PCP
PLANT & CELL PHYSIOLOGY

Beyond the Standard GWAS—A Guide for Plant Biologists

Pieter Clauw[1,†], Thomas James Ellis[1,†], Hai-Jun Liu[1,2] and Eriko Sasaki[3,*]

[1]Gregor Mendel Institute of Molecular Plant Biology, Austrian Academy of Sciences, Vienna BioCenter (VBC), Dr. Bohr-Gasse 3, Vienna 1030, Austria
[2]Yazhouwan National Laboratory, Sanya 572024, China
[3]Faculty of Science, Kyushu University, 744, Motooka, Nishi-ku, Fukuoka 819-0395, Japan
[†]These authors contributed equally to this work.

*Corresponding author: E-mail, sasaki.eriko.997@m.kyushu-u.ac.jp

(Received 26 May 2024; Accepted 10 July 2024)

Classic genome-wide association studies (GWAS) look for associations between individual single-nucleotide polymorphisms (SNPs) and phenotypes of interest. With the rapid progress of high-throughput genotyping and phenotyping technologies, GWAS have become increasingly powerful for detecting genetic determinants and their molecular mechanisms underpinning natural phenotypic variation. However, GWAS frequently yield results with neither expected nor promising loci, nor any significant associations. This is often because associations between SNPs and a single phenotype are confounded, for example with the environment, other traits or complex genetic structures. Such confounding can mask true genotype–phenotype associations, or inflate spurious associations. To address these problems, numerous methods have been developed that go beyond the standard model. Such advanced GWAS models are flexible and can offer improved statistical power for understanding the genetics underlying complex traits. Despite this advantage, these models have not been widely adopted and implemented compared to the standard GWAS approach, partly because this literature is diverse and often technical. In this review, our aim is to provide an overview of the application and the benefits of various advanced GWAS models for handling complex traits and genetic structures, targeting plant biologists who wish to carry out GWAS more effectively.

Keywords: Advanced GWAS • Complex traits • Genetic architecture • Multiple traits

Introduction

Understanding the genetic basis of phenotypic diversity is a central question in biology. Genome-wide association studies (GWAS) use samples from natural populations and cultivars to identify associations between genetic variants and traits, and have become increasingly powerful with advances in high-throughput genotyping and phenotyping technologies (Dhondt et al. 2013, Ellegren 2014, Gill et al. 2022). GWAS of morphological and physiological traits have helped elucidate the genetic variants underlying biological pathways (Atwell et al. 2010), identifying variants associated with susceptibility and response to disease (Todesco et al. 2010, Demirjian et al. 2023), pinpoint targets for selective breeding (El-Soda et al. 2015, Albert et al. 2016, Yano et al. 2016), and illuminate the forces of selection in natural populations (Li et al. 2010, Fournier-Level et al. 2011, Josephs et al. 2017, Rees et al. 2020). GWAS have great potential for revealing the genetic basis of traits and understanding the interaction between genetic variation and environments.

The vast majority of GWAS have used a simple but powerful statistical model to relate genotypes to phenotypes. Under this 'standard GWAS' model (Supplementary note box **1**), an association is calculated between a single-nucleotide polymorphism (SNP) and a single phenotype for each SNP in turn. This standard GWAS is widely applied and has been the subject of several comprehensive reviews (e.g. Korte and Farlow 2013, Sul et al. 2018, Uffelmann et al. 2021). However, GWAS often yield results that are challenging to interpret, such as an absence of genetic associations at all, associations in regions with no clear link to the trait, or associations that cannot be validated. This is often because this simple model is insufficient to address the biological question at hand. In particular, GWAS relies on natural variation, which is often more genetically complex than that the standard model assumes (**Fig. 1**). First, there are often complex patterns of correlation between multiple traits and between traits and the environment (Devlin and Roeder 1999, Dickson et al. 2010, Platt et al. 2010). Second, there may be multiple segregating haplotypes that can obscure true patterns of associations at individual SNPs. Third, phenotypes are often measured with substantial noise, while the real effect sizes at individual SNPs are often small. These factors can cause true associations to be missed. They may lead to spurious associations, resulting in considerable wasted time and effort validating them

Plant Cell Physiol. 00(00): 1–13 (2024) doi:https://doi.org/10.1093/pcp/pcae079, Advance Access publication on 11 July 2024, available online at https://academic.oup.com/pcp

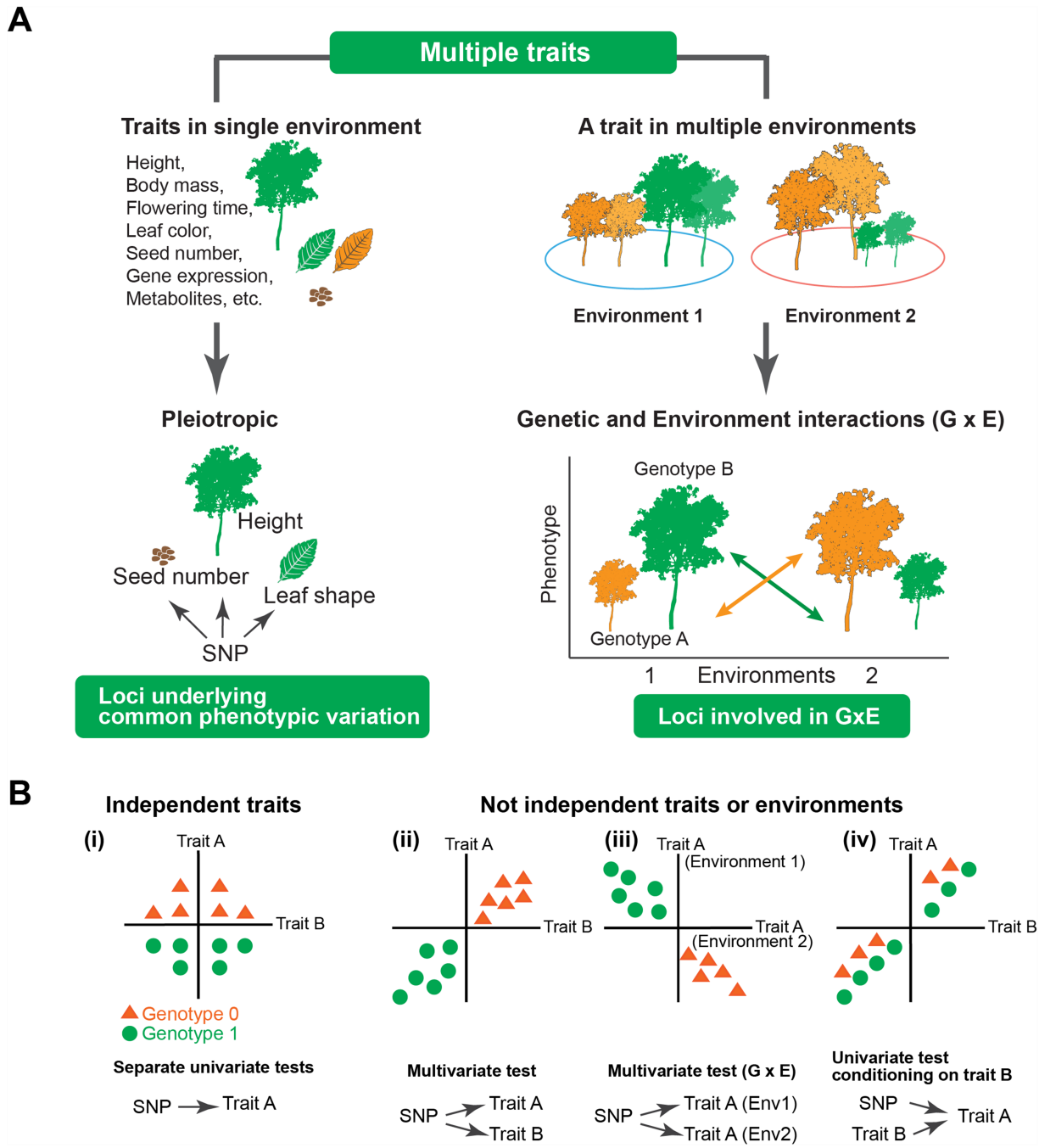

Fig. 1 The concept of multiple traits. (A) Examples of multiple traits: pleiotropic effects and genetic and environmental interactions (G x E) to be applied in multiple-trait analyses. (B) Four scenarios of association between a trait and a given SNP. Each point represents an individual with a shape (circle and triangle) corresponding to the genotype at a causal SNP. (i) Traits A and B are not correlated, and the SNP only affects trait A. Separate univariate tests detect SNPs underlying variation for each trait. (ii) Traits A and B are correlated when measured in the same environment, and the same SNP affects both traits (pleiotropic effect). A multivariate test of traits A and B can detect SNPs underlying both traits. (iii) Trait A varies between environments because the SNP genotypes respond differently in the two environments (GxE), leading to a correlation between environments. A multivariate test of a candidate SNP on trait A measured in different environments can distinguish the common effect in both environments as well as the effect in each environment separately (GxE). (iv) Trait A is regulated by both a causal SNP and an additional trait B. Individuals with genotype 0 at the SNP (triangle) have higher values for trait A than individuals with genotype 1 in a way that would be masked if the correlation between traits A and B were ignored. In this case, a univariate test conditioning on trait B can detect the SNP showing a trait-specific effect on trait A.

