

游离 DNA 甲基化测序文库构建方法及应用

技术领域

本申请涉及生物技术，并且更特别地涉及游离 DNA 的表观遗传学分析方法。

背景技术

细胞外游离 DNA (cell-free DNA, cfDNA) 一般而言是在细胞新陈代谢、胞吐或凋亡 (apoptotic) 和坏死 (necrotic) 期间，从细胞内部释放到体液或环境中的 DNA 片段。血浆、血清、尿液、唾液、羊水等材料中含有微量的游离 DNA，反映了体内遗传变异、疾病、衰老等信息。对游离 DNA 进行提取、文库构建、测序和分析作为一种有效的无创/微创检测手段广泛应用于临床诊断和科学研究的各个方面，如癌症筛查及诊断、产前诊断、植入前诊断、衰老研究等领域。游离 DNA 携带的表观遗传信息如 DNA 甲基化/羟甲基化组学特征、片段组学特征越来越被认为是侦测、诊断及/或监测癌症等各种疾病的有价值的分子标记。

DNA 甲基化 (DNA methylation) 是生物体内在 DNA 上发生甲基化化学修饰的现象，其化学本质是：在胞嘧啶碱基的嘧啶环第 5 位碳原子上，原有的一个氢原子被替换成甲基，形成 5-甲基胞嘧啶 (5-methylcytosine)。这种 DNA 修饰在不改变 DNA 序列的前提下，影响了基因表达与调控、细胞分裂与分化、细胞的代谢与生长等诸多生命过程，最终影响着生物个体的生长发育、代谢繁衍、疾病发生发展和衰老死亡，是一种表观遗传机制。如果需要对 DNA 序列上甲基化修饰进行高通量测序 (Next Generation Sequencing, NGS)，必须先特异性地将具有/不具有甲基化的胞嘧啶转化为其他碱基，才能区分甲基化修饰。

片段组学信息是对游离 DNA 的片段化特征进行分析得到的信息集合，例如游离 DNA 在基因组上的分布特征、片段长度与起始终止位置、片段末端碱基特征、片段不平齐特征等信息。研究表明，游离 DNA 倾向于在某些基因组区域或元件处被切割，所以片段组学信息可以反映染色质的开放和封闭、核小体占位、转录因子结合、表观遗传学修饰等状态，是关键的表现基因组学信息之一，也是监测和诊断癌症等各种疾病的有价值的分子标记[1]。如果要在高通量测序时获取片段组学信息，必须保留 DNA 片段的完整性。

然而，由于游离 DNA 是在细胞内/外受到强烈的 DNA 酶降解作用形成的，具有以下几个特性导致对游离 DNA 检测困难：

1. cfDNA 含量少 (例如每毫升血浆仅含 cfDNA 几纳克)。这要求检测方法必须适应 1ng 甚至更低的 cfDNA 起始量，需要具有更高建库效率来提高 cfDNA 检测灵敏度。

2. cfDNA 片段短 (100bp 到 200bp 左右)。首先，DNA 短片段在建库过程中极易因纯化操作而丢失，更无法承受化学转化导致的二次断裂；其次，短片段无法应用转座酶、随

机引物、打断连接等全基因组建库方法，应用靶向 PCR 等靶向扩增建库方法也有一定的困难和局限。

3. cfDNA 往往化学结构损伤严重。cfDNA 的损伤来自于多个过程：细胞生理、病理状态产生的 DNA 损伤；细胞死亡过程中由核酸酶消化产生的损伤；在体液环境中(如血液循环)持续被游离核酸酶等活性物质攻击产生的损伤；在血浆等溶液中保存时间过长、反复冻融等过程产生的损伤。已有研究发现，97~98% 的 cfDNA 都具有缺口，是发生在 DNA 双链上的单链磷酸二酯键断裂缺口[2]。使用单链建库方法，cfDNA 会在双链变性解离后在缺口处断裂，产生碎片[3]。

现有的 DNA 甲基化测序文库构建方法可分为以下四种：（1）基于重亚硫酸盐转化的双链建库方法[4]；（2）基于重亚硫酸盐转化的单链建库方法(如 Swift Biosciences 试剂盒 Accel-NGS® Methyl-Seq DNA Library Kit, Cat. No. 30024)；（3）基于酶转化法的双链建库方法(如 New England Biolabs, Inc. EM-seq 技术 WO2017075436A1 及其商业化试剂盒 E7120; WO2019136413A1 的 TAPS 技术; WO2021077415 的 Cabernet 技术)；（4）基于酶转化法的单链建库方法(如 CN114032287A)。但是，以上四种方法都无法完全解决游离 DNA 甲基化检测的难点。

当前广泛使用的 DNA 甲基化检测方法——重亚硫酸盐转化法(Bisulfite conversion)——需要 DNA 在高盐(9M)、高酸(pH = 5)、高温(50~90℃)的严酷环境下反应数个小时，会使得 DNA 变性、断裂，绝大部分 DNA 会在转化反应过程中降解丢失。DNA 的大量断裂和损失，使得重亚硫酸盐转化法无法应对 cfDNA 低起始量建库要求，且加剧了 cfDNA 的短片段化和缺口损伤，所以很难基于重亚硫酸盐转化法研发高灵敏的游离 DNA 甲基化检测技术。

近年来基于酶学转化法的 DNA 甲基化测序技术逐渐发展成熟，例如 EM-seq 技术、TAPS 技术和 Cabernet 技术等。酶学转化法摒弃了严酷的化学处理过程，避免在转化时对 DNA 造成断裂和降解，建库效率更高、灵敏度更高、文库均一性更高。但是现有的酶学转化法专利和商品都无法全部解决游离 DNA 甲基化检测的全部问题。对于 cfDNA 起始量少的问题，EM-seq 技术(WO2017075436A1)和 TAPS 技术(WO2019136413A1)最低起始量都在 5~10 纳克，无法对游离 DNA 进行高灵敏度测序；Cabernet 技术(WO2021077415)可以进行低起始量 DNA 的建库，甚至可以对单细胞进行高灵敏的建库，但无法使用转座酶应用于短片段的游离 DNA，特别是对于 cfDNA 缺口问题，Cabernet 技术尚未提供解决方法。

cfDNA 存在的化学损伤问题，对于不检测甲基化的常规测序，使用双链建库，缺口会被末端修复过程抹平，不影响测序，未引起重视。在传统的重亚硫酸甲基化测序中也未被引起重视，是因为使用单链建库，会使 cfDNA 在缺口处断裂丢失，但甲基化测序数据本身不会有错误。基于新型酶转化法的常规方法是双链建库，确实观察到严重的甲基化抹除，产生严重的甲基化丢失现象。例如，Erger 等[5]应用双链建库方法构建的甲基化

