

• 论 著 •

基于 GEO 数据库筛选胶质母细胞瘤相关差异基因及信号通路*

韩 阳

首都医科大学附属北京友谊医院检验科, 北京 100050

摘要:目的 使用基因表达谱数据库(GEO)快速筛选出与胶质母细胞瘤的发生、发展过程密切相关的差异基因及信号通路。方法 利用生物信息学分析方法在 GEO 数据库中选择 GSE31262 数据集作为主要分析对象,包括 5 个成人神经干细胞个体样本和 9 个胶质母细胞瘤干细胞个体样本。从这些样本中筛选潜在的差异基因,使用京都基因和基因组百科全书(KEGG)通路分析与基因本体论(GO)富集分析对筛选到的差异基因进行对比。在 STRING 在线数据库(<https://string-db.org/>)中,对筛选出来的差异表达基因所编码的蛋白,构建蛋白质-蛋白质相互作用(PPI)分析网络,从而进一步筛选确定评分水平最高的前 10 位基因。通过癌症基因组图谱(TCGA)、免疫细胞浸润分析及受试者工作特征(ROC)曲线分析等方法,分析差异基因与胶质母细胞瘤预后情况的相关性及其诊断价值。结果 通过筛选 GSE31262 数据集得到差异基因共 2 692 个(上调基因 1 182 个,下调基因 1 510 个)。KEGG 通路分析和 GO 富集分析的结果发现,上述差异基因主要涉及细胞器裂变、核分裂、胚胎器官发育、染色体分离、轴突生成、胶质细胞再生,主要参与细胞内吞作用、溶酶体、MAPK 信号通路、细胞周期、人乳头瘤病毒感染、人类免疫缺陷病毒感染等过程。在 PPI 网络中筛选出的评分居前 10 位的关键差异基因分别为 CDK1、CCNB1、CCNA2、BUB1、KIF11、TOP2A、CCNB2、CHEK1、RRM2 和 ASPM,均与胶质母细胞瘤预后有关($P < 0.000 1$)。免疫细胞浸润分析表明,Th2 细胞的浸润水平与 10 个基因的表达呈正相关,巨噬细胞与 RRM2 的表达呈负相关,与其他 9 个基因的表达呈正相关。ROC 曲线分析显示,10 个基因在胶质母细胞瘤中均有较高的诊断价值。结论 筛选得到的 10 个差异基因为 CDK1、CCNB1、CCNA2、BUB1、KIF11、TOP2A、CCNB2、CHEK1、RRM2 和 ASPM,上述基因有望成为胶质母细胞瘤诊断及治疗的潜在分子靶点。

关键词:胶质母细胞瘤; 生物信息学; 差异基因; 信号通路**DOI:**10.3969/j.issn.1673-4130.2023.20.008**中图法分类号:**R739.4**文章编号:**1673-4130(2023)20-2473-06**文献标志码:**A

Key genes and signaling pathways of glioblastoma based on GEO database*

HAN Yang

Department of Clinical Laboratory, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China

Abstract: **Objective** To screen the genes and signaling pathways closely related to the occurrence and development of glioblastoma in Gene Expression Omnibus (GEO) database by bioinformatics analysis, and to provide new ideas for further diagnosis and treatment. **Methods** GSE31262 gene expression profile data set was selected as analysis objects in GEO database, including 5 adult neural stem cell samples and 9 glioblastoma stem cell samples. The differentially expressed genes were screened out, and gene ontology (GO) enrichment analysis and genome database (KEGG) pathway analysis were performed for the differentially expressed genes. In the STRING online database, protein-protein interaction (PPI) network was constructed for the encoding proteins of the above screened differentially expressed genes, and the top 10 differentially expressed genes were screened. The relationship between key genes and prognosis of glioblastoma patients was analyzed from the Cancer Genome Atlas (TCGA) database. Immune infiltration analysis and receiver operating characteristic (ROC) curve analysis were also performed. **Results** Totally 2 692 samples were selected from GSE31262 dataset, among which 1 182 were up-regulated and 1 510 down-regulated. GO enrichment analysis and KEGG pathway analysis showed that it was mainly involved in organelle fission, nuclear division, embryonic organ development, chromosome separation, axon generation, glial cell regeneration, protein localization in the cell periphery, etc. It is also mainly involved in cell endocytosis, lysosome, MAPK signaling pathway, cell

