BMC Proceedings

PROCEEDINGS Open Access

A novel method to detect rare variants using both family and unrelated case-control data

Tao Feng, Robert C Elston, Xiaofeng Zhu*

From Genetic Analysis Workshop 17
Boston, MA, USA. 13-16 October 2010

Abstract

To detect rare variants associated with a phenotype, we develop a novel statistical method that can use both family and unrelated case-control data. Unlike the currently existing methods, we first use family data to calculate weights to be given to rare variants, differentiating between concordantly affected and discordant sib pairs. These weights are then used in an association test applied to the unrelated case-control data. We applied the proposed method to the simulated sequencing data in Genetic Analysis Workshop 17 and identified two genes associated with the disease.

Background

Genome-wide association studies, which are based on the common disease/common variants assumption, have successfully identified susceptibility loci for complex traits. However, the variants discovered through these studies explain only a modest portion of the trait variability [1]. With the new technological advances, it has been suggested that it is time to shift the search from common variants of modest effect to rarer variants of large effect by effectively searching the full genome [2]. Rare variants may hold promise to predict individual risk for personalized medicine because of their large effect, although it has been argued that common variants illuminate the biologic pathways that underlie diseases [3].

Bodmer and Tomlinson [4] suggested that a set of low-frequency variants from different genes can account for a significant proportion of the variability of relatively common diseases. To achieve reasonable statistical power, it is critical to define the rare variants and test them collectively. The existing statistical methods in the literature mainly collapse rare variants [5]. Madsen and Browning [6] proposed using the inverse of the variance of the minor allele frequency (MAF) in control subjects as a weight and then collapsing the weighted rare variants.

Briefly, for the *i*th individual Madsen and Browning [6] define a genetic score:

$$X_i = \sum_{j=1}^{L} w_j g_{ij}, \tag{1}$$

where L is the number of variants, g_{ij} is the genotypic score, and w_j is the weight for the *j*th single-nucleotide polymorphism (SNP). The weight w_j is defined as the inverse of the *j*th SNP's standard deviation estimated in control subjects when the corresponding MAF is less than α (such as 0.02) and 0 otherwise. Then the Wilcoxon rank sum test is applied to do the association test. Madsen and Browning rank the genetic scores, calculate the sum of the ranks for case subjects as:

$$X = \sum_{i \in \text{cases}} \text{rank}(X_i), \tag{2}$$

and calculate the p-value using a permutation strategy. That is, they permute disease status among individuals 1,000 times to compute 1,000 statistics X, denoted X_1^*, X_2^*, …, $X_{1,000}^*$. Then they use the sample mean $\hat{\mu}$ and

* Correspondence: xzhu1@darwin.EPBI.cwru.edu
Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Cleveland, OH 44106, USA

the standard deviation $\hat{\sigma}$ of X_1^*, X_2^*, ..., $X_{1,000}^*$ to calculate the test statistic:

$$Z = \frac{X - \hat{\mu}}{\hat{\sigma}}, \tag{3}$$

which follows approximately a standard normal distribution under the null hypothesis. Madsen and Browning [6] demonstrated that this weighted-sum method is more powerful than the collapsing method [5].

Recently, we have demonstrated that family data are useful for searching for rare variants [7,8], because the rare variants can be substantially enriched among segregating family members. Here, we present a statistical method to test rare variants by using both family and unrelated case-control sequencing data.

Methods

Defining a weight for each SNP using the family data set

We start by assuming that we have L variants belonging to a group (gene, pathway, specific genomic region, etc.). We have shown that family data, such as affected sib pairs, have enriched information for detecting rare disease-associated variants because the same disease variants are more likely to segregate within a family [7]. Thus each family can be considered homogeneous; that is, affected family members share the same or allelic disease variants, the latter being an idea used in traditional linkage analysis. Our method uses either affected sib pairs or discordant sib pairs to determine weights for such rare variants.

Assume that a SNP has two alleles A and a, where A always refers to the minor allele. For the ith sib pair and the jth SNP, we define a genotype score $\tilde{g}_{ij}$. Let α be a predefined threshold. We always let $\tilde{g}_{ij} = 0$ if a SNP has a MAF greater than α. For those SNPs with MAF ≤ α, we define $\tilde{g}_{ij}$ as follows. For affected sib pairs, $\tilde{g}_{ij} = 1$ if both sibs carry A, $\tilde{g}_{ij} = 0$ if neither sib carries A, and $\tilde{g}_{ij} = K / L$ if one sib carries A and the other does not, where K is the number of other SNPs in the region with MAF $<\alpha$ carried by the other sib. For discordant sib pairs, $\tilde{g}_{ij} = 1$ if the affected sib carries allele A and the unaffected sib does not, $\tilde{g}_{ij} = 0$ if the affected sib does not carry A, and $\tilde{g}_{ij} = K / L$ if both sibs carry A, where K is the number of alleles with MAF $<\alpha$ present at the other SNPs carried by the affected sib.

