遗传 HEREDITAS (Beijing) 2013 年 6 月, 35(6): 685－694
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2013.00685

基因组规模 DNA 甲基化测序数据处理及其表观遗传分析

王庭璋, 单杲, 徐建红, 薛庆中

浙江大学农业与生物技术学院, 杭州 310058

摘要: 鉴定 DNA 甲基化胞嘧啶(mC)并能制作基因组规模甲基化图谱的新方法——BS-Seq, 最近已被开发, 它是基于新一代高通量测序结合 DNA 亚硫酸氢盐转换技术, 不仅可以从基因组规模洞察不同生物之间在 DNA 甲基化水平和模式上的差异, 也能从不同基因组区域, 包括基因、外显子、重复序列等方面, 阐明 DNA 甲基化环境和核苷酸偏好上的保守性, 加深理解 DNA 胞嘧啶(C)甲基化在调控基因表达和沉默转座子等重复序列中所起的表观遗传学影响。文章举例介绍了 DNA 甲基化位点数据预处理的具体步骤, 通过处理分别将参考序列中的胞嘧啶(C)替换成胸腺嘧啶(T), 鸟嘌呤(G)替换成腺嘌呤(A), 而将读序列中的胞嘧啶(C)替换为胸腺嘧啶(T)。文章综述了全基因组 DNA 甲基化分析的主要内容, 包括: (1)不同序列环境下的胞嘧啶甲基化; (2)全基因组上的甲基化的分布情况; (3)DNA 甲基化环境和核苷酸的偏好; (4)DNA-蛋白质互作位点上的 DNA 甲基化; (5)不同基因结构元件的胞嘧啶甲基化程度。DNA 甲基化分析技术为研究不同物种的表观基因组, 环境和表观互作提供了强大的工具, 并为进一步发展人体疾病诊断和治疗方法提供理论基础。

关键词: 新一代测序; DNA 甲基化; BS-Seq; 数据处理; 表观遗传学

Genome-scale sequence data processing and epigenetic analysis of DNA methylation

WANG Ting-Zhang, SHAN Gao, XU Jian-Hong, XUE Qing-Zhong

College of Agriculture and Biotechnology, *Zhejiang University*, *Hangzhou* 310058, *China*

Abstract: A new approach recently developed for detecting cytosine DNA methylation (mC) and analyzing the genome-scale DNA methylation profiling, is called BS-Seq which is based on bisulfite conversion of genomic DNA combined with next-generation sequencing. The method can not only provide an insight into the difference of genome-scale DNA methylation among different organisms, but also reveal the conservation of DNA methylation in all contexts and nucleotide preference for different genomic regions, including genes, exons, and repetitive DNA sequences. It will be helpful to under-

收稿日期: 2012–10–24; 修回日期: 2012–12–27
基金项目: 国家重点基础研究发展计划(973 计划)项目(编号：2010CB126205)和国家自然科学基金项目(编号：31171165)资助
作者简介: 王庭璋, 博士后, 研究方向：生物信息学。Tel: 0571-88982406; E-mail: wtzhzhtw@gmail.com
通讯作者: 徐建红, 博士, 研究员, 研究方向：基因组学与分子生物学。E-mail: jhxu@zju.edu.cn
薛庆中, 教授, 研究方向：基因组学, 遗传学。E-mail: xueqingzhong@hotmail.com
网络出版时间: 2013-2-27 9:19:28
URL: http://www.cnki.net/kcms/detail/11.1913.R.20130227.0919.001.html

stand the epigenetic impacts of cytosine DNA methylation on the regulation of gene expression and maintaining silence of repetitive sequences, such as transposable elements. In this paper, we introduce the preprocessing steps of DNA methylation data, by which cytosine (C) and guanine (G) in the reference sequence are transferred to thymine (T) and adenine (A), and cytosine in reads is transferred to thymine, respectively. We also comprehensively review the main content of the DNA methylation analysis on the genomic scale: (1) the cytosine methylation under the context of different sequences; (2) the distribution of genomic methylcytosine; (3) DNA methylation context and the preference for the nucleotides; (4) DNA- protein interaction sites of DNA methylation; (5) degree of methylation of cytosine in the different structural elements of genes. DNA methylation analysis technique provides a powerful tool for the epigenome study in human and other species, and genes and environment interaction, and founds the theoretical basis for further development of disease diagnostics and therapeutics in human.

Keywords: next-generation sequencing (NGS); DNA methylation; BS-Seq; data processing; epigenetics

生物体因 DNA 分子甲基化(DNA methylation)所引起的遗传特性变化，属于一种重要的表观遗传标记(Epigenetic marker)。最近证实，DNA 甲基化的维护和组蛋白修饰存在关联[1]。甲基化过程中其 DNA 序列并没有发生改变，但在胞嘧啶(C)上添加了甲基。这种遗传修饰不仅普遍存在于哺乳动物细胞 CpG 环境中，并且在多能胚胎干细胞和植物细胞的非 CpG 对称的环境中也会发生(如 CHG 和 CHH，其中 H 代表 A、C 或者 T)[2~4]。研究表明，DNA 甲基化不仅对于人类和哺乳动物的发育和疾病等方面产生了至关重要的影响，最近发现，植物杂种优势的分子机理也与 DNA 甲基化关联[5]。Shen 等[6]指出植物杂种优势可能是由于杂种 F_1 基因组的 DNA 甲基化水平增加，改变了植物昼夜节律所致。因此，深入研究甲基化的途径和揭示其调控机制一直是表观遗传学研究热门的话题。生物体内许多内源性基因，无论是启动子区域或转录区域内均会产生甲基化，它与转录水平高度相关，但和基因表达往往不是直接相关的[2~4]；精确辨认 DNA 甲基化模式非常复杂困难，特别是基因组规模 DNA 甲基化分析一直未能完美解决。直到 2006 年 Zhang 等[2]绘制出拟南芥(*Arabidopsis thaliana*)甲基化胞嘧啶单碱基分辨率图谱，从中可以精确地测量全基因组范围内胞嘧啶甲基化的组成与分布，研究各种 DNA 甲基化突变体对全基因组甲基化模式的影响，发现植物体内胞嘧啶的甲基化表现为 CpG(或 CpNpG)、CpNpNp 形式(C 与 G 分别是胞嘧啶和鸟嘌呤，p 是磷酸根，N 是任意的碱基)。

