iPINBPA: AN INTEGRATIVE NETWORK-BASED FUNCTIONAL MODULE DISCOVERY TOOL FOR GENOME-WIDE ASSOCIATION STUDIES

LILI WANG

School of Computing, Queen's University
25 Union Street, Goodwin Hall, Kingston, Ontario, K7L 3N6, Canada
Email: lili@cs.queensu.ca

PARVIN MOUSAVI

School of Computing, Queen's University
25 Union Street, Goodwin Hall, Kingston, Ontario, K7L 3N6, Canada
Email: pmousavi@cs.queensu.ca

SERGIO E. BARANZINI

Department of Neurology, University of California San Francisco
675 Nelson Rising Lane, Room 215, San Francisco, CA 94158, USA
Email: sebaran@cgl.ucsf.edu

We introduce the integrative protein-interaction-network-based pathway analysis (iPINBPA) for genome-wide association studies (GWAS), a method to identify and prioritize genetic associations by merging statistical evidence of association with physical evidence of interaction at the protein level. First, the strongest associations are used to weight all nodes in the PPI network using a *guilt-by-association* approach. Second, the gene-wise converted p-values from a GWAS are integrated with node weights using the Liptak-Stouffer method. Finally, a greedy search is performed to find enriched modules, *i.e.*, sub-networks with nodes that have low p-values and high weights. The performance of iPINBPA and other state-of-the-art methods is assessed by computing the concentrated receiver operating characteristic (CROC) curves using two independent multiple sclerosis (MS) GWAS studies and one recent ImmunoChip study. Our results showed that iPINBPA identified sub-networks with smaller sizes and higher enrichments than other methods. iPINBPA offers a novel strategy to integrate topological connectivity and association signals from GWAS, making this an attractive tool to use in other large GWAS datasets.

1. Introduction

In the last decade, Genome-wide association studies (GWAS) have been a powerful tool to identify statistically significant differences in allelic frequencies between cases and controls at each tested single nucleotide polymorphism (SNP) for hundreds of phenotypes.[1] In order to consider a signal of genome-wide significance, a Bonferroni correction is usually applied ($p\text{-value} < 5 \times 10^{-8}$ for 1 million markers) under the assumption of independence among SNPs.[2] While this method ensues a low ratio of false positives, it inevitably increases the ratio of false negatives, thus neglecting a sizable proportion of risk SNPs and limiting the overall utility of this approach. Furthermore, the results of GWAS do not directly provide any functional information of the variants. Recent advances in our understanding of biological networks, especially the large-scale human protein interaction network (PIN), have enabled its use to investigate statistically modest associations in GWAS in the context of functional modules (referred to as pathways) to elucidate the underlying molecular mechanisms of several human diseases.[3-5]

Pathway analysis approaches can be divided into three broad classes. The first class of methods attempts to compute the over-representation of a given list of genes in gene ontology (GO) or pre-computed pathway databases (*e.g.,* KEGG, Biocarta, *etc*.). Examples include DAVID[6] and INRICH.[7] The second class of methods involve functional class scoring (FCS) approaches, such as GenGen,[8] SSEA[9] and PARIS.[10] The input data to these tools is SNP-based statistics, such as *p-values*. FCS methods aggregate SNP-wise statistics into a single score for each pre-defined pathway. While potentially revealing, both the first and second classes of methods ignore the functional connections between genes and assume independence between pathways. A third class is composed of network-based analyses, and largely overcomes the assumption of pathway independence. These approaches commonly use a scaffold of protein interactions to build connections between gene products, where nodes represent proteins and edges represent physical or functional interactions between pairs of proteins. Rather than focusing on individual markers, network-based analysis methods take into account multiple loci in the context of molecular networks. Due to this critical feature, these methods can afford to use sub-genome-wide statistical significance and yet increase the power to detect new associations and functional relationships between genes in complex traits.[5]

To date, several network-based methods have been proposed to identify functional modules in the form of sub-networks of a given larger ensemble. For example, protein interaction network-based pathway analysis (PINBPA) of GWAS data was developed to identify over-represented modules in a large multiple sclerosis (MS) GWAS.[5] This approach, adapted from a similar method for gene expression analysis, uses a greedy algorithm[11] to identify sub-networks based on aggregated gene-wise statistics. Dense module searching of GWAS (dmGWAS) also extensively searches for sub-networks enriched with low p-value genes in GWAS datasets.[12] Aside from using the human PIN as a scaffold, neither PINBPA nor dmGWAS exploit topological properties of this network. Another tool, called network interface miner for multigenic interactions (NIMMI),[13] combines topological connectivity with association signals from GWAS. In this tool, sub-networks are generated for each node by adding their neighbors up to the second-order. Nodes are weighted using a modified Google PageRank algorithm, and then the pre-generated sub-networks are scored.[13] More recently, the disease association protein-protein link evaluator (DAPPLE)[14] was reported to prioritize novel associations in Crohn's disease and rheumatoid arthritis datasets. Using a fixed PIN,

