BMC Pregnancy and Childbirth

RESEARCH

Open Access

Identification of hub genes, diagnostic model, and immune infiltration in preeclampsia by integrated bioinformatics analysis and machine learning

Yihan Zheng[1†], Zhuanji Fang[2†], Xizhu Wu[1], Huale Zhang[2] and Pengming Sun[3*]

Abstract

Purpose This study aimed to identify novel biomarkers for preeclampsia (PE) diagnosis by integrating Weighted Gene Co-expression Network Analysis (WGCNA) with machine learning techniques.

Patients and methods We obtained the PE dataset GSE25906 from the gene expression omnibus (GEO) database. Analysis of differentially expressed genes (DEGs) and module genes with Limma and Weighted Gene Co-expression Network analysis (WGCNA). Candidate hub genes for PE were identified using machine learning. Subsequently, we used western-blotting (WB) and real-time fluorescence quantitative (qPCR) to verify the expression of *F13A1* and *SCCPDH* in preeclampsia patients. Finally, we estimated the extent of immune cell infiltration in PE samples by employing the CIBERSORT algorithms.

Results Our findings revealed that *F13A1* and *SCCPDH* were the hub genes of PE. The nomogram and two candidate hub genes had high diagnostic values (AUC: 0.90 and 0.88, respectively). The expression levels of *F13A1* and *SCCPDH* were verified by WB and qPCR. CIBERSORT analysis confirmed that the PE group had a significantly larger proportion of plasma cells and activated dendritic cells and a lower portion of resting memory CD4+T cells.

Conclusion The study proposes F13A1 and SCCPDH as potential biomarkers for diagnosing PE and points to an improvement in early detection. Integration of WGCNA with machine learning could enhance biomarker discovery in complex conditions like PE and offer a path toward more precise and reliable diagnostic tools.

Keywords WGCNA, LASSO, Machine learning, CIBERSORT algorithms, Preeclampsia

†Yihan Zheng and Zhuanji Fang contributed equally to this work.

*Correspondence:
Pengming Sun
fmsun1975@fjmu.edu.cn
[1]Department of Anesthesiology, Fujian Maternity and Child Health Hospital, College of Clinical Medicine for Obstetrics & Gynecology and Pediatrics, Fujian Medical University, Fuzhou, Fujian 350001, China
[2]Department of Obstetrics, Fujian Maternity and Child Health Hospital, College of Clinical Medicine for Obstetrics & Gynecology and Pediatrics, Fujian Medical University, Fuzhou, Fujian 350001, China
[3]Department of Gynecology, Fujian Maternity and Child Health Hospital, College of Clinical Medicine for Obstetrics & Gynecology and Pediatrics, Fujian Medical University, Fuzhou, Fujian 350001, China

Introduction

Preeclampsia is a pregnancy-specific disease that complicates up to 5–8% of all pregnancies and is considered one of the leading causes of maternal morbidity and mortality worldwide [1]. With its exact cause remaining unknown, preeclampsia is assumed to arise from the interaction between environmental and genetic factors. Low levels of *PlGF* and high levels of *sFlt-1* have been identified in previous studies as good predictors of PE and adverse pregnancy outcomes in the latter half of pregnancy but have poor performance in identifying women in early pregnancy or those with mild symptoms [2]. Because no specific biomarkers are available to accurately diagnose preeclampsia, it is difficult to effectively prevent or treat the disease [3–5].

The primary challenge for the biomarker discovery of PE is the disease complexity and single biomarker inability to provide diagnosis with enough reliability, especially at early stages of pregnancy [6, 7]. Much of the current research on biomarkers for PE has focused on those that reflect vascular dysfunction, immune dysregulation, and placental abnormalities [8]. However, most of these studies fail in their representation of the complete spectrum of molecular changes in PE and hence yield biomarkers with generally poor sensitivity and specificity, especially in early or mild forms of the disease. Therein lies the critical need for more sensitive and specific biomarkers to identify PE sufficiently early that interventions would be most effective.

Only recently has progress in bioinformatics and machine learning enabled surmounting some of these challenges [9]. Such complex disease studies as PE are specifically suitable for the machine learning methods-random forest and LASSO regression [10, 11]. Gene expression data is of high dimensionality and requires methods that can handle large datasets to manage and analyze it. These models can therefore handle such complexity by selecting, out of the huge amount of data developed in a transcriptomic study, the most relevant features-that is, genes. Random forest is an ensemble learning technique very effective in classifying complex data sets and ranking gene importance based on their contribution to the prediction [12]. By contrast, LASSO regression is useful in feature selection via the application of a regularization penalty, which will help in identifying the most predictive genes while at the same time minimizing overfitting [13]. In fact, these two approaches are particularly valuable in the context of PE, since high-dimensional gene expression data will have to be analyzed carefully in order to bring to light the key biomarkers most relevant for diagnosis.

Apart from these, some of the bioinformatics methods, such as Weighted Gene Co-expression Network Analysis integrated with machine learning models, can further provide deep insight into the molecular mechanisms of PE [14]. WGCNA allows for identifying co-expressed gene modules that may be of biological relevance for the disease [15], while ML techniques like random forest and LASSO can further prioritize the genes based on diagnostic potential. This integrated approach, in addition to enhancing robustness in biomarker discovery, further allows accuracy and reliability of potential biomarkers for clinical use.

Most biomarker studies of PE, although much improved, have been limited by a narrow set of biomarkers relied upon or a lack of comprehensive insight into the molecular pathways of the disease [16]. Herein, we aim to overcome these shortcomings by using a combined approach of WGCNA and machine learning algorithms for the identification of biomarkers for PE. In this work, we have focused on two most promising biomarkers from our analysis, *F13A1* and *SCCPDH*. We integrated bioinformatics and machine learning techniques in identifying these biomarkers, which were further validated by western blotting and qPCR in an independent cohort. Our findings indicate that *F13A1* and *SCCPDH* have diagnostic advantages compared to already existing biomarkers and thus may enable more accurate detection of PE even at its early or mild stages.

We performed transcriptomics data analysis to find the potential novel biomarkers in preeclampsia by combining bioinformatics analysis with machine learning techniques. Bioinformatics analysis in this study included collecting and analyzing publicly available data on preeclampsia, including gene expression data of PE patients and controls (GEO: GSE25906) [17]. GSE25906 has been selected because this dataset contains whole gene expression profiles from placental samples of both PE and control pregnancies, thus allowing a robust comparison of the gene expression patterns between these groups. Besides, many studies have used this dataset before. Hence, this will be a good benchmark for us in terms of validation and comparability to existing literature on biomarkers of PE. Gene Ontology enrichment and Gene Set Enrichment Analysis were applied in this paper to identify the disrupted biological processes and pathways of PE, respectively. With LASSO regression analysis and random forest algorithms, we identified hub gene expression signatures related to preeclampsia. The immune cell infiltrations in placental samples were further investigated to deeply analyze the underlying mechanism of PE and to analyze the relationship between different immune factors and hub genes. The study presents new insights into the molecular mechanism of preeclampsia and points to new biomarkers that may form a basis for early diagnosis of the disease.

Materials and methods

Data collection

The PE dataset used in this study (GSE25906) was obtained from the Gene Expression Omnibus (GEO) database. This dataset included gene expression profiles of placental samples from PE patients ($n=23$) and normal pregnant women ($n=37$). (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA135777).

Differential gene expression analysis

Limma is a differential expression method based on a generalized linear model [18]. Here, we used the R software package limma (version 3.40.6) for differential analysis to obtain differential genes between different PE samples and controls. The DEGs were identified based on the false discovery rate (FDR < 0.05) and absolute log2-fold change (FC > 0.58) criteria.

Weighted gene co-expression network analysis

WGCNA was used to identify modules of co-expressed genes and further identify the hub genes important for PE diagnosis. "WGCNA" package in R was employed to carry out this analysis [19]. The key steps performed in the WGCNA analysis were as follows: 1. **Data Preprocessing:** First, the genes with low variability were excluded by calculating the median absolute deviation (MAD) for each gene. Only the top 50% of genes with the highest MAD were retained to reduce noise. 2. **Soft-Thresholding Power Selection:** A soft thresholding power (β) was determined to ensure the network exhibited scale-free topology, a characteristic of biological networks. We selected $\beta = 3$ based on an analysis of scale independence and mean connectivity, which showed a scale-free R^2 value of 0.9. This choice was critical to creating a reliable adjacency matrix, emphasizing strong correlations while penalizing weaker ones. 3. **Topological Overlap Matrix (TOM):** The adjacency matrix was transformed into a TOM, which represents the shared connectivity between gene pairs, and a dissimilarity measure was calculated. 4. **Module Identification:** Gene modules (clusters of co-expressed genes) were identified using hierarchical clustering and the dynamic tree cut algorithm. A minimum module size of 30 genes was set to ensure the robustness of the obtained modules. Modules were given unique colors for visualization. 5. **Correlation analysis:** Module eigengene (the first principal component) was calculated to represent overall expression pattern of genes within one module. The correlations of module eigengenes to clinical traits, such as PE versus control, were determined to identify modules significantly associated with PE.