* 基金项目:首都医科大学附属北京友谊医院种子计划人才项目(YYZZ202249)。

作者简介:韩阳,男,主管技师,主要从事临床生化免疫检验研究。△ 通信作者,E-mail:hanyangyouyi@163.com。

cycle, human papillomavirus infection, human immunodeficiency virus infection, etc. In the PPI network, the top 10 key genes were CDK1, CCNB1, CCNA2, BUB1, KIF11, TOP2A, CCNB2, CHEK1, RRM2 and ASPM. The 10 differentially expressed genes were associated with patient prognosis ($P < 0.0001$). Immune infiltration analysis showed that the infiltration level of Th2 cells was positively correlated with the expression of 10 genes, while macrophages were negatively correlated with the expression of RRM2, and significantly positively correlated with the expression of 9 other genes. Mast cells showed a significant negative correlation with the expression of ASPM, BUB1, CDK1, CCNA2, CCNB2, CHEK1, KIF11, TOP2A, and RRM2. Dendritic cells showed a significant negative correlation with nine other genes except for RRM2. ROC curve analysis showed that all the 10 genes had good independent diagnostic value in glioblastoma. **Conclusion** Differential expression gene profiles exist in tumor tissues and normal tissues of glioblastoma patients. The key genes screened out are CDK1, CCNB1, CCNA2, BUB1, KIF11, TOP2A, CCNB2, CHEK1, RRM2 and ASPM. The above genes could be used as potential molecular markers of glioblastoma.

Key words: glioblastoma; bioinformatics; differentially expressed genes; signaling pathway

神经胶质瘤起源于神经胶质细胞,是目前最常见的颅内恶性肿瘤之一,占原发性中枢神经系统肿瘤的45%左右^[1],年发病率在3/10万~8/10万^[2]。有研究认为,胶质瘤治疗极其困难,预后尤为不佳,大约77%的恶性胶质瘤患者在确诊后1年内死亡^[3-4]。其中,胶质母细胞瘤是最为常见的脑胶质瘤类型,是所有肿瘤中预后最差的类型之一。因此,研究胶质母细胞瘤的早期干预靶点具有重要意义。

随着基因芯片及二代高通量测序技术的不断发展,越来越多的研究利用基因测序技术与生物信息学分析方法,通过筛选与疾病发生发展相关的差异基因,为疾病的机制研究寻找新的靶标和研究思路^[5-6]。胶质母细胞瘤属于多基因疾病,其发生和恶性进展与遗传因素密切相关^[7]。因此,本研究拟通过基因表达谱数据库(GEO)筛选出胶质母细胞瘤相关的差异表达基因,探讨这些潜在致病基因与胶质母细胞瘤患者诊断、预后的相关性,为寻找新的标志物指明方向。

1 资料与方法

1.1 资料来源 本研究对 GEO 数据库进行检索,选择“human”为物种类别,检索关键词设定为“Glioma”,分析对象为 GSE31262 基因表达谱数据集,该数据集主要包括 5 个成人神经干细胞个体样本和 9 个胶质母细胞瘤干细胞个体样本^[8]。

1.2 差异表达基因筛选 采用 R 语言软件中的 limma 程序包,首先对所得到的全部数据进行标准化处理;其次,对比两组胶质母细胞瘤干细胞样本和正常成人神经干细胞样本得到的基因数据,并从这批数据中寻找差异有统计学意义的差异表达基因。判定标准为 $|\log_2(FC)| > 1$ 及校正后 $P < 0.05$ 。

1.3 差异表达基因的功能富集分析 使用基因本体论(GO)富集分析和京都基因和基因组百科全书(KEGG)通路分析方法预测潜在功能及信号通路。

1.4 蛋白质-蛋白质相互作用(PPI)网络构建及差异基因筛选 在 STRING 在线数据库(<https://string-db.org/>)中,对筛选出的差异基因所编码的蛋白进行 PPI 网络构建,筛选 combined score > 0.4 的相互作用

蛋白。使用 Cytoscape 软件分析 PPI 结果,采用 Degree 算法和 Cytohubba 插件,得到评分最高的前 10 位差异基因。

1.5 差异基因的批量生存分析 利用癌症基因图谱(TCGA)数据库检索到 666 例胶质母细胞瘤患者的转录组测序结果及其临床数据资料。根据基因表达水平差异,将筛选出的差异基因分为高、低表达组。

1.6 免疫浸润分析 使用 R GSVA 程序包中提供的 ssGSEA 算法,通过识别 24 种免疫细胞标志物^[9],评估胶质母细胞瘤组织的免疫细胞浸润情况。结果采用“ggplot2”进行可视化,所有分析和可视化均在 R 4.2.1 中进行。

