

网络出版时间: 2019-7-15 18:06 网络出版地址: <http://kns.cnki.net/kcms/detail/34.1065.R.20190715.0922.014.html>

基因芯片筛选肝细胞性肝癌的差异表达基因和通路

江 勇, 钱叶本, 陈 朋, 朱 良

摘要 目的 通过基因芯片和生物信息学的分析方法, 筛选与肝癌发病相关的差异表达基因和通路, 为进一步研究肝癌发生发展过程中的相关机制提供依据。方法 收集 2 对肝癌组织和对应的癌旁组织, 提取总 RNA, 并与基因芯片行杂交检测, 分析得到差异表达基因。使用 DAVID 数据库, 对差异基因行 GO 功能注释和 KEGG 富集分析; 应用 String 蛋白互作数据库分析差异表达基因的蛋白互作关系, 构建蛋白互作网络并通过 Cytoscape 行可视化分析, 运用网络分析插件 CytoHubba 筛选核心基因, 最后结合 TCGA (The Cancer Genome Atlas) 数据库中肝癌基因表达数据, 初步探究核心基因在肝癌中的表达情况。结果 按设定的标准, 共筛选得到 4 324 个差异表达基因, 上调表达的有 2 552 个, 下调表达的有 1 772 个。GO 分析显示差异基因主要参与血小板衍生生长因子结合、受体抑制剂和拮抗剂活性等相关的分子功能; KEGG 富集分析提示这些基因与 TGF- β 、Rap1、PI3K-Akt 以及趋化因子等信号通路相关。蛋白互作网络分析筛选得到 6 个核心基因, TCGA 数据库验证发现核心基因 DCN、COL1A1 和 COL1A2 在肝癌组织中的表达存在显著差异 ($P < 0.000 1$)。结论 GO 功能注释和 KEGG 富集分析揭示了肝癌潜在的发病机制, 核心基因作为肝癌的研究提供新方向, 有可能成为肝癌诊断和治疗的潜在靶点。

关键词 肝癌; 基因芯片; 差异表达基因; 核心基因; TCGA
中图分类号 R 735.7

文献标志码 A **文章编号** 1000-1492(2019)08-1237-06
doi: 10.19405/j.cnki.issn1000-1492.2019.08.014

肝细胞性肝癌 (hepatocellular carcinoma, HCC) 是原发性肝癌的主要类型, 肝癌发病率居全球肿瘤的第 5 位, 然而仅我国肝癌年发病例数就达到全世界的 55%, 且新发病例数仍在逐年增加^[1]。复发和转移是困扰肝癌患者治疗和预后的主要难点。因此, 深入研究肝癌发生发展过程中的相关机制, 积极探讨其差异表达的基因和通路, 有助于进一步认识肝癌的发病机制, 提升肝癌的治疗水平和改善患者的预后。

2019-03-04 接收

基金项目: 安徽省自然科学基金(编号: 1508085MH173)

作者单位: 安徽医科大学第一附属医院普外科, 合肥 230022

作者简介: 江 勇, 男, 硕士研究生;

钱叶本, 男, 博士, 教授, 主任医师, 硕士生导师, 责任作者,

E-mail: qianyeben@hotmail.com

基因芯片具有高通量、高灵敏度和自动快速等优点, 可平行对比肿瘤组织及其对应的正常组织的基因表达水平, 极大地提高了差异表达基因的筛选效率^[2], 有利于深入认识肿瘤的发病机制, 提升肝癌患者早期诊断、治疗和预后的水平。该研究通过分析肝癌组织及其对应的癌旁组织的基因芯片, 筛选得到差异表达基因; 进一步通过生物信息学的方法, 对筛选得到的差异基因进行聚类和功能富集分析, 同时构建蛋白互作网络, 得到核心基因。最后通过 TCGA 数据库中的肝癌基因表达数据 RNA-se-queing 初步探究核心基因的表达水平, 为肝癌的诊断和治疗提供潜在的生物标记。

