DOI: 10.1534/genetics.110.125070

An Optimal Weighted Aggregated Association Test for Identification of Rare Variants Involved in Common Diseases

Jae Hoon Sul,* Buhm Han,* Dan He* and Eleazar Eskin*,†,1

*Computer Science Department and †Department of Human Genetics, University of California, Los Angeles, California 90095

Manuscript received November 12, 2010
Accepted for publication February 14, 2011

ABSTRACT

The advent of next generation sequencing technologies allows one to discover nearly all rare variants in a genomic region of interest. This technological development increases the need for an effective statistical method for testing the aggregated effect of rare variants in a gene on disease susceptibility. The idea behind this approach is that if a certain gene is involved in a disease, many rare variants within the gene will disrupt the function of the gene and are associated with the disease. In this article, we present the rare variant weighted aggregate statistic (RWAS), a method that groups rare variants and computes a weighted sum of differences between case and control mutation counts. We show that our method outperforms the groupwise association test of Madsen and Browning in the disease-risk model that assumes that each variant makes an equally small contribution to disease risk. In addition, we can incorporate prior information into our method of which variants are likely causal. By using simulated data and real mutation screening data of the susceptibility gene for ataxia telangiectasia, we demonstrate that prior information has a substantial influence on the statistical power of association studies. Our method is publicly available at http://genetics.cs.ucla.edu/rarevariants.

OVER the past few years, genome-wide association studies (GWAS) have identified many disease-causing variants (Corder *et al.* 1993; Bertina *et al.* 1994; Altshuler *et al.* 2000). Most of these studies are conducted by collecting common variants and perform a series of single-marker tests where each variant is tested individually to discover associations. However, only a small portion of disease heritability is explained by common variants, and several recent studies consider rare variants that collectively affect diseases (Cohen *et al.* 2004; Fearnhead *et al.* 2004; Kryukov *et al.* 2007; Romeo *et al.* 2007; Blauw *et al.* 2008; Bodmer and Bonilla 2008; Gorlov *et al.* 2008; International Schizophrenia Consortium 2008; Ji *et al.* 2008; Walsh *et al.* 2008; Xu *et al.* 2008). Since each rare variant is present in only a small number of individuals, single-marker tests have low power to identify these variants involved in disease. Hence, groupwise association tests that group rare variants in genes have received considerable attention as methods that increase the power of studies on rare variants, and a number of methods have been proposed such as the cohort allelic sums test (CAST) (Morgenthaler and Thilly 2007), the combined multivariate and collapsing (CMC) method (Li and Leal 2008), a weighted-sum statistic (Madsen and Browning 2009), and recently a variable-threshold approach (Price *et al.* 2010).

A groupwise association test is more complex than a single SNP association because there are many different ways of combining information across multiple variants. How the information from different variants is combined affects the statistical power of the association test, which also depends on the actual effect sizes of the variants on the disease phenotype. The challenge in developing groupwise association testing methods is that the underlying disease-risk model is not known.

In this article, we focus on a disease-risk model that is motivated by filling a blind spot in traditional GWAS. In this model, all variants including common variants make an equally small contribution to disease risk; that is, rarer variants are assumed to have higher effect sizes than common variants. Since each variant contributes only a small amount to the total disease risk, the single-marker test is not likely to detect associations in this disease-risk model, and thus this model describes associations usually not found in traditional GWAS. This is the same model discussed in Madsen and Browning (2009). Under this model, a weighted-sum statistic by Madsen and Browning (MB) is shown to be more powerful than other grouping methods such as CAST and CMC (Madsen and Browning 2009).

We propose a new method for the groupwise association test called the rare variant weighted aggregate statistic (RWAS). RWAS computes a weighted sum of differences between case and control mutation counts where weights are estimated from data to increase power of studies. The

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.125070/DC1.

[1]*Corresponding author:* Department of Computer Science and Department of Human Genetics, Mail Code 1596, 3532-J Boelter Hall, University of California, Los Angeles, CA 90095-1596.
E-mail: eeskin@cs.ucla.edu

optimal weights that maximize the power can be derived when the effect sizes of variants are known. When the true effect sizes are not known, RWAS approximates the optimal weights under the assumption that each variant makes an equal contribution to population disease risk. Simulations show that RWAS outperforms MB and the approximated weights achieve nearly the same power as the optimal weights under this assumed disease model. We also show how prior information on whether a variant is likely to be involved in a disease can be incorporated into RWAS. We first show through simulations that prior information greatly influences the statistical power of studies. Then, by using the real mutation screening data of the susceptibility gene for ataxia telangiectasia along with information of how likely a variant is to be deleterious (Tavtigian *et al.* 2009), we demonstrate that prior information plays a key role in this association study and RWAS is able to successfully detect the association in real data. The software package implementing RWAS is publicly available at http://genetics.cs.ucla.edu/rarevariants.

METHODS

Optimal weighted aggregate statistic: We consider an association study in which multiple variants within a gene affect the trait. For each variant, a difference in mutation counts between case and control individuals is computed, and a weighted sum of differences is used as a statistic for the group. This is in fact equivalent to computing a weighted sum of *z*-scores of variants where the *z*-score of a variant is computed from an allele frequency difference between cases and controls (Eskin 2008; Han *et al.* 2008; Zaitlen *et al.* 2010).

