·论　　著·

基于 GEO 数据库筛选阿尔茨海默病的关键基因*

李方舟[1],郗雪艳[1],杜伯雨[1,2],戴文敬[1△]
(湖北医药学院:1. 生物医药研究院;2. 基础医学院,湖北 十堰 442000)

[摘　要]　**目的**　应用生物信息学方法筛选阿尔茨海默病的关键基因。**方法**　从 GEO 数据库下载 GSE5281 和 GSE138260 数据集,用 R 语言对数据集进行去批次效应;WGCNA 分析构建共表达网络,筛选与疾病相关基因模块并绘制基因聚类图和相关性图;以 WGCNA 筛选出的模块基因为对象进行差异表达基因分析,绘制火山图及热图;对差异表达基因进行 Lasso 回归分析,筛选关键基因;对差异表达基因进行京都基因和基因组数据库(KEGG)和基因本体(GO)富集分析,绘制气泡图。**结果**　通过 WGCNA 分析获得包含 704 个基因的 brown 模块与疾病高度相关;brown 模块差异表达分析得到 39 个差异表达基因,其中 10 个下调基因,29 个上调基因;进一步 Lasso 回归后筛选出 9 个关键基因。GO 和 KEGG 富集分析表明,差异表达基因在矿物质元素的吸收、氧化还原驱动的活性跨膜转运蛋白等方面富集。**结论**　GEO 数据库初步筛选出潜在的阿尔茨海默病关键基因,但还需进一步实验验证。

[关键词]　阿尔茨海默病;　生物信息学;　GEO 数据库;　差异表达基因;　富集分析
DOI:10.3969/j.issn.1009-5519.2023.17.003　　**中图法分类号**:Q-31
文章编号:1009-5519(2023)17-2893-06　　**文献标识码**:A

Screening for the key genes of Alzheimer's disease based on GEO database*

LI Fangzhou[1],*XI Xueyan*[2],*DU Boyu*[1,2],*DAI Wenjing*[1△]
(1. *Institute of Biological Medicine*;2. *School of Basic Medical Sciences*,*Hubei Medical University*,*Shiyan*,*Hubei* 442000,*China*)

[**Abstract**]　**Objective**　To screen key genes in Alzheimer's disease by bioinformatics method. **Methods**　The data of GSE5281 and GSE138260 were downloaded from the Gene Expression Omnibus(GEO) database,and the R software was used to carry out batch effect on the data set. The co-expression network was constructed by WGCNA analysis,and the disease-related gene modules were screened and the gene cluster and correlation maps were made. The differentially expressed genes(DEGs) were analyzed by using the module genes screened by WGCNA,and volcano map and heatmap were drawn. The DEGs were analyzed by Lasso regression and the key genes were screened. Then Kyoto Encyclopedia of Genes and Genomes(KEGG) and gene ontology(GO) enrichment analysis of DEGs were performed and bubble charts were drawn. **Results**　The brown gene module which include 704 genes correlated with Alzheimer's disease through WGCNA analysis. A total of 39 DEGs were identified by brown module differential expression analysis,including 10 down-regulated genes and 29 up-regulated genes. A total of 9 key genes were screened by Lasso analysis. GO and KEGG enrichment analysis showed that these DEGs were significantly enriched in mineral absorption,oxidoreduction-driven active transmembrane transporter,etc. **Conclusion**　Potential key genes of Alzheimer's disease have been preliminarily screened out based on GEO database,but need further experimental verification.

[**Key words**]　Alzheimer's disease;　Bioinformatics;　Gene Expression Omnibus database;　Differentially expressed genes;　Enrichment analysis

阿尔茨海默病(AD)是一种由阿诺斯·阿尔茨海默(Alois Alzheimer)发现、埃米尔·克雷佩林(Emil Kraepelin)命名的常见并伴随缓慢进展的神经性退行痴呆[1-2]。AD 主要以淀粉样 β 肽沉聚在大脑最容易

*　**基金项目**:湖北医药学院 2021 启动金项目(2021QDJZR026)。
作者简介:李方舟(1995—),硕士研究生,助理实验师,主要从事生物化学与分子生物学方面的研究。　△　**通信作者**,E-mail:645619727@qq.com。

受影响的部位，例如大脑内侧颞叶、皮层而形成的神经斑块和神经纤维缠结为特征[3]。AD 是一个全球性的健康难题，影响着全世界范围内近 5 000 万人口的健康，是造成人类痴呆的主要原因。根据预测，AD 的患病人数将会在 10 年之后翻倍，并在 2050 年达到近 1.5 亿[4-5]。因此，AD 的诊断与治疗方法的开发应用变得尤为重要。

AD 的病理特征主要分为两大类：(1)通过积聚而造成的正向损伤，如神经纤维缠结、淀粉样斑块和其他在 AD 患者脑内发现的沉积物；(2)由于萎缩而造成的负向损伤，如神经细胞、轴突、树突、海马体等大面积萎缩[6-8]。

