论著·临床研究

应用生物信息学方法筛选结直肠癌关键基因*

朱义芳[1]，戴红梅[2]，张峪涵[3]，魏丹凤[4]，潘逼然[4]，刘　蕾[4]，张彤彤[4]，郭元彪[4]，刘华伟[2△]
（1. 四川省骨科医院检验科，四川成都 610041；重庆医科大学附属成都第二临床学院/成都市第三人民医院：2. 检验科；3. 神经内科；4. 实验医学研究部，四川成都 610031）

摘　要：目的　采用 4 个芯片数据集来挖掘结直肠癌差异基因，通过基因注释和蛋白相互作用网络构建找到关键基因。**方法**　下载 GEO 芯片数据 GSE4398、GSE21815、GSE32323、GSE44076，筛选结直肠癌和正常组织间差异表达的基因，采用 R3.4.4 软件进行数据处理和分析，通过取交集获得候选的差异基因，通过 Funrich 软件进行基因功能分析，通过 String 和 Cytoscape 软件进行基因编码蛋白的相互作用分析。**结果**　通过差异基因分析并取 4 个 GEO 数据集的交集，一共获得 430 个差异表达基因，其中表达上调的基因 277 个，表达下调的基因 153 个。基因富集分析发现，表达上调的基因主要位于细胞核、细胞浆、微管、中心体和细胞外；主要参与纺锤体组装；主要参与调节趋化因子活性，在细胞周期、有丝分裂 G1-G1/S 期、M-M/G1 期、G2/M 期、DNA 破坏关键节点及 DNA 复制等信号通路中富集。153 个表达下调的基因主要富集在细胞外、参与代谢过程，发挥的分子功能主要是催化活性和配体依赖性核受体活性，下调基因没有富集在某条信号通路上。通过蛋白互作网络初步鉴定了 CDK1、CCNB1、TOP2A、MAD2L1、TTK、BUB1B、AURKA、RRM2、UBE2C、ASPM 等结直肠癌相关的关键基因。**结论**　通过基因芯片结合生物信息学方法，发现了结直肠癌相关的关键基因，有助于明确结直肠癌发病的分子机制，这些关键基因也可作为结直肠癌的治疗靶点。

关键词：结直肠癌；　基因芯片；　关键基因；　生物信息学
DOI：10.3969/j.issn.1673-4130.2019.14.014　　**中图法分类号：**R735.34
文章编号：1673-4130(2019)14-1721-05　　**文献标识码：**A

Screening key genes in colorectal cancer with bioinformatic methods*

ZHU Yifang[1], *DAI Hongmei*[2], *ZHANG Yuhan*[3], *WEI Danfeng*[4], *PAN Biran*[4], *LIU Lei*[4], *ZHANG Tongtong*[4], *GUO Yuanbiao*[4], *LIU Huawei*[2△]
(1. *Department of Clinical Laboratory, Sichuan Provincial Orthopedic Hospital, Chengdu, Sichuan* 610041, *China*; 2. *Department of Clinical Laboratory*; 3. *Department of Neurology*; 4. *Medical Research Center, the Second Affiliated Clinical College of Chongqing Medical University/the Third People's Hospital of Chengdu, Chengdu, Sichuan* 610031, *China*)

Abstract: Objective　To integrate four microarray sets to screen differentially expressed genes and to find key genes with gene annotation and protein-protein interaction network construction. **Methods**　Datasets were download from GSE4398, GSE21815, GSE32323, GSE44076, data processing and analysis process were done with R3.4.4 software, the differentially expressed genes overlaps were outputted, gene function analysis were completed with Funrich software, further protein-protein interaction were done with String and Cytoscape software. **Results**　A total of 430 differentially expressed genes were obtained by differential gene analysis with overlaps of four GEO datasets, of which 277 genes were up-regulated and 153 genes were down-regulated. Up-regulated genes were mainly enriched in the nucleus and cytoplasm, tubes, central body and cells; mainly involved in spindle assembly; mainly involved in regulating chemokine activity, cell cycle G1-G1/S phase and mitosis, M-M/G1 and G2/M phase, DNA damage checkpoint and DNA replication signaling pathway. The down-regulated genes were mainly concentrated in the extracellular and involved in the metabolic process, and their molecular functions were mainly catalytic activity and ligand-dependent nuclear receptor activity. The down-regulated genes were not concentrated in a certain signaling pathway. CDK1, CCNB1,

