doi:10.3969/j.issn.1672-5565.2014.03.10

癌症相关的 DNA 甲基化连锁区域

王丽波,王　芳*,张　岩*
(哈尔滨医科大学生物信息科学与技术学院,黑龙江 哈尔滨 150086)

摘　要:DNA 甲基化是重要的表观遗传标记之一,在转录调控中起直接作用。DNA 甲基化的异常与癌症的发生发展密切相关。高通量测序使得在单碱基分辨率下检测全基因组的 DNA 甲基化水平成为可能。本文基于临近 CpGs 位点甲基化水平的相关性挖掘 DNA 甲基化连锁区域。结果发现 DNA 甲基化连锁区域的甲基化水平和模式在癌症中存在异常,而且显著富集到分化/发育相关的生物学功能。DNA 甲基化连锁区域的挖掘有助于对具有生物学功能的表观遗传标记的进一步理解,有助于对癌症诊断的表观遗传标记的挖掘。

关键词:DNA 甲基化;连锁不平衡;癌症;相关系数
中图分类号:R978.1+6　　**文献标志码**:A　　**文章编号**:1672-5565(2014)-03-213-05

DNA methylation linkage disequilibrium block are associated with cancer

WANG Libo, WANG Fang*, ZHANG Yan*
(*Department of Bioinformatics*, *Harbin Medical University*, *Harbin* 150086, *China*)

Abstract: DNA methylation is one of the important epigenetic markers which plays a direct role in transcriptional regulation. Abnormal of DNA methylation is closely associated with cancer development. High throughput sequencing technology has made it possible to measure genome-wide DNA methylation level based on single base resolution. We identified DNA methylation linkage disequilibrium blocks that showed strong correlation of DNA methylation between adjacent CpGs. We found that the methylation levels and patterns of block in cancer were significantly different from normal, and enriched in differentiation/development biological functions. The identification of DNA methylation block will help further understanding of epigenetic makers having biological functions, and even the mining of epigenetic biomarkers for cancer diagnosis.

Keywords: DNA methylation; Linkage disequilibrium; Cancer; Correlation coefficient

DNA 甲基化是表观遗传的重要修饰之一,并被广泛研究。DNA 甲基化一般发生于 CG 相连的二核苷酸部位(CpGs),通过改变染色质结构、DNA 构造和稳定性等对基因表达具有重要的调控作用[1]。随着表遗传学的发展,人们认识到肿瘤不仅是遗传性疾病,同时也是由 DNA 甲基化异常引起的基因调控失常的表观遗传性疾病[2]。人类基因组 DNA 存在广泛的甲基化修饰。在早期发育阶段,甲基化和去甲基化的交替进行是细胞得以生长和分化的关键程序,且在细胞正常发育以及保持基因组稳定性中起着至关重要的作用。正常细胞内,启动子区的 CPG 岛呈非甲基化状态,而大部分散在分布的 CpG 岛二核苷酸多发生甲基化[3]。肿瘤中常伴随基因组整体甲基化水平降低和某些基因 CpG 岛区域甲基化水平异常升高(如抑癌基因),并且这两种变化可在一种肿瘤中同时发生。基因组整体甲基化水平降低可导致原癌基因活化等,进一步促进了肿瘤的发生。基因启动子区的 CpG 岛发生异常高甲基化可导致基因转录沉默,使重要基因如抑癌基因等表达极度降低或不表达,进而也促进了肿瘤细胞的形

收稿日期:2013-11-01;**修回日期**:2013-11-22.
基金项目:哈尔滨医科大学大学生创业基金资助。
作者简介:王丽波,女,本科生,研究方向:计算表观遗传学;E-mail: wanglibo930@gmail.com.
*通信作者:张岩,女,博士,教授,研究方向:计算表观遗传学;E-mail: yanyou1225@163.com.
王芳,女,硕士,讲师,研究方向:计算表观遗传学;E-mail: wangfang@ems.hrbmu.edu.cn.

