DOI:10.3872/j.issn.1007-385x.2019.02.008

·基础研究·

肺腺癌相关基因的生物信息学分析

高强,钟英英,丁华杰,叶云(广西科技大学 生物与化学工程学院,广西 柳州 545006)

[摘 要] **目的**:通过生物信息学分析基因表达谱,获取肺腺癌相关基因及信号通路。**方法**:从GEO数据库下载GSE40791、GSE68571、GSE43458和GSE18842表达数据,将4个微阵列数据集整合获得肺腺癌相关差异表达基因,利用STRING数据库为差异表达基因构建肺腺癌蛋白-蛋白互相作用网络,并挖掘肺腺癌网络中基因模块及关键基因。通过DAVID对各基因模块进行基因富集分析,发掘基因模块在肺腺癌细胞中所执行的调控功能及模块中关键基因与患者的预后关系。**结果**:初步筛查获得肺腺癌相关37个上调基因、120个下调基因,并成功构建蛋白-蛋白相互作用网络,通过MCODE算法在蛋白-蛋白相互作用网络中构建基因模块以及计算关键基因(KIF14,SEPP1,SPP1,RBP4),最终获得的4个模块分别参与细胞周期、血凝变化、细胞黏附和细胞代谢的调控。经验证4个关键基因在肺腺癌和正常组织中有明显表达差异(P<0.05)。生存分析显示KIF14的表达对肺腺癌的预后有显著影响(P<0.01),SEPP1、SPP1对患者生存率有显著影响(P<0.05),RBP4对患者的生存率影响无统计学意义(P>0.05)。**结论**:通过生物信息方法分析肺腺癌和癌旁正常组织的差异基因表达,最终筛选出3个差异表达非常显著且对患者预后影响明显的基因,对肺腺癌的诊断和预后治疗提供了新思路,提高肺腺癌机制的研究效率。

[关键词] 肺腺癌;基因表达谱;差异基因
[中图分类号] R730.5;R734 [文献标识码] A [文章编号] 1007-385X(2019)02-0190-06

Bioinformatic analysis on related genes of lung adenocarcinoma

GAO Qiang, ZHONG Yingying, DING Huajie, YE Yun(College of Biological and Chemical Engineering, Guangxi University of Science and Technology, Liuzhou 545006, Guangxi, China)

[Abstract] Objective: To indentify the candidate genes and signaling pathways in lung adenocarcinoma by analyzing gene profiles with bioinformatics. **Methods:** The expression profiles of GSE40791, GSE68571, GSE43458, and GSE18842 were down-loaded from the Gene Expression Omnibus (GEO) database. The four microarray datasets were integrated to obtain the differentially expressed genes related to lung adenocarcinoma. STRING database was used to construct the protein-protein interaction (PPI) network of differentially expressed genes, and to further explore the gene modules and the key genes. DAVID was used to perform the gene enrichment analysis of each gene module, and to explore the regulatory function of each gene module in adenocarcinoma cells, as well as the relationship between the key genes in the module and the prognosis of the patients. **Results:** Thirty-seven up-regulated genes and 120 down-regulated genes were obtained from the primary screen, and the protein-protein interaction(PPI) network was successfully constructed. According to MCODE algorithm, we constructed gene modules and calculated the core genes (KIF14, SEPP1, SPP1, RBP4) in the PPI network. Finally, four modules were proved to be involved in regulation of cell cycle, blood coagulation, cell adhesion and cell metabolism, and four key genes were proved to be differentially expressed between lung adenocarcinoma tissues and normal tissues (all P<0.05). Survival analysis showed that expressions of KIF14, SEPP1 and SPP1 had significant effect on the prognosis of lung adenocarcinoma (P<0.01 or P<0.05), while RBP4 exerted insignificant difference in the survival rate of lung adenocarcinoma patients (P>0.05). **Conclusion:** With bioinformatics, three differentially expressed genes between lung adenocarcinoma tissues and normal adjacent tissues were finally screened out and proved to be closely related to the prognosis of patients, which provided new thoughts in the diagnosis and prognosis prediction of lung adenocarcinoma and improved the study efficiency on the mechanism of lung

