Genomics Proteomics Bioinformatics 12 (2014) 210–220

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

REVIEW

Pathway-based Analysis Tools for Complex Diseases: A Review

Lv Jin [1], Xiao-Yu Zuo [2], Wei-Yang Su [3], Xiao-Lei Zhao [1], Man-Qiong Yuan [4], Li-Zhen Han [2], Xiang Zhao [1], Ye-Da Chen [1], Shao-Qi Rao [1,2,4,*]

[1] *Institute for Medical Systems Biology, and Department of Medical Statistics and Epidemiology, School of Public Health, Guangdong Medical College, Dongguan 523808, China*
[2] *Department of Medical Statistics and Epidemiology, School of Public Health, Sun Yat-Sen University, Guangzhou 510080, China*
[3] *Community Health Service Management Center of Panyu District, Guangzhou 511400, China*
[4] *Department of Statistical Sciences, School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou 510275, China*

Received 21 June 2014; revised 30 August 2014; accepted 4 September 2014
Available online 28 October 2014

Handled by Andreas Keller

KEYWORDS

Complex disease;
Pathway-based analysis;
Algorithms;
Software and databases

Abstract Genetic studies are traditionally based on single-gene analysis. The use of these analyses can pose tremendous challenges for elucidating complicated genetic interplays involved in complex human diseases. Modern pathway-based analysis provides a technique, which allows a comprehensive understanding of the molecular mechanisms underlying complex diseases. Extensive studies utilizing the methods and applications for pathway-based analysis have significantly advanced our capacity to explore large-scale omics data, which has rapidly accumulated in biomedical fields. This article is a comprehensive review of the pathway-based analysis methods—the powerful methods with the potential to uncover the biological depths of the complex diseases. The general concepts and procedures for the pathway-based analysis methods are introduced and then, a comprehensive review of the major approaches for this analysis is presented. In addition, a list of available pathway-based analysis software and databases is provided. Finally, future directions and challenges for the methodological development and applications of pathway-based analysis techniques are discussed. This review will provide a useful guide to dissect complex diseases.

Introduction

The etiology for complex human disease is complicated, which involves numerous genes, environmental factors and their interactions [1]. Yet until recently, the genetic basis for most complex diseases has been largely unknown, with just a list of genes identified accounting for very little of the diseases in

* Corresponding author.
E-mail: raoshaoq@gdmc.edu.cn (Rao SQ).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2014.10.002

the population [2]. Genetic approaches that explore the hereditary variants for complex human diseases have significantly changed from family-based linkage studies, which traditionally mapped Mendelian disorders, to population-based association studies, which were aimed at capturing both common and rare variants for the complex diseases. In the last decade, following the International HapMap Project [3], the development of industrial high-throughput genotyping platforms has led to large-scale genome-wide association studies (GWAS), which are now commonly used to determine the genetic basis for the complex human diseases [4,5].

The methods used to analyze large-scale genetic data are significantly behind the rapid advances in the industrial omics technology. Traditional genetic analysis to explore likely single genes or SNPs associated with the disease only identifies a small proportion of the susceptible genetic variants and contributes to a limited understanding of complex diseases. In addition, current popular single-point analysis of GWAS data suffers from the low replication and validation rate [1,6,7]. There is a growing consensus that genetic risk to complex disease is mostly contributed by multiple genes of small or moderate effect factors through their sophisticated interactions acting in a modular fashion, rather than by the mutations of individual genes [5,7]. Hence, to further interpret the underlying molecular mechanisms that cause complex diseases, systematic dissection of the interactions between the individual disease genes as well as their functionalities is essential [6,8].

