Identification of Immune-Associated Genes in Diagnosing Polycystic Ovary Syndrome with Metabolic Syndrome by Weighted Gene Co-expression Network Analysis

Yi Xu
Zhongda Hospital, Southeast University

Yihui Gu
The First People's Hospital of Lianyungang

Yang Shen
Zhongda Hospital, Southeast University

Wen Feng (fengw125@126.com)
The First People's Hospital of Lianyungang

Research Article

Keywords: PCOS, MS, WGCNA, immune infiltration, diagnosis

Posted Date: June 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3044016/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

**Background:** Dysregulated immune system and metabolic dysfunction take an essential impart on the pathogenesis of polycystic ovary syndrome (PCOS) and metabolic syndrome (MS). The purpose of this study is to identify the key immune-associated genes of PCOS patients with MS.

**Methods:** Three PCOS and one MS dataset from the Gene Expression Omnibus (GEO) database were downloaded. After the data sets were combined with the de-batch effect, analysis was conducted using methods such as differentially expressed genes (DEGs), weighted gene co-expression networks (WGCNA), functional enrichment, protein-protein interaction (PPI) network construction, and LASSO regression, which were used to identify candidate immune-related essential genes for diagnosing the combination of these two diseases. Evaluate the obtained genes through ROC and ultimately explore abnormalities in PCOS through immune cell infiltration analysis.

**Results:** The merged PCOS dataset contains 1513 differentially expressed genes (DEGs), with 691 identified by MS. DEGs of MS were primarily enriched in immune regulation and metabolic dysfunction. The crossed gene functions of DEGs in PCOS and module genes in MS are also mainly enriched in the immune system and metabolic dysfunction. We screened 35 node genes from 57 cross genes through the PPI network and selected six candidate genes using the lasso regression method. Five genes analyzed by the ROC curve had high diagnostic values (RAB4B, SFP91, ARHGAP33, LIME1, and SHMT2) (area under the curve from 0.72 to 0.89). Finally, an imbalance in the proportion of immune cells in PCOS was also observed.

**Conclusion:** Five candidate genes were identified as the final diagnostic markers. Our conclusion can provide potential diagnostic candidate genes for patients of MS with PCOS.

Introduction

About 5-18% of women are affected by Polycystic Ovary Syndrome (PCOS)
[1, 2]. The pathogeny is complicated, including genetic and epigenetic susceptibility, hypothalamus and ovarian dysfunction, androgen overexposure, insulin resistance, and obesity-related mechanisms[3, 4]. Metabolic syndrome (MS) is a chronic non-infectious syndrome, and its clinical features include a series of risk factors, including insulin resistance, elevated blood pressure, abdominal obesity, abnormal Blood glucose, and plasma lipids[5, 6]. Therefore, PCOS and MS are closely related but are different entities. In addition, they are both immune-related diseases.

Existing studies have proved that metabolic abnormality correlated factors or their genetic markers are higher in PCOS patients[7, 8]. The incidence rates of glucose intolerance and diabetes[9], atherogenic dyslipidemia[10], systemic inflammation[4], non-alcoholic liver disease[11], and coagulation disorders[12] are higher in PCOS women. According to previous studies, the related pathological changes of MS are closely related to immune system dysfunction[13] and metabolic abnormality[14] related molecules. Therefore,
immune system disorders and metabolic abnormality molecules may contribute to the accurate diagnosis of PCOS patients with MS.

There is controversy about the diagnostic criteria of adolescent PCOS patients. Some diagnostic criteria conflict with the normal physiology of the mature reproductive system, and the diagnostic criteria used to diagnose adult women do not apply to adolescents\textsuperscript{[1]}. PCOS is an essential factor leading to women's infertility\textsuperscript{[15]}. There is a close relationship between PCOS and MS\textsuperscript{[16, 17]}, so it is vital to find sensitive and specific diagnostic tools for PCOS early, especially for adolescent patients with MS.

Nevertheless, there is limited research on the diagnostic criteria for MS combined with PCOS to date. Previously, some researchers explored the plasma biomarkers of MS in PCOS patients by metabonomics\textsuperscript{[18]}. However, the diagnostic genes related to transcriptome have not been explored. Here, we first downloaded the data sets GSE34526, GSE5850, GSE102293(PCOS), and GSE98895(MS) from the Gene Expression Synthesis (GEO) database, identified differentially expressed genes (DEGs) by Limma and found significant module genes by weighted gene co-expression network analysis (WGCNA). Functional enrichment analysis, construction of protein-protein interaction network, The Least Absolute Shrinkage and Selection Operator (LASSO), and analysis of immune cell infiltration are identified the related genes, receiver operating characteristic (ROC) curve is used to evaluate the hub immune-related diagnostic biomarkers of PCOS, and MS. This research may be conducive to determine the immune-related potential diagnostic markers of PCOS in MS patients.