(Beavis 1994, Xu 2003, Platt et al. 2010). In these cases, the simple association between one SNP and one trait assumed by the standard GWAS is not a good model to understand the true genetic basis of the trait.

Fortunately, numerous methods have been developed that go beyond the standard model to address some of these problems (Tibbs Cortes et al. 2021). By accounting for the structure of the data more realistically, these methods have great

potential to identify genetic associations more accurately and efficiently. However, since the literature is diverse and often technical, these methods remain underutilized. In this review, we aim to highlight three broad groups of approaches that go beyond the standard GWAS (Supplementary note box **1**), which we believe are particularly relevant to questions in plant biology. This review is aimed at researchers without a strong statistics background but are nevertheless familiar with the standard GWAS and wish to go further. First, we highlight how modeling multiple traits in a single analysis can increase statistical power and interpretability. Second, we discuss what can be done to investigate an apparent association once one has been identified. Finally, we discuss what interesting conclusions may be drawn even if a study has found no peaks of association. Our goal is to build an intuition into why these methods are useful rather than go into statistical details.

Combining Multiple Traits in a Single Analysis

It is common for biological phenotypes to be correlated. When those traits share a genetic basis, the loci involved are said to be pleiotropic (**Fig. 1A**; Stearns 2010). Pleiotropy is usually thought of as reflecting correlations between traits in the same organism, such as vegetative size and reproductive output. Nevertheless, the idea is equally applicable to a single phenotype measured in multiple environments (Falconer 1952). In these cases, using GWAS to directly assess pleiotropic relationships or how phenotypes depend on the environment can be helpful in addressing the underlying biological questions. In this section we illustrate some ways to incorporate multiple traits into GWAS, focusing on (i) joint analysis of multiple phenotypes, (ii) how phenotypes change across environments and (iii) accounting for correlated traits not directly of interest.

Joint analysis of multiple traits

When we analyze the association between loci and multiple phenotypes in a single model, statistical power usually increases compared to multiple analyses of individual phenotypes (Stephens 2013). This gain in power in such 'multitrait' analyses comes from directly modeling the correlation in residual errors between traits (**Fig. 1**). Here we highlight several of the most popular multitrait models that are suitable as the number of traits increases from two to thousands. We focus on methods that estimate associations with multiple traits that are also able to account for genome-wide relatedness (Supplementary note box **1**). A detailed review and comparison of 10 related methods are given by Porter and O'Reilly (2017).

Building on methods for handling many traits in quantitative genetics, Korte et al. (2012) described a multiple-trait mixed model (MTMM) that linked multivariate regression with the population structure control needed for GWAS. For pairs of traits, MTMM estimates two separate effects for each SNP: the common genetic effect of the SNP on both traits, and a trait-specific effect. MTMM is implemented in LIMIX (Lippert et al. 2014). Zhou and Stephens (2014) extended this idea to allow for more than two phenotypes in a fully multivariate framework in the software package GEMMA. It can often help to transform phenotypes so that they are on the same scale (Schielzeth 2010). A good example of these approaches is that of Thoen et al. (2017), who identified loci associated with 30 stress responses and the shared genetic architectures in *Arabidopsis thaliana*. Associations were stronger and effect sizes were larger in multitrait compared to single-trait analyses.

It is increasingly feasible to generate datasets with hundreds or even thousands of traits, including phenomes from large-scale phenotyping technologies as well as genome-wide molecular phenotypes, such as the transcriptome, metabolome or epigenome. These phenotypes are typically regulated as networks, and a major goal is to understand the genetic regulation of these networks (Eichten et al. 2013, Fu et al. 2013, Schmitz et al. 2013, Dubin et al. 2015, Kawakatsu et al. 2016, Zhu et al. 2018). The scale of these datasets brings a substantial computational and multiple-testing burden that require different assumptions and approaches (Petretto et al. 2010, Ferguson et al. 2012, Flutre et al. 2013, Li et al. 2018). For example, the Multivariate Adaptive Shrinkage (MASH) approach of Urbut et al. (2019) addresses the computational and multiple-testing burdens by breaking up the task into two stages. MASH first estimates SNP effects on each trait separately. It then updates these initial values based on their standard errors and the correlation between them in a Bayesian framework to gain a more realistic picture of the relationship between SNPs and all traits combined. They applied this method to investigate how the association between local SNPs and gene expression varies across 44 human tissues, and found substantial heterogeneity in SNP effects across tissues.

Meta-analysis is an alternative approach for examining shared and trait-specific genetic effects as a post hoc analysis (Munafò and Flint 2004, Evangelou and Ioannidis 2013). Multivariate analyses can be effective but they require datasets in which all phenotypes have been measured for the same set of genotypes in order to fit a single model. Meta-analysis approach integrates the evidence for an association at each SNP across multiple univariate GWAS. Building on classical meta-analysis, the simplest approach is to sum negative log *P*-values, which is tantamount to asking whether the SNP shows associations with any of the datasets, making it a candidate for further investigation. Several alternative approaches have been developed to test more sophisticated null hypotheses (Evangelou and Ioannidis 2013), and many bioinformatics tools and software are available (Purcell et al. 2007, Mägi and Morris 2010). Examples of this method include summarizing pleiotropic effects of 234 agronomic traits in Sorghum (Mural et al. 2021), DNA methylation levels in 308 families of transposons in *A. thaliana* (Sasaki et al. 2019), comprehensive seed phenotypes in cowpea (Lo et al. 2019) and growth-related traits for four unrelated populations in Eucalyptus (Müller et al. 2019). One issue is that meta-GWAS essentially treats component studies as independent, and the resulting summary statistics will be biased if this is

not true. This is a particular concern for meta-GWAS on molecular phenotypes, which are often strongly correlated. While further development is clearly needed, meta-GWAS are still useful tools for generating hypotheses about interesting loci which can then be validated by further work.

Another approach is to simplify the data to one or a handful of dimensions prior to performing GWAS. It may be possible to synthesize multiple related traits into a single 'function-value trait' (Gomulkiewicz et al. 2018), which can then be analyzed as a single trait. More generally, principal component analysis (PCA) summarizes multivariate phenotypic data into a smaller set of variables that are orthogonal (i.e. not correlated) with one another (Pearson 1901, Ringnér 2008). This has the advantages that (i) the multiple testing problem is reduced (Weller et al. 1996), (ii) results may be more robust since skewed original phenotypic variations tend to be synthesized into a normal distribution (Kumar et al. 2022) and (iii) single-trait standard GWAS can be applied to these transformed phenotypes. Single-trait models for PCA-transformed traits have been widely applied, for example for flowering time in rice (Yano et al. 2019), microelement accumulation in maize (Ma et al. 2021), inflorescence and leaf architecture in maize (Rice et al. 2020), and root-system architecture in *A. thaliana* (Julkowska et al. 2017). The disadvantage of this approach is that the resulting principal components are synthetic traits and it can be difficult to interpret their biological meaning.

Interactions between genotype and the environment

Quantitative phenotypes typically depend at least to some extent on the environment. The environment may affect all phenotypes in a similar way (a direct environmental effect) but there can also be genotype-specific responses to each environment (a genotype-by-environment interaction, or GxE; **Fig. 1A**). For example, we would expect that crop yield would be reduced across genotypes when plants are exposed to a pathogen, but particular genotypes may be susceptible or resistant. An important example is when genotypes show increased yield or fitness in the region they were bred or evolved than do foreign genotypes grown at the same location, but reduced yield or fitness at other sites. GxE is thus an important concept in agriculture and environmental adaptation (Kawecki and Ebert 2004).