1 材料与方法

1.1 病例资料 根据医学伦理要求和组织样本取材标准, 2 对肝癌组织及其对应的癌旁正常肝脏组织来自于安徽医科大学第一附属医院肝胆胰外科, 组织标本取材于肝癌根治性切除术后组织。经过病理诊断后, 确诊为肝细胞性肝癌。患者术前未进行过放疗、化疗或者介入治疗等操作。组织样本放入液氮中保存。该研究已经得到安徽医科大学第一附属医院伦理委员会的批准(批文号: 2018-13-22)。

1.2 基因芯片 本研究使用的基因芯片为 Agilent Human Gene Expression (4 × 44 k), 包括 41 000 条探针, 由伯豪生物技术公司提供, 实际操作流程参照 Agilent 表达谱芯片标准流程: 总 RNA 反转录得到双链 cDNA, 再进一步用 Cyanine-3-CTP (Cy3) 标记得到 cRNA。将标记好的 cRNA 和芯片杂交, 洗脱结束后, 利用 Agilent Scanner G2505C (Agilent Technologies) 扫描得到原始图像。

1.3 方法

1.3.1 组织样本 RNA 提取和纯化 肝癌及其癌旁组织的总 RNA 通过 QIAGEN RNeasy[®] Mini Kit(货号: #74106) 试剂盒提取和纯化, 同时纯化得到的 RNA 通过 NanoDrop ND-2000 (Thermo Scientific) 进行定量检测, RNA 完整性经 Agilent Bioanalyzer 2100 (Agilent Technologies) 检测。

1.3.2 芯片杂交与扫描 总 RNA 通过反转录得到

双链 cDNA, 再进一步用 Cyanine-3-CTP (Cy3) 荧光染料标记双链 cDNA, 得到 cRNA。经 QIAGEN RNeasy® Mini Kit 试剂盒纯化后, 将标记好的 cRNA 置于 60 °C 温水浴 30 min, 随后冰水浴 1 min, 进行片段化处理。随后将芯片与 cRNA 在 65 °C、10 r/min 的条件下, 于杂交炉中滚动杂交 17 h。杂交结束后, 进行芯片洗脱。最后利用 Agilent Scanner G2505C (Agilent Technologies) 扫描得到原始图像, 通过 Feature Extraction 软件 (version 10.7.1.1, Agilent Technologies) 处理原始图像, 得到原始数据。

1.3.3 芯片数据预处理和差异基因分析 利用 Genespring 软件 (version 12.5; Agilent Technologies) 对芯片原始数据进行标准化处理。标准化后的数据进行过滤, 在用于比较的每组样本中, 至少有 1 组完全标记为已被检测出的探针, 进行后续分析。利用基因表达倍数变化值 (Fold change) 进行差异基因筛选, 筛选标准: Fold change 的绝对值 ≥ 3.0 。

1.3.4 GO 功能注释和 KEGG 富集分析 使用 DAVID (version 6.7, <http://david.abcc.ncifcrf.gov/>) 在线分析平台^[9], 对筛选得到的差异基因在基因本体 (gene ontology, GO) 中注释, 并进行 KEGG (kyoto encyclopedia of genes and genomes, KEGG) 功能富集分析, 以判定差异基因主要影响的生物学功能和通路。 $P < 0.05$ 表示差异有统计学意义。

1.3.5 蛋白互作网络的构建和筛选核心蛋白 通过 String 蛋白互作数据库 (the Search Tool for the Retrieval of Interacting Genes, STRING, <http://string-db.org/>) 分析差异表达基因的蛋白互作关系^[10], 并以此来构建蛋白互作网络, 设置阈值条件为综合相关性评分 > 0.4 。将 String 蛋白互作数据库分析得到的结果导入 Cytoscape 软件, 通过网络分析插件 CytoHubba 计算节点的边 (Degree, 即互作连线的数量), 筛选得到网络中心节点 (Hub Node)。中心节点对应的基因即为核心基因, 同时也称作核心蛋白。

1.3.6 核心基因的表达验证 肝癌原始的基因表达数据 RNA-sequencing 下载于 TCGA 数据库, 通过 Perl 处理下载的原始数据, 生成基因矩阵。再将 Ensembl Gene ID 转换为基因名称 (Gene symbol), 得到最终可分析的基因表达数据, 包括 374 例肝癌组织和 50 例正常肝脏组织, 初步探究核心基因在肝癌组织中的表达情况。

