5-2016

General Approach for Combining Diverse Rare Variant Association Tests Provides Improved Robustness Across a Wider Range of Genetic Architectures

Brian Greco
University of Michigan-Ann Arbor

Allison Hainline
Vanderbilt University

Jaron Arbet
University of Minnesota - Twin Cities

Kelsey Grinde
University of Washington - Seattle Campus

Alejandra Benitez
University of California - Berkeley

See next page for additional authors

General Approach for Combining Diverse Rare Variant Association Tests Provides Improved Robustness Across a Wider Range of Genetic Architectures

Abstract
The widespread availability of genome sequencing data made possible by way of next-generation technologies has yielded a flood of different gene-based rare variant association tests. Most of these tests have been published because they have superior power for particular genetic architectures. However, for applied researchers it is challenging to know which test to choose in practice when little is known *a priori* about genetic architecture. Recently, tests have been proposed which combine two particular individual tests (one burden and one variance components) to minimize power loss while improving robustness to a wider range of genetic architectures. In our analysis we propose an expansion of these approaches, yielding a general method that works for combining any number of individual tests. We demonstrate that running multiple different tests on the same dataset and using a Bonferroni correction for multiple testing is never better than combining tests using our general method. We also find that using a test statistic that is highly robust to the inclusion of non-causal variants (Joint-infinity) together with a previously published combined test (SKAT-O) provides improved robustness to a wide range of genetic architectures and should be considered for use in practice. Software for this approach is supplied. We support the increased use of combined tests in practice-- as well as further exploration of novel combined testing approaches using the general framework provided here--to maximize robustness of rare-variant testing strategies against a wide range of genetic architectures.

Keywords
next-generation sequencing, genome-wide association studies, case-control

Disciplines
Genetics and Genomics

Authors
Brian Greco, Allison Hainline, Jaron Arbet, Kelsey Grinde, Alejandra Benitez, and Nathan L. Tintle

Title: A general approach for combining diverse rare variant association tests provides improved robustness across a wider range of genetic architectures

Running title: General approach for combining rare variant tests

Authors: Brian Greco[1], Allison Hainline[2], Jaron Arbet[3,4], Kelsey Grinde[5,6], Alejandra Benitez[7], Nathan Tintle[8]

1. Department of Biostatistics, University of Michigan, Ann Arbor, MI
2. Department of Biostatistics, Vanderbilt University, Nashville, TN
3. Department of Statistics, Winona State University, Winona, MN
4. Department of Biostatistics, University of Minnesota, Minneapolis, MN
5. Department of Mathematics, Statistics and Computer Science, St. Olaf College, Northfield, MN
6. Department of Biostatistics, University of Washington, Seattle, WA
7. Department of Biostatistics, U.C. Berkeley, Berkeley, CA
8. Department of Mathematics, Statistics and Computer Science, Dordt College, Sioux Center, IA

Corresponding author:

Dr. Nathan Tintle, Department of Mathematics, Statistics and Computer Science, Dordt College, 498 4th Ave. NE, Sioux Center, IA 51250. Phone: 712-722-6264. Fax: 712-722-6035. Email: nathan.tintle@dordt.edu

Abstract (250 word max; currently 239)

The widespread availability of genome sequencing data made possible by way of next-generation technologies has yielded a flood of different gene-based rare variant association tests. Most of these tests have been published because they have superior power for particular genetic architectures. However, for applied researchers it is challenging to know which test to choose in practice when little is known *a priori* about genetic architecture. Recently, tests have been proposed which combine two particular individual tests (one burden and one variance components) to minimize power loss while improving robustness to a wider range of genetic architectures. In our analysis we propose an expansion of these approaches, yielding a general method that works for combining any number of individual tests. We demonstrate that running multiple different tests on the same dataset and using a Bonferroni correction for multiple testing is never better than combining tests using our general method. We also find that using a test statistic that is highly robust to the inclusion of non-causal variants (Joint-infinity) together with a previously published combined test (SKAT-O) provides improved robustness to a wide range of genetic architectures and should be considered for use in practice. Software for this approach is supplied. We support the increased use of combined tests in practice-- as well as further exploration of novel combined testing approaches using the general framework provided here--to maximize robustness of rare-variant testing strategies against a wide range of genetic architectures.

Key words: next-generation sequencing; genome-wide association studies; case-control

Introduction:

Numerous tests of genotype-phenotype association for rare variants have been proposed, all of which attempt to combine signals at multiple variant sites within a gene into a single, powerful gene-based test of association. According to recent work, which test is most powerful is highly dependent upon the true genetic architecture of the phenotype [1,2]. The challenge for the applied researcher is to know which test to choose, given limited information about the true genetic architecture of disease.

A general understanding of test behavior can be obtained by noting the existence of two broad classes of tests (length and joint) among the many tests proposed to date[1]. Length tests (alternatively: burden, collapsing, linear; for example, CMC[3]) attempt to enhance the genotype-phenotype signal in a region of interest by collapsing variant measurements into a single measure of rare variant "burden," which is then tested for association with a phenotype of interest. They are called length tests because they can be interpreted geometrically as testing for a difference in the lengths of the minor allele frequency vectors between cases and controls. These tests tend to be powerful when the proportion of causal variants is large and the effects of the causal variants are similar[1]. Joint tests (alternatively: variance components, quadratic; for example, SKAT[4]) combine the strength of evidence of individual phenotype-variant associations across the variants in a region of interest and tend to be powerful when there are larger proportions of non-causal variants and there is more variation in the effects of causal variants[1]. Joint tests are so named because they simultaneously test for differences between the lengths of the minor allele frequency

vectors in cases and controls, as well as testing for a non-zero angle between the vectors. A full discussion and classification of existing tests is available elsewhere[1,2].

