Bioinformatics Advance Access published January 9, 2017

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

SigMod

SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network

Yuanlong Liu[1,2,*], Myriam Brossard[1,2], Damian Roqueiro[3], Patricia Margaritte-Jeannin[1,2], Chloé Sarnowski[1,2], Emmanuelle Bouzigon[1,2], Florence Demenais[1,2]

[1]INSERM, Genetic Variation and Human Diseases Unit, UMR-946, Paris, France
[2]Université Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, Paris, France
[3]Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zurich, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Dr. Jonathan Wren

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Apart from single marker-based tests classically used in genome-wide association studies (GWAS), network-assisted analysis has become a promising approach to identify a set of genes associated with disease. To date, most network-assisted methods aim at finding genes connected in a background network, whatever the density or strength of their connections. This can hamper the findings as sparse connections are non-robust against noise from either the GWAS results or the network resource.

Results: We present SigMod, a novel and efficient method integrating GWAS results and gene network to identify a strongly interconnected gene module enriched in high association signals. Our method is formulated as a binary quadratic optimization problem, which can be solved exactly through min-cut algorithms. Compared to existing methods, SigMod has several desirable properties: (i) edge weights quantifying confidence of connections between genes are taken into account, (ii) the selection path can be computed rapidly, (iii) the identified gene module is strongly interconnected, hence includes genes of high functional relevance, and (iv) the method is robust against noise from either the GWAS results or the network resource. We applied SigMod to both simulated and real data. It was found to outperform state-of-the-art network-assisted methods in identifying disease-associated genes. When SigMod was applied to childhood-onset asthma GWAS results, it successfully identified a gene module enriched in consistently high association signals and made of functionally related genes that are biologically relevant for asthma.

Availability: An R package SigMod is available at: https://github.com/YuanlongLiu/SigMod

Contact: yuanlong.liu@inserm.fr, florence.demenais@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have achieved considerable success in genetic analysis of complex traits. Thousands of single nucleotide polymorphisms (SNPs) associated with human traits and diseases have been identified since the first GWA study was published (Klein *et al*., 2005) (http://www.genome.gov/gwastudies/). However, the single marker analysis commonly used in GWAS has limitations. Under the very conservative genome-wide significant level of $P = 5 \times 10^{-8}$, only a few of the most significant signals are reported, while many polymorphisms with small marginal effects are missed. The reported SNPs often explain a limited part of the genetic component of a disease or trait (Maher, 2008; Eichler *et al*., 2010).

To overcome these limitations, a variety of knowledge-based methods have been proposed for integrative and joint analysis of multiple genes. Examples include, but are not limited to the gene set enrichment analysis (GSEA) methods that identify biological pathways enriched in association signals (Subramanian *et al*., 2005); text-mining methods that build links between genes from scientific literature (Raychaudhuri *et al*., 2009), etc. Among these approaches stands the network-assisted analyses that overlay gene-level *P*-values onto a gene network (GeneNet) to search for connected genes (also known as gene module) enriched in association signals. The rationale behind this is the principle of "guilt-by-association", which states that genes (or gene products) connected in a network are usually participating in the same, or related, cellular functions (Oliver, 2000; Wolfe *et al*., 2005; Li *et al*., 2015). Although a number of methods have been developed for this purpose (Ideker *et al*., 2002; Cabusora *et al*. 2005; Jia *et al*., 2011), they often search modules using heuristic or greedy algorithms, hence cannot guarantee to identify the module enriched in highest signals, and are prone to include biologically irrelevant genes by chance. Also, many of them have the limitation that edge weights are not taken into account during the module searching process, although edge weights represent the confidence or strength of connections between genes and can contain useful information.

Azencott *et al*. (2013) have proposed a module searching method that overcomes some of these limitations. Their method, named SConES, was originally developed for identifying a set of SNPs that are maximally associated with a phenotype and tend to be connected in an underlying network. SConES formulates the module searching task as a binary optimization problem that can be solved exactly and efficiently via graph min-cut algorithms. It also allows incorporating edge weights, making it more robust to false connections. Nevertheless, SConES sets its tuning parameters via a cross-validation strategy that requires using raw genotype data, and therefore cannot be applied to studies in which only summary-level statistics are available, as it is often the case in large genetic consortiums. Also, as indicated in their paper, SConES may select several disconnected subnetworks along with multiple isolated nodes, which may lead to an overall low interconnection among selected nodes. These disconnected subnetworks and especially the isolated nodes, are likely to be less functionally related to the other nodes and the selected module may be less associated with disease as compared to a module whose nodes are strongly connected.

In this paper, we propose a novel method SigMod that has the ability to select a Strongly Interconnected Gene MODule maximally associated with the disease. We formulate this module selection task as an optimization problem similar to SConES, but we incorporate a modification in the objective function to explicitly encourage the overall strong interconnection among selected genes. We believe that a set of strongly interconnected genes are more functionally related and biologically relevant. We show that our method has the same advantage as SConES in terms of allowing incorporation of edge weights, and can also be solved exactly and efficiently via graph min-cut algorithms. In addition, we propose an algorithm to compute the module selection path, which provides the ability to trace the selection change and to select a desirable amount of genes. We also develop a parameter setting strategy to identify the optimal selection. Our strategy does not require using raw genotype data, hence can be applied to a broader range of studies than SConES. We evaluated SigMod using both simulated and real data, and made comparisons with SConES and another popular network-based method dmGWAS (Jia *et al*., 2011). The results showed our method is more powerful in identifying a module made of functionally relevant genes and enriched in consistent association signals.

2 Methods

SigMod aims to identify a disease-associated gene module using two types of input data: a list of gene-level *P*-values obtained from GWAS SNP-level *P*-values, and a GeneNet. To get gene-level *P*-values, SNPs need to be first assigned to genes using dbSNP and RefSeq genes with genomic coordinates in the corresponding genome build, but methods vary according to the choice of gene boundaries that can be strictly limited to the start and stop positions of the genes, or extended beyond these positions up to 500 kb. This SNP to gene assignment issue has been previously debated in Jia and Zhao (2014) and will be further discussed in Section 5. Once SNPs have been assigned to genes, gene-level *P*-values, which represent the significance of gene-disease associations, are computed from GWAS SNP-level *P*-values using any gene-based method that has been previously proposed (e.g., Liu *et al*., 2010; Lamparter *et al*., 2011; Li *et al*., 2011). One of the most popular gene-based methods consists of using the best SNP *P*-value assigned to a gene but this *P*-value needs to be corrected for variation in gene length (as explained in Section 4.2). The GeneNet represents the biological knowledge of gene-gene relationships, such as physical interactions between gene products (proteins), gene co-expression or co-occurrence of gene-related terms in the literature. Each connection can have a weight that measures the confidence or strength of the connection. This type of information can be derived from experiments like co-expression analysis or retrieved from databases such as STRING (Szklarczyk *et al*., 2014).

