On Rare-Variant Analysis in Population-Based Designs: Decomposing the Likelihood to Two Informative Components

Sungho Won^{a, b} Youngdoe Kim^a Christoph Lange^{c–g}

^aDepartment of Applied Statistics, and ^bThe Research Center for Data Science, Chung-Ang University, Seoul, Korea; ^cGerman Center for Neurodegenerative Diseases, and ^dInstitute for Genomic Mathematics, University of Bonn, Bonn, Germany; ^eDepartment of Biostatistics, Harvard School of Public Health, ^fHarvard Medical School, and ^gCenter for Genomic Medicine, Brigham and Women's Hospital, Boston, Mass., USA

Key Words

Rare variants · Association analysis · Burden test

Abstract

Various analytical approaches have been suggested for the characterization of rare variants. One main approach is to collapse the genetic information of rare variants in a region and to construct an overall test statistic. Here, we proposed a new approach based on collapsed genotype scores. By utilizing the information of the association signal that is ignored in collapsing methods, i.e. the configuration of rare alleles, we constructed a more powerful test and compared it with existing rare-variant approaches. With extensive simulation studies, we showed that our method performs better than existing approaches, and we applied our method to a sequencing study of nonsyndromic cleft lip illustrating the practical advantages of the proposed method.

© 2014 S. Karger AG, Basel

Introduction

The development of statistical tools for the detection of associations between genetic variants and common diseases has been a major research focus in genetic epidemiology. Under the common disease common variant (CDCV) hypothesis, association methodology has continuously been developed and improved. However, even though genome-wide association studies based on CDCV have successfully identified a number of disease susceptibility variants (DSVs), these findings explain only a fraction of the expected heritability. Recent theoretical work and empirical data [1, 2] suggest that the common disease rare variant (CDRV) hypothesis is a plausible explanation for the so-called ‘missing heritabilities’ [3, 4]. The work of Bansal et al. [5] also suggests that the recent expansion of the human population may have resulted in a large number of rare DSVs, and therefore, rare-variant association analysis may play a key role in upcoming genetic association analyses.

CDRV assumes that the risk for a common disease is influenced by multiple rare variants with strong genetic effects, implying that only a few cases share the same

causal variants. With the arrival of high-throughput sequencing data, rare variants are the focus of genetic association studies. However, because single variant analysis has insufficient power to address analysis questions posed by the CDRV hypothesis, various collapsing approaches [6–10] that aggregate the genetic information for rare variants in a specified region and test their associations with the phenotype of interest have been suggested. For instance, it is unclear how to score collapsed genotypes, i.e. by indicator variables or allele counts. The most efficient scoring algorithm depends on the unknown disease model, and if multiple rare variants independently affect the disease in an additive way, the latter would be an efficient analysis strategy. A recent investigation [8] has found that the latter may be a robust choice for various disease models. Additionally, weighting rare alleles can increase the statistical efficiency, and if we let p_j be the minor allele frequency (MAF) for rare variant j , the most popular choice of weight has been $[p_j(1 - p_j)]^{-1/2}$ [7, 11]. This weighting scheme is efficient if the expected MAF differences in cases and controls are uniformly distributed. Other important issues include variations in effect sizes and their directions, and the MAF threshold choice for collapsing rare variants. In the presence of both protective and deleterious variants, the collapsed genotype scores between cases and controls are similar, and the drop in the statistical power of the collapsing approach is substantial. Some statistical methods, such as the Cochran-Mantel-Haenszel statistic [12], C-alpha test [13], and SKAT-O test [14], have been proposed in order to reduce the sensitivity of the collapsing approach to the presence of rare DSVs with different effect sizes and directions. Moreover, the choice of the allele-frequency threshold for collapsing rare variants also influences the efficiency of association analyses [9], and Price et al. [9] proposed to use a data-driven allele-frequency threshold that maximizes statistical efficiency.

