网络出版时间:2022-1-20 14:39 网络出版地址:https://kns.cnki.net/kcms/detail/34.1065.R.20220119.1133.012.html

生物信息学预测橘皮素治疗结直肠癌核心基因

宋子健[1],李建伟[1],胡和智[2]

摘要 **目的** 通过生物信息学分析筛选橘皮素治疗结直肠癌(CRC)的mRNA核心基因并利用生存分析方法验证mRNA核心基因对CRC的预测效果。**方法** 使用DMSO溶剂和橘皮素处理CRC的HCT116细胞48 h后进行RNA测序。将测序结果进行预处理和差异表达分析。从差异表达基因中筛选lncRNA关键基因,建立相应的lncRNA-miRNA-mRNA调控网络,得到mRNA关键基因。通过基因本体论(GO)分析、京都基因组百科全书(KEGG)通路分析和蛋白质互作网络(PPI)分析,得到mRNA核心基因,并利用生存分析方法对mRNA核心基因进行验证。**结果** 差异表达分析后筛选出197个lncRNA差异表达基因、128个miRNA差异表达基因和1 938个mRNA差异表达基因。利用lncRNA-miRNA-mRNA调控网络筛选得到5个lncRNA关键基因和117个mRNA关键基因。关键基因的GO分析和KEGG通路分析结果表明它们富集在与CRC密切相关的功能和通路上,通过PPI网络分析得到6个mRNA核心基因,采用生存分析方法验证发现其中有3个mRNA核心基因(*FOS*、*CCND2*、*MXD1*)与CRC密切相关。**结论** 通过生物信息学方法分析橘皮素治疗CRC的分子机制,筛选出3个差异表达非常显著且对患者预后影响明显的基因,为CRC的诊断、治疗和预后治疗提供了新思路。

关键词 结直肠癌;橘皮素;生物信息学;核心基因
中图分类号 R 574.62
文献标志码 A **文章编号** 1000-1492(2022)02-0229-06
doi:10.19405/j.cnki.issn1000-1492.2022.02.013

作为第三大恶性肿瘤的结直肠癌(colorectal cancer, CRC)具有恶性程度高、病程进展迅速、易复发和转移等特点,对人类健康和生命安全构成重大威胁[1]。目前对于CRC的分子机制和核心基因的不完全了解阻碍了对CRC的各项研究。橘皮素是一类天然黄酮类物质,属于黄酮类化合物,广泛存在于芸香科植物川橘果皮、酸橙果皮和柑橘茎叶中,目前已被证实具有抑制细菌和抗肿瘤等药理作用[2]。因此,该研究利用橘皮素治疗结直肠癌的RNA-seq数据预测核心基因,并利用生存分析对其进行预后分析,为CRC的诊断及治疗药物的研制提供新的作用靶点。

2021-10-20 接收
基金项目:国家自然科学基金(编号:81672113、62072154)
作者单位:[1] 河北工业大学人工智能与数据科学学院,天津 300401
[2] 河北工业大学廊坊分校,廊坊 065099
作者简介:宋子健,男,硕士研究生;
胡和智,男,讲师,责任作者,E-mail:huhezhi@hebut.edu.cn

1 材料与方法

1.1 数据资料收集 使用DMSO溶剂和橘皮素药物处理CRC的HCT116细胞48 h,使用TRIzol提取实验组和溶剂对照组细胞的总RNA,使用逆转录试剂盒将其逆转为cDNA,进行RNA测序。最后,将细胞样本分为3个橘皮素实验组和3个无橘皮素对照组。

1.2 差异基因筛选方法 本研究筛选差异表达使用R语言中的程序包对橘皮素实验组和无橘皮素对照组的RNA-seq数据进行基因差异表达分析,差异基因的筛选标准为$|\log2FC|>1$和$P<0.05$。

