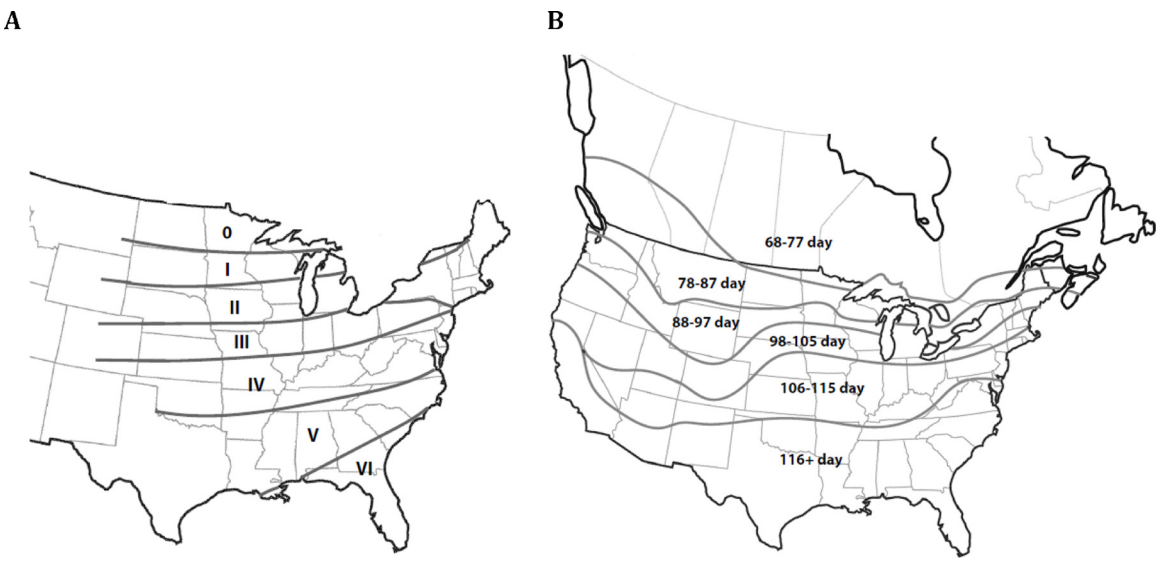

Fig. 8. Soybean (**A**) and maize (**B**) production and variety development are organized by maturity zones. For example, soybean maturity groups that are adapted to seven major zones, notated as 0-VI [141]. Varieties are developed to demonstrate the best performance in their respective zones, creating a high level of genotype-environment covariation. Nested Association Mapping panel populations are developed to increase power and precision of QTL analysis. In maize, 25 diverse varieties were each crossed to B73 [142]. When such a diverse panel is grown in a narrow range of environments, it is inevitable that some of the varieties will not be grown in the "real" environments for which they were optimized. These individuals will then not express their true genetic potential, and the true genetic effect will be confounded by the genotype by environment covariation. A basic question yet to be answered is: What is the ideal environment to test a NAM population for accurately estimating the genetic effects?.

Studies adopting strategy 2 deploy a non-linear term that brings observations from the various time points into the same scale. Functional modeling and random regression are key methodologies for longitudinal GWAS analysis [151], which simplify the detection and interpretation of QTL by capturing dynamic patterns [152]. Particularly in genomic prediction [153], random regression is a popular methodology in animal breeding [154] with increasing popularity in plants [155].

In addition to longitudinal traits, e.g., repeated measurements of biomass over time, the total genotypic value of a variety can be represented as the summation of genotypic competency related to different traits that express over time in the lifetime of plants. For example, the total genotypic value of a line can be regarded as the additive effects of nutrient and water acquisition by root, photosynthesis by leaf, and transport efficiency and grain fill. Proper strategies are needed to dissect plant competency across different traits and over different stages.

20. Statistical methodologies

Due to the growing popularity of GWAS analysis, various statistical methods have been developed over time to suit specific needs and to overcome specific limitations. The most common methodological framework relies on parametric linear models, either by testing one marker at a time, or all markers at once.

Single marker analysis (SMA) is the backbone of testing marker significance and effect. When structure is absent, statistical methods as simple as ANOVA and *t*-test may suffice. The use of generalized linear models (GLM) increased in popularity as it provides some control over population structure with fast computational efficiency, and it was not restricted to traits that are normally distributed [156]. However, in highly structured population, GLM methods show inflated significance levels leading to inadequate control of false positives and loss of power due to relatedness among the individuals in the population [33]. Mixed linear models (MLM) were proposed to overcome some limitations of GLM [33], providing more power, flexibility, and stringent control of population structure [157], as well as being able to detect some rare variants [158].

Yu et al., [33] presented a population structure (Q) + kinship (K) mixed linear model (MLM) to control the confounding effect of Q and K. The kinship matrix can be derived from pedigree data or through available marker data as described by VanRaden [37]. The original Q + K model suggested by Yu et al., needed computational power to estimate variance components in every iteration. Kang et al. [159], proposed a computationally superior method, efficient mixed-model association (EMMA), that simplified the relationship matrix compared to method described by Yu et al. [33] and was able to control false-positives. Since a large number of markers are tested using the same model structures, by chance alone, a proportion of markers that are significant will be false-positives. To account for the multiple-testing problem, SMA requires some statistical controls such as Bonferroni correction and False discovery rate (FDR). Bonferroni correction, that adjust the p-value by simply dividing it with the number of tests (markers), is highly conservative leading to false negatives and loss of power in the test. FDR adjusts the p-value by dividing it with the number of tests (markers) times the proportion of false positives allowed in the GWAS study. FDR is therefore, less conservative than Bonferroni method and is the most commonly used method in SMA GWAS studies. MLM are usually limited to Gaussian process, so Bayesian counterparts are needed for associations of categorical and ordinary traits [160].

