doi:10.11659/jjssx.10E024091

·基础研究·

生物信息学分析结直肠癌预后不良相关的基因

王立杰,陈 思,常 程,高凯月,张海佳 (吉林大学白求恩第一医院普通外科中心结直肠肛门外科,吉林 长春 130000)

[摘 要] **目的** 通过生物信息学方法筛选与结直肠癌(CRC)预后不良相关的基因。**方法** 从基因表达综合数据库(GEO)数据库中获取GSE74602、GSE110223、GSE113513和GSE141174的基因表达谱数据集,包含65例CRC组织和65例正常组织样本。通过GEO2R工具和Venn软件,筛选CRC组织与正常组织之间的差异表达基因(DEGs),并利用Venn软件在DEGs中提取一致性基因。通过STRING数据库构建蛋白质-蛋白质相互作用(PPI)网络识别核心基因,并利用Kaplan-Meier Plotter和GEPIA分析核心基因。再采用京都基因与基因组百科全书(KEGG)通路富集分析13个选定基因。**结果** 在4个数据集中共得到171个DEGs,其中上调基因148个,下调基因23个。上调的DEGs在氧化还原过程、碳酸氢盐运输、消化、离子跨膜运输和单碳代谢过程中富集;下调DEGs主要富集于细胞增殖的正调控。生存曲线分析显示,87个基因中有30个与生存预后差显著相关。GEPIA分析显示,与正常组织相比,30个基因中有13个在CRC组织中高表达,其中MYC和FGFR3在CRC通路中显著富集。**结论** 本研究筛选出MYC和FGFR3在CRC中显著上调且与预后不良密切相关,二者可能成为CRC患者潜在预后标志物和治疗靶点。

[关键词] 结直肠癌;差异表达基因;蛋白质;预后;生物信息学分析
[中图分类号] R735.3 [文献标志码] A [收稿日期] 2024-10-24

Bioinformatics analysis of genes associated with poor prognosis in colorectal cancer

WANG Li-jie, CHEN Si, CHANG Cheng, GAO Kai-yue, ZHANG Hai-jia (Department of Colorectal & Anal Surgery, General Surgery Center, First Bethune Hospital of Jilin University, Changchun Jilin 130000, China)

Abstract: Objective To screen genes associated with poor prognosis of colorectal cancer (CRC) through bioinformatics analysis. **Methods** The gene expression profiles of GSE74602, GSE110223, GSE113513 and GSE141174 were obtained from Gene Expression Omnibus (GEO) database, including samples from 65 CRC tissues and 65 normal tissues. Differentially expressed genes (DEGs) between CRC tissues and normal tissues were screened out by GEO2R tool and Venn software, and the consistent genes were extracted from DEGs by Venn software. A protein-protein interaction (PPI) network was constructed by the STRING database to identify key genes, which were analyzed by Kaplan-Meier Plotter and GEPIA. Kyoto encyclopedia of genes and genomes (KEGG) pathway enrichment analysis was conducted on the 13 selected genes. **Results** There were 171 DEGs obtained from the four datasets, including 148 up-regulated genes and 23 down-regulated genes. Up-regulated DEGs were enriched in the redox processes, bicarbonate transport, digestion, ion trans-membrane transport, and one-carbon metabolism; and down-regulated DEGs were enriched in the positive regulation of cell proliferation. The survival curve analysis showed that 30 of the 87 genes were significantly associated with poor survival prognosis. GEPIA showed that 13 of the 30 genes were highly expressed in CRC tissues compared to normal tissues, among which MYC and FGFR3 markedly enriched in the CRC pathway. **Conclusion** This study identifies that MYC and FGFR3 are significantly up-regulated in CRC and closely associated with poor prognosis, suggesting that they may serve as potential prognostic markers and therapeutic targets for CRC patients.