1.7 统计学处理 采用 Graphpad Prism 8.4 统计软件进行数据分析。采用 R 语言软件中的 ggplot 和 pheatmap 程序包绘制差异表达基因的火山图及热图,采用 ClusterProfiler 程序包进行基因富集的分析及绘制图谱,两组比较采用 t 检验,多组间比较采用方差分析。使用 R 语言软件中的 pROC 程序包进行受试者工作特征(ROC)曲线分析,分析各差异基因诊断胶质母细胞瘤的曲线下面积(AUC)。分析结果用 ggplot2 进行可视化。绘制 Kaplan-Meier 生存分析曲线,比较两组患者生存率的差异并计算风险比(HR),Kaplan-Meier 曲线的 P 和 $HR(95\%CI)$ 通过 Log-rank 检验和单变量 COX 回归分析得出。以 $P < 0.05$ 为差异有统计学意义。

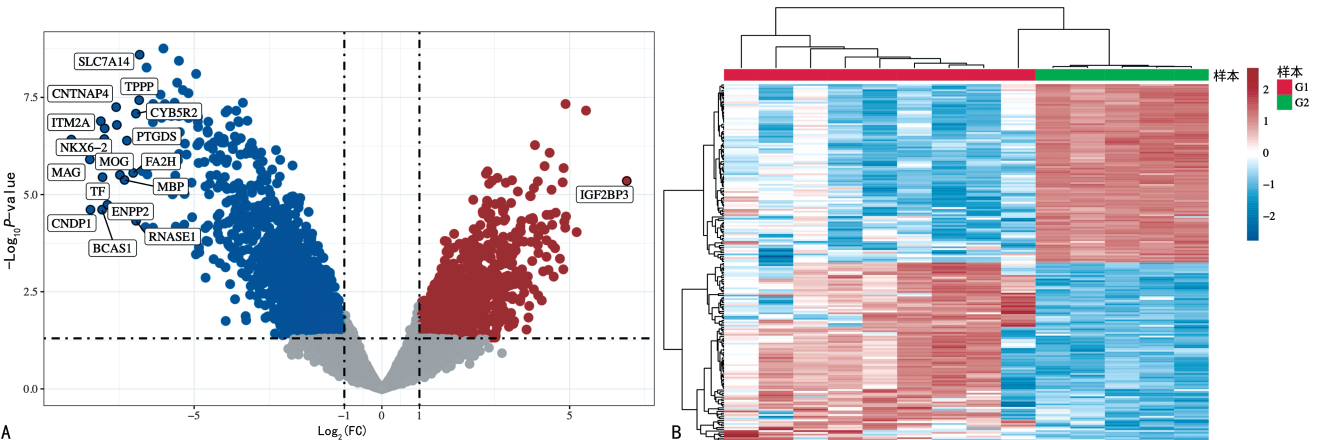
2 结果

2.1 差异基因的筛选结果 选取 GSE31262 数据集为分析对象,分析 5 个成人神经干细胞个体样本和 9 个胶质母细胞瘤干细胞个体样本,从上述样本中的全部基因中筛选出 $|\log_2(FC)| > 1$ 的差异表达基因 2 692 个,其中表达上调 1 182 个、表达下调 1 510 个。差异基因火山图及聚类热图结果见图 1。

2.2 差异基因的 GO 富集分析和 KEGG 通路分析结果 筛选出的上调、下调差异表达基因的相关功能分析结果见图 2。KEGG 通路分析结果显示,筛选到的差异表达基因主要参与细胞内吞作用、溶酶体、

