ORIGINAL ARTICLE

Identification of key genes for type 1 diabetes mellitus by network-based guilt by association

Shan-Shan Li[1]
Jia-Mei Tian[2]
Tong-Huan Wei[3]
Hao-Ren Wang[4]

1. Department of Endocrinology, Linyi People's Hospital, Linyi 276000, China
2. Department of Pediatric Internal Medicine, Linyi People's Hospital, Linyi 276000, China
3. Department of Internal Medicine, The People's Hospital of Linyi Hi-Tech Industrial Development Zone, Linyi 276000, China
4. Department of Internal Medicine, Linyi Luozhuang Central Hospital, Linyi 276000, China

http://dx.doi.org/10.1590/1806-9282.66.6.778

SUMMARY

OBJECTIVE: *This study aimed to propose a co-expression-network (CEN) based gene functional inference by extending the "Guilt by Association" (GBA) principle to predict candidate gene functions for type 1 diabetes mellitus (T1DM).*

METHODS: *Firstly, transcriptome data of T1DM were retrieved from the genomics data repository for differentially expressed gene (DEGs) analysis, and a weighted differential CEN was generated. The area under the receiver operating characteristics curve (AUC) was chosen to determine the performance metric for each Gene Ontology (GO) term. Differential expression analysis identified 325 DEGs in T1DM, and co-expression analysis generated a differential CEN of edge weight > 0.8.*

RESULTS: *A total of 282 GO annotations with DEGs > 20 remained for functional inference. By calculating the multifunctionality score of genes, gene function inference was performed to identify the optimal gene functions for T1DM based on the optimal ranking gene list. Considering an AUC > 0.7, six optimal gene functions for T1DM were identified, such as regulation of immune system process and receptor activity.*

CONCLUSIONS: *CEN-based gene functional inference by extending the GBA principle predicted 6 optimal gene functions for T1DM. The results may be potential paths for therapeutic or preventive treatments of T1DM.*

KEYWORDS: *Diabetes mellitus, type 1. Protein binding. Genetic association studies. Genetics.*

INTRODUCTION

Type 1 diabetes mellitus (T1DM) is a disorder of glucose homeostasis characterized by progressive insulin deficiency, which results in hyperglycemia that develops as a result of autoimmune destruction of the pancreatic β-cell[1]. The morbidity of T1DM has increased worldwide in the last decades, especially in childhood and developed countries[2].

Increasingly, high-throughput genome-wide association studies have resulted in a paradigm shift in the way that researchers view complex diseases. As a powerful approach for molecular research, high-throughput genome-wide analysis has been applied for unprecedented discovery in various diseases. Transcriptome analysis has yielded huge

DATE OF SUBMISSION: 30-Dec-2019
DATE OF ACCEPTANCE: 19-Jan-2020
CORRESPONDING AUTHOR: Hao-Ren Wang
No.422 Luosi Road, Luozhuang District, Linyi 276000, Shandong, China. Tel/Fax: 86-0539-7088535
E-mail: lss201910@163.com

amounts of genes influencing the likelihood of developing T1DM[3-5]. Understanding the function of uncharacterized genes is one of the major challenges of biology[6,7]. While most biological functions arise from integrated activities between many genes, making gene function prediction complex[8]. Gene interaction usually involves participation in the same or related cellular functions. Thus, gene interactions can be used to infer gene functional relationships. The function of a protein can be inferred by observing whether it interacts with another protein of known function, which is an example of the "guilt by association" concept[9,10]. Gene networks have been widely used to predict gene function using the neighbor voting algorithm, a basic application of the "guilt by association" principle[9,11].

In this study, we performed a network-based gene function inference to explore informative genes and gene functions involved in the development of T1DM by expanding the "guilt-by-association" method.

METHODS

Our gene function prediction procedure contained two main core steps: network characterization across thousands of gene ontology (GO) annotation sets using a fully vectorized neighbor voting algorithm, and gene multifunctionality assessment to determine the optimal gene functions.

