多组学数据分析识别阿尔茨海默病的因果基因*

南方医科大学公共卫生学院生物统计学系(510515) 谢晨曦 张瑞祥 游祖明 周基元△

【提 要】 **目的** 通过整合多组学数据使用全转录组关联研究(transcriptome-wide association study,TWAS)识别潜在的阿尔茨海默病的因果基因,并探索其潜在的致病机制。**方法** 基于大脑中的转录组汇总数据和蛋白质组汇总数据以及与阿尔茨海默病相关的全基因组关联研究(genome-wide association studies,GWAS)汇总数据,使用PMR-Egger方法进行TWAS或蛋白质组关联研究(proteome-wide association study,PWAS)以寻找阿尔茨海默病的潜在因果基因,再使用贝叶斯共定位进一步评估GWAS与eQTL(expression quantitative trait locus)/pQTL(protein quantitative trait locus)共享相同变异位点的概率,并对最终得到的基因进行蛋白质相互作用(protein-protein interaction,PPI)网络分析以及富集分析,探索其潜在的致病机制。**结果** 该研究在大脑蛋白质组中共发现15个潜在的因果基因,在大脑转录组中发现6个潜在的因果基因,其中大脑转录组以及蛋白质组存在一个共同的因果基因*MTCH2*。此外,这20个因果基因中的15个基因在PPI网络中形成了以*APOE*为核心的基因群,并富集于有统计学差异的14个GO条目。**结论** 该研究发现了阿尔茨海默病的*MTCH2*等20个潜在的因果基因以及14个潜在的生物通路,为阿尔茨海默病的实验研究、预防策略和临床治疗提供了依据。

【关键词】 阿尔茨海默病 全转录组关联研究 全蛋白组关联研究 因果基因
【中图分类号】 R749.1 【文献标志码】 A **DOI** 10.11783/j.issn.1002-3674.2026.05.005

Identifying Causal Genes of Alzheimer's Disease by Integrating Multi-omics Data

Xie Chenxi, Zhang Ruixiang, You Zuming, et al(*Department of Biostatistics, School of Public Health, Southern Medical University, Guangzhou* 510515)

【**Abstract**】 **Objective** This study aims to identify potential causal genes for Alzheimer's disease (AD) by integrating multi-omics data and using the transcriptome-wide association study (TWAS) method, and to explore the underlying pathogenic mechanisms. **Methods** We used summary data from the brain transcriptome and proteome, along with genome-wide association study (GWAS) summary data related to Alzheimer's disease. The PMR-Egger method was employed for TWAS or proteome-wide association study (PWAS) to identify potential causal genes for AD. Bayesian colocalization was then used to assess the probability of shared genetic variants between GWAS and eQTL (expression quantitative trait locus)/pQTL (protein quantitative trait locus). The identified genes were further analyzed using protein-protein interaction (PPI) network analysis and enrichment analysis to explore their potential pathogenic mechanisms. **Results** 15 potential causal genes in the brain proteome and 6 potential causal genes in the brain transcriptome were found. Among them, *MTCH2* was identified in both the brain transcriptome and proteome. Additionally, 15 genes among these 20 genes formed a gene cluster centered around *APOE* in the PPI network and were enriched in 14 statistically significant GO biological processes. **Conclusion** We identified 20 potential causal genes, such as *MTCH2*, and 14 pathways related to AD, providing valuable insights for experimental research, preventive strategies and clinical treatments for AD.

【**Key words**】 Alzheimer's disease; Transcriptome-wide association study; Proteome-wide association study; Causal gene

阿尔茨海默病(Alzheimer's disease, AD)是最常见的与年龄相关的神经退行性疾病,随着人口的老龄化,痴呆症患者的数量急剧增加[1]。有研究数据表明,全球约有5500万人患有AD,预计到2050年,AD患者的数量将接近1.5亿[2]。由于个体认知功能的受损,AD患者的生活质量下降,这将给家庭以及社会带来沉重的疾病负担[3]。然而,目前其尚无有效的治疗方法或治愈方法,且其致病的生理机制尚不完全清楚。因此,寻找和AD相关的致病机制和潜在的危险因素至关重要。

