Published in final edited form as:

Stat Med. 2014 May 20; 33(11): 1816–1827. doi:10.1002/sim.6068.

A Powerful Association Test of Multiple Genetic Variants Using a Random-Effects Model

K. F. Cheng^{1,*}, J. Y. Lee², W. Zheng³, and Chun Li^{4,5}

¹Biostatistics Center and Department of Public Health, Taipei Medical University, Taiwan

²Graduate Institute of Statistics, National Central University, Chungli, Taiwan

³Vanderbilt Epidemiology Center, Department of Medicine, Vanderbilt University, Nashville, TN, USA

⁴Center for Human Genetics Research, Vanderbilt University, Nashville, TN, USA

⁵Department of Biostatistics, Vanderbilt University, Nashville, TN, USA

Abstract

There is an emerging interest in sequencing-based association studies of multiple rare variants. Most association tests suggested in the literature involve collapsing rare variants with or without weighting. Recently, a variance-component score test, SKAT, was proposed to address the limitations of collapsing method. Although SKAT was shown to outperform most of the alternative tests, its applications and power might be restricted and influenced by missing genotypes. In this paper, we suggest a new method based on testing whether the fraction of causal variants in a region is zero. The new association test, T_{REM} , is derived from a random-effects model, allows for missing genotypes and the choice of weighting function is not required when common and rare variants are analyzed simultaneously. We performed simulations to study the type I error rates and power of four competing tests under various conditions on the sample size, genotype missing rate, variant frequency, effect directionality, and the number of non-causal rare variant and/or causal common variant. The simulation results showed that T_{REM} was a valid test and less sensitive to the inclusion of non-causal rare variants and/or low effect common variants, or to the presence of missing genotypes. When the effects were more consistent in the same direction, T_{REM} also had better power performance. Finally, an application to the Shanghai breast cancer study showed that rare causal variants at the FGFR2 gene were detected by T_{REM} and SKAT, but T_{REM} produced more consistent results for different sets of rare and common variants.

Keywords

association test; random-effects model; rare variant; sequencing-based study

*Correspondence to: Professor K.F. Cheng, Biostatistics Center, Taipei Medical University, Taipei, Taiwan. kfcheng@tmu.edu.tw.

Web Resources

The software T_{REM} is freely available at <http://www2.cmu.edu.tw/~biostat/downloads/KFCpro/KFCprogram.htm>

1. Introduction

Common disease-common variant (CDCV) is a popular hypothesis in human genetics. It states that genetic susceptibility to common diseases is mainly due to alleles with moderate frequencies in the population. Another competing hypothesis is the common disease-rare variant (CDRV) hypothesis. It states that there is extreme allelic and locus heterogeneity and multiple rare variants underlie susceptibility to complex diseases [1,2]. Many complex diseases have been studied using a genome-wide association approach, which focused mainly on very common single nucleotide variants (SNVs) with minor allele frequency (MAF) >5%. Unfortunately, many associated SNVs identified by genome-wide association studies (GWASs) have very small effects and explain only a small fraction of disease heritability [3]. Recently, Cirulli and Goldstein [4] evaluated the evidence for an important role of rare variants of major effects in common diseases and suggested the hypothesis that multiple rare gene variants, each with moderate to high penetrance, could be a primary player for common diseases [5–7]. Dickson et al. [8], on the other hand, used computer simulations to show that rare genetic variants can easily create a phenomenon called synthetic association that is credited to common variants. If synthetic association is true, it may render targeted resequencing efforts based on the CDCV theory to discover causal variants unsuccessful [9–13], because the disease is caused by multiple strong-effect variants each carried by only a few people. The inadequacy of CDCV hypothesis was also pointed out in a recent paper by McClellan and King [14].

Although it is still unclear how exactly rare and common variants affect complex diseases, some results have shown that the combination of GWAS and sequencing could be a good technique for studying diseases [15]. Bansal et al. [16] pointed out that many sequencing studies, focusing on specific genes, have already shown collections of rare variants can indeed be associated with particular diseases. In this process, one powerful approach to identifying rare variants is to perform a composite test on variants grouped by genes or pathways. This has motivated several variant collapsing methods [17–22] to enrich association signals and at the same time reduce degrees of freedom.

The concept of collapsing methods is based on grouping the frequency counts of the variants in a gene or region and comparing the information between the cases and controls. Among these methods, the combined multivariate and collapsing (CMC) [17] and weighted sum (WS) [18] tests have received the most attention. The CMC test is defined by collapsing variants in subgroups according to allele frequencies and treating them as common variants which are then analyzed together with other common variants using multivariate analysis, such as Hotelling's T-squared test. In WS method, the individual is scored by a weighted sum of mutation counts of all variants. The WS statistic is the standardized rank sum of the scores for the affected individuals. These two methods were shown in simulations to have far better power performance than the usual single- and multi-marker tests [17–18]. In particular, the power of these tests increased as the number of causal variants increased while single- and multi-marker tests can have an inverse relationship. However, collapsing SNVs can attenuate the effect of causal variants when non-causal variants are included in the analysis, and can result in scores that are dominated by non-causal variants if the majority of variants included in analysis are non-causal. As a result, the power of these tests

