评述

人类基因组 DNA 甲基化数据分析的研究现状

凡时财^{①②}, 李承哲^①, 裴云飞^③

① 电子科技大学自动化工程学院, 成都 611731;

② 四川省大数据与智慧信息系统协同创新中心, 成都 611731;

③ 清华大学自动化系, 北京 100084

E-mail: shicaifan@uestc.edu.cn

收稿日期: 2015-01-12; 接受日期: 2015-03-09; 网络版发表日期: 2015-05-06

国家自然科学基金(批准号: 61105003)和中央高校基础研究基金(批准号: A03007023801210)资助项目

doi: 10.1360/N052015-00009

摘要 作为人类基因组最为典型的表观遗传现象, DNA 甲基化在多种关键生理活动中扮演重要角色. 系统分析基因组尺度的 DNA 甲基化概况意义重大. 从 CpG 岛等基本定义出发, 阐述了高通量 DNA 甲基化的检测技术以及针对芯片技术与下一代测序技术的低水平数据处理方法; 重点对比了基于机器学习理论对 CpG 位点及 CpG 岛甲基化水平的预测算法, 以及所利用的特征对预测效果的影响与发展趋势; 并对 DNA 差异甲基化在组织特异性、癌症等多种疾病中的计算分析进行了全面的综述.

关键词DNA 甲基化
检测技术
预测方法
差异甲基化
CpG 岛

DNA 甲基化是表观遗传学中最早被人们认识的现象之一, 曾被称为有别于 A, C, G, T 的“第 5 种碱基”^[1]. 在人类基因组中, 约 70%~80% 的 CpG 处于甲基化状态, 即呈现整体甲基化的特点, 同时这种甲基化模式存在时间和空间上的差异性. 时间上的差异性表现为, 精子和卵子结合成为受精卵以后, 合子将经历整体去甲基化的过程, 受精卵植入子宫壁以后又会经历基因组从头甲基化的过程^[2]. 空间上的差异性表现为非甲基化的 CpG 不是均匀分布在基因组上, 而是成簇出现形成 CpG 密度较高的非甲基化区域, 即 CpG 岛. DNA 甲基化与基因组印迹^[3]、干细胞分化^[4]、胚胎发育^[2]以及包含癌症在内的多种疾病的发生与发展密切相关^[5,6], 因此, 对于 DNA 甲基化的深入研究受到广泛关注, 在 2013 年汤森路透的生物科学领域重点研究方向排名中, DNA 甲基化被认为是最前沿的研究方向. 目前, 研究重点也从早期对于单

个基因或者位点甲基化水平检测发展到对多个细胞和组织甲基化图谱的检测与分析. 人类表观基因组协会于 2000 年 10 月发起了人类表观基因组计划 (Human Epigenome Project), 旨在识别并理解人类主要生理组织基因的 DNA 甲基化模式, 该计划于 2006 年完成了 12 个人体组织第 6, 20 和 22 号染色体的检测工作^[7]. 美国国立卫生研究院 (National Institutes of Health, NIH) 2008 年启动的表观基因组学路线图计划 (Roadmap Epigenomics Program) 目标之一是拟对 100 种以上的人体细胞和组织的 DNA 甲基化图谱进行检测和研究, 并发现差异的 DNA 甲基化模式^[8].

本文作者于 2009 年撰写了 DNA 甲基化的生物信息学研究情况的综述性文章^[9], 近几年, Fouse^[10], Laird^[11], Bock^[12], Ma^[13] 及 Wilhelm-Benartzi^[14] 等人分别对基因组尺度的多种 DNA 甲基化实验检测方法、前期数据处理方法及在癌症中的应用进行了专题性

引用格式: 凡时财, 李承哲, 裴云飞. 人类基因组 DNA 甲基化数据分析的研究现状. 中国科学: 生命科学, 2015, 45: 450–459

Fan S C, Li C Z, Pei Y F. DNA methylome data analysis in human genome. SCIENTIA SINICA Vitae, 2015, 45: 450–459, doi: 10.1360/N052015-00009

的综述. 随着表观遗传学的快速发展, 特别是 DNA 甲基化图谱研究的进一步深入, 亟需对高通量 DNA 甲基化的实验检测与处理方法、计算预测方法及 DNA 差异甲基化的计算分析方法等方面的最新进展进行系统梳理, 便于学者对该领域的最新研究成果与进展有系统的了解与追踪, 并开展进一步深入的研究.

从 DNA 甲基化等基本定义入手, 重点分析了 CpG 岛的多种定义方法以及搜索算法, 再系统描述了目前常用的多种高通量 DNA 甲基化检测方法; 在对 DNA 甲基化水平的预测研究中, 分别对 CpG 位点及 CpG 岛的甲基化水平预测方法进行了对比分析; 对 DNA 甲基化在不同组织之间、癌症之间以及其他疾病间差异模式的研究状况进行了阐述; 最后对 DNA 甲基化研究中需要进一步研究的问题进行了展望.