DAPPLE builds direct and indirect networks among a list of genes or SNPs and computes the probabilities that those connections may have arisen by chance. However, DAPPLE does not take into account gene-wise or SNP-wise GWAS statistics.

In this paper, we introduce the integrative protein-interaction-network based pathway analysis (iPINBPA), a novel network-based pathway analysis strategy. This approach, based on the same principles of PINBPA, integrates topological connectivity among genes in PIN space and the association signals from GWAS to extensively search for sub-networks enriched in significant GWAS signals. We tested the performance of iPINBPA against PINBPA and dmGWAS using two independent well-powered datasets in multiple sclerosis (MS). Our results show that iPINBPA can identify sub-networks involving MS genes with much higher precision than the other tested methods.

2. Data and Methods

2.1. *Data*

The data used in this work include two independent GWAS data sets in MS, human PIN and benchmark MS genes for evaluation of our proposed method.

2.1.1. *GWAS data sets*

Two large-scale GWAS data sets have been used to evaluate the proposed method. The first data set is a meta-analysis (denoted as *Meta2.5*) of seven independent moderately powered GWAS and one meta-analysis and includes *137,432* SNPs mapped to *17,425* unique genes for *5,545* cases and *12,153* controls.[15] The second data set is the largest GWAS in MS to date (denoted as *WTCCC2),* and is composed of *137,457* SNPs mapped to *17,401* unique genes in *9,772* cases and *17,376* controls.[16] For both data sets, the SNP-wise statistical significance (SNP-wise p-value), was transformed into gene-wise significance (gene-wise p-value), using the versatile gene-based association study (VEGAS) tool.[17] If the gene-wise $p\text{-}value \leq 0.05$ in GWAS data set, the corresponding gene is defined as nominally significant. There are *1,982* unique nominally significant genes in *WTCCC2* and *1,690* in the *Meta2.5* data set.

2.1.2. *Human protein interaction network*

We used a high-confidence, manually curated human protein interaction network previously reported,[5] which is composed of *8,960* proteins and *27,724* interactions. The PIN is represented as an undirected graph.

2.1.3. *Benchmarking*

We assess the performance of our proposed method for two sets of benchmark genes: 1) the *45* genes emerging from GWAS as of 2011 (denoted as *WTCCC2* genes), and 2) the *135* genes from the most recent MS study, the ImmunoChip custom genotyping array[18] (denoted as *iChip* genes).

2.2. *Method*

Given a human protein interaction network, and gene-wise p-values from a GWAS data set, iPINBPA detects enriched sub-networks in three steps as shown in Fig. 1. First, for a given GWAS data set, the nominally significant genes (gene-wise *p-value* ≤ 0.05) are selected as seed genes, and then nodes in the network are weighted (*i.e.* network smoothing) via the random walk with restart algorithm according to their connectivity to seed genes based on *guilt-by-association*. Second, a network score is defined by the combination of the gene-wise p-values with node weights using the Liptak-Stouffer method.[19] The background distribution for the network score is calculated using random sampling for various network sizes. Finally, a heuristic algorithm extensively searches for modules enriched in genes with low p-values and high weights, *i.e.*, high network score. We will explain each step in detail in the following subsections.

