Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants

Iuliana Ionita-Laza,[1,6,*] Seunggeun Lee,[2,6] Vlad Makarov,[1] Joseph D. Buxbaum,[3,4,5] and Xihong Lin[2,*]

Recent developments in sequencing technologies have made it possible to uncover both rare and common genetic variants. Genome-wide association studies (GWASs) can test for the effect of common variants, whereas sequence-based association studies can evaluate the cumulative effect of both rare and common variants on disease risk. Many groupwise association tests, including burden tests and variance-component tests, have been proposed for this purpose. Although such tests do not exclude common variants from their evaluation, they focus mostly on testing the effect of rare variants by upweighting rare-variant effects and downweighting common-variant effects and can therefore lose substantial power when both rare and common genetic variants in a region influence trait susceptibility. There is increasing evidence that the allelic spectrum of risk variants at a given locus might include novel, rare, low-frequency, and common genetic variants. Here, we introduce several sequence kernel association tests to evaluate the cumulative effect of rare and common variants. The proposed tests are computationally efficient and are applicable to both binary and continuous traits. Furthermore, they can readily combine GWAS and whole-exome-sequencing data on the same individuals, when available, and are also applicable to deep-resequencing data of GWAS loci. We evaluate these tests on data simulated under comprehensive scenarios and show that compared with the most commonly used tests, including the burden and variance-component tests, they can achieve substantial increases in power. We next show applications to sequencing studies for Crohn disease and autism spectrum disorders. The proposed tests have been incorporated into the software package SKAT.

Introduction

The rapid development of sequencing technologies has led to the launch of numerous sequencing studies for many complex traits.[1] In addition to discovery of common variants, usually defined as those having a population frequency of 5% or higher, sequencing allows discovery of low-frequency and rare variants as well. The relative contribution of rare and common variants to disease risk is unknown for many traits, but it is reasonable to assume that a combination of rare and common variants influences the risk of many complex diseases. Recent studies have shown that novel, rare, low-frequency, and common variants can all be contributory variants at the same disease locus.[2–4]

Over the past several years, genome-wide association studies (GWASs) have led to the identification of many common genetic variants associated with risk of diverse complex traits.[5] Although the variants identified so far usually explain only a small to modest part of the estimated heritability for a given trait, it has been shown for several traits, including schizophrenia (MIM 181500), bipolar disorder (MIM 125480), autism (MIM 209850), and human height (MIM 606255), that many common variants with small effects might cumulatively explain a substantial proportion of the heritability.[6–9]

The main strategy employed by GWASs has been to evaluate each variant individually with a univariate statistic, such as the Cochran-Armitage test for trend.[10] Such a variant-by-variant analysis has been shown to be underpowered for rare variants, and consequently, many groupwise association tests have been proposed,[11–23] including burden and variance-component (e.g., SKAT) tests.[22] Most groupwise association tests use a weighting scheme that upweights the contribution of rare variants and downweights the contribution of common variants,[12] and they thereby mostly test for the effect of rare variants. However, the relative influence of rare and common variants is not known a priori for any disease-related gene, and such a weighting scheme can lead to loss of power when common variants in a region under investigation are also associated with disease. Under commonly used simulation scenarios, the genetic variance explained by common variants in a small genetic region can be higher than that explained by rare variants in the region (see Appendix A). Currently, rare and common variants are tested separately with the use of different testing strategies (as described above). However, because the overall goal is to identify genes that contain disease risk variants, be they rare or common, it is desirable to test for the combined effect of rare and common variants with a unified statistical test that allows both rare and common variants to contribute fully to the overall test statistic.

In this paper, we develop omnibus procedures to test for the effect of both common and rare genetic variants on a trait of interest. We first revisit the definition of common

[1]Department of Biostatistics, Columbia University, New York, NY 10032, USA; [2]Department of Biostatistics, Harvard University, Boston, MA 02115, USA; [3]Seaver Autism Center, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [4]Departments of Psychiatry, Neuroscience, and Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [5]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[6]These authors contributed equally to this work
*Correspondence: ii2135@columbia.edu (I.I.-L.), xlin@hsph.harvard.edu (X.L.)
http://dx.doi.org/10.1016/j.ajhg.2013.04.015.

and rare variants in the context of sequencing data and propose a separation threshold that depends on the sample size. We then propose several tests for combining burden and variance-component test statistics for rare and common variants. These tests are applicable to both binary and continuous traits and population- and family-based designs. We show applications to sequencing studies for Crohn disease (MIM 266600) and autism spectrum disorders (ASDs).

Currently, whole-genome sequencing is very expensive for large-scale association studies. Instead, whole-exome sequencing (WES) focuses on a gene's protein-coding components, which represent about 1% of the whole genome. On the other hand, many variants identified in GWASs are in noncoding regions, as might be expected from the prevalence of noncoding variants assayed in these studies.[24] Indeed, according to the Illumina Gene Annotation files of Human1Mduo and HumanOmni5-4 arrays, over 90% of the variants on the array are in intronic or intergenic regions and only 3% and 8% are located in coding and exonic regions, respectively. Because many WES studies are performed on individuals with existing GWAS data, an important application of the proposed methods is in combining GWAS data with WES data on the same individuals. Another application is to the study of deep resequencing of GWAS loci, where rare, low-frequency, and common contributory variants are expected to coexist.[2]

Material and Methods

Definition of a Threshold to Partition Variants into Rare and Common Variants

As we mentioned in the Introduction, most of the existing sequence-based association tests use a weighting function (usually depending on the variant frequency) that upweights the contribution of rare variants and correspondingly downweights the contribution of common variants.[12,22] Such a weighting scheme is necessary if both rare and common variants are to be included together in the study of rare-variant effects (otherwise, common variants dominate). However, when common variants are important for disease risk, such an approach is likely to lose power. A different way to combine rare and common variants together is to first partition variants into two separate groups—rare and common—and then combine the results from association tests with variants in the two groups, e.g., with the use of combined multivariate collapsing (CMC).[11] In CMC, rare variants (e.g., those with a minor allele frequency [MAF] < 0.01) are collapsed together, whereas each common variant forms a separate group. Results from rare and common variants are then combined with the use of a multivariate Hotelling's T-Square statistic. Even though this approach involves the ad hoc choice of a frequency cutoff, it has the advantage of allowing both rare and common variants to better contribute to the overall test for the effect in the region, although a large number of degrees of freedom (df) are used for common variants.

A commonly used approach in the literature is to use a fixed-frequency threshold T, e.g., 0.01, to partition variants into rare and common groups. Variants with a sample frequency less than 0.01 are treated as *rare*, those with a frequency between 0.01 and 0.05 are treated as *low frequency*, and the rest are considered *common*. A different approach is to define the threshold as a function of the total sample size. Intuitively, a variant with frequency 0.01 is rare in a small data set of 500 individuals but is quite common in a much larger data set with, say, 100,000 individuals. One large sample theory threshold[25] is to take

$$T = \frac{1}{\sqrt{2n}},$$

where n is the number of individuals in the study. Specifically, if one defines a variant as being common if it can be analyzed by itself with moment-based statistics (such as sample mean), then a natural asymptotic threshold is $1/\sqrt{2n}$, which is proportional to the SD of the sample mean. Note that this threshold only depends on the total sample size. It is not an optimal separation cutoff, given that such a cutoff would necessarily depend also on the true disease model, which is unknown to us.

In this setting, variants with MAF $\leq 1/\sqrt{2n}$ are considered rare, whereas variants with MAF $\geq 1/\sqrt{2n}$ are considered common. When $n = 500$, then $T = 0.031$. When $n = 10{,}000$, then $T = 0.007$. In the Results, we perform sensitivity analyses to investigate how this threshold compares with commonly used thresholds, such as 0.01 or 0.05, under several disease models.

Testing for the Overall Effect of Rare and Common Variants

To test for the overall effect of rare and common variants, we consider here several possible approaches that make use of the previously developed burden and variance-component (SKAT) tests for rare and common variants.[18,22] One simple approach is based on Fisher's method of combining the p values from the rare and common variant tests. Alternative approaches are based on combining the test statistics directly by using weighted-sum statistics. We start with this latter family of tests and then describe Fisher's combination method.

