Published in final edited form as:

Ann Hum Genet. 2012 January ; 76(1): 74–85. doi:10.1111/j.1469-1809.2011.00684.x.

Testing Rare Variants for Association with Diseases: a Bayesian Marker Selection Approach

Lei Zhang¹, Yu-Fang Pei^{1,*}, Rong Hai^{1,2}, and Hong-Wen Deng^{1,*,3,4}

¹Center of System Biomedical Sciences, University of Shanghai for Science and Technology, Shanghai 200093, P. R. China

²The Affiliated hospital of Inner Mongolia medical College Hohhot, Inner Mongolia, 010050, P. R. China

³Department of Biostatistics, Tulane University, New Orleans, LA 70112, USA

⁴Laboratory of Molecular and Statistical Genetics and Key Laboratory of Protein Chemistry and Developmental Biology of Ministry of Education, College of Life Sciences, Hunan Normal University, 410081 Changsha, Hunan, People's Republic of China

Abstract

It has been a research focus to uncover the genetic determination of complex diseases caused by rare variants. As the vast majority of genomic variants represent background variation, highlighting potentially causal mutations through weighting scheme is critical to the success of rare variants aimed association studies. In this study, we propose a novel Bayesian marker selection approach to perform weighting-based association test. In this approach, individual association signal and its direction are used to weight variants. In addition, the predicted biological function of variants is taken as prior information to direct the selection of likely causal variants. Simulation studies show that the proposed method has improved power over several existing methods in certain conditions. Analyses of two empirical datasets demonstrate its applicability.

Keywords

weighting; Bayesian marker selection; rare variants; association

Introduction

Though common variant aimed genome-wide association studies (GWASes) have identified hundreds of single nucleotide polymorphisms (SNPs) associated with a variety of complex diseases and traits (<http://www.genome.gov/gwastudies>), it is still an open question that the identified variants explain only a small portion of phenotypic variation (Manolio *et al.*, 2009). Until recently, it has been recognized that the “missing” heritability may be accounted for by variants with lower frequencies but with a much larger number across the human genome. A variety of theoretical and experimental studies have shown that multiple rare variants are as important as common variants in determining the etiology of complex diseases and the variation of traits (Pritchard, 2001, Pritchard & Cox, 2002, Cohen *et al.*, 2004, Cohen *et al.*, 2006, Ahituv *et al.*, 2007a, Kryukov *et al.*, 2007, Ji *et al.*, 2008, Tarpey *et al.*, 2009), and extensive evidence continues to emerge. Thus, it has been a research focus

*Corresponding to: Yu-fang Pei, Ph. D. Tel: 86-21-65710369 pyf0419@gmail.com Fax: 86-21-65710108 Hong-Wen Deng, Ph.D. Tel: 504-988-5164 hdeng2@tulane.edu Fax: 504-988-1706.

to uncover the genetic determination caused by rare variants for most complex diseases. To depict an exhaustively unbiased map of rare variants across the genome, the 1000 genomes project (Siva, 2008) has been launched with the aim of sequencing at least 1,000 human genomes in various populations. The densest map of rare variants that the project identifies, together with advances in next-generation sequencing technologies (Tucker *et al.*, 2009), will promote the next-wave rare variants aimed association studies by sequencing a portion or even the entire of the genome in a large number of individuals.

Due to allelic heterogeneity (Pritchard & Cox, 2002) and rarity of variants, conventional one-by-one statistical analyses are inefficient for rare variants. As solutions, a variety of specific methods have been proposed (Madsen & Browning, 2009, Bansal *et al.*, 2010, Bhatia *et al.*, 2010, Han & Pan, 2010, Li *et al.*, 2010, Liu & Leal, 2010, Price *et al.*, 2010, Zhang *et al.*, 2010b, Zhang *et al.*, 2010a, Zhu *et al.*, 2010, Ionita-Laza *et al.*, 2011, Sul *et al.*, 2011, Yi & Zhi, 2011). One of these proposed analyzing strategies is to take the genomic region that contains multiple rare variants as the unit of analyses (Li & Leal, 2008). This is achieved by merging the information contained in multiple variants together and examining the association by the merged signal. Ideally, rare variants to be merged should be homogeneous in the sense that all are irrelevant to or associated with the phenotype, and in the latter case, in the same direction. Unfortunately, the genome is well known to be “noisy” by the fact that the vast majority of variants represent background variation rather than having any phenotypic effect even in causal genes. Furthermore, though the majority of newly arising mutations are believed to be deleterious, some others still have chances to be beneficial (Nejentsev *et al.*, 2009). Thus, merging all rare variants equally regardless of their relative phenotypic impacts may produce an unsatisfactory “signal-to-noise” ratio.