1.4 统计学处理 应用 Graph pad 6.02 软件进行统计学处理, 数据均以 $\bar{x} \pm s$ 表示, 组间比较采用 t 检

验。超几何分布应用于差异基因 GO 功能注释和 KEGG 富集的统计学分析。 $P < 0.05$ 为差异有统计学意义。

2 结果

2.1 差异表达基因 差异表达基因以 Fold change ≥ 3.0 为筛选标准, 在肝癌及癌旁组织中共筛选出 4 324 个差异基因, 其中上调表达的有 2 552 个基因, 下调表达的有 1 772 个基因。挑选差异程度最大的前 100 个基因, 热图用于展示这些基因在肝癌组织和正常组织的表达情况, 见图 1, 红色代表高表达; 绿色代表低表达。本研究的基因芯片结果已上传到 gene expression omnibus (GEO) 数据库中 (编号: GSE117361)。

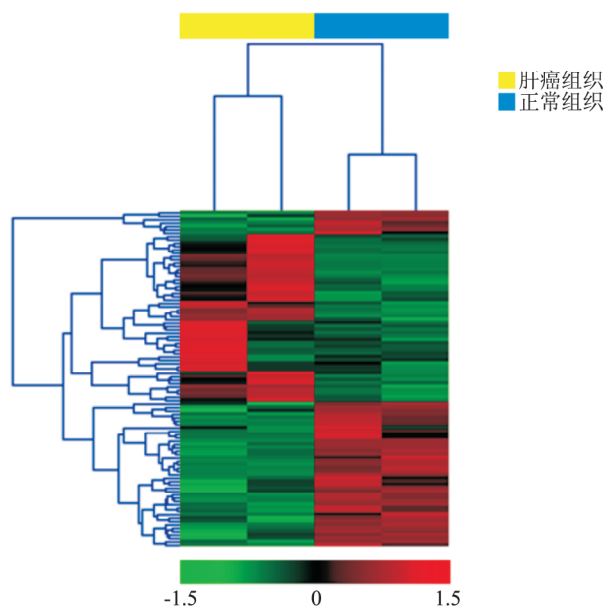


图1 前 100 个差异表达基因的热图

2.2 差异表达基因的 GO 功能注释 GO 功能注释的结果表明, 差异表达基因在分子功能、生物过程及细胞组成的分布是存在差异的。根据显著富集 GO 结果, 发现差异表达基因主要富集在血小板衍生生长因子结合、受体抑制剂和拮抗剂活性等相关的分子功能, 同时也富集于调节细胞对血管内皮生长因子反应等生物过程和胶原蛋白复合物等细胞组成, 见图 2。

2.3 差异表达基因的 KEGG 富集分析 差异基因的 KEGG 信号通路富集分析结果表明, 在显著富集 KEGG 的结果中, 包括 TGF- β 信号通路、Rap1 信号通路、PI3K-Akt 信号通路、细胞黏附、ECM-受体相互作用以及趋化因子信号通路等, 与肝癌发生进展密