测序文库数据显示, 平均甲基化水平从文库读长开始约 80%持续降低至 40%, 在后续 Read2 观察到普遍的甲基化水平抹除, 但作者未在文中讨论该异常。在竹石生物公开的 DNA 甲基化测序文库及其构建方法和检测方法 (CN114032287A) 中, 作者观察到了甲基化水平随文库读长降低的异常现象, 但将原因简单归因为末端单链缺失, 并使用单链建库方法来绕开问题, 并未真正解决问题, 单链状态下缺口处的断裂, 无法准确获得片段组学信息, 且单链连接效率不高, 建库产出低。因此, 迄今为止, 学界并未意识到并解决单链缺口的存在对 DNA 甲基化检测的影响。

综上所述, 目前还没有一种建库方法, 可以综合解决游离 DNA 甲基化检测面临的模板起始量低、片段短、缺口严重的问题, 无法获得完整的 DNA 文库, 以准确测量游离 DNA 甲基化和获取片段组学信息。

发明内容

本发明人在深入研究 cfDNA 特性和建库方式原理后, 针对 cfDNA 甲基化检测过程中存在的问题分析原因, 并进行一系列针对性的改进, 提供了新的 cfDNA 甲基化测序方法。

在一方面, 本文提供了一种对用于构建测序文库的生物样品进行前处理的方法, 其包括:

S1: 以 DNA 连接酶修补所述样品中包含的 DNA 片段中的单链缺口; 以及

S2: 任选地, 以 DNA 聚合酶将所述 DNA 片段上的末端缺口补齐,

其中所述测序文库用于在测序后获得所述生物样品的 DNA 甲基化信息数据或者获得所述生物样品的 DNA 甲基化信息和片段组学信息数据。

在一些实施方案中, 所述 DNA 片段的长度在 50 bp 至 1000 bp 之间; 优选地, 在 100 bp 至 500 bp 之间; 更优选地, 在 100 bp 至 350 bp 之间。

在一些实施方案中, 所述方法还包括在步骤 S1 之前片段化所述生物样品中包含的 DNA 分子, 以产生所述 DNA 片段。

在一些实施方案中, 所述连接酶具有 3', 5'-磷酸二酯键催化活性。

在一些实施方案中, 所述连接酶选自 HiFi-Taq 连接酶、T4 DNA 连接酶、Taq DNA 连接酶、*E.coli* DNA 连接酶和 T7 DNA 连接酶。

在一些实施方案中, 所述生物样品选自动物体液、细胞培养液或自然界生物样品, 包括但不限于外周血、血浆、血清、尿液、粪便、唾液、脑脊液、淋巴液、肺泡灌洗液、羊水、囊胚腔液、细胞培养液、胚胎培养液、微生物培养基、土壤浸出液和骨粉浸出液。

在一些实施方案中, 所述 DNA 片段包括但不限于细胞外游离 DNA、细胞核外游离 DNA、片段化 DNA, 例如来自癌症受试者的细胞外游离 DNA。

在一些实施方案中, 所述样品中的 DNA 片段含量不高于 10ng, 或不高于 1 ng, 例如 0.1 ng。

在一些实施方案中, 所述 DNA 甲基化包括但不限于胞嘧啶的甲基化和/或羟甲基化。

另一方面，本文提供了一种构建测序文库的方法，其包括：

1) 提供生物样品，并采用上述前处理方法对所述生物样品进行前处理，以获得不带有 3', 5'-磷酸二酯键缺口的 DNA 片段；

2) 利用所述不带有 3', 5'-磷酸二酯键缺口的 DNA 片段构建所述测序文库；

其中所述测序文库用于在测序后获得所述生物样品的 DNA 甲基化信息数据或者获得生物样品的 DNA 甲基化信息和片段组学信息数据。

在一些实施方案中，步骤 2)还包括：

a) 对所述 DNA 片段的两端进行末端补平处理及添加接头序列，优选地，所述接头序列中胞嘧啶部分或全部为甲基化胞嘧啶和/或羟甲基化胞嘧啶；以及

b) 通过酶转化法将所述 DNA 片段中的胞嘧啶转化为尿嘧啶，在后续扩增和测序中被识别为胸腺嘧啶，甲基化胞嘧啶则在后续扩增和测序中被识别为胞嘧啶。

在一些实施方案中，步骤 2)还包括：

a) 对所述 DNA 片段的两端进行末端补平处理及添加接头序列；

b) 通过酶转化法将所述末端 DNA 片段中的甲基化胞嘧啶转化为尿嘧啶或二氢尿嘧啶，所述尿嘧啶或二氢尿嘧啶在后续扩增和测序中被识别为胸腺嘧啶，未甲基化胞嘧啶在后续扩增和测序中被识别为胞嘧啶。

在一些实施方案中，所述酶转化法采用 EM-seq 转化法或 TAPS 转化法。

在一些实施方案中，所述方法还包括在步骤 a)后向所述生物样品中添加携带 DNA。

在一些实施方案中，所述携带 DNA 可以为任何不包括所述接头序列的 DNA，优选地，所述携带 DNA 的片段大小为 100-500 bp。

另一方面，本文提供了一种通过上述方法获得的测序文库。

另一方面，本文提供了一种确定生物样品中包含的 DNA 片段中单链缺口位置的方法，其包括：

1) 将所述生物样品分为两份：生物样品 A 和生物样品 B；

2) 使用前述前处理方法处理生物样品 A 以制备第一测序文库；使用前述前处理方法处理生物样品 B，条件是在所述前处理步骤中不使用 DNA 连接酶，以制备第二测序文库；

3) 基于所述第一测序文库和所述第二测序文库所获得的 DNA 甲基化信息差异来确定所述单链缺口位置。

另一方面，本文提供了一种在受试者中鉴定健康状况的方法，其包括：

1) 提供来自所述受试者的生物样品；

2) 采用前述前处理方法构建测序文库；以及

3) 对所述测序文库测序，或在杂交捕获富集后进行测序，

其中根据测序数据获得所述生物样品的 DNA 甲基化信息数据或所述生物样品的 DNA 甲基化信息数据和片段组学信息数据，并与群体中正常 DNA 甲基化信息数据和/或片段组学信息数据进行比较，以确定所述受试者的健康状况。

本文提供的上述方法，可用于细胞外游离 DNA 的表观遗传学分析。本文提供的方法还可用于基因组学、医学、诊断学和表观遗传学研究领域。

附图说明

图 1 为本文提供的甲基化文库构建方法的流程示意图。

图 2 为 EM-seq 转化法的流程示意图。

图 3 为 TAPS 转化法的流程示意图。

图 4 显示了使用 DNA Ligase 处理的方法(实线)与不进行处理的方法(虚线)对人外周血游离 DNA 进行甲基化建库的文库甲基化偏差分析结果。沿每条文库片段的 5'到 3'方向(横轴)，其 CpG 甲基化率的平均值(纵轴)会发生变化。

