Statistical Science
2014, Vol. 29, No. 2, 302–321
DOI: 10.1214/13-STS456

Pooled Association Tests for Rare Genetic Variants: A Review and Some New Results

Andriy Derkach, Jerry F. Lawless and Lei Sun

Abstract. In the search for genetic factors that are associated with complex heritable human traits, considerable attention is now being focused on rare variants that individually have small effects. In response, numerous recent papers have proposed testing strategies to assess association between a group of rare variants and a trait, with competing claims about the performance of various tests. The power of a given test in fact depends on the nature of any association and on the rareness of the variants in question. We review such tests within a general framework that covers a wide range of genetic models and types of data. We study the performance of specific tests through exact or asymptotic power formulas and through novel simulation studies of over 10,000 different models. The tests considered are also applied to real sequence data from the 1000 Genomes project and provided by the GAW17. We recommend a testing strategy, but our results show that power to detect association in plausible genetic scenarios is low for studies of medium size unless a high proportion of the chosen variants are causal. Consequently, considerable attention must be given to relevant biological information that can guide the selection of variants for testing.

Key words and phrases: Linear statistics, quadratic statistics, score tests, weighting, power, next generation sequencing, complex traits.

1. INTRODUCTION

Genome-wide association studies (GWAS) have identified numerous genetic variants (single nucleotide polymorphisms, or SNPs) that are associated with complex human traits [e.g., Manolio, Brooks and Collins (2008), Hindorff et al. (2009)]. However, because of their limited sample sizes, such studies are effective only at identifying common variants, that is, for which the minor allele frequency (MAF) is not too small (e.g., MAF $\geq$5% for sample size $\sim$2000). In addition, variants that have been identified through GWAS explain only small fractions of the estimated trait heritabilities. There is now much interest in understanding the role of rare variants (as represented by SNPs with small MAFs), but because they are rare it is difficult to detect associations with specific traits [e.g., Bansal et al. (2010); Asimit and Zeggini (2010)]. Next generation sequencing (NGS) can produce detailed information on rare variants but studies involving large numbers of individuals are not yet practical due to cost, heterogeneity and other concerns. Attention has consequently focused on methods that combine information across multiple rare SNPs in a genomic region (see Section 6 for discussion on the practical choice of a genomic region and SNPs within the region for analysis and its impact on the statistical inference). This area is the focus of our article. Our purpose is to review methods of testing for association between rare

Andriy Derkach is Graduate Student, Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, Ontario, Canada 1M5S 3G3. Jerry F. Lawless is Distinguished Professor Emeritus, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 and Professor, Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario, Canada M5T 3M7. Lei Sun is Associate Professor, Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, Ontario, Canada 1M5S 3G3 and Associate Professor, Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario, Canada M5T 3M7 (e-mail: sun@utstat.toronto.edu).

variants and a trait, unify the different methods, and give some new results.

To motivate our discussion, we refer to data from the Genetic Analysis Workshop 17 (GAW 17) [Almasy et al. (2011), 1000 Genomes Project Consortium (2010)]. These data include real sequence data (SNP genotypes) obtained from the 1000 Genomes Project, and simulated phenotype data (trait values) simulated by the GAW 17 committee. We focus here on a single quantitative trait, Q2. The values of Q2 and other traits were simulated for each person using normal linear regression models that included the SNP effects and, in some cases, additional covariates. Details concerning the simulation of trait values are given by Almasy et al. (2011). For Q2 the regression model involved effects for 72 SNPs within 13 genes, with MAFs ranging from 0.07% to 17.07%. Our objective is to look for evidence of associations between rare variants and Q2.

Papers that propose pooled association testing strategies for rare variants include Morgenthaler and Thilly (2007), Li and Leal (2008), Madsen and Browning (2009), Bansal et al. (2010), Han and Pan (2010), Hoffmann, Marini and Witte (2010), Morris and Zeggini (2010), Price et al. (2010), Yi and Zhi (2011), Neale et al. (2011), Wu et al. (2011), Sul, Buhm and Eleazar (2011) and Lee, Wu and Lin (2012). This previous work has provided many tests but insight into settings when a method will perform well, indifferently or poorly is still limited. Recently, Basu and Pan (2011) and Ladouceur et al. (2012) conducted extensive empirical evaluation (simulation) studies and reached a similar conclusion 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, directions of the associations (deleterious, protective or both), and the relationship between variant frequencies and genetic effects [Ladouceur et al. (2012)]. However, the joint effects of these factors have not been quantified analytically. Moreover, the test procedures assume that SNPs have been placed in groups, with pooling and testing carried out for SNPs within a given group. There are various ways SNPs might be grouped and this will affect the three factors mentioned. Ways of grouping SNPs are currently being studied in connection with the recent Genetic Analysis Workshop 18 (GAW 18) and elsewhere.

In this paper we consider tests for genotype–phenotype association within a unified framework. Most existing test statistics are either linear statistics that are powerful against specific association alternatives [e.g., Morgenthaler and Thilly (2007), Li and Leal (2008), Morris and Zeggini (2010), Madsen and Browning (2009) and Price et al. (2010)] or quadratic statistics that have reasonable power across a wide range of alternatives [e.g., Neale et al. (2011), Wu et al. (2011), Lee, Wu and Lin (2012)]. We study both classes of statistics theoretically and empirically and provide several new insights. In particular, we examine the (asymptotic or exact) powers of various tests as a function of the three factors above. We deal with both categorical and quantitative traits, and allow trait-dependent selection of individuals in a study as well as nonindependent SNPs. We conduct novel simulation studies that complement other recent empirical investigations and shed new light on methods' comparison. We also discuss so-called optimality of tests and indicate what this means in practical settings.

A feature of many of the linear statistics and of the quadratic statistics of Wu et al. (2011) and Lee, Wu and Lin (2012) is the use of weights associated with individual SNPs, because of the suggestion that rarer variants tend to have larger genetic effects. We demonstrate that even if this assumption is true, using weights inversely proportional to MAFs can in some cases have an adverse effect. We also show that for linear statistics, methods of weight selection based on estimated effects [e.g., Han and Pan (2010), Yi and Zhi (2011), Hoffmann, Marini and Witte (2010), Lin and Tang (2011)] are similar to using quadratic statistics.

A referee has stressed the importance of several caveats concerning the type of data considered in the paper, and hence the "success" of testing procedures such as discussed here. First, errors in sequencing data commonly occur. Methods for addressing this have not yet been well studied in the present context, and we assume that genotypes are as given. Methods used in other contexts [Daye, Li and Wei (2012), Skotte, Korneliussen and Albrechtsen (2012)] are typically based on estimated sequencing error probabilities, but we note that their accuracy is not well established in specific settings. A second caveat is that the identification of rare variants is difficult because of their low frequency, and because sequencing errors can substantially affect the estimation of small MAFs. They can also lead to a SNP that is actually monomorphic being identified as a rare polymorphic SNP in some instances. Finally, the nature and level of heritability explained by rare variants is at this point speculative and it is unclear whether major successes will occur from the approaches considered here. We take pains in the

paper to consider a broad range of genetic models but we cannot of course answer questions about the scientific fundamentals.

The remainder of the paper is organized as follows. Section 2 introduces the framework for testing the association between a group of rare variants and a general trait, reviews tests that have been proposed along with analytical results relating the power of linear and quadratic statistics to the various factors, and considers adjustment for covariates. Section 3 presents theoretical power calculations for normally distributed traits that clarify when various methods will do well and the effects of using weights. Section 4 gives numerical results based on large-scale simulation studies of over 10,000 different models for both quantitative and binary traits. Section 5 examines the GAW17 quantitative trait Q2 and sequence data from the 1000 Genomes Project. Section 6 concludes with some recommendations for pooled testing. Online supplementary materials [Derkach, Lawless and Sun (2013a)] include details specific about test statistics and additional tables and figures for the power comparison studies.

2. SCORE TESTS FOR ASSOCIATION

2.1 No Covariate Adjustment

We assume that a group of J SNPs and a trait Y are under consideration. The objective is to test whether there is association between Y and one or more of the SNPs. For a set of n unrelated individuals, let Y_i be the measured trait value for individual i and $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. Let X_{ij} denote the SNP genotype for individual i, $i = 1, \ldots, n$ and $j = 1, \ldots, J$; for simplicity we assume that X_{ij} denotes whether the rare allele is present ($X_{ij} = 1$) or absent ($X_{ij} = 0$) and let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})'$. It is straightforward to consider the case where X_{ij} is the number of copies (0, 1 or 2) of the rare allele for SNP j, but there will be no or very few individuals with two rare alleles in a study of current typical size. We assume for now that there is no adjustment for covariates, since many papers address only this case. However, covariate adjustment is often important and we consider it in Section 2.4.

Our interoest is in testing the null hypothesis

$$H_0 : \mathbf{Y} \text{ and } \mathbf{X} \text{ are independent.} \tag{2.1}$$

Most proposed methods for testing H_0 are based on statistics that are (weighted) linear or quadratic combinations of statistics S_j which measure association between Y and SNP j, $j = 1, \ldots, J$. Without loss of generality, we assume that S_j is such that under the null $E[S_j] = 0$ and $\mathrm{Var}(S_j) = \sigma_{0j}^2$, and under alternatives $E[S_j] = \mu_j$ and $\mathrm{Var}(S_j) = \sigma_j^2$. To facilitate further discussion, we assume that Y is defined so that a SNP with $\mu_j > 0$ is termed deleterious, with $\mu_j < 0$ is protective, and with $\mu_j = 0$ is neutral; both deleterious and protective SNPs are causal variants. Let $\mathbf{S} = (S_1, \ldots, S_J)'$ and $E[\mathbf{S}] = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)'$, and assume for simplicity that the hypothesis of no association (2.1) is equivalent to the null hypothesis

$$H_0 : \boldsymbol{\mu} = \mathbf{0}. \tag{2.2}$$

There are various options for S_j, but the approaches referred to in Section 1 can almost all be expressed in terms of statistics of the form

$$S_j = \sum_{i=1}^{n} (Y_i - \overline{Y}) X_{ij}, \quad j = 1, \ldots, J, \tag{2.3}$$

where $\overline{Y} = \sum_{i=1}^{n} Y_i / n$ [e.g., see Lin and Tang (2011); Basu and Pan (2011)]. The S_j arise as score statistics in regression models for the two important cases where Y_i is normally distributed and binary, respectively. They also arise from Poisson models for counts and for other models in the linear exponential family [e.g., Lee, Wu and Lin (2012)]. For completeness, we outline this for the binary case in the supplementary materials [Derkach, Lawless and Sun (2013a)]. Other statistics, for example, Wald or likelihood ratio statistics, could be used (see Section 2.4), but score statistics are almost universally used in this area, and we focus on them. We note that the score statistics have the advantage of requiring only estimates obtained under the null hypothesis. In some contexts it is also useful to replace $Y_i - \overline{Y}$ in (2.3) with some other function α_i of either Y_i or its rank, with $\sum_{i=1}^{n} \alpha_i = 0$. It should be noted that genotypes X_{ij}, $j = 1, \ldots, J$, are not assumed to be mutually independent in the subsequent development.