First, we assume that there are M rare variants in a group given $N/2$ case and $N/2$ control individuals. Let p_i denote population minor allele frequency (MAF) of variant i, and let $\widehat{p}_i^{+}$ and $\widehat{p}_i^{-}$ denote the observed MAF of case and control individuals in the sample, respectively. Then, the *z*-score of variant i (or the association statistic at variant i), denoted z_i, is calculated as

$$z_i = \frac{\widehat{p}_i^{+} - \widehat{p}_i^{-}}{\sqrt{2/N}\sqrt{\widehat{p}_i^{\pm}\left(1-\widehat{p}_i^{\pm}\right)}} \quad \left(\text{where } \widehat{p}_i^{\pm} = \frac{\widehat{p}_i^{+} + \widehat{p}_i^{-}}{2}\right). \tag{1}$$

The *z*-score approximately follows a normal distribution with variance equal to 1 and with mean equal to $\lambda_i\sqrt{N}$ [called the noncentrality parameter (NCP)],

$$\begin{aligned} z_i &\sim \mathcal{N}\left(\lambda_i\sqrt{N}, 1\right) \\ \lambda_i\sqrt{N} &= \frac{p_i^{+} - p_i^{-}}{\sqrt{2p_i^{\pm}\left(1-p_i^{\pm}\right)}}\sqrt{N} \quad \left(\text{where } p_i^{\pm} = \frac{p_i^{+} + p_i^{-}}{2}\right), \end{aligned} \tag{2}$$

where p_i^{+} and p_i^{-} are the true MAF of case and control individuals, respectively. Denoting γ_i as relative risk of variant i, p_i^{+} and p_i^{-} are

$$p_i^{+} = \frac{\gamma_i p_i}{(\gamma_i - 1)p_i + 1} \tag{3}$$

$$p_i^{-} = p_i \text{ (assuming the disease prevalence is very small).} \tag{4}$$

Let w_i be a weight of variant i. Then, a weighted sum of *z*-scores (S) and its distribution are

$$S = \frac{\sum_{i=1}^{M} w_i z_i}{\sqrt{\sum_{i=1}^{M} w_i^2}} \sim \mathcal{N}\left(\frac{\sqrt{N}\sum_{i=1}^{M} w_i\lambda_i}{\sqrt{\sum_{i=1}^{M} w_i^2}}, 1\right). \tag{5}$$

The greatest power is achieved when the NCP is maximized, which is equivalent to maximizing the $\sum w_i\lambda_i/\sum w_i^2$ term. Using the Cauchy–Schwartz inequality, the NCP is maximized when $w_i = \lambda_i$. Therefore, the optimal weight for variant i is λ_i, and we call the weighted association method based on the optimal weights the optimal weighted aggregate statistic (OWAS). The OWAS is optimal under any disease-risk models, but determining optimal weights requires knowledge of relative risk and population MAF of variants according to the definitions of λ_i, p_i^{+}, and p_i^{-} (Equations 2–4). We can estimate the population MAF from observed MAF of case and control individuals (see supporting information, File S1 for details), but obtaining or estimating relative risk is often not easy. We note that if the numbers of cases ($N^{+}/2$) and controls ($N^{-}/2$) are unequal, we replace $\sqrt{N}$ above with $\sqrt{2N^{+}N^{-}/(N^{+}+N^{-})}$ and replace $\widehat{p}_i^{\pm}$ and $p_i^{\pm}$ in Equations 1 and 2 with $(N^{+}\widehat{p}_i^{+} + N^{-}\widehat{p}_i^{-})/(N^{+}+N^{-})$ and $(N^{+}p_i^{+} + N^{-}p_i^{-})/(N^{+}+N^{-})$, respectively, and the above results hold.

RWAS: Setting the weights for OWAS requires knowledge of the effect sizes that are unknown. To set the weights for our method without knowledge of the effect sizes, we assume a disease-risk model in which all variants have constant population attributable risk (PAR). In this model, each group of variants has a certain level of the group PAR, and each variant in the group has the same marginal PAR. Let ω denote the marginal PAR that is the group PAR divided by the number of causal variants in a group. Given ω and p_i of variant i, its relative risk, γ_i, is

$$\gamma_i = \frac{\omega}{(1-\omega)p_i} + 1. \tag{6}$$

Then, it follows from Equations 3 and 4 that

$$\begin{aligned} p_i^{+} &= \omega + p_i(1-\omega) \\ p_i^{-} &= p_i. \end{aligned} \tag{7}$$

The optimal weights (Equation 2) can be written as

$$\lambda_i = \frac{\omega(1-p_i)}{\sqrt{2p_i^{\pm}\left(1-p_i^{\pm}\right)}} \approx \omega\sqrt{\frac{1-p_i}{p_i}} \quad \left(\text{assuming } p_i^{\pm} \approx p_i\right). \tag{8}$$

Since ω in Equation 8 is fixed for all variants, we can ignore it and derive an analytically approximated form of the optimal weights as

$$w_i = \sqrt{\frac{1-p_i}{p_i}}. \tag{9}$$

We call the weighted sum of z-scores whose weights are approximated in Equation 9 the RWAS. The statistic of RWAS, S_{RWAS}, can be formulated as

$$S_{\mathrm{RWAS}} = \frac{\sum w_i z_i}{\sqrt{\sum w_i^2}} \approx \frac{\sum\left(\left(\hat{p}_i^{+} - \hat{p}_i^{-}\right)/\hat{p}_i^{\pm}\right)}{\sqrt{2/N}\sqrt{\sum\left(\left(1-\hat{p}_i^{\pm}\right)/\hat{p}_i^{\pm}\right)}} \sim \mathcal{N}\left(\frac{\sum\left(\left(p_i^{+} - p_i^{-}\right)/p_i^{\pm}\right)}{\sqrt{2/N}\sqrt{\sum\left(\left(1-p_i^{\pm}\right)/p_i^{\pm}\right)}}, 1\right)\left(\text{assuming } \hat{p}_i^{\pm} \approx p_i\right). \tag{10}$$

We compare S_{RWAS} to the standard normal distribution to obtain a P-value.

