Published in final edited form as:
Hum Genet. 2014 February ; 133(2): 125–138. doi:10.1007/s00439-013-1377-1.

Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives

Peilin Jia and
Department of Biomedical Informatics, Vanderbilt University School of Medicine, 2525 West End Avenue Suite 600, Nashville, TN, USA

Zhongming Zhao
Department of Biomedical Informatics, Vanderbilt University School of Medicine, 2525 West End Avenue Suite 600, Nashville, TN, USA, Department of Psychiatry, Vanderbilt University School of Medicine, 1601 21st Ave S, Nashville, TN 37212, USA, Department of Cancer Biology, Vanderbilt University School of Medicine, 2220 Pierce Ave. Nashville, TN 37232, USA

Abstract

Genome-wide association studies (GWAS) have rapidly become a powerful tool in genetic studies of complex diseases and traits. Traditionally, single marker-based tests have been used prevalently in GWAS and have uncovered tens of thousands of disease-associated SNPs. Network-assisted analysis (NAA) of GWAS data is an emerging area in which network-related approaches are developed and utilized to perform advanced analyses of GWAS data in order to study various human diseases or traits. Progress has been made in both methodology development and applications of NAA in GWAS data, and it has already been demonstrated that NAA results may enhance our interpretation and prioritization of candidate genes and markers. Inspired by the strong interest in and high demand for advanced GWAS data analysis, in this review article, we discuss the methodologies and strategies that have been reported for the NAA of GWAS data. Many NAA approaches search for subnetworks and assess the combined effects of multiple genes participating in the resultant subnetworks through a gene set analysis. With no restriction to pre-defined canonical pathways, NAA has the advantage of defining subnetworks with the guidance of the GWAS data under investigation. In addition, some NAA methods prioritize genes from GWAS data based on their interconnections in the reference network. Here, we summarize NAA applications to various diseases and discuss the available options and potential caveats related to their practical usage. Additionally, we provide perspectives regarding this rapidly growing research area.

Introduction

Genome-wide association studies (GWAS) have become a powerful tool to study the genetic architectures of complex diseases and traits in humans. Since the first GWA study was published (Klein et al. 2005), more than 1650 publications of GWA studies, which have pinpointed tens of thousands of single nucleotide polymorphisms (SNPs), have been reported and deposited into the GWAS Catalog (as of August 1, 2013, http://www.genome.gov/gwastudies/) (Hindorff et al. 2009). Conventional GWAS analysis requires disease-associated SNPs to reach the genome-wide significance level (e.g., $p < 5 \times 10^{-8}$) due to multiple test correction. However, this stringent requirement has excluded

Corresponding author: Zhongming Zhao, Phone: +1-615-343-9158, Fax: +1-615-936-8545, zhongming.zhao@vanderbilt.edu.

Competing interests
The authors declare that they have no competing interests.

many genuinely associated SNPs that have moderate or weak association signals (Wang et al. 2010). As a result, the discovered significant SNPs might only explain a small proportion of genetic risks for most complex diseases or traits, leaving the question of “missing heritability” open to further investigation (Manolio et al. 2009). A number of contributing factors have been hypothesized, such as joint effects of multiple SNPs/genes (Cantor et al. 2010; Wang et al. 2010), epistasis effects (e.g., SNP-SNP interactions and gene-gene interactions) (Hu et al. 2011; McKinney et al. 2009; McKinney and Pajewski 2011), epigenetic regulations, gene-environment interactions, and joint effects of rare variants and common/rare variants (Gibson 2011).

Among these expanded approaches to search for missing heritability, gene set analysis (GSA) of GWAS data, which assesses the combined effects of multiple SNPs/genes, has been of great interest to researchers due to its improved power and biologically interpretable results (Lee et al. 2011; Wang et al. 2010). The rationale of GSA recognizes that while many variants might contribute to complex diseases, with each variant individually creating weak to moderate effects, their combined effects could be significant (Gibson 2010; Yang et al. 2010; Yang et al. 2011). Throughout this review, GSA is denoted as an analytical approach that conducts an enrichment test of a set of genes. The gene set can be defined using canonical pathways (Kanehisa et al. 2010), Gene Ontology (GO) categories (Ashburner et al. 2000), or subnetworks. Pathway-based analysis (PBA) is one type of GSA that uses canonical pathways, GO biological process categories, or other pathway annotations as its gene set unit. So far, PBA approaches have been well-developed for GWAS analyses and successfully applied to the study of various complex diseases (Askland et al. 2009; Askland et al. 2012; Chen et al. 2010; Elbers et al. 2009; Fehringer et al. 2012; Holmans et al. 2009; Jia et al. 2012a; Jia et al. 2010b; Perry et al. 2009; Wang et al. 2007; Wang et al. 2009; Wang et al. 2011a). PBA utilizes pre-defined gene sets based on functional annotations, thereby providing well-annotated biological information and regulation. Readers may refer to several recent review articles (Ramanan et al. 2012; Wang et al. 2010; Wang et al. 2011b) for more detailed discussions.

In contrast to PBA’s reliance on pre-defined gene sets, network-assisted analysis (NAA) can define gene sets flexibly and dynamically using subnetwork search algorithms. Enrichment tests may also be included in NAA. In NAA’s structure, genetic data is overlaid onto a reference network. Then, a search for subnetworks is conducted with the guidance of the GWAS data. The first network-assisted study was reported in an analysis of multiple sclerosis (MS) in 2009 (Baranzini et al. 2009) using the plugin package jActiveModule in the software tool Cytoscape (Shannon et al. 2003). This plugin package was originally developed for the network analysis of gene expression. The authors superimposed genetic association data (*p*-values) onto a human protein-protein interaction (PPI) network and successfully identified subnetworks that were enriched with MS-associated genes as well as immune- and neurodevelopment-related pathways. Inspired by this successful application, several methods and tools have gained rapid publications. These methods vary in their data management, subnetwork search algorithm, hypothesis testing, and result interpretation. Considering the broadness of network approaches, NAA of GWAS can reach beyond the GSA principle of gene set-based enrichment analysis. Disease Association Protein-Protein Link Evaluator (DAPPLE) is one such expanded approach (Rossin et al. 2011). A set of association loci is required as DAPPLE’s input. Then, this method prioritizes genes based on connectivity in the PPI network (see details below). With the growing awareness of NAA’s advantages, many original GWA or genetic studies have included such analyses to help interpret their association results (Arning et al. 2012; Neale et al. 2012; Xu et al. 2012).

While promising, current development of NAA approaches for GWAS data is still in an early stage. In this article, we aim to review the strategies and methodologies in this emerging field of NAA of GWAS data, the issues that have arisen from real applications, and the challenges that accompany the deluge of new data. Notably, we discuss analyses that are based on networks in which nodes are typically proteins (e.g., PPI networks) rather than other genetic or statistical interaction systems such as SNP-SNP interactions, gene-gene interactions, or interactions built on SNP and expression quantitative trait locus (eQTL) relationships. The methods discussed in this review provide guidelines for future network-assisted analyses of GWAS data.

Demand and rationale in network-assisted analysis of GWAS data

As high-throughput technologies advance rapidly, many GWAS datasets have been published, and many more are expected in the near future. The demand on GSA, PBA, and NAA is strong, as investigators now often search for association signals beyond single markers. Both pathway and network approaches explore combined genetic association signals based on the existing biological knowledgebase. One advantage of these approaches is that the interpretation of the results (enriched pathways or networks) is straightforward. Both NAA and PBA each have their own strengths and weakness. Table 1 summarizes several representative features that distinguish NAA from PBA. PBA has been successfully applied in the study of many diseases. The enriched pathways uncovered by PBA provide insights that might be missed by single marker analysis. However, PBA has several limitations. (1) Results gleaned using different methods differ greatly (Fehringer et al. 2012; Wang et al. 2013). (2) Pathway annotations used in PBA are limited to our current knowledge. For poorly-studied diseases such as dental caries, there is a lack of appropriate canonical biological pathways for PBA. (3) Current pathway annotations cover many pre-defined pathways that may be too general in their delivery of disease-related biological functions (Ruano et al. 2010; Sun 2012). (4) Connections among many genes are lacking within major annotation databases used for PBA, such as the GO database (Farber 2013).

