Research article

Identification of diagnostic biomarkers of gestational diabetes mellitus based on transcriptome gene expression and alternations of microRNAs

XIA Xuemei* and HU Xuemei

People’s Hospital of Lishui City, Zhejiang Province, China

* Correspondence: Email: 18957093513@163.com.

Abstract: Background: Gestational diabetes mellitus (GDM), characterized by glucose intolerance during pregnancy, poses substantial health risks for both mothers and infants due to the interplay of insulin resistance and β-cell dysfunction. Molecular biomarkers, including SNPs, microRNAs (miRNAs), and proteins, have been linked to GDM development during pregnancy. Notably, miRNA-mediated regulation of gene expression holds pivotal roles in metabolic disorders. This study aims to identify diagnostic biomarkers for GDM and establish a diagnostic model. Methods: Firstly, gene expression data from GDM samples (N = 9) and normal samples (N = 9) were sourced from the Gene Expression Omnibus (GEO) database. Subsequently, the limma package was employed to discern differentially expressed genes (DEGs), with subsequent functional and enrichment analyses executed using the clusterProfiler package. A comprehensive exploration of genes significantly correlated with GDM was undertaken via weighted gene co-expression network analysis (WGCNA). The construction of a protein-protein interaction (PPI) network was facilitated by STRING, while visualization of hub genes was achieved through Cytoscape. Moreover, the miRNA-mRNA network was established using StarBase. Concurrently, immune infiltration significantly correlated with hub genes was identified. Results: In this study, 209 DEGs between normal and GDM samples were identified, and these genes were associated with collagen containing extracellular matrix heparin binding and axon guidance, etc. Then, 18 modules were identified by WGCNA and the brown module including 212 genes had a significantly negative correlation with GDM (r = −0.66, P = 0.003). Additionally, five low gene expressions (CXCL12, MEF2C, MMP2, SOX17 and THBS2) and two high gene expressions (BMP4 and SFRP5) were identified as GDM related hub genes. Moreover, hub genes regulated by alternations of miRNAs were established and three hub genes (CXCL12, MEF2C and THBS2) were negatively correlated with activated Natural Killer (NK) cells while two hub genes (BMP4 and SFRP5) were positively correlated with activated NK cells. Conclusions: This study offers novel hub genes that...
could contribute to the diagnostic approach for GDM, potentially shedding light on the intricate mechanisms underpinning GDM's developmental pathways.

**Keywords:** GDM; miRNA; diagnostic markers; hub genes; immune infiltration

---

1. **Introduction**

Gestational diabetes mellitus (GDM) is a significant problem that occurs during pregnancy and is characterized by the presence of glucose intolerance and persistent insulin resistance [1]. Women grappling with a diagnosis of Gestational Diabetes Mellitus (GDM) confront a heightened vulnerability to the onset of Type 2 Diabetes Mellitus (T2DM) and cardiovascular ailments. These implications hold the potential to profoundly impact the future well-being of the offspring [2]. Diet and lifestyle changes, metformin, glyburide or insulin are generally recommended to treat GDM but have a limited efficacy [3,4]. At present, quantitative fetal fibronectin (fFN) and cervical length (CL) screening are common methods for clinical prediction of GDM. However, there are certain shortcomings in methods: the individual differences, the lack of specific and accurate quantitative standards for predicting GDM, the lag, the occurrence of adverse events such as infections, and low sensitivity and specificity of fFN. Due to the lack of diagnostic markers and treatment approaches for GDM, there is an increasing incidence of GDM, and GDM is related to the high risk of future cardiovascular events [5]. Therefore, to screen the early diagnostic biomarkers of GDM is important to prevent the development of GDM.