These problems lend themselves well to multitrait approaches such as those described above because they can directly estimate genetic, environmental and GxE effects. This has been applied, for example, to gene expression in different environments (Lippert et al. 2014, Clauw et al. 2016), GxE of drought responses in *A. thaliana* and tomato (El-Soda et al. 2015, Albert et al. 2016) and temperature-dependent flowering time in *A. thaliana* (Sasaki et al. 2015). An alternative approach is to directly estimate a measure of plasticity (Valladares et al. 2006, Filiault and Maloof 2012), and use this directly as a trait in a univariate GWAS. For example, Morrison and Linder (2014) did not find loci showing significant GxE interaction germination traits in *A. thaliana* in a multitrait model, but did identify significant genetic associations with reaction norms (**Fig. 1A**; a simple measure of the difference in phenotype between environments) for the same traits. This illustrates that these two approaches may capture different aspects of the data.

The inclusion of covariates: a double-edged sword

As previously mentioned, GWAS rely on natural variation, which is often confounded by spatial and environmental variables, which can lead to spurious genetic associations. For example, commercially important reproductive phenotypes in rice are strongly confounded with local adaptation and flowering time (Crowell et al. 2016). Just as we can adjust for confounding due to population structure (Supplementary note box **1**), GWAS can adjust for other sources of confounding by including additional information as covariates (**Fig. 1B**). Including covariates differs from the multiple-trait models described above in that the former includes additional explanatory variables in the model, while the latter includes additional response variables (**Fig. 1B**). In the rice example, including flowering time as a covariate in a GWAS of reproductive traits revealed additional genetic associations without the need for increased sample sizes (Crowell et al. 2016). Likewise, methylation of different sequence motifs in *A. thaliana* is partially regulated by the same pathways, and accounting for this allowed for the detection of quantitative trait locus (QTL) where none were found before (Sasaki et al. 2022). In both examples, the significant associations included known candidate genes, indicating that including these covariates yielded biologically meaningful results.

However, it is important to be aware that inappropriate covariates can also reduce power to detect true associations, or even amplify spurious associations (Mefford and Witte 2012, Pirinen et al. 2012, Stephens 2013). Whether or not to include covariates depends crucially on the causal relationships between variables, in particular whether the confounding variable is causative for the phenotype of interest or not. However, causal inference is challenging, and it is not always clear what the optimal model should be. We refer the reader to Stephens (2013) for a detailed discussion of this issue and to McElreath (2018) for an introduction to causal modeling. The inclusion of covariates in GWAS should be planned with care.

Following up on Associations

When a GWAS identifies one or more regions of the genome showing significant associations with a trait, what should be done next? Since a genetic association is merely a correlation, there is no substitute for validating the association with experimental evidence, such as mutants, crosses or allele swapping. Nevertheless, there are some statistical approaches that can be used to gain further insight into your initial results. In this section, we detail three of these approaches, focusing on (i) fine mapping to narrow down candidate causal variants, (ii)

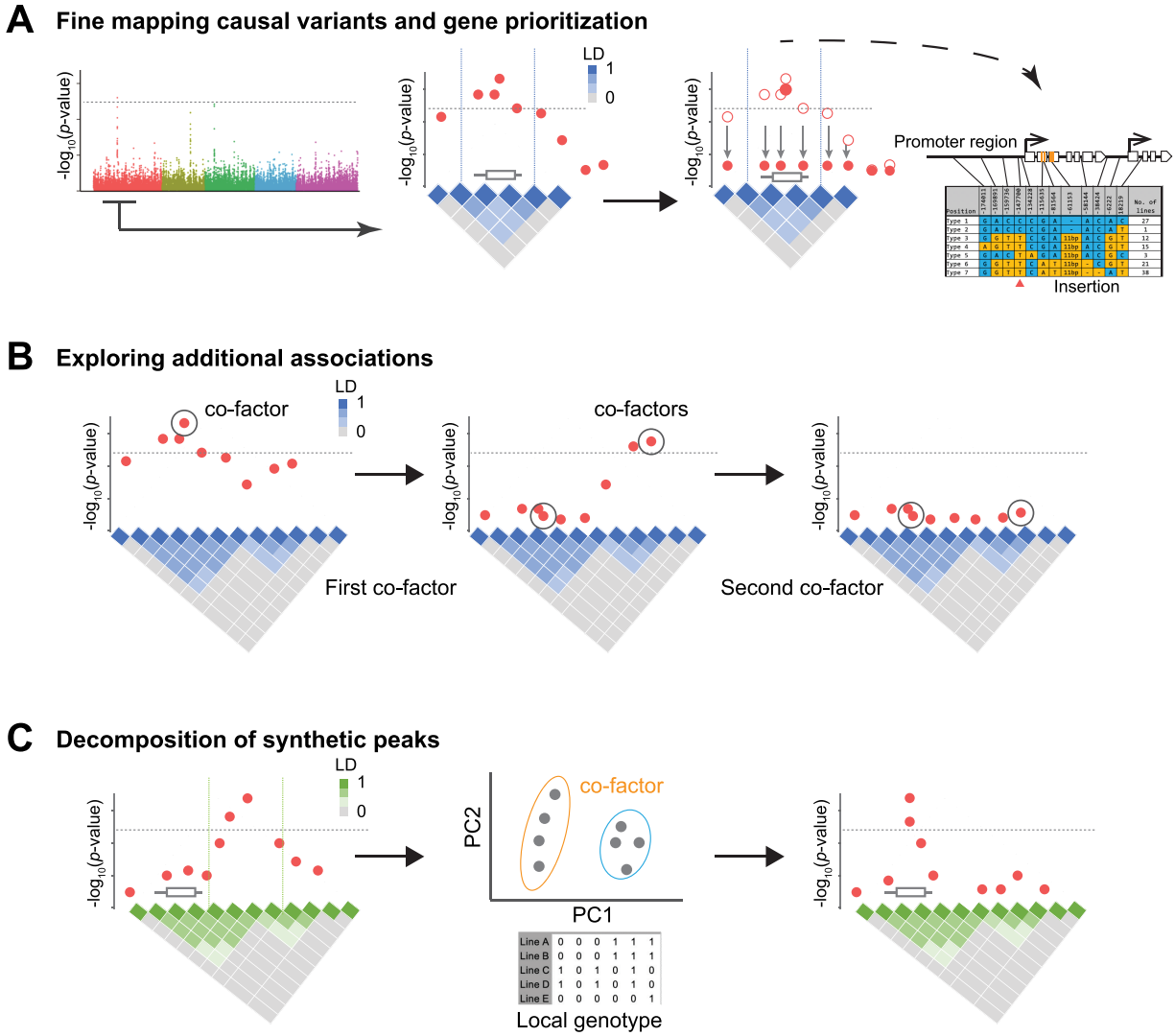

Fig. 2 Following up on associations. (A) Fine mapping: target regions for fine mapping are determined based on significant peaks from GWAS results. Genome regions around the peak are analyzed according to genetic structures, represented by linkage disequilibrium (LD). After narrowing down the region, a penalized model identifies a handful of candidate SNPs by estimating the effects of all selected SNPs at once, and penalizing or shrinking the original effects (open points) toward zero (filled points), leaving only those SNPs that explain the phenotype the best. Original *P*-values (filled plots) at most loci go to zero (open plots). Alignments, including indels and gene annotations, help to infer the biological mechanism driving the association with the phenotype. The table indicates alignments of the target regions and triangles below the table indicate candidate SNPs. (B) Exploring additional associations. After identifying an association with an initial GWAS, the genotype at the most strongly associated SNP is used as a cofactor in a second GWAS. This helps identify additional associations independent of the initial association and eliminates many associations in LD. Iterating this process can detect additional independent SNPs contributing to phenotypic variation in the targeted region. (C) Decomposition of synthetic peaks. An example of a synthetic peak linking multiple haplotypes containing causal alleles. PCA of genotypes around the peak reveals the genetic structure. Adding PC values as covariates in the model corrects the local genetic structure, and the association indicates a more accurate position of the causal variants.

using initial results to identify additional associations and (iii) assessment of whether associations are spurious (**Fig. 2**).

