DOI: 10.16781/j.CN31-2187/R.20240049

·论　著·

基于加权基因共表达网络分析挖掘子痫前期的诊断标志物

姚瑞倩[1,2]，喻　东[3,4*]，薛　赓[2*]
1. 上海理工大学健康科学与工程学院，上海 200093
2. 海军军医大学（第二军医大学）基础医学院医学遗传学教研室，上海 200433
3. 海军军医大学（第二军医大学）转化医学研究中心精准医学教研室，上海 200433
4. 上海市细胞工程重点实验室，上海 200433

［摘要］ **目的**　通过生物信息学分析和机器学习模型挖掘公共数据库中的有效信息，识别子痫前期相关的候选基因，以提高子痫前期早期诊断的准确性并为发病机制和诊疗研究提供靶点。**方法**　从基因表达综合数据库中检索子痫前期患者和正常孕妇胎盘组织样本的 RNA-seq 数据集，利用生物信息分析工具完成数据下载、质量控制、比对及定量后获得基因表达矩阵。采用 DESeq2 1.38.3 工具筛选差异表达基因，通过基因本体和京都基因与基因组百科全书数据库确定富集通路，利用加权基因共表达网络分析（WGCNA）构建共表达网络，利用随机森林算法建立机器学习预测模型。**结果**　4 个数据集 156 例孕妇（70 例子痫前期患者、86 例正常孕妇）胎盘组织样本共筛选出 49 个共有差异表达基因，这些基因显著富集在细胞外区域、卵泡刺激素分泌的正向调节通路、激素活性通路及细胞因子－细胞因子受体相互作用等信号通路。通过 WGCNA 将 49 个差异表达基因分为 7 个共表达模块，鉴定出与子痫前期高度相关的关键模块，并筛选出 6 个候选关键基因，分别为 fms 相关受体酪氨酸激酶 1（*FLT1*）、冠毛素 2（*PAPPA2*）、蛋白磷酸酶 1 调节抑制因子亚基 1C（*PPP1R1C*）、肌球蛋白Ⅶ B（*MYO7B*）、长基因间非蛋白编码 RNA 2009（*LINC02009*）和抑制素亚基 α（*INHA*）。基于这 6 个关键基因构建的随机森林模型对子痫前期有较好的预测价值（AUC＝0.978）。**结论**　子痫前期可能与激素分泌、免疫反应、血管生成因子、妊娠相关血浆蛋白、抑制素等有关，相关基因或可成为子痫前期诊断的候选标志物。

［关键词］ 子痫前期；生物标志物；加权基因共表达网络分析；随机森林模型
［引用本文］ 姚瑞倩，喻东，薛赓．基于加权基因共表达网络分析挖掘子痫前期的诊断标志物［J］．海军军医大学学报，2024，45（12）：1529-1539．DOI: 10.16781/j.CN31-2187/R.20240049.

Mining diagnostic markers of preeclampsia based on weighted gene co-expression network analysis

YAO Ruiqian[1,2], YU Dong[3,4*], XUE Geng[2*]
1. School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
2. Department of Medical Genetics, College of Basic Medical Sciences, Naval Medical University (Second Military Medical University), Shanghai 200433, China
3. Department of Precision Medicine, Center of Translational Medicine, Naval Medical University (Second Military Medical University), Shanghai 200433, China
4. Shanghai Key Laboratory of Cell Engineering, Shanghai 200433, China

［**Abstract**］ **Objective**　To mine valid information in public databases through bioinformatics analysis and machine learning models and to identify candidate genes related to preeclampsia, so as to improve the accuracy of early diagnosis and provide targets for pathogenesis, diagnosis and treatment research. **Methods**　The RNA-seq datasets of placental tissue samples of preeclampsia patients and healthy pregnant women were retrieved from the Gene Expression Omnibus, and the gene expression matrix was obtained after data download, quality control, comparison and quantification through bioinformation analysis. The differentially expressed genes were screened by DESeq2 1.38.3, the enrichment pathway was determined using Gene Ontology and Kyoto Encyclopedia of Genes and Genomes, the co-expression network was constructed using weighted gene co-expression network analysis (WGCNA), and the machine learning prediction model was established by random forest algorithm. **Results**　A total of 49 common differentially expressed genes were screened from placental tissue samples of 156 pregnant women (70 preeclampsia patients and 86 healthy pregnant women) in 4 datasets and they were significantly enriched in extracellular regions, positive regulation pathway of follicle-stimulating hormone secretion, hormone activity pathway, and

［收稿日期］ 2024-01-19　［接受日期］ 2024-08-26
［基金项目］ 国家自然科学基金面上项目（81971402）. Supported by General Program of National Natural Science Foundation of China (81971402).
［作者简介］ 姚瑞倩，硕士生．E-mail: yaorq999@163.com
*通信作者（Corresponding authors）. Tel: 021-81871659, E-mail: yudong615@126.com; Tel: 021-81871054, E-mail: xg_smmu@hotmail.com

cytokine-cytokine receptor interaction pathway, etc. The 49 differentially expressed genes were categorized into 7 co-expression modules by WGCNA, and key modules highly related to preeclampsia were identified. Six candidate key genes (fms related receptor tyrosine kinase 1 [*FLT1*], pappalysin 2 [*PAPPA2*], protein phosphatase 1 regulatory inhibitor subunit 1C [*PPP1R1C*], myosin Ⅶ B [*MYO7B*], long intergenic non-protein coding RNA 2009 [*LINC02009*], and inhibin subunit α [*INHA*]) were screened. The random forest model based on these 6 key genes had good predictive value for preeclampsia (area under curve was 0.978). **Conclusion**　Preeclampsia may be associated with genes for hormone secretion, immune response, angiogenic factors, pregnancy-associated plasma proteins, and inhibin, and these genes may be candidate diagnostic markers of preeclampsia.

［**Key words**］ preeclampsia; biomarkers; weighted gene co-expression network analysis; random forest model

［**Citation**］ YAO R, YU D, XUE G. Mining diagnostic markers of preeclampsia based on weighted gene co-expression network analysis[J]. Acad J Naval Med Univ, 2024, 45(12): 1529-1539. DOI: 10.16781/j.CN31-2187/R.20240049.

子痫前期（preeclampsia）是妊娠期特有的一种多系统进展性疾病，其特点是妊娠 20 周以后出现新发高血压和蛋白尿，或出现新发高血压和终末器官功能障碍伴或不伴蛋白尿，占所有妊娠的 2%～8%[1]。在临床实践中，子痫前期通常分为早发型（妊娠 34 周内）和晚发型（妊娠 34 周后）、轻度和重度（基于血压、临床表现和蛋白尿程度）[2]。研究证实从妊娠早期开始使用阿司匹林可以降低子痫前期的患病率[3-4]。然而，目前对阿司匹林用药的适宜人群、开始和结束用药的时机及剂量等仍在不断探索中，阿司匹林在临床研究中的应用并未如预期可明显降低子痫前期的发病率，其用于预防子痫前期仍存在一定的局限性[5]，当前唯一有效治疗子痫前期的方法是终止妊娠。因此对子痫前期的早期预测和诊断极其重要，了解子痫前期发生和发展的分子机制可能会改善治疗现状。

