

第 1 章 引 言

1.1 问题的提出

多态性 CpG 位点的 DNA 甲基化是基因表达的潜在调控物,与正常生理学功能以及肿瘤等疾病的发生发展都有密不可分的关系。全基因组 DNA 甲基化谱被应用于多种癌症亚型的分子标志鉴别研究中,然而,对甲基化组的更加深入、细致的泛癌症定量描述与分析还较有限。启动子区 DNA 甲基化组是相对稳定的还是变异程度很高的? 高变异的启动子区 CpG 位点有没有生物学意义? 启动子区 DNA 甲基化和拷贝数目多态性这两个基因表达的重要决定因素,它们之间的关系是相互加强的还是减弱的? 如果将多种癌症的 DNA 甲基化谱放在一起进行聚类分析,同一种癌症的所有肿瘤是会完全聚在一起亦或有多大程度的混杂? 聚类得到的不同聚簇有什么异同以及这些差异是否有生物学意义? 与正常组织相比,肿瘤组织的启动子区 DNA 甲基化组有什么特点? 肿瘤之间的异质性在多大程度上与来源组织有关,又有多大程度是由癌症的调节紊乱导致的? 肿瘤组织相对于正常组织的差异甲基化 CpG 位点有什么特征?

启动子区域的 DNA 甲基化通常被认为是基因沉默的标志,也有研究报道 CpG 位点甲基化可以通过招募特异性结合蛋白质来促进转录。那么启动子区 CpG 位点甲基化与基因表达的相关性分布究竟如何呢? 对于那些强正相关或者强负相关的“CpG 位点-基因对”,与 CpG 位点的固有特征有关系吗? 对于与基因表达呈不同相关性的 CpG 位点,它们到各种基因组特征的距离有差异吗? 与多个基因组区域的重合度或者共定位相似吗? 另外,与 CpG 位点甲基化呈不同相关性的基因,与这些基因的表达水平有什么关系? 这两类基因有什么功能差异呢? 启动子区 CpG 位点甲基化在转录因子对基因表达的调控过程中可能会起到什么作用? 诸如此类的细节问题,是本研究所要了解 and 深入挖掘的。

1.2 选题背景及意义

CpG 二核苷酸位点的 DNA 甲基化标记在多潜能性、发育和多种疾病中都扮演重要的角色(Robertson, 2005; Smith et al., 2013)。多项证据表明,启动子区 CpG 位点的甲基化状态改变能够驱动很多癌症相关基因的表达失调,因此是肿瘤发生和发展过程中的主要驱动因素之一(Chatterjee et al., 2012; Jones, 2012)。全基因组 DNA 甲基化谱被广泛用于多种癌症亚型的分子标志物中,包括乳腺癌(The Cancer Genome Atlas Research, 2012b)、肺癌(The Cancer Genome Atlas Research, 2012a, 2014)和卵巢癌(The Cancer Genome Atlas Research, 2011)等。然而,对于 DNA 甲基化组的更加细致的泛癌症定量分析还很有限。之前有研究开发了一种二重聚类的数据整合策略(cluster of cluster assignments, COCA),用于 12 种癌症的泛癌症多组学数据聚类,发现了大规模的癌症的汇聚(Hoadley et al., 2014)。然而,尽管 DNA 甲基化组在这个数据整合分析中被使用了,却不是研究的重点,也没有起到不可或缺的作用(Hoadley et al., 2014)。

癌症基因组图谱(The Cancer Genome Atlas, TCGA)是由美国国家癌症研究所(National Cancer Institute, NCI)和美国国家人类基因组研究所(National Human Genome Research Institute, NHGRI)合作的项目,提供了 33 种癌症种类的外显子组、单核苷酸多态性、甲基化、RNA 表达、小 RNA 和临床诊断数据,极大便利了癌症研究团体对于癌症的预防、诊断和治疗,也使得泛癌症的比较、整合分析成为可能。因为有配对肿瘤组织和正常组织的大样本规模、癌症类型覆盖广泛、可靠、一致的高通量数据处理流程以及患者的生存数据,TCGA 提供了一个价值很高的资源,使研究者可以在泛癌症水平上设计特定的大数据分析,从而挖掘有生理学意义的生物学洞见。

