

基于 GEO 数据库整合 miRNA-mRNA 表达谱筛选卵巢癌的关键基因分子及生物信息分析

李 超, 朱晓丹, 张玲华, 杨兴坤 (佛山市妇幼保健院, 广东佛山 528000)

摘要: 目的 卵巢癌是妇产科恶性肿瘤死亡的主要原因之一, 但其致病分子机制还未被清晰阐明。该研究通过整合生物信息学方法, 旨在挖掘其潜在的关键基因分子及生物学功能, 以便更全面地揭示其发病机制。**方法** 从 GEO (Gene Expression Omnibus) 数据库下载 miRNA 和 mRNA 表达谱芯片集。通过 R 语言“limma”包筛选差异表达的 miRNA 和 mRNA。通过 FunRich 软件对筛选的差异 miRNA 进行靶标预测, 并与筛选的差异 mRNA 取交集得到共有差异基因。通过 R 语言“clusterProfiler”包对共有差异基因进行富集分析和通路注释以挖掘其生物学功能。运用 string 数据库和 cytoscape 软件进一步构建 miRNA-靶基因调控网络, 并鉴定分子网络中的关键基因分子。**结果** 共筛选鉴定出 167 个共有差异基因。富集分析表明共有差异基因主要涉及细胞外组织、胚胎器官发育、突触后特化、胶原三聚体和 DNA 结合转录激活等生物过程; 通路注释表明共有差异基因主要参与蛋白质的消化吸收和松弛素信号通路行为。分子网络分析筛选鉴定了 10 个候选关键基因, 并发现 miR-29c-3p, miR-1271-5p 和 miR-133b 抑癌分子与共有差异基因存在最广泛的靶向关系, 处于调控中枢核心地位。上述关键基因分子在卵巢癌发生发展中扮演了重要角色。**结论** 该研究采用的系统整合方法学和鉴定的关键基因分子有助于揭示卵巢癌的潜在致病机制, 也为卵巢癌的早期筛查提供新的候选标志物。

关键词: 卵巢癌; 差异表达基因; 生物信息分析; 分子网络; GEO。

中图分类号: R737.31; R730.43 文献标识码: A 文章编号: 1671-7414 (2021) 05-038-05

doi:10.3969/j.issn.1671-7414.2021.05.008

Screening Key Biomarkers and Bioinformatic Analysis of Ovarian Cancer by Integrated miRNA-mRNA Expression Profiles Based on GEO Database

LI Chao, ZHU Xiao-dan, ZHANG Ling-hua, YANG Xing-kun

(Foshan Women and Children Hospital, Guangdong Foshan 528000, China)

Abstract: Objective Ovarian cancer is one of the leading causes of death in gynecological malignancies, of which molecular mechanism hasn't been elucidated clearly yet. The research aims to reveal the potential key molecular and biological processes of ovarian cancer by means of integrated bioinformatics, in order to more fully clarify its pathogenesis. **Methods** The microarray sets of miRNA and mRNA expression profiles were downloaded from the GEO (Gene Expression Omnibus) database. The differentially expressed miRNAs and mRNAs were screened by the “limma” package of R language. The target prediction was performed on the differentially expressed miRNAs identified by the FunRich program and overlapped differentially expressed genes (DEGs) were obtained combined with miRNA and mRNA datasets. The overlapped DEGs in the network were analyzed to explore the biological processes involved by enrichment and pathway analysis by the “clusterProfiler” package of the R language. The regulatory network of miRNA-gene was further constructed by string database and cytoscape software, and the key molecular were identified in the molecular protein-protein interaction (PPI) network among DEGs. **Results** A total of 167 overlapped DEGs were identified. The enrichment showed that the overlapped DEGs were mainly involved in process named extracellular related organization, embryonic organ development, postsynaptic specialization, collagen trimer and DNA-binding transcription activator. The pathway analysis showed that these DEGs were involved in protein digestion and absorption and relaxin signaling pathway. The PPI molecular network identified 10 key genes, and found that miR-29c-3p, miR-1271-5p and miR-133b, existed the most extensive targeting relationship with overlapped DEGs, being three key miRNAs of the regulatory network, which played the role of tumor suppressor. These key molecules may play an important role in the occurrence and development of ovarian cancer. **Conclusion** The methodology used and identification of key molecules in this study contributed to understanding the pathogenesis of ovarian cancer and providing new candidate biomarkers for early screening of ovarian cancer.