[基金项目] 广西自然科学基金资助项目(No.2017GXNSFAA198325);广西高校中青年教师基础能力提升资助项目(No.2017KY353);2017年度广西糖资源与加工重点实验室开放课题资助项目(No.2016TZYKF06,No.GXTZY201704);广西科技大学硕士生创新资助项目(No.GKYC201718)。Project supported by the Guangxi Natural Science Foundation (No.2017GXNSFAA198325), the Project of Improving Basic Ability of Young and Middle-aged Scholars in Guangxi Colleges (No. 2017KY353), the Key Laboratory for Processing of Sugar Resources of Guangxi (No.2016TZYKF06,No.GXTZY201704),and the Graduate Innovation Program of Guangxi University of Science and Technology (No. GKYC201718)
[作者简介] 高强(1992-),男,硕士生,主要从事肿瘤相关基因的生物信息学研究,E-mail: 1192242515@qq.com
[通信作者] 叶云(YE Yun, corresponding author), 博士,教授,硕士生导师,主要从事肿瘤相关基因的生物信息学研究,E-mail: yunyeg@gxust.edu

adenocarcinoma.

[Key words] lung adenocarcinoma; gene expression profile; differentially expressed genes

[Chin J Cancer Biother, 2019, 26(2): 190-195. DOI：10.3872/j.issn.1007-385X.2019.02.008]

肺腺癌（lung adenocarcinoma）属于非小细胞肺癌，是世界上最常见的癌症之一，在一些地区肺腺癌发病率在肺癌中居于首位，并且肺腺癌发病率及病死率极高[1-2]。肺腺癌的前期诊断以及预后预测均较困难，一些治疗方案对患者生存率的提高也并不明显，所以深入对肺腺癌机制的研究非常重要[3]。基因芯片是一种高通量和系统性的研究技术，能检测和分析不同组织的差异表达基因。随着实验技术的不断革新和发展，生物数据的急剧膨胀，以基因组和功能基因为主要研究对象的生物信息学迅速发展，尤其是应用生物信息学方法发现新基因并对其展开功能研究[4-5]。目前，基因表达谱芯片在肿瘤发生机制、早期诊断、肿瘤基因分型、指导治疗、评估预后等领域得到了普遍的应用[6-7]。本研究采用多种生物信息学方法从GEO数据库提取肺腺癌基因表达数据，通过差异分析筛选出相关的潜在基因，并结合生存时间以及生存状态进行生存期分析，探讨相关基因的表达与肺腺癌的关系，为进一步研究肺腺癌发生发展的作用机制提供线索和依据。

1 材料与方法

1.1 数据获取与基因筛选

GEO数据库(https://www.ncbi.nlm.nih.gov/)下载Affymetrix基因表达谱数据(GSE40791, GSE18842, GSE43458,GSE68571)，详细芯片平台及样本分类见表1。

利用基于R语言分析程序GEO2R，分别对4个基因表达谱数据进行分析，经过*t*检验的方法，定义校正后$P<0.05$和$\log_2FC>1$的差异表达基因有统计意义。利用Venny 2.1.0(http://bioinfogp. cnb. csic. es / tools/venny/)分析差异基因，获得共同上、下调差异基因。

表1 GEO数据信息(*n*)

Tab.1 GEO data information (*n*)

Data set	Platform	Normal	Tumor	Sample source
GSE40791	GPL570	100	94	Lung tissue
GSE18842	GPL570	45	46	Lung tissue
GSE43458	GPL6244	30	80	Lung tissue
GSE68571	GPL80	10	86	Lung tissue

1.2 差异基因编码蛋白间相互作用的分析

利用STRING（https://string-db.org/cgi/input.pl）对差异基因编码蛋白进行相互作用分析。基于已经建立的差异基因蛋白-蛋白相互作用网络，通过Cytoscape 3.6.1中MCODE插件基于K-Core算法发掘肺腺癌蛋白-蛋白相互作用网络中不同功能的基因模块，将这些复杂的小模块从整个网络中抽提出来，再单独进行功能分析。本研究选取MCODE分数大于3.0（含）的模块进行后续研究，即MCODE通过基于密度的非交叠式聚类算法，对网络图中的各个节点信息进行计算，由种子节点进行扩展，逐步加入其邻居节点中符合要求的节点，从而形成一个功能模块，而后将此种子节点位置基因作为核心基因进行后续研究[8]。

