Detecting Rare Variants for Quantitative Traits Using Nuclear Families

Wei Guo\textsuperscript{a, b}, Yin Yao Shugart\textsuperscript{a}

\textsuperscript{a}Division of Intramural Division Program, National Institute of Mental Health, National Institute of Health, Bethesda, Md., USA; \textsuperscript{b}Key Laboratory for Applied Statistics of Ministry of Education and School of Mathematics and Statistics, Northeast Normal University, Changchun, China

Key Words
Rare variants · Nuclear families · Quantitative traits · Weights · Penalized regression

Abstract
With the advent of sequencing technology opening up a new era of personal genome sequencing, huge amounts of rare variant data have suddenly become available to researchers seeking genetic variants related to human complex disorders. There is an urgent need for the development of novel statistical methods to analyze rare variants in a statistically powerful manner. While a number of statistical tests have already been developed to analyze collapsed rare variants identified by association tests in case-control studies, to date, only two \textit{FBAT} tests-for-rare (described in the updated \textit{FBAT} version v2.0.4) have applied collapsing methods analogously in family-based designs. For further research in this area, this study aims to introduce three new beta-determined weight tests for detecting rare variants for quantitative traits in nuclear families. In addition to evaluating the performance of these new methods, it also evaluates that of the two \textit{FBAT} tests-for-rare, using extensive simulations of situations with and without linkage disequilibrium. Results from these simulations suggest that the four tests using beta-determined weights outperform the two collapsing methods used in \textit{FBAT} (-v0 and -v1). In addition, both the linear combination method (detailed in the \textit{FBAT} menu v2.0.4) and the multiple regression method (mixing LASSO and Ridge penalties) performed better than the other two beta-determined weight tests we proposed. Following testing and evaluation, we submitted four new beta-determined weight methods of statistical analysis in a computer program to the Comprehensive R Archive Network (CRAN) for general use.

Background

Genome-wide association studies revealed that the already identified genetic variants that have been associated with human common disorders can explain only a small portion of the heritability [1], and that 'the missing heritability' might be due to the joint contributions from multiple rare variants which could not be generated by the earlier platforms of genome-wide association studies [2]. In addition to the rare variants, the other factors including copy number variants, polygenes of modest effects, unknown epistasis effects, potential genetic interactions, gene-environment interactions, and epigenetic effects may also contribute to the currently missing heri-
tability. By investigating the effect of rare variants, we hope to discover more biological factors which are relevant to disease etiology.

With the advent of DNA sequencing technology, next-generation sequence has become a promising tool to detect both rare and common variants underlying common human disorders including psychiatric disorders, cardiac vascular diseases, and cancer. As thoroughly discussed by Bansal et al. [3], in addition to the need for significant improvements in sequence technologies in terms of accuracy and coverage, effective analytical tools are also highly desired for data analysis and for making the biological link between genotypic and phenotypic measures. Recently, a number of statistical methods have been developed for teasing out the contribution of rare variants to the development of complex traits. For case-control designs, current approaches mainly focus on collapsing the rare variants to achieve higher statistical power. Of the methods evaluated by Bansal et al. [3], we will discuss a few that are more relevant to the methods we have developed or evaluated.

The Combined Multivariate and Collapsing test [4] is among the first approaches proposed for the analysis of rare variants. This test compares the number of non-synonymous alleles in different phenotypic groups. Under the Combined Multivariate and Collapsing test, however, variants are either included or excluded in the test statistic on the basis of an assumed allele frequency threshold, regardless of their actual numbers in cases or controls. Other researchers suggest using frequency-weighted tests with larger weights assigned to rare variants on the assumption that the rare variants are more likely to influence the disorder of interest [5]. This method obviates the need to choose thresholds for allele frequencies. The later Kernel-Based Adaptive Cluster test also uses a weighting scheme, which in this case reflects the effect size of individual variants [6]. Yet another approach combines rare variants into a single test by selecting an allele frequency threshold based on the observed data. This variable threshold approach [7] is favored by the researchers who viewed determining an optimal weighting scheme or allele frequency threshold as an impossible task. (For comprehensive reviews of these collapsing methods, see [3, 8, 9].)

In addition to the methods described above, penalized regression can be used to analyze rare variants while avoiding the singular design matrix that can occur with sequence variants that are very rare. Two groups dealing with genetic applications [10, 11] have developed a novel use of regularized regression methods to handle rare issues in association analysis. One study focused on determining rare haplotypes [10], while the other aimed at distinguishing risk and neutral variants [11]. To date, however, only two FBAT tests for rare variants (described in the updated version v2.0.4 of FBAT and in Yip et al. [12]) have sought to apply analogous collapsing methods to analyze family-based designs for rare variants.

In this study, we show how penalized regression can be applied in a family-based design, as detailed below. The main disadvantage of regression-based association methods is that, when variants are very rare, the matrix of genotypes may become singular. LASSO and ELASTICNET (a mix of LASSO and Ridge) are both well-known tools that shrink regression coefficients to reduce data dimensions. We first discuss using the LASSO method (in the Ŝlasso test) to analyze rare variants in family-based designs for quantitative traits. We then explore how the ELASTICNET penalty (in the Ŝelasticnet test) combines the strengths of the quadratic regularization with those of adaptively weighted LASSO shrinkage. Because it has been well documented that family-based designs are more robust than case-control designs to population stratification, we used family-based designs in this study. Our goal here is to implement three existing and three novel methods of rare analysis in a family-based design (which comprises both genotypic and phenotypic data for two sequenced parents and their offspring), and then to compare their performance.