Keywords: colorectal cancer; differentially expressed genes; protein; prognosis; bioinformatical analysis

结直肠癌(colorectal cancer, CRC)是一种常见的消化道恶性肿瘤,其发病率和病死率长期位居恶性肿瘤前列。在我国,CRC的发病率与病死率持续攀升,已成为威胁国民身心健康的重大公共卫生问题[1-2]。然而,目前CRC的预后评估与治疗靶点选择仍存在局限性,因此亟需探索更可靠的预后生物标志物,为优化治疗策略、阐明潜在发病机制提供理论依据[3-6]。

随着生物信息学技术与基因芯片技术的成熟应用,通过高通量检测筛选差异表达基因(differentially expressed genes, DEGs)已成为肿瘤机制研究的重要手段。该技术不仅具有高效可靠的优势,还推动了大量基因表达数据的积累与公共数据库的建立,为挖掘肿瘤相关分子机制提供了丰富的数据源[7]。以往针对CRC的生物信息学研究证实,综合运用生物信息学分析方法可有效揭示疾病潜在机制,为标志物筛选与靶点发现奠定基础[7]。

本研究基于公共数据库资源开展CRC相关生物信息学分析,从基因表达综合数据库(Gene Expression Omnibus, GEO)中筛选合适数据集;筛选CRC组织与

[基金项目] 吉林省科技厅资助项目(20240304063SF)
[通信作者] 张海佳,E-mail:zhj296064233@jlu.edu.cn

正常组织中的DEGs，随后进行全域功能富集分析（Gene Ontology, GO）及京都基因与基因组百科全书（Kyoto encyclopedia of genes and genomes，KEGG）通路分析；进一步通过构建蛋白质-蛋白质相互作用（protein-protein interaction，PPI）网络识别核心基因，并利用Kaplan-Meier Plotter在线数据库对核心DEGs进行预后价值评估。通过上述生物信息学数据挖掘，旨在筛选出可靠的CRC预后生物标志物，为其早期筛查、预后评估及治疗靶点开发提供新的分子依据与理论支持。

1 材料与方法

1.1 微阵列数据信息

本研究的基因表达数据来源于美国国家生物技术信息中心基因表达综合数据库，共纳入4个CRC相关数据集，分别为GSE74602、GSE110223、GSE113513和GSE141174，所有数据均包含CRC组织与正常组织的基因表达谱信息。各数据集的平台及样本信息如下：GSE74602基于GPL6104平台（Illumina HumanRef-8 v2.0 Expression BeadChip），包含30对样本（30例CRC组织和30例正常组织），数据来源于新加坡基因研究所；GSE141174基于GPL6104平台，包含8对样本（8例CRC组织和8例正常组织），数据来源于瑞典临床微生物学实验室；GSE110223基于GPL96平台（[HG-U133A] Affymetrix Human Genome U133A Array），包含13对样本（13例CRC组织和13例正常组织），数据来源于希腊生物、药物化学与生物技术研究所；GSE113513基于GPL15207平台（[PrimeView] Affymetrix Human Gene Expression Array），包含14对样本（14例CRC组织和14例正常组织），数据来源于福建中医药大学。

1.2 DEGs的筛选与处理

采用GEO2R在线工具[8]对上述4个数据集进行差异表达分析，以筛选CRC组织与正常组织间的DEGs，筛选标准为：|logFC|>1，调整P<0.05。通过Venn在线工具对筛选得到的DEGs进行交集分析，获取4个数据集共有的DEGs。根据logFC值对共同DEGs进行分类：logFC>0的基因为上调基因，logFC<0的基因为下调基因。

1.3 GO功能富集分析与KEGG通路富集分析

GO功能富集分析通过定义基因及其RNA或蛋白质产物的功能，挖掘高通量转录组或基因组数据中独特的生物学特性[9]，主要包括分子功能（molecular function，MF）、细胞成分（cellular component，CC）、生物过程（biological process，BP）三个维度。KEGG通路富集分析是整合基因组、疾病、生物通路、药物及化学物质信息的数据库集合[10]，用于解析基因参与的信号通路。本研究采用DAVID在线生物信息学工具[11]对共同DEGs进行上述富集分析，该工具可通过对大量基因或蛋白质的注释挖掘其核心功能，最终可视化呈现显著富集的BP、MF、CC条目及KEGG通路。