Fine mapping causal variants and gene prioritization

A region associated with a phenotype may contain hundreds of SNPs in linkage disequilibrium with the causal variant. Having identified initial associations, a next step might be to refine or 'fine map' the set of SNPs that are likely to be causal variants responsible for the phenotype. In GWAS, this is often done statistically. For example, penalized models estimate effects of all variants in a region at once in a way that penalizes or 'shrinks' the association at most loci to zero, leaving only one or a handful of SNPs with non-zero associations (**Fig. 2A**). A popular penalized model is lasso regression; see Spain and Barrett (2015), Schaid et al. (2018) and Uffelmann et al. (2021) for an

overview of related methods. Caution is required in interpreting the results as indicating that any variant has a direct causal effect on a trait. This assumes that a peak reflects a single causal SNP, that this SNP has been genotyped, and that the population is homogeneous. In reality, a causal variant is often an ungenotyped structural variant (**Fig. 2A**), there may be several causal mutations nearby one another, and patterns of association may be complicated by local or global genetic structure (Larsson et al. 2013, Hormozdiari et al. 2014, Spain and Barrett 2015). Nevertheless, with due care fine mapping can be a useful tool in narrowing down candidate variants for further investigation.

After fine-mapping, the next step is inferring the biological cause of the phenotype, according to the selected SNPs. This includes predicting the potential impact of variants on protein function or the disruption of regulatory elements. This process is challenging, but necessary for selecting SNPs for building biological hypotheses and experimentally validating those hypotheses. Currently, accumulating biological resources, including detailed gene annotation and population-level gene expression data, are available to further narrow down the candidates (Broekema et al. 2020, Uffelmann et al. 2021). In these cases, it is impossible to be sure about causality using GWAS alone, and it is therefore wise to follow-up on associations with additional data.

Exploring additional associations

In the standard GWAS, we typically test the association at one SNP at a time (Supplementary note box **1**). However, there may be multiple SNPs with substantial, independent effects on the trait, but which are correlated with one another due to physical linkage or population structure. In this case, the effects of these SNPs can obscure one another (Segura et al. 2012, Yang et al. 2012). A solution to this problem is to repeat the GWAS including the genotype at the most strongly associated SNP as a cofactor in a multilocus mixed model (Segura et al. 2012). This method often reveals additional peaks that were previously masked (**Fig. 2B**). For example, Dubin et al. (2015) identified a genetic association with DNA methylation close to the methyltransferase gene *CHROMOMETHYLASE 2* (CMT2). A subsequent GWAS using the genotype at that SNP as a cofactor revealed a second association at a nearby locus. Variants at these two loci were in perfect linkage disequilibrium, indicating that there had been two independent rounds of selection at this gene. Including genotypes as cofactors can be done manually with any GWAS software that accepts cofactors. Alternatively, an automated stepwise screening procedure is available in LIMIX (Lippert et al. 2014).

While useful, users should be aware that multilocus procedures are typically tantamount to stepwise regression, which has received substantial criticism (e.g. Harrell 2015). Nevertheless, as long as GWAS is performed with these caveats in mind, and especially when resulting peaks are independently validated, MLMM is a very useful tool to clarify genetic associations (Segura et al. 2012).

Decomposition of synthetic peaks

If a trait is controlled by multiple, locally, clustered loci, a non-causal SNP may often show a stronger association with the phenotype than any of the causal alleles (**Fig. 2C**). A common scenario is that causal alleles are only weakly associated with the phenotype because they are at low frequency, whereas certain non-causal alleles are at higher frequency but are linked to multiple causal alleles, and so 'absorb' the effects of those linked alleles (Devlin and Roeder 1999, Dickson et al. 2010, Platt et al. 2010). Such spurious associations are well known as 'synthetic peaks' or 'ghost peaks', and are usually caused by genetic heterogeneity, when multiple haplotypes segregate in a region that have not been broken up by recombination (Bergelson and Roux 2010, Platt et al. 2010). In addition to genetic heterogeneity, a recent study suggested that synthetic peaks reflect a signal of epistasis between SNPs (Liu et al. 2024).

Synthetic associations are most often detected by careful examination of association patterns and haplotype structures around significant peaks. This may reveal that the region of association is especially wide, that there are multiple peaks close to one another, or that the region includes a known candidate gene, but some distance from the strongest association. GWAS in *A. thaliana* have provided many examples of these patterns, including life history traits (Atwell et al. 2010, Kerdaffrec et al. 2016, Sasaki et al. 2021), and agronomic traits in tomato (Lin et al. 2014) and rice (Huang et al. 2010, Yano et al. 2016). Hidden haplotype structures can also be revealed by PCA of the SNP matrix in the region (e.g. Todesco et al. 2020; Sasaki et al. 2021) or the use of machine learning (Liu et al. 2024).

Once evidence for a synthetic association has been uncovered, the next step is to re-examine genetic associations within haplotype groups. This can be conducted either manually or statistically. For example, Yano et al. (2016) identified a genetic association with heading date in rice, which was close to but did not include the candidate gene *HEADING DATE 1* (*Hd1*), a flowering-time regulator. However, when the samples were split into subpopulations based on *Hd1* haplotype they did recover a genetic association at the *HD1* locus. Similarly, this stratification can be conducted statistically by including haplotypes as cofactors in a second GWAS analysis, as described in the previous section (Kerdaffrec et al. 2016, Sasaki et al. 2021).

When there are No Significant Associations

It may be the case that even a large, well-designed GWAS returns no significant genetic associations at all. In such cases, it can be tempting, if dispiriting, to conclude that the GWAS 'failed'. However, more than a century of work indicates that many heritable traits should be influenced by a large number of loci, each making a small contribution (Barton, Etheridge and Véber, 2017, Galton 1877, Fisher 1918). This is especially true for traits under natural or artificial selection, because selection quickly removes variation at these loci. With this in mind, the absence of strong genetic associations simply indicates that there are no alleles of large effect segregating in the sample, and it is

important to be aware that this is a perfectly valid conclusion to reach. Rather, this indicates that the interesting questions lie in the relationship between phenotypes and the relatedness between individuals. Alternatively, multiple alleles within a single gene resulting from independent selection events can disrupt true associations (Atwell et al. 2010). The absence of significant associations does not necessarily mean that a GWAS has 'failed'.

In this section, we outline steps that may be taken to follow-up on a GWAS that did not find strong genetic associations. First, there is much that can be learnt about the genetics of quantitative traits by focusing on phenotypes only; see Falconer and Mackay (1996) and Lynch and Walsh (1998) for an introduction to the topic, and Sella and Barton (2019) for a thorough review of the biology of quantitative genetic variation in the GWAS era. Here we highlight practical steps that may be taken that use genetic information directly, focusing on (i) how to quantify the extent to which a trait has any genetic basis at all, (ii) how to partition genetic signals from different parts of the genome and whether (iii) population structure or (iv) genetic heterogeneity obscures a true association (**Fig. 3**).

Quantifying the genetic basis of a trait

Heritability describes the proportion of overall trait variation that is due to genetic differences between individuals. This can be viewed as a direct quantitative estimate of the correlation between phenotype and relatedness. For example, the heritability of flowering time in *A. thaliana* and rice is >0.9, indicating that more than 90% of the variation is due to genetic differences (Sasaki et al. 2015). However, only a handful of significant associations with individual loci could be detected by GWAS, and their joint allelic effects explain only a part of total heritability (Yu et al. 2002, Li et al. 2010, 1001 Genomes Consortium 2016). This discrepancy indicates genetic variation in flowering time is due to many alleles with small effect sizes. On the other hand, low heritability estimates indicate that either the trait has a weak genetic basis, and/or that the trait is strongly influenced by the environment or is measured with substantial error (Houle 1992). It may be possible to improve the estimate of heritability by reviewing the study design to remove environmental effects and reduce measurement error, which may in turn allow genetic association to be detected. In this way, heritability is a useful step in determining how much GWAS can tell us, and highlights the need for careful study design.

