Open Journal of Statistics, **2011, 1, 15-18**
doi:10.4236/ojs.2011.11002 Published Online April 2011 (http://www.SciRP.org/journal/ojs)

A Gene Score Test for Disease Association with Multiple Genes

Changchun Xie
Population Genomics Program, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada.
E-mail: *xiech@mcmaster.ca*

Received March 14, 2011; *revised March* 30, 2011; *accepted April* 10, 2011

Abstract

The traditional method for creating a gene score to predict a given outcome is to use the most statistically significant single nucleotide polymorphisms (SNPs) from all SNPs which were tested. There are several disadvantages of this approach such as excluding SNPs that do not have strong single effects when tested on their own but do have strong joint effects when tested together with other SNPs. The interpretation of results from the traditional gene score may lack biological insight since the functional unit of interest is often the gene, not the single SNP. In this paper we present a new gene scoring method, which overcomes these problems as it generates a gene score for each gene, and the total gene score for all the genes available. First, we calculate a gene score for each gene and second, we test the association between this gene score and the outcome of interest (i.e. trait). Only the gene scores which are significantly associated with the outcome after multiple testing correction for the number of gene tests (not SNPs) are considered in the total gene score calculation. This method controls false positive results caused by multiple tests within genes and between genes separately, and has the advantage of identifying multi-locus genetic effects, compared with the Bonferroni correction, false discovery rate (FDR), and permutation tests for all SNPs. Another main feature of this method is that we select the SNPs, which have different effects within a gene by using adjustment in multiple regressions and then combine the information from the selected SNPs within a gene to create a gene score. A simulation study has been conducted to evaluate finite sample performance of the proposed method.

Keywords: Gene Score, SNP, GWAS, Permutation, GLM, Multiple Regression, Sum Test

1. Introduction

Due to rapid developments in high-throughput genetic technologies, genome-wide association studies (GWAS) have become common. The success of GWAS depends on genotyping a large number of SNPs (i.e. 500,000 to 1 million) and determining which of these SNPs are significantly associated with the outcome of interest. It is expected that genotyping more SNPs should lead to more accurate gene localization. However, the benefit of the increased number of SNPs is reduced either by multiple testing correction if the SNPs are tested one at a time, or by the increased number of degrees of freedom in the statistical test if multiple regression or haplotype analysis is used. Wang and Elston [1] describe the two possibly conflicting goals of "catching information" and "cutting the cost" of multiple testing or the large number of degrees of freedom. They suggested using a weighted score test (WST) which has only one degree of freedom to achieve the two goals simultaneously. However, Chapman and Whittaker [2] show that, if some of the coded SNPs are positively correlated with the outcome, while others are negatively correlated with the outcome, the WST may have low power. Pan [3] suggested an alternative approach called the sum test, which also has only 1 degree of freedom. Like WST, the sum test has the same problem of sign (i.e. some coded SNPs are positively correlated with the outcome and the others are negatively correlated with the outcome). To overcome the limitation of the sum test, Pan [3] proposed five tests, which are closely related to each other. There is a heuristic solution to the sign problem, discussed by Wang and Elston [1]: before using the WST (or the sum test), one needs to adjust the coding of SNPs so that all SNPs are positively correlated with the outcome. The sign problem is not the only limitation of the sum test. It uses information from all of the SNPs, although the majority of the SNPs might not be associated with the trait, and their use may reduce the power. To increase power and reduce false positive results caused by multiple tests and dependence among test statistics, Gu *et al.* [4] proposed a modified forward multiple regression approach. They chose the SNP with the maximum order statistics in the regression model if its P-value is less than a pre-specified α level and then retained the selected SNP in the regression model and looked for the second SNP among the SNPs with the largest 5% of order statistics in the previous step of regression. Repeating this procedure until no more SNP could be selected. Their simulation studies show that the modified forward multiple regression approach has higher power than the Bonferroni and false discovery rate (FDR) procedure for detecting moderate and weak genetic effects.

For multiple testing correction, the current methods such as Bonferroni correction, FDR and permutation tests, do not consider the situation described in **Figure 1,** where an association pattern of 4 SNPs is provided. All of those methods will choose SNP2 before SNP3 since SNP2 has a smaller P-value than SNP3, although SNP2 might be significant just because it is highly linked with SNP1 and has nothing to do with the trait as shown in **Figure 1**. This predicament motivates the development of the new gene score method.

