(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116875669 A
(43)申请公布日 2023.10.13

(21)申请号 202310683121.X
(22)申请日 2020.08.17
(30)优先权数据
62/887,987 2019.08.16 US
62/970,586 2020.02.05 US
62/991,891 2020.03.19 US
63/019,790 2020.05.04 US
63/051,210 2020.07.13 US
(62)分案原申请数据
202080004977.4 2020.08.17
(71)申请人 香港中文大学
地址 中国香港新界
(72)发明人 卢煜明 赵慧君 陈君赐 江培勇 郑淑恒 彭文磊 谢安仪
(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
专利代理师 王达佐 洪欣
(51)Int.Cl.
C12Q 1/6869(2018.01)
G16B 20/00(2019.01)
G16B 20/10(2019.01)
G16B 20/20(2019.01)
G16B 40/20(2019.01)
G16B 40/10(2019.01)
G16B 15/00(2019.01)
G16B 30/00(2019.01)

权利要求书4页 说明书96页 附图148页

(54)发明名称

测定核酸的碱基修饰

(57)摘要

本文描述了利用碱基修饰的测定来分析核酸分子和获取核酸分子分析数据的系统和方法。碱基修饰可以包含甲基化。用于测定碱基修饰的方法可以包含使用来源于测序的特征。这些特征可以包含来自测序碱基的光信号的脉冲宽度、碱基的脉冲间持续时间以及碱基的组成。可以对机器学习模型进行训练以使用这些特征来检测所述碱基修饰。单倍型之间的相对修饰或甲基化水平可以指示病症。修饰或甲基化状态还可以用于检测嵌合分子。

CN 116875669 A

1.用于检测核酸分子中的核苷酸的甲基化的方法,所述方法包括:

接收第一多个第一数据结构,所述第一多个第一数据结构中的每个第一数据结构对应于在多个第一核酸分子中的相应核酸分子中进行测序的核苷酸的相应窗口,其中所述相应窗口包括至少两个核苷酸,其中通过测量对应于所述核苷酸的光信号脉冲对所述第一核酸分子中的每一个进行测序,其中每个第一核酸分子的每个窗口中的靶位置处的核苷酸中的所述甲基化具有已知的第一状态,每个第一数据结构包括以下性质的值:

对于所述窗口内的每个核苷酸:

所述核苷酸的一致性,

所述核苷酸的相对于相应窗口内的所述靶位置的位置,

对应于所述核苷酸的所述脉冲的宽度,以及

脉冲间持续时间,其表示对应于所述核苷酸的所述脉冲与对应于相邻核苷酸的脉冲之间的时间,

存储多个第一训练样本,每个第一训练样本包含所述第一多个第一数据结构之一和指示所述靶位置处的所述核苷酸的所述甲基化的所述第一状态的第一标记,以及

当所述第一多个第一数据结构被输入到模型时,通过基于所述模型的与所述第一标记的对应标记匹配或不匹配的输出来优化所述模型的参数,使用所述多个第一训练样本来训练所述模型,其中所述模型的输出指明相应窗口中的所述靶位置处的所述核苷酸是否具有所述甲基化。

2.根据权利要求1所述的方法,其进一步包括:

接收第二多个第二数据结构,所述第二多个第二数据结构中的每个第二数据结构对应于在多个第二核酸分子中的相应核酸分子中进行测序的核苷酸的相应窗口,其中每个第二核酸分子的每个窗口内的靶位置处的核苷酸中的所述甲基化具有已知的第二状态,每个第二数据结构包括性质与所述第一多个第一数据结构相同的值;

存储多个第二训练样本,每个第二训练样本包含所述第二多个第二数据结构之一和指示所述靶位置处的所述核苷酸的所述第二状态的第二标记;

其中:

所述第一状态或所述第二状态是所述甲基化存在,而另一种状态是所述甲基化不存在,并且

训练所述模型进一步包括当所述第二多个第二数据结构被输入到所述模型时,通过基于所述模型的与所述第二标记的对应标记匹配或不匹配的输出优化所述模型的参数来使用所述多个第二训练样本。

3.根据权利要求2所述的方法,其中所述多个第一核酸分子与所述多个第二核酸分子相同。

4.根据权利要求2所述的方法,其中所述多个第一核酸分子是使用多重置换扩增、通过第一类型的甲基化核苷酸产生的,并且其中所述多个第二核酸分子是使用多重置换扩增、通过第一类型的非甲基化核苷酸产生的。

5.根据权利要求1所述的方法,其中所述光信号是来自染料标记的核苷酸的荧光信号。

6.根据权利要求1所述的方法,其中与所述第一多个第一数据结构相关联的每个窗口包括每个第一核酸分子的第一链上的4个连续核苷酸。

7.分析生物体的生物样本的方法,所述生物体在第一染色体区域中具有第一单倍型和第二单倍型,所述生物样本包含DNA分子,所述方法包括:

分析来自所述生物样本的多个DNA分子,其中分析DNA分子包含:

鉴定所述DNA分子在参考人类基因组中的位置;

测定所述DNA分子的相应等位基因;以及

确定所述DNA分子在一个或多个基因组位点处是否被甲基化;

鉴定所述第一染色体区域的第一部分的一个或多个杂合基因座,每个杂合基因座包含所述第一单倍型中的对应的第一等位基因和所述第二单倍型中的对应的第二等位基因;

鉴定所述多个DNA分子的第一集合,每个DNA分子:

定位于所述一个或多个杂合基因座中的任一个杂合基因座处,

包含所述杂合基因座的所述对应的第一等位基因,并且

包含N个基因组位点中的至少一个基因组位点,N是大于或等于一的整数;

使用所述多个DNA分子的所述第一集合测定所述第一单倍型的所述第一部分的第一甲基化水平;

鉴定所述多个DNA分子的第二集合,每个DNA分子:

定位于所述一个或多个杂合基因座中的任一个杂合基因座处,

包含所述杂合基因座的所述对应的第二等位基因,并且

包含所述N个基因组位点中的至少一个基因组位点;

使用所述多个DNA分子的所述第二集合测定所述第二单倍型的所述第一部分的第二甲基化水平;

使用所述第一甲基化水平和所述第二甲基化水平计算参数值;

将所述参数值与参考值进行比较;以及

使用所述参数值与所述参考值的比较来确定所述生物体中的病症的分类。

8.检测生物样本中的嵌合分子的方法,所述方法包括:

对于来自所述生物样本的多个DNA分子中的每个DNA分子:

对所述DNA分子执行单分子测序以获得序列读段,所述序列读段提供了N个位点中的每个位点处的甲基化状态,N为5或更大,其中所述序列读段的甲基化状态形成甲基化谱式;

使所述甲基化谱式滑过对应于嵌合分子的一个或多个参考谱式,所述嵌合分子具有来自参考人类基因组的两个组成部分的两个部分,所述一个或多个参考谱式包含甲基化状态与非甲基化状态之间的变化;以及

鉴定所述甲基化谱式与所述一个或多个参考谱式中的第一参考谱式之间的匹配位置,所述匹配位置鉴定所述序列读段中的所述参考人类基因组的所述两个组成部分之间的接合点;以及

输出所述接合点作为嵌合分子中的基因融合的位置。

9.用于确定样本核酸分子是胎儿来源还是母体来源的方法,其中所述样本核酸分子是细胞游离的并且从怀有胎儿的雌性受试者的生物样本获得,所述方法包括:

对于所述多个样本核酸分子中的每个样本核酸分子:

(a)接收通过测量对应于所述样本核酸分子中测序的核苷酸的光信号脉冲而获得的数据,并从所述数据获得以下性质的值:

对于每个核苷酸：

所述核苷酸的一致性，

所述核苷酸在所述样本核酸分子中的位置，

对应于所述核苷酸的所述脉冲的宽度，以及

脉冲间持续时间，其表示对应于所述核苷酸的所述脉冲与对应于相邻核苷酸的脉冲之间的时间；

(b)对于在所述样本核酸分子中测序的核苷酸的一个或多个核苷酸中的每一个，创建输入数据结构，所述输入数据结构包括在所述样本核酸分子中测序的核苷酸的窗口，其中所述窗口包括至少两个核苷酸，并且其中对于窗口内的每个核苷酸，所述输入数据结构包括以下性质：

所述核苷酸的一致性，

所述核苷酸相对于所述窗口内的靶位置的位置，

对应于所述核苷酸的所述脉冲的宽度，以及

脉冲间持续时间，

(c)对于所述样本核酸分子中测序的核苷酸的一个或多个核苷酸中的每一个，将输入数据结构输入到模型中，所述模型被配置为：

接收所述输入数据结构，所述输入数据结构使用通过测量对应于所述核苷酸的光信号脉冲而获得的数据来生成，

将所述模型的优化参数应用于所述输入数据结构的性质；和

输出甲基化状态，所述甲基化状态指定在相应窗口中的靶位置处的核苷酸是否具有甲基化；和

(d)对于所述样本核酸分子中测序的核苷酸中的一个或多个核苷酸中的每一个，使用所述模型确定当核苷酸在所述输入数据结构中的窗口内的靶位置时核苷酸中是否存在所述甲基化的甲基化状态；和

(e)使用所述一个或多个核苷酸的甲基化状态来确定所述样本核酸分子是胎儿来源的还是母本来源的。

10.确定在胎儿基因组的区域中是否存在拷贝数异常的方法，所述方法包括：

对于多个样本核酸分子中的每个样本核酸分子，其中所述多个样本核酸分子是细胞游离的并且从怀有胎儿的雌性受试者的生物样本获得：

(a)接收通过测量对应于所述样本核酸分子中测序的核苷酸的光信号脉冲而获得的数据，并从所述数据获得以下性质的值：

对于每个核苷酸：

所述核苷酸的一致性，

所述核苷酸在所述样本核酸分子中的位置，

对应于所述核苷酸的所述脉冲的宽度，以及

脉冲间持续时间，其表示对应于所述核苷酸的所述脉冲与对应于相邻核苷酸的脉冲之间的时间；

(b)对于在所述样本核酸分子中测序的核苷酸的一个或多个核苷酸中的每一个，创建输入数据结构，所述输入数据结构包括在所述样本核酸分子中测序的核苷酸的窗口，其中

所述窗口包括至少两个核苷酸，并且其中对于所述窗口内的每个核苷酸，所述输入数据结构包括以下性质：

所述核苷酸的一致性，

所述核苷酸相对于所述窗口内的靶位置的位置，

对应于所述核苷酸的所述脉冲的宽度，以及

脉冲间持续时间；

(c)对于所述样本核酸分子中测序的核苷酸的一个或多个核苷酸中的每一个，将输入数据结构输入到模型中，所述模型被配置为：

接收所述输入数据结构，所述输入数据结构使用通过测量对应于所述核苷酸的光信号脉冲而获得的数据来生成，

将所述模型的优化参数应用于所述输入数据结构的性质；和

输出甲基化状态，所述甲基化状态指定在相应窗口中的靶位置处的核苷酸是否具有甲基化；和

(d)对于所述样本核酸分子中测序的核苷酸中的一个或多个核苷酸中的每一个，使用所述模型确定当核苷酸在输入数据结构中的窗口内的靶位置时核苷酸中是否存在甲基化的甲基化状态；和

(e)将所述样本核酸分子鉴定为与所述区域对齐；

使用所述多个样本核酸分子中的每个样本核酸分子的所述一个或多个核苷酸的甲基化状态来确定所述区域的甲基化水平；和

使用所述甲基化水平确定拷贝数异常是否存在于胎儿基因组的所述区域中。

测定核酸的碱基修饰

[0001] 相关申请的交叉引用

[0002] 本申请是申请号为202080004977.4的中国专利申请的分案申请，并且要求以下临时申请的优先权益：2020年7月13日提交的名称为“测定核酸的碱基修饰(DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS)”的美国临时申请第63/051,210号；2020年5月4日提交的名称为“测定核酸的碱基修饰(DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS)”的美国临时申请号第63/019,790号；2020年3月19日提交的名称为“测定核酸的碱基修饰(DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS)”的美国临时申请第62/991,891号；2020年2月5日提交的名称为“测定核酸的碱基修饰(DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS)”的美国临时申请第62/970,586号；2019年8月16日提交的名称为“测定核酸的碱基修饰(DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS)”的美国临时申请第62/887,987号，所有这些美国临时申请的全部内容通过引用并入本文用于所有目的。

背景技术

[0003] 核酸中碱基修饰的存在因不同的生物体而异，所述生物体包含病毒、细菌、植物、真菌、线虫、昆虫和脊椎动物(例如人)等。最常见的碱基修饰是在不同位置处向不同DNA碱基添加甲基，即所谓的甲基化。已经在胞嘧啶、腺嘌呤、胸腺嘧啶和鸟嘌呤上找到甲基化，如5mC(5-甲基胞嘧啶)、4mC(N4-甲基胞嘧啶)、5hmC(5-羟甲基胞嘧啶)、5fC(5-甲酰基胞嘧啶)、5caC(5-羧基胞嘧啶)、1mA(N1-甲基腺嘌呤)、3mA(N3-甲基腺嘌呤)、7mA(N7-甲基腺嘌呤)、3mC(N3-甲基胞嘧啶)、2mG(N2-甲基鸟嘌呤)、6mG(O6-甲基鸟嘌呤)、7mG(N7-甲基鸟嘌呤)、3mT(N3-甲基胸腺嘧啶)和4mT(O4-甲基胸腺嘧啶)。在脊椎动物基因组中，5mC是最常见的碱基甲基化类型在其之后是鸟嘌呤(即形成CpG的形式)。

[0004] DNA甲基化对哺乳动物的发育至关重要，并且在基因表达和沉默、胚胎发育、转录、染色质结构、X染色体失活、防止重复元件的活性、在有丝分裂期间维持基因组稳定性以及调节亲源基因组印记方面具有显著作用。

[0005] DNA甲基化以协同的方式在启动子和增强子的沉默中发挥着许多重要的作用(Robertson,2005;Smith和Meissner,2013)。已经发现许多人类疾病与DNA甲基化的异常相关，包含但不限于致癌过程、印记基因失调性疾病(例如，贝克威思-威德曼综合征(Beckwith-Wiedemann syndrome)和普拉德-威利综合征(Prader-Willi syndrome))、重复元件不稳定相关的疾病(例如，脆性X综合征)、自身免疫性疾病(例如，全身性红斑狼疮)、代谢障碍(例如，I型和II型糖尿病)、神经障碍、衰老等。

[0006] 对DNA分子进行甲基化组修饰的精确测量将具有许多临床意义。一种广泛使用的测量DNA甲基化的方法是通过使用亚硫酸氢盐测序(BS-seq)(Lister等人，2009;Frommer等人，1992)。在此方法中，首先用亚硫酸氢盐处理DNA样本，非甲基化胞嘧啶(即C)转化为尿嘧啶，而甲基化的胞嘧啶保持不变。然后通过DNA测序分析经亚硫酸氢盐修饰的DNA。在另一种方法中，在亚硫酸氢盐转化之后，然后使用能够区分不同甲基化谱的亚硫酸氢盐转化的DNA

的引物对经过修饰的DNA进行聚合酶链式反应(PCR)扩增(Herman等人,1996)。后一种方法被称为甲基化特异性PCR。

[0007] 此类基于亚硫酸氢盐的方法的一个缺点是,已经报道了亚硫酸氢盐转化步骤会使大多数经过处理的DNA显著降解(Grunau,2001)。另一个缺点是亚硫酸氢盐转化步骤将产生强烈的CG偏差(Olova等人,2018年),从而导致通常具有非均质甲基化状态的DNA混合物的信噪比降低。此外,由于在亚硫酸氢盐处理期间DNA的降解,亚硫酸氢盐测序将不能对长DNA分子进行测序。因此,需要测定核酸碱基的修饰,而无需事先进行化学反应(例如,亚硫酸氢盐转化)和核酸扩增(例如,使用PCR)。

发明内容

[0008] 已经开发了一种新方法,在一个实施例中,所述新方法使得能够测定核酸中的碱基修饰(如5mC)而无需模板DNA预处理(如酶和/或化学转化,或蛋白质和/或抗体结合)。尽管此类模板DNA预处理对于所述碱基修饰的测定不是必需的,但在所示出的实例中,某些预处理(例如,用限制酶消化)可能有助于增强本发明的各方面(例如,允许富集CpG位点以供分析)。本公开中呈现的实施例可以用于检测不同类型的碱基修饰,例如,包含但不限于4mC、5hmC、5fC和5caC、1mA、3mA、7mA、3mC、2mG、6mG、7mG、3mT和4mT等。此类实施例可以利用-在测序过程中所产生的特征,例如受各种碱基修饰影响的动力学特征,以及测定甲基化状态的靶位置周围的窗口中的核苷酸的一致性。

[0009] 本发明的实施例可以用于但不限于单分子测序。一种类型的单分子测序是单分子实时测序,其中对单个DNA分子测序的进度进行实时监测。一种类型的单分子实时测序是由太平洋生物科学公司(Pacific Biosciences)使用其单分子实时(SMRT)系统进行商业化的测序。方法可以使用来自测序碱基的信号的脉冲宽度、碱基的脉冲间持续时间(IPD)和碱基的一致性,以便检测碱基或相邻碱基中的修饰。另一个单分子系统是基于纳米孔测序的系统。纳米孔测序系统的一个实例是由牛津纳米孔技术公司(Oxford Nanopore Technologies)进行商业化的系统。

[0010] 已经开发的方法可以作为检测生物样本中的碱基修饰以出于各种目的(包含但不限于研究和诊断目的)评估所述样本中的甲基化谱的工具。检测到的甲基化谱可以用于不同的分析。甲基化谱可以用于检测DNA的组织来源(例如,母亲或胎儿、器官、细菌或从癌症患者血液中富集的肿瘤细胞中获得的DNA)。对组织中异常甲基化谱的检测有助于鉴定个体的发育障碍,鉴定和预测肿瘤或恶性肿瘤。

[0011] 本发明的实施例可以包含分析生物体的单倍型的相对甲基化水平。两种单倍型之间甲基化水平的失衡可以用于确定病症的分类。较高的失衡可以指示存在病症或更严重的病症。所述病症可以包含癌症。

[0012] 单分子的甲基化谱式可以鉴定嵌合体和杂合DNA。嵌合分子和杂合分子可以包含来自两种不同基因、染色体、细胞器(例如,线粒体、细胞核、叶绿体)、生物体(哺乳动物、细菌、病毒等)和/或物种的序列。检测嵌合分子或杂合DNA分子的接合点可以允许检测各种病症或疾病(包含癌症、产前或先天性病症)的基因融合。

[0013] 参考以下详细描述和附图,可以更好地理解本发明的实施例的本质和优点。

附图说明

[0014] 图1展示了根据本发明的实施例的携带碱基修饰的分子的SMRT测序。

[0015] 图2展示了根据本发明的实施例的携带甲基化和非甲基化CpG位点的分子的SMRT测序。

[0016] 图3展示了根据本发明的实施例的脉冲间持续时间和脉冲宽度。

[0017] 图4示出了根据本发明的实施例的用于检测碱基修饰的DNA的沃森链(Watson strand)的测量窗口的实例。

[0018] 图5示出了根据本发明的实施例的用于检测碱基修饰的DNA的克里克链(Crick strand)的测量窗口的实例。

[0019] 图6示出了根据本发明的实施例的通过组合来自DNA的沃森链和其互补的克里克链的数据来检测任何碱基修饰的测量窗口的实例。

[0020] 图7示出了根据本发明的实施例的通过组合来自DNA的沃森链和其附近区域的克里克链的数据来检测任何碱基修饰的测量窗口的实例。

[0021] 图8示出了根据本发明的实施例的用于测定CpG位点处的甲基化状态的沃森链、克里克链和这两条链的测量窗口的实例。

[0022] 图9示出了根据本发明的实施例的构建用于对碱基修饰进行分类的分析、计算、数学或统计模型的一般程序。

[0023] 图10示出了根据本发明的实施例的对碱基修饰进行分类的一般程序。

[0024] 图11示出了根据本发明的实施例的使用具有已知的沃森链甲基化状态的样本构建用于对CpG位点处的甲基化状态进行分类的分析、计算、数学或统计模型的一般程序。

[0025] 图12示出了根据本发明的实施例的对未知样本的沃森链的甲基化状态进行分类的一般程序。

[0026] 图13示出了根据本发明的实施例的使用具有已知的克里克链甲基化状态的样本构建用于对CpG位点处的甲基化状态进行分类的分析、计算、数学或统计模型的一般程序。

[0027] 图14示出了根据本发明的实施例的对未知样本的克里克链的甲基化状态进行分类的一般程序。

[0028] 图15示出了根据本发明的实施例的使用来自沃森链和克里克链的具有已知甲基化状态的样本构建用于对CpG位点处的甲基化状态进行分类的统计模型的一般程序。

[0029] 图16示出了根据本发明的实施例的对来自沃森链和克里克链的未知样本的甲基化状态进行分类的一般程序。

[0030] 图17A和17B示出了根据本发明的实施例的用于测定甲基化的训练数据集和测试数据集的性能。

[0031] 图18示出了根据本发明的实施例的用于测定甲基化的训练数据集和测试数据集的性能。

[0032] 图19示出了根据本发明的实施例的用于测定甲基化的训练数据集和测试数据集在不同测序深度下的性能。

[0033] 图20示出了根据本发明的实施例的用于测定甲基化的不同链的训练数据集和测试数据集的性能。

[0034] 图21示出了根据本发明的实施例的用于测定甲基化的不同测量窗口的训练数据

集和测试数据集的性能。

[0035] 图22示出了根据本发明的实施例的仅使用下游碱基来测定甲基化的不同测量窗口的训练数据集和测试数据集的性能。

[0036] 图23示出了根据本发明的实施例的仅使用上游碱基来测定甲基化的不同测量窗口的训练数据集和测试数据集的性能。

[0037] 图24示出了根据本发明的实施例的使用与在训练数据集中使用不对称侧翼尺寸的下游和上游碱基相关联的动力学谱式的甲基化分析的性能。

[0038] 图25示出了根据本发明的实施例的使用与在测试数据集中使用不对称侧翼尺寸的下游和上游碱基相关联的动力学谱式的甲基化分析的性能。

[0039] 图26示出了根据本发明的实施例的关于CpG位点处的甲基化状态分类的特征的相对重要性。

[0040] 图27示出了根据本发明的实施例的在不使用脉冲宽度信号的情况下用于甲基化检测的基于基序的IPD分析的性能。

[0041] 图28是根据本发明的实施例的使用经受了甲基化分析的胞嘧啶上游的2-nt和下游的6-nt的主成分分析技术的图。

[0042] 图29是根据本发明的实施例的使用主成分分析的方法与使用卷积神经网络的方法之间的性能比较图。

[0043] 图30示出了根据本发明的实施例的仅使用上游碱基来测定甲基化的不同分析、计算、数学或统计模型的训练数据集和测试数据集的性能。

[0044] 图31A示出了根据本发明的实施例的通过全基因组扩增产生具有未甲基化腺嘌呤的分子的一种方法的实例。

[0045] 图31B示出了根据本发明的实施例的通过全基因组扩增产生具有甲基化腺嘌呤的分子的一种方法的实例。

[0046] 图32A和32B示出了根据本发明的实施例的未甲基化数据集与甲基化数据集之间的沃森链的模板DNA中经过测序的A碱基的脉冲间持续时间(IPD)值。

[0047] 图32C示出了根据本发明的实施例的用于测定沃森链的甲基化的ROC特性曲线。

[0048] 图33A和33B示出了根据本发明的实施例的未甲基化数据集与甲基化数据集之间的克里克链的模板DNA中经过测序的A碱基的脉冲间持续时间(IPD)值。

[0049] 图33C示出了根据本发明的实施例的用于测定克里克链的甲基化的ROC特性曲线。

[0050] 图34展示了根据本发明的实施例的沃森链的6mA测定。

[0051] 图35展示了根据本发明的实施例的克里克链的6mA测定。

[0052] 图36A和图36B示出了根据本发明的实施例的使用基于测量窗口的卷积神经网络模型对uA数据集与mA数据集之间的沃森链的经过测序的A碱基进行甲基化的测定概率。

[0053] 图37示出了根据本发明的实施例的使用基于测量窗口的CNN模型对沃森链的经过测序的A碱基进行6mA检测的ROC曲线。

[0054] 图38示出了根据本发明实施例的基于IPD度量的6mA检测与基于测量窗口的6mA检测之间的性能比较。

[0055] 图39A和39B示出了根据本发明实施例的使用基于测量窗口的CNN模型对uA数据集与mA数据集之间的克里克链的那些经过测序的A碱基进行甲基化的测定概率。

[0056] 图40示出了根据本发明的实施例的使用基于测量窗口的CNN模型对克里克链的经过测序的A碱基进行的6mA检测的性能。

[0057] 图41示出了根据本发明的实施例的包含沃森链和克里克链的分子中跨A碱基的甲基化状态的实例。

[0058] 图42示出了根据本发明的实施例的通过选择性地使用mA数据集中IPD值大于其第10百分位的A碱基来增强训练的实例。

[0059] 图43是根据本发明的实施例的mA数据集中未甲基化腺嘌呤的百分比对每个孔中子读段(subreads)的数量的图。

[0060] 图44示出了根据本发明的实施例的测试数据集中双链DNA分子的沃森链与克里克链之间的甲基腺嘌呤谱式。

[0061] 图45是表格,其示出了根据本发明的实施例的训练数据集和测试数据集中的完全未甲基化的分子、半甲基化分子、完全甲基化的分子和具有交错甲基腺嘌呤谱式的分子的百分比。

[0062] 图46展示了根据本发明的实施例的具有关于腺嘌呤位点的完全未甲基化分子的分子、半甲基化分子、完全甲基化的分子以及具有交错的甲基腺嘌呤谱式的分子的代表性实例。

[0063] 图47示出了根据本发明的实施例的含有CpG岛(如黄色阴影所示)的长读段(6,265bp)的实例。

[0064] 图48是表格,其示出了根据本发明的实施例的通过太平洋生物科学公司SMRT测序对9个DNA分子进行测序并且与印记区域重叠。

[0065] 图49示出了根据本发明的实施例的基因组印记的实例。

[0066] 图50示出了根据本发明的实施例的用于测定印记区域中的甲基化谱式的实例。

[0067] 图51示出了根据本发明的实施例的推导出甲基化水平的新方法与常规亚硫酸氢盐测序的比较。

[0068] 图52示出了根据本发明的实施例的血浆DNA的甲基化检测的性能。(A)预测的甲基化概率与通过亚硫酸氢盐测序定量的甲基化水平范围之间的关系。(B)根据本公开中呈现的实施例通过太平洋生物科学公司(PacBio)测序测定的甲基化水平(y轴)与以10Mb分辨率通过亚硫酸氢盐测序定量的甲基化水平(x轴)之间的相关性。

[0069] 图53示出了根据本发明的实施例的太平洋生物科学公司SMRT测序与BS-seq之间的Y染色体的基因组呈现(GR)的相关性。

[0070] 图54示出了根据本发明的实施例的基于CpG区块的甲基化检测的实例,其中每个CpG区块含有一系列CpG位点。5mC:甲基化;C:未甲基化。

[0071] 图55示出了根据本发明的实施例的使用基于CpG区块的方法对人DNA分子的甲基化训练和测试。(A)训练数据集中的性能。(B)独立测试数据集中的性能。

[0072] 图56A和56B示出了根据本发明的实施例的肿瘤组织中的拷贝数变化。

[0073] 图57A和57B示出了根据本发明的实施例的肿瘤组织中的拷贝数变化。

[0074] 图58示出了使用根据本发明的实施例推导的甲基化水平从孕妇血浆中确定血浆DNA组织来源示意图。

[0075] 图59示出了根据本发明的实施例的通过Y染色体读段推导出的胎盘对母本血浆

DNA的贡献与胎儿DNA浓度之间的相关性。

[0076] 图60示出了根据本发明的实施例的总结来自不同人类组织DNA样本的测序数据的表格。

[0077] 图61示出了根据本发明的实施例的分析甲基化谱式的各种方式的图示。

[0078] 图62A和62B示出了根据本发明的实施例的通过亚硫酸氢盐测序和单分子实时测序定量的全基因组水平下的甲基化密度的比较。

[0079] 图63A、63B和63C示出了根据本发明的实施例的通过亚硫酸氢盐测序和单分子实时测序定量的总体甲基化水平的不同相关性。

[0080] 图64A和64B示出了根据本发明的实施例的肝细胞癌(HCC)细胞系和来自健康对照受试者的血沉棕黄层样本的1-Mnt分辨率下的甲基化谱式,所述受试者的甲基化水平通过亚硫酸氢盐测序和单分子实时测序来测定。

[0081] 图65A和65B示出了根据本发明的实施例的针对HCC细胞系(HepG2)和来自健康对照受试者的血沉棕黄层样本的通过亚硫酸氢盐测序和单分子实时测序测定的1-Mnt分辨率下甲基化水平的散点图。

[0082] 图66A和66B示出了根据本发明的实施例的针对HCC细胞系(HepG2)和来自健康对照受试者的血沉棕黄层样本的通过亚硫酸氢盐测序和单分子实时测序测定的100-knt分辨率下甲基化水平的散点图。

[0083] 图67A和67B示出了根据本发明的实施例的通过亚硫酸氢盐测序和单分子实时测序测定甲基化水平的HCC肿瘤组织和邻近正常组织在1-Mnt分辨率下甲基化谱式。