NAA provides complementary approaches that may expand the current advanced analysis of GWAS data. NAA relies on network annotations that present functional interactions among genes and their products. A variety of networks are available, such as PPI networks, gene co-expression networks, and regulatory networks. Compared to PBA, NAA is more flexible when defining gene sets throughout the whole network, allowing for the generation and testing of various types of hypotheses. Nevertheless, NAA is not flawless. The PPI data used in NAA tends to contain false positives and is dynamic under specific cellular conditions. The resultant subnetworks typically do not specify biological function as explicitly as canonical pathways.

The rationale behind NAA is the principle of “guilt-by-association (GBA)” (Farber 2013; Lee et al. 2011), which states that genes (or genes’ products) that are interconnected in the network are more likely to share the same or a similar function (Gillis and Pavlidis 2012). Accordingly, studying association signals pulled from GWAS data in the context of a biological network will likely help to uncover disease susceptibility genes and their interactions. A straightforward design for NAA of GWAS data is to search for subnetworks that are both enriched with association signals and contain functional relationships. Several methods have been published based on this concept. In the next section, we discuss the issues and features of these methods, thereby aiming to provide guidelines for the design and implementation of NAA of GWAS data.

Design and methodological issues

As illustrated in Figure 1, in general, the NAA of GWAS data has the following steps. First, GWAS data is preprocessed to compute SNP-based and gene-based statistical values, which represent their association levels related to the phenotype/disease under investigation.

Second, the genetic signals (SNP- or gene-based *p* values or scores) are overlaid onto the network in a quantitative or qualitative manner. Third, a network-based integrative analysis is performed to explore the topological characteristics of the network, study the connections among genes of interest, and search for subnetworks, among other purposes. Finally, the results are postprocessed and evaluated through the use of significance tests, module selection, follow-up functional analyses of subnetwork genes, and other approaches.

Preprocess GWAS data: gene-based p-value computation

In GWA studies, association tests are performed for SNP markers, yet PPI network annotation data is based on genes' products (proteins). Hence, one challenge in NAA is to map SNPs to genes and then compute gene-based statistical values. Accordingly, only SNPs mapped to genic regions are considered when estimating gene-based statistical values, while intergenic SNPs are typically discarded due to the nature of GSA. A genic region is defined by the gene (from gene's start point to end point) plus its boundary regions (the flanking regions both upstream and downstream of the gene). SNPs that are mapped to their nearest gene region would belong to that gene. In practice, the length of the boundary regions varies from 0 to 500 kb. Using a schizophrenia GWAS dataset genotyped on the Affymetrix 6.0 chip, we previously demonstrated that the total number of informative genes (genes with at least one SNP in the GWAS dataset) increased from ~18,000 to ~20,000 when the gene boundary increased from 5 to 100 kb (Jia et al. 2011a). The SNPs mapped to genic regions increased accordingly from ~307,000 to 509,000 out of the 725,000 genotyped SNPs. Despite this increase, Lee et al. (2011) showed that a change of boundaries from 0 to 250 kb did not significantly affect the power of the related network analysis.

As shown above, the SNP-gene mapping process will result in a gene having multiple SNPs with genotype data. Thus, it is necessary to estimate a gene-based *p*-value to represent the overall association signal to the gene. So far, there have been many methods to calculate this estimation. They can be categorized into two classes depending on whether it requires raw genotyping data. The first class of methods does not require raw genotyping data. Rather, they use the SNP *p*-values to compute gene-based *p*-values (Curtis et al. 2008; Li et al. 2011; Liu et al. 2010; Wang et al. 2007; Zaykin et al. 2007). The most popular method in this class defines a gene-based *p*-value by using the smallest *p*-value from all SNPs mapped to its genic region (hereafter denoted as the MinP method). The advantages of the MinP method are its easy implementation and its sensitivity in capturing association signals. However, one disadvantage is that genes represented by more SNPs on genotyping arrays (SNP density, which is usually correlated to gene length) are more likely to have small *p*-values (Askland et al. 2012; Jia et al. 2012b), thereby introducing biases. Several methods were recently developed to overcome this major disadvantage of the MinP method, such as the Fisher's method (Curtis et al. 2008), the truncated-product *p*-value method (Zaykin et al. 2007; Zaykin et al. 2002), the extended Simes' procedure (Li et al. 2011; Wang et al. 2007), and the versatile gene-based test (VEGAS) (Liu et al. 2010). Among them, VEGAS performs a gene-based association test by considering either all SNPs that are mapped to the gene or a subset of SNPs ranked by significance (through function "Top-x% test" in VEGAS) while accounting for gene length and linkage disequilibrium (LD) structures. Given that a gene has *n* SNPs for the computation of the gene-based *p*-value, VEGAS simulates an *n*-element, multivariate, normally distributed vector using both SNPs' *p*-values and pairwise LD values among these *n* SNPs. VEGAS has been implemented in both online and stand-alone tools. It is easy to use, and its computational time is moderate. One disadvantage is that VEGAS only allows for autosomal genes due to inaccurate LD information from sex chromosomes (Liu et al. 2010).

The second class of methods requires individual genotype and phenotype data and involves various regression-related strategies (Chen et al. 2010; Wang and Abbott 2008; Wu et al. 2010b). Many early works used multivariate regression to compute gene-based p-values. These methods involve a high dimensionality of data (thus, high degree of freedom) and ignore LD information (Mukhopadhyay et al. 2010; Wessel and Schork 2006). Recently, several sophisticated approaches have been proposed to deal with these problems, such as the kernel linear regression (Wu et al. 2010b) and penalized regression (Chen et al. 2010) methods. Because this review article focuses on NAA approaches, readers who are interested in gene based *p*-value methods are referred to several recent reviews for more details (Dering et al. 2011; Ma et al. 2013; Wu and Cui 2013).

Computing gene-based *p*-values is critical because it is the first step in both PBA and NAA of GWAS datasets. As shown in Figure 2, if the MinP method is used, a permutation-based test is strongly suggested in the downstream pipeline to adjust for potential biases derived from SNP density and/or gene size. Alternatively, if a method can take into account the potential biases in the gene-based *p*-value computing step, the significance test in the downstream pipeline would not be necessary for the purpose of adjusting for potential biases (e.g., through permutation test). It is worth noting that intergenic SNPs, with the exception of those located in the boundary region of genes, are largely discarded. Furthermore, genes that currently do not have appropriate pathway or network annotations are excluded in PBA or NAA.

Features of network data

A variety of molecular networks have been actively explored in biological systems, such as the PPI network, metabolic network, regulatory network, gene co-expression network, and so on. We refer to a functional protein network to represent any interactions among proteins, including physical interactions, co-expression, and co-occurrence in literatures. The functional protein network represents a global view of connectivity among genes in terms of their protein interactions. It has been widely used in NAA of GWAS data, primarily because it is straightforward in its examination of the joint genetic effects of multiple related genes. Among the available functional networks, the PPI network has been the most popular resource for NAA. The major sources of the protein network data include (1) large-scale experiments, (2) computational predictions, and (3) literature mining and curation. All of these sources of network data have been exploited in NAA of GWAS. Table 2 lists four specific datasets, which either have previously been applied to the NAA of GWAS data or are suitable for direct applications in disease studies.

One caveat of PPI data is that most currently available PPI data sets are static and ignore the temporal and spatial dependence of protein interactions in real biological systems. Most of the currently available PPI data derives from large-scale high-throughput experiments *in vitro* and may not represent tissue-specific information (Barshir et al. 2013). In addition, PPI data is currently far from being complete. It accumulates and changes over time. For example, the BioGRID database (Chatr-Aryamontri et al. 2013), a comprehensive database that collects genetic and protein interaction data from human and model organisms, updates its interaction data monthly. Finally, PPI data is not error-free and may not specifically reflect real cellular conditions.

In addition to the direct and physical interactions gleaned from PPI data, indirect interactions relying on functional correlations could be useful for GWAS analysis (Barabasi et al. 2011; Braun 2012; Lage et al. 2007; Vidal et al. 2011). These indirect interactions include tissue-specific gene co-expression (Lee et al. 2011), literature co-citation, and genetic interactions (Hu et al. 2011), among others. Compared to physical interactions, these functional interactions are more biologically relevant to diseases or traits and, thus, instrumental to the discovery of meaningful subnetworks enriched with GWAS signals. For example, disease-relevant tissue data, such as gene co-expression correlations, may be of special interest due to its tissue-specific information, as demonstrated in specific studies of obstructive sleep apnea (Liu et al. 2011) and autism (Ben-David and Shifman 2012). Details of these data resources can be found in additional reviews elsewhere (Barabasi et al. 2011; Vidal et al. 2011).