1.3 基因富集分析

对每一模块分别进行富集分析。利用DAVID (https://david.ncifcrf.gov/home. jsp) 对每一模块差异基因进行功能注释，包括分子功能、细胞组成与生物过程的GO功能富集分析，以及KEGG通路富集分析，定义为$P<0.05$。并通过差异基因的GO分析及KEGG分析，对差异基因的GO条目以及KEGG条目进行分类，寻找不同的差异基因可能和哪些基因功能和细胞信号通路的改变有关[9]。

1.4 肺腺癌芯片分析关键基因的差异表达

GEPIA（http://gepia.cancer-pku.cn/）是由北京大学开发的基因表达谱动态数据分析数据库，通过此数据库发掘肺腺癌关键基因在肺腺癌与正常肺组织中的差异表达，筛选条件为LUAD数据集，定义为$|Log2FC|\geq1$、$P<0.01$。

1.5 通过GEPIA数据库对患者生存期进行分析

通过GEPIA数据库对肺腺癌数据进行在线生存分析，筛选条件为LUAD数据集，95%置信区间，时间轴单位为月。基因表达差异采用*t*检验，在肺腺癌中表达量与预后的关系采用Log-rank检验，以Logrank

P<0.05或P<0.01表示差异有统计学意义。

2 结　果

2.1 初步筛查到的肺腺癌中的差异表达基因

对4个数据样本进行分析，分别获得差异倍数在2倍以上的上、下调差异基因，通过Venny 2.1.0分别对4组数据的上调基因和下调基因进行分析，获得4组数据：共同上调基因37个，共同下调基因120个。结果见图1。

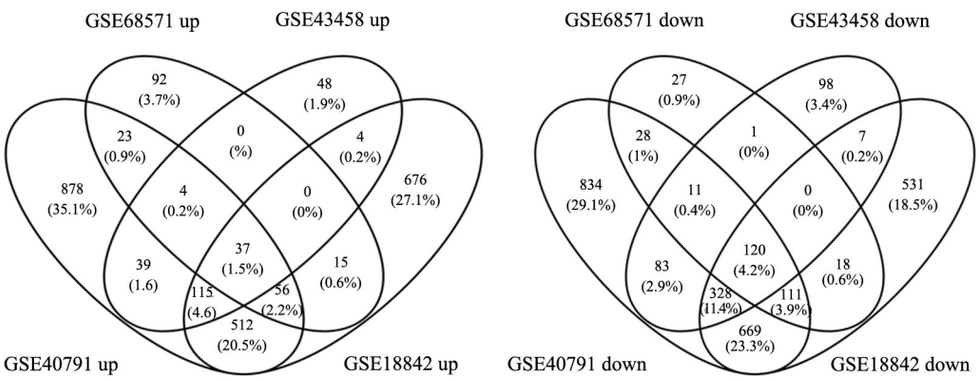


图1 差异表达基因韦恩图

Fig.1 Venn diagrams of the differentially expressed genes identified in four datasets

2.2 肺腺癌相关差异表达基因编码蛋白-蛋白相互作用网络

STRING数据库收录了大量已知蛋白-蛋白相互作用关系，利用该数据库对差异基因编码蛋白构建蛋白-蛋白相互作用进行分析，结果（图2）显示，部分蛋白相对集中，即这些蛋白相互作用主要集中在与细胞相关的功能与通路，有助于对细胞的功能进行系统层面的理解。采用Cytoscape 3.6.1中MCODE插件分析互作网络，定义网络中的核心模块，共筛选出4个主要基因模块：模块A包含14个基因91个相互作用关系（MCODE score=14.0），模块B包含9个基因36个相互作用关系（MCODE score=9.0），模块C包含6个基因10个互作关系（MCODE score=4.0），模块D包含9个基因16个相互作用关系（MCODE score=4.0），结果见图3。根据MCODE算法构建的基因模块中，KIF14，SEPP1，SPP1，RBP4分别为扩展出基因模块的种子节点位置基因，即为关键基因，模块A中关键基因KIF14，模块B中关键基因SEPP1，模块C中关键基因SPP1，模块D中关键基因RBP4。通过分析基因模块内部的功能关系，预测潜在靶标、解释作用机理，有助于加快药物研究的进展。