[0084] 图68A和68B示出了根据本发明的实施例的针对HCC肿瘤组织和邻近正常组织的通过亚硫酸氢盐测序和单分子实时测序测定的1-Mnt分辨率下甲基化水平的散点图。

[0085] 图69A和69B示出了根据本发明的实施例的针对HCC肿瘤组织和邻近正常组织的通过亚硫酸氢盐测序和单分子实时测序测定的100-knt分辨率下甲基化水平的散点图。

[0086] 图70A和70B示出了根据本发明的实施例的通过亚硫酸氢盐测序和单分子实时测序测定甲基化水平的HCC肿瘤组织和邻近正常组织在1-Mnt分辨率下甲基化谱式。

[0087] 图71A和71B示出了根据本发明的实施例的针对HCC肿瘤组织和邻近正常组织的通过亚硫酸氢盐测序和单分子实时测序测定的1-Mnt分辨率下甲基化水平的散点图。

[0088] 图72A和72B示出了根据本发明的实施例的针对HCC肿瘤组织和邻近正常组织的通过亚硫酸氢盐测序和单分子实时测序测定的100-knt分辨率下甲基化水平的散点图。

[0089] 图73示出了根据本发明的实施例的肿瘤抑制基因CDKN2A附近的甲基化异常谱式的实例。

[0090] 图74A和74B示出了根据本发明的实施例的通过单分子实时测序检测的差异甲基化区域。

[0091] 图75示出了根据本发明的实施例的使用单分子实时测序的HCC组织与邻近非肿瘤组织之间的乙型肝炎病毒DNA的甲基化谱式。

[0092] 图76A示出了根据本发明的实施例的使用亚硫酸氢盐测序的来自患有肝硬化但未患HCC的患者的肝组织中乙型肝炎病毒DNA的甲基化水平。

[0093] 图76B示出了根据本发明的实施例的使用亚硫酸氢盐测序的HCC组织中乙型肝炎病毒DNA的甲基化水平。

[0094] 图77展示了根据本发明的实施例的甲基化单倍型分析。

[0095] 图78示出了根据本发明的实施例的根据一致性序列测定的经过测序的分子的长度分布。

[0096] 图79A、79B、79C和79D示出了根据本发明的实施例的印记区域中的等位基因甲基化谱式的实例。

[0097] 图80A、80B、80C和80D示出了根据本发明的实施例的非印记区域中的等位基因甲基化谱式的实例。

[0098] 图81示出了根据本发明的实施例的等位基因特异性片段的甲基化水平的表格。

[0099] 图82示出了根据本发明的实施例的用于使用甲基化谱确定怀孕期间胎盘来源的血浆DNA的实例。

[0100] 图83展示了根据本发明的实施例的胎儿特异性DNA甲基化分析。

[0101] 图84A、84B和84C示出了根据本发明的实施例的SMRT-seq的跨不同试剂盒的不同测量窗口尺寸的性能。

[0102] 图85A、85B和85C示出了根据本发明的实施例的SMRT-seq的跨不同试剂盒的不同测量窗口尺寸的性能。

[0103] 图86A、86B和86C示出了根据本发明的实施例的通过亚硫酸氢盐测序和SMRT-seq(Sequel II测序试剂盒2.0)定量的总体甲基化水平的相关性。

[0104] 图87A和87B示出了根据本发明的实施例的各种肿瘤组织与配对的邻近非肿瘤组织之间的总体甲基化水平的比较。

[0105] 图88示出了根据本发明的实施例的使用根据环状一致性序列(CCS)测定的序列上下文来测定甲基化状态。

[0106] 图89示出了根据本发明的实施例的使用根据CCS测定的序列上下文来检测甲基化CpG位点的ROC曲线。

[0107] 图90示出了根据本发明的实施例的在没有CCS信息并且与参考基因组事先未比对的情况下检测甲基化CpG位点的ROC曲线。

[0108] 图91示出了根据本发明的实施例的制备用于单分子实时测序的分子的实例。

[0109] 图92示出了根据本发明的实施例的CRISPR/Cas9系统的图示。

[0110] 图93示出了根据本发明的实施例的用于引入跨越所关注的末端封闭分子的两个切口的Cas9复合物的实例。

[0111] 图94示出了根据本发明的实施例的通过亚硫酸氢盐测序和单分子实时测序测定的Alu区域的甲基化分布。

[0112] 图95示出了根据本发明的实施例的通过使用来自单分子实时测序的结果的模型测定的Alu区域的甲基化水平的分布。

[0113] 图96示出了根据本发明的实施例的组织和组织中Alu区域的甲基化水平的表格。

[0114] 图97示出了根据本发明的实施例的使用与Alu重复序列有关的甲基化信号对不同癌症类型的聚类分析。

[0115] 图98A和98B示出了根据本发明的实施例的在涉及全基因组扩增和M.SsssI处理的测试数据集中读段深度对总体甲基化水平量化的影响。

[0116] 图99示出了根据本发明的实施例的使用不同的子读段深度截止值由SMRT-seq

(Sequel II测序试剂盒2.0)和BS-seq测定的总体甲基化水平之间的比较。

[0117] 图100是表格,其示出了根据本发明的实施例的子读段深度对SMRT-seq(Sequel II测序试剂盒2.0)和BS-seq的两个测量结果之间的甲基化水平的相关性的影响。

[0118] 图101示出了根据本发明的实施例的由Sequel II测序试剂盒2.0产生的数据中关于片段长度的子读段深度分布。

[0119] 图102示出了根据本发明的实施例的检测核酸分子中核苷酸的修饰的方法。

[0120] 图103示出了根据本发明的实施例的用于检测核酸分子中核苷酸的修饰的方法。

[0121] 图104展示了根据本发明的实施例的基于相对单倍型的甲基化失衡分析。

[0122] 图105A和105B是单倍型区块的表格,其示出了根据本发明的实施例的针对情况TBR3033的与邻近非肿瘤组织DNA相比肿瘤DNA中的Hap I与Hap II之间的差异甲基化水平。

[0123] 图106是单倍型区块的表格,其示出了根据本发明的实施例的针对情况TBR3032的与邻近正常组织DNA相比肿瘤DNA中的Hap I与Hap II之间的差异甲基化水平。

[0124] 图107A是表格,其根据本发明的实施例基于由Sequel II测序试剂盒2.0产生的数据总结了示出肿瘤与邻近非肿瘤组织之间的两种单倍型之间的甲基化失衡的单倍型区块的数量。

[0125] 图107B是表格,其根据本发明的实施例基于由Sequel II测序试剂盒2.0产生的数据总结了示出在不同肿瘤阶段的肿瘤组织中的两种单倍型之间的甲基化失衡的单倍型区块的数量。

[0126] 图108展示了根据本发明的实施例的基于相对单倍型的甲基化失衡分析。

[0127] 图109示出了根据本发明的实施例的对具有第一单倍型和第二单倍型的生物体中的病症进行分类的方法。

[0128] 图110展示了根据本发明的实施例的产生人组成部分被甲基化而小鼠组成部分未被甲基化的人-小鼠杂合片段。

[0129] 图111展示了根据本发明的实施例的产生人组成部分未被甲基化而小鼠组成部分被甲基化的人-小鼠杂合片段。

[0130] 图112示出了根据本发明的实施例的连接之后DNA混合物(样本MIX01)中DNA分子的长度分布。

[0131] 图113展示了根据本发明的实施例的第一DNA(A)和第二DNA(B)结合在一起的接合点区域。

[0132] 图114展示了根据本发明的实施例的DNA混合物的甲基化分析。

[0133] 图115示出了根据本发明的实施例的样本MIX01中的CpG位点被甲基化概率的箱线图。

[0134] 图116示出了根据本发明的实施例的在样本MIX02交叉连接之后DNA混合物中DNA分子的长度分布。

[0135] 图117示出了根据本发明的实施例的样本MIX02中的CpG位点被甲基化概率的箱线图。

[0136] 图118是表格,其根据本发明的实施例比较了通过亚硫酸氢盐测序和太平洋生物科学公司测序测定的MIX01的甲基化。

[0137] 图119是表格,其根据本发明的实施例比较了通过亚硫酸氢盐测序和太平洋生物

科学公司测序测定的MIX02的甲基化。

[0138] 图120A和120B示出了根据本发明的实施例的MIX01和MIX02的仅人和仅小鼠DNA的5Mb分类中的甲基化水平。

[0139] 图121A和121B示出了根据本发明的实施例的用于人-小鼠杂合DNA片段的人组成部分和小鼠组成部分的5Mb分类中的MIX01和MIX02的甲基化水平。

[0140] 图122A和122B是代表性图,其示出根据本发明的实施例的单个人-小鼠杂合分子中的甲基化状态。

[0141] 图123示出了根据本发明的实施例的检测生物样本中嵌合分子的方法。

[0142] 图124展示了根据本发明的实施例的测量系统。

[0143] 图125示出了可与根据本发明的实施例的系统和方法一起使用的示例计算机系统的框图。

[0144] 图126示出了根据本发明的实施例的使用DNA末端修复和A-加尾的基于MspI的靶向单分子实时测序。

[0145] 图127A和127B示出了根据本发明的实施例的MspI消化片段的长度分布。

[0146] 图128示出了根据本发明的实施例的具有某些选定长度范围的DNA分子的数量的表格。

[0147] 图129是根据本发明的实施例的限制酶消化之后CpG岛内的CpG位点的覆盖百分比对DNA片段的长度的图。

[0148] 图130示出了根据本发明的实施例的不使用DNA末端修复和A-加尾的基于MspI的靶向单分子实时测序。

[0149] 图131示出了根据本发明的实施例的衔接子自连接的可能性降低的基于MspI的靶向单分子实时测序。

[0150] 图132是根据本发明的实施例的通过基于MspI的靶向单分子实时测序测定的胎盘样本与血沉棕黄层DNA样本之间的总体甲基化水平的图。

[0151] 图133示出了根据本发明的实施例的使用由基于MspI的靶向单分子实时测序测定的胎盘样本和血沉棕黄层样本的DNA甲基化谱的胎盘样本和血沉棕黄层样本的聚类分析。

[0152] 术语

[0153] “组织”对应于集合在一起作为功能单元的一组细胞。单个组织中可以存在超过一种类型的细胞。不同类型的组织可以由不同类型的细胞(例如,肝细胞、肺泡细胞或血细胞)组成,但是也可以对应于来自不同生物体的组织(母体与胎儿;接受移植的受试者的组织;被微生物或病毒感染的生物体的组织)或对应于健康细胞与肿瘤细胞。“参考组织”可以对应于用于测定组织特异性甲基化水平的组织。来自不同个体的相同组织类型的多个样本可以用于测定所述组织类型的组织特异性甲基化水平。

[0154] “生物样本”是指取自人类受试者的任何样本。生物学样本可以是组织活检、细针抽取或血细胞。样本还可以是例如孕妇的血浆或血清或尿液。还可以使用粪便样本。在各个实施例中,来自孕妇的已经富集了细胞游离DNA的生物样本(例如,通过离心方案获得的血浆样本)中的大部分DNA可以是细胞游离的,例如,大于50%、60%、70%、80%、90%、95%或99%的DNA可以是细胞游离的。离心方案可以包含例如在3,000g×10分钟下获得流体部分,并在例如在30,000g下再-离心另外10分钟以除去残留的细胞。在某些实施例中,在3,000g

离心步骤之后,可以随后(例如,使用孔径直径为5μm或更小的过滤器)过滤流体组成部分。

[0155] “序列读段”是指从核酸分子的任何组成部分或全部进行测序的一串核苷酸。例如,序列读段可以是从核酸片段测序的短核苷酸串(例如,20-150个)、在核酸片段的一个或两个末端处的短核苷酸串,或生物样本中存在的整个核酸片段的测序。序列读段可以通过多种方式获得,例如使用测序技术或使用探针,例如通过杂合阵列或捕获探针或扩增技术,如聚合酶链式反应(PCR)或使用单引物的线性扩增或等温扩增。

[0156] “子读段”是由一条环化DNA模板链中的所有碱基产生的序列,所述序列已被DNA聚合酶复制到一条连续链中。例如,子读段可以对应于一条环化DNA模板DNA链。在此类实例中,在环化之后,一个双链DNA分子将具有两个子读段:每个测序通道一个。在一些实施例中,例如由于存在测序错误,因此所产生的序列可以包含一条链中所有碱基的子集。

[0157] “位点”(也被称为“基因组位点”)对应于单个位点,所述位点可以是单个碱基位置或一组相关的碱基位置,例如,CpG位点或较大的一组相关的碱基位置。“基因座”可以对应于包含多个位点的区域。基因座可以仅包含一个位点,这将使所述基因座在所述上下文中等同于一个位点。

[0158] “甲基化状态”是指给定位点处的甲基化状态。例如,位点可以是甲基化的、未甲基化的,或者在某些情况下是不确定的。

[0159] 每个基因组位点(例如,CpG位点)的“甲基化指数”可以指在所述位点处显示出甲基化的DNA片段(例如,如根据序列读段或探针所测定的)占覆盖所述位点的读段总数的比例。“读段”可以对应于从DNA片段获得的信息(例如,位点处的甲基化状态)。可以使用在一个或多个位点处优先与特定甲基化状态的DNA片段杂合的试剂(例如引物或探针)来获得读段。通常,此类试剂在通过根据其甲基化状态差异地修饰或差异地识别DNA分子的工艺处理后被施加,所述甲基化状态是例如亚硫酸氢盐转化、或甲基化敏感的限制酶、或甲基化结合蛋白、或抗甲基胞嘧啶抗体、或识别甲基胞嘧啶和羟甲基胞嘧啶的单分子测序技术(例如单分子实时测序和纳米孔测序(例如来自牛津纳米孔技术公司))。

[0160] 区域的“甲基化密度”可以指区域内示出甲基化的位点处的读段数除以覆盖所述区域中的所述位点的读段的总数。位点可以具有特定的特性,例如,是CpG位点。因此,区域的“CpG甲基化密度”可以指示出CpG甲基化的读段数除以覆盖所述区域中的CpG位点(例如,特定CpG位点、CpG岛内或更大区域内的CpG位点)的读段的总数。例如,可以根据在CpG位点处进行亚硫酸氢盐处理(对应于甲基化胞嘧啶)后未转化的胞嘧啶的总数测定人类基因组中每100kb组距的甲基化密度,作为由映射到100kb区域的序列读段覆盖的所有CpG位点的比例。这种分析也可以针对其它组距长度(例如500bp、5kb、10kb、50kb或1Mb等)进行。区域可以是整个基因组或染色体或染色体的组成部分(例如,染色体臂)。当区域仅包含CpG位点时,所述CpG位点的甲基化指数与所述区域的甲基化密度相同。“甲基化的胞嘧啶的比例”可以指区域中在所分析的胞嘧啶残基(即包含CpG上下文之外的胞嘧啶)的总数内示出被甲基化(例如,在亚硫酸氢盐转化之后未转化)的胞嘧啶位点“C's”的数量。甲基化指数、甲基化密度、在一个或多个位点处甲基化的分子计数以及在一个或多个位点处甲基化的分子(例如,胞嘧啶)的比例是“甲基化水平”的实例。除了亚硫酸氢盐转化之外,可以使用本领域技术人员已知的其它方法来查询DNA分子的甲基化状态,所述方法包含但不限于对甲基化状态敏感的酶(例如,甲基化敏感型限制酶)、甲基化结合蛋白、使用对甲基化状态敏感的平台

的单分子测序(例如,纳米孔测序(Schreiber等人《美国国家科学院院刊(Proc Natl Acad Sci)》2013;110:18910-18915)和(例如来自太平洋生物科学公司的)单分子实时测序(Flusberg等人《自然方法(Nat Methods)》2010;7:461-465))。

[0161] “甲基化组”提供了基因组中多个位点或基因座处的DNA甲基化量的量度。甲基化组可以对应于基因组的全部、基因组的大部分或基因组的一个或多个相对较小部分。

[0162] “妊娠血浆甲基化组”是从妊娠动物(例如,人类)的血浆或血清中测定的甲基化组。妊娠血浆甲基化组是细胞游离甲基化组的实例,因为血浆和血清包含细胞游离DNA。妊娠血浆甲基化组还是混合甲基化组的实例,因为其是来自体内不同器官或组织或细胞的DNA的混合物。在一个实施例中,此类细胞是造血细胞,其包含但不限于红系(即红细胞)谱系、髓系谱系(例如,中性粒细胞和其前体)和巨核细胞谱系的细胞。在妊娠期间,血浆甲基化组可以包含来自胎儿和母体的甲基化组信息。“细胞甲基化组”对应于从患者的细胞(例如,血细胞)中测定的甲基化组。血细胞的甲基化组称为血细胞甲基化组(或血液甲基化组)。

[0163] “甲基化谱”包含与多个位点或区域的DNA或RNA甲基化相关的信息。与DNA甲基化相关的信息可以包含但不限于CpG位点的甲基化指数、区域中CpG位点的甲基化密度(缩写为MD)、相邻区域内的CpG位点的分布、包含多于一个CpG位点的区域内的每个单独CpG位点的甲基化谱式或水平以及非CpG甲基化。在一个实施例中,甲基化谱可以包含多于一种类型的碱基(例如胞嘧啶或腺嘌呤)的甲基化或未甲基化谱式。基因组中的大部分的甲基化谱可以被视为相当于甲基化组。哺乳动物基因组中的“DNA甲基化”通常是指在CpG二核苷酸中向胞嘧啶残基的5’碳上添加甲基(即5-甲基胞嘧啶)。DNA甲基化可能在其它上下文中的胞嘧啶中发生,例如,CHG和CHH,其中H是腺嘌呤、胞嘧啶或胸腺嘧啶。胞嘧啶甲基化还可以呈5-羟甲基胞嘧啶形式。还报道了非胞嘧啶甲基化,如N^6-甲基腺嘌呤。

[0164] “甲基化谱式”是指甲基化和未甲基化碱基的顺序。例如,甲基化谱式可以是单个DNA链、单个双链DNA分子或另一种类型的核酸分子上的甲基化碱基的顺序。作为实例,三个连续的CpG位点可能具有以下甲基化谱式中的任一种:UUU、MMM、UMM、UMU、UUM、MUM、MUU或MMU,其中“U”指示未甲基化位点,并且“M”指示甲基化位点。当将此概念扩展到包含但不限于甲基化的碱基修饰时,将使用术语“修饰谱式”,所述术语是指经过修饰的和未经修饰的碱基的顺序。例如,修饰谱式可以是单个DNA链、单个双链DNA分子或另一种类型的核酸分子上的经过修饰的碱基的顺序。作为实例,三个连续的潜在地可修饰的位点可能具有以下修饰谱式中的任一种:UUU、MMM、UMM、UMU、UUM、MUM、MUU或MMU,其中“U”指示未经修饰的位点,并且“M”指示经过修饰的位点。不以甲基化为基础的碱基修饰的一个实例是氧化变化,如在8-氧代鸟嘌呤中。

[0165] 术语“高甲基化的”和“低甲基化的”可以指如通过单分子甲基化水平测量的单个DNA分子的甲基化密度,例如,分子内甲基化碱基或核苷酸的数量除以所述分子内的可甲基化的碱基或核苷酸的总数。高甲基化分子是单分子甲基化水平处于或高于阈值的分子,所述阈值可以根据应用而定义。阈值可以是5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或95%。低甲基化分子是单分子甲基化水平处于或低于阈值的分子,所述阈值可以根据应用而定义,并且可以根据应用而变化。阈值可以是5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或95%。

[0166] 术语“高甲基化的”和“低甲基化的”还可以指如通过这些分子的多分子甲基化水平测量的DNA分子群体的甲基化水平。高甲基化分子群体是多分子甲基化水平处于或高于阈值的群体，所述阈值可以根据应用而定义，并且可以根据应用而变化。阈值可以是5％、10％、20％、30％、40％、50％、60％、70％、80％、90％或95％。低甲基化分子群体是多分子甲基化水平处于或低于阈值的分子，所述阈值可以根据应用而定义。阈值可以是5％、10％、20％、30％、40％、50％、60％、70％、80％、90％和95％。在一个实施例中，可以将分子群体与一个或多个选定基因组区域进行比对。在一个实施例中，一个或多个选定基因组区域可以与如癌症、遗传病、印记基因失调性疾病、代谢障碍或神经疾病等疾病有关。一个或多个选定基因组区域的长度可以为50个核苷酸(nt)、100nt、200nt、300nt、500nt、1000nt、2knt、5knt、10knt、20knt、30knt、40knt、50knt、60knt、70knt、80knt、90knt、100knt、200knt、300knt、400knt、500knt或1Mnt。

[0167] 术语“测序深度”是指基因座被与所述基因座进行比对的序列读段所覆盖的次数。所述基因座可以与核苷酸一样小，或者与染色体臂一样大，或者与整个基因组一样大。测序深度可以表示为50x、100x等，其中“x”是指基因座被序列读段覆盖的次数。测序深度也可以应用于多个基因座或整个基因组，在此情况下，x可以指分别对基因座或单倍体基因组或整个基因组进行测序的平均次数。超深测序可以指测序深度是至少100x。

[0168] 如本文所使用的，术语“分类”是指与样本的特定性质相关联的任何一个或多个数字或其它一个或多个字符。例如，“+”符号(或词语“阳性”)可以表示样本归类为具有缺失或扩增。分类可以是二元(例如，阳性或阴性)或具有更多分类等级(例如，1到10或0到1的标度)。

[0169] 术语“截止值”和“阈值”是指在操作中使用的预定数字。例如，截止长度可以指超过一定长度则不包含片段的长度。阈值可以是高于或低于一定值应用特定分类的值。这些术语中的任一个可以在这些上下文中的任一种中使用。截止值或阈值可以是“参考值”或者可以从表示特定类别或区分两个或多个类别的参考值中得出。如本领域技术人员将理解的，可以以各种方式确定此类参考值。例如，可以针对具有不同已知分类的两个不同的受试者确定度量，并且可以选择参考值作为一个分类的代表(例如，平均值)或度量的两个集群之间的值(例如，选择以获得期望的灵敏度和特异性)。作为另一实例，可以基于样本的统计分析或模拟来确定参考值。

[0170] 术语“癌症等级”可以指是否存在癌症(即，存在或不存在)、癌症的阶段、肿瘤的尺寸、是否存在转移、身体的总肿瘤负荷、癌症对治疗的反应和/或癌症严重程度的其它度量(例如癌症复发)。癌症等级可以是数字或其它标记，如符号、字母和颜色。所述等级可以是零。癌症等级还可以包含恶化前或癌前病状(状态)。癌症等级可以以各种方式使用。例如，筛查可以检查癌症是否存在于以前不知道患有癌症的人身上。评估可以调查被诊断出患有癌症的人，以监测癌症随着时间的进展、研究治疗的有效性或确定预后。在一个实施例中，预后可以表示为患者死于癌症的可能性，或在特定持续时间或时间之后癌症进展的可能性，或癌症转移的可能性或程度。检测可以意指“筛查”或可以意指检查具有癌症提示性特征(例如症状或其它阳性测试)的人是否患有癌症。

[0171] “病理等级”(或病症等级)可以指与生物体相关联的病理的数量、程度或严重性，其中所述等级可以如以上针对癌症所描述的等级。病理的另一个实例是对移植器官的排

斥。其它示例病理可以包含基因印记病症、自身免疫攻击(例如,损害肾脏的狼疮性肾炎的或多发性硬化症)、炎性疾病(例如,肝炎)、纤维化过程(例如,肝硬化)、脂肪浸润(例如,脂肪肝疾病)、退行性过程(例如阿尔茨海默氏病)和缺血性组织损伤(例如,心肌梗塞或中风)。受试者的健康状态可以被认为是无病理分类。

[0172] “妊娠相关病症”包含以母本和/或胎儿组织中基因相对表达水平异常为特征的任何病症。这些病症包含但不限于先兆子痫、宫内发育迟缓、侵入性胎盘形成、早产、新生儿溶血病、胎盘功能不全、胎儿水肿、胎儿畸形、HELLP综合征、全身性红斑狼疮和其它母体免疫性疾病。

[0173] 缩写“bp”是指碱基对。在一些情况下,“bp”可以用来表示DNA片段的长度,即使DNA片段可以是单链的且不包含碱基对。在单链DNA的上下文中,“bp”可以解释为提供核苷酸长度。

[0174] 缩写“nt”是指核苷酸。在一些情况下,“nt”可以用来表示碱基单元中单链DNA的长度。而且,“nt”可以用来表示相对位置,如被分析的基因座的上游或下游。在涉及技术概念化、数据呈现、处理和分析的一些上下文中,“nt”和“bp”可以互换使用。

[0175] 术语“序列上下文”可以指一段DNA中的碱基组成(A、C、G或T)和碱基顺序。这样一段DNA可以围绕碱基,此段DNA可围绕作为碱基修饰分析目标的碱基。例如,序列上下文可以指经受碱基修饰分析的碱基的上游和/或下游的碱基。

[0176] 术语“动力学特征”可以指来源于测序的特征,包含来自单分子实时测序的特征。此类特征可以用于碱基修饰分析。示例动力学特征包含上游和下游序列上下文、链信息、脉冲间持续时间、脉冲宽度和脉冲强度。在单分子实时测序中,持续监测聚合酶活性对DNA模板的影响。因此,从此类测序中产生的测量结果可以被认为是动力学特征,例如,核苷酸序列。

[0177] 术语“机器学习模型”可以包含基于使用样本数据(例如,训练数据)对测试数据进行预测的模型,并且因此可以包含监督学习。通常使用计算机或处理器开发机器学习模型。机器学习模型可以包含统计模型。

[0178] 术语“数据分析框架”可以包括将数据作为输入,且随后输出输出预测结果算法和/或模型。“数据分析框架”的实例包含统计模型、数学模型、机器学习模型、其它人工智能模型和其组合。

[0179] 术语“实时测序”可以指在测序所涉及的反应过程期间涉及数据收集或监测的技术。例如,实时测序可以涉及光学监测或拍摄DNA聚合酶掺入新碱基。

[0180] 术语“约(about)”或“大约(approximately)”可以意指在特定值的一个可接受的误差范围内,如本领域一般熟悉此项技术者所测定的,这将部分地取决于如何测量或测定所述值,即,测量系统的局限性。例如,根据本领域的实践,“约”可以意指在1个或大于1个标准偏差内。可替代地,“约”可以意指给定值的最多20%、最多10%、最多5%或最多1%的范围。可替代地,特别是对于生物系统或过程,术语“约”或“大约”可以意指在值的数量级内、值的5倍内,且更佳在2倍内。当在本申请和权利要求书中描述特定值时,除非另外指出,否则应假设术语“约”表示所述特定值在可接受的误差范围内。术语“约”可以具有本领域一般熟悉此项技术者通常理解的含义。术语“约”可以指±10%。术语“约”可以指±5%。

具体实施方式

[0181] 实现碱基修饰(包含甲基化碱基)的无亚硫酸氢盐测定是不同研究工作的主题,但没有一个被证明在商业上是可行的。最近,已经发表了一项使用温和条件进行5mC和5hmC碱基转化的用于检测5mC和5hmC的无亚硫酸氢盐方法(Y.Liu等人,2019)。此方法涉及多个步骤的酶促和化学反应,包含十-十一易位(TET)氧化、吡啶硼烷还原和PCR。转化反应的每个步骤的效率以及PCR偏差会对5mC分析的最终准确性产生不利影响。例如,已经报道了5mC转化率为约96%,其中假阴性率为约3%。这种表现可能会限制人们检测基因组中甲基化的某些细微变化的能力。另一方面,酶促转化无法在整个基因组中获得同样好的表现。例如,5hmC的转化率比5mC的转化率低8.2%,并且非CpG的转化率比CpG情形的转化率低11.4%(Y.Liu等人,2019)。因此,理想的情况是开发用于测量天然DNA分子的碱基修饰而无需任何事先的转化(化学转化或酶促转化或其组合)步骤,甚至无需扩增步骤的方法。

[0182] 存在许多概念验证研究(Q.Liu等人,2019;Ni等人,2019),其中通过长读段纳米孔测序方法(例如,使用牛津纳米孔技术公司开发的系统)产生的电信号使得能够使用深度学习方法来检测甲基化状态。除牛津纳米孔公司外,还有其它允许长读段的单分子测序方法。一个实例是单分子实时测序。单分子实时测序的一个实例是将太平洋生物科学公司SMRT系统商业化。由于单分子实时测序(例如,太平洋生物科学SMRT系统)的原理与基于非光学的纳米孔系统(例如,牛津纳米孔技术公司)的原理不同,因此针对此类基于非光学的纳米孔系统开发的碱基修饰检测方法不能用于单分子实时测序。例如,非光学纳米孔系统不是为了捕获通过(单分子实时测序所采用的,如太平洋生物科学公司SMRT系统所采用的)基于固定化DNA聚合酶的DNA合成产生的荧光信号的谱式而设计的。作为进一步的实例,在牛津纳米孔测序平台中,每个测得的电事件都与k-mer(例如,5-mer)相关联(Q.Liu等人,2019)。然而,在太平洋生物科学公司SMRT测序平台中,每个荧光事件通常与单个掺入的碱基相关联。此外,单个DNA分子将在太平洋生物科学公司SMRT测序中进行多次测序,包含沃森链和克里克链。相反,对于牛津纳米孔公司的长读段测序方法,沃森链和克里克链中的每一个各执行一次序列读出。

