Recent Developments in Statistical Methods for GWAS and High-throughput Sequencing Association Studies of Complex Traits

Duo Jiang[a*] and Miaoyan Wang[b]

[a]Department of Statistics, Oregon State University Corvallis, OR 97330, USA
jiangd@stat.oregonstate.edu; 1 (541)737-1989

[b]Department of Statistics, University of Wisconsin–Madison
Madison, WI 53706, USA

*Address for correspondence:
Duo Jiang
Department of Statistics
Oregon State University
239 Weniger Hall
Corvallis, OR 97330
Phone : (541) 737-1989 Fax : (541) 737-3489
jiangd@stat.oregonstate.edu

Recent Developments in Statistical Methods for GWAS and High-throughput Sequencing Studies of Complex Traits

ARTICLE HISTORY
Compiled July 4, 2018

ABSTRACT
The advent of large-scale genetic studies has helped bring a new era of biomedical research on dissecting the genetic architecture of complex human disease. Genome-wide association studies (GWASs) and next-generation sequencing studies are two popular tools for identifying genetic variants that are associated with complex traits. This article overviews some of the most important statistical tools for analyzing data from these two types of studies, with an emphasis on single-SNP tests for common variants and region-based tests for rare variants. We compare various statistical methods for common and rare variants in humans, and describe some critical considerations to guide the choice of an analysis method. Also discussed are the related topics of sample ascertainment, missing heritability, and multiple testing correction, as well as some remaining analytical challenges presented by complex trait association mapping using genomic data obtained via high-throughput technologies.

KEYWORDS
Genome-wide association studies; high-throughput sequencing; complex traits; association analysis; statistical tests

1. Introduction

Genetic association analysis is a widely-used tool for dissecting the genetic basis of human diseases and complex traits. Recent advances in high-throughput genotyping and sequencing allow the identification of disease susceptibility variants in exquisitely fine detail. A genome-wide association study (GWAS) examines a high-density set of genetic markers and performs an unbiased, genome-wide search of association signals. This pipeline has led to many exciting findings of common genes underlying complex human diseases, such as asthma, diabetes, cardiovascular disease, and psychiatric illness [34,84,97,131]. More recently, whole-genome sequencing (WGS) collects human genetic sequence variations across the entire spectrum of allele frequencies, thereby enabling the investigation of rare variants that are usually missed in a GWAS [35].

Data from genetic association studies often feature complex dependence structure and high-dimensionality. With the large number of genetic variations in the genome, it is a challenging task to distinguish the true causal polymorphisms from the background noise or confounding effects. Developing powerful, scalable statistical tools that exploit the biological implications in these large-scale datasets is of great importance to both the statistics and the genetics communities. Despite a large body of literature, there is still little consensus on a single most appropriate statistical procedure for genetic association analysis.

In this article, we survey recent advances in statistical methods for genetic associa-

tion studies, illustrating the challenges presented by genomic data obtained via high-throughput technologies. Our emphasis is on association testing methods for common and rare variants in humans. In Section 2, we outline the key components of a GWAS analysis. In Section 3, we focus on recently developed statistical methods for association testing of common genetic variants. We discuss several approaches to accounting for confounding factors and to testing for association with a quantitative-trait and a binary trait. Also discussed are the related topics of case-control ascertainment, retrospective analysis and missing heritability. In Section 4, we review recently developed statistical methods for rare-variant association testing in high-throughput sequencing studies and some analytical issues that arise.

2. Overview of GWAS Analysis

The primary goal of GWASs is to identify genetic variants that contribute towards the phenotypic variation of complex traits. A GWAS uses chip arrays to type hundreds of thousands of single-nucleotide polymorphisms (SNPs) across a large number of individuals, and then assesses the correlation between SNP genotypes and the trait of interest. A typical genome-wide association analysis involves at least the following three broadly defined steps: (i) data quality control; (ii) association testing; (iii) results interpretation. In this section we briefly introduce these steps, with a more detailed discussion of step (ii) deferred to Section 3.

2.1. *Quality Control for Genomic Data*

Quality control (QC) usually involves filtering out (i.e., removing) SNPs with low genotyping accuracy. QC is an important step to minimize potential false findings in GWASs. Common SNP filters include missing call rate (MCR), minor allele frequency (MAF), and Hardy-Weinberg equilibrium (HWE) [92]. These QC filters are informative indicators of genotyping quality: extreme deviation from HWE could reflect genotyping error [117]; high rate of missingness suggests poor genotype probe performance [80]; SNPs with low MAF are more prone to genotyping error as many calling algorithms perform poorly with rare alleles [92].

Genotype imputation is often carried out in GWASs to allow better use of the typed SNPs. Using external resources such as HapMap [22] and data from the 1000 Genomes Project [21], one can impute the unmeasured genotypes based on known linkage disequilibrium (LD) structure and haplotype frequencies. For tightly linked markers, genotype imputation can be reasonably reliable. After imputation, an additional quality control step is often required to remove SNPs with low imputation certainty. In principle, imputed SNPs should be analyzed separately from genotyped SNPs [92], because uncertainty in imputation needs to be accounted for when performing association tests. There are several methods specifically designed for testing association with imputed SNPs, which exploit their posterior genotype probabilities [8,62,72,104]. In this article, our main focus will be on the analysis of genotyped SNPs only.

2.2. *Association Testing*

After the completion of QC, statistical analysis is performed to detect the association between a SNP and a trait. GWASs are primarily based on the common disease–

common variants (CDCV) hypothesis [35,124], which postulates that the complex disease is largely caused by common genetic variants with moderate effects, each of which explains a certain proportion of the phenotypic variance. Under the CDCV hypothesis, the most popular strategy for identifying associations is to conduct a series of single-SNP association tests, in which each SNP is tested for association separately from the other SNPs with a given trait. A proper choice of association tests should take into account many factors, such as the possible population and family structure in the sample, the type of the trait variable, presence of ascertainment in the sampling design, etc. In Section 3, we will survey current statistical methods to tackle these challenges.

2.3. *Reporting Association Results*

Once a statistical test is chosen and performed in the previous step, each SNP will produce a test statistic measuring its association with the trait of interest and a p-value measuring the statistical significance. Manhattan and quantile-quantile (Q-Q) plots are useful tools for visualizing GWAS results and for model diagnostics. A Manhattan plot is a scatter plot showing the levels of statistical significance by chromosomal locations. SNPs in the entire GWAS analysis are laid out on the x-axis in genomic order based on which chromosome a SNP belongs to and its location on the chromosome, and the y-axis represents the p-value of each SNP on the negative logarithm scale. Visual inspection of the peaks in a Manhattan plot facilitates the detection of genomic regions with strong association signals. A Q-Q plot is another commonly examined graphical representation of GWAS results, which shows the empirical distribution of the observed p-values against the theoretical distribution under the null hypothesis of no association. In practice, a vast majority of SNPs are expected to be unassociated with the trait, so the bulk of the points should fall on or close to the $y = x$ line (called the reference line) until the end (Fig. 1A). A global deviation from the reference line (Fig. 1B) usually indicates inadequate control of population and/or family structure or the presence of other confounding factors.

In GWASs, a large number of hypothesis tests are carried out, and this leads to a multiple testing problem. Using a 5% significance threshold, we would expect to incorrectly reject the null $.05N$ times, where N is the number of tests performed across the genome. For a typical GWAS, N ranges from hundreds of thousands to over a million, which would lead to a daunting number of false discoveries if the 5% significance threshold is imposed on nominal p-values. To control the genome-wide error rate, a widely used solution is to use the Bonferroni procedure: based on an estimate of one million independent SNPs across the human genome, the Bonferroni correction at level 0.05 yields the significance threshold 5×10^{-8}, known as the "genome-wide significance level."

This Bonferroni correction is usually considered conservative, and other alternative solutions to the multiple testing issue have been explored, including false discovery rate (FDR) [13] and permutation procedures [132]. FDR estimates the proportion of false positives among the tests that are declared as significant, and an FDR-controlling procedure will result in fewer false negatives. Permutation is another approach to significance assessment in a GWAS. While somewhat computationally intensive, permutation is a flexible and robust way to generate the empirical null distribution of test statistics while taking into account the LD patterns among SNPs [1]. We note that research on these alternatives approaches is still in its infancy, and the genome-wide

significance threshold of 5×10^{-8} remains a commonly agreed-upon criterion adopted by most studies.

3. Single-SNP Association Testing

In this section, we will discuss some challenges for single-SNP association testing in GWAS data and review some recently developed methods to tackle these challenges.

3.1. *Controlling Sample Structure Confounding*

In genetic association mapping, a common confounding factor is sample structure, which is a term referring to family or ancestral relationships within a sample that are due to both known and unknown structure. Various forms of sample structure are widespread in genetic association studies, including population stratification/admixture, family relatedness and cryptic relatedness. In the presence of sample structure, the independence assumption made by many standard statistical techniques may break down, leading to severely compromised performance and reduced reliability of the tests.

Many genetic studies include family members with known pedigree relationships. Family-based designs have long been popular in traditional genetic studies, and the samples collected for those studies are often included in current association analysis. Moreover, including family members can increase the power of detecting association due to enrichment of disease-associated SNPs among relatives. With related individuals, it is well known that dependence resulting from family structure needs to be accounted for to ensure that association tests have properly controlled type 1 error [85]. Carefully adjusting for familial correlation can also improve power [118].

In addition to pedigree correlation, another well-known source of sample structure concerns latent relatedness among sampled individuals that is not due to known family relationships. Seen from evolutionary history, all members of the human race are mutually related to a varying extent through a giant genealogy, although the underlying genealogical structure is usually unobserved for a sample except in pedigree-based studies. When some pairs of individuals are more closely related than others in a sample, the heterogeneity in the amount of relatedness can give rise to confounding in association testing. Intuitively, relatedness among individuals introduces correlation not only to the observed genotypes on the tested SNP (or group of SNPs), but also to the overall genome-wide variation (known as genetic background), which may in turn produce correlation in the phenotype of interest. This simultaneous correlation with the genotype and the phenotype can act as a confounder, which, if not properly accounted for, creates deviation from the null hypothesis (of no genetic association) across the genome and leads to genomic inflation of association signals.

One example of latent relatedness is population stratification, which arises when the sample includes individuals from multiple population subgroups. Distinct subpopulations often have their own distinctive genetic backgrounds shared by individuals from the same ancestral group, which leads to many traits being correlated with ancestry. Meanwhile, the ancestry differences also result in allele frequency differences between subpopulations. This gives rise to the situation of confounding, in which the ancestry (confounder) correlates with both the genotype (predictor) and the phenotype (response). For example, in case-control studies, association tests can be roughly considered a comparison of genotype distribution between phenotype groups. When both

phenotypic and genotypic distributions vary by subpopulation, genetic variants that are not directly associated with the trait can generate spurious association signals if population substructure is not properly corrected for. Similar confounding effects can occur with admixed populations, in which each individual is genetically a mixture of multiple ancestral populations with the mixing proportions varying across the sample. Another type of hidden sample structure is termed cryptic relatedness, defined to be family relationships not explicitly observed in the sample.

