doi:10.3969/j.issn.1672-5565.2014.01.12

高通量 DNA 甲基化数据的处理和分析方法

王心宇,许颖出,刘洪波,王　芳,张　岩*,苏建忠*
(哈尔滨医科大学生物信息科学与技术学院,黑龙江 哈尔滨 150080)

摘　要:DNA 甲基化作为一种表观遗传学修饰,在调控基因表达、X 染色体失活、印记基因等方面都发挥着重要的作用。不同的 DNA 甲基化的预处理方法结合二代测序产生了大量的高通量甲基化数据,这些数据的存储、处理和分析是当前亟需解决的问题。在本文中,总结了目前存在的三种高通量 DNA 甲基化检测技术(限制性内切酶法,亲和纯化法,重亚硫酸盐转换法),以及针对这些技术产生的高通量数据开发的存储、处理和分析工具。另外,还注重介绍了单碱基水平的 DNA 甲基化检测技术,BS-Seq 的测序原理、数据处理流程以及后续的分析工具。

关键词:DNA 甲基化;高通量;二代测序;BS-Seq
中图分类号:N37　　**文献标志码**:A　　**文章编号**:1672-5565(2014)-01-072-05

Methods of bioinformatics analysis for high-throughput sequencing of DNA methylation

WANG Xinyu, XU Yingchu, LIU Hongbo, WANG Fang, ZHANG Yan*, SU Jianzhong*
(*College of Bioinformatics Science and Technology*, *Harbin Medical University*, *Harbin* 150086, *China*)

Abstract: DNA methylation is an important epigenetic modification and plays crucial roles in regulating gene expression, X chromosome activation and imprinting genes. Several pretreatment approaches of DNA methylation combined with next-generation sequencing have generated enormous high-throughput data. How to store, process and analyze large volume of raw data produced are in an urgent requirement. Here, we summarized three high-throughput sequencing technologies of DNA methylation and the relative bioinformatics tools. Furthermore, we highlight the principles and the methods of data processing of the combination of bisulfite treatment of DNA and high-throughput sequencing data (BS-seq).

Keywords: DNA methylation; High-throughput; NGS; BS-Seq

DNA 甲基化是重要的表观遗传学修饰之一,以往的研究表明,DNA 甲基化在细胞发育和分化、调控基因表达、X 染色体失活、基因沉默、疾病的发生等方面扮演着重要的角色[1-3]。在真核生物中,通常是 CpG 二核苷酸中胞嘧啶的第五个碳原子上发生了甲基化(5mC),胞嘧啶甲基化也可能会发生在 CHG 和 CHH(H 是除 G 外的任意一种核苷酸)上。全基因组的甲基化水平呈现双峰分布,而且低甲基化的区域多数是在 CpG 二核苷酸聚集区域(CpG 岛)[4]。以往的研究发现,位于启动子区域的高甲基化的 CpG 岛与基因的沉默有关,可能是因为 DNA 甲基化阻碍转录因子结合而直接抑制了基因转录。在过去的几十年里,由于实验技术和费用的限制,DNA 甲基化的数据往往只检测了基因组的局部区域,而且是低通量的数据。

二代测序技术的发展极大地推动了表观遗传调控机制的研究,基于二代测序技术发展起来的 DNA 甲基化的检测技术为 DNA 甲基化的研究提供了大量的高通量、全基因组的 DNA 甲基化数据。这些高通量数据的产生使得 DNA 甲基化研究的重点

收稿日期:2013-10-10;**修回日期**:2013-11-20.
基金项目:黑龙江省大学生创新创业训练计划(201210226018),国家自然基金青年项目(61203262)。
作者简介:王心宇,男,本科生,研究方向:计算表观遗传学;E-mail:wangxinyuhs@126.com.
*通信作者:张岩,女,教授,研究方向:计算表观遗传学、生物信息学; E-mail: yanyou1225@163.com;
苏建忠,男,讲师,研究方向:计算表观遗传学、生物信息学; E-mail: jianzhongsu82@gmail.com.