到目前为止，关于 AD 的发病进展及发病机制提出了一些假设，但具体病因和疾病进展机制还有待证明。关于 AD 的病因提出了 2 条主要的假说，胆碱成因假说和淀粉样蛋白成因假说。胆碱成因假说认为胆碱功能受损是造成 AD 的关键因素；淀粉样蛋白成因则认为淀粉样蛋白生成及修饰过程中有异，产生淀粉样蛋白异构体是 AD 的主要发病原因[5,9-10]。作为一种多因素疾病，其病程的进展与多种风险因素相关，如年龄的增长、遗传、头部损伤、血管疾病、细菌或病毒的感染、重金属等环境因素等[5]。其中最为主要的风险因素是年龄，绝大多数 AD 患者的年龄均在 65 岁以上，年轻人(30 岁左右)除非是家族遗传性 AD，否则基本不会患有这种疾病[11]。衰老是迟发性 AD 的最大危险因素，占 AD 病例的 95%以上。但近期确诊 1 例排除已有基因突变和家族性 AD 的 19 岁 AD 患者，表明 AD 将不再局限于老年人[12]。

截至目前，AD 没有治愈的方法，只有一些改善症状的治疗手段[13-14]。而最大限度地减轻 AD 对患者损害的方法是在 AD 进展为轻度症状前给予患者神经性保护的药物[15]。所以对潜在 AD 患者的早期诊断是缓解疾病症状极为关键的影响因素。2011 年美国国家衰老研究所阿尔茨海默病协会提出了新的诊断标准，这包括临床症状及生物标志物的共同诊断[5]。AD 有 2 类生物标志物：(1)可以通过正电子成像术和脑脊液中检测到的脑淀粉样蛋白标志物；(2)神经元损伤标志物，如脑脊液 tau 蛋白、与代谢相关的氟脱氧葡萄糖(FDG)及通过核磁成像技术直接观测到的大脑萎缩等[16-18]。

过去的生物信息学分析仅仅分析筛选了 AD 的差异性表达基因作为 AD 的诊断标志物，如 EGFR、CD44、BCL2L1、HGG4、LPP、CTAGE 等[19-20]。为了进一步了解 AD 的发病原因及发病机制，发掘 AD 的特征标志物，提高诊断效率，本研究综合 WGCNA、差异性分析及 Lasso 回归分析，基于 GEO 数据库中 AD 患者组及对照组基因表达图谱，用 R 语言更准确地筛选 AD 的关键基因及信号通路，以达到初步筛选 AD 关键基因、开阔疾病的诊断思路、开发有效治疗方法的目的。

1　资料与方法

1.1　数据来源及去批次　研究所用数据来源于美国国立生物技术中心的 GEO 数据库。以“Alzheimer′s disease”“Homosapiens”检索高通量测序数据集。筛选出注释平台分别为 GPL570、GPL27556 的 2 个数据集 GSE5281 和 GSE138260。运用 R 语言对 2 个数据集中的数据进行 ID 转换、数据合并，并对 GEO 2 组数据集进行去批次运算，以去除 2 组数据的批次效应，增加接下来生物信息学分析的准确性。

1.2　WGCNA 筛选与疾病相关基因　用 R 语言进行 WGCNA 分析，排除异常信息及异常样本，构建共表达网络，将基因分为不同的模块。不同基因模块与表型数据关联分析，计算筛选出与患病相关性最高的基因模块。输出这个模块基因的表达数据集以进行后续生物信息学分析。

1.3　基因表达差异性分析　用 R 语言对数据按照表型进行分组后，进行表达差异性分析，筛选出 $|\log FC|>1.2$ 且 $P<0.05$ 的差异表达基因。

1.4　Lasso 回归进一步筛选关键基　应用 R 语言对筛选出的差异表达基因进行 Lasso 回归筛选与表型相关基因。

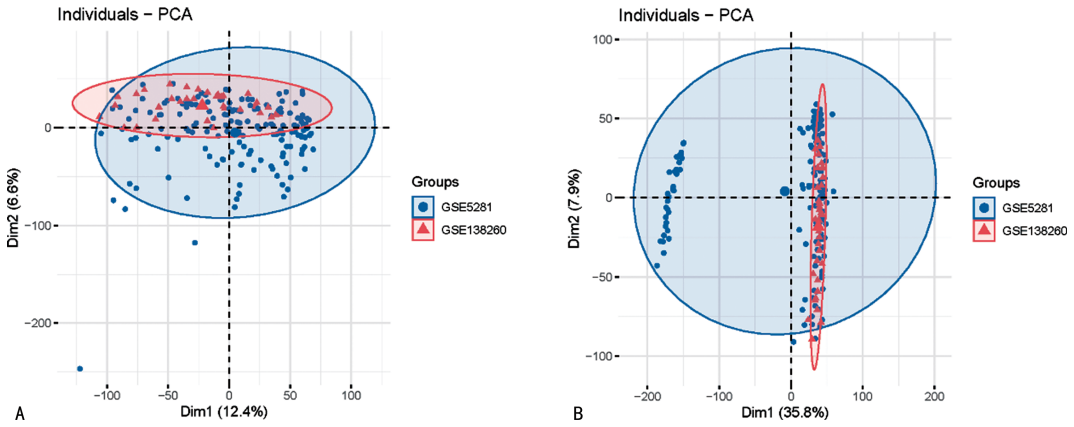
1.5　功能富集分析　应用 R 语言对差异表达基因进行京都基因和基因组数据库(KEGG)和基因本体(GO)富集分析。KEGG 富集分析可以用于分析筛选出基因可能的生物学功能和其涉及的相关信号通路；而 GO 富集分析则可用于分析基因的相关功能，又可分为生物过程(BP)、分子功能(MF)和细胞成分(CC)。

