

肝癌相关差异表达基因的生物信息学分析及蛋白互作网络构建

曹丹, 余乐, 王丽春[△]

四川大学华西医院 感染性疾病中心(成都 610041)

【摘要】 目的 建立生物信息学分析筛选肝癌基因芯片差异表达基因的方法,助力肝癌发病分子机制的研究。**方法** 通过 R 语言软件分析肝癌芯片数据 GSE45436 中肝癌组织和癌旁组织差异表达的基因, DAVID 软件对差异表达基因进行 GO 功能富集及 KEGG 通路富集分析, String 和 Cytoscape 软件分析关键基因和模块。**结果** 共筛选出 375 个差异表达明显的基因,其中表达上调和下调的分别有 99 个和 296 个。差异基因功能分析主要涉及细胞周期、p53 信号通路、补体途径、细胞色素 P450 代谢通路等。蛋白质-蛋白质相互作用(protein-protein interaction, PPI)网络显示拓扑异构酶 II α (TOP2A) 可能与肝癌的发生发展最为相关。**结论** 本研究采用基因芯片结合生物信息学方法,构建了肝癌差异表达基因编码的蛋白互作网络。TOP2A 可能是肝癌相关的核心基因。

【关键词】 生物信息学分析 肝癌 差异基因表达 功能富集

Bioinformatics Analysis of Hepatocellular Carcinoma Related Gene and Construction of a Protein-Protein Interaction Network CAO Dan, YU Le, WANG Li-chun[△]. Center of Infectious Diseases, West China Hospital, Sichuan University, Chengdu 610041, China

[△] Corresponding author, E-mail: mindywang0218@163.com

【Abstract】 Objective To identify key genes associated with hepatocellular carcinoma (HCC) through analyzing the functions of differentially expressed genes (DEGs) and the interactions of their encoded proteins. **Methods** The microarray dataset GSE45436 was downloaded from the Gene Expression Omnibus (GEO) database. The DEGs in hepatocellular carcinoma and adjacent tissues were analyzed using the R software. Bioinformatics tools DAVID, STRING, GEPIA, Cytoscape, cBioPortal were applied to analyze the biological functions of the DEGs and their encoded protein interactions. **Results** A total of 375 DEGs were identified, consisting of 296 downregulated genes and 99 upregulated genes. The enriched functions and pathways of the DEGs included cell cycle, p53 signaling pathway, complement activation, and metabolism of xenobiotics by cytochrome P450. The PPI analysis showed that TOP2A might be involved in the carcinogenesis of HCC. **Conclusions** Differentially expressed genes in hepatocellular carcinoma and adjacent tissues and their encoded protein interactions revealed by the bioinformatics analysis provide guidance for further research on the molecular mechanism and targeted therapy of HCC. TOP2A may play a key role in HCC.

【Key words】 Bioinformatics analysis Hepatocellular carcinoma Differentially expressed genes Enriched functions and pathways

肝细胞肝癌已成为全球最常见的恶性肿瘤之一,其死亡率位列于癌症相关肿瘤死亡率第三位,严重威胁着人类健康^[1]。肝癌的病因及确切分子机制尚不完全清楚,目前认为其发病是复杂的生物学过程,受遗传因素和环境因素双重影响。据流行病学及相关研究表明,肝炎病毒感染、酒精、黄曲霉素、亚硝胺类物质等都与肝癌发病有关。目前越来越多的研究表明,基因的异常表达与突变参与了肝癌的发生和发展,包括细胞周期蛋白 D1 (CCND1),表皮生长因子受体 (EGFR)、c-myc、ras 等,CCND1 基因 G870A 多态性可能增加中国人群乙肝相关肝癌的风险,EGFR 的慢性刺激在肝癌的发生和发展中起关键作用,c-myc mRNA 和蛋白表达随肝癌的恶性

程度而增加,ras 通路的异常激活可能在肝癌中也有重要作用^[2-5]。然而,由于早期缺乏有效的诊断方法,肝癌的死亡率仍然很高。因此,了解肝癌发生、增殖和复发的确切分子机制,对明确肝癌的治疗策略至关重要。

