专家介绍

魏晟,男,副教授,博士生导师,华中科技大学同济医学院公共卫生学院流行病与卫生统计学系主任。中华预防医学会流行病学青年委员会常委、湖北省预防医学会流行病学分会常委、美国AACR 学会会员。长期从事肿瘤分子流行病学研究,已在国内外专业期刊发表学术论文 60 余篇,担任《International Journal of Cancer》、《Carcinogenesis》、《Environmental and Molecular Mutagenesis》、《Cancer》等专业杂志的审稿人。现主持国家自然科学基金面上项目 1 项,作为主要课题参加者,参加欧盟第七框架项目、教育部创新团队、国家"十一五"科技攻关项目以及多个省部级课题。研究成果曾获湖北省科技进步二等奖。

基于 TCGA 数据库挖掘肺腺癌预后相关的甲基化位点和基因

王 可,赵荣仙,杨素莲,孟 瑜,魏 晟*

(华中科技大学同济医学院公共卫生学院流行病与卫生统计学系,湖北 武汉 430030)

[摘 要] **目的**:通过对 TCGA(The Cancer Genome Atlas)数据库的数据挖掘,扫描全基因组范围内的肺腺癌预后相关的甲基化位点。**方法**:采用 2016 年 4 月从 TCGA 网站下载的肺腺癌预后数据及基于 Illumina Methylation 450 芯片的全基因组甲基化数据,经过数据连接,纳入同时有临床预后信息和甲基化数据的肺腺癌患者 232 例。采用 Cox 比例风险模型分析各位点甲基化水平对肺腺癌总生存率的影响,并评估其与对应基因 mRNA 表达的相关性,进一步评估 mRNA 表达对肺腺癌预后的影响。**结果**:纳入的 232 例肺腺癌患者的平均年龄为(64.823 ± 9.300)岁,平均生存时间为(20.217 ± 17.067)个月。本研究发现肺腺癌预后相关最显著的甲基化位点为位于基因 EHBP1 区域的 cg03955927 (P=1.98×10^{-7}),对应的 HR 值及 95%可信区间为 0.605(0.501~0.731)。筛选的肺腺癌预后关联性最强的前 20 个甲基化位点中,17 个位点的高甲基化水平为肺腺癌预后的保护因素,其他 3 个为危险因素。5 个甲基化位点的甲基化水平与对应基因的 mRNA 表达相关,其中 cg12013757 对应的 KRI1 基因的 mRNA 表达与肺腺癌的预后有关(HR:1.316,95%CI:1.109~1.561,P=0.001 6)。**结论**:本研究通过对 TCGA 数据库的挖掘,初步发现 KRI1 基因的甲基化位点的甲基化水平对肺腺癌的预后有影响,可以作为为肺癌预后的生物标志物进一步研究。

[关键词] 肺腺癌;TCGA;甲基化;mRNA;预后

[中图分类号] R734.2 [文献标志码] A [文章编号] 1007-4368(2016)06-665-05

doi:10.7655/NYDXBNS20160605

DNA methylations associated with survival of lung adenocarcinoma with TCGA database

Wang Ke,Zhao Rongxian,Yang Sulian,Meng Yu,Wei Sheng*

(*Department of Epidemiology and Biostatistics,School of Public Health,Tongji Medical college,Huazhong University of Science and Technology,Wuhan 430030,China*)

[**Abstract**] **Objective**:The aim of this study was to investigate the DNA methylation sites associated with the prognosis of patients with lung adenocarcinoma in the whole genome-level with the data-mining for TCGA(The Cancer Genome Atlas)public database. **Methods**:The clinical data of lung adenocarcinoma patients was downloaded from TCGA database in April,2016,as well as genome-wide DNA methylation data with Illumina's Infinum Human Methylation 450 Bead Chips(HM450). After data linked,232 cases of lung adenocarcinoma with the information of both clinic and methylation were finally included in this study. A Cox's proportional hazards regression analysis was conducted to assess the relationship between DNA methylation levels or mRNA expression and the overall survival of lung adenocarcinoma patients,further evaluate the correlation between DNA methylation and mRNA expression,as

[基金项目] 国家自然科学基金(81172754)