Model and Notations

We assume that n subjects are sequenced in a region (e.g., a gene) that has m variants: m_1 rare variants and m_2 common variants ($m = m_1 + m_2$). Let X be the $n \times m$ genotype matrix. We consider regression model

$$g[E(Y_i)] = \alpha_0 + \mathrm{C}_i\alpha + \mathrm{X}_i\beta, \qquad \text{(Equation 1)}$$

where $g(\cdot)$ is a link function and can be set to be the identity function when traits are continuous or the logistic function when traits are dichotomous; $\alpha = (\alpha_1, \ldots, \alpha_p)'$ are regression coefficients for the covariates, $\mathrm{C}_i = (C_{i1}, \ldots, C_{ip})$, that we want to adjust for. $\mathrm{X}_i = (X_{i1}, \ldots, X_{im})$ is the vector of genotypes for the i^{th} individual, and Y_i is its trait value. $\beta = (\beta_1, \ldots, \beta_m)'$ are regression coefficients for the m genetic variants. We assume that β is a random variable with $E(\beta_j) = 0$, $\mathrm{Var}(\beta_j) = w_j^2\tau$, and $\mathrm{corr}(\beta_j, \beta_k) = \rho$ for different j and k. For testing the null hypothesis of no genetic effects,

$$H_0 : \beta = 0,$$

the variance-component score statistic has been proposed as[21,22]

$$Q_\rho = \left(\mathrm{Y} - \widehat{\mu_0}\right)' \mathrm{K}_\rho \left(\mathrm{Y} - \widehat{\mu_0}\right),$$

where

$$\mathrm{K}_\rho = \mathrm{XWR}_\rho\mathrm{WX}',$$

in which $R_\rho = (1-\rho)I + \rho 11'$ specifies an exchangeable correlation matrix and $W = \mathrm{diag}(w_1, \ldots, w_m)$ is a diagonal weight matrix; for a dichotomous trait, $\widehat{\mu_0}$ is a vector of estimated probabilities of Y under the null model. Although this class of tests is more general, we restrict attention to two commonly used tests, the burden test (when $\rho = 1$) and the SKAT test (when $\rho = 0$). These score statistics are easily computed and can be written simply as

$$\text{SKAT}: \quad Q_{\rho=0} = \sum_{j=1}^{m} w_j^2 \left[\sum_{i=1}^{n} \left(Y_i - \widehat{\mu_{i,0}} \right) X_{ij} \right]^2 \text{ and}$$

$$\text{Burden}: \quad Q_{\rho=1} = \left[\sum_{j=1}^{m} w_j \sum_{i=1}^{n} \left(Y_i - \widehat{\mu_{i,0}} \right) X_{ij} \right]^2.$$

A weighting scheme that upweights rare variants and downweights common variants has been proposed for testing for rare-variant effects: $w_j = Beta(\widehat{p_j}, 1, 25)$, where $\widehat{p_j}$ is the MAF estimated on the basis of all subjects for variant j. The null distribution of Q_ρ is approximated by a mixture of χ_1^2 distributions. Davies' method[26] or moment matching can be employed for calculating the p value.

Here, we propose several tests that explicitly separate rare and common variants. Let X_1 be the $n \times m_1$ genotype matrix of rare variants and X_2 be the $n \times m_2$ genotype matrix of common variants. First, we rewrite the regression model in Equation 1 as

$$g[E(Y_i)] = \alpha_0 + C_i\alpha + X_{1i}\beta_1 + X_{2i}\beta_2, \qquad \text{(Equation 2)}$$

where X_{1i} is the genotype vector of rare variants and X_{2i} is the genotype vector of common variants for the i^{th} individual. $\beta_1 = (\beta_{11}, \ldots, \beta_{1m_1})'$ and $\beta_2 = (\beta_{21}, \ldots, \beta_{2m_2})'$ are coefficient vectors for rare and common variants, respectively. The null hypothesis of no genetic effects in the region corresponds to

$$H_0 : \beta_1 = 0, \beta_2 = 0.$$

Combined Sum Test of Rare- and Common-Variant Effects

In order to test for the joint effect of rare and common variants in a region, we combine score test statistics for rare and common variants as a weighted sum. Suppose that β_{1j} is a random variable with $E(\beta_{1j}) = 0$, $\mathrm{Var}(\beta_{1j}) = w_{1j}^2(1-\phi)\tau$, and $\mathrm{corr}(\beta_{1j}, \beta_{1k}) = \rho_1$ for different j and k. Similarly, we assume that β_{2j} is a random variable with $E(\beta_{2j}) = 0$, $\mathrm{Var}(\beta_{2j}) = w_{2j}^2\phi\tau$, and $\mathrm{corr}(\beta_{2j}, \beta_{2k}) = \rho_2$. The null hypothesis of $\beta_1 = \beta_2 = 0$ is equivalent to $\tau = 0$. A score test statistic with given (ϕ, ρ_1, ρ_2) is

$$\begin{aligned} Q_{\phi,\rho_1,\rho_2} &= (1-\phi)(Y - \widehat{\mu}_0)' K_{1,\rho_1} (Y - \widehat{\mu}_0) + \phi (Y - \widehat{\mu}_0)' K_{2,\rho_2} (Y - \widehat{\mu}_0) \\ &= (1-\phi) Q_{\text{rare}} + \phi Q_{\text{common}}, \end{aligned}$$

which is a weighted sum of rare- and common-variant test statistics and has weight parameter ϕ, where $K_{1,\rho_1} = X_1 W_1 R_{\rho_1} W_1 X_1'$ and $K_{2,\rho_2} = X_2 W_2 R_{\rho_2} W_2 X_2'$. A simple approach is to select ϕ such that the rare and common variants contribute equally to the test statistics. In particular, we choose $\phi = \mathrm{SD}[Q_{\text{rare}}]/(\mathrm{SD}[Q_{\text{rare}}] + \mathrm{SD}[Q_{\text{common}}])$ so that $(1-\phi)Q_{\text{rare}}$ and ϕQ_{common} have the same variance. The two weight matrices for rare and common variants are general and can accommodate a large family of possible weights. In this paper, we use different weight functions for rare and common variants. In particular, for rare variants we use the same weights as proposed in the original SKAT tests, i.e., $w_{1j} = Beta(\widehat{p_j}, 1, 25)$. However, for common variants, this weighting scheme does not work because it assigns almost zero weight to common variants (e.g., $w = 0.0004$ for a MAF of 0.30 but $w = 7.28$ for a MAF of 0.05). Instead, for common variants, we use $w_{2j} = Beta(\widehat{p_j}, 0.5, 0.5)$,[12] which slowly decreases with increasing MAF. For example, for MAF = 0.05, $w = 1.46$, for MAF = 0.10, $w = 1.06$, for MAF = 0.30, $w = 0.69$, and for MAF = 0.5, $w = 0.64$.

Table 1. Sequence-Based Association Tests: Existing Tests and the Proposed RC-SKAT Tests

Method	Name	Description
Burden	B	original burden test
SKAT	S	original SKAT test
CMC	C	combined multivariate collapsing test
SKAT-NW	S_{NW}	original SKAT test with no variant weighting
Burden-C	B_C	combined sum test with burden tests for rare and common variants
SKAT-C	S_C	combined sum test with SKAT tests for rare and common variants
Burden-A	B_A	adaptive sum test with burden tests for rare and common variants (rare variants are projected over the common variants)
SKAT-A	S_A	adaptive sum test with SKAT tests for rare and common variants (rare variants are projected over the common variants)
Burden-F	B_F	Fisher's method with burden tests for rare and common variants
SKAT-F	S_F	Fisher's method with SKAT tests for rare and common variants

For a given ϕ, the null distribution of Q_{ϕ,ρ_1,ρ_2} is a mixture of χ_1^2 distributions, and a p value can be obtained efficiently as follows. Let $(\lambda_1, \ldots, \lambda_m)$ denote the eigenvalues of the following matrix,

$$(1-\phi)P^{1/2} K_{1,\rho_1} P^{1/2} + \phi P^{1/2} K_{2,\rho_2} P^{1/2},$$

where $P = \widehat{V}^{-1} - \widehat{V}^{-1}\tilde{C}(\tilde{C}'\widehat{V}^{-1}\tilde{C})^{-1}\tilde{C}'\widehat{V}^{-1}$ is an $n \times n$ matrix, $\tilde{C}$ is an $n \times (p+1)$ matrix equal to $[\underline{1}\ C]$, and $\widehat{V}$ is a diagonal matrix of the variance of Y under the null hypothesis. It has been shown that Q_{ϕ,ρ_1,ρ_2} follows a mixture of χ_1^2 distributions,

$$\sum_{j=1}^{m} \lambda_j \chi_{1,j}^2,$$

where $\chi_{1,j}^2$ are independent and identically distributed (i.i.d.) chi-square random variables with df = 1.[27] Asymptotic p values can be computed with Davies' method or moment matching.[18,22] We refer to these tests as burden-C if $\rho_1 = \rho_2 = 1$ and SKAT-C if $\rho_1 = \rho_2 = 0$ (see Table 1).