在过去的几十年中,基因芯片技术和生物信息学分析已广泛用于基因组水平遗传变异的筛查。本研究筛选出肝癌芯片数据 GSE45436 中肝癌组织和癌旁组织差异表达的基因,采用生物信息学方法,分析与肝癌相关基因的功能,同时构建蛋白质-蛋白质相互作用(PPI)网络,旨在为肝癌发生、发展的分子机制的进一步研究提供理论基础。

1 材料和方法

1.1 芯片来源

[△] 通信作者, E-mail: mindywang0218@163.com

肝癌芯片数据 GSE45436 来源于 NCBI(美国国立生物技术信息中心)的 GEO 数据库(<http://www.ncbi.nlm.nih.gov/geo>)。芯片为全基因组的 RNA 表达芯片。该芯片数据包含 95 例肝癌组织和 39 例癌旁组织样本。

1.2 数据下载及处理

下载肝癌组织芯片 GSE45436 的 CEL 数据压缩包和探针文件后,通过 R 语言软件包对原始数据进行标准化及表达值计算,再以 $|\lg FC| > 2$ 、 $P < 0.01$ 的标准筛选出显著差异表达基因进行后续分析。

1.3 GO 分析和 KEGG 通路富集分析

DAVID(<http://david.ncifcrf.gov>)是一个联合了生物学数据和分析工具的在线生物信息数据库,能够为基因和蛋白列表提供注释信息,从而使用户得到相应的生物学信息。KEGG 是用于理解高通量实验技术所产生的大规模分子数据集的高级功能和生物系统的数据库。GO 是一种生物信息学工具,用于注释基因并分析这些基因的生物过程。我们将差异基因导入至 DAVID 数据库中,以 $P < 0.05$ 作为入选标准,对差异基因进行 GO 功能富集及 KEGG 通路富集分析。

1.4 PPI 的构建

String 数据库(<https://string-db.org/>)是一个了解蛋白质相互作用的数据库。我们将差异基因导入 String 数据库后得到的 PPI 网络图结果再导入至 Cytoscape 软件中进行可视化,经 MCODE 插件筛选功能模块进行分析,最后再次使用 DAVID 数据库对功能模块中所包含的基因进行 KEGG pathway 分析。

1.5 核心基因的筛选和分析

根据节点度选择核心基因,利用 GEPIA(<http://gepia.cancer-pku.cn/index.html>)数据库分析核心基因在肝癌组织和正常组织的表达水平,同样利用 GEPIA 绘制核心基因的 Kaplan-Meier 生存曲线。

2 结果

2.1 肝癌与癌旁组织的差异表达基因

通过对基因芯片 GSE45436 进行数据分析,结果(图 1)显示有 375 个表达差异明显的基因,表达上调和下调的分别为 99 个和 276 个。

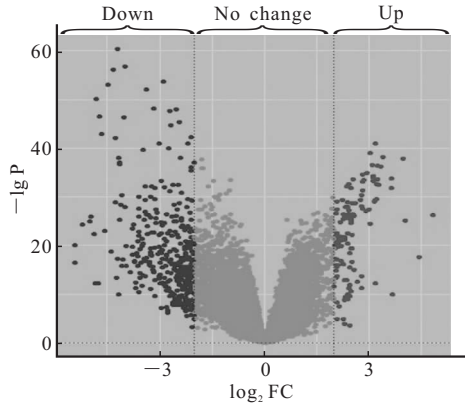


图 1 差异表达基因火山图
Fig 1 Volcano plot of DEGs

2.2 差异表达基因的 GO 功能富集分析和 KEGG pathway 分析

使用 DAVID 数据库分析差异表达基因的分子功能、生物学过程和生物学组成。GO 分析结果(表 1)表明,差异基因的生物过程的变化在氧化还原过程、细胞分裂、细胞增殖、细胞周期调控、补体活化、