However, despite all these methodological improvements, statistical challenges still remain for the analysis of whole genome sequencing data. For genome-wide next-generation sequencing studies, the number of markers surpasses 100,000,000, and investigations identifying efficient statistics are necessary to reduce the incidence of false-negative findings. In this report, we propose a new method for rare-variant association analysis. The proposed method consists of two independent statistics that are combined into an overall statistic. Our simulation studies showed that the proposed method was the most efficient in our simulation setting, and its application to real data illustrated the value of this approach in practice.

Methods

Let us assume that there are m rare variants in a genetic region that independently affect disease susceptibility. The variants can reside within a gene or a genomic region. Each variant has two alleles, denoted as A_j and B_j , where A_j is the rare allele with a frequency p_j at variant j in the population. The rare allele frequencies in cases and controls are denoted as p_j^a and p_j^u , respectively. We let x_{ij} be the number of A_j alleles in individual i at variant j , where $i = 1, \dots, n$, and x_{ij} can be 0, 1, or 2. The variable y_i indicates the disease status, i.e. 1 for cases and 0 for controls. The numbers of cases and controls are $n_a = \sum_i y_i$ and $n_u = \sum_i (1 - y_i)$, respectively. If we assume that the weight for variant j is w_j and $w_1^2 + \dots + w_m^2 = 1$, the collapsed genotype score X_i of individual i for all variants is $\sum_j w_j x_{ij}$. Furthermore, the collapsed genotype scores at variant j for cases and controls are $D_j = w_j \sum_i x_{ij} y_i$ and $d_j = w_j \sum_i x_{ij} (1 - y_i)$, respectively, and thus, the total collapsed genotype scores in cases and controls, g_a and g_u , become

$$g_a = \sum_{i=1}^n y_i X_i = \sum_{j=1}^m D_j, \quad g_u = \sum_{i=1}^n (1 - y_i) X_i = \sum_{j=1}^m d_j.$$

Two Statistically Independent Components for Rare-Variant Analysis

The analysis of rare variants is often based on collapsed genotypes, but additional information for the statistical analysis can be provided by using the component orthogonal to collapsed genotypes. Collapsed genotypes for cases and controls are denoted by g_a and g_u , respectively, and the likelihood for the component orthogonal to collapsed genotypes is obtained by conditioning on g_a and g_u . Therefore, the likelihood describing the relationship between multiple rare alleles and the disease status can be decomposed into a product of two conditional likelihoods

$$\begin{aligned} L(x_{ij}, y_i, i = 1, \dots, n, j = 1, \dots, m | n_a, n_u, m) \\ \propto L(g_a, g_u | g_a + g_u, n_a, n_u, m) L(x_{ij}, y_i, i = 1, \dots, n, \\ j = 1, \dots, m | g_a, g_u, n_a, n_u, m). \end{aligned}$$

Here, we will provide a statistic for each likelihood, and these two statistics will be combined into an overall statistic. The likelihood in the left part of the equation corresponds to the collapsed genotypes and that in the right part corresponds to the component orthogonal to collapsed genotypes, i.e. the rare allele configurations. For the derivation of our statistic, we assume that $w_j = 1$. However, because our final inference is based on permutations, our statistical test will be valid with different choices of weights.

First, we focused on $L(g_a, g_u | g_a + g_u, n_a, n_u, m)$ which is often used in collapsing methods. In particular, the distribution of $g_a + g_u$, which is not informative for genetic association, is conditioned to extract the nuisance parameter. The sum of binomial distributions with rare success probabilities has been shown to converge to the Poisson distribution under both independence and dependence of x_{ij} [15, 16], and thus, we can assume that $\sum_i y_i x_{ij}$ approximately follows the Poisson distribution. We let $\lambda_a = p_1^a + \dots + p_m^a$ and $\lambda_u = p_1^u + \dots + p_m^u$. If rare variants are in linkage equilibrium, and $w_j = 1$, we have

$$\begin{aligned} g_a = \sum_{i=1}^n y_i X_i = \sum_{j=1}^m \sum_{i=1}^n x_{ij} y_i \sim \text{Poisson}(n_a \lambda_a) \text{ and} \\ g_u = \sum_{i=1}^n (1 - y_i) X_i = \sum_{j=1}^m \sum_{i=1}^n x_{ij} (1 - y_i) \sim \text{Poisson}(n_u \lambda_u). \end{aligned}$$