图 5 显示了游离 DNA 的片段分布规律，其中存在以约 170bp 为周期的峰。

图 6 显示了不同建库方法产生的文库片段分布。以相同文库 DNA 投入量进行毛细管凝胶电泳，分析 1~6000bp 范围内 DNA 片段分布分析。结果显示本发明方法(上图)的文库体现出周期性的 DNA 片段峰谱，符合外周血游离 DNA 的特点；而 Swift Biosciences, Accel-NGS® Methyl-Seq DNA Library Kit 试剂盒(下图)，则只有一个主峰、没有周期性峰，且含有大量的非文库杂峰。

具体实施方式

除非另有说明，本文使用的所有技术和科学术语具有本领域普通技术人员所通常理解的含义。

“DNA 连接酶”在本文中指用于修补双链 DNA 分子中的单链缺口(nick)的酶。DNA 连接酶通常具有 3', 5'-磷酸二酯键催化活性。在一个实例中，该 DNA 连接酶可催化缺口处的 5'磷酸基团与 3'羟基生成磷酸二酯键。DNA 连接酶通常不会在缺口处添加新的核苷酸。

“DNA 聚合酶”在本文中指，在引物存在时，能够以一条 DNA 链为模板在引物的 3'端添加 dNTP (脱氧核糖核苷酸)以合成其互补链的酶。DNA 聚合酶通常还可具有 3'-5'外切酶活性，在合成时起校对作用，以及 5'-3'外切酶活性，起切除修复作用。

“DNA 片段”在本文中指 DNA 短片段，例如长度在 50 bp 至 700 bp 之间，例如 100 bp 至 500 bp，尤其是 100 bp 至 350 bp。生物样品中包含的 DNA 片段通常是不均一的，即长度不等，相应地，上述长度可指其这些 DNA 片段的长度均值。这些 DNA 片段可能具有不同的序列，例如来自同一生物体基因组的不同区域，或者甚至来自不同的生物体。DNA 片段可以是带有单链切口的，也可以是平末端或非平齐末端(带有 3'突出或 5'突出)。

“平末端 DNA 片段”在本文中指末端不带有 3'突出或 5'突出的双链 DNA 片段。

“DNA 甲基化”在本文中指 DNA 分子或 DNA 片段中胞嘧啶碱基被修饰为 5-甲基胞嘧啶(5mC)。脊椎动物中的 DNA 甲基化一般发生在 CpG 位点(即 DNA 序列中胞嘧啶后紧连鸟嘌呤的位点)，经 DNA 甲基转移酶催化将胞嘧啶转化为 5-甲基胞嘧啶。人类基因组中

大多 CpG 位点已被甲基化，但是在某些特定区域，如富含胞嘧啶(C)和鸟嘌呤(G)的 CpG 岛则通常未被甲基化。CpG 甲基化可影响相关基因的转录活性，例如，甲基化可抑制抑癌基因，而去甲基化则刺激某些癌基因的表达，这些情况下都有可能导致癌症发生。另外，尽管发生率低于 5mC，也有少数胞嘧啶碱基被修饰为 5-羟甲基胞嘧啶(5hmC)、5-乙酰基化(5fC)、5-羧基化(5caC)。本文中提及甲基化时，也可以指修饰为 5hmC、5fC、5caC，除非上下文另有说明。

“DNA 甲基化信息”在本文中是指关于 DNA 分子或 DNA 片段中甲基化状况的信息，包括但不限于，甲基化位点、甲基化水平、甲基化方式(5mC 或 5hmC)等。本文中“甲基化水平”，也可称为“甲基化程度”，指样本中某一特定甲基化位点被甲基化修饰的比例(或频率)。可通过多种方式来检测一个位点是否被甲基化。常用方法包括化学或酶转化法，在转化过程中甲基化胞嘧啶和未甲基的胞嘧啶中的一种转化为尿嘧啶(U)或者在碱基配对方式上基本上等同于尿嘧啶的碱基(例如二氢尿嘧啶，DHU)。在随后的扩增过程中，对应的尿嘧啶作为胸腺嘧啶(T)与腺嘌呤(A)配对，最终结果是该甲基化位点的胞嘧啶或甲基化胞嘧啶在检测结果(如测序结果)中以胸腺嘧啶体现出来。通过与参考序列对比，就能够确定 DNA 分子或 DNA 片段中的胞嘧啶是否被甲基化。该参考序列可以为来自相同样品但未经上述转化的序列，或者健康群体中的对应序列。另外，如下文有描述的，还可通过一些方法来对 5mC 和 5hmC 进行区分。目前，已经将 DNA 甲基化信息广泛用于癌症(例如，肺癌、乳腺癌、肝癌、结直肠癌等)筛查和诊断中，尤其是早期筛查和诊断。

“片段组学信息”在本文中是指对样品中游离 DNA 的片段化特征进行分析得到的信息集合，例如游离 DNA 的在基因组上的分布特征、片段长度与起始终止位置、片段末端碱基特征、片段不平整特征等信息。近期已有一些文献报道将其用于癌症(如肺癌)筛查。

“酶转化法”在本文中是指借助于特定酶的催化作用对胞嘧啶或甲基化胞嘧啶进行修饰，以便在随后的检测中可区分其甲基化和未甲基化状态，或者区分 5mC 和 5hmC。本领域已知一些酶学转化方法，包括，例如但不限于，EM-seq 转化法和 TAPS 转化法。酶转化法中通常将这些酶与一些化学试剂配合使用。

“EM-seq 转化法”在本文中是指 New England Biolabs 公司开发的通过酶学转化将甲基化胞嘧啶和未甲基化胞嘧啶进行区分的技术。其利用双加氧酶 TET (包括 TET1、TET2、TET3 等)将 5mC 和 5hmC 修饰为 5-羧基胞嘧啶(5caC)，接着采用胞苷脱氨酶(cytidine deaminase)对未甲基化的胞嘧啶脱氨，转化为尿嘧啶。所用的胞苷脱氨酶包括 APOBEC 蛋白家族成员，例如 APOBEC 3A。在利用 DNA 糖基转移酶(GT)的情况下，可进一步区分 5mC 和 5hmC。EM-seq 转化法的流程可参见图 2。关于 EM-seq 转化法的更详细信息可参见 PCT 申请公开 WO2017075436A，在此通过引用将其全文并入本文。