成[4-5]。P16INK4a 是一种细胞周期调控蛋白,通过与细胞周期蛋白依赖激酶 CDK4 及 CDK6 结合而抑制后者的蛋白激酶活性,从而抑制细胞的增殖。而 P16INK4a 基因启动子 5'端的 CpG 岛甲基化或外显子 1α 的 CpG 甲基化可导致 p16 表达缺失,从而导致该基因的失活,促进了癌症的形成,这一基因的灭活主要与胃癌的发生相关[6]。随着高通量测序技术的发展,单碱基分辨率下检测 DNA 甲基化的水平已经得以实现,促进了全基因组范围更高精度甲基化水平和模式的分析。重亚硫酸氢钠测序技术的短序列片段中包含多个 CpG 位点,而且这些位点之间的甲基化水平高度连锁,即其中一个 CpG 位点的甲基化改变能够通过另一 CpG 位点的甲基化变化来解释[7]。本文基于临近 CpG 位点之间的 DNA 甲基化水平高度相关这一假设,挖掘 DNA 甲基化连锁区域有助于挖掘基因组中有功能的甲基化区域,进一步理解这些区域在癌症中的改变模式,有助于挖掘癌症的表观遗传学诊断标记。

1 材料和方法

1.1 材料

UCSC 的 encode 数据库中(http://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html)下载 RRBS 的 DNA 基甲基化数据,包括 52 个正常样本(胚胎干细胞、成纤维细胞、主动脉平滑肌细胞、肾、脑、血、肝、肺、胰腺、心脏、胎盘、骨骼肌、皮肤、胃、睾丸、子宫、B-淋巴细胞、星形胶质细胞、成骨细胞)和 23 个癌症样本(白血病、肺癌组织、子宫颈癌、肝癌、乳腺腺癌、神经母细胞瘤、大肠腺癌、子宫内膜腺癌、前列腺癌、胚胎性癌、卵巢腺癌、胰腺癌、脑肿瘤、神经细胞株)[8]。在每个样本中将多次的生物学重复进行合并,同一个 CG 位点的甲基化水平取均值。统计所有 CG 位点所在的参考基因组位置(Hg19)、覆盖度以及相应的 DNA 甲基化水平。

1.2 方法

1.2.1 皮尔森相关系数的计算

根据 CpG 位点所在的参考基因组的位置从小到大进行排序,然后分别提取每个 CpG 位点对应的正常和癌症样本中的甲基化水平,分别构成正常和癌症的 DNA 甲基化水平向量。在正常和癌症样本中,基于 pearson 相关系数计算临近一个 CpG 位点之间的相关系数,公式如下:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}。\qquad(1)$$

其中,n 代表正常(癌症)样本的个数;x_i 和 y_i 分别代表第 i 个正常(癌症)样本中 CpG 位点及下一个 CpG 的甲基化水平;$\bar{x}$ 和 $\bar{y}$ 分别代表正常(癌症)样本中两个 CpG 位点 DNA 甲基化水平的均值。

1.2.2 相关系数阈值的确定

为了确定临近 CpG 位点之间相关系数的阈值,本文从基因组中随机抽取两个 CpG 点按照上述公式计算其相关系数。定义错误发现率(*FDR*)的公式如下:

$$FDR=\frac{\#(r_{\text{random}}>r_0)}{\#(r_{\text{observe}})}>r_0。\qquad(2)$$

其中,分子表示随机情况下相关系数大于 r_0 的数目;分母表示真实情况下相关系数大于 r_0 的数目。根据 $FDR=0.01$ 确定相关系数的阈值。当 CpG 对之间的相关系数高于此阈值时则认为二者的 DNA 甲基化水平连锁,否则为不相关。

大量的含有少量 CpG 位点的区域被获得。这些含有少量的 CG 位点的区域,临近之间的相关性与样本数量的偶然因素相关,尤其是含有两个 CG 位点的区域。我们认为只有多个 CG 位点相邻并且具有高度连锁的甲基化模式的区域才具有调控的功能。为了确定区域内含有的 CG 位点的数目的阈值,打乱了临近 CpG 位点的样本标签,重新计算 r 值。然后根据阈值筛选 DNA 甲基化区域,得到随机情况下 DNA 甲基化 block 所含有的 CG 位点的数目的零分布。

