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

associations of a single marker with a disease. The list of significant SNPs from several studies may have little overlap. Therefore, replication of association findings at the SNP level can be difficult if there are redundant genes with similar roles present [18]. In comparison, pathway-based approaches that utilize information from multiple loci in a functional unit could produce more stable and robust results than single gene analyses do [7,31]. Fourth, the ultimate purpose of genetic studies of complex diseases is to decipher the path from genotype to phenotype. In spite of the conduct of extensive studies in search for genes causing complex diseases, connections between DNA variations and complex phenotypes, which are essential for unraveling pathogeneses of complex diseases and predicting variation in human health, have remained elusive. In this sense, pathway-based approaches provide a complementary role to single-point analysis for interpreting the molecular paths underlying human diseases.

Pathway-based analysis methods

According to the strategies for handling the multivariate genetic data for pathways, we classify pathway-based analysis methods into four groups. Instead of individually reviewing each pathway analysis approach, our goal here is to illustrate the algorithms of the representative methods for each group, as shown in **Table 1**, and discuss their relative merits.

Over-representation analysis

ORA, often called functional enrichment analysis, is the earliest pathway-based analysis approach to identify an over-represented pathway with a list of susceptible genes obtained by using traditional statistical tests for contingency tables (*e.g.*, Fisher's exact test, see Table 1) [32]. ORA for SNP data starts by selecting SNPs and mapping the interesting SNPs to the corresponding genes. This initial selection process is based on whether a SNP is mapped to the pathway or whether the SNP is susceptible to the disease [32]. Depending on the results, ORA builds a 2×2 contingency table to conduct a hypergeometric test [32]. The corresponding P value of a given pathway (k_i) is computed by:

$$P(k_i) = 1 - \sum \frac{\binom{M}{n}\binom{N-M}{n-m}}{\binom{N}{n}},$$

where N is the number of all the studied genes, n is the total number of the risk genes, M is the total number of genes in a pathway k_i and m is the number of risk genes contained in pathway k_i.

ORA is the most widely used functional analysis method because it is easily performed. However, it has several limitations. First, ORA considers that each individual gene is of equal importance, which is often not so in biology. Second, the gene (or SNP) list for ORA is usually based on a stringent significant threshold, which can be a salient issue when the number of genes or SNPs analyzed is very large. For GWAS data, statistical power for identifying significant genes or SNPs is limited and therefore the list is often incomplete. Third, construction of the list of statistically significant genes based on the univariate analysis of individual genes does not permit results for genes in a pathway to reinforce each other for detecting an over-represented pathway.

Gene set-based scoring

Gene set-based scoring approaches cover a range of methods that are directly extended from ORA, in which each individual gene is not assumed to be equivalent, instead their importance is ranked by some statistics or P values. Some non-parametric rank sum statistics like Kolmogorov–Smirnov statistic or the Wilcoxon rank sum is used to assess the overall effect of a gene set on a biological phenotype [2,12,33]. The abovementioned ORA is the simplest case in that an equal weight is assumed for all the genes included in a gene set.

The earliest application of a gene-set based scoring approach is the analysis of genome-wide expression profiles [12], in which Subramanian et al. described a powerful analytical method called GSEA for interpreting gene expression data. The method derives its power by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location or regulation. The rationale is: if a set of functionally-related genes (*e.g.*, a module or a pathway) is correlated with a disease phenotype, there is a trend that the set enriches in a certain area of the ranked gene list according to their differential expression between the sample classes [12]. This idea can be directly borrowed for pathway-based analysis of GWAS data by using a ranked SNP/gene list according to their statistical significance in association with a disease phenotype. Then, similarly, an enrichment score (ES), a running-sum statistic, is calculated for a pathway-based gene set, by walking down the list. This statistic reflects the degree to which a gene

Table 1 Algorithms and their applications in pathway-based analysis

Algorithm	Core method	Data types	Refs.
Over-representation analysis	Fisher's exact test	SNP	[32]
Gene set-based scoring	GSEA	Microarray/SNP	[2,12]
	GSA	SNP	[15]
	SRT	SNP	[16]
	LRpath	Microarray	[17]
Multivariate approaches	A two-stage approach	SNP	[18]
	SPCA	SNP	[19]
	Logistic kernel machine regression	Microarray/SNP	[38,39]
Topological-based analysis	SPIA	Microarray	[20,46]
	CliPPER	Microarray	[21]

Note: GSEA, gene set enrichment analysis; GSA, gene set analysis; SRT, SNP-ratio test; SPCA, supervised principal component analysis; SPIA, signaling pathway impact analysis.

set is overrepresented at the extremes (top or bottom) of the entire ranked list. ES is the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov–Smirnov-like statistic. Finally, we can assess significance of the analyzed pathway by using some permutation techniques and adjust for multiple testing, as described previously [34].

