Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

From association to prediction: statistical methods for the dissection and selection of complex traits in plants

Alexander E Lipka[1], Catherine B Kandianis[2,3], Matthew E Hudson[1], Jianming Yu[4], Jenny Drnevich[5], Peter J Bradbury[6] and Michael A Gore [3]

Quantification of genotype-to-phenotype associations is central to many scientific investigations, yet the ability to obtain consistent results may be thwarted without appropriate statistical analyses. Models for association can consider confounding effects in the materials and complex genetic interactions. Selecting optimal models enables accurate evaluation of associations between marker loci and numerous phenotypes including gene expression. Significant improvements in QTL discovery via association mapping and acceleration of breeding cycles through genomic selection are two successful applications of models using genome-wide markers. Given recent advances in genotyping and phenotyping technologies, further refinement of these approaches is needed to model genetic architecture more accurately and run analyses in a computationally efficient manner, all while accounting for false positives and maximizing statistical power.

Addresses
[1] University of Illinois, Department of Crop Sciences, Urbana, IL 61801, USA
[2] Michigan State University, Department of Biochemistry and Molecular Biology, East Lansing, MI 48824, USA
[3] Cornell University, Plant Breeding and Genetics Section, School of Integrative Plant Science, Ithaca, NY 14853, USA
[4] Iowa State University, Department of Agronomy, Ames, IA 50011, USA
[5] University of Illinois, High Performance Biological Computing Group and the Carver Biotechnology Center, Urbana, IL 61801, USA
[6] United States Department of Agriculture (USDA) – Agricultural Research Service (ARS), Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853, USA

Corresponding author: Lipka, Alexander E (alipka@illinois.edu)

Current Opinion in Plant Biology 2015, **24**:110–118

This review comes from a themed issue on **Genome studies and molecular genetics**

Edited by **Insuk Lee** and **Todd C Mockler**

For a complete overview see the Issue and the Editorial

Available online 17th March 2015

http://dx.doi.org/10.1016/j.pbi.2015.02.010

Introduction

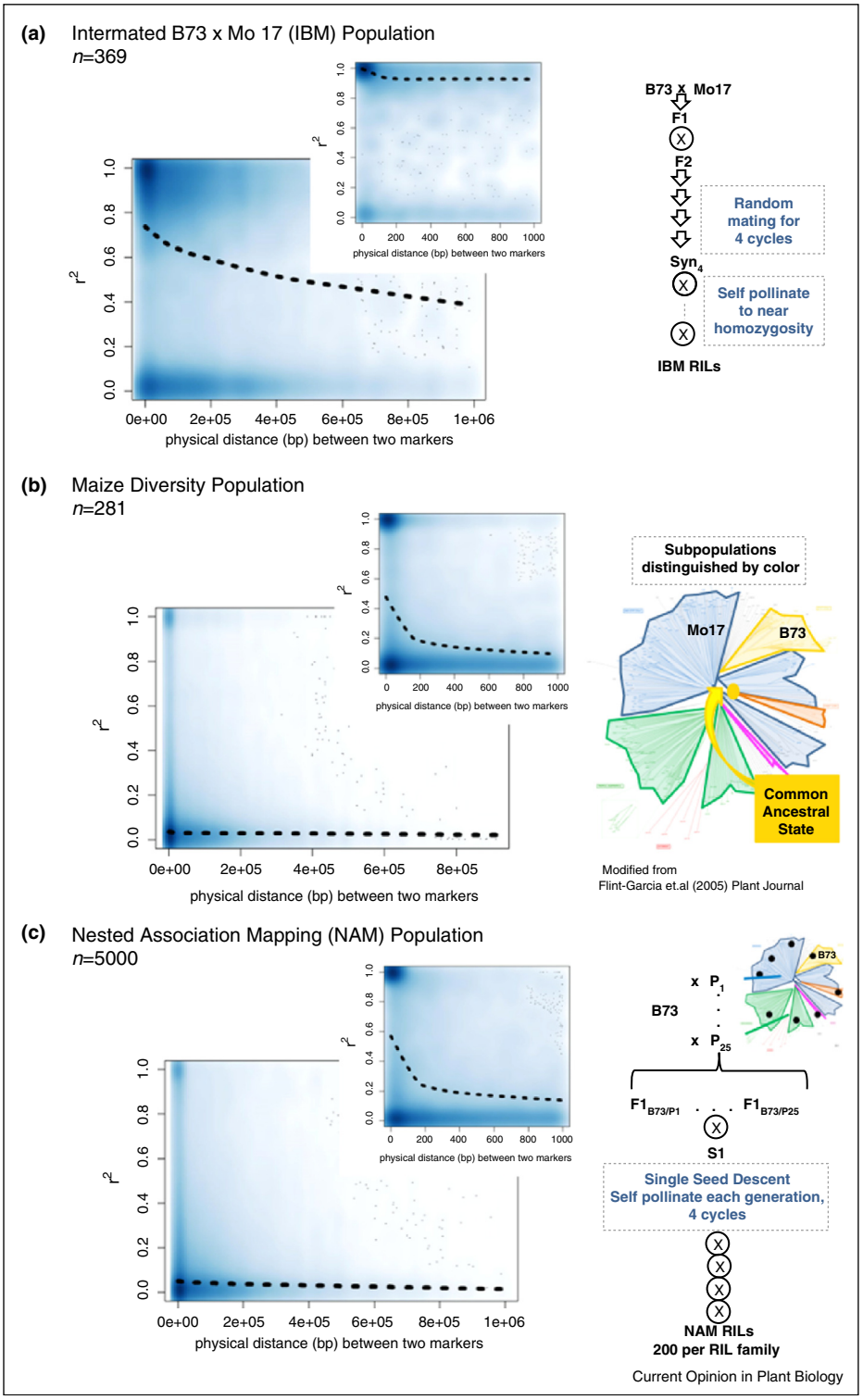
The ability to understand the genetic basis of biological phenomena requires the development and refinement of statistical approaches that identify associations between genetic markers and phenotypes. As genotypic data become easier to obtain, a more complete and complex landscape of genetic diversity is revealed, requiring more comprehensive statistical models that have the capacity to distinguish true biological associations from false positives arising from population structure and linkage disequilibrium (LD). Applications for the association of genetic markers with phenotypes have also expanded. In particular, the development of models predicting phenotypic values with genome-wide data sets can significantly accelerate plant breeding cycles. Although computationally efficient approaches have been developed to fit these models, the full potential of such analytical approaches has yet to be realized.

