Association Tests Using Common and Rare Variants and family data

Renfang Jiang[*] Jianping Dong[†] Yilin Dai[‡]

Abstract

Main method of dealing with rare variants in association testing is to collapse rare variants to form a super variant. Some problems remain. Collapsing many non-causal variants will introduce noise and reduce power of tests. Collapsing methods can be seriously impaired by misclassification of collapsing regions. Collapsing deleterious and protective variants together will also reduce power of tests. The classification of rare variants is subjective, if only rare variants are included in a study, some important genetic information may be left out because of this.

We propose a test using both common and rare variants. A forward selection method will be used to exclude non-causal variants in study. The selection is based on the correlation coefficient for each SNP with the trait. The proposed tests perform well in different scenarios.

We also propose a family based test, which uses genetic information from within-family variation and between -family variation. The test will not only avoid population stratification, but also increase power when population stratification is not severe.

Key Words: collapsing methods, rare variants, family data

1. Rare variants

Common SNPs can only explain a small proportion of the observed heritable variability. People usually think a SNP is rare if its minor allele frequency is less than 0.01. A commonly used method in dealing with rare variants is to collapse rare variants in a given region into one variant, the detection of the collapsed rare variants becomes easier. Many different ways of collapsing have been proposed. Some uses a indicator function on the rare variants in the region. It counts the number of people with at least one rare variants. Cohort allelic sum test (Morgenthaler and Thilly 2007) compares number of individuals with rare mutations between cases and controls. Some uses sum of rare variants instead of the indicator function. It counts the number of rare mutations in a region for each individual. Some puts weights on rare variants, and uses a weighted sum of rare variants. A choice of weights is allele frequencies. Combined multivariate collapsing (Li and Leal 2008) is a multivariate test with common variants and collapsed scores of rare variants. Weighted sum statistic (Madsen and Browning 2009) collapses both rare and common variants by adding different weights based on allele frequencies. Another choice of weights is odds ratios. In ORWSS (Feng, Elston and Zhu 2011) weights are calculated based on odds ratios. An implicit assumption of the collapsing of rare variants in a region is that these rare variants more or less are all causal variants. However, when this is not the case, collapsing many non-causal variants will introduce noise and reduce power of tests. Collapsing methods can be seriously impaired by misclassification of collapsing regions (Li and Leal 2008).

1.1 Methods

Collapsing deleterious and protective variants together will reduce power of tests. The classification of rare variants is subjective, if only rare variants are included in a study,

[*] Michigan Technological University
[†] Michigan Technological University
[‡] Michigan Technological University

some important genetic information may be left out because of this. Our goal is to develop a new test to dress these problems.

A forward selection method will be used to exclude non-causal variants in the study. The selection is based on the correlation coefficient for each SNP with the trait. A weighted sum approach in collapsing rare variants. The deleterious and protective components are separated by the correlation coefficients of SNPs and the trait. Step 1. Forward selection on common SNPs with sum collapsing. Step 2. Forward selection on rare SNPs with weighted sum collapsing. Step 3. Repeat step 2 for rare SNPs without bases from common SNPs of step 1. The results are denoted as S(+,both), S(-,both) and S(+,rare)-S(-,rare). Step 4. Let S(wSC) be the one among S(+,both), S(-, both), S(+, rare), and S(-,rare) with the largest correlation coefficient with the trait vector. Let S(wSCd) be the one of S(+,both)-S(-,both) and S(+,rare)-S(-,rare) with the largest correlation coefficient with the trait vector. The test statistics are constructed by using logistic regression model if the traits are qualitative; while a regression model will be used if the traits are quantitative. Finally, the p-value is calculated by permutation procedure. Two tests are proposed: BwSC (weighted selective collapsing) using S(wSC) and BwSCd using S(wSCd).

