Robust and Powerful Tests for Rare Variants Using Fishers Method to Combine Evidence of Association From Two or More Complementary Tests

ANDRIY DERKACH

Department of Statistical Sciences,

University of Toronto, Toronto, ON, M5S 3G3, Canada

JERRY F. LAWLESS

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

LEI SUN

Division of Biostatistics, Dalla Lana School of Public Health,

University of Toronto, Toronto, ON, M5T 3M7, Canada

E-mail: sun@utstat.toronto.edu

Summary

Many association tests have been proposed for rare variants, but the choice of a powerful test is uncertain when there is limited information on the underlying genetic model. Proposed methods use either linear statistics, which are powerful when most variants are causal and have the same direction of effect, or quadratic statistics, which are more powerful in other scenarios. To achieve robustness, it is natural to combine the evidence of association from two or more complementary tests. To this end, we consider the minimum-p and Fisher's methods of combining P-values from linear and quadratic statistics. Extensive simulation studies show that both methods are robust across models with varying proportions of causal, deleterious, and protective rare variants, allele frequencies, and effect sizes. When the majority ($>$75%) of the causal effects are in the same direction (deleterious or protective), Fisher's method consistently outperforms the minimum-p and the individual linear and quadratic tests, as well as the optimal sequence kernel association test, SKAT-O. When the individual test has moderate power, Fisher's test has improved power for 90% of the $\sim$5000 models considered, with $>$20% relative efficiency gain for 40% of the models. The maximum absolute power loss is 8% for the remaining 10% of the models. An application to the GAW17 quantitative trait Q2 data based on sequence data of the 1000 Genomes Project shows that, compared with linear and quadratic tests, Fisher's test has comparable power for all 13 functional genes and provides the best power for more than half of them.

Keywords: Robust methods; Fisher's method; Rare variants; Complex traits; Next-generation sequencing; 1000 genome project

This is the peer reviewed version of the following article: "Derkach, A., Lawless, J.F. and Sun, L. (2013). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epidemiology*, 37 (1), 110–121", which has been published in final form at DOI: 10.1002/gepi.21689. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

1 INTRODUCTION

Rare variants play an important role in studies of complex human diseases and traits, and next-generation sequencing technology provides rich data for analysis [Cirulli and Goldstein, 2010]. This recent focus on rare variants has produced numerous genotype-phenotype association testing strategies based on aggregating information across multiple SNPs. They include, among others, proposals by Morgenthaler and Thilly [2007], Li and Leal [2008], Madsen and Browning [2009], Bansal et al. [2010], Han and Pan [2010], Hoffmann et al. [2010], Morris and Zeggini [2010], Price et al. [2010], Yi and Zhi [2011], Neale et al. [2011], Wu et al. [2011], Lin and Tang [2011], Lee et al. [2012a]. Basu and Pan [2011] and Derkach et al. [2012] review the many test statistics that have been proposed. Their work shows that the tests can be considered within a unified framework, with methods divided into two classes: tests based on linear composite statistics, which are powerful against very specific alternative hypotheses [e.g., Li and Leal, 2008; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Morris and Zeggini, 2010; Price et al., 2010] and tests based on quadratic statistics, which are designed to have reasonable power across a wide range of alternatives [e.g., Lee et al., 2012a; Neale et al., 2011; Wu et al., 2011]. Several papers have also considered using adaptive weighting of the rare variants under study, based on the observed phenotype and genotype data. For linear statistics [e.g., Han and Pan, 2010; Hoffmann et al., 2010; Lin and Tang, 2011; Yi and Zhi, 2011], it can be shown analytically that these adaptive methods are operationally similar to using quadratic statistics, unless (correct) prior information on SNP effect is available [Derkach et al., 2012]. We note as well that score statistics obtained from random effect regression models lead to quadratic statistics [Basu and Pan, 2011; Goeman et al., 2006].

Tests within the linear class or the quadratic class perform rather similarly, but there are substantial differences in power between linear and quadratic tests [e.g., Basu and Pan, 2011; Derkach et al., 2012; Han and Pan, 2010; Lin and Tang, 2011]. More specifically, linear tests can outperform quadratic tests if all or almost all SNPs under consideration are causal variants and their effects are in the same direction. However, tests based on linear statistics can perform poorly when there are both protective and deleterious SNPs, and more generally, when a substantial portion of the SNPs is neutral. It is becoming evident that "the power of recently proposed statistical methods depend strongly on the underlying hypotheses concerning the relationship of phenotypes with each of these three factors [proportions of causal variants, and direction of the associations (deleterious, protective, or both)]. No method demonstrates consistently acceptable power despite this large sample size, and the performance of each method depends upon the underlying assumption of the relationship between rare variants and complex traits", as concluded by Ladouceur et al. [2012]. Robustness, therefore, is critical and consequential when our knowledge about the genetic architecture of rare variants is still incomplete [Cirulli and Goldstein, 2010].

Basu and Pan [2011] recommended that both linear and quadratic statistics be used in settings where prior information is limited. In this report, we propose hybrid association test statistics that borrow strength from each class of tests by combining them via Fisher's method or the minimum-p approach. We show that both hybrid statistics are robust across genetic models with respect to power, and in some situations Fisher's statistics can outperform linear and quadratic statistics with a relative efficiency gain of more than 100%. The advantages of the proposed methods are demonstrated through extensive simulation

studies of over 10,000 different models with varying proportions of causal, deleterious, and protective rare variants; variant frequencies; effect sizes; and the relationships between variant frequencies and effect sizes, for studies of both binary and quantitative traits, as well as an application to the Genetic Analysis Workshop 17 (GAW17) simulated quantitative trait Q2 data based on the mini-exome sequence data that were provided by the 1000 Genomes Project [Almasy et al., 2011; 1000 Genomes Project Consortium, 2010]. We also compare the two tests with SKAT-O [Lee et al., 2012a], the optimal sequence kernel association test that uses the minimal P-value of a family of tests that are based on weighted averages of a linear statistic and the original SKAT quadratic statistic [Wu et al., 2011], and we show that the proposed hybrid statistics have better power than SKAT-O.

2 Materials and Methods

2.1 Methods

To formulate the testing problem, we assume that a group of J SNPs labeled $j = 1, \ldots, J$ and a (quantitative or binary) phenotype Y for n subjects are under consideration. Let Y_i be the phenotype value, and X_{ij} be the genotype value of the ith subject representing the number of copies of the rare allele for the jth SNP. In practice, X_{ij} = 0 or 1 because of the low frequency of the rare allele. Most statistics [e.g., Lin and Tang, 2011; Neale et al., 2011; Wu et al., 2011] for testing association between Y_i and X_{ij} in the absence of other factors can be written in terms of

$$S_j = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) X_{ij}}{\sqrt{(\sum_{i=1}^n X_{ij})\,(1 - \sum_{i=1}^n X_{ij}/n)}} = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) X_{ij}}{\sqrt{m_j(1 - m_j/n)}}\ , \quad j = 1, \ldots, J\ , \tag{1}$$

where $m_j = \sum_{i=1}^n X_{ij}$ is the total number of copies of the rare allele of SNP j, approximately equal to the number of subjects carrying the rare allele of SNP j. The statistic $S = (S_1, \ldots, S_J)'$ is also the scaled score statistic from linear or logistic regression models relating Y_i and $X_i = (X_{i1}, \ldots, X_{iJ})'$ [e.g., Basu and Pan, 2011; Lin and Tang, 2011; Wu et al., 2011], and it has an expectation of 0 under the null hypothesis of no association, H_0: Y_i is independent of X_i .