Differential co-expression network from transcriptome data

Here, the transcriptome data of T1DM retrieved from the public functional genomics data repository Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/), under the accessing number of GSE55098[12], were utilized to determine gene differential expression and gene co-expression. The microarray data were obtained from peripheral blood mononuclear cells of 12 patients with newly diagnosed T1DM and 10 normal controls and presented on the GPL570 (Affymetrix U133 Plus 2.0) platform. Detailed sample characteristics and microarray experiments have been shown in a previous study[12]. Gene expression levels were normalized using the robust multiarray average (RMA) procedure and normalized using the median method.

Gene differential expressions between T1DM and normal controls were measured by Linear Models for Microarray Data (Limma) package in R. By assimilating a set of gene-specific t-tests and a Benjamini-Hochberg false-discovery-rate (FDR) based method[13] to adjust the p-values, differentially expressed genes were determined under the threshold values of $p < 0.05$ and $|\log2FoldChange| \geq 2$. Then, a network was generated based on these differentially expressed genes, and the Spearman's correlation coefficient (SCC), a measure of the correlation between two genes, was used to re-weight the gene network. In the differential co-expression network, SCC gives a value of edge connection between -1 and +1 inclusive, and the SCC absolute value was considered as the weight value of the edge; a weight value close to 0 represents a weaker connection between two genes, and a weight value close to 1, a stronger connection between two genes.

In this representation of the differential co-expression network, each row and column indicated a node, and the connection between the two was indicated by the corresponding entry in the adjacency matrix. Moreover, Cytoscape (http://cytoscape.org/) was employed to visualize the differential co-expression network.

Furthermore, topological centrality is effective for identifying essential molecules in well-characterized interaction networks. Here we analyzed the topological centrality of the differential co-expression network. Degree quantifies the local topology of each node, by summing up the number of its adjacent nodes. Genes with high node degrees tend to be associated with many functions. Thus, degree centrality of the differential co-expression network was investigated.

GO annotation

After representing the differential co-expression network as a matrix, we performed a gene attribute analysis to determine the gene set label vectors and assess gene multifunctionality. The GO consortium (http://geneontology.org/) contains 19,003 human GO terms, covering 18,402 genes. To improve the prediction performance, only GO terms with differentially expressed genes > 20 remained in the subsequent analysis.

Network-based gene function prediction

After obtaining both the network and the annotations, we assessed the network through its topology and the annotations through gene multifunctionality calculation.

Neighbor voting for gene function prediction

Neighbor voting algorithm based on the "guilt-by-association" principle was employed to perform the gene function prediction. In the "guilt-by-association" principle, genes with shared functions are preferentially connected. In a network, genes with similar neighbors may share common properties, giving rise to a prediction metric based on the similarity of neighbors. In this study, we applied the neighbor voting algorithm based on the differential co-expression network and GO annotations. Specifically, we hid a subset of gene labels in one GO term and assessed whether the remaining genes in this GO term could predict the identities of the hidden genes using information inferred from the differential co-expression network.

Multifunctionality assessment

Gene multifunctionality refers to genes possessing multiple molecular functions, each of which can be characterized by the set of genes inferred to be interacting in a particular biological context[14]. Moreover, node degree is unambiguously linked to multifunctionality, and multifunctional genes often are expected to exhibit a higher node degree. As with node degree, multifunctionality is also a key factor in explaining the results of gene function prediction. Given a gene i, its multifunctionality was defined as:

$$MF_i = \sum_{k|i \in GO_k} \frac{1}{n_k * n'_k}$$

Where nk is the number of genes within the GO term k, and n'k is the number of genes outside the GO term k. The more highly annotated or multifunctional a gene is, the higher the chances of predicting them as good candidates for having any annotation. A comparison of gene multifunctionality with the neighbor voting performance AUC presented an indication of the degree to which generic predictions dominate results. Thus, as a control of the annotations in the neighbor voting algorithm, we performed a gene multifunctionality assessment using GO annotations and generated a ranked score for each gene.