Pathway-based analysis is an effective technique that overcomes the limitations of the current single-locus methods. This procedure provides a comprehensive understanding of the molecular mechanisms that cause complex diseases [2]. Principally, a pathway-based approach is similar to the Gene Ontology (GO) analysis [9]. However, the pathway analysis is more specific and detailed, and it tests the association between a pathway, which comprises a set of functionally-related genes, and a disease phenotype. Its capacity of capturing biological interaction among genes and improving power and robustness has been well recognized [10,11]. The early application of pathway-based approaches was extended directly from the Gene Set Enrichment Analysis (GSEA) in microarray data analysis [2,12] and now it has evolved in several directions [13,14]. Moreover, varieties of set-based methods with similar ideas have been developed, such as the gene set analysis (GSA) [15], SNP-ratio test (SRT) [16] and LRpath, a logistic regression-based method for pathway (or gene set) analysis [17]. Methods that focus on the original data instead of statistical results have also been developed and these techniques test the joint distribution of the multi-locus data or extract the principal components from the original data, such as in the linear combination test (LCT) [18] and supervised principal component analysis (SPCA) [19]. Recently, some topological methods to parse the internal information of pathway (*e.g.*, signaling pathway impact analysis (SPIA) [20] and CliPPER [21]) have also been developed. In short, pathway-based analysis has gradually become an advanced way to the analysis of complex diseases [22].

With the methodological advance, application of pathway-based analysis to unravel complex human diseases has also entered a new era [23,24]. Several studies have demonstrated that pathway-based analysis is superior when it is applied to large-scale genetic datasets for rheumatoid arthritis (RA) [18,24], type 2 diabetes (T2D) [25], schizophrenia [13], Parkinson's disease [26], *etc*. In addition, tracing the shared pathways among several pathologies tends to be an ongoing interest of disease pleiotropism, for example, the study of genetic links between RA and systemic lupus erythematosus [27], schizophrenia and T2D [28].

This article is a comprehensive review of the pathway-based analysis methods. The general concepts and principles for the pathway-based analysis are introduced and then, a comprehensive review of the major approaches for this type of analysis is presented. In addition, a list of available pathway-based analysis software and databases is provided. Finally, future directions and challenges for the methodological development and applications of pathway-based analysis techniques are discussed.

Pathway-based analysis: general concepts and principles

Currently, there are a variety of pathway-based approaches, which correspond to different research designs and data types. In this article, we focus on SNP/GWAS-derived pathway analysis, but we also include some classical tools for analysis of microarray, as principally they can be easily extended to other data types. Despite some differences in methods for pathway prioritization or null hypotheses to be tested, the basic principle is largely the same, *i.e.*, a pathway-based analysis relies on the use of a testing strategy that targets damaged functionalities, which can produce the outward disease phenotype. It is increasingly recognized that the genetic variations occurring at multiple loci often perturb signal transduction, regulatory and metabolic pathways, resulting in detrimental changes in phenotype [18]. Therefore, pathway-based methods are aimed at analyzing a predetermined aggregation of genes (or SNPs) (alternatively called a gene set) that are contained in a functional unit as defined by prior biological knowledge (*e.g.*, Kyoto Encyclopedia of Genes and Genomes (KEGG), see http://www.genome.jp/kegg/). Depending on whether the individual genotype data or single-point SNP *P* values (often obtained by single-point association test) are used, varieties of methods, such as over-representation analysis (ORA), gene set analysis for 'results' data [29], principal components or regressions for the individual data [19,30]and topology-based analysis [20] (see next section for details), are proposed to combine information from multiple genetic loci within a pathway to assess its overall association with a phenotype.

Compared to single gene analysis methods, pathway-based approaches appear to be well suited for analysis of massive GWAS data, either from biological or statistical considerations [23]. First, since pathway-based approaches focus on sets of genes instead of individual genes, dimension reduction is automatically achieved. Consequently, pathway-based analysis unlikely suffers from the issue of the multiple-test corrections when a large number of SNPs are examined. Second, common diseases often arise from the joint action of multiple SNPs/genes within a pathway. Although each single SNP may confer only a small disease risk, their joint actions are likely to have a significant role in the development of disease. If one only considers the most significant SNPs, the genetic variants that jointly have significant risk effects but make only a small contribution if individually will be missed. Third, locus heterogeneity, in which alleles at different loci cause disease in different populations, will increase the difficulty in replicating