Presently, a multitude of studies have put forth the notion that genetic and miRNA alterations assume pivotal roles in the genesis of GDM [6,7]. For example, the increase of IL-38 in the placentas or serum may contribute to the development of GDM [8]. Furthermore, the diminished presence of glucagon-like peptide-1 (GLP-1) undermines the functionality of β-cells, thereby establishing a link to the pathogenesis of GDM [9]. miRNAs, small single-stranded non-coding RNAs (consisting of 18 to 25 nucleotides), are pivotal players in the post-transcriptional modulation of gene expression. They show promise as potential diagnostic biomarkers for GDM [10]. For example, microRNAs (miRNAs) play a role in both trophoblast proliferation and differentiation, as well as in the control of insulin secretion and glucose transport in pregnant women [11]. The potential biomarker for gestational diabetes mellitus (GDM) in the first trimester could be the increased levels of miR-223 and miR-23a. This finding suggests the possibility of a new early non-invasive diagnostic tool for GDM [12]. Therefore, to screen the alternations of transcriptome gene and miRNAs related to GDM is important.

The analysis of high throughput sequencing data provides disease diagnostic markers, which has emerged as promising diagnostic and therapeutic tools for GDM [13,14]. To illustrate, a total of 465 genes exhibiting differential expression (DEGs) have been meticulously identified. This revelation not only imparts novel facets to the diagnostic landscape of GDM but also holds the potential to enrich the realm of personalized GDM treatments [15]. Among these, six distinct genes have been singled out, bearing the promise of serving as valuable biomarkers for the purpose of GDM diagnosis [16]. In our study, we identified DEGs related to GDM and analyzed the biological roles of these DEGs. Then, we obtained 7 hub genes and established mRNA-miRNA network related to GDM. Finally, we identified immune infiltration, which was significantly correlated with hub genes.
2. **Methods**

2.1. **Acquisition the data of GDM**

Two gene expression datasets, namely GSE49524 (platform: GPL7020 NuGO array human NuGO_Hs1a520180) and GSE51546 (platform: GPL10558 Illumina HumanHT-12 V4.0 expression bead chip), were obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo). A total of eighteen samples of Umbilical Cord Blood (UCB) were collected, consisting of 9 samples from women diagnosed with gestational diabetes mellitus (GDM) and 9 samples from women with normal pregnancies.

2.2. **Analysis of differentially expressed genes**

As per the platform annotation data, the probes inherent to both platforms were translated into corresponding gene symbols. In instances where a single gene was associated with multiple probes, the average expression value of these probes was employed to represent the gene’s expression level. To mitigate the influence of batch discrepancies between the two platforms, Surrogate Variable Analysis (SVA 3.28.0) was undertaken. Subsequently, the gene expression profile post-batch effect removal was established. In adherence to the criteria (|logFC| > 0.5 and p < 0.05), the differentially expressed genes (DEGs) existing between the GDM and normal samples were meticulously extracted using the limma R package. These DEGs were then visualized through both a heat map and a volcano map, offering insightful representations of their distinct patterns.

2.3. **Enrichment analysis**

The GO and KEGG enrichment of DEGs were performed by the cluster Profiler R package. For GO enrichment analysis, the functional enrichment of BP, CC and MF is plotted according to adj.p < 0.05. For KEGG enrichment analysis, the pathway enrichment is drawn according to p < 0.05.

2.4. **Weighted gene co-expression network analysis**

The expression matrix is arranged in descending order according to the variance, and the first 5000 genes with high variance are selected. Key modules and related genes among 5000 genes were selected by the R package “WGCNA”. In short, the adjacency matrix is converted into a topological overlap matrix (TOM) to divide the genes into different modules. The soft threshold was set to be 5 (R2 = 0.93 and the minimum number of modules = 30) and the key module related to GDM was obtained.

2.5. **PPI network construction**

According to the criteria of gene importance (GS) > 0.4 and module importance (MM) > 0.6, 212 genes were selected from the key modules. Intersecting the key module genes and 209 DEGs, a total of 74 genes were obtained. The protein-protein interactions (PPI) network was constructed by STRING (http://www.string-db.org). In short, the relationships of genes were identified by STRING and
visualized by the Cystoscope plug-in. According to the MCODE plug-in, the key module gene of the network is identified, which is the hub gene.

