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 . . . 1
2. Methodological issues . . . 2
   2.1. From SNPs to genes . . . 2
   2.2. From genes to gene sets . . . 2
   2.3. Formulating hypothesis . . . 3
   2.4. Constructing test statistics . . . 3
   2.5. Potential sources of bias . . . 4
   2.6. Assessing statistical significance . . . 5
3. Several areas for improving gene set analysis of GWAS . . . 6
4. Summary and perspectives . . . 6
Acknowledgments . . . 7
References . . . 7

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 © 2011 Elsevier Inc. All rights reserved.
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

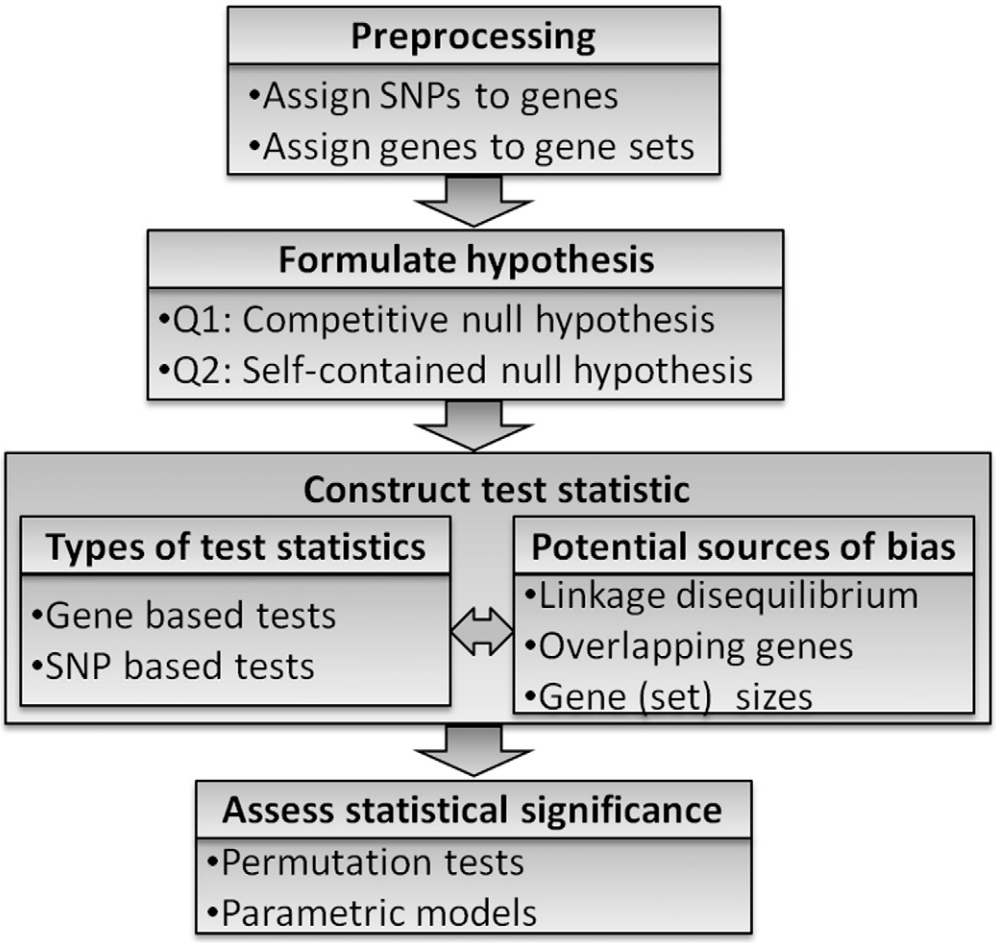

Fig. 1. Work flow for gene set analysis of GWAS datasets.

than from gene regulation as in typical biological pathways. It highlights the degree of incompleteness of our current knowledge about the human genes and their regulation.

2.3. Formulating hypothesis

In the analysis of gene expression data, Tian et al. [27] formulated two statistical hypotheses for testing coordinated association between groups of genes with a phenotype of interest. In the context of GWAS analysis, they are

> Competitive null hypothesis (Q1) — *The genes in a gene-set show the same magnitude of associations with the disease phenotype compared with genes in the rest of the genome;*
> Self-contained null hypothesis (Q2) — *The genes in a gene-set are not associated with the disease phenotype.*

A third null hypothesis (Q3) – *none of the gene sets considered is associated with the phenotype* – has also been proposed recently [28,29]. In contrast to Q1 and Q2, which test for individual gene sets, Q3 tests the entire dataset. For tests of individual gene sets, Goeman and Buhlmann [30] classified tests corresponding to Q1 and Q2 as competitive and self-contained tests, respectively. While a competitive test compares disease association test statistics for genes in the gene set versus that for genes in the rest of the genome, a self-contained test directly tests gene set association with disease and does not depend on genes outside the gene set. Table 1 lists some examples of competitive tests for gene set analysis of GWAS, including GSEA, over-representation analysis based on Fisher's exact test (hypergeometric test) and their extensions such as ALIGATOR [31] and GSA-SNP [32]. Table 2 lists some examples of self-contained tests, including the SNP Ratio Test [6], GRASSS [33] and the SPCA method [12]. When the "real" causal SNPs are fully contained in one particular gene set, testing Q1 and Q2 are approximately the same. However, when SNPs in multiple gene sets are associated with the disease or when causal genes are shared by multiple gene sets, using competitive tests that compare gene set association signals with the rest of the genome may result in loss of power [8,34]. For example, Tintle et al. [35] found the SUMSTAT statistics (based on the MAX–MIN statistic [36]) performed better than GSEA and Fisher's exact test.