Approximation of MB to a sum of z-scores: Our methods (OWAS and RWAS) adopt a weighted sum of z-scores approach, and MB can also be approximated as a weighted sum of z-scores with weights equal to 1 (or an unweighted sum of z-scores). MB computes a statistic, denoted as z_{MB}, as follows. It can be decomposed into a sum of z_{MB_i} over M variants where i corresponds to the ith variant and M is the number of variants,

$$z_{\mathrm{MB}} = \frac{x - \hat{\mu}}{\hat{\sigma}} = \sum_{i=1}^{M} z_{\mathrm{MB}_i} = \sum_{i=1}^{M} \frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i}, \tag{11}$$

where for variant i, x_i is the sum of ranks or genetics scores of cases, and $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the average and standard deviation of x_i in the null distribution, respectively. We use the sum of genetic scores of cases as x_i since its power is very similar to the power of the sum of ranks (Madsen and Browning 2009).

Then, x_i, $\hat{\mu}_i$, and $\hat{\sigma}_i$ can be approximated as

$$x_i = \frac{\sqrt{N}}{2}\frac{\hat{p}_i^{+}}{\sqrt{\hat{p}_i^{-}\left(1-\hat{p}_i^{-}\right)}}, \quad \hat{\mu}_i = \frac{\sqrt{N}}{2}\frac{p_i}{p_i(1-p_i)}, \quad \hat{\sigma}_i = \sqrt{\frac{1}{2}} \tag{12}$$

(see File S1 for details) and the standardized statistic at variant i, z_{MB_i}, can be derived as

$$z_{\mathrm{MB}_i} = \frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i} \approx \frac{\left(\sqrt{n}/2\right)\left(\hat{p}_i^{+}/\sqrt{\hat{p}_i^{-}\left(1-\hat{p}_i^{-}\right)} - p_i/\sqrt{p_i(1-p_i)}\right)}{1/\sqrt{2}}$$

$$\approx \sqrt{\frac{N}{2}}\frac{\hat{p}_i^{+} - \hat{p}_i^{-}}{\sqrt{\hat{p}_i^{-}\left(1-\hat{p}_i^{-}\right)}}\left(\text{assuming } p_i \approx \hat{p}_i^{-}\right).$$

Finally, a sum of z_{MB_i} over M variants is equivalent to the original statistic of MB:

$$z_{\mathrm{MB}} = \frac{x - \hat{\mu}}{\hat{\sigma}} = \sum_{i=1}^{M} z_{\mathrm{MB}_i} = \sum_{i=1}^{M} \frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i} = \sum_{i=1}^{M} \sqrt{\frac{N}{2}}\frac{\hat{p}_i^{+} - \hat{p}_i^{-}}{\sqrt{\hat{p}_i^{-}\left(1-\hat{p}_i^{-}\right)}}. \tag{13}$$

Note that Equation 13 shows MB is an unweighted sum of z-scores. One difference between z_{MB_i} in Equation 13 and the z-score used in our methods (Equation 1) is the way the population MAF is estimated, which appears in the denominator of the z-score; MB estimates it only from control individuals, but we estimate it from all case and control individuals (see File S1 for details).

RWAS with prior information: RWAS can be directly extended to incorporate prior knowledge about the degree that each variant is believed to be causal. Note that the underlying truth is that each variant is either causal or not. Thus, let V^i be the variable indicating the "causal status" of variant i, such that $V^i = 1$ if variant i is causal and $V^i = 0$ if not. Let $V = \{V^1, \ldots, V^M\}$ denote the causal statuses of all M variants. V can have 2^M possible values. Let υ_j be the jth value of 2^M possible values. That is, $\upsilon_j = \{\upsilon_j^1, \ldots, \upsilon_j^M\}$ is an ordered set of 0 and 1 that represents a specific scenario of causal statuses.

Assume that we have prior knowledge that the probability of variant i being causal is c_i. Then, the probability of each scenario υ_j can be computed as

$$P(\upsilon_j) = \prod_{i=1}^{M} c_i^{\upsilon_j^i}(1-c_i)^{1-\upsilon_j^i}. \tag{14}$$

Then, the expected noncentrality parameter of the weighted sum of z-score statistics is

$$E[\mathrm{NCP}] = \sum_{j=1}^{2^M} P(\upsilon_j)\sqrt{N}\,\frac{\sum_{i=1}^{M} w_i(\upsilon_j^i \lambda_i)}{\sqrt{\sum_{i=1}^{M} w_i^2}} \tag{15}$$

$$= \frac{\sum_{i=1}^{M} c_i w_i \lambda_i}{\sqrt{\sum_{i=1}^{M} w_i^2}}. \tag{16}$$

The Cauchy–Schwartz inequality shows that this quantity is maximized when $w_i = c_i\lambda_i$. Thus, the prior knowledge $\{c_i\}$ can be easily incorporated into the RWAS by multiplying the prior probability into each weight.

Web resources: The software package for RWAS is publicly available online at http://genetics.cs.ucla.edu/rarevariants.