肿瘤组织的 DNA 甲基化模式面临着大规模的紊乱,这种失调反映在甲基化谱和其他组学数据里,进而被 TCGA 等项目所测定。本研究即是在 TCGA 项目提供的甲基化组、转录组、基因组、生存资料等公开数据的基础上,结合 DNA 元素百科全书(Encyclopedia of DNA Elements, ENCODE)项目、基因表达综合数据库(Gene Expression Omnibus, GEO)等其他层次的数据,通过整合分析的手段了解 DNA 甲基化组在癌症中的角色。通过在泛癌症上下文中探究肿瘤间 DNA 甲基化的异质性,发现了新的癌症种

类汇聚和癌症亚型分类。启动子区 CpG 位点甲基化的多态性富集在转录因子基因中,提示了转录异质性的可能的来源。“CpG 位点-基因对”的关联性分析发掘出与前人的假设相反的、非典型的基因激活通路。总之,本研究对癌症中启动子区 DNA 甲基化水平进行了综合全面的研究,针对癌症基因组和表观遗传组产生了一系列有洞见的资源,是进一步探究肿瘤发生和发展的基因表达调控的基础。

1.3 文献综述

1.3.1 基因组 DNA 甲基化模式

核苷酸甲基化是一种可逆性地标记基因组 DNA 的分子途径。原核生物可以通过甲基化胸腺嘧啶或胞嘧啶来识别和降解入侵 DNA、追踪错配修复和基因组重复事件(Arber et al.,1962; Wion et al.,2006)。而在真核生物中,DNA 甲基化只能发生在胞嘧啶残基。本研究是关于人类癌症的整合组学项目,因此本书只介绍胞嘧啶甲基化模式。

1.3.1.1 CpG 位点甲基化

DNA 甲基化可以改变调控区域的功能状态,但是无法改变“沃森-克里克”(Watson-Crick)碱基配对,因此是典型的“表观遗传”标记。“CpG”这个标记指的是从 DNA 序列的 5'端到 3'端,分别是胞嘧啶和鸟嘌呤,中间被一个磷酸基团相连,同时也能与 C : G 配对区分开来。1975 年,两篇关键文献独立地提出 CpG 二核苷酸上下文中胞嘧啶残基的甲基化可以作为脊椎动物的表观遗传标记(Holliday et al.,1975; Riggs,1975)。这些文献指出,序列可以被从头(*de novo*)甲基化,甲基化可以通过体细胞分裂而遗传(需要一种识别半甲基化 CpG 回文序列的酶),甲基基团的存在可以被 DNA 结合蛋白质解读等,这些基本知识到现在仍然是认识 DNA 甲基化功能的框架。很多进化枝都会有基因组的甲基化,不过在其中脊椎动物是最特别的,因为它们的胞嘧啶甲基化是发生在整个基因组的,仅有启动子区域处于非甲基化状态。图 1.1 比较了典型的非脊椎动物和脊椎动物基因组中甲基化状态的胞嘧啶碱基的分布。

基因组中分布广泛的、默认状态的 DNA 甲基化,导致了不充分碱基切除修复,以及随后的 CpG 位点缺失,并且是随着进化时间而越发严重的。

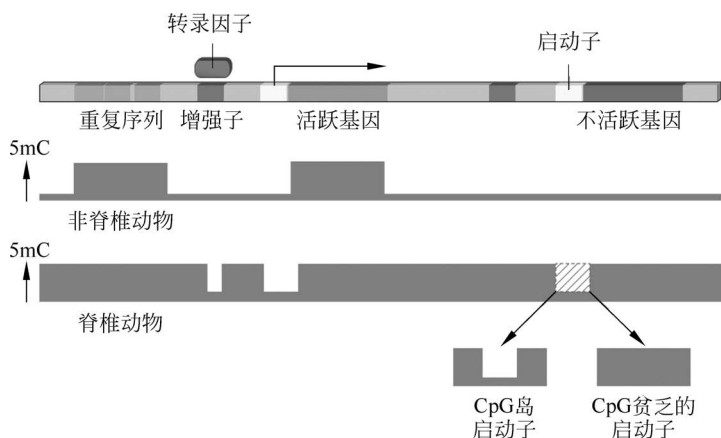


图 1.1 典型的非脊椎动物和脊椎动物基因组中胞嘧啶甲基化的分布 (Schubeler, 2015)