In the following sections, we will first introduce the formulation of SigMod, and then provide an efficient and exact algorithm to solve the optimization problem. Afterwards we will present a tuning parameter setting strategy to find the parameters leading to an optimal gene module selection. A flowchart summarizing these steps is shown in Figure 1.

2.1 Formulation of the SigMod method

We first transform gene-level *P*-values into scores by $z = \Phi^{-1}(1-P)$, where $\Phi^{-1}(\cdot)$ is the inverse normal distribution function. These gene scores are overlaid onto the GeneNet to build a scored GeneNet, denoted as $G = (\boldsymbol{V}, \boldsymbol{A})$, where $\boldsymbol{V}$ are nodes representing genes, and $\boldsymbol{A}$ is the weighted adjacency matrix representing connections among genes. We define $\boldsymbol{u}$ as a vector of binary variables indicating whether a gene $\boldsymbol{V}_p$ is selected ($\boldsymbol{u}_p = 1$) or not ($\boldsymbol{u}_p = 0$). We formulate this selection task as an optimization problem that maximizes the following objective function:

$$f(\boldsymbol{u}) = \boldsymbol{z}^T\boldsymbol{u} + \lambda \boldsymbol{u}^T \boldsymbol{A}\boldsymbol{u} - \eta \|\boldsymbol{u}\|_0 . \tag{1}$$

The first component $\boldsymbol{z}^T\boldsymbol{u}$ defines the joint effect of the gene module on the phenotype (disease) by summing up the scores of its gene members. The second component $\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u}$ quantifies its connection strength as the summed edge weights in the module, since $\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} = \sum_{p,q} \boldsymbol{A}_{pq}\boldsymbol{u}_p\boldsymbol{u}_q$. The third component is the sparsity regularizer controlling the size of the gene module, where the module size is represented by $\|\boldsymbol{u}\|_0$, i.e., the number of non-zero elements in $\boldsymbol{u}$. λ and η are positive tuning parameters specifying the importance of the corresponding components. Therefore, we are able to select a strongly interconnected gene module enriched in high association signals, by choosing proper parameters and solving the optimization problem:

$$\arg\max_{\boldsymbol{u}} f(\boldsymbol{u}) . \tag{2}$$

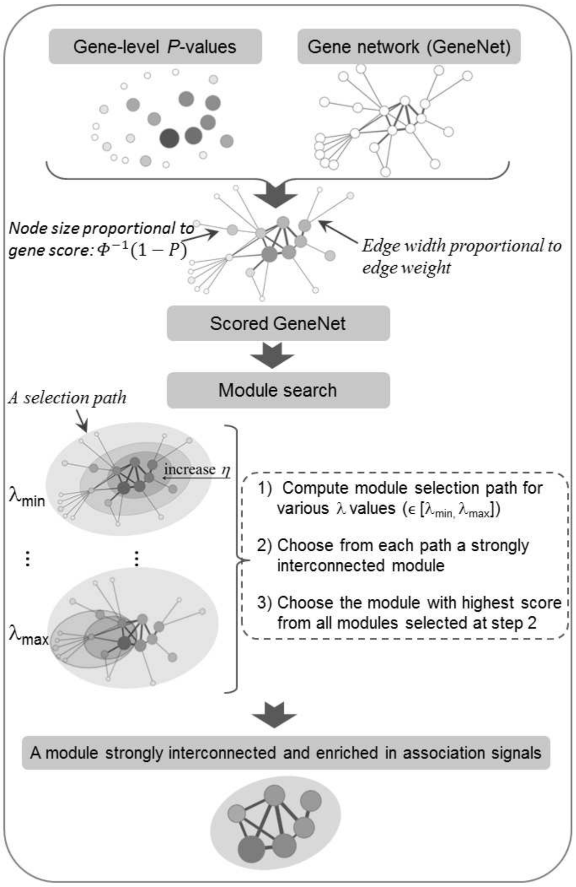
Gene-level *P*-values
Gene network (GeneNet)
Node size proportional to gene score: $\Phi^{-1}(1-P)$
Edge width proportional to edge weight
Scored GeneNet
Module search
A selection path
increase η
$\lambda_{\min}$
$\lambda_{\max}$
1) Compute module selection path for various λ values ($\in [\lambda_{\min}, \lambda_{\max}]$)
2) Choose from each path a strongly interconnected module
3) Choose the module with highest score from all modules selected at step 2
A module strongly interconnected and enriched in association signals

Fig. 1. Workflow of SigMod. SigMod takes a list of gene-level *P*-values computed from genome-wide association studies (GWAS) and a gene network (GeneNet) as input. The gene-level *P*-values are converted into scores and overlaid onto the GeneNet to build a scored network. SigMod identifies a module that is strongly interconnected and enriched in high association signals from this network using a 3-step procedure, as outlined in this figure and detailed in the text (Section 2.3.2). The λ in this figure is the connectivity parameter that controls the balance between module score and module connectivity. The selection path in the figure represents the sequence of distinct modules selected by increasing the sparsity parameter η from a starting value to $+\infty$, as described in Section 2.3.2.

Note that our formulation differs from the formulation of SConES (Azencott *et al.*, 2013). SConES selects genes by maximizing the objective function defined as $g(\boldsymbol{u}) = \boldsymbol{z}^T\boldsymbol{u} - \lambda\boldsymbol{u}^T\boldsymbol{L}\boldsymbol{u} - \eta\|\boldsymbol{u}\|_0$, where $\boldsymbol{L}$ is the Laplacian matrix defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, and $\boldsymbol{D}$ is the diagonal matrix of weighted node degrees, i.e. $\boldsymbol{D}_{pp} = \boldsymbol{d}_p := \sum_q \boldsymbol{A}_{pq}$. The difference between the two objective functions $f(\boldsymbol{u})$ and $g(\boldsymbol{u})$ is in the second component, which leads to different behaviors of each method. Specifically, SConES incorporates the Laplacian matrix to encourage adjacent nodes to be selected together. However, this does not guarantee the overall strong interconnection among selected nodes. Contrariwise, the SigMod formulation incorporates the adjacency matrix to explicitly encourage selection of strongly interconnected nodes. More specifically, since $\boldsymbol{A} = \boldsymbol{D}$ - $\boldsymbol{L}$, it has

$$\begin{aligned} f(\boldsymbol{u}) &= \boldsymbol{z}^T\boldsymbol{u} + \lambda\boldsymbol{u}^T(\boldsymbol{D}-\boldsymbol{L})\boldsymbol{u} - \eta\|\boldsymbol{u}\|_0 \\ &= (\boldsymbol{z}+\lambda\boldsymbol{d})^T\boldsymbol{u} - \lambda\boldsymbol{u}^T\boldsymbol{L}\boldsymbol{u} - \eta\|\boldsymbol{u}\|_0 \ . \\ &= g(\boldsymbol{u}) + \lambda\boldsymbol{d}^T\boldsymbol{u} \end{aligned}$$

Therefore for each given λ, additional scores $\lambda\boldsymbol{d}$ are added to the nodes in SigMod compared to SConES. Nodes with higher degrees are thereby given more preferences. Additional differences between these two methods will be presented in the following sections.