In genetic studies, population structure and family relationships may or may not be known. For samples with fully known family structures, the pedigree-based kinship matrix may be used to adjust for familial correlation [2,85]. To deal with unobserved sample structure, a useful strategy is to utilize the genetic variants themselves to infer the hidden structure among the sampled individuals. Most existing methods fall into two distinctive categories. The first class of methods consider population membership in stratified samples or ancestry proportions in admixed samples as unobserved covariates. These methods usually involve two steps, the first of which tries to explicitly reconstruct the unmeasured variables based on ancestry informative SNPs or genome-wide data. In the second step, the inferred ancestry is included in a regression model as one or more fixed effects to correct the test statistics for sample structure. A well-known approach is principal component analysis (PCA), which identifies ancestry differences among individuals using genome-wide SNP data. Price et al. [94] developed EIGENSTRAT, a method that uses the top principal components (PCs) of the genetic covariance matrix as surrogates for ancestry which are included as covariates in a regression-based association model. Examples of other methods in this category include a model-based clustering tool for detecting population structure [96], methods related to PCA [15,144,150], and various other methods [4,59,143]. This approach is effective in adjusting for confounding if population stratification is the only kind of structure present in the sample. However, these methods suffer from the drawback of not being able to handle samples containing additional complexity such as (known and/or unknown) related individuals [95].

The second category of methods view the hidden sample structure as producing dependence among observations. These methods therefore try to estimate the dependence structure and model it as correlation among sampled individuals. A type of method within this category that has recently gained much popularity is the linear mixed model (LMM) approach. LMM models the phenotypic distribution using a mixture of fixed effects and random effects. The fixed effects include the genetic variant being tested and other relevant covariates, such as age and sex, that need to be adjusted for. The random effects include additive polygenic effects whose covariance matrix is assumed to be proportional to a genetic relatedness matrix (GRM). The GRM is usually estimated from genome-wide SNP data intended to capture the overall covariance among individuals due to population structure (ancestry difference), family structure, and cryptic relatedness. Because of the flexibility to simultaneously account for various types of sample structure including cryptic relatedness as well as population stratification and admixture, LMM has emerged as the method of choice in genetic association analysis.

One of the major differences between the PCA-based and LMM-based methods is whether to treat population structure as a low-dimensional fixed effect or part of a high-dimensional random effect. Recent work [43] reveals that the structure matrices caused by population stratification/admixture (ancestry difference) and caused by kinship (family or cryptic relatedness) have distinguished properties: population stratification is a low-dimensional process embedded in a high-dimensional space so

a small number of PCs are adequate to capture the structure, whereas kinship is a high-dimensional/full-rank process which cannot be captured by a small number of PCs. This provides a rationale for the out-performance of LMM over PCA when the sample contains family or cryptic relatedness. On the other hand, modeling population structure as part of a random effect might lead to insufficient correction at SNPs having strong ancestry differentiation. This observation gives rise to some hybrid methods combining the strengths of PCA and LMM. For example, Thornton et al. [120] proposed to simultaneously estimate population-structure PCs and pairwise kinship coefficients. Then, an LMM is fitted by including population-structure PCs to account for ancestry and by including pairwise kinship estimates to account for family relatedness [20]. This hybrid approach demonstrates better control of type 1 error inflation for highly differentiated SNPs compared to the standard LMM [20,95].

3.2. *Linear Mixed Models for Quantitative-Trait Analysis*

A general paradigm for testing for the association between a phenotype of interest and the genotype at a SNP is regression analysis. Ordinary linear models, as we have argued in Section 3.1, are vulnerable to spurious associations in the presence of population and family structure. Here we present the commonly used mixed-effect model primarily designed for association analysis on a quantitative trait.

Consider a sample of n individuals, and let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ be the phenotype vector on the sample. Let $\boldsymbol{W}$ be an $n \times k$ covariate matrix encoding in its columns k covariates including the intercept term. Also let $\boldsymbol{G}^{\text{test}} = (G_1^{\text{test}}, \ldots, G_n^{\text{test}})^T$ be the genotype vector. The standard LMM for a quantitative trait takes the form

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{G}^{\text{test}}\gamma + \boldsymbol{\varepsilon}, \quad \text{where} \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma_a^2\boldsymbol{\Phi} + \sigma_e^2\boldsymbol{I}), \tag{1}$$

where $\boldsymbol{W}\boldsymbol{\beta}$ represents the fixed effects of covariates including the intercept, $\boldsymbol{G}^{\text{test}}\gamma$ is the effect of the SNP currently being tested, σ_a^2 and σ_e^2 are variance component parameters corresponding to additive polygenic effects and i.i.d. environmental errors, respectively, and $\boldsymbol{\Phi}$ is a GRM quantifying the overall genetic similarity between any pair of individuals in the sample. The matrix $\boldsymbol{\Phi}$ is assumed to be either known based on the pedigree or can be estimated (discussed in the next paragraph). To test a biallelic SNP, G_i^{test} is often encoded as 0, 1, or 2, according to whether the individual i has 0, 1, or 2 copies of the minor allele at the SNP. It is common to assume an additive allele effect; that is, γ is a scalar association parameter of interest. Then testing for genetic association amounts to the hypothesis test

$$H_0 \colon \gamma = 0 \quad \text{vs.} \quad H_\alpha \colon \gamma \neq 0.$$

In the context of genetic association analysis, a number of statistical tests, including Wald, Rao's score and likelihood ratio tests, have been proposed.

The matrix $\boldsymbol{\Phi}$ in (1) is designated to model the genetic covariance among individuals due to known or unknown sample structure that is not captured by $\boldsymbol{W}$. When complete pedigree information is available, a commonly used option for $\boldsymbol{\Phi}$ is the pedigree-based kinship matrix [2]. When $\boldsymbol{\Phi}$ is obtained from pedigrees, one implicitly assumes that the pedigree information is reliable and the founders are independently drawn from a single population [2]. For samples containing cryptic relatedness and/or hidden population structure, $\boldsymbol{\Phi}$ can be estimated using genotype data at a large number of SNPs across

the genome. Kang et al. [53] proposed an empirical GRM $\hat{\mathbf{\Phi}}$, where the (i, j)th entry of the matrix is estimated by

$$\hat{\mathbf{\Phi}}_{ij} = \frac{1}{S} \sum_{s=1}^{S} \frac{(G_i^s - 2\hat{p}_s)(G_j^s - 2\hat{p}_s)}{2\hat{p}_s(1 - \hat{p}_s)},$$

where $s = 1, \ldots, S$ indexes the SNPs being considered across the genome, G_i^s and G_j^s are the genotypes of individuals i and j at SNP s, and $\hat{p}_s$ is the sample average estimator of the minor allele frequency of SNP s. The use of this empirical GRM in LMM has been demonstrated to be effective in correcting for a variety of types of population and family structure in association analysis.

The LMM has historically been viewed as a theoretically appealing but computationally demanding approach [33]. In a typical GWAS, model (1) needs to be analyzed for millions of different $\boldsymbol{G}^{\text{test}}$'s, one for each SNP. Owning to the recent development of efficient algorithms, LMM has now become feasible for large cohort studies with up to half a million individuals [67]. A number of LMM software packages have been developed for large-scale GWASs (see Table 1). Algorithms such as EMMAX [53] and GRAMMAR-Gamma [115] assume constant variance component parameters σ_a^2 and σ_e^2 across tested SNPs, and estimate them under the null only once per genome-wide scan. This approximation seems reasonable, at least at the initial stage of analysis, considering that most SNP effects are small. Other algorithms such as GEMMA [147] and FaST-LMM [64] implement an exact analysis which re-estimates the variance component parameters for each tested SNP. In addition, different software packages also vary in their strategies for constructing the empirical GRM: GCTA-LOGO [140] performs a leave-one-chromosome-out (LOCO) scheme to avoid proximal contamination, in which only SNPs not located on the same chromosome as the tested SNP are included in the calculation of the GRM; FaST-LMM [64] selects a subset of SNPs through FaST-LMM-Select [65] for GRM estimation. Most packages separate GRM estimation from the association testing, thus allowing users to read in an externally estimated GRM as desired. A recent study [30] compared the performance of different LMM-based methods, including GTAM [2], EMMAX [53], GRAMMAR-Gamma [115], FaST-LMM [64], GEMMA [147] and MASTOR [49], in GWASs with related individuals. The results show a strong concordance in the association signals across different packages, suggesting that the software choice may be more subject to computational considerations such as speed and memory usage.

3.3. *Binary-Trait Association Analysis*

3.3.1. Generalized Linear Mixed Models and the Quasi-likelihood Approach

We now turn our attention to association analysis on a binary trait. Many human GWASs are conducted on traits that come in the form of a binary variable, e.g., the presence or absence of a specific disease. A binary trait takes on two possible values, $Y_i \in \{0, 1\}$. Unlike a normal random variable, a binary random variable has its mean restricted between 0 and 1 and admits a particular relationship between its mean and variance specified by $\sigma_i^2 = \mu_i(1 - \mu_i)$. To accommodate these features, a generalized linear model (GLM) seems a natural option when the sample consists of only unrelated

individuals in the absence of population structure,

$$\begin{aligned} &Y_i|\boldsymbol{W},\boldsymbol{G} \sim \text{Bernoulli}(\mu_i), \text{ independently,} \\ &\text{logit}(\mu_i) = \boldsymbol{W}_i\boldsymbol{\beta} + \boldsymbol{G}_i^{\text{test}}\gamma, \ i = 1,\ldots,n, \end{aligned} \tag{2}$$

where μ_i is the mean of Y_i conditional on $\boldsymbol{W}$ and $\boldsymbol{G}$, $\boldsymbol{W}_i\boldsymbol{\beta}$ represents the covariate effects for individual i, γ is the association parameter of interest, and $\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$ is the link function. In general, the logit link can be replaced by other link functions such as probit. The use of logit link enables the interpretation of the model coefficients in terms of log odds ratios, whereas the probit link corresponds to the classical liability-threshold model [133].

