European Journal of Obstetrics & Gynecology and Reproductive Biology 186 (2015) 91–96

Contents lists available at ScienceDirect

European Journal of Obstetrics & Gynecology and Reproductive Biology

journal homepage: www.elsevier.com/locate/ejogrb

Predicting of disease genes for gestational diabetes mellitus based on network and functional consistency

Qingying Zhang [1,2], Mulan He [1,2], Jue Wang [1,2], Shuangping Liu [1,2], Haidong Cheng [1,2,*], Yan Cheng [1,2,*]

[1] Department of Obstetrics, Obstetrics and Gynecology Hospital of Fudan University, Shanghai 200090, China
[2] Shanghai Key Laboratory of Female Reproductive Endocrine Related Diseases, Shanghai 200090, China

ARTICLE INFO

Article history:
Received 25 September 2014
Received in revised form 11 December 2014
Accepted 18 December 2014

Keywords:
Gestational diabetes mellitus
Disease genes
Interaction network
Functional enrichment
Differentially expressed genes

ABSTRACT

Objective: Gestational diabetes mellitus (GDM) is a world-widely prevalent disease with adverse outcomes. This study aims to identify its disease genes through bioinformatics analysis.
Study design: The raw gene expression profiling (ID: GSE19649) was downloaded from Gene Expression Omnibus database, including 3 GDM and 2 healthy control specimens. Then limma package in R was utilized to identify differentially expressed genes (DEGs, criteria: p value <0.05 and $|\log_2 FC| > 1$). Simultaneously, known disease genes of GDM were downloaded from Online Mendelian Inheritance in Man database. Then, DEGs and known disease genes were uploaded to STRING to investigate their protein–protein interactions (PPIs). Gene pairs with confidence score >0.8 were utilized to construct PPI network. Furthermore, pathway and functional enrichment analyses were performed through KOBAS (criterion: p value <0.05) and DAVID (The Database for Annotation, Visualization and Integrated Discovery) software (criterion: false discovery rate <0.05), respectively.
Results: A total of 404 DEGs were identified, including 273 up-regulated and 131 down-regulated DEGs. Moreover, 68 known disease genes of GDM were obtained. Then, 190 gene pairs were identified to significantly interact with each other. After deleting PPIs between DEGs, PPI network was constructed, consisting of 115 gene pairs. Furthermore, genes in PPI network were significantly enriched in 10 functions and 8 pathways.
Conclusion: Based on PPI network and functional consistency, 6 candidate genes of GDM were considered to be candidate disease genes of GDM, including *CYP1A1*, *LEPR*, *ESR1*, *GYS2*, *AGRP*, and *CACNA1G*. However, further studies are required to validate these results.

Introduction

Gestational diabetes mellitus (GDM) is formally defined as glucose intolerance with onset or first recognition during pregnancy, affecting 2–5% of all pregnant women [1]. GDM patients have a higher incidence of pre-eclampsia during pregnancy and type 2 diabetes mellitus within 5–16 years after delivery than healthy pregnant women [2]. Babies of GDM patients are typically at increased risk of macrosomia, stillbirth, neonatal hypoglycemia, neonatal cardiac dysfunction, childhood obesity, and type 2 diabetes [2,3]. Due to its world-wide prevalence and adverse outcomes, GDM is in need of early detection and management which require full understanding of the pathogenic mechanism of GDM [4].

Genetic, social, and environmental factors, like family heredity, age, lifestyle, and body mass index significantly influence the predisposition to GDM, indicating that GDM is a complex human disease [5]. Furthermore, genetic studies suggest that GDM is a multigenic disease, in which variants in multiple genes interact with various environmental factors to cause this disease [3]. Thus, it is necessary to identify disease genes of GDM and investigate their molecular etiology. It is confirmed that GDM is caused by enhanced insulin resistance and pancreatic β cell dysfunction [6], through the variant genes coding for the insulin secretion, insulin and insulin signaling, lipid and glucose metabolism, and maturity-onset diabetes of the young [3,7]. Furthermore, several microarray studies have been conducted to investigate the dysregulated genes in GDM patients, and genes like *COL1A1* (collagen, type I, alpha 1), *FABP4* (fatty acid binding protein 4, adipocyte), *GHR* (growth hormone

* Corresponding authors at: Department of Obstetrics, Obstetrics and Gynecology Hospital of Fudan University, Lane 128 Shenyang Road, Yangpu District, Shanghai 200090, China. Tel.: +86 021 33189900; fax: +86 021 33189900.
E-mail address: hdcheng_2003@163.com (Y. Cheng).

http://dx.doi.org/10.1016/j.ejogrb.2014.12.016
0301-2115/

receptor), *S100A8* (S100 calcium binding protein A8), *CALM1* (calmodulin 1 (phosphorylase kinase, delta)), *MIF* (macrophage migration inhibitory factor), *FLT1* (fms-related tyrosine kinase 1), *TNF* (tumor necrosis factor), *LEP* (leptin), *IFNG* (interferon, gamma), *IL1B* (interleukin 1, beta), *HLA-G* (major histocompatibility complex, class I, G), *CD48* (CD48 molecule), *VAV3* (vav 3 guanine nucleotide exchange factor), *PTPN6* (protein tyrosine phosphatase, non-receptor type 6), and *IL15* (interleukin 15) have been identified to be involved in GDM [8–10]. However, the disease genes of GDM have not yet been identified completely.