2.4. Constructing test statistics

A test statistic can be constructed with units based on either gene or SNP association signals. We refer to them as gene-based and SNP-based methods, respectively. In the former, the *P*-values of the SNPs located within each gene are summarized by gene-level association measures first, and, then, the gene-level *P*-values are used to calculate gene set test scores. The power of these methods mainly depends on the proportion of the genes (for gene-based methods) or SNPs (for SNP-based methods) with strong association signals in the gene set. In practice, several studies reported that gene-based methods may have more power [5,37]; this is because only a few SNPs, which are often located on different genes, may contribute to disease risk (or are in LD with causal variants).

However, in gene-based methods, a consensus has not been reached on the best strategy for SNP information reduction within each gene. A common and simple approach is to represent each gene using the most significant SNP. Since only one SNP *P*-value is used to represent each gene, the potential effects of multiple association signals for the gene

Table 1
Some examples of competitive tests, which compare disease associations for the genes in a gene-set with genes in the rest of the genome.

Reference	Year	Software	Input data for the method	Condense SNP information within each gene	Gene set test statistic	Significance assessment
Gene-based methods						
Wang et al. [5]	2007	GSEA http://www.openbioinformatics.org/gengen	Genotype	Most significant SNP *P*-value; Sime's combination test	Modified Kolmogorov–Smirnov (KS) statistic	Sample permutations
Askland et al. [77]	2009	EVA (Exploratory Visual Analysis) http://www.exploratoryvisualanalysis.org/	SNP *P*-values	Most significant SNP *P*-value	Fisher's exact test	Hypergeometric distribution
Guo et al. [51]	2009		SNP *P*-values	Most significant SNP *P*-value	Modified Kolmogorov–Smirnov statistic	SNP permutations
Holmans et al. [31]	2009	ALIGATOR (Association List Go AnnoTatOR) http://x004.psycm.uwcm.ac.uk/~peter/	SNP *P*-values	Most significant SNP *P*-value with correction for gene size	Modified Fisher's Exact test	Gene re-samplings
Freudenberg et al. [47]	2010		Genotype	Most significant SNP *P*-value	Odds ratio for the presence of SNP associations; number of loci in a category that have SNP associations	Sample permutations
Jia et al. [26]	2010	dmGWAS http://bioinfo.mc.vanderbilt.edu/dmGWAS.html	Genotype	Most significant SNP *P*-value	Z-score	Gene randomization and sample permutations
Luo et al. [70]	2010		SNP *P*-values	Linear combination test; quadratic test; decorrelation test of SNP *P*-values	Linear combination test; quadratic test; decorrelation test of gene *P*-values	Normal or Chi-square distribution
Nam et al. [32]	2010	GSA-SNP http://gsa.muldas.org	SNP *P*-values	Second best SNP *P*-value	Z-statistic, maxmean statistic [36], and modified KS statistic [5]	Gene re-samplings and sample permutations
Peng et al. [41]	2010		SNP *P*-values	Fisher's combined *P*-value; Sidak's correction to the most significant SNP; Sime's combination test; or FDR method	Fisher's exact test	Hypergeometric distribution
Zhang et al. [87]	2010	*i*-GSEA4GWAS http://gsea4gwas.psych.ac.cn/	SNP *P*-values	Most significant SNP *P*-value	Modified Kolmogorov–Smirnov (KS) statistic	SNP permutations
SNP-based methods						
Holden et al. [88]	2008	GSEA-SNP http://nr.no/pages/samba/area_emr_smbi_gseasnp	Genotype		Modified KS statistic	Sample permutations
Schwarz et al. [89]	2008	SNPtoGO http://webtools.imbs.uni-luebeck.de/snptogo	SNP *P*-values		Fisher's exact test	Hypergeometric distribution
Medina et al. [90]	2009	GESBAP (GEne Set Based Analysis of Polymorphisms) http://bioinfo.cipf.es/gesbap/www/index.jsp	SNP *P*-values		Sequential applications of Fisher's exact test on different partitions of the gene list	Hypergeometric distribution corrected by FDR[91]

may be missed. In addition, because longer genes are more likely to have significant *P*-values, this approach may inflate the association test statistic for gene sets that have many long genes. Multiple comparison procedures, such as Sidak's correction [38], Simes' correction [39], or False Discovery Rate (FDR) [40], can be used to adjust the most significant *P*-value (for the number of SNPs located on the gene), but representing genes with the corrected *P*-values may give overly conservative gene set testing results [5,41].

Recently, Ballard et al. [42] compared seven multi-marker association tests, including single marker analysis using the best-scoring SNP, and found that principal component regression [43] is the most powerful among them. In addition, several recent studies also proposed using a subset of SNPs with the lowest *P*-values. The selection of the SNP subset can be based on a fixed truncation point [44–46] or data adaptive thresholds [12]. It has been shown that the SNP selection process can improve power over other approaches that either include all SNPs or use only the most significant SNPs [12,37].

2.5. Potential sources of bias

When scoring gene sets, several sources of potential bias need to be considered:

1) Linkage disequilibrium patterns. Because markers in high LD may originate from a single association signal, an effective strategy may involve down-weighting *P*-values from regions with high LD compared to regions with relatively independent association signals. To this end, strategies have been proposed to group markers in high LD as a "proxy cluster" [8] or use LD blocks from the HapMap database as units of analysis [47] and then assign a single *P*-value for each cluster or LD block.
2) Overlapping genes. Another related and potentially serious problem may result from overlapping genes. When several functionally related genes in a gene set are clustered locally, careful attention should be paid to the SNPs mapped to overlapping genes. When selecting one or more of the most significant SNPs to represent each gene, gene set significance may be driven by only a few of these SNPs, because the significant SNPs mapped to multiple genes could be included multiple times. For example, in our analysis of the GAIN schizophrenia dataset [11], the "starch and sucrose metabolism gene set (HSA00500)" included several genes located closely on the chromosome (e.g., *UGT1A1, UGT1A3, UGT1A4, UGT1A5, UGT1A6, UGT1A7, UGT1A8, UGT1A9, UGT1A10*). When the most significant SNP was used to represent the association signal of each gene, most of the genes in the cluster were represented by the same SNP, which had the *P*-value 6.502×10^{-4}. Therefore, when this SNP has a small *P*-value, the gene set would likely be identified as a significant gene set, while, in fact, the results of multiple significant genes in the gene set were driven by one highly significant SNP located on multiple genes.