RESULTS

Power comparison between RWAS and MB: We evaluate the power of our novel method, RWAS, in

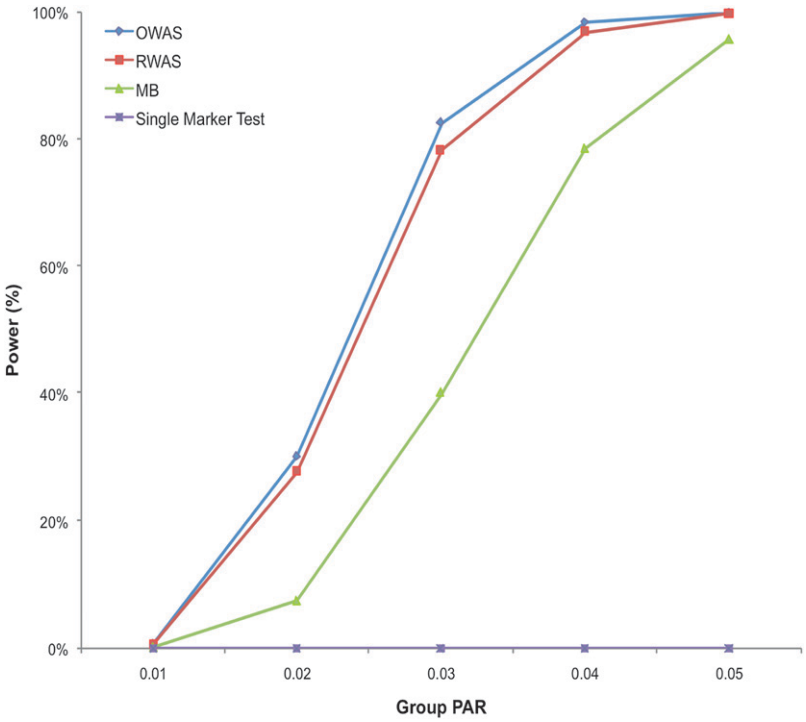

FIGURE 1.—Power comparison in the constant PAR model. There are five group PAR levels (1%, 2%, 3%, 4%, and 5%), and for each group PAR, 10,000 data sets were generated. Each data set contained 1000 case and 1000 control individuals having 100 variants (50 D variants and 50 N variants). Four different methods (RWAS, OWAS, MB, and the single-marker test) were tested, and their power was estimated as the number of significant data sets among the 10,000 data sets using a significance threshold of 2.5×10^{-6}.

the constant PAR disease-risk model where all variants have the same PAR. This was the model used to estimate the power of MB, and MB was shown to be more powerful than other competing methods (MADSEN and BROWNING 2009). Throughout all experiments, we use the sum of genetic scores of case individuals as a statistic for MB, rather than using the sum of ranks of cases suggested by Madsen and Browning. One reason is that both sums yield similar results (MADSEN and BROWNING 2009), and another reason is that the sum of genetic scores allows RWAS and MB to be compared in the same sum of *z*-scores framework (see METHODS for approximation of MB to a sum of *z*-scores method). The power of RWAS is also compared to the power of OWAS that is the optimal weighted sum of *z*-scores and from which the weights of RWAS are derived. OWAS uses the effect sizes of variants for its weights, and hence the power of OWAS can be thought of as the upper bound of power that can be achieved in the weighted sum of *z*-scores approach. In this experiment, OWAS knows the group PAR that generated data sets (see below), computes relative risk of each variant using Equation 6, and estimates population MAF as described in File S1.

We use exactly the same simulation parameters as in Madsen and Browning to estimate the power of methods. In the simulations, a total of 10,000 data sets are generated, each with 1000 case and 1000 control individuals having 100 variants. The power of a method is estimated as the number of significant data sets among the 10,000 data sets using a significance threshold of 2.5×10^{-6} based on the Bonferroni correction assuming 20,000 genes genome-wide. Among 100 variants, 50 variants are disease-risk contributing variants (D variants) and 50 variants are disease-risk neutral variants (N variants). For each variant, we sample its MAF in controls using Wright's formula (WRIGHT 1931; EWENS 2004) with the same parameter values as in Madsen and Browning (see MADSEN and BROWNING 2009 for details). According to Equation 6, relative risk of D variants is calculated from MAF of variants in controls and the marginal PAR that is the group PAR divided by the number of D variants while relative risk of N variants is 1. MAF of variants in cases can then be calculated using relative risk and MAF of variants in controls according to Equation 3. We independently sample mutations of each variant in case and control individuals according to its MAF in cases and controls, respectively.

The results of power simulations demonstrate that RWAS consistently outperforms MB when the group PAR varies from 1% to 5% (Figure 1). For example, at the group PAR of 3%, RWAS has 78% power while MB has 40% power. The power simulations also show that the power of RWAS is very close to the power of OWAS. Although OWAS has higher power than RWAS at all group PAR levels, the difference in power between the two methods is small; the power of RWAS is ~2–4% smaller than that of OWAS. Therefore, the analytical approximation of the optimal weights in RWAS reduces its power by only a small amount in this disease model, and it can achieve high power even if it is not given the true effect sizes of variants.