Later, Wang et al. [2] extended the aforementioned GSEA to analyze the GWAS data and explicitly formulate the ES computation. Supposed that N genes, which are each represented by a SNP, have their statistical values ranked from the largest to smallest, and the list is denoted by $r_{(1)}, \ldots, r_{(N)}$. A weighted Kolmogorov–Smirnov-like running-sum statistic tests for the over-represented genes within a given gene set S (*e.g.*, a pathway composed of N_{H} genes G_j) and is calculated by:

$$ES(S) = \max_{1 \leqslant i \leqslant N} \left(\sum_{G_j \in S,\ j \leqslant i} \frac{|r(j)^p|}{N_{\mathrm{R}}} - \sum_{G_j \notin S,\ j \leqslant i} \frac{1}{N - N_{\mathrm{H}}} \right),$$

where $N_R = \sum_{G_j \in S} |r(j)|^p$ and p is a parameter that gives more weight to the genes with extreme statistical values [2].

Evidently, in gene set-based scoring approach, each individual gene is no longer considered of equal importance and it uses more information than ORA to analyze pathways. Gene set-based scoring approaches (GSEA and its derivatives) differ from the previous ORA in two important aspects. First, it considers all of the genes/SNPs in an experiment, not only those above an arbitrary cutoff in terms of expression fold-change or association significance. Second, it assesses the significance by permuting the phenotype class labels, which preserves gene–gene correlations and, thus, provides a more accurate null model. Nevertheless, gene set-based scoring approach relies on the results of single-point analysis for each genetic locus, in essence it is a univariate analysis. It does not explicitly configure the sophisticated interplays between genes contained in a pathway.

Multivariate approaches

A two-stage approach

A two-stage approach, proposed by several scientists [18], is aimed at tackling several challenges inherent in the aforementioned ORA and gene set-based scoring approaches. The first challenge is how to represent a gene in GWAS. Wang et al. [2] suggested to choose the most significant SNP from each gene as a representative. But, in GWAS, a gene often contains a variable number of SNPs. The genes that contain a number of SNPs jointly having significant risk effects, but individually making only a small contribution, will be missed in such representation. The second challenge is how to deal with correlations among SNPs and genes. Owing to linkage disequilibrium (LD), there may be high correlations among some SNPs. The statistics that were used by Wang and colleagues [2] for testing association of a pathway with the disease do not take correlations among SNPs into account.

To solve these problems, Luo et al. [18] considered three basic units of association analysis—SNP, gene and pathway—and suggest a two-stage (gene and pathway) GWAS. In gene and pathway-based GWAS, each gene is represented by all SNPs of the gene, which are either located within the gene or are not >500 kb away from the gene. Unlike the aforementioned ORA and gene set-based scoring approaches, in which one examines whether significantly-associated genes are overrepresented in the set of genes to be analyzed, the authors formulated the gene and pathway-based GWAS as the problem to jointly test for association of multiple SNPs within the gene or multiple genes within the pathway with disease. As the proposed two-stage analysis makes full use of the correlation structures between multiple SNPs within a gene or multiple genes within a pathway, it demonstrated better repeatability and reliability in identifying insightful pathways or gene groups related to the development of complex diseases in several large independent genome-wide association studies [18,19,29,35–37].

In order to combine a set of dependent P values of SNPs into an overall significance level for a gene or a set of dependent P values of genes into an overall significance level for a pathway, Luo et al. [18] proposed three novel statistics. These include LCT, quadratic test (QT) and the decorrelation test (DT). LCT takes a linear combination of P values for all SNPs within the gene or a linear combination of statistics for testing association of the genes within the pathway; QT is based on a quadratic form and the test statistic thus follows asymptotically a central χ^2 distribution; finally, in DT, to combine dependent P values, these dependent variables are first transformed into independent variables and then independent variables are combined. For technical details of the three test statistics, readers can consult with the original publication [18].

Supervised principal component analysis

Similarly, to deal with the problems of multicollinearity encountered in a pathway-based analysis, Chen et al. [19] proposed to apply the SPCA model to pathway-based SNP association analysis to test the association between a group of SNPs and variation in disease outcome. The idea behind the SPCA model is that within a biological pathway, genetic variations in a subset of SNPs, each contributing a modest amount to disease predisposition, work together to disrupt normal biological processes. Given a gene set defined by *a priori* knowledge for pathways (*e.g.*, KEGG database), SNPs on an array are first mapped to groups of genes within each pathway. Then a subset of SNPs that is most significantly associated with disease outcome is selected to estimate the latent variable through PCA of this subset. Finally, to identify pathways associated with disease outcome, the authors [19] proposed to test the association between the estimated latent variable and disease outcome using a linear model. In the proposed model, the estimated latent variable is an optimal linear combination of a selected subset of SNPs; therefore, the proposed SPCA model fully utilizes information from both disease-predisposing and disease-protective SNPs in a pathway.

Logistic kernel machine

The logistic kernel machine, first proposed by Liu et al. [38] for analysis of genome-wide expression profiles, may be the first unified approach for multi-dimensional parametric and nonparametric modeling of the pathway effect. This model links the disease to covariates parametrically, and to genes within a genetic pathway nonparametrically using kernel machines.