Many authors have considered linear test statistics for H_0 (2.2) of the form

$$W_L = \sum_{j=1}^{J} w_j S_j = \mathbf{w}'\mathbf{S}, \tag{2.4}$$

where the weights w_js are specified nonnegative values and $\mathbf{w} = (w_1, \ldots, w_J)'$. Basu and Pan (2011) provided a review, and we note two important cases: Morgenthaler and Thilly (2007) considered the "cohort allelic sums test" (CAST) where each $w_j = 1$, and Madsen and Browning (2009) based w_j on the (estimated) MAF, with larger weights for SNPs with smaller MAF. The rationale for the latter weights is

that causative SNPs would be subject to "purifying selection" and so be rarer in the population than neutral SNPs, but evidence for this so far seems slight. We also note that because the MAFs have to be estimated, sequencing errors as discussed in Section 1 can have an effect; we assume (idealistically) that such errors have not occurred. Price et al. (2010) also considered "threshold" versions in which $w_j > 0$ only if the estimated MAF is below a specified threshold (e.g., 1% or 5%). Such linear composite statistics can have good power against association alternatives where $\mu_j \geq 0$, with $\mu_j > 0$ for some subset of $\{j = 1, \ldots, J\}$. However, their power may be poor for alternatives where both positive and negative values of μ_j are possible, and when only a small proportion of the J SNPs are causal and have $\mu_j > 0$ [Neale et al. (2011), Basu and Pan (2011)]. The effects of association direction on different statistics are studied in Sections 3 and 4.

Many authors have also considered quadratic statistics,

$$W_Q = \mathbf{S}' A \mathbf{S}, \tag{2.5}$$

where A is a positive definite (or semi-definite) symmetric matrix. One common choice is $A = \Sigma_0^{-1}$, where Σ_0 is a known or estimated covariance matrix for $\mathbf{S}$ under H_0; this gives a Hotelling statistic,

$$W_H = \mathbf{S}' \Sigma_0^{-1} \mathbf{S}. \tag{2.6}$$

Other quadratic statistics include the "SSU" statistic of Pan (2009) and the "C-alpha" statistic of Neale et al. (2011) which are based on $A = I$, the $J \times J$ identity matrix; the "SKAT" statistic of Wu et al. (2011) uses $A = \text{diag}\{a_1, \ldots, a_J\}$, where the a_js are weights that depend on the MAFs via a *Beta* function. The linear statistic W_L in (2.4) can also be expressed in quadratic form, since W_L^2 is equivalent to (2.5) with $A = \mathbf{w}\mathbf{w}'$. However, note that A is no longer positive definite in this case. Quadratic statistics arise naturally from regression models relating Y and X_j as we discus below. Finally, we remark that recent work has considered combining evidence from linear and quadratic statistics [e.g., Lee, Wu and Lin (2012) and Derkach, Lawless and Sun (2013b)]. We discuss this in Section 6, but focus on individual linear and quadratic statistics here (Table 1).

2.2 Distributions of Linear and Quadratic Statistics Under Normality

It is instructive to consider the case where $\mathbf{S}$ is normally distributed. For both binary and quantitative traits, the vectors $\mathbf{S}$ are all at least asymptotically normal, and analytical derivations of power and discussions of optimality rely on this assumption [e.g., Lin and Tang (2011); Lee, Wu and Lin (2012)]. The

TABLE 1

Summary of different association tests for analyzing rare variants. This is not an exhaustive list of all existing tests (see Sections 2 and 6 for additional examples). Tests derived from random effect models and adaptive linear models are operationally similar to quadratic tests (see Section 2.3 for discussion). Details of the notation: see Section 2.1. Briefly, $\mathbf{S} = (S_1, \ldots, S_J)'$ is a vector of test statistics for a group of J rare variants, $\mathbf{w} = (w_1, \ldots, w_J)'$ is a vector of weights, A is a positive definite (or semi-definite) symmetric matrix, Σ_0 is a known or estimated covariance matrix for $\mathbf{S}$, p_j is the minor allele frequency (MAF) of SNP j, $f(p_j) = 1/\sqrt{p_j(1-p_j)}$ in Weighted-sum of Madsen and Browning (2009), $f(p_j)$ depends on the MAF via a Beta distribution in SKAT of Wu et al. (2011), and p_L and p_Q are the p-values from chosen Linear and Quadratic tests

Class of tests		
Linear	**Quadratic**	**Combined/Hybrid**
$W_L = \mathbf{w}'\mathbf{S}$	$W_Q = \mathbf{S}'A\mathbf{S}$	$H(W_L, W_Q)$
Example of specific tests		
$\mathbf{w} = \mathbf{1}$ (CAST, W_{L1})	$A = I$ (SSU and C-alpha, W_C)	$\max_w\{W_L\}$ (EREC)
Morgenthaler and Thilly (2007)	Pan (2009), Neale et al. (2011)	Lin and Tang (2011)
$w_j = f(p_j)$ (Weighted-sum, W_{Lp})	$A = \text{diag}\{a_j\}$, $a_j = f(p_j)$ (SKAT)	$\max_{\rho\in[0,1]}(\rho W_L + (1-\rho)W_Q)$ (SKAT-O)
Madsen and Browning (2009)	Wu et al. (2011)	Lee, Wu and Lin (2012)
$w_j = 0$ if $p_j >$ threshold (Threshold)	$A = \Sigma_0^{-1}$ (Hotelling, W_H)	$-2\log(p_L) - 2\log(p_Q)$ (Fisher's method), $\min(p_L, p_Q)$ (minimum-p)
Price et al. (2010)	Basu and Pan (2011)	Derkach, Lawless and Sun (2013b)

case where $\mathbf{S}$ is normal in finite samples also is well known in connection with tests for a multivariate normal mean $\boldsymbol{\mu}$; see, for example, Mardia, Kent and Bibby (1979), Chapter 5.

Suppose that under H_1 for which $\boldsymbol{\mu} \neq \mathbf{0}$ the distribution of $\mathbf{S}$ is (exactly or asymptotically) multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix Σ, $\mathbf{S} \sim N(\boldsymbol{\mu}, \Sigma)$. For simplicity we assume that Σ is known; this is allowable for asymptotic results which we focus on here. In finite samples where Y given X is normal, the effect of estimating Σ is to replace normal and chi-square distributions below with t and F distributions, respectively. With J fixed and n going to infinity, these converge to the normal and chi-square distributions we consider.

Let $\lambda_1, \ldots, \lambda_J$ be the eigenvalues of $\Sigma^{1/2} A \Sigma^{1/2}$ and P be the $J \times J$ orthogonal matrix whose columns are the corresponding eigenvectors. Then the following distributional results hold [e.g., Rao (1973), Section 3b.4]:

(i) W_Q is distributed as a linear combination of independent noncentral χ^2_1 random variables,

$$W_Q \sim \sum_{j=1}^{J} \lambda_j \chi^2_{1,\mathrm{nc}_j}, \tag{2.7}$$

where $\chi^2_{k,r}$ denotes a noncentral χ^2 random variable with k degrees of freedom and noncentrality parameter r, and $\mathrm{nc}_j = (\{P' \Sigma^{-1/2} \boldsymbol{\mu}\}_j)^2$.

(ii) If $A = \Sigma^{-1}$, then $W_Q \sim \chi^2_{J,\mathrm{nc}}$ with $\mathrm{nc} = \boldsymbol{\mu}' \Sigma^{-1} \boldsymbol{\mu}$. If $\Sigma = \Sigma_0$, then W_Q is the Hotelling statistic (2.6).

(iii) $Z_L^2 = W_L^2/(\mathbf{w}' \Sigma \mathbf{w}) = (\mathbf{w}'\mathbf{S})^2/(\mathbf{w}' \Sigma \mathbf{w}) \sim \chi^2_{1,\mathrm{nc}}$ with $\mathrm{nc} = (\mathbf{w}'\boldsymbol{\mu})^2/(\mathbf{w}' \Sigma \mathbf{w})$ when $\Sigma = \Sigma_0$. When this is not true, then the distribution of Z^2 is a multiple of the noncentral χ^2_1 random variable.

(iv) Under the null hypothesis H_0: $\boldsymbol{\mu} = \mathbf{0}$, $W_L^2/(\mathbf{w}' \Sigma_0 \mathbf{w})$ is a χ^2_1 random variable; W_Q is a linear combination of independent χ^2_1 random variables with each $\mathrm{nc}_j = 0$ in (2.7).

It should be noted that no adjustment is needed to reflect the fact that $\mathbf{w}$ may involve estimated MAFs. This is because the distributional results are based on the sampling distribution of Y given X_{ij}, where estimates of MAFs are functions of X alone and so are treated as fixed in this section. We return to this point in Section 4.1, and we also note in Section 4.2 that complications arise when retrospective (case–control) studies are used with binary responses. These results allow the power against a simple alternative hypothesis H_1 with a specified $\boldsymbol{\mu} \neq \mathbf{0}$ to be calculated for any linear test statistic (2.4) or quadratic test statistic (2.5). Critical values for a test of H_0: $\boldsymbol{\mu} = \mathbf{0}$ are obtained according to (iv). Software exists for the computation of probabilities associated with linear combinations of central or noncentral χ^2_1 random variables, for example, the *CompQuadForm* package in R [Duchesne and Lafaye de Micheaux (2010)]. In particular, we note that:

(a) For a size α test using the linear statistic W_L in (2.4) or, equivalently, Z_L^2 in (iii) above, the α critical value is $\chi^2_1(1-\alpha)$, the $1-\alpha$ quantile for the χ^2_1 distribution. (The test is two-sided to allow for either positive or negative W_L under H_1.) The power against H_1 when $\Sigma = \Sigma_0$ is

$$P\big(\chi^2_{1,\mathrm{nc}_L} > \chi^2_1(1-\alpha)\big) \quad \text{where } \mathrm{nc}_L = (\mathbf{w}'\boldsymbol{\mu})^2/(\mathbf{w}' \Sigma \mathbf{w}). \tag{2.8}$$

(b) For a size α test using the Hotelling statistic W_H in (2.6), the α critical value is $\chi^2_J(1-\alpha)$. The power against H_1 in the case where $\Sigma = \Sigma_0$ is

$$P\big(\chi^2_{J,\mathrm{nc}_H} > \chi^2_J(1-\alpha)\big) \quad \text{where } \mathrm{nc}_H = \boldsymbol{\mu}' \Sigma^{-1} \boldsymbol{\mu}. \tag{2.9}$$

The specific power of both statistics depends on $\boldsymbol{\mu}$ and on the distribution of $\mathbf{S}$ under H_1, however, some general features can be seen. For simplicity, suppose $\Sigma = \Sigma_0$ and that Σ is diagonal (SNPs are independent). The quadratic statistic W_H (2.6) is a reasonable choice when both deleterious ($\mu_j > 0$) and protective ($\mu_j < 0$) SNPs are plausible, because nc_H is a function of the μ_j^2. The statistic W_H can be decomposed as $W_H = Z_L^2 + R$, where Z_L and R are independent under H_1, and $R \sim \chi^2_{J-1,\mathrm{nc}_R}$ with $\mathrm{nc}_R = \mathrm{nc}_H - \mathrm{nc}_L = \boldsymbol{\mu}' \Sigma^{-1} \boldsymbol{\mu} - (\mathbf{w}'\boldsymbol{\mu})^2/(\mathbf{w}' \Sigma \mathbf{w})$. The linear statistic W_L is optimal when $\mathrm{nc}_R = 0$, but the advantage of W_L over the quadratic statistic W_H disappears as nc_R increases. We will discuss this in Sections 3 and 4.