*通信作者(Corresponding author),E-mail:Weisheng@mail.tjmu.edu.cn

well as the relationship between mRNA expression and lung adenocarcinoma survival. **Results**: The mean age and survival time of 232 people with lung adenocarcinoma was (64.823 ± 9.300)years and (20.217 ± 17.067)months. cg03955927 located in EHBP1 was found to be the strongest methylation site in this study which was associated with the survival of lung adenocarcinoma, adjusted HRs(hazard ratio)were 0.605 (0.501-0.731), P=1.98×10^{-7}. Among the strongest 20 methylation sites, the high methylation level of 17 sites were considered as protective factors for the survival of lung adenocarcinoma and that of other 3 sites were risk factors. The methylation levels of 5 methylation sites affect the targeted gene's mRNA expression. In addtion, the mRNA expression of KRI1 was associated with the survival of lung adenocarcinoma, adjusted HRs were 1.316 (1.109-1.561)with a P value of 0.001 6. **Conclusion**: With TCGA data mining, we found that methylation site in KRI1 gene region is highly related to lung adenocarcinoma prognosis, as further study biomarker related to prognosis of lung cancer.

[**Key words**] lung adenocarcinoma; TCGA; methylation; mRNA; prognosis

[Acta Univ Med Nanjing, 2016, 36(06): 665-669]

目前肺癌是人类最常见的恶性肿瘤之一,且其发病率和病死率呈逐渐上升趋势[1-2]。目前研究认为肺癌的发生发展是多因素、多阶段以及多基因改变共同作用的过程,其中抑癌基因的甲基化失活是其重要机制之一[3-5]。已有研究发现某些甲基化位点可以影响基因的表达,且与肺癌的预后有关[6-7]。然而,当前这些研究大多只针对某一染色体某一区域或某个基因片段进行甲基化分析,尚缺少全基因组范围水平对肺癌预后影响的研究。为此,本研究通过对 TCGA(The Cancer Genome Atlas)公共数据库的数据挖掘,探索全基因组范围内甲基化与肺腺癌预后的相关性,发现与甲基化相关的 mRNA,进而影响肺腺癌预后的基因,为今后肺腺癌预后相关标志物的研究提供科学依据。

1 材料和方法

1.1 材料

本研究使用的数据库是 2016 年 4 月从 TCGA 网站下载的肺腺癌的临床预后数据及基于美国 Illumina 公司 Methylation 450 芯片检测的全基因组甲基化在 level3 的数据,经过数据连接,保留同时有临床预后信息和甲基化数据的肺腺癌患者 232 例。mRNA 数据来自 TCGA 中肺腺癌 RNA 测序数据。

1.2 方法

本次纳入的 TCGA 甲基化数据库,选用了 Illumina Human Methylation 450 微阵列平台的 level3 的甲基化芯片数据。在该检测平台上,每个样本对应的 Beta 值将映射到基因组上(甲基化位点/基因)并进行 DNA 甲基化分析。此平台可检测人类全基因组近 450 000 个甲基化位点, 具有单碱基的分辨率,全面覆盖了 96%的 CpG 岛。mRNA 表达水平数据来自肺腺癌组织的 RNA 测序数据。

1.3 统计学方法

连续性变量用均数 ± 标准差($\bar{x} \pm s$)表示,分类型变量用样本量(构成比)表示。采用 Cox 比例风险模型分析全基因组各甲基化位点甲基化水平对肺腺癌总生存率的影响, 校正后的 P值≤1×10^{-8} 为对肺腺癌预后影响有统计学意义。采用线性相关模型来评估不同甲基化水平与其对应基因 mRNA 表达的相关性。进一步将 mRNA 水平进行三分位数分组,运用多元 Cox 比例风险模型对 mRNA 表达水平与肺腺癌预后之间的关系进行分析。模型中调整了年龄、性别、种族、吸烟以及肺腺癌病理分期 5 个变量。所有分析采用 SAS9.4 软件进行。除全基因组预后分析外,$P \leqslant 0.05$ 为差异有统计学意义。

