A Guide to Genome-Wide Association Mapping in Plants

Liana T. Burghardt,[1] Nevin D. Young,[2] and Peter Tiffin[1]

[1]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, Minnesota
[2]Department of Plant Pathology, University of Minnesota, St. Paul, Minnesota

Genome-wide association studies (GWAS) have developed into a valuable approach for identifying the genetic basis of phenotypic variation. In this article, we provide an overview of the design, analysis, and interpretation of GWAS. First, we present results from simulations that explore key elements of experimental design as well as considerations for collecting the relevant genomic and phenotypic data. Next, we outline current statistical methods and tools used for GWA analyses and discuss the inclusion of covariates to account for population structure and the interpretation of results. Given that many false positive associations will occur in any GWA analysis, we highlight strategies for prioritizing GWA candidates for further statistical and empirical validation. While focused on plants, the material we cover is also applicable to other systems. © 2017 by John Wiley & Sons, Inc.

Keywords: GWAS • genomics • QTL mapping • genotype • phenotype • association mapping

How to cite this article:
Burghardt, L.T., Young, N.D., and Tiffin, P. 2017. A guide to genome-wide association mapping in plants. *Curr. Protoc. Plant Biol.* 2:22-38.
doi: 10.1002/cppb.20041

INTRODUCTION

Identifying gene function has been a long-standing goal in biology. For plant scientists, identifying gene function not only can advance our understanding of basic biology, but also can provide the resources for crop improvement. Until recently, forward genetics—mutagenesis followed by phenotypic screens to identify individuals that have a phenotype distinct from the wild-type—have been the primary approach for identifying the functional importance of genes. Forward genetics can be effective at identifying genes of major effect but is unlikely to identify genes with subtle effects. Forward genetics also fails to provide direct information on whether allelic diversity at these loci is responsible for naturally occurring variation. In recent years, genome-wide association studies (GWAS) and biparental quantitative trait locus (QTL) mapping (Mackay et al., 2009) have developed into valuable approaches for identifying gene function. These approaches complement forward genetics and have the potential to identify genes of more subtle effects, as well as genes that comprise segregating alleles responsible for phenotypic variation (Chan et al., 2011; Ogura and Busch, 2015).

GWAS are distinct from biparental QTL mapping in two important ways. First, traditional QTL mapping relies on mapping populations derived from only two parents; therefore, only a limited amount of the genetic variation can be assayed. Second, relatively few recombination events occur during the generation of a biparental mapping population. This means it can be hard to locate QTL with high resolution, and thus labor-intensive fine mapping is required after a QTL is identified. By contrast, GWAS include population-scale samples and thereby assay a wide swath of natural variation. GWAS also leverage the history of recombination events across a lineage thus allowing finer resolution of QTL location.

In this article we provide an overview of the design and interpretation of GWAS. After a brief overview of the approach, we present

Current Protocols in Plant Biology 2:22-38, March 2017
Published online March 2017 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cppb.20041

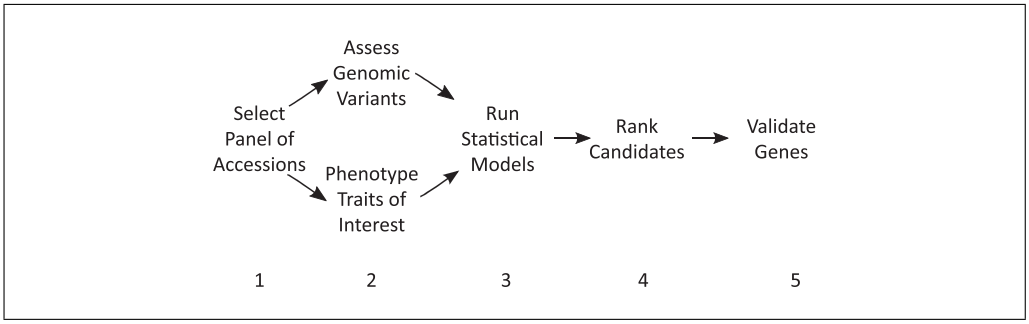


Figure 1 Basic approach for conducting a GWAS. (1) Select a panel of diverse genotypes. This can represent a range-wide collection or a collection of germplasm relative to a specific breeding program. For a handful of species already genotyped, HapMap panels are already available. (2) Collect genotypic and phenotypic data on every accession. The more precise the phenotype data and the more high density the genotype data the better. (3) Perform statistical analyses linking genomic variants to phenotypic variation. Analyses, usually linear models or χ^2-like tests, can be implemented using readily available software. (4) Prioritize candidates for follow-up work; strength of statistical association, complementary mapping data, biological knowledge, genomic annotation, and expression data can all be valuable. (5) Validate candidates using an independent GWAS panel or reverse genetics.

results from simulations that explore experimental design and outline considerations for the collection of genomic and phenotypic data. We then discuss statistical analyses, the interpretation of results, and how to prioritize and validate GWAS candidates. Lastly, we highlight both limitations and the future of GWAS. We focus our discussion on the study of plants, but because the important issues to consider when conducting GWAS are not system-specific, most of what we cover is equally applicable to animals or microbes. We do not cover several other important topics including the strengths and weaknesses of GWAS, findings from previous GWAS, or in-depth treatments of statistical analyses; these topics have been covered by recent reviews and perspectives (Korte and Farlow, 2013; Huang and Han, 2014; Ogura and Busch, 2015; Goddard et al., 2016).

OUTLINE OF THE BASIC APPROACH

Conducting a GWAS is conceptually straightforward (Fig. 1; see Table 1 for a glossary of commonly used terms in GWAS): (1) collect both phenotypic data on trait(s) of interest and genotypic data on a large collection of accessions; (2) evaluate the strength of the association between phenotypic variation and each of the genomic variants using χ^2-like or (generalized) linear model analyses; and (3) interpret the results of the analyses to identify a set of genes with segregating alleles that are candidates responsible for phenotypic variation. Because of statistical limitations, genetic variants identified through GWA analyses are considered candidates, and robust support for a candidate gene having causative phenotypic effects requires validation through independent data or experimentation. While, conceptually straightforward, the details of the statistical analyses and the proper interpretation of the results can be complicated.

The power of GWAS to identify genes of interest will be a function of five important factors: the genetic complexity of a trait, the heritability of that trait, the number of accessions assayed, the relatedness of those accessions, and the density of the genomic variants. GWAS will have the greatest power to identify a causative gene (i.e., one that has segregating alleles that contribute to phenotypic variation) when there is no confounding between relatedness and phenotypic variation, few loci of large effect are responsible for phenotypic variation, and most of the phenotypic variation is the result of genetic rather than environmental variation (i.e., high heritability). On the other end of the spectrum, the most challenging situation for finding genes that contribute to phenotypic variation is when the sample is highly structured genetically, a trait has low heritability (i.e., phenotypic variation is primarily due to environmental variation), and the genetic basis of the trait is complex (i.e., many genes contribute to variation with each gene contributing only a small amount to the genetic variance of that trait).

FACTORS AFFECTING GWAS PERFORMANCE

The potential for a GWAS to identify genes of interest is intricately related to the size of the

Table 1 Glossary of Commonly Used Terms

Term	Definition
Alleles	One of two (or more) alternate forms of a gene or genetic locus
Accessions	Collections of germplasm or tissue
Effect size	A measure of the size of the difference between groups regardless of the sample size or variance
Epistasis	In the most general sense, this describes any interaction among genes; but, for quantitative geneticists it specifically refers to when two loci do not have an additive effect on phenotype
False positives (Type 1 error)	The incorrect rejection of a true null hypothesis
Genetic complexity	The number and effect sizes of the genetic variants underlying phenotypic variation in a trait
Heritability	The proportion of phenotypic variation that is attributable to genotype in a given environment and sample
Linkage disequilibrium (LD)	The non-random association of alleles at different loci
Mendelian trait	Any phenotype controlled by a locus (or loci) whose effect can be described by the principles of inheritance outlined by Mendel
Minor allele frequency (MAF)	The frequency at which the second most common allele occurs in a given sample
Quantitative trait	A measurable phenotype that depends on the cumulative actions of many genes
Quantitative trait loci (QTL)	A section of DNA that correlates with variation in a phenotype
Recombination	The exchange of genetic material either between multiple regions of a chromosome or across chromosomes
P-value	The probability of obtaining a result equal to or "more extreme" than the observed value when the null hypothesis is true

panel of accessions, the proportion of variance explained by a locus, and the heritability of the trait. We conducted simulations to illustrate how each of these factors affects the statistical power to detect a single, known causative variant. In these simulations we assume each genomic variant (i.e., single nucleotide polymorphism [SNP]) is independent of other variants, has additive gene action, is the causative variant assayed, and is found in individuals that are all equally distantly related to one another. This last assumption allows us to omit the statistical covariates that account for population structure that are often included in analysis of GWA (see Statistical Analysis, Interpretation of *P*-Values, and Model Covariates). The absence of population structure also allows us to use a simple t-test for evaluating the statistical strength of a genotype-phenotype association. The results of the simulations (Fig. 2) highlight several important considerations related to the design and interpretation of GWAS:

1. When trait heritability is high, there is greater power to identify a locus underlying variation in that trait than when heritability is low.
2. There will be less power to identify a locus responsible for a small amount of genetic variation than a large amount of variation (Fig. 2, columns). Thus, genetically complex traits—those where many loci explain small amounts of the phenotypic variation—will show much weaker statistical signals at the contributing loci than genetically simple traits.
3. The more genotypes included in your panel, the more power to detect a given locus (Fig. 2, rows).
4. While the *P*-values in this simulation may look heartening, at least when genes are of large effect and traits have high heritability, they are based on a single statistical test for a known causative SNP. In reality, we will not know which SNP

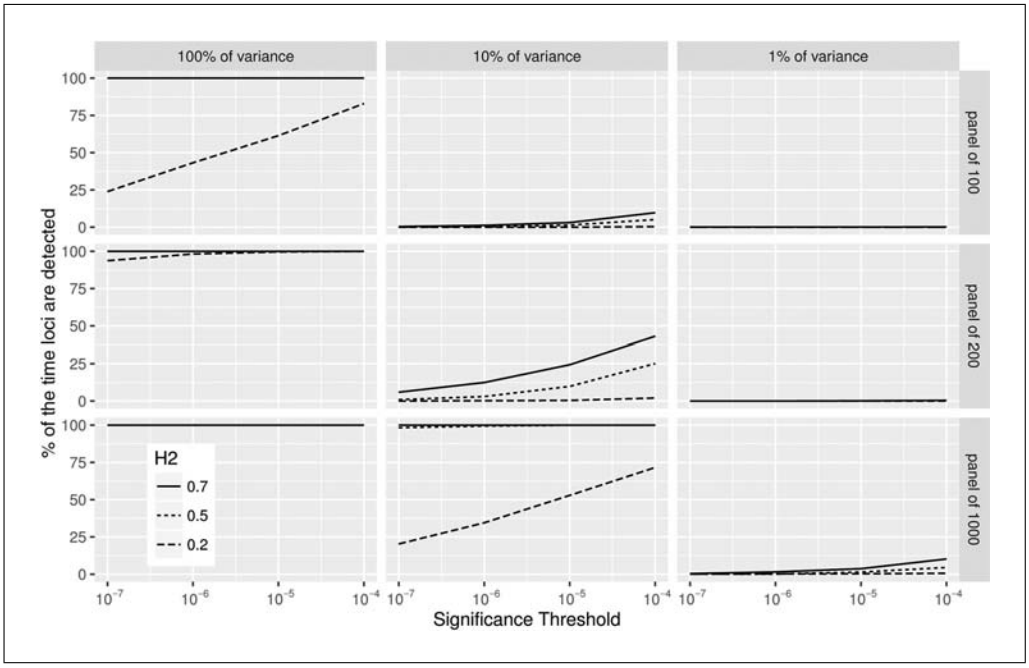


Figure 2 Percentage of time (vertical axis) a single locus underlying phenotypic variation would have a statistical signal that is greater than specific significance thresholds (horizontal axis). The statistical strength of an association depends on trait heritability (line type), the percent of genetic variation explained by the locus (columns), and the number of genotypes used (rows). The results are from 1000 simulations with the following assumptions: the locus has two alleles present at equal frequencies, no population structure, normally distributed phenotypes, and no epistasis. We used a t-test to determine *P*-values without adjustment for genome-wide testing.

is the causative one, and an enormous number of statistical tests will be performed on loci with no effect. Thus, GWAS have a major issue with putative associations that end up being false positives. For instance, if there are 1 million SNPs and thus 1 million statistical tests are run, by chance alone we expect 10,000 SNPs to have *P*-values $<10^{-2}$ (0.01) and 10 SNPS to have *P*-values $<10^{-5}$ (0.00001). These false positives can make identifying the causative SNP illustrated here like finding a needle in a haystack.

THE IMPORTANCE OF PHENOTYPE: PICKING TRAITS AND MINIMIZING ALTERNATE SOURCES OF VARIANCE

It is important to think about phenotyping for two reasons. One reason is simply that trait heritability can depend on the quality of the phenotypic data; the more precisely phenotyping is done, the higher the heritability and the more power to detect causal variants. It is also important to remember that heritability of a trait is specific to a population and environment in which it is measured (Falconer and Mckay, 1996), and therefore identifying causative loci may depend on phenotyping environment.

The second reason for thinking carefully about phenotype is because the probability of GWAS succeeding is directly related to the number and effect sizes of genes responsible for variation: The fewer the genes and the larger the effects, the more power for identifying the genes. Consistent with the expectation that it is easier to identify loci underlying variation in traits with a simple genetic basis, some of the strongest statistical signals are for metabolic traits for which there is a single enzyme or transporter that is responsible for distinct phenotypes (Lipka et al., 2013; Chen et al., 2014; Matsuda et al., 2015; Strauch et al., 2015; Shakoor et al., 2016). By contrast, we should not expect GWAS to provide extremely strong signals for genes that contribute to variation in highly complex multigenic quantitative traits such as height. In fact, height in both maize and humans is estimated to be affected by many genes of small (<1%) effect (Yang et al., 2010; Peiffer et al., 2014). This does not mean that GWAS should be limited to traits with variation determined by genes of major effect. Rather, researchers investigating the genetics of "quantitative" traits should not expect the same strong statistical support

that can be found when mapping the genetic basis of phenotypes that are determined by a single gene. Therefore, for quantitative traits like these, it might be particularly important to integrate multiple types of data to prioritize putatively causative genes for subsequent statistical or functional analyses (see Prioritizing GWAS Candidates).

There is, however, an alternative approach for researchers interested in identifying the genetic basis of variation in truly complex traits. Many ecologically or economically important quantitative traits, such as height, fitness, or yield, are cumulative traits that integrate the effects of many upstream traits. Because of their cumulative nature, a large portion of the genome can be implicated in the downstream phenotype. In these cases it may be more fruitful to focus on the genetic basis of the individual upstream traits because these upstream traits may be simpler than the downstream traits to which they contribute. Imagine variation in a composite trait is due to 100 genes (probably fewer than reality), each of equal effect. For success, GWAS then would need to have the statistical power to identify genes with a 1% effect size. If, however, one knows of a highly heritable trait that itself explains 10% of the variation in yield, one could try mapping the upstream trait. If that trait is affected by 10 genes that each explain 10% of the variation, then GWAS will have much greater power to detect them, and ultimately one will have identified multiple loci that contribute to the composite trait of interest (Fig. 2, column 2 versus column 3).

CONSIDERATIONS FOR GENERATING GENOMIC DATA AND CALLING VARIANTS

To associate phenotypic variation with genomic variation, one needs to genotype the entire panel of accession that will be used for the association analyses. The key question here is: What level of variant coverage is needed to ensure a high probability of finding a causative variant? In short, the more variants the better. Thus, full genome sequencing is by far the best option when feasible. Below we outline the statistical reasons for this, provide empirical examples of why high variant density is important, and discuss some of the tools available for imputing missing data when coverage is low or uneven across the genome.

The simulations presented in Figure 2 were run with the assumption that the causative variant was included in the analyses; in a perfect world with complete genome coverage, all causative variants would have been assayed. However, if a causative variant is not assayed, all is not lost. The potential for identifying a "missed" causative variant depends on whether it is in strong linkage disequilibrium (LD; i.e., it is "tagged" by one or more assayed variants). LD is a measure of the non-random association of variants at different loci. Importantly, LD is not the same as physical linkage; many allelic variants that are in close physical proximity have low LD either because of recombination or simply because the variants are not at equal frequencies. Although there are several measures of LD, the one most relevant to GWAS is r^2—the squared correlation coefficient between allelic frequency at each of two loci (Gaut and Long, 2003; Slatkin, 2008). Unfortunately, most variants in the genome are not in high LD with any other variants, even when those variants are physically closely linked (e.g., Remington et al., 2001; Branca et al., 2011). Thus the probability of capturing a causative allele or a variant in LD with the causative allele will be greatest with more complete genomic data.