2.6. hub genes analysis

To find out hub gene, its expression level in GDM and control was analyzed, then its ROC was estimated. Finally, according to the miRNA-mRNA in the StarBase database (http://starbase.sysu.edu.cn/index.php), the number of prediction programs was set as 5, and the miRNA corresponding to the hub gene was analyzed then the miRNA-mRNA network was drawn.

2.7. CIBERSORT algorithm

The immunological markers of each sample were assessed using the CIBERSORT algorithm (R script v1.03), which allowed for the estimation of the abundance of different cell types within the mixed cell population. The findings demonstrate the spatial arrangement of immune cells and the variations in immune infiltration, as well as the statistical significance assessed by the Wilcoxon rank sum test.

2.8. Statistical analysis

Utilizing the “survival” data package as the foundation, we formulated a Cox regression model. To perform a non-parametric statistical hypothesis test for the comparison of two groups, we employed the Wilcoxon rank sum test. Conversely, the Kruskal-Wallis test was used for scenarios involving two or more categories. Significance was established at P < 0.05.

3. Results

3.1. Data preprocessing

Initially, the gene expression profiles sourced from the GSE49524 and GSE51546 datasets, encompassing the two distinct platforms (GPL7020 and GPL10558), were amalgamated. This integration process focused on retaining solely the samples corresponding to GDM and normal conditions. As a result, a cumulative total of 18 samples emerged, equally split between 9 GDM and 9 normal samples. The shared genes, amounting to 15220, across both platforms were singled out, culminating in the construction of a unified expression profile that encapsulated the data from both platforms. To address any potential batch effect stemming from the divergent platform data, a meticulous analysis was conducted using an R package. Following the identification of this batch effect, measures were taken to eliminate its influence, ultimately yielding a refined gene expression profile that was poised for further comprehensive analysis (as depicted in Figure 1).
Figure 1. Removal of batch effects. (A and B) The correction of batch effects arising from the two platforms was undertaken through the utilization of the SVA R package, employing the normal Q-Q plot method. (C and D) Utilizing the SVA R package, the correction of batch effects between the two platforms was accomplished through the application of the inverse gamma Q-Q plot method.

3.2. Screening of differentially expressed genes

A comprehensive tally of 209 Differentially Expressed Genes (DEGs), comprising 135 upregulated genes and 73 downregulated genes, were successfully pinpointed through the utilization of the limma R package. The resulting ensemble of DEGs showcased distinct patterns of expression disparity between the GDM samples and their normal counterparts. The heat map of the DEGs is shown in Figure 2A and a volcano plot is shown in Figure 2B.
Figure 2. The differentially expressed genes (DEGs) were graphically portrayed through heat maps and volcano plots. In the heat maps, genes marked in red signify up-regulation, while those in blue denote down-regulation. Conversely, in the volcano plots, genes indicated in red depict up-regulation, and those in green denote down-regulation. (A and B) Displayed are both the heat map and the volcano plot detailing the DEGs observed between GDM samples and their normal counterparts.

3.3. Functional enrichment analysis of DEGs

GO Functional Enrichment Analysis and KEGG Pathway Enrichment Analysis were systematically executed on the roster of DEGs through the utilization of the cluster profile methodology. The ensuing graphical representations, featured in Figure 3, spotlight the top 10 outcomes for both GO functional enrichment analysis and KEGG pathway enrichment analysis. These outcomes, indicative of the enrichment analysis, unveiled a conspicuous trend. Specifically, a majority of the DEGs exhibited enrichment across diverse GO terms. The findings from the enrichment analysis indicated that a majority of the DEGs exhibited enrichment in several GO terms, including those related to the extracellular matrix including collagen and binding to heparin. The GO functional enrichment analysis and KEGG pathway enrichment analysis of the DEGs were carried out by cluster profile. The top 10 GO functional enrichment analysis diagram and KEGG pathway enrichment diagrams of DEGs are displayed in Figure 3. The results of the enrichment analysis suggested that most of the DEGs were enriched in several GO terms, such as collagen containing extracellular matrix and heparin binding. Furthermore, KEGG analysis indicated ribosome and oxidative phosphorylation in Axon guidance, MAPK signaling pathway, and Fluid shear stress as key enriched pathways.
3.4. WGCNA analysis