In the updated version of the FBAT software (v2.0.4), both unweighted (FBAT-v0) and weighted (FBAT-vl) sum tests are considered for analyzing rare variants. The weight is based on allele frequency, similar to the weighted sum statistic [5]. In addition, the linear combination (LC) method, which also uses the beta-determined weights, has been previously implemented in the FBAT software for common variants. FBAT-LC has been demonstrated to be a powerful test for analyzing multiple markers using the estimated regression coefficients by single marker regression as the weights [13].

To date, no study has evaluated the performance of FBAT-LC in the context of rare variants. For the sake of comparison, the power of FBAT-LC is also investigated with minor adjustments for the rare variants (to be detailed later). In addition, we propose three Ŝ tests using the regression coefficients estimated by multiple regression as the weights to detect rare variants for quantitative traits in nuclear families. The performances of three newly proposed methods and of the FBAT methods (FBAT-v0, FBAT-vl, and FBAT-LC) are all investigated using simulations.
Our paper is structured as follows. We review the existing FBAT tests (FBAT-v0, FBAT-v1, and FBAT-LC) for their application to rare variants. Then, we introduce our proposed tests in the application of multiple regression ($S_\beta$) and regularized regression ($S_{\beta \text{lasso}}$ and $S_{\beta \text{elasticnet}}$) in a two-generation family setting. Particularly, we evaluate the performance of the test statistics based on simulations. Finally, we summarize the advantages and disadvantages of all the methods investigated in the Results and Discussion sections.

**Methods**

A number of statistical tests have already been developed for analyzing collapsed rare variants in association test data from case-control studies. Because the family-based design is more robust to population stratification than the case-control design, however, we chose it for this study. Recognizing that FBAT-LC (implemented in v2.0.4) has recently been used to collapse underlying rare variants in family-based designs, we here propose the novel approach of using regression coefficients as weights. In the past, FBAT-LC was used to analyze common variants associated with quantitative traits. However, when we applied FBAT-LC to rare variants with very low frequencies ($<5$ observations for a particular variant), FBAT-LC failed to generate results (this failure likely stemmed from non-convergence of regression). To avoid this problem, we modified the FBAT-LC test ($mFBAT-CL$) and implemented this modified version in our *betalab* program. We intended to explore the other methods such as $S_\beta$ using regression coefficients as weights estimated by multiple marker regression. As noted above, our new collapsing methods use LASSO and ELASTICNET, two well-known tools for reducing data dimension by shrinking regression coefficients. It is our expectation that, where variants are very rare, these two new penalized methods ($S_{\beta \text{lasso}}$ and $S_{\beta \text{elasticnet}}$) may have advantages over FBAT-LC. The following beta-determined weight methods including $mFBAT-CL$, $S_\beta$, $S_{\beta \text{lasso}}$ and $S_{\beta \text{elasticnet}}$ were therefore selected for testing and evaluation.

**Statistical Methods**

For the readers’ convenience, we recall the notion of Xu et al. [13] and adopt all the equation given by the original authors. Let $M$ be the total number of variants in the gene of interest. Let $N$ be the total number of families, and $n_i$ be the number of offspring in the $i$-th family. Let $T_{ij}$ and $X_{ij}^k$ denote the trait value and the genotype score at the $k$-th marker of the $i$-th family. Let $\tilde{z}_i$ denote the trait value and the genotype score estimated from the sample. This single marker test statistic can be written as

$$\hat{z}_i = \sum_i T_{ij} (X_{ij}^k - E(X_{ij}^k)),$$

which describes the correlation between the trait values and the genotype score at the $k$-th marker if $E(T_{ij}) = 0$, and

$$\hat{z}_i / \sqrt{\text{Var}(\hat{z}_i)}$$

is the standardized $FBAT$ test statistic and approximately $N(0, 1)$ when the sample size $N$ is reasonably large. It was clearly illustrated that the expectation of the genotype score $E(X_{ij}^k)$ is calculated with respect to parental genotypes at the $k$-th variant in the $i$-th family under the null distribution of parental random transmission using Mendel’s laws. When one or both parents are missing, $E(X_{ij}^k)$ could be calculated using the sufficient statistics for parental genotypes [14].

In order to test for the rare variant association, we collapsed $M$ variants according to specific weights defined by us as follows:

$$\sum_i \sum_j T_{ij} (X_{ij}^k - E(X_{ij}^k)) = \sum_i w_i \hat{z}_i = W^{T} \hat{Z},$$

where the weight $W = (w_1, w_2, ..., w_M)^T$ is later defined in different forms. Here, the general form for the $S$ test statistic can be written as

$$S = \frac{W^{T} \hat{Z}}{\sqrt{W^{T} \Sigma W}},$$

where $\Sigma$ is the estimated covariance matrix of $\hat{Z}$ and can be calculated from the empirical estimator of the covariance matrix $\sigma$ with elements given by

$$\sigma_{ml} = \sum_i \left( \sum_j T_{ij} (X_{ij}^m - E(X_{ij}^m)) \sum_j T_{ij} (X_{ij}^l - E(X_{ij}^l)) \right),$$

where $m = 1, \ldots, M$ and $l = 1, \ldots, M$. The $S$ statistic follows the $N(0, 1)$ distribution approximately when $N$ is large. As suggested by Rakovski et al. [15], let $D$ be the diagonal matrix with elements equal to $\text{Var}(\hat{z}_i)$, and the corresponding adjusted variance matrix $\Sigma$ is defined by

$$\Sigma' = D^{-1/2} \left[ \text{Diag}(\Sigma^{-1})^{-1/2} \sum \text{Diag}(\Sigma^{-1})^{-1/2} D^{-1/2} \right] D^{-1/2},$$

which is a better empirical covariance estimate and is therefore used in the $S$ test in this paper. The $p$ value of the $S$ test defined in equation 3 is calculated by $2[1 - \Phi(S)]$, where $\Phi$ is the distribution function of $N(0, 1)$, and $\Phi(S)$ is the probability that a standard normal distributed variable is less than the absolute value of $S$. Thus, the $S$ tests become two-sided tests (different from the ones implemented in FBAT). In addition, we would like to note that equations 1, 2, 3, and 5 are identical to the ones proposed by Xu et al. [13].