由目标基因 DNA 甲基化的检测转移到了全基因组 DNA 甲基化高通量数据的检测、存储、处理和分析上。近几年,研究者构建了多个 DNA 甲基化数据库,开发了大量的 DNA 甲基化高通量数据的处理和分析工具,使得深入的表观遗传调控机制的研究成为可能。

1　基于二代测序技术的 DNA 甲基化检测技术

1.1　DNA 甲基化预处理方法

甲基化后的胞嘧啶(5 mC)与普通的胞嘧啶(C)在 DNA 序列上并无差异,如果直接使用 DNA 测序,将无法区分测得的胞嘧啶 C 是 C 还是 5 mC。所以检测 DNA 甲基化需要首先对待检测的 DNA 序列中胞嘧啶进行预处理,将非甲基化的胞嘧啶 C 与甲基化的胞嘧啶 5 mC 区分开来,目前的 DNA 甲基化预处理方式主要分为三种:

(1) 限制性内切酶法(Endonuclease digestion)

限制性内切酶法是指利用甲基化限制性内切酶(HpaII, MspI 和 HhaI 等)在各自的识别位点对甲基化的胞嘧啶有不同的敏感性来检测 CpG 的甲基化[5]。限制性内切酶法结合二代测序的技术有 MRE-seq,MCA-seq,MSCC 和 HELP-seq。尽管限制性内切酶测序法成本低、高效,然而由于检测的 CpG 位点局限于酶切位点附近,基因组覆盖率低,另外还存在 CpG 偏好性、酶切不完全导致的假阳性等问题,使用这种方法检测 DNA 甲基化的研究越来越少。

(2) 亲和纯化法(Affinity enrichment)

亲和纯化是利用甲基化 CpG 结合蛋白(MBD)或者对 5mC 特异的抗体来亲和提纯甲基化区域。MeDIP-seq 和 MBD-seq 是最常用的两种结合亲和纯化和二代测序技术的 DNA 甲基化检测方法。基于测序的亲和纯化法能够快速、低成本地检测全基因组范围内的甲基化水平,然而它只能获得区域的甲基化水平,特别是 MeDIP-seq 偏向于 CpG 富集的区域,分散的低密度的甲基化位点可能被识别成非甲基化区域,目前还没有能够去除掉这种偏性的生物信息学方法。

(3) 重亚硫酸盐转换法(Bisulphite conversion)

重亚硫酸盐转换结合二代测序技术是目前最精准的 DNA 甲基化检测方法,能够检测单碱基水平的甲基化状态,被称为 DNA 甲基化检测的“金标准”。对基因组中未发生甲基化的胞嘧啶进行重亚硫酸盐处理,将其转换成 U,经 PCR 扩增后变成 T,重亚硫酸盐转换对甲基化的胞嘧啶不起作用。通过结合二代测序,即可绘制出单碱基分辨率的全基因组 DNA 甲基化图谱。目前常用的重亚硫酸盐转换结合二代测序技术的 DNA 甲基化检测技术有 BS-seq 和 RRBS 等。

1.2　二代测序技术

在使用 DNA 甲基化预处理区分出未甲基化的胞嘧啶和甲基化的胞嘧啶后,再使用二代测序技术检测 DNA 序列,来获取胞嘧啶上的甲基化状态。

目前二代测序技术主要分为三个平台:Roche、Illumina、SOLiD。其中每种测序平台又拥有多种系统,比如 Illumina 就有 HiSeq、GAIIx 等系统。不同的测序技术在测得的 read 长度、精确性、通量都有差异,适用于不同的研究目的需要。

1.3　高通量 DNA 甲基化检测技术进展

结合二代测序技术和 DNA 甲基化预处理的 DNA 甲基化检测方法,在近几年获得了大量的全基因组的 DNA 甲基化测序数据。

国外很多实验室产生了大量、精准的高通量 DNA 甲基化数据,例如,Lister 等人于 2008 年检测的拟南芥全基因组甲基化谱和 2009 年测得的人类全基因组甲基化谱[6-7],Stadler 等人于 2011 年测定了小鼠胚胎干细胞和神经前体细胞的全基因组甲基化谱等[8]。国内近年来也产生了大量的高通量 DNA 甲基化数据,例如,2010 年,中科院昆明研究所,华大基因和上海交通大学癌症表观遗传中心等九家科研机构联合测定了桑蚕的单碱基水平的 DNA 甲基化谱,王俊教授课题组测定的人类完全分化的血细胞的全基因组 DNA 甲基化谱等。这些全基因组水平的 DNA 甲基化数据为表观遗传调控机制的研究提供了数据资源。

2　DNA 甲基化数据储存和可视化

目前研究者构建了各种各样的数据库来存储世界范围的各大实验室和科研机构产生的高通量 DNA 甲基化数据,便于数据的查询、下载、可视化分析及全球化的资源共享。从第一个 DNA 甲基化的公共数据库 MethDB 由 Grunau 等人于 2001 年构建以来,已有多个和 DNA 甲基化相关的数据库被开发,例如,NCBI 的存储表观遗传修饰数据的 Epigenomics,主要包括 DNA 甲基化、组蛋白修饰和非编码 RNA 等数据。PubMeth 是结合文本的基因注释信息的 DNA 甲基化数据库。DiseaseMeth 储存 72 种人类疾病相关的 DNA 甲基化的数据库,并实现了统计学分析及可视化[9]。