The purple module, which showed the strongest correlation with PE, correlation coefficient = 0.65, P = 1.9e-8, was selected for further analysis. This module was hypothesized to contain genes most relevant to PE pathology.

Functional enrichment analysis

Functional enrichment analysis was performed to identify the biological functions and pathways associated with the hub genes. The Gene Ontology (GO) database was used for functional enrichment analysis [20]. The functional enrichment analysis was performed using the R package "clusterProfiler". The criteria for functional enrichment analysis using Gene Ontology (GO) analysis were set as a p value < 0.05. The analysis was performed based on the intersection of DEGs and the most significant module genes. The results of GO enrichment analysis were visualized via the Sangerbox platform [21].

Machine learning algorithms

LASSO regression and RF were used to identify candidate hub genes for the diagnosis of PE. LASSO regression is a popularly used method for feature selection and model fitting [22], while RF is a popularly used method for classification and feature selection [23]. The performance of the algorithms was assessed by the AUC. The LASSO regression and RF analyses were performed using the "glmnet" [24] and "randomForest" [25] R packages, respectively. Those genes present in both the LASSO and RF analyses were considered as potential hub genes for the diagnosis of PE.

Validation via western blotting and qPCR

For external validation, placental tissue samples were obtained from 12 women who delivered at Fujian Maternal and Child Health Hospital between April and July 2023. The cohort included six PE patients and six controls. Inclusion criteria for PE patients were based on the International Society for the Study of Hypertension in Pregnancy (ISSHP) guidelines, which include hypertension (≥ 140/90 mmHg) after 20 weeks of gestation and evidence of end-organ damage or proteinuria (> 300 mg/24-hour urine), and Early-Onset PE was diagnosed before 34 weeks of gestation. Late-Onset PE ($n=3$) was diagnosed after 34 weeks of gestation [26]. Controls were healthy pregnant women with no history of hypertension or pregnancy complications. Exclusion criteria included chronic hypertension, multiple pregnancies, or autoimmune diseases.

Demographic and clinical characteristics, including age, body mass index (BMI), mode of delivery, blood pressure, and proteinuria levels, were collected to ensure representativeness.

Western Blotting Protein was extracted from placental tissue, and the expression of *F13A1* and *SCCPDH* was evaluated using antibodies specific to these proteins.

F13A1 antibody is derived from Abcam, article number AB179444, with a dilution ratio of 1:10,000 and a molecular weight of 80 kDa. *SCCPDH* antibody is derived from Abcam, article number AB185709, with a dilution ratio of 1:2000 and a molecular weight of 47 kDa. (Supplementary Table 3).

qPCR RNA was extracted from placental tissues and reverse-transcribed into cDNA. Gene expression levels were quantified using primers specific to *F13A1* and *SCCPDH*. Primer sequences and reaction conditions are detailed in Supplementary Tables 4–5.

This study was approved by the Ethics Committee of Fujian Maternity and Child Health Hospital (2022KYLLRD01038).

Diagnostic model

A nomogram was constructed using the hub genes identified in the machine learning algorithms. The nomogram was constructed using the R package "rms" [27]. The diagnostic value of the nomogram was evaluated using the receiver operating characteristic (ROC) curve. The nomogram assigns a score to each of the candidate genes, represented by "Points." The total score, or "Total Points," is calculated as the sum of the scores of all the candidate genes. This information can then be used to help diagnose PE by considering the overall score of the candidate genes.

Immune Cell Infiltration Analysis

In this study, the proportion of immune cells in PE and control samples was determined by using CIBERSORT, a computational method that employs gene expression profiles to identify immune cell proportions [28]. The "Cibersort" R package was utilized to perform the immune cell infiltration analysis. The bar plot was utilized to visually represent the proportion of each type of immune cell in the various samples, and the vioplot was used to compare the proportion of different immune cells between the PE and control groups. Additionally, the correlation between 22 types of infiltrating immune cells was depicted using a heatmap, which was created with the "corrplot" R package [29].

Statistical analysis

Statistical analysis and visualization were performed using R version 4.1.3. The ROC curve and the calculation of AUC along with its 95% confidence interval were established using SPSS Version 26.0 (IBM Corporation, Armonk, NY, USA). The comparison of the proportions of various immune cells between the control group and the AVC group was carried out using Student's t test, which was performed using GraphPad Prism Version 8.3.0 (GraphPad Software, San Diego, CA, USA). Image J software was used to measure the gray value of each sample in each group. SPSS was used to analyze the differences between groups. All statistical tests were two-sided and *P*-value less than 0.05 was considered statistically significant.

Results

Identification of differentially expressed genes

A visual representation of the study's methodology is depicted in Fig. 1, starting with gene expression data obtained from the placentas of 23 PE patients and 37 controls. A total of 45 DEGs were identified in the PE dataset using the Limma method, of which 30 were upregulated and 15 were downregulated. Compared to the original publication's 128 DEGs, which used unsupervised hierarchical clustering for analysis [17], Limma has an advantage in comparing differential gene expression between two groups of samples. It uses a linear model to estimate the mean-variance relationship of the data and then applies empirical Bayes methods to obtain moderated t-statistics and p values for each gene [30]. The heatmap and volcano plot of PE DEGs are shown in Fig. 2A and B, respectively.

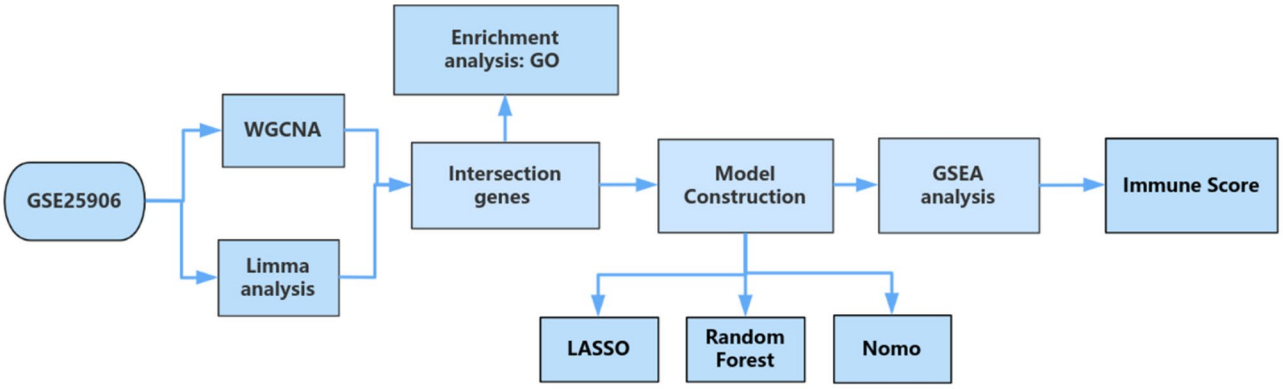

Fig. 1 The workflow of the analyses

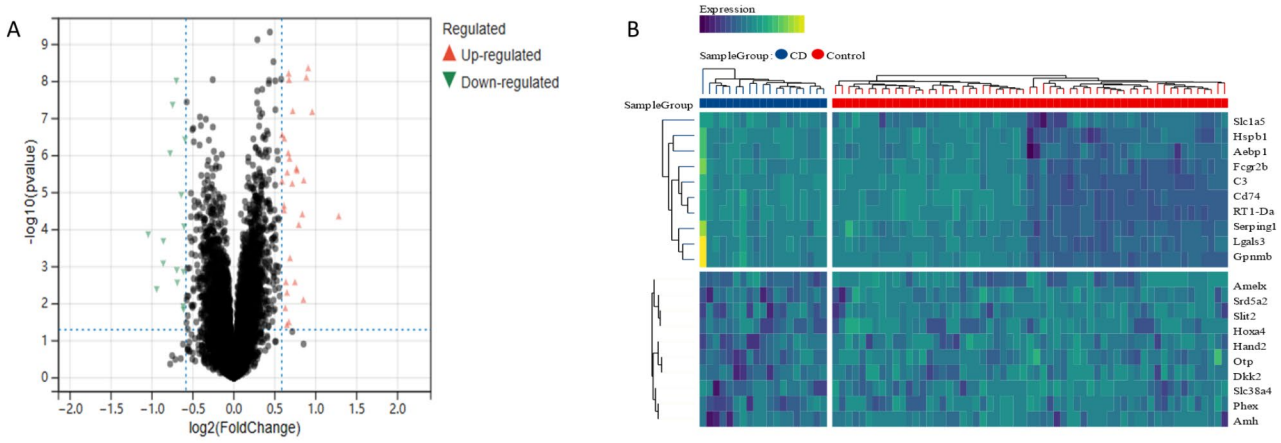

Fig. 2 Differentially expressed genes between PE and control samples. (**A**) Genes in red represent significantly high expression in PE, genes in green represent significantly high expression in control, and gray indicate insignificant changes. (**B**) The heatmap shows the all the 45 genes significantly highly expressed in PE or control samples

Interpretation The identified DEGs likely represent molecular disruptions associated with PE, including genes involved in immune regulation, vascular remodeling, and placental function. These disruptions underline the complexity of PE pathology, which involves both maternal and placental factors.

