Identifying Rare Variants Associated with Complex Traits via Sequencing

Bingshan Li,[1] Dajiang J. Liu,[2] and Suzanne M. Leal[3]

[1]Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee
[2]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan
[3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas

ABSTRACT

Although genome-wide association studies have been successful in detecting associations with common variants, there is currently an increasing interest in identifying low-frequency and rare variants associated with complex traits. Next-generation sequencing technologies make it feasible to survey the full spectrum of genetic variation in coding regions or the entire genome. The association analysis for rare variants is challenging, and traditional methods are ineffective, however, due to the low frequency of rare variants, coupled with allelic heterogeneity. Recently a battery of new statistical methods has been proposed for identifying rare variants associated with complex traits. These methods test for associations by aggregating multiple rare variants across a gene or a genomic region or among a group of variants in the genome. In this unit, we describe key concepts for rare variant association for complex traits, survey some of the recent methods, discuss their statistical power under various scenarios, and provide practical guidance on analyzing next-generation sequencing data for identifying rare variants associated with complex traits. *Curr. Protoc. Hum. Genet.* 78:1.26.1-1.26.22. © 2013 by John Wiley & Sons, Inc.

Keywords: complex traits • rare variants • association tests • aggregation analysis • exome • sequencing

INTRODUCTION

To date, numerous genes involved in disease and trait etiology have been identified through linkage and association studies. For Mendelian diseases, usually linkage analysis is used to localize the genomic regions harboring causal variants, and then fine mapping methods are used to pinpoint causal genes and variants (see *UNIT 1.19* for linkage analysis). Thus far, the underlying genetic cause of ~3,500 Mendelian disorders is known (*http://www.ncbi.nlm.nih.gov/omim*). Unlike Mendelian diseases, which are caused by rare high-penetrant genetic variants, the genetic basis of complex traits remains largely unknown. Until recently, the pursuit of an understanding of the genetic etiology of complex traits has been almost solely based on the common-disease common-variants (CDCV) hypothesis (Smith and Lusis, 2002; Hirschhorn and Daly, 2005; Iyengar and Elston, 2007; Schork et al., 2009), which asserts that common complex diseases are due to common variants—e.g., minor allele frequency (MAF) >0.05—with usually little or no allelic heterogeneity within a locus. Single-nucleotide polymorphisms (SNPs) are the most prevalent form of common genetic variation and become de facto markers for localizing common disease-causing variants. Because neighboring common variants can be in strong linkage disequilibrium (LD) and thus provide redundant information, it is only necessary to survey a subset of SNPs (i.e., tag-SNPs) to achieve genome-wide coverage (see also *UNIT 1.4*). When disease-causing common variants are in strong LD with one or more tagSNPs, the genetic effects of causal variants lead to association signals that are observable in tagSNPs; this is the basis for Genome-Wide Association Studies (GWAS), which makes genotyping genome-wide tagSNPs on thousands of samples using microarray chips a cost-effective strategy. By design, this is an indirect mapping strategy that rarely directly tests causal variants. To date, >8,500 associated SNPs have been identified for a variety of complex traits (the

Current Protocols in Human Genetics 1.26.1-1.26.22, July 2013
Published online July 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471142905.hg0126s78

National Human Genome Research Institute Catalog of Published Genome-Wide Association Studies, *http://www.genome.gov/gwastudies*). Nevertheless, most identified common associated SNPs only have weak genetic effects (e.g., odds ratio <1.5), and collectively these identified variants only account for a small proportion of heritability for most complex traits (Maher, 2008; Manolio et al., 2009), suggesting that other major mechanisms are involved in the genetic etiology of complex traits. Of great interest is the common disease rare variants (CDRV) hypothesis (Smith and Lusis, 2002; Iyengar and Elston, 2007; Schork et al., 2009), which asserts that common diseases are due to rare variants, and that contrary to the CDCV hypothesis, rare variants often exhibit extreme allelic heterogeneity. To date, several studies have unraveled functional roles of rare variants in complex traits (Cohen et al., 2004; Ahituv et al., 2007; Romeo et al., 2007; Bodmer and Bonilla, 2008; Ji et al., 2008), making studying the role of rare variants in complex trait etiology an attractive avenue to pursue. Rare variants are inefficiently tagged in GWAS due to the low correlation between rare variants, e.g., MAF <1%, and common tagSNPs with a much higher MAF, e.g., >5% (Li and Leal, 2008). Therefore, associations with causal rare variants will usually be missed in GWAS. Sequencing can uncover the full spectrum of genetic variation and is the optimal approach to identifying rare genetic variants for association studies. Due to the advancement of cost-effective next-generation sequencing (NGS) technologies (Mardis, 2008; Shendure and Ji, 2008), it is now feasible to sequence whole exomes (i.e., the protein coding regions) in hundreds or even thousands of samples, and soon it will be practical to sequence whole genomes. Because of the continuously decreasing cost of sequencing, our understanding of the allelic architecture of complex traits will accumulate gradually in the near future, when more whole-genome or -exome sequencing studies are carried out.

Traditional analysis methods used in GWAS are single-marker tests, where SNPs are tested individually and multiple testing is corrected to control the family-wise error rate (FWER). Due to the extensive LD among common SNPs, permutation-based approaches can be used to account for the inter-SNP correlation. For GWAS, however, using permutation-based methods to control for FWER is computationally intensive, and so it is not practical for large-scale studies. An alternative approach is to estimate the effective number of independent tests based on the LD patterns and to use the Bonferroni approach to control for FWER. In practice, a *p*-value of $<5 \times 10^{-8}$ is used to declare genome-wide statistical significance based on European populations, corresponding to correcting for one million independent tests (Dudbridge and Gusnanto, 2008). This cutoff is still used even if less or more than one million tests are performed. Although testing individual variants for association can also be done for rare variants, it is inevitably underpowered (Li and Leal, 2008), unless the sample size is excessively large. To address the challenges, numerous novel analysis strategies have been developed to identify rare variants associated with complex traits. This battery of new methods, often referred to as collapsing, group-wise, or pooled approaches, generally involves aggregating multiple rare variations in a gene or region or any arbitrary set in the genome, and testing the association effect of the group of rare variants as a whole. In this commentary, we describe some key concepts related to rare variants, summarize recently developed aggregation methods, provide practical guidance on the analysis of rare variation obtained from sequencing, and discuss further challenges that need to be addressed for rare variant association studies.

KEY CONCEPTS

LD, Association Mapping, and Rare Variants

Recall that linkage disequilibrium (LD) is the nonrandom association of alleles at different loci, and the co-occurrence of alleles at two loci on the same haplotype is either more or less frequent than expected from random pairing of the two alleles based on their allele frequencies (Hartl and Clark, 2007). Let *A* and *a* denote the two alleles at locus 1, and *B* and *b* represent the two alleles at locus 2, with corresponding allele frequencies of p_A, $p_a = 1 - p_A$ for locus 1 and p_B, $p_b = 1 - p_B$ for locus 2. Without loss of generality (WLOG), we assume that the minor alleles at the two loci are *A* and *B*, respectively, and $p_A < = p_B$. Denote the frequency of the two-locus haplotype H_{AB} as p_{AB}. Then the LD between the two loci is $D_{AB} = p_{AB}$-p_A*p_B, where p_A*p_B is the expected frequency of H_{AB} under the assumption of no association between the two loci (i.e., no LD).

The concept of LD is the basis for LD-based indirect association mapping, where an investigator analyzes a variant that is in LD with the causal variant instead of directly analyzing the causal variant. For example, when the underlying causal allele is A at locus 1 and $D_{AB}! = 0$, the frequency of allele A in cases is different from that in controls. Due to the LD between A and B, we will also observe differential frequencies of allele B in cases versus controls. Therefore it is not necessary to directly test the causal SNPs, and LD-based association mapping is the strategy that was most commonly employed prior to the sequencing era. There are different measures of LD, and for association studies the most relevant measure is r^2, which is the correlation coefficient between alleles at the two loci. It can be calculated as $r^2 = D_{AB}{}^2/p_A(1 - p_A)p_B (1 - p_B) = (p_{AB} - p_A{*}p_B)^2/ p_A(1 - p_A)p_B (1 - p_B)$. There is a simple relationship between the sample size and r^2 when the underlying disease model is multiplicative; that is, for fixed statistical power, the sample size required is inversely proportional to r^2 (Pritchard and Przeworski, 2001). For example, the sample size needs to be doubled to achieve the same power if the r^2 between the tag SNP and the casual variant is 0.5. It is obvious that the higher r^2 is, the better power for association studies.

There is extensive LD in the human genome. The HapMap project (Altshuler et al., 2010) comprehensively surveyed the LD among common variants in various populations. A necessary condition to achieve high r^2 is the similarity of the allele frequencies at two loci. For given allele frequencies, the maximum r^2 is achieved, but not necessarily equal to 1, when the two minor alleles are on the same haplotype and the haplotypes were not broken by historical recombination events. In this case, the two loci are in complete LD (i.e., D' = 1) and only three haplotypes (H_{AB}, H_{aB}, and H_{BB}, with $p_{AB} = p_A$) are observed in the population. The maximum r^2 between the two loci is $(p_{AB}\text{-}p_A{*}p_B)^2/p_A(1 - p_A)p_B (1 - p_B) = p_A(1 - p_B)/p_B(1 - p_A)$. Only when $p_A = p_B$, i.e., the variant frequency is exactly the same for the two loci, $r^2 = 1$; in this case, only two haplotypes (H_{AB} and H_{ab}) are observed and the two loci are in perfect LD. When the allele frequencies of two loci are very different, the r^2 will never be very large. To see this, let's assume $p_A << p_B$ and $1 - \mathrm{p_A} \approx 1$. Then the maximum $r^2 \approx p_A/p_B{*} (1 - p_B)$, which is $<<1$, since p_A is very small compared with p_B. This indicates that causal rare variants are mostly likely to be missed in GWAS for single-marker tests, since GWAS chips are designed to include predominantly common variants (e.g., MAF > 0.05).

When considering haplotypes across multiple SNPs, some rare causal variants are likely to be tagged by rare haplotypes, and methods based on rare haplotype analysis (Li et al., 2010a; Zhu et al., 2010) may be used to identify such signals. It is expected, however, that the majority of unobserved causal rare variants reside on common haplotypes, and rare haplotype-based methods may tag only a small proportion of causal rare variants. Other strategies are therefore in order for mapping rare variants associated with complex traits. Let's assume A and B are the minor alleles of two rare variants, that is, $p_A << 1$, $p_B << 1$, and $1 - p_A \approx 1$, $1 - p_B \approx 1$. We also assume WLOG that the A allele was introduced in the population later than allele B. By chance, it is more likely that allele A occurred on the haplotypes not carrying the B allele. In this case there are three haplotypes (H_{Ab}, H_{aB}, and H_{ab}) after the introduction of the A allele, and $r^2 = (0 - p_A{*}p_B)^2/p_A(1 - p_A)p_B(1 - p_B) \approx \mathrm{p_A p_B} \approx 0$. Due to the extremely weak r^2 between rare variants, it is clear that LD-based indirect mapping via traditional genotyping is not a viable option for rare variants.