When the sample contains population structure and family relatedness, the association analysis based on model (2) is sensitive to unmeasured confounding factors. Given the success of LMMs in quantitative-trait analysis, current GWASs frequently fit LMMs (1) to binary traits. This is in the spirit of Armitage's test [5], in which the binary trait $\boldsymbol{Y}$ is treated as if it were quantitative. However, the use of linear models for binary traits relies on a generally invalid assumption that ignores the intrinsic heteroscedasticity in the binary data caused by the variance-to-mean relationship $\sigma_i^2 = \mu_i(1-\mu_i)$. In the presence of important covariates such as population stratification, the homoscedasticity assumption in LMMs is violated [50,83]. As a consequence, fitting LMMs to binary traits could lead to inflated type 1 error and reduced power in the association testing.

A statistically more justified modeling option is to incorporate random effects into the GLM given in (2) in order to account for the polygenic effects resulting from sample structure. The generalized linear mixed model (GLMM) combines GLM and LMM by including a random effect in the mean model,

$$\begin{aligned} &Y_i|\boldsymbol{W},\boldsymbol{G},\boldsymbol{\eta} \sim \text{Bernoulli}(\mu_i), \text{ independently,} \\ &\text{with } \text{logit}(\mu_i) = \boldsymbol{W}_i\boldsymbol{\beta} + \boldsymbol{G}_i^{\text{test}}\gamma + \eta_i \quad \text{with } \boldsymbol{\eta} = (\eta_1,\ldots,\eta_n)^T \sim N(\mathbf{0}, \sigma_a^2\boldsymbol{\Phi}), \end{aligned}$$

where $\boldsymbol{\eta}$ is the random effect due to the additive polygenic effects, σ_a^2 is the variance of the random effect, and $\boldsymbol{\Phi}$ is the GRM as in Section 3.2. Although GLMM is well justified for modeling a binary trait, its use in large-scale GWAS is limited due to high computational cost. Fitting a GLMM usually involves an unstable and often intractable high-dimensional integration over the distribution of the random effects. Recent attempts to overcome this problem approximate this integral by introducing some regularity or stochasticity in the fitting algorithm. For example, GMMAT [17] takes a penalized log-likelihood approach to obtain the shrinkage estimators in a flexible and fast algorithm, and GLOGS [108] adopts a sampling-importance-resampling approach to approximate the MLE estimators. Both software packages perform score-based association tests and are applicable to large-scale binary-trait GWASs. A potential drawback of these algorithms is that the approximation necessary for speeding up computations may lead to compromised estimation accuracy. For example, in the statistics literature, it is known that penalized methods may yield estimates that are biased towards zero when used on binary data [101].

In contrast to GLMM, the quasi-likelihood model is another approach to modeling binary traits in samples subject to population or family structure. It specifies only the first two moments of $\boldsymbol{Y}|(\boldsymbol{W},\boldsymbol{G})$ rather than a full probability distribution. Because the random effects are implicitly integrated out in the model, the quasi-likelihood

approach bypasses the need for high-dimensional integration. In the context of association analysis, the quasi-likelihood model is expressed as

$$\begin{aligned}\text{logit}(\mu_i) &= \boldsymbol{W}_i\boldsymbol{\beta} + \boldsymbol{G}_i^{\text{test}}\gamma,\ i = 1, \ldots, n \\ \text{Var}(\boldsymbol{Y}|\boldsymbol{W}, \boldsymbol{G}) &= \boldsymbol{M}^{1/2}\boldsymbol{\Sigma}\boldsymbol{M}^{1/2},\ \text{with } \boldsymbol{\Sigma} = \xi\boldsymbol{I} + (1-\xi)\boldsymbol{\Phi},\end{aligned}$$

where μ_i is the mean of Y_i conditional on $\boldsymbol{W}$ and $\boldsymbol{G}$, $\boldsymbol{M}^{1/2}$ denotes an $n \times n$ diagonal matrix with ith diagonal element $\sqrt{\mu_i(1-\mu_i)}$, γ is the association parameter of interest, $\xi \in [0, 1]$ is a variance component parameter, and $\boldsymbol{\Phi}$ is the GRM as before. Note that $\boldsymbol{\Sigma}$ is pre- and post-multiplied by the diagonal matrix $\boldsymbol{M}^{1/2}$ to respect the Bernoulli variance $\mu_i(1-\mu_i)$. In the quasi-likelihood framework, parameter estimation does not require the maximization of a fully specified likelihood function, which can be computationally burdensome, but instead can be achieved efficiently by solving estimating equations [52,127,145]. This feature makes quasi-likelihood an appealing approach to modeling non-Gaussian data with complex dependent structure.

3.3.2. Case-Control Ascertainment

Non-random ascertainment arises frequently in genetic studies, and it needs special attention for binary-trait association analysis. In binary-trait GWASs, there are two common sampling designs: (i) case-control studies where individuals are sampled on the basis of the phenotype (case-control ascertainment) or on the basis of the phenotype and clinical covariates (case-control-covariate ascertainment); (ii) prospective cohort studies where individuals are randomly sampled from the base population prior to the onset of disease/disorder. GWASs for low-prevalence diseases often adopt a case-control design wherein affected individuals (cases) are over-sampled relative to the disease prevalence. Compared to a prospective cohort design, a case-control design poses additional challenges to association modeling. In general, the joint distribution of $(\boldsymbol{Y}, \boldsymbol{G}, \boldsymbol{W})$ in a case-control ascertained sample differs from what it would be in a simple random sample [50]. For example, unequal case-control sampling ratios across subpopulations may introduce population confounding, even if disease prevalence is the same in all subpopulations [17]. Ascertainment can also create a spurious correlation between high-risk genotypes and high-risk covariates in case-control samples due to the enrichment of both risk genotypes and covariates in the cases [74,91].

Recently, several methods have been proposed in the attempt to allow case-control ascertainment in binary-trait GWASs. LEAP [130] and LTMLM [41] account for over-sampling of cases by using an externally obtained disease prevalence to adjust the heritability estimate. Both methods fit a liability threshold linear mixed model to the case-control data and proceed to test for association using latent liability estimates. However, as pointed out by the authors, the presence of related individuals in the sample could lead to biased liability estimates. LT-Fam [41] is an extension of LTMLM to the family-based case-control ascertainment, and it chooses to use published heritability estimates to avoid the biased heritability estimation in the liability scale. Instead, GMMAT [17] and CARAT [52] fit logistic mixed models to the case-control data. These two methods do not require the specification of a disease prevalence and are generally applicable to samples with family and population structure including population stratification caused by unequal case-control ratio across subpopulations. To guard against the case-control ascertainment, CARAT takes one step further: it uses a retrospective model to calibrate the association test statistic. As we will see in

the next section, retrospective analysis brings a number of advantages over the standard prospective analysis, including robustness to case-control ascertainment and to phenotype model misspecification.

3.4. *Retrospective vs. Prospective Analysis*

Standard association methods model the phenotype as a response variable while using genotypes and covariates as predictors. This approach is termed a prospective analysis, as opposed to a retrospective analysis that we will describe later. The prospective model is biologically easier to interpret as it describes how the phenotype is influenced by the genetic marker and other covariates for individuals randomly sampled from a population. When statistical inference is based on a phenotype model, careful and accurate specification of the phenotype distribution is often required in order to produce well-calibrated and powerful association tests.

However, in practice, it is typically unknown what the underlying distribution of the phenotype is and how it relates to the covariates. Consequently, phenotype model misspecification is ubiquitous. This includes, for example, failure to include important non-genetic effects, neglecting epistasis or gene-environment interactions, ignoring one or more variance components, and phenotype-based ascertainment in the sampling design. Performing statistical inference based on a misspecified model can lead to improper control of type 1 error and/or reduced power. In fact, several studies have shown that the standard (prospective) LMM produced inflated test statistics if additional important variance components were missed out in the model [110,122].

A possible remedy for phenotype model misspecification is to use a retrospective model for association analysis; that is, we treat genotype as a response variable and model the distribution of the genotype conditional on the phenotype and covariates. While, in theory, model misspecification can lead to compromised performance in both retrospective and prospective methods, it tends to be less of a concern for retrospective analysis due to two reasons [50]: First, it has been shown that a correctly specified prospective model for $\boldsymbol{Y}|(\boldsymbol{W}, \boldsymbol{G}^{\text{test}})$ in an ascertained sample becomes misspecified with either case-control ascertainment or case-control-covariate ascertainment, whereas a retrospective model for $\boldsymbol{G}^{\text{test}}|(\boldsymbol{Y}, \boldsymbol{W})$ is unaffected; Second, it can be argued that, the phenotype distribution is intrinsically more challenging to correctly specify than the genotype distribution, because Mendelian inheritance is highly informative about how the genotypes may be distributed under the null. Recent work has established the connection between the prospective and retrospective models. Under suitable assumptions, the association parameter of interest can be identified in the conditional mean of $\boldsymbol{G}^{\text{test}}|(\boldsymbol{W}, \boldsymbol{Y})$. For example, the retrospective version of LMM given in Equation (1) models the genotype $\boldsymbol{G}^{\text{test}}$ as a random drawn from a distribution with null covariance $\text{Var}_0(\boldsymbol{G}^{\text{test}}|\boldsymbol{W}, \boldsymbol{Y}) = \sigma_s^2\boldsymbol{\Phi}$ and mean

$$\mathbb{E}(\boldsymbol{G}^{\text{test}}|\boldsymbol{W}, \boldsymbol{Y}) = p_s\mathbf{1} + \delta\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\beta}), \tag{3}$$

where p_s is the allele frequency and σ_s^2 is the variance of $\boldsymbol{G}^{\text{test}}$ for an outbred individual. In this case, we wish to test the null hypothesis $\delta = 0$. (See [49,126] for more details on justification.) Although obtaining the full distribution of $\boldsymbol{G}^{\text{test}}$ is not straightforward, the first two moments are enough to construct a quasi-likelihood score test statistic for δ. The retrospective formulation for binary-trait association modeling has been proposed similarly [107,118,119].

Retrospective modeling is an appealing approach both in theory and in practice.

Since retrospective inference is conditional on the phenotype, the resulting test statistic is robust to phenotype model misspecification and phenotype-based or phenotype-and-covariate-based ascertainment [49]. This feature is extremely useful in binary-trait association studies in which non-random ascertainment is a common practice. In addition, the score statistics derived from (3) have the same calibration factor across SNPs except for a scalar multiplier $\hat{\sigma}_s^2$. Making use of this fact facilitates the efficient implementation of a genome-wide association scan that scales linearly in the number of individuals and linearly in the number of SNPs [52,67,115].

