Genomics 98 (2011) 1–8

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Review

Gene set analysis of genome-wide association studies: Methodological issues and perspectives

Lily Wang [a,*], Peilin Jia [b,c], Russell D. Wolfinger [d], Xi Chen [e], Zhongming Zhao [b,c,f,**]

[a] Department of Biostatistics, Vanderbilt University, Nashville, TN 37232, USA
[b] Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA
[c] Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN 37232, USA
[d] SAS Institute Inc., Cary NC 27513, USA
[e] Division of Cancer Biostatistics, Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA
[f] Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

ARTICLE INFO

Article history:
Received 6 December 2010
Accepted 15 April 2011
Available online 30 April 2011

Keywords:
Genome-wide association study
Gene set
Pathway
Gene-set enrichment analysis
Statistical significance
Complex disease

ABSTRACT

Recent studies have demonstrated that gene set analysis, which tests disease association with genetic variants in a group of functionally related genes, is a promising approach for analyzing and interpreting genome-wide association studies (GWAS) data. These approaches aim to increase power by combining association signals from multiple genes in the same gene set. In addition, gene set analysis can also shed more light on the biological processes underlying complex diseases. However, current approaches for gene set analysis are still in an early stage of development in that analysis results are often prone to sources of bias, including gene set size and gene length, linkage disequilibrium patterns and the presence of overlapping genes. In this paper, we provide an in-depth review of the gene set analysis procedures, along with parameter choices and the particular methodology challenges at each stage. In addition to providing a survey of recently developed tools, we also classify the analysis methods into larger categories and discuss their strengths and limitations. In the last section, we outline several important areas for improving the analytical strategies in gene set analysis.

Contents

1. Introduction

Recently, genome-wide association studies (GWAS), which typically test disease associations with half to a few million single nucleotide polymorphisms (SNPs) across the human genome in hundreds to thousands of samples, have successfully identified many genetic variants contributing to the susceptibilities of complex diseases. However, the variants identified so far, individually or in

* Correspondence to: L. Wang, Department of Biostatistics, Vanderbilt University School of Medicine, S2323 Medical Center North, Nashville, TN 37232, USA. Fax: +1 615 343 4924.
** Correspondence to: Z. Zhao, Department of Biomedical Informatics, Vanderbilt University School of Medicine, 2525 West End Avenue, Suite 600, Nashville, TN 37203, USA. Fax: +1 615 936 8545.
E-mail addresses: lily.wang@vanderbilt.edu (L. Wang), zhongming.zhao@vanderbilt.edu (Z. Zhao).

0888-7543/$ – see front matter
doi:10.1016/j.ygeno.2011.04.006

combination, account for only a small proportion of the inherited component of disease risk [1]. A possible explanation is that due to the large number of genetic polymorphisms examined in GWAS and the massive amount of tests conducted, real but weak associations are likely to be missed after multiple comparison adjustment (e.g., corrected by half a million tests in a typical GWAS).

To help prioritize association signals from GWAS and to better understand the biological themes underlying complex diseases, gene set analysis has become increasingly popular. Instead of conducting analysis for single SNPs or single genes, gene set analysis tests disease association with genetic variants in a group of functionally related genes, such as those belonging to the same biological pathway. One possible cause of complex diseases is the changes in activities of biological pathways: where there are a number of mutations in different genes, each contributes a modest amount to disease predisposition and work together to cause disruptions in normal biological processes.

Current approaches for gene set analysis are still in an early stage of development. When different analysis methods are used, the resulting significant gene sets often vary substantially, even when the same dataset is used [2,3]. One possible reason might be the lack of statistical power in the tests, which are often borrowed from gene set analysis for microarray gene expression data. For many diseases, compared to the amount of differentiation in gene expression levels, effect sizes for SNPs that contribute to disease risk or are in linkage disequilibrium (LD) with the causal variants are typically much smaller. In a recent simulation study [4], we found for gene sets consisting of markers weakly associated with disease (nominal P-value<0.05), all three gene set analysis methods examined – Gene Set Enrichment Analysis (GSEA) [5], Fisher's exact test, and SNP Ratio Test [6] – lacked statistical power for detecting disease associated gene sets. Several recent studies also indicated that gene set analysis results are often prone to sources of bias including gene set size, LD patterns and overlapping genes [3,5,7,8]. Before gene set based approaches are used to draw significant conclusions, the limitations in these methods must be addressed first.