Linkage disequilibrium between rare variants may not generate high correlations, and thus, this derivation seems reasonable. In the absence of association between any rare variants and affection status, $\lambda^a = \lambda^u$, and we have

$$\begin{aligned} L(g_a, g_u | g_a + g_u, n_a, n_u, m) \\ &= \frac{P\left(\sum_{i=1}^n y_i X_i = g_a, \sum_{i=1}^n (1 - y_i) X_i = g_u\right)}{P\left(\sum_{i=1}^n X_i = g_a + g_u\right)} \\ &= \frac{e^{-n_a \lambda} (n_a \lambda_a)^{g_a} / g_a! e^{-n_u \lambda} (n_u \lambda_u)^{g_u} / g_u!}{e^{-n \lambda} (n \lambda)^{g_a + g_u} / (g_a + g_u)!} \\ &= \frac{(g_a + g_u)!}{g_a! g_u!} \left(\frac{n_a}{n}\right)^{g_a} \left(\frac{n_u}{n}\right)^{g_u} \sim B\left(g_a + g_u, \frac{n_a}{n}\right). \end{aligned}$$

If the numbers of cases and controls are the same, n_a/n becomes 1/2. As a result, we could consider the following statistic based on $L(g_a, g_u | g_a + g_u, n_a, n_u, m)$

$$p_{T_1} = 2 \min \left\{ P\left(\text{Binom}\left(g_a + g_u, \frac{n_a}{n}\right) \geq g_a\right), P\left(\text{Binom}\left(g_a + g_u, \frac{n_a}{n}\right) \leq g_a\right) \right\}.$$

Our second test statistic is independent of p_{T_1} and can be obtained from $L(x_{ij}, y_i, i = 1, \dots, n, j = 1, \dots, m | g_a, g_u, n_a, n_u, m)$. If we assume that rare variants are independent, and $w_j = 1$, the conditional probability of $D_1, \dots, D_m$, given g_a , is

$$\begin{aligned} &\frac{P(D_1, D_2, \dots, D_m)}{\sum_{D'_1 + D'_2 + \dots + D'_m = g_a} P(D'_1, D'_2, \dots, D'_m)} \\ &= \left(\frac{g_a}{D_1, D_2, \dots, D_m} \right) (\pi_1^a)^{D_1} \dots (\pi_m^a)^{D_m}, \\ &\text{where } \pi_j^a = \frac{p_j^a}{p_1^a + \dots + p_m^a}, \end{aligned}$$

and this probability is denoted by $P(D_1, D_2, \dots, D_m | g_a, \pi_1^a, \dots, \pi_m^a)$. In particular, π_j^a for each rare variant is similar, and $P(D_1, D_2, \dots, D_m | g_a, \pi_1^a, \dots, \pi_m^a)$ for cases is asymptotically equal to the probability of the configuration of rare alleles as follows

$$p_{T_2} = P(D_1, D_2, \dots, D_m | g_a; 1/m, \dots, 1/m).$$

It is important to note that the rare allele configuration across the genomic region has not been taken into account in previous collapsing approaches [6–10], but p_{T_2} could still be informative in some situations. For instance, assuming that there is at least one DSV, the chance of observing one or more rare alleles at disease susceptibility loci is much higher than the chance at a null locus for multiplicative diseases, which would contribute to the nonrandom configuration of rare alleles. In the simulated data, we confirmed the informativity of p_{T_2} for the analysis of rare variants, and the configuration of rare alleles combined with genomic distance information has been also shown to be informative [17]. There-

fore, we combined p_{T_1} and p_{T_2} into our overall statistic. In particular, if $w_j \neq 1$, it should be noted that D_j in p_{T_2} is defined by

$$D_j = w_j \sum_{i=1}^n x_{ij} y_i,$$

and permutations have to be applied for the statistical validity of p_{T_2} .

