

基于生物信息学和机器学习的阿尔兹海默症诊断模型构建及免疫分析

徐林瑞¹, 张译予¹, 崔家齐¹, 丛显铸¹, 李爽¹, 葛佳瑜¹, 孔雨佳¹, 王素珍¹, 石福艳¹, 王金荣²
(1. 山东第二医科大学公共卫生学院卫生统计学教研室, 山东 潍坊 261053; 2. 山东第二医科大学附属医院中医科, 山东 潍坊 261041)

[摘要] **目的:** 利用生物信息学技术和机器学习 (ML) 算法筛选阿尔兹海默症 (AD) 相关基因并构建其诊断模型, 探讨AD患者的免疫学特征, 为AD诊断提供新的生物标志物。**方法:** 从基因表达综合 (GEO) 数据库中下载AD相关的基因表达数据集GSE125583, 通过差异分析获得差异表达基因 (DEGs), 借助基因本体论 (GO) 功能富集分析和京都基因与基因组百科全书 (KEGG) 信号通路富集分析探讨DEGs的生物学功能及信号通路, 并绘制蛋白-蛋白相互作用 (PPI) 网络, 采用Cytoscape软件和最小绝对收缩和选择算子 (LASSO) 回归、极限梯度提升 (XGBoost) 和随机森林 (RF) 3种ML算法对枢纽 (Hub) 基因进行筛选, 将筛选后的Hub基因通过RF构建AD诊断模型并进行特征重要性排序, 以测试集评价AD诊断模型和关键基因的效能。采用单样本基因集富集分析 (ssGSEA) 对AD组与对照组进行免疫细胞浸润分析。**结果:** 差异分析共筛选出1 287个DEGs。GO功能富集分析, DEGs主要参与神经信号、突触和囊泡等相关的生物学功能; KEGG信号通路富集分析, DEGs主要在离子转运、神经递质和配体门控等通路上富集。3种ML算法共筛选出9个交集Hub基因。AD诊断模型, 对AD诊断性能最高的前4个关键基因分别为腺苷酸环化酶激活多肽1 (ADCYAP1)、脑源性神经营养因子 (BDNF)、血小板衍生生长因子受体 β (PDGFRB) 和趋化因子受体4 (CXCR4), 对应受试者工作特征 (ROC) 的曲线下面积 (AUC) 值分别为0.852、0.795、0.820和0.756; 模型的AUC值为0.828, 准确率为81.25%, 灵敏度为84.40%, 特异度为71.43%。免疫细胞浸润分析, AD组织中巨噬细胞、单核细胞、各种自然杀伤 (NK) 细胞和淋巴细胞浸润程度较高, 其中, NK细胞/自然杀伤T (NKT) 细胞和浆细胞样树突状细胞与4个关键基因显著相关 ($P<0.05$)。**结论:** 基于生物信息学技术与ML算法筛选出的特征基因对AD具有一定的诊断能力, ADCYAP1等基因可能会成为AD诊断的潜在生物标志物, 对AD的早期防治具有重要意义。

[关键词] 生物信息学; 机器学习; 阿尔兹海默症; 诊断模型; 腺苷酸环化酶激活多肽1基因

[中图分类号] R749.16 **[文献标志码]** A

[收稿日期] 2024-10-16 **[录用日期]** 2025-05-08

[基金项目] 国家自然科学基金面上项目 (81872719); 国家自然科学基金青年科学基金项目 (81803337); 山东省科技厅自然科学基金面上项目 (ZR2023MH034); 山东省潍坊市中医药科研立项项目 (第二类) (WFZYY2024-2-001)

[作者简介] 徐林瑞 (2000—), 男, 山东省泰安市人, 在读硕士研究生, 主要从事卫生统计学方面的研究。

[通信作者] 王素珍, 教授 (E-mail: wangsz@sdsu.edu.cn);

王金荣, 副主任护师 (E-mail: wjrwl@126.com)

©《吉林大学学报(医学版)》编辑部, 开放获取遵循CC BY-NC-ND协议。

© Editorial Board of Journal of Jilin University (Medicine Edition). Open access under CC BY-NC-ND license.

Construction of diagnostic model for Alzheimer's disease and immune analysis based on bioinformatics and machine learning

XU Linrui¹, ZHANG Yiyu¹, CUI Jiaqi¹, CONG Xianzhu¹, LI Shuang¹, GE Jiayu¹, KONG Yujia¹,
WANG Suzhen¹, SHI Fuyan¹, WANG Jinrong²

(1. Department of Health Statistics, School of Public Health, Shandong Second Medical University, Weifang 261053, China; 2. Department of Traditional Chinese Medicine, Affiliated Hospital, Shandong Second Medical University, Weifang 261041, China)

ABSTRACT Objective: To screen the Alzheimer's disease (AD)-related genes and construct its diagnostic model using bioinformatics technology and machine learning (ML) algorithms, to discuss the immunological characteristics of AD patients, and to provide novel biomarkers for AD diagnosis. **Methods:** The AD-related gene expression dataset GSE125583 was downloaded from the Gene Expression Omnibus (GEO) database. Differentially expressed genes (DEGs) were identified through differential analysis. Gene Ontology (GO) functional enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway enrichment analyses were performed to explore the biological functions and signaling pathways of DEGs. A protein-protein interaction (PPI) network was constructed, and hub genes were screened using Cytoscape software combined with three ML algorithms: Least Absolute Shrinkage and Selection Operator (LASSO), eXtreme Gradient Boosting (XGBoost), and Random Forest (RF). The screened hub genes were utilized to build an AD diagnostic model via RF, followed by feature importance ranking. The model's efficacy and key genes were evaluated using a test set. Single-sample gene set enrichment analysis (ssGSEA) was used for immune cell infiltration analysis between AD group and control group. **Results:** Differential analysis identified 1 287 DEGs. The GO functional enrichment analysis results revealed that DEGs were primarily involved in biological functions related to neural signaling, synapses, and vesicles. KEGG signaling pathway enrichment analysis indicated significant enrichment of DEGs in ion transport, neurotransmitter, and ligand-gated channel pathways. Nine overlapping hub genes were screened by the three ML algorithms. In the AD diagnostic model, the top four key genes with highest diagnostic performance were adenylyl cyclase-activating polypeptide 1 (*ADCYAP1*), brain-derived neurotrophic factor (*BDNF*), platelet-derived growth factor receptor β (*PDGFRB*), and C-X-C motif chemokine receptor 4 (*CXCR4*), with corresponding area under the curve (AUC) values of 0.852, 0.795, 0.820, and 0.756, respectively. The model achieved an AUC of 0.828, accuracy of 81.25%, sensitivity of 84.40%, and specificity of 71.43%. The immune cell infiltration analysis results demonstrated higher infiltration of macrophages, monocytes, natural killer (NK) cells, and lymphocytes in AD tissue. Among these, NK/natural killer T (NKT) cells and plasmacytoid dendritic cells showed significant correlations with the four key genes ($P < 0.05$). **Conclusion:** The feature genes screened based on bioinformatics and ML exhibit diagnostic potential for AD. Genes such as *ADCYAP1* may serve as potential biomarkers for AD diagnosis, offering significant implications for early prevention and treatment.