Stanton-Geddes et al. (2013) provide an empirical example of the expected performance of GWAS when reduced-representation genomic data are used. These authors compared GWAS results when using variants from whole-genome sequencing to in-silico reduced-representation data sets in which 1 SNP was assayed from every 1 kb across the genome. Notably, analyses of the reduced-representation data identified only a subset of the candidates identified with the full sequence data. Only 9 to 17 of top 20 associations within the reduced-representation data were within 20 kb of the top 200 candidates identified using full-sequence data, and only 16 to 30 of the top 50 candidates overlapped with the 200 full-sequence based candidates. Results from actual reduced-representation data, in which SNPs will not be evenly distributed across the genome as in standard applications of GBS/RADseq (genotyping by sequencing/restriction site–associated DNA sequencing), would be expected to provide even poorer sensitivity (Lowry et al., 2016).

This comparison underscores the point that while representation data might be useful for identifying some genes contributing to phenotypic variation, they will provide a far less complete picture than full-genome sequence data. In practice, obtaining full-genome sequence data (for many individuals) is still beyond the budget for many projects—particularly for organisms with large genome

sizes. Therefore, if conducting GWAS is important and full genome sequence data are not available, researchers should consider focusing on transcriptome data (Lowry et al., 2016) or on candidate genes that can be assayed using gene capture (Zhou and Holliday, 2012; Suren et al., 2016).

Even if full genomes are sequenced, unequal coverage will typically result in many polymorphic sites that are not covered in all accessions. This missing data will reduce the statistical power to detect associations. There are multiple approaches available for imputing the variant state at sites for which there are missing data (Halperin and Stephan, 2009; Marchini and Howie, 2010; van Leeuwen et al., 2015). These approaches vary in terms of computational demands and the types of data needed to infer missing data. Some approaches, such as those implemented in Beagle (Browning and Browning, 2007, 2016), rely on data from pedigrees. Others rely on inferred linkage disequilibrium from population data such as PHASE (Stephens and Scheet, 2005) and fastPHASE (Scheet and Stephens, 2006); some, on a high-quality training set of known linkage such as IMPUTE (Howie et al., 2009); and others, on the very simple (and fast) approach of assuming a heterozygous state and filling in the missing data with average phenotypic value, such as GAPIT (Lipka et al., 2012), which may or may not be appropriate given your system.

STATISTICAL ANALYSIS, INTERPRETATION OF P-VALUES, AND MODEL COVARIATES

The statistical analyses employed in association analyses boil down to a linear model (for continuously variable traits) or χ^2-like tests (for categorical; i.e., case-control). If one has a million SNPs, one can think of the GWA as conducting one million regressions (or χ^2 tests) in which the phenotype is the response variable, the genomic variant is an explanatory variable, and one or more covariates might be included to reduce the effects of unequal relatedness among accessions. The probability of obtaining the measured association between the phenotype and a genomic variant by chance—the *P*-value—can then be used to rank the statistical support for a variant being an important contributor to phenotypic variation. Given the very large number of tests performed, there will always be many low *P*-values that are false positives; they will be putative associations due to chance alone. As outlined above, this is because for each causal variant there will be a large number of variants tested that have no effect on the phenotype.

The *P*-values of phenotype-variant associations are usually presented using Manhattan plots (Fig. 3) in which negative $\log_{10}$ *P*-values (vertical axis) are plotted against genomic position (horizontal axis). Genomic regions with high peaks of SNPs that stand out against the background level of *P*-values are the genomic positions with the strongest signals of contribution to phenotypic variation. *P*-values are generally treated as the probability associated with rejecting a null hypothesis by chance. Dating back to the time of Fisher (1925), biologists often use a nominal *P*-value of 0.05 as a critical value for rejecting a null hypothesis. In GWAS, we are not really testing a hypothesis (i.e., "I hypothesize that the alleles segregating at this locus are responsible for phenotypic variation in a trait of interest"). After all, if we know that a trait has a non-zero heritability, then we know that there are genes that underlie variation in that trait, yet it is quite common for GWAS to be conducted for which none of the variants tested have a $P < 0.05$ after correcting for the millions of tests conducted. Because of the complications from a strict interpretation of *P*-values when applied to GWAS, we think there is merit in considering *P*-values as a metric to select a pool of variants for further study rather an as a criterion that must meet a predetermined critical threshold to be labeled significant.

There are many software packages available for conducting GWAS (Table 2). In general, any of these packages can be used effectively on any dataset although there are some differences with regard to input formats, whether they are designed to handle continuous (e.g., TASSEL, EMMAX) or binomial response variables GMMAT (Chen et al., 2016), whether they incorporate epistasis such as lrgpr (Hoffman et al., 2014) or prior information on different categories of SNPs BayesR (Moser et al., 2015; Goddard et al., 2016), and how they handle missing data.

Regardless of what software is used, ensuring that the statistical model is an accurate description of the data is a key—and somewhat complicated component—to identifying promising candidates. One can get an idea of model fit using quantile-quantile (Q-Q) plots of the observed values $-\log_{10}$ *P*-values versus the expected values $-\log_{10}$ *P*-values (Yang et al., 2011; Chen et al., 2016). If the empirical *P*-values deviate drastically from null expectations, then the *P*-values generated from the analyses may be greatly inflated or deflated

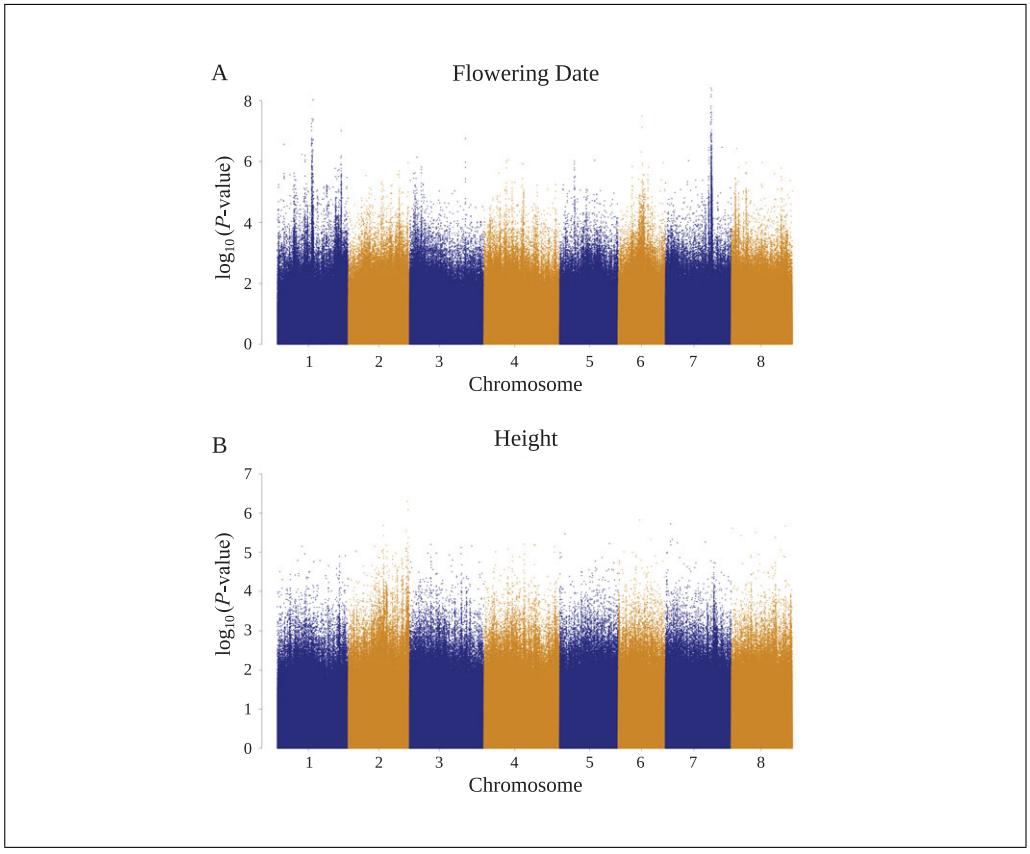


Figure 3 Manhattan plots are standardly used for visualizing the results of GWAS. The horizontal axis is chromosomal location, and the vertical axis is typically the negative $\log_{10}$ of the *P*-value for each locus (so that higher values have greater statistical support). Each dot represents a single nucleotide polymorphism (SNP; or other genomic variant). Some traits, such as flowering time shown in (**A**) have large, wide peaks whereas other traits such as height (**B**) have many small peaks. The lack of strong associations of height (B) with the genetic markers could be due to a lack of genetic variation for the trait or indicate a complex genetic basis—as is likely the case with height.