**Existing Methods**

The different weighting schemes are investigated for the test statistic $S$ (as shown in equation 3) implemented in the FBAT software. The first one is the $FBAT$-v0 test. Under this scheme, the weights equal 1, i.e. $w_k = 1$, $k = 1, \ldots, M$, and $v0$ is the unweighted sum statistics in the FBAT software. The second is the $FBAT$-v1 test, in which the weights are defined as

$$w_k = 1 / \sqrt{\xi p_k (1 - p_k)}, \quad k = 1, \ldots, M,$$

where $\xi$ is the total number of nuclear families, and $p_k$ is the allele frequency for the $k$-th variant estimated from the sample. This weight follows the weighting scheme used by Madsen and Browning [5] in which they estimated the allele frequencies using controls sampled from the general population. Note that both $FBAT$-v0 and $FBAT$-v1 are two-sided tests which have now been implemented in the FBAT software.

In addition to the frequency-determined weights, the beta-determined weights estimated from the regression model are also
the natural candidates for weighting rare variants, which are similar to the ones used in some other methods [9, 13]. Using the regression model, the effect size of each marker can be estimated from the parental genotypes and the offspring’s trait value. Similar to the FBAT-LC methods [13], the weights are obtained from the following model:

\[ E(T_j) = \alpha + \beta E(X_{ij}^k), \]  

where \( i = 1, 2, ..., N; j = 1, 2, ..., n; k = 1, 2, ..., M, \) and \( E(X_{ij}^k) = X_{ij}^k \) for parents and children without parents. Parental phenotypic values are not used in the above regression model. When \( w_k = \hat{\beta}_k \) for all \( k = 1, 2, ..., M, \) the \( S \) test becomes the FBAT-LC test in the FBAT software. While dealing with rare variants, it is possible that the estimated coefficient is unavailable. When the \( \hat{\beta}_k \) is unavailable in some rare variants, we set \( w_k \) as zero. Further, the FBAT-LC test here is modified to be a two-sided test under the framework of the \( S \) test, while the original FBAT-LC test in the FBAT software is a one-sided test. Therefore, the modified test is named mFBAT-LC in this study.

Novel Methods

In the first \( S \) test, we propose the weights \( w_k = \hat{\beta}_k, k = 1, 2, ..., M, \) where \( \hat{\beta}_k \) are estimated from the multiple marker regression model:

\[ E(T_j) = \alpha + \sum_{k=1}^{M} \beta_k E(X_{ij}^k), \]  

where \( i = 1, 2, ..., N; j = 1, 2, ..., n; \) The test statistic \( S \) under this weighting scheme is called \( S_\text{LM} \) test. Second, in the case of the scenario that the sequence variants are rare, penalized regression is considered to avoid the singular design matrix. In the second test, the weights \( w_k = \hat{\beta}_k, k = 1, 2, ..., M, \) where \( \hat{\beta}_k \) are estimated from the multiple marker regression model having the LASSO penalty. In the LASSO regression model, the \( \hat{\beta}_k \) are the least square estimates that minimize the loss of function

\[ \sum_i \sum_j [E(T_j) - \alpha - \sum_k \beta_k E(X_{ij}^k)]^2 + \lambda \sum_k |\beta_k|, \]  

where \( \lambda \) is the tuning parameter chosen using AIC criteria. The test statistic \( S \) under this weight scheme is called \( S_\text{lasso} \) test. In the third test, the weights \( w_k = \hat{\beta}_k, k = 1, 2, ..., M, \) where \( \hat{\beta}_k \) are estimated by the multiple marker regression model having the ELASTICNET penalty, which is a mixed penalty of \( l_1 \) (LASSO) and \( l_2 \) (Ridge). In the ELASTICNET regression, the \( \hat{\beta}_k \) are the least square estimates that minimizes the loss of function

\[ \sum_i \sum_j [E(T_j) - \alpha - \sum_k \beta_k E(X_{ij}^k)]^2 + \lambda \left[ \frac{1}{2} \sum_k \beta_k^2 + \theta \sum_k |\beta_k| \right], \]  

where \( \theta \) is set as 0.05 and \( \lambda \) is the tuning parameter chosen using AIC criteria. The ELASTICNET penalty can stabilize the \( l_1 \) regularization path [16]. Under this circumstance, the ELASTICNET penalty is very close to the Ridge penalty. Both LASSO and ELASTICNET methods have now been implemented in the R package glmnet [17, 18]. The test statistic \( S \) under this weight scheme is called \( S_\text{elasticnet} \) test.

We would like to further clarify that the FBAT-\( v0 \) test is an unweighted test, while the FBAT-\( v1 \) test uses frequency-determined weights. Additionally, the four beta-determined weight tests, \( LC \), \( S_\beta \), \( S_\text{lasso} \), and \( S_\text{elasticnet} \) use the estimated \( \hat{\beta}_k \) from the regression model as the weights. Note that all six tests compared here are two-sided tests. The \( FBAT-v0 \) and \( FBAT-v1 \) tests have already been implemented in the FBAT software. The \( S_\beta \), \( S_\text{lasso} \), and \( S_\text{elasticnet} \) tests are newly proposed tests. The FBAT-LC test was initially implemented in FBAT, but it cannot generate output when the estimated regression coefficient is not available due to the rareness of the variants. To be noticed, we modified the estimated coefficients by replacing NA with 0 and implemented the mFBAT-LC method in our betafam program.