全基因组关联研究(genome-wide association study,GWAS)已经确定了很多和疾病/性状相关联的单核苷酸多态性(single nucleotide polymorphism, SNP)位点。然而,通过GWAS识别得到的SNP位点可能存在于非编码区,未能直接参与到基因的转录以及蛋白质的表达过程中。因此,仅凭借GWAS结果难以确定这些基因是否为因果基因,也难以确定其潜在的调节机制。全转录组关联研究(transcriptome-wide association study, TWAS)与蛋白质组关联研究(proteome-wide association study,PWAS)已经被广泛应用于寻找疾病因果基因之中[4-5],TWAS和PWAS分别通过分析转录组水平以及蛋白质组学水平基因(蛋白质)与疾病的关系,来寻找潜在的因果基因。本研究整合了eQTL(expression quantitative trait locus)汇总数据与pQTL(protein quantitative trait locus)汇总数据以及GWAS汇总数据,旨在识别潜在的对AD具有调控作用的基因,并探索其潜在的生物过程。除此以外,我们还使用了一种贝叶斯共定位方法(Bayesian

*基金项目:国家自然科学基金(82574202)
△通信作者:周基元,E-mail:zhoujiyuan5460@hotmail.com

test for colocalization, Coloc)[6]来进一步筛选TWAS/PWAS得到的因果基因，以确保GWAS数据与eQTL/pQTL数据是由同一致病位点所导致。总而言之，我们使用TWAS/PWAS方法以及Coloc共定位方法将GWAS汇总数据以及多维的数量性状汇总数据结合起来，来探索特定的途径以及潜在的因果基因。

在先前的寻找与AD致病相关基因的研究中，Wingo等[4]与Gockley等[5]使用了FUSION方法来进行TWAS以及PWAS分析，而Wang等[7]使用了MR-JTI（Mendelian randomization-joint tissue imputation）方法，通过分析基因表达数据或蛋白质组学数据寻找AD的因果基因，然而FUSION方法与MR-JTI方法在控制SNP位点的水平多效性方面存在局限性。另一方面，注意到一些SNP表现出水平多效性效应，并且其广泛分布于基因组中，影响广泛的性状[8]。因此，在我们的研究中将使用对水平多效性进行了建模的PMR-Egger方法来对多维数据进行TWAS和PWAS分析。

材料与方法

1. 数据来源

本研究所使用的GWAS汇总数据来源于国际阿尔茨海默病基因组学项目的第一阶段（international genomics of Alzheimer's project, IGAP）[9]，通过对遗传位点与AD做关联研究，筛选出潜在的与AD相关联的SNP位点。IGAP提供了大规模的欧洲血统个体的GWAS汇总数据，其中包含17008个AD病例以及37154个对照，共7055881个SNP位点。我们选择了一组来自基因型组织表达项目（genotype-tissue expression, GTEx）中大脑转录组的汇总数据，其包含280001个基因的4240231个SNP位点的总共10859309条eQTL数据[10]。另外，我们使用了一组来自1277个个体的大脑蛋白质组研究的pQTL汇总数据作为PWAS的分析数据集[11]，其中包含8458个基因的4539190个SNP位点的总共17327047条pQTL数据。由于没有获取相应的个体基因型数据，我们选择了来自千人基因组计划相同人群的基因面板来计算SNP之间的LD（linkage disequilibrium）矩阵。

2. 统计方法

（1）TWAS/PWAS

2020年，Yuan等[12]提出了PMR-Egger方法，其不同于FUSION等两阶段模型的TWAS方法，PMR-Egger使用似然框架来建模，以减小两阶段框架中参数估计的不确定性。除此之外，PMR-Egger还对水平多效性进行建模，可以直接检验水平多效性。其针对汇总数据的建模过程如下：

$$\hat{\beta}_x = \beta + \varepsilon_1 \quad (1)$$

$$\hat{\beta}_y = \alpha\beta + \gamma + \varepsilon_2 \quad (2)$$

其中$\hat{\beta}_x$和$\hat{\beta}_y$分别是eQTL/pQTL数据的效应量和GWAS数据的效应量，β为需要估计的SNP对基因表达量的效应量，α为基因表达量对结局的效应量，γ为水平多效性的效应量，ε_1和ε_2表示残差项。通过对（1）和（2）两式在似然框架下进行联合建模，计算效应量α和γ以及相应的P值。