1.2.3 DNA 甲基化连锁区域的定义

如果临近的 CpG 位点之间甲基化水平的相关系数大于阈值,则将 CpG 连接然后向下一个 CpG 位点延伸,直到相关系数小于阈值则延伸停止。该区域被定义为 DNA 甲基化连锁区域。计算该区域中所有 CpG 位点在所有样本中的甲基化水平的均值,定义为 DNA 甲基化连锁区域的甲基化水平。

2 结果分析

2.1 DNA 甲基化连锁区域的挖掘

本文分别从正常样本和疾病样本中获得 902 825,920 516 个 CpG 位点,全基因上临近 CpG 位点之间的距离分布显示大部分 CpG 位点之间的距离不超过 100 bp。分别计算临近位点的皮尔森相关系数,根据 $FDR=0.01$ 确定皮尔森相关系数平方的阈值为 0.75。如果临近的 CpG 位点之间的相关系数超过阈值则将其相连并向下延伸,直到相关系数的平方小于 0.75 为止,得到的区域被认为是 DNA 甲基化连锁区域。最终,737 个 DNA 甲基化连锁区域在正常样本中获得,3 384 个 DNA 甲基化连

锁区域在癌症样本中获得。DNA 甲基化连锁区域发现在正常样本和癌症样本中甲基化连锁区域的长度没有差别(见图 1A),但是区域内所含的 CG 位点的数目、平均甲基化水平以及 R^2 存在显著差异(见图 1B,1C,1D),而且区域内的 R^2 并没有随着区域长度的增加而降低(见图 1D)。意味着在癌症中临近的 CpG 位点之间倾向更强的连锁程度,而且这种 DNA 甲基化连锁区域倾向于发生在 CpG 密集的区域倾向发生高甲基化变异,暗示着连锁区域内的 CpG 位点可能共同发生异常导致癌症的发生。