Linear statistics that have been proposed take the general form

$$W_L = \sum_{j=1}^{J} w_j S_j = \mathbf{w}'\mathbf{S}\ . \tag{2}$$

where w_j is a prespecified weight for SNP j and $w = (w_1, \ldots, w_J)'$. Quadratic statistics take the form

$$W_Q = \mathbf{S}' A \mathbf{S}\ , \tag{3}$$

where A is a positive definite (or semidefinite) symmetric matrix [Basu and Pan, 2011; Derkach et al., 2012; Lin and Tang, 2011; Wu et al., 2011].

Let p L be the two-sided P-value obtained from W_L to allow for either positive or negative association statistic under the alternative hypothesis. Let p_Q be the P-value obtained from W_Q as $Prob(W_Q \geq$ observed value) under H_0, because for quadratic statistics only large values of W_Q provide evidence against H_0. To combine information from the linear and quadratic statistics, we propose to use Fisher's method to combine P-values from W_L and W_Q, using

$$W_F = -2\log(p_L) - 2\log(p_Q) \tag{4}$$

as the association test statistic. Large values of W_F correspond to small values of p_L and/or p_Q and indicate evidence against the null hypothesis of no association. If p_L and p_Q are independent under H_0, then W_F is distributed as χ^2_4. However, W_L and W_Q (thus p_L and p_Q) are not independent except asymptotically when $J \to \infty$. Simulations described below and Supporting Information Tables S1 and S2 show that the χ^2_4 approximation is inadequate for realistic settings. Thus we assess statistical significance in finite samples using a novel permutation distribution approach described in the Appendix.

Another way to combine evidence from two or more tests is through the minimum-p approach. Here we consider

$$W_M = \min(p_L, p_Q) \ . \tag{5}$$

The minimum-p principle has been proposed by many authors, [e.g., Lee et al., 2012a; Lin and Tang, 2011], each considering a different set of tests. For example, the recent SKAT-O statistic [Lee et al., 2012a] is the minimal P-value of a family of tests that are based on weighted averages of a linear statistic and the quadratic SKAT statistic, with weights ranging from 0 (using the quadratic statistic only) to 1 (using the linear statistic only). EREC [Lin and Tang, 2011] can be considered as a special case of SKAT-O. We focus on the simple minimum-p statistic of (5) and the Fisher's statistics, but we compare them with SKAT-O in the Discussion section below and in the Supporting Information.

For symmetric case-control studies or studies of normally distributed traits, the statistic W_M is asymptotically distributed as $J \to \infty$ under H_0 as $\min\{U_1, U_2\}$, where U_1 and U_2 are two independent Unif(0,1) variables. As with W_F , asymptotic approximations may not provide satisfactory P-values in many practical situations, and once again we rely on permutation-based P-values.

The general problem of combining information from test statistics has been studied by several authors [e.g., Loughin, 2004; Owen, 2009; Stouffer, 1949]. They show that no single approach is best (most powerful) under all circumstances. [Owen, 2009] gave a careful comparison of (4) and (5), and found that if one of the original statistics has low power and the other high power, W_M is a better hybrid statistic. On the other hand, if both tests have reasonable power, Fisher's statistic W_F is a better choice.

2.2 Simulation Models

We conducted extensive simulation studies to examine the finite sample performance of linear, quadratic, minimum-p, and Fishers test statistics. We considered association studies for both quantitative and binary traits. Here, we focus on quantitative traits for which simulations can be conducted efficiently to study a large number (over 10,000) of genetic models. Results from case-control studies are provided as Supporting Information and are discussed below in the final Discussion section.

To provide numerical comparisons of power, we consider for simplicity the case where X_j indicates the presence (1) or absence (0) of the minor allele, and let

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_J X_{iJ} + e_i \ , \ \text{ for } \ i = 1, \ldots, n \ , \tag{6}$$

with $e_i \sim N(0, \sigma^2)$, so the hypothesis of no association H_0 becomes $\beta = (\beta_1, \ldots, \beta_J)' = \mathbf{0}$. We first assume that the X_{ij}'s are mutually independent Bernoulli variables with $P(X_{ij} = 1) = p_j$, approximately twice the minor allele frequency (MAF) of SNP j, for $j = 1, \ldots, J$. We later consider X_{ij} obtained from the sequence data of the 1000 Genomes Project [1000 Genomes Project Consortium, 2010], which are not mutually independent.

The specific linear and quadratic statistics considered here are W_L in (2) with $\mathbf{w} = \mathbf{1}$,

$$W_L = \mathbf{1}'\mathbf{S} = \sum_{j=1}^{J} S_j \ , \tag{7}$$

and W_Q in (3) with $A = I$,

$$W_Q = \mathbf{S}'\mathbf{S} = \sum_{j=1}^{J} S_j^2 \ , \tag{8}$$

where S_j is defined in (1). Because S_j is an MAF-scaled score statistic, the linear statistic (7) is the same as the MAF-weighted linear statistic proposed by [Madsen and Browning, 2009] and the quadratic statistic (8) is equivalent to Hotelling's statistic [Derkach et al., 2012]. Although other linear and quadratic statistics could be considered, we focus on W_L and W_Q because within-class difference in power is substantially smaller than the between-class difference [Basu and Pan, 2011; Derkach et al., 2012; Lin and Tang, 2011], and the latter is the subject of interest here. Moreover, extensive simulation results in Basu and Pan [2011] show that W_L and W_Q have good power among the classes of linear and quadratic statistics, respectively.

Under the normal model (6), the work of Derkach et al. [2012] shows that, when m_j in (1) equals its expected value np_j,

$$S_j \sim N(\mu_j, \sigma^2) \ ,$$
$$\mu_j = \sqrt{n}\beta_j\sqrt{p_j(1-p_j)} \ ,$$

and the variance explained by each SNP j is approximately

$$EV_j = \frac{\left(\beta\sqrt{p_j(1-p_j)}\right)^2}{\sigma^2}$$

Moreover, W_Q has a noncentral χ^2_{J,nc_Q} distribution under an alternative H_1 for which $\beta \neq \mathbf{0}$ and the power of W_Q is a function of the noncentrality parameter, which is

$$nc_Q = \sum_j \frac{\mu_j^2}{\sigma^2} = n\sum_j EV_j \ .$$

W_L^2 has a noncentral χ^2_{1,nC_L} distribution under H_1 and

$$nc_L = \frac{\left(\sum_j \mu_j\right)^2}{\sum_j \sigma^2} = n\frac{\left(\sum_j sign(\beta_j)\sqrt{EV_j}\right)^2}{J} \ .$$

Therefore, the power of both statistics depends only on sample size n, the number of SNPs J, variance explained by each SNP EV_j, its direction of effect $sign(\beta_j)$ (for linear statistics only), and type 1 error α. It is important to note that effect size β_j, MAF $p_j/2$, and the total phenotypic variance σ^2 do not directly affect power once EV_j is known or specified. However, also note that $EV_j = (\beta_j\sqrt{p_j(1-p_j)})^2/\sigma^2$, and therefore the models considered here implicitly assume that, for a specified EV_j level, variants with smaller MAFs tend to have bigger effect sizes (i.e., the "MAF-effect-dependent" model assumption). Later we discuss additional simulation studies assuming that MAFs and effect sizes are mutually independent (i.e., the "MAF-effect-independent" model assumption).

