·技术与方法·

DNA 甲基化数据分析方法和软件应用*

付利娟[1,2]，夏映曦[3]，何俊琳[1]，刘学庆[1]，陈雪梅[1]，王应雄[1]，丁裕斌[1△]
（重庆医科大学：1. 公共卫生学院；2. 中医药学院 400016；3. 重庆江陵医院 400021）

摘 要：目的 分析 DNA 甲基化芯片实验过程质量控制方法、数据统计分析要点及实验结果的验证和数据的可视化处理。**方法** 利用文献、DNA 甲基化实验数据探讨 DNA 甲基化研究中的方法学。**结果** DNA 甲基化芯片初筛异常过程应多在步骤质量控制工作中，包括 DNA 片段化、免疫共沉淀阳性对照的选择、去除原始扫描噪音信号和数据均一化处理。DNA 甲基化芯片的结果可采用常用的甲基化特异性 PCR（MSP）和甲基化测序 PCR（BSP），引物设计软件包括 Methprimer 和 Methyl Primer Express。DNA 甲基化芯片分析数据的可视软件为 Signal map；BSP 结果的可视化可采用 Windows 系统下的执行软件 QUMA 和 BISMA。**结论** DNA 甲基化研究，应从多角度控制实验的设计和数据的产生及结果的分析。

关键词：DNA 甲基化；软件；质量控制；数据分析
doi：10.3969/j.issn.1671-8348.2012.17.018 **文献标识码：**A **文章编号：**1671-8348（2012）17-1719-03

Data analysis and its analytical softs application on DNA methylation in tumor research*

Fu Lijuan[1,2]，*Xia Yingxi*[3]，*He Junlin*[1]，*Liu Xueqing*[1]，*Chen Xuemei*[1]，*Wang Yingxiong*[1]，*Ding Yubin*[1△]
（1. *School of Public Health*；2. *College of Chinese Traditional Medicine*，*Chongqing Medical University*，*Chongqing* 400016，*China*；3. *Chongqing Jiangling Hospital*，*Chongqing* 400021，*China*）

Abstract：Objective To analyze the quality control method，data analytic key points，results confirmation and data visualization processing during the experimental process of DNA methylation chip. **Methods** We used published paper and our original research data to explore the methods used in DNA methylation analysis. **Results** The quality control includes DNA segmentation，positive control selection，noise signal removing and data normalization. Mehtylation specific PCR（MSP）and bisulfite sequencing PCR（BSP）were needed in validation of MeDIP-Chip array results. Methprimer and Methyl Primer Express were used for primer designing. MeDIP-Chip array results were visualized by signal map and BSP result could be analyzed by QUMA and BISMA**Conclusion** In DNA methyaltion research，multifactorial quality control in MeDIP-chip，design and data analysis is necessary.

Key words：DNA methylation； software； data analysis； quality control

DNA 甲基化作为一种重要的表观修饰方式，它可在不改变基因序列的情况，调控基因的转录，近年来已成为生命研究的热点之一[1]。DNA 甲基化一旦发生紊乱，可导致包括肿瘤、胚胎发育、老年化进程以及自身免疫性在内的多种疾病状态[2]。由于 CpG 岛甲基化所致的抑癌基因转录失活是一个可逆转的基因修饰过程，且该逆转过程（CpG 岛去甲基化）可直接恢复抑癌基因功能，因此，DNA 去甲基化调控抑癌基因功能的研究已成为肿瘤基因治疗的新型手段之一[3]。DNA 甲基化的研究手段多样，其中，DNA 甲基化芯片属高通量、高效率的研究手段之一[4]，在 DNA 甲基化研究中应用非常广泛，对研究者的要求亦较高。从 DNA 甲基化基因芯片设计、芯片数据的质量控制、后期的数据分析、数据的 DNA 甲基化特异性 PCR、COBRA、BSP 测序等验证方法到数据的可视化显示，需要研究者熟悉诸多软件的使用。本研究将 DNA 甲基化研究中的质量控制、数据分析过程以及常用的软件使用予以介绍，并探讨这些数据分析过程中应注意的地方。