1.3 lncRNA关键基因筛选 根据基因种类,从筛选出的差异基因集中分别提取lncRNA差异表达基因集、miRNA差异表达基因集和mRNA差异表达基因集。对lncRNA差异表达基因集采用如下方法筛选得到lncRNA关键基因:①提高筛选标准,以$|\log2FC|>2$和$P<0.05$为新阈值,分别筛选出差异表达更为显著的lncRNA差异表达基因集和miRNA差异表达基因集,分别记为集合A和集合B,通过归并排序算法对两个集合均按照P值由小到大进行排序;②从StarBase数据库[3]中收集与miRNA差异表达基因集B中miRNA存在调控关系的lncRNAs,记为lncRNA基因集C;③将集合A与集合C取交集,得到的lncRNA基因集记为集合D;④利用GEPIA数据库[4]中的临床数据对集合D中的lncRNAs进行生存分析,得到有显著预后价值的lncRNA关键基因集。显著lncRNA能够通过大数据网站(如GEPIA等)预测它们和临床病理参数的关系,以供后续研究使用。

1.4 lncRNA-miRNA-mRNA调控网络构建 首先,利用DIANA网站[5]获取lncRNA-miRNA调控关系数据,然后通过miRDB网站[6]获取miRNA-mR-

NA 调控关系数据。利用 Cytoscape3.7.2 软件[7]构建 lncRNA-miRNA-mRNA 调控网络，将其中的 mRNAs 记为 mRNA 关键基因，以便后续研究。

1.5 基因功能注释和通路富集分析 为了更深层次了解橘皮素在治疗 CRC 中的调控功能，对 mRNA 关键基因集进行基因本体论（gene ontology，GO）分析和京都基因组百科全书（Kyoto encyclopedia of genes and genomes，KEGG）通路分析，从而找到差异基因的分子功能、参与的主要生物过程以及它们所属的代谢通路等。DAVID 网站[8]是一个生物信息数据库，它整合了生物学数据和分析工具，主要用于 GO 富集分析和 KEGG 通路分析。将 mRNA 关键基因集导入到 DAVID 数据库中进行分析，相关阈值设置为 $P<0.05$、Kappa Score = 0.5、Min Level = 5 和 Max Level = 8。

1.6 mRNA 核心基因集筛选 通过 STRING 数据库，收集 mRNA 关键基因集的蛋白质互作网络（PPI）数据，PPI 分数设置为 0.4，并使用 Cytoscape3.7.2 软件构建相应的 PPI 网络[9]。基于 MCODE 算法和 cytoHubba 插件中拓扑分析方法 Degree、MNC 和 EPC 分别对 PPI 网络进行分析。MCODE 算法中的阈值设为：Node Density Cutoff = 0.1、Node Score Cutoff = 0.2、K-Core = 2、Max. Depth = 100；3 个拓扑分析方法的阈值设置为：Degree Score≥3、MNC≥2、EPC≥4.8，最终得到 mRNA 核心基因集。

1.7 统计学处理 通过 GEPIA 数据库对 mRNA 核心基因集进行在线生存分析，验证其有效性。数据集限定为 COAD 和 READ 数据集，时间轴单位设置为月；基因表达差异采用 t 检验，在 CRC 中表达量与预后的关系采用 Log-rank 检验，以 $P<0.05$ 表示差异有统计学意义。核心 mRNA 能够通过大数据网站（如 GEPIA 等）预测它们和临床病理参数的关系，以供后续研究使用。

2 结果

2.1 差异基因筛选 RNA 测序后共有 21 460 条基因数据，根据阈值 $P<0.05$，$|\log2FC|>1$ 对基因数据进行差异表达分析，共得到 2 614 个差异表达基因，上调差异表达基因 1 711 个，下调差异表达基因 903 个。其中，含有 197 个 lncRNA 差异表达基因，128 个 miRNA 差异表达基因，1 938 个 mRNA 差异表达基因。差异表达基因的火山图如图 1 所示，图中红色为上调基因，绿色为下调基因。