Genotyping technologies built on next-generation sequencing (NGS) have made it possible to obtain an unprecedented number of genetic markers and have enabled accurate quantification of gene expression at low cost. However, NGS has introduced new statistical challenges that need to be addressed, including consideration of rare alleles, the increase in computational time, and appropriate treatment of the multiple testing problem. In this review, we provide guidelines and suggestions for conducting an optimal statistical analysis that addresses these issues. Additionally, we highlight the most promising statistical approaches that should become more widespread in the plant genetics research community.

Genome-wide association study (GWAS): current practices and future perspectives

The two most widely used data sets for studying genetic variability are those derived from biparental crosses (e.g. F_2 populations or recombinant inbred lines [RILs], Figure 1a) and those that consist of individuals assembled with complex relatedness or geographical origin (e.g. diversity populations, Figure 1b). These two data sets differ with respect to the number of recombination events they capture (reviewed in [1••]). While biparental crosses only exploit recent recombination events that occurred during the establishment of the population, diversity populations (also called diversity panels) have captured all historical recombination events occurring during the evolution of the sampled individuals. Diversity populations are typically obtained by collecting as many informative samples as possible, with the intention that they

Figure 1



Linkage disequilibrium (LD) plots of squared correlations of allele frequencies (r^2, *y*-axis) against physical distance between SNP sites along maize chromosome 6 (*x*-axis) in three populations. Differing in crossing scheme, the populations include: **(a)** the intermated B73 × Mo17 (IBM) population, a biparental population designed to increase the frequency of recombinants for improved genetic resolution by successively random mating the F_2 population before selfing to derive intermated recombinant inbred lines (IRILs); **(b)** a maize diversity population, assembled to include a broad representation of allelic diversity and historical recombination events; **(c)** the nested association mapping (NAM) population, a set of RIL families derived from the common B73 parent and one of 25 different, diverse founder inbred lines, which combines recent and historical

are representative of the population under study (e.g. [2–4]). High-density genetic markers are then used to genotype the diversity population and to identify significant associations between marker genotypes and a phenotype of interest. Because it can be expected that sequences physically close to the markers are likely to be in LD with them, the markers need to cover the LD structure of the genome. In particular, they need to be in strong LD with the causative variants. For this reason, the design of both the optimal population and the marker distribution are both species-specific and population-specific.

It is also possible to create experimental crosses that utilize both historical and recent recombination events, for example the nested association mapping (NAM) design pioneered in maize [5•,6•] (Figure 1c). The NAM design involves crossing multiple diverse inbred lines to a common parent, usually a reference with a fully sequenced genome (e.g. B73 in maize), that ultimately yields multiple RIL families which share a common parent for joint linkage-association mapping of complex trait variation. This type of mapping population provides a common metric (i.e. the common parent) against which alleles from all non-common parents can be compared, and increases the statistical power for genetic mapping. On the basis of experience with the US maize NAM population [5•,7], it is recommended that NAM designs should allocate resources such that the number of diverse parents is higher than that in the original design. A larger number of parents enable the population to capture a greater number of historical recombination events, resulting in a community resource that should achieve a balance between diversity among parents and adequate sampling within each cross, thus allowing higher precision and power in mapping.