To evaluate the accuracy of the asymptotic null distributions of the proposed hybrid statistics, W_F and W_M, in finite samples in terms of J, we considered $J = 10, 20, 30, 40, 50$, or 100 and $EV_j \equiv 0$, for all $j = 1, \ldots, J$. (Sample size n is not a factor under the normal assumption and under H_0.) For each

combination, we generated $\mathbf{S} = (S_1, \ldots, S_J)'$ from $\mathbf{S} = N(0, \sigma^2 I)$ independently 10^6 times. Without loss of generality, we assumed $\sigma^2 = 1$ because σ^2 does not affect the size or power of the tests once EV_j is specified. For each simulated replicate, we calculated p_L for W_L based on $N(0,1)$ and p_Q for W_Q based on χ^2_J, then combined the two P-values to obtain W_F and W_M. Finally, we calculate p_F for W_F based on χ^2_4 and p_M for W_M based on $\min(U_1, U_2)$, with U_1 and U_2 assumed independent. Note that the χ^2_4 approximation for W_F and $\min(U_1, U_2)$ for W_M are only used here for assessing the accuracy of the asymptotic null distributions. For all power comparisons below, we used empirical critical values that are based on the results from this study for quantitative traits or on the permutation method (Appendix) for case-control and application studies.

Table 1: Parameters and parameter values of simulated models for studies of quantitative or binary traits. The MAF-effect-dependent model assumes that variants with smaller MAFs tend to have bigger effect sizes; the MAF-effect-independent model assumes that MAF and effect sizes are mutually independent

	Parameters	Parameter Values
n	sample size ($n_{case} = n_{control} = n/2$ for binary traits)	500, 1000 or 2000
J	total number of SNPs	Unif {10, 20, 30, 40, 50}
p_C	proportion of the causal SNPs	Unif (0.1, 1)
J_C	number of the causal SNPs, an integer closest to $J \cdot p_C$	
p_D	proportion of the deleterious SNPs among the causal ones	Unif (0.75, 1)
J_D	number of the deleterious SNPs, an integer closest to $J_C \cdot p_D$	
p_P	proportion of the protective SNPs among the casual ones, $1 - p_D$	
J_P	number of the protective SNPs, $J_C - J_D$	
p_N	proportion of the neutral SNPs, $1 - p_C$	
J_N	number of the neutral SNPs, $J - J_D - J_P$	
	Quantitative traits under the MAF-effect dependent assumption; 10,000 independently simulated models	
EV_j	the variance explained by SNP j ($EV_j = \beta_j^2 p_j(1-p_j)$)	
	for neutral SNPs	0
	for causal SNPs	Unif (0.001, 0.0025)
	Quantitative traits under the MAF-effect independent assumption; 10,000 independently simulated models	
p_j	approximately twice the MAF of SNP j	Unif (0.005, 0.02)
β_j	regression coefficient in (6) of SNP j	
	for neutral SNPs	0
	for causal SNPs	Unif (0.45, 0.5) or Unif (-0.5, -0.45)
		(The resulting EV_js in the range 0.001 to 0.0049)
	Binary traits under the MAF-effect dependent assumption; 500 independently simulated models	
p_j	approximately twice the MAF of SNP j	Unif (0.005, 0.02)
e_j^β	OR of SNP j	
	for neutral SNPs	1
	for causal SNPs	$C/\sqrt{p_j(1-p_j)}, C = 4\sqrt{0.005(1-0.005)}$
		(The resulting ORs in the range 2 (or 1/2) to 4 (or 1/4)
	Binary traits under the MAF-effect independent assumption; 500 independently simulated models	
p_j	approximately twice the MAF of SNP j	Unif (0.005, 0.02)
e_j^β	OR of SNP j	
	for neutral SNPs	1
	for causal SNPs	Unif (2, 4) or Unif (1/2, 1/4)

To evaluate power of the statistics under a broad range of scenarios, we independently generated 10,000 different models as described in Table 1 for studies of quantitative traits under the MAF-effect-dependent assumption. For each combination of parameter values (i.e., one of the 10,000 models), we generated S_j from $N(0, \sigma^2)$ for J_N neutral variants, from $N(\sqrt{n}\sqrt{EV_j}, \sigma^2)$ for J_D deleterious variants, and from $N(-\sqrt{n}\sqrt{EV_j}, \sigma^2)$ for J_P protective variants, independently 10,000 times (i.e., 10,000 data replicates for each of the 10,000 models). We calculated p_L for W_L based on $N(0,1)$, p_Q for W_Q based on χ^2_J, then combined the two P-values to obtain W_F and W_M. We then determined if p_F for W_F and p_M for W_M were less than a given α value by comparing W_F and W_M with the empirical critical values

from Supporting Information Tables S2 and S4, respectively. This ensures that the tests have the correct type 1 error α. Finally, we estimated power by the proportion of the 10,000 data replicates that had p_L, p_Q, p_M, and p_F less than α, respectively, for W_L, W_Q, W_M, and W_F.

2.3 Application Data

Similar to many earlier studies of rare variants, the simulating models considered so far assumed that genotypes of a group of rare variants X_j, $j = 1, \ldots, J$, are mutually independent, although the tests themselves do not require this. One rationale is that rare variants act independently. The independence assumption also allows for evaluation of a large number of different models, as well as systematic presentation and understanding of the results. However, the general conclusions made so far are not affected by the independence assumption. As a proof of principle, we also analyzed the GAW17 data for which multiple phenotypes were simulated based on the "mini-exome" sequence data provided by the 1000 Genomes Project [Almasy et al., 2011; 1000 Genomes Project Consortium, 2010].

We analyzed quantitative trait Q2 that is influenced by 72 SNPs in 13 genes but not by other covariates [Almasy et al., 2011]. We used data from the $n = 321$ unrelated Asian subjects (Han Chinese, Denver Chinese, and Japanese). Because we excluded SNPs that had MAF >5% or were monomorphic within the Asian sample, VNN1 had no causal rare variant but it was kept in the analysis to serve as a negative control. The choice to focus on SNPs with MAF $\leq$ 5% was made because this thresholding (almost) does not affect the number of causal SNPs (70 of the 72 causal SNPs have MAF $\leq$ 5% in the range of (0.16.1.4%)), but it reduces the number of neutral SNPs in a gene, so that the proportion of the causal variants is high enough to have meaningful power comparisons for at least some of the 13 genes (Table 2).

The GAW17 data include 200 replicates (same genotype data but different phenotype data independently simulated based on the true genotype-phenotype association model), and for each, we calculated permutation-based P-values for the four tests using the method described in the Appendix, based on 104 permutations. We estimated power for $\alpha = 0.05$ by the proportion of the 200 replicates for which the empirical P-values are ≤ 0.05 for each test. The choice of the liberal type 1 error level 0.05 was based on the overall low power of detecting these genes due to small sample size, small genetic effect, extremely small MAF, or the low proportion of the causal variants within a gene. Because power estimated from 200 replicates is highly variable, comparisons should be focused on the first group of eight genes for which the maximum power is 10% or more.

3 Results

3.1 Simulation Results

The empirical type 1 error rates for W_L and W_Q, as expected, were very close to the nominal level because of the assumption of normality (results not shown). Therefore, p_L and p_Q used to obtain the W_F and W_M statistics were "honest" P-values. However, W_F has a slight inflation of type 1 error around 0.06 for $\alpha = 0.05$, and a large inflation of $5 \cdot 10^{-4}$ for $\alpha = 10^{-4}$, worse for smaller J, and better for bigger J as expected (Supporting Information Table S1). This reflects the nonindependency of W_L and W_Q under H_0 with small J. The empirical type 1 error rates for W_M were consistent with the nominal levels considered (Supporting Information Table S3). However, this does not hold in general, e.g., for non-normally distributed traits with small sample sizes or case-control studies. Therefore, in practice P-values should be obtained empirically. The Appendix provides an efficient permutation scheme that

Table 2: Power of the four test statistics applied to the GAW17 sequence data provided by the 1000 Genomes Project. The four statistics are linear W_L in (7), quadratic W_Q in (8), minimum-p W_M in (5), and Fisher's W_F in (4). The 13 genes presented here are all the causal genes for simulated quantitative trait Q2. VNN1 does not have causal variants because one of the two causal variants has MAF 26% and the other is not polymorphic within the Asian sample ($n = 321$). VNN1 is kept in the analysis to serve as a negative control. All causal variants were designed by GAW17 to have the same direction of effects (minor alleles were associated with higher Q2 values). The average genetic effect is the average of regression coefficient β values of the causal variants used to simulate Q2 (effects are independent of populations by the GAW17 design). Genes are ordered according to the maximum power of Fisher's test. Powers shown vary considerably due to inherent factors and estimation based only on 200 replicates, and the 13 genes are separated into different groups