1 材料与方法

1.1 材料 DNA 甲基化原始芯片数据，甲基化测序 PCR（bisulfite sequence PCR，BSP）数据，分析所需各种在线、本地安装软件，如 Signal Map、UCSC Genome Browser、Methyprimer、Methyl Primer Express 等。

1.2 方法 采用文献学习及软件学习法，分析实验过程中质量控制的必要方法，统计分析各种实验数据，进行引物设计以及研究数据的可视化处理等。

2 结 果

2.1 芯片的设计与质量控制 目前常用的商业 DNA 甲基化芯片主要由 Roche-nimblegen 和 Agilent 两个公司生产。芯片包括 Chip-on-Chip 和 MeDIP-Chip 芯片，根据实验设计的需要，可选择不同的类型。这两种较常用的甲基化芯片类型包括多种不同分辨率的芯片，芯片杂交的探针既可囊括基因组 CpG 区和启动子区，亦可专门针对启动子区的 DNA 甲基化。以 MeDIP-Chip 芯片为例，整个 DNA 甲基化芯片实验应包括如下质控步骤：（1）超声打断基因组产生的片段应在 200～1 000 bp范围内；（2）甲基免疫共沉淀过程质控应选择明确的甲基化区域，如印记基因 Xist 做阳性对照，同时选择如 Actb，Aprt 等基因作为非甲基化区域的对照；（3）通过对基因芯片扫描的原始数据进行分析，校正异常杂交信号，去除噪音信号，并通过对信号点（MA-plot）的分布明确信号值的均一性，进一步采用相关分析判断重复实验的再现性和配对样本间的相关性；（4）数据分析过程质量控制，首先要进行数据的均一化处理以

* **基金项目：**重庆市生殖健康与出生缺陷重点实验室开放课题（0801）；重庆市科委项目（CSCT2009BB5271）。 △ **通讯作者，**Tel：13220293739；E-mail：dingyb@gmail.com。

表 1 MethyPrimer 设计的 ALKBH3 甲基化 PCR 引物

基因 ALKBH3	引物序列	退火温度(℃)	片段大小(bp)
甲基化引物(M)	F:5′-ATT ATT CGG ATT GAG GAT TGC-3′	63.0	125
	R:5′-GAA ACC TTA AAA ATA AAA CAC CGA C-3′		
非甲基化引物(U)	F:5′-GGA TTA TTT GGA TTG AGG ATT GT-3′	63.2	128
	R:5′-CAA AAC CTT AAA AAT AAA ACA CCA AC-3′		

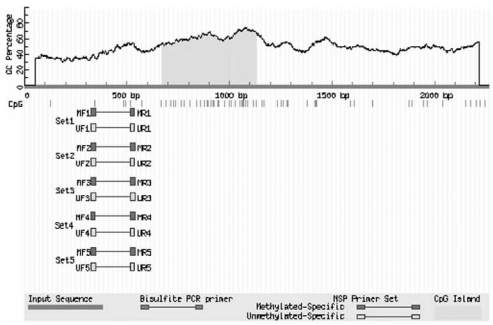
表 2 Methyl Primer Express 设计的 ALKBH3 甲基化 PCR 引物

基因 ALKBH3	引物序列	退火温度(℃)	片段大小(bp)
甲基化引物(M)	F:5′-TGA TTA GGT TTT TTA GGC GC-3′	60.36	173
	R:5′-TCC GCA ATC TAT AAT CGA AAC-3		
非甲基化引物(U)	F:5′-GGT GAT TAG GTT TTT TAG GTG T-3′	57.50	173
	R:5′-TCC ACA ATC TAT AAT CAA AAC CT-3′		

判断出不同芯片间的 DNA 甲基化差异,其次是对明确的区域和整个基因组的差异甲基化区域进行判别,这一过程在 Roche-Nimblegen 中主要由 NimbleScan v2.5 软件完成[5]。

2.2 甲基化数据分析 DNA 甲基化芯片数据结果,除了可进一步进行统计学分析外,差异甲基化基因启动子或 CpG 岛的可视化,如 Roche-Nimblegen 公司的数据可采用 Signal Map 进行阅读,即导入注释数据和 GFF 格式的 Peak 数据和 log2IP/input 数据后,可根据 NimbleScan 输出的统计结果,查找差异 DNA 甲基化基因的位置、大小、转录起始与终止区域、TSS 点以及 Log2IP/input 值(图 1)[6]。