由于患有子痫前期的孕妇症状通常在分娩后缓解，且分娩后可以检测到胎盘的组织病理学变化，因此胎盘功能不全一直被认为是导致子痫前期的根本原因，胎盘标志物对预测子痫前期可能具有特异性和灵敏性。由于样本量小或数据分析不充分，许多芯片研究未能确定独特的胎盘分子标志物。转录组测序（RNA-seq）可对包括编码和非编码转录本在内的转录组进行全测序，对探索疾病机制和生物标志物有参考价值[6]。Kaartokallio 等[7]针对子痫前期胎盘的 RNA-seq 数据进行了分析，观察到子痫前期患者的胎盘存在血管功能和免疫平衡紊乱，并鉴定出了一些或许可以预测和诊断子痫前期的差异表达基因。Ren 等[8]通过 RNA-seq 发现早发型和晚发型重度子痫前期的分子机制不同，晚发型轻度子痫前期可能没有胎盘特异性致病因素。这些研究表明分析子痫前期胎盘组织的转录组数据将有助于发现子痫前期发展的分子机制，以及筛选出预测、诊断子痫前期的标志物。

随着各种生物信息学工具和公共数据库的出现，研究人员利用生物信息学方法能够高效且经济地从高通量测序数据中挖掘出疾病的致病基因及有潜力的诊断靶标。一般的生物信息学方法在处理高维数据时往往存在困难，因为高维数据之间的相关性和交互作用复杂且难以解释。加权基因共表达网络分析（weighted gene co-expression network analysis，WGCNA）是一种描述大样本基因间强关联性的高级生物信息学方法[9]。WGCNA 的独特优势在于能将基因表达数据转化为共表达模块，从而深入了解可能导致相关表型特征的信号网络[10]。WGCNA 可用于寻找与疾病高度相关的基因模块，并能够将模块的特征基因或关键基因与样本的表型特征联系起来。该方法被广泛应用于各种疾病研究，对鉴定候选生物标志物或治疗靶点有很大帮助[11-13]。另外机器学习算法能够从新视角分析大群体基因测序或微阵列数据，随机森林算法的优势在于其分类表现较好，抗过拟合能力较强，具有较好的鲁棒性，对噪声和异常值有较好的容忍性[14]。来自美国国家生物技术信息中心（National Center for Biotechnology Information，NCBI）的基因表达综合数据库（Gene Expression Omnibus，GEO）是当今最大、最全面的公共测序数据资源，包含了多种疾病的多组学高通量测序的原始数据，提供了大量的数据资源。

本研究整合了来自 NCBI GEO 的 4 个 RNA-seq 数据集（每个数据集都包含子痫前期孕妇的胎盘组织样本和正常孕妇的胎盘组织样本），对比鉴定了 4 个数据集共有的差异表达基因并对这些差异表达基因进行了功能富集，然后通过 WGCNA 构建了子

痫前期的共表达网络，鉴定出与子痫前期高度相关的关键基因模块，而且这些关键基因在随机森林预测模型中得到验证。

1　材料和方法

1.1　数据集选择与预处理　从 NCBI GEO 中检索获得子痫前期患者的胎盘组织 RNA-seq 数据集，分别为 GSE114691、GSE186257、GSE148241 和 GSE218039。使用 SRA-Toolkit 软件下载原始数据 SRA 文件，通过 Fastq-dump 命令获得 FASTQ 文件。使用 FastQC 软件对原始数据进行质量评估，并通过 trim_galore 软件进行质量控制。然后利用 Hisat2 软件将质控后的序列与人类基因组序列（GRCh38）进行比对，使用 FeatureCounts 软件对基因表达进行定量，最终获得基因表达矩阵。

1.2　统计学处理　使用 R 4.2.3 软件对测序数据进行统计分析。使用 R 4.2.3 软件中 DESeq2 1.38.3 包鉴定子痫前期患者与正常孕妇胎盘组织中的差异表达基因，$|\log_2(FC)|\geqslant 1$（其中 FC 为差异倍数）且校准 $P<0.05$ 被认为是差异表达的基因。使用 R 4.2.3 软件 randomForest 4.7-1.1 包建立预测子痫前期的随机森林模型，训练集与测试集分别为每个数据集中 70% 和 30% 的数据，通过 Predict 函数计算模型的预测概率分值，并使用 R 4.2.3 软件 pROC 1.18.2 包绘制模型的 ROC 曲线，计算 AUC 值以衡量模型的预测性能。

1.3　基因富集分析　利用注释、可视化和集成发现数据库（Database for Annotation，Visualization，and Integrated Discovery；DAVID）（*https://david.ncifcrf.gov/*）对差异表达基因进行基因本体（Gene Ontology，GO）分析，确定基因富集的生物学过程、分子功能和细胞组分[15-16]。在 KOBAS 3.0 网站（*http://kobas.cbi.pku.edu.cn/*）进行京都基因与基因组百科全书（Kyoto Encyclopedia of Genes and Genomes，KEGG）通路富集分析[16-17]。通过 R 4.2.3 软件 ggplot2 包将 $P<0.05$ 的通路可视化于气泡图中。

1.4　WGCNA 共表达网络分析　基因表达数据通过方差稳定变换（variance stabilizing transformation，VST）及对数转换后，利用 R 4.2.3 软件 WGCNA 1.72-1 包构建共表达网络。具有相似表达模式的差异表达基因被归入一个模块，每个模块被赋予一种颜色。通过计算基因的模块隶属度（module membership，MM）和基因显著性（gene significance，GS）识别与临床表型相关的关键基因[18]。然后利用 modulePreservation 函数计算保守性 Z-summary 得分，以剔除保守性较差的模块。最后使用 Cytoscape 3.10.0 软件对关键基因网络进行可视化。

2　结　果

2.1　研究队列及基因表达分析　通过检索 NCBI GEO 共收集到 4 个与子痫前期相关的数据集，其中 2 个数据集明确了样本的亚型分别为重度子痫前期和早发型重度子痫前期（表 1）。4 个数据集共包括 156 例孕妇胎盘组织样本的 RNA-seq 数据，其中 70 例为子痫前期患者样本、86 例为正常孕妇样本。主成分分析结果显示，子痫前期患者与正常孕妇的胎盘组织样本可分为 2 个明显的聚类，说明子痫前期患者与正常孕妇胎盘组织间的基因表达差异显著（图 1A）；而且在所有数据集中显著上调的差异表达基因数量均多于显著下调的差异表达基因（图 1B）。

表 1　4 个子痫前期相关数据集的样本信息

Tab 1　Basic information of 4 preeclampsia-related datasets

Item	GSE114691	GSE148241	GSE186257	GSE218039
Assay type	RNA-seq	RNA-seq	RNA-seq	RNA-seq
Tissue source	Placenta	Placenta	Placenta	Placenta
Sample size, *n*				
Total	41	41	44	30
Preeclampsia	20	9	26	15
Control	21	32	18	15
Subtype of preeclampsia	—	Early-onset severe preeclampsia	Severe preeclampsia	—
Total read count	47 401	50 134	54 249	49 779

“—”: Not mentioned.