Construction of subnetworks

In general, NAA of GWAS data starts with certain pre-defined “seed” genes from the original GWAS data. Seed genes can range from a small set of pre-selected interesting genes (e.g., genes with $p < 5\times10^{-8}$) to all genes mapped onto a reference network (i.e., iteratively take each gene as a seed). NAA expands from the seed genes to a subnetwork by searching the whole network and then assessing the statistical significance of the resultant candidate subnetworks. We summarized the subnetwork construction approaches in the following five points.

Binary versus quantitative—Each gene extracted from the GWAS data is denoted by a statistical value (e.g., gene-based *p*-value) to represent its disease-association level. Depending on how these quantitative genetic signals are used, NAA methods can be categorized into two groups (Figure 4a). The first group dichotomizes genes by selecting a set of interesting genes (e.g., genes with $p < 5\times10^{-8}$) and uses these interesting genes as input query genes (i.e., seed genes) (Rossin et al. 2011). In this group of methods, GWAS data is only used to select interesting genes, and the subsequent analyses generally do not require further information from the GWAS dataset (Rossin et al. 2011). The second group of methods overlays the quantitative signal of each gene onto the whole network to build a node-weighted network. The node weights are transformed from the association significance values (i.e., gene-based *p*-values) (Akula et al. 2011; Baranzini et al. 2009; Jia et al. 2011b; Liu et al. 2011). Then, the search process is guided by the node weights.

Topologically oriented versus genetically oriented—One advantage of network analysis is that the topological structure of a network reflects biological interactions among nodes (proteins encoded by genes). Previous studies in network medicine have explored a number of similarly expressed hypotheses based on the principle of guilt-by-association, such as “proteins involved in the same disease are more likely to interact with each other,” or, conversely, “proteins closely located in the network tend to lead to similar disease phenotypes” (Barabasi et al. 2011). Depending on how the topological structure is applied within the analysis of GWAS data, NAA methods can be categorized as topologically oriented, genetically oriented, or a combination of both (Figure 4b). Topologically oriented methods leverage topological features such as degrees and shortest paths to build disease-subnetworks (Lee et al. 2011). As illustrated in Figure 4b, in every subnetwork-expansion step, those nodes that have the most important topological signatures (e.g., those nodes with the largest degree) tend to be selected. Because of this feature, the resultant subnetwork gleaned using a topologically oriented method greatly depends on the reference network. Thus, topological features may be included in computation of subnetwork scores in these methods. Genetically oriented methods use genetic signals to guide the search for disease subnetworks (Baranzini et al. 2009; Jensen et al. 2011; Jia et al. 2011b). In the expansion step, genetically oriented methods recruit the nodes that have the most significant association signals (Figure 4b). Subnetwork scores are typically computed by combining the genetic signals of all the nodes in the subnetwork in genetically oriented methods. A third design leverages features from both types of methods. For instance, Akula et al. (2011) calculated a combined score of a subnetwork using the Liptak-Stouffer method:

$z_{comb}=\frac{\sum w_i z_i}{\sqrt{\sum w_i^2}}$, where z_i is converted from a gene-based *p*-value of the component genes in the subnetwork, and w_i is the gene-based weight obtained from the modified Google PageRank algorithm. Thus, this approach accounts for both topological and genetic characteristics of the genes (Akula et al. 2011).

Global versus local search—Given a node-weighted reference PPI network, the search path of a disease-related subnetwork may either reach all the nodes in the network or be restricted to reach a certain distance from the seed genes (Figure 4c). Starting with the seed genes, a global search theoretically allows for access to every node in the network as long as there are available paths (Baranzini et al. 2009; Jia et al. 2011b; Lee et al. 2011). For local searches, only the nodes and edges that are located within pre-defined distances, e.g., 2 steps away from the query seed genes, are explored and evaluated (Akula et al. 2011; Rossin et al. 2011).

Prioritization versus combination—Some of the published methods prioritize candidate genes, while others assess the combined effect of a set of genes without accounting for the particular rank of each individual gene in the resultant subnetwork. The main purpose of gene prioritization is to find disease candidate genes. Many methods for gene prioritization (Bornigen et al. 2012; Kohler et al. 2008; Sun et al. 2009) can be adapted to GWAS data analysis. Starting with a list of reference genes (e.g., known disease genes), these gene prioritization methods uncover the features of the reference genes and rank the candidate genes to identify new, potentially disease-associated genes. In GWA studies with a multi-stage design, gene prioritization is particularly useful in the selection of candidate genes/markers at the initial stage. Then, the prioritized genes are recruited for deep and dense genotyping in the later stage(s) (e.g., stage 2 or stage 3 for replication). Representative prioritization methods, such as GRAIL (Raychaudhuri et al. 2009) and DAPPLE (Rossin et al. 2011), have the functions to facilitate candidate gene prioritization. On the other hand, methods that focus on combined gene effects evaluate genes in a subnetwork as a whole (Akula et al. 2011; Jia et al. 2011b) with no particular rank for each gene in the subnetwork. A permutation test is often implemented in these methods to assess disease associations.

Direct versus indirect interactions—Most network-based methods used for the analysis of GWAS data are built on direct interactions among proteins (Akula et al. 2011; Baranzini et al. 2009; Jia et al. 2011b; Liu et al. 2011). Recently, indirect connections have drawn increasing attention. These connections describe the closeness of nodes in a network rather than relying on direct interactions. Figure 4d illustrates the scenarios of direct and indirect interactions. For instance, two disconnected proteins in the network (nodes a and b) that are highly reachable through the paths between them are more likely to be involved in the same disease than two proteins that have only sparse paths between them (Figure 4d). The diffusion kernel-based network model is one of the emerging methods that incorporate indirect interactions (Barabasi et al. 2011; Kohler et al. 2008; Kondor and Lafferty 2002). This type of method has been widely used in network medicine but, to date, has been rarely applied in GWAS data. One pioneering application by Lee et al. (2011) explored six network methods to assess their capability to boost the statistical power of GWA studies and detect disease candidate genes. Considering their strength in the measurement of the global distance in a network (Kohler et al. 2008; Nitsch et al. 2009), kernel-based models warrant future expansion and applications.

Statistical significance tests

After obtaining candidate subnetworks, an important next step is to evaluate and select subnetworks by utilizing appropriate significance tests. Two hypotheses have been widely formulated and followed in most GSA of GWAS data (Wang et al. 2010; Wang et al. 2011b). The competitive null hypothesis (Q1) states that the genes in a gene set have the same association levels with the investigated disease compared with the genes outside of the gene set. The self-contained null hypothesis (Q2) states that the genes in a gene-set are not associated with the given disease (Figure 3). Because of the dynamics in subnetwork searches and due to different network algorithm features, the traditional multiple testing correction methods (e.g., Bonferroni correction) are generally not applicable to NAA. In practice, estimation of the null distribution through randomization or permutation tests is popular to assess the significance of the results. In general, a randomization test is typically adopted in Q1-related methods, while permutation analysis is more appropriate for assessing the results obtained by Q2-related methods (Figure 3).

In randomization tests, random networks are first generated to estimate the null distribution. Then, the results obtained in the actual reference network are assessed for significance by comparing them with the null distribution. The approaches to generate random networks are quite diverse, including various node- and edge-related randomization strategies. A number of previous investigations have proved that the whole human PPI network behaves in a scale-free fashion. One feature of scale-free networks is that a small portion of nodes have numerous interacting partner nodes (i.e., “hub” nodes/genes). Accordingly, they are more likely to be included in subnetworks by chance in NAA. However, when generating random network data, many methods ignore the topological features of the nodes and treat all nodes equally. To control the potential biases, Li et al. (2009) suggested the generation of topologically-matched random networks according to degree and clustering coefficients. In another work by Rossin et al. (2011), random networks were built using matching node degrees.

