Chapter 5

Methods and Tools in Genome-wide Association Studies

Anja C. Gumpinger, Damian Roqueiro, Dominik G. Grimm, and Karsten M. Borgwardt

Abstract

Many traits, such as height, the response to a given drug, or the susceptibility to certain diseases are presumably co-determined by genetics. Especially in the field of medicine, it is of major interest to identify genetic aberrations that alter an individual's risk to develop a certain phenotypic trait. Addressing this question requires the availability of comprehensive, high-quality genetic datasets. The technological advancements and the decreasing cost of genotyping in the last decade led to an increase in such datasets. Parallel to and in line with this technological progress, an analysis framework under the name of *genome-wide association studies* was developed to properly collect and analyze these data. Genome-wide association studies aim at finding statistical dependencies—or associations—between a trait of interest and point-mutations in the DNA. The statistical models used to detect such associations are diverse, spanning the whole range from the frequentist to the Bayesian setting.

Since genetic datasets are inherently high-dimensional, the search for associations poses not only a statistical but also a computational challenge. As a result, a variety of toolboxes and software packages have been developed, each implementing different statistical methods while using various optimizations and mathematical techniques to enhance the computations.

This chapter is devoted to the discussion of widely used methods and tools in genome-wide association studies. We present the different statistical models and the assumptions on which they are based, explain peculiarities of the data that have to be accounted for and, most importantly, introduce commonly used tools and software packages for the different tasks in a genome-wide association study, complemented with examples for their application.

Key words Genome-wide association studies, Missing heritability, Linkage disequilibrium, Phenotypes, Univariate mapping, Population structure correction, Genomic inflation, Multilocus mapping, Multiple hypothesis correction, Meta-analysis, GWAS tools

1 Introduction

1.1 Genome-wide Association Studies: An Overview

Genome-wide association studies (GWAS) have become a valuable tool to identify associations between genetic variants in a group of individuals and a phenotype present in these individuals. The phenotype in question can be a trait such as height, the presence or absence of a disease, the response to a drug treatment, or any

Louise von Stechow and Alberto Santos Delgado (eds.), *Computational Cell Biology: Methods and Protocols*, Methods in Molecular Biology, vol. 1819, https://doi.org/10.1007/978-1-4939-8618-7_5,

other phenotype of interest. The genetic variants used in GWAS are primarily single-nucleotide polymorphisms (SNPs), which correspond to single base-pairs in the DNA that are known to vary between individuals. The goal of a genome-wide association study is to determine which SNPs are associated with the phenotype in a statistically significant manner.

Historically, and prior to the emergence of GWAS, *linkage mapping* was a technique used to detect genetic markers that segregated within families affected by rare diseases. Linkage mapping was successful for rare and Mendelian diseases, for example, at identifying loci associated to Huntington's disease [1] and at detecting mutations in the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene in patients with cystic fibrosis [2]. Nevertheless, for complex diseases such as type II diabetes and schizophrenia, it is the cumulative effect of dozens or hundreds of variants throughout the entire genome—each with a small effect on the phenotype—that confer a greater risk of developing the disease. This is the reason why linkage mapping was less successful in the realm of complex diseases [3, 4], which in turn allowed GWAS to rise to prominence as a tool to identify associations between a trait and genetic variants of smaller effects. There are numerous reasons why GWAS, in contrast to linkage mapping, are better equipped to detect these associations: (a) in GWAS, hundreds of thousands to millions of SNPs are surveyed, (b) GWAS do not necessarily rely on pedigree information, and (c) GWAS have larger sample sizes than linkage mapping studies and, thus, have more power to detect associations [5].

Whether one intends to perform GWAS in plants to increase crop yields, or in livestock to identify genes associated with economically important traits such as fertility, or in humans to find SNPs associated with common diseases, this chapter serves as a guide to all the theoretical aspects of GWAS. The chapter also contains detailed protocols on how to conduct GWAS with different tools and how to overcome potential pitfalls in the analysis.

1.2 Performing GWAS

The haploid human genome comprises approximately three billion base-pairs, 3% of which show variation among individuals (estimate based on the 84.7 million reported SNPs by the 1000 Genomes project [6]). These base-pairs that vary across a population represent the single-nucleotide polymorphisms mentioned in the previous section. Endeavors such as the original HapMap project [7], and the more recent 1000 Genomes project [6] have aimed at genotyping a multitude of human genomes to detect and annotate genetic variation among individuals.

For performing a genome-wide association study, we assume that two types of data are readily available for all individuals in the study: their phenotype and genotypes. The latter can be obtained

through state-of-the-art sequencing technologies [8], or through a genotyping array [9].

In a traditional (univariate) genome-wide association study, a measure of association or statistical dependence between each individual SNP and the phenotype is computed. Then a p-value is derived for each association score, which represents the probability of observing an association signal of the same strength or stronger under the null hypothesis of no association between the SNP and the phenotype. If the p-value falls below a predefined significance threshold α, commonly 0.01 or 0.05, the null hypothesis is rejected, which means that there is an association between the SNP and the phenotype. Despite the strong evidence against the null hypothesis in this case, there remains a chance of $\alpha^*100\%$ that the low p-value is purely due to random chance and that the detected association is therefore a false positive result. Avoiding false positive findings is among the major challenges in GWAS.

1.3 Challenges in GWAS

1.3.1 Avoiding False Positive Findings

In GWAS, typically hundreds of thousands to millions of SNPs are tested simultaneously for association with the phenotype. Since each of these hypotheses are rejected at a significance level α—typically, $\alpha = 0.05$—this multiplicity of tests can lead to the reporting of spurious associations if no correction for multiple hypothesis testing is performed. There is rich statistical literature on various methods to correct for multiple hypothesis testing [10] and the most prevalent ones, namely **family wise error rate** and **false discovery rate**, are discussed in detail later in this chapter.

Another possible source of false positive findings is the presence of confounders, such as environmental factors, population or family structure, cryptic relatedness, age, and gender. Autoimmune diseases in humans are a good example of how gender can be a confounder as approximately 80% of patients affected by an autoimmune disease are women [11]. Population structure, on the other hand, refers to differences in allele frequencies between groups of individuals in a study due to systematic ancestry differences [12], such as geographic proximity or individuals sharing the same ethnicity. Not correcting for these confounding factors may lead to false positive findings: SNPs that seem to be associated to the phenotype while they are actually associated with the confounding factor, e.g., with population structure.

Similar to the importance of avoiding false positive findings in GWAS, it is equally crucial to avoid missing true associations, the so-called false negatives. False negatives occur when the statistical signal of the marker is not strong enough to reach genome-wide significance. Possible reasons for this are (a) little evidence in the data to support the statistical association, e.g., because of a small sample size, or (b) the significance threshold being too low. In general, there is a trade-off between the number of false positives and false negatives and we defer this discussion

to Subheading 2.7.1 when we address the topic of multiple-hypothesis testing correction.

1.3.2 Missing Heritability

Over the last decade, GWAS have been successfully performed on different organisms, such as *Arabidopsis thaliana* [13–15], rice [16], fruit flies [17], mice [18], and humans [19–21]. In humans, particularly in the field of autoimmune and metabolic diseases, GWAS have revealed important insights into the genetic mechanisms of disease development and progression [4]. Despite these and many other successes, the association signals detected in univariate GWAS often explain only a small fraction of the total phenotypic variability. This phenomenon has been referred to as **missing heritability** [4, 22–24].

In the literature, different strategies have been proposed to discover the missing heritability. One class of approaches aims at changing the hypothesis underlying GWAS by considering the joint effects of multiple SNPs [25]. Examples covered in this chapter are (a) the search for nonlinear SNP-SNP interactions [26–28], also known as epistasis, (b) the joint analysis of SNPs overlapping with genes [29–32], and (c) the interaction of SNPs when superimposed on a biological network [33–36].

Another class of approaches tries to alleviate the burden of multiple hypothesis testing by attempting to increase the per-hypothesis significance threshold while decreasing the number of hypothesis to be tested. Examples of these approaches that are also covered in this chapter are (a) gene-based approaches [29–32] as well as (b) methods analyzing intervals of genetic markers [37], or (c) clustering with subsequent hierarchical testing [38].

1.4 Outline of the Chapter

This chapter provides both a theoretical and practical guide to GWAS. In Subheading 2, *Methods and Definitions*, we summarize important theoretical concepts of GWAS. We present formal definitions and discuss different types of GWAS as well as the evaluation of their results. In Subheading 3, *Tools and Software*, we introduce different software tools and packages that are commonly used to perform GWAS. The presentation of each tool is accompanied by a tutorial on how to use it. We conclude the chapter by reviewing the development of GWAS over the last decade, and highlighting current challenges and future directions of research.

In addition to the contents of this chapter, we provide a virtual machine (VM) with preinstalled tools and all the necessary scripts to run the examples in Subheading 3. The blue boxes in Subheading 3 provide step-by-step protocols on how to run the scripts on the VM. The VM and sample scripts can be accessed online at: https://www.bsse.ethz.ch/mlcb/gwas. The link also provides additional details on how to install the VM and how to use the sample scripts.

2 Methods and Definitions

Before delving into the methodological aspects of GWAS, we first introduce the concept of linkage disequilibrium followed by a discussion on how genomic data are frequently encoded and preprocessed. We then cover the main statistical models for univariate association testing, i.e., when each SNP is analyzed separately. We proceed to complement the ideas presented in univariate methods by discussing more sophisticated scenarios in which methods that rely on gene-based analyses and interactions between genetic markers are considered. We conclude this theoretical section by providing details on how the results of GWAS can be evaluated and combined.

2.1 Linkage Disequilibrium

When working with genetic data, a concept of utmost importance that needs to be taken into consideration is that of **linkage disequilibrium** (LD). For a given population, two SNPs are said to be in LD when their genotypes are correlated in a way that will not arise by chance. From a biological perspective, LD between loci is determined by local recombination rates. As a result of this, older populations that went through a higher number of recombination events show different LD patterns than younger populations do. In humans, for example, LD patterns vary between different ethnic groups [7]. One of the key technological achievements that greatly facilitated GWAS was the development of SNP arrays [7, 39]. SNP arrays do not survey the entirety of SNPs in a genome, but instead rely on information about regions of SNPs in high LD. From each of these regions, a tag SNP is selected and ultimately genotyped under the assumption that, if the tag SNP is associated to the phenotype, then the tag SNP is in high LD with the presumably causal SNP [3].

2.2 Data Encoding and Preprocessing Steps

2.2.1 Encoding of Genotypes and Phenotypes

Genotype data can have different formats, depending on the genotyping platform that was used. In what is normally considered to be *raw* data, a SNP for a given individual is represented by its two alleles (for diploid cells) in the letters A, C, G, or T (*see* Fig. 1a) for each of the four nucleotides. Many of the statistical models described in this section require that the raw allele information be encoded as a numerical value. This is referred to as the **encoding** of the genotype. There exist different models in which genotypes can be encoded and each of them reflects prior assumptions about the underlying mode of inheritance, such as the additive, the dominant and the recessive encoding (*see* Fig. 1b). Of these models, the most commonly used one is the additive model, in which each SNP is represented by the number of its minor alleles, i.e., 0, 1, or 2 [40].

The phenotype can be either qualitative or quantitative. The former refers to phenotypes that have discrete class labels for each

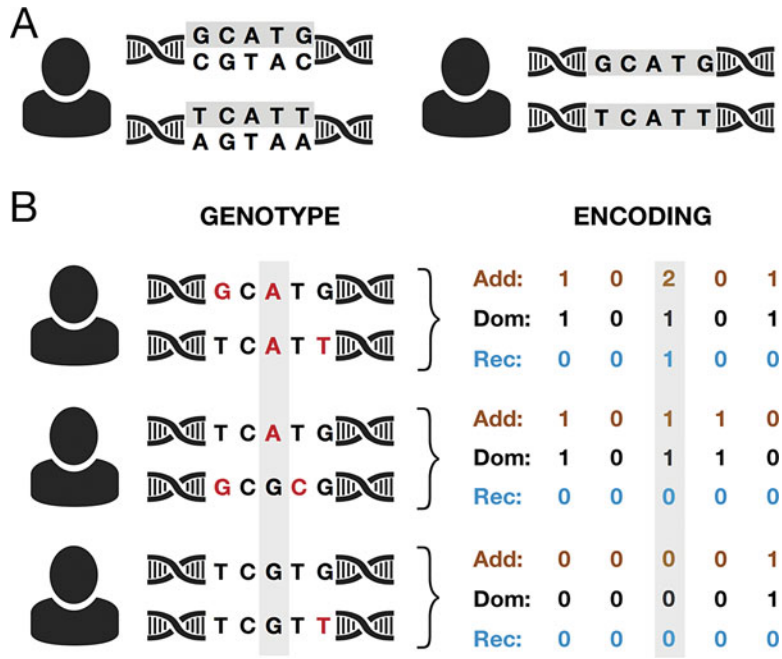

Fig. 1 Data representation in GWAS for a diploid individual. (**a**) For an individual, the genotype is represented by the alleles on two strands of DNA, one from each chromosome. (**b**) Three different encoding schemes (additive, dominant, and recessive) for SNP data. The alleles highlighted in red are the minor alleles. The additive encoding counts the number of minor alleles. The dominant encoding maps the homozygous major alleles to 0 and all other genotypes to 1. The recessive encoding maps the homozygous minor allele to 1, and all other genotypes to 0

individual, for example case/control. Quantitative phenotypes, on the other hand, are represented by a real number. Typical examples of this type of phenotype are height, body-mass index, or blood pressure [41]. The phenotype being qualitative or quantitative is an important criterion for the choice of the statistical method in a genome-wide association study.

