

基于GEO数据库对子痫前期关键基因筛选及诊断模型构建的研究

李剑霞, 李 蕾*

呼和浩特市妇幼保健院(呼和浩特市妇女儿童医院), 内蒙古 呼和浩特

收稿日期: 2026年6月13日; 录用日期: 2026年7月6日; 发布日期: 2026年7月14日

摘 要

目的: 旨在挖掘子痫前期(Preeclampsia, PE)患者相较正常孕妇胎盘组织中的关键差异表达基因及其临床诊断价值, 并探索性构建候选诊断模型。方法: 从Gene Expression Omnibus (GEO)公共数据库下载GSE74341、GSE10588和GSE25906三个PE胎盘组织基因表达数据集, 筛选差异基因, 采用生物信息学技术进行基因本体(Gene Ontology, GO)过表征分析和基因集富集分析(Gene Set Enrichment Analysis, GSEA)分析差异基因功能。以LASSO回归构建主诊断模型, 随机森林和支持向量机作为敏感性分析模型; 通过嵌套五折交叉验证、独立外部验证和留一队列交叉验证评估模型性能。结果: 共筛选出686个显著差异基因, 其中上调347个、下调339个。GO富集分析显示差异基因主要参与氧水平响应(response to oxygen levels)、缺氧反应(response to hypoxia)、氧含量降低响应(response to decreased oxygen levels)和骨化(ossification)等生物过程, GSEA分析提示脂肪因子信号通路、细胞周期和细胞黏附分子相互作用等通路呈趋势性富集。LASSO算法构建了由YWHAB、TIMM22、TMEM45A、KARS、QPCT、SDHC、ARHGEF16、CCNH、CLN6、CNOT2、COPS5、EGFR和EPC2组成的13基因诊断面板, 其中YWHAB、TIMM22和TMEM45A为交叉验证中稳定入选的基因(选择频率 $\geq 80\%$)。模型在发现队列内嵌套交叉验证AUC = 0.937 (95%CI 0.870~0.989), 敏感性为0.875, 特异性为0.944, 准确率为0.912, Brier评分为0.139; 模型在GSE25906独立外部队列上AUC = 0.659 (95%CI 0.510~0.803), 敏感性为0.696, 特异性为0.703, 准确率为0.700, Brier评分为0.259; 三队列LOCO-CV平均AUC = 0.654 (范围0.513~0.846)。结论: 本研究在PE胎盘组织中筛选出以YWHAB、TIMM22和TMEM45A为核心的候选差异表达基因, 并构建了以13个基因为核心的LASSO候选诊断面板。该模型在发现队列内部具有良好判别能力, 但外部队列和留一队列验证提示跨队列泛化能力有限, 仍需在更大规模、表型一致且平台标准化的独立队列中进一步验证。

关键词

子痫前期, 生物信息分析, 差异表达基因, LASSO回归

*通讯作者。

Research on Key Gene Screening and Diagnostic Model Construction for Pre-Eclampsia Based on the GEO Database

Jianxia Li, Qiang Li*

Hohhot Maternal and Child Health Care Hospital (Hohhot Women and Children's Hospital), Hohhot Inner Mongolia

Received: June 13, 2026; accepted: July 6, 2026; published: July 14, 2026

Abstract

Objective: To identify key differentially expressed genes (DEGs) in placental tissues from patients with preeclampsia (PE) compared with normal pregnant women, evaluate their clinical diagnostic value, and exploratorily construct candidate diagnostic models. **Methods:** Three PE placental tissue gene expression datasets, GSE74341, GSE10588, and GSE25906, were downloaded from the Gene Expression Omnibus (GEO) public database. DEGs were screened, and bioinformatics approaches were used to perform Gene Ontology (GO) over-representation analysis and Gene Set Enrichment Analysis (GSEA) to investigate their biological functions. A primary diagnostic model was constructed using LASSO regression, with random forest and support vector machine models used for sensitivity analysis. Model performance was evaluated using nested five-fold cross-validation, independent external validation, and leave-one-cohort-out cross-validation. **Results:** A total of 686 significant DEGs were identified, including 347 upregulated and 339 downregulated genes. GO enrichment analysis showed that these DEGs were mainly involved in biological processes such as response to oxygen levels, response to hypoxia, response to decreased oxygen levels, and ossification. GSEA suggested trend-level enrichment of pathways including adipocytokine signaling, cell cycle, and cell adhesion molecule interactions. A 13-gene diagnostic panel was constructed using the LASSO algorithm, consisting of YWHAB, TIMM22, TMEM45A, KARS, QPCT, SDHC, ARHGEF16, CCNH, CLN6, CNOT2, COPS5, EGFR, and EPC2. Among these, YWHAB, TIMM22, and TMEM45A were stable genes across folds, with selection frequencies of $\geq 80\%$. In nested cross-validation within the discovery cohort, the model achieved an AUC of 0.937, 95% CI: 0.870~0.989, with sensitivity of 0.875, specificity of 0.944, accuracy of 0.912, and a Brier score of 0.139. In the independent external validation cohort GSE25906, the model achieved an AUC of 0.659, 95% CI: 0.510~0.803, with sensitivity of 0.696, specificity of 0.703, accuracy of 0.700, and a Brier score of 0.259. The mean AUC in three-cohort leave-one-cohort-out cross-validation was 0.654, with a range of 0.513~0.846. **Conclusion:** This study identified candidate DEGs centered on YWHAB, TIMM22, and TMEM45A in PE placental tissues and constructed a 13-gene LASSO-based candidate diagnostic panel. The model showed good discriminative performance within the discovery cohort; however, external validation and leave-one-cohort-out validation indicated limited cross-cohort generalizability. Further validation in larger independent cohorts with consistent phenotypes and standardized platforms is warranted.

