整合生物信息学法与加权基因共表达网络分析联合分析脑胶质瘤特征基因

黄晓珊[1]，曾慧[1]，周碧兰[1]，彭电[1]，刘治芳[1]，鲍美华[2*](1.长沙卫生职业学院，长沙 410000；2.长沙医学院，长沙 410000)

摘要：目的 利用脑胶质瘤(glioblastoma, GBM)芯片数据，采用整合生物信息学法和加权基因共表达网络分析(weighted gene correlation network analysis, WGCNA)法，寻找肿瘤发病特征基因、关键通路以及转录调控机制。**方法** 利用 GEO 数据库中高通量基因芯片数据，通过整合生物信息学法筛选出差异基因；利用 WGCNA 分析 GBM 关键基因 hub 基因；采用维恩分析法，整合这些差异基因与 hub 基因，筛选出 GBM 特征基因。采用基因本体论(Gene Ontology，GO)基因功能注释，京都基因和基因组数据库(Kyoto Encyclopedia of Genes and Genomes，KEGG)通路富集，分析 GBM 特征基因所富集的功能和通路。利用 Kaplan-Meier 分析特征基因与 GBM 生成率的关系。利用基因云生物信息平台(GCBI)，分析调控这些特征基因的转录因子。**结果** 经分析，发现 273 个特征基因。这些特征基因主要可影响离子通道、蛋白激酶、γ-氨基丁酸受体功能。*CHD5*，*SYP*，*PHYHIP* 表达水平与 GBM 生存率显著相关；转录因子 Sp1、Sp3、REST 可能是调控这些特征基因的关键因子。**结论** 本研究从多种角度定义了 GBM 的特征基因及调控机制，为其精准治疗提供了依据。

关键词：脑胶质瘤；差异基因；生物信息学；加权基因共表达网络；转录因子

中图分类号：R966 **文献标志码：**A **文章编号：**1007-7693(2020)19-2311-06

DOI: 10.13748/j.cnki.issn1007-7693.2020.19.002

引用本文：黄晓珊，曾慧，周碧兰，等. 整合生物信息学法与加权基因共表达网络分析联合分析脑胶质瘤特征基因[J]. 中国现代应用药学, 2020, 37(19): 2311-2316.

Identification of Characteristic Genes in Glioblastoma by Integrated Bioinformatics and Weighted Gene Correlation Network Analysis

HUANG Xiaoshan[1], ZENG Hui[1], ZHOU Bilan[1], PENG Dian[1], LIU Zhifang[1], BAO Meihua[2*](*1.Changsha Health Vocational College, Changsha 410000, China; 2.Changsha Medical University, Changsha 410000, China)*

ABSTRACT: OBJECTIVE To identify the key characteristic genes, pathways and transcriptional regulatory mechanisms of tumors using glioblastoma(GBM) chip data, integrated bioinformatics methods and weighted gene correlation network analysis (WGCNA). **METHODS** High throughput chip data was downloaded from GEO database. The integrated bioinformatics methods were used to identify the differentially expressed genes. WGCNA was used to analyze the hub genes, which was the key gene in GBM. Differentially expressed genes and hub genes were integrated by the Venn analysis, and the characteristic genes in GBM was identified. The Gene Ontology(GO) and Kyoto Encyclopedia of Genes and Genomes(KEGG) enrichment were used to analyze the functions and pathways enriched by GBM characteristic genes. Kaplan-Meier analysis was used to evaluate the association between the levels of the characteristic genes and patients' overall survival. The online tool Gene-cloud of Biotechnology Information(GCBI) was used to analyze the transcription factors regulating these characteristic genes. **RESULTS** Two hundred and seventy-three characteristic genes were identified. These genes most likely affected the functions of ion channel, protein kinase and GABA receptor. The expression level of *CHD5*, *SYP*, and *PHYHIP* were significantly related to the overall survival of GBM patients. Transcription factor Sp1, Sp3, and REST might be the key transcription factors for these characteristic genes. **CONCLUSION** The present study identifies the characteristic genes of GBM and their regulatory mechanism by various bioinformatics methods. The results may provide new basis for the precise treatment of GBM.