To distinguish those likely causative variants from background variation, a variety of weighting schemes have been proposed in literature (Madsen & Browning, 2009, Han & Pan, 2010, Liu & Leal, 2010, Price *et al.*, 2010, Zhang *et al.*, 2010b, Zhang *et al.*, 2010a, Ionita-Laza *et al.*, 2011, Sul *et al.*, 2011, Yi & Zhi, 2011). Therein, Madsen and Browning (2009) weight each variant according to its population frequency so that rarer variants obtain a heavier weight. The hypothesis underlying this weighting is that rarer variants render a larger per-allele phenotypic effect. Price *et al.* (2010) extend this work by weighting only variants with MAFs below a specified threshold, but they examine all MAF thresholds encountered in the sample. Despite their potential merit in enlarging mutation signals, MAF-based weighting schemes may be limited by the fact that the relationship between the frequency of an allele and its causative status is not well established. For instance, it is unlikely to judge an allele as a causative one if it is even rarer in normal population than in affected individuals. A more plausible source of weighting scheme may thus base weight on individual association signal as well as its direction. However, a technical problem for this approach is that in conventional frequentist manner, association signals are difficult to measure precisely for rare variants. The problem could be circumvented in a Bayesian view of point. In the Bayesian framework, the posterior distribution of association could be identifiable as long as an appropriate prior distribution is given (Yi & Zhi, 2011).

Besides the data itself, another source of information that could be used to select potentially causal variants is the prior knowledge on each variant. Unlike neutral genetic markers, each variant in gene coding regions has biological impact on protein structure and function. The functional impact could be studied and predicted *a priori* by a variety of bioinformatic algorithms (Adzhubei *et al.*, 2010, Ng & Henikoff, 2003, Ng & Henikoff, 2006, Mooney & Altman, 2003, Reumers *et al.*, 2006). Incorporating prior knowledge acquired by these algorithms into weighting on likely causal variants has the potential to improve the performance of association tests (Price *et al.*, 2010). However, the improvement is heavily dependent on the accuracy of prediction algorithms (Zhang *et al.*, 2010b). Thus, it is

preferable to take the predicted functional impact as prior distribution rather than the sole source to weight variants.

In this study, in order to improve the power of rare variants aimed association test, we propose a novel Bayesian method that integrates these two sources of information. Specifically, they are modeled into a single Bayesian marker selection framework to instruct the selection of likely causal variants. On the one hand, each rare variant is weighted according to its posterior association distribution. On the other hand, the predicted functional impact is modeled as prior distribution to facilitate highlighting likely causal variants. We will demonstrate the utility of our method by extensive simulation studies as well as by analyzing two real datasets.

Methods

Definition

Throughout this study, we assume that rare variants are independent and present no interaction effect. Suppose a number of N individuals, including N_d cases and N_u controls, are sequenced at a genomic region of interest, and a total of L rare variants (MAF<5%) are discovered. Let d_l be the total number of rare alleles carried by cases at the l th variant, and let $\mathbf{D} = (d_1, \dots, d_L)$. Denote $\mathbf{U} = (u_1, \dots, u_L)$ for controls in the same manner. In addition, let $\mathbf{R} = (r_1, \dots, r_L)$ represent the number of rare alleles carried by both cases and controls. The L variants are partitioned into three groups: group 0 contains variants that are irrelevant to the disease; group 1 contains variants at which rare alleles increase the risk of the disease, and group 2 contains variants at which rare alleles decrease the risk of the disease. Let $I_l = 0, 1, 2$ indicate the membership of the l th variants, and let $\mathbf{I} = (I_1, \dots, I_L)$. The genotypes of cases in the three groups are denoted by $\mathbf{D}_0$, $\mathbf{D}_1$, and $\mathbf{D}_2$, and those of controls are denoted by $\mathbf{U}_0$, $\mathbf{U}_1$, and $\mathbf{U}_2$, respectively.

Model

We extend a Bayesian marker partition model (Zhang & Liu, 2007, Tang *et al.*, 2009) to study the membership of variants. Variants in group 0 are irrelevant to the disease, so that rare allele at each variant has the same distribution between cases and controls. Suppose rare alleles follows a binomial distribution, and let $\boldsymbol{\theta}_0 = (\theta_l, I_l=0)$ be the vector of rare allele frequency for variants in group 0. The likelihood of genotype data of the sample in group 0 is written by

$$P(\mathbf{D}_0, \mathbf{U}_0 | \boldsymbol{\theta}_0) = \prod_{l=1}^L C_N^{r_l} \theta_l^{r_l} (1 - \theta_l)^{N-r_l},$$

where $C_N^{r_l}$ is the binomial coefficient. Assuming that each θ_l has a prior $\text{beta}(\alpha_l, \beta_l)$ distribution, we integrate out $\boldsymbol{\theta}_0$ and obtain the marginal probability of genotypes in group 0,

$$P(\mathbf{D}_0, \mathbf{U}_0 | \mathbf{I}) = \prod_{l=1}^L C_N^{r_l} \frac{B(\alpha_l + r_l, \beta_l + N - r_l)}{B(\alpha_l, \beta_l)}, \quad (1)$$

where $B(\cdot)$ denotes the Beta function.

Group 1 contains variants at which rare alleles increase the risk of disease, so that rare allele at each variant has a larger frequency in cases than in controls. Let $\boldsymbol{\theta}_1 = (\theta_l, I_l=1)$ be rare allele frequencies for cases in group 1. To restrict each allele frequency in controls to be

lower than that in cases, we introduce a vector of factor $\xi_1 = \{\xi_l \in (0, 1), l=1\}$ so that allele frequency at each variant in controls $p_l = \theta_l \times \xi_l$. The factor ξ_l reflects the magnitude of discrepancy in allele frequency between cases and controls. In fact, $1/\xi_l$ approximates the odd ratio for very rare allele. Given θ_1 and ξ_1 , the likelihood of genotypes in group 1 is written by

$$P(D_1, U_1 | \theta_1 \xi_1) = \prod_{l=1} \left[C_{N_d}^{d_l} \theta_l^{d_l} (1 - \theta_l)^{N_d - d_l} \cdot C_{N_u}^{u_l} (\theta_l \xi_l)^{u_l} (1 - \theta_l \xi_l)^{N_u - u_l} \right].$$