2.3 肺腺癌相关基因参与的调控作用

DAVID富集分析结果（表3）显示，模块A主要参与细胞周期的调控；模块B主要参与血凝变化的调控；模块C主要参与细胞黏附的调控；模块D主要参与细胞代谢途径。

2.4 肺腺癌中4种关键基因的表达差异

通过肺腺癌患者与正常肺组织样本比较KIF14、SEPP1、SPP1、RBP4基因的表达差异，相对于正常组织KIF14与SPP1在肺腺癌组织中明显高表达，而相对于正常组织SEPP1和RBP4在肺腺癌组织中明显低表达（图4）。未来这些基因能够成为治疗靶点，可以通过逆转基因的异常表达来实现。

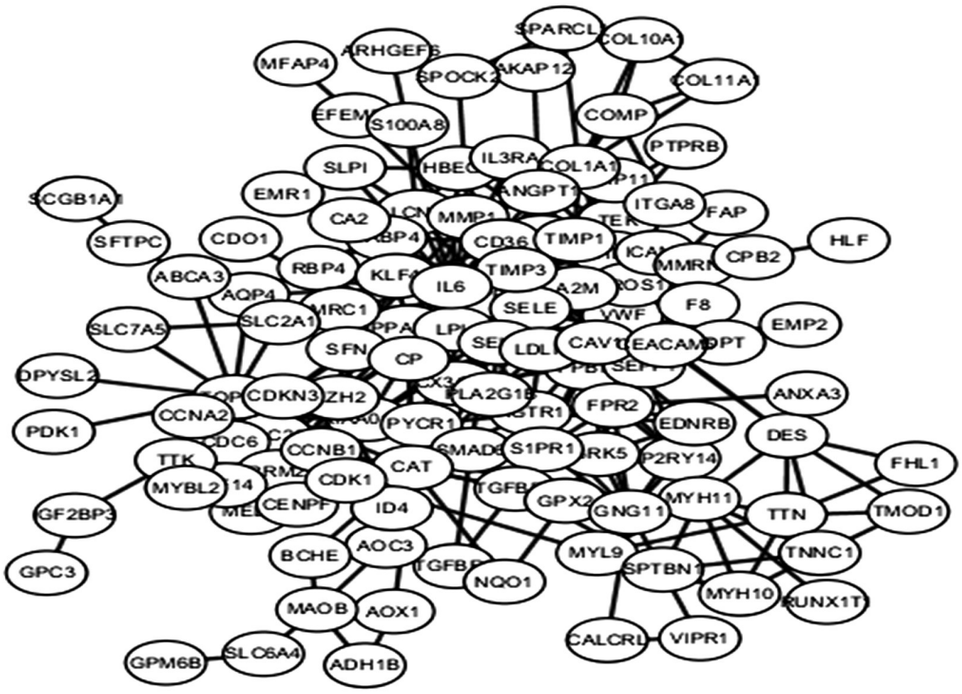
2.5 肺腺癌中4种关键基因与患者的预后关系

KIF14表达水平对患者的总生存时间有着显著影响(P<0.01)。而SEPP1、SPP1对患者生存率也有明显影响(P<0.05)，RBP4对患者的生存率影响无统计学意义。结果见图5。

3 讨　论

对于包括肺腺癌在内的癌症治疗方法在不断更新，并且取得了很大的进展，但是预后仍然不太理想[10]；且大部分已知的抗肿瘤药物除了对肿瘤细胞具有抑制作用外，对正常细胞的毒副反应也较大[11]。因此，对于肺腺癌发生发展机制的研究尤为重要，既可以帮助深入了解肺腺癌的发病机制，又可以协助抗肿瘤药物的研发。本研究通过生物信息学分析获得了肺腺癌组织与正常肺组织的差异表达基因，并建立蛋白-蛋白互相作用网络，继而筛选蛋白网络中的基因模块，发掘在肺腺癌中发挥重要调控作用的关