1.2 Simulation results

Data used in the simulation are generated following previous studies (Pan and Shen 2011 ,Wang and Elston 2007). The target region contains four observed common SNPs and an unobserved common SNP. It also contains 28 observed rare SNPs, and 8 of them are randomly chosen as causal rare SNPs. Allele frequencies of common SNPs are randomly chosen between 0.1 and 0.3; allele frequencies of rare SNPs are randomly chosen between 0.001 and 0.005. The covariance between observed common SNPs is 0.4, and the covariance between observed common SNP and the unobserved common SNP is 04a, where a=1 or -1 with equal chance. Covariance between rare SNPs Zi and Zj is $0 : 4^{|i-j|}, 1 \leq i, j \leq 28$. Five hundred cases and five hundred controls are simulated with one thousand replicates. The significant level is 0.05 for all scenarios. Type I error rates are correct in all simulations. The powers are shown in tables 1-4. The first letter in the names of the tests is either B or R, B stands for using both common and rare SNPs, R stands for using rare SNPs only. After the first letter, the lower case letters describe the ways of collapsing. For example, ind means collapsing using indicator function, sum means collapsing using sum function, wsum means collapsing using weighted sum function, wor means collapsing using weighted sum function with odds ratios as weights, and w means our weighted sum function. After that SC means selective collapsing. Two more tests are added in tables 2 and 4. They are Cs and Cm, Cs is the single marker test for common SNPs with Bonferroni correction, and Cm is multiple marker test for common SNPs. The proposed tests are BwSC and BwSCd.

2. Family data

Although many disease-associated common variants have been discovered through genome-wide association studies, much of the genetic effects of complex diseases have not been explained. Population-based association studies are vulnerable to population stratification. A possible solution is to use family-based tests. However, if tests only estimate the genetic effect from the within-family variation to avoid population stratification, they may ignore the useful genetic information from between-family variation and lose power.

Table 1: No common SNPs effect, the effects of rare SNPs are in the same direction

odds ratio	1.3	1.6	1.9	2.2	2.5	2.8	3.1
BwSC	0.316	0.509	0.654	0.775	0.892	0.927	0.970
BwSCd	0.201	0.340	0.445	0.586	0.734	0.825	0.885
Rind	0.227	0.376	0.522	0.630	0.737	0.810	0.851
Rsum	0.245	0.424	0.570	0.670	0.778	0.846	0.888
Bind	0.129	0.204	0.318	0.419	0.522	0.623	0.698
Bsum	0.147	0.243	0.343	0.470	0.565	0.674	0.751
RindSC	0.295	0.420	0.589	0.726	0.834	0.884	0.954
RsumSC	0.298	0.425	0.588	0.731	0.834	0.894	0.946
Bwsum	0.302	0.474	0.631	0.710	0.810	0.875	0.931
Bwor	0.090	0.170	0.226	0.295	0.416	0.408	0.580

Table 2: weak common SNPs effect, the effects of rare SNPs are in the same direction

odds ratio	1.3	1.6	1.9	2.2	2.5	2.8	3.1
BwSC	0.344	0.538	0.631	0.778	0.850	0.935	0.954
BwSCd	0.210	0.395	0.484	0.625	0.661	0.822	0.848
Rind	0.237	0.394	0.472	0.600	0.715	0.785	0.843
Rsum	0.247	0.418	0.543	0.636	0.747	0.811	0.869
Bind	0.278	0.364	0.436	0.517	0.618	0.677	0.760
Bsum	0.298	0.384	0.461	0.562	0.668	0.735	0.795
RindSC	0.236	0.430	0.565	0.702	0.781	0.888	0.910
RsumSC	0.238	0.446	0.605	0.705	0.815	0.892	0.920
Bwsum	0.341	0.534	0.658	0.703	0.846	0.870	0.911
Bwor	0.253	0.312	0.344	0.475	0.456	0.582	0.648
Cs	0.163	0.157	0.144	0.164	0.174	0.191	0.193
Cm	0.195	0.199	0.193	0.207	0.212	0.228	0.238

Table 3: No common SNPs effect, the effects of rare SNPs are in different directions

odds ratio	1.3	1.6	1.9	2.2	2.5	2.8	3.1
BwSC	0.135	0.148	0.200	0.227	0.297	0.373	0.465
BwSCd	0.134	0.197	0.250	0.340	0.391	0.441	0.558
Rind	0.062	0.058	0.089	0.095	0.118	0.129	0.164
Rsum	0.054	0.062	0.092	0.083	0.113	0.118	0.158
Bind	0.062	0.060	0.059	0.074	0.085	0.010	0.128
Bsum	0.062	0.059	0.065	0.073	0.090	0.101	0.117
RindSC	0.090	0.150	0.214	0.221	0.314	0.352	0.395
RsumSC	0.094	0.151	0.202	0.210	0.335	0.353	0.449
Bwsum	0.107	0.096	0.096	0.136	0.179	0.221	0.270
Bwor	0.090	0.126	0.133	0.165	0.211	0.222	0.255