The nonparametric genetic pathway effect allows for possible interactions among the genes within the same pathway and a complicated relationship between the genetic pathway and the outcome. Later, Wu et al. [39] extended this semiparametric regression model to analyze GWAS data. This type of genome-wide pathway-based analysis proceeds via a two-step procedure. First, SNPs are assigned to SNP sets based on certain meaningful biological criteria (*e.g.*, KEGG pathway classifications or other genomic features). Then, tests for the association between each genomic feature and a disease phenotype are performed using a logistic kernel machine-based multimarker statistic as described below.

For a population-based case-control GWAS in which n independent subjects are genotyped, individual SNPs are first grouped into a SNP set belonging to a pathway (more accurately, genes annotated to these SNPs belong to a pathway). For a pathway-affiliated SNP set containing p SNPs, z_{i1}, z_{i2}, …, z_{ip} are the genotyped values of SNPs for the i^{th} subject ($i = 1, \ldots,n$), and y_i is the disease status for the i^{th} subject ($y_i = 0$ and 1 for controls and cases, respectively). Let x_{i1}, x_{i2}, …, x_{im} denote the values of the covariates and $z_{ij} = 0$, 1 and 2, corresponding to homozygotes for the major allele, heterozygotes and homozygotes for the minor allele, respectively. For the i^{th} individual, the semiparametric model is given by

$$\text{logitP}(y_i = 1) = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_m x_{im} + h(z_{i1}, z_{i2}, \cdots, z_{ip}),$$

where α_0 is an intercept term and α_1, …, α_{m} are regression coefficients of the covariates. The general function $h(\cdot)$ is arbitrary and is defined by a positive, semi-definite kernel function $K(\cdot,\cdot)$, and y_i is influenced by z_{i1}, z_{i2}, …, z_{ip} through $h(\cdot)$ [39].

Choosing an optimal kernel function is the key to analyzing $h(\cdot)$. A desired model can be specified by changing the choice of $K(\cdot,\cdot)$. Essentially, $K(\cdot,\cdot)$ is a function that projects the genotype data from the original space to another space and then $h(\cdot)$ is modeled linearly in this new space, such that if one considers $h(\cdot)$ in the original space, it can be highly nonlinear. More intuitively, however, $K(\cdot,\cdot)$ can be viewed as a function that measures the similarity between two individuals on the basis of the genotypes of the SNPs in the SNP set. The common choices for $K(\cdot,\cdot)$ are the linear, Gaussian and identical-by-state (IBS) kernels. Finally, the genetic effect of the pathway specified by $h(\cdot)$ is tested by a variance component score, which follows a scaled χ^2 distribution.

Topological-based analysis

Recently a new group of methods for pathway-based analysis emerges as topological-based approaches, aiming at explicitly incorporating the dependent structure among genes highlighted by the topology of pathways. Although most methods of this category are developed for analyzing gene expression data, they virtually can be extended to other data types (*e.g.*, GWAS) fairly easily. Unlike the aforementioned approaches that consider only the number of genes and their "expression" in a pathway, topological-based approaches combine the conventionally measured molecular data and also the structural information of the pathway provided by biological databases. A large number of publicly available pathway knowledge bases provide information beyond simple lists of genes for each pathway. These knowledge bases, including KEGG [40], Reactome [41], MetaCyc [42], RegulonDB [43], BioCarta (http://www.biocarta.com) and PantherDB [44], also provide information about gene products that interact with each other in a given pathway, how they interact (*e.g.*, activation or inhibition) and where they interact (*e.g.*, cytoplasm or nucleus).

Topological-based methods are essentially the same as the aforementioned multivariate approaches. The key difference between the two methodological groups is the use of pathway topology to compute gene-level statistics. Massa et al. [45] proposed to use graphical Gaussian models that exploit the graphical evidence of a pathway. This method converts a pathway into a graphical mode and then compares gene sets, which are defined by the pathway. The topological analysis is focused on the strength of the links among genes of a pathway between two phenotypic groups, which is analogous to the aforementioned logical kernel machine modeling. Thus, both the relationship between the genes (*i.e.*, their strength) determined by their topology and experimental data (microarrays and GWAS data) are used to analyze this pathway. Later on, Martini et al. [21] proposed CliPPER, an empirical two-step method, for the identification of significant signal transduction paths within significantly-altered pathways. The initial step uses the aforementioned method to test the entire pathway, which identifies the subgroups of the genes (*i.e.*, signal paths) that makes the entire structure different. The P value of the test determining whether two graphical Gaussian models are homoscedastic as the weight is collected to compute the relevance of each path. Then, a junction tree is reconstructed to identify related pathways with means or covariance matrices that are significantly different between biological statuses.

A recent impact factor analytic approach called SPIA was proposed by Tarca et al. [20,46], which attempts to capture several aspects of the data: changes in gene expression, the pathway enrichment and the topology of signaling pathways. This method considers the structure and dynamics of an entire pathway by incorporating a number of important biological factors, including changes in gene expression, types of interactions and the positions of genes in a pathway. In brief, SPIA models a signaling pathway as a graph, where nodes represent genes and edges represent interactions between them. Furthermore, it defines a gene-level statistic, called perturbation factor (PF) of a gene, as a sum of its measured change in expression and a linear function of the PFs of all genes in a pathway. The impact factor of a pathway (pathway-level statistic) is defined as a sum of PFs of all genes in a pathway.