键基因靶点。通过肺腺癌基因富集分析结果显示，在4个模块中，模块A主要参与细胞周期的调控，模块A中KIF14的高表达会影响细胞分裂，促进肿瘤细胞的分裂增殖[12]，有报道[13-17]，在乳腺癌、肝癌、胃癌、卵巢癌、神经胶质瘤等多种肿瘤细胞中KIF14基因高表达，并促进肿瘤细胞的异常增殖。而KIF14对肺腺癌发生发展及其对患者预后的影响却无明确报道，甚至有文献报道[18]，KIF14及其编码蛋白在肺腺癌中或具有促进与抑制癌细胞增殖的双重作用。因此KIF14基因及编码蛋白可作为今后肺腺癌实验验证研究的关注点。模块B中SEPP1编码蛋白是血浆硒的主要储存形式，包含几乎50%的总血浆硒，且在部分人群中SEEP1水平较低时患肺癌的危险增加[19]。但是在不同地区SEPP1的影响结果也不同，有可能成为肺癌潜在的发生标志物。模块C中关键基因SPP1已有明确报道其基因编码蛋白含有的GRGDS序列与细胞的黏附作用有关，对肿瘤细胞的转移侵袭产生影响[20]，进一步验证了本研究的结果。模块D中，RBP4可能对2型糖尿病的发生起作用，参与异常的糖代谢，而能量代谢的重构是肿瘤的标志之一。尽管目前还未见RBP4参与肺腺癌糖代谢异常的有关报道，但RBP4也可能成为肺腺癌的治疗靶点，值得进一步深入探究。



Circle indicates protein; the line between circles indicates protein-protein interaction;the more the lines between proteins, the stronger the interaction between proteins

图2 肺腺癌差异基因编码蛋白-蛋白相互作用网络

Fig.2 The PPI interaction network of differentially expressed genes in adenocarcinoma

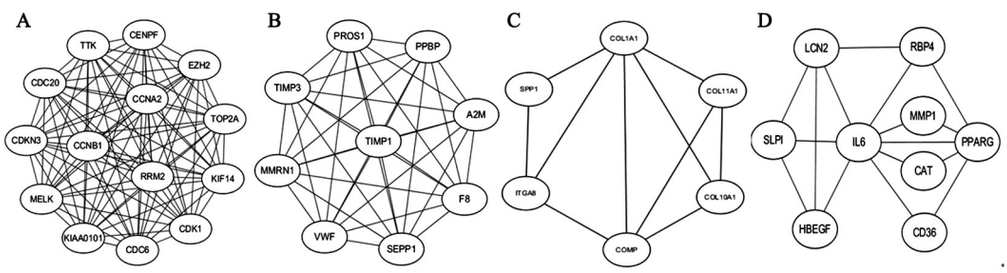


Figure 3 shows the gene modules constructed by MCODE algorithm. The key gene is the seed node gene that extends the gene module; KIF14, SEPP1, SPP1 and RBP4 are the key genes in gene module A, module B, module C and module D, respectively