随着新一代测序技术的开发，使用亚硫酸氢盐处理 DNA 结合鸟枪法测序已成为研究胞嘧啶 DNA 甲基化的新方法，被称为 BS-Seq[7]。例如，Meissner 等[3]研究了哺乳动物细胞基因组规模的 DNA 甲基化谱，指出 DNA 甲基化模式与组蛋白的甲基化模式具有更紧密的相关性，CpG 岛甲基化在细胞分化过程中会发生广泛的变化，它是动态的表观遗传标记。Lister 等[8]发表了第一张哺乳动物基因组的全基因组单碱基甲基化胞嘧啶的分辨率图谱，指出在人类细胞内，约 1%的 DNA 碱基会受到甲基化。在成熟体细胞组织中，较易发生 CpG 甲基化；而胚胎干细胞中则以非 CpG 甲基化(即 CHG 和 CHH)较常见。目前在许多动植物物种上，也已应用 BS-Seq 方法制作出单碱基分辨率水平的甲基化图谱。

本文简介了 DNA 甲基化检测的主要方法，通过举例较详细地介绍了甲基化测序数据处理的基本步骤，并对已报道的不同生物甲基化图谱数据分析结果及其生物学意义加以综述。

1 DNA 甲基化检测的方法

全基因组 DNA 甲基化分析的关键是区分甲基化和非甲基化的胞嘧啶，它主要基于以下 3 种方法：①酶切法：用甲基化敏感的限制性内切酶(如 *Hpa* Ⅱ、*MSP* Ⅰ、*Mcr*BC 等)酶切(Digestion with methylation-sensitive restriction enzymes)，可以酶切未甲基化的 DNA 片段，从而使基因组内的甲基化 DNA 片段得以富集。当 CpG 位点没有被甲基化时，酶切割 DNA 和随后的 PCR 扩增废止。但是，如果 CpG 位

点被甲基化，酶就不能切割和通过 PCR 扩增 DNA 链。因此 PCR 条带的存在或不存在与在一个特定 CpG 部位甲基化的存在或不存在关联。②免疫沉淀法：亲和纯化(Affinity purification)，使用抗体对甲基化胞嘧啶，甲基结合域(MBD)或其他的蛋白质结构域进行免疫沉淀，使基因组 DNA 甲基化或未被甲基化馏分[9]。③亚硫酸氢盐转化(Bisulfite conversion)法，先将基因组 DNA 片段变性，然后用亚硫酸氢盐处理，可以将其中未甲基化的胞嘧啶(Cytosine, C)转换成尿嘧啶(Uracil, U)，再通过 PCR 技术扩增后把尿嘧啶转换成胸腺嘧啶(Thymine, T)。相反，未转化的甲基胞嘧啶，最终以胞嘧啶形式被检测到。值得注意的是，亚硫酸氢盐转化后，DNA 链不再是互补链，由于亚硫酸氢盐转化和高通量测序技术结合可以将确定的甲基化图谱分辨率[10]提升到单碱基水平，此外，通过引物设计可以检测特异链的甲基化状态[11]。因此，亚硫酸氢盐转化被视为辨认任何 DNA 序列的胞嘧啶甲基化状态的“金标准”，BS-Seq 成为目前最有用和最广泛使用的 DNA 甲基化分析技术[12]。BS-Seq 文库的产生包括以下 4 个步骤：基因组 DNA 片段的获得、甲基化测序接头连接、凝胶纯化、亚硫酸氢盐转化和 PCR 扩增。

2　BS-Seq 数据处理

通过基因组 DNA 片段的获得、甲基化测序接头连接、凝胶纯化、亚硫酸氢盐(BS)转化和 PCR 扩增等 4 个步骤，获得 BS-Seq 文库。

2.1　读序列和参考基因组数据处理流程

构建文库后，可以应用 Illumina/Solexa，或者 Roche/454 测序仪进行高通量测序。

2.2　读序列和参考基因组的获取

由 Illumina 分析流程(pipeline)产生的 BS-Seq 读序列(reads)，其格式为 fastq。科研人员一般会申请将测序仪每次运行的所获的原始读序列数据整体提交到 NCBI 的 SRA 数据库保存，每个数据都有其登录号(Accession Number)以便查询和索取。用户通过网址(ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRRxxx/SRRxxxyyy/SRRxxxyyy.sra)输入登录号，即可获取读序列数据。例如，从 SRA 数据库获得两个水稻组织甲基化实验数据：SRR059000 ~ SRR059009[13]，在 SRRxxxyyy 数据中，SRR 为登记号，(SRR059000)，前面 3 位数字 059 为(xxx)来自于水稻胚(embryo)组织，后面 3 位数字 000 至 009 等 10 个数字串(yyy))均取自胚乳(Endosperm)组织。

从 Internet 网络公共数据库或特定物种数据库中可获取上已发布或已完成测序物种的全基因组序列，它们通常作为基因组研究的参考序列使用。例如，水稻基因组参考序列版本 7.0[14] 可从网址(ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/)获取。

