NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH Public Access
Author Manuscript
IEEE Int Workshop Genomic Signal Process Stat. Author manuscript; available in PMC 2012 August 14.

Published in final edited form as:
IEEE Int Workshop Genomic Signal Process Stat. 2011 ; : 131–134. doi:10.1109/GENSiPS.2011.6169462.

Network-assisted Causal Gene Detection in Genome-wide Association Studies: An Improved Module Search Algorithm

Peilin Jia[1] and **Zhongming Zhao**[1,2,3,*]

[1]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

[2]Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

[3]Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

Abstract

The recent success of genome-wide association (GWA) studies has greatly expanded our understanding of many complex diseases by delivering previously unknown loci and genes. A large number of GWAS datasets have already been made available, with more being generated. To explore the underlying moderate and weak signals, we recently developed a network-based dense module search (DMS) method for identification of disease candidate genes from GWAS datasets, leveraging on the joint effect of multiple genes. DMS is designed to dynamically search for the best nodes in a step-wise fashion and, thus, could overcome the limitation of pre-defined gene sets. Here, we propose an improved version of DMS, the topologically-adjusted DMS, to facilitate the analysis of complex diseases. Building on the previous version of DMS, we improved the randomization process by taking into account the topological character, aiming to adjust the bias potentially caused by high-degree nodes in the whole network. We demonstrated the topologically-adjusted DMS algorithm in a GWAS dataset for schizophrenia. We found the improved DMS strategy could effectively identify candidate genes while reducing the burden of high-degree nodes. In our evaluation, we found more candidate genes identified by the topologically-adjusted DMS algorithm have been reported in the previous association studies, suggesting this new algorithm has better performance than the unweighted DMS algorithm. Finally, our functional analysis of the top module genes revealed that they are enriched in immune-related pathways.

Keywords

dmGWAS; dense module search; GWAS; schizophrenia; network; gene set enrichment analysis

1 Introduction

Building on the "common-disease-common-variant (CDCV)" hypothesis, genome-wide association (GWA) studies have been popular during the past several years to identify disease associated markers and genes. So far, more than 200 diseases or traits have been studied by GWAS, and more than 4500 markers have been identified for disease association [1]. Despite the success of GWA studies, these findings only represent a small proportion of the genetic factors involved in complex diseases, while a great number of disease-associated

©2011 IEEE

[*]Corresponding author: zhongming.zhao@vanderbilt.edu.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

markers and their interactions have not been discovered yet. There are many explanations of the possible missingness or lack of power, such as rare variations and epigenetic factors that are difficult to detect in GWA studies, either because of the GWAS design (i.e., common variants) or the limitation of computation power in genome-wide SNP interactions. Among these explanations, one important factor is that most genome-wide significant markers identified so far are observed to have modest effects on individual odds ratios that are typically less than 1.2 [2]. To overcome this limitation (i.e., no strong genetic effect of the markers), gene-set based methods have recently been proposed to investigate the joint effect of multiple functionally related genes [3–5]. It has been implicated that these methods could have improved power in uncovering a set of genes significantly enriched with association signals, and such genes could exist in the same pathway or functional group [4].

We recently developed a network-based algorithm, the dense module search (DMS), to identify subnetworks significantly enriched with association signals from GWAS dataset(s) for a specific disease/trait [6]. DMS could overcome the disadvantages of pre-defined pathways, which are limited to *a priori* knowledge and only include approximately 30% of human genes, a problem frequently observed in standard gene-set based analysis. DMS allows a dynamic search of *de novo* gene sets consisting of genes interconnected in the context of a protein-protein interaction (PPI) network. Our previous applications of DMS to GWAS datasets for breast cancer and pancreatic cancer have demonstrated that DMS could successfully identify a set of genes in the form of subnetworks, which are significantly enriched and associated with the diseases. We also showed that DMS could greatly improve power when we compared the performance with other methods such as Gene Set Enrichment Analysis (GSEA), which is currently the major algorithm for gene-set based analysis of GWAS datasets [7].