“TAPS (TET-assisted pyridine borane sequencing)转化法”与 EM-seq 转化法有些类似，其也通过酶学方法将 5mC 和 5hmC 修饰为 5-羧基胞嘧啶(5caC)，但接着其不进行脱氨步骤，而是使用还原剂吡啶硼烷(pyridine borane)将 5caC 转变成二氢尿嘧啶(DHU)，同时未

甲基化的胞嘧啶保持不变。在后续的复制或扩增过程中，DHU 作为尿嘧啶被聚合酶识别，与 A 配对。这样，存在 5mC 和 5hmC 修饰的位点作为 T 被检测到，不存在修饰的 C 依然作为 C 被检测到。类似地，在使用 DNA 糖基转移酶(GT)或 KRuO_4 的情况下，可进一步区分 5mC 和 5hmC。TAPS 转化法的流程可参见图 3。关于 TAPS 转化法的更详细信息可参见 PCT 申请公开 WO2019136413A，在此通过引用将其全文并入本文。

“健康状况”在本文中指受试者的健康状态，包括是否患有某种疾病、患有某种疾病的风险大小、是否适合于某种治疗剂或治疗手段、疾病预后情况，等等。

“受试者”在本文中指动物，例如哺乳动物，包括(但不限于)人类、啮齿动物、猿猴、猫科动物、犬科动物、马科动物、牛科动物、猪科动物、绵羊、山羊、哺乳类实验动物、哺乳类农畜、哺乳类运动动物和哺乳类宠物。受试者可为雄性或雌性且可为任何适龄受试者，包括婴儿、幼年、青年、成年和老年受试者。在一些实例中，受试者为患者。在特定实例中，受试者为人类，诸如人类患者。该术语通常可与“患者”、“检测对象”、“治疗对象”等互换使用。

在本发明的一个方面，提供了对准备进行测序文库构建的样品进行前处理的方法。该前处理包括对样品中含有的 DNA 片段中的单链缺口进行修补，然后再进行常规的末端修复。修补缺口可采用 DNA 连接酶，末端修复可采用 DNA 聚合酶进行。之所以需要在常规末端修复前进行缺口修补，是因为发明人首次意识到并通过研究证实，这些单链缺口的存在是导致甲基化信息丢失的重要原因，特别是对于细胞外游离 DNA (cfDNA)，例如循环肿瘤 DNA (circulating tumor DNA, ctDNA)来说。仅进行常规修复步骤，所用的 DNA 聚合酶通过其 5'-3'外切酶活性和聚合酶活性进行了后缺口下游的链替换，从而抹去了原有链中包含的甲基化信息。在修补缺口后再进行 DNA 片段的末端修复，不会出现链替换过程，也就能够保留原来的甲基化信息。

在本发明的另一方面，提供了构建测序文库的方法，其包括在对样品进行上述前处理后，进行测序文库的构建。在一些实施方案中，在该文库构建过程中还包括利用酶转化法修饰其中的胞嘧啶(C)或甲基化胞嘧啶(包括 5mC 和/或 5hmC)，从而在后续的检测步骤能够将甲基化和未甲基化胞嘧啶进行区分或者将 5mC 和 5hmC 进行区分，形成 DNA 甲基化测序文库。将该测序文库上机测序后可获得相应的甲基化信息数据。领域技术人员可理解，采用本发明的方法，测序后不仅可获得甲基化信息数据，还可同时获得片段组学信息数据。这是由于本建库技术是基于 cfDNA 双链片段进行甲基化 NGS 检测文库构建，并且修补单链缺口后通过酶学转化区分甲基化状态，基本上不损害样品中原有的 DNA 片段的完整性，保留了其原有的片段组学信息，包括片段长度分布特征、片段长度等。

在本发明的另一方面，通过使用携带 DNA 可进一步降低本发明方法对样品中 DNA 片段含量的要求，直至 DNA 含量低至 0.1 ng，甚至更低。

在本发明的另一方面，通过添加或不添加 DNA 连接酶导致的甲基化信息差异来确定生物样品中包含的 DNA 片段中单链缺口位置。可理解的是，通常并非在缺口位置处正好

存在 5mC 或 5hmC, 因此, 在大多情况下, 该缺口位置可认为是缺口位置范围, 例如, 可认为缺口位置位于第一位置和其下游的第二位置之间: 第一位置为添加或不添加 DNA 连接酶均可检测到的甲基化位置, 第二位置为检测到甲基化差异的位置(例如添加 DNA 连接酶可检测到甲基化, 而不添加连接酶则检测不到甲基化)。

在本发明的另一方面, 可通过获得的甲基化信息来判断受试者的健康状况, 尤其是, 将甲基化信息与片段组学信息结合, 更准确地判断受试者健康状况。

可理解地, 本发明要解决的技术问题包括但不限于:

1. 提供一种游离 DNA 的高通量二代测序文库构建及 NGS 测序方法, 可以同时准确测定 DNA 甲基化和片段组学信息;
2. 方法需要有很高的灵敏度, 对低至 1ng 及 1ng 以下的低起始量样本都能进行高效的 NGS 文库构建;
3. 方法需要保留游离 DNA 链的完整性, 保证测序文库中的片段组学信息不被破坏;
4. 避免因为游离 DNA 上的缺口损伤导致的 DNA 断裂和甲基化信息丢失;
5. 技术路线简单改动后可以实现 DNA 羟甲基化等任意甲基化信息与片段组学信息的测定。

本发明提供的方法可用于但不限于细胞外游离 DNA 的表观遗传学分析, 能够实现上述目标。该方法可包括如下的一个或多个特征:

1. 使用特定 DNA 连接酶, 避免因 DNA 缺口损伤导致的甲基化丢失。对于含有缺口的 DNA, 本方法使用一种或几种特定的 DNA 连接酶, 如 Taq DNA ligase, 将模板 DNA 双链上的磷酸二酯键缺口进行修补, 防止后续过程中 DNA 链的断裂丢失和/或甲基化信息的抹除。经过缺口修复处理的 DNA 模板在经过应用酶学转化法的双链建库后, 文库 3'端不再产生严重的甲基化信息的抹除, 检测到的甲基化率恢复至平均甲基化率附近。

2. 使用携带 DNA 辅助对微量 DNA 进行保护, 实现对微量 DNA 进行有效文库构建和高灵敏的甲基化测序。对于本方法在酶学转化法的基础上进行改进, 参考专利公开 WO2021077415 的 Cabernet 技术, 在模板 DNA 添加文库接头后、酶学转化前, 加入物理化学性质相似但不带接头序列的携带 DNA (carrier DNA), 可以在遇到纯化等容易发生 DNA 损失或降解的步骤中, 携带 DNA 成为主要被丢失的部分, 有效避免模板 DNA 的损失。另外在转化结束后省去单链纯化步骤, 直接将 PCR 扩增反应体系加入, 以最大程度避免纯化过程造成的 DNA 模板的损失。