Adaptive Sum Test of Rare- and Common-Variant Effects

Above, we have chosen the weight parameter ϕ such that rare and common variants contribute equally to the overall test statistic. An alternative approach is to compute p values for varying values of ϕ and use the minimum p value as a test statistic. This approach can be potentially more powerful if the overall effect sizes of rare and common variants are very different, for example, when only rare variants in the region are associated or when only common variants are associated with a trait. However, for this type of adaptive test, asymptotic p values cannot be obtained easily because of

the potential correlation that exists between rare and common variants.

Here, we propose the following adaptive approach instead. First, we linearly transform Equation 2 via projection. The transformed model is

$$g[E(Y_i)] = \alpha_0 + C_i\alpha + \tilde{X}_{1i}\gamma_1 + X_{2i}\gamma_2, \qquad \text{(Equation 3)}$$

where $\tilde{X}_1 = (\tilde{X}'_{11}, \ldots, \tilde{X}'_{1n})' = (I - M)X_1$ and $M = X_2(X_2'X_2)^{-1}X_2'$ is an $n \times n$ projection matrix onto the column space of X_2; $\gamma_1 = (\gamma_{11}, \ldots, \gamma_{1m_1})'$ and $\gamma_2 = (\gamma_{21}, \ldots, \gamma_{2m_2})'$ are regression coefficients of the transformed model. Note that $\tilde{X}_1$ corresponds to the residuals by performing a linear regression of each component of X_{1i} on X_{2i}. We assume that γ_{1j} is a random variable with $E(\gamma_{1j}) = 0$, $\mathrm{Var}(\gamma_{1j}) = w_{1j}^2(1-\phi)\tau$, and $\mathrm{corr}(\gamma_{1j}, \gamma_{1k}) = \rho_1$ for different j and k. Similarly, γ_{2j} is a random variable with $E(\gamma_{2j}) = 0$, $\mathrm{Var}(\gamma_{2j}) = w_{2j}^2\phi\tau$, and $\mathrm{corr}(\gamma_{2j}, \gamma_{2k}) = \rho_2$. The null hypothesis of $\beta_1 = \beta_2 = 0$ in the original Equation 2 is identical to $\tau = 0$ in the transformed Equation 3. A score test statistic with given (ϕ, ρ_1, ρ_2) is

$$\begin{aligned} Q^*_{\phi,\rho_1,\rho_2} &= (1-\phi)(Y - \hat{\mu}_0)'\tilde{K}_{1,\rho_1}(Y - \hat{\mu}_0) + \phi(Y - \hat{\mu}_0)'K_{2,\rho_2}(Y - \hat{\mu}_0) \\ &= (1-\phi)Q^*_{\mathrm{rare}} + \phi Q^*_{\mathrm{common}}, \end{aligned}$$

where $\tilde{K}_{1,\rho_1} = (I - M)X_1W_1R_{\rho_1}W_1X_1'(I - M)$ and $K_{2,\rho_2} = X_2W_2R_{\rho_2}W_2X_2'$.

We propose the adaptive test,

$$T = \min_{0 \le \phi \le 1} p_\phi(\rho_1, \rho_2),$$

where $p_\phi(\rho_1, \rho_2)$ is the p value for Q^*_{ϕ,ρ_1,ρ_2}. Test statistic T can be obtained by a simple grid search, $0 = \phi_1 < \ldots < \phi_b = 1$. In simulation studies and real data analysis, we used a grid of five values (0, 0.25, 0.50, 0.75, and 1). Because both Q^*_{rare} and Q^*_{common} follow a mixture of chi-square distributions and they are independent, the null distribution of T can be easily obtained (see Appendix A). We refer to these tests as burden-A if $\rho_1 = \rho_2 = 1$ and SKAT-A if $\rho_1 = \rho_2 = 0$ (see Table 1).

In this approach, the rare variants are projected on the common variants, a procedure that is similar to the practice of including GWAS signals as covariates in order to test whether there is any effect that rare variants contribute beyond the common variant effects. An alternative approach would be to project common variants on the rare variants. We evaluate both of these tests in our simulation studies.

Fisher's Combination Method

An alternative approach is to combine the p values from the rare- and common-variant tests instead of combining test statistics. Let p_{rare} and p_{common} be the corresponding p values from the tests with rare variants only and common variants only, respectively. Then, we consider the following test statistic:

$$Q_{F,\rho_1,\rho_2} = -2\log_e(p_{\mathrm{rare}}) - 2\log_e(p_{\mathrm{common}}).$$

Under the null hypothesis, both $-2\log_e(p_{\mathrm{rare}})$ and $-2\log_e(p_{\mathrm{common}})$ are distributed as chi-square variables with 2 df. Fisher's combination method assumes that the statistics to be combined are independent, and in that case, the distribution of Q_{F,ρ_1,ρ_2} is a chi-square with 4 df. However, because the rare- and common-variant statistics might be correlated, the distribution of Q_{F,ρ_1,ρ_2} is more complicated. According to Brown,[28] it can be approximated by a weighted chi-square distribution, $c\chi^2_f$, and a p value can be calculated by moment matching. More precisely, we have

$$\begin{aligned} &E(Q_{F,\rho_1,\rho_2}) = 4 \text{ and} \\ &\mathrm{Var}(Q_{F,\rho_1,\rho_2}) = 4 + 2\mathrm{cov}(-2\log_e(p_{\mathrm{rare}}), -2\log_e(p_{\mathrm{common}})), \end{aligned}$$

where the covariance between $-2\log_e(p_{\mathrm{rare}})$ and $-2\log_e(p_{\mathrm{common}})$ is approximated by quadratic functions of the correlation between the rare- and common-variant statistics, denoted by r. More precisely, as in Brown,[28]

$$\mathrm{cov}(-2\log_e(p_{\mathrm{rare}}), -2\log_e(p_{\mathrm{common}})) = \begin{cases} r(3.25 + 0.75r) & 0 \le r \le 1 \\ r(3.27 + 0.71r) & -0.5 \le r \le 0 \end{cases}$$

Although this result for the covariance assumes a joint multivariate normal density for the variables, in our applications, this approximation worked well for our situation (as a result of the small correlation that exists between rare- and common-variant statistics). The correlation (r) between the two quadratic forms (for rare and common variants) can be calculated analytically (see Appendix A). We approximate the distribution of Q_{F,ρ_1,ρ_2} by using $c\chi^2_f$, in which we estimate c and f by matching the first two moments. This way, we obtain

$$f = \frac{2E(Q_{F,\rho_1,\rho_2})^2}{\mathrm{Var}(Q_{F,\rho_1,\rho_2})}$$

$$c = \frac{\mathrm{Var}(Q_{F,\rho_1,\rho_2})}{2E(Q_{F,\rho_1,\rho_2})}.$$

We refer to these tests as burden-F if $\rho_1 = \rho_2 = 1$ and SKAT-F if $\rho_1 = \rho_2 = 0$ (see Table 1).

Results

Simulated Data

We simulated sequence data on 10,000 haplotypes in one genomic region of length 1 Mb under a coalescent model by using the software package COSI.[29] The model used in the simulations was the calibrated model for the European population. For our purposes, we randomly sampled small subregions of size 5 or 25 kb and simulated data sets with n = 1,000–5,000 individuals (equal number of cases and controls).

We considered several disease models that involve a mixture of common and rare disease risk variants. In our simulated disease models (Table 2), we simulated rare risk variants from those variants with a MAF $\le$ 0.01 and common risk variants with MAF $>$ 0.01. For all models, we assumed that a small percentage of the variants in a region are associated with disease with effect sizes as described in Table 2. Models 1–5 assume that all disease-associated variants confer risk, whereas model 6 assumes a mixture of risk and protective variants. Because the number of common variants in a given region is generally much smaller than the number of rare variants, in order to increase the contribution from common variants, for model 2, we assumed that although only 10%–30% of rare variants are associated, 50% of the common variants are associated with disease.

Table 2. Six Disease Models

Model	Description
1	10%–30% of rare variants have an ORR = 2
	10%–30% of common variants have an ORC = 1.1
2	10%–30% of rare variants have an ORR= 2
	50% of common variants have an ORC = 1.1
3	10%–30% of all variants have an OR = $e^{0.2\lvert(\log_{10}(\mathrm{MAF}))\rvert}$
4	10%–30% of rare variants have an ORR = 2
	no common associated variants
5	10%–30% of common variants have an ORC = 1.2
	no rare associated variants
6	10%–30% of rare variants have an ORR = 2
	10%–30% of common variants have an ORC = 1.2
	30% of associated variants are protective (and have an ORR = 0.5 and an ORC = 0.84)

Abbreviations are as follows: ORR, OR for disease-associated variants with MAF $\leq$ 0.01; and ORC, OR for disease-associated variants with MAF $>$ 0.01.