表 1 肝癌相关差异表达基因的 GO 功能分析结果
Table 1 GO function analysis of DEGs related to HCC

Term	Count	P	Term	Count	P
Molecularfunction			GO:0008284-positive regulation of cell proliferation	17	0.04
GO:0005509-calcium binding	27	0.006	GO:0008283-cell proliferation	15	0.02
GO:0005506-iron binding	26	0.007	GO:0006956-complement activation	15	3.83×10^{-9}
GO:0019901-protein kinase binding	18	0.002	GO:0051726-regulation of cell cycle	11	3.1×10^{-4}
GO:0004497-monooxygenase activity	17	2.52×10^{-14}	GO:0000086-G2/M transition of mitotic cell cycle	11	6.85×10^{-4}
GO:0019825-oxygen binding	15	3.3×10^{-13}	GO:0006260-DNA replication	9	0.02
GO:0003824-catalytic activity	13	5.14×10^{-4}	Cellular component		
GO:0008395-steroid hydroxylase activity	10	1.32×10^{-9}	GO:0070062-extracellular exosome	110	1.01×10^{-12}
GO:0008392-arachidonic acid epoxigenase activity	10	2.59×10^{-12}	GO:0005576-extracellular region	79	9.11×10^{-14}
GO:0070330-aromatase activity	8	8.92×10^{-7}	GO:0005615-extracellular space	67	6.69×10^{-12}
GO:0005518-collagen binding	5	0.03	GO:0031090-organelle membrane	22	2.34×10^{-17}
Biological processes			GO:0031012-extracellular matrix	13	0.02
GO:0055114-oxidation-reduction process	49	4.87×10^{-15}	GO:0030496-midbody	12	5.5×10^{-5}
GO:0051301-cell division	32	1.81×10^{-11}	GO:0000776-kinetochore	10	3.3×10^{-5}
GO:0007067-mitotic	25	6.29×10^{-10}	GO:0005579-membrane attack complex	5	5.17×10^{-6}
GO:0006955-immune response	21	6.89×10^{-4}			

免疫反应等;分子功能的变化主要集中在蛋白激酶、胶原蛋白、钙铁结合、酶的活性调控;生物学组成分析显示这些基因大多参与细胞外分泌体、细胞外基质、细胞外区等的组成。KEGG pathway 分析(表 2)表明,差异基因主要涉及细胞周期、p53 信号通路、营养物质代谢、补体途径、细胞色素 P450 代谢通路、胆汁酸的合成和胆汁分泌、化学物质致癌等。

2.3 关键基因分析

见表 3、图 2、图 3。根据 PPI 网络图的结果,我们依据蛋白相互作用关系节点度的高低,筛选出 15 个基因进一步分析。这些基因与其他基因存在较为紧密的相互作用关系,相互作用节点度均大于 60,

表 3 节点度最高的 15 个关键基因

Table 3 TOP 15 hub genes with higher degree of connectivity

Gene	Degree	Primary function
TOP2A	87	TOP2A alters the topologic states of DNA and acts as a target for several anticancer agents and mutations of this gene have been associated with drug resistance
CCNB1	69	CCNB1 can regulate the cell cycle progression
NDC80	67	NDC80 encodes a component of the NDC80 kinetochore complex that organize and stabilize microtubule-kinetochore interactions
CDK1	67	CDK1 can regulate the cell cycle progression, apoptosis and carcinogenesis of tumor cells
CCNB2	64	CCNB2 is associated with tumor invasion, metastasis and poor prognosis
CDKN3	64	CDKN3 encodes a cyclin-dependent kinase inhibitor, and dephosphorylate CDK2 kinase, thus prevent the activation of CDK2 kinase
CCNA2	63	CCNA2 can regulate the cell cycle progression
CDC20	63	CDC20 can regulate the cell cycle progression. High expression of CDC20 is associated with development and progression of HCC
AURKA	63	AURKA encodes a cell cycle-regulated kinase that appears to be involved in microtubule formation and/or stabilization at the spindle pole during chromosome segregation
BUB1	62	BUB1 encodes a serine/threonine-protein kinase that play a central role in mitosis
BUB1B	62	BUB1B encodes a kinase involved in spindle checkpoint function and then ensure the proper separation of chromosomes
BIRC5	62	BIRC5 may prevent apoptotic cell death and is highly expressed in most tumors
TTK	62	TTKC can regulate cell proliferation
ZWINT	62	ZWINT encodes different spliced transcript variant.
KIF11	61	KIF11 encodes a motor protein that regulate chromosome positioning, centrosome separation and establishing a bipolar spindle during cell mitosis

