Published in final edited form as:

Curr Protoc Hum Genet. ; 84: 1.28.1–1.28.9. doi:10.1002/0471142905.hg0128s84.

Statistical Methods for Genome-Wide and Sequencing Association Studies of Complex Traits in Related Samples

Timothy Thornton*

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

Abstract

Genome-wide association studies (GWAS) and sequencing studies are routinely conducted for the identification of genetic variants that are associated with complex traits. Many genetic studies for association mapping include related individuals. When relatives are included in an association analysis, familial correlations must be appropriately taken into to ensure correct type I error and to increase power. This unit provides an overview of statistical methods that are available for GWAS and sequencing association studies of complex traits in samples with related individuals.

Keywords

Relatedness; GWAS; Sequence; Association Mapping; Family Data; Complex Traits; Mixed Models; Genome-wide association studies

Introduction

Recent advances in high-throughput genotyping and whole-genome/whole-exome sequencing technologies have facilitated unprecedented opportunities for the identification of genetic variants underpinning complex phenotypes. Many genetic association studies of complex traits sample related individuals as it is well known that samples with relatedness can offer advantages over samples that have only unrelated individuals. For example, the power to detect genetic association with a complex trait may be increased when including affected individuals with affected relatives as they have a higher expected frequency of causal variants compared to randomly sampled affected individuals (Teng and Risch, 1999). In sequencing studies, including relatives also allows for more reliable detection of sequencing error as well as increased power for identifying rare variants with extremely low frequencies that may be segregating in ascertained pedigrees but whose effects may not be detectable in unrelated population-based samples (Kazma and Bailey, 2011; Jiang and McPeck, 2013).

When related individuals are included in a genetic association study, familial correlations must be taken into account in order to ensure validity of the association results. This unit provides an overview of statistical methods for association testing in related samples. Methods for the analysis of genetic variants from genome-wide association studies (GWAS)

*Correspondence: tathornt@u.washington.edu.

and sequencing studies will be discussed, and the focus will be on approaches that are applicable to completely general samples with related individuals, i.e., methods that are appropriate for arbitrary combinations of family and unrelated samples. (For an overview of association methods that have been developed for family-based designs, see Unit 1.12 (Ewens, et al., 2008).) Table 1 provides a summary of the methods highlighted in this unit, and details of the methods are described in the following sections.

Key Concepts

GWAS and Sequencing Association Studies

The “common disease / common variant” hypothesis that common genetic variants in a population are largely responsible for common, complex diseases has been the basis for thousands of GWAS and candidate gene studies (Visscher et al., 2012). GWAS have successfully identified thousands of genetic variants that are associated with a variety of complex diseases and quantitative traits. Most genome-wide association studies perform single-variant association tests, where each variant in a genome-screen is tested individually for association with a phenotype.

The uncovered variants from GWAS have largely been of small effect and explain only a small fraction of trait heritability. Rare variants, defined here as variants with minor allele frequencies less than 1%-5%, likely play a significant role in many complex traits, and may explain some of the “missing heritability” not explained by the common variants identified through GWAS (Li and Leal, 2008; Manolio et al., 2009). Detecting rare variant associations from GWAS data is difficult due to rare variants having low linkage-disequilibrium (LD) with common variants on the single-nucleotide polymorphism (SNP) genotyping arrays used in GWAS. Thanks to advancements in high-throughput sequencing technologies (Shendure and Ji, 2008; Mardis 2008) rare variants across the genome can be identified from sequence data, and whole-genome and whole-exome sequencing studies are now routinely conducted for the identification of rare-variants that are involved with complex traits.

Measures of Genetic Relatedness

Statistical methods for detecting genetic associations in samples with related individuals use genetic relatedness measures to account for correlated genotypes among relatives in the analysis. Identity-by-descent (IBD) sharing probabilities and kinship coefficients are commonly used genetic relatedness measures for pairs of individuals. Alleles at a genetic locus that are inherited copies of the same ancestral allele are said to be IBD. The term IBD is generally used to refer to recent ancestry as opposed to ancient, common ancestry. Pairs of non-inbred individuals (i.e., individuals with unrelated parents) can share 2, 1, or 0 alleles IBD. For a pair of individuals i and j , we denote δ_{ij}^2 , δ_{ij}^1 , and δ_{ij}^0 to be the probability that i and j share 2, 1, and 0 alleles IBD, respectively, at a locus. The kinship coefficient for i and j , which we denote as ϕ_{ij} , is defined to be the probability that a random allele selected from i and a random allele selected from j at a locus are IBD. The kinship coefficient for a pair of individuals is defined to be the probability that a random allele selected from the first individual in the pair and a random allele selected from second individual in the pair at a