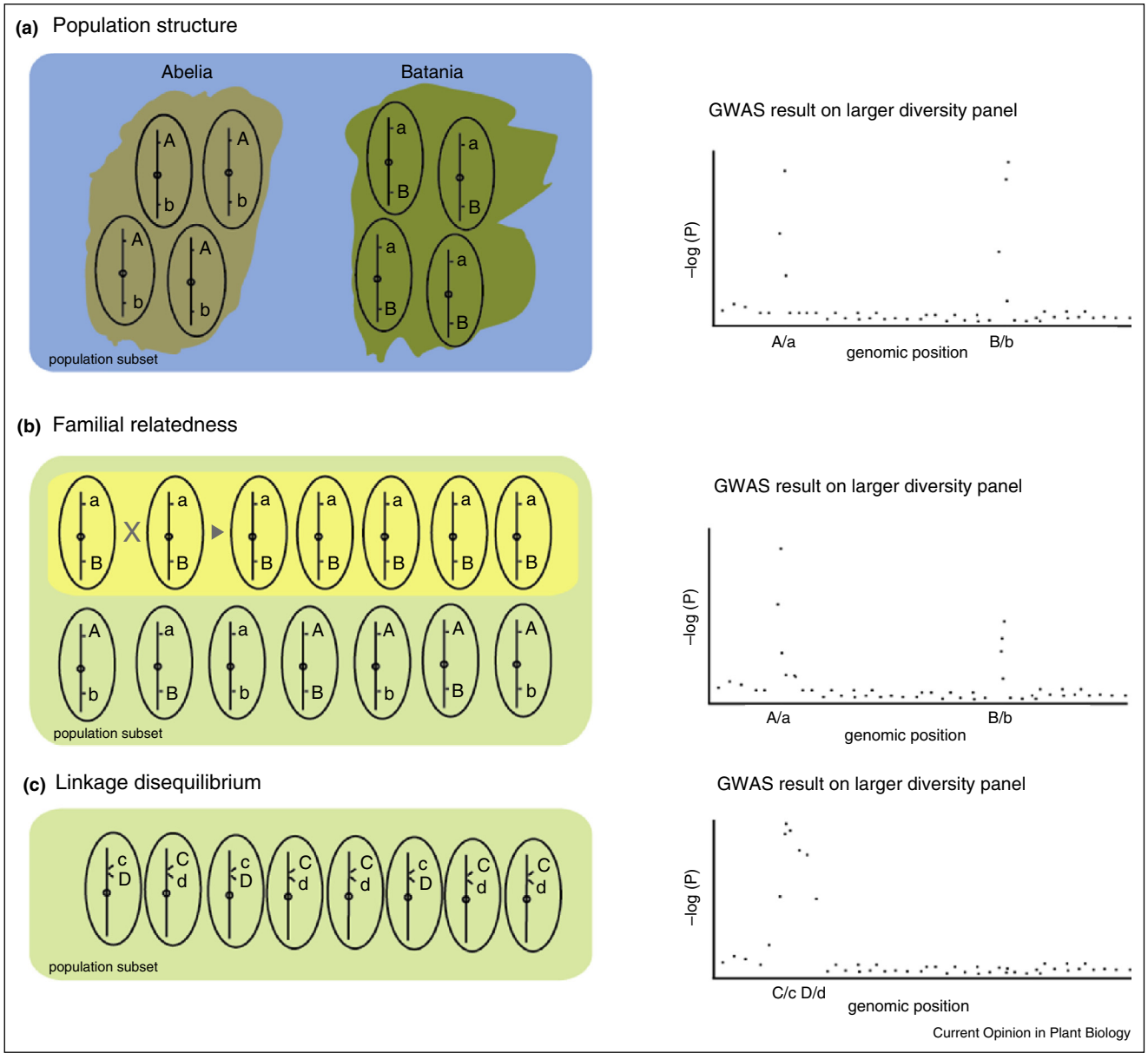
It is possible to consider the statistical analyses conducted in biparental crosses, diversity populations, and NAM designs as one fundamental approach, namely quantification of the associations between genetic markers and a phenotype. However, while biparental populations have a clearly defined population makeup defined by the crossing scheme, population structure and relatedness in diversity populations (Figure 2a and b) can be important sources of false signals. For this reason, the basic statistical model used in a QTL analysis is typically expanded in a GWAS with covariates for structure and kinship (reviewed in [8]). There are several approaches that use genetic markers to quantify the population structure present in a diversity population, some common ones being STRUCTURE (a Bayesian clustering approach) [9], principal component analysis (PCA) [10], and a discriminant analysis of principal components (DAPC) [11]. Commonly, covariates from STRUCTURE and principal components (PCs) from the PCA are included as fixed effects in the GWAS model to account for population structure (DAPC is promising but relatively untried for plant GWAS). Because STRUCTURE and PCA covariates perform roughly equally, we recommend using PCs to generate covariates as the computational resources needed for a PCA are much less than for STRUCTURE, which can be prohibitive for large marker data sets. If the definition of subpopulations provided by STRUCTURE is needed but high-performance computing is not available, DAPC provides a more efficient way to cluster the individuals. For each tested trait, we recommend conducting a model selection approach (e.g. BIC model selection available in [12]) to determine the optimal number of fixed effect covariates needed in the GWAS model [8].

Particularly in crop systems, germplasm accessions can often share recent common ancestry, which if not taken into account can also lead to spurious association signals (Figure 2b). For this reason, the inclusion of a kinship matrix into the GWAS model is often more important and powerful than the inclusion of fixed-effect population structure covariates for reducing false positives. There are several approaches available to calculate a kinship matrix that use marker information to estimate familial relatedness among the individuals in a diversity population. Of these, the approach from [13] is preferred because it has a biologically relevant interpretation, namely it uses identity-by-state to estimate identity-by-descent. Although correction for population structure and kinship removes a substantial amount of spurious signals in a GWAS, it must be noted that false positives arising from population structure may not always be completely controlled, particularly when the genetic divergence between subpopulations is extensive (e.g. [14]), and that there is a tremendous loss of statistical power for adaptive traits that are highly correlated with population structure.