或许与肝癌的发病密切相关。这些基因的名称、节点度和主要功能如表 3 所示。在这些基因之中,拓扑异构酶 II α (TOP2A)的节点度最高,值为 87,提示它可能在肝癌的发生发展中起重要作用。利用 GEPIA 数据库分析 TOP2A 在肝癌组织(369 例)和正常组织(160 例)的表达水平,结果如图 2 所示,以正常组织中 TOP2A 的表达水平为参照,TOP2A 表达水平在肝癌组织中增高,差异有统计学意义($P < 0.05$)。最后我们利用 GEPIA 数据库绘制 182 例 TOP2A 高表达肝癌组织与 182 例 TOP2A 低表达肝癌组织的 Kaplan-Meier 生存曲线,结果显示 TOP2A 高表达的肝癌患者的总生存率和无病生存率低于低表达的患者,差异有统计学意义($P < 0.05$,图 3)。

表 2 肝癌相关差异表达基因的 KEGG pathway 分析
Table 2 KEGG pathway analysis of DEGs related to HCC

Term	Count	P
hsa01100;Metabolic pathways	76	1.33×10^{-10}
hsa00830;Retinol metabolism	19	2.83×10^{-13}
hsa05204;Chemical carcinogenesis	19	1.36×10^{-11}
hsa04610;Complement and coagulation cascades	17	1.2×10^{-10}
hsa00980;Metabolism of xenobiotics by cytochrome P450	17	3.68×10^{-10}
hsa04110;Cell cycle	16	5.18×10^{-11}
hsa04976;Bile secretion	12	5.56×10^{-6}
hsa00140;Steroid hormone biosynthesis	11	7.26×10^{-6}
hsa04115;p53 signaling pathway	11	2.73×10^{-5}
hsa00350;Tyrosine metabolism	9	6.88×10^{-6}
hsa00071;Fatty acid metabolism	9	4.11×10^{-5}
hsa00010;Glycolysis/gluconeogenesis	8	0.004

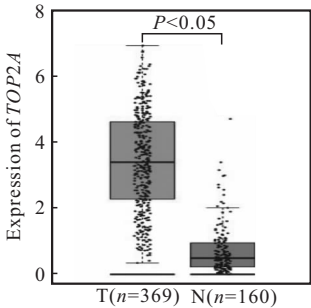


图 2 TOP2A 在肝癌肿瘤组织及正常组织中的表达水平
Fig 2 Expression of TOP2A in tumor (T) and normal tissues (N) of hepatocellular carcinoma

2.4 PPI 功能模块分析

PPI 网络通过在线网站 String 分析得到,随后我们根据 Cytoscape 软件中 MCODE 插件对 PPI 网

络进行聚类分析后得到不同的 PPI 功能模块(图 4),再通过 DAVID 在线分析对功能模块中所包含的基因进行 KEGG pathway 分析,结果显示(表 4 和表 5)这两个模块主要与细胞周期、p53 信号通路、细胞色素 P450 代谢通路有关。