1.4 PPI网络构建

采用STRING在线数据库[12]构建共同DEGs的PPI网络，以评估基因编码蛋白质之间的潜在相互作用关系，网络构建的富集显著性阈值设为PPI富集P<0.05，并通过该工具可视化蛋白质间的相互作用关系。

1.5 生存分析及核心基因RNA测序表达

采用Kaplan-Meier Plotter在线数据库[13]对筛选出的核心DEGs进行预后价值评估。该数据库整合了EGA、TCGA及GEO（仅限Affymetrix微阵列）的基因表达数据与临床预后信息，通过计算*Log-rank P*值和风险比及其95%置信区间，筛选出具有显著预后意义（P<0.05）的核心基因并可视化呈现。为进一步验证核心DEGs的表达趋势，利用GEPIA在线工具分析其在GTEx项目与TCGA数据库中数千例样本中的RNA测序表达数据[14]。

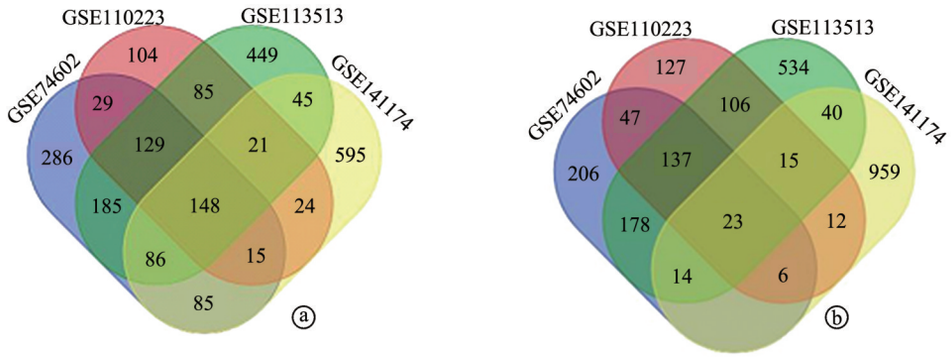
2 结果

2.1 CRC中DEGs的鉴定

共纳入65例CRC组织和正常组织。通过GEO2R在线工具，从GSE74602、GSE110223、GSE113513和GSE141174数据集中分别提取了1 643、1 269、3 497和2 180个DEGs。进一步使用Venn图软件对4个数据集的DEGs进行交集分析，最终在CRC组织中共鉴定出171个常见DEGs（均满足|logFC|>1），其中上调基因148个，下调基因23个（图1）。

2.2 CRC中DEGs的GO功能富集分析及KEGG通路富集分析

通过DAVID软件对171个DEGs进行GO功能富集分析显示，DEGs在BP、CC及MF中均存在显著富集。在BP方面，上调的DEGs主要富集于氧化还原过程、碳酸氢盐运输、消化、离子跨膜运输和单碳代谢过程，下调的DEGs则主要富集于细胞增殖的正调控、细胞对胰岛素刺激的反应和细胞外基质。在CC方面，上调的DEGs富集于细胞外泌体、膜整体成分、质膜、胞外空间和质膜整体成分，下调的DEGs分布于细胞质、胞外区域和细胞外基质。在MF方面，上调的DEGs富集于锌离子结合、氧化还原酶活性、转运蛋白活性和激素活性。KEGG通路富集分析结果显示，上调的DEGs主要富集在代谢途径、矿物质吸收、氮代谢和胰腺分泌中，而下调的DEGs主要富集在叶酸参与的一碳代谢，见表1。