KEYWORDS: glioblastoma; differentially expressed genes; bioinformatics; weighted gene correlation network; transcription factor

基金项目：国家自然科学基金项目(81670427)；湖南省自然科学基金(2019JJ40330，2018JJ5072)；湖南省教育厅项目(20C0142，15C0157)；长沙市科技计划项目(kq1801057)

作者简介：黄晓珊，女，硕士 Tel: (0731)84015809 E-mail: huangxiaoshan0123@yeah.net *通信作者：鲍美华，女，博士，教授 Tel: (0731)88602602 E-mail: mhbao78@163.com

脑胶质瘤(glioblastoma，GBM)是最常见的一种中枢神经系统恶性肿瘤，约占所有原发性恶性脑部肿瘤的 80%。目前对于 GBM 的主要治疗方法有手术、化疗、放疗等。由于其进展快、预后差，其 5 年生存率仅为 5%，患者中位生存期仅 14 个月[1]。深入研究 GBM 的分子机制，发现 GBM 预后的关键分子特征，对改善其预后，控制其进展具有重要意义。

近年来，随着分子生物学的发展，测序及基因芯片等手段被大量应用于生命科学领域。大数据的收集为肿瘤相关基因筛选、分子功能预测、药物靶点研究和分子治疗等，提供了有效的方法和手段。在 GBM 中，已有研究报道了芯片筛选结果，以及 GBM 预后相关的基因如 *MEOX2*、*HMGB1* 等[2-3]。然而，由于单个芯片结果使用的技术平台、临床样本不同，导致现有的分析结果常存在一定的局限性和不一致性。整合生物信息学方法可以对多个数据集进行整合分析，以应用于多种肿瘤标记基因的筛选[4]。加权基因共表达网络分析(weighted gene correlation network analysis，WGCNA)是利用基因表达模式的不同，挖掘出相似表达模式的基因，定义为模块的一种算法。同一模块的基因很可能是功能紧密相关的或同一条信号通路的成员，具有特定的意义。将整合生物学分析得到的差异基因(differentially expressed genes，DEGs)与 WGCNA 分析得到的关键基因进行整合，获得与疾病临床特征密切相关的、更加可靠的特征基因，对于 GBM 的分子分型、治疗和预后将有重要意义。

本研究对基因表达数据库 (gene expression omnibus，GEO)数据库中的已知芯片数据进行分析。通过整合生物信息学法整合多个芯片数据得到 DEGs，并利用 WGCNA 筛选关键基因 hub 基因，将 DEGs 与 hub 基因综合分析，得到 GBM 的特征基因。并对这些特征基因进行功能富集，以建立针对 GBM 的分子表达特征基因谱，识别关键分子事件和路径。并对特征基因的调控网络进行分析，为 GBM 的精准医疗提供理论依据。

1　材料与方法

1.1　数据下载及预处理

利用 NCBI(National Center for Biotechnology Information)平台下的 GEO 数据库(http://www.ncbi.nlm.nih.gov/geo/)下载 Affymetrix 芯片数据集 GSE104291、GSE30563、GSE4290 及 GSE50161。4 个数据集共含 118 个 GBM 样本，41 个正常脑组织样本。利用 R 软件的软件包进行数据标准化预处理。

1.2　DEGs 基因筛选

采用 R 软件的 limma 软件包对 4 个数据集的 DEGs 进行筛选，筛选标准：$P<0.05$，差异倍数>2.0。采用 RRA 软件包对 4 个数据集的 DEGs 进行整合分析，并用热图可视化。

1.3　WGCNA 分析 hub 基因

利用 R 软件的 WGCNA 软件包对 GSE4290 的 4 890 个表达数据进行共表达网络的构建。简要步骤如下：构建基因共表达相似性矩阵，确定软阈值后再将相似性矩阵转换为邻接矩阵。之后将邻接矩阵转换成拓扑矩阵，基于拓扑覆盖法(topological overlap measure，TOM)，使用层次聚类法，对基因进行聚类，并按照混合动态剪切树的方法确定基因模块。通过模块与表型相关性分析、模块中基因与表型相关性分析，最终得到与 GBM 相关性最高的 hub 基因。

1.4　GBM 相关特征基因筛选

采用韦恩图对“1.2”项筛选出的 DEGs 和“1.3”项分析出的 hub 基因进行分析，筛选二者重合的基因为 GBM 特征基因。

1.5　基因功能注释和通路分析

对于“1.4”项筛选出的特征基因，利用 DAVID 网站生物信息资源数据库(https://david.ncifcrf.gov/)中的基因本体论(gene ontology，GO)和京都基因和基因组数据库(Kyoto Encyclopedia of Genes and Genomes，KEGG)进行功能和通路富集分析。