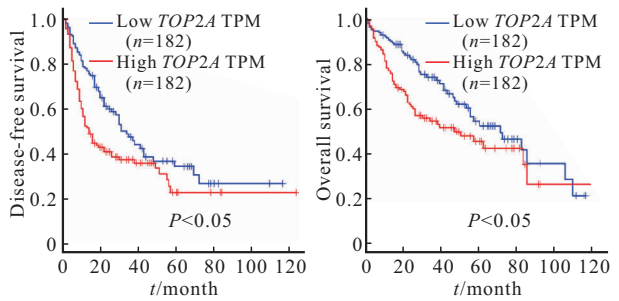
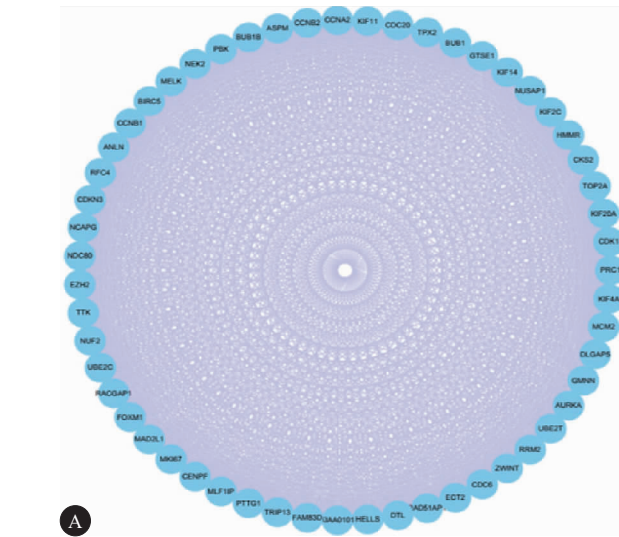


图 3 TOP2A 对肝癌患者预后的评估

Fig 3 Prognosis of patients with hepatocellular carcinoma predicted by TOP2A



3 讨论

肝癌是世界第五大常见恶性肿瘤,近年来死亡率呈上升趋势。然而,肝癌的分子机制仍不清楚。目前高通量基因芯片等现代生物技术以及生物信息学应用能够从分子水平更加深入的探索疾病的发生发展。本文采用生物信息学方法对 GEO 数据库下载的肝癌和癌旁组织的生物芯片数据进行数据分析,包括 GO 功能富集、KEGG pathway 分析、PPI 网络分析、核心基因的筛选及其 Kaplan-Meier 生存曲线绘制。

在筛选出的 375 个表达差异明显的基因中,表达上调有 99 个,表达下调有 276 个,通过 GO 分析,这些差异基因主要参与氧化还原过程、补体活化、细胞分裂、细胞增殖、细胞周期调控等生物学过程,而

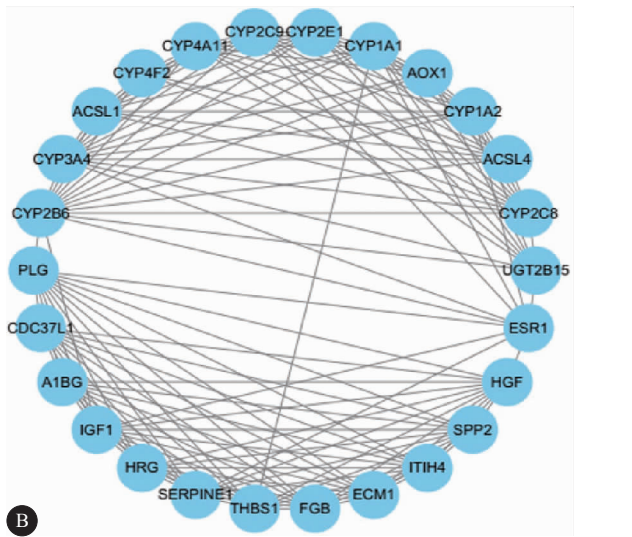


图 4 功能模块图:模块 1(A)和模块 2(B)

Fig 4 Function module diagram: module 1 (A) and module 2 (B)

表 4 模块 1 的 KEGG pathway 分析

Term	P
Cell cycle	1.21×10^{-14}
Progesterone-mediated oocyte maturation	4.89×10^{-6}
Oocyte meiosis	1.48×10^{-5}
p53 signaling pathway	4.27×10^{-5}