Thus, sequencing seems to be the optimal approach to uncover and identify associated rare variants. Since association tests are performed directly on potentially causal variants, sequencing-based association analysis is a direct mapping approach, but without the need for the fine mapping step following LD-based association studies. Traditional Sanger sequencing (Sanger et al., 1977) is laborious, low-throughput, and expensive. On the other hand, NGS technologies make it feasible to sequence targeted regions, exomes, and whole genomes of thousands of samples, and hold great promise for genetic studies of complex traits. The 1000 Genomes Project (1000 Genomes Project Consortium, 2010; Abecasis et al., 2012) utilized NGS platforms to provide a comprehensive catalog of genetic variation in various populations. NGS is routinely used for genetic studies of various traits and is expected to reveal a comprehensive allelic architecture and genetic etiology for complex traits in the near future.

Sequencing Strategies

For sequencing studies, one of the key factors that determines the accuracy and completeness of the underlying genetic variation spectrum is the sequencing depth of coverage, which is defined as the average number of reads mapping to each position in the genome (see *UNITS 18.2, 18.3, & 18.4*). For example, let the total number of sequenced reads be N, each with R bases. The depth of coverage can be calculated as $N*R/L$, where L is the length of the genomic regions. For genome-level coverage, $L = \sim 3 \times 10^9$, the total length of the genome. For exome-level coverage, L is the total length of corresponding exonic target regions for a particular capturing technology. Although high-depth (e.g., >30×) whole-genome sequencing (WGS) is ideal for complete surveys of genomic variants, high-coverage sequence data is associated with high costs. To date, this strategy is still not practical for sequencing the whole genome of thousands of samples.

Two alternative strategies, each with specific goals, have been commonly used in current large-scale sequencing studies: low-depth WGS and high-depth exome sequencing. Low-depth WGS is designed to sequence the whole genome to 4× to 6× and is a cost-effective approach to having whole-genome coverage of genetic variation. Due to insufficient coverage of each position for inferring underlying genotypes, a haplotype-based approach, e.g., MaCH/Thunder (Li et al., 2010b; Li et al., 2011; Howie et al., 2012) or IMPUTE-2 (Howie et al., 2009; Howie et al., 2012) is required to jointly call genotypes, utilizing the LD among variants across the genome. Due to extensive LD among common variants, this approach is best suited for studying common and low-frequency variants (e.g., MAF > 1%). As previously discussed, however, the LD (in terms of r^2) among rare variants or between rare and common variants is extremely low and the LD-based joint calling may not be ideal for very rare variants (e.g., MAF < 0.5%). Alternatively, exome sequencing selectively sequences to a high depth the coding regions of the genome (∼30 megabases) (Ng et al., 2009). Although the exome occupies only ∼1% of the genome, it is estimated to harbor ∼85% of disease-causing variants (Choi et al., 2009). This number may change dramatically, however, when noncoding regions are extensively studied through WGS. Exome sequencing starts with the targeted capturing of coding sequences. Next, the enriched exomes are sequenced to deep coverage (e.g., >100×). Since exome sequencing is still considerably less expensive than WGS and promises to identify causal coding variants, exome sequencing is currently the most popular approach for studying the genetic etiology of Mendelian and complex traits. Although both rare and common variants in the coding region can be accurately inferred from sequencing, it for the most part excludes noncoding regions and is likely to miss regulatory variants. With the continuously decreasing cost of NGS, in the next few years high-coverage WGS will become a common way of studying the noncoding portion of the genome.

Likelihood Models for Genetic Association

Let n be the number of individuals in a sample, c be the number of covariates to be included, and k be the total number of variants to be tested in a gene. In this unit we use "gene" to refer to any collection of variants that are to be analyzed together, e.g., a gene, a region, a pathway, or any arbitrary set of variants in the genome. For $i = 1, \ldots, n$, let y_i be the phenotype of the ith individual; for $i = 1, \ldots, n$, $j = 1, \ldots, k$, let X_{ij} denote the number of rare alleles the ith individual carries at the jth variant; for $i = 1, \ldots, n$, $j = 1, \ldots, c$, let Z_{ij} denote the value of the jth covariate of the ith individual. We represent the genotype and covariate data of the ith individual in vector form

$$X_i = \begin{bmatrix} X_{i1} \\ \cdots \\ X_{ik} \end{bmatrix} \qquad Z_i = \begin{bmatrix} 1 \\ Z_{i1} \\ \cdots \\ Z_{ic} \end{bmatrix}$$

Similarly, we let X_j denote the genotypes of the jth variant across all samples, and use $\mathbf{X}$ and $\mathbf{Z}$ to represent the data matrix of genotypes and covariates respectively, where $\mathbf{X}$ is an n by k matrix and $\mathbf{Z}$ is an n by c matrix. We can represent the genotype-phenotype relationship in the regression framework in the following way:

$$y_i = \boldsymbol{\beta}' X_i + \boldsymbol{\gamma}' Z_i + \varepsilon = \sum_{j=1}^{k} \beta_j X_{ij} + \sum_{j=1}^{c} \gamma_j Z_{ij} + \varepsilon \qquad (1)$$

$$\mathrm{logit}(p_i) = \boldsymbol{\beta}' X_i + \boldsymbol{\gamma} Z_i = \sum_{j=1}^{k} \beta_j X_{ij} + \sum_{j=1}^{c} \gamma_j Z_{ij} \qquad (2)$$

Model (1) is for normally distributed quantitative traits and model (2) is for dichotomous traits, where $\boldsymbol{\beta}$ is the vector of the genetic effects of $\mathbf{X}_i$ and $\boldsymbol{\gamma}$ is a vector of the effects of covariates $\mathbf{Z}_i$ on y_i. For dichotomous traits, let p_i denote the probability of being a case given the ith individual's genotypes and covariates, and

$$\text{logit}(p_i) = \ln(p_i)/(1-p_i)$$

In this setup, the intercept is included in the $\boldsymbol{\gamma}$ vector. The coefficients $\boldsymbol{\beta}$'s are the log odds ratios (ORs) of the rare alleles for case/control data and the additive effects for quantitative traits. Here we include in the model all variants, including the noncausal ones that are to be analyzed together. For noncausal variants, the $\boldsymbol{\beta}$'s are zero and pose no modeling difficulties. Under the null hypothesis that $\mathbf{X}_i$ is not associated with y_i, $\boldsymbol{\beta} = 0$, i.e., $\beta_1 = \cdots = \beta_k = 0$. For binary traits assuming a logistic model (2) the likelihood for the ith individual is:

$$L_i(\boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'X_i + \gamma' Z_i}}{1 + e^{\boldsymbol{\beta}'X_i + \gamma' Z_i}}$$

For quantitative traits assuming a linear model (1) the likelihood for the ith individual is:

$$L_i(\boldsymbol{\beta}) = \left(\sigma\sqrt{2\pi}\right)^{-1} e^{-\frac{(y_i-u)^2}{2\sigma^2}}$$

where $\boldsymbol{\mu}$ is the population mean of the trait. Combining all individual data, the likelihood is:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta})$$

To avoid redundant discussions for both dichotomous and quantitative traits, we will use dichotomous traits to describe association studies, unless otherwise specified. Most of these principles directly apply to quantitative traits as well. Based on the likelihood models, commonly used hypothesis testing approaches for genetic effects $\boldsymbol{\beta}$ are described below.

Likelihood ratio tests. Let L_1 be the maximum likelihood in the full model over the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and L_0 be the maximum likelihood in the null model over the parameter space $\boldsymbol{\gamma}$ while fixing $\beta_1 = \cdots = \beta_k = 0$. The Likelihood Ratio Test (LRT) statistic $\lambda = -2\ln(L_0/L_1)$ follows a χ^2_{k} distribution with k degrees of freedom (d.f.) asymptotically. An asymptotical p value can be obtained by comparing the LRT statistic with the χ^2_{k} distribution for large sample sizes. When $k > 1$, this tests for the overall effects of all SNPs as a whole but not individual SNP effects.

Wald tests. When maximizing likelihood over the parameters in the full model, the maximum likelihood estimates (MLEs) of β_j's and their corresponding standard errors can be obtained in standard statistical packages. Let $\hat{\beta}_i$ denote the MLE of β_i and $\text{SE}(\hat{\beta}_i)$ denote the standard error of $\hat{\beta}_i$. The Wald statistic $w = \beta_i/\text{SE}(\beta_i)$ follows a standard normal asymptotically, and an asymptotic p value can be calculated by comparing the statistic with the standard normal distribution. The Wald test can be carried out for any of the β_i's conditional on covariates and other SNPs.

Score tests. Let $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ be the combined vector of the parameters of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and $A_i = (X_i, Z_i)$ be the combined vector of X_i and Z_i. Then the score statistic for $\boldsymbol{\alpha}$ is:

$$U_{\boldsymbol{\alpha}} = \sum_{i=1}^{n} \frac{\partial \ln(L_i(\boldsymbol{\alpha}))}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n} (y_i - \tilde{y}_i) A_i$$

where $\tilde{y}_i$ is the expected phenotype for the ith individual. Under the null hypothesis its expectation is zero, and its covariance matrix is:

$$V_{\boldsymbol{\alpha}} = \sum_{i=1}^{n} \frac{\partial^2 \ln(L_i(\boldsymbol{\alpha}))}{\partial \boldsymbol{\alpha}^2} = \sum_{i=1}^{n} \tilde{y}_i (1 - \tilde{y}_i) A_i A_i'$$

Since we are only interested in testing the SNP effects $\boldsymbol{\beta}$, the covariate effects $\boldsymbol{\gamma}$ are considered as nuisance parameters. To eliminate $\boldsymbol{\gamma}$, let $\tilde{y}_i$ be the fitted phenotype values after regressing out the covariates, i.e., $\tilde{y}_i = \text{logit}^{-1}$ $\hat{\boldsymbol{\gamma}}$ Z_i). The score vector for $\boldsymbol{\beta}$ is:

$$U_{\beta} = \sum_{i=1}^{n} (y_i - \tilde{y}_i) X_i$$

Under the null, U_β follows a multivariate normal distribution asymptotically: $U_\beta \sim N_k(0, V_\beta)$, where V_β is the covariance matrix of U_β under the null. V_β can be obtained from V_α as $V_\beta = V_{\beta\beta} - V_{\beta\gamma} V_{\gamma\gamma}^{-1} V_{\gamma\beta}$, where $V_{\beta\beta}$, $V_{\beta\gamma}$, and $V_{\gamma\gamma}$ are corresponding submatrices of V_α. To test the null hypothesis that $\boldsymbol{\beta} = 0$, a score test statistic can be computed as $\text{S} = U_\beta V^{-1}{}_\beta U'{}_\beta$, which follows asymptotically a χ^2_{k} with k d.f. As for LRT, the score test is only testing the overall effect of all SNPs but not individual SNP effects.