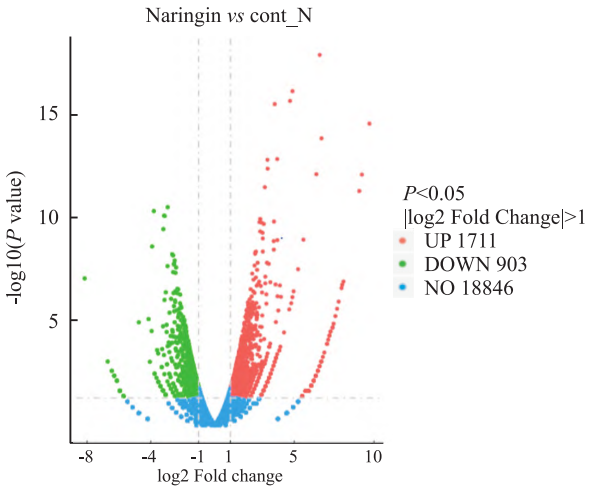

图 1 RNA-seq 数据的差异表达分析火山图

2.2 lncRNA 关键基因集 以阈值 $|\log2FC|>2$ 且 $P<0.05$ 为标准对 197 个 lncRNA 差异表达基因再次进行筛选，得到 116 个 lncRNA 基因，命名为集合 A；以阈值 $|\log2FC|>2$ 且 $P<0.05$ 为标准对 miRNA 差异表达基因集进行筛选，得到 92 个 miRNA 基因，命名为集合 B；从 StarBase 数据库中共收集到 65 个与集合 B 存在调控关系 lncRNA 基因，命名为集合 C；对集合 A 和集合 C 取交集，得到 32 个 lncRNA 基因，命名为集合 D；利用 GEPIA 数据库进行生存分析，共得到 5 个具有显著预后价值的 lncRNA 基因（*MALAT*1、*NEAT*1、*LINC*00342、*LINC*01133、*LINC*00662），记为 lncRNA 关键基因集，生存曲线如图 2 所示。由图 2 可知，5 个 lncRNA 关键基因的 Logrank P 均 <0.05，*MALAT*1（Logrank $P=0.022$）、*NEAT*1（Logrank $P=0.013$）、*LINC*00342（Logrank $P=0.035$）、*LINC*01133（Logrank $P=0.036$）和 *LINC*00662（Logrank $P=0.045$），这证明 5 个 lncRNA 关键基因均对患者的生存周期有显著影响。

2.3 lncRNA-miRNA-mRNA 调控网络 针对 5 个 lncRNA 关键基因，从 DIANA 网站获取 651 条与它们相关的 lncRNA-miRNA 调控关系数据，从 miRDB 数据库获取 419 条与它们相关的 miRNA-mRNA 调控关系数据。基于以上两组调控关系数据，利用 Cytoscape3.7.2 软件的 Merge 功能构建相应的 lncRNA-miRNA-mRNA 调控网络。lncRNA-miRNA-mRNA 网络由 1 057 个节点和 909 条边组成，共有 117 个 mRNA 关键基因。调控网络如图 3 所示。