**Materials and methods**

**Microarray Data**

Three raw datasets (GSE34526\textsuperscript{[19]} GSE5850\textsuperscript{[20]} GSE102293\textsuperscript{[21]}) of PCOS datasets were selected, including gene expression data, and one dataset(GSE98895\textsuperscript{[22]}) of MS were downloaded from the GEO database(https://www.ncbi.nlm.nih.gov/geo/\textsuperscript{[23]}) (Figure 1).

**Merge datasets, control batch and preprocessing**

Merge the datasets and control the batch effect\textsuperscript{[24]}. Limma fits a linear equation to the expression of each gene by a generalized linear model. We use the R software package Limma (version 3.40.6) to analyze the differences and get the differential genes between PCOS and the normal group. We set the expression spectrum data set according to |log2 times change (FC)| > 1.5 and p-value < 0.05 as the standard for identifying deg with the Limma package\textsuperscript{[25]}.  

**Weighted Gene Co-Expression Network Analysis**

Weighted gene co-expression network analysis (WGCNA) is suitable for studying biological networks based on the pairwise correlation between variables\textsuperscript{[26]}. On the basis of gene correlation, it allows us to define modules (gene clusters) and explore the correlation between co-expressed modules. Also, WGCNA
can be used as a feature selection method to analyze the potential relationship between phenotypes and gene modules to help screen out gene networks associated with specific phenotypes.

First, we use the gene expression profile, eliminate the top 50% of genes with the smallest MAD, and remove the outliers and samples. Further, use WGCNA to build a scale-free co-expression network.

**Functional Enrichment Analysis**

KEGG rest API was used to obtain the latest gene annotation\(^{[27]}\), Using the GO annotations of genes in the R software package (version 3.1.0) as the background and conducting enrichment analysis using the R software package custerProfiler (version 3.14.3) to obtain the results of gene set enrichment\(^{[29]}\). Set the minimum gene set to 5 and the maximum gene set to 5000, P value of < 0.05 and an FDR of < 0.25, where considered statistically significant. Interacted analysis was conducted between the DEGs of MS and the most significant module genes of MS's WGCNA, and the DEGs of PCOS have interacted with the most significant mode genes of MS's WGCNA. Finally, the results obtained from the two gene sets were subjected to GO and KEGG analysis.

**Protein–Protein Interaction Network Construction**

The interaction between proteins is essential at the cellular functional level. To further explore the protein network of interacted gene expression between PCOS' DEGs and MS' GCNA, the String database was used to establish the protein-protein interaction (PPI) network structure was established\(^{[30]}\). Select genes with interactions in the PPI network for score sorting, and perform subsequent Lasso regression analysis on genes with values>0.

**LASSO regression analysis**

Lasso Regression is a variant of linear regression, which constrains the model through specific regularization methods to avoid overfitting and improve generalization ability\(^{[31]}\). The R software package glmnet was used to integrate survival time, survival status, and gene expression data, and regression analysis was performed using the lasso cox regression method. In addition, we also set up a 3-fold cross-validation to obtain the optimal model. The \(\lambda\) value was set at 0.087; five genes were ultimately identified.

**Receiver Operating Characteristic Evaluation**

The R software package proc (version 1.17.0.1) was used for ROC analysis to calculate the area under the curve (AUC value) and its 95% confidence interval. Generally speaking, AUC>0.7 is considered an excellent diagnostic value\(^{[32]}\).

**Immune Infiltration Analysis**

CIBERSORT is a cell-type component identification method based on gene expression profile data. It can predict the proportion of various cell subpopulations in the sample, thereby analyzing the number and
proportion of different types of cells in a particular organization or sample\cite{33}. The “Cibersort” R software package was used to analyze the proportion of immune cell infiltration between PCOS patients and control groups in order to compare the differences in the proportion of various types of immune cells between PCOS and control groups. The comparison of the proportion of different types of immune cells was visualized by Vioplot\cite{34}.

**Statistical Analysis**

In this research, all data analysis was conducted using R software. A value with P<0.05 is considered significant. ROC analysis was used to obtain the AUC value.