The underlying mixed linear model used to detect statistical associations in a plant GWAS has remained unchanged since [15••]. However, many approaches have been developed since then to reduce the computational burden of testing the hundreds of thousands of markers in a typical GWAS (e.g. [16,17]), mainly through efficiently estimating variance components. Of these, the most widely used approaches in plant GWAS include efficient mixed-model association (EMMA) [18], EMMA

(Figure 1 Legend Continued) recombination events to maximize statistical power and mapping resolution while capturing moderate levels of allelic diversity. Using non-imputed SNP markers from the July 2012 All Zea GBS Final Build publicly available on Panzea.org, LD is estimated as r^2 between pairs of SNP sites based on 50-site windows along chromosome 6. Loess curves (black dotted line) approximate the decrease of r^2 (i.e. the LD decay) over two contrasting physical distances, specifically over 1 million bp (main figure) and over 1 thousand bp (inset). Plots show that LD decays with increasing physical distance, and with the increased number of recombination events.

Figure 2



(a) Common sources of association between traits and non-causal alleles in an uncorrected genome-wide association study (GWAS). A diversity population of wild accessions from two imaginary geographically isolated islands, Abelia and Batania, is used for GWAS on drought tolerance. Abelia has a dry, arid climate and Batania has a humid climate where fungi are common pathogens. A cartoon subset of this imaginary diversity population is shown. One allele in the population, A, confers drought tolerance, the cognate allele a does not. At a second, unlinked locus, the allele B confers fungus resistance, b does not. In the uncorrected GWAS, a false association between the B/b locus and drought tolerance is observed because of the population structure. The alleles are fixed in both subpopulations, thus association appears as two separate peaks of comparable significance on the Manhattan plot. Such structure can arise as a result of diversifying selection as well as geographic isolation, and fixation of the alleles is an extreme case. **(b)** Imaginary worldwide farmers and breeders have historically cultivated the same species as in (a), and a germplasm collection of crop varieties from other locations is genotyped. The germplasm collection is used as a diversity population for the same drought trait. In this case the population is not geographically or otherwise isolated, but many cultivar accessions in the panel are related. The green box represents a cartoon subset of the panel, and the yellow box a set of related accessions descended from two accessions from Batania, both with the aB haplotype. Since the progeny of the cross are still fixed at the two loci, and the experimenters do not correct for kinship, a false association is again seen for allele B/b in the Manhattan plot of the uncorrected GWAS for drought tolerance. **(c)** A well-mixed population contains two different loci, C/c (causal) and D/d (non-causal), which are in strong linkage disequilibrium (LD). Consequently, a strong association between the trait and a broad peak spanning the causal polymorphism at the C/c locus and the linked but non-causal D/d locus is observed. Because in this case we understand that the peak is attributable to the C/c locus only, the width of the peak suggests that the resolution of the GWAS is limited by LD. Thus regions or genomes with high LD can both limit resolution and allow discovery of associations using smaller numbers of markers.

eXpedited (EMMAX) [19], population parameters previously determined (P3D) [20], and the compressed mixed linear model [20,21]. One predominant factor leading to the popularity of these approaches is that they are freely available in easy-to-use software packages including Trait Analysis by aSSociation, Evolution and Linkage (TASSEL) [22] and Genome Association and Prediction Integrated Tool (GAPIT) [12]. Although there are more advanced approaches, they are not as widely used in the plant research community. Of these, the genome-wide efficient mixed-model association (GEMMA) approach [23] should become more widely adopted because it efficiently refits the mixed model at every marker, resulting in exact estimates of marker effects (rather than approximate estimates in EMMAX or P3D).

Linkage disequilibrium is fundamentally important to the use of any GWAS method where the genotyping does not cover every sequence variant in the genome (Figure 2c). Nonetheless, LD can hinder the ability to identify causal variants [24••]. This is also exemplified in a phenomenon called synthetic associations [25•,26•], when multiple low-frequency causative variants (i.e. allelic heterogeneity) spanning large distances are associated with a common variant that is detected in a GWAS. Unlike other sources of false positives, there is no known solution to synthetic associations. Moreover, the impact of synthetic associations is not improved by higher density marker genotyping methods, and is potentially exacerbated in newly developed augmented approaches [27,28] that simultaneously account for the multiple association signals underlying a trait. One example of these augmented approaches is the multi-locus mixed model (MLMM) [27], which conducts stepwise regression to identify multiple association signals while retaining the fixed and random effects from the model of [15••] to account for false positives. Although these models can simultaneously account for the effects of multiple loci, it is increasingly thought that epistasis can account for a substantial amount of 'missing heritability' [29], a phenomenon that is still often cited to show the challenge in dissecting complex traits [4]. Several new algorithms [30–32] are making it possible to utilize high-performance computing to test for two-way epistasis between all pairs of loci. In the absence of such computational resources, it is possible to narrow the search down to an *a priori* set of loci. For example, it is reasonable to test for two-way epistasis between all markers in the final model from the MLMM approach. Finally, the use of multivariate approaches [33] such as the multivariate linear mixed models (mvLMMs) [34] and the multi-trait mixed model (MTMM) [35] that associate genetic markers with several correlated traits will become crucial as high-throughput phenotyping approaches (e.g. [36,37]) become more widespread.