Recent papers have proposed combining test statistics across both the length and joint classes to yield more powerful test statistics [1,5–8]. Results from these papers demonstrate how to combine a single version of a length test with a single version of a joint test[5], how to use a weighting strategy to find the optimal weighted combination of two particular length and joint test statistics[6], and that different weighted combinations of particular length and joint tests can be more powerful than single tests for different genetic architectures[1]. Overall, these combined testing approaches show improved power against a wider range of genetic architectures when compared to using either statistic separately [1,5–7].

In general, any approach that combines a single length test and a single joint test will have a limited range of situations in which it is powerful. In particular, the combined test can only be powerful in cases where either of the two individual tests being combined is powerful. The combined test will lack power where the two tests being combined, simultaneously, lack power (but potentially where another, powerful, alternative test exists). For example, a recent paper suggested novel test statistics which may provide increased power when a large proportion of non-causal variants is present in the gene[1], but current test-combining strategies have not evaluated this class of alternatives. Thus, more general test-combining strategies are needed in order to potentially yield more powerful results when the component tests being combined are powerful for a wide range of genetic architectures.

In this paper we will demonstrate how to combine an arbitrarily large and diverse set of gene-based rare variant test statistics using an efficient permutation strategy. We then simulate a wide range of genetic architectures and evaluate the performance of two different methods of combining tests (Fisher's, minimum p-value) when combined tests involve many different types of tests, including those using a variety of norms. We explore which combinations of tests are ideal and when.

Methods

General strategy for combining tests

We propose the following approach for combining *p*-values from *k* different gene-based rare variant tests. For a gene of interest, calculate $\boldsymbol{f}^+$and $\boldsymbol{f}^-$**,** where $\boldsymbol{f}^+$ is a vector of observed allele frequencies, $(f_1^+, f_2^+, \ldots, f_m^+)$, in the cases, across the *m* variant sites in the gene and where $f_j^+ = \frac{c_j^+}{2N^+}$, letting c_j^+indicate the total number of minor alleles in the cases at site *j*, and N^+ be the number of cases in the sample. Vector $\boldsymbol{f}^-$ holds similar definitions for the controls.

After computing $\boldsymbol{f}^+$and $\boldsymbol{f}^-$, find the *p*-value for each of the *k* different gene-based rare variant tests, yielding a vector of *p*-values, $\boldsymbol{p}=(p_1, p_2, \ldots, p_k)$, for each gene of interest (see *Rare variant tests* section for details). The vector $\boldsymbol{p}$ is used to generate a test statistic, $S_k=f(\boldsymbol{p})$, which summarizes the strength of evidence across $\boldsymbol{p}$; essentially, the combined strength of evidence of genotype-phenotype association across the entire set of *k* tests. We consider two different ways of computing S_k. The first is the Fisher's combined *p*-value

test statistic and is computed as $F_k = \sum_{i=1}^{k} -2\log(p_i)$. We note that if the k tests were mutually independent, the distribution of F_k would follow a chi-squared distribution; however, that is likely not the case in practice. Instead, we assess significance of F_k using the permutation strategy described in the following section.

The second summary statistic is the minimum *p*-value, ***Min(p)***, with significance assessed using the permutation strategy described in the following section. For comparison, we also compute significance of the *Min(p)* statistic using a Bonferroni correction approach where the summary statistic is deemed significant if *Min(p)* is less than α/k, for some *a priori* specified α.

Description of the permutation strategy

For a general univariate summary statistic S_k of vector $\boldsymbol{p}$ (in our case either F_k or *Min(p)*), statistical significance can be assessed by permuting phenotype status, performing k tests on the permuted data, recomputing S_k on each permutation, and calculating the percent of times that permuted values of S_k are greater than the observed S_k. Recently[5], an efficient permutation strategy for assessing the significance of a test S_k with k=2 (one length and one joint) test was proposed. We extend the approach for any number of gene-based tests k of any type. The extended approach is to: (1) Calculate the observed value of S_k as a function of $\boldsymbol{p}$, where $\boldsymbol{p}$ is the vector of *p*-values for each of the $i=1,\ldots,k$ tests being combined. (2) Permute the phenotype and re-compute test statistics, $t_i^*(l)$, under permutation for each of the $i=1,\ldots k$ tests and for each of $l=1,\ldots,P$ permutations (where P is large), yielding $t_i^* = (t_i^*(1), t_i^*(2), \ldots, t_i^*(P))$, a vector of permuted test statistics for test *i. Note: These are the*

same P permutations for all tests. (3) Calculate $Rank(t_i^*(l))$, the rank of each of the test statistics in vector t_i^* for each of the *i=1,…,k* tests, where $Rank(t_i^*(l))$=1 for the largest value of $t_i^*(l)$ and $Rank(t_i^*(l))$=*P* for the smallest value of $t_i^*(l)$. (4) Calculate an empirical *p*-value for each of the permuted test statistics as $p_i^*(l) = Rank(t_i^*(l))/P$. (5) An empirical null distribution (no genotype-phenotype association) for *S* is computed by calculating the value of $S_k(l)$ from the vector of *p*-values $\boldsymbol{p}^*(\boldsymbol{l}) = (p_1(l), p_2(l), \dots, p_k(l))$, for each permutation *l=1,…,P.* (6) The significance of S_k is computed by calculating the percentage of $S_k(l)$ values that are larger than S_k, out of the set of *P* phenotype permutations.

A few additional comments are worthwhile. First, the procedure can be modified in a straightforward manner for two-sided tests (either individual or combined), by looking at both tails of the empirical null distribution of statistics. Second, for individual tests based on asymptotic distributions, steps (3) and (4) are merely replaced by using the asymptotic distribution to calculate the $p_i(l)$. Finally, and importantly, we note that the use of the same *P* permutations in step (5*)* is needed in order to properly model the correlation structure between tests and generate an appropriate null distribution for S_k.