2.3　读序列和参考基因组的序列替换

对于由 BS-Seq 方法产生的 DNA 甲基化的原始读序列，需要进行以下 3 个步骤的处理。①删除读序列中低质量碱基(PHRED 记分≤2)之后的所有碱基。②搜索并删除读序列两端的接头寡核苷酸。③将读序列中的胞嘧啶碱基(C)替换为胸腺嘧啶(T)(图 1)。

同时，还要对参考基因组序列进行两种对应的处理。一是将参考序列中的胞嘧啶(C)替换成胸腺嘧啶(T)，如图 1 中 5′-ATCG-3′替换为 5′-ATTG-3′；二是将参考序列中的鸟嘌呤(G)替换成腺嘌呤(A)，图 1 中 5′-ATCG-3′则替换为 5′-ATCA-3′，其反向互补序列则为 5′-TGAT-3′。

通过上述生物信息技术处理，就能获得全基因组甲基化位点及其甲基化水平的信息。

2.4　读序列在参考基因组上的作图

经过上述 BS 转化后，源于 Watson 链(正义链)的读序列都被作图到转化后无胞嘧啶的参考序列上，而来自于 Crick 链(反义链)的读序列则定位到无鸟嘌呤的参考序列上(图 1)。至于将很短的读序列定位到极长的参考序列上，则需要借助 Bowtie 软件[15]。基因组某个区域可能出现多个相同或不同的读序列，读序列出现次数(即覆盖度)通常不应低于 10×。当对同一份实验材料进行两次独立实验时，为提高位点上的覆盖度，可将两次实验数据整合并分析。

2.5　甲基化胞嘧啶位点的识别

亚硫酸氢盐转化效率以及测序的错误会对胞嘧

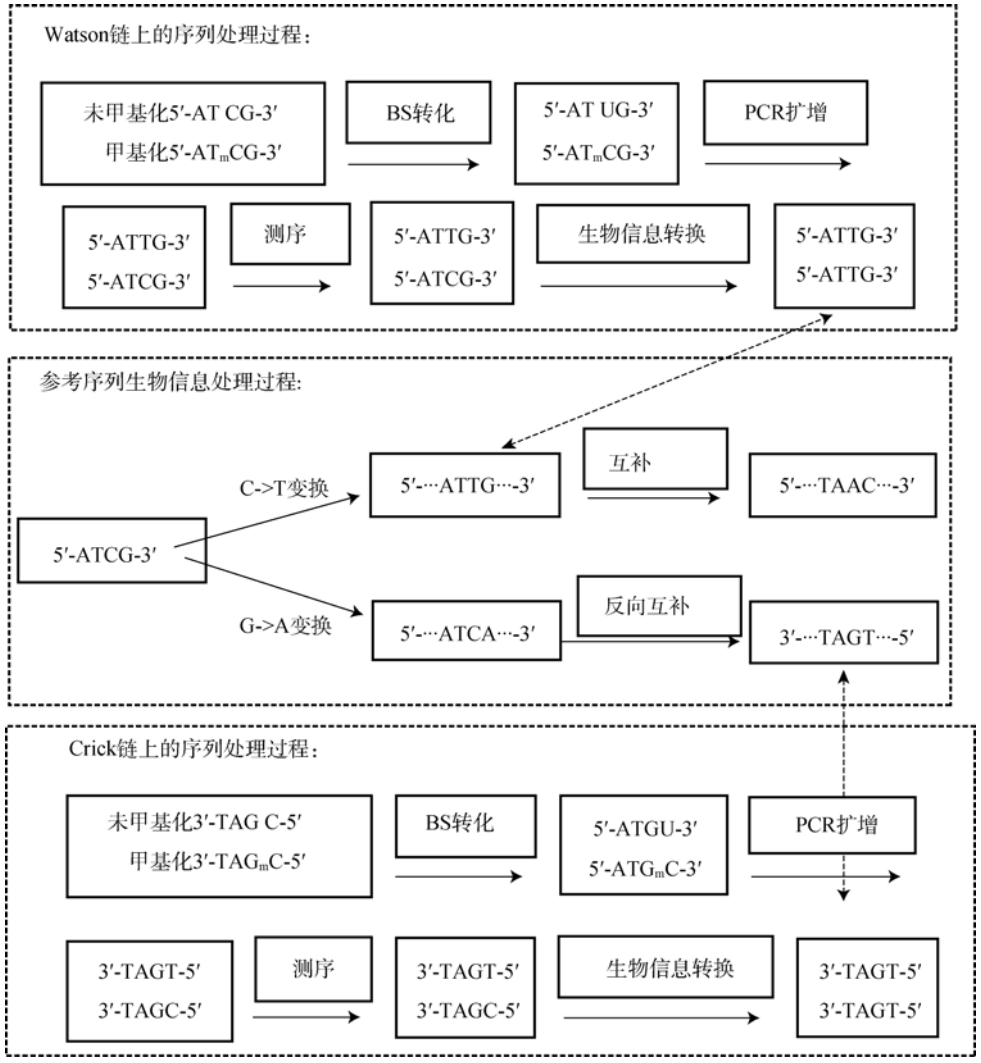

图 1　BS-Seq 测序和数据处理流程图

实线箭头表示信息处理的主要步骤，虚线双箭头表示读序列和参考序列的比对。Watson 链上产生的读序列(未甲基化的 5′-ATCG-3′和甲基化的 5′-ATmCG-3′)通过 C→T 变换后，都以 5′-ATTG-3′形式定位到参考序列上。Crick 链上序列(未甲基化的 5′-CGAT-3′和甲基化的 5′-mCGAT-3′)经 C→T 变换后，也都以 5′-TAGT-3′定位到参考序列的反向互补序列上。mC 表示甲基化细胞嘧啶。