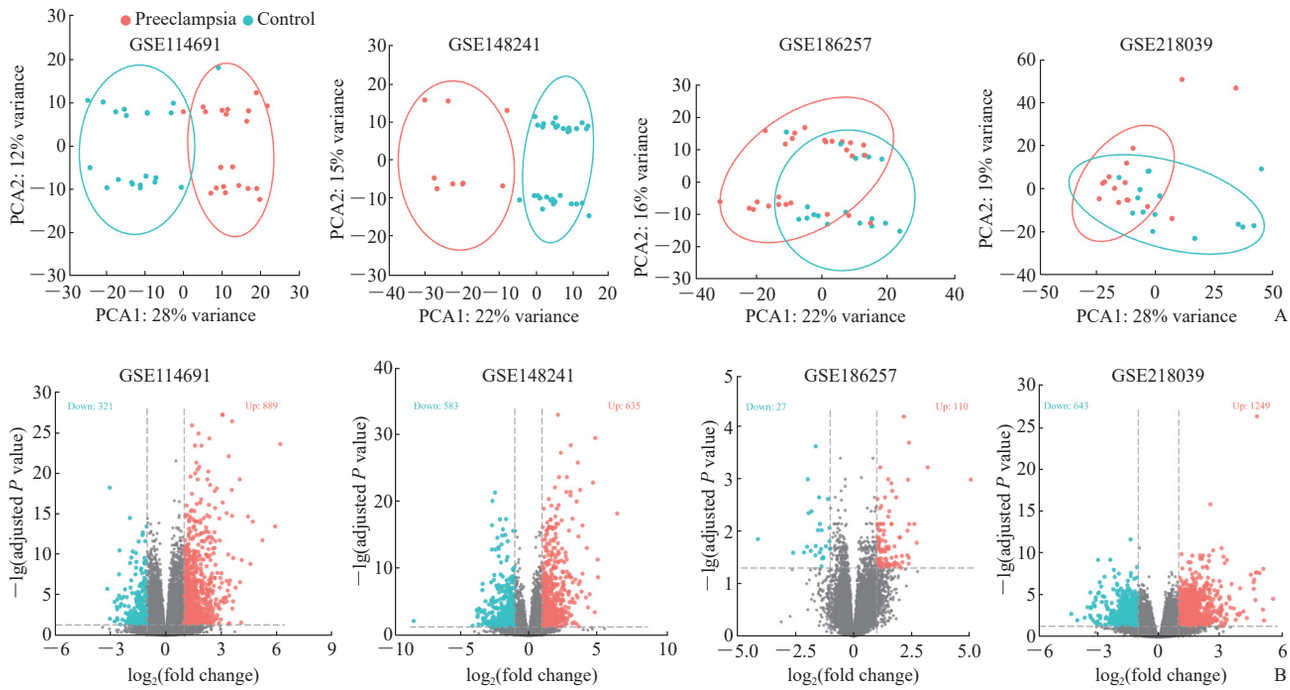

图 1　4 个数据集的主成分分析（A）和差异表达基因的火山图（B）

Fig 1　Principal component analysis (PCA) (A) and volcano plot of differentially expressed genes (B) of 4 datasets

2.2　*差异表达基因的通路富集分析*　有 49 个差异表达基因是 4 个数据集共有的，其中 44 个表达上调、5 个表达下调（表 2、图 2A）。GO 功能分析结果显示，在生物学过程中这些差异表达基因显著富集在卵泡刺激素分泌的正向调节通路；在细胞组分中，有 10 个基因［蛋白 C 受体（protein C receptor，*PROCR*）、嗅觉调节素样蛋白 3（olfactomedin like 3，*OLFML3*）、瘦素（leptin，*LEP*）、谷氨酰胺酰肽环转移酶（glutaminyl-peptide cyclotransferase，*QPCT*）、黄体生成素亚基β（luteinizing hormone subunit β，*LHB*）、HtrA 丝氨酸肽酶 4（HtrA serine peptidase 4，*HTRA4*）、抑制素亚基α（inhibin subunit α，*INHA*）、卵泡抑素样蛋白 3（follistatin like 3，*FSTL3*）、冠毛素 2（pappalysin 2，*PAPPA2*）和糖基磷脂酰肌醇锚定高密度脂蛋白结合蛋白 1（glycosylphosphatidylinositol anchored high density lipoprotein binding protein 1，*GPIHBP1*）］富集在细胞外区域，除 *OLFML3* 外其余 9 个基因均表达上调；在分子功能中，差异表达基因显著富集在激素活性通路（图 2B）。KEGG 通路富集分析结果显示，最显著的是细胞因子－细胞因子受体相互作用通路，C-X3-C 基序趋化因子受体 1（C-X3-C motif chemokine receptor 1，*CX3CR1*）、*LEP* 和 *INHA* 均富集在该通路上（图 2C）。这些结果说明，在子痫前期患者与正常孕妇胎盘组织之间的差异表达基因可能与激素分泌、免疫应答与免疫调节等有关。

2.3　*加权基因共表达网络的建立*　为进一步探讨这 49 个差异表达基因与疾病表型的相关性，选择疾病样本量最大的数据集 GSE186257 进行 WGCNA，挖掘可能与子痫前期发生和发展密切相关的共表达模块。共构建了 7 个共表达模块，这些模块均独立于其他模块，可见棕色和蓝色模块与胎儿性别相关（均 $P=0.01$，图 3A），并且在这 2 个模块中 *INHA*、长基因间非蛋白编码 RNA 2009（long intergenic non-protein coding RNA 2009，*LINC02009*）和 MIR31 宿主基因（MIR31 host gene，*MIR31HG*）是与胎儿性别相关的重要基因（均 GS＞0.5，图 3B）。然而，这 7 个共表达模块与患有子痫前期的孕妇是否早产并不相关（图 3A）。此外，红色模块和蓝色模块之间具有强相关性（图 3C），并且这 2 个模块内的 8 个基因［fms 相关受体酪氨酸激酶 1（fms related receptor tyrosine kinase 1，*FLT1*）、*INHA*、肌球蛋白Ⅶ B（myosin Ⅶ B，*MYO7B*）、蛋白磷酸酶 1 调节抑制因子亚基 1C（protein phosphatase 1 regulatory inhibitor subunit 1C，*PPP1R1C*）、*LINC02009*、*PAPPA2*、*FSTL3* 和 *LEP*］与这 2 个模块高度相关（均 MM＞0.8，图 3D）。

表 2　4 个子痫前期相关数据集共有的 49 个差异表达基因的表达趋势及差异倍数

Tab 2　Expression trends and fold change of 49 common differentially expressed genes in 4 preeclampsia-related datasets