a:上调 DEGs;b:下调 DEGs

图1　Venn 图分析结果

表1　CRC 中 DEGs 的 KEGG 通路富集分析

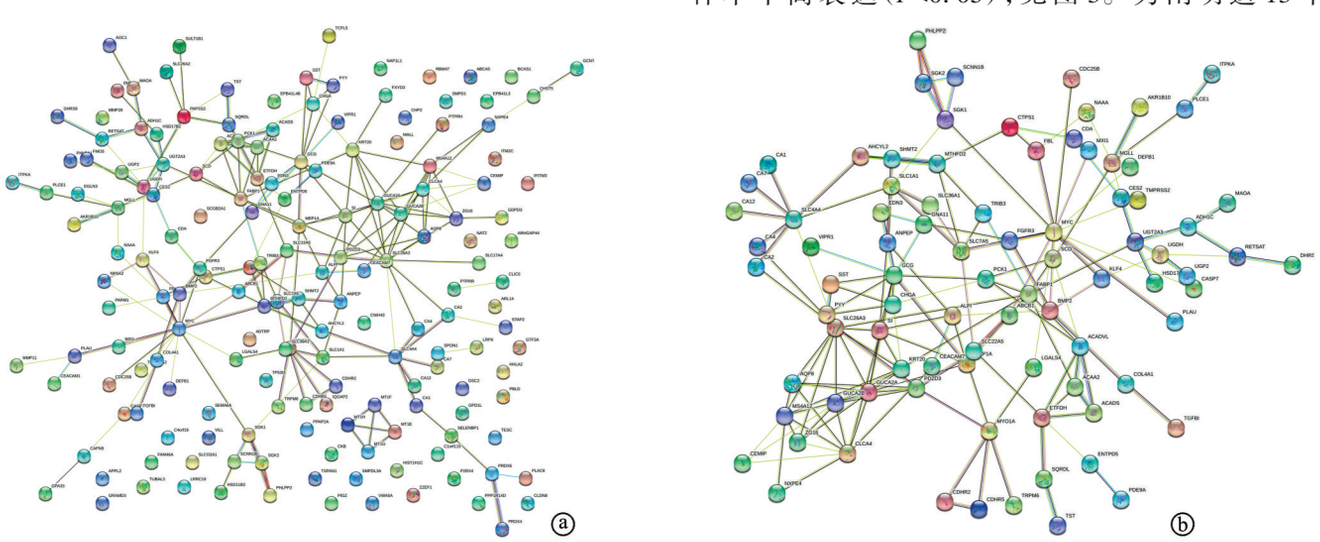
基因	术语	基因数量	基因占比(%)	*P*	校正后 *P*
上调基因	hsa00910:氮代谢	5	3.378	0.000	0.005
	hsa01100:代谢途径	30	20.270	0.000	0.005
	hsa04978:矿物质吸收	6	4.054	0.000	0.006
	hsa04964:近端小管碳酸氢盐回收	4	2.703	0.002	0.075
	hsa00920:硫代谢	3	2.027	0.004	0.121
	hsa00040:戊糖和葡萄糖醛酸的相互转化	4	2.703	0.006	0.143
	hsa00071:脂肪酸降解	4	2.703	0.012	0.239
	hsa00983:药物代谢-其他酶类	4	2.703	0.016	0.267
	hsa04972:胰腺分泌	5	3.378	0.023	0.338
	hsa00561:Glycerolipid 新陈代谢	4	2.703	0.031	0.393
	hsa00830:视黄醇新陈代谢	4	2.703	0.039	0.443
	hsa00982:药物代谢-细胞色素 P450	4	2.703	0.045	0.443
	hsa04976:胆汁分泌	4	2.703	0.047	0.443
	hsa00052:半乳糖代谢	3	2.027	0.048	0.443
下调基因	hsa00670:叶酸参与的一碳代谢	2	8.696	0.031	1.000

2.3　PPI 网络

171 个 DEGs 导入 PPI 中,该蛋白互作包括 171 个互作蛋白和 204 条连接线,其中涵盖 23 个下调基因和 148 个上调基因(图 2a)。然后进行 STRING 分析,结果显示 171 个互作蛋白中共识别出 87 个相互关联的核心蛋白(图 2b)。

