

生物信息分析

DOI: 10.13406/j.cnki.cyxh.002700

基于生物信息学方法筛选结直肠癌转移的相关基因

丘新泽, 罗世波, 陶坤林, 闭彩莹, 吴江妮, 黄杰安, 刘诗权

(广西医科大学第二附属医院消化内科, 南宁 530000)

【摘要】目的:应用生物信息学对基因芯片数据进行分析,探讨结直肠癌的转移机制,寻找潜在的转移及预后标记物。**方法:**从 GEO 数据库选取 GSE41568、GSE68468 进行分析,使用 R 语言 limma 包筛选结直肠原发癌与转移瘤之间的差异基因,获得 2 个数据集的共同差异基因;运用 clusterProfiler 包进行功能富集分析并进行可视化;通过 STRING 数据库、Cytoscape 软件进行蛋白质互作网络分析及关键模块的筛选;使用 TCGA 数据库的结直肠癌数据对差异基因进行表达分析和生存分析。**结果:**共筛选出 108 个共同差异基因;通过基因富集分析发现,差异基因主要富集在补体及凝血级联等通路及生物学过程中;通过蛋白质互作网络分析发现 3 个关键模块;基于 TCGA 数据对共同差异基因分析发现,CLCA1、COLEC11、FCGBP、PDZD2、SERPINA1、SPINK4 共 6 个基因在 I-Ⅱ 和 Ⅲ-Ⅳ 期结直肠癌的表达有显著差异 ($P<0.05$),并且与结直肠癌的预后显著相关 ($P<0.05$)。**结论:**通过生物信息学筛选出 108 个共同差异基因及 3 个关键模块,并得到 6 个预后相关基因,为研究结直肠癌的转移机制及预后、靶向治疗提供了一定的理论支持。

【关键词】结直肠癌;转移;生物信息学分析;差异基因**【中图分类号】**R318.04**【文献标志码】**A**【收稿日期】**2020-08-13

Identification of related genes associated with colorectal cancer metastasis based on bioinformatics analysis

Qiu Xinze, Luo Shibo, Tao Kunlin, Bi Caiying, Wu Jiangni, Huang Jiean, Liu Shiquan

(Department of Gastroenterology, The Second Affiliated Hospital of Guangxi Medical University)

【Abstract】Objective: To explore the mechanisms of colorectal cancer (CRC) metastasis and find the potential biomarkers associated with metastasis and prognosis by analyzing the gene profiles with bioinformatics analysis. **Methods:** The gene profiles of GSE41568, GSE68468 were selected from the GEO database; R software with the limma package was used to analyze the data and screen the differentially expressed genes between primary tumors and metastatic tumors, and we got the common differentially expressed genes. ClusterProfiler package was used for the biological function enrichment analysis and visualization; STRING database and Cytoscape software were used to construct protein-protein interaction (PPI) network and identify the key modules; TCGA database was applied for the expression analysis and survival analysis of common differentially expressed genes. **Results:** We identified 108 common differentially expressed genes between primary tumors and metastatic tumors. Enrichment analysis indicated that the genes were enriched in complement and coagulation cascades pathways and biological processes, et al. We also found 3 key modules by PPI network analysis. The analysis of common differentially expressed genes based on TCGA data showed that there were significant differences in the expression of CLCA1, COLEC11, FCGBP, PDZD2, SERPINA1, and SPINK4 between the stage I-Ⅱ and Ⅲ-Ⅳ of CRC ($P<0.05$), with significant correlation with the prognosis of CRC ($P<0.05$). **Conclusion:** With bioinformatics analysis, a total of 108 common differentially expressed genes, 3 key modules and 6 genes associated with prognosis have been screened out, thereby providing the theoretical support for research on mechanisms of metastasis, prognosis prediction and targeted therapies of CRC.