To construct an overall association test based on p_{T_1} and p_{T_2} , we calculate the sum of the weighted Z-statistics, as suggested by Liptak [18]. Liptak's method has been shown to be the most efficient if the noncentrality parameter for each p value under the alternative hypothesis is known [19]. If we let Z_α be the α quantile of the standard normal distribution, our overall association test statistic T under the independence of p_{T_1} and p_{T_2} is obtained by

$$T = l_1 Z_{p_{T_1}} + l_2 Z_{p_{T_2}}, l_1^2 + l_2^2 = 1.$$

The noncentrality parameter divided by the standard deviation for each p value under the alternative hypothesis has been suggested as the most efficient choice for l_1 and l_2 [19], but those are unknown for the proposed two statistics. With extensive simulation studies, we found that $2/\sqrt{5}$ and $1/\sqrt{5}$ are usually the most efficient choices for the simulated data, and therefore, this choice was used in our simulations and data analysis. In particular, linkage disequilibrium and the weighted genotype score could generate correlations between markers even though their effects are minimal. Therefore, the overall p value is calculated by a permutation test that randomly assigns disease status to subjects, while keeping their genotypes fixed.

Simulation Settings

For our simulations, haplotypes were simulated using the software *cosi* [20], based on the coalescent model. In the coalescent model for *cosi*, we assumed that the mutation rate was 1.5×10^{-8} , and 5,000 haplotypes with 50,000 base pairs were generated. Then, we randomly selected m variants for which the MAFs were <0.05 . Pairs of haplotypes that were randomly chosen with replacement were used to derive individual genotypes in the region. We considered additive and dominant disease modes of inheritance, and the disease prevalence was assumed to be 0.15. Sampling was repeated until the given numbers of cases and controls were obtained. If we denoted the relative risk of the homozygous disease genotype at rare variant j by λ_j , we sampled λ_j from $U(1.0, 1.5)$ for the statistical efficiency evaluation; otherwise, it was assumed to be 1. The relative risk of the heterozygous disease genotype for each variant was calculated under the additive and dominant disease modes of inheritance. Therefore, the penetrance for each individual was

$$\begin{cases} \exp\left[\beta_0 + \sum_{j=1}^m \log(\lambda_j \cdot x_{ij} / 2)\right] & \text{for additive disease} \\ \exp\left[\beta_0 + \sum_{j=1}^m \log(\lambda_j \cdot I(x_{ij} > 0))\right] & \text{for dominant disease.} \end{cases}$$

Here, we chose β_0 , which resulted in 0.15 disease prevalence.

For comparison with the existing methods, the proposed method that was denoted by T was compared with the combined multivariate and collapsing (CMC) method [6], the variable threshold approach (VTA) [9], and rank-based Z-test (RBZ) [7]. For the CMC method, since genotypes are coded as 1 if rare alleles are observed and otherwise as 0, Pearson's χ^2 test is applied to summary contingency table. For the VTA method, among the various allele-

Table 1. Empirical type 1 error

<i>m</i>	<i>n</i>	α level	T	CMC	VTA	RBZ	LR	C-alpha	SKAT-O
5	1,000	0.05	0.0546	0.0536	0.0559	0.0539	0.0517	0.0554	0.0541
	1,000	0.01	0.0103	0.0100	0.0110	0.0118	0.0102	0.0108	0.0123
	2,000	0.05	0.0507	0.0530	0.0502	0.0538	0.0503	0.0517	0.0527
	2,000	0.01	0.0093	0.0100	0.0095	0.0103	0.0093	0.0083	0.0098
10	1,000	0.05	0.0503	0.0463	0.0523	0.0472	0.0503	0.0508	0.0515
	1,000	0.01	0.0102	0.0087	0.0125	0.0096	0.0102	0.0095	0.0115
	2,000	0.05	0.0492	0.0497	0.0502	0.0527	0.0498	0.0465	0.0493
	2,000	0.01	0.0093	0.0102	0.0113	0.0098	0.0100	0.0082	0.0095

Empirical type 1 error estimates for the proposed method are calculated from 10,000 replicates. p values are obtained with 1,000 permuted samples. The total sample size is *n*, and *m* indicates the number of rare variants in a genetic region.