2　结　　果

2.1　GSE5281 和 GSE138260 数据合并及对数据进行去批次效应　GSE5281 数据集包含 74 例正常和 84 例 AD 患者的基因表达信息；GSE138260 数据集则包含 19 例正常和 17 例 AD 患者的基因表达信息。2 组数据未处理的数据点散乱(图 1B)，进行去批次效应后数据点相对集中在一个范围内(图 1A)；减小后续分析的误差。

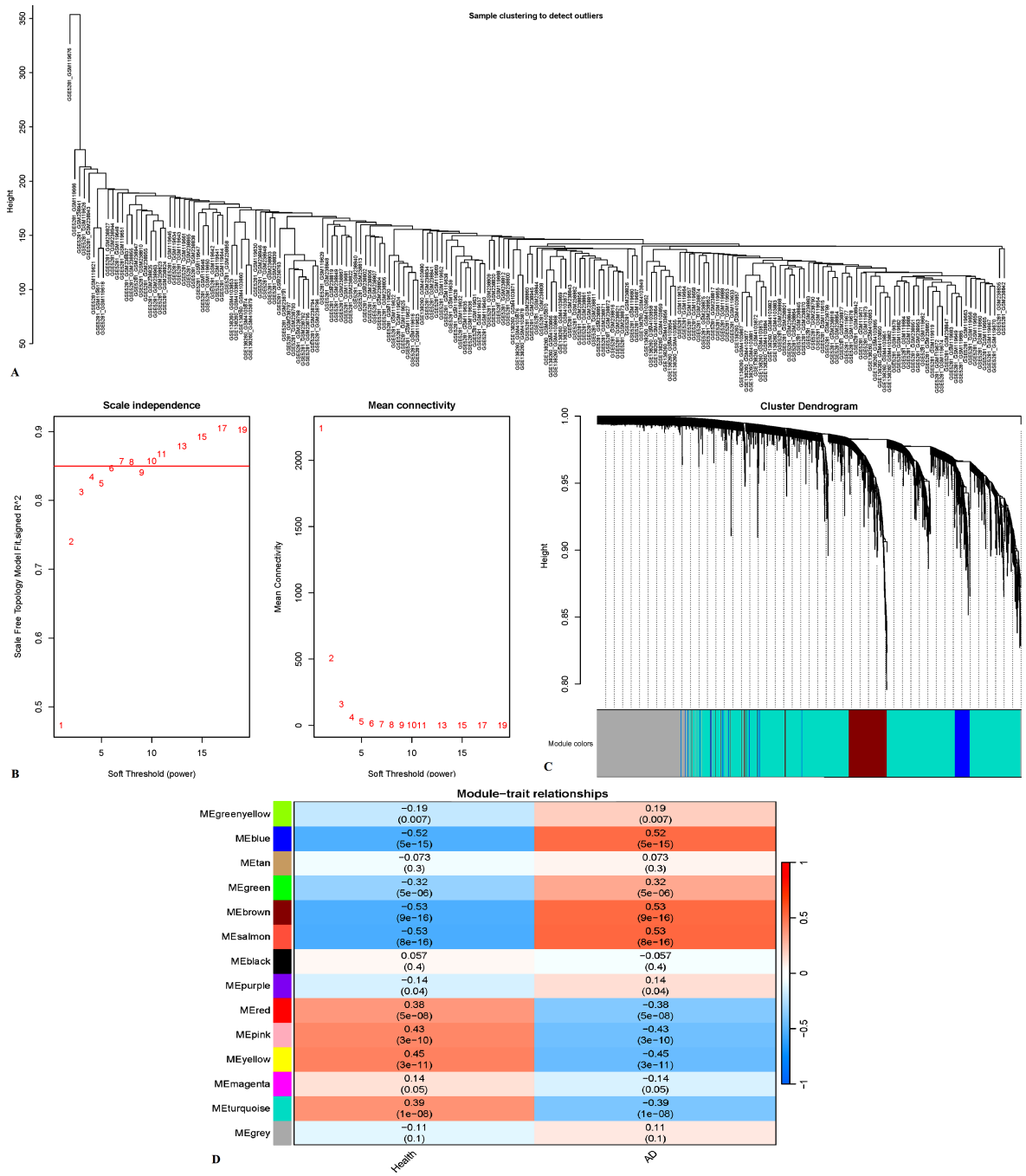
2.2　WGCNA 筛选与疾病相关的基因集　数据集去批次效应后，R 语言排除异常离群样本 GES5281_GSM119676(图 2A)。通过表达矩阵与表型数据的共同载入，确定软阈值为 8，构建表达网络(图 2B)。