Simulation

For the sake of simplicity, our simulations avoid using population genetic models to generate genotypic data. Our intention here is to keep our simulation models simple enough for us to monitor statistical behavior by changing certain parameters. The genetic models used here resemble those assumed by Li and Leal [4] and Lin and Tang [9], although it can be debated whether those models are in fact realistic. Under the assumption of no linkage disequilibrium (LD) between the adjacent variants (for parameter settings 1 and 2), the parental genotypes are randomly sampled under the Hardy-Weinberg equilibrium assumption based on given allele frequencies. Further, the genotypes of the offspring at each marker are assumed to be independently transmitted from the parental genotype without considering the LD structure. Under the assumption of LD (under parameter setting 3), the haplotypes of parents are assigned according to the haplotype frequencies, and then the parental haplotypes are randomly transmitted to the offspring without recombination. Two hundred and fifty nuclear families with two sibs (the total sample size is 1,000) are randomly sampled from the population. Finally, the quantitative trait values for each individual in the family are assigned to be a random variable according to the following trait model:

\[ T_j = \sum_{k=1}^{M} \beta_k X_{ij}^k + \epsilon_j, \]  

where \( \epsilon_j \) are independent and identically distributed and follow the standard normal distribution. \( X_{ij}^k \)'s are genotype scores simulated in the previous steps, and \( \beta_k \) will be specified in the simulation parameters.

Parameter Setting 1

The number of variants, \( M \), is set as 10, 50, 100, 200, and 500 in groups 1–5, respectively. We consider the four combinations of allele frequencies when \( M = 10 \) as follows: (1) \( p_k = 0.0005, k = 1, ..., 10, \) with a total frequency (the sum for 10 variants) of 0.05%; (2) \( p_k = 0.005, k = 1, ..., 10, \) with a total frequency of 0.5%; (3) \( p_k = 0.01k, k = 1, ..., 10, \) with a total frequency of 5.5%, and (4) \( p_k = 0.005k, k = 1, ..., 10, \) with a total frequency of 27.5%. When the number of variants, \( M \), is 50, 100, and 200, we repeat the above-defined allele frequencies for \( M/10 \) times. These four sets of allele frequencies are indexed by \( \pi_1, \pi_2, \pi_3, \) and \( \pi_4 \), respectively, in an increasing order of the total frequency of variants for each \( M \). When \( M = 500 \), the allele frequencies are taken to be \( p_k = 0.0001k, k = 1, ..., 100, \) and repeat the frequencies four times for the remaining 400 variants, and this set is indexed by \( \pi_5 \).
Five combinations of the $\beta$s in the trait model are considered as the following: (1) $\beta_k = 0$ $(k = 1,...,M)$ for investigating the type I error; (2) $\beta$s are frequency determined and defined as follows:

$$\beta_k = \begin{cases} \frac{x}{\sqrt{p_k(1 - p_k)}}, & \text{when } k = \max(10, 0.2M), \\ 0, & \text{elsewhere,} \end{cases}$$

where $x$ satisfies $\sum_{k} x_k = 10$, which is chosen such that the power is reasonably high for the purpose of comparison; (3) $\beta$s are equal, and $\beta_k = 1$ when $k \leq \max(10, 0.2M)$ and $\beta_k = 0$ for elsewhere; (4) 90% of the $\beta$s are positive and 10% are negative, and the $\beta$s are defined as follows:

$$\beta_k = \begin{cases} x, & \text{when } k \text{ is not multiple of } 10, \\ -x/2, & \text{when } k \text{ is multiple of } 10, \\ 0, & \text{elsewhere,} \end{cases}$$

where $x$ is chosen as 1 for the same purpose, and (5) 50% of the $\beta$s can be positive and the other 50% are negative, and the $\beta$s are defined as follows:

$$\beta_k = \begin{cases} x, & \text{when } k \text{ is an odd, } k \leq \max(10, 0.2M), \\ -x, & \text{when } k \text{ is an even, } k \leq \max(10, 0.2M), \\ 0, & \text{elsewhere,} \end{cases}$$

where $x$ is also chosen as 1. Under such a parameter setting, we simulated 10 risk variants (the effect size for these variants can be either positive or negative) when $M = 10$, and only 20% of the variants are considered 'risk variants' with same frequencies with neutral variants when $M = 100$, 200, and 500. These five sets of $\beta$s are indexed by $\psi_1,..., \psi_5$, respectively.

**Parameter Setting 2 for Effects of the Proportion of Risk Variants**

To evaluate the effect of the proportion of risk variants, we set the allele frequencies according to $\pi$, which is defined in parameter set 1. The $\beta$s in the trait model are considered as follows:

$$\beta_k = \begin{cases} x, & \text{when } k \text{ is an odd, } k \leq 500 \tau, \\ -x, & \text{when } k \text{ is an even, } k \leq 500 \tau, \\ 0, & \text{elsewhere,} \end{cases}$$

where $x$ is also chosen as 1 and $\tau = 5, 10, 20, 30, 40\%$, respectively, indicating the proportion of risk variants.

**Parameter Setting 3 for Effects of LD**

Consider 10 variants with equal allele frequencies of 0.005. Ten haplotypes are constructed as follows: $1101000000, 0110000000, 1010000000, 0101000000, 0010000000, 0001000000, 0000100000, 0000010000, 0000001000, 0000000100, 0000000010, 0000000001, 0000000000, 0000000000, 0000000000, 0000000000, 0000000000, 0000000000, 0000000000, 0000000000, 0000000000, 0000000000, 0000000000$. Four indicators of the minor allele and 0 indicates the major allele. Let $R_{ij}^2$ be the $R^2$ values between variant $i$ and variant $j$, where $i, j = 1,..., 10$. There are two strong pair-wise LD ($R_{23}^2 = R_{32}^2 = 0.638$) and two weak pair-wise LD ($R_{23}^2 = R_{32}^2 = 0.038$). For generating the data, the haplotypes of the parents are assigned according to the haplotype frequency and the assumption of Hardy-Weinberg equilibrium. Parental haplotypes are then randomly transmitted to the offspring without recombination. Five thousand replicates are generated, and the power is evaluated for a significance level of 0.01.