Table 2. Empirical power estimates for each component

<i>m</i>	<i>n</i>	p_{T_1}	p_{T_2}	T	LR
20	500	0.159	0.102	0.162	0.158
	1,000	0.212	0.116	0.224	0.213
	1,500	0.274	0.125	0.300	0.262
	2,000	0.340	0.128	0.353	0.339
	2,500	0.371	0.149	0.381	0.367
	3,000	0.440	0.158	0.475	0.438
30	500	0.181	0.115	0.196	0.198
	1,000	0.296	0.111	0.320	0.307
	1,500	0.407	0.147	0.431	0.407
	2,000	0.498	0.182	0.518	0.498
	2,500	0.542	0.173	0.559	0.545
	3,000	0.627	0.208	0.649	0.624

Empirical power estimates for the proposed method are calculated from 1,000 replicates. p values for each replicate are obtained with 1,000 permuted samples. The total sample size is *n*, the disease prevalence is assumed to be 0.15, and *m* indicates the number of rare variants in a genetic region. p_{T_1} and p_{T_2} indicate the tests based on the collapsed genotypes and configuration of each rare variant, respectively. 20 and 30 rare variants whose MAFs were <0.05 were collapsed, and 50% of rare variants had deleterious effects on disease.

frequency threshold choices for collapsing rare variants, the threshold that results in the minimum p value is chosen and the permutation-based p value is calculated. For the RBZ method, each individual is ranked with the rare allele counts, and the sum of ranks for cases is compared with that for controls. We also considered the logistic regression (LR) where the disease status is regressed on the weighted sum of genotypes [8] and variance component approaches, such as the C-alpha [13] and the SKAT-O tests

[14]. The C-alpha method compares the magnitude of inflated variance between cases and controls, and the SKAT-O method combines C-alpha and LR with optimal weight. For all methods, each rare variant was weighted using

$$w_j = 1 / \sqrt{p_j^u (1 - p_j^u)},$$

and permutation tests were used to ensure the statistical validity for simulations with small sample sizes.

Results

Evaluation Using Simulated Data

The empirical type 1 error estimates were calculated from 10,000 replicates at the 0.05 and 0.01 significance levels, and in each replicate, we assumed $n = n_a + n_u$ and $n_a = n_u$. We assumed that there were 5 and 10 rare variants in a gene, and the p values were calculated with 1,000 permuted samples. Our results show that the proposed method preserves the nominal significance levels under the null hypothesis (table 1).

For statistical efficiency evaluation, we calculated the empirical power estimates based on 1,000 replicates for various sample sizes under the additive and dominant disease modes of inheritance. Table 2 shows the empirical power estimates of the two components of *T*, p_{T_1} and p_{T_2} . We randomly selected 20 and 30 rare variants whose MAFs were <0.05, and 50% of rare variants were assumed to have a deleterious effect on disease. Our results showed that p_{T_1} was more informative than p_{T_2} , and the performance of p_{T_1} was similar to that of LR. This result illustrates that our first component p_{T_1} seems to correspond to the collapsed genotypes, and p_{T_2} that corresponds to

Fig. 1. Empirical power estimates under the additive disease mode of inheritance, when rare variants whose MAFs were <0.05 were collapsed. Empirical power estimates for the proposed method at the 0.05 significance level were calculated from 1,000 replicates. p values were obtained with 1,000 permuted samples. 10 (**a**, **b**), 20 (**c**, **d**), or 30 (**e**, **f**) rare variants were collapsed. n indicates the number of cases and controls, and the disease prevalence was assumed to be 0.15.

the rare allele configuration is still informative, even though the power improvement of T by p_{T_2} is not substantially large.