Weighted gene co-expression network analysis and key module identification

As a complementary approach that focuses on gene expression correlation, we applied WGCNA to determine the strongest gene modules with member gene expression associated with PE. A "soft" threshold of $\beta=3$ (scale-free $R^2=0.9$) was selected based on the average connectivity and scale independence (as depicted in Fig. 3A and B). The clustering dendrogram of the PE and control samples is shown in Fig. 3C. As a result of the threshold, six gene co-expression modules (GCMs) were generated and are displayed in different colors in Fig. 3D and F. The correlation between PE and the GCMs is shown in Fig. 3E, with the purple module (comprising 44 genes) demonstrating the highest correlation (PE coefficient=0.65, P=19e-8) and thus being identified as the central module for further analysis. The correlation between module membership and gene significance in the purple and black modules was calculated, and a significant positive correlation was observed between them (r=0.58, 0.46, respectively), as depicted in Fig. 3G and H. Thus, the genes in the purple module were found to be most significantly related to PE.

Interpretation The strong association between the purple module and PE suggests that genes in this module may play key roles in disease development. Functional enrichment analysis revealed involvement in biological processes like activin binding and regulation of gonadotropin secretion, highlighting their potential roles in placental hormone regulation and immune response.

Functional enrichment analysis of genes associated with preeclampsia

To determine the reliability of the results in reflecting the pathogenesis of PE, we conducted a functional enrichment analysis based on the genes that were shared from the differential expression analysis and WGCNA module genes. Our analysis resulted in the identification of 10 common genes (CGs) from the intersection of 45 DEGs and 17 genes in the purple module (as depicted in Fig. 4A). The 10 genes are *RDH13*, *ENG*, *F13A1*, *DNAJC3*, *JAK1*, *TUBA1A*, *STRADBP1*, *SCCPDH*, *SPAG4*, *and TUBAP2.*

GO analysis showed that the CGs were significantly enriched in biological process (BP) terms, including "activin binding", "regulation of follicle-stimulating hormone secretion" and "positive regulation of gonadotropin secretion" (Fig. 4B). The enrichment analysis revealed that the CGs of PE were mainly related to balancing hormones and supporting reproductive function.

Interpretation These results underscore the importance of placental signaling pathways in PE and suggest that targeting these processes could offer therapeutic or diagnostic opportunities.

Identification of candidate hub genes via machine learning

Independently, we also applied LASSO regression and RF machine learning algorithms to evaluate candidate genes for their potential as PE diagnostic tools. The results of these algorithms are shown in Fig. 5A and B for LASSO regression and Fig. 5C and D for RF. The LASSO regression algorithm identified 10 potential candidate biomarkers, while the RF algorithm ranked genes based on

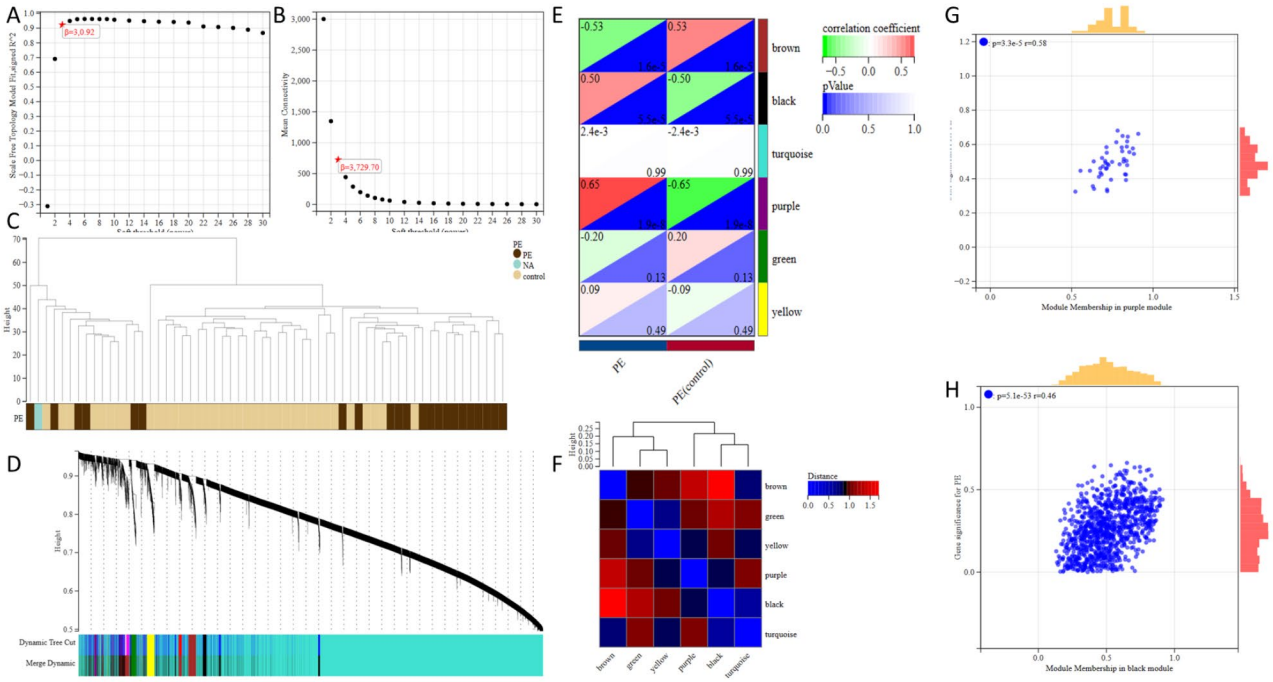

Fig. 3 Results of the WGCNA. (**A**) β = 3 was selected as the soft threshold with the combined analysis of scale independence and average connectivity. (**B**) The corresponding mean connectivity values at different soft threshold powers. (**C**) Clustering dendrogram of the PE and control samples. (**D**) Cluster dendrogram of genes. (**E**) Correlations between different modules and PE. Red indicates a positive correlation, and green represents a negative correlation. (**F**) Gene co-expression modules represented by different colors under the gene tree. (**G**, **H**) Correlation of module membership and gene significance in the purple and black module

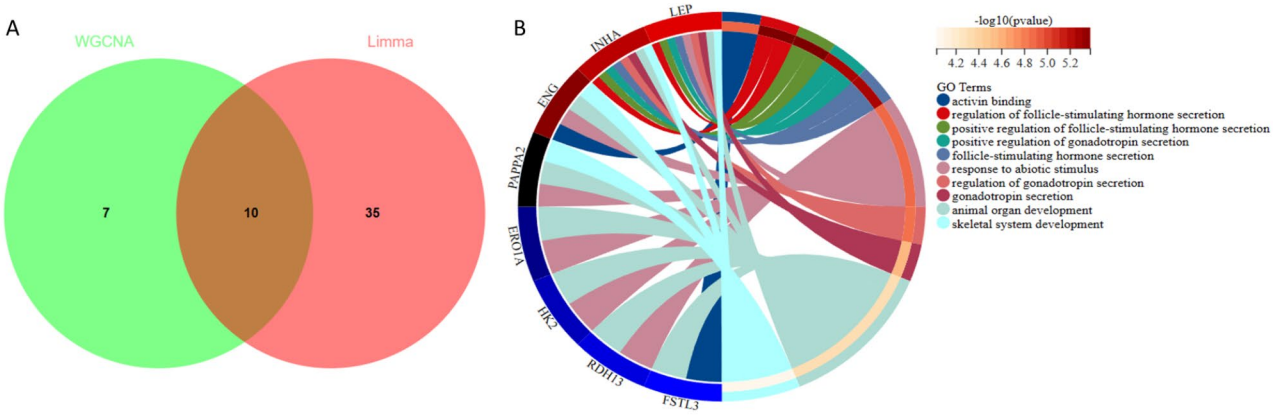

Fig. 4 Intersection genes and GO analysis. (**A**) Ten hub genes were obtained by taking the intersections of the DEGs and ME purple module genes of the WGCNA. (**B**) Biological processes in which the common genes were involved

their importance. ROC showed that RF model has good predict ability (Fig. 5E), and the multidimensional scale diagram indicated the separating capacity of the model (Fig. 5F).