In the context of score tests for association, parameter estimation is necessary only under the null hypothesis of no association. This motivates another class of association tests obtained by combining a prospective model with a null retrospective model, including MASTOR [49], CARAT [52] and CERAMIC [145]. To illustrate the idea of this approach, we start by noting that a number of prospective association statistics (such as GTAM [2] and GRAMMAR-Gamma [115]) can be written as

$$T_{\text{pro}} = \frac{\left(\boldsymbol{Y}^T\boldsymbol{V}^{-1}\boldsymbol{G}^{\text{test}}\right)^2}{\text{Var}_{0,\boldsymbol{Y}}(\boldsymbol{Y}^T\boldsymbol{V}^{-1}\boldsymbol{G}^{\text{test}})}, \tag{4}$$

where $\boldsymbol{V}^{-1}$ is a certain choice of matrix that does not depend on $\boldsymbol{Y}$ or $\boldsymbol{G}^{\text{test}}$, and $\text{Var}_{0,\boldsymbol{Y}}$ denotes the null variance taken with respect to $\boldsymbol{Y}$ under the model (1). In the light of retrospective modeling, one could replace the calibration factor in the denominator, $\text{Var}_{0,\boldsymbol{Y}}(\boldsymbol{Y}^T\boldsymbol{V}^{-1}\boldsymbol{G}^{\text{test}})$, by $\text{Var}_{0,\boldsymbol{G}^{\text{test}}}(\boldsymbol{Y}^T\boldsymbol{V}^{-1}\boldsymbol{G}^{\text{test}})$, where $\text{Var}_{0,\boldsymbol{G}^{\text{test}}}$ denotes the null variance taken under a null model for $\boldsymbol{G}^{\text{test}}$ conditional on $\boldsymbol{Y}$ and $\boldsymbol{W}$. For example, the following null genotype model [119] can be used,

$$\mathbb{E}_0(\boldsymbol{G}^{\text{test}}|\boldsymbol{W},\boldsymbol{Y}) = p_s\mathbf{1}, \text{ and } \text{Var}_0(\boldsymbol{G}^{\text{test}}|\boldsymbol{W},\boldsymbol{Y}) = \sigma_s^2\boldsymbol{\Phi}, \tag{5}$$

where $\boldsymbol{\Phi}$ is a GRM assumed to be the same across SNPs whereas the scalar multiplier σ_s^2 is allowed to vary across SNPs. Combining (4) and (5) leads to a retrospective association statistic

$$T_{\text{retro}} = \frac{\left(\boldsymbol{Y}^T\boldsymbol{V}^{-1}\boldsymbol{G}^{\text{test}}\right)^2}{\text{Var}_{0,\boldsymbol{G}^{\text{test}}}(\boldsymbol{Y}^T\boldsymbol{V}^{-1}\boldsymbol{G}^{\text{test}})} = \frac{\left(\boldsymbol{Y}^T\boldsymbol{V}^{-1}\boldsymbol{G}^{\text{test}}\right)^2}{\sigma_s^2\boldsymbol{Y}^T\boldsymbol{V}^{-1}\boldsymbol{\Phi}\boldsymbol{V}^{-1}\boldsymbol{Y}}. \tag{6}$$

Under the null, T_{retro} follows a χ_1^2 distribution. It is worth noting that the model given in (5) makes relatively weak assumptions about the genotype distribution under the null and are satisfied by a variety of common models for population structure as well as for family relatedness [119]. Moreover, the test based on (6) remains valid even when the phenotype model is misspecified and with either random or phenotype-based ascertainment [49,52,145]. As a result, the retrospective way of assessing significance enables robust control of type 1 error. Meanwhile, compared with a purely retrospective analysis, the combination approach inherits the flexibility of a prospective analysis for modeling various types of phenotype data.

3.5. *Extensions and Challenges*

Existing GWAS methods typically look for genetic association on a single-SNP basis, i.e., each SNP is tested individually for its marginal association with a particular phenotype of interest, with all other SNPs ignored. While this approach has successfully

identified thousands of high-risk SNPs, a considerable proportion of heritability remained unexplained for most complex diseases. The missing heritability problem [29] has motivated more sophisticated strategies for examining genetic association. One possibility is that the missing heritability stems from the polygenic nature of the complex trait. To this end, a number of multi-locus association methods have been developed. Some methods propose to test the joint effects of multiple genetic variants in a gene or pathway with the trait [89,102,135]. This approach bears some similarity to the SNP-set test in the rare-variant analysis (see Section 4), and can greatly complement the single-SNP approach in GWASs. Other methods use a different strategy, by starting with the entire set of SNPs as predictors in a regression model and employing a variable selection technique to attain a reduced model with a subset of SNPs. The variable selection criteria usually involve some parameter regularization like LASSO [12,70,137] or utilize sparse Bayesian variable selection [37,45] to reduce the model complexity. The selected SNPs can then be used in further analysis, for example, to test for gene-gene interactions [46,127].

In addition to multi-locus analysis, multi-phenotype analysis is emerging as a powerful tool in association studies. Increasing evidence shows that a single genetic variant or a set of genetic variants can affect multiple traits at the same time, a phenomenon called pleiotropy [23,106]. This brings about the concept of phenome-wide association study (PheWAS) [14] – a reversal of the GWAS paradigm – in which a single genetic variant is tested for an association with a broad range of human phenotypes. A number of multi-phenotype methods are available: MQFAM [32] uses canonical correlation analysis to identify a linear combination of the traits that maximizes the correlation with the tested genetic variant; MV-BIMBAM [109] provides a Bayesian model comparison approach for multivariate association analysis; PHENI [24] applies a Bayesian multiple-phenotype mixed model for imputing and analyzing multiple phenotypes; MV-LMM [148] is the multivariate analogue of EMMA [54], which fits a multivariate linear mixed model to multiple, possibly related phenotypes. Despite the growing body of toolkits, simulated data suggested that no single method performs best under all scenarios [14]. Development of the PheWAS approach is still in its infancy, and unlocking the full potential of PheWAS for the characterization of the complex human genotype-phenotype relationship remains challenging.

4. Rare-variant Association Studies Using High-throughput Sequencing

4.1. *Background*

Although GWASs have successfully identified more than 10,000 SNPs associated with complex human traits and diseases [131], much of the heritability for these traits remains unaccounted for [71]. To date, the genetic markers targeted by GWAS are predominantly common variants (MAF $\geq$ 5%). However, rare variants (MAF $<$ 5%) are highly abundant in the human genome representing 95% of the genetic variability [82]. Previous studies [29,71] have suggested that these rare variants, while largely unrepresented in GWAS, may partially contribute to the missing heritability unexplained by GWAS findings [29,71]. Both evolutionary theory and empirical studies indicate that deleterious mutations undergo purifying selection and therefore tend to be rare [57,68,149]. There is also recent evidence that rare genetic variations are implicated in complex diseases [38,90,99].

Recent advances in high-throughput sequencing technologies have opened up new

opportunities for detecting rare-variant association with complex traits. Unlike array-based genotyping used in GWASs, high-throughput sequencing technologies do not rely on probes for preselected targets, thereby facilitating the identification of a plethora of rare genetic variations in the genome. In the past few years, the sequencing cost has been reduced dramatically and is still falling. As a result, whole-genome and whole-exome sequencing studies have been increasingly deployed as a popular tool to understand the contribution of rare variants to complex traits.

4.2. *Region-based Association Analysis for Rare Variants*

With increasing availability of high-throughput sequencing data, there is pressing demand to develop powerful association tests for rare genetic variants. While analysis of rare-variant associations presents some of the same analytical problems posed by GWAS, it also faces some unique challenges. In GWASs, the standard approach to association testing is the single-SNP method, in which SNPs are tested one at a time. The statistical power of such a test, for a fixed effect size, declines as the MAF of the tested SNP decreases. As a result, the single-SNP method may suffer from substantial power loss when used to analyze rare variants [9] due to the rarity of individuals carrying the mutant alleles. Moreover, the genome-wide significance threshold of $\alpha = 5 \times 10^{-8}$ commonly adopted in GWAS corresponds to approximately a million independent loci in the human genome, which is an estimate based on the total number of common SNPs and the LD structure exhibited by common genetic variation [22]. However, rare variants are far more abundant in the genome and less correlated with each other than common SNPs, which results in a more severe multiple testing burden. As a consequence, a more stringent significance threshold may be needed, leading to further power loss [7]. Nonetheless, single-variant tests can still be a useful tool for rare-variant analysis. The potential power loss may be mitigated for studies with very large sample sizes and rare variants whose effects sizes are very large [86].

A common strategy to improve power in rare-variant association analysis is to perform a region-based analysis, in which information across putative causal variants in a predefined genetic region is aggregated to test for association with the phenotype. Because the genetic regions are frequently defined by individual genes, such tests are often also referred to as gene-based tests. The regions can also be chosen using other functional annotation (see Section 4.4.1). These regions are then considered the units of association tests, and the analysis goal is to detect whether a group of variants within a region are jointly associated with the trait of interest. This aggregation strategy can improve statistical power by (i) accumulating the single-variant effects to boost association signal, and (ii) relieving multiple testing burden by reducing the total number of tests performed.

In recent years, there have been numerous methods developed for region-based tests. Here, we review some of the most commonly used types of methods. We consider a study with n individuals and a genetic region of interest in which m rare variants will be aggregated in a test. Let Y_i denote the phenotype of individual i, and $G_{ij} = 0, 1$ or 2 denote the genotype of individual i at variant j coded by the number of rare alleles carried by the individual.

4.2.1. Burden Tests

One broad class of region-based methods [61,69,75,93], typically referred to as "burden tests," involve collapsing multiple rare-variant sites in a region into a single variable,

representing a genetic burden score. The idea behind burden tests is that the presence and/or greater abundance of rare mutations in an individual confers a genetic burden that tends to add to disease susceptibility. A simple way to define the genetic burden score is

$$B_i = \sum_{j=1}^{m} w_j G_{ij},$$

where w_j is a prespecified weight for variant j that reflects the prior information on how likely the variant is to be associated with the phenotype and how strong the effect is. Then association is tested between the trait and the burden score in a univariate fashion, under a regression model or using nonparametric procedures. For example, for a quantitative trait, one can use a linear model where the trait is the response variable and the explanatory variables include the burden score and possible covariates. Then, to test for genetic association, one can consider the score statistic for the effect of the burden score on the trait, and the resulting test statistic is equivalent to

$$T_{\text{burden}} = (\sum_{j=1}^{m} w_j S_j)^2,$$

where S_j is the marginal score statistic of variant j. Significance of the test statistic can be assessed by comparing its value to a chi-square distribution with 1 degree of freedom.