Table 2
Some examples of self-contained tests, which test for disease associations for genes in a gene-set directly.

Reference	Year	Software	Input data for the method	Condense SNP information within each gene	Gene set test statistic	Significance assessment
Gene-based methods						
Yu et al. [37]	2009		SNP *P*-values	Adaptive rank truncated product statistic (ARTP) method	Adaptive rank truncated product statistic (ARTP)	An efficient single-level permutation algorithm
Chen et al. [33]	2010	GRASS (Gene set Ridge regression in ASsociation Studies) http://linchen.fhcrc.org/grass.html	Genotype	Principal components		Sample permutations
SNP-based methods						
Dinu et al. [92]	2007		Genotype		U-statistic [93]	Sample permutations
Chai et al. [34]	2009		Genotype		Fisher's combined *P*-value, corrected by Brown's approximation	Chi-square distribution
O'Dushlaine et al. [6]	2009	SNP Ratio Test http://sourceforge.net/projects/snpratiotest/	Genotype		SNP Ratio Test	Sample permutations
De la Cruz et al. [94]	2009		Genotype		Fisher's combined *P*-value, with rank truncation and weights	Sample permutations
Chen et al. [12]	2010		Genotype		Supervised Principal Components	Mixture distribution
Eleftherohorinou et al. [73]	2010		Genotype		Cumulative trend test statistics — sum of all single SNP *P*-values in the gene set	Fit skewed normal distribution to 1000 sample permutations
Ruano et al. [68]	2010		Genotype		Fisher's combined *P*-value	Sample permutations
Wang et al. [55]	2011		SNP *P*-values		*t*-statistic in mixed model	Empirical null distribution

3) Gene set size and gene length. Finally, as mentioned above, in order to score gene sets in an unbiased manner, all selection processes (e.g., selecting the most significant SNPs to represent each gene and selecting the most significant genes to represent each gene set) need to be accounted for in the final gene set analysis. For example, when a gene is represented by the signal of a single SNP from the gene region, the potential effects of multiple association signals for the gene may be missed. Furthermore, because longer genes are more likely to have nominally significant *P*-values, choosing the most significant SNP to represent each gene may inflate the association test statistic for gene sets that have many long genes. Several recent studies assessed the impact of gene length on gene set analysis results and proposed new resampling based strategies for the correction of such bias [48,49].

There are other biases that might affect gene set analysis results including annotation biases from different databases. Two examples are: 1) genes that have been well studied are more thoroughly annotated; and 2) there may be discrepancies of gene set definitions in different databases (e.g., KEGG [20] vs. BioCyc [50]). Furthermore, for enrichment based methods that test the competitive null hypothesis, the choice of the background genes for the enrichment test is a critical factor and a potential bias.

2.6. Assessing statistical significance

To preserve LD patterns, permutations of sample labels are typically employed to establish null distribution of gene set scores. However, several difficulties remain for the application of permutation tests to GWAS.

First, a typical GWAS measures a half million or more SNPs on hundreds or even thousands of samples. The recalculation of a gene set score for each permutation is extremely computationally intensive, especially for competitive tests based on markers from the entire genome. To reduce the amount of computation, several researchers explored assessing gene set significance by resampling genes or SNPs [31,32,51]. It has been suggested that apart from genomic regions that exhibit long range LD (e.g., the Major Histocompatibility Complex (MHC) region), SNPs located on different genes may have little LD [31,33]. Another permutation scheme introduced recently is restandardization, which combines sample label permutation and gene resampling [36,52]. The idea of restandardization is that, while permuting sample labels preserve the correlation structure between genes, the null distribution based on sample permutation approximates the theoretical null distribution $(0,1)$ [53]. However, this distribution ignores the empirical mean and standard deviation of the gene set statistic, which can be approximated more closely by resampling genes. Therefore, for each sample permutation, the mean and standard deviation from gene resampling are used to restandardize the permutation value. Specifically, the restandardized permutation value is computed as $S^{**} = \mu^{+} + \frac{\sigma^{+}}{\sigma^{*}}(S^{*} - \mu^{*})$ where (μ^{+}, σ^{+}) and (μ^{*}, σ^{*}) are the mean and standard deviation of gene set scores obtained from resampling sets of genes or permuting sample labels, respectively [36].

Second, it is not straight forward to model the hierarchical structure in gene sets: SNPs lie within genes, which lie within gene sets using permutation tests. To this end, an efficient algorithm that uses single level permutation iterations to achieve the goal of the multiple-level permutation procedure has been recently proposed [37].

Third, to increase sample size, many GWAS were conducted at multiple study sites, often with different sampling designs. Permutation tests rely on exchangeability of the permuted units. To avoid misleading results, careful consideration is required to account for data structure in complex study designs [54].

An alternative strategy is to employ more flexible parametric models. For GWAS with case–control designs, we have explored modeling disease associations with gene sets using a class of statistical models called mixed effects models [55,56]. In addition to the fixed effects that model the mean structure (e.g., overall association for a group of genes), these models also include random effects that account for variance and covariance structures in the dataset. Future studies include assessing the feasibility of these models for GWAS with more complex designs.

Additionally, Bayesian methods have recently been proposed for genetic association studies [57–60]. These methods can be extended to combine association signals across SNPs and genes in the same pathway. For example, Stephens et al. [61] performed a SNP set analysis for the association between polymorphisms of the *HNF1A* gene and plasma C-reactive protein (CRP) concentration [62] using Bayesian regression approach. This approach was implemented in the software BIMBAM [59].