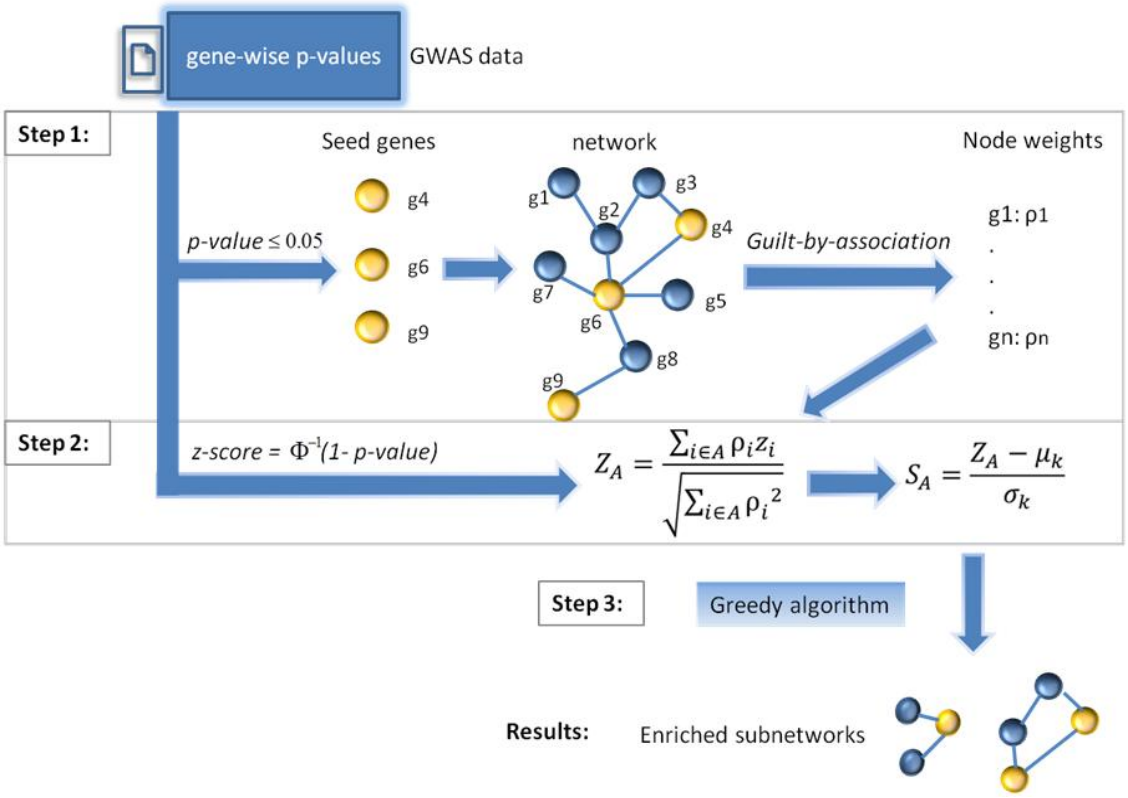

Fig. 1 Work flow of iPINBPA

2.2.1. *Random walk with restart*

Based on the assumption of *guilt-by-association*, Köhler *et al.*[20] developed a random walk with restart method to prioritize disease-associated genes. In this method, a walker starts moving from a seed node to connecting neighbors randomly. Nodes in the network are scored according to the probabilities of the walker reaching them at the end of the process. iPINBPA extends Köhler's approach by weighting the edge e_{ij} connecting n_i and n_j using corresponding gene-wise *p-values* as: $W_{ij} = ((1 - p_i) + (1 - p_j))/2$, where p_i and p_j are gene-wise p-values of n_i and n_j, and normalizing the adjacency matrix W by its columns. A score vector is calculated after each step of the walker as follows:

$$P(t) = (1 - r)W \cdot P(t - 1) + rP(0) \quad (1)$$

where $P(t)$ is the score after t steps of walking, and r is the restart ratio. The initial score vector $P(0)$ represents *a-priori* knowledge of genes (in our case, nominal significance), where *1* is assigned for seed genes and *0* for the rest. Finally, all nodes are scored according to their values in the vector $P(T)$, which quantitatively measures the topological connection to seed genes.

As mentioned above, iPINBPA requires a group of seed genes to start random walks. In this study, we used the nominally significant genes (gene-wise *p-value* $\leq$ *0.05*) in the GWAS data set as seed genes. This network smoothing step refines the searching for enriched sub-networks, as nominally significant genes will be assigned higher scores than the non-significant ones.

2.2.2. *Network scoring*

In the second step of our approach, each p-value p_i is transformed into its standard normal deviate z_i using the inverse normal CDF: $z_i = \Phi^{-1}(1 - p_i)$, and then a score for a network *A* containing *k* nodes is defined using a weighted *Z* transform test[19] (also called Liptak-Stouffer formula), as shown in Equation (2). By using this formula, the gene-wise significance from GWAS is combined with node connectivity to known disease genes. According to this algorithm, nodes with low *p-value* and close to known associated genes will score higher.

$$Z_A = \frac{\sum_{i \in A} P(T)_i z_i}{\sqrt{\sum_{i \in A} P(T)_i^2}} \quad (2)$$