*　**基金项目：**成都市科技局科研项目(2018-YF05-00669-SN)；成都市卫健委科研项目(2018056, 2018069, 2016005)；四川省卫健委科研项目(18PJ126, 18PJ112)。

作者简介：朱义芳，女，主管技师，主要从事肿瘤基因功能研究。　△　**通信作者**，E-mail：1959519239@qq.com。

本文引用格式：朱义芳，戴红梅，张峪涵，等. 应用生物信息学方法筛选结直肠癌关键基因[J]. 国际检验医学杂志，2019，40(14)：1721-1725.

TOP2A, MAD2L1, TTK, BUB1B, AURKA, RRM2, UBE2C, ASPM were identified as key genes related to colorectal cancer by protein interaction network. **Conclusion** The key genes of colorectal cancer can be obtained by the combination of genechip and bioinformatics analysis, which can help us determining the potential molecular mechanism of colorectal cancer, These candidate genes also can be used as therapeutic targets for colorectal cancer.

Key words: colorectal cancer; gene chip; key genes; bioinformatics

结直肠癌是第 3 位常见的恶性肿瘤，在全球肿瘤相关性死亡的原因中也位列第 3 位[1]。2018 年美国结直肠癌新发病例数预计超过 14 万，死亡病例数超过 5 万[2]。结直肠癌发病机制复杂，包括多个基因、多条通路的交互作用[3]。到目前为止，结直肠癌仍是全球医疗的重大难题，仍缺乏系统的、整体的理解其发病的分子机制。传统的单基因检测的研究，虽然能发现某些基因在肿瘤形成发展中发挥的具体作用，但不能全面的挖掘出结直肠癌形成过程中更为广泛存在的多个基因和通路的改变。近年来，随着基因芯片技术在肿瘤中的广泛应用，大量的芯片数据产生，其中大部分数据被储存在公共数据库中未被挖掘。整合并重新分析这些数据可为新的研究提供线索，为全面分析结直肠发病的分子机制提供便利[4]。

本研究选取了 4 个 GEO 芯片数据集来鉴定结直肠癌中差异表达的基因，这些差异表达基因可作为潜在的结直肠癌标志物。进一步的功能富集分析可阐明这些差异表达基因的生物学功能，信号通路分析可明确它们调控结直肠癌的信号通路的关键分子，为揭示结直肠癌发病机制奠定基础。

1 材料与方法

1.1 芯片数据信息 从美国国立生物技术信息中心(NCBI)的 GEO 芯片数据库中选取了 GSE9348[5]、GSE21815[6-7]、GSE32323[8]、GSE44076[9-12] 4 个芯片数据集，从中获取结直肠癌和正常或相邻黏膜组织的基因表达谱。GSE9348 采用美国昂菲公司人基因组 U133 Plus 2.0 芯片 GPL570 平台，包括 70 例早期结直肠癌和 12 例健康对照组织。GSE21815 采用美国安捷伦公司人基因组 4×44K G4112F 芯片 GPL6480 平台，包括 132 例结肠癌患者和 9 例正常对照组织。GSE32323 芯片检测采用美国昂菲公司人基因组 U133 Plus 2.0 芯片 GPL570 平台，包括 17 例配对的结直肠癌和非结直肠癌组织。GSE44076 采用美国昂菲公司人基因组 U219 芯片 GPL13667 平台，样本来源于 98 例结肠癌患者肿瘤和相邻正常黏膜组织及 50 例健康对照者的结肠组织。

1.2 数据处理 下载芯片数据压缩包和探针文件，通过 R3.4.4 软件的 RMA 算法对芯片数据进行标准化，采用 $|\log FC|>1$，$P<0.05$ 的入选标准，利用 R 语言的 limma 包筛选出差异表达基因进行进一步分析。

1.3 差异基因筛选 将 4 个 GEO 芯片数据集筛选出的在肿瘤中上调和下调基因分别导入在线软件 VENNY 2.1 (http://bioinfogp.cnb.csic.es/tools/venny/index.html)，通过取交集，获得在 4 个芯片数据集中基因表达均发生改变的差异基因。