For a dichotomous trait, we assumed the following logistic model:

$$\mathrm{logit}[P(Y_i = 1)] = \alpha_0 + \beta_1 X_{i1} + \ldots + \beta_m X_{im}.$$

α_0 was chosen such that the disease prevalence was 0.05. We compared the proposed combination tests with three of the most commonly used tests—burden, SKAT, and CMC[11]—as well as the SKAT test with no variant weighting (see Table 1 for a description of these tests).

Type 1 Error

To evaluate the type 1 error for the proposed combination (RC-SKAT) methods, we simulated data under the null model ($\mathrm{logit}[P(Y_i = 1)] = \alpha_0$) for n = 1,000–2,000 and regions of size 5 and 25 kb (we did not simulate the n = 5,000 scenario because of the prohibitive computational cost for very small significance values). Results based on 10^7 simulations are shown in Table 3. The type 1 error for all the proposed methods agrees well with the expectation for $\alpha = \{0.05, 0.01, 0.0001, 2.5 \times 10^{-6}\}$. Note that $\alpha = 2.5 \times 10^{-6}$ is an exome-wide significance level of 0.05 when 20,000 genes are simultaneously evaluated.

Power with Different Frequency Cutoffs

Because the proposed combination tests require an explicit partition of the genetic variants in a region into rare and common variants, we first evaluated the sensitivity of using several different separation cutoffs to the true disease model. We compared the power of the proposed methods under three disease models when using different separation cutoffs. The disease models are as follows. The first two are disease models 1 and 3 in Table 2. The third model is similar to model 1 in Table 2 but has the same odds ratio (OR) of 1.5 for all risk variants (hence, there is no difference in effect size between rare and common risk variants).

Furthermore, for each of these models, we considered four possible scenarios: (1) all risk variants have a MAF $<$ 0.005, (2) all risk variants have MAF $<$ 0.01, (3) all risk variants have MAF $<$ 0.05, and (4) all risk variants have MAF $>$ 0.05. For example, for model 1 and scenario (1) above, all

Table 3. Type 1 Error for the Proposed RC-SKAT Combination Tests

Length	*n*	α	Burden-A	SKAT-A	Burden-C	SKAT-C	Burden-F	SKAT-F
5 kb	1,000	5.0×10^{-2}	4.9×10^{-2}	3.9×10^{-2}	4.9×10^{-2}	4.6×10^{-2}	5.0×10^{-2}	4.9×10^{-2}
		1.0×10^{-2}	9.6×10^{-3}	7.1×10^{-3}	9.7×10^{-3}	8.8×10^{-3}	1.0×10^{-2}	1.0×10^{-2}
		1.0×10^{-4}	8.2×10^{-5}	6.1×10^{-5}	8.6×10^{-5}	7.4×10^{-5}	1.3×10^{-4}	1.3×10^{-4}
		2.5×10^{-6}	1.6×10^{-6}	2.2×10^{-6}	1.6×10^{-6}	1.9×10^{-6}	2.6×10^{-6}	3.0×10^{-6}
5 kb	2,000	5.0×10^{-2}	4.9×10^{-2}	4.3×10^{-2}	4.9×10^{-2}	4.7×10^{-2}	5.0×10^{-2}	4.9×10^{-2}
		1.0×10^{-2}	9.8×10^{-3}	8.2×10^{-3}	9.8×10^{-3}	9.2×10^{-3}	1.0×10^{-2}	9.9×10^{-3}
		1.0×10^{-4}	8.7×10^{-5}	8.6×10^{-5}	8.9×10^{-5}	8.2×10^{-5}	1.1×10^{-4}	1.2×10^{-4}
		2.5×10^{-6}	1.9×10^{-6}	2.2×10^{-6}	1.8×10^{-6}	1.4×10^{-6}	4.0×10^{-6}	3.4×10^{-6}
25 kb	1,000	5.0×10^{-2}	4.9×10^{-2}	3.7×10^{-2}	4.9×10^{-2}	4.6×10^{-2}	5.0×10^{-2}	4.9×10^{-2}
		1.0×10^{-2}	9.8×10^{-3}	6.6×10^{-3}	9.7×10^{-3}	8.6×10^{-3}	1.1×10^{-2}	9.9×10^{-3}
		1.0×10^{-4}	9.1×10^{-5}	6.6×10^{-5}	8.3×10^{-5}	6.9×10^{-5}	1.5×10^{-4}	1.3×10^{-4}
		2.5×10^{-6}	1.7×10^{-6}	2.1×10^{-6}	1.4×10^{-6}	1.7×10^{-6}	4.2×10^{-6}	1.9×10^{-6}
25 kb	2,000	5.0×10^{-2}	5.0×10^{-2}	4.1×10^{-2}	4.9×10^{-2}	4.7×10^{-2}	5.0×10^{-2}	4.8×10^{-2}
		1.0×10^{-2}	9.9×10^{-3}	7.8×10^{-3}	9.8×10^{-3}	9.2×10^{-3}	1.0×10^{-2}	9.9×10^{-3}
		1.0×10^{-4}	9.8×10^{-5}	8.9×10^{-5}	1.0×10^{-4}	8.6×10^{-5}	1.4×10^{-4}	1.2×10^{-4}
		2.5×10^{-6}	2.4×10^{-6}	3.2×10^{-6}	2.1×10^{-6}	2.0×10^{-6}	4.8×10^{-6}	3.9×10^{-6}

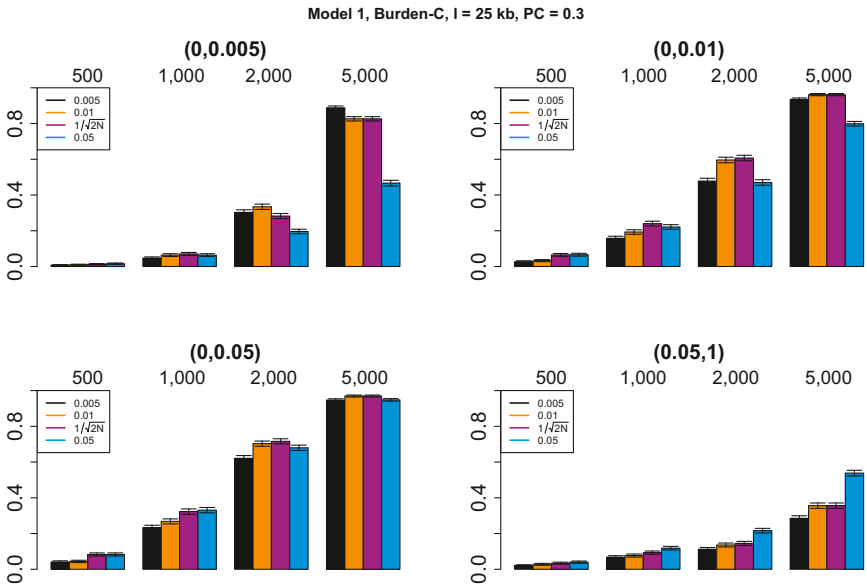

Figure 1. Power of the Burden-C Test with Different Frequency Thresholds to Separate Rare from Common Variants

Power ($\alpha = 2.5 \times 10^{-6}$) of using different frequency thresholds to separate rare and common variants (fixed values 0.005, 0.01, and 0.05 versus $1/\sqrt{2n}$) for the proposed combination method burden-C for model 1 (in Table S1) and for $n = 500$, 1,000, 2,000, and 5,000 in a region of size 25 kb. The proportion of associated variants (PC) in the region is 30%. The sample-dependent threshold $1/\sqrt{2n}$ is: 0.03 ($n = 500$), 0.02 ($n = 1{,}000$), 0.015 ($n = 2{,}000$), and 0.01 ($n = 5{,}000$).

risk variants have a MAF < 0.005 and an OR $= 2$. A summary of these models is given in Table S1, available online. We compared the power of the proposed combination methods (e.g., burden-C and SKAT-C) when using conventional cutoffs (such as 0.005, 0.01, and 0.05) to separate rare and common variants versus the sample-dependent cutoff of $1/\sqrt{2n}$. All power calculations are based on 1,000 simulated data sets for each scenario.