2.2.2 Data Preprocessing and Quality Control

Before we can start testing for associations between SNPs and the phenotype of interest, certain preprocessing tasks are commonly applied to minimize the chances of reporting false results.

Transformation of Phenotypes

Some statistical methods make a particular assumption about the distribution of the phenotype and its noise. For example, when applying linear regression in GWAS, one of the assumptions of the method is that, given the genotype, the phenotype follows a Gaussian distribution. In practice, however, this assumption rarely holds. Therefore, a common preprocessing task is to transform or normalize the phenotypic data such that it follows the expected distribution of the statistical model.

Filtering Using Hardy–Weinberg Equilibrium

The Hardy–Weinberg equilibrium (HWE) is a model that allows for the prediction of genotype frequencies from one generation to the next. SNPs in GWAS are expected to be in HWE and any deviation can be assessed through a statistical test with a common threshold on the p-value of 1e−06 [42] (this threshold is commonly used in humans, for other organisms the consensus threshold may be different). If an SNP is found not to be in HWE, this is normally because of a sampling or a pure genotyping error. When this occurs, the SNP is removed from the analysis in a preprocessing step.

Filtering by Minor Allele Frequency

The minor allele frequency (MAF) is the frequency of the less common allele at a biallelic locus [3]. SNPs that have low MAF with values smaller than 0.05 or 0.01, the so-called rare variants, are commonly excluded from standard GWAS [5]. Unless sample sizes are very large or the effects of the rare variants are high, standard GWAS techniques are underpowered to detect associations with rare variants [5, 43]. There is an entire class of methods collectively known as burden tests that are well-suited for rare variants [44–46]. In addition to burden tests, there are other methods such as the C-alpha test [47] or SKAT [43] which have proven to be effective at analyzing rare variants. This chapter focuses on methods that utilize common variants and the reader is referred to refs. [43–47] for more details about the analysis of rare variants.

Filtering by Missing Genotypes

Another important quality control step that needs to be performed before conducting GWAS is to exclude from the analysis: (a) individuals with a large number of missing genotypes, and (b) SNPs with a high rate of missing genotypes across all individuals [42]. The first case is normally a consequence of poor DNA quality or low DNA concentration. This affects the accuracy of SNP calling algorithms, which then report large numbers of missing SNPs for the individual. In the second case, when a SNP has a low call rate, i.e., a high missing rate, it is considered of low quality and is excluded from the analysis to avoid false positives [48]. There are well-defined protocols to impute the values of missing SNPs, and we refer the interested reader to [49] for more details.

2.3 Concepts in Univariate GWAS

After quality control and preprocessing we are in a position to start the analysis of the data. We herein describe the methods that are most commonly used to test SNPs for association to the phenotype. Since all SNPs are tested independently, these methods belong to the class of single-locus or univariate GWAS [3].

2.3.1 Two Sample Tests

When analyzing a qualitative phenotype, as it is done in a case/control study, a common way to test an SNP for its

association to the phenotype is to create a contingency table by counting the allele frequencies in cases and controls. This contingency table can then be used to test for association using a discrete test statistic, such as Fisher's exact test [50] or a χ^2 test [51].

2.3.2 Linear Models

Linear models are often used to test a single SNP for its association with a phenotype [42]. The underlying assumption in linear models is that the phenotype can be modeled as an additive (linear) combination of the genotype values of the SNP and (possibly) covariates, such as age or gender, each of them with certain effect sizes. Moreover, linear models contain a residual term that captures noise in the data. Let us consider a dataset with n individuals, where for each individual we have: (a) a phenotype value, (b) a genotype value for a given SNP, and (c) p covariates. Then, a linear model can be described as follows:

$$\boldsymbol{y} = \beta_0 \overrightarrow{1} + \beta_1 \boldsymbol{x}_\mathbf{G} + \boldsymbol{X}_\mathbf{C} \boldsymbol{\beta}_2^T + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y}$ represents the $n \times 1$ vector of phenotypes, $\overrightarrow{1}$ is an $n \times 1$ vector of ones, $\boldsymbol{x}_\mathbf{G}$ is the $n \times 1$ vector of the genotypes for a given SNP, $\boldsymbol{X}_\mathbf{C}$ is the $n \times p$ matrix of covariates and $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}$, $\boldsymbol{\beta}_2 \in \mathbb{R}^{1 \times p}$ are the offset, the genotype and the covariate effects, respectively. Additionally, $\epsilon \in \mathbb{R}^{n \times 1}$ is the vector of residuals, which are assumed to follow a known probability distribution, allowing the model parameters β_0, β_1, and $\boldsymbol{\beta}_2$ to be estimated using probabilistic techniques such as maximum likelihood estimation (MLE) [52]. The inclusion of the covariates $\boldsymbol{X}_\mathbf{C}$ into the model accounts for known factors that might have an influence on the phenotype, such as age or gender. It is common to reformulate Eq. 1 as:

$$\boldsymbol{y} = \boldsymbol{X} \boldsymbol{\beta}^T + \boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{X} = \left[\overrightarrow{1}, \boldsymbol{x}_\mathbf{G}, \boldsymbol{X}_\mathbf{C}\right] \in \mathbb{R}^{n \times (2+p)}$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \boldsymbol{\beta}_2] \in \mathbb{R}^{1 \times (2+p)}$. The matrix $\boldsymbol{X}$ is referred to as the design matrix. When applying linear models to test SNPs for their association with the phenotype, an individual linear model has to be fitted for each SNP. Depending on the number of SNPs, this can be computationally expensive.

Linear Regression

A linear regression model assumes that the residuals, and therefore also the phenotype $\boldsymbol{y}$, are normally distributed given the genotype $\boldsymbol{x}_\mathbf{G}$, making it applicable to quantitative phenotypes. Using methods such as MLE, the parameters β_0, β_1, and $\boldsymbol{\beta}_2$ in the model Eq. 1 can be estimated from the data. Once the model is fitted, the parameter β_1 describes the effect the SNP has on the phenotype. The parameter's deviation from zero is an indicator of the effect

size: the larger its absolute value is, the higher the contribution of $\boldsymbol{x}_{\mathbf{G}}$ to the phenotype is. With a statistical test, it is possible to assess if the deviation from zero is statistically significant at a predefined significance level.

Logistic Regression

Logistic regression is a special case of linear models when the phenotype is qualitative, as in a case/control study [52].

While in linear regression the phenotype is modeled through a linear combination of genetic and covariate effects, in logistic regression the logarithm of the odds is modeled as:

$$\log\left(\frac{\boldsymbol{P}\left(\boldsymbol{y}=1\right)}{1-\boldsymbol{P}\left(\boldsymbol{y}=1\right)}\right)=\beta_0\overrightarrow{1}+\beta_1\boldsymbol{x}_{\mathbf{G}}+\boldsymbol{X}_{\mathbf{C}}\boldsymbol{\beta}_2^{\boldsymbol{T}} \tag{3}$$

where all parameters are defined as in Eq. 1. Logistic regression models are commonly used to model the probability of a target variable to take a specific value (in the case of GWAS, the phenotype). For a more extensive coverage on logistic regression, the reader is referred to [52].

Linear Mixed Models

A further variation of linear regression is the linear mixed model (LMM). Over the last few years LMMs have gained popularity in the field of GWAS [53]. They constitute a flexible framework for the analysis of genetic data, with a linear combination of fixed and random effects accounting for the phenotypic variation. While in linear regression the parameters of the model are fixed, in LMMs some parameters are assumed to follow a Gaussian distribution, the so called random effects. In many applications, the genetic variant and the covariates are modeled as fixed effects, while the genetic similarity among samples is modeled as a random effect, adding an additional term to the model in Eq. 2:

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}^{\boldsymbol{T}}+\boldsymbol{\gamma}+\epsilon \tag{4}$$

where $\boldsymbol{\gamma}=\boldsymbol{X}_{\mathbf{S}}\boldsymbol{\beta}_3^{\boldsymbol{T}}$ is called a random effect, with each parameter in the vector $\boldsymbol{\beta}_3\in\mathbb{R}^{1\times w}$ drawn from the same Gaussian distribution, and $\boldsymbol{X}_{\mathbf{S}}$ is the $n\times w$ design matrix of the random effect. This is equivalent to $\boldsymbol{\gamma}$ being drawn from a multivariate Gaussian distribution with a covariance matrix proportional to $\boldsymbol{K}=\boldsymbol{X}_{\mathbf{S}}\boldsymbol{X}_{\mathbf{S}}^{\boldsymbol{T}}\in\mathbb{R}^{n\times n}$ [54]. As with linear regression, this results in a phenotype with a Gaussian distribution and therefore in a model applicable to quantitative traits.

Bayesian Mixed Models

From a Bayesian point of view, the Gaussian distribution of $\boldsymbol{\beta}_3$ in the LMM of Eq. 4 can be interpreted as a prior on the parameter vector $\boldsymbol{\beta}_3$. This prior can be replaced by other distributions, thereby incorporating different prior assumptions into the model and giving rise to Bayesian mixed models. The choice of the

prior distribution has implications on the estimation of the model parameters. For example in BOLT-LMM [54], SNP effects are modeled as random effects, and the Gaussian prior on $\boldsymbol{\beta}_3$ in Eq. 4 is replaced with a mixture of two Gaussians, keeping the parameter estimation simple due to the convenient mathematical representation of the Gaussian.

2.4 Population Structure Correction

As mentioned in the introduction, the presence of population structure in the data can lead to false positive associations. A common way to assess the degree of population structure in the data is to compute the genomic inflation factor λ_{GC} [55]. It describes the deviation of the median of the observed test statistics from the median of the expected test statistics. Since the distribution of the test statistics under the null hypothesis is known (either because there is a closed form representation, or because it has been derived empirically using permutation testing), so is its median. Considering the assumption that most of the SNPs are not associated with the trait, λ_{GC} should be close to one if no confounder is present [56]. If λ_{GC} is inflated (larger than 1) or deflated (lower than 1), a correction for population structure is recommended to avoid false positives or false negatives, respectively. The inflation or deflation of *p*-values can be easily visualized in a Q-Q plot (refer to Subheading 2.7.2 for more details about these plots). The three most common correction approaches are based on (a) the principal components (PCs) of a genetic similarity matrix [12], (b) the genomic inflation factor λ_{GC} [57], and (c) a combination of an LMM with the genetic similarity matrix as random effect [58, 59].

2.4.1 Principal Components of the Genetic Similarity Matrix

For a genetic dataset with n individuals, the **genetic similarity matrix** corresponds to the $n \times n$ matrix that captures the similarity between the individuals based on their SNP information (genotypes). To compute this matrix, assume $\boldsymbol{X}_{\mathbf{G}}$ to be the $n \times d$ matrix containing d SNPs for n individuals. Prior to the computation of the similarity matrix, each SNP in $\boldsymbol{X}_{\mathbf{G}}$ is commonly centered around its mean and normalized by its allele-frequencies [12], resulting in the matrix $\widehat{\boldsymbol{X}}_{\mathbf{G}}$. Then, the similarity matrix is computed as:

$$\boldsymbol{K} = \frac{1}{n}\widehat{\boldsymbol{X}}_{\mathbf{G}}\widehat{\boldsymbol{X}}_{\mathbf{G}}^{\boldsymbol{T}} \in \mathbb{R}^{n \times n} \tag{5}$$

It has been shown that including the leading PCs of the genetic similarity matrix $\boldsymbol{K}$ as covariates in a linear model corrects for population structure [12]. Nevertheless, it is not a priori clear how many PCs should be included as covariates. Commonly, the association analysis is repeated multiple times, with an increasing number of leading PCs included as covariates. For each run, the

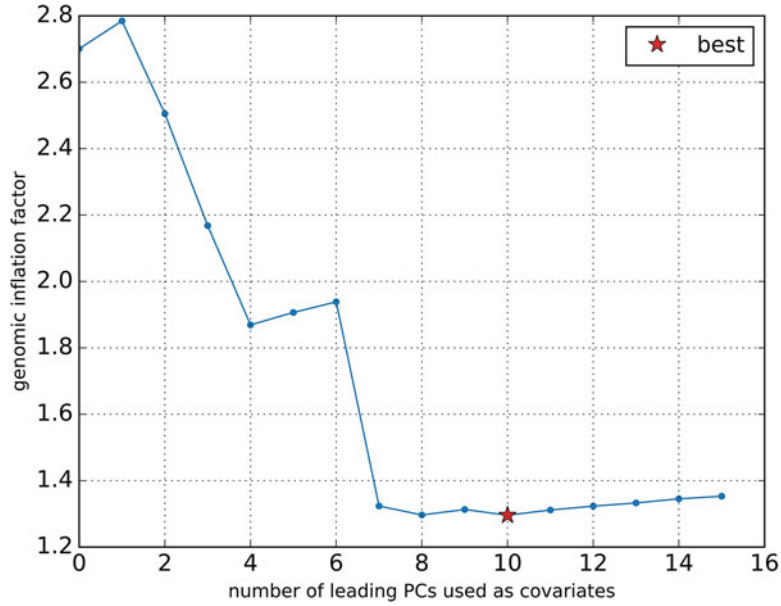

Fig. 2 The genomic inflation factor, when different numbers of leading PCs are included as covariates into a linear regression model

genomic inflation is computed and plotted against the number of PCs in the model (*see* Fig. 2). Finally, the number of leading PCs that are used is the one that yields a genomic inflation factor λ_{GC} close to 1.