In this review, we discuss the detailed procedures for gene set analysis, along with parameter choices and the particular methodological challenges at each stage. In addition to providing a survey of recently developed tools, we also classify the analysis methods into larger categories and discuss their strengths and limitations. As many new methods are expected to be developed quickly due to the strong demand of initial and secondary (or advanced) analysis of numerous GWAS datasets, our goal is not to provide a comprehensive list of gene set analysis methods. Instead, we aim to provide readers with some of our insights so that they can assess and then use the most appropriate methods for their specific needs. In the last section, we outline several important areas for improving the analytical strategies in gene set analysis. Other recent reviews on gene set analysis of GWAS are Wang et al. (2010) [9] and Cantor et al. (2010) [7].

2. Methodological issues

Fig. 1 outlines the critical steps for assessing statistical significance of disease associations with gene sets: 1) Preprocess data and define the gene sets to be tested, 2) formulate a hypothesis, 3) construct corresponding statistical tests, and 4) assess the statistical significance of the study results. We next discuss each of these steps in order.

2.1. From SNPs to genes

When defining gene boundaries, different criteria (e.g., 500 kb [5], 200 kb [10], 20 kb [11], and 5 kb [12] in both upstream and downstream of the gene coding regions) have been proposed in the literature. Considering LD and gene regulation pattern, investigators often define a gene region to include both the genic region (core part) and the boundary regions (upstream and downstream of the gene). More sophisticated approaches, such as including SNPs that are in LD with the gene, have also been developed [13,14]. These strategies aim to cover SNP markers that play regulatory roles in gene expression and/or link to causal variants within the same LD block. However, these approaches also include more irrelevant SNPs. Thus, they may not only dilute potential signal strength for a gene set but also increase computational burden dramatically, especially for gene sets with a large number of genes. One potentially promising strategy is to take advantage of the information from gene expression studies. Veyrieras et al. [15] estimated that the majority of genetic variants influencing gene expression are located within 20 kb of the genes. Recently, to identify T2D associated pathways, Zhong et al. [16] assessed the impact of the SNPs on gene expressions in liver and adipose tissues and summarized each gene by the SNP significantly associated with the gene's transcript abundance. For general reference, Gamazon et al. [17] developed the SCAN database, which provides information on mapping genetic variants associated with gene expression based on the samples in the HapMap project [18,19]. More comprehensive databases will be developed in the future, for example, those for expression quantitative trait loci (eQTL, regions of the genome that impact gene expression) measured in disease relevant tissues. We expect that utilizing the information from gene expression studies will improve the power of the gene set analysis approach for GWAS.

2.2. From genes to gene sets

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [20] and Gene Ontology (GO) [21] are frequently used gene set annotation databases. When GO terms are used, gene sets categorized into biological process categories have often been selected for gene set analysis, since the other two categories (molecular function and cellular components) are not similar to the typical biological pathways such as those from KEGG. The MSigDB database [22] includes comprehensive gene sets from both the KEGG and GO databases, as well as from other sources such as chromosome and cytogenetic band regions, gene sets collected from expert knowledge in literature, *cis*-regulatory motifs, and co-expressed cancer-associated genes. In addition, other sources such as the PANTHER Classification System [23] and REACTOME [24] also provide publicly available gene set information. Note that GO terms are organized in a hierarchical structure, and substantial overlap of component genes is expected between parent and child nodes. The MSigDB collection has partially solved this problem by removing the gene sets that have the same member genes with their parent nodes or their sibling nodes.

Redundancy among gene sets has often been observed because, by their nature, gene sets such as pathways are biological systems in which a gene may function in multiple ways and thus may appear multiple times in functional gene sets. Although at the systems biology level this reflects the crosstalk between gene sets and the complexity of biological systems, it causes an overlap of member genes and redundant information among gene sets, thus making the results of gene set analysis more difficult to interpret.

Another issue is that gene set annotation is still incomplete. So far, only about 5000 human genes have been annotated to the KEGG pathways, which are most frequently used in the literature. Thus, in gene set analysis of GWAS, all non-annotated genes will be automatically filtered out. A potential improvement is to use protein–protein interaction (PPI) data. As of March 4, 2010, there were approximately 11,000 proteins included in an integrated PPI network analysis platform, Protein Interaction Network Analysis (PINA), which collected and annotated six other public PPI databases (MINT, IntAct, DIP, BioGRID, HPRD, and MIPS/MPact) [25]. This provides much more annotation information about human proteins than does KEGG, and has been used for dense-module searching (DMS) of enriched association signals from one or multiple GWAS datasets [26]. Another advantage in the DMS approach is its flexibility in defining gene set size, which overcomes a potential limitation of the fixed size in KEGG or other biological pathways. However, DMS utilizes the information only from PPIs, rather