Bioinformatics, 31(15), 2015, 2591–2594
doi: 10.1093/bioinformatics/btv150
Advance Access Publication Date: 24 March 2015
Applications Note

OXFORD

Genetics and population analysis

EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles

Quan Wang[1], Hui Yu[1], Zhongming Zhao[1,2,3,4,*] and Peilin Jia[1,2,*]

[1]Department of Biomedical Informatics, [2]Center for Quantitative Sciences, [3]Department of Psychiatry and [4]Department of Cancer Biology, Vanderbilt University, Nashville, TN 37232, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 6, 2014; revised on February 18, 2015; accepted on March 11, 2015

Abstract

Summary: We previously developed dmGWAS to search for dense modules in a human protein–protein interaction (PPI) network; it has since become a popular tool for network-assisted analysis of genome-wide association studies (GWAS). dmGWAS weights nodes by using GWAS signals. Here, we introduce an upgraded algorithm, EW_dmGWAS, to boost GWAS signals in a node- and edge-weighted PPI network. In EW_dmGWAS, we utilize condition-specific gene expression profiles for edge weights. Specifically, differential gene co-expression is used to infer the edge weights. We applied EW_dmGWAS to two diseases and compared it with other relevant methods. The results suggest that EW_dmGWAS is more powerful in detecting disease-associated signals.
Availability and implementation: The algorithm of EW_dmGWAS is implemented in the R package *dmGWAS_3.0* and is available at http://bioinfo.mc.vanderbilt.edu/dmGWAS.
Contact: zhongming.zhao@vanderbilt.edu or peilin.jia@vanderbilt.edu
Supplementary information: Supplementary materials are available at *Bioinformatics* online.

1 Introduction

Over the past decade, genome-wide association studies (GWAS) have successfully uncovered many susceptibility loci for common diseases. However, the identified loci only explain a small portion of the genetic risk (Jia *et al.*, 2011). It is challenging to uncover the remaining risky loci as their association signals are likely to be moderate or weak. One potential solution to this challenge is to incorporate other functional information, such as protein–protein interaction (PPI) networks, to investigate joint association signals beyond single markers (Jia *et al.*, 2012).

We previously developed a network-assisted approach, dmGWAS, to address this problem (Jia *et al.*, 2011). dmGWAS applies a greedy algorithm to search for dense modules in a PPI network in which nodes are weighted by using GWAS signals. After its initial release, dmGWAS received much attention from the research community and has become quite popular for network-assisted analysis of GWAS signals.

Motivated by the strong demand, we enhanced dmGWAS by integrating and enabling gene expression profiles to assign edge weights, and denoted the new algorithm EW_dmGWAS. Differential gene co-expression (DGCE), which measures the change of gene co-expression between case and control samples, is one important feature of transcriptional information, reflecting cellular dynamics and contributing to pathogenesis (Yu *et al.*, 2013). Thus, we utilize DGCE to infer the weight of each edge and combine the association signals of its two nodes to assess the overall disease risk of network modules within the human PPI network.

2 Methods

In brief, EW_dmGWAS integrates GWAS signals and gene expression profiles to extract dense modules from a background PPI network. Node weights are derived from GWAS signals and edge weights are derived from gene expression profiles. The module score

is a combination of node weight and edge weight. The aim of EW_dmGWAS is to locally identify the modules with maximum scores.

2.1 Defining node weight

To determine disease association at the gene level, we first mapped the Single Nucleotide Polymorphism (SNP) *P*-values from GWAS onto gene-based *P*-values. Then, we defined node weight by nodeweight $(v) = \varphi^{-1}(1-p)$, where p denotes the gene-based *P*-value of node v, and φ is the standard normal distribution function.