3. Several areas for improving gene set analysis of GWAS

Although the underlying principle that many functionally related genes collectively contribute to overall disease susceptibility is simple and appealing, as we described earlier, the complexities in GWAS dataset structure raise many technical issues. Several areas of improvement for gene set analysis especially worth noting are as follows.

1) Improve statistical power for detecting disease associated gene sets. Nearly all current methods treat every gene equally when constructing gene set statistics, a more powerful strategy would involve weighing genes and SNPs within a gene set differentially by leveraging a priori biological information, such as that from expression quantitative trait loci (eQTL) studies [16] or network topology [26,63–66].
   In addition, improving SNP coverage and, thus, the number of informative genes may also be beneficial, although it will also increase the computational burden. Holmans et al. [31] performed an imputation analysis for un-typed SNPs using genotype information from the HapMap samples. They demonstrated that the imputation analysis could improve the power for detecting bipolar disorder (BPD) associated gene sets using their ALIGATOR method. Better refined gene set definitions [67,68] that group genes according to well-defined biological information may also be beneficial. For example, Low et al. [67] divided the estrogen metabolic pathway into three sub-pathways involved in androgen synthesis, androgen-to-estrogen conversion and estrogen removal and then found only SNPs within the androgen-to-estrogen conversion pathway were significantly associated with breast and endometrial cancer susceptibilities.
2) Develop strategies for the assessment and comparison of gene set analysis methods. When assessing the performance of a method, it is important to ensure that the proportion of false positive findings from the test is as expected. Null gene sets can be generated by randomly simulating disease outcomes without using any genotype data [55], or by randomly sampling genes from a GWAS dataset [3,4]. Next, one can plot a histogram of the estimated *P*-values for these "null" gene sets. These *P*-values are expected to roughly follow a uniform distribution. It is desirable to have a method whose type I error is equal to or less than the significance cutoff (e.g., 0.05).
   Similarly, to compare the power of different methods, one can randomly sample disease associated genes (with different strengths of associations) from a GWAS dataset or generate disease outcome based on genetic models with various parameters indicating strengths of associations [12,55]. Benchmark GWAS datasets for diseases with well known biological basis, such as Crohn's Disease (CD), would also be useful for evaluating and comparing gene set analysis methods.As an example, Ballard et al. [69] compared two gene set analysis methods based on their applications to three CD datasets. Although most GWAS gene set analyses are discovery projects, careful attention still needs to be paid to guard against spurious findings so that resources can be efficiently allocated to subsequent genotyping, re-sequencing and functional studies.
   As mentioned above, these biases may stem from gene length (the number of SNPs in a gene), gene set size (the number of genes in a gene set), overlapping genes, LD patterns, and population stratifications. In addition, any selection process during data processing (e.g., selecting the most significant SNP to represent each gene) should be accounted for in the final tests. The impact of several potential sources of bias needs to be evaluated for gene set analysis methods. When two or more GWAS datasets are available for the same disease or phenotype, to minimize the bias, we suggest that investigators use one dataset as the discovery dataset and the other (s) as validation dataset(s) [26].
3) Assess the stability of gene set testing results. In addition to power and type I error rate, another important aspect is the stability of the significance testing results. Different sets of samples would give different results due to sampling variations. When different sub-samples from a homogenous population are taken, a method with small variance, and thus stable results across the sub-samples, would be desirable.One strategy is to take sub-samples from all the samples, conduct gene set testing for each subsample, and evaluate the stability of gene set *P*-values based on their changes in rank ordering in different sub-samples. One possible cause for instability of the results in genetic association studies is genetic heterogeneity, in which different variants may account for disease status or trait level in different patients.
   To address this problem, several investigators have hypothesized 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 [31,32,51,70]. More studies are needed to evaluate and test this hypothesis, which has already been validated in gene expression studies [71]. Note that replication and stability assessments are most meaningful when type I error rate for a method is preserved, so applying a method with severe downward biased *P*-values to two datasets would not constitute a valid replication [72].
4) Develop threshold-free procedures. To improve stability of results, one strategy is to develop threshold-free procedures with few, if any, a priori selected parameters. For example, in the commonly used over-representation analysis, a significance threshold is first selected and used to classify whether or not genes are significantly associated with a particular disease, followed by comparing the proportion of disease associated genes in the gene set with the proportion in the rest of the genome by Fisher's exact test. The identification of an optimal threshold is often a difficult task. Holmans et al. [31] suggested that investigators apply a range of cutoff values and then select the cutoff value that gives the most significant increase in over-represented gene sets. A more comprehensive approach, albeit computationally intensive, is to choose a threshold value that could make a reasonable compromise among power, type I error rate, and stability of gene set analysis results using a cross-validation scheme.

4. Summary and perspectives

In summary, recent studies [11,73–84] have repeatedly demonstrated that gene set analysis is a promising approach for analyzing and interpreting GWAS datasets in order to better understand the genetic architecture underlying complex diseases. In this paper, we have provided an up-to-date review of the current progress, as well as the limitations in gene set analysis methods for GWAS. The power and potential performance of these methods may be further improved by integrating additional biological and environmental information at the systems level. For example, network-based approaches that combine association signals in GWAS with local PPI information can help account for gene–gene interactions and identify genes playing central roles in protein networks by interconnecting many disease genes that are weakly associated with disease themselves [26,63–66]. Similarly, analysis that models gene pathways with environmental interactions will help investigators identify novel genes with weak marginal effects that act jointly with exposure factors [85]. As many more GWAS datasets are expected to be generated in the near future,

meta-analyses, which integrate multiple independent GWAS datasets, can be included in gene set analysis methods to increase sample size and power [86]. We hope this review and discussion on the methodological issues on gene set analysis of GWAS will help investigators to find better solutions, understand potential biases, and make gene set analysis more practical and beneficial for understanding genetic variants conferring disease risks in GWAS.