1.6　转录因子分析

为进一步分析基因调控网络，本研究对“1.4”项筛选出的特征基因进行了转录因子分析。采用基因云生物信息平台(GCBI) (https://www.gcbi.com.cn)分析基因上游−2 000 bp 到 500 bp 的转录因子位点，并将数据可视化。

1.7　特征基因与生存率的关系分析

为进一步分析特征基因与 GBM 生存率的关系，本研究利用 TCGA-GBM 数据库，获取特征基因的表达情况和患者生存数据，并利用 PROGgeneV2 在线工具绘制 Kaplan-Meier 曲线。同时利用 Oncomine(https://www.oncomine.org)分析正常脑组织和 GBM 组织中特征基因的表达水平。

2 结果

2.1 DEGs 筛选结果

通过整合生物信息学方法，对数据集 GSE104291、GSE30563、GSE4290 和 GSE50161 进行整合分析，结果发现 321 个差异表达的基因，其中 57 个显著上调，264 个显著下调。其中 20 个显著上调和 20 个显著下调的基因见图 1。

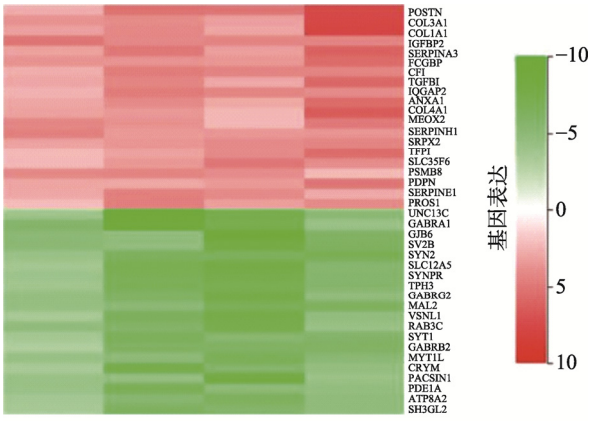

图 1 20 个显著上调和 20 个显著下调的差异基因热图

Fig. 1 Heatmap image of the top 20 upregulated and the top 20 downregulated integrated DEGs

2.2 Hub 基因分析结果

WGCNA 分析结果见图 2，青绿色和蓝色模块与疾病相关性最高。其中青绿色模块含 3 293 个基因，蓝色模块含 387 个基因。

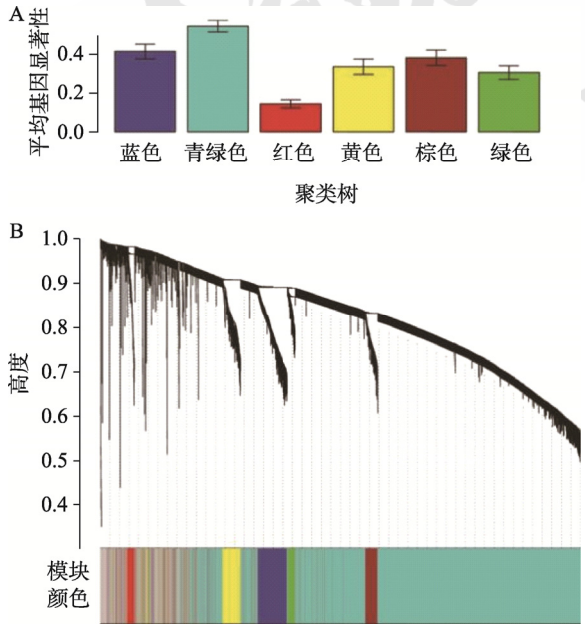

图 2 WGCNA 分析与脑胶质瘤相关的 hub 基因

A–各模块与 GBM 的关系；B–聚类树分析图。

Fig. 2 WGCNA analysis of the hub genes in GBM

A–relationship between each module and GBM; B–clustering tree diagram.