**Results**

As noted above, this study evaluates the power of existing statistical methods and offers several innovative new methods for the analysis of family-based data for rare genetic variants associated with complex human disease. Using simulated genetic models with different parameters, we compared the performance of the existing LC method (with a suggested modification) and three novel methods. While the modified two-sided FBAT-LC ($m$FBAT-LC) was expected to do well, we also explored how other multiple regression methods might be used to describe relationships between quantitative traits and genetic variants. Under the assumption a trait is caused by multiple variants in an additive fashion, for example, the ordinary least-squares estimator for the multiple linear regression model is the best linear unbiased estimator of the coefficients. To clarify, ‘best’ here means giving the lowest possible mean-squared error of the estimate.

Following, we present the type I error and power (under parameter setting 1), then we evaluate the effects of the proportion of risk variants (parameter setting 2) and investigate the effects of LD simulated between markers (parameter setting 3). In parameter settings 1 and 2, no LD was assumed between any of the markers. In parameter setting 1, we define five situations, each of which has a set of specified allele frequencies with a defined effect size. In parameter setting 2, we simulate a special situation with 500 variants to increase variety in the proportion of risk variants in a specified genetic model. In parameter setting 3, we take LD into account by constructing a haplotype in which the extent of LD is simulated in four pairs of variants ($r^2$ is 0.038 in two pairs and 0.638 in the other two pairs).

**Type I Error Rates and Power under Parameter Setting 1**

The type I error rates are investigated for the $LC$, $S_\beta$, $S$lasso, $S$elasticnet, FBAT-v0, and FBAT-v1 tests under ($\pi_1$, $\pi_2$, $\pi_3$, $\pi_4$) parameter sets for 10, 50, 100, 200, and 500 markers, respectively. As shown in table 1, the type I errors are acceptable when the nominal level is 0.05 or 0.01. The power of all six tests is shown in table 2. First, the power comparison is of interest between the frequency-determined weight methods and beta-determined weight methods. As it very much depends on the number of variants ($M$), the performance of the $S_\beta$ test depending on the multiple regression will be discussed separately. In most simulation conditions, the three beta-determined weight tests $LC$, $S$lasso and $S$elasticnet
outperform the two **FBAT** tests except when \( M = 10 \) and all 10 markers are considered to be risk variants. For example, under \( \pi_1 \) when \( M = 10 \) of \( \psi_1 \), the power of **FBAT-v0** and **FBAT-v1** is 13.2 and 6.7%, respectively (shown in table 2). However, the power of the other four tests is 100%. When \( \beta_s \) are in both directions as in \( \psi_1 \) to \( \psi_2 \), the frequency-determined weight methods are not able to recognize the directions; therefore, the two **FBAT** tests did not perform as well as the beta-determined tests. We also recognized that the **FBAT-v0** and **FBAT-v1** tests were not capable of assigning the zero weights to neutral variants when 80% of the variants were neutral in group 2–4 (\( M > 10 \)). Thus, we can conclude that, under most situations, **mFBAT-**LC, \( S_{\beta,lasso} \), and \( S_{\beta,elasticnet} \) provide better power than the two **FBAT** tests (**FBAT-v0** and **FBAT-v1**) because the beta-determined weight-based tests can estimate the weights reflecting the effect size of variants in both directions and can assign zero weights to neutral variants.

Second, the performance of the four beta-determined weight tests is investigated. The **mFBAT-**LC and \( S_{\beta,elasticnet} \) tests perform similarly and provide better power than the \( S_\beta \) and \( S_{\beta,lasso} \) tests in most of the situations compared. \( S_\beta \) has slightly better power than **mFBAT-**LC only when \( M = 10 \), as shown in table 2; however, the power difference is not statistically significant for the two tests. The \( S_\beta \) and \( S_{\beta,lasso} \) tests provide similar power when \( M \) is equal to 10 and 50; however, they fail to gain power as good as the **mFBAT-**LC and **S_{\beta,elasticnet}** tests when \( M \) is equal to 100 and larger. Particularly, \( S_\beta \) has almost no power in the extreme case when \( M = 500 \), and \( S_{\beta,lasso} \) lose power gradually as \( M \) increases from 100 to 500. Due to the small sample size relative to the large number of markers, the estimated coefficients from the regression model appear to be unstable.

Third, the power improves as the total frequency of risk variants increases. However, with the exception of \( \psi_2 \), the power of \( \pi_2 \) is slightly higher than \( \pi_1 \).