2.4.2 Adjusting Test Statistics for Genomic Inflation

Devlin et al. [55, 57] have suggested to adjust the raw test statistics with the genomic inflation factor λ_{GC}. Dividing each test statistic by λ_{GC} and subsequently computing the p-value from the adjusted test statistic reduces the genomic inflation:

$$T_{\text{adjusted}} = \frac{T_{\text{raw}}}{\lambda_{GC}}.$$

2.4.3 Correction for Population Structure in LMMs

LMMs have been successful at correcting for population structure, family relatedness and cryptic relatedness in GWAS [53]. The correction is achieved by explicitly accounting for structure among individuals in the model, e.g., by including a random effect with covariance proportional to the genetic similarity matrix defined in Eq. 5. Depending on the type of relatedness between individuals, coupling more than one genetic similarity matrix, and if necessary, additionally including PCs as fixed effects into the model, can increase the power of GWAS [59]. An often-encountered downside of LMMs for GWAS is the computationally demanding task of obtaining the genetic similarity matrix. This problem has been addressed in different approaches, such as [58, 60–62], which are based on approximations of the matrix, or on computationally efficient exact methods.

2.5 Gene-Based Approaches

In contrast to the univariate methods previously described, gene-based approaches aim at deriving p-values of association, not for single SNPs but for genes. Gene-based approaches rely on first mapping SNPs to genes, followed by specific methods to test the resulting genes for association. Both of these topics are discussed below.

2.5.1 Mapping SNPs to Genes

In gene-based GWAS, a gene is represented by the set of SNPs that overlaps with it. Moreover, it is a common practice to also include SNPs in close proximity to the gene (between 20 and 50 kb, upstream and downstream) [25, 63, 64]. These SNPs are theorized to have regulatory effects on gene expression, thereby affecting the phenotype. Nevertheless, special care has to be taken when expanding the width of this margin as it increases the chance of assigning the same SNP to multiple genes. This results in a violation of the assumption that genes are independent, and might lead to inflated association signals [65].

2.5.2 Association Testing of Genes

There exist two main approaches for association testing of a gene: (a) two-step approaches, which consist of computing univariate test statistics for each SNP in the gene and, subsequently, collapsing them into one test statistic for the whole gene, and (b) one-step approaches, in which all SNPs in the gene are used simultaneously to derive a test statistic [64, 66].

Two-Step Methods

The first step consists of computing the p-values, $p_{g,1}, \ldots, p_{g,m}$, of the m SNPs that overlap with the gene g, derived from the univariate test statistics $s_{g,1}, \ldots, s_{g,m}$. In the second step, the test-statistics are combined using one of the following approaches:

(a) Minimum p-value [32, 64]: The gene p-value is set to the minimum of the SNP p-values overlapping with the gene, i.e., $p_g = \min(p_{g,1}, \ldots, p_{g,m})$.

(b) Sum of test statistic [29, 32]: If the test statistics $s_{g,1}, \ldots, s_{g,m}$ come from a χ^2 test with one degree of freedom, the gene test statistic s_g can be computed as $s_g = \sum_{i=1}^{m} s_{g,i}$, which follows a χ^2-distribution with m degrees of freedom, assuming independence between the statistics. In reality, this assumption of independence does not hold due to LD, which requires the development of methods taking LD into account.

(c) Average test statistic [29, 67]: Computation of the average over the k most significant test statistics, i.e., the gene test statistic corresponds to $s_g = \frac{1}{k} \sum_{i=1}^{k} s_{g,i}$. Since the null-distribution of this statistic cannot be analytically derived, a p-value of association is obtained by performing permutation testing.

One-Step Methods

Alternatively, all the SNPs overlapping with a gene can be tested simultaneously in one test by using the linear mixed model framework. For this purpose, the model in Eq. 1 has to be adapted such that the genotype vector $\boldsymbol{x}_{\mathbf{G}} \in \mathbb{R}^{n\times 1}$ is replaced by a genotype matrix $\boldsymbol{X}_{\mathbf{G}} \in \mathbb{R}^{n\times m}$, and the m columns in the matrix correspond to the genotypes of the m SNPs mapped to the gene [68].

Both one- and two-step approaches have their respective advantages. The one-step approaches, in contrast to the two-step approaches, make no prior assumptions about the direction of effects of the SNPs harbored in the gene. An advantage of two-step approaches is that they do not require access to the original genotype data, but it is sufficient to use the summary statistics of a univariate analysis for each SNP. We refer the reader to [64] for a more comprehensive discussion about these methods.

2.6 Detection of Interactions Between Genetic Loci

As mentioned in Subheading 1.3.2, missing heritability refers to the gap between the heritability of a trait and the phenotypic variation that can be accounted for by association signals from univariate GWAS. One of the possible sources of missing heritability is hypothesized to be the fact that univariate GWAS do not account for nonlinear interactions between loci [69]. Thus, extensions that take into account interactions between SNPs or genes have been proposed to complement the existing univariate approaches. Below we discuss two of these extensions.

2.6.1 Epistasis

There exist various definitions of epistasis, with the most commonly used one referring to the deviation from additivity in a linear model [26]. This can be captured in a statistical model as follows:

$$\boldsymbol{y} = \beta_0 \overrightarrow{1} + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \beta_{12} \boldsymbol{x}_1 \boldsymbol{x}_2 + \epsilon \tag{6}$$

In model Eq. 6, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ correspond to the genotype at SNP_1 and SNP_2, respectively. The multiplicative term $\boldsymbol{x}_1\boldsymbol{x}_2$ is the element-wise product of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and corresponds to the epistatic interaction between the two loci. The parameter β_{12} is the effect size that will be tested for deviation from zero to obtain a p-value of the epistatic interaction. The statistical test is analogous to the one in the linear models explained earlier. For a more comprehensive review of different test statistics used in the analysis of genome-wide interactions, the reader is referred to [70].

2.6.2 Network-Based Approaches

To complement the analysis of epistasis, one can explore higher-order interactions between SNPs in the hope that a group of them can be found to act in concert to disrupt a certain biological process. Testing all possible combinations of SNPs at a genome-wide level for association is computationally infeasible. Therefore, the class of methods known as **network-based GWAS** is frequently used to minimize the search space by incorporating

prior knowledge in the form of biological networks and testing only interactions between markers in these networks [25].

Commonly used networks are for example protein–protein interaction networks [71–74] in which a node corresponds to a gene (or its product, a protein) and an edge represents any type of interaction between the nodes at both ends. The interactions can be derived from different sources, which renders some interactions as having higher confidence than others. While some interactions are derived from repositories of experiments, others are predicted or manually curated from the literature [72, 73]. In general, these interactions are not context specific and hold for a variety of tissues and cell types. Exploring and mining these networks with respect to a phenotype is the goal of network-based GWAS. Most methods that were developed to include network information aim at finding dense subgraphs (called modules) within the network that are associated to the phenotype [33–36]. Since these networks are based on genes/proteins, a mapping between genes/proteins and SNPs needs to take place before testing for association.

2.7 Evaluation of GWAS

Once a genome-wide association study has been performed, there are additional post-processing steps that are normally performed on the association signals. These steps include, but are not limited to: (a) correcting p-values for multiple hypothesis testing, (b) visualizing the results, (c) (potentially) merging results obtained in an individual genome-wide association study with a larger meta-analysis of the same phenotype, and (d) searching for a biological interpretation of the final results. Each of these steps is described in the subsections below.

2.7.1 Multiple Hypothesis Testing

The analysis of large numbers of SNPs results in the simultaneous execution of equally large numbers of tests. This gives rise to a problem known as **multiple hypothesis testing** (MHT) [75]. As an example, a univariate genome-wide association study with d SNPs will result in d tests, each of them tested at a significance level α. On average, $d \times \alpha$ tests will be false positives and as d is normally in the order of 10^5 or 10^6, it is indispensable to apply a correction to the p-values to avoid reporting large numbers of false positives. The two most prominent approaches to correct for multiple hypothesis testing are controlling the family-wise error rate (FWER) and the false discovery rate (FDR).

Controlling the Family-Wise Error Rate (FWER)

The FWER is defined as the probability of obtaining one or more false positives. The most common method for controlling the FWER at level α is the Bonferroni correction [76], which guarantees the FWER to be smaller than α, with the per-test threshold $\delta = \alpha / d$.

Controlling the False Discovery Rate (FDR)

The FDR is defined as the expectation of the false discovery proportion, which in turn is the proportion of false associations among all significant associations. The most common method for controlling the FDR is the Benjamini–Hochberg procedure [77], although there exist other approaches as well, such as Benjamini–Yekutieli [78] and Storey–Tibshirani [79].

Comparison of FWER and FDR

The difference between the FWER and the FDR lies in the number of false positives one is willing to accept in the outcome. Controlling the FWER at a 5% significance level means that there is at most a 5% chance that one or more of the significant hits are false positives. In order to achieve this, only loci with strong association signals are deemed significant by the Bonferroni threshold. While this reduces the number of false positives, the number of false negatives tends to increase. When using the FDR to control for MHT, it means that on average up to 5% of the significant results might actually be false positives. While this approach is less conservative than the Bonferroni correction, the number of false positives might be higher and with more loci being considered significant, the number of false negatives tends to decrease.

2.7.2 Visualization of GWAS Results

Manhattan Plots

In a Manhattan plot [42], each dot corresponds to one genetic marker (SNP). Its genetic location is indicated on the x-axis, and the $-\log_{10}$ of its p-value on the y-axis (see plots in second column of Fig. 3a–d). This transformation results in SNPs with low p-values (and therefore stronger association) to have high values on the y-axis. Due to LD, SNPs in close proximity to each other show similar association to the phenotype, resulting in spikes in close proximity to SNPs with low p-values. This resembles the Manhattan skyline and gave rise to the term Manhattan plot.

Quantile-Quantile Plots

Under the null-hypothesis of no association, p-values follow a uniform distribution. Quantile-quantile (Q-Q) plots illustrate this expected distribution of p-values compared to the observed distribution. In a Q-Q plot, each dot corresponds to one SNP, and its position corresponds to its expected p-value (x-axis) against its observed p-value (y-axis) in $-\log_{10}$ space (see plots in first column of Fig. 3).

Under the general assumption in GWAS that only a small portion of the SNPs are associated to the phenotype [56], the majority of the expected and observed values should coincide (i.e., lie on the bisecting line of the plot). Deviations for a high number of markers, especially in the range of intermediate to high p-values, indicate the presence of confounders that artificially inflate the p-values (e.g., Fig. 3a). This inflation can be caused, for example, by population structure or cryptic relatedness among the individuals. As mentioned in Subheadings 1.3.1 and 3.4—for

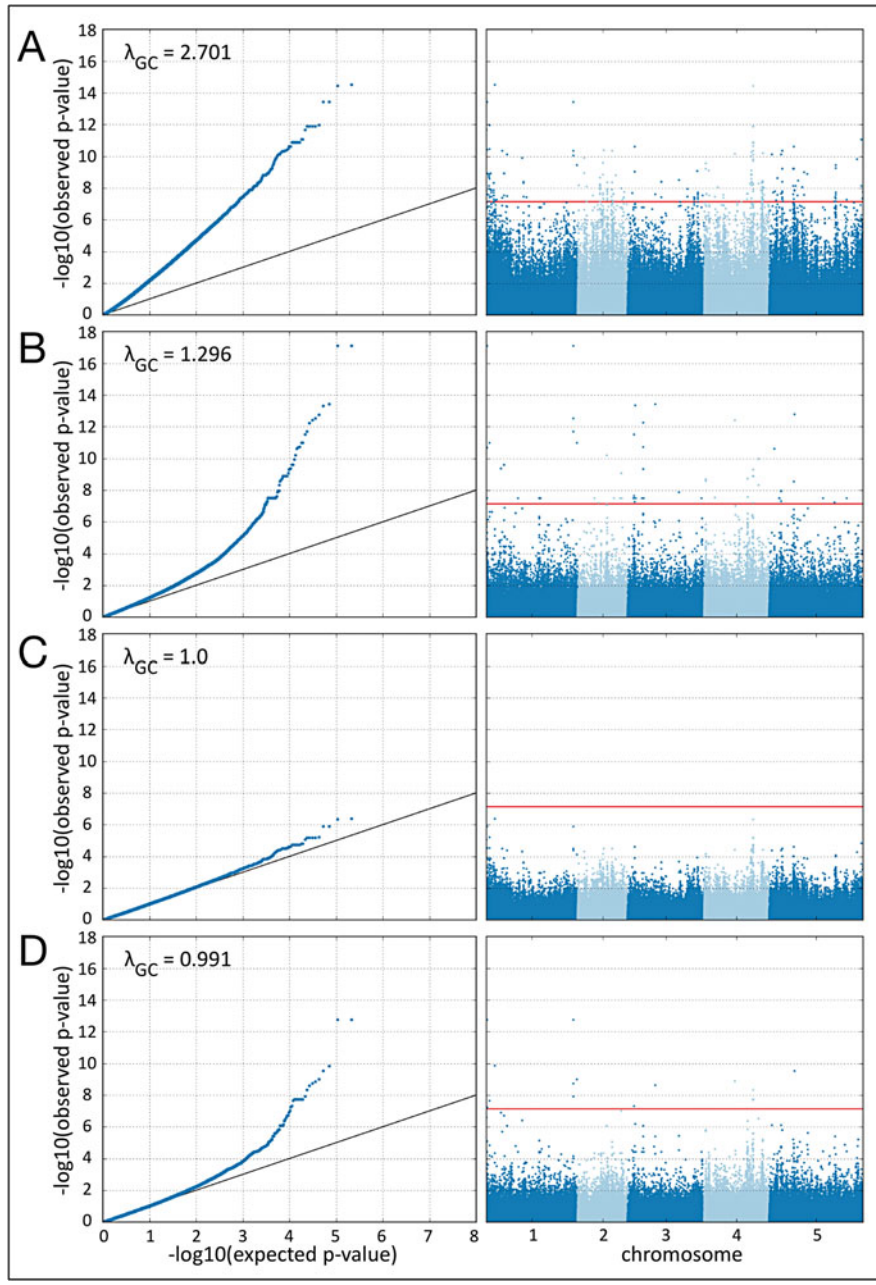

Fig. 3 Visualization of results for population structure correction with Q-Q plots (left) and Manhattan plots (right) for the *A. thaliana* dataset with the "FT Field" phenotype. The red line in the Manhattan plots represents the Bonferroni threshold. (**a**) Baseline: *p*-values derived with linear regression without any correction. (**b**) Using the ten leading PCs of the genetic similarity matrix as covariates in the linear regression. (**c**) *p*-values after correction with the genomic inflation factor λ_{GC}. (**d**) *p*-values generated with FaST-LMM

population structure—a correction needs to be performed in order to avoid reporting false positive results.