2.4 mRNA 关键基因的 GO 和 KEGG 分析结果 通过 DAVID 数据库对 117 个 mRNA 关键基因进

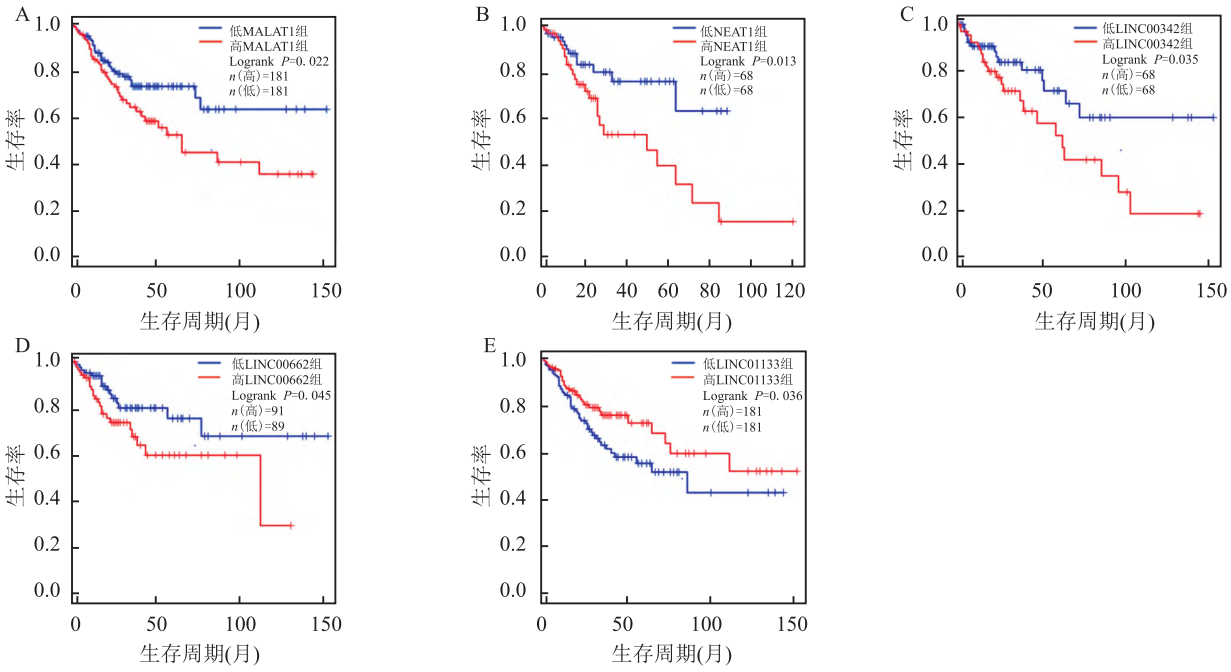

图 2　5 个 lncRNA 关键基因的生存曲线图

A:*MALAT*1 基因生存曲线图;B:*NEAT*1 基因生存曲线图;C:*LINC*00342 基因生存曲线图;D:*LINC*00662 基因生存曲线图;E:*LINC*01133 基因生存曲线图

行 GO 富集分析,在生物过程方面,mRNA 关键基因功能主要集中于 RNA 聚合酶 II 启动子转录的调控和 TOR 信号通路等生物过程,表 1 给出了生物过程中差异性显著的前 5 个条目。KEGG 通路分析结果表明 CRC 的发生发展与 PI3K-Akt 信号通路密切相关。表 2 给出差异性显著的前 5 条 KEGG 通路分析通路。通过 GO 富集分析和 KEGG 通路分析证实 mRNA 关键基因集与 CRC 密切相关。

表 1　差异性显著的前 5 个 GO 条目

GO 号	GO 条目	*P* 值
GO:0006357	RNA 聚合酶 II 启动子转录的调控	1.75×10^{-4}
GO:0006950	响应胁迫	7.30×10^{-4}
GO:0031929	TOR 信号	3.81×10^{-3}
GO:0048015	磷脂酰肌醇介导的信号传导	5.53×10^{-3}
GO:0048661	平滑肌细胞增殖的正调控	7.50×10^{-3}

表 2　差异显著的前 5 个 KEGG 通路分析条目

KEGG 号	KEGG 通路	*P* 值
hsa04151	粘着斑	9.34×10^{-4}
hsa04068	FoxO 信号通路	3.53×10^{-3}
hsa05205	癌症中的蛋白聚糖	4.12×10^{-3}
hsa04151	PI3K-Akt 信号通路	4.52×10^{-3}
hsa04110	细胞周期	1.50×10^{-2}