基金项目: 佛山市遗传病精准诊断工程技术研究中心项目 (NO.2020001003953), 佛山市科学技术局审批。

作者简介: 李超 (1990-), 男, 医学硕士, 主管检验师, 研究方向: 肿瘤和遗传病的细胞与分子诊断, E-mail:lichao1990@hotmail.com。

通讯作者: 杨兴坤, 女, 医学遗传学博士, 副主任医师, 研究方向: 肿瘤遗传咨询和罕见病的基因诊断, E-mail:yangxingkun@126.com。

Keywords: ovarian cancer; DEGs; bioinformatics analysis; molecular network; GEO.

卵巢癌是妇产科恶性肿瘤死亡的主要原因之一,最新数据显示2018年全球有295 414例新发患者^[1]。卵巢癌的总体五年生存率低于45%,主要是由于诊断较晚已发生远处转移。目前临床实验室采用的CA125等标志物缺乏特异性,超声检查尤其易漏诊早期病例^[2-4]。因此卵巢癌新型标志物的再挖掘,疾病关键基因的再鉴定是其预防控制需要重点解决的难题。

近年来表达谱的广泛应用积累了海量的基因组学数据,这有赖被深入挖掘和解读。而生物信息学能够从整体层面揭示复杂疾病的分子靶标,属于肿瘤等研究领域较为革新的技术手段。检索国内外近三年的相关研究,已有表达谱芯片与生物信息学联合挖掘卵巢癌关键基因的多篇文章发表^[5-8]。然而绝大多数课题组选取的均为同一类型的芯片数据集,存在准确度不高而产生假阳性的问题,故可能难以获得可靠结果。miRNA是一种非编码RNA,主要通过碱基配对与mRNA内的互补序列结合,导致mRNA沉默而负向调控下游基因的表达,广泛参与细胞发育、分化和细胞周期^[9]。多项研究发现几乎所有类型的肿瘤均可发生miRNA失调,从而影响靶标基因的表达^[10-11]。因此本研究采用卵巢癌miRNA-mRNA表达谱数据集进行联合分析,在国内还尚属首次研究,具有较好的科学性和创新性。本文通过整合生物信息学手段,构建分子蛋白调控网络并鉴定卵巢癌致病关键分子及其涉及的生物学功能,将为阐明卵巢癌的发生发展调控机制提供一些科学精准的理论依据。

1 材料与方法

1.1 资料来源 GEO (<https://www.ncbi.nlm.nih.gov/geo/>) 是一个公共的功能基因组学数据库,由美国国立生物技术信息中心维护,收录了全球大量的高通量基因表达数据。本研究从中检索卵巢癌的表达谱芯片,限定实验类型为肿瘤组织对比正常组织。最终确定了GSE119055和GSE66957两个符合条件的数据集。前者为miRNA表达谱,包含了3个正常组织和6个肿瘤组织。后者为mRNA表达谱,包含了12个正常组织和57个肿瘤组织。本研究资料来源于公开数据库,无需医学伦理学再审查。

1.2 方法与统计学分析

1.2.1 差异基因:针对入选的mRNA数据集,根据其平台注释文件,将探针ID转换为对应的基因名称;采用R语言的“limma”包分析挑选差异基因^[12],定义命令包中的参数 $\text{adj.P.Val} < 0.05$ 且 $\log\text{FC}$ 绝对值 > 1.5 的基因为具有统计学意义的差异表达基因。利用R包“org.Hs.eg.db”转换基因名称为对应的基因

ID^[13],以便后续进行富集分析。miRNA数据集采用同样方法和参数标准进行处理。

1.2.2 靶标预测:采用FunRich软件进行筛选的差异miRNA的下游靶标预测,将预测的基因列表结果与mRNA数据集筛选到的差异基因取交集,得到共有差异基因用于后续的调控网络、关键基因鉴定和功能富集等过程分析^[14]。