2.2 Defining edge weight

We used the change of gene co-expression between case and control samples to infer edge weight. Specifically, let r_{case} and r_{control} represent the Pearson's correlation coefficient (PCC) of gene expression in case and control samples, respectively, and let n_{case} and n_{control} represent the sample size, respectively. We first used the Fisher transformation [Equation (1)] and then Fisher's test of difference between two conditions [Equation (2)] to define a new statistic X:

$$F(x) = \frac{1}{2} ln \frac{1+x}{1-x} \ , \tag{1}$$

$$X = \frac{F(r_{\text{case}}) - F(r_{\text{control}})}{\sqrt{\frac{1}{n_{\text{case}}-3} + \frac{1}{n_{\text{control}}-3}}}. \tag{2}$$

The newly defined statistic X approximately follows the standard normal distribution (Hou *et al.*, 2014). Accordingly, we defined edge weight as edge weight$(e) = \varphi^{-1}[1 - 2*(1 - \varphi(|X|))]$.

2.3 Defining module score

To quantitatively evaluate the density of highly weighted nodes and edges within a module, we defined the module score S by

$$S = \lambda \frac{\sum_{e \in E} \text{edgeweight}(e)}{\sqrt{\text{No. of } E}} + (1 - \lambda) \frac{\sum_{v \in V} \text{nodeweight}(v)}{\sqrt{\text{No. of } V}}, \tag{3}$$

where E and V represent the edges and nodes of the module, and λ is a parameter between 0 and 1 to balance GWAS and gene expression signals.

2.4 Module search

We implemented a greedy algorithm to search for dense modules as follows.

1. Assign a seed module M and calculate the module score S_m of M. Initially, the seed module is a single gene.
2. Examine all the first order neighbors of M, and identify the neighbor node $N_{\max}$ that generates the maximum increment of the module score.
3. Add $N_{\max}$ to the current module M if the score increment is greater than $S_m \times r$, where r is a parameter that decides the magnitude of increment.
4. Repeat steps 1–3 until no more neighbors can be added.

2.5 Normalization of module score

In order to evaluate the significance of the identified modules, we used a randomization-based method to obtain the background distribution of the module scores. Specifically, for a module M with K nodes, we randomly generated a sub-network with the same size, and calculated the score $S_m(\pi)$ of this sub-network. We repeated this process 10 000 times and denoted the mean and standard deviation of $S_m(\pi)$ as μ and σ. The module score was normalized by $S_N = (S_m - \mu)/\sigma$, and S_N was used to determine the significance of the identified modules.

3 Implementation and application

The algorithm of EW_dmGWAS is implemented in the R package *dmGWAS_3.0* and is available at http://bioinfo.mc.vanderbilt.edu/dmGWAS. It takes three types of data as input: a list of genes with association *P*-values, gene expression profiles in both case and control samples, and a human PPI network. For mapping SNP *P*-values from GWAS onto gene based *P*-values, multiple tools are available (Ballard *et al.*, 2010). In our implementation, we utilized versatile gene-based association study (VEGAS) (Liu *et al.*, 2010) for this purpose. r and λ are two parameters that need to be determined in EW_dmGWAS. r is suggested to be 0.1, as was used in our previous version (Jia *et al.*, 2011). For λ, we proposed the following approach. We randomly extracted 10 000 sub-networks from the background network and then obtained the magnitude ratios (mr) by comparing the edge weight part and the node weight part of the sub-networks:

$$mr = \left| \frac{\sum_{e \in E} \text{edgeweight}(e)}{\sqrt{\text{No. of } E}} \Big/ \frac{\sum_{v \in V} \text{nodeweight}(v)}{\sqrt{\text{No. of } V}} \right|. \tag{4}$$

λ was estimated by $1/(1 + \text{median}(mr))$ (Ma *et al.*, 2011). Alternatively, in our R package, we leave the options open for users to provide other values of r and λ according to their expertise. The output of EW_dmGWAS is a list of identified modules, ordered by the normalized module score S_N.