Rare variant tests

We explored combinations of different gene-based rare variant tests which were selected to represent a variety of different approaches for evaluating genotype-phenotype associations. We define $\|\boldsymbol{x}\|_p = \left(\sum_{j=1}^{m} |x_i|^p\right)^{1/p}$ as the *p*-norm for a vector $\boldsymbol{x} = (x_1, x_2, \dots, x_m)$. The individual rare variant tests we considered were: (1) *Sequence Kernel Adaptive Test*

(*SKAT*) [4]. *SKAT* is essentially equivalent to $Q_S = \|\boldsymbol{f}^+ - \boldsymbol{f}^-\|_2{}^1$ with an asymptotic distribution used for statistical significance-- a joint test using the 2-norm. (2) *Combined Multivariate and Collapsing Test* (*CMC*)[3]. When all variants are collapsed, *CMC* can be viewed as essentially equivalent to $Q_B = \|\boldsymbol{f}^+\|_1 - \|\boldsymbol{f}^-\|_1{}^1$ with significance assessed using an asymptotic distribution-- a length test using a 1-norm. In our analysis we collapsed all variants because our simulations focused on variants with population minor allele frequency less than 1%. (3) *Sequence Kernel Adaptive Test-Optimal* (*SKAT-O*) [6,7]. *SKAT-O* combines SKAT and a general burden test (CMC) by the optimal weight ρ, such that $Q_\rho = \rho Q_B + (1-\rho)Q_S$ yields the minimum p-value and uses an asymptotic distribution to assess statistical significance. (4) Length tests with different norms (*L(p)*) [1], which test for differences in the lengths of the minor allele frequency vectors between cases and controls. We considered four versions of length tests of the form $L(p) = \|\boldsymbol{f}^+\|_p - \|\boldsymbol{f}^-\|_p$, with significance assessed via phenotype permutation. The four versions were generated by considering different values of the norm, *p*, *p*=1, 2, 4 and ∞, where $\|\boldsymbol{x}\|_\infty = \max(\text{abs}(x_i))$. (5) Joint tests with different norms (*J(p)*) [1], which simultaneously test for differences in the lengths and for a non-zero angle between the two allele frequency vectors. We considered four versions of joint tests of the form $J(p) = \|\boldsymbol{f}^+ - \boldsymbol{f}^-\|_p$, with significance assessed via phenotype permutation. We used four different values of *p*, *p*=1, 2, 4 and ∞. Higher normed tests are more robust to the inclusion of non-causal variants[1]. Thus, we considered a total of 11 individual gene-based variant tests (*SKAT (a 2-norm joint test), SKAT-O (a combined test), CMC (a 1-norm length test), L(1), L(2), L(4), L(∞), J(1), J(2), J(4), J(∞)*.

We then combined subsets of the 11 individual gene-based rare tests using both the Fisher's and *Min(p)* approaches (see *Methods: General strategy for combining tests section*). The 8 different combinations of tests we considered were: (1) *Length tests with different norms (L(1), L(2), L(4), L(∞)*) (*CT1*), (2) *Joint tests with different norms (J(1), J(2), J(4), J(∞)) (CT2),*(3) *Similar length tests (CMC, L(1)) (CT3),* (4) *Similar joint tests (SKAT, J(2)) (CT4),* (5) *Typical length-joint combined test (SKAT, CMC) (CT5),* (6) *Length and joint tests across norms (L(1), L(2), L(4), L(∞),(J(1), J(2), J(4), J(∞)*) (*CT6*), (7) *Length and joint with some norms (L(1), L(4), J(1), J(4) (CT7),* (8) *More robust SKAT-O (SKAT-O, J(∞)) (CT8).* A brief rationale for the inclusion of each test is provided in Table 1.

Simulations

We conducted two main simulation studies as part of our analysis. In the first simulation, we explored the general behavior of the *Fisher's* and *Min(p)* approaches across a variety of different numbers of tests, correlation structures and power settings using generalized gene-based test statistics. In the second simulation we simulated data according to *a priori* specified genetic disease models and applied the gene-based rare variant tests of association described in the previous section.

Simulation #1: Investigating the behavior of Min(p) and Fisher's

Data was simulated from multivariate normal random variables, $T \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (*MVN* = Multivariate Normal), using R^9 where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$ and the $k \times k$ covariance matrix

$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,k} \\ \rho_{2,1} & 1 & \rho_{2,3} & \cdots \\ \cdots & \rho_{3,2} & 1 & \rho_{k-1,k} \\ \rho_{k,1} & \cdots & \rho_{k,k-1} & 1 \end{pmatrix}$. Each multivariate normal sample represents a vector of test statistics, *T*, from *k* different gene-based rare variant tests, where H_0: $\boldsymbol{\mu} = \mathbf{0}$, H_a: at least one of $\mu_1, \mu_2, \ldots, \mu_k > 0$ and $\rho_{i,j}$ is a measure of correlation between tests *i* and *j*. We consider all possible combinations of the following parameters: (1) Number of tests, *k*, equal to 2, 4, 6, 10 and 20 (2) $\rho_{i,j}$ = 0, 0.25, 0.50, 0.75, 0.90 and 0.99 between the test statistics of two tests *i*,*j*. Note: we specified the correlation ρ between test statistics, however the corresponding correlations between p-values are quite similar (details not shown). (3) (a) H_0: $\boldsymbol{\mu} = (\mu_1 = 0, \mu_2 = 0, \ldots, \mu_k = 0)$ (b) An H_a where all tests perform equally well: $\boldsymbol{\mu} = (\mu_1 = 2, \mu_2 = 2, \ldots, \mu_k = 2)$. We note that the approximate power of each individual test, *i*, under the alternative hypothesis ($\mu_i = 2$) is equal to $P(Z > z_\alpha - \mu_i) = P(Z > -0.355) = 0.64$, where $Z \sim$ Normal(0,1) at a significance level of 5% ($z_\alpha = 1.645$) for a one-sided upper-tailed test, representing a moderately powered test. We also considered lower significance levels of 0.01, 0.001 and 0.0001, which yield individual test power of 37%, 14% and 4%, respectively.