There are two main approaches to estimating heritability. Classical quantitative genetic approaches use prior information about relatedness, for example by quantifying the variance in phenotype of individuals within and between multiple families or genotypes (Falconer and Mackay 1996). This can be done without genotype data and has a relatively straightforward interpretation, but may not always be possible to estimate. In contrast, so-called SNP heritability or pseudo-heritability estimates relatedness based on shared SNPs, and uses this to estimate the correlation with phenotype (Kang et al. 2010, Yang et al. 2010) (**Fig. 3A**). This is estimated by building a matrix of relatedness between all pairs of individuals, and using this to fit a random effect describing the variance in the phenotype explained by relatedness (**Fig. 3A**). This is very similar to population structure adjustment using a relatedness matrix in the standard GWAS, but without any main effects of individual SNP effects. This approach was motivated by the failure of conventional GWAS to identify variants affecting human height (the so-called 'missing heritability' debate); by taking all loci into account at once with a relatedness matrix, a much greater proportion of variance in height could be explained (Yang et al. 2010). The interpretation of SNP heritability is more complicated than classical heritability because it is sensitive to the effect sizes of causative SNPs. In particular, it may not be a good heritability estimate when the phenotype is controlled by a small number of loci (Yang et al. 2017) because SNP heritability assumes effects are spread fairly evenly across the genome.

Partitioning genetic variation across the genome

SNP heritability measures the relationship between differences in phenotype and relatedness across the whole genome (**Fig. 3A**). This idea can be taken further by partitioning the genome into units of interest, building separate matrices of relatedness for each unit and asking how much of the variance in phenotype is explained by each (Visscher et al. 2007, Yang et al. 2011). This efficiently describes the aggregate effect of all SNPs at once where the effects of any individual SNP would be too small to be detected. For example, Meng et al. (2016) compared the variance in gene expression explained by SNPs in *cis* and *trans* to each gene, as well as DNA methylation level at the gene, and found a primary role for *trans* effects. This approach can also be expanded to test polygenic GxE (Lippert et al. 2014). For example, Sasaki et al. (2015) found strong GxE effects in flowering time phenotypes for *A. thaliana* accessions grown at two temperatures, but the genetic basis for the variation was only revealed by taking the aggregate effects of many loci into account at once.

Population structure masks associations

An inherent challenge in GWAS is to account for population structure (**Fig. 3B**). On one hand, doing so is essential because not accounting for population structure generates false-positive associations (Kang et al. 2008, Yu et al. 2006, Vilhjálmsson and Nordborg 2013). On the other hand, this can obscure true associations that are correlated with population structure (Korte and Farlow 2013). A simple approach is to compare GWAS models with or without correction for population structure (Atwell et al. 2010). Another is to run separate GWAS on distinct subpopulations of the data set (Lopez-Arboleda et al. 2021). Although this likely entails a substantial loss of sample size, it may allow for the detection of alleles segregating within a population without overcorrecting for differences between populations (Sasaki et al. 2015, Gloss et al. 2022). This may itself reveal different evolutionary histories among populations (Lopez-Arboleda et al. 2021).

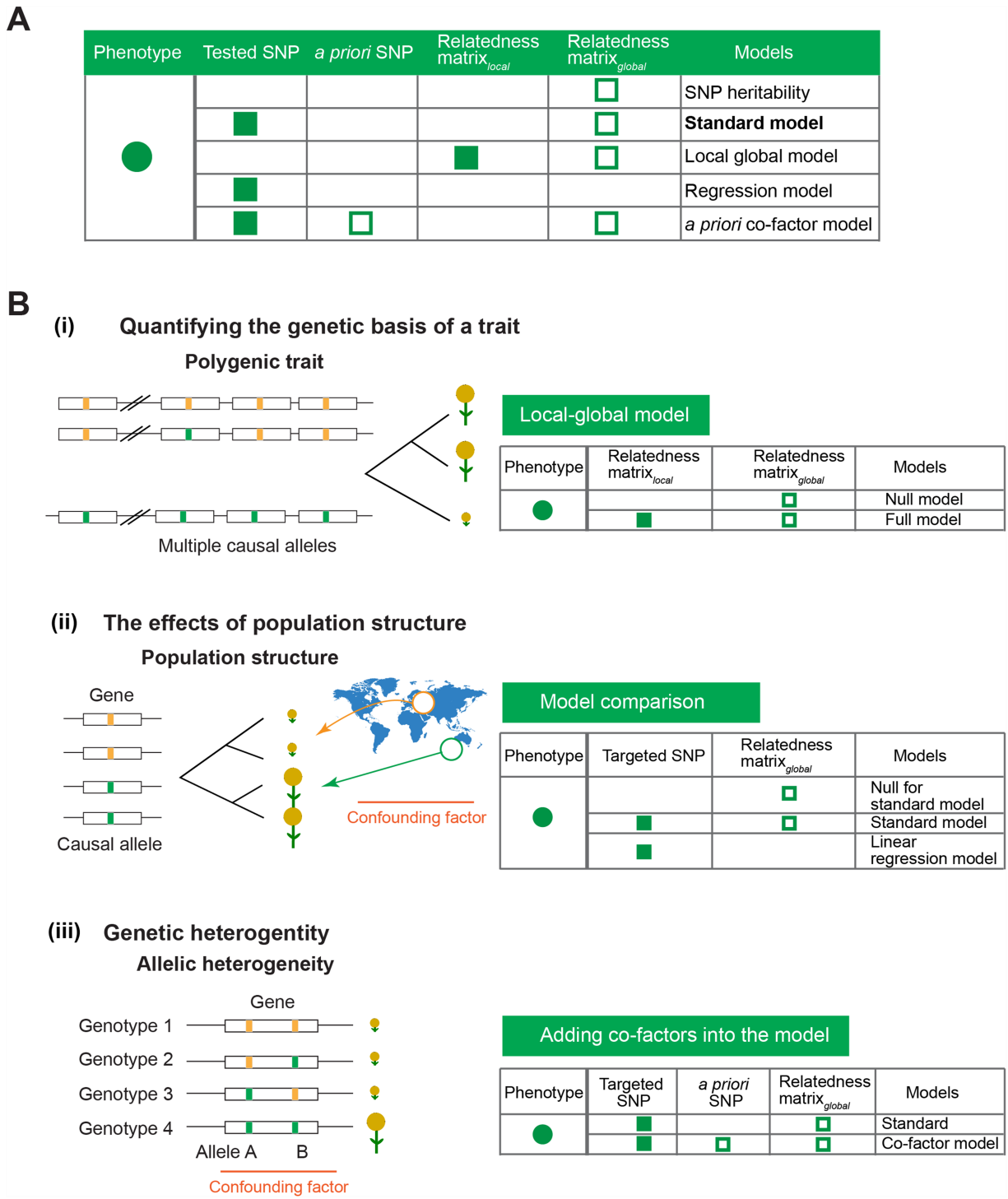

Fig. 3 Mapping approaches if there seem to be no significant associations. (A) Summary of basic models. Each row indicates models. Phenotype is the dependent variable (circle), and the others are independent variables to be tested (filled squares) and the correction (open squares). (B) Examples of factors masking significant associations. (i) Quantifying the genetic basis of a trait. For polygenic traits, clusters of SNPs each with small effects on the trait can be assessed using the local–global model. This model compares the variance in the trait explained by a relatedness matrix based on SNPs in a small region of the genome to that explained by a relatedness matrix based on genome-wide SNPs. (ii) The effects of population structure. When causal variants are correlated with population structure, then accounting for this structure with a relatedness matrix can obscure the association at these variants. These associations may be revealed by comparing GWAS with and without the correction for population structure. (iii) Genetic heterogeneity. When there are multiple causal SNPs that are confounded by complex haplotype structures incorporating *a priori* SNPs as a cofactor in the models help to identify associations at a more accurate position of the causal variants. SNP effects are tested with and without *a priori* information.

Population structure causes false-positive associations because it generates linkage disequilibrium between loci across the genome. Several methods aim to model this linkage directly by first identifying other SNPs in linkage disequilibrium with a test SNP and then recalculating the relatedness matrix excluding these SNPs (Listgarten et al. 2012, Wang et al. 2014). Fixed and random model Circulating Probability Unification (FarmCPU) takes these ideas a step further by combining the use of linked SNPs as cofactors and accounting for these SNPs in the relatedness matrix (Liu et al. 2016). This is done by alternately identifying associated SNPs, adjusting the relatedness matrix based on those SNPs, then testing associations at each SNP again, and so on, until no further improvement is possible.

