BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 23 2008, pages 2784–2785
doi:10.1093/bioinformatics/btn516

Genetics and population analysis

GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies

Marit Holden[1], Shiwei Deng[2], Leszek Wojnowski[2] and Bettina Kulle[3,4,*]

[1]Norwegian Computing Center, Oslo, Norway, [2]Department of Pharmacology, University of Mainz, Mainz, Germany, [3]Epi-Gen, Faculty Division Akershus University Hospital and [4]Department of Biostatistics, University of Oslo, Oslo, Norway

Received on August 7, 2008; revised and accepted on October 2, 2008

Advance Access publication October 14, 2008

Associate Editor: Martin Bishop

ABSTRACT

The power of genome-wide SNP association studies is limited, among others, by the large number of false positive test results. To provide a remedy, we combined SNP association analysis with the pathway-driven gene set enrichment analysis (GSEA), recently developed to facilitate handling of genome-wide gene expression data. The resulting GSEA-SNP method rests on the assumption that SNPs underlying a disease phenotype are enriched in genes constituting a signaling pathway or those with a common regulation. Besides improving power for association mapping, GSEA-SNP may facilitate the identification of disease-associated SNPs and pathways, as well as the understanding of the underlying biological mechanisms. GSEA-SNP may also help to identify markers with weak effects, undetectable in association studies without pathway consideration. The program is freely available and can be downloaded from our website.

Contact: bkulle@medisin.uio.no

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Genome-wide association studies are currently considered to be the most promising approach for the identification of genes underlying complex traits relevant to the general population, and for fine mapping of complex disease loci (Hirschhorn and Daly, 2005; Risch and Merikangas, 1996). A major unresolved problem is the large number of false positive associations resulting from the high number of statistical tests. This problem needs to be addressed by replication in other cohorts, which are not always available, or through laborious verification by non-clinical experimentation (Wang *et al*., 2005).

Recently, several approaches have been developed to improve the interpretation of the results of genome-wide gene expression analyses. One such approach is supervised clustering, with gene expression changes analyzed for groups of genes with common underlying biology (Curtis *et al*., 2005). An interesting implementation of this principle is the gene set enrichment analysis (GSEA), which focuses on groups of genes that share a biochemical or cellular function, chromosomal location or regulation. GSEA allows for a description of gene expression changes on the level of such signaling pathways or regulatory networks and thus greatly increases their comprehension (Subramanian *et al*., 2005).

We describe a GSEA extension, GSEA-SNP, which applies pathway-driven analysis to genome-wide SNP association studies. The method rests on the assumption that SNPs underlying a disease phenotype are enriched in genes constituting a signaling pathway or those with a common regulation or chromosomal localization. A similar method has recently been proposed by Wang and colleagues (2007). A significant difference is their use of the χ^2-test whereas we employ a more powerful, genotype-based test (the MAX-test) together with a common allele-based test. In addition, Wang *et al*. consider the highest statistical value found among all SNPs mapped to a given gene as the statistical value of the gene. In contrast, we use all SNPs available in the corresponding dataset for a given gene. Furthermore, we address the problem of possible dependency between the SNPs by permutation testing. Most importantly, the code of our program is freely available.

2 ALGORITHM

GSEA-SNP analyses genome-wide SNP data from two classes of samples, such as cases and controls. The following description is very similar to the original description of the GSEA method as applied to gene expression data (Subramanian *et al*., 2005). The main differences are that (i) the original GSEA method compares a *gene list* with several *gene sets*, while GSEA-SNP compares a *SNP list* with several *SNP sets* where each SNP set is obtained from a gene set constituting e.g. a signaling pathway and (ii) the original GSEA method uses a continuous metric to determine the degree of association with the binary phenotype, as appropriate for gene expression data. The GSEA-SNP method utilizes an allele- or genotype-based statistic, which is more appropriate for the categorical SNP data. The recommended genotype-based test is the MAX-test which calculates the maximum test statistics of the three Cochrane–Armitage trend statistics according to different weights with regard to three different ways of inheritance (recessive, dominant and additive) (Freidlin *et al*., 2002). Alternatively, the user may calculate the standard χ^2 allele-based test statistics.