Results for the first model, with a 25 kb genetic region, and the burden-C test are shown in Figure 1 (the results for the SKAT-C test are in Figure S1; additional results, including other simulation scenarios, are given in the Supplemental Data). The optimal cutoff ultimately depends both on sample size (n) and on the true disease model (i.e., the joint distribution of true risk allele frequencies and effect sizes, which is unknown to us). When sample sizes are large (e.g., $n = 5{,}000$), the optimal threshold correlates well with the underlying disease model. For example, for model 1, true risk variants have an OR $= 2$ if their MAF ≤ 0.01 and an OR $= 1.1$ if their MAF > 0.01. For $n = 5{,}000$, if the risk variants have a MAF ≤ 0.005 or 0.01, the best separation cutoffs are 0.005 and $0.01 = 1/\sqrt{2n}$; using 0.05 as a cutoff causes a significant loss in power (Figures 1 and S1). Conversely, when all risk variants have a MAF > 0.05, a cutoff of 0.05 tends to perform the best. In spite of this expected dependence of the optimal separation threshold on the true disease model when sample sizes are large, the sample-dependent threshold tends to perform consistently well across the simulated scenarios. When sample sizes are small (e.g., $n = 500$), the dependence of the best cutoff on the true disease model is less clear, but because of the increased variance in the observed frequencies of risk alleles, a sample-dependent cutoff as proposed here can be considered. In what follows, for the sake of fixation, we will report power for these combination tests by using only the $1/\sqrt{2n}$ threshold.

Comparison of Power across Different Tests

In Figures 2 and 3, we report the power of the proposed RC-SKAT methods and of the existing methods (burden test, SKAT, and CMC) for the disease models in Table 2 and genetic regions of size 25 kb. We also compare with the power of the original SKAT test, but without variant weighting. Overall, for all the models that include common associated variants (models 1–3, 5, and 6), the combination methods outperform existing methods, oftentimes in a substantial way. Even when all disease-associated variants are rare (model 4), the proposed combination methods perform similarly to the existing methods, suggesting that applying the proposed methods in that case causes no or little efficiency loss (Figure 3A). However, when all risk variants are common (model 5), the RC-SKAT approaches outperform the existing methods substantially (Figure 3B). The proposed combination methods outperform CMC across all six models. The same is true when they are compared with the original SKAT test with no variant weighting.

For models 1–4, which include only risk variants, SKAT tests tend to be more powerful than the corresponding burden tests when the proportion of risk variants is small (e.g., 10%); however, burden tests become more powerful than SKAT tests when the proportion of risk variants is large (e.g., 30% or more). Note that for models 1–3, which include both rare and common risk variants, SKAT-F (Fisher) test tends to have better power than the burden-F test regardless of whether the proportion of associated variants is 10% or 30%. For model 5, which includes only common risk variants, the SKAT tests tend to perform better than the burden tests regardless of the proportion of risk variants (probably as a result of the fact that only a small proportion of variants in a given region are common). The same holds true for model 6, which includes both risk and protective variants (Figure 3C). Note that CMC also performs better than existing burden and SKAT tests for this model. Although the CAST rare-variant statistic employed by CMC loses power when there is a mixture of risk and protective variants in the region, the

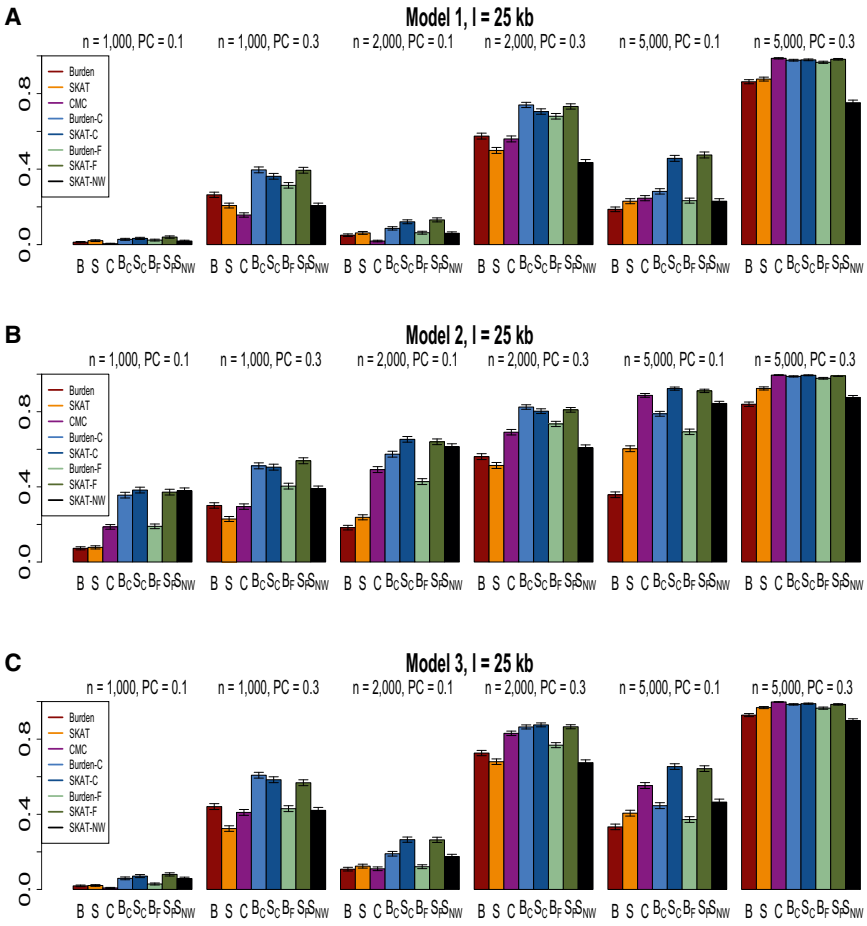

Figure 2. Power for Models 1–3 in 25 kb Regions

Power ($\alpha = 2.5 \times 10^{-6}$) of the tests in Table 1 for a region of size 25 kb (l = 25 kb) across disease models 1–3 in Table 2 for n = 1,000, 2,000, and 5,000 and two different values for the proportion of associated variants in a region: 10% (i.e., PC = 0.1) or 30% (i.e., PC = 0.3).

Hotelling's statistic (which underlies the CMC test) for rare and common variants performs well in such settings. The original SKAT test without variant weighting tends to perform worse than the proposed combination tests, with the exception of model 5, which only includes common risk variants. However, compared with the proposed tests, SKAT without weighting suffers substantial loss of power for the remaining five models, and especially for model 4, which has only rare risk variants.

The adaptive tests (burden-A and SKAT-A, obtained by either the projection of rare variants over common variants or the other way around) tend to perform worse than the burden-C and SKAT-C tests (see Figures S2 and S3).

Results for a region of size 5 kb were similar and are shown in Figures S6 and S7.

Crohn Disease *NOD2* Sequence Data

We applied our RC-SKAT methods to sequencing data for *NOD2* (MIM 605956) from 453 Crohn disease cases and 103 healthy controls.[30] In total, 60 single-nucleotide variations, nine of which have a frequency greater than 5%, have been identified (in exons and all of the intron-exon junctions). Because only pooled frequency counts were available for each variant, we generated simulated sequencing data for 453 cases and 103 controls, consistent with the observed counts, and we assumed the correlation structure between the common variants as in HapMap 3 (CEU [Utah residents with ancestry from northern and western Europe from the CEPH collection] population). Rare variants were assumed to be independent of each other and the common variants.

We first performed a common-variant analysis by using the trend test in PLINK;[31] six of the nine common variants were found to be associated ($p < 0.05$), and the smallest p value was 1.1×10^{-4} (Table 4). We then performed gene-based tests. We applied three of the commonly used tests for sequence data (burden, SKAT, and CMC), the burden and SKAT tests without variant weighting, and the new RC-SKAT tests to this case-control data set. We used different frequency thresholds, including the proposed $1/\sqrt{2n} = 0.029$, 0.01, and 0.05, to separate the variants into rare and common. The results are given in Table 5. As shown, several of our combination tests resulted in exome-wide-significant p values (e.g., the p value for SKAT-C was 1.7×10^{-7} when the threshold was $1/\sqrt{2n}$), as did CMC (p value of 1.7×10^{-7}). Both the original burden and SKAT tests produced only modest p values ($>5.0 \times 10^{-4}$). Similarly, the original burden and SKAT tests with no variant weighting resulted in p values $> 1.0 \times 10^{-5}$. Because it is known that both common and rare variants in *NOD2* independently affect disease risk[2] (see also Table 4), it is not surprising that combination tests such as those discussed here perform better than existing tests, such as the burden and SKAT tests, which focus primarily on detecting rare risk variants.