In the genetic burden score, various weights have been proposed. For example, using equal weights for all variants [76] amounts to simply counting how many rare variants each individual bears. Alternatively, weights can also be constructed on the basis of the MAF, p_j, of a variant in the form of (i) $w_j = 1/\sqrt{p_j(1-p_j)}$, which upweights rarer variants [69], (ii) $w_j = I(p_j < t)$, which retains only the SNPs whose MAFs are below a threshold [61], or (iii) $w_j = \sqrt{p_j(1-p_j)}$, which is derived assuming equal contributions to population disease risk from all variants [111]. The variants can also be weighted based on their sequencing quality [6]. Another approach is to construct the weights based on predicted functions of the variants [58], such as bioinformatic annotations of the impact of an amino acid change. Many other variations of the burden tests exist. These include two of the earliest burden methods, CAST [75] and CMC [61], which set $B_i = 1$ whenever at least one rare variant is present for individual i and $B_i = 0$ otherwise.

A limitation of burden tests is that they make strong underlying assumptions about the configuration of the genetic effects across variants. Their power relies on most of the pooled variants being causal, and their effects being mostly in the same direction and of similar magnitude. Violations of those assumptions are highly likely in practice and can result in substantial power loss [51,60,81]. Intuitively, if the S_j's have different signs in equation (4.2.1), then the association signals of individual variants may cancel each other out when they are aggregated via the sum of the S_j's.

A handful of attempts have been made to couple a burden test with data-drive procedures in order to make the method more flexible. The Variable Threshold (VT) [93] builds on the hypothesis that, due to purifying selection, there is an MAF threshold below which a variant will have a much higher likelihood to be functional, and that power can be gained by pooling only such variants and ignoring the others. Since the MAF threshold is unknown, VT adaptively chooses a working threshold that yields

the most significant association result, with the p-value of the resulting test statistic assessed by permutations. Methods have also been proposed to construct the variant-specific weights w_j's in a data-adaptive way. Han & Pan [40] developed a method that allows the sign of the weight w_j to be either positive or negative depending on the direction of the estimated marginal association of variant j from the same dataset. Extensions of it have been proposed to allow the magnitudes of the weights to be also estimated from the data [44,63,141]. It has been shown both theoretically and in simulation studies that a burden test with data-adaptive weights is similar to a variance component test [27]. To some extent, data-driven procedures can relax the assumptions underlying a classical burden test. It should be noted that the significance of an adaptive test needs to counterbalance the fact that additional parameters are chosen based on the data. Consequently, the enhanced flexibility of data-driven tests may come at the price of some power loss when too many additional parameters are adaptively determined, e.g., when the parameters start to model the noise rather than a systematic trend in the data. Another limitation of adaptive burden tests is that many of them do not have an analytical null distribution and thus need resampling techniques (e.g., permutations or bootstrapping) to assess p-values.

4.2.2. Variance Component Tests

A second broad class of region-based methods [56,81,88,134], which are called "variance component (VC) tests," consider the genetic effect of a rare variant as a random effect and model the distribution of the effects across a set of variants being tested. Compared to the burden test, which models the rare-variant effects as fixed, this random-effect approach allows the genetic effects to have varying sizes and magnitudes across variants. This can accommodate the scenario where a mixture of protective, deleterious and non-causal rare alleles are included in the test. Then, the VC method tests for non-zero variance of the distribution of the random rare-variant effects by aggregating individual variant statistics measuring the strength of association at each site. This is in contrast to the approach taken by a burden test which directly aggregates the genotypes of individual sites.

We use a linear model set-up to illustrate the idea of a VC test. For individual i, let Y_i be a quantitative trait of interest, $\boldsymbol{X}_i$ be a covariate (row) vector including the intercept, $\boldsymbol{G}_i = (G_{i1}, \cdots, G_{im})$ be the genotype (row) vector for the m rare-variant sites, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)^T$ be a vector of the phenotypic effects of the m variants. Given the variant effect vector $\boldsymbol{\beta}$, a model for Y_i is

$$Y_i|\boldsymbol{\beta}, \boldsymbol{X}_i, \boldsymbol{G}_i \sim N(\boldsymbol{X}_i\boldsymbol{\gamma} + \boldsymbol{G}_i\boldsymbol{\beta}, \sigma^2),$$

where $\boldsymbol{\gamma}$ is the vector of covariate effects, and σ^2 is the variance of environmental effects. For a binary trait, a logistic model can be used instead. A VC method typically assumes that the variant effects β_j's are independent random effects coming from a common distribution with zero mean and variance σ^2_β. Variant-specific weights can also be incorporated in the model by assuming $\mathrm{Var}(\beta_j) = w_j\sigma^2_\beta$. Then, the problem of testing rare-variant association can be solved by testing $H_0\colon \sigma^2_\beta = 0$ against $H_1\colon \sigma^2_\beta > 0$. For example, SKAT [134] uses weights from the density of a Beta distribution evaluated

at the MAF of a variant, and its test statistic takes the form

$$T_{\text{vc}} = \sum_{j=1}^{m} w_j^2 S_j^2,$$

where S_j is the marginal score statistic of variant j alone. It is noteworthy that by first squaring the S_j's rather than directly summing them as is done in a burden test, this test statistic allows both positively and negatively associated variants to contribute to the aggregated signal without canceling out each other. Under the null hypothesis of no association, the asymptotic distribution of T_{vc} is a linear combination of independent 1-degree-of-freedom chi-square distributions, which can be used to assess p-values under a large sample size. Other methods for p-value calculation have also been proposed to improve calibration of the test statistic [60,142].

Examples of VC tests proposed prior to SKAT include SumSqU [88], which assumes equal weights for all variants, and C-alpha [81], which can be seen as a special case of SKAT for binary traits with no covariates incorporated. Another related type of methods are kernel-based methods, many of which can be derived from a model similar to what is assumed in a VC method [79,136], but with different assumptions made on the variance structure of the variant effects.

4.2.3. Omnibus Tests

Burden tests and VC tests tend to perform well in different scenarios. VC tests are more powerful than burden tests when the pooled variants include both negatively and positively associated variants as well as neutral variants, whereas burden tests tend to outperform VC tests when the proportion of causal variants is relatively high among the pooled variants and the variant effects have similar directions and magnitudes [60, 134]. In practice, there is typically little prior knowledge about the configuration of the genetic effects of the pooled variants, and the genetic architecture of the trait can vary from one genetic region to another. Therefore, it is of interest to seek an omnibus test that unifies the burden tests and the VC tests for enhanced robustness. SKAT-O [60] is such a method attempting to join the strengths of the two approaches, with a mixture of the two adaptively balanced by the data. It considers a class of convex combinations of the burden and SKAT test statistics given by

$$T_\rho = \rho T_{\text{burden}} + (1-\rho) T_{\text{vc}},$$

for $0 \leq \rho \leq 1$. To obtain an optimal test statistic, SKAT-O proposes to choose ρ in a data-driven way to minimize the p-value of the corresponding T_ρ. The significance of the resulting test statistic can be evaluated analytically without the need for permutations. By optimizing over the parameter ρ, SKAT-O adaptively balances between burden tests and SKAT in order to achieve robustness for widely ranging genetic architecture.

Other methods joining burden and VC tests have been also proposed. Derkach et al. [26] proposed to first perform the two tests separately and then combine the p-values using Fisher's method:

$$T_{\text{Fisher}} = -2\log(p_{\text{burden}}) - 2\log(p_{\text{vc}}),$$

where p_{burden} and p_{vc} are the p-values of the burden and the VC tests, respectively.

The significance of the test statistic T_{Fisher} is then assessed by permutation. Wang et al. [129] proposed to jointly test for the common association across rare-variant sites and the individual deviations from the common effect using a score test. A related method was proposed by Sun et al. [112], which allows the common effect across variants to depend on covariates and known variant characteristics.

An important advantage of an omnibus test is its robustness to deviations from the assumptions underlying either the burden test or a VC test. For example, SKAT-O is shown to have good power in scenarios where either burden test or SKAT performs poorly, and may even outperform both under certain settings. However, when the actual genetic architecture largely aligns with what is assumed in the burden or VC tests, an omnibus test may suffer some power loss due to the price paid by optimizing over a wider class of tests (as in SKAT-O) or losing extra degrees of freedom (as in the method by Want et al. [129]). Another disadvantage of some omnibus tests is the need for permutation procedures to assess significance, which may be computationally intensive and hard to generalize to samples with population or family structure.

4.2.4. Other Tests

Many other rare-variant association testing methods have been proposed that do not fit into the aforementioned categorization. In particular, an important factor that affects the power of a test is the sparsity of association signals, as reflected by the fraction of causal variants among the pooled variants. A number of attempts have been made to detect associations when single-variant association signals are sparse. The EC method [18] combines the single-variant test statistics in an exponential way:

$$T_{\text{EC}} = \sum_{j=1}^{m} \exp\{Z_j^2/2\},$$

where Z_j is a test statistic for variant j that follows a standard normal distribution under the null (e.g., Z_j given by $Z_j^2 = S_j^2/\text{Var}(S_j)$, where S_j is the score statistic). This is in contrast to the burden test, which combines single-variant statistics linearly (Equation (4.2.1)), and the VC test, which combines them quadratically (Equation (4.2.2)). The exponential combination scheme serves to amplify the effects from large Z_j's and thus helps boost power under scenarios of sparse association signals, where only a very low fraction of the pooled variants are causal or have large effects. Permutations are needed to compute the p-values in EC.

Recently, the higher criticism (HC) [28] has been explored and extended as a statistical tool to detect rare-variant associations that are highly sparse and weak [10,11,78,138]. HC is a statistical method that was originally proposed to aggregate information across a large number of independent test statistics to test the joint null hypothesis against the alternative that a sparse set of signals are present. To address the unique needs of region-based rare-variant association analysis, adaptations of HC have been proposed to accommodate correlations between single-variant statistics induced by linkage disequilibrium [10], to allow analytic calculations of p-values that do not rely on high-dimension asymptotics [11], and to be applicable to binary regression problems [78].

4.3. *Methods For Samples with Family Structure*

Related individuals are frequently included in high-throughput sequencing studies. Samples of family members can yield several advantages in sequencing studies compared with samples with unrelated individuals. Unlike the population-based design in which the effect of a very rare causal variant can be challenging to capture due to the extremely low frequency of observing the mutant allele, family-based samples can be enriched with such variants and thereby yield improved power [55]. Moreover, sampling individuals from a known pedigree offers unique opportunities to impute relatives' sequence data from a small number of sequenced individuals by leveraging the pedigree information [39,123,126]. For a fixed sequencing cost, this can further improve power. In addition, including related individuals allows for more reliable methods to assess data quality and to detect and correct for sequencing errors [100,146].