3 高通量 DNA 甲基化数据的处理和分析

结合二代测序技术和 DNA 甲基化预处理的方法,在近几年产生了大量的全基因组的 DNA 甲基化测序数据。然而,因为存在多种测序技术以及多种 DNA 甲基化预处理的技术,这些高通量的数据的存储、处理和分析是目前 DNA 甲基化研究的一个难点和热点。目前常见的高通量 DNA 甲基化数据检测,处理和分析的流程如图 1 所示。

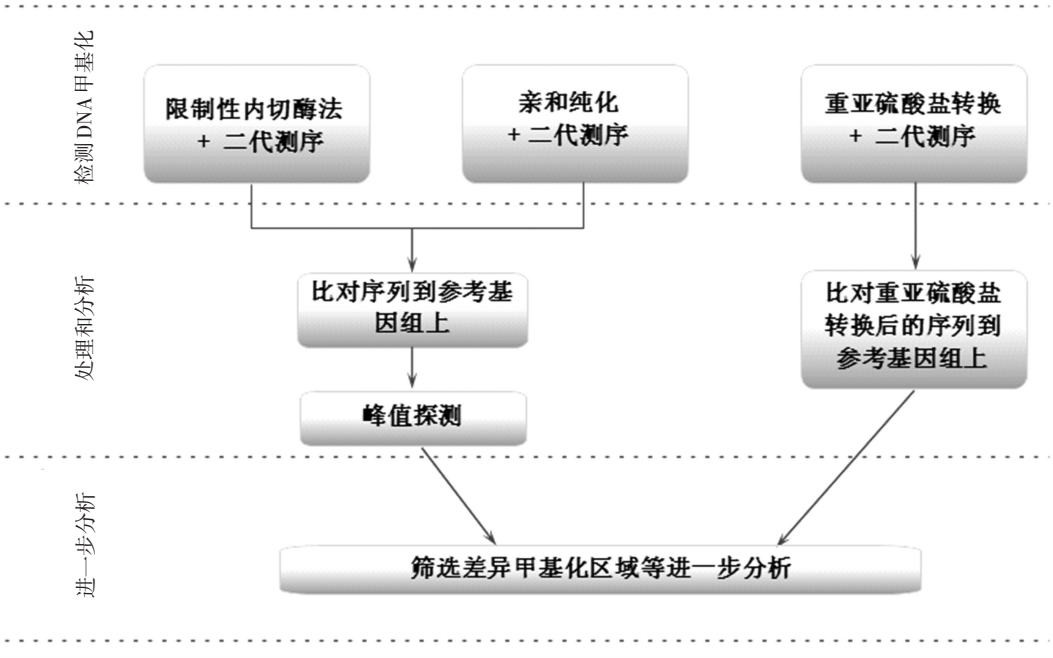

图 1 高通量 DNA 甲基化测序数据的检测,处理和分析的方法及软件

Fig.1 Methods of detection and software packages of analysis for high-throughput sequencing of DNA methylation

3.1 DNA 甲基化序列数据处理的挑战

3.1.1 甲基化预处理方法的差异和测序技术的差异

MeDIP-seq 和 MBD-seq 只能检测某个区域的甲基化状态,而 BS-Seq、RRBS 方法能够测得单碱基水平的甲基化状态。不同的 DNA 甲基化检测方法测得的数据也存在差异,需要不同的处理和分析方法。

3.1.2 MBD-Seq、MeDIP-Seq 数据处理的挑战

MBD-Seq 和 MeDIP-Seq 测得的序列数据可以使用 Bowtie、SOAP 等短序列比对软件直接比对到参考基因组上,用映射到某个区域的 reads 数目来反应这个区域的甲基化程度[10-11]。然而,这两种测序方法检测的区域偏向 CpG 密集的甲基化区域。当某个甲基化区域的 CpG 分散时,很有可能被视为非甲基化区域。基因组的不同区域上 CpG 密度分布是不均匀的,因而需要开发新的生物信息学方法来校正,以获取基因组范围内准确的甲基化水平。

3.1.3 BS-Seq、RRBS 数据处理的挑战

BS-Seq 和 RRBS 可以直接测得单个胞嘧啶的甲基化状态,准确性很高。然而,因为经过重亚硫酸盐转换之后,DNA 的序列发生了改变(C 变成了 T,mC 和其他碱基保持不变),不能够直接比对到参考基因组上。另外,与 Illumina 直观的碱基序列不同,SOLiD 测序将 reads 利用颜色空间进行编码,将每一个碱基与它邻近的碱基用一种颜色表示。碱基序列比对的工具不适用于 SOLiD 测序产生的序列。

