BMC Proceedings

PROCEEDINGS Open Access

Gene-based Higher Criticism methods for large-scale exonic single-nucleotide polymorphism data

Shiquan He, Zheyang Wu*

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

Abstract

In genome-wide association studies, gene-based methods measure potential joint genetic effects of loci within genes and are promising for detecting causative genetic variations. Following recent theoretical research in statistical multiple-hypothesis testing, we propose to adapt the Higher Criticism procedures to develop novel gene-based methods that use the information of linkage disequilibrium for detecting weak and sparse genetic signals. With the large-scale exonic single-nucleotide polymorphism data from Genetic Analysis Workshop 17, we show that the new Higher-Criticism-type gene-based methods have higher statistical power to detect causative genes than the minimal *P*-value method, ridge regression, and the prototypes of Higher Criticism do.

Background

Genetic factors within genes, pathways, or other physical and functional genome segments often jointly affect the pathogenesis of a disease. Compared with single-locus association analysis in genome-wide association studies, gene-based multiple-loci analysis explores the joint effect of genetic factors related to these functional units and thus likely better reveals the biomedical mechanism of complex traits. Because gene-based methods are potentially more powerful in discovering the complex genetic signals of common diseases, they are considered promising for the study of genetic associations [1]. In recent literature, a number of gene- and pathway-based methods have been studied [2-10]. However, they either assume independent loci [6,9] or have no theoretical justification for ways of considering linkage disequilibrium (LD) [3]. Because of the physical or genetic closeness of the grouped loci, LD is a significant and informative characteristic of genomic segments of genes. LD is particularly significant among genetic factors obtained from next-generation sequence data, for example, rare single-nucleotide polymorphism (SNP) variations located within the exome, which can hardly be replaced by tagging SNPs without losing information. Thus, to better reveal causative genes of complex diseases, investigators strive to properly use the information of LD by developing methods based on a solid statistical theory of multiple-hypothesis testing.

In the latest development of multiple-hypothesis testing theory, the Higher Criticism (HC) and innovated Higher Criticism (iHC) methods have proved to be asymptotically the most efficient hypothesis-testing procedures [11,12]. These methods sustain the lowest boundary for the weakness and sparsity of signals, below which signals are statistically undetectable. This property is important for genome-wide association studies because the unfound genetic factors are likely to be sparse (i.e., true genetic factors are only a small proportion of all considered factors) and weak (i.e., statistical associations of true genetic factors are weak at the population level because of small genetic effects or rare variations). Furthermore, HC-type procedures can incorporate the correlations into test statistics to provide higher statistical power in many circumstances. Based on this celebrated advance of multiple-testing theory, we propose to develop more powerful HC-type gene-based methods that incorporate LD information within genes. Using the large-scale exonic SNP data from Genetic Analysis Workshop 17 (GAW17), we compare the

* Correspondence: zheyangwu@wpi.edu
Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280, USA

prototypes and extensions of HC-type methods with the single-SNP method and ridge regression. The results show that the original HC and iHC methods are not superior to the single-SNP method and ridge regression, but two new methods, eigen-decomposition Higher Criticism (eHC) and innovated Higher Criticism based on marginal estimation of genetic effects (iHCM), show better performance.

Methods

We applied six gene-based methods to analyze the GAW17 data of 697 unrelated individuals from the pilot3 study of the 1000 Genomes Project. Each method was applied to each of the 200 replicates of simulated Q1, Q2, and binary traits. Their ability to detect causative genes was compared on the basis of the average performance over the 200 replicates. Genotype was coded as the number of major alleles at each SNP.

Method 1: minimal *P*-value method

The minimal *P*-value method is a gene detection method that uses single-SNP analysis. The association of individual SNPs in a gene are tested, and the smallest *P*-value (or, equivalently, the test statistic with the largest magnitude) is used to study the significance of that gene. For quantitative traits, the *T*-test statistic from a simple regression model is used. For the binary outcome, a *Z* statistic is used [13].