2.7.3 Meta-Analysis

Historically, the meta-analysis was developed as a tool to combine results from similar clinical trials [80]. After the advent of GWAS, meta-analysis has proved to be a robust methodology to combine results obtained from different studies. As each individual genome-wide association study has normally a modest sample size, a meta-analysis of multiple GWAS has the ability to increase the overall power and to reduce false positives [81].

It is a common practice in a meta-analysis of GWAS to pool the association signals detected in different studies without explicitly using the underlying genetic data. This is another aspect that makes meta-analysis such an appealing method, as access to genotype data is frequently regulated by strict privacy protections. Individuals that join a study may consent for their genetic data to be used in that specific study, but only allow for summary statistics to be disseminated among other research groups. These summary statistics are the association signals—there is no genotype information—that are aggregated in a meta-analysis.

There exist different ways to integrate signals from different studies. The most commonly used techniques are Fisher's method, Stouffer's method [82], and Stouffer's weighted method, as well as approaches based on fixed and random effect models [83]. The decision of which method is most appropriate to combine GWAS results heavily depends on the underlying assumptions of the studies at hand. The reader is referred to [81] for a comprehensive review of meta-analysis in GWAS.

2.7.4 Implications of Significant Findings

In the case of finding a truly significant SNP there are two different association outcomes: the SNP can either be causal and the association is called a **direct association**, or the SNP is in high LD with the causal variant, in which case one speaks of an **indirect association** [84]. Indirect associations occur when the causal SNP is not genotyped in the study, and the statistical test picks up the signal of the tag SNP marking the LD pattern that includes the causal variant. It is important to notice that without further experiments direct and indirect associations cannot be distinguished.

Another aspect of finding a significant variant is its biological implication. In many cases, markers that are deemed significant in a GWAS do not overlap with coding regions of a gene, but lie in intergenic regions [66]. It is common practice to map the significant SNP to genes that lie in close proximity, and assess if the gene is known to be involved in phenotype-specific pathways or functions.

3 Tools and Software

The previous section presented models, methods and equations that constitute the theoretical foundations of GWAS. There is a plethora of toolboxes and software packages that have been developed for all the different facets of GWAS previously described. While some of these tools are flexible and allow the user to conduct GWAS using a variety of models, others are more specialized and tailored for a specific subset of methods.

Here we start by summarizing the different tools and software in an attempt to provide a high-level view of the functional groups to which they belong. The remainder of this section presents details of a selected subset of tools, including clear protocols on how to run them. The vast majority of these tools are installed in the virtual machine (VM) that accompanies this book chapter and we encourage the reader to run the protocols we present here on the VM. The VM also contains a wiki page that facilitates the navigation through all sample scripts with their respective output and plots. Please refer to Subheading 1.4 for more details about the VM.

PLINK [67] is one of the most widely used software packages for GWAS. It allows the user to perform different kinds of analysis on SNP data, including univariate GWAS using two-sample tests (Subheading 2.3.1) and linear regression models (Subheading 2.3.2), as well as set-based tests (Subheading 2.5) and epistasis screenings (Subheading 2.6.1). Another flexible framework for deciphering the architecture of complex traits is GCTA [85], which started as a project to estimate phenotypic variation from SNP data and has subsequently been extended to accommodate more functionality, including GWAS using linear mixed models (Subheading 2.3.2). In addition to GCTA there are various toolboxes and software packages that implement different approaches to association testing with linear mixed models [53], among them FaST-LMM [58], EMMAX [86], GEMMA [62], and GRAMMAR-Gamma [87] as well as an extension to the Bayesian setting (Subheading 2.3.2) called BOLT-LMM [54].

Moving from the SNP to the gene level, there exist a variety of tools implementing different gene-based tests (Subheading 2.5). Examples of these are VEGAS [29] and PASCAL [32], which follow a two-step approach (Subheading 2.5.2), or MAGMA [88] and FaST-LMM-set [68] that are based on linear models (Subheading 2.3.2).

Another branch of software tools—those that implement network-based approaches (Subheading 2.6.2)—allow for the joint test of multiple variants by including prior knowledge in the form of biological networks. Some examples of such tools are SConES [35], dmGWAS [33, 36] and DAPPLE [34]. Other methods, that

also analyze sets of SNPs but that were not covered in Subheading 2, have the particularity of defining the sets to test on the fly. Two examples of these methods are FAIS [37] and hierGWAS [38].

Different tools and software packages not only distinguish themselves by the association methods and tests they provide, but also by their computational efficiency. In conducting univariate GWAS, for example, a tool can choose to test SNPs sequentially or to perform many association tests in parallel as there is no impediment to test multiple SNPs simultaneously. There are levels of parallelization that the user can easily implement, e.g., run a univariate genome-wide association study of human data on each of the 22 chromosomes separately (most GWAS in humans are performed on autosomal chromosomes and exclude chromosomes X, Y, and SNPs from the mitochondrion). But it is the level of parallelization offered by the tool that can clearly improve its runtime and computational efficiency. Some tools allow multi-threading on CPUs while others were designed to run on graphics processing units (GPUs) to achieve maximal parallelization of tasks. Good examples of the latter are the tools that provide efficient implementations of methods to search for epistasis (Subheading 2.6.1) such as EPIBLASTER [27] and GLIDE [28].

Our summary of tools and software would not be complete if we did not mention the GWAS workbenches that are available as online resources. Notable examples are EMMA [60], DGRP2 [17], Matapax [89], GWAPP [90] and easyGWAS [91]. In essence, they allow the user to perform GWAS, analyze and (in certain cases) annotate the results, all within a web server. The functionality provided by these web tools differs, but their main advantage consists in abstracting the user from the tedious work of having to conduct the analyses in their local installation.

In the following subsections, we give a descriptive introduction to commonly used tools for GWAS. Additionally, we provide examples that illustrate how the different types of GWAS introduced in Subheading 2 can be performed. These examples are presented as short snippets of code. Most of these examples require the specification of input data files by giving the complete path to the file. To facilitate the understanding of our examples, we assume a hypothetical file called mydata.txt. In referring to this file we use the following convention:

- Data directory or path: /home/gwasuser/data, the directory where the file is stored. In code snippets, the path will be stored in a variable, for example, $path.
- Full path: /home/gwasuser/data/mydata.txt, is the fully qualified name of the file in the file system, obtained simply by joining the data directory and filename.
- Extension of the filename: .txt, which normally is used to indicate the type of file. In this example, .txt means ASCII text.

Table 1
Execution times, on the VM, of all the scripts presented in Subheading 3

Script name	Execution time (hh:mm:ss)
example_3.1_binary2plain.sh	0:00:01
example_3.1_plain2binary.sh	0:00:02
example_3.2_preprocessing.sh	0:00:01
example_3.3.1_linear.sh	0:00:21
example_3.3.1_model.sh	0:00:01
example_3.3.1_assoc.sh	0:00:12
example_3.3.1_logistic.sh	0:00:20
example_3.3.2_lmm.sh	0:01:55
example_3.4.1.1_compute_pcs_plink.sh	0:00:02
example_3.4.1.2_compute_pcs_eigensoft.sh	0:00:25
example_3.4.1_correction_with_pcs.sh	0:05:20
example_3.4_generate_plots.sh	0:03:27
example_3.5.1.1_set_flag.sh	0:09:00
example_3.5.1.1_make_set_flag.sh	0:10:37
example_3.5.1.2_vegas.sh	No runtime
example_3.6.1.2_epistasis.sh	29:40:01
example_3.7.1_dmgwas.sh	0:10:25

The times are in hours:minutes:seconds and represent the actual elapsed time by the script (obtained with the Linux command time, field real)

- Prefix of the filename: mydata, obtained by removing the extension from the filename.

Some of the examples in this section will have combinations of the items above, such as $path/mydata.txt (to specify the full path) or $path/mydata (as it is the case in many tools that exclude the extension and only require the path and prefix of the file). In addition to these code snippets, we provide sample scripts that can be executed on the VM. Table 1 shows the running times on the VM of all the scripts presented in Subheadings 3.1 through 3.7.

3.1 Data Formats in GWAS

The file format that has become a *de facto* standard for GWAS is the file format used by PLINK [67]. This is commonly referred to as **PLINK format**. Genotype data, in PLINK format, can be stored as two different types of files: plain text and binary. Text files, normally, are delimited by a white space or tab and have the extensions .ped and .map (Fig. 4a, b). Each line in a .ped

file corresponds to one individual. The first six columns of the file contain the family ID, individual ID, paternal ID, maternal ID, gender, and phenotype, in that order. The remaining columns contain the genotype information. To complement the .ped file, the .map file contains one line for each SNP, indicating the SNP location and its identifier (if given). The ordering of the SNPs in the .ped file (as columns) matches the ordering of the SNPs in the .map file (as rows). Storing genotype data as plain text implies that the .ped file is both large in size (in the order of dozens of GB) and cumbersome to process. This is the reason why the second type of PLINK file, the binary file, is the one that is most commonly used. The PLINK binary file has the extension .bed and contains, essentially, the same genotype information as the .ped file, albeit in binary format. Full datasets for GWAS, in PLINK binary format, consist of three files with identical prefix but with different extensions: .bed, .bim, and .fam (Fig. 4c, d). The .bim and .fam files are delimited by a white space. A .bim file contains information about the SNPs in the study (similarly to a .map file). The .fam file has the same information stored in the first six columns as a .ped file.

Example of Data Conversion on the VM (using PLINK)

To run an example of data conversion in which a PLINK file in binary format is converted to text format, navigate to the directory containing the sample code by typing at the prompt "cd $EXAMPLES/3.1_data_formats". Execute the script by typing "./example_3.1_binary2plain.sh". Alternatively, if the conversion is from plain text to binary, you can execute another of the sample scripts by typing "./example_3.1_plain2binary.sh". These scripts will perform data conversion on a sample dataset of the plant *Arabidopsis thaliana* (abbreviated as *A. thaliana*) included in the VM.