In our own perception, topological-based analysis is superior to other methods because it also considers the internal structure of the pathway, reflecting its own property of the pathway, *i.e.*, it gains power from pathway topology. An additional advantage is that it has the merit of sound biological interpretation due to the very nature of this methodology. However, one obvious limitation of this analysis is that these methods are largely empirical, thus hard to prove. In addition, there is a dearth of available software and platforms for implementation, although some Bioconductor packages or web tools (*e.g.*, *graphite* [47,48]) are released. Finally, current topological-based methods only handle the static properties of the network topology; thus, they have the inability for a dynamic systems model.

Available resources for pathway-based analysis

Pathway resources have been rapidly accumulated, which has in turn facilitated the development of pathway-based approaches. The term "pathway-based" refers to the basic

analysis unit that is not a gene or a SNP, but a pathway. A pathway is usually defined as a set of related functional genes. The composition of a pathway can be artificially defined or acquired from several public pathway databases. These public databases, include KEGG (website: http://www.genome.jp/kegg/pathway.html) [49], BioCarta and Pathway Interaction Database (PID, website: http://pid.nci.nih.gov), are aimed at providing different pathway repositories based on their functional categorizations (*e.g.*, metabolic pathways and regulatory pathways). Generally, these databases can be grouped into four categories depending on what the researcher wants to emphasize [10]: the metabolic pathway databases, signal transduction pathway databases, protein–protein interaction pathway databases and transcriptional regulation pathway databases, respectively. More detailed descriptions about these categories and the long list of the corresponding databases are presented in **Table 2**.

In the last decade, software and web tools to utilize these databases for pathway-based analysis of omics data were rapidly developed with advances in database construction, and consequently have greatly promoted application of these public resources in the biomedical fields. **Table 3** lists the common software and web platforms mainly used for analysis of microarray data, and **Table 4** lists the available pathway-based analytical tools and web servers designed for pathway analysis on GWAS data. Also, their unique features, functionalities, and databases for annotations are given.

Challenges and future direction

Biological challenges

Pathway-based analysis has significantly enhanced our capacity to explore large-scale omic data, providing an invaluable tool for identifying the damaged functionalities involved in complex diseases. However, we should be cautious on several challenges and limits inherent in this knowledge-guided analysis, which can be divided into two broad categories: (i) knowledge biases and (ii) methodological challenges. Pathway-based analysis relies on acquiring knowledge from gene or pathway databases which provide us varieties of information about genes and how these genes interact with each other. Nevertheless, our knowledge about the existing genes or biological pathways is incomplete [22]. Although a few prominent pathways are well studied, our knowledge of majority of biological pathways (including the limited number of pathways documented in various pathway databases) remains largely fragmented. Because most of pathway-based approaches still rely on mapping multiple SNPs to a single gene, followed by gene-to-pathway mapping, it is often the case in a pathway-based association study that a large number of SNPs or genes fail to be mapped to their corresponding genes or pathways. Apparently, to make full use of biological knowledge to gain power for a pathway-based analysis, we need knowledge bases with high-resolution annotations, in particular for analysis of next-generation sequencing (NGS) data. In addition to the incomplete annotations, many of the existing annotations are of low quality and may be inaccurate. Annotations inferred from indirect evidence (*e.g.*, computationally predicted) are considered to be of lower quality than those derived from direct experimental evidence. Finally, current diagrams for pathways are largely inferred from transcriptional relationships among genes or at protein level, and knowledge for many other regulatory factors (*e.g.*, post transcriptional modifiers, epigenetic factors and environmental triggers) is generally lacking. Hence, for improved power of pathway-based approaches to find damaged functionalities, more integrated databases, which comprehensively review the link of genetic variants, pathways and the environmental triggers, and which jointly analyze their interplays and their contributions to disease, have been called [8,26,50,51]. The Human Genome Epidemiology Network (HuGENet) project is dedicated to this end [52].

Methodological challenges

Methodological challenges for pathway-based approaches in analysis of GWAS or transcriptome have been extensively reviewed in several previous reports [22,53–55]. Hence, we only discuss several key issues here. The first key issue is to improve statistical power for detecting damaged functionalities, which may be achievable by developing more proper weighing schemes for SNPs within a gene or genes within a pathway