3. 重亚硫酸盐转化会导致片段组学信息丢失, 而本方法应用酶转化法和双链 DNA 建库策略, 在准确测定 DNA 甲基化信息的同时保留 DNA 片段组学信息。且在利用 DNA 连接酶处理时不影响片段组学信息的保留。

方法概述

(1) 游离 DNA 的提取

游离 DNA 是指游离于细胞外或细胞核外的 DNA，可以提取自多种体液、体外细胞培养液、自然环境等生物材料，包括但不限于：外周血、血浆、血清、尿液、粪便、唾液、脑脊液、淋巴液、肺泡灌洗液、羊水、囊胚腔液、细胞培养液、胚胎培养液、微生物培养基、土壤浸出液、骨粉浸出液等。游离 DNA 可以经过提取、纯化得到。

用于下游测序分析的游离 DNA 总量可以低至皮克、纳克级。

(2) DNA 的缺口修复

该步骤使用特定的连接酶(Ligase)，可以是但不限于任意一种或几种修复试剂和/或 DNA 连接酶的组合：HiFi-Taq ligase、T4 DNA Ligase、Taq DNA Ligase、T7 DNA Ligase、E.coli DNA ligase 等。

DNA 连接酶和/或 DNA 修复试剂将模板 DNA 双链上的缺口进行修补，防止后续过程中 DNA 链的断裂丢失和 / 或甲基化信息的抹除。

(3) DNA 两端加入文库接头序列(adaptor)

使用聚合酶等将双链末端缺口补齐，在一些实例中，在末端加入一个 A(腺嘌呤)碱基，方便接头连接。

该步骤使用 DNA 连接酶将带有部分/全部胞嘧啶甲基化或羟甲基化修饰的特定序列的 DNA 接头与模板 DNA 两端进行连接。

在一些实例中，DNA 接头序列上可带有生物素标记，方便后续纯化。

在一些实例中，DNA 接头序列上可以引入样品条码和单 DNA 分子条码，对 DNA 模板来源进行标记，增加检测准确性，也允许多个来源的样本在后续步骤混合。

(4) 加入 Carrier DNA

为了防止后续多个步骤中微量模板 DNA 的损失，在模板 DNA 溶液中加入数倍于模板 DNA 量的性质相似的携载 DNA (carrier DNA)，可以在遇到纯化等容易发生 DNA 损失或降解的步骤中成为主要被丢失的部分，有效避免模板 DNA 的损失。

由于 carrier DNA 不带有特定序列的 DNA 接头序列，不会在后续文库扩增和测序中被检测到。

(5) 甲基化转化

此步骤使用酶学转化法对 DNA 上的胞嘧啶、甲基化胞嘧啶和/或羟甲基化胞嘧啶进行特异性转化，使得胞嘧啶和甲基化和或羟甲基化胞嘧啶的碱基配对规则发生改变，在后续扩增和测序过程中对应不同的碱基，从而得以区分。

此步骤可以使用 New England Biolabs 公司发明的“EM-seq”技术。在另一个例子中可以使用 TAPS 技术。

(6) 文库扩增和测序

在甲基化转化时，DNA 经历热变性和碱基配对关系改变，为单链或部分单链状态。单链 DNA 在纯化过程中亲和力和弱，导致大量损失。本方法可省去单链纯化步骤，直接将 PCR 扩增反应体系加入，以最大程度避免模板 DNA 在纯化过程中的损失。

文库将进行 PCR 扩增反应，扩增引物含有与上述接头序列和测序仪适配序列互补的序列，还可以引入样品条码。引物可以与模板 DNA 两端所加入的接头序列互补配对，在扩增的同时形成期望的测序文库。

测序文库在纯化后进行定量和质检，交由测序仪进行测序。

(7) 测序结果分析

在采用 EM-seq 技术的甲基化转化的实例中，胞嘧啶转化为胸腺嘧啶，甲基化胞嘧啶保留为胞嘧啶；而采用 TAPS 技术的甲基化转化实例中，甲基化胞嘧啶转化为胸腺嘧啶，胞嘧啶不变。通过与参考基因组的比对，可以得到 DNA 甲基化信息。由于本方法在甲基化转化之前、游离 DNA 保持双链状态时加入文库接头，所以 DNA 片段信息得以保留，通过与参考基因组的比对，可以得知游离 DNA 的原 DNA 片段长度、片段分布、上下游断点位置、末端碱基模式等片段组学信息。上述分析过程可参考图 1。

本发明方法的其他应用

游离 DNA 的 DNA 羟甲基化与片段组学测序

在哺乳动物体内，DNA 羟甲基化含量比 DNA 甲基化低，但具有不同的生物学意义。

在上述甲基化转化过程中，不区分 DNA 甲基化与羟甲基化，如果需要仅测量羟甲基化，可以对 5mC 和 5hmC 进行区分，例如只需要在转化过程中不加入 TET 氧化酶，保留其他酶学反应即可。

基于靶向 PCR 的富集方法

富集感兴趣的片段，可以节省测序成本、提高目的基因的检测灵敏度。

在上文方法概述部分步骤(6)中，使用模板 DNA 的两端接头互补的引物进行 PCR 将对全部模板 DNA 进行扩增；如果需要富集目的基因片段，则可以使用靶向 PCR 的方法进行扩增。设计含有与甲基化转化后靶向基因片段互补的序列的靶向引物，投入 PCR 反应体系以进行靶向基因序列的扩增。

靶向引物可以含有/不含有与测序仪兼容的部分或全部接头序列。

可以仅投入上游或下游引物，得到一端为靶向引物、一端为上文方法概述部分步骤(3)所加入接头序列的片段。

上游与下游引物也可以末端相连、构建在一条 DNA 上，形成锁式引物。

可以先投入默认接头序列引物，扩增后再投入靶向引物；也可以同时投入；也可以不投入默认接头序列引物，仅使用靶向引物。

基于杂交捕获的富集方法

在上文方法概述部分步骤(6)中，可以在扩增前或扩增后进行杂交捕获富集。需要设计含有与甲基化转化后靶向基因片段互补的序列的探针。

杂交捕获富集使用生物素等标记的单链 DNA/RNA 探针与文库混合，经过热变性和复性，与靶向序列互补的文库会与探针杂交，杂交后的探针可以使用带有链霉亲和素的磁珠等方法进行捕获，得到富集后的文库。

