

Rare Variant Association Analysis: Beyond Collapsing Approaches

Han Chen and Josée Dupuis

Introduction

Because most studies do not have sufficient power to detect association with rare single nucleotide variants (SNVs), a number of approaches to jointly analyze SNVs have been proposed. The earlier approaches consisted of simply counting the number of rare alleles within a gene or pathway carried by each participant, and evaluating whether the count of rare alleles was associated with a trait or disease of interest. More sophisticated approaches followed, introducing weights to allow for some SNVs to have larger effects on the trait, and using of different definition of “rare” based on minor allele frequencies, described in detail in Chap. 13. However, these approaches had highest power when all rare SNVs had the same direction of effect on the trait studied, meaning that all SNVs were either detrimental or beneficial, and were seriously underpowered in situations where both detrimental and beneficial SNVs had an influence on the trait of interest, or a large proportion of SNVs were neutral.

To remedy the shortcoming of the earlier collapsing approaches, a number of methods allowing for different direction of effects were proposed and have been evaluated in simulation settings. In the next section, we outline these approaches, with emphasis on their commonality, advantages, and disadvantages in the analysis of rare SNVs.

H. Chen • J. Dupuis (✉)

Boston University School of Public Health, Boston, MA, USA

e-mail: dupuis@bu.edu

Methods

All approaches described in this section start from the following basic model:

$$g[E(Y_i)] = \gamma_0 + \sum_c \gamma_c z_{ic} + f(G_i) \quad (1)$$

where Y_i is the trait of interest, either a quantitative trait or a binary disease indicator, z_{ic} is the value of the c th covariate in individual i , γ_c is the effect of the c th covariate on the trait Y , G_i is the genotype at all SNVs within a functional unit (gene or pathway) for individual i , and $f(G_i)$ is a function on the genotypes. The function $g(\cdot)$ is a generalized linear model link function. For example, one may use the logit link function for binary traits and the identity link for quantitative traits.

More specifically, if $f(\cdot)$ is a linear function, then

$$g[E(Y_i)] = \gamma_0 + \sum_c \gamma_c z_{ic} + \sum_j \beta_j G_{ij} \quad (2)$$

where G_{ij} is the number of rare alleles carried by individual i at SNV j and β_j is the effect of SNV j on the trait.

In joint tests of association, the typical hypothesis of interest can be written as $H_0: \beta_j = 0$ for all j , although the specific form of the null hypothesis and the choice of test statistic vary according to the approach. For example, a general collapsing test statistic may be obtained by setting $\beta_j = \beta w_j$, where w_j is a weight assigned to the j th SNV. The w_j are assumed to be known, although in practice they are often estimated from the observed data. When assuming $\beta_j = \beta w_j$, (2) can be written as

$$g[E(Y_i)] = \gamma_0 + \sum_c \gamma_c z_{ic} + \beta \sum_j w_j G_{ij} \quad (3)$$

and the null hypothesis becomes $H_0: \beta = 0$. A Wald test, score test, or likelihood ratio test can be used to test the null hypothesis in a regression context. Using the notation and model defined in (1), we describe a number of methods for joint analysis of rare SNVs that go beyond the collapsing methods described in Chap. 13.

The Data-Adaptive Sum (aSum) Test

The data-adaptive sum (aSum) test proposed by Han and Pan (2010) is one of the earliest approaches developed for the scenario when both deleterious and protective SNVs are present. The original model used by Han and Pan reduces to (3) without covariates although it is simple to extend the approach to include covariates. The novelty of Han and Pan's approach rests in the definition of the vector of weight w_j , which depends on the observed data in the following way. Han and Pan defined $\hat{\beta}_{Mj}$

as the estimate of the effect of SNV j in the model with a single SNV included (M stands for marginal model), and P_{Mj} as the p -value for the test $H_0: \beta_{Mj}=0$. Then, for a pre-specified cutoff α_0 , Han and Pan suggested setting $w_j=-1$ if $\hat{\beta}_{Mj} < 0$ and $P_{Mj} \leq \alpha_0$, and $w_j=1$ otherwise. The choice of threshold α_0 will influence the power of the test. In the case of $\alpha_0=0$, all $w_j=1$ and the approach reduces to an unweighted collapsing test, where the rare SNV count is tested for association with a trait. In the case of $\alpha_0=1$, w_j is set to the sign of $\hat{\beta}_{Mj}$, the marginal effect of each SNV.

Han and Pan recommended using a score test to evaluate the association between $\sum_j w_j G_{ij}$ and the trait of interest. However, because the w_j 's are selected based on the significance and sign of the single SNV estimated effects, using the asymptotic distribution to assess the significance of the score test would lead to inflated type-I error rate. To surmount this problem, Han and Pan proposed a permutation approach, where phenotypes (and covariates if applicable) are permuted among unrelated individuals and the procedure is repeated, selecting the most appropriate w_j for each permuted dataset and computing the score statistic for association. Because significance thresholds in gene-based genome-wide studies are typically in the order of 10^{-6} , a large number of permutations would need to be performed in order to get accurate permutation p -values, which could render this procedure impractical. To alleviate this issue, Han and Pan evaluated a second approach to estimate the significance of their adaptive test by assuming that the distribution of the score statistic follows a shifted chi-square distribution of the form $a\chi_1^2 + b$, where a and b are parameters estimated from the permutation distribution. Estimation of a and b can be performed with a few hundred permutations, and this greatly increases the efficiency of the procedure. In their evaluation, Han and Pan used only 100 permutations to estimate a and b , and compared the p -value obtained under the shifted χ_1^2 assumption to a more typical permutation test with thousands of permutations.