Interpretation *F13A1*, a coagulation-related gene, and *SCCPDH*, involved in amino acid metabolism, represent distinct but potentially complementary aspects of PE pathology. *F13A1's* role in blood coagulation aligns with the hypercoagulable state observed in PE, while *SCCPDH's* involvement in metabolic pathways suggests links to placental stress and dysfunction.

The expression of *F13A1* and *SCCPDH* in PE placenta tissue

The baseline data table presents the demographic and clinical characteristics of the early-onset preeclampsia (PE), late-onset PE, and control groups. The **early-onset PE** group had a mean age of 27.7 years, BMI of 29.9 kg/m^2, systolic blood pressure of 151 mmHg, diastolic blood pressure of 94.3 mmHg, and 24-hour urine protein levels

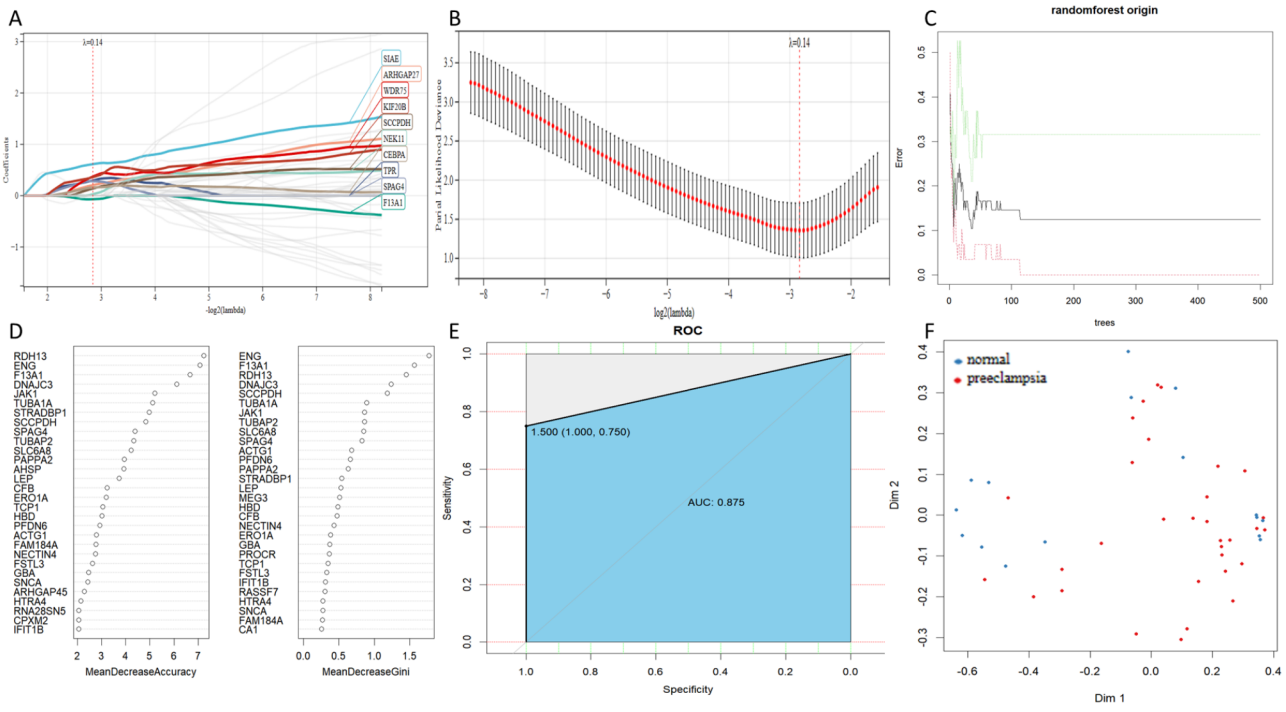

Fig. 5 Machine learning in screening candidate diagnostic biomarkers for PE: (**A**, **B**) Biomarkers screening in the Lasso model. The number of genes ($n=10$) corresponding to the lowest point of the curve is the most suitable for PE diagnosis. (**C**, **D**) The random forest algorithm shows the error in PE; control group and genes were ranked based on the importance score. (**E**) ROC showed that RF model's AUC = 0.875, and (**F**) the multidimensional scale diagram indicated the separating capacity of the model

of 320 mg. The **late-onset PE** group had a mean age of 31.0 years, BMI of 31.3 kg/m², systolic blood pressure of 152.3 mmHg, diastolic blood pressure of 96.0 mmHg, and 24-hour urine protein levels of 350 mg. In contrast, the **control group** had a mean age of 30.3 years, BMI of 28.5 kg/m², systolic blood pressure of 120 mmHg, diastolic blood pressure of 76.5 mmHg, and 24-hour urine protein levels of 116.3 mg (Supplementary Table 1). Under WB, we observed the gray value of the *F13A1* was down-regulated in the PE group and up-regulated in the control group, in contrast, *SCCPDH* was up-regulated in PE group and down-regulated in control group (Fig. 6A, B). Similarly, by qPCR we found that *F13A1* was down-regulated in the PE group and up-regulated in the control group, while *SCCPDH* was up-regulated in the PE group and down-regulated in the control group (Fig. 6C, D). Therefore, we validated our previous findings using placenta tissue obtained from Fujian Maternity and Child Health Hospital.

Interpretation The downregulation of *F13A1* may contribute to impaired vascular remodeling, a hallmark of PE, while *SCCPDH* upregulation could reflect metabolic adaptations or stress responses in the placenta. These findings support their roles as biomarkers and potential contributors to PE pathology.

Diagnostic value assessment

Overlapping between the top 10 genes from RF and the 10 potential candidate genes led to the identification of two final candidate genes (*F13A1 and SCCPDH*) (Fig. 7A). The nomogram was constructed based on the two candidate genes (Fig. 7B), and an ROC curve was established to assess the diagnostic specificity and sensitivity of each gene. The AUC and 95% CI were as follows: *F13A1* (AUC: 0.90, 95% CI 0.82–0.98) and *SCCPDH* (AUC: 0.88, 95% CI: 0.79–0.97) (Fig. 7C, D).

Interpretation The high AUC values for *F13A1* and *SCCPDH* highlight their potential as accurate biomarkers for diagnosing PE.

Expression of hub genes and biological pathway enrichment

The expression of *F13A1* was downregulated in the PE group, while *SCCPDH* was upregulated in the PE group (Fig. 8A). Gene Set Enrichment Analysis (GSEA) [31] of hub genes has shown that they are associated with tissue function, immune regulation and the circulatory system, including gap junction, the TGF-β signaling pathway and cardiac muscle contraction (Fig. 8B, C).