locus are IBD. The kinship coefficient can be written as a function of IBD sharing

probabilities, where for outbred individuals i and j , $\phi_{ij} = \frac{1}{2}\delta_{ij}^2 + \frac{1}{4}\delta_{ij}^1$. Table 2 gives IBD sharing probabilities and kinship coefficients for a few outbred relatives pairs at autosomal loci.

Association methods have also been developed for the analysis of X-chromosome variants in related samples, and these methods use X-linked kinship coefficients to account for IBD sharing on the X among relatives. Unlike the autosomal kinship coefficient, kinship coefficients for a relative pair on the X-chromosome depends on the sexes of the individuals. (See Thornton et al. (2012) and Schaid et al. (2013) for more details on X-linked kinship coefficients.)

Population Structure Confounding in Genetic Association Studies

Genetic association studies often have ancestry differences among sample individuals, or population structure. It is well known that failure to appropriately account for population structure can lead to spurious association in both GWAS and sequencing association studies. Confounding due to ancestry is a serious concern for genetic association studies. (See Unit 1.22, Edwards and Gao, 2012, for additional details on identifying and correcting for population structure in unrelated samples.) Different ethnic groups often share distinct dietary habits and other lifestyle characteristics that leads to many traits of interest being correlated with ancestry and/or ethnicity. Genetic loci across the genome that are not involved with a trait but have allele frequency differences among different ethnic groups that are represented in the sample can have highly significant associations. Principal components analysis (PCA) (Price et al., 2007) has been the prevailing approach for population structure correction in GWAS and sequencing association studies with unrelated samples. In samples with related individuals and population structure, however, the top principle components from PCA are not able to appropriately account for sample structure due to the correlated genotypes among relatives in association studies (Thornton and McPeck 2010). Appropriate methods for genetic association testing in the presence of population and pedigree structure in both GWAS and sequencing studies are discussed in the next section.

Strategic Approach

We consider genetic association testing of genetic variants from GWAS and sequencing studies in a variety of sample settings with related individuals. An overview of statistical methods for single-variant association testing, such as those commonly used in GWAS, and association testing with groups of variants in a genetic region of interest, such as methods used in sequence association studies of rare variants, is provided.

Single-SNP Association Testing with Known Pedigrees

Consider a genetic association study for a sample of n individuals from a single population with the kinship assumed to be known (e.g. having a defined pedigree structure). Assume that the phenotype of interest is a quantitative trait and let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ denote the $n \times 1$ phenotype vector, where Y_i is the quantitative trait value for individual i . Consider testing an autosomal SNP for association with the phenotype, where $\mathbf{g} = (g_1, \dots, g_n)^T$ is an $n \times 1$ vector

of the genotypes at the SNP for the n sample individuals, where $g_i = 0, 1$, or 2 , according to whether individual i has, respectively, $0, 1$ or 2 copies of the reference allele at the SNP. Let $\mathbf{X}$ be a $n \times k$ matrix of covariates that are relevant to the phenotype, where the matrix of k covariates includes a column of 1's for the intercept. The following linear mixed-effects (LME) model has been proposed for genetic association testing of a fixed SNP effect while simultaneously accounting for polygenic random effects that are correlated among related individuals and random residual effects accounting for environment and/or measurement error:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{g}\alpha + \delta + \varepsilon \quad \text{Equation 1}$$