1 基本定义

1.1 DNA 甲基化

人类基因组中, DNA 甲基化主要指在 DNA 甲基化转移酶(DNA methyltransferase, DNMT)作用下, CpG 双核苷酸胞嘧啶的第 5 位碳原子上共价结合一个甲基基团^[15]. DNA 甲基化状态的形成主要包含 2 种形式: 在 DNA 甲基化转移酶 DNMT3a 和 DNMT3b 催化作用下的从头甲基化(*de novo* DNA methylation)以及在甲基化酶 DNMT1 参与下的维持性甲基化(maintenance DNA methylation).

1.2 DNA 去甲基化

DNA 去甲基化也被称为 DNA 甲基化丢失(loss of DNA methylation), 即甲基基团从胞嘧啶上消失的过程. 包含主动去甲基化与被动去甲基化 2 种模式.

1.3 CpG 岛

CpG 岛最早是由 Tykocinski 和 Max^[16]发现并定义的, 指包含限制性内切酶 *Hpa* II 结合位点的小区域, 因此也被称为 *Hpa* II 小片段岛. 后来, Gardiner-Garden 和 Frommer^[17]给出了量化的定义: 长度不小于 200 bp、GC 含量不小于 50%、CpG 实际含量与期望含量之比不小于 0.6 的区域. 由于上述定义包含了许多短的重序列, Takai 和 Jones^[18]将 CpG 岛进行了重新的定义: 长度不小于 500 bp、GC 含量不小于

55%、CpG 实际含量与期望含量之比不小于 0.65. 基于上述定义, 出现了一些用以提供 CpG 岛检测的方法. 例如, CpGProD 采用滑动窗口提取 CpG 岛, 能显示 CpG 岛与转录起始位点间的位置关系^[19]; CpG Island searcher 对间隔小于 100 bp 的 CpG 岛进行合并, 并提供 CpG 双核苷酸分布的图示^[18].

Hackenberg 等人^[20]认为, CpG 岛中 CpG 双核苷酸之间的距离分布会与其余 DNA 序列有所不同, 因此将在假设检验中显著聚集的区域定义为 CpG 岛, 此方法(CpGcluster)的检测效率极高; Bock 等人^[21]则引入多种表观遗传特征(如 DNA 甲基化、组蛋白修饰等)作为是否是 CpG 岛的判断指标, 通过量化 CpG 岛的“强度”进行定义, 此种定义更具有生物学的价值; Su 等人^[22]通过采用相邻 CpG 位点的互信息开发了功能 CpG 岛的预测方法 CpG_MI, 获得了最高的预测精度; Chuang 等人^[23]结合转录因子起始位点开发了 CpG 岛的分析预测平台 CpGPAP, 并提供了在线预测服务. Wu 等人^[24]基于隐马尔科夫模型, 采用概率打分的方式对特定序列是否为 CpG 岛进行了计算分析, 这种非二值化的定义可便于用户灵活应用.

相对而言, Gardiner-Garden 和 Frommer^[17]给出的 CpG 岛定义使用最为广泛. 但由于 CpG 岛本身属于具有特定生物意义的术语, 因此, 具体采用何种定义以及何种搜索策略, 则需要研究者结合自己的研究目的加以灵活选择.

2 DNA 甲基化的高通量检测方法

全基因组水平的 DNA 甲基化的研究得益于各种高通量 DNA 甲基化实验检测手段的发展与进步, 如今的 DNA 甲基化方法已经实现了单个碱基分辨率的甲基化水平检测. 总体而言, DNA 甲基化的实验检测可以分为前期的样品处理以及第二阶段目标序列定位与甲基化水平量化 2 个步骤.

前期的样品处理根据所用试剂的不同, 可以分为甲基化敏感的限制性内切酶(methylation-sensitive restriction enzymes, MRE)、5-甲基胞嘧啶抗体(5-methylcytosine antibody)、结合甲基化 DNA 的蛋白(methylated DNA binding proteins)、结合非甲基化 DNA 的蛋白(proteins binding unmethylated DNA)和亚硫酸氢钠(sodium bisulfite)等. 甲基化敏感的限制性内切酶相对便宜, 也是较早应用于 DNA 甲基化检测的试剂, 但是特定的内切酶只能识别相应序列背景

条件下的 CpG 位点, 该缺陷可以通过多种限制性内切酶组合的方法对样本进行并行处理, 实现 CpG 位点的识别与剪切. 5-甲基胞嘧啶抗体能够在不区别 DNA 背景序列的情况下实现对甲基化 DNA 区域的识别与绑定, 这类试剂购买与使用都很方便, 该方法的特点是对 CpG 含量较高的区域具有较强的结合能力. 此方法的检测分辨率由扩增前随机打断的 DNA 片段长度决定. 亚硫酸氢钠法被称为 DNA 甲基化检测的金标准, 它通过把非甲基化的胞嘧啶转化为尿嘧啶, 进而在 PCR 扩增过程中变为胸腺嘧啶, 而甲基化的胞嘧啶保持不变. 亚硫酸氢钠法具有单碱基分辨率的优点, 并且获得的是 CpG 甲基化的绝对水平.

DNA 甲基化实验检测的目标序列定位与甲基化水平的量化大致可以分为基于凝胶电泳的分析(gel-based)、基于芯片的分析(array-based)以及基于下一代测序技术(next-generation sequencing-based)的分析. 同第一阶段的多种试剂方法相结合, 形成了多种检测方法. 其中的代表性方法见表 1.