In permutation tests, randomized phenotype data is generated to estimate the null distribution, e.g., by swapping case and control labels in the GWAS data. Then, subnetworks were assessed using both the actual and permuted GWAS datasets. Permutation tests, although often computationally intensive and requiring raw genotype data, have been found to be effective to adjust for various biases, such as the long gene bias, as discussed in a recent survey (Jia et al. 2010a). Permutation tests are especially applicable when the MinP method is used. Due to its computational intensity, a permutation test cannot be practically implemented in most online tools. However, it can be conveniently incorporated in stand-alone computer programs. We expect that the development of future NAA tools for GWAS data will take into consideration both computational efficiency and practical convenience.

Analysis procedures

GWAS datasets collected from different consortia, ethnic populations, or technical platforms are now available for many complex diseases or traits. New approaches are needed to take advantage of the multiple datasets for the same diseases or traits. In a recent publication, Jia et al. (2012c) proposed an approach using two schizophrenia GWAS datasets, one as discovery and another as an evaluation dataset, to find subnetworks that are consistently highly scored in both datasets. This approach may be applied to the same analysis method for multiple GWAS datasets (cross-dataset validation), different methods for the same GWAS dataset (cross-method validation), or both.

NAA Methods

There are a number of publications regarding the network analysis of GWAS data. These publications are either methodology-oriented or discovery-oriented (Table 3). Below, we selected several representative approaches for detailed discussion.

Methods based on jActiveModule

jActiveModule is a plugin package in the software tool Cytoscape (Shannon et al. 2003) and was initially developed for gene expression data analysis (Ideker et al. 2002). The original model works on a node-weighted PPI network where node weights are represented by *z*-scores converted from gene-based *p*-values, e.g., $z = \Phi^{-1}(1-p)$. The underlying algorithm of jActiveModule relies on a greedy search. The algorithm starts with "seed" genes and recruits the best nodes in the neighborhood of the seeds in each expansion step. Module expansion stops when it does not satisfy the pre-defined criteria, e.g., if the module score increase is too small to reach the cutoff value. Then, modules are evaluated using a randomization test. In gene expression data and in most applications of jActiveModule to GWAS data, the module score (z_m) is computed by combining node weights using the Liptak-Stouffer *z*-score method (Ideker et al. 2002), $z_m = \sum z_i / \sqrt{k}$, where *k* is the number of nodes in the subnetwork, and z_i is the node weight. Recent studies have started to incorporate genotyping data features in the computation of network scores based on either the correlations among genes/SNPs (Pedroso et al. 2012) or the topological features of gene products (Akula et al. 2011).

jActiveModule is a well-developed model in Cytoscape, and it only requires a gene-based *z*-score. Thus, it can be easily adapted to GWAS data analysis. For example, node weights can be computed based on GWAS *p*-values instead of gene expression *p*-values. A protocol called PANOGA describes the details of using jActiveModule for GWAS data (Burcu and Osman Ugur 2012). So far, jActiveModule has been one of the most widely used network methods in GWAS data analysis (Table 4), with applications in multiple sclerosis (Baranzini et al. 2009), obstructive sleep apnea (Liu et al. 2011), and rheumatoid arthritis (Burcu and Osman Ugur 2011).

While it is easy to run, jActiveModule was developed originally as a tool to analyze gene expression data rather than GWAS data. This fact creates several potential issues. First, in gene expression data, the majority of genes are not differentially expressed (DE) and these gene *p*-values are evenly distributed between 0 and 1, with an exception of DE genes. Therefore, the converted *z*-scores are expected to evenly fall into the intervals that are less than 0 and greater than 0. In GWAS data, however, the SNP *p*-values, instead of gene *p*-values, are presumably evenly distributed between 0 and 1, with the exception of disease-associated SNPs. Depending on how the gene-based *p*-values are computed, the results could be substantially skewed. For example, in the MinP method, gene-based *p*-values are biased towards 0, causing an uneven distribution of the converted *z*-scores (Jia and Zhao 2012). This outcome, in turn, results in extremely high module scores, making the selection of modules difficult (Baranzini et al. 2009; Jia et al. 2011b). To solve this problem, multiple strategies may be considered. For instance, gene-based *p*-values can be computed using more sophisticated methods rather than the MinP method (see above). Alternatively, additional steps can be introduced in the downstream analysis to adjust for biases (Figure 3) (Jia et al. 2012c). It should also be considered that, since jActiveModule was originally developed for analyzing gene expression data, the permutation test for GWAS data is not implemented in jActiveModule. When users perform an association test (Q2), additional work is needed to implement a permutation test on the resultant modules from jActiveModule.

dmGWAS

dmGWAS is the first tool to implement a network-based method specifically for GWAS data analysis (Jia et al. 2011b). It allows for a quantitative global search using the association signals for guidance. Although built initially on the design of jActiveModule, dmGWAS includes functions to facilitate the analysis of genetic data, such as the computation of gene-based *p*-values and implementation of permutation tests. dmGWAS implemented several methods for computing gene-based *p*-values, including the MinP method, the Simes' method (Wang et al. 2007), the Fisher's method, and the method that uses the smallest gene-based false discovery data value (Jia et al. 2011b; Peng et al. 2010). However, for sophisticated methods such as VEGAS (Liu et al. 2010), the user needs to compute gene-based *p*-values using independent tools first. Overall, the NAA strategy implemented in dmGWAS is to search for subnetworks that are significantly associated with a specific disease or trait.

NIMMI

The software tool Network Interface Miner for Multigenic Interactions (NIMMI) (Akula et al. 2011) applies a local search to find subnetworks that have high combined scores. NIMMI is the first NAA tool to integrate topological features with genetic signals. It computes a combined score using the Liptak-Stouffer method, i.e., $Z_{comb}=\sum w_i z_i / \sqrt{\sum w_i^2}$, where z_i is the *z*-score converted from gene-based *p*-values, and w_i represents topological features computed using a modified Google PageRank method (Akula et al. 2011). NIMMI requires nodes in the subnetworks to be within a certain distance (*d*) from the seed nodes. A distance of 2 was suggested in the original work (Akula et al. 2011) (Figure 4c). One disadvantage of NIMMI is that the resultant subnetworks contain all the nodes that are *d* steps away from the seed without further filtering. Correspondingly, the subnetworks would include many unrelated nodes and tend to be too large to prioritize genes for follow up investigations.

DAPPLE

DAPPLE performs local searches and evaluates whether genes located in association loci are significantly connected in the PPI network. In implementation, DAPPLE takes associated loci, or the genes located within the loci, as input seed genes. Importantly, DAPPLE's search steps do not rely on quantitative genetic data. Two sets of interactions are defined in DAPPLE: (1) direct interactions among any two proteins encoded by the genes located in association loci and (2) indirect interactions, in which two association proteins are linked through a third protein. DAPPLE also implements a series of comprehensive randomization tests, including node-label randomization, to assess the significance of the interactions. Because DAPPLE only needs a set of interesting genes as its input, it can be easily extended to many other cases for the study of candidate genes that are not necessarily derived from GWAS (Neale et al. 2012; Sanders et al. 2012). In addition, DAPPLE provides a convenient online tool for users to upload the candidate gene list(s) for analysis.

One limitation of DAPPLE is that it heavily relies on how the association loci are defined, as only genes located in these loci are used. So far, 5×10^{-8} has been widely accepted as the genome-wide significance cutoff *p*-value to select disease-associated SNPs in GWA studies. However, for studies in which no significant loci were identified by this stringent threshold (Sullivan et al. 2008), the selection of association loci could be arbitrary. Of note, the GWAS signals are only used in the first stage to define input genes in DAPPLE. In fact, the search step for interactions in DAPPLE is independent of GWAS data but depends on the reference PPI network.

Perspectives for future work

With the rapid accumulation of GWAS data, investigators are now facing more challenges related to advanced analyses of multiple GWAS datasets for complex diseases or traits. These datasets were generated using different platforms, were sourced from different ethnic populations, or were initially analyzed by different groups or consortia. NAA approaches are expected to address issues involving multiple datasets, multi-dimensional data types, and different types of variants. Previous studies have demonstrated the utilization of multiple GWAS datasets to validate the discovered pathways in PBA (Fehringer et al. 2012; Wang et al. 2009). Based on these studies, NAA is suggested to incorporate cross-dataset designs as well. For example, candidate subnetworks identified in one GWAS dataset (discovery stage) could be further validated in an independent GWAS dataset (evaluation stage) for the same disease or trait using the same NAA method (Jia et al. 2012c).