Gene ID|name	Trend	log_2(fold change)			
		GSE114691	GSE148241	GSE186257	GSE218039
ENSG00000054179|*ENTPD2*	Up	1.59	2.57	2.37	1.96
ENSG00000061656|*SPAG4*	Up	2.25	2.38	1.08	2.05
ENSG00000070404|*FSTL3*	Up	3.60	4.89	2.00	1.69
ENSG00000101000|*PROCR*	Up	1.73	1.97	1.08	1.33
ENSG00000102755|*FLT1*	Up	3.08	3.62	1.63	2.69
ENSG00000104826|*LHB*	Up	1.75	1.80	1.25	1.75
ENSG00000105205|*CLC*	Up	1.25	1.34	1.14	1.95
ENSG00000115828|*QPCT*	Up	1.95	2.26	1.66	1.17
ENSG00000116183|*PAPPA2*	Up	2.42	2.51	1.08	2.07
ENSG00000123999|*INHA*	Up	2.40	1.29	1.09	1.93
ENSG00000130822|*PNCK*	Up	2.66	2.78	1.42	1.68
ENSG00000148488|*ST8SIA6*	Up	1.82	1.22	1.02	1.19
ENSG00000149256|*TENM4*	Up	1.60	1.88	1.18	1.24
ENSG00000150722|*PPP1R1C*	Up	2.03	3.09	1.49	2.56
ENSG00000162753|*SLC9C2*	Up	1.46	2.44	1.41	1.84
ENSG00000165810|*BTNL9*	Up	3.59	2.56	1.54	2.30
ENSG00000168490|*PHYHIP*	Up	1.87	1.71	1.32	1.25
ENSG00000169495|*HTRA4*	Up	3.58	4.73	2.34	2.48
ENSG00000169994|*MYO7B*	Up	2.37	3.08	1.28	2.28
ENSG00000171124|*FUT3*	Up	3.43	2.43	2.71	2.78
ENSG00000171889|*MIR31HG*	Up	1.41	2.07	1.03	3.27
ENSG00000174697|*LEP*	Up	4.59	6.50	2.61	5.17
ENSG00000177628|*GBA1*	Up	1.42	1.69	1.03	1.10
ENSG00000186806|*VSIG10L*	Up	1.04	1.93	1.04	1.86
ENSG00000214946|*TBC1D26*	Up	2.68	3.53	1.43	2.10
ENSG00000237949|*LOC102724768*	Up	2.21	1.97	1.45	1.95
ENSG00000247095|*MIR210HG*	Up	2.38	2.82	1.09	1.88
ENSG00000277494|*GPIHBP1*	Up	2.19	1.68	1.53	1.85
ENSG00000283646|*LINC02009*	Up	4.01	3.53	1.78	4.56
ENSG00000287315|*LOC101927401*	Up	1.72	1.47	1.07	2.27
ENSG00000224658	Up	6.20	3.79	2.38	2.92
ENSG00000226022	Up	1.49	2.28	1.83	1.56
ENSG00000233002	Up	3.30	3.96	1.54	1.89
ENSG00000237514	Up	2.19	2.74	1.20	2.28
ENSG00000241219	Up	2.22	1.73	1.12	1.04
ENSG00000254592	Up	1.13	1.72	1.08	1.67
ENSG00000255104	Up	2.41	3.45	3.18	3.04
ENSG00000279393	Up	2.53	2.98	1.17	1.84
ENSG00000284309	Up	1.92	1.84	1.02	1.58
ENSG00000285424	Up	1.58	1.99	1.35	1.28
ENSG00000286754	Up	2.29	2.16	1.10	3.11
ENSG00000287233	Up	2.08	1.09	1.37	1.74
ENSG00000290531	Up	2.21	1.21	1.09	1.93
ENSG00000291137	Up	1.01	1.50	1.15	1.41
ENSG00000116774|*OLFML3*	Down	−1.19	−1.01	−1.10	−1.49
ENSG00000117090|*SLAMF1*	Down	−3.02	−2.00	−1.99	−2.98
ENSG00000162706|*CADM3*	Down	−1.31	−2.01	−1.18	−1.71
ENSG00000168329|*CX3CR1*	Down	−1.14	−1.43	−1.47	−1.66
ENSG00000233725|*NRAD1*	Down	−1.17	−2.02	−1.80	−2.67

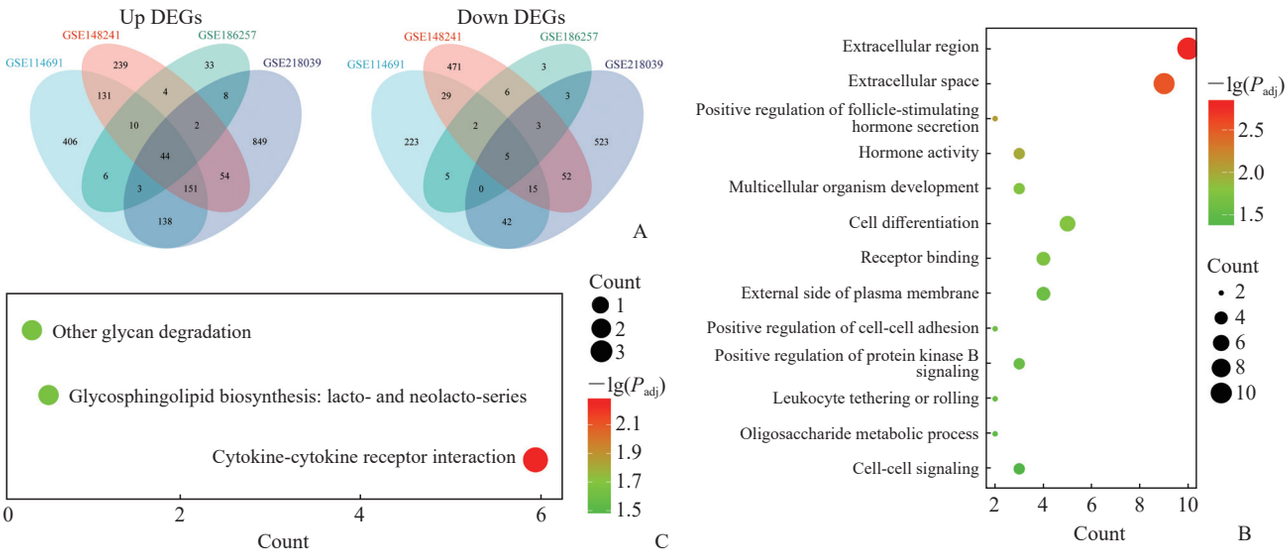

图 2　4 个子痫前期相关数据集中 DEG 的维恩图（A）及 GO（B）、KEGG（C）富集分析结果

Fig 2　Venn diagram (A) and GO (B) and KEGG (C) enrichment analysis results of DEGs in 4 preeclampsia-related datasets

DEG: Differentially expressed gene; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; P_{adj}: Adjusted P value.

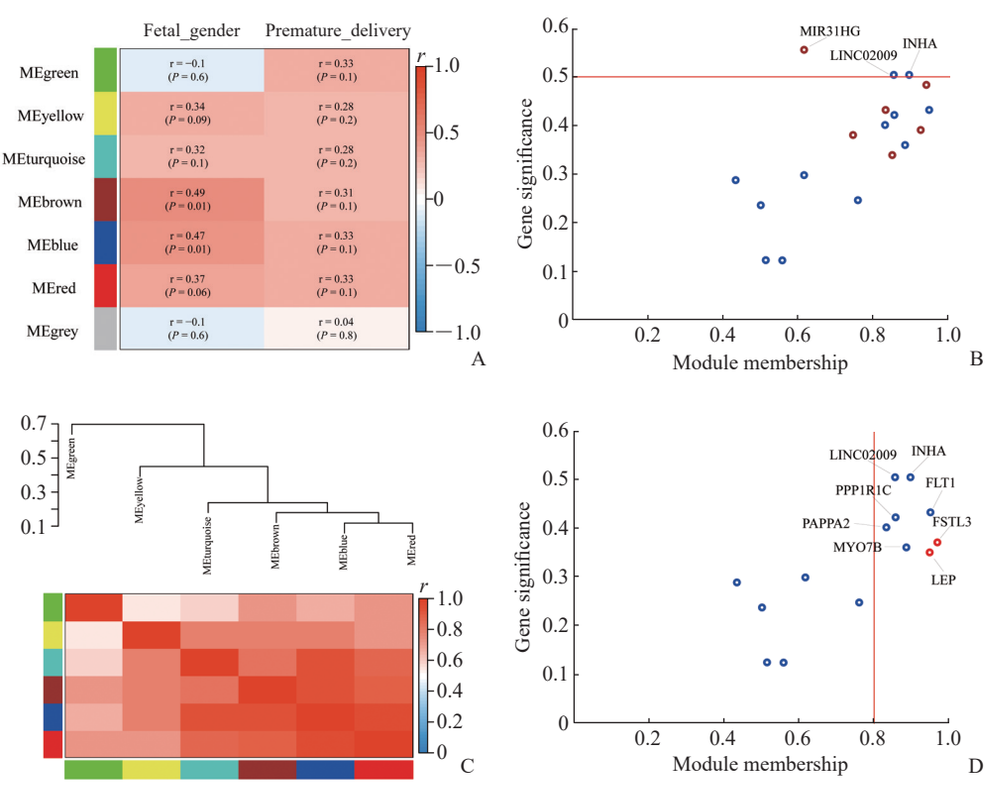

图 3　子痫前期相关数据集 GSE186257 加权基因共表达网络的建立

Fig 3　Establishment of a weighted gene co-expression network for preeclampsia-related dataset GSE186257