(Fig. 4). One reason the empirical distribution may deviate from expectations is because phenotypic data do not match the assumptions of the statistical model used for the analyses (i.e., a normal distribution for the linear models used in ANOVAs). Therefore, just as is standard practice in conducting analyses of variance, it may be beneficial to transform the phenotypic values so that the distribution of residual values is more or less normally distributed (Goh and Yap, 2009). An alternative solution is to use a generalized linear model (GLM) that allows for response variables (phenotypes) that are not normally distributed. See Chen et al. (2016) for an example of a GLM developed for use with a binomial response variable.

A second, more complicated and more often discussed reason that the distribution of observed *P*-values may deviate from expectations is population structure among the accessions (i.e., not all accessions are equally related to one another; Vilhjálmsson and Nordborg, 2013). Unfortunately, population structure is present in almost all samples. Population structure can be problematic for two reasons. One reason is that if phenotype data are not collected under controlled conditions, then more closely related individuals are more likely to have a shared environment and thus a more similar phenotype than more distantly related individuals. In other words, phenotype will be correlated with relatedness. For most plant studies, unlike human studies, phenotype data have been collected from plants grown in a common environment, so we don't have to worry about the this problem. The second reason population structure can inflate the number of false positives is that if a phenotype covaries with population structure, then many non-causative SNPs will covary with the phenotype simply due to relatedness. Given the near ubiquity of isolation by distance (Meirmans, 2012), population structure is a potentially important problem for most plant GWAS.

Table 2 Commonly Used Packages for Conducting GWAS (Circa 2016)[a]

Package	Description	Web site
TASSEL	Variety of algorithms MLM, GLM, weighted MLM, genomic selection, fast association; supports P3D compression; can process GBS data; designed to determine dominance/additivity of effects; user-friendly GUI	http://www.maizegenetics.net/tassel
GAPIT	Package that can perform MLM and EMMA; supports P3D and EMMAx; works via R language	http://www.maizegenetics.net/gapit
EMMAX	Efficient mixed-model method for large genomic datasets; command-line interface only	http://genetics.cs.ucla.edu/emmax/index.html
GEMMA	Standard/multivariate/Bayesian linear mixed-model framework; estimates quantitative genetic traits and proportioning of variance; command-line interface only	http://www.xzlab.org/software.html
ANGSD	Useful when genotypic states are not known with certainty; measures population genetic parameters; command-line interface only	http://www.popgen.dk/angsd/index.php/ANGSD
Plink	Wide-ranging toolset for conducting GWAS; originally designed for human genome data; command-line interface only	http://pngu.mgh.harvard.edu/~purcell/plink/
Lrgpr	Allows for testing of G×G and G×E; works via R language	http://lrgpr.r-forge.r-project.org/

[a]G×G, genotype × genotype interaction; G×E, genotype × environment interaction; GBS, genotyping by sequence; GLM, general linear model; GUI, graphical user interface; GWAS, genome-wide association study; MLM, mixed linear model.

To reduce the effects of population structure, GWA analyses usually include covariates to remove the association between population structure and phenotypic variation statistically. The measures of relatedness used as covariates can be kinship (K) matrix, subpopulation membership (i.e., Q matrix) inferred using STRUCTURE (Pritchard et al., 2000; Raj et al., 2014) or similar program (i.e., INSTRUCT), principal components (Price et al., 2006), or a genetic relatedness matrix (GRM) as is used in the human literature (Speed and Balding, 2015). All commonly used GWAS software easily incorporates these covariates.

Although the approaches for accounting for population structure are well developed, it is not always clear how many, or even what covariates/measures of population structure, should be included in an analysis. Yu et al. (2006) recommended that a K matrix be used when there is familial structure only, a Q matrix when sampling is strongly structured across populations, and both Q and K matrices when there are multiple levels of relatedness. These recommendations were not, however, based on simulations or robust interrogation of diverse datasets. In practice, the best approach for minimizing the effects of population structure will depend on demographic history, mating system, and the specific structure of the sample used (Zhang et al., 2008).

A second challenge is that statistical covariates that reduce the effects of population structure can also remove statistical support for causative loci contributing to phenotypic variation. In the first published genome-scale association analyses of *A. thaliana*, Atwell et al. (2010) showed that including a Q matrix to control for population structure resulted in the loss of statistical support for known genes that contribute to flowering time variation. Similarly the first major GWAS in rice found that accounting for population structure reduced their power to identify known associations (Zhao et al., 2011). Perhaps even worse, using covariates when it is unnecessary can actually inadvertently increase the strength of false associations (Yang et al., 2011).

Given the complications of population structure it seems worth considering the value of reporting results from analyses that correct and also do not correct for population structure. For traits that do not covary with population structure, removing the signal

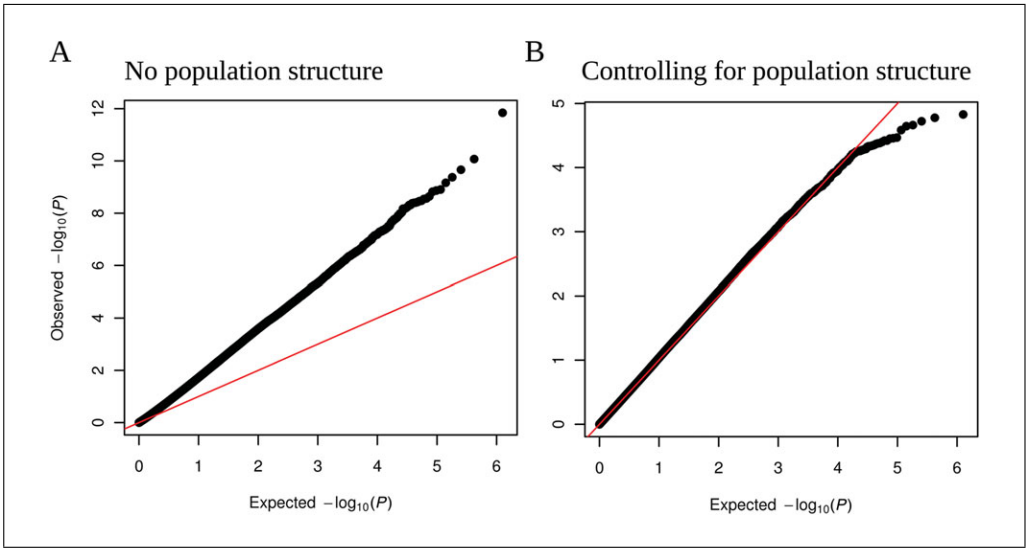


Figure 4 Q-Q plots show the expected distribution of *P*-values and the observed *P*-values. Strong signals of population structure can create deviations from this expectation (**A**). If the dots fall below the red line, it indicates that fewer associations are being found than expected; while if the dots fall above the line, it indicates more associations are being found than expected. Including covariates to control for population structure can improve the match between observed and expected distributions (**B**). While not depicted here, the wrong model type can also cause deviations from expectations. Example is from a GWAS analysis performed on height in Medicago truncatula and is modified from Figure S3 in Stanton-Geddes et al. (2013).

of population structure will remove many false-positives from analyses. Moreover, correcting for population structure is important if *P*-values are going to be interpreted as true probabilities of obtaining a specific association by chance alone. However, for phenotypes that covary with population structure, the SNPs that make the greatest contribution to the phenotype of interest might still be expected to be enriched among the most extremes *P*-values.

PRIORITIZING GWAS CANDIDATES

If GWAS are being conducted to identify functionally important genes, then moving beyond the statistical associations generated from GWAS will require functional validation. Choosing which GWAS-identified candidates to pursue for functional validation requires prioritization. What are the factors an investigator can use to sort through the initial list and thereby focus their attention on those most promising? Below we outline some considerations including screening based on the strength of signal from surrounding markers, minor allele frequency (MAF), evolutionary signatures, and a priori knowledge of biology.

Perhaps the most important information to use when identifying candidates for evaluation is a priori knowledge of biology. Simply examining gene annotations or conducting a gene ontology (GO) search is possibly the most basic line of evidence to bring to prioritizing, though still with powerful predictive value. Increasingly, however, the independent evidence that provides the best value during prioritization lies in expression data (Ritchie et al., 2015). At the simplest level, candidates that are expressed in relevant tissue(s) should be prioritized above others that lack expression support (Houston et al., 2015). The expanding array of gene expression "atlases" (e.g., Petryszak et al., 2016) readily provide this kind of data, though targeted qPCR could easily be performed if needed. Expression data can even be extended beyond the mere question of appropriate location or stage. The creation of expression networks enables the discovery of interconnected webs of genes that have a priori evidence for playing a role in the phenotype of interest. Such expression networks can then be integrated together with the list of GWA candidate loci, highlighting candidates with the best chances of biological relevance (Chan et al., 2011; Bunyavanich et al., 2014; Jia and Zhao, 2014; Schaefer et al., 2016). At the same time, other types of networks, such as protein-protein interaction networks, have the potential to provide similar insights during the GWA prioritization process.