1.4 基因注释分析 采用 Funrich 3.1.3 软件进行基因注释分析，分别显示差异基因的细胞组成(CC)、分子功能(MF)、生物学过程(BP)、信号通路(BPA)，结果根据-Log10(P)值的大小排序，同时呈现富集的基因占总体的比例。

1.5 PPI 网络构建 将表达差异基因导入 String 10.5 在线分析网站(https://string-db.org/)，获得蛋白相互作用的数据，然后通过 Cytoscape 3.6.1 软件对结果进行可视化和进一步分析。

2 结 果

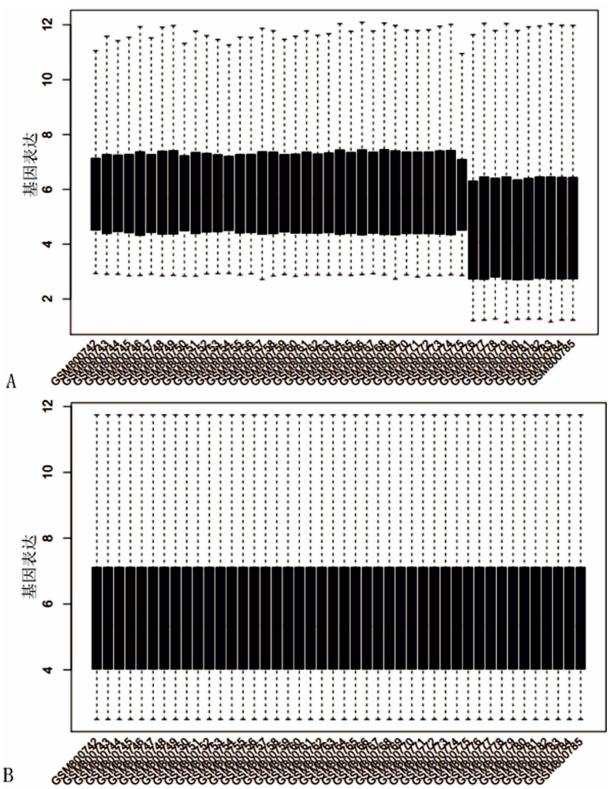
2.1 芯片数据标准化 对基因芯片数据的标准化处理，主要目的是消除由于实验技术所导致的表达量的变化，并且使各个样本和平行实验的数据处于相同的水平，从而得到具有生物学意义的基因表达量的变化。以 GSE32323 为例，该芯片数据在标准化前的箱线图，见图 1A。各样本的基因表达不在一条水平线上，通过分位数标准化后，将 34 例样本的芯片结果调整到同一水平，见图 1B。

2.2 差异基因筛选 通过差异基因分析，GSE9348 芯片中获得 1 355 个表达上调的基因，1 735 个表达下调的基因，GSE21815 芯片中获得 7 005 个表达上调的基因，490 个表达下调的基因，从 GSE32323 芯片中获得 722 个表达上调的基因，490 个表达下调的基因，GSE44076 芯片分析获得 821 个表达上调基因，873 个表达下调的基因。4 个 GEO 数据集取交集分别得到了表达上调的基因 277 个，见图 2A；表达下调的基因 153 个，见图 2B。

其中表达上调的基因包括参与细胞分裂周期的基因如 CDC6、CDC25B、CDCA5、CDCA7、GTF2IRD1 等，与细胞黏附功能相关的分子 CDH3、CLDN1 等，参与肿瘤转移的基质金属蛋白酶家族分子 MMP1、MMP3、MMP7、MMP7、MMP12 等。表达下调的基因包括参与机体代谢的 GPAT3、B3GNT7、AHCYL2 等，以及参与黏液分泌和免疫反应的 ADAMDEC1、CLCA1、CLCA4 等。