Here, we proposed topologically-matched normalization strategy in the DMS method to substantially improve the DMS algorithm. The major aim is to adjust the potential bias when applying network-based methods. For example, it has been found cancer proteins have more interactors because of their functional importance [8–9]. Correspondingly, high-degree proteins would have more chance to be selected in network analysis, especially when performing step-wise searching and expanding strategies in DMS. Consequently, cancer genes have often been observed in candidate gene prioritization studies when investigators studied other diseases such as psychiatric disorders. In this study, we demonstrated that our topologically-adjusted DMS method could efficiently reduce the bias caused by the degree and, thus, substantially improve the power to detect true candidate genes in a psychiatric disorder.

II. Material and Methods

A. The CATIE GWAS Dataset

The Clinical Antipsychotic Trial of Intervention Effectiveness (CATIE) project is a multiphase randomized controlled trial. It was genotyped by Perlegen Sciences using the Affymetrix 500K and Perlegen's custom 164K chip. A detailed description of the samples can be found in reference [10]. We accessed this dataset (Distribution 7.0) from http://www.nimhgenetics.org/ through NIMH approval. Only the Causation samples were used. The pipeline of quality control, including the selection of samples and markers, was described in references [6, 11–12]. In summary, a total of 738 schizophrenia patients and 733 controls, and 446,225 SNPs were used in this study.

To facilitate the network-based strategy, SNP markers were further mapped to genes based on their genomic coordinates. Because some SNPs may execute their function through regulation or other functional effects, we included an extended region of 20kb both upstream

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and downstream of each gene. Finally, gene-wise *P* values were computed using the SNP with the smallest *P* value among all the SNPs mapped to a gene. This approach resulted in a total of 19,310 genes.

We used a comprehensive human PPI network downloaded from the Protein Interaction Network Analysis (PINA) platform [13] (March 4, 2010). Six public PPI databases have been collected in this dataset (MINT, IntAct, DIP, BioGRID, HPRD, and MIPS/MPact). After overlaying the GWAS data onto this network, we removed the nodes (genes) that were not genotyped or failed to be mapped to human protein coding genes or the edges indicating self-interaction. We had 9,227 nodes and 43,869 edges.

B. Weighted DMS

We overlay the GWAS data onto the PPI network by assigning each node a *z*-score based on the *P* value of its encoding gene, i.e., $z_i = \Phi^{-1}(1 - P_i)$ for the i^{th} node, where Φ^{-1} is the inverse normal cumulative density function and P_i is the corresponding gene-wise *P* value from the GWAS dataset.

To adjust the bias caused by nodes with many interactors, we proposed the topologically-matched randomization method in comparison to the previous regular randomization. We used the parameter of degree to describe the topological characteristics of the network, defined as the number of interactors for each node in the overall network. After examining the degree distribution of the working interactome, we categorized the nodes into four groups according to their degrees: group A with degree range between $0–2^2$, group B $2^2–2^4$, group C $2^4–2^6$, and group D $>2^6$.

The detailed module search strategy can be referred to in our previous work [6]. Briefly, the strategy has the following steps.

Step 1. Candidate module construction. Each node in the network is used as the seed to search for a local highest-scored module. During the module expansion, the neighborhood nodes with ≤ 2 steps away from the seed module are iteratively evaluated, and the node with the maximum score increase is selected. The module expansion is terminated when no neighborhood node can increase the module score by $Z_{(m+1)} \geq Z_m \times (1 + r)$, where $r = 0.1$ is used in this work and can be adjusted appropriately in different DMS applications.

Step 2. Module normalization. To determine whether the module score is higher than expected by chance, a total of 1,000 topologically-matched random networks are generated. More specifically, for each module, we first counted the number of nodes in each of the four degree groups as defined above, and then randomly selected the same number of nodes for each group from the working interactome, resulting in a topologically-matched random module. In each random process, module scores are computed for all the candidate modules generated in step 1. Finally, a normalized score is computed for each module by $Z_N = \frac{Z_m - mean(Z_m(\pi))}{sd(Z_m(\pi))}$, where $Z_m(\boldsymbol{\pi})$ is the score of the $\boldsymbol{\pi}^{th}$ randomization, and Z_m is the observed module score.