Keywords

Preeclampsia, Bioinformatics Analysis, Differentially Expressed Genes, LASSO Regression



1. 引言

子痫前期(Preeclampsia, PE)是妊娠期特有的多系统功能障碍性疾病,以妊娠 20 周后新发高血压和蛋白尿为主要特征,可进展为子痫抽搐、早产、孕产妇死亡等严重并发症,是全球孕产妇和围产儿死亡的主要原因之一[1][2]。据统计,PE 的发病率高达 4%~8%,即使采取预防措施,仍可能对母婴健康产生短期和长期的负面影响[3][4]。然而,由于 PE 受到胎盘缺血、氧化应激、炎症反应等多重因素的影响,其确切发病机制至今尚未完全阐明[5]。

胎盘作为母体与胎儿间物质交换的关键媒介,对于胎儿的生长发育和正常妊娠的维持至关重要,大量研究已证实胎盘缺血缺氧、血管内皮功能障碍、炎症反应过度激活以及免疫调节失衡等因素共同参与了 PE 的病理过程[6]-[8]。因此,通过研究胎盘组织,探究 PE 的发病机制,寻找具有高度敏感性与特异性的分子标志物,构建高效的诊断模型,对于 PE 的早期发现与及时干预具有重大意义。

近年来,高通量测序技术的快速发展为揭示 PE 的分子机制及构建诊断模型提供了新的思路和方法。本研究基于 Gene Expression Omnibus (GEO)数据库中的基因表达数据,运用生物信息学方法分析 PE 与正常妊娠者(Normal Control, NC)的差异表达基因,通过基因本体(Gene Ontology, GO)和基因集富集分析(Gene Set Enrichment Analysis, GSEA)揭示其涉及的生物学过程和信号通路。本研究基于 GEO 数据库中 PE 与 NC 胎盘组织表达谱数据,采用差异表达分析、GO 富集分析和 GSEA 探索 PE 相关分子改变,并以最小绝对收缩与选择算子(Least Absolute Shrinkage and Selection Operator regression, LASSO)回归构建候选诊断模型。与单一训练-测试划分相比,本研究进一步采用嵌套交叉验证、独立外部验证和留一队列交叉验证评估模型的稳定性和跨队列泛化能力,以降低小样本高维建模中常见的数据泄露和过拟合风险。

2. 资料与方法

2.1. 基因微阵列数据下载

以“Preeclampsia”为关键词从 GEO 数据库(<https://www.ncbi.nlm.nih.gov/geo/>)进行检索,筛选样本为 PE 与 NC 胎盘组织的基因微阵列数据,最终纳入三个独立队列:GSE74341 (PE 15 例,NC 10 例)、GSE10588 (PE 17 例,NC 26 例)和 GSE25906 (PE 23 例,NC 37 例)。其中, GSE74341 与 GSE10588 合并作为发现队列(PE 32 例,NC 36 例,共 68 例),合并后采用 sva 程序包中的 ComBat 方法校载,经 log₂ 转换、标准化、探针注释和基因符号汇总后进入后续分析;当多个探针对应同一基因时,保留表达方差较大的探针代表该基因。

2.2. 差异基因筛选

基于预处理后的 GSE74341 和 GSE10588 表达矩阵,取两个队列共有基因合并,采用 ComBat 方法进行批次效应校正。运用 R 语言“stats”程序包进行主成分分析(Principal Component Analysis, PCA),评估样本整体分布。采用 R 语言“limma”程序包进行线性模型分析,以 Benjamini-Hochberg 方法校正多重检验,筛选阈值设定为 FDR < 0.05 且 |log₂FC| > 0.5,获得差异表达基因,通过火山图将差异基因分布可视化。

2.3. 聚类与功能富集分析

本研究使用 R 语言“pheatmap”程序包对差异基因的表达数据进行聚类分析, 通过热图呈现 PE 与 NC 样本间的基因表达模式差异, 系统揭示 PE 相关的生物学通路变化。功能富集分析采用 clusterProfiler 程序包完成, 包括基因本体(Gene Ontology, GO)过表征分析和基因集富集分析(Gene Set Enrichment Analysis, GSEA)。GO 过表征分析覆盖生物过程(Biological Process, BP)、细胞组分(Cellular Component, CC)和分子功能(Molecular Function, MF)。GSEA 以发现队列全部共有基因按 limma 统计量排序后的基因列表作为输入, 参考 GO 本体和 KEGG 通路基因集进行富集分析。多重检验校正采用 Benjamini-Hochberg 方法; GO 过表征分析以校正后 $P < 0.05$ 为显著阈值, GSEA 结果同时以 P 值和 FDR 校正结果。

2.4. 诊断模型构建与多层次验证

以筛选得到的显著差异表达基因为候选基因, 并按调整后 P 值和 $|\log_2FC|$ 排序, 选取前 200 个基因作为候选基因。模型构建采用 LASSO 逻辑回归作为主模型, 随机森林和支持向量机作为敏感性分析模型。

为降低小样本高维建模中的过拟合和数据泄露风险, 采用嵌套五折交叉验证(nested 5-fold cross-validation)评估发现队列模型性能。外层 5 折用于获得样本预测结果, 内层 5 折用于 LASSO 的 lambda 超参数调优; lambda 选择采用 lambda.1se 规则, 以获得相对稀疏且稳定的模型。每个外层训练折中, 从候选特征空间内进一步筛选前 80 个特征进入建模, 外层测试 fold 仅用于一次预测, 不参与任何特征选择、参数调优或模型训练。各外层 fold 中被 LASSO 选中的基因记录选择频率, 选择频率 $\geq 60\%$ 的基因定义为稳定候选标志物。

跨队列泛化能力采用留一队列交叉验证(leave-one-cohort-out cross-validation, LOCO-CV)评估。每次保留 GSE74341、GSE10588 或 GSE25906 中一个队列作为外部测试集, 其余两个队列合并并经 ComBat 校正后作为训练集, 固定训练流程后在被保留队列上评估模型性能。独立外部验证采用 GSE25906 完整队列, 固定全数据训练得到的最终 LASSO 模型进行预测, 不进行重训练或阈值再优化。跨队列预测前, 对表达矩阵进行队列内基因水平 z-score 标准化, 以减小平台间表达尺度差异对预测概率的影响。

模型性能采用 ROC 曲线下面积(Area Under the Curve, AUC)及 bootstrap 95%置信区间(95% Confidence Interval, 95% CI)进行评估, 并在约登(Youden)指数最优阈值下计算敏感性(Sensitivity)、特异性(Specificity)、准确率(Accuracy)和 Brier 评分。模型校准能力通过校准曲线评估, 潜在临床净获益通过决策曲线分析(Decision Curve Analysis, DCA)评估。最终 LASSO 模型的基因系数用于构建列线图(Nomogram), 以提高模型解释性。