2 结 果

2.1 纳入分析的人群基本特征

纳入的 232 例肺腺癌患者中,男 105 例(45.3%),女 127 例(54.7%);平均年龄(64.823 ± 9.300)岁;白种人 190例(81.9%),非白种人 42 例(18.1%);现在吸烟的有 49 例(21.3%),从不吸烟的有 30 例(13.0%),已戒烟的有 151 例(65.7%)。肺腺癌病理分期 1 期 142 例 (61.2%),2 期 54 例 (23.3%),3 期 36 例(15.5%);其中 1 期包括Ⅰ、ⅠA、ⅠB,2 期包括ⅡA、ⅡB,3 期包括ⅢA、ⅢB、Ⅳ。平均生存时间(20.217±17.067)个月。其中有 12 例肺腺癌患者在随访结束时死亡,占总人群的 5.2%。

2.2 与肺腺癌预后相关的甲基化位点

本研究采用 Cox 比例风险模型的方法对肺腺癌的预后情况进行分析。模型中调整了年龄、性别、种族、吸烟以及肺腺癌病理分期 5 个变量。将甲基化位点按照 Cox 生存分析结果的 P 值从小到大进行排序,选取 P 值最小的前 20 个甲基化位点(表 1)。

位于基因 LOC100132215/EHBP1 的甲基化位点 cg03955927 与肺腺癌的预后分析显示 P 值最小(P=1.98×10^{-7}),对应的 HR 值为0.605(0.501~0.731)。所有 20 个甲基化位点中,17个甲基化位点对应的 HR 值 0.61~0.65,提示这些甲基化位点的高甲基化水平可能是肺腺癌预后的保护性因素。另外 3 个位点 cg15414833、cg19640339、cg02525822(分别位于基因 RUNDC3B/ABCB1、EIF2B4/SNX17、KLHL36)的高甲基化水平为影响肺腺癌预后的危险因素,HR 值 1.52~1.58。

表 1　与肺腺癌预后相关的前 20 个甲基化位点

Table 1　The strongest 20 methylation sites associated with lung adenocarcinoma survival

甲基化位点	基因	染色体	染色体位置	HR(95%CI)[a]	P 值
cg03955927	LOC100132215,EHBP1	2	63272232	0.605(0.501~0.731)	1.98×10^{-7}
cg26741954	GRWD1	19	48954979	0.608(0.504~0.734)	2.08×10^{-7}
cg21927420	MAD1L1	7	2106929	0.640(0.532~0.769)	2.05×10^{-6}
cg11323433	CARS2	13	111301576	0.638(0.529~0.769)	2.41×10^{-6}
cg09781414	WDR90	16	715207	0.643(0.535~0.774)	2.98×10^{-6}
cg22257667	LRRC8A	9	131670714	0.639(0.529~0.773)	4.05×10^{-6}
cg03833378	ZNF500	16	4802600	0.642(0.531~0.776)	4.70×10^{-6}
cg09228454	SMUG1	12	54577528	0.645(0.534~0.778)	4.88×10^{-6}
cg10043427	SDK1	7	4167809	0.650(0.540~0.783)	5.47×10^{-6}
cg15414833	RUNDC3B,ABCB1	7	87257767	1.583(1.298~1.930)	5.62×10^{-6}
cg12013757	KRI1	19	10668565	0.656(0.547~0.787)	5.76×10^{-6}
cg24438277	AP3D1	19	2120959	0.653(0.543~0.785)	5.98×10^{-6}
cg09939191		14	103377153	0.657(0.547~0.788)	6.05×10^{-6}
cg14018434	SLC2A8	9	130161758	0.658(0.549~0.789)	6.19×10^{-6}
cg02525822	KLHL36	16	84682731	1.519(1.267~1.821)	6.25×10^{-6}
cg19640339	EIF2B4,SNX17	2	27593343	1.542(1.278~1.862)	6.31×10^{-6}
cg12533565		7	6694373	0.647(0.535~0.782)	6.77×10^{-6}
cg20309371	MAD1L1	7	2135948	0.645(0.532~0.781)	7.29×10^{-6}
cg15478515	SDK1	7	4219174	0.648(0.536~0.783)	7.38×10^{-6}
cg21186438	TPCN2	11	68851932	0.646(0.533~0.783)	8.14×10^{-6}