Assume a prior $\text{beta}(\alpha_l, \beta_l)$ distribution for each θ_l , and a prior $\text{beta}(\gamma_{l1}, \gamma_{l2})$ distribution for each ξ_l ; we integrate out θ_1 and ξ_1 and obtain the marginal probability of genotypes in group

$$P(D_1, U_1 | \mathbf{I}) = \prod_{l=1} \left\{ C_{N_d}^{d_l} C_{N_u}^{u_l} \left[\sum_{i=0}^{N_u - u_l} C_{N_u - u_l}^i \frac{B(\alpha_l + N_u + d_l - i, \beta_l + N_d - d_l + i)}{B(\alpha_l, \beta_l)} \cdot \frac{B(\gamma_{l1} + u_l, \gamma_{l2} + N_u - u_l - i)}{B(\gamma_{l1}, \gamma_{l2})} \right] \right\}. \quad (2)$$

Group 2 contains variants at which rare alleles decrease the risk of disease, so that rare allele at each variant has a larger frequency in controls than in cases. In a similar manner to group 1, let $\theta_2 = (\theta_l, l=2)$ be rare allele frequencies of controls in group 2, and let $\xi_2 = \{\xi_l \in (0, 1), l=2\}$ so that the allele frequency in cases equals $\theta_l \times \xi_l$. Assume a prior $\text{beta}(\alpha_l, \beta_l)$ distribution for each θ_l , and a prior $\text{beta}(\gamma_{l1}, \gamma_{l2})$ distribution for each ξ_l ; we integrate out θ_2 and ξ_2 and obtain the marginal probability of the genotypes in group 2,

$$P(D_2, U_2 | \mathbf{I}) = \prod_{l=2} \left\{ C_{N_u}^{u_l} C_{N_d}^{d_l} \left[\sum_{i=0}^{N_d - d_l} C_{N_d - d_l}^i \frac{B(\alpha_l + N_d + u_l - i, \beta_l + N_u - u_l + i)}{B(\alpha_l, \beta_l)} \cdot \frac{B(\gamma_{l1} + d_l, \gamma_{l2} + N_d - d_l - i)}{B(\gamma_{l1}, \gamma_{l2})} \right] \right\}. \quad (3)$$

Letting the prior distribution of $\mathbf{I}$ be $P(\mathbf{I})$, we combine formulas (1) – (3) and obtain the posterior distribution of $\mathbf{I}$,

$$P(\mathbf{I} | \mathbf{D}, \mathbf{U}) \propto P(\mathbf{I}) P(\mathbf{D}_0, \mathbf{U}_0 | \mathbf{I}) P(\mathbf{D}_1, \mathbf{U}_1 | \mathbf{I}) P(\mathbf{D}_2, \mathbf{U}_2 | \mathbf{I}).$$

In the above formula, $P(\mathbf{I})$ quantifies our prior knowledge of each variants belonging to which group. Denote p_{l0} , p_{l1} , and p_{l2} the prior probabilities for the l th variant belonging to each of the three groups, then $P(\mathbf{I})$ has the following form,

$$P(\mathbf{I}) = \prod_{l=1}^L p_{l0}^{I_{l=0}} p_{l1}^{I_{l=1}} p_{l2}^{I_{l=2}},$$

where $I_l=k$ is an indicator variable which equals 1 if the statement holds and 0 otherwise.

Testing Association

The posterior distribution $P(\mathbf{I}|\mathbf{D},\mathbf{U})$ contains all information provided by the combination of prior knowledge and data together. The three probabilities corresponding to each variant facilitate statistical inference for various purposes. Here we focus on variants in group 1 and test their association with the disease in a frequentist manner. We judge a variant to belong to group 1 if its posterior probability in group 1 is larger than that in groups 0 and 2. For each individual j , define a genetic score

$$g_j = \sum_{l=1} g_{jl},$$

where $g_{jl}=0, 1$, or 2 encodes the number of rare alleles carried by the individual at the l th variant. All genetic scores over cases and controls are ranked ascendingly and the sum of ranks of cases is obtained, termed r_D . A rank-sum test (Lehmann, 1999) could be performed but the result will be biased due to the Bayesian marker selection procedure. Therefore, we instead take r_D as the statistic and perform a series of Monte Carlo simulations to assess its significance. Specifically, we perform a number of $k=10,000$ Monte Carlo permutations each by permuting case/control status. In each permuted dataset, we re-perform the Bayesian variable selection procedure and obtain a value of r_D . An exact p-value is obtained by dividing the number of replicates in which the value of r_D is larger than that of the unpermuted sample by the total number of replicates. We term the test Bayesian Marker Selection Test (BMST).

Model Parameters

The parameters of Bayesian marker selection model include the prior probability $P(\mathbf{I})$; the hyper-parameters α_l and β_l for allele frequency distribution; the hyper-parameters γ_{l1}, γ_{l2} for the distribution of discrepancy of allele frequency.