where β is a $k \times 1$ vector of covariate effects, α is the (scalar) association parameter of interest measuring the effect of genotype on phenotype with α equal to 0 when there is no SNP effect, and δ and ε are length n vectors of random polygenic effects with $\delta \sim N(\mathbf{0}, \sigma_a^2 \Phi)$ and $\varepsilon \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$, where σ_a^2 represents additive genetic variance, σ_e^2 represents environmental variance, Φ is a known kinship matrix with (i, j) -th entry equal $2\phi_{ij}$ (twice the kinship coefficient for individuals i and j), and $\mathbf{I}_n$ is an n -dimensional identity matrix. The GTAM association statistic (Abney et al., 2002) is a score statistic for the LME model given by equation 1 for testing the null hypothesis $H_0: \alpha = 0$ versus the alternative hypothesis is $H_A: \alpha \neq 0$, where the nuisance parameters β , σ_a^2 , and σ_e^2 are estimated under the null hypothesis by using either maximum likelihood or restricted maximum likelihood (REML) estimation. Similar LME approaches to GTAM have also been proposed for genetic association testing with a single SNP in related samples with known pedigrees (Abecasis et al., 2000; Jakobsdottir and McPeck, 2013).

The previously mentioned LME regression methods for quantitative are based on an assumption that the quantitative trait of interest is normally distributed. Many quantitative traits of interest, however, are not normally distributed and the GQLS method (Feng et al. 2011) has been proposed for association mapping of general quantitative traits in related individuals. GQLS is a retrospective analysis approach that treats genotype data at an autosomal marker as the random variable and phenotype as a covariate in a generalized linear regression model. A retrospective analysis allows for the phenotype to have any continuous distribution, and thus is applicable to general quantitative traits, unlike the mixed-model methods.

Statistical methods for dichotomous trait association mapping in related samples have also been proposed. Now consider association testing with a dichotomous phenotype $D = (D_1, \dots, D_n)^T$, where D_i is equal to 1 if individual i is affected with the trait and 0 otherwise. For case-control association testing of $\mathbf{D}$ and a SNP genotype vector $\mathbf{g} = (g_1, \dots, g_n)^T$ in related samples, Slager and Schaid (2001) proposed an extension of the Armitage trend test (Sasieni 1997) for unrelated samples where relatedness among sample individuals is accounted for in the variance calculation of the test statistic by using kinship coefficients. Similar case-control association tests for related samples were also proposed by Bourgain et al. (2004). Thornton and McPeck. (2007) developed the MQLS case-control association test, which is an optimal score statistic for detecting genetic associations in related samples when a SNP

has a small effect size, such as the very modest effect sizes that are expected for complex traits. Statistical methods are also available for case-control association testing with haplotypes in related samples (Browning et al. 2005; Wang and McPeck, 2009).

GWAS with Related Samples and Cryptic Structure

GWAS often have underlying sample structures due to both relatedness and population structure. The statistical methods discussed in the previous subsection may not be valid when there is cryptic relatedness and/or ancestry differences among sample individuals. For complex trait association mapping with hidden structure in sample, Kang et al (2010) proposed an LME method named EMMAX, which can be viewed as an extension of the previously discussed GTAM method to samples with cryptic sample structure. EMMAX is based on the LME model given by equation (1) but with the known genetic relatedness matrix Φ used for the variance-covariance matrix of the polygenic random effects replaced with an empirical genetic relatedness matrix (GRM), $\hat{\Phi}$, that is estimated using high-density SNP genotyping data from across the autosomal chromosomes. The EMMAX approach is a powerful and effective method of choice for association mapping of quantitative traits in the presence of both population structure and relatedness, and a number of similar computationally efficient LME approaches for large-scale GWAS have also been proposed including GEMMA (Zhou and Stephens, 2010), GRAMMAR-Gamma (Svishcheva et al. 2012), and FaST-LMM (Lippert et al., 2011).

For valid case-control association testing in samples with both population and pedigree structure, the ROADTRIPS method (Thornton and McPeck, 2010) can be used. Similar to the previously mentioned LME methods for cryptic structure, ROADTRIPS uses a covariance matrix estimated from genome-screen data to correct for unknown structure and essentially extends the aforementioned case-control association methods with known pedigree structure to the contexts when sample structure, pedigree and population, is partially or completely unknown.