A: Correlation analysis of the modules with fetal gender and premature delivery; B: A scatterplot of module membership and gene significance in the blue and brown modules, with the color of the points corresponding to the color of the module to which they belong; C: Module-eigengene (ME) adjacency heatmap; D: A scatterplot of module membership and gene significance in the blue and red modules, with the color of the points corresponding to the color of the module to which they belong. ME is defined as the first principal component of a co-expression module matrix. INHA: Inhibin subunit α; LINC02009: Long intergenic non-protein coding RNA 2009; MIR31HG: MIR31 host gene; FLT1: Fms related receptor tyrosine kinase 1; MYO7B: Myosin Ⅶ B; PPP1R1C: Protein phosphatase 1 regulatory inhibitor subunit 1C; PAPPA2: Pappalysin 2; FSTL3: Follistatin like 3; LEP: Leptin.

2.4　关键基因模块的验证　为了评估 GSE186257 网络模块在其他数据集上的表达模式，使用 modulePreservation 函数计算模块的保守性 Z-summary 得分（图 4A）。蓝色模块在其他 3 个数据集（GSE114691、GSE148241、GSE218039）均得分较高，表明该基因模块在不同数据集中均表现稳定。然而，Z-summary 得分较低的红色和绿色模块在不同数据集中保守性较低，表明这 2 个基因模块可能能够较好地区分子痫前期的不同亚型。基于以上结果，推测蓝色模块是与子痫前期高度相关的关键模块。蓝色模块内的 12 个基因中有 4 个基因（*FLT1*、*PAPPA2*、*INHA* 和 *MYO7B*）表达量相对较高（图 4B），并且网络拓扑图显示这 4 个基因之间存在强相关性（图 4C），提示 *FLT1*、*PAPPA2*、*INHA* 和 *MYO7B* 这 4 个基因可能是子痫前期网络中的关键基因。

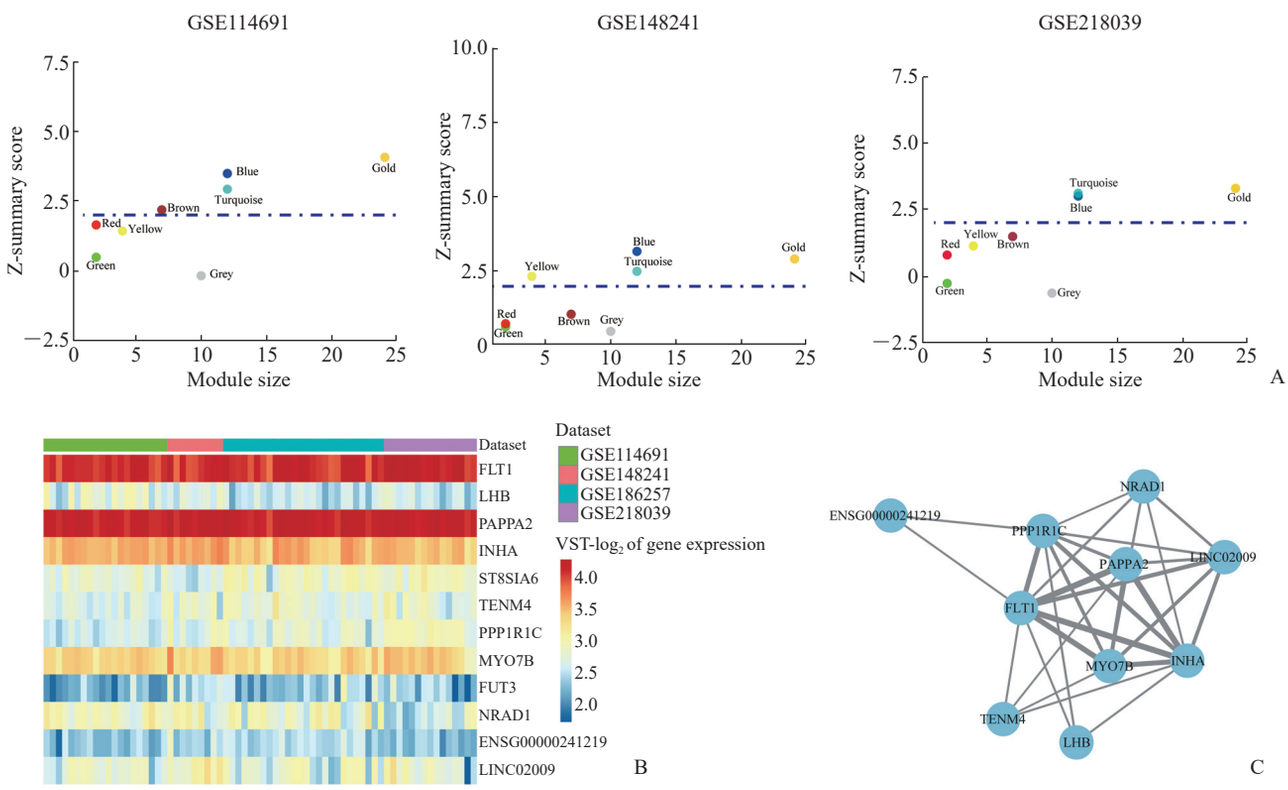

图 4　数据集 GSE186257 的共表达模块在数据集 GSE114691、GSE148241 和 GSE218039 中的验证

Fig 4　Validation of gene co-expression modules of GSE186257 in GSE114691, GSE148241, and GSE218039

A: The Z-summary score of the co-expression modules of preeclampsia-related dataset GSE186257 in the other 3 datasets (GSE114691, GSE148241, and GSE218039) (below the blue line represents no conservatism, i.e. Z-summary score＜2); B: The heatmap of differentially expressed genes in the blue module; C: Visual network of blue module (the thickness of the line represents the weight of the network between genes, with thicker lines indicating stronger relationships between genes on both sides). VST: Variance stabilizing transformation; FLT1: Fms related receptor tyrosine kinase 1; LHB: Luteinizing hormone subunit β; PAPPA2: Pappalysin 2; INHA: Inhibin subunit α; ST8SIA6: ST8 α-N-acetyl-neuraminide α-2,8-sialyltransferase 6; TENM4: Teneurin transmembrane protein 4; PPP1R1C: Protein phosphatase 1 regulatory inhibitor subunit 1C; MYO7B: Myosin Ⅶ B; FUT3: Fucosyltransferase 3; NRAD1: Non-coding RNA in aldehyde dehydrogenase 1A pathway; LINC02009: Long intergenic non-protein coding RNA 2009.