**Results**

**Intersection between data sets and control batch**

Figure 2A shows the intersection between 3 PCOS data sets (GSE34526, GSE5850, GSE102293). Boxplots of gene expressions before and after standardization for 3 data sets, respectively (Figure 2B,2C). UMAP of gene expressions before and after standardization for 3 data sets, respectively (Figure 2D,2E). After controlling the batch effect, the data distribution between each data set tends to be consistent.

**Identifying differential genes**

Using the Limma package, 1513 DEGs were identified in PCOS datasets, of which 781 were up-regulated, and 732 were down-regulated. The thermal map and volcanic map of PCOS DEGs are shown in Figures 3 A, B. Regarding the MS data set, 1449 DEGs (844 up-regulated and 605 down-regulated) were screened out (Figure 3 C, D).

**Weighted Gene Co-Expression Network Analysis**

WGCNA is used to identify the most significantly correlated module in MS, displayed as a yellow module. We chose b=14 (scaleless R2=0.9) as the 'soft' threshold. (Figures 4A, B). The clustering dendrogram of the MS group and the control group are shown in Figure 4C, which further generated ten gene co-expression modules, each with different colors (Figure 4 D, E). The correlation between MS and module genes is shown in Figure 4F, with the yellow module (691 genes) containing the highest coefficient (correlation coefficient=0.74, p=5.2 * 10−8). Therefore, further analysis will be conducted on the yellow module in the future. The members and genes in the MS yellow module have the highest significance (r=0.62) (Figure 4G). Therefore, the correlation between this module gene and MS is most significant. Therefore, the correlation between this module gene and MS is most significant.

**Functional Enrichment Analysis of MS**
In order to further explore whether there is a certain correlation between the pathogenesis of MS and PCOS, we conducted an interactive analysis between the DEGs genes of MS and the critical module genes of WGCNA, followed by functional enrichment of the interacted genes. By crossing 1449 DEG and 691 genes in the yellow module, a total of 108 common genes were screened out. Finally, 108 interacting genes were screened out (Figure 5A).

CGs were primarily enriched in the “PD-L1 expression and PD-1 checkpoint pathway in cancer,” “Fatty acid metabolism,” “Th1 and Th2 cell differentiation”, “Metabolic pathways,” and “Endocrine resistance” from KEGG analysis (Figure 5B). CGs were mainly enriched in biological process (BP) terms, including “protein localization to postsynaptic specialization membrane,” “regulation of alpha-beta T cell proliferation,” “response to interleukin-3,” and “CD4-positive, alpha-beta T cell proliferation” from GO analysis (Figure 5C). the CGs were mainly located in the “mRNA cleavage and polyadenylation specificity factor complex,” “mediator complex”, and “integral component of Golgi membrane” from cellular component (CC) ontology (Figure 5D). “GTPase activity,” “carbohydrate derivative binding,” “immunoglobulin binding,” and “immunoglobulin binding” were the most significant items in CGs from Molecular function (MF) analysis (Figure 5E). CGs of MS were mainly related to immunity and metabolic reaction.

Functional Enrichment Analysis of PCOS With MS and Identification of Candidate Genes

In order to further explore the association between MS essential genes and the pathogenesis of PCOS, 57 genes were ultimately selected by interacting with the DEGs of PCOS and the critical module genes of MS (Figure 6A). Fifty-seven genes were primarily enriched in the “Metabolic pathways,” “Antigen processing and presentation,” and “Allograft rejection”; all of the above ontologies were intimately related to the immune function and metabolic pathway from KEGG enrichment analysis (Figure 6D). Also, genes were enriched in “Allograft rejection,” “glycosyl compound metabolic process,” and “interleukin-12-mediated signaling pathway” (BP); “extracellular exosome,” “proteasome complex,” and “insulin-responsive compartment” (CC); and “lipid binding,” “intramolecular oxidoreductase activity, interconverting keto- and enol-groups,” and “acid phosphatase activity” (MF) from GO analysis (Figures 6E–G).

The enriched pathway of interacting genes is closely related to immunity, and then node genes that interact with the genes are mined through PPI protein networks for subsequent filtering analysis. Figure 6B shows a close interaction between 35 genes; The genes are sorted according to the counts in Figure 6C.

LASSO regression and ROC analysis

Figure 7A and B show that the LASSO regression algorithm identified six biomarkers with diagnostic values (RAB4B, SFP91, ARHGAP33, LIME1, ARL6PI, and SHMT2). Subsequent ROC analysis was performed, and the ARL6PI gene (AUC=0.65) was removed. The AUC values of the other five genes were all>0.7 (Figure C-H).