2.3 Additional Considerations: Optimality, Random Effect Models, Adaptive Linear Models, *p*-Values and Permutation Distribution

A number of authors [e.g., Lee, Wu and Lin (2012), Neale et al. (2011), Lin and Tang (2011)] have claimed to obtain "optimal" tests. This is theoretically possible if we specify a suitable family of test statistics, but for this to be of practical use we must have strong prior knowledge about the alternative hypothesis. For example, among the class of linear statistics (2.4), maximal

power is obtained when $\mathbf{w} = \Sigma^{-1}\boldsymbol{\mu}$. When the S_js are independent so that $\Sigma = \operatorname{diag}\{\sigma_1^2, \ldots, \sigma_J^2\}$, this gives $w_j = \mu_j/\sigma_j^2$. This linear statistic is (asymptotically) optimal among all tests of fixed size based on $\mathbf{S}$, assuming $\boldsymbol{\mu}$ is known. Quadratic statistics (2.5) for which A has rank 2 or more can never be optimal against a specific alternative $(\boldsymbol{\mu}, \Sigma)$. However, quadratic tests can maintain reasonable power over wide ranges of alternatives, whereas a linear statistic's power can be poor except near a specific alternative. Goeman, van de Geer and van Houwelingen (2006) and other authors have discussed optimality of score statistics coming from random effects models, but these results are also based on averaging over a family of alternatives, which may or may not be plausible in a given setting. For example, quadratic statistics (2.5) can be obtained from random effect regression models in which Y is related to $\mathbf{X}$ through a linear function $\boldsymbol{\beta}'\mathbf{X}$ and the $J \times 1$ regression coefficient $\boldsymbol{\beta}$ is a random vector with mean $\mathbf{0}$ and covariance matrix τA. The hypothesis $\tau = 0$ then corresponds to H_0 in (2.1) and a score statistic for testing it is [Goeman, van de Geer and van Houwelingen (2006), Basu and Pan (2011)]

$$W_Q' = \tfrac{1}{2}\mathbf{S}'A\mathbf{S} - \tfrac{1}{2}\operatorname{trace}(A\Sigma_0). \tag{2.10}$$

Using W_Q' is equivalent to using W_Q in (2.5) when Σ_0 is known. The first term in (2.10) also arises from other score tests in generalized linear models [Lee, Wu and Lin (2012)]. In general, Σ_0 (and A) involve estimates and asymptotic distributions for W_Q are used to get p-values. The asymptotic distributions are typically of the form (2.7), but with the λ_j involving estimates. We comment further on the calculation of p-values at the end of this section.

Some authors [e.g., Han and Pan (2010), Hoffmann, Marini and Witte (2010), Lin and Tang (2011)] have proposed two-stage or other adaptive approaches in which the weighting vector $\mathbf{w}$ for W_L in (2.4) is chosen after preliminary examination of the direction of S_j or an estimate of its effect based on the observed data, in a hope of choosing an "optimal" weight. However, such an approach cannot on its own (i.e., without the use of additional information from other sources) improve globally the linear statistics. In fact, if we choose the $\mathbf{w}$ that maximizes the standardized linear test statistic (2.4), then we end up with the quadratic statistic (2.6). In particular [e.g., Mardia, Kent and Bibby (1979), page 127, or Li and Lagakos (2006), Section 3],

$$\sup_{\mathbf{w}}\left\{\frac{W_L^2}{\operatorname{Var}(W_L)}\right\} = \sup_{\mathbf{w}}\left\{\frac{(\mathbf{w}'\mathbf{S})^2}{\mathbf{w}'\Sigma\mathbf{w}}\right\} = \mathbf{S}'\Sigma^{-1}\mathbf{S} = W_H,$$

where the maximizing vector is $\mathbf{w} = \Sigma^{-1}\mathbf{S}$. This helps explain why Basu and Pan (2011) found that adaptive procedures did not perform as well as one might have hoped.

Lin and Tang (2011) have proposed a test statistic $T_{\max}$ based on the maximum of a specified set of K linear statistics, each with different weights, $T_k^2 = (\mathbf{w}_k'\mathbf{S})^2/(\mathbf{w}_k'\Sigma\mathbf{w}_k)$. We do not consider such statistics here, but it is clear that their performance will depend on the choice of "appropriate" weighting vectors $\mathbf{w}_k$. When there is little prior information and the $\mathbf{w}_k$s are selected to cover a wide range of alternatives, it seems likely that $\max(T_k^2)$ would be similar to W_H. A similar suggestion involving quadratic statistics is made by Lee, Wu and Lin (2012). In practice, there is often very limited prior information about the nature of $\boldsymbol{\mu}$, especially concerning which SNPs might be causal, so one cannot be confident that a linear test statistic will be effective, nor which quadratic statistics might be the best. Sections 3 and 4 investigate situations in which specific statistics will be more powerful.

To achieve reasonable power, sample sizes have to be rather large, as we discuss in Section 4. The calculation of p-values, critical values or power is often based on large sample approximations given by normal and chi-square distributions in Section 2.2. In general, this requires estimation of matrices Σ_0 and A (as do test statistics themselves) but with consistent estimators the limiting distributions provide adequate approximation for sufficiently large samples. In general, a consistent estimator of Σ_0 for $\mathbf{S}$ given by (2.3) is

$$\hat{\Sigma}_0 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} X_c' X_c, \tag{2.11}$$

where X_c' has (i, j) entry $X_{ij} - \overline{X}_j$ (where $\overline{X}_j = \sum_{i=1}^n X_{ij}/n$). However, because events with $X_{ij} = 1$ are rare, the distribution of $\mathbf{S}$ can be quite nonnormal even in rather large samples, and more accurate ways to calculate p-values and critical values are needed, especially for quadratic statistics. Some authors [e.g., Lee, Wu and Lin (2012)] have given skewness or kurtosis adjustments that seem to improve accuracy in certain settings. More generally, however, we can obtain p-values (and study power) by simulation. When there is no adjustment for covariates, the permutation distribution of $\mathbf{S} = (S_1, \ldots, S_J)'$ is typically used [e.g., Basu and Pan (2011)]; this is the distribution that arises from randomly permuting the Y_is and assigning them to the $\mathbf{X}_i$s. This also applies when Y is a discrete variable, when X_{ij}s are correlated within individuals (e.g., due to linkage disequilibrium, LD) and when sampling of

the individuals is Y-dependent. More generally, when there are covariates present, we may need to rely on bootstrap simulations. We comment on this in the following section.

2.4 Adjustment for Covariates

Lin and Tang (2011) and Wu et al. (2011) have stressed that adjustment for covariates and population stratification will be important in many contexts involving rare variants. In this case we use regression models; for illustration, we consider the case of a binary trait. Suppose that in addition to the genotype vector $\mathbf{X}_i$ there is a vector $\mathbf{v}_i$ of covariates that may be related to a binary trait Y_i. Then a logistic regression model

$$\begin{aligned}\Pr(Y_i = 1|\mathbf{X}_i, \mathbf{v}_i) \\ = \frac{\exp(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + \boldsymbol{\gamma}'\mathbf{v}_i)}{1+\exp(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + \boldsymbol{\gamma}'\mathbf{v}_i)} = \mu_i\end{aligned} \tag{2.12}$$

might be considered, and a test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$ can be carried out. For testing rare variants some authors have replaced the term $\boldsymbol{\beta}'\mathbf{X}_i$ in (2.12) with βr_i, where $r_i = \sum_{j=1}^{J} X_{ij}$ is the total number of rare variants per individual [e.g., Morris and Zeggini (2010); Yilmaz and Bull (2011)], but this corresponds to using a linear statistic in previous sections and can be ineffective. We consider the case where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ in order to examine settings for which causal SNPs may be either deleterious or beneficial. Consideration of the power of alternative tests in large samples parallels the discussion in Section 2.2, as follows.

Let $\hat{\boldsymbol{\beta}}$ be the estimator of $\boldsymbol{\beta}$ based on the model in question and assume that under $H_0 : \boldsymbol{\beta} = \mathbf{0}$, the asymptotic distribution of $\sqrt{n}\hat{\boldsymbol{\beta}}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix Σ. Following Li and Lagakos (2006), we consider a sequence of contiguous alternatives

$$H_1^{(n)} : \boldsymbol{\beta} = \mathbf{b}/\sqrt{n}, \tag{2.13}$$

where $\mathbf{b} = (b_1, \ldots, b_J)'$ is a specified vector. Under this sequence as $n \to \infty$ the distribution of $\sqrt{n}\hat{\boldsymbol{\beta}}$ approaches a multivariate normal distribution with mean $\mathbf{b}$ and covariance matrix Σ. Thus, asymptotic power for a test statistic can be computed in the same way as in Section 2.3. Li and Lagakos (2006) compare the quadratic Wald test statistic $W = \hat{\boldsymbol{\beta}}'\hat{\Sigma}^{-1}\hat{\boldsymbol{\beta}}$, where $\hat{\Sigma}$ is a consistent estimate of Σ under H_0, with linear statistics $Z = \mathbf{a}'\hat{\beta}$. These are analogous to (2.6) and (2.4), respectively. The likelihood score statistic for testing $\boldsymbol{\beta} = \mathbf{0}$ is an alternative to the Wald statistic; it is easily found as [e.g., Lin and Tang (2011)]

$$\mathbf{U} = \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)\mathbf{X}_i, \tag{2.14}$$

where $\hat{\mu}_i = e^{\hat{\beta}_0 + \hat{\boldsymbol{\gamma}}'\mathbf{v}_i}/(1 + e^{\hat{\beta}_0 + \hat{\boldsymbol{\gamma}}'\mathbf{v}_i})$ and $\hat{\beta}_0$, $\hat{\boldsymbol{\gamma}}$ are estimated from (2.12) when $\beta = \mathbf{0}$. It also follows from standard maximum likelihood large sample theory that the covariance matrix of $\mathbf{U}$ under H_0 is estimated consistently by

$$\begin{aligned}\hat{\Sigma}_U = \widehat{\mathrm{Var}}(\mathbf{U}) = \left(\sum_{i=1}^{n} \hat{\sigma}_i^2 \mathbf{X}_i\mathbf{X}_i'\right) \\ - \left(\sum_{i=1}^{n} \hat{\sigma}_i^2 \mathbf{X}_i\tilde{\mathbf{v}}_i'\right)\left(\sum_{i=1}^{n} \hat{\sigma}_i^2 \tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i'\right)^{-1} \\ \cdot \left(\sum_{i=1}^{n} \hat{\sigma}_i^2 \tilde{\mathbf{v}}_i\mathbf{X}_i'\right),\end{aligned} \tag{2.15}$$

where $\hat{\sigma}_i^2 = \hat{\mu}_i(1 - \hat{\mu}_i)$ and $\tilde{\mathbf{v}}_i = (1, \mathbf{v}_i')'$. These correspond to results given by Lin and Tang (2011), who consider linear statistics based on linear combinations of the elements $U_1, \ldots, U_J$ of $\mathbf{U}$. The statistic (2.14) and variance estimate (2.15) are given here for prospective sampling but can be shown to apply under case–control sampling. As in Sections 2.1–2.3, test statistics such as $W_H^* = \mathbf{U}'\hat{\Sigma}_U^{-1}\mathbf{U}$ and $W_L^* = (\mathbf{w}'\mathbf{U})/(\mathbf{w}'\hat{\Sigma}_U^{-1}\mathbf{w})$, which correspond to W_H and W_L in preceding sections, can be used. When there are no covariates $\mathbf{v}_i$, it is readily seen that (2.14) reduces to (2.3) and that (2.15) equals $(n-1)/n$ times (2.11). It should be noted that when covariates $\mathbf{v}_i$ are present, the normal approximations considered earlier apply, but the permutation distribution p-values do not unless the $\mathbf{X}_i$s are independent of the $\mathbf{v}_i$. Lin and Tang (2011) suggest a parametric bootstrap as an alternative, based on randomly generating response Y_is from the fitted null model based on $\hat{\beta}_0$, $\hat{\boldsymbol{\gamma}}$.