Autism *LRP2* Sequence Data

We then applied the proposed methods to a second sequencing data set for ASDs. *LRP2* (MIM 600073) is a gene that resides in a region linked to ASD on chromosome 2q. Recently, on the basis of three independent data sets, Ionita-Laza et al.[32] have found evidence that rare variants associated with ASD cluster in a small region of this gene. Moreover, three publications focused on de novo mutations have reported additional supporting evidence for the role of this gene in ASD and intellectual

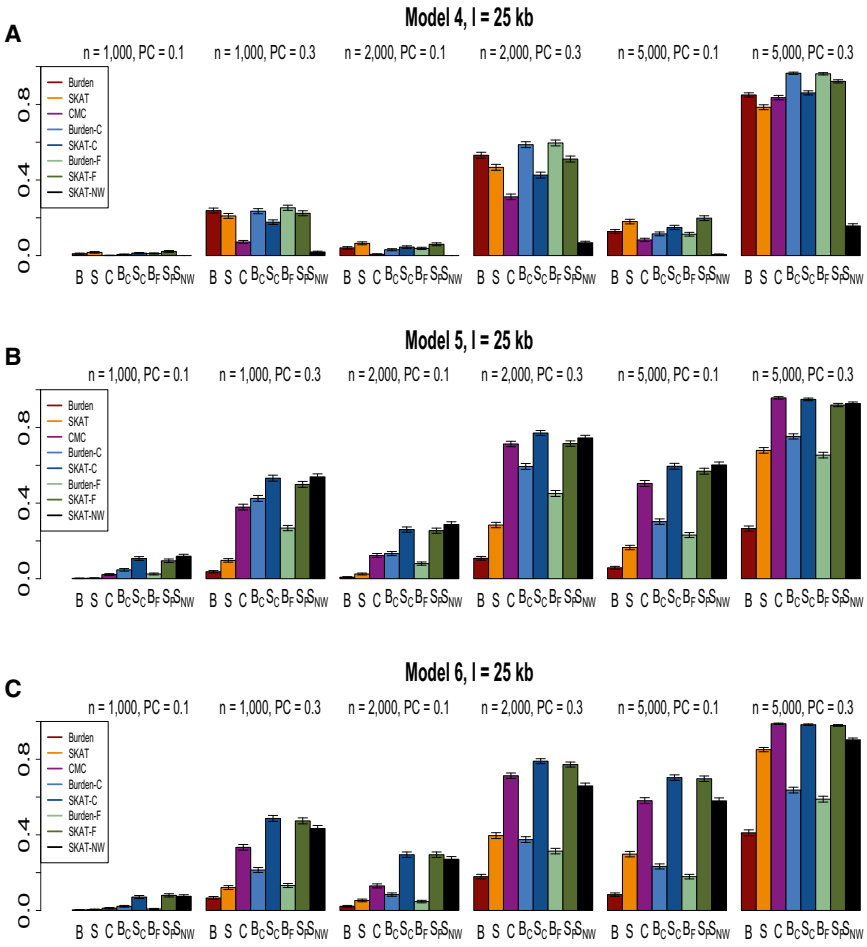

Figure 3. Power for Models 4–6 in 25 kb Regions

Power ($\alpha = 2.5 \times 10^{-6}$) of the tests in Table 1 for a region of size 25 kb (l = 25 kb) across disease models 4–6 in Table 2 for n = 1,000, 2,000, and 5,000 and two different values for the proportion of associated variants in a region: 10% (i.e., PC = 0.1) or 30% (i.e., PC = 0.3).

disability.[33–35] In Ionita-Laza et al.,[32] we observed that, in addition to the cluster of rare disease-associated variants, common variants are also associated with ASD. Hence, for *LRP2*, there is evidence from different studies for the contribution of de novo, rare, and common variants to autism.

We applied the RC-SKAT tests in Table 1 to a data set consisting of 430 cases and 379 controls sequenced in the exonic regions of *LRP2* (more details about this data set are reported in Appendix A). The results of several existing tests and the proposed RC-SKAT tests are given in Table 5. As shown, the existing tests, including burden, SKAT and CMC, resulted in marginally nonsignificant p values for this gene, suggesting no significant rare-variant effects. The proposed combination tests resulted in several significant p values, given that the signals in this gene mainly come from common variants. For example, the p value for burden-C was 6.1×10^{-4} when $T = 1/\sqrt{2n} = 0.024$, whereas for SKAT-C, the corresponding p value was 2.3×10^{-2}. Because the common *LRP2* variants associated with ASD tend to have similar effect sizes and the same direction of effect, it is expected that the burden tests perform better than the SKAT tests. The original burden and SKAT tests without weighting performed similarly to the proposed combination tests for this gene.

Discussion

We have proposed sequence kernel association tests that test for the contribution of both rare and common genetic variants to risk of complex diseases. Unlike most existing tests that upweight the contribution of rare variants and downweight the contribution of common variants, the proposed tests assess the contribution of rare and common variants separately and then combine the corresponding test statistics or the p values by using either an equal weight or an adaptive weight. As with the existing burden and SKAT tests, it is easy to incorporate covariates, including principal components, to adjust for population stratification.

The proposed RC-SKAT tests are based on first partitioning all the variants into rare and common variants. Such partitioning is based on a frequency threshold. Usually, this threshold is chosen as a fixed value, e.g., 0.01 or 0.05. We have suggested here another possible threshold that depends on the sample size, namely, $1/\sqrt{2n}$. Therefore, when the sample size is small, the separating threshold will be higher than for a larger sample size. Although this threshold is clearly not optimal in that it only depends on sample size and not on the underlying effect-size distribution, it can serve as a lower bound on the possible cutoffs to be considered. This ensures that variants that occur only a few times in a modest data set are not classified as common (which can lead to loss of power). Furthermore, this threshold can be used for identifying additional "common" variants (e.g., those variants that have a frequency below 0.01 and that are traditionally defined as rare) that can be tested individually for their effect on the trait. In practice, we suggest using this sample-size-dependent threshold, along with several larger fixed thresholds. Although it is possible to adaptively select the threshold (as in Price et al.[13]), such an approach requires permutations that are computationally intensive and might be problematic when there are covariates to be adjusted for.

The idea of separating variants into rare and common has been proposed before, most notably in the CMC test.

Table 4. p Values from the Cochran-Armitage Trend Test for the Common *NOD2* Variants Significantly Associated with Crohn Disease

NOD2 Variant	f_A	f_U	p	OR
c.3020insC	0.10	0.02	0.0001	5.98
c.802C>T	0.41	0.27	0.0008	1.84
c.1377C>T	0.41	0.28	0.0016	1.79
c.2722G>C	0.06	0.01	0.0035	6.59
c.2104C>T	0.10	0.04	0.0074	2.65
c.33G>T in 5′ UTR	0.41	0.33	0.033	1.46

Significance is defined as $p < 0.05$. Abbreviations are as follows: f_A, estimated frequency of minor allele in cases; f_U, estimated frequency of minor allele in controls; and OR, estimated odds ratio.

We have compared the proposed combination tests with CMC and have shown that the proposed tests tend to have higher power than CMC uniformly across the simulated scenarios. CMC loses power especially when the regions tested are large (e.g., 25 kb) because of the increased degrees of freedom for Hotelling's T^2 and as such resulted in a nonsignificant p value for *LRP2* (a large gene with length ~200 kb). More recently, Cardin et al.[36] have proposed a hierarchical modeling approach for joint association between rare and common variants and binary traits. This approach is based on averaging over prior distributions of effect sizes, which are unknown to us, and therefore performance might be sensitive to such assumptions. Furthermore, the approach is computationally intensive, which makes it difficult to evaluate in large-scale power simulations of the type performed here. We have, however, considered several additional combination tests, including a test based on the minimum of p values from a burden or SKAT rare-variant test and individual p values for common variants (more details are in Appendix A). However, none of these more straightforward alternatives resulted in better power than the new tests we have proposed here (Figures S4 and S5).

Two of the proposed combination methods (burden-A and SKAT-A) are based on the adaptive approach that selects the optimal weight for combining common- and rare-variant test statistics. To compute p values analytically, it relies on the projection of the common-variant information out from the rare-variant information. Although the projection allows fast computation, it can weaken the association signal. In our simulation experiments, these approaches tended to be less powerful than the other combination methods. In practice, we suggest using burden-C and SKAT-C or alternatively burden-F and SKAT-F, similar to the suggestion made by others[37,38] of using both a linear (i.e., burden) test and a quadratic (i.e., SKAT) test in the context of association testing with rare variants.