STRATEGIC APPROACH

Traditional Approaches

In GWAS, the most commonly used analysis approach is single-marker tests, where a statistical test is carried out for each marker and a threshold of 5 × 10^{-8} is used to correct for multiple testing to control FWER at the genome level. In GWAS and candidate gene studies, in addition to single-marker tests, if the hypothesis is whether a gene harbors association signals, multimarker tests are often applied to jointly test the overall effect of the markers in a gene or region as a whole. When the CDRV hypothesis holds, although both can be applied to rare variants, the following sections demonstrate that both approaches are underpowered.

Single-marker tests

The simplest approach to genome-wide analysis is to analyze individual variants separately. Without any model assumptions, a 2 d.f. Pearson χ^2 test can be performed on a 2 by 3 contingency table to compare the frequencies of three genotypes of a variant in cases versus controls. For rare variants, the frequency of homozygous rare alleles may be very low and Pearson χ^2 tests may have inflated type I error. One remedy is to group the rare homozygotes and the heterozygotes together (i.e., assuming a dominant model). After the grouping, a 1 d.f. Pearson χ^2 test can be performed and is expected to achieve improved power over a 2 d.f. test. For complex traits, an additive genetic model is usually assumed for the three genotypes, where carrying an extra copy of the variant allele increases the genetic risk. A convenient way to code the genotype is 0, 1, 2 for genotypes carrying 0, 1, or 2 rare alleles, respectively. To test for association in regression models, only one variant can be included in (1) or (2), and the test for $\beta = 0$ can be carried out either through a Wald test, an LRT, or a score test (equivalent to the commonly used Cochran-Amitage test for trend), all with 1 d.f.

The 0, 1, 2 coding for the genotype is not the most powerful approach for all scenarios. If prior knowledge is available, other coding approaches can be used to reflect the genetic effect of each genotype. For example, a 0, 0, 1 coding is for the recessive model, where carrying one copy of a variant allele does not increase disease risk, and 0, 1, 1 is for the dominant model, where the increase in disease risk is the same for heterozygous and homozygous variant carriers. Other methods can be used as well to represent complex models. Although an additive model is unlikely to be strictly correct, the trend test (i.e., score test) is not to test the linearity, and as long as there is a trend—which is likely to hold for complex traits—the trend test is expected to be robust and achieve increased power due to the parsimony of the model. For rare variants uncovered through sequencing, it is likely that association tests will be carried out directly on causal variants, and therefore flexible genetic models can be used if prior knowledge is available for specific diseases or variants.

Multimarker tests

Oftentimes the interest is to test whether multiple variants in a gene, region, or any collection of variants as a whole are associated with the phenotype. This can be achieved using LRT or score tests discussed previously. These multimarker tests can only jointly test the effect of all markers as a whole, and if the null hypothesis is rejected, it is not known which variants are associated with the phenotypes. It is possible that all of them or only a subset might be associated with the phenotype. To pinpoint associated variants, a single-marker test may be needed to examine individual variant effects.

Limitations of traditional methods when applied to rare variants

The performance of various testing strategies is heavily influenced by the underlying genetic models, and both the power and type I error can be dramatically different under CDCV and CDRV hypotheses. Let's use a gene as an example. Two key features that are different in the two hypotheses influence power. First, under the CDCV hypothesis, it is most likely that there will be only one causal variant per gene. Extreme allelic heterogeneity in a gene is often the case when the CDRV hypothesis holds, however; that is, for the CDCV hypothesis only, one causal variant contributes to the association signal, while for the CDRV hypothesis, multiple rare variants independently influence the phenotype. The second difference is that common variants in a gene are often in strong LD, so that multiple common variants can be used to tag the underlying causal common variant, while rare variants are often weakly correlated. Given these differences, the power for single-marker tests of rare variants is low for three reasons. First, very few individuals in a sample carry rare alleles at single-variant sites and therefore the association signal is weak due to low frequencies. Second, in the presence of allelic heterogeneity, distinct causal variants in a gene are observed in affected samples and the association signals of individual variants are weakened by one another (Slager et al., 2000). Finally, rare variants are drastically more abundant than common ones (Keinan and Clark, 2012; Nelson et al., 2012; Tennessen et al., 2012) and are only weakly correlated, resulting in a severe penalty to correct for multiple testing. All of these reasons make the single-marker test an unfavorable approach.

For multimarker tests, allelic heterogeneity poses a less severe problem than for single-marker tests, since multiple causal variants jointly contribute to the association signal. As a result, multimarker tests can be more powerful than single-marker tests (Li and Leal, 2008). There is a large penalty in terms of degrees of freedom for multimarker tests, however, due to the excess of rare variants to be tested within a gene and because the power is degraded with increasing numbers of rare variants. Although ultimately single-marker tests can be used to pinpoint individual causal rare variants when the sample size is sufficiently large, currently this is not feasible due to the prohibitive cost of sequencing thousands of samples. A promising strategy for the modest sample size of current sequencing studies is to test for rare variant associations by aggregating multiple rare variants across a gene. This approach is described in detail in the following sections.

For rare variants, not only the statistical power but also the type I error rate is negatively affected. Due to the low frequency of rare variants, the sparsity of the data may make the asymptotic results inaccurate for modest sample sizes. For example, likelihood ratio tests for case/control data are often anticonservative due to the numerical instability of the likelihood maximization, and conversely Wald and scores tests are often conservative. In such situations, permutation is usually carried out to obtain empirical *p* values. Even after aggregating multiple rare variants (see the following sections), the cumulated allele frequency may not be sufficient for asymptotic results to hold and permutation is often required to calculate empirical *p* values.

Aggregation Association Analysis for Rare Variants

The goal is to test whether multiple rare variants in a gene as a whole are associated with the phenotype. This class of tests is often referred to as aggregation association tests. The major advantage of this approach is the achievement of dimension reduction through aggregating multiple rare variants into a single unit of analysis. Specifically, in the regression framework (1) or (2), where X_i represents the genotype of the *k* rare variants carried by the *i*th individual in a gene, the key is to reduce the dimension of $\boldsymbol{\beta}$ from *k* to 1 or a small number. A variety of aggregation methods have been proposed (see review papers by Asimit and Zeggini, 2010; Bansal et al., 2010; Dering et al., 2011; Stitziel et al., 2011; Ladouceur et al., 2012), and we present some of them in the follow categories.

1. Burden tests

The strategy of this category of aggregation association methods is to test whether there is an excess of rare variants in cases or controls. A general approach is to collapse or aggregate multiple rare variants in a gene into a single "super" variant and then to perform the association tests on this single super variant. Such an aggregation, if done properly, can achieve these benefits: when summed, low frequencies of multiple rare variants increase the overall frequency of the super variant; and the degrees of freedom are reduced from *k* to 1. This results in both an enrichment of signals and a reduction of dimensionality. Formally, this aggregation can be represented in this regression model:

$$\text{logit}(p_i) = \beta_\alpha \delta(X_{il}, \ldots, X_{ik}) + \sum_{j=1}^{c} \gamma_j Z_{ij}$$

where $\boldsymbol{\delta}(X_{i1}, \ldots X_{ik})$ is a function that summarizes multiple rare variants into a single number, which represents a single super variant, and β_a is the genetic effect of the super variant. By aggregating, the original null hypothesis of $\beta_1 = \cdots \beta_k = 0$ is equivalent to the null hypothesis $\beta_a = 0$. Now the association test of multiple rare variants becomes a single d.f. test, greatly reducing dimensionality. The central component of burden tests is the construction of the aggregation function $\boldsymbol{\delta}(X_{i1}, \cdots X_{ik})$. Although an appropriately constructed aggregation can increase power, the inclusion of noncausal variants can dramatically reduce the power. Several aggregation approaches and their performance in various scenarios are discussed below.

Indicator function. The simplest collapsing way is the use of an indicator function (Li and Leal, 2008),

$$\delta(X_{i1}, \ldots, X_{ik}) = \begin{cases} 1 & \sum_{j=1}^{k} X_{ij} > 0 \\ 0 & otherwise \end{cases}$$

This simple approach codes 1 for individuals that carry one or more rare alleles within the tested genetic region and zero if all variants are major alleles. For this method, the association test is transformed into testing whether the frequency of rare variant carriers in cases is different from that in controls. Single-marker tests using LRT, Wald, or score statistics can be carried out as previously described. In addition, a 2 by 2 table can be constructed with numbers of rare variant carriers and noncarriers, and a Fisher exact test can be performed. Although simple, this strategy has an intuitive interpretation in terms of OR of rare variant carriers. Even if other more involved methods are used, this simple counting can serve as an estimate of the overall genetic effect of rare variants.

Variant counting. The use of an indicator collapsing method indicates that most likely one individual carries only one rare variant. For larger genes or when multiple genetic regions are analyzed as a single unit, however, the probability that an individual carries more than one rare variant increases. If it is assumed that individuals with more than one rare variant have an increased risk of being affected, ignoring this information may reduce the power to detect an association. A simple extension is to count the number of rare variants each individual carries (Li and Leal, 2009), i.e.:

$$\delta(X_{il},\ldots,X_{ik}) = \sum_{j=1}^{k} X_{ij}$$

where X_{ij} is coded as the number of rare alleles the ith individual carries at the jth variant site. As in the single-marker approach, Wald, score, or LRT tests can be used. In this aggregation, no simple contingency tables can be tabulated because of the potential LD between rare variants. The estimated β_a can be interpreted as the log(OR) per rare variant in a gene on average.

Weighted sum statistic. It is likely that different rare variants have differential genetic effects. For example, causal variants with strong deleterious effects are under strong purifying selection and are therefore more likely to be rare (Gorlov et al., 2008; Keinan and Clark, 2012; Tennessen et al., 2012), and nonsynonymous variants are more likely than synonymous variants to affect the gene function. In either the indicator-collapsing or variant-counting approach, such information is ignored, and power loss is expected when rare variants to be collapsed have different effects. To take this into account, Madsen and Browning proposed a weighted sum statistic (WSS) to aggregate multiple rare variants, i.e.,

$$\delta(X_{il},\ldots,X_{ik}) = \sum\nolimits_{j=1}^{k} w_j X_{ij}$$

where w_j is the weight assigned to the jth rare variant. Specifically, a frequency-dependent weighting is used:

$$w_j = 1/\sqrt{p_j(1-p_j)}$$

where p_j is the allele frequency of the jth rare variant in controls and estimated as:

$$p_j = \frac{m_j^U + 1}{2n_j^U + 2}$$

in which m_j^U is the number of minor alleles of the jth variant in controls and n_j is the number of controls with nonmissing data. Here the numbers 1 and 2 are used to avoid the estimation of zero frequency, which can cause numerical instability. In WSS, rarer variants are up-weighted so that rare alleles contribute more to the test statistic. The genetic scores are then ranked and the WSS is calculated as the sum of the ranks of the cases. Since the frequency estimation depends on the phenotype, i.e., only unaffected individuals are used, permutation is performed to obtain empirical p values by permuting case/control status.