KEYWORDS Bioinformatics; Machine learning; Alzheimer's disease; Diagnostic model; adenylyl cyclase-activating polypeptide 1 gene

阿尔兹海默症 (Alzheimer's disease, AD) 是一种慢性神经退行性疾病, 其主要表现为记忆减退及行动和认知功能障碍, 是导致老年痴呆的最普遍

原因^[1]。随着人口老龄化的加剧, AD发病率逐年升高^[2], 预计到2050年, 全球AD患病人数将达到1.52亿, 其中发展中国家的增幅最大^[3-4]。AD发

病机制复杂多样, 包括胆碱能降低、Tau蛋白异常和 β -淀粉样蛋白 (amyloid β -protein, A β) 沉积等多方面^[4-5], 但其作用机制尚未完全明确。目前, AD的诊断方法主要有正电子发射型计算机断层显像 (positron emission computed tomography, PET) 和脑脊液 (cerebrospinal fluid, CSF) 蛋白分析方法, PET诊断准确性虽高, 但因其高成本而未广泛应用, 而CSF蛋白分析因技术时间的限制, 难以实现早期诊断^[6]。因此, 寻找AD诊断的潜在生物标志物十分重要。

生物信息学技术和机器学习 (machine learning, ML) 算法的快速发展为AD的研究带来了新机遇^[7-8]。利用生物信息学技术结合ML算法构建AD诊断模型, 有利于识别AD的潜在生物标志物。基于此, 本研究首先对AD组和对照组的基因表达数据进行差异分析和富集分析, 初步筛选可能与AD相关的差异表达基因 (differentially expressed genes, DEGs), 识别DEGs参与的生物学现象、代谢和信号传导途径等, 揭示AD发生发展过程中的分子机制; 借助蛋白-蛋白相互作用 (protein-protein interaction, PPI) 网络分析寻找DEGs中的枢纽 (Hub) 基因, 采用最小绝对收缩和选择算子 (least absolute shrinkage and selection operator, LASSO) 回归、极限梯度提升 (extreme gradient boosting, XGBoost) 和随机森林 (random forest, RF) 3种ML算法对Hub基因进一步筛选, 取3种算法交集的Hub基因作为AD关键基因, 通过RF构建AD诊断模型并对关键基因进行重要性排序, 以测试集对模型性能进行评价; 利用单样本基因集富集分析评估AD组和对照组样本的免疫细胞浸润情况^[9-10], 并对AD特征基因与单样本基因集富集分析 (single sample Gene Set Enrichment Analysis, ssGSEA) 中的28种免疫细胞亚型进行相关性分析, 为AD的早期筛查、临床诊断和个性化治疗提供参考。

1 资料与方法

1.1 数据来源和处理 AD基因表达数据集GSE125583来源于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 (Gene Expression Omnibus, GEO) 数据库, 该数据集包含219例AD患者和70例神经学正常的冷冻梭形回组织。采用R软件对原始数据集进行标准化、数据格式转

换、基因ID转换和质量控制等处理, 剔除表达水平低或不表达的劣质基因, 按照1:1的近似比例划分为训练集和测试集, 其中, 训练集包含110例AD组织和35例正常组织, 测试集包含109例AD组织和35例正常组织。

1.2 差异分析 采用R软件中的“DESeq2”“edgeR”“limma”3种R软件包分别对训练集的基因表达数据进行组间差异分析, 筛选DEGs, 3种R软件包筛选DEGs的标准: $|\text{Log}_2\text{倍数变化}(\text{fold change, FC})| > 0.5$, 且校正后的 $P < 0.05$ 。其中, FC为AD组与对照组组间基因表达水平的比值, 若 $|\text{Log}_2\text{FC}| > 0.5$ 且 $\text{Adjust } P < 0.05$, 则为上调基因; 若 $|\text{Log}_2\text{FC}| < -0.5$ 且 $\text{Adjust } P < 0.05$, 则为下调基因。

1.3 富集分析 通过R软件中的“clusterProfiler”包^[11]对3种差异分析共同筛选出的交集DEGs进行基因本体论 (Gene Ontology, GO) 功能富集分析和京都基因与基因组百科全书 (Kyoto Encyclopedia of Genes and Genomes, KEGG) 信号通路富集分析。其中, GO功能富集分析按照生物过程 (biological process, BP)、细胞组分 (cell component, CC) 和分子功能 (molecular function, MF) 对DEGs分别进行注释及分类; KEGG信号通路富集分析则是根据信号通路对DEGs进行注释和分类; 2种富集分析的筛选标准均为 $\text{Adjust } P < 0.05$ 。

1.4 PPI网络构建与Hub基因筛选 采用STRING数据库 (<http://cn.string-db.org/>) 对交集DEGs构建PPI网络, 最低要求互作分数设置为0.7, PPI网络可视化采用Cytoscape 3.9.0软件^[12-13], 利用该软件中的CytoHubba插件筛选出边缘渗透组分 (edge percolated component, EPC) 得分最高的前100位基因作为Hub基因, 并对其中前10位Hub基因进行相关性分析。其中EPC通过考虑节点在其邻近区域的连接性来评估该节点的重要性, EPC值越高, 说明该节点连接性越强, 重要性越高。

1.5 基于多种ML算法的Hub基因筛选 对EPC得分前100位的Hub基因, 采用LASSO回归、XGBoost和RF 3种ML算法分别筛选可能与AD相关的基因。其中, 基于R软件中的“glmnet”包构建LASSO模型, 模型中变量系数为0, 则对应基因被剔除; 基于R软件中的“xgboost”包构建

XGBoost模型,以增益(Gain)值衡量基因的重要性,筛选出Gain值最高的前20位Hub基因;基于R软件中的“randomForest”包构建RF模型,以平均基尼指数下降(mean decrease gini, MDG)作为变量重要性排序指标,筛选出MDG最高的前20位Hub基因;对3种ML算法筛选出的Hub基因取交集并借助韦恩图实现可视化。

1.6 AD诊断模型构建和性能评价 以3种ML算法共同筛选出的Hub基因作为自变量,以是否为AD组织作为因变量,构建RF模型。以平均准确度下降(mean decrease accuracy, MDA)和MDG评价Hub基因对AD的贡献程度,将Hub基因作为AD关键基因,基于测试集评价模型的综合性能,以受试者工作特性(receiver operating characteristic, ROC)曲线评估关键基因在AD诊断中的效能并计算模型的准确率、灵敏度和特异度等指标,进一步评价模型性能。

1.7 免疫细胞浸润分析 采用ssGSEA并结合基因表达数据比较AD组织和对照组织中免疫细胞亚型的浸润丰度,对28种免疫细胞亚型进行自相关性分析,以AD诊断模型中的关键基因联合免疫细胞进行相关性分析。基于免疫细胞基因集cellMarker并利用R软件中的“gsva”包量化免疫细胞浸润丰度,以“ggcor”包实现相关性分析可视化。