示例的基因组区域包括具有近端(启动子)和远端(增强子)调控区域的活跃基因和不活跃基因。灰色条棒的高低指示了每个区域观察到的 DNA 甲基化的相对比例。CpG 岛通常与启动子区域重合,一般是非甲基化状态的,而 CpG 贫乏的启动子在不激活时是甲基化状态的。
5mC: 5-甲基胞嘧啶

具体来说,胞嘧啶自发脱氨基后变成尿嘧啶,能够被尿嘧啶 DNA 糖苷酶(uracil-DNA glycosylase,UDG)高效地去除,这样维持了 DNA 稳定。但是,处于甲基化状态的胞嘧啶脱氨基得到胸腺嘧啶,这是正确的基因组碱基,不过会造成 T : G 错配。细胞中存在胸腺嘧啶 DNA 糖苷酶(thymine-DNA glycosylase,TDG)和具有甲基 CpG 结合结构域的 MBD4 蛋白质(methyl-CpG-binding domain protein 4),能够靶向这种特殊的错配并去除胸腺嘧啶(Cortazar et al.,2011)。然而,这个矫正机制的效率不够高,而 T : G 错配能很快地导致突变,产生 C 到 T 的转换(transition)。这种转换是人类疾病中以及亲缘关系近的哺乳动物中最常见的突变,佐证了基因组 DNA 甲基化的代价是突变负荷的增加。除哺乳动物以外,其他脊椎动物(如鱼和非洲爪蟾)中,基因组的 CpG 位点的缺失就没有那么明显了(Long et al.,2013b)。

在脊椎动物中,可遗传的甲基化只发生在 CpG 二核苷酸上。CpG 序列的对称性使得这种甲基化标记可以在细胞分裂中传递(Song et al.,2011)。其他序列的甲基化广泛存在于植物和某些真菌中(Cokus et al.,2008; Rountree et al.,1997),在哺乳动物中也有一些报道(Lister et al.,2009)。人的干细胞中也发现了 CHG 位点和 CHH 位点(H 代表 A、C 或

T)的甲基化,它们富集在与基因表达直接相关的基因体区,而在蛋白质结合位点和增强子区缺失(Lister et al.,2009)。非 CpG 甲基化的水平在分化过程中下降,在诱导多潜能干细胞中恢复,说明其在多潜能状态的起源与维持中的关键角色(Laurent et al.,2010; Lister et al.,2009),不过这种非 CpG 甲基化的机制仍然不清楚(Laurent et al.,2010),只是有观察表明,非 CpG 岛甲基化比 CpG 岛甲基化的动态性更强、更有组织特异性(Portela et al.,2010)。

1.3.1.2 CpG 岛及含有 CpG 岛的启动子

1.3.1.1 节中提到,DNA 甲基化几乎只存在于 CpG 二核苷酸中。而 CpG 二核苷酸在全基因的分布有两个特点:第一非常稀少,比例约为 1%,远低于根据 CC 含量所计算出的预期比例;第二分布不均匀,倾向于以成簇的形式出现,被称为“CpG 岛”(CpG island,CGI),剩下的部分则缺失 CpG 位点。CpG 岛是一段超过 200 个 bp、大约 1 kb 长的区域,其中 G+C 含量至少 50%,并且 CpG 频率的观察值与统计学期望值的比至少为 0.6。因为甲基胞嘧啶可以自发地被酶催化脱氨基化,因此基因组 CpG 的缺失应当是由生殖细胞系中甲基化序列的脱氨基导致的;CpG 岛的存在则意味着它们可能在生殖细胞系中从不或者仅仅瞬时地甲基化(Smallwood et al.,2011)。大约 60%的人类基因启动子与 CpG 岛有关联(Jones,2012),也有人认为这个比例低至 55%(Han et al.,2011)或高达 75%(Bestor et al.,2015)。通常 CpG 岛在正常细胞中是非甲基化的,其中又有约 6%在发育或者分化过程中有组织特异性的甲基化模式(Straussman et al.,2009)。而尽管这个甲基化的密度有双峰的分布,具有中间甲基化状态的区域同样存在(Takai et al.,2002)。