Burden and Kernel Rare-Variant Association Tests

The single-variant quantitative and case-control tests discussed in the previous two subsections have been shown to work well for detecting genetic associations with common variants. Single-marker analyses, however, have essentially no power for detecting associations with rare variants due to their low frequencies. A popular strategy for detecting rare variant associations from sequencing data is to jointly consider all rare variants in a genetic region or gene for association testing. Most association methods for rare variants can be classified into two groups: burden association tests and kernel association tests (Lee et al. 2012; Schaid et al. 2013). Burden tests aggregate all rare variants across a genetic region in an individual into a single value, where the variant sites can be unweighted (Liu and Leal, 2010), weighted according to minor allele frequencies (Madsen and Browning, 2009), or assigned weights based on other features of the data such as functionality (Liu and Leal, 2010; Lin and Tang, 2011). A sampled individual's aggregate value across a genetic region can be viewed as the individual's genetic "burden" score, and burden tests evaluate the association of the burden scores (or the variant-sum across the region) and phenotype.

Kernel association methods do not aggregate variants into a single value as the burden tests do. Instead, kernel-based methods, such as the widely-used SKAT method (Wu et al., 2011), use a LME approach that aggregates individual variant statistics that measure strength of association with each variant site, where a kernel matrix specifying genetic similarities in the region for pairs of subjects is included in the analysis. More details on the kernel matrix are given in the next subsection.

Burden tests have higher power than kernel-based methods in unrelated samples when a large proportion of the rare variants in a region are causal and with effects on phenotype that are in the same direction, while kernel-based methods have higher power when a large proportion of the variants are non-causal or if there is a combination of both risk and protective variants in a region that are influencing the phenotype (Wu et al., 2011). The SKAT-O (where the “O” stands for optimal) method (Lee et al. 2012) is a unified rare-variant association approach that combines features of both burden and kernel methods for detecting genetics associations, and the method has high-power in a variety of rare-variant association settings.

Association Testing with Sequencing Data in Related Samples

Burden and kernel-based approaches for rare variant association testing have largely been developed for unrelated samples. Recently, statistical methods have been developed for testing associations with rare variants from sequencing data in samples that include related individuals. Chen et al. (2013) proposed the FamSKAT method, which can be viewed as an extension of the sequence kernel association test (SKAT) method for unrelated individuals to pedigree samples. Let $\mathbf{Y}$ and $\mathbf{X}$ be the previously defined vector of quantitative traits and matrix of covariates, respectively, for a sample of n individuals, and let $\mathbf{G}$ be an $n \times m$ matrix containing the genotype values for the sample individuals at m variant sites in a genetic region of interest (possibly consisting of rare variants or a combination of rare and common variants from sequencing data) with elements G_{ij} having values of 0, 1, or 2 for the number of minor alleles that individual i has at the j -th variant site. The LME model of FamSKAT for jointly testing a group of variants for association with a quantitative phenotype in a related sample is

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{G}\mathbf{W}\gamma + \delta + \varepsilon \quad \text{Equation 2}$$

where γ is a $m \times 1$ vector of random effects for the variant sites in the genetic region of interest with $\gamma \sim N(\mathbf{0}, \tau \mathbf{I}_m)$, τ is the variance component for the variant sites, $\mathbf{I}_m$ is an m -dimensional identity matrix, $\mathbf{W}$ is a diagonal matrix of prespecified weights for each marker or variant sites (based on some particular features of the variants, such as minor or allele frequencies or functional information), and the remaining parameters are defined exactly as in the LME model given by Equation 1, i.e., β is a $k \times 1$ vector of covariate effects, and δ and ε are random polygenic and environmental effects, respectively, with $\gamma \sim N(\mathbf{0}, \sigma_a^2 \Phi)$ and $\varepsilon \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$. Note that m SNP effects of the variant sites in the LME model of Equation 2 are modeled as random effects which is a different than the proposed LME model of Equation 1 where the SNP effect is modeled as a fixed effect. The FamSKAT association

statistic is a score statistic for testing the null hypothesis $H_0: \tau = 0$ versus $H_A: \tau > 0$ for the LME model of Equation 2, which is equivalent to testing the null hypothesis of $H_0: \gamma = \mathbf{0}$ versus $H_A: \gamma \neq \mathbf{0}$, but without requiring an m -degree of freedom test for detecting genetic associations for all variant sites, which would have little to no power when testing multiple variants with small effect sizes. Note that the variance-covariance matrix of $\mathbf{Y}$ is a function of the linear “kernel” matrix $\mathbf{K} = \mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}^T$, which is a matrix of genetic similarity measures for pairs of individuals across the genetic region. Non-linear kernel matrices have also been proposed for rare variant association testing (Wu et al. 2011).