2 结果

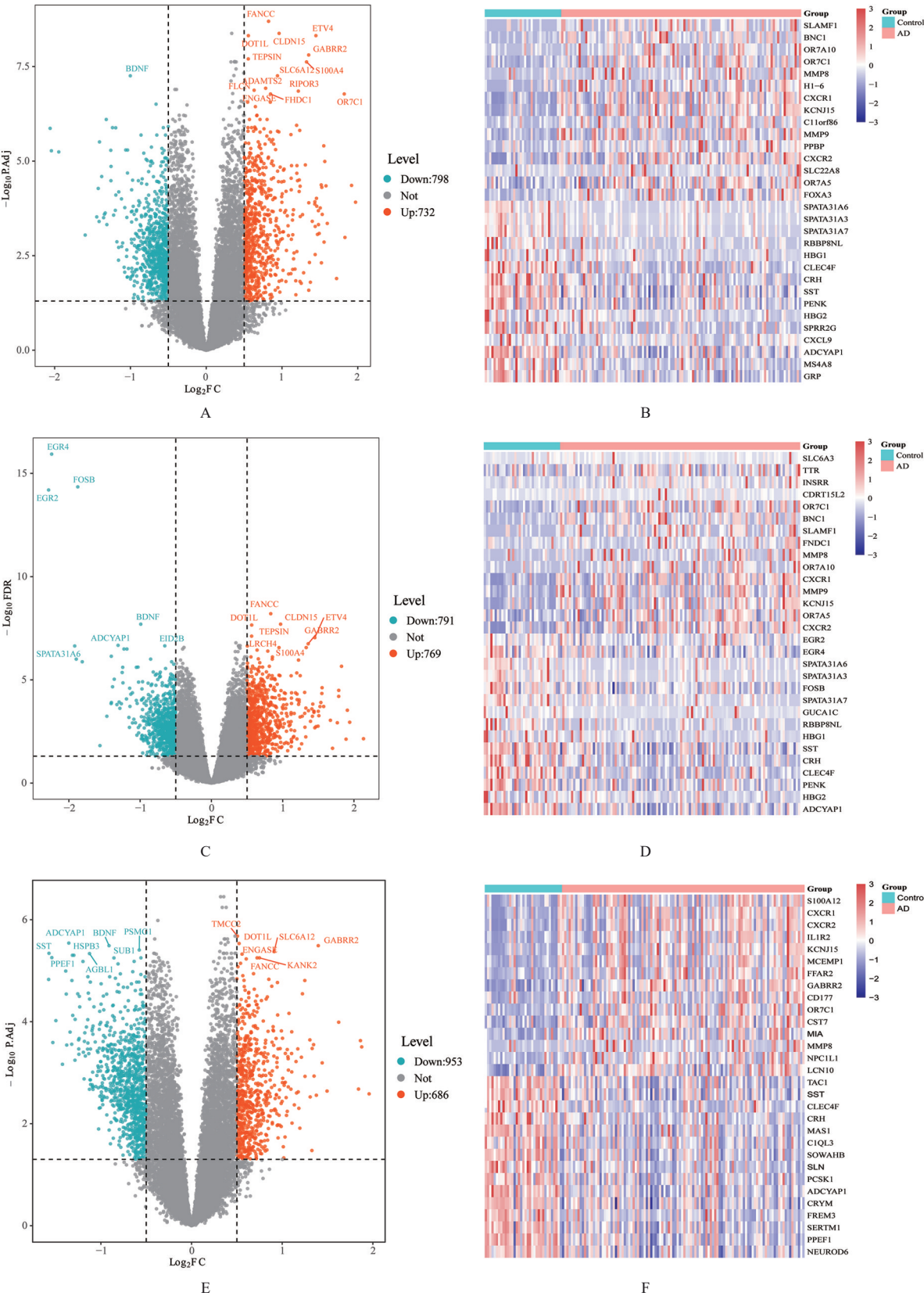
2.1 DEGs的筛选 3种R软件包对DEGs的筛选结果:“DESeq2”包的差异分析共筛选出1 530个DEGs,其中上调基因732个,下调基因798个(图1A和1B);“edgeR”包的差异分析共筛选出1 560个DEGs,其中上调基因769个,下调基因791个(图1C和1D);“limma”包的差异分析共筛选出1 639个DEGs,其中上调基因686个,下调基因953个(图1E和1F);3种R软件包取交集后,共筛选出1 287个DEGs,其中上调基因575个(图2A),下调基因712个(图2B)。

2.2 基于DEGs的GO功能富集分析和KEGG信号通路富集分析 对DEGs分别进行GO功能富集分析和KEGG信号通路富集分析,按照筛选条件得到385个BP、96个CC、80个MF和25个KEGG信号通路。GO功能富集分析结果显示:BP主要为信号释放、跨突触信号传导调节、膜电位调节、突触内囊泡介导的转运和神经递质转运等过程

(图3A);CC主要为谷氨酸能突触、神经元细胞体、突触前膜、胞吐囊膜和 γ -氨基丁酸能突触等组成(图3B);MF主要为单原子离子通道活性、无机阴离子跨膜转运蛋白活性、神经递质受体活性、突触后神经递质受体活性和配体门控单原子阴离子通道活性等功能(图3C)。KEGG信号通路富集分析结果显示:DEGs主要富集在神经活性配体-受体相互作用、磷脂酰肌醇3激酶(phosphatidylinositol 3-kinase, PI3K)/蛋白激酶B(protein kinase B, AKT)信号通路、环磷酸腺苷信号通路、 γ -氨基丁酸能突触通路和5-羟色胺能突触通路等(图3D)。

2.3 PPI网络构建和Hub基因筛选 基于筛选出的DEGs,利用STRING数据库并结合Cytoscape 3.9.0软件对DEGs构建PPI网络(图4A),将EPC得分最高的前100位基因作为Hub基因,EPC得分前10位的基因分别为脑源性神经营养因子(brain derived neurotrophic factor, BDNF)、C-X-C基序趋化因子配体(C-X-C motif chemokine ligand 8, CXCL) 8、胰岛素样生长因子1(insulin-like growth factor 1, IGF1)、转化生长因子 β 1(transforming growth factor beta1, TGFB1)、基质金属蛋白酶9(matrix metalloproteinase 9, MMP-9)、趋化因子受体4(C-X-C motif chemokine receptor 4, CXCR4)、G蛋白亚基 γ 2(G protein subunit gamma 2, GNG2)、白蛋白(albumin, ALB)、生长抑素(somatostatin, SST)和神经肽Y(neuropeptide Y, NPY),前10位Hub基因的PPI网络是由10个节点和24条边组成(图4B),前10位Hub基因之间存在一定的相关性(图4C)。

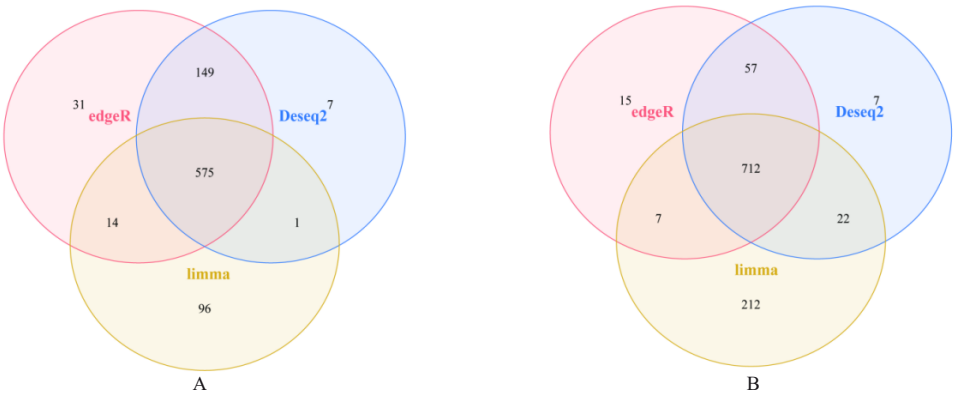
2.4 基于多种ML算法筛选Hub基因 以LASSO回归、XGBoost和RF 3种ML算法对EPC得分前100位的Hub基因进一步筛选。通过R软件绘制LASSO回归的系数路径图和LASSO回归的交叉验证曲线(图5A和图5B),10折交叉验证确定最佳正则化参数 λ (Lambda)为0.006 245 4,对应的 $\text{Log}(\lambda)$ 为-5.075 91。根据 λ 确定LASSO回归模型中变量的回归系数(S),以S绝对值大小反映各Hub基因对AD的重要性,共筛选出成纤维细胞生长因子9(fibroblast growth factor 9, FGF9)和胃泌素释放肽(gastrin releasing peptide, GRP)等39个较为重要的基因(图5D);XGBoost算法筛选过程中,根据Gain值共筛选出IGF2和血小板衍生生长因子受体 β (platelet derived growth factor receptor



A: Volcanic map of DESeq2; B: Heatmap of DEGs based on DESeq2; C: Volcanic map of edgeR; D: Heatmap of DEGs based on edgeR; E: Volcanic map of limma; F: Heatmap of DEGs based on limma.