3.2 DNA 甲基化序列数据处理分析的研究现状

研究者已经开发的峰度探测软件包括 MACS, USeq, PeakSeq, FindPeaks, BayesPeak 等,其中 MACS 是目前最常用的峰值探测工具。然而,目前仍没有专门处理 MBD-seq 数据的工具或软件来降低或去除 CpG 密度对 MBD-seq 产生数据的影响。

研究者基于短序列匹配算法(Bowtie,SOAP 等)开发了 10 多种专门处理重亚硫酸盐转换后的 reads 的比对工具和算法,比如 Bismark, MethylCoder, BRAT, BSMAP, BS Seeker, B-SOLADA, SOCS-B, BatMeth, RMAP-BS, FadE 等[12-14]。其中,Bismark 是最常用的碱基序列比对工具,FadE, B-SOLADA, SOCS-B, BatMeth 是可以处理颜色空间编码的 reads。如表 1 所示。

表 1　2011~2012 年 BS-Seq 分析软件包比较

Table1　Comparison of software packages for BS-Seq analysis from 2011 to 2012

	MethylCoder	B-SOLADA	Bismark	FadE	BatMeth
比对工具	Bowtie/GSNAP	Bowtie	Bowtie	-	-
语　言	Python	Python	Perl	-	-
支持序列类型	单/双末端	单末端	单/双末端	-	-
输　出	比对结果,甲基化水平	比对结果,甲基化水平	比对结果,甲基化水平	-	-
特　点	无偏比对,高效率	高效率	无偏比对,高效率	信息熵	适用于 SOLiD 和 Illumina
时　间	2011	2011	2011	2012	2012
形成软件	是	是	是	否	否
被引用次数	15	8	86	-	2

4　BS-Seq 的数据处理及分析

4.1　BS-Seq 的原理

BS-Seq 先利用重亚硫酸盐转换将普通的胞嘧啶变为 U,而甲基化的胞嘧啶保持不变,然后使用 PCR 扩增使得 U 变成 T。对转换和扩增后的 DNA 序列进行测序,将得到的 DNA 序列与参考基因组进行比较。认为 C-C 配对(参考基因组上在某个位置上是 C,测得的 reads 在该位置上也是 C)的就是甲基化的胞嘧啶,C-T 配对的是非甲基化的胞嘧啶。如图 2 所示。

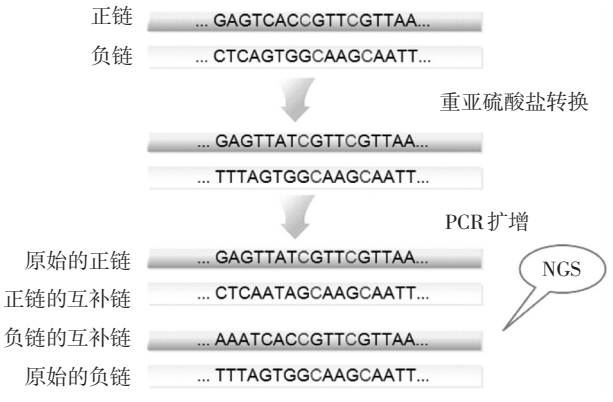

图 2　BS-Seq 原理

Fig.2　BS-Seq protocol

4.2　BS-Seq 数据处理流程

使用 BS-Seq 测得的序列数据通常为 fastq 或 fasta 格式。从序列数据中获得单个胞嘧啶的甲基化水平一般包括以下几个步骤,如图 3 所示:

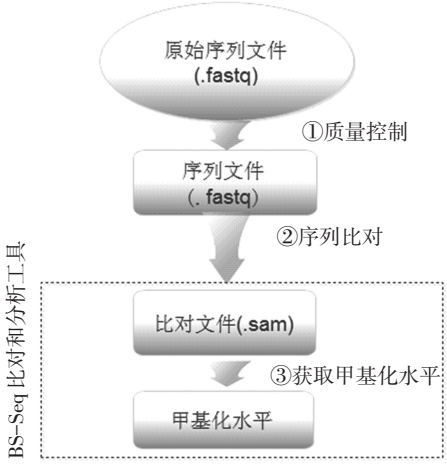

图 3　BS-Seq 数据处理流程

Fig.3　Recommended workflow for the analysis of BS-Seq data

(1)序列的质量控制。对于真实的数据,当 reads 的长度增加时,测序的错误率倾向于升高。另外,reads 上包含的引物会降低匹配到基因组上的准确率。因此,有时候会对序列数据进行碱基质量分数控制、修剪引物等处理。