1.2.3 调控网络:根据miRNA负向调控靶基因的原理,纳入调控网络的miRNA与靶基因表达变化要求负相关,筛选出表达情况符合条件的miRNA-靶基因组合来构建调控网络。该构建鉴定过程采用cytoscape3.7.1软件进行可视化,直观剖析该调控网络。

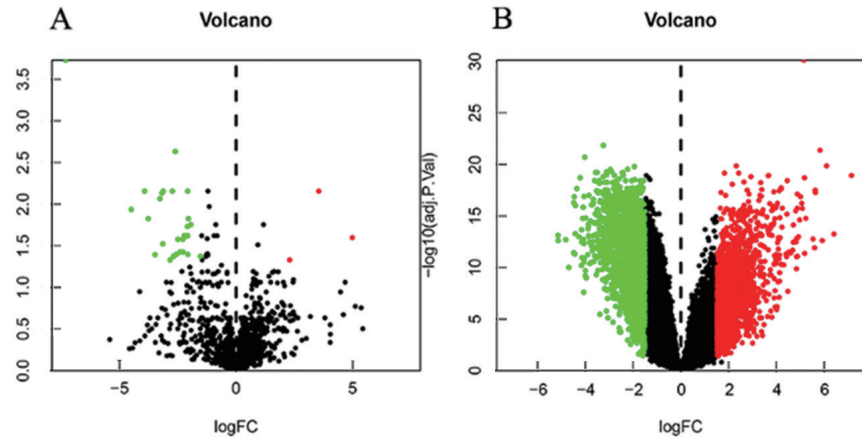
1.2.4 关键基因:String (<https://string-db.org/>) 数据库是用于研究基因编码蛋白之间的相互作用网络,用于挖掘核心的调控关键基因。通过string进行共有差异基因的分子蛋白互作网络分析。接着采用cytoscape3.7.1版本中的“cytohubba”模块进行关键基因的识别^[15],数据导入来源为string数据库获得的网络结果文件,选择MCC算法,得到排名前10的关键基因。

1.2.5 功能注释:利用R包“clusterProfiler”进行共有差异基因的GO (Gene Ontology) 富集和KEGG (Kyoto Encyclopedia of Genes and Genomes, 京都基因与基因组百科全书) 通路功能注释^[13]。GO富集分为三大类,分别是:分子生物学功能 (Molecular Function, MF)、生物学过程 (Biological Process, BP) 和细胞学组分 (Cellular Components, CC)。KEGG是一个整合了基因组、化学和系统功能信息的综合数据库,被广泛用于基因通路的富集注释。设置命令包中的筛选参数 $\text{pvalueCutoff} = 0.05$ 为具有生物统计学意义。

2 结果

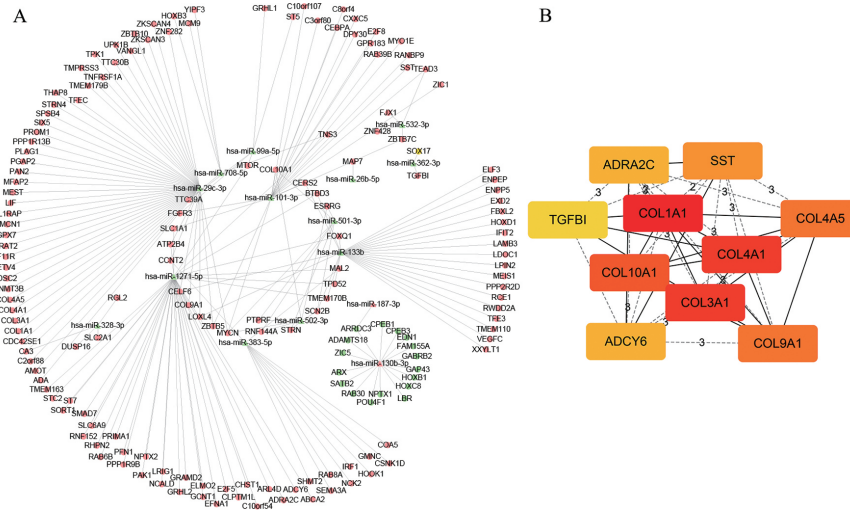
2.1 差异基因 GSE119055数据集共鉴定出31个差异表达的miRNA,其中28个为下调的miRNA,3个为上调的miRNA。GSE66957数据集共鉴定出3 844个差异表达的基因,其中1 995个为下调的基因,1 849个为上调的基因。差异表达分子分布图见图1。