After generating 10,000 multivariate normal random samples for each combination of simulation parameters, we computed the *p*-value of each test statistic, T_i, for each of the 10,000 samples, by finding $1 - \phi(T_i)$ where $\phi()$ is the cumulative distribution function (CDF) of a standard, normal distribution. We then applied *Min(p)* and Fisher's methods to each set of p-values, with significance assessed by comparing alternative hypothesis values of *Min(p)* and Fisher's statistics to the simulated distributions of these statistics under the

null hypothesis. The power of each approach (*Min(p)* and Fisher's) for each simulation setting is estimated by dividing the fraction of significant (α=0.05, 0.01, 0.001 or 0.0001) statistics by 10,000 (the number of independent samples). We then conducted a follow-up simulation in which we varied the number of tests, *k* (*k*=2, 4, 6, 10 and 20), fixed $\rho_{i,j} = 0$ between two tests i,j and then varied the number of tests for which $\mu_i = 2$ from 1 to 10, with the remaining tests having $\mu_i = 0$. Full results from these simulations, which include observed correlations between p-values for all settings illustrating the approximately equivalent correlations between test statistics and p-values, are available in *Supplemental Tables* 1a-1c.

Simulation #2- Investigating the behavior of combinations of gene-based rare variant tests across different genetic disease models

We simulated data to represent a variety of different genetic disease models. In all simulations, we considered a sample size of 2,000 individuals split evenly between cases and controls. We then simulated data across all possible combinations of the following parameters: (1) Number of single nucleotide variants (SNVs) (32 or 64) (2) Proportion of non-causal SNVs (0, ¼, ½, ¾, 7/8, 15/16, 31/32, 63/64, 1) (3) Proportion of causal SNVs that increase disease risk (0, ¼, ½, ¾, 1), with the remaining causal SNVs causing a decline in disease risk (4) Relative risk of causal, risk-increasing SNVs (1.1, 1.5 and 2.0). To investigate impact on test performance in the presence of risk-reducing SNVs, some simulation settings included risk-reducing SNVs with relative-risk 0.5. Furthermore, SNV minor allele frequencies were simulated in a three to one ratio of less common (0.1% population minor allele frequency) to more common (1% minor allele frequency) SNVs

spread evenly across all non-causal and causal SNVs. We note that when the number of SNVs is not divisible by 4, a single 1% minor allele frequency SNV is assigned before generating up to 3 additional 0.1% minor allele frequency SNVs. Thus, there were a total of 2 (number of SNVs) x 9 (proportion of non-causal) x 5 (proportion of risk increasing SNVs) x 3 (relative risk of risk increasing SNVs) settings, of 270 possible simulation settings. However, some of the combinations are redundant or impossible; removing these cases yields 197 total simulation settings considered in our analysis.

Five-hundred samples were generated at each simulation setting, with each of the 20 individual tests and each of the 11 combined tests applied to each sample, and separate p-values for *Min(p)* (permutation p-value) and Fisher's for each combined test. Empirical power estimates are computed as the percentage of p-values less than 0.05 (nominal alpha), giving power estimates within $2\sqrt{\frac{0.5(1-0.5)}{500}} \approx 4\%$ of the true power 95% of the time. For the Bonferroni testing approach, we deem the test significant if at least one of the individual test *p*-values in the set is below the Bonferonni correct alpha value of 0.05/k. Where needed, 500 permutations were used to assess statistical significance for individual and combined tests.

To further explore test performance at significance levels commonly used in practice, additional simulations were conducted. In particular, 16 of the settings described above were investigated using 50,000 permutations at significance levels of 10^{-4}, 10^{-3} and 10^{-2}. Fourteen of these settings represented situations in which causal variants were present (32 total SNPs with 1,2 or 4 causal variants; 64 total SNPs with 1,2,4 or 8 causal variants),

where all causal variants have RR=2 (7 cases) or 3 (7 cases); 200 simulations were conducted at each setting. Two settings represented situations in which no causal variants were present (32 total SNPs and 64 total SNPs), and used 840 and 460 total simulations at each setting, respectively.

Application

As a proof of concept, we applied select gene-based tests to data from Genetic Analysis Workshop 17. The data consists of real genotype data (from the 1000 Genomes Project consortium) on which a disease phenotype was simulated[10]. We considered 25 genes which were known to contain causal variants for the simulated disease phenotype and showed variation in the sample of n=321 unrelated Asian subjects. Given the small sample size and low power in this dataset[5], final disease status for each of the 321 individuals was averaged across 200 independent phenotype simulations, with individuals who were diseased in at least 100 of the 200 independent simulations identified as 'diseased,' and the rest not. As has been done previously[5], we used a significance level of 0.05 for this analysis.

Results

General patterns in the performance of Min(p) and Fisher's methods (Simulation #1)

We start by exploring the general behavior of Min(p) and Fisher's method across a generic set of k tests, with different correlation structure and test performance (*Simulation #1* described earlier). The goal of this analysis is to provide an intuitive sense of how the number of tests, correlation between tests and individual test performance is related to the performance of *Min(p)* and Fisher's method in a well-understood environment. Detailed

simulation results are provided in *Supplemental Tables 1a-1c*. *Supplemental table 1a* illustrates that the type I error rate is controlled across all simulation settings and significance levels.