通过 WGCNA 分析基因共被分为 14 个模块(图 2C)；是否患病与 14 个基因模块的相关性计算表明 brown 模块的 704 个基因与 AD 的相关性最高，其相关性系数为 0.53，P 值为 9e-16(图 2D)。



注：A. GEO 归一化数据图；B. GEO 原始数据图。

图 1　数据集去批次效应图



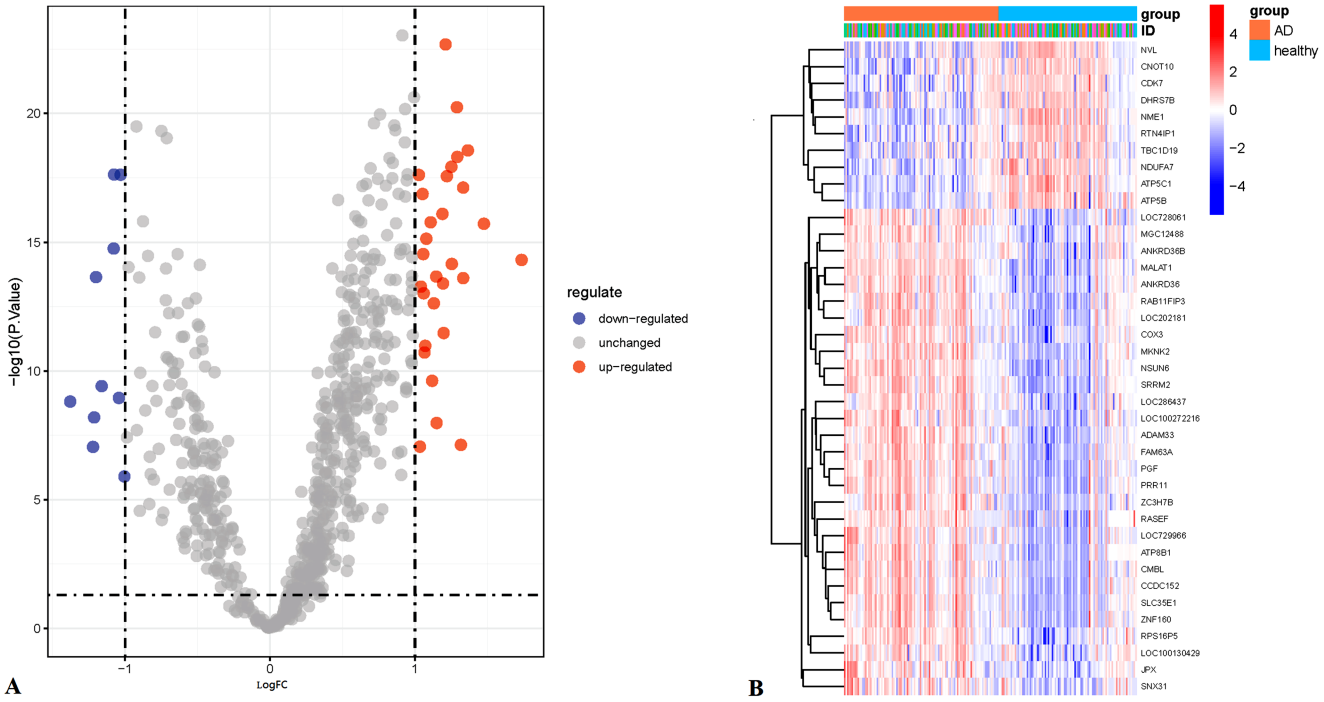
注：A. GEO 样本聚类树；B. AD 与健康的模块-性状关系图；C. 尺度独立性和平均连接阈值图；D. 基因模块聚类树。

图 2　WGCNA 分析样本聚类图及软阈值图

2.3　筛选疾病相关基因中差异性表达的基因　以 brown 模块中 704 个基因的表达矩阵为对象，以 |logFC|>1.2 且 $P<0.05$ 为阈值筛选表达差异的基因并作图。结果显示共有 39 个表达差异的基因，其中 10 个下调基因，29 个上调基因（图 3A、B）。