2.2 共有差异基因与调控网络 根据筛选的差异miRNA和mRNA列表集,结合FunRich软件预测得到的差异miRNA的靶基因,取重叠的交集基因得到共有差异基因文件。将结果文件导入cytoscape软件可视化分析,结果如图2A所示。其中miR-29c-3p, miR-1271-5p和miR-133b处于网络调控的中枢地位,调控的靶基因数量最多。



A: 差异表达的 miRNA; B: 差异表达的基因。横轴: $\log_2(FC)$; 纵轴: $-\log_{10}(\text{adj.P.Val})$ 。FC 表示肿瘤样本对比正常组织的表达变化倍数, adj.P.Val 表示校准后的 P 值。绿色: 差异表达下调; 红色: 差异表达上调。

图1 差异表达分子的火山图分布情况



A: miRNA-靶基因调控网络筛选图。椭圆表示差异靶基因, 三角形表示差异 miRNA, 线代表调控存在靶向调控关系。红色表示表达上调, 绿色表示表达下调。B: 关键共有差异基因鉴定网络图。红色越深, 得分越高, 生物统计学意义越显著。

图2 关键网络基因的筛选与鉴定图

2.3 关键基因鉴定 通过 string 数据库和“cytohubba”模块鉴定共有差异基因中的关键基因结果见表1, 均为过表达差异基因。根据最新的 MCC 算法筛选前10位关键基因, 分别为 COL1A1, COL3A1,

COL4A1, COL10A1, COL9A1, COL4A5, SST, ADRA2C, ADCY6 和 TGFBI。值得注意的是, COL1A1, COL3A1 和 COL4A1 得分最高, 处于网络的关键核心位置, 见图2B。