Most NAA approaches discussed in this review focus on the integration of GWAS data with PPI data. In addition to PPIs, other data, such as gene expression data at pathway levels (Xiong et al. 2012; Zhang et al. 2012) and co-expression networks (Farber 2010), can also be used in NAA. One recent study constructed regulatory networks consisting of SNPs and genes as nodes and their correlations as edges (Zhu et al. 2012). Furthermore, a protein-interaction-network-based pathway analysis (PINBPA) was proposed and successfully applied to two large genetic studies of multiple sclerosis (International Multiple Sclerosis Genetics Consortium 2013). Many network-related approaches that have been applied to other biological data, such as somatic mutation data (Ng et al. 2012; Vaske et al. 2010), mRNA abundance data (Chuang et al. 2007), and methylation data, may be adapted and incorporated in the NAA of GWAS data. With the rapid accumulation of genetic and genomic data, multi-dimensional integration is expected to gain more attention in the near future than canonical single domain based approaches.

Expanding NAA to incorporate a full spectrum of genetic variants, including common, rare, and *de novo* SNPs, is another potential direction. Advances in next-generation sequencing technologies have revolutionized biological and biomedical research, including related search methods for association signals in complex diseases and traits. The high volume of single-nucleotide variants uncovered in each single human genome or exome creates both challenges and opportunities for genetic studies. The richness and variety of hypothesis testing methods in network medicine provide multiple directions to facilitate the study of genetic variants within network data. Rare variants are typically computed for each gene by collapsing them in gene regions. These gene-based *p*-values from rare variants can be integrated into the PPI network using the same approach that is used for common variants. For instance, one could search for subnetworks that are enriched with rare variants for a disease. It is also of interest to test whether the genes pulled based on rare variant associations are connected via PPIs in a non-random manner (Neale et al. 2012; Sanders et al. 2012). Gilman et al. (2012) designed an approach named NETBAG for NAA of rare *de novo* variants (Gilman et al. 2011), and they later successfully extended their approach to diverse types of genetic variants. These works demonstrate the tendency of genetic variants to converge on functional gene networks in schizophrenia. With this proof of concept, new NAA approaches that incorporate various genetic data are expected to be in high demand in the near future.

Conclusion

In summary, we discussed the main features of network-based analyses of GWAS data and highlighted key issues in the design, implementation, and application of these network-based methods. So far, the available methods vary in their hypothesis testing, search algorithm, integration strategy, and additional useful features. Specifically, we reviewed four NAA methods in detail. We expect more GWAS data at both genotyping array and sequencing levels to be generated in the near future. These new datasets, along with existing GWAS datasets, will help us to fine-tune network-based methods. Furthermore, these enhanced methods and computational tools will further help to uncover genes and subnetworks that might play a role in complex diseases or traits.

Acknowledgments

This work was partially supported by National Institutes of Health grants (R01LM011177, R21HG006037, R03CA167695, and R03DE022093) and a 2010 NARSAD Young Investigator Award (to P. J.). We thank Ms. Rebecca H. Posey and Dr. Qingguo Wang for their valuable proofreading assistance.

References

Akula N, Baranova A, Seto D, Solka J, Nalls MA, Singleton A, Ferrucci L, Tanaka T, Bandinelli S, Cho YS, Kim YJ, Lee JY, Han BG, McMahon FJ. A network-based approach to prioritize results from genome-wide association studies. PLoS One. 2011; 6:e24220. [PubMed: 21915301]

Arning A, Hiersche M, Witten A, Kurlemann G, Kurnik K, Manner D, Stoll M, Nowak-Gottl U. A genome-wide association study identifies a gene network of ADAMTS genes in the predisposition to pediatric stroke. Blood. 2012; 120:5231–5236. [PubMed: 22990015]

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–9. [PubMed: 10802651]

Askland K, Read C, Moore J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum Genet. 2009; 125:63–79. [PubMed: 19052778]

Askland K, Read C, O'Connell C, Moore JH. Ion channels and schizophrenia: a gene set-based analytic approach to GWAS data for biological hypothesis testing. Hum Genet. 2012; 131:373–391. [PubMed: 21866342]

Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011; 12:56–68. [PubMed: 21164525]

Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet. 2009; 18:2078–2090. [PubMed: 19286671]

Baranzini SE, Srinivasan R, Khankhanian P, Okuda DT, Nelson SJ, Matthews PM, Hauser SL, Oksenberg JR, Pelletier D. Genetic variation influences glutamate concentrations in brains of patients with multiple sclerosis. Brain. 2010; 133:2603–2611. [PubMed: 20802204]

Barshir R, Basha O, Eluk A, Smoly IY, Lan A, Yeger-Lotem E. The TissueNet database of human tissue protein-protein interactions. Nucleic Acids Res. 2013; 41:D841–844. [PubMed: 23193266]

Ben-David E, Shifman S. Networks of neuronal genes affected by common and rare variants in autism spectrum disorders. PLoS Genet. 2012; 8:e1002556. [PubMed: 22412387]

Bornigen D, Tranchevent LC, Bonachela-Capdevila F, Devriendt K, De Moor B, De Causmaecker P, Moreau Y. An unbiased evaluation of gene prioritization tools. Bioinformatics. 2012; 28:3081–8. [PubMed: 23047555]

Braun P. Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays. Proteomics. 2012; 12:1499–1518. [PubMed: 22589225]

Burcu B-G, Osman Ugur S. A new methodology to associate SNPs with human diseases according to their pathway related context. PLoS One. 2011; 6:e26277. [PubMed: 22046267]

Burcu, B-G.; Osman Ugur, S. Identification of SNP Targeted Pathways From Genome-wide Association Study (GWAS) Data. 2012. Published online 29 May 2012

Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet. 2010; 86:6–22. [PubMed: 20074509]

Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M. The BioGRID interaction database: 2013 update. Nucleic Acids Res. 2013; 41:D816–823. [PubMed: 23203989]

Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. Am J Hum Genet. 2010; 86:860–871. [PubMed: 20560206]

Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007; 3:140. [PubMed: 17940530]

Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, Abecasis GR, Barrett JC, Behrens T, Cho J, De Jager PL, Elder JT, Graham RR, Gregersen P, Klareskog L, Siminovitch KA, van Heel DA, Wijmenga C, Worthington J, Todd JA, Hafler DA, Rich SS, Daly MJ. Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet. 2011; 7:e1002254. [PubMed: 21852963]

Curtis D, Vine AE, Knight J. A simple method for assessing the strength of evidence for association at the level of the whole gene. Adv Appl Bioinform Chem. 2008; 1:115–120. [PubMed: 21918610]

Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. Genet Epidemiol. 2011; 35(Suppl 1):S12–17. [PubMed: 22128052]

Detera-Wadleigh SD, Akula N. A systems approach to the biology of mood disorders through network analysis of candidate genes. Pharmacopsychiatry. 2011; 44(Suppl 1):S35–42. [PubMed: 21547870]

Elbers CC, van der Schouw YT, Wijmenga C, Onland-Moret NC. 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. 2009; 58:e9. author reply e10. [PubMed: 19720820]

Farber CR. Identification of a gene module associated with BMD through the integration of network analysis and genome-wide association data. J Bone Miner Res. 2010; 25:2359–2367. [PubMed: 20499364]

Farber CR. Systems-level analysis of genome-wide association data. G3 (Bethesda). 2013; 3:119–129. [PubMed: 23316444]

Fehringer G, Liu G, Briollais L, Brennan P, Amos CI, Spitz MR, Bickeboller H, Wichmann HE, Risch A, Hung RJ. Comparison of pathway analysis approaches using lung cancer GWAS data sets. PLoS One. 2012; 7:e31816. [PubMed: 22363742]

Garcia-Alonso L, Alonso R, Vidal E, Amadoz A, de Maria A, Minguez P, Medina I, Dopazo J. Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. Nucleic Acids Res. 2012; 40:e158. [PubMed: 22844098]

Gibson G. Hints of hidden heritability in GWAS. Nat Genet. 2010; 42:558–560. [PubMed: 20581876]

Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2011; 13:135–145. [PubMed: 22251874]

Gillis J, Pavlidis P. "Guilt by association" is the exception rather than the rule in gene networks. PLoS Comput Biol. 2012; 8:e1002444. [PubMed: 22479173]

Gilman SR, Chang J, Xu B, Bawa TS, Gogos JA, Karayiorgou M, Vitkup D. Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. Nat Neurosci. 2012; 15:1723–1728. [PubMed: 23143521]

Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron. 2011; 70:898–907. [PubMed: 21658583]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009; 106:9362–9367. [PubMed: 19474294]

Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Owen MJ, O'Donovan MC, Craddock N. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet. 2009; 85:13–24. [PubMed: 19539887]

Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, Moore JH. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. BMC Bioinformatics. 2011; 12:364. [PubMed: 21910885]

Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002; 18(Suppl 1):S233–S240. [PubMed: 12169552]

International Multiple Sclerosis Genetics Consortium. Network-Based Multiple Sclerosis Pathway Analysis with GWAS Data from 15,000 Cases and 30,000 Controls. Am J Hum Genet. 2013

Jensen MK, Pers TH, Dworzynski P, Girman CJ, Brunak S, Rimm EB. Protein interaction-based genome-wide analysis of incident coronary heart disease. Circ Cardiovasc Genet. 2011; 4:549–556. [PubMed: 21880673]

Jia P, Liu Y, Zhao Z. Integrative pathway analysis of genome-wide association studies and gene expression data in prostate cancer. BMC Syst Biol. 2012a; 6(Suppl 3):S13. [PubMed: 23281744]

Jia P, Tian J, Zhao Z. Assessing gene length biases in gene set analysis of Genome-Wide Association Studies. Int J Comput Biol Drug Des. 2010a; 3:297–310. [PubMed: 21297229]

Jia P, Wang L, Fanous AH, Chen X, Kendler KS, Zhao Z. A bias-reducing pathway enrichment analysis of genome-wide association data confirmed association of the MHC region with schizophrenia. J Med Genet. 2012b; 49:96–103. [PubMed: 22187495]

Jia P, Wang L, Fanous AH, Pato CN, Edwards TL, Zhao Z. Network-assisted investigation of combined causal signals from genome-wide association studies in schizophrenia. PLoS Comput Biol. 2012c; 8:e1002587. [PubMed: 22792057]

Jia P, Wang L, Meltzer HY, Zhao Z. Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. Schizophr Res. 2010b; 122:38–42. [PubMed: 20659789]

Jia P, Wang L, Meltzer HY, Zhao Z. Pathway-based analysis of GWAS datasets: effective but caution required. Int J Neuropsychopharmacol. 2011a; 14:567–572. [PubMed: 21208483]

Jia P, Zhao Z. Searching joint association signals in CATIE schizophrenia genome-wide association studies through a refined integrative network approach. BMC Genomics. 2012 1471-2164-13-S5–S15.

Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics. 2011b; 27:95–102. [PubMed: 21045073]

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010; 38:D355–D360. [PubMed: 19880382]

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308:385–389. [PubMed: 15761122]

Kohler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008; 82:949–958. [PubMed: 18371930]

Kondor, RI.; Lafferty, JD. Diffusion kernels on graphs and other discrete input spaces. Proceedings of the Nineteenth International Conference on Machine Learning; Morgan Kaufmann Publishers Inc; 2002. p. 315-322.

Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol. 2007; 25:309–316. [PubMed: 17344885]

Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011; 21:1109–1121. [PubMed: 21536720]

Li J, Zimmerman LJ, Park BH, Tabb DL, Liebler DC, Zhang B. Network-assisted protein identification and data interpretation in shotgun proteomics. Mol Syst Biol. 2009; 5:303. [PubMed: 19690572]

Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet. 2011; 88:283–293. [PubMed: 21397060]

Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG, Macgregor S. A versatile gene-based test for genome-wide association studies. Am J Hum Genet. 2010; 87:139–145. [PubMed: 20598278]

Liu Y, Patel S, Nibbe R, Maxwell S, Chowdhury SA, Koyuturk M, Zhu X, Larkin EK, Buxbaum SG, Punjabi NM, Gharib SA, Redline S, Chance MR. Systems biology analyses of gene expression and genome wide association study data in obstructive sleep apnea. Pac Symp Biocomput. 2011:14–25. [PubMed: 21121029]

Ma L, Clark AG, Keinan A. Gene-based testing of interactions in association studies of quantitative traits. PLoS Genet. 2013; 9:e1003321. [PubMed: 23468652]

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]

McKinney BA, Crowe JE, Guo J, Tian D. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. PLoS Genet. 2009; 5:e1000432. [PubMed: 19300503]

McKinney BA, Pajewski NM. Six Degrees of Epistasis: Statistical Network Models for GWAS. Front Genet. 2011; 2:109. [PubMed: 22303403]

Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genet Epidemiol. 2010; 34:213–221. [PubMed: 19697357]

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, DePristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH Jr, Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012; 485:242–245. [PubMed: 22495311]

Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, Benz C, Haussler D, Stuart JM. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics. 2012; 28:i640–i646. [PubMed: 22962493]

Nitsch D, Tranchevent LC, Thienpont B, Thorrez L, Van Esch H, Devriendt K, Moreau Y. Network analysis of differential expression for the identification of disease-causing genes. PLoS One. 2009; 4:e5526. [PubMed: 19436755]

Pedroso I, Lourdusamy A, Rietschel M, Nothen MM, Cichon S, McGuffin P, Al-Chalabi A, Barnes MR, Breen G. Common genetic variants and gene-expression changes associated with bipolar disorder are over-represented in brain signaling pathway genes. Biol Psychiatry. 2012; 72:311–317. [PubMed: 22502986]

Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI, Xiong M. Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet. 2010; 18:111–117. [PubMed: 19584899]

Perry JR, McCarthy MI, Hattersley AT, Zeggini E, Weedon MN, Frayling TM. Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. Diabetes. 2009; 58:1463–1467. [PubMed: 19252133]

Ragnedda G, Disanto G, Giovannoni G, Ebers GC, Sotgiu S, Ramagopalan SV. Protein-protein interaction analysis highlights additional loci of interest for multiple sclerosis. PLoS One. 2012; 7:e46730. [PubMed: 23094030]

Raj T, Shulman JM, Keenan BT, Chibnik LB, Evans DA, Bennett DA, Stranger BE, De Jager PL. Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. Am J Hum Genet. 2012; 90:720–726. [PubMed: 22482808]

Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends Genet. 2012; 28:323–332. [PubMed: 22480918]

Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 2009; 5:e1000534. [PubMed: 19557189]

Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, Cotsapas C, Daly MJ. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. 2011; 7:e1001273. [PubMed: 21249183]

Ruano D, Abecasis GR, Glaser B, Lips ES, Cornelisse LN, de Jong AP, Evans DM, Davey Smith G, Timpson NJ, Smit AB, Heutink P, Verhage M, Posthuma D. Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. Am J Hum Genet. 2010; 86:113–125. [PubMed: 20060087]

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, Meyer KA, Bilguvar K, Mane SM, Sestan N, Lifton RP, Gunel M, Roeder K, Geschwind DH, Devlin B, State MW. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012; 485:237–241. [PubMed: 22495306]

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]

Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, Wagner M, Lee S, Wright FA, Zou F, Liu W, Downing AM, Lieberman J, Close SL. Genomewide association for schizophrenia in the CATIE study: results of stage 1. Mol Psychiatry. 2008; 13:570–584. [PubMed: 18347602]

Sun J, Jia P, Fanous AH, Webb BT, van den Oord EJ, Chen X, Bukszar J, Kendler KS, Zhao Z. A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. Bioinformatics. 2009; 25:2595–6602. [PubMed: 19602527]

Sun YV. Integration of biological networks and pathways with genetic association studies. Hum Genet. 2012

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010; 26:i237–i245. [PubMed: 20529912]

Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. Cell. 2011; 144:986–998. [PubMed: 21414488]

Wang K, Abbott D. A principal components regression approach to multilocus genetic association studies. Genet Epidemiol. 2008; 32:108–118. [PubMed: 17849491]

Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 2007; 81:1278–1283. [PubMed: 17966091]

Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 2010; 11:843–854. [PubMed: 21085203]

Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, Sleiman PM, Imielinski M, Glessner J, Hou C, Wilson DC, Walters T, Kim C, Frackelton EC, Lionetti P, Barabino A, Van Limbergen J, Guthery S, Denson L, Piccoli D, Li M, Dubinsky M, Silverberg M, Griffiths A, Grant SF, Satsangi J, Baldassano R, Hakonarson H. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am J Hum Genet. 2009; 84:399–405. [PubMed: 19249008]