QTL mapping from independent studies also can provide strong guidance about what

associations are important and worth pursuing. QTL mapping based on biparental crosses is probably the most road-tested. In one of the original studies examining GWA mapping of disease resistance genes in Arabidopsis, the value of independent QTL mapping to confirm putative SNP associations was tested critically (Nemri et al., 2010). Indeed, simply using structured GWA mapping panels consisting of nested association mapping (NAM; Kump et al., 2011) or MAGIC (Kover et al., 2009) populations provides independent QTL mapping evidence. Based on results like these, candidate GWA loci that are confirmed by independent QTL mapping rise to the top during the prioritization step. However, pursuing a full-scale QTL mapping experiment, even when a mapping population derived from contrasting parents already exists, is a formidable and time-consuming commitment. There is also no guarantee the allele(s) uncovered by GWA will be segregated in the QTL mapping population.

Additionally, candidate regions often differ wildly in the number of SNPs that indicate strong associations. Sometimes isolated SNP markers show strong evidence of being associated with the phenotype of interest, whereas other times there are clusters, or peaks, of strongly associated tightly-linked SNPs (see Fig. 4A). While the observation of an isolated candidate SNP does not necessarily indicate a lack of phenotypic importance, particularly if marker coverage is low in a given area, there are some evolutionary reasons that clusters and peaks might be better indicators of promising regions of the genome. One possible reason for a clear peak is that there are multiple genes in that region that influence the phenotype. On the other hand, peaks may also reflect strong patterns of linkage disequilibrium that could be indicative of a recent selective sweep, partial sweep, or history of balancing selection (Barton and Turelli, 1989; Yeaman and Whitlock, 2011).

Finally, minor allele frequencies can provide some guidance to which associations may be spurious. The rarer the allele the more susceptible the locus is to being an outlier due to phenotypic sampling variance.

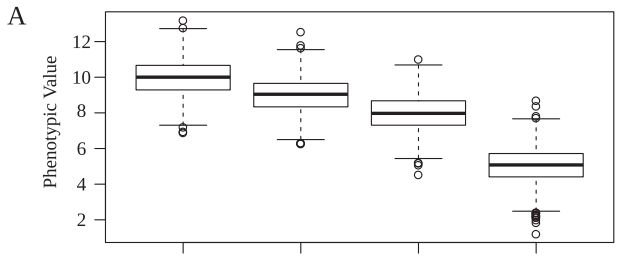
VALIDATION OF CANDIDATES

Once a manageable number of candidates have been identified, the next step in a complete GWA discovery chain will be validation (i.e., verifying the candidate locus is causally connected to the phenotype). The statistical associations identified through GWA will be plagued by false positives and false negatives (Bergelson and Roux, 2010). Therefore, strong claims that a candidate identified through GWA actually contributes to phenotypic variation requires validation. Both statistical and experimental approaches for validation are available.

To validate a GWA candidate statistically, one could repeat the analyses using an independent sample from the same species (NCI-NHGRI Working Group et al., 2007) or a related lineage (i.e., phylo-GWAS; Pease et al., 2016). Obtaining a strong false positive in two or more independent samples is highly unlikely. This expectation is only true, however, if the second GWA is conducted using a panel that does not suffer from the same sources of bias as the first panel (i.e., does not have the same population structure). Otherwise, the false positives may be repeated due to the same underlying bias. When using a statistical approach for validation, researchers should realize that the probability/effect size obtained from a first analysis are likely to be inflated, due to the Beavis effect or winner's curse (explored in Zöllner and Pritchard, 2007).

If an independent population sample is available, identifying the same gene in a second GWAS certainly provides strong evidence for causation. However, the failure to identify the variant in a second panel does not necessarily mean that the gene identified in the first analysis is not functionally important (Liu et al., 2008; Greene et al., 2009). Because variation in quantitative traits can be achieved through many genetic combinations, the combinations of genes responsible for impactful phenotypic variation in one population need not be the same set of genes responsible in a second population (Symonds et al., 2005).

The gold standard of validation is empirical confirmation. If a researcher is lucky, a GWA candidate might have been implicated as being functionally important through previous forward genetic screens. If a knockout of a gene produces a phenotype, it seems reasonable that allelic variation at that locus should also affect phenotypic variation. However, because of the complex genetic basis and low heritability of many traits explored with GWA, previous mutant screens may not have had the power to detect them. The reason for this is that there is typically no replication in forward genetic mutant screens. Each individual is compared to wild-type independently, and because there is always some amount of phenotypic variation among wild-type plants due to the



B

SD from Wild-type	Sample Size	% of Exp. Diff at p<.05	% of Exp. Diff at p<.01
1	5	26	1
1	10	55	26
1	20	88	69
2	5	76	41
2	10	99	91
2	20	100	100
5	5	100	99
5	10	100	100
5	20	100	100

Figure 5 Validating the phenotypic effect of loci of small effect requires larger sample sizes than loci of large effect. (**A**) Phenotypic distributions for 1000 wild-type individuals (far left) and for 100 individuals which have mutations that cause a one, two, or five standard deviation (SD) reduction in phenotypic value. (**B**) Indicates the likelihood of detecting with a t-test the true phenotypic difference between wild-type and each of the mutants illustrated in panel A given three different experiment sizes. Validating a loci that shifts the mean by one standard deviation requires sample sizes >20 to ensure a phenotypic difference that can be detected with $P < 0.05$. In contrast, it is easy to validate loci of large effect even with a small sample size. Phenotypic data of wild-type and each mutant were drawn 1000 times at each sample size, and t-tests were used to test for phenotypic differences between mutant and wild-types.

environment, individuals with mutations that have more subtle effects may not fall outside of the wild-type distribution. An additional limitation of forward screens is that only a subset of allelic effects can be observed on a single genetic background.

Fortunately, candidate gene function can now also be validated through a host of reverse genetic approaches, including transposable element insertion libraries, RNAi, and various genome editing methods including CRISPR/Cas9. Most significant are strategies based on CRISPR technology (Lowder et al., 2015). CRISPRs have the potential to target a specific single base pair or segments of DNA many kb in length. CRISPRs have been successfully used to modify phenotype in *Agrobacterium rhizogenes* roots as well as in regenerants from stable transformation (Ron et al., 2014). Typically, there are few or no off-site modifications (Bortesi et al., 2016; Wolt et al., 2016), something that is significant when testing candidate loci. Moreover, the underlying CRISPR sequence can be crossed out in just one generation, so mutant individuals can be tested more rigorously in comparison with isogenic controls as part of a GWA candidate validation experiment.

Because the genes identified through GWAS are likely to have subtle effects relative to those identified through forward genetic screens, screening of mutants generated by reverse approaches must be carefully designed and controlled. Because of the small effect sizes, large numbers of replicates will generally be required to obtain statistically convincing evidence of an effect. The exact size of the experiments used for validation is difficult to dictate, although the smaller the effect, the larger the experiment needs to be (Fig. 5). A gene that causes a 1 standard deviation decrease in phenotype requires more than 20 replicates to consistently determine that it is different from wild-type at $P < 0.05$. Lastly, it is important to remember that allelic effects can depend on the genetic background. For instance, in Arabidopsis (Vaistij et al., 2013) a loss of function mutation of the transcription

factor SPATULA in one genetic background leads to increased seed dormancy while loss of function in a different accession leads to decreased seed dormancy.

Our own work evaluating GWAS candidates underlying variation in nodule production in *Medicago truncatula* exemplifies the need for validation and replication. We tested ten candidate genes representing six genomic regions that were narrowed down on the basis of the strength of statistical association, proximity to annotated gene models, and expression patterns. Using forward genetics we validated three candidate genes from three different regions—thus validating a candidate from half of the regions examined (Curtin et al., 2017). The phenotypic data from these validation studies show clear overlap of mutant and wild-type distributions (as in Fig. 5), but statistically significant support for a difference between the classes was detected because an average of 22 replicates of each wild-type and mutant plant were used for the validation assays.

The rapid advancement of genome editing makes it seem likely that within a few years one might be able to apply genome editing to evaluate the function of tens if not hundreds of candidates identified through GWAS. The potential for relatively high-throughput functional validation would not only make GWAS a powerful approach for prioritizing candidate genes for targeting through gene editing, but such large-scale validation would be invaluable for evaluating the strength of GWAS and for providing insight into the genomic basis of quantitative traits.