啶位点的甲基化识别产生影响，因此，需要保证目标区域具有足够的测序深度。如已发布的全基因组甲基化图谱中，人[8]和拟南芥[7]的基因组覆盖倍数分别超过了 14 和 15 倍。如要获得更准确的数据，则需采用更复杂的模型，如 Lister 等[8]使用了二项分布(Binomial distribution)模型(B(n, p))，这里，n 是二项分布中的实验次数，表示读序列深度(Read depth，即覆盖该位点的读序列总数)，ρ是测序胞嘧啶出现的概率，表示非甲基化 Lambda 基因组上实际测得的胞嘧啶位点数，并采用 1%错误率(False discovery rate, FDR)对其进行矫正。通过二项分布模型估算出每个参考序列位点上的最少胞嘧啶数。当目标区域位点上测得的胞嘧啶数量多于最少胞嘧啶值时，就被视为甲基化位点。

2.6　胞嘧啶甲基化绝对含量(mC)和相对含量(mC/C)的估算

甲基化水平通常用含甲基化胞嘧啶的读序列数占覆盖对应位点上所有读序列的百分比计算。因此，对于特定的胞嘧啶位点而言，0 表示不存在甲基化，100%则表示该位点完全被甲基化，0~100%之间则表示被甲基化的程度。比较不同区域内的甲基化水平，

则需统计该区域内所有胞嘧啶位点甲基化水平的平均值。计算基因组胞嘧啶的甲基化含量时，通常将目标区域从 5′端到 3′端划分为适当数目的框(bin)，小框大小设为 100 bp。绝对甲基化含量(mC)的计算方法是将所有甲基化类型(mCG, mCHG 或 mCHH)的总数除以小框的大小，即区域的长度；相对甲基化含量(mC/C)是将对应甲基化类型的绝对含量除以目标区域内该类型的胞嘧啶位点总数。

3　BS-Seq 数据后续分析

综合已发表的文献，通常 BS-Seq 数据经上述预处理后，可进行以下后续分析，进而探索其中蕴含的生物学意义。

3.1　全基因组中不同序列环境下的胞嘧啶甲基化

生物基因组内存在 3 种不同甲基化序列环境(context)：CG、CHG 和 CHH(这里 H 表示 A、C 或者 T 中的任何一个碱基)，表 1 列出了 8 种真核生物的胞嘧啶甲基化水平[16]，通常 CG 甲基化平均水平(39.8%)远高于 CHG(6.51%)和 CHH(1.37%)(表 1)[2, 17]。不同物种的甲基化水平差异明显，统计其 CG 占全基因组的比率可见，如蜜蜂(*Apis mellifera*)和衣藻(*Chlamydomonas reinhardtii*)甲基化水平很低分别为 0.93%和 5.38%，而斑马鱼(*Danio rerio*)和小鼠(*Mus musculus*)甲基化水平却分别高达 80.3%和 74.2%(表 1)，拟南芥(*Arabidopsis thaliana*)、海鞘(*Ciona intestinalis*)、毛果杨(*Populus trichocarpa*)、水稻(*Oryza sativa*)等物种呈中等甲基化水平，它们的 CG 比率变动在 22.3%~59.4%之间。上述这些差异强烈暗示，物种的甲基化水平是受遗传控制的[18]。保持 CG 甲基化是通过 DNA 甲基转移酶 DNMT1 实现；而 CHH 甲基化和部分的 CHG 甲基化，则由 Dnmt3 实现。在模式植物拟南芥中较高水平的 CHG，由植物特异性的甲基转移酶 CMT3 保持[19, 20]。

3.2　全基因组胞嘧啶甲基化的分布情况

通过甲基化敏感的限制性内切酶作图和亚硫酸氢盐测序方法，发现胞嘧啶甲基化在全基因组内会呈现全局性(Globally)和镶嵌性两种分布模式。脊椎动物和基因组较大的植物，如玉米(*Zea mays*)中，除了活性基因启动子周围的 CpG 岛外，大部分区域 CG 都被高度甲基化，使其甲基化区域连续发生，因而其 DNA 甲基化呈现全局性分布[21~23]。然而，镶嵌性分布模式是由于其部分基因主体(Gene body)上的胞嘧啶被甲基化，但转座子和重复序列上的胞嘧啶甲基化则受到 RNA 介导机制的制约，如基因组较小的植物(如拟南芥)，其转座子等元件上不存在特异性地甲基化，而在无脊椎动物中(如后口动物和昆虫)[24]和真菌基因组中，只有重复的 DNA 序列被甲基化[25]。在所有真核生物中，植物的 DNA 甲基化水平最高，胞嘧啶被甲基化高达 50%[26]。这可能是由于许多转座子区域已被甲基化[27,28]。

在哺乳动物胎儿肺成纤维细胞(IMR90)细胞中，DNA 甲基化几乎是在完全对称的 CG 环境下发生(占 99.98%)，CHG 和 CHH 的甲基化几乎缺失，不过，在胚胎干细胞(H1)阶段具有非 CG 甲基化的特性，能检测到少量的 CHG 和 CHH 甲基化，可以看出，这两种类型细胞的遗传差异[8]。从全基因组 DNA 甲基化分析看，每条染色体内 DNA 甲基化密度显示很大变化。通常在着丝粒、端粒等异染色质区域 DNA 甲基

表 1　8 个真核生物的胞嘧啶甲基化水平[16]

物种	拉丁学名	CG(%)	CHG(%)	CHH(%)	平均值(%)
蜜蜂	*Apis mellifera*	0.93	0.26	0.17	0.45
衣藻	*Chlamydomonas reinhardtii*	5.38	2.59	2.49	3.49
拟南芥	*Arabidopsis thaliana*	22.3	5.92	1.51	9.91
海鞘	*Ciona intestinalis*	31.1	0.17	0.12	10.46
毛果杨	*Populus trichocarpa*	41.9	20.9	3.25	22.02
水稻	*Oryza sativa*	59.4	20.7	2.18	27.43
小鼠	*Mus musculus*	74.2	0.3	0.29	24.93
斑马鱼	*Danio rerio*	80.3	1.22	0.91	27.48
平均值(%)		39.4	6.51	1.37	