2.5　关键基因预测子痫前期模型的建立　在基因模块中，与模块高度相关的同时与性状也高度相关的基因被定义为该模块的关键基因。通过绘制 MM 与 GS 散点图筛选出蓝色模块的 6 个关键基因，分别为 *FLT1*、*PAPPA2*、*PPP1R1C*、*MYO7B*、*LINC02009* 和 *INHA*（MM＞0.8、GS＞0.3，图 5A）。将这 6 个基因纳入随机森林模型，ROC 曲线分析结果显示，在测试集中，所建立的模型对子痫前期有较好的预测价值（AUC＝0.978，图 5B）。以上结果表明，基于 *FLT1*、*PAPPA2*、*PPP1R1C*、*MYO7B*、*LINC02009* 和 *INHA* 这 6 个基因所建立的预测模型有助于诊断子痫前期。

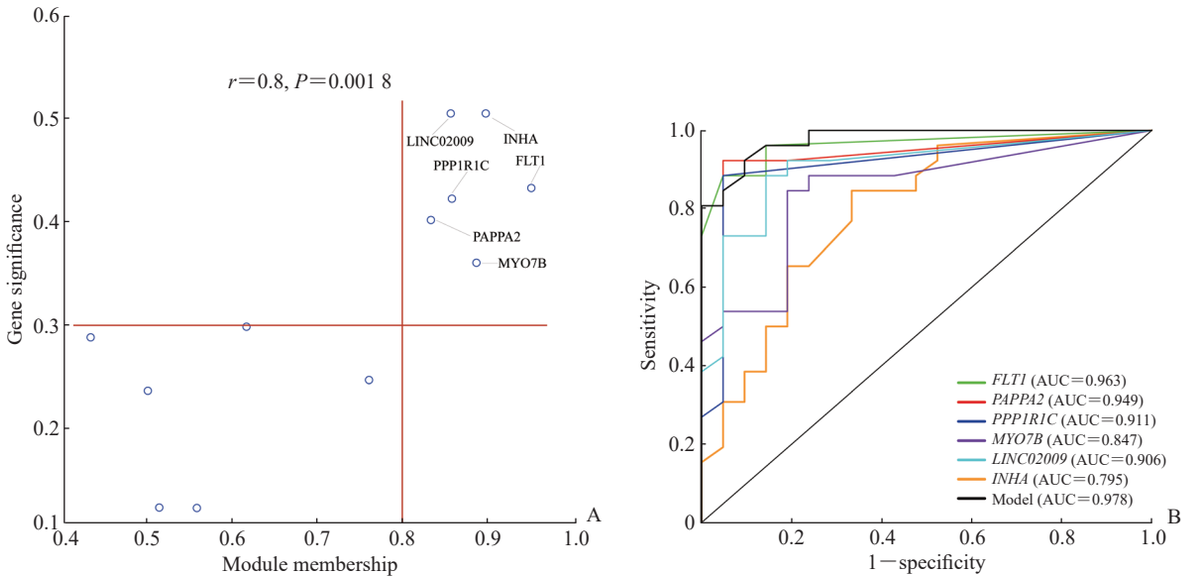

图 5　子痫前期关键基因的筛选及随机森林模型的预测效能评价

Fig 5　Screening of key genes related to preeclampsia and prediction efficiency evaluation of random forest model

A: A scatterplot of module membership and gene significance in the blue module; B: ROC curves of *FLT1*, *PAPPA2*, *PPP1R1C*, *MYO7B*, *LINC02009*, *INHA*, and the random forest model in predicting preeclampsia. FLT1: Fms related receptor tyrosine kinase 1; PAPPA2: Pappalysin 2; PPP1R1C: Protein phosphatase 1 regulatory inhibitor subunit 1C; MYO7B: Myosin Ⅶ B; LINC02009: Long intergenic non-protein coding RNA 2009; INHA: Inhibin subunit α; ROC: Receiver operating characteristic; AUC: Area under curve.

3　讨　论

子痫前期是导致孕产妇和胎儿发病与死亡的主要原因之一[19]，当前唯一有效治疗子痫前期的方法是终止妊娠，因此开发早期预测及诊断子痫前期的标志物至关重要[20]。胎盘功能不全被认为是导致孕妇患子痫前期的根本原因，然而源自胎盘的分子机制在很大程度上仍不为人所知。RNA-seq 技术的发展使许多基因的表达得以测量，通过数据挖掘方法确定正常组织和患者组织之间的差异表达基因有助于了解疾病的发病机制[21]。然而，在目前的许多研究中数据分散及数据量较小的问题普遍存在。因此，本研究整合并深度挖掘了 4 个 RNA-seq 公共数据集，获得了更全面的生物学信息，有助于理解子痫前期的发生、发展机制。一般的数据挖掘方法缺乏对大规模的高维数据的系统性分析，WGCNA 通过基因之间的相关系数构建分层聚类树，根据聚类树的不同分支将大量基因分为不同的基因模块，从而评估基因模块与临床特征间的关联，该方法在筛选疾病特征标志物和潜在靶点中表现出较其他分析方法更明显的优势[22]。机器学习算法在数据挖掘中的应用也为探究疾病潜在的治疗靶点提供了支撑[23]。本研究通过结合 WGCNA 与机器学习算法挖掘潜在的子痫前期诊断标志物，并证实这些基因对子痫前期的诊断价值，同时证明本研究所使用挖掘方法的可靠性。

本研究鉴定出的 49 个共有差异表达基因显著富集在卵泡刺激素分泌的正向调节通路。卵泡刺激素作用于其受体在刺激卵泡发育和成熟中起着关键作用[24]，并且在子痫前期患者的胎盘样本中卵泡刺激素受体 mRNA 的表达水平显著低于正常孕妇的胎盘样本[25]。细胞外区域是最显著富集且基因数量最多的细胞组分。细胞外区域对细胞维持生理功能发挥着至关重要的作用，包括细胞间信号传递、细胞黏附、细胞外基质形成等，已有研究表明存在于细胞外区域的基因如高迁移率族蛋白 B1（high mobility group box 1，*HMGB1*）具有促炎作用且与子痫前期相关[26]。在本研究中显著富集在细胞外区域的 *PROCR*、*HTRA4* 和 *LHB* 已在子痫前期相关研究中被报道。PROCR 是一种跨膜糖蛋白，其在胎盘滋养层细胞中表达下调且与子痫前期有关[27]。HTRA4 和 LHB 在子痫前期患者的胎盘绒毛组织中均明显上调，而且已有研究证实相较于健康人群，LHB 在子痫前期患者的血浆中显著上调[28]。

上述结果提示激素分泌和免疫反应可能在子痫前期发病机制中起着重要作用。KEGG 通路富集分析结果显示，差异表达基因显著富集在细胞因子－细胞因子受体相互作用等信号通路，富集在该通路上的基因 *CX3CR1* 是一种促血管生成因子，CX3CR1 表达可能与子痫前期胎盘血管网发育不全有关[29]。这一结果表明血管生成功能障碍可能是子痫前期的发病机制之一[30]。