2.5　mRNA 核心基因集　将 117 个 mRNA 关键基因上传至 STRING 数据库,构建对应的 PPI 网络,将结果导出并保存为 tsv 格式。利用 Cytoscape3.7.2 软件将该 PPI 网络可视化,由 48 个节点和 45 条边组成,如图 4 所示。利用 MCODE 算法和 cytoHubba 插件中 Degree、MNC 及 EPC 拓扑分析方法对 PPI 网络进行分析。MCODE 算法分析结果如图 5 所示,拓扑分析方法前 10 名结果如表 3 所示。利用 Python 语言的 numpy 程序包对 MCODE 算法的结果基因集和拓扑分析方法的结果基因集取交集,得到 6 个 mRNA 核心基因(*FOS*、*GADD45A*、*CCND2*、*MYCN*、*BACH1* 和 *MXD1*)。这些核心基因在橘皮素治疗 CRC 中起到重要的调控作用,有成为生物标志物和药物靶点的潜力。

表 3　拓扑分析方法前 10 名结果

基因	MNC	Degree	EPC
FOS	3	6	5.545
CCND2	3	3	5.148
HSP90B1	2	2	4.812
CHD1	2	4	4.742
BACH1	2	2	4.724
MIA3	2	3	4.844
BTAF1	2	2	4.655
MYCN	2	3	4.544
MXD1	2	4	5.9
GADD45A	2	2	4.311