We propose a method, which controls false positive results caused by multiple tests within genes and between genes separately. First, the SNPs within a gene compete with each other by using Gu e*t al.*'s [4] modified forward multiple regression method, which has more power to detect multiple weak genetic factors than FDR and Bonferroni. We then combine the information from the selected SNPs within a gene to create a gene score, which has only 1 degree of freedom. To avoid the sign problem, we follow an approach, which is similar to Wang and Elston's [1] approach. We adjust the coding of SNPs so that all SNPs are positively associated with the trait. Finally, the genes compete with each other by using gene scores and Bonferroni correction. As shown in **Figure 1**, SNP2 will compete with SNP1 first, instead of competing with SNP3 directly. Since we compare the gene score of gene 1 with that of gene 2, SNP3 in gene 2 might be chosen before SNP1 in gene 1 if the joint effect of gene 2 is stronger than the joint effect of gene 1. This is another advantage of our method compared with current methods including Gu et al's modified forward multiple regression methods for all SNPs.

2. Methods

Let X_{ij} denote the locus score [5], defined as the number of risk alleles (0,1, or 2) for SNP *j* ($j = 1,2, \cdots, L_i$) for gene *i* ($I = 1,2, \cdots, K$) carried by an individual. *L* ($L = L_1 + L_2 + \cdots + L_K$) would denote the total number of SNPs. Suppose the trait value is *Y* and a test is conducted for SNP *j* in gene *i* by using a generalized linear model (GLM) [6]:

$$h\left[E\left(Y\right)\right] = \beta_0 + X_{ij}\beta_1 \qquad (1)$$

where *E*(*Y*) is the mean of *Y*, *h*() is the link function. Typical link functions include the identity link for a continuous normally distributed outcome and a logit link for binary traits. The model can be conditional

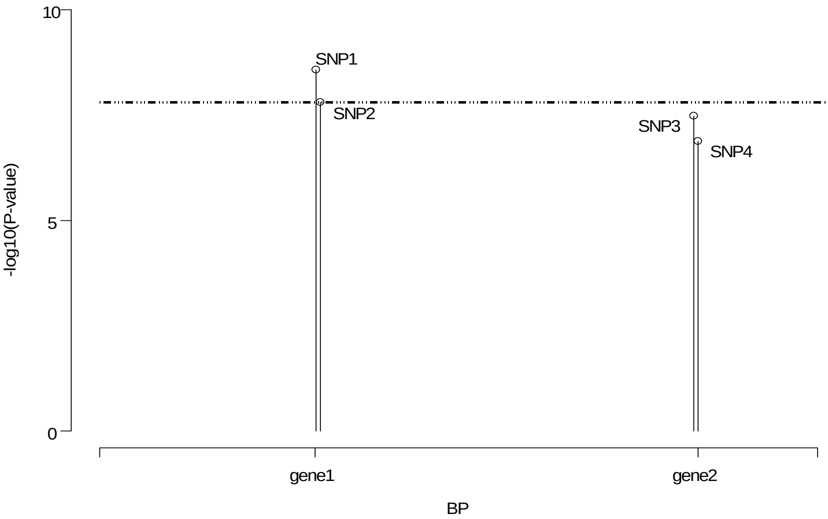

Figure 1. An association pattern of 4 SNPs within 2 genes: SNP2 has smaller P-value than SNP3, however, SNP2 might be significant just because it is in high linkage disequilibrium with SNP1 and has nothing to do with the trait. SNP3 might be more important than SNP2.

GLM for a matched data set or adjusted for some covariates such as age and sex. This test yields P-value p_{ij} (j = 1,2, $\cdots$,L_i, I = 1,2, $\cdots$,K). Let $p_{i(1)} = \min(p_{ij}, j = 1,2, \cdots,L_i)$, whose corresponding SNP is denoted by SNP$_{i(1)}$ and corresponding locus score is denoted by $X_{i(1)}$. If $p_{i(1)}$ > a pre-specified α, the gene score for gene i is 0. If $p_{i(1)}$ ≤ a pre-specified α, we ask whether there is another SNP (in gene i) which is associated with the trait after the effect of SNP$_{i(1)}$ is accounted for by using the model:

$$h\left[E(Y)\right] = \beta_0 + X_{i(1)}\beta_1 + X_{ij}\beta_2 \qquad (2)$$

where SNP j does not include the selected SNP$_{\mathrm{i(1)}}$ and the SNPs with P-value > 0.05. The P-value of this test is denoted by $\mathrm{p_{ij}}^{(2)}$. Let $\mathrm{p_{i(1)}}^{(2)} = \mathrm{min_j}(\mathrm{p_{ij}}^{(2)})$, whose corresponding SNP is denoted by SNP$_{\mathrm{i(2)}}$ and corresponding locus score is denoted by X$_{\mathrm{i(2)}}$. If SNP$_{\mathrm{i(2)}}$ ≤ the pre-specified α, we will search the third SNP (in gene i) which is associated with the trait after the effect of SNP$_{\mathrm{i(1)}}$ and SNP$_{\mathrm{i(2)}}$ is accounted for by using the model:

$$h\left[E(Y)\right] = \beta_0 + X_{i(1)}\beta_1 + X_{i(2)}\beta_2 + X_{ij}\beta_3 \qquad (3)$$

where the SNP j does not include the selected SNP$_{\mathrm{i(1)}}$, SNP$_{\mathrm{i(2)}}$ and the SNPs with P-value > 0.05 in model (2). These steps continue until no further SNP can be found in gene i. Suppose we select SNP$_{\mathrm{i(1)}}$, SNP$_{\mathrm{i(2)}}$, $\cdots$, SNP$_{\mathrm{i(s)}}$ from gene i, the gene score for gene i can be defined as

$$\mathrm{G_i} = X_{\mathrm{i(1)}} + X_{\mathrm{i(2)}} + \cdots + X_{\mathrm{i(s)}} \quad . \qquad (4)$$

Now we focus on the gene instead of SNPs within the gene. In order to obtain the association between the gene and the trait, we use the model

$$h\left[E(Y)\right] = b_0 + G_i b_1 \qquad (5)$$

This model is similar to the SUM test [3], having only one degree of freedom. However, this model does not have the sign problem limitation that the SUM test does, because we use X$_{\mathrm{ij}}$, the locus score, which is the number of risk alleles. Unlike the SUM test, this method uses only selected SNPs to remove noise. To adjust for the multiple testing of multiple genes, we use Bonferroni correction (P-value ≤ α/ K) by assuming there is no linkage disequilibrium (LD) between different genes. Here we consider all genotyped genes, including those genes, whose gene scores = 0 because of no selected SNPs. Suppose there are t gene scores (for example, G$_1$, G$_2$, $\cdots$, G$_{\mathrm{t}}$) that are significant after Bonferroni correction. The total gene score is defined as

$$G = G_1 + G_2 + \cdots + G_t \qquad (6)$$

3. Simulation

To evaluate the performance of the proposed method, we conducted a simulation study to compare power and average number of false positives to detect associated genes for our method, the Bonferroni procedure, the FDR method and modified forward multiple regression method. The simulated data set is generated to have a similar structure to that of the genotype data from the INTERHEART genetics study [7]. On average, there are 15 SNPs per gene and 100 genes in the simulated data set. We picked one gene with 8 SNPs and the other gene with 21 SNPs and then we picked 2 SNPs from each gene. Let the probability of complex disease for the tth subject follow a logistic function

$$P_t = 1 - 1/\left[1 + \exp\left(-2.5 + 0.5\,X_{11t} + 0.3\,X_{12t} + 0.25\,X_{21t} + 0.2\,X_{22t}\right)\right]$$

where X$_{\mathrm{ijt}}$ denotes the number of risk alleles (0,1, or 2) for SNP j (j = 1, 2) for gene i (I = 1, 2) carried by subject t. Using Bernoulli distribution, the disease status of each subject was generated based on the probability of disease for the subject. From the simulated data set, we randomly selected 1000 cases (with disease) and 1000 controls (without disease) to form a data set. To obtain 1000 replicates, the simulated data set was generated 1000 times. From each data set, 1000 cases and 1000 control was randomly selected. We choose α as: 1) α = 0.05/15 (note that 15 is the average number of SNPs per gene); 2) α = 0.05/ # of SNPs within gene. For the first case, α is the same for all genes, while for the second case, α is different for the genes with different numbers of SNPs. The power calculated for detecting each causal gene is the number of times detected in 1000 replicates divided by 1000. The average number of false positives (ANFP) in each replicate is calculated by dividing the number of total false-positive genes found in 1000 replicates by 1000. For Bonferroni, FDR and modified forward multiple regression, there are only tests for SNPs , not for genes. We define a gene “detected” if one or more SNPs in the gene are significant. The results of this simulation are listed in **Table 1** and **Table 2**. Our proposed method (GST) has much higher power than Bonferroni, FDR and modified forward multiple regression.