The GSEA-SNP method starts with ranking the SNPs into an ordered list *L*. For this purpose, the appropriate test statistic is computed for each SNP, determining its degree of association with

[*]To whom correspondence should be addressed.

the binary phenotype. The most associated SNPs (i.e. those with low P-values) are situated at the top of the list L.

The following steps are carried out on the list L, and on an a priori defined gene set G_i representing genes, for example, in a common cytogenetic band, in a metabolic pathway or sharing a regulatory motif in the promoter region. From the gene set G_i, a corresponding SNP set S_i is obtained. The assignment of SNPs to genes is done by the user. Different assignments (e.g. dbSNP databank-based or customized) can easily be implemented. S_i contains all SNPs from genes in the G_i set simultaneously contained in the list L. In the following, an enrichment score (ES_i) is calculated for the SNP set S_i. The ES_i value reflects the degree to which a set S_i is over-represented at the top of the list L (see figure in Supplementary Material). ES_i is calculated by screening this list from the top to the bottom and increasing (decreasing) a running-sum statistic when encountering a SNP (not) in S_i. The amount of increase when encountering a SNP in S_i depends on the value of the corresponding test statistics of this SNP, while the amount of decrease for a SNP not in S_i is constant. The constant is chosen such that the running-sum statistic becomes 0 when reaching the end of the ordered list. Next, the nominal significance-level p_i of ES_i is estimated for each SNP set S_i by applying a permutation test to the binary phenotype. Importantly, this approach preserves the correlation between the SNPs from the underlying set. The last step adjusts the significance levels p_i of ES_i for multiple testing. To this end, the enrichment scores ES_i are first normalized according to the size of the SNP set S_i which yields a normalized enrichment score NES_i. The proportion of false positives is controlled by calculating the false discovery rate (FDR) (Ge *et al.*, 2003) for each SNP set S_i. The application of FDR is a natural choice because genome-wide association studies are primarily used to generate hypotheses.

The above approach is called the *SNPs-in-pathways* approach. Alternatively, SNPs can be analyzed in the classic way, i.e. one-by-one, without any a priori defined gene set. The latter procedure consists of an allele- or genotype-based test instead of a full GSEA-SNP procedure, followed by FDR-adjustment, and it is called *SNPs-one-by-one*.

3 ADDITIONAL MATERIAL

The GSEA-SNP method has been implemented in R (www.r-project.org) by extending the original GSEA code. There are two main changes/extensions of the original code: implementation of SNP data-handling procedures and implementation of association tests for SNP data. The gene sets used for testing were downloaded from http://www.broad.mit.edu/gsea/. The results of testing on a set of 52 cases and 52 controls genotyped for almost 11 000 SNPs are provided in the Supplementary Material. The GSEA-SNP program can be downloaded from our webpage http://www.nr.no/pages/samba/area_emr_smbi_gseasnp.

ACKNOWLEDGEMENTS

We thank the GSEA team for the permission to modify the original GSEA-code. We thank Arnoldo Frigessi for useful discussions.

Funding: The Research Council of Norway.

Conflict of Interest: none declared.

REFERENCES

Curtis,R.K. *et al*. (2005) Pathways to the analysis of microarray data. *Trends Biotechnol.*, **23**, 429–435.

Freidlin,B. *et al*. (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum. Hered.*, **53**, 146–152.

Ge,Y.C. *et al*. (2003) Resampling-based multiple testing for microarray data analysis. *Test*, **12**, 1–77.

Hirschhorn,J.N. and Daly,M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.

Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Wang,K. *et al*. (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.

Wang,W.Y.S. *et al*. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109–118.