The prior probability is a set of probability vectors quantifying our *a priori* belief of each variant being causal. Incorporating predicted functional impact by specialized algorithms has the ability to improve the power of association test, though it is largely dependent on the accuracy of prediction algorithms. When prior knowledge is not available, we set $p_{l0}=p_{l1}=p_{l2} = 1/3$.

The distribution of rare allele frequency is estimated by analyzing the real sequence data produced by 1000 genomes pilot project (Durbin *et al.*, 2010). We focus on the exon dataset, which has the deepest coverage and most number of sequenced individuals, of the European (CEU+TSI) population. Maximal likelihood estimation on all available data gives the estimate of Beta distribution $(\alpha, \beta) = (0.309, 2.024)$ (see the Appendix for the detail). Inference with only rare variants (MAF<5%) shifts the distribution to rare variants further with $(\alpha, \beta) = (0.328, 2.496)$. This latter estimate will be used as the prior distribution for all rare variants.

The discrepancy of allele frequency between cases and controls reflects variant effect on the phenotype, and its distribution is quite subjective. We set the parameters $(\gamma_{l1}, \gamma_{l2}) = (0.5, 0.5)$ for all variants. Additional analyses on several other parameters give similar results.

Computing Issue

Since rare variants are assumed to be independent, the high-dimensional posterior distribution of $\mathbf{I}$ can be transformed into the product of a series of posterior distributions of individual variants. For each individual variant, the posterior probability vector can be exactly computed. Therefore, no computation-demanding sampling procedure will be required. The most expensive computation is the calculation of a series of beta functions in equations (1)-(3). For the sake of speeding, the intermediate results are stored for later access.

Comparison with Existing Methods

For comparison, we include several existing methods into analyses. These methods include: collapsing (Li & Leal, 2008); Weighted-Sum (WS) (Madsen & Browning, 2009); Variable Threshold weighted sum (VT) (Price *et al.*, 2010); Rare Cover (RC) (Bhatia *et al.*, 2010); Kernel-Based Adaptive Cluster (KBAC) (Liu & Leal, 2010); and Replication-based Weighting (RW) (Ionita-Laza *et al.*, 2011). For WS and VT, we invoke the R package *rareVariantTests.R* provided in (Price *et al.*, 2010) and retrieve the p-values *score3p* and *score4p*, respectively. For KBAC, we invoke the program *kbac* provided in (Liu & Leal, 2010). For RW, we invoke the program *test_rare* provided in (Ionita-Laza *et al.*, 2011). The other tests are implemented by our in-house software. All parameters are set to default values suggested by respective authors. Specifically, for RC, the convergence cut-off is set to 0.5, and the window size to 5kb. Notice that default settings may not represent the optimal performance of these methods.

Simulation Studies

To systematically investigate the performance of the proposed method under a variety of conditions, we conducted a series of simulation studies. We focused on the European population and simulated a gene coding region by a demographic model (Williamson *et al.*, 2005), which can account for the change in size of the population. This complex model assumes the European population experienced a bottleneck, followed by a long duration of recovery, then by a rapid expansion until to the modern population. It has a total of seven parameters: the effective size of ancestral (7947), bottleneck (262), recovery (7019), and current populations (52907); the duration (in unit of generation) of bottleneck (84), recovery (4557), and recent expansion (576), respectively. The values of these parameters were estimated in (Boyko *et al.*, 2008). In addition, both mutation and recombination rates were set to 1.0×10^{-8} per site. To account for the effect of natural selection, we added a distribution of selection coefficient to newly arising mutations, where the distribution was taken from a “neutral+gamma” form. The parameters of the distribution were estimated in (Boyko *et al.*, 2008). Specifically, a proportion of 0.239 of newly arising mutations were modeled as neutral, while the remainder were suffering negative selection with selection coefficients drawn from the gamma distribution $\text{Gamma}(1.02, 800)$. The gene was simulated with two lengths: 10 kb and 20 kb. The simulation was performed by an extensive forward simulation program *sfs_code* (Hernandez, 2008). The command implementing the simulation was `sfs_code 1 1 -t 0.00031788 -r 0.00031788 -B 10 -L 1 k --popSize 7947 -Td 0 0.032968 -Td 0.005285 26.79 -Td 0.291997 7.53768 -TE 0.328237 --sampSize 10000 --outfile out.txt -W 2 0.239 0 1 1.02 800 --popFreq freq.txt`, where *k* was replaced by either gene length in unit of base pair.

The output from *sfs_code* contained all derived mutations and their frequencies, which were taken as the template to simulate genotypes. We simulated a pool of 10,000 quantitatively phenotyped individuals and selected cases/controls from two tails of phenotypic distribution. Genotypes for each individual were simulated according to allele frequencies with the assumption of Hardy-Weinberg equilibrium. Common variants ($\text{MAF} \geq 5\%$) and mono-sites,

which present no variation in simulated sample, were removed. On simulating phenotype, a set C of causative variants was randomly selected with a specified proportion. For individual j , define a genetic risk score as

$$r_j = \sum_{l \in C} g_{jl}.$$

The quantitative phenotype was generated by a linear regression equation

$$y_j = \mu + \beta r_j + e_j,$$

where μ was the grand phenotypic mean; e_j was the random error which followed a normal distribution with zero mean. The coefficient β was determined by gene heritability h^2 , which was defined as

$$h^2 = \frac{\beta^2 \text{var}(r)}{\beta^2 \text{var}(r) + \text{var}(e)},$$

where $\text{var}(r)$ was the variance of risk score across the sample, and $\text{var}(e)$ was the variance of random error.