3.2 Data Preprocessing with PLINK

Most of the data preprocessing in GWAS can be easily performed with PLINK. Depending on the filtering criteria to use (*see* Subheading 2.2.2) different flags must be specified. As an example, the command:

```
plink --bfile $path/mydata --make-bed
      --out $path/mydata_filtered --maf 0.01
      --hwe 1e-6
```

processes the binary genotype data stored in mydata.bed (with its accompanying files mydata.bim and mydata.fam), and generates a filtered output file mydata_filtered.bed (again, with the accompanying files mydata_filtered.bim and mydata_filtered.fam). These new files contain genotype data that passed two filtering criteria: (a)

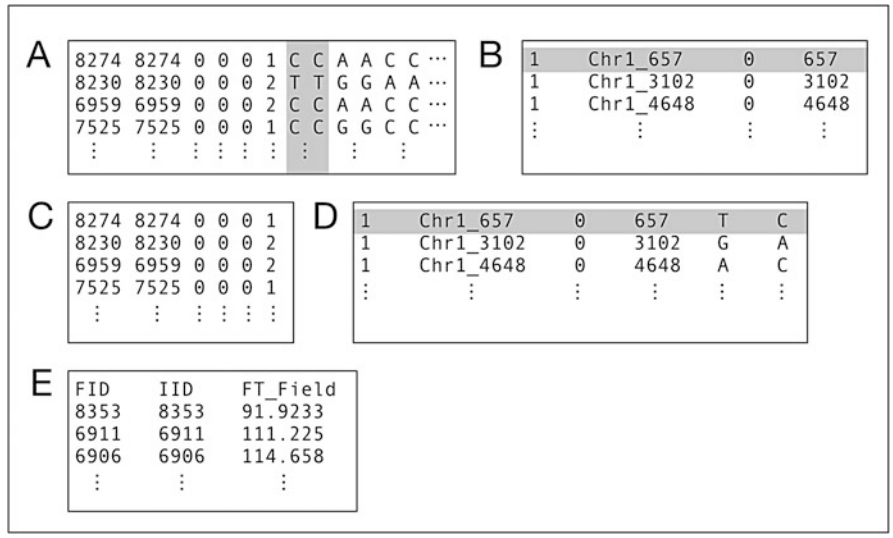

Fig. 4 The different types of PLINK files for the *A. thaliana* dataset. (**a**) .ped file: Each line corresponds to one sample. The first six columns are the family ID, individual ID, paternal ID, maternal ID, sex, and phenotype, respectively. The remaining columns are the SNP data, with two columns per SNP, representing to the two alleles (data for one SNP is highlighted in grey). (**b**) The .map file, in which each line represents one SNP. The columns correspond to the chromosome, the SNP identifier, the genetic distance in morgans (optional), and the base-pair position, respectively. The SNP highlighted in grey corresponds to the one highlighted in (**a**). (**c**) The .fam file corresponds to the first six columns of the .ped file. (**d**) Example of the .bim file. It has the same format as the .map file plus two additional columns indicating the two alleles of the SNP. (**e**) Example of the phenotype file with the "FT Field" phenotype. The columns correspond to the family ID, the individual ID and the phenotype value

Table 2
Flags to invoke different data filtering in PLINK and their commonly used values

Flag	Description	Standard value
--maf {x}	minor allele frequency (MAF); SNPs with lower MAF than {x} are excluded from the analysis	0.01 or 0.05
--hwe {x}	*p*-value threshold of HWE below which SNPs are excluded	1e–6
--mind {x}	Samples with more than {x} * 100% missing genotypes are excluded	0.1
--geno {x}	SNPs with more than {x} * 100% missing values are excluded	0.1

SNPs with minor allele frequencies larger than 1% and (b) SNPs in Hardy–Weinberg equilibrium at a 1e−6 significance threshold. *See* Table 2 for additional filtering flags.

Example of Data Preprocessing on the VM (using PLINK)

To run an example of data preprocessing on the *A. thaliana* dataset, navigate to the directory containing the code by typing at the prompt "cd $EXAMPLES/3.2_data_preprocessing". Execute the script by typing "./example_3.2_preprocessing.sh".

Table 3
Flags to invoke commonly used association tests in PLINK and the phenotypes for which the tests are appropriate

Flag	Description	Phenotype
--assoc	1-degree-of-freedom χ^2 test	case/control
--model	1df dominant, 1df recessive, 2df genotypic, Cochran-Armitage trend	case/control
--logistic	Logistic regression	case/control
--linear	Linear regression	quantitative

3.3 Univariate Association Studies

As described in Subheading 2.3, a univariate association study tests each SNP separately for association with the phenotype. Below we present in more detail two tools to conduct univariate association studies: (a) PLINK, which implements a variety of methods, and (b) FaST-LMM, designed to support different linear mixed model approaches.

3.3.1 Univariate GWAS Using PLINK

A standard case/control association study can be run with PLINK by using the following command:

Basic Analysis

```
plink --bfile $path/mydata_filtered
      --out $path/mydata_filtered --assoc
```

The `--assoc` flag calls a 1-degree-of-freedom χ^2 test on the SNPs in the binary dataset mydata_filtered.bed (with its .bim and .fam files) and writes the output to mydata_filtered.assoc. Other than the χ^2 test, PLINK can be used to find associations with models such as linear or logistic regression. Table 3 lists commonly used flags for univariate testing implemented in PLINK.

Additional Options

PLINK offers a variety of additional flags. Explaining all of them is beyond the scope of this chapter, so we focus on the ones that are of specific use in standard GWAS.

A useful flag that can be added is `--adjust`. It will generate an additional output file with the suffix .adjusted in its filename. This file contains the raw *p*-values as well as the *p*-values after correction for multiple hypothesis testing.

Upon availability, including covariates such as age, gender or sex into a genome-wide association study reduces the amount of false positives. When a linear model is chosen, i.e., a linear or logistic regression, covariates specified in a covariate file can be included using the `--covar` flag, followed by the full path filename of the covariate file (*see* Fig. 4e). The `--covar-number` and `--covar-name` flags allow the user to select a subset of all the covariates in the file. These flags must be followed by the indices of the columns or the names of the chosen covariates. In case the name of a covariate is used, it must match the name of a column in the header of the file.

In addition to the *p*-values computed from the theoretical null-distributions of the test, empirical *p*-values can be obtained via permutation testing. PLINK offers two options for this, namely the `perm` method which performs adaptive Monte Carlo permutation testing, or the `mperm={number of permutations}` method, that computes a max(T) permutation. Both methods can be called with PLINK by adding them directly after the model in the command line, for example:

```
plink --bfile $path/mydata_filtered
      --out $path/mydata_filtered
      --assoc perm plink
      --bfile $path/mydata_filtered
      --out $path/mydata_filtered
      --assoc mperm=1000
```

Both calls result in output files with the additional extensions perm and mperm, respectively.

By default, PLINK uses the phenotype that is specified in the .fam file. Nevertheless, a different phenotype can be specified by using the `--pheno` flag followed by the full path to a text file containing the phenotypes of interest, *see* Fig. 4e for the file format. In case `--pheno` is given, the phenotype specified in the .fam file will be ignored.

3.3.2 Univariate GWAS Using FaST-LMM

FaST-LMM (short for Factored Spectrally Transformed Linear Mixed Models) [58] is a method and software package for association studies using LMMs. FaST-LMM computes a genetic similarity matrix and includes it in the linear model via a random effect, thereby correcting for structure among the samples that potentially could cause false positives. By default, and in order to avoid proximal contamination, FaST-LMM uses a **leave-out-one-chromosome** approach. This means that the genetic similarity matrix is computed from SNPs on all chromosomes, except for

those on the chromosome containing the SNP to be tested for association. Here we explain how to work with the Python implementation of FaST-LMM.

Basic Analysis

To run a basic association analysis with FaST-LMM, a .bed file containing the SNPs to be tested is needed, as well as a separate text file with the phenotypes. FaST-LMM is then invoked from Python with the `single_snp` function from the fastlmm package (https://pypi.python.org/pypi/fastlmm):

```
single_snp(test_snps=$path/mydata,
    pheno=$path/mypheno.txt,
    output_file_name=$path/myoutput.txt)
```

This will perform the association study for the genotype data in mydata.bed and the phenotype specified in mypheno.txt. The output will be written to myoutput.txt. FaST-LMM can be applied to both, quantitative and qualitative phenotypes without having to specify this in the function call.

Additional Options

As mentioned above, FaST-LMM always includes at least one genetic similarity matrix via a random effect in the model. By setting the argument `K0` as a .bed file, these data will be used to compute the first genetic similarity matrix. If `K0` is not specified, the data in `test_snps` will be used for its computation.

In addition to the first genetic similarity matrix, a second one can be included by adding the parameter `K1` followed by a .bed file containing the genotypes for its computation. Similar to `K0`, `K1` will also be added as a random effect to the model.

Furthermore, FaST-LMM allows for the inclusion of covariates as fixed effects, using the `covar` argument. The format of the covariate file is the standard PLINK covariate file format (*see* Fig. 4e). In contrast to PLINK it is not possible to specify which covariates to use, but all covariates in the file will be included.

3.3.3 Other Tools

Besides PLINK and FaST-LMM there exist many different tools for performing univariate GWAS that implement different approaches and methodologies. Among the most widely used ones are the following:

(a) GCTA (Genome-wide Complex Trait Analysis) [85] is a flexible toolbox for the analysis of complex traits that comprises a high number of functionalities, such as estimation of genetic relationships, estimation of phenotypic variance, data transformation, and GWAS with linear mixed models.

(b) BOLT-LMM [54] is a method for GWAS using Bayesian mixed models with a mixture-of-Gaussians prior.

(c) EMMAX (Efficient Mixed-Model Association eXpedited) [86] constitutes a toolbox for GWAS with linear mixed models.

Examples of Univariate GWAS

To run the different approaches of univariate GWAS on the *A. thaliana* dataset, navigate to the directory containing the code by typing at the prompt "cd $EXAMPLES/3.3_univariate_gwas". There are four scripts to perform a univariate analysis with PLINK, each with a different model, and one script to perform an analysis with FaST-LMM. To run any of them, type "./example_3.3.1_{plink_model}.sh" or "./example_3.3.2_lmm.sh".

3.4 Population Structure Correction

The three methods to correct for population structure introduced in Subheading 2.4 are based on (a) the principal components (PCs) of the genetic similarity matrix, (b) the genomic inflation factor, and (c) the application of LMMs.

3.4.1 Correction Using Principal Components of the Genetic Similarity Matrix

The main task in this approach to correct for population structure is the computation of the leading principal components of the genetic similarity matrix. Once the PCs have been computed, they can be included in any statistical model for GWAS that allows for covariates, *see* Subheading 3.3 for details. Here we present how the PCs can be computed using PLINK and EIGENSTRAT [12]. The main difference between the two methods is that EIGENSTRAT allows for automatic outlier removal in addition to providing an efficient approximation for very large datasets. In the examples shown below, the PCs obtained from both tools are comparable.

Computation of the Principal Components Using PLINK

The following PLINK command generates the leading `num_pcs` principal components of the genetic similarity matrix, computed from the genotype data in mydata.bed:

```
plink --bfile $path/mydata
      --out $path/mypcs --pca num_pcs header
```

It generates the two output files mypcs.eigenvec and mypcs.eigenval, containing the leading eigenvectors (corresponding to the PCs) and the associated eigenvalues, respectively. The `header` modifier adds a header line to the eigenvector file, making it directly usable for tools such as PLINK or FaST-LMM.

Computation of the Principal Components Using EIGENSTRAT

Another tool for the computation of the principal components of the similarity matrix described in Eq. 5 is EIGENSTRAT, implemented in EIGENSOFT [92]. It contains the function smartpca.perl, which can be called from the command line as follows:

```
smartpca.perl -i $path/mydata
    -a $path/mydata -b $path/mydata -k {num_pcs}
    -o $path/myoutput -p $path/myplot
    -l $path/mylog -e $path/myeigenvalues
```

The `-i`, `-a`, and `-b` flags expect the full path filenames of the .bed, .bim and .fam files, respectively. With the `-k` flag, the default number of ten PCs can be changed. With the `-o` flag, the output prefix is specified; `-p` is followed by the filename of an output plot displaying the data along the first two PCs. The `-l` flag specifies the filename of the log-file, and the `-e` flag corresponds to the filename to which the eigenvalues should be written. In order to use the PCs as covariates with PLINK or FaST-LMM, they have to be brought into the PLINK phenotype file format (Fig. 4e). We provide a function for this in our allgwas module on the VM located under /home/gwasuser/tools/allgwas.

3.4.2 Correction Using the Genomic Inflation Factor

Computing *p*-values that are corrected for population structure by using the genomic inflation can be done with PLINK by adding the `--adjust` flag to the GWAS command, for example:

```
plink --bfile $path/mydata
      --out $path/myoutput --assoc --adjust
```

In addition to the standard output file myoutput.assoc, this generates a second output file myoutput.assoc.adjusted that contains the raw *p*-values and different types of adjusted *p*-values. Among these, the *p*-values adjusted for population structure with the genomic inflation factor are indicated in the header line by "GC".

3.4.3 Correction Using LMMs

Due to the success of LMMs in GWAS, many toolboxes that implement this method have been developed. One of them is FaST-LMM which, as discussed in Subheading 3.3.1, automatically computes the genetic similarity matrix using a **leave-out-one-chromosome** technique and includes it as the covariance matrix of a random effect into the model.

3.4.4 Comparison of the Different Approaches for population Structure Correction

To illustrate the different correction methods for population structure, we apply them to the *A. thaliana* dataset, together with the quantitative phenotype "FT Field" which measures the number of days to flowering of plants grown in the field [11]. Figure 3a displays the baseline, i.e., the *p*-values obtained from a linear regression without any form of correction for population structure. It is obvious from the Q-Q plot and Manhattan plot that the *p*-values are inflated. In the Q-Q plot this inflation manifests itself by the deviation of the *p*-values from the bisecting line, whereas

in the Manhattan plot one can recognize the inflation by the large number of significant SNPs. This contradicts the prior assumption that only few loci are associated to the phenotype.

We compute the PCs of the genetic similarity matrix and subsequently use them as covariates in the linear regression model. The genomic inflation factor λ_{GC} shows the least deviation from one when using the leading ten PCs as covariates (*see* Fig. 2). Including them as covariates in the linear regression model results in the *p*-values shown in Fig. 3b. Although the inflation measured by λ_{GC} drops closer to one, the *p*-values remain inflated and still a large number of the SNPs are significant after Bonferroni correction.

When correcting for population structure using λ_{GC}, the inflation decreases substantially, leading to no genome-wide significant results (*see* Fig. 3c). The problem with this approach is that, although the inflation vanishes, actually there is no correction. The ranking of the SNPs with respect to their *p*-value remains unchanged. This means that the SNPs with the lowest *p*-values prior to the correction will remain the ones with lowest *p*-values after correction, despite the fact that they might actually be associated to population structure.