Statistical corrections can alleviate, but are unlikely to completely eliminate, confounding with population structure, especially when this structure is strong. In these cases, it may be worth considering an alternative design that reduces population structure experimentally by crossing. Popular designs include backcrossing diverse genotypes to a single parent (nested association mapping; Yu et al. 2008), crossing multiple parental genotypes to each other (Kover et al. 2009, Liu et al. 2020), or combining data from multiple bi-parental crosses (Xiao et al. 2016). These designs can be seen as combining the advantages of high genetic diversity of natural populations with reduced population structure from crossing. Nevertheless, they require substantial effort to set up, cannot easily be augmented with additional samples as they become available, and may not be feasible in many species. Plant species are particularly amenable to these designs because they can often be inbred and seeds stored and reused. Plants also often show substantial population structure in nature, and so experimental crosses are often of great benefit in elucidating the genetic basis of traits (Kitony 2023).

Genetic heterogeneity may mask associations

We previously described how genetic heterogeneity can cause spurious genetic associations where none truly exists (**Fig. 2C**). It may also be that multiple causal variants within a single gene are segregating in population, but the true signal of each is diluted by genetic heterogeneity. This is known as allelic heterogeneity, and the classic example is in flowering time in *A. thaliana* (**Fig. 3B**). *FRIGIDA* (*FRI*), the major determinant of flowering time, controls *FLOWERING LOCUS C* (*FLC*), a suppressor of flowering time. Multiple independent loss-of-function alleles in *FRI* have arisen that dramatically shorten flowering time (Shindo et al. 2005, Fulgione et al. 2022). This means that each allele, while occurring at low frequency, is only partly associated with flowering time but strongly associated with population structure. This means that these associations have been challenging to detect with GWAS (Atwell et al. 2010).

Identifying allelic heterogeneity that masks associations is challenging, but if prior information about haplotypes is available, this can be included in the analysis to refine associations. For example, two commonly used lab strains of *A. thaliana*, Columbia and Landsberg erecta, are known to harbor independent loss-of-function mutations at *FRI*, but neither of these associations were found using a standard GWAS of flowering time phenotypes (Atwell et al. 2010). Including the haplotype state at these alleles as a cofactor in a GWAS on flowering time improved the associations (**Fig. 3B**). In the absence of prior knowledge, an alternative approach is gene-set analysis (de Leeuw et al. 2015). Rather than looking for associations with individual SNPs, this instead focuses on associations with entire genes. In a first step, this performs PCA of SNPs from an entire gene. If there is substantial structuring into distinct haplotypes, this should reflect a lot of the variation between genotypes and should explain the strongest principal components. The resulting principal components are then used as pseudo-genotypes to look for associations with the phenotype (de Leeuw et al. 2015).

Conclusions and Perspectives

With the rapid advancement of high-throughput genotyping and phenotyping technologies, GWAS has become increasingly powerful. The flexible GWAS models introduced in this review represent robust tools for elucidating the molecular and evolutionary basis of plants shaped by natural conditions.

This success has relied on a simple correlation between SNPs and traits of interest. However, this relationship is often distorted by confounding with other variables. Two sources of confounding have come up again and again in this review. First, SNPs in natural populations and cultivars always show some degree of linkage disequilibrium. Over short scales, alleles are arranged into haplotypes, causing correlations between nearby SNPs. Over longer scales, there will be correlations between SNPs due to population structure or selection. Thus, GWAS panels are fundamentally different from mutant screenings that use a single genetic background. Second, traits are often correlated with other biological traits or environmental variables. If ignored, these correlations can cause real associations to be missed, or spurious associations to be identified (**Fig. 2**). A common feature of many of the approaches we have outlined is that they aim to directly model the relationships between SNPs, traits of interest and confounders, and thereby increase the power to detect true genetic associations (**Fig. 3**).

Nevertheless, a major challenge is that the true causal relationship between variables is not known and often not obvious. Despite this, there are two steps that can be taken at different stages of a project. The first is to ensure that the experimental design is robust as it can be. High-quality phenotype data and designed GWAS panels make GWAS results more reliable (Myles et al. 2009, Ogura and Busch 2015). It is worth taking time to think through potential confounding variables and their possible relationships with the phenotype of interest, and planning how they can be accounted for experimentally or statistically (Stephens 2013). Collecting multiple relevant traits from the samples simultaneously allows for flexibility in the choice of analysis and reducing potential statistical issues (Stephens 2013;

see also Supplementary note box **1**). Note that there may often be multiple biologically plausible hypotheses and that embracing this is both legitimate and wise (Burnham and Anderson 2002, Betini et al. 2017).

The other is to approach analysis with a data exploration mind-set. Since confounding can take many forms that are difficult to predict from the outset, it can be useful to try several approaches and compare the results (**Figs. 1B**, **2**, and **3**). Some of the tools described here may be better or worse at describing different aspects of the data as they enable the modeling of relationships between genetic and phenotypic variables. For example, single and multitrait models can be seen as complementary approaches, and it can be worthwhile trying both. Likewise, confounding due to genetic heterogeneity is typically revealed by careful exploration of underlying haplotype structures. It is important to note that exploration should be done with care—simply trying different analyses until a desirable result is found is tantamount to *P*-value hacking, and liable to generate incorrect conclusions. It is better to remember that GWAS are best viewed as hypothesis-generating exercises, and that initial genetic associations are the starting point to explore and validate these hypotheses in more detail.

In many different fields, GWAS applications have brought us great new biological insights. The potential for continued discovery is vast, and increased usage of more advanced GWAS methods will further our understanding of the genetic regulation of phenotypic variation. The ongoing development of innovative methodologies will allow for asking unanswered questions that are currently limited by our computational capacities.

Supplementary Data

Supplementary data are available at *PCP* online.

Data Availability

There are no data to be declared.

Funding

Japan Society for the Promotion of Science (JP20K26671 and JP21H02538 to E.S.).

Acknowledgments

We would like to thank Magnus Nordborg for his continuous discussion and support in building the body of knowledge represented here. We also thank Tal Dahan-Meir, members of the Nordborg lab, and Takehiko Ogura for their critical reading and valuable comments on this manuscript.

Author Contributions

P.C., T.E. and E.S. planned the design, and P.C., T.E., H.L. and E.S. wrote the manuscript.

Disclosures

The authors have no conlicts of interest to declare.

References

1001 Genomes Consortium (2016) 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166: 481–491.

Albert, E., Segura, V., Gricourt, J., Bonnefoi, J., Derivot, L. and Causse, M. (2016) Association mapping reveals the genetic architecture of tomato response to water deficit: focus on major fruit quality traits. *J. Exp. Bot.* 67: 6413–6430.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., et al. (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465: 627–631.

Barton, N.H., Etheridge, A.M. and Véber, A. (2017) The infinitesimal model: definition, derivation, and implications. *Theor. Popul. Biol.* 118: 50–73.

Beavis, W.D. (1994) The power and deceit of QTL experiments: lessons from comparitive QTL studies. *In Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference* pp. 250–266. American Seed Trade Association, Washington, DC.

Bergelson, J. and Roux, F. (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nat. Rev. Genet.* 11: 867–879.

Betini, G.S., Avgar, T. and Fryxell, J.M. (2017) Why are we not evaluating multiple competing hypotheses in ecology and evolution? *R. Soc. Open Sci.* 4: 160756.

Broekema, R.V., Bakker, O.B. and Jonkers, I.H. (2020) A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* 10: 190221.

Burnham, K.P. and Anderson, D.R. (2002) Advanced issues and deeper insights. *In Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. pp. 267–351. Springer, New York.

Clauw, P., Coppens, F., Korte, A., Herman, D., Slabbinck, B., Dhondt, S., et al. (2016) Leaf growth response to mild drought: natural variation in Arabidopsis sheds light on trait architecture. *Plant Cell* 28: 2417–2434.

Crowell, S., Korniliev, P., Falcão, A., Ismail, A., Gregorio, G., Mezey, J., et al. (2016) Genome-wide association and high-resolution phenotyping link *Oryza sativa* panicle traits to numerous trait-specific QTL clusters. *Nat. Commun.* 7: 10527.

de Leeuw, C.A., Mooij, J.M., Heskes, T. and Posthuma, D. (2015) MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* 11: e1004219.

Demirjian, C., Vailleau, F., Berthomé, R. and Roux, F. (2023) Genome-wide association studies in plant pathosystems: success or failure? *Trends Plant Sci.* 28: 471–485.

Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics* 55: 997–1004.

Dhondt, S., Wuyts, N. and Inzé, D. (2013) Cell to whole-plant phenotyping: the best is yet to come. *Trends Plant Sci.* 18: 428–439.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. and Goldstein, D.B. (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol.* 8: e1000294.

