BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 16 2011, pages 2319–2320
doi:10.1093/bioinformatics/btr380

Systems biology

Advance Access publication June 23, 2011

eResponseNet: a package prioritizing candidate disease genes through cellular pathways

Jialiang Huang[1,2,3], Yi Liu[1,2], Wei Zhang[1,2,3], Hong Yu[1,2,3] and Jing-Dong J. Han[1,*]

[1]Chinese Academy of Sciences Key Laboratory of Computational Biology, Chinese Academy of Sciences-Max Planck Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China [2]Center of Molecular Systems Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China and [3]The Graduate University of Chinese Academy of Sciences, Beijing 100062, China

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Although genome-wide association studies (GWAS) have found many common genetic variants associated with human diseases, it remains a challenge to elucidate the functional links between associated variants and complex traits.

Results: We developed a package called eResponseNet by implementing and extending the existing ResponseNet algorithm for prioritizing candidate disease genes through cellular pathways. Using type II diabetes (T2D) as a study case, we demonstrate that eResponseNet outperforms currently available approaches in prioritizing candidate disease genes. More importantly, the package is instrumental in revealing cellular pathways underlying disease-associated genetic variations.

Availability: The eResponseNet package is freely downloadable at http://hanlab.genetics.ac.cn/eResponseNet.

Contact: jdhan@picb.ac.cn

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

Received on March 29, 2011; revised on June 6, 2011; accepted on June 19, 2011

1 INTRODUCTION

Genome-wide association studies (GWAS) have uncovered hundreds of common genetic variants contributing to human diseases. Various methods have been developed to translate statistical associations into biological functions, such as identification of expression quantitative trait loci (eQTL) using expression profiles (Doss *et al.*, 2005; Heinig *et al.*, 2010) or prediction of disease causative genes linked with single nucleotide polymorphisms (SNPs) using biological networks (Lage *et al.*, 2007; Wu *et al.*, 2008).

However, it is still an enormous gap to understand the biological mechanisms by which each DNA variation affects disease risk (Frazer *et al.*, 2009). Here, to address this gap, we developed a computational package by implementing and extending the ResponseNet algorithm. We showed that the extended package 'eResponseNet' can effectively prioritize candidate disease genes linked to disease-associated SNPs and uncover cellular pathways underlying disease-associated human genetic variations.

*To whom correspondence should be addressed.

2 METHODS

2.1 Implementing the ResponseNet

We implemented the ResponseNet algorithm, originally designed to identify pathways linking genetic screen hits to transcriptional changes (Yeger-Lotem *et al.*, 2009), using C/perl language and GLPK package (www.gnu.org/software/glpk). The ResponseNet algorithm is a minimum-cost flow optimization algorithm where flow goes from source nodes (genetic hits) to sink nodes (differentially expressed genes) through a weighted network, where each node or edge has been assigned a weight according to its confidence; edges are associated with a capacity that limits the flow and with a cost. The problem can be described as a linear programming formula that minimizes the overall cost of the network when distributing the maximal flow from source nodes to sink nodes (Yeger-Lotem *et al.*, 2009)

$$\text{Min}\left(\sum_{i\in V, j\in V} -\log(w_{ij})\cdot f_{ij}\right) - \left(\gamma\cdot \sum_{i\in \text{Source}} f_{Si}\right)$$

where w_{ij} and f_{ij} represents the weight and the flow of edge connecting gene i and j, respectively; $-\log(w_{ij})$ represents the cost of the edge; S is an auxiliary node pointing to the source nodes. The choice of the tuning parameter γ primarily determines the size and the quality of the output subnetwork. Higher γ values will allow more connections between the source and the sink genes but with lower confidence (Yeger-Lotem *et al.*, 2009).

2.2 Extending the ResponseNet

We designed the eResponseNet package to prioritize candidate disease genes or genes associated with a certain phenotype/trait, meanwhile, to predict the molecular interactions among these genes leading to the disease or trait. Instead of fixing the parameter γ to a default value in ResponseNet, we scan through a range of γ values (between 4 and 10 with increments of 0.05) to identify a series of optimal subnetworks from a weighted interactome. Then, to rank each candidate gene's relevance to the disease (network), the candidate genes was assigned a minimal parameter γ when the candidate gene first appears in an optimal subnetwork. The smaller the minimal parameter γ, the higher the priority is given to the gene. A summary of setting parameter γ and the main differences between eResponseNet and ResponseNet is presented in the Supplementary Material.