2.4　Lasso 回归筛选关键基因与验证　通过 Lasso 回归构建表型模型，Lasso 算法推荐有 2 个阈值；lambda.min 对应出 9 个关键基因，lambda.1se 对应出 7 个关键基因（图 4A、B）。R 语言对模型进行自我预测，选择曲线下面积（AUC）值更接近 1 的 lambda.min。筛选出 9 个关键基因为 MALAT1、NSUN6、SRRM2、ATP5B、SLC35E1、MKNK2、ZC3H7B、CMBL、JPX（表 1）。通过受试者操作特征曲线（ROC 曲线）验证 Lasso 回归筛选的基因，其 AUC 值均在 0.74 以上（图 4E、F），证明这 9 个基因均可作为潜在 AD 的生物标志物。其中，MALAT1、NSUN6、SRRM2、SLC35E1、MKNK2、ZC3H7B、CMBL、JPX 这 8 个基因在 AD 中表达上调；ATP5B 这个基因在 AD 中表达下调（图 5）。

2.5　KEGG 和 GO 的信号通路富集分析　对差异表达的基因进行 KEGG 富集分析和 GO 富集分析显示，KEGG 富集分析显示这 39 个基因主要参与矿物质元素的吸收、近端小管碳酸氢盐回收等通路调控（图 6A）；GO 富集分析表明，差异表达基因的主要分子功能与氧化还原驱动的活性跨膜转运蛋白、磷脂酰胆碱翻转酶活性等相关（图 6B）。

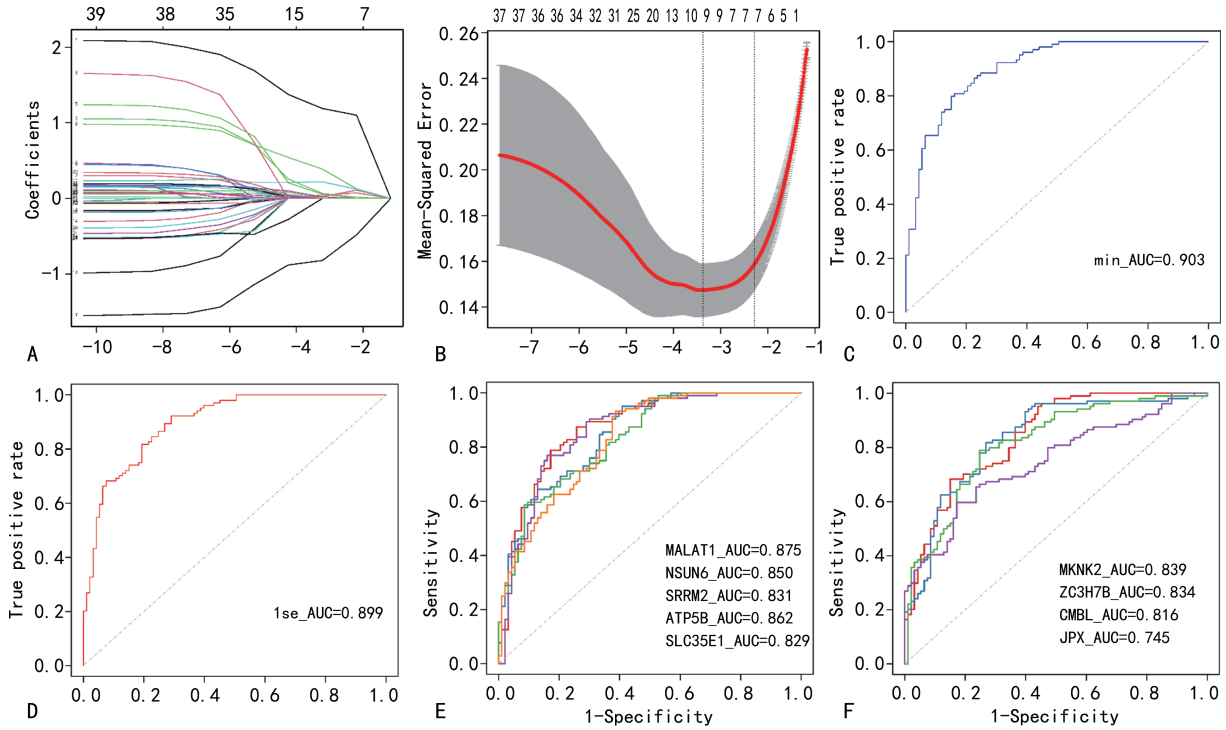


注：A. 火山图；B. 热图。

图 3　差异表达基因火山图和热图

表 1　Lasso 筛选的 12 个关键基因

基因	样本	AD *vs*. 健康（logFC）	AUC
metastasis associated lung adenocarcinoma transcript 1	MALAT1	1.211	0.875
NOP2/Sun RNA methyltransferase 6	ITPKB	1.288	0.850
serine/arginine repetitive matrix 2 factor F	SRRM2	1.364	0.831
ATP synthase, H+ transporting mitochondrial	ATP5B	−1.031	0.862
solute carrier family 35 member E1	SLC35E1	1.026	0.829
MAPK interacting serine/threonine kinase 2	MKNK2	1.219	0.839
zinc finger CCCH-type containing 7B	ZC3H7B	1.052	0.834
carboxymethylenebutenolidase homolog	CMBL	1.078	0.816
JPX transcript	JPX	1.118	0.745