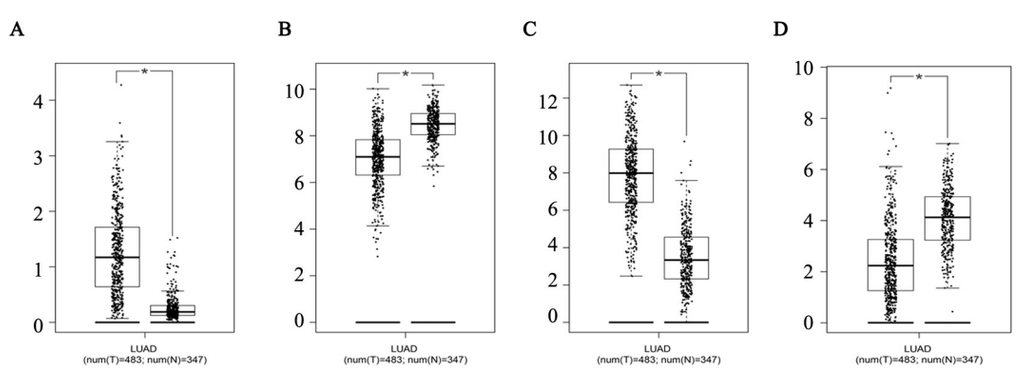
图3 蛋白-蛋白相互作用网络中的基因模块

Fig.3 Gene modules in PPI network

表3 蛋白-蛋白相互作用网络基因富集分析

Tab.3 Gene enrichment analysis of PPI network

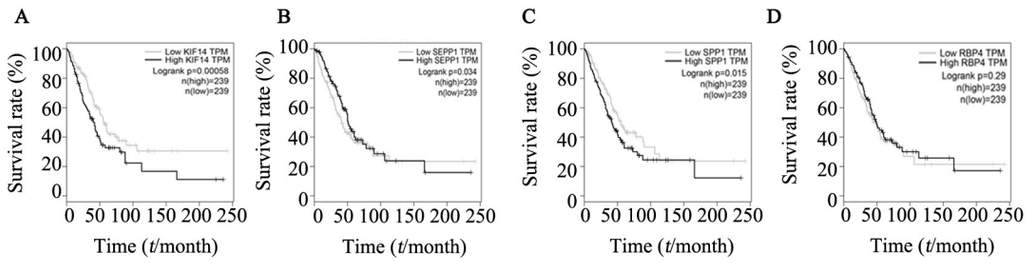
Rank	Category	Pathway	Gene	*P*
Module A	KEGG	Cell cycle	6	0.01
		P53 signaling pathway	3	0.01
	GO	ATP binding	6	0.01
		Regulation of G2/M transition of mitotic cell cycle	2	0.01
		Spindle midzone	2	0.01
		Positive regulation of cytokinesis	2	0.05
		Nucleus	6	0.05
		Cell division	2	0.05
Module B	KEGG	Complement and coagulation cascades	4	0.01
	GO	Blood coagulation	3	0.01
		Extracellular exosome	5	0.01
		Platelet activation	2	0.01
		Basement membrane	2	0.01
		Blood microparticle	2	0.05
Module C	KEGG	ECM-receptor interaction	4	0.01
		Focal adhesion	4	0.01
		PI3K-Akt signaling pathway	4	0.01
	GO	Proteinaceous extracellular matrix	3	0.01
		Extracellular matrix structural constituent	2	0.01
		Cell adhesion	2	0.05
Module D	KEGG	PPAR signaling pathway	3	0.01
		Pathways in cancer	3	0.05
	GO	Glucose homeostasis	3	0.01
		Extracellular space	5	0.01
		Response to retinoic acid	2	0.01



KIF14(A) and SPP1(C) are highly expressed while SEPP1(B) and RBP4(D) are low expressed in lung adenocarcinoma tissues compared with normal tissues（all $P<0.05$）

图4 肺腺癌和正常肺组织中KIF14(A), SEPP1(B), SPP1(C), RBP4(D)表达的差异

Fig.4 Differential expression of KIF14(A), SEPP1(B), SPP1(C), RBP4(D) in lung adenocarcinoma and normal tissues



The expression of KIF14 has significant effect on prognosis of lung adenocarcinoma (Logrank *P*=0.00058). SEPP1 and SPP1 have significant impact on patients' survival (Logrank *P*=0.034, Logrank *P*=0.015), and there was no statistically significant effect of RBP4 on patients' survival (Logrank *P*=0.29)

图5 肺腺癌中KIF14(A)、SEPP1(B)、SPP1(C)、RBP4(D)表达与患者预后的生存曲线

Fig. 5 KIF14(A), SEPP1(B), SPP1(C), RBP4(D) expression and prognosis survival curve of lung adenocarcinoma

随着高通量技术的高速发展，可获得大量与肺腺癌相关的研究数据，本研究利用现有的公共数据库和分析工具整合了多组肺腺癌研究数据并进行统一处理，从基因层面分析肺腺癌的相关分子机制，得到一些可靠的结果，为今后肺腺癌发生发展机制和治疗的深入研究提供有价值的信息，为新药的研发提供新的研究思路与实验切入点。

[参 考 文 献]

[1] TORRE L A, BRAY F, SIEGEL R L, et al. Global cancer statistics, 2012 [J]. Ca A Cancer J Clin, 2015, 65(2): 87-108. DOI：10.3322/caac.21262.

[2] 周宝森. 女性肺腺癌危险因素分析[J]. 中国公共卫生, 2000, 16(6): 536-539. DOI: 10.11847/zgggws2000-16-06-49.

[3] 全斌, 喻艳林. 肺结核合并肺癌的发生机制研究进展[J]. 山东医药, 2015, 55(24): 104-106. DOI: 10.3969/j.issn.1002-266X.2015.24.047.

[4] GU C, SHEN T. cDNA microarray and bioinformatic analysis for the identification of key genes in Alzheimer's disease[J]. Int J Mol Med, 2014, 33(2): 457-461. DOI: 10.3892/ijmm.2013.1575.

[5] 李瑶. 基因芯片数据分析与处理 [M]. 北京：化学工业出版社, 2006: 7-9.