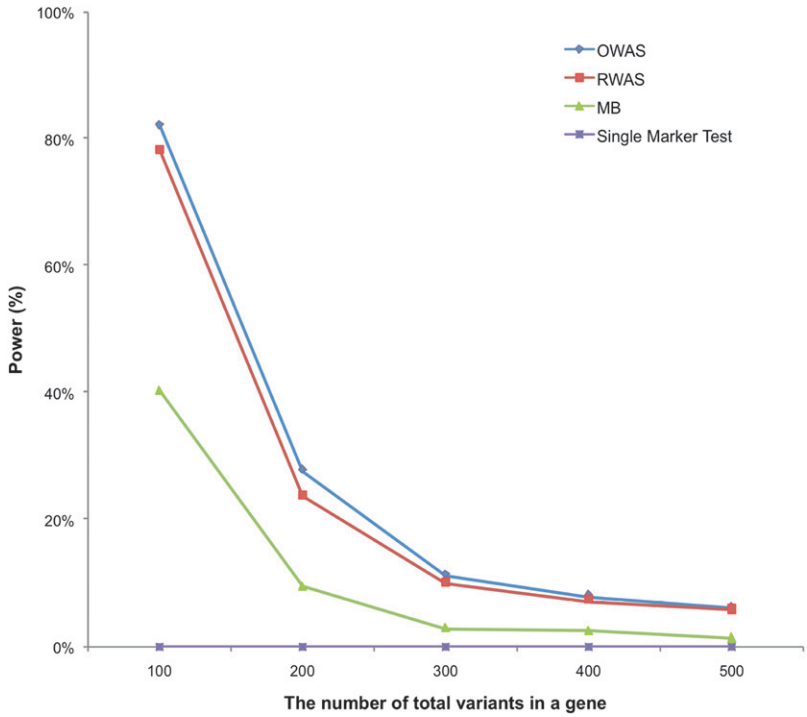

FIGURE 2.—Power comparison with the different numbers of total variants in a gene. We simulated five different numbers of total variants: 100, 200, 300, 400, and 500. All simulations had 50 causal variants, the group PAR of 3%, and 1000 case and 1000 control individuals. We created 10,000 data sets for each of five different simulations. The plot shows the power of RWAS, OWAS, MB, and the single-marker test.

Type I error rates of RWAS and OWAS: To check whether type I error rates (false positive rates) of RWAS and OWAS are correctly controlled, we create 100 million data sets without any causal variant. Each data set has 1000 cases and 1000 controls with 100 variants, and we measure type I error rates of RWAS and OWAS on the 100 million null data sets under three different significance thresholds: 0.05, 0.01, and 2.5×10^{-6}. The reason why we use a very large number of data sets is because the significance threshold for power is very low (2.5×10^{-6}). The proportion of significant data sets is an estimate of the type I error rate for each method.

The type I error rates for RWAS are 0.0503, 0.0089, and 1.2×10^{-7}, and those for OWAS are 0.0502, 0.0091, and 1.8×10^{-7} for the significance thresholds of 0.05, 0.01, and 2.5×10^{-6}, respectively. This indicates that the type I error rates are correctly controlled for RWAS and OWAS when the significance thresholds are 0.05 and 0.01. When the significance threshold is 2.5×10^{-6}, RWAS and OWAS both have lower type I error rates than the expected rate.

Power of RWAS with the different numbers of variants: Since the number of variants in a gene may be >100, we evaluate effects of the number of variants in a gene on the power of groupwise tests. We create five different data sets with five different numbers of total variants: 100, 200, 300, 400, and 500. In all five data sets, the number of causal variants is 50, and the group PAR is 3%. The number of case and control individuals is the same as in the previous experiment.

Figure 2 shows that as the number of total variants in a gene increases, the power of all methods decreases. For example, when a gene contains 100 variants, RWAS achieves 78% power while it has 6% power when there are 500 variants in a gene. This is because there are more noncausal variants in a gene as the number of variants increases. A large number of noncausal variants reduce our ability to detect causal variants and power of the groupwise tests.

Power of RWAS with prior information: Prior information can reduce or remove the influence of noncausal variants, and in this experiment we observe how prior information influences the power of RWAS. The prior information we consider is the probability of a variant being causal to a disease, denoted as c_i. We generate data sets with predefined true c_i values, and we evaluate how the power of RWAS changes when different prior information is given to RWAS. We first generate data sets that contain 100 variants split into two groups, each with 50 variants. We set c_i of the first group to 0.8 and c_i of the second group to 0.2. Then, five different types of prior information are given to RWAS: (1) "correct c_i" that is equivalent to true c_i of data sets, (2) "uniform incorrect c_i" in which $c_i = 1$ for all variants, (3) "three-fourths correct c_i" that corresponds to three-fourths of true c_i of the first and second groups, (4) "one-half correct c_i" that matches one-half of true c_i of the first and second groups, and (5) "very incorrect c_i" in which c_i of the first and second groups is 0.2 and 0.8, respectively, which is opposite to true c_i of data sets. The single-marker test and MB are also tested to compare their power to RWAS.

We follow the same experimental framework as in the previous experiment in this power simulation with two