Table 2 Pathway databases

Category	Description	Examples	URL
Metabolic pathways	Primarily contain a series of biochemical reactions, especially the chemical modifications of the small molecule substrates of enzymes	KEGG	http://www.genome.ad.jp/kegg/
		BioCyc	http://www.biocyc.org/
		BIOPATH	http://www.mol-net.de/databases/biopath.html
		EMP	http://emp.mcs.anl.gov/
Signal transduction pathways	Describe the information spread from one part or sub-process of the cell to another, generally through a series of covalent modifications of protein	CSNDB	http://geo.nihs.go.jp/csndb/
		SPAD	http://www.grt.kyushu-u.ac.jp/spad/
		TransPath	http://transpath.gbf.de/
		BBID	http://bbid.grc.nia.nih.gov/
Protein–protein interaction pathways	Focus on interactions between proteins, most of which are derived from various large-scale experimental methods	CYGD	http://mips.gsf.de/genre/proj/yeast/index.jsp
		DIP	http://dip.doe-mbi.ucla.edu/
		GRID	http://biodata.mshri.on.ca/grid/servlet/Index
		HPRD	http://www.hprd.org/
Transcriptional regulation pathway	Mainly concern the relationships between transcription factors and the corresponding genes they regulate	STKE	http://www.stke.org/
		BTITE	http://www.genome.ad.jp/brite/
		TRANSFAC	http://transfac.gbf.de/
		CST	http://www.cellsignal.com/

Table 3 The common software and web platforms used for microarray data analysis

Software	Feature	Annotations	URL
PathMAPA	A tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis, based on expression data	Local databases	http://bioinformatics.med.yale.edu/pathmapa.htm
MetaCore	Based on a high-quality, manually-curated database, MetaCore is an integrated software suite for functional analysis of microarray, metabolic, SAGE, proteomics, NGS, copy number variation, siRNA, microRNA and screening data	MetaRodent, MetaLink, MetaSearch	http://www.genego.com/
Ingenuity Pathway Analysis (IPA)	A comprehensive software/database search tool for finding functions and pathways for specific biological states	GO, KEGG, BIND	https://www.ingenuity.com/
ePath3D	An easy-to-use and powerful software for creating and managing illustrated 3D pathways for publications and presentations	eProtein, ePathway	http://www.proteinlounge.com/epath3d/
Pathway Builder	An online pathway drawing tool which is the fastest and easiest method of creating signal transduction pathways, enabling the users to design their own project or use pre-made pathway templates to help get them started	GenBank, Uniprot/Swiss-Prot, TrEMBL, KEGG, ENZYME, *etc.*	http://www.pathwaybuilder.com/
Interactive Pathways Explorer (iPath)	A web-based tool for the visualization, analysis and customization of various pathways maps from KEGG. The recently-released version 2 could deal with metabolic pathway, regulatory pathway and biosynthesis of secondary metabolites	KEGG	https://pathwayexplorer.genome.tugraz.at
GSEA-P & R-GSEA	GSEA-P is a desktop application for Gene Set Enrichment Analysis, with a friendly graphic interface. R-GSEA is provided as a standalone R program.	MSigDB, Gene Set Cards, GEO	http://www.broadinstitute.org/gsea/
DAVID	A tool for augmenting and integrating functional annotations from other databases	KEGG,GO	http://david.abcc.ncifcrf.gov/
MetaCyc	Applications include serving as an encyclopedia of metabolism, providing a reference data set for the computational prediction of metabolic pathways in sequenced organisms, supporting metabolic engineering and helping to compare biochemical networks	KEGG, BioCyc, EcoCyc	http://metacyc.org/
Reactome	Intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education	KEGG	http://www.reactome.org/
GenMAPP	Designed to visualize gene expression and other genomic data on maps representing biological pathways and groupings of genes	GenMAPP, GO	http://www.genmapp.org
FunCluster	An integrative tool for analyzing gene co-expression networks from microarray expression data; the analytic model implemented in the library involves two abstraction layers: transcriptional and functional (biological roles)	GO, KEGG	http://corneliu.henegar.info/FunCluster.htm
Graphite web	A novel web tool for pathway analyses, consisting of topological-based analysis and network visualization for gene expression data of both microarray and RNA-seq experiments	KEGG, Reactome	http://graphiteweb.bio.unipd.it/

Table 4 The available analytical tools and web servers designed for pathway analysis on GWAS data