2.5. 统计学分析

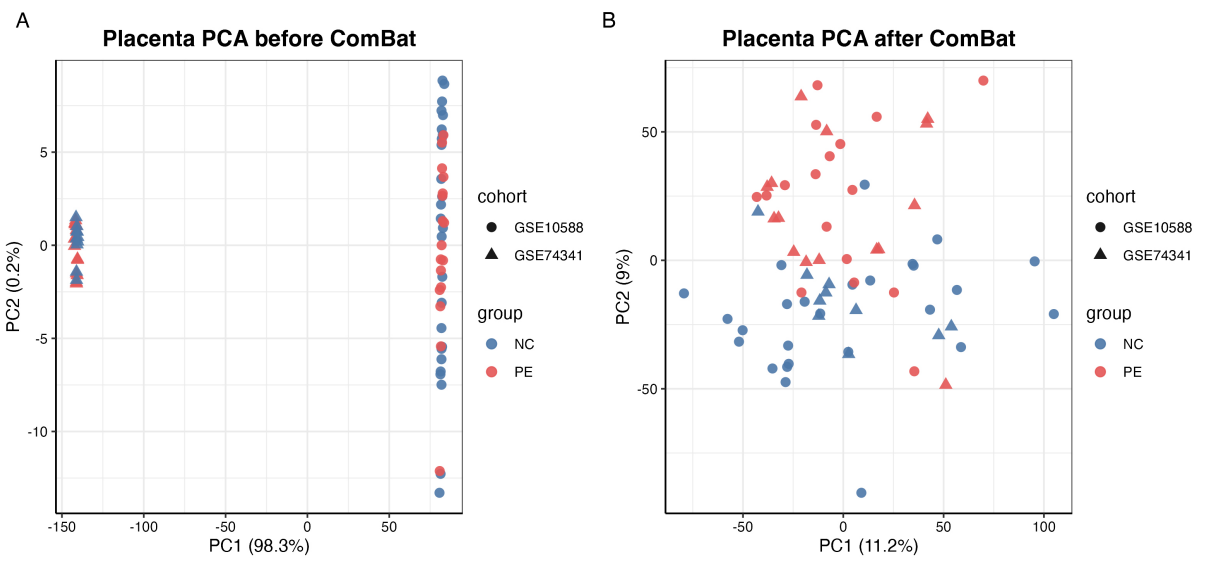
所有数据分析均采用 R 软件完成。差异表达基因筛选以 $FDR < 0.05$ 且 $|\log_2FC| > 0.5$ 为统计学阈值。模型 AUC 的 95%置信区间通过 bootstrap 重抽样获得; 模型在 Youden 指数最优阈值下报告敏感性、特异性、准确率和 Brier 评分。不同模型 AUC 的比较采用 DeLong 检验。双侧 $P < 0.05$ 认为差异具有统计学意义。

3. 结果

3.1. 差异基因分析

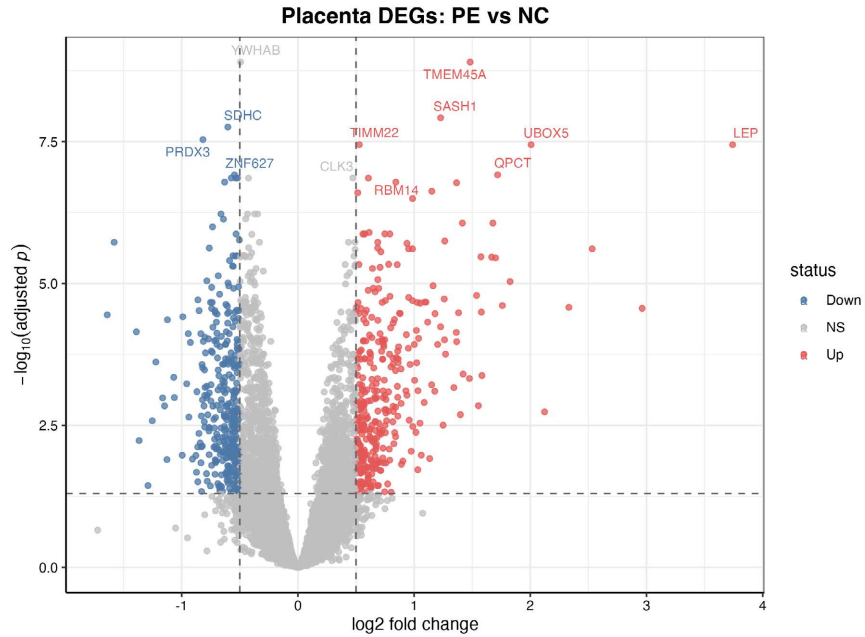
PCA 结果显示, ComBat 校正前样本在主成分空间中存在明显队列分离, 提示不同 GEO 队列间存在批次效应; 经 ComBat 校正后, 队列来源造成的分离明显减弱, 样本分布更适合后续合并分析, 见图 1。基于校正后的发现队列(NC = 36, PE = 32, $n = 68$), 在 11,965 个共有基因中共筛选出 686 个显著差异表达

基因($FDR < 0.05$ 且 $|\log_2FC| > 0.5$), 其中上调 347 个、下调 339 个。火山图显示差异基因在 PE 与 NC 之间的整体分布, 见图 2。差异基因聚类热图显示, PE 与 NC 样本在主要差异基因表达模式上呈现可辨别差异, 见图 3。