Acknowledgments

We thank two anonymous reviewers for their helpful comments and Rebecca Hiller Posey for critically reading and improving an earlier draft of the manuscript. The work of LW was partially supported by NICHD grant 5P30 HD015052-25 and NIH grant 1P50 MH078028-01A1. The work of XC was partially supported by NCI grant 5P30CA068485-13. The work of ZZ was partially supported by NIH grants R21AA017437, P20AA017828 and R01MH083094, the Vanderbilt-Ingram Cancer Center Core grant P30CA68485, and a 2009 NARSAD Maltz Investigator Award.

References

[1] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, J.H. Cho, A.E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C.N. Rotimi, M. Slatkin, D. Valle, A.S. Whittemore, M. Boehnke, A.G. Clark, E.E. Eichler, G. Gibson, J.L. Haines, T.F. Mackay, S.A. McCarroll, P.M. Visscher, Finding the missing heritability of complex diseases, Nature 461 (2009) 747–753.

[2] C.C. Elbers, Y.T. van der Schouw, C. Wijmenga, N.C. Onland-Moret, Comment on: Perry et al. (2009) interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. Diabetes;58:1463–1467, Diabetes 58 (2009) e9 (author reply e10).

[3] C.C. Elbers, K.R. van Eijk, L. Franke, F. Mulder, Y.T. van der Schouw, C. Wijmenga, N.C. Onland-Moret, Using genome-wide pathway analysis to unravel the etiology of complex diseases, Genet. Epidemiol. 33 (2009) 419–431.

[4] P. Jia, L. Wang, H.Y. Meltzer, Z. Zhao, Pathway-based analysis of GWAS datasets: effective but caution required, Int. J. Neuropsychopharmacol. (2011), doi: 10.1017/S1461145710001446 (Epub ahead of print December 16, 2010).

[5] K. Wang, M. Li, M. Bucan, Pathway-based approaches for analysis of genomewide association studies, Am. J. Hum. Genet. 81 (2007) 1278–1283.

[6] C. O'Dushlaine, E. Kenny, E.A. Heron, R. Segurado, M. Gill, D.W. Morris, A. Corvin, The SNP ratio test: pathway analysis of genome-wide association datasets, Bioinformatics 25 (2009) 2762–2763.

[7] R.M. Cantor, K. Lange, J.S. Sinsheimer, Prioritizing GWAS results: a review of statistical methods and recommendations for their application, Am. J. Hum. Genet. 86 (2010) 6–22.

[8] M.G. Hong, Y. Pawitan, P.K. Magnusson, J.A. Prince, Strategies and issues in the detection of pathway enrichment in genome-wide association studies, Hum. Genet. 126 (2009) 289–301.

[9] K. Wang, M. Li, H. Hakonarson, Analysing biological pathways in genome-wide association studies, Nat. Rev. Genet. 11 (2010) 843–854.

[10] J.R. Perry, M.I. McCarthy, A.T. Hattersley, E. Zeggini, M.N. Weedon, T.M. Frayling, Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach, Diabetes 58 (2009) 1463–1467.

[11] P. Jia, L. Wang, H.Y. Meltzer, Z. Zhao, Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data, Schizophr. Res. 122 (2010) 38–42.

[12] X. Chen, L. Wang, B. Hu, M. Guo, J. Barnard, X. Zhu, Pathway-based analysis for genome-wide association studies using supervised principal components, Genet. Epidemiol. 34 (2010) 716–724.

[13] W.S. Bush, G. Chen, E.S. Torstenson, M.D. Ritchie, LD-spline: mapping SNPs on genotyping platforms to genomic regions using patterns of linkage disequilibrium, BioData Min. 2 (2009) 7.

[14] M.G. Hong, Y. Pawitan, P.K. Magnusson, J.A. Prince, Strategies and issues in the detection of pathway enrichment in genome-wide association studies, Hum. Genet. 126 (2009) 289–301.

[15] J.B. Veyrieras, S. Kudaravalli, S.Y. Kim, E.T. Dermitzakis, Y. Gilad, M. Stephens, J.K. Pritchard, High-resolution mapping of expression-QTLs yields insight into human gene regulation, PLoS Genet. 4 (2008) e1000214.

[16] H. Zhong, X. Yang, L.M. Kaplan, C. Molony, E.E. Schadt, Integrating pathway analysis and genetics of gene expression for genome-wide association studies, Am. J. Hum. Genet. 86 (2010) 581–591.

[17] E.R. Gamazon, W. Zhang, A. Konkashbaev, S. Duan, E.O. Kistner, D.L. Nicolae, M.E. Dolan, N.J. Cox, SCAN: SNP and copy number annotation, Bioinformatics 26 (2010) 259–262.

[18] D.M. Altshuler, R.A. Gibbs, L. Peltonen, E. Dermitzakis, S.F. Schaffner, F. Yu, P.E. Bonnen, P.I. de Bakker, P. Deloukas, S.B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, K. Chang, A. Hawes, L.R. Lewis, Y. Ren, D. Wheeler, D.M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J.M. Korn, K. Kristiansson, C. Lee, S.A. McCarrol, J. Nemesh, A. Keinan, S.B. Montgomery, S. Pollack, A.L. Price, N. Soranzo, C. Gonzaga-Jauregui, V. Anttila, W. Brodeur, M.J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, Q. Zhang, M.J. Ghori, R. McGinnis, W. McLaren, F. Takeuchi, S.R. Grossman, I. Shlyakhter, E.B. Hostetter, P.C. Sabeti, C.A. Adebamowo, M.W. Foster, D.R. Gordon, J. Licinio, M.C. Manca, P.A. Marshall, I. Matsuda, D. Ngare, V.O. Wang, D. Reddy, C.N. Rotimi, C.D. Royal, R.R. Sharp, C. Zeng, L.D. Brooks, J.E. McEwen, Integrating common and rare genetic variation in diverse human populations, Nature 467 (2010) 52–58.