图1 AD组和对照组样品中DEGs表达水平

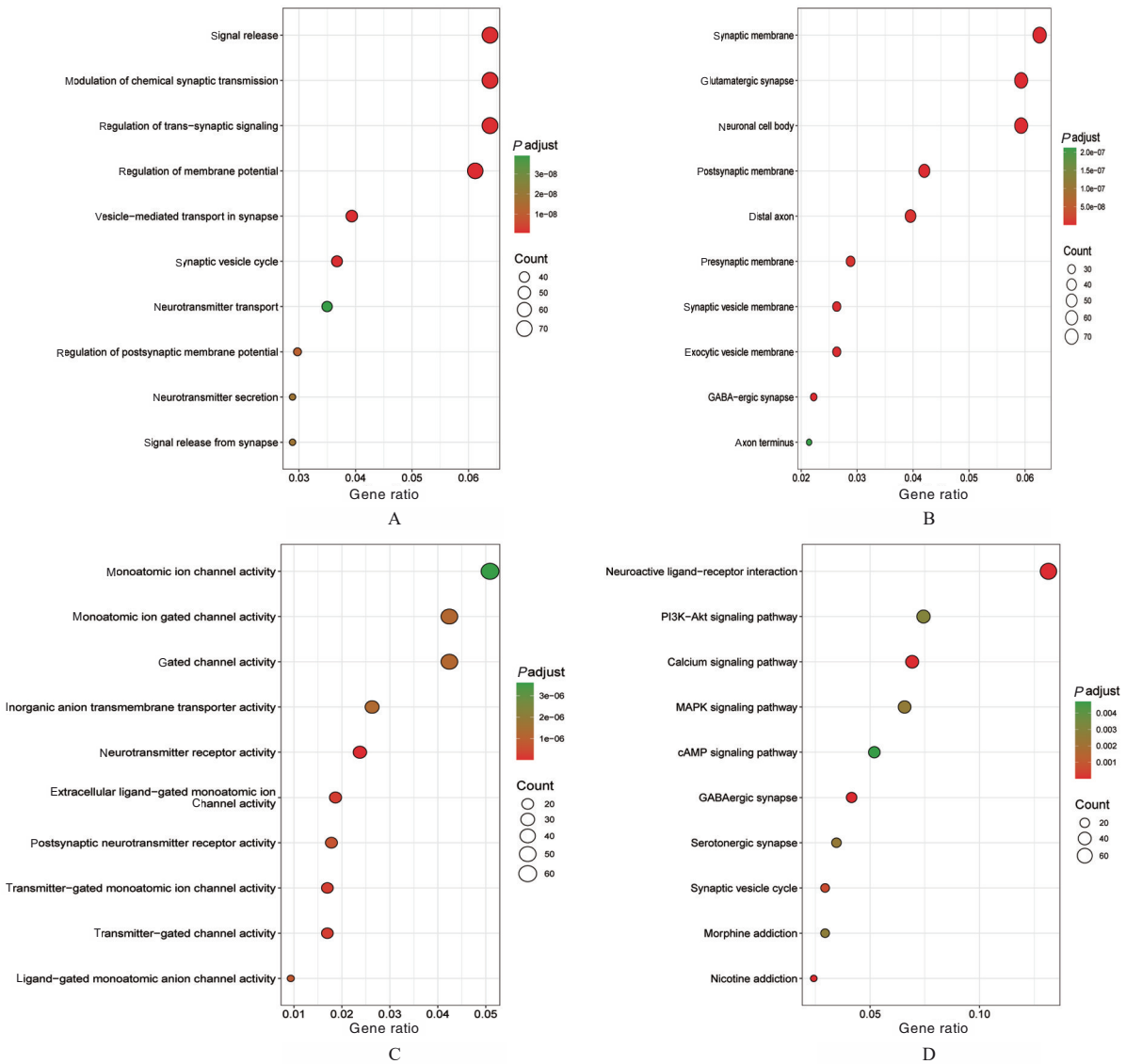
Fig. 1 Expression levels of DEGs in samples in AD and control groups



A: Venn diagram of up-regulated genes; B: Venn diagram of down-regulated genes.

图2 3种R软件包共同筛选到的交集DEGs韦恩图

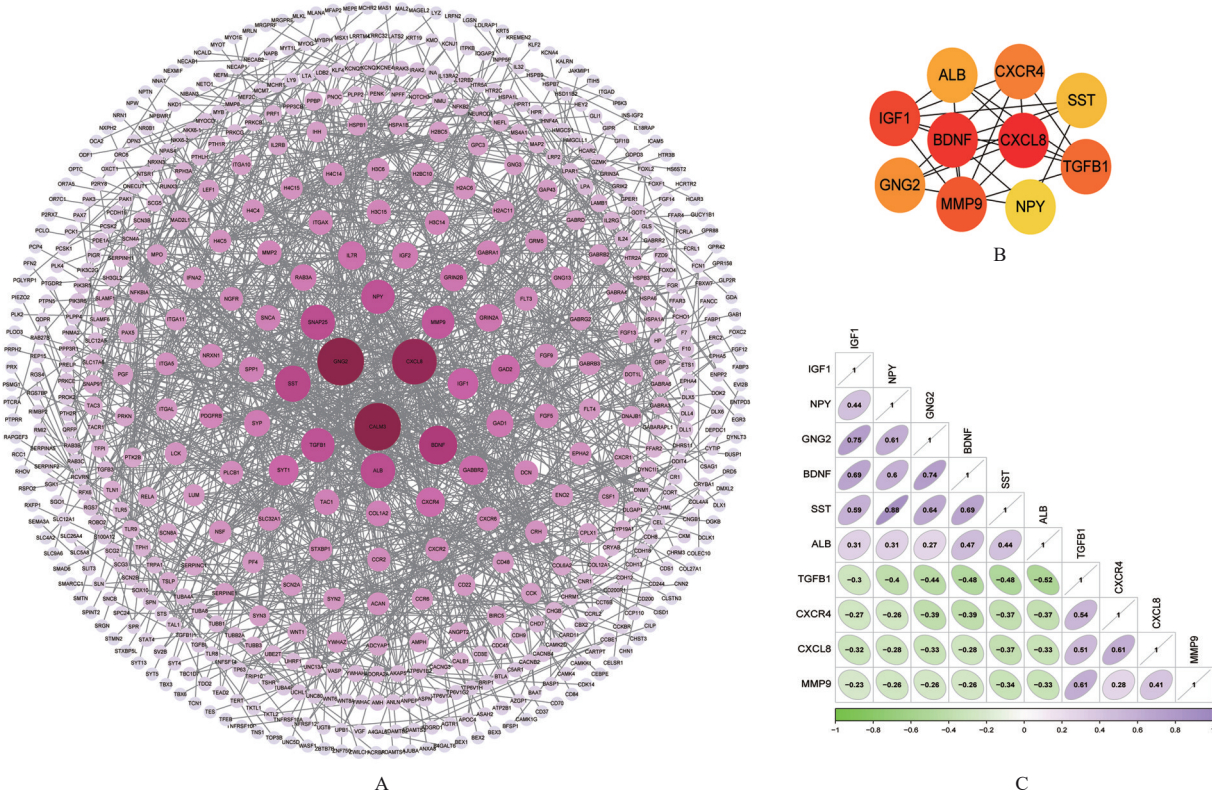
Fig. 2 Venn diagrams of DEGs intersected by three R software packages together



A—C: GO functional enrichment analysis of DEGs; A: BP; B: CC; C: MF; D: KEGG signaling pathway enrichment analysis.

图3 DEGs的GO功能富集分析和KEGG信号通路富集分析

Fig. 3 GO functional and KEGG signaling pathway enrichment analysis on DEGs



A: PPI network of DEGs; B: Top 10 key genes of DEGs in PPI network; C: Correlation analysis on key genes.