2.3 GO 分析和信号通路富集分析 为了更系统全面的了解上述差异基因的细胞定位、分子功能、参与的生物学过程及信号通路，采用 Funrich 3.1.3 软件将差异基因进行了基因富集分析并采用 GraphPad

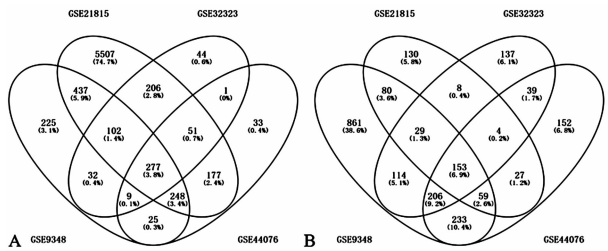
Prism 作图。结果发现，表达上调的基因主要分布于细胞核和细胞外，其基因占比分别为 20.6%（$P<0.001$）和 22.6%（$P=0.004$），在细胞浆、微管、中心体、微管中的基因数量少，基因占比分别为 3.5%（$P<0.001$）、11.5%（$P=0.003$）、4.8%（$P=0.002$），结果见图 3A；纺锤体的完整性决定了染色体分裂的正确性，上调差异基因主要参与纺锤体组装其基因占比为0.7%（$P=0.04$），因而它们表达上调引起细胞异常分裂是结直肠癌发生的关键因素，见图 3B；趋化因子在免疫监视过程中发挥重要作用，免疫监视功能过低，异常细胞可逃过监视形成肿瘤，上调差异基因的分子功能就是主要富集在调节趋化因子活性，其基因占比为 2.6%（$P=0.003$），见图 3C；肿瘤恶性增生主要表现为细胞增殖失控，上调差异基因主要富集在与细胞增殖相关的信号通路上，其中细胞周期、有丝分裂 G1-G1/S 期、M-M/G1 期、G2/M 期 DNA 破坏关键节点及 DNA 复制的基因占比分别是 21.8%（$P<0.001$）、12.8%（$P<0.001$）、15.8%（$P<0.001$）、6%（$P<0.001$）、15.8%（$P<0.001$），结果见图 3D。



注：A 为标准化前；B 为标准化后

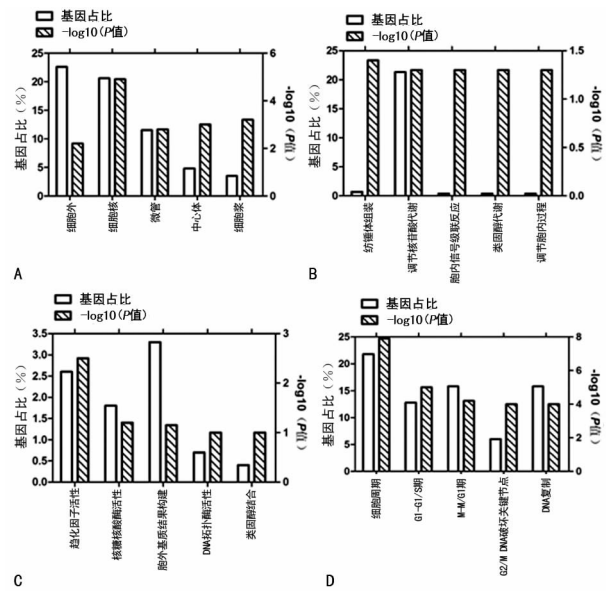
图 1　芯片数据标准化

153 个表达下调的基因主要富集在细胞外（基因占比 = 28.5%，$P<0.001$）、参与代谢过程（基因占比 = 19.2%，$P=0.029$）；机体内大多数化学反应都是催化反应，而下调的基因在调控催化活性方面发挥着重要作用，其基因占比为 10.3%（$P=0.006$），核受体通过调控靶基因从而影响肿瘤细胞的药物敏感性，下调的基因可调控配体依赖性核受体活性，基因占比为 2.7%（$P=0.039$），从而影响肿瘤的治疗效果；这些基因参与多条信号通路，但没有富集在某条信号通路上（$P>0.05$），见图 4。