表 5 模块 2 的 KEGG pathway 分析

Term	P
Metabolic pathways	1.39×10^{-5}
Metabolism of xenobiotics by cytochrome P450	2.98×10^{-8}
Chemical carcinogenesis	4.79×10^{-8}
p53 signaling pathway	0.01

细胞周期失控和细胞异常增殖与肿瘤的发生发展密切相关^[6]。最近有研究也提出补体活化也有促进肿瘤的作用^[7]。此外,正常情况下,机体的氧化系统和还原系统处于有机平衡,即机体的氧化还原过程处于动态平衡过程,一旦氧化系统和还原系统失去平衡,细胞氧化还原环境遭受破坏,可表现出代谢异常和功能异常,持续不平衡则可能导致肿瘤的发生^[8]。KEGG pathway 结果表明差异基因大多涉及细胞周期、p53 信号通路、营养物质代谢、补体途径、细胞色素 P450 代谢通路等。因为 TOP2A(值为 87)在 PPI 网络分析中的节点度最高,提示它可能在肝癌的发生发展中起重要作用,因此我们重点分析了

TOP2A 基因。TOP2A 基因编码的 TOP2A 蛋白参与 DNA 的复制转录、损伤修复过程,控制和改变 DNA 的拓扑状态^[9]。据报道,在肝癌组织中 TOP2A 过表达与微血管浸润、缩短患者生存时间、化疗耐药和复发相关^[10]。此外, TOP2A 在肺癌、食管癌和乳腺癌中也有过表达,因此考虑 TOP2A 可作为肿瘤诊断、治疗和预后的生物标志物。此外,细胞分裂周期蛋白 20(CDC20)在肝癌组织中呈上调趋势,研究表明 CDC20 的高表达与肿瘤分化及 TMN 分级相关,同时 CDC20 高表达预示着肝癌预后不良^[11]。杆状病毒 IAP 重复序列(BIRC5)在肝癌细胞系中过表达,抑制 BIRC5 过表达可促进肝癌细胞凋亡^[12]。细胞周期蛋白依赖性激酶 1(CDK1)调节细胞周期进程并与细胞周期蛋白 B(Cyclin B)结合形成 CDK1-Cyclin B 激酶复合体,miR-582-5p 可以通过抑制 CDK1 表达调控肝癌的进展^[13]。

从 PPI 网络图中筛选出的两个功能模块分析表明,肝癌的发生与细胞周期、p53 信号通路、细胞色素 P450 代谢通路相关。细胞周期失控和细胞异常增殖与肿瘤的发生发展密切相关。p53 信号通路中,p53 作为抑癌基因能够维持细胞基因组稳定,在细胞周期和细胞凋亡调控中发挥重要作用,p53 信号通路异常能促进肝癌细胞恶性增殖^[14]。另外据报道,大多数肝癌病理样本中均存在 p53 基因突变和缺失,这可能与肝癌的发生相关^[15]。细胞色素 P450 代谢通路在外源物质的代谢中起着重要作用,其中细胞色素 P450 在化学物质致癌中参与肿瘤发生^[16]。

本研究亦存在不足之处。首先,该芯片肝癌数据均来自于同一医院,为单中心研究,不能代表所有肝癌患者,因此在其他地区和种族的 HCC 患者中有必要进一步验证。其次,该芯片数据并未上传临床资料,因此我们无法探讨 TOP2A 与临床因素(例如临床分期、肝功能、Child 分级、甲胎蛋白、乙肝病毒感染)的关系,后期拟以本院肝癌患者作为研究对象,结合多种临床因素进一步探讨 TOP2A 在肝癌中的具体作用。再者,本研究采用生物信息学方法发现了 TOP2A 等关键基因,以及细胞周期、p53 信号通路、细胞色素 P450 代谢通路可能与肝癌的发生和预后相关,未来需要进行相关实验,以验证本研究的预测结果。