Step 3. Module selection. We select the modules that are scored within the top 1% quantile of the module score distribution and merge them to construct a disease-specific subnetwork. This subnetwork is enriched with association signals from the GWAS dataset used.

We named this improved version of the DMS algorithm “topo-adjusted DMS” for search disease-specific modules and subnetworks from GWAS dataset(s).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

III. Results

We applied the topo-adjusted DMS algorithm to the CATIE GWAS dataset for schizophrenia. We obtained a total of 8,545 modules and, based on that, selected the top 86 modules (1% of all modules) according to their score Z_N values. By combining these modules, we first constructed a subnetwork including 179 nodes (genes) in the 86 modules. Visual examination of this network showed that two genes, *PTP4A3* and *IKBKE*, showed extremely high degree and betweenness values, another topological character defined as the number of shortest path going through a node. Each of these two genes introduced many nodes that are weakly connected to the network by only one edge (PPI). Thus, these two genes, as well as those connected to the subnetwork only through these two genes, were manually removed from the subsequent analysis, as they are likely false positive nodes that escaped the topo-adjusted randomization process. As a result, a total of 143 nodes were included in the final subnetwork, and we denoted it as the schizophrenia-specific subnetwork (Figure 1).

To demonstrate the improvement of the weighted DMS method, we applied the previous unweighted DMS method to the same GWAS dataset. It generated a subnetwork with 86 modules and 173 genes. To evaluate their performance, we downloaded schizophrenia candidate genes from the SZGene database [14] (as of January 26, 2011). This database manually collected genes that have been studied for association with schizophrenia through traditional association studies as well as GWA studies. These genes (we denoted as "szgenes") can serve as gold positives. We found a total of 23 szgenes were included in the weighted DMS results (16.08%), while only 20 szgenes were included in the unweighted DMS results (11.56%) (Table I). The result indicates that the topo-adjusted DMS could greatly improve the proportion of known candidate genes for schizophrenia; thus, it has high sensitivity in searching for candidate genes.

Further examination of the module genes showed that a total of 114 genes (79.72%) had *P* values < 0.05. The remaining "non-significant" genes were recruited mainly because they were located in a low-*P*-value environment and they interacted with genes with small *P* values. For example, the genes *NCK1* and *MED28* did not show significant association with schizophrenia individually; however, each of them was observed to interact with the protein products of other important genes for schizophrenia. NCK1 interacts with four proteins encoded by szgenes, *DTNBP1*, *FYN*, *PLCG1*, and *PRX*, and MED28 interacts with two such proteins encoded by *FYN* and *GRB2*. This further demonstrated the capability of DMS to identify joint effect of multiple genes, despite that they may not be significantly associated with the disease on their own. More importantly, selection of these genes did not rely on their pre-defined pathways.

As shown in Figure 1, the resultant subnetwork includes many interesting genes related to schizophrenia. The nodes labeled in red are those that have been studied in previous association studies or recent GWA studies. The genes *AKT1* [15], *DTNBP1* [16], *DLG2* [17], *GRID1* [18], *NTRK3* [19] and *YWHAZ* [20] have been well-studied in schizophrenia, most of which have positive associations.

We then examined the pathways that are significantly enriched with the subnetwork genes by using the Ingenuity Pathway Analysis (IPA) system [21]. For the canonical pathways, the most significantly enriched ones include TGF-β Signaling, RAR Activation, 14-3-3-mediated Signaling, Neurotrophin/TRK Signaling, and PPARα/RXRα Activation (Table II), most of which are related to immune and inflammation system, supporting the immune-involved hypothesis of schizophrenia.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

IV. Discussion

In this study, we present an improved network-assisted algorithm to identify candidate genes from GWAS dataset through an optimization towards the maximum joint effects of a set of genes. Compared to standard gene-set based analysis, the dense module search method introduces flexibility through a dynamic search of the best node in each step, and, thus, it is granted the opportunity to identify *de novo* gene sets (i.e., modules) in the context of the whole human interactome. The proposed randomization process in this study takes topological characteristics into account and further improves its sensitivity to search for weak to moderate genes. Our application of the topo-adjusted DMS to the schizophrenia CATIE GWAS dataset successfully demonstrated the superiority of our topo-adjusted DMS method, as the module genes are more likely to have positive association signals from previous studies.