	SNP Distribution	Ave. MAF of	Avg. Effect of	Power of the Four Tests			
Gene	J_C, J_N	J_C, J_N	J_C	Linear	Quadratic	Minimum-p	Fisher's
8 genes for which the maximum power is 10% or more							
SIRT1	4, 7	0.27%, 0.22%	0.71	0.44	0.39	0.43	0.50
BCHE	5, 10	0.22%, 0.19%	0.72	0.29	0.39	0.39	0.45
PDGFD	3, 6	0.78%, 0.65%	0.74	0.29	0.35	0.38	0.43
SREBF1	4, 5	0.39%, 0.40%	0.52	0.29	0.15	0.24	0.26
GCKR	1, 0	1.21%, NA	0.38	0.25	0.25	0.25	0.25
VLDLR	4, 6	0.19%, 1.64%	0.75	0.12	0.09	0.12	0.13
PLAT	4, 7	0.39%, 0.49%	0.68	0.13	0.13	0.11	0.13
RARB	1, 5	0.78%, 0.90%	0.64	0.06	0.14	0.12	0.11
4 genes for which the maximum power is 10% or less							
INSIG1	3, 1	0.16%, 3.42%	0.20	0.06	0.03	0.03	0.05
VNN3	2, 2	0.16%, 2.57%	0.37	0.03	0.04	0.04	0.04
LPL	1, 4	0.16%, 0.23%	0.73	0.02	0.05	0.03	0.03
VWF	1, 3	0.16%, 1.90%	0.34	0.02	0.01	0.01	0.02
1 gene for which there is no polymorphic rare causal variants in the Asian sample							
VNN1	0, 3	NA, 0.31%	NA	0.02	0.05	0.04	0.05

provides correct P-values for W_L, W_Q, W_F, and W_M simultaneously, which is used for our simulation and application studies.

Figure 1 shows the empirical power of the four test statistics compared to the maximum power. Sample size is 1000 and type 1 error is $\alpha = 10^{-4}$. (Results for n = 500 and 2000 at $\alpha = 10^{-4}$ are in Supporting Information Figs. S2 and S3, respectively and are characteristically similar; results for other α levels, 0.05, 10^{-2}, and 10^{-3} are also similar and not shown.) Several observations can be made from Figure 1.

- The maximum power is often achieved by the Fisher's test: this occurs in 75% of the 10,000 simulated models.
- Both linear and quadratic tests have large variability in power reflecting the wide variation in the simulated models. Power of the linear test is more than 5% below the maximum power for 52% of the 10,000 models; when the maximum power is around 60%, power of the linear test can be as low as 15%. Power of the quadratic test is more than 5% below the maximum power for 53% of the 10,000 models; when the maximum power is around 60%, power of the quadratic test can be as low as 20%.
- Both Fisher's and minimum-p test statistics are robust in terms of power. However, Fisher's test consistently outperforms the minimum-p test, and it can have substantially better power than the individual linear and quadratic tests.
- When either the linear or quadratic test has moderate power, Fisher's test has improved power. For example, among the 10,000 models simulated, the power of linear or quadratic test is at least 20% for 4,903 models, and Fisher's test has improved power for 90% of the 4,903 models. The relative efficiency gain is at least 20% for 40% of the 4,903 models (and at least 50% for 10% of the 4,903 models). Among the 380 of the 4,903 models for which Fisher's test has less power, the maximum absolute power loss is 8%.

To better understand the impact of the various parameters, Figure 2 presents the same results from a different perspective showing the individual power as a function of the number of causal variants J_C (large scale of the X-axis) and the number of deleterious variants J_D (small scale of the X-axis), when the total number of rare variants is $J = 30$ (see Supporting Information Figs. S1a-S1d for J = 10, 20, 40, and 50). It is clear that power of all tests highly depend on the percentage of causal SNPs in the group of SNPs investigated. For example, among the 10,000 models simulated, to achieve power of 50% or greater, the average proportion of causal SNPs is 81% (SE = 13%, min = 42%) for the linear test, 81% (SE = 12%, min = 50%) for the quadratic test, 80% (SE = 14%, min = 42%) for the minimum-p test, and 77% (SE = 15%, min = 36%) for Fisher's test. For the 2005 models with $J = 30$ shown in Figure 2, a proportion of 80% being causal means $J_C = J \cdot p_C = 24$ on the large scale of the X-axis.

To further demonstrate the differential consequences of effect directions on different tests, Figure 3 shows the individual power for the 75 models that have $J_C = 24$ casual variants of the $J = 30$ total variants. The X-axis in Figure 3 shows the number of deleterious variants out of the 24 casual ones, ranging from $J_D = J_C \cdot p_D = 24 \cdot 75\% = 18$ to $J_D = 24 \cdot 100\% = 24$. Although the linear test can outperform the quadratic test by a large margin for some models, it is highly sensitive to the direction of effects. For example, for models in Figure 3 where all 24 causal SNPs are deleterious ($J_D = 24$, $J_P = 0$), power of the linear test is over 90% compared to $\sim$60% for the quadratic test (power of the minimum-p and Fisher's tests are also over 90%). However, if 4 of the 24 causal SNPs are protective