2.4　核心基因分析

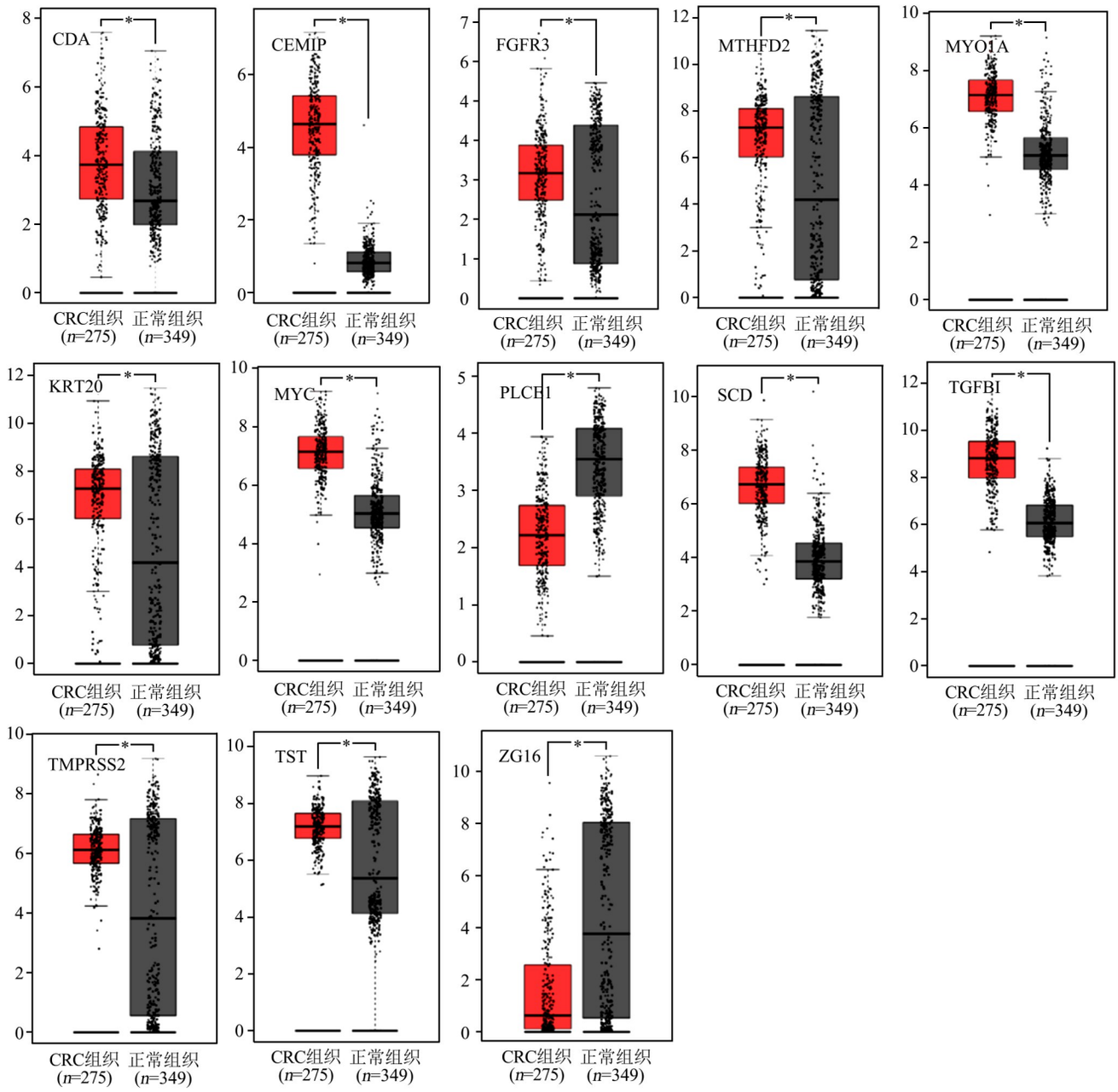
通过 Kaplan-Meier Plotter 分析 PPI 筛选出的 87 个核心蛋白基因的生存曲线数据发现,表达 CEMIP、CDA、FGFR3、KRT20、MTHFD2、MYC、MYO1A、SCD、PLCE1、TGFBI、TMPRSS2、TST、ZG16、HSD17B2、TRPM6、CES2、AHCYL2、BMP2、SLC22A5、SI、PHLPP2、AKR1B10、DHRS9、MGLL、SLC36A1、NXPE4、SLC1A1、ANPEP、CEACAM7、ALPI 的患者生存率较低,表达其余 57 个基因的患者生存率无显著变化(P<0.05)。利用 GEPIA 挖掘出 30 个基因在 CRC 样本和正常样本中差异表达,与正常样本相比,30 个基因中有 13 个在 CRC 样本中高表达(P<0.05),见图 3。为阐明这 13 个



a:171 个 DEGs 蛋白互作图;b:87 个核心蛋白互作图　节点代表蛋白;边表示蛋白的相互作用

图2　STRING 在线数据库构建 PPI 网络

DEGs可能参与的生物学通路，通过DAVID数据库对KEGG通路进行了显著性富集分析，结果显示，MYC和FGFR3基因在CRC通路中显著富集（MYC $P<0.05$，FGFR3 $P<0.01$）。