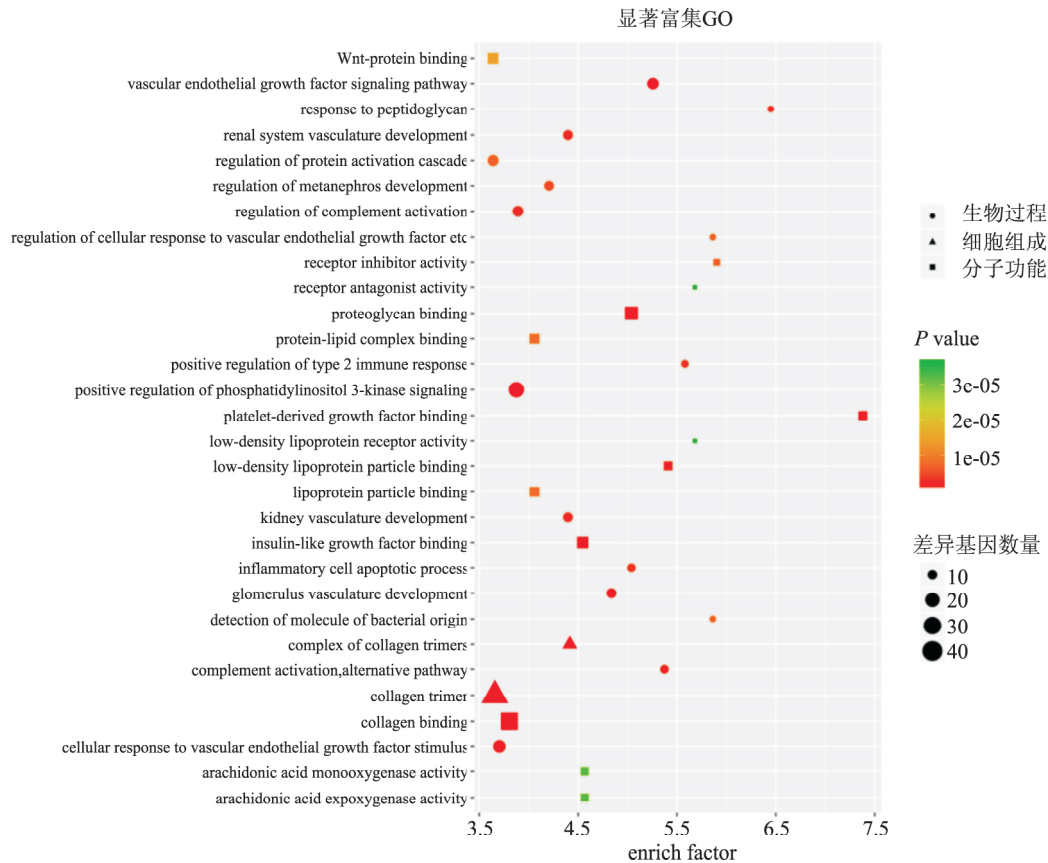


图2 GO功能注释

切相关,这进一步说明在肝癌的发病过程中多条信号通路失控,可能在疾病进展中起到十分重要的作用,见图3。

2.4 差异基因相互作用网络分析 通过 String 蛋白互作数据库分析处理差异基因,构建蛋白互作网络,见图4。根据节点数目排序得到最相关的6个核心蛋白,对应的基因分别为 DCN、TNF、COL1A1、COL3A1、FN1 和 COL1A2。

2.5 核心基因的表达验证 通过 TCGA 数据库中 374 例肝癌组织和 50 例正常肝脏组织的基因表达数据,初步探究 6 个核心基因 DCN、TNF、COL1A1、COL3A1、FN1 和 COL1A2 的表达情况,见图5。结果表明, COL1A1 ($P < 0.0001$) 和 COL1A2 ($P < 0.0001$) 在肝癌组织中的表达较正常组织显著上调, DCN 在肝癌组织中的表达较正常组织显著下调 ($P < 0.0001$); 而 TNF ($P = 0.0522$), FN1 ($P = 0.1030$) 和 COL3A1 ($P = 0.7675$) 的表达无明显差异。

3 讨论

中国是肝癌发病率最高的国家,与乙肝病毒感

染密切相关,其具体的发病机制不明。肝癌患者的确诊往往都在晚期,失去了根治性手术治疗的机会,且肝癌术后易复发,极大地影响了患者的预后。基因芯片技术可探索肿瘤发病过程中多种基因的表达变化及相互之间的调控关系,这有利于理解肝癌的发病机制,进一步提高诊治水平,改善预后。

本研究收集了新鲜的肝癌及癌旁组织,深入分析其基因芯片的结果,发现了 4 324 个差异表达基因,其中上调表达的有 2 552 个基因,下调表达的有 1 772 个基因。这一结果说明了肝癌发病过程中的基因转录变化,其发病机制可能与特定基因和表观遗传的变化密切相关。通过 GO 功能注释和 KEGG 富集分析差异表达基因潜在的生物学作用,结果发现差异表达基因主要富集于血小板衍生生长因子结合等分子功能,血管内皮生长因子反应等生物过程和胶原蛋白复合物等细胞组成; KEGG 富集信号通路显示,差异表达基因主要参与了 TGF- β 、Rap1、PI3K-Akt 和细胞黏附等信号通路,这些信号通路的变化与肝癌发生发展密切相关。TGF- β 信号通路参与了肝癌发生发展的每个阶段,其在肝纤维化的进展中发挥重要作用;同时 TGF- β 在肝癌中表达上

