DOI:10.3872/j.issn.1007-385x.2020.08.011

·临床研究·

基于GEO数据库筛选胃癌分子标志物

胡鑫¹,高悦¹,田欣涓²,韩明盛¹,李瑞华¹,付西峰³,马艳琴¹(1.山西农业大学生命科学学院,山西 太谷 030801;2.中国药科大学 药学院,江苏 南京 211198;3.山西白求恩医院 普通外科,山西 太原 030032)

[摘 要] **\$\mathbf{n}**\$\$\vec{h}\$

[关键词] 胃癌;分子标志物;GEO数据库;差异表达基因 [中图分类号] R735.2; R730 [文献标识码] A [文章编号] 1007-385X(2020)08-0903-08

Screening of molecular markers in gastric cancer based on GEO database

HU Xin¹, GAO Yue¹, TIAN Xinjuan², HAN Mingsheng¹, LI Ruihua¹, FU Xifeng³, MA Yanqin¹ (1. College of Life Science, Shanxi Agricultural University, Taigu 030801, Shanxi, China, 2. College of Pharmacy, China Pharmaceutical University, Nanjing 211198, Jiangsu, China, 3. Department of General Surgery, Shanxi Bethune Hospital, Taiyuan 030032, Shanxi, China)

[Abstract] Objective: Bioinformatics combined with Gene Expression Omnibus (GEO) was used to screen key genes involved in the development of gastric cancer in order to obtain molecular markers for diagnosis, target selection and prognosis prediction of gastric cancer. Methods: The chip data sets related to gastric cancer (GC) from the GEO database were downloaded, and differentially expressed genes (DEG) were screened. Functional enrichment analysis on DEG was performed, and protein-protein interaction network (PPI) was constructed to screen key genes. Then, co-expression networks were further constructed, and survival curves were drawn and hierarchical clustering analysis was performed. Results: A total of 261 GC-related DEGs were selected, and 14 key genes were obtained through analysis, which were PLOD1, PLOD3, COL1A1, COL1A2, COL2A1, COL3A1, COL4A1, COL4A2, COL8A1, COL12A1, COL15A1, ITGA2, LUM and SERPINH1. Key genes are mainly involved in biological processes such as generation of collagen fiber tissues, extracellular matrix tissues, extracellular structure tissues, skin morphogenesis, collagen biosynthesis and vascular development. Survival curve analysis showed that the change in the expression of ITGA2 (*P*=0.0679) also showed a correlation with the reduction of disease-free survival in gastric cancer patients. Compared with normal gastric tissues, hierarchical cluster analysis showed that the expressions of genes PLOD1, PLOD3, COL3A1, ITGA2, COL1A1, COL1A1, COL4A1, LUM, COL12A1, SERPINH1 and COL8A1 in GC tissues were up-regulated. Conclusion: The key genes obtained after screening can be used as potential molecular markers for early diagnosis, treatment target selection and prognosis judgment of gastric cancer, which provide reference for subsequent research.

[Key words] gastric cancer; molecular marker; Gene Expression Omnibus (GEO) database; differentially expressed gene

[Chin J Cancer Biother, 2020, 27(8): 903-910. DOI: 10.3872/j.issn.1007-385X.2020.08.011]

[通信作者] 马艳琴(MA Yanqin, corresponding author), 女,博士,教授,硕士生导师,主要从事生物化学相关研究, E-mail: mayanqin466@163.com

 $- \oplus$

[[]基金项目] 山西省自然科学基金资助项目(No.201901D111412,201901D111232);山西省研究生教育创新项目资助(No.2020SY197)。Project supported by the Natural Science Foundation of Shanxi Province (No. 201901D111412, 201901D111232), and the Graduate Education Innovation Project of Shanxi Province (No.2020SY197)

[[]作者简介] 胡鑫(1995-),女,硕士生,主要从事生物信息学研究,E-mail:1499640944@qq.com

· 904 ·

胃癌(gastric cancer,GC)是消化道最常见的恶性 肿瘤之一,其发病率和病死率在所有癌症中均位列 第二,且中晚期GC患者所占比例较高,有不断上升 的趋势,疗效极差凹。预计2020年中国GC发病率为 24.30/10万^[2],可见我国GC防控局势严峻。目前,GC 的诊断主要通过内镜检查、影像学检查并结合肿瘤 标志物筛查。GC治疗方式主要有外科手术、内镜、 化学药物、基因及综合治疗等。肿瘤分子标志物是 肿瘤细胞或组织自身产生和释放的某种物质,常以 抗原、酶、激素等代谢产物的形式存在于肿瘤细胞内 或宿主体液内¹³。发现GC高危预警、早期诊断和有 效治疗的分子标志物,对GC的临床治疗具有重要意 义^[4]。本研究从基因表达数据库(Gene Expression Omnibus,GEO)中下载与GC有关的数据集,筛选差 异表达基因(differentially expressed gene, DEG),对 DEG做相关生物信息学分析,筛选可能参与GC发生 发展的关键基因(key gene),以期为GC的诊断、治疗 及预后判断提供更多有效的靶点。