【Key words】colorectal cancer; metastasis; bioinformatics analysis; differentially expressed gene**作者介绍:**丘新泽, Email: qiu xinze@hotmail.com,

研究方向: 消化系统肿瘤的生物信息学分析及临床研究。

通信作者:刘诗权, Email: poempower@163.com。**基金项目:**国家自然科学基金资助项目(编号:81460380); 广西自然

科学基金资助项目(编号:2017GXNSFAA198019)。

优先出版: <https://kns.cnki.net/kcms/detail/50.1046.r.20201112.1729.006.html>

(2020-11-13)

结直肠癌是世界上最普遍的癌症之一。全球癌症统计报告显示,结直肠癌发病率及死亡率分别居世界第 3 位和第 2 位^[1]。随着外科手术、化疗、靶向药物等治疗手段的进步,结直肠癌患者的 5 年生存率逐渐上升,但在晚期癌症患者中,其 5 年生存率仅 12%~14%^[2]。肝脏及肺脏是结直肠癌远处转移的好发器官^[3]。结直肠癌转移与长链非编码 RNA、微小 RNA、肿瘤微环境构建、上皮-间充质转化等密切相关^[4]。基因芯片是 20 世纪末出现的一种基因检测技术,能够检测上万基因转录本的表达水平,现已较为成熟。虽然已有关于结直肠癌转移的基因芯片研究,但都具有一定的局限性,如样本量不足、芯片数据单一、差异分析方法欠妥等。本研究纳入样本数较多的结直肠癌芯片数据,发掘出与结直肠癌转移相关的差异基因,进行富集分析和蛋白质互作网络分析,并基于肿瘤基因组图谱(the cancer genome atlas,TCGA)的二代测序数据对差异基因进行表达分析和生存分析,为研究结直肠癌的转移机制及潜在治疗靶点提供一定的根据。

1 材料与方法

1.1 数据来源

从基因表达数据库(gene expression omnibus,GEO)中选取编号为 GSE41568、GSE68468 的基因芯片数据集。GSE41568 基于 GPL570 平台(Affymetrix Human Genome U133 Plus 2.0 Array)测得。本研究纳入 39 例原发癌及 88 例肝肺转移瘤样本^[5];GSE68468 由 GPL96 平台(Affymetrix Human Genome U133A Array)测得,纳入结直肠癌原发灶 186 例及肝肺转移瘤 67 例^[6]。通过 TCGA 数据库下载结直肠癌表达数据及相应的临床数据,其中有完整临床分期及生存资料的样本共 583 例。

1.2 差异表达分析

基于 R 语言(版本:4.0.0),使用 limma 包分别对 2 个基因芯片数据集进行差异分析,筛选出原发癌和转移瘤的差异基因^[7]。差异基因筛选阈值为 $|\log_2FC| > 1$ 并且经 FDR 校正后 $P < 0.05$ 。通过绘制 Venn 图取 2 个基因芯片差异基因的交集,以获得它们的共同差异基因进行后续分析。

1.3 基因功能富集分析

使用 clusterProfiler 包对获得的共同差异基因进行基因本体论(gene ontology,GO)功能富集和 KEGG 通路富集,以明确基因参与的生物学功能及生物学通路等,筛选阈值为经 FDR 校正后 $P < 0.05$ ^[8]。基因本体论从 3 个方面对基因进行功能注释,包括分子功能、生物学过程、细胞组分^[9]。

1.4 蛋白质互作网络分析

利用 STRING 在线数据库(<http://string-db.org>, 版本:11.0)和 Cytoscape(版本:3.8.0)软件对差异基因进行蛋白质互作(protein-protein interaction,PPI)网络分析。单个基因的边数越多,越能提示其在蛋白质互作网络中的核心地位。Cytoscape 的 MCODE 插件能够根据基因间的联系程度挖掘出其中的关键模块,以深入理解分子间的调控及作用机制。关键模块筛选阈值设为节点数及评分均大于 5。

1.5 共同差异基因的表达分析及生存分析

将 TCGA 的结直肠癌样本按照临床分期分为 I-II 和 III-IV 期 2 组,比较共同差异基因在 2 组样本的表达水平。以基因表达水平的中位数作为截断值,将结直肠癌样本分为高表达组和低表达组,通过 survival 包进行生存分析。使用 R 语言进行统计学分析,差异表达分析使用两样本 Wilcoxon 秩和检验;生存分析使用 log-rank 检验。检验水准 $\alpha = 0.05$ 。

2 结果

2.1 共同差异基因的筛选

通过对原发癌和转移瘤的基因进行差异表达分析,GSE41568 数据集中有 333 个差异基因;GSE68468 数据集集中有 163 个差异基因。通过 Venn 图对 2 个数据集的差异基因结果取交集,共得到 108 个共同差异基因(图 1)。