The alternative framework to SMA (GLM and MLM) are whole-genome regressions (WGR) methods, popularly deployed for genomic predictions [161]. Xu [111,112] proposed the first WGR for detecting QTL, a linear model where each marker was treated as an individual random effect. The methodology quickly gained popularity on Bayesian framework [162] due to the flexibility of modeling complex linear models and the possibility of performing variable selection [163]. WGR methods condition the statistical evaluation of each SNP to every other SNP, eliminating the need for multiple-testing corrections [164]. WGR is a promising framework for GWAS analysis, often referred to as the next-generation of GWAS models [165]. Recently, multi-locus mixed models such as MLMM (Multi-Locus Mixed-Model [166];) and FarmCPU (Fixed and random model Circulating Probability Unification; 2016), that tests multiple markers simultaneously have been suggested to have more power and better false positive control compared to SMA. In multi-locus mixed model methodologies, marker testing is performed

in two steps. At first, an initial test is performed to identify significant markers known as pseudo-quantitative trait nucleotides (pseudo-QTN). Secondly, using the pseudo-QTN as covariates, the remaining SNPs are tested for their significance. Among the several MLM and multi-locus mixed model methodologies, FarmCPU possess higher power and better control of false-positives in GWAS studies [167,168].

Contrary to parametric methods (SMA and WGR), random forest regression (RF) is a nonparametric procedure based on resampling multiple decision trees [169]. RF has been utilized as flexible framework for GWAS [170]. RF is designed to be agnostic to the distribution of the trait, no assuming normality and being able to handle categorical traits [171]. Instead of estimating the effect of individual markers to the phenotype, RF is based on decision trees, being able to detect marker associations that rely on additive, dominant and epistatic effects, and capture haplotypes of high complexity [172]. However, the interpretation of QTL detected through RF can be challenging, as this methodology does not provide the explicit mechanism by which the MTA occurs [173].

21. Summary

GWAS methods can be a powerful tool to uncover causal genetic polymorphisms in plants, as long as the methods are applied correctly and within effective experimental settings. Identification of causal polymorphisms has and will continue to aid breeders in developing improved varieties to meet the food needs of an ever-increasing world population. While GWAS is a good first step towards the discovery and deployment of key genes, more and better research is necessary to evaluate the reproducibility and transferability of GWAS results, especially across environments and genetic backgrounds. Further development towards optimal experimental settings for GWAS analysis will undoubtedly require an interdisciplinary approach. We believe that a successful GWAS project is achieved in a collaborative environment with scientists from various fields. The identification of key traits to perform GWAS, proper analytical methods, generating the necessary genetic resource for mapping, and choosing adequate genotyping platform are the key components under the control of breeders and geneticist.

Declaration of Competing Interest

The authors have no conflict of interest.

Acknowledgement

Financial support from USDA Hatch grant 1013073 via Purdue College of Agriculture to MM is greatly appreciated.