1 材料与方法

1.1 数据的获取

从 GEO 数 据 库 (http://www.ncbi.nlm.nih.gov/ geo/) 中获取与 GC 相关的 3 组数据集:GSE33651、 GSE26899和GSE2701。GSE33651共有 52个样本, 包括GC组40个、正常组12个;GSE26899共有108个 样本,包括GC组96个、正常组12个;GSE2701共有 49个样本,包括GC组44个、正常组5个。

1.2 GC相关DEG的筛选

利用 GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/)以|log₂FC|>1且校正后P值(adj.P value)<0.05 为筛选标准分别对三组数据集进行 DEG 筛选。利用 R 软件包绘制火山图。为降低假阳性率,对3组数据 的 DEG 取交集绘制维恩图。

1.3 GO和KEGG富集分析

利用 DAVID(https://david.ncifcrf.gov/)对 DEG 进行基因本体(Gene Ontology,GO)功能注释,同时进行京都基因与基因组百科全书(Kyoto Encyclopedia of Genes and Genomes,KEGG)通路分析,以P<0.05为截取标准。

 蛋白互作网络(protein-protein interaction network, PPI)构建及关键基因获取

利用 STRING 数据库(https://string-db.org/)对 DEG 进行 PPI 构建,将此 PPI 数据导入 Cytoscape 软件,利用 MCODE 插件筛选 DEG 中的关键基因。筛选条件为 MCODE score>5,degree cut-off=2,node score cut-off= 0.2,Max depth=100,k score=2,获得关键基因。

1.5 关键基因共表达分析及生存率分析

利用 cBioPortal 在线分析工具(https://www.cbioportal.org)对筛选得到的关键基因进行共表达网络分 析、整体生存率及无病生存率分析。

1.6 关键基因生物学过程分析

利用 Cytoscape 中的 BINGO (Biological Networks Gene Oncology Tool)插件分析关键基因参与 的生物学过程和功能。

1.7 关键基因的层次聚类分析

使用 UCSC (University of Colifornia Santa Cruz, http://genome.ucsc.edu/)对关键基因进行层次聚类, 分析关键基因与肿瘤发生或性别等的关系。

2 结 果

2.1 筛选得出216个GC相关DEG

通过 GEO2R 对 3 组数据集进行分析,得到表达 上、下调的 DEG,绘制火山图(图1)。数据集 GSE33651共筛选出6067个DEG,其中上调基因4075 个,下调基因1992个(图1A);数据集GSE26899 共筛选出2248个DEG,其中上调基因1066个,下 调基因1182个(图1B);数据集GSE2701共2627个 DEG,其中上调基因1540个,下调基因1087个(图 1C)。将3个数据集筛选结果取交集制作维恩图(图 1D),得到差异共表达基因216个,其中上调基因153个、 下调基因108个。

2.2 DEG涉及的生物学过程与信号通路

GO分析结果(表1)表明,261个差异共表达基因 主要在对药物的反应、防御病毒、细胞增殖、各种器 官组织的发育、血管生成、胶原蛋白的合成与分解、 氧化还原等生物学过程富集。分子功能的变化主要 集中在蛋白质结合、蛋白质同源二聚活性、钙离子结 合上。细胞组成变化主要集中在细胞外基质(extracellular matrix,ECM)、细胞质核周区、细胞器膜、外 泌体、基底膜、内质网及溶酶体等。KEGG通路分析 结果(表2)表明,DEG主要富集于ECM受体相互作 用、蛋白质消化吸收、局部黏附、阿米巴病、PI3K-Akt 信号通路及癌症通路等。

2.3 PPI的构建及关键基因的获取

 \oplus

通过 STRING 网站对差异共表达基因构建 PPI (图2A),可见该网络有多个关键模块。

通过 Cytoscape 中的 MCODE 插件寻找密集连接 的区域,筛选 PPI 中最重要的模块,共获得 14 个关键 基因,分别是:PLOD1、PLOD3、COL1A1、COL1A2、 COL2A1、COL3A1、COL4A1、COL4A2、COL8A1、 COL12A1、COL15A1、ITGA2、LUM、SERPINH1(图 2B)。