Normal linear regression models for quantitative variables Y also produce score statistics of the form (2.14) with $\hat{\mu}_i = \hat{\beta}_0 + \hat{\gamma}'\mathbf{v}_i$, as do certain other generalized linear models [Lee, Wu and Lin (2012)]. It should be mentioned that in the case of quantitative Y-dependent sampling and models with supplementary covariates $\mathbf{v}_i$ as in (2.12), adjustments to estimating functions [e.g., Huang and Lin (2007); Yilmaz and Bull (2011)] are needed; this is beyond our present scope, but we note that statistics like (2.14) arise once again [Barnett, Lee and Lin (2013)].

3. NORMALLY DISTRIBUTED TRAITS

3.1 Distributions of the Linear and Quadratic Statistics

To provide more insights on the effects of the choice of linear vs. quadratic statistics and the use of weights on power, it is helpful to consider genetic scenarios described by a normal linear model,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_1 X_{iJ} + e_i \quad \text{for } i = 1, \ldots, n, \tag{3.1}$$

with $e_i \sim N(0, \sigma^2)$ and the X_{ij}s mutually independent Bernoulli variables with $P(X_{ij} = 1) = p_j$, approximately twice the MAF of SNP j, $j = 1, \ldots, J$. The score statistic $\mathbf{S} = (S_1, \ldots, S_J)'$ with

$$S_j = \sum_{i=1}^{n} (Y_i - \overline{Y}) X_{ij} = \sum_{i=1}^{n} (X_{ij} - \overline{X}_j) Y_i \tag{3.2}$$

arises from maximum likelihood theory for testing $H_0 : \boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)' = \mathbf{0}$, as noted in Section 2.4. Normal models are widely used for quantitative traits such as blood pressure or lipid levels. Due to the normality of Y, the distribution of S_j given the genotypes is $S_j \sim N(m_j(1 - m_j/n)\beta_j, m_j(1 - m_j/n)\sigma^2)$, where $m_j = \sum_{i=1}^{n} X_{ij}$. For any given sample the m_j are treated as fixed values, and for simplicity we consider the case where m_j is equal to its expected value np_j so that

$$\mathbf{S} \sim N(\boldsymbol{\mu}, \Sigma), \tag{3.3}$$

where $\boldsymbol{\mu} = (np_1(1 - p_1)\beta_1, \ldots, np_J(1 - p_J)\beta_J)'$ and $\Sigma = \operatorname{diag}\{np_1(1-p_1)\sigma^2, \ldots, np_J(1-p_J)\sigma^2\}$. As earlier, we ignore the small effects due to the need to estimate σ^2 in large samples.

Here and in simulations below, we consider settings according to the variation of Y explained by the set of SNPs. Under model (3.1), the total phenotypic variation explained by the J SNPs is

$$\begin{aligned} EV &= \frac{\operatorname{Var}(E[Y|\mathbf{X}])}{\operatorname{Var}(Y)} = \frac{\sum_{j=1}^{J} p_j(1 - p_j)\beta_j^2}{\sum_{j=1}^{J} p_j(1 - p_j)\beta_j^2 + \sigma^2} \\ &\approx \sum_{j=1}^{J} p_j(1 - p_j)\beta_j^2/\sigma^2 \\ &= \sum_{j=1}^{J} EV_j, \end{aligned} \tag{3.4}$$

where $EV_j = p_j(1-p_j)\beta_j^2/\sigma^2$ is the "Explained Variation" by SNP j. The approximation assumes that the phenotypic variation explained by genetic factors is small, which is in agreement with current data. The distribution of $W_L = \mathbf{w}'\mathbf{S}$ is $N(n\sum_{j=1}^{J} w_j p_j(1 - p_j)\beta_j, n\sum_{j=1}^{J} w_j^2 p_j(1 - p_j)\sigma^2)$, and

$$W_L^2 \Big/ \left(\sum_{j=1}^{J} w_j^2 p_j(1 - p_j)\sigma^2 \right) \sim \chi^2_{1,\mathrm{nc}_L}, \tag{3.5}$$

where

$$\begin{aligned} \mathrm{nc}_L &= n\frac{(\sum_{j=1}^{J} w_j p_j(1 - p_j)\beta_j/\sigma)^2}{\sum_{j=1}^{J} w_j^2 p_j(1 - p_j)} \\ &= n\frac{(\sum_{j=1}^{J} w_j \operatorname{sign}(\beta_j)\sqrt{p_j(1 - p_j)}\sqrt{EV_j})^2}{\sum_{j=1}^{J} w_j^2 p_j(1 - p_j)}. \end{aligned} \tag{3.6}$$

Similarly, assuming $A = \operatorname{diag}\{a_1, \ldots, a_J\}$ where the a_js can also be interpreted as weights for quadratic statistics $W_Q = \mathbf{S}'A\mathbf{S}$, we have

$$W_Q \sim \sum_{j=1}^{J} \lambda_j \chi^2_{1,\mathrm{nc}_j}, \tag{3.7}$$

where

$$\begin{aligned} &\lambda_j = a_j np_j(1 - p_j)\sigma^2 \quad \text{and} \\ &\mathrm{nc}_j = np_j(1 - p_j)\beta_j^2/\sigma^2 = nEV_j. \end{aligned} \tag{3.8}$$

3.2 Effects of Weights and Genetic Factors on Power

We consider for discussion two linear statistics $W_L = \mathbf{w}'\mathbf{S}$: W_{L1} with $w_j \equiv 1$ [Morgenthaler and Thilly (2007)] and W_{Lp} with $w_j = 1/\sqrt{p_j(1 - p_j)}$ [Madsen and Browning (2009)]. We also consider two quadratic statistics $W_Q = \mathbf{S}'A\mathbf{S}$: W_C with $A = I$ ($a_j \equiv 1$) (C-alpha) and the Hotelling W_H with $A = \Sigma^{-1}$ ($a_j = 1/(np_j(1 - p_j)\sigma^2)$). We note that the p_j are actually the values $\hat{p}_j = m_j/n$, but $\hat{p}_j = p_j$ here since we are considering the situation where the values of m_j are equal to their expected values np_j. From (3.5)–(3.8) we then have

$$W_{L1}^2 \Big/ \left(\sum_{j=1}^{J} p_j(1 - p_j)\sigma^2 \right) \sim \chi^2_{1,\mathrm{nc}_{L1}},$$

where

$$\begin{aligned} \mathrm{nc}_{L1} &= n\frac{(\sum_{j=1}^{J} p_j(1 - p_j)\beta_j/\sigma)^2}{\sum_{j=1}^{J} p_j(1 - p_j)} \\ &= n\frac{(\sum_{j=1}^{J} \operatorname{sign}(\beta_j)\sqrt{p_j(1 - p_j)}\sqrt{EV_j})^2}{\sum_{j=1}^{J} p_j(1 - p_j)}, \end{aligned} \tag{3.9}$$

$$W_{Lp}^2/(J\sigma^2) \sim \chi^2_{1,\mathrm{nc}_{Lp}},$$

where

$$\begin{aligned}\mathrm{nc}_{Lp} &= n\frac{(\sum_{j=1}^J \sqrt{p_j(1-p_j)}\beta_j/\sigma)^2}{J} \\ &= n\frac{(\sum_{j=1}^J \operatorname{sign}(\beta_j)\sqrt{EV_j})^2}{J},\end{aligned} \tag{3.10}$$

$$W_C \sim \sum_{j=1}^J (np_j(1-p_j)\sigma^2)\chi^2_{1,\mathrm{nc}_j},$$

where $\mathrm{nc}_j = np_j(1-p_j)\beta_j^2/\sigma^2 = nEV_j$ as in equation (3.8), and

$$W_H \sim \chi^2_{J,\mathrm{nc}},$$

where $\mathrm{nc} = \sum_{j=1}^J \mathrm{nc}_j = n\sum_{j=1}^J EV_j \approx nEV$.

The above results show that the power of W_H depends (approximately) just on the total explained variation EV and sample size n, and it is not sensitive to the direction of the SNP effects [$\operatorname{sign}(\beta_j)$] nor the MAF p_j. Although the C-alpha statistic W_C uses "equal" weights for all SNPs, its power depends not only on the EV_js and n but also on the p_js, because the corresponding coefficients for the linear combination of independent χ^2_{1,nc_j} are proportional to $p_j(1-p_j)$, essentially giving *smaller weight to rarer variants*. The test statistic W_C has been found powerful in a wide range of settings for binary phenotypes [e.g., Neale et al. (2011), Basu and Pan (2011)]. For the most part, the settings investigated were ones where the regression coefficients β_js in a model for Y given $\mathbf{X}$ were unrelated to the p_js. In that case EV_j and nc_j tend to be smaller for rarer variants and a smaller weight is preferred. However, if larger $|\beta_j|$s are more likely to be found among rarer variants, then W_H could be more powerful than W_C. Simulations in Section 4 confirm this.

Powers of the linear statistics depend on the effect directions and on the weights. The effect of using weights inversely proportional to p_j [e.g., $W_{Lp} = \mathbf{w}'S$ with $w_j = 1/\sqrt{p_j(1-p_j)}$] is unclear, because nc_{Lp} in (3.10) is not necessarily bigger than nc_{L1} in (3.10) for W_{L1} with equal weights, even if rarer variants tend to have bigger genetic effects in terms of larger $|\beta|$ values. We provide numerical results on the power of W_{L1}, W_{Lp}, W_C and W_H under various conditions in Section 4 for studies of both quantitative and binary traits.

3.3 Additional Theoretical Results with More General Settings

Here we investigate the effects of dependency between genotypes. Due to genetic linkage, rate of recombination, genetic selection and other factors, genotypes of SNPs from the same chromosomal region may not be independent of each other at the population level, that is, $P(X_{ij}X_{ij'}) \neq P(X_{ij})P(X_{ij'})$. This phenomenon is also known as linkage disequilibrium [e.g., Reich et al. (2001)]. Similar to the previous section, we discuss results based on linear normal model (3.1) and score statistic $\mathbf{S} = (S_1, \ldots, S_J)'$ in (3.2). This statistic can be rewritten in vector form as

$$\mathbf{S} = X_c'\mathbf{Y}, \tag{3.11}$$

where X_c has (i, j) entry $X_{ij} - \overline{X}_j$ (where $\overline{X}_j = \sum_{i=1}^n X_{ij}/n$). Due to normality of Y, the distribution of $\mathbf{S}$ given genotypes $\mathbf{X}$ is multivariate normal,

$$\mathbf{S} \sim N(\boldsymbol{\mu}, \Sigma), \tag{3.12}$$

where $\boldsymbol{\mu} = E(\mathbf{S}) = X'X_c\boldsymbol{\beta} = X_c'X_c\boldsymbol{\beta}$ and $\operatorname{Var}(\mathbf{S}) = \Sigma = \sigma^2 X_c'X_c$. We denote $n\hat{\Sigma}_X = X_c'X_c$, an estimate of the covariance matrix of genotypes $\mathbf{X}$ and so $\boldsymbol{\mu} = n\hat{\Sigma}_X\boldsymbol{\beta}$ and $\Sigma = \sigma^2 n\hat{\Sigma}_X$. Under mutually independent genotypes, matrix Σ_X is approximately diagonal, $n\hat{\Sigma}_X = \operatorname{diag}\{m_1(1-m_1/n), \ldots, m_J(1-m_J/n)\}$, and we provided insights on the effect of the choice of linear and quadratic statistics for this covariance structure in Section 3.2. Here we give additional results for the general covariance structure. Similar to the previous sections, m_j and $m_{lj} = \sum_{i=1}^n X_{il}X_{ij}$ are treated as fixed values, and for simplicity we consider the case where m_j is equal to its expected value np_j and m_{lj} is equal to its expected value np_{lj}, where $p_{lj} = P(X_{il} = 1, X_{ij} = 1)$.