其中, Illumina 公司基于 Infinium 平台的 450K 甲基化芯片在各种疾病、组织特异性表达等多种领域研

究中应用极为广泛^[33,41]. 基于芯片数据获得的 DNA 甲基化数据同以往用以检测基因表达以及转录水平的芯片数据类似, 需要进行图像处理以及数据归一化等处理步骤^[42,43]. 特别是针对 450K 芯片数据, 开发了若干用以数据预处理及数据分析的工具软件, 具体见表 2. 经过上述方法处理后的芯片数据用各个探针的相对甲基化水平 β 或者经过对数处理后的 M 值表示.

基于下一代测序技术的检测方法因为检测精度较高, 愈发受到行业内的关注. 全基因组的亚硫酸氢钠测序法(whole-genome shotgun bisulphate sequencing, WGBS)是完全实现全基因组单碱基分辨率的甲基化检测方法^[40], 该方法是目前进行多组织、多疾病中 DNA 甲基化模式认知的最可靠手段. 但该测序方法极其昂贵, 而且亚硫酸氢钠法处理后样本的序列复杂度降低. 还有其他相对性价比较高的基于测序的 DNA 甲基化检测方法, 如基于限制性内切酶的测序方法(MRE-Seq)^[34,51]、基于 5-甲基胞嘧啶单克隆抗体的测序方法(MeDIP-Seq)、简化代表性亚硫酸氢钠测序技术(reduced representation bisulfite sequ-

表 1 高通量 DNA 甲基化检测的代表性方法

方法	甲基化敏感的限制性内切酶	5-甲基胞嘧啶抗体	亚硫酸氢钠
基于凝胶电泳的方法	基因组限制性酶切扫描法(restriction landmark genomic scanning, RLGS) ^[25,26]		Sanger BS(bisulfite-sequencing) ^[17]
基于芯片技术的方法	MCAM(methylated CpG island amplification and microarray) ^[27] CHARM(comprehensive high-throughput arrays for relative methylation) ^[24] HELP(hpall tiny fragment enrichment by ligation-mediated PCR) ^[28]	MeDIP(methylated DNA immunoprecipitation) ^[29] mDIP(methyl DNA immunoprecipitation) ^[30] MIRA(methylated-CpG island recovery assay) ^[31]	GoldenGate ^[32] Infinium ^[33]
基于下一代测序技术的方法	Methyl-seq ^[34] HELP-seq ^[35] MSCC(methylation sensitive cut counting) ^[36]	MeDIP ^[7] MethylCAP-seq(methyl capture sequencing) ^[37]	简化代表性亚硫酸氢钠测序(RRBS) ^[38] BC-seq ^[39] 全基因组的亚硫酸氢钠测序法(WGBS) ^[40]

表 2 DNA 甲基化芯片数据预处理的主要方法

方法	链接	主要功能	开发语言
IMA ^[44]	http://ima.r-forge.r-project.org/	异常探针数据剔除, 芯片数据片内、片间归一化	R
Lumi ^[45]	www.bioconductor.org	背景颜色的校正, 芯片数据片内、片间归一化	R
MethLAB ^[46]	http://genetics.emory.edu/conneely/MethLAB	片内以及批次进行预处理, 并与相关的表型进行关联分析	R
minfi ^[47]	http://bioconductor.org/packages/release/bioc/html/minfi.html	芯片数据前期预处理、质量评估以及对特异甲基化位点进行分析与绘图	R
ChAMP ^[48]	http://www.bioconductor.org/packages/release/bioc/html/ChAMP.html	芯片数据的归一化, 质量控制, 差异甲基化区域的获取以及拷贝数异常的分析	R
wateRmelon ^[49]	http://mammoet.iop.kcl.ac.uk/i450/	探针数据偏差校正, 芯片数据片内、片间归一化	R
Methylumi ^[50]	http://www.bioconductor.org/	专注于荧光颜色层面的背景建模, 属于最前端的数据处理工作	R

encing, RRBS)等. RRBS 基于 *Msp* I 酶切原理与下一代测序技术结合, 实现单碱基的甲基化图谱检测^[38,52]. 针对 Bisulfite-seq 实验获取单碱基甲基化水平的关键步骤是序列的比对技术. 目前, 面向该实验手段开发的序列比对算法、软件有 10 余种, 其比对策略总体上可以分为通配符策略(wildcard strategy)与三碱基策略(three-letter strategy) 2 大类. 通配符策略的主要特点是允许读段中的 T 作为通配符比对到基因组上的 C 或者 T 的位置, 但是读段中的 C 必须比对到基因组上的 C 的位置, 相关的软件主要有 RMAPBS^[46], BSMAP^[40]和 Last^[53]等, 此策略得到的 DNA 甲基化水平可能会偏高. 三碱基策略的主要特点是将读段以及参照基因组序列中的 C 和 T 看作同一字符, 即将 4 个碱基转化为 3 个碱基, 再利用常规的序列比对软件进行比对. 该策略可以实现甲基化水平的无偏估计, 主要的比对软件有 BS Seeker^[54,55], Bismark^[56], Brat^[57]和 MethylCoder^[58]等.