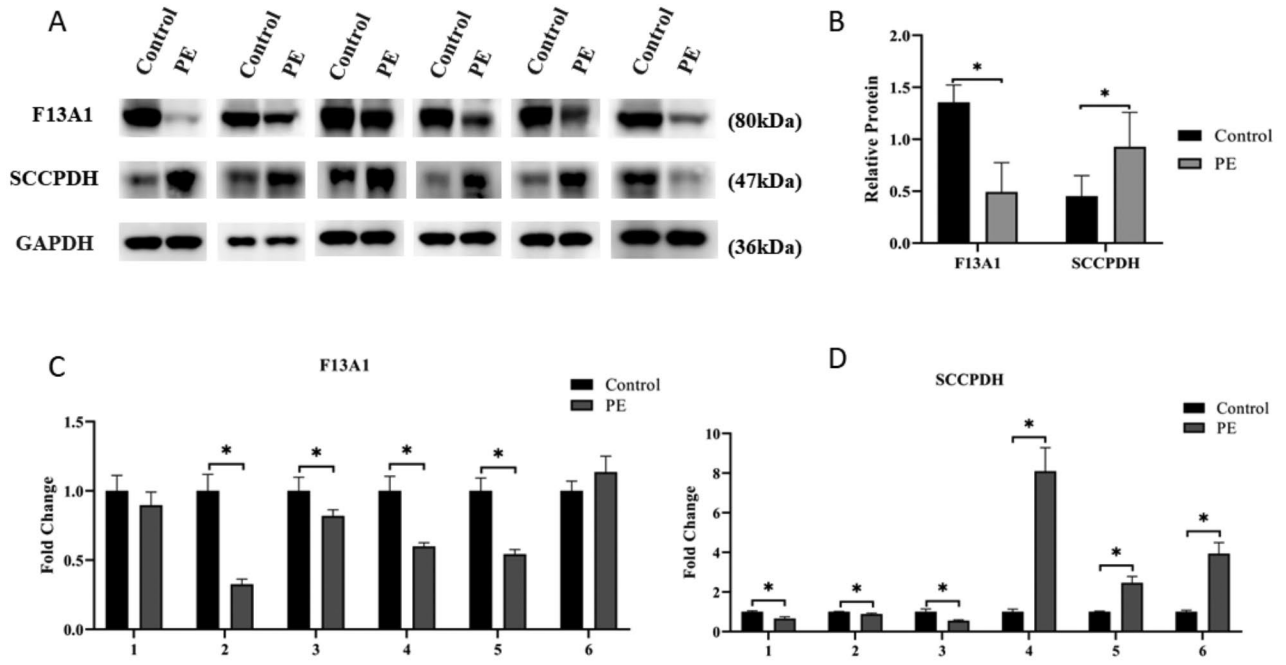

Fig. 6 The expression of *F13A1* and *SCCPDH* in placenta tissue. (**A**) Expression of *F13A1* and *SCCPDH* in western blotting, the lanes represent samples from two groups (6 vs. 6 comparison). Each lane came from a separate blot run under identical conditions. (**B**) The difference in gray value of western blotting. (**C**) Expression of *F13A1* in qPCR. (**D**) Expression of *SCCPDH* in qPCR

Interpretation The differential expression of *F13A1* and *SCCPDH* suggests their involvement in distinct but interconnected aspects of PE pathology.

Immune cell infiltration analysis

Because the bulk placenta was used for the gene expression analysis, we wanted to test whether the difference in gene expression was related to cell population change. To address this issue, we performed cell composition deconvolution analysis using CIBERSORT [28] on 23 PE samples and 37 control samples. The results for the 22 blood cell types are displayed in a histogram (Fig. 9A) and a boxplot (Fig. 9B). Figure 8C showed data matrix correlation analysis and its heat map. The data showed that the PE group had a significantly larger proportion of plasma cells ($P=0.02$) and activated dendritic cells ($P=0.02$) and a lower portion of resting memory CD4+T cells ($P=0.03$) than the control group.

Interpretation These findings highlight an altered immune landscape in PE, characterized by enhanced inflammatory responses and impaired adaptive immunity. The interactions between immune cells and hub genes like *F13A1* and *SCCPDH* could provide further insights into the immune dysregulation in PE.

Biomarker interaction hypotheses

Given the somewhat different roles of *F13A1* and *SCCPDH*, these markers may also act in intersecting pathways to contribute to PE. For example, downregulation of *F13A1* might impede coagulation and vascular stability, while upregulation of *SCCPDH* may signal metabolic shifts attempting to compensate for placental stress. Both pathways could intersect at the point of immune dysfunction to worsen clinical manifestations of PE. Further studies are required to know how these biomarkers might interact within the molecular landscape of PE.

Summary of key findings

The Supplementary Table 2 summarizes the identified hub genes, their expression trends, and biological relevance.

Discussion

The results of our study demonstrate the potential of using bioinformatics analysis and machine learning algorithms in identifying novel biomarkers for the diagnosis of preeclampsia. By combining different computational techniques, we were able to identify two genes (*F13A1* and *SCCPDH*, AUC: 0.90 and 0.88, respectively) that have high diagnostic value for PE patients. The nomogram constructed using these two hub genes could be used as a diagnostic tool for PE.

The identification of *F13A1* and *SCCPDH* as potential biomarkers for PE is an important finding. Coagulation Factor XIII A Chain (*F13A1*) is a serine protease inhibitor that has been shown to be involved in various

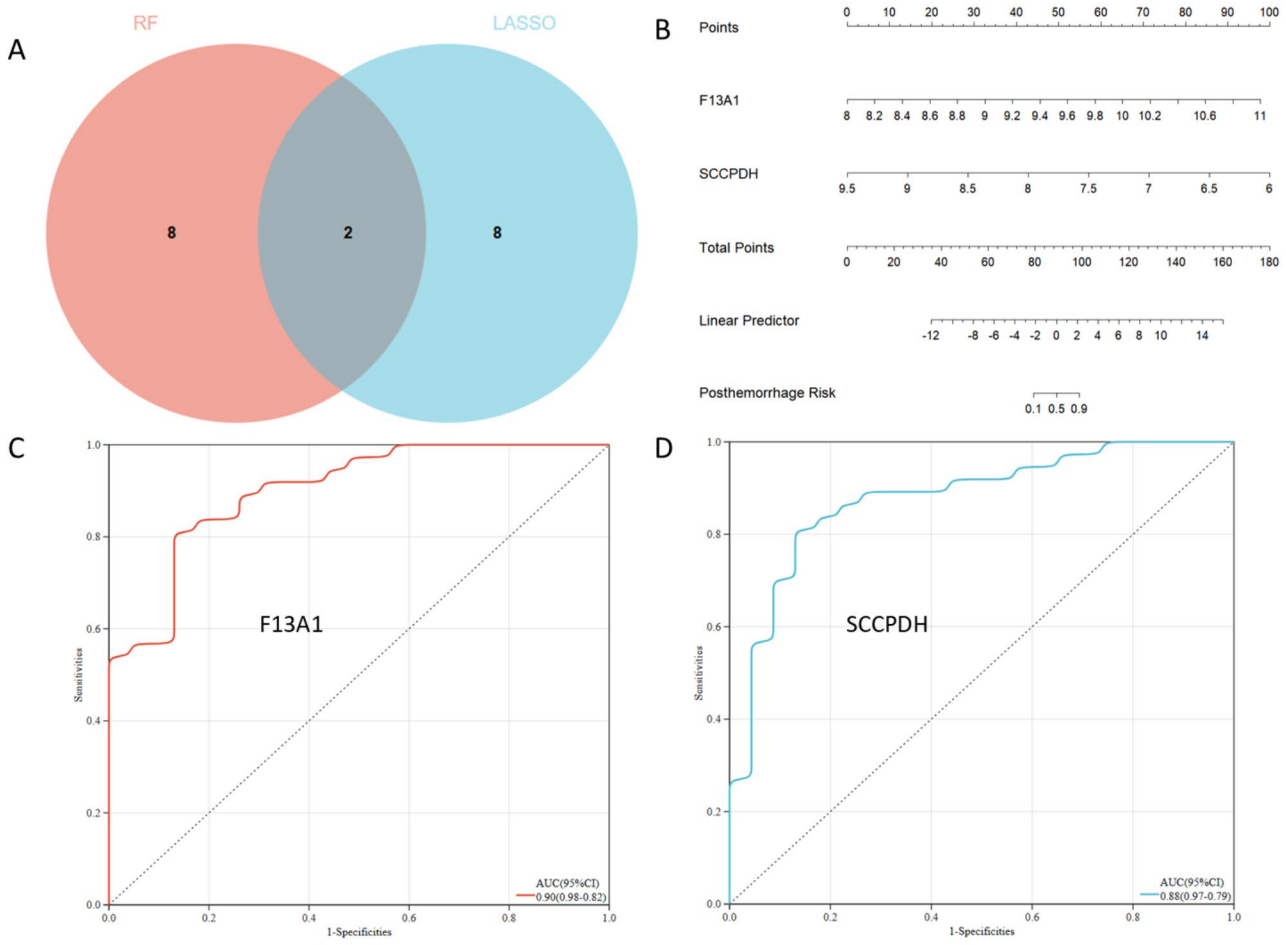

Fig. 7 Nomogram construction and the diagnostic value evaluation. (**A**) Venn diagram shows that two candidate diagnostic genes are identified via the above two algorithms. (**B**) The visible nomogram for diagnosing PE. (**C**, **D**) The ROC curve of each hub gene

physiological processes, including the regulation of blood coagulation and fibrinolysis [32]. Previously, Yu Shangguan [33] and Epiney et al. [34]. also found that *F13A1* is differentially expressed (decreased) in placental tissues of individuals with preeclampsia compared to controls.

Hofbauer cells have thus been implicated in influencing both the trophoblast invasion and angiogenesis known to take place throughout normal placental development [35]. Downregulation of *F13A1* in PE may compromise these functions, thereby predisposing to abnormal vascular remodeling, decreased placental perfusion, and increased inflammation characteristic of the condition [36, 37]. Further investigations are required on the mechanistic role of *F13A1* in Hofbauer cells, especially regarding the influence on trophoblast-Hofbauer cell crosstalk and its subsequent effects on angiogenic-immune pathways. Such interactions can be further elucidated using immunohistochemical analyses and functional assays, which may unravel new therapeutic targets for the management of PE.