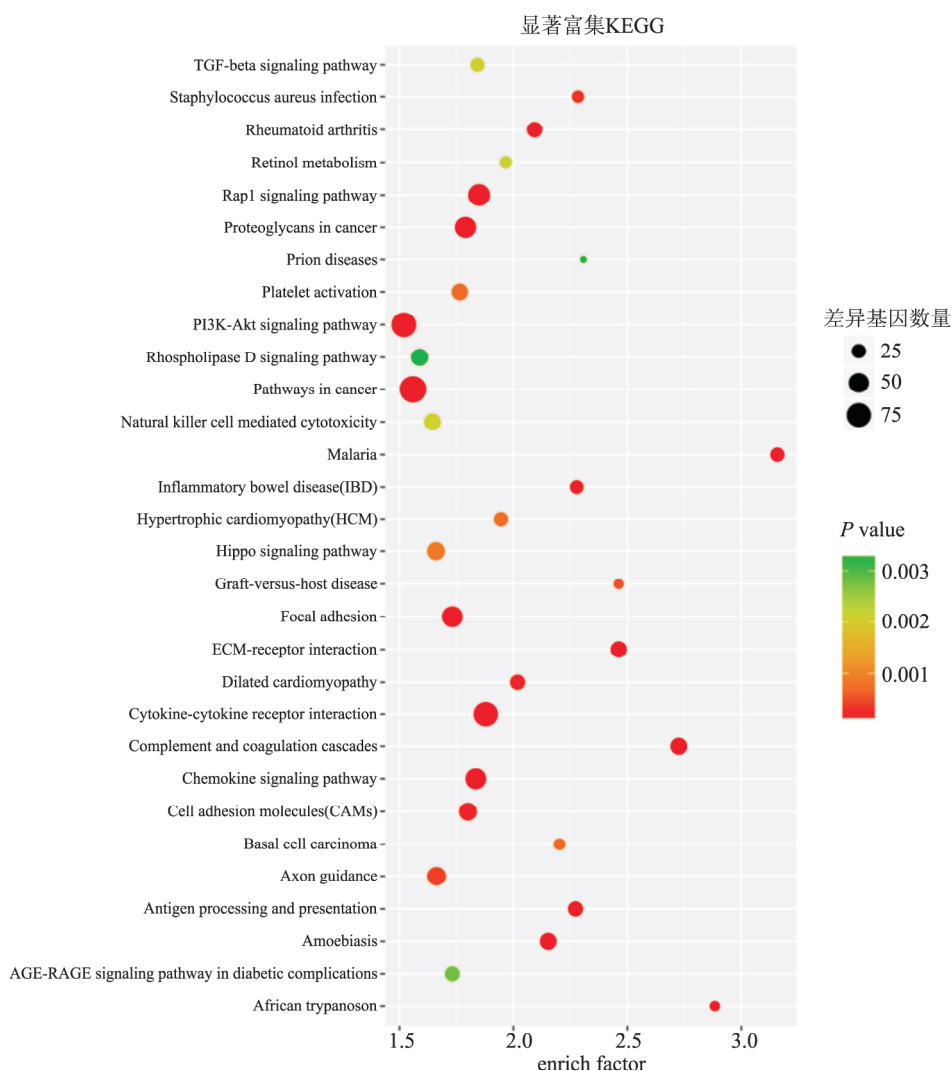


图3 KEGG 通路富集分析

调,与肝癌预后直接相关^[8]。Rap1 是 Ras 家族单体 G 蛋白的成员之一,在肿瘤发生发展中扮演重要的角色,已经得到大家的共识^[9]。研究^[10]表明 Rap1 可调节肿瘤细胞黏附、迁移和侵袭能力;在不同肿瘤类型中,Rap1 信号通路的激活呈现出截然不同的表现,尤其是影响肿瘤转移的作用,依然有待于进一步研究。Rap1 的激活对于 EGFR 介导胰腺癌的转移过程中是必不可少的,同时肿瘤的生长不受影响^[11]。PI3K-Akt 主要通过酪氨酸激酶受体传递生长因子的信号被激活^[12],可调控肿瘤细胞的增殖和凋亡,与肝癌的发病过程密切相关^[13]。由此说明,差异表达基因参与了肝癌发病过程中的多种细胞生物学过程,有待进一步研究其具体的生物学机制。