3 DNA 甲基化水平预测

采用机器学习的方法对 DNA 甲基化进行预测是 DNA 甲基化生物信息学研究的一大热点, 也是目前全基因组 DNA 甲基化实验检测手段费用相对高昂的情况下, 了解与分析多组织、多细胞、全基因组 DNA 甲基化模式较为便捷的手段. 因此, DNA 甲基化水平预测方法的精度与准确度成为衡量计算方法是否能够推广的关键因素. DNA 甲基化水平的预测主要包括对 CpG 位点及 CpG 岛(或者是 CpG 富集的区域)甲基化水平的预测.

3.1 CpG 位点甲基化水平预测

2005 年, Bhasin 等人^[59]在 *FEBS Letter* 杂志上发表了第一篇 CpG 位点甲基化状态预测的文章, 他们选取 CpG 位点周围 39 bp 的序列特征, 并用长度为 5 的二进制代码进行编码, 采用支持向量机进行预测, 获得了约 84% 的曲线下面积(area under curve, AUC)准确率. Stevens 等人^[60]通过融入 MeDIP-seq 与 MRE-seq 数据特征对人类基因组 CpG 位点的甲基化状态进行预测, 创新性开发了条件随机场模型, 该模型的预测结果与 WGBS 数据的相似度很高, 并采用实验测序的方法验证了该算法的有效性.

3.2 CpG 岛甲基化水平预测

相比而言, 对于 CpG 岛甲基化状态的预测工作更多. Feltus 等人^[61]早在 2003 年就尝试采用 DNA 序列特征对 CpG 岛的甲基化状态进行预测, 得到了 82% 的正确率, 并发现了 7 种区别 2 类 CpG 岛的序列模式. 他们的研究还发现, 易于甲基化和不易于甲基化的 CpG 岛在长度、GC 含量、CpG 比例等方面没有显著差异, 由于该结论来自生物体外的过量 DNA 甲基化转移酶的处理数据, 因此和来自体内的甲基化数据得到的结论有所悖离. Fang 等人^[62]开发了 CpG 岛片段甲基化状态预测工具 MethCGI, 通过引入转录因子结合位点信息, 该模型准确率达到 84%, 对 MethDB 中来自多个组织的甲基化数据进行预测, 也获得了较好的效果. 之后, Cokus 等人^[63]通过结合伪三核苷酸组合(pseudo trinucleotide composition)得到的 DNA 物理化学特性构建了朴素贝叶斯模型用以预测 CpG 岛的甲基化状态, 获得了 88% 的准确率, 表明了该方法的有效性.

随着诸如组蛋白修饰等其他表观基因组学的发展, 相关的表观遗传信息也被用以 DNA 甲基化的预测中. Bock 等人^[64]通过增加预测的组蛋白修饰状态以及染色质结构特征, 对 CpG 岛甲基化进行了预测, 并开发了在线预测工具 EpiGRAPH, 可实现对待分析序列的各种特征提取以及基于支持向量机、随机森林以及 logistic 回归等方法的甲基化状态预测, 在业内具有良好的影响力. 本课题组^[65,66]通过融入组蛋白甲基化修饰数据, 对 CpG 岛的甲基化状态进行了预测, 达到了当时最高的预测效果, 同时发现与验证了部分组蛋白修饰与 DNA 甲基化的关系: H3K4me1, H3K4me2, H3K4me3 在非甲基化 CpG 岛附近富集度较高, 表明部分与基因表达正相关的组蛋白修饰常常协同出现以保证具有活性的染色质区域得以保持稳定的活性状态. 研究表明, 组蛋白修饰数据特征对 CpG 岛甲基化状态预测的精度有较大贡献, 但由于目前公开发表的基因组水平的组蛋白修饰特征数据并不多, 因此也局限了此类方法在多个细胞组织中的应用.

综上所述, DNA 甲基化预测的工作主要在预测所采用的特征以及分类器的选择上具有一定差异, 具体总结见表 3. 待分析 CpG 位点或 CpG 岛局部区域的序列特征(k-mer、重复序列特征、转录因子结合位点、序列保守性等)对其甲基化水平具有较强的影

表 3 DNA 甲基化水平预测工具概况

采用的特征	使用的分类器	工具
DNA 序列特征	支持向量机 ^[59,62]	http://bio.dfci.harvard.edu/Methylator/MethCGI (http://bioinfo.au.tsinghua.edu.cn/MethCGI/MethCGI.html)
	logistic 回归 ^[69,70]	HDMFINDER
	线性判别分析 ^[61,70]	
	朴素贝叶斯模型 ^[63]	
DNA 序列特征+组蛋白修饰特征	随机森林, logistic 回归, 支持向量机 ^[64]	EpiGRAPH(http://epigraph.mpi-inf.mpg.de/WebGRAPH/)
	支持向量机 ^[65,66,71]	CpGMethPred (http://users.ece.gatech.edu/~hzheng7/CGIMetPred.zip)
DNA 序列特征+部分甲基化水平特征	条件随机场 ^[60]	MethylCRF(http://methylCRF.wustl.edu)