The WSS as originally proposed cannot account for covariate effects and permutation is needed to obtain empirical p values, which is computationally expensive. The same aggregation can be implemented in regression models and has been extended to general score tests (Lin and Tang, 2011). Such a setup can readily incorporate covariates and efficiently obtain estimates of genetic effects. To obtain asymptotic p values in regression models, frequency estimates should not be dependent upon phenotypes, since inflation of type I error is expected if frequencies are estimated based on controls only.

Weighting scheme. The assumption of up-weighting rarer variants is that rarer variants are more likely to have larger effects. The frequency-based weighting scheme proposed in WSS is arbitrary and may have reduced power when the weighting is far away from the true relationship. To see how weighting can affect the power and how the upper bound of power can be achieved by optimal weighting schemes, we can compare the weighted sum approach with the true model. The weighted sum model is as follows:

$$\begin{aligned}\text{logit}(p_i) &= \beta_a \sum_{j=1}^{k} w_j X_{ij} \\ &\quad + \sum_{j=1}^{c} \gamma_j Z_{ij} \\ &= \sum_{j=1}^{k} \beta_a w_j X_{ij} \\ &\quad + \sum_{j=1}^{c} \gamma_j Z_{ij}\end{aligned} \qquad (3)$$

Assuming the true model is (2), if we compare (3) to (2), we will see that when $w_j = \beta_j/\beta_a$, (3) is recovered to the true model (2). This indicates that when w_j for the *j*th variant is assigned proportional to its true genetic effect, the aggregation can achieve the optimal power (Lin and Tang, 2011). Conversely, improper weighting schemes that dramatically deviate from this relationship are detrimental to the power. Ideally, a zero weight should be assigned to nonassociated variants. Accidentally up-weighting instead of down-weighting noncausal variants amplifies noise and reduces power. Theoretically, no uniformly most powerful tests exist for this multidimensional problem (Cox and Hinkley, 1979), and for a fixed weighting scheme the power depends on the alternative hypothesis (i.e., the true genetic model). Empirical evaluation of rare variant analysis methods in various scenarios is consistent with the theory (Ladouceur et al., 2012). Although it is generally impossible to know a priori the true model, the ability to utilize prior knowledge to assign weights that are close to the true genetic model is the key to achieving increased power of aggregation analysis. It should be noted that it is not the absolute value of the weight but rather the ratio of the weights for rare variants that determines the power, since the model is unchanged if weights are multiplied by a nonzero constant and β_a is divided by the same constant. Although optimal weighting is not possible in reality, it is helpful to calculate the relative ratio of weights assigned to rare variants when designing weighting schemes. A weighting scheme that generates extremely large ratios is questionable and can lead to a great decrease in power.

Madsen and Browning used a frequency-related weighting scheme. This weighting scheme will achieve increased power when

$$w_j = 1/\sqrt{p_j(1-p_j)} \propto \log(OR_j)$$

for the *j*th variant. Although rarer variants are more likely to be functional due to strong purifying selection, it may be difficult to justify this relationship between ORs and frequencies. The variable threshold (VT) method (Price et al., 2010) proposed to explicitly incorporate functional prediction scores from PolyPhen-2 (Adzhubei et al., 2010; Ramensky et al., 2002) as weights for individual variants in the aggregation testing. Assuming that functional prediction scores reflect the genetic effect of individual variants on the trait under study, this approach has the potential to increase power compared with collapsing or equal weighting schemes. Some tools specifically generate predictive functional scores for nonsynonymous variants (Ferrer-Costa et al., 2005; Bromberg and Rost, 2007), while others are more general tools that can assess the potential of disease causing or evaluate the sequence conservation across species (Cooper et al., 2005; Siepel et al., 2005; Schwarz et al., 2010). Since these scores are from external sources and are not dependent upon phenotype and genotype data, asymptotic results hold for large sample sizes. One of the challenges is that functional prediction scores from different bioinformatics tools are often not consistent, and it is unclear how to integrate the inconsistent predictions into the analysis. Additionally, even if bioinformatics could predict with high accuracy that a variant is causal for one phenotype, it does not guarantee that it is causal for the trait under study.

2. Mixed-effects models

For case/control studies, burden tests look for an enrichment of rare variants in cases compared with controls for risk alleles, or an excess of protective alleles in controls compared with cases. Burden tests will achieve greatest power when all causal variants have the same direction of genetic effects. When a portion of causal variants has effects in opposite directions, i.e., protective and detrimental or increasing and decreasing quantitative trait values, aggregating variants will weaken the overall association signal, resulting in reduced power to detect an association. In an extreme scenario, when half of the variants decrease disease risk and the other half increase disease risk, the association signal can be completely cancelled out. Theoretically, this can be solved by assigning negative weights to protective variants. In reality, however, it is impossible to decide which set of variants has risk effects and which set is protective. New methods have been developed to deal with this situation.

The C-alpha test (Neale et al., 2011) is one of the early methods that was developed to tackle this problem for case/control data. It compares the observed variance of allele counts to the expected variance under the null hypothesis of no association for dichotomous traits. Let the number of observed rare alleles be n_j for the *j*th variant. When no variants are associated with the phenotype, the rare allele count at the *j*th variant sites in cases follows a binomial distribution (n_j, p_j), where $p_j = p_0$ for all $j = 1,\ldots, k$ and p_0 is the expected proportion of allele count in cases (e.g., p_0=0.5 when the numbers of cases and controls are equal). When causal variants have different effect sizes and directions, not all p_j's are equal to p_0 and the data are a mixture of binomial distributions. Since any mixture of binomial distributions creates overdispersion, testing the increased variance over its expected value under the null serves as the foundation for the C-alpha test.

To also address the problem when variants have effects in opposite directions, the sequence kernel association test (SKAT) (Wu et al., 2011) was developed in a more general framework. Additionally, it has been shown that C-alpha is a special case of SKAT (Wu et al., 2011). Compared with current implementations of the C-alpha test, SKAT has the flexibility to accommodate such additional features as including covariates (e.g., principal components for adjusting population stratification); accounting for LD among variants; allowing for the analysis of both qualitative and quantitative traits; weighting of variants based on frequencies or functional prediction scores; and handling complex genetic models (e.g., epistasis effects) (Wu et al., 2011). SKAT is a variance-component score test in a multiple regression model (1) or (2). SKAT assumes that each β_j follows an arbitrary distribution with a mean of zero and a variance of $w_j\tau$, where τ is a variance component and w_j is the weight for the *j*th variant. Under this setup, the original null hypothesis is equivalent to H_0: $\tau = 0$. A variance component score test in a mixed-effects model can be used to test this hypothesis. The SKAT score statistic is defined as $Q = (\mathbf{y} - \tilde{\mathbf{y}})' \, \mathbf{K}(\mathbf{y} - \tilde{\mathbf{y}})$, where $\mathbf{K} = \mathbf{XWX}$', $\tilde{\mathbf{y}}$ is the predicted mean of the phenotype under the null hypothesis, i.e., fitting (1) or (2) without β_j's, as described before, **W** is diag($w_1, \ldots ,w_k$) with w_j being the weight for the *j*th variant. Since it is a score test, it can be efficiently computed to obtain asymptotic *p* values (see Wu et al., 2011, for details).

An attractive feature of SKAT is that flexible genetic models and prior knowledge can be incorporated in the **K** matrix. **K** is an *n* by *n* matrix, with the (*i*, *i*')-th entry equal to $K(X_i, X_i')$, representing the genetic similarity of the *i*th and the *j*th individuals. *K*(.,.) is called a kernel function and different kernels can be constructed depending on hypotheses about genetic models of variants for specific studies and genes. The simplest kernel is the weighted linear kernel, i.e.,

$$K(X_i, X_i') = \sum\nolimits_{j=1}^{k} w_j X_{ij} X_{i'j}$$

As discussed for WSS, a good weighting scheme can increase power, while one that does not reflect the true underlying genetic model can reduce power. If equal weights are used and no covariates are included for case control data, SKAT is equivalent to C-alpha (Wu et al., 2011). In the original paper, the authors proposed to use a beta distribution to specify weights because it is flexible enough to accommodate a wide range of weights. Specifically,

$$\sqrt{w_j} = Beta(MAF_j, a_1, a_2)$$

where a_1 and a_2 are prespecified parameters for a beta distribution and MAF_j is the rare allele frequency estimated across both cases and controls. The authors suggested $a_1 = 1$ and $a_2 = 25$, which allows for increasing weights for rare variants and decreasing weights for common variants. Other values can also be used for different prior knowledge; for example, $a_1 = a_2 = 1$ corresponds to assigning equal weight for all variants, and $a_1 = a_2 = 0.5$ specifies:

$$\sqrt{w_j} = \sqrt{MAF_j(1 - MAF_j)}$$

which put strong weights on rare variants. There may not be a simple relationship between the allele frequency and the genetic effect, and other prior information can be incorporated to guide the weighting, such as the functional prediction scores discussed in the VT test. Other complex kernels can also be constructed to accommodate more complex models such as epistasis. Wu et al. (2011) provides details for interested users.

SKAT was proposed as a variance component score test in a mixed-effects model and has a connection with the score test in a fixed-effect regression model. For the linear kernel, i.e.,

$$K(X_i, X_i') = \sum_{j=1}^{k} w_j X_{ij} X_{i'j}$$

it can be shown that

$$Q = \sum_{j=1}^{k} w_j S_j^2$$

where

$$S_j = X_j'(\mathbf{y} - \tilde{\mathbf{y}})$$

is the individual score statistic of the jth variant (Wu et al., 2011). Similarly, the WSS score statistic is:

$$S = \sum_{i=1}^{n} (y_l - \tilde{y}_i) \sum_{j=1}^{k} w_j X_{ij}$$
$$= \sum_{j=1}^{k} w_j X_j'(\mathbf{y} - \tilde{\mathbf{y}}) = \sum_{j=1}^{k} w_j S_j$$

SKAT also has connections with other tests; see Pan (2009), for more discussion. For single-marker tests, i.e., $k = 1$, SKAT is equivalent to single-marker score tests. When $k > 1$, their behaviors become different, and the relative power depends on genetic models and how weights are assigned. For example, when two rare variants have opposite directions of effect, it is most likely that the score statistic of the risk allele is positive while the score statistic of the other variant is negative. The association signal is cancelled out when positive weights are assigned to both variants, resulting in reduced power, as shown in **Weighting scheme**. On the other hand, the squared score statistics in SKAT eliminate the direction issue and all variants contribute positive scores to the test statistic. Therefore, SKAT can gain more power compared with burden tests in the presence of opposite directions of genetic effects. However, there is a trade-off between power and robustness. When a large proportion of rare variants has the same direction of effects, SKAT will be less powerful than burden methods. In general, it is challenging for investigators to know when to apply SKAT or burden tests due to the complexity of the genetic basis of complex traits. It is plausible—especially for dichotomous traits—that genetic variants in the same gene are likely to affect the gene function in a similar fashion, and burden tests may be more powerful if the analysis unit is a gene.