2.2 Optimization algorithm

We show that the optimization of Equation (2) can be solved exactly using a similar graph min-cut approach as presented in Azencott *et al.* (2013). To achieve this, we construct an augmented network of G (denoted as G_{st}), by first adding two artificial nodes s and t to G, then redefining its adjacency matrix as $\boldsymbol{B}$:

$$\begin{cases} \boldsymbol{B}_{pq} = \lambda\boldsymbol{A}_{pq} \\ \boldsymbol{B}_{sp} = (\boldsymbol{z}_p + \lambda\boldsymbol{d}_p - \eta)\times I(\boldsymbol{z}_p + \lambda\boldsymbol{d}_p \ge \eta) \quad \text{for } 1 \le p,q \le n. \\ \boldsymbol{B}_{tp} = (\eta - \boldsymbol{z}_p - \lambda\boldsymbol{d}_p)\times I(\boldsymbol{z}_p + \lambda\boldsymbol{d}_p < \eta) \end{cases} \qquad (3)$$

Definition 1. Given a network $\mathrm{G} = (\boldsymbol{V}, \boldsymbol{A})$, *for any* $s, t \in \boldsymbol{V}$, *a s-t cut* $C = \{\boldsymbol{X}, \bar{\boldsymbol{X}}\}$ *is defined as a node partition of* $\boldsymbol{V}$ *such that: (1)* $\boldsymbol{X} \cup \bar{\boldsymbol{X}} = \boldsymbol{V}$; *(2)* $s \in \boldsymbol{X}$ *and* $t \in \bar{\boldsymbol{X}}$.

Definition 2. A s-t cut is called a s-t min-cut if $\kappa(C)$ *is minimized, where* $\kappa(C)$ *is the capacity of a s-t cut C, defined as* $\kappa(C) = \sum_{\boldsymbol{V}_p \in \boldsymbol{X}, \boldsymbol{V}_q \in \bar{\boldsymbol{X}}} \boldsymbol{A}_{pq}$.

Therefore according to Proposition (1) in Azencott *et al.* (2013), if $C^* = (\boldsymbol{X}^*, \overline{\boldsymbol{X}^*})$ is a *s-t* min-cut of G_{st}, then $\boldsymbol{u}^*$ is the solution of the optimization problem of Equation (2), where $\boldsymbol{u}_p^* = 1$ if $\boldsymbol{V}_p \in \boldsymbol{X}^*$, and $\boldsymbol{u}_p^* = 0$ otherwise. Hence solving the optimization problem is equivalent to finding a *s-t* min-cut on the augmented network G_{st}. Thus any *s-t* min-cut algorithm can be applied to find the solution.

2.3 Determination of the tuning parameters η and λ

The SigMod objective function Equation (1) includes two tuning parameters, η and λ, that need to be determined. To find the parameter values leading to an optimal gene module selection, we first propose a path algorithm that allows computing all distinct selections at a given λ while varying η over a range of values. Based on this algorithm, we provide a procedure to find the tuning parameters that can lead to the optimal gene module selection. These different steps are described as follows.

2.3.1 Computing the selection path at any given λ value

For a given value of λ, the module selection by solving Equation (2) has the *nesting property* that $S(\eta_1) \subseteq S(\eta_0)$ if $\eta_1 > \eta_0$, where $S(\eta)$ represents the module selected by setting the sparsity parameter as η (see Supplementary Materials for proof). Therefore increasing η results in removing genes from a previously selected module. To conveniently trace this selection change, we develop the *path algorithm* that allows computing the sequence of distinct modules selected by increasing η from $\eta_{\min}$ to $\eta_{\max}$ $(0 \le \eta_{\min} < \eta_{\max})$. We denote this sequence as $\mathrm{P} = \langle S(\eta_{\min}),\ldots,S(\eta_{\max})\rangle$ and call it as the selection path over $[\eta_{\min}, \eta_{\max}]$. Note that these modules are nested according to the nesting property, i.e., $S(\eta_{\min}) \supseteq \cdots \supseteq S(\eta_{\max})$. An example of selection path is given in Figure S1.

Our path algorithm aims to compute P efficiently. It is developed by exploring the property of *s-t* min-cut on the augmented graph G_{st}, since computing $S(\eta)$ is equivalent to finding the *s-t* min-cut as stated in Section 2.2. We define the capacity function $\kappa^*(\eta)$ as the capacity of the *s-t* min-cut on G_{st}, where the capacity of a cut is defined in Definition 2 (Section 2.2). It is apparent that $\kappa^*(\eta)$ is a continuous and piecewise linear function of η. Its slope changes at either a break-point or a change-point, where a value of η is a break-point if it leads to the change of selection, and is a change-point if it causes the rewiring of an edge of G_{st} from *s* to *t* according to Equation (3). Thus, computing the selection path is equivalent to finding all break-points of $\kappa^*(\eta)$, which can be achieved by correcting $\kappa^*(\eta)$ at each change-point to transform it to a concave function, then applying the iterative contraction algorithm described in Gallo *et al*. (1989). Once all break-points are obtained, the selection path can be computed by setting η at each of the break-points and solving the problem defined by Equation (2). A detailed description of this algorithm is presented in Supplementary Materials.

We also notice that the module selection by solving Equation (2) has the *memoryless property*, that if a gene is not selected by setting η at some value (e.g., $\eta = \eta_{\min}$), then it can be removed from the GeneNet when computing the selection at a η value greater than $\eta_{\min}$. The mathematical description of this property and its proof is given in Supplementary Materials (Proposition 1). This property can be utilized to speed up the computation of selection path over $[\eta_{\min}, \eta_{\max}]$, using the following two-step procedure:

- Step 1: compute $S(\eta_{\min})$ on the complete network G;
- Step 2: compute selection path over $[\eta_{\min}, \eta_{\max}]$ on the subnetwork G_{sub} induced by the genes in $S(\eta_{\min})$.

This speed-up strategy makes the computation of selection path more efficient, especially when the size of $S(\eta_{\min})$ is far less than the total amount of genes in the whole network. It is the case for many studies in which only a small portion of genes are intended to be selected while the majority of genes are left out at the first stage of the selection process.