After the phenotype was generated, a specified number of individuals with highest phenotypes were selected as cases and an equal number of individuals with lowest phenotypes were selected as controls. As a baseline, we set the proportion of causative variants to be equal to 30%, the gene heritability to 0.5%, and the number of cases/controls to 1,000/1,000.

To evaluate the effect of predicted functional impact of variants on power estimation, we simulated functional impacts given by prediction algorithms. The accuracy of a particular prediction algorithm is measured by its sensitivity and specificity. Here the sensitivity is defined as the probability that a causative variant is correctly predicted, and the specificity as the probability that a non-causative variant is incorrectly predicted as a causative one. Assumed that the sensitivity and the specificity were α and β . We used a probability p of being causative as the continuous measure given by prediction algorithm. Let O be an indicator of predicted causative variant, and let R be an indicator of causative variant. Specifically, we had $P(O=1 | R=1) = \alpha$, and $P(O=1 | R=0) = \beta$. Given $R=1$ or 0 , O was drawn from a Bernoulli distribution $\text{Ber}(\alpha)$ or $\text{Ber}(\beta)$. We then simulated p from the Uniform distribution $U(t, 1)$ or $U(0, t)$ given $O=1$ or 0 , where t is a threshold to declare causative variant by prediction algorithm. In this simulation we set $t=0.5$. Given p , the probability that a variant is either neutral or protective was set to $(1-p)/2$.

One explicit assumption in our Bayesian model is the independence among variants. To test its robustness to the violation of this assumption, we also simulated variants with linkage disequilibrium (LD) structures. Specifically, we added the following step before generating the phenotype. A specified proportion of 30% of variants were randomly selected, and all subjects' genotypes at each variant were permuted so that the rare allele was in complete LD ($D'=1$) with the rare allele at its nearest neighboring variant. Subsequent simulations were the same as above.

The simulation studies were conducted at a variety of parameter settings. At each setting, 10,000 replicates were performed to estimate power rate. Both type-I error and power were

estimated at two significance levels: 0.01 and 0.001. Model parameters for the proposed test were set to the following: the prior distribution of rare allele was Beta (0.328, 2.496). The prior distribution of discrepancy between cases and controls was Beta (0.5, 0.5). The prior weights of alleles being risk, neutral, and protective were (0.33, 0.33, 0.33) in the absence of prior knowledge.

Applications

We re-analyzed two publicly available studies published by Cohen et al. (2004, 2006). In the first study, in order to examine the association of rare variants with HDL-C level in three candidate genes (*ABCA1*, *APOA1*, and *LCAT*), Cohen et al. (2004) sequenced the three genes in one group consisting of 128 individuals with low HDL-C levels and another group of 128 individuals with high HDL-C levels. To validate their findings, they sequenced the same three genes in a second sample consisting of 155 individuals with low HDL-C levels and 108 individuals with high HDL-C levels. These two samples are combined and analyzed together. As the authors did, we combine the three genes together and analyze rare variants that are unique to either high or low group. The resultant data include 35 variants that are exclusive in low group and 5 variants that are exclusive in high group.

In the second study, Cohen et al. (2006) used the same strategy and sequenced the gene *NPC1L1* in one group consisting of 256 individuals with low Ca:L ratios and another group of 256 individuals with high Ca:L levels, in order to examine the association of identified rare variants with LDL-C level. Again, we analyze rare variants unique to either high or low group, as the authors did. A total of 25 rare variants were analyzed, among which 20 are exclusive in low group and the other 5 are exclusive in high group.

Results

Type-I Error

Table 1 lists type-I error rates for various tests. It is obvious that all tests have correct error rates that are close to target levels, indicating their robustness. Though each test except Collapsing uses a particular weighting scheme in favor of enlarging association, they correct the selection bias through permutations, thus resulting in an unbiased test. All tests are still valid in the presence of LD among variants.

Power Estimate

Figure 1 depicts power estimate for 10 kb gene length in the absence (upper panel) or presence of LD (lower panel). As expected, all tests have improved power with increased gene heritability. Among the tests, the proposed BMST has the highest power when the heritability is below 0.6; otherwise RW is more powerful. Their powers are followed by KBAC and VT, where KBAC is more powerful than VT in most cases. They are then followed by WS and RC, which generally have approximately equal power. At last, Collapsing has the lowest power in most cases. For BMST, the heritability at which the gene is detected with 80% power is approximately 0.5%, while that for RW, KBAC, VT, RC, WS are 0.6, 0.6, 0.7, 0.9, and 1.0%, respectively. Collapsing can only achieve power around 70% even at a heritability as high as 1.0%. In the presence of LD, the trends of power for various tests are similar to that in the absence of it.

Figure 2 displays power estimate for 20 kb gene length. Again, all tests have improved power with increased heritability. Among the tests, BMST has the uniform highest power in all settings, followed by RW, then by KBAC and VT, whose powers are approximately equal. Among the other three tests whose powers are less than the above ones, WS is superior to both RC and Collapsing, where the latter two tests perform generally equally.

Again, power estimates in the presence of LD do not change significantly (lower panel). Therefore, we report results only in the absence of LD in what follows.