2.3 GBM 相关特征基因及功能注释结果

进一步对整合生物信息学法筛选出的 321 个 DEGs 与 WGCNA 法分析得到的青绿色、蓝色模块的 hub 基因进行韦恩图分析，青绿色模块与 DEGs 有 256 个基因重合，蓝色模块与 DEGs 有 17 个基因重合。这 273 个基因为 GBM 特征基因。进一步对这 273 个特征基因进行基因注释，发现这些基因主要与蛋白结合(包括离子通道结合、蛋白激酶结合等)、蛋白激酶活性、γ–氨基丁酸(γ-aminobutyric acid，GABA)受体活性、神经突触传递、吗啡成瘾通路等分子行为有关，结果见图 3~4。

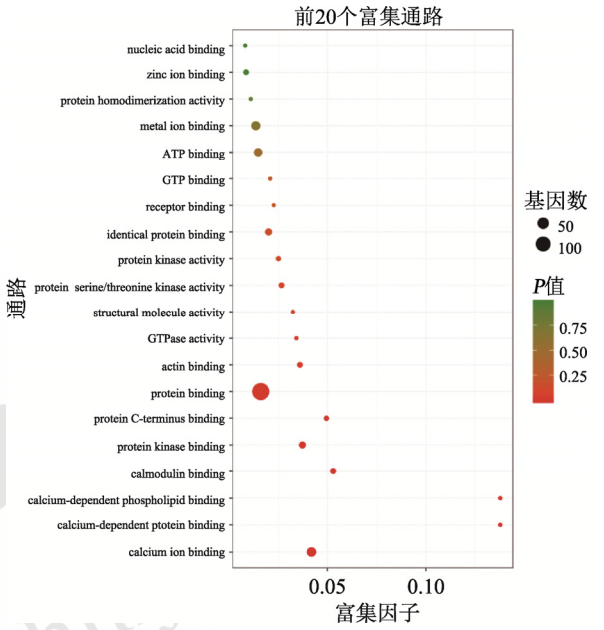

图 3 GO 基因功能富集分析图

Fig. 3 Bubble plots of GO enrichment

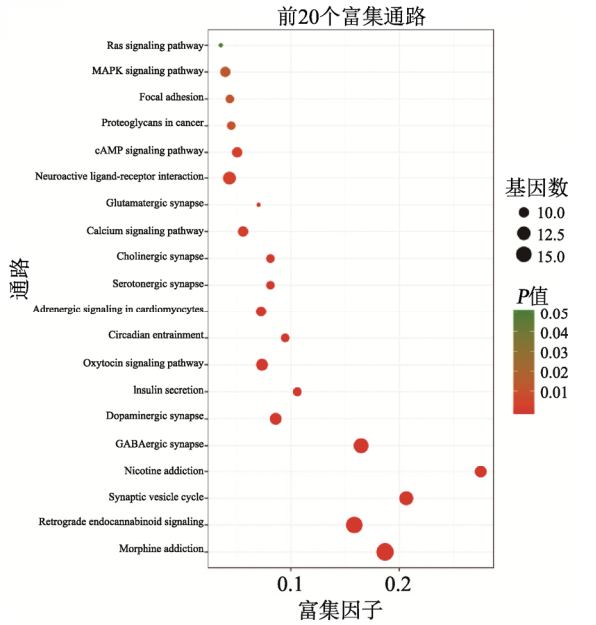

图 4 KEGG 基因通路富集分析图

Fig. 4 Bubble plots of KEGG enrichment

2.4 转录因子分析结果

转录因子调控网络见图 5，这些特征基因(宝蓝色)可被多种转录因子调控(绿色)。其中，转录因子 Sp1、Sp3、REST 可同时调控多个基因的表达，而 SERPINE1 可被 SP1 等 20 个转录因子调控。