To determine the significance of the network score calculated above, we performed a random sampling[11] of gene sets of size $k \in [1, 500]$ for 1000 times. For gene sets at size *k*, we computed their scores Z_A, then calculated the mean of network score μ_k and the standard deviation σ_k. The adjusted network score is defined as:

$$S_A = \frac{Z_A - \mu_k}{\sigma_k} \quad (3)$$

2.2.3. *Greedy algorithm*

The last step of iPINBPA is to find locally optimal sub-networks according to the adjusted network scores. A greedy algorithm starts searching for the optimal sub-network *G* for each node v_{start} in the network. It searches all neighbors of *G* as long as their shortest path to v_{start} is less than or equal to *2*, if adding a neighbor increases the network score S_G , then add the neighbor with the largest increase. It stops adding until there is no increasing of S_G . Then it starts searching any node inside *G* as long as this node is not v_{start} and removable, which means *G* is still a connected sub-network after removing this node. If removing a node will increase S_G , then remove the one with the largest increase. The algorithm stops searching until there is no increasing of S_G . The pseudo code of this algorithm is as follows:

(1) $G \leftarrow \{v_{start}\}$

(2) For each neighbor node *v* of *G* and depth <= *2*:
(3) Calculate score S'_G if add *v* into *G*
(4) If $\max(S'_G) > S_G$ then:
(5) Add the corresponding node v_{max} into *G*
(6) Go back to step (2)
(7) Else:
(8) For each node *v* in *G* except v_{start}:
(9) Calculate score S''_G if remove *v* from *G*
(10) If $\max(S''_G) > S_G$ then:
(11) If the corresponding node v'_{max} is removable:
(12) Remove v'_{max} from *G*
(13) Go back to step (8).
(14) Else:
(15) Return *G*

2.2.4. *Parameters*

We chose the network's characteristic path length (*4.38* for the used PIN) as the default time step (*T = 5*). The second parameter of random walk is the restart ratio *r*, which weights prior knowledge. As there is no standard criterion to select the restart ratio, we set up the default value as *0.5*. Furthermore, we tested iPINBPA with different restart ratios and the corresponding performance is discussed in section 3.4.

2.2.5. *Evaluation*

To evaluate the performance of iPINBPA, we tested two sets of reported benchmark genes. As shown in a previous study,[5] if we define association regions (blocks) composed of significant genes (gene-wise *p-value* $\leq$ *0.05*), there are *665* association blocks containing *1,982* unique genes in the *WTCCC2* data set, and *612* blocks containing *1,690* unique genes in the Meta2.5 data set. The size of associated blocks vary from *1* to *>100*, thus posing a challenge to quantitatively compare the prediction or enrichment performance for each association block. In this study, we applied iPINBPA to identify sub-networks in a high-confidence PIN and a GWAS data set, and ranked genes using their highest network score in descending order. For genes having the same network score, they were ranked by their gene-wise p-values in ascending order. Based on the ranking, CROC curves[21] were computed to assess the efficiency of iPINBPA in identifying the benchmark genes. CROC curves use an exponential function ($f(x) = (1 - e^{-\alpha x})/(1 - e^{-\alpha})$) to magnify any relevant portion of the corresponding ROC by an appropriate continuous transformation of the coordinates. CROC curves have been shown to be more effective than ROC curves to measure the ability of methods in drug discovery and gene prediction.[21] In our case, the early retrieval performance is also adequate, as we only consider the top scored/ranked nodes or sub-networks, and the size of benchmark genes is less than one percent of the total number of genes in the network.

3. Results

Based on its predecessor (PINBPA), iPINBPA introduces node-weighting by means of significant disease-related genes and integrates these weights with gene-based significance into a score, which is further

normalized by network size (see methods). We applied the iPINBPA approach to two independent large-scale GWAS datasets in multiple sclerosis (MS) (*Meta2.5* and *WTCCC2*), and benchmarked its performance against other established methods on the same input data.

In pathway analysis of GWAS, it is necessary to compute gene-wise (rather than SNP-wise) significance. Given that most associations fall outside coding regions, allocating a significant finding to a gene is not always straightforward. One common strategy is to assign the significant association to the closest gene, taking into account recombination hotspots. However, due to linkage disequilibrium (LD), it is not unusual to find that several genes map within the "area of influence" of the lead SNP. While usually the closest gene to the lead SNP is assigned, in reality, patterns of extended LD make it impossible to assign any given gene within that area with certainty.