Other rare-variant association test for unrelated samples have also recently been extended to related samples. Jiang and McPeck (2014) proposed the MONSTER method, which is an extension of the previously mentioned SKAT-O method to pedigree-based samples, and both Chen et al. (2013) and Schaid et al. (2013) developed burden rare-variant tests that are valid for samples with related individuals.

X-chromosome Association Testing

Statistical methods for detecting genetic association have primarily been developed for the analysis of markers on the autosomes, with significantly less attention given to the analysis of markers on the X chromosome, despite the potential for identifying X-linked genes that influence complex traits. Autosomal association methods, however, will typically not be valid for X chromosome analysis. Some of the methodological challenges for the X include (1) accounting for X chromosome copy number differences in females and males and (2) appropriately adjusting for genetic correlations among sample individuals on the X, including pedigree structure, which can be quite different from the autosomes for relatives. A few statistical methods have been proposed for association testing with X-linked markers. Zheng et al. (2007) and Clayton (2008) proposed methods for X chromosome case-control association analysis in samples of unrelated individuals from a single population. For case-control association testing with markers on the X chromosome in samples with related individuals, Thornton et al. (2012) developed the XM test, which is an extension of the MQLS autosomal association method to the X-chromosome. Schaid et al. (2013) developed both kernel and burden tests for X-linked rare variation association testing in related samples with sequencing data.

Commentary

Many genetic association studies include individuals with some degree of relatedness, and this unit highlights a variety of statistical methods and approaches for associations testing of dichotomous and quantitative traits with related individuals. The focus is on methods that are applicable to general samples with related individuals, including samples with known and/or cryptic population and pedigree structure.

Linear mixed effects (LME) methods have emerged as a powerful and effective approach for association mapping with quantitative traits in samples with population structure as well as family structure or cryptic relatedness. A large and continually expanding array of LME methods are available for GWAS with sample structures. A recent study (Euhansunthornwattana et al., 2014) compared the performance of a variety of different LME

approaches (and software implementations). There was a very high concordance of the association results across the different LME approaches, thus indicating that the choice of LME implementation for a given study can largely be based on considerations relating to computation time, memory usage, and convenience.

Existing LME methods can suffer a substantial loss in power for detecting genetic associations in ascertained case-control GWAS due to ascertainment bias where disease cases are overly represented in a sample relative to disease prevalence (Yang et al. 2014). For genetic association testing in case-control samples with partially or completely unknown sample structure, the ROADTRIPS method is recommended, while other approaches, such as the MQLS method, can be used to detect association in samples from a single population with known pedigree relationships.

Statistical methods for sequencing association studies are still in their infancy. To date, there are only a limited number of methods available for rare variant association testing in related samples. Schaid et al. (2013) and Chen et al. (2013) proposed rare variant burden tests adapted for related samples, while the FamSKAT method (Chen et al., 2013) extended the SKAT rare variant association method for independent samples to pedigrees. The aforementioned burden tests are expected to have higher power than FamSKAT in related samples when most variants in a genetic region are causal and the effects are in the same direction, while FamSKAT is expected to be more powerful than the burden tests when a large fraction of the variants in a genetic region are either noncausal or when there are multiple causal variants in the region with effects that are in different directions. The recently proposed MONSTER method (Jiang and McPeck, 2014) provides a unified approach that optimally combines features of the burden tests and FamSKAT for high power in both scenarios.

Care must be taken in using PCA analysis for correcting for population structure in genetic association studies. PCA and related approaches for population structure inference are not able to incorporate rare variants in the analysis due to the standard genetic relatedness matrix becoming unstable with low minor allele frequency variants. As a result, sequencing studies often use principal components that are calculated from common genetic variants to correct for population structure. Recent work (O'Conner et al., 2013), however, has shown that population structure for rare-variants can be quite different from than the structure of common-variants. In addition, variants that are very rare are often private to a single population. Additional research and methodology development is required to obtain appropriate correction of population structure for rare variant association analyses from sequencing data as well as to understand the impact of population structure on rare variants.