化的密度较高[29, 30]。致使这些区域中的转座子不易移动，从而对基因组的完整性起保护作用[5]。在邻近基因的启动子内还发现有 CG 二核苷酸甲基化群，通常被称为 CpG 岛[31, 32]。

在植物细胞中，DNA 甲基化可以在胞嘧啶碱基内所有序列环境发生，包括：对称 CG 和 CHG 环境和不对称 CHH 环境[18]。全基因组 CG、CHG 和 CHH 背景的 DNA 甲基化水平，分别约为 24%、6.7%和 1.7%[29]。在拟南芥中，位于染色体近着丝粒区域，重复序列富集，CG、CHG 和 CHH 甲基化程度也较高，显示高度的相关性[7]。此外，所有类型的甲基化水平都和其序列的长度呈强烈的正相关。但是，反向重复序列中两侧翼 DNA 甲基化水平呈现逐渐减少的趋势。

3.3 DNA 甲基化环境和核苷酸的偏好

3 种 DNA 甲基化环境下其胞嘧啶甲基化水平差异十分悬殊，据报道，在拟南芥基因组中，CG 环境下的胞嘧啶甲基化水平的高低差距可以达到 13 倍，CHG 甲基化为 11 倍，CHH 则超过 900 倍。在研究不同 DNA 甲基化环境下核苷酸的偏好时，通常会观察到 CG、CHG 和 CHH 及其紧邻的 7-mer(单元)或 9-mer 核苷酸的变化，如图 2 中显示了 7-mer 碱基含量(从-2 到+4)。3 种核苷酸的偏好也有差异，如 CG 甲基化环境中其上下游通常检测到碱基 A 或 T，而 CHG 和 CHH 环境中其 H 位置的碱基通常为 C[8]。在小麦(*Triticum aestivum*)胚芽中也曾发现类似情况，即 CAG 和 CTG 位点甲基化水平比 CCG 位点高[33]。

拟南芥非 CG 甲基化环境(CHG 和 CHH)上游偏爱 TA 二核苷酸，但在 mCG 位点没有出现局部序列富集，推测非 CG 甲基化位点并不影响其局部序列富集。和拟南芥不同，人类胚胎干细胞(H1)未观察到 CHG 和 CHH 的上游对 TA 二核苷酸的偏爱(图 2)。此外，CHG 和 CHH 后面的碱基是 A 和 T 相对较多，在哺乳动物 DNMT3 甲基转移酶体外研究中也观察到此序列偏好。

为识别基因组 DNA 甲基化相邻位点之间的距离偏好，通常会重点分析基因组内含子区域的非

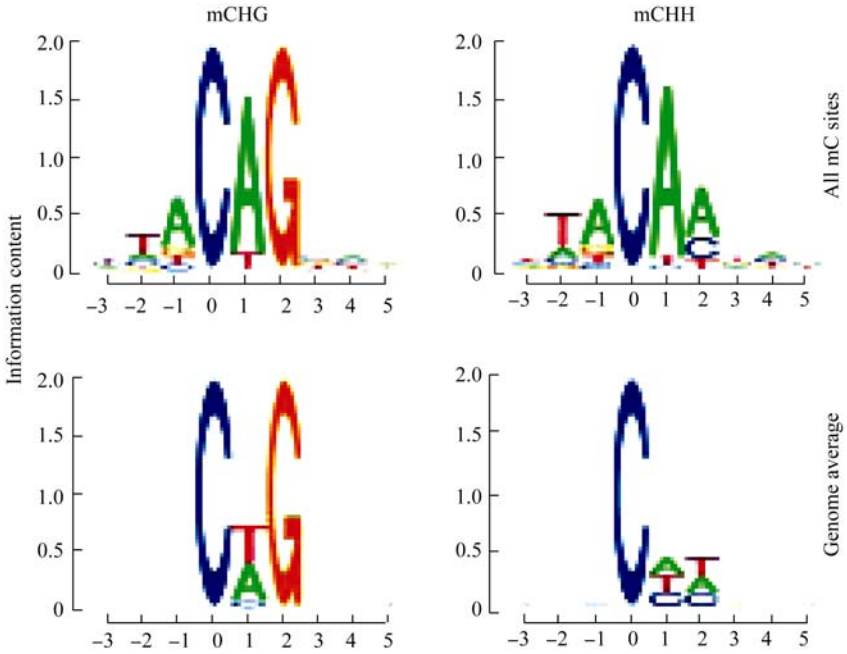

图 2 人类胚胎干细胞非 CG DNA 的甲基化位点近端序列的标志图(Logo)

横坐标为 mCHG 和 mCHH 上下游紧邻的 7 个 mer，纵坐标为碱基含量。上方为所有胞嘧啶甲基化所处的序列环境。H 指的是腺嘌呤(A)、胸腺嘧啶(T)和胞嘧啶(C)中的任何一个碱基。mCHG 中的 H 偏好于腺嘌呤(A)，mCHG 上游偏好 TA(胸腺嘧啶和腺嘌呤)，但下游无碱基偏好。mCHH 中第一个 H 偏好于腺嘌呤(A)，第二个 H 首先偏好于腺嘌呤(A)，其次为胞嘧啶(C)，mCHH 上游序列偏好 TA，下游也无碱基偏好。下方为基因组平均的序列环境。mCHG 位点中的 H 常为胸腺嘧啶(T)和腺嘌呤(A)，上下游均无碱基偏好。mCHH 位点中的 H 以及上下游碱基均无明显偏好。