大部分研究聚焦在转录起始位点(transcription start site,TSS)处的 CpG 岛。当 CpG 岛启动子(指启动子区域包含 CpG 岛,下同)基因处于激活状态的时候,它们的启动子通常具有核小体缺失区域(nucleosome-depleted region,NDR)(指 DNA 没有缠绕成核小体,而是相对游离存在)的特征,并且核小体缺失区域通常被包含组蛋白变体 H2A.Z 的核小体包围,具有 H3K4me3 的组蛋白修饰标记(Kelly et al.,2010)。CpG 岛启动子可以被多种机制抑制,例如多梳蛋白质(Polycomb proteins)介导的抑制:编码胚胎发育的主要调控因子的基因,例如 MYOD1(myogenic differentiation 1)基因或 PAX6(paired box 6)基因,在胚胎干细胞和分化细胞中都被多梳复

合物抑制而不被表达；它们的转录起始位点处有核小体，被 H3K27me3 标记，这些通常都与非激活基因相关(Taberlay et al., 2011)。

虽然 DNA 甲基化会阻碍 CpG 岛启动子的转录起始，使得所连锁的基因稳定地被抑制，如图 1.2(a) 所示；然而，一些被抑制的基因的启动子区 CpG 岛并没有甲基化。甲基化状态的 CpG 岛通常只位于处于长期稳定的抑制状态的基因，例如印记(imprinting)基因(基因组印记是指两个亲代等位基因中的一个被过甲基化，导致只有一个等位基因表达)(Kacem et al., 2009; Li et al., 1993)、X 染色体失活基因(X-chromosome inactivation)(雌性哺乳动物体细胞的两条 X 染色体中的一条被稳定地沉默，从而实现剂量补偿)(Jaenisch et al., 2003; Reik et al., 2005)，以及只在生殖细胞中表达、在体细胞中很有可能不应当表达的基因。CpG 岛的 DNA 甲基化带来的稳定的抑制可以持续超过 100 年，并且对 CpG 岛的存在没有影响，因为体细胞中这段区域中的脱氨基事件不会通过生殖细胞传递到下一代。

另外，DNA 甲基化不是仅仅出现在 CpG 岛中。术语“CpG 海滨”(CpG island shores)指的是距离 CpG 岛极近的(2 kb 以内)、有较低的 CpG 密度的区域。CpG 海滨的甲基化与转录的失活密切相关，大多数组织特异性的 DNA 甲基化看起来不是发生在 CpG 岛，而是发生在 CpG 海滨(Doi et al., 2009; Irizarry et al., 2009)，如图 1.2(b)所示。差异甲基化的 CpG 海滨在人与小鼠中保守，能够区分特异性的组织。此外，重编程中 70% 的差异甲基化区域是与 CpG 海滨有关联的(Doi et al., 2009; Ji et al., 2010)。

1.3.1.3 非 CpG 岛及 CpG 贫乏启动子

基因组中除了成簇存在的 CpG 岛，还有些 CpG 位点是分散存在的。除了 CpG 岛启动子，人类基因组中剩下的启动子的 CpG 位点相对贫乏，因此称为“CpG 贫乏启动子”(CpG-poor promoter)，也有研究称之为“非 CpG 岛启动子”(non-CGI promoters)。虽然在所有基因中，含有 CpG 贫乏启动子的比例不到一半，但是 CpG 岛启动子主要出现在很多组织的看家基因(housekeeping gene)中，而在组织特异性表达的基因中，CpG 贫乏启动子的比例超过了一半(Saxonov et al., 2006)。也有研究发现，组织特异性 RNA 聚合酶经常结合到 CpG 贫乏启动子上(Barrera et al., 2008)，这些都说明了 CpG 贫乏启动子在终末分化组织中起到表观遗传调控的重要作用。