Given that a GWAS with optimized statistical power consists of hundreds to thousands of individual genotypes at tens of thousands to several millions of SNP markers, it should maximize the amount of information gained by also measuring as many phenotypes as possible. In addition to standard quantitative traits, gene expression levels may also be analyzed as phenotypes. The spread of microarray technology in the early 2000s led to 'eQTL' studies looking for associations between thousands of markers and thousands of gene expression levels [38••,39]. Thorough reviews of eQTL issues and pitfalls can be found elsewhere [40–44]. The technology of choice has changed to direct sequencing of mRNA (i.e. RNA-Seq) due to known problems with the hybridization method of arrays [45,46] and the decrease in cost of RNA-Seq. Nevertheless the total cost of sequencing mRNA from hundreds to thousands of individuals may still be prohibitive for most agricultural studies and the thousands of genes tested can make it difficult to find true associations among all the false positives [47]. Instead, expression measurements of a targeted subset of genes, perhaps defined using a pilot RNA-Seq study, are possibly a more accessible option with medium-throughput qPCR (e.g. Fluidigm). However, regardless of whether the gene expression measurements come from microarrays, RNA-Seq or qPCR, they are no different from any other phenotype. Thus a GWAS for gene expression should use the standard statistical models employed in a typical GWAS (e.g. [15••]) instead of those available in computational tools designed for an eQTL analysis, many of which do not account for population structure and relatedness [48,49].

The ability of a GWAS to identify loci associated with a phenotype will depend on the sample size and marker density. However even with sufficient marker density, it may not be possible to tag all possible sources of genetic variation present in a species using only SNPs. Thus, one major challenge with a GWAS is the marker technology itself; markers capable of tagging copy number variation, epigenetic variation, and transposons need to become further developed and more widely incorporated into the GWAS model. Although the inclusion of such markers could result in the identification of more loci, it may be nearly impossible to locate causal polymorphisms proximal to centromeres or other regions of recombination suppression or greater than expected LD. Another major issue limiting the ability of a GWAS to identify loci is the stringent correction of the multiple testing problem, which arises because of the number of markers that are tested [50]. To address this issue, some studies have augmented their GWAS with a targeted approach that tests only markers located within genomic regions proximal to an *a priori* set of candidate genes with subtle effects likely to control variation for studied phenotypes [51]. It is possible to obtain this set using either QTL from previous studies (e.g. [52,53]) or through grouping genes together using a function or network analysis (e.g. [54,55]). In addition to the marker data, an important

factor contributing to the detection of loci in a GWAS is the ability to reduce variation in phenotypic data attributable to non-genetic components. The inclusion of more replications, more environments, and the use of optimal laboratory techniques typically result in the reduction of this source of variation. Finally, it is important to recognize that the models used in a GWAS have assumptions that need to be met. If they are not, alternate statistical models (e.g. negative binomial regression model) or alternate approaches to identify genomic regions associated with a trait (e.g. a meta analysis) may be considered. While some statistical barriers may have existing, straightforward solutions, collaboration with biostatisticians during the entire course of the study may ensure that the most appropriate statistical analysis is being conducted.

Genomic selection to improve plant breeding practices

One of the most exciting new approaches is genomic selection (GS) [56•,57•], which uses statistical models to predict which individuals will have optimal phenotypes based on marker data. This approach holds great promise for plant breeding efforts because it can theoretically achieve multiple cycles of selection in the amount of time required to complete one cycle using phenotype-based selection approaches [58•,59•]. In contrast to the statistical models typically used in a GWAS, GS models include all genome-wide markers that pass quality control standards. Thus GS models use both large-effect and small-effect loci to predict genomic estimated breeding values. Given the number of markers (p) and sample size (n) in a typical diversity population, GS models usually have more predictors than the sample size (a $p \gg n$ model), resulting in an infinite number of possible marker effect estimates [60•]. To address this, a multitude of frequentist and Bayesian penalized regression approaches have been evaluated for GS. Two recent reviews comprehensively summarize most of the statistical models that have been assessed for GS, the general guidelines for reporting results, and how to accurately quantify prediction accuracies [61••,62••]. In practice, most statistical models in GS produce prediction accuracies that are relatively indistinguishable from one another [63•]. Therefore we recommend using the RR-BLUP model [56•,64] because it is one of the most computationally efficient approaches and is implemented in an R package, rrBLUP [65]. Some practical considerations for GS breeding programs include the selection of an optimal training population for fitting the initial GS model and the need to refit the GS model after several generations of selection to account for changes in allele frequencies, LD, and population structure [59•,66•]. Ultimately, GS has the potential to complement phenotypic selection by allowing breeders to select for traits that are difficult to phenotype and to accomplish selection cycles in non-adaptive environments such as winter nurseries.

Maximizing genomic information from latest sequencing technologies

The advent of NGS has made it possible to affordably obtain markers with genome-wide coverage using techniques as appropriate including genotyping-by-sequencing (GBS) [67•], restriction site associated DNA sequencing (RAD-seq) [68], target region resequencing or whole-genome resequencing from any species. Previously, markers for a GWAS were obtained through high-density SNP arrays such as diversity array technology (DArT) [69], Illumina Infinium or Affymetrix, which still dominate human studies but are being replaced by NGS for most plant studies. Given an appropriate experimental design, it is reasonable to assume that a GWAS will detect genetic signals associated with a heritable trait controlled by several large-effect loci. The number of genetic markers needed to achieve this goal depends on many species-specific and experiment-specific factors, the most important being genome size, rate of LD decay, and magnitude of the QTL effect. An ideal data set in the perfect world would include every polymorphism in the genome as a marker and include enough individuals to enable identification of causal mutations of every gene responsible for the variation of a given trait. While marker coverage is a major experimental consideration, the shortcomings of available sequencing technologies providing marker data points is an equally important concern. In general, NGS-based approaches yield more markers than DArT or SNP arrays; however, NGS-derived markers tend to have higher missing rates, especially for technologies such as GBS or RAD-seq. Array-based techniques such as DArT and Illumina use only a subset of the individuals, often not those being studied, to obtain polymorphisms fixed on the array and can thus be subject to severe ascertainment bias. Such bias results in overestimation of the divergence between individuals that are genetically similar to those used to obtain the markers. Another important factor is the availability of reference genomes in the species under study. A single reference genome also results in ascertainment bias in the markers, with lines most related to the reference genome typically having the lowest missing rate. A reference genome is required for imputation approaches for which marker order must be known (e.g. Beagle4 [70] and fast inbred line library imputation [FILLIN] [71]). Challenges remain with imputation techniques, for example distinguishing between missing data for biological reasons (e.g. presence/absence variation) and those arising from sampling variation.