References

[1] J. Thoday, Location of polygenes, Nature 191 (1961) 368–370.
[2] S. Tanksley, H. Medina-Filho, C. Rick, Use of naturally-occurring enzyme variation to detect and map genes controlling quantitative traits in an interspecific backcross of tomato, Heredity 49 (1982) 11–25.
[3] A. Paterson, E. Lander, J. Hewitt, S. Peterson, S. Lincoln, S. Tanksley, Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms, Nature 335 (1988) 721–726.
[4] S. McCough, R. Doerge, QTL mapping in rice, Trends Genet. 11 (1995) 482–487.
[5] W.D. Beavis, QTL analyses: power, precision, and accuracy. Molecular dissection of complex traits, 1998, 145-162, in: A.H. Paterson (Ed.), Molecular Dissection of Complex Traits, CRC Press, Boca Raton, 1998.
[6] A. DeWan, M. Liu, S. Hartman, S. Zhang, D. Liu, C. Zhao, et al., *Htra1* promoter polymorphism in wet age-related macular degeneration, Science 314 (2006) 989–992.
[7] C. Ku, E. Loy, Y. Pawitan, K. Chia, The pursuit of genome-wide association studies: where are we now? J. Hum. Genet. 55 (2010) 195–206.
[8] P. Visscher, M. Brown, M. McCarthy, J. Yang, Five years of GWAS discovery, Am. J. Hum. Genet. 90 (2012) 7–24.
[9] P. Burton, D. Clayton, L. Cardon, N. Craddock, P. Deloukas, A. Duncanson, et al., Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, Nature 447 (2007) 661–678.
[10] A. Korte, A. Farlow, The advantages and limitations of trait analysis with GWAS: a review, Plant Methods 9 (2013) 29.
[11] A. Kraakman, R. Niks, P. Van den Berg, P. Stam, F. Van Eeuwijk, Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars, Genetics 168 (2004) 435–446.
[12] K. Zhao, M. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, et al., An Arabidopsis example of association mapping in structured samples, PLoS Genet. 3 (1) (2007).
[13] J. Roy, K. Smith, G. Muehlbauer, S. Chao, T. Close, B. Steffenson, Association mapping of spot blotch resistance in wild barley, Mol. Breed. 26 (2010) 243–256.
[14] X. Huang, Y. Zhao, C. Li, A. Wang, Q. Zhao, W. Li, et al., Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm, Nat. Genet. 44 (2012) 32–39.
[15] N. Ranc, S. Muñ os, J. Xu, M. Le Paslier, A. Chauveau, R. Bounon, et al., Genome-wide association mapping in tomato (*Solanum lycopersicum*) is possible using genome admixture of *Solanum lycopersicum* var. cerasiforme, G3 Genes| Genomes| Genet. 2 (2012) 853–864.
[16] M. Wang, N. Jiang, T. Jia, L. Leach, J. Cockram, R. Waugh, et al., Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars, Theor. Appl. Genet. 124 (2012) 233–246.
[17] M. Alvarez, T. Mosquera, M. Blair, The use of association genetics approaches in plant breeding, Plant Breed. Rev. 38 (2014) 17–68.
[18] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, et al., The new NHGRI- EBI catalog of published genome-wide association studies (GWAS Catalog), Nucleic Acids Res. 45 (2017) D896–D901.
[19] B. Brachi, G. Morris, J. Borevitz, Genome-wide association studies in plants: the missing heritability is in the field, Genome Biol. 12 (2011) 232.
[20] X. Huang, B. Han, Natural variations and genome-wide association studies in crop plants, Annu. Rev. Plant Biol. 65 (2014) 531–551.
[21] J. Yang, N.A. Zaitlen, M.E. Goddard, P.M. Visscher, A.L. Price, Advantages and pitfalls in the application of mixed-model association methods, Nat. Genet. 46 (2014) 100–106.
[22] R. Bernardo, Breeding for Quantitative Traits in Plants, 2nd ed., Stemma Press, Woodbury, MN, 2010.
[23] J. McClellan, M. King, Genetic heterogeneity in human disease, Cell 141 (2010) 210–217.
[24] P. Marjoram, A. Zubair, S.V. Nuzhdin, Post-GWAS: where next? More samples, more SNPs or more biology? Heredity 112 (2014) 79–88.
[25] F. Frommlet, M. Bogdan, D. Ramsey, Phenotype and Genotypes: The Search for Influential Genes, Springer, London, 2016, https://doi.org/10.1007/978-1-4471-5310-8.
[26] A. Xavier, W.M. Muir, B. Craig, K.M. Rainey, Walking through the statistical black boxes of plant breeding, Theor. Appl. Genet. 129 (2016) 1933–1949.
[27] I. Tattersall, Human origins: out of Africa, Proc. Natl. Acad. Sci. 106 (2009) 16018–16021.
[28] N. Bandillo, D. Jarquin, Q. Song, R. Nelson, P. Cregan, J. Specht, A. Lorenz, A population structure and genome-wide association analysis on the USDA soybean germplasm collection, Plant Genome 8 (2015), https://doi.org/10.3835/plantgenome2015.04.0024 Article #3.
[29] S. Wright, Breeding structure of populations in relation to speciation, Am. Nat. 74 (1940) 232–248.
[30] H. Innan, Y. Kim, Pattern of polymorphism after strong artificial selection in a domestication event, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 10667–10672.
[31] J. Doebley, B. Gaut, B. Smith, The molecular genetics of crop domestication, Cell 127 (2006) 1309–1321.
[32] K.R. Thornton, J.D. Jensen, C. Becquet, P. Andolfatto, Progress and prospects in mapping recent selection in the genome, Heredity 98 (2007) 340–348.
[33] J. Yu, G. Pressoir, W. Briggs, I. Bi, M. Yamasaki, J. Doebley, et al., A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, Nat. Genet. 38 (2006) 203–208.
[34] S. Wright, The genetical structure of populations, Ann. Hum. Genet. 15 (1949) 323–354.
[35] T. Thornton, J. Bermejo, Local and global ancestry inference and applications to genetic association analysis for admixed populations, Genet. Epidemiol. 38 (Suppl 1) (2014) S5–S12.
[36] S. Turner, L. Armstrong, Y. Bradford, C. Carlson, D. Crawford, A. Crenshaw, et al., Quality control procedures for genome-wide association studies, Curr. Protoc. Hum. Genet. 68 (SUPPL) (2011) 1–19, https://doi.org/10.1002/0471142905.hg0119s68.
[37] P. VanRaden, Genomic measures of relationship and inbreeding, Interbull Bull. (2007) 33–36.
[38] T.L. Odong, J. Van Heerwaarden, J. Jansen, T.J. van Hintum, F.A. Van Eeuwijk, Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? Theor. Appl. Genet. 123 (2011) 195–205.
[39] J.K. Pritchard, M. Stephens, N.A. Rosenberg, P. Donnelly, Association mapping in structured populations, Am. J. Hum. Genet. 67 (2000) 170–181.
[40] J. Stephan, O. Stegle, A. Beyer, A random forest approach to capture genetic effects in the presence of population structure, Nat. Commun. 6 (2015) 7432, https://doi.org/10.1038/ncomms8432.
[41] N. Patterson, A.L. Price, D. Reich, Population structure and eigenanalysis, PLoS Genet. 2 (2006) e190.
[42] A. Price, N. Patterson, R. Plenge, M. Weinblatt, N. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies, Nat. Genet. 38 (2006) 904–909.
[43] D. Lin, D. Zeng, Correcting for population stratification in genomewide association studies, J. Am. Stat. Assoc. 106 (2011) 997–1008.
[44] M. Mohammadi, J. Endelman, S. Nair, S. Chao, S. Jones, G. Muehlbauer, et al., Association mapping of grain hardness, polyphenol oxidase, total phenolics, amylose content, and β-glucan in us barley breeding germplasm, Mol. Breed. 34 (2014) 1229–1243.