When all tests are powerful

When all tests being combined have good power (64% at $\alpha = 0.05$), both the Fisher's and *Min(p)* approaches yield increased power as the number of tests being combined increases. However, Fisher's method tends to outperform *Min(p),* with the magnitude of the power gain for Fisher's relative to *Min(p)* decreasing as the correlation between tests increases, and the power of combined, highly correlated tests equal to the power of a single test--approximately 64% (see *Supplemental Table 1b* and *Figure 1*). In situations where all tests are powerful, *Min(p)* ignores the power from all the tests but one, forgoing the opportunity to improve the power by combining tests and yielding lower power overall as compared to Fisher's approach. Similar results are observed for other significance levels.

When some tests are powerful

When we varied the number of powerful (good) tests (power=64% at $\alpha = 0.05$) and under-powered (bad) tests (power=5%=type I error rate) we found that *Min(p)* outperforms Fisher's if there is only one good test in the set, with the magnitude of improvement increasing as the number of bad tests increases (for example, see *Figure 2,* similar results are observed for other significance levels, see *Supplemental Table 1c*.). When there are two good tests in the set, Fisher's does better when there are few bad tests, but as more and more bad tests are added to the set, *Min(p)* gains an advantage over Fisher's. In general, *Min(p)* outperforms Fisher's when the proportion of bad tests in the set is large. The

impact of correlation between tests on these relationships can be inferred from the previous section.

Performance of combined tests on simulated phenotype-genotype data (Simulation #2)

Type I error simulation

The type I error simulation showed general control of the type I error rate across all individual tests and combined tests considered here, with the lone exception being the Bonferroni method, which was, as expected, often conservative. Detailed Type I error simulation results are in *Supplemental tables 2a and 2b*. Additional simulations at lower significance levels ($1x10^{-2}$, $1x10^{-3}$, $1x10^{-4}$) also showed control of the type I error rate in all cases (detailed results not shown).

Min(p) beats Bonferroni every time

Across the 197 simulation settings and 8 combined tests (1576 possibilities; see *Supplemental Table 3*), as well as all follow-up simulations at lower significance levels, there were only 10 times where power of the Bonferroni approach exceeded the power of the *Min(p)* approach, doing so only minimally (ranging from 0.002 to 0.004); well within the range of expected variation due to simulation. Thus, it is safe to conclude that *Min(p)* will always be better than Bonferroni. We do not consider the Bonferroni approach in subsequent analyses.

Improving a combined test with additional tests

We explored 8 different combined tests. Rationale and summaries of performance are provided in Table 1. In general, the results of the second simulation study confirmed results of the first simulation study with regards to the use of *Min(p)* or Fisher's and how many tests to combine. In short, (1) combining tests that are powerful in different situations will generally be advantageous (e.g., CT6, CT7 and CT8), (2) *Min(p)* outperforms Fisher's combining method when there is a mix of powerful and non-powerful tests being combined (e.g., CT5, CT6, CT7) and (3) combining highly correlated tests has little benefit (e.g., CT2, CT3, CT4). These results held true even at lower significance levels (see *Supplemental Table 4*)

Robust test statistic

As shown in Table 1, CT8 yielded the best overall performance, with the Fisher's method performing slightly better than the *Min(p)* method across all simulation settings; CT6 and CT7 also performed quite well. Across the 197 simulation settings, CT8 (combination of SKAT-O and $J(\infty)$) yielded power no more than 5% smaller than SKAT-O power in 87.3% (Fisher's; 172/197) and 83.2% (*Min(p)*; 164/197) of simulation settings. The power of CT8 was never worse than 10% less than SKAT-O power. However, the combined test was sometimes substantially better than SKAT-O, as shown in *Table 2*. In particular, since $J(\infty)$ is robust to the inclusion of high proportions of non-causal variants, CT8 is more robust to the inclusion of non-causal variants than SKAT-O alone. $J(\infty)$, however, performs more poorly than SKAT-O and most other tests when the proportion of causal variants in a gene is moderate (see *Supplemental Table 3*, which provides the full results for all simulation settings, for details). Finally, Figures 3 and 4 illustrate the performance of the methods at a

low significance level, showing similar results at a relative risk of 2. We note that the power is not very high in this case. Supplemental figures 2 and 3 illustrate the same performance using a relative risk of 3, yielding larger power.

The performance of the Fisher's combination approach was generally better than the *Min(p)* approach of CT8 as shown in Tables 1 and 2. In a head to head comparison, the Fisher's approach yielded better power than the *Min(p)* approach in more than twice as many simulations (119 vs. 45 settings), though power gains were only modestly better (average power gain 1.8% vs. 1%), with a max power difference of only 5.2%. Table 2 also illustrates the relatively good performance of CT6 and 7 in this subset of simulation settings.

Application to data from Genetic Analysis Workshop 17

The p-values for four tests (SKAT-O, *J(∞)*) and both the Fisher's and *Min(p)* versions of CT8) which were applied to 25 genes containing at least one causal variant are provided in *Supplemental Table 5*. Six genes are significant ($p<0.05$) using SKAT-O alone and four genes are significant using *J(∞)* alone (three genes are significant using both approaches), for a total of seven genes identified by at least one of the two individual testing methods. The *Min(p)* version of CT8 identified all seven of the genes as significant and Fisher's identified five of the seven as significant, while the remaining two were borderline significant ($p<0.07$), demonstrating that the combined methods are robust. In particular, we note that the *PIK3C3* gene was significant using the *J(∞)* approach ($p=0.035$), but not

SKAT-O (p=0.056), and was significant for both combined tests (*Min(p)* p-value=0.041, Fisher's p-value=0.035).