Similar to the previous section, we consider settings according to the variation of Y explained by the set of SNPs. Under model (3.1) and covariance structure Σ_X, the total phenotypic variation explained by the J SNPs is

$$\begin{aligned}EV &= \frac{\operatorname{Var}(E[Y|\mathbf{X}])}{\operatorname{Var}(Y)} = \frac{\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}}{\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta} + \sigma^2} \\ &\approx \frac{\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}}{\sigma^2}\end{aligned} \tag{3.13}$$

when explained variation is small. One should note that when genotypes are not mutually independent, the total explained variation by J SNPs is not approximately equal to the sum of the individual explained variations as in (3.4).

Again we consider the two linear statistics $W_{L1} = \mathbf{w}'\mathbf{S}$ with $w_j = 1$, W_{Lp} with $w_j = 1/\sqrt{p_j(1-p_j)}$ and two quadratic statistic $W_Q = \mathbf{S}'A\mathbf{S}$: W_C with $A = I$ (C-alpha) and Hotelling W_H with $A = \Sigma$. We note again that we are considering the situation where the values of m_j and m_{lj} are equal to their expected values np_j and np_{lj}, respectively, thus, $\hat{p}_j = p_j$ and $\hat{\Sigma}_X = \Sigma_X$. Let $\mathbf{U}\Lambda\mathbf{U}'$ be the eigendecomposition of matrix Σ_X, where $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_J\}$ consists of the eigenvalues of Σ_X and $\mathbf{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_J\}$ is an orthogonal matrix constructed from corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_J$. Based on the derivations in Section 2.2, the following distributional results hold:

(i) $W_{L1}^2/(\sigma^2\mathbf{1}'\Sigma_X\mathbf{1}) \sim \chi^2_{1,\text{nc}}$, with noncentrality parameter $\text{nc} = n\frac{(\mathbf{1}'\Sigma_X\boldsymbol{\beta})^2}{\sigma^2\mathbf{1}'\Sigma_X\mathbf{1}}$.

(ii) $W_{Lp}^2/(\sigma^2\mathbf{w}'\Sigma_X\mathbf{w}) \sim \chi^2_{1,\text{nc}}$, with noncentrality parameter $\text{nc} = n\frac{(\mathbf{w}'\Sigma_X\boldsymbol{\beta})^2}{\sigma^2\mathbf{w}'\Sigma_X\mathbf{w}}$ and $\mathbf{w} = (1/\sqrt{p_1(1-p_1)}, \ldots, 1/\sqrt{p_J(1-p_J)})'$.

(iii) $W_C \sim \sum_{j=1}^J \lambda_j \chi^2_{1,\text{nc}_j}$, with $\text{nc}_j = n\lambda_j(\mathbf{u}_j'\boldsymbol{\beta})^2/\sigma^2$.

(iv) $W_H \sim \chi^2_{\text{rank}(\Sigma_X),\text{nc}} = \sum_{j=1}^J I(\lambda_j > 0)\chi^2_{1,\text{nc}_j}$, with $\text{nc}_j = n\lambda_j(\mathbf{u}_j'\boldsymbol{\beta})^2$ and $\text{nc} = \sum_{j=1}^J n\lambda_j(\mathbf{u}_j'\boldsymbol{\beta})^2/\sigma^2 = n\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta}/\sigma^2 \approx nEV$.

The power of the Hotelling statistic W_H again depends solely on (approximate) explained variation by the J SNPs and $\text{rank}(\Sigma_X) = \sum_{j=1}^J I(\lambda_j > 0)$. If two different sets of J SNPs explain the same total phenotypic variation, then the power for W_H is the same for those two sets regardless of the correlation structure between SNPs, provided the corresponding Σ_Xs have the same rank. This also implies that when two sets of J SNPs explain the same total phenotypic variation, the Hotelling statistic is more powerful for the set of SNPs where Σ_X has lower rank. A second conclusion is that power of the other three statistics depends on the covariance structure of the SNPs, $\hat{\Sigma}_X$, and their effects $\boldsymbol{\beta}$. In fact, when two sets of J SNPs explain the same total phenotypic variation and one of the sets consists of mutually independent SNPs, the power of these three tests for the set of independent SNPs is not necessary larger than the power for another set of SNPs with a different covariance structure. This is confirmed by our empirical evaluations presented in supplementary materials [Derkach, Lawless and Sun (2013a)].

4. NUMERICAL POWER COMPARISONS

We conducted extensive and novel simulation studies to examine the finite sample performance of linear and quadratic statistics. Since there is little background information suggesting what genetic scenarios are most plausible, we generated data from over 10,000 different genetic models that involve varying proportions of protective, deleterious and neutral variants, variant frequencies, effect sizes, and relationships between variant frequencies and effect sizes. Careful analysis of the results provides considerable insight into the performance of different statistics. The statistics considered here are the two linear statistics, $W_{L1} = \mathbf{1}'\mathbf{S}$, $W_{Lp} = \mathbf{w}'\mathbf{S}$, where $w_j = 1/\sqrt{p_j(1-p_j)}$, and two quadratic statistics $W_C = \mathbf{S}'I\mathbf{S}$ and $W_H = \mathbf{S}'\Sigma^{-1}\mathbf{S}$, as discussed in Section 3.2 and Table 1. Estimation of the p_j is discussed in Sections 4.1 and 4.2 below.

We studied both quantitative and binary traits. Table 2 describes the simulation models considered. For each type of trait, we considered two types of scenarios, S1 ("MAF-effect independent") assumes that $|\beta_j|$ (the size of the genetic effect) of a causal SNP j is unrelated to p_j (approximately twice the MAF), and S2 ("MAF-effect dependent") assumes that $|\beta_j|$ is inversely related to p_j. For normally distributed quantitative traits, the MAF-effect dependent models were simulated by directly specifying the phenotypic variance explained by SNP j, $EV_j = (\beta_j\sqrt{p_j(1-p_j)})^2/\sigma^2$, and without loss of generality we take $\sigma^2 = 1$. We did not restrict all causal variants to have the same direction of effect, but assumed that the majority of the causal variants have the same direction with $p_D = J_D/J_C$ ranging from 75% to 100%, a reasonable assumption based on what has been reported in the literature. (We also simulated models where p_C ranges from 50% to 75%; the linear statistics performed poorly and were dominated by the quadratic statistics, as one would expect.) Here we assume that the genotypes of different SNPs are mutually independent, but Section 5 considers possibly nonindependent genotypes obtained from sequence data of the 1000 Genomes Project [1000 Genomes Project Consortium (2010)]. We also conducted additional simulation studies examining the effect of dependency between SNPs on power, supporting conclusions made in Section 3.3 above.

4.1 Quantitative Traits

We first considered the normal linear model in (3.1) for which results in Section 3.2 give the power of the different statistics. Results presentation and discussion focus on $n = 1000$ and type 1 error $\alpha = 10^{-4}$. (Other n and α values were also considered, but results are

TABLE 2

*Parameters and parameter values of simulated models for studies of quantitative or binary traits. Scenario S*1 (*MAF-effect independent*) *assumes MAFs and effect sizes are mutually independent. Scenario S*2 (*MAF-effect dependent*) *assumes that variants with smaller MAFs tend to have bigger effect sizes*

	Parameters	**Parameter values**
n	Sample size ($n_{\text{case}} = n_{\text{control}} = n/2$ 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 scenario S1 (*MAF-effect independent*)*;* 10,000 *independently simulated models*	
p_j	Approximately twice the MAF of SNP j	Unif(0.005, 0.02)
β_j	Regression coefficient in (3.1) 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)
	Quantitative traits under scenario S2 (*MAF-effect dependent*)*;* 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)
	Binary traits under scenario S1 (*MAF-effect independent*)*;* 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)
	Binary traits under scenario S2 (*MAF-effect dependent*)*;* 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)

qualitatively similar across tests.) The choice of $\alpha = 10^{-4}$ is to reflect the fact that testing would typically be conducted for multiple genetic regions. Table 2 shows the combination of factors and indicates how data from 10,000 different models were generated.

For each of the 10,000 randomly generated genetic models we used critical values according to the exact distributions in Section 3.1 to compute power. Specifically, for each model we considered a sample of size $n = 1000$ for which the m_j equaled their expected values np_j. Thus, $\hat{p}_j = p_j$ for each SNP and the J by J covariance matrix Σ in (3.3) equals $\text{diag}\{np_j(1 - p_j)\sigma^2\}$ under both the null ($\boldsymbol{\beta} = 0$) and alternative hypothesis represented by the genetic model. Since n is large, we ignored the effect of estimating σ^2 (as in Section 3.1) and used the true value $\sigma^2 = 1$; this has a negligible effect on power. The use of $\hat{p}_j = p_j$ deserves discussion, since in practice the value $\hat{p}_j$ will vary from sample to sample. However, they are functions only of the covariates X_{ij} and so no adjustments to the distribution in Section 3.1 are needed. However, the power provided by using (3.5) or (3.7) with the p_j estimated with $\hat{p}_j$ are conditional, that is, they apply to samples with the described set of values m_j. Unconditional power is also of interest; this reflects sampling variation in the m_j (and $\hat{p}_j$). Unconditional power is calculated (or estimated) by averaging conditional powers for the case where $m_j = np_j$ in this sec-

tion. In the supplementary materials [Derkach, Lawless and Sun (2013a)] we provide some unconditional power values. We find that differences with the conditional powers are small (see Figures S6 and S7).

For visual display, Figure 1 shows the within-class power comparisons (linear W_{Lp} vs. linear W_{L1}, and quadratic W_H vs. quadratic W_C) of the four tests for 1000 models randomly selected from the 10,000 independently generated models. In view of the wide variations in model parameters, powers of the tests vary widely across the 1000 models. For each model, powers of the two linear statistics are similar and likewise for powers of the two quadratic statistics. Moreover, under scenario S1 [Figure 1(a)] neither statistic within each class dominates the other across the 1000 models. However, under scenario S2 [Figure 1(b)], the Hotelling statistic performs better than the C-alpha statistic for almost all models, as our earlier comments in Section 3.2 suggest. In this case, we also see that the linear statistic using weights inversely proportional to MAFs does not always lead to a better power even when the assumption that rarer variants have bigger effects is in fact true here [Figure 1(b)].