Table 4: weak common SNPs effect, the effects of rare SNPs are in different directions

odds ratio	1.3	1.6	1.9	2.2	2.5	2.8	3.1
BwSC	0.133	0.182	0.256	0.332	0.357	0.479	0.480
BwSCd	0.190	0.217	0.308	0.386	0.468	0.568	0.548
Rind	0.045	0.077	0.068	0.103	0.115	0.120	0.157
Rsum	0.054	0.074	0.062	0.091	0.109	0.126	0.154
Bind	0.200	0.184	0.200	0.198	0.244	0.255	0.233
Bsum	0.190	0.182	0.200	0.197	0.243	0.226	0.229
RindSC	0.068	0.122	0.176	0.241	0.270	0.359	0.387
RsumSC	0.094	0.119	0.193	0.254	0.273	0.371	0.390
Bwsum	0.100	0.114	0.164	0.172	0.193	0.236	0.272
Bwor	0.201	0.245	0.260	0.311	0.334	0.398	0.405
Cs	0.156	0.131	0.155	0.139	0.186	0.149	0.146
Cm	0.211	0.185	0.214	0.192	0.221	0.211	0.190

2.1 Methods

In family-based association studies, FBAT, a general unified approach, has been proposed to permit any type of genetic models, a general family design, different phenotypes and multiple markers (Laird, Horvath, Xu 2000). Family-based tests are generally robust to population stratification and those tests can avoid any population bias in other standard designs. Recently, the multi-marker test $FBAT_{MM}$ (Rakovski et al 2007), which is similar to the Hotelling T^2 test, has been proposed for family-based studies. Another multi-marker test $FBAT_{LC}$ (Xu et al 2006) linearly combines single-marker test statistics using data-driven weights derived by conditional mean model (Lange et al 2003). The weights are least square estimates of genetic effects. The data-driven weights are regarded as fixed for FBAT. These two methods have been implemented in the program FBAT, which has been widely used in family-based association studies. The data-driven weights in $FBAT_{LC}$ are the estimates of genetic effect considering between-family variation. It is a biased estimator and is sensitive to population structure. We investigate the data-driven weights used in $FBAT_{LC}$ and provide a new methodology to analyze the multiple correlated markers for family-based association studies. We use $FBAT_{WS}$ to denote the new test. It is based on weighted sum of two association tests. One of which estimates the genetic effect from both within-family and between-family variation and the other is from within-family variation only. The weights are computed automatically based on a measure of the population stratification strength in family data. The proposed method can capture more important information from multiple loci in the family data while maintaining robustness to population stratification. Due to population stratification and linkage disequilibrium which cause a bias for the estimate, a permutation procedure is employed conditional on the traits, parental genotypes, and haplotypes.

The general idea of FBAT (Laird, Horvath, Xu 2000) is to regard the offspring genotype as random conditional on the traits and parental genotypes. The test statistic is computed from the distribution of offspring genotype under the null hypothesis. Let T_{ij} denote the coded trait for the jth offspring in the ith family and X_{ijk} denote the coded genotype score for the kth marker of the jth offspring in the ith family, where $i = 1, \ldots, M$, $j = 1, \ldots, N$, and $k = 1, \ldots, K$. Following the standardized FBAT (Laird, Horvath, Xu 2000), let

$$U_{ik} = \sum_j T_{ij}(X_{ijk} - E(X_{ijk})),$$

$$V_{ik} = var(U_{ik}) = \sum_j \sum_l T_{ij} T_{il} cov(X_{ijk}, X_{ilk}).$$

With a large number of families, FBAT statistic for the kth marker:

$$Z_k = \frac{\sum_i U_{ik}}{\sqrt{\sum_i V_{ik}}}$$

is approximately $N(0, 1)$.

Another approach to the multi-marker family-based association testing is to linearly combine single-marker test statistics using data-driven weights ($FBAT_{LC}$) (Xu et al 2006). Conditional on the traits and parental genotypes, the weights can be derived by the conditional mean model of trait T for the kth marker as follows:

$$E(T_{ij}) = \alpha_k + \beta_k f(X_{ijk})$$

where $f(X_{ijk}) = E(X_{ijk})$ for offspring in the informative families and $f(X_{ijk}) = X_{ijk}$ for the others (include offspring in the non-informative families and all parents). Let $w = (w_1, \dots, w_k)$ where $w_k = \hat{\beta}_k / SE(\hat{\beta}_k)$ is the standardized least square estimator of β_k. Then the multi-marker $FBAT_{LC}$ test statistic:

$$FBAT_{LC} = \frac{w^T Z}{\sqrt{w^T \Sigma w}}$$

is approximately $N(0, 1)$, where $Z = (Z_1, \dots, Z_k)^T$ is the vector of single FBAT test statistic and Σ can be derived from the conditional pairwise haplotype distribution in offspring or from the empirical estimator of the covariance matrix (Rakovski et al 2007).