[0183] 据报道聚合酶动力学会受到大肠杆菌序列中甲基化状态的影响(Flusberg等人,2010)。先前的研究表明,当与检测6mA、4mC、5hmC和8-氧代鸟嘌呤相比时,使用单分子实时测序的聚合酶动力学来推导单分子中特定CpG的甲基化状态(5mC对C)更具挑战性。原因是甲基基团很小且朝向主沟并且不参与碱基配对,从而导致由5mC引起的动力学中非常细微的扰动(Clark等人,2013)。因此,用于测定单分子水平处胞嘧啶的甲基化状态的方法很少。

[0184] 铃木(Suzuki)等人开发了一种试图结合相邻CpG位点的脉冲间持续时间(IPD)比率以提高鉴定那些位点的甲基化状态的置信度的算法(铃木等人,2016)。然而,此算法仅允许预测完全甲基化或完全未甲基化的基因组区域,但是缺乏测定中间甲基化谱式的能力。

[0185] 关于单分子实时测序,目前的方法仅单独使用一个或两个参数,由于5-甲基胞嘧啶与胞嘧啶之间的测量差异,在检测5mC时获得的准确性非常有限。例如,Flusberg等人证明了IPD在包含N6-甲基腺苷、5-甲基胞嘧啶和5-羟甲基胞嘧啶的碱基修饰中发生了变化。然而,未发现测序动力学的脉冲宽度(PW)具有显著影响。因此,在其用于预测碱基修饰的方法中,以检测N6-甲基腺苷为例,仅使用IPD而不使用PW。

[0186] 在同一组的后续出版物中(Clark等人,2012;Clark等人,2013),在用于检测5-甲

基胞嘧啶的算法中引入了IPD但未引入PW。在Clark等人2012年的研究中,5-甲基胞嘧啶的检出率范围在未将其转化为5-甲基胞嘧啶的情况下仅为1.9%到4.3%。此外,在Clark等人2013年的研究中,作者进一步重申了5-甲基胞嘧啶的动力学特征的微妙之处。为了克服检测5-甲基胞嘧啶的低灵敏度,Clark等人进一步开发了一种方法,所述方法使用十-十一易位(Tet)蛋白将5-甲基胞嘧啶转化为5-羧甲基胞嘧啶以便提高5-甲基胞嘧啶的灵敏度(Clark等人,2013),因为由5-羧基胞嘧啶引起的IPD变化比由5-甲基胞嘧啶引起的要多得多。

[0187] 在Blow等人的最新报告中,Flusberg等人先前描述的基于IPD比率的方法被用于以每个生物体130倍的读段覆盖率检测217种细菌和13种古细菌物种中的碱基修饰(Blow等人,2016)。在其鉴定的所有碱基修饰中,只有5%涉及5-甲基胞嘧啶。他们将5-甲基胞嘧啶的这种低检出率归因于用于检测5-甲基胞嘧啶的单分子实时测序的低灵敏度。在大多数细菌中,DNA甲基转移酶(MTases)靶向一组序列基序以在这些基序中的几乎所有基序处进行甲基化(例如,在大肠杆菌中,5'-GmATC-3'被Dam甲基化或者5'-CmCWGG-3'被Dcm甲基化),这些基序位点中只有一小部分基序位点保持未甲基化(Beaulaurier等人,2019)。此外,使用基于IPD的方法对经过或未经Tet蛋白处理的5'-CCWGG-3'基序中第二个C的甲基化状态进行分类,得出5-甲基胞嘧啶的检出率分别为95.2%和1.9%(Clark等人,2013)。总体而言,没有事先进行碱基转化的IPD方法(例如,使用Tet蛋白)遗漏了大多数5-甲基胞嘧啶。

[0188] 在上述研究中(Clark等人,2012;Clark等人,2013;Blow等人,2016),使用了基于IPD的算法而没有考虑候选碱基修饰所在的序列上下文。其它小组已尝试考虑核苷酸的序列上下文检测碱基修饰。例如,Feng等人使用分层模型来分析IPD,以在相应的序列上下文中检测4-甲基胞嘧啶和6-甲基腺苷(Feng等人,2013)。然而,在其方法中只考虑了所关注的碱基处的IPD和与所述碱基邻近的序列上下文,但是没有使用与所关注的碱基邻近的所有相邻碱基的IPD信息。另外,算法中未考虑PW,并且他们没有提供有关5-甲基胞嘧啶的检测的任何数据。

[0189] 在另一项研究中,Schadt等人开发了一种被称为条件随机场的统计方法,用于分析所关注的碱基和相邻碱基的IPD信息,以确定所关注的碱基是否为5-甲基胞嘧啶(Schadt等人,2012)。在这项工作中,还通过将碱基输入到方程式中来考虑这些碱基之间的IPD相互作用。然而,并没有在方程式中输入核苷酸序列(即A、T、G或C)。当应用所述方法测定M.Sau3AI质粒的甲基化状态时,即使在800倍的质粒序列的序列覆盖率下,ROC曲线下的面积也接近0.5。而且,在其方法中,他们在分析中没有考虑PW。

[0190] 在Beckman等人的另一项研究中,比较了基因组中的靶细菌基因组与(例如,通过全基因组扩增获得的)完全未甲基化基因组之间共享相同的4-nt或6-nt基序的所有序列的IPD(Beckman等人,2014)。此类分析的目的仅在于鉴定将可能更频繁地受到碱基修饰影响的基序。在研究中,仅考虑了潜在地经过修饰的碱基的IPD,而不考虑相邻碱基的PW或IPD。其方法不能提供关于单个核苷酸的甲基化状态的信息。

[0191] 总而言之,这些先前仅利用IPD或结合相邻核苷酸中的序列信息来分组数据的尝试不能以有意义或实际的准确性测定5-甲基胞嘧啶的碱基修饰。在Goui1等人的最新综述中,作者推断,由于信噪比低,使用单分子实时测序来检测单个分子中的5-甲基胞嘧啶是不准确的(Goui1等人,2019)。在这些先前的研究中,使用动力学特征进行全基因组甲基化组

分析是否可行尚不明确，尤其是对于如人类基因组、癌症基因组或胎儿基因组等复杂的基因组而言。

[0192] 与先前的研究相比，本公开中描述的方法的一些实施例以针对测量窗口内的每个碱基的IPD、PW和序列上下文为基础。推断出如果可以使用多个指标的组合，例如，同时利用包含上游和下游序列上下文、链信息、IPD、脉冲宽度以及脉冲强度的特征，也许能够实现准确测量单碱基分辨率下的碱基修饰(例如，mC检测)。序列上下文是指一段DNA中的碱基组成(A、C、G或T)和碱基顺序。这样一段DNA可以围绕一个碱基，这个碱基可被用于进行碱基修饰分析或作为碱基修饰分析的目标。在一个实施例中，所述一段DNA可以在经受碱基修饰分析的碱基的近侧。在另一个实施例中，所述一段DNA可以远离经受碱基修饰分析的碱基。所述一段DNA可以在经受碱基修饰分析的碱基的上游和/或下游。

[0193] 在一个实施例中，用于碱基修饰分析的上游和下游序列上下文、链信息、IPD、脉冲宽度以及脉冲强度的特征被称为动力学特征。

[0194] 本公开中呈现的实施例可以用于从但不限于细胞系、来自生物体的样本(例如，实体器官、实体组织、通过内窥镜检查获得的样本、来自孕妇的血液或血浆或血清或尿液、绒毛膜绒毛活检等)、从环境中获得的样本(例如细菌、细胞污染物)、食物(例如，肉类)中获得的DNA。在一些实施例中，本公开中呈现的方法还可以在首先富集基因组的一部分的步骤之后例如使用杂合探针(Albert等人，2007;Okou等人，2007;Lee等人，2011)或基于物理分离(例如，基于长度等)的方法应用，或者在限制酶消化(例如MspI)或基于Cas9的富集(Watson等人，2019)之后应用。尽管本发明不需要酶促转化或化学转化来起作用，但是在某些实施例中，可以包含此类转化步骤以进一步增强本发明的性能。

[0195] 本公开的实施例使得在检测碱基修饰或测量修饰水平时准确性或实用性或便利性有所改善。可以直接检测修饰。实施例可以避免酶促转化或化学转化，所述转化可能不会保留所有修饰信息以供检测。另外，某些酶促转化或化学转化可能与某些类型的修饰不兼容。本公开的实施例还可以避免通过PCR进行的扩增，所述扩增可能不会将碱基修饰信息转移到PCR产物。另外，DNA的两条链可以一起测序，由此使来自一条链的序列与其互补序列能够与另一条链配对。相比之下，PCR扩增会分开双链DNA的两条链，因此这种序列配对很困难。

[0196] 在有或没有酶促转化或化学转化的情况下测定的甲基化谱可以用于分析生物样本。在一个实施例中，甲基化谱可以用于检测细胞DNA的来源(例如，母本或胎儿、组织、病毒或肿瘤)。对组织中异常甲基化谱的检测有助于鉴定个体发育障碍以及鉴定和预测肿瘤或恶性肿瘤。单倍型之间的甲基化水平失衡可以用于检测病症，包含癌症。单个分子中的甲基化谱式可以(例如，通过遗传或基因组操作)鉴定嵌合(例如，在病毒与人之间)和杂合DNA(例如，在天然基因组中通常未融合的两个基因之间)；或两个物种之间。

[0197] 甲基化分析可以通过增强训练来改进，所述增强训练可以包含缩小训练集中使用的数据。可以靶向特定区域进行分析。在实施例中，此类靶向可以涉及单独或与一种或多种其它试剂结合的酶，所述酶可以基于其序列切割DNA序列或基因组。在一些实施例中，所述酶是识别和切割一种或多种特定DNA序列的限制酶。在其它实施例中，可以组合使用多于一种具有不同识别序列的限制酶。在一些实施例中，限制酶可以基于识别序列的甲基化状态切割或不切割。在一些实施例中，所述酶是CRISPR/Cas家族中的一种。例如，可以使用

CRISPR/Cas9系统或其它基于导向RNA(即,与互补靶DNA序列结合并在过程中引导酶在靶基因组位置处发挥作用的短RNA序列)的系统来靶向所关注的基因组区域。在一些情况下,无需与参考基因组比对即可进行甲基化分析。

[0198] I.单分子实时测序的甲基化检测

[0199] 本公开的实施例允许直接检测碱基修饰而无需酶促转化或化学转化。通过单分子实时测序获得的动力学特征(例如,序列上下文、IPD和PW)可以通过机器学习进行分析,以开发用于检测修饰存在或修饰不存在的模型。修饰水平可以用于测定DNA分子的来源或病症的存在或水平。

[0200] 出于说明目的,以太平洋生物科学公司SMRT测序作为单分子实时测序的实例,DNA聚合酶分子定位在作为零模波导(ZMW)孔的底部。ZMW是一种用于将光限制在较小观察体积内的纳米光子装置,所述观察体积可以是直径非常小的孔并且不允许光在检测波长范围内传播,使得只有来自固定化聚合酶掺入的染料标记的核苷酸的光信号的发射可在低且恒定的背景信号下被检测到(Eid等人,2009)。DNA聚合酶催化荧光标记的核苷酸掺入到互补核酸链中。

[0201] 图1示出了对携带碱基修饰的分子进行单分子环状一致性测序的实例。分子102、104和106携带碱基修饰。可以将DNA分子(例如,分子106)与发夹衔接子连接以形成连接的分子108。随后,连接分子108可以形成环化分子110。环化分子可以与固定化DNA聚合酶结合并且可以启动DNA合成。不携带碱基修饰的分子也可以被测序。

[0202] 图2示出了通过单分子实时测序进行测序的携带甲基化和/或未甲基化CpG位点的分子的实例。首先将DNA分子与发夹衔接子连接以形成环状分子,所述环状分子可以与固定化DNA聚合酶结合并启动DNA合成。在图2中,将DNA分子202与发夹衔接子连接以形成连接的分子204。然后,连接的分子204形成环化分子206。没有CpG位点的分子也可以被测序。环化分子206包含未甲基化CpG位点208,所述位点仍然可以被测序。

[0203] 一旦DNA合成开始,荧光染料标记的核苷酸将被基于环状DNA模板的固定化聚合酶掺入到新合成的链中,从而引起光信号的发射。因为DNA模板已被环化,所以整个环形DNA模板将多次通过聚合酶(即DNA模板中的一个核苷酸将被多次测序)。从所述过程中产生的序列被称为子读段,其中环化DNA模板中的所有碱基全部通过DNA聚合酶。ZMW中的一个分子将产生多个子读段,因为聚合酶可以连续围绕整个环状DNA模板多次。在一个实施例中,子读段可以仅包含序列子集、碱基修饰或环状DNA模板的其它分子信息,因为在一个实施例中存在测序错误。

[0204] 如图3所展示的,所得荧光脉冲的到达时间和持续时间将允许测量聚合酶动力学。脉冲间持续时间(IPD)是两个发射脉冲之间的时间周期长度的度量,所述发射脉冲中的每一个都提示了新生链中掺入的荧光标记的核苷酸(图3)。如图3所示,脉冲宽度(PW)是与关于碱基调用的脉冲的持续时间相关联的反映聚合酶动力学的另一度量。PW可以是信号峰值高度(即,所掺入的染料标记的核苷酸的荧光强度)的0%处的脉冲的持续时间。在一个实施例中,PW可以由例如但不限于信号峰值高度的5%、10%、20%、30%、40%、50%、60%、70%、80%或90%处的脉冲的持续时间来限定。在一些实施例中,PW可以是峰下的面积除以信号峰值高度。

[0205] 此类聚合酶动力学(如IPD)已被证明受合成序列和微生物序列(例如大肠杆菌)中

的碱基修饰的影响，如N6-甲基腺嘌呤(6mA)、5-甲基胞嘧啶(5mC)和5-羟甲基胞嘧啶(5hmC)(Flusberg等人，2010)。Flusberg等人在2010年的研究中未使用序列上下文和IPD作为独立输入来检测修饰，这导致模型缺乏实际有意义的检测准确性。Flusberg等人仅使用序列上下文来确认GATC中发生6mA。Flusberg等人未提及将序列上下文与IPD结合使用作为检测甲基化状态的输入。

[0206] 当仅使用IPD信号时，赋予互补链中5-甲基胞嘧啶新碱基掺入的弱扰动使得即使对于相对简单的微生物基因组而言甲基化判读也极具挑战性，因为据报道甲基化基序$C^{m}CWGG$的检出率仅为1.9％到4.3％(Clark等人，2013)。例如，太平洋生物科学公司提供的分析软件包(SMRT Link v6.0.0)无法执行5mC分析。此外，SMRT Link v5.1.0的先前版本要求在甲基化分析之前使用Tet1酶将5mC转化为5-羧基胞嘧啶(5caC)，因为与5caC相关联的IPD信号将得到增强(Clark等人，2013)。因此，没有研究表明使用单分子实时测序以全基因组方式分析人类基因组的天然DNA的可行性就不足为奇了。

[0207] II.测量窗口谱式和机器学习模型

[0208] 需要在不进行酶促或化学转化修饰和/或碱基的情况下检测碱基中的修饰的技术。如本文所述的，可以使用通过对靶碱基周围的碱基进行单分子实时测序获得的动力学特征数据检测靶碱基中的修饰。动力学特征可以包含脉冲间持续时间、脉冲宽度和序列上下文。可以获得靶碱基上游和下游的一定数量的核苷酸的测量窗口的这些动力学特征。这些特征(例如，在测量窗口中的特定位置处)可以用于训练机器学习模型。作为样本制备的实例，DNA分子的两条链可以通过发夹衔接子连接，由此形成环状DNA分子。环状DNA分子允许获得沃森链和克里克链中的任一个或两者的动力学特征。可以基于测量窗口中的动力学特征来开发数据分析框架。然后可以使用此数据分析框架来检测包含甲基化的修饰。本节描述了用于检测修饰的各种技术。

[0209] A.使用单链

[0210] 如图4所示，作为实例，从太平洋生物科学公司SMRT测序获得了沃森链的子读段以分析一个特定碱基的碱基修饰状态。在图4中，可以将来自经受碱基修饰分析的碱基每侧的3个碱基定义为测量窗口400。在一个实施例中，将这7个碱基(即3-核苷酸(nt)上游和下游序列和一个用于碱基修饰分析的核苷酸)的序列上下文、IPD和PW编译为2维(即2-D)矩阵作为测量窗口。在所示实例中，测量窗口400仅针对沃森链的一个子读段。本文描述了其它变型。

[0211] 矩阵的第一行402表示所研究的序列。在矩阵的第二行404中，0的位置表示用于碱基修饰分析的碱基。-1、-2和-3的相对位置分别指示经受碱基修饰分析的碱基上游的位置1-nt、2-nt和3-nt。+1、+2和+3的相对位置分别指示经受碱基修饰分析的碱基下游的位置1-nt、2-nt和3-nt。每个位置包含2列，所述列包含对应的IPD值和PW值。以下4行(行408、412、416和420)分别对应于链(例如，沃森链)中4种类型的核苷酸(A、C、G和T)。矩阵中IPD值和PW值的存在取决于在特定位置处对哪种对应的核苷酸类型进行了测序。如图4所示，在0的相对位置处，IPD值和PW值示出在指示沃森链中“G”的行中，这表明在所述位置的序列结果中被判读为鸟嘌呤。列中与经过测序的碱基不对应的其它网格将被编码为“0”。作为实例，对应于2-D数字矩阵(图4)的序列信息对于沃森链将是5’-GATGACT-3’。

[0212] 如图5所描绘的一个实施例中所示，测量窗口可以应用于来自克里克链的数据。通

过单分子实时测序获得克里克链的子读段以分析一种特定碱基的碱基修饰状态。在图5中，可以将来自经受碱基修饰分析的碱基每侧的3个碱基和经受碱基修饰分析的碱基定义为测量窗口。在一个实施例中，将这7个碱基(即3-核苷酸(nt)上游和下游序列和一个用于碱基修饰分析的核苷酸)的序列上下文、IPD、PW编译为2维(即2-D)矩阵作为测量窗口。矩阵的第一行指示所研究的序列。在矩阵的第二行中，0的位置表示用于碱基修饰分析的碱基。-1、-2和-3的相对位置分别指示经受碱基修饰分析的碱基上游的位置1-nt、2-nt和3-nt。+1、+2和+3的相对位置分别指示经受碱基修饰分析的碱基下游的位置1-nt、2-nt和3-nt。每个位置包含2列，所述列包含对应的IPD值和PW值。以下4行对应于此链(例如克里克链)中的4种类型的核苷酸(A、C、G和T)。矩阵中IPD值和PW值的存在取决于在特定位置处对哪种对应的核苷酸类型进行了测序。如图5所示，在0的相对位置处，IPD值和PW值示出在指示克里克链中“T”的行中，这表明在所述位置的序列结果中被判读为胸腺嘧啶。列中与经过测序的碱基不对应的其它网格将被编码为“0”。作为实例，对应于2-D数字矩阵(图5)的序列信息对于克里克链将是5’-ACTTAGC-3’。

[0213] B.使用沃森链和克里克链

[0214] 图6示出了可以以结合沃森链和其互补克里克链的数据的方式实施测量窗口的实施例。如图6所示，通过单分子实时测序获得了沃森链和克里克链的子读段以分析一种特定碱基的修饰。在一个实施例中，来自环状DNA模板的克里克链的测量窗口与来自沃森链的测量窗口互补，对其进行碱基修饰分析。在图6中，可以将来自经受碱基修饰分析的沃森链中第一碱基每侧的3个碱基和第一碱基定义为第一测量窗口。可以将来自克里克链中第二碱基每侧的3个碱基和第二碱基定义为第二测量窗口。第二碱基与第一碱基互补。在一个实施例中，将来自沃森链和克里克链的这7个碱基(即3-核苷酸(nt)上游和下游序列和一个用于碱基修饰分析的核苷酸)的序列上下文、IPD、PW编译为2维(即2-D)矩阵。来自沃森链和克里克链的这些测量窗口分别被视为第一测量窗口和第二测量窗口。

[0215] 沃森链和克里克链的矩阵的第一行指示所研究的序列。在沃森链的矩阵的第二行中，0的位置表示用于碱基修饰分析的第一碱基。克里克链的矩阵的第二行所示的0的位置表示与第一碱基互补的第二碱基。-1、-2和-3的相对位置分别指示第一碱基和第二碱基上游的位置1-nt、2-nt和3-nt。+1、+2和+3的相对位置分别指示第一碱基和第二碱基下游的位置1-nt、2-nt和3-nt。来源于沃森链和克里克链的每个位置将对应于2个列，所述列包含对应的IPD值和PW值。沃森链和克里克链的矩阵中的以下4行分别对应于特定链(例如，克里克链)中4种类型的核苷酸(A、C、G和T)。矩阵中IPD值和PW值的存在取决于在特定位置处对哪种对应的核苷酸类型进行了测序。

[0216] 如图6所示，在0的相对位置处，IPD值和PW值显示在指示沃森链中“A”和克里克链中“T”的行中，这表明在沃森链和克里克链的所述位置处的序列结果中分别被判读为腺嘌呤和胸腺嘧啶。列中与经过测序的碱基不对应的其它网格将被编码为“0”。作为实例，对应于沃森链的2-D数字矩阵(图6)的序列信息将是5’-ATAAGTT-3’。对应于克里克链的2-D数字矩阵(图6)的序列信息将是5’-AACTTAT-3’。

[0217] 如本实例所示，可以合并来自沃森链和克里克链的数据以形成新的矩阵，也可以将其视为测量窗口。此新矩阵可以用作用于训练机器学习模型的单个样本。因此，新矩阵中的所有值都可以被视为单独的特征，但是例如当使用卷积神经网络(CNN)时，2D矩阵中的特

定位置可能会产生影响。可以通过矩阵中的非零条目来传递不同链的各个位置处的序列上下文。

[0218] 图7示出了可以以来自沃森链和克里克链的数据不是彼此完全互补的位置的方式来实施测量窗口。如图7所示，第一测量窗口是5'-ATAAGTT-3'；并且第二个测量窗口是5'-GTAACGC-3'。在一些实施例中，沃森链和克里克链可以彼此移位，使得位置不互补。

[0219] 图8示出了测量窗口可以用于分析CpG位点处的甲基化状态。0的位置对应于CpG位点的胞嘧啶，并且因此在两条链之间存在位移，使得两条链的C都在0位置。因此，包含在测量窗口中的来自沃森链和克里克链的序列中只有一部分彼此互补。在其它实施例中，测量窗口中来自沃森链和克里克链的所有序列可以彼此互补。在其它实施方式中，测量窗口中来自沃森链和克里克链的序列均不互补。

[0220] 在一个实施例中，对于测量窗口，经受碱基修饰分析的碱基周围的DNA段的长度可以是非对称的。例如，所述碱基上游的X-nt和下游的Y-nt可以用于碱基修饰分析。X可以包含但不限于0、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、100、150、200、300、400、500、1000、2000、4000、5000和10000；Y可以包含但不限于0、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、100、150、200、300、400、500、1000、2000、4000、5000和10000。

[0221] C.训练模型和检测修饰

[0222] 图9示出了关于如何使用测量窗口来测定任何碱基修饰的一般过程。对已知未经修饰的和经过修饰的DNA样本进行单分子实时测序。经过修饰的DNA(例如，经过修饰的分子902)意指碱基(例如，碱基904)在位点处具有修饰(例如，甲基化)。未经修饰的DNA(例如，未经修饰的分子906)意指碱基(例如，碱基908)在位点处不具有修饰。两组DNA都可以人工创建或加工以形成经过修饰的/未经修饰的DNA。

[0223] 在阶段910，随后可以对样本进行单分子实时测序。作为SMRT测序的一部分，环状分子可以通过反复地穿过固定化DNA聚合酶进行多次测序。每次获得的序列信息将被视为子读段。因此，一个环状DNA模板将产生多个子读段。可以使用例如但不限于BLASR将测序子读段与参考基因组比对(Mark J Chaisson等人，《BMC生物信息学(BMC Bioinformatics)》2012;13:238。在各个其它实施例，BLAST(Altschul SF等人，《分子生物学期刊(J Mol Biol.)》1990;215(3):403-410)、BLAT(Kent WJ，《基因组研究(Genome Res.)》2002;12(4):656-664)、BWA(Li H等人，《生物信息学(Bioinformatics)》2010;26(5):589-595)、NGMLR(Sedlazeck FJ等人，《自然方法(Nat Methods)》2018;15(6):461-468)、LAST(Kielbasa SM等人，《基因组研究》2011;21(3):487-493)和Minimap2(Li H，《生物信息学》2018;34(18):3094-3100)可以用于将子读段与参考基因组比对。比对可以允许来自多个子读段的数据被组合(例如，平均)，因为相同位置的每个子读段中的数据可以被识别。

[0224] 在阶段912，从比对结果中获得了进行碱基修饰分析的碱基周围的IPD、PW和序列上下文。在阶段914，以某种结构(例如但不限于如图9所示的2-D矩阵)记录IPD、PW和序列上下文。

[0225] 在阶段916，使用包含具有已知碱基修饰的参考动力学谱式衍生的分子的多个2-D

矩阵来训练分析、计算、数学或统计模型。在阶段918，通过训练结果产生统计模型。为了简单起见，图9仅示出了通过训练开发的统计模型，但是可以开发任何模型或数据分析框架。示例数据分析框架包含机器学习模型、统计模型和数学模型。统计模型可以包含但不限于线性回归、逻辑回归、深度递归神经网络(例如，长短期记忆，LSTM)、贝叶斯分类器、隐马尔可夫模型(HMM)、线性判别分析(LDA)、k均值聚类、基于密度的带噪声应用空间聚类(DBSCAN)、随机森林算法和支持向量机(SVM)。进行碱基修饰分析的碱基周围的DNA段可能是所述碱基上游的X-nt和下游的Y-nt，即“测量窗口”。

[0226] 可以在训练过程中使用数据结构，因为正确的输出(即，修饰状态)是已知的。例如，与来自一条或多条沃森链和/或克里克链的碱基上游和下游的3-nt相对应的IPD、PW和序列上下文可以用于构建2-D矩阵，所述矩阵用于训练对碱基修饰进行分类的一个或多个统计模型。以此方式，训练可以提供一种模型，该模型可以对具有先前已知状态的核酸的位置处的碱基修饰进行分类。

[0227] 图10示出了关于从携带已知状态的碱基修饰的DNA样本中学习的一个或多个统计模型如何能够检测碱基修饰的一般过程。对具有未知状态的碱基修饰的样本进行SMRT测序。使用例如上述技术将测序子读段与参考基因组进行比对。另外或替代地，子读段可以彼此比对。其它实施例可以仅使用一个子读段或独立地对其进行分析，从而不进行比对。

[0228] 对于进行碱基修饰分析的碱基，可以使用训练步骤(图9)中使用的相当的测量窗口从比对结果中的沃森链和/或克里克链获得IPD、PW和序列上下文，并与所述碱基相关联。在另一个实施例中，训练与测试程序之间的测量窗口将是不同的。例如，训练与测试程序之间的测量窗口的尺寸可能是不同的。那些IPD、PW和序列上下文将转换为2-D矩阵。可以将测试样本的此类2-D矩阵与参考动力学特征进行比较以确定碱基修饰。例如，可以通过从训练样本中学到的一个或多个统计模型将测试样本的2-D矩阵与参考动力学特征进行比较，使得可以确定测试样本中核酸分子的位点处的碱基修饰。统计模型可以包含但不限于线性回归、逻辑回归、深度递归神经网络(例如，长短期记忆，LSTM)、贝叶斯分类器、隐马尔可夫模型(HMM)、线性判别分析(LDA)、k均值聚类、基于密度的带噪声应用空间聚类(DBSCAN)、随机森林算法和支持向量机(SVM)。

[0229] 图11示出了关于如何利用所述方法对CpG位点的甲基化状态进行分类的一般过程。对已知在CpG位点处未甲基化和甲基化的DNA样本进行单分子实时测序。将测序子读段与参考基因组进行了比对。使用了沃森链数据。

[0230] 从比对结果中，获得了进行甲基化分析的CpG位点处的胞嘧啶周围的IPD、PW和序列上下文并记录在特定结构中，例如但不限于图11所示的2-D矩阵。使用了许多包含参考动力学谱式衍生的具有已知甲基化状态的分子的2-D矩阵来训练一个或多个统计模型。被调查的碱基周围的一段DNA可能是所述碱基上游的X-nt和下游的Y-nt，即“测量窗口”。X可以包含但不限于0、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、100、150、200、300、400、500、1000、2000、4000、5000和10000；Y可以包含但不限于0、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、100、150、200、300、400、500、1000、2000、4000、5000和10000。在一个实施例中，与来自沃森链的碱基上游和下

游的3-nt相对应的IPD、PW和序列上下文可以用于构建2-D矩阵，所述矩阵用于训练对碱基修饰进行分类的一个或多个统计模型。

[0231] 图12示出了对未知样本的碱基修饰进行分类的一般程序。对未知甲基化状态的样本进行了单分子实时测序。将测序子读段与参考基因组进行了比对。

[0232] 对于比对结果中CG位点的胞嘧啶，可以使用在训练步骤(图11)中应用的相当的测量窗口从沃森链获得与被研究修饰的碱基相关联的IPD、PW和序列上下文。那些IPD、PW和序列上下文可以转换为2-D矩阵。将测试样本的此类2-D矩阵与图11中所展示的参考动力学谱式进行比较以测定甲基化状态。X11