### Table 1. Type I error rates

<table>
<thead>
<tr>
<th>No. of variants</th>
<th>Allele frequencies</th>
<th>Significance level of 0.05 (0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( mFBAT-<strong>LC</strong> )</td>
<td>( S_\beta )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( S_{\beta,lasso} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( S_{\beta,elasticnet} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>**FBAT-**v0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>**FBAT-**v1</td>
</tr>
<tr>
<td>( M = 10 )</td>
<td>( \pi_1 ): 0.0005k, ( k = 1,..., 10 )</td>
<td>0.04 (0.005)</td>
</tr>
<tr>
<td></td>
<td>( \pi_2 ): 0.005</td>
<td>0.04 (0.005)</td>
</tr>
<tr>
<td></td>
<td>( \pi_3 ): 0.001k, ( k = 1,..., 10 )</td>
<td>0.047 (0.009)</td>
</tr>
<tr>
<td></td>
<td>( \pi_4 ): 0.005k, ( k = 1,..., 10 )</td>
<td>0.046 (0.006)</td>
</tr>
<tr>
<td>( M = 50 )</td>
<td>( \pi_1 ): 0.0005k, ( k = 1,..., 10 )</td>
<td>0.047 (0.007)</td>
</tr>
<tr>
<td></td>
<td>( \pi_2 ): 0.005</td>
<td>0.047 (0.007)</td>
</tr>
<tr>
<td></td>
<td>( \pi_3 ): 0.001k, ( k = 1,..., 10 )</td>
<td>0.046 (0.01)</td>
</tr>
<tr>
<td></td>
<td>( \pi_4 ): 0.005k, ( k = 1,..., 10 )</td>
<td>0.052 (0.009)</td>
</tr>
<tr>
<td>( M = 100 )</td>
<td>( \pi_1 ): 0.0005k, ( k = 1,..., 10 )</td>
<td>0.059 (0.01)</td>
</tr>
<tr>
<td></td>
<td>( \pi_2 ): 0.005</td>
<td>0.049 (0.01)</td>
</tr>
<tr>
<td></td>
<td>( \pi_3 ): 0.001k, ( k = 1,..., 10 )</td>
<td>0.047 (0.012)</td>
</tr>
<tr>
<td></td>
<td>( \pi_4 ): 0.005k, ( k = 1,..., 10 )</td>
<td>0.052 (0.01)</td>
</tr>
<tr>
<td>( M = 200 )</td>
<td>( \pi_1 ): 0.0005k, ( k = 1,..., 10 )</td>
<td>0.05 (0.009)</td>
</tr>
<tr>
<td></td>
<td>( \pi_2 ): 0.005</td>
<td>0.052 (0.009)</td>
</tr>
<tr>
<td></td>
<td>( \pi_3 ): 0.001k, ( k = 1,..., 10 )</td>
<td>0.045 (0.007)</td>
</tr>
<tr>
<td></td>
<td>( \pi_4 ): 0.005k, ( k = 1,..., 10 )</td>
<td>0.05 (0.009)</td>
</tr>
<tr>
<td>( M = 500 )</td>
<td>( \pi_1 ): 0.0001k, ( k = 1,..., 100 )</td>
<td>0.052 (0.007)</td>
</tr>
<tr>
<td></td>
<td>( \pi_2 ): 0.005</td>
<td>0.052 (0.001)</td>
</tr>
</tbody>
</table>

The nominal level is 0.05 and 0.01, and the results are based on 5,000 replicates. The number of variants, \( M \), is set as 10, 50, 100, 200, and 500. The four sets of allele frequencies are considered as follows, when \( M = 10 \), \( \pi_1 \) \( p_k = 0.0005k \), \( k = 1,..., 10 \); \( \pi_2 \) \( p_k = 0.005 \), \( k = 1,..., 10 \); \( \pi_3 \) \( p_k = 0.001k \), \( k = 1,..., 10 \); and \( \pi_4 \) \( p_k = 0.005k \), \( k = 1,..., 10 \). When \( M = 50 \), 100, and 200, the four sets \( \pi_1-\pi_4 \) are defined to repeat the allele frequencies defined above for \( M/10 \) times. The allele frequencies \( \pi_3 \) are taken to be \( p_k = 0.0001k \), \( k = 1,..., 100 \), and repeat this frequencies four times for the remaining 400 variants. All \( \beta_s \) are set as zero.
### Table 2. Power for different allele frequency and beta settings under the significance level of 0.05

<table>
<thead>
<tr>
<th>No. of variants</th>
<th>Allele frequencies</th>
<th>mFBAT-LC</th>
<th>$S_\beta$</th>
<th>$S_{\beta,\text{larsso}}$</th>
<th>$S_{\beta,\text{elasticnet}}$</th>
<th>FBAT-v0</th>
<th>FBAT-v1</th>
</tr>
</thead>
</table>

#### $\mu$: frequency-determined $\beta$

**$M = 10$**
- $(\pi_1): 0.005k, k = 1, \ldots, 10; 0.005$
  - $= 10$  
  | $0.326$ | $0.327$ | $0.315$ | $0.311$ | $0.722$ | $0.705$

**$M = 50$**
- $(\pi_1): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.738$ | $0.750$ | $0.731$ | $0.731$ | $0.956$ | $0.945$

**$M = 100$**
- $(\pi_1): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.880$ | $0.884$ | $0.878$ | $0.879$ | $0.972$ | $0.972$

**$M = 200$**
- $(\pi_1): 0.005k, k = 1, \ldots, 10; 0.005$
  | $1.000$ | $1.000$ | $1.000$ | $1.000$ | $1.000$ | $1.000$

**$M = 500$**
- $(\pi_1): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.238$ | $0.227$ | $0.224$ | $0.228$ | $0.246$ | $0.230$

#### $\mu$: equal $\beta$

**$M = 10$**
- $(\pi_1): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.441$ | $0.445$ | $0.426$ | $0.424$ | $0.879$ | $0.846$

**$M = 50$**
- $(\pi_1): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.844$ | $0.852$ | $0.845$ | $0.845$ | $0.993$ | $0.984$

**$M = 100$**
- $(\pi_1): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.859$ | $0.864$ | $0.846$ | $0.849$ | $0.983$ | $0.975$

**$M = 200$**
- $(\pi_1): 0.005k, k = 1, \ldots, 10; 0.005$
  | $1.000$ | $1.000$ | $1.000$ | $1.000$ | $1.000$ | $1.000$

**$M = 500$**
- $(\pi_1): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.378$ | $0.346$ | $0.379$ | $0.382$ | $0.422$ | $0.404$