Method 2: ridge regression

Ridge regression is a natural and powerful approach for testing the association between a response and a group of correlated covariates. Here, we apply ridge regression to detect gene-based association, accommodating LD among SNPs within genes. To our best knowledge, we did not find in the literature any application of ridge regression to testing for association of grouped SNPs. Our simulations (not reported here) show that in many situations ridge regression has a higher statistical power than many LD-incorporating methods proposed in the literature, such as principal components analysis [2,7], linear and quadratic combination tests, and the decorrelation test [3]. To estimate the vector of ridge regression coefficients, we obtain the tuning parameter by minimizing the cross-validation prediction (R function lm.ridge [14]). The sum of the squared residuals describing the goodness-of-fit of the model is treated as the score for each gene. For simplicity, the same procedure is applied to both quantitative and binary traits.

Method 3: Higher Criticism

The Higher Criticism is a multiple comparisons procedure initialized by Tukey [15] and further developed by Donoho and Jin [11]. We adapt this procedure as a gene-based method to detect the significance of the association between a gene and a trait. For genes containing multiple SNPs, the rejection of the association between a gene and a trait is equivalent to the rejection of the joint null hypothesis between multiple SNPs in the gene and the trait. Let $p_{(1)} \le p_{(2)} \le \ldots \le p_{(L)}$ be the ordered *P*-values of *L* individual SNPs in a gene, as described in method 1. The HC statistic for this gene is:

$$\mathrm{HC} = \max_{\left\{ j:\frac{1}{L}\le p_{(j)}\le\frac{1}{2} \right\}} \left\{ \frac{L^{1/2}\left(\frac{j}{L} - p_{(j)} \right)}{\left[p_{(j)}\left(1-p_{(j)}\right) \right]^{1/2}} \right\}. \quad (1)$$

The idea of HC is to find the largest standardized difference between the observed and expected fractions of significance under the joint null hypothesis over a range of significance levels. It provides proper family-wise error control for simultaneously testing multiple hypotheses. Under the assumption of independence among tests, Donoho and Jin [11] proved that HC is optimal in the sense that among all testing procedures, HC provides the lowest boundary for the region of amplitude versus sparsity in which the procedure has asymptotically full statistical power. In the GAW17 data, because most genes have a small number of SNPs, we replace the lower bound of the significance range $1/L$ with 10^{-10}.

Method 4: innovated Higher Criticism

To exploit the potential advantages that correlations among test statistics can offer, Hall and Jin [12] developed innovated Higher Criticism (iHC) to incorporate the information of correlations. They showed that iHC can provide higher power when the correlation matrix is a Toeplitz matrix or when the correlation decays at a polynomial rate along with the distance between the correlated tests. To adapt iHC into the gene-based detection, we use the following procedure:

Step 1. Jointly estimate the vector of regression coefficients for the SNPs in a gene, such that $\hat{\beta} \sim N\left(\beta, \Sigma\right)$. For quantitative traits, the least-squares estimation is applied so that:

$$\hat{\Sigma} = \sigma^2 \left(X'X \right)^{-1}, \quad (2)$$

where *X* is the design matrix of genotypes. Because σ^2 is unique for a specific outcome variable, it will not affect the comparison of the statistics and can thus be treated as having a value of 1. For the binary trait, we fit a multicovariate logistic regression to get the maximum-likelihood estimator of the coefficient vector and its variance matrix.

Step 2. Let:

$$U = (u_{kj})_{\{1\le k,\, j\le n\}} \tag{3}$$

be an inverse matrix (lower triangular) of the Cholesky factorization of $\hat{\Sigma}$, that is, $U\hat{\Sigma}U' = I$. We trim U by making the lower off-diagonals zero if they are equal to or more than b rows away from the central diagonal. Specifically, the resulting matrix is:

$$\tilde{U} = (\tilde{u}_{kj})_{\{1\le k,\, j\le n\}}, \tag{4}$$

where for some bandwidth $1 \le b \le n$, $\tilde{u}_{kj} = u_{kj}$ if $k - b + 1 \le j \le k$, and $\tilde{u}_{kj} = 0$ otherwise. Hall and Jin [12] suggested using use b = log n. Here we choose b as the largest integer equal to or less than log n.