（2）贝叶斯共定位

Coloc共定位通过贝叶斯方法来评估两个相关的信号是由于同一致病变异所导致的概率[6]。我们使用了默认的先验概率$P_1 = 10^{-4}$，$P_2 = 10^{-4}$，$P_{12} = 10^{-5}$，其中P_1是给定SNP与AD相关联的概率，P_2是给定SNP是显著QTL的概率，P_{12}是给定SNP既与AD关联，同时也与eQTL/pQTL相关联的概率。检验五个互斥的假设：①没有找到任何性状（H_0）的因果SNP；②只有性状1具有因果SNP（H_1）；③只有性状2具有因果SNP（H_2）；④两个性状都有一个因果SNP，但两个因果SNP不同（H_3）；⑤这两个性状都有一个因果SNP，并且共享相同的SNP（H_4）。我们主要关注最后一个假设H_4，后验概率（posterior probability, PP）用于量化对H_4的支持（表示为PP4）。我们筛选出PP4≥0.7的基因。

（3）PPI网络及富集分析

STRING在线工具整合了实验数据、公共数据库以及文献信息等来构建蛋白质相互作用（protein-protein interaction, PPI）网络[13]。其中，蛋白质-蛋白质的边代表蛋白质之间功能相关联，其中边的颜色代表了数据的不同来源。我们通过在线工具STRING（https://cn.string-db.org）对PWAS最终得到的基因进行了PPI网络分析。

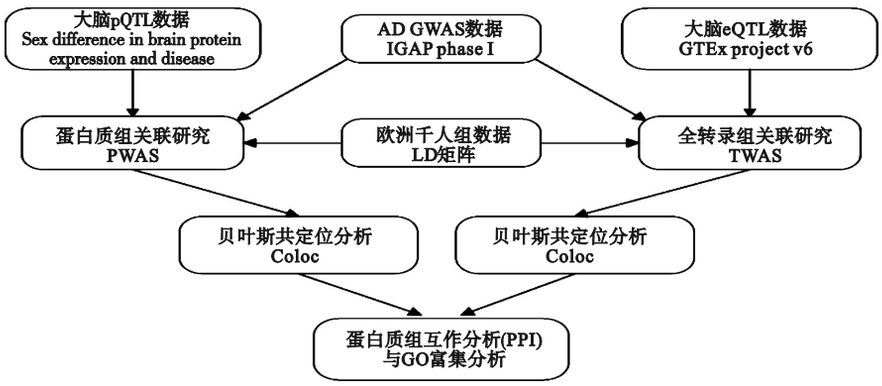
（4）研究流程

本研究采用似然框架下的TWAS方法以及贝叶斯共定位方法来寻找AD的因果基因。具体来说，我们首先筛选出GWAS汇总数据以及eQTL/pQTL汇总数据中共同存在的SNP位点，然后通过PMR-Egger方法来进行TWAS/PWAS分析，寻找潜在的因果基因，最后再对筛选得到的GWAS汇总数据以及eQTL/pQTL汇总数据进行贝叶斯共定位分析，筛选出共定位概率大于0.7的基因，如图1所示。

结　果

1. AD的TWAS以及Coloc分析结果

我们对TWAS结果得到的基因的自由度为1的卡方分布的P值进行了Bonferroni校正，将其设置为3.87×10^{-6}，并对得到的26个显著的基因进行了Coloc共定位，筛选出后验概率大于0.7的基因，最终我们鉴定得到了6个基因（*CLU*、*CEACAM19*、*NDUFS3*、*MADD*、*SLC39A23-AS1*、*MTCH2*），结果见图2。