3 RESULTS

3.1 Prioritizing candidate disease genes

We used type II diabetes (T2D) as a study case (Supplementary Material) to test the performance of the eResponseNet package

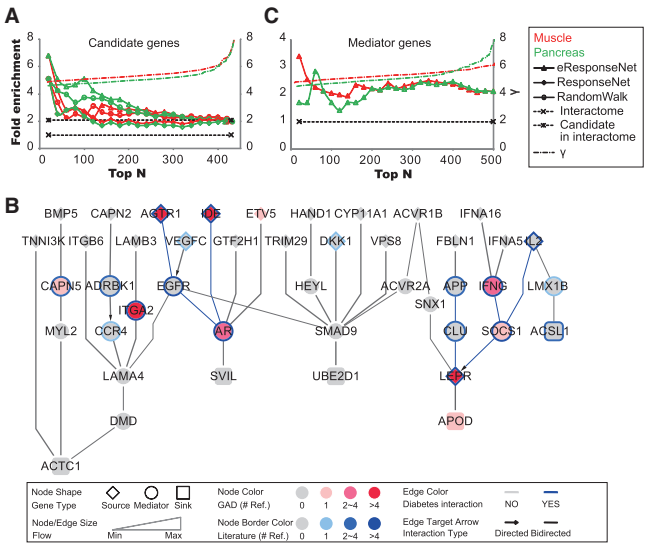

Fig. 1. Performance of eResponseNet. (**A**) Prioritizing T2D candidate genes. Fold enrichment of GAD-annotated T2D genes (left *y*-axis) is plotted for top N candidate genes (*x*-axis) ranked by various methods linking them to differentially expressed genes detected in muscle or pancreas, compared with GWAS candidate genes in the interactome. The minimal parameter γ values of genes incorporated in a subnetwork are indicated by the right *y*-axis. (**B**) The partial network identified by eResponseNet with GWAS candidate genes as source nodes and differentially expressed genes in diabetic versus normal muscle samples as sink nodes at $\gamma = 5$. (**C**) Fold enrichment of GAD-annotated T2D genes in top N mediator genes identified by eResponseNet compared with the genes in the full network.

in prioritizing candidate disease genes. All T2D candidate genes surrounding GWAS loci were considered as sources, while significantly differentially expressed genes detected by microarrays as sinks. A series of subnetworks were calculated from a weighted human interactome network integrated from three sources KEGG, HPRD and STRING, upon varying the parameter γ. Each of the candidate genes was assigned a minimal parameter γ (Section 2).

As our expectation, the genes assigned with lower γ are preferentially more enriched in genetic association database (GAD) (Becker *et al.*, 2004)-annotated T2D genes (Supplementary Material) in the subnetworks (Supplementary Fig. S1). The performance of prioritizing candidate disease genes by the minimal parameter γ in eResponseNet outperforms the gene ranking criteria by the amount of flow used in the original ResponseNet algorithm (Fig. 1A and Supplementary Table S1). Furthermore, in our dataset, its performance is even better than the random walk (Kohler *et al.*, 2008) (Fig. 1A), a best-performing algorithm in prioritizing candidate disease genes. By using the genes co-cited with diabetes ('literature diabetes genes', Supplementary Material) as another golden standard positive (GSP) set less biased for GWAS loci, we also confirmed that the performance of eResponseNet was not depending on the particular datasets or GSPs (Fig. 1A and Supplementary Fig. S2A).

3.2 Uncovering cellular pathways underlying T2D-associated genetic variations

We further examined whether the subnetworks identified by eResponseNet can explain the molecular mechanisms of how these DNA variations affect disease risk. To this end, we found that the mediator genes connecting sources and sinks included many well-known T2D causative genes, such as AR, IFNG and SOCS1 (Fig. 1B and Supplementary Fig. S2B), suggesting that some of the genes within T2D-associated loci (e.g. IDE and LEPR) affect T2D susceptibility by acting through or together with the known T2D causative genes. Indeed, our predicted pathways are consistent with reported mouse knock-out phenotypes and molecular interactions in the existing literature (Supplementary Material). In addition, the pathways also linked these genes to other genes previously unknown to affect T2D susceptibility, which are potentially new mediators to the disease. Overall, compared with those in full network, the mediator genes identified by eResponseNet were on average 2-fold more enriched for both the GAD and the literature diabetes genes, with the latter GSP which favors the mediator nodes more than the source nodes (Fig. 1C and Supplementary Fig. S2C).

4 CONCLUSION

In this article, we present eResponseNet, a package demonstrated effective for prioritizing candidate disease genes and identifying their underlying molecular pathways to human disease. This method apparently can be generalized to analyzing any complex disease and can be used to link any two layers of 'omics' data, including not only the SNPs to gene expression (as described here), but also epigenomic or microRNA data to gene expression, gene expression to proteomic or metabolic data and so on.

The eResponseNet output files can be directly visualized in Cytoscape (Supplementary Material).

Funding: China Natural National Science Foundation (Grant #30890033 and 91019019), Chinese Ministry of Science and Technology (Grant #2011CB504206), Chinese Academy of Sciences (Grant #KSCX2-EW-R-02 and KSCX2-EW-J-15), and Stem cell leading project (XDA01010303) to J.-D.J.H.

Conflict of Interest: none declared.

REFERENCES

Becker,K.G. *et al.* (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
Doss,S. *et al.* (2005) Cis-acting expression quantitative trait loci in mice. *Genome Res.*, **15**, 681–691.
Frazer,K.A. *et al.* (2009) Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, **10**, 241–251.
Heinig,M. *et al.* (2010) A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature*, **467**, 460–464.
Kohler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, **25**, 309–316.
Wu,X. *et al.* (2008) Network-based global inference of human disease genes. *Mol Syst Biol*, **4**, 189.
Yeger-Lotem,E. *et al.* (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet*, **41**, 316–323.