We demonstrated EW_dmGWAS in breast cancer (BC) and schizophrenia (SCZ), respectively. As a comparison, we applied three other methods, including the previous version of dmGWAS, the guilt-by-rewiring (GBR) method (Hou *et al.*, 2014) and MetaRanker 2.0 (Pers *et al.*, 2013), to the same datasets. GBR and MetaRanker 2.0 are similar to EW_GWAS in that they both incorporate GWAS signals and gene expression profiles to identify candidate disease genes (Supplementary Note). The BC GWAS data were obtained from the National Cancer Institute Cancer Genetics Markers of Susceptibility project (CGEMS) (Hunter et al., 2007), and gene expression data were downloaded from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). The SCZ GWAS data were obtained from the Genetic Association Information Network (GAIN) (Jia *et al.*, 2012), and gene expression data were downloaded from the public Gene Expression Omnibus (GEO) database (GSE21138). The PPI network was obtained from the Protein Interaction Network Analysis (PINA) platform (Wu *et al.*, 2009). Details of the data and analyses are provided in the Supplementary Note.

Both EW_dmGWAS and dmGWAS reported a list of dense modules as output. As suggested in our previous study (Jia *et al.*, 2011), we chose the candidate genes residing in the top 1% of modules for evaluation. For the BC dataset, EW_dmGWAS and dmGWAS reported 128 and 100 candidate genes, respectively. The output of both GBR and MetaRanker 2.0 is a list of prioritized genes. We therefore chose the top 128 genes prioritized by the two methods for comparison. We collected 517 genes (as of August, 2014) from the Cancer Gene Census category (CGC, http://cancer.sanger.ac.uk/cancergenome/projects/census/) as a benchmark to evaluate the candidate genes reported by the different methods. Among the candidate genes identified by each method, EW_dmGWAS, dmGWAS, GBR and MetaRanker 2.0 identified 14, 1, 4 and 3 CGC genes,

Table 1. Enriched KEGG pathways of candidate genes in the BC dataset

Enriched KEGG pathway	Number of genes	Adjusted *P*-value*
dmGWAS		
Metabolic pathway	11	4.40×10^{-5}
EW_dmGWAS		
Pathways in cancer	10	8.22×10^{-7}
RIG-I-like receptor signaling pathway	6	1.11×10^{-6}
Neurotrophin signaling pathway	7	1.71×10^{-6}
Tight junction	7	2.23×10^{-6}
Hepatitis C	7	2.48×10^{-6}
ErbB signaling pathway	6	3.78×10^{-6}
Endocytosis	7	3.80×10^{-5}
Adherens junction	5	4.08×10^{-5}
GnRH signaling pathway	5	2.00×10^{-4}
Leukocyte transendothelial migration	5	4.00×10^{-4}
Focal adhesion	6	5.00×10^{-4}
Jak-STAT signaling pathway	5	1.50×10^{-3}
Calcium signaling pathway	5	3.00×10^{-3}
Chemokine signaling pathway	5	4.50×10^{-3}
GBR		
Regulation of actin cytoskeleton	6	7.40×10^{-5}
Neurotrophin signaling pathway	5	7.40×10^{-5}
Axon guidance	5	7.40×10^{-5}
Pathways in cancer	6	5.00×10^{-4}
MetaRanker 2.0		
Purine metabolism	5	3.00×10^{-4}
Pathways in cancer	6	6.00×10^{-4}
Metabolic pathways	9	$6..40 \times 10^{-3}$

**P*-values were adjusted by Bonferroni correction.

respectively (Supplementary Table S2), suggesting EW_dmGWAS is more powerful in identifying disease-related genes (all *P*-values $<$ 0.05, binomial test). We further performed the pathway enrichment analysis of candidate genes (Table 1) using WebGestalt (Zhang *et al.*, 2005). Among the significant pathways (adjusted *P*-value $<$ 0.01), candidate genes reported by EW_dmGWAS were found to be enriched in the most cancer related pathways, including 'pathways in cancer', 'ErbB signaling pathway' and 'Jak-STAT signaling pathway'. Candidate genes reported by dmGWAS are only enriched in one significant pathway ('metabolic pathway'), which is not directly related to BC. 'Pathways in cancer' is also enriched in the candidate genes identified by GBR and MetaRanker 2.0. However, the candidate genes within 'pathways in cancer' are fewer compared with those identified by EW_dmGWAS.