[19] K.A. Frazer, D.G. Ballinger, D.R. Cox, D.A. Hinds, L.L. Stuve, R.A. Gibbs, J.W. Belmont, A. Boudreau, P. Hardenbol, S.M. Leal, S. Pasternak, D.A. Wheeler, T.D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, J. Zhou, S.B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R.C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, W. Sun, H. Wang, Y. Wang, X. Xiong, L. Xu, M.M. Waye, S.K. Tsui, H. Xue, J.T. Wong, L.M. Galver, J.B. Fan, K. Gunderson, S.S. Murray, A.R. Oliphant, M.S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.F. Olivier, M.S. Phillips, S. Roumy, C. Sallee, A. Verner, T.J. Hudson, P.Y. Kwok, D. Cai, D.C. Koboldt, R.D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.C. Tsui, W. Mak, Y.Q. Song, P.K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, et al., A second generation human haplotype map of over 3.1 million SNPs, Nature 449 (2007) 851–861.

[20] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, Y. Yamanishi, KEGG for linking genomes to life and the environment, Nucleic Acids Res. 36 (2008) D480–D484.

[21] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (2000) 25–29.

[22] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A 102 (2005) 15545–15550.

[23] H. Mi, Q. Dong, A. Muruganujan, P. Gaudet, S. Lewis, P.D. Thomas, PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium, Nucleic Acids Res. 38 (2010) D204–D210.

[24] I. Vastrik, P. D'Eustachio, E. Schmidt, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney, L. Stein, Reactome: a knowledge base of biologic pathways and processes, Genome Biol. 8 (2007) R39.

[25] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T.P. Makela, S. Hautaniemi, Integrated network analysis platform for protein–protein interactions, Nat. Methods 6 (2009) 75–77.

[26] P. Jia, S. Zheng, J. Long, W. Zheng, Z. Zhao, dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks, Bioinformatics 27 (2011) 95–102.

[27] L. Tian, S.A. Greenberg, S.W. Kong, J. Altschuler, I.S. Kohane, P.J. Park, Discovering statistically significant pathways in expression profiling studies, Proc Natl Acad Sci U S A 102 (2005) 13544–13549.

[28] I. Dinu, J.D. Potter, T. Mueller, Q. Liu, A.J. Adewale, G.S. Jhangri, G. Einecke, K.S. Famulski, P. Halloran, Y. Yasui, Gene-set analysis and reduction, Brief. Bioinform. 10 (2009) 24–34.

[29] D. Nam, S.Y. Kim, Gene-set approach for expression pattern analysis, Brief. Bioinform. 9 (2008) 189–197.

[30] J.J. Goeman, P. Buhlmann, Analyzing gene expression data in terms of gene sets: methodological issues, Bioinformatics 23 (2007) 980–987.

[31] P. Holmans, E.K. Green, J.S. Pahwa, M.A. Ferreira, S.M. Purcell, P. Sklar, M.J. Owen, M.C. O'Donovan, N. Craddock, Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder, Am. J. Hum. Genet. 85 (2009) 13–24.

[32] D. Nam, J. Kim, S.Y. Kim, S. Kim, GSA-SNP: a general approach for gene set analysis of polymorphisms, Nucleic Acids Res. 38 (Suppl) (2010) W749–W754.

[33] L.S. Chen, C.M. Hutter, J.D. Potter, Y. Liu, R.L. Prentice, U. Peters, L. Hsu, Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data, Am. J. Hum. Genet. 86 (2010) 860–871.

[34] H.S. Chai, H. Sicotte, K.R. Bailey, S.T. Turner, Y.W. Asmann, J.P. Kocher, GLOSSI: a method to assess the association of genetic loci-sets with complex diseases, BMC Bioinformatics 10 (2009) 102.

[35] N.L. Tintle, B. Borchers, M. Brown, A. Bekmetjev, Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16, BMC Proc. 3 (Suppl 7) (2009) S96.

[36] B. Efron, R.J. Tibshirani, On testing the significance of sets of genes, Ann. Appl. Stat. 1 (2007) 107–129.

[37] K. Yu, Q. Li, A.W. Bergen, R.M. Pfeiffer, P.S. Rosenberg, N. Caporaso, P. Kraft, N. Chatterjee, Pathway analysis by adaptive combination of P-values, Genet. Epidemiol. 33 (2009) 700–709.

[38] Z. Sidak, Rectangular confidence regions for the means of multivariate normal distributions, J. Am. Stat. Assoc. 62 (1967) 626–633.

[39] R.J. Simes, An improved Bonferroni procedure for multiple tests of significance, Biometrika 73 (1986) 751–754.

[40] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. B 57 (1995) 289–300.

[41] G. Peng, L. Luo, H. Siu, Y. Zhu, P. Hu, S. Hong, J. Zhao, X. Zhou, J.D. Reveille, L. Jin, C.I. Amos, M. Xiong, Gene and pathway-based second-wave analysis of genome-wide association studies, Eur. J. Hum. Genet. 18 (2010) 111–117.

[42] D.H. Ballard, J. Cho, H. Zhao, Comparisons of multi-marker association methods to detect association between a candidate region and disease, Genet. Epidemiol. 34 (2010) 201–212.

[43] K. Wang, D. Abbott, A principal components regression approach to multilocus genetic association studies, Genet. Epidemiol. 32 (2008) 108–118.