We also considered simulations with sample sizes $n = 500$ and 2000, to see the effect on the linear versus quadratic statistic comparison. For simplicity we show plots for W_{L1} and W_H; plots for W_{Lp} and W_C are very similar. Figure 2 and Table 3 show that which type of statistic is better depends on the sample size and the model parameters. When $n = 500$, both the

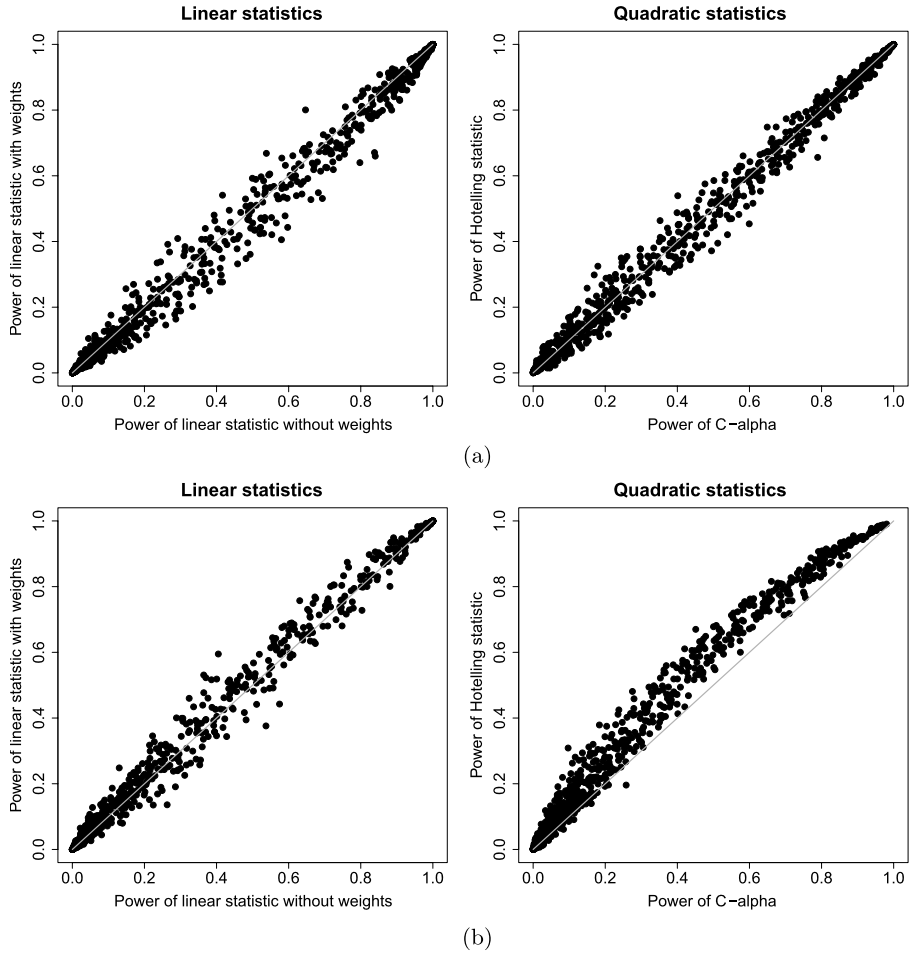

FIG. 1. *Within-class power comparison of the four statistics for* 1000 *independently generated models for studies of QUANTITATIVE traits under* (a) *scenario S*1 (*MAF-effect independent*) *and* (b) *scenario S*2 (*MAF-effect dependent*) *as described in Table 2. The four statistics are the two linear statistics* $W_L = (w_1, \ldots, w_J)'\mathbf{S}$ *in* (2.4): *"without weights"* W_{L1} *where* $w_j \equiv 1$ *and "with weights"* W_{Lp} *where* $w_j = 1/\sqrt{p_j(1-p_j)}$, *and two quadratic statistics* $W_Q = \mathbf{S}'A\mathbf{S}$ *in* (2.5): *the C-alpha statistic* W_C *where* $A = I$ *and the Hotelling statistic* W_H *where* $A = \Sigma_S^{-1}$. *Sample size* $n = 1000$ *and type* 1 *error* $\alpha = 10^{-4}$. *The set of* 1000 *models presented here is a random subset of all the* 10,000 *models independently generated.*

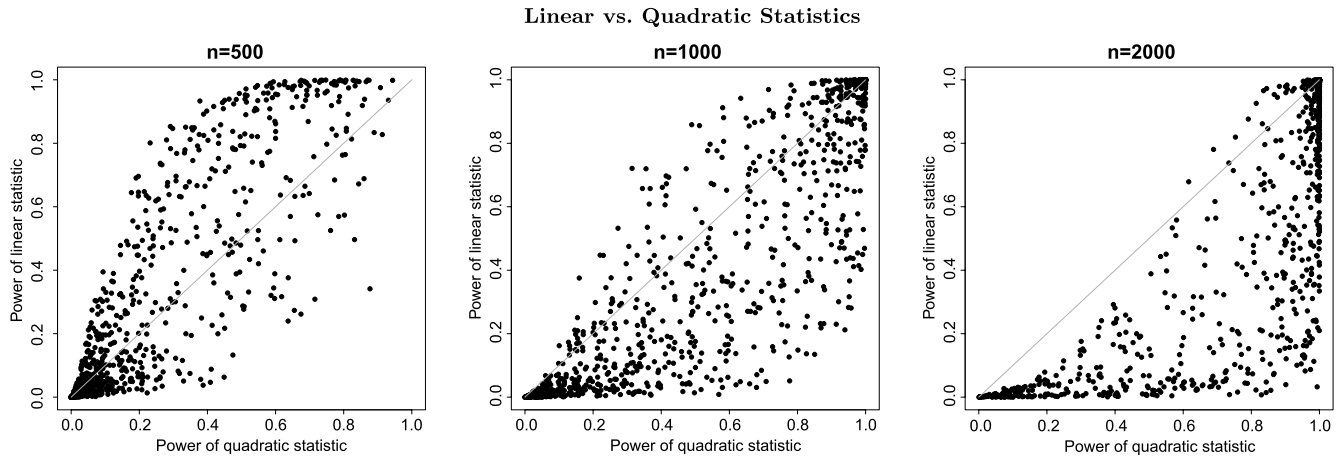

FIG. 2. *Between-class power comparison of the linear statistic W_{L1} vs. the quadratic Hotelling statistic W_H for studies of QUANTITATIVE traits under scenario S*1 (*MAF-effect independent*). *Other details see Figure* 1.

linear and quadratic statistics have low power (more than 65% of the 1000 models have power <20%; Table 3). In that case, good power (80%) is achieved only for those models with high proportions of causal SNPs (among which the proportion of deleterious SNPs is at least 75% by study design); the linear statistic is better than the quadratic statistic. However, as n increases, the quadratic statistic displays good power across many models and by $n = 2000$ dominates the linear statistic for most of the models. Similar conclusions can be made based on results from the models simulated under scenario S2 (see supplementary materials Figure S1 [Derkach, Lawless and Sun (2013a)]).

To better understand the impact of the various model parameters on different statistics, Figure 3 presents power from a different perspective showing the individual power of the linear statistic W_{L1} [Figure 3(a)] and the quadratic statistic W_H [Figure 3(b)] 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 $J = 30$ under the scenario S1. Results for scenario S2 are in supplementary materials Figure S2; results for $J = 10$, 20, 40 and 50 are qualitatively similar and not shown. It is clear that the power of both tests depends highly on the percentage of causal SNPs in the group of SNPs investigated. For example, among the 10,000 models giving power of 50% or greater, the average proportion of causal SNPs (p_C) is 81% (SE = 13% and min = 42%) for the linear test and 81% (SE = 12% and min = 50%) for the quadratic test. The powers for the quadratic statistics vary much less than those for the linear statistics; this is due to the latter's need for both p_C and p_D (the proportion of

TABLE 3

*Breakdown of the power of the linear statistic W_{L1} and the quadratic Hotelling statistic W_H under scenario S*1 (*MAF-effect independent*). *Proportions of the* 1000 *models in Figure* 2 *that have power in the specified ranges. For other details see Figures* 1 *and* 2 *legends*

	Power range				
Sample size	**0–20%**	**20–40%**	**40–60%**	**60–80%**	**80–100%**
	Proportion of the models in power range; W_{L1}				
$n = 500$	0.66	0.11	0.06	0.06	0.11
$n = 1000$	0.46	0.11	0.08	0.07	0.28
$n = 2000$	0.30	0.08	0.06	0.07	0.49
	Proportion of the models in power range; W_H				
$n = 500$	0.68	0.14	0.09	0.07	0.02
$n = 1000$	0.32	0.13	0.10	0.10	0.35
$n = 2000$	0.10	0.07	0.06	0.07	0.70

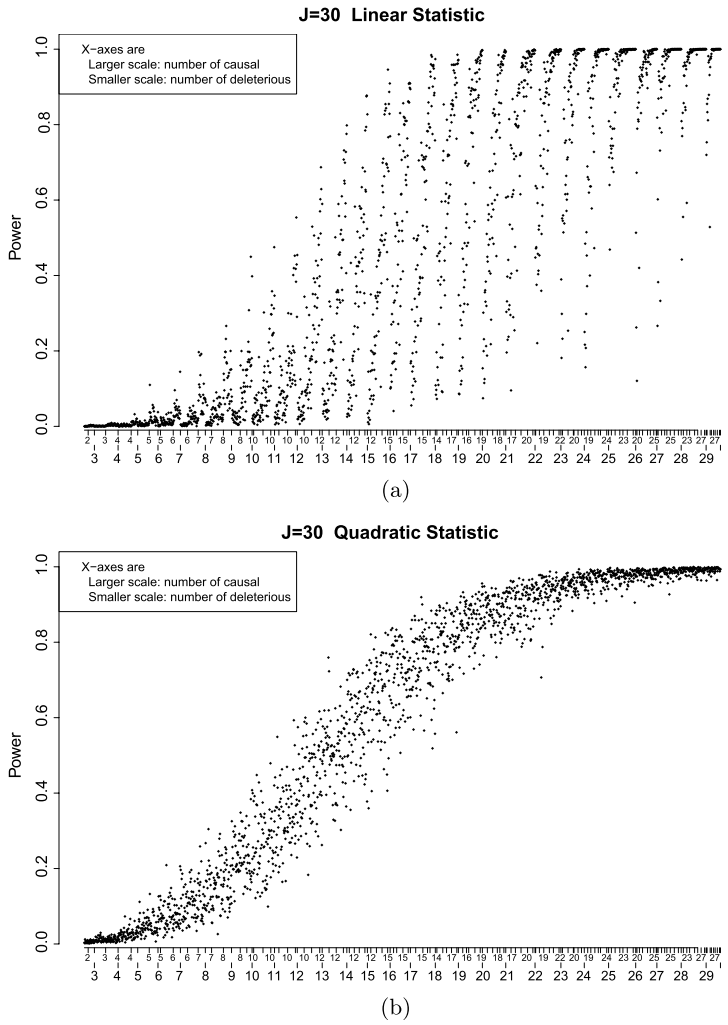

FIG. 3. *Individual power of* (a) *the linear statistic* W_{L1} *and* (b) *the quadratic Hotelling statistic* W_H *for studies of QUANTITATIVE traits under scenario S*1 (*MAF-effect independent*) *for models with* $J = 30$ *total number of rare variants. The large scale of the X-axis shows the number 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 in the range of* $J_D = J_C \cdot p_D = J_C \cdot 75\%$ *to* $J_D = J_C \cdot 100\%$, *depending on the actual number of causal variants in a model. The* 2005 *models shown here are the models with* $J = 30$ *among the* 10,000 *models generated as described in Table* 2. *Sample size* $n = 1000$ *and type* 1 *error* $\alpha = 10^{-4}$.

deleterious SNPs among the causal ones) being close to 1 in order to achieve high power.

To examine the effect of correlation between SNPs on power, we conducted additional simulation studies. Briefly, we considered two types of correlation scenarios (D1: correlation among casual variants and D2: correlation between causal and neutral variants) and compared power of the four tests (W_{L_1}, W_{L_p}, W_C, W_H) to the independence case, under two different assumptions of the corresponding genetic effects (E1: total explained variation by all causal variants is fixed and E2: the regression coefficient β_js are fixed). Under E1, neither correlation structure affects power of W_H; however, D1 increases power of the other three tests while D2 can increase or decrease power. Under E2, D1 increases power of all four tests; D2 once again can increase or decrease power. Details of the simulation study design and results (Figures S8–S11) are in the supplementary material [Derkach, Lawless and Sun (2013a)].