[45] A. Poets, M. Mohammadi, K. Seth, H. Wang, T. Kono, Z. Fang, et al., The effects of both recent and long-term selection and genetic drift are readily evident in North American barley breeding populations, G3 Genes| Genomes| Genet. 6 (2016) 609–622.

[46] B.S. Weir, C.C. Cockerham, Estimating F-statistics for the analysis of population structure, Evolution 38 (1984) 1358–1370.

[47] E. Jorgenson, J. Witte, Coverage and power in genomewide association studies, Am. J. Hum. Genet. 78 (2006) 884–888.

[48] R. Brenchley, M. Spannagl, M. Pfeifer, G. Barker, R. D'Amore, A. Allen, et al., Analysis of the bread wheat genome using whole-genome shotgun sequencing, Nature 491 (2012) 705–710.

[49] P. Schnable, D. Ware, R. Fulton, J. Stein, F. Wei, S. Pasternak, et al., The B73 maize genome: complexity, diversity, and dynamics, Science 326 (2009) 1112–1115.

[50] Y. Bian, Q. Yang, P. Balint-Kurti, R. Wisser, J. Holland, Limits on the reproducibility of marker associations with southern leaf blight resistance in the maize nested association mapping population, BMC Genomics 15 (1068) (2014) doi.org/10.1186/1471-2164-15-1068.

[51] S. Gabriel, S. Schaffner, H. Nguyen, J. Moore, J. Roy, B. Blumenstiel, et al., The structure of haplotype blocks in the human genome, Science 296 (2002) 2225–2229.

[52] G. Johnson, L. Esposito, B. Barratt, A. Smith, J. Heward, G. Di Genova, et al., Haplotype tagging for the identification of common disease genes, Nat. Genet. 29 (2001) 233–237.

[53] K. Ding, I. Kullo, Methods for the selection of tagging SNPs: a comparison of tagging efficiency and performance, Eur. J. Hum. Genet. 15 (2007) 228–236.

[54] A. Elmas, T. Yang, X. Wang, D. Anastassiou, Discovering genome-wide tag SNPs based on the mutual information of the variants, PLoS One 11 (12) (2016).

[55] C.S. Carlson, M.A. Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, D.A. Nickerson, Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium, Am. J. Hum. Genet. 74 (1) (2004) 106–120.

[56] X. Ke, M. Miretti, J. Broxholme, S. Hunt, S. Beck, D. Bentley, et al., A comparison of tagging methods and their tagging space, Hum. Mol. Genet. 14 (2005) 2757–2767.

[57] K. Ahmadi, M. Weale, Z. Xue, N. Soranzo, D. Yarnall, J. Briley, et al., A single-nucleotide polymorphism tagging set for human drug metabolism and transport, Nat. Genet. 37 (2005) 84–89.

[58] A. Lorenz, M. Hamblin, J. Jannink, Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley, PLoS One 5 (11) (2010).

[59] C. Maldonado, F. Mora, C. Scapim, M. Coan, Genome-wide haplotype-based association analysis of key traits of plant lodging and architecture of maize identifies major determinants for leaf angle: hapLA4, PLoS One 14 (3) (2019).

[60] A. N'Diaye, J. Haile, A. Cory, F. Clarke, J. Clarke, R. Knox, C. Pozniak, 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 (1) (2017).

[61] R. Contreras-Soto, F. Mora, M. de Oliveira, W. Higashi, C. Scapim, I. Schuster, A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis, PLoS One 12 (2) (2017).

[62] F. Li, W. Wen, J. Liu, Y. Zhang, S. Cao, Z. He, et al., Genetic architecture of grain yield in bread wheat based on genome-wide association studies, BMC Plant Biol. 19 (1) (2019) 168.

[63] D. Falconer, T. Mackay, R. Frankham, Introduction to Quantitative Genetics, 4th ed, (1996).

[64] E. Cirulli, D. Goldstein, Uncovering the roles of rare variants in common disease through whole-genome sequencing, Nat. Rev. Genet. 11 (2010) 415–425.

[65] S. Dickson, K. Wang, I. Krantz, H. Hakonarson, D.B. Goldstein, Rare variants create synthetic genome-wide associations, PLoS Biol. 8 (1) (2010) e1000294, , https://doi.org/10.1371/journal.pbio.1000294.

[66] P. Zheng, W. Allen, K. Roesler, M. Williams, S. Zhang, J. Li, et al., A phenylalanine in *DGAT* is a key determinant of oil content and composition in maize, Nat. Genet. 40 (2008) 367–372.

[67] J. Cook, M. McMullen, J. Holland, F. Tian, P. Bradbury, J. Ross-Ibarra, et al., Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels, Plant Physiol. 158 (2012) 824–834.

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

[69] A. Karn, J. Gillman, S. Flint-Garcia, Genetic analysis of teosinte alleles for kernel composition traits in maize, G3 Genes| Genomes| Genet. 7 (2017) 1157–1164.