[0233] 图13和图14示出了来自克里克链的动力学特征可以用于如上所阐述的类似于沃森链的程序的训练和测试程序。一个或多个统计模型可以是相同或不同的模型。当模型不同时，其可以用于获得独立的分类，所述分类可以进行比较，例如，如果所述模型一致，则鉴定为修饰状态。如果所述模型不一致，则可以鉴定为未分类状态。当所述模型是相同模型时，可以将数据组合成单个数据结构，例如图6中的矩阵。

[0234] 图15和图16示出了来自沃森链和克里克链的动力学特征可以用于如上所阐述的训练和测试程序。对已知在CpG位点处未甲基化和甲基化的DNA样本进行单分子实时测序。将测序子读段与参考基因组进行比对，尽管子读段彼此之间的比对是可能的，如本文所述的其它方法可以做到。

[0235] 对于比对结果中的子读段，获得了经受甲基化分析的CpG位点的胞嘧啶周围的IPD、PW和序列上下文。因为DNA分子是通过使用两个发夹衔接子环化的(例如，遵循SMRTBell模板制备方案)，所以可以对环形分子进行多次测序，由此产生分子的多个子读段。子读段可以用于产生环形一致性测序(CCS)读段。通常，对于本文所述的所有方法，一个ZMW可以产生多个子读段，但仅对应于一个CCS读段。

[0236] 在一些实施例中，完全未甲基化的数据集可以通过对人DNA片段进行PCR来创建。例如，完全甲基化的数据集可以通过经CpG甲基转移酶M.SssI处理的人DNA片段产生，其中所有CpG位点均假定为是甲基化的。在其它实例中，可以使用另一种CpG甲基转移酶，如M.MpeI。在其它实施例中，具有已知甲基化状态的合成序列或具有不同甲基化水平的预先存在的DNA样本，或通过限制酶切割甲基化和未甲基化DNA分子随后进行连接而产生的杂合甲基化状态(其将产生一定比例的嵌合甲基化/未甲基化DNA分子)可以用于训练甲基化预测模型或分类器。

[0237] 动力学谱式(包含序列上下文、IPD和脉冲宽度(PW))的转换可以是用于分析CG位点处的甲基化状态的包括来自沃森链和克里克链的特征的2-D矩阵，如图15所展示的。此方法使得能够准确地捕获由甲基化胞嘧啶以及其附近的序列上下文引起的细微动力学变化。与本文所述的各种方法中的任何方法一样，对于子读段中存在的每个CpG，可以使用(例如，CpG位点的胞嘧啶上游和下游的3个碱基)测量窗口进行后续分析，从而使得总共7个核苷酸(包含CpG位点的胞嘧啶)被一起分析。可以计算这7个核苷酸中每个碱基的IPD和PW。为了捕获归因于动力学变化的序列上下文，可以将IPD和PW信号编译为特定的碱基判读、相对测序位置和链信息，如图15所示。为了简单起见，此类数据结构被称为动力学2-D数字矩阵。

[0238] 此类2-D数字矩阵类似于“2-D数字图像”。例如，2-D数字矩阵的第一行包含进行甲基化分析的CpG基因座的胞嘧啶周围的相对位置，其中所述胞嘧啶位点的上游和下游具有

3-nt。0的位置表示甲基化待测定的胞嘧啶位点。-1和-2的相对位置指示所讨论的胞嘧啶上游的1-nt和2-nt。+1和+2的相对位置指示将使用的胞嘧啶下游的1-nt和2-nt。每个位置将对应于2列，所述列包含对应的IPD值和PW值。每行对应于沃森链和克里克链中4种类型的核苷酸(A、C、G和T)。矩阵中IPD值和PW值的填充取决于在特定位置处的测序结果(即子读段)中预设了哪种对应的核苷酸类型。

[0239] 如图15所示，在0的相对位置处，IPD值和PW值显示在沃森链中的“C”行中，这表明在所述位置处判读为胞嘧啶。列中与经过测序的碱基不对应的其它网格将被编码为“0”。作为实例，对应于2-D数字矩阵(图15)的序列信息对于沃森链和克里克链分别是5’-ATACGTT-3’和5’-TAACGTA-3’。在这种情况下，沃森链和克里克链中CpG位点的胞嘧啶侧翼的上游和下游序列将是不同的。由于CpG位点处的甲基化在沃森链与克里克链之间是对称的(Lister等人，2009)，因此在一个优选实施例中，两条链中的动力学用于训练甲基化预测模型。在另一个实施例中，沃森链和克里克链可以用于分别训练甲基化预测模型。

[0240] 考虑到单分子实时测序的高通量，在一个实施例中，深度学习算法(例如卷积神经网络(CNN))(LeCun等人，1989)可能适合于区分甲基化CpG和未甲基化CpG。另外或替代地，还可以使用其它算法，例如但不限于线性回归、逻辑回归、深度递归神经网络(例如，长短期记忆，LSTM)、贝叶斯分类器、隐马尔可夫模型(HMM)、线性判别分析(LDA)、k均值聚类、基于密度的带噪声应用的空间聚类(DBSCAN)、随机森林算法和支持向量机(SVM)等。训练可以分别使用沃森链和克里克链，也可以在组合的新矩阵中使用，如图6-8所描述的。

[0241] 动力学谱式的另一个转换可以是N维矩阵。N可以例如是1、3、4、5、6和7。例如，3-D矩阵将是根据被分析的DNA段的串联CG位点的数量分层的2-D矩阵的堆叠，其中第3维将是所述DNA段中串联CG位点的数量。在一些实施例中，(例如，通过脉冲的峰值高度或通过脉冲信号下方的面积来测量的)脉冲强度或脉冲幅度也可以并入矩阵中。脉冲强度(图3中脉冲峰值幅度的度量)可以被添加到与原始2-D矩阵顶部的PW值和IPD值相关联的列邻近的额外列，或者被添加到第3维以形成3-D矩阵。

[0242] 作为进一步的实例，8(行)x21(列)的2D矩阵可以被转换成包括168个元素的1-D矩阵(即向量)。并且可以扫描此1-D矩阵，例如以执行CNN或其它模型化。作为另一个实例，方法可以将8x21的2-D矩阵拆分为多个较小的矩阵，例如两个4x21的2-D矩阵。将这两个较小的矩阵在竖直方向上放在一起可以提供3-D矩阵(即x＝21，y＝4，z＝2)。方法可以扫描第1个2-D矩阵，并且然后扫描第2个2-D矩阵，以形成用于机器学习的数据呈现。可以进一步拆分数据以形成更高维度的矩阵。另外，可以将二级结构信息添加到数据结构，例如2-D矩阵顶部的额外矩阵(1-D矩阵)。此类额外矩阵可以编码测量窗口内的每个碱基是否涉及二级结构(例如茎-环结构)，例如，涉及“茎”的碱基被编码为0且涉及“环”的碱基被编码为1。

[0243] 在一个实施例中，单个DNA分子内CpG位点的甲基化状态可以表示为基于统计模型被甲基化概率，而不是给出“甲基化”或“未甲基化”的定性结果。概率为1指示，基于统计模型，CpG位点可以被认为是甲基化的。概率为0指示，基于统计模型，CpG位点可以被视为是未甲基化的。在随后的下游分析中，可以使用截止值对基于概率特定CpG位点是归类为“甲基化”还是“未甲基化”进行分类。截止值的可能值包含5％、10％、15％、20％、25％、30％、35％、40％、45％、50％、55％、60％、65％、70％、75％、80％、85％、90％或95％。CpG位点被甲基化的预测概率大于预定义的截止值可以被归类为“甲基化”，而CpG位点被甲基化概率不

大于预定义的截止值可以被归类为“未甲基化”。所需的截止值将使用例如ROC曲线分析训练数据集而获得。

[0244] 图16示出了对来自沃森链和克里克链的未知样本的甲基化状态进行分类的一般程序。对具有未知甲基化状态的样本进行单分子实时测序。测序子读段可以与其它方法一样与参考基因组进行比对或彼此比对,以测定给定位置的一致性值(例如,平均值、中值、众数或其它统计值)。如图所示,两条链的测量结果可以合并为单个2D矩阵。

[0245] 对于比对结果中CG位点的胞嘧啶,可以使用训练步骤(图16)中应用的相当的测量窗口(CpG位点胞嘧啶上游和下游的3-nt)从沃森链获得与被调查修饰的碱基相关联的IPD、PW和序列上下文,但是可以使用不同尺寸的窗口。可以将测试样本的此类2-D矩阵与图16中所展示的参考动力学谱式进行比较以测定甲基化状态。

[0246] III.用于甲基化检测的示例模型训练

[0247] 为了测试所提出方法的可行性和有效性,在进行单分子实时测序之前,制备了通过M.SssI处理(甲基化文库)和PCR扩增(未甲基化文库)的胎盘DNA文库。分别获得了甲基化和未甲基化文库的44,799,736和43,580,452个子读段,所述子读段分别对应于421,614和446,285个环状一致性序列(CCS)。作为结果,每个分子在甲基化和未甲基化文库中的测序中值为34和32次。数据集由通过太平洋生物科学公司Sequel测序试剂盒3.0制备的DNA产生。此试剂盒被开发用于原始的太平洋生物科学公司Sequel测序仪。为了区分Sequel与其后续版本Sequel II,在本文中将原始Sequel称为Sequel I。因此,Sequel测序试剂盒3.0在本文中将被称为Sequel I测序试剂盒3.0。针对Sequel II测序仪设计的测序试剂盒包含也在本公开中描述的Sequel II测序试剂盒1.0和Sequel II测序试剂盒2.0。

[0248] 使用了从甲基化和未甲基化文库产生的经过测序的分子的50%来训练统计模型(并使用剩余的50%进行验证),所述统计模型在这种情况下是卷积神经网络(CNN)模型。作为实例,CNN模型可以具有一个或多个卷积层(例如,1D或2D层)。卷积层可以使用一个或多个不同的过滤器,其中每个过滤器使用对特定矩阵元素的局部(例如,相邻或周围)矩阵值进行操作的内核,由此向特定矩阵元素提供新的值。一种实施方案使用两个1D卷积层(每个层具有100个内核尺寸为4的过滤器)。可以分别应用过滤器且随后进行组合(例如,以加权平均的方式)。所得矩阵可以小于输入矩阵。

[0249] 卷积层之后可以是ReLU(整流线性单元)层,所述ReLU之后可以是丢弃率为0.5的丢弃层。ReLU是可以对从一个或多个卷积层产生新矩阵(图像)的各个值进行运算的激活函数的实例。还可以使用其它激活函数(例如,sigmoid、softmax等)。可以使用此类层中的一个或多个层。丢弃层可以在ReLU层或最大池化层上使用,并充当防止过度拟合的正则化层。丢弃层可以在训练过程中使用,以作为训练的一部分在一个优化过程(例如,减少成本、函数损失)的不同迭代期去忽略不同的(例如,随机)值。

[0250] 可以在ReLU层之后使用最大池化层(例如,池的尺寸为2)。最大池化层的作用类似于卷积层,但是可以取输入被内核重叠的区域的最大值,而不是在输入与内核之间取点乘积。可以使用一个或多个另外的卷积层。例如,可以进一步使用丢弃率为0.5的丢弃层将来自池化层的数据输入到另外两个1D卷积层(例如,每个层具有128个内核尺寸为2的过滤器,随后是ReLU层)。使用了池大小为2的最大池化层。最后,可以使用完全连接的层(例如,具有10个神经元,随后是ReLU层)。具有一个神经元的输出层之后可以是sigmoid层,由此产生甲

基化概率。可以对层、过滤器和内核尺寸的各种设置进行调整。在此训练数据集中，使用了来自甲基化和未甲基化文库的468,596和432,761个CpG位点。

[0251] A.训练数据集和测试数据集的结果

[0252] 图17A示出了训练数据集中每个单个DNA分子中每个CpG位点被甲基化的概率。甲基化文库中甲基化的概率比未甲基化文库高得多。因为被甲基化概率的截止值为0.5，94.7％的未甲基化CpG位点被正确预测为未甲基化，并且84.7％的甲基化CpG被正确预测为甲基化。

[0253] 图17B示出了测试数据集的性能。使用由训练数据集训练的模型来预测来自甲基化和未甲基化文库的独立测试数据集中的469,729和432,024个CpG位点的甲基化状态。因为被甲基化概率的截止值为0.5,94.0％的未甲基化CpG位点被正确预测为未甲基化，并且84.1％的甲基化CpG位点被正确预测为甲基化。这些结果表明，使用新型动力学转换结合序列上下文可以测定(例如来自人类受试者的)DNA中的甲基化状态。

[0254] 藉由在模型中包括特征的子集，评估每个特征(序列上下文，IPD以及PW)在预测CpG甲基化状态方面的能力。在训练数据集中，具有(i)仅序列上下文、(ii)仅IPD和(iii)PW的模型的AUC值分别为0.5、0.74和0.86。而结合IPD和序列上下文以0.86的AUC提高了性能。序列上下文(“Seq”)、IPD和PW的组合分析以0.94的AUC显著提高了性能(图18A)。独立测试数据集的性能与训练数据集相当(图18B)。

[0255] 将CpG位点的子读段深度定义为覆盖子读段和其周围10bp的子读段的平均数量。如图19A和19B所示，CpG位点的子读段深度越高，将实现的甲基化检测的准确性就越高。例如，如测试数据集中所示(图19B)，如果每个CpG位点的深度至少为10，则预测甲基化状态的AUC将为0.93。然而，如果每个CpG位点的子读段深度至少为300，则预测甲基化状态的AUC将为0.98。另一方面，即使深度为1，也可以获得0.9的AUC，这表明我们的方法可以使用低测序深度来实现甲基化预测。

[0256] 为了测试链信息对甲基化分析性能的影响，分别根据本公开中呈现的实施例使用源自沃森链和克里克链的序列上下文、IPD和PW进行训练。图20A和图20B示出了使用单链(即沃森链或克里克链)进行训练和测试是可行的，因为在训练数据集和测试数据集中AUC可以达到最多0.91和0.87。使用包含沃森链和克里克链的两条链(例如，如图6-8所描述)将产生最佳性能(在训练数据集和测试数据集中，AUC分别为:0.94和0.90)，这表明链信息对实现最佳性能将是重要的。

[0257] 进一步测试了CpG位点上游和下游的核苷酸的不同数目，以研究此参数如何影响在本公开中提出的根据本公开中呈现的实施例的性能。图21A和21B示出了在CpG的上下文中胞嘧啶上游和下游的核苷酸的数量将影响甲基化预测的准确性。例如，出于说明目的，考虑但不限于被分析的胞嘧啶上游和下游的2个核苷酸(nt)、3nt、4nt、6nt、8nt、10nt、15nt和20nt，使用被分析的胞嘧啶上游和下游的2nt的方法的AUC在训练数据集和测试数据集中仅为0.50，而使用被调查的胞嘧啶上游和下游的15nt的方法的AUC在训练数据集和测试数据集中将增加到0.95和0.92。这些结果表明，改变被分析的胞嘧啶侧翼上游和下游区域的长度将允许找出最佳性能。在一个实施例中，如图21B所示，将使用胞嘧啶上游和下游的3nt来测定甲基化状态，这可以实现0.89的AUC。

[0258] 在一个实施例中，可以使用被调查的胞嘧啶侧翼的不对称序列来根据本公开中呈

现的实施例执行分析。例如，可以使用胞嘧啶上游的2nt结合下游的1nt、3nt、4nt、5nt、6nt、7nt、8nt、9nt、10nt、11nt、12nt、13nt、14nt、15nt、16nt、17nt、18nt、19nt、20nt、25nt、30nt、35nt和40nt；可以使用胞嘧啶上游的3nt结合下游的1nt、2nt、4nt、5nt、6nt、7nt、8nt、9nt、10nt、11nt、12nt、13nt、14nt、15nt、16nt、17nt、18nt、19nt、20nt、25nt、30nt、35nt和40nt；可以使用胞嘧啶上游的4nt结合下游的1nt、2nt、3nt、5nt、6nt、7nt、8nt、9nt、10nt、11nt、12nt、13nt、14nt、15nt、16nt、17nt、18nt、19nt、20nt、25nt、30nt、35nt和40nt。作为另一个实例，可以使用胞嘧啶下游的2nt结合上游的1nt、3nt、4nt、5nt、6nt、7nt、8nt、9nt、10nt、11nt、12nt、13nt、14nt、15nt、16nt、17nt、18nt、19nt、20nt、25nt、30nt、35nt和40nt；可以使用胞嘧啶下游的3nt结合上游的1nt、2nt、4nt、5nt、6nt、7nt、8nt、9nt、10nt、11nt、12nt、13nt、14nt、15nt、16nt、17nt、18nt、19nt、20nt、25nt、30nt、35nt和40nt；可以使用胞嘧啶下游的4nt结合上游的1nt、2nt、3nt、5nt、6nt、7nt、8nt、9nt、10nt、11nt、12nt、13nt、14nt、15nt、16nt、17nt、18nt、19nt、20nt、25nt、30nt、35nt和40nt。在某些实施例中，通过利用与胞嘧啶上游的n-nt和下游的m-nt相关联的IPD、PW、链信息和序列上下文，可以提高测定甲基化状态的准确性。此类变化的测量窗口可以应用于其它类型的碱基修饰分析，如5hmC、6mA、4mC和oxoG，或本文公开的任何修饰。此类变化的测量窗口可以包含DNA二级结构分析，如G-四联体和茎环结构。以上说明了此类实例。还可以添加此类二级结构信息作为矩阵中的另一列。

[0259] 图22A和图22B示出了使用仅与至少3个碱基的下游碱基相关联的动力学谱式来测定甲基化状态是可行的。根据本公开中呈现的实施例，通过使用与胞嘧啶和其下游的3、4、6、8和10个碱基相关联的特征，在训练数据集中，测定甲基化状态的AUC分别为0.91、0.92、0.94、0.94和0.94；在测试数据集中，AUC分别为0.87、0.88、0.90、0.90和0.90。

[0260] 然而，图23A和图23B示出如果仅使用与上游碱基相关联的特征，则分类能力似乎会降低其区分甲基化状态的能力。对于2到10个上游碱基，训练数据集和测试数据集中的AUC均为0.50。

[0261] 图24和图25示出上游碱基和下游碱基的不同组合将允许在测定甲基化状态中实现最佳分类能力。例如，与胞嘧啶上游的8个碱基和下游的8个碱基相关联的特征将在此数据集中实现最佳性能，其中训练数据集和测试数据集中的AUC分别为0.94和0.91。

[0262] 图26示出了关于CpG位点处的甲基化状态分类的特征的相对重要性。括号中的“W”和“C”指示链信息，“W”用于沃森链并且“C”用于克里克链。使用随机森林测定每个特征的重要性，包含序列上下文、IPD和PW。随机森林树分析显示，IPD和PW的特征重要性在受到调查的胞嘧啶的下游达到峰值，这表明对分类能力的主要贡献是受到调查的胞嘧啶下游的IPD和PW。

[0263] 随机森林由多个决策树构成。在决策树的构建期间，使用基尼不纯度来确定应针对决策节点采用哪种决策逻辑。对最终分类结果影响较大的重要特征可能在距决策树的根部更近的节点中，而对最终分类结果影响较小的不重要特征可能在距根部更远的节点中。因此，可以通过计算相对于随机森林中的所有决策树的根部的平均距离来估计特征重要性。

[0264] 在一些实施例中，沃森链与克里克链之间的CpG位点处的甲基化判读的一致性可以进一步用于提高特异性。例如，可能需要将示出甲基化的两条链都称为甲基化状态，并且

将示出未甲基化的两条链都称为未甲基化状态。由于已知CpG位点处的甲基化通常是对称的,因此从每条链中进行确认可以提高特异性。

[0265] 在各个实施例中,来自整个分子的总体动力学特征可以用于测定甲基化状态。例如,在单分子实时测序期间,整个分子中的甲基化将影响整个分子的动力学。通过对整个模板DNA分子的测序动力学(包含IPD、PW、片段长度、链信息和序列上下文)进行建模,可以提高关于分子是甲基化还是未甲基化的分类的准确性。作为实例,测量窗口可以是整个模板分子。IPD、PW或其它动力学特征的统计值(例如,平均值、中值、众数、百分位等)可以用于测定整个分子的甲基化。

[0266] B.其它分析技术的局限性

[0267] 据报道,基于IPD对特定序列基序中特定C的甲基化检测是非常低的,例如,灵敏度仅为1.9%(Clark等人,2013)。还尝试了在不使用PW度量并且仅对IPD使用截止值而不对本文所述的数据结构使用的情况下通过将不同的序列基序与IPD结合来重现此类分析。例如,提取了被调查的CpG侧翼上游和下游的3-nt。所述CpG的IPD根据以所述CpG为中心的6-nt侧翼序列(即分别是上游和下游3nt)的上下文而分为不同的组(对于6个位置为4096个组)。使用ROC研究了相同序列基序内甲基化CpG与未甲基化CpG之间的IPD。例如,比较了未甲基化"AATCGGAC"基序和甲基化"AAT^{m}CGGAC"基序中CpG的IPD,从而示出了AUC为0.48。因此,相对于使用各种的实施例,在特定序列组中使用截止值表现较差。

[0268] 图27示出了用于进行甲基化检测而不使用脉冲宽度信号的以上基于基序的IPD分析的性能(Beckmann等人,《BMC生物信息学》2014)。竖直柱状图表示跨被分析的CpG位点侧翼的不同k-mer基序的平均AUC(即,被调查的CpG位点周围的碱基数)。图27示出了发现跨不同k-mer基序(例如,所讨论的CpG位点周围的2-mer、3-mer、4-mer、6-mer、8-mer、10-mer、15-mer、20-mer)的甲基化与未甲基化胞嘧啶之间的基于IPD的辨别能力的平均AUC小于60%。这些结果表明,在给定的基序上下文中考虑候选核苷酸的IPD而不考虑相邻核苷酸的IPD(Flussberg等人,2010)的效果将不如本文公开的测定CpG甲基化的方法。

[0269] 还测试了Flusberg等人的研究中的方法(Flusberg等人,2010)。分析了总共5,948,348个DNA片段,所述片段是经受甲基化分析的胞嘧啶上游的2-nt和下游的6-nt。有2,828,848个甲基化片段和3,119,500个未甲基化片段。如图28所示,发现了使用IPD和PW从主成分分析中推导出的信号在具有甲基化胞嘧啶(mC)和未甲基化胞嘧啶(C)的片段之间有很大的重叠,这表明Flusberg等人描述的方法缺乏实际意义上的准确性。这些结果表明,在Flussberg等人的研究(Flussberg等人,2010)中使用的在碱基和相邻碱基处线性组合PW值和IPD值的主成分分析无法可靠或有意义地区分5-甲基胞嘧啶和未甲基化胞嘧啶。

[0270] 图29示出了在涉及IPD和PW的Flussberg等人的研究(Flussberg等人,2010)中使用了两种主成分的基于主成分分析的方法的AUC(AUC:0.55)比涉及IPD和PW以及本公开中所示的序列上下文的基于卷积神经网络的方法(AUC:0.94)的准确性要差得多。

[0271] C.其它数学/统计模型

[0272] 在另一个实施例中,其它数学/统计模型(例如包含但不限于随机森林和逻辑回归)可以通过调整以上开发的特征来训练。对于CNN模型,训练数据集和测试数据集是由经过M.SssI处理(甲基化)和PCR扩增(未甲基化)的DNA构建的,用于训练随机森林(Breiman,2001)。在此随机森林分析中,用6个特征来描述每个核苷酸:IPD、PW和对碱基身份进行编码

的4组分二进制向量。在此类二进制向量中，A、C、G和T分别用[1,0,0,0]、[0,1,0,0]、[0,0,1,0]和[0,0,0,1]进行编码。对于每个被分析的CpG位点，将其上游和下游的10nt信息并入到两条链中，从而形成252维(252-D)向量，其中每个特征代表一个维度。上述具有252-D向量的训练数据集用于训练随机森林模型以及逻辑回归模型。经过训练的模型用于预测独立测试数据集中的甲基化状态。随机森林由100个决策树构成。在树构建期间，使用了引导程序样本。在分割每个决策树的节点时，采用了基尼不纯度来确定最佳分割，并且在每个分割中将最多考虑15个特征。而且，决策树的每片叶子都必须包含至少60个样本。

[0273] 图30A和图30B示出了使用随机森林和逻辑回归进行甲基化预测的方法的性能。图30A示出了训练数据集中CNN、随机森林和逻辑回归的AUC值。图30B示出了测试数据集中CNN、随机森林和逻辑回归的AUC值。在训练数据集和测试数据集中，使用随机森林的方法的AUC分别达到0.93和0.86。

[0274] 用相同的252-D向量描述的训练数据集用于训练逻辑回归模型。经过训练的模型用于预测独立测试数据集中的甲基化状态。具有L2正则化的逻辑回归模型(Ng和Y.,2004)与训练数据集拟合。如图30A和图30B所示，在训练数据集和测试数据集中，使用逻辑回归的方法的AUC分别达到0.87和0.83。

[0275] 因此，这些结果表明，除了CNN之外的某些模型(例如但不限于随机森林和逻辑回归)可以用于使用在本公开中开发的特征和分析方案的甲基化分析。这些结果还表明，根据本公开的实施例实施的在测试数据集中的AUC为0.90的CNN(图30B)优于随机森林(AUC：0.86)和逻辑回归(AUC:0.83)。

[0276] D.核酸的6mA修饰的测定

[0277] 除了甲基化CpG，本文所述的方法还可以检测其它DNA碱基修饰。例如，可以检测包含6mA形式的甲基化腺嘌呤。

[0278] 1.使用动力学特征和序列上下文的6mA检测

[0279] 为了评估用于测定核酸的碱基修饰的所公开实施例的性能和实用性，进一步分析了N6-腺嘌呤甲基化(6mA)。在一个实施例中，通过用未甲基化腺嘌呤(uA)、未甲基化胞嘧啶(C)、未甲基化鸟嘌呤(G)和未甲基化胸腺嘧啶(T)进行全基因组扩增，(例如，从胎盘组织中提取的)大约1ng的人DNA被扩增以获得100ng的DNA产物。

[0280] 图31A示出了一种通过全基因组扩增产生具有未甲基化腺嘌呤的分子的方法的实例。在图中，“uA”表示未甲基化腺嘌呤，并且“mA”表示甲基化腺嘌呤。使用耐核酸外切酶的硫代磷酸酯修饰的随机六聚体作为引物执行全基因组扩增，所述六聚体在基因组上随机结合，从而允许聚合酶(例如Phi29 DNA聚合酶)扩增DNA(例如，通过等温线性扩增)。在阶段3102，双链DNA变性。在阶段3106，当许多随机六聚体(例如，3110)退火到变性模板DNA(即单链DNA)时引发扩增反应。如3114所示，当链3118的六聚体介导的DNA合成在5’到3’方向上进行并到达下一个六聚体介导的DNA合成位点时，聚合酶置换新合成的DNA链(3122)并继续链延伸。经置换的链变成了再次与随机六聚体结合的单链DNA模板，并且可能引发新的DNA合成。等温过程中重复的六聚体退火和链置换将导致经扩增的DNA产物的高产率。此处描述的扩增可以被归入多重置换扩增(MDA)技术。

[0281] 经扩增的DNA产物被进一步片段化成例如但不限于长度为100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、5kb、10kb、20kb、30kb、40kb、50kb、60kb、

70kb、80kb、90kb、100kb或其它期望长度范围的片段。片段化过程可以包含酶消化、雾化、流体动力剪切和超声处理等。作为结果，最初的碱基修饰(如6mA)可以通过用未甲基化A(uA)进行全基因组扩增几乎被消除。图31A示出了两条链均具有未甲基化A的DNA产物的可能片段(3126、3130和3134)。对此类不具有mA的全基因组经扩增的DNA产物进行单分子实时测序以产生uA数据集。

[0282] 图31B示出了一种通过全基因组扩增产生具有甲基化腺嘌呤的分子的方法的实例。在图中，“uA”表示未甲基化腺嘌呤，并且“mA”表示甲基化腺嘌呤。通过用6mA和未甲基化C、G和T进行全基因组扩增，大约1ng人DNA被扩增以获得10ng DNA产物。可以通过一系列化学反应产生甲基化腺嘌呤(J D Engel等人《生物化学杂志(J Biol Chem.)》1978;253:927-34)。如图31B所展示的，使用耐核酸外切酶的硫代磷酸酯修饰的随机六聚体作为引物执行全基因组扩增，所述六聚体在基因组上随机结合，从而允许聚合酶(例如Phi29 DNA聚合酶)扩增DNA(例如通过等温线性扩增)，类似于图31A。耐外切核酸酶的硫代磷酸酯修饰的随机六聚体对校对DNA聚合酶的3’→5’外切核酸酶活性具有抵抗力。因此，在扩增期间，将保护随机六聚体免于降解。

[0283] 当许多随机六聚体退火到变性模板DNA(即单链DNA)时引发扩增反应。当六聚体介导的DNA合成在5’到3’方向上进行并到达下一个六聚体介导的DNA合成位点时，聚合酶置换新合成的DNA链并继续链延伸。经置换的链变成了再次与随机六聚体结合的单链DNA模板，并且引发新的DNA合成。等温过程中重复的六聚体退火和链置换将导致经扩增的DNA产物的高产率。