不进行缺口修复以得到缺口位置信息

如果不实施上文方法概述部分步骤(2)中的连接酶处理，配合使用合适的聚合酶进行修复，也可以配合使用带有修饰的核苷酸，缺口下游甲基化信息将被改变。通过分析甲基化被抹除/改变的序列，可以得到缺口位置。

本发明公开了可以用于带有单链缺口的 DNA 的高灵敏 DNA 甲基化测序技术。该技术首先针对 DNA 缺口损伤进行修复，确保了 DNA 甲基化信号不会丢失和误读，其次该技术使用双链 DNA 为模板进行建库，克服单链建库效率低下问题，之后使用大幅优化调整过的酶学转化方法对甲基化信息进行碱基配对规则的转化，配合携带 DNA 在反应和纯化过程中对模板 DNA 的保护，最终以极低起始量的 DNA 为模板得到高质量、高灵敏的甲基化信息测序文库，同时保留 DNA 的片段组学信息。

本发明方法的优点包括但不限于：

1. DNA 最低起始量：低至 0.1ng 甚至更低；
2. 文库产量：可提高 4 倍；
3. 文库比对率：从 40%提高至 70%；
4. 文库 3'端连片 DNA 甲基化丢失长度：从 170bp 降至 40bp 以内。

本发明方法可应用在癌症早筛、癌症基因检测、产前诊断、植入前诊断、遗传诊断等多种液体活检诊断产品中，也为相关领域的科学研究提供了有力的技术支持。

以下通过具体实施例来进一步说明本发明。

实施例 1：

本实施例以外周血 cfDNA 为例描述测序文库制备过程。

1.1 提取 cfDNA

cfDNA 的提取可以采用现有技术中的任何标准手段进行。

用 5mL EDTA 静脉穿刺真空采血管(紫盖)采集外周血 5mL，在 4 小时内进行下述血浆分离：室温 1600xg 10min 水平转子低速离心。小心取出上层血浆，转移至 1.5mL 管中。用高速离心机 4℃ 6000xg 离心 10min，转移上清到新的 1.5mL 管中，保存在-80℃冰箱或干冰上。采用核酸提取或纯化试剂盒(QIAGEN, QIAamp® Circulating Nucleic Acid, 货号 55114)，从分离的血浆中提取得到 cfDNA，通常 1ml 血浆可以得到 5~20ng cfDNA。

1.2 cfDNA 缺口修复

配置 Ligase 反应体系，此处以 Taq DNA Ligase 为例，在纯化后的 cfDNA 溶液中加入下表 1 试剂。其中 Sonicated Spike-in DNA 为全基因组 CpG 甲基化 pUC19 DNA(NEB E7122)及全基因组 CpG 非甲基化的 lambda DNA(NEB E7123)等体积混合，使用超声波打断至 300bp 后稀释至 0.2ng/μL 备用。

表 1 连接酶预混液组成

Ligase 预混液	体积 μL
H ₂ O	6.3
Ligase 10x reaction buffer	2.5
Sonicated Spike-in DNA	0.2
ligase	1
	10

短暂混匀并离心后，在 PCR 仪上进行 37℃~60℃ 10~30min 反应。

1.3 末端修复和接头连接

配置 End Prep 反应体系，在上一步的修复产物中，按顺序加入试剂盒 NEBNext Ultra II End Repair/dA-Tailing Module(NEB E7546L)的试剂(表 2)：

表 2 末端修复试剂

试剂名称	体积/ μL
End Prep reaction buffer	3.5
End Prep Enzyme	1.5
总体积 μL	5

短暂混匀并离心后，在 PCR 仪上进行：20℃ 30min，65℃ 30min 反应。

反应结束后加入全胞嘧啶甲基化修饰的接头序列。为与后续建库引物互补，一对建议的接头序列为：5'- A [5mC] A [5mC] T [5mC] T T T [5mC] [5mC] [5mC] T A [5mC] A [5mC] G A [5mC] G [5mC] T [5mC] T T [5mC] [5mC] G A T [5mC] T-3' (SEQ ID NO: 1)和 5'-磷酸基-G A T [5mC] G G A A G A G [5mC] A [5mC] A [5mC] G T [5mC] T G A A [5mC] T [5mC] [5mC] A G T [5mC] A-3' (SEQ ID NO: 2)。引物稀释至总浓度 15 μM ，加入 1.25 μL 并混匀。

之后按顺序加入试剂盒 NEBNext Ultra II Ligation Module(NEB E7595L)的以下试剂(表 3)：

表 3 添加接头序列试剂

试剂名称	体积/ μL
Ligation enhancer	0.5
Ultra II Ligation	15
总体积 μL	15.5

短暂混匀并离心后，在 PCR 仪上进行：20℃ 反应 15min。

1.4 加入 Carrier DNA 并纯化

上一步的连接产物中加入 Carrier DNA 1 μL ，短暂混匀并离心。

Carrier DNA 的制备方法为：将 lambda DNA(NEB N3011)使用超声波打断至 300bp 后稀释至 25ng/ μL 备用。

使用 1.8 倍体积的 SPRI 或 Ampure XP 磁珠进行 DNA 纯化，使用水进行洗脱得到纯化后的 DNA。

1.5 DNA 甲基化修饰的酶学转化

将上一步得到的纯化 DNA 使用 NEBNext® Enzymatic Methyl-seq Conversion Module (NEB, E7125L)试剂盒进行酶学转化。包括 DNA 的 TET 氧化反应与糖基化保护、DNA 热变性、APOBEC 脱氨基反应。在 APOBEC 脱氨基反应后, 不进行纯化, 直接进行下一步反应。

1.6 文库 PCR 扩增

将上一步 APOBEC 脱氨基反应结束后的液体加入带有文库接头和 Index 标签的 PCR 引物, 如加入 5μL 的 NEBNext® Multiplex Oligos for Illumina 一对引物。再加入与已有液体等体积的 2x Q5U master mix (NEB M0597L)。短暂混匀并离心后在 PCR 仪上进行:

98℃ 30s; 7 个循环的 98℃ 10s、62℃ 30s、72℃ 90s; 65℃ 5min, 保存于 4℃。

使用 1.1 倍体积的 SPRI 或 Ampure XP 磁珠进行 DNA 纯化, 使用 1x TE buffer 进行洗脱得到纯化后的 DNA, 可保存于-20℃。

1.7 文库测序与分析

建好的文库经浓度定量和片段分布质检后, 在 illumina 测序仪上进行双端测序。测序下机数据使用 Bismark 软件进行比对和分析(为规避甲基化率分布不均的碱基, Read1 的前~10 个碱基和 Read2 的前~40 个碱基的甲基化率将不计入甲基化分析结果), 得到全基因组甲基化数据。经过生物信息学分析, 文库的在基因组上的分布特征、片段长度与起始终止位置、5'末端碱基分布特征等信息为片段组学信息。