*：$P<0.05$

图3 13个高表达基因

3 讨论

为了寻找更有用的CRC预后生物标志物，本研究基于GSE74602、GSE110223、GSE113513和GSE141174 4个基因图谱数据集，并采用生物信息学方法系统分析CRC组织和正常组织中的DEGs，以筛选潜在的预后标志物及治疗靶点。

本研究研究通过GEO2R和Venn软件筛选出了171个DEGs，其中148个上调，23个下调。然后，采用DAVID方法进行GO功能富集分析和KEGG通路富集分析，结果显示，上调的DEGs在氧化还原过程、细胞外泌体、锌离子结合、代谢途径、矿物质吸收中富集，提示这些基因可能通过调控细胞代谢和微环境稳态参与CRC的发生；下调的DEGs在细胞增殖的正调控中富集，提示DEGs可能通过促进细胞增殖参与CRC的发生发展。为进一步挖掘核心基因，利用STRING在线数据库构建171个节点和204条边的PPI网络，为

筛选关键功能基因提供了相互作用关系基础。采用Kaplan-Meier Plotter对87个基因进行预后评估,发现30个DEGs与CRC患者预后不良显著相关。经GEPIA数据库验证发现,这30个DEGs中有13个在CRC组织中高表达。其中,MYC、FGFR3基因在CRC通路中显著富集,提示二者可能是CRC预后评估的关键指标及潜在的治疗靶点。

MYC作为经典癌基因,在CRC组织中高表达且被广泛激活,其通过调控细胞生长、侵袭和转移参与CRC的进展,多年来一直被作为抗肿瘤治疗的研究热点[15]。已有研究证实,MYC介导的致癌信号通路与CRC的发生密切相关,且MYC家族在多种癌症的发生发展中发挥核心作用[16]。FGFR3作为酪氨酸激酶家族成员,通过调控细胞增殖、细胞迁移和血管生成等过程影响肿瘤进展。研究表明,靶向抑制FGFR家族成员可有效减少CRC细胞的增殖、转移和侵袭,如黄连根有效成分通过靶向FGFR1抑制CRC细胞增殖[17],miR-99a-SP通过负向调控FGFR3发挥抑癌作用,均支持FGFR3作为CRC的治疗靶点[18]。

综上,本研究通过综合生物信息学分析,筛选出MYC和FGFR3 2个在CRC中显著上调且与预后不良相关的DEGs。二者不仅在CRC进展中通过调控关键信号通路发挥重要作用,且在膀胱癌通路中富集[19-21],提示其可能参与跨肿瘤类型的共同调控机制。这一发现为CRC的预后评估提供了潜在生物标志物,同时为开发靶向MYC和FGFR3的CRC治疗策略奠定了理论基础,后续需通过体内外实验进一步验证其功能及作用机制。

[参考文献]

[1] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA Cancer J Clin, 2018, 68(6): 394-424. doi: 10. 3322/caac. 21492.

[2] Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015[J]. CA Cancer J Clin, 2016, 66(2): 115-132. doi: 10. 3322/caac. 21338.

[3] Szasz AM, Lanczky A, Nagy A, et al. Cross validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1, 065 patients[J]. Oncotarget, 2016, 7(31): 49322-49333. doi: 10. 18632/oncotarget. 10337.

[4] Matsuoka T, Yashiro M. Novel biomarkers for early detection of gastric cancer[J]. World J Gastroenterol, 2023, 29(17): 2515-2533. doi: 10. 3748/wjg. v29. i17. 2515.