Han and Pan performed extensive simulation studies, showing that their approach outperforms collapsing tests in many scenarios. Although the evaluation of aSum using the reduced number of permutations and the shifted χ_1^2 assumption appears to yield the correct type-I error, they cautioned that this approach should be more thoroughly studied and that the permutation distribution without this shifted χ_1^2 assumption is preferable, when feasible, to assess the significance of the test statistic. Given that Han and Pan explored the accuracy of the shifted χ_1^2 distribution at the $\alpha=0.05$ level only, and not in the tail when the accuracy is most important, this warning by the authors seems warranted.

The greatest advantage of the Han and Pan's approach is the gain in power over collapsing approaches when both deleterious and protective SNVs influence the trait of interest. However, there are a number of shortcomings to the approach. First, the permutation procedure greatly increases the computational burden. Second, the method is only applicable to unrelated individuals because the permutation procedure assumes that observations are interchangeable, and hence independent. This assumption will be violated in family samples and may be too restrictive in unrelated samples with cryptic relatedness, as would be present in population isolates. Finally, Han and Pan's approach will be most powerful when all SNVs

have the same magnitude of effects because of the simple +1/−1 weighting scheme. Because it is expected that some SNVs may have a large effect on the trait of interest, and that some SNVs may have no effect at all, a number of approaches were proposed to address this weakness.

Step-Up Method

Hoffmann et al. (2010) proposed a general step-up approach to allow SNVs to have different effect on the trait, taking into consideration that some SNVs may have no effect at all. Model (3) is also the basis for the step-up approach, although their original model does not allow for inclusion of covariates. However, the approach could easily accommodate covariate adjustments. Again, the difference in the step-up approach from other proposed rare SNV methods comes down to 1) the choice of test statistic, and 2) the choice of weights w_j .

To evaluate the association between SNVs and trait, Hoffman et al. (2010) suggested using the score test with empirically derived variance:

$$T = \frac{\left[\sum_{i=1}^N U_i \right]^2}{\sum_{i=1}^N U_i^2}$$

where $U_i = (Y_i - \bar{Y}) \sum_{j=1}^J w_j (G_{ij} - 2\hat{p}_j)$, $\bar{Y} = \sum_{i=1}^N Y_i / N$, and \hat{p}_j is the estimated minor allele frequency of SNV j $\left(\hat{p}_j = \sum_{i=1}^N G_{ij} / [2N] \right)$. This statistic can be computed efficiently for both binary and quantitative traits. Inclusion of covariates can be accommodated by replacing \bar{Y} by $\hat{\mu}_i$, where $\hat{\mu}_i = \hat{\gamma}_0 + \sum_c \hat{\gamma}_c z_{ic}$ for continuous traits and $\hat{\mu}_i = \text{logit}^{-1} \left(\hat{\gamma}_0 + \sum_c \hat{\gamma}_c z_{ic} \right)$ for binary traits, with $\hat{\gamma}_0$ and $\hat{\gamma}_c$ estimated under the null hypothesis of no rare variant influence on the trait. When the weights are known, this score statistic follows asymptotically a χ^2 distribution. However, the optimum weighting scheme is usually unknown, and the authors proposed various ways of setting the weights w_j to maximize power.

Hoffman et al. (2010) proposed to use weights of the form $w_j = a_j s_j v_j$, where a_j is a continuous weight, s_j depends on the direction of effect, as in the Han and Pan's approach above, and v_j is an indicator variable specifying whether SNV j belongs in the model (i.e., has a nonzero effect on the trait studied). This model addresses two of the shortcomings of the Han and Pan's approach. First, it takes into account that some SNVs may be "noise" SNV and have no effect on the trait. Secondly, the a_j 's allow for SNVs to have different effect sizes. Although this is a very general model, one has to define a_j , s_j , and v_j in order to perform a test of

association. In the next paragraph, we describe some options that Hoffmann et al. (2010) proposed for setting the components a_j , s_j , and v_j .

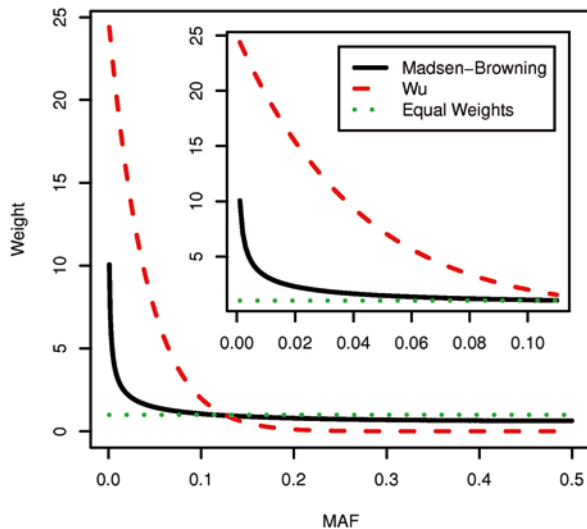
The term a_j allows for SNVs to have different magnitude of effect on the trait. If one assumes that rarer SNVs have a larger effect on the trait, a natural choice for a_j is the Madsen–Browning weight function (2009) that depends on the allele frequency and are proportional to $\frac{1}{\sqrt{\hat{p}_j(1-\hat{p}_j)}}$, where \hat{p}_j is the estimated minor

allele frequency of SNV j . A more general form for a_j is the beta function with parameters α and β . Setting $\alpha=\beta=1/2$ is equivalent to the Madsen–Browning weight (2009). Wu et al. (2011) proposed using $\alpha=1$ and $\beta=25$. If one assumes that all SNVs have the same effect on the trait, then one should give equal weights to all SNVs by setting $a_j=1$. A comparison of these three weighting schemes is presented in Fig. 1.