It is challenging to compare different pathway analysis methods because of the lack of accurate knowledge of complex traits and the incomplete human PIN. Since DAPPLE and NIMMI do not accept a user-defined network and DAPPLE only accepts a short list of SNPs or genes (up to *500*), it is not possible to directly compare these methods to iPINBPA. Thus, we compared iPINBPA to PINBPA and to dmGWAS. We performed three different tests: (1) Prediction of *WTCCC2* genes using *Meta2.5* data; (2) Prediction of *iChip* genes using *WTCCC2* data; and (3) Significantly enriched networks from both GWAS data sets.

3.1. *Prediction of WTCCC2 genes using Meta2.5 data*

We first tested the ability of each method to identify the *WTCCC2* genes using *Meta2.5* data. There are *45* *WTCCC2* genes previously identified in GWAS studies (*24* of them are represented in our network). *Meta2.5* data are aggregated from seven moderately powered GWAS and one meta-analysis before the completion of *WTCCC2*. *Meta2.5* GWAS data set contains weaker association signals than *WTCCC2* GWAS data set (*26* SNPs with *p-value* $<5 \times 10^{-8}$ and *1,690* nominally significant genes in *Meta2.5*, but *57* associated SNPs and *1,982* nominally significant genes in *WTCCC2*).

We measured the fold enrichment of AUC score of each method compared to a random classifier. As shown in Fig. 2A, iPINBPA (*fold enrichment= 5.858*) performs marginally better than PINBPA (*fold enrichment= 5.386*) and significantly better than dmGWAS (*fold enrichment = 3.646*), with $\alpha = 14, f(0.05) = 0.5$.

3.2. *Predicting iChip genes using WTCCC2 data*

We also tested the ability of each method to identify the latest MS genes reported in a recent study using the ImmunoChip (*iChip*) custom genotyping array[18] using *WTCCC2*. In this test, a total of *135* genes were associated with MS (Although the total number of reported associated loci is *110*, some SNPs map to more than one gene). Of these *135* iChip MS genes, *42* genes were *WTCCC* genes (*23* of them are represented in our network), and thus *93* genes (*54* genes represented in our network) found in *iChip* are novel. As shown in Fig. 2B, iPINBPA (*fold enrichment = 6.22*) performs better than PINBPA (*fold enrichment = 5.211*) and dmGWAS (*fold enrichment = 2.818*) in the prediction of *iChip* genes, with $\alpha = 14, f(0.05) = 0.5$.