can be dramatically reduced when many non-causal rare variants are included or when the range of variant frequency is wide with varied levels of effect. Many alternative methods are variations of these methods [19–22]. To address these limitations, a sequence kernel association test (SKAT) was recently proposed for testing association while easily adjusting for covariates [23]. In SKAT, all variant effects are considered as random with zero means and weighted variances. It is the usual variance-component score test based on a mixed model. However, to compute SKAT statistic (and Hotelling's statistic in the CMC method) the genotype data are required to be complete, rendering it inefficient as some variants or samples have to be excluded from the analysis. The SKAT using the imputed data based on the observed allele frequencies and Hardy-Weinberg equilibrium has been suggested. This simple approach can control type I error rate when covariates are not involved, but its power loss can be substantial.

We propose a new approach to detecting association signals for a group of variants. Our method does not compare genetic information directly between the case and control groups. Instead, we test whether the fraction of causal variants equals zero. We demonstrate that this approach is more powerful than most of the well known rare-variant association tests and more robust to the inclusion of non-causal variants or high frequency but low effect variants, and to the presence of missing genotypes. In addition, if variants have different effect directions (that is, some variants are protective and others deleterious) but more consistent in one direction, T_{REM} also outperforms other competing tests. These results were based on the analysis of extensive simulations given in our study. We also applied our test and the competing tests to the Shanghai breast cancer study and showed our test and SKAT were able to detect association signals from variants at *FGFR2* gene while the CMC and WS tests failed to detect any association signals. Detailed analysis for various sets of common and rare variants, however, showed that T_{REM} produced more consistent results in the sense that the range of p-values was smaller. This indicates that comparing with other competing tests, the new test is indeed more robust.

2. Methods

Suppose we have genotypes at K SNVs in a gene region. At the k -th ($k = 1, \dots, K$) variant site, the minor allele is denoted as a_k and the major allele as A_k . We assume that N_{1k} cases with genotype data at the k -th variant are observed, and among them, there are X_{1k} cases each carrying at least one copy of a_k . The corresponding numbers in the controls are denoted by N_{0k} and X_{0k} . Let p_{1k} (or p_{0k}) be the probability of a case (or control) subject carrying at least one copy of a_k . Using these probabilities, the effect of the k -th variant on the disease outcome can be measured by $\theta_k = p_{1k} / (p_{0k} + p_{1k})$. If the k -th variant increases the disease risk, then $p_{1k} > p_{0k}$, and $0.5 < \theta_k < 1$. If the variant is non-causal, then $p_{1k} = p_{0k}$ and $\theta_k = 0.5$. The relationship between the effect parameter and usual disease-genotype odds ratio OR_k is given by

$$\theta_k = OR_k / \{OR_k(1 + p_{0k}) + (1 - p_{0k})\}. \quad (1)$$

Most of the variants in a gene region have little or no effects on the disease. In this paper, we describe the effect heterogeneity of variants by using a distribution. This is in effect to analyze variants based on a random-effects model [24] which uses a two-level hierarchical modeling. The distribution of data constitutes the first level of the modeling with variant-dependent parameter. In our case, the first level distribution is the conditional probability of X_{1k} given θ_k :

$$p(X_{1k}|\theta_k) = \binom{n_{1k}}{x_{1k}} p_{1k}^{x_{1k}} (1 - p_{1k})^{n_{1k} - x_{1k}} = \binom{n_{1k}}{x_{1k}} \left(p_{0k} \frac{\theta_k}{1 - \theta_k} \right)^{x_{1k}} \left(1 - p_{0k} \frac{\theta_k}{1 - \theta_k} \right)^{n_{1k} - x_{1k}}. \quad (2)$$

The distribution of the effect parameters, or random-effects, constitutes the second level of the modeling. Here, we assume that random effects θ_k 's follow a mixture distribution $(1 - \alpha) I_{0.5} + \alpha U(0.5, 1)$, where $I_{0.5}$ is a point mass distribution at 0.5 and $U(0.5, 1)$ is a uniform distribution over 0.5 and 1. The mixture distribution specifies that $100(1 - \alpha)\%$ of the variants are non-causal and $100\alpha\%$ of the variants are causal with effect levels evenly distributed between 0.5 and 1. Thus testing association of the variants with disease risk is equivalent to testing $\alpha = 0$.