CG 甲基化，因为内含子不会经受编码蛋白质的选择性胁迫。通过 BS-Seq 发现人类基因组 mCHH 环境中具有明显的 8 碱基周期性[7, 8]，而在 mCHG 位点中其周期性的 8 个碱基不连续，通常会因其他碱基插入而被分开[8]。拟南芥基因组在 mCHH 环境中，8 碱基周期性也很明显，并和单圈 DNA 螺旋对应，推测在植物和动物之间可能存在共同的从头甲基化的分子机制。人类基因组中沉积不同的 mCHG 和 mCHH 相对间距模式，暗示 DNMT3A 可能通过不同途径对 mCHG 和 mCHH 中的胞嘧啶进行甲基化[8]，因此，将这几种非 CG 甲基化类型进行子分类是十分必要的。

3.4　DNA-蛋白质互作位点上的 DNA 甲基化

近年来，应用 ChIP-seq 可以鉴定细胞中蛋白质-DNA 相互作用的位点，如 Nanog、SOX2、KLF4、Oct4 等蛋白质和增强子蛋白(TAF1, P300)对 DNA 甲基化的影响[8]。从图 3a 可见，通常在非 CG 的环境下，观察到邻近的 TSS 位点时其甲基化相对密度谱会明显减少。基于 H3K4mel 和 H3K27ac 区域 ChIP-seq 平均富集度的检测可以把增强子位点分为 3 个类型(即 IMR90 特异性型、H1 特异性型和 H1 和 IMR90 细胞共有型，图 3b)。它们的 CG 和非 CG 甲基化密度有差异，尤其是非 CG 甲基化密度。IMR90 特异性增强子型在非 CG 甲基化密度上位点和侧翼 5 kb 间无明显差异，而 H1 特异性增强子型和 H1 和 IMR90 细胞共有型中其位点和侧翼 5 kb 间出现明显 DNA 甲基化密度低谷(图 3b) [34]。真核生物中的多功能转录因子 CTCF 是一种高度保守的多锌指、DNA 结合核蛋白。最新发现 41%的 CTCF 结合位点具有不同的甲基化状态，并且甲基化变化主要发生在 2 个识别序列关键的核苷酸位置。在正常细胞与肿瘤细胞之间 CTCF 结合模式明显不同，研究表明，在肿瘤细胞中 CTCF 结合程度减弱与 DNA 甲基化程度的增强往往关联[35]。

在拟南芥中蛋白编码基因体内，CG 甲基化水平远高于 CHG 和 CHH，并呈现一个高峰，与此相反，短干扰 RNA(siRNAs)在基因体内明显减弱，形成一个低谷，这种反相关与 RNA 指导的 DNA 甲基化的分子特性[18]，表明增强子和基因体具有不同的 CG 与非 CG 基因甲基化模式。

3.5　不同基因及其结构元件的胞嘧啶甲基化程度

基于 BS-Seq 提供的单碱基水平的甲基化图谱，还可以通过 R 绘图程序画出不同基因功能元件，包括启动子、外显子、内含子、UTR 区中相对甲基化密度分布图。由图 4 可以观察到在启动子(启动子涵盖转录起始位点上游 2 kb)和转录起始位点(TSS)甲基化相对密度较低，接近 TSS 处呈明显下降，但没有完全枯竭。而到 5′UTR 明显回升；外显子、内含子和 3′UTR 域中的非 CG 甲基化(mCHG 和 mCHH)密度比启动子、转录起始位点和 5′UTR 区域高出约两倍。有趣的是，外显子区 CG 和 CHH 甲基化密度十分接近，进入内含子区和 3′UTR 区后，CHH 甲基化密度又明显下降[8]。通过对更多地物种进行 BS-Seq 测序，Feng 等[16]发现基因主体上的甲基化在动植物之间是很保守的，而且大部分甲基化还偏好于外显子区域。

4　结　语

在基因组上检测 DNA 甲基化位点通常采用 3 种方法。基于酶切方法仅能识别一部分位点的甲基化，且不能识别甲基化在染色体上的具体位置。免疫沉淀的方法虽不受序列的限制，但未达到单碱基水平的高分辨率，因而不能识别甲基化位点所处的序列环境，难以检测低水平的甲基化位点，还存在偏好序列的富集[35]；随着新一代测序技术(大规模平行，末端配对，长读序列测序)的发展，BS-Seq 方法已在很多动植物物种中应用于制作全基因组甲基化图谱。这里，我们首先介绍了全基因组规模 DNA 甲基化测序数据处理的基本方法，然后，进一步对表观遗传后续分析加以综述。真核生物基因组胞嘧啶甲基化分析表明，不同生物甲基化水平有明显差异，是受其不同的遗传背景控制的。脊椎动物和无脊椎动物全基因组的胞嘧啶甲基化分布呈现了两种不同模式，前者为全局性，后者为镶嵌性。和哺乳动物不同，植物 DNA 甲基化主要发生在转座子和其他重复的 DNA 元件中[2]。植物基因组与动物以及人类基因组在 DNA 甲基化环境和核苷酸的偏好上有相似之处，如在 mCHH 环境中位点具有明显的周期性，它们可能存在共同的从头甲基化的分子机制，但其甲基化途径有所差异。此外，基因组内不同基因功