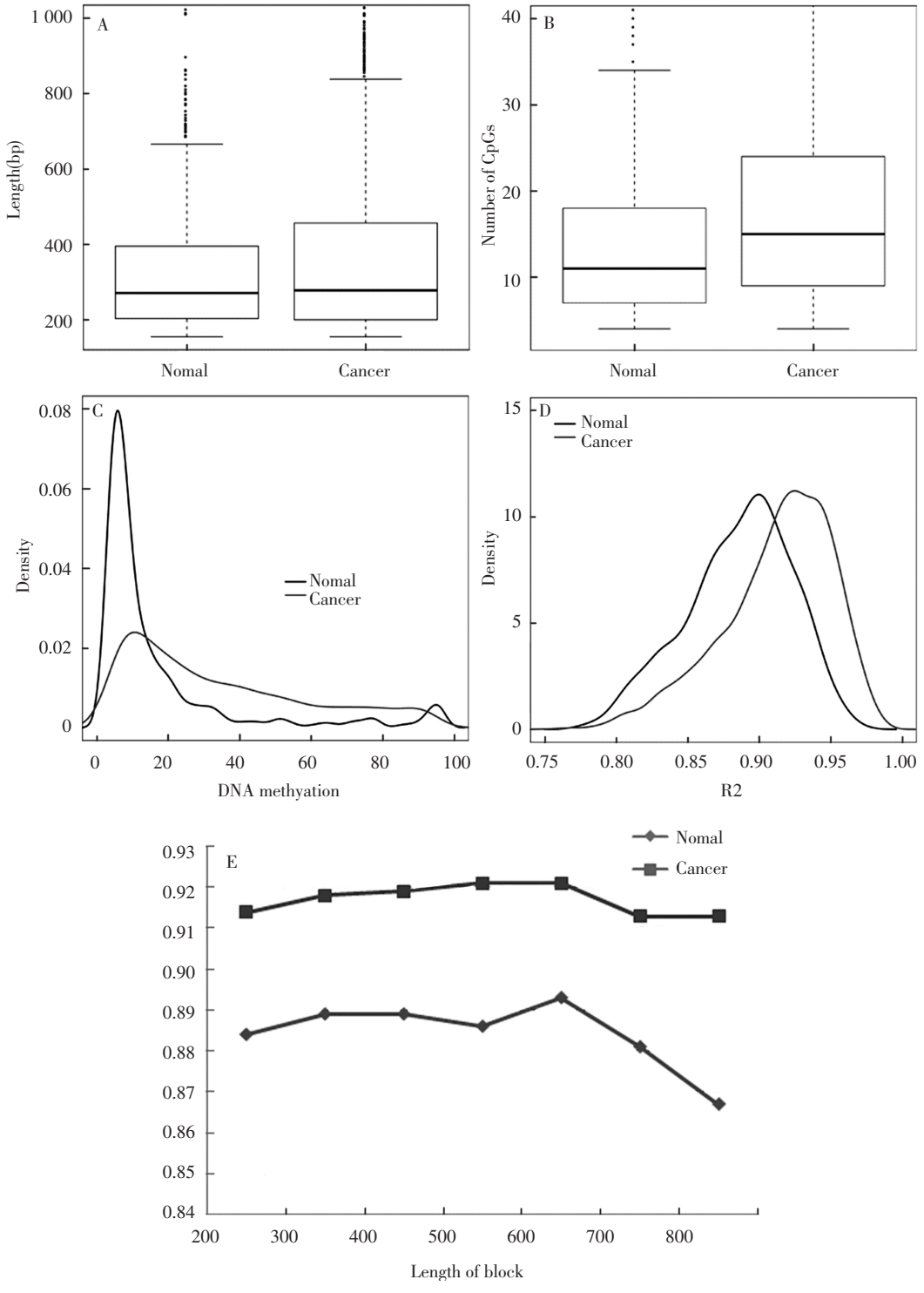

图 1　DNA 甲基化连锁区域的特征

Fig. 1　Characteristic of DNA methylation block

注:A:全基因组内临近的 CpG 位点之间的距离;B:DNA 甲基化连锁区域长度的分布;C:DNA 甲基化连锁区域内 CG 位点的数目;D:DNA 甲基化连锁区域内甲基化水平的分布;E:DNA 甲基化连锁区域长度与 R^2 的关系。

Notes: A: Distance of adjacent CGs in genome wide; B: Length distribution of DNA methylation block; C: Number of CGs in DNA methylation block; D: Distribution of DNA methylation in DNA methylation block; E: Relationship between R^2 and length of block.

2.2 DNA 甲基化连锁区域的生物学意义

为了进一步研究 DNA 甲基化连锁区域的生物学功能及意义,分别将正常样本和癌症样本的 DNA 甲基化连锁区域进行基因本体论(GO)的功能富集分析。如果一个 DNA 甲基化连锁区域的上下游 500 bp范围内存在基因,则该基因被认为是 DNA 甲基化连锁区域的相关基因。我们在正常样本中找到 617 个相关基因,在癌症样本中找到 2 575 个相关基因。将 DNA 甲基化连锁区域的相关基因采用 DAVID 工具(http://david.abcc.ncifcrf.gov/)进行基因功能富集分析,多重检验矫正之后的显著性水平定义为 0.01。癌症样本和正常样本中显著性水平最高的前 10 个功能(见图 2A,B),结果显示正常和癌症的 DNA 甲基化连锁区域都富集到分化\发育以及表达调控的功能,尤其是在癌症中与神经元的发育和分化相关。此外,癌症中 DNA 甲基化连锁区域的 KEGG 富集分析显示富集到癌症通路和细胞形成通路(见图 2C)。结果表明,癌症中 DNA 甲基化连锁区域可能促使癌症的发生。