The other components of the weighting function, s_j , allow for SNVs to have different direction of effect ($s_j = -1$ or $+1$). Values of s_j are usually set based on the observed data. As described above, Han and Pan (2010) proposed an approach for setting s_j based on the sign (and significance) of the regression coefficient. Hoffmann et al. (2010) proposed a modified approach that is computationally more efficient when there are no covariates. For binary traits, $s_j = -1$ when the SNV is more prevalent in controls, and $+1$ otherwise. For continuous traits, s_j is the sign of the correlation coefficient between the additively coded SNV and trait.

The final components of the weighting function, v_j , determine which SNVs are allowed to enter the model and hence are assumed to influence the trait. Values of v_j may be determined using prior information, such as functional annotation (e.g., v_j for non-synonymous SNVs), or may be data-driven (e.g., $v_j = 1$ if $\hat{p}_j < 0.01$). Hoffmann et al. (2010) proposed an iterative procedure for setting v_j called the

Fig. 1 The weighting schemes as a function of minor allele frequency (MAF)



“step-up” approach that is akin to forward selection in regression. First, all models with only one SNV are evaluated and the model with the largest score statistic is selected. Then, all models including that first selected SNV and one other SNV are evaluated. The score statistic for the best model with two SNVs is compared to the score statistic including only the best SNV; if the model with two SNVs has a higher score statistic than the model with one SNV, the procedure continues including SNVs in the model in this iterative fashion until the score statistic no longer increases.

Statistical significance of the final score statistic is evaluated empirically, permuting the trait values among all individuals and performing the step-up procedure for each permutation. The final p -value is the proportion of permutation datasets with a score statistic higher than the observed score statistic. The procedure has been implemented in an R package *thgenetics* (<http://cran.r-project.org/web/packages/thgenetics/index.html>). Although the R package does not allow for covariate adjustment, there is nothing in the theoretical development of the approach that would prevent inclusion of covariates. The R package is fairly efficient when analyzing a moderate number of SNVs (~ 20), but becomes highly computationally intensive with larger number of SNVs (~ 100) although the implementation allows the users to analyze subsets of SNVs that are then combined into a single test statistic (the “pathway” option).

The step-up approach is very general and encompasses many of the previously described collapsing tests and approaches. For example, if s_j is set according to the sign and significance of the regression coefficient from the marginal model, with $v_j = 1$, we are back to the Han and Pan’s approach. If the a_j is set to the Madsen–Browning weights, also with $v_j = 1$, then we get the Madsen–Browning test. However, the permutation procedure required by both the Hoffman et al. and the Han and Pan’s approaches poses a challenge for their genome-wide implementation. Moreover, the permutation approach is valid when observations are independent and therefore not appropriate for family samples without omitting related samples or adapting the permutation procedure to account for correlated observations, an issue that remains a challenge.

Sequence Kernel Association Test

Despite both the Han and Pan (2010) and Hoffman et al. (2010) approaches not having an implicit assumption that all SNV effects are in the same direction, the computational limitation imposed by the required permutation procedure is a drawback. Wu et al. (2011) proposed the sequence kernel association test (SKAT), a method that accommodates SNVs with different direction of effects and does not require permutation. SKAT is based on model (2), and the null hypothesis of interest is $H_0: \beta_j = 0$ for all j . However, because β_j cannot be reliably estimated for rare SNVs, Wu et al. (2011) assume that each β_j follows an arbitrary distribution with a mean of zero and a variance of $w_j^2\tau$, where w_j is a known weight for SNV j and τ is a variance component.

A test of $H_0: \beta_j=0$ for all j is equivalent to testing $H_0: \tau=0$. Wu et al. (2011) propose to perform a test of this latter hypothesis using a variance-component score test for a mixed model, assuming γ_c are fixed effects and β_j are random effects. The test statistic $Q = (Y - \hat{\mu}) GWWG'(Y - \hat{\mu})$, where $\hat{\mu}$ is the predicted mean of Y under the null hypothesis, $K = GWWG'$ is the weighted linear kernel matrix, W is a matrix whose diagonal elements are w_j and non-diagonal element are 0, and G is the $N \times J$ matrix of additively coded genotype. Note that $\hat{\mu} = \hat{\gamma}_0 + \sum_c \hat{\gamma}_c z_c$ for continuous traits and $\hat{\mu} = \text{logit}^{-1} \left(\hat{\gamma}_0 + \sum_c \hat{\gamma}_c z_c \right)$ for binary traits. In the special case where Y is binary and there are no covariates, the SKAT statistic is equivalent to the C-alpha test proposed by Neale et al. (2011). In the C-alpha statistics, each rare SNV has the same probability of occurring in cases and controls under the null hypothesis of no association. Excess occurrence in cases or in controls is taken as evidence for association. A measure of excess occurrence is aggregated over all SNVs to create the C-alpha statistic. The SKAT statistic can be seen as a generalization of the C-alpha test, allowing for continuous traits and covariates or equivalently, the C-alpha test is a special case of a SKAT statistic. Under the null hypothesis, Q follows a weighted

sum of χ_1^2 statistics, $Q \sim \sum_{j=1}^J \lambda_j \chi_{1,j}^2$ with λ_j estimated from the eigenvalues of a function of the weighted genotype covariance matrix. Therefore, evaluation of the significance of Q can be achieved analytically without resorting to permutation. The Q statistic can be re-written as the sum of the score test for each individual SNV:

$Q = \sum_{j=1}^J w_j^2 S_j^2$, where $S_j = \sum_{i=1}^N G_{ij} (Y_i - \hat{\mu}_i)$. When using equal weights (W is the identity matrix, all $w_j=1$), the SKAT statistic is equivalent to the sum of squares of the marginal score statistics (SumSqU, or SSU) proposed by Pan (2009). This form of the Q statistic is extremely useful when analyzing multiple cohorts. For example, one could use inverse variance weighted meta-analysis to obtain a pooled estimate of the score statistic for each variant, and use the meta-analyzed scores in the computation of the Q statistics. Similarly, the asymptotic distribution of the meta-analyzed Q could be obtained by pooling the genotype covariance matrix to evaluate significance.

