Testing Genetic Effects of Rare and Common Variants Together in Association Studies

Renfang Jiang* Jianping Dong† Yilin Dai‡

Abstract

We develop a robust strategy to analyze association between complex traits and genetic regions using both common and rare variants. Our goals are to treat common and rare variants separately, to avoid including non-causal variants in the study, and to distinguish the effects of deleterious and protective variants. Simulation results showed that the test has high power in different situations.

Key Words: rare variants, genetic association study, common variants, collapsing, common diseases

1. Background

In past decades many genetic variants have been discovered for being associated with some common diseases in large scale genome wide association studies. Most of these genetic variants are common SNPs. However, these common SNPs can only explain a small proportion of the observed heritable variability. Following the development of the new generation sequencing techniques, a new hypothesis about the cause of common diseases have been emerged: many common diseases are caused by many rare genetic variants. People usually think a SNP is rare if its minor allele frequency is less than 0.01. Since they are rare, it is often difficult to detect a single rare SNP using small or medium size of samples. Many methods have been proposed to detect rare variants. One of these methods 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). Moreover, 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. Our goal is to develop a new test to dress these problems.

*Michigan Technological University, Houghton, MI 49931
†Michigan Technological University, Houghton, MI 49931
‡Michigan Technological University, Houghton, MI 49931

2. Method

In order to partially eliminate the subjectiveness of the definition of rare variants, we will use 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. We take a weighted sum approach in collapsing rare variants. A weight for each rare SNP is calculated based on an empirical estimate of the probability that an individual with the mutation will have the disease. The deleterious and protective components are separated by the correlation coefficients of SNPs and the trait. Our main goal is to detect disease associated rare variants. Common SNPs are included just in case some genetic information is contained in these common SNPs, and common SNPs are not our main target. When we include common SNPs, we do not want to increase the degrees of freedom and noise/signal ratio. To achieve this, we make two collapsings, one is for all selected common and rare SNPs, the other is for rare SNPs only. We pick one from the two according to the correlation coefficients of them with the trait. The deleterious and protective components are treated the same way.

The detailed procedure is the following.

Step 1. Forward selection on common SNPs with sum collapsing.

Let S(+) be the deleterious component, and let S(-) be the protective component. The initial value of S(+) and S(-) are zero vectors. If we can find a common SNP such that when the genotype vector of the SNP is added to S(+), the correlation coefficient of S(+) and the trait vector is the largest, we add that genotype vector to S(+). Similarly, if we can find a common SNP such that when the genotype vector of the SNP is added to S(-), the correlation coefficient of S(-) and the trait vector is the smallest, we add genotype vector of that SNP to S(-). Repeat the above procedure until all common SNPs are collapsed or improvement is too small.

Step 2. Forward selection on rare SNPs with weighted sum collapsing.

First we calculate weights for rare SNPs. Let q_k be the proportion of people having both disease and the rare mutation at the kth SNP among the people having the rare mutation at the kth SNP. Let $p_k = |q_k - 0.5|$. The weight of the kth SNP is $w_k = p_k K / \sum_j p_j$, where K is the number of rare SNPs. Note that q_k is an empirical estimate of the probability that a individual with the mutation at the kth SNP will have the disease. We then repeat Step 1 for rare SNPs and collapsing them with their weights. The results are denoted as S(+, both) and S(-,both).

Step 3. Repeat step 2 for rare SNPs without bases from common SNPs of step 1, and the results are S(+, rare) and 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).

3. Simulation results

Data 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;

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

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 Z_i and Z_j 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. We compare our tests with other eight tests. The names of these tests are defined in the following ways. 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.

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

4. Discussions and conclusions

In table 1, there is no common SNPs effects. Although the proposed tests using both common and rare SNPs, they still outperform the tests using only rare SNPs. The reason is the selective collapsing. Two tests with selective collapsing are close but not as good as the proposed tests because the proposed weight function is better than theirs. In tables 3 and 4, the proposed tests are much better than others because the proposed tests are the only ones separating deleterious and protective variants. In tables 2 and 4, both single marker test Cs and the multiple marker tests Cm for common SNPs perform poorly because the information from common SNPs along is not strong enough. In conclusion, the proposed tests perform well in all the above scenarios no matter common SNPs have an effect or not, and no matter there is protective SNPs or not.

REFERENCES

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.

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.

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

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.

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

Wang T, Elston RC (2007), "Improved power by use of a weighted score test for linkadge disequilibrium mapping, *Am J Hum Genet*, 80(2):353-360.

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).

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