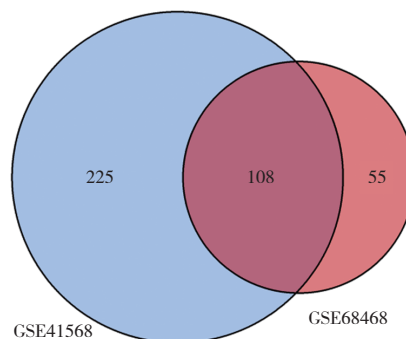


图 1 差异表达基因的 Venn 图

2.2 共同差异基因富集分析

GO 功能富集分析结果显示,这些基因主要介入血小板脱粒、体液免疫反应、急性炎症反应、补体激活、急性期反应等生物过程(图 2A),参与酶抑制剂活性、肽酶抑制剂活性等分子功能(图 2B),富集在血液微粒、含胶原的细胞外基质、内质网腔、囊泡腔等部位(图 2C)。KEGG 通路富集分析发现,共同差异基因主要集中在补体和凝血级联、药物代谢-细胞色素 P450 酶、胆固醇代谢、细胞色素 P450 酶代谢外源物、胱病毒病等通路(图 2D)。共有 24 个差异基因富集在补体和凝血级联通路中,通过 KEGG 通路图可以显示出这些基因在该通路的发挥的作用(图 3)。