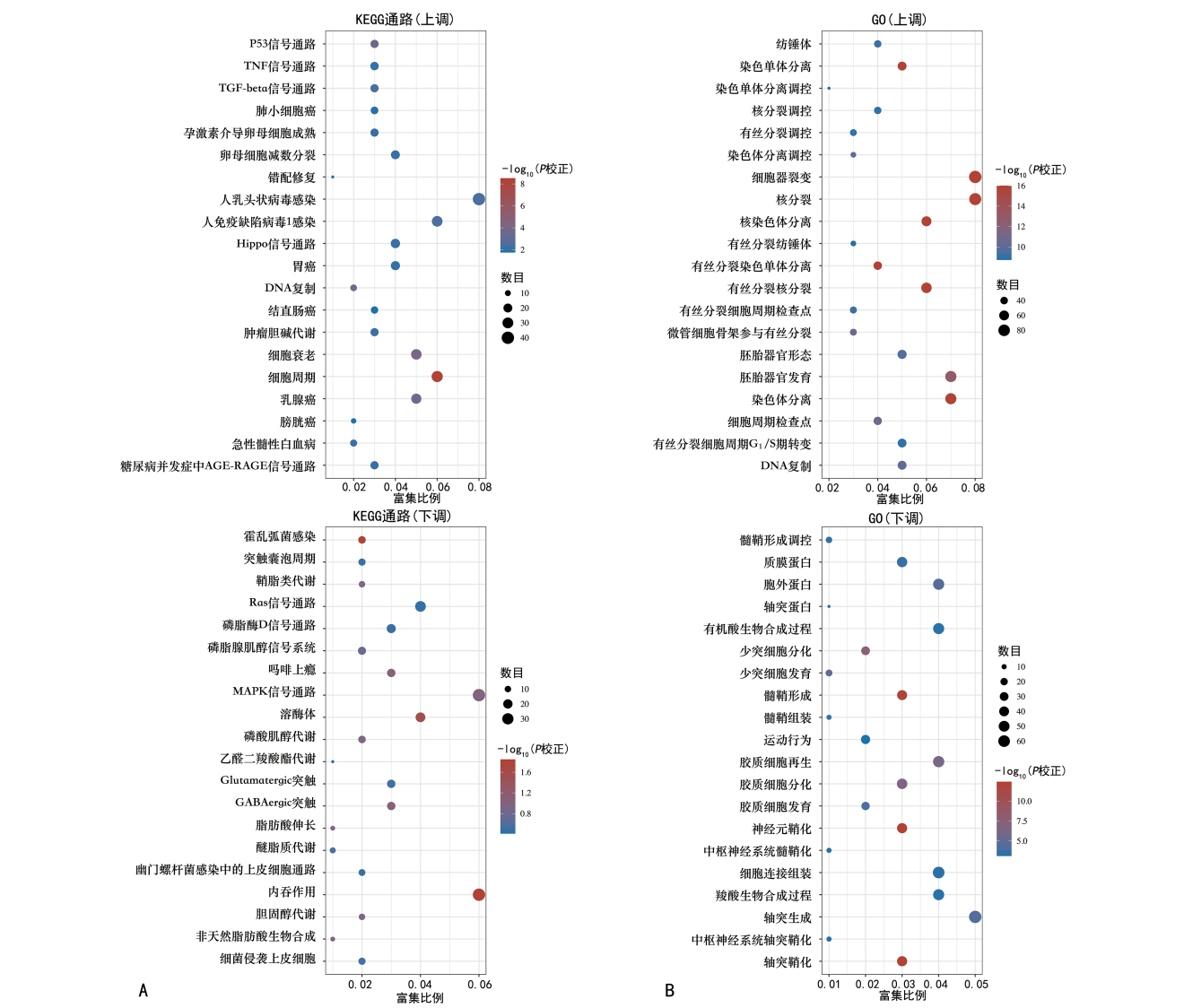
MAPK 信号通路、细胞周期、人乳头瘤病毒感染、人类免疫缺陷病毒感染等;GO 富集分析结果显示,差异表达基因主要涉及细胞器裂变、核分裂、胚胎器官发育、

染色体分离、轴突生成、胶质细胞再生、蛋白质定位于细胞外围等。



注:A 为差异基因火山图;●代表上调基因,●代表下调基因,●为无差异基因;B 为差异基因聚类热图;蓝色为下调基因,红色为上调基因,G1 为胶质母细胞瘤干细胞样本,G2 为正常成人神经干细胞。