Although the data-driven weights are independent of Z under H_0 because the FBAT test is computed conditional on traits and on parental genotypes, the power of $FBAT_{LC}$ will be highly dependent on the estimate of the optimal weights. In the conditional mean model, the weights are estimates of genetic effects using population data, which can be regarded as estimates of the genetic effects using between-family variation. It has been shown that this estimator is biased unless there is no population stratification. Intuitively, the more accurate the estimate is, the closer the weights to the optimal weights, and the more power the test can gain. However it will lose power if the effect of population stratification is significant. Thus, we proposed a new multi-marker test $FBAT_{WS}$ using adaptive weights to combine two test statistics based on the estimate of the existing population stratification.

The strength of population stratification will be measured by

$$v = \frac{1}{k} \sum_k \frac{D_k - E(D_k)}{SD(D_k)}$$

where $D_k = |Z_k - w_k|$ for $k = 1, \dots, K$. Then the test statistic can be written as:

$$FBAT_{WS} = \frac{1}{1+v} w^T Z + \frac{v}{1+v} Z^T Z$$

Under the null hypothesis: no genetic effect and no population stratification, Z_k and w_k are independent standard normal random variables. Therefore, D_k is a folded normal random variable with $E(D_k) = 2/\sqrt{\pi}$ and $Var(D_k) = 2 - 4/\pi$. It is clear that the strength of population stratification increases as D_K increases. When population stratification is strong, $FBAT_{WS}$ will automatically put more weight on the second term to maintain robustness against spurious positives. On the other hand, when the effect of population stratification is relatively weak, $FBAT_{WS}$ will automatically put more weight on the first term to make use of both sources of genetic variation: between-family and within-family. In latter case, the degrees of freedom of the test will be reduced, and power of the test will be increased. Because LD structure will be maintained in the permutation procedure, in order to improve the computational efficiency, $FBAT_{WS}$ does not consider LD structures. The second term $Z^T Z$ can be written as:

$$Z^T Z = U^T diag(V)^{-1} U$$

where $U = (\sum_i U_{i1}, \ldots, \sum_i U_{ik})$ and $V = (v_{k_1 k_2})$. This is an empirical estimator of the covariance matrix Σ, where

$$v_{k_1 k_2} = \sum_i (\sum_j T_{ij}[X_{ijk_1} - E(X_{ijk_1})] \sum_j T_{ij}[X_{ijk_2} - E(X_{ijk_2})]).$$

Therefore, the second term $Z^T Z$ is one of the asymptotic tests in (Pan 2009), which has been proposed recently to gain more power under strong LD structures. When the parental haplotypes are known, a permutation procedure will be employed to compute the p-value of $FBAT_{WS}$. For each child with fixed trait in any family, each parental haplotype is transmitted to the child with equal probability, so that, for any given parental hypostyles, there are four different permutations of the data. When the parental haplotypes are unknown, inferring haplotype is needed. There are several methods to infer haplotypes. For example, Thunder (Li et al 2010), Beagle (Browning, Browning 2009), Impute2 (Montgomery et al 2013), and SNPtools (Wang et al 2013). Haplotype can also be inferred by using sequencing reads (Delaneau et al 2013).

2.2 Simulation results

In the simulation study, we compare the power of the proposed test $FBAT_{WS}$ with the following three FBAT tests: (1) the single-marker test with Bonferroni multiple testing adjustment $FBAT_B$, the Bonferroni adjusted p-value $P_{adj} = 1-(1-P_{min})^K$, where P_{min} is the minimal p-value among the single-marker tests (2) the multi-marker test $FBAT_{MM}$ (Rakovski, Xu, Lazarus, Blacker, Laird 2007), which is similar to the Hotelling T^2 test, (3) the multi-marker test $FBAT_{LC}$ (Xu, Rakovski, Xu, Laird 2006) that linearly combines the single-marker test statistics using data-driven weights.

One goal of the simulation study is to examine whether the proposed multi-marker test is robust to the underlying LD structure. We consider six different LD structures and assume additive genetic effect.