3. Data-driven approaches

To carry out aggregation analyses and to increase statistical power, several criteria need to be determined. These criteria include, for example, the frequency cutoff of rare variants and the weight for each variant. Usually these criteria are either prespecified or come from external sources. The choices are usually arbitrary, however, and may not be proper for some traits or genes. An alternative is to let the data drive these choices—that is, to select appropriate criteria based on the phenotype and genotype data under study. Since this class of methods uses the same data for both feature selection and hypothesis testing, permutation procedures are usually needed to obtain empirical p values. We describe a few of these methods and discuss their performance.

Variable-threshold method. For a complex trait, it is likely that causal variants span a wide spectrum of allele frequencies and that the allelic architecture varies widely from trait to trait. Although variants with allele frequencies of less than 0.01 are commonly used in practice for aggregation analyses, there is no clear biological justification for this threshold. An improper threshold cutoff may dramatically reduce power by excluding causal variants and including noncausal variants. To avoid the arbitrary specification of frequency cutoffs, the variable-threshold method (Price et al., 2010) was proposed to automatically select the "optimal" frequency threshold for rare variants and include the variants with frequencies below this threshold for aggregation analyses. Specifically, for a given weighting scheme (e.g., functional prediction, inverse of allele frequency, or no weighting), a statistic S_T is calculated for each threshold T in the range between the lower bound T_L and the upper bound T_U, and the maximum of $S_{\max}$ and the corresponding threshold $T_{\max}$ are recorded. Since $S_{\max}$ depends on phenotype and genotype, the distribution of $S_{\max}$ under the null is generally unknown and permutations are required to assess the significance of $S_{\max}$. Specifically, the same procedure is carried out to obtain $\mathrm{S}_{\text{max\_perm}}$ for each permutated data set and $S_{\max}$ is compared with the distribution of the $S_{\text{max\_perm}}$ to obtain empirical p values. Although a different statistic was used in the original publication of the VT method, the standard score statistic in a regression model can be used to carry out the same VT testing procedure because of its desirable statistical properties (Lin and Tang, 2011).

The VT method essentially performs many tests to find the optimal cutoff and uses permutation to correct the multiple testing to obtain empirical p values. When $T_{\max}$ is far from a user-specified cutoff, VT is expected to achieve increased power by finding the proper threshold, and if $T_{\max}$ is close to the prespecified threshold, VT suffers loss of power because of correcting for multiple testing. To reduce the search space, it may be desirable to set T_L and T_U accordingly to confine the search in a smaller range. For example, it may not be necessary or desirable to include in the aggregation association test variants with frequencies >0.1. VT also explicitly incorporates functional prediction scores in weighting variants, e.g., PolyPhen-2 scores (Ramensky et al., 2002; Adzhubei et al., 2010), and proper weighting can reciprocally increase the effectiveness of selecting the optimal threshold.

Adaptive weighting methods. Although allele frequency and function prediction scores are used for weighting, they may not reflect true genetic models. Adaptive weighting methods have been proposed to utilize phenotype and genotype data to guide the weighting of variants. The estimated regression coefficients (EREC) method (Lin and Tang, 2011) first estimates the regression coefficients in the general model (1) and (2) and then incorporates these estimates of individual coefficients in the weighting scheme. As described before, the optimal weights are the true β_j's, which are unknown. It is tempting to use the maximum-likelihood estimates $\hat{\beta}_j$'s to guide the assignment of the weights of individual variants. In the EREC approach, it was proposed to use $w_j = \hat{\beta}_j + \boldsymbol{\delta}$ as the weight for the *j*th variant, where $\boldsymbol{\delta}$ is a constant. The value of $\boldsymbol{\delta}$ is arbitrary and the EREC method recommends that it be set to 1 for dichotomous and 2 for quantitative traits when sample size is <2000. The behavior of the EREC method depends on the choice of the constant $\boldsymbol{\delta}$. When $\boldsymbol{\delta} = 0$, it is equivalent to assigning $\hat{\beta}_j$ as the weight of the *j*th variant. Since $\hat{\beta}_j$'s are maximum-likelihood estimates, plugging $\hat{\beta}_j$'s as the weights in model (3) results in the maximum likelihood of the original model (2). Therefore, testing based on such a weighting scheme is asymptotically equivalent to the multimarker LRT or score test with *k* d.f. At the other extreme, when $\delta >> \hat{\beta}_j$'s, all weights are close to a constant and this is equivalent to the variant-counting approach. Therefore, EREC can be viewed as a method that falls between multimarker tests and the variant-count approach. Further modification can be made such that $\boldsymbol{\delta}$ is no longer a constant, but variable for different variants when desired, reflecting the prior belief of the genetic effects. The EREC method is expected to achieve both robustness due to its feature of multimarker tests as well as increased power, owing to its collapsing functionality. It is not clear, however, how the selection of $\boldsymbol{\delta}$ values affects the statistical power for various scenarios.

Other data-driven methods use similar approaches for dynamic weight assignment. For example, the kernel-based adaptive cluster method (Liu and Leal, 2010a) and the data adaptive sum test (Han and Pan, 2010) assign weights adaptively to variants based upon variant counts from the data. The RareCover (Bhatia et al., 2010) method uses a variable-selection approach to select the optimal set of variants that maximize the burden test statistics (i.e., finding optimal assignment of zero weights to a subset of variants). A general class of adaptive methods has also been proposed (Pan and Shen, 2011). Statistical significance for these tests is usually evaluated by permutation. Interested readers can refer to the original papers of the authors cited in this paragraph for details.

4. Hybrid methods

Both burden tests and nonburden methods (e.g., multimarker tests and SKAT) have increased power for certain genetic models and are underpowered in other situations, and all have reduced power when noncausal variants are included in the analysis. Hybrid methods have been proposed to combine methods that are powerful for variants with the same effects and methods that are robust when either noncausal variants or variants with opposite effects are present. We will describe in this section the combined multivariate and collapsing (CMC) method (Li and Leal, 2008) and the SKAT-O approach (Lee et al., 2012a; Lee et al., 2012b).

CMC. The CMC method combines the burden tests and multivariate tests explicitly to achieve both increased power and robustness. It intuitively collapses subsets of *k* rare variants and then jointly tests the collapsed subsets in a multivariate test. Based on the regression framework (2), an example of the CMC method is as follows:

$$\begin{aligned}\operatorname{logit}(p_i) = {} & \beta_1^{CMC}\delta_1(X_{ij}, j \in \Delta_1) \\ & + \beta_2^{CMC}\delta_2(X_{ij}, j \in \Delta_2) \\ & + \sum_{j \in k-\Delta_1-\Delta_2} \beta_j X_{ij} \\ & + \sum_{j=1}^{c} \gamma_j Z_{ij}\end{aligned}$$

In the above modeling, Δ_1 and Δ_2 are two sets of rare variants that are to be aggregated, and $\beta_1{}^{\mathrm{CMC}}$ and $\beta_2{}^{\mathrm{CMC}}$ are the genetic effects of the collapsed super variants in the two subsets. The CMC method proposes to use the indicator function for collapsing and can be implemented using other collapsing approaches such as weighted sum score statistic. The null hypothesis in the CMC method becomes:

$$H_0 : \beta_1^{CMC} = \beta_2^{CMC} = \boldsymbol{\beta}_{j \in k - \Delta_1 - \Delta_2} = 0$$

A multiple d.f. LRT or score test can be carried out to jointly test the hypothesis. Here the dimension reduction is achieved for the rare variants in Δ_1 and Δ_2 for increased power while the multivariate tests of the subsets achieve robustness. It has been shown that common noncausal variants have greater detrimental effects on power of burden tests than that of multivariate tests (Li and Leal, 2008), and if different frequencies are to be analyzed in a gene, collapsing only rare variants using CMC is expected to be robust. Other criteria can also be used to collapse subsets of rare variants, e.g., rare functional variants affecting splicing and stop codons may be collapsed in one subset, while less dramatic changes like missense variants are collapsed in another. The CMC method is a flexible framework in that both the complete collapsing and multimarker tests without collapsing are special cases of CMC at two extremes. Its flexibility often requires appropriate user-defined subsets, however, which may not be obvious in reality.

SKAT-O. SKAT was developed to tackle the problem that both risk and protective alleles are present in a gene but prior knowledge is rarely available about the directionality of causal variants. Lee et al. developed the SKAT-Optimal test, which combined a burden test and SKAT in a single framework (Lee et al., 2012a,b). Recall that in SKAT each β_j is assumed to follow an arbitrary distribution with a mean of zero and a variance of $w_j\tau$ (see Mixed-effects models). All $\boldsymbol{\beta}_j$'s are assumed to be independent in SKAT. The new class of tests is formulated as a generalized family of SKAT through a family of kernels that incorporate a correlation structure among variant effects. Specifically, Lee et al. used an exchangeable correlation structure and the correlation matrix of β_j's is $\mathbf{R}_\rho = (1 - \boldsymbol{\rho})\mathbf{I} + \boldsymbol{\rho}\mathbf{11}'$, where $\mathbf{I}$ is a k by k identical matrix with 1 on the diagonal and zero otherwise, and $\mathbf{11}'$ is a k by k matrix with all entries being 1. This matrix is a compound symmetry correlation structure with 1 on the diagonal and $\boldsymbol{\rho}$ for all off-diagonal entries. The statistic for dichotomous traits in model (2) is:

$$Q_\rho = (\mathbf{y} - \tilde{\mathbf{y}})' K_\rho (\mathbf{y} - \tilde{\mathbf{y}})$$

This is similar to the original SKAT statistic with the exception that the kernel is replaced by:

$$K_\rho = \mathbf{XWR}_\rho \mathbf{WX}'$$

By separating the $\mathbf{R}_\rho$ matrix into the sum of two parts, it can be seen that $\mathbf{Q}_\rho$ is a linear combination of a burden test and the SKAT, i.e.,

$$Q_\rho = (1 - \rho)Q_{SKAT} + \rho Q_{burden}$$

The statistic is calculated as:

$$Q_{optimal} = \min_{0<\rho<1} p_\rho$$

where p_ρ is the p value calculated for a specific $\boldsymbol{\rho}$. SKAT-O uses a grid search approach to find the best $\boldsymbol{\rho}$ value that minimizes p_ρ: set a grid $0 < \boldsymbol{\rho}_1 < \ldots < \boldsymbol{\rho}_n < 1$, calculate $p_{\rho_l}, \ldots, p_{\rho_n}$ and obtain:

$$Q_{optimal} = \min\{p_{\rho_l}, \ldots, p_{\rho_n}\}$$

For large sample sizes, the p value of Q_{optimal} is derived analytically to evaluate the significance. Simulation studies suggest that SKAT-O outperformed SKAT and burden tests in a wide range of scenarios (Lee et al., 2012a, b). The correlation $\boldsymbol{\rho}$ determines the relative contribution of either test to the SKAT-O statistic. When $\boldsymbol{\rho} = 0$, it reduces to a burden test, when $\boldsymbol{\rho} = 1$, it is equivalent to SKAT, and when $0 < \boldsymbol{\rho} < 1$ it achieves the unification of these two kinds of tests. Since $\boldsymbol{\rho}$ is estimated from data, SKAT-O is also a data-driven approach. Using a similar argument, EREC can also be viewed as a data-driven hybrid method. Like other data-driven approaches, SKAT-O also involves multiple testing (i.e., searching for the optimal $\boldsymbol{\rho}$). Due to the correction of multiple testing, SKAT-O will be less powerful than both SKAT and burden tests when the true $\boldsymbol{\rho}$ is close to zero or one.