响,特别是上下游 400 bp 的序列特征^[67]。然而, DNA 序列相同的不同组织在甲基化模式上存在差异,这表明仅从序列特征进行甲基化水平预测具有较大局限性。再者,还有研究表明,表观遗传修饰的改变可能会先于基因组序列本身的改变^[68]。因此,其他表观遗传形式(组蛋白修饰、核小体等特征)较序列特征而言,与 DNA 甲基化水平有更强的特异相关性,体现了 DNA 甲基化与其他表观遗传形式相辅相成的关系。此外,若干非全基因组尺度的甲基化检测方法所提供的甲基化数据本身也可以作为特征对全基因组的其他位点进行预测,并具有应用于不同组织预测的潜能。可以预见,Stevens 等人^[60]的研究思路将成为后续的研究热点。

4 DNA 差异甲基化的研究

一直以来, DNA 差异甲基化 (differential methylation)被认为是影响基因特异表达的重要因素之一^[72],对其研究的关注度也不断升温。研究方法也从单位点的对比分析发展为对多个体、多年龄段、多组织、多种细胞系及多种疾病的基因组尺度的研究。

对于 DNA 差异甲基化研究,需要首先探讨以下 3 个问题:多种组织/疾病样本甲基化数据的可靠性,待考察目标区域的确定,以及用什么方法定义差异。

为提高数据的可靠性,第 2 小节中对于基因芯片的各种片内归一化、片间归一化方法以及多批次校正的处理方法需进一步关注。确定目标区域的方法中,滑窗法最为简单^[40],但该方法在全基因组甲基化的数据集上显得力不从心。VanderKraats 等人^[73]则采用针对基因启动子区附近的固定区域进行差异化检测

的方法,此方法实现简单,但检测的敏感度受选定的区域长度影响较大。Su 等人^[74]提出基于甲基化热点并延伸的方法确定考察区域,此方法获得的区域可根据 DNA 甲基化的具体情况灵活改变,但是算法过程比较复杂,计算速度相对较慢。

目前采用的定义差异甲基化的方法在组织差异化、癌症等疾病的差异甲基化研究中皆有应用。

4.1 组织差异甲基化研究

Byun^[75], Nagae^[76]和 Fernandez^[77]等人分别在各自的研究中通过非监督的聚类等方法发现了能够区分和反映不同组织特性的 DNA 甲基化特征,这些差异甲基化的位点通常分布在非 CpG 岛的 5'端的尾部 (non-CpG island 5' end)^[78]。Zhang 等人^[79]采用基于香农熵的分析方法,对多组织中 DNA 差异表达的阈值进行了概率建模分析,得到了更为合理的划分阈值,他们发现,基于该阈值得到的组织差异表达区域在基因功能注释分析中,具有更好的生物学解释。Wang 课题组^[80]采用统计分析方法结合 MeDIP-seq 和 MRE-seq 数据(简称 M&M 方法),对 19 个人类 DNA 甲基化谱样本(来自 4 个组织的 9 个细胞系)进行了分析。他们构建一系列的假设统计分析模型,通过结合 DNA 甲基化水平、CpG 含量等信息得到了更为准确的差异甲基化区域。结果表明,差异表达的 DNA 甲基化区域能够较好的对应上不同的个体、细胞类型以及组织类型。

4.2 癌症差异甲基化研究

TCGA 协会(The Cancer Genome Atlas Consortium)对多于 5000 个组织的 DNA 甲基化数据进行了

检测, 抑癌基因启动子区 CpG 岛的超甲基化被认为是很多癌症产生的重要因素^[81,82], DNA 甲基化谱具有区分癌症类型、特定癌症的亚型从而实现诊断、个性化治疗、预后以及临床试验等方面的潜在价值. Turcan 等人^[83]对百余个实验样本的 DNA 甲基化谱数据, 采用方差分析法、FDR 校正后的显著性阈值发现, *IDH1* 基因的突变会导致整个 DNA 甲基化模式的变化, 从而诱发神经胶质瘤; 此外, 利用甲基化数据实现了神经胶质瘤亚型的划分. Zhang 课题组^[13]先后开发了基于分级聚类以及基于协方差分析的统计模型实现差异甲基化检测, 在多种癌症的 450K 芯片数据上获得了较好的结果^[80]. Fernandez 等人^[77]对 855 个原发癌的 DNA 甲基化数据分析也发现, 不同组织来源的原发癌的 DNA 甲基化模式有显著差异, 会在聚类热图中表现出明显的模式特征, 在多种假设检验下(Fisher's exact test, Kruskal-Wallis test, Wilcoxon test)均表现出显著性极高的差异. 但是, 目前针对癌症的研究多采用来自细胞系的样本, 并且所研究的癌症类型并没有同 TCGA 组织中研究的癌症类型相统一.

4.3 其他疾病中差异甲基化研究

目前有关 DNA 异常甲基化模式的报道多集中在癌症类疾病中, 其他非癌症类疾病中的状况由于可利用的数据相对较少, 研究并不深刻. 从已有的报道中可发现, Desplats 等人^[84]在路易体痴呆疾病中发现维持性 DNA 甲基化转移酶(DNMT1)表达水平降低, 从而导致脑组织中甲基化水平的整体偏低, 进而诱发该疾病. Javierre 等人^[85]采用 GoldenGate 检测技术对 30 对患有自身免疫性疾病的同卵双胞胎及 30 对年龄性别相当的正常个体进行研究, 通过散点图、*t* 检验以及 GO(gene ontology)分析发现, 系统性红斑狼疮的发生对应很大数量基因的 DNA 甲基化状态的改变. Nakano 等人^[86]基于 450K 芯片数据对 6 个类风湿性关节炎成纤维样滑膜细胞组织与 5 个骨关节炎成纤维样滑膜细胞组织进行检测, 通过超甲基化位点富集度(enrichment)的显著性分析(假设为超几何分布), 提取出 1859 个差异甲基化的位点, 再结合 KEGG(Kyoto encyclopedia of genes and genomes)通路及 GO 注释分析结果, 验证了结果的有效性, 基于支持向量机的分类结果则进一步证实了上述方法的正