图 1 差异基因的筛选结果

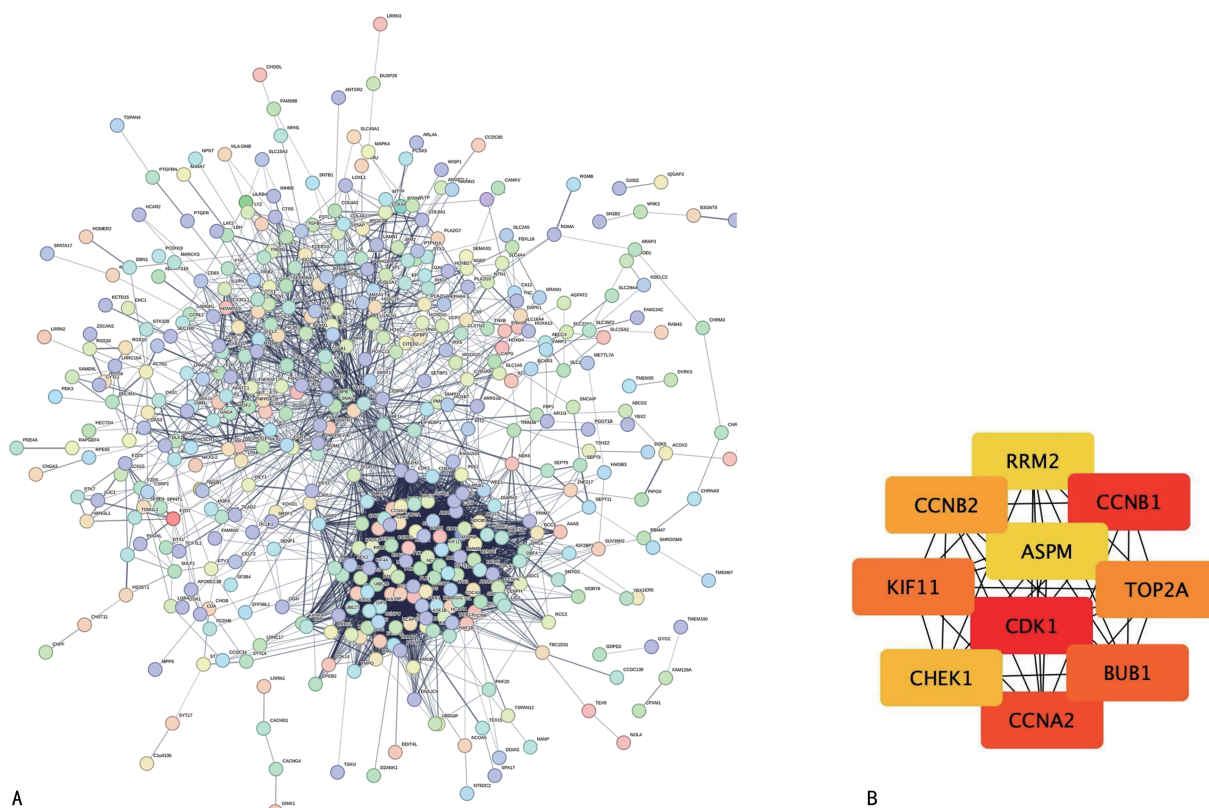


注:A 为 KEGG 通路分析;B 为 GO 富集分析。

图 2 差异基因的 KEGG 通路分析和 GO 富集分析结果

2.3 差异基因蛋白网络分析 采用 STRING 在线数据库构建 PPI 网络,对 528 个 $|\log_2(FC)|$ 大于 2 的上调基因进行 STRING 蛋白质关系分析,利用 PPI 网络

筛选出差异基因,评分居前 10 位的差异基因主要包括 CDK1、CCNB1、CCNA2、BUB1、KIF11、TOP2A、CCNB2、CHEK1、RRM2 和 ASPM。见图 3。



注:A 为差异表达基因编码蛋白的 PPI 网络;B 为评分最高的前 10 位差异基因。

图 3 差异基因编码蛋白的 PPI 网络及差异基因的筛选

2.4 生存分析 根据基因表达水平差异,将筛选出的 10 个差异基因分为高、低表达组,比较两组患者生存率的差异,结果显示,上述差异基因均与疾病预后相关($P < 0.0001$),见表 1。10 个差异基因表达水平越高,HR 越大,胶质母细胞瘤患者预后越差;其中,CCNA2 基因预测胶质母细胞瘤患者预后的 HR 最大,为 5.563 4(4.073 7,7.598 0),差异有统计学意义($P < 0.0001$)。