Replication

To rule out the possibility of spurious associations due to confounding factors, e.g., population stratification and sequencing batch effects, it is critical to replicate findings in independent samples. As discussed above, there are no uniformly most powerful tests for this type of analysis strategy. To potentially reduce false negatives, a viable approach is to carry out different tests in the initial study. If applying multiple tests is not corrected for in the initial study, there will be inflation of type I error rates. Therefore, replication studies are extremely important for the confirmation of association findings.

Two replication approaches can be employed. A simple strategy is to genotype the variants discovered in the initial study in an independent panel and perform appropriate rare variant aggregation tests on these genotyped variants (Liu and Leal, 2010b). This variant-based approach is cost effective and can quickly generate genotype data on a large replication sample. However, if causal rare variants have not been uncovered in the initial study, these causal variants will be missed in the replication panel and a loss of power is to be expected. An alternative strategy is to sequence the genes or regions of interest in a replication sample and to carry out association tests on the variants uncovered in the new sample (Liu and Leal, 2010b). This sequence-based approach is likely to uncover additional causal rare variants, but is more expensive and time consuming. Which replication strategy to choose depends on specific study goals and designs. For example, if the intent is to identify a more complete allelic spectrum for clinical applications, the sequence-based strategy serves as a better approach than the variant-based approach. Through extensive simulations, it has been demonstrated that the sequence-based approach is usually more powerful than variant-based replication; however, for most situations the difference is not dramatic (Liu and Leal, 2010b). Given the low cost of genotyping compared with sequencing, a variant-based strategy is attractive for large-scale replication studies of complex traits.

Estimates of Genetic Effects

After a genetic association is identified, it is desirable to estimate genetic effects for the identified association and quantify the explained genetic variance, in addition to the p-values that are usually reported. In rare variant burden tests, since multiple variants are aggregated as a single super variant, the slope parameter β_a in the regression model measures the change of mean trait value per unit of change in the burden score. For example, it is shown that for quantitative traits β_a is the weighted average of the individual β's (Pan, 2009). For the same data set, different aggregation strategies (indicator function, differential weights in weighted sum approaches) will lead to different estimates of the genetic effects. Therefore, they do not have a straightforward interpretation. On the other hand, the phenotypic variance explained by the burden score has a natural interpretation. In fact, it's shown that the locus genetic variance explained by the burden score will always be lower than the true genetic variance, unless optimal weights are assigned (Liu and Leal, 2012a). This suggests that some proportion of the heritability may still be missing and not explained by the burden analysis, even if an association is established between the gene and the phenotype. It is still necessary to pinpoint causal variants to more precisely estimate the contribution of causal rare variants to complex trait etiology.

Sequencing analysis pipeline for rare variant association

Previous sections presented various analysis methods for rare variant association studies. For most practical applications, sequencing data will be used to identify genes with associated rare variants. In this section we describe a practical pipeline for analyzing sequencing data to identify rare variant associations, focusing primarily on exome sequencing.

1. Variant calling

To identify rare variant association, it is critical to accurately call variants from NGS data. NGS reads are usually short, with moderate error rates. For example, typical reads from NGS platforms (e.g., Illumina) are around 100 bps with error rates ~0.5% to 1% per base. It is important to recognize that these imperfect sequences may lead to incorrect variant calls. The first step is to align short reads to the human reference genome. A variety of software is available (Nielsen et al., 2011), and BWA is a widely used tool (Li and Durbin, 2010). After the initial alignment, reads around short insertions or deletions need to be realigned, duplicated reads need to be removed, and raw base quality scores need to be recalibrated (DePristo et al., 2011; Nielsen et al., 2011). These preprocessing steps are meant to generate the accurate alignment of bases with calibrated quality scores. After these steps are taken, it is helpful to generate summary statistics about the alignment, such as the fraction of reads mapped to the target regions, distribution of depths on the target regions, base and mapping quality scores, etc. Contamination may also be checked based on the alignment (Jun et al., 2012). Outlier samples may be identified and removed from downstream analyses. After the sample cleanup, the next step is to identify variant sites and individual genotypes from the aligned bases across study samples. The standard variant calling tools—e.g., GATK (DePristo et al., 2011) and SAMtools (Li et al., 2009)—are likelihood-based and generate quality scores for variant calling. The current calling algorithms recommend joint calling of multiple samples together to increase the accuracy for common variants and decrease the false positive rare variant calls (DePristo et al., 2011; Li, 2011). If the samples are related, PolyMutt (Li et al., 2012a) or TrioCaller (Chen et al., 2013) can be used to perform family-aware variant calling. After this step, an initial Variant Call Format (VCF) (Danecek et al., 2011) file is generated to store all variant sites and individual genotypes with quality scores to indicate the confidence of the calling.

2. Annotation

This step aims to annotate all identified variant sites with functional features. Several software packages (Liu et al., 2011; Wang et al., 2010) are available. In this section, we will describe ANNOVAR (Wang et al., 2010), a tool that can integrate multiple databases and annotate variants with a variety of functional information. For gene-centric features, it can annotate variants as synonymous, nonsysnonymous, stop gain/loss, splicing, 5'UTR and 3'UTR, and intronic. It generates the corresponding positions and amino acid changes for coding variants. These annotations are based on transcripts and some variants may have multiple annotations for different transcripts in the same gene or transcripts from overlapping genes. For non-synonymous variants, it also provides functionality prediction scores from a variety of prediction algorithms, including PolyPhen-2 (Ramensky et al., 2002; Adzhubei et al., 2010), SIFT (Ng and Henikoff, 2003), LRT (Chun and Fay, 2009), and MutationTaster (Schwarz et al., 2010). For all variants, ANNOVAR outputs sequence conservation scores such as GERP++ (Cooper et al., 2005) and PhastCon (Siepel et al., 2005). Other information includes dbSNP IDs, allele frequencies in the 1000 Genomes Project (1000 Genomes Project Consortium, 2010; Abecasis et al., 2012), and the NHLBI-Exome Sequencing Project from the Exome Variant Server (EVS) (Emond et al., 2012; Tennessen et al., 2012). All these categories of information are useful for selecting promising variants for the analysis and construction of sensible weighting schemes in rare variant aggregation analysis.

3. Quality assessment of variant calling

It is common practice to filter out false positive variant calls using machine learning approaches (Abecasis et al., 2012; 1000 Genomes Project Consortium, 2010; DePristo et al., 2011) by using features that are predictive of misalignment of reads (e.g., mapping quality, sequence repeats, mappability). After filtering "bad" calls, it is helpful to check the Ti/Tv ratio, i.e., the ratio of numbers of transitions (A<->G and C<->T) versus transversions (all other nucleotide changes) from the reference alleles. There are more possible Tv's than Ti's, and if false variant calling is random regardless of Ti or Tv changes, we expect that the Ti/Tv ratio will be ~0.5. Since transitions occur more easily than transversions, a Ti/Tv ratio higher than 0.5 is expected. On the genome level, the observed Ti/Tv ratio is ~2.2 to 2.3, and for coding variants the Ti/Tv ratio is slightly over 3 (Abecasis et al., 2012; 1000 Genomes Project Consortium, 2010; Tennessen et al., 2012). A significantly reduced Ti/Tv ratio indicates an excess of false positive variant calls. The Ti/Tv ratio analysis is particularly important for checking the quality of novel variant sites that are not present in public databases (Ng et al., 2009).

4. Aggregation analysis

After generating a clean set of genotype calls with functional features, the key is to perform association analyses to identify associated rare variants. Although the focus is on aggregation analysis, it is always desirable to perform single-marker tests to identify relatively common variants with larger genetic effects. As we show in the discussion of various aggregation analyses, there is no single method that is superior to other methods in all scenarios. The performance is largely dependent on the true underlying genetic models, which are unknown. Here we provide some practical strategies that we hope are useful in genetic association studies. This is still an active research area and practitioners are encouraged to apply appropriate approaches and adapt new advances for their studies.

For gene-based analyses, the first step is to determine which variants to include in the aggregation analysis. If only rare variants are to be included, a set of prespecified allele frequency cutoffs may be used, e.g., MAF 0.05 to MAF 0.01, or the VT method may be used. To select a cutoff, it is worth noting that approaches based on the estimates from controls need to use permutation to calculate *p* values to avoid inflation of type I error rates. This is because under the null hypothesis that rare variants are not associated with the trait, under this selection criterion the expected frequency in controls is less than that in cases, which is the allele frequency in the population. The next step is to determine which functional variants are included. In practice, the first priorities may be the analysis of stop gain/loss, splicing and missense variants, and performing gene-based analyses on this category of variants. For dichotomous traits, it is natural to apply burden tests to test for an enrichment of these "functional" variants in cases or controls, with possible weighting of each variant based on such prior knowledge as functional prediction scores. Before applying weights to variants, it is desirable to check the ratio of the weights so that the range is compatible with complex traits. For example, it is hardly believable that the OR of one variant is hundreds of times higher than that of another one, and if noncausal variants are accidently more highly weighted than others, the association signal will be dramatically diluted by noise. For quantitative traits, either in random or extreme sampling designs in which SKAT-O can be applied, although rare variants that influence the quantitative trait values in different directions may be expected, it is biologically plausible for a large proportion of causal variants to have effects in the same direction, and burden tests should also be considered. For complex diseases with a strong indication of specific genetic models, specific aggregation tests may be appropriately constructed. For example, if a recessive model is suggested, a compound heterozygote modeling (e.g., the collapsing of heterozygotes where rare alleles are on different haplotypes) can be used to test for highly conserved functional variants, e.g., splicing sites and stop gain/loss variants.