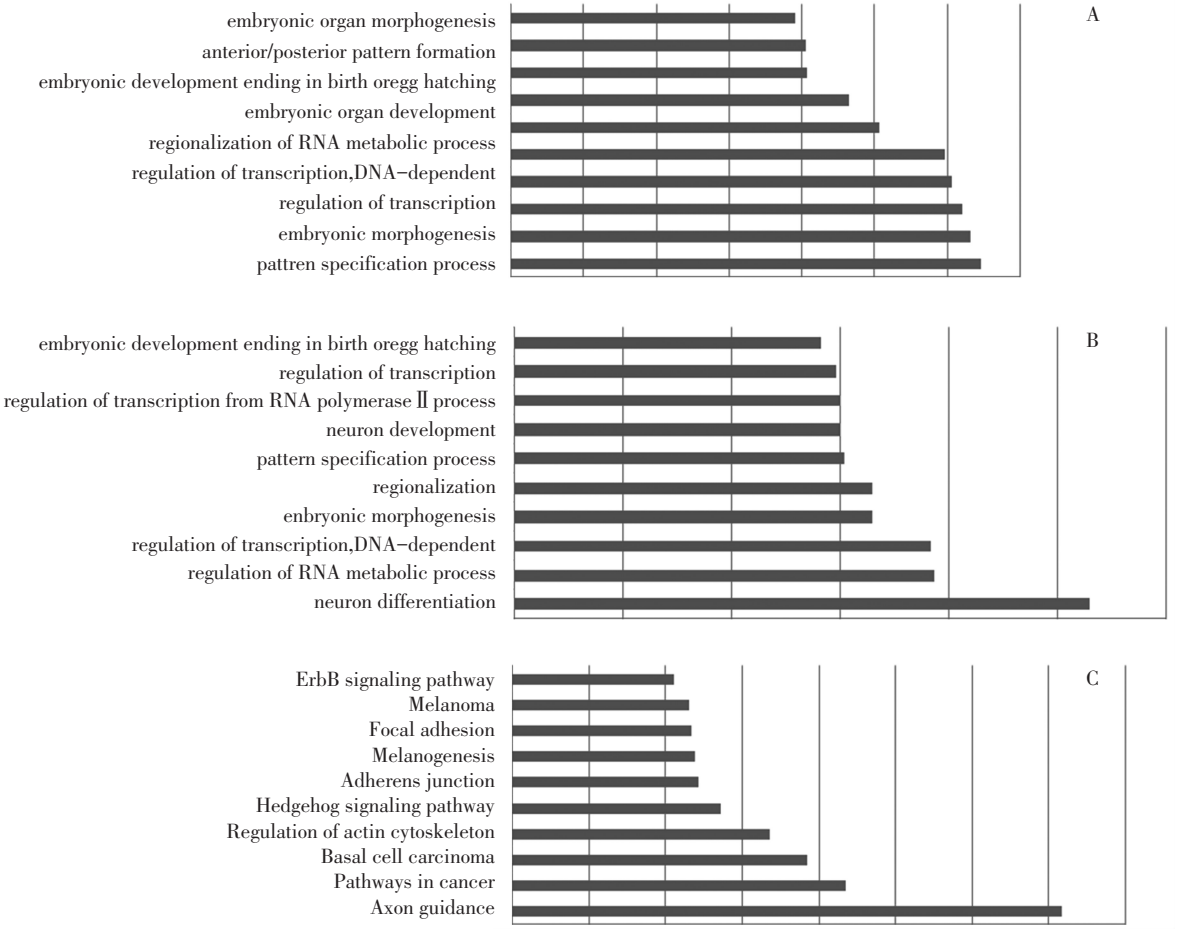

图 2 DNA 甲基化连锁区域的功能富集

Fig.2 Functional enrichment of DNA methylation block

注:A:正常的 GO 富集结果;B:癌症的 GO 富集结果;C:癌症的 KEGG 富集结果;x 轴表示富集分析的 P 值以 **10** 为底的负对数。

Notes:A:GO enrichment results of normal;B:GO enrichment results of cancer;C:KEGG enrichment results of cancer.

2.3 DNA 甲基化连锁区域在癌症中的异常模式

DNA 甲基化连锁区域尽管在正常样本和癌症样本中均存在很强的连锁程度,但是在两类样本中呈现的不同甲基化水平和模式。以 HIC1 基因为例,该基因对生长调节和肿瘤的抑制具有重要作用。位于该基因中超甲基化区域的缺失与肿瘤、Miller-Dieker 综合征存在至关重要的联系。图 3 显示,在本研究中该基因位于 chromosome 17p13.3 区域,在正常样本和癌症样本中存在 DNA 甲基化连锁区域。该连锁区域在正常和癌症样本中均呈现了紧密的连锁程度(r^2=0.780,0.798),然而该区域在两类样本中的甲基化模式存在显著差异。该区域的甲基化水平在癌症样本中显著高于正常样本,而且连锁程度的变异小于正常样本。这意味着该连锁区域的异常甲基化可能与癌症相关,甚至可以作为癌症的表观遗传诊断标记。