[44] J. Hoh, A. Wille, J. Ott, Trimming, weighting, and grouping SNPs in human case–control association studies, Genome Res. 11 (2001) 2115–2119.
[45] F. Dudbridge, B.P. Koeleman, Rank truncated product of P-values, with application to genomewide association scans, Genet. Epidemiol. 25 (2003) 360–366.
[46] D.V. Zaykin, L.A. Zhivotovsky, P.H. Westfall, B.S. Weir, Truncated product method for combining P-values, Genet. Epidemiol. 22 (2002) 170–185.
[47] J. Freudenberg, A.T. Lee, K.A. Siminovitch, C.I. Amos, D. Ballard, W. Li, P.K. Gregersen, Locus category based analysis of a large genome-wide association study of rheumatoid arthritis, Hum. Mol. Genet. 19 (2010) 3863–3872.
[48] N. Bonifaci, B. Gorski, B. Masojc, D. Wokolorczyk, A. Jakubowska, T. Debniak, A. Berenguer, J. Serra Musach, J. Brunet, J. Dopazo, S.A. Narod, J. Lubinski, C. Lazaro, C. Cybulski, M.A. Pujana, Exploring the link between germline and somatic genetic alterations in breast carcinogenesis, PLoS One 5 (2010) e14078.
[49] P. Jia, J. Tian, Z. Zhao, Assessing gene length biases in gene set analysis of genome-wide association studies, Int. J. Comput. Biol. Drug Des. 3 (2011) 297–310.
[50] P.D. Karp, C.A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin, N. Lopez-Bigas, Expansion of the BioCyc collection of pathway/genome databases to 160 genomes, Nucleic Acids Res. 33 (2005) 6083–6089.
[51] Y.F. Guo, J. Li, Y. Chen, L.S. Zhang, H.W. Deng, A new permutation strategy of pathway-based approach for genome-wide association study, BMC Bioinformatics 10 (2009) 429.
[52] M. Ackermann, K. Strimmer, A general modular framework for gene set enrichment analysis, BMC Bioinformatics 10 (2009) 47.
[53] B. Efron, Microarrays, empirical Bayes, and the two-groups model, Stat. Sci. 23 (2008) 1–47.
[54] G.A. Churchill, R.W. Doerge, Naive application of permutation testing leads to inflated type I error rates, Genetics 178 (2008) 609–610.
[55] L. Wang, P. Jia, R.D. Wolfinger, X. Chen, B.L. Grayson, T.M. Aune, Z. Zhao, An efficient hierarchical generalized linear mixed model for testing disease association with biological pathways in genome-wide association studies, Bioinformatics 27 (2011) 686–692.
[56] C.E. McCulloch, S.R. Searle, Generalized, Linear and Mixed Models, John Wiley & Sons, Inc., 2001
[57] D.J. Lunn, J.C. Whittaker, N. Best, A Bayesian toolkit for genetic association studies, Genet. Epidemiol. 30 (2006) 231–247.
[58] J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, A new multipoint method for genome-wide association studies by imputation of genotypes, Nat. Genet. 39 (2007) 906–913.
[59] B. Servin, M. Stephens, Imputation-based analysis of association studies: candidate regions and quantitative traits, PLoS Genet. 3 (2007) e114.
[60] J. Wakefield, A Bayesian measure of the probability of false discovery in genetic epidemiology studies, Am. J. Hum. Genet. 81 (2007) 208–227.
[61] M. Stephens, D.J. Balding, Bayesian statistical methods for genetic association studies, Nat. Rev. Genet. 10 (2009) 681–690.
[62] A.P. Reiner, M.J. Barber, Y. Guan, P.M. Ridker, L.A. Lange, D.I. Chasman, J.D. Walston, G.M. Cooper, N.S. Jenny, M.J. Rieder, J.P. Durda, J.D. Smith, J. Novembre, R.P. Tracy, J.I. Rotter, M. Stephens, D.A. Nickerson, R.M. Krauss, Polymorphisms of the HNF1A gene encoding hepatocyte nuclear factor-1 alpha are associated with C-reactive protein, Am. J. Hum. Genet. 82 (2008) 1193–1201.
[63] S.E. Baranzini, N.W. Galwey, J. Wang, P. Khankhanian, R. Lindberg, D. Pelletier, W. Wu, B.M. Uitdehaag, L. Kappos, M.S.A.C. Gene, C.H. Polman, P.M. Matthews, S.L. Hauser, R.A. Gibson, J.R. Oksenberg, M.R. Barnes, Pathway and network-based analysis of genome-wide association studies in multiple sclerosis, Hum. Mol. Genet. 18 (2009) 2078–2090.
[64] J.W. Baurley, D.V. Conti, W.J. Gauderman, D.C. Thomas, Discovery of complex pathways from observational data, Stat. Med. 29 (2010) 1998–2011.
[65] W. Pan, Network-based model weighting to detect multiple loci influencing complex diseases, Hum. Genet. 124 (2008) 225–234.
[66] L. Chen, L. Zhang, Y. Zhao, L. Xu, Y. Shang, Q. Wang, W. Li, H. Wang, X. Li, Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways, Bioinformatics 25 (2009) 237–242.
[67] Y.L. Low, Y. Li, K. Humphreys, A. Thalamuthu, H. Darabi, S. Wedren, C. Bonnard, K. Czene, M.M. Iles, T. Heikkinen, K. Aittomaki, C. Blomqvist, H. Nevanlinna, P. Hall, E.T. Liu, J. Liu, Multi-variant pathway association analysis reveals the importance of genetic determinants of estrogen metabolism in breast and endometrial cancer susceptibility, PLoS Genet. 6 (2010) e1001012.
[68] D. Ruano, G.R. Abecasis, B. Glaser, E.S. Lips, L.N. Cornelisse, A.P. de Jong, D.M. Evans, G. Davey Smith, N.J. Timpson, A.B. Smit, P. Heutink, M. Verhage, D. Posthuma, Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability, Am. J. Hum. Genet. 86 (2010) 113–125.
[69] D. Ballard, C. Abraham, J. Cho, H. Zhao, Pathway analysis comparison using Crohn's disease genome wide association studies, BMC Med. Genomics 3 (2010) 25.