As with GWAS data, when related individuals are present in high-throughput sequence data, association tests need to account for the family structure in order to ensure adequate control of type 1 error and to improve power. To achieve this, statistical methods have been developed in recent years for rare-variant association detection in completely general designs containing related individuals. Such methods are appropriate for arbitrary combinations of related and unrelated samples, including small outbred pedigrees and unrelated individuals, complex inbred pedigrees, as well as population-based samples that include cryptic relatedness. FamSKAT [16] is a method that extends SKAT to account for familial correlation by incorporating a polygenic variance component in the VC model. More specifically, it assumes a multivariate normal distribution on $\boldsymbol{Y}$ given $\boldsymbol{\beta}$, with the mean structure given in Equation (4.2.2), the covariance structure given by

$$\mathrm{Var}(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{G}) = \sigma_e^2 \boldsymbol{I} + \sigma_a^2 \boldsymbol{\Phi},$$

with the same assumptions on $\boldsymbol{\beta}$ as in the VC model for unrelated individuals. The same authors proposed famBT [16], an extension of the burden test to family samples. Schaid et al. [103] also developed extensions of the burden and VC tests to pedigree data. As is the case for samples with unrelated individuals, for family samples, VC and burden tests tend to perform well in different scenarios depending on the underlying genetic architecture of the trait in the test region. To obtain a more robust method, MONSTER is a method that joins the strengths of famVC and famBT by extending SKAT-O to samples that include related individuals. Other types of tests have also been extended to related individuals. Choi et al. [19] developed retrospective rare-variant association tests for family samples, and Zhu & Xiong [151] developed an extension to the CMC method. In addition to methods applicable to general study designs that include relatives, various approaches have been developed specifically for analysis of family-based designs [25,42,47,128].

4.4. *Additional Considerations on the Analysis of Sequence Data*

4.4.1. Which Variants to Aggregate?

In region-based rare variant tests, the set of variants that are grouped in a region constitute the analysis unit. To select which variants are grouped into a unit, an important issue is to determine the regions within which variants are to be pooled. A popular option, particularly in whole-exome sequencing studies, is to collapse over an individ-

ual gene. A benefit of gene-based tests is their natural interpretability, because genes are considered the functional units of heredity and many genes have well-annotated functional information. An alternative approach is to construct the regions using sliding windows of a fixed chromosomal length or number of variants [113]. Compared with gene-based tests, this strategy has the advantage of being able to include intergenic variants. But it is often unclear how to choose the window size, and it can be challenging to interpret significant regions.

Another important consideration is whether to include all variants in a region in a test or to include only a subset of variants. For example, the subset of variants can be chosen to be all the coding variants or only the nonsynonymous coding variants. In addition, many bioinformatic tools are available to predict the impact of a DNA mutation on the amino acid sequence, the functional role of the DNA mutation or its evolutionary conservativenss [3,114,116,125]. Such predictions can be used to refine the subset of variants included in a test. For example, variants that are predicted to be likely detrimental can be prioritized. The bioinformatic predictions can also be used to weight the variants.

4.4.2. Population Stratification

In addition to family structure, another frequent source of confounding in sequencing studies is population stratification. To adjust for population structure in rare-variant association testing, some authors have proposed to use similar strategies taken in GWASs: PCs can be included as covariates in a regression model; alternatively, in a rare-variant association test carried out in samples with possibly related individuals, the pedigree-based kinship matrix can be replaced by a GRM estimated from genome-wide data [103]. However, caution must be exercised when using these approaches. Because rare genetic mutations often reflect recent evolutionary history and display a different stratification pattern from common SNPs, it is unclear whether methods designed to control stratification for GWAS remain effective for rare-variant analysis [66,73,87]. Using rare variants to estimate PCs and the GRM can be unstable due to the low MAFs [121]. Further methodological research is needed to address the problem of how population structure can be effectively adjusted for in rare-variant association testing.

4.4.3. Multiple Testing Correction in Rare-Variant Association Analysis

As mentioned previously, the GWAS significance threshold of $\alpha = 5 \times 10^{-8}$ may not be appropriate for sequencing studies due to the greater abundance of rare variants than common SNPs and their unique LD structure. For rare-variant analysis, there are still no community standards for rare-variant analysis on a genome-wide significance threshold in place that comparable to that in GWAS. This is partly due to the fact that the number of tests performed depends on a variety of factors such as the sequencing platform used, the depth of coverage, the size and ancestry of the sample, and which and how variants are aggregated in a region-based test.

For single-variant tests, a range of thresholds have been suggested from 1×10^{-9} to 3.75×10^{-7} under various scenarios and based on different assumptions [31,98,105]. For gene-based tests, it has been suggested that a reasonable genome-wide significance threshold is $\alpha = 2.5 \times 10^{-6}$, based on Bonferroni correction corresponding to approximately 20,000 genes in the human genome [77]. A limitation of this threshold is that it does not take into account the correlations between tests performed for individuals

genes [36]. With such correlations, the effective number tests are expected to be much smaller than 20,000, potentially justifying a less stringent significance threshold and thereby allowing for higher power. Moreover, it is not straightforward what threshold should be used for other variant grouping schemes used in a region-based test. Xu et al. [139] proposed a threshold empirically ascertained from whole-genome sequence data on chromosome 3 from the UK10K study, but their study targeted a specific implementation of sliding-window tests only and was based on data of European ancestry.

Given the scope and depth of current literature, there is not enough evidence to comprehensively assess the performance of the proposed significance thresholds when applied to new rare-variant analysis. It remains an open research question what the best strategy is for handling multiple testing in high-throughput sequencing data.

5. Discussion

The past decade has witnessed a rapid expansion of big data in biomedical research. In particular, high-throughput genotyping and sequencing techniques are routinely utilized to study the genetic basis of complex human disease, and they generate massive genetic/genomic data that pose tremendous analytical challenges. Here, we have reviewed some commonly-used statistical methods for the analysis of GWAS and high-throughput sequencing studies, and have discussed some of the issues and challenges that arise in such analysis. While many methods have been proposed for genetic association testing, the performance of a given method depends on the underlying genetic architecture of a complex trait, and there is usually no single method that is optimal across all scenarios. By comparing various methods, we have discussed some important considerations for choosing a statistical test.

Another prevailing theme in our review is how to address the statistical dependence in genetic association analysis; this includes, e.g., dependence among sampled individuals, dependence between loci across the genome, dependence between multiple phenotypes, etc. Neglecting or not properly accounting for such dependencies could lead to inflated type 1 error and/or reduced power in association testing. Although methods such as PCA and LMM have demonstrated powerful in accounting for sample structure in common-variant association analysis, their usefulness in rare-variants analysis remains to be fully demonstrated.

Most of the work we have described are for (common or rare) single-nucleotide variants, which have been the focus of the majority of genetic association studies. Other types of DNA sequence variations such copy number variation and inversions are less investigated but have been shown in many studies to be important components for complex diseases [48]. Moreover, integrating multiple types of omics data such as transcriptomic and metabolomic data offers great promise to further our understating of biological processes underlying complex diseases.

Figures and Tables

Figure 1. Example Q-Q plots from simulated data when (A) no confounders are present to inflate type 1 error and (B) confounders are present, leading to a global deviation from the null hypothesis of no association.

Table 1. Comparison of softwares for large-scale GWASs

Package	Approach	GRM Estimation	Complexity
EMMAX [53]	LMM (approximate)	using genome-wide SNPs	$O(MN^2)$
GRAMMAR-Gamma [115]	LMM (approximate)	using genome-wide SNPs	$O(MN^2)$ for building GRM and $O(MN)$ for association tests
GEMMA [147]	LMM (exact)	using genome-wide SNPs	$O(MN^2)$
FaST-LMM [64]	LMM (exact)	using SNPs selected through FaST-LMM-Select procedure	$O(MN^2)$ if $M > N$ or $O(M^2N)$ if $M \leq N$
GCTA-LOGO [140]	LMM (approximate)	via the LOCO analysis	$O(MN^2)$
GTAM [2]	LMM (approximate)	using kinship from known pedigree	$O(MN^2)$
MASTOR [49]	LMM (retrospective)	using kinship from known pedigree	$O(MN^2)$ for building GRM and O(MN) for association tests
BOLT-LMM [67]	LMM (approximate)	using genome-wide SNPs or via the LOCO analysis	$O(MN^{1.5})$

Note: N is the number of individuals and M is the number of SNPs.

References

[1] Mark Abney. Permutation testing in the presence of polygenic variation. *Genetic Epidemiology*, 39(4):249–258, 2015.

[2] Mark Abney, Carole Ober, and Mary Sara McPeek. Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *The American Journal of Human Genetics*, 70(4):920–934, 2002.

[3] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010.

[4] D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

[5] Peter Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[6] Jennifer L Asimit, Aaron G Day-Williams, Andrew P Morris, and Eleftheria Zeggini. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Human Heredity*, 73(2):84–94, 2012.

[7] Paul L Auer and Guillaume Lettre. Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, 7(1):16, 2015.

[8] Yurii S Aulchenko, Maksim V Struchalin, and Cornelia M van Duijn. Probabel package for genome-wide association analysis of imputed data. *BMC Bioinformatics*, 11(1):134, 2010.

[9] Vikas Bansal, Ondrej Libiger, Ali Torkamani, and Nicholas J Schork. Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11):773–785, 2010.

[10] Ian Barnett, Rajarshi Mukherjee, and Xihong Lin. The generalized higher criticism for testing SNP-set effects in genetic association studies. *Journal of the American Statistical Association*, 2016.

[11] Ian J Barnett and Xihong Lin. Analytic P-value calculation for the higher criticism test in finite d problems. *Biometrika*, 101(4):964, 2014.

[12] Saonli Basu, Wei Pan, Xiaotong Shen, and William S Oetting. Multilocus association testing with penalized regression. *Genetic Epidemiology*, 35(8):755–765, 2011.

[13] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[14] William S Bush, Matthew T Oetjens, and Dana C Crawford. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nature Reviews Genetics*, 17(3):129–145, 2016.

[15] H.-S. Chen, X. Zhu, H. Zhao, and S. Zhang. Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Annals of Human Genetics*, 67(3):250–264, 2003.

[16] Han Chen, James B Meigs, and Josée Dupuis. Sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology*, 37(2):196–204, 2013.

[17] Han Chen, Chaolong Wang, Matthew P Conomos, Adrienne M Stilp, Zilin Li, Tamar Sofer, Adam A Szpiro, Wei Chen, John M Brehm, Juan C Celedón, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4):653–666, 2016.

[18] Lin S Chen, Li Hsu, Eric R Gamazon, Nancy J Cox, and Dan L Nicolae. An exponential combination procedure for set-based association tests in sequencing studies. *The American Journal of Human Genetics*, 91(6):977–986, 2012.