2.3.2 Hierarchical procedure to find the tuning parameters leading to an optimal gene module selection

As mentioned above, the module selection in SigMod depends on two parameters η and λ. The selection as a function of η can be tracked through the selection path at any given value of λ. The parameter λ, which allows a balance between the module score and module connectivity, needs to be chosen carefully. On one hand, if λ is too small, the selection mainly focuses on gene scores while it ignores the connections among genes. This results in the top scored genes scattered in the network to be selected, whichever their connections. On the other hand, if λ is too big, the network topology dominates the selection, while the gene scores do not influence the module selection. This leads to a set of most strongly interconnected genes to be selected, whichever their association scores. Since the goal of our method is to find a gene module that is strongly interconnected and is enriched in high association signals, we propose the following procedure to set the parameters properly:

- Step 1: do an exhaustive search for k equally spaced λ values in $[\lambda_{\min}, \lambda_{\max}]$. Compute for each λ the selection path $\mathrm{P}(\lambda)$ to collect all modules with module size less than max_select;
- Step 2: compute the size difference between consecutive modules in each path $\mathrm{P}(\lambda)$, i.e. $\Delta s_i = |S_i| - |S_{i+1}|$, where S_i is the i^{th} module in the path. Then choose S_{i^*} within each path, where $i^* = \max\{i \mid \Delta s_i \ge \tau\}$;
- Step 3: remove genes not connected to others in each S_{i^*}. Choose from all resulting S_{i^*} the one with highest standardized score as final selection, denoted as S^*.

In Step 1, we explore the module selection for k different λ values. For each value, we calculate its selection path $\mathrm{P}(\lambda)$ to collect all distinct modules whose number of genes is less than max_select (specified by the user). This can be achieved by starting at a trial value $\eta = \eta_0$ and computing the path over the sparsity range $[\eta_0, \infty]$. If $|S(\eta_0)| < max\_select$, decrease η and compute the path in the extended range, until the size of the largest selected module surpasses max_select. The range $[\lambda_{\min}, \lambda_{\max}]$ should be broad enough, so that an optimal selection is contained in these paths. Though exhaustive search is potentially expensive, the incorporation of our speed-up path algorithm can largely reduce the computational burden.

In Steps 2, the goal is to find a *local optimum* module within each path, where by *local optimum* we mean the selected module is strongly interconnected and enriched in high scores relative to that path. We identify this *local optimum* by examining the size difference between consecutive modules in P, i.e., $|S_1| - |S_2|$, $|S_2| - |S_3|$, etc. This is because, by our formulation, if the connectivity regularizer does not have an effect, the genes will be removed one by one from the module; while if the regularizer has an effect, some strongly interconnected genes are non-separable and are removed together, which corresponds to a large size jump (τ) between consecutive selections in the selection path, as shown in Figure S2. We select the smallest module in the path that contains such non-separable genes (by choosing $i^* = \max\{i \mid \Delta s_i \ge \tau\}$). We set $\tau = 5$ by default, but it can be adjusted based on actual situation.

In the final step, we first remove the genes that are not connected to any other gene in each *local optimum* module. Then we choose from these local optima the one with highest standardized score, where the standardized score of a module S is defined as

$$z^*(S) = \frac{z(S) - |S| \times \hat{\mu}}{\sqrt{|S|}\hat{\sigma}}.$$

Here $z(S) = \sum_{s \in S} z_s$. $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and standard deviation of all gene scores in GeneNet.

A summary of this procedure is shown in Figure 1. Through this hierarchical procedure we increase the possibility to find the true disease-associated gene module.

3 Implementation

We implemented our method in an R package *SigMod* (available at https://github.com/YuanlongLiu/SigMod). SigMod takes a list of gene-level *P*-values and a GeneNet as input. Each connection in the GeneNet can be assigned a weight to quantify the confidence or strength of the connection. When the weight of a connection is unavailable, it can be specified as 1 or 0 to indicate presence or absence of the connection.

The *SigMod* package consists of the main function *select_subnet* to solve the optimization problem of Equation (2); the *selection_path* function to calculate the selection path as described in Section 2.3.1; and additional functions to help identify the optimal module selection. We use the *graph.maxflow* function in R package *igraph 0.7.1* (Csardi and Nepusz, 2006) to find the *s-t* min-cut. It implements the Goldberg-Tarjan Push-Relabel algorithm (Goldberg and Tarjan, 1988), and has the smallest known time complexity of $\mathrm{O}\left(n_1 n_2 \log\left(n_1^2 / n_2\right)\right)$, where n_1 is the number of genes in GeneNet and n_2 is the number of connections.

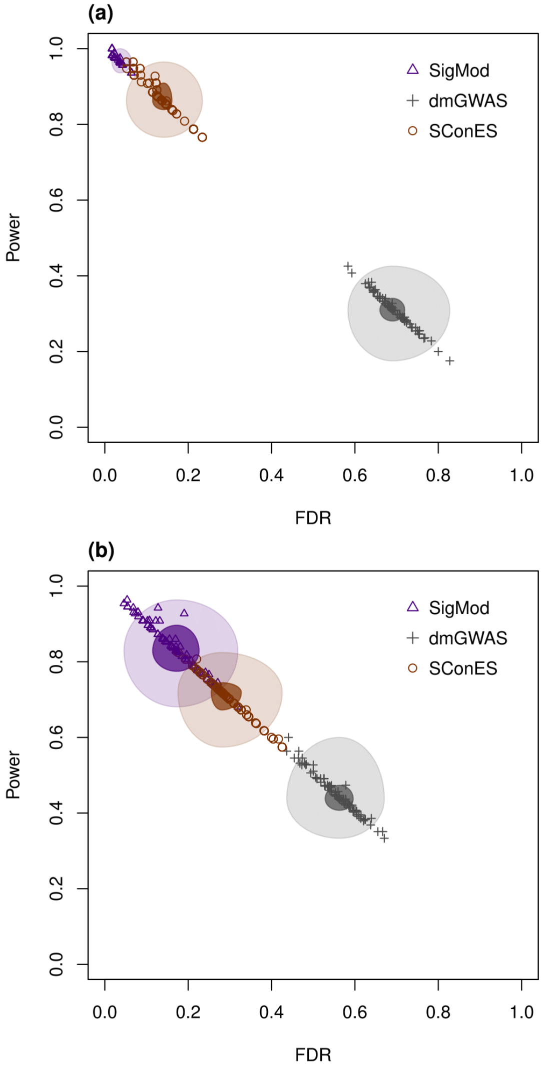

Fig. 2. False discovery rate (FDR) versus power of three network analysis methods applied to simulated data. The results of 20 replicates of five causal modules are aggregated. Five-number statistics (minimum, first quartile, median, third quartile, and maximum) of each quantity are shown by ellipse plot (Tomizono, 2013). Plot (a) shows the results without adding noise to the GWAS data or GeneNet. Plot (b) shows the results with noise added to both GWAS data and GeneNet.