Dubin, M.J., Zhang, P., Meng, D., Remigereau, M.-S., Osborne, E.J., Paolo Casale, F., et al. (2015) DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. *Elife* 4: e05255.

Eichten, S.R., Briskine, R., Song, J., Li, Q., Swanson-Wagner, R., Hermanson, P.J., et al. (2013) Epigenetic and genetic influences on DNA methylation variation in maize populations. *Plant Cell* 25: 2783–2797.

Ellegren, H. (2014) Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* 29: 51–63.

El-Soda, M., Kruijer, W., Malosetti, M., Koornneef, M. and Aarts, M.G.M. (2015) Quantitative trait loci and candidate genes underlying genotype

by environment interaction in the response of *Arabidopsis thaliana* to drought. *Plant Cell Environ*. 38: 585–599.
Evangelou, E. and Ioannidis, J.P.A. (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet*. 14: 379–389.
Falconer, D.S. (1952) The problem of environment and selection. *Am Nat* 86: 293–298.
Falconer, D.S. and Mackay, T.F.C. (1996) Introduction to Quantitative Genetics, 4th edn. Longman, Harlow, England.
Ferguson, J.P., Cho, J.H. and Zhao, H. (2012) A new approach for the joint analysis of multiple ChIP-seq libraries with application to histone modification. *Stat. Appl. Genet. Mol. Biol*. 11: Article 1.
Filiault, D.L. and Maloof, J.N. (2012) A genome-wide association study identifies variants underlying the *Arabidopsis thaliana* shade avoidance response. *PLoS Genet*. 8: e1002589.
Fisher, R.A. (1918) The correlation between relatives under the supposition of mendelian inheritance. *Trans. R. Soc. Edinb*. 52: 399–433.
Flutre, T., Wen, X., Pritchard, J. and Stephens, M. (2013) A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet*. 9: e1003486.
Fournier-Level, A., Korte, A., Cooper, M.D., Nordborg, M., Schmitt, J. and Wilczek, A.M. (2011) A map of local adaptation in *Arabidopsis thaliana*. *Science* 334: 86–89.
Fu, J., Cheng, Y., Linghu, J., Yang, X., Kang, L., Zhang, Z., et al. (2013) RNA sequencing reveals the complex regulatory network in the maize kernel. *Nat. Commun*. 4: 2832.
Fulgione, A., Neto, C., Elfarargi, A.F., Tergemina, E., Ansari, S., Göktay, M., et al. (2022) Parallel reduction in flowering time from *de novo* mutations enable evolutionary rescue in colonizing lineages. *Nat. Commun*. 13: 1461.
Galton, F. (1877) Typical laws of heredity. *Nature* 15: 492–95, 512–14, 532–33.
Gill, T., Gill, S.K., Saini, D.K., Chopra, Y., de Koff, J.P. and Sandhu, K.S. (2022) A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. *Phenomics* 2: 156–183.
Gloss, A.D., Vergnol, A., Morton, T.C., Laurin, P.J., Roux, F. and Bergelson, J. (2022) Genome-wide association mapping within a local *Arabidopsis thaliana* population more fully reveals the genetic architecture for defensive metabolite diversity. *Philos. Trans. R. Soc. Lond. B. Biol. Sci*. 377: 20200512.
Gomulkiewicz, R., Kingsolver, J.G., Carter, P.A. and Heckman, N. (2018) Variation and evolution of function-valued traits. *Annu. Rev. Ecol. Evol. Syst*. 49: 139–164.
Harrell, F.E. (2015) Multivariable Modeling Strategies. In: Regression Modeling Strategies. *Springer Series in Statistics* Springer, Cham.
Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. and Eskin, E. (2014) Identifying causal variants at loci with multiple signals of association. *Genetics* 198: 497–508.
Houle, D. (1992) Comparing evolvability and variability of quantitative traits. *Genetics* 130: 195–204.
Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet*. 42: 961–967.
Josephs, E.B., Stinchcombe, J.R. and Wright, S.I. (2017) What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? *New Phytol*. 214: 21–33.
Julkowska, M.M., Koevoets, I.T., Mol, S., Hoefsloot, H., Feron, R., Tester, M.A., et al. (2017) Genetic components of root architecture remodeling in response to salt stress. *Plant Cell* 29: 3198–3213.
Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y., Freimer, N.B., et al. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet*. 42: 348–354.
Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., et al. (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178: 1709–1723.
Kawakatsu, T., Huang, -S.-S.C., Jupe, F., Sasaki, E., Schmitz, R.J., Urich, M.A., et al. (2016) Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* 166: 492–505.
Kawecki, T.J. and Ebert, D. (2004) Conceptual issues in local adaptation. *Ecol. Lett*. 7: 1225–1241.
Kerdaffrec, E., Filiault, D.L., Korte, A., Sasaki, E., Nizhynska, V., Seren, Ü., et al. (2016) Multiple alleles at a single locus control seed dormancy in Swedish *Arabidopsis*. *Elife* 5: e22502.
Kitony, J.K. (2023) Nested association mapping population in crops: current status and future prospects. *J. Crop. Sci. Biotechnol*. 26: 1–12.
Korte, A. and Farlow, A. (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9: 29.
Korte, A., Vilhjálmsson, B.J., Segura, V., Platt, A., Long, Q. and Nordborg, M. (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics* 44: 1066–1071.
Kover, P.X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I.M., Purugganan, M.D., et al. (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet*. 5: e1000551.
Kumar, K., Anjoy, P., Sahu, S., Durgesh, K., Das, A., Tribhuvan, K.U., et al. (2022) Single trait versus principal component based association analysis for flowering related traits in pigeonpea. *Sci. Rep*. 12: 10453.
Larsson, S.J., Lipka, A.E. and Buckler, E.S. (2013) Lessons from *Dwarf8* on the strengths and weaknesses of structured association mapping. *PLoS Genet*. 9: e1003246.
Li, G., Shabalin, A.A., Rusyn, I., Wright, F.A. and Nobel, A.B. (2018) An empirical Bayes approach for multiple tissue eQTL analysis. *Biostatistics* 19: 391–406.
Li, Y., Huang, Y., Bergelson, J., Nordborg, M. and Borevitz, J.O. (2010) Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A*. 107: 21199–21204.
Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., et al. (2014) Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet*. 46: 1220–1226.
Lippert, C., Casale, F.P., Rakitsch, B. and Stegle, O. (2014) LIMIX: genetic analysis of multiple traits. *bioRxiv*. 003905.
Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E. and Heckerman, D. (2012) Improved linear mixed models for genome-wide association studies. *Nat. Methods* 9: 525–526.
Liu, H.-J., Swarts, K., Xu, S., Yan, J. and Nordborg, M. (2024) On the contribution of genetic heterogeneity to complex traits. *bioRxiv*. 2024.03.27.586967.
Liu, H.-J., Wang, X., Xiao, Y., Luo, J., Qiao, F., Yang, W., et al. (2020) CUBIC: an atlas of genetic architecture promises directed maize improvement. *Genome Biol*. 21: 20.
Liu, X., Huang, M., Fan, B., Buckler, E.S. and Zhang, Z. (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet*. 12: e1005767.
Lo, S., Muñoz-Amatriaín, M., Hokin, S.A., Cisse, N., Roberts, P.A., Farmer, A.D., et al. (2019) A genome-wide association and meta-analysis reveal regions associated with seed size in cowpea [Vigna unguiculata (L.) Walp]. *Theor. Appl. Genet*. 132: 3079–3087.
Lopez-Arboleda, W.A., Reinert, S., Nordborg, M. and Korte, A. (2021) Global genetic heterogeneity in adaptive traits. *Mol. Biol. Evol*. 38: 4822–4831.
Lynch, M. and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc, Sunderland, MA.
Ma, L., Qing, C., Zhang, M., Zou, C., Pan, G. and Shen, Y. (2021) GWAS with a PCA uncovers candidate genes for accumulations of microelements in maize seedlings. *Physiol. Plant* 172: 2170–2180.