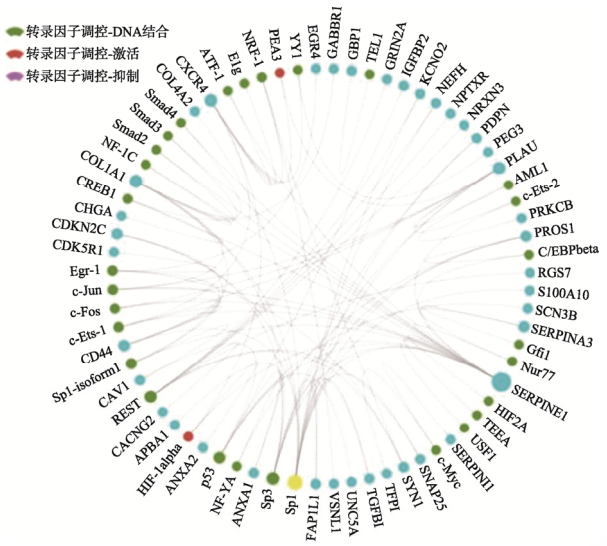

图 5 转录因子与特征基因调控网络

Fig. 5 Transcription factors and characteristic gene regulation networks

2.5 特征基因与 GBM 预后的关系

通过对特征基因与患者生存的 Kaplan-Meier 分析，发现 *CHD5*、*SYP*、*PHYHIP* 基因水平与 GBM 生存率显著相关，结果见图 6A。高表达 *CHD5*、*SYP*、*PHYHIP* 的患者生存时间明显高于低表达的患者。进一步分析表明，GBM 患者 *CHD5*、*SYP* 和 *PHYHIP* 的水平明显低于正常对照组，结果见图 6B。

3 讨论

GBM 是一种发生于脑内的最常见的肿瘤。阐明 GBM 的分子特征，对其治疗靶点和预后标记的发现具有重要意义。现有的芯片研究为进一步分析 GBM 的分子特征提供了可能。

生物信息学是结合生命科学和计算机科学的一门新兴学科。它通过综合利用生物学、计算机科学和信息技术，揭示大量而复杂的生物数据所赋有的生物学奥秘。整合生物信息学法采用整合分析的方法，更为科学的分析多个数据集，获得更可靠的数据[5]；WGCNA 通过聚类分析的方式，挖掘具有相似特征的数据。将两者结合，对芯片数据进行分析，可得到更为可靠，且与疾病特征密切相关的特征基因。

本研究通过对 GEO 数据库的 4 个数据集进行整合分析，并联合 WGCNA 方法，得到 273 个特征基因。对这些基因进行功能注释，发现这些基

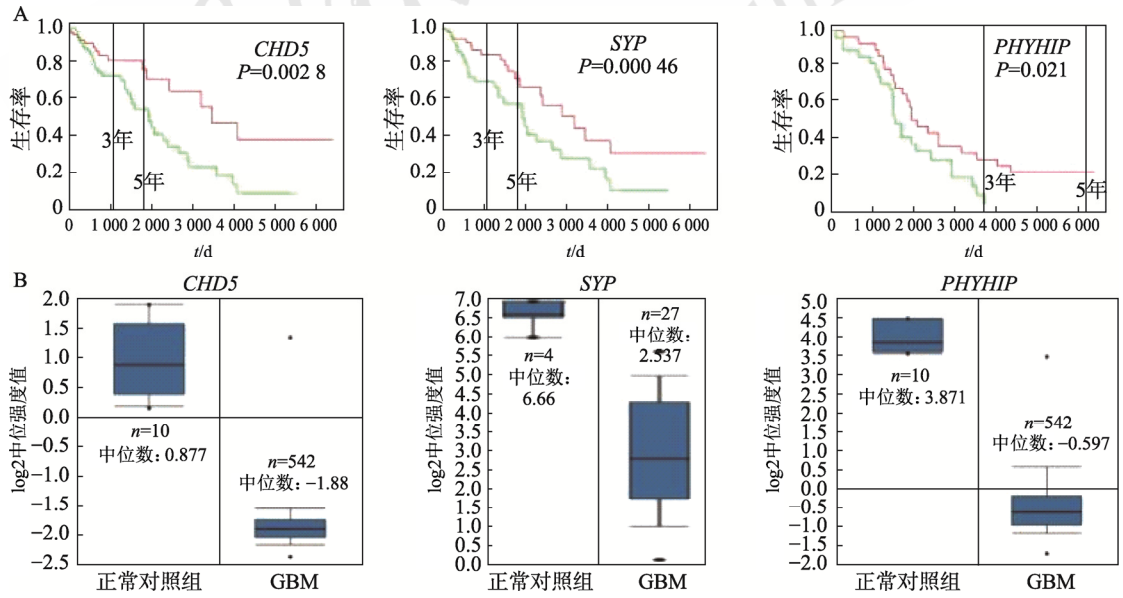

图 6 特征基因 *CHD5*、*SYP* 和 *PHYHIP* 表达水平与 GBM 预后的关系

A–特征基因 *CHD5*、*SYP* 和 *PHYHIP* 表达水平与 GBM 预后的 Kaplan-Meier 生存曲线(红色：高表达，*n*=103；绿色：低表达，*n*=103)；B–特征基因 *CHD5*、*SYP* 和 *PHYHIP* 在正常脑组织和 GBM 组织中的表达水平。

Fig. 6 Relationship between the expression level of characteristic genes *CHD5*, *SYP* and *PHYHIP* and the prognosis of GBM

A–survival analysis by Kaplan-Meier method for *CHD5*, *SYP* and *PHYHIP* (red: high expression, *n*=103; green: low expression, *n*=103); B–expression of *CHD5*, *SYP* and *PHYHIP* in GBM tissues and normal tissues.