图4 DEGs的PPI网络构建和Hub基因相关性分析

Fig. 4 PPI network construction and Hub gene correlation analysis on DEGs

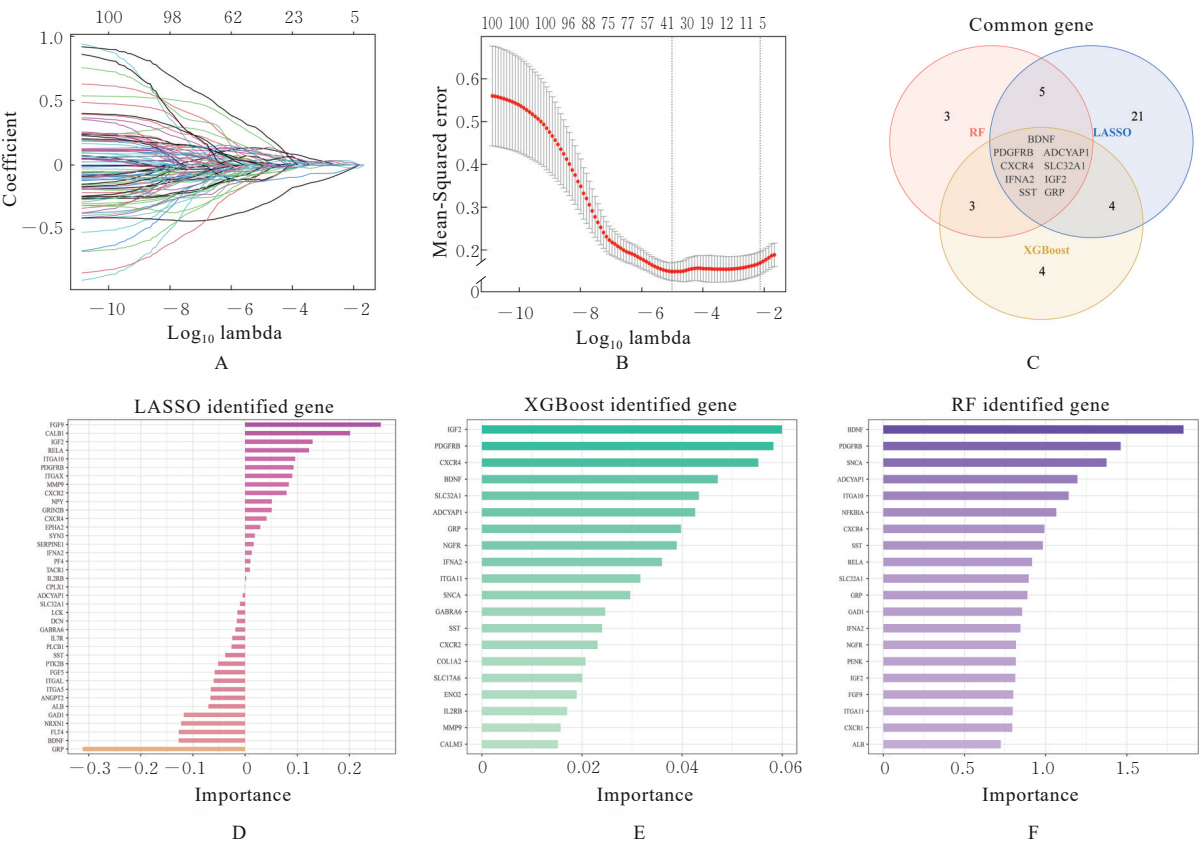
beta, *PDGFRB*) 等 20 个基因 (图 5E); RF 算法筛选中, 根据 MDG 值共筛选出 *BDNF* 和 *PDGFRB* 等 20 个基因 (图 5F); 3 种 ML 算法共筛选出 9 个交集 Hub 基因, 分别为 *BDNF*、*PDGFRB*、腺苷酸环化酶激活多肽 1 (adenylate cyclase activating polypeptide 1, *ADCYAP1*)、*CXCR4*、溶质载体家族 32 成员 1 (solute carrier family 32 member 1, *SLC32A1*)、干扰素 $\alpha 2$ (interferon Alpha 2, *IFNA2*)、*IGF2*、*SST* 和 *GRP* (图 5C)。

2.5 AD 诊断模型的构建及评价 将 9 个 Hub 基因作为 AD 诊断的关键基因, 采用 RF 算法构建 AD 诊断模型。利用 10 折交叉验证, 确定最优 mtry (指定节点中用于二叉树的变量数) 为 2, 最佳 ntree (构成森林的决策树的个数) 为 500; MDG 和 MDA 均位列前 5 位的关键基因有 4 个: *BDNF*、*ADCYAP1*、*PDGFRB* 和 *CXCR4* (图 6A 和 6B); 模型的 AUC 值为 0.828, 准确率为 81.25% (图 6C), 灵敏度为 84.40%, 特异度为 71.43% (图 6D); 关键基因 *BDNF*、*ADCYAP1*、*PDGFRB*

和 *CXCR4* 的 AUC 值分别为 0.795、0.852、0.820 及 0.756 (图 7A~7D)。

2.6 AD 组和对照组脑组织的免疫细胞浸润情况

与对照组织比较, AD 组中活化 B 淋巴细胞、活化 CD4+T 淋巴细胞、CD56 强 NK/CD56 弱阳性/NK 细胞、中央记忆 CD4+/CD8+T 淋巴细胞、效应记忆 CD8+T 淋巴细胞、 $\gamma\delta$ T 细胞、幼稚 B 细胞、巨噬细胞、肥大细胞、髓源抑制性细胞、记忆 B 细胞、单核细胞、NK/NKT 细胞、中性粒细胞、浆细胞样树突细胞 (dendritic cell, DC)、调节性 T 淋巴细胞、1 型辅助性 T 细胞/17 型辅助性 T 细胞浸润程度较高 ($P < 0.05$), 而效应记忆 CD4+T 淋巴细胞及 Th2 细胞浸润程度较低 ($P < 0.05$) (图 8A); 免疫细胞自相关分析结果显示: 中央记忆 CD4+/CD8+T 淋巴细胞与骨髓源性抑制细胞和 NK 细胞相关性较强, 单核细胞、Th2 细胞和嗜酸性细胞与其他免疫细胞相关性较弱 (图 8B); *ADCYAP1*、*CXCR4*、*PDGFRB* 和 *BDNF* 与 NK/NKT 细胞和浆细胞样树突状细胞均呈显著正相关关系 ($P < 0.01$) (图 8B)。



A: A coefficient profile plot was produced against the Log₁₀ Lambda sequence; B: Optional process diagram of cross-validation parameter (λ). The abscissa is the number of non-zero coefficients in the model. As lambda increases, Mean-Squared error of the model gradually decreases, 39 genes with nonzero coefficients were selected by optimal lambda(λ); C: Venn Diagram of LASSO, RF, and XGBoost ML algorithms for screening Hub genes; D: LASSO regression screening for Hub genes; E: XGBoost screening for Hub genes; F: Random Forest screening for Hub genes.