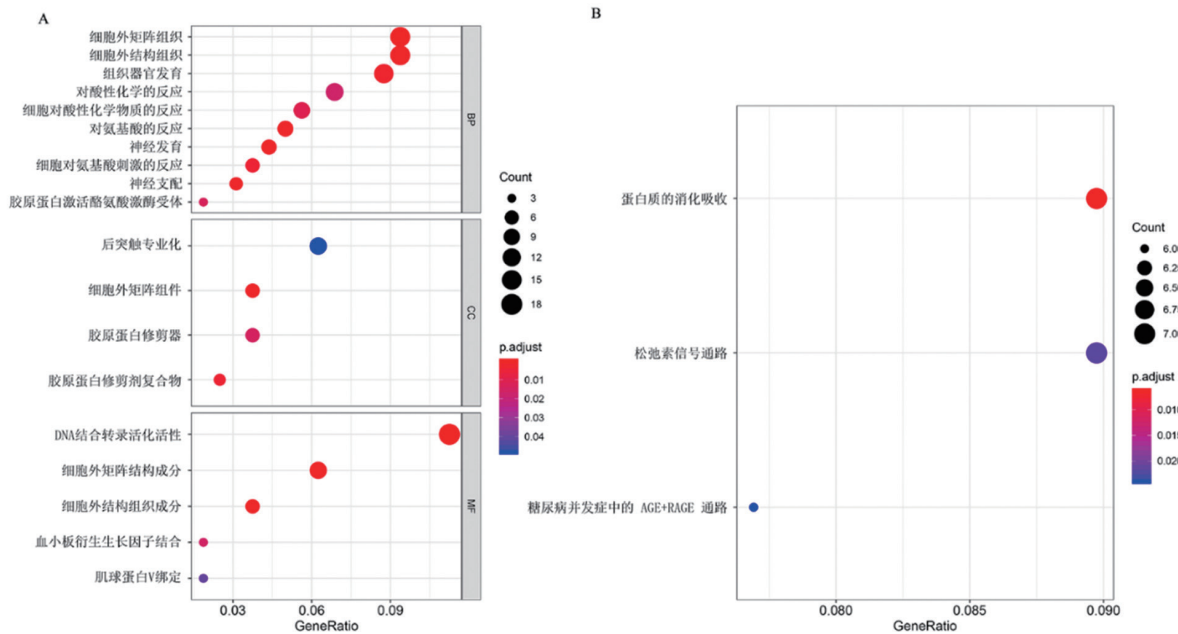
表1 处于分子蛋白互作网络中枢的前10位关键基因功能情况

No	Gene	Full name	Gene functions
1	COL1A1	collagen type I alpha 1 chain	编码 I 型胶原蛋白的亲 $\alpha 1$ 链, 存在于大多数结缔组织中, 与肺癌和食管癌有关。
2	COL3A1	collagen type III alpha 1 chain	编码 III 型胶原蛋白的亲 $\alpha 1$ 链, 基因突变与动脉瘤有关。
3	COL4A1	collagen type IV alpha 1 chain	编码 IV 型胶原蛋白 α 蛋白, 异常基因表达与大脑、肾脏和肌肉缺陷有关。
4	COL10A1	collagen type X alpha 1 chain	编码 X 型胶原蛋白的 α 链, 异常表达可能与肺癌有关。
5	COL9A1	collagen type IX alpha 1 chain	编码 IX 型胶原蛋白的三个 α 链之一, 存在两个可变单体。
6	COL4A5	collagen type IV alpha 5 chain	编码 IV 型胶原蛋白的六个亚单位之一, 遗传变异与遗传性肾炎有关。
7	SST	somatostatin	属于激素生长激素, 影响正常细胞和肿瘤细胞的增殖。
8	ADRA2C	adrenoceptor alpha 2C	G 蛋白偶联受体超级家庭的成员, 主要负责调解神经递质传播。
9	ADCY6	adenylate cyclase 6	编码腺素环化酶系列蛋白质的成员, 这是合成环性 AMP 所需的。
10	TGFBI	transforming growth factor beta induced	编码含有 RGD 的蛋白质, 抑制细胞黏附, 促进细胞增殖和迁移。

注: 本表注释主要参考 NCBI (<https://www.ncbi.nlm.nih.gov/gene>) 中较明确的结论。

2.4 功能富集注释 采用“clusterProfiler”包的 GO 富集分析和 KEGG 通路注释结果见图 3。富集分析表明共有差异基因主要涉及细胞外组织、胚胎器官

发育、突触后特化、胶原三聚体和 DNA 结合转录激活等过程。KEGG 通路分析表明这些差异基因主要参与蛋白质的消化吸收和松弛素信号通路行为。



A: GO 富集气泡图: 横轴为基因占比, 纵轴为富集到的条目描述信息。点的颜色对应 p.adjust 的值, 而点大小对应该条目下的差异基因个数。
B: KEGG 通路气泡图: 横轴为基因占比, 纵轴为富集的通路描述信息, 点的颜色对应 p.adjust 的值, 而点的大小对应通路下的差异基因个数。

图 3 差异表达基因的功能富集分析图

3 讨论

卵巢癌我国的死亡率居女性恶性肿瘤之首, 但相关致病分子机制尚未完全阐明。需要指出, 乳腺癌易感基因 (breast Cancer susceptibility gene, BRCA) 是早已被明确鉴定的卵巢癌关键基因之一, 但卵巢癌 BRCA 变异频率介于 3%~27%, 我国卵巢癌患者 BRCA 变异率为 28.45%。BRCA 基因检测为卵巢癌预防、诊断和治疗提供了精准建议。但仍需要鉴定新型关键分子进行联合筛查剩余大部分病例以此提高诊断率, 这正是本研究的出发点。

综合目前的研究^[11, 16-17], 除了致病基因外, miRNA 等非编码 RNA 表达失调也能够广泛介导各种类型的恶性肿瘤。本研究为了提高预测准确度, 整合了卵巢癌 miRNA-mRNA 表达谱来鉴定共有差异表达基因。根据 miRNA 与靶 mRNA 互补结合, 负向调控基因的原理, 我们构建了 miRNA- 基因调控网络并鉴定出 3 个具有中枢地位的抑癌关键 miRNA: miR-29c-3p, miR-1271-5p 和 miR-133b。同时通过 cytoscape 可视化网络中的共有差异基因, 预测筛选出 10 个关键基因, 即 COL1A1, COL3A1, COL4A1, COL10A1, COL9A1, COL4A5, SST, ADRA2C, ADCY6 和 TGFBI。查询 NCBI 数据库发现这 10 个基因绝大多数都参与了某些肿瘤发生。我们通过 R 包注释了共有差异基因涉及的生物过程, 发现 11.3% (18/160) 的共有