Software	Feature	Annotations	URL
INRICH	A pathway-based genome-wide association (GWA) analysis tool that tests for enriched association signals of predefined gene sets across independent genomic intervals	KEGG, GO	http://atgu.mgh.harvard.edu/inrich/
GeSBAP	A simple implementation tool of the GSA strategy for the analysis of GWA studies, provides a significant increase in the power testing for this type of studies	GO, BioCarta, KEGG	http://bioinfo.cipf.es/gesbap/
GSEA-SNP	A program for the identification of disease-associated SNPs and pathways, the understanding of the underlying biological mechanisms, and the identification of markers with weak effects, undetectable in association studies without pathway consideration		http://nr.no/pages/samba/area_emr_smbi_gseasnp
GSA-SNP	A standalone software that implements three widely-used GSA methods: Z-statistic method, Restandardized GSA and GSEA for pathway analysis, providing a fast computation and an easy-to-use interface	dbSNP, GO, MSigDB, GeneCards	http://gsa.muldas.org
dmGWAS	Designed to identify significant protein–protein interaction (PPI) modules and the candidate genes, specifically adapted for GWAS datasets, including data preparation, integration, searching and validation in GWAS permutation data	GO,MINT, IntAct, DIP, BioGRID, HPRD,MIPS/MPact	http://bioinfo.mc.vanderbilt.edu/dmGWAS.html
PLINK	A whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally-efficient manner		http://pngu.mgh.harvard.edu/purcell/plink/
Path	A valuable tool for investigating gene–gene interactions in large genetic association studies, designed to help researchers interface their data with biological information from several bioinformatics resources	NCBI, OMIM, KEGG, UCSC dbSNP, IIDB, PharmGKB, *etc*.	http://genapha.icapture.ubc.ca/PathTutorial
Pathway-PDT	A new tool to perform pathway analysis using the framework of Pedigree Disequilibrium Test (PDT) for general family data, combining more information from raw genotypes in general nuclear families	KEGG	https://sourceforge.net/projects/pathway-pdt/
ALIGATOR	A program for testing for Gene Ontology categories over-represented on a list of significant SNPs from a GWA analysis, defining significant SNPs by prespecified *P* value cut-off and then counting significant genes in each pathway	GO, dbSNP, NCBI	http://x004.psycm.uwcm.ac.uk/≃peter/
GenGen	A suite of free software tools to facilitate the analysis of high-throughput genomics data sets	KEGG, BioCarta, GO	http://www.openbioinformatics.org/gengen/
SNP ratio test	Assess significance of enrichment of significant associations in GWA studies and can be applied to pathways such as KEGG/GO or user-defined pathways to test specific hypotheses	KEGG, GO	http://sourceforge.net/projects/snpratiotest/
i-GSEA4GWAS	A tool for SNP label permutation, assign SNPs to genes, calculate modified GSEA enrichment score	KEGG, BioCarta, GO	http://gsea4gwas.psych.ac.cn/
ICSNPathway	A tool for comprehensive interpretation of GWAS data by integrating LD analysis, functional SNP annotation and pathway-based analysis	KEGG, BioCarta, GO	http://icsnpathway.psych.ac.cn/

by leveraging *a priori* biological information from multiple sources or integrating different omic data (*e.g.*, GWAS, microarrays and proteomics). For a pathway-based GWAS, improving SNP coverage and, thus, the number of informative genes may also be a feasible solution. For multivariate approaches, imputation for un-typed SNPs or other covariates is an effective way to increase sample size for improved power. Better refined gene set definitions that group genes according to well-defined biological information or by identifying the signal paths within a pathway using the aforementioned topology-based approaches [20,21,45,46] may also be beneficial. The second key issue is to eliminate or reduce the biases introduced during various steps of a pathway-based analysis. For example, large genes containing many SNPs are more likely to contain significant SNPs by chance alone and so are large pathways containing large genes. Some permutation-based approaches that control for gene size by comparing the actual association data with the distribution of association statistics generated from randomly-permuted data sets are expected to reflect chance-based confounding effects, including biases introduced by gene size [55]. However, the permutation procedures *per se*, if not designed properly, can also introduce some bias [53]. For instance, permutation of SNPs, which is often used in *P* value-based approaches, can disrupt LD patterns between SNPs and may not generate the correct null distribution. For raw genotype-based approaches, permutation of phenotypes (binary traits or quantitative traits) may not generate the correct null distribution either, as explained previously [53]. Furthermore, no matter whether the SNPs or phenotypes are being permuted, the sampling units are assumed to be independent and identically distributed, which may not be the case, as gene–gene interactions may play an important role in disease susceptibility and study participants might be distantly related. The third key issue is to enhance robustness of pathway-based approaches or repeatability of the significance testing results obtained for pathways. One possible cause for poor repeatability of the results in pathway-based analysis is genetic heterogeneity, in which different variants may account for disease status or trait level in different patients. In addition, the different properties of statistical tests on a disease architecture with no major-effect [53] or arbitrary thresholds used to assess significance of a SNP, gene or gene set may also lead to poor repeatability [54]. In conclusion, although some scientists believe that results from testing gene sets rather than from individual markers would be more stable across different samples in the population and, thus, easier to replicate [56,57], enhancing robustness of pathway-based approaches or repeatability of the significance testing results in analysis of pathways still remains to be a difficult task.

Finally, it is worth noting that the recently-developed topological-based methods for analysis of microarrays have the potential to significantly advance the current methods for the pathway-based analysis for GWAS. The classical ORA [58] or GESA [11,12] pay little attention to the contribution of gene or pathway's architecture on their biological activity [8,51], while the multivariate approaches developed so far only consider the statistical correlations between SNPs within a gene or between genes within a pathway. It can be imagined that the underlying genetic architecture for most complex phenotypes is far more complicated than any mathematical models could fully accommodate. For example, the commonly-used Pearson's correlation metric only captures the linear dependence between genes in a pathway, while in reality the biological relationships between genes may not follow linearity. In our own perceptive, incorporating the internal structure or topological properties and environmental triggers into pathway-based analyses is essential for fully assessing the susceptibility of a pathway to a disease. In addition to the merits aforementioned, topology-based approaches have the potential to model and analyze dynamic responses or model effects of external stimuli [22]. All these factors would render topology-based approaches to be a next-generation benchmark methodology for pathway-based analysis of both microarrays and GWAS data.