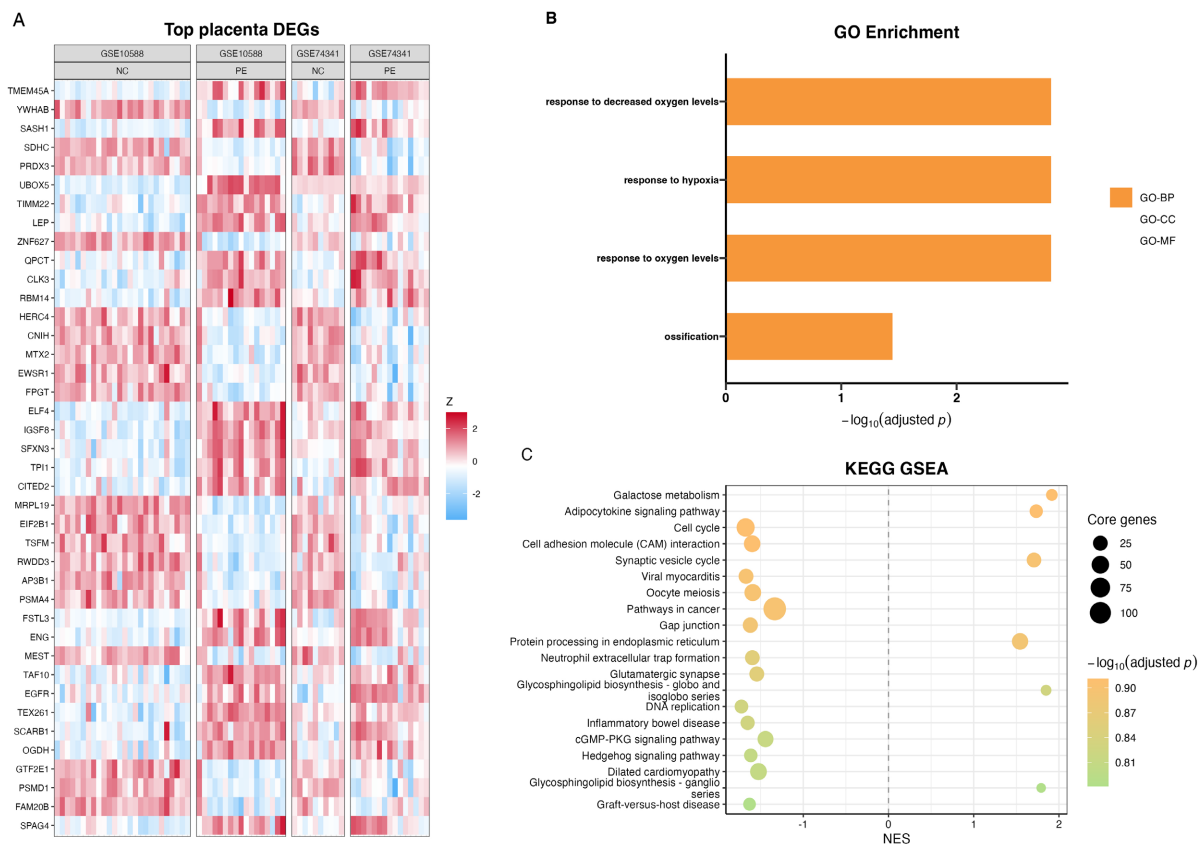
图注: (A) ComBat 校正前, 样本沿 PC1 按队列分离, 提示存在显著批次效应; (B) 使用 sva 包 ComBat 函数校正后, 两队列样本在 PCA 空间充分混合, 支持后续合并差异分析。

Figure 1. Batch effect correction for the combined discovery cohort of GSE74341 and GSE10588
图 1. GSE74341 与 GSE10588 合并发现队列的批次效应校正



图注: 横坐标为 $\log_2$ fold change (PE vs NC), 纵坐标为 $-\log_{10}(\text{adjusted } p\text{-value})$ 。差异分析采用 limma 经验贝叶斯方法, 多重检验校正使用 Benjamini-Hochberg FDR。在 $FDR < 0.05$ 且 $|\log_2FC| > 0.5$ 阈值下共鉴定 686 个差异表达基因, 其中上调 347 个(红色), 下调 339 个(蓝色), 未达阈值基因以灰色显示。

Figure 2. Volcano plot showing differentially expressed genes
图 2. 差异表达基因火山图



图注: (A) 按 $|\log_2\text{FC}|$ 排序的 Top 50 差异基因表达 heatmap, 每行为一个基因, 每列为一个样本, 颜色映射为 z-score 标准化后的表达量; (B) GO 过表达分析 BP/CC/MF 三个本体的 Top 富集条目, 横轴为 $-\log_{10}(\text{adjusted } p\text{-value})$; (C) KEGG-GSEA 通路气泡图, 气泡大小为 leading-edge 基因数, 颜色映射为 $-\log_{10}(\text{FDR})$ 。

Figure 3. Clustering analysis and functional enrichment of differentially expressed genes in the discovery cohort
图 3. 发现队列差异基因聚类与功能富集

3.2. 差异表达基因的 GO 及 GSEA 富集分析

对发现队列 686 个差异表达基因进行 GO 过表达分析。结果显示, BP 层面主要富集于氧水平响应(response to oxygen levels)、缺氧反应(response to hypoxia)、氧含量降低响应(response to decreased oxygen levels)和骨化(ossification)等条目, 提示 PE 胎盘组织中缺氧应激相关转录改变较为突出。上述结果与 PE 中胎盘灌注不足、缺血缺氧和氧化应激参与疾病发生发展的经典认识一致。GSEA 进一步从全基因排序层面评估通路改变。KEGG-GSEA 结果显示, 脂肪因子信号通路(Adipocytokine signaling pathway)呈上调趋势, 细胞周期(Cell cycle)和细胞黏附分子相互作用(Cell adhesion molecule interaction)呈下调趋势。GO-GSEA 结果显示, 激素水平调控、细胞聚集、血小板源性生长因子刺激反应、肽分泌和血管生成相关细胞迁移等条目存在富集, 提示 PE 胎盘中内分泌调控、细胞黏附迁移和血管生成过程可能存在系统性改变, 见图 3。

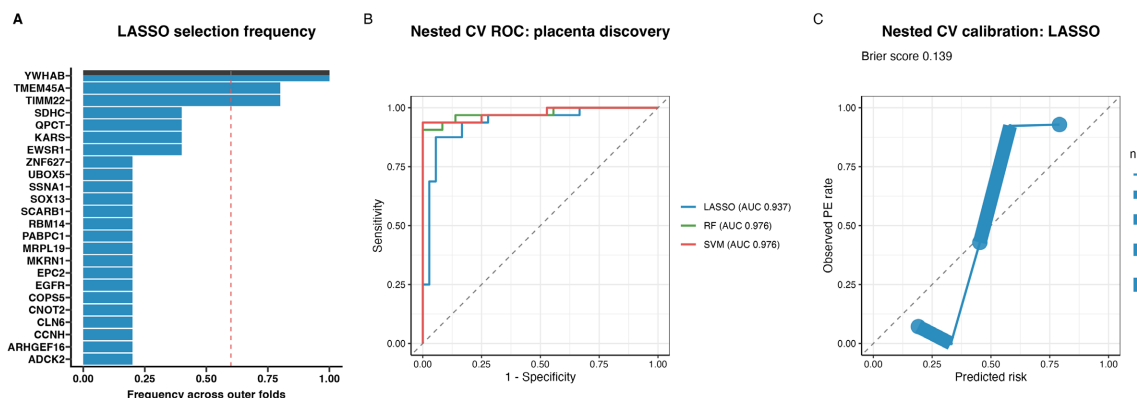
3.3. 诊断模型构建与多层次验证

以差异表达基因中排序前 200 个基因作为候选基因, 采用 LASSO 逻辑回归在嵌套五折交叉验证框架下进行特征筛选和模型训练。LASSO 在不同外层折内的基因选择频率显示, YWHAB 在 5/5 个外层折中被选中, 选择频率为 100%; TIMM22 和 TMEM45A 均在 4/5 个外层折中被选中, 选择频率为 80%。因