Comparison of power at 10 and 20 kb gene lengths (Figure 3) shows that power is decreased with longer gene for most tests; however different tests have different sensitivities. RC is most sensitive to gene length by most power loss, followed by KBAC, Collapsing, RW, and VT. For BMST and WS, power is nearly unchanged under the two lengths.

When predicted functional impacts of individual variants are informative, the usage of such prior knowledge could improve performance of tests, as displayed in Figure 4, in which the sensitivity and specificity of prediction are set to 0.6 and 0.2. All the four tests that incorporate prior information into weighting variants have improved power. The magnitude of power improvement for BMST is around 3%, slightly lower than that for the other tests, which are 6% for VT, 5% for WS, and 6% for RW.

We also investigate the performance of tests when functional prediction is non-informative. To simulate this scenario, we generate for each variant a continuous prediction measure from the Uniform distribution $U(0, 1)$, independent of its causative status. As listed in table 2, all the four tests have power lost, where the magnitude of the loss is approximately 8% for BMST, 10% for VT, 16% for WS, and 12% for RW.

Computation Time

Computation time was measured by analyzing a simulated sample with 1,000/1,000 individuals and 10 kb gene length. The number of permutations was set to 10,000. Running on a single Intel 2.8GHz CPU, the computation time was 69s for BMST, 41s for RC, 13s for KBAC, 5s for RW, 2s for VT, 1s for WS, and <1s for Collapsing. Scaling the configuration to a computer cluster with 8 or 16 CPUs on each computing node, BMST could complete a genome-wide scan (~20,000 genes) within days, thus making it applicable to large-scale association scans.

Applications

As applications, we re-analyzed the two studies described in the above section and examined the association of identified rare variants with the two traits HDL-C and LDL-C, respectively. The analyses are listed in Table 3. Obviously, all tests have the ability to detect the association. However, BMST has lower p-values than most of the other tests, implicating its efficacy.

We then predicted the functional impact of each variant and incorporated prediction results into test. Variants were predicted by a popular server PolyPhen-2 (Adzhubei *et al.*, 2010), which outputs one of three categories for each variant: “benign”, “Possibly damaging”, and “Probably damaging”. We assign “benign” variants a prior probability 1/3, “Possibly damaging” variants 1/2, and “Probably damaging” variants 5/6, of being causative. PolyPhen-2 could not give a prediction for non-sense mutations. As non-sense mutations disrupt protein structure, they tend to damage protein function. Therefore, we divided them into the “Probably damaging” category and assign a prior probability 5/6. The association results are listed in Table 4. For BMST, association signals are boosted in both samples, indicating the value of using prior information. For the other tests, their association signals are boosted in LDL-C sample, but remain unchanged or instead got elevated in HDL-C sample. Scrutiny of this sample indicates that several mutations in case group are predicted as “benign”. Weighting variants directly by these predictions may reduce association signal in case group. Instead, by specifying these predictions as prior distribution, BMST still has chance to pick them up through posterior distribution.

Discussion

In this study, we have proposed a novel method to test the association of rare variants and complex diseases. By modeling in a Bayesian marker selection framework, we incorporated two sources of information to highlight likely causal variants: individual association signal and prior knowledge of functional impact. Simulation studies showed that the proposed method outperformed several existing methods in certain conditions. The proposed method was applicable to large-scale association scan.

Interaction between rare variants could also be a cause of complex diseases (Liu & Leal, 2010), and it is easy to extend the proposed method to include interaction effect (Zhang & Liu, 2007). However, there is a fundamental distinction in the analyses of interaction effects for rare and common variants. For two-way interaction model, three genotypes at each locus correspond to nine possible two-locus genotype combinations. For common variants, each combination could be interpreted as interaction effect. For rare variants, however, because it is rare alleles that are suspected to be causative, only combinations between rare alleles could be interpreted as interaction effect. This interpretation reduces the number of combination of interest from nine to four. Furthermore, the occurrence of homozygous mutant genotype is much rarer, so only one joint combination is left to be of primary interest to detect interaction effect.

For modeling simplicity, we assume independence among rare variants. However, our method is still valid when this assumption is violated, as demonstrated by our simulation results. We also notice that most of the existing grouping-based tests (except KBAC) implicitly assume independent effects among variants regardless of the presence of LD, because for these tests when mutation signals of multiple variants are merged, the patterns of LD among them are also collapsed. In fact, assuming the rare allele at a second variant to be in LD with the causal rare allele is equivalent to assuming two causal alleles.

In this study, we implement a one-sided test by examining the existence of causal variants. The test of protective variants is straightforward to implement in the same manner. Alternatively, a two-sided test could be constructed by combining association signals of these two groups of variants together. Two-sided test is expected to be more powerful than one-sided test in the presence of both causal and protective effects. However, most of current empirical studies focus only on identifying disease-causing mutations, left the research of distributions of protective variants sparse.

Our method is applicable to analyzing arbitrary genomic regions, though we focus our analyses on gene-coding regions. The prior information of variants at non-coding regions could be evaluated and predicted by studying ortholog sequences. However, it is still challenging to specify a proper prior distribution according to the conservation of variants, because the relationship between sequence conservativeness and phenotype variation are not well matched (Ahituv *et al.*, 2007b).

Common variants are an important source of genetic variation for complex diseases as well. When common and rare variants are analyzed together, a multivariate method should be adopted. Specifically, the synthesized mutation signal by the group of rare variants and common variants are analyzed jointly in a multivariate model.