As shown in Table I, the weighted DMS method greatly increased the coverage rate of szgenes from 11.56% to 16.08%. At the molecular level, the weighted DMS method successfully recovered several well-studied genes for schizophrenia and other psychiatric disorders, including *AKT1*, *DTNBP1*, *GRID1*, and *NTRK3*. Interestingly, gene *GRB2* was also found in the top list. We have recently identified *GRB2* as a candidate gene for schizophrenia by using a completely different network approach [12, 22] and have successfully validated several SNPs located in *GRB2* to be positively associated with schizophrenia patients in the Irish Case Control Study of Schizophrenia (ICCSS) sample [23]. Here, using an independent GWAS dataset, we recovered *GRB2* as a strong schizophrenia candidate gene again. Furthermore, as shown in Figure 1, GRB2 functions as a "hub" node by interacting with several proteins that are encoded by strong schizophrenia candidate genes like *PLCG1* [24] and *FYN* [25]. Due to space limitations, we are unable to discuss more details of these genes, or other promising genes in this schizophrenia-specific subnetwork. However, the wealth of information included in the subnetwork provided valuable insights for interpretation of previous identified genes as well as a promising list for future validation.

In summary, we proposed an effective network-based algorithm to identify candidate genes from GWAS datasets and demonstrated it in a schizophrenia GWAS dataset. The R package of the DMS method, *dmGWAS*, can be found on our website (http://bioinfo.mc.vanderbilt.edu/dmGWAS.html). The user may utilize it for the analysis of other GWAS datasets.

Acknowledgments

We would like to thank Drs. Patrick F. Sullivan and Edwin van den Oord for valuable assistance in the CATIE GWAS data process. CATIE dataset use was granted through the National Institute of Mental Health (NIMH) Schizophrenia Genetics Initiative. This work was supported by NIH grant (1R03DE022093), VA grant (RFP CX-09-008), 2009 NARSAD Maltz Investigator Award (to ZZ) and 2010 NARSAD Young Investigator Award (to PJ).

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA. 2009; vol. 106:9362–9367. [PubMed: 19474294]
2. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011; vol. 43:513–518. [PubMed: 21614091]
3. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 2010; vol. 11:843–854. [PubMed: 21085203]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4. Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z. Pathway analysis of genome-wide association studies: methodological issues and perspectives. Genomics. 2011; vol. 98:1–8.

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

6. 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. 2011; vol. 27:95–102. [PubMed: 21045073]

7. Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am J Hum Genet. 2007; vol. 81:1278–1283. [PubMed: 17966091]

8. Xia J, Sun J, Jia P, Zhao Z. Do cancer proteins really interact strongly in the human protein-protein interaction network? Comput Biol Chem. 2011; vol. 35:121–125. [PubMed: 21666777]

9. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics. 2006; vol. 22:2291–2297. [PubMed: 16844706]

10. Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, et al. Genomewide association for schizophrenia in the CATIE study: results of stage 1. Mol Psychiatry. 2008; vol. 13:570–584. [PubMed: 18347602]

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

12. Sun J, Jia P, Fanous AH, Webb BT, van den Oord EJ, Chen X, et al. A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. Bioinformatics. 2009; vol. 25:2595–2602. [PubMed: 19602527]

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

14. Allen NC, Bagade S, McQueen MB, Ioannidis JP, Kavvoura FK, Khoury MJ, et al. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. Nat Genet. 2008; vol. 40:827–834. [PubMed: 18583979]

15. Ikeda M, Iwata N, Suzuki T, Kitajima T, Yamanouchi Y, Kinoshita Y, et al. Association of AKT1 with schizophrenia confirmed in a Japanese population. Biol Psychiatry. 2004; vol. 56:698–700. [PubMed: 15522255]