确性. 研究发现, 与类风湿性关节炎相关的重要基因表现出欠甲基化状态, 而 *TGFB2* 与 *FOXO1* 等基因则表现为超甲基化. 进一步的通路分析表明, 差异甲基化的基因同类关节炎的发病机理密切相关.

综上, 在正态分布、超几何分布假设下进行 *t* 检验或者方差分析, 或者基于非参数的假设检验计算显著性指标是目前常用的定义差异的方法, 基于香农熵的方法则为差异化方法的研究提供了新的方向. 结合待分析的组织来源及样本数量对结果进行有效校正, 是显著性评判中需重点考虑的问题. 采用分级聚类等非监督学习的方法则为差异位点在多种类别中的数值显示提供了直观的图示工具^[87,88], 结合信号通路及 GO 分析则为进一步验证分析所得到的差异甲基化位点提供了生物学依据. 此外, 基于混杂细胞的 DNA 甲基化数据如何分解得到单一细胞的甲基化水平数据进而开展差异甲基化的研究也是亟需解决的问题^[71].

5 展望

虽然 DNA 甲基化在近 10 年的研究与分析中取得了长足进步与丰硕成果, 但是离系统性地理解 DNA 甲基化的发生、变化及功能的目标还非常遥远. 全基因组单碱基精度的 DNA 甲基化图谱是探索这一问题的必经途径. 因此, 改进与发展以下一代测序技术为基础的性价比更高的高通量检测方法显得尤为重要. 通过对更多正常组织与细胞、多种癌症以及癌症多个阶段的 DNA 甲基化图谱的检测, 剖析 DNA 甲基化在细胞分化、疾病的发生与发展中的作用对于干细胞研究、癌症的诊断与治疗等都有积极作用. 对于从事计算分子生物学、生物信息学的科学工作者而言, 更需要在数据挖掘方面开发新的软件工具包. 例如, 如何减小因实验样本来源、样本批次及实验噪声所导致的人为偏差, 基于标准格式定义的 DNA 甲基化图谱分析软件快速准确地得到 DNA 甲基化水平, 通过结合多种 DNA 甲基化检测技术获得的甲基化数据对 DNA 甲基化进行准确的预测, 融合多种表观遗传信息进行多源数据的关联分析与推理, 开发更为准确有效的识别具有功能意义的差异甲基化区域等. 总之, 只有来自多学科的科研工作者深度合作, 才能开启 DNA 甲基化研究的新的篇章.

参考文献

- 1 Wyatt G R. Occurrence of 5-methylcytosine in nucleic acids. *Nature*, 1950, 166: 237–238
- 2 Reik W, Santos F, Dean W. Mammalian epigenomics: reprogramming the genome for development and therapy. *Theriogenology*, 2003, 59: 21–32
- 3 Barlow D P. Genomic imprinting: a mammalian epigenetic discovery model. *Annu Rev Genet*, 2011, 45, 45: 379–403
- 4 Meissner A. Epigenetic modifications in pluripotent and differentiated cells. *Nat Biotechnol*, 2010, 28: 1079–1088
- 5 Jones P A, Baylin S B. The fundamental role of epigenetic events in cancer. *Nat Rev Genet*, 2002, 3: 415–428
- 6 Feinberg A P. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 2007, 447: 433–440
- 7 Eckhardt F, Lewin J, Cortese R, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, 2006, 38: 1378–1385
- 8 Ziller M J, Gu H C, Muller F, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 2013, 500: 477–481
- 9 Fan S C, Zhang X G. Progress of bioinformatics study in DNA methylation. *Prog Biochem Biophys*, 2009, 36: 143–150
- 10 Fouse S D, Nagarajan R P, Costello J F. Genome-scale DNA methylation analysis. *Epigenomics*, 2010, 2: 105–117
- 11 Laird P W. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genetics*, 2010, 11: 191–203
- 12 Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet*, 2012, 13: 705–719
- 13 Ma X T, Wang Y W, Zhang M Q, et al. DNA methylation data analysis and its application to cancer research. *Epigenomics*, 2013, 5: 301–316
- 14 Wilhelm-Benartzi C S, Koestler D C, Karagas M R, et al. Review of processing and analysis methods for DNA methylation array data. *Brit J Cancer*, 2013, 109: 1394–1402
- 15 Bird A P. CpG islands as gene markers in the vertebrate nucleus. *Trends Genet*, 1987, 3: 342–347
- 16 Tykocinski M L, Max E E. CG dinucleotide clusters in *MHC* genes and in 5' demethylated genes. *Nucleic Acids Res*, 1984, 12: 4385–4396
- 17 Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*, 1987, 196: 261–282
- 18 Takai D, Jones P A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci USA*, 2002, 99: 3740–3745
- 19 Ponger L, Mouchiroud D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, 2002, 18: 631–633
- 20 Hackenberg M, Previti C, Luque-Escamilla P L, et al. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, 2006, 7: 446
- 21 Bock C, Walter J, Paulsen M, et al. CpG island mapping by epigenome prediction. *PLoS Comput Biol*, 2007, 3: e110
- 22 Su J Z, Zhang Y, Lv J, et al. CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes. *Nucleic Acids Research*, 2010, 38: e6
- 23 Chuang L Y, Yang C H, Lin M C, et al. CpGPAP: CpG island predictor analysis platform. *BMC Genetics*, 2012, 13: 13
- 24 Wu H, Caffo B, Jaffee H A, et al. Redefining CpG islands using hidden Markov models. *Biostatistics*, 2010, 11: 499–514
- 25 Hatada I, Hayashizaki Y, Hirotsume S, et al. A genomic scanning method for higher organisms using restriction sites as landmarks. *Proc Natl Acad Sci USA*, 1991, 88: 9523–9527
- 26 Ohgane J, Aikawa J I, Ogura A, et al. Analysis of CpG islands of trophoblast giant cells by restriction landmark genomic scanning. *Dev Genet*, 1998, 22: 132–140
- 27 Toyota M, Ho C, Ahuja N, et al. Identification of differentially methylated sequences in colorectal cancer by methylated CpG island amplification. *Cancer Res*, 1999, 59: 2307–2312
- 28 Khulan B, Thompson R F, Ye K, et al. Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res*, 2006, 16: 1046–1055
- 29 Weber M, Hellmann I, Stadler M B, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*, 2007, 39: 457–466
- 30 Keshet I, Schlesinger Y, Farkash S, et al. Evidence for an instructive mechanism of *de novo* methylation in cancer cells. *Nat Genet*, 2006, 38: 149–153
- 31 Rauch T A, Zhong X Y, Wu X W, et al. High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. *Proc Natl Acad Sci USA*, 2008, 105: 252–257
- 32 Bibikova M, Lin Z W, Zhou L X, et al. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res*, 2006, 16: 383–393
- 33 Bibikova M, Lin Z, Zhou L, et al. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res*, 2006, 16:

383–393

- 34 Brunner A L, Johnson D S, Kim S W, et al. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res*, 2009, 19: 1044–1056
- 35 Oda M, Glass J L, Thompson R F, et al. High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res*, 2009, 37: 3829–3839
- 36 Berman B P, Weisenberger D J, Laird P W. Locking in on the human methylome. *Nat Biotechnol*, 2009, 27: 341–342
- 37 Rauch T, Wang Z D, Zhang X M, et al. Homeobox gene methylation in lung cancer studied by genome-wide analysis with a microarray-based methylated CpG island recovery assay. *Proc Natl Acad Sci USA*, 2007, 104: 5527–5532
- 38 Meissner A, Gnirke A, Bell G W, et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res*, 2005, 33: 5868–5877
- 39 Hodges E, Smith A D, Kendall J, et al. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res*, 2009, 19: 1593–1605
- 40 Lister R, Pelizzola M, Dowen R H, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 2009, 462: 315–322
- 41 Lokk K, Modhukur V, Rajashekar B, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol*, 2014, 15: r54
- 42 Fortin J P, Labbe A, Lemire M, et al. Functional normalization of 450K methylation array data improves replication in large cancer studies. *Genome Biol*, 2014, 15: 503
- 43 Dedeurwaerder S, Defrance M, Bizet M, et al. A comprehensive overview of Infinium HumanMethylation450 data processing. *Brief Bioinform*, 2014, 15: 929–941
- 44 Wang D, Yan L, Hu Q, et al. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics*, 2012, 28: 729–730
- 45 Du P, Kibbe W A, Lin S M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 2008, 24: 1547–1548
- 46 Kilaru V, Barfield R T, Schroeder J W, et al. MethLAB: a graphical user interface package for the analysis of array-based DNA methylation data. *Epigenetics*, 2012, 7: 225–229
- 47 Maksimovic J, Gordon L, Oshlack A. SWAN: subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol*, 2012, 13: R44
- 48 Morris T J, Butcher L M, Feber A, et al. ChAMP: 450K chip analysis methylation pipeline. *Bioinformatics*, 2014, 30: 428–430
- 49 Pidsley R, Wong C C Y, Volta M, et al. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, 2013, 14: 293
- 50 Triche T J, Weisenberger D J, Van Den Berg D, et al. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res*, 2013, 41: e90
- 51 Ball M P, Li J B, Gao Y, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol*, 2009, 27: 361–368
- 52 Meissner A, Mikkelsen T S, Gu H C, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 2008, 454: 766–770
- 53 Frith M C, Mori R, Asai K. A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic Acids Res*, 2012, 40: e100
- 54 Chen P Y, Cokus S J, Pellegrini M. BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, 2010, 11: 203
- 55 Guo W L, Fizev P, Yan W H, et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*, 2013, 14: 774
- 56 Krueger F, Andrews S R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 2011, 27: 1571–1572
- 57 Harris E Y, Ponts N, Le Roch K G, Lonardi S. BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. *Bioinformatics*, 2012, 28: 1795–1796
- 58 Pedersen B, Hsieh T F, Ibarra C, et al. MethylCoder: software pipeline for bisulfite-treated sequences. *Bioinformatics*, 2011, 27: 2435–2436
- 59 Bhasin M, Zhang H, Reinherz E L, et al. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett*, 2005, 579: 4302–4308
- 60 Stevens M, Cheng J B, Li D, et al. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res*, 2014, 23: 1541–1553