One crucial aspect that needs to be studied more thoroughly is the contribution of rare alleles to complex trait variation [72]. All markers below a certain minor allele frequency (MAF) threshold are typically removed because most of the approaches used in a GWAS are not appropriate for analyzing them. As such, it becomes nearly impossible to assess the contribution of rare

variants to phenotypic variation using traditional approaches outside of creating a biparental cross that segregates for the rare variant (thus rendering it common in the population under study). However, improvements in genotyping technologies and decreases in sequencing costs have resulted in diversity populations with exceptionally large sample sizes, providing the ability to include markers with lower MAFs in statistical analysis. Coupled with new statistical approaches specifically designed to study rare alleles [72], it is anticipated that future association studies could further elucidate the effects of rare variants on phenotypic variation.

Concluding remarks

The current statistical approaches that associate genetic markers to phenotypes are sufficient to identify genomic signals of moderate to large effect and predict phenotypic values accurately enough for GS to make significant genetic gains in plant breeding programs. However, current studies usually lack the statistical power and mapping resolution to detect causative variants controlling a trait and could be prone to false positives. It is therefore important to understand the shortcomings of these approaches and to identify potential sources of limitations, especially with respect to NGS technologies. Continued improvement of these statistical approaches will enable the modeling of biological phenomena to be more accurate and will result in implementations that are more computationally efficient and available in software that produces easily interpretable results. The technology for genotype and phenotype data collection is also progressing extremely fast, and statistical approaches must constantly be developed and adapted to use the latest technologies. To ensure that these goals are realized, it is critical that researchers working in our field continue to work together in an interdisciplinary environment.

Acknowledgements

This research was supported by National Science Foundation awards #0922493 and #1238142, University of Illinois starting funds (A.E.L.), and Cornell University startup funds (M.A.G.). We acknowledge the assistance of Patrick J. Brown in providing insight into NGS technologies and Christine H. Diepenbrock for invaluable feedback on the GS section.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. •• Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang ZW, Costich DE, Buckler ES: **Association mapping: critical considerations shift from genotyping to experimental design**. *Plant Cell* 2009, **21**:2194-2202.

This paper describes practical considerations and challenges for conducting association mapping.

2. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng DZ, Platt A, Tarone AM, Hu TT *et al.*: **Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines**. *Nature* 2010, **465**:627-631.

3. Flint-Garcia SA, Thuillet AC, Yu JM, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES: **Maize association population: a high-resolution platform for quantitative trait locus dissection**. *Plant J* 2005, **44**:1054-1064.

4. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA *et al.*: **Comprehensive genotyping of the USA national maize inbred seed bank**. *Genome Biol* 2013, **14**.

5. • McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li HH, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C *et al.*: **Genetic properties of the maize nested association mapping population**. *Science* 2009, **325**:737-740.

The authors describe genetic attributes of the maize nested association mapping population.

6. • Yu JM, Holland JB, McMullen MD, Buckler ES: **Genetic design and statistical power of nested association mapping in maize**. *Genetics* 2008, **178**:539-551.

The authors conduct computer simulations in the maize nested association mapping panel and assess its ability to detect QTL.

7. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC *et al.*: **The genetic architecture of maize flowering time**. *Science* 2009, **325**:714-718.

8. Zhu CS, Gore MA, Buckler ES, Yu J: **Status and prospects of association mapping in plants**. *Plant Genome* 2008, **1**:5-20.

9. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: **Association mapping in structured populations**. *Am J Hum Genet* 2000, **67**:170-181.

10. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nat Genet* 2006, **38**:904-909.

11. Jombart T, Devillard S, Balloux F: **Discriminant analysis of principal components: a new method for the analysis of genetically structured populations**. *BMC Genet* 2010, **11**.

12. Lipka AE, Tian F, Wang QS, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang ZW: **GAPIT: genome association and prediction integrated tool**. *Bioinformatics* 2012, **28**:2397-2399.

13. Loiselle BA, Sork VL, Nason J, Graham C: **Spatial genetic-structure of a tropical understory shrub, *Psychotria officinalis* (rubiaceae)**. *Am J Bot* 1995, **82**:1420-1425.

14. Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S: **Genetic structure and diversity in *Oryza sativa* L.** *Genetics* 2005, **169**:1631-1638.

15. •• Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB *et al.*: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness**. *Nat Genet* 2006, **38**:203-208.

This paper introduces the unified mixed linear model for genome-wide association studies, which is one of the most widely used and effective statistical models to account for spurious associations.

16. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D: **FaST linear mixed models for genome-wide association studies**. *Nat Methods* 2011, **8**:833-835.

17. Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z: **A SUPER powerful method for genome wide association study**. *PLOS ONE* 2014, **9**:e107684.