[0284] 经扩增的DNA产物被进一步片段化成例如但不限于长度为100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、5kb、10kb、20kb、30kb、40kb、50kb、60kb、70kb、80kb、90kb、100kb或其它长度组合的片段。如图31B所示，经扩增的DNA产物将包含跨每条链的腺嘌呤位点的不同形式的甲基化谱式。例如，双链分子的两条链都可以相对于腺嘌呤(分子I)被甲基化，当两条链在全基因组扩增期间从DNA合成中衍生时会产生所述腺嘌呤。

[0285] 作为另一个实例，双链分子的一条链可以包含在腺嘌呤位点(分子II)的交错甲基化谱式。交错甲基化谱式被定义为包含DNA链中存在的甲基化和未甲基化碱基的混合物的谱式。在以下实例中，使用包含DNA链中存在的甲基化和未甲基化腺嘌呤的混合物的交错腺嘌呤甲基化谱式。这种类型的双链分子(分子II)可能是由于含有未甲基化腺嘌呤的未甲基化六聚体与DNA链结合并引发DNA延伸而产生的。将对此类含有具有未甲基化腺嘌呤的六聚体的经扩增的DNA产物进行测序。可替代地，这种类型的双链分子(分子II)将由来自含有未甲基化腺嘌呤的原始模板DNA的片段化DNA引发，因为此类片段化DNA可以作为引物与DNA链结合。将对此类含有部分具有未甲基化腺嘌呤的原始DNA的经扩增的DNA产物进行测序。由于未甲基化六聚体引物仅是所得DNA链的一小部分，因此大多数片段将仍包含6mA。

[0286] 作为另一个实例，双链DNA分子的一条链可以在腺嘌呤位点被甲基化，而另一条链可以是未甲基化的(分子III)。当提供不具有甲基化腺嘌呤的原始DNA链作为模板DNA分子以产生具有甲基化腺嘌呤的新链时，可能会产生这种类型的双链分子。

[0287] 两条链都可以是未甲基化的(分子IV)。这种类型的双链分子可能是由于两条不具有甲基化腺嘌呤的原始DNA链的重新退火而产生的。

[0288] 片段化过程可以包含酶消化、雾化、流体动力剪切和超声处理等。此类全基因组经扩增的DNA产物可能主要在A位点被甲基化。对具有mA的此DNA进行单分子实时测序以产生mA数据集。

[0289] 对于uA数据集,使用单分子实时测序对262,608个中值长度为964bp的分子进行了测序。中值子读段深度为103x。在这些子读段中,48%可以使用BWA比对软件与人类参考基因组进行比对(Li H等人《生物信息学》2009;25:1754-60)。作为实例,可以采用Sequel II系统(太平洋生物科学公司)来进行单分子实时测序。使用SMRTbell快速模板制备试剂盒2.0(太平洋生物科学公司)对片段化DNA分子进行单分子实时(SMRT)测序模板构建。用SMRT Link v8.0软件(太平洋生物科学公司)计算了测序引物退火和聚合酶结合条件。简而言之,将测序引物v2退火到测序模板,并且然后使用Sequel II结合和内部控制试剂盒2.0(太平洋生物科学公司)将聚合酶与模板结合。在Sequel II SMRT Cell 8M上执行测序。用Sequel II测序试剂盒2.0(太平洋生物科学公司)在Sequel II系统上收集了30个小时的测序影像。

[0290] 对于mA数据集,使用单分子实时测序对804,469个中值长度为826bp的分子进行了测序。中值子读段深度为34x。在这些子读段中,27%可以使用BWA比对软件与人类参考基因组进行比对(Li H等人《生物信息学》2009;25:1754-60)。

[0291] 在一个实施例中,以链特异性方式分析了包含但不限于IPD和PW的动力学特性。对于从沃森链得出的测序结果,使用从uA数据集中随机选择的644,318个不具有甲基化的A位点和从mA数据集中随机选择的718,586个具有甲基化的A位点来构成训练数据集。使用此类训练数据集来建立用于区分甲基化腺嘌呤与未甲基化腺嘌呤的分类模型和/或阈值。测试数据集由639,702个不具有甲基化的A位点和723,320个具有甲基化的A位点构成。此类测试数据集用于验证从训练数据集推导出的模型/阈值的性能。

[0292] 分析了源自沃森链的测序结果。图32A示出了uA和mA数据集的训练数据集的脉冲间持续时间(IPD)值。对于训练数据集,观察到mA数据集中的经过测序的A位点的IPD值(中值:1.09;范围:0-9.52)比uA数据集中的IPD值(中值:0.20;范围:0-9.52)高(P值<0.0001;曼-惠特尼U检验(Mann Whitney U test))。

[0293] 图32B示出了uA和mA数据集的测试数据集的IPD。当研究测试数据集中经过测序的A位点的IPD值时,观察到mA数据集中的IPD值比uA数据集中的IPD值高(中值1.10对0.19;P值<0.0001;曼-惠特尼U检验)。

[0294] 图32C示出了使用IPD截止值的ROC曲线下的区域。真阳性率在y轴上,假阳性率在x轴上。使用对应的IPD值区分具有和不具有甲基化的模板DNA分子中的经过测序的A碱基时,AUC下的面积对于训练数据集和测试数据集均为0.86。

[0295] 除了来自沃森链的结果之后,还分析了源自克里克链的测序结果。图33A示出了uA和mA数据集的训练数据集的IPD值。对于训练数据集,观察到mA数据集中的经过测序的A位点的IPD值(中值:1.10;范围:0-9.52)比uA数据集中的IPD值(中值:0.19;范围:0-9.52)高(P值<0.0001;曼-惠特尼U检验)。

[0296] 图34B示出了uA和mA数据集的测试数据集的IPD值。与uA数据集相比,在测试数据集的mA数据集中也观察到了经过测序的A位点的更高IPD值(中值1.10对0.19;P值<0.0001;曼-惠特尼U检验)。

[0297] 图33C示出了ROC曲线下的区域。真阳性率在y轴上,假阳性率在x轴上。使用对应的

IPD值区分具有和不具有甲基化的模板DNA分子中的经过测序的A碱基时,ROC曲线(AUC)值下的面积对于训练数据集和测试数据集分别为0.86和0.87。

[0298] 图34示出了根据本发明的实施例的使用测量窗口对沃森链进行6mA测定的图示。这种测量窗口可以包含如IPD和PW以及附近的序列上下文等动力学特征。6mA的测定可以类似于甲基化CpG的测定来执行。

[0299] 图35示出了根据本发明的实施例的使用测量窗口对克里克链进行6mA测定的图示。这种测量窗口可以包含如IPD和PW以及附近的序列上下文等动力学特征。

[0300] 作为实例,使用了来自被分析的模板DNA中经过测序的A碱基每侧的10个碱基来构建测量窗口。根据本文公开的方法,使用包含IPD、PW和序列上下文的特征值来训练使用卷积神经网络(CNN)的模型。在其它实施例中,统计模型可以包含但不限于线性回归、逻辑回归、深度递归神经网络(例如,长短期记忆,LSTM)、贝叶斯分类器、隐马尔可夫模型(HMM)、线性判别分析(LDA)、k均值聚类、基于密度的带噪声应用空间聚类(DBSCAN)、随机森林算法和支持向量机(SVM)等。

[0301] 图36A和图36B示出了使用基于测量窗口的CNN模型对uA数据集与mA数据集之间的沃森链的经过测序的A碱基进行甲基化的测定概率。图36A示出了从训练数据集中学习到的CNN模型。作为实例,CNN模型使用了两个1D卷积层(每个层具有64个内核尺寸为4的过滤器,随后是ReLU(整流线性单元)层),随后是丢弃率为0.5的丢弃层。使用了池大小为2的最大池化层。随后,所述最大池化层流入两个1D卷积层(每个层具有128个内核尺寸为2的过滤器,随后是ReLU层),进一步使用丢弃率为0.5的退出层。使用了池大小为2的最大池化层。最后,具有10个神经元的完全连接层后接ReLU层,其中具有一个神经元的输出层后接sigmoid层,从而得出甲基化概率。可以对层、过滤器、内核尺寸的其它设置进行调整,例如,如本文针对其它甲基化(例如,CpG)所描述的。在关于沃森链的测序结果的此训练数据集中,使用了来自未甲基化和甲基化文库的644,318个和718,586个A碱基。

[0302] 基于CNN模型,对于沃森链相关数据,来自mA数据库的模板DNA分子中的经过测序的A碱基在训练数据集和测试数据集中都产生了比uA数据集种存在的A碱基高得多的甲基化概率(P值<0.0001;曼-惠特尼U检验)。对于训练数据集,uA数据集中A位点上甲基化的中值概率为0.13(四分位距,IQR:0.09-0.15),而mA数据集中的所述值为1.000(IQR:0.998-1.000)。

[0303] 图36A示出了针对测试数据集测定的甲基化概率。对于测试数据集,uA数据集中A位点上甲基化的中值概率为0.13(IQR:0.10-0.15),而mA数据集中的所述值为1.000(IQR:0.997-1.000)。图36A和36B示出了可以对基于测量窗口的CNN模型进行训练以检测测试数据集中的甲基化。

[0304] 图37是使用基于测量窗口的CNN模型对沃森链的经过测序的A碱基进行6mA检测的ROC曲线。真阳性率在y轴上,且假阳性率在x轴上。所述图示出了对于由沃森链测序结果组成的训练数据集和测试数据集,使用CNN模型区分具有和不具有甲基化的经过测序的A位点的AUC值分别为0.94和0.93。这表明使用沃森链的数据用本文的公开内容来测定A位点上的甲基化状态是可行的。如果使用0.5的测定甲基化概率作为截止值,则6mA检测可以达到99.3%的特异性和82.6%的灵敏度。图37示出了可以使用基于测量窗口的CNN模型以高特异性和灵敏度来检测6mA。可以将模型的准确性与仅使用IPD度量的技术进行比较。

[0305] 图38示出了基于IPD度量的6mA检测与基于测量窗口的6mA检测之间的性能比较。灵敏度绘制在y轴上,并且特异性绘制在x轴上。图38示出了使用根据本文的公开内容的基于测量窗口的6mA分类的性能(AUC:0.94)优于仅使用IPD度量的常规方法(AUC:0.87)(P值<0.0001;德龙氏试验(DeLong's test))。基于测量窗口的CNN模型胜过基于IPD度量的检测。

[0306] 图39A和39B示出了使用基于测量窗口的CNN模型对uA数据集与mA数据集之间的克里克链的那些经过测序的A碱基进行甲基化的测定概率。图39A示出了训练数据集,并且图39B示出了测试数据集。两个图都在y轴上绘制了甲基化概率。图39A和39B示出了基于CNN模型,对于克里克链相关数据,来自mA数据库的模板DNA分子中的经过测序的A碱基在训练数据集和测试数据集中都产生了比uA数据集中存在的A碱基高得多的甲基化概率(P值<0.0001;曼-惠特尼U检验)。

[0307] 图40示出了使用基于测量窗口的CNN模型对克里克链的经过测序的A碱基进行的6mA检测的性能。真阳性率在y轴上。假阳性率在x轴上。图40示出了对于由克里克链测序结果组成的训练数据集和测试数据集,使用CNN模型区分具有和不具有甲基化的经过测序的A位点的AUC值分别为0.95和0.94。还显示出使用本文公开的CNN方法的性能(AUC:0.94)优于仅使用IPD度量的方法的性能(0.87)(P值<0.0001)。结果表明,使用克里克链的数据用本文的公开内容来测定A位点上的甲基化状态是可行的。如果使用0.5的测定甲基化概率作为截止值,则6mA检测可以达到99.3%的特异性和83.0%的灵敏度。图40示出了可以使用基于测量窗口的CNN模型以高特异性和灵敏度来检测6mA。

[0308] 图41示出了包含沃森链和克里克链的分子中的A碱基的甲基化状态的实例。白点表示未甲基化腺嘌呤。黑点表示甲基化腺嘌呤。带点的水平线表示双链DNA分子的链。分子1示出了沃森链和克里克链均被测定为A碱基是未甲基化的。分子2示出了沃森链几乎全部是未甲基化的,而克里克链几乎全部是甲基化的。分子3示出了沃森链和克里克链均被测定为A碱基几乎全部是甲基化的。

[0309] 2.使用选择性数据集的增强训练

[0310] 如图36A、36B、39A和39B所示,在mA数据集中,跨模板DNA分子中的经过测序的A碱基的甲基化概率呈双峰分布。换句话说,mA数据集中存在一些具有uA信号的分子。mA数据集中存在完全未甲基化分子和半甲基化分子进一步证明了这一点(图41)。一个可能的原因可能是DNA模板中具有uA的分子在全基因组扩增后仍在mA数据集中占相当大的比例,因为具有6mA的分子会导致全基因组扩增步骤期间DNA扩增效率降低。这一解释得到了以下事实的支持:在相同的扩增条件下,用6mA扩增的1ng基因组DNA将仅产生10ng DNA产物,而用未甲基化A扩增的1ng基因组DNA将产生100ng DNA产物。因此,对于mA数据集,腺嘌呤通常是未甲基化的原始模板DNA分子(例如0.051%)(Xiao CL等人《分子细胞(Mol Cell.)》2018;71:306-318)将大约占腺嘌呤总量的10%。

[0311] 在一个实施例中,当试图对用于区分mA与uA的CNN模型进行训练时,将选择性地使用在mA数据集中具有相对较高的IPD值的那些A碱基,以便减少uA数据对训练用于mA检测的模型的影响。只能使用IPD值大于某个截止值的A碱基。截止值可以对应于百分位。在一个实施例中,将使用mA数据集中IPD值大于第10百分位处的值的那些A碱基。在一些实施例中,将使用IPD值大于第1、第5、第15、第20、第30、第40、第50、第60、第70、第80、第90或第95百分位处的值的那些A。百分位可以基于来自一个参考样本或多个参考样本中的所有核酸分子的

数据。

[0312] 图42示出了通过选择性地使用mA数据集中IPD值大于其第10百分位的A碱基来增强训练的性能。图42示出了y轴上的真阳性率和x轴上的假阳性率。所述图示出了通过使用mA数据集中IPD值大于第10百分位的A碱基来训练CNN模型，在区分mA与uA碱基时，AUC将增加到0.98，这优于在训练前没有根据IPD值进行选择的情况下由数据训练的模型(AUC：0.94)。这表明，使用IPD值选择mA位点以创建训练数据集将有助于提高辨别能力。

[0313] 为了进一步证实mA数据集中存在具有uA碱基的分子，假设mA数据集中uA的百分比将在具有较多子读段的孔中富集，因为与不具有6mA的分子相比，分子中存在的6mA将在产生新链时减缓聚合酶延伸。

[0314] 图43示出了mA数据集中未甲基化腺嘌呤的百分比对每个孔中子读段的数量的图。y轴示出了mA数据集中uA的百分比。x轴示出了每个孔中子读段的数量。在去除IPD值低于第10百分位的A位点后，使用通过使用mA位点训练的增强模型重新分析测试数据集。观察到随着每孔子读段数量的增加，uA逐渐增加(即从14.6增加到55.05％)，包含从每个测序孔1到10个子读段增加到每孔10到20个子读段、到每孔40到50个子读段、每孔60到70个子读段以及大于70个子读段。因此，具有大量子读段的孔往往具有较低的mA。A的甲基化可以延迟测序反应的进展。因此，具有较高子读段深度的测序孔将更有可能相对于A是未甲基化的。可以使用与分子相关联的子读段数量的截止值来利用此行为以检测未甲基化分子，例如，超过70个子读段可以被鉴定为多数未甲基化。

[0315] 图44示出了测试数据集中双链DNA分子的沃森链与克里克链之间的甲基腺嘌呤谱式。A的甲基化是不对称的，并且因此两条链之间的行为是不同的。大多数分子由于mA的掺入被甲基化，其中一些残留的A是未甲基化的。y轴示出了克里克链的甲基腺嘌呤水平。x轴示出沃森链的甲基腺嘌呤水平。每个点表示双链分子。使用通过选定的mA位点训练的增强模型，双链分子可以根据如下每条链的甲基化水平被分为不同的组：

[0316] (a)对于双链DNA分子，沃森链和克里克链的甲基腺嘌呤水平均大于0.8。对于腺嘌呤位点，这种双链分子被定义为完全甲基化的分子(图44，区域A)。链的甲基腺嘌呤水平被定义为在所述链的全部A位点中被测定为甲基化的A位点的百分比。

[0317] (b)对于双链DNA分子，一条链的甲基腺嘌呤水平大于0.8，而另一条链的甲基腺嘌呤水平小于0.2。对于腺嘌呤位点，这种双链分子被定义为半甲基化分子(图44，区域B1和B2)。

[0318] (c)对于双链DNA分子，沃森链和克里克链的甲基腺嘌呤水平均小于0.2。对于腺嘌呤位点，这种双链分子被定义为完全未甲基化的分子(图44，区域C)。

[0319] (d)对于双链DNA分子，沃森链和克里克链的甲基腺嘌呤水平不属于a组、b组和c组。对于腺嘌呤位点，这种双链分子被定义为具有交错甲基化谱式的分子(图44，区域D)。交错甲基化谱式被定义为存在于DNA链中的甲基化和未甲基化腺嘌呤的混合物。

[0320] 在一些其它实施例中，用于定义未甲基化链的甲基腺嘌呤水平的截止值可以是但不限于小于0.01、0.05、0.1、0.2、0.3、0.4和0.5。用于定义甲基化链的甲基腺嘌呤水平的截止值可以是但不限于大于0.5、0.6、0.7、0.8、0.9、0.95和0.99。

[0321] 图45是表格，其示出了训练数据集和测试数据集中的完全未甲基化的分子、半甲基化分子、完全甲基化的分子和具有交错甲基腺嘌呤谱式的分子的百分比。测试数据集中

的分子可以被归类为关于腺嘌呤位点的完全未甲基化的分子(7.0%)、半甲基化分子(9.8%)、完全甲基化的分子(79.4%)和具有交错甲基腺嘌呤谱式的分子(3.7%)。这些结果与训练数据集中示出的结果相当,对于所述训练数据集,存在关于腺嘌呤位点的完全未甲基化的分子(7.0%)、半甲基化分子(10.0%)、完全甲基化的分子(79.4%)和具有交错甲基腺嘌呤谱式的分子(3.6%)。

[0322] 图46展示了具有关于腺嘌呤位点的完全未甲基化分子的分子、半甲基化分子、完全甲基化的分子以及具有交错的甲基腺嘌呤谱式的分子的代表性实例。白点表示未甲基化腺嘌呤。黑点表示甲基化腺嘌呤。带点的水平线表示双链DNA分子的链。

[0323] 在实施例中,可以通过增加用于训练CNN模型的6mA碱基的纯度来提高区分甲基化与未甲基化腺嘌呤的性能。为此,可以增加DNA扩增反应的持续时间,使得增加的新产生的DNA产物可以稀释由原始DNA模板贡献的未甲基化腺嘌呤的作用。在其它实施例中,可以在用6mA进行DNA扩增期间掺入生物素化的碱基。可以使用链霉亲和素包被的磁珠将用6mA新产生的DNA产物拉下并富集。

[0324] 3.6mA甲基化谱的用途

[0325] DNA6mA修饰存在于细菌、古细菌、原生生物和真菌的基因组中。(Didier W等人《自然评论:微生物学(Nat Rev Micorbiol.)》2009;4:183-192)。另据报道,人类基因组中存在6mA,占总腺嘌呤的0.051%(Xiao CL等人《分子细胞(Mol Cell.)》2018;71:306-318)。考虑到人类基因组中6mA的含量低,在一个实施例中,可以通过在全基因组扩增步骤中调整dNTP混合物中6mA的比例(N表示未经修饰的A、C、G和T)来创建训练数据集。例如,可以使用1:10、1:100、1:1000、1:10000、1:100000或1:1000000的6mA与dNTP的比率。在另一个实施例中,腺嘌呤DNA甲基转移酶M.EcoGII可以用于创建6mA训练数据集。

[0326] 胃癌和肝癌组织中6mA的含量较低,并且这种6mA的下调与肿瘤发生的增加有关(Xiao CL等人《分子细胞》2018;71:306-318)。另一方面,据报道,胶质母细胞瘤中存在更高水平的6mA(Xie等人《细胞(Cell)》2018;175:1228-1243)。因此,本文公开的用于6mA的方法对于研究癌症基因组学将是有用的(Xiao CL等人《分子细胞》2018;71:306-318;Xie等人《细胞》2018;175:1228-1243)。另外,发现了6mA在哺乳动物的线粒体DNA中更为普遍和丰富,这显示出与低氧有关(Hao Z等人《分子细胞》2020;doi:10.1016/j.molcel.2020.02.018)。因此,本公开中用于6mA检测的方法对于研究如怀孕、癌症和自身免疫疾病等不同临床条件下的线粒体应激反应将是有用的。

[0327] IV.结果和应用

[0328] A.检测甲基化

[0329] 对于不同的生物样本和基因组区域,使用上述方法检测CpG位点处的甲基化。作为实例,使用单分子实时测序用孕妇血浆中的细胞游离DNA进行的甲基化测定通过使用亚硫酸氢盐测序的甲基化测定进行验证。甲基化结果可以用于不同的应用,包含测定拷贝数和诊断疾病。以下所述的方法不限于CpG位点,并且还可以应用于本文所述的任何修饰。

[0330] 1.胎盘组织中长DNA分子的甲基化检测

[0331] 单分子实时测序可以对长度为千碱基的DNA分子进行测序(Nattestad等人,2018)。使用此处描述的发明对CpG位点的甲基化状态进行解密将允许通过协同利用单分子实时测序的长读段信息来推断甲基化状态的单倍型信息。为了证明推断长读段甲基化状态

及其单倍型信息的可行性，对具有被28,913,838个子读段覆盖的478,739个分子的胎盘组织DNA进行了测序。有7个分子的长度大于5kb。每个分子平均被3个子读段覆盖。

[0332] 图47示出了沿长度为6,265bp的长DNA分子(即单倍型模块)的甲基化状态，所述长DNA分子在ZMW孔编号为m54276_180626_162240/40763503的ZMW中进行测序并且被定位到人类基因组中chr1:113246546-113252811的基因组位置。“-”表示非CpG核苷酸；“U”表示CpG位点处的未甲基化状态；并且“M”表示CpG位点处的甲基化状态。用黄色突出显示的区域4710指示CpG岛区域，所述区域已知通常是未甲基化的(图47)。推导出所述CpG岛中的大多数CpG位点是未甲基化的(96％)。相反，推导出CpG岛以外的CpG位点的75％是未甲基化的。这些结果表明，CpG岛以外的甲基化水平(例如CpG岛岸/岛架)高于CpG岛的甲基化水平。在所述CpG岛以外的区域中以单倍型布置的甲基化和未甲基化状态的混合物将指示甲基化谱式的可变性。此类观察总体上符合当前的理解(Zhang等人，2015；Feinberg和Irizarry，2010)。因此，本公开内容使得能够沿包含甲基化和未甲基化状态的长分子判读不同的甲基化状态，这意味着可以对甲基化状态的单倍型信息进行定相。单倍型信息是指一段连续的DNA上CpG位点的甲基化状态的连接。

[0333] 在一个实施例中，可以在此使用这种方法来分析沿单倍型的甲基化状态，以检测和分析印记区域。印记区域经受表观遗传调控，所述表观遗传调控以亲源的方式导致甲基化状态。例如，一个重要的印记区域位于人类染色体11p15.5上，并含有印记基因IGF2、H19和CDKN1C(P57^{kip2})，所述印记基因是胎儿生长的强调节因子(Brioude等人，《自然评论：内分泌学(Nat Rev Endocrinol.)》2018；14:229-249)。印记区域的遗传和表观遗传异常可能与疾病有关。贝-维综合征(BWS)是一种过度生长综合征，其中患者在儿童早期常常表现出巨舌症、腹壁缺损、偏侧发育过度、腹腔器官增大和胚胎肿瘤的风险增加。BWS被认为是由11p15.5区域内的遗传或表观遗传缺陷引起的(Brioude等人，《自然评论：内分泌学》2018；14:229-249)。位于H19与IGF2之间的被称为ICR1的区域(印记控制区1)在父本等位基因上有差异甲基化。ICR1指导IGF2亲源特异性表现。因此，ICR1中的遗传和表观遗传异常将导致IGF2的异常表达，这是引起BWS的可能原因之一。因此，沿印记区域检测甲基化状态将具有临床意义。

[0334] 从管理当前报道的印记基因的公共数据库(http://www.geneimprint.org/)中下载了92个印记基因的数据。这些印记基因上游和下游5kb的区域用于进一步分析。在这些区域中，有160个CpG岛与这些印记基因相关联。从胎盘样本中获得了324,248个环状一致性序列。去除低质量和与CpG岛重叠的区域短(例如小于相关CpG岛的长度的50％)的环状一致性序列后，获得了与对应于8个印记基因的9个CpG岛重叠的9个环状一致性序列。

[0335] 图48是表格，其显示9个DNA分子通过单分子实时测序进行测序并与印记区域重叠，所述印记区域包含H19、WT1-AS、WT1、DLK1、MEG3、ATP10A、LRRTM1和MAGI2。第6列包含与涉及印记区域的CpG岛重叠的DNA段。“U”表示CpG上下文中的未甲基化胞嘧啶；“M”表示CpG上下文中的甲基化胞嘧啶。“*”表示测序结果中未覆盖的CpG位点；“-”表示来自非CpG位点的核苷酸；如果分子与单核苷酸多态性(SNP)重叠，则在括号中指出基因型。第7列指示整个分子的甲基化状态。如果根据本公开中呈现的实施例示出了大部分CpG位点(例如大于50％)被甲基化，则分子可以被称为甲基化，否则将被称为未甲基化。

[0336] 在9个DNA分子中，5个DNA分子(55.6％)被称为甲基化，这与50％的DNA分子将被甲

基化的预期并没有显著偏离。如图48的表的第6列所示，大多数CpG位点以一致的方式被示出为甲基化或未甲基化，即，作为甲基化单倍型。一个实施例是，如果根据本公开中呈现的实施例示出了大部分CpG位点(例如大于50％)被甲基化，则分子可以被称为甲基化，否则将被称为未甲基化。可以使用用于测定分子是甲基化的还是未甲基化的其它截止值，例如但不限于分子中至少10％、20％、30％、40％、50％、60％、70％、80％、90％和100％的CpG位点经分析被视为甲基化。

[0337] 在另一个实施例中，可以使用同时包括至少一个SNP和至少一个CpG位点分析的分子来确定区域是否可能与印记区域相关联或者已知印记基因是否异常(例如印记丢失)。出于说明的目的，图49示出了来自印记区域的第一分子携带等位基因“A”；并且来自所述印记区域的第二分子携带等位基因“G”。假设印记区域是父本印记的，则来自母本单倍型的第一分子是完全未甲基化的；并且来自父本单倍型的第二分子是完全甲基化的。在一个实施例中，这种假设将提供甲基化状态的基本事实，从而允许测试根据本公开中呈现的实施例的碱基修饰检测的性能。

[0338] 图49示出了用于测定印记区域中的甲基化谱式的实例。提取生物样本中的DNA并与发夹衔接子连接以形成环状DNA分子。关于那些环状DNA分子的序列信息和碱基修饰(例如CpG位点处的甲基化状态)是未知的。那些环状DNA分子经过了单分子实时测序。在将子读段定位到参考基因组之后确定了源自那些环状DNA分子的每个子读段中碱基的IPD、PW和序列上下文。另外，测定了那些分子的基因型。根据本公开中呈现的实施例将与CG位点相关联的测量窗口中的IPD、PW和序列上下文与参考动力学谱式进行比较，以测定每个CpG的甲基化状态。如果两个具有不同等位基因的分子以一个完全未甲基化而另一个完全甲基化的方式示出不同的甲基化谱式，则与这两个分子相关联的基因组区域将是印记区域。在一个实施例中，如果此类基因组区域碰巧是已知的印记区域，例如，如图49中所展示的，则这两个分子的甲基化谱式与正常情况下预期的甲基化谱式(即基本事实)一致。这可以表明根据本公开中呈现的实施例的用于甲基化状态分类的方法的准确性。在一个实施例中，根据本公开中呈现的实施例测得的甲基化谱式与预期的甲基化谱式之间的衍生将指示印记异常，例如，印记丢失。

[0339] 图50示出了用于测定印记区域中的甲基化谱式的实例。在一个实施例中，可以通过分析跨某个谱系树的所述区域的甲基化谱式来进一步测定印记谱式。例如，可以对跨父本基因组、母本基因组和后代的甲基化谱式和等位基因信息执行分析。这种谱系树可以进一步包含父本或母本祖父的基因组、父本或母本祖母的基因组或其它相关基因组。在另一个实施例中，此类分析可以扩展到特定群体中的三人组家庭(母亲、父亲和孩子)数据集，例如，从而根据本文中呈现的实施例获得每个个体的甲基化和基因型信息。

[0340] 如分类后所示，可以测定基因型(框中的等位基因)和甲基化状态。对于每个分子，可以在每个位点提供甲基化谱式(例如，全部甲基化或全部未甲基化)，以便鉴定分子是从哪个亲本继承的。或者，可以测定甲基化密度，并且一个或多个截止值可以对分子是高甲基化的(例如，>80％或其它％并且来自一个亲本)还是低甲基化的(例如，<20％或其它％并且来自另一个亲本)进行分类。