a:调整了年龄、性别、种族、吸烟、肺腺癌病理分期。

2.3　甲基化水平与 mRNA 表达的相关关系

不同位点的甲基化水平与相关基因 mRNA 表达的相关分析结果见表 2。从表中可以看出,调整年龄、性别、种族、吸烟以及肺腺癌病理分期 5 个变量后,甲基化位点所在的 18 个基因中,甲基化水平与其对应基因的 mRNA 表达水平呈显著相关的基因有 6 个:GRWD1(cg26741954)、ZNF500(cg03833378)、SMUG1(cg09228454)、SDK1(cg10043427,cg15478515)、RUNDC3B(cg15414833)、KRI1(cg12013757)。其中,甲基化水平与 GRWD1、ZNF500、SMUG1、RUNDC3B、KRI1 基因的 mRNA 表达呈正相关,与基因 SDK1 的 mRNA 表达呈负相关。

表 2　甲基化水平与 mRNA 表达的相关分析

Table 2　The correlation analysis between DNA methylation levels and mRNA expression

甲基化位点	基因	相关系数[a]	P 值
cg03955927	EHBP1	−0.048	0.505 4
cg26741954	GRWD1	−0.241	0.000 7
cg21927420	MAD1L1	0.069	0.342 1
cg11323433	CARS2	−0.051	0.476 5
cg09781414	WDR90	−0.054	0.452 9
cg22257667	LRRC8A	0.021	0.770 0
cg03833378	ZNF500	−0.195	0.006 4
cg09228454	SMUG1	−0.215	0.002 6
cg10043427	SDK1	0.208	0.003 6
cg15414833	RUNDC3B	−0.158	0.027 9
cg15414833	ABCB1	−0.042	0.559 7
cg12013757	KRI1	−0.152	0.034 2
cg24438277	AP3D1	−0.007	0.926 3
cg14018434	SLC2A8	0.001	0.986 0
cg02525822	KLHL36	−0.062	0.393 0
cg19640339	EIF2B4	0.102	0.157 1
cg19640339	SNX17	0.017	0.809 2
cg20309371	MAD1L1	−0.005	0.942 0
cg15478515	SDK1	0.203	0.004 5
cg21186438	TPCN2	−0.047	0.510 9

a:调整了年龄、性别、种族、吸烟、肺腺癌病理分期。

2.4　相关基因 mRNA 表达对肺腺癌预后的影响

本研究采用 Cox 比例风险模型分析各基因 mRNA 表达水平与肺腺癌预后之间的关系(表 3)。模型中调整了年龄、性别、种族、吸烟以及肺腺癌病理分期 5 个变量。与肺腺癌预后相关的基因有 WDR90(HR:1.318,95% CI:1.114~1.559,P=0.001 3)、KRI1(HR:1.316,95%CI:1.109~1.561,P=0.001 6)和

KLHL36(HR:0.803,95%CI:0.680~0.949,*P*=0.009 8)。

综合上述表格的分析结果发现,位于基因 KRI1 的甲基化位点 cg12013757 与肺腺癌的预后有关(HR:0.656,95% CI:0.547 ~0.787,$P=5.76\times10^{-6}$),且 cg12013757 位点与 mRNA 的表达相关,相关系数为-0.152,*P*=0.0342。另外,基因 KRI1 的 mRNA 表达水平可以影响肺腺癌的预后(HR:1.316,95%CI:1.109~1.561,*P*=0.001 6)。

表 3 mRNA 表达对肺腺癌预后的影响

Table 3 Influence of mRNA expression on the overall survival of lung adenocarcinoma patients