Step 3. We get matrix $\bar{U}$ by normalizing each column of $\tilde{U}$ such that the L2 norm of each column of $\bar{U}$ is 1. Let $V = \bar{U}'U$.

Step 4. The iHC statistic is the HC statistic based on:

$$p_j = P\left\{\left|N(0,1)\right| \ge \left|\left(V\beta\right)_j\right|\right\}, \qquad 1 \le j \le n. \tag{5}$$

Method 5: eigen-decomposition Higher Criticism

The key idea of iHC is to transform the correlated tests into independent tests so that the HC procedure is valid to apply. The transformation process incorporates the correlation information into the new test statistics. Specifically, Cholesky decomposition allows $V\hat{\beta}$ to be approximately $N(V\beta, I)$, where I is an identity matrix. Note that when the trimming bandwidth b = 1, we have $V = U$. Here we consider eigen-decomposition, which can also provide the transformation into independence. The steps are as follows.

Step 1. Estimate the coefficients $\hat{\beta}$ and the corresponding covariance $\hat{\Sigma}$ of the SNPs in a gene as in step 1 of method 4.

Step 2. We take eigen-decomposition:

$$\hat{\Sigma} = \hat{\sigma}^2\left(X'X\right)^{-1} = Q\Lambda Q', \tag{6}$$

$$\hat{\Sigma}^{-1/2} = Q\Lambda^{-1/2}Q', \tag{7}$$

where Q is a matrix whose columns are eigenvectors of $\hat{\Sigma}$ and Λ is a diagonal matrix whose diagonal elements are the eigenvalues corresponding to the eigenvectors. Note that we have:

$$\Sigma^{-1/2}\hat{\beta} \sim N\left(\Sigma^{-1/2}\beta, I\right), \tag{8}$$

which satisfies the assumption of independent normality in the theoretical research of Donoho and Jin [11].

Step 3. The eHC statistic is the HC statistic based on the *P*-values:

$$p_j = P\left\{\left|N(0,1)\right| \ge \left|\left(\hat{\Sigma}^{-1/2}\hat{\beta}\right)_j\right|\right\}, \qquad 1 \le j \le n. \tag{9}$$

Method 6: innovated Higher Criticism based on marginal estimation

Because many SNPs in the GAW17 data have rare minor alleles, the joint estimation of genetic effects by multiple covariate (logistic) regression is unstable (imagine a high-dimensional contingency table of many cells being 0). To accommodate this issue, we studied innovated Higher Criticism based on marginal estimation of genetic effects (iHCM). Specifically, we obtain the test statistics for SNPs in a gene using single-SNP analysis (as in method 1, *T*-test statistics are used for quantitative traits; *Z* statistics are used for the binary disease liability).

For both quantitative and binary traits, we consider two approaches to estimate the correlations among the test statistics. In the first approach, from each data replicate we calculate one set of marginal *T* or *Z* statistics and then combine the 200 sets of test statistics from the 200 replicates to calculate the Pearson correlation coefficients. In the second approach, we approximate the correlation between two test statistics by using the correlation between the genotypes of these two SNPs. The second method is not quite proper, especially under nonnull situations, but it has been widely used in the literature [6,16]. Based on our experience, the first method is slightly better. The results reported later in this paper are based on the first estimation approach. Using $\hat{\beta}$ to denote the marginal *T*- or *Z*-test statistics and $\hat{\Sigma}$ to denote the estimated covariance matrix of $\hat{\beta}$ (because the marginal test statistics are asymptotically standard normal distributed under the null hypothesis), we applied steps 2–4 from method 4 to obtain the iHCM statistic.