Software

Software written for R[9] is available for free download on the research group's software page (http://www.dordt.edu/academics/programs/math/statgen/software.shtml). All individual and combined tests considered here are included.

Discussion

We have proposed a general and flexible method for combining different rare variant tests of association to potentially improve robustness across a wide range of genetic architectures while minimizing power loss through the addition of multiple tests. A naïve approach to combining tests is to use a Bonferroni correction after applying multiple different rare variant tests to the same data. However, Bonferroni is often conservative, especially when tests being combined are correlated, and we demonstrated that the *Min(p)* approach is always more powerful because it empirically estimates the appropriate correlation structure. Thus, in practice, researchers should never run multiple (k>1) gene-based tests on the same dataset and then apply a stricter Bonferroni correction strategy ($\alpha/(k$*genes)) to their dataset. The *Min(p)* approach proposed here will always be more powerful than such an approach.

We also showed that while the *Min(p)* approach is sometimes optimal, the Fisher's method offers advantages over *Min(p)* in some cases because it combines separate signals

into a combined signal when tests are well-powered and the correlation between tests is low. However, we've shown that when combining tests with lower power, *Min(p)* improves to the point of being better than Fisher's method in some cases. In short, *Min(p)* ignores the 'noise' of low powered tests, while Fisher's averages low powered tests into the signal. Furthermore, as the correlation between well-powered tests increases, *Min(p)* also gains power relative to Fisher's. Ultimately, the answer to whether *Min(p)* or Fisher's provides more power is dependent upon the underlying power and correlation structure of the tests being combined. However, combining highly correlated tests is not advantageous either. The most benefit is obtained by combining disparate tests-- as we illustrated by combining *J(∞)* with SKAT-O--to yield a more robust and powerful test. Across simulation settings considered here the Fisher's approach for the SKAT-O/ *J(∞)* combined test was somewhat more robust than the *Min(p)* approach and so is recommended for use in practice.

More broadly than either *Min(p)* or Fisher's, our method is flexible enough to consider any of the numerous other choices for S_k, which is simply a function of the vector of *p*-values from the *k*-tests being combined, $\boldsymbol{p}=(p_1, p_2, \ldots, p_k)$. We have focused on Fisher's and *Min(p)* because they represent two extreme approaches: Fisher's is a weighted average of all the p-values, and *Min(p)* only uses a single value from the vector. Furthermore, both approaches are popular since, when tests are independent, each has fairly well understood asymptotic properties. More research is needed to explore additional possibilities. We note that while we restricted our analysis to case-control study designs, the results are directly applicable to results for quantitative traits.

A key advantage to the combined testing approach comes when evaluating multiple genes and/or multiple phenotypes. In these cases, *a priori*, there may be little information about which individual test is most powerful given the wide range of potential genetic architectures. The best test strategy will be one which provides an optimal tradeoff of power loss and robustness. Namely, for any particular genetic architecture, an individual test can be constructed with better power than any combined test. However, individual tests may be powerful against only a small set of genetic architectures. Thus, a combined test may trade off (vs. an individual test) small amounts of power against some genetic architectures for large improvements in power versus other genetic architectures.

One area of application we have explored is the straightforward application of our approach to gene-based rare variant tests that use thresholds (e.g., CMC[3] which thresholds on Minor Allele Frequency, or the Odds Ratio Weighted Sum Statistic[11] with thresholds on empirical odds ratio) to generate variable threshold tests in a straightforward manner. In short, simply combine the same test across multiple thresholds to yield an optimally robust test (detailed results not shown).

With this in mind, how should a researcher utilize combined tests in practice? Prior work[5–7] has shown that combined tests can be considered 'optimal,' however, these approaches have been limited to combining *L(1)* and *J(2)* tests. In this paper we have shown that combining other disparate tests can be advantageous (e.g., combining SKAT-O, itself a combination of *L(1)* and *J(2)*, with *J(∞)*). For example, we showed that the inclusion of a higher norm test can provide increased robustness to the inclusion of non-causal variants. In practice, we recommend including *J(∞)* in a combined test with *L(1)*

and *J(2)* (e.g., SKAT-O with *J(∞)*) to maximize robustness to the inclusion of non-causal variants in cases where little prior knowledge exists to prioritize potential causal SNPs and/or it is anticipated that a high proportion of SNPs included in the test may be non-causal. However, further analysis of simulated data with larger sample sizes, additional variation in causal variant risk distribution, etc., and which builds on our analysis of real genotype data from Genetic Analysis Workshop 17, is warranted. This exploration is especially needed given recent results yielding moderately sized relative risks, even for rare variants, in practice.

Conclusions

Combined testing approaches offer a general and appealing alternative to individual, gene-based rare variant tests of association which may be optimized only for particular genetic architectures. We have demonstrated that the loss of power from the addition of one or two disparate tests may be offset by improved power for a wider range of genetic architectures. We also identified a particular combined test with good properties. As additional, novel, rare-variant tests are developed they should be evaluated for possible combination with existing tests to yield maximally robust testing approaches.

Acknowledgments

This work was funded by the National Human Genome Research Institute (R15HG006915). We acknowledge the use of the Hope College parallel computing cluster for assistance in data analysis. We also acknowledge funding of Genetic Analysis Workshop 17 (NIH R01 GM031575), and the preparation of the Simulated Exome Data

Set, which was supported in part by NIH R01 MH059490 and used sequencing data from the 1000 Genomes Project.

Conflict of Interest

The authors declare no conflict of interest.

Supplemental materials

Supplemental information is available at the European Journal of Human Genetic website.

References

1 Liu K, Fast S, Zawistowski M, Tintle NL. A geometric framework for evaluating rare variant tests of association. *Genet Epidemiol* 2013; **37**: 345–57.