18. Kang H, Zaitlen N, Wade C, Kirby A, Heckerman S, Daly M, Eskin E: **Efficient control of population structure in model organism association mapping**. *Genetics* 2008, **178**:1709-1723.

19. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E: **Variance component model to account for sample structure in genome-wide association studies**. *Nat Genet* 2010, **42**:348 U110.

20. Zhang ZW, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu JM, Arnett DK, Ordovas JM *et al.*: **Mixed linear model approach adapted for genome-wide association studies**. *Nat Genet* 2010, **42**:355-360.

21. Li M, Liu X, Bradbury P, Yu J, Zhang YM, Todhunter RJ, Buckler ES, Zhang Z: **Enrichment of statistical power for genome-wide association studies**. *BMC Biol* 2014, **12**:73.

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

23. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies**. *Nat Genet* 2012, **44**:821 U136.

24. •• Platt A, Vilhjalmsson BJ, Nordborg M: **Conditions under which genome-wide association studies will be positively misleading**. *Genetics* 2010, **186**:1045-1052.
Using simulation studies, the authors describe situations in which a genome-wide association study will yield misleading results about the location of putative genes.

25. • Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB: **Rare variants create synthetic genome-wide associations**. *PLoS Biol* 2010, **8**:e1000294.
Synthetic associations are described and assessed using both simulated and real data. The authors conclude that synthetic associations might be underlying many of the signals identified in a GWAS.

26. • Orozco G, Barrett JC, Zeggini E: **Synthetic associations in the context of genome-wide association scan signals**. *Hum Mol Genet* 2010, **19**:R137-R144.
Using both simulation studies and empirical data, the authors conclude that synthetic associations do not underlie most GWAS signals.

27. Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, Long Q, Nordborg M: **An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations**. *Nat Genet* 2012, **44**:825 U144.

28. Pérez P, de los Campos G: **Genome-wide regression & prediction with the BGLR statistical package**. *Genetics* 2014, **198**:482-495.

29. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A *et al.*: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461**:747-753.

30. Gyenesei A, Moody J, Semple CAM, Haley CS, Wei W: **High-throughput analysis of epistasis in genome-wide association studies with BiForce**. *Bioinformatics* 2012, **28**:1957-1964.

31. Goudey B, Rawlinson D, Wang Q, Shi F, Ferra H, Campbell RM, Stern L, Inouye MT, Ong CS, Kowalczyk A: **GWIS – model-free, fast and exhaustive search for epistatic interactions in case-control GWAS**. *BMC Genomics* 2013, **14**.

32. Schupbach T, Xenarios I, Bergmann S, Kapur K: **FastEpistasis: a high performance computing solution for quantitative trait epistasis**. *Bioinformatics* 2010, **26**:1468-1469.

33. Wisser RJ, Kolkman JM, Patzoldt ME, Holland JB, Yu J, Krakowsky M, Nelson RJ, Balint-Kurti PJ: **Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a GST gene**. *Proc Natl Acad Sci USA* 2011, **108**:7339-7344.

34. Zhou X, Stephens M: **Efficient multivariate linear mixed model algorithms for genome-wide association studies**. *Nat Methods* 2014, **11**:407-409.

35. Korte A, Vilhjalmsson BJ, Segura V, Platt A, Long Q, Nordborg M: **A mixed-model approach for genome-wide association studies of correlated traits in structured populations**. *Nat Genet* 2012, **44**:1066-1071.

36. Moore CR, Johnson LS, Kwak IY, Livny M, Broman KW, Spalding EP: **High-throughput computer vision introduces the time axis to a quantitative trait map of a plant growth response**. *Genetics* 2013, **195**:1077-1086.

37. Andrade-Sanchez P, Gore M, Heun J, Thorp K, Carmo-Silva A, French A, Salvucci M, White J: **Development and evaluation of a field-based high-throughput phenotyping platform**. *Funct Plant Biol* 2014, **41**:68-79.

38. •• Doerge RW: **Mapping and analysis of quantitative trait loci in experimental populations**. *Nat Rev Genet* 2002, **3**:43-52.
A definitive review of QTL analysis approaches, with applications to microarrays.

39. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G *et al.*: **Genetics of gene expression surveyed in maize, mouse and man**. *Nature* 2003, **422**:297-302.

40. Majewski J, Pastinen T: **The study of eQTL variations by RNA-seq: from SNPs to phenotypes**. *Trends Genet* 2011, **27**:72-79.

41. Cubillos FA, Coustham V, Loudet O: **Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants**. *Curr Opin Plant Biol* 2012, **15**:192-198.

42. Ernst CW, Steibel JP: **Molecular advances in QTL discovery and application in pig breeding**. *Trends Genet* 2013, **29**:215-224.

43. Nica AC, Dermitzakis ET: **Expression quantitative trait loci: present and future**. *Philos Trans R Soc B* 2013, **368**.

44. Westra HJ, Franke L: **From genome to function by studying eQTLs**. *Biochim Biophys Acta* 2014, **1842**:1896-1902.

45. Verdugo RA, Farber CR, Warden CH, Medrano JF: **Serious limitations of the QTL/microarray approach for QTL gene discovery**. *BMC Biol* 2010, **8**:96.

46. Ramasamy A, Trabzuni D, Gibbs JR, Dillman A, Hernandez DG, Arepalli S, Walker R, Smith C, Ilori GP, Shabalin AA *et al.*: **Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies**. *Nucleic Acids Res* 2013, **41**.