差异基因跟 DNA 转录激活功能相关 (见图 3A)。很明显可以合理推论出这些差异 miRNA 和差异基因的异常表达很可能介导了卵巢癌的发生和发展。针对挖掘到的充当抑癌角色的三个关键 miRNA: miR-29c-3p, miR-1271-5p 和 miR-133b, 乃是本文最重要的发现之一。检索 miRbase 和 genecards 等权威数据库发现 miR-29c-3p 和 miR-1271-5p 的分子功能均包括细胞群增殖负调控、细胞迁移负调控和基因表达负调控; 而 miR-133b 通过与 mRNA 3' - UTR 结合, 参与转录后基因沉默。这些专业数据库得到的证据反映了本研究生物信息挖掘的科学性, 也揭示了本文挖掘到的关键标志物调控卵巢癌的致病分子机制。同时结合近年来发表的经实验验证类的相关研究发现, HU Zhenhua 等^[18]人通过体外裸鼠移植模型证实过表达的 miR-29c-3p 通过下调 FOXP1/ATG14 通路抑制自噬, 表明 miR-29c-3p 是卵巢癌的新型靶标。WANG Lin 等^[19]人通过 qRT-PCR 验证证实 miR-1271-5p 在卵巢癌组表达显著降低; DU 等^[20]人利用 miR-1271-5p 模拟物和抑制剂进行过表达和敲除实验, 并通过双重荧光素酶活性分析证明了 miR-1271-5p 直接靶向 SPIN1 抑制乳腺癌的增殖和发展。LIU Xiaoqin 等^[21]人发现 miR-133b 通过靶向表皮生长因子受体介导 Akt 和 Erk1/2 失活抑制卵巢癌细胞的增殖和侵袭。这些涉及到三个关键 miRNA 分子的研究结论都与我们的研究中揭

示的关键 miRNA 表达情况和靶标预测结果相符。

综上所述,结合已发表的研究,显示出本研究通过整合 miRNA-mRNA 数据集挖掘卵巢癌疾病关键分子的可靠性。当然这些潜在关键分子需要更深入的体外实验鉴定其所涉及的相互作用靶标和调控通路。虽然如此,表达谱的生物信息学预测还是提供了很好的方法,缩小了体外实验研究的范围,节省了宝贵的资源。未来,我们相信研究人员利用计算生物学实施肿瘤大数据战略,或许能够一次性揭示众多复杂多样的肿瘤发生的关键分子。