Large-scale genomic technologies have completely transformed human genetic studies of complex traits. The available statistical methods for the analysis of whole-genome sequence data, however, have limitations, and more powerful and computationally efficient methods will be required for the detection of both rare and common causal genetic variants of small effect sizes that have not yet been uncovered by existing methodology. At the same time, statistical methods that can leverage both the rich GWAS SNP genotyping data, including data from the many large genotyped cohorts and family-based studies, and the large amounts

of sequencing data being generated will be essential for a greater understanding of the genetic factors that are contributing to complex diseases and quantitative traits.

Acknowledgments

This study was supported in part by the National Institutes of Health grants K01 CA148958 and P01 HG0099568

Literature Cited

- Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet.* 2000; 66:279–292. [PubMed: 10631157]
- Abney M, Ober C, McPeck MS. Quantitative- trait homozygosity and association mapping and empirical genome-wide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *Am J Hum Genet.* 2002; 70:920–934. [PubMed: 11880950]
- Browning SR, Briley JD, Briley LP, Chandra G, Charnecki JH, et al. Case-control single-marker and haplotypic association analysis of pedigree data. *Genet Epidemiol.* 2005; 28:110–122. [PubMed: 15578751]
- Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol.* 2013; 37:196–204. [PubMed: 23280576]
- Clayton D. Testing for association on the X chromosome. *Biostatistics.* 2008; 9:593–600. [PubMed: 18441336]
- Edwards TL, Gao X. Methods for Detecting and Correcting for Population Stratification. *Curr Protoc Hum Genet.* 2012; 73:1.22.1–1.22.14. 1.22.
- Eu-ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SM, Blackwell JM, Cordell HJ, Wellcome Trust Case Control Consortium 2. Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. *PLoS Genetics.* 2014; 10:e1004445. [PubMed: 25033443]
- Ewens WJ, Spielman RS, Kaplan NL, Gao X, Morris RW, Martin ER. Disease Associations and Family-Based Tests. *Curr Protoc Hum Genet.* 2008; 58:1.12.1–1.12.24. 1.12.
- Feng Z, Wong WW, Gao X, Schenkel F. Generalized genetic association study with samples of related individuals. *Ann Appl Stat.* 2011; 5:2109–2130.
- Jakobsdottir J, McPeck MS. MASTOR: Mixed-Model Association Mapping of Quantitative Traits in Samples with Related Individuals. *American Journal of Human Genetics.* 2013; 92:652–666. [PubMed: 23643379]
- Jiang D, McPeck MS. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genetic Epidemiology.* 2014; 38:10–20. [PubMed: 24248908]
- Kang HM, Sul JH, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics.* 2010; 42:348–354. [PubMed: 20208533]
- Kazma R, Bailey JN. Population-based and family-based designs to analyze rare variants in complex diseases. *Genet Epidemiol.* 2011; 35(Suppl 1):S41–S47. [PubMed: 22128057]
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet.* 2012a; 91:224–237. [PubMed: 22863193]
- Li B, Leal SM. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to the Analysis of Sequence Data. *American Journal of Human Genetics.* 2008; 83:311–321. [PubMed: 18691683]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38:904–909. [PubMed: 16862161]
- Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet.* 2011; 89:354–367. [PubMed: 21885029]

- Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nature Methods*. 2011; 8:833–835. [PubMed: 21892150]
- Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet*. 2010; 6(10):e1001156. [PubMed: 20976247]
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009; 5(2):e1000384. [PubMed: 19214210]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. *Nature*. 2009; 461:747–753. [PubMed: 19812666]
- Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*. 2008; 9:387–402. [PubMed: 18576944]
- O'Connor TD, Kiezun A, Bamshad M, Rich SS, Smith JD, et al. Fine-scale patterns of population stratification confound rare variant association tests. *PLoS One*. 2013; 8:e65834. [PubMed: 23861739]
- Risch N, Merikangas K. The Future of Genetic Studies of Complex Human Diseases. *Science*. 1996; 273:1516–1517. [PubMed: 8801636]
- Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics*. 1997; 53:1253–1261. [PubMed: 9423247]
- Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SM. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet Epidemiol*. 2013; 37:409–418. [PubMed: 23650101]
- Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008; 26:1135–1145. [PubMed: 18846087]
- Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. Rapid variance components-based method for whole-genome association analysis. *Nature Genetics*. 2012; 44:1166–1170. [PubMed: 22983301]
- Teng J, Risch N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex diseases. II. Individual genotyping. *Genome Res*. 9(1999):234–241. [PubMed: 10077529]
- Thornton T, McPeck MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am J Hum Genet*. 2007; 81:321–337. [PubMed: 17668381]
- Thornton T, McPeck MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet*. 2010; 86(2):172–184. [PubMed: 20137780]
- Thornton T, Zhang Q, Cai X, Ober C, McPeck MS. XM: association testing on the X-chromosome in case-control samples with related individuals. *Genet Epidemiol*. 2012; 36(5):438–450. [PubMed: 22552845]
- Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. *American Journal of Human Genetics*. 2012; 90:7–24.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011; 89:82–93. [PubMed: 21737059]
- Wang Z, McPeck MS. An Incomplete-Data Quasi-Likelihood Approach to Haplotype-Based Genetic Association Studies on Related Individuals. *Journal of the American Statistical Association*. 2009; 104, 1251–1260.
- Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*. 2014; 46:100–106. [PubMed: 24473328]
- Zheng G, Joo J, Zhang C, Geller NL. Testing association for markers on the X chromosome. *Genet Epidemiol*. 2007; 31:834–843. [PubMed: 17549761]
- Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*. 2012; 44:821–824. [PubMed: 22706312]

Table 1
Summary Table of Association Methods for General Samples with Relatedness

Methods	Sample	Population Structure Adjustment	Genetic Data	Phenotypes	Methodology
GTAM (Abney et al., 2002); MASTOR (Jakobson et al., 2013)	Known Pedigrees	No	GWAS SNPs	Quantitative	Linear mixed effect (LME) model with known kinship matrix to account for pedigree relationships
EMMAX (Kang et al., 2010); FaST-LMM (Lippert et al., 2011); GRAMMAR-Gamma (Svishcheva et al., 2012); GEMMA (Zhou and Stephens, 2012)	Known or Cryptic Relatedness	Yes	GWAS SNPs	Quantitative	LME model with an empirical genetic relatedness matrix (GRM) to account for cryptic pedigree and population structure
GQLS (Feng et al., 2011)	Known Pedigrees	No	GWAS SNPs	General Traits	Regression Model with pedigree-based kinship coefficients for relatedness adjustment
MQSL (Thornton and McPeck 2007); Corrected-Arimage Trend test (Slager and Schaid, 2001); WQLS (Bourgain et al. 2003)	Known Pedigrees	No	GWAS SNPs	Case-Control	Regression model with adjustment for relatedness via a known kinship matrix
XM (Thornton and McPeck, 2012)	Known Pedigrees	No	X-linked GWAS SNPs	Case-Control	Regression model with X-linked kinship coefficients to account for known relatedness on the X
ROADTRIPS (Thornton and McPeck, 2010)	Known or Cryptic Relatedness	Yes	GWAS SNPs	Case-Control	Regression model with adjustment for cryptic population and pedigree structure via an empirical GRM
FamSKAT (Chen et al., 2013); MONSTER (Jiang and McPeck, 2014)	Known Pedigrees	No	Sequencing data	Quantitative	LME kernel-based model with known kinship matrix to adjust for pedigree relatedness.
Statistical methods of Schaid et al. (2013)	Known or Cryptic Relatedness	Yes	Autosomal or X-linked sequencing data	Case-Control	Kernel and burden methods with either known or empirical autosomal and X-linked kinship coefficients for relatedness adjustment

Table 2
Kinship Coefficients and IBD Sharing Probabilities for Outbred

Outbred Relationship	ϕ (Kinship Coefficient)	δ_2 (Probability of IBD=2)	δ_1 (Probability of IBD=1)	δ_0 (Probability of IBD=0)
Monozygotic-Twins	0.5	1	0	0
Parent-Offspring	0.25	0	1	0
Full-Siblings	0.25	0.25	0.5	0.25
Second-degree	0.125	0	0.5	0.5
Third-degree	0.0625	0	0.25	0.75
Fourth-degree	0.03125	0	0.125	0.875
Unrelated pairs	0	0	0	1

In the table, ϕ is the kinship coefficient and δ_0 , δ_1 , δ_2 , are the probability of sharing 2, 1, and 0 alleles IBD, respectively.