2 Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet* 2014; **95**: 5–23.

3 Li B, Leal SM. Methods for detecting associations with rare variants for common diseases : application to analysis of sequence data. *Am J Hum Genet* 2008; : 311–321.

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

5 Derkach A, Lawless JF, Sun L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genet Epidemiol* 2013; **37**: 110–21.

6 Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson D a *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012; **91**: 224–37.

7 Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing assocation studies. *Biostatistics* 2012; **13**: 762–775.

8 Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol* 2013; **37**: 334–44.

9 R. 2013.www.r-project.org.

10 Almasy L, Dyer TD, Peralta JM, Kent JW, Charlesworth JC, Curran JE *et al.* Genetic Analysis Workshop 17 mini-exome simulation. *BMC Proc* 2011; **5 Suppl 9**: S2.

11 Feng T, Elston RC, Zhu X. Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genet Epidemiol* 2011; **35**: 398–409.

Figures

Figure 1. Power of combined testing approaches as the correlation between powerful tests increases.

All *k* tests being combined have individual power of 64.5% at a significance level of 0.05. When combining multiple powerful tests, and no tests with low power, the Fisher's method is always more powerful than the *Min(p)* method since all tests contribute to the power of the combined test for Fisher's method, but only a single test contributes to the *Min(p)* approach. As the correlation between powerful tests increases all combined tests converge to the power of a single test (64.5%). In general, combining more powerful tests increases power. Similar patterns are observed with lower significance levels (see *Supplemental Tables* 1a-1c).

Figure 2. Power of combined testing approaches as the number of poorly performing tests increases

Of the *k* tests being combined either 1, 2 or 4 tests are 'good' (having power = 64.5%), while the remainder perform poorly (power = 5%, the type I error rate). When there is only one powerful test, the *Min(p)* method outperforms Fisher's method, but when there are four 'good' (powerful) tests, Fisher's test outperforms *Min(p)*. The breakeven point is shown when there are two good tests and we see that Fisher's is better when there are 10 or fewer tests, but *Min(p)* is better when there are 20 total tests being combined. This figure only illustrates cases where there is no correlation between tests. The impact of correlation between tests can be inferred from Figure 1. Similar patterns are observed with lower significance levels (see *Supplemental Tables* 1a-1c).

Figure 3. Power of single and combined gene-based rare variant tests (32 SNVs)

Power of five different tests (3 individual and 2 combined) in the presence of high percentage of non-causal variants and at a significance level of $1x10^{-4}$. The relative risk of the causal SNVs in the set of 32 SNVs is 2, with 1000 cases and 1000 controls. The combined test using either the *Min(p)* or Fisher's approaches is a robust alternative to individual tests.

Figure 4. Power of single and combined gene-based rare variant tests (64 SNVs)

Power of five different tests (3 individual and 2 combined) in the presence of high percentage of non-causal variants and at a significance level of $1x10^{-4}$. The relative risk of the causal SNVs in the set of 64 SNVs is 2, with 1000 cases and 1000 controls. The combined test using either the *Min(p)* or Fisher's approaches is a robust alternative to individual tests.

Table 1. Overview of combined test rationale and performance

Test	Tests combined	Rationale	Avg. corr.[1]	Observed Strengths	Observed Weaknesses	Overall performance[2]
CT1	L(1), L(2), L(3), L(∞)	Assess robustness against non-causal variants and tradeoff with number of tests combined	0.53	Minimal	Poor performance with risk-reducing variants	Poor (*Min(p)*: 38.6%) (Fisher's: 47.2%)
CT2	J(1), J(2), J(3), J(∞)	Assess robustness against non-causal variants and tradeoff with number of tests combined	0.87	Handles risk reducing variants	Redundant	Good (*Min(p)*: 8.9%) (Fisher's: 7.9%)
CT3	CMC, L(1)	Assess impact of combining highly correlated tests	0.92	Minimal	Poor performance with risk-reducing variants; redundant	Poor (*Min(p)*: 42.7%) (Fisher's: 42.2%)
CT4	SKAT, J(2)	Assess impact of combining highly correlated tests	0.99	Handles risk reducing variants	Redundant	Good (*Min(p)*: 7.4%) (Fisher's: 7.4%)
CT5	SKAT, CMC	Assess a 'standard' combination of tests	0.46	Fairly robust	Lacks robustness to high proportion of non-causal variants	Good (*Min(p)*: 7.5%) (Fisher's: 8.1%)
CT6	L(1), L(2), L(3), L(∞),J(1), J(2), J(3), J(∞)	Assess robustness against non-causal variants and tradeoff with number of tests combined	0.54	Fairly robust	Some poorly performing tests (e.g., length tests) make Fisher's perform suboptimally	Good (*Min(p)*: 5.6%) (Fisher's: 9.2%)
CT7	L(1), L(4), J(1), J(4)	Assess robustness against non-causal variants and tradeoff with number of tests combined	0.50	Fairly robust	Fairly good performance, though Fisher's performs a bit poorer due to length tests	Very good (*Min(p)*: 4.9%) (Fisher's: 6.5%)
CT8	SKAT-O, J(∞)	Assess ability to create a more robust SKAT-O	0.77	More robust to inclusion of noncausal variants	Slightly lower power than SKAT-O when few non-causal variants	Very good (*Min(p)*: 4.6%) (Fisher's: 3.0%)

1. Average pairwise correlation across all pairs of tests in the combined test. See *Supplemental Figure 1* for complete matrix of pairwise correlations.
2. Percent of simulations in which method had at least 5% lower power than other methods.