[70] W. Bodmer, C. Bonilla, Common and rare variants in multifactorial susceptibility to common diseases, Nat. Genet. 40 (2008) 695–701.

[71] G. Gibson, Rare and common variants: twenty arguments, Nat. Rev. Genet. 13 (2012) 135–145.

[72] J.K. Pritchard, Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. 69 (2001) 124–137.

[73] M. McCarthy, J. Hirschhorn, Genome-wide association studies: potential next steps on a genetic journey, Hum. Mol. Genet. 17 (2008) R156–R165.

[74] M. Momen, A.A. Mehrgardi, M.A. Roudbar, A. Kranis, R.M. Pinto, B.D. Valente, G. Morota, G.J. Rosa, D. Gianola, Including phenotypic causal networks in genome-wide association studies using mixed effects structural equation models, bioRxiv 2018 (2018) 251421, , https://doi.org/10.1101/251421.

[75] J. Zhu, Analysis of conditional genetic effects and variance components in developmental genetics, Genetics 141 (1995) 1633–1639.

[76] H. Zhang, J. Chen, R. Li, Z. Deng, K. Zhang, B. Liu, et al., Conditional QTL mapping of three yield components in common wheat (*Triticum aestivum* l.), Crop J. 4 (2016) 220–228.

[77] S. Banerjee, B. Yandell, N. Yi, Bayesian quantitative trait loci mapping for multiple traits, Genetics 179 (2008) 2275–2289.

[78] Z. Deng, Z. Liang, L. Bin, K. Zhang, J. Chen, H. Qu, et al., Conditional QTL mapping of sedimentation volume on seven quality traits in common wheat, J. Integr. Agric. 12 (2013) 2125–2133.

[79] Y. Uga, K. Sugimoto, S. Ogawa, J. Rane, M. Ishitani, N. Hara, et al., Control of root system architecture by deeper rooting 1 increases rice yield under drought conditions, Nat. Genet. 45 (2013) 1097–1102.

[80] J. Pritchard, M. Przeworski, Linkage disequilibrium in humans: models and data, Am. J. Hum. Genet. 69 (2001) 1–14.

[81] F. Cole, L. Kauppi, J. Lange, I. Roig, R. Wang, S. Keeney, et al., Homeostatic control of recombination is implemented progressively in mouse meiosis, Nat. Cell Biol. 14 (2012) 424–430.

[82] K. Hillers, A. Villeneuve, Chromosome-wide control of meiotic crossing over in *C. elegans*, Curr. Biol. 13 (2003) 1641–1647.

[83] E. Martini, R. Diaz, N. Hunter, S. Keeney, Crossover homeostasis in yeast meiosis, Cell 126 (2006) 285–295.

[84] G. Sidhu, C. Fang, M. Olson, M. Falque, O. Martin, W. Pawlowski, Recombination patterns in maize reveal limits to crossover homeostasis, Proc. Natl. Acad. Sci. 112 (2015) 15982–15987.

[85] W. Hanneman, M. Legare, S. Sweeney, J. Schimenti, Cisplatin increases meiotic crossing-over in mice, Proc. Natl. Acad. Sci. 94 (1997) 8681–8685.

[86] I. Kovalchuk, O. Kovalchuk, V. Kalck, V. Boyko, J. Filkowski, M. Heinlein, et al., Pathogen- induced systemic plant signal triggers DNA rearrangements, Nature 423 (2003) 760–762.

[87] J. Lucht, B. Mauch-Mani, H. Steiner, J. Metraux, J. Ryals, B. Hohn, Pathogen stress increases somatic recombination frequency in Arabidopsis, Nat. Genet. 30 (2002) 311–314.

[88] F. Zemp, C. Sidler, I. Kovalchuk, Increase in recombination rate in *Arabidopsis thaliana* plants sharing gaseous environment with X-ray and UVC-irradiated plants depends on production of radicals, Plant Signal. Behav. 7 (2012) 782–787.

[89] J. Chinnici, Modification of recombination frequency in drosophila. I. Selection for increased and decreased crossing over, Genetics 69 (1971) 71–83.

[90] B. Charlesworth, D. Charlesworth, Genetic variation in recombination in *Drosophila*. Ii. Genetic analysis of a high recombination stock, Heredity 54 (1985) 85–98.

[91] A. Kong, G. Thorleifsson, H. Stefansson, G. Masson, A. Helgason, D. Gudbjartsson, et al., Sequence variants in the *rnf212* gene associate with genome-wide recombination rate, Science 319 (2008) 1398–1401.

[92] C. Sandor, W. Li, W. Coppieters, T. Druet, C. Charlier, M. Georges, Genetic variants in *rec8*, *rnf212*, and *prdm9* influence male recombination in cattle, PLoS Genet. 8 (2012) e1002854.

[93] S. Johnston, C. Berenos, J. Slate, J. Pemberton, Conserved genetic architecture underlying individual recombination rate variation in a wild population of soay sheep (*Ovis aries*), Genetics 203 (2016) 583–598.

[94] C. Chen, W. Zhang, L. Timofejeva, Y. Gerardin, H. Ma, The Arabidopsis rock-n-rollers gene encodes a homolog of the yeast atp-dependent DNA helicase *mer3* and is required for normal meiotic crossover formation, Plant J. 43 (2005) 321–334.

[95] R. Mercier, S. Jolivet, D. Vezon, E. Huppe, L. Chelysheva, M. Giovanni, et al., Two meiotic crossover classes cohabit in Arabidopsis: one is dependent on *mer3*, whereas the other one is not, Curr. Biol. 15 (2005) 692–701.