The RC-SKAT tests are applicable to population-based designs with binary or quantitative traits. Both burden tests and SKAT tests have been recently extended to

Table 5. Sequence-Based Association Test Results for *NOD2* and *LRP2*

Gene	Cutoff	p Value by Method									
		Burden	SKAT	CMC	Burden-A	SKAT-A	Burden-C	SKAT-C	Burden-F	SKAT-F	Burden-NW	SKAT-NW
NOD2	0.01	1.06×10^{-3}	5.17×10^{-4}	2.99×10^{-7}	3.60×10^{-6}	1.87×10^{-7}	2.41×10^{-6}	2.05×10^{-7}	1.20×10^{-6}	2.50×10^{-7}	1.1×10^{-5}	2.0×10^{-5}
	$1/\sqrt{2n}$	1.06×10^{-3}	5.17×10^{-4}	1.68×10^{-7}	1.79×10^{-5}	2.22×10^{-7}	4.40×10^{-6}	1.70×10^{-7}	3.40×10^{-6}	4.60×10^{-8}	1.1×10^{-5}	2.0×10^{-5}
	0.05	1.06×10^{-3}	5.17×10^{-4}	1.68×10^{-7}	1.79×10^{-5}	2.22×10^{-7}	4.40×10^{-6}	1.70×10^{-7}	4.70×10^{-6}	8.20×10^{-8}	1.1×10^{-5}	2.0×10^{-5}
LRP2	0.01	2.78×10^{-1}	4.79×10^{-1}	3.50×10^{-1}	2.56×10^{-3}	2.64×10^{-2}	4.59×10^{-3}	3.79×10^{-2}	7.90×10^{-4}	1.10×10^{-2}	1.3×10^{-4}	5.0×10^{-3}
	$1/\sqrt{2n}$	2.78×10^{-1}	4.79×10^{-1}	2.12×10^{-1}	2.01×10^{-4}	1.99×10^{-2}	6.13×10^{-4}	2.36×10^{-2}	3.60×10^{-4}	9.30×10^{-3}	1.3×10^{-4}	5.0×10^{-3}
	0.05	2.78×10^{-1}	4.79×10^{-1}	8.21×10^{-2}	5.65×10^{-4}	1.66×10^{-2}	9.98×10^{-4}	1.90×10^{-2}	4.60×10^{-4}	1.20×10^{-2}	1.3×10^{-4}	5.0×10^{-3}

family-based designs,[39] and hence, the proposed combination tests are also applicable to family-based designs. Such family-based tests are transmission-disequilibrium types of tests, and are hence robust to population stratification.

WES is only able to survey variation in coding regions, thereby missing a lot of the variation in noncoding regions. It is not uncommon for a study to contain both GWAS data and WES data on the same individuals. The proposed methods can take advantage of the common variation on the GWAS arrays and combine with WES data to increase the power to identify genes containing variation associated with complex diseases. Furthermore, the proposed tests are applicable to deep-resequencing data for GWAS loci.

In summary, we have proposed tests for both rare- and common-variant effects and have shown that they are more powerful than existing groupwise association tests under a wide range of scenarios. The proposed tests are implemented in the software package SKAT.

Appendix A

Variance Explained by Common versus Rare Risk Variants

Although the relative contribution of rare and common variants to risk of complex diseases is unknown for most complex traits, it is expected that both common and rare variants are important. Even though rare variants are more likely to be functional and are expected to have higher effects than common variants, common variants can account for a substantial proportion of the genetic variance. In commonly used simulation experiments with 25 kb genetic regions that include 10% rare and common disease risk variants (Table A1), the genetic variance explained by common variants can be higher than that explained by rare variants (Figure A1).

Asymptotic Null Distribution of the Adaptive Sum Test of Rare- and Common-Variant Effects

It is easy to show that

$$Q^*_{\text{rare}} = \kappa_1 \sim \sum_{q=1}^{m_1} \lambda_{1q}\chi^2_1 \text{ and}$$
$$Q^*_{\text{common}} = \kappa_2 \sim \sum_{q=1}^{m_2} \lambda_{2q}\chi^2_1,$$

where $\{\lambda_{1q}\}$ are eigenvalues of $P^{1/2}K_{1,\rho_1}P^{1/2}$ and $\{\lambda_{2q}\}$ are eigenvalues of $P^{1/2}\tilde{K}_{2,\rho_2}P^{1/2}$, in which $P = \hat{V}^{-1} - \hat{V}^{-1}\tilde{C}(\tilde{C}'\hat{V}^{-1}\tilde{C})^{-1}\tilde{C}'\hat{V}^{-1}$, where $\tilde{C}$ is an $n\times(q+1)$ matrix equal to $[\underline{1}\ C]$ and $\hat{V}$ is a diagonal matrix of the variance of Y under the null hypothesis. We approximate the mixture of chi-square distributions by using moment matching. If we let $q(\phi_k)$ be the $(1-T)^{\text{th}}$ percentile of the distribution of $Q^*_{\phi_k,\rho_1,\rho_2}$ for each ϕ_k in our grid, then the p value of T can be calculated as

Table A1. Simulation Models for Investigating the Variance Explained by Rare versus Common Variants

Model	Description
1	10% of rare variants have an ORR = 2
	10% of common variants have an ORC = 1.1
2	10% of rare variants have an ORR= 2
	10% of common variants have an ORC = 1.2
3	10% of all variants have an OR = $e^{0.2\lvert(\log_{10}(\text{MAF}))\rvert}$

$$\begin{aligned} p &= 1 - P\left[Q^*_{\phi_1,\rho_1,\rho_2} < q(\phi_1), \ldots, Q^*_{\phi_b,\rho_1,\rho_2} < q(\phi_b)\right] \\ &= 1 - P[(1-\phi_1)\kappa_1 + \phi_1\kappa_2 < q(\phi_1), \ldots, (1-\phi_b)\kappa_1 \\ &\quad + \phi_b\kappa_2 < q(\phi_b)] \\ &= 1 - P\left[\kappa_1 < \min_{0\le\phi\le1} \frac{q(\phi) - \phi\kappa_2}{1-\phi}\right] \\ &= 1 - E\left[P\left(\kappa_1 < \min_{0\le\phi\le1} \frac{q(\phi) - \phi\kappa_2}{1-\phi} \mid \kappa_2\right)\right]. \end{aligned}$$

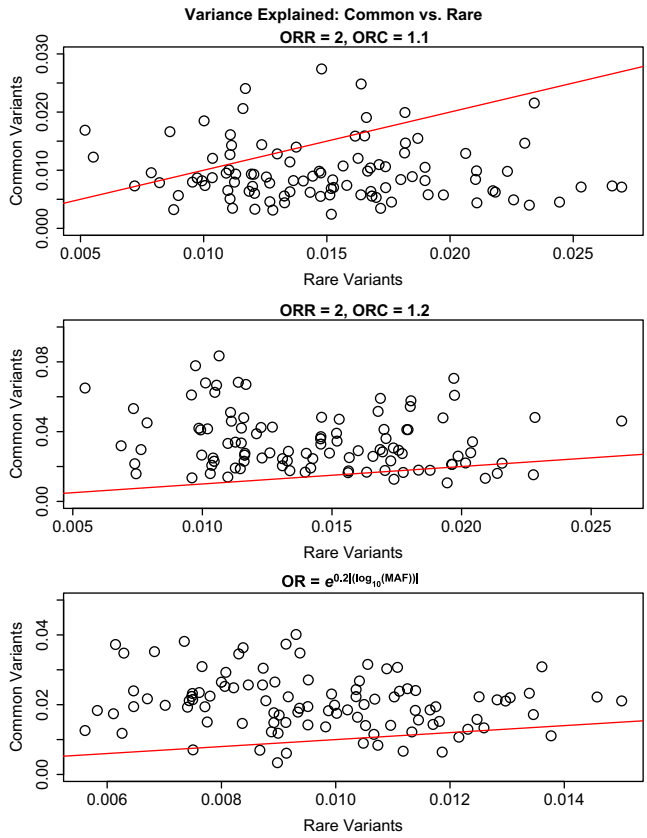

Figure A1. Variance Explained by Rare versus Common Risk-Associated Variants

Variance explained by common associated variants versus rare associated variants in a 25 kb region (based on 100 random regions). Ten percent of all variants in the region are associated with the trait. Each dot corresponds to one random 25 kb region. Variance explained by common variants (MAF $> 1/\sqrt{2n}$) is on the y axis, and variance explained by rare variants (MAF $< 1/\sqrt{2n}$) is on the x axis.