Table 2. Power of common gene-based rare variant tests and novel combined tests across select settings

			Power									
			Single tests				*Combined tests*					
Total number of variants	Number of risk inc. variants (RR[1])	Number of risk dec. variants (RR[1])	CMC	SKAT	SKAT-O	*J(∞)*	CT6: *Min(p)*	CT6: Fisher's	CT7: *Min(p)*	CT7: Fisher's	CT8: *Min(p)*	CT8: Fisher's
64 variants in the gene	32 (2)	0	**1**	**1**	**1**	**0.99**	**1**	**1**	**1**	**1**	**1**	**1**
	16 (2)	0	0.85	**0.97**	**0.988**	0.926	**0.988**	**0.994**	**0.99**	**0.996**	**0.986**	**0.984**
	8 (2)	0	0.346	**0.786**	**0.774**	0.784	**0.82**	**0.872**	0.808	**0.866**	0.798	**0.83**
	4 (2)	0	0.122	0.416	0.338	**0.548**	**0.506**	**0.52**	0.44	0.428	**0.504**	**0.502**
	2 (2)	0	0.086	0.368	0.304	**0.5**	**0.454**	**0.462**	**0.424**	0.408	**0.472**	**0.466**
	1 (2)	0	0.1	0.392	0.322	**0.514**	0.488	**0.53**	0.442	0.432	**0.478**	**0.496**
32 variants in the gene	16 (2)	0	**0.986**	**0.99**	**0.996**	0.944	**0.996**	**0.998**	**0.994**	**0.996**	**0.994**	**0.996**
	8 (2)	0	0.632	**0.8**	**0.816**	0.772	**0.83**	**0.88**	0.82	**0.872**	0.812	0.824
	4 (2)	0	0.206	0.546	0.526	**0.602**	**0.602**	**0.622**	**0.574**	**0.586**	**0.584**	**0.596**
	2 (2)	0	0.194	0.49	0.446	**0.59**	0.51	**0.56**	0.51	0.504	**0.548**	**0.574**
	1 (2)	0	0.132	0.494	0.426	**0.58**	**0.554**	**0.58**	0.52	0.528	**0.536**	**0.558**
32 variants in the gene	24 (1.1)	8 (0.5)	0.066	**0.322**	0.266	0.22	**0.256**	0.214	**0.268**	0.212	**0.266**	**0.29**
	18 (1.1)	6 (0.5)	0.082	**0.336**	0.25	0.234	**0.234**	0.188	**0.242**	0.2	**0.248**	**0.288**
	12 (1.1)	4 (0.5)	0.05	**0.184**	**0.134**	**0.134**	**0.122**	**0.114**	**0.126**	0.108	**0.154**	**0.164**
	6 (1.1)	2 (0.5)	0.074	**0.16**	**0.12**	**0.152**	**0.108**	**0.104**	**0.118**	**0.108**	**0.126**	**0.148**
	3 (1.1)	1 (0.5)	0.074	**0.152**	**0.126**	**0.144**	**0.132**	**0.13**	**0.136**	**0.122**	**0.138**	**0.146**
64 variants in the gene	48 (1.1)	16 (0.5)	0.108	**0.518**	0.418	0.248	**0.348**	0.302	**0.362**	**0.35**	**0.364**	**0.392**
	36 (1.1)	12 (0.5)	0.068	**0.404**	0.322	0.24	**0.3**	**0.27**	**0.312**	**0.288**	**0.302**	**0.314**
	24 (1.1)	8 (0.5)	0.052	**0.254**	**0.208**	0.19	**0.2**	0.162	**0.204**	0.174	**0.214**	**0.232**
	12 (1.1)	4 (0.5)	0.048	**0.134**	**0.116**	**0.116**	**0.1**	**0.11**	**0.102**	**0.108**	**0.126**	**0.132**
	6 (1.1)	2 (0.5)	0.07	**0.086**	**0.07**	**0.086**	**0.078**	**0.078**	**0.072**	**0.08**	**0.07**	**0.088**
	3 (1.1)	1 (0.5)	0.05	**0.1**	**0.078**	**0.11**	**0.078**	**0.078**	**0.08**	**0.066**	**0.092**	**0.104**

Bold indicates tests that are within 5% of optimal for single tests or within 5% of optimal for combined tests.

[1]RR=Relative risk of causal variants

Supplementary Materials

Supplemental Figure 1 – Correlation between p-values of different tests is considered across different genetic architectures. Joint tests and SKAT like tests are highly correlated, as are Length and CMC tests.

Supplemental Figure 2 - Power of five different tests (3 individual and 2 combined) in the presence of high numbers of non-causal variants and at a significance level of $1x10^{-4}$. The relative risk of the causal SNVs in the set of 32 SNVs is 3, with 1000 cases and 1000 controls. The combined test using either the *Min(p)* or Fisher's approaches is a robust alternative to individual tests.

Supplemental Figure 3 - Power of five different tests (3 individual and 2 combined) in the presence of high numbers of non-causal variants and at a significance level of $1x10^{-4}$. The relative risk of the causal SNVs in the set of 64 SNVs is 3, with 1000 cases and 1000 controls. The combined test using either the *Min(p)* or Fisher's approaches is a robust alternative to individual tests.

Supplemental Tables 1a, 1b and 1c – Simulation results for null hypothesis (*Table 1a*), alternative hypothesis (*Table 1b*) and mixed hypothesis (*Table 1c*) situations involving generic combinations of two or more gene-based rare variant tests with different correlations and power and using different test combination strategies, across four different significance levels.

Supplemental Table 2a and 2b – Type I error rates of individual and combined tests across a variety of simulation settings. Type I error rates are generally maintained.

Supplemental Table 3- The power of each combined and individual test for all 197 simulation settings

Supplemental Table 4 – The power of combined and individual tests at lower significance levels

Supplemental Table 5- The p-values of four different gene-based rare variant tests on 25 genes from GAW17