4 Results

We evaluated the performance of SigMod using both simulated and real datasets. We downloaded a comprehensive human GeneNet from the STRING database version 10 (Szklarczyk *et al*., 2014), which contains information on various types of connections among genes. This GeneNet includes 19,247 genes and 4,274,001 edges. Each edge represents a known or predicted interaction between genes or gene products (proteins), including direct (physical) and indirect (functional) associations derived from four sources including systematic genome comparisons, high-throughput experiments, co-expression and previous knowledge from literature. Each edge in the STRING GeneNet is assigned a weight varying from 0 to 1, which represents the combined confidence of the connection between two genes derived from different sources of information.

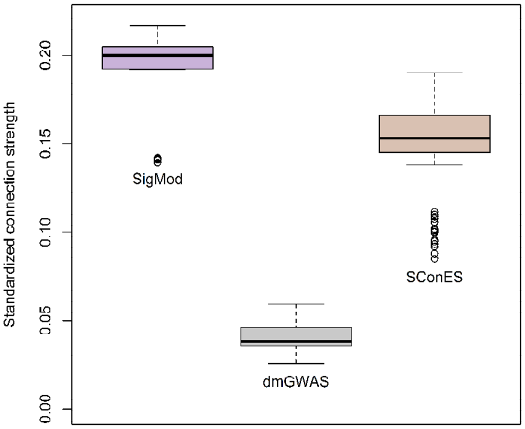

Fig. 3. Box plots of the standardized connection strength (ρ) of gene modules identified by three network analysis methods (SigMod, SConES and dmGWAS). The results of 20 replicates of five causal modules were aggregated.

4.1 Results of the simulated data

We first conducted simulations using the STRING GeneNet. We chose five strongly interconnected gene modules identified by CFinder (Adamcsek *et al*., 2006) as candidate causal modules (Figure S3). The sizes of these modules ranged from 47 to 87.

In each simulation, a single module was set as the causal module. We followed the proposal of Rajagopalan and Agarwal (2005) to set *P*-values of the genes belonging to the causal module to be uniformly distributed between 0 and 10^{-3}. *P*-values of other genes were uniformly distributed between 0 and 1. We set $[\lambda_{\min}, \lambda_{\max}] = [0.005, 0.05]$ and computed selection paths for $k = 100$ λ values in this range. Other parameters were set as $\tau = 5$ and $max\_select = 1000$.

We compared our method with two state-of-the-art module search methods dmGWAS (Jia *et al*., 2011) and SConES (Azencott *et al*., 2013). The dmGWAS method identifies gene modules by starting from each gene in the GeneNet and repeatedly adding neighboring genes that generate the maximum increment of the module score ($z(S) = \sum_{s \in S} z_s$). Module growth terminates if adding neighboring genes does not yield more than r% (r=10 by default) increment of the score. As in dmGWAS the number of genes to be selected is determined by the user, we selected approximately the same number of genes as that of the causal module under study. To do so, we first set parameters to their default values to generate raw modules. Then we ordered the raw modules according to their module scores. Top modules were selected sequentially until the cumulative size of these modules exceeded that of the causal module. The SConES method, as described in Section 2.1, selects genes by maximizing the objective function $g(\boldsymbol{u})$. It should be noticed that its original implementation uses a cross-validation approach to set tuning parameters, which does not apply to our study as raw genotype data are not used. Nonetheless, according to the relationship between $f(\boldsymbol{u})$ and $g(\boldsymbol{u})$ described in Section 2.1, it is straightforward that our path algorithm can also be applied to SConES. Thereby, we computed its selection paths using the same λs as for SigMod. In each path, we chose the first module selection whose size exceeded that of the causal module. Among these selections we chose the one with largest standardized score.

We ran 20 repetitions for each of the five candidate gene modules (hence 20×5 experiments for each method). We computed the power (fraction of causal module genes selected) and false discovery rate (FDR, fraction of selected genes that are not causal) of each experiment. SigMod has systematically higher power and lower FDR over all experiments, as presented in Figure 2 (results are aggregated for all experiments; see Figure S4 for individual results). SConES has lower power and higher FDR than SigMod while dmGWAS performs worst in these simulations. We further compared the standardized connection strength of the selected modules, defined as $\rho = 2\omega / m(m-1)$, where m is the module size; ω is the sum of pairwise edge weights in the module. As shown in Figure 3 and Figure S5, the connection strengths of gene modules selected by SigMod are much higher than the other two methods.

The performance of these methods against noise was also evaluated. Two sources of noise were considered simultaneously. The first one is standard Gaussian noise added to the scores of the causal module genes. The second noise is added to the topology of GeneNet by randomly rewiring 5% of the edges, where at a rewire step, two edges $V_1 \sim V_2$, $V_3 \sim V_4$ becomes $V_1 \sim V_4$, $V_3 \sim V_2$. This rewire process keeps the distribution of node degree unchanged. We observed that SigMod still has the best performance among the three methods, with an average power of 0.83 and FDR of 0.18 (Figure 2 and Figure S4). Interestingly, dmGWAS has an improved performance when noise is added (higher power and lower FDR). This is because it selects genes with highest scores. By adding Gaussian noise to the scores, genes with increased score are more likely to be selected.

4.2 Identification of a gene module associated with childhood-onset asthma

We applied SigMod to the meta-analysis results of 18 childhood-onset asthma GWASs, which were part of the European GABRIEL asthma consortium (Moffatt *et al*., 2010). The data are described in detail elsewhere (Moffatt *et al*., 2010). In order to check the consistency of results of SigMod, we used a discovery-evaluation scheme. Therefore, the 18 childhood-onset asthma GWASs were randomly split into two groups of 9 GWASs while preserving a similar sample size for the two groups: 3,031 cases/2,893 controls in the first group for discovery and 2,679 cases/3,364 controls in the second group for evaluation. In each group, a meta-analysis was applied to 2,370,689 single SNP association statistics (Hapmap2-imputed SNPs after quality control), using a random-effects model implemented in STATA V12 (distributed by Stata Corporation, College Station, Texas, USA). The results of these two meta-analyses (SNP-level *P*-values) were respectively named META1 and META2.

To aggregate SNP-level results into genes, SNPs were mapped to genes (between the start site and 3'-untranslated region of each gene) using dbSNP Build 132 and human Genome Build 37.1, making a total of 24,120 genes with at least one SNP mapped. Each gene-level *P*-value was taken as the best SNP *P*-value among all SNPs mapped to the gene, and was further corrected for gene length using permutations. We applied the Circular Genomic Permutation (CGP) approach that can preserve linkage disequilibrium (LD) among SNPs when permuting SNP-level statistics (Cabrera *et al*., 2012). It was shown to have similar performance to the highly time-consuming gold standard of phenotype permutation (Brossard *et al*., 2013). These corrected gene-level *P*-values were converted to scores by inverse normal transformation. The scores were mapped to the STRING-based GeneNet to build a scored network, which consisted of 15,724 genes and 3,055,850 edges.