注：A. 系数分布图；B. Lasso 交叉验证曲线；C、D. min，1se ROC 曲线；E. MALAT1、NSUN6、SRRM2、ATP5B、SLC35E1 ROC 曲线；F. MKNK2、ZC3H7B、CMBL、JPX 曲线。

图 4　Lasso 回归分析图

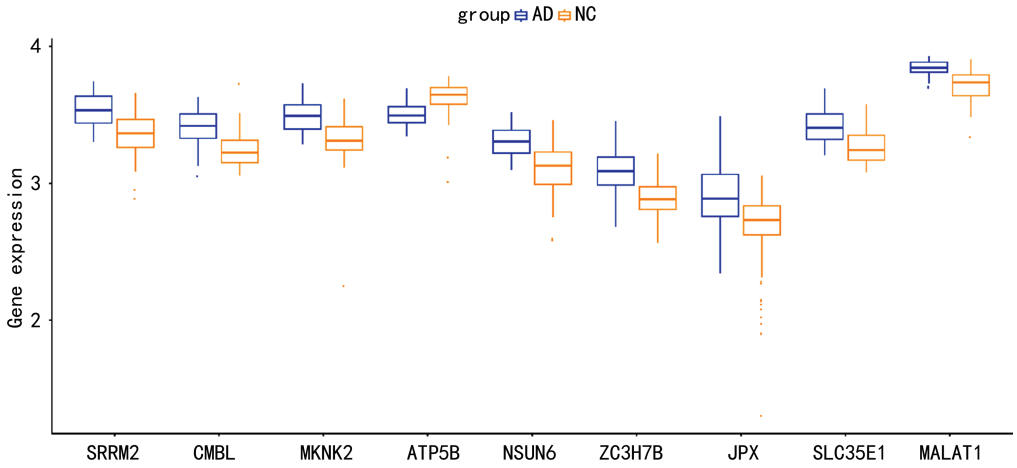
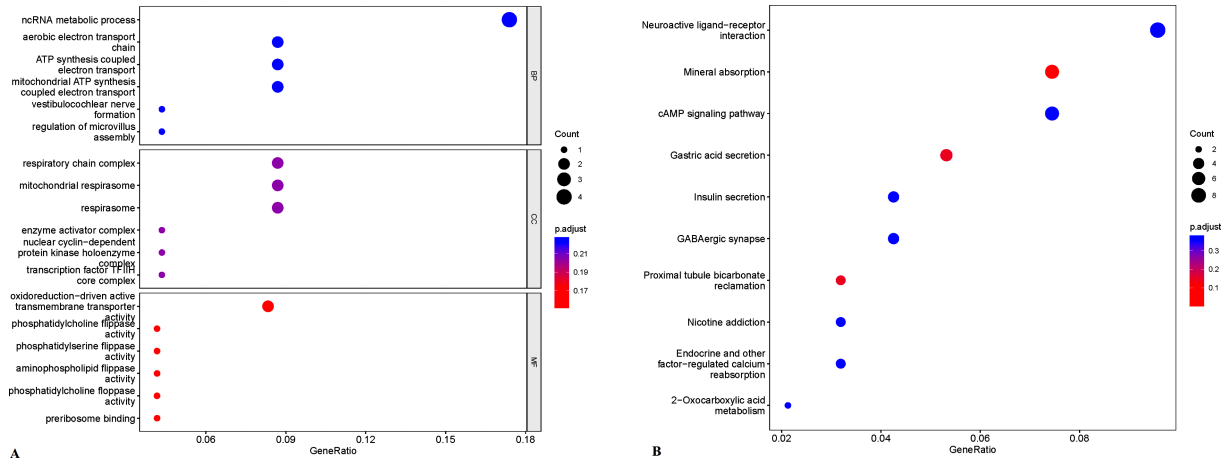