For the SCZ dataset, EW_dmGWAS, dmGWAS, GBR and MetaRanker 2.0 reported 65, 105, 65 and 65 candidate genes, respectively. We utilized the 38 manually curated SCZ core genes as a benchmark to evaluate the results. These 38 genes have been commonly considered as candidate genes in expert reviews or have shown significant results in meta-analysis studies (Jia *et al.*, 2010). Although the candidate genes reported by dmGWAS, GBR and MetaRanker 2.0 do not overlap with the 38 SCZ core genes, EW_dmGWAS reported 2 SCZ core genes (Supplementary Table S3, all *P*-values $<$ 0.05, binomial test). Table 2 summarizes the enriched pathways in the SCZ dataset. Interestingly, two SCZ related pathways are enriched in the candidate genes identified by EW_dmGWAS, including 'Endocytosis' and 'Neuroactive ligand-receptor interaction'. Recent studies have shown that 'Neuroactive ligand-receptor interaction' plays an important role in the antipsychotic treatment response (Adkins *et al.*, 2012), while

Table 2. Enriched KEGG pathways of candidate genes in the SCZ dataset

Enriched KEGG pathway	Number of genes	Adjusted *P*-value*
dmGWAS		
Metabolic pathway	17	2.16×10^{-9}
EW_dmGWAS		
Protein processing in endoplasmic recticulum	7	1.83×10^{-8}
Endocytosis	5	2.07×10^{-5}
Neuroactive ligand-receptor interaction	5	5.83×10^{-5}
GBR[&]		
–	–	–
MetaRanker 2.0		
–	–	–

**P*-values were adjusted by Bonferroni correction. [&]No significant results.

'Endocytosis' has been implicated as the common pathophysiology underlying SCZ (Zhao *et al.*, 2014). In contrast, no interesting pathways were found to be enriched in the candidate genes reported by dmGWAS, GBR and MetaRanker 2.0.

Collectively, these results demonstrate that gene expression data are an informative complement to GWAS signals to dissect the underlying genetic architecture, and EW_dmGWAS is a powerful network tool for genetic association studies in the research community.

Acknowledgements

We thank the numerous users for their valuable feedback.

Funding

National Institutes of Health Grants (R01LM011177, R01MH095621, P50CA095103, P50CA098131 and P30CA068485), Ingram Professorship Funds (to Z.Z.), and 2010 National Alliance for Research in Schizophrenia and Affective Disorders Yong Investigator Award (to P.J.).

Conflict of Interest: none declared.

References

Adkins,D.E. *et al.* (2012) SNP-based analysis of neuroactive ligand-receptor interaction pathways implicates PGE2 as a novel mediator of antipsychotic treatment response: data from the CATIE study. *Schizophr. Res.*, **135**, 200–201.

Ballard,D.H. *et al.* (2010) Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet. Epidemiol.*, **34**, 201–212.

Hou,L. *et al.* (2014) Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. *Hum. Mol. Genet.*, **23**, 2780–2790.

Hunter,D.J. *et al.* (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, **39**, 870–874.

Jia,P. *et al.* (2010) SZGR: a comprehensive schizophrenia gene resource. *Mol. Psychiatry.*, **15**, 453–462.

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

Jia,P. *et al.* (2012) Network-assisted investigation of combined causal signals from genome-wide association studies in schizophrenia. *PLoS Comput. Biol.*, **8**, e1002587.

Liu,J.Z. *et al.* (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.*, **87**, 139–145.

Ma,H. *et al.* (2011) COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics*, **27**, 1290–1298.

Pers,T.H. *et al.* (2013) MetaRanker 2.0: a web server for prioritization of genetic variation data. *Nucleic Acids Res.*, **41**, W104–W108.

Wu,J. *et al.* (2009) Integrated network analysis platform for protein–protein interactions. *Nat. Methods*, **6**, 75–77.

Yu,H. *et al.* (2013) Dynamic protein interaction modules in human hepatocellular carcinoma progression. *BMC Syst. Biol.*, **7** (Suppl 5), S2.

Zhang,B. *et al.* (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, **33**, W741–W748.

Zhao,Z. *et al.* (2014) Transcriptome sequencing and genome-wide association analyses reveal lysosomal function and actin cytoskeleton remodeling in schizophrenia and bipolar disorder. *Mol Psychiatry*. [Epub ahead of print].