This can be obtained by computationally efficient one-dimensional integration.

Correlation between Two Quadratic Forms

We use the following relations to calculate the correlation between two quadratic forms. Let us assume that $E[Y] = 0$. Let $\Sigma = \mathrm{Var}[Y]$. Let A and B be real and symmetric $n \times n$ matrices. Then

$$E[Y'AY] = \mathrm{tr}(A\Sigma) \text{ and}$$
$$\mathrm{Cov}[Y'AY \cdot Y'BY] = 2\mathrm{tr}(A\Sigma B\Sigma).$$

Data Generation and Processing for the Broad AASC Data

The ASD case-control data set has been sequenced as part of the American Recovery and Reinvestment Act (ARRA) Autism Sequencing Collaboration (AASC). WES of the samples was carried out at the Broad Institute as follows.

The SureSelect v.2 Human exon Agilent 38 Mb exon-capture kit was used for library enrichment (Agilent Technologies). After capture, another round of ligation-mediated PCR was performed for increasing the quantity of DNA available for sequencing. All libraries were sequenced with an Illumina HiSeq2000 according to the manufacturer's (Illumina) instructions for paired-end 100 bp reads. After sequencing, the data were put through a computational pipeline for WES data processing and analysis according to the general workflow adopted by the 1000 Genomes Project.[40] First, the alignment of raw sequence reads to the human reference genome sequence (NCBI GRCh37) was performed with a fast lightweight Burrows-Wheeler Aligner[41] and a binary version of SAMtools.[42] The Genome Analysis Toolkit (GATK)[43] was then used for base-quality recalibration and local realignment for minimizing base-calling error and mapping error, respectively. SNPs were called with GATK for all samples jointly. Only variants passing GATK standard filters were considered for analysis. Resulting calls were annotated with Snpeff[44] and GATK VariantAnnotator tools.

Additional Gene-Based Tests

In addition to the tests described in Table 1, we also considered several additional tests, as follows:

1. SKAT-rare: the original SKAT test restricted to rare variants (those variants with MAF $< 1/\sqrt{2n}$).
2. SKAT-common: the original SKAT test restricted to common variants (those variants with MAF $> 1/\sqrt{2n}$) and with weight $Beta(\widehat{p_j}, 0.5, 0.5)$.
3. SKAT-NW: the original SKAT test for all variants in a region and with no variant weighting.
4. LRT-common: the likelihood ratio test with common variants only.
5. LRT: in which the minimum of the p values of SKAT-rare and LRT-common tests is multiplied by 2.
6. Min-p: in which the minimum of the p values of SKAT-rare and of common variants is multiplied by 2 (calculating the p value for common variants requires that the minimum p value of the p values for individual common variants be adjusted for the effective number of variants, as described in Gao et al.[45]).

For LRT and Min-p, the minimum of the p values of common and rare variants tests is multiplied by 2 so that the number of tests can be adjusted for. Because the correlation between rare and common variants tends to be low, this adjustment is largely accurate.

Supplemental Data

Supplemental Data include 32 figures and 1 table and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

The research was partially supported by National Science Foundation grant DMS-1100279 and National Institutes of Health grants R01MH095797 and 1R03HG005908 (to I.I.-L.), National Institutes of Health grant K99-HL113164 (to S.L.), a Seaver Foundation grant and National Institutes of Health grants MH089025 and MH100233 (to J.D.B.), and National Institutes of Health grants R37 CA076404 and P01CA134294 (to S.L. and X.L.).

Received: January 28, 2013
Revised: March 20, 2013
Accepted: April 18, 2013
Published: May 16, 2013

Web Resources

The URLs for data presented herein are as follows:

HapMap 3, http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/omim
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/
SKAT, www.hsph.harvard.edu/xlin/software.html

References

1. Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nat. Genet. *44*, 623–630.
2. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al.; National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC); United Kingdom Inflammatory Bowel Disease Genetics Consortium; International Inflammatory Bowel Disease Genetics Consortium. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066–1073.
3. Asselbergs, F.W., Guo, Y., van Iperen, E.P., Sivapalaratnam, S., Tragante, V., Lanktree, M.B., Lange, L.A., Almoguera, B.,

Appelman, Y.E., Barnard, J., et al.; LifeLines Cohort Study. (2012). Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. Am. J. Hum. Genet. *91*, 823–838.

4. Diogo, D., Kurreeman, F., Stahl, E.A., Liao, K.P., Gupta, N., Greenberg, J.D., Rivas, M.A., Hickey, B., Flannick, J., Thomson, B., et al.; Consortium of Rheumatology Researchers of North America; Rheumatoid Arthritis Consortium International. (2013). Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. Am. J. Hum. Genet. *92*, 15–27.
5. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.
6. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
7. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
8. Klei, L., Sanders, S.J., Murtha, M.T., Hus, V., Lowe, J.K., Willsey, A.J., Moreno-De-Luca, D., Yu, T.W., Fombonne, E., Geschwind, D., et al. (2012). Common genetic variants, acting additively, are a major source of risk for autism. Mol Autism *3*, 9.
9. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., and Wray, N.R.; Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ); International Schizophrenia Consortium (ISC); Molecular Genetics of Schizophrenia Collaboration (MGS). (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. *44*, 247–250.
10. Agresti, A. (2002). Categorical Data Analysis, Second Edition (Gainesville, FL: John Wiley & Sons).
11. 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.
12. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
13. Price, A.L., Kryukov, G.V., de Bakker, P.I., 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.
14. Liu, D.J., and Leal, S.M. (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*, e1001156.
15. 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.
16. Ionita-Laza, I., Buxbaum, J.D., Laird, N.M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. *7*, e1001289.
17. 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.
18. 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.
19. 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.
20. Tzeng, J.Y., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M.I., Sale, M.M., Worrall, B.B., Hsu, F.C., Thomas, D.C., and Sullivan, P.F. (2011). Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. Am. J. Hum. Genet. *89*, 277–288.
21. 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.; NHLBI GO Exome Sequencing Project—ESP Lung Project Team. (2012). 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.
22. Lee, S., Wu, M.C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. Biostatistics *13*, 762–775.
23. Chen, L.S., Hsu, L., Gamazon, E.R., Cox, N.J., and Nicolae, D.L. (2012). An exponential combination procedure for set-based association tests in sequencing studies. Am. J. Hum. Genet. *91*, 977–986.
24. Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M., and Gerstein, M.B. (2010). Annotating non-coding regions of the genome. Nat. Rev. Genet. *11*, 559–571.
25. Cai, T., Jeng, J., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *73*, 629–662.
26. Davies, R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika *64*, 247–254.
27. Zhang, D., and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. Biostatistics *4*, 57–74.
28. Brown, M. (1975). 400: A method for combining non-independent, one-sided tests of significance. Biometrics *31*, 987–992.
29. Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Res. *15*, 1576–1583.
30. Lesage, S., Zouali, H., Cézard, J.P., Colombel, J.F., Belaiche, J., Almer, S., Tysk, C., O'Morain, C., Gassull, M., Binder, V., et al.; EPWG-IBD Group; EPIMAD Group; GETAID Group. (2002). CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. Am. J. Hum. Genet. *70*, 845–857.
31. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
32. Ionita-Laza, I., Makarov, V., and Buxbaum, J.D.; ARRA Autism Sequencing Consortium. (2012). Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked

and associated with autism spectrum disorders, in three datasets. Am. J. Hum. Genet. *90*, 1002–1013.
33. Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., et al. (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285–299.
34. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.
35. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. *367*, 1921–1929.
36. Cardin, N.J., Mefford, J.A., and Witte, J.S. (2012). Joint association testing of common and rare genetic variants using hierarchical modeling. Genet. Epidemiol. *36*, 642–651.
37. Derkach, A., Lawless, J.F., and Sun, L. (2012) Assessment of Pooled Association Tests for Rare Genetic Variants within a Unified Framework. arXiv, arXiv:1205.4079, http://arxiv.org/abs/1205.4079.
38. Basu, S., and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. Genet. Epidemiol. *35*, 606–619.
39. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D., and Lin, X. (2013). Family-based association tests for sequence data, and comparisons with population-based association tests. Eur. J. Hum. Genet. http://dx.doi.org/10.1038/ejhg.2012.308.
40. 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., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.
41. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.
42. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
43. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
44. Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.
45. Gao, X., Starmer, J., and Martin, E.R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet. Epidemiol. *32*, 361–369.