为了进一步探索与子痫前期高度相关的关键基因，本研究通过 WGCNA 将差异表达基因分为 7 个共表达模块，结果提示蓝色与棕色模块中的 *INHA*、*LINC02009* 和 *MIR31HG* 与胎儿性别相关。研究表明孕早期维生素 D 水平对胎儿性别有一定影响[31]，Murata 等[32]发现在成年雌鼠的垂体前叶中 INHA 和维生素 D 具有强正相关性，提示 INHA 与子痫前期患者的胎儿性别有一定关联。本研究结果还提示，蓝色和红色模块中的 *FSTL3*、*INHA* 和 *LEP* 与模块高度相关。先前的研究表明，FSTL3 和 INHA 在子痫前期孕妇胎盘样本中明显上调，被定义为枢纽基因，且在预测子痫前期的 logistic 回归模型中表现优秀[33]。此外，*INHA* 被认为是母体子痫前期的易感基因，可能通过高度失调的免疫和炎症反应促进子痫前期的发展[34]。LEP 也被多项研究证明是与子痫前期发病相关的关键基因[35-37]。蓝色模块进一步被鉴定为与子痫前期高度相关的关键模块，*FLT1*、*PAPPA2*、*PPP1R1C*、*MYO7B*、*LINC02009* 和 *INHA* 是与模块高度相关且与性状显著相关的关键基因，其中 *FLT1*、*PAPPA2* 和 *INHA* 在所有样本中的表达量相对较高。这些基因之间具有强相互作用，基于此构建的子痫前期的随机森林模型表现优秀（AUC＝0.978）。FLT1、PAPPA2 和 INHA 已被研究证实与子痫前期密切相关。FLT1 是一种血管内皮生长因子受体（vascular endothelial growth factor receptor，VEGFR），研究报道 *VEGFR* 基因与血管生长和内皮功能障碍有关，这可能在一定程度上解释了子痫前期的发生[38]。多项研究显示，可溶性 FLT1 在子痫前期早筛和诊断中具有重要价值[39-40]。PAPPA2 编码蛋白质分裂形成的胰岛素样生长因子结合蛋白 5 在子痫前期患者的血浆中显著上调[41-43]。胰岛素样生长因子在刺激绒毛外滋养层侵袭和子宫螺旋动脉重塑过程中发挥着重要作用，PAPPA2 在子痫前期患者的血浆中表达增加被认为与子宫胎盘缺血有关[44]。一些血清学研究也证明了 PAPPA2 早期预测子痫前期的潜在价值[45-46]。上述结果表明，本研究所使用的方法挖掘出的大量靶标与子痫前期的发生、发展高度相关。因此，尽管目前仅在少数研究中涉及的 MYO7B[7]，以及尚未有明确研究报道的 PPP1R1C 和 LINC02009 同样值得关注，它们也可能参与了子痫前期的发生，并对患者的早期诊断及治疗有潜在价值。

综上所述，本研究利用来自 NCBI GEO 的子痫前期患者胎盘组织的 RNA-seq 数据，从 4 个数据集中初步筛选出 49 个共有差异表达基因，并明确它们的功能富集途径与激素分泌和免疫反应等相关；进一步通过 WGCNA 筛选出共表达网络中的 6 个关键基因（*FLT1*、*PAPPA2*、*PPP1R1C*、*MYO7B*、*LINC02009* 和 *INHA*），并在随机森林模型中证实了它们作为子痫前期早期筛查与诊断分子标志物的潜力。这些包括血管生成因子、妊娠相关血浆蛋白、抑制素在内的分子或许能够成为子痫前期早期筛查、诊断的候选标志物，一些新发现的关键基因也可能为子痫前期的治疗提供新的靶点。本研究同时证实 WGCNA 方法能够显著增强数据挖掘的系统性并提高效率，在筛选疾病诊断标志物和治疗靶点方面的具有较高的优越性和可靠性。

本研究具有一定的局限性，一是本研究的 4 个数据集之间存在一定异质性，尽管本研究选择了 4 个数据集的共有差异表达基因，但在后续分析中仍应继续扩大样本量，增加结果的可靠性和可重复性；二是本研究筛选出的候选基因在未来需要更多的工作来验证它们的实际临床应用价值，并探索它们在子痫前期发病相关通路中的调控作用。

［参 考 文 献］

［1］ Gestational hypertension and preeclampsia: ACOG practice bulletin summary, number 222[J]. Obstet Gynecol, 2020, 135(6): 1492-1495. DOI: 10.1097/AOG.0000000000003892.

［2］ ROBERTS J M, RICH-EDWARDS J W, MCELRATH T F, et al. Subtypes of preeclampsia: recognition and determining clinical usefulness[J]. Hypertension, 2021, 77(5): 1430-1441. DOI: 10.1161/HYPERTENSIONAHA.120.14781.

［3］ ROLNIK D L, WRIGHT D, POON L C, et al. Aspirin versus placebo in pregnancies at high risk for preterm preeclampsia[J]. N Engl J Med, 2017, 377(7): 613-622.

DOI: 10.1056/NEJMoa1704559.

[4] ROBERGE S, NICOLAIDES K, DEMERS S, et al. The role of aspirin dose on the prevention of preeclampsia and fetal growth restriction: systematic review and meta-analysis[J]. Am J Obstet Gynecol, 2017, 216(2): 110-120.e6. DOI: 10.1016/j.ajog.2016.09.076.

[5] 赫英东，陈倩. 阿司匹林预防子痫前期的局限性和临床应用选择[J]. 中国实用妇科与产科杂志，2021，37(5): 519-522. DOI: 10.19538/j.fk2021050104.

[6] RAO M S, VAN VLEET T R, CIURLIONIS R, et al. Comparison of RNA-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies[J]. Front Genet, 2018, 9: 636. DOI: 10.3389/fgene.2018.00636.

[7] KAARTOKALLIO T, CERVERA A, KYLLÖNEN A, et al. Gene expression profiling of pre-eclamptic placentae by RNA sequencing[J]. Sci Rep, 2015, 5(1): 14107. DOI: 10.1038/srep14107.

[8] REN Z, GAO Y, GAO Y, et al. Distinct placental molecular processes associated with early-onset and late-onset preeclampsia[J]. Theranostics, 2021, 11(10): 5028-5044. DOI: 10.7150/thno.56141.

[9] LANGFELDER P, HORVATH S. WGCNA: an R package for weighted correlation network analysis[J]. BMC Bioinformatics, 2008, 9: 559. DOI: 10.1186/1471-2105-9-559.

[10] WAN Q, TANG J, HAN Y, et al. Co-expression modules construction by WGCNA and identify potential prognostic markers of uveal melanoma[J]. Exp Eye Res, 2018, 166: 13-20. DOI: 10.1016/j.exer.2017.10.007.

[11] ZHOU J, GUO H, LIU L, et al. Construction of co-expression modules related to survival by WGCNA and identification of potential prognostic biomarkers in glioblastoma[J]. J Cell Mol Med, 2021, 25(3): 1633-1644. DOI: 10.1111/jcmm.16264.

[12] LIN J, MENG Y, SONG M F, et al. Network-based analysis reveals novel biomarkers in peripheral blood of patients with preeclampsia[J]. Front Mol Biosci, 2022, 9: 757203. DOI: 10.3389/fmolb.2022.757203.

[13] LIU Z, LI M, HUA Q, et al. Identification of an eight-lncRNA prognostic model for breast cancer using WGCNA network analysis and a Cox-proportional hazards model based on L1-penalized estimation[J]. Int J Mol Med, 2019: 1333-1343. DOI: 10.3892/ijmm.2019.4303.

[14] BREIMAN L. Random forests[J]. Mach Learn, 2001, 45(1): 5-32. DOI: 10.1023/A: 1010933404324.

[15] SHERMAN B T, HAO M, QIU J, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)[J]. Nucleic Acids Res, 2022, 50(W1): W216-W221. DOI: 10.1093/nar/gkac194.

[16] 宋英娜，杨剑秋，刘俊涛，等. 早发型重度子痫前期孕妇胎盘组织中差异表达基因的研究[J]. 中华妇产科杂志，2014，49(7): 501-505. DOI: 10.3760/cma.j.issn.0529-567x.2014.07.006.

[17] BU D, LUO H, HUO P, et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis[J]. Nucleic Acids Res, 2021, 49(W1): W317-W325. DOI: 10.1093/nar/gkab447.

[18] SONG Z Y, CHAO F, ZHUO Z, et al. Identification of hub genes in prostate cancer using robust rank aggregation and weighted gene co-expression network analysis[J]. Aging, 2019, 11(13): 4736-4756. DOI: 10.18632/aging.102087.

[19] Gestational hypertension and preeclampsia: ACOG practice bulletin, number 222[J]. Obstet Gynecol, 2020, 135(6): e237-e260. DOI: 10.1097/AOG.0000000000003891.