RESULTS

Objects

We firstly obtained the accessible expression data of T1DM to identify the differentially expressed genes and generate the gene co-expression. We constructed the co-expression adjacency matrix of 325 differentially expressed genes (covering 52,650 interactions), where the entry indicated the connection between two genes (Figure 1A).

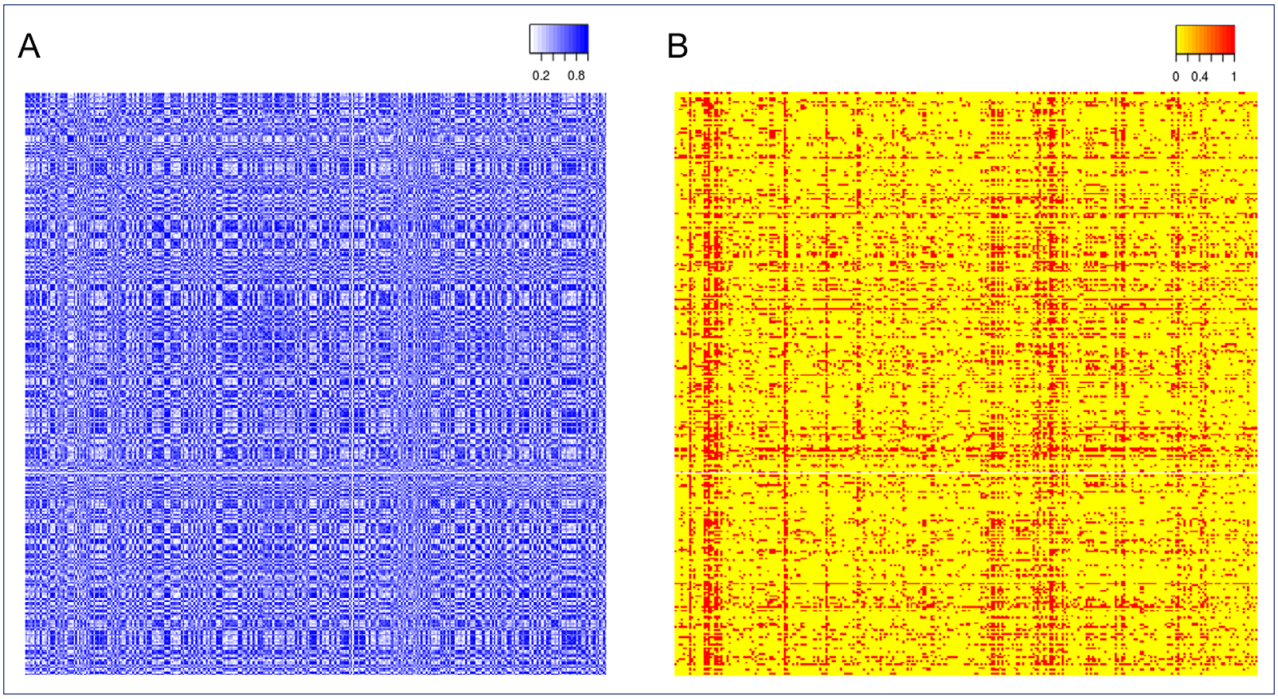


FIGURE1. Two necessary objects for the network-based function inference by extending the "guilt by association" method. A: Differential co-expression network matrix; B: Gene set annotation vectors.

FIGURE 2. A: The distribution of node degree of the co-expression network. B: Differential co-expression network with edge weight > 0.8 and node degree > 1.