4.2 Binary Traits

Here, we provide detailed numerical results for case–control studies involving a binary trait Y, where a normal approximation for $\mathbf{S}$ might not be adequate. As in Section 4.1, we examine the performance of W_{L1},

W_{Lp}, W_C and W_H. We assume that the distribution of Y_i given $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})'$ is Bernoulli with

$$\text{(4.1)}\quad \text{Prob}(Y_i = 1|\mathbf{X}_i) = \frac{\exp(\beta_0 + \sum \beta_j X_{ij})}{1 + \exp(\beta_0 + \sum \beta_j X_{ij})},$$

and that the X_{ij}s in the population are mutually independent Bernoulli variables with $P(X_{ij} = 1) = p_j$ for $j = 1, \ldots, J$. We first used asymptotic distributions for the linear and quadratic statistics provided in Section 2.3 to obtain p-values, and we evaluated type I error rate and obtained empirical critical values for each of the four tests (supplementary materials Table S1). In this case the test statistics are based on (2.3) with the covariance matrix given by (A.3) in the supplementary materials [Derkach, Lawless and Sun (2013a)]. Unlike the quantitative traits above, the SNP genotypes X_{ij} here vary from sample to sample and thus so do the values $\hat{p}_j$ $(j = 1, \ldots, J)$. Supplementary Table S1 shows that normal approximations are satisfactory for the linear statistics but chi-square approximations for the quadratic statistic produce p-values (and thus critical values) that are much too conservative. We conducted simulations to assess power under different scenarios, using empirical critical values for the quadratic statistics. The simulation of case–control data is discussed in the supplemetary materials [Derkach, Lawless and Sun (2013a)]. Given the amount of computation required, we considered 500 models randomly generated under each of the two MAF-effect scenarios described in Table 2.

Results in Figure 4 are slightly different from those in Figure 1 for quantitative traits. Under scenario S1

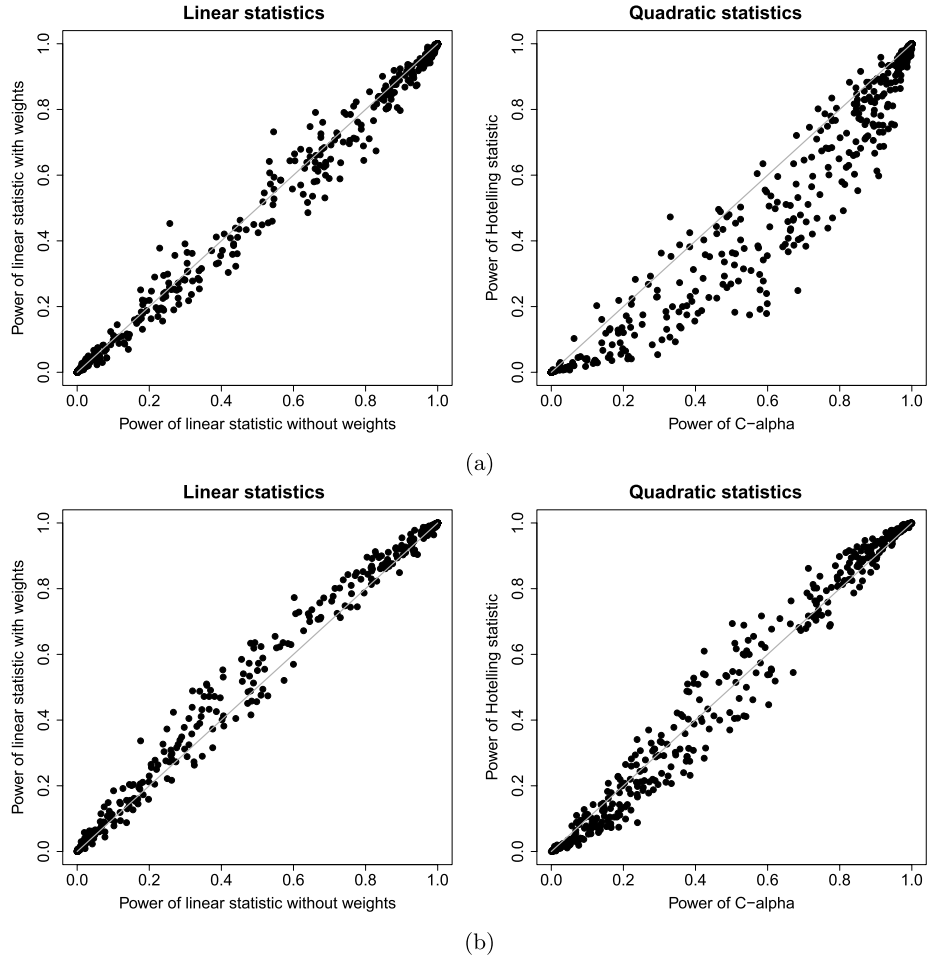

FIG. 4. *Within-class power comparison of the four statistics for* 500 *independently generated models for studies of BINARY traits under* (a) *scenario S*1 (*MAF-effect independent*) *and* (b) *scenario S*2 (*MAF-effect dependent*) *as described in Table* 2. *The four statistics are the two linear statistics* $W_L = (w_1, \ldots, w_J)'\mathbf{S}$ *in* (2.4): *"without weights"* W_{L1} *where* $w_j \equiv 1$ *and "with weights"* W_{Lp} *where* $w_j = 1/\sqrt{p_j(1-p_j)}$, *and two quadratic statistics* $W_Q = \mathbf{S}'A\mathbf{S}$ *in* (2.5): *the C-alpha statistic* W_C *where* $A = I$ *and the Hotelling statistic* W_H *where* $A = \Sigma_S^{-1}$. *Sample size* $n = 1000$ *and type* 1 *error* $\alpha = 10^{-4}$.

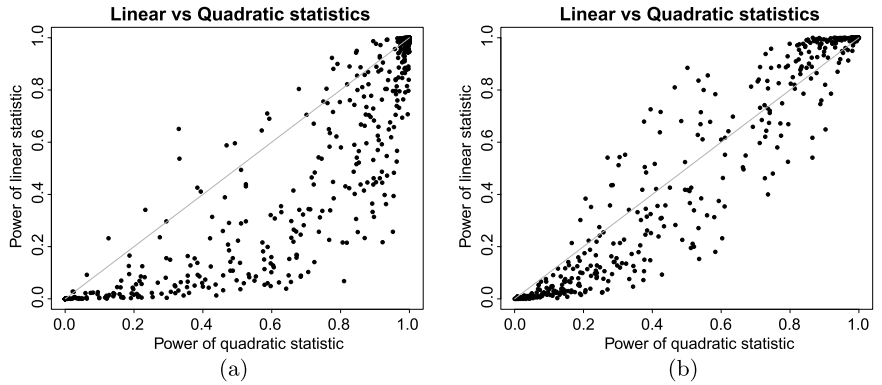

FIG. 5. *Between-class power comparison of the two statistics for* 500 *independently generated models for studies of BINARY traits under* (a) *scenario S*1 (*MAF-effect independent*) *and* (b) *scenario S*2 (*MAF-effect dependent*) *as described in Table* 2. *The linear statistic is* W_{L1} *and the quadratic statistic is C-alpha statistic* W_C. *For other details see Figure* 4 *legends*.

[Figure 4(a), left panel], neither of the two linear statistics dominates the other, which is similar to the case for quantitative traits [Figure 1(a), left panel]. Between the two quadratic statistics [Figure 4(a), right panel], W_C is more powerful than W_H; this is consistent with the findings of Basu and Pan (2011) discussed in Section 3.2. However, the systematic power difference between W_C and W_H is absent under scenario S2 [Figure 4(b), right panel]. This supplements the picture provided by Basu and Pan (2011), who did not consider cases where genetic effects are inversely proportional to MAFs, and it supports our earlier comment that the relative performance of W_C and W_H depends on the relationship between SNP effects and MAFs.

Under the MAF-effect dependent assumption, the linear statistic W_{Lp} appears to be consistently better than W_{L1} across the 500 models [Figure 4(b), left panel]. However, we emphasize that the apparent better power for W_{Lp} is mainly driven by the use of true variant frequency p_j values in the weight specification, $w_j = 1/\sqrt{p_j(1-p_j)}$. These would be unavailable to us in a real situation. In practice, how to estimate p_j can have major impacts on the validity of the test as well as on power. Some authors have suggested using the control sample only [e.g., Madsen and Browning (2009)], but it is not clear if the standard permutation-based approach for p-value estimation as used here is still valid. An additional concern for this approach is the possibility of a deleterious effect. In that case, which subsample is the proper "control" sample is not clear. If both cases and controls were used to estimate p_j, $\hat{p}_j$ would tend to be bigger than p_j for a causal SNP j because of the oversampling of cases, while $\hat{p}_{j'}$ is expected to be $p_{j'}$ for a neutral SNP j'. Consequently, using $w_j = 1/\sqrt{\hat{p}_j(1-\hat{p}_j)}$ downweights a causal SNP compared to a neutral one with the same frequency, resulting in loss of power. This is clear from the results shown in supplementary materials Figure S3 for both the MAF-effect independent and dependent scenarios. The practical use of weights, particularly for linear statistics, therefore, must be carefully considered in the case–control setting.

Figure 5 compares the power of W_{L1} and W_C across the 500 models. Under scenario S1 [Figure 5(a)], the quadratic statistic has better power than the linear statistic for the majority of the models. Under scenario S2 [Figure 5(b)], among the models with power less than 50%, the quadratic statistic has better power, but among the models with higher power, the linear statistic is more often better.

5. APPLICATION TO THE GAW17 DATA

The numerical studies in the previous section focused on mutually independent SNPs, although the tests themselves do not require this [see supplementary materials (Derkach, Lawless and Sun (2013a)] for additional simulation studies on dependent SNPs). To consider settings where this might not be so along with real sequence data, we examined real human sequence data [1000 Genomes Project Consortium (2010)] that were used to generate the GAW17 phenotype data [Almasy et al. (2011)] introduced in Section 1.

We consider here quantitative trait Q2 which is influenced by 72 SNPs in 13 genes but not by other covariates; recall from Section 1 that traits were simulated, so it is known which SNPs are causal. To assess the performance of association statistics, we carried out

"pseudo power" comparisons by determining the p-values for each of four test statistics, across each of the 13 genes, using the 200 replicate samples available (same genotype data but different phenotype data, independently simulated, based on the true genotype–phenotype association model).

We used data from the $n = 321$ unrelated Asian subjects (Han Chinese, Denver Chinese and Japanese) and excluded SNPs that had MAF $>5\%$ or were monomorphic within the Asian sample. Gene *VNN1* had no causal rare variant but it was kept in the analysis to serve as a negative control. The threshold MAF $\leq 5\%$ does not reduce the number of causal SNPs much (70 of the 72 causal SNPs have MAF $\leq 5\%$), but it reduces the number of neutral SNPs in a gene and therefore increases power.

For each of the 200 replicates, we calculated permutation-based p-values for the four statistics, W_{L1}, W_{Lp}, W_C and W_H (see Table 1). We estimated power for $\alpha = 0.05$ by the proportion of the 200 replicates for which the empirical p-values were ≤ 0.05 for each test. For each sample, gene and statistic combination, the p-value for the null hypothesis of no association was obtained from the permutation distribution by randomly generating 10,000 permutations of each replicate sample.

The choice of the liberal type 1 error $\alpha = 0.05$ was based on the low power of detecting genetic effects of sizes represented by the simulation models, with a sample of 321 people. Table 4 summarizes the rare variants for the 13 genes and gives the empirical power for each statistic. Only the first group of 9 genes have maximum power above 10%.