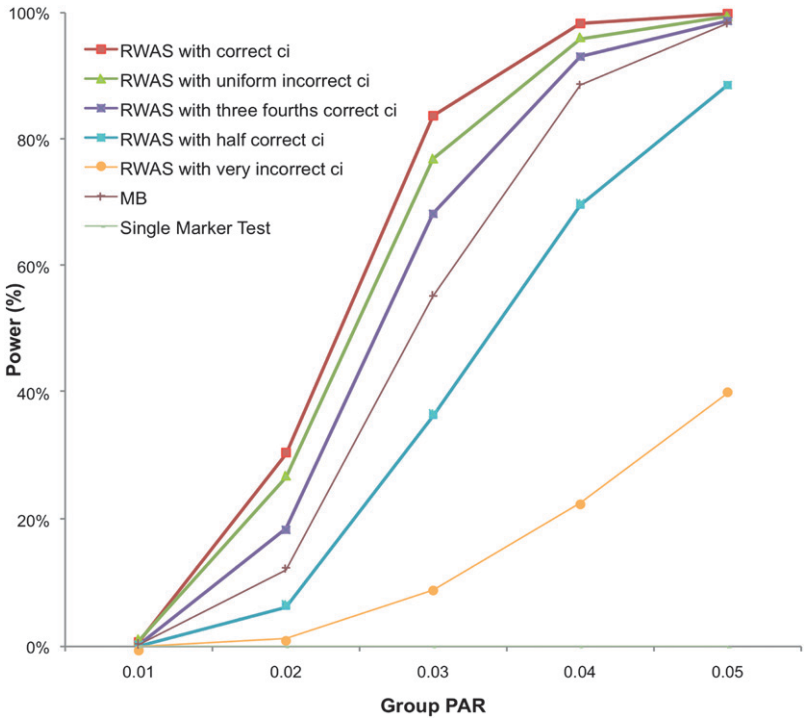

FIGURE 3.—Power of RWAS with different prior information. For each group PAR, 10,000 data sets were generated, and each data set contained 1000 case and 1000 control individuals having 100 variants with predefined true c_i values. c_i of 50 variants was 0.8, and c_i of the other 50 variants was 0.2. Five different types of prior information were given to RWAS: "correct c_i" (same c_i as true c_i of data sets), "uniform incorrect c_i" ($c_i = 1$ for all variants), "three-fourths correct c_i" (three-fourths of true c_i), "one-half correct c_i" (one-half of true c_i), and "very incorrect c_i" (opposite c_i to true c_i of data sets). The single-marker test, MB, and RWAS with the five different types of prior information were tested.

changes. The first change is that we have two different c_i values assigned to the two groups of variants as mentioned earlier. For each data set, a variant is causal with the probability proportional to its c_i. Relative risk of a causal variant is given by Equation 6 whereas a noncausal variant has relative risk of 1. The other change is that the same set of control MAFs is assigned to the two groups: MAFs of 50 variants in control individuals are sampled using Wright's formula and assigned to each group. The reason is that we want to observe only the effect of prior information on the power of studies, but the power is also dependent on MAF of variants.

Results show that the power of RWAS with correct c_i is always the highest among different prior information applied to RWAS (Figure 3). By knowing the correct prior information, the power increases as much as 7%; at the group PAR of 3%, the power of RWAS with correct c_i is 84% while the power of RWAS with uniform incorrect c_i is 77%. However, if RWAS is given incorrect prior information, it may suffer power loss as the power of RWAS with very incorrect c_i is >70% lower than the power of RWAS with correct c_i at the group PAR of 3% and 4%. This shows that when prior information is not very accurate, RWAS may achieve higher power by assuming that every variant is causal. The results also indicate that as RWAS is given more correct prior information, its power increases: the power of RWAS with three-fourths correct c_i is higher than the power of RWAS with one-half correct c_i. Results of the experiment demonstrate that prior information may considerably influence the power of studies and higher power can be achieved by knowing correct prior information.

RWAS with prior information on real mutation screening data: To evaluate RWAS and effects of prior information on real sequencing data, we use mutation screening data of the susceptibility gene for ataxia telangiectasia (TAVTIGIAN *et al.* 2009). This gene is called *ATM*, and it is also known as an intermediate-risk gene for breast cancer. TAVTIGIAN *et al.* (2009) collected data from seven *ATM* mutation screening studies in breast cancer cases and controls as well as data from their own mutation screening, which resulted in collecting 2531 case and 2245 control individuals (called "*bona fide* case–control studies"). They further increased the number of cases and controls by adding 17 case-only or control-only mutation screening studies, but we focus on the *bona fide* case–control studies in our experiment because adding the case-only and control-only studies does not yield substantial changes in results (TAVTIGIAN *et al.* 2009).

Tavtigian *et al.* discovered 170 rare missense variants in the *ATM* data set and used the missense analysis programs, Align-GVGD (TAVTIGIAN *et al.* 2006) and SIFT (NG and HENIKOFF 2003), to find how likely each variant is to be deleterious. Align-GVGD categorizes variants into seven grades: C0 (most likely neutral), C15, C25, C35, C45, C55, and C65 (most likely deleterious). Since the absolute deleteriousness of grades is not reported, we arbitrarily assign c_i of 0.05, 0.2, 0.35, 0.5, 0.65, 0.8, and 0.95 to the seven grades, respectively. SIFT yields scores for variants ranging from 1.00 (most likely neutral) to 0.00 (most likely deleterious) in steps of 0.01. There is a predefined threshold (0.05) in SIFT scores such that variants whose SIFT scores are ≤0.05

are considered deleterious while other variants are considered neutral. Hence, we assigned c_i of 1 to variants with SIFT scores ≤ 0.05 and c_i of 0 to other variants.

We first apply RWAS to the case–control studies without prior information, and RWAS yields a *P*-value of 0.3946. The *P*-value indicates no significant difference in mutation counts between cases and controls, and Tavtigian *et al.* also reported that they did not find a significant association by comparing frequency in cases *vs.* controls or by using CMC (Tavtigian *et al.* 2009). However, when RWAS is applied with prior information from Align-GVGD, it yields a *P*-value of 0.0078, which indicates a significant association between rare variants and the disease. The result is consistent with the results of Tavtigian *et al.* (2009); a significant *P*-value was obtained by performing a log-linear trend test with output of Align-GVGD. Therefore, this suggests that prior information may be useful in association studies and that RWAS can be applied to detect an association in real data.

Interestingly, RWAS reports a nonsignificant *P*-value of 0.0881 when using SIFT scores as prior information while Tavtigian *et al.* (2009) found a significant association with SIFT scores. It may be because the binary classification of variants according to SIFT scores is not as informative as output of Align-GVGD in predicting how likely each variant is to be causal. In other words, variants that are considered deleterious (SIFT scores ≤ 0.05) may be deleterious to different degrees, but SIFT scores do not capture this. The relative degree of a variant's deleteriousness is important in RWAS because more deleterious variants receive higher weights. Hence, this experiment shows that methods to determine prior information of variants play a key role in the real data analysis, and different prior information may yield different results.