Approaches for comparison among genes

We consider two permutation-based approaches to compare the significance of different genes adjusted for the distinct number of SNPs and their correlation structures. In each permutation, we randomly shuffle the response and then calculate all the test statistics from methods 1–6 for all genes. Let S_i and S'_{ij}, i = 1, ..., I, j = 1, ..., J, denote any test statistic from methods 1–6 for the ith gene from the original data and from the jth permutation. The first approach calculates the empirical *P*-values for each gene:

$$p_i = \frac{\text{The size of set } \left\{S'_{ij} \ge S_i,\, j = 1, \ldots, J\right\}}{J}. \tag{10}$$

The second approach normalizes the scores from Eq. (10) to make them directly comparable among genes. Specifically, the normalized scores are:

$$S_i^* = \frac{S_i - \text{mean}(S_i')}{\text{SD}(S_i')}, \quad (11)$$

where $S_i' = (S_{ij}')_{1 \le j \le J}$ is a vector of the test statistics of gene i from permutations. For both approaches, the number of permutations is J = 10,000. To simplify the computation, the statistics obtained from permuting replicate 1 are also used to get the empirical P-values and the normalized scores for other replicates, because the null distributions for all replicates are the same.

Note that our knowledge of the underlying causative genes was used only for the purpose of evaluating the performance of the gene detection methods. The design of the analysis methods does not rely on knowing which genes are true.

Results

The genotype data were cleaned in the following way before analysis. If the vectors of SNP genotypes in a gene were linearly dependent, we removed SNPs with the smallest minor allele frequencies (MAFs) until the remaining SNP genotype vectors became linearly independent. This process avoided overfitting the models and retained the full genetic information of each gene because the deletion did not change the linear space expanded by the SNP genotype data. We analyzed 23,980 SNPs after 507 redundant SNPs (not including any causative SNPs) were removed. Most of the remaining SNPs had very small MAF: Both the minimum and the 1st quantile were 0.000717, the median was 0.002152, the mean was 0.030630, the 3rd quantile was 0.01076, and the maximum was 0.4993.

The number of SNPs and the LD structure within genes are two critical factors for gene-based methods. The GAW17 data contain 3,205 genes. For the distribution of SNP numbers within genes, both the minimum and the 1st quantile were 1, the median was 2, the mean was 7.5, the 3rd quantile was 8, and the maximum was 205. The distribution of the absolute Pearson correlation coefficients between all pairwise SNPs within genes also skewed to the right: The median was 0.008752, the mean was 0.05682, the 3rd quantile was 0.04558, and the maximum was 0.9606. Figure 1 shows the distributions of the correlations between SNP pairs that are located within the same genes and have distances 1 (adjacent), 2 (separated by one SNP), or 3 (separated by two SNPs). As shown in Figure 1, there is an overall trend of correlation decay when the distance between the SNPs increases. To some extent, this observation justifies the application of the iHC procedure, which asymptotically performs well when the correlation decays at a polynomial rate.

We compared the six methods described earlier according to how well they ranked genes. A gene with a smaller empirical P-value (or a larger standardized score) was more statistically significant and thus had a higher rank (i.e., a smaller value of rank). We prefer methods that give higher ranks to causative genes. For the quantitative traits Q1, Q2, and the binary disease liability, the table in additional file 1 lists the average and standard deviation of ranks (over 200 replicates) of causative genes based on empirical P-values. Different methods may provide distinct advantages for different genes. The means and standard deviations for the average ranks over all