通过蛋白互作网络分析,本研究得到了可能与肝癌发生密切相关的 6 个核心蛋白,对应的基因分别为 DCN、TNF、COL1A1、COL3A1、FN1 和 COL1A2。

进一步分析 TCGA 数据库中肝癌的基因表达数据,结果表明 DCN、COL1A1 和 COL1A2 在肝癌组织中的表达存在显著差异 ($P < 0.0001$),而 TNF、FN1 和 COL3A1 的表达无明显差异。核心蛋白聚糖 (decorin, DCN) 是存在于细胞外基质的一类富含亮氨酸的小分子蛋白聚糖,研究表明 DCN 可抑制多种肿瘤细胞的增殖,可与多种细胞表面受体结合,介导肿瘤抑制的作用,包括对自噬和炎症的刺激作用以及对血管生成和肿瘤发生的抑制作用。肝硬化是肝癌重要的癌前病变,研究发现 DCN 在肝硬化组织的表达显著高于肝癌及癌旁组织,而 DCN 在肝癌组织中的表达显著低于癌旁组织;同时,DCN 高表达的肝癌患者中位生存期更长,提示 DCN 可作为肝癌的预后标志物^[14]。DCN 可调控胶原纤维的生成,特异性地阻断 TGF- β 信号通路,发挥抗纤维化的生物活性^[15]。COL1A1 和 COL1A2 基因分别编码 I 型胶