此, YWHAB、TIMM22 和 TMEM45A 被定义为交叉验证中稳定入选的基因。基于全发现队列训练得到的最终 LASSO 模型包含 13 个基因: YWHAB、TIMM22、TMEM45A、KARS、QPCT、SDHC、ARHGEF16、CCNH、CLN6、CNOT2、COPS5、EGFR 和 EPC2。最终模型截距为 0.088, 各基因系数分别为 YWHAB = -0.889、TIMM22 = 0.760、TMEM45A = 0.639、KARS = -0.655、QPCT = 0.293、SDHC = -0.388、ARHGEF16 = 0.088、CCNH = -0.086、CLN6 = -0.086、CNOT2 = -0.014、COPS5 = -0.321、EGFR = 0.341 和 EPC2 = -0.227。

LASSO 主模型的嵌套五折交叉验证 AUC = 0.937 (95% CI 0.870~0.989), Youden 最优阈值为 0.490, 敏感性为 0.875, 特异性为 0.944, 准确率为 0.912, Brier 评分为 0.139, 见图 4。作为敏感性分析, 随机森林模型的嵌套交叉验证 AUC = 0.976 (95% CI 0.934~1.000), 敏感性为 0.906, 特异性为 1.000, 准确率为 0.956, Brier 评分为 0.061; 支持向量机模型的嵌套交叉验证 AUC = 0.976 (95% CI 0.934~1.000), 敏感性为 0.938, 特异性为 1.000, 准确率为 0.971, Brier 评分为 0.046。考虑到 LASSO 模型具有更好的稀疏性和可解释性, 本文将其作为主模型进行后续外部验证和生物学解释。在 GSE25906 独立外部验证队列中, 固定发现队列训练得到的 LASSO 模型进行预测, 结果显示 AUC = 0.659 (95% CI 0.510~0.803), Youden 最优阈值为 0.600, 敏感性为 0.696, 特异性为 0.703, 准确率为 0.700, Brier 评分为 0.259, 见图 5。该结果提示模型在独立队列中仍保留一定区分能力, 但外部性能明显低于发现队列内部嵌套交叉验证结果, 说明其跨平台、跨人群和跨表型应用时仅具有有限至中等的可迁移性, 尚不能作为稳定的临床诊断模型直接推广。

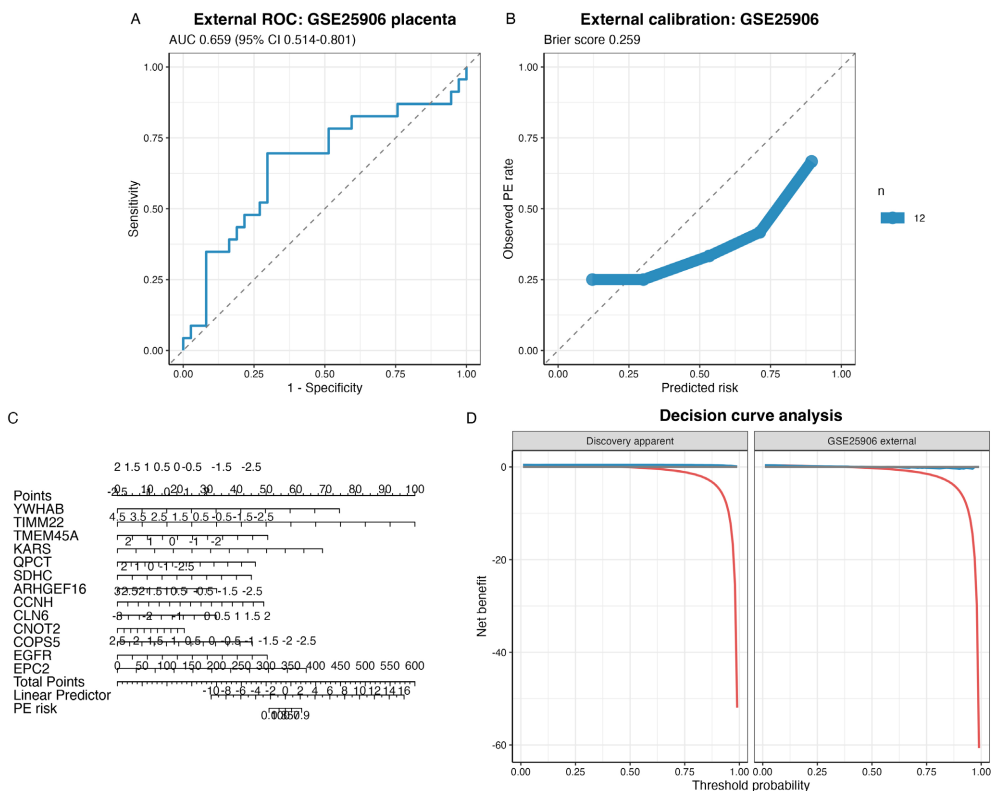
进一步采用三队列留一队列交叉验证评估跨队列泛化能力。以 GSE74341 为 held-out cohort 时, 模型 AUC = 0.513 (95% CI 0.287~0.733), 敏感性为 0.333, 特异性为 1.000, 准确率为 0.600, Brier 评分为 0.307; 以 GSE10588 为 held-out cohort 时, 模型 AUC = 0.846 (95% CI 0.701~0.955), 敏感性为 0.706, 特异性为 0.962, 准确率为 0.860, Brier 评分为 0.182; 以 GSE25906 为 held-out cohort 时, 模型 AUC = 0.602 (95% CI 0.445~0.749), 敏感性为 0.261, 特异性为 0.973, 准确率为 0.700, Brier 评分为 0.340。三队列 LOCO-CV 平均 AUC = 0.654, 范围为 0.513~0.846, 见图 6。该结果提示 PE 胎盘转录组诊断模型存在明显跨队列异质性, 模型在不同平台、不同临床表型构成和不同样本来源队列之间的迁移能力并不稳定。