To infer the proportion α , we need the marginal probability of X_{1k} , which is given by

$$\begin{aligned} p(X_{1k}|\alpha) &= (1 - \alpha) \binom{n_{1k}}{x_{1k}} p_{0k}^{x_{1k}} (1 - p_{0k})^{n_{1k} - x_{1k}} \\ &\quad + 2\alpha \binom{n_{1k}}{x_{1k}} \int_{\frac{1}{2}}^1 \left(\frac{p_{0k}\theta_k}{1 - \theta_k} \right)^{x_{1k}} \left(1 - \frac{p_{0k}\theta_k}{1 - \theta_k} \right)^{n_{1k} - x_{1k}} d\theta_k \\ &= (1 - \alpha) \binom{n_{1k}}{x_{1k}} p_{0k}^{x_{1k}} (1 - p_{0k})^{n_{1k} - x_{1k}} \\ &\quad + \alpha \frac{2(1 + p_{0k})p_{0k}}{(1 - p_{0k})} \binom{n_{1k}}{x_{1k}} \int_{p_{0k}}^1 \frac{t^{x_{1k}}(1 - t)^{n_{1k} - x_{1k}}}{(p_{0k} + t)^2} dt. \end{aligned} \quad (3)$$

The last equality is derived by taking $\frac{p_{0k}\theta_k}{1 - \theta_k} = t$ in the integral. p_{0k} 's are estimated from the control data as follows: $p_{0k} = X_{0k}/N_{0k}$ if $X_{0k} > 0$ and $p_{0k} = 1/(2N_{0k})$ if $X_{0k} = 0$. We use the

pseudo-likelihood (since p_{0k} 's are estimated) function $L(\alpha) = \prod_{k=1}^K p(X_{1k}|\alpha)$ to derive a likelihood ratio statistic $2\log \{L(\hat{\alpha})/L(0)\}$ for testing $\alpha=0$, where $\hat{\alpha}$ is the maximal likelihood estimate based on $L(\alpha)$. Note that the numbers of case (or control) genotypes at variants are not required to be equal. This means that missing genotypes are allowable. Also note that $L(\alpha)$ may not be the true joint likelihood of the genotype data, since the SNVs are from the same gene region and their genotype data may be correlated. However, if all variants are rare then genotype data are frequently considered as approximately independent. In this paper, we treat $L(\alpha)$ as a working likelihood and use the usual permutation argument to calculate p-value. Therefore, our test is still valid even the independence condition fails to

hold. The validity of this approach was confirmed by our simulation results. Our likelihood ratio test is denoted by T_{REM} , since it is derived from a random-effects model.

The formulation of our random effects model is based on the assumption that causal variants for a disease are deleterious. If variants are expected to be protective, we can use the same model but the definitions of major and minor alleles need to be exchanged. Further, in formulating the second-stage model, we have used $U(0.5,1)$ to model the distribution for the effects of causal variants. This means that if a variant is causal, then its effect size will take any value between 0.5 and 1 with equal chance. In principle, this model is reasonable when one has no prior knowledge about the magnitude of the effect. However, quite often, one is able to specify an upper bound for the odds ratios, denoted by OR^* . In this situation, the

effect size θ_k has an upper bound $B_k = \frac{OR^*}{OR^* + 1 + p_{0k}(OR^* - 1)} < 1$, and it is more reasonable to use $U(0.5, B_k)$ to model the distribution of random effects. This is in effect allowing random effects to take values closer to their true levels with higher probability. We expect that T_{REM} has better performance if a sharper and true upper bound is used. However, according to our simulation results, if the number of causal variants is small the improvement is prudent even when the sharpest upper bound is used. Another interesting feature of using this distribution is that if the k -th variant has a minor allele frequency greater than that for the l -th variant, then $B_k < B_l$. This means that the random-effects model actually specifies relatively smaller mean effects for higher frequency variants and larger mean effects for smaller frequency variants. Thus, although it is not explicitly stated, weighting has been automatically used in our modeling of random effects. In contrast, if the same $U(0.5,1)$ is used for modeling all distributions of random effects, high and low frequency variants are weighted equally.

We suggest setting $OR^*=5$ because that most odds ratios discussed in the literatures were smaller than this value. All simulations reported in this paper were conducted with this default choice. Free software, T_{REM} , written in R code for computing the test statistic and its p-value, can be downloaded from <http://www2.cmu.edu.tw/~biostat/downloads/KFCpro/KFCprogram.htm>.

3. Simulations

We performed simulations to study the properties of four competing tests based on the sequence-level data. The four tests were CMC, WS, SKAT and T_{REM} . All variants with allele frequencies below 0.01 were collapsed into one group in computing the CMC test. The p-values of WS and T_{REM} were determined by permuting the diseased/non-diseased status and those for CMC and SKAT were determined by using asymptotic distributions.

We used coalescent simulation model with similar LD pattern, recombination rate, and population history for the Europeans to simulate realistic sequence-level genetic data. We randomly selected one gene region with 250 kb in length and used COSI [25] to generate a pool of 25000 haplotypes. From this pool of haplotypes, we calculated the allele frequencies of all variants.

We considered the region containing K variants. Among them, the first M variants were assumed to be causal and the rest were non-causal. In order to evaluate type I error rate of each test under the null hypothesis and missing genotype, we consider the region composed of 2–200 rare variants with frequencies between 0.1% and 1% and missing rate 10%, 30% or 50%.