基因	甲基化位点	HR(95%CI)[a]	*P* 值
EHBP1	cg03955927	1.131(0.955~1.339)	0.153 6
GRWD1	cg26741954	1.021(0.861~1.211)	0.809 4
MAD1L1	cg21927420; cg20309371	1.095(0.929~1.290)	0.277 9
CARS2	cg11323433	0.989(0.835~1.172)	0.900 2
WDR90	cg09781414	1.318(1.114~1.559)	0.001 3
LRRC8A	cg22257667	1.038(0.883~1.221)	0.651 5
ZNF500	cg03833378	1.159(0.977~1.375)	0.090 4
SMUG1	cg09228454	0.992(0.839~1.173)	0.924 1
SDK1	cg10043427; g15478515	0.981(0.827~1.165)	0.830 8
RUNDC3B	cg15414833	0.987(0.830~1.173)	0.881 2
ABCB1	cg15414833	1.082(0.917~1.278)	0.349 8
KRI1	cg12013757	1.316(1.109~1.561)	0.001 6
AP3D1	cg24438277	1.092(0.924~1.291)	0.301 1
SLC2A8	cg14018434	1.014(0.863~1.191)	0.865 1
KLHL36	cg02525822	0.803(0.680~0.949)	0.009 8
EIF2B4	cg19640339	0.904(0.765~1.068)	0.235 5
SNX17	cg19640339	0.859(0.726~1.017)	0.077 5
TPCN2	cg21186438	1.171(0.995~1.379)	0.057 5

a:调整了年龄、性别、种族、吸烟、肺腺癌病理分期。

3 讨 论

本研究通过对 TCGA 数据库进行挖掘,初步发现 20 个与肺腺癌预后关联性最强的相关甲基化位点,在此基础上分析这些甲基化位点与 mRNA 表达的相关关系,以及对应基因 mRNA 表达水平对肺腺癌预后的影响,最后发现位于 KRI1 基因的甲基化位点的甲基化水平不仅与肺腺癌的预后有关,而且与对应基因的 mRNA 水平有关,对应基因的 mRNA 表达也与肺腺癌的预后有关。由于样本量的限制,本研究在进行甲基化数据挖掘时发现的最小 *P* 值为 1.98×10^{-7},未达到全基因组关联分析的显著性水平。但结合有关的分析结果,本研究发现的相关甲基化位点的甲基化水平及相关基因与肺腺癌预后的相关性仍具有生物学机制的基础。

DNA 甲基化主要通过对 CpG 序列的胞嘧啶进行甲基化修饰来调控基因的表达,基因 CpG 岛高甲基化常导致基因转录沉默,使抑癌基因、致癌基因及 DNA 调控基因等失去功能,引起染色体不稳定,从而使正常细胞或肿瘤细胞生长分化调控失常,导致或抑制肿瘤的发生[8-11]。本研究发现若干基因的高甲基化水平可以改善肺腺癌的预后,相关功能研究支持这一相关性。比如 EHBP1 是 EH 域结合蛋白 1,其包含 11 个转录因子结合位点,可调节肌动蛋白动力学以及网格蛋白介导的内吞作用[12]。目前已有研究发现 EHBP1 在细胞中的高表达可导致肌动蛋白进行广泛重组[13],EHBP1 基因的高甲基化使该基因的活性被抑制,进一步促进肿瘤细胞的凋亡,阻碍肿瘤的发生发展[14]。GRWD1 是一种附着于组蛋白的蛋白质,在整个基因组中起着调节染色体、维护基因组完整性的作用[15]。GRWD1 基因的异常甲基化可抑制其活性,研究显示 GRWD1 在肿瘤细胞中的高表达可促进肿瘤细胞生长,而低表达则会导致肿瘤细胞 DNA 损伤[16],故 GRWD1 的异常甲基化可能是肿瘤发生发展过程中的一个保护性因素。然而,本研究发现的若干基因的甲基化水平与对应基因的 mRNA 表达相关性并不强,其具体机制仍有待研究。

作为本研究发现的新甲基化位点cg12013757,位于 KRI1 基因区域。该基因是 40S 核糖体生物合成过程中的关键因素,同时也是维持细胞活性的必要基因[17]。KRI1 基因在遗传、功能和生物学机制等方面与 KRR1 基因相互作用,进而形成一个核糖体合成所必需的复合体[18]。有研究显示,当通过实验使 KRI1 的表达关闭时,多核糖体和 40S 核糖体的合成减少,提示 KRI1 的异常表达可能,对细胞活性产生较大影响[19]。目前关于 KRI1 基因的报道较少,其与肿瘤发生发展的相关生物学机制仍不清楚,需要进一步研究。

总之,本研究采用现有的 TCGA 数据库,在全基因组水平对肺腺癌预后相关的甲基化位点进行了初步挖掘,发现一些新的肺腺癌预后相关的甲基化位点,对今后肺癌预后的预测研究有一定的理论指导意义。