2.4 基因共表达分析及生存曲线的构建

使用 cBioPortal 网站分析上述 14 个关键基因及 其共表达网络,结果见图3。图中粗体黑色圆圈节点 代表本试验中的关键基因,细框黑色节点代表其共 表达基因。从图中可以看出,部分基因有着紧密的 联系,分子网络密集。

利用cBioPortal中的数据分析关键基因表达改

变对 GC 患者整体生存率和无进展生存率的影响,结果见图4和图5。基因 COL3A1(P=0.0241)表达的改变显著降低了 GC 患者整体生存率;基因 IT-GA2(P=0.0679)的表达改变与GC 患者的无进展生存率降低相关;其他基因表达的改变对GC 患者的整体生存率和无进展生存率无明显影响。





图1 GC相关DEG的筛选

Fig.1 Screening of DEGs related to GC

GO ID	Description	Count	Р
GO:0042493	Response to drug	13	1.07E-05
GO:0051607	Defense response to virus	10	1.12E-05
GO:0001501	Skeletal system development	7	0.00109
GO:0008283	Cell proliferation	11	0.00111
GO:0005518	Collagen binding	9	3.64E-08
GO:0001525	Angiogenesis	15	4.11E-06
GO:0055114	Oxidation-reduction process	18	0.005086
GO:0070062	Extracellular exosome	84	3.38E-12
GO:0005764	Lysosome	8	0.038789
GO:0005604	Basement membrane	9	1.38E-05
GO:0031090	Organelle membrane	5	0.032981
GO:0048471	Perinuclear region of cytoplasm	16	0.027347
GO:0005783	Endoplasmic reticulum	22	0.005795
GO:0005515	Protein binding	159	1.31E-05
GO:0042803	Protein homo dimerization activity	23	8.04E-04
GO:0042802	Identical protein binding	26	7.29E-05
GO:0005509	Calcium ion binding	23	6.33E-04

表1 GC样本中DEG的GO分析 Tab.1 GO analysis of DEGs in gastric cancer samples

Tab.2 KEGG analysis of DEGs in gastric cancer samples				
Pathway ID	Description	Count	Р	
Hsa04512	ECM receptor interaction	12	2.35E-09	
Hsa04974	Protein digestion and absorption	8	5.13E-05	
Hsa04510	Focal adhesion	11	8.88E-05	
Hsa05200	Pathways in cancer	10	0.03053	
Hsa04145	Phagosome	7	0.040735	
Hsa05146	Amoebiasis	9	3.91E-04	
Hsa04151	PI3K-Akt signaling pathway	12	0.030495	





图 2 PPI 网络图(A)及紧密连接度最高的基因模块(B) Fig.2 PPI network (A) and the gene module with the highest connection degree (B)



Blue arrow: interaction; Green arrow: state change;

Pink line: other 图 3 cBioPortal 分析共表达网络 Fig.3 Key genes and their co-expression genes were analyzed using cBioPortal

2.5 关键基因的生物学过程分析

对关键基因参与的生物学过程分析结果如图6 所示,节点颜色深浅代表显著性大小,节点大小表示 本体中涉及的基因数量。获得的14个关键基因主要 参与胶原纤维的组织、ECM的组织、细胞外结构的组 织、皮肤形态发生、胶原的生物合成过程及血管的发 育等生物学过程。

2.6 关键基因的层次聚类分析

经UCSC对关键基因进行层次聚类分析,结果见 图7。GC患者中男性的发病人数明显多于女性。与 正常胃组织相比,GC组织中基因PLOD1、PLOD3、 COL3A1、ITGA2、COL1A2、COL1A1、COL4A1、 LUM、COL12A1、SERPINH1、COL8A1表达上调,表 明这些基因的表达增加可能与GC的发生有很大的 关系。

3 讨论

GC是常见的恶性肿瘤之一,具有易复发、早转移、高死亡率的特点^[5]。为深入了解其中的分子机制,筛选可用于GC诊断和药物治疗的靶点的关键基因,本研究采用生物信息学方法对GEO数据库中的3组GC相关数据集进行分析,得到差异共表达基因216个。通过构建DEG的PPI并分析连接最紧密的模块,共获得14个关键基因。这些关键基因主要参与胶原纤维的组织、ECM的组织、细胞外结构的组织、皮肤形态发生、胶原的生物合成过程及血管的发育等生物学过程。