To study the effect of variant frequency we considered two scenarios: (1) 5 causal variants ($K=M=5$) in the region with frequencies equal to 3%, 5% (10%, 40%) or between 0.1% and 1%, (2) 25 variants in the region with 5 rare causal variants and 20 non-causal variants with frequencies equal to 3%, 5% (10%, 40%) or between 0.1% and 1%. In order to study the power performance and its impact by the inclusion of the non-causal variants, we considered the region with 10 causal variants with OR s between 1.8 and 2.0 and 0%, 50%, 67%, 75%, 80%, 90%, or 95% of non-causal variants. To study the impact of the effect directionality on the power performance, we also considered the regions with 100%, 80% or 60% of the causal variants being deleterious and rest being protective. To study the impact of missing genotypes on the power performance, we considered the scenarios with 10%, 30% or 50% missing rate for all subjects at each locus. In all scenarios, rare variants had minor allele frequencies (MAF_s) ranging from 0.1% to 1%. High frequency (common) variants had MAF either 3% or 5% (10% or 40%). Among randomly selected rare variants in our simulations, we found that the largest minor allele frequency (MAF_{max}) was 0.0936 and the smallest minor allele frequency (MAF_{min}) was 0.00120. The OR of a causal variant was determined according to its MAF and the OR -range (OR_{min} to OR_{max}) through the formula:

$$OR = OR_{min} + (1 - \frac{MAF - MAF_{min}}{MAF_{max} - MAF})(OR_{max} - OR_{min}).$$

This means that the causal variant had smaller OR if it had larger MAF . In our simulations, most of the selected rare variants for study had very small linkage disequilibrium ($LD=r^2$). For example, more than 95% of the randomly selected variants had LD s smaller than 1%. However, there are cases that LD s of the rare variants are high ($>1\%$). To study the impact of LD , we also add a small simulation, with specially selected high LD rare variants (more than 50% of their r^2 greater than 1%) generated from COSI, to investigate the type I errors and power of the competing tests.

The joint genotype frequency of the K variants was determined from the pool of haplotypes. Given the genotype data ($G_1, \dots, G_M, \dots, G_K$) of K variants, where dominant genotype coding was adopted, the corresponding disease status D (1 for diseased and 0 for non-diseased) was determined by the disease prevalence $\Pr(D=1)=0.01$ and risk model

$$\Pr(D=1|G_1, \dots, G_K) = \frac{\exp(\alpha + \sum_{k=1}^K \beta_k G_k)}{1 + \exp(\alpha + \sum_{k=1}^K \beta_k G_k)},$$

where β_k was the log- OR of the k th causal variant.

To study the null performance and power of the four tests under each simulation condition, we generated 5000 replicates of n_1 case data and n_0 control data and used sample size $n=n_0=n_1=500$, or 1000 in most simulations. In each replication, the case and control data were sampled until the pre-specified sample size was reached. Also, 2000 permutation steps were performed in each replication to determine p-values of the WS and T_{REM} tests. Type I error rate or power were estimated by the proportion of replicates with a p-value ≤ 0.05 . Upper bound $OR^*=5$ was used for computing T_{REM} in all simulations.

4. Results

4.1 Empirical Type I Error Rates

Figure I shows the simulated type I error rates of the four competing tests under various numbers (2–200) of non-causal variants and genotype missing rates. The results indicate that all four tests are able to control their type I error rates near the desired 5% level even genotype missing rate is as large as 50%. The average and range of the simulated type I error rates of the T_{REM} test are 5.0% and 4.2%–5.8%, respectively. The corresponding values for SKAT, WS and CMC tests are 4.8% and 3.3%–5.7%, 5.0% and 4.4%–5.6%, and 5.0% and 4.3%–5.5%, respectively. Notice that even sample size is small (500), the four tests still have very reasonable performance in controlling type I error rate. SKAT is slightly conservative when there are only two variants.

4.2 Empirical Power When High Frequency Variants Were Included

High frequency variants tend to distort the performance of the collapsing methods. Table I shows some power results for the four tests when there are high frequency causal or non-causal variants included in the region. We assumed that the high frequency causal variants had odds ratios 1.2–1.4 and their MAFs were either 3% or 5% (10% or 40%). In the case of 3% or 5% MAFs, the results indicate that the inclusion of two high frequency but low effect causal variants tend to lower the power of all tests. However, by increasing sample size, the power loss becomes much smaller. In the case of small sample size, T_{REM} has the best power performance while the CMC and WS tests perform equally well as the second best. If the sample size is moderate or large, all tests tend to have similar power. For example, the largest percentage of power loss for T_{REM} is 11.5% when the sample size is 500, but reduced to 1.5 % when sample size is increased to 1000. That is, for moderate sample size, the inclusion of high frequency but low effect variants in the region does not cause significant power gain or loss if T_{REM} is used. We also observe similar effect for other tests. In the case of including high frequency (MAF 3% or 5%) but non-causal variants in the study, the power of T_{REM} or SKAT is also found to be almost unchanged even when sample size is small. Nevertheless, the power performance of CMC and WS tests may be changed. We find the power loss is 21.7% for the CMC test and 27% for the WS test when sample size is 500. But they are reduced to 9%, and 16.4%, respectively, when sample size is increased to 1000.