Let:

$$G_j = \frac{1}{N_{\text{sib}}} \sum_{i=1}^{N_{\text{sib}}} \tilde{g}_{ij}, \tag{4}$$

where N_{sib} is the number of sib pairs. For the jth SNP, define:

$$b_j = \frac{G_j}{[p_j(1-p_j)]^{1/2}}, \tag{5}$$

where p_j is the MAF of the SNP, estimated from the case and control subjects combined. We rank the b_j in descending order and give the jth SNP weight $w_j = b_j$ if b_j falls in the top quartile and weight $w_j = 0$ if not. The proposed weight incorporates information about the variants shared (not shared) by affected (discordant) sib pairs.

Performing association tests

After defining the weights for individual SNPs, using either affected or discordant sib pairs, we use the weights to test for association between the phenotype and each set of variants in a group. We assume that we have N_D unrelated case subjects and N_C unrelated control subjects. Let k be the kth unrelated individual. For each SNP j (j = 1, ..., L), we let w_j be the weight as found earlier using the sib pairs. We use the w_j to calculate the kth individual's genetic score, similar to the idea of Madsen and Browning [6]; that is,

$$\gamma_k = \sum_{j=1}^{L} w_j g_{kj}, \tag{6}$$

where g_{kj} is the number of minor alleles in SNP j for individual k. We then define:

$$\bar{\gamma}_D = \frac{1}{N_D} \sum_{k=1}^{N_D} \gamma_k = \frac{1}{N_D} \sum_{k=1}^{N_D} \sum_{j=1}^{L} w_j g_{kj}, \tag{7}$$

where the summation is over the unrelated case subjects. Similarly, we define:

$$\bar{\gamma}_C = \frac{1}{N_C} \sum_{k=1}^{N_C} \gamma_k = \frac{1}{N_C} \sum_{k=1}^{N_C} \sum_{j=1}^{L} w_j g_{kj}, \tag{8}$$

calculated for the unrelated control subjects. Our null hypothesis is that no marker is associated with the phenotype; that is, we have $E[\bar{\gamma}_D] = E[\bar{\gamma}_C]$, and so we define our test statistic for association as:

$$T = \frac{\bar{\gamma}_D - \bar{\gamma}_C}{(S_{\bar{\gamma}_D} + S_{\bar{\gamma}_C})^{1/2}}, \tag{9}$$

where $S_{\bar{\gamma}_D}$ and $S_{\bar{\gamma}_C}$ are estimates of the sample variances $\mathrm{var}(\bar{\gamma}_D)$ and $\mathrm{var}(\bar{\gamma}_C)$ in case and control subjects, respectively.

Application to the Genetic Analysis Workshop 17 sequence data set

The Genetic Analysis Workshop 17 (GAW17) simulated sequence data set includes both family and unrelated individuals data, which is ideal for the proposed method. Each replicate is composed of 697 individuals in the family data set and 697 individuals in the unrelated individuals data set. There are 200 replicates in total, all with the same genotype data but with phenotype data independently simulated across the replicates.

We first performed an analysis using only one replicate, which was selected to be the 82nd replicate. To calculate the weights, we clustered the sib pairs into affected and discordant sib pairs. Based on the 82nd phenotype replicate, we identified 38 affected and 22 discordant sib pairs. We then set α = 0.01 as the MAF cutoff to define the rare variants and calculated weights based on the identified 38 affected sib pairs using the Eq. (4-5).

The 82nd phenotype replicate includes 209 unrelated case subjects and 488 unrelated control subjects. This data set was used to test for association in each of 3,205 genes.

We compared the power of the proposed method with Madsen and Browning's method using the same unrelated case-control data. The same MAF threshold of α = 0.01 was used to define rare variants for Madsen and Browning's method.

We expect there to be little or no power for detecting rare variants using a single replicate because of the small sample size. We therefore included additional replicates until we reached a sample size of 400 affected sib pairs and 2,400 case subjects and 2,400 control subjects. We used 400 affected sib pairs, which was suggested by our previous study [7]. Although the genotypes are the same for the different replicates, the genotype-phenotype association is simulated independently for the different replicates. Thus the way we increased the sample size will have little impact on our power comparison.

Results

We applied the proposed method and Madsen and Browning's method to 3,205 genes using the 82nd replicate. As expected, we found virtually no power for either method.

We next increased the sample size to 400 affected sib pairs to calculate the weights and an additional 2,000 case subjects and 2,000 control subjects for the association test using our proposed method. For comparison, we used 2,400 case subjects and 2,400 control subjects for Madsen and Browning's method. Thus the total sample size is the same for both methods. Figure 1 presents the result for testing 3,205 genes. The horizontal line indicates the 5% significance level after adjusting for 3,205 tests. After correcting for multiple comparisons, we observed 14 and 7 genes reaching significance for the proposed and Madsen and Browning's methods, respectively; 2 of 14 and 1 of 7 significant genes are real causal signals.