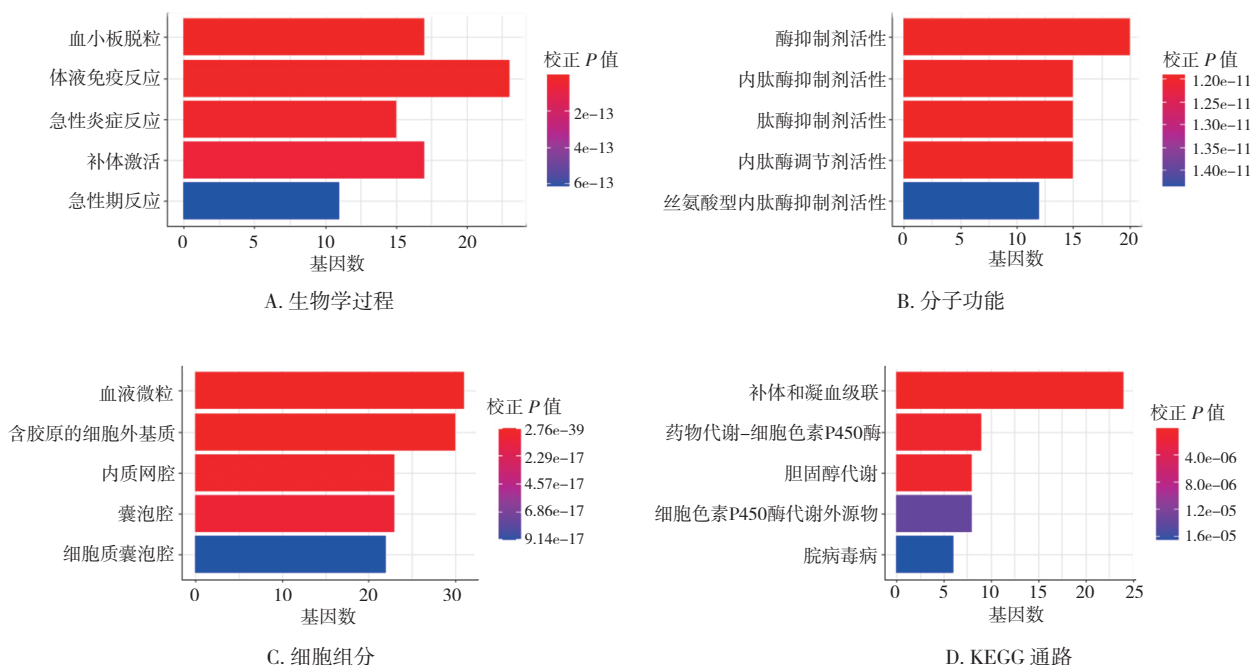


图2 GO及KEGG通路富集分析

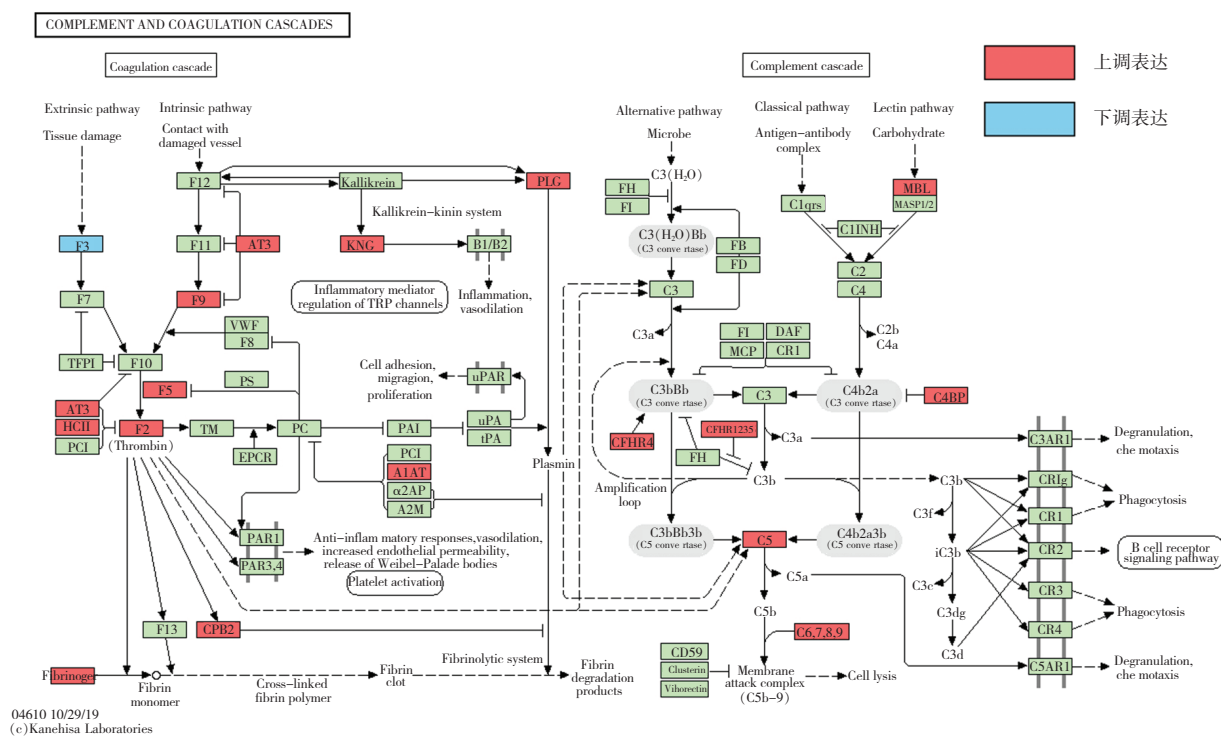


图3 差异表达基因参与凝血及补体级联通路

2.3 蛋白质互作网络分析与关键模块的筛选

使用 STRING 在线数据库和 Cytoscape 软件构建蛋白质互作网络,包括 99 个节点和 1 018 个边缘(图 4A)。本研究通过 MCODE 筛选出了 3 个关键模块:模块 1(评分:28.065),模块 2(评分:7.750),模块 3(评分:5.200);模块 1 中的黄色基因是富集在补体和凝血级联通路中的差异基因,在模块 1 基因中的占比为 34.38%(图 4B)。

2.4 共同差异基因的表达分析及预后分析

通过 TCGA 的结直肠癌数据对共同差异基因进行分析。结果显示,共有 6 个基因在 I-Ⅱ 和 Ⅲ-Ⅳ 期结直肠癌中的表达差异有统计学意义,并且表达水平和结直肠癌患者总生存时间显著相关,其中 PDZD2、COLEC11 在 Ⅲ-Ⅳ 期结直肠癌中高表达,而 CLCA1、SERPINA1、FCGBP、SPINK4 则为低表达(图 5)。