In the case of including common causal variants with higher MAFs 10% or 40% in the region, the power of all tests tend to be greater than that with MAFs 3% or 5%. That is, the negative impact of including common causal variants (if there is any) in study is smaller if

the variants have larger MAFs. If the common variants are non-causal, the power of all tests, excepting WS, is also greater if larger MAFs are used. However, for the WS test, its effects are reversed. It is more powerful if the common variants have smaller MAFs. That is, the negative impact of including non-causal common variants in the region is greater if the variants have larger MAFs.

4.3 Empirical Power When Non-causal Variants Were Included and/or Effects Have Different Directions

The collapsing methods are designed to aggregate effects regardless of the effect directions. Thus, if the effects are in different directions, they may be cancelled out and the tests are underpowered. Figure II presents results showing the joint effect on power due to the inclusion of non-causal variants and causal variants with different effect directions. We find out that if all effects or most effects are in the positive directions (more than 80% of the causal variants were deleterious in our simulations), T_{REM} has the best power performance. If effects are less consistent in one direction (only 60% of the causal variants are deleterious and 40% are protective), SKAT becomes better than T_{REM} unless sample size is small (500). The results indicate that the power of any competing test is simultaneously influenced by the number of non-causal variants, direction-consistency level (number of protective variants/number of deleterious variants), and sample size. In general, the CMC and WS tests are less robust to the inclusion of non-causal variants. The power loss becomes severe as the number of non-causal variants grows. Although the power loss can be recovered by increasing sample size, however, the improvement is very slow. In contrast, T_{REM} and SKAT are more robust to the inclusion of non-causal variants, if all effects are consistent in one direction. However, in the case of small sample size, T_{REM} is better than SKAT. As sample size grows, the power losses of the two tests become negligible.

Figure II also indicates that the impact of non-causal variants might be intensified by the direction-consistency level of the effects. For example, when the sample size is small (500), the power reduction of any test is at least 50% if less than 80% of the causal variants are deleterious. However, T_{REM} and SKAT are relatively better than the CMC and WS tests in this case. The power loss due to different effect directions can be recovered back if larger sample size (> 2000) is used and T_{REM} or SKAT is applied in association studies. If the CMC or WS test is applied, one would need very large sample size to detect association signals.

4.4 Empirical Power When Genotypes Were Missing

Table II presents results showing how the power of the competing tests is influenced by the number of non-causal variants, genotype missing rate and sample size. It is clear that the power performance might be influenced by the presence of missing genotypes. In principle, the power loss is greater when the missing rate is higher. Despite of this, we find out that T_{REM} is very much robust to the joint effect of non-causal variant and genotype missing when sample size is moderate (1000). Under all conditions given in the simulation study and sample size 1000, the power range of T_{REM} is 0.915–0.995. Taking our simulation variations into account, this range is relatively small. The corresponding ranges for SKAT, CMC and WS test are 0.485–0.976, 0.275–0.998, and 0.533–0.998, respectively. The CMC

and WS tests are low powered because of the inclusion of non-causal variants not because of genotype missing. On the other hand, SKAT is low powered because of high genotype missing rate not because of the inclusion of many non-causal variants. When sample size is small (500), the power range for T_{REM} is 0.713–0.931. The corresponding ranges for SKAT, CMC and WS test are 0.248–0.771, 0.155–0.919, and 0.293–0.923, respectively. This indicates that when sample size is small, the power loss of SKAT, due to missing genotypes, is also the greatest.

4.5 Empirical Type I errors and Power When the Variants Have High LD Level

In our previous simulations, we found that 95.56% and 98.13% of the r^2 were less than 0.01, when the numbers of variants under study were 10 and 200, respectively. This is because that the rare variants have very small MAFs and in general, their LDs tend to be small. However, there are exceptional cases where variants are in higher LD level. To study the effect of LD, we selected variants in different LD levels for further simulations. Table III presents type I errors and power of the four competing tests when the variants are in high or low LDs and sample size $n=500$. First, under all cases, the type I errors of all tests were very near to the designated 5% level. The reported type I errors ranged from 0.047 to 0.051 for all tests. Next, in the case of high LD causal variants, we found that the power of all tests was greater if the non-causal variants were also in high LD. However, if the LD level of the non-causal variants was lower, the power of CMC and WS tests became smaller, but the power of SKAT and T_{REM} was almost unchanged. That is, SKAT and T_{REM} were more robust to the LD level of non-causal variants. The power of all tests was smaller if the causal variants were in low LD. For example, the power of SKAT dropped from 0.983 to 0.732 when the low LD causal variants were used instead of high LD causal variants. In contrast, the power of T_{REM} dropped from 0.993 to 0.775.

4.6 The Shanghai Breast Cancer Study

We evaluated our method and the alternative methods using genotype data from the Shanghai Breast Cancer Study, a large case-control study of breast cancer risk in East Asian women. A total of 2774 cases and 2177 controls were genotyped using the Affymetrix SNP 6.0 array in a genome-wide association study (GWAS). Detailed genotyping methods and stringent QC criteria were described in Zheng et al²⁷. We focused on the FGFR2 gene as it is well known to be associated with breast cancer risk [26,27]. We selected all SNPs from our GWAS data that passed QC, had minor allele frequency (MAF) ≥ 0.05 , and were within 250 kb of rs10510102, a SNP previously associated with breast cancer. There were 32 SNPs that qualified the criteria. SNP rs10510102 was not included as its MAF was >0.05 .