We applied SigMod to the META1 discovery set. We used the same parameter settings as described in simulations, i.e. $\lambda_{\min} = 0.005$, $\lambda_{\max} = 0.05$, $k = 100$, $\tau = 5$ and $max\_select = 1000$. We identified a strongly interconnected gene module of 190 genes and 1,295 connections (Figure S6).

4.2.1 Enrichment of the identified gene module in high association signals

The selected gene module has a standardized score of 36.09, which is significantly higher than the scores of 100,000 random modules (each has the same number of genes as in the identified module) sampled from the scored GeneNet ($P < 10^{-5}$; Figure S7). All module genes have significant *P*-values ($P \leq 0.05$), ranging from 5.48×10^{-6} to 1.88×10^{-2}. These *P*-values are ranked at the top of the whole gene list, with highest rank of 1 and lowest rank of 581 (Table S1).

We then evaluated whether the selected gene module was enriched in consistent association signals, by computing its score using META2 dataset. The gene module had a standardized score of 5.85, which was again significantly higher than scores of 100,000 randomly generated modules ($P < 10^{-5}$; Figure S7). This shows the ability of SigMod to select a module displaying consistent association signals.

4.2.2 Association of the identified gene module with asthma

The association of the identified module with childhood-onset asthma was evaluated through CGP permutation of SNP *P*-values that can preserve the genomic structure, using META1 and META2 respectively. For each evaluation, a total of 100,000 CGP samples were generated and scores of the identified gene module were recomputed using these samples. The observed score of the identified module was significantly higher than those obtained from the permutation samples ($P < 10^{-5}$ evaluated using either META1 or META2) (Figure S8). This shows the gene module is significantly associated with childhood-onset asthma.

4.2.3 Functional clustering and annotations of genes belonging to the identified gene module

Our method is based on the “guilt by association” principle. To explore the functional relatedness of genes belonging to the selected module, we used the gene functional classification tool of the DAVID Bioinformatics Resource (Huang *et al*., 2009). This tool generates a gene-to-gene similarity matrix based on shared functional annotation profiles using over 75,000 terms from 14 annotation sources and classifies highly related genes into functionally related groups. We identified nine functional gene clusters of which seven included genes having strong connections within our selected module (Figure 4 for these seven groups and Figure S9 for the additional two groups). Altogether the nine functionally related groups included 68 out of the 190 module genes (36%). The function of each gene cluster was annotated by the most representative gene ontology (GO) category shared by all genes within a cluster and with highest (or close to highest) enrichment in these genes. For the seven clusters with strong gene-gene connections, these GO categories corresponded to the MHC protein complex, known to be associated with many immune-related diseases including asthma, and potentially novel mechanisms such as nucleosome assembly, regulation of ubiquitin-protein ligase activity, protein catabolic process, zinc ion binding, as well as regulation of transcription (clusters 6 and 7) which plays a key role in autoimmune diseases (Farh *et al*., 2015) that share susceptibility loci with asthma (Welter *et al*., 2014).

Finally, we performed KEGG pathway enrichment analysis to further annotate the module genes. We used the enrichKEGG function of the R package clusterProfiler (Yu *et al*., 2012), which interrogates KEGG on the fly to get the latest pathway information. We found 15 pathways

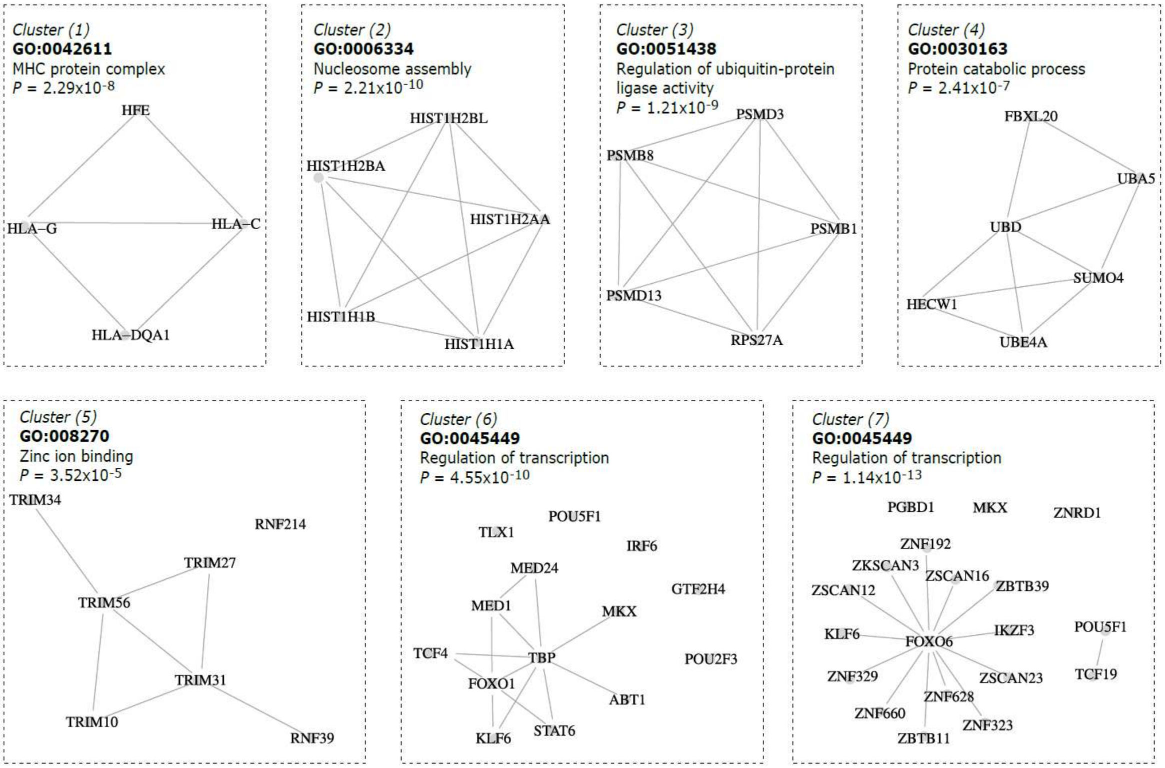

Fig. 4. Seven strongly interconnected functional gene clusters identified by DAVID in the selected gene module associated with childhood-onset asthma. The main function of each cluster is represented by the gene ontology (GO) category that has highest enrichment in the cluster genes. The *P*-values correspond to the significance of enrichment of the shown GO term in the corresponding gene cluster.

(Table S2) significantly enriched in genes from the identified module (FDR<0.05). Of particular interest is that five KEGG pathways are related to virus infection, which supports previous findings of the modulating effect of genetic variants associated with asthma at the 17q21 locus on the association of asthma with viral infections (Smit *et al*., 2010; Çalışkan *et al*., 2013). Moreover, the antigen presentation pathway was already identified by DAVID as the MHC complex GO, and the Inflammatory Bowel Disease and Type 1 Diabetes pathways represent two auto-immune diseases that share susceptibility loci with asthma (Welter *et al*., 2014). All of this adds further evidence that the selected gene module includes genes of functional relevance for asthma.