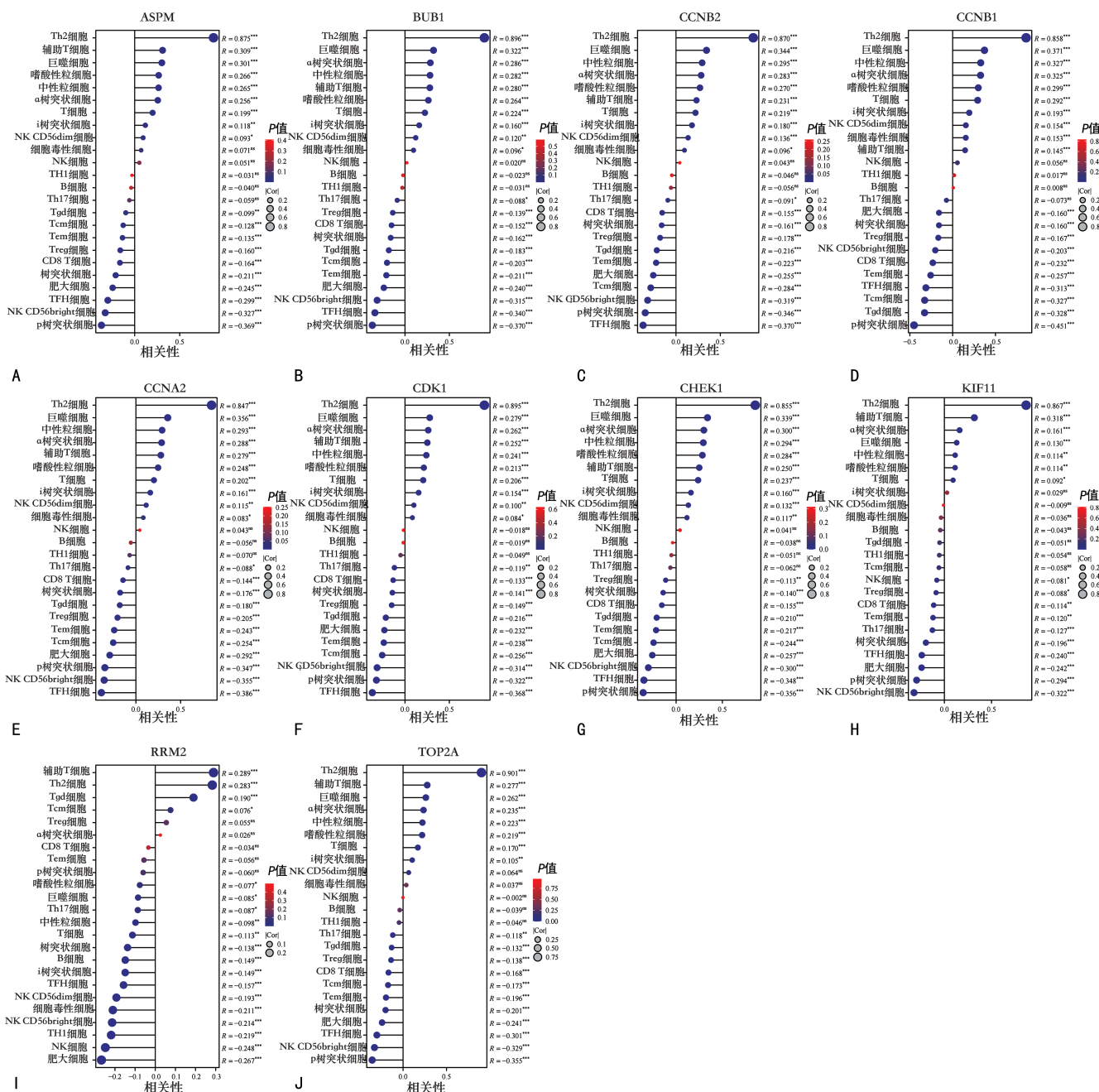
2.5 免疫浸润分析 本研究使用 ssGSEA 算法探索了胶质母细胞瘤组织的 24 种免疫细胞浸润水平和 10 个 hub 基因的表达水平之间的关系。Th2 细胞的浸润水平与 10 个基因的表达呈明显正相关,巨噬细胞与 RRM2 的表达呈负相关,与其他 9 个基因的表达呈明显正相关。肥大细胞与 ASPM、BUB1、CDK1、CCNA2、CCNB2、CHEK1、KIF11、TOP2A、RRM2 的表达呈明显负相关($|R| > 0.2$)。树突状细胞与除 RRM2 外的其他 9 个基因呈明显负相关($|R| > 0.2$)。见图 4。

2.6 ROC 曲线分析 ROC 曲线分析结果显示,10 个基因诊断胶质母细胞瘤的 AUC 及 95%CI 如下:CDK1,0.944 (95%CI: 0.910 ~ 0.978);CCNB1,0.918 (95%CI: 0.827 ~ 1.000);CCNA2,0.904

(95%CI: 0.841 ~ 0.968);BUB1,0.966 (95%CI: 0.936 ~ 0.997);KIF11,0.952 (95%CI: 0.924 ~ 0.980);TOP2A,0.971(95%CI:0.952~0.989);CCNB,20.918(95%CI:0.827~1.000);CHEK1,0.936 (95%CI: 0.874 ~ 0.998);RRM2,0.834 (95%CI: 0.734 ~ 0.934);ASPM,0.905 (95%CI: 0.858 ~ 0.952)。这提示 10 个基因在胶质母细胞瘤中均有较高的诊断价值。