This data set allows us to study the association of breast cancer risk with rare variants at the FGFR2 gene. The numbers in the parentheses of Table IV are the number of SNPs with higher MAF in the cases (numerator) and the number of SNPs with higher frequency in the controls (denominator). MAF cut-off = 1 % means that variants with MAF less than or equal to 1% were included in the analyses. Other cut-off had similar definition. From Table IV, we see that T_{REM} has detected association signals with p-values 0.016–0.0445 for cut-offs 2%–5%, while SKAT, WS and CMC tests have p-values 0.0130–0.0600, 0.1658–0.7978, and 0.0284–0.8513, respectively. These results indicate that T_{REM} has the smallest p-value range

and hence is more robust. According to Table III, we also suspect that only the variants with MAFs between 1% and 2% are associated with the breast cancer. The other variants are suspected to be non-causal.

5. Discussion

There are many real examples showing that complex diseases can be caused by rare variants. However, analyzing rare variants with the usual method one by one is unlikely to detect any association signal, due to the rarity of the variants and presence of too many variants in the gene regions. This leads one to consider tests based on collapsing a set of variants in a gene region. However, these methods also have limitations because of large portion of gene regions that harbor many non-causal variants can create noise and reduce the power of signal detection. In addition, deleterious and protective variants may coexist in the same region and thus if collapsing method is applied on these variants, their combined effect may become too small for signal detection unless sample size is very large. To circumvent this difficulty, some authors [25] proposed a two-stage approach. Their idea is to divide variants into two subsets first: one in which each variant had a greater variant frequency in the cases than in the controls and the other in which each variant had a greater frequency in the controls. Second, applied the CMC or WS tests on the two subsets and obtained two p-values. The final p-value was the minimal of the two adjusted by Bonferroni correction. Unfortunately, this test was conservative and their power property not well studied when large portion of the variants were non-causal.

SKAT was a recently suggested test that addressed above limitations. It is a flexible, computationally efficient regression approach and can be adjusted for covariates such as principle components to account for population stratification. Another similar regression approach was given by Lin and Tang [29]. Both approaches require the choice of weight function for the regression parameters and complete data. In the case of missing genotypes, a simple imputation method based on the observed allele frequencies and Hardy-Weinberg equilibrium condition was suggested for SKAT [23]. However, when covariates are included in the regression model or LDs between variants are not weak, using this simple imputation method should be careful. Genotype missing is common in sequencing studies. The missing rate can be anywhere, probably depending on the depth of the sequencing. The missing rate could be high when depth is low. Since the 1000 genome projects are generating low depth coverage data, thus the issue of missing genotypes is important. According to our simulation study, when missing rate was high, the power loss of SKAT was also high even covariates were not included. More studies are needed in order to fully understand the properties of SKAT when missing rate is high or when covariates are included. In the future, we plan to address this issue when covariates are included in the analysis. We also plan to compare the performances of SKAT and T_{REM} when missing data are imputed using 1000 genomes as reference panel prior to the analysis or other type of “correct” information is applied for imputation..

There are six major factors influencing the power performance of the tests: number of non-causal variants, number of causal variants and their effect sizes, variant frequency, effect directionality, missing genotypes and sample size. In this paper, we have proposed a new

association test to address these issues. One distinguishing feature of T_{REM} is that it is not designed to compare genetic information between cases and controls as other existing tests, or using the collapsing concept. Also, the new test does not assume regression model as that given in SKAT. Instead, we propose to test whether the fraction of functional variants in a region equals zero. We have previously noted that our test has implicitly used a weighting strategy such that the variants with smaller MAFs have been assigned larger upper bounds for their random effects. However, if necessary, we also can assign different weights to the different log-likelihood contributions of SNVs. The performance of this approach deserves study in the future.

We have performed extensive simulations to compare the performance of four competing tests. We have shown that T_{REM} could properly controlled type I error rate and was better than the competing tests in power when sample size was small. In particular, when all causal variants were more consistent in one direction, our test had better performance. As sample size grew, T_{REM} and SKAT tended to have similar performance but T_{REM} was slightly inferior to SKAT when the effect directions were less consistent. This is because that SKAT assumed random effect followed a zero mean and weighted variance distribution and thus it is more powerful when the effects had different directions. But it is less powerful when all effects were in the same direction. Since the effects are often in the same direction and SKAT is very sensitive to genotype missing, our test deserves recommendation in real applications.

T_{REM} also depends on the choice of OR^* . We took $OR^*=5$ in all simulations and its performance seemed acceptable. However, we have noticed that its performance may be further improved by using smaller OR^* . Unreported simulation results showed that its power gain is greater when the number of causal variants is larger. To address the issue of optimal OR^* , however, more study is needed.