图 5　9 个关键基因的表达箱线图



注：A. KEGG 分析气泡图；B. GO 分析气泡图。

图 6　富集分析图

3　讨　论

为了更准确地了解 AD 的内在发病机制，挖掘其生物标志物，本研究使用现代生物信息学方法，从 GEO 数据库 2 个数据集 GSE5281 和 GSE138260 的 AD 患者与健康对照组的基因表达数据进行 R 语言综合分析，筛选 AD 患病关键基因及对基因进行富集分析。首先 R 语言数据合并后，对合并数据进行标准化处理；WGCNA 分析对基因表达数据中的基因分为 14 个模块，对模块和患病与否进行相关性分析筛选出 704 个关键基因。随后，对这 704 个关键基因进行差异性分析进一步筛选出上调 29 个、下调 10 个，共 39 个差异表达关键基因。本研究构建了表型模型，利用 Lasso 回归分析最终筛选出 MALAT1、NSUN6、SRRM2、ATP5B、SLC35E1、MKNK2、ZC3H7B、CMBL、JPX 9 个关键基因，其可能是潜在的 AD 生物标志物。ATP5B 在 AD 患者中显著性下调，ATP5B 参与多种细胞功能，包括腺苷基核糖核苷酸结合活性、血管抑素结合活性和质子转运 ATP 酶活性等，参与脂质代谢过程[21-22]。MALAT1、NSUN6、SRRM2、SLC35E1、MKNK2、ZC3H7B、CMBL、JPX 8 个基因在 AD 患者中表达上调，其中 MALAT1 是多种基因的转录调节因子，并参与调控细胞周期[23]，NSUN6、SRRM2、SLC35E1、MKNK2 等基因均与基因的转录调控相关。这些基因参与调控 AD 的具体机制及其在其中起到的作用需要进一步的实验验证。

本研究对差异性表达基因的富集分析显示，这些关键基因与矿物质元素的吸收、近端小管碳酸氢盐回收等通路相关，其分子功能主要涉及氧化还原驱动的活性跨膜转运蛋白、磷脂酰胆碱翻转酶活性过程，提示了在 AD 疾病的进展中，微量元素的吸收、氧化还原等代谢反应、脂质代谢可能起到关键作用。

本研究虽然以 GEO 数据库中的 2 个数据集为研究对象，筛选出了潜在的 AD 诊断标志物，即 MALAT1、NSUN6、SRRM2、ATP5B、SLC35E1、MKNK2、ZC3H7B、CMBL、JPX，为 AD 的诊断、机制和治疗靶点提供了新的思路，但需实验进一步验证。生物信息学综合差异基因表达、WGCNA 及 Lasso 回归分析的筛选方法可极大地缩短疾病关键基因、生物标志物的选择确认，有助于揭示疾病的内在分子机制，从而开发更加精准的诊断方法与更加有效的治疗方式。

参考文献

[1] CIPRIANI G, DOLCIOTTI C, PICCHI L, et al. Alzheimer and his disease: A brief history[J]. Neurol Sci, 2011, 32(2): 275-279.

[2] BLENNOW K, LEON M, ZETTERBERG H J L. Alzheimer's disease[J]. World J Biol Psychiatry, 2006, 368(9533): 387-403.

[3] DE-PAULA V J, RADANOVIC M, DINIZ B S, et al. Alzheimer's disease[J]. Subcell Biochem, 2012, 65: 329-352.

[4] HODSON R. Alzheimer's disease[J]. Nature, 2018, 559(7715): S1.

[5] BREIJYEH Z, KARAMAN R J M. Comprehensive review on Alzheimer's disease: Causes and treatment[J]. Molecules, 2020, 25(24): 5789.

[6] SERRANO-POZO A, FROSCH M P, MASLIAH E, et al. Neuropathological alterations in Alzheimer disease[J]. Cold Spring Harb Perspect Med, 2011, 1(1): a006189-a006189.

[7] SPIRES-JONES T, HYMAN B J N. The intersection of amyloid beta and tau at synapses in Alzheimer's disease[J]. Neuron, 2014, 82(4): 756-771.

[8] SINGH S K, SRIVASTAV S, YADAV A K, et al. Overview of Alzheimer's disease and some therapeutic approaches targeting Aβ by using several synthetic and herbal compounds[J]. Oxid Med Cell Longev, 2016: 7361613.

[9] RICHARD A. Risk factors for Alzheimer's disease[J]. Folia Neuropathologica, 2019, 57(2): 87-105.

[10] ANAND P, SINGH B. A review on cholinesterase inhibitors for Alzheimer's disease[J]. Arch Pharm Res, 2013, 36(4): 375-399.

[11] GUERREIRO R, BRAS J. The age factor in Alzheimer's disease[J]. Genome Med, 2015, 7: 106.

[12] JIA J, ZHANG Y, SHI Y, et al. A 19-year-old adolescent with probable Alzheimer's disease[J]. J Alzheimers Dis, 2023, 91(3): 915-922.

[13] YIANNOPOULOU K G, PAPAGEORGIOU S G. Current and future treatments in Alzheimer disease: An update[J]. J Cent Nerv Syst Dis, 2020, 12: 1179573520907397.

[14] LIVINGSTON G, HUNTLEY J, SOMMERLAD A, et al. Dementia prevention, intervention, and care: 2020 report of the lancet commission[J]. Lancet, 2020, 396(10248): 413-446.

[15] DEKOSKY S T, MAREK K J S. Looking backward to move forward: Early detection of neurodegenerative disorders[J]. Science, 2003, 302(5646): 830-834.