表 1 差异基因表达水平与患者预后的关系

基因	P	HR(95%CI)
CDK1	<0.0001	4.640 4(3.435 3,6.268 2)
CCNB1	<0.0001	5.545 1(4.040 9,7.609 1)
CCNA2	<0.0001	5.563 4(4.073 7,7.598 0)
BUB1	<0.0001	4.638 7(3.434 1,6.265 7)
KIF11	<0.0001	2.710 7(2.061 2,3.565 0)
TOP2A	<0.0001	4.425 7(3.282 6,5.967 0)
CCNB2	<0.0001	5.517 6(4.030 9,7.552 6)
CHEK1	<0.0001	4.528 8(3.353 0,6.116 9)
RRM2	<0.0001	4.677 5(3.467 7,6.309 5)
ASPM	<0.0001	4.405 9(3.278 9,5.920 2)



注:A~J 分别为 ASPM、BUB1、CCNB2、CCNB1、CCNA2、CDK1、CHEK1、KIF11、RRM2、TOP2A 基因表达相关性分析; * $P < 0.05$; * * $P < 0.01$; * * * $P < 0.001$ 。

图 4 胶质母细胞瘤组织的 24 种免疫细胞浸润水平与 10 个差异基因表达水平之间的关系

3 讨论

现有研究表明约 30% 左右的脑肿瘤为神经胶质瘤,占恶性中枢神经系统肿瘤的 80%^[10]。其中,胶质母细胞瘤是最为常见的脑胶质瘤类型,是所有肿瘤中预后最差类型之一。胶质母细胞瘤的发病机制较为复杂,多数研究针对多种胶质母细胞瘤相关基因进行了研究^[11]。然而,胶质母细胞瘤发生的确切机制、表观遗传特征及基因分子特征仍未阐明。随着生物信息学技术的持续发展,通过多种富集分析方法筛选疾病差异基因信息逐渐成为寻找疾病诊断和治疗靶点的新思路^[12]。

本研究选取 GSE31262 基因表达谱数据集作为分析对象,分析了 5 个成人神经干细胞个体样本和 9

个胶质母细胞瘤干细胞个体样本,从中筛选出了 2692 个差异表达基因。通过对这些差异基因进行 KEGG 通路分析,本研究发现上述差异基因主要参与细胞内吞作用、溶酶体、MAPK 信号通路、细胞周期、人乳头瘤病毒感染、人类免疫缺陷病毒感染等。GO 富集分析显示这些差异表达基因主要涉及细胞器裂变、核分裂、胚胎器官发育、染色体分离、轴突生成、胶质细胞再生、蛋白质定位于细胞外围等。其次,本研究采用 Cytoscape 软件分析所得到的各个差异基因之间的关系,并筛选出了可能在胶质母细胞瘤发生发展中发挥关键作用的 10 个差异基因(CDK1、CCNB1、CCNA2、BUB1、KIF11、TOP2A、CCNB2、CHEK1、RRM2、ASPM)。进一步在 TCGA 数据库中分析上述 10 个差异

基因的表达水平与胶质母细胞瘤患者预后的关系,结果显示,上述 10 个差异基因均与患者预后有关($P < 0.0001$)。与低表达组患者比较,高表达组患者的风险比明显升高,临床预后较差。其中,CCNA2 基因预示胶质母细胞瘤患者预后的风险比值最高,该结果提示 CCNA2 基因可能与胶质母细胞瘤患者预后的关系最为密切。

CCNA2 基因定位于人染色体 4 号染色体, q27 区,全长 7 489 bp,在人体多种器官组织中表达^[13]。GAO 等^[14]研究发现,CCNA2 基因在多种肿瘤中表达上调,SAVIRANTA 等^[15]研究显示 CCNA2 基因在癌症转化和进展中发挥关键作用。然而,目前鲜有研究表明 CCNA2 基因与胶质母细胞瘤患者的不良预后有直接关系。本研究通过生物信息学分析发现,CCNA2 基因是胶质母细胞瘤 PPI 信号通路中的差异基因之一,并与胶质母细胞瘤患者的预后相关。目前,有研究发现,CCNA2 基因具有促进胶质母细胞瘤细胞增殖、迁移和侵袭的作用^[16]。这提示 CCNA2 基因在胶质母细胞瘤的发生发展中可能发挥重要作用,为 CCNA2 与胶质母细胞瘤预后相关提供了理论依据,其更深的潜在机制值得进一步探索。