In the study of rare variants, their LDs are usually small. This is because that the variants' MAFs are very small. However, there are few rare variants having relatively high LD. If one has such causal variants, we found that SKAT and T_{REM} were more powerful in detecting the association signals and also robust to the LD level of the non-causal variants. Under this situation, the performance of T_{REM} relative to SKAT was still better. Finally, we emphasize that the CMC and WS tests were developed specifically for the next generation sequencing data. If we focused only on the rare variants with MAFs between 0.05% and 0.1%, unreported simulation results also indicated that T_{REM} had better power performance.

The random-effects model approach is a powerful method for detecting rare variants associated with disease. The next step after signal detection is to perform replication study. Since the associated region may contain many non-causal variants, it is also important to have a strategy to identify more important variants for further study. Currently, we are using the concept of the random-effects model to solve this problem.

Acknowledgments

The research of K. F. Cheng was supported in part by the National Science Council of Taiwan (NSC100-2118-M-039-001-MY3). The research of C. Li was supported in part by the National Institutes of Health grant R01HG004517.

References

1. Iyengar SK, Elston RC. The genetic basis of complex traits: Rare variants or “common gene, common disease”? *Methods Mol. Biol.* 2007; 376:71–84. [PubMed: 17984539]
2. Smith DJ, Lusk AJ. The allelic structure of common disease. *Hum. Mol. Genet.* 2002; 11:2455–2461. [PubMed: 12351581]
3. Maher B. Personal genomes: the case of missing heritability. *Nature.* 2008; 456:18–21. [PubMed: 18987709]
4. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat.Rev.* 2010; 11:415–425.
5. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* 2008; 40:695–701. [PubMed: 18509313]
6. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* 2009; 19:212–219. [PubMed: 19481926]
7. Pritchard JA. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* 2001; 69:124–137. [PubMed: 11404818]
8. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol.* 2010; 8:e1000294. [PubMed: 20126254]
9. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.* 2008; 17:R156–R165. [PubMed: 18852205]
10. Lowe CE, Cooper JD, Brusko T, Walker NM, Smyth DJ, et al. Large scale genetic fine mapping and genotype-phenotype associations implicate polymorphism IL2RA region in type I diabetes. *Nat. Genet.* 2007; 39:1074–1082. [PubMed: 17676041]
11. Burfoot RK, Jensen CJ, Field J, Stankovich J, Varney MD, et al. SNP mapping and candidate gene sequencing in the class I region of the HLA complex: searching for multiple sclerosis susceptibility genes in Tasmanians. *Tissue Antigens.* 2008; 71:42–50. [PubMed: 17971048]
12. Hafler JP, Maier LM, Cooper JD, Plagnol V, Hinks A, et al. CD226 Gly307Ser association with multiple autoimmune diseases. *Genes. Immun.* 2009; 10:5–10. [PubMed: 18971939]
13. Deloukas, P. On behalf of the Wellcome Trust Case Control Consortium. High throughput approaches to fine mapping in regions of confirmed associations. Presentation at the 58th Annual Meeting of the American Society of Human Genetics; November 13, 2008; Philadelphia Pennsylvania. 2008.
14. McClellan J, King MC. Genetic heterogeneity in human disease. *Cell.* 2010; 144:210–217. [PubMed: 20403315]
15. Johansen, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* 2010; 42:684–687. [PubMed: 20657596]
16. Bansal V, Ondrej L, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nature Rev. Genet.* 2010; 11:773–785. [PubMed: 20940738]
17. Li B, Leal S. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 2008; 83:311–321. [PubMed: 18691683]
18. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5:e1000384. [PubMed: 19214210]
19. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, et al. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 2010; 86:832–838. [PubMed: 20471002]

20. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.* 2007; 615:28–56. [PubMed: 17101154]
21. 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]
22. Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am. J. Hum. Genet.* 2010; 87:604–617. [PubMed: 21070896]
23. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association test (SKAT). *Am. J. Hum. Genet.* 2011; 89:82–93. [PubMed: 21737059]
24. Laird N, Ware H. Random-effects models for longitudinal data. *Biometrics.* 1982; 38:963–974. [PubMed: 7168798]
25. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 2005; 15:1576–1583. [PubMed: 16251467]
26. Hunter DJ, et al. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* 2007; 39:870–874. [PubMed: 17529973]
27. Zheng W, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* 2009; 41:324–328. [PubMed: 19219042]
28. Zhang L, Pei YF, Li J, Papasian CJ, Deng HW. Efficient utilization of rare variants for detection of disease-related genomic regions. *PLoS ONE.* 2010; 5(12):e14288. [PubMed: 21170328]
29. Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* 2011; 89:354–367. [PubMed: 21885029]

Figure I.

Type I error rates of the tests. R is the probability of missing a genotype randomly at each variant for each subject. The case and control sample sizes are both equal n.

Figure II.

The power of the tests when the effects of the variants are in different directions. P^* is the proportion of positive log-ORs. The case and control sample sizes are both equal n .