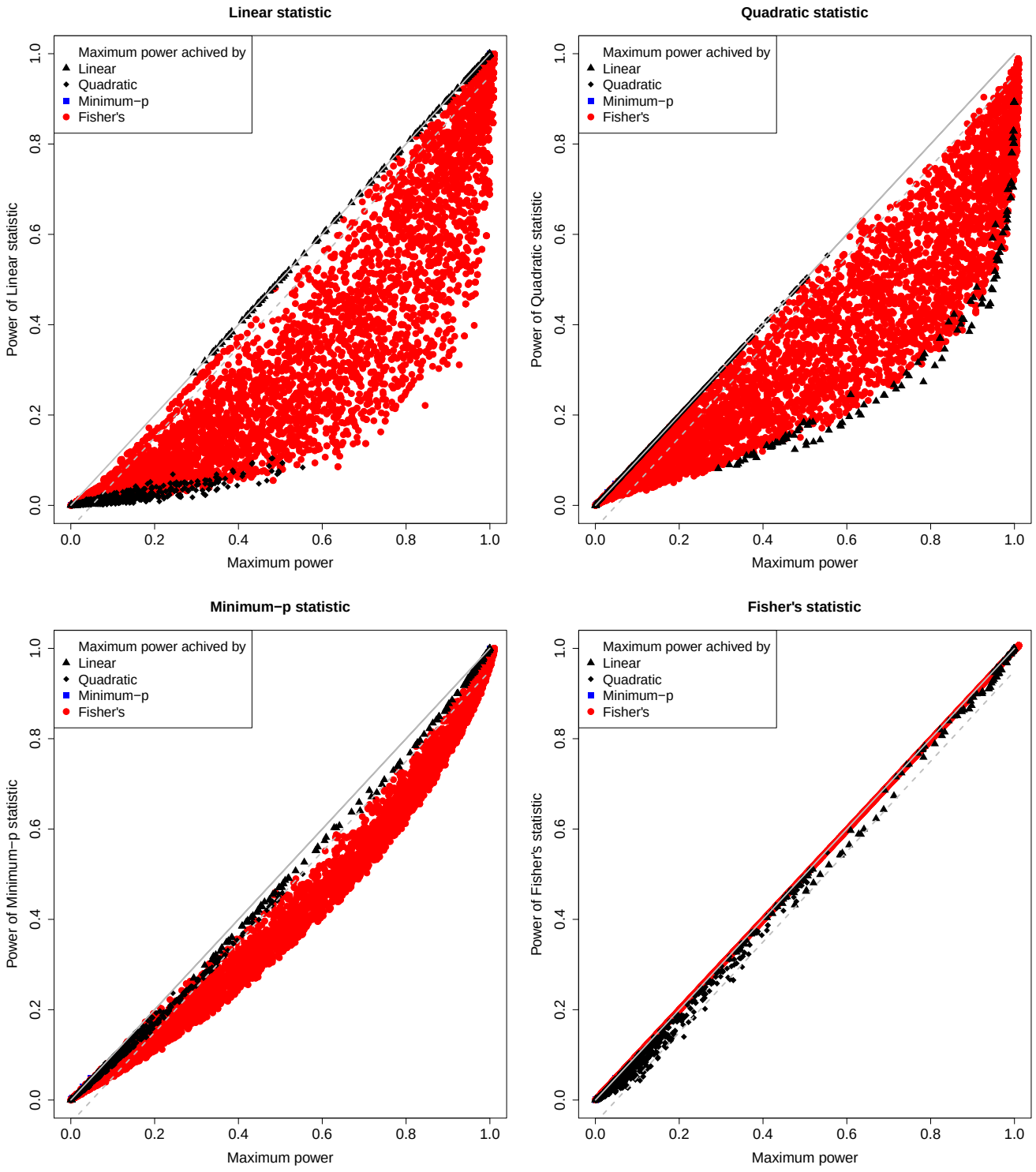

Figure 1: Empirical power of the four test statistics compared to the maximum power for 10,000 independently generated models for studies of quantitative traits under the MAF-effect-dependent assumption as described in Table 1. The four statistics are linear W_L in (7), quadratic W_Q in (8), minimum-p W_M in (5), and Fisher's W_F in (4). For each genetic model, the maximum power among the four statistics and the statistic that provides the maximum power are recorded (black triangle for W_L , black diamond for W_Q, blue square for W_M, and red circle for W_F), and it is compared with power of individual statistics. Sample size n = 1000 and type 1 error $\alpha = 10^{-4}$. Results for n = 500 and 2000 are in Supporting Information Figures. S2 and S3, respectively.

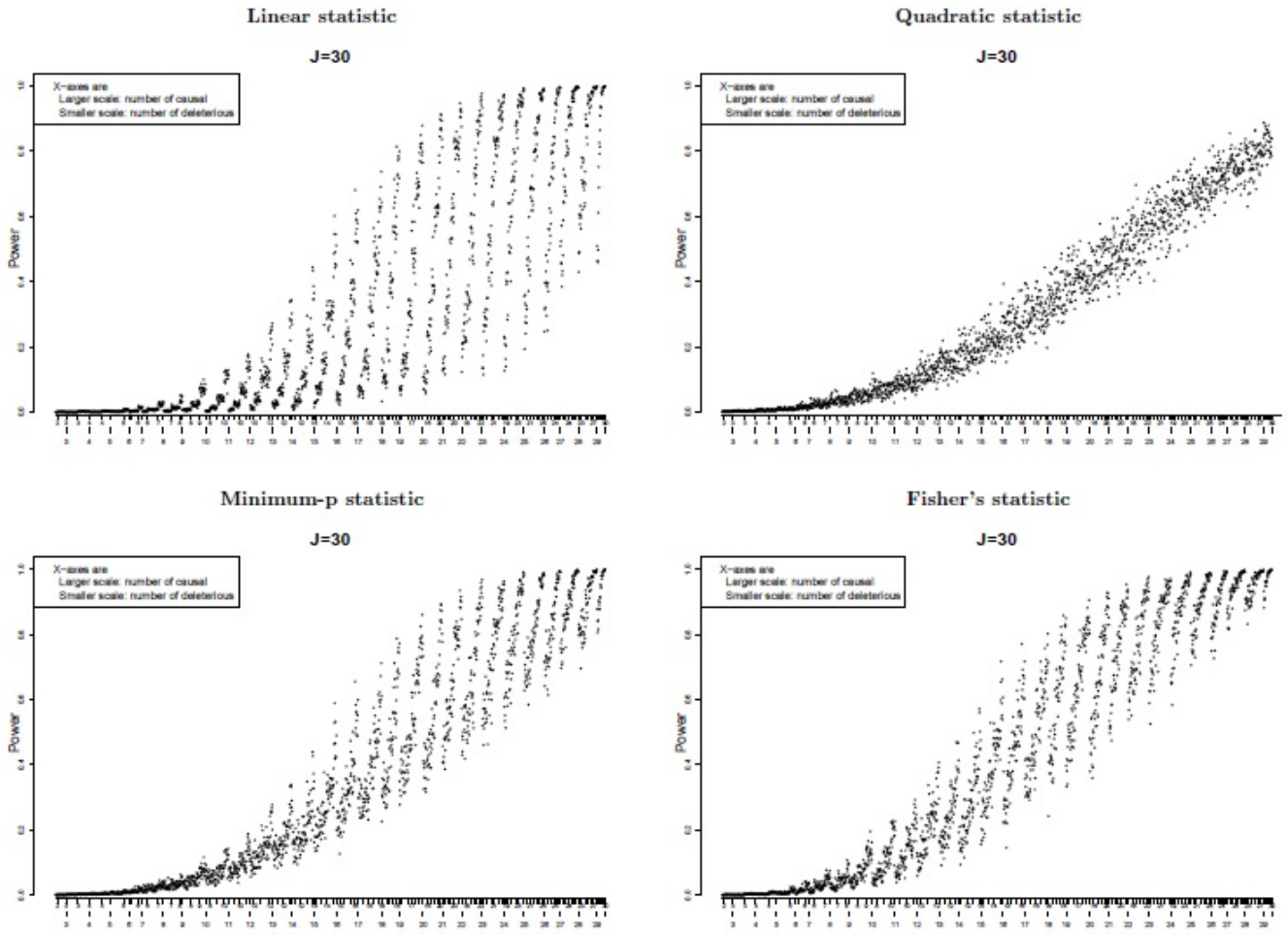

Figure 2: Empirical individual power of the four test statistics for the 2005 of the 10,000 models in Figure 1 with $J = 30$ total number of rare variants. The large scale of the X-axis shows the number of causal variants in the range of $J_C = J \cdot p_C = 30 \cdot 10\% = 3$ to $J_C = 30 \cdot 100\% = 30$. The small scale of the X-axis shows the number of deleterious variants J_D out of the total of J_C causal variants in the range of $J_D = J_C \cdot p_C = J_c \cdot 75\%$ to $J_D = J_C \cdot 100\%$, depending on the actual number of causal variants in a model. The 2005 models are a subset of the 10,000 models generated as described in Table 1 and Figure 1. Results of other J values are in Supporting Information Figures. S1a–S1.