ECM 是肿瘤微环境中重要的成分,其蛋白质的 沉积和交联为癌细胞增殖、迁移及浸润提供了必要 的生物化学和生物物理的支持。胶原蛋白是 ECM 的 主要成分,胶原蛋白的交联与 ECM 的硬度有关,能够 增强癌细胞的迁移、侵袭和粘附性^[6-7]。研究^[8-9]表明, 肿瘤微环境中胶原蛋白的状态(含量、成熟度、形态 和结构)发生变化,胶原的特征可作为预测 GC 早期 淋巴结转移的独立指标。有研究^[10]发现,COL12A1 表达升高与GC的发生、侵袭、转移及临床分期高度 相关。本研究中筛选到的14个关键基因中有9个属 于胶原蛋白基因,层次聚类分析表明,基因COL1A1、 COL1A2、COL3A1、COL4A1、COL8A1、COL12A1的 表达增加与GC的发生有很大的关系。进一步对GC 患者生存率的分析表明 COL3A1 表达的改变与 GC 患者的不良预后高度相关。上述研究结果提示本研 究所筛选的一组胶原蛋白作为 GC 检测、预后判断及 治疗靶标的应用价值。







Fig.5 Progression-free survival analysis of key genes by cBioPortal

PLOD编码赖氨酰羟化酶,对于胶原蛋白的生物 合成、交联和沉积至关重要^[11],人类基因组PLOD包 括PLOD1~3三个成员。膀胱癌中检测到PLOD1异 常表达,并且通过siRNA介导的敲除或PLOD1抑制 剂处理显著降低了膀胱癌细胞的恶性表型^[12]。最新 的一项生物信息研究^[13]表明,PLOD家族基因表达的 升高与GC患者生存率呈高度负相关。SERPINH1定 位于内质网,作为胶原蛋白特异的分子伴侣参与胶 原蛋白的生物合成,进而促进ECM形成。在GC中 表达明显增高,且与肿瘤浸润深度、预后明显相关, 可作为癌症标志物^[14-15]。ITGA2编码整联蛋白α2亚 单位,可和β1亚单位形成异二聚体的穿膜受体,与



图 6 Cytoscape 分析关键基因参与的生物学过程 Fig.6 Biological process network of key genes analyzed by Cytoscape



图 7 关键基因的层次聚类分析 Fig.7 Hierarchical clustering analysis of key genes

ECM的胶原蛋白、层粘联蛋白等结合。α2β1整联蛋 白在多种癌细胞中高表达,从而促进癌细胞的迁移、 浸润及血管新生^[16-17]。通过特异性抗体阻断ITGA2 可以抑制体外GC细胞的增殖和迁移^[18]。LUM是富 含白氨酸的小蛋白酶家族的成员,作为肿瘤基因和 肿瘤抑制基因具有双重作用^[19]。研究^[20]表明,成纤维 细胞来源的LUM通过激活整联蛋白β1介导的FAK 信号通路而促进GC细胞的增殖、迁移和浸润。本研 究中聚类分析表明,与正常组织相比,GC组织中 PLOD1、PLOD3、SERPINH1、LUM、ITGA2基因表达 显著升高。生存曲线分析表明,ITGA2表达升高的 GC患者无病生存率明显低于低表达的GC患者。这 些结果提示这5个基因可作为GC治疗的潜在靶点及 预后判断的分子标志物。

综上所述,本研究运用生物信息学的方法对GC 芯片数据进行挖掘,筛选出可能参与GC发生和发展 的关键基因,这些基因作为潜在分子标志物将有助 于GC的早期诊断、治疗靶点选择和预后判断,并为 后续的研究提供指导。此外,本研究为利用生物信 息学方法对现有数据的二次挖掘,为阐明GC发生、 浸润、转移等机制还需进一步的实验室和临床验证。

[参考文献]