In conclusion, we propose a novel method to test the association of multiple rare variants and complex diseases. The program implementing our method is available upon request from the authors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Lei Zhang is supported by National Natural Science Foundation of China project (31100902). Hong-Wen Deng is supported by Shanghai Leading Academic Discipline Project (S30501) and startup fund from Shanghai University of Science and Technology, grants from NIH (P50AR055081, R01AG026564, R01AR050496, RC2DE020756, R01AR057049, and R03TW008221) and the Franklin D. Dickson/Missouri Endowment and the Edward G. Schlieder Endowment. The 1000 genomes project is acknowledged for the access of the pilot exon project datasets. Dr. Jonathan C. Cohen is acknowledged for the access of the summary datasets in the two of his published papers.

References

- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010; 7:248–9. [PubMed: 20354512]
- Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, Yosef N, Ruppin E, Sharan R, Vaisse C, Sunyaev S, Dent R, Cohen J, Mcpherson R, Pennacchio LA. Medical sequencing at the extremes of human body mass. *Am J Hum Genet*. 2007a; 80:779–91. [PubMed: 17357083]
- Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. Deletion of ultraconserved elements yields viable mice. *PLoS Biol*. 2007b; 5:e234. [PubMed: 17803355]
- Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet*. 2010; 11:773–85. [PubMed: 20940738]
- Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, Bafna V. A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol*. 2010; 6:e1000954. [PubMed: 20976246]
- Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*. 2008; 4:e1000083. [PubMed: 18516229]
- Cohen JC, Boerwinkle E, Mosley TH Jr. Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med*. 2006; 354:1264–72. [PubMed: 16554528]
- Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, Mcpherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*. 2004; 305:869–72. [PubMed: 15297675]
- Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, Mcvean GA. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467:1061–73. [PubMed: 20981092]
- Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered*. 2010; 70:42–54. [PubMed: 20413981]
- Hernandez RD. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*. 2008; 24:2786–7. [PubMed: 18842601]
- Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet*. 2011; 7:e1001289. [PubMed: 21304886]
- Ji W, Foo JN, O'roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet*. 2008; 40:592–9. [PubMed: 18391953]
- Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet*. 2007; 80:727–39. [PubMed: 17357078]
- Lehmann, EL. *Elements of Large Sample Theory*. Springer; 1999. p. 174

- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008; 83:311–21. [PubMed: 18691683]
- Li Y, Byrnes AE, Li M. To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. *Am J Hum Genet.* 2010; 87:728–35. [PubMed: 21055717]
- Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.* 2010; 6:e1001156. [PubMed: 20976247]
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5:e1000384. [PubMed: 19214210]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature.* 2009; 461:747–53. [PubMed: 19812666]
- Mooney SD, Altman RB. MutDB: annotating human variation with functionally relevant data. *Bioinformatics.* 2003; 19:1858–60. [PubMed: 14512363]
- Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science.* 2009; 324:387–9. [PubMed: 19264985]
- Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003; 31:3812–4. [PubMed: 12824425]
- Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet.* 2006; 7:61–80. [PubMed: 16824020]
- Price AL, Kryukov GV, De Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010; 86:832–8. [PubMed: 20471002]
- Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet.* 2001; 69:124–37. [PubMed: 11404818]
- Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet.* 2002; 11:2417–23. [PubMed: 12351577]
- Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics.* 2006; 22:2183–5. [PubMed: 16809394]
- Siva N. 1000 Genomes project. *Nat Biotechnol.* 2008; 26:256. [PubMed: 18327223]
- Sul JH, Han B, He D, Eskin E. An Optimal Weighted Aggregated Association Test for Identification of Rare Variants Involved in Common Diseases. *Genetics.* 2011
- Tang W, Wu X, Jiang R, Li Y. Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy. *PLoS Genet.* 2009; 5:e1000464. [PubMed: 19412524]
- Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, O'meara S, Latimer C, Dicks E, Menzies A, Stephens P, Blow M, Greenman C, Xue Y, Tyler-Smith C, Thompson D, Gray K, Andrews J, Barthorpe S, Buck G, Cole J, Dunmore R, Jones D, Maddison M, Mironenko T, Turner R, Turrell K, Varian J, West S, Widaa S, Wray P, Teague J, Butler A, Jenkinson A, Jia M, Richardson D, Shepherd R, Wooster R, Tejada MI, Martinez F, Carvill G, Goliath R, De Brouwer AP, Van Bokhoven H, Van Esch H, Chelly J, Raynaud M, Ropers HH, Abidi FE, Srivastava AK, Cox J, Luo Y, Mallya U, Moon J, Parnau J, Mohammed S, Tolmie JL, Shoubbridge C, Corbett M, Gardner A, Haan E, Rujirabanjerd S, Shaw M, Vandeleur L, Fullston T, Easton DF, Boyle J, Partington M, Hackett A, Field M, Skinner C, Stevenson RE, Bobrow M, Turner G, Schwartz CE, Gecz J, Raymond FL, Futreal PA, Stratton MR. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet.* 2009; 41:535–43. [PubMed: 19377476]
- Tucker T, Marra M, Friedman JM. Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet.* 2009; 85:142–54. [PubMed: 19679224]

- Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U S A*. 2005; 102:7882–7. [PubMed: 15905331]
- Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol*. 2011; 35:57–69. [PubMed: 21181897]
- Zhang L, Pei YF, Li J, Papasian CJ, Deng HW. Efficient utilization of rare variants for detection of disease-related genomic regions. *PLoS One*. 2010a; 5:e14288. [PubMed: 21170328]
- Zhang L, Pei YF, Li J, Papasian CJ, Deng HW. Improved detection of rare genetic variants for diseases. *PLoS One*. 2010b; 5:e13857. [PubMed: 21079782]
- Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nat Genet*. 2007; 39:1167–73. [PubMed: 17721534]
- Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol*. 2010; 34:171–87. [PubMed: 19847924]

Figure 1.