[0341] 2.cfDNA分子的甲基化检测

[0342] 作为另一个实例，细胞游离DNA(cfDNA)甲基化也已被越来越多地视为无创产前检

测的重要分子信号。例如,已经表明,来自携带组织特异性甲基化的区域的cfDNA分子可以用于测定孕妇血浆中如中性粒细胞、T细胞、B细胞、肝脏、胎盘等不同组织的贡献比例(Sun等人,2015)。还证明了使用孕妇血浆DNA甲基化来检测21三体综合症的可行性(Lun等人,2013)。母本血浆中的cfDNA分子以166bp的中值长度被片段化,这比以大约500bp长度进行人工片段化的大肠杆菌DNA要短得多。据报道,cfDNA是非随机片段化的,例如,血浆DNA的末端基序与比如来自胎盘的组织起源有关。细胞游离DNA的此类特性特征使序列上下文与人工片段化的大肠杆菌DNA截然不同。因此,目前尚不清楚此类聚合酶动力学是否能够定量地推导通常细胞游离DNA分子的甲基化水平。例如通过使用从上述组织DNA分子训练的甲基化预测模型,本专利申请中的公开内容将适用于但不限于孕妇血浆中的细胞游离DNA甲基化分析。

[0343] 使用单分子实时测序,对具有男性胎儿的孕妇的六个血浆DNA样本进行了测序,中值为30,738,399个子读段(范围:1,431,215-105,835,846),对应于中值为111,834个CCS(范围:61,010-503,582)。每个血浆DNA的测序中值为262次(范围:173-320)。数据集由通过Sequel I测序试剂盒3.0制备的DNA产生。

[0344] 为了评估cfDNA分子的甲基化检测,使用了亚硫酸氢盐测序(Jiang等人,2014)来分析上述6个孕妇血浆DNA样本的甲基化。获得了中值为6,600万的配对末端读段(5,800-8,200万个配对末端读段)。发现总体甲基化中值为69.6%(67.1%-72.0%)。

[0345] 图51示出了通过新方法和常规亚硫酸氢盐测序推导的甲基化水平的比较。y轴是根据本专利申请中呈现的实施例预测的甲基化水平。x轴是通过亚硫酸氢盐测序推导的甲基化水平。分析了通过单分子实时测序产生的血浆DNA结果的中值为314,675的CpG位点(范围:144,546-1,382,568)。预计为甲基化的CpG位点的中值比例为64.7%(范围:60.8%-68.5%),这似乎与通过亚硫酸氢盐测序推导的结果相当。如图51所示,通过单分子实时测序推导的总体甲基化水平与当前甲基化预测方法和亚硫酸氢盐测序之间存在良好的相关性(r:0.96,p值=0.0023)。

[0346] 由于亚硫酸氢盐测序的深度浅,因此推导人类基因组中每个CpG的甲基化水平(即被甲基化的经过测序的CpG的分率)可能不太可靠。替代地,通过聚合覆盖基因组区域的CpG位点的读取信号来计算一些具有多个CpG位点的区域的甲基化水平,在所述基因组区域中,任何两个连续的CpG位点都在50nt内并且CpG位点的数量至少为10。经过测序的胞嘧啶在区域中跨CpG位点的经过测序的胞嘧啶和胸腺嘧啶的总和中的百分比指示了所述区域的甲基化水平。根据区域甲基化水平将区域分为不同的组。随着甲基化水平的增加,从先前的训练数据集(即组织DNA)中学习的模型所预测的甲基化概率相应地增加(图52A)。这些结果进一步表明了使用单分子实时测序来预测孕妇的cfDNA分子的甲基化状态的可行性和有效性。图52B示出了根据本公开中呈现的实施例使用单分子实时测序估计的10Mb基因组窗口中的甲基化水平与通过亚硫酸氢盐测序估计的结果具有良好的相关性(r=0.74;p值<0.0001)。

[0347] 图53示出了通过单分子实时测序测得的孕妇母本血浆中Y染色体的基因组呈现(GR)与通过BS-seq测得的结果具有良好的相关性(r=0.97;p值=0.007)。这些结果表明,单分子实时测序还可以准确定量来源于如胎盘等非造血组织的DNA分子,所述非造血组织的贡献DNA通常占少数。换句话说,本公开证明了在测序之前同时分析天然分子的拷贝数异常和甲基化状态而无需任何碱基转化和扩增的可行性。

[0348] 3.基于CpG区块的方法

[0349] 一些实施例可以对含有多个CpG位点的多个基因组区域执行甲基化分析，所述多个CpG位点例如但不限于2个、3个、4个、5个、10个、20个、30个、40个、50个、100个CpG位点等。这种基因组区域的长度可以是例如但不限于50nt、100nt、200nt、300nt和500nt等。此区域中CpG位点之间的距离可以是例如但不限于10nt、20nt、30nt、40nt、50nt、100nt、200nt、300nt等。在一个实施例中，可以在50nt内合并任何两个连续的CpG位点以形成CpG区块，使得此区块中的CpG位点的数量超过10个。在这种基于区块的方法中，可以将多个区域组合到一个表示为单个矩阵的窗口中，从而有效地将这些区域一起处理。

[0350] 作为一个实例，如图54所示，使用与CpG区块相关联的所有子读段的动力学进行甲基化分析。将所述区块中每个CpG侧翼的上游和下游10nt的投影IPD谱相对于CpG位点进行人工比对，以计算平均IPD谱(图54)。词语“投影”意指已将子读段动力学信号与讨论中的每个对应的CpG位点进行比对。使用CpG区块的平均IPD谱来训练模型(例如，使用人工神经网络，简称为ANN)，以鉴定每个区块的甲基化状态。ANN分析包含一个输入层、两个隐藏层和一个输出层。每个CpG区块以将输入到ANN的21个IPD值的特征向量为特征。第一隐藏层包含10个以ReLu为激活函数的神经元。第二隐藏层包含5个以ReLu为激活函数的神经元。最后，输出层包含1个以Sigmoid为激活函数的神经元，其将输出甲基化概率。示出甲基化概率>0.5的CpG位点被视为甲基化，否则被视为未甲基化。可以使用平均IPD谱分析整个分子的甲基化状态。如果一定数量的高于阈值(例如，0、1、2、3等)的位点被甲基化或者如果分子具有一定的甲基化密度，则整个分子可以被认为是甲基化的。

[0351] 未甲基化和甲基化文库中有9,678个和9,020个CpG区块，所述区块中的每一个至少含有10个CpG位点。那些CpG区块覆盖了未甲基化和甲基化文库的176,048个和162,943个CpG位点。如图55A和图55B所示，在预测训练数据集和测试数据集的甲基化状态时，可以获得大于90％的总体准确性。然而，此类依赖于CpG区块的实施例将大大减少能够被评估的CpG的数量。根据定义，对CpG位点的最少数量的要求会将甲基化分析限制在某些特定的基因组区域(例如，优先分析CpG岛)。

[0352] B.起源或病症的确定

[0353] 甲基化谱可以用于检测组织起源或确定病症的分类。甲基化谱分析可以与其它临床数据结合使用，包含成像、常规血液检查和其它医学诊断信息。甲基化谱可以使用本文所述的任何方法测定。

[0354] 1.拷贝数异常的测定

[0355] 本节示出了SMRT对于测定拷贝数是准确的，并且因此可以同时分析甲基化谱和拷贝数谱。

[0356] 已经证明了可以通过肿瘤组织的测序来揭示拷贝数异常(Chan(2013))。此处，示出了可以通过使用单分子实时测序对肿瘤组织进行测序来鉴定与癌症相关联的拷贝数异常。例如，对于案例TBR3033，分别获得了肿瘤DNA和其配对的邻近非肿瘤肝组织DNA的589,435个和1,495,225个一致性序列(用于构建每个一致性序列的子读段的最低要求是5)。数据集由通过Sequel II测序试剂盒1.0制备的DNA产生。在一个实施例中，基因组在计算机模拟中被分为2Mb窗口。计算了映射到每个窗口的一致性序列的百分比，从而得到2Mb分辨率的基因组呈现(GR)。GR可由一个位置的许多读段来确定，这些读段由全基因组的总读段来

进行标准化。

[0357] 图56A示出了使用单分子实时测序的肿瘤与其配对的邻近非肿瘤组织DNA之间的GR比率。y轴示出了肿瘤DNA与配对的邻近正常组织DNA之间的拷贝数比率，并且x轴示出了包含染色体1到22的每个2Mb窗口的基因组分类指数。对于此图，GR比率高于所有2Mb窗口的第95百分位的区域被归类为具有拷贝数增加，而GR比率低于所有2Mb窗口的第5百分位的区域被归类为具有拷贝数缺失。观察到，染色体13具有拷贝数缺失，而染色体20具有拷贝数增加。这种增加和缺失是正确的结果。

[0358] 图56B示出了使用亚硫酸氢盐测序的肿瘤与其配对的邻近非肿瘤组织之间的GR比率。y轴示出了肿瘤DNA与配对的邻近正常组织DNA之间的拷贝数比率，并且x轴示出了包含染色体1到22的每个2Mb窗口的基因组分类指数。图56A中通过单分子实时测序鉴定的拷贝数变化在图56B中匹配的亚硫酸氢盐测序结果中得到验证。

[0359] 对于案例TBR3032，分别获得了肿瘤DNA和其配对的邻近非肿瘤组织DNA的413,982个和2,396,054个一致性序列(用于构建每个一致性序列的子读段的最低要求是5)。在一个实施例中，基因组在计算机模拟中被分为2Mb窗口。计算了映射到每个窗口的一致性序列的百分比，即2Mb基因组表示(GR)。

[0360] 图57A示出了使用单分子实时测序的肿瘤与其配对的邻近非肿瘤组织DNA之间的GR比率。y轴示出了肿瘤DNA与配对的邻近正常组织DNA之间的拷贝数比率，并且x轴示出了包含染色体1到22的每个2Mb窗口的基因组分类指数。对于此图，GR比率高于所有2Mb窗口的第95百分位的区域被归类为具有拷贝数增加，而GR比率低于所有2Mb窗口的第5百分位的区域被归类为具有拷贝数缺失。观察到，染色体4、6、11、13、16和17具有拷贝数缺失，而染色体5和7具有拷贝数增加。

[0361] 图57B示出了使用亚硫酸氢盐测序的肿瘤与其配对的邻近非肿瘤组织之间的GR比率。y轴示出了肿瘤DNA与配对的邻近正常组织DNA之间的拷贝数比率，并且x轴示出了包含染色体1到22的每个2Mb窗口的基因组分类指数。图57A中通过单分子实时测序鉴定的拷贝数变化在图57B中匹配的亚硫酸氢盐测序结果中得到验证。

[0362] 因此，可以同时分析甲基化谱和拷贝数谱。在此例示中，由于肿瘤组织的肿瘤纯度通常不总是100％，因此扩增的区域将相对增加肿瘤DNA贡献，而缺失的区域将相对减少肿瘤DNA贡献。因为肿瘤基因组的特征是整体低甲基化，所以与缺失的区域相比，扩增的区域将进一步降低甲基化水平。作为说明，对于案例TBR3033，使用本发明测量的染色体22(拷贝数增益)的甲基化水平为48.2％，这低于染色体3(拷贝数损失)的甲基化水平(甲基化水平：54.0％)。对于案例TBR3032，使用本发明测量的染色体5p臂(拷贝数增益)的甲基化水平为46.5％，这低于染色体5q臂(拷贝数损失)的甲基化水平(甲基化水平:54.9％)。

[0363] 2.孕妇血浆DNA组织映射

[0364] 如图58所示，我们推断，甲基化分析的准确性将使得能够将孕妇的血浆DNA甲基化谱与不同参考组织(例如，肝脏、中性粒细胞、淋巴细胞、胎盘、T细胞、B细胞、心脏、大脑等)的甲基化谱进行比较。因此，可以使用以下程序来推导来自不同细胞类型的孕妇血浆DNA库中的DNA贡献。根据本公开中呈现的实施例测定的DNA混合物(例如血浆DNA)的CpG甲基化水平被记录在载体(X)中，并且跨不同组织的检索到的参考甲基化水平被记录在基质(M)中，所述基质可以通过但不限于亚硫酸氢盐测序来定量。不同组织对DNA混合物的比例贡献(p)

可以通过但不限于二次规划来求解。此处，使用数学方程式来说明不同器官对所分析的DNA混合物的比例贡献的推导。DNA混合物中不同位点的甲基化密度与跨不同组织的对应位点的甲基化密度之间的数学关系可以表示为：

[0365] $\bar{X}_i = \sum_k (P_k \times M_{ik})$，

[0366] 其中$\bar{X}_i$表示DNA混合物中CpG位点i的甲基化密度；P_k表示细胞类型k对DNA混合物的比例贡献；M_{ik}表示细胞类型k中CpG位点i的甲基化密度。当位点的数量等于或大于器官的数量时，可以测定个体P_k的值。为了提高信息量，对跨所有参考组织类型的甲基化水平显示出较小可变性的CpG位点被丢弃。在一个实施例中，使用了一组特定的CpG位点来执行分析。例如，那些CpG位点的特征是跨不同组织的甲基化水平的变异系数(CV)大于30％并且组织中最大甲基化水平与最小甲基化水平之间的差异超过25％。在一些其它实施例中，还可以使用5％、10％、20％、30％、40％、50％、60％、80％、90％、100％、110％、200％、300％等的CV；并且可以使用超过5％、10％、15％、20％、25％、30％、40％、50％、60％、70％、80％、90％、100％等的组织中最大甲基化水平与最小甲基化水平之间的差异。

[0367] 算法中可以包含另外的标准以提高准确性。例如，所有细胞类型的聚合贡献将被限制为100％，即

[0368] $\sum_k P_k = 100\%$。

[0369] 此外，所有器官的贡献都必须是非负的：

[0370] $P_k \geq 0, \ \forall k$

[0371] 由于生物学变异，观察到的整体甲基化谱式可能与从组织的甲基化推导的甲基化谱式不完全相同。在这种情况下，将需要进行数学分析来测定各个组织最可能的比例贡献。就这一点而言，DNA中观察到的甲基化谱式与从组织推导的甲基化谱式之间的差异用W表示：

[0372] $$W = \bar{X}_i - \sum_k (P_k \times M_{ik})$$

[0373] 可以通过最小化W来测定每个P_k的最有可能的值，所述W是观察到的甲基化谱式与推导的甲基化谱式之间的差异。可以使用数学算法来求解此方程式，例如通过但不限于使用二次规划、线性/非线性回归、期望最大化(EM)算法、最大似然算法、最大后验估计和最小二乘法。

[0374] 如图59所示，使用图58中呈现的血浆DNA组织映射方法，观察到胎盘DNA对携带男性胎儿的孕妇的母本血浆的贡献与通过Y染色体读段估计的胎儿DNA分率具有良好的相关性。此结果表明了使用动力学追踪孕妇血浆DNA的起源组织的可行性。

[0375] 3.区域甲基化水平定量

[0376] 本节描述了用于测定选定基因组区域的代表性甲基化水平的技术，其可以使用相对低水平的测序来完成。可以使用甲基化位点的数量和甲基化位点的总数来测定每条链或每个分子或每个区域的甲基化水平。还分析了各种组织的甲基化水平。

[0377] 我们对11份人类组织DNA样本进行了测序，得到每份样本中值为3,070万的子读段(范围：910万-8,860万)，所述中值可以与人类参考基因组(hg19)进行比对。每份样本的子读段都是从中值为380万的太平洋生物科学公司单分子实时(SMRT)测序孔(范围：110-1,

150万)中产生的,所述孔中的每个孔含有至少一个可以与人类参考基因组比对的子读段。平均而言,SMRT孔中的每个分子平均测序9.9次(范围:6.5-13.4次)。人类组织DNA样本包含1份怀孕受试者的母本血沉棕黄层样本、1份胎盘样本、2份肝细胞癌(HCC)肿瘤组织、与前述2份HCC组织配对的2份邻近非肿瘤组织、4份来自健康对照受试者的血沉棕黄层样本(M1和M2来自男性受试者;F1和F2来自女性受试者)、1份HCC细胞系(HepG2)。图60示出了测序数据汇总的细节。

[0378] 图60示出了第一列中的不同组织组和第二列中的样本名称。“子读段总数”指示从SMRT孔中产生的序列的总数,包含从沃森链和克里克链产生的序列。“定位子读段”列出了可以与人类参考基因组比对的子读段的数量。“子读段可定位性”是指可以与人类参考基因组比对的子读段的比例。“每个SMRT孔的平均子读段深度”指示从每个SMRT孔产生的平均子读段数。“SMRT孔的数量”是指产生可检测的子读段的SMRT孔的数量。“可定位孔”指示含有至少一个可比对子读段的孔的数量。“可定位孔比例(%)”是含有至少一个可比对子读段的孔的比例。

[0379] a)甲基化水平和谱式分析技术

[0380] 在一个实施例中,可以测量单个核酸链(例如DNA或RNA)的甲基化密度,所述甲基化密度被定义为链内甲基化碱基的数量除以所述链内可甲基化碱基的总数。此测量还被称为“单链甲基化水平”。由于单分子实时测序平台可以从双链DNA分子的两条链中的每条链中获得测序信息,因此在本公开的上下文中,此单链测量是特别可行的。这通过在准备测序文库时使用发夹衔接子而得以促进,以便将双链DNA分子的沃森链和克里克链以环状形式连接并一起测序。实际上,此结构还使得同一双链DNA分子的配对沃森链和克里克链在同一反应中测序,使得可以单独测定并直接比较任何双链DNA分子的沃森链和克里克链上对应互补位点的甲基化状态(例如,图20A和20B)。

[0381] 使用其它技术无法轻松实现这些基于链的甲基化分析。因为在不使用本申请中公开的直接甲基化分析方法的情况下,将需要应用另一种方法,例如通过亚硫酸氢盐转化,将甲基化碱基与未甲基化碱基区分开。亚硫酸氢盐转化需要用亚硫酸氢钠处理DNA,使得甲基化胞嘧啶和未甲基化胞嘧啶可以分别被区分为胞嘧啶和胸腺嘧啶。在许多亚硫酸氢盐转化方案的变性条件下,双链DNA分子的两条链彼此解离。在许多测序应用中,使用例如Illumina平台,亚硫酸氢盐转化的DNA随后通过聚合酶链反应(PCR)进行扩增,所述聚合酶链反应涉及将双链DNA解离成单链。

[0382] 通过Illumina测序,可以在亚硫酸氢盐转化之前使用甲基化衔接子制备无PCR测序文库。即使使用此策略,双链DNA分子的每条DNA链也将被随机选择以用于流动细胞中的桥扩增。由于测序的随机性,因此不可能将来自同一DNA分子的每条链在同一反应中进行测序。即使在同一运行中分析了来自同一基因座的一个以上序列读段,也没有简单的方法来测定两个读段是来自一个双链DNA分子的配对沃森链和克里克链中的每一个,还是来自两个不同的双链DNA分子。此类考虑是重要的,因为在本发明的某些实施例中,双链DNA分子的两条链可能展现出不同的甲基化谱式。当测量多条核酸链(例如DNA或RNA)的单链甲基化密度时,还可以基于图61中关于“所关注的基因组区域的甲基化水平”的概念和方程式来测定“多链甲基化水平”。

[0383] 图61示出了分析甲基化谱式的各种方法。具有未知序列和甲基化信息的双链DNA

分子(X)与衔接子连接,所述衔接子在一个实例中形成发夹环结构。作为结果,在此实例中,DNA分子的两条单链(包含沃森X(a)链和克里克X(b)链)在物理上以环状形式合在一起。可以使用本公开中描述的方法(例如,使用动力学、电子、电磁、光信号或来自测序仪的其它类型的物理信号)来获得沃森链和克里克链两者中位点的甲基化状态。可以在同一反应中调查环化DNA分子中的沃森链和克里克链。测序后,修剪掉衔接子序列。

[0384] 通过分析可以测定不同的甲基化水平。在图61的(I)中,可以分析仅单链分子(如X(a)或X(b))的甲基化谱式。此分析可以被称为单链甲基化谱式分析。分析可以包含但不限于测定位点甲基化状态或甲基化谱式。在图61中,单链分子X(a)示出了甲基化谱式5'-UMMUU-3',其中"U"指示未甲基化位点并且"M"指示甲基化位点,而互补单链分子X(b)示出了甲基化谱式3'-UMUUU-5'。因此,X(b)具有与X(a)不同的甲基化谱式。X(a)和X(b)的对应单链甲基化水平分别为40%和20%。

[0385] 相反,如(II)中所示,可以分析单个双链DNA分子水平的甲基化谱式(即考虑沃森链和克里克链的甲基化谱式)。此分析可以被称为单分子双链DNA甲基化谱式分析。此示例性分子X的单分子双链DNA甲基化水平为30%。此分析的一种变体,即将来自沃森链和克里克链的动力学信号组合以分析修饰。具体地,由于CpG位点上的甲基化通常是对称的,因此在测定位点的甲基化状态之前,可以将来自沃森链和克里克链的动力学信号组合到一个位点。在一些情况下,使用从分子的沃森链和克里克链结合的动力学信号测定碱基修饰的性能将优于单独使用单链的动力学信号测定碱基修饰的性能。例如,如图20B所示,与单独使用单链(AUC:0.85)相比,组合使用来自包含沃森链和克里克链的两条链的动力学信号会产生更大的AUC(0.90)。

[0386] 在图61的(III)中,测定了所关注的基因组区域的甲基化水平,其中携带不同分子长度和不同数量的可甲基化位点(例如CpG位点)的不同DNA分子可以有助于所关注的基因组区域。此分析可以被称为多链甲基化水平分析。术语"多链"可以指多个单链DNA分子、或多个双链DNA分子、或其任何组合。在此实例中,有三个覆盖所关注的基因组区域的双链DNA分子:分子"X"、"Y"和"Z",每个分子具有"a"和"b"链。此区域的对应甲基化水平为9/28,即32%。要分析的基因组区域的长度可以为1nt、10nt、20nt、30nt、40nt、50nt、100nt、1knt(千核苷酸,即一千个核苷酸)、2knt、3knt、4knt、5knt、10knt、20knt、30knt、40knt、50knt、100knt、200knt、300knt、400knt、500knt、1Mnt(兆核苷酸,即100万个核苷酸)、2Mnt、3Mnt、4Mnt、5Mnt、10Mnt、20Mnt、30Mnt、40Mnt、50Mnt、100Mnt或200Mnt。基因组区域可以是染色体臂或整个基因组。

[0387] 还可以在测定分子中位点的甲基化状态后测定甲基化谱式。例如,在单个双链DNA分子上有三个连续的CpG位点的情况下,沃森链和克里克链中的每一个上的甲基化谱式可以针对这三个位点被显示为甲基化(M)、未甲基化(N)和甲基化(M)。例如对于沃森链,此谱式MNM可以被称为此区域的沃森链的"甲基化单倍型"。由于DNA甲基化维持活性的存在,双链DNA分子的沃森链和克里克链的甲基化谱式可能是彼此互补的。例如,如果CpG位点在沃森链上是甲基化的,则克里克链上的互补CpG位点也可以是甲基化的。类似地,沃森链上的未甲基化CpG位点可以与克里克链上的未甲基化CpG位点互补。

[0388] 在一个实施例中,可以测量单个DNA分子的甲基化水平,所述甲基化水平被定义为分子内甲基化碱基或核苷酸的数量除以所述分子内可甲基化碱基或核苷酸的总数。此测量

还被称为“单分子甲基化水平”。由于单分子实时测序平台可能具有较长的读段长度,因此此单分子测量在当前公开的上下文中可能特别有用。当测量多个DNA分子的单分子甲基化水平时,还可以基于图61中的概念和方程式测定“多分子甲基化水平”。例如,“多分子甲基化水平”可以是单分子甲基化水平的平均值或中值。

[0389] 在一些实施例中,可以在DNA分子上分析一种或多种遗传多态性(例如单核苷酸多态性(SNP))连同分子上位点的甲基化状态,从而揭示所述分子的遗传和表观遗传信息。此类分析将揭示所分析DNA分子的“阶段性甲基化单倍型”。阶段性甲基化单倍型分析可用于例如对母本血浆(含有携带母本和胎儿遗传和表观遗传特征的细胞游离DNA分子的混合物)中基因组印记和细胞游离核酸的研究。

[0390] b)甲基化结果的比较

[0391] 使用亚硫酸氢盐测序和使用本公开所述的单分子实时测序来测定图60的表格中的组织的全基因组水平下的甲基化密度。图62A示出了在y轴上通过亚硫酸氢盐测序定量的甲基化密度和在x轴上的组织类型。图62B示出了在y轴上通过本公开所述的单分子实时测序定量的甲基化密度和在x轴上的组织类型。

[0392] 图62A示出了使用亚硫酸氢盐测序(即对样本进行亚硫酸氢盐转化并且然后经受Illumina测序)的跨不同组织的甲基化密度(Lister等人,《自然(Nature)》2009;462:315-322),所述不同组织包含HepG2、HCC肿瘤组织、与HCC肿瘤邻近的匹配的正常肝组织(即邻近正常组织)、胎盘组织和血沉棕黄层样本。HepG2显示了最低的甲基化水平,其中甲基化水平为40.4%。血沉棕黄层样本显示了最高的甲基化水平,其中甲基化水平为76.5%。发现HCC肿瘤组织的平均甲基化密度(51.2%)低于匹配的邻近正常组织的平均甲基化密度(71.0%)。这与HCC肿瘤与邻近正常组织相比在全基因组水平下低甲基化的预期是一致的(Ross等人《表观基因组学(Epigenomics)》2010;2:245-69)。数据集由通过Sequel II测序试剂盒1.0制备的DNA产生。

[0393] 使用单分子实时测序和根据本公开的方法对相同组织的部分进行甲基化分析。结果示于图62B中。使用本公开的单分子实时测序方法的甲基化分析能够示出HepG2细胞系的甲基化程度最低,其次是所分析的HCC肿瘤组织,并且然后是胎盘组织。邻近非肿瘤肝组织样本的甲基化程度高于包含HCC和胎盘组织的其它组织,其中血沉棕黄层的甲基化程度最高。

[0394] 图63A、63B和63C示出了通过根据本文所述的方法的亚硫酸氢盐测序和单分子实时测序定量的总体甲基化水平的相关性。图63A示出了在x轴上通过亚硫酸氢盐测序定量的甲基化水平和在y轴上使用本文所述的方法通过单分子实时测序定量的甲基化水平。黑色实线是拟合的回归线。虚线是两个测量结果相等的地方。

[0395] 根据本文公开的发明,亚硫酸氢盐测序与单分子实时测序之间的甲基化水平具有非常高的相关性($r=0.99$;P值<0.0001)。这些数据指示,使用由此公开的单分子实时测序方法的甲基化分析是测定组织之间甲基化水平的有效手段,并且能够比较这些组织之间的甲基化状态和甲基化谱。对于甲基化水平的两个量度,我们注意到图63A中的回归线的斜率从一个度量偏离。这些结果表明,与常规的大规模并行亚硫酸氢盐测序相比,在使用根据本公开的单分子实时测序测定甲基化水平时,两个测量结果之间可能存在偏差(在一些情况下,此偏差可以被称为偏置)。

[0396] 在一个实施例中，可以使用线性或LOESS(局部加权平滑)回归对偏置进行定量。作为实例，如果将大规模并行亚硫酸氢盐测序(Illumina)视为参考，则可以使用回归系数来转换通过根据本公开的单分子实时测序测定的结果，从而协调不同平台之间的读数。在图63A中，线性回归公式为Y＝aX+b，其中“Y”表示通过根据本公开的单分子实时测序测定的甲基化水平；“X”表示由亚硫酸氢盐测序测定的甲基化水平；“a”表示回归线的斜率(例如，a＝0.62)；“b”表示y轴的截距(例如，b＝17.72)。在这种情况下，将通过(Y-b)/a计算通过单分子实时测序测定的经过协调的甲基化值。在另一个实施例中，可以使用两个测量结果之间的偏差(ΔM)和两个测量结果的对应平均值($\bar{M}$)的关系，所述偏差和所述对应平均值由以下公式(1)和(2)定义：

[0397] $\Delta M = S - \text{基于亚硫酸氢盐的甲基化}$，(1)

[0398] $\bar{M} = \frac{S + \text{基于亚硫酸氢盐的甲基化}}{2}$，(2)

[0399] 其中“S”表示通过根据本发明的单分子实时测序测定的甲基化水平，并且“基于亚硫酸氢盐的甲基化”表示通过亚硫酸氢盐测序测定的甲基化水平。

[0400] 图63B示出了ΔM与$\bar{M}$之间的关系。两个测量结果的平均值($\bar{M}$)绘制在x轴上，并且两个测量结果之间的偏差(ΔM)绘制在y轴上。虚线表示水平横跨零的线，所述线上的数据点表明两个测量结果之间没有差异。这些结果表明，偏差根据平均值而变化。两个测量结果的平均值越高，偏差的幅度就越大。ΔM值的中值为-8.5％(范围：-12.6％到+2.5％)，这表明两种方法之间存在差异。

[0401] 图63C示出了x轴上的两个测量结果的平均值($\bar{M}$)和y轴上的相对偏差(RD)。相对偏差由以下公式定义：

[0402] $RD = \frac{\Delta M}{\bar{M}} \times 100\%$，(3)。

[0403] 虚线表示水平横跨零的线，所述线上的数据点表明两个测量结果之间没有差异。这些结果表明，相对偏差根据平均值而变化。两个量度的平均值越大，相对偏差的幅度就越大。RD值的中值为-12.5％(范围：-18.1％到+6.0％)。