Extension to analysis of NGS data

Although pathway-based approaches are initially developed for expression microarray data, and later extended to analysis of GWAS data, they may have good potential for a broader application to other omic data like NGS data. Recently, increasing attention is being placed on comprehensive exploration of the genomic variants as high-throughput sequencing has become a feasible solution in practice. The high resolution genome-wide sequence maps provided by these advanced sequencing technologies enable us to examine every detail in sequence variations including SNPs that is used for the traditional GWA studies. In spite of some specific challenges for pathway-based analysis of the massive but more informative sequence data (see the previous review by Wang et al. [53]), NGS data would provide golden opportunities to expand the capacity and power of pathway-based analysis or gene set analysis in general to formulate and test the global hypothesis on disease susceptibility. First, the NGS data could provide necessary coverage to capture all potential genomic subsets [59]. Such refined subsets of candidate genomic regions would allow us to examine all possible genomic structures and their biological roles. Second, the NGS data could provide clues for rare variant discovery. Rare variants and their interactions with common variants and the environment have been shown to contribute to the heterogeneity of several complex diseases [60]. These rare variants may help us to purify the population and to enhance power of a pathway-based analysis on the NGS data. Although the NGS data hold several extended promises, there are several outstanding statistical and practical considerations for performing a pathway-based (or a set-based) analysis of such huge data. To name a few, pathway-based analysis of NGS data requires more efficient statistical methods for detecting genome-wide rare variants, good computational capacity for handling large-scale datasets, improved genotype calling rate for sequencing data and finally the precise functional annotations. Nevertheless, with advent of 1000 Genomes Project [61] and improvements in both high-throughput sequencing techniques and data analytic tools, we believe that it will become feasible to perform sequencing-based GWA studies. It is delightful to see that the first methodological article for pathway analysis on NGS data is published this year. Zhao et al. [62] proposed a novel pathway analysis approach, called smoothed functional principal component analysis (SFPCA), for pathway-based association analysis on NGS data. We anticipate that more and more pathway-based methodologies for dealing with this new data type will appear in the near future. Also, we notice that a *de novo* sequencing technology, known as Single Molecule Real Time

(SMRT) DNA sequencing [63] is invented last year. It has super capacity in fast sequencing DNA, RNA and methylated DNA (10 times faster than current sequencing facilities). As a result, we can foresee that large amount of sequencing data will be available, providing unprecedented opportunities for fully exploring the power of pathway-based approaches for dissecting complex diseases.

Competing interests

The authors declare no conflict of interests.

Acknowledgements

This study was supported in part by the National Natural Science Foundation of China (Grant Nos. 31071166 and 81373085), Natural Science Foundation of Guangdong Province (Grant No. 8251008901000007), Science and Technology Planning Project of Guangdong Province (Grant No. 2009A030301004), Dongguan City Science and Technology Project (Grant No. 2011108101015) and the Guangdong Medical College Funds (Grant Nos. JB1214, XG1001, XZ1105 and STIF201122).

References

[1] Panoutsopoulou K, Zeggini E. Finding common susceptibility variants for complex disease: past, present and future. Brief Funct Genomic Proteomic 2009;8:345–52.
[2] Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 2007;81:1278–83.
[3] International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–61.
[4] Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. Nature 2004;429:446–52.
[5] Freimer NB, Sabatti C. Human genetics: variants in common diseases. Nature 2007;445:828–30.
[6] Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 2009;10:392–404.
[7] Thomas D. Gene–environment-wide association studies: emerging approaches. Nat Rev Genet 2010;11:259–72.
[8] Moore JH, Williams SM. Epistasis and its implications for personal genetics. Am J Hum Genet 2009;85:309–20.
[9] Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, et al. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet 2009;85:13–24.
[10] Curtis RK, Oresic M, Vidal-Puig A. Pathways to the analysis of microarray data. Trends Biotechnol 2005;23:429–35.
[11] Tilford CA, Siemers NO. Gene set enrichment analysis. Methods Mol Biol 2009;563:99–121.
[12] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
[13] Jia P, Wang L, Meltzer HY, Zhao Z. Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. Schizophr Res 2010;122:38–42.
[14] Yang W, de las Fuentes L, Davila-Roman VG, Charles Gu C. Variable set enrichment analysis in genome-wide association studies. Eur J Hum Genet 2011;19:893–900.
[15] Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, Tarraga J, et al. Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. Nucleic Acids Res 2009;37:W340–4.
[16] O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, et al. The SNP ratio test: pathway analysis of genome-wide association datasets. Bioinformatics 2009;25:2762–3.
[17] Sartor MA, Leikauf GD, Medvedovic M. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics 2009;25:211–7.
[18] Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M. Genome-wide gene and pathway analysis. Eur J Hum Genet 2010;18:1045–53.
[19] Chen X, Wang L, Hu B, Guo M, Barnard J, Zhu X. Pathway-based analysis for genome-wide association studies using supervised principal components. Genet Epidemiol 2010;34:716–24.
[20] Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, et al. A systems biology approach for pathway level analysis. Genome Res 2007;17:1537–45.
[21] Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res 2013;41:e19.
[22] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 2012;8:e1002375.
[23] Wilke RA, Mareedu RK, Moore JH. The pathway less traveled: moving from candidate genes to candidate pathways in the analysis of genome-wide data from large scale pharmacogenetic association studies. Curr Pharmacogenomics Person Med 2008;6:150–9.
[24] Giacomelli L, Covani U. Bioinformatics and data mining studies in oral genomics and proteomics: new trends and challenges. Open Dent J 2010;4:67–71.
[25] Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, et al. Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol 2009;33:419–31.
[26] Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, de Andrade M, et al. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet 2007;3:e98.
[27] Lee YH, Bae SC, Choi SJ, Ji JD, Song GG. Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis. Mol Biol Rep 2012;39:10627–35.
[28] Liu Y, Li Z, Zhang M, Deng Y, Yi Z, Shi T. Exploring the pathogenetic association between schizophrenia and type 2 diabetes mellitus diseases based on pathway analysis. BMC Med Genomics 2013;6:S17.
[29] Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet 2010;18:111–7.
[30] Ballard D, Abraham C, Cho J, Zhao H. Pathway analysis comparison using Crohn's disease genome wide association studies. BMC Med Genomics 2010;3:25.
[31] Wu MC, Lin X. Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways. Stat Methods Med Res 2009;18:577–93.
[32] Chen L, Zhang L, Zhao Y, Xu L, Shang Y, Wang Q, et al. Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways. Bioinformatics 2009;25:237–42.