In figures 1 and 2, we randomly selected 10, 20, and 30 variants whose MAFs were <0.05 : 10 rare variants for panels a and b, 20 rare variants for panels c and d, and 30 rare variants for panels e and f. We assumed that 5 and 7 variants were deleterious in panels a and b, respectively, 10 and 14 in panels c and d, respectively, and 15 and 20

in panels e and f, respectively. We assumed that the disease modes of inheritance for each variant were additive in figure 1 and dominant in figure 2. In figure 3, we randomly selected 50, 70, and 90 variants whose MAFs were <0.05 : 50 rare variants for panels a and b, 70 rare variants for panels c and d, and 90 rare variants for panels e and f. We assumed that 15 and 25 variants are deleterious in panels a and b, respectively, 25 and 35 in panels c and d, respectively, and 25 and 45 in panels e and f, respectively.

Fig. 2. Empirical power estimates under the dominant disease mode of inheritance, when rare variants whose MAFs were <0.05 were collapsed. Empirical power estimates for the proposed method at the 0.05 significance level were calculated from 1,000 replicates. p values were obtained with 1,000 permuted samples. 10 (**a**, **b**), 20 (**c**, **d**) or 30 (**e**, **f**) rare variants were collapsed. n indicates the number of cases and controls, and the disease prevalence was assumed to be 0.15.

The additive disease mode of inheritance was assumed for figure 3. The remaining rare variants had no effect on the disease status in figures 1–3. In figures 1 and 2, the proportions of deleterious rare variants were approximately 50 and 70%. In figure 3, the proportions of deleterious rare variants were approximately 30 and 50%. We generated 1,000 replicates, and the empirical powers were estimated as the proportions of replicates where the p values were less than the 0.05 nominal significance level.

Our results showed that the proposed method always performed better than the other approaches, followed by LR. The CMC method was the least powerful in our simulation setting, but CMC is expected to be relatively efficient, particularly in a situation where rarer variants have stronger effects on the disease [21]. SKAT-O, an optimized statistic of a collapsing method and the C-alpha method [13] were expected to have performances similar to those of collapsing methods, such as LR. How-

Fig. 3. Empirical power estimates under the additive disease mode of inheritance, when rare variants whose MAFs were <0.05 were collapsed. Empirical power estimates for the proposed method at the 0.05 significance level were calculated from 1,000 replicates. p values were obtained with 1,000 permuted samples. 50 (a, b), 70 (c, d) or 90 (e, f) rare variants were collapsed. n indicates the number of cases and controls, and the disease prevalence was assumed to be 0.15.

ever, the differences between LR and SKAT-O may be explained by the fact that LR and SKAT-O are based on Wald and Score test statistics, respectively. Similar results have been observed in the situation where the proportions of deleterious rare variants were 10% (result not shown). The empirical power improvement of the proposed method to the second-most efficient method, LR, ranged from 5 to 38%, and compared to the other approaches, the empirical power improvements ranged

from 10 to 100%. In addition, the empirical powers for the dominant disease mode of inheritance were usually larger than those for the additive disease mode of inheritance. For the dominant model, the genetic effects of heterozygous and homozygous genotypes were assumed to be the same, and thus, the effect size of the heterozygous genotype became larger. Therefore, the rarity of the homozygous genotype may improve the power for dominant diseases.

Table 3. Results for the association of RISKINT with nonsyndromic CL/P

Weight	Threshold	T	CMC	VTA	RBZ	LR	C-alpha	SKAT-O
1	0.01	0.0120	0.0990	0.1920	0.0603	0.0620	0.0260	0.0260
	0.05	0.0860	0.2920	0.2190	0.1748	0.1980	0.6540	0.2682
$1/\sqrt{\hat{p}_j^u(1-\hat{p}_j^u)}$	0.01	0.0590	0.1130	0.1660	0.0655	0.0500	0.0250	0.0241
	0.05	0.0780	0.2720	0.2290	0.2086	0.0910	0.6680	0.0754

The association of RISKINT with nonsyndromic CL/P is tested. p values are obtained with 5,000 permuted samples. Two choices of weights and thresholds for collapsing are considered. The statistics with the minimal p value for each case are shown in bold.