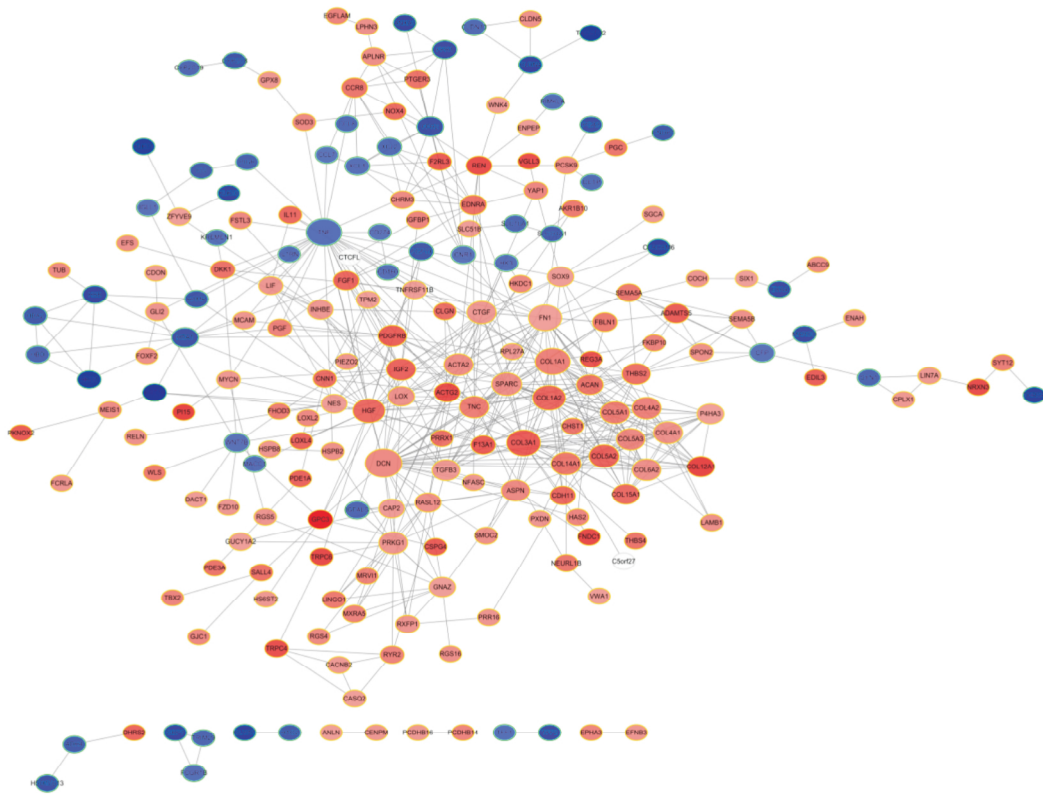


图4 蛋白相互作用网络图

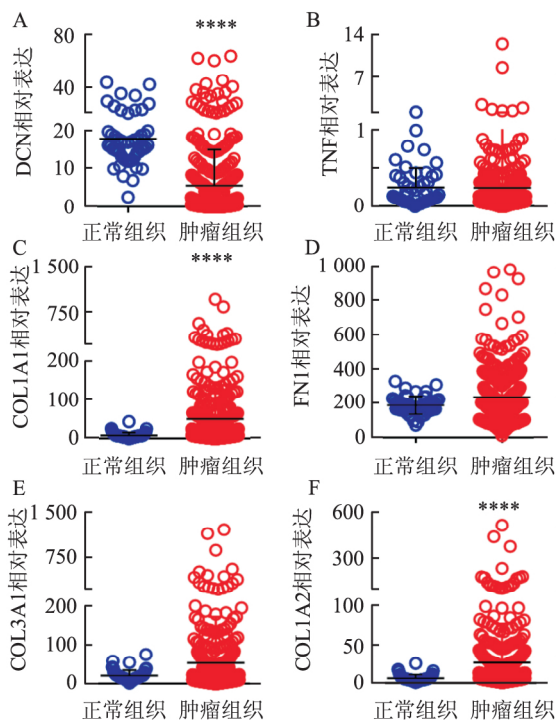


图5 TCGA 数据库初步验证核心基因的表达

A: DCN; B: TNF; C: COL1A1; D: FBN1; E: COL3A1; F: COL1A2; 与

正常组织比较: **** $P < 0.0001$

原蛋白的 pro- $\alpha 1$ 链和 pro- $\alpha 2$ 链,该胶原蛋白的三螺旋结构包括两条 $\alpha 1$ 链和一条 $\alpha 2$ 链,是细胞外基质的主要蛋白。COL1A1 在侵袭性肝癌中表达上调,此外,COL1A1 和 COL1A2 基因促进了 I 型胶原蛋白的合成,有助于 HCC 的侵袭和转移^[3]。COL1A1 的启动子发生甲基化,导致其在肝癌组织中表达下调,该表观遗传改变可能作为 HCC 的预后标志物^[4]。微小 RNA (MicroRNA) let-7g 可抑制肝癌的转移,研究发现其抑制作用部分是通过靶向作用于 COL1A2 实现的^[5]。

综上所述,肝癌发病过程中涉及多基因、多通路的改变,本研究通过基因芯片筛选肝癌的差异表达基因和通路,有助于进一步理解肝癌的发病机制。同时,通过 TCGA 数据库初步验证了核心基因的表达,为肝癌提供了新的研究方向,有待于进一步研究其具体的功能和作用。

参考文献

- [1] GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death,1990-2013: a systematic analysis for the Global Burden of Disease Study 2013 [J]. Lancet,2015, 385(9963):117-71.