[5] Shiotani A, Iishi H, Uedo N, et al. Histologic and serum risk markers for noncardia early gastric cancer[J]. Int J Cancer, 2005, 115(3): 463-469. doi: 10. 1002/ijc. 20852.

[6] Duffy MJ, Lamerz R, Haglund C, et al. Tumor markers in colorectal cancer, gastric cancer and gastrointestinal stromal cancers: European group on tumor markers 2014 guidelines update[J]. Int J Cancer, 2014, 134(11): 2513-2522. doi: 10. 1002/ijc. 28384.

[7] Isella C, Terrasi A, Bellomo SE, et al. Stromal contribution to the colorectal cancer transcriptome[J]. Nat Genet, 2015, 47(4): 312-319. doi: 10. 1038/ng. 3224.

[8] Davis S, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor[J]. Bioinformatics, 2007, 23(14): 1846-1847. doi: 10. 1093/bioinformatics/btm254.

[9] Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool fortheunification ofbiology. The Gene Ontology Consortium[J]. Nat Genet, 2000, 25(1): 25-29. doi: 10. 1038/75556.

[10] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes[J]. Nucleic Acids Res, 2000, 28(1): 27-30. doi: 10. 1093/nar/28. 1. 27.

[11] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources[J]. Nat Protoc, 2009, 4(1): 44-57. doi: 10. 1038/nprot. 2008. 211.

[12] Szklarczyk D, Franceschini A, Wyder S, et al. STRINGv10: protein-protein interaction networks, integrated over the tree of life[J]. Nucleic Acids Res, 2015, 43(Database issue): D447-D452. doi: 10. 1093/nar/gku1003.

[13] Matsuoka T, Yashiro M. Bioinformatics analysis and validation of potential markers associated with prediction and prognosis of gastric cancer[J]. Int J Mol Sci, 2024, 25(11): 5880-5903. doi: 10. 3390/ijms25115880.

[14] Tang Z, Li C, Kang B, et al. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses[J]. Nucleic Acids Res, 2017, 45(1): 98-102. doi: 10. 1093/nar/gkx247.

[15] Guo Y, Ye Q, Deng P, et al. Spermine synthase and MYC cooperate to maintain colorectal cancer cell survival by repressing Bim expression[J]. Nat Commun, 2020, 11(1): 3243. doi: 10. 1038/s41467-020-17067-x.

[16] Bian X, Wang X, Zhang Q, et al. The MYC paralog-PARP1 axis as a potential therapeutic target in MYC paralog-activated small cell lung cancer[J]. Front Oncol, 2020, 10: 565820. doi: 10. 3389/fonc. 2020. 565820.

[17] Zhao R, Bu YC, Wei L, et al. Acetylshikonin suppressed growth of colorectal tumour tissue and cells by inhibiting the intracellular kinase, T-lymphokine-activated killer cell-originated protein kinase[J]. Br J Pharmacol, 2020, 77(10): 2303-2319. doi: 10. 1111/bph. 14981.

[18] Wang X, Yin Z, Zhao Y, et al. Identifying potential prognostic biomarkers in head and neck cancer based on the analysis of microRNA expression profiles in TCGA database[J]. Mol Med Rep, 2020, 21(3): 1647-1657. doi: 10. 3892/mmr. 2020.

[19] 钟 键,金正贤,卞卫星,等. 膀胱癌组织FGFR3基因突变和蛋白表达与预后关系研究[J]. 中华肿瘤防治杂志,2020,27(2):130-135. doi:10. 16073/j. cnki. cjcpt. 2020. 02. 07.

[20] Wang Y, Ju L, Wang G, et al. DNA polymerase POLD1 promotes proliferation and metastasis of bladder cancer by stabilizing MYC[J]. Nat Commun, 2023, 14(1): 2421. doi: 10. 1038/s41467-023-38160-x.

[21] 李小航,李安安,朱宝兴,等. 膀胱尿路上皮癌中TP53、FGFR3和PIK3CA基因突变的临床病理意义[J]. 临床与实验病理学杂志,2023,39(6):668-674. doi:10. 13315/j. cnki. cjcep. 2023. 06. 006.

(编辑:刘 艺)