More generally, instead of the linear function in model (2), SKAT can also take a more flexible function $f(G_i)$ in model (1), thus allowing for interactions among variants. Assuming the vector $f(G)$ of size N follows a distribution with mean 0 and covariance matrix τK , the test statistic $Q = (Y - \hat{\mu})' K (Y - \hat{\mu})$ may be used to evaluate the null hypothesis $H_0: \tau = 0$.

SKAT offers many advantages over other approaches. First, the computational efficiency that results from using asymptotic rather than empirical distribution of the test statistic under the null hypothesis makes it feasible to apply to genome-wide studies. Moreover, the robustness of the test statistic to the direction and magnitude of effects offers increased power in scenarios where both deleterious and protective SNVs are at play. However, when most SNVs have the same direction of effect, SKAT has been shown to be less powerful than a simpler burden tests. For this reason, a combination of burden test and SKAT statistic may offer better power.

SKAT-O

When most SNVs included in the analysis are functionally related to the trait of interest and have the same direction of effect, then a burden test may outperform SKAT. Lee et al. (2012) proposed an extension to the SKAT statistic to deal with this scenario. They proposed a different class of kernels to use in the SKAT test, and the resulting Q statistic derived from this class of kernels is equivalent to a linear combination of the burden test and SKAT statistics:

$$Q_\rho = (1 - \rho)Q_{\text{SKAT}} + \rho Q_{\text{burden}}, \quad \text{with } 0 \leq \rho \leq 1.$$

When $\rho=0$, Q_ρ reduces to the SKAT statistic; when $\rho=1$, Q_ρ reduces to the burden test statistic $Q_{\text{burden}} = \left(\sum_{j=1}^J w_j S_j \right)^2$, which is the square of the score test statistic for $H_0: \beta=0$ in model (3). For a fixed value of ρ , the distribution of Q_ρ follows a weighted sum of χ_1^2 distribution, with weights estimated from the eigenvalues of a function of the weighted genotype covariance matrix. However, Lee et al. (2012) suggested a data-driven approach to setting the value of ρ to optimize power by finding the minimum p -value over all values of ρ . They provide a procedure to evaluate the significance of this new test statistic that takes into consideration the fact that the p -value was minimized over ρ , a nuisance parameter which is present only under the alternative hypothesis. Again, the procedure does not require permutation and is highly computationally efficient. The name of this new procedure is SKAT-O, where ‘‘O’’ stands for optimized. Via simulations, Lee et al. (2012) showed that this procedure has close to equivalent power to the burden test when a large proportion of the SNVs have the same direction of effect, and power close to the original SKAT statistic in the context of SNVs with different direction of effect.

Wang et al. (2012) also proposed a joint test (Score-Joint), combining a burden test, equivalent to the square-root of Q_{burden} above, with a test of the variance component parameters τ defined in the SKAT section. Compared with SKAT-O, it is a joint test on two parameters, and it requires permutation to evaluate significance.

The SKAT-O statistic offers some power advantage over the original SKAT procedure when the proportion of influential SNVs is large and most SNVs have the same direction of effect, at small cost of some added complexity in computation.

Score-Seq

Lin and Tang (2011) proposed a slightly different procedure to test for association between a group of SNVs and a trait of interest. As for all the previous approaches, the basis of the method is model (2). In the same spirit as many of the collapsing approaches, Lin and Tang’s approach assumes that $\beta_j = \beta w_j$, where w_j is the weight assigned to SNV j , and the model reduces to model (3) previously described. To test

the null hypothesis $H_0: \beta=0$, assuming a known vector of weights w , Lin and Tang derived the score statistic, which is of the form $U = \sum_{i=1}^N (Y_i - \hat{\mu}_i) G_i w$ with variance $V = \hat{\sigma}_0^2 \left\{ \sum_{i=1}^N (G_i w)^2 - N^{-1} \left(\sum_{i=1}^N G_i w \right)^2 \right\}$ when there is no covariates involved (but a more complex form with covariates). Note that $\hat{\sigma}_0^2 = \bar{Y}(1 - \bar{Y})$ for binary trait and the estimated variance of Y for quantitative trait. The statistic $T = U / \sqrt{V}$ may be used to determine if the rare SNVs have an effect on the trait Y . The power of the test will depend on the choice of w , with optimum power achieved when $w_j = \beta_j$, the true (but unknown) value of the effect size parameter. Lin and Tang’s approach differs from the typical weighted collapsing method in setting the values of the weight vector. When considering weighting schemes, Lin and Tang proposed two ways to achieve maximum power: (1) Maximizing the test statistic over multiple weight vectors and (2) setting weights from the Estimated REgression Coefficients (EREC). We describe both sets of weights below.

Maximizing the Test Statistic Over Multiple Weight Vectors

Given L weight vectors, w^1, \dots, w^L , each of length J , that include the weights for each of the J SNVs in the analysis, one can compute L score statistics (T_1) to test the association between the trait and the weighted genotypes formed by Gw^l . Ling and Tang suggested using the maximum test statistic over all weight vectors ($T_{\max} = \max |T_1|$) to test for association between the SNVs and the trait. They derive the asymptotic distribution of T_{\max} by assuming that the T_1 statistics follow a multivariate normal distribution with mean 0, and with an estimated covariance matrix that can be computed from the data and weight vectors. Significance of the test can be evaluated asymptotically using the equation:

$$\Pr(T_{\max} > t_{\max}) = 1 - \Pr(|T_1| < t_{\max}, \dots, |T_L| < t_{\max}).$$

For example, one could evaluate the T statistic for equal weight ($w_j = 1$ for all j), the Madsen–Browning weight and the Wu weight, and use the maximum statistic over these three weight vectors, taking into account that the statistic was maximized over three weight functions when evaluating significance. This may offer increased power over collapsing approaches using a single set of weights. One could also define the weights with a variable threshold based on allele frequencies to determine inclusion of SNVs, and maximize over multiple allele frequency thresholds. This is akin to the variable threshold (VT) test proposed by Price et al. (2010), with the added advantage that significance may be evaluated without the need for computationally intensive permutations.