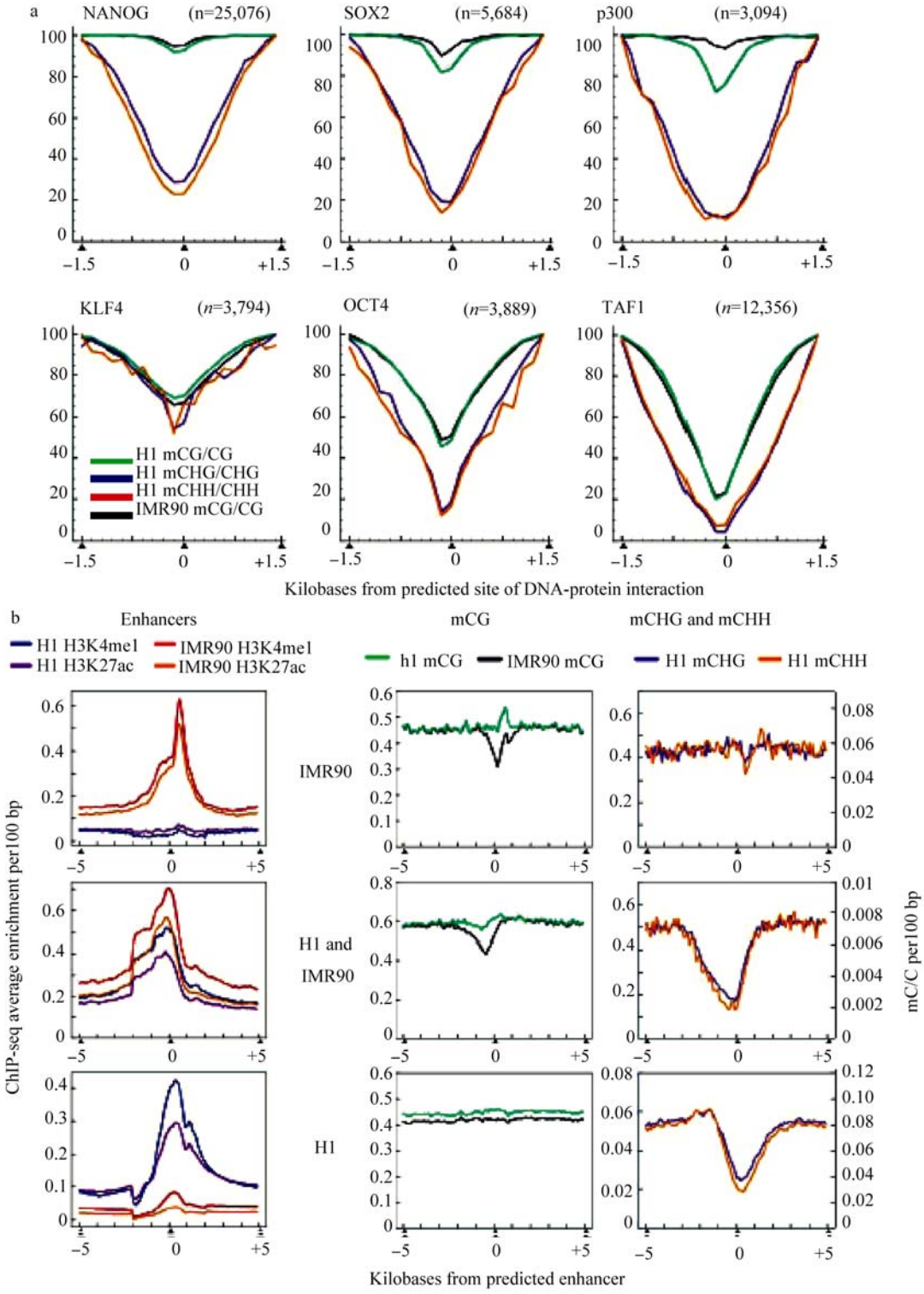

图 3 蛋白质 DNA 互作位点 DNA 甲基化密度[81]

a：蛋白质 DNA 互作位点上游 1.5 kb 和下游 1.5 kb 间的 DNA 甲基化密度。X 轴：与互作位点的距离; Y 轴：mC/C 归一化值。b：3 类增强子位置(即 IMR90 特异性的(上图)、H1 和 IMR90 细胞共有的(中图)和 H1 特异性的(下图)。X 轴：与增强子位点的距离; Y 轴：每 100 kb Chip-seq 平均富集(左), 每 100 kb mC/C(右)。

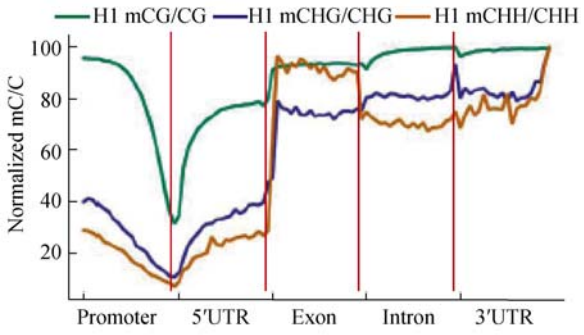

图 4　人类胚胎干细胞(H1)细胞非 **CG** 甲基化密度[8]
X 轴自左向右：启动子、外显子、内含子、UTR; Y 轴归一化 MC/C。

能元件区间的相对甲基化密度也具差异。最新发现肿瘤细胞中重要转录因子 CTCF 与染色质 DNA 之间的相互作用和染色质 DNA 甲基化关联[35]，这两个不同层次的表观基因组学数据的有机地整合，将为筛选诱导肿瘤发生的表观遗传靶点提供可能。DNA 甲基化是一个十分复杂而富有挑战性的课题，进一步挖掘其蕴藏的生物学意义，不仅能丰富真核基因表达调控的科学理论体系，并具有潜在的临床转化应用价值。

参考文献(References):

[1] Rothbart SB, Krajewski K, Nady N, Tempel W, Xue S, Badeaux AI, Barsyte-Lovejoy D, Martinez JY, Bedford MT, Fuchs SM, Arrowsmith CH, Strahl BD. Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat Struct Mol Biol*, 2012, 19(11): 1155–1160. DOI