图注: (A) 外层 5 折交叉验证中 LASSO 各基因的入选频率, 红色虚线标示 60% 稳定性阈值; YWHAB (100%)、TIMM22 (80%)、TMEM45A (80%) 为最稳定的候选标志物。 (B) 嵌套 5 折交叉验证的合并 ROC 曲线, LASSO 在所有外层折叠的预测合并后 AUC = 0.937 (95% CI 0.870~0.989), 同步比较的 Random Forest 与 SVM-radial AUC 分别为 0.976 与 0.976, DeLong 检验三者两两差异均不显著 ($p > 0.10$), 最终选用 LASSO 因其稀疏可解释。 (C) LASSO 嵌套交叉验证的预测概率校准曲线, 横轴为预测风险, 纵轴为观察 PE 比例, 对角虚线为完美校准参照, Brier score = 0.139。

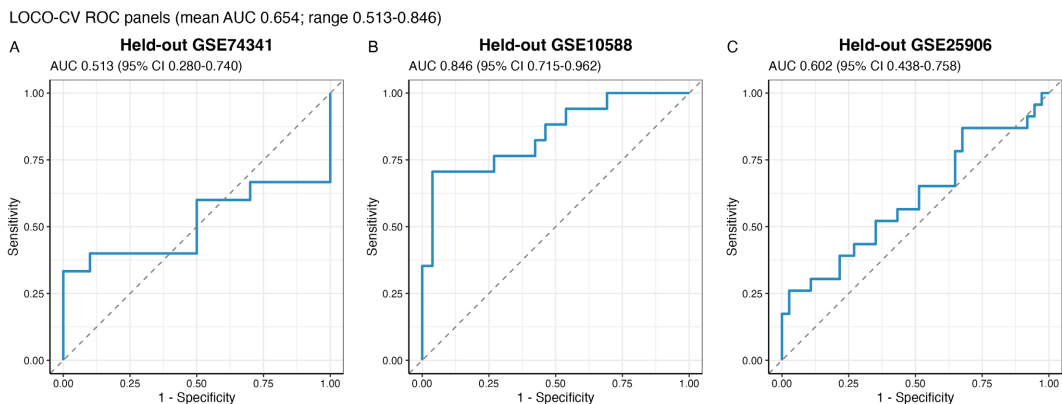
Figure 4. Nested cross-validation results of the LASSO logistic regression model in the discovery cohort

图 4. LASSO 逻辑回归在发现队列内的嵌套交叉验证结果



图注: (A) 将发现队列锁定的 13 基因 LASSO 模型直接应用于 GSE25906 数据集, 样本表达进行队列内 z-score 标准化以处理跨平台异质性, 外部 AUC = 0.659 (95% CI 0.510~0.803), sensitivity = 0.696, specificity = 0.703, accuracy = 0.700。(B) 外部验证校准曲线, Brier score = 0.259, 提示模型概率校准在外部队列存在一定偏移。(C) 基于发现队列拟合的 13 基因列线图, 各基因得分加和后映射至 PE 风险概率, 可作为临床辅助决策的可视化工具。(D) 决策曲线分析, 左面板为发现队列内部 net benefit, 右面板为 GSE25906 外部 net benefit; 模型在临床相关阈值区间内优于 treat-all 与 treat-none 策略。

Figure 5. Independent external validation based on the GSE25906 dataset and construction of the final diagnostic model
图 5. GSE25906 独立外部验证及最终诊断模型



图注: 图(A), (B)和(C)依次为 GSE74341、GSE10588、GSE25906 作为 held-out 队列时的 ROC 曲线, 对应 AUC 分别为 0.513、0.846、0.602, 平均 AUC = 0.654。结果表明模型在跨平台、跨人群条件下可迁移性有限, 且不同队列间差异明显, GSE10588 的 AUC 最高, 而 GSE74341 的 AUC 最低, 提示训练队列与目标队列在临床表型、对照构成和平台设计上的匹配程度可能影响模型迁移表现。

Figure 6. Leave-one-cohort-out cross-validation evaluates cross-cohort generalization performance
图 6. 留一队列交叉验证评估跨队列泛化性能

4. 讨论

PE 作为妊娠期严重且高发的并发症之一, 其发病机制复杂, 早期诊断困难。传统的诊断依赖临床症状进行判断, 但症状出现时, 母婴已受到疾病带来的危害, 故实现早期精准诊断, 采取及时有效干预措施, 降低不良妊娠结局的发生风险至关重要。因此, 寻找敏感特异的分子标志物, 构建高效的诊断模型, 成为 PE 研究领域的热点。本研究基于 GEO 数据库胎盘组织转录组数据, 采用严格的批次校正、嵌套交叉验证、外部验证和留一队列验证策略, 筛选 PE 相关候选标志物并评估其诊断模型表现。

通过差异表达分析和功能富集分析揭示了 PE 胎盘组织中以缺氧应激为核心的转录组改变。GO 富集结果显示, 差异基因显著富集于 response to oxygen levels、response to hypoxia 和 response to decreased oxygen levels 等条目, 这与 PE 中胎盘灌注不足、缺血缺氧和氧化应激参与疾病发生发展的经典模型一致[9]。与此同时, 骨化、细胞迁移、激素调控和血管生成相关过程的改变提示 PE 胎盘中可能存在滋养层细胞分化迁移异常、母胎界面重塑不足和内分泌调控紊乱[10]。KEGG-GSEA 结果显示脂肪因子信号通路上调趋势以及细胞周期、细胞黏附分子相互作用下调趋势与 PE 中代谢异常、滋养层增殖受限和细胞黏附侵袭能力下降的病理特征具有生物学一致性。因此, 富集结果支持胎盘缺氧应激 - 滋养层细胞功能异常 - 血管与代谢调控紊乱这个 PE 胎盘病理机制[11]。

通过 LASSO 嵌套交叉验证构建了基于 13 个基因的诊断模型, 其中 YWHAB、TIMM22 和 TMEM45A 在外层交叉验证中具有较高选择稳定性, 是最值得优先关注的候选标志物。YWHAB 编码 14-3-3 蛋白 β 亚型, 14-3-3 蛋白家族可通过结合磷酸化丝氨酸/苏氨酸位点参与细胞周期、凋亡、迁移和信号转导调控。已有研究显示, YWHAB 可在妊娠早期胎盘中由经典启动子和 LINE-1 替代启动子表达, 且其蛋白产物与胎盘高表达并参与滋养层分化的蛋白存在功能联系[12]。TIMM22 编码线粒体内膜转位酶 22 复合体的重要组成部分, 参与核编码线粒体内膜蛋白的导入和线粒体稳态维持。TIM22 复合体已被证明是线粒体内膜蛋白转运的重要分子机器[13], 而 PE 胎盘中线粒体异常、氧化应激和能量代谢紊乱已有较充分证据支持[14][15]。TMEM45A 是缺氧相关跨膜蛋白。既往研究显示 TMEM45A 在低氧条件下表达升高, 并参与肿瘤细胞对缺氧环境的适应和化疗耐受[16]。同时有研究报道在严重 PE 胎盘中包括 FLT1、LEP、HK2、FSTL3、ENG、TMEM45A、ARHGEF4 和 HTRA1 在内的经典 PE 相关代谢、缺氧和血管生成基因上调[17]。将 TMEM45A 选入模型, 与 GO 富集分析中缺氧反应相关条目的显著富集相互印证, 提示其可能反映 PE 胎盘缺氧应激相关转录改变。