One of the greatest advantages of this approach is the ability to evaluate empirically the significance of the test statistics when multiple weight functions are evaluated.

In practice, because the trait etiology is often unknown and one does not know, a priori, which rare SNVs influence the trait, investigators often evaluate multiple weight functions, which may involve restricting which SNVs are included in the test, based on function or other annotation, or by relaxing the definition of “rare” to allow common SNVs to be included. However, correction for multiple testing is often performed using a simple Bonferroni correction, leading to overly conservative tests because the correction does not take the correlation of the test statistics into consideration. The ability to properly correct for multiple testing induced by the evaluation of multiple weights function is a great addition to the literature.

Nevertheless, the approach would still have low power in the presence of both deleterious and protective rare SNVs, prompting Lin and Tang to explore a different approach to determine the optimal weight vector.

Estimated REgression Coefficients

As noted earlier, the most powerful test would be obtained by setting $w_j = \beta_j$, the true but unknown value of the parameter. While β_j may be estimated from the data, it will likely be poorly estimated because of the low frequency of the tested alleles. Lin and Tang suggested setting $w_j = \hat{\beta}_j + \delta$, where δ is a given constant. This is similar to Han and Pan’s earlier approach, where w_j was dependent on the significance and sign of the beta estimate, although Han and Pan (2010) ignored the magnitude of the effect estimates. Because the data is used in setting the optimum weights, significance is evaluated using a permutation approach, where the phenotype value Y (and covariates if applicable) are permuted among individuals, and both weights and test statistics are recomputed with permuted data. It is important to permute both trait and covariates together; the null hypothesis is evaluated by breaking the relationship between genotype and trait, but keeping the relationship between the trait and covariates intact. Lin and Tang implemented this approach into the software Score-Seq, with an adaptive permutation test that selects fewer permutation iterations for large p -values but increases the number of permutation iterations to get more precision for low p -values.

The authors recommend setting $\delta = 1$ for binary traits and $\delta = 2$ for standardized quantitative traits when the sample size is less than 2,000. The authors have not explored the effect of varying δ on power.

The authors compared the multiple weight evaluation approach and Estimated REgression Coefficients (EREC) method with other available methods, namely the collapsing approach by Madsen and Browning (2009), the variable threshold approach proposed by Price et al. (2010), and SKAT. They showed the advantage of evaluating multiple weight functions over most collapsing tests when all SNVs had the same direction of effect. They also showed that EREC has a clear advantage over SKAT when all SNVs have the same direction of effect with no neutral SNVs included, a fact that was acknowledged by Wu et al. (2011) and remediated with the introduction of the SKAT-O statistic. In the presence of both deleterious

and protective SNVs, Lin and Tang (2011) also demonstrated an advantage of EREC over the SKAT statistic, claiming that the gain in power is due to the overly conservative asymptotic evaluation of the significance of SKAT statistic, while their permutation evaluation is not conservative. However, they acknowledge that the SKAT method is more computationally efficient than the EREC test.

Kernel-Based Adaptive Cluster

Liu and Leal (2010) proposed the kernel-based adaptive cluster (KBAC) approach, which classifies genotypes into groups based on multi-locus genotype patterns. Their method can be formulated using model (1) defined earlier.

For a set of J variants, there are at most 3^J genotype groups. However, when testing rare variants, the number of observed genotype groups may drop dramatically because of the low minor allele frequency and linkage disequilibrium. Given J SNVs, the $M + 1$ distinct genotype patterns are denoted by P_0, P_1, \dots, P_M , and P_0 represents a pattern with no rare alleles. Using the model defined in (1), Liu and Leal (2010) let $f(G_i) = \eta K_m$ for individual i with genotype pattern P_m , where the kernel K_m is estimated from the data. The null hypothesis $H_0: \eta = 0$ is evaluated using a score test to determine if there is some association between genotype patterns and phenotype. Because the kernel is data-driven, a permutation procedure is implemented for p -value evaluation.

Liu and Leal proposed (2010) three types of kernels for case–control designs: hyper-geometric kernel, marginal binomial kernel, and asymptotic normal kernel. Their evaluation of the approach focused on the hyper-geometric kernel, defined as

$$K_m = \sum_{r \in \{0, 1, \dots, N_m^1\}} \frac{\binom{N_m}{r} \binom{N - N_m}{N^1 - r}}{\binom{N}{N^1}}$$

where N^1 and N^0 are the number of cases and controls, respectively, with $N = N^1 + N^0$, and N_m is the number of individuals with genotype pattern P_m among which there are N_m^1 cases and N_m^0 controls. The kernel is different from the kernel in SKAT, because it is data-driven and depends on the genotype–trait relationship. Appropriate kernels for quantitative trait analyses were not proposed.

When there are no covariates, the score statistic from the logistic regression model (1) reduces to the KBAC statistic (up to a constant scalar):

$$\text{KBAC} = \left(\sum_{m=1}^M K_m \left(\frac{N_m^1}{N^1} - \frac{N_m^0}{N^0} \right) \right)^2,$$

Basu and Pan (2011) suggested that KBAC might not perform well when there are both deleterious and protective variants, and when the proportion of causal variants is small. However, compared with other approaches, KBAC is attractive in rare variants association analysis because it allows for interactions among variants, by testing genotype patterns of multiple variants as a group, rather than simply summing up genotypes or test statistics from individual variants.