Immune Cell Infiltration Analysis
From the above conclusion, we observe that the enriched functions of MS-related genes are closely related to the pathogenesis of PCOS. Both diseases are highly enriched in immune function, so that we will conduct immune-related analysis on PCOS below. The proportion of 22 kinds of immune cells in the PCOS group and control group samples is different, which is shown in the bar plot (Figure 8A). The boxplot demonstrated that PCOS patients had a higher level of monocytes, M0 macrophages, activated dendritic cells, and activated mast cells, and a lower level of native CD4 T cells (Figure 8B).

Correlation analysis between immune cells revealed that resting memory CD4 T cells were positively associated with activated dendritic cells (r = 0.81). That activated memory CD4 T cells were positively related to δγ T cells (r = 0.78), whereas native B cells were negatively related to neutrophils (r = −0.68) (Figure 8B). In summary, the proportion of immune cell infiltration in patients with polycystic ovary syndrome significantly changes, and the correlation between these changes is significant, which may be a potential regulatory point for the treatment of polycystic ovary syndrome.

Discussion

PCOS is one of the leading causes of infertility in women. There is no previous study to find diagnostic genes using the transcriptome data of two diseases. Recent studies have identified several biomarkers for PCOS diagnosis, such as sCD36[35], SHBG[36], and some adipocytokines (adiponectin, visfatin, and vaspin and apelin), copeptin, irisin, PAI-1 and zonulin[37]. In this study, we used a series of integrated bioinformatics analyses to identify candidate biomarkers for PCOS in MS patients. Our final result was the identification of five essential immune-related candidate genes (RAB4B, SFP91, ARHGAP33, LIME1, and SHMT2).

We first selected candidate genes through methods such as differential genes and WGCNA to demonstrate a close correlation between the pathogenesis of the two diseases. Then, we used LASSO regression and ROC analysis to ultimately evaluate the accuracy of the candidate genes. When the AUC value of the ROC curve 0.7, it implies accuracy[38]. When the gene expression is high, we can determine whether PCOS is present based on the gene expression results of MS patients.

Ras-Related Protein Rab-4B (RAB4B) is a member of the RAS superfamily of small GTPases involved in vesicular trafficking[39]. Its functions mainly focus on the innate immune system and protein metabolism. RAB4B is a small GTP enzyme that primarily maintains the localization of glucose transporter GLUT4 in adipocytes in terms of function. The RAB4B dysfunction of T cells will lead to an imbalance of Treg/Th17 in fat, leading to pathological changes such as adipose tissue dysfunction and insulin resistance[40]. Large amounts of studies show that strong correlation between insulin resistance, lipid metabolic dysfunction, and polycystic ovarian syndrome (PCOS)[3, 40]. RAB4B is also overexpressed in PCOS. Thus we suggest that RAB4B could represent a potential diagnostic target for PCOS in MS patients.
ZFP91 can cause abnormal expression of FOXA1, leading to increased proliferation and migration activity and increased chemotherapy resistance\textsuperscript{[41]}. ZFP91 is related to drug resistance of T-cell lymphoma\textsuperscript{[42]}. Thus, ZFP91 significantly regulates the homeostasis and function of T cells\textsuperscript{[41]}. Peripheral blood Tregs in patients with PCOS decreased compared to the normal group\textsuperscript{[43]}. ZFP91 take a significant impact on tumorigenesis and NF-κB pathways\textsuperscript{[44]}. Many signal pathways are affected by OS, such as Keap1-Nrf2, NF-κB, FOXO, and MAPK pathways\textsuperscript{[44]}. Oxidative stress (OS) is closely related to PCOS\textsuperscript{[45]}. ZFP91 is overexpressed in MS patients with PCOS. Therefore, we speculate that ZFP91 may induce immune function changes and pathological processes such as OS, leading to PCOS. In addition, we infer that ZFP91 can be used as a biomarker for diagnosis.

The pathways involved in the ARHGAP33 (Rho GTPase Activating Protein 33) gene include Rho GTPase signaling and RAC1 GTPase cycling\textsuperscript{[46]}. It may be involved in the process of intracellular transportation\textsuperscript{[47]}, and participate in the process of insulin-regulating glucose transport\textsuperscript{[46]}. Previously, a large number of experimental and clinical data showed that compared with healthy controls, PCOS patients had abnormal endometrial expression disorder, increased endometrial insulin resistance, impaired glucose transport and utilization, followed by chronic inflammation, immune dysfunction, changes in uterine blood vessel distribution and other abnormal changes\textsuperscript{[1]}. Therefore, we speculate that ARHGAP33 can be used as a biomarker for diagnosis.