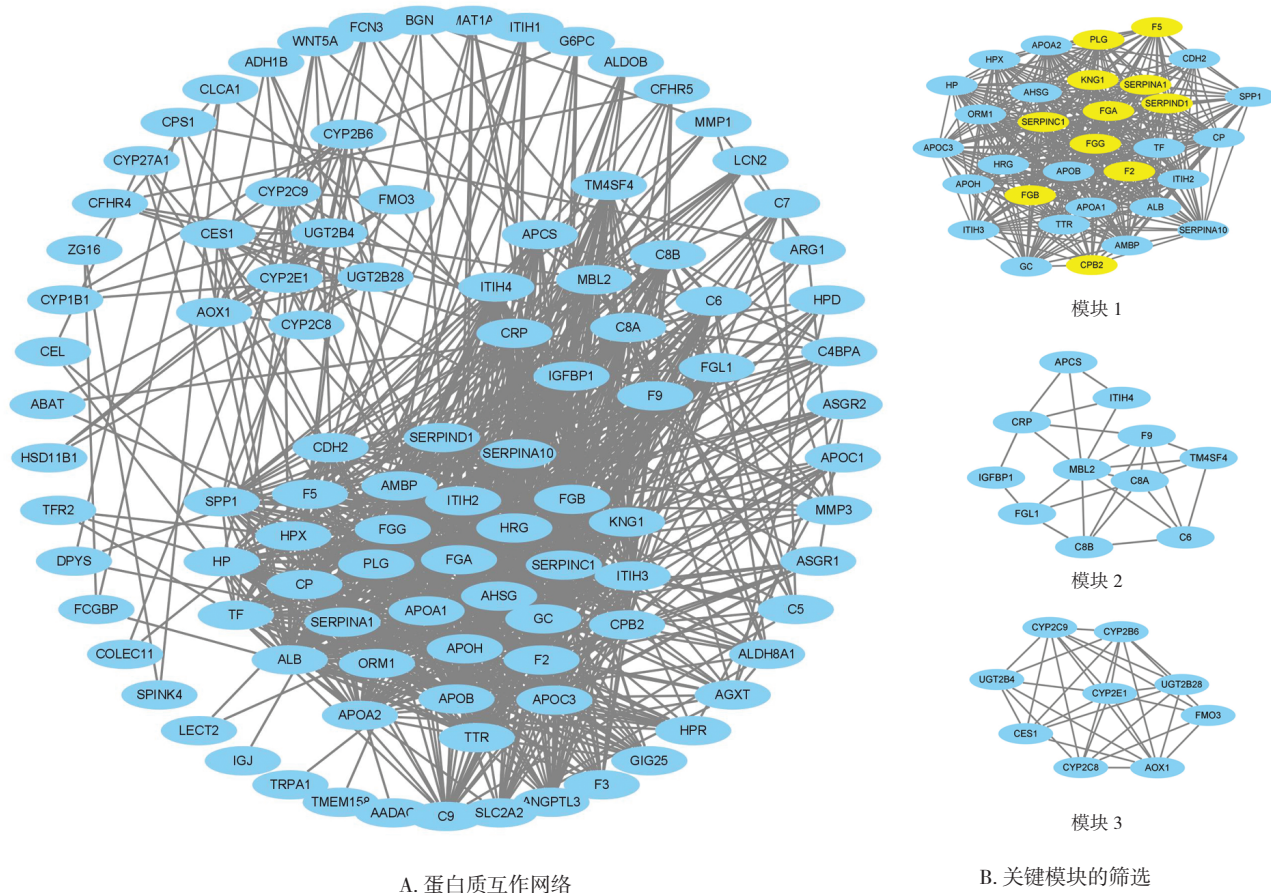


图 4 构建 PPI 网络并筛选出 3 个关键模块

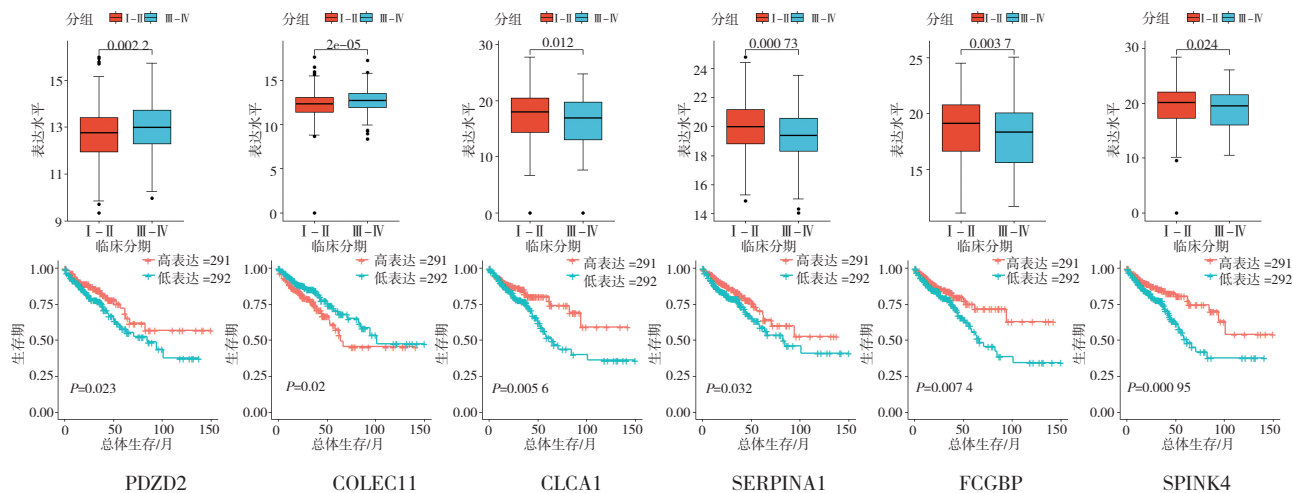


图 5 通过 TCGA 数据对共同差异基因进行表达分析及预后分析筛选出 6 个基因

3 讨论

转移是结肠直肠癌患者预后不良的重要危险因素,目前结肠直肠癌转移的机制尚未完全阐释清楚。因

此,筛选结肠直肠癌转移相关基因,进而探索结肠直肠癌转移的机制,寻找其治疗靶点尤为重要。基因芯片技术在基因功能、转录调控、分子标记物等研究中被广泛使用,能够高通量检测基因的表达量。通过对芯片数据进行生物信息学分析,为进一步探索

结直肠癌的转移机制及预后治疗提供了方向。

本研究基于 R 语言对 GSE41568、GSE68468 进行差异表达分析,共获得 108 个共同差异基因。通过 GO 及 KEGG 富集分析发现,差异基因与补体级联、凝血级联的激活密切相关。同时,模块 1 的较多基因和补体及凝血级联通路相关,模块 1 基因可能和补体及凝血通路的激活密切相关。肿瘤微环境影响肿瘤的生长和迁徙,而补体系统在其中起着重要且复杂的作用^[10]。大多数情况下,补体激活产生的炎症反应对清除人体内抗原或肿瘤细胞等具有帮助。但是,进一步研究发现,某些肿瘤也能够从补体的激活中获益。补体级联的激活促进了肿瘤血管生成,并且构建了免疫抑制环境,进而影响肿瘤细胞的存活及迁移^[11]。补体 5(complement 5,C5)是补体系统的成员之一,C5a 是 C5 的裂解片段。