Discussion

All approaches described in this chapter use the same underlying model linking a trait to rare SNV genotypes, described in (1). Many other approaches for rare variant analyses have been proposed in the literature. Two examples of non-regression-based approach include the replication-based test (RBT) proposed by Ionita-Laza et al. (2011) and the functional principal component analysis (FPCA) introduced by Luo et al. (2011)

The RBT was developed for case–control designs and looks for more frequent occurrences of mutations in either cases or controls. Enrichment in cases is measured by a weighted sum of indicators of higher allele frequency in cases compared to controls, where the weights are data-driven and are higher for variants with larger difference in allele frequency between cases and controls. Because rare variants may be protective, a similar statistic for enrichment in controls is computed, and the RBT statistic is defined as the maximum of the two enrichment statistics. Statistical significance is evaluated by permutation. Compared with burden tests, RBT is less sensitive to the presence of both deleterious and protective variants, but power is reduced when the proportion of causal variants is low.

Luo et al. (2011) proposed the FPCA approach, which takes both rare variants and their genomic locations into consideration. From a functional data analysis point of view, they treat the positions as a continuous variable and define the genotype of each individual as a function of positions. By using data reduction and smoothing techniques, FPCA overcomes the high-dimensionality and multicollinearity issues in multivariate tests and collapsing methods, and is less sensitive to sequence errors and missing data. However, the multivariate nature of the Hotelling's T^2 test performed after reducing the dimension of genotype data using principal components may hamper power over lower dimensional methods described in this chapter. When the correlation between rare variants is low, FPCA introduces extra computational burden, but may not have much power gain compared to multivariate tests on the original genotype data. Also, FPCA does not adjust for covariates and is not directly applicable to quantitative traits, although such extensions would be straightforward.

In an ideal world, one would have infinite data and would be able to assess the effect on the phenotype of each rare variant individually. However, because of limits in sample sizes imposed by budget constraints and also simply by the availability of cases for certain rare diseases, getting reliable estimate of the effect of rare SNV on

Table 1 Summary of non-collapsing rare variant association analysis approaches

Test	Binary	Quantitative	Covariates	p -Value	References
SSU	Yes	Yes	Yes	Analytical	Pan (2009)
aSum	Yes	Yes	Yes	Permutation	Han and Pan (2010)
KBAC	Yes	No	Yes	Permutation	Liu and Leal (2010)
Step-up	Yes	Yes	Yes	Permutation	Hoffmann et al. (2010)
RBT	Yes	No	No	Permutation	Ionita-Laza et al. (2011)
C-alpha	Yes	No	No	Either	Neale et al. (2011)
FPCA	Yes	No	No	Analytical	Luo et al. (2011)
SKAT	Yes	Yes	Yes	Analytical	Wu et al. (2011)
Score-Seq	Yes	Yes	Yes	Analytical	Lin and Tang (2011)
EREC	Yes	Yes	Yes	Permutation	Lin and Tang (2011)
Score-Joint	Yes	Yes	Yes	Permutation	Wang et al. (2012)
SKAT-O	Yes	Yes	Yes	Analytical	Lee et al. (2012)

the quantitative trait or disease of interest is often not feasible. Therefore, additional assumptions are needed in order to identify rare SNVs associated with a phenotype. The rare variant approaches included in this chapter differ in their assumptions. Obviously, the closer the assumptions are to the “truth,” the more effective the approaches will be at identifying SNVs and genes that are important in disease etiology. The most powerful approach will often depend on the true trait model, which unfortunately remains unknown for most traits under investigations. To a lesser extent, the choice of test statistic will also affect the ability to identify the causal variants. Table 1 summarizes the non-collapsing rare variants association analysis approaches mentioned in this chapter. Below we discuss differences between the approaches presented in this chapter, and how these differences may affect the ability to identify SNVs and genes influencing a quantitative trait or disease of interest.

Test Statistic and Evaluation of Statistical Significance

The approaches described in this chapter differ by the test statistic used to evaluate the null hypothesis of no association. However, they all have one thing in common: they strive to use computationally efficient statistics that can be computed genome-wide. All approaches use a score test because it is less computationally intensive than a likelihood ratio test. Moreover, all approaches strive for efficient evaluation of their score test.

In aSum, although permutation is required for evaluation of the score statistic, Han and Pan (2010) investigated ways to decrease the computational burden of their permutation procedure. Because very small p -values are required when analyzing multiple genomic regions, a large number of permutations are typically required to estimate such small p -values. Han and Pan (2010) investigated approximation to the

permutation distribution by a scaled non-central chi-square, and used a small number of permutations to estimate the scaling and shift parameters.

Hoffmann et al. (2010) also used a score statistic and permutation. To improve upon Han and Pan's method in terms of computation efficiency, they determine the direction of effect based on the correlation coefficient, doing away with formal testing of each variant. In addition, they implemented an adaptive permutation approach, where a few initial permutations are used to assess the p -value, and additional permutations are performed only when the p -value is below a certain threshold. While it is certainly feasible to apply Hoffman et al.'s approach to a large number of genomic regions across the genome, the computational burden of the permutation approach prevents large-scale simulation evaluation of the approach.

The SKAT and SKAT-O statistics are also score tests, but with the advantage that statistical significance can be evaluated theoretically, without requiring time consuming permutation. However, Lin and Tang noted that SKAT can be conservative, and suggested that permutation evaluation could improve power, especially for small samples.

While both SKAT and EREC offer a general framework to test for association between a group of SNVs and a trait using a score test, the difference in their underlying assumptions lead to a different score statistics: SKAT assumes that β_j follows a distribution with mean 0 and variance $w_j^2\tau$, while Lin and Tang (2011) assumes that β_j is of the form βw_j . Both methods are univariate tests, but τ is a variance parameter with one-sided alternative in SKAT, and β is a location parameter with two-sided alternative in Lin and Tang (2011), leading to different statistics with different distributions. While the significance of both score statistics may be evaluated empirically, Lin and Tang further propose to set the weights empirically, and because the data is used in setting weights, asymptotic evaluation is no longer possible.