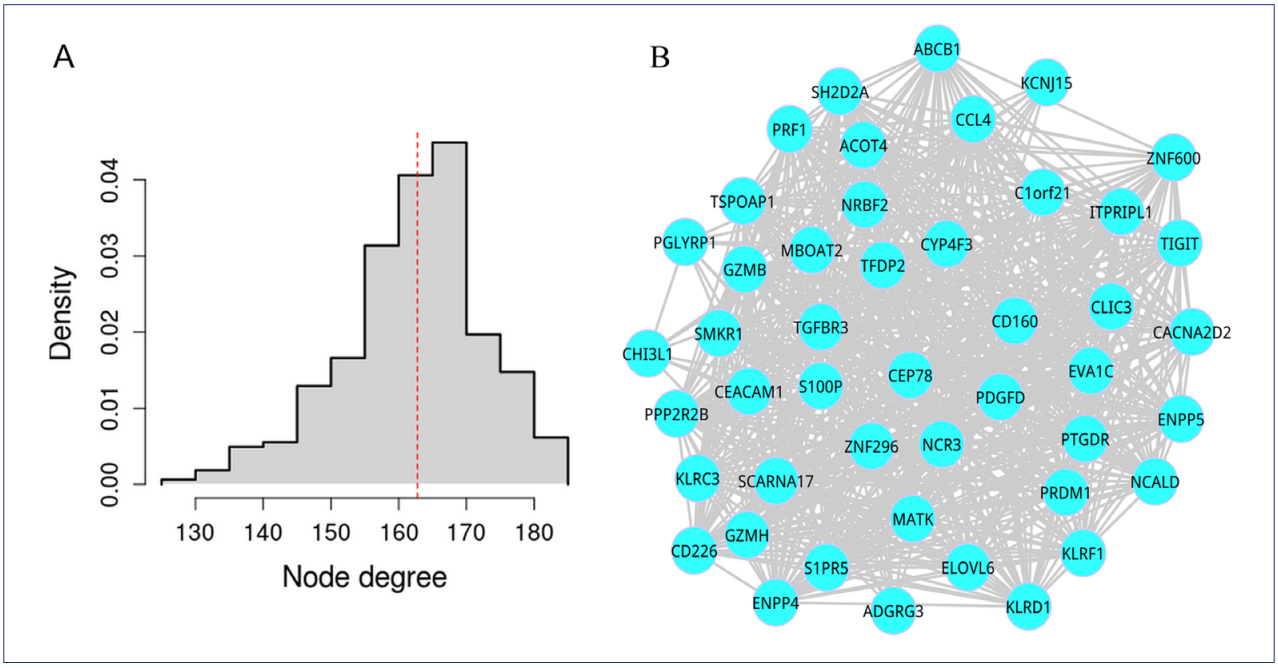
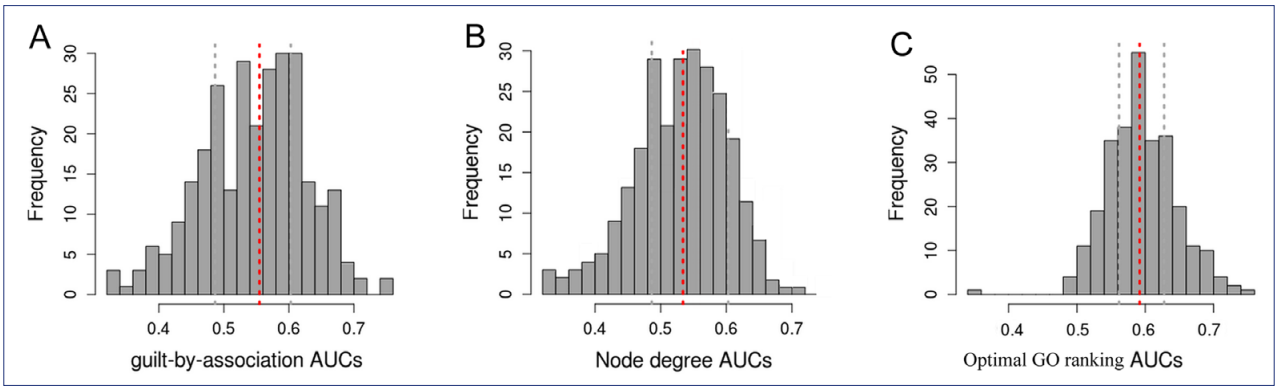


FIGURE 3. The distributions of the area under the receiver operating characteristics curve (AUC) scores. A: Distribution of AUC scores from the neighbor voting algorithm. B: Distribution of AUC scores for node degree ranking. C: Distribution of AUC scores from multifunctionality assessment. red: median, grey: inter-quartile ranges.