[16] MCKHANN G M, KNOPMAN D S, CHERTKOW H, et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease

(下转第 2904 页)

content_5358604. htm.

[2] 国务院深化医药卫生体制改革领导小组. 国务院深化医药卫生体制改革领导小组印发关于以药品集中采购和使用为突破口进一步深化医药卫生体制改革若干政策措施的通知[EB/OL]. (2019-11-20)[2022-10-03]. http://www.gov.cn/xinwen/2019-12/03/content_5457859.htm.

[3] 于保荣. 未来 5～10 年中国医疗保障制度的设计与思考:《中共中央国务院关于深化医疗保障制度改革的意见》的解读[J]. 卫生经济研究,2020,37(4):3-7.

[4] 国务院办公厅. 国务院办公厅关于推动药品集中带量采购工作常态化制度化开展的意见[EB/OL]. 2021-01-28[2022-10-06]. https://www.gov.cn/zhengce/content/2021-01/28/content_5583305.htm.

[5] 卢耀恩,石金铭,桑青原,等. 区块链技术应用于互联网医疗的研究热点与趋势分析:基于 CiteSpace 软件的可视化计量分析[J]. 中国医院管理,2022,42(11):18-22.

[6] 杨照,刘扬,江滨. 抗菌药物实施集中带量采购之思考[J]. 中国卫生,2021(8):65-66.

[7] 陈昊,饶苑弘. 化学药品注射剂一致性评价与开展带量采购的思考[J]. 中国新药杂志,2020,29(8):864-868.

[8] 薛天祺,路云,常峰. 国家药品集中带量采购中选结果及采购规则优化方向分析[J]. 卫生经济研究,2022,39(5):12-16.

[9] 刘洁兰,阮智慧,张天南,等. 抗肿瘤药品集中带量采购政策在某三级甲等中医院的应用实效分析[J]. 中国医疗管理科学,2022,12(5):24-27.

[10] 屈茹楠,高岸,范国荣,等. 国家带量采购政策对上海某院原研和仿制降压药使用状况的影响[J]. 中国药业,2022,31(15):10-15.

[11] 汪江涛,丁伯平,魏成成,等. 药品带量采购对芜湖市中医医院质子泵抑制剂使用的影响[J]. 现代药物与临床,2022,37(7):1603-1611.

[12] 赵耀伟,王成亮,闫彬,等. 基于间断时间序列的带量采购政策对中标他汀类药物使用影响分析[J]. 医药导报,2022,41(8):1234-1238.

[13] 杨琪,果伟,刘珊珊. 药品带量采购对某医院抗精神病药原研药和仿制药使用情况影响[J]. 中国医院药学杂志,2021,41(4):400-403.

[14] 向左娟,陈小娟,胡晓杰,等. 药品集中带量采购后某院抗菌药物使用及细菌耐药率变迁情况[J]. 中南药学,2022,20(11):2678-2683.

[15] 赵洁,李巍,王皋俊. 价值医疗视角下国家药品集中带量采购在某公立医院的实施效果评价[J]. 中国药房,2021,32(19):2410-2414.

(收稿日期:2022-11-22　修回日期:2023-04-25)

(上接第 2898 页)

[J]. Alzheimers Dement,2011,7(3):263-269.

[17] MAYEUX R,STERN Y J N O A. Epidemiology of Alzheimer's disease[J]. Cold Spring Harb Perspect Med,2012,2(8):a006239.

[18] YAARI R,FLEISHER A S,TARIOT P N. Updates to diagnostic guidelines for Alzheimer's disease[J]. Prim Care Companion CNS Disord,2011,13(5):PCC.11f01262.

[19] 宋祯彦,余婧萍,贺春香,等. 不同阶段阿尔茨海默病患者海马 CA1 区基因表达的生物信息学分析[J]. 世界科学技术:中医药现代化,2019,21(9):1791-1798.

[20] 李长征,王慧,吕蔚然,等. 阿尔茨海默病相关基因的生物信息学分析[J]. 中华神经医学杂志,2012,11(8):3.

[21] ANDERSSON U,ANTONICKA H,HOUSTEK J,et al. A novel principle for conferring selectivity to poly(A)-binding proteins: Interdependence of two ATP synthase beta-subunit mRNA-binding proteins[J]. Biochem J,2000,346 Pt 1(Pt 1):33-39.

[22] XU Y T,TAN H J,LIU K F,et al. Targeted inhibition of ATP5B gene prevents bone erosion in collagen-induced arthritis by inhibiting osteoclastogenesis[J]. Pharmacol Res,2021,165:105458.

[23] ZHANG T,LUO J Y,LIU F,et al. Long noncoding RNA MALAT1 polymorphism predicts MACCEs in patients with myocardial infarction[J]. BMC Cardiovasc Disord,2022,22(1):152.

(收稿日期:2023-03-31　修回日期:2023-06-10)