[参考文献]

[1] Eberhardt WE, Stuschke M. Multimodal treatment of non-small-cell lung cancer[J]. Lancet, 2015, 386 (9998): 1018-1020

[2] 夏 宁,张 宇,郝可可,等. 血清肿瘤标志物联合检测

诊断肺癌的临床应用研究[J]. 南京医科大学学报(自然科学版),2015,35(12):1784-1786

[3] Feng H,Zhang Z,Wang X,et al. Identification of DLC-1 expression and methylation status in patients with non-small-cell lung cancer[J]. Mol Clin Oncol,2016,4(2):249-254

[4] Guo F,Guo L,Li Y,et al. MALAT1 is an oncogenic long non-coding RNA associated with tumor invasion in non-small cell lung cancer regulated by DNA methylation[J]. Int J Clin Exp Pathol,2015,8(12):15903-15910

[5] Li J,Jia XF,Liu J,et al. Relationship of EGFR DNA methylation with the severity of non-small cell lung cancer[J]. Genet Mol Res,2015,14(4):11915-11923

[6] Feng H,Zhang Z,Qing X,et al. Promoter methylation of APC and RAR-β genes as prognostic markers in non-small cell lung cancer (NSCLC)[J]. Exp Mol Pathol,2016,100(1):109-113

[7] Zhang X,Yang X,Wang J,et al. Down-regulation of PAX6 by promoter methylation is associated with poor prognosis in non small cell lung cancer[J]. Int J Clin Exp Pathol,2015,8(9):11452-11457

[8] Terry MB,Delgado-Cruzata L,Vin-Raviv N,et al. DNA methylation in white blood cells:association with risk factors in epidemiologic studies[J]. Epigenetics,2011,6(7):828-837

[9] Schübeler D. Function and information content of DNA methylation[J]. Nature,2015,517(7534):321-326

[10] Harada H,Miyamoto K,Yamashita Y,et al. Prognostic signature of protocadherin 10 methylation in curatively resected pathological stage I non-small-cell lung cancer[J]. Cancer Med,2015,4(10):1536-1546

[11] 桂　珍, 严　枫. UHRF1 在表观遗传调控和肿瘤诊治中的研究进展[J]. 南京医科大学学报(自然科学版),2016,36(2):129-134

[12] Ghalali A,Wiklund F,Zheng H,et al. Atorvastatin prevents ATP-driven invasiveness via P2X7 and EHBP1 signaling in PTEN-expressing prostate cancer cells[J]. Carcinogenesis,2014,35(7):1547-1555

[13] Guilherme A,Soriano NA,Bose S,et al. EHD2 and the novel EH domain binding protein EHBP1 couple endocytosis to the actin cytoskeleton[J]. J Biol Chem,2004,279(11):10593-10605

[14] Giagtzoglou N,Li T,Yamamoto S,et al. Drosophila EHBP1 regulates Scabrous secretion during Notch-mediated lateral inhibition[J]. J Cell Sci,2013,126 (Pt 16):3686-3696

[15] Gratenstein K,Heggestad AD,Fortun J,et al. The WD-repeat protein GRWD1:potential roles in myeloid differentiation and ribosome biogenesis[J]. Genomics,2005,85(6):762-773

[16] Sugimoto N,Maehara K,Yoshida K,et al. Cdt1-binding protein GRWD1 is a novel histone-binding protein that facilitates MCM loading through its influence on chromatin architecture[J]. Nucleic Acids Res,2015,43(12):5898-5911

[17] Zheng S,Lan P,Liu X,et al. Interaction between ribosome assembly factors Krr1 and Faf1 is essential for formation of small ribosomal subunit in yeast[J]. J Biol Chem,2014,289(33):22692-22703

[18] You KT,Park J,Kim VN. Role of the small subunit processome in the maintenance of pluripotent stem cells[J]. Genes Dev,2015,29(19):2004-2009

[19] Sasaki T,Toh-E A,Kikuchi Y. Yeast Krr1p physically and functionally interacts with a novel essential Kri1p, and both proteins are required for 40S ribosome biogenesis in the nucleolus[J]. Mol Cell Biol,2000,20(21):7971-7979

[收稿日期] 2016-05-10