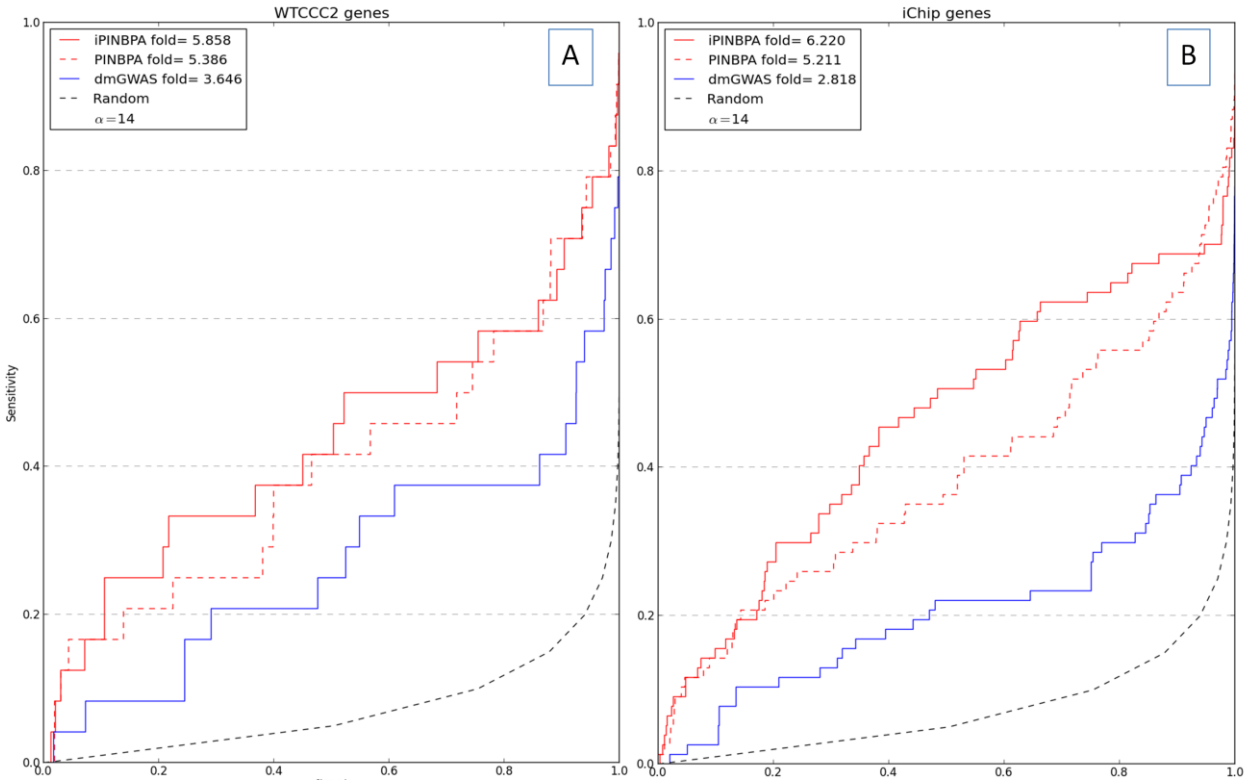

Fig. 2 CROC curves of Meta2.5 and WTCCC GWAS data sets

3.3. *Significantly enriched networks*

As the primary goal of our approach is to identify the enriched pathways for the given GWAS data set, we selected the top scored sub-networks (*score* > *3* and *size* ≥ *5*) from each method. For this analysis we also tested NIMMI, which returns sub-networks with *p-values*. For NIMMI, the sub-networks with *p-value* < *0.0013* (equivalent z-score to the other methods) were selected. As shown in Table 1, iPINBPA is more sensitive to GWAS signals and identifies smaller networks, resulting in higher precision. By overlapping the selected networks from both *WTCCC2* and *Meta2.5*, iPINBPA identified *1,299* genes (including *17 WTCCC2* genes and *44 iChip* genes), PINBPA identified *5,047* genes (including *23 WTCCC2* genes and *69 iChip* genes), dmGWAS identified *7,634* genes (including *24 WTCCC2* genes and *77 iChip* genes). NIMMI identified *4,832* genes (including *19 WTCCC2* genes and *49 iChip* genes). Altogether, iPINBPA achieved the highest precision for both sets of benchmark genes.

To evaluate the biological significance of the *1,299* candidate associated genes reported by iPINBPA, we tested their functional annotation clustering using the online tool DAVID. The KEGG pathways in the cluster with the highest enrichment score (*8.94*) are listed in Table 2. While the precise etiology of MS is still unclear, it has been consistently described as a T-cell-mediated autoimmune disease. As such, it is not surprising that related KEGG pathways such as allograft rejection, type *1* diabetes mellitus, graft-versus-host disease, and thyroid disease are significantly enriched. This result suggests that genes prioritized by iPINBPA are consistent with the biological functions likely implicated in MS pathogenesis.

Table 1. Stats of top scored sub-networks from iPINBPA, PINBPA and dmGWAS

	iPINBPA		**PINBPA**		**dmGWAS**		**NIMMI**	
GWAS data set	WTCCC2	Meta2.5	WTCCC2	Meta2.5	WTCCC2	Meta2.5	WTCCC2	Meta2.5
# networks	1496	1295	4079	4080	7109	7000	402	400
# total nodes	2163	1938	6012	6079	7665	7643	4950	4979
# overlap of nodes	1299		5047		7634		4832	
Precision (# *WTCCC2* genes)	0.013 (17)		0.005 (23)		0.003 (24)		0.004 (19)	
Precision (# *iChip* genes)	0.034 (44)		0.014 (69)		0.01 (77)		0.01 (49)	

Table 2. Functional annotation clusters of 1299 genes selected from iPINBPA in DAVID

KEGG Pathway	**Count**	**P-Value**	**Benjamini**
Allograft rejection	24	4.7E-12	3.5E-11
Type I diabetes mellitus	24	4.0E-10	2.1E-9
Graft-versus-host disease	22	3.5E-9	1.6E-8
Autoimmune thyroid disease	23	2.6E-7	9.8E-7

3.4. *Tuning parameters*

The restart ratio *r* in random walk with restart can be tuned by the user. We tested iPINBPA with different restart ratios (*0.1*, *0.3*, *0.5*, *0.7*, and *0.9*) and evaluated its performance as shown in Fig. *3*.