(2)序列比对。BS-Seq 产生的序列与基因组上的原始序列存在差异(普通 C 变为 T,互补链上的 G 变成了 A),需要使用 BS-Seq 特有的序列比对软件(Bismark 等),将 BS-Seq 产生的序列数据比对到参考基因组上。

(3)产生甲基化水平。从 reads 的基因组位置中获得每个胞嘧啶的甲基化 reads 数和非甲基化 reads 数。然后使用公式 M/(U+M)计算某个胞嘧啶的甲基化水平,U 和 M 分别是在这个胞嘧啶上的非甲基化 reads 数和甲基化 reads 数。

4.3 甲基化水平的后续分析

将单个胞嘧啶上的测序信息转换成了[0,1]的DNA 甲基化水平后,研究者开发了一系列的 DNA 甲基化数据分析工具,实现从 DNA 甲基化水平中寻找甲基化模式和统计学分析等功能,以方便实验生物学家进行进一步的 DNA 甲基化调控机制的研究。

张岩教授课题组于 2012 年开发了一个可视化工具 CpG_MPs,可以从标准化后的 DNA 甲基化水平中筛选甲基化区域和非甲基化区域[15]。Altuna 等人也于 2012 年开发了一个 R 包,实现了对 DNA 甲基化水平的样本质量可视化、差异甲基化分析、功能注释等功能[16]。

5 总 结

基于二代测序技术的 DNA 甲基化检测方法极大地推动了 DNA 甲基化的研究。研究者基于这些技术产生的高通量数据开发了一系列的生物信息学工具,然而,仍然有许多问题需要解决。目前已经开发了许多种工具可以处理和分析 BS-Seq 数据,然而对于 MBD-Seq 和 MeDIP-Seq,虽然也有一些工具,但却还无法解决 CpG 密度偏性的问题。对于 BS-Seq 的数据,颜色空间编码的碱基序列比对的精度和效率依然是一项挑战。

参考文献(References)

[1] LAIRD P W. Principles and challenges of genomewide DNA methylation analysis[J]. Nature reviews. Genetics, 2010 , 11, 191-203.

[2] BIRD A.DNA methylation patterns and epigenetic memory [J]. Genes & development, 2002 , 16, 6-21.

[3] GORE A, LI Z, FUNG H L, et al. Somatic coding mutations in human induced pluripotent stem cells[J]. Nature, 2011, 471, 63-67.

[4] SU Jianzhong, ZHANG Yan, LÜ Jie, et al.CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes[J]. Nucleic Acids Res,2009,38, e6.

[5] ZILBERMAN D, HENIKOFF S. Genome-wide analysis of DNA methylation patterns[J]. Development, 2007, 134, 3959-3965.

[6] LISTER R, PELIZZOLA M, DOWEN R H, et al. Human DNA methylomes at base resolution show widespread epigenomic differences[J]. Nature,2009,462,315-22.

[7] LISTER R, O'MALLEY R C, TONTI-FILIPPINI J, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis[J]. Cell,2008,133,523-36.

[8] STADLER M B, MURR R, BURGER L, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions[J].Nature,2011,484, 550.

[9] LÜ Jie, LIU Hongbo, SU Jianzhong, et al. DiseaseMeth: a human disease methylation database[J]. Nucleic Acids Res,2012,40,D1030-1035.

[10] LI Ruiqiang, YU Chang, LI Yingrui, et al. SOAP2: an improved ultrafast tool for short read alignment [J]. Bioinformatics, 2009, 25, 1966-1967.

[11] LANGMEAD B, TRAPNELL C, POP M,et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome[J]. Genome biology, 2009, 10, R25.

[12] KRUEGER F, KRECK B, FRANKE A, et al. DNA methylome analysis using short bisulfite sequencing data [J], Nat Methods ,2012,9, 145-151.

[13] LIM J Q, TENNAKOON C, LI G, et al. BatMeth: improved mapper for bisulfate sequencing reads on DNA methylation[J] , Genome Biology,2012, 13:R82.

[14] SOUAIAIA T, ZHANG Z, CHEN T.FadE: whole genome methylation analysis for multiple sequencing platforms [J]. Nucleic Acids Res,2012, 41,e14.

[15] SU Jianzhong, YAN Haidan, WEI Yanjun,et al. CpG_MPs: identification of CpG methylation patterns of genomic regions from high- throughput bisulfite sequencing data[J].Nucleic Acids Res,2012, 41,e4.

[16] AKALIN A, KORMAKSSON M, LI S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles [J], Genome Biology, 2012, 13:R87.