According to $R^2 > 0.9$, the average connectivity $\approx 0$, the soft threshold is determined to be 5 (Figure 4A). When the soft threshold is 5, $R^2$ is 0.93 and slope is $-1.57$ (Figure 4B). Cluster genes are based on soft thresholds (Figure 4C), and draw a heat map of module feature relationships (Figure 4D). Each row represents a module feature gene, and the columns represent sample features. We visualize the module and histogram of gene meaning (Figure 4E), and the brown module is selected as the key module. The gene importance (GS) and module importance (MM) in the brown module are showed by a scatter plot (Figure 4F). The brown module comprises 364 genes in total.

**Figure 3.** An examination of the modular gene’s functional enrichment. (A–D) The 10 most significant GO functions and KEGG pathways of DEGs.
Figure 4. Weighted correlation network analysis. (A) Sample clustering to detect outliers. (B) Sample dendrogram and trait heatmap. (C) Analysis of network topology for various soft-thresholding powers. (D) Clustering dendrogram of genes. (E) Heatmap of the module-trait relationships. (F) Gene significance across modules. (G) brown module.
3.5. **PPI analysis**

Based on GS > 0.4 and MM > 0.6 for the genes of the brown module, 212 genes were screened. Taking the intersection of the result of WGCNA and DEGs, 74 intersection genes were obtained (Figure 5A). Construct a PPI model by the STRING database (Figure 5B) and analyze the PPI network by Cytoscape’s MCODE plug-in (Figure 5C), and obtain 7 hub genes, namely *BMP4, CXCL12, MEF2C, MMP2, SFRP5, SOX17* and *THBS2*.

**Figure 5.** Protein-protein interaction (PPI) network. (A) Intersection of modular genes and DEGs. (B) Protein-protein interaction network of the intersection genes. (C) Hub gene screen.
3.6. *Hub gene analysis and mRNA-miRNA network construction*

We generate box diagrams for the seven hub genes based on their expression levels in GDM and the control group. Five low-expression genes (CXCL12, MEF2C, MMP2, SOX17, and THBS2) and two high-expression genes (BMP4 and SFRP5) were identified as hub DEGs related to GDM (Figure 6). Further examination of the ROC curves of the seven hub genes revealed that their AUC areas were all greater than 0.7 (Figure 7). According to the StarBase database, we map the mRNA-miRNA network (Figure 8).

**Figure 6.** The expression of the diagnostic markers.
Figure 7. The ROC curve of the diagnostic markers.

Figure 8. mRNA-miRNA network construction.
3.7. **hub genes associations with immune infiltrates**

A deeper exploration was undertaken to scrutinize the intricate interplay between hub genes and immune infiltrations within the GDM microenvironment. In contrast to the immune infiltrations observed in the normal samples, a discernible augmentation in activated NK cells was apparent. Of particular significance were the associations observed between specific hub genes and activated NK cells. In this context, three hub genes, namely CXCL12, MEF2C, and THBS2, exhibited an inverse correlation with activated NK cells. Conversely, two other hub genes, BMP4 and SFRP5, displayed a positive correlation with the activation status of NK cells. The elucidation of these intricate relationships, as encapsulated in Figure 9, provides a deeper understanding of the dynamic interactions occurring within the GDM microenvironment, shedding light on potential regulatory mechanisms between hub genes and the immune response, specifically the involvement of activated NK cells.