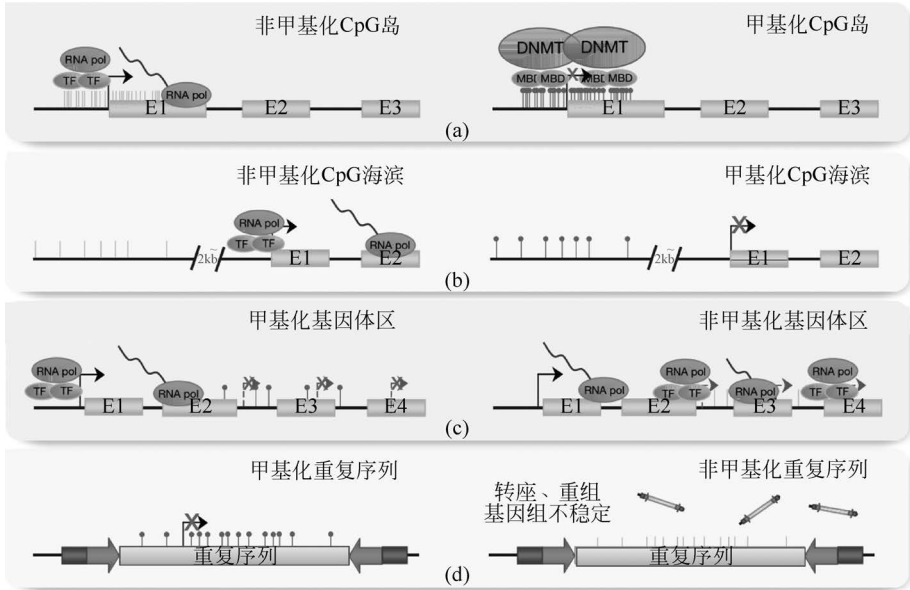


图 1.2 基因组 DNA 甲基化模式 (Portela et al., 2010)

基因组不同的区域都有 DNA 甲基化, 甲基化模式的改变会导致疾病, 左栏和右栏分别是正常状态下和失调状态下的模式。(a) 基因启动子上的 CpG 岛在正常状态下是非甲基化的, 能够允许转录, 异常的超甲基化导致转录失活; (b) 在 CpG 海滨处的甲基化模式与 CpG 岛是类似的; (c) 当甲基化发生在基因体区时, 促进了转录, 阻止了可疑的转录起始; 在疾病中, 基因体区被去甲基化, 使得转录在不正确的位置起始; (d) 重复序列一般是高甲基化的, 阻止了内寄生序列的激活; 在疾病中这个模式被改变, 造成染色质不稳定、转位和基因断裂。TF: 转录因子; RNA pol: RNA 聚合酶; DNMT: DNA 甲基转移酶; MBD: 具有甲基化 CpG 结合结构域的蛋白质; E1, E2, E3, E4: 外显子

CpG 贫乏启动子在很多特性上与 CpG 岛启动子迥然不同 (Deaton et al., 2011), 例如, 在起始转录方面, CpG 贫乏启动子内包含更多的 TATA box, 转录起始位点周围有更多的转录因子结合模体 (Baek et al., 2007); 系统地鉴别哺乳动物转录本的 5' 端, 发现在 CpG 岛启动子处的转录倾向于从一个宽泛的区域起始, 而在 CpG 贫乏启动子处的起始形成一个尖锐的峰 (Carninci et al., 2006)。在染色质组织方面, CpG 贫乏启动子需要 SWI/SNF 染色质重塑复合物来激活 (Ramirez-Carrozzi et al., 2009), 具有 H3K36me2 组蛋白修饰标记 (Blackledge et al., 2010), 且与 H3K27me3 标记没有关联 (Mikkelsen et al., 2007; Mohn et al., 2008)。更细致的比较见文献 Vavouri et al. (2012)。

相对于 CpG 岛启动子, CpG 贫乏启动子的甲基化水平的波动是可观的(Nagae et al., 2011)。具有 CpG 贫乏启动子的基因如果是在原生生殖细胞中表达的, 那么在其转录起始位点处是非甲基化状态; 而如果是只在胚胎干细胞或者组织特异性表达的基因, 通常在精子中甲基化, 但在卵母细胞或表达的体细胞中不甲基化(Farthing et al., 2008)。著名的例子包括编码 OCT4 和 NANOG 转录因子的基因, 它们是维持干细胞状态所必需的, 有研究表明, OCT4 基因和 NANOG 基因启动子可能经历了 AID 胞嘧啶核苷脱氨基酶(Bhutani et al., 2010; Popp et al., 2010)和/或 TET3 甲基胞嘧啶双加氧酶(Gu et al., 2011)的主动去甲基化。另外还有一类组织特异性的基因在精子和胚胎干细胞中甲基化, 而且只在这些基因有表达的特定组织中去甲基化(Han et al., 2011)。