[96] G. Chietera, F. Chardon,) Natural variation as a tool to investigate nutrient use efficiency in plants, Nutrient Use Efficiency in Plants, Springer, 2014, pp. 29–50.

[97] M. Zhu, S. Zhao, Candidate gene identification approach: progress and challenges, Int. J. Biol. Sci. 3 (2007) 420–427.

[98] J. Alonso, A. Stepanova, T. Leisse, C. Kim, H. Chen, P. Shinn, et al., Genome-wide insertional mutagenesis of *Arabidopsis thaliana*, Science 301 (2003) 653–657.

[99] J. Jeon, S. Lee, K. Jung, S. Jun, D. Jeong, J. Lee, et al., T-DNA insertional mutagenesis for functional genomics in rice, Plant J. 22 (2000) 561–570.

[100] P. Hilson, J. Allemeersch, T. Altmann, S. Aubourg, A. Avon, J. Beynon, et al., Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications, Genome Res. 14 (2004) 2176–2189.

[101] K. Krasileva, H. Vasquez-Gross, T. Howell, P. Bailey, F. Paraiso, L. Clissold, et al., Uncovering hidden variation in polyploid wheat, Proc. Natl. Acad. Sci. 114 (2017) E913–E921.

[102] A. Malzahn, L. Lowder, Y. Qi, Plant genome editing with TALEN and CRISPR, Cell Biosci. 7 (2017) 21, https://doi.org/10.1186/s13578-017-0148-4.

[103] B. Collard, D. Mackill, Marker-assisted selection: an approach for precision plant breeding in the twenty-first century, Philos. Trans. R. Soc. Lond. B Biol. Sci. 363 (2008) 557–572.

[104] T. Wurschum, Mapping QTL for agronomic traits in breeding populations, Theor. Appl. Genet. 125 (2012) 201–210.

[105] J.-L. Jannink, M. Bink, R. Jansen, Using complex plant pedigrees to map valuable genes, Trends Plant Sci. 6 (2001) 337–342.

[106] R. Bernardo, Molecular markers and selection for complex traits in plants: learning from the last 20 years, Crop Sci. 48 (2008) 1649–1664.

[107] J. Holland, Epistasis and plant breeding, Plant Breed. Rev. 21 (2001) 27–92.

[108] B. Vilhja´lmsson, M. Nordborg, The nature of confounding in genome-wide association studies, Nat. Rev. Genet. 14 (2013) 1–2.

[109] J. Bocianowski, Epistasis interaction of QTL effects as a genetic parameter influencing estimation of the genetic additive effect, Genet. Mol. Biol. 36 (2013) 93–100.

[110] A. Xavier, S. Xu, W.M. Muir, K.M. Rainey, NAM: association studies in multiple populations, Bioinformatics 31 (2015) 3862–3864.

[111] S. Xu, Mapping quantitative trait loci by controlling polygenic background effects, Genetics 195 (2013) 1209–1222.

[112] S. Xu, Estimating polygenic effects using markers of the entire genome, Genetics 163 (2013) 789–801.

[113] S. Xu, An empirical Bayes method for estimating epistatic effects of quantitative trait loci, Biometrics 63 (2007) 513–521.

[114] B. Mathew, J. Léon, W. Sannemann, M.J. Sillanpää, Detection of Epistasis for