#### $\nu$: 90% positive and 10% negative $\beta$

**$M = 10$**
- $(\pi_2): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.287$ | $0.289$ | $0.292$ | $0.287$ | $0.222$ | $0.217$

**$M = 50$**
- $(\pi_2): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.597$ | $0.588$ | $0.561$ | $0.564$ | $0.551$ | $0.521$

**$M = 100$**
- $(\pi_2): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.997$ | $0.951$ | $0.979$ | $0.972$ | $0.919$ | $0.901$

**$M = 200$**
- $(\pi_2): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.997$ | $0.951$ | $0.999$ | $0.995$ | $0.919$ | $0.940$

**$M = 500$**
- $(\pi_2): 0.005k, k = 1, \ldots, 10; 0.005$
  | $0.973$ | $0.118$ | $0.968$ | $0.985$ | $0.976$ | $0.952$

Fourth, although in theory FBAT-v1 is expected to outperform FBAT-v0 in $\psi_2$, $v1$ and $v0$ have similar power (table 2). For the equal weights in $\psi_0$, FBAT-v0 appears to be slightly more powerful than FBAT-v1.

**Effects of the Proportion of Risk Variants under Parameter Setting 2**

To evaluate the effect of the proportion of risk variants under the parameter set 2, the power for the proportion of risk variants is shown in figure 1. In general, all tests gain power when the proportion of risk variants increases. The $mFBAT-LC$, $S_{\beta lasso}$ and $S_{\beta elasticnet}$ gain reasonable power when more risk variants are included in the simulated genes. However, FBAT-v0, FBAT-v1, and $S_{\beta}$ do not seem to benefit much from the increased number of risk variants.

Further, our simulation results clearly indicate that the $S_{\beta}$ model fails to obtain good estimates when there are more than 100 variants, i.e. the ratio of the number of markers to the sample size in regression is larger than 20% (fig. 1). We would like to state that this ratio of 20% is only a rough threshold based upon the simulations we conducted.

**Effects of LD under Parameter Setting 3**

In the presence of LD, type I errors are slightly lower than the nominal level of 0.01, as shown in table 3 ($\psi_1$). For equal $\beta$s under $\psi_3$, the power of beta-determined weight-based methods improves significantly (6–14%) in the presence of LD. By contrast, since LD has no obvious effect on the allele frequencies in the sampled individuals, frequency-determined weight-based methods such as...
FBAT-v0 and FBAT-v1 have similar power in the presence of LD. We need to point out here that the nature of $\psi_2$ is equivalent to that of $\psi_3$ because of the assumed equal allele frequencies under these two settings. In Table 3, under $\psi_2/\psi_3$, $\psi_4$, and $\psi_5$, the power of all six tests improve slightly in our simulation.

**Discussion**

Our goal is to propose three new family-based methods to test for associations between rare variants. We favor the family design for two reasons: (1) it is well accepted that these methods are more robust to population structure than case-control-based methods, and (2) if a rare variant is observed in a family, identity-by-descent sharing among family members may cause it to be enriched more remarkably in frequency-based than in population-based designs. The family design is therefore expected to be a more powerful way to test for association between rare variants and phenotypes related to human disorders. In family data, moreover, potential sequencing errors which are still a concern in sequencing data generated by current platforms can be eliminated by checking for Mendelian consistency.

With the aim of improving statistical tools available for the analysis of rare variants in family data, this study modifies an existing LC method, proposes three novel methods, and uses simulated genetic models with different parameters to test and compare their performance. We expect the modified two-sided FBAT-LC (mFBAT-LC) to work well where a gene of interest contains both positive and negative variants. We also expect (see Methods section above) the new multiple regression models proposed to prove more suitable for the description of relations between quantitative traits and genetic variants.

Although the idea for collectively testing for association with a group of rare variants has been well accepted by the field of statistical genomics, different ways of combining genotypic information can still be considered. The most popular weights in collapsing methods assign larger weights to rare variants [5], which is an optimal weight under the assumption that the rarer variant carries a higher disease risk. However, the frequencies of the variants may not be the only factor that determines the strength of genetic effect. Therefore, we are employing a few different approaches such as mFBAT-LC, $S_{\beta}$, $S_{\text{lasso}}$, and $S_{\text{elasticnet}}$ which considers weighting individual nucleotides by their putative functional effects (here we refer to the beta-determined weights used in this paper). The mFBAT-LC method uses the single marker regression method to estimate the effect sizes, which was already implemented in the FBAT software, but the other three $S$ tests proposed use the multiple regression method to estimate the effect sizes. The significant advantage here is that family-based studies can use the independent information from parents to estimate the effect sizes, which are the weights used in collapsing. Therefore, the $S$ statistic follows normal distribution and thus avoids the need for permutation to obtain the statistical p values while such permutation is required in the case-control design.

In this simulation, the type of regression coefficients and size of gene (number of variants) varied, and performances are evaluated with and without LD present (in parameter setting 3). For comparison, $\beta$s in the simulation model are assumed to be frequency determined, in the same direction, and, respectively, both positive and negative. Of the six methods compared, the mFBAT-LC and $S_{\text{elasticnet}}$ approaches are shown to perform well in all simulated situations and are therefore recommended for testing for association between rare variants and complex human traits. We also note that: (1) where most variants in a gene are risk variants with effect sizes in the

![Image](https://example.com/image.png)
same direction, the $FBAT$-v0 and $FBAT$-v1 methods are reasonably powerful; (2) four of the new beta-determined weight-based methods, which can estimate effect size by independent regression of parental genotypes only, overcome limitations caused by difference of effect direction or size to achieve reasonable power, and (3) computation speed (particularly with a large number of genetic variants) is faster using the implemented regularized regression approach than the approach implemented in $mFBAT$-LC.