关于 CpG 贫乏启动子的甲基化水平与所在基因的表达水平之间的相关性, 结论还不是特别确切。最开始, 一项全基因组范围的研究发现非 CpG 岛的甲基化与表达不存在反相关性(Weber et al., 2005); 后来另一项研究重新分析了数据, 认为基因表达与甲基化之间的负相关在整个基因组中都是非常明显的(Gal-Yam et al., 2008)。随后的研究表明, 对于 CpG 岛启动子, DNA 甲基化是基因失活的充分不必要条件, 而对于 CpG 贫乏启动子, DNA 甲基化不阻碍表达(Weber et al., 2007)。针对在 CpG 贫乏的启动子区的 DNA 甲基化丢失和转录激活的相关性, Bestor 等认为这只是反映了转录激活的结果而非原因(Bestor et al., 2015), 因为这与 Sp1 等转录因子结合之后周边的甲基化丢失的报道是一致的(Brandeis et al., 1994; Macleod et al., 1994; Matsuo et al., 1998; Mummaneni et al., 1998)。因此, 他们提出的假设是, CpG 岛启动子在所有表达状态下都是非甲基化的; 而 CpG 贫乏的启动子在漫长的非激活状态下漂变(drift)为部分甲基化状态, 而在转录激活的时候去除了甲基化, 这便是转录诱导的去甲基化假设(Bestor et al., 2015)。

CpG 岛之外的基因体区甲基化的功能是沉默逆转录病毒、LINE1 元件、Alu 元件以及其他 DNA 重复元件的主要机制(Yoder et al., 1997)。甲基化阻碍了这些元件的转录起始, 同时允许宿主基因的转录通过 CpG 岛之外的基因体区甲基化, 如图 1.2(d)所示。另外, 重复元件中的 CpG 位点深度甲基化能够阻止造成染色质不稳定、转位和基因断裂的内部寄生序列的重新激活, 从而维护染色体稳定性(Esteller, 2007a)。最后, 基因体区的甲基化对于引发癌症的体细胞和生殖细胞系突变也有贡献(Rideout et al., 1990)。

1.3.2 DNA 甲基化的设定、维持和去除

1.3.2.1 DNA 甲基化的设定

DNA 甲基化是由 DNA 甲基转移酶催化介导的、将甲基基团从 S-腺苷蛋氨酸转移到 DNA 的过程。DNA 甲基转移酶的全称是 DNA methyltransferase, 简称 DNMT(DNA MTase)。哺乳动物中 DNA 甲基转移酶家族有 5 个成员, 分别是 DNMT1、DNMT2、DNMT3A、DNMT3B 和 DNMT3L, 但是只有 DNMT1、DNMT3A 和 DNMT3B 有甲基转移酶的活性。其中 DNMT3A 和 DNMT3B 能够从头甲基化 DNA, 而 DNMT1 的作用是维持甲基化(图 1.3)。

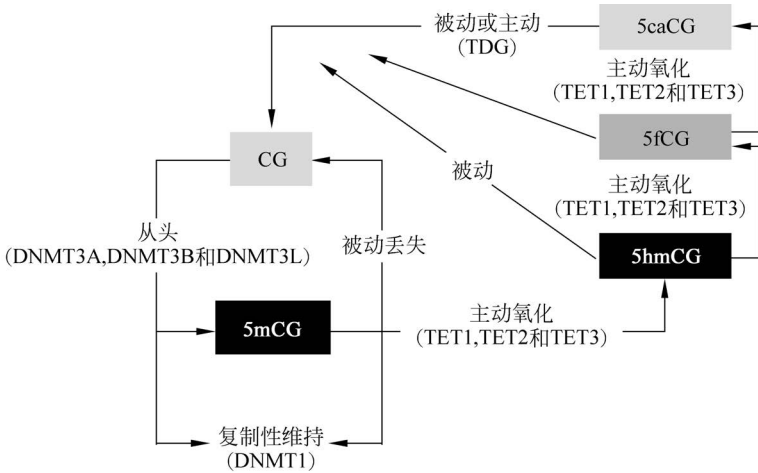


图 1.3 胞嘧啶甲基化的设定、维持和去除机制(Schubeler, 2015)