In this study, the gene expression data of GDM patients and healthy controls were analyzed statistically to screen differentially expressed genes (DEGs). Then, the PPI (protein–protein interaction) network of DEGs and known disease genes downloaded from OMIM (Online Mendelian Inheritance in Man) database was constructed. After pathway and functional enrichment analyses, genes with interactions and similar functions were identified, supplying a platform for the early diagnosis of GDM and the development of individualized treatment.

Data and methods

Microarray data

The raw gene expression profiling (ID: GSE19649) [10] and probe annotation files were both downloaded from National Center of Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) [11]. A total of 5 chips were available, including 3 of pooled GDM specimens and 2 of pooled healthy control specimens. All of the patients and healthy controls were Chinese pregnant women matched for age, BMI (body mass index), gestational age, and gestational diabetes screen. The platform used here is GPL7350Aalborg University Illumina human-6 v2.0 expression beadchip.

Known disease genes

Known disease genes of GDM (defined as set A) were downloaded from OMIM database [12]. OMIM is one of the most authoritative databases of human genes and genetic phenotypes, including the records of text, picture, and Ref. [12].

Data preprocessing and DEGs screening

The downloaded raw gene expression data were preprocessed. Firstly, the probe-level symbols were converted into gene-level symbols. The expression values of probes corresponding to a specific gene were averaged to obtain the gene expression value. Secondly, gene expression values were $\log_2$ transformed [13] and standardized by using the preprocessCore package in R [14]. Then, the limma package in R [15] was applied to DEGs (defined as set B) screening, and p value <0.05 as well as $|\log_2 FC| > 1$ were set as the screening criteria.

Hierarchical clustering analysis of DEGs

To determine the specificity of DEGs between the specimens of GDM patients and healthy controls, the pheatmap package in R was utilized to conduct bidirectional hierarchical clustering analysis (BHCA) [16,17]. Based on the Euclidean distances of expression values, the genes with similar expression patterns were clustered by BHCA.

Construction of PPI network

The union gene set of set A and B was uploaded to STRING (Search Tool for the Retrieval of Interacting Genes) [18], which calculates the confidence scores for protein–protein interactions between gene pairs. Only the gene pairs with confidence score >0.8 were utilized to construct PPI network. After deleting interactions between DEGs, the remaining network was visualized by using Cytoscape software [19].

Functional enrichment analysis

To investigate the bio-functions of genes in PPI network, GO (Gene Ontology) functional enrichment analysis was performed based on the online software DAVID (The Database for Annotation, Visualization and Integrated Discovery) [20]. Based on hypergeometric distribution, DAVID takes the genes with similar or related functions as a whole set [21]. In this analysis, false discovery rate (FDR) <0.05 was set as the criterion.

Pathway enrichment analysis

To investigate the pathways involving genes in PPI network, KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis was performed by using KOBAS (KEGG Orthology Based Annotation System) server based on cumulative hypergeometric distribution [22]. The criterion for this analysis was set as p value <0.05.

Results

DEGs between specimens of GDM patients and healthy controls

The microarray data of GDM and control specimens after normalization is shown in Fig. 1. After DEGs screening, a total of 404 DEGs (p value <0.05 and $|\log_2 FC| > 0.5$) were identified, including 273 up-regulated and 131 down-regulated DEGs (Supplementary file 1). Moreover, 68 known disease genes of GDM were obtained from OMIM database (Supplementary file 2).

Hierarchical clustering analysis of DEGs

According to gene expression patterns, DEGs were clustered through BHCA, after which, GDM and control specimens could be distinguished clearly (Fig. 2). It is indicated that the identified DEGs were significantly specific for GDM.

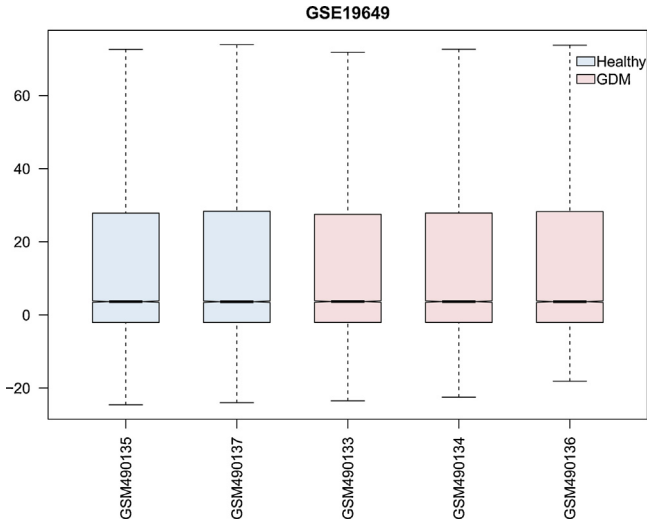

Fig. 1. Microarray data after normalization. The lines in box are coincident, indicating that these chips have been highly normalized.