- 61 Feltus F A , Lee E K, Costello J F, et al. Predicting aberrant CpG island methylation. *Proc Natl Acad Sci USA*, 2003, 100: 12253–12258
- 62 Fang F, Fan S C, Zhang X G, et al. Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, 2006, 22: 2204–2209
- 63 Cokus S J, Feng S, Zhang X, et al. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature*, 2008, 452: 215–219
- 64 Bock C, Halachev K, Buch J, et al. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. *Genome Biol*, 2009, 10: R14
- 65 Fan S C, Zhang M Q, Zhang X G. Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochem Biophys Res Commun*, 2008, 374: 559–564
- 66 凡时财, 邹见效, 徐红兵, 等. 人类基因组 CpG 岛甲基化概况的预测. *科学通报*, 2010, 55: 1329–1334
- 67 Fan S, Fang F, Zhang X, et al. Putative zinc finger protein binding sites are over-represented in the boundaries of methylation-resistant CpG islands in the human genome. *PLoS One*, 2007, 2: e1184
- 68 Baylin S B, Herman J G, Graff J R, et al. Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv Cancer Res*, 1998, 72: 141–196
- 69 Yang Y, Nephew K, Kim S. A novel k-mer mixture logistic regression for methylation susceptibility modeling of CpG dinucleotides in human gene promoters. *BMC Bioinformatics*, 2012, 13: S15
- 70 Das R, Dimitrova N, Xuan Z Y, et al. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci USA*, 2006, 103: 10713–10716
- 71 Zheng X Q, Zhao Q, Wu H J, et al. MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biol*, 2014, 15: 419
- 72 Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev*, 2002, 16: 6–21
- 73 VanderKraats N D, Hiken J F, Decker K F, et al. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic Acids Res*, 2013, 41: 6816–6827
- 74 Su J Z, Yan H D, Wei Y J, et al. CpG_MPs: identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic Acids Res*, 2013, 41: e41
- 75 Byun H M, Siegmund K D, Pan F, et al. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum Mol Genet*, 2009, 18: 4808–4817
- 76 Nagae G, Isagawa T, Shiraki N, et al. Tissue-specific demethylation in CpG-poor promoters during cellular differentiation. *Hum Mol Genet*, 2011, 20: 2710–2721
- 77 Fernandez A F, Assenov Y, Martin-Subero J I, et al. A DNA methylation fingerprint of 1628 human samples. *Genome Res*, 2012, 22: 407–419
- 78 Illingworth R S, Bird A P. CpG islands—‘A rough guide’. *FEBS Lett*, 2009, 583: 1713–1720
- 79 Zhang Y, Liu H B, Lv J, et al. QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res*, 2011, 39: e58
- 80 Li D, Zhang B, Xing X, et al. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. *Methods*, 2015, 72: 29–40
- 81 Jones P A, Baylin S B. The epigenomics of cancer. *Cell*, 2007, 128: 683–692
- 82 Esteller M. Molecular origins of cancer: epigenetics in cancer. *New Engl J Med*, 2008, 358: 1148–1159
- 83 Turcan S, Rohle D, Goenka A, et al. *IDH1* mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*, 2012, 483: 479–483
- 84 Desplats P, Spencer B, Coffee E, et al. α -synuclein sequesters Dnmt1 from the nucleus: a novel mechanism for epigenetic alterations in Lewy body diseases. *J Biol Chem*, 2011, 286: 9031–9037
- 85 Javierre B M, Fernandez A F, Richter J, et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res*, 2010, 20: 170–179
- 86 Nakano K, Whitaker J W, Boyle D L, et al. DNA methylome signature in rheumatoid arthritis. *Ann Rheum Dis*, 2013, 72: 110–117
- 87 Assenov Y, Muller F, Lutsik P, et al. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods*, 2014, 11: 1138–1140
- 88 Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*, 2012, 13: R87

DNA Methylome Data Analysis in Human Genome

FAN ShiCai^{1,2}, LI ChengZhe¹ & PEI YunFei³

1 College of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China;

2 Innovation Institute of Big Data and Intelligent Information System of Sichuan Province, Chengdu 611731, China;

3 Department of Automation, Tsinghua University, Beijing 100084, China

As the most common type of epigenomic modifications in human genome, DNA methylation plays crucial roles in many key biological activities. It is meaningful to investigate the DNA methylation landscape. In this review, some fundamental and controversial definitions are introduced in the first part, and the genome-wide detection and analysis methods are investigated, especially the array-based and sequence-based methods; the computational prediction models of methylation levels of CpG sites and CpG islands are highlighted in detail, and the vital influence of applied features for prediction performances is analyzed. Finally, the computational analyses of differential methylation in tissues and various diseases including cancer are discussed.

DNA methylation, detection technology, prediction, differential methylation, CpG island

doi: 10.1360/N052015-0009