4.2.4 Comparison of results using SigMod, dmGWAS and SConES

For purpose of comparison, we also applied dmGWAS and SConES to the META1 dataset to identify modules. We used the same strategy as described in the simulation study to select approximately the same number of genes as selected by SigMod. We compared the identified modules for their enrichment of association signals (quantified by the module score $z*$), and evaluated the replicability of these signals in the independent META2 dataset.

As shown in Table 1, the module identified by SConES has a slightly lower score than the module selected by SigMod. All genes of SigMod and SConES modules have a significant *P*-value ($P \leq 0.05$), hence are likely to be bona fide genes. Comparatively, the module identified by dmGWAS has a score that is twice as small as the SigMod module score. Also only half of its module genes have a significant *P*-value. This shows dmGWAS has a lower ability to identify genes having strong association signals. This is likely because: dmGWAS uses a heuristic search algorithm that does not guarantee the maximization of the module score; while SigMod and SConES use exact algorithms to ensure the maximization.

When these modules were evaluated for replication of results using the independent META2 dataset, the module identified by SigMod again had the highest score (see Table 1). Specifically, 30 genes out of the 190 genes were significant when evaluated from META2 (Table S1), hence were significant in both META1 and META2 and are thus of biological interest. These module genes account for almost half of the 70 genes in the GeneNet that are significant in both datasets, demonstrating the ability of SigMod to identify genes displaying consistently high association signals. Comparatively, the signals in the module identified by dmGWAS or by SConES were less replicated, as indicated in Table 1 by the module score and the number of significant genes evaluated using META2. Specifically, 18 out of 190 genes identified by SConES from META1 remained significant in META2. This lower replication rate (60% of the SigMod replication rate) may be due to the lower overall interconnection among genes selected by SConES. As shown in Table 1, the number and strength of connections between genes in the SConES module are both 18% of the values observed in the SigMod module. These genes with lower overall connection strength are likely to be less functionally relevant, and to have a less consistent joint effect on disease.

As for computational efficiency, all three methods (with SConES using our tuning parameter setting strategy) have comparable run time of ~3h on a server (2.66 GHz Intel® Xeon® Processor X5650 and 160 GB of RAM).

Table 1. Comparison of the performance of SigMod with dmGWAS and SConES in identifying a gene module associated with childhood-onset asthma. Various features of the identified gene module were compared, including the number of genes and edges, the connection strength (ρ), the standardized module score (z^*), and the number of nominally significant genes (#sig) in the module. META1 and META2 are the two datasets consisting of SNP-level *P*-values obtained from meta-analyses of childhood-asthma GWAS.

				META1		META2	
	#genes	#edges	ρ	z^*	#sig	z^*	#sig
SigMod	190	1295	0.022	36.08	190	5.85	30
SConES	190	232	0.004	35.52	190	4.14	18
dmGWAS	191	679	0.011	17.18	92	3.65	25

5 Discussion and conclusion

Network-assisted analysis of GWAS data to identify gene modules enriched in high association signals has received increasing attention over the last decade. In this article, we proposed a novel method SigMod, tailored for such purpose. SigMod takes a gene network and a list of gene-level *P*-values as input. The gene network can be retrieved from databases or derived from experiments that are best suitable to the study. In our application to the asthma data, we chose the STRING network that has the advantage of integrating connection information from various sources. The gene-level *P*-values, which represent the significance of gene-disease association, can be computed from GWAS SNP-level *P*-values using any proper gene-based methods (e.g., Liu *et al*., 2010; Lamparter *et al*., 2011). In our study, gene-level *P*-values were chosen as the best SNP *P*-value in a gene and were corrected for gene length using Circular Genomic Permutation that can preserve the LD pattern between SNPs. One challenge in network-assisted analysis is the assignment of SNPs to genes, as discussed in Jia and Zhao's review (Jia and Zhao, 2014). In our study, we used a stringent definition of gene boundaries, which were represented by the start site and 3'-untranslated region of each gene to reduce false positives. Although gene boundaries can be extended to a few kilobases both upstream and downstream of a gene, it was shown that a change of boundaries from 0 to 250 kb did not significantly affect the power of the related network analysis (Lee *et al*, 2011), although this needs to be further confirmed. Moreover, extension of boundaries to flanking regions of a gene may increase the degree of overlap of nearby genes and thus the number of wrong SNP-to-gene assignments. More sophisticated SNP to gene annotation strategies that take into account functional information, such as gene expression through expression quantitative trait loci (eQTLs), or that define a regulatory domain for each gene (McLean *et al*, 2010), may be considered. However, the performance of such annotation strategies with respect to the classical ones need to be further assessed.

SigMod selects a strongly interconnected gene module enriched in association signals by optimizing a binary quadratic objective function. We showed the optimization problem can be solved exactly through graph min-cut algorithms. We also designed a path algorithm that allows computing the selection path at any given λ value. This provides the flexibility to select an appropriate number of genes. In combination with the path algorithm, we proposed a strategy that enables choosing proper parameters to keep a balance between module score and module connectivity. This strategy does not require using raw genotype data. We believe that a proper parameter setting strategy is as important as the formulation of the objective function, as inappropriate parameters can lead to unwanted results, especially for network-assisted analysis where numerous gene modules can be selected. Comparatively, in the original SConES method the parameters are determined using a cross-validation approach, which cannot be applied to situations where raw genotype data are unavailable, as often encountered.

In comparison to previous approaches that only require the selected genes being connected in a network, SigMod encourages selecting genes having overall strong interconnection. This emphasis is well grounded as the identified module is more robust against noise. In particular, genes that have some false connections in the selected module may still be kept in the module after removing such connections, whereas for a loosely interconnected module, removal of false connections may destroy the module structure. Also, a strong interconnection among genes can reflect close functional relationships, as implied by the "guilt by association" principle and demonstrated by our application to the asthma dataset.

SigMod has a different focus compared with SConES. Specifically, SConES focuses on co-selection of adjacent nodes rather than the overall strong interconnection among selected nodes. The node preference between SigMod and SConES is also different. SConES favors low degree nodes while SigMod rewards nodes of higher degrees, as indicated in Section 2.1. We believe that rewarding high degree nodes is particularly suitable for some applications. It has been widely observed that many disease-causing genes have high degrees in a gene network, especially those playing a central role in complex diseases (Lee *et al.* 2013; Xiong *et al*., 2014). These genes can even show higher connectivity in an integrated gene network (e.g. STRING) that aggregates connection information from various sources. Although SigMod rewards genes of higher degree, the scale of rewarding is controlled by a tuning parameter λ. This parameter keeps the balance between the module score and the connectivity, which can be chosen properly using our parameter setting strategy. The validity of this strategy was verified in the simulation study and in the application to asthma GWAS data, where all selected genes were nominally significant (after correction for gene length) and were ranked at the top of the gene list in the whole network (Table S1). We did not observe any gene was selected just because it is a hub gene even when it had a very low score.