除三个稳定基因外, EGFR 在胎盘中表达较高, 并参与胎盘生长、滋养层融合和血管相关调控; 已有综述专门讨论 EGFR 在正常妊娠和 PE 中的作用[18], 且研究报道 PE 中表皮生长因子信号系统表达降低[19]。QPCT 编码 glutaminyl-peptide cyclotransferase, 已有研究将 Qpct 鉴定为胎盘特异性印记基因[20], 近期单细胞转录组研究进一步显示 Qpct 缺失可导致小鼠胎盘形成和血管生成异常[21]。KARS 编码 lysyl-tRNA synthetase, KARS/KRS 已被报道可作为分泌性促炎信号分子诱导炎症反应[22], 并与血管炎症和动脉粥样硬化相关[23]。因此, KARS 在本模型中的生物学解释应限于蛋白翻译、炎症和血管应激的间接线索, 不能写成已有 PE 直接证据。但是模型基因的统计选择并不等同于因果机制确认。本研究的 LASSO 模型反映的是在胎盘转录组数据中具有分类贡献的一组基因组合, 其生物学合理性来自既往胎盘发育、缺氧应激、线粒体功能、滋养层生物学和血管炎症文献的交叉支持。未来仍需通过 qRT-PCR、Western blot、免疫组化、滋养层细胞功能实验和多中心前瞻性队列进一步验证这些基因的表达差异、细胞定位和功能作用。

本研究的外部验证和 LOCO-CV 结果提示, PE 胎盘转录组模型的跨队列可迁移性有限。造成这一现象的首要原因应是临床和表型异质性, 而不仅是平台差异。GSE74341 同时包含 EOPE 和 LOPE 病例, 并

设置 preterm control 与 term control, 且病例和对照在孕周背景上并不完全一致; GSE10588 则主要为 severe PE 与 normal labor 对照, GEO 中缺乏逐样本孕周、胎儿生长受限、分娩方式等信息; GSE25906 虽提供孕周、胎儿性别、引产方式和 batch, 但未标注 EOPE、LOPE 或严重程度。不同 PE 亚型、孕周跨度、胎儿生长情况、疾病严重程度和采样背景均可能影响胎盘基因表达, 从而削弱模型迁移能力。技术因素同样存在, 包括 GPL16699、GPL2986 和 GPL6102 三种平台差异, 以及 GSE74341 作为三队列中唯一的双通道数据与其余两套单通道强度数据之间在数据尺度和设计上的根本差异。从 LOCO-CV 看, GSE10588 的 AUC 最高, 可能与其 severe PE 信号较强、病例定义相对集中有关; GSE74341 的 AUC 最低, 则可能与 EOPE/LOPE 混合、preterm/term 对照并存及 pooled reference 设计共同有关。

本研究仍存在一定局限性。首先, 尽管本研究纳入三个独立胎盘组织队列, 并采用嵌套交叉验证、独立外部验证和 LOCO-CV 进行多层次评估, 但模型在跨队列验证中表现出明显异质性。LASSO 模型在发现队列内部 nested CV 中 AUC 为 0.937, 而在 GSE25906 外部验证中 AUC 为 0.659, LOCO-CV 平均 AUC 为 0.654, 提示不同 GEO 队列之间的平台差异、样本处理差异和 PE 临床表型构成差异会显著影响模型泛化能力。其次, 本研究基于胎盘组织转录组数据, 胎盘样本通常在分娩后获取, 因此该模型目前更适合定位为胎盘分子病理研究和候选诊断标志物筛选工具, 不能直接等同于产前无创筛查模型。第三, 本研究尚未在母体外周血、血清、尿液或其他临床可及样本中验证该基因面板的表达和诊断价值, 未来需要评估这些候选基因是否可转化为产前可检测标志物。第四, 本研究未进行实验验证, 候选基因的表达差异、蛋白水平变化、组织定位及其对滋养层细胞增殖、迁移、侵袭、线粒体功能和缺氧应答的影响仍需进一步实验研究确认。

本研究基于三个 GEO 胎盘组织队列, 采用 ComBat 批次效应校正、limma 差异表达分析、LASSO 逻辑回归、嵌套五折交叉验证、独立外部验证和留一队交叉验证, 构建了 PE 胎盘组织诊断模型, 并构建出以 YWHAB、TIMM22 和 TMEM45A 为核心的 13 基因诊断模型。该模型在发现队列内部具有良好判别能力, 但在外部队列和 LOCO-CV 中表现为中等且存在明显异质性, 提示 PE 胎盘分子诊断模型仍需在更大规模、表型定义一致、平台标准化的多中心队列中进一步验证。本研究为 PE 胎盘分子标志物筛选和诊断模型构建提供了方法学参考和候选基因基础。