Subsequently, these GO terms were represented as a binary vector, where each entry corresponded to a differentially expressed gene, with a 1 indicating that the differentially expressed gene was a member of this GO term, and a 0 if it was not (Figure1B).

Differential co-expression network

After generating a co-expression matrix of 325 differentially expressed genes, gene interactions with weight values < 0.8 were removed, and the remaining gene interactions were used to construct the differential co-expression network, including 45 differentially expressed genes and 675 interactions (Figure 2A). A degree centrality analysis was performed for the co-expression network and illustrated the distribution of node degree in Figure2B.

Gene function inference

Generally, genes with similar neighbors may share common properties. Thus, we performed gene function inference for T1DM based on the differential co-expression network. A node degree analysis is an important assessment of the network. Genes with high node degrees tend to be involved in many GO annotations set. Thus, we analyzed the gene node degree and calculated the node degree AUC (Figures 3A and 3B). A total of 176 terms were identified with AUC > 0.5. Moreover, one GO term defense response to another organism (AUC = 0.718) showed good performance with AUC > 0.7.

Based on the optimal ranking gene list, we generated the distribution of AUCs for 282 GO terms (Figure 3C). Go terms with AUC > 0.7 were defined

as the optimal gene functions for T1DM, including regulation of immune system process (AUC = 0.741), positive regulation of immune system process (AUC = 0.738), system process (AUC = 0.725), signal transducer activity (AUC = 0.717), transmembrane signaling receptor activity (AUC = 0.713), and receptor activity (AUC = 0.713).

DISCUSSION

In this study, we proposed a co-expression network-based gene functional inference by extending the "Guilt by Association" strategy to predict candidate gene functions for T1DM from the GO consortium, which is important for revealing the molecular mechanisms and future applications of therapeutic decisions.

By extending the "Guilt by Association" strategy on the differential co-expression network, we generated several optimal gene functions for T1DM by assessing gene multifunctionalities, such as regulation of the immune system process and receptor activity. Regulation of the immune system process is defined as any process that modulates the frequency, rate, or extent of an immune system process. It is well known that T1DM is a chronic autoimmune disorder with the destruction of pancreatic β cells in genetically predisposed individuals with impaired immune regulation. Previous studies have revealed altered immune regulation in patients with T1DM[15,16]. Dys-regulation of the immune system contributes to the breakdown of immune regulation, leading to T1DM[17]. Eizirik et al.[18] indicated that innate immunity and inflammatory mediators play important roles in the process of T1DM. Moreover, several genetic variants in T1DM have been proven to have functional features of impaired immune regulation[19]. Regulation of the immune system process might enable investigators to restore immune imbalances with therapeutic interventions.

CONCLUSIONS

Using a co-expression network-based gene functional inference based on the "Guilt by Association" principle, our study predicted 6 optimal gene functions related to the regulation of immune system process and receptor activity, which might lead to potential paths for therapeutic or preventive treatments of T1DM and its complications.

Conflict of interest

None declared

Author Contributions

Conceptualization, Hao-Ren Wang; formal analysis, Shan-Shan Li; writing and original draft preparation, Hao-Ren Wang; writing and review and editing, Jia-Mei Tian; supervision, Tong-Huan Wei; funding acquisition, Hao-Ren Wang.