In our simulations we found SigMod outperforms SConES and another state-of-the-art method dmGWAS. It has the best power and lowest false discovery rate. This high performance was preserved in presence of noise from both GWAS results and network information, demonstrating its robustness. Further application of SigMod to childhood-onset asthma GWAS results successfully identified a gene module significantly associated with disease. The analyses of functional relationships among genes highlighted known asthma-related gene functions and novel ones which allow generating new hypotheses regarding the mechanisms underlying asthma pathogenesis. Though the module identified by SConES was also enriched in high association signals in the META1 discovery dataset, these signals were less well replicated in the independent META2 dataset. A possible explanation is that the genes in the SConES module are less connected than those identified by SigMod, as reflected by the overall connection measure (ρ). They are thus likely to be less functionally related and may have a less consistent joint effect on disease. This emphasizes again the importance of favouring strong interconnection as achieved by SigMod.

To our knowledge, our method is one of the very few methods in related work that both take edge weights into account and can be solved using exact algorithms. As there are emerging approaches to define connections among genes (e.g., physical or functional, experiment verified or computational based interaction), edge weights are an important indicator of the confidence or strength of the connection. For those methods that do not incorporate edge weight, an arbitrary hard cutoff has to be given to define the presence or absence of a connection, which can lose useful information.

Our current formulation of SigMod did not take into account the LD pattern that may exist among SNPs belonging to adjacent genes or gene clusters in a chromosomal region that may share similar functions. This may cause over selection of genes belonging to such clusters. However, when many genes possess high scores but are in the same LD interval, the algorithm picks automatically those having stronger connections with other genes located in different chromosomal regions. This matches the concept of Taşan *et al*. (2015) that genes with more connections are of higher importance. Nonetheless, SigMod is different from their approach, in that the algorithm decides itself the optimal number of genes to be selected in a LD interval, instead of given a "prix fixe" constraint to select only one gene from it. We believe this is more rational as it is generally unsure whether there is only one causal gene in a LD interval.

In conclusion, we proposed an exact and efficient method SigMod for integrative analysis of GWAS data with network-based knowledge. Our method enables to find a functionally relevant gene module enriched with high association signals. It is robust against noise from either the GWAS results or the background network. Though our method is especially designed for identifying a gene module associated with disease (or trait), it can be applied to any other network-assisted feature selection problem of the same concept.

Acknowledgements

Funding: This work was funded by the Marie Curie Initial Training Network MLPM2012, Grant No. 316861. This work was also supported by the French National Agency for Research ANR-USPC-2012-EDAGWAS, ANR-11-BSV1-027-GWIS-AM and ANR-15-EPIG-0004-05 and a contract from the European Commission (018996).

Conflict of Interest: none declared.

References

Adamcsek, B. *et al.* (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**(8), 1021–1023.

Azencott, C.-A. *et al.* (2013). Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, **29**(13), i171–i179.

Brossard, M. *et al.* (2013). Comparison of permutations strategies to assess gene-set significance in gene-set-enrichment analysis. *Abstracts from the 22nd Annual Meeting of the International Genetic Epidemiology Society*. Page 15.

Cabrera, C. P. *et al.* (2012). Uncovering networks from genome-wide association studies via circular genomic permutation. *G3: Genes | Genomes | Genetics*, **2**(9), 1067–1075.

Cabusora,L. *et al*. (2005). Differential network expression during drug and stress response. *Bioinformatics*, **21**, 2898–2905.

Çalışkan, M. *et al.* (2013). Rhinovirus wheezing illness and genetic risk of childhood- onset asthma. *New England Journal of Medicine*, **368**(15), 1398–1407.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, **1695**(5), 1-9.

Eichler, E. E. *et al.* (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, **11**(6), 446–450.

Farh, K. K.-H. *et al.* (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**(7539), 337–343.

Gallo, G. *et al.* (1989). A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, **18**(1), 30–55.

Goldberg, A. V. and Tarjan, R. E. (1988). A new approach to the maximum-flow problem. *Journal of the ACM (JACM)*, **35**(4), 921–940.

Huang, D. W. *et al.* (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, **4**(1), 44–57.

Ideker, T. *et al.* (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**(suppl 1), S233–S240.

Jia, P. *et al.* (2011). dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, **27**(1), 95–102.

Jia, P. and Zhao, Z. (2014). Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Human genetics*, **133**(2), 125–138.

Klein, R. J. *et al.* (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**(5720), 385–389.

Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput Biol*, **12**(1), e1004714.

Lee, I, *et al*. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research* **21**(7), 1109-1121.

Lee, Y. *et al*. (2013). Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases. *Journal of the American Medical Informatics Association*, **20**(4), 619–629.

Li, M.-X. *et al.* (2011). GATES: a rapid and powerful gene-based association test using extended Simes procedure. *The American Journal of Human Genetics*, **88**(3), 283–293.

Li, Z.-C. *et al.* (2015). Identification of drug–target interaction from interactome network with "guilt-by-association" principle and topology features. *Bioinformatics*, **32**(7), 1057–1064.

Liu, J. Z. *et al.* (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, **87**(1), 139–145.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, **456**(7218), 18–21.

McLean, C. Y. *et al*. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, **28**(5), 495-501.

Moffatt, M. F. *et al.* (2010). A large-scale, consortium-based genomewide association study of asthma. *New England Journal of Medicine*, **363**(13), 1211–1221.

Oliver, S. (2000). Proteomics: guilt-by-association goes global. *Nature*, **403**(6770), 601–603.

Rajagopalan, D. and Agarwal, P. (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, **21**, 788–793.

Raychaudhuri, S. *et al* (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet*, **5**(6), e1000534.

Smit, L. *et al.* (2010). 17q21 variants modify the association between early respiratory infections and asthma. *European Respiratory Journal*, **36**(1), 57–64.

Subramanian, A. *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43), 15545–15550.

Szklarczyk, D. *et al.* (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, page gku1003.

Taşan, M. *et al.* (2015). Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nature methods*, **12**(2), 154–159.

Tomizono, S. (2013). *elliplot: Ellipse Summary Plot of Quantiles*. R package version 1.1.1.

Welter, D. *et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, **42**(D1), D1001–D1006.

Wolfe, C. J. *et al.* (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC bioinformatics*, **6**(1), 227.

Xiong, W. *et al*. (2014). The centrality of cancer proteins in human protein-protein interaction network: a revisit. *International journal of computational biology and drug design*, **7**(2-3), 146–156.

Yu, G. *et al.* (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, **16**(5), 284–287.