Saccharine dehydrogenase (*SCCPDH*) is an enzyme involved in the metabolism of lysine, an essential amino acid. *SCCPDH* and *F13A1* are both involved in the response to elevated platelet cytosolic Ca2+ [38]. The function of *SCCPDH* and its role in PE are not well understood. Some studies have demonstrated in animal models that *SCCPDH* is involved in the regulation of chronic stress, which contributes to anxiety depression [39]. There is evidence to suggest that anxiety and depression are prevalent in women with preeclampsia and that these mental health conditions can have an impact on the onset and course of the disease [40, 41].

A plausible hypothesis is that *F13A1* downregulation and *SCCPDH* upregulation may converge on immune regulation pathways, influencing the inflammatory milieu in PE. Dysregulated interactions between trophoblast cells and maternal immune cells could create a feedback loop of vascular and metabolic dysfunction, further impairing placental function. Future studies should explore these interactions to elucidate the mechanisms driving PE and identify potential therapeutic targets.

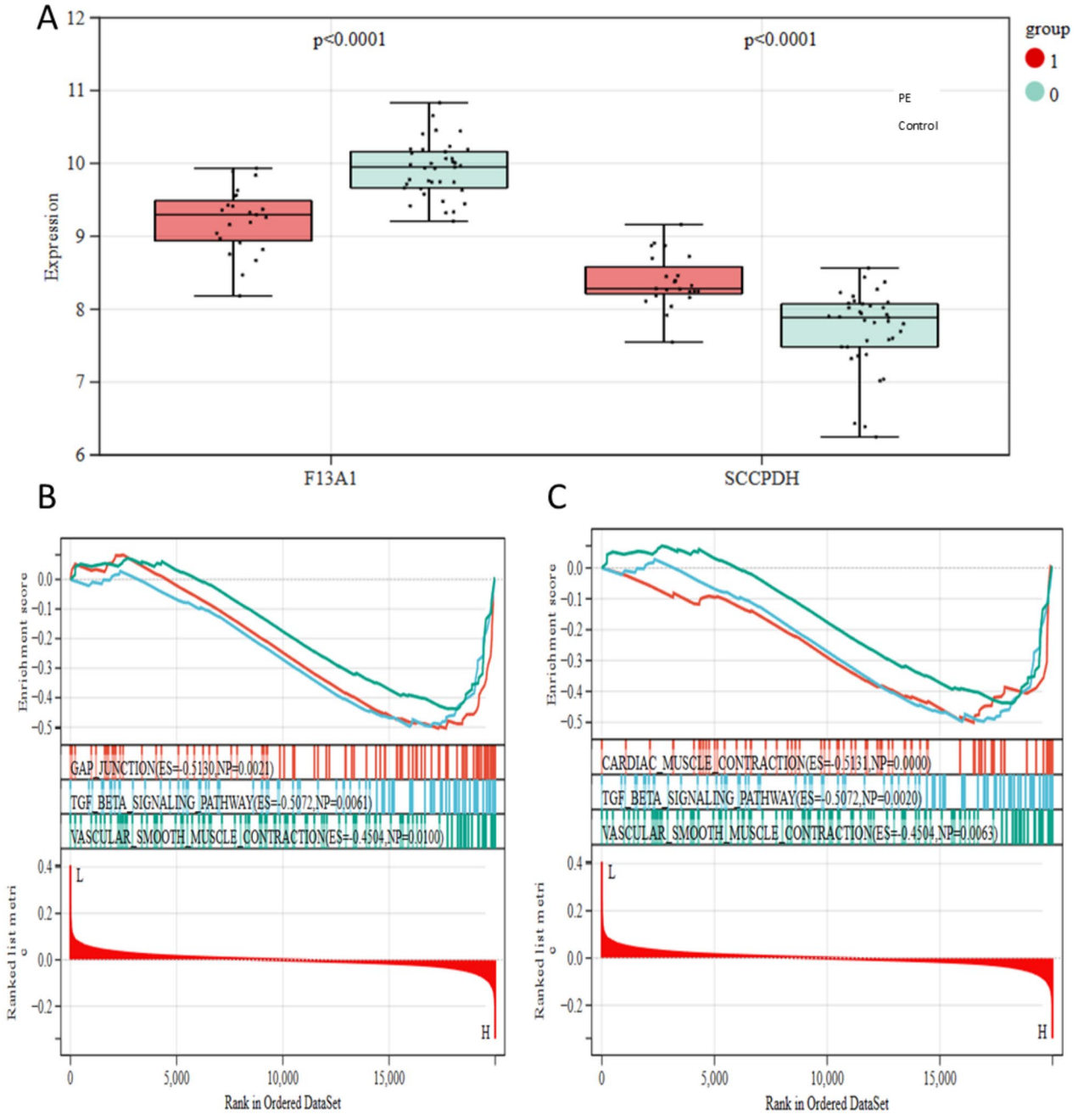

Fig. 8 Expression of the top two candidate genes. Expression of hub genes in the PE and control groups (**A**), and GSEA revealed the enriched pathway of hub genes. (**B**) *F13A1*; (**C**) *SCCPDH*

In addition to individual genes, our study also identified pathways affected in PE. The relationship between the TGF-β signaling pathway and preeclampsia is not fully understood, but TGF-β signaling is believed to play a role in the development of the condition [42]. TGF-β is a signaling molecule that regulates various physiological processes, including inflammation and the immune response, and it has been shown to be involved in the development of hypertension [43, 44] and renal dysfunction [45]. Some studies have suggested that TGF-β signaling may contribute to the pathogenesis of preeclampsia by disrupting the normal functioning of maternal blood vessels, leading to increased blood pressure and decreased blood flow to the uterus and placenta [46, 47]. Dysfunctional TGF-β may lead to incomplete normal invasion of trophoblasts, causing superficial placentation and impaired uteroplacental perfusion-one of the characteristic features of PE [48]. Also, TGF-β signaling can contribute to the immune cell

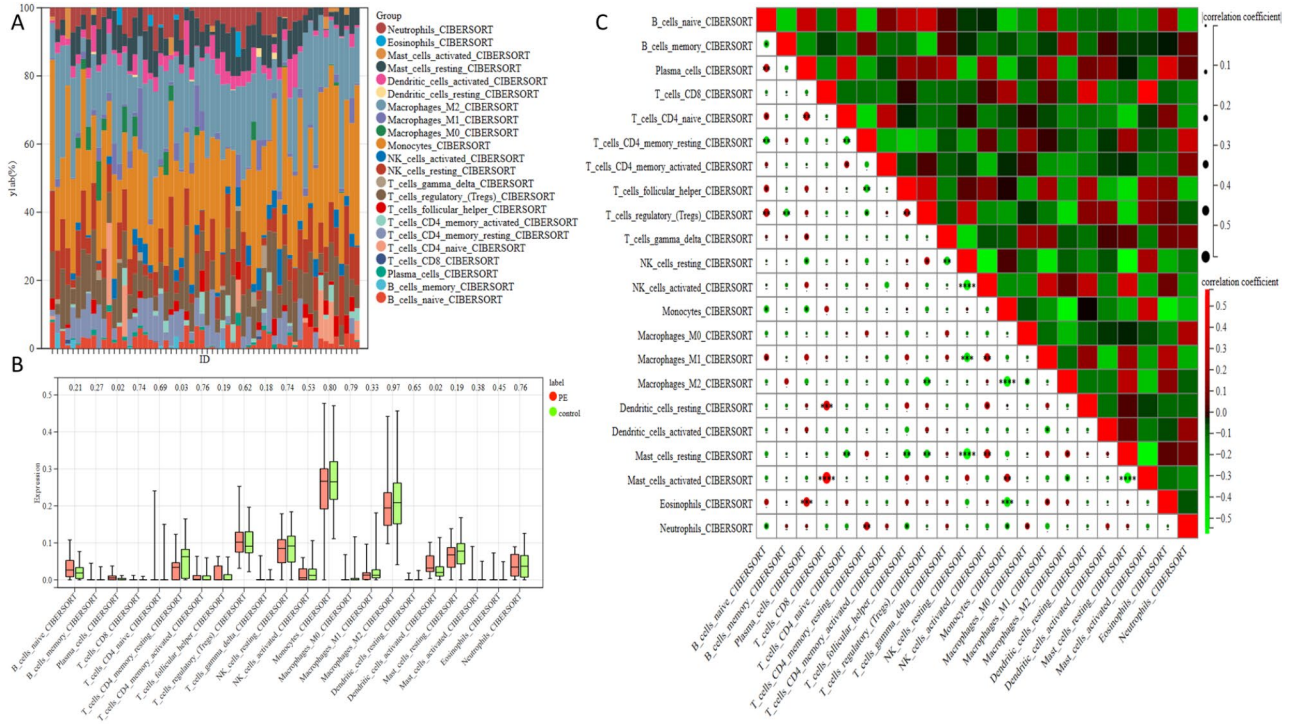

Fig. 9 Immune cell compositions between PE and Control samples. (**A**) The relative percentage of 22 immune cells in each sample. (**B**) Differences in immune infiltration between PE and Control samples. (**C**) Data matrix correlation analysis and its heat map

infiltration characteristic of the pro-inflammatory state in this condition [49]. Therefore, more research is needed to fully understand the role of TGF-β signaling in the development of preeclampsia.