Bareke H 等^[12]认为,C5a 低表达有抗肿瘤作用,而高表达则促进肿瘤进展。有研究表明,C5a 在调节 M2 型巨噬细胞中起关键作用,而 M2 型巨噬细胞进一步通过核因子- κ B 信号通路促进结肠癌的肝转移^[13]。另 1 项研究发现,C5a/C5aR 通路通过激活 PI3K/AKT 信号传导,进而抑制 p21/p-p21 的表达,导致胃癌的发生发展^[14]。凝血级联中的凝血酶及纤维蛋白在一些恶性肿瘤中处于高表达状态。Kuang MY^[15]、Adams GN^[16]、Wojtukiewicz MZ^[17]等认为,凝血酶增加癌细胞的黏附能力,使癌细胞跨内皮迁移,并促进肿瘤血管生成,进而影响癌症进展。同时,凝血级联的激活也进一步促进了补体级联的激活^[18]。补体与凝血级联在结直肠癌转移中的具体机制尚不清楚。本研究通过富集分析及 PPI 网络分析为此提供了一定的实验线索。

通过 TCGA 数据库进一步分析发现,6 个基因与结直肠癌的进展和预后密切相关。CLCA1 蛋白是钙激活氯离子通道蛋白家族的成员。有研究表明,CLCA1 的表达与结直肠癌的预后相关,CLCA1 可能通过抑制 Wnt/ β -catenin 信号传导通路和上皮间充质转化过程而发挥抑癌作用^[19-20]。PDZD2 是一种含 PDZ 结构域的分泌肽。有研究报道,PDZD2 在前列腺癌、乳腺癌、肝癌中起肿瘤抑制作用^[21]。Tam CW 等^[22]的研究则表明,PDZD2 通过减慢前列腺癌细胞进入细胞周期 S 期而发挥抗肿瘤作用;PDZD2 及其