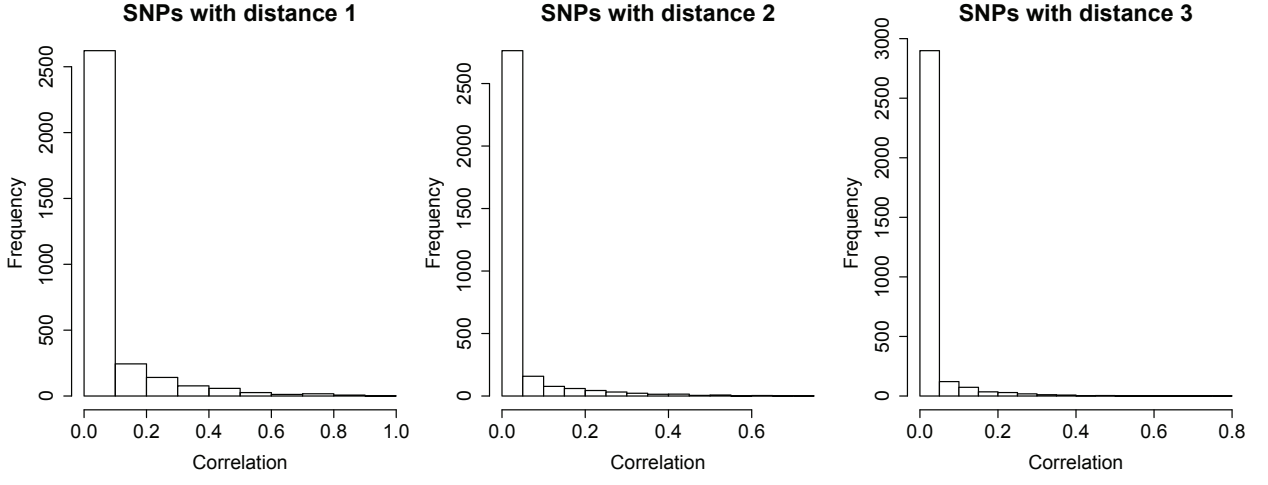

Figure 1 Distribution of correlation magnitude between SNPs. We count the correlation between a pair of SNPs if they are located in the same gene and have a distance 1 (adjacent), 2 (separated by one SNP), or 3 (separated by two SNPs).

causative genes are also given in the table in additional file 1 and provide an overall evaluation for different methods. The smallest average ranks are boldfaced to highlight the best performances. From the table in additional file 1 we can see that the newly developed HC-based methods eHC and iHCM have better performance. Particularly for Q1, Q2, and the binary trait, the iHCM method is uniformly the best. For Q2, the eHC method also performs well. The original HC method [11] and the iHC method [12] are not superior to the minimal *P*-value method and ridge regression. Our results for the comparison of the standardized scores to the ranked genes give similar patterns.

Discussion

For the large-scale exonic SNP data of unrelated individuals and the 200 simulation replicates of quantitative and binary traits, most of the true genes are not ranked at the top. Taking the binary trait as an example, we find that the mean of the average ranks obtained from the minimal *P*-value method is even larger than a value that would be obtained by random choice (≈1,600, or half the total gene number). One of the potential causes is rare variations. In the GAW17 data most of the causative SNPs have very small MAFs ($\ll 0.01$). SNPs with rare variations are difficult to detect for two reasons. First, the statistical signals are weak; and, second, the traditional association measurements, such as those based on regression or logistic regression, do not provide stable estimates, which is evidenced by the large standard deviations for the ranks of true genes over 200 replicates in Table 1.

For the weak signal problem, HC-type methods are ideal for detecting weak and sparse signals. However, the stability problem affects any method developed on the basis of unstable estimations. Some methods that combine rare variations to stabilize the estimations (e.g., weighted-sum methods [17] or collapsing methods [18]) can be applied to improve the HC-type methods.

There are several gaps between GAW17 data and the assumptions of the HC-type methods. First, the theoretical study for HC-type methods is based on asymptotics and assumes a large number of tests in each group. However, the GAW17 data have relatively small numbers of SNPs in genes. Second, the iHC method assumes that the correlation matrix is a Toeplitz matrix or has elements that decay at a polynomial rate. Figure 1 does show an overall trend of decay, but the real correlations in specific genes are much more complicated. We expect that the original iHC method could perform better for those data from genome-wide association studies that are closer to these assumptions when they have more SNPs contained in genes and more consistent correlation decay (e.g., among further separated SNPs).