There is some evidence to suggest that gap junctions may play a role in the regulation of blood flow and blood pressure, which are key features of preeclampsia [50, 51]. Gap junctions allow the communication between cells for the exchange of ions, metabolites, and signaling molecules [52]. Any changes in the activity of gap junctions in trophoblasts would perturb placental homeostasis, leading to failure of nutrient and oxygen transport across the placenta to the fetus, and may further contribute to the increased systemic vascular resistance of PE [53]. Such investigations could incorporate therapeutic options with a view to mitigating the progression of PE.

Our study highlights the importance of dysregulated immune cells in the pathogenesis of PE. The results showed that various immune cells were dysregulated in PE patients, which is consistent with previous studies that have suggested that an abnormal immune response is a key contributor to the development of PE [54, 55]. This highlights the need for further studies to investigate the underlying mechanisms of immune cell dysregulation in PE, which could lead to the development of new therapies for this disease. In future studies, it would be important to perform additional bioinformatics analysis and machine learning on larger and more diverse datasets to further refine and validate our results. In addition, functional studies should be performed to gain a deeper understanding of the biological mechanisms underlying the candidate biomarkers. These studies have the potential to significantly improve our understanding of preeclampsia and to lead to new and more effective strategies for its prevention and treatment.

Limitations: First, the sample size of the dataset used in this study was comparatively small, which may limit the generalization of the results. Second, this study relied solely on data from placental tissues, which itself introduces a number of biases. However, while instructive for disease mechanisms, placental tissue does not provide access to systemic biomarkers, which are more readily available for clinical diagnostics, for instance, in maternal blood or urine. Furthermore, most gene expression profiles in the placenta reflect late-stage PE, which limits their applicability for early detection. These findings need to be validated in easily accessible biological specimens for eventual clinical use. Moreover, neither *F13A1* nor *SCCPDH* has been validated as a diagnostic biomarker across populations. Biomarker expression may, in fact, differ according to patient demographics, including ethnicity, age, or body mass index, and co-morbidities such as diabetes or chronic hypertension. These factors could affect the molecular profiles of PE to an extent that the expression of both *F13A1* and *SCCPDH*, or their diagnostic performance, becomes modified. It would, therefore,

be important that their validity is studied across a wide variety of populations and different geographic regions to establish the general applicability of such biomarkers. Stratified analyses may refine their diagnostic utility in specific subgroups.

Conclusion

This work demonstrates the diagnostic value of *F13A1* and *SCCPDH* as new biomarkers for PE disease and depicts their roles in disease pathology. Integration of WGCNA and machine learning introduces a solid framework for biomarker explorations that are deeply related to the insight into molecular mechanisms of PE. Once their limitations are overcome and findings validated in diverse populations, these biomarkers may mark a sea change in early diagnosis and improvement in patient outcomes.

Abbreviations

PE	Preeclampsia
DEGs	Differentially Expressed Genes
WGCNA	Weighted Gene Co-expression Network Analysis
GEO	Gene Expression Omnibus
WB	Western Blotting
F13A1	Coagulation Factor XIII A Chain
SCCPDH	Saccharopine Dehydrogenase
AUC	Area Under the Receiver Operating Characteristic Curve
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
RF	Random Forest
LASSO	Least Absolute Shrinkage and Selection Operator
ROC	Receiver Operating Characteristic
SPSS	Statistical Package for the Social Sciences
CI	Confidence Interval

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12884-024-07028-3.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

The authors would like to express their sincere gratitude to Professor Deyou Zheng at the Einstein School of Medicine for his invaluable support and assistance with bioinformatics analysis and machine learning methods in this study. Professor Zheng's expertise and experience played a crucial role in the design of the experiments and the processing of the data, enabling the authors to gain deeper insights into the phenomena under investigation.

Author contributions

YZ conceptualized the study, conducted the methodology, performed formal analysis, and drafted the original manuscript. ZF was responsible for data curation and contributed to the writing of the original draft. XW contributed to visualization and participated in the investigation. HZ assisted with visualization and took part in reviewing and editing the manuscript. PS (corresponding author) supervised the project, provided resources, secured funding, and contributed to the conceptualization, writing, and review of the manuscript. All authors read and approved the final manuscript.

Funding

Fujian Natural Science Foundation Project, 2022J01425. Fujian Natural Science Foundation Project, 2022J011042. Joint Funds for the innovation of science and Technology, Fujian province, 2023Y939.

Data availability

The PE dataset used in this study (GSE25906) was obtained from the Gene Expression Omnibus (GEO) database. This dataset included gene expression profiles of placental samples from PE patients ($n = 23$) and normal pregnant women ($n = 37$). (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA135777). The authors declare that all data needed to evaluate the conclusions in the paper are present in the supplementary materials. The data can be provided by Yihan Zheng pending scientific review and a completed material transfer agreement. Requests for the data should be submitted to: zyh@fjmu.edu.cn. The original, unprocessed Western blot images, are available in the Supplementary material in this manuscript submission.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the principles outlined in the Declaration of Helsinki. Ethical approval for the collection and use of human tissue samples was obtained from the Ethics Committee of Fujian Maternity and Child Health Hospital (Approval Number: 2022KYLLRD01038). All participants provided informed consent prior to inclusion in the study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 14 September 2024 / Accepted: 2 December 2024
Published online: 21 December 2024

References

1. Jung E, Romero R, Yeo L, Gomez-Lopez N, Chaemsaithong P, Jaovisidha A, et al. The etiology of preeclampsia. Am J Obstet Gynecol. 2022;226(2s):S844–66.
2. Chappell LC, Cluver CA, Kingdom J, Tong S. Pre-eclampsia. Lancet (London England). 2021;398(10297):341–54.
3. Bian X, Biswas A, Huang X, Lee KJ, Li TK, Masuyama H, et al. Short-term prediction of adverse outcomes using the sFlt-1 (Soluble fms-Like Tyrosine kinase 1)/PlGF (placental growth factor) ratio in Asian Women with suspected Preeclampsia. Hypertension. 2019;74(1):164–72.
4. Verlohren S, Droge LA. The diagnostic value of angiogenic and antiangiogenic factors in differential diagnosis of preeclampsia. Am J Obstet Gynecol. 2022;226(2S):S1048–58.
5. Chaemsaithong P, Sahota DS, Poon LC. First trimester preeclampsia screening and prediction. Am J Obstet Gynecol. 2022;226(2S):S1071–Se972.
6. Jairajpuri DS, Almawi WY. MicroRNA expression pattern in pre-eclampsia (Review). Mol Med Rep. 2016;13(3):2351–8.
7. Wang Z, Zhao G, Zeng M, Feng W, Liu J. Overview of extracellular vesicles in the pathogenesis of preeclampsia†. Biol Reprod. 2021;105(1):32–9.
8. Torres-Torres J, Espino YSS, Martinez-Portilla R, Borboa-Olivares H, Estrada-Gutierrez G, Acevedo-Gallegos S et al. A narrative review on the pathophysiology of Preeclampsia. Int J Mol Sci. 2024;25(14).
9. Zhou Y, Shi W, Zhao D, Xiao S, Wang K, Wang J. Identification of Immune-Associated genes in diagnosing aortic valve calcification with metabolic syndrome by Integrated Bioinformatics Analysis and Machine Learning. Front Immunol. 2022;13:937886.
10. Liu K, Fu Q, Liu Y, Wang C. An integrative bioinformatics analysis of microarray data for identifying hub genes as diagnostic biomarkers of preeclampsia. Biosci Rep. 2019;39(9).
11. Wang H, Zhang Z, Li H, Li J, Li H, Liu M, et al. A cost-effective machine learning-based method for preeclampsia risk assessment and driver genes discovery. Cell Bioscience. 2023;13(1):41.
12. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Brief Bioinform. 2023;24(2).