裂解产物 sPDZD2 通过 p53 依赖性和非 p53 依赖性通路抑制肿瘤的生长,并提出了利用 sPDZD2 来治疗癌症的新思路。SERPINA1 是一种丝氨酸蛋白酶抑制剂,已经发现其与肺癌、胃癌及结肠癌的进展相关^[23-25]。有研究发现,Snail 作用于 SERPINA1 增强子,上调 SERPINA1 表达,进而增加纤维连接蛋白的表达,促进结直肠癌的侵袭转移^[26]。FCGBP 是一种能和 IgG 的 Fc 片段互相作用的结合蛋白,介导维持正常肠黏膜结构^[27]。Harada N 等^[28]研究提示,FCGBP 可能在肠道免疫反应和炎症中起重要作用。COLEC11 是凝集素蛋白家族的成员之一,与补体凝集素途径的激活密切相关^[29-30]。SPINK4 是一种包含 kazal 结构域的丝氨酸蛋白酶抑制剂。Xie MZ 等^[31]研究表明,SPINK4 可作为一种血清标记物在结直肠癌的诊断中发挥作用。FCGBP、COLEC11、SPINK4 这 3 个基因目前缺乏癌症领域研究,仍需后续研究进一步探索其在结直肠癌转移中的作用。

综上所述,本研究运用生物信息学方法,发现了与结直肠癌转移相关的 108 个基因。通过富集分析得到的生物学过程及通路信息,以及 PPI 分析发现的 3 个关键模块,为结直肠癌转移机制的实验研究提供了一定的思路及理论根据。同时,通过 TCGA 数据库进一步分析得到的 6 个基因可能是结直肠癌诊断及靶向治疗的关键所在,仍有必要进一步挖掘其潜在价值。

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.
- [2] Siegel RL, Miller KD, Fedewa SA, et al. Colorectal cancer statistics, 2017[J]. CA Cancer J Clin, 2017, 67(3): 177-193.
- [3] Riihimäki M, Hemminki A, Sundquist J, et al. Patterns of metastasis in colon and rectal cancer[J]. Sci Rep, 2016, 6: 29765.
- [4] 赵亮, 丁彦青. 结直肠癌转移机制研究进展[J]. 浙江大学学报(医学版), 2014, 43(4): 486-493.
- [5] Zhao L, Ding YQ. Progression in metastasis of colorectal carcinoma[J]. J Zhejiang Univ Med Sci, 2014, 43(4): 486-493.
- [6] Gavert N, Sheffer M, Raveh S, et al. Expression of L1-CAM and ADAM10 in human colon cancer cells induces metastasis[J]. Cancer Res,

2007,67(16):7703–7712.

[6] Lu M,Zessin AS,Glover W,et al. Activation of the mTOR pathway by oxaliplatin in the treatment of colorectal cancer liver metastasis[J]. PLoS One,2017,12(1):e0169439.

[7] Ritchie ME,Phipson B,Wu D,et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies[J]. Nucleic Acids Res,2015,43(7):e47.

[8] Yu GC,Wang LG,Han YY,et al. clusterProfiler:an R package for comparing biological themes among gene clusters[J]. OMICS,2012,16(5):284–287.

[9] Ashburner M,Ball CA,Blake JA,et al. Gene ontology:tool for the unification of biology. The Gene Ontology Consortium[J]. Nat Genet,2000,25(1):25–29.

[10] Roumenina LT,Daugan MV,Petitprez F,et al. Context-dependent roles of complement in cancer[J]. Nat Rev Cancer,2019,19(12):698–715.

[11] Okrój M,Potempa J. Complement activation as a helping hand for inflammophilic pathogens and cancer[J]. Front Immunol,2019,9:3125.

[12] Bareke H,Akbuga J. Complement system's role in cancer and its therapeutic potential in ovarian cancer[J]. Scand J Immunol,2018,88(1):e12672.

[13] Piao CM,Zhang WM,Li TT,et al. Complement 5a stimulates macrophage polarization and contributes to tumor metastases of colon cancer[J]. Exp Cell Res,2018,366(2):127–138.