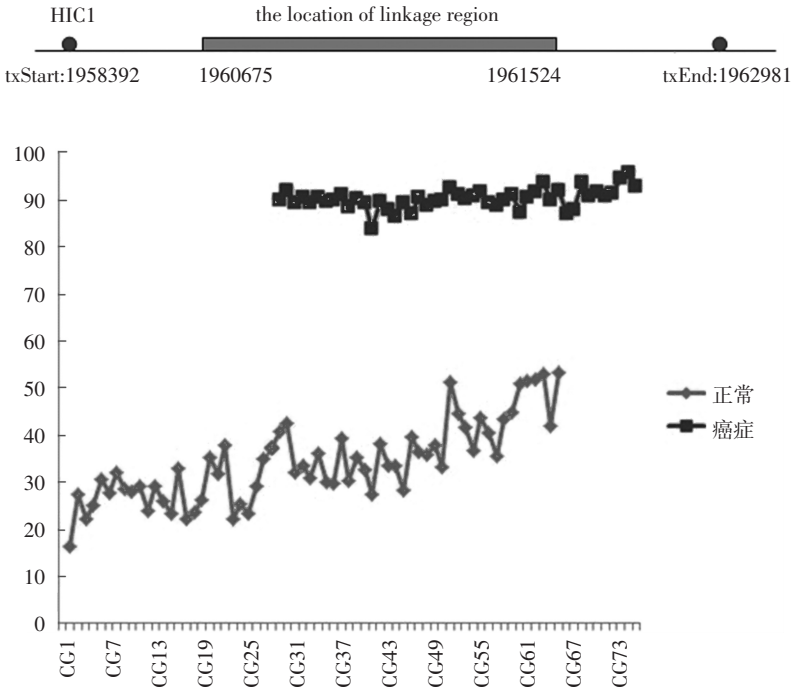

图 3　位于 HIC1 基因内的 DNA 甲基化连锁区域

Fig.3　DNA methylation block located within HIC1 gene

3　结　论

近年来,随着表观遗传学的发展,越来越多的研究表明 DNA 甲基化的异常与癌症的发生发展密切相关。本文的结果显示 DNA 甲基化连锁区域与癌症的关联不仅仅体现在甲基化水平上而且体现在甲基化模式上。挖掘 DNA 甲基化连锁区域有助于挖掘基因组中有功能的甲基化区域,而这些区域在癌症中的改变模式有助于挖掘癌症的表观遗传学诊断标记。希望能为研究者开启一个新的角度去探索 DNA 的甲基化水平与癌症发生的联系,进而对疾病能够更好的进行诊断和治疗。

参考文献(References)

[1] CHRISTOPHER G B, SARAH F, CECILIA M L, et al. Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus [J]. PLoS One, 2010, 5 (11): e14040.

[2] LUDVÍKOVÁ M, PESTA M, HOLUBEC L J, et al. New aspects of tumor pathobiology [J]. Ceskoslovenská Patologie, 2009, 45(4): 94.

[3] BANERJEE, HIRENDRA N, MUKESH V. Epigenetic mechanisms in cancer [J]. Biomarkers, 2009, 3(4): 397-410.

[4] QURESHI, SOHAIL A, MUHAMMED U B,et al. Utility of DNA methylation markers for diagnosing cancer [J]. International Journal of Surgery, 2010,8(3): 194-198.

[5] 吴川清,陶凯雄.内皮素 B 受体基因甲基化与肿瘤关系的研究进展[J]. 世界华人消化杂志, 2010, 18 (23): 2448-2452.
WU Chuanqing,TAO Kaixiong.Research progress of endothelin B receptor gene methylation and cancer [J]. World Journal of Gastroenterology, 2010, 18(23): 2448-2452.

[6] MERLO, ADRIAN,JAMES G H, et al. 5′ CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers [J]. Nature Medicine, 1995,1(7): 686-692.

[7] SHOEMAKER, ROBERT, DENG Jie, et al. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome [J]. Genome Research, 2010, 20(7): 883-889.

[8] ROSENBLOOM, KATE R, TIMOTHY R D, et al. ENCODE whole-genome data in the UCSC genome browser [J]. Nucleic Acids Research, 2010, 38 (suppl 1): D620-D625.