In cases where reverse genetic approaches are not possible, other gene manipulation and cell biological strategies can still be used to pursue the question of candidate gene validation. For example, detailed examination of gene expression in the context of the phenotype of interest can provide support and better insight into biological function. Overexpression and reporter experiments together with RNA in situ hybridization potentially drill down into gene function, and depending upon the nature of the phenotype that is being targeted in the GWA mapping, provide insights into the relationship of the candidate gene to the trait of interest.

LIMITATIONS OF GWAS

GWAS have proven to be effective at identifying genes contributing to phenotypic variation in several plant species including maize, *Arabidopsis thaliana*, poplar, pine, rice, and *Medicago truncatula*, among others (Zhu et al., 2008). Nevertheless, identifying some genes that contribute to variation is far from identifying the complete genomic basis of variation, and GWAS have several important limitations. First, unless conducted with tens of thousands of accessions, GWAS are unlikely to identify genes of truly small effect (Bush and Moore, 2012). Second, recessive alleles are usually not identified because heterozygotes will have the same phenotype as that of the homozygous dominant, though this is less of an issue for inbred lines or highly selfing species. Third, epistatic interactions—where an allelic effect on a phenotype is influenced by allelic variation at other loci (Phillips, 2008)—are statistically and computationally complicated because of the enormous number of tests required to conditionally test loci against each other. Nevertheless, recent methods using two stage testing procedures (Dai et al., 2012) are gaining traction. The potential importance of epistasis is illustrated by an examination of GWAS candidates for the "mustard oil bomb" in Brassicaceae (Brachi et al., 2015). Traditional GWAS analyses also appear to have limited power for identifying genes responsible for phenotypic plasticity (GxE; Sasaki et al., 2015), something that is potentially important for crop breeding and conservation in a changing world (El-Soda et al., 2014).

THE FUTURE OF GWAS

As a fundamental part of the toolkit for identifying naturally variable genomic regions underlying phenotypic variation, GWAS are now being creatively applied to myriad problems. For instance, the genomic and phenotypic data used for GWAS can be used to estimate heritability without the need for pedigree information (Speed and Balding, 2015). Similarly, the data used for GWAS can also be used to predict phenotypes using linear models, a process that is central to breeding through genomic selection (Goddard, 2009; Jannink et al., 2010; de los Campos et al., 2013). Genomic selection can greatly speed up selection cycles and therefore provide an economical and efficient approach for crop and tree improvement (Heslot et al., 2015). While GWAS and genomic selection are normally applied to achieve distinct goals, effect sizes from genomic predication and GWAS have recently been cross-referenced to identify novel candidates for morphological traits in Arabidopsis

(Kooke et al., 2016), and both approaches have been leveraged to identify targets for improving the resistance of Cassava to mosaic virus (Wolfe et al., 2016).

Similarly, many types of phenotypes can be used as a target for association analysis including less standard ones such as climate at the site of origin (Yoder et al., 2014), gene expression levels and co-expression networks (Chan et al., 2011; Schaefer et al., 2016), ionomes (Shakoor et al., 2016), or even principle component axes of multivariate traits (Aschard et al., 2014). Furthermore, advances in sequencing technology and variant detection algorithms offer the potential to test previously unexplored types of genomic variants, including copy number variants and structural variants. Increased statistical power can be leveraged by looking at enrichment of associations in a pathway of interest as opposed to each variant separately (Lipka et al., 2013). Lastly, the GWA concept is expanding from the frequentist approach to fitting models to Bayesian (Li et al., 2011) and machine learning approaches (Szymczak et al., 2009; Stephan et al., 2015).

Regardless of the specific phenotypes or whether GWAS is being used as a stand-alone approach or together with other omic technologies, robust experimental design, careful analysis, and informed interpretation of results are important for successful identification of genes underlying phenotype variation.

ACKNOWLEDGEMENTS

The authors would like to thank J. Guhlin, B. Epstein, S. J. Curtin, and J.-M. Michno for useful conversations. This work was supported, in part, by NSF award IOS-1237993.

LITERATURE CITED

Aschard, H., Vilhjálmsson, B.J., Greliche, N., Morange, P.E., Trégouët, D.A., and Kraft, P. 2014. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am. J. Hum. Genet.* 94:662-676. doi: 10.1016/j.ajhg.2014.03.016.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., Jiang, R, Muliyati, N.W., Zhang, X., Amer, M.A., Baxter, I., Brachi, B., Chory, J, Dean, C., Debieu, M., de Meaux, J., Ecker, J.R., Faure, N., Kniskern, J.M., Jones, J.D., Michael, T., Nemri, A., Roux, F., Salt, D.E., Tang, C., Todesco, M., Traw, M.B., Weigel, D., Marjoram, P., Borevitz, J.O., Bergelson, J., and Nordborg, M. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465:627-631. doi: 10.1038/nature08800.

Barton, N.H. and Turelli, M. 1989. Evolutionary quantitative genetics: How little do we know? *Annu. Rev. Genet.* 23:337-370. doi: 10.1146/annurev.ge.23.120189.002005.

Bergelson, J. and Roux, F. 2010. Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nat. Rev. Genet.* 11:867-879. doi: 10.1038/nrg2896.

Bortesi, L., Zhu, C., Zischewski, J., Perez, L., Bassié, L., Nadi, R., Forni, G., Lade, S.B., Soto, E., Jin, X., Medina, V., Villorbina, G., Muñoz, P., Farré, G., Fischer, R., Twyman, R.M., Capell, T., Christou, P., and Schillberg, S. 2016. Patterns of CRISPR/Cas9 activity in plants, animals and microbes. *Plant Biotechnol. J.* 1-14. doi: 10.1111/pbi.12634.

Brachi, B., Meyer, C.G., Villoutreix, R., Platt, A., Morton, T.C., Roux, F., and Bergelson, J. 2015. Coselected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 112:4032-4037. doi: 10.1073/pnas.1421416112.

Branca, A., Paape, T.D., Zhou, P., Briskine, R., Farmer, A.D., Mudge, J., Bharti, A.K., Woodward, J.E., May, G.D., Gentzbittel, L., Ben, C., Denny, R., Sadowsky, M.J., Ronfort, J., Bataillon, T., Young, N.D., and Tiffin, P. 2011. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proc. Natl. Acad. Sci. U.S.A.* 108:E864-870. doi: 10.1073/pnas.1104032108.

Browning, S.R. and Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81:1084-1097. doi: 10.1086/521987.

Browning, B.L. and Browning, S.R. 2016. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* 98:116-126. doi: 10.1016/j.ajhg.2015.11.020.

Bunyavanich, S., Schadt, E.E., Himes, B.E., Lasky-Su, J., Qiu, W., Lazarus, R., Ziniti, J.P., Cohain, A., Linderman, M., Torgerson, D.G., Eng, C.S., Pino-Yanes, M., Padhukasahasram, B., Yang, J.J., Mathias, R.A., Beaty, T.H., Li, X., Graves, P., Romieu, I., Navarro Bdel, R., Salam, M.T., Vora, H., Nicolae, D.L., Ober, C., Martinez, F.D., Bleecker, E.R., Meyers, D.A., Gauderman, W.J., Gilliland, F., Burchard, E.G., Barnes, K.C., Williams, L.K., London, S.J., Zhang, B., Raby, B.A., and Weiss, S.T. 2014. Integrated genome-wide association, coexpression network, and expression single nucleotide polymorphism analysis identifies novel pathway in allergic rhinitis. *BMC Med. Genomics* 7:48. doi: 10.1186/1755-8794-7-48.

Bush, W.S. and Moore, J.H. 2012. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* 8:e1002822. doi: 10.1371/journal.pcbi.1002822.

Chan, E.K.F., Rowe, H.C., Corwin, J.A., Joseph, B., and Kliebenstein, D.J. 2011. Combining

genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. *PLoS Biol*. 9:e1001125. doi: 10.1371/journal.pbio.1001125.

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., Liu, X., Zhang, H., Dong, H., Zhang, W., Zhang, L., Yu, S., Wang, G., Lian, X., and Luo, J. 2014. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46:714-721. doi: 10.1038/ng.3007.

Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedón, J.C., Redline, S., Papanicolaou, G.J., Thornton, T.A., Laurie, C.C., Rice, K., and Lin, X. 2016. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* 98:653-666. doi: 10.1016/j.ajhg.2016.02.012.

Curtin, S.J., Tiffin, P., Guhlin, J., Trujillo, D.I., Burghardt, L.T., Atkins, P., Baltes, N.J., Denny, R., Voytas, D.F., Stupar, R.M., and Young, N.D. 2017. Validating GWAS candidates by characterizing genes that control quantitative variation in nodulation. *Plant Physiol.* [ePub ahead of print] doi: 10.1104/pp.16.01923.