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-A Cancer Journal for Clinicians, 2018, 68(6): 394-424.
- [2] NASRIN R, MOHAMMAD H. Current advancement on diagnosis of ovarian cancer using biosensing of CA 125 biomarker: Analytical approaches(Review)[J]. TrAC - Trends in Analytical Chemistry, 2018, 108:1-12.
- [3] 马同敏, 赵志强. 血清 CA125, STIP1 和 IGF- I 联合检测对卵巢癌的早期诊断价值研究 [J]. 现代检验医学杂志, 2018, 33(6): 50-54, 58.
MA Tongmin, ZHAO Zhiqiang. Clinical value of combined detection of serum CA125, STIP1 and IGF-I levels in early diagnosis of ovarian cancer [J]. Journal of Modern Laboratory Medicine, 2018, 33(6): 50-54, 58.
- [4] 付妮娜, 范婧晖, 钟慧. CA125, HE4 和 MMP7 联合检测在 I 型和 II 型上皮性卵巢癌患者诊断中的价值 [J]. 现代检验医学杂志, 2019, 34(6): 109-112, 156.
FU Nina, FAN Jinghui, ZHONG Hui. Diagnostic value of combined detection of CA125, HE4 and MMP7 in type I and II epithelial ovarian cancer [J]. Journal of Modern Laboratory Medicine, 2019, 34(6): 109-112, 156.
- [5] 吕博文, 钱钧, 王杰, 等. miR-455-5 p 在上皮性卵巢癌中表达研究及其靶基因功能分析 [J]. 实用肿瘤学杂志, 2019, 33(2): 115-121.
LÜ Bowen, QIAN Jun, WANG Jie, et al. Expression of miR-455-5p in epithelial ovarian cancer and its target gene function [J]. Practical Oncology Journal, 2019, 33(2): 115-121.
- [6] 郭世民, 李红梅. 利用生物信息学发现基因 HADH 与 SRC 为卵巢癌潜在治疗靶基因 [J]. 临床医药文献电子杂志, 2017, 4(97): 19011-19012, 19028.
GUO Shimin, LI Hongmei. Identification of HADH and SRC as prognostic biomarkers for ovarian cancer based on bioinformatics approaches [J]. Journal of Clinical Medical Literature (Electronic Edition), 2017, 4(97): 19011-19012, 19028.
- [7] YANG X, ZHU S, LI L, et al. Identification of differentially expressed genes and signaling pathways in ovarian cancer by integrated bioinformatics analysis(Article)[J]. Onco Targets and Therapy, 2018, 11:1457-1474.
- [8] FENG Hao, GU Zhongyi, LI Qin, et al. Identification of significant genes with poor prognosis in ovarian cancer via bioinformatical analysis(Article)[J]. Journal of Ovarian Research, 2019, 12(1):35.
- [9] DALMAY T. Mechanism of miRNA-mediated repression of mRNA translation.[J]. Essays in Biochemistry, 2013, 54 : 29-38.
- [10] LUJAMBIO A, LOWE S W. The microcosmos of cancer[J]. Nature, 2012, 482(7385): 347-355.
- [11] RUPAIMOOLE R, CALIN G A, LOPEZ-BERESTEIN G, et al. miRNA deregulation in cancer cells and the tumor microenvironment[J]. Cancer Discovery, 2016, 6(3): 235-246.
- [12] SMYTH G K. LIMMA: Linear Models for Microarray Data [M]. GENTLEMAN R, CAREY V J, HUBER W, et al, Bioinformatics and Computational Biology Solutions Using R and Bioconductor, New York: Springer, 2005: 397-420.
- [13] YU Guangchuang, WANG Ligen, HAN Yanyan, et al. clusterProfiler: an R package for comparing biological themes among gene clusters[J]. OMICS-A Journal of Integrative Biology, 2012, 16(5): 284-287.
- [14] PATHAN M, KEERTHIKUMAR S, CHISANGA D, et al. A novel community driven software for functional enrichment analysis of extracellular vesicles data[J]. Journal of Extracellular Vesicles, 2017, 6(1): 1321455.
- [15] CHIN C H, CHEN S H, WU H H, et al. cytoHubba: identifying hub objects and sub-networks from Complex interactome[J]. BMC Systems Biology, 2014, 8(Suppl 4): S11.
- [16] REDDY K B. MicroRNA (miRNA) in cancer[J]. Cancer Cell International, 2015, 15(1): 38.
- [17] TANG Wei, WAN Shixiang, YANG Zhen, et al. Tumor origin detection with tissue-specific miRNA and DNA methylation markers[J]. Bioinformatics (Oxford, England), 2018, 34(3): 398-406.
- [18] HU Zhenhua, CAI Mingbo, ZHANG Ying, et al. miR-29c-3p inhibits autophagy and cisplatin resistance in ovarian cancer by regulating FOXF1/ATG14 pathway[J]. Cell Cycle, 2020, 19(2):193-206.
- [19] WANG Lin, ZHU Miaojun, REN Aimin, et al. A ten-microRNA signature identified from a genome-wide microRNA expression profiling in human epithelial ovarian cancer[J]. PLoS One, 2014, 9(5): e96472.
- [20] DU H Y, LIU B. MiR-1271 as a tumor suppressor in breast cancer proliferation and progression via targeting SPIN1[J]. European Review for Medical and Pharmacological Sciences, 2018, 22(9): 2697-2706.
- [21] LIU Xiaoqin, LI Guohong. MicroRNA-133b inhibits proliferation and invasion of ovarian cancer cells through Akt and Erk1/2 inactivation by targeting epidermal growth factor receptor[J]. International Journal of Clinical and Experimental Pathology, 2015, 8(9): 10605-10614.

收稿日期: 2021-02-22

修回日期: 2021-04-16