此外,本研究进一步分析了胶质母细胞瘤组织的 24 种免疫细胞浸润水平和 10 个 hub 基因的表达水平之间的关系,为其在胶质母细胞瘤发生发展中的具体作用和机制提供思路。本研究 ROC 曲线分析也进一步证实了这 10 个基因在胶质母细胞瘤中的诊断价值。

综上所述,本研究利用 GEO 数据库筛选出 10 个与胶质母细胞瘤密切相关的基因,对其进行深入研究可能有助于进一步阐释胶质母细胞瘤发生、发展的机制,并且 10 个差异基因有望成为评估胶质母细胞瘤患者诊断和预后的新的标志物。然而,本研究也存在一些局限性,如尚未深入的分子生物学研究验证上述发现。因此,未来研究方向为进一步扩充胶质母细胞瘤样本库,深入分析上述差异基因与疾病进展的关系。

参考文献

- [1] BRAT D J, VERHAAK R G, ALDAPE K D, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas[J]. *N Engl J Med*, 2015, 372(26): 2481-2498.
- [2] SIEGEL R L, MILLER K D, JEMAL A. Cancer statistics, 2019[J]. *CA Cancer J Clin*, 2019, 69(1): 7-34.
- [3] HAINFELLNER J, LOUIS D N, PERRY A, et al. Letter in response to David N Louis, et al, International Society of Neuropathology-Haarlem Consensus Guidelines for nervous system tumor classification and grading, brain pathology[J]. *Brain Pathol*, 2014, 24(6): 671-672.
- [4] LIU Y, YAN W, ZHANG W, et al. MiR-218 reverses high invasiveness of glioblastoma cells by targeting the oncogenic transcription factor LEF1[J]. *Oncol Rep*, 2012, 28(3): 1013-1021.
- [5] 陈玉升, 郭杨, 申汉威, 等. 胶质瘤差异表达基因筛选、功能富集和相关信号通路生物信息学分析[J]. *中华医学杂志*, 2019, 99(29): 2311-2314.
- [6] 王培昌, 侯王付. 基于 GEO 数据库筛选阿尔茨海默病的关键基因及信号通路[J]. *检验医学*, 2022, 37(7): 664-668.
- [7] MALMER B, ISELIUS L, HOLMBERG E, et al. Genetic epidemiology of glioma[J]. *Br J Cancer*, 2001, 84(3): 429-434.
- [8] SANDBERG C J, ALTSCHULER G, JEONG J, et al. Comparison of glioma stem cells to neural stem cells from the adult human brain identifies dysregulated Wnt-signaling and a fingerprint associated with clinical outcome[J]. *Exp Cell Res*, 2013, 319(14): 2230-2243.
- [9] BINDEA G, MLECNIK B, TOSOLINI M, et al. Spatio-temporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer[J]. *Immunity*, 2013, 39(4): 782-795.
- [10] RASMUSSEN B K, HANSEN S, LAURSEN R J, et al. Epidemiology of glioma: clinical characteristics, symptoms, and predictors of glioma patients grade I - IV in the the Danish Neuro-Oncology Registry[J]. *J Neurooncol*, 2017, 135(3): 571-579.
- [11] DIWANJI T P, ENGELMAN A, SNIDER J W, et al. Epidemiology, diagnosis, and optimal management of glioma in adolescents and young adults[J]. *Adolesc Health Med Ther*, 2017, 8: 99-113.
- [12] 刘音, 杨涛, 谢裕赛, 等. 基于 GEO 数据库筛选狼疮性肾炎的关键基因和信号通路[J]. *上海交通大学学报(医学版)*, 2021, 41(6): 749-755.
- [13] KO E, KIM Y, CHO E Y, et al. Synergistic effect of Bcl-2 and cyclin A2 on adverse recurrence-free survival in stage I non-small cell lung cancer[J]. *Ann Surg Oncol*, 2013, 20(3): 1005-1012.
- [14] GAO T, HAN Y, YU L, et al. CCNA2 is a prognostic biomarker for ER+ breast cancer and tamoxifen resistance[J]. *PLoS One*, 2014, 9(3): e91771.
- [15] SAVIRANTA P, KALLIONIEMI O. Human Protein Atlas charts a diverse terrain[J]. *Trends Biotechnol*, 2006, 24(5): 195-197.
- [16] JIANG F, LUO F, ZENG N, et al. Characterization of fatty acid metabolism-related genes landscape for predicting prognosis and aiding immunotherapy in glioma patients[J]. *Front Immunol*, 2022, 13: 902143.

(收稿日期: 2023-01-02 修回日期: 2023-06-26)