Next, our simulation study will be based on real LD structure. We download the Haplotypes data from 170 unrelated samples of JPT+CHB (Japanese in Tokyo, Japan + Han Chinese in Beijing, China) in the HapMap3 Phased Haplotypes. We consider three genes CHI3L2 (in the region of 15.78kb), CTLA4 (in the region of 10kb) and IL21R (in the region of 47.69kb), which have also been analyzed in other simulation studies (Chapman, Whittaker 2008, Jiang, Dong, Dai 2009, Wang, Abbott 2008, Wang, Elston 2007). Their LD pattern can be visualized on the HapMap site. We perform the simulation study using SNPs with minor allele frequency (MAF) > 0.01, and we remove the redundant SNPs that are perfectly correlated with other SNPs. We have 12 SNPs left for CHI3L2, seven SNPS for CTLA4 and 10 SNPs for IL21R. We calculate haplotype frequencies from the samples of each gene and generate the parents of each family based on the known haplotype frequencies. The disease marker is randomly chosen as unobserved SNP. Other SNPs are

Table 5: Type I error rates for four FBAT tests in simulated structures

LD structures	LD=L1	LD=L2	LD=L3	LD=L4	LD=L5	LD=L6
B	0.047	0.036	0.051	0.042	0.052	0.039
MM	0.047	0.045	0.068	0.054	0.057	0.050
LC	0.050	0.057	0.058	0.045	0.055	0.047
WS	0.052	0.052	0.059	0.038	0.052	0.048

observed as haplotype data and the quantitative phenotypes of offspring in each family are generated from a quantitative phenotype model. Two scenarios (500 trios under one population and two populations) are considered in the simulation study with 1000 simulation replicates and a significance level of 0.05. To generate quantitative phenotypes for samples from one population, let $\mu_p = 0$; for samples from two distinct populations, let μ_p be 0.5 or -0.5.

Type I error rate for the case of six mimicked LD structures is shown in Table 1. All tests have a correct Type I error rate. It is expected that the proposed method will have a correct Type I error rates due to the permutation procedure. The result of power comparison is shown in Figure 2.

Four FBAT tests are considered for power comparisons with six different LD structures. The unobserved casual SNP has an equal chance to be positively or negatively correlated to those observed SNPs in all scenarios. In Figure 2, $FBAT_B$ (B), $FBAT_{MM}$ (MM), $FBAT_{LC}$ (LC), and $FBAT_{WS}$ (WS) are indicated by the blue dot-dashed line, the green dotted line, the red dash line, and the black solid line, respectively. In the first simulation study, the goal is to compare the performance of the proposed method with other FBAT methods. We fix the window size for each scenario and assume the sample come from the same population. An examination of the results show that $FBAT_{WS}$ has a consistently higher power in all cases, followed by $FBAT_{LC}$, $FBAT_{MM}$, and $FBAT_B$. $FBAT_B$ is considered as the most conservative test in this study, because the independent assumption is violated. $FBAT_{MM}$ improves the power by considering the variance-covariance matrix. On the other hand, it also suffer from the relatively high degrees of freedom, especially when the region under consideration is large. $FBAT_{LC}$ with one degrees of freedom improves the power by using the optimal weights to combine single-marker tests and overcomes the degrees of freedom problem raised by $FBAT_{MM}$. In a genetic region with strong LD, we do not have any clue of how the underlying casual marker is related to the observed SNPs. The optimal weights in $FBAT_{LC}$ are biased estimates of genetic effects (Abecasis, Cardon, Cookson 2000). Therefore, using incorrect estimation of genetic effect as weights in $FBAT_{LC}$ will lose some power. $FBAT_{WS}$ improve the power by not only considering the optimal weights to combine single-marker tests like $FBAT_{LC}$, but also automatically adjusting the weights based on the estimate of the genetic effect from between-family variants and within-family variants.

Type I error rates for the simulated HapMap data on CHI3L2, IL21R, and CTLA4 are given in Table 2. Type I error rate of all tests are well controlled under 0.05 level of significance. We also found that $FBAT_B$ has a lower type 1 error rate than other tests, because the strong LD structure existed in all three regions.

The results of power comparison in one population and two populations are shown in Figures 3 and 4. The underlying casual marker is randomly selected each time, which make the LD structures relatively complicated in these scenarios.

Four FBAT tests are considered for power comparisons under different LD structures

Table 6: Type I error rates of four FBAT tests using HapMap data, $*$ denote the case in the mixed populations of two.