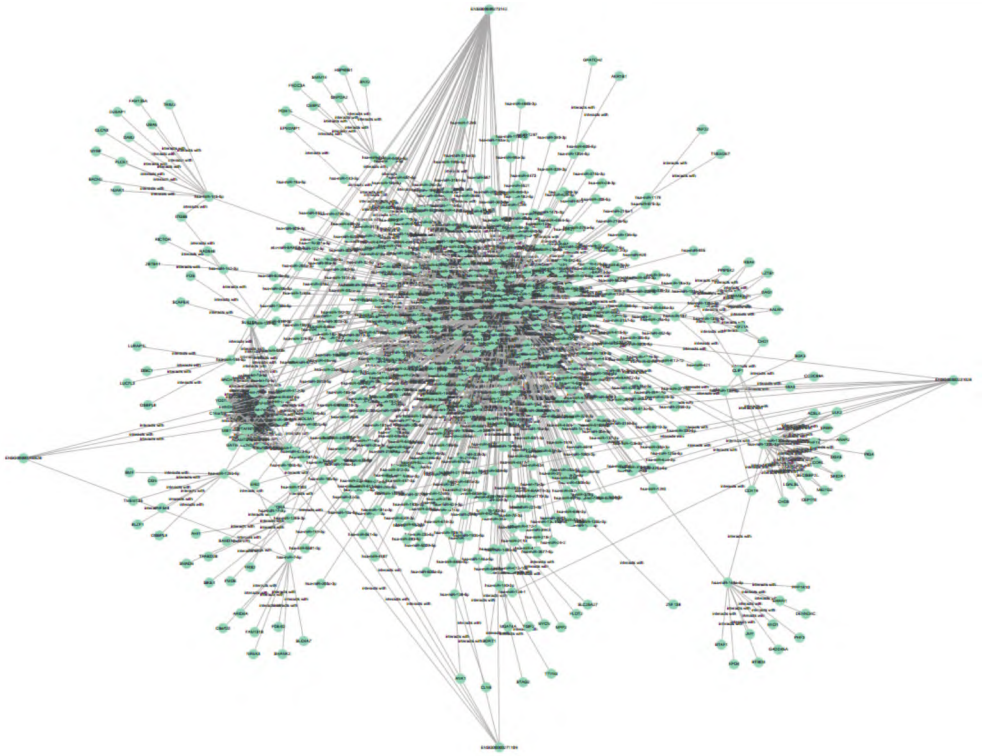

图 3 lncRNA 关键基因对应的 lncRNA-miRNA-mRNA 调控网络

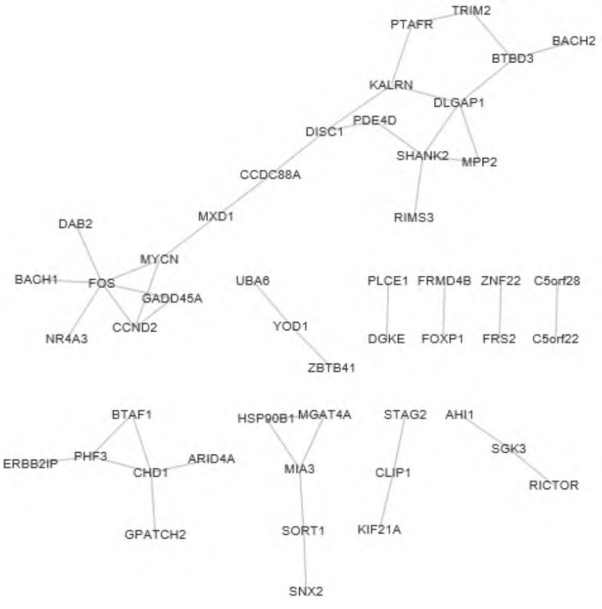

图 4 mRNA 关键基因对应的 PPI 网络

2.6 mRNA 核心基因与患者的预后关系 利用 GEPIA 数据库对 6 个 mRNA 核心基因进行生存分析,各自的生存曲线图如图 6 所示。其中,*FOS*、*CCND2* 和 *MXD1* 表达水平对患者的总生存时间有着显著影响($P < 0.05$)。而 *GADD45A*、*MYCN* 和 *BACH1* 对患者的生存率影响差异无统计学意义。

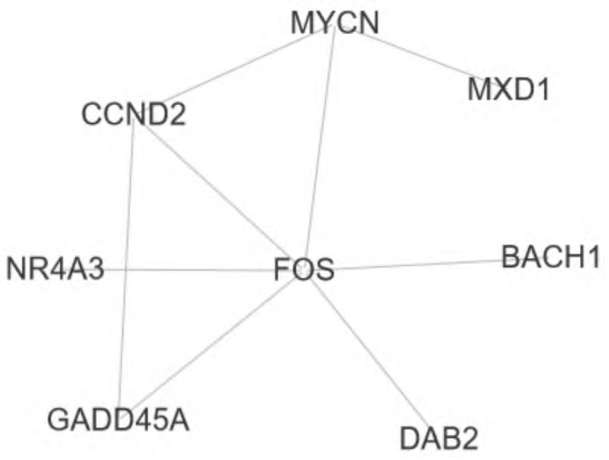

图 5 MCODE 分析结果图

3 讨论

CRC 的发生与外界环境、行为方式、遗传等多种因素密切相关,尽管目前对 CRC 的研究已取得了较大进步,但是 CRC 的预后仍然效果不佳。随着测序技术的飞速发展和生物信息技术的不断突破,CRC 的分子机制研究成为了当前的一个热点,寻找 CRC 诊断及预后的生物标志物和药物靶点为 CRC 的诊疗提供了新的思路。橘皮素已被证实具有抑制