[70] L. Luo, G. Peng, Y. Zhu, H. Dong, C.I. Amos, M. Xiong, Genome-wide gene and pathway analysis, Eur. J. Hum. Genet. 18 (2010) 1045–1053.
[71] T. Manoli, N. Gretz, H.J. Grone, M. Kenzelmann, R. Eils, B. Brors, Group testing for pathway analysis improves comparability of different microarray datasets, Bioinformatics 22 (2006) 2500–2506.
[72] P. Kraft, S. Raychaudhuri, Complex diseases, complex genes: keeping pathways on the right track, Epidemiology 20 (2009) 508–511.
[73] H. Eleftherohorinou, V. Wright, C. Hoggart, A.L. Hartikainen, M.R. Jarvelin, D. Balding, L. Coin, M. Levin, Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases, PLoS One 4 (2009) e8068.
[74] T.G. Lesnick, S. Papapetropoulos, D.C. Mash, J. Ffrench-Mullen, L. Shehadeh, M. de Andrade, J.R. Henley, W.A. Rocca, J.E. Ahlskog, D.M. Maraganore, A genomic pathway approach to a complex disease: axon guidance and Parkinson disease, PLoS Genet. 3 (2007) e98.
[75] J.R. Perry, M.I. McCarthy, A.T. Hattersley, E. Zeggini, C. Wellcome Trust Case Control, M.N. Weedon, T.M. Frayling, Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach, Diabetes 58 (2009) 1463–1467.
[76] A. Torkamani, E.J. Topol, N.J. Schork, Pathway analysis of seven common diseases assessed by genome-wide association, Genomics 92 (2008) 265–272.
[77] K. Askland, C. Read, J. Moore, Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission, Hum. Genet. 125 (2009) 63–79.
[78] J.C. Lambert, B. Grenier-Boley, V. Chouraki, S. Heath, D. Zelenika, N. Fievet, D. Hannequin, F. Pasquier, O. Hanon, A. Brice, J. Epelbaum, C. Berr, J.F. Dartigues, C. Tzourio, D. Campion, M. Lathrop, P. Amouyel, Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis, J. Alzheimers Dis. 20 (2010) 1107–1118.
[79] J. Li, K. Humphreys, T. Heikkinen, K. Aittomaki, C. Blomqvist, P.D. Pharoah, A.M. Dunning, S. Ahmed, M.J. Hooning, J.W. Martens, A.M. van den Ouweland, L. Alfredsson, A. Palotie, L. Peltonen-Palotie, A. Irwanto, H.Q. Low, G.H. Teoh, A. Thalamuthu, D.F. Easton, H. Nevanlinna, J. Liu, K. Czene, P. Hall, A combined analysis of genome-wide association studies in breast cancer, Breast Cancer Res. Treat. 126 (2011) 717–727.
[80] I. Menashe, D. Maeder, M. Garcia-Closas, J.D. Figueroa, S. Bhattacharjee, M. Rotunno, P. Kraft, D.J. Hunter, S.J. Chanock, P.S. Rosenberg, N. Chatterjee, Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade, Cancer Res. 70 (2010) 4453–4459.
[81] K. Wang, H. Zhang, S. Kugathasan, V. Annese, J.P. Bradfield, R.K. Russell, P.M. Sleiman, M. Imielinski, J. Glessner, C. Hou, D.C. Wilson, T. Walters, C. Kim, E.C. Frackelton, P. Lionetti, A. Barabino, J. Van Limbergen, S. Guthery, L. Denson, D. Piccoli, M. Li, M. Dubinsky, M. Silverberg, A. Griffiths, S.F. Grant, J. Satsangi, R. Baldassano, H. Hakonarson, Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease, Am. J. Hum. Genet. 84 (2009) 399–405.
[82] D.I. Chasman, On the utility of gene set methods in genomewide association studies of quantitative traits, Genet. Epidemiol. 32 (2008) 658–668.
[83] P. Jia, J.M. Ewers, Z. Zhao, Prioritization of epilepsy associated candidate genes by convergent analysis, PLoS One 6 (2011) e17162.
[84] C. O'Dushlaine, E. Kenny, E. Heron, G. Donohoe, M. Gill, D. Morris, A. Corvin, Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility, Mol. Psychiatry 16 (2011) 286–292.
[85] D. Thomas, Gene-environment-wide association studies: emerging approaches, Nat. Rev. Genet. 11 (2010) 259–272.
[86] A.V. Segre, L. Groop, V.K. Mootha, M.J. Daly, D. Altshuler, Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits, PLoS Genet. 6 (2010).
[87] K. Zhang, S. Cui, S. Chang, L. Zhang, J. Wang, i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study, Nucleic Acids Res. 38 (2010) W90–W95.
[88] M. Holden, S. Deng, L. Wojnowski, B. Kulle, GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies, Bioinformatics 24 (2008) 2784–2785.
[89] D.F. Schwarz, O. Hadicke, J. Erdmann, A. Ziegler, D. Bayer, S. Moller, SNPtoGO: characterizing SNPs by enriched GO terms, Bioinformatics 24 (2008) 146–148.
[90] I. Medina, D. Montaner, N. Bonifaci, M.A. Pujana, J. Carbonell, J. Tarraga, F. Al-Shahrour, J. Dopazo, Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies, Nucleic Acids Res. 37 (2009) W340–W344.
[91] F. Al-Shahrour, L. Arbiza, H. Dopazo, J. Huerta-Cepas, P. Minguez, D. Montaner, J. Dopazo, From genes to functional classes in the study of biological systems, BMC Bioinformatics 8 (2007) 114.
[92] V. Dinu, H. Zhao, P.L. Miller, Integrating domain knowledge with statistical and data mining methods for high-density genomic SNP disease association analysis, J. Biomed. Inform. 40 (2007) 750–760.
[93] D.J. Schaid, S.K. McDonnell, S.J. Hebbring, J.M. Cunningham, S.N. Thibodeau, Nonparametric tests of association of multiple genes with human disease, Am. J. Hum. Genet. 76 (2005) 780–793.
[94] O. De la Cruz, X. Wen, B. Ke, M. Song, D.L. Nicolae, Gene, region and pathway level analyses in whole-genome studies, Genet. Epidemiol. 34 (2009) 222–231.