总之,我们通过生物信息学分析确定了差异表达的基因,它们可能与肝癌的发生、发展和预后相关。本研究共筛选出 375 个差异基因和 15 个关键

基因, TOP2A 可能是肝癌相关的核心基因。为了得到更准确的相关性结果,我们需要在后期进行一系列的验证实验,以验证这种预测结果。

参 考 文 献

- [1] EL-SERAG HB, RUDOLPH KL. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology*, 2007, 132(7): 2557-2576.
- [2] HU Z, ZHOU Z, XIONG G, *et al.* Cyclin D1 G870A polymorphism and the risk of hepatocellular carcinoma in a Chinese population. *Tumour Biol*, 2014, 35(6): 5607-5612.
- [3] BERASAIN C, CASTILLO J, PRIETO J, *et al.* New molecular targets for hepatocellular carcinoma: the ErbB1 signaling system. *Liver Int*, 2007, 27(2): 174-185.
- [4] GAN FY, GESELL MS, ALOUSI M, *et al.* Analysis of ODC and c-myc gene expression in hepatocellular carcinoma by in situ hybridization and immunohistochemistry. *J Histochem Cytochem*, 1993, 41(8): 1185-1196.
- [5] NEWELL P, TOFFANIN S, VILLANUEVA A, *et al.* Ras pathway activation in hepatocellular carcinoma and antitumoral effect of combined sorafenib and rapamycin *in vivo*. *J Hepatol*, 2009, 51(4): 725-733.
- [6] KIM I, XU W, REED JC. Cell death and endoplasmic reticulum stress: disease relevance and therapeutic opportunities. *Nat Rev Drug Discov*, 2008, 7(12): 1013-1030.
- [7] 宁 昌, 陈文伟, 黎 燕. 补体系统与肿瘤的关系研究进展. *国际药学研究杂志*, 2014, 41(5): 516-521.
- [8] 杨梦祺, 刘盼盼, 黄 蓬. 肿瘤氧化还原代谢与干预. *中国生化药物杂志*, 2016, 36(9): 16-23.
- [9] WATANUKI A, OHWADA S, FUKUSATO T, *et al.* Prognostic significance of DNA topoisomerase II alpha expression in human hepatocellular carcinoma. *Anticancer Res*, 2002, 22(2B): 1113-1119.
- [10] WONG N, YEO W, WONG WL, *et al.* TOP2A overexpression in hepatocellular carcinoma correlates with early age onset, shorter patients survival and chemoresistance. *Int J Cancer*, 2009, 124(3): 644-652.
- [11] LI J, GAO JZ, DU JL, *et al.* Increased CDC20 expression is associated with development and progression of hepatocellular carcinoma. *Int J Oncol*, 2014, 45(4): 1547-1555.
- [12] CAO L, LI C, SHEN S, *et al.* OCT4 increases BIRC5 and CCND1 expression and promotes cancer progression in hepatocellular carcinoma. *BMC Cancer*, 2013, 13(1): 82 [2013-02-22]. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583731/pdf/1471-2407-13-82.pdf>. doi: 10.1186/1471-2407-13-82.
- [13] ZHANG Y, HUANG W, RAN Y, *et al.* miR-582-5p inhibits proliferation of hepatocellular carcinoma by targeting CDK1 and AKT3. *Tumour Biol*, 2015, 36(11): 8309-8316.
- [14] MENG X, FRANKLIN DA, DONG J, *et al.* MDM2-p53 pathway in hepatocellular carcinoma. *Cancer Res*, 2014, 74(24): 7161-7167.
- [15] 赵相轩, 温 锋, 毛晓楠. 细胞凋亡在原发性肝癌治疗中的作用研究进展. *实用肿瘤学杂志*, 2016, 30(5): 448-452.
- [16] MOLINA-ORTIZ D, CAMACHO-CARRANZA R, GONZ LEZ-ZAMORA JF, *et al.* Differential expression of cytochrome P450 enzymes in normal and tumor tissues from childhood rhabdomyosarcoma. *PLoS One*, 2014, 9(4): e93261 [2014-04-03]. <https://dx.doi.org/10.1371/journal.pone.0093261>.

(2018-06-17 收稿, 2018-10-11 修回)

编辑 吕 熙