[6] LUZZI V I, HOLTSCHLAG V, WATSON M A. Gene expression profiling of primary tumor cell populations using laser capture microdissection, rna transcript amplification, and genechip® microarrays[J]. Methods Mol Biol, 2005, 293: 187-207. DOI：10.1385/1-59259 -853-6:187.

[7] 李燕妮, 齐士勇, 颜艳, 等. 基于Keap1基因多态性的肾透明细胞癌分子标志物的研究[J]. 中国肿瘤生物治疗杂志, 2017, 24(3): 278-283. DOI:10.3872/j.issn.1007-385X.2017.03. 011.

[8] RADICCHI F, CASTELLANO C, CECCONI F, et al. Defining and identifying communities in networks[J]. P NATL ACAD SCI USA, 2004, 101(9): 2658-2663. DOI: 10.1073/pnas. 0400054101.

[9] 张健, 王冬, 于景翠. 胃癌中miRNA功能相关的信号通路[J]. 中国肿瘤生物治疗杂志, 2017, 24(11): 1331-1335. DOI: 10.3872/j.issn.1007-385X.2017.11.016.

[10] MASTERS G A, JOHNSON D H, TEMIN S. Systemic therapy for stage IV non-small-cell lung cancer: american society of clinical oncology clinical practice guideline update[J]. J Oncol Pract, 2017, 33 (30): 832-837. DOI: 10.1200/JCO.2015.62.1342.

[11] FU Q, YANG F, ZHAO J, et al. Bioinformatical identification of key pathways and genes in human hepatocellular carcinoma after CSN5 depletion[J]. Cell Signal, 2018, 49: 79-86. DOI: 10.1016/j.cellsig.2018.06.002.

[12] HANAHAN D, WEINBERG R A. Hallmarks of cancer: the next generation[J]. Cell, 2011, 144(5): 646-676. DOI: 10.1016/j. cell. 2011.02.013.

[13] WANG Q, WANG L, LI D, et al. Kinesin family member 14 is a candidate prognostic marker for outcome of glioma patients[J]. Cancer Epidemiol, 2013, 37(1): 79-84. DOI: 10.1016/j.canep. 2012.08.011.

[14] SHAH S P, ROTH A, GOYA R, et al. The clonal and mutational evolution spectrum of primary triple negative breast cancers[J]. Nature, 2012, 486(7403): 395-399. DOI: 10.1038/nature10933.

[15] YANG T, ZHANG X B, ZHENG Z M. Suppression of KIF14 expression inhibits hepatocellular carcinoma progression and predicts favorable outcome[J]. Cancer Sci, 2013, 104(5): 552-557. DOI: 10.1111/cas.12128.

[16] 陈宗营, 马恒. KIF14在胃癌中的表达及其意义[J]. 中国现代普通外科进展, 2011, 14(12): 941-944. DOI: 10.3969/j. issn. 1009-9905.2011.12.005.

[17] TH RIAULT B L, PAJOVIC S, BERNARDINI M Q, et al. Kinesin family member 14: an independent prognostic marker and potential therapeutic target for ovarian cancer[J]. INT J Cancer, 2012, 130(8): 1844-1854. DOI: 10.1002/ijc.26189.

[18] HUNG P F, HONG T M, HSU Y C, et al. The motor protein kif14 inhibits tumor growth and cancer metastasis in lung adenocarcinoma [J]. PLoS One, 2013, 8(4): e61664[2018-12-23]. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061664. DOI：10. 1371/ journal.pone.0061664.

[19] EPPLEIN M, BURK R F, CAI Q, et al. A prospective study of plasma selenoprotein p and lung cancer risk among low-income adults [J]. Cancer Epidemiol Biomarkers Prev, 2014, 23(7): 1238-1244. DOI: 10.1158/1055-9965.EPI-13-1308.

[20] OLDBERG A, FRANZ N A, HEINEG RD D. Cloning and sequence analysis of rat bone sialoprotein (osteopontin) cDNA reveals an ARG-Gly-Asp cell-binding sequence[J]. Proc Natl Acad Sci USA, 1986, 83(23): 8819-8823. DOI: 10.1073/pnas.83.23.8819.

[收稿日期] 2018-12-17 [修回日期] 2019-01-12
[本文编辑] 阮芳铭