[19] Sungkyoung Choi, Sungyoung Lee, Markus M Nöthen, Christoph Lange, Taesung Park, and Sungho Won. FARVAT: a family-based rare variant association test. *Bioinformatics*, 30(22):3197–205, 2014.

[20] Matthew P Conomos, Cecelia A Laurie, Adrienne M Stilp, Stephanie M Gogarten, Caitlin P McHugh, Sarah C Nelson, Tamar Sofer, Lindsay Fernández-Rhodes, Anne E Justice, Mariaelisa Graff, et al. Genetic diversity and association studies in US hispanic/latino populations: applications in the hispanic community health study/study of latinos. *The American Journal of Human Genetics*, 98(1):165–184, 2016.

[21] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[22] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[23] Chris Cotsapas, Benjamin F Voight, Elizabeth Rossin, Kasper Lage, Benjamin M Neale, Chris Wallace, Gonçalo R Abecasis, Jeffrey C Barrett, Timothy Behrens, Judy Cho, et al. Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet*, 7(8):e1002254, 2011.

[24] Andrew Dahl, Valentina Iotchkova, Amelie Baud, Åsa Johansson, Ulf Gyllensten, Nicole Soranzo, Richard Mott, Andreas Kranis, and Jonathan Marchini. A multiple-phenotype imputation method for genetic studies. *Nature Genetics*, 48(4):466–472, 2016.

[25] Gourab De, Wai-Ki Yip, Iuliana Ionita-Laza, and Nan Laird. Rare variant analysis for family-based design. *PLoS One*, 8(1):e48495, 2013.

[26] Andriy Derkach, Jerry F Lawless, and Lei Sun. Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epidemiology*, 37(1):110–121, 2013.

[27] Andriy Derkach, Jerry F Lawless, Lei Sun, et al. Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science*, 29(2):302–321, 2014.

[28] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, pages 962–994, 2004.

[29] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450, 2010.

[30] Jakris Eu-Ahsunthornwattana, E Nancy Miller, Michaela Fakiola, Selma MB Jeronimo, Jenefer M Blackwell, Heather J Cordell, Wellcome Trust Case Control Consortium 2, et al. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet*, 10(7):e1004445, 2014.

[31] João Fadista, Alisa K Manning, Jose C Florez, and Leif Groop. The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, 24(8):1202–1205, 2016.

[32] Manuel AR Ferreira and Shaun M Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–133, 2009.

[33] Ronald A Fisher. Xv.-the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02):399–433, 1919.

[34] Christian Fuchsberger, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, Davis J McCarthy, et al. The genetic architecture of type 2 diabetes. *Nature*, 536(7641), 2016.

[35] Greg Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145, 2012.

[36] Celia MT Greenwood, ChangJiang Xu, and Antonio Ciampi. Significance thresholds for rare variant signals. In *Assessing Rare Variation in Complex Traits*, pages 169–183. Springer, 2015.

[37] Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, pages 1780–1815, 2011.

[38] Julius Gudmundsson, Patrick Sulem, Daniel F Gudbjartsson, Gisli Masson, Bjarni A Agnarsson, Kristrun R Benediktsdottir, Asgeir Sigurdsson, Olafur Th Magnusson, Sigurjon A Gudjonsson, Droplaug N Magnusdottir, et al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature Genet-*

ics, 44(12):1326–1329, 2012.

[39] A. Gusev, M.J. Shah, E.E. Kenny, A. Ramachandran, J.K. Lowe, J. Salit, C.C. Lee, E.C. Levandowsky, T.N. Weaver, Q.C. Doan, et al. Low-pass genome-wide sequencing and variant inference using identity-by-descent in an isolated human population. *Genetics*, 190(2):679–689, 2012.

[40] Fang Han and Wei Pan. A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, 70(1):42–54, 2010.

[41] Tristan J Hayeck, Po-Ru Loh, Samuela Pollack, Alexander Gusev, Nick Patterson, Noah A Zaitlen, and Alkes L Price. Mixed model association with family-biased case-control ascertainment. *The American Journal of Human Genetics*, 100(1):31–39, 2017.

[42] Zongxiao He, Brian J O'Roak, Joshua D Smith, Gao Wang, Stanley Hooker, Regie Lyn P Santos-Cortez, Biao Li, Mengyuan Kan, Nik Krumm, Deborah A Nickerson, et al. Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *The American Journal of Human Genetics*, 94(1):33–46, 2014.

[43] Gabriel E Hoffman. Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PLoS One*, 8(10):e75707, 2013.

[44] Thomas J Hoffmann, Nicholas J Marini, and John S Witte. Comprehensive approach to analyzing rare genetic variants. *PLoS One*, 5(11):e13584, 2010.

[45] Clive J Hoggart, John C Whittaker, Maria De Iorio, and David J Balding. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet*, 4(7):e1000130, 2008.

[46] Hung Hung, Yu-Ting Lin, Penweng Chen, Chen-Chien Wang, Su-Yun Huang, and Jung-Ying Tzeng. Detection of gene–gene interactions using multistage sparse and low-rank regression. *Biometrics*, 72(1):85–94, 2015.

[47] Iuliana Ionita-Laza, Seunggeun Lee, Vladimir Makarov, Joseph D Buxbaum, and Xihong Lin. Family-based association tests for sequence data, and comparisons with population-based association tests. *European Journal of Human Genetics*, 21(10):1158–1162, 2013.

[48] Iuliana Ionita-Laza, Angela J Rogers, Christoph Lange, Benjamin A Raby, and Charles Lee. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*, 93(1):22–26, 2009.

[49] Johanna Jakobsdottir and Mary Sara McPeek. MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. *The American Journal of Human Genetics*, 92(5):652–666, 2013.

[50] Duo Jiang, Joelle Mbatchou, and Mary Sara McPeek. Retrospective association analysis of binary traits: overcoming some limitations of the additive polygenic model. *Human Heredity*, 80(4):187–195, 2015.

[51] Duo Jiang and Mary Sara McPeek. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genetic Epidemiology*, 38(1):10–20, 2014.

[52] Duo Jiang, Sheng Zhong, and Mary Sara McPeek. Retrospective binary-trait association test elucidates genetic architecture of crohn disease. *The American Journal of Human Genetics*, 98(2):243–255, 2016.

[53] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010.

[54] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[55] R. Kazma and J.N. Bailey. Population-based and family-based designs to analyze rare variants in complex diseases. *Genetic Epidemiology*, 35(S1):S41–S47, 2011.

[56] C Ryan King, Paul J Rathouz, and Dan L Nicolae. An evolutionary framework for association testing in resequencing studies. *PLoS Genet*, 6(11):e1001202, 2010.

[57] Gregory V Kryukov, Len A Pennacchio, and Shamil R Sunyaev. Most rare missense al-

leles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics*, 80(4):727–739, 2007.

[58] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature Protocols*, 4(7):1073–1081, 2009.

[59] A.B. Lee, D. Luca, L. Klei, B. Devlin, and K. Roeder. Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology*, 34(1):51–59, 2010.

[60] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, ESP Lung Project Team, David C Christiani, Mark M Wurfel, Xihong Lin, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.

[61] B. Li and S.M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[62] Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. Genotype imputation. *Annual Review of Genomics and Human Genetics*, 10:387–406, 2009.

[63] Dan-Yu Lin and Zheng-Zheng Tang. A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, 89(3):354–367, 2011.

[64] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

[65] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9(6):525–526, 2012.

[66] Qianying Liu, Dan L Nicolae, and Lin S Chen. Marbled inflation from population structure in gene-based association studies with rare variants. *Genetic Epidemiology*, 37(3):286–292, 2013.

[67] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, 2015.

[68] Daniel G MacArthur, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K Pickrell, Stephen B Montgomery, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, 2012.

[69] B.E. Madsen and S.R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384, 2009.

[70] Nathalie Malo, Ondrej Libiger, and Nicholas J Schork. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *The American Journal of Human Genetics*, 82(2):375–385, 2008.

[71] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[72] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913, 2007.

[73] Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, 44(3):243–246, 2012.

[74] Joel Mefford and John S Witte. The covariate's dilemma. *PLoS Genet*, 8(11):e1003096, 2012.

[75] S. Morgenthaler and W.G. Thilly. A strategy to discover genes that carry multi-allelic

or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1):28–56, 2007.

[76] Andrew P Morris and Eleftheria Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2):188–193, 2010.

[77] Loukas Moutsianas, Vineeta Agarwala, Christian Fuchsberger, Jason Flannick, Manuel A Rivas, Kyle J Gaulton, Patrick K Albers, Gil McVean, Michael Boehnke, David Altshuler, et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet*, 11(4):e1005165, 2015.

[78] Rajarshi Mukherjee, Natesh S Pillai, and Xihong Lin. Hypothesis testing for high-dimensional sparse binary regression. *Annals of Statistics*, 43(1):352, 2015.

[79] Indranil Mukhopadhyay, Eleanor Feingold, Daniel E Weeks, and Anbupalam Thalamuthu. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genetic Epidemiology*, 34(3):213–221, 2010.

[80] Benjamin M Neale and Shaun Purcell. The positives, protocols, and perils of genome-wide association. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147(7):1288–1294, 2008.

[81] Benjamin M Neale, Manuel A Rivas, Benjamin F Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M Purcell, Kathryn Roeder, and Mark J Daly. Testing for an unusual distribution of rare variants. *PLoS Genet*, 7(3):e1001322, 2011.

[82] Matthew R Nelson, Daniel Wegmann, Margaret G Ehm, Darren Kessner, Pamela St Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104, 2012.

[83] Marc Nerlove and S James Press. *Univariate and multivariate log-linear and logistic models*, volume 1306. Rand Corporation Santa Monica, CA, 1973.

[84] The Network, Pathway Analysis Subgroup of the Psychiatric Genomics Consortium, et al. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience*, 18(2):199–209, 2015.

[85] D.L. Newman, M. Abney, M.S. McPeek, C. Ober, and N.J. Cox. The importance of genealogy in determining genetic associations with complex traits. *American Journal of Human Genetics*, 69(5):1146–1148, 2001.

[86] Dan L Nicolae. Association tests for rare variants. *Annual Review of Genomics and Human Genetics*, 17:117–130, 2016.

[87] Timothy D O'Connor, Adam Kiezun, Michael Bamshad, Stephen S Rich, Joshua D Smith, Emily Turner, Suzanne M Leal, Joshua M Akey, et al. Fine-scale patterns of population stratification confound rare variant association tests. *PLoS One*, 8(7):e65834, 2013.

[88] Wei Pan. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology*, 33(6):497–507, 2009.

[89] Wei Pan, Il-Youp Kwak, and Peng Wei. A powerful pathway-based adaptive test for genetic association with common or rare variants. *The American Journal of Human Genetics*, 97(1):86–98, 2015.