16. Guo AY, Sun J, Riley BP, Thiselton DL, Kendler KS, Zhao Z. The dystrobrevin-binding protein 1 gene: features and networks. Mol Psychiatry. 2009; vol. 14:18–29. [PubMed: 18663367]

17. MacLaren EJ, Charlesworth P, Coba MP, Grant SG. Knockdown of mental disorder susceptibility genes disrupts neuronal network physiology in vitro. Mol Cell Neurosci. 2011; vol. 47:93–99. [PubMed: 21440632]

18. Fallin MD, Lasseter VK, Avramopoulos D, Nicodemus KK, Wolyniec PS, McGrath JA, et al. Bipolar I disorder and schizophrenia: a 440-single-nucleotide polymorphism screen of 64 candidate genes among Ashkenazi Jewish case-parent trios. Am J Hum Genet. 2005; vol. 77:918–936. [PubMed: 16380905]

19. Otnaess MK, Djurovic S, Rimol LM, Kulle B, Kahler AK, Jonsson EG, et al. Evidence for a possible association of neurotrophin receptor (NTRK-3) gene polymorphisms with hippocampal function and schizophrenia. Neurobiol Dis. 2009; vol. 34:518–524. [PubMed: 19344762]

20. Jia Y, Yu X, Zhang B, Yuan Y, Xu Q, Shen Y. An association study between polymorphisms in three genes of 14-3-3 (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein) family and paranoid schizophrenia in northern Chinese population. Eur Psychiatry. 2004; vol. 19:377–379. [PubMed: 15363479]

21. Ingenuity Pathway Analysis system. Available: http://www.ingenuity.com/

22. Sun J, Jia P, Fanous AH, Oord Evd, Chen X, Riley BP, et al. Schizophrenia gene networks and pathways and their applications fornovel candidate gene selection. PLoS ONE. 2010; vol. 5:1–9.

23. Sun J, Wan C, Jia P, Fanous AH, Kendler KS, Riley BP, et al. Application of systems biology approach identifies and validates GRB2 as a risk gene for schizophrenia in the Irish Case Control Study of Schizophrenia (ICCSS) sample. Schizophr Res. 2011; vol. 125:201–208. [PubMed: 21195589]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

24. Jungerius BJ, Hoogendoorn ML, Bakker SC, Van't Slot R, Bardoel AF, Ophoff RA, et al. An association screen of myelin-related genes implicates the chromosome 22q11 PIK4CA gene in schizophrenia. Mol Psychiatry. 2008; vol. 13:1060–1068. [PubMed: 17893707]

25. Carter CJ. Schizophrenia susceptibility genes directly implicated in the life cycles of pathogens: cytomegalovirus, influenza, herpes simplex, rubella, and Toxoplasma gondii. Schizophr Bull. 2009; vol. 35:1163–1182. [PubMed: 18552348]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