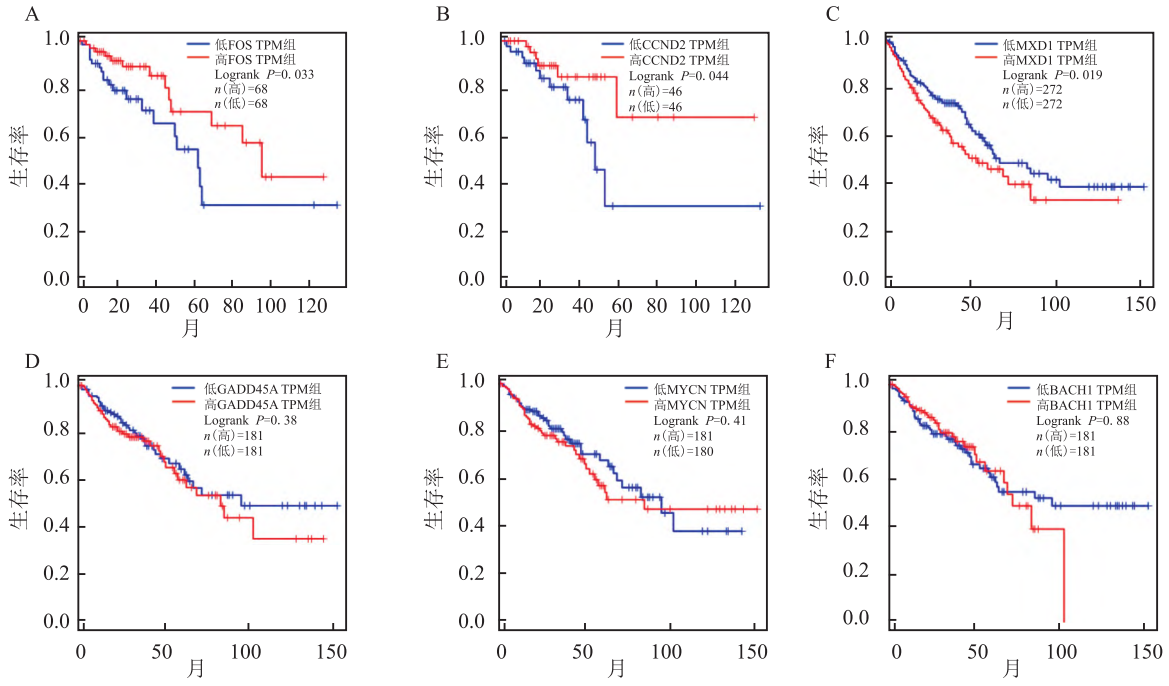

图6 6个mRNA核心基因表达与患者预后的生存曲线

A:*FOS*基因生存曲线图;B:*CCND2*基因生存曲线图;C:*MXD1*基因生存曲线图;D:*GADD45A*基因生存曲线图;E:*MYCN*基因生存曲线图;F:*BACH1*基因生存曲线图

细菌和抗肿瘤等药理作用,但对橘皮素治疗CRC的分子机制却不甚了解。因此,本文以橘皮素治疗CRC的RNA-seq数据为研究对象,通过生物信息学分析方法,筛选出117个mRNA关键基因。GO富集分析和KEGG通路分析结果表明,关键基因均富集在与CRC相关的功能和通路上。其中,KEGG分析结果表明PI3K-Akt信号通路与CRC密切相关。研究[10]表明,SPOCK1在CRC细胞系中过表达,沉默SPOCK1可逆转CRC细胞中的EMT过程,显著减弱了迁移/侵袭,抑制体外增殖和体内肿瘤的生长。敲除SPOCK1明显降低了HCT116细胞中p-PI3K和p-Akt的蛋白表达水平。此外,mRNA关键基因还显著富集在乙型肝炎、肺结核、胰腺癌、甲型流感、前列腺癌等多种疾病的相关信号通路,提示橘皮素对多种疾病具有治疗作用,为今后的相关研究提供了新的思路。

通过生存分析,从mRNA关键基因中筛选出3个与CRC预后密切相关的核心基因(*FOS*、*CCND2*和*MXD1*)。其中,*FOS*和*CCND2*已被文献[11-12]证实与CRC有密切关系。有研究[13]表明,*MXD1*参与了乳腺癌癌细胞的增殖和转移过程,在乳腺癌组织中*MXD1*表达显著下调,并影响了乳腺癌患者的预后。因此,课题组推断*MXD1*也极有可能与CRC的发生和发展相关。综上所述,3个核心基因有成为生物标志物和药物靶点的可能,对CRC的发病机制及治疗提供了新的思路,也为CRC药物靶点研究提供重要参考。

参考文献

[1] Cancho V G, Bazán J L, Dey D K. A new class of regression model for a bounded response with application in the study of the incidence rate of colorectal cancer[J]. Stat Methods Med Res, 2020, 29(7): 2015-33.

[2] Dey D K, Chang S N, Vadlamudi Y, et al. Synergistic therapy with tangeretin and 5-fluorouracil accelerates the ROS/JNK mediated apoptotic pathway in human colorectal cancer cell[J]. Food Chem Toxicol, 2020, 143:111529.