Aggregation tests can also be applied to pathways or gene sets. It may not be desirable, however, to aggregate all rare variants, since the total number of rare variants may be too excessive to detect the association signal contributed by a smaller number of causal variants. The investigation of power in various tests in realistic simulations is lacking in this setup. It is expected that a few genes in a pathway or gene set may harbor causal variants, and the CMC method or SKAT-O may be used to guard against the noise in the noncausal genes. Prior knowledge from other resources (e.g., gene expression data) or data-driven approaches are a viable way to subselect promising genes (e.g., genes expressed in appropriate tissues) in order to reduce the dimension prior to aggregation analyses.

After obtaining exome-wide *p* values, we recommend drawing a QQ plot to check the behavior of all test methods. Systematic deviations from what is expectated indicate issues. If inflation of type I error is observed, it can be caused by the use of an anticonservative test, e.g., LRT for extremely rare variants (Li and Leal, 2008). Conversely, a deflation can be observed when conservative tests (Fisher exact tests or score tests when the sample size is not large) are used. These can sometimes be circumvented by obtaining empirical *p*-values via permutation. Confounding factors, such as population stratification or sequencing batch effects, can generate false association signals, and permutation will not resolve these issues. In such situations it is important to explore the data to identify and correct the confounding factors so as to avoid spurious associations (see Commentary at the end of this unit for more details).

5. Follow-up studies

It is important to carry out follow-up studies to confirm any significant findings in the initial sequencing study, and this is particularly true for rare variant associations. It is of particular interest to assay the function of candidate genes on phenotypes, but this strategy is time- and labor-consuming. The most economical approach is to replicate candidate genes in an independent sample. Given the effectiveness of the variant-based approach (Liu and Leal, 2010b), this strategy is more practical to assay thousands of samples. Custom chips can be designed to target the top candidate genes. Of particular note is the exome-chip design, which includes on the array ~240,000 nonsynonymous and splice site variants identified by sequencing >12,000 individuals (Do et al., 2012; *http://genome.sph.umich.edu/wiki/Exome_Chip_Design*). Although this exome chip can be used for replication, currently it is also used for primary association studies of coding variants. The exome chip is being genotyped on >1,000,000 individuals with phenotype data for a wide variety of traits (Do et al., 2012). For replication studies, ideally the same analysis strategy that is used in the discovery panel to select top candidate genes would be applied to the replication data, although it is also desirable to explore other genes for additional signals in the exome-chip data.

Software packages for rare variant analyses

Most original papers describing analysis methods provide software for carrying out the proposed methods. General tools that implement a battery of published methods are also available. Some of the methods can be easily implemented in statistical packages such as R software (R Development Core Team, 2008). Due to the complexity of the rare variant analyses, currently available tools may not fulfill specific analysis needs for specific studies. In such situations it is desirable to implement custom-designed approaches in R, to, for example, carry out specific analyses. The following list names a few software packages that implement most of the methods discussed in this unit:

- PLINK/SEQ, a package implementing a variety of methods (*http://atgu.mgh.harvard.edu/plinkseq/*)
- EPACTS, a package implementing a variety of methods (*http://genome.sph.umich.edu/wiki/EPACTS*)
- SKAT-related packages (Lee et al., 2012a, 2012b; Wu et al., 2011: *http://www.hsph.harvard.edu/research/skat/*)
- SCORE-Seq (Lin and Tang, 2011: *http://www.bios.unc.edu/∼dlin/software/SCORE-Seq/*)
- SimRare, a tool for the simulation and evaluation of various methods (Li et al., 2012b: *http://code.google.com/p/simrare/*)
- Variant Association Tools (San Lucas et al., 2012: *http://varianttools.sourceforge.net/Association/HomePage*)

COMMENTARY

Study designs

In this unit we focus on unrelated case/control or quantitative designs. Other study designs provide attractive alternatives. For example, family studies were largely ignored in the GWAS because of the low power for common variants. With the advent of rare variant searches, however, there is a resurgence of family studies, and the question of whether family or unrelated designs are more powerful for identifying rare variants with larger genetic effects is still being debated. It is argued that collecting families with multiple affected individuals can enrich causal rare variants, and sequencing such families is expected to achieve improved power over unrelated designs (Cirulli and Goldstein, 2010; Peng et al., 2010). A particular advantage of family studies is the ease of replication of rare variant findings. For example, a much reduced sample size is needed to ascertain additional family members of those individuals who carry candidate rare variants for replication. On the other hand, it requires a large sample to observe enough copies of rare variants in unrelated individuals. Family samples, however, are much more difficult to collect. It is likely that in the future, both designs will be carried out and that they will prove to complement each other. Although we only discussed methods that are devised for unrelated designs, some can be extended to family studies. A few other methods are available as well (Fang et al., 2012; Zhu and Xiong, 2012).

For quantitative traits, sampling individuals with extremely low or high phenotypes is more powerful than random sampling (Huang and Lin, 2007; Barnett et al., 2013; Liu and Leal, 2012b). This extreme sampling strategy has been successful in identifying rare variants in sequencing studies (Cohen et al., 2004, 2005, 2006; Romeo et al., 2007). One simple analysis approach is to treat the two extremes as cases and controls, for which all methods discussed in this unit are readily applicable. This simple strategy ignores the information carried in individual phenotypes, however. For traits that are normally distributed and sampling that is based on phenotype value cutoffs only, the extreme phenotypes follow truncated normal distributions. In such cases, likelihood models that are conditional on extreme sampling are expected to achieve improved power (Barnett et al., 2013; Huang and Lin, 2007; Liu and Leal, 2012b). In reality, however, investigators should be cautious about applying such methods if the traits do not follow a normal distribution or if additional criteria are used to select extremes. In these situations the extreme phenotypes may not follow truncated normal distributions, and it is recommended that case/control methods be used to gain more robustness.

Confounding factors for rare variant associations

In addition to the statistical challenges of rare variant association analyses, to avoid spurious associations, two confounding factors—namely, batch effects of sequencing and population stratification—are worth further discussion. These confounding effects have not been investigated extensively but their impact on association results can be substantial. Advanced methods are needed for studies in which such confounding effects are present.

Batch effects of sequencing refer to the differential variant and genotype calls in cases versus controls that are caused by any possible confounding factors, such as DNA sources, sequencing technologies, sequencing depth, calling approaches, and postprocessing. For example, sequencing depth has been shown to be such a confounding factor in the 1000 Genomes Project data (Abecasis et al., 2012; 1000 Genomes Project Consortium, 2010). It is not uncommon for different sequencing strategies to be used in the same study. For example, newer technologies with reduced error rates and increased coverage are used for later phases of sequencing, resulting in better genotype calls; different capturing kits have a more dramatic confounding effect due to varying capturing evenness and target regions. On their own, these factors can introduce systematic differences in cases and controls and in combination may lead to strong batch effects that generate spurious associations. It is generally true that genotypes of rare variants are more difficult to infer from sequencing than common variants. For a few methods in which the rarer a variant is, the more heavily it is up-weighted, it is unclear how the potential false rare variant calls may affect the power and false positive associations, and this warrants additional investigation. In the benign case where the heterogeneity of the sequencing strategies does not lead to confounding in well-designed studies, power loss is expected if these differences are ignored. It is helpful to explore batch effects—using, for example, principal component analysis (PCA)—and to take appropriate steps once confounding factors are identified. Unless sequencing technologies maintain high accuracy as they mature, differential sequencing platforms (e.g., reagents, software pipelines, etc.) are likely to be used in the same study and advanced methods that take into account such sequencing effects may be needed to increase the analysis power while controlling for batch effects.

The problem of population stratification in association studies is well recognized, and effective methods based on PCA (Price et al., 2006) and variance component models (Kang et al., 2010) are routinely applied to GWAS data. For rare variants, however, it is less clear how population stratification may affect association analyses. Several recent large-scale sequencing studies reveal that a vast majority of variants are rare, and the excess of rare variants is due to recent explosive human population growth (Keinan and Clark, 2012; Nelson et al., 2012). The departure of population growth from equilibrium skews the patterns of genetic variation and makes the modeling of population genetics more challenging. Excessive mutations introduced after the split of modern populations obscure the association studies; for example, study samples regarded as homogenous for GWAS may show differential patterns in the spectrum of rare variants. Simulation studies showed that the impact of rare variant population stratification on association mapping can be stronger than that of common variants (Mathieson and McVean, 2012). Commonly used approaches for correcting population stratification in GWAS—including methods based on PCA and variant-component models—may not always be effective in correcting the population stratification of rare variants (Mathieson and McVean, 2012). Although PCA has been shown to be effective on some data (Zhang et al., 2013), definitive conclusions require more studies, and the development of rare variant analysis methods can clearly benefit from an understanding of genetic variations that are caused by the recent explosion in human population growth.

Concluding remarks

In this unit we have described various rare variant analysis methods, their statistical features and scope of application, and discussed challenges of rare variant analysis from several perspectives. We also outline a pipeline for sequencing analysis to identify rare variant associations. It is clear that no consensus can be reached on standard approaches for aggregation analyses of rare variants, however. It is up to investigators to select appropriate analysis strategies that are tailored to their studies. Forthcoming results from ongoing studies will further our understanding of the architecture of complex traits, which will in turn help develop better analysis strategies. We hope that this unit serves as a general platform for introducing this emerging field and provides useful guidelines for rare variant association analysis.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Wei Pan, Division of Biostatistics, School of Public Health, University of Minnesota, for helpful discussions. This research is supported by an internal development fund from Vanderbilt University (B.L) and National Institutes of Health grants MD005964 and HL102926 (S.M.L.).

LITERATURE CITED

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061-1073.

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56-65.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. 2010. A method and server for predicting damaging missense mutations. *Nat. Methods* 7:248-249.

Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hébert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S., Yosef, N., Ruppin, E., Sharan, R., Vaisse, C., Sunyaev, S., Dent, R., Cohen, J., McPherson, R., and Pennacchio, L.A. 2007. Medical sequencing at the extremes of human body mass. *Am. J. Hum. Genet.* 80:779-791.