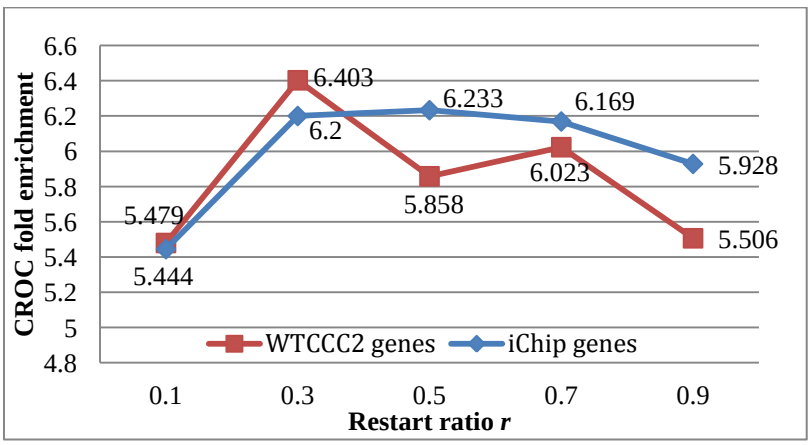

Fig. 3 CROC fold enrichments of different values of restart ratio *r*

In addition, we also tested iPINBPA with different cutoffs of selecting seed genes to start random walks, which controls the sensitivity of iPINBPA indirectly. By default, we used the nominally significant genes (gene-wise *p-value* $\leq$ *0.05*). If a more stringent cutoff is used to select fewer number of seed genes, iPINBPA usually returns smaller sub-networks. Table 3 shows the sizes and precision of top selected sub-networks from iPINBPA with different cutoffs.

Table 3. Stats of top selected sub-networks from iPINBPA with different cutoffs

	p-value ≤ 0.01		**p-value ≤ 0.005**		**p-value ≤ 0.001**	
GWAS data set	WTCCC2	Meta2.5	WTCCC2	Meta2.5	WTCCC2	Meta2.5
# networks	1732	1224	1691	1293	1547	1458
# total nodes	2108	1522	2000	1535	1774	1617
Mean of network size (node) (std)	17.25 (15.49)	12.71 (8.78)	12.6 (11.12)	10.24 (4.36)	9.95 (4.52)	7.7 (2.32)
Mean of network size (edge) (std)	26.91 (34.93)	16.22 (16.87)	16.21 (23.8)	12.38 (7.73)	10.68 (6.6)	8.37 (3.67)
# overlap of nodes	1133		1106		1082	
Precision (# *WTCCC2* genes)	0.014 (16)		0.013 (14)		0.01 (11)	
Precision (# *iChip* genes)	0.036 (41)		0.034 (38)		0.028 (30)	

4. Discussion

GWAS have been extremely successful in identifying thousands of associations in hundreds of complex traits. Due to the extensive statistical adjustments needed to avoid type *1* errors, type *2* errors are necessarily a consequence of GWAS studies, thus limiting their effectiveness. Furthermore, typically, only a few markers are replicated in any given GWAS. Effective post-GWAS analysis methods can help prioritize associations using additional sources of evidence and are becoming a useful complementary strategy to the standard analytical pipeline.

Here we introduced a novel network-based pathway analysis strategy for GWAS, which integrates topological connectivity in a PIN and the association signals from GWAS to detect significant sub-networks and also prioritize genes associated with a complex disease. The main feature of iPINBPA is the strategy we employed to identify enriched sub-networks by merging evidence from multiple sources. To our knowledge, this is the first method that integrates node weighting with a greedy search for significant sub-networks. Comparisons with different data sets and methods have demonstrated that our integrative approach dramatically improves the performance in predicting novel associations. The increase of prediction precision comes mostly from the fact that, unlike in the classical approach, potential associations with no biological relationships to statistically confirmed associations are down-weighted in this approach.

Given the multi-dimensional nature of GWAS data, it is not uncommon to see a low precision in prioritizing novel associations through network-based pathway analysis. The identified sub-networks presented here are around nodes with quite significant p-values, thus the overlap of these sub-networks lends additional support to our methods. By incorporating additional information (*e.g.,* regulatory, cell-specific expression, *etc.*), the precision of network-based pathway analysis would be improved gradually.

Unlike dmGWAS, iPINBPA and PINBPA use VEGAS to map SNPs to genes. For the analysis of GWAS data, the mapping of SNPs to genes is an open challenge. In this paper, we focus on the comparison of methodology and performance of different network-based analysis methods. We did not address potential variations emerging from using different strategies of mapping SNPs to genes; the default mapping recommended for each method was utilized.