RESUMO

OBJETIVO: *O objetivo deste estudo é realizar uma inferência funcional genética baseada na rede de coexpressão (CEN), expandindo o escopo do princípio de "Culpa por Associação" (GBA - Guilt by Association) para prever as funções genéticas do diabetes mellitus tipo 1 (T1DM).*

MÉTODOS: *Primeiro, os dados transcritos do T1DM foram recuperados do repositório de dados genômicos para a análise dos genes diferenciais (DEGs), e foi gerada uma CEN diferencial ponderada. A área sob a curva ROC (AUC) foi escolhida para determinar a métrica de desempenho para cada termo de Ontologia Genética (GO). A análise da expressão diferencial identificou 325 DEGs no T1DM, e a análise de coexpressão gerou uma CEN diferencial com aresta de peso >0,8.*

RESULTADOS: *Um total de 282 anotações de GO com DEGs >20 foram mantidas para inferência funcional. Ao calcular a pontuação de multifuncionalidade dos genes, a inferência da função genética foi realizada para identificar as funções genéticas ideais para T1DM com base na lista de classificação genética ideal. Considerando um valor de AUC >0,7, foram identificadas seis funções genéticas ideais para a T1DM, tais como a regulação do processo imunológico e da atividade dos receptores.*

CONCLUSÕES: *A inferência funcional genética baseada em CEN, ao expandir o princípio de GBA, previu seis funções genéticas ideais para o T1DM. Os resultados podem ser caminhos potenciais para tratamentos terapêuticos ou preventivos do T1DM.*

PALAVRAS-CHAVE: *Diabetes mellitus tipo 1. Ligação proteica. Estudos de associação genética. Genética.*

REFERENCES

1. Atkinson MA, Eisenbarth GS. Type 1 diabetes: new perspectives on disease pathogenesis and treatment. Lancet. 2001;358(9277):221-9.
2. Gale EA. The rise of childhood type 1 diabetes in the 20th century. Diabetes. 2002;51(12):3353-61.
3. Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, Glessner JT, et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature. 2007;448(7153):591-4.
4. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009;324(5925):387-9.
5. Eizirik DL, Sammeth M, Bouckenooghe T, Bottu G, Sisino G, Igoillo-Esteve M, et al. The human pancreatic islet transcriptome: expression of candidate genes for type 1 diabetes and the impact of pro-inflammatory cytokines. PLoS Genet. 2012;8(3):e1002552.
6. Janitz M. Assigning functions to genes: the main challenge of the post-genomics era. Rev Physiol Biochem Pharmacol. 2007;159:115-29.
7. Morris RJ. Thy-1, a pathfinder protein for the post-genomic era. Front Cell Dev Biol. 2018;6:173.
8. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402(6761 Suppl):C47-52.
9. Oliver S. Guilt-by-association goes global. Nature. 2000;403(6770):601-3.
10. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol. 2000;18(12):1257-61.
11. Peña-Castillo L, Tasan M, Myers CL, Lee H, Joshi T, Zhang C, et al. A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biol. 2008;9(Suppl 1): S2.
12. Yang M, Ye L, Wang B, Gao J, Liu R, Hong J, et al. Decreased miR-146 expression in peripheral blood mononuclear cells is correlated with ongoing islet autoimmunity in type 1 diabetes patients 1miR-146. J Diabetes. 2015;7(2):158-65.
13. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res. 2001;125(1-2):279-84.
14. Gillis J, Pavlidis P. The impact of multifunctional genes on "guilt by association" analysis. PLoS One. 2011;6(2):e17258.
15. Zóka A, Műzes G, Somogyi A, Varga T, Szémán B, Al-Aissa Z, et al. Altered immune regulation in type 1 diabetes. Clin Dev Immunol. 2013;2013:254874.
16. Mejía-León ME, Barca AM. Diet, microbiota and immune system in type 1 diabetes development and evolution. Nutrients. 2015;7(11):9171-84.
17. Dwyer CJ, Ward NC, Pugliese A, Malek TR. Promoting immune regulation in type 1 diabetes using low-dose interleukin-2. Curr Diab Rep. 2016;16(6):46.
18. Eizirik DL, Colli ML, Ortis F. The role of inflammation in insulitis and beta-cell loss in type 1 diabetes. Nat Rev Endocrinol. 2009;5(4):219-26.
19. Roep BO, Tree TI. Immune modulation in humans: implications for type 1 diabetes mellitus. Nat Rev Endocrinol. 2014;10(4):229-42.