[0404] 据报道，常规的全基因组亚硫酸氢盐测序(Illumina)引入了显著偏差的序列输出并且高估了总体甲基化，其在定量特定基因组区域处的方法之间的甲基化水平中存在实质性变化(Olova等人《基因组生物学(Genome Biol.)》2018;19:33)。本文公开的方法可以在没有亚硫酸氢盐转化的情况下执行，所述亚硫酸氢盐转化会急剧降解DNA，并且可以在没有PCR扩增的情况下执行，所述PCR扩增可能会使过程复杂化或者可能在测定甲基化水平时引入其它错误。

[0405] 图64A和64B示出了1Mb分辨率下的甲基化谱式。图64A示出了HCC细胞系(HepG2)的甲基化谱式。图64B示出了来自健康对照受试者的血沉棕黄层样本的甲基化谱式。染色体表意文字(每个图中最外面的环)按顺时针方向从短臂末端排列到长臂末端。外部的第二个环(也被称为中间环)示出了通过亚硫酸氢盐测序测定的甲基化水平。最里面的环示出了通过根据本公开的单分子实时测序测定的甲基化水平。甲基化水平分为5个等级，即，0-20％(浅绿色)、20-40％(绿色)、40-60％(蓝色)、60-80％(浅红色)和80-100％(红色)。如图64A和64B中，1Mb分辨率下的甲基化谱在亚硫酸氢盐测序(中间轨迹)与根据本公开的单分子实时

测序(最里面的轨迹)之间是一致的。示出了母本血沉棕黄层样本的甲基化水平高于HCC细胞系(HepG2)。

[0406] 图65A和65B示出了在1Mb分辨率下测量的甲基化水平的散点图。图65A示出了HCC细胞系(HepG2)的甲基化水平。图65B示出了来自健康对照受试者的血沉棕黄层样本的甲基化水平。对于图65A和图65B这两个图,通过亚硫酸氢盐测序定量的甲基化水平在x轴上,并且通过根据本公开的单分子实时测序测量的甲基化水平在y轴上。实线是拟合的回归线。虚线是两种测量技术相等的地方。对于HCC细胞系,通过单分子实时测序以1Mb分辨率测定的甲基化水平与通过亚硫酸氢盐测序测量的甲基化水平具有良好的相关性(r＝0.99;P<0.0001)(图65A)。还观察到了来自血沉棕黄层样本的数据的相关性(r＝0.87,P<0.0001)(图65B)。

[0407] 图66A和66B示出了在100kb分辨率下测量的甲基化水平的散点图。图66A示出了HCC细胞系(HepG2)的甲基化水平。图66B示出了来自健康对照受试者的血沉棕黄层样本的甲基化水平。对于图66A和图66这两个图,通过亚硫酸氢盐测序定量的甲基化水平在x轴上,并且通过根据本公开的单分子实时测序测量的甲基化水平在y轴上。实线是拟合的回归线。虚线是两种测量技术相等的地方。当分析的分辨率增加到每个100kb(或100-knt)窗口时,还观察到在1Mb(或1-Mnt)分辨率下两种方法之间的甲基化定量测量结果之间的高度的相关性。所有这些数据指示,本公开的单分子实时方法是用于定量基因组区域内的甲基化水平或甲基化密度的有效工具,所述甲基化水平或甲基化密度在不同程度的分辨率下变化,例如在1Mb(或1-Mnt)或100kb(或100-knt)下变化。数据还指示,本发明是用于评估区域之间或样本之间的甲基化谱或甲基化谱式的有效工具。

[0408] 图67A和67B示出了1Mb分辨率下的甲基化谱式。图67A示出了HCC肿瘤组织(TBR3033T)的甲基化谱式。图67B示出了邻近正常组织(TBR3033N)的甲基化谱式。染色体表意文字(每个图中最外面的环)按顺时针方向从短臂末端排列到长臂末端。外部的第二个环(也被称为中间环)示出了通过亚硫酸氢盐测序测定的甲基化水平。最里面的环示出了通过根据本公开的单分子实时测序测定的甲基化水平。甲基化水平分为5个等级,即,0-20%(浅绿色)、20-40%(绿色)、40-60%(蓝色)、60-80%(浅红色)和80-100%(红色)。如图67A所示,可以检测到HCC肿瘤组织DNA(TBR3033T)中的低甲基化,所述HCC肿瘤组织DNA可以与图67B中的邻近正常肝组织DNA(TBR3033N)区分开。通过亚硫酸氢盐测序(中间轨迹)和根据本公开的单分子实时测序(最里面的轨迹)测定的甲基化水平和谱式是一致的。示出了邻近正常组织DNA的甲基化水平高于HCC肿瘤组织DNA的甲基化水平。

[0409] 图68A和68B示出了在1Mb分辨率下测量的甲基化水平的散点图。图68A示出了HCC肿瘤组织(TBR3033T)的甲基化水平。图68B示出了邻近正常组织的甲基化水平。对于图68A和图68B这两个图,通过亚硫酸氢盐测序定量的甲基化水平在x轴上,并且通过根据本公开的单分子实时测序测量的甲基化水平在y轴上。实线是拟合的回归线。虚线是两种测量技术相等的地方。对于HCC肿瘤组织DNA,通过单分子实时测序以1Mb分辨率测量的甲基化水平与通过亚硫酸氢盐测序测定的甲基化水平具有良好的相关性(r＝0.96;P值<0.0001)(图68A)。来自邻近正常肝组织样本的数据也是相关的(r＝0.83,P值<0.0001)(图68B)。

[0410] 图69A和69B示出了在100kb分辨率下测量的甲基化水平的散点图。图69A示出了HCC肿瘤组织(TBR3033T)的甲基化水平。图69B示出了邻近正常组织(TBR3033N)的甲基化水

平。对于图69A和图69B这两个图，通过亚硫酸氢盐测序定量的甲基化水平在x轴上，并且通过根据本公开的单分子实时测序测量的甲基化水平在y轴上。实线是拟合的回归线。虚线是两种测量技术相等的地方。当以更高的分辨率(例如，100kb窗口)执行甲基化水平的测量时，也观察到在1Mb分辨率下两种方法之间甲基化定量数据的这种高度相关性。

[0411] 图70A和70B示出了其它肿瘤组织和正常组织在1Mb分辨率下的甲基化谱式。图70A示出了HCC肿瘤组织(TBR3032T)的甲基化谱式。图70B示出了邻近正常组织(TBR3032N)的甲基化谱式。染色体表意文字(每个图中最外面的环)按顺时针方向从短臂末端排列到长臂末端。外部的第二个环(也被称为中间环)示出了通过亚硫酸氢盐测序测定的甲基化水平。最里面的环示出了通过根据本公开的单分子实时测序测定的甲基化水平。甲基化水平分为5个等级，即，0-20％(浅绿色)、20-40％(绿色)、40-60％(蓝色)、60-80％(浅红色)和80-100％(红色)。如图70A所示，可以检测到HCC肿瘤组织DNA(TBR3032T)中的低甲基化，所述HCC肿瘤组织DNA可以与图70B中的邻近正常肝组织DNA(TBR3032N)区分开。通过亚硫酸氢盐测序(中间轨迹)和施用本发明的单分子实时测序(最里面的轨迹)测定的甲基化水平和谱式是一致的。示出了邻近正常组织DNA的甲基化水平高于HCC肿瘤组织DNA的甲基化水平。

[0412] 图71A和71B示出了在1Mb分辨率下测量的甲基化水平的散点图。图71A示出了HCC肿瘤组织(TBR3032T)的甲基化水平。图71B示出了邻近正常组织的甲基化水平。对于图71A和图71B这两个图，通过亚硫酸氢盐测序定量的甲基化水平在x轴上，并且通过根据本公开的单分子实时测序测量的甲基化水平在y轴上。实线是拟合的回归线。虚线是两种测量技术相等的地方。对于HCC肿瘤组织DNA，通过单分子实时测序以1Mb分辨率测量的甲基化水平与通过亚硫酸氢盐测序测定的甲基化水平具有良好的相关性($r=0.98$;$P<0.0001$)(图71A)。来自邻近正常肝组织样本的数据也是相关的($r=0.87$,$P<0.0001$)(图71B)。

[0413] 图72A和72B示出了在100kb分辨率下测量的甲基化水平的散点图。图72A示出了HCC肿瘤组织(TBR3032T)的甲基化水平。图72B示出了邻近正常组织(TBR3032N)的甲基化水平。对于图72A和图72B这两个图，通过亚硫酸氢盐测序定量的甲基化水平在x轴上，并且通过根据本公开的单分子实时测序测量的甲基化水平在y轴上。实线是拟合的回归线。虚线是两种测量技术相等的地方。当以更高的分辨率(例如，100kb窗口)执行甲基化水平的测量时，也观察到在1Mb分辨率下两种方法之间甲基化定量数据的这种高度相关性。

[0414] 4.肿瘤与邻近正常组织之间的差异甲基化区域

[0415] 经常在癌症基因组的区域中发现甲基化组异常。此类异常的一个实例是选定基因组区域的低甲基化和高甲基化(Cadieux等人《癌症研究(Cancer Res.)》2006;66:8469-76;Graff等人《癌症研究》1995;55:5195-9;Costello等人《自然遗传学(Nat Genet.)》2000;24:132-8)。另一个实例是选定基因组区域中甲基化和未甲基化碱基的异常谱式。本节示出了测定甲基化的技术可以用于在分析肿瘤时执行定量分析和诊断。

[0416] 图73示出了肿瘤抑制基因CDKN2A附近的甲基化异常谱式的实例。用蓝色突出显示并加下划线的坐标指示CpG岛。黑色实心点指示甲基化位点。空心点指示未甲基化位点。每条带点的横线右侧括号中的数字指示片段的长度、单分子甲基化密度和CpG位点的数量。例如，(3.3kb,MD:17.9％,CG:39)意指此片段的长度为3.3kb，此片段的甲基化水平为17.9％并且CpG位点的数量为39。MD表示甲基化密度。

[0417] 如图73所示，CDKN2A(细胞周期蛋白依赖性激酶抑制剂2A)基因对包含INK4A(p16)

和ARF(p14)的两种蛋白质进行编码,从而充当肿瘤抑制剂。在邻近肿瘤组织的非肿瘤组织中,有两个分子(分子7301和分子7302)覆盖了与CDKN2A基因重叠的区域。分子7301和分子7302的单个双链DNA分子的甲基化水平分别示出为17.9%和7.6%。相反,发现肿瘤组织中存在的分子7303的单个双链DNA分子的甲基化水平为93.9%,这远高于配对的邻近非肿瘤组织中存在的分子的甲基化水平。另一方面,还可以使用邻近肿瘤组织的非肿瘤组织中存在的分子7301和7302来计算多链甲基化水平。作为结果,多链甲基化水平为9.7%,其低于肿瘤组织的水平(93.9%)。不同的甲基化水平表明,可以使用单个双链分子甲基化水平和/或多链甲基化水平来检测或监测如癌症等疾病。

[0418] 图74A和图74B示出了根据本发明的实施例的通过单分子实时测序检测的差异甲基化区域。图74A示出了癌症基因组中的低甲基化。图74B示出了癌症基因组中的高甲基化。x轴指示CpG位点的坐标。用蓝色突出显示并加下划线的坐标指示CpG岛。黑色实心点指示甲基化位点。空心点指示未甲基化位点。每条带点的横线右侧括号中的数字指示片段的长度、片段级甲基化密度和CpG位点的数量。例如,(3.1kb,MD:88.9%,CG:180)意指此片段的长度为3.1kb,此片段的甲基化密度为88.9%并且CpG位点的数量为180。

[0419] 图74A示出了靠近GNAS基因的区域,所述区域与邻近正常肝组织相比在HCC肿瘤组织中显示出更多的低甲基化片段。图74B示出了靠近ESR1基因的区域,所述区域在HCC组织中显示出高甲基化片段,但是来自配对的邻近非肿瘤组织的与对应区域比对的DNA片段反而显示了低甲基化。如图74B所示,当癌症样本与非癌症样本比较时,单个DNA分子的甲基化谱或甲基化单倍型足以揭示那些基因组区域(即GNAS和ESR1)的异常甲基化状态。

[0420] 这些数据指示,此处公开的单分子实时测序甲基化分析可以测定单个DNA片段上每个CpG位点(无论是甲基化还是未甲基化)的甲基化状态。单分子实时测序的读段长度比Illumina测序的读段长度要长得多(约千碱基长),所述Illumina测序每个读段通常跨越100-300nt的长度(De Maio等人《微生物基因组(Micob Genom.)》2019;5(9))。将单分子实时测序的长读段长度性质与此处公开的甲基化分析方法相结合,可以容易地测定沿任何单个DNA分子存在的多个CpG位点的甲基化单倍型。甲基化谱是指CpG位点从基因组的一个坐标到连续DNA段内(例如,在同一条染色体上,或在细菌质粒内,或在病毒基因组中的单个DNA段内)的另一个坐标的甲基化状态。

[0421] 因为单分子实时测序无需事先扩增就可以单独分析每个DNA分子,所以针对任何单个DNA分子测定的甲基化谱实际上是甲基化单倍型,意思是CpG位点从同一DNA分子的一端到另一端的甲基化状态。如果从同一基因组区域测序一个或多个分子,则可以使用如图61所示的同一公式从多个DNA片段的数据中聚合基因组区域中跨所有经过测序的CpG位点的每个CpG位点的甲基化%(即甲基化水平或甲基化密度)。可以针对提供经过测序的基因组区域的甲基化谱的所有经过测序的CpG位点报告每个CpG位点的甲基化%。可替代地,可以从经过测序的基因组区域内的所有读段和所有位点集合数据以提供区域的一个甲基化%值,即以图64到72所示的计算1Mb或1kb区域的甲基化水平的相同方式。

[0422] 5.病毒DNA甲基化分析

[0423] 本节示出了本公开的甲基化技术可以用于准确地测定病毒DNA中的甲基化水平。

[0424] 图75示出了使用单分子实时测序的两对HCC组织样本与邻近非肿瘤组织样本之间的乙型肝炎病毒DNA的甲基化谱式。每个箭头表示HBV基因组中的基因注释。带有“P”、“S”、

“X”和“C”的箭头分别指示有关HBV基因组的基因注释:编码聚合酶、表面抗原、X蛋白和核心蛋白。鉴定了一个片段(分子I),其长度为1,183bp,来源于邻近非肿瘤组织,跨越以虚线矩形突出显示的从2,278到3,141的HBV基因组,示出甲基化水平为12%。还鉴定了来源于肿瘤组织长度为3,215bp、2,961bp和3,105bp的三个片段(分子II、III和IV)。其中,HCC肿瘤中的两个片段(分子III和IV)与非肿瘤组织中的分子I跨越的HBV基因组区域重叠。与虚线矩形(HBV基因组位置:2,278-3,141)中突出显示的HBV区域的低甲基化水平(12%)相反,HCC组织中的那些片段(分子III和IV)的甲基化水平较高(即24%和30%)。这些结果表明,使用单分子实时测序的方法可测定病毒基因组中的甲基化谱式,并且能够鉴定HCC与非HCC组织之间的HBV的差异甲基化区域(DMR)。因此,使用根据本公开的单分子实时测序来测定跨病毒基因组的甲基化状态将提供一种使用组织活检来研究临床相关性的新工具。

[0425] 此DMR区域碰巧与基因P、C和S重叠。据报道,与具有HBV感染但不具有癌症的肝组织相比,此区域在HCC组织中也显示为高甲基化(Jain等人《科学报告(Sci Rep.)》2015;5:10478;Fernandez等人《基因组研究(Genome Res.)》2009;19:438-51)。

[0426] 汇集了四名患有肝硬化但无HCC患者肝组织的亚硫酸氢盐测序结果,获得了1,156个HBV片段用于甲基化分析。图76A示出了患有肝硬化但无HCC患者的肝组织中乙型肝炎病毒DNA的甲基化水平。另外,汇总了15名患者的HCC肿瘤组织的亚硫酸氢盐测序结果,获得了736个HBV片段用于甲基化分析。图76B示出了HCC肿瘤组织中乙型肝炎病毒DNA的甲基化水平。如图76A和图76B所示,还通过大规模并行亚硫酸氢盐测序观察到了在HCC组织中甲基化水平高于肝硬化肝组织的HBV的DMR区域(HBV基因组位置:1,982-2,435)。这些结果表明,用于测定病毒基因组的甲基化状态的方法是有效的。

[0427] 6.变体相关甲基化分析

[0428] 不同的等位基因可以与不同的甲基化谱相关联。例如,印记基因可能具有一个甲基化水平高于另一个等位基因的等位基因。本节示出了甲基化谱可以用于区分某些基因组区域中的等位基因。

[0429] 一个含有单个DNA模板的单分子实时测序孔将产生许多子读段。子读段包含动力学特征[例如脉冲间持续时间(IPD)和脉冲宽度(PW)]和核苷酸组合物。在一个实施例中,来自一个单分子实时测序孔的子读段可以用于产生一致性序列(还被称为环形一致性序列,CCS),所述一致性序列可以显著减少测序误差(例如错配、插入或缺失)。本文描述了CCS的其它细节。在一个实施例中,可以使用与人类参考基因组比对的那些子读段来构建一致性序列。在另一个实施例中,可以通过将子读段映射到同一单分子实时测序孔中最长的子读段来构建一致性序列。

[0430] 图77展示了阶段性甲基化单倍型分析的原理。实心圆点表示被归类为甲基化的CpG位点。空心圆点表示被归类为未甲基化的CpG位点。

[0431] 如图77中的一个实施例所示,子读段与人类参考基因组比对。将来自一个单分子实时测序孔的比对子读段折叠以形成一致性序列。通常可以使用跨每个比对位置的子读段中存在的最频繁的核苷酸来测定一致性序列。因此,可以从一致性序列中鉴定出核苷酸变体,包含但不限于单核苷酸变体、插入和缺失。根据本公开,可以使用由核苷酸变体标记的同一分子中的平均IPD和PW来测定甲基化谱式。因此,可以进一步测定变体相关的甲基化谱式。同一分子中的甲基化状态可以被视为甲基化单倍型。甲基化单倍型可能不容易直接由

两个或更多个短DNA分子构建,因为可能没有分子标志物来区分两个或多个片段化短DNA分子是源自原始单个分子还是由两个或更多个不同的原始分子贡献的。合成长读段技术(如10X基因组学公司(10X Genomics)开发的连接读段测序)提供了一种可能性,其将单个长DNA分子分布到一个分区(如液滴)中,并用相同的分子条形码序列标记来源于所述长DNA分子的短DNA分子。然而,此条形码步骤涉及无法保留原始甲基化状态的PCR扩增。

[0432] 此外,如果试图使用亚硫酸氢盐处理长DNA分子,则亚硫酸氢盐处理之前的第一步涉及在破坏性条件下的DNA变性,从而将双链DNA变成单链DNA,因为亚硫酸氢盐在某些化学条件下只能作用于单链DNA分子。此DNA变性步骤会将长DNA分子降解为短片段,从而导致原始甲基化单倍型信息的丢失。基于亚硫酸氢盐的甲基化分析的第二个缺点是在亚硫酸氢盐转化步骤中将双链DNA变性为单链DNA,即沃森链和克里克链。对于分子,对沃森链进行测序的可能性为50%,并且对克里克链进行测序的可能性为50%。在数百万条沃森链和克里克链中,同时对分子的沃森链和克里克链进行测序的可能性极低。即使假设分子的沃森链和克里克链都被测序,仍然不可能绝对地确定此类沃森链和克里克链是源自原始单个片段还是由两个或更多个不同的原始片段贡献的。Liu等人最近介绍了一种用于检测甲基化胞嘧啶和羟甲基胞嘧啶的无亚硫酸氢盐测序方法(Liu等人《自然生物技术(Nat Biotechnol.)》2019;37:424-429)所述方法在温和条件下使用基于十-十一易位(TET)酶的转化,从而导致较少DNA降解。然而,所述方法涉及两个连续的酶促反应步骤。酶促反应的任一步骤的低转化率都将显著影响总转化率。另外,即使对于这种用于检测甲基化胞嘧啶的无亚硫酸氢盐的测序方法,在测序结果中仍然难以区分分子的沃森链和克里克链。

[0433] 相反,在本发明的实施例中,分子的沃森链和克里克链通过钟形衔接子共价连接形成环状DNA分子。作为结果,分子的沃森链和克里克链在同一反应孔中进行测序,并且可以确定每条链的甲基化状态。

[0434] 本发明的实施例的一个优点是能够确定长的连续DNA分子(例如长度为千碱基或千核苷酸)的甲基化和遗传(即序列)信息。使用短读测序技术产生此类信息更加困难。对于短读段测序技术,必须使用遗传或表观遗传学特征的支架将多个短读段上的测序信息结合起来,使得可以推导出长段甲基化和遗传信息。然而,由于此类遗传或表观遗传锚之间的距离,这在许多情况下可能证明是具有挑战性的。例如,平均每1kb有一个SNP,而目前的短读段测序技术通常可以每个读段测序高达300nt,即使是配对末端形式也可以测序600nt。

[0435] 在一个实施例中,变体相关的甲基化单倍型分析可以用于研究印记基因中的甲基化谱式。印记区域以亲源的方式经受表观遗传调控(例如CpG甲基化)。例如,对图60的表格中的一个血沉棕黄层DNA样本(M2)进行测序以获得约1.52亿个子读段。对于此样本,53%的单分子实时测序孔产生了至少一个可与人类参考基因组比对的子读段。每个SMRT孔的平均子读段深度为7.7x。总共获得了约300万个一致性序列。约91%的参考基因组至少被一致性序列覆盖一次。对于覆盖区域,测序深度为7.9x。数据集由通过Sequel II测序试剂盒1.0制备的DNA产生。

[0436] 图78示出了根据一致性序列测定的经过测序的分子的长度分布,其中中值长度为6,289bp(范围:66-198,109bp)。片段长度(bp)在x轴上示出并且与片段长度相关的频率(%)在y轴上示出。

[0437] 图79A、79B、79C和79D示出了印记区域中的等位基因甲基化谱式的实例。x轴指示

CpG位点的坐标。用蓝色突出显示并加下划线的坐标指示CpG岛。黑色实心点指示甲基化CpG位点。空心点指示未甲基化CpG位点。嵌入每个水平系列的实心点和空心点(即CpG位点)之间的字母指示SNP位点处的等位基因。每条水平系列点右侧括号中的数字指示片段的长度、片段级甲基化密度和CpG位点的数量。例如,(10.0kb,MD:79.1%,CG:139)表明对应片段的长度为10.0kb,片段的甲基化密度为79.1%并且CpG位点的数量为139。虚线矩形描绘了每个基因内甲基化差异最大的区域。

[0438] 图79A示出了源自SNURF基因的中值长度为11.2kb(范围:1.3-25kb)的11个经过测序的片段。SNURF基因是母本印记的,这意味着个体从母体继承的基因的拷贝是甲基化和转录沉默的。如图79A所示,在虚线矩形中,C等位基因相关片段是高度甲基化的,而T等位基因相关片段是高度未甲基化的。高度甲基化可以指示超过70%、80%、90%、95%或99%的位点被甲基化。可以在包含PLAGL1(图79B)、NAP1L5(图79C)和ZIM2(图79D)的其它印记基因中观察到等位基因特异性甲基化谱式。图79B示出了对于PLAGL1,T等位基因相关片段是高度未甲基化的,而C等位基因相关片段是高度甲基化的。图79C示出了对于NAP1L5,C等位基因相关片段是高度未甲基化的并且所述T等位基因相关片段是高度甲基化的。图79D示出了对于ZIM2,C等位基因相关片段是高度未甲基化的并且所述T等位基因相关片段是高度甲基化的。

[0439] 图80A、80B、80C和80D示出了非印记区域中的等位基因甲基化谱式的实例。x轴指示CpG位点的坐标。用蓝色突出显示并加下划线的坐标指示CpG岛。黑色实心点指示甲基化CpG位点。空心点指示未甲基化CpG位点。嵌入每个水平系列的实心点和空心点(即CpG位点)之间的字母指示单核苷酸多态性(SNP)位点处的等位基因。每条水平系列点右侧括号中的数字指示片段的长度、片段级甲基化密度和CpG位点的数量。虚线矩形指示随机选择的区域,用于计算括号中报告的甲基化密度。与图79A-79D中的结果相反,非印记基因中不存在此类可观察到的等位基因甲基化谱式。图80A示出了在chr7区域中没有不同的等位基因甲基化谱式。图80B示出了在chr12区域中没有不同的等位基因甲基化谱式。图80C示出了在chr1区域中没有不同的等位基因甲基化谱式。图80D示出了在另一个chr1区域中没有不同的等位基因甲基化谱式。

[0440] 图81示出了具有等位基因特异性片段的甲基化水平的表格。第一列列出了“印记基因”和“随机选择区域”的类别。第二列列出了特定基因。第三列列出了基因中SNP的第一等位基因。第四列列出了基因中SNP的第二等位基因。第五列示出了与第一等位基因连接的片段的甲基化水平。第六列示出了与第二等位基因连接的片段的甲基化水平。对于那些印记基因(P值=0.03),与等位基因2连接的片段(平均值:88.6%;范围:84.6%-91.1%)的甲基化水平比那些与等位基因1连接的片段(平均值:12.2%;范围7.6-15.7%)要高得多,这表明存在等位基因特异性甲基化。相反,那些随机选择区域之间的甲基化水平没有显著变化(P值=1),这表明不存在等位基因特异性甲基化。

[0441] 7.妊娠期细胞游离DNA分析

[0442] 在此例示中,证明了此处公开的方法适用于分析从至少有一个胎儿的孕妇获得的血浆或血清中的细胞游离核酸。在怀孕期间,在母本循环中发现了来自胎盘细胞的细胞游离DNA和细胞游离RNA分子。此类胎盘来源的细胞游离核酸分子在母本血浆或循环的细胞游离胎儿核酸中也被称为细胞游离胎儿核酸。在母本细胞游离核酸的背景中,母本血浆中存

在细胞游离胎儿核酸。例如，循环的细胞游离胎儿DNA分子在母本血浆和血清中细胞游离母本DNA的背景中作为次要物种存在。

[0443] 为了区分细胞游离母本DNA与母本血浆或血清中的细胞游离胎儿DNA，已知可以使用遗传或表观遗传手段或其组合。在遗传学上，胎儿基因组与母本基因组的不同之处在于父本遗传的胎儿特异性SNP等位基因、父本遗传的突变或新生突变。在表观遗传学上，与母本血细胞的甲基化组相比，胎盘甲基化组通常是低甲基化的(Lun等人《临床化学(Clin Chem.)》2013;59:1583-94)。因为胎盘是细胞游离胎儿DNA的主要贡献者，而母本血细胞是细胞游离母本DNA在母本循环(血浆或血清)中的主要贡献者，所以与血浆或血清中的细胞游离母本DNA相比，细胞游离胎儿DNA分子通常是低甲基化的。与母本血细胞相比，存在胎盘是高甲基化的特定的基因组基因座。例如，RASSF1A的启动子和外显子1区域在胎盘中比在母本血细胞中的甲基化程度更高(Chiu等人《美国病理学杂志(Am J Pathol.)》2007;170:941-950)。因此，与来自相同基因座的循环的细胞游离母本DNA相比，来自此RASSF1A基因座的循环的细胞游离胎儿DNA将是高甲基化的。

[0444] 在实施例中，可以基于两个循环核酸池之间的不同甲基化状态将细胞游离胎儿DNA与细胞游离母本DNA分子区分开。例如，发现沿细胞游离DNA分子的CpG位点大部分是未甲基化的，此分子很可能来自胎儿。如果发现沿细胞游离DNA分子的CpG位点大部分是甲基化的，此分子很可能来自母体。本领域的技术人员已知有若干种方法来确定此类分子是否确实来自胎儿或母体。一种方法是将经过测序的分子的甲基化谱式与胎盘或母本血细胞中对应基因座的已知甲基化谱进行比较。

[0445] 图82示出了用于使用甲基化谱确定怀孕期间血浆DNA的胎盘来源的实例。用蓝色突出显示并加下划线的坐标指示CpG岛。黑色实心点指示甲基化位点。空心点指示未甲基化位点。每条带点的横线附近的括号中的数字指示片段的长度、单分子甲基化密度和CpG位点的数量。

[0446] 如图82所示，如果母本血浆细胞游离DNA分子与RASSF1A的启动子区域(已知在胎盘组织中是特异甲基化的区域)比对并且使用本发明的方法产生的测序数据是高甲基化的，则此分子可能来源于胎儿或胎盘。相反，示出低甲基化的分子很可能来源于母本背景DNA(主要是造血起源)。

[0447] 图83展示了用于胎儿特异性甲基化分析的方法。所述方法包含利用含有胎儿特异性SNP等位基因或胎儿特异性突变(例如，父本遗传或自然新生)的测序分子。当鉴定出此类胎儿特异性遗传特征时，存在于同一细胞游离DNA分子上的碱基的甲基化状态反映了细胞游离胎儿DNA或胎盘甲基化组的甲基化谱。当血浆细胞游离DNA测序揭示母本基因组中不存在的等位基因或突变时(例如，通过分析母本基因组DNA)，或通过分析父本DNA或已知在家族中传播时(例如，通过分析来自先证者的DNA)，胎儿特异性遗传特征可以被发现。

[0448] 可以通过分析那些携带与母本基因组中纯合等位基因不同的等位基因的DNA片段来测定胎儿特异性DNA分子的甲基化。可以期望胎儿DNA分子的甲基化低于母本DNA分子的甲基化。