Dai, J.Y., Kooperberg, C., Leblanc, M., and Prentice, R.L. 2012. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* 99:929-944. doi: 10.1093/biomet/ass044.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., and Calus, M.P.L. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193:327-345. doi: 10.1534/genetics.112.143313.

El-Soda, M., Malosetti, M., Zwaan, B.J., Koornneef, M., and Aarts, M.G.M. 2014. Genotype × environment interaction QTL mapping in plants: Lessons from Arabidopsis. *Trends Plant Sci*. 19:390-398. doi: 10.1016/j.tplants.2014.01.001.

Falconer, D.S. and Mckay, T.E.C. 1996. Introduction to Quantitative Genetics, 4th ed. Pearson, London.

Fisher, R.A. 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, U.K.

Gaut, B.S. and Long, A.D. 2003. The lowdown on linkage disequilibrium. *Plant Cell* 15:1502-1506. doi: 10.1105/tpc.150730.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* 136:245-257. doi: 10.1007/s10709-008-9308-0.

Goddard, M.E., Kemper, K.E., MacLeod, I.M., Chamberlain, A.J., and Hayes, B.J. 2016. Genetics of complex traits: Prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc. R. Soc. B Biol. Sci.* 283:1173-1186. doi: 10.1098/rspb.2016.0569.

Goh, L. and Yap, V.B. 2009. Effects of normalization on quantitative traits in association test. *BMC Bioinformatics* 10:415. doi: 10.1186/1471-2105-10-415.

Greene, C.S., Penrod, N.M., Williams, S.M., and Moore, J.H. 2009. Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS One* 4:e5639. doi: 10.1371/journal.pone.0005639.

Halperin, E. and Stephan, D.A. 2009. SNP imputation in association studies. *Nat. Biotechnol.* 27:349-351. doi: 10.1038/nbt0409-349.

Heslot, N., Jannink, J.-L., and Sorrells, M.E. 2015. Perspectives for genomic selection applications and research in plants. *Crop Sci*. 55:1-12. doi: 10.2135/cropsci2014.03.0249.

Hoffman, G.E., Mezey, J.G., and Schadt, E.E. 2014. Lrgpr: Interactive linear mixed model analysis of genome-wide association studies with composite hypothesis testing and regression diagnostics in R. *Bioinformatics* 30:3134-3135. doi: 10.1093/bioinformatics/btu435.

Houston, K., Burton, R.A., Sznajder, B., Rafalski, A.J., Dhugga, K.S., Mather, D.E., Taylor, J., Steffenson, B.J., Waugh, R., and Fincher, G.B. 2015. A genome-wide association study for culm cellulose content in barley reveals candidate genes co-expressed with members of the cellulose synthase a gene family. *PLoS One* 10:e0130890. doi: 10.1371/journal.pone.0130890.

Howie, B.N., Donnelly, P., and Marchini, J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 5:e1000529. doi: 10.1371/journal.pgen.1000529.

Huang, X. and Han, B. 2014. Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* 65:531-551. doi: 10.1146/annurev-arplant-050213-035715.

Jannink, J.-L., Lorenz, A.J., and Iwata, H. 2010. Genomic selection in plant breeding: From theory to practice. *Brief. Funct. Genom. Proteom.* 9:166-177. doi: 10.1093/bfgp/elq001.

Jia, P. and Zhao, Z. 2014. Network-assisted analysis to prioritize GWAS results: Principles, methods and perspectives. *Hum. Genet.* 133:125-138. doi: 10.1007/s00439-013-1377-1.

Kooke, R., Kruijer, W., Bours, R., Becker, F., Kuhn, A., van de Geest, H., Buntjer, J., Doeswijk, T., Guerra, J., Bouwmeester, H., Vreugdenhil, D., and Keurentjes, J.J. 2016. Genome-wide association mapping and genomic prediction elucidate the genetic architecture of morphological traits in *Arabidopsis thaliana*. *Plant Physiol*. 170:pp.00997.2015. doi: 10.1104/pp.15.00997.

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

Kover, P.X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I.M., Purugganan, M.D., Durrant, C., and Mott, R. 2009. A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*.

PLoS Genet. 5:e1000551. doi: 10.1371/journal.pgen.1000551.

Kump, K.L., Bradbury, P.J., Wisser, R.J., Buckler, E.S., Belcher, A.R., Oropeza-Rosas, M.A, Zwonitzer, J.C., Kresovich, S., McMullen, M.D., Ware, D., Balint-Kurti, P.J., and Holland, J.B. 2011. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* 43:163-168. doi: 10.1038/ng.747.

Li, J., Das, K., Fu, G., Li, R., and Wu, R. 2011. The Bayesian lasso for genome-wide association studies. *Bioinformatics* 27:516-523. doi: 10.1093/bioinformatics/btq688.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., and Zhang, Z. 2012. GAPIT: Genome association and prediction integrated tool. *Bioinformatics* 28:2397-2399. doi: 10.1093/bioinformatics/bts444.

Lipka, A.E., Gore, M.A., Magallanes-Lundback, M., Mesberg, A., Lin, H., Tiede, T., Chen, C., Buell, C.R., Buckler, E.S., Rocheford, T., and DellaPenna, D. 2013. Genome-wide association study and pathway level analysis of tocochromanol levels in maize grain. *G3: Genes Genomes Genetics* 3:1287-1299. doi: 10.1534/g3.113.006148.

Liu, Y.J., Papasian, C.J., Liu, J.F., Hamilton, J., and Deng, H.W. 2008. Is replication the gold standard for validating genome-wide association findings? *PLoS One* 3:e4037. doi: 10.1371/journal.pone.0004037.

Lowder, L.G., Zhang, D., Baltes, N.J., Paul, J.W., Tang, X., Zheng, X., Voytas, D.F., Hsieh, T.-F., Zhang, Y., and Qi, Y. 2015. A CRISPR/Cas9 toolbox for multiplexed plant genome editing and transcriptional regulation. *Plant Physiol*. 169:971-985. doi: 10.1104/pp.15.00636.

Lowry, D.B., Hoban, S., Kelley, J.L., Lotterhos, K.E., Reed, L.K., Antolin, M.F., and Storfer, A. 2016. Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour.* [ePub ahead of print] doi: 10.1111/1755-0998.12596.

Mackay, T.F.C., Stone, E.A, and Ayroles, J.F. 2009. The genetics of quantitative traits: Challenges and prospects. *Nat. Rev. Genet.* 10:565-577. doi: 10.1038/nrg2612.

Marchini, J. and Howie, B. 2010. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11:499-511. doi: 10.1038/nrg2796.

Matsuda, F., Nakabayashi, R., Yang, Z., Okazaki, Y., Yonemaru, J.I., Ebana, K., Yano, M., and Saito, K. 2015. Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. *Plant J*. 81:13-23. doi: 10.1111/tpj.12681.

Meirmans, P.G. 2012. The trouble with isolation by distance. *Mol. Ecol.* 21:2839-2846. doi: 10.1111/j.1365-294X.2012.05578.x.

Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., and Visscher, P.M. 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet*. 11:e1004969. doi: 10.1371/journal.pgen.1004969.

NCI-NHGRI Working Group, Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., Bailey-Wilson, J.E., Brooks, L.D., Cardon, L.R., Daly, M., Donnelly, P., Fraumeni, J.F. Jr, Freimer, N.B., Gerhard, D.S., Gunter, C., Guttmacher, A.E., Guyer, M.S., Harris, E.L., Hoh, J., Hoover, R., Kong, C.A., Merikangas, K.R., Morton, C.C., Palmer, L.J., Phimister, E.G., Rice, J.P., Roberts, J., Rotimi, C., Tucker, M.A., Vogan, K.J., Wacholder, S., Wijsman, E.M., Winn, D.M., and Collins, F.S. 2007. Replicating genotype-phenotype associations. *Nature* 447:655-660. doi: 10.1038/447655a.

Nemri, A., Atwell, S., Tarone, A.M., Huang, Y.S., Zhao, K., Studholme, D.J., Nordborg, M., and Jones, J.D.G. 2010. Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. *Proc. Natl. Acad. Sci. U.S.A.* 107:10302-10307. doi: 10.1073/pnas.0913160107.

Ogura, T. and Busch, W. 2015. From phenotypes to causal sequences: Using genome wide association studies to dissect the sequence basis for variation of plant development. *Curr. Opin. Plant Biol.* 23:98-108. doi: 10.1016/j.pbi.2014.11.008.

Pease, J.B., Haak, D.C., Hahn, M.W., and Moyle, L.C. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol*. 14:e1002379. doi: 10.1371/journal.pbio.1002379.