KBAC classifies individuals into different groups based on genotype patterns, and performs a test on the difference between the proportions of each genotype group in cases and in controls. The test is similar to a weighted χ^2 test of independence. Liu and Leal (2010) used permutation to evaluate statistical significance. Noting that the original KBAC statistic suffers when there are both deleterious and protective variants within a particular genotype pattern, Basu and Pan (2011) proposed a modified statistic to overcome this issue. KBAC is distinctive in rare variant analysis by allowing for interactions, but it may suffer from loss of power when the proportion of non-causal variants is high, as the number of genotype patterns increases dramatically.

Missing Data and Imputing Rare SNVs

While most of the methods discussed in this chapter have been evaluated using targeted or exome sequencing, application of the methods could be extended to imputed genotypes. Rare SNVs are often poorly imputed in unrelated samples because of the low linkage disequilibrium with nearby SNVs. However, familial

transmission information, if available, may improve imputation. In model (2), the genotypes could be defined as the expected number of rare alleles, or dosage, instead of a three category variable indicating the number of rare alleles a person carries. Theoretically, all approaches described in this chapter based on model (2) can accommodate the use of dosage genotype, although not all software implementation can do so.

In the regression framework of model (2), missing genotypes are not allowed. One needs to either exclude observations with one or more missing genotypes, or impute such missing data. As the number of SNVs included in the analysis increases, excluding observations with missing genotypes will greatly reduce the sample size and power, even when the genotyping call rate is high. Therefore, most software includes some approaches for imputing the missing values. For rare SNVs, one option would be to set all missing genotypes to the homozygous major allele, which is the most likely genotype. This imputation scheme is easy to implement. However, for more common SNVs, it will create bias in allele frequency estimates, which in turn could result in false positive results if the missing rate differs in cases and controls. For this reason, SNVs with high missing rates are often omitted from analysis. A second approach to fill in missing genotypes is to impute the mean genotype value, or dosage, which is equal to twice the rare allele frequency. While this will not bias the estimate of allele frequency, this may cause other types of bias. For example, if the missingness is not random and participants with missing data are more likely to be from the case or control set, or if they have lower or higher trait values, then imputing the average dosage may create false association because most observations will have a genotype of 0 rare allele, while missing observations will have a dosage value of twice the rare allele frequency. This could be more pronounced if the imputation is performed in cases and control separately. The third option is to impute the missing data using information on nearby SNV and familial transmission, if available. This approach capitalizes on linkage disequilibrium at nearby SNVs to more precisely impute missing genotypes. Unfortunately, this type of imputation works best for common SNV, but imputation quality for rare SNV can be poor, especially if no familial information is available. Again, SNVs with differential missingness in cases and controls, or missingness pattern related to a quantitative trait studied, could lead to false positive errors. To avoid such bias one can omit SNVs with high missing rate, but also test for differential missingness in cases or controls, or for association between proportion missing genotypes and a quantitative trait. Wu et al. (2011) showed that for small amount of missingness, imputing to the most likely genotypes did not decrease power considerably.

Choice of Weights to Maximize Power

Weighting schemes are used in most rare variant methods to try to improve power to detect association between SNVs and trait. To reach maximum power, a weighting scheme should give close to zero weights to SNVs without effect on the trait,

and weights proportional to the effect size for associated SNVs. Because it is believed that rarer SNVs will have a larger effect on the trait, several proposed weighting schemes depend on the rare allele frequency, such as the Madsen–Browning and Wu weights. Madsen–Browning weights decrease much more rapidly than the Wu weight as the minor allele frequency increases; see Fig. 1. As a consequence, the effect of including more common SNVs when using the Madsen–Browning weight should be small, while more common SNVs would contribute more substantially to the test statistic under the Wu or equal weighting scheme. Hoffman et al. (2010) also described ways to include functional annotation in determining the weights, assuming that SNVs that are more likely to be damaging or functionally important would have a larger effect on the trait. Such annotation can also be incorporated in the SKAT and score-seq weighting schemes, although the Han and Pan +1/−1 cannot be easily generalized to take functional annotation into consideration. FPCA can also take weighted genotypes instead of original additive genotypes and calculate principal components. As prior information becomes more precise, methods that can incorporate information on function annotation will be most useful.

Which SNVs to Include in Association Testing

Ideally, only SNVs influencing the trait of interest would be evaluated for association with the trait. Unfortunately, one does not know, a priori, which SNVs are causal or in LD with causal SNVs, and which SNVs have no effect on the trait. Inclusion of “noise” SNVs will lower the power of the test, as will failure to include some causal SNVs. Therefore, one has to strike a balance between including too many SNVs, with some noise SNVs, and too few SNVs, missing important variants. There are two separate issues to deciding which subset of SNVs to include in a test: (1) definition of the genomic region and (2) selection of SNVs within a region.

While one may wish to evaluate large regions for association, inclusion of too many SNVs, many likely to have no effect on the trait, will impede the ability to detect true associations. Therefore, it is common to divide large genomic regions into smaller analysis units. A natural unit of analysis is a gene level, or if a finer division is sought, exons or transcripts may be used to define a genomic region of interest. However, most Genome-Wide Association Study (GWAS) findings map outside of gene regions, and investigators may wish to evaluate rare SNV in the region around GWAS findings (Hindorff et al. 2009). Genomic region boundary could be based on conserved regions across species, recombination estimates around the GWAS finding, or more agnostically based on sliding windows across the region of interest. The sliding window approach could easily be accommodated in the framework from Lin and Tang (2011), where the test statistic used is the maximum test statistic over a number of weight functions. One can think of a sliding window as putting a weight of zero to all SNVs outside the window being considered, and use the method describe in Lin and Tang to get the significance of the maximum test statistic over multiple windows within a region. As a clearer picture emerges of how

rare SNVs influence traits, we will be able to use prior information to determine the best size and boundaries to define genomic regions for investigation. In the meantime, one has to explore various ways of defining genomic regions in order to maximize the chance to detect true associations.