表 4 测序结果以如下一条 reads 为例:

Read 编号	A00682:655:HKFCVDSX2:2:1205:32524:18490_1:N:0:CTC AGAA+AACTTGC
测序得到的序列为:	TTTTAAAAATATTTAAGTAAAGTAGAGAATGTAGAA ATGTTATAGATTATATTTTTTGATTATGATATAATAA AATTAGAAATTATAGTATGGAAATTTAAAAGTTTTTT TTTAAATAATTTTATGTTAAAAAGAAATTTAGGTCGG GT (SEQ ID NO: 3) 以及 ATGTTTGTAATTTTAGTATTTTGGGAGGTTAAGGTGG GTAGGTTATTTGAGGTTAGTAGTTTAAGATTAGTTTC GTTAATATGGTAATATTTTGTTTTTATTAATAAATA AAAATTAGTTGGGTTTGGTGGTTTATGTTTGTAATTT TA (SEQ ID NO: 4)
甲基化特征为:	hhh.....h..h.....h.....h.x.....h.x.....h.h..h..... x..h.....h..h.....h.....h.....hx...xZ.... 和 ...hx.....hx....h.....hh.....x...h.h.....x.x....h....hx..hh.Z..h.. h.....h.hhx...h.h..h.....h.....x....hx.....h.h...hx.....hhx 其中“.”代表非胞嘧啶碱基, z 代表 CpG 位点的胞嘧啶, h 和 x 代表非 CpG 位点的胞嘧啶, 小写字母代表该胞嘧啶 未被甲基化, 大写字母代表被甲基化, 本 read 中所有 CpG 位点均检测到甲基化。
比对到基因组上的起始位置 为:	第 1 号染色体上的 100915 位置(使用 hg38 参考基因组)。

比对到基因组上的终止位置为:	第 1 号染色体上的 101074 位置(使用 hg38 参考基因组)。
片段长度为:	159bp
5'末端碱基为:	CCC

测序结果包含数千万至数十亿条上述信息, 测序量随测序仪器和芯片选择有所不同。

实施例 2 DNA 缺口修复比较

为评估本方法可以弥补现有酶学转化建库方法的甲基化丢失缺陷, 进行平行对照比较测试:

从健康人血浆中提取游离 DNA 后, 分成两组进行实验:

实验组: 使用 DNA Ligase 处理(参照实施例 1 中步骤 1.2: DNA 缺口修复反应)

对照组: 不使用 DNA ligase 处理(作为对照组, 以水替代 DNA Ligase)

将两组游离 DNA 进行相同的双链建库步骤:

使用试剂 NEBNe.t Ultra II End Repair/dA-Tailing Module(NEB E7546L)进行末端修饰, 使用试剂 NEBNe.t Ultra II Ligation Module(NEB E7595L)与试剂 NEBNe.t® EM-seq™ Adaptor(NEB E7165)进行接头连接。

加入接头后的模板 DNA, 经过磁珠纯化, 使用 NEBNe.t® Enzymatic Methyl-seq Conversion Module(NEB, E7125L)试剂盒进行酶学转化处理。转化处理后的 DNA 使用 2. Q5U master mi.(NEB M0597L)进行 PCR 文库扩增。

构建完成的文库在 illumina 测序仪上进行双端 150bp 测序, 使用 Bismark 软件进行比对和分析, 并对文库甲基化偏差(M-bias)进行统计和分析, 沿文库片段 5'到 3'方向, 对每个读长位置的平均甲基化率进行画图, 评估文库甲基化水平是否平均。结果显示在图 4 中。

如图所示, 不对游离 DNA 进行缺口修复, 经过甲基化双链建库后, 测序文库中 DNA 平均甲基化水平从 Read1 的约 130bp 开始, 沿着文库片段的 5'到 3'方向一直延伸到 Read2, 产生越来越显著的甲基化丢失(图中虚线), 在 Read2 测序数据中尤为明显, 平均丢失了真实值的 40%甲基化率。实验组经过 DNA ligase 修复处理(图中实线), 修复了 Read2 上的甲基化丢失(文库 3'最末端 30bp 甲基化率下降是由于游离 DNA 双链模板 3'末端本身具有缺失产生的)。

该数据说明, 本文提供的方法可以弥补因游离 DNA 缺口损伤导致的甲基化丢失。

实施例 3 在获得甲基化信息的同时可获得片段组学信息

为评估本方法相对现有方法的优越性, 对照测试本发明技术及商业化的试剂盒。

实验组: 本发明技术方法, 具体操作详见实施例 1;

对比组: Swift Biosciences (试剂盒厂家货号 33024 Accel-NGS® Methyl-Seq DNA Library Kit, 具体操作详见试剂盒操作手册。

从健康人血浆中提取游离 DNA 后，以相同游离 DNA 起始量(3ng)、相同 PCR 扩增轮数(7 轮)进行两种不同方法的建库测试。对文库进行测序和分析后，使用 Bismark 软件对比对效率进行评估。结果如表 5 所示。

表 5 本发明技术方法与 Swift 方法建库比较

	本发明技术方法	Swift 方法
文库产量	83.4ng	21.15ng
文库比对率	70%	44%
文库重复率	15%	22%
片段分布	有核小体周期峰	无核小体周期峰

具体地，对建库后的 DNA 产出量进行 Qubit 定量，与 Swift 试剂盒相比，本发明技术方法的文库产量为 83.4ng，远高于 Swift 方法构建的文库产量；本发明技术方法测序文库的比对率(align rate)为 70%，比 swift 文库的 44%高出近一倍，说明文库有效性好；本方法构建的测序文库重复率(duplication rate)为 15%，显著低于 swift 文库的 22%的重复率，说明文库中重复片段更少、质量更好；本方法保留了片段组学信息(见图 5 和图 6)。使用安捷伦脉冲场自动化电泳片段分析系统(Agilent Femto Pulse)，对建库前的 DNA 片段进行分析，结果如图 5 所示，游离 DNA 本身具有以约 170bp 为周期的特征峰谱线，为典型的核小体占位导致的片段化特征；使用安捷伦片段分析系统(Agilent 5200 Fragment Analyzer System)，对建库后的片段进行分析，结果显示(见图 6)，与 Swift 试剂盒构建的文库相比，本发明提供的建库测序方法构建的测序文库含有游离 DNA 特征的片段化分布，可以更好地分析得到片段组学数据。