因主要与蛋白结合(包括金属离子结合、蛋白激酶结合等)、蛋白激酶活性、GABA 受体活性及神经突触传递、吗啡成瘾通路等分子行为有关。

离子通道包括钙、钾、氯等通道，与肿瘤的发生发展关系密切。离子通道通过调节细胞膜电位、细胞周期、细胞体积、细胞内钙浓度等影响肿瘤细胞的增殖与凋亡，调控肿瘤的发展与转移。离子通道阻断剂可能成为肿瘤治疗的新靶标[6]。本研究发现的特征基因中有影响钙离子通道的功能的基因 *SYT*、*CABP1*、*MICU3* 等；影响钾离子通道的功能的基因 *KCNJ3/4/6/9* 等；影响氯离子通道的功能基因 *GABR* 等。它们可能通过影响离子通道功能，进而影响 GBM 的发生发展，也可能是药物作用的关键靶点。

蛋白激酶与肿瘤发生发展的关系密切。蛋白激酶家族如蛋白激酶 C 家族、促分裂原活化的蛋白激酶(mitogen activated protein kinase，MAPK)家族与肿瘤细胞的增殖、分化、凋亡、迁移、多药耐药性和功能同步化密切相关[7-8]。本研究发现 *PRKCZ/B*、*PAK1/3/6*、*MAP3K9* 等基因是 GBM 的特征基因。它们可能通过影响肿瘤细胞增殖和耐药的多重途径，影响肿瘤的发生发展，也可能是药物作用的新靶标。

GABA 及其受体通路与肿瘤增殖和侵袭也密切相关。GABA 是哺乳动物中枢神经系统重要的抑制性神经递质。研究表明 GABA 不仅与神经细胞间的信息传递相关，还与多种肿瘤的增殖与侵袭相关。GABA 与其受体结合，可调控基质金属蛋白酶、细胞内钙浓度、MAPK 通路等过程，是肿瘤治疗的潜在靶点[9]。本研究分析发现，在 GBM 中，GABA 受体 A1/A2/A4/ B2/B3/D/G2 发生明显改变，且为特征基因。在 KEGG 通路分析中，它们富集于吗啡成瘾通路。这些基因可能通过多种机制影响 GBM 的发生。

CHD 是染色质调节域蛋白家族成员，编码染色质重塑及 DNA 螺旋酶蛋白 5。CHD5 在大脑中高度表达。它通过改变神经元基因、转录因子和 SWI/SNF 重构酶的表达水平，参与神经再生、衰老、阿尔茨海默病等过程[10]。也有研究表明，CHD5 具有肿瘤抑制作用。在 GBM 中，CHD5 低表达与较差的预后相关[11]。

SYP 是编码脑和内分泌细胞中的突触素的基因。由于突触素在突触中广泛分布，是突触数量的标记物。突触素的变化可引起小鼠行为的改变[12]。在肿瘤中，突触素被认为与结直肠癌的生存呈正相关，并在肿瘤治疗过程中发挥作用[13]。

PHYHIP 又称 PAHX-AP1，是脑特异性蛋白，位于 8 号染色体的 p 臂上，编码 phytanoyl-CoA 2-羟化酶相互作用蛋白。乳腺癌患者的 8 号染色体 p 臂经常丢失，从而导致乳腺癌患者的 PHYHIP 表达下调。有研究证实，PHYHIP 与 PAHX 相互作用，促进唐氏综合征患者双特异性酪氨酸磷酸化调控激酶 1A(DYRK1A)的核转位[14]。然而，PHYHIP 在肿瘤中的功能并不清楚。本研究表明，*PHYHIP* 是 GBM 癌变过程中的一个关键基因，且是 GBM 预后的一个新的标记物。