图 1 甲基化数据分析图



深色部分,分别位于 672～996 bp 区域和 1 001～1 131 bp 区域。

图 2 Methyprimer 预测出 ALKBH3 基因的两个 CpG 岛图

2.3 MSP 引物设计

2.3.1 基因的外显子区查找 可在 University of California, Santa Cruz 分校的 UCSC Genome Browser 数据库(http://genome.ucsc.edu/cgi-bin/hgGateway)搜索[7]。除了搜索启动子区,研究者还可以根据目的基因甲基化所在位置,选择 5′-UTR 和外显子区。具体搜索的方法及限制,可使用搜索引擎搜索如下关键词“UCSC 启动子查找”。应注意的是,UCSC Genome Browser 注释数据库有 hg16、17、18 和 19 版,在搜索时,应注意选择搜索的数据库版本与 DNA 甲基化芯片数据的注释数据库版本相对应。除了 UCSC 数据库外,NCBI 的 Mapview(http://www.ncbi.nlm.nih.gov/mapview/index.html)亦可以搜索启动子区。搜索引擎的选择,通常是根据芯片结果注释时所采用的数据库来决定的。更多的情况下,芯片注释使用的数据库是 UCSC Genome Browser。

2.3.2 引物设计软件 甲基化芯片结果验证最常用的方法是甲基化 PCR(methylation specific PCR,MSP)和硫化测序 PCR(bisfulfite sequencing PCR,BSP)。甲基化引物设计是 MSP 和 BSP 中的关键。研究者最常用的甲基化引物设计软件是在线 Methyprimer(http://www.urogene.org/methprimer/index1.html)[8]。研究者可将已知的启动子区拷贝到该软件的窗口后,选择 CpG 岛的大小、限制 GC 含量等限制条件后,即可自行设计 MSP 或 BSP 引物。通常情况下 Methyprimer 会在 CpG 岛区域设计引物,但有些基因的引物设计结果却并不在软件预测的 CpG 岛区(图 2),如 Alkylation Repair Homolog 3(ALKBH3)基因。将该基因启动子区、5′ UTR 区和 CDs 区序列后拷贝到 Methyprimer 后,软件预测出了两个 CpG 岛,分别位于 672～996 区域和 1 001～1 131 区域,设计出的 5 对 MSP 引物均全部位于 325～542 区域内,而非 CpG 岛区域。因此,这类基因引物的设计就需要研究者先根据自己的知识经验来限定 CpG 岛区,再依据甲基化引物设计的要求自行设计。DNA 甲基化引物设计的原则主要有:(1)引物扩增区域最好位于转录起始位点(transcription start site,TSS)250 bp 以内;(2)引物至少应包括 3 个以上(多数情况下 4 个或更多)CpG;(3)预测的退火温度大于 55 ℃[9]。根据上述要求,设计的 ALKBH3 基因引物见表 1。令一款由 Applied Biosystems 公司开发的免费软件 Methyl Primer Express (https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=602121&tab=Overview)[10],可本地安装后使用。该软件进行 CpG 岛预测后,能够准确地设计出位于 CpG 岛区域内的引物及其扩增区(图 3)。引物设计时,软件还会提

醒使用者选择哪个 CpG 岛来设计引物，设计出的引物与 Methprimer 人工设定的区域很接近。这个软件比较简单易用，推荐初学者使用这一软件。熟练者，可将二者结合使用。利用 Methyl Primer Express 设计 ALKBH3 MSP 引物（表 1、2）。设计好的甲基化引物可通过 Blast（http://medgen.ugent.be/methBLAST/）进一步验证，确保其目标扩增序列的特异性。此外，Ugent 网站 http://medgen.ugent.be/methprimerdb/search_primers.php 为研究者提供了部分基因甲基化启动子序列，这些序列均是被研究者实验过程所验证的引物。