Results in Table 4 are consistent with our previous conclusions: (i) linear tests with and without weights based on MAF vary in relative power but not substantially; (ii) quadratic statistics W_C and W_H also have slightly variable relative power; (iii) between-class performance is highly variable. As expected, linear statistics outperform quadratic statistics if the proportion of causal variants is not too low (e.g., genes *SIRT*1 and *SREBF*1), but the pattern can be reversed if this is not the case, even when the effects in this data are all in the same direction (e.g., *BCHE* and *RARB*).

6. DISCUSSION AND RECOMMENDATIONS

We have reviewed and studied tests of association between rare variants and phenotypes within a unified

TABLE 4

*Power of the four test statistics applied to the GAW*17 *sequence data provided by the* 1000 *Genomes Project. The* 13 *genes presented here are all the causal genes for simulated quantitative trait Q*2. *VNN*1 *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* ($n = 321$). *VNN*1 *is kept in the analysis to serve as a negative control. All causal variants were designed by GAW*17 *to have the same direction of effects* (*minor alleles were associated with higher Q*2 *values*). *The average genetic effect is the average of regression coefficient* β *values of the causal variants used to simulate Q*2 (*effects are independent of populations by the GAW*17 *design*). *Genes are ordered according to the maximum power of the four tests which is bolded. Powers shown vary considerably due to inherent factors and estimation based only on* 200 *replicates, and the* 13 *genes are separated into different groups*

Gene	SNP distribution J_C, J_N	Ave. MAF of J_C, J_N	Avg. effect of J_C	Power			
				Linear W_{Lp}	Linear W_{L1}	Quadratic W_C	Quadratic W_H
9 genes for which the maximum power is 10% or more							
*SIRT*1	4, 7	0.27%, 0.22%	0.71	**0.44**	0.40	0.25	0.39
BCHE	5, 10	0.22%, 0.19%	0.72	0.29	0.35	**0.43**	0.39
PDGFD	3, 6	0.78%, 0.65%	0.74	0.29	0.43	**0.45**	0.35
*SREBF*1	4, 5	0.39%, 0.40%	0.52	**0.49**	0.47	0.18	0.28
GCKR	1, 0	1.21%, NA	0.38	**0.25**	**0.25**	**0.25**	**0.25**
RARB	1, 5	0.78%, 0.90%	0.64	0.06	0.03	0.07	**0.14**
PLAT	4, 7	0.39%, 0.49%	0.68	**0.13**	**0.13**	0.06	**0.13**
VLDLR	4, 6	0.19%, 1.64%	0.75	**0.12**	0.08	0.06	0.09
*VNN*3	2, 2	0.16%, 2.57%	0.37	0.03	**0.10**	0.06	0.04
3 genes for which the maximum power is 10% or less							
*INSIG*1	3, 1	0.16%, 3.42%	0.20	0.06	0.06	0.04	0.03
LPL	1, 4	0.16%, 0.23%	0.73	0.02	0.03	0.06	0.05
VWF	1, 3	0.16%, 1.90%	0.34	0.02	0.01	0.03	0.01
1 gene for which there is no polymorphic rare causal variants in the Asian sample							
*VNN*1	0, 3	NA, 0.31%	NA	0.02	0.02	0.04	0.05

framework which gives theoretical insights about the performance of the methods (Table 1). Tests can have greatly varying power depending on the total number of rare variants, the numbers of deleterious, protective and neutral variants, the effect directions and the relationship between the effect sizes and the MAFs of causal variants. When substantial numbers of both deleterious and protective SNPs are present, quadratic test statistics are much better. They can also outperform linear statistics in settings where causal SNPs are all deleterious (or all protective), but a substantial fraction of the SNPs are not associated with the phenotype. However, our results also indicate that power to detect moderate levels of association is not high unless sample sizes are very large or a high proportion of the chosen SNPs are causal. Sequencing errors and other caveats concerning the data will further decrease power. Cases where power is substantial for smaller studies are predominantly ones where SNPs are almost all deleterious or all beneficial, and it is the linear test statistics that achieve highest power. Consequently, the definition of a chromosomal region and selection of SNPs within the region are critical to statistical inference regardless of the specific test used. In practice, a chromosomal region can be a gene, coding region of a gene or other types of genetic unit (e.g., a group of SNPs that are in moderate or strong linkage disequilibrium of each other); selection of SNPs within a region can be also based on relevant biological information since not all SNPs are equal a priori (e.g., some SNPs are believed to be more important than others based on functional genomic annotation). Different choices could lead to different statistical power [e.g., King, Rathouz and Nicolae (2010), Derkach et al. (2014)].

Our work complements that of Basu and Pan (2011), and a brief comparison is useful. They found similar results to ours in simulation studies for case–control scenarios, concerning the performance of linear statistics. Among the quadratic statistics, they found that the C-alpha/SSU type statistic $W_C = \mathbf{S}'I\mathbf{S}$ was generally the best and superior to the Hotelling statistics $\mathbf{S}'\Sigma^{-1}\mathbf{S}$. However, their simulation scenarios did not include cases where larger causal effects are associated with SNPs having smaller MAFs. Our numerical studies [scenario S2 under the MAF-effect dependent assumption in Table 2; Figure 1(b) for quantitative traits and Figure 4(b) for binary traits] and investigation of GAW17 data (Table 4) indicate the importance of the MAF-effect independent or nonindependent assumption on the choice of a good test statistic.

As an approach to rare variant testing in the absence of strong prior information, we support the recommendation of Basu and Pan (2011) to perform tests using both linear and quadratic statistics. In Derkach, Lawless and Sun (2013b) we investigated tests based on Fisher's method and the minimum-p method [e.g., Owen (2009)] for combing p-values from linear and quadratic statistics. Such tests were shown to be robust across the wide range of models considered here, in the sense of achieving power that is close to that of the better of a linear and quadratic statistic in a given setting. Comparisons were also made with the recent SKAT-O statistic of Lee, Wu and Lin (2012), which considers the minimum p-value across a class of statistics. The overall conclusion is that the Fisher's method outperforms the individual linear and quadratic tests as well as the minimum p-value approach, when the majority of the causal variants has the same direction of effect; however, the minimum p-value is better if (approximately) half of the causal variants are deleterious and the other half are protective.

It is beyond our scope here, but an empirical assessment of test statistics that involve covariate adjustment would be valuable. In addition, accurate and computationally efficient methods of obtaining p-values deserve attention. Parametric bootstrap simulation [e.g., Lin and Tang (2011)] can be used when sampling of individuals is random, but when it is trait-dependent matters are more complicated. In the case–control simulation for binary traits, for example, the sampling is effectively for X_i and other covariates v_i given Y_i. Methods that avoid detailed modeling of the distribution of (X_i, v_i) are desired. Empirical assessment is also difficult for family based association studies when samples are correlated. We hope to report on this in a future communication.

Finally, we reiterate our remarks made in Section 1 concerning the potential effects of sequencing errors. A realistic assessment of their scope and impact is called for.

ACKNOWLEDGMENTS

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) and the Canadian Institutes of Health Research (CIHR) grants to LS, NSERC to JFL, the

Ontario Graduate Scholarship (OGS) and the CIHR Strategic Training for Advanced Genetic Epidemiology (STAGE) fellowship to AD, University of Toronto. *Conflict of Interest*: None declared.

SUPPLEMENTARY MATERIAL

Pooled Association Tests for Rare Genetic Variants: A Review and Some New Results (DOI: 10.1214/13-STS456SUPP; .pdf). The supplementary materials include derivation of the permutation distribution of **S** for general traits, analytical results and simulation details for study of binary traits, simulation details for study of the effect of correlation between SNPs on power, and an additional 1 table and 11 figures for the studies of type 1 error rates and power for both quantitative and binary traits, for both MAF-effect independent and dependent scenarios, and for both independent and dependent rare variants.

REFERENCES

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

Almasy, L., Dyer, T. D., Peralta, J. M., Kent, J. W., Charlesworth, J. C., Curran, J. E. and Blangero, J. (2011). Genetic Analysis Workshop 17 mini-exome simulation. *BMC Proc.* **5 Suppl 9** S2.

Asimit, J. and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* **44** 293–308.

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

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

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

Daye, Z. J., Li, H. and Wei, Z. (2012). A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Res.* **40** e60.

Derkach, A., Lawless, J. F. and Sun, L. (2013a). Supplement to "Pooled association tests for rare genetic variants: A review and some new results." DOI:10.1214/13-STS456SUPP.

Derkach, A., Lawless, J. F. and Sun, L. (2013b). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genet. Epidemiol.* **37** 110–121.

Derkach, A., Lawless, J. F., Merico, D., Paterson, A. D. and Sun, L. (2014). Evaluation of gene-based association tests for analyzing rare variants using Genetic Analysis Workshop 18 data. *BMC Proc.* **8 Suppl 1** S9.

Duchesne, P. and Lafaye de Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Comput. Statist. Data Anal.* **54** 858–862. MR2580921

Goeman, J. J., van de Geer, S. A. and van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 477–493. MR2278336

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

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106** 9362–9367.

Hoffmann, T. J., Marini, N. J. and Witte, J. S. (2010). Comprehensive approach to analyzing rare genetic variants. *PLoS ONE* **5** e13584.

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

King, C. R., Rathouz, P. J. and Nicolae, D. L. (2010). An evolutionary framework for association testing in resequencing studies. *PLoS Genet.* **6** e1001202.

Ladouceur, M., Dastani, Z., Aulchenko, Y. S., Greenwood, C. M. T. and Richards, J. B. (2012). The empirical power of rare variant association methods: Results from sanger sequencing in 1998 individuals. *PLoS Genet.* **8** e1002496.

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

Li, Q. H. and Lagakos, S. W. (2006). On the relationship between directional and omnibus statistical tests. *Scand. J. Stat.* **33** 239–246. MR2279640

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

Lin, D.-Y. and Tang, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics* **89** 354–367.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5** e1000384.

Manolio, T. A., Brooks, L. D. and Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118** 1590–1605.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, Waltham, MA.

Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **615** 28–56.

Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34** 188–193.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D. et al. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet.* **7** e1001322.

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *Ann. Statist.* **37** 3867–3892. MR2572446

PAN, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.* **33** 497–507.

PRICE, A. L., KRYUKOV, G. V., DE BAKKER, P. I., PURCELL, S. M. et al. (2010). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics* **86** 832–838.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, Hoboken, NJ. MR0346957

REICH, D. E., CARGILL, M., BOLK, S., IRELAND, J., SABETI, P. C., RICHTER, D. J., LAVERY, T., KOUYOUMJIAN, R., FARHADIAN, S. F., WARD, R. and LANDER, E. S. (2001). Linkage disequilibrium in the human genome. *Nature* **411** 199–204.

SKOTTE, L., KORNELIUSSEN, T. S. and ALBRECHTSEN, A. (2012). Association testing for next-generation sequencing data using score statistics. *Genet. Epidemiol.* **36** 430–437.

SUL, J. H., BUHM, H. and ELEAZAR, E. (2011). Increasing power of groupwise association test with likelihood ratio test. *J. Comput. Biol.* **18** 1611–1624.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence Kernel association test. *The American Journal of Human Genetics* **89** 82–93.

YI, N. and ZHI, D. (2011). Bayesian analysis of rare variants in genetic association studies. *Genet. Epidemiol.* **35** 57–69.

YILMAZ, Y. E. and BULL, S. B. (2011). Are quantitative trait-dependent sampling designs cost-effective for analysis of rare and common variants? *BMC Proc.* **5 Suppl 9** S111.