转录因子调控是基因调控的关键机制，也是肿瘤治疗的重要靶点。本研究发现转录因子 Sp1、Sp3、REST 可同时调控多个特征基因的表达，而 SERPINE1 可被 SP1 等 20 个转录因子调控。影响这些转录因子可能是调控这些基因表达，进而实现肿瘤治疗的手段之一。

本研究虽然筛选出了一些特征基因，但仍存在一些局限性，如芯片数据需要进一步的实验验证；芯片纳入的患者及健康对照者性别、年龄、身体状况和地域分布等基本信息缺乏；芯片的 mRNA 检测结果与蛋白表达的一致性需要进一步明确；用于表达量分析的正常对照样本数偏少等。因此，GBM 的特征基因仍需要更全面的数据支撑。

综上所述，本研究利用多种生物信息学方法，筛选了 GBM 的特征基因，并从不同角度定义了 GBM 发病的分子机制，提出了可能的药物治疗靶点及基因调控方式，为 GBM 的分子分型和精准治疗提供了思路。

REFERENCES

[1] OSTROM Q T, BAUCHET L, DAVIS F G, et al. Response to “the epidemiology of glioma in adults: A ‘state of the science’ review” [J]. Neuro-oncology, 2015, 17(4): 624-626.

[2] WANG L N, WEI B, HU G Z, et al. Screening of differentially expressed genes associated with human glioblastoma and functional analysis using a DNA microarray [J]. Mol Med Rep, 2015, 12(2): 1991-1996.

[3] CANCER GENOME ATLAS RESEARCH NETWORK. Comprehensive genomic characterization defines human glioblastoma genes and core pathways [J]. Nature, 2008, 455(7216): 1061-1068.

[4] NI M W, LIU X K, WU J R, et al. Identification of candidate

biomarkers correlated with the pathogenesis and prognosis of non-small cell lung cancer via integrated bioinformatics analysis [J]. Front Genet, 2018(9): 469. Doi:10.3389/fgene.2018.00469.

[5] ZHANG Y, ZHONG L, LOU Q, et al. Study on the HPV-positive oropharyngeal cancer features gene based on GEO database by bioinformatics [J]. Chin J Mod Appl Pharm (中国现代应用药学), 2018, 35(5): 638-641.

[6] 韩璐, 张红, 狄翠霞, 等. 离子通道与肿瘤关系研究现状 [J]. 生理科学进展, 2014, 45(3): 225-229.

[7] SHAO Y, SU Y, YAO M. Updated correlation of the protein kinase C family with cancer [J]. J Me Postgraduates, 2009, 22(6): 661-664.

[8] 刘伟, 黄玮, 李瑞琴. MAPK/ERK 信号传导通路与肿瘤发生的相关机制研究进展[J]. 中国现代医药杂志, 2016, 18(8): 97-100.

[9] LIAO Y, WANG F, CHEN L. Relationship of γ-aminobutyric acid and its receptors with proliferation and invasion of tumor [J]. Chin J Cancer Biother, 2009, 16(1): 93-96.

[10] POTTS R C, ZHANG P S, WURSTER A L, et al. CHD5, a brain-specific paralog of Mi2 chromatin remodeling enzymes, regulates expression of neuronal genes [J]. PLoS One, 2011, 6(9): e24515.

[11] WANG L, HE S M, TU Y Y, et al. Downregulation of chromatin remodeling factor CHD5 is associated with a poor prognosis in human glioma [J]. J Clin Neurosci, 2013, 20(7): 958-963.

[12] SCHMITT U, TANIMOTO N, SEELIGER M, et al. Detection of behavioral alterations and learning deficits in mice lacking synaptophysin [J]. Neuroscience, 2009, 162(2): 234-243.

[13] MAEDA I, TAJIMA S, ARIIZUMI Y, et al. Can synaptophysin be used as a marker of breast cancer diagnosed by core-needle biopsy in epithelial proliferative diseases of the breast [J]. Pathol Int, 2016, 66(7): 369-375.

[14] LEE Z H, KIM H, AHN K Y, et al. Identification of a brain specific protein that associates with a refsum disease gene product, phytanoyl-CoA alpha-hydroxylase [J]. Brain Res Mol Brain Res, 2000, 75(2): 237-247.

收稿日期：2019-11-08
(本文责编：李艳芳)