Figure 3d shows the *p*-values obtained from the LMM approach implemented in FaST-LMM. In comparison to the baseline and the PC approach, this results in the least inflation.

Examples of population structure correction

To compute the principal components of the genetic similarity matrix using PLINK or EIGENSOFT for the *A. thaliana* dataset, navigate to the directory containing the code by typing at the prompt "cd $EXAMPLES/3.4_population_structure_correction" and perform the computation by typing "./example_3.4.1.1_compute_pcs_plink.sh" or "./example_3.4.1.2_compute_pcs_eigensoft.sh" at the command prompt. When executing "./example_3.4.1.2_compute_pcs_eigensoft.sh", the inflation factor for different numbers of leading PCs is computed, and Fig. 2 is generated. The script example_3.4.4_generate_plots.sh generates the individual plots in Fig. 3. To run it, type "./example_3.4.4_generate_plots.sh".

3.5 Gene-Based Tests

In Subheading 2.5 two different approaches for gene-based testing were introduced, namely the two-step approaches that combine univariate SNP test-statistics into test-statistics for genes, and the one-step approaches that consider all SNPs mapped to a gene jointly to derive a *p*-value for that gene.

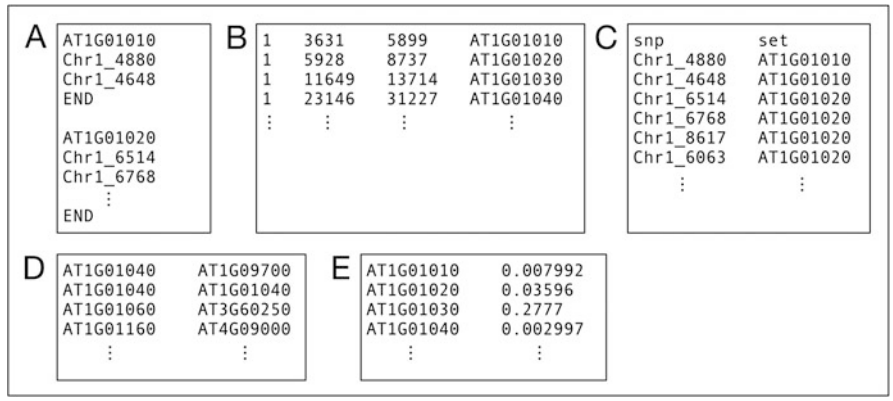

Fig. 5 Examples of file types for gene-based and network-based GWAS for the *A. thaliana* dataset. (**a**) PLINK's *set* file. Each gene is indicated by the gene-name, followed by the SNPs in the gene. The end of the gene-set is marked by END. (**b**) Gene annotation file: each line describes a gene, the columns correspond to the chromosome, the starting and ending position of the gene, as well as the name of the gene. (**c**) Gene file required for FaST-LMM-set: each line corresponds to an SNP and the gene, the SNP is mapped to. (**d**) dmGWAS network interaction file: each row corresponds to an interaction between the two genes in the first and second column. (**e**) dmGWAS gene *p*-value file: each row corresponds to one gene, the first column is the gene name, the second column is the *p*-value of the gene

3.5.1 Two-Step Approaches for Gene-Based Testing

Gene-Based Testing Using PLINK

PLINK computes a gene *p*-value using the *average test statistic* approach. By default, it will use the five most significant univariate test statistics for this average, after omitting SNPs in high LD (default >0.5). Since the distribution under the null hypothesis is not known, the method relies on permutation testing. The number of permutations can be adjusted using the `--mperm` flag, with a default value of 1000. The basic command to run a gene-based test with PLINK is the following:

```
plink --bfile $path/mydata
      --out $path/gene_results
      --set-test --assoc --mperm 1000
```

In addition to these flags, the information of the SNP-to-gene assignment has to be included. There are two different ways to do this:

(a) The first option is to generate a set file that assigns the SNPs to the genes (*see* Fig. 5a), and to pass its full path filename to PLINK with the `--set` flag, i.e.

```
plink --bfile $path/mydata
      --out $path/gene_results
      --set-test --assoc --mperm 1000
      --set $path/myset.set
```

(b) The second option is to use gene-range lists, and include them with the `--make-set` flag (*see* Fig. 5b). This allows the specification of borders around the genes, and SNPs lying within these borders will also be assigned to the gene. The flag `--make-set-borders` lets the user define the size of the borders in kilobases (kb):

```
plink --bfile $path/mydata
      --out $path/gene_results
      --set-test --assoc --mperm 1000
      --make-set $path/hg_built.txt
      --make-set-borders 20
```

For both options (a) and (b), the user can change the maximum number of SNPs used for the average by including the flag `--set-max`. The flag `--set-p` determines the maximum univariate *p*-value an SNP can have to be considered in the average. The `--set-p` flag overrules the `--set-max` flag in the sense that SNPs with *p*-values above the set-p threshold will never be included, even if this implies including fewer than set-max SNPs in the average.

Gene-Based Testing Using VEGAS

VEGAS (Versatile Gene-based Association Study) [29] is a command line tool to compute gene-based *p*-values using the *sum of test statistics* approach. As explained in Subheading 2.5.2, SNPs in close proximity commonly show high LD, violating the assumption of independence of the test-statistics in the sum. In order to obtain the null distribution while taking the correlation between SNPs into account and thus obtain valid *p*-values, permutation testing is indispensable, but it comes at the cost of a high computational effort. VEGAS circumvents this computational burden by replacing the permutations with draws from a *p*-value distribution that accounts for the observed LD pattern (Monte Carlo simulations).

VEGAS is applicable to human data only, since it expects SNP-identifiers to follow the *rs#* naming convention. The tool can be used either online or as a downloadable command line tool. Here, we explain briefly how the downloadable version of VEGAS is used. The following command invokes a basic run:

```
vegas $path/mydata.txt -pop {population}
    -out $path/myoutput
```

where mydata.txt is a two-column file delimited by a white space, containing the univariate *p*-values for the SNPs, and `{population}` corresponds to the name of a reference population from which LD should be estimated. The output will be saved to myoutput.out.

As an alternative to the `-pop` flag, the user can invoke VEGAS with the `-custom` flag, pointing to a .bed file. When this flag is set, LD is estimated from the genotypes in the .bed file.

3.5.2 One-Step Approaches for Gene-Based Testing

Gene-Based Testing Using FaST-LMM-Set

In FaST-LMM-set, sets of genetic variants are tested for their association with a phenotype by implementing the one-step method introduced in Subheading 2.5.2. Defining each gene (plus a border region) as a set, FaST-LMM-set fits an LMM with the genotype-matrix as the covariance matrix of a random effect. In contrast to the univariate implementation (Subheading 3.4.2), a random effect to model the relatedness among the individuals is optional and will not be added by default to the model. Upon availability, covariates can be included as a fixed effect. Once a model for a gene is fitted, a likelihood ratio or score test is applied to test the gene effect.

The Python implementation of FaST-LMM-set requires a file containing the mapping of SNPs to genes (*see* Fig. 5c). FaST-LMM-set can be invoked using the Python function `snp_set` from the fastlmm package with the following command in Python:

```
snp_set(test_snps=$path/mydata,
    set_list=$path/set_file.txt,
    pheno=$path/mypheno.txt,
    output_file_name=$path/myoutput.txt)
```

where the genotype data is stored in mydata.bed, the phenotype information in the file mypheno.txt, and the SNP-set information in set_file.txt. This will generate an output file myoutput.txt, which contains the statistics and *p*-values of the genes.

The user has the option to include covariates with the `covar` argument. Furthermore, an additional random effect can be included to model the relatedness among the samples. To do so, the `G0` argument has to be set to point to a .bed file that contains the SNPs from which genetic similarity should be estimated. Both arguments `covar` and `G0` help reduce the bias of confounders on the association results. The statistical test used can be changed using the test argument, with two settings available: the score test, invoked with `test='sc_davies'`, or the likelihood ratio test (default), invoked with `test='lrt'`.

Examples for Gene-based Testing

To perform gene-based testing on the *A. thaliana* dataset, navigate to the directory containing the code by typing "cd $EXAMPLES/3.5_gene_based_testing" at the prompt. The two different PLINK approaches can be executed by typing "./example_3.5.1.1_make_set_flag.sh" or "./example_3.5.1.1_set_flag.sh". Typing "./example_3.5.2.1_fastlmm_set.sh" runs the FaST-LMM-set method.

In addition, we provide a script to run VEGAS on a PLINK toy-dataset. The script does not execute because it requires supplementary files that do not fit in the VM. Nevertheless, the script is available as an example of how to invoke VEGAS after running a univariate genome-wide association study with PLINK. It can be found at "$EXAMPLES/example_3.5.1.2_vegas.sh".

3.6 Epistasis Search

The search for epistatic interactions is computationally more challenging than the search for single associations. In this search, every locus will be tested against every other locus, which squares the number of models to fit compared to univariate GWAS. In order to alleviate this burden, it is common practice to first run a univariate analysis and only use high ranking SNPs for an epistasis screening. However, this approach might not detect interactions between variants with moderate to low effect sizes. Another way to deal with the computational complexity of the problem is to rely on parallel programming, e.g., on graphics processing units (GPUs). Here, we describe how a search for epistatic interactions can be performed with PLINK and briefly mention two other approaches based on GPU-computing.

3.6.1 Two-Locus Epistasis Search Using PLINK

PLINK provides two functions to run an epistasis screening, namely the fast-epistasis method and the epistasis method.

The fast-epistasis constitutes a fast scan for epistatic interactions only with a qualitative phenotype, e.g., in a case/control study. By default, every two loci are first collapsed from a 3×3 contingency table into a 2×2 table separately for cases and controls. Then, for the two classes, the odds ratios are computed based on the tables, and their difference is tested for significance. While this is a fast method, it does not take the individual effect of each locus into account, and should therefore rather be used for a first screening before carrying out a proper epistasis study.

The second option, epistasis, is based on the linear model described in Eq. 6. As opposed to fast-epistasis, it works with qualitative and quantitative phenotypes. It requires fitting a model for each combination of SNPs and is thus much more time consuming than the fast-epistasis method. *See* Table 1 for a comparison of the execution times between the two methods.

Epistasis Screening Using Fast-Epistasis

A fast-epistasis run can be invoked with the following command:

```
plink --bfile $path/mydata
      --out $path/myoutput --fast-epistasis
```

where the genotype data is contained in mydata.bed (with its corresponding .bim and .fam files). This will produce two output files, myoutput.epi.cc and myoutput.epi.cc.summary. The former is a text file where each line corresponds to one SNP-pair with an

epistasis p-value smaller than 0.0001 (this threshold can be modified using the `--epi1` flag; for small datasets with a small number of pairs it can be set to 1). In the file myoutput.epi.cc.summary each line corresponds to one SNP. It contains information on how often the SNP occurred in an epistatic interaction with p-value lower than 0.01 (column with header "N_SIG", this threshold can be adapted with `--epi2`) and the interaction partner with which the lowest epistasis p-value was reached.

Epistasis Screening Using Epistasis

Analogously to fast-epistasis, one can run a full epistasis screen based on linear/logistic regression with the following command:

```
plink --bfile $path/mydata
      --out $path/myoutput --epistasis
```

This generates two output files: myoutput.epi.qt and myoutput.epi.qt.summary for a quantitative trait, or myoutput.epi.cc and myoutput.epi.cc.summary for a case/control phenotype. These output files report the same statistics as in the fast-epistasis case, and the flags `--epi1` and `--epi2` behave in the same way.

3.6.2 Other Tools

Apart from PLINK, other commonly used tools for the detection of epistatic interactions are:

(a) EPIBLASTER [27]: GPU-based tool to detect epistatic interactions in case/control datasets. The detection of two-locus interactions is based on a two-stage approach, where in the first stage SNP-SNP pairs are selected based on the difference in correlation to the phenotype in cases versus controls. In the second stage, selected pairs are tested using logistic regression with a likelihood ratio test.

(b) GLIDE [28]: GPU-based epistasis tool based on linear regression and a 4-degree-of-freedom t-test. It is applicable to quantitative and qualitative phenotypes and has high performance due to an efficient GPU implementation.

Examples for Epistasis Searches

To perform an epistasis search on the *A. thaliana* dataset, navigate to the directory containing the code by typing at the prompt "cd $EXAMPLES/3.6_epistasis". The two different PLINK-approaches can be run by typing "./example_3.6.1.1_fast_epistasis.sh" or "example_3.6.1.2_epistasis.sh". Please note that performing epistasis search is time-consuming.

3.7 Network-Based GWAS

3.7.1 Dense Module Searching for Genome-Wide Association Studies (dmGWAS)

dmGWAS [33, 36] is a method implemented in an R package to find modules within a biological interaction network that are enriched with low *p*-value genes. The search for these modules follows a greedy strategy [93], i.e., the algorithm iteratively adds the gene with the lowest *p*-value to the current module. This corresponds to making the locally optimal decision at each step, which might not necessarily lead to a globally optimal decision. While it allows for an efficient exploration of large networks, it comes with the downside of not analyzing all possible modules.