结 果

图 1　TWAS 分析与 PWAS 分析流程图

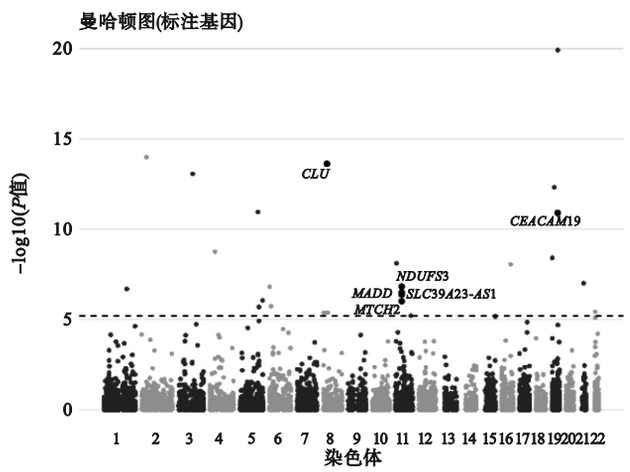


图 2　TWAS 分析结果

2. AD 的 PWAS 以及 Coloc 分析结果

与 TWAS 流程相同，我们对筛选得到的基因的卡方分布（自由度为 1）的 *P* 值进行了 Bonferroni 校正，将校正后的显著性水平设定为 5.91×10^{-6}，结果见图 3。得到了 198 个显著的因果基因，我们使用了 Coloc 共定位，最终得到了 15 个因果基因（*APOE*、*TMEM*126*B*、*APOC*2、*ZMAT*2、*CD*2*AP*、*MTCH*2、*CELF*1、*ISYNA*1、*HARS*2、*CD*14、*EPHX*2、*NECAB*1、*PLBD*2、*ABCB*9、*MTSS*2），其中包括已经确认和 AD 具有较强的因果效应的基因 *APOE* 和 *APOC*2 以及一个与 TWAS 同时找到了的基因 *MTCH*2。

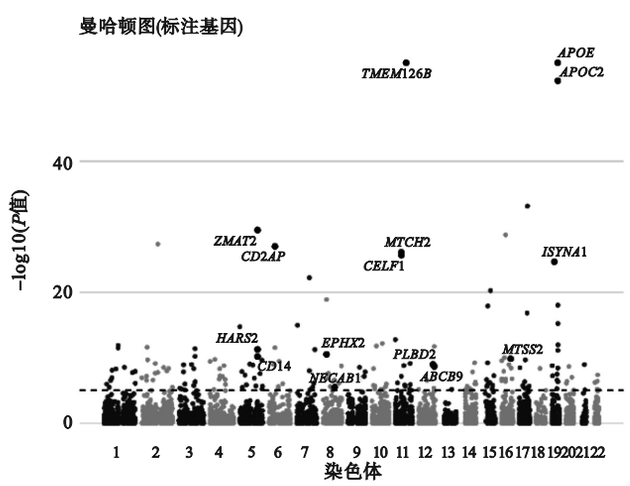


图 3　PWAS 分析结果

3. PPI 网络分析与 GO 富集分析结果

我们对 TWAS 以及 PWAS 得到的共 20 个基因进行了 PPI 网络分析，为了突出潜在的蛋白互作关系，该网络删除了独立的蛋白（基因），如图 4 所示。图 4 中红色、蓝色、粉色以及紫色的点分别代表着富集到的不同的条目（GO−term）。其中红色代表着逆转胆固醇转运过程（*APOC*2、*APOE*、*CLU*），其 *FDR*（false discovery rate）值为 0.0156，蓝色表示内吞作用的调节过程（*CLU*、*APOC*2、*APOE*、*CD*2*AP*、*CD*14），其 *FDR* 值为 0.0156，黄色（*CLU*、*APOC*2、*APOE*）和紫色（*CLU*、*CD*2*AP*、*CD*14、*APOE*、*APOC*2）分别为高密度脂蛋白颗粒和中密度脂蛋白颗粒的细胞成分，其 *FDR* 值分别为 0.0076、0.0081。我们发现 PPI 网络形成了一个以 *APOE* 为核心的群落，而 *APOE* 是一个已知的与 AD 和低密度脂蛋白水平相关的基因，其在中枢神经系统中起着重要作用。除此之外，我们还发现 TWAS 与 PWAS 共同的因果基因（*MTCH*2）与 *TMEM*126*B* 基因存在蛋白的互作关系，其综合交互得分为 0.695，表明 *MTCH*2 与 *TMEM*126*B* 可能存在潜在的交互机制对 AD 产生影响。