图5 基于多种ML算法的Hub基因筛选
Fig. 5 Hub gene screening based on multiple ML algorithms

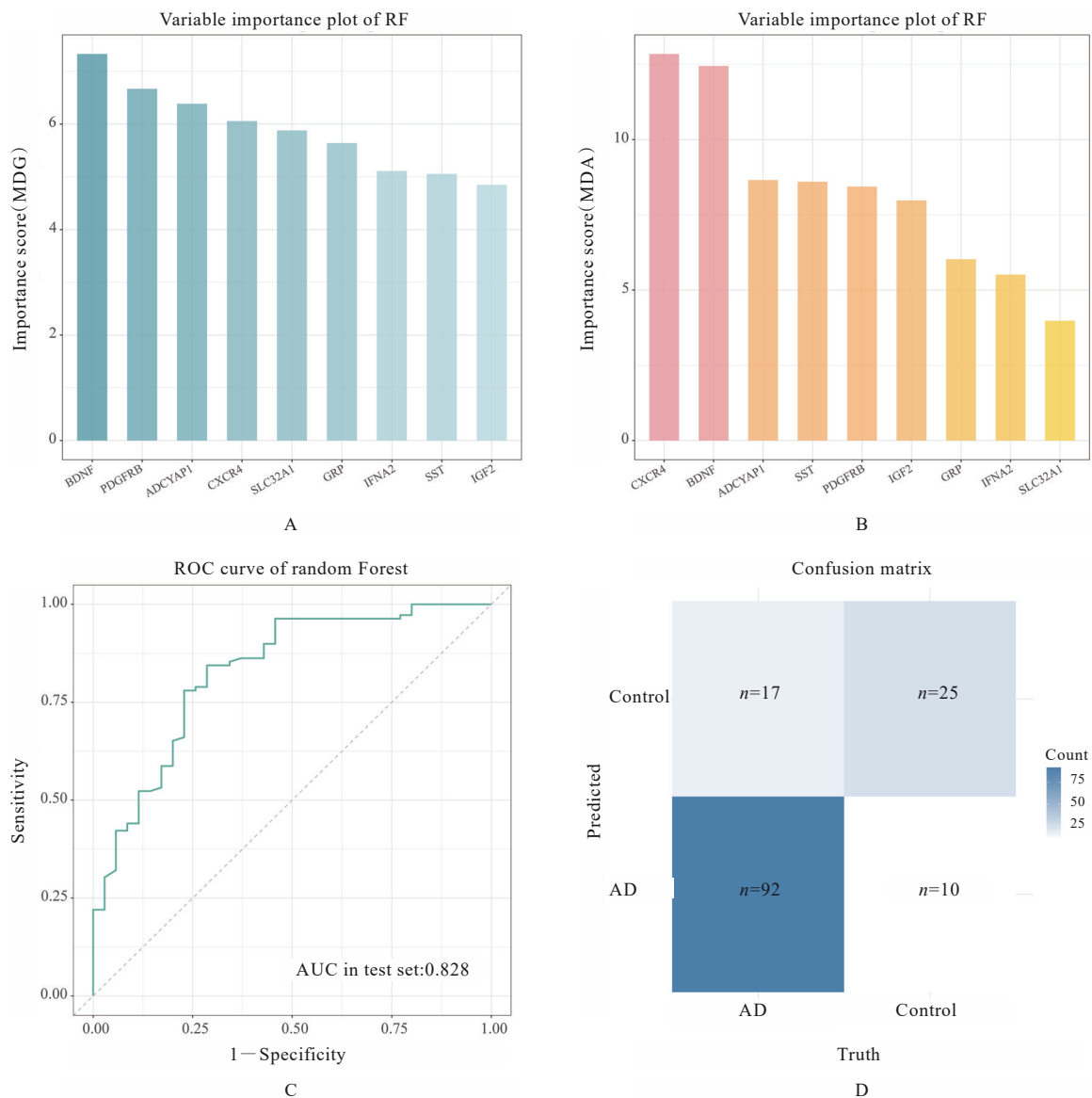
3 讨论

AD在痴呆类病例中占60%~80%，是最常见的痴呆类疾病^[14]。迄今为止，影响AD发病的因素十分复杂，而AD发病风险60%~80%取决于遗传因素^[15]。本研究从遗传学角度出发，利用生物信息学技术和ML算法对AD组及对照组的转录组数据进行分析，构建AD诊断模型，筛选关键基因，发掘AD相关的潜在分子标志物。

本研究采用“DESeq2”“edgeR”“limma”3种差异分析R软件包来筛选DEGs^[16-18]，综合3种R软件包的优势，提高DEGs筛选的可靠性和稳定性。最终共筛选出1287个DEGs，其中上调575个，下调712个，将其用于后续分析以降低筛选假阳性，提高准确性。

为揭示DEGs的生物学功能和信号通路，本研究对1287个DEGs分别进行了GO功能富集分析

和KEGG信号通路富集分析。GO功能富集分析结果显示：DEGs主要参与神经信号传递、神经递质转运、突触功能和膜电位调节等BP。有研究者^[19]指出：谷氨酸作为人脑中兴奋性神经递质，在介导认知和行为功能等方面具有重要作用，而谷氨酸能突触的丧失被认为与AD的临床症状密切相关；DEGs主要定位在囊泡相关成分、突触相关成分和神经元特定成分等CC。研究^[20]显示：小鼠模型中囊泡谷氨酸转运蛋白1（vesicular glutamate transporters 1, VGLUT1）表达明显增加，具有广泛的A β 沉积，而A β 沉积正是AD的主要标志之一，提示囊泡及突触相关DEGs可能与AD发病机制高度相关；DEGs主要富集在跨膜离子转运功能、神经递质和配体门控的信号传导功能上。研究者^[21]分析了AD组和正常对照组研究对象的小脑转录谱，结果显示：活性离子跨膜转运蛋白活性基



A: Sequencing of gene characteristic importance based on Mean Decrease Accuracy (MDA) in Random Forest; B: Sequencing of gene characteristic importance based on Mean Decrease Gini (MDG) in Random Forest; C: Receiver Operating Characteristic Curve of AD diagnostic model; D: Confusion matrix.

图6 AD诊断模型构建和性能评价

Fig. 6 Construction and performance evaluation of AD diagnostic model

因的过度表达，可能会导致细胞内外离子浓度失衡，进而造成肌肉收缩、神经传导过程紊乱，最终引起AD。本研究中KEGG信号通路富集分析结果显示：DEGs主要调控神经递质通路、细胞内信号转导通路和离子信号通路。研究^[22]显示：PI3K-AKT等细胞内信号通路可以作为防治AD的天然靶点。

为构建AD诊断模型，本研究首先基于DEGs构建PPI网络，初步筛选Hub基因，随后以LASSO回归、XGBoost和RF对Hub基因进一步筛选，最

终，利用RF构建AD诊断模型并对关键基因进行重要性排序，以测试集评价模型性能，结果显示：BDNF、ADCYAP1、PDGFRB和CXCR4均位列MDA及MDG排序的前5位，4个关键基因和模型的AUC均大于0.75，模型的准确率为81.25%，灵敏度为84.40%，特异度为71.43%，表明模型的诊断性能良好，关键基因对AD具有一定的诊断能力。ADCYAP1是一种腺苷酸环化酶激活肽，其受体可激活环磷酸腺苷（cyclic adenosine monophosphate, cAMP）信号通路介导的蛋白酶