We would like to point out that the performance of the $S_\beta$ test based on the multiple regression model needs to be interpreted with caution. If the true $\beta$s are known under an additive model and used as the weights, the $S$ test is undoubtedly the best. This was clearly observed in the simulation studies (data not shown). In theory, if the sample size is large enough, the multiple regression model has the best linear unbiased estimates and is expected to outperform the single marker regression model. However, due to the small sample size relative to the large number of markers, the estimated coefficients from the regression model tend to be unstable. Further, the variance of the estimates can be very large due to the rarity of the genetic variants. This motivated us to choose the penalized regression methods in this study.

Our simulation results also indicate that the $S_\beta$ model fails to gain good estimates when the number of markers is greater than 50, i.e. the ratio of the number of markers to the sample size in regression is larger than 5% (table 2). We would like to note that the ratio of 5% is only a rough threshold based upon the simulations we conducted. Obviously, it is difficult to determine the real cutoff since it largely depends on the rank of $X'X$ in the regression that is directly related to the allele frequencies, the number of markers, and the sample size. For this reason, the regularized regression methods are employed to shrink the variance of $\beta$ estimates by inducing a penalty term. The first one is the LASSO penalty, but we find that LASSO shrinks many coefficients of the very rare variants to zero and loses power when the very rare variants are risk variants. Although the LASSO method holds promises to improve power to detect rare variants, the nature of LASSO can be too stringent when the allele frequencies of risk variants are extremely rare, which limit the merits under the performed simulations. The second popularly used regularization method is the Ridge regression [19], which takes $l_2$ regularized regression to accommodate variant correlations. Because we desire to shrink some coefficients of neutral variants to zero to reduce the data dimension, the ELASTICNET penalty, a mixture of LASSO and Ridge penalties, is considered.

The LD effect is considered in this study. Li and Leal [4] pointed out that it is reasonable to assume the independence of the variants within a locus. They further discussed that ‘if this assumption (of LD) is violated and two functional variants are on the same haplotype, the power is increased, because there is a higher probability of carrying more than one functional variant that increases the probability of an individual being a case.’ Our simulation results seem to support their statement. The power increase is also observed in all beta-determined weight methods in the presence of LD.

### Table 3. Power for LD effects under the significance level of 0.01

<table>
<thead>
<tr>
<th>$\beta$ model</th>
<th>LC</th>
<th>$S_\beta$</th>
<th>$S_\beta$.lasso</th>
<th>$S_\beta$.elasticnet</th>
<th>$FBAT$-v0</th>
<th>$FBAT$-v1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without LD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\psi_1$</td>
<td>0.0086</td>
<td>0.0082</td>
<td>0.0048</td>
<td>0.0046</td>
<td>0.0052</td>
<td>0.0066</td>
</tr>
<tr>
<td>$\psi_2/\psi_3$</td>
<td>0.6540</td>
<td>0.6860</td>
<td>0.6637</td>
<td>0.6627</td>
<td>0.1489</td>
<td>0.9020</td>
</tr>
<tr>
<td>$\psi_4$</td>
<td>0.6530</td>
<td>0.6650</td>
<td>0.6416</td>
<td>0.6456</td>
<td>0.1616</td>
<td>0.7680</td>
</tr>
<tr>
<td>$\psi_5$</td>
<td>0.7320</td>
<td>0.7420</td>
<td>0.7190</td>
<td>0.7160</td>
<td>0.2407</td>
<td>0.0090</td>
</tr>
<tr>
<td>With LD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\psi_1$</td>
<td>0.0066</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0053</td>
<td>0.0037</td>
<td>0.0114</td>
</tr>
<tr>
<td>$\psi_2/\psi_3$</td>
<td>0.7144</td>
<td>0.7566</td>
<td>0.7427</td>
<td>0.7459</td>
<td>0.2843</td>
<td>0.8982</td>
</tr>
<tr>
<td>$\psi_4$</td>
<td>0.6828</td>
<td>0.7142</td>
<td>0.7040</td>
<td>0.7108</td>
<td>0.2592</td>
<td>0.7918</td>
</tr>
<tr>
<td>$\psi_5$</td>
<td>0.6948</td>
<td>0.7470</td>
<td>0.7385</td>
<td>0.7389</td>
<td>0.2380</td>
<td>0.0080</td>
</tr>
</tbody>
</table>

The four combinations of four types of $\psi$ represent four different beta parameter settings, $\psi_1$, $\psi_2/\psi_3$, $\psi_4$, $\psi_5$, which define $\beta_0$ frequency-determined, equal, 90% positive and 10% negative $\beta$s, and 50% positive and 50% negative $\beta$s, respectively. The power is based on 1,000 replications.
To benefit potential users, we have implemented the mFBAT-LC, Sβ, Sβlasso, and Sβelasticnet methods in our betafam software coded by R language. Finally, we would like to note a limitation of the software we developed: only nuclear family data, parents with one or more offspring, are considered in the current version. This work will be furthered in the near future.

Web Resources

URL for the FBAT software is http://www.biostat.harvard.edu/_FBAT/FBAT.htm
URL for the R glmnet package is http://cran.rproject.org/web/packages/glmnet/index.html
The betafam package is submitted to the Comprehensive R Archive Network (CRAN): http://cran.r-project.org/

Acknowledgements

Before Dr. Guo moved to the Intramural Research Program at the National Institute of Mental Health, this work was partially supported by the National Natural Science Foundation of China (grant No. 10901031).

We thank 2 anonymous reviewers for their insightful comments and suggestions.

The views expressed in this presentation do not necessarily represent the views of the NIMH, NIH, HHS, or the United States Government.

References

8 Asimit J, Zeggini E: Rare variant association analysis methods for complex traits. Annu Rev Genet 2010;44:293–308.