As input, dmGWAS requires two files: the first one, edge_file.txt, is a two-column, tab-delimited text file containing the network. Each line corresponds to one edge, represented by its two adjacent nodes (*see* Fig. 5d). The other file, pvalue_file.txt, is also a two-column, tab-delimited text file, containing the *p*-values assigned to each node (first column is the node and the second column the *p*-value) (*see* Fig. 5e). *P*-values can be obtained, for example, by implementing one of the methods described in Subheading 3.5. The following commands in R convert the text files to R-readable tables that can be used with dmGWAS:

```
network <- read.delim($path/edge_file.txt,
    as.is=T, header=F)
pvalues <- read.delim($path/pvalue_file.txt,
    as.is=T, header=F)
```

In case the files contain a header in the first line, the header flag should be set to `header=T`. With the data in this format, dmGWAS can be invoked with the following command:

```
res.list <- dms(network, pvalues,
    expr1=NULL, expr2=NULL, d=2, r=0.1)
```

The arguments `expr1` and `expr2` correspond to files containing the gene-expression for cases and controls, respectively, and are used to compute edge-weights. We will not use this functionality here and refer the interested reader to [36]. There are two more parameters, `d` and `r`, that affect the behavior of the greedy search algorithm. Following the recommendation of the authors [33, 36], these values should be set to `d=2` and `r=0.01`.

3.7.2 Other Methods for Network-Based GWAS

(a) SConES (Selecting Connected Explanatory SNPs) [35] is a method implemented in MATLAB to efficiently detect sets of SNPs that are connected in an underlying network and have maximal association to a phenotype, while meeting sparsity and connectivity constraints.

(b) DAPPLE (Disease Association Protein-Protein Link Evaluator) [34] is based on testing the hypothesis that modules of

associated genes are more densely connected than expected by random chance. In order to test this hypothesis, permutation testing is applied.

Examples for Network-based GWAS

To perform a network-based GWAS on the *A. thaliana* dataset, navigate to the directory containing the code by typing at the prompt "cd $EXAMPLES/3.7_network_based". It contains a script to run dmGWAS which can be executed by typing "./example_3.7.1_dmgwas.sh" at the prompt.

3.8 Other Types of GWAS

3.8.1 Hierarchical GWAS (hierGWAS)

Buzdugan et al. [38] presented a hierarchical testing procedure for SNP data using a linear regression coupled with a multiple sample splitting procedure for fitting. By clustering SNPs according to a predefined criterion, and hierarchically testing those clusters from largest (all SNPs) to smallest (single SNPs), they developed a framework for testing the cluster-tree and correcting for MHT. Their approach is implemented in the R package *hierGWAS*.

3.8.2 Combinatorial Association Mapping

The goal of combinatorial association mapping is to determine the statistical association of higher-order interactions between genetic markers and a phenotype of interest. When making a naive attempt to test all higher-order interactions between SNPs one faces two main obstacles: firstly, the vast number of association tests that need to be performed creates a computational bottleneck and, secondly, there is a sharp decrease in statistical power due to corrections for MHT (e.g., Bonferroni correction). Newly developed algorithms based on significant pattern mining provide an alternative to mitigate these computational and statistical challenges. These algorithms can be used to perform region-based association studies [94] and to detect higher-order epistatic interactions of genetic variants [95], while allowing for the correction of categorical covariates such as age or gender. These approaches are implemented in the Python and R package CASMAP.

3.9 Online Tools

Apart from downloadable tools, some online tools have been developed to allow the user to perform GWAS on web-servers. Since these web-servers provide user-friendly graphical interfaces, they are usually more intuitive to use than command line tools or Python/R packages, which require the user to have some basic knowledge in computer science. Due to current legislation however, users are not allowed to upload and analyze human genetic data with those tools.

3.9.1 GWAPP

GWAPP [90] is a web application for performing GWAS in the model organism *A. thaliana*. There are a variety of accessions

available on the web-server that can be analyzed with a user-specified phenotype. GWAPP implements three methods to test for associations, namely linear regression, an accelerated mixed model and a Wilcoxon rank sum test [96]. Correction for MHT is done using the Benjamini–Yekutieli method to control the FDR [78]. Interactive Manhattan plots, LD structure plots, and different statistics (depending on the model) are generated as final analysis results.

3.9.2 easyGWAS

easyGWAS [91] is a cloud platform for computing, sharing and comparing the results of GWAS on user-specific datasets and phenotypes, as well as publicly available genotype and phenotype data from model organisms, such as *Arabidopsis thaliana* [13, 15, 97–99] and *Pristionchus pacificus* [100]. Based on the phenotype and genotype, the user can choose between various state-of-the-art GWAS methods, including the Wilcoxon rank-sum test, linear/logistic regression and FaST-LMM [58]. easyGWAS allows the user to apply different genotype encodings of the data (*see* Fig. 1b) and phenotype transformations. Furthermore, the user has the possibility to also perform meta-analysis on already computed GWAS or compare its results using a novel intersection analysis. In addition, the user can choose between four different multiple hypothesis correction methods: Bonferroni [76], Benjamini–Hochberg [77], Benjamini–Yekutieli [78], and Storey–Tibshirani [79]. easyGWAS offers dynamic visualizations to facilitate the analysis of GWAS results. Some examples are: interactive Manhattan plots with clickable SNPs, LD plots with gene annotations and chord diagrams to illustrate associated SNPs that are shared between different GWAS.

4 Remarks and Conclusions

Genome-wide association studies have been instrumental in identifying common genetic variants that correlate with specific traits or diseases. The early successes in GWAS, in particular at identifying new susceptibility loci for different types of cancer [101], led some companies to produce and market risk tests that would allow an individual to assess their risk level of developing certain diseases. This was cause for alarm at the time, especially because most of the risk alleles in the reported SNPs only conferred very small increments in risk, between 1.1 and 1.5-fold [102]. As a result, some of these risk tests, such as one that was produced in 2008 for breast cancer, drew intense criticism [103]. Essentially, the test was based on seven SNPs, none of them located on the genes BRCA1 and BRCA2 whose mutations account for up to 40% of familial breast cancer. Although the seven SNPs had been reported in GWAS as being associated to breast cancer, the scientific consensus

was that in order to effectively assess risk in a patient, more associated variants needed to be discovered. Events like this, coupled with the fact that the SNPs reported in early GWAS had small effects and explained a small proportion of the heritability [23], launched a debate over the ability and efficacy of GWAS to elucidate the genetic architecture of complex diseases [4]. Nevertheless, and despite the early criticisms, GWAS in humans have invariably advanced our understanding of numerous complex diseases and traits [104–106]. The sheer amount of studies in the *GWAS Catalog*—a resource that contains all published studies for humans [107]—attests to the success of GWAS. At the time of writing this chapter, the catalog contained more than 2700 studies on 2000 traits with more than 30,000 associations reported between SNPs and traits [108].

After more than a decade of research in this area, however, the fact remains that most human SNPs reported in GWAS have small effect-sizes and explain a small proportion of the total heritability [109]. In attempting to address one of the potential reasons of the missing heritability, namely the limited sample size, researchers have conducted GWAS on increasingly larger sample sizes. This was achieved either by increasing the number of individuals in a single genome-wide association study, or by combining results of different GWAS through meta-analysis. As a result, the prediction of "larger sample sizes would yield stronger association signals" [4] finally materialized. A case in point are the studies of schizophrenia [110]. The gene ZNF804A was deemed borderline significant in a study of 479 cases and 2937 controls [111], but the association signal of the gene reached statistical significance in a meta-analysis of 21,274 cases and 38,675 controls [112]. Moreover, the number of loci known to be associated to schizophrenia jumped from 30 to 108 after a large collaborative effort based on a meta-analysis of 36,989 cases and 113,075 controls [104]. This has led to new ambitious initiatives like the one at the Centers for Common Disease Genomics (news release about funding: https://www.genome.gov/27563570/) which aims at sequencing up to 200,000 individuals in hope of elucidating the genomic architecture of complex diseases (with an initial focus on autism and cardiovascular diseases).

The slow pace at which results obtained from GWAS have been applied in clinical settings still remains a point of contention and represents an important aspect in need of future improvement. A possible explanation for this delay is related to structural hurdles in the healthcare system. As an example, in the USA there is an average delay of 17 years between the discovery and adoption of best care practices in the clinic [113]. However, there are cases in which GWAS findings could make their way to the clinic, for example: (a) using GWAS results for risk prediction of type 1 diabetes, especially since early detection can avoid the irreversible

depletion of β-cells [114], (b) the repurposing of approved drugs that target genes identified by GWAS, in particular given the costly and lengthy pipeline required to develop new drug targets [115], and (c) the screening of patients, to avoid adverse drug reactions, by using SNPs from GWAS conducted in the field of pharmacogenomics [116].

Although the first GWAS were conducted in humans, the last decade has shown a steady increase of GWAS in model organisms such as *D. melanogaster* [117], *A. thaliana* [15], *C. elegans* [118], *M. musculus* [119], and *Canis lupus familiaris* [120]. The latter in particular, illustrates the success of GWAS in organisms with limited genetic diversity such as dog breeds. The studies of canine diseases whose phenotypes closely resemble those in humans, as it is the case for obsessive-compulsive disorder (OCD), yielded insights into the genetic underpinnings of such diseases in humans. In OCD, for example, statistically significant associations reported in dogs, have functional associations to top SNPs—that did not reach genome-wide significance—in humans [121]. The importance of these developments cannot be overstated, especially because, as opposed to human genetics, model organisms (a) interact with an environment in which the researcher can exert maximum control, and (b) are the ideal vehicle to perform functional validation of candidate genes identified in GWAS. This type of validation, in turn, paves the way for the discovery of where the missing heritability of a trait may lie [122].

In addition to model organisms, GWAS have been widely adopted in crops and livestock. For crops, prominent examples are maize [123], rice [16], tomato [124], and grapevine [125]. In these cases, GWAS have been used in concert with other genetic tools to increase crop yield by making breeding more efficient [126]. Similarly, GWAS in livestock have been instrumental in identifying variants that are associated to economically important traits such as meat quality, milk yield, egg production, and others [127].

Despite a number of successes mentioned above, various challenges in GWAS still lie ahead. One important challenge is, and will continue to be, disentangling the missing heritability in humans. A promising line of research, and one that can potentially lead to a reduction of the missing heritability, is the search for association in high-order SNP interactions. In a genome-wide association study with d SNPs the number of subsets of SNPs that can potentially exhibit association to the phenotype is 2^d. Current approaches limit this vast combinatorial search space either by looking at 2-locus interactions (epistasis) or by superimposing biological networks (network-based approaches). To expand the search space, algorithms that perform pattern mining coupled with sophisticated methods of correction for multiple hypothesis testing have been proposed [128–130]. These methods are based on item-set mining

and do not require a priori information about which items can interact with each other. This, in turn, allows them to potentially consider any number of interacting items. It is important to note that an item, in the context of a genome-wide association study, corresponds to a SNP with a minor allele genotype. In the realm of GWAS, another important aspect to consider is the presence of confounders, such as population structure, that can be prevalent in the data. Pattern mining approaches that can account and correct for confounders are essential for identifying SNPs that are jointly associated to the phenotype while reducing the number of false positives [95]. Future research in this area should focus on (a) enabling the abovementioned methods to handle data that must not necessarily be binary, and (b) scaling the methods to make them applicable to GWAS with large sample sizes and where d is in the order of hundreds of thousands or even millions.

Genome-wide association studies, coupled with the ever-decreasing costs in DNA sequencing and subsequent genotyping are painting a bright future ahead of us. Due to their pervasive of their pervasive use in a wide variety of organisms and traits, understanding the methods, assumptions, and statistical models behind GWAS has become a necessary skill of every researcher working with genomic data.