An inherent limitation of all approaches using protein networks, is that interactions have only been described for a subset of all known proteins. Furthermore, if only high confidence interactions are taken into account as described in this study, approximately only half of all proteins are represented in the network. This necessarily places an upper boundary to the number of successful predictions any of these methods can make. With new and more accurate techniques to determine protein interactions, this limitation may be overcome in the near future. Another potential restriction of these methods is that they use global interactions, when actually tissue specific interactions might be more appropriate. Several efforts are currently underway to develop tissue-specific protein interactions that, together with knowledge about the organ/tissue compromised in a given disease, could be incorporated into network analysis of GWAS in the future. Furthermore, with the incorporation of genome-wide regulatory data (*e.g.,* ENCODE, Epigenomics Roadmap, *etc.*), it will be possible to derive cell specific networks. This will greatly enhance the performance of this approach, as it will enable the incorporation of pathophysiologically relevant and disease-specific data.

The integrative strategy we proposed in this study is generic can be readily applied to any disease or biological datasets, *e.g.,* gene expression datasets and proteomic data, as long as quantitative gene-wise or protein-wise statistical measures and putative disease genes are available.

5. Acknowledgements

This work was partially supported by Natural Sciences and Engineering Research Council of Canada, the Ontario Early Researcher Award, and grants from the National Multiple Sclerosis Society (R01NS049477). SEB is a Harry Weaver Neuroscience Fellow of the National Multiple Sclerosis Society.

References

1 T. A. Manolio, *N. Engl. J. Med.,* **363**, 166-176 (2010).

2 G. S. Barsh, G. P. Copenhaver, G. Gibson and S. M. Williams, *PLoS Genet.,* **8**, e1002812 (2012).

3 K. Wang, M. Li and H. Hakonarson, *Nat. Rev. Genet.,* **11**, 843-854 (2010).

4 S. E. Baranzini, N. W. Galwey, J. Wang, et al, *Hum. Mol. Genet.,* **18**, 2078-2090 (2009).

5 International Multiple Sclerosis Genetics Consortium, *Am. J. Hum. Genet.,* (2013).

6 W. Huang da, B. T. Sherman and R. A. Lempicki, *Nucleic Acids Res.,* **37**, 1-13 (2009).

7 P. H. Lee, C. O'Dushlaine, B. Thomas and S. M. Purcell, *Bioinformatics,* **28**, 1797-1799 (2012).

8 K. Wang, M. Li and M. Bucan, *Am. J. Hum. Genet.,* **81**, 1278-1283 (2007).

9 L. Weng, F. Macciardi, A. Subramanian, et al, *BMC Bioinformatics,* **12**, 99-2105-12-99 (2011).

10 B. L. Yaspan, W. S. Bush, E. S. Torstenson, et al, *Hum. Genet.,* **129**, 563-571 (2011).

11 T. Ideker, O. Ozier, B. Schwikowski and A. F. Siegel, *Bioinformatics,* **18 Suppl 1**, S233-40 (2002).

12 P. Jia, S. Zheng, J. Long, W. Zheng and Z. Zhao, *Bioinformatics,* **27**, 95-102 (2011).

13 N. Akula, A. Baranova, D. Seto, et al, *PLoS One*, **6**, e24220 (2011).

14 E. J. Rossin, K. Lage, S. Raychaudhuri, et al, *PLoS Genet.,* **7**, e1001273 (2011).

15 N. A. Patsopoulos, Bayer Pharma MS Genetics Working Group, Steering Committees of Studies Evaluating IFNbeta-1b and a CCR1-Antagonist, et al, *Ann. Neurol.,* **70**, 897-912 (2011).

16 International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, S. Sawcer, et al, *Nature,* **476**, 214-219 (2011).

17 J. Z. Liu, A. F. McRae, D. R. Nyholt, et al, *Am. J. Hum. Genet.,* **87**, 139-145 (2010).

18 International Multiple Sclerosis Genetics Consortium (IMSGC), A. H. Beecham, N. A. Patsopoulos, et al, *Nat. Genet.,* (2013).

19 M. C. Whitlock, *J. Evol. Biol.,* **18**, 1368-1373 (2005).

20 S. Köhler, S. Bauer, D. Horn and P. Robinson, *The American Journal of Human Genetics,* **82**, 949-958 (2008).

21 S. J. Swamidass, C. A. Azencott, K. Daily and P. Baldi, *Bioinformatics,* **26**, 1348-1356 (2010).