LD structures	CHI3L2	CTLA4	IL21R	CHI3L2$*$	CTLA4$*$	IL21R$*$
B	0.023	0.024	0.027	0.029	0.026	0.034
MM	0.049	0.036	0.041	0.051	0.040	0.042
LC	0.044	0.035	0.042	0.045	0.050	0.039
WS	0.040	0.037	0.037	0.037	0.041	0.054

of three genes CHI3L2 (in the region of 15.78kb), CTLA4 (in the region of 10kb) and IL21R (in the region of 47.69kb). The unobserved casual SNP is randomly selected in all scenarios. In Figures 3 and 4, $FBAT_B$ (B), $FBAT_{MM}$ (MM), $FBAT_{LC}$ (LC), and $FBAT_{WS}$ (WS) are denoted by the blue dot-dashed line, the green dotted line, the red dash line, and the black solid line, respectively.

We consider all samples from one population first. $FBAT_{WS}$ has a relatively high power in most scenarios. For gene CHI3L2, where SNPs are dense and highly correlated with each other, $FBAT_{WS}$ is the most powerful test, followed by $FBAT_{LC}$, $FBAT_{MM}$ and $FBAT_B$ when the heritability is relatively low. As heritability increases, $FBAT_{MM}$ achieves the highest power and $FBAT_{WS}$ is the second among all tests. This implies $FBAT_{WS}$ is more sensitive to the genetic effect with low heritability. $FBAT_{MM}$ is adept to deal with genetic region with strong LD and high heritability. For the gene CTLA4, where the number of markers is relatively small and LD pattern is relatively weak, $FBAT_{WS}$ is again the most powerful test, followed by $FBAT_{LC}$, $FBAT_B$ and $FBAT_{MM}$. For the gene IL21R, where SNPs are loose and LD pattern is relatively weak, $FBAT_{WS}$ is the most powerful test, followed by $FBAT_B$, $FBAT_{LC}$, and $FBAT_{MM}$. For genetic region with weak LD like CTLA4 and IL21R, $FBAT_{MM}$ lose its potential power due to the issue of degrees of freedom. In all scenarios of two populations, the results are similar that $FBAT_{WS}$ is the most powerful test except for simulated data based on gene CTLA4 with high heritability. In practice, most undiscovered genetic variants have low heritability. The power of tests depends on the LD patter. In general, $FBAT_{WS}$ automatically adjusted the weights to combine the estimates of genetic effect from various source of genetic variants, therefore is a powerful test for family-based association studies. It is robust to population stratification and the underlying LD structure. Our simulated results demonstrate that $FBAT_{WS}$ is a potentially powerful test among multi-marker tests.

REFERENCES

Abecasis GR, Cardon LR, Cookson WOC (2000), “A general test of association for quantitative traits in nuclear families”. *American Journal of Human Genetics*, 66(1):279-292.

Browning BI, Browning SR (2009),“A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals”. *American Journal of Human Genetics*, 84:210-223.

Chapman J, Whittaker J (2008), “Analysis of multiple SNPs in a candidate gene or region”. *Genetic Epidemiology*, 32(6):560-566.

Cheung CY, Thompson EA, Wijsman EM (2013), “GIGI: an approach to effective imputation of dense genotypes on large pedigrees”. *American Journal of Human Genetics*, 92(4): 504-516.

Cobat A, Abel L, Alcais, A, Schurr E (2014), “A general efficient and flexible approach for genome-wide association analyses of imputed genotypes in family-based designs”. *Genetic Epidemiology*, 38(6): 560-571.

Dai Y, Jiang R, Dong J (2012), “Weighted selective collapsing strategy for detecting rare and common variants in genetic association study,” *BMC Genetics* 13:7

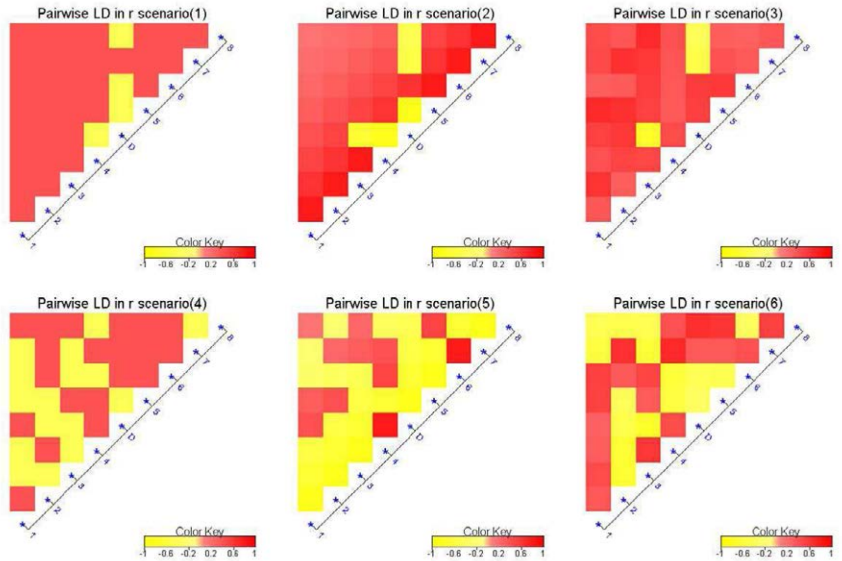

Figure 1: LD Structures for Simulation.