Power with 10 kb gene length.

A pool of 10,000 quantitatively phenotyped individuals is generated, from which 1,000/1,000 individuals with extreme phenotypes are selected as cases/controls. A gene with 10 kb length is simulated, with gene heritability varying from 0.1% to 1.0%. Power is estimated at significance level 1% with 10,000 replicates. Both the absence (Upper panel A) and presence (Lower panel B) of LD among rare variants are simulated.

Figure 2.
Power with 20 kb gene length.
See legend for figure 1 for detail.

Figure 3.
Comparison of power between 10 and 20 kb gene lengths.

Figure 4.

Power when prior knowledge is used.

Prior knowledge is acquired by simulating the predicted functional impact for each variant. The sensitivity and specificity of prediction are set to 0.6 and 0.2. Power is estimated at 10 kb (Upper panel) and 20 kb (Lower panel) gene lengths at significance level 0.01.

Table 1

Type-I error rates

Test	Target level			
	1.0%		0.10%	
	10k	20k	10k	20k
LD is absent				
BMST	0.9	1.2	0.10	0.11
KBAC	1.2	1.1	0.11	0.10
RC	0.8	1.3	0.08	0.14
VT	1.0	0.9	0.12	0.09
WS	0.9	1.2	0.10	0.13
RW	1.1	1.2	0.08	0.11
Collapsing	0.7	1.1	0.06	0.10
LD is present				
BMST	1.4	1.1	0.07	0.10
KBAC	1.3	1.5	0.08	0.13
RC	1.2	0.4	0.11	0.06
VT	1.4	0.8	0.09	0.06
WS	1.5	1.0	0.08	0.13
RW	1.4	1.0	0.23	0.12
Collapsing	0.6	1.0	0.09	0.09

Notes: One-thousand cases and 1,000 controls are simulated at two gene lengths: 10 and 20 kb. Type-I error rates are estimated at target levels 1% and 0.1% by performing 10,000 replicates. Abbreviations: BMST, the proposed Bayesian marker selection test; KBAC, the kernel-based adaptive cluster test (Liu & Leal, 2010); RC, the rare cover test (Bhatia *et al.*, 2010); VT, the variable threshold test (Price *et al.*, 2010); WS, weighted sum test (Madsen & Browning, 2009); RW, replication based weighting test (Ionita-Laza *et al.*, 2011); Collapsing, collapsing test (Li & Leal, 2008)

Table 2

Power estimate when prior knowledge is non-informative

Length	Test				
	BMST	VT	WS	RW	
10kb	No prior	0.81	0.71	0.60	0.78
	Prior	0.73	0.62	0.45	0.66
20kb	No prior	0.81	0.68	0.60	0.73
	Prior	0.74	0.57	0.44	0.62

Notes: A pool of 10,000 quantitatively phenotyped individuals is generated, from which 1,000/1,000 individuals with extreme phenotypes are selected as cases/controls. Gene heritability is set to 0.5%. Prior probability is simulated from the Uniform distribution $U(0, 1)$, independent of the causal status of variants. Power is estimated at significance level 1% with 10,000 replicates.

Table 3

Analyses of two real datasets

Sample	Test					
	BMST	KBAC	VT	WS	RC	Collapsing RW
HDL-C	2.0E-07	2.0E-07	3.0E-07	4.0E-07	4.0E-07	5.0E-07
LDL-C	1.0E-05	2.0E-05	8.0E-05	6.0E-05	7.0E-04	1.0E-04

Notes: For HDL-C sample, 283 individuals with low HDL-C levels and 236 individuals with high levels were sequenced at three genes (*ABCA1*, *APOA1*, and *LCAT*). A total of 40 rare variants in the three genes were combined and analyzed together. Of these variants, 35 are exclusive in low group and the other 5 are exclusive in high group. For LDL-C sample, 256 individuals with low Ca:L ratios and 256 individuals with high ratios were sequenced at gene *NPC1L1*. A total of 25 rare variants were combined and analyzed together. Of these variants, 20 are exclusive in low group and the other 5 are exclusive in high group. P-values were estimated by 1×10^8 permutation replicates.

Table 4

Analyses of two real datasets when prior knowledge is used

Sample	Test			
	BMST	VT	WS	RW
HDL-C	8.9E-08	3.0E-07	4.0E-07	2.0E-06
LDL-C	2.0E-06	2.0E-06	3.0E-06	8.2E-05

Notes: The functional impact of each variant was predicted by PolyPhen-2 (Adzhubei *et al.*, 2010). Three categories: “Benign”, “Possibly damaging”, and “Probably damaging”, given by PolyPhen-2 were converted into three prior probabilities: 1/6, 3/6, and 5/6. P-values were estimated by 1×10^7 permutation replicates, but for BMST, 1×10^8 replicates were performed.