Flowering Time Using Bayesian Multilocus Estimation in a Barley MAGIC Population, Genetics 208 (2008) 525–536.
[115] E. Sasaki, P. Zhang, S. Atwell, D. Meng, M. Nordborg, "Missing" G x E variation controls flowering time in *Arabidopsis thaliana*, PLoS Genet. 11 (2015) e1005597.
[116] D. Crews, Phenotypic plasticity: functional and conceptual approaches, Am. J. Hum. Biol. 17 (2005) 124–125.
[117] S. Grenier, P. Barre, I. Litrico, Phenotypic plasticity and selection: nonexclusive mechanisms of adaptation, Scientifica (2016) 2016.
[118] J. Guntrip, R. Sibly, Phenotypic plasticity, genotype-by-environment interaction and the analysis of generalism and specialization in *callosobruchus maculatus*, Heredity 81 (1998) 198–204.
[119] S. Via, R. Lande, Genotype-environment interaction and the evolution of phenotypic plasticity, Evolution 39 (1985) 505–522.
[120] D. Ebert, L. Yampolsky, A. Van Noordwijk, Genetics of life history in *daphnia magna*. ii. Phenotypic plasticity, Heredity 70 (1993) 344–344.
[121] V. Sadras, G. Rebetzke, Plasticity of wheat grain yield is associated with plasticity of ear number, Crop Pasture Sci. 64 (2013) 234–243.
[122] V. Sadras, C. Lawson, Genetic gain in yield and associated changes in phenotype, trait plasticity and competitive ability of south Australian wheat varieties released between 1958 and 2007, Crop Pasture Sci. 62 (2011) 533–549.
[123] X. Chen, F. Zhao, S. Xu, Mapping environment-specific quantitative trait loci, Genetics 186 (2010) 1053–1066.
[124] J. Crossa, From genotype x environment interaction to gene x environment interaction, Curr. Genomics 13 (2012) 225–244.
[125] A. Korol, Y. Ronin, E. Nevo, Approximate analysis of QTL-environment interaction with no limits on the number of environments, Genetics 148 (1998) 2015–2028.
[126] C. Te´tard-Jones, M. Kertesz, R. Preziosi, Quantitative trait loci mapping of phenotypic plasticity and genotype–environment interactions in plant and insect performance, Philos. Trans. R. Soc. Lond., B, Biol. Sci. 366 (2011) 1368–1379.
[127] J.H. Sul, M. Bilow, W.Y. Yang, E. Kostem, N. Furlotte, D. He, E. Eskin, Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models, PLoS Genet. 12 (2016) e1005849.
[128] A. Xavier, D. Jarquin, R. Howard, V. Ramasubramanian, J.E. Specht, G.L. Graef, W.D. Beavis, B.W. Diers, Q. Song, P.B. Cregan, R. Nelson, K.M. Rainey, Genome-Wide analysis of grain yield stability and environmental interactions in a multiparental soybean population, G3 Genes| Genomes| Genet. 8 (2018) 519–529.
[129] H. Raman, R. Raman, N. Coombes, J. Song, S. Diffey, A. Kilian, et al., Genome-wide association study identifies new loci for resistance to *Leptosphaeria maculans* in canola, Front. Plant Sci. 7 (2016) Article #1513 DOI.org/10.3389/fpls.2016.01513.
[130] M. Maccaferri, J. Zhang, P. Bulli, Z. Abate, S. Chao, D. Cantu, et al., A genome-wide association study of resistance to stripe rust (*Puccinia striiformis* f. sp. tritici) in a worldwide collection of hexaploid spring wheat (*Triticum aestivum* l.), G3: Genes Genomes Genet. 5 (2015) 449–465.
[131] J. Pace, N. Lee, H. Naik, B. Ganapathysubramanian, T. Lubberstedt, Analysis of maize (*Zea mays* L.) seedling roots with the high-throughput image analysis tool ARIA (automatic root image analysis), PLoS One 9 (2014) e108255.
[132] F. Biscarini, P. Cozzi, L. Casella, P. Riccardi, A. Vattari, G. Orasen, et al., Genome-wide association study for traits related to plant and grain morphology, and root architecture in temperate rice accessions, PLoS One 11 (2016) e0155425.
[133] J. Markham, Measuring plant neighbour effects, Funct. Ecol. 10 (1996) 548–549.
[134] J. Mutic, J. Wolf, Indirect genetic effects from ecological interactions in *Arabidopsis thaliana*, Mol. Ecol. 16 (2007) 2371–2381.
[135] M. El-Soda, M. Malosetti, B. Zwaan, M. Koornneef, M. Aarts, Genotype × environment interaction QTL mapping in plants: lessons from Arabidopsis, Trends Plant Sci. 19 (2014) 390–398.
[136] H. Gauch, P. Rodrigues, J. Munkvold, E. Heffner, M. Sorrells, Two new strategies for detecting and understanding QTL × environment interactions, Crop Sci. 51 (2011) 96–113.
[137] M. Ungerer, S. Halldorsdottir, M. Purugganan, T. Mackay, Genotype-environment interactions at quantitative trait loci affecting inflorescence development in *Arabidopsis thaliana*, Genetics 165 (2003) 353–365.
[138] D. Yang, R. Jing, X. Chang, W. Li, Identification of quantitative trait loci and environmental interactions for accumulation and remobilization of water-soluble carbohydrates in wheat (*Triticum aestivum* L.) stems, Genetics 176 (2007) 571–584.
[139] X. Wang, Y. Pang, J. Zhang, Q. Zhang, Y. Tao, B. Feng, et al., Genetic background effects on QTL and QTL × environment interaction for yield and its component traits as revealed by reciprocal introgression lines in rice, Crop J. 2 (2014) 345–357.
[140] C. Cavanagh, M. Morell, I. Mackay, W. Powell, From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants, Curr. Opin. Plant Biol. 11 (2008) 215–221.
[141] Y. Zhang, J. Liu, Bayesian inference of epistatic interactions in case-control studies, Nat. Genet. 39 (2007) 1167–1173.
[142] J. Yu, J. Holland, M. McMullen, E. Buckler, Genetic design and statistical power of nested association mapping in maize, Genetics 178 (2008) 539–551.
[143] U. Wintersberger, On the origins of genetic variants, FEBS Lett. 285 (1991) 160–164.