DISCUSSION

In this article, we presented the RWAS to detect associations with a group of rare variants. We first developed the OWAS that maximizes the power of studies under the weighted sum of *z*-score statistics, but we need to know the effect sizes of variants to use OWAS. We then developed RWAS by analytically approximating the optimal weights, and it can be applied without the knowledge of effect sizes. The simulations demonstrate that RWAS outperforms a weighted sum statistic by Madsen and Browning (2009) in the same disease-risk model discussed in Madsen and Browning (2009). The simulations also show that the power of RWAS is very close to the power of OWAS, suggesting that RWAS achieves nearly optimal power in the disease-risk model we focused on.

We then extended RWAS to incorporate prior information of variants, and we considered the probability of a variant being causal to a disease as prior information in this article. To determine effects of prior information on association studies, we used both simulated data and real mutation screening data for the susceptibility gene for ataxia telangiectasia. The results of simulated data show that power can be increased by incorporating correct prior information, and this is confirmed in the real data since RWAS is able to detect an association in the real data with prior information while it is not able to do so without the information. Hence, this suggests that it would be advantageous to incorporate prior information into association studies and RWAS can be used to find associations in such association studies.

Many studies suggest that rare variants are not in linkage disequilibrium with each other (Pritchard 2001; Pritchard and Cox 2002; Li and Leal 2008). To compute the *P*-values, our statistic assumes that these variants are independent. However, in the case that the rare variants are linked, we can apply a permutation test to obtain *P*-values to apply the method.

J.H.S., D.H., B.H., and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, and 0916676 and National Institutes of Health grants K25-HL080079 and U01-DA024417. B.H. is supported by a Samsung Scholarship. This research was supported in part by the University of California, Los Angeles subcontract of contract N01-ES-45530 from the National Toxicology Program and National Institute of Environmental Health Sciences to Perlegen Sciences.

LITERATURE CITED

Altshuler, D., J. N. Hirschhorn, M. Klannemark, C. M. Lindgren, M. C. Vohl *et al.*, 2000 The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat. Genet. **26:** 76–80.

Bertina, R. M., B. P. C. Koeleman, T. Koster, F. R. Rosendaal, R. J. Dirven *et al.*, 1994 Mutation in blood coagulation factor V associated with resistance to activated protein C. Nature **369:** 64–67.

Blauw, H. M., J. H. Veldink, M. A. van Es, P. W. van Vught, C. G. J. Saris *et al.*, 2008 Copy-number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen. Lancet Neurol. **7:** 319–326.

Bodmer, W., and C. Bonilla, 2008 Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet. **40:** 695–701.

Cohen, J. C., R. S. Kiss, A. Pertsemlidis, Y. L. Marcel, R. McPherson *et al.*, 2004 Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science **305:** 869–872.

Corder, E. H., A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell *et al.*, 1993 Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science **261:** 921–923.

Eskin, E., 2008 Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. Genome Res. **18:** 653–660.

Ewens, W. J., 2004 *Mathematical Population Genetics*, Ed. 2. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Fearnhead, N. S., J. L. Wilding, B. Winney, S. Tonks, S. Bartlett *et al.*, 2004 Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. Proc. Natl. Acad. Sci. USA **101:** 15992–15997.

Gorlov, I. P., O. Y. Gorlova, S. R. Sunyaev, M. R. Spitz and C. I. Amos, 2008 Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am. J. Hum. Genet. **82:** 100–112.

Han, B., H. M. Kang, M. S. Seo, N. Zaitlen and E. Eskin, 2008 Efficient association study design via power-optimized tag SNP selection. Ann. Hum. Genet. **72:** 834–847.

International Schizophrenia Consortium, 2008 Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature **455:** 237–241.

Ji, W., J. N. Foo, B. J. O'Roak, H. Zhao, M. G. Larson *et al.*, 2008 Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat. Genet. **40:** 592–599.

Kryukov, G. V., L. A. Pennacchio and S. R. Sunyaev, 2007 Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am. J. Hum. Genet. **80:** 727–739.

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

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

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

Ng, P. C., and S. Henikoff, 2003 SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. **31:** 3812–3814.

Price, A. L., G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples *et al.*, 2010 Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. **86:** 832–838.

Pritchard, J. K., 2001 Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. **69:** 124–137.

Pritchard, J. K., and N. J. Cox, 2002 The allelic architecture of human disease genes: common disease-common variant. . .or not? Hum. Mol. Genet. **11:** 2417–2423.

Romeo, S., L. A. Pennacchio, Y. Fu, E. Boerwinkle, A. Tybjaerg-Hansen *et al.*, 2007 Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat. Genet. **39:** 513–516.

Tavtigian, S. V., A. M. Deffenbaugh, L. Yin, T. Judkins, T. Scholl *et al.*, 2006 Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J. Med. Genet. **43:** 295–305.

Tavtigian, S. V., P. J. Oefner, D. Babikyan, A. Hartmann, S. Healey *et al.*, 2009 Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. Am. J. Hum. Genet. **85:** 427–446.

Walsh, T., J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce *et al.*, 2008 Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science **320:** 539–543.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Xu, B., J. L. Roos, S. Levy, E. J. Van Rensburg, J. A. Gogos *et al.*, 2008 Strong association of de novo copy number mutations with sporadic schizophrenia. Nat. Genet. **40:** 880–885.

Zaitlen, N., B. Paaniuc, T. Gur, E. Ziv and E. Halperin, 2010 Leveraging genetic variability across populations for the identification of causal variants. Am. J. Hum. Genet. **86:** 23–33.