[3] Li J H, Liu S, Hui Z, et al. StarBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data[J]. Nucleic Acids Res, 2014, 42(Database issue): D92-7.

[4] 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(W1): W98-W102.

[5] Karagkouni D, Paraskevopoulou M D, Chatzopoulos S, et al. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions[J]. Nucleic Acids Res, 2018, 46(D1): D239-45.

[6] Chen Y, Wang X. MiRDB: an online database for prediction of

functional microRNA targets[J]. Nucleic Acids Res, 2020, 48(D1): D127 - 31.

[7] Smoot M E, Ono K, Ruscheinski J, et al. Cytoscape 2.8: new features for data integration and network visualization[J]. Bioinformatics, 2011, 27(3): 431 - 2.

[8] Huang da W, Sherman B T, Lempicki R A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources[J]. Nat Protoc, 2009, 4(1): 44 - 57.

[9] 汪圣毅, 张永红, 闫亚飞, 等. 胃癌化疗抵抗基因 PMP22 下游的生物信息学机制[J]. 安徽医科大学学报, 2019, 54(4): 509 - 14.

[10] Wang N, Yang L, Dai J, et al. 5-FU inhibits migration and invasion of CRC cells through PI3K/AKT pathway regulated by MARCH1[J]. Cell Biol Int, 2020, 45(2): 368 - 81.

[11] 屈 潇. C-Myb 靶向 c-fos 调控结直肠癌生长和转移的机制研究[D]. 合肥: 安徽医科大学, 2019.

[12] Li W C, Wu Y Q, Gao B, et al. MiRNA-574-3p inhibits cell progression by directly targeting CCND2 in CRC[J]. Biosci Rep, 2019, 39(12): BSR20190976.

[13] Zhang X, Zhao H, Zhang Y, et al. The microRNA-382-5p/MXD1 axis relates to breast cancer progression and promotes cell malignant phenotypes[J]. J Surg Res, 2020, 246: 442 - 9.

Prediction of core genes in the treatment of colorectal cancer with naringin using bioinformatics

Song Zijian[1], Li Jianwei[1], Hu Hezhi[2]

([1]*School of Artificial Intelligence, Hebei University of Technology, Tianjin* 300401;
[2]*Langfang Branch, Hebei University of Technology, Langfang* 065099)

Abstract ***Objective*** To screen the mRNA core genes of naringin in the treatment of colorectal cancer (CRC) by bioinformatics analysis, and to verify the predictive effect of mRNA core genes on CRC by survival analysis. ***Methods*** The HCT116 cells of CRC were treated with DMSO solvent and naringin for 48 h and then RNA sequencing was conducted. The sequencing results were preprocessed and their differentially expressed genes were analyzed. The key lncRNAs were screened from differentially expressed genes, and the corresponding lncRNA-miRNA-mRNA regulatory network was established. With gene ontology (GO) analysis, Kyoto encyclopedia of genes and genomes (KEGG) pathway analysis and Protein-Protein Interaction (PPI) network analysis, the core mRNAs were obtained and verified by survival analysis method. ***Results*** Ultimately, 197 differentially expressed lncRNAs, 128 differentially expressed miRNAs and 1 938 differentially expressed mRNAs were screened. Based on lncRNA-miRNA-mRNA regulatory network, 5 key lncRNAs and 117 key mRNAs were screened. The results of GO analysis and KEGG pathway analysis showed that they were mainly enriched in the functions and pathways closely related to CRC. In the end, 6 mRNA core genes were obtained by PPI network analysis, and 3 core mRNAs (*FOS*, *CCND2*, *MXD1*) were gained by survival analysis, which closely resembled CRC. ***Conclusion*** The molecular mechanism of naringin in the treatment of CRC is analyzed by means of bioinformatics, 3 core mRNAs with significant differences are screened out and they all have an important impact on the prognoses of patients, and the study will provide new ideas for the diagnosis, treatment and prognosis of CRC.

Key words colorectal cancer; naringin; bioinformatics; core genes