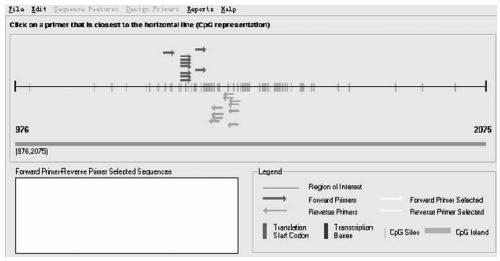


图 3　Methyl Primer Express 预测 CpG 岛和设计的甲基化引物起始位点图

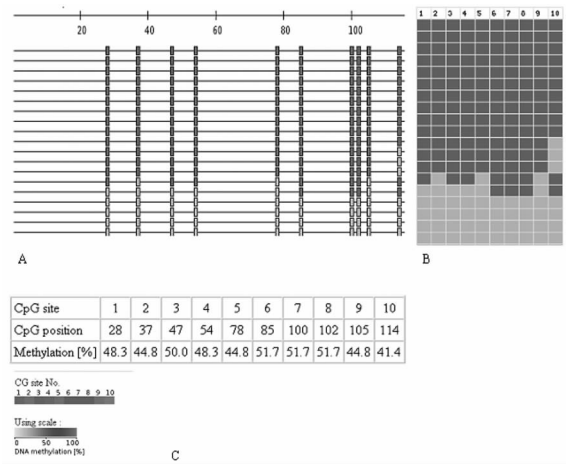
2.4　BSP 结果的可视化　目前，多款软件被用于 BSP 结果的可视化和 CpG 甲基化位点的统计分析，包括 BiQ analyzer，BISMA（Bisulfite Sequencing DNA Methylation Analysis）和 QUMA Quantification Tool for Methylation Analysis。BiQ analyzer 可视化分析 CpG 位点功能较强，但是在 CpG 甲基化与非甲基化的模式作图和甲基化数据分析上有明显不足[11]。BISMA 和 QUMA 在 CpG 差异甲基化作图与数据分析上各具优势。

2.4.1　QUMA（http://quma.cdb.riken.jp/）　QUMA 是一款使用方便、集成多个分析功能、基于网络的 CpG 甲基化测序结果分析软件，它可以整齐地排列测序的原始结果，分析甲基化图谱，并进行统计学比较，检验测序质量以及现实可视化的甲基化模式[6]。利用该网站提供样本数据，分析结果见图 4。

Sequence name	mismatch (gap) / alignment length (% identity)	Me-CpG (%)	unconverted (% converted)	Methylation pattern (or reason for the exclusion
Gm9_J1_seq_10	1 (1) / 453 (99.8)	6 (31.6)	2/131 (98.5)	○○○○○○○○○○○●○●●●●●
Gm9_J1_seq_03	0 (0) / 453 (100.0)	6 (31.6)	0/131 (100.0)	●●○○○○○○○●○○○○●●●●
Gm9_J1_seq_14	0 (0) / 453 (100.0)	8 (42.1)	0/131 (100.0)	●●●○○○○○○○●○○○○●●○
Gm9_J1_seq_12	2 (2) / 454 (99.6)	9 (47.4)	0/130 (100.0)	●○●●○○○○○○●○○○○●●○
Gm9_J1_seq_04	0 (0) / 453 (100.0)	12 (63.2)	0/131 (100.0)	●○○○●○○○●○●●●●●●●●
Gm9_J1_seq_16	2 (0) / 453 (99.6)	15 (78.9)	1/131 (99.2)	●●●●●●○●●●●●●○●○●●
Gm9_J1_seq_08	0 (0) / 453 (100.0)	16 (84.2)	0/131 (100.0)	●●●●●●●○●●●●○●●●○●
Gm9_J1_seq_01	1 (0) / 453 (99.8)	18 (94.7)	0/130 (100.0)	●●○●●●●●●●●●●●●●●●
Gm9_J1_seq_07	1 (1) / 453 (99.8)	18 (94.7)	0/130 (100.0)	●●○●●●●●●●●●●●●●●●
Gm9_J1_seq_11	1 (0) / 453 (99.8)	18 (94.7)	1/131 (99.2)	●●○●●●●●●●●●●●●●●●
Gm9_J1_seq_05	3 (0) / 453 (99.3)	18 (100.0)	2/131 (98.5)	●●×●●●●●●●●●●●●●●●
Gm9_J1_seq_02	1 (0) / 453 (99.8)	19 (100.0)	1/131 (99.2)	●●●●●●●●●●●●●●●●●●
Gm9_J1_seq_13	1 (1) / 454 (99.8)	19 (100.0)	1/131 (99.2)	●●●●●●●●●●●●●●●●●●