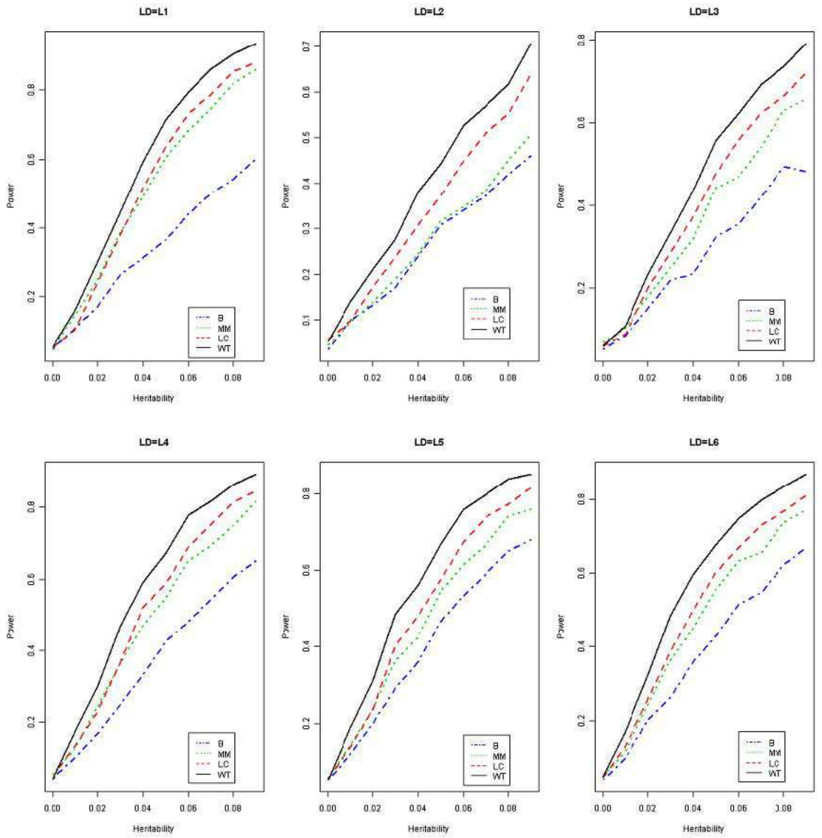

Figure 2: Power comparisons using simulated data.

Dai Y, Guo L, Dong J, Jiang R (2011), “Improved power by collapsing rare and common variants based on a data-adaptive forward selection strategy,” *BMC Proceedings*, 5(Suppl 9):(S114).

Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J (2013), “Haplotype estimation using sequencing reads”. *Genetic Epidemiology*, 93:687-696.

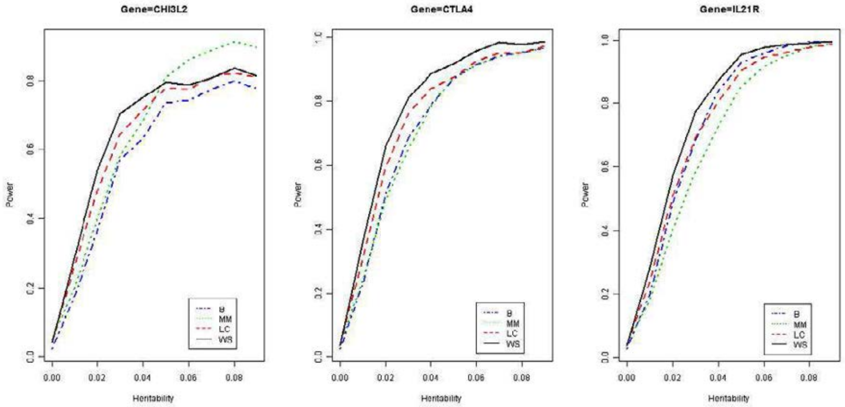

Figure 3: Power comparisons using HapMap data.