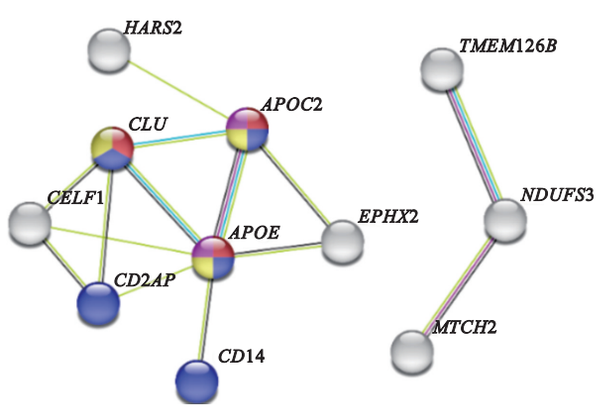


图 4　PPI 网络

通过对 TWAS 和 PWAS 的结果进行富集分析，15 个基因被富集到了 14 个条目（GO−term）中，其主要体现在生物过程（biological process，BP）和细胞组成（cellular component，CC）两个方面，如图 5 所示。除了

图4中所包含的以*APOE*为核心的与脂质蛋白相关的条目外，*MTCH2*也被发现富集于脂质稳态的条目中，这些生物机制可能与AD密切相关。

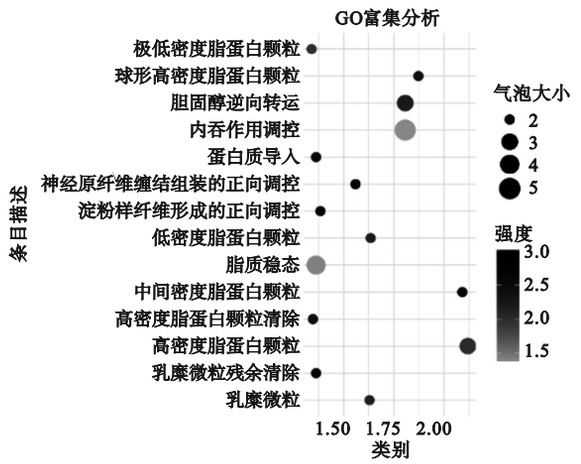


图5 GO富集分析结果

讨 论

本研究通过整合大脑组织的转录组以及蛋白质组学汇总数据，进行了TWAS分析和PWAS分析，一共确定了20个阿尔茨海默病潜在的因果基因。跟现有研究相比，本研究在方法学整合与数据分析深度上进行了针对性的优化：首先，本研究纳入了迄今规模较大的大脑蛋白质组学数据集（n=1277），既往研究从未使用过基于该数据集得到的汇总数据进行分析，因此，本研究较既往同类研究可显著提升统计效力，为发现AD潜在因果基因提供了更可靠的数据基础；其次，本研究使用了PMR-Egger模型系统校正了SNP位点上水平多效性的干扰，并进一步使用了Coloc共定位模型进行验证，增强了TWAS分析因果推断结果的可靠性。在本研究中，*MTCH2*基因在大脑转录组以及蛋白质组中均表现出了因果效应，表明了其在mRNA和蛋白质水平上的一致性。通过PPI网络分析我们发现一个以*APOE*基因为核心的群落，通过对因果基因进行GO富集分析我们发现AD的发病机制可能与胆固醇的逆向转运、内吞作用的调节以及脂蛋白颗粒相关。