图 4　QUMA 图示分析甲基化测序结果

2.4.2　BISMA（http://biochem.jacobs-university.de/BDPC/BISMA/）　BISMA 是一款目前功能更为全面，可视化效果和统计学数据分析，最优秀的 DNA 甲基化测序数据可视化分析软件。它可快速抽取上传的 txt 或 ABI 测序格式的原始数据文件，辅助分析序列方向，高度自动化的进行复杂计算，去除载体序列，结果分析快速准确。同时还可判别测序结果的质量、亚硫酸盐转化效率、检测碱基缺失或丢失和过滤 N 位的甲基化。在质量控制和数据处理能力较高的情况下，分析并展示 CpG 甲基化模式，并首次在同类软件中支持重复序列的分析[7]，见图 5。



A：CpG 甲基化位点的核酸位置；B：基因的甲基化模式；C：每个 CpG 位点受到甲基化比例。

图 5　BISMA 分析甲基化测序数据统计结果

3　讨　论

DNA 甲基化是生命活动过程中常见的表观遗传修饰方式之一[12-14]。DNA 甲基化异常分两种类型，一种是 CpG 岛超甲基化（hypermethylation），另一种是低甲基化（hypomethylation）[15]。DNA 甲基化异常与许多种类型的疾病发生相关，营养、环境因素同样可影响 DNA 甲基化状态[2]。此外，由于 DNA 甲基化的异常状态是一种可逆转的生物学行为，因此，DNA 甲基化研究成为目前各中疾病发生研究领域的热点之一[16]。DNA 甲基化研究，无论是针对某个生理过程还是疾病发生的机制探索，均是系统工作，需从甲基化芯片设计开始，到 MSP、BSP 等验证，甚至还包括功能实验验证等方面，进行周详的设计与计划。尤其是在芯片实验过程中，多个涉及的质量控制过程的步骤尤为重要，事关整个实验的成败，因此，应在芯片实验的整个过程执行严格的质量控制工作[5]。DNA 甲基化芯片的验证过程主要包括 MSP 和 BSP，引物的设计亦是实验中的关键，选择更好的软件进行引物的设计，并优化设计好的引物和 PCR 反应条件是实验成功的前提。BSP 结果的可视化，有助于读者更直观地了解甲基化测序验证结果。因此，DNA 甲基化研究，应从多角度控制实验的设计和数据的产生及结果的分析。

参考文献：

[1] Feng S, Jacobsen SE, Reik W. Epigenetic reprogramming in plant and animal development[J]. Science, 2010, 330(6004): 622-627.

[2] Morgan HD, Santos F, Green K, et al. Epigenetic reprogramming in mammals[J]. Human Molecular Genetics, 2005, 14(1): R47-58.

[3] Rodenhiser D, Mann M. Epigenetics and human disease: translating basic biology into clinical applications[J]. Canadian Medical Association journal, 2006, 174(3): 341-348.

[4] Dunwell T, Hesson L, Rauch TA, et al. A Genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers[J]. Molecular Cancer, 2010, 9: 44.

（下转第 1726 页）

参考文献：

[1] Phillips LR, Crist J. Social relationships among family caregivers: a cross-cultural comparison between Mexican Americans and non-Hispanic White caregivers[J]. J Transcult Nurs, 2008, 19(4): 326-337.

[2] Lee CC, Czaja SJ, Schulz R. The moderating influence of demographic characteristics, social support, and religious coping on the effectiveness of a multicomponent psychosocial caregiver intervention in three racial ethnic groups [J]. Journal of Gerontology: Psychological Sciences, 2010, 65B(2): 185-194.

[3] Vitacca M, Isimbaldi C, Mainini A, et al. The frail person and his caregiver: cure, care or simultaneous care? A conceptual article[J]. Journal of Medicine and the Person, 2011(9): 6-12.

[4] Statutory SC. Social workers: stress, job satisfaction, coping, social support and individual differences[J]. British Journal of Social Work, 2008, 38: 1173-1193.

[5] 张作记. 行为医学量表手册[J]. 中国行为医学科学, 2001, 特刊: 41-42.