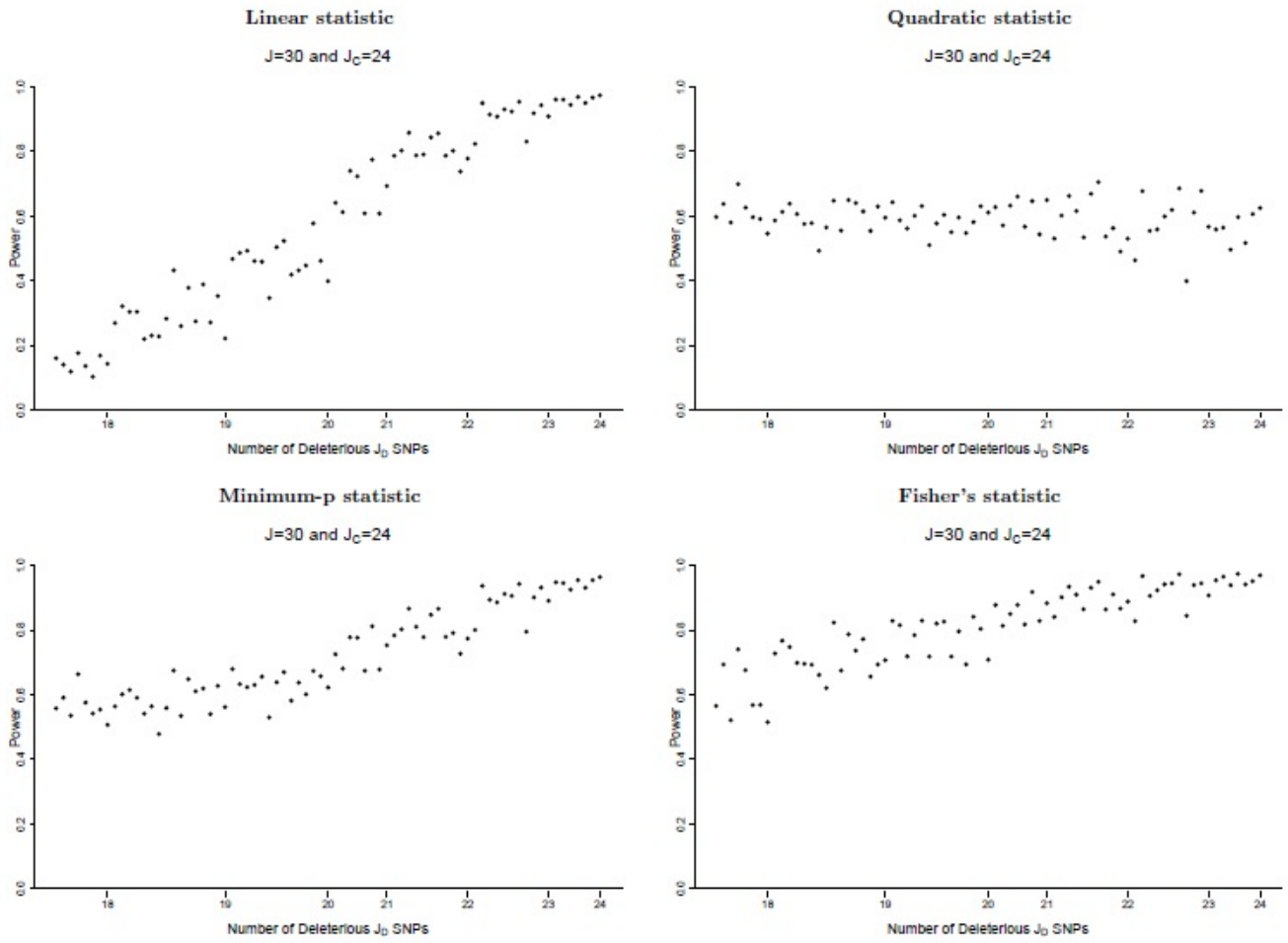

Figure 3: Empirical individual power of the four test statistics for the 75 of the 2005 models in Figure 2 with $J_C = 24$ causal variants out of total $J = 30$ rare variants. The X-axis shows the number of deleterious variants J_D out of the total of $J_C = 24$ causal variants in the range of $J_D = J_C \cdot p_D = 24 \cdot 75\% = 18$ to $J_D = 24 \cdot 100\% = 24$. Other details are Figures. 1 and 2 and Table 1.

($J_D = 20$, $J_P = 4$), power for the linear test drops to ∼40% while power of the quadratic test remains at ∼60% (power of the minimum-p is ∼60% and power of Fisher's test is ∼70%). Both the minimum-p and Fisher's tests are robust because they combine information from the complementary linear and quadratic tests, and the relationship between their power and the various parameters is similar. One noticeable difference is that power of the minimum-p test is constrained between power of the linear and quadratic tests, while Fisher's test does not have such a restriction.

3.2 Application Results

Results in Table 2 are consistent with our previous conclusions. (1). Performance of individual linear and quadratic tests can be highly variable depending on the model. For example, power of the linear and quadratic tests for SIRT1 are 44% and 39%, respectively, while power of the two tests for BCHE are 29% and 39%, respectively. The example of BCHE also shows that even when all causal variants have the same direction of effect, quadratic tests can outperform linear tests if the proportion of the causal variants is low (5 out 15 for BCHE). (2). The minimum-p and Fisher's hybrid statistics are both robust, and Fisher's test consistently outperforms the minimum-p test. (3). Fisher's test not only provides comparable power for all genes analyzed but often it is the most powerful test, with appreciable power; e.g., the power of Fisher's test is 50% for SIRT1 and 45% for BCHE.

Some authors have reported problems related to the GAW17 data and some of the published analyses. For example, Tintle et al. [2011] noted Two main causes emerged: population stratification and long-range correlation (gametic phase disequilibrium) between rare variants. These issues however do not affect our analyses, because we used the samples from the Asian population only and we assessed the statistical significance of the tests using a permutation-based method as described in the Appendix.

4 Discussion

As discussed above, the genetic models considered in Table 1 and Figure 1 directly specify EV_j, the variance explained by a rare variant, which implicitly assumes that rarer variants have bigger genetic effects. We also investigated models where MAFs and genetic effects are independent of each other (Table 1; Quantitative traits under the MAF-effect-independent assumption). Results are presented in Supporting Information Figure S4 and are very similar to those in Figure 1.

We also conducted extensive simulation studies of case-control studies. Briefly, the distribution of Y_i given X_i is Bernoulli with

$$Prob(Y_1 = 1|X_i) = \frac{e^{\beta_0+\sum_j \beta_j X_{ij}}}{1 + e^{\beta_0+\sum_j \beta_j X_{ij}}} ,$$

where X_{ij} are mutually independent Bernoulli variables as in the quantitative setting. Without loss of generality, $\beta_0 = -2.1922$ so that $Prob(Y_1 = 1|X_i = 0) = 0.1$. Other parameters are described in Table 1, separately under the MAF-effect-dependent or -independent assumption. For each combination of parameter values, a sample was obtained by first generating genotype X_i for each subject. The case ($Y = 1$) and control ($Y = 0$) status were assigned based on the probabilities from the logistic model, $Prob(Y_i = 1|X_i)$, allowing for the case-control design. This was done independently 1000 times to estimate power of the four tests for each of the 500 models independently generated. Due to non-normality, p_L and p_Q, P-values of the linear and quadratic statistics, were also obtained empirically via 106 permutations (see the Appendix). Results are in Supporting Information Figure S5 (MAF-effect dependent) and Supporting Information Figure S6 (MAF-effect independent), and they are similar to each other and similar to those in Figure 1 for quantitative traits.

The models considered so far do not restrict all causal variants to have the same direction of effect, but do assume the majority of the causal variants have the same direction (without loss of generality, deleterious) with $p_D = J_D/J_C \sim \text{Unif}(0.75, 1)$. Although this is more plausible than the scenario when deleterious and protective variants are equally likely among the causal ones, we also investigated models for which $p_D \sim \text{Unif}(0.5, 0.75)$; all other parameters were generated as described in Table 1. For such models, the linear test has little power in most cases while the quadratic test has much better power, as expected. Results of quantitative traits under the MAF-effect-dependent assumption are in Figure 4 (results of other types of studies as described in Table 1 are characteristically similar and not shown). The maximum power was achieved by the quadratic test in 94% of 10,000 simulated models; for the remaining 6% of the models the maximum power was achieved by Fisher's test. Consequently, although both minimum-p and Fisher's tests are reasonably robust, the minimum-p statistic is close to best statistic for each model and is a better hybrid statistic, consistent with the findings of Owen [2009].