LIME1 encodes a transmembrane adapter protein that can bind to the Src family kinases Lck and Lyn, thereby connecting downstream signaling pathways after T and B cell signal stimulation\textsuperscript{[48,49]}. It participates in signal transduction mediated by BCR and TCR. In the absence of TCR signaling, it may be involved in CD4-mediated inhibition of T-cell activation. T-cell dysfunction may play a role in the pathogenesis of PCOS. Therefore, LIME1 is obviously related to PCOS.

Previous studies have demonstrated that immune dysfunction and metabolic disorders are closely related to the pathological mechanism of PCOS\textsuperscript{[3,7]}. An abundant number of studies have demonstrated that insulin resistance and immune system dysfunction play an important role in PCOS. According to our study, the pathway functions enriched by DEGs in PCOS patients are mainly in immune function and metabolism. PCOS patients had a higher level of monocytes, M0 macrophages, activated dendritic cells, and activated mast cells, and a lower level of native CD4 T cells, which is consistent with previous research results. Above all, identifying biomarkers can provide precise treatment for the diagnosis and targeted treatment of PCOS.

**Declarations**

**Conflict of Interest**

There are no conflicts of interest regarding the publication of this article.

**Authors' contributions**
Yi Xu conceived and designed the study, drafted the article, revised the article critically, and be responsible for materials.

Yihui Gu had final approval of the submitted versions.

**Funding**

This study was supported by National Natural Science Foundation of China (No. 82072078), Jiangsu Province Key Research and Development Project (SBE2020741118), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJCX22_0070).

**Ethics approval and consent to participate**

Not applicable.

**Availability of data and materials**

The following information was supplied regarding data availability: Data is available at the GEO database (https://www.ncbi.nlm.nih.gov/geo/).

**Acknowledgements**

Not applicable

**Consent for publication**

Not applicable

**References**


**Figures**
Figure 1

Flowchart. GSE, gene expression omnibus series; DEGs, differentially expressed genes; Limma, linear models for microarray data; WGCNA, weighted gene co-expression network analysis.
Figure 2

Before and after Controlling batch effects on PCOS datasets

(A) Combined histogram of three data sets for PCOS. (B) Boxplot before batch control. (C) Boxplot after batch control. (D) UMAP before batch control. (E) UMAP after batch control.
Figure 3

Differentially expressed genes identification from the integrated PCOS dataset and MS dataset.

(A) Heatmap of the integrate three datasets of GSE34526 GSE5850 GSE102293. (B) Volcano map of the integrate three datasets of GSE34526 GSE5850 GSE102293. (C) Heatmap of GSE98895. (D) Volcano map of GSE98895.
Figure 4

Module gene of MS obtained by WGCNA method.

(A, B) Analysis of combining scale independence with average connectivity. (C) Clustering dendrogram between MS and control groups. (D) Co-expression module of gene tree. (E) Heatmap of eigengene adjacency. (F) Heatmap of the association between modules and MS. The yellow module is shown to be significantly related to MS. (G) Correlation diagram between module members and gene significance in yellow module.
Figure 5

Enrichment and WGCNA analysis.

(A) 108 genes interacted by MS DEGs and WGCNA yellow module genes of venn diagram. (B) KEGG pathway analysis of the intersected genes. (C–E) GO analysis of the intersected genes, including biological process (BP), cellular component (CC), and molecular function (MF), respectively from GO analysis.
Figure 6

Enrichment analysis of intersected genes between WGCNA yellow module gene and DEGs of PCOS.

(A) Venn diagram shows that 57 genes were intersected by DEGs of PCOS and yellow module gene of WGCNA. (B) PPI network reveals the interaction of 35 genes. (C) The bar chart shows the gene nodes of
35 genes in PPI network. (D) KEGG analysis of 57 intersected genes. (E–G) GO analysis of 57 intersected genes.

Figure 7

Lasso regression analysis.
(A, B) Biomarkers screening in the Lasso model, six diagnostic genes suitable for PCOS patients with MS were screened out. (C–H) The ROC curve of each candidate gene (RAB4B, SFP91, ARHGAP33, LIME1, ARL6PI, and SHMT2) show the diagnostic value, the AUC value of five genes is > 0.7, except ARL6PI.

Figure 8

Immune cell infiltration analysis of PCOS.
(A) The barplot shows the proportion of 22 kinds of immune cells in the merged PCOS datasets. (B) Immune checkpoints between PCOS and control groups visualized by the vioplo diagram. (C) Correlation between 22 kinds of immune cells.