Wang L, Jia P, Wolfinger RD, Chen X, Grayson BL, Aune TM, Zhao Z. An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. Bioinformatics. 2011a; 27:686–692. [PubMed: 21266443]

Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z. Gene set analysis of genome-wide association studies: methodological issues and perspectives. Genomics. 2011b; 98:1–8. [PubMed: 21565265]

Wang Q, Jia P, Wang L, Feingold E, Cuenco KT, Marazita ML, Zhao Z. Association signals unveiled by a comprehensive gene set enrichment analysis of dental caries genome-wide association studies. PLoS One. 2013; 8:e72653. [PubMed: 23967329]

Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet. 2006; 79:792–806. [PubMed: 17033957]

Wu C, Cui Y. Boosting signals in gene-based association studies via efficient SNP selection. Brief Bioinform. 2013

Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. Genome Biol. 2010a; 11:R53. [PubMed: 20482850]

Wu J, Vallenius T, Ovaska K, Westermarck J, Makela TP, Hautaniemi S. Integrated network analysis platform for protein-protein interactions. Nat Methods. 2009; 6:75–77. [PubMed: 19079255]

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010b; 86:929–942. [PubMed: 20560208]

Xiong Q, Ancona N, Hauser ER, Mukherjee S, Furey TS. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome Res. 2012; 22:386–397. [PubMed: 21940837]

Xu B, Ionita-Laza I, Roos JL, Boone B, Woodrick S, Sun Y, Levy S, Gogos JA, Karayiorgou M. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nat Genet. 2012; 44:1365–1369. [PubMed: 23042115]

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42:565–569. [PubMed: 20562875]

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, Visscher PM. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011; 43:519–525. [PubMed: 21552263]

Zaykin DV, Zhivotovsky LA, Czika W, Shao S, Wolfinger RD. Combining p-values in large-scale genomics experiments. Pharm Stat. 2007; 6:217–226. [PubMed: 17879330]

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining P-values. Genet Epidemiol. 2002; 22:170–185. [PubMed: 11788962]

Zhang M, Liang L, Morar N, Dixon AL, Lathrop GM, Ding J, Moffatt MF, Cookson WO, Kraft P, Qureshi AA, Han J. Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma. Hum Genet. 2012; 131:615–623. [PubMed: 22006220]

Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, Tu Z, Brem RB, Bumgarner RE, Schadt EE. Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. PLoS Biol. 2012; 10:e1001301. [PubMed: 22509135]

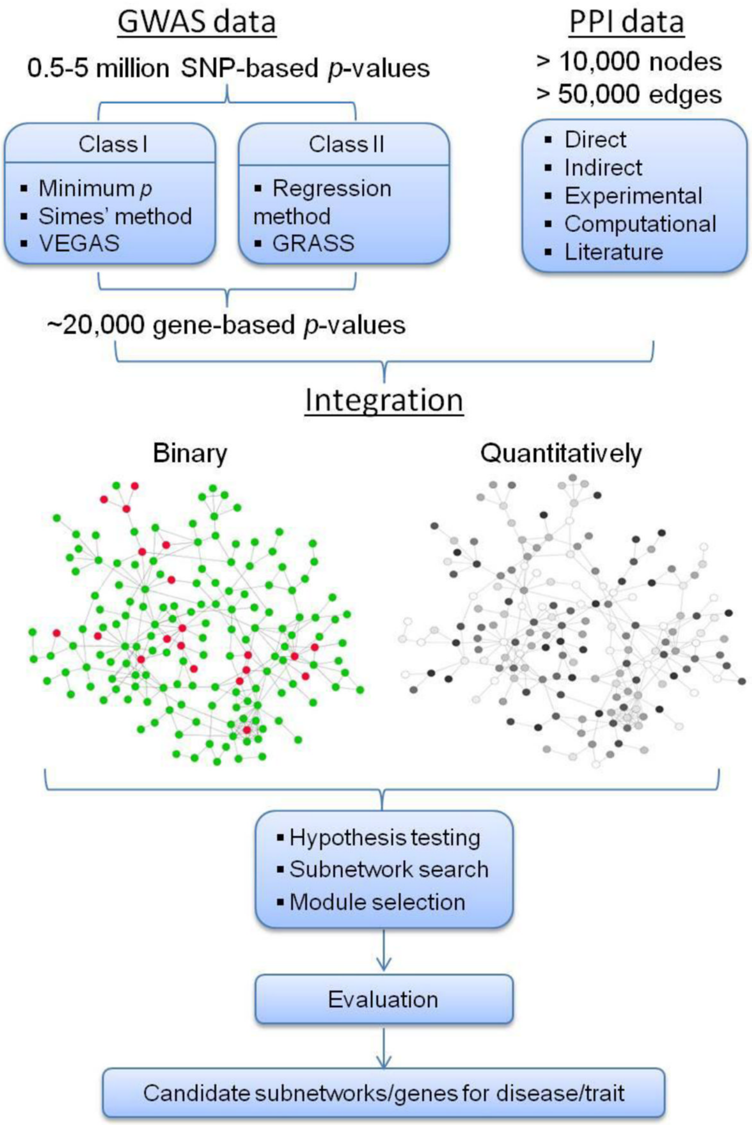

Figure 1.

A general pipeline for network-assisted analysis of data from genome-wide association studies (GWAS).

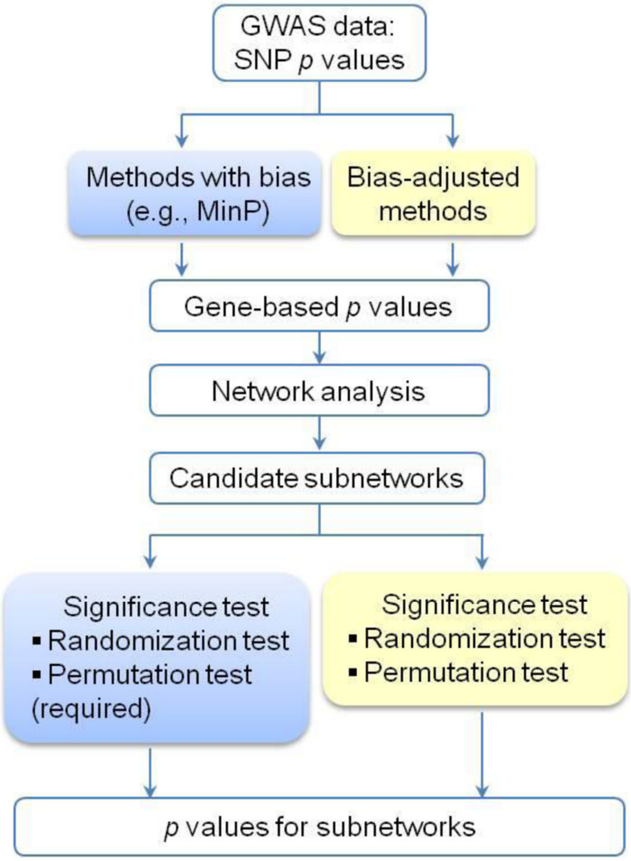

Figure 2.

Overall design of network analysis of GWAS data that addresses potential biases. On the top, the methods in the blue box (e.g., the MinP method) may introduce biases in the computation of gene-based statistics. Therefore, adjustments should be made in downstream steps (blue box in the bottom) by applying a permutation test or other tests. The methods in the yellow box compute gene-based statistics by considering the effects of gene length, LD structure, and SNP density. Thus, an adjustment is not required in those downstream analyses (yellow box in the bottom).

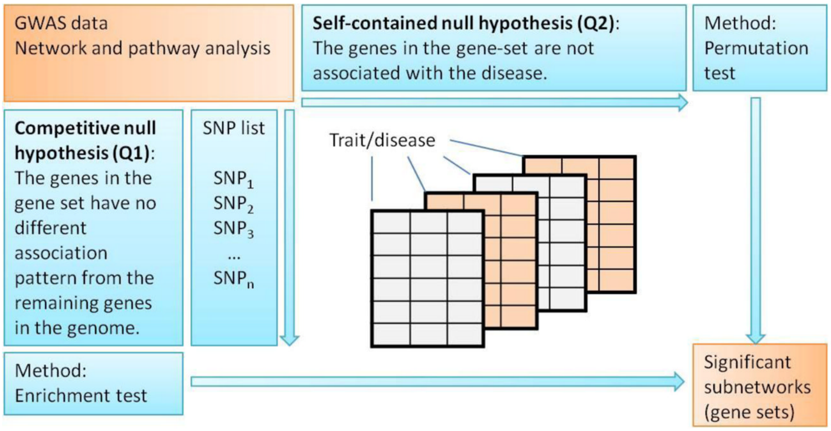
GWAS data
Network and pathway analysis
Self-contained null hypothesis (Q2):
The genes in the gene-set are not associated with the disease.
Method:
Permutation test
Competitive null hypothesis (Q1):
The genes in the gene set have no different association pattern from the remaining genes in the genome.
SNP list
SNP_1
SNP_2
SNP_3
...
SNP_n
Trait/disease
Method:
Enrichment test
Significant subnetworks (gene sets)

Figure 3.