We next examined the 15 causal genes generated in the simulations (Table 1). Both methods detected *PICK3C2B*, but our proposed method resulted in a much smaller *p*-value (8.67×10^{-10} vs. 1.48×10^{-5}). Furthermore, the proposed method identified *HSP90AA1* ($p = 8.63 \times 10^{-6}$), which was missed by Madsen and Browning's method.

Discussion and conclusions

In this paper, we propose a novel method to analyze the GAW17 data set. Unlike the existing methods, the proposed method calculates the weights using either affected or discordant sib pairs. The proposed method requires that both the case and control groups and the family members are genotyped for the same set of SNPs or are resequenced in the same region. Compared with Madsen and Browning's method, the proposed method detects the true causal gene *HSP90AA1*, which was

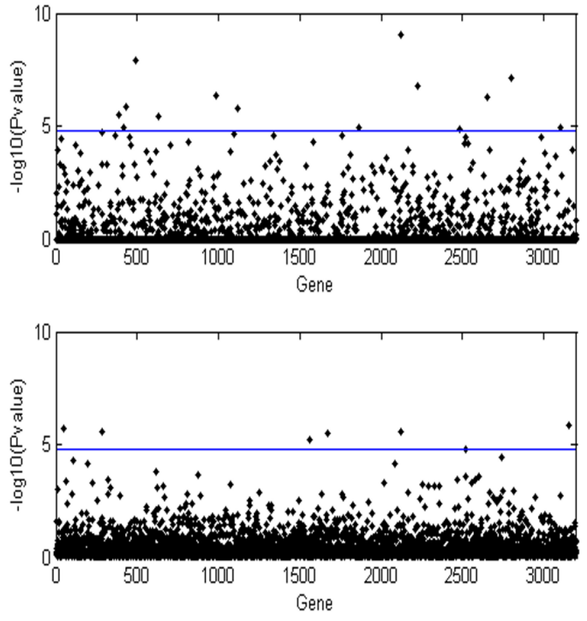

Figure 1 Manhattan plot of $-\log_{10}$(*p*-value) for each single gene after pooling data to reach a larger sample size The upper panel is the result of our proposed method, and the lower panel is the result of Madsen and Browning's method.

Table 1 Causal gene with tested *p*-value

Gene	*p*-value (Madsen and Browning method)	*p*-value (our method)
AKT3	0.173	1.0
BCL2L11	0.0555	1.0
ELAVL4	0.90	0.0283
HSP90AA1	0.737	8.63×10^{-6}
NRAS	0.459	1.0
PIK3C2B	1.48×10^{-5}	8.67×10^{-10}
PIK3C3	0.796	1.0
PIK3R3	0.732	0.236
PRKCA	0.564	1.0
PRKCB1	0.684	0.236
PTK2	0.868	1.0
PTK2B	0.433	1.0
RRAS	0.480	1.0
SHC1	0.925	1.0
SOS2	0.465	1.0

The causal genes of latent liability disease and their *p*-values found using Madsen and Browning's method and our proposed method.

missed by Madsen and Browning's method. Our method demonstrates that incorporating family data can potentially improve statistical power to detect rare variants in an association analysis, which is consistent with the result of Zhu et al. [7].

Acknowledgments
This work was supported by National Institutes of Health grant HL086718 from the National Heart, Lung, and Blood Institute, National Human Genome Research Institute grant HG003054, and National Center for Research Resources grant RR03655.
This article has been published as part of *BMC Proceedings* Volume 5 Supplement 9, 2011: Genetic Analysis Workshop 17. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/5?issue=S9.

Authors' contributions
XZ designed the study. TF performed the data analysis. TF and XZ drafted the manuscript. RCE revised the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that there are no competing interests.

Published: 29 November 2011

References

1. Visscher PM: **Sizing up human height variation.** *Nat Genet* 2008, **40**:489-490.
2. Goldstein DB: **Common genetic variation and human traits.** *New Engl J Med* 2009, **360**:1696-1698.
3. Hirschhorn JN: **Genomewide association studies: illuminating biologic pathways.** *New Engl J Med* 2009, **360**:1699-1701.
4. Bodmer W, Tomlinson I: **Rare genetic variants and the risk of cancer.** *Curr Opin Genet Dev* 2010, **20**:262-267.
5. 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.
6. Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet* 2009, **5**:e1000384.
7. Zhu X, Feng T, Li Y, Lu Q, Elston RC: **Detecting rare variants for complex traits using family and unrelated data.** *Genet Epidemiol* 2010, **32**:171-187.
8. Feng T, Zhu X: **Genome-wide searching of rare genetic variants in WTCCC data.** *Hum Genet* 2010, **128**:269-280.

doi:10.1186/1753-6561-5-S9-S80
Cite this article as: Feng *et al.*: **A novel method to detect rare variants using both family and unrelated case-control data.** *BMC Proceedings* 2011 **5**(Suppl 9):S80.