[6] 陈向明. 质的研究方法与社会科学研究[M]. 北京: 教育科学出版社, 2000: 171, 96, 401-408.

[7] Nancy B, Susan KG. The practice of nursing research [M]. 3 rd Edition. WB Saunders Company, 1998: 542-547.

[8] 丛丽, 赵光红. 护理人员核心能力与社会支持相关性研究[J]. 中国护理管理, 2010, 10(12): 21-23.

[9] 仲稳山, 陈武英, 蔡蕾. 护士长社会支持与应对方式研究[J]. 精神医学杂志, 2009, 22(2): 135-137.

[10] Brehm SS. 亲密关系[M]. 郭辉, 译. 北京: 人民邮电出版社, 2005: 325.

[11] Almeida J, Subramanian SV, Kawachi I, et al. Is blood thicker than water? Social support, depression and the modifying role of ethnicity/nativity status[J]. J Epidemiol Community Health, 2011, 65(1): 51-56.

[12] 张杉杉, 李敬雅. 城市低保人员的社会支持系统分析[J]. 人口与经济, 2011(1): 51-56.

[13] Schwartz SJ. The applicability of familism to diverse ethnic groups: a preliminary study[J]. J Soc Psychol, 2007, 147(1): 101-118.

[14] Dunn MG, O'Brien KM. Psychological Health and Meaning in Life Stress, Social Support, and Religious Coping in Latina/Latino Immigrants[J]. Hispanic Journal of Behavioral Sciences, 2009, 31(2): 204-227.

[15] 宁艳花, 张琳, 姚丽, 等. 银川市老年人社会支持状况的调查[J]. 中国老年学志, 2011, 31(3): 487-490.

(收稿日期：2011-10-09 修回日期：2011-11-22)

(上接第 1721 页)

[5] Pölmke N, Santacruz D, Walter J. Comprehensive analysis of DNA-methylation in mammalian tissues using MeDIP-chip[J]. Methods, 2010, 53(2): 175-184.

[6] Sörensen AL, Jacobsen BM, Reiner AH, et al. Promoter DNA Methylation Patterns of Differentiated Cells Are Largely Programmed at the Progenitor Stage[J]. Molecular Biology of the Cell, 2010, 21: 2066-2077.

[7] Zhu JC, Sanborn JZ, Benz S, et al. The UCSC Cancer Genomics Browser[J]. Nature Methods, 2009, 6: 239-240.

[8] Li J, Gao F, Li N, et al. An improved method for genome wide DNA methylation profiling correlated to transcription and genomic instability in two breast cancer cell lines [J]. BMC Genomics, 2009, 10: 223.

[9] Shames DS, Girard L, Gao B, et al. A Genome-Wide Screen for Promoter Methylation in Lung Cancer Identifies Novel Methylation Markers for Multiple Malignancies[J]. PLoS Med, 2006, 3(12): 2244-2263.

[10] Okamoto J, Hirata T, Chen Z, et al. EMX2 is epigenetically silenced and suppresses growth in human lung cancer [J]. Oncogene, 2010, 29(44): 5969-5975.

[11] Bock C, Reither S, Mikeska T, et al. BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing[J]. Bioinformatics, 2005, 21(21): 4067-4068.

[12] Kumaki Y, Oda M, Okano M. QUMA: quantification tool for methylation analysis [J]. Nucleic Acids Research, 2008, 36: W170-175.

[13] Rohde C, Zhang YY, Reinhardt R, et al. BISMA -Fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences [J]. BMC Bioinformatics, 2010, 11: 230.

[14] Pulukuri SM, Patibandla S, Patel J, et al. Epigenetic inactivation of the tissue inhibitor of metalloproteinase-2 (TIMP-2) gene in human prostate tumors[J]. Oncogene, 2007, 26: 5229-5237.

[15] Cindy D, Davis, Eric O, et al. DNA Methylation, Cancer Susceptibility, and Nutrient Interactions [J]. Exp Biol Med, 2004, 229: 988-995.

[16] Ramchandani S, Bhattacharya SK, Cervoni N, et al. DNA methylation is a reversible biological signal[J]. Proc Natl Acad Sci USA, 1999, 96(11): 6107-6112.

(收稿日期：2011-10-09 修回日期：2011-11-22)