13. Dong B, Liu X, Yu S. Utilizing machine learning algorithms to identify biomarkers associated with diabetic nephropathy: a review. Med (Baltim). 2024;103(8):e37235.
14. Ding Y, Yang X, Han X, Shi M, Sun L, Liu M, et al. Ferroptosis-related gene expression in the pathogenesis of preeclampsia. Front Genet. 2022;13:927869.
15. Dai Y, Sun X, Wang C, Li F, Zhang S, Zhang H, et al. Gene co-expression network analysis reveals key pathways and hub genes in Chinese cabbage (Brassica rapa L.) during vernalization. BMC Genomics. 2021;22(1):236.
16. Rybak-Krzyszkowska M, Staniczek J, Kondracka A, Bogusławska J, Kwiatkowski S, Góra T et al. From biomarkers to the molecular mechanism of Preeclampsia-A Comprehensive Literature Review. Int J Mol Sci. 2023;24(17).
17. Tsai S, Hardison NE, James AH, Motsinger-Reif AA, Bischoff SR, Thames BH, et al. Transcriptional profiling of human placentas from pregnancies complicated by preeclampsia reveals disregulation of sialic acid acetylesterase and immune signalling pathways. Placenta. 2011;32(2):175–82.
18. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
20. The Gene Ontology C. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(D1):D330–8.
21. Shen W, Song Z, Zhong X, Huang M, Shen D, Gao P, et al. Sangerbox: a comprehensive, interaction-friendly clinical bioinformatics analysis platform. iMeta. 2022;1(3):e36.
22. Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. Biomed Eng Online. 2018;17(Suppl 1):131.
23. Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D, Marshall S. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol Meas. 2014;35(11):2191–203.
24. Zhang M, Zhu K, Pu H, Wang Z, Zhao H, Zhang J, et al. An Immune-related signature predicts survival in patients with lung adenocarcinoma. Front Oncol. 2019;9:1314.
25. Alderden J, Pepper GA, Wilson A, Whitney JD, Richardson S, Butcher R, et al. Predicting pressure Injury in critical care patients: a machine-learning model. American journal of critical care: an official publication. Am Association Critical-Care Nurses. 2018;27(6):461–8.
26. Magee LA, Brown MA, Hall DR, Gupte S, Hennessy A, Karumanchi SA, et al. The 2021 International Society for the Study of Hypertension in Pregnancy classification, diagnosis & management recommendations for international practice. Pregnancy Hypertens. 2022;27:148–69.
27. Pan X, Jin X, Wang J, Hu Q, Dai B. Placenta inflammation is closely associated with gestational diabetes mellitus. Am J Translational Res. 2021;13(5):4068–79.
28. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.
29. Hu K. Become competent within one day in Generating boxplots and Violin plots for a novice without prior R experience. Methods Protocols. 2020;3(4).
30. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14(1):91.
31. Tilford CA, Siemers NO. Gene set enrichment analysis. Methods in molecular biology (Clifton, NJ). 2009;563:99–121.
32. Ma S, Chen C, Liang Q, Wu X, Wang X, Wu W, et al. Phenotype and genotype of FXIII deficiency in two unrelated probands: identification of a novel F13A1 large deletion mediated by complex rearrangement. Orphanet J Rare Dis. 2019;14(1):182.
33. Shangguan Y, Wang Y, Shi W, Guo R, Zeng Z, Hu W, et al. Systematic proteomics analysis of lysine acetylation reveals critical features of placental proteins in pregnant women with preeclampsia. J Cell Mol Med. 2021;25(22):10614–26.
34. Epiney M, Ribaux P, Arboit P, Irion O, Cohen M. Comparative analysis of secreted proteins from normal and preeclamptic trophoblastic cells using proteomic approaches. J Proteom. 2012;75(6):1771–7.
35. Thomas JR, Appios A, Zhao X, Dutkiewicz R, Donde M, Lee CYC et al. Phenotypic and functional characterization of first-trimester human placental macrophages, Hofbauer cells. J Exp Med. 2021;218(1).
36. Soendergaard C, Kvist PH, Seidelin JB, Nielsen OH. Tissue-regenerating functions of coagulation factor XIII. J Thromb Haemostasis: JTH. 2013;11(5):806–16.
37. Bazzan E, Casara A, Radu CM, Tinè M, Biondini D, Faccioli E, et al. Macrophages-derived factor XIII links coagulation to inflammation in COPD. Front Immunol. 2023;14:1131292.
38. Science WI. Pathway network for Response to elevated platelet cytosolic Ca2 + SuperPath 2023 [updated 2023 January 09. https://pathcards.genecards.org/card/response_to_elevated_platelet_cytosolic_ca2
39. Tian F, Liu D, Chen J, Liao W, Gong W, Huang R, et al. Proteomic response of rat pituitary under chronic mild stress reveals insights into vulnerability and resistance to anxiety or depression. Front Genet. 2021;12:751999.
40. Bilbul M, Caccese C, Horsley K, Gauvreau A, Gavanski I, Montreuil T, et al. Maternal anxiety, depression and vascular function during pregnancy. J Psychosom Res. 2022;154:110722.
41. Roberts L, Henry A, Harvey SB, Homer CSE, Davis GK. Depression, anxiety and posttraumatic stress disorder six months following preeclampsia and normotensive pregnancy: a P4 study. BMC Pregnancy Childbirth. 2022;22(1):108.
42. Kamrani A, Soltani-Zangbar MS, Shiri S, Yousefzadeh Y, Pourakbari R, Aghebati-Maleki L, et al. TIGIT and CD155 as Immune-Modulator receptor and Ligand on CD4(+) T cells in Preeclampsia patients. Immunol Investig. 2022;51(4):1023–38.
43. Guignabert C, Humbert M. Targeting transforming growth factor-β receptors in pulmonary hypertension. Eur Respir J. 2021;57(2).
44. Goumans MJ, Ten Dijke P. TGF-β signaling in Control of Cardiovascular function. Cold Spring Harb Perspect Biol. 2018;10(2).
45. Meng XM, Nikolic-Paterson DJ, Lan HY. TGF-β: the master regulator of fibrosis. Nat Rev Nephrol. 2016;12(6):325–38.
46. Liu YH, Zheng L, Cheng C, Li SN, Shivappa N, Hebert JR et al. Dietary inflammatory index, inflammation biomarkers and preeclampsia risk: a hospital-based case-control study. Br J Nutr. 2022;18:1–9.
47. Xu XH, Jia Y, Zhou X, Xie D, Huang X, Jia L, et al. Downregulation of lysyl oxidase and lysyl oxidase-like protein 2 suppressed the migration and invasion of trophoblasts by activating the TGF-β/collagen pathway in preeclampsia. Exp Mol Med. 2019;51(2):1–12.
48. Okae H, Toh H, Sato T, Hiura H, Takahashi S, Shirane K, et al. Derivation Hum Trophoblast Stem Cells Cell stem cell. 2018;22(1):50–e636.
49. Yi M, Wu Y, Niu M, Zhu S, Zhang J, Yan Y et al. Anti-TGF-β/PD-L1 bispecific antibody promotes T cell infiltration and exhibits enhanced antitumor activity in triple-negative breast cancer. J Immunother Cancer. 2022;10(12).
50. Lall VK, Bruce G, Voytenko L, Drinkhill M, Wellershaus K, Willecke K, et al. Physiologic regulation of heart rate and blood pressure involves connexin 36-containing gap junctions. FASEB Journal: Official Publication Federation Am Soc Experimental Biology. 2017;31(9):3966–77.
51. Ni X, Li XZ, Fan ZR, Wang A, Zhang HC, Zhang L, et al. Increased expression and functionality of the gap junction in peripheral blood lymphocytes is associated with hypertension-mediated inflammation in spontaneously hypertensive rats. Cell Mol Biol Lett. 2018;23:40.
52. Wei CJ, Xu X, Lo CW. Connexins and cell signaling in development and disease. Annu Rev Cell Dev Biol. 2004;20:811–38.
53. Winterhager E, Kidder GM. Gap junction connexins in female reproductive organs: implications for women's reproductive health. Hum Reprod Update. 2015;21(3):340–52.
54. Bu C, Wang Z, Ren Y, Chen D, Jiang SW. Syncytin-1 nonfusogenic activities modulate inflammation and contribute to preeclampsia pathogenesis. Cell Mol Life Sci. 2022;79(6):290.
55. Luo F, Yue J, Li L, Mei J, Liu X, Huang Y. Narrative review of the relationship between the maternal-fetal interface immune tolerance and the onset of preeclampsia. Annals Translational Med. 2022;10(12):713.

Publisher's note