[90] Gina M Peloso, Paul L Auer, Joshua C Bis, Arend Voorman, Alanna C Morrison, Nathan O Stitziel, Jennifer A Brody, Sumeet A Khetarpal, Jacy R Crosby, Myriam Fornage, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *The American Journal of Human Genetics*, 94(2):223–232, 2014.

[91] Matti Pirinen, Peter Donnelly, and Chris CA Spencer. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics*, 44(8):848–851, 2012.

[92] Monnat Pongpanich, Patrick F Sullivan, and Jung-Ying Tzeng. A quality control algorithm for filtering SNPs in genome-wide association studies. *Bioinformatics*,

26(14):1731–1737, 2010.

[93] Alkes L Price, Gregory V Kryukov, Paul IW de Bakker, Shaun M Purcell, Jeff Staples, Lee-Jen Wei, and Shamil R Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838, 2010.

[94] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

[95] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[96] J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[97] Childhood Asthma Management Program, Carole Ober, Dan L Nicolae, Mexico City Childhood Asthma Study (MCAAS), et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse north american populations. *Nature Genetics*, 43(9):887–892, 2011.

[98] Sara Leslie Pulit, Paul I Wen de Bakker, et al. The multiple testing burden in sequencing-based disease studies of global populations. *bioRxiv*, page 053264, 2016.

[99] Manuel A Rivas, Mélissa Beaudoin, Agnes Gardet, Christine Stevens, Yashoda Sharma, Clarence K Zhang, Gabrielle Boucher, Stephan Ripke, David Ellinghaus, Noel Burtt, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genetics*, 43(11):1066–1073, 2011.

[100] J.C. Roach, G. Glusman, A.F. Smit, C.D. Huff, R. Hubley, P.T. Shannon, L. Rowen, K.P. Pant, N. Goodman, M. Bamshad, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639, 2010.

[101] German Rodriguez and Noreen Goldman. Improved estimation procedures for multilevel models with binary response: a case-study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(2):339–355, 2001.

[102] Daniel J Schaid, Shannon K McDonnell, Scott J Hebbring, Julie M Cunningham, and Stephen N Thibodeau. Nonparametric tests of association of multiple genes with human disease. *The American Journal of Human Genetics*, 76(5):780–793, 2005.

[103] Daniel J Schaid, Shannon K McDonnell, Jason P Sinnwell, and Stephen N Thibodeau. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genetic Epidemiology*, 37(5):409–418, 2013.

[104] Bertrand Servin and Matthew Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, 3(7):e114, 2007.

[105] Rafal S Sobota, Daniel Shriner, Nuri Kodaman, Robert Goodloe, Wei Zheng, Yu-Tang Gao, Todd L Edwards, Christopher I Amos, and Scott M Williams. Addressing population-specific multiple testing burdens in genetic association studies. *Annals of Human Genetics*, 79(2):136–147, 2015.

[106] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, 2013.

[107] Minsun Song, Wei Hao, and John D Storey. Testing for genetic associations in arbitrarily structured populations. *Nature Genetics*, 47(5):550–554, 2015.

[108] Stephen A Stanhope and Mark Abney. GLOGS: a fast and powerful method for GWAS of binary traits with risk covariates in related populations. *Bioinformatics*, 28(11):1553–1554, 2012.

[109] Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PLoS One*, 8(7):e65245, 2013.

[110] Jae Hoon Sul, Michael Bilow, Wen-Yun Yang, Emrah Kostem, Nick Furlotte, Dan He, and Eleazar Eskin. Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PLoS Genet*, 12(3):e1005849, 2016.

[111] Jae Hoon Sul, Buhm Han, Dan He, and Eleazar Eskin. An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genetics*, 188(1):181–188, 2011.

[112] Jianping Sun, Yingye Zheng, and Li Hsu. A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology*, 37(4):334–344, 2013.

[113] Yun Ju Sung, Keegan D Korthauer, Michael D Swartz, and Corinne D Engelman. Methods for collapsing multiple rare variants in whole-genome sequence data. *Genetic Epidemiology*, 38(S1):S13–S20, 2014.

[114] Shamil R Sunyaev. Inferring causality and functional significance of human coding dna variants. *Human Molecular Genetics*, pages 21(R1): R10–R17, 2012.

[115] Gulnara R Svishcheva, Tatiana I Axenovich, Nadezhda M Belonogova, Cornelia M Van Duijn, and Yurii S Aulchenko. Rapid variance components-based method for whole-genome association analysis. *Nature Genetics*, 44(10):1166–1170, 2012.

[116] Jacob A Tennessen, Abigail W Bigham, Timothy D OConnor, Wenqing Fu, Eimear E Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69, 2012.

[117] Yik Y Teo, Andrew E Fry, Taane G Clark, ES Tai, and Mark Seielstad. On the usage of hwe for identifying genotyping errors. *Annals of Human Genetics*, 71(5):701–703, 2007.

[118] Timothy Thornton and Mary Sara McPeek. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *The American Journal of Human Genetics*, 81(2):321–337, 2007.

[119] Timothy Thornton and Mary Sara McPeek. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *The American Journal of Human Genetics*, 86(2):172–184, 2010.

[120] Timothy Thornton, Hua Tang, Thomas J Hoffmann, Heather M Ochs-Balcom, Bette J Caan, and Neil Risch. Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91(1):122–138, 2012.

[121] Timothy A Thornton. Statistical methods for genome-wide and sequencing association studies of complex traits in related samples. *Current Protocols in Human Genetics*, pages 1–28, 2015.

[122] George Tucker, Po-Ru Loh, Iona M MacLeod, Ben J Hayes, Michael E Goddard, Bonnie Berger, and Alkes L Price. Two-variance-component model improves genetic prediction in family datasets. *The American Journal of Human Genetics*, 97(5):677–690, 2015.

[123] L.H. Uricchio, J.X. Chong, K.D. Ross, C. Ober, and D.L. Nicolae. Accurate imputation of rare and common variants in a founder population from a small number of sequenced individuals. *Genetic Epidemiology*, 36(4):312–319, 2012.

[124] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[125] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, 2010.

[126] Miaoyan Wang, Johanna Jakobsdottir, Albert V Smith, and Mary Sara McPeek. G-STRATEGY: Optimal selection of individuals for sequencing in genetic association studies. *Genetic Epidemiology*, 40(6):446–460, 2016.

[127] Miaoyan Wang, Fabrice Roux, Claudia Bartoli, Carine Huard-Chauveau, Christopher Meyer, Hana Lee, Dominique Roby, Mary Sara McPeek, and Joy Bergelson. Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proceedings of the National Academy of Sciences*, page 201710980, 2018.

[128] Xuefeng Wang, Zhenyu Zhang, Nathan Morris, Tianxi Cai, Seunggeun Lee, Chaolong Wang, W Yu Timothy, Christopher A Walsh, and Xihong Lin. Rare variant association test in family-based sequencing studies. *Briefings in Bioinformatics*, page doi: 10.1093/bib/bbw083, 2016.

[129] Yuanjia Wang, Yin-Hsiu Chen, and Qiong Yang. Joint rare variant association test of

the average and individual effects for sequencing studies. *PLoS One*, 7(3):e32485, 2012.

[130] Omer Weissbrod, Christoph Lippert, Dan Geiger, and David Heckerman. Accurate liability estimation improves power in ascertained case-control studies. *Nature Methods*, 12(4):332–334, 2015.

[131] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, 2014.

[132] Peter H Westfall and S Stanley Young. *Resampling-based multiple testing: Examples and methods for P-value adjustment*, volume 279. John Wiley & Sons, 1993.

[133] Sewall Wright. An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics*, 19(6):506, 1934.

[134] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[135] Michael C Wu, Peter Kraft, Michael P Epstein, Deanne M Taylor, Stephen J Chanock, David J Hunter, and Xihong Lin. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

[136] Michael C Wu, Arnab Maity, Seunggeun Lee, Elizabeth M Simmons, Quaker E Harmon, Xinyi Lin, Stephanie M Engel, Jeffrey J Molldrem, and Paul M Armistead. Kernel machine SNP-set testing under multiple candidate kernels. *Genetic Epidemiology*, 37(3):267–275, 2013.

[137] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

[138] Zheyang Wu, Yiming Sun, Shiquan He, Judy Cho, Hongyu Zhao, Jiashun Jin, et al. Detection boundary and higher criticism approach for rare and weak genetic effects. *The Annals of Applied Statistics*, 8(2):824–851, 2014.

[139] ChangJiang Xu, Ioanna Tachmazidou, Klaudia Walter, Antonio Ciampi, Eleftheria Zeggini, and Celia MT Greenwood. Estimating genome-wide significance for whole-genome sequencing studies. *Genetic Epidemiology*, 38(4):281–290, 2014.

[140] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–106, 2014.

[141] Nengjun Yi and Degui Zhi. Bayesian analysis of rare variants in genetic association studies. *Genetic Epidemiology*, 35(1):57–69, 2011.

[142] Xiang Zhan, Anna Plantinga, Ni Zhao, and Michael C Wu. A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics*, page 10.1111/biom.12684, 2017.

[143] J. Zhang, P. Niyogi, and M.S. McPeek. Laplacian eigenfunctions learn population structure. *PLoS One*, 4(12):e7928, 2009.

[144] S. Zhang, X. Zhu, and H. Zhao. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic Epidemiology*, 24(1):44–56, 2003.

[145] Sheng Zhong, Duo Jiang, and Mary Sara McPeek. CERAMIC: Case-control association testing in samples with related individuals, based on retrospective mixed model analysis with adjustment for covariates. *PLoS Genet*, 12(10):e1006329, 2016.

[146] B. Zhou and A.S. Whittemore. Improving sequence-based genotype calls with linkage disequilibrium and pedigree information. *The Annals of Applied Statistics*, 6(2):457–475, 2012.

[147] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, 2012.

[148] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms

for genome-wide association studies. *Nature Methods*, 11(4):407–409, 2014.

[149] Qianqian Zhu, Dongliang Ge, Jessica M Maia, Mingfu Zhu, Slave Petrovski, Samuel P Dickson, Erin L Heinzen, Kevin V Shianna, and David B Goldstein. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *The American Journal of Human Genetics*, 88(4):458–468, 2011.

[150] X. Zhu, S. Zhang, H. Zhao, and R.S. Cooper. Association mapping, using a mixture model for complex traits. *Genetic Epidemiology*, 23(2):181–196, 2002.

[151] Yun Zhu and Momiao Xiong. Family-based association studies for next-generation sequencing. *The American Journal of Human Genetics*, 90(6):1028–1045, 2012.