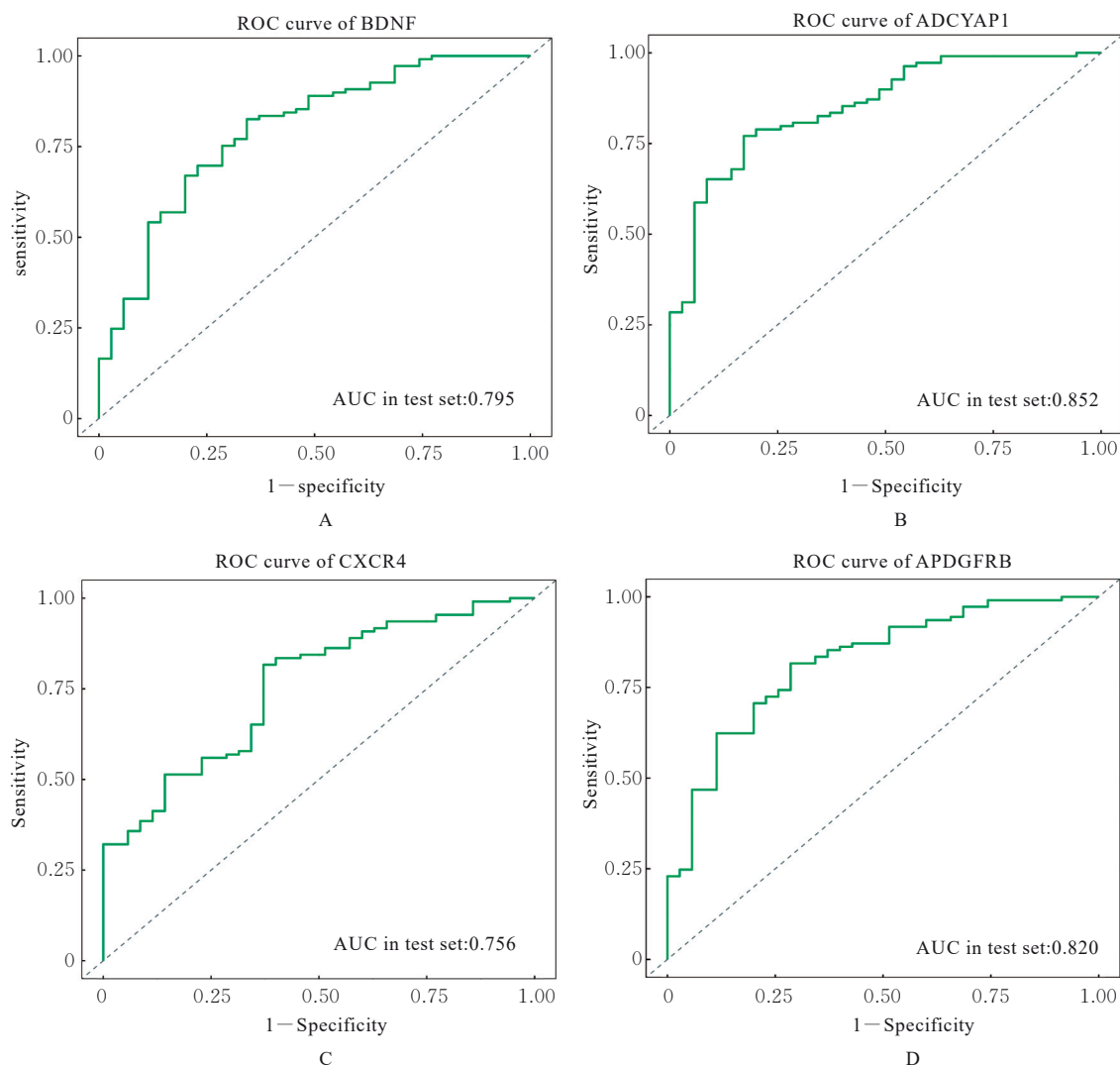


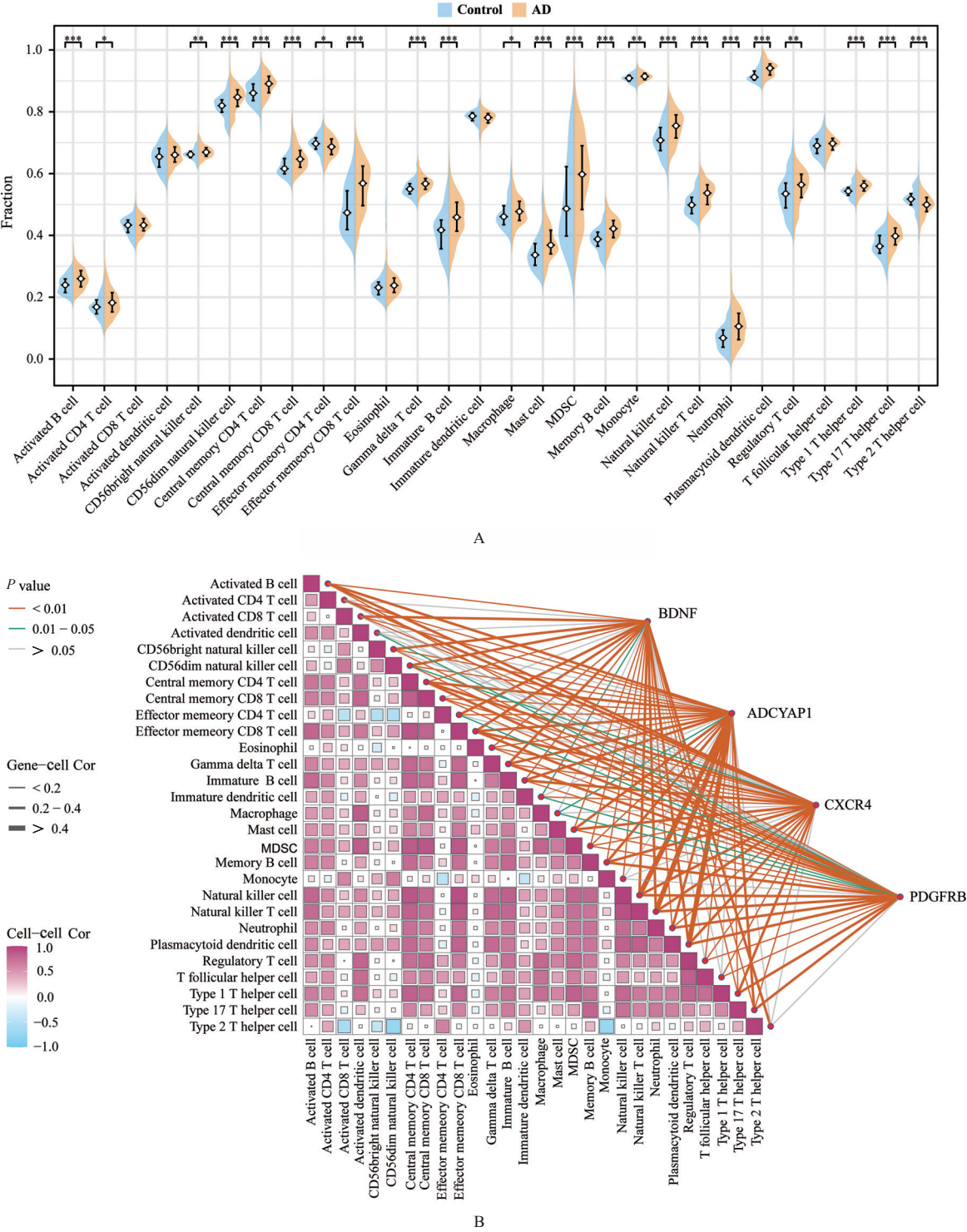
图7 关键基因的AD诊断能力

Fig. 7 Diagnostic capacities of key genes for AD

体活性,促进Tau蛋白水解,可能成为AD和Tau蛋白病变治疗的新靶点^[23]。研究^[24]显示:ADCYAP1编码的垂体腺苷酸环化酶激活肽(pituitary adenylate cyclase activating polypeptide, PACAP)具有调节神经递质和抗氧化应激的作用,其通过调节ADCYAP1表达水平来影响自身的生成,故在缓解AD方面具有一定的治疗价值。BDNF属于神经营养因子蛋白质家族的一员,研究^[25-27]显示:BDNF可通过酪氨酸蛋白激酶B(tyrosine protein kinase B, TrkB)激活PI3K/AKT等信号通路来抑制神经元凋亡,促进突触的形成和重塑,增强突触传递效率,对AD具有保护作用;PDGFRB是指血小板衍生生长因子受体 β ,其可调节脑内血管生成和修复,通过测量CSF中可溶性PDGFRB水平来衡量血脑屏障(blood-brain

barrier, BBB)的受损程度,是判断认知功能障碍的生物标志物之一^[28];CXCR4属于CXC趋化因子受体家族,与特异性配体CXCL12结合可激活AKT和cAMP响应元件结合蛋白(cAMP responsive element binding protein, CREB)等,调节细胞存活和增殖、神经保护及可塑等,从而影响衰老及AD,故CXCR4可能会成为治疗AD的新靶点和生物标志物^[29]。

为进一步探讨AD患者的免疫学特征,本研究进行了免疫细胞浸润分析,结果显示:与对照组脑组织比较,AD组脑组织中多数免疫细胞活性显著上调,包括巨噬细胞、单核细胞、辅助性T细胞、各种NK细胞和淋巴细胞等,表明在AD进展中可能伴随着免疫系统的激活。研究^[30-31]证实:在多种AD小鼠模型中,外周1型和外周17型辅助性细



A: Immune cell infiltration analysis results of AD group and normal controls group in GSE125583 dataset(* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$);
B: Correlation and auto-correlation analysis between 28 immune cell subtypes and characteristic genes of AD.