[14] Chen J,Li GQ,Zhang L,et al. Complement C5a/C5aR pathway potentiates the pathogenesis of gastric cancer by down-regulating p21 expression[J]. Cancer Lett,2018,412:30–36.

[15] Kuang MY,Peng YZ,Tao XT,et al. FGB and FGG derived from plasma exosomes as potential biomarkers to distinguish benign from malignant pulmonary nodules[J]. Clin Exp Med,2019,19(4):557–564.

[16] Adams GN,Rosenfeldt L,Frederick M,et al. Colon cancer growth and dissemination relies upon thrombin,stromal PAR-1,and fibrinogen [J]. Cancer Res,2015,75(19):4235–4243.

[17] Wojtukiewicz MZ,Hempel D,Sierko E,et al. Thrombin—unique coagulation system protein with multifaceted impacts on cancer and metastasis[J]. Cancer Metastasis Rev,2016,35(2):213–233.

[18] Dzik S. Complement and coagulation:cross talk through time[J]. Transfus Med Rev,2019,33(4):199–206.

[19] Chen TJ,He HL,Shiue YL,et al. High chloride channel accessory 1 expression predicts poor prognoses in patients with rectal cancer receiving chemoradiotherapy[J]. Int J Med Sci,2018,15(11):1171–1178.

[20] Li XF,Hu WX,Zhou JJ,et al. CLCA1 suppresses colorectal can-

cer aggressiveness via inhibition of the Wnt/beta-catenin signaling pathway[J]. Cell Commun Signal,2017,15(1):38.

[21] Tam CW,Liu VWS,Leung WY,et al. The autocrine human secreted PDZ domain-containing protein 2(sPDZD2) induces senescence or quiescence of prostate,breast and liver cancer cells via transcriptional activation of p53[J]. Cancer Lett,2008,271(1):64–80.

[22] Tam CW,Cheng AS,Ma RM,et al. Inhibition of prostate cancer cell growth by human secreted PDZ domain-containing protein 2,a potential autocrine prostate tumor suppressor[J]. Endocrinology,2006,147(11):5023–5033.

[23] Higashiyama M,Doi O,Kodama K,et al. An evaluation of the prognostic significance of alpha-1-antitrypsin expression in adenocarcinomas of the lung:an immunohistochemical analysis[J]. Br J Cancer,1992,65(2):300–302.

[24] Tahara E,Ito H,Taniyama K,et al. Alpha1-antitrypsin,alpha1-antichymotrypsin,and alpha2-macroglobulin in human gastric carcinomas:a retrospective immunohistochemical study[J]. Hum Pathol,1984,15(10):957–964.

[25] Karashima S,Kataoka H,Itoh H,et al. Prognostic significance of alpha-1-antitrypsin in early stage of colorectal carcinomas[J]. Int J Cancer,1990,45(2):244–250.

[26] Kwon CH,Park HJ,Choi JH,et al. Snail and serpinA1 promote tumor progression and predict prognosis in colorectal cancer[J]. Oncotarget,2015,6(24):20312–20326.

[27] Johansson MEV,Thomsson KA,Hansson GC. Proteomic analyses of the two mucus layers of the colon barrier reveal that their main component,the Muc2 mucin,is strongly bound to the Fcgbp protein[J]. J Proteome Res,2009,8(7):3549–3557.

[28] Harada N,Iijima S,Kobayashi K,et al. Human IgGfC binding protein(FcγBP) in colonic epithelial cells exhibits mucin-like structure [J]. J Biol Chem,1997,272(24):15232–15241.

[29] Aagaard JB,Bjerrum KB,Graversen JH,et al. Novel monoclonal antibodies against native human collectin L1,using the heteromeric complex CL-LK as immunogen[J]. Monoclon Antibodies Immunodiagn Immunother,2018,37(5):207–211.

[30] Gong KJ,Xia M,Wang YQ,et al. Importance of glycosylation in the interaction of Tamm-Horsfall protein with collectin-11 and acute kidney injury[J]. J Cell Mol Med,2020,24(6):3572–3581.

[31] Xie MZ,Li KZ,Li JL,et al. Association and diagnostic value of serum SPINK4 in colorectal cancer[J]. PeerJ,2019,7:e6679.

(责任编辑:唐秋姗)