[33] Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 2005;21:1943–9.
[34] Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res 2001;125:279–84.
[35] De la Cruz O, Wen X, Ke B, Song M, Nicolae DL. Gene, region and pathway level analyses in whole-genome studies. Genet Epidemiol 2010;34:222–31.
[36] Hong MG, Pawitan Y, Magnusson PK, Prince JA. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. Hum Genet 2009;126:289–301.
[37] Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, et al. Pathway analysis by adaptive combination of P-values. Genet Epidemiol 2009;33:700–9.
[38] Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics 2008;9:292.
[39] Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 2010;86:929–42.
[40] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
[41] Joshi-Tope G, Vastrik I, Gopinath GR, Matthews L, Schmidt E, Gillespie M, et al. The genome knowledgebase: a resource for biologists and bioinformaticists. Cold Spring Harb Symp Quant Biol 2003;68:237–43.
[42] Karp PD, Riley M, Paley SM, Pellegrini-Toole A. The MetaCyc database. Nucleic Acids Res 2002;30:59–61.
[43] Huerta AM, Salgado H, Thieffry D, Collado-Vides J. RegulonDB: a database on transcriptional regulation in *Escherichia coli*. Nucleic Acids Res 1998;26:55–9.
[44] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003;13:2129–41.
[45] Massa MS, Chiogna M, Romualdi C. Gene set analysis exploiting the topology of a pathway. BMC Syst Biol 2010;4:121.
[46] Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, et al. A novel signaling pathway impact analysis. Bioinformatics 2009;25:75–82.
[47] Sales G, Calura E, Cavalieri D, Romualdi C. Graphite – a Bioconductor package to convert pathway topology to gene network. BMC Bioinformatics 2012;13:20.
[48] Sales G, Calura E, Martini P, Romualdi C. Graphite web: web tool for gene set analysis exploiting pathway topology. Nucleic Acids Res 2013;41:W89–97.
[49] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;32:D277–80.
[50] Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics 2010;26:445–55.
[51] Minguez P, Dopazo J. Functional genomics and networks: new approaches in the extraction of complex gene modules. Expert Rev Proteomics 2010;7:55–63.
[52] Lin BK, Clyne M, Walsh M, Gomez O, Yu W, Gwinn M, et al. Tracking the epidemiology of human genes in the literature: the HuGE Published Literature database. Am J Epidemiol 2006;164:1–4.
[53] Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet 2010;11:843–54.
[54] Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z. Gene set analysis of genome-wide association studies: methodological issues and perspectives. Genomics 2011;98:1–8.
[55] Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends Genet 2012;28:323–32.
[56] Nam D, Kim J, Kim SY, Kim S. GSA-SNP: a general approach for gene set analysis of polymorphisms. Nucleic Acids Res 2010;38:W749–54.
[57] Guo YF, Li J, Chen Y, Zhang LS, Deng HW. A new permutation strategy of pathway-based approach for genome-wide association study. BMC Bioinformatics 2009;10:429.
[58] Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global functional profiling of gene expression. Genomics 2003;81:98–104.
[59] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008;26:1135–45.
[60] Gottlieb B, Beitel LK, Alvarado C, Trifiro MA. Selection and mutation in the ''new'' genetics: an emerging hypothesis. Hum Genet 2010;127:491–501.
[61] Siva N. 1000 Genomes project. Nat Biotechnol 2008;26:256.
[62] Zhao J, Zhu Y, Boerwinkle E, Xiong M. Pathway analysis with next-generation sequencing data. Eur J Hum Genet 2014. http://dx.doi.org/10.1038/ejhg.2014.121.
[63] Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 2013;10:563–9.