References

1. MacDonald ME, Novelletto A, Lin C et al (1992) The Huntington's disease candidate region exhibits many different haplotypes. Nat Genet 1:99–103
2. Kerem B-S (1989) Identification of the cystic fibrosis gene: genetic analysis. Trends Genet 5:363
3. Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. PLoS Comput Biol 8:e1002822
4. Visscher PM, Brown MA, McCarthy MI et al (2012) Five years of GWAS discovery. Am J Hum Genet 90:7–24
5. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888
6. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74
7. Gibbs RA, Belmont JW, Hardenbol P et al (2003) The international HapMap project. Nature 426:789–796
8. Davey JW, Hohenlohe PA, Etter PD et al (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510
9. Fan J-B, Chee MS, Gunderson KL (2006) Highly parallel genomic assays. Nat Rev Genet 7:632–644
10. Dudoit S, van der Laan MJ (2008) Multiple hypothesis testing. In: Multiple testing procedures with applications to genomics. Springer, New York, NY, pp 1–47
11. Fairweather D, Frisancho-Kiss S, Rose NR (2008) Sex differences in autoimmune disease from a pathological perspective. Am J Pathol 173:600–609
12. Price AL, Patterson NJ, Plenge RM et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909
13. Atwell S, Huang YS, Vilhjálmsson BJ et al (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631
14. Meijón M, Satbhai SB, Tsuchimatsu T et al (2014) Genome-wide association study using cellular traits identifies a new regulator of root development in Arabidopsis. Nat Genet 46:77–81
15. Alonso-Blanco C, Andrade J, Becker C et al (2016) 1,135 Genomes reveal the global pat-

tern of polymorphism in *Arabidopsis thaliana*. Cell 166:481–491
16. Zhao K, Tung C-W, Eizenga GC et al (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat Commun 2:467
17. Mackay TFC, Richards S, Stone EA et al (2012) The *Drosophila melanogaster* genetic reference panel. Nature 482:173–178
18. Kirby A, Kang HM, Wade CM et al (2010) Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. Genetics 185:1081–1095
19. Scott LJ, Mohlke KL, Bonnycastle LL et al (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341–1345
20. Chasman DI, Schürks M, Anttila V et al (2011) Genome-wide association study reveals three susceptibility loci for common migraine in the general population. Nat Genet 43:695–698
21. Freilinger T, Anttila V, de Vries B et al (2012) Genome-wide assoiation analysis identifies susceptibility loci for migraine without aura. Nat Genet 44:777–782
22. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era — concepts and misconceptions. Nat Rev Genet 9:255–266
23. Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–753
24. Lee SH, Wray NR, Goddard ME et al (2011) Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88:294–305
25. Pedroso I, Breen G (2011) Gene set analysis and network analysis for genome-wide association studies. Cold Spring Harb Protoc 2011:pdb.top065581
26. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11:2463–2468
27. Kam-Thong T, Czamara D, Tsuda K et al (2011) EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. Eur J Hum Genet 19:465–471
28. Kam-Thong T, Azencott C-A, Cayton L et al (2012) GLIDE: GPU-based linear regression for detection of epistasis. Hum Hered 73:220–236
29. Liu JZ, Mcrae AF, Nyholt DR et al (2010) A versatile gene-based test for genome-wide association studies. Am J Hum Genet 87:139–145
30. Listgarten J, Lippert C, Heckerman D (2013) FaST-LMM-select for addressing confounding from spatial structure and rare variants. Nat Genet 45:470–471
31. Lippert C, Xiang J, Horta D et al (2014) Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. Bioinformatics 30:3206–3214
32. Lamparter D, Marbach D, Rueedi R et al (2016) Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput Biol 12:e1004714
33. Jia P, Zheng S, Long J et al (2011) dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics (Oxford, England) 27:95–102
34. Rossin EJ, Lage K, Raychaudhuri S et al (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet 7:e1001273
35. Azencott C-A, Grimm D, Sugiyama M et al (2013) Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics 29:i171–i179
36. Wang Q, Yu H, Zhao Z et al (2015) EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. Bioinformatics (Oxford, England). 31:2591–2594
37. Llinares-López F, Grimm DG, Bodenham DA et al (2015) Genome-wide detection of intervals of genetic heterogeneity associated with complex traits. Bioinformatics 31:i240–i249
38. Buzdugan L, Kalisch M, Navarro A et al (2016) Assessing statistical significance in multivariable genome wide association analysis. Bioinformatics 32:1990–2000
39. Matsuzaki H, Dong S, Loi H et al (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods 1:109–111
40. Clarke GM, Anderson CA, Pettersson FH et al (2011) Basic statistical analysis in genetic case-control studies. Nat Protoc 6:121–133
41. Plomin R, Haworth CMA, Davis OSP (2009) Common disorders are quantitative traits. Nat Rev Genet 10:872–878

42. Power RA, Parkhill J, de Oliveira T (2017) Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet 18:41–50
43. Wu MC, Lee S, Cai T et al (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89:82–93
44. Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615:28–56
45. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5:e1000384
46. Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34:188–193
47. Neale BM, Rivas MA, Voight BF et al (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7:e1001322
48. Anderson CA, Pettersson FH, Clarke GM et al (2010) Data quality control in genetic case-control association studies. Nat Protoc 5:1564–1573
49. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499–511
50. Fisher RA (1925) Statistical methods for research workers. Genesis Publishing Pvt Ltd., Edinburgh
51. Pearson K (1900) X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag Ser 5(50):157–175
52. Fahrmeir L, Kneib T, Lang S et al (2013) Regression: models, methods and applications. Springer Science & Business Media, New York, NY
53. Yang J, Zaitlen NA, Goddard ME et al (2014) Advantages and pitfalls in the application of mixed-model association methods. Nat Genet 46:100–106
54. Loh P-R, Tucker G, Bulik-Sullivan BK et al (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 47:284–290
55. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004
56. Yang J, Weedon MN, Purcell S et al (2011) Genomic inflation factors under polygenic inheritance. Eur J Hum Genet 19:807–812
57. Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60:155–166
58. Lippert C, Listgarten J, Liu Y et al (2011) FaST linear mixed models for genome-wide association studies. Nat Methods 8:833–835
59. Widmer C, Lippert C, Weissbrod O et al (2014) Further improvements to linear mixed models for genome-wide association studies. Sci Rep 4:6874
60. Kang HM, Zaitlen NA, Wade CM et al (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723
61. Zhang Z, Ersoz E, Lai C-Q et al (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360
62. Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44:821–824
63. Veyrieras J-B, Kudaravalli S, Kim SY et al (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4:e1000214
64. Mooney MA, Nigg JT, McWeeney SK et al (2014) Functional and genomic context in pathway analysis of GWAS data. Trends Genet 30:390–400
65. Sedeño-Cortés AE, Pavlidis P (2014) Pitfalls in the application of gene-set analysis to genetics studies. Trends Genet 30:513–514
66. Ballard DH, Cho J, Zhao H (2010) Comparisons of multi-marker association methods to detect association between a candidate region and disease. Genet Epidemiol 34: 201–212
67. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575
68. Listgarten J, Lippert C, Kang EY et al (2013) A powerful and efficient set test for genetic markers that handles confounders. Bioinformatics 29:1526–1533

69. Zuk O, Hechter E, Sunyaev SR et al (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci 109:1193–1198
70. Ueki M, Cordell HJ (2012) Improved statistics for genome-wide interaction analysis. PLoS Genet 8:e1002625
71. Cowley MJ, Pinese M, Kassahn KS et al (2012) PINA v2.0: mining interactome modules. Nucleic Acids Res 40:D862
72. Szklarczyk D, Franceschini A, Kuhn M et al (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39:D561–D568
73. Franceschini A, Szklarczyk D, Frankild S et al (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41:D808–D815
74. Li T, Wernersson R, Hansen RB et al (2017) A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods 14:61–64
75. Johnson RC, Nelson GW, Troyer JL et al (2010) Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics 11:724
76. Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilita. Libreria internazionale Seeber, Firenze
77. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Methodol 57:289–300
78. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188
79. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci 100:9440–9445
80. Thompson JR, Attia J, Minelli C (2011) The meta-analysis of genome-wide association studies. Brief Bioinform 12:259–269
81. Evangelou E, Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14:379–389
82. Stouffer SA, Suchman EA, DeVinney LC et al (1949) The American soldier: adjustment during army life. In: Studies in social psychology in World War II, vol 1. Princeton University Press, Princeton, NJ
83. Borenstein M, Hedges LV, Higgins JPT et al (2010) A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Syn Methods 1:97–111
84. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108
85. Yang J, Lee SH, Goddard ME et al (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82
86. Kang HM, Sul JH, S.K. Service et al (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–354
87. Svishcheva GR, Axenovich TI, Belonogova NM et al (2012) Rapid variance components-based method for whole-genome association analysis. Nat Genet 44:1166–1170
88. de Leeuw CA, Mooij JM, Heskes T et al (2015) MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 11:e1004219
89. Childs LH, Lisec J, Walther D (2012) Matapax: an online high-throughput genome-wide association study pipeline. Plant Physiol 158:1534–1541
90. Seren Ü, Vilhjálmsson BJ, Horton MW et al (2012) GWAPP: a web application for genome-wide association mapping in Arabidopsis. Plant Cell 24:4793–4805
91. Grimm DG, Roqueiro D, Salome P et al (2017) easyGWAS: a cloud-based platform for comparing the results of genome-wide association studies. Plant Cell 29:5
92. Galinsky KJ, Bhatia G, Loh P-R et al (2016) Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. Am J Hum Genet 98:456–472
93. Cormen TH, Leiserson CE, Rivest RL et al (2009) Introduction to algorithms. MIT Press, Cambridge, MA
94. Llinares-López, Papaxanthos L, Bodenham D, Roqueiro D (2017) COPDGene Investigators, Karsten Borgwardt; Genome-wide genetic heterogeneity discovery with categorical covariates. Bioinformatics 33(2): 1820–1828
95. Papaxanthos L, Llinares-Lopez F, Bodenham D et al (2016) Finding significant combinations of features in the presence of categorical covariates. In: Lee DD, Sugiyama M, Luxburg UV et al (eds) Advances in

neural information processing systems, vol 29. Curran Associates, Inc, Red Hook, NY, pp 2271–2279
96. Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1:80–83
97. Horton MW, Hancock AM, Huang YS et al (2012) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat Genet 44:212–216
98. Seymour DK, Chae E, Grimm DG et al (2016) Genetic architecture of nonadditive inheritance in *Arabidopsis thaliana* hybrids. Proc Natl Acad Sci 113:E7317–E7326
99. Seren Ü, Grimm D, Fitz J et al (2017) AraPheno: a public database for *Arabidopsis thaliana* phenotypes. Nucleic Acids Res 45:D1054–D1059
100. McGaughran A, Rödelsperger C, Grimm DG et al (2016) Genomic profiles of diversification and genotype-phenotype association in Island nematode lineages. Mol Biol Evol 33:2257–2272
101. Easton DF, Eeles RA (2008) Genome-wide association studies in cancer. Hum Mol Genet 17:R109–R115
102. Kraft P, Hunter DJ (2009) Genetic risk prediction–are we there yet? N Engl J Med 360:1701–1703
103. Couzin J (2008) DNA test for breast cancer risk draws criticism. Science 322:357–357
104. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological insights from 108 schizophrenia-associated genetic loci. Nature 511:421–427
105. Wood AR, Esko T, Yang J et al (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46:1173–1186
106. Fuchsberger C, Flannick J, Teslovich TM et al (2016) The genetic architecture of type 2 diabetes. Nature 536:41–47
107. Welter D, MacArthur J, Morales J et al (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42:D1001–D1006
108. T. Burdett, P.N. Hall, E. Hastings, et al. The NHGRI-EBI catalog of published genome-wide association studies. www.ebi.ac.uk/gwas.
109. Gusev A, Bhatia G, Zaitlen N et al (2013) Quantifying missing heritability at known GWAS loci. PLoS Genet 9:e1003993
110. Bergen SE, Petryshen TL (2012) Genome-wide association studies (GWAS) of schizophrenia: does bigger lead to better results? Curr Opin Psychiatry 25:76–82
111. O'Donovan MC, Craddock N, Norton N et al (2008) Identification of loci associated with schizophrenia by genome-wide association and follow-up. Nat Genet 40:1053–1055
112. Williams HJ, Norton N, Dwyer S et al (2011) Fine mapping of ZNF804A and genome wide significant evidence for its involvement in schizophrenia and bipolar disorder. Mol Psychiatry 16:429–441
113. Richardson WC, Berwick DM, Bisgard J et al (2001) Crossing the quality chasm: a new health system for the 21st century. Institute of Medicine, National Academy Press, Washington, DC
114. Manolio TA (2013) Bringing genome-wide association findings into clinical use. Nat Rev Genet 14:549–558
115. Lencz T, Malhotra AK (2015) Targeting the schizophrenia genome: a fast track strategy from GWAS to clinic. Mol Psychiatry 20:820–826
116. Chan SL, Jin S, Loh M et al (2015) Progress in understanding the genomic basis for adverse drug reactions: a comprehensive review and focus on the role of ethnicity. Pharmacogenomics 16:1161–1178
117. Huang W, Massouras A, Inoue Y et al (2014) Natural variation in genome architecture among 205 *Drosophila melanogaster* genetic reference panel lines. Genome Res 24:1193–1208
118. Andersen EC, Gerke JP, Shapiro JA et al (2012) Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat Genet 44:285–290
119. Farber CR, Bennett BJ, Orozco L et al (2011) Mouse genome-wide association and systems genetics identify Asxl2 as a regulator of bone mineral density and osteoclastogenesis. PLoS Genet 7:e1002038
120. Hayward JJ, Castelhano MG, Oliveira KC et al (2016) Complex disease and phenotype mapping in the domestic dog. Nat Commun 7:10460
121. Tang R, Noh HJ, Wang D et al (2014) Candidate genes and functional noncoding variants identified in a canine model of obsessive-compulsive disorder. Genome Biol 15:R25

122. Flint J, Eskin E (2012) Genome-wide association studies in mice. Nat Rev Genet 13:807–817
123. Li H, Peng Z, Yang X et al (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45:43–50
124. Lin T, Zhu G, Zhang J et al (2014) Genomic analyses provide insights into the history of tomato breeding. Nat Genet 46:1220–1226
125. Nicolas SD, Péros J-P, Lacombe T et al (2016) Genetic diversity, linkage disequilibrium and power of a large grapevine (Vitis vinifera L) diversity panel newly designed for association studies. BMC Plant Biol 16:74
126. Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol 65:531–551
127. Sharma A, Lee JS, Dang CG et al (2015) Stories and challenges ofgenome wide association studies in livestock — a review. Asian Australas J Anim Sci 28: 1371–1379
128. Terada A, Okada-Hatakeyama M, Tsuda K et al (2013) Statistical significance of combinatorial regulations. Proc Natl Acad Sci 110:12996–13001
129. Minato S, Uno T, Tsuda K et al (2014) A fast method of statistical assessment for combinatorial hypotheses based on frequent itemset enumeration. In: Machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, pp 422–436
130. Llinares-López F, Sugiyama M, Papaxanthos L et al (2015) Fast and memory-efficient significant pattern mining via permutation testing. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Sydney, NSW, pp 725–734