Table 1

The power of the tests when high frequency causal and non-causal variants are included.¹

Number of high frequency causal variants ²		Number of low frequency causal variants	Sample size					
			500			1000		
			WS	CMC	T _{REM}	SKAT	WS	T _{REM}
0	5		0.955	0.956	0.863	0.950	0.998	0.998
1	4		0.937	0.942	0.865	0.952	0.999	0.998
			(0.953)	(0.954)	(0.912)	(0.974)	(0.999)	(1.000)
2	3		0.748	0.751	0.693	0.840	0.959	0.983
			(0.905)	(0.865)	(0.829)	(0.944)	(0.988)	(0.999)
Number of high frequency non-causal variants ³		Number of low frequency non-causal variants	CMC	WS	SKAT	T _{REM}	CMC	WS
0	20		0.665	0.667	0.720	0.776	0.920	0.916
1	19		0.562	0.579	0.701	0.766	0.867	0.851
			(0.566)	(0.448)	(0.719)	(0.771)	(0.860)	(0.716)
2	18		0.521	0.487	0.691	0.764	0.837	0.766
			(0.520)	(0.309)	(0.720)	(0.769)	(0.835)	(0.464)

¹The first high frequency variant has MAF 3% and the second high frequency variant has MAF 5%. The numbers in the parentheses are the power of the tests when the first high frequency variant has MAF 10% and the second high frequency variant has MAF 40%. High frequency causal variants have ORs between 1.2 and 1.4.

²The low frequency causal variants have MAFs between 0.1% and 1% with ORs between 2.4 and 2.6.

³The total number of variants is 25. Among them, there are 20 non-causal variants and 5 causal variants. All low frequency variants have MAFs between 0.1% and 1%. The causal variants have low frequencies with ORs between 2.4 and 2.6.

Table II

The power of the tests when genotype data are missing.¹

Sample size									

Sample size									
500					1000				
Genotype missing rate	Proportion of causal variants ²	CMC	WS	SKAT	T _{REM}	CMC	WS	SKAT	T _{REM}
50%	33%	0.592	0.604	0.370	0.775	0.865	0.872	0.708	0.951
	25%	0.550	0.587	0.396	0.789	0.824	0.851	0.727	0.960
	20%	0.464	0.499	0.361	0.769	0.765	0.799	0.717	0.952
	10%	0.324	0.398	0.464	0.806	0.564	0.681	0.807	0.965
	5%	0.198	0.313	0.435	0.770	0.324	0.534	0.784	0.951
	100%	0.766	0.773	0.286	0.798	0.967	0.967	0.566	0.963
	50%	0.612	0.615	0.248	0.713	0.895	0.899	0.485	0.922
	33%	0.534	0.547	0.259	0.723	0.840	0.847	0.523	0.930
	25%	0.515	0.536	0.266	0.749	0.806	0.821	0.549	0.933
	20%	0.439	0.465	0.249	0.726	0.735	0.758	0.511	0.915
	10%	0.317	0.381	0.328	0.778	0.556	0.652	0.656	0.947
	5%	0.195	0.293	0.317	0.746	0.361	0.533	0.640	0.937

² All variants have MAFs between 0.1% and 1%. Among them, only 10 variants are causal with ORs 1.8–2.0.

Table III

The Power and type I errors of the tests when the variants have low or high LDs *

LD (r^2)		CMC	WS	SKAT	T _{REM}
5 causal variants	5 non-causal variants				
High ¹	High ¹	0.944 (0.050)	0.955 (0.052)	0.983 (0.048)	0.993 (0.045)
High ²	Low ²	0.848 (0.050)	0.873 (0.051)	0.979 (0.047)	0.987 (0.049)
Low ³	High ³	0.635 (0.050)	0.645 (0.051)	0.732 (0.048)	0.775 (0.049)

* 10 variants were used in the simulation. All variant allele frequencies were between 0.1% and 1%. In the power simulation, the ORs of the causal variants were between 2.4 and 2.6. In the type I error simulation, the ORs of the causal variants were 1. The simulated type I errors were given within the parentheses.

¹ Variants 1–5 were designated as causal variants, having 60% of their $r^2 > 0.01$, max $r^2 = 0.281$. Variants 6–10 were designated as non-causal variants, having 50% of their $r^2 > 0.01$, max $r^2 = 0.706$.

² Variants 1–5 were designated as causal variants, having 60% of their $r^2 > 0.01$; max $r^2 = 0.281$. Variants 6–10 were designated as non-causal variants, having 10% of their $r^2 > 0.01$.

³ Variants 1–5 were designated as causal variants, having 10% of their $r^2 > 0.01$. Variants 6–10 were designated as non-causal variants having 60% of their $r^2 > 0.01$ and max $r^2 = 0.281$.

Table IV

P-values of the tests applied to the Shanghai breast cancer study

Gene		MAF cut-off				
		1%	2%	3%	4%	5%
FGFR2 (rs10510102)	# variants	14(7/7)	21(14/7)	23(14/9)	27(16/11)	32(17/15)
	WS	0.8615	0.0284	0.1552	0.3544	0.8513
	CMC	0.8621	0.1658	0.3019	0.6093	0.7978
	SKAT	0.0545	0.0130	0.0185	0.0360	0.0600
	T _{REM}	0.2065	0.0160	0.0185	0.0285	0.0445