Once genomic regions have been selected, one needs to determine which SNVs within the region to include in an association test. Burden tests often restrict analyses to SNVs with a low rare allele frequency, using threshold of 1, 2, or 5 %, and similar thresholds may be applied to the methods in this chapter. The optimal allele frequency threshold will depend on the frequency of the true causal SNVs, and using a too stringent threshold will omit important SNVs and reduce power, while a threshold that is too liberal will include too many noise SNVs and also decrease power. Variable threshold approaches, such as the one developed by Price et al. (2010) or by Lin and Tang (2011), can overcome the issue of having to evaluate a single allele frequency threshold. Functional annotation may also be used to try to identify SNVs that are more likely to influence the trait. However, recent publications indicate that there are a lot of functional elements outside of genes, so restricting analyses to protein-altering SNVs may miss important functional variants. Other measures of potential functionality, such as how conserved the region around the SNVs is in other species may be fruitful. An alternative is to include all SNVs within a region, but to use a weighting scheme to up-weight SNVs that are more likely to influence to trait based on annotation, and to down-weight SNVs that are most likely neutral. Hoffmann et al.'s approach gives specific examples regarding inclusion of prior annotation information in the evaluation of the null hypothesis of no association. Incorporating this information can be easily done by using different weighting schemes in the SKAT, Score-Seq, or FPCA framework. Obviously, as our functional annotation improves, our ability to detect genes and SNVs influencing the trait will also improve.

Meta-analysis

Another consideration when selecting the most suitable approach for analysis of rare SNVs is the availability of meta-analysis approaches. In the GWAS context, most discoveries were achieved after the formation of large consortia, where meta-analysis of many cohorts uncovered loci with smaller effect on the traits of interest. The need for larger sample sizes may be even more pronounced in the analysis of rare SNVs, where a single cohort may have very few individuals carrying rare alleles for a particular SNV, so that joining forces with other studies will be crucial for discoveries of rare SNV association. Because all approaches provide evaluation of the significance of an association test in the form of a p -value, one can use a p -value-based approach, such as the Fisher or Stouffer approach, for combining results from multiple cohorts. However, methods that directly combine the beta estimates from model (2) may offer improved efficiency (Lee et al. 2013; Liu et al. 2014). Development of efficient meta-analysis approaches will be important in our quest to identify rare variants influencing traits of interest.

Other Types of Traits

While most traits studied fall in two categories, binary disease status or continuous measurements such as blood pressure, lipid levels, or fasting glucose, other phenotypes of interest may be time-to-event or ordinal/categorical measures. For example, one may be interested in studying time to recurrence of cancer, or age of development of type 2 diabetes. Some psychiatric disorders may have multiple levels of severity and may be best coded as ordinal variables (American Psychiatric Association 2000). Some of the approaches above naturally extend to other types of phenotypes. For example, Lin and Tang provide details on the application of their approach for time-to-event data, although their software implementation does not include this option. Chen et al. (2014) extended the SKAT statistic for survival traits. Other approaches, such as Han and Pan's or Hoffmann et al.'s method, could easily accommodate survival and ordinal traits using the typical regression framework (Cox proportional hazard model for survival and generalized linear model for ordinal data) because the significance is evaluated using permutation. The limiting factor is incorporation of these options into user-friendly and computationally efficient software that are easily accessible to investigators with these types of data.

Exome sequencing, exome chip, and whole genome sequencing have opened the floodgate on rare variants that were not investigated in the earlier GWAS era, when most studies focused on SNVs with frequency $>1\%$. Our success in identifying key genes that influence diseases and traits of interest will rest on the appropriate use of statistical tools, and gathering as much knowledge as possible on the potential function of the variants under study. Hopefully, the combination of these tools will lead to exciting new discoveries and will further our understanding of the architecture of complex traits.

References

- American Psychiatric Association (2000) Diagnostic and statistical manual of mental disorders: DSM-IV-TR®. APA, Washington, DC
- Basu S, Pan W (2011) Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 35(7):606–619
- Chen H, Lumley T, Brody J, Heard-Costa NL, Fox CS, Cupples LA et al (2014) Sequence kernel association test for survival traits. *Genet Epidemiol* 38:191–197
- Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 70(1):42–54
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* 106(23):9362–9367
- Hoffmann TJ, Marini NJ, Witte JS (2010) Comprehensive approach to analyzing rare genetic variants. *PLoS One* 5(11), e13584
- Ionita-Laza I, Buxbaum JD, Laird NM, Lange C (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet* 7(2), e1001289

- Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13(4):762–775
- Lee S, Teslovich TM, Boehnke M, Lin X (2013) General framework for meta-analysis of rare variants in sequencing association studies. *Am J Hum Genet* 93(1):42–53
- Lin DY, Tang ZZ (2011) A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89(3):354–367
- Liu DJ, Leal SM (2010) 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(10), e1001156
- Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S et al (2014) Meta-analysis of gene-level tests for rare variant association. *Nat Genet* 46:200–204
- Luo L, Boerwinkle E, Xiong M (2011) Association studies for next-generation sequencing. *Genome Res* 21(7):1099–1108
- Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5(2), e1000384
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M et al (2011) Testing for an unusual distribution of rare variants. *PLoS Genet* 7(3), e1001322
- Pan W (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* 33(6):497–507
- Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ et al (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86(6):832–838
- Wang Y, Chen Y, Yang Q (2012) Joint rare variant association test of the average and individual effects for sequencing studies. *PLoS One* 7(3), e32485
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89(1):82–93