[20] BOKSLAG A, VAN WEISSENBRUCH M, MOL B W, et al. Preeclampsia; short and long-term consequences for mother and neonate[J]. Early Hum Dev, 2016, 102: 47-50. DOI: 10.1016/j.earlhumdev.2016.09.007.

[21] DENG N, LEI D, HUANG J, et al. Circular RNA expression profiling identifies hsa_circ_0011460 as a novel molecule in severe preeclampsia[J]. Pregnancy Hypertens, 2019, 17: 216-225. DOI: 10.1016/j.preghy.2019.06.009.

[22] ZHANG B, HORVATH S. A general framework for weighted gene co-expression network analysis[J]. Stat Appl Genet Mol Biol, 2005, 4: Article17. DOI: 10.2202/1544-6115.1128.

[23] KUMAR N, NARAYAN DAS N, GUPTA D, et al. Efficient automated disease diagnosis using machine learning models[J]. J Healthc Eng, 2021, 2021: 9983652. DOI: 10.1155/2021/9983652.

[24] YU H N, RICHARDSON T E, NATARAJA S, et al. Discovery of substituted benzamides as follicle stimulating hormone receptor allosteric modulators[J]. Bioorg Med Chem Lett, 2014, 24(9): 2168-2172. DOI: 10.1016/j.bmcl.2014.03.018.

[25] LI M, JIA Y, LING Y, et al. Reduced expression of follicle stimulating hormone receptor mRNA and protein in pregnancies complicated by pre-eclampsia[J]. Mol Med Rep, 2017, 16(1): 367-372. DOI: 10.3892/mmr.2017.6599.

[26] PRADERVAND P A, CLERC S, FRANTZ J, et al. High mobility group box 1 protein (HMGB-1): a pathogenic role in preeclampsia?[J]. Placenta, 2014, 35(9): 784-786. DOI: 10.1016/j.placenta.2014.06.370.

[27] WANG H, WANG P, LIANG X, et al. Down-regulation of endothelial protein C receptor promotes preeclampsia

by affecting actin polymerization[J]. J Cell Mol Med, 2020, 24(6): 3370-3383. DOI: 10.1111/jcmm.15011.

[28] LAPAIRE O, GRILL S, LALEVEE S, et al. Microarray screening for novel preeclampsia biomarker candidates[J]. Fetal Diagn Ther, 2012, 31(3): 147-153. DOI: 10.1159/000337325.

[29] SZEWCZYK G, PYZLAK M, PANKIEWICZ K, et al. The potential association between a new angiogenic marker fractalkine and a placental vascularization in preeclampsia[J]. Arch Gynecol Obstet, 2021, 304(2): 365-376. DOI: 10.1007/s00404-021-05966-3.

[30] REDMAN C W G, STAFF A C. Preeclampsia, biomarkers, syncytiotrophoblast stress, and placental capacity[J]. Am J Obstet Gynecol, 2015, 213(4 Suppl): S9-S11, S9.e1. DOI: 10.1016/j.ajog.2015.08.003.

[31] VESTERGAARD A L, ANDERSEN M K, OLESEN R V, et al. High-dose vitamin D supplementation significantly affects the placental transcriptome[J]. Nutrients, 2023, 15(24): 5032. DOI: 10.3390/nu15245032.

[32] MURATA T, CHIBA S, KAWAMINAMI M. Changes in the expressions of annexin A1, annexin A5, inhibin/activin subunits, and vitamin D receptor mRNAs in pituitary glands of female rats during the estrous cycle: correlation analyses among these factors[J]. J Vet Med Sci, 2022, 84(8): 1065-1073. DOI: 10.1292/jvms.22-0141.

[33] HUANG S, CAI S, LI H, et al. Prediction of differentially expressed genes and a diagnostic signature of preeclampsia via integrated bioinformatics analysis[J]. Dis Markers, 2022, 2022: 5782637. DOI: 10.1155/2022/5782637.

[34] YONG H E J, MELTON P E, JOHNSON M P, et al. Genome-wide transcriptome directed pathway analysis of maternal pre-eclampsia susceptibility genes[J]. PLoS One, 2015, 10(5): e0128230. DOI: 10.1371/journal.pone.0128230.

[35] CHEN S, KE Y, CHEN W, et al. Association of the LEP gene with immune infiltration as a diagnostic biomarker in preeclampsia[J]. Front Mol Biosci, 2023, 10: 1209144. DOI: 10.3389/fmolb.2023.1209144.

[36] PENG Y, HONG H, GAO N, et al. Bioinformatics methods in biomarkers of preeclampsia and associated potential drug applications[J]. BMC Genomics, 2022, 23(1): 711. DOI: 10.1186/s12864-022-08937-3.

[37] MOHAMAD M A, MOHD MANZOR N F, ZULKIFLI N F, et al. A review of candidate genes and pathways in preeclampsia-an integrated bioinformatical analysis[J]. Biology, 2020, 9(4): 62. DOI: 10.3390/biology9040062.

[38] DUAN W, XIA C, WANG K, et al. A meta-analysis of the vascular endothelial growth factor polymorphisms associated with the risk of pre-eclampsia[J]. Biosci Rep, 2020, 40(5): BSR20190209. DOI: 10.1042/BSR20190209.

[39] STEPAN H, GALINDO A, HUND M, et al. Clinical utility of sFlt-1 and PlGF in screening, prediction, diagnosis and monitoring of pre-eclampsia and fetal growth restriction[j]. Ultrasound Obstet Gynecol, 2023, 61(2): 168-180. DOI: 10.1002/uog.26032.

[40] SOUDERS C A, MAYNARD S E, YAN J, et al. Circulating levels of sFlt1 splice variants as predictive markers for the development of preeclampsia[J]. Int J Mol Sci, 2015, 16(6): 12436-12453. DOI: 10.3390/ijms160612436.

[41] MACINTIRE K, TUOHEY L, YE L, et al. PAPPA2 is increased in severe early onset pre-eclampsia and upregulated with hypoxia[J]. Reprod Fertil Dev, 2014, 26(2): 351-357. DOI: 10.1071/RD12384.

[42] WAGNER P K, OTOMO A, CHRISTIANS J K. Regulation of pregnancy-associated plasma protein A2 (PAPPA2) in a human placental trophoblast cell line (BeWo)[J]. Reprod Biol Endocrinol, 2011, 9: 48. DOI: 10.1186/1477-7827-9-48.

[43] KEIKKALA E, FORSTÉN J, RITVOS O, et al. Serum Inhibin-A and PAPP-A2 in the prediction of pre-eclampsia during the first and second trimesters in high-risk women[J]. Pregnancy Hypertens, 2021, 25: 116-122. DOI: 10.1016/j.preghy.2021.05.024.

[44] LAMALE-SMITH L M, GUMINA D L, KRAMER A W, et al. Uteroplacental ischemia is associated with increased PAPP-A2[J]. Reprod Sci, 2020, 27(2): 529-536. DOI: 10.1007/s43032-019-00050-3.

[45] NAGALLA S R, JANAKI V, VIJAYALAKSHMI A R, et al. Glycosylated fibronectin point-of-care test for diagnosis of pre-eclampsia in a low-resource setting: a prospective Southeast Asian population study[J]. BJOG, 2020, 127(13): 1687-1694. DOI: 10.1111/1471-0528.16323.

[46] WANG J, HU H, LIU X, et al. Predictive values of various serum biomarkers in women with suspected preeclampsia: a prospective study[J]. J Clin Lab Anal, 2021, 35(5): e23740. DOI: 10.1002/jcla.23740.

［本文编辑］　杨亚红