本研究发现*MTCH2*、*MADD*、*NDUFS3*、*CELF1*等基因是AD的高风险基因，其中*MTCH2*在表达水平以及蛋白质组学中与AD显著相关。*MTCH2*对AD的影响除了通过基因表达影响性状外，其还可能通过SNP位点的多效性影响AD。在11号染色体上的基因簇中，*MTCH2*编码了线粒体内膜转运蛋白中的一种，这些蛋白与线粒体介导的细胞死亡、脂肪细胞分化、胰岛素敏感性相关，并与体质指数的增加有遗传关联[14]。在一项小鼠的实验中，*MTCH2*被发现对线粒体凋亡具有调节作用，其会促进钙离子诱导的神经元损伤，并且在小鼠前脑中敲除*MTCH2*基因会导致钙离子处理能力以及线粒体运动性缺陷，进而导致海马体突触减少、运动协调性下降[15]。研究发现*MADD*、*MTCH2*基因的表达水平与$A\beta_{42}$的生成密切相关：敲低*MTCH2*与*MADD*可显著降低$A\beta_{42}$水平，而过表达该基因则会导致$A\beta_{42}$的积累效应[16]。通过PPI分析，我们发现*MTCH2*基因与*NDUFS3*基因可能存在潜在生物过程从而导致AD。此外，*MADD*与*NDUFS3*基因均与*CELF1*基因座存在关联，*CELF1*基因座上的多个基因可能发挥协同作用影响AD[16]。已经有证据表明*NUP160*会导致蛋白质运输异常，从而进一步导致神经退行性疾病，在本研究中发现的*MTCH2*与*NDUFS3*被发现与*NUP160*基因表达在正常人脑中高度相关，而这种相关性在具有AD的个体中丢失[17]。*NDUFS3*基因可能通过影响能量代谢进而对AD产生影响，具体而言，*NDUFS3*是NADH脱氢酶复合体的核心亚基之一，而进一步的研究表明，抑制复合体Ⅰ与认知功能、突触可塑性、p-tau相关[18]。研究表明AD患者脑组织中*CLU*基因的mRNA表达会显著升高，且*CLU*蛋白被证实可与$A\beta$相互作用并影响纤维的形成[19]。

在已经确定的风险基因中，*APOE*基因仍然是最强和最普遍的，其中*APOE*的ε_4等位基因显著增加了AD风险。*APOE*可通过影响大脑中β淀粉样蛋白代谢、tau蛋白磷酸化等途径从而导致AD[20]。除此以外，*APOE*作为大脑中主要的脂质运转蛋白，可介导脂质以及胆固醇的代谢[21]，而缺乏促进胆固醇外排可能导致细胞内积累从而导致细胞毒性。Claes等[22]研究发现，当人类iPSC（induced pluripotent stem cell）衍生的异种移植小胶质细胞移植到淀粉样蛋白小鼠模型的大脑中后，这些细胞会在靠近淀粉样蛋白斑块的位置积累大量脂滴，而远离斑块的区域则没有这种现象。我们的研究发现，*APOC2*基因与AD存在因果关联。然而此前关于*APOC2*基因与AD之间存在关联的研究较少，有研究指出在调整*APOE*基因型后，*APOC2*附近基因座与脑淀粉样血管病相关[23]。此外，*TOMM40-APOE-APOC2*区域的DNA甲基化水平在AD患者与认知正常人群间存在差异，提示*APOC2*可能在AD的发病机制中发挥潜在作用[24]。

综上所述，本研究通过使用大脑的多组学数据，寻找了AD的潜在因果基因，并通过GO富集分析、PPI网络分析，进一步探索了其潜在的作用机制以及相关蛋白的互作关系。本研究可为AD的发病机制提供新的方向以及证据，且部分基因可能成为潜在的治疗靶点。然而，本研究仍然存在以下局限性。首先，基因表达调控过程高度复杂，本研究仅基于大脑转录组学和蛋白质组学汇总数据进行了分析，未能全面涵盖可

变剪切、代谢组等其他生物学过程的影响。其次，受限于当前公开的大规模 AD 基因组数据（如 IGAP）仍以欧洲人群为主，本研究暂时缺乏同等规模的亚洲人群数据验证，尽管基于欧洲人群数据的分析具有较高的统计效力，并在一定程度上揭示了跨种族的遗传风险基因（如 *APOE*），但未来仍需通过多族裔队列进一步验证。此外，由于缺乏个体层面的数据，我们难以控制环境因素等混杂变量的影响，也无法整合多组织多组学数据以更全面地解析 AD 的发病机制。最后，本研究主要基于生物信息学分析，尚未对所鉴定的因果基因进行深入的实验验证，未来需结合功能实验进一步确认其分子机制。

参 考 文 献

[1] Shi Y, Lin F, Li Y, et al. Association of pro-inflammatory diet with increased risk of all-cause dementia and Alzheimer's dementia: a prospective study of 166377 UK Biobank participants[J]. BMC Medicine, 2023,21(1):266.

[2] Khan S, Barve KH, Kumar MS. Recent advancements in pathogenesis, diagnostics and treatment of Alzheimer's disease[J]. Current Neuropharmacology, 2020,18(11):1106-1125.