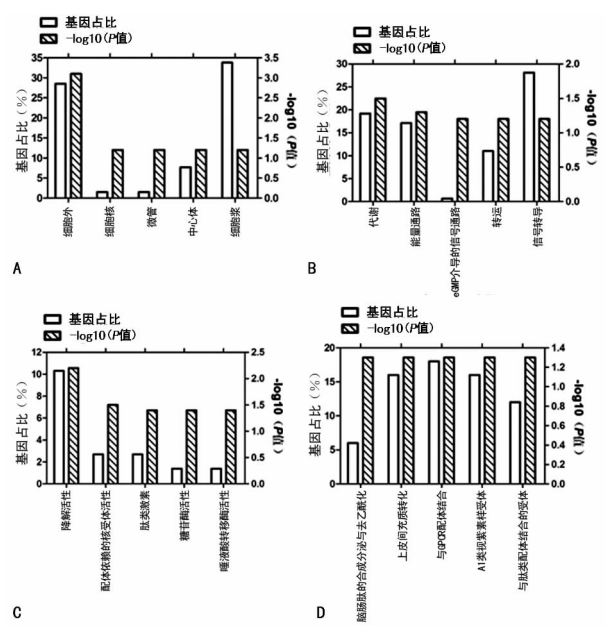
注：A 为表达上调的基因；B 为表达下调的基因

图 2　差异基因韦恩图



注：A 为细胞定位；B 为生物学功能；C 为分子功能；D 为生物学通路

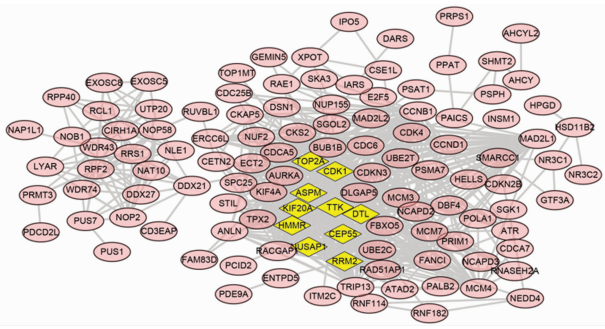
图 3　基因注释分析结直肠癌中表达上调的基因



注：A 为细胞定位；B 为生物学功能；C 为分子功能；D 为生物学通路

图 4　基因注释分析结直肠癌中表达下调的基因

2.4　蛋白互作网络构建　构建结直肠癌差异表达基因对应的蛋白的相互作用关系，有助于系统的研究疾病分子机制，找到结直肠癌相关的关键基因。通过 string 软件获得蛋白质相互作用结果，选取相互作用强（联合分数≥0.7）的蛋白质，再用 cytoscape 软件构建蛋白互作网络，去除无相互作用的基因，得到了一份包含 213 个差异表达基因及 1025 条相互作用关系的蛋白作用网络。见图 5。根据互作节点的数量的降序排列，筛选出其中的前十位基因，见表 1，它们是结直肠癌的关键基因。



注：节点表示在结直肠癌中差异基因对应的蛋白产物，其中关键基因用矩形显示；两节点间的线表示两节点对应的蛋白之间有相互作用

图 5　差异基因的蛋白质相互作用网络

表 1　蛋白相互作用筛选出的 10 个关键基因

排序	基因名	互作节点数（个）	结直肠癌中表达水平
1	CDK1	54	上调
2	CCNB1	47	上调
3	TOP2A	47	上调
4	MAD2L1	43	上调
5	TTK	40	上调
6	BUB1B	39	上调
7	AURKA	38	上调
8	RRM2	34	上调
9	UBE2C	34	上调
10	ASPM	33	上调

3　讨　　论

迄今为止，已有许多的研究者进行了大量的基础和临床研究，来揭示结直肠癌形成和进展的原因和机制，但全球结直肠癌的发病率和病死率仍居高不下，主要原因是大部分聚焦在单个遗传学事件或结果来源于单个队列研究[13]。本研究整合了 4 个 GEO 数据集，利用生物信息学的方法进行深度分析，首先，鉴定出了 430 个差异表达基因，包括 277 个表达上调的基因和 153 个表达下调的基因。表达上调的基因主要位于细胞核、细胞浆等，主要参与纺锤体组装，组装过程的异常可引起染色体异常分裂从而发生癌变[14]。其分子功能主要为调节趋化因子活性，参与细胞周期及 DNA 复制等，从而参与调控肿瘤免疫监视[15]，参与肿瘤细胞的迁移、增殖及凋亡[16]。下调基因主要富集在细胞外，参与代谢过程，发挥催化活性、配体依赖性核受体活性等作用，这些都是肿瘤发生发展的重要原因[17]。