图8 免疫细胞浸润分析

Fig. 8 Analysis on immune cell infiltration

胞与炎症细胞因子的释放有关。研究^[32]显示：NK细胞可以产生白细胞介素（interleukin, IL）3、IL-5和IL-10等细胞因子，而IL-10的稳定对帕金森

森症、骨关节炎和AD等炎症性疾病至关重要，因此NK细胞可能通过其释放的细胞因子影响AD的进展。研究^[33-34]显示：AD患者中，单核-巨噬细

胞通过BBB进入脑组织,参与清除A β 斑块,但持续的单核-巨噬细胞活化可能会导致慢性炎症状态,从而促进A β 的沉积和神经元损伤。因此,单核-巨噬细胞在AD进展中可能具有双重作用:既参与病理物质的清除,也可能通过引发慢性炎症而加剧疾病进展。

综上所述,本研究利用生物信息学技术和ML算法逐步筛选AD相关*Hub*基因,并进一步构建AD诊断模型,挖掘AD相关的关键基因,探讨AD患者的免疫学特征,为其诊断提供了新的分子标志物。然而,本研究也存在局限性,首先,病例与对照组的比例不均衡,可能会影响基因筛选的准确性;其次,本研究是基于公共数据库进行的,缺乏关于*Hub*基因的验证性实验。因此,仍需要采用比例适宜的随机对照试验并结合临床数据验证,以期AD的诊断和防治提供更有力的依据。

利益冲突声明:

所有作者声明不存在利益冲突。

作者贡献声明:

徐林瑞参与研究选题、数据分析和论文撰写,张译予、崔家齐、丛显铸、李爽和葛佳瑜参与数据整理、论文修改及图表编辑,孔雨佳、王素珍、石福艳和王金荣参与研究设计及论文审阅。

参考文献

- [1] LANE C A, HARDY J, SCHOTT J M. Alzheimer's disease[J]. Eur J Neurol, 2018, 25(1): 59-70.
- [2] KHAN S, BARVE K H, KUMAR M S. Recent advancements in pathogenesis, diagnostics and treatment of Alzheimer's disease[J]. Curr Neuroparmacol, 2020, 18(11):1106-1125.
- [3] HU Y, ZHOU G, ZHANG C, et al. Identify compounds' target against Alzheimer's disease based on in-silico approach[J]. Curr Alzheimer Res, 2019, 16(3): 193-208.
- [4] TWAROWSKI B, HERBET M. Inflammatory processes in Alzheimer's disease-pathomechanism, diagnosis and treatment: A review[J]. Int J Mol Sci, 2023, 24(7): 6518.
- [5] BREIJYEH Z, KARAMAN R. Comprehensive Review on Alzheimer's disease: causes and treatment[J]. Molecules, 2020, 25(24): 5789.
- [6] BATEMAN R J, BARTHÉLEMY N R, HORIE K. Another step forward in blood-based diagnostics for Alzheimer's disease[J]. Nat Med, 2020, 26(3): 314-316.
- [7] MERKIN A, KRISHNAMURTHI R, MEDVEDEV O N. Machine learning, artificial intelligence and the prediction of dementia[J]. Curr Opin Psychiatry, 2022, 35(2): 123-129.
- [8] SUN X, ZHANG H, YAO D, et al. Integrated bioinformatics analysis identifies *Hub* genes associated with viral infection and Alzheimer's disease[J]. J Alzheimers Dis, 2022, 85(3): 1053-1061.
- [9] HUANG R, WANG W, CHEN Z, et al. Identifying immune cell infiltration and effective diagnostic biomarkers in Crohn's disease by bioinformatics analysis[J]. Front Immunol, 2023, 14: 1162473.
- [10] WANG W, CHEN Z, HUA Y. Bioinformatics prediction and experimental validation identify a novel cuproptosis-related gene signature in human synovial inflammation during osteoarthritis progression[J]. Biomolecules, 2023, 13(1): 127.
- [11] WU T, HU E, XU S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data[J]. Innovation(Camb), 2021, 2(3): 100141.
- [12] MOUSAVIAN Z, KHODABANDEH M, SHARIFIZARCHI A, et al. StrongestPath: a Cytoscape application for protein-protein interaction analysis[J]. BMC Bioinformatics, 2021, 22(1): 352.
- [13] 时 承, 姚旭峰. 机器学习用于帕金森病诊断的研究进展[J]. 中国医学物理学杂志, 2024, 41(5): 640-645.
- [14] 龚玉芳, 邓海霞, 卢 燕, 等. 基于生物信息学建立与验证心肌梗死铁死亡基因诊断模型及免疫分析[J/OL]. 中国胸心血管外科临床杂志: 1-10[2025-08-10]. <https://link.cnki.net/urlid/51.1492.R.20240321.1137.030>.
- [15] ROSTAGNO A A. Pathogenesis of Alzheimer's disease[J]. Int J Mol Sci, 2022, 24(1): 107.
- [16] SCHELTENS P, DE STROOPER B, KIVIPELTO M, et al. Alzheimer's disease[J]. Lancet, 2021, 397(10284): 1577-1590.
- [17] HU R, ZHANG W, HAN Z, et al. Identification of immune-related target and prognostic biomarkers in PBMC of hepatocellular carcinoma [J]. BMC Gastroenterol, 2023, 23(1): 234.
- [18] LIU X, ZHAO J, XUE L, et al. A comparison of transcriptome analysis methods with reference genome[J]. BMC Genomics, 2022, 23(1): 232.
- [19] SOFER T, KURNIANSYAH N, AGUET F, et al. Benchmarking association analyses of continuous exposures with RNA-seq in observational studies [J]. Brief Bioinform, 2021, 22(6): bbab194.
- [20] PINKY P D, PFITZER J C, SENFELD J, et al.

- Recent insights on glutamatergic dysfunction in Alzheimer's disease and therapeutic implications[J]. *Neuroscientist*, 2023, 29(4): 461-471.
- [21] TIMMER N M, METAXAS A, VAN DER STELT I, et al. Cerebral level of vGlut1 is increased and level of glycine is decreased in TgSwDI mice[J]. *J Alzheimers Dis*, 2014, 39(1): 89-101.
- [22] D'ANGIOLINI S, BASILE M S, MAZZON E, et al. In silico analysis reveals the modulation of ion transmembrane transporters in the cerebellum of Alzheimer's disease patients[J]. *Int J Mol Sci*, 2023, 24(18): 13924.
- [23] LONG H Z, CHENG Y, ZHOU Z W, et al. PI3K/AKT signal pathway: a target of natural products in the prevention and treatment of Alzheimer's disease and Parkinson's disease[J]. *Front Pharmacol*, 2021, 12: 648636.
- [24] SCHALER A W, RUNYAN A M, CLELLAND C L, et al. PAC1 receptor-mediated clearance of tau in postsynaptic compartments attenuates tau pathology in mouse brain[J]. *Sci Transl Med*, 2021, 13(595): eaba7394.
- [25] SZEGE CZKI V, HORVÁTH G, PERÉNYI H, et al. Alzheimer's disease mouse as a model of testis degeneration[J]. *Int J Mol Sci*, 2020, 21(16): 5726.
- [26] CHOI S H, BYLYKBASHI E, CHATILA Z K, et al. Combined adult neurogenesis and BDNF mimic exercise effects on cognition in an Alzheimer's mouse model[J]. *Science*, 2018, 361(6406): eaan8821.
- [27] COLUCCI-D'AMATO L, SPERANZA L, VOLPICELLI F. Neurotrophic factor BDNF, physiological functions and therapeutic potential in depression, neurodegeneration and brain cancer[J]. *Int J Mol Sci*, 2020, 21(20): 7777.
- [28] GAO L, ZHANG Y, STERLING K, et al. Brain-derived neurotrophic factor in Alzheimer's disease and its pharmaceutical potential[J]. *Transl Neurodegener*, 2022, 11(1): 4.
- [29] SHI H, KORONYO Y, RENTSENDORJ A, et al. Identification of early pericyte loss and vascular amyloidosis in Alzheimer's disease retina[J]. *Acta Neuropathol*, 2020, 139(5): 813-836.
- [30] WANG Q L, FANG C L, HUANG X Y, et al. Research progress of the CXCR4 mechanism in Alzheimer's disease[J]. *Ibrain*, 2022, 8(1): 3-14.
- [31] BRYSON K J, LYNCH M A. Linking T cells to Alzheimer's disease: from neurodegeneration to neurorepair[J]. *Curr Opin Pharmacol*, 2016, 26: 67-73.
- [32] CAO W, ZHENG H. Peripheral immune system in aging and Alzheimer's disease[J]. *Mol Neurodegener*, 2018, 13(1): 51.
- [33] MOLLAZADEH H, CICERO A F G, BLESSO C N, et al. Immune modulation by curcumin: The role of interleukin-10[J]. *Crit Rev Food Sci Nutr*, 2019, 59(1): 89-101.
- [34] TAYLOR X, CLARK I M, FITZGERALD G J, et al. Amyloid- β (A β) immunotherapy induced microhemorrhages are associated with activated perivascular macrophages and peripheral monocyte recruitment in Alzheimer's disease mice[J]. *Mol Neurodegener*, 2023, 18(1): 59.