![Diagram](image)

**Figure 9.** Evaluation and visualization of immune cell infiltration. (A) CIBERSORT analysis of immune cell infiltration between GDM samples and control samples. (B) Heat map illustrating the correlations among 19 distinct immune cell types. The dimensions of the colored squares reflect the intensity of the correlation, with shades of blue indicating positive correlations and shades of red denoting negative correlations. The depth of color saturation, particularly in darker shades, corresponds to the heightened strength of the observed correlations. (C) Violin plots depicting the distribution of 19 distinct immune cell types. The presence of red markers within the plots signifies variations in infiltration levels between the two compared groups. (D) The “Gene” module provides a visual representation of the relationship between gene expression and immune infiltration.
4. Discussion

Alterations in gene and miRNA expression hold pivotal roles in the landscape of pregnancy-related diseases [17,18]. These changes in gene and miRNA expression profiles have the potential to serve as crucial biomarkers, facilitating the differentiation between GDM and normal samples. Their utility extends to encompass diverse clinical applications, encompassing accurate clinical diagnosis, treatment evaluation, and prognostic assessment [19,20].

GEO is an international public database that contains high throughput microarray and next-generation sequence (NGS) functional genomic data sets submitted by the research community. To identify the gene expression and miRNA expression alterations of GDM, data of mRNA microarrays (GSE49524 and GSE51546) from the GEO database were systematically analyzed between GDM and controls.

Within our study, a total of 209 DEGs were discerned between GDM samples and their normal counterparts. Among these, 135 exhibited upregulation, while 73 displayed downregulation. Employing DAVID analysis, a comprehensive portrayal emerged, depicting these 209 genes predominantly engaging in enriched biological processes uniquely relevant to the context of GDM. KEGG analysis revealed that DEGs are enriched in the MAPK signaling pathway. The MAPK signaling pathway is involved in the protective effect of metformin on renal dysfunction in GDM [21]. In addition, five down-regulated genes, including CXCL12, MEF2C, MMP2, SOX17 and THBS2 and two up-regulated genes including BMP4 and SFRP5 were identified as GDM related hub but their functions in GDM were poorly understood. In the context of GDM, a notable increase in the serum levels of CXCL12, signifying heightened angiogenesis, was observed in comparison to normal pregnancies. This elevation was accompanied by a significant association between GDM and the augmented presence of CXCL12 [22,23]. Furthermore, the perturbed expression of the PGC-1α domain MEF2C was noted in the contrast between 3890 complication-free pregnancies and 441 pregnancies characterized by complications, including GDM. This aberration in expression was implicated in trophoblast invasion and differentiation, thereby exerting influence on placental development [24]. The study unveiled noteworthy distinctions in THBS-2 levels between fetal and maternal compartments (p = 0.013 and 0.0014) within Hyperglycemia in Pregnancy (HIP) and non-HIP contexts, positioning THBS2 as a discernible risk factor for the onset of HIP [25]. The activity of MMP-2 regulates stromal remodeling by the increase of TGFβ-1 activation and oxidative stress, which is associated with the mother subjected to GDM [26]. The BMP4/NOX-1/COX-2 signaling pathway is involved in GDM-related hypertension and overexpression of BMP4 could lead to hypertension by impairing endothelial function in pregnancy [27]. GDM is associated with impaired maternal immune responses. Consistent with our research, Thiago [28] found that NK cells in the peripheral blood of GDM is frequently evaluated. CD16+CD56+NK cells is low in GDM, while the CD16−CD56+ cells is high in GDM [29].

Several limitations are inherent to our study. Notably, due to data availability constraints, we were unable to comprehensively analyze potential associations between hub genes and essential clinical parameters, as well as prognosis.

In addition, the effects of abnormal expression of hub genes were not validated in clinical experiments. Moreover, the miRNA targeting hub genes were confirmed. Overall, further study is required to validate these genes.
Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Availability of data and material

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgment

We sincerely appreciate the investigators and authors who have contributed to this field. This research was supported by the Zhejiang Provincial Medical and Health Technology Project (Grant No. 2022498503).

Conflict of interest

The authors declare there are no competing interests.

Author contributions

Xuemei Xia, analysis and collection of the data, drafting the manuscript; Xuemei Hu, conception and design, analysis and explanation of the data, revising the manuscript. All the authors have read and approved the final manuscript.

References