Mägi, R. and Morris, A.P. (2010) GWAMA: software for genome-wide association meta-analysis. *BMC Bioinf*. 11: 288.
McElreath, R. (2018) *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, New York.
Mefford, J. and Witte, J.S. (2012) The covariate's dilemma. *PLoS Genet*. 8: e1003096.
Meng, D., Dubin, M., Zhang, P., Osborne, E.J., Stegle, O., Clark, R.M., et al. (2016) Limited contribution of DNA methylation variation to expression regulation in *Arabidopsis thaliana*. *PLoS Genet*. 12: e1006141.
Morrison, G.D. and Linder, C.R. (2014) Association mapping of germination traits in *Arabidopsis thaliana* under light and nutrient treatments: searching for G×E effects. *G3* 4: 1465–1478.
Müller, B.S.F., de Almeida Filho, J.E., Lima, B.M., Garcia, C.C., Missiaggia, A., Aguiar, A.M., et al. (2019) Independent and joint-GWAS for growth traits in Eucalyptus by assembling genome-wide data for 3373 individuals across four breeding populations. *New Phytol*. 221: 818–833.
Munafò, M.R. and Flint, J. (2004) Meta-analysis of genetic association studies. *Trends Genet*. 20: 439–444.
Mural, R.V., Grzybowski, M., Miao, C., Damke, A., Sapkota, S., Boyles, R.E., et al. (2021) Meta-analysis identifies pleiotropic loci controlling phenotypic trade-offs in sorghum. *Genetics* 218: iyab087.
Myles, S., Peiffer, J., Brown, P.J., Ersoz, E.S., Zhang, Z., Costich, D.E., et al. (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21: 2194–2202.
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.
Pearson, K. (1901) LIII. On lines and planes of closest fit to systems of points in space. *London Edinburgh Dublin Philos. Mag. J. Sci*. 2: 559–572.
Petretto, E., Bottolo, L., Langley, S.R., Heinig, M., McDermott-Roe, C., Sarwar, R., et al. (2010) New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Comput. Biol*. 6: e1000737.
Pirinen, M., Donnelly, P. and Spencer, C.C.A. (2012) Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat. Genet*. 44: 848–851.
Platt, A., Vilhjálmsson, B.J. and Nordborg, M. (2010) Conditions under which genome-wide association studies will be positively misleading. *Genetics* 186: 1045–1052.
Porter, H.F. and O'Reilly, P.F. (2017) Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Sci. Rep*. 7: 38837.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. 81: 559–575.
Rees, J.S., Castellano, S. and Andrés, A.M. (2020) The genomics of human local adaptation. *Trends Genet*. 36: 415–428.
Rice, B.R., Fernandes, S.B. and Lipka, A.E. (2020) Multi-trait genome-wide association studies reveal loci associated with maize inflorescence and leaf architecture. *Plant Cell Physiol*. 61: 1427–1437.
Ringnér, M. (2008) What is principal component analysis? *Nat. Biotechnol*. 26: 303–304.
Sasaki, E., Gunis, J., Reichardt-Gomez, I., Nizhynska, V. and Nordborg, M. (2022) Conditional GWAS of non-CG transposon methylation in *Arabidopsis thaliana* reveals major polymorphisms in five genes. *PLoS Genet*. 18: e1010345.
Sasaki, E., Kawakatsu, T., Ecker, J.R. and Nordborg, M. (2019) Common alleles of *CMT2* and *NRPE1* are major determinants of CHH methylation variation in *Arabidopsis thaliana*. *PLoS Genet*. 15: e1008492.
Sasaki, E., Köcher, T., Filiault, D.L. and Nordborg, M. (2021) Revisiting a GWAS peak in *Arabidopsis thaliana* reveals possible confounding by genetic heterogeneity. *Heredity* 127: 245–252.
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.
Schaid, D.J., Chen, W. and Larson, N.B. (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet*. 19: 491–504.
Schielzeth, H. (2010) Simple means to improve the interpretability of regression coefficients. *Methods Ecol. Evol*. 1: 103–113.
Schmitz, R.J., He, Y., Valdes-Lopez, O., Khan, S.M., Joshi, T., Urich, M.A., et al. (2013) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res*. 23: 1663–1674.
Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., et al. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet*. 44: 825–830.
Sella, G. and Barton, N.H. (2019) Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annu. Rev. Genomics Hum. Genet*. 20: 461–493.
Shindo, C., Aranzana, M.J., Lister, C., Baxter, C., Nicholls, C., Nordborg, M., et al. (2005) Role of *FRIGIDA* and *FLOWERING LOCUS C* in determining variation in flowering time of Arabidopsis. *Plant Physiol*. 138: 1163–1173.
Spain, S.L. and Barrett, J.C. (2015) Strategies for fine-mapping complex traits. *Hum. Mol. Genet*. 24: R111–9.
Stearns, F.W. (2010) One hundred years of pleiotropy: a retrospective. *Genetics* 186: 767–773.
Stephens, M. (2013) A unified framework for association analysis with multiple related phenotypes. *PLoS One* 8: e65245.
Sul, J.H., Martin, L.S. and Eskin, E. (2018) Population structure in genetic studies: confounding factors and mixed models. *PLoS Genet*. 14: e1007309.
Thoen, M.P.M., Davila Olivas, N.H., Kloth, K.J., Coolen, S., Huang, -P.-P., Aarts, M.G.M., et al. (2017) Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. *New Phytol*. 213: 1346–1362.
Tibbs Cortes, L., Zhang, Z. and Yu, J. (2021) Status and prospects of genome-wide association studies in plants. *Plant Genome* 14: e20077.
Todesco, M., Balasubramanian, S., Hu, T.T., Traw, M.B., Horton, M., Epple, P., et al. (2010) Natural allelic variation underlying a major fitness trade-off in *Arabidopsis thaliana*. *Nature* 465: 632–636.
Todesco, M., Owens, G.L., Bercovich, N., Légaré, J.S., Soudi, S., Burge, D. O., et al. (2020) Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*. 584: 602–607.
Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., et al. (2021) Genome-wide association studies. *Nat. Rev. Methods Primers*. 1: 1–21.
Urbut, S.M., Wang, G., Carbonetto, P. and Stephens, M. (2019) Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet*. 51: 187–195.
Valladares, F., Sanchez-Gomez, D. and Zavala, M.A. (2006) Quantitative estimation of phenotypic plasticity: bridging the gap between the evolutionary concept and its ecological applications. *J. Ecol*. 94: 1103–1116.
Vilhjálmsson, B.J. and Nordborg, M. (2013) The nature of confounding in genome-wide association studies. *Nat. Rev. Genet*. 14: 1–2.
Visscher, P.M., Macgregor, S., Benyamin, B., Zhu, G., Gordon, S., Medland, S., et al. (2007) Genome partitioning of genetic variation for height from 11,214 sibling pairs. *Am. J. Hum. Genet*. 81: 1104–1110.
Wang, Q., Tian, F., Pan, Y., Buckler, E.S. and Zhang, Z. (2014) A SUPER powerful method for genome wide association study. *PLoS One* 9: e107684.
Weller, J.I., Wiggans, G.R., Vanraden, P.M. and Ron, M. (1996) Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. *Theor. Appl. Genet*. 92: 998–1002.

Xiao, Y., Tong, H., Yang, X., Xu, S., Pan, Q., Qiao, F., et al. (2016) Genome-wide dissection of the maize ear genetic architecture using multiple populations. *New Phytol.* 210: 1095–1106.

Xu, S. (2003) Theoretical basis of the beavis effect. *Genetics* 165: 2259–2268.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42: 565–569.

Yang, J., Ferreira, T., Morris, A.P., Medland, S.E. Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* 44: 369–75, S1–3.

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., et al. (2011) Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* 43: 519–525.

Yang, J., Zeng, J., Goddard, M.E., Wray, N.R. and Visscher, P.M. (2017) Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* 49: 1304–1310.

Yano, K., Morinaka, Y., Wang, F., Huang, P., Takehara, S., Hirai, T., et al. (2019) GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proc. Natl. Acad. Sci. U. S. A.* 116: 21262–21267.

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

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

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

Yu, S.B., Li, J.X., Xu, C.G., Tan, Y.F., Li, X.H. and Zhang, Q. (2002) Identification of quantitative trait loci and epistatic interactions for plant height and heading date in rice. *Theor. Appl. Genet.* 104: 619–625.

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

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., et al. (2018) Rewiring of the fruit metabolome in tomato breeding. *Cell* 172: 249–261.e12.

Plant Cell Physiol. 00(00): 1–13 (2024) doi:https://doi.org/10.1093/pcp/pcae079, Advance Access publication on 11 July 2024, available online at https://academic.oup.com/pcp