参考文献

- [1] 金璟, 何路路, 陈园, 等. 子痫前期胎盘组织中差异表达基因及其临床意义[J]. 广东医学, 2023, 44(9): 1142-1149.
- [2] Poon, L.C., Shennan, A., Hyett, J.A., et al. (2019) The International Federation of Gynecology and Obstetrics (FIGO) Initiative on Pre-Eclampsia: A Pragmatic Guide for First-Trimester Screening and Prevention. *International Journal of Gynecology & Obstetrics*, **145**, 1-33.
- [3] (2020) Gestational Hypertension and Preeclampsia: ACOG Practice Bulletin Summary, Number 222. *Obstetrics & Gynecology*, **135**, 1492-1495.
- [4] 苏朔瑶. 基于生物信息学筛选子痫前期潜在生物标志物及临床风险预测模型的构建与验证[D]: [硕士学位论文]. 长春: 吉林大学, 2024.
- [5] Melchiorre, K., Giorgione, V. and Thilaganathan, B. (2022) The Placenta and Preeclampsia: Villain or Victim? *American Journal of Obstetrics and Gynecology*, **226**, S954-S962. <https://doi.org/10.1016/j.ajog.2020.10.024>
- [6] Hu, M., Li, J., Baker, P.N. and Tong, C. (2022) Revisiting Preeclampsia: A Metabolic Disorder of the Placenta. *The FEBS Journal*, **289**, 336-354. <https://doi.org/10.1111/febs.15745>
- [7] Rana, S., Lemoine, E., Granger, J.P. and Karumanchi, S.A. (2019) Preeclampsia: Pathophysiology, Challenges, and Perspectives. *Circulation Research*, **124**, 1094-1112. <https://doi.org/10.1161/circresaha.118.313276>
- [8] Ma'ayeh, M. and Costantine, M.M. (2020) Prevention of Preeclampsia. *Seminars in Fetal and Neonatal Medicine*, **25**, Article ID: 101123. <https://doi.org/10.1016/j.siny.2020.101123>
- [9] Redman, C.W.G. and Sargent, I.L. (2000) Placental Debris, Oxidative Stress and Pre-Eclampsia. *Placenta*, **21**, 597-602. <https://doi.org/10.1053/plac.2000.0560>

- [10] Lyall, F., Bulmer, J.N., Duffie, E., Cousins, F., Theriault, A. and Robson, S.C. (2001) Human Trophoblast Invasion and Spiral Artery Transformation: The Role of PECAM-1 in Normal Pregnancy, Preeclampsia, and Fetal Growth Restriction. *The American Journal of Pathology*, **158**, 1713-1721. [https://doi.org/10.1016/s0002-9440\(10\)64127-2](https://doi.org/10.1016/s0002-9440(10)64127-2)
- [11] Ives, C.W., Sinkey, R., Rajapreyar, I., Tita, A.T.N. and Oparil, S. (2020) Preeclampsia—Pathophysiology and Clinical Presentations: JACC State-of-the-Art Review. *Journal of the American College of Cardiology*, **76**, 1690-1702. <https://doi.org/10.1016/j.jacc.2020.08.014>
- [12] Demeneva, V.V., Tolmacheva, E.N., Nikitina, T.V., Sazhenova, E.A., Yuriev, S.Y., Makhmutkhodzhaev, A.S., *et al.* (2023) Expression of the NUP153 and YWHAB Genes from Their Canonical Promoters and Alternative Promoters of the LINE-1 Retrotransposon in the Placenta of the First Trimester of Pregnancy. *Vavilov Journal of Genetics and Breeding*, **27**, 63-71. <https://doi.org/10.18699/vjgb-23-09>
- [13] Valpadashi, A., Callegari, S., Linden, A., Neumann, P., Ficner, R., Urlaub, H., *et al.* (2021) Defining the Architecture of the Human TIM22 Complex by Chemical Crosslinking. *FEBS Letters*, **595**, 157-168. <https://doi.org/10.1002/1873-3468.13978>
- [14] Vangrieken, P., Al-Nasiry, S., Bast, A., Leermakers, P.A., Tulen, C.B.M., Schiffrers, P.M.H., *et al.* (2021) Placental Mitochondrial Abnormalities in Preeclampsia. *Reproductive Sciences*, **28**, 2186-2199. <https://doi.org/10.1007/s43032-021-00464-y>
- [15] Hu, X.Q. and Zhang, L. (2022) Mitochondrial Dysfunction in the Pathogenesis of Preeclampsia. *Current Hypertension Reports*, **24**, 157-172. <https://doi.org/10.1007/s11906-022-01184-7>
- [16] Flamant, L., Roegiers, E., Pierre, M., Hayez, A., Sterpin, C., De Backer, O., *et al.* (2012) TMEM45A Is Essential for Hypoxia-Induced Chemoresistance in Breast and Liver Cancer Cells. *BMC Cancer*, **12**, Article No. 391. <https://doi.org/10.1186/1471-2407-12-391>
- [17] Aisagbonhi, O., Bui, T., Nasamran, C.A., St. Louis, H., Pizzo, D., Meads, M., *et al.* (2023) High Placental Expression of FLT1, LEP, PHYHIP and IL3RA—In Persons of African Ancestry with Severe Preeclampsia. *Placenta*, **144**, 13-22. <https://doi.org/10.1016/j.placenta.2023.10.008>
- [18] Clemente, L. and Bird, I.M. (2023) The Epidermal Growth Factor Receptor in Healthy Pregnancy and Preeclampsia. *Journal of Molecular Endocrinology*, **70**, e220105. <https://doi.org/10.1530/jme-22-0105>
- [19] Armant, D.R., Fritz, R., Kilburn, B.A., Kim, Y.M., Nien, J.K., Maihle, N.J., *et al.* (2015) Reduced Expression of the Epidermal Growth Factor Signaling System in Preeclampsia. *Placenta*, **36**, 270-278. <https://doi.org/10.1016/j.placenta.2014.12.006>
- [20] Guo, J., He, H., Liu, Q., Zhang, F., Lv, J., Zeng, T., *et al.* (2015) Identification and Epigenetic Analysis of a Maternally Imprinted Gene Qpct. *Molecules and Cells*, **38**, 859-865. <https://doi.org/10.14348/molcells.2015.0098>
- [21] Guo, J., Zheng, J., Li, R., Yao, J., Zhang, H., Wang, X., *et al.* (2025) Single-Cell Transcriptome Analysis Reveals Abnormal Angiogenesis and Placentation by Loss of Imprinted GlutaminyI-Peptide Cyclotransferase. *Journal of Zhejiang University-Science B*, **26**, 589-608. <https://doi.org/10.1631/jzus.b2400099>
- [22] Park, S.G., Kim, H.J., Min, Y.H., Choi, E., Shin, Y.K., Park, B., *et al.* (2005) Human Lysyl-tRNA Synthetase Is Secreted to Trigger Proinflammatory Response. *Proceedings of the National Academy of Sciences*, **102**, 6356-6361. <https://doi.org/10.1073/pnas.0500226102>
- [23] Yun, K., Seo, Y., Oh, Y., Kim, S., Maeng, H.S., Cho, J.G., *et al.* (2025) Lysyl-tRNA Synthetase 1 Promotes Atherogenesis via Autophagy-Related Secretion and Inflammation. *Scientific Reports*, **15**, Article No. 11099. <https://doi.org/10.1038/s41598-025-96046-y>