[144] D. Torkamaneh, J. Laroche, A. Tardivel, L. O'Donoughue, E. Cober, I. Rajcan, et al., Comprehensive description of genome-wide nucleotide and structural variation in short-season soybean, Plant Biotechnol. J. 16 (2018) 749–759, https://doi.org/10.1111/pbi.12825.
[145] N.D. Young, P. Zhou, K.A. Silverstein, Exploring structural variants in environmentally sensitive gene families, Curr. Opin. Plant Biol. 30 (2016) 19–24.
[146] R.K. Saxena, D. Edwards, R.K. Varshney, Structural variations in plant genomes, Brief. Funct. Genomics 13 (2014) 296–307.
[147] A. Xavier, B. Hall, A.A. Hearst, K.A. Cherkauer, K.M. Rainey, Genetic architecture of phenomic-enabled canopy coverage in Glycine max, Genetics 206 (2017) 1081–1089.
[148] A.S. Kaler, J.D. Ray, W.T. Schapaugh, M.K. Davies, C.A. King, L.C. Purcell, Association mapping identifies loci for canopy coverage in diverse soybean genotypes, Mol. Breed. 38 (2018) 50.
[149] M.M. Muraya, J. Chu, Y. Zhao, A. Junker, C. Klukas, J.C. Reif, T. Altmann, Genetic variation of growth dynamics in maize (*Zea mays* L.) revealed through automated non-invasive phenotyping, Plant J. 89 (2017) 366–380.
[150] K. Sikorska, N.M. Montazeri, A. Uitterlinden, F. Rivadeneira, P.H. Eilers, E. Lesaffre, GWAS with longitudinal phenotypes: performance of approximate procedures, Eur. J. Hum. Genet. 23 (2015) 1384–1391.
[151] C. Ning, H. Kang, L. Zhou, D. Wang, H. Wang, A. Wang, J. Fu, S. Zhang, J. Liu, Performance gains in genome-wide association studies for longitudinal traits via modeling time-varied effects, Sci. Rep. 7 (2017), https://doi.org/10.1038/s41598-017-00638-2 Article #590.
[152] R. Wu, M. Lin, Functional mapping—how to map and study the genetic architecture of dynamic complex traits, Nat. Rev. Genet. 7 (2006) 229–237.
[153] H. Kang, L. Zhou, R. Mrode, Q. Zhang, J.F. Liu, Incorporating the single-step strategy into a random regression model to enhance genomic prediction of longitudinal traits, Heredity 119 (2017) 459–467.
[154] L.R. Schaeffer, Strategy for applying genome-wide selection in dairy cattle, J. Anim. Breed. Genet. 123 (2006) 218–223.
[155] J. Sun, J.E. Rutkoski, J.A. Poland, J. Crossa, J.L. Jannink, M.E. Sorrells, Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield, Plant Genome 10 (2017) 1–12.
[156] I.C. Marschner, A.C. Gillett, Relative risk regression: reliable and flexible methods for log-binomial models, Biostatistics 13 (2011) 179–192.
[157] Z. Zhang, E. Ersoz, C.Q. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, P.J. Bradbury, J. Yu, D.K. Arnett, J.M. Ordovas, E.S. Buckler, Mixed linear model approach adapted for genome-wide association studies, Nat. Genet. 42 (2010) 355–360.
[158] S. Lee, M.C. Wu, X. Lin, Optimal tests for rare variant effects in sequencing association studies, Biostatistics 13 (2012) 762–775.
[159] H. Kang, N. Zaitlen, C. Wade, A. Kirby, D. Heckerman, M. Daly, E. Eskin, Efficient control of population structure in model organism association mapping, Genetics 178 (2008) 1709–1723, https://doi.org/10.1534/genetics.107.080101.
[160] X. Wang, V.M. Philip, G. Ananda, C.C. White, A. Malhotra, P.J. Michalski, K.R. Karuturi, S.R. Chintalapudi, C. Acklin, M. Sasner, D.A. Bennett, A Bayesian framework for generalized linear mixed modeling identifies new candidate loci for late-onset Alzheimer's disease, Genetics 209 (2018) 51–64.
[161] G. de los Campos, J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, M.P. Calus, Whole-genome regression and prediction methods applied to plant and animal breeding, Genetics 193 (2013) 327–345.
[162] N. Yi, S. Xu, Bayesian LASSO for quantitative trait loci mapping, Genetics. 179 (2) (2008) 1045–1055.
[163] R.B. O'Hara, M.J. Sillanpää, A review of Bayesian variable selection methods: what, how and which, Bayesian Anal. 4 (2009) 85–117.
[164] R.L. Fernando, D. Garrick, Bayesian methods applied to GWAS, Genome-Wide Association Studies and Genomic Prediction, Humana Press, Totowa, NJ, 2013, pp. 237–274.
[165] E. De Maturana, N. Ibáñez-Escriche, Ó González-Recio, G. Marenne, H. Mehrban, S. Chanock, M. Goddard, N. Malats, Next generation modeling in GWAS: comparing different genetic architectures, Hum. Genet. 133 (2014) 1235–1253.
[166] V. Segura, B. Vilhjálmsson, A. Platt, et al., An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations, Nat. Genet. 44 (2012) 825–830, https://doi.org/10.1038/ng.2314.
[167] X. Liu, M. Huang, B. Fan, E. Buckler, Z. Zhang, Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies, PLoS Genet. 12 (2) (2016) e1005767, , https://doi.org/10.1371/journal.pgen.1005767.
[168] A. Kaler, J. Gillman, T. Beissinger, L. Purcell, Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize, Front. Plant Sci. 10 (2020) 1794, https://doi.org/10.3389/fpls.2019.01794.
[169] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[170] B.A. Goldstein, E.C. Polley, F.B. Briggs, Random forests for genetic association studies, Stat. Appl. Genet. Mol. Biol. 10 (2011) 1544–6115, https://doi.org/10.2202/1544-6115.1691.
[171] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning; Data Mining, Inference and Prediction, ISBN 978-0-387-84858-7 (2008).
[172] V. Botta, G. Louppe, P. Geurts, L. Wehenkel, Exploiting SNP correlations within random forest for genome-wide association studies, PLoS One 9 (2014) e93379.
[173] Y.A. Meng, Y. Yu, L.A. Cupples, L.A. Farrer, K.L. Lunetta, Performance of random forest when SNPs are in linkage disequilibrium, BMC Bioinformatics 10 (2009) 78.