Conclusions

Gene-based methods that properly incorporate linkage disequilibrium information can improve the detection of causative genes. In Genetic Analysis Workshop 17, we studied the large-scale exonic SNP data of unrelated individuals and the 200 simulation replicates of quantitative traits (Q1, Q2) and the binary trait. Two novel HC-type gene-based methods, eHC and iHCM, have better performances than the minimal *P*-value method (a benchmark), ridge regression (a good method accommodating linkage disequilibrium), and the original HC methods (HC and iHC) in that eHC and iHCM lead to higher average ranks of all causative genes.

Additional material

Additional file 1: Average and standard deviation of ranks of true genes over 200 replicates for Q1, Q2, and the binary trait.

Acknowledgments

We are grateful to the Computing and Communications Center at Worcester Polytechnic Institute for computational support.
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

SH and ZW participated in the design of the study, performed the statistical analysis, drafted the manuscript, read and approved the final manuscript.

Competing interests

The authors declare that there is no competing interest.

Published: 29 November 2011

References

1. Neale BM, Sham PC: **The future of association studies: gene-based analysis and replication.** *Am J Hum Genet* 2004, **75**:353-362.
2. Ballard DH, Cho J, Zhao H: **Comparisons of multi-marker association methods to detect association between a candidate region and disease.** *Genet Epidemiol* 2010, **34**:201-212.
3. Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M: **Genome-wide gene and pathway analysis.** *Eur J Hum Genet* 2010, **18**:1045-1053.
4. Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A: **Association tests using kernel-based measures of multi-locus genotype similarity between individuals.** *Genet Epidemiol* 2010, **34**:213-221.
5. Li M, Wang K, Grant SF, Hakonarson H, Li C: **ATOM: a powerful gene-based association test by combining optimally weighted markers.** *Bioinformatics* 2009, **25**:497-503.
6. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, *et al*: **Gene and pathway-based second wave analysis of genome-wide association studies.** *Eur J Hum Genet* 2010, **18**:111-117.
7. Wang K, Abbott D: **A principal components regression approach to multilocus genetic association studies.** *Genet Epidemiol* 2008, **32**:108-118.
8. Yang HC, Hsieh HY, Fann CS: **Kernel-based association test.** *Genetics* 2008, **179**:1057-1068.
9. Wang K, Li M, Bucan M: **Pathway-based approaches for analysis of genomewide association studies.** *Am J Hum Genet* 2007, **81**:1278-1283.
10. Hoh J, Wille A, Ott J: **Trimming, weighting, and grouping SNPs in human case-control association studies.** *Genome Res* 2001, **11**:2115-2119.
11. Donoho D, Jin J: **Higher criticism for detecting sparse heterogeneous mixtures.** *Ann Stat* 2004, **32**:962-994.

12. Hall P, Jin J: **Innovated Higher Criticism for detecting sparse signals in correlated noise.** *Ann Stat* 2010, **38**:1686-1732.
13. Zuo Y, Zou G, Zhao H: **Two-stage designs in case-control association analysis.** *Genetics* 2006, **173**:1747-1760.
14. R Development Core Team: **R: A language and environment for statistical computing.** Vienna, R Foundation for Statistical Computing; 2010.
15. Tukey JW: **T13 N: The higher criticism.** Course notes, Princeton University; 1976.
16. Chen L, Tong T, Zhao H: **Considering dependence among genes and markers for false discovery control in eQTL mapping.** *Bioinformatics* 2008, **24**:2015-2022.
17. Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet* 2009, **5**:e1000384.
18. Dering C, Pugh E, Ziegler A: **Statistical analysis of rare sequence variants: an overview of collapsing methods.** *Genet Epidemiol* 2011.

doi:10.1186/1753-6561-5-S9-S65
Cite this article as: He and Wu: Gene-based Higher Criticism methods for large-scale exonic single-nucleotide polymorphism data. *BMC Proceedings* 2011 **5**(Suppl 9):S65.