[2] Zhang XY, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell*, 2006, 126(6): 1189–1201. DOI

[3] Meissner A, Mikkelsen TS, Gu HC, Wernig M, Hanna J, Sivachenko A, Zhang XL, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 2008, 454(7205): 766–770. DOI

[4] Vaughn MW, Tanurdžić M, Lippman Z, Jiang HM, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, Colot V, Doerge RW, Martienssen RA. Epigenetic natural variation in *Arabidopsis thaliana*. *PLoS Biol*, 2007, 5(7): e174. DOI

[5] He GM, Elling AA, Deng XW. The epigenome and plant development. *Annu Rev Plant Biol*, 2011, 62(1): 411–435. DOI

[6] Shen HS, He H, Li JG, Chen W, Wang XC, Guo L, Peng ZY, He GM, Zhong SW, Qi YJ, Terzaghi W, Deng XW. Genome-wide analysis of DNA methylation and gene expression changes in two *Arabidopsis* ecotypes and their reciprocal hybrids. *Plant Cell*, 2012, 24(3): 875–892. DOI

[7] Cokus SJ, Feng SH, Zhang XY, Chen ZG, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature*, 2008, 452(7184): 215–219. DOI

[8] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 2009, 462(7271): 315–322. DOI

[9] Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet*, 2005, 37(8): 853–862. DOI

[10] Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res*, 2009, 19(6): 959–966. DOI

[11] Clark SJ, Statham A, Stirzaker C, Molloy PL, Frommer M. DNA methylation: bisulphite modification and analysis. *Nat Protoc*, 2006, 1(5): 2353–2364. DOI

[12] Zilberman D, Henikoff S. Genome-wide analysis of DNA methylation patterns. *Development*, 2007, 134(22): 3959–3965. DOI

[13] Zemach A, Kim MY, Silva P, Rodrigues JA, Dotson B, Brooks MD, Zilberman D. Local DNA hypomethylation activates genes in rice endosperm. *Proc Natl Acad Sci USA*, 2010, 107(43): 18729–18734. DOI

[14] Ouyang S, Zhu W, Hamilton J, Lin HN, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res*, 2007, 35(Database issue): D883–D887. DOI

[15] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2: Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 2012, 9(4): 357–359. DOI

[16] Feng SH, Cokus SJ, Zhang XY, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, Ukomadu C, Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE.

Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci USA*, 2010, 107(19): 8689–8694. DOI

[17] Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet*, 2007, 39(1): 61–69. DOI

[18] Henderson IR, Jacobsen SE. Epigenetic inheritance in plants. *Nature*, 2007, 447(7143): 418–424. DOI

[19] Chan SW, Henderson IR, Jacobsen SE. Gardening the genome: DNA methylation in *Arabidopsis thaliana*. *Nat Rev Genet*, 2005, 6(5): 351–360. DOI

[20] Goll MG, Bestor TH. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem*, 2005, 74(1): 481–514. DOI

[21] Macleod D, Clark VH, Bird A. Absence of genome-wide changes in DNA methylation during development of the zebrafish. *Nat Genet*, 1999, 23(2): 139–140. DOI

[22] Stancheva I, El-Maarri O, Walter J, Niveleau A, Meehan RR. DNA methylation at promoter regions regulates the timing of gene activation in *Xenopus laevis* embryos. *Dev Biol*, 2002, 243(1): 155–165. DOI

[23] Estécio MR, Gharibyan V, Shen LL, Ibrahim AE, Doshi K, He R, Jelinek J, Yang AS, Yan PS, Huang TH, Tajara EH, Issa JP. LINE-1 hypomethylation in cancer is highly variable and inversely correlated with microsatellite instability. *PLoS One*, 2007, 2(5): e399. DOI

[24] Tweedie S, Charlton J, Clark V, Bird A. Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol Cell Biol*, 1997, 17(3): 1469–1475. DOI

[25] Selker EU, Tountas NA, Cross SH, Margolin BS, Murphy JG, Bird AP, Freitag M. The methylated component of the Neurospora crassa genome. *Nature*, 2003, 422(6934): 893–897. DOI

[26] Montero LM, Filipski J, Gil P, Capel J, Martínez-Zapater JM, Salinas J. The distribution of 5'methylcytosine in the nuclear genome of plants. *Nucleic Acids Res*, 1992, 20(12): 3207–3210. DOI

[27] Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR. Maize genome sequencing by methylation filtration. *Science*, 2003, 302(5653): 2115–2117. DOI

[28] SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL. Nested retrotransposons in the intergenic regions of the maize genome. *Science*, 1996, 274(5288): 765–768. DOI

[29] Gonzalo S, Jaco I, Fraga MF, Chen TP, Li E, Esteller M, Blasco MA. DNA methyltransferases control telomere length and telomere recombination in mammalian cells. *Nat Cell Biol*, 2006, 8(4): 416–424. DOI

[30] Steinert S, Shay JW, Wright WE. Modification of subtelomeric DNA. *Mol Cell Biol*, 2004, 24(10): 4571–4580. DOI

[31] Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet*, 2009, 10(5): 295–304. DOI

[32] Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nature Rev Genet*, 2008, 9(6): 465–476. DOI

[33] Gruenbaum Y, Naveh-Many T, Cedar H, Razin A. Sequence specificity of methylation in higher plant DNA. *Nature*, 1981, 292(5826): 860–862. DOI

[34] Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. CTCF mediates methylation-sensitive enhancer-blocking activity at the *H19*/*Igf2* locus. *Nature*, 2000, 405(6785): 486–489. DOI

[35] Wang H, Maurano MT, Qu HZ, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, Thurman RE, Kaul R, Myers RM, Stamatoyannopoulos JA. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res*, 2012, 22(9): 1680–1688. DOI