- [6] Owzar K, Barry W T, Jung S H, et al. Statistical challenges in pre-processing in microarray experiments in cancer [J]. *Clin Cancer Res*, 2008, 14(19): 5959–66.
- [7] 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.
- [8] Szklarczyk D, Morris J H, Cook H, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible [J]. *Nucleic Acids Res*, 2017, 45(D1): D362–8.
- [9] Lin T H, Shao Y Y, Chan S Y, et al. High serum transforming growth factor- β 1 levels predict outcome in hepatocellular carcinoma patients treated with sorafenib [J]. *Clin Cancer Res*, 2015, 21(16): 3678–84.
- [10] Bailey C L, Kelly P, Casey P J. Activation of Rap1 promotes prostate cancer metastasis [J]. *Cancer Res*, 2009, 69(12): 4962–8.
- [11] Zhang Y L, Wang R C, Cheng K, et al. Roles of Rap1 signaling in tumor cell migration and invasion [J]. *Cancer Biol Med*, 2017, 14(1): 90–9.
- [12] Huang M, Anand S, Murphy E A, et al. EGFR-dependent pancreatic carcinoma cell metastasis through Rap1 activation [J]. *Oncogene*, 2012, 31(22): 2783–93.
- [13] Zhang L, Zhou F, ten Dijke P. Signaling interplay between transforming growth factor- β receptor and PI3K/AKT pathways in cancer [J]. *Trends Biochem Sci*, 2013, 38(12): 612–20.
- [14] Hartmann W, Digon-Sontgerath B, Koch A, et al. Phosphatidylinositol 3-kinase/AKT signaling is activated in medulloblastoma cell proliferation and is associated with reduced expression of PTEN [J]. *Clin Cancer Res*, 2006, 12(10): 3019–27.
- [15] Shang Y K, Li F, Zhang Y, et al. Systems analysis of key genes and pathways in the progression of hepatocellular carcinoma [J]. *Medicine*, 2018, 97(23): e10892.
- [16] Shen Y, Zeglinski M R, Turner C T, et al. Topical small molecule granzyme B inhibitor improves remodeling in a murine model of impaired burn wound healing [J]. *Exp Mol Med*, 2018, 50(5): 68.
- [17] Zekri A N, El-Sisi E R, Abdallah Z F, et al. Gene expression profiling of circulating CD133(+) cells of hepatocellular carcinoma patients associated with HCV infection [J]. *J Egypt Natl Canc Inst*, 2017, 29(1): 19–24.
- [18] Hayashi M, Nomoto S, Hishida M, et al. Identification of the collagen type 1 alpha 1 gene (COL1A1) as a candidate survival-related factor associated with hepatocellular carcinoma [J]. *BMC cancer*, 2014, 14: 108.
- [19] Ji J, Zhao L, Budhu A, et al. Let-7g targets collagen type I alpha2 and inhibits cell migration in hepatocellular carcinoma [J]. *J Hepatol*, 2010, 52(5): 690–7.

Identification of differentially expressed genes and pathways in hepatocellular carcinoma by using microarray

Jiang Yong, Qian Yeben, Chen Peng, et al

(Dept of General Surgery, The First Affiliated Hospital of Anhui Medical University, Hefei 230022)

Abstract *Objective* To further investigate the pathogenesis of hepatocellular carcinoma (HCC), we screened the differentially expressed genes and pathways by using microarray and bioinformatics analysis method. *Methods* The microarray data of 2 pairs HCC tumor tissues and corresponding adjacent normal tissues were and to analyze differentially expressed genes (DEGs). GO functional and KEGG enrichment analysis for DEGs were performed to use DAVID database. The protein-protein interaction (PPI) network was constructed by the String protein interaction database. In addition, the PPI network was visualized by Cytoscape, and the hub genes were screened by using CytoHubba App. Finally, we preliminarily investigated the hub genes expression in HCC by analyzing the RNA-sequencing data deposited in The Cancer Genome Atlas (TCGA). *Results* According to the established standard, we identified 4 324 DEGs including 2 552 up-regulated genes and 1 772 down-regulated genes. GO analysis showed that DEGs were mainly involved in the platelet-derived growth factor binding (PDGF), receptor inhibitor and antagonist activity; KEGG enrichment analysis indicated that these genes were related to TGF- β , Rap1, PI3K-Akt, and chemokine signaling pathway. Six hub genes related to HCC were screened from PPI network. Moreover, there were significant differences in the hub genes expression of DCN, COL1A1, and COL1A2 in HCC tissues confirmed by the analysis of TCGA database ($P < 0.0001$). *Conclusion* GO functional and KEGG enrichment analysis may reveal the potential pathogenesis of HCC. The hub genes provide new clues for the research of HCC, and may become potential targets for the diagnosis and treatment of HCC.

Key words hepatocellular carcinoma; microarray; differentially expressed genes; hub gene; TCGA