[3] Sun Y, Zhu J, Zhou D, et al. A transcriptome-wide association study of Alzheimer's disease using prediction models of relevant tissues identifies novel candidate susceptibility genes[J]. Genome Med, 2021,13(1):141.

[4] Wingo AP, Liu Y, Gerasimov ES, et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis[J]. Nat Genet, 2021,53(2):143-146.

[5] Gockley J, Montgomery KS, Poehlman WL, et al. Multi-tissue neocortical transcriptome-wide association study implicates 8 genes across 6 genomic loci in Alzheimer's disease[J]. Genome Med, 2021,13(1):76.

[6] Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics[J]. PLoS Genetics, 2014,10(5):e1004383.

[7] Wang YH, Luo PP, Geng AY, et al. Identification of highly reliable risk genes for Alzheimer's disease through joint-tissue integrative analysis[J]. Front Aging Neurosci, 2023,15:1183119.

[8] Verbanck M, Chen CY, Neale B, et al. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases[J]. Nat Genet, 2018,50(5):693-698.

[9] Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74046 individuals identifies 11 new susceptibility loci for Alzheimer's disease[J]. Nat Genet, 2013,45(12):1452-1458.

[10] Qi T, Wu Y, Zeng J, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood[J]. Nat Commun, 2018,9(1):2282.

[11] Wingo AP, Liu Y, Gerasimov ES, et al. Sex differences in brain protein expression and disease[J]. Nat Med, 2023,29:2224-2232.

[12] Yuan Z, Zhu H, Zeng P, et al. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies[J]. Nat Commun, 2020,11(1):3861.

[13] Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets[J]. Nucleic Acids Res, 2019,47(D1):D607-D613.

[14] Escott-Price V, Bellenguez C, Wang LS, et al. Gene-wide analysis detects two new susceptibility genes for Alzheimer's disease[J]. PLoS One, 2014,9(6):e94661.

[15] D'Orsi B, Niewidok N, Düssmann H, et al. Mitochondrial carrier homolog 2 functionally co-operates with BH3 interacting-domain death agonist in promoting Ca^{2+}-induced neuronal injury[J]. Frontiers in Cell and Developmental Biology, 2021,9:750100.

[16] Xu M, Liu Q, Bi R, et al. Coexistence of multiple functional variants and genes underlies genetic risk locus 11p11.2 of Alzheimer's disease[J]. Biological Psychiatry, 2023,94(9):743-759.

[17] Karch CM, Ezerskiy LA, Bertelsen S, et al. Alzheimer's disease risk polymorphisms regulate gene expression in the ZCWPW1 and the CELF1 loci[J]. PLoS One ,2016,11(2):e0148717.

[18] Ruisch IH, Widomska J, De Witte W, et al. Molecular landscape of the overlap between Alzheimer's disease and somatic insulin-related diseases[J]. Alzheimers Res Ther, 2024,16(1):239.

[19] Karch CM, Goate AM. Alzheimer's disease risk genes and mechanisms of disease pathogenesis[J]. Biol Psychiatry, 2015,77(1):43-51.

[20] Liu CC, Zhao N, Yamaguchi Y, et al. Neuronal heparan sulfates promote amyloid pathology by modulating brain amyloid-β clearance and aggregation in Alzheimer's disease[J]. Science Translational Medicine, 2016,8(332):332ra44.

[21] Raulin AC, Martens YA, Bu G. Lipoproteins in the Central Nervous System: from biology to pathobiology[J]. Annual Review of Biochemistry, 2022,91:731-759.

[22] Claes C, Danhash EP, Hasselmann J, et al. Plaque-associated human microglia accumulate lipid droplets in a chimeric model of Alzheimer's disease[J]. Mol Neurodegeneration, 2021,16(1):50.

[23] Shade LMP, Katsumata Y, Abner EL, et al. GWAS of multiple neuropathology endophenotypes identifies new risk loci and provides insights into the genetic risk of dementia[J]. Nat Genet, 2024,56(11):2407-2421.

[24] Shao Y, Shaw M, Todd K, et al. DNA methylation of TOMM40-APOE-APOC2 in Alzheimer's disease[J]. J Hum Genet, 2018,63(4):459-471.

（责任编辑：张　悦）