- [1] 张玲倩, 卢宁. miR-200c 在胃癌早期诊断中的作用研究现状及展 望[J]. 世界华人消化杂志, 2019, 27(06): 382-388. DOI: 10.11569/ wcjd.v27.i6.382.
- [2] 杨之洵,郑荣寿,张思维,等.中国胃癌发病趋势及预测[J].中国肿 瘤, 2019, 28(5): 321-326.
- [3] 杭佑宝,姜从桥. 胃癌相关肿瘤标志物的现状和研究进展[J]. 中华 全科医学, 2012, 10(04):610-611. DOI: 10.16766/j.cnki.issn.1674-4152.2012.04.087.
- [4] 丽敏,刘险峰,孙鲁山,等.胃癌分子标志物在临床中的研究进展 [J]. 中国医学工程, 2018, 26(09): 25-29. DOI: 10.19338/j. issn.1672-2019.2018.09.007.
- [5] AKAGI T, SHIRAISHI N, KITANO S. Lymph node metastasis of gastric cancer[J]. Cancers (Basel), 2011, 3(2): 2141-2159. DOI: 10.3390/cancers3022141.
- [6] DU H Z, PANG M, HOU X Y, et al. PLOD2 in cancer research[J]. Biomed Pharmacother, 2017, 90: 670-676. DOI: 10.1016/j. biopha.2017.04.023.
- [7] PEINADO H, MORENO-BUENO G, HARDISSON D, et al. Lysyl oxidase-like 2 as a new poor prognosis marker of squamous cell carcinomas[J]. Cancer Res, 2008, 68(12): 4541-4550. DOI: 10.1158/ 0008-5472.CAN-07-6345
- [8] ZHOU Z H, JI C D, XIAO H L, et al. Reorganized collagen in the tumor microenvironment of gastric cancer and its association with prognosis [J]. J Cancer, 2017, 8(8): 1466-1476. DOI:10.7150/jca.18466.
- [9] CHEN D X, CHEN G, JIANG W, et al. Association of the collagen signature in the tumor microenvironment with lymph node metastasis in early gastric cancer[J/OL]. JAMA Surg, 2019, 154(3): e185249[2020-02-13]. https://www.ncbi.nlm.nih.gov/pmc/articles/

PMC6439641/. DOI:10.1001/jamasurg.2018.5249.

- [10] JIANG X X, WU M J, XU X, et al. COL12A1, a novel potential prognostic factor and therapeutic target in gastric cancer[J]. Mol Med Rep, 2019, 20(4): 3103-3112. DOI:10.3892/mmr.2019.10548.
- [11] QI Y F, XU R. Roles of PLODs in collagen synthesis and cancer progression[J]. Front Cell Dev Biol, 2018, 6: 66. DOI: 10.3389/ fcell.2018.00066.
- [12] YAMADA Y, KATO M, ARAI T, et al. Aberrantly expressed PLOD1 promotes cancer aggressiveness in bladder cancer: a potential prognostic marker and therapeutic target[J]. Mol Oncol, 2019, 13(9): 1898-1912. DOI:10.1002/1878-0261.12532.
- [13] LI S S, LIAN Y F, HUANG Y L, et al. Overexpressing PLOD family genes predict poor prognosis in gastric cancer[J]. J Cancer, 2020, 11(1): 121-131. DOI:10.7150/jca.35763.
- [14] QI Y J, ZHANG Y, PENG Z Q, et al. SERPINH1 overexpression in clear cell renal cell carcinoma: association with poor clinical outcome and its potential as a novel prognostic marker[J]. J Cell Mol Med, 2018, 22(2): 1224-1235. DOI:10.1111/jcmm.13495.
- [15] ZHENG S X, YANG L H, DAI Y S, et al. Screening and survival analysis of hub genes in gastric cancer based on bioinformatics[J]. J Comput Biol, 2019, 26(11): 1316-1325. DOI:10.1089/cmb.2019.0119.
- [16] GONG J, LU X Y, XU J, et al. Coexpression of UCA1 and ITGA2 in pancreatic cancer cells target the expression of miR-107 through focal adhesion pathway[J]. J Cell Physiol, 2019, 234(8): 12884-12896. DOI:10.1002/jcp.27953.
- [17] XU Y, SHEN L F, LI F J, et al. MicroRNA-16-5p-containing exosomes derived from bone marrow-derived mesenchymal stem cells inhibit proliferation, migration, and invasion, while promoting apoptosis of colorectal cancer cells by downregulating ITGA2[J]. J Cell Physiol, 2019, 234(11): 21380-21394. DOI:10.1002/jcp.28747.
- [18] CHUANG Y C, WU H Y, LIN Y L, et al. Blockade of ITGA2 induces apoptosis and inhibits cell migration in gastric cancer[J/OL]. Biol Proced Online, 2018, 20: 10[2020-02-13]. https://www.ncbi.nlm. nih. gov/pmc/articles/PMC5928594/. DOI: 10.1186/s12575-018-0073-x.
- [19] CHEN X W, LI X, HU X J, et al. LUM expression and its prognostic significance in gastric cancer[J/OL]. Front Oncol, 2020, 10: 605 [2020-02-13]. https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC7242722/. DOI:10.3389/fonc.2020.00605.
- [20] WANG X F, ZHOU Q, YU Z J, et al. Cancer-associated fibroblastderived Lumican promotes gastric cancer progression via the integrin β1-FAK signaling pathway[J]. Int J Cancer, 2017, 141(5): 998-1010. DOI:10.1002/ijc.30801.

[收稿日期] 2020-03-03

[修回日期] 2020-06-29

[责任编辑] 黄静怡

 \oplus