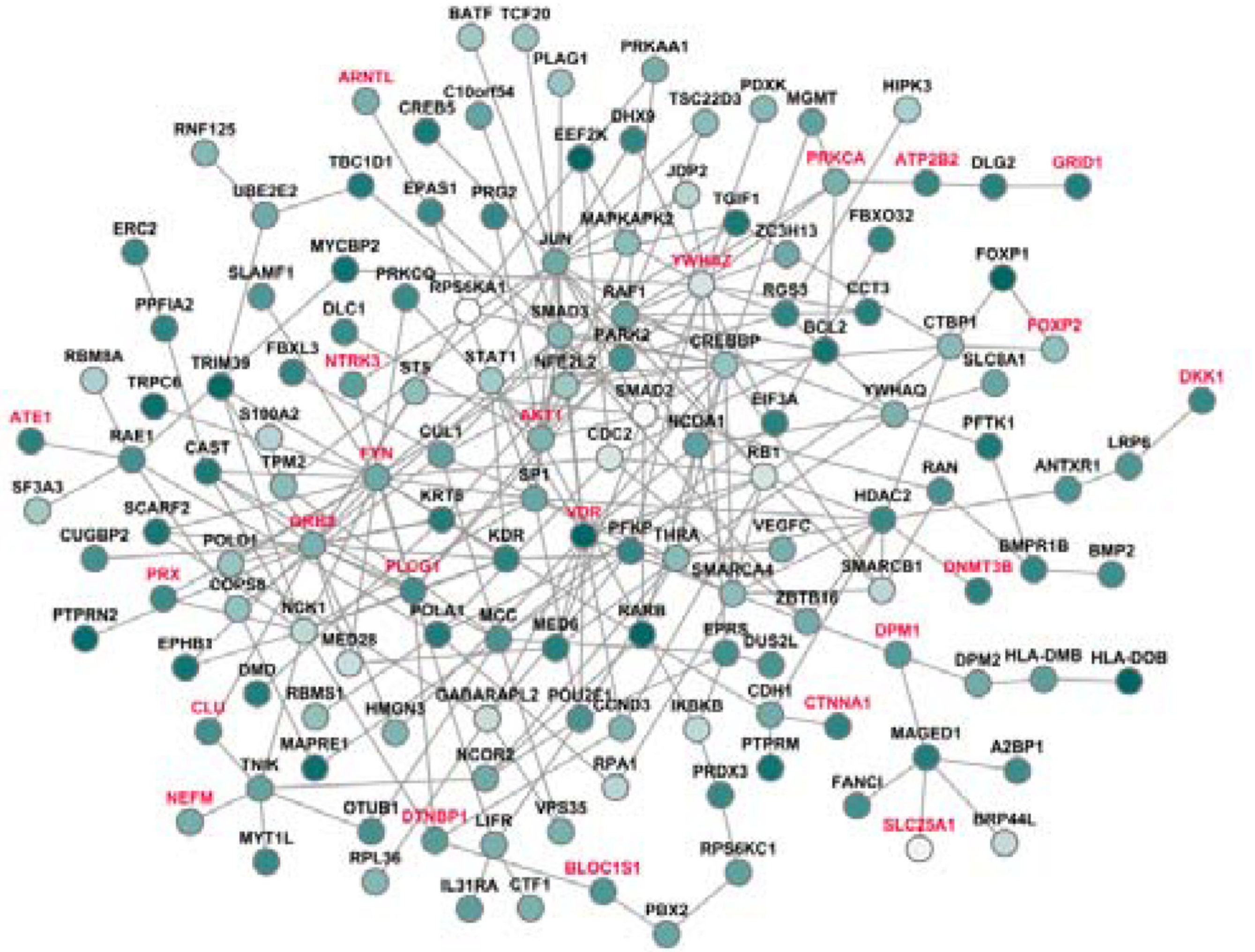

Figure 1.

Module gene based subnetwork for schizophrenia. Nodes in red are szgenes. The node darkness is proportinal to the *z*-score.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

TABLE I

Summary of DMS Results

	Unweighted DMS	**Topo-adjusted DMS**
# modules	86	86
# module genes	173	143
# szgenes in modules	20	23
Proporntion of szgenes	11.56%	16.08%

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE II

Significantly Enriched Pathways in the Subnetwork

Ingenuity Canonical Pathways	P_{BH}[a]	O[b]
TGF-β signaling	1.95×10^{-8}	11
RAR activation	1.95×10^{-8}	14
Molecular mechanisms of cancer	2.88×10^{-8}	18
Prolactin signaling	2.88×10^{-8}	10
Glucocorticoid receptor signaling	2.45×10^{-7}	15
14-3-3-mediated signaling	1.91×10^{-6}	10
Neurotrophin/TRK signaling	3.80×10^{-6}	8
PPARα/RXRα activation	5.37×10^{-6}	11
Pancreatic adenocarcinoma signaling	6.92×10^{-6}	9
Regulation of IL-2 expression in activated and anergic T lymphocytes	7.76×10^{-6}	8
Thrombopoietin signaling	8.13×10^{-6}	7

[a] BH: multiple testing correction by Benjamini & Hochberg (1995).

[b] Observed number of module genes in the category.