Next, we evaluated the sensitivity of the proposed method to different ranges of MAFs (online suppl. fig. 1 and 2, see www.karger.com/doi/10.1159/000357643). We randomly selected 5 or 10 rare variants whose MAFs were between 0.001 and 0.02. We assumed that 2 and 4 of them were deleterious in panels a and b, respectively, and 5 and 7 were deleterious in panels c and d, respectively. The remaining variants had no effect on the disease status. We calculated the empirical power estimates at the 0.05 significance level for the various sample sizes under the additive (online suppl. fig. 1) and dominant (online suppl. fig. 2) disease modes of inheritance. Our results showed that the proposed method still performs best in this scenario. The difference in performance between the proposed method and other methods was similar when rarer MAF thresholds were chosen and the dominant disease mode of inheritance was assumed.

Applications to Real Data

Cleft lip with or without cleft palate (CL/P) is one of the most common congenital malformations in the European population and can be part of a complex malformation syndrome or an isolated anomaly. The prevalence of cleft lip is 1/700 to 1/1,000, and nonsyndromic cases explain around 70% of all CL/P. Our resequencing study consisted of 192 cases with nonsyndromic CL/P and 192 controls in Germany and neighboring countries. Twenty-nine patients had two or more relatives with nonsyndromic CL/P in at least two consecutive generations, and all patients were examined by one of two medical geneticists to exclude any underlying syndrome. Controls were not screened for cleft status because a substantial drop in power would not be generated due to the very low prevalence of orofacial clefting in the general population [22].

Evidence of genetic association for the candidate gene sequenced in this study came from a previous genome-wide association study [23]. Three loci showed evidence

suggesting an association at the genome-wide significance level, and among these loci, the region located at 8q24 near rs987525 (hereafter called RISKINT) was resequenced. The entire coding region and untranslated regions were sequenced using an Applied Biosystems 3130XL Genetic Analyzer.

In total, 13 SNPs were identified from sequencing, and the proposed method was applied to the rare variants. Among the 13 SNPs, the MAFs of 6 SNPs were <0.05, and the MAFs of 4 SNPs were <0.01. We conducted association analyses with these rare variants. There were 16 cases and 14 controls with missing genotypes, and their genotypes were assumed to be 0. p values were calculated with 5,000 permuted samples, and we considered 1 and

$$1/\sqrt{\hat{p}_j^u(1-\hat{p}_j^u)}$$

as weights. Table 3 shows the results for two choices of thresholds and weights, and the statistics with the minimal p value for each case are shown in bold. Although no significant results were observed considering the multiple testing issue, table 3 shows that the significant results, at the 0.05 significance level, were found by the proposed method and SKAT-O, and the smallest p value of RISKINT with nonsyndromic CL/P was obtained with the proposed method, when we used $w_j = 1$ and the 0.01 allele-frequency threshold. SKAT-O also provided a significant association, when the 0.01 allele-frequency threshold and

$$1/\sqrt{\hat{p}_j^u(1-\hat{p}_j^u)}$$

as the weight were used; however, it is less significant than with the proposed method using $w_j = 1$. RISKINT has been reported to be highly associated with nonsyndromic CL/P from genome-wide association studies. Therefore, we can conclude that this result illustrates the practical advantage of the proposed method.

Discussion

In this report, we proposed a new approach for rare-variant association analysis that utilizes additional information about the association and is, consequently, more powerful than other existing methods. The proposed method consists of two informative components. The first component uses collapsed genotypes, and the second is based on the configuration of rare alleles. While existing methods have focused only on the former, the latter has not been generally considered, and in this study, we showed that this second component can be informative. We showed that combining the two components into an overall statistic led to an improvement in the statistical power compared to general collapsing approaches. However, even though our results from simulated data showed that Z-scores weighted by $2/\sqrt{5}$ and $1/\sqrt{5}$ were usually efficient, there is no uniformly most powerful method to combine two independent statistics, and further investigations are necessary to provide appropriate robustness for various disease models.