Peiffer, J.A., Romay, M.C., Gore, M.A., Flint-Garcia, S.A., Zhang, Z., Millard, M.J., Gardner, C.A., McMullen, M.D., Holland, J.B., Bradbury, P.J., and Buckler, E.S. 2014. The genetic architecture of maize height. *Genetics* 196:1337-1356. doi: 10.1534/genetics.113.159152.

Petryszak, R., Keays, M., Tang, Y.A., Fonseca, N.A., Barrera, E., Burdett, T., Füllgrabe, A., Fuentes, A.M., Jupp, S., Koskinen, S., Mannion, O., Huerta, L., Megy, K., Snow, C., Williams, E., Barzine, M., Hastings, E., Weisser, H., Wright, J., Jaiswal, P., Huber, W., Choudhary, J., Parkinson, H.E., and Brazma, A. 2016. Expression Atlas update—An integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res*. 44:D746-D752. doi: 10.1093/nar/gkv1045.

Phillips, P.C. 2008. Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* 9:855-867. doi: 10.1038/nrg2452.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904-909. doi: 10.1038/ng1847.

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.

Raj, A., Stephens, M., and Pritchard, J.K. 2014. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* 197:573-589. doi: 10.1534/genetics.114.164350.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M., and Buckler, E.S. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. U.S.A.* 98:11479-11484. doi: 10.1073/pnas.201394398.

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., and Kim, D. 2015. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* 16:85-97. doi: 10.1038/nrg3868.

Ron, M., Kajala, K., Pauluzzi, G., Wang, D., Reynoso, M.A., Zumstein, K., Garcha, J., Winte, S., Masson, H., Inagaki, S., Federici, F., Sinha, N., Deal, R.B., Bailey-Serres, J., and Brady S.M. 2014. Hairy root transformation using Agrobacterium rhizogenes as a tool for exploring cell type-specific gene expression and function using tomato as a model. *Plant Physiol*. 166:455-469. doi: 10.1104/pp.114.239392.

Sasaki, E., Zhang, P., Atwell, S., Meng, D., and Nordborg, M. 2015. "Missing" G x E variation controls flowering time in *Arabidopsis thaliana*. *PLoS Genet.* 11:e1005597. doi: 10.1371/journal.pgen.1005597.

Schaefer, R.J., Michno, J.-M., and Myers, C.L. 2016. Unraveling gene function in agricultural species using gene co-expression networks. *Biochim. Biophys. Acta* S1874-9399:30166-30163. doi: 10.1016/j.bbagrm.2016.07.016.

Scheet, P. and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78:629-644. doi: 10.1086/502802.

Shakoor, N., Ziegler, G., Dilkes, B.P., Brenton, Z., Boyles, R., Connolly, E.L., Kresovich, S., and Baxter, I.R. 2016. Integration of experiments across diverse environments identifies the genetic determinants of variation in Sorghum bicolor seed element composition. *Plant Physiol*. 170:1989-1998. doi: 10.1104/pp.15.01971.

Slatkin, M. 2008. Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9:477-485. doi: 10.1038/nrg2361.

Speed, D. and Balding, D.J. 2015. Relatedness in the post-genomic era: Is it still useful? *Nat. Rev. Genet.* 16:33-44. doi: 10.1038/nrg3821.

Stanton-Geddes, J., Paape, T., Epstein, B., Briskine, R., Yoder, J., Mudge, J., Bharti, A.K., Farmer, A.D., Zhou, P., Denny, R., May, G.D., Erlandson, S., Yakub, M., Sugawara, M., Sadowsky, M.J., Young, N.D., and Tiffin, P. 2013. Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in Medicago truncatula. *PloS One* 8:e65688. doi: 10.1371/journal.pone.0065688.

Stephan, J., Stegle, O., and Beyer, A. 2015. A random forest approach to capture genetic effects in the presence of population structure. *Nat. Commun.* 6:7432. doi: 10.1038/ncomms8432.

Stephens, M. and Scheet, P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* 76:449-462. doi: 10.1086/428594.

Strauch, R.C., Svedin, E., Dilkes, B., Chapple, C., and Li, X. 2015. Discovery of a novel amino acid racemase through exploration of natural variation in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 112:11726-11731. doi: 10.1073/pnas.1503272112.

Suren, H., Hodgins, K.A., Yeaman, S., Nurkowski, K.A., Smets, P., Rieseberg, L.H., Aitken, S.N., and Holliday, J.A. 2016. Exome capture from the spruce and pine giga-genomes. *Mol. Ecol. Resour.* 16:1136-1146. doi: 10.1111/1755-0998.12570.

Symonds, V.V., Godoy, A.V., Alconada, T., Botto, J.F., Juenger, T.E., Casal, J.J., and Lloyd, A.M. 2005. Mapping quantitative trait loci in multiple populations of Arabidopsis thaliana identifies natural allelic variation for trichome density. *Genetics* 169:1649-1658. doi: 10.1534/genetics.104.031948.

Szymczak, S., Biernacka, J.M., Cordell, H.J., González-Recio, O., König, I.R., Zhang, H., and Sun, Y.V. 2009. Machine learning in genome-wide association studies. *Genet. Epidemiol.* 33:51-57. doi: 10.1002/gepi.20473.

Vaistij, F.E., Gan, Y., Penfield, S., Gilday, A.D., Dave, A., He, Z., Josse, E.M., Choi, G., Halliday, K.J., and Graham, I.A. 2013. Differential control of seed primary dormancy in Arabidopsis ecotypes by the transcription factor SPATULA. *Proc. Natl. Acad. Sci. U.S.A.* 110:10866-10871. doi: 10.1073/pnas.1301647110.

van Leeuwen, E.M., Kanterakis, A., Deelen, P., Kattenberg, M.V., Genome of the Netherlands Consortium, Slagboom, P.E., de Bakker, P.I., Wijmenga, C., Swertz, M.A., Boomsma, D.I., van Duijn, C.M., Karssen, L.C., and Hottenga, J.J. 2015. Population-specific genotype imputations using minimac or IMPUTE2. *Nat. Protoc.* 10:1285-1296. doi: 10.1038/nprot.2015.077.

Vilhjálmsson, B.J. and Nordborg, M. 2013. The nature of confounding in genome-wide association studies. *Nat. Rev. Genet.* 14:1-2. doi: 10.1038/nrg3382.

Wolfe, M.D., Rabbi, I.Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., Lozano, R., del Carpio, D.P., Ramu, P., and Jannink, J.-L. 2016. Genome-wide association and prediction reveals the genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. *Plant Genome* 9:1-248. doi: 10.3835/plantgenome2015.11.0118.

Wolt, J.D., Wang, K., Sashital, D., and Lawrence-Dill, C.J. 2016. Achieving plant CRISPR

targeting that limits off-target effects. *Plant Genome* 9:1-8. doi: 10.3835/plantgenome2016.05.0047.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., and Visscher, P.M. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42:565-569. doi: 10.1038/ng.608.

Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A.V, Ingelsson, E., O'Connell, J.R., Mangino, M., Mägi, R., Madden, P.A., Heath, A.C., Nyholt, D.R., Martin, N.G., Montgomery, G.W., Frayling, T.M., Hirschhorn, J.N., McCarthy, M.I., Goddard, M.E., and Visscher, P.M., and GIANT Consortium. 2011. Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* 19:807-812. doi: 10.1038/ejhg.2011.39.

Yeaman, S. and Whitlock, M.C. 2011. The genetic architecture of adaptation under migration-selection balance. *Evolution* 65:1897-1911. doi: 10.1111/j.1558-5646.2011.01269.x.

Yoder, J.B., Stanton-Geddes, J., Zhou, P., Briskine, R., Young, N.D., and Tiffin, P. 2014. Genomic signature of adaptation to climate in Medicago truncatula. *Genetics* 196:1263-1275. doi: 10.1534/genetics.113.159319.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., and Buckler, E.S. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38:203-208. doi: 10.1038/ng1702.

Zhang, F., Wang, Y., and Deng, H.W. 2008. Comparison of population-based association study methods correcting for population stratification. *PLoS One* 3:e3392. doi: 10.1371/journal.pone.0003392.

Zhao, K., Tung, C.W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., Norton, G.J., Islam, M.R., Reynolds, A., Mezey, J., McClung, A.M., Bustamante, C.D., and McCouch, S.R. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* 2:467. doi: 10.1038/ncomms1467.

Zhou, L. and Holliday, J.A. 2012. Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC Genomics* 13:703. doi: 10.1186/1471-2164-13-703.

Zhu, C., Gore, M., Buckler, E.S., and Yu, J. 2008. Status and prospects of association mapping in plants. *Plant Genome J.* 1:5-20. doi: 10.3835/plantgenome2008.02.0089.

Zöllner, S. and Pritchard, J.K. 2007. Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* 80:605-615. doi: 10.1086/512821.