47. Churchill GA, Doerge RW: **Naive application of permutation testing leads to inflated type I error rates**. *Genetics* 2008, **178**:609-610.

48. Wright FA, Shabalin AA, Rusyn I: **Computational tools for discovery and interpretation of expression quantitative trait loci**. *Pharmacogenomics* 2012, **13**:343-352.

49. Battle A, Montgomery SB: **Determining causality and consequence of expression quantitative trait loci**. *Hum Genet* 2014, **133**:727-735.

50. Bush WS, Moore JH: **Genome-wide association studies**. *PLoS Comput Biol* 2012, **8**:e1002822.

51. Mackay TFC: **The genetic architecture of quantitative traits**. *Annu Rev Genet* 2001, **35**:303-339.

52. Lipka AE, Gore MA, Magallanes-Lundback M, Mesberg A, Lin HN, Tiede T, Chen C, Buell CR, Buckler ES, Rocheford T *et al.*: **Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain**. *G3: Genes Genomes Genet* 2013, **3**:1287-1299.

53. Owens BF, Lipka AE, Magallanes-Lundback M, Tiede T, Diepenbrock CH, Kandianis CB, Kim E, Cepela J, Mateos-Hernandez M, Buell CR *et al.*: **A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels**. *Genetics* 2014, **198**:1699-1716.

54. Lantieri F, Jhun MA, Park J, Park T, Devoto M: **Comparative analysis of different approaches for dealing with candidate regions in the context of a genome-wide association study**. *BMC Proc* 2009, **3**:S93.

55. Wang K, Li MY, Hakonarson H: **Analysing biological pathways in genome-wide association studies**. *Nat Rev Genet* 2010, **11**:843-854.

56. • Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps**. *Genetics* 2001, **157**:1819-1829.
This paper describes the use of genome-wide markers to predict phenotypic values. Both frequentist and Bayesian (BayesA and BayesB) approaches are discussed.

57. • Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: **Invited review: genomic selection in dairy cattle: progress and challenges**. *J Dairy Sci* 2009, **92**:433-443.
This is a thorough review of genomic selection and its use in dairy cattle.

58. • Bernardo R, Yu JM: **Prospects for genomewide selection for quantitative traits in maize**. *Crop Sci* 2007, **47**:1082-1090.
The authors compare genomic selection to marker-assisted recurrent selection in maize, and conclude that genomic selection is superior.

59. • Heffner EL, Sorrells ME, Jannink JL: **Genomic selection for crop improvement**. *Crop Sci* 2009, **49**:1-12.
The authors review various genomic selection approaches and consider its applications to crop breeding.

60. • Gianola D: **Priors in whole-genome regression: the Bayesian alphabet returns**. *Genetics* 2013, **194**:573-596.
This paper reviews some of the most popular Bayesian approaches for genomic selection, and discusses the pitfalls of using them to make inferences on genetic architecture.

61. •• Daetwyler H, Calus M, Pong-Wong R, de los Campos G, Hickey J: **Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking**. *Genetics* 2013, **193**:347-365.
The authors describe guidelines for reporting genomic selection results that should more widely adapted.

62. •• de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL: **Whole-genome regression and prediction methods applied to plant and animal breeding**. *Genetics* 2013, **193**:327-345.
This article gives an overview of the genomic selection models available, describes some practical considerations when assessing results, and describes several lessons learned from simulation studies and real data analyses.

63. • Heslot N, Yang HP, Sorrells ME, Jannink JL: **Genomic selection in plant breeding: a comparison of models**. *Crop Sci* 2012, **52**:146-160.
This paper assesses the predictive ability of various genomic selection models, and concludes that similar prediction accuracies are obtained for many of the models.

64. Piepho HP: **Ridge regression and extensions for genomewide selection in maize**. *Crop Sci* 2009, **49**:1165-1176.

65. Endelman JB: **Ridge regression and other kernels for genomic selection with R Package rrBLUP**. *Plant Genome* 2011, **4**:250-255.

66. • Lorenz AJ, Chao SM, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink JL: **Genomic selection in plant breeding: knowledge and prospects**. *Adv Agron* 2011, **110**:77-123.
A book chapter describing statistical approaches and quantitative genetics considerations for genomic selection.

67. • Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species**. *PLoS ONE* 2011, **6**:e19379.
The authors describe an approach for constructing genotyping-by-sequencing libraries using restriction enzymes.

68. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA: **Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags**. *PLoS Genet* 2010, **6**:e1000862.

69. Akbari M, Wenzl P, Caig V, Carling J, Xia L, Yang SY, Uszynski G, Mohler V, Lehmensiek A, Kuchel H *et al.*: **Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome**. *Theor Appl Genet* 2006, **113**:1409-1420.

70. Browning BL, Browning SR: **Improving the accuracy and efficiency of identity-by-descent detection in population data**. *Genetics* 2013, **194**:459-471.

71. Swarts K, Li HH, Romero Navarro JA, An D, Romay MC, Hearne S, Acharya C, Glaubitz JC, Mitchell SE, Elshire RJ *et al.*: **Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants**. *Plant Genome* 2014.

72. Zhu CS, Li XR, Yu JM: **Integrating rare-variant testing, function prediction, and gene network in composite resequencing-based genome-wide association studies (CR-GWAS)**. *G3: Genes Genomes Genet* 2011, **1**:233-243.