Furthermore, there are some additional limitations to the proposed method. First, the effects of each rare variant on disease status are usually unknown, and when the protective and deleterious rare variants are collapsed, the power drop for the proposed method could be substantial. Some modifications can be made to the proposed method to take this into consideration. For instance, as was suggested by Han and Pan [24], if the estimated regression coefficients for some rare variants are negative and the p values are <0.1 , their major alleles, instead of rare alleles, could be used to collapse genotypes, and the association of the collapsed genotype scores with the trait can be inferred using a permutation test. Alternatively, using some cases and controls, we could estimate the ex-

pected difference and direction of each rare variant, and the selected model could be applied to the remaining cases/controls. Second, our statistic cannot be utilized if the phenotype of interest is continuous or if some covariants affect the phenotype. Thus, an extension is necessary in such a scenario. For example, p_{T_1} based on the binomial distribution can be replaced by a logistic model, and for testing the configuration of rare alleles, i.e. p_{T_2} , the likelihood ratio test can be used. Third, we found that the permuted p values from 1,000 iterations for a gene containing 30 rare variants took 6.3 s with our R program on an Intel(R) Xeon(R) CPU E7-4850 @ 2.00GHz, which indicates that genome-wide analysis is computationally intensive. However other methods, such as SKAT-O, do not require permutation, and the proposed method is computationally less efficient. Fourth, our second statistic can be uninformative in certain situations. For instance, if the effect of each rare variant on disease is similar, rare alleles in cases may be randomly aligned. However, this scenario may not be common, and therefore, it is worthwhile to use the second component.

In summary, we showed that the proposed two-component method is efficient for identifying rare variants. The application of this method to real data illustrated its practical value, providing interesting findings, and the R code for the proposed method can be downloaded from <http://cau.ac.kr/~swon/software/software.php>.

Acknowledgement

This research was supported by the Chung-Ang University Research Scholarship grants in 2011. The authors would like to thank the anonymous referees for their helpful suggestions and comments.

References

- 1 Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB: Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010;8:e1000294.
- 2 Goldstein DB: The importance of synthetic associations will only be resolved empirically. *PLoS Biol* 2011;9:e1001008.
- 3 Manolio TA, Collins FS, Cox NJ, et al: Finding the missing heritability of complex diseases. *Nature* 2009;461:747–753.
- 4 Visscher PM, Hill WG, Wray NR: Heritability in the genomics era – concepts and misconceptions. *Nat Rev Genet* 2008;9:255–266.
- 5 Bansal V, Libiger O, Torkamani A, Schork NJ: Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 2010;11:773–785.
- 6 Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;83:311–321.
- 7 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009;5:e1000384.
- 8 Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010;34:188–193.
- 9 Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR: Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010;86:832–838.
- 10 Zhu X, Feng T, Li Y, Lu Q, Elston RC: Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol* 2010;34:171–187.
- 11 Ionita-Laza I, Buxbaum JD, Laird NM, Lange C: A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet* 2011;7:e1001289.

- 12 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.
- 13 Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ: Testing for an unusual distribution of rare variants. *PLoS Genet* 2011;7:e1001322.
- 14 Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X: Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;91:224–237.
- 15 Hodges JL, Le Cam L: The poisson approximation to the poisson binomial distribution. *Ann Math Stat* 1960;31:737–740.
- 16 Le Cam L: An approximation theorem for the poisson binomial distribution. *Pacific J Math* 1960;10:1181–1197.
- 17 Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, Fimmers R, Silverman EK, Pagano M, Mangold E, Lange C: 'Location, Location, Location': a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. *Bioinformatics* 2012;28:3027–3033.
- 18 Liptak T: On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl* 1958;3:171.
- 19 Won S, Morris N, Lu Q, Elston RC: Choosing an optimal method to combine p-values. *Stat Med* 2009;28:1537–1553.
- 20 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.
- 21 Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB: The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. *PLoS Genet* 2012;8:e1002496.
- 22 Moskvina V, Holmans P, Schmidt KM, Craddock N: Design of case-controls studies with unscreened controls. *Ann Hum Genet* 2005;69:566–576.
- 23 Mangold E, Ludwig KU, Birnbaum S, et al: Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet* 2010;42:24–26.
- 24 Han F, Pan W: A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 2010;70:42–54.