In some settings, a test of no association may be based on a regression model with several environmental or population stratification covariates [e.g., Lin and Tang, 2011]. Because adjusting for covariates is performed at the individual linear and quadratic test statistic level, the calculation of the proposed hybrid statistics remains the same as in (5) for the minimum-p statistic and in (4) for Fisher's statistic. However, covariates adjustment could affect the computation of P-values for the hybrid tests. Simple permutation procedures are not valid unless SNP genotypes are independent of both the response Y and covariates. Several authors [e.g., Lee et al., 2012a; Lin and Tang, 2011] have proposed parametric bootstrap to obtain P-values in the presence of covariates. The parametric bootstrap approach combined with the joint resampling methodology as discussed in the Appendix can be used for our hybrid test statistics. For large sample size, an alternative for obtaining valid P-values is numerical calculation. In that case, vector $\mathbf{S}$ is (approximately) distributed as multivariate normal and hence can be generated. Further research is needed on robust ways to obtain P-values in the presence of covariates.

The goal of this study is to show that combing evidence of association from complementary linear and quadratic tests can lead to robust and more powerful tests. As a proof of principle, we used the MAF-weighted linear statistic W_L in (7) and Hotelling's statistic W_Q in (8), and we assumed that there are no other influencing factors. However, the concept can be extended to any two or more complementary tests. The power of such hybrid statistics depends on the power of the original individual tests and the dependency between the tests under the null hypothesis of no association H_0. An interesting question is whether one could further improve robustness by combining the P-values from the minimum-p and Fisher's tests, however, this is beyond the scope of this work.

Recently, Lee et al. [2012a] developed an optimal sequence kernel association test, SKAT-O, extending the work of Wu et al. [2011] that proposed the quadratic SKAT test. The SKAT-O uses the minimum-p principle by considering a family of tests based on weighted averages of a linear statistic and SKAT. We evaluated the empirical power of SKAT-O using the existing R-package [Lee et al., 2012b] under the various scenarios as outlined in Table 1. In the case of quantitative traits, we observed the Fisher's statistic performs better than SKAT-O, when the proportion of deleterious SNPs is from Unif(0.75, 1) (Supporting Information Fig. S7, $n = 1000$, $\alpha = 10^{-4}$). Similar results were also seen in the case-control study where we used SKAT-O with suggested adjustment for small sample size (Supporting Information Fig. S9, $n_{cases} = n_{controls} = 500$, $\alpha = 10^{-4}$). However, when the proportion of deleterious SNPs is from Unif (0.5,0.75), the proposed minimum-p statistic outperforms SKAT-O and Fisher's statistic because the latter two statistics lose power when one of the tests has little to no power (Supporting Information Fig. S8 for quantitative traits and Supporting Information Fig. S10 for binary traits). We also observed that SKAT-O has a slight inflated empirical type 1 error, therefore results for SKAT-O presented here are a little too optimistic. Nevertheless, the general conclusions hold.

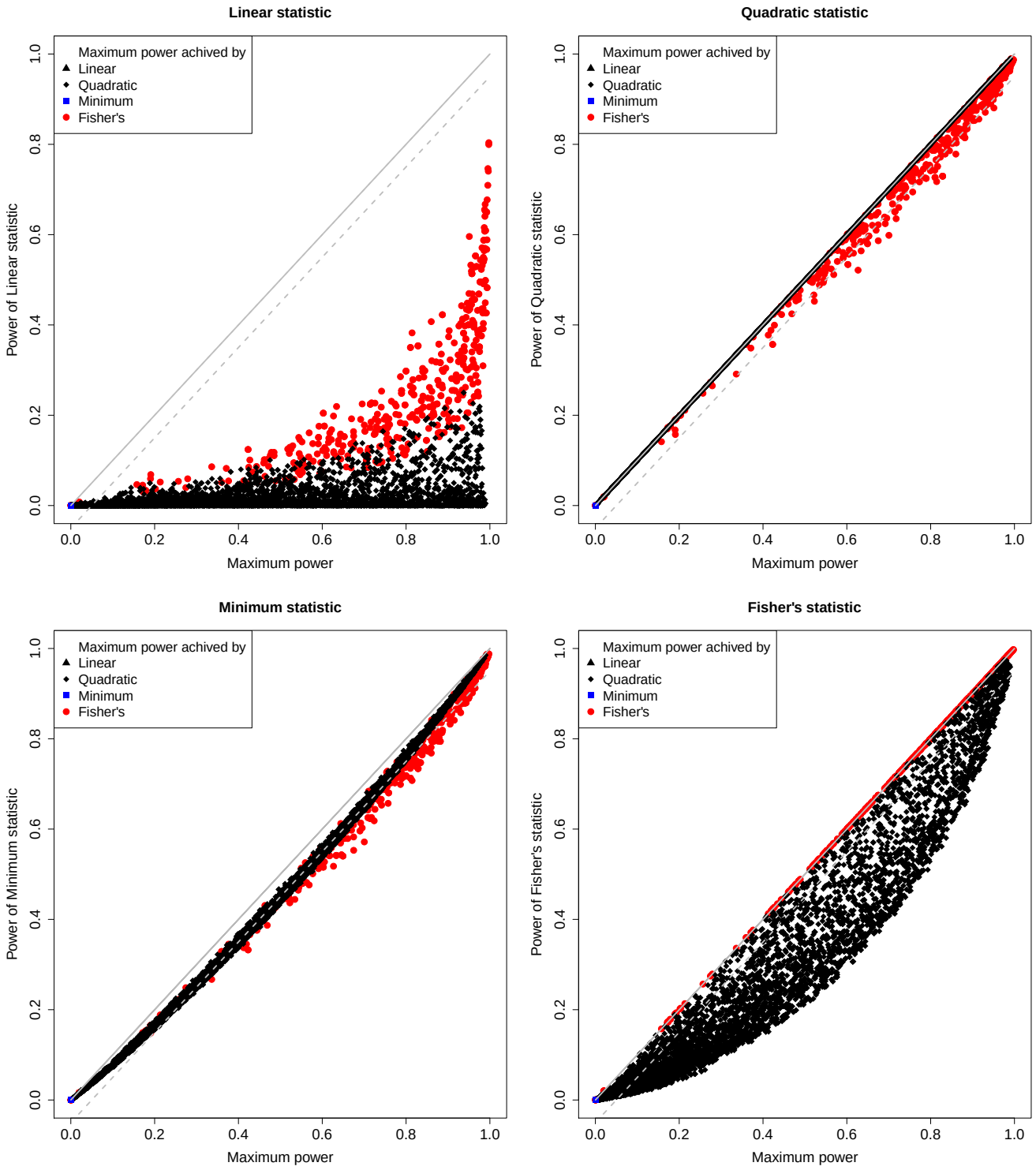

Figure 4: Empirical power of the four test statistics compared to the maximum power for 10,000 independently generated models for which the proportion of deleterious SNPs among the causal ones is generated from Unif (0.50, 0.75). All other parameters are described in Table 1 for studies of quantitative traits under the MAF-effect-dependent assumption, and they are the same as the ones used for Figure 1.

In summary, the proposed minimum-p and Fisher's hybrid test statistics provide much needed robustness in terms of power for association tests of rare variants, by combining information from the complementary linear and quadratic test statistics. Statistical significance of the hybrid statistics can be obtained efficiently using the same permutation-based method often required for the existing linear and quadratic statistics, without the need for additional permutations. The minimum-p statistic is attractive if one believes that causal rare variants are equally likely to be deleterious and protective. However, for the plausible scenario when the majority of the causal variants have the same direction of effect (either deleterious or protective), Fisher's test consistently outperforms methods that use the minimum-p principle [e.g., the simple minimum-p test considered here and SKAT-O Lee et al., 2012a], and it often provides considerably better power than the individual linear and quadratic tests. The general concept of using Fisher's method to combine information from two or more existing but complementary methods applied to the same data, beyond the traditional setting of meta-analysis of multiple data resources, can be readily extended and is useful in many other scientific studies.