CDK1 是调控 G2-M 关键节点的重要基因，在结直肠癌患者组织中检测到 CDK1 高表达，且 CDK1 核浆比越高，患者预后越差[18]。CCNB1 是调节细胞周期的重要基因，结直肠癌细胞中高表达的 CCNB1 可促进肿瘤细胞增殖和肿瘤生长[19]。MAD2L1 也是调控细胞有丝分裂的关键分子，已有研究发现该基因在肝癌中的异常高表达与患者的生存时间呈负相关，下一步可作为结直肠癌治疗的靶点[20]。在结直肠癌患者组织中检测到 TOP2A 基因表达增加，细胞实验发现，敲降 TOP2A 可抑制结肠癌细胞的增殖和侵袭能力[21]。TTK 是纺锤体组装关键节点，已有研究发现 TTK 在结肠癌组织中高表达，TTK 过表达结肠癌细胞可抵抗细胞凋亡[22]，TTK 还可通过线粒体调节肿瘤细胞的活力[23]。BUB1B 也同样参与纺锤体组装[24]，本研究首次提出该基因在结直肠癌中表达上调，其具体的作用机制尚无研究。AURKA 基因也已证实在结直肠癌患者组织中表达上调，该基因在细胞分裂和染色体稳定性发挥重要作用，可作为结直肠癌患者的预后标志物[25]。RRM2 基因过表达与肿瘤的侵袭性和化疗药物抵抗相关，可作为结直肠癌治疗的靶分子[26]。UBE2C 在结肠癌患者中高表达，可作为其诊断的标志物，研究表明，抑制 UBE2C 能减缓结直肠癌细胞生长速度，增加细胞对化疗药物的敏感性，因而可开发相应的分子靶向药物用于结直肠癌患者个体化治疗[27]。ASPM 是调控正常有丝分裂纺锤体功能的关键基因，可影响 DNA 双链断点的修复，能作为化疗药物的靶点[28]，但目前尚无研究报道该基因与结直肠癌的关系，研究者首次发现该基因在结直肠癌中表达上调。

4　结　　论

本研究联合 4 张结直肠癌基因芯片数据集，采用生物信息学的分析方法，获得了 430 个差异基因，构建了蛋白互作网络，最终获得 10 个关键基因，主要和细胞周期、纺锤体组装、染色体稳定性、肿瘤细胞侵袭和化疗药物耐药有关。这些发现有助于理解结直肠癌的成因和潜在的分子机制，筛选出的基因可作为结直肠癌治疗的靶点。

参考文献

[1] ARNOLD M, SIERRA M S, LAVERSANNE M, et al. Global patterns and trends in colorectal cancer incidence and mortality[J]. Gut, 2017, 66(4): 683-691.

[2] SIEGEL R L, MILLER K D, JEMAL A. Cancer statistics, 2018[J]. CA Cancer J Clin, 2018, 68(1): 7-30.

[3] TO K K, TONG C W, WU M, et al. MicroRNAs in the

prognosis and therapy of colorectal cancer: From bench to bedside[J]. World J Gastroenterol, 2018, 24(27): 2949-2973.

[4] LIANG B, LI C, ZHAO J. Identification of key pathways and genes in colorectal cancer using bioinformatics analysis[J]. Med Oncol, 2016, 33(10): 111-126.

[5] HONG Y, DOWNEY T, EU K W, et al. A metastasis-prone signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics[J]. Clin Exp Metastasis, 2010, 27(2): 83-90.

[6] IWAYA T, YOKOBORI T, NISHIDA N, et al. Downregulation of miR-144 is associated with colorectal cancer progression via activation of mTOR signaling pathway[J]. Carcinogenesis, 2012, 33(12): 2391-2397.

[7] KOGO R, SHIMAMURA T, MIMORI K, et al. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers[J]. Cancer Res, 2011, 71(20): 6320-6326.

[8] KHAMAS A, ISHIKAWA T, SHIMOKAWA K, et al. Screening for epigenetically masked genes in colorectal cancer Using 5-Aza-2′-deoxycytidine, microarray and gene expression profile[J]. Cancer Geno Prote, 2012, 9(2): 67-75.

[9] CLOSA A, CORDERO D, SANZ-PAMPLONA R, et al. Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis[J]. Carcinogenesis, 2014, 35(9): 2039-2046.