本发明首次指出，双链建库产生的严重的甲基化缺失问题是由于单链缺口和末端单链缺失共同产生的，末端单链缺失属于原始信息丢失，无法挽回；而单链缺口造成了大部分读长的甲基化被抹除。其原理为：双链建库步骤中，需要先对双链 DNA 模板进行末端修复，形成平末端或带有 3'-A 粘末端的 DNA 双链，然后再连接接头序列。而末端修复所使用的 DNA 合成酶具有自缺口开始合成的活性(Extension from Nick)，会从缺口位置开始、沿着 DNA 的 5'到 3'方向将下游原有 DNA 模板链置换或酶解，以人为提供的无修饰碱基来合成替代链，导致连片的 DNA 甲基化信息抹除缺失，并且这种甲基化擦除是无法在后期通过生物信息学等手段复原，因为缺口位置的随机性，无法使用一致的标准准确判断有效甲基化信息部分。本文提供的前处理方法在建库前对缺口进行修复，解决了大范围的甲基化丢失的问题(图 4)。

参考文献：

1. Lo, Y. M. D., Han, D. S. C., Jiang, P. & Chiu, R. W. K. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* 372, (2021).
2. Sanchez, C., Roch, B., Mazard, T., Blache, P., Dache, Z., Pastor, B., Pisareva, E., Tanos, R., & Thierry, A. R. (2021). Circulating nuclear DNA structural features, origins, and complete

size profile revealed by fragmentomics. JCI insight, 6(7), e144561. <https://doi.org/10.1172/jci.insight.144561>.

3. (Sanchez, C., Snyder, M. W., Tanos, R., Shendure, J., & Thierry, A. R. (2018). New insights into structural features and optimal detection of circulating tumor DNA determined by single-strand DNA analysis. NPJ genomic medicine, 3, 31. <https://doi.org/10.1038/s41525-018-0069-0>).

4. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008; 133:523–36.

5. Erger, F., Nörling, D., Borchert, D., Leenen, E., Habbig, S., Wiesener, M. S., Bartram, M. P., Wenzel, A., Becker, C., Toliat, M. R., Nürnberg, P., Beck, B. B., & Altmüller, J. (2020). cfNOMe — A single assay for comprehensive epigenetic analyses of cell-free DNA. Genome Medicine, 12(1), 54. <https://doi.org/10.1186/s13073-020-00750-5>.

权 利 要 求 书

1. 对用于构建测序文库的生物样品进行前处理的方法，包括：
S1：以 DNA 连接酶修补所述样品中包含的 DNA 片段中的单链缺口；以及
S2：以 DNA 聚合酶将所述 DNA 片段上的末端缺口补齐，
其中所述测序文库用于在测序后获得所述生物样品的 DNA 甲基化信息数据或者获得生物样品的 DNA 甲基化信息和片段组学信息数据；
其中步骤 S1 中所述单链缺口为 3', 5'-磷酸二酯键缺口。
2. 如权利要求 1 所述的方法，其中所述 DNA 片段的长度在 50 bp 至 1000 bp 之间。
3. 如权利要求 1 所述的方法，还包括在步骤 S1 之前片段化所述生物样品中包含的 DNA 分子，以产生所述 DNA 片段。
4. 如权利要求 1 所述的方法，其中所述连接酶具有 3', 5'-磷酸二酯键催化活性。
5. 如权利要求 1 或 4 所述的方法，其中所述连接酶选自 HiFi-Taq 连接酶、T4 DNA 连接酶、Taq DNA 连接酶、*E.coli* DNA 连接酶和 T7 DNA 连接酶。
6. 如权利要求 1 所述的方法，其中所述生物样品选自细胞培养液或自然界生物样品。
7. 如权利要求 1 所述的方法，其中所述生物样品选自外周血、血浆、血清、尿液、粪便、唾液、脑脊液、淋巴液、肺泡灌洗液、羊水、囊胚腔液、细胞培养液、胚胎培养液、微生物培养基、土壤浸出液和骨粉浸出液。
8. 如权利要求 7 所述的方法，其中所述 DNA 片段选自细胞外游离 DNA 和片段化 DNA。
9. 如权利要求 1 所述的方法，其中所述样品中的 DNA 片段含量不高于 10 ng。
10. 如权利要求 1 所述的方法，其中所述 DNA 甲基化包括胞嘧啶的甲基化和/或羟甲基化。
11. 构建测序文库的方法，包括：

1) 提供生物样品, 并采用权利要求 1-10 任一项所述的方法对所述生物样品进行前处理, 以获得不带有 3', 5'-磷酸二酯键缺口的 DNA 片段;

2) 利用所述 DNA 片段构建所述测序文库;

其中所述测序文库用于在测序后获得所述生物样品的 DNA 甲基化信息数据或者获得生物样品的 DNA 甲基化信息和片段组学信息数据。

12. 如权利要求 11 所述的方法, 其中步骤 2)还包括:

a) 对所述 DNA 片段的两端进行末端补平处理及添加接头序列, 其中所述接头序列中胞嘧啶部分或全部为甲基化胞嘧啶和/或羟甲基化胞嘧啶; 以及

b) 通过酶转化法将所述 DNA 片段中的胞嘧啶转化为尿嘧啶, 在后续扩增和测序中被识别为胸腺嘧啶, 甲基化胞嘧啶则在后续扩增和测序中被识别为胞嘧啶;

或者

a) 对所述 DNA 片段的两端进行末端补平处理及添加接头序列;

b) 通过酶转化法将所述 DNA 片段中的甲基化胞嘧啶转化为尿嘧啶或二氢尿嘧啶, 所述尿嘧啶或二氢尿嘧啶在后续扩增和测序中被识别为胸腺嘧啶, 未甲基化胞嘧啶在后续扩增和测序中被识别为胞嘧啶。

13. 如权利要求 12 所述的方法, 其中所述酶转化法采用 EM-seq 转化法或 TAPS 转化法。

14. 如权利要求 12 所述的方法, 其中还包括在步骤 a)后向所述生物样品中添加携带 DNA。

15. 如权利要求 14 所述的方法, 其中所述携带 DNA 不含有所述接头序列, 并且所述携带 DNA 的片段大小为 100-500 bp。

16. 确定生物样品中包含的 DNA 片段中单链缺口位置的方法, 包括:

1) 将所述生物样品分为两份: 生物样品 A 和生物样品 B;

2) 使用权利要求 1-10 任一项所述的方法处理生物样品 A 以制备第一测序文库; 使用权利要求 1-10 任一项所述的方法处理生物样品 B, 条件是步骤 S1 中不使用所述 DNA 连接酶, 以制备第二测序文库; 以及

3) 基于所述第一测序文库和所述第二测序文库所获得的 DNA 甲基化信息差异来确定所述单链缺口位置,

其中所述单链缺口为单链 3', 5'-磷酸二酯键缺口。