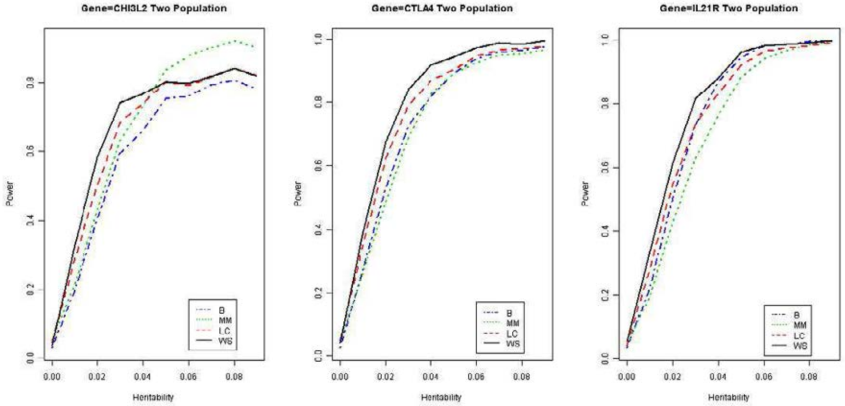

Figure 4: Power comparisons using HapMap data.

Feng T, Elston RC, Zhu X (2011),“ Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS),” *Genet Epidemiol*, 35(5):398-409.

He Z, O'Roak B, Smith JD, Wang G, Hooker S, Santos-Cortez RLP, Li B, Kan M, Krumm N, Nickerson DA, Shendure J, Eichler EE, Leal SM (2014), “Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data”. *American Journal of Human Genetics*, 94:33-46.

Jiang RF, Dong JP, Dai YL (2009), “Improving power in genetic-association studies via wavelet transformation”. *MC Genetics* 10:53.

Jiang Y, Satten GA, Han Y, Epstein MP, Heinzen EL, Goldstein DB, Allen AS (2014), “Utilizing population controls in rare-variant case-parent association tests”. *American Journal of Human Genetics*; 94:845-853.

Laird NM, Horvath S, Xu X (2000), “Implementing a unified approach to family-based tests of association”. *Genetic Epidemiology*; 19:S36-S42.

Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM (2003), “Using the noninformative families in family-based association tests: A powerful new testing strategy”. *American Journal of Human Genetics*;73(4):801-811.

Lee S, Abecasis GR, Boehnke M, Lin X (2014), “Rare-variant association study designs and statistical tests”. *American Journal of Human Genetics*, 95:5-23.

Li BS, Leal SM (2008), “Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data,” *Am J Hum Genet*, 83(3):311-321.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010), “MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes”. *Genetic Epidemiology*, 34:816-834.

Madsen BE, Browning SR (2009), “A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic,” *Plos Genet* 2009., 5(2)

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al(2013), “1000 Genomes project consortium. (2013). The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes". *Genome Res.* 23:749-761.

Morgenthaler S, Thilly WG (2007)," A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST)," *Mutat Res-Fund Mol M 2007*, 615(1-2):28-56.

Pan W (2009), "Asymptotic Tests of Association with Multiple SNPs in Linkage Disequilibrium". *Genetic Epidemiology*;33(6):497-507.

Pan W, Shen X (2011), "Adaptive tests for association analysis of rare variants," *Genet Epidemiol*, 35(5):381-388.

Rakovski CS, Xu X, Lazarus R, Blacker D, Laird NM (2007), "A new multimarker test for family-based association studies". *Genetic Epidemiology*;31(1):9-17.

Saad M, Wijsman E (2013), "Power of family-based association designs to detect rare variants in large pedigrees using imputed genotypes". *Genetic Epidemiology*, 38(1):1-9.

Wang K, Abbott D (2008), "A principal components regression approach to multilocus genetic association studies". *Genetic Epidemiology*; 32(2):108-118.

Wang T, Elston RC (2007), "Improved power by use of a weighted score test for linkage disequilibrium mapping". *American Journal of Human Genetics*, 80(2):353-360.

Wang X, Lee S, Zhu X, Redline S, Lin X (2013), "GEE-based SNP set association test for continuous and discrete traits in family-based association studies", *Genetic Epidemiology*, 37(8):778-786.

Wang Y, Lu J, Yu J, Gibbs RA, Yu F (2013), "An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data". *Genome Res.*, 23:833-842.

Xu X, Rakovski C, Xu XP, Laird N (2006), "An efficient family-based association test using multiple markers". *Genetic Epidemiology*, 30(7):620-626.

Yu Z (2012), "Family-based association tests using genotype data with uncertainty". *Biostatiistics*, 13(2):228-240.