[10] CORDERO D, SOLE X, CROUS-BOU M, et al. Large differences in global transcriptional regulatory programs of normal and tumor colon cells[J]. BMC Cancer, 2014, 14(1): 708-719.

[11] SANZ-PAMPLONA R, BERENGUER A, CORDERO D, et al. Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer[J]. Mol Cancer, 2014, 13(1): 46-64.

[12] SOLE X, CROUS-BOU M, CORDERO D, et al. Discovery and validation of new potential biomarkers for early detection of colon cancer[J]. PLoS One, 2014, 9(9): e106748- e106758.

[13] DUFFY M J. Use of biomarkers in screening for cancer[J]. Adv Exp Med Biol, 2015, 867(1): 27-39.

[14] BROWN A, GEIGER H. Chromosome integrity checkpoints in stem and progenitor cells: transitions upon differentiation, pathogenesis, and aging[J]. Cell Mol Life Sci, 2018, 4(6): 211-224.

[15] POPOVIC A, JAFFEE E M, ZAIDI N. Emerging strategies for combination checkpoint modulators in cancer immunotherapy[J]. J Clin Invest, 2018, 128(8): 3209-3218.

[16] WANG C, LIU Z, XU Z, et al. The role of chemokine receptor 9/chemokine ligand 25 signaling: from immune cells to cancer cells[J]. Oncol Lett, 2018, 16(2): 2071-2077.

[17] 王水良. 核受体的研究进展[J]. 遗传学报, 2004, 4(31): 420-429.

[18] SUNG W W, LIN Y M, WU P R, et al. High nuclear/cytoplasmic ratio of Cdk1 expression predicts poor prognosis in colorectal cancer patients[J]. BMC cancer, 2014, 14(1): 951.

[19] XIAO Z, XUE J, GU W Z, et al. Cyclin B1 is an efficacy-predicting biomarker for Chk1 inhibitors[J]. Biomarkers, 2008, 13(6): 579-596.

[20] LI Y, BAI W, ZHANG J. MiR-200c-5p suppresses proliferation and metastasis of human hepatocellular carcinoma (HCC) via suppressing MAD2L1[J]. Biomed Pharm, 2017, 92(1): 1038-1044.

[21] ZHANG R, XU J, ZHAO J, et al. Proliferation and invasion of colon cancer cells are suppressed by knockdown of TOP2A[J]. J Cell Biochem, 2018, 1(8): 1-8.

[22] LING Y, ZHANG X, BAI Y, et al. Overexpression of Mps1 in colon cancer cells attenuates the spindle assembly checkpoint and increases aneuploidy[J]. Biochem Biophysical Res, 2014, 450(4): 1690-1695.

[23] ZHANG X, LING Y, GUO Y, et al. Mps1 kinase regulates tumor cell viability via its novel role in mitochondria[J]. Cell Death Dis, 2016, 7(7): 2292-2303.

[24] HAHN M M, VREEDE L, BEMELMANS S A, et al. Prevalence of germline mutations in the spindle assembly checkpoint gene BUB1B in individuals with early-onset colorectal cancer[J]. Genes Chrom Cancer, 2016, 55(11): 855-863.

[25] KOH H M, JANG B G, HYUN C L, et al. Aurora kinase a is a prognostic marker in colorectal adenocarcinoma[J]. J Pathol Transl Med, 2017, 51(1): 32-39.

[26] NANA A W, WU S Y, YANG Y S, et al. Nano-Diamino-Tetrac (NDAT) enhances resveratrol-induced antiproliferation by action on the RRM2 pathway in colorectal cancers[J]. Horm Cancer, 2018, 5(7): 25-34.

[27] CACCIOLA N A, CALABRESE C, MALAPELLE U, et al. UbcH10 expression can predict prognosis and sensitivity to the antineoplastic treatment for colorectal cancer patients[J]. Mole Carci, 2016, 55(5): 793-807.

[28] KATO T A, OKAYASU R, JEGGO P A, et al. ASPM influences DNA double-strand break repair and represents a potential target for radiotherapy[J]. Int J Radiat Biol, 2011, 87(12): 1189-1195.

(收稿日期: 2019-01-20　修回日期: 2019-04-21)