4. Discussion

In this paper, we have proposed a new method to create a gene score for each gene and then a total gene score for all the genes tested. Compared with the traditional gene score method, this method has the advantage of using SNPs, which may have weak single effects, yet strong joint effects. This method controls false positive results caused by multiple tests within genes and between genes separately, which has the advantage of identifying multi-locus genetic effects, compared with the Bonferroni correction, FDR, and permutation tests for all

Table 1. Power and average number of false positives to detect the associated genes (for GST, α=0.05/the average number of SNPs per gene).

Method	Power (1000 cases and 1000 controls)		ANFP
	Gene 1	Gene 2	
Bonferroni ($\alpha = 6.6 \times 10^{-5}$)	49.5%	15.0%	0.068
FDR ($q = 0.046$)	42.8%	13.1%	0.068
MFMR ($\alpha = 6.4 \times 10^{-5}$)	49.1%	15.1%	0.068
GST ($\alpha = 3.3 \times 10^{-3}$)	54.1%	18.7%	0.068

FDR, false discovery rate; MFMR, modified forward multiple regression; GST, gene score test; ANFP, average number of false positives; q is the controlled q-value level.

Table 2. Power and average number of false positives to detect the associated genes (for GST, α=0.05/the number of SNPs within the gene).

Method	Power (1000 cases and 1000 controls)		ANFP
	Gene 1	Gene 2	
Bonferroni ($\alpha = 4.5 \times 10^{-5}$)	43.7%	13.1%	0.049
FDR ($q = 0.034$)	38.9%	11.4%	0.049
MFMR ($\alpha = 4.4 \times 10^{-5}$)	43.5%	12.5%	0.049
GST (α = 0.05/# of SNPs within gene)	63.3%	15.1%	0.049

FDR, false discovery rate; MFMR, modified forward multiple regression; GST, gene score test; ANFP, average number of false positives; q is the controlled q-value level.

SNPs as shown in Figure 1. Unlike the sum test, which counts all the SNPs (even when the majority of SNPs are not associated with the trait), our method removes these SNPs with a resultant increase in power. Our method can be easily generalized to consider interaction between SNPs within each gene and the interaction between genes. Our method can also be modified to develop a weighted gene score by using genetic effect sizes as the weights if the estimates of the true genetic effect sizes are reliable and accurate. Our simulation shows that our proposed method (GST) has much higher power than FDR to detect associated genes. Further research is required to assess gene scores that include genes, which are not independent because of LD between the genes. One solution might be to combine the genes with strong LD into one gene cluster and then use the gene cluster to replace the genes with strong LD. Another solution might be to use permutation tests instead of Bonferroni correction for gene scores.

5. References

[1] T. Wang and R. C. Elston, "Improved power by use of a weight score test for linkage disequilibrium mapping," The American Journal of Human Genetics, Vol. 80, 2007, pp. 353-360.

[2] J. M. Chapman and J. Whittaker, "Analysis of multiple SNPs in a candidate gene or region," Genetic Epidemiology, Vol. 32, 2008, pp. 560-566.

[3] W. Pan, "Asymptotic tests of association with multiple SNPs in linkage disequilibrium," Genetic Epidemiology, Vol. 33, 2009, pp. 497-507.

[4] X. Gu, R. F. Frankowski, G. L. Rosner, M. Relling, B. Peng and C. I. Amos, "A modified forward multiple regression in high-density genome-wide association studies for complex traits," Genetic Epidemiology, Vol. 33, 2009, pp. 518-525.

[5] S. R. Seaman and B. Muller-Myhsok, "Rapid simulation of P values for product methods and multiple-testing adjustment in association studies," The American Journal of Human Genetics, Vol. 76, 2005, pp. 399-408.

[6] P. McCullagh and J. A. Nelder, "Generalized Linear Models," London: Chapman & Hall, 1983.

[7] S. S. Anand, C. Xie, G. Pare, A. Montpetit, S. Rangarajan, M. J. McQueen, H. J. Cordell, B. Keavney, S. Yusuf, T. J. Hudson and J. C. Engert, "Genetic variants associated with myocardial infarction risk factors in over 8000 individuals from five ethnic groups: the INTERHEART genetic study," Circulation Cardiovascular Genetics, Vol. 2, No. 1, 2009, pp. 16-25.