ACKNOWLEDGEMENTS

The authors would like to thank the Genetic Analysis Workshop 17 (GAW17) committee and the 1000 Genomes Project for providing the GAW17 application data, and Dr. Andrew Paterson for insightful discussions. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC; 250053-2008) and the Canadian Institutes of Health Research (CIHR; MOP 84287) grants to L.S., NSERC to J.F.L., the Ontario Graduate Scholarship (OGS) and the CIHR Strategic Training for Advanced Genetic Epidemiology (STAGE) fellowship to A.D., University of Toronto. The authors have no conflict of interest to declare.

APPENDIX

An efficient permutation-based method that provides empirical P-values, simultaneously, for the linear, quadratic, minimum-p, and Fisher's tests

Here, we describe an efficient permutation-based method that provides empirical P-values, p_L, p_Q, p_M, and p_F, simultaneously, for tests based on W_L , W_Q , W_M, and W_F, respectively. This approach is novel, but similar in spirit to methods used for nonparametric estimation of copula functions [Genest and Rivest, 1993].

For a given dataset, let $Y = (Y_1, \ldots, Y_n)'$ be the (binary or quantitative) phenotype values for n subjects, $X_j = (X_{1j}, \ldots, X_{nj})'$, $j = 1, \ldots, J$ be the corresponding genotype values for a group of J SNPs under study. Let $W_{L,obs}$ be the observed linear statistic calculated, e.g., as in Equation (7), and $W_{Q,obs}$ be the observed quadratic statistic calculated, e.g., as in Equation (8). Due to sparsity or non-normality, asymptotic approximations often do not provide satisfactory P-values in practical settings, so resampling-based methods are recommended by many authors [e.g., Basu and Pan, 2011; Lin and Tang, 2011; Neale et al., 2011].

To preserve the possible dependence present in the observed genotypes between SNPs, a permuted dataset under the null of no association is obtained by permuting the phenotype. Let Y^k, $k = 1, \ldots, K$ be the K independently permuted phenotype vectors and W_L^k and W_Q^k be the corresponding linear and quadratic statistics for the kth permuted dataset. "Honest" P-values for the linear and quadratic tests

using the observed data are obtained, respectively, as

$$\begin{aligned} p_L &= \sum_k I(W^2_{L,k} \geq W^2_{L,obs})/K \ , \\ p_Q &= \sum_k I(W^2_{Q,k} \geq W^2_{Q,obs})/K \ , \end{aligned}$$

where $I(\cdot)$ indicates if the statistic from the kth permutated sample is greater than or equal to the observed statistic. The observed test statistics of W_M and W_F are, respectively,

$$\begin{aligned} W_{M,obs} &= \min(p_L, p_Q) \ , \\ W_{F,obs} &= -2\log(p_L) - 2\log(p_Q) \ . \end{aligned}$$

To empirically assess the statistical significance of $W_{M,obs}$ and $W_{F,obs}$ without additional permutations, let

$$p^k_L = Rank(|W^k_L|)/K, \quad k = 1, \ldots, K \ ,$$

be the empirical P-value of the linear test using the kth permuted sample, where $Rank(|W^k_L|)$ is the rank of $|W^k_L|$ among all K linear statistics calculated based on the K permuted samples. (Other choices are possible, e.g., $p^k_L = (Rank(|W^k_L|) - 0.5)/K$ but results are not practically different; P-values for linear statistics are two-sided to allow for either positive or negative association statistic under the alternative hypothesis.) Similarly, we calculate

$$p^k_Q = Rank(W^k_Q)/K \ , \quad k = 1, \ldots, K \ .$$

The W_M and W_F statistics using the K permuted datasets are, respectively,

$$\begin{aligned} W^k_M &= \min(p^k_L, p^k_Q) \ , \quad k = 1, \ldots, K \ , \\ W^k_F &= -2\log(p^k_L) - 2\log(p^k_Q) \ , \quad k = 1, \ldots, K \ . \end{aligned}$$

Finally, "honest" P-values of the minimum-p and Fisher tests for the observed data are obtained, respectively, as

$$\begin{aligned} p_M &= \sum_k I(W^k_M \leq W_{M,obs})/K \ , \\ p_F &= \sum_k I(W^k_F \geq W_{F,obs})/K \ . \end{aligned}$$

The size of K depends on the particular application. For the GAW17 data, $K = 10^4$ because P-values were large and power were assessed at $\alpha = 0.05$. For the simulation studies, $K = 10^6$ because power was assessed at α as low as 10^{-4}. For a more stringent type 1 error control (e.g., $10^{-6} = 0.05/50,000$ genes or bins/groups of rare variants) suitable for whole-genome analysis of rare variants, computational burden can be an issue, common to all genome-wide analyses that require permutations to assess statistical significance. For extremely sparse data, permutation-based methods are also known to be conservative. For example, in the extreme case of a case-control study of one single SNP, if there was only one copy of of the rare allele present in the sample, there would be only two distinct test statistics among all possible permutated datasets, resulting in permutation-based P-values being 0, 0.5, or 1. Randomized P-values are often recommended to circumvent the problem, but additional research is needed.

REFERENCES

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

Almasy L, Dyer T, Peralta J, Kent J, Charlesworth J, Curran J, Blangero J. 2011. Genetic analysis workshop 17 mini-exome simulation. *BMC Proceedings* 5:S2.

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11:773–185.

Basu S, Pan W. 2011. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 35:606–619.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 11:415–425.

Derkach A, Lawless JF, Sun L. 2012. Assessment of pooled association tests for rare genetic variants within a unified framework. arXiv:1205.4079 [stat.ME]; submitted for publication.

Genest C, Rivest LP. 1993. Statistical inference procedures for bivariate Archimedean Copulas. *J Am Stat Assoc* 88:1034–1043.

Goeman JJ, Van De Geer SA, Van Houwelingen HC. 2006. Testing against a high dimensional alternative. *J R Stat Soc: B (Statistical Methodology)* 68:477–493.

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

Hoffmann TJ, Marini NJ, Witte JS. 2010. Comprehensive approach to analyzing rare genetic variants. *PLoS ONE* 5:e13584.

Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB. 2012. The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. *PLoS Genet* 8:e1002496.

Lee S, Lin X, Wu MC. 2012a. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*.

Lee S, Miropolsky L, Wu M. 2012b. SKAT: SNP-set (Sequence) Kernel Association Test. R package version 0.76.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83:311–321.

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

Loughin TM. 2004. A systematic comparison of methods for combining p-values from independent tests. *Comput Stat Data Analysis* 47:467–485.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* 5: e1000384.

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mut Res* 615:28–56.

Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34:188–193.

Neale BM, Rivas MA, Voight BF, Altshuler D., Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet* 7:e1001322.

Owen AB. 2009. Karl Pearsons meta-analysis revisited. *Ann Statistics* 37:3867–3892.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei L-J, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86:832–838.

Stouffer S. 1949. The American Soldier: Adjustment during Army life. The American Soldier. Princeton, NJ: Princeton University Press.

Tintle N, Aschard H, Hu I, Nock N, Wang H, Pugh E. 2011. Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 Genomes Project exon sequencing data in unrelated individuals: summary results from Group 7 at Genetic Analysis Workshop 17. *Genet Epidemiol* 35:S56–S60.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the Sequence Kernel Association Test. *Am J Hum Genet* 89:82–93.

Yi N, Zhi D. 2011. Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol* 35:57–69.