Communicating editor: F. Zou

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.125070/DC1

An Optimal Weighted Aggregated Association Test for Identification of Rare Variants Involved in Common

Jae Hoon Sul, Buhm Han, Dan He and Eleazar Eskin

DOI: 10.1534/genetics.110.125070

FILE S1

MATERIALS AND METHODS

Estimation of population MAF in OWAS

There can be several ways to estimate population MAF. For example, MB estimates it from control individuals (MADSEN and BROWNING 2009). We choose to estimate population MAF in the following way. Denoting p_i as population MAF of variant i, we first assume that its true overall sample frequency is equal to observed overall sample frequency.

$$p_i^+ + p_i^- = \hat{p}_i^+ + \hat{p}_i^- \tag{17}$$

p_i^+ and p_i^- are defined in terms of p_i (Equations 3, 4), and we can rewrite (Equation 17) as

$$\frac{\gamma_i p_i}{(\gamma_i - 1)p_i + 1} + p_i = 2\hat{p}_i^{\pm} \quad \left(\text{where } \hat{p}_i^{\pm} = \frac{\hat{p}_i^+ + \hat{p}_i^-}{2} \right) \tag{18}$$

We can compute p_i in terms of γ_i and $\hat{p}_i^{\pm}$ by finding the root of (Equation 18).

$$p_i = \frac{b + \sqrt{b^2 + 8(\gamma_i - 1)\hat{p}_i^{\pm}}}{2(\gamma_i - 1)} \quad \text{where } b = 2\hat{p}_i^{\pm}(\gamma_i - 1) - (\gamma_i + 1) \tag{19}$$

Approximation of x_i, $\hat{\mu}_i$ and $\hat{\sigma}_i$ of MB

In this section, we show that x_i, $\hat{\mu}_i$ and $\hat{\sigma}_i$ of MB can be approximated as (Equation 12). First, MB calculates a weight of variant i ($\hat{w}_i$) as

$$\hat{w}_i = \sqrt{N \cdot q_i(1 - q_i)} \quad \text{where } q_i = \frac{m_i^U + 1}{2n_i^U + 2} \tag{20}$$

N is the total number of case and control individuals, m_i^U is the number of mutations for variant i in control individuals, and n_i^U is the number of control individuals.

MB then calculates the genetic score (γ_j) of each individual j.

$$\gamma_j = \sum_{i=1}^{M} \frac{I_{ij}}{\hat{w}_i}$$

where M is the number of variants, and I_{ij} is the number of mutations observed in individual j at variant i. MB ranks all individuals (both cases and controls) by their genetic scores and calculates the sum of the ranks of cases as its test statistic (x).

$$x = \sum_{j \in \text{cases}} \text{rank}(\gamma_j)$$

Madsen and Browning reports that x can also be computed using the sum of genetic scores instead of the sum of ranks, and the two methods have very similar power. Hence, we will compute x as the sum of genetic scores.

$$x = \sum_{j \in \text{cases}} \gamma_j$$

First, we observe that the sum of genetic scores of cases is equivalent to the sum of observed MAF of each variant in cases divided by the weight of the variant. In other words, we sum the number of mutations per variant instead of the number of mutations per individual.

$$\sum_{j \in \text{cases}} \gamma_j = \sum_{i=1}^{M} \frac{\mathcal{N}/2 \cdot \hat{p}_i^+}{\sqrt{\mathcal{N} q_i (1-q_i)}} \tag{21}$$

Assuming $q_i \approx \hat{p}_i^-$ since q_i is an estimate of MAF of variant i in controls, the statistic of variant i, x_i, in (Equation 21) is

$$x_i = \frac{\sqrt{\mathcal{N}}}{2} \frac{\hat{p}_i^+}{\sqrt{\hat{p}_i^- (1-\hat{p}_i^-)}} \tag{22}$$

Next, we derive the statistic of the null distribution denoted as x_i^*. First $\hat{p}_i^+$ and $\hat{p}_i^-$ have the following distribution under the null distribution.

$$\hat{p}_i^+ \sim \mathsf{N}\left(p_i, \frac{p_i(1-p_i)}{\mathcal{N}/2} \right) \tag{23}$$

$$\hat{p}_i^- \sim \mathsf{N}\left(p_i, \frac{p_i(1-p_i)}{\mathcal{N}/2} \right) \tag{24}$$

By multiplying $\hat{p}_i^+$ in (Equation 23) by $\frac{\sqrt{\mathcal{N}}}{2\sqrt{\hat{p}_i^- (1-\hat{p}_i^-)}}$ and assuming $\hat{p}_i^- \approx p_i$, we can derive x_i^* that is approximately equivalent to x_i in (Equation 22). x_i^* and its distribution are then

$$x_i^* = \frac{\sqrt{\mathcal{N}}}{2} \frac{\hat{p}_i^+}{\sqrt{\hat{p}_i^- (1-\hat{p}_i^-)}} \approx \frac{\sqrt{\mathcal{N}}}{2} \frac{\hat{p}_i^+}{\sqrt{p_i(1-p_i)}} \sim \mathsf{N}\left(\frac{\sqrt{\mathcal{N}}}{2} \frac{p_i}{p_i(1-p_i)}, \frac{1}{2} \right) \tag{25}$$

Thus, the mean ($\hat{\mu}_i$) of x_i^* is $\frac{\sqrt{\mathcal{N}}}{2} \frac{p_i}{p_i(1-p_i)}$, the standard deviation ($\hat{\sigma}_i$) is $\sqrt{1/2}$, and x_i is (Equation 22).

LITERATURE CITED

MADSEN, B. E. , and S. R. BROWNING, 2009 A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384.