CpG 二核苷酸不同的甲基化状态以及设定、维持和去除标记的生物学通路。DNMT: DNA 甲基转移酶; TDG: 胸腺嘧啶 DNA 糖基化酶; TET: 甲基胞嘧啶双加氧酶; CG: CpG 二核苷酸; 5mCG: 5-甲基 CG; 5hmCG: 5-羟甲基 CG; 5fCG: 5-甲酰 CG; 5caCG: 5-羧基 CG

在胚胎干细胞中, DNMT3A 和 DNMT3B 高表达, 而在分化细胞中下调, 因此被认为是在胚胎发育过程中负责建立甲基化的模式(Esteller, 2007b), DNMT3A 对于造血干细胞的分化也是必需的(Challen et al., 2012)。DNMT3L 是 DNMT3A 和 DNMT3B 的无催化活性的同源物, 也被证明在建立母体基因组印记中起作用(Bourc'his et al., 2001)。基因组印记发生在配子形成过程中, 作为 DNMT3A 和 DNMT3B 通用的刺激因子,

DNMT3L 在这时表达,并与它们在细胞核中互作、共定位(Chen et al., 2005; Holz-Schietinger et al., 2010)。另外, DNMT2 包含了具有 DNA 甲基转移酶家族特征的所有催化模体,以往认为它几乎没有任何 DNMT 活性,后来也有报道称它能甲基化 tRNA^{Asp} (Goll et al., 2006)。

需要注意的是,虽然 DNMT3A 和 DNMT3B 诱导了从头甲基化,但是基因沉默的机制是非常复杂的,有很多层次、需要多个酶参与, DNA 甲基化是否在其中是最早期的事件,仍然有争议。Ooi 等的工作显示,表达 DNMT3L 的细胞通过两个 DNMT3A 和两个 DNMT3L 分子的四聚复合物获得了从头甲基化的能力,在这个过程中还需要一个核小体。因为活跃的转录起始位点缺失核小体,缺乏从头甲基化的这种底物,因此 DNA 甲基化不能用来起始基因沉默机制(Ooi et al., 2007)。这个发现被另一项研究所验证:使用维甲酸(retinoid acid)来诱导分化的胚胎癌细胞后, *OCT4* 基因的远端增强子和 *NANOG* 基因的启动子处首先出现了一个核小体,接着 DNMT3A 被招募到核小体,随后才是从头甲基化以及 *OCT4* 基因被沉默(You et al., 2011)。尽管在不表达 DNMT3L 的细胞中是否有相似的事件发生还不清楚,这些结果说明至少在一部分细胞中, DNA 甲基化不是基因沉默的初始事件。

1.3.2.2 DNA 甲基化的维持

维持性 DNA 甲基转移酶 DNMT1 是细胞中最丰富的 DNA 甲基转移酶,主要是在细胞周期的 S 期转录;它也有从头甲基化的活性,但是相对而言,对半甲基化状态的 DNA 有 30~40 倍的倾向性,因此可以甲基化半保守 DNA 复制过程中产生的半甲基化位点。在 S 期, DNA 聚合酶处理因子 PCNA 蛋白质(proliferating cell nuclear antigen)可以与 DNMT1 互作,增加其对新合成的 DNA 的亲合性,确保其在复制叉的定位(Chuang et al., 1997)。UHRF1 蛋白质(ubiquitin-like plant homeodomain and RING finger domain-containing protein 1)也可以行使类似的功能,通过有很强半 CpG 结合倾向性的 SET 结构域和 RING 结构域将 DNMT1 拉到半甲基化 DNA 处(Bostick et al., 2007)。

然而,从头甲基化与维持甲基化的工作划分并不总是非常清晰,仅凭 DNMT1 本身并不能够维持在已建立的 DNA 甲基化模式中,也需要 DNMT3A 和 DNMT3B 的持续参与(Jones et al., 2009),它们能够甲基化那些在复制叉处被 DNMT1 漏掉的位点(Jones et al., 2009)。这 3 个 DNA 甲