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P.E., Altshuler, D.M., Gibbs, R.A., de Bakker, P.I.W., Deloukas, P., Gabriel, S.B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L.R., Ren, Y., Wheeler, D., Gibbs, R.A., Muzny, D.M., Barnes, C., Darvishi, K., Hurles, M., Korn, J.M., Kristiansson, K., Lee, C., McCarrol, S.A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S.B., Pollack, S., Price, A.L., Soranzo, N., Bonnen, P.E., Gibbs, R.A., Gonzaga-Jauregui, C., Keinan, A., Price, A.L., Yu, F., Anttila, V., Brodeur, W., Daly, M.J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S.F., Zhang, Q., Ghori, M.J.R., McGinnis, R., McLaren, W., Pollack, S., Price, A.L., Schaffner, S.F., Takeuchi, F., Grossman, S.R., Shlyakhter, I., Hostetter, E.B., Sabeti, P.C., Adebamowo, C.A., Foster, M.W., Gordon, D.R., Licinio, J., Manca, M.C., Marshall, P.A., Matsuda, I., Ngare, D., Wang, V.O., Reddy, D., Rotimi, C.N., Royal, C.D., Sharp, R.R., Zeng, C., Brooks, L.D., and McEwen, J.E. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52-58.

Asimit, J. and Zeggini, E. 2010. Rare variant association analysis methods for complex traits. *Ann. Rev. Genet.* 44:293-308.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N.J. 2010. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* 11:773-785.

Barnett, I.J., Lee, S., and Lin, X. 2013. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genet. Epidemiol.* 37:142-151.

Bhatia, G., Bansal, V., Harismendy, O., Schork, N.J., Topol, E.J., Frazer, K., and Bafna, V. 2010. A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comp. Biol.* 6:e1000954.

Bodmer, W. and Bonilla, C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* 40:695-701.

Bromberg, Y. and Rost, B. 2007. SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 35:3823-3835.

Chen, W., Li, B., Zeng, Z., Sanna, S., Sidore, C., Busonero, F., Kang, H.M., Li, Y., and Abecasis, G.R. 2013. Genotype calling and haplotyping in parent-offspring trios. *Genome Res.* 23:142-151.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., and Lifton, R.P. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 106:19096-19101.

Chun, S. and Fay, J.C. 2009. Identification of deleterious mutations within three human genomes. *Genome Res.* 19:1553-1561.

Cirulli, E.T. and Goldstein, D.B. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* 11:415-425.

Cohen, J., Pertsemlidis, A., Kotowski, I.K., Graham, R., Garcia, C.K., and Hobbs, H.H. 2005. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. *Nat. Genet.* 37:161-165.

Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305:869-872.

Cohen, J.C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G.L., Grundy, S.M., and Hobbs, H.H. 2006. Multiple rare variants in *NPC1L1* associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl. Acad. Sci. U.S.A.* 103:1810-1815.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 15:901-913.

Cox, D.R. and Hinkley, D.V. 1979. Theoretical Statistics, Chapman and Hall, London, England.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., and Durbin, R. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., and Daly, M.J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43:491-498.

Dering, C., Hemmelmann, C., Pugh, E., and Ziegler, A. 2011. Statistical analysis of rare sequence variants: An overview of collapsing methods. *Genet. Epidemiol.* 35:S12-S17.

Do, R., Kathiresan, S., and Abecasis, G.R. 2012. Exome sequencing and complex disease:

Practical aspects of rare variant association studies. *Hum. Mol. Genet.* 21:R1-R9.

Dudbridge, F. and Gusnanto, A. 2008. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* 32:227-234.

Emond, M.J., Louie, T., Emerson, J., Zhao, W., Mathias, R.A., Knowles, M.R., Wright, F.A., Rieder, M.J., Tabor, H.K., Nickerson, D.A., Barnes, K.C., Gibson, R.L., and Bamshad, M.J. 2012. Exome sequencing of extreme phenotypes identifies *DCTN4* as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. *Nat. Genet.* 44:886-889.

Fang, S., Sha, Q., and Zhang, S. 2012. Two adaptive weighting methods to test for rare variant associations in family-based designs. *Genet. Epidemiol.* 36:499-507.

Ferrer-Costa, C., Gelpi, J.L., Zamakola, L., Parraga, I., de la Cruz, X., and Orozco, M. 2005. PMUT: A web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* 21:3176-3178.

Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., and Amos, C.I. 2008. Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 82:100-112.

Han, F. and Pan, W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* 70:42-54.

Hartl, D.L. and Clark, A.G. 2007. Principles of Population Genetics, 4th ed. Sinauer Associates, Sunderland, Massachusetts.

Hirschhorn, J.N. and Daly, M.J. 2005. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6:95-108.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44:955-959.

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

Huang, B.E. and Lin, D.Y. 2007. Efficient association mapping of quantitative trait loci with selective genotyping. *Am. J. Hum. Genet.* 80:567-576.

Iyengar, S.K. and Elston, R.C. 2007. The genetic basis of complex traits: Rare variants or “common gene, common disease”? *Methods Mol. Biol.* 376:71-84.

Ji, W., Foo, J.N., O’Roak, B.J., Zhao, H., Larson, M.G., Simon, D.B., Newton-Cheh, C., State, M.W., Levy, D., and Lifton, R.P. 2008. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.* 40:592-599.

Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, R.P. 2012. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* 91:839-848.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-y., Freimer, N.B., Sabatti, C., and Eskin, E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42:348-354.

Keinan, A. and Clark, A.G. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336:740-743.

Ladouceur, M., Dastani, Z., Aulchenko, Y.S., Greenwood, C.M.T., and Richards, J.B. 2012. The empirical power of rare variant association methods: Results from Sanger sequencing in 1,998 individuals. *PLoS Genet.* 8:e1002496.

Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., and Lin, X. 2012a. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91:224-237.

Lee, S., Wu, M.C., and Lin, X. 2012b. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13:762-775.

Li, B. and Leal, S.M. 2008. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* 83:311-321.

Li, B. and Leal, S.M. 2009. Discovery of rare variants via sequencing: Implications for the design of complex trait association studies. *PLoS Genet.* 5:e1000481.

Li, B., Chen, W., Zhan, X., Busonero, F., Sanna, S., Sidore, C., Cucca, F., Kang, H.M., and Abecasis, G.R. 2012a. A likelihood-based framework for variant calling and *de novo* mutation detection in families. *PLoS Genet.* 8:e1002944.

Li, B., Wang, G., and Leal, S.M. 2012b. SimRare: A program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits. *Bioinformatics* 28:2703-2704.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987-2993.

Li, H. and Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.

Li, Y., Byrnes, A.E., and Li, M. 2010a. To identify associations with rare variants, just WHaIT: *W*eighted *h*aplotype *a*nd *i*mputation-based *t*ests. *Am. J. Hum. Genet.* 87:728-735.

Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. 2010b. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34:816-834.

Li, Y., Sidore, C., Kang, H.M., Boehnke, M., and Abecasis, G.R. 2011. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.* 21:940-951.

Lin, D.-Y. and Tang, Z.-Z. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* 89:354-367.

Liu, D.J. and Leal, S.M. 2010a. 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.* 6:e1001156.

Liu, D.J. and Leal, S.M. 2010b. Replication strategies for rare variant complex trait association studies via next-generation sequencing. *Am. J. Hum. Genet.* 87:790-801.

Liu, D.J. and Leal, S.M. 2012a. Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. *Am. J. Hum. Genet.* 91:585-596.

Liu, D.J. and Leal, S.M. 2012b. A unified framework for detecting rare variant quantitative trait associations in pedigree and unrelated individuals via sequence data. *Hum. Hered.* 73:105-122.

Liu, X., Jian, X., and Boerwinkle, E. 2011. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 32:894-899.

Maher, B. 2008. Personal genomes: The case of the missing heritability. *Nature* 456:18-21.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A., and Visscher, P.M. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747-753.

Mardis, E.R. 2008. Next-generation DNA sequencing methods. *Ann. Rev. Genom. Hum. Genet.* 9:387-402.

Mathieson, I. and McVean, G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* 44:243-246.

Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet.* 7:e1001322.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St. Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L., Woollard, P., Topp, S., Hall, M.D., Nangle, K., Wang, J., Abecasis, G., Cardon, L.R., Zöllner, S., Whittaker, J.C., Chissoe, S.L., Novembre, J., and Mooser, V. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337:100-104.

Ng, P.C. and Henikoff, S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31:3812-3814.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., Bamshad, M., Nickerson, D.A., and Shendure, J. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272-276.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Rev. Genet.* 12:443-451.

Pan, W. 2009. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.* 33:497-507.

Pan, W. and Shen, X. 2011. Adaptive tests for association analysis of rare variants. *Genet. Epidemiol.* 35:381-388.

Peng, B., Li, B., Han, Y., and Amos, C.I. 2010. Power analysis for case-control association studies of samples with known family histories. *Hum. Genet.* 127:699-704.

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

Price, A.L., Kryukov, G.V., de Bakker, P.I.W., Purcell, S.M., Staples, J., Wei, L.-J., and Sunyaev, S.R. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86:832-838.

Pritchard, J.K. and Przeworski, M. 2001. Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.* 69:1-14.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Ramensky, V., Bork, P., and Sunyaev, S. 2002. Human non-synonymous SNPs: Server and survey. *Nucleic Acids Res.* 30:3894-3900.

Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H., and Cohen, J.C. 2007. Population-based resequencing of *ANGPTL4* uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.* 39:513-516.

San Lucas, F.A., Wang, G., Scheet, P., and Peng, B. 2012. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with *variant tools*. *Bioinformatics* 28:421-422.

Sanger, F., Nicklen, S., and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74:5463-5467.

Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. 2009. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* 19:212-219.

Schwarz, J.M., Rödelsperger, C., Schuelke, M., and Seelow, D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7:575-576.

Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. *Nat. Biotechnol.* 26:1135-1145.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., and Haussler, D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-1050.

Slager, S.L., Huang, J., and Vieland, V.J. 2000. Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genet. Epidemiol.* 18:143-156.

Smith, D.J. and Lusis, A.J. 2002. The allelic structure of common disease. *Hum. Mol. Genet.* 11:2455-2461.

Stitziel, N.O., Kiezun, A., and Sunyaev, S. 2011. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* 12:227.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H.M., Jordan, D., Leal, S.M., Gabriel, S., Rieder, M.J., Abecasis, G., Altshuler, D., Nickerson, D.A., Boerwinkle, E., Sunyaev, S., Bustamante, C.D., Bamshad, M.J., and Akey, J.M. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64-69.

Wang, K., Li, M., and Hakonarson, H. 2010. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89:82-93.

Zhang, Y., Guan, W., and Pan, W. 2013. Adjustment for population stratification via principal components in association analysis of rare variants. *Genet. Epidemiol.* 37:99-109.

Zhu, X., Feng, T., Li, Y., Lu, Q., and Elston, R.C. 2010. Detecting rare variants for complex traits using family and unrelated data. *Genet. Epidemiol.* 34:171-187.

Zhu, Y. and Xiong, M. 2012. Family-based association studies for next-generation sequencing. *Am. J. Hum. Genet.* 90:1028-1045.