Two types of hypothesis testing in GWAS gene set enrichment analysis. Starting with the original GWAS data, the competitive null hypothesis (Q1) tests whether the genes in a gene set have the same association pattern compared with the remaining genes in the genome. The self-contained null hypothesis (Q2) tests whether the genes in a gene set are associated with the studied disease or not. To test Q1, a randomization of SNP markers or genes is typically adopted, and to test Q2, a permutation of case/control labels is utilized.

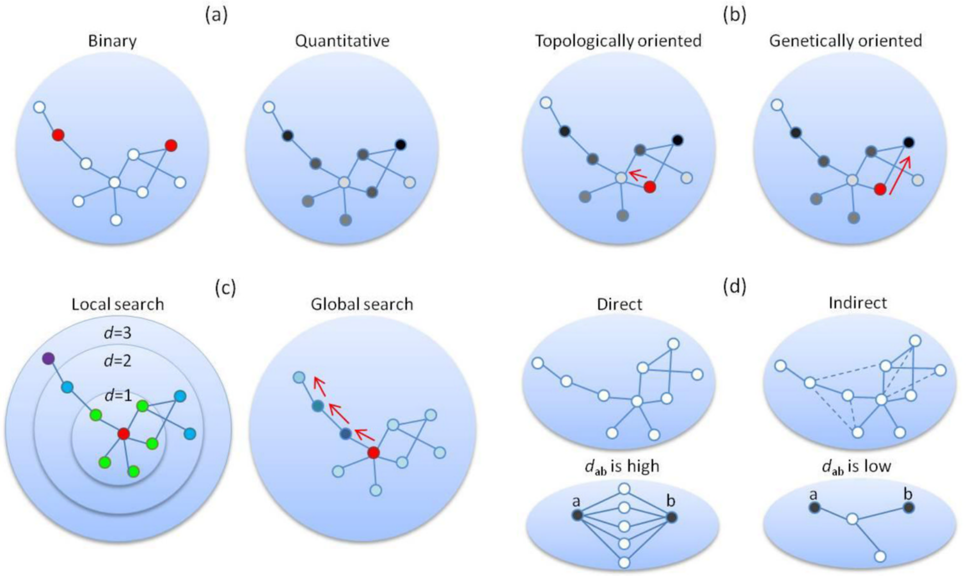

Figure 4.

Algorithms in network-assisted analysis of GWAS data. Nodes represent genes. Red nodes denote seed genes. (a) Two representations of gene association signals: binary (left) versus quantitative (right). Gene-based association signals are indicated by color and darkness: red/white for seed/non-seed genes (left) and darkness level proportional to the association strength (right). (b) Search methods guided by topological characteristics (left) or genetic signals (right). Starting from a seed gene (in red), the former method searches for candidate genes with significant topological characteristics (e.g., genes with high degree), while the latter method searches for candidate genes with strong association signals (nodes in dark color). (c) Search that is restricted by local paths or global paths. Nodes are categorized according to their distance (*d*, defined by shortest path) from the seed gene (in red), e.g., nodes with *d*=1 (in green), *d*=2 (in blue), and *d*=3 (in purple). In local searches, only nodes that are located within a pre-defined distance (e.g., *d*≤2) from the seed genes are considered. In global searches, the search space is not defined by *d* but could theoretically access every node in the network. (d) Search methods in the space of direct and indirect (dash lines) interactions. Suppose d_{ab} represents the closeness between two nodes, a and b, in the network. When many paths exist between the nodes, these nodes are considered to be close to each other in network space, and their closeness (d_{ab}) is high. On the other hand, if there are few paths (in the bottom right panel, only one path exists between nodes a and b), the d_{ab} of these nodes is low.

Table 1

Features of pathway-based and network-assisted analysis approaches

Feature	Pathway-based analysis	Network-assisted analysis
Gene set annotation	Pathways or gene sets from knowledgebase	A reference network
Gene set component genes	Fixed	Dynamic
Search for sub-pathways/subnetworks	No	Yes
Available for topological characteristics	No	Yes
Assess combined association signals	Yes	Available but not required
Function	Focus on one specific function	No particular function boundaries
Correlation amongst genes (or molecules)	Not always available, e.g., GO does not have interactions amongst genes (or molecules)	Available
Hypothesis testing	Q1 vs. Q2*, pre-defined pathways	Multiple and dynamic

*Details of Q1 and Q2 are provided in Figure 3 and the main text.

Table 2

Summary of protein networks used in GWAS data analysis

Source	Curation from literature	Computational prediction	High-throughput experiment	Direct	Indirect		Reference
					Co-expression	Phenotype	
PINA			•	•			(Wu et al. 2009)
HumanNet	•	•	•	•	•		(Lee et al. 2011)
Lage et al.	•	•	•	•		•	(Lage et al. 2007)
Wu et al.	•	•	•	•	•		(Wu et al. 2010a)

Table 3

Published network-based methods for GWAS data analysis

Method	Genetic measurement	Strategy	Association test	Randomization test	Software	Reference
jActiveModule	Quantitative	Combinatory	No	Yes	Stand-alone	(Baranzini et al. 2009; Burcu and Osman Ugur 2011; Liu et al. 2011)
dmGWAS	Quantitative	Combinatory	Yes	Yes	Stand-alone	(Jia et al. 2011b)
NIMMI	Quantitative	Combinatory	No	No	Stand-alone	(Akula et al. 2011)
DAPPLE	Binary	Prioritization	No	Yes	Online	(Rossin et al. 2011)
Lee et al.	Quantitative	Prioritization	No	NA	NA	(Lee et al. 2011)
NetworkMiner	Quantitative	Combinatory	No	Yes	Online	(Garcia-Alonso et al. 2012)

Table 4

Network-assisted GWAS studies

Phenotype/disease	Method	Type	Epub. date	Reference
Multiple sclerosis	jActiveModule	Methodology	Mar. 2009	(Baranzini et al. 2009)
Multiple sclerosis	jActiveModule	Application	May 2010	(Baranzini et al. 2010)
Breast cancer and pancreatic cancer	dmGWAS	Methodology	Nov. 2010	(Jia et al. 2011b)
Obstructive sleep apnea	jActiveModule	Application	Dec. 2010	(Liu et al. 2011)
Rheumatoid arthritis, Crohn’s disease, height, lipids, and type 2 diabetes	DAPPLE	Methodology	Jan. 2011	(Rossin et al. 2011)
Crohn’s disease and type 2 diabetes	Label propagation algorithm	Methodology	May 2011	(Lee et al. 2011)
Seven immune-mediated and autoimmune diseases	DAPPLE	Application	Aug. 2011	(Cotsapas et al. 2011)
Mood disorders	Michigan Molecular Interactions (MiMI)	Application	2011	(Detera-Wadleigh and Akula 2011)
Height, Crohn’s disease	NIMMI	Methodology	Sep. 2011	(Akula et al. 2011)
Rheumatoid arthritis	PANOGA	Application	Oct. 2011	(Burcu and Osman Ugur 2011, 2012)
Bipolar disorder	Greedy search modified from jActiveModule	Methodology	Apr. 2012	(Pedroso et al. 2012)
Bipolar disorder	NetworkMiner	Methodology	Jun. 2012	(Garcia-Alonso et al. 2012)
Schizophrenia	dmGWAS	Application	Jul. 2012	(Jia et al. 2012c)
Alzheimer disease	DAPPLE	Application	Apr. 2012	(Raj et al. 2012)
Multiple sclerosis	DAPPLE	Application	Oct. 2012	(Ragnedda et al. 2012)
Pediatric stroke	dmGWAS	Application	Dec. 2012	(Arning et al. 2012)