[0449] 作为实例，对一名孕妇的血沉棕黄层DNA和其匹配的胎盘DNA进行测序，以分别获得59x和58x单倍型基因组覆盖率。总共鉴定了822,409个信息性SNP，其中母体是纯合子并且胎儿是杂合子。通过单分子实时测序，在母本血浆(M13160)中发现了2,652个胎儿特异性

片段和24,837个共享片段(即携带共享等位基因的片段;主要是母本起源)。胎儿DNA分率为19.3%。根据本公开,推导了那些胎儿特异性片段和共享片段的甲基化谱。作为结果,发现胎儿特异性片段的甲基化水平为57.4%,而共享片段的甲基化水平为69.9%。这一发现与目前的知识一致,即孕妇血浆中胎儿DNA的甲基化水平低于母本DNA(Lun等人《临床化学》2013;59:1583-94)。

[0450] 甲基化谱式可以用于诊断或监测目的。例如,母本血浆样本的甲基化谱已用于确定胎龄(https://www.ncbi.nlm.nih.gov/pubmed/27979959)。一种应用是质量控制步骤。另一个潜在的应用是监测妊娠的"生物"与"时间"年龄。此应用可以用于早产的检测或风险评估。其它实施例可以用于分析母本血液中的胎儿细胞。在仍其它实施例中,此类胎儿细胞可以通过基于抗体的方法或通过使用细胞标志物的选择性染色(例如,在细胞表面上或细胞质中)来鉴定,或者通过流式细胞术或显微操作或显微切割或物理方法(例如,通过腔室、表面或容器的不同流速)来富集。

[0451] C.使用不同试剂的甲基化检测

[0452] 本节示出了甲基化技术并不限于特定的试剂系统。

[0453] 使用不同的试剂系统执行甲基化分析,以确认可以应用技术。作为实例,使用Sequel II系统(太平洋生物科学公司)执行SMRT-seq以进行单分子实时测序。使用SMRTbell快速模板制备试剂盒2.0(太平洋生物科学公司)对经过剪切的DNA分子进行单分子实时(SMRT)测序模板构建。用SMRT Link v8.0软件(太平洋生物科学公司)计算了测序引物退火和聚合酶结合条件。简而言之,将测序引物v2退火到测序模板,并且然后使用Sequel II结合和内部控制试剂盒2.0(太平洋生物科学公司)将聚合酶与模板结合。在Sequel II SMRT Cell 8M上执行测序。用Sequel II测序试剂盒2.0(太平洋生物科学公司)在Sequel II系统上收集了30个小时的测序影像。在其它实施例中,其它化学试剂和反应缓冲液将用于SMRT序列。在一个实施例中,聚合酶将根据其甲基化状态具有沿DNA模板链掺入核苷酸的不同动力学特征(Huber等人《核酸研究(Nucleic Acids Res.)》2016;44:9881-9890)。在本公开中,除非另有说明,否则使用测序引物v1产生结果。

[0454] 为了证明本发明在本文所述公开内容中使用不同试剂的用途,分析了基于不同测序试剂盒产生的SMRT-seq数据,包含但不限于Sequel I测序试剂盒3.0、RS II、Sequel II测序试剂盒1.0和Sequel II测序试剂盒2.0。RS II每个SMRT单元包含150,000ZMW。Sequel每个SMRT单元使用1,000,000ZMW。Sequel II通过两个测序试剂盒(1.0和2.0)每个SMRT单元使用800万ZMW。此分析涉及两个数据集。基于全基因组扩增后的DNA制备了第一数据集,其表示未甲基化状态。基于M.SsssI甲基转移酶处理后的DNA制备了第二类型数据集,其表示甲基化状态。这些数据使用Sequel测序仪中的Sequel测序试剂盒3.0;Sequel II测序仪中的Sequel II测序试剂盒1.0和Sequel II测序试剂盒2.0产生。因此,获得了具有用不同试剂(例如聚合酶)产生的动力学特征的三个数据集。将每个数据集分为训练数据集和测试数据集以使用根据本公开的CNN模型来评估性能。

[0455] 1.测量窗口

[0456] 图84A、84B和84C示出了SMRT-seq在包括全基因组扩增数据(未甲基化CpG位点)和经过M.SsssI处理的数据(甲基化CpG位点)的训练数据集中跨不同试剂盒的不同测量窗口尺寸的性能。真阳性率绘制在y轴上,并且假阳性率绘制在x轴上。图84A示出了基于Sequel

测序试剂盒3.0产生的SMRT-seq数据。图84B示出了基于Sequel II测序试剂盒1.0产生的SMRT-seq数据。图84C示出了基于Sequel II测序试剂盒2.0产生的SMRT-seq数据。在图中，“-”指示被分析的CpG胞嘧啶位点的上游信号。“+”指示被分析的CpG胞嘧啶位点的下游信号。例如，“-6nt”表示被分析的CpG胞嘧啶位点的6nt上游信号。“+6nt”表示被分析的CpG胞嘧啶位点的6nt下游信号。“±6nt”指示包含被分析的CpG胞嘧啶位点的6nt上游信号和6nt下游信号两者(即，CpG胞嘧啶位点侧翼总共有12nt序列)。

[0457] 对于基于Sequel测序试剂盒3.0的训练数据集，如图84A所示，使用包括被分析的CpG胞嘧啶上的信号和所述胞嘧啶位点(由-6nt表示)的6nt上游信号(例如，IPD、PW、相对位置和序列组成)的测量窗口，AUC值为0.50表明在区分甲基化CpG胞嘧啶与未甲基化CpG胞嘧啶时没有辨别能力。然而，对于基于Sequel II测序试剂盒1.0和2.0的训练数据集，对应的AUC值为0.62(图84B)和0.75(图84C)。这些数据表明，SMRT-seq中使用的不同试剂具有不同的内在动力学特性。这些数据表明，本文公开的方法易于适应不同试剂的使用。此外，随着试剂的进一步发展，例如使用不同的聚合酶和其它化学物质，可以潜在地提高检测碱基修饰的准确性。

[0458] 作为另一个实例，对于基于Sequel测序试剂盒3.0的训练数据集，如图84A所示，使用包括CpG胞嘧啶位点(由-10nt表示)的10bp上游信号的测量窗口，AUC值为0.50表明在区分甲基化CpG胞嘧啶与未甲基化CpG胞嘧啶时没有辨别能力。然而，对于基于Sequel II测序试剂盒1.0和2.0的训练数据集，对应的AUC值为0.66(图84B)和0.79(图84C)，所述值与包括6nt上游信号的测量窗口相比有所提高。这些数据证实了用于SMRT-seq的不同试剂具有不同的内在动力学特性。这些数据表明，本文公开的方法易于适应不同试剂的使用。

[0459] 与具有上游信号的测量窗口相反，具有下游信号的测量窗口可以使得分类性能更大地提高。例如，对于基于Sequel测序试剂盒3.0的训练数据集，如图84A所示，使用包括CpG胞嘧啶位点的6nt下游信号(+6nt)的测量窗口，0.94的AUC值比使用6nt上游信号的值(AUC：0.5)要大得多。对于基于Sequel II测序试剂盒1.0和2.0的训练数据集，对应的AUC值分别为0.95(图84B)和0.92(图84C)，示出了与包括6nt上游的测量窗口相比有所提高。这些数据表明，与序列上下文关联的动力学特征将使用但不限于CNN模型来提高分类能力。这些数据还表明，通过调整测量窗口，本文的公开内容将适用于通过不同的试剂和测序条件(例如，不同的聚合酶、其它化学试剂，其浓度和测序反应参数(例如，持续时间))产生的数据集。使用包含CpG胞嘧啶位点的10nt下游信号的测量窗口可以通过分析得出类似的结论(图84A、84B和84C)。

[0460] 在另一个实施例中，可以使用包括被分析的胞嘧啶上的信号以及所述胞嘧啶的上游和下游信号的测量窗口。例如，如图84A、84B和84C所示，使用包括6nt上游信号和6nt下游信号(用±6nt表示)的测量窗口，发现基于Sequel测序试剂盒3.0、Sequel II测序试剂盒1.0和2.0的训练数据集的AUC值分别为0.94、0.95和0.92。使用包括10nt上游信号和10nt下游信号(用±10nt表示)的测量窗口，发现基于Sequel测序试剂盒3.0、Sequel II测序试剂盒

[0461] 1.0和2.0的训练数据集的AUC值分别为0.94、0.95和0.94。这些数据表明，本文的公开内容将广泛适用于由不同试剂和测序反应参数产生的数据集。

[0462] 图85A、85B和85C示出了当应用从训练数据集训练的CNN模型时，从跨不同测序试

剂盒具有不同测量窗口的测试数据集获得了结果。真阳性率绘制在y轴上,并且假阳性率绘制在x轴上。图例中的标记等同于84A、84B和84C中使用的标记。图85A示出了基于Sequel测序试剂盒3.0产生的SMRT-seq数据。图85B示出了基于Sequel II测序试剂盒1.0产生的SMRT-seq数据。图85C示出了基于Sequel II测序试剂盒2.0产生的SMRT-seq。训练数据集中得出的所有结论可以在训练过程中不涉及的这些独立测试数据集中进行验证。另外,在三个独立的测试数据集中,对涉及Sequel II测序试剂盒1.0和2.0的两个数据集(2/3)的分析示出,使用包含10nt上游和下游信号(表示为±10nt)的测量窗口的性能要优于其它两个。

[0463] 2.与亚硫酸氢盐测序的比较

[0464] 图86A、86B和86C示出了通过亚硫酸氢盐测序和SMRT-seq(Sequel II测序试剂盒2.0)定量的总体甲基化水平的相关性。图86A示出了在y轴上通过SMRT-seq定量的百分比形式的甲基化水平。图86B示出了在x轴上通过亚硫酸氢盐测序定量的百分比形式的甲基化水平。黑线是拟合的回归线。虚线是两个度量相等的对角线。图86B示出了布兰德-奥特曼图(Bland-Altman plot)。x轴指示根据本公开的SMRT-seq和亚硫酸氢盐测序定量的甲基化水平的平均值。y轴指示根据本公开的SMRT-seq与亚硫酸氢盐测序之间的甲基化水平差异(即太平洋生物科学公司甲基化-基于亚硫酸氢盐的甲基化)。虚线对应于水平横跨零的线,在所述线上两个度量之间没有差异。偏离虚线的数据点表明度量之间存在偏差。图86C示出了相对于通过亚硫酸氢盐测序定量的值的百分比变化。x轴指示根据本公开的SMRT-seq和亚硫酸氢盐测序定量的甲基化水平的平均值。y轴指示两个度量之间的甲基化水平差异相对于甲基化水平平均值的百分比。虚线对应于水平横跨零的线,在所述线上两个度量之间没有差异。偏离虚线的数据点表明度量之间存在偏差。

[0465] 对于图86A,线性回归公式为Y＝aX+b,其中“Y”表示根据本公开通过SMRT-seq测定的甲基化水平;“X”表示由亚硫酸氢盐测序测定的甲基化水平;“a”表示回归线的斜率(如a＝1.45);“b”表示y轴的截距(如b＝-20.98)。在这种情况下,通过SMRT-seq测定的甲基化值将通过(Y-b)/a计算。此图示出了与Sequel II测序试剂盒1.0一样,对于Sequel II测序试剂盒2.0,通过SMRT-seq测定的甲基化水平可以转换为通过亚硫酸氢盐测序测定的甲基化水平,反之亦然。

[0466] 图86B是示出了根据本公开的SMRT-seq与亚硫酸氢盐测序之间的甲基化定量偏差的布兰德-奥特曼图,其中x轴指示通过根据本公开的SMRT-seq和亚硫酸氢盐测序定量的甲基化水平的平均值,并且y轴指示通过根据本公开的SMRT-seq和亚硫酸氢盐测序定量的甲基化水平的差异。两个测量结果之间的中值差异为-6.85%(范围:-10.1%-1.7%)。通过本公开定量的甲基化水平相对于通过亚硫酸氢盐测序定量的值的中值百分比变化为-9.96%(范围:-14.76%-3.21%)。差异根据平均值而变化。两个度量的平均值越高,偏差就越大。

[0467] 图86C示出了与图86B相同的数据,但不同之处在于甲基化水平除以两个甲基化水平的平均值。图86C还示出了随着两个量度的平均值越高,偏差就越大。

[0468] 误差可能与亚硫酸氢盐测序有关并且与通过SMRT-seq的方法无关。据报道,传统的全基因组亚硫酸氢盐测序(Illumina)引入了显著偏差的序列输出并且高估了总体甲基化,其在定量特定基因组区域的方法之间的甲基化水平时存在显著差异(Olova等人《基因组生物学》2018;19:33)。本文公开的实施例具有许多示例性优点,由此其可以在不进行亚硫酸氢盐转化的情况下执行,所述亚硫酸氢盐转化会急剧降解DNA,并且可以在不进行PCR

扩增的情况下执行。

[0469] 3.组织起源

[0470] 根据本公开的实施例使用单分子实时测序(SMRT-seq,太平洋生物科学公司)执行跨各种癌症类型的甲基化分析。用于SMRT-seq的癌症类型包含但不限于结直肠癌(n＝3)、食道癌(n＝2)、乳腺癌(n＝2)、肾细胞癌(n＝2)、肺癌(n＝2)、卵巢癌(n＝2)、前列腺癌(n＝2)、胃癌(n＝2)和胰腺癌(n＝1)。还包含其匹配的邻近非肿瘤组织以进行SMRT-seq。数据集由通过Sequel II测序试剂盒2.0制备的DNA产生。

[0471] 图87A和87B示出了各种肿瘤组织与配对的邻近非肿瘤组织之间的总体甲基化水平的比较。百分比形式的甲基化水平在y轴上。在图87A中,通过SMRT-seq定量甲基化水平。在图87B中,通过亚硫酸氢盐测序定量甲基化水平。组织(即肿瘤组织或邻近的非肿瘤组织)的类型在x轴上。不同的符号表示不同的起源组织。

[0472] 图87A示出了肿瘤组织(包含乳腺癌、结肠直肠癌、食道癌、肝癌、肺癌、卵巢癌、胰腺癌、肾细胞癌和胃癌)的总体甲基化水平显著低于相应的非肿瘤组织(分别包含乳腺、结肠、食管、肝、肺、卵巢、胰腺、前列腺、肾和胃等)(P值＝0.006,配对的样本威尔克森(Wilcoxon)符号等级测试)。肿瘤与配对的非肿瘤组织之间甲基化水平的中值差异为-2.7%(IQR:-6.4%～-0.8%)。

[0473] 图84B证实了肿瘤组织中较低的甲基化水平。因此,这些结果表明,可以通过根据本公开的SMRT-seq准确地测定跨越各种癌症类型和组织的甲基化谱式,这意味着本公开在组织活检的基础上广泛应用于癌症的早期检测、预后、诊断和治疗。跨各种肿瘤类型的甲基化水平降低的不同程度可能表明甲基化谱式与癌症类型相关,从而可以确定癌症起源的组织。

[0474] D.增强检测和其它技术

[0475] 在一些实施例中,可以使用以下参数中的一个或多个来执行碱基修饰(例如,甲基化)的分析:序列上下文、IPD和PW。IPD和PW可以通过测序反应测定,而无需与参考基因组进行比对。单分子实时测序方法的各方面可以进一步增强确定序列上下文、IPD和PW的准确性。一方面是环状一致性测序的性能,其中测序模板的特定部分可以被多次测量,因此允许通过这些多个读数基于平均值或分布来测量序列上下文、IPD和PW。在某些实施例中,在没有比对过程的情况下对碱基修饰进行分析可以提高计算效率、减少周转时间并且可以减少分析成本。尽管可以在没有比对过程的情况下执行实施例,但是在其它实施例中,可以使用比对过程并且所述比对过程可能是优选的,例如,如果使用比对过程来确定检测到的碱基修饰的临床或生物学意义(例如,如果肿瘤抑制剂为高甲基化);或者,如果使用比对过程来选择对应于某些所关注的基因组区域的测序数据的子集以进行进一步分析。对于期望来自选定基因组区域的数据的实施例,这些实施例可能需要使用一种或多种酶或基于酶的方法来靶向此类区域,所述方法可以在基因组的所关注的区域(例如,限制酶或CRISPR-Cas9系统)中切割。CRISPR-Cas9系统可能优于基于PCR的方法,因为PCR扩增通常不会保留有关DNA碱基修饰的信息。可以分析此类(生物信息学[例如,通过比对]或通过如CRISPR-Cas9等方法)选定的区域的甲基化水平以提供关于组织起源、胎儿病症、妊娠疾病和癌症的信息。

[0476] 1.不与参考基因组比对的情况下使用子读段进行甲基化分析

[0477] 在实施例中,可以在不与参考基因组比对的情况下使用测量窗口来执行甲基化分

析,所述测量窗口包括来自子读段的动力学特征和序列上下文。如图88所示,源自零模波导(ZMW)的子读段用于构建一致性序列8802(还被称为环状一致性序列,CCS)。计算了CCS中每个位置的平均动力学值,包含但不限于PW值和IPD值。基于所述CpG位点的上游和下游序列根据CCS确定了CpG位点周围的序列上下文。因此,将构造如本公开中所定义的测量窗口用于训练,其中测量窗口包含根据具有相对于CCS的动力学特征的子读段的PW、IPD值和序列上下文。此程序避免了子读段与参考基因组的比对。

[0478] 为了测试图88中所示的原理,使用了601,942个源自全基因组扩增DNA的未甲基化CpG位点和163,527个源自经过CpG甲基转移酶(例如M.SssI)处理的DNA的甲基化CpG位点,从而形成训练数据集。使用了546,393个源自全基因组扩增DNA的未甲基化CpG位点和193,641个源自经过CpG甲基转移酶(例如M.SssI)处理的DNA的甲基化CpG位点,从而形成测试数据集。数据集由通过Sequel II测序试剂盒2.0制备的DNA产生。

[0479] 如图89所示,在一个实施例中,使用动力学特征和与子读段和CCS相关联的序列上下文来训练卷积神经网络(CNN)模型以测定甲基化,可以分别获得0.94和0.95的AUC值,用于区分测试和训练数据集中的甲基化CpG位点与未甲基化CpG位点。在其它实施例中,可以使用其它神经网络模型、深度学习算法、人工智能和/或机器学习算法。

[0480] 如果将甲基化概率的截止值设定为0.2,则在检测甲基化CpG位点时可以获得82.4%的灵敏度和91.7%的特异性。这些结果说明,可以使用具有动力学特征的子读段来区分甲基化和未甲基化CpG位点,而无需事先与参考基因组进行比对。

[0481] 在另一个实施例中,为了测定跨CpG位点的甲基化状态,还可以直接从子读段中使用动力学特征以及序列上下文而无需CCS信息和事先与参考基因组进行比对。使用了动力学特征(包含在跨越子读段中存在的CpG上游的20-nt和下游的20-nt的位置的PW值和IPD值)来训练CNN模型以测定甲基化状态。如图90所示,根据本公开的实施例,在训练数据集和测试数据集中检测甲基化CpG位点,使用与子读段相关的动力学特征的ROC曲线的AUC分别为0.70和0.69。这些数据表明,使用本公开的实施例使用与子读段相关的动力学特征来推断DNA分子的甲基化谱式是可行的,但无需事先比对和构建一致性序列。然而,在此实施例中测定甲基化的性能不如在本公开中组合利用比对信息或一致性序列的实施例。可以设想的是,产生子读段和动力学值的增强的精度将改善使用子读段和其相关动力学特征测定碱基修饰的性能。

[0482] 2.使用靶向的单分子实时测序对缺失区域进行甲基化分析

[0483] 本文描述的方法还可以用于分析一个或多个选定基因组区域。在一个实施例中,可以首先通过杂合方法来富集一个或多个所关注的区域,所述杂合方法允许来自一个或多个所关注的区域的DNA分子与具有互补序列的合成寡核苷酸杂合。对于使用本文所述方法进行的碱基修饰分析,靶DNA分子在经受测序之前不能通过PCR进行扩增,因为原始DNA分子中的碱基修饰信息不会转移到PCR产物中。已经开发了若干方法以在不执行PCR扩增的情况下富集这些靶区域。

[0484] 在另一个实施例中,可以通过使用CRISPR-Cas9系统来富集一个或多个靶区域(Stevens等人《公共科学图书馆:综合(PLOS One)》2019;14(4):e0215441;Watson等人《实验室研究(Lab Invest)》2020;100:135-146)。在一个实施例中,首先将DNA样本中DNA分子的末端去磷酸化,以使其不易直接连接至测序衔接子。然后,一个或多个所关注的区域被

Cas9蛋白和向导RNA(crRNA)引导以形成双链切口。然后将两侧都带有双链切口的一个或多个所关注的区与所选测序平台指定的测序衔接子连接。在另一个实施例中,可以用核酸外切酶处理DNA,使得未被Cas9蛋白结合的DNA分子被降解(Stevens等人《公共科学图书馆:综合》2019;14(4):e0215441)。由于这些方法不涉及PCR扩增,因此可以对具有碱基修饰的原始DNA分子进行测序,并测定碱基修饰。在一个实施例中,此方法可以用于靶向许多共享同源序列的区域,例如长散布核元件(LINE)重复序列。在一个实例中,此类分析可以用于分析母本血浆中循环的细胞游离DNA,以检测胎儿非整倍体(Kinde等人《公共科学图书馆:综合》2012;7(7):e41162。

[0485] 如图91所示,靶向单分子实时测序可以通过使用CRISPR(聚集的规则间隔的短回文重复序列)/Cas9(CRISPR相关蛋白质9)系统来实现。将携带5'-磷酰基(即5'-P)和3'羟基(即3'-OH)的DNA片段(例如分子9102)进行末端封闭处理,从而去除5'-P并将3'-OH与双脱氧核苷酸(即ddNTP)连接。因此,末端已被修饰的所得分子(例如分子9104)不能与衔接子连接用于随后的DNA文库制备。然而,末端封闭的分子经受了CRISPR/Cas9系统介导的靶标特异性裂解,从而将5'-P和3'-OH末端引入到所关注的分子。携带5'-P和3'-OH末端的此类新裂解的DNA分子(例如,分子9106)获取了与发夹衔接子连接形成环状分子9108的能力。未经连接的衔接子、线性DNA和仅携带一个切割位点的分子用核酸外切酶III和VII进行消化。作为结果,与两个发夹衔接子连接的分子被富集,并经受单分子实时测序。根据本公开中存在的实施例,这些靶分子适合于碱基修饰分析(即,靶向单分子实时测序)。

[0486] 如图92所示,CRISPR/Cas9系统中的Cas9蛋白与向导RNA(即gRNA)相互作用,所述向导RNA包含CRISPR RNA(crRNA,负责DNA靶向)和反式激活crRNA(tracrRNA,负责与Cas9形成复合物)(Pickar-Oliver等人《自然评论:分子细胞生物学(Nat Rev Mol Cell biol.)》2019;20:490-507)。曲线形状表示Cas9蛋白,这是一种使用CRISPR序列作为向导来识别和切割与CRISPR序列的一部分互补的特定DNA链的酶。crRNA退火到tracrRNA。在一个实施例中,合成单个RNA序列包含crRNA和tracrRNA序列,其被称为单向导RNA(sgRNA)。crRNA中的片段(被称为间隔子序列)将引导Cas9蛋白通过与目标区域的互补碱基配对来识别和切割双链DNA(dsDNA)的特定链。在一个实施例中,间隔子序列与靶向dsDNA之间的互补性不涉及错配。在另一个实施例中,间隔子序列和靶向dsDNA之间的互补碱基配对将允许错配。例如,错配的数量是但不限于1、2、3、4、5、6、7、8等。在一个实施例中,CRISPR序列将是可编程的,这取决于切割效率、特异性、灵敏度以及不同CRISPR/Cas复杂设计的复用能力。

[0487] 如图93所展示的,设计了一对靶向跨越人类基因组中Alu元件的两个切口的CRISPR/Cas9复合物。"XXX"指示Cas9核酸酶切割位点两侧的三个核苷酸。"YYY"指示与"XXX"互补的三个对应的核苷酸。5'-NGG表示前间隔子序列邻近基序(PAM)序列。在其它CRISPR/Cas系统中,PAM序列可以不同,并且Cas核酸酶切割位点两侧的序列可以不同。在此图中,Alu区域的长度为223bp。有1,175,329个Alu区域,每个区域都包含人类基因组中此类Alu元素的同源物。此Alu元素中有中值为5个的CpG位点(范围:0-34)。作为实例,此设计包含36-nt crRNA,其包含20-nt间隔子序列。详细的gRNA序列信息如下所示:

[0488] 用于引入第一切口的第一CRISPR/Cas9复合物:(所有5'到3'的序列)

[0489] crRNA:GCCUGUAAUCCCAGCACUUUGUUUUAGAGCUAUGCU

[0490] tracrRNA:

[0491] AGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUU

[0492] 用于引入第二切口的第二CRISPR/Cas9复合物：

[0493] crRNA:AGGGUCUCGCUCUGUCGCCCGUUUUAGAGCUAUGCU

[0494] tracrRNA：

[0495] AGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUU

[0496] crRNA分子退火到tracrRNA(如67-nt)以形成gRNA的主链。具有经过设计的gRNA的Cas9核酸酶可以以一定水平的特异性切割含有靶向切割位点的末端封闭分子的两条链。人类基因组中有116,184个所关注的Alu区域,所述区域应该由经过设计的CRISPR/Cas9复合物切割。因此,Cas9复合物靶向切割后的那些Alu区域可以与发夹衔接子连接。与发夹衔接子连接的那些分子可以通过单分子实时测序进行测序。可以以靶向方式测定那些Alu区域的甲基化谱式。在一个实施例中,来自两个Cas9复合物的间隔子序列可以与双链DNA底物的相同链(例如沃森链或克里克链)碱基配对。在一个实施例中,来自两个Cas9复合物的gRNA中的间隔子序列可以与双链DNA底物的不同链碱基配对。例如,Cas9复合物中的一个间隔子序列与双链DNA底物的沃森链互补,并且Cas9复合物中的另一个间隔子序列与双链DNA底物的克里克链互补,反之亦然。

[0497] 在一个实施例中,与发夹衔接子连接的DNA分子呈环状形式,其将抵抗核酸外切酶消化。因此,可以用核酸外切酶(例如核酸外切酶III和VII)处理与衔接子连接的DNA产物,以除去线性DNA(例如脱靶DNA分子)。使用核酸外切酶的这一步骤可以进一步富集靶分子。要测序的靶分子的长度取决于由一种或多种Cas9核酸酶引入的两个切割位点之间的跨度长度,例如,包含但不限于10bp、20bp、30bp、40bp、50bp、100bp、200bp、300bp、400bp、500bp、1000bp、2000bp、3000bp、4000bp、5000bp、10kb、20kb、30kb、40kb、50kb、100kb、200kb、300kb、500kb和1Mb。

[0498] 作为实例,使用具有靶向Alu区域的gRNA的Cas9,使用单分子实时测序对来自人肝细胞癌(HCC)肿瘤组织样本的187,010个分子进行测序。其中,113,491个分子携带靶向切口(即,目标切割率约为分子的60.7%)。数据集由通过Sequel II测序试剂盒2.0制备的DNA产生。换句话说,在此实例中,由Cas9复合物引入到所关注的分子的切割位点的特异性为60.7%。在其它实施例中,由Cas9或其它Cas复合物引入到所关注的分子的切割位点的特异性可以变化,包含但不限于1%、5%、10%、20%、30%、40%、50%、60%、70%、80%、90%和100%。IPD值、PW值和来源于未与参考基因组比对的CCS和子读段的序列上下文用于测定Alu序列中CpG位点的甲基化状态。

[0499] 如图94所示,观察到由亚硫酸氢盐测序和根据本公开的单分子实时测序测定的甲基化水平之间的类似甲基化分布。图94示出了亚硫酸氢盐测序和单分子实时测序(太平洋生物科学公司)的甲基化密度(百分比)的直方图。y轴指示具有x轴上示出的特定甲基化密度的样本中分子的比例。此结果表明,使用Cas9介导的靶向单分子实时测序来测定甲基化谱式是可行的。此结果还表明,可以使用子读段相关的动力学特征(包含PW值和IPD值)来测定甲基化,而无需与参考基因组进行比对。如图94所示,观察到相当数量的Alu区域显示出低甲基化,这与癌症基因组将在Alu重复区域去甲基化的现有知识一致(Rodriguez等人《核酸研究》2008;36:770-784)。

[0500] 图95示出了由根据本公开的单分子实时测序测定的甲基化水平在y轴上的分布和

由亚硫酸氢盐测序测定的甲基化密度在x轴上的分布。如图95所示，根据亚硫酸氢盐测序的结果，Alu区域的甲基化水平被分为5类，即0-20％、20％-40％、40％-60％、60％-80％和80％-100％。我们的模型使用测量窗口进一步测定了同一组Alu区域的甲基化水平，所述测量窗口包含每个Alu区域类别的动力学特征和序列上下文(y轴)。由我们的模型测定的甲基化水平分布根据跨分类类别的甲基化水平的升序逐渐增加。同样，这些结果表明，使用Cas9介导的靶向单分子实时测序来测定甲基化谱式是可行的。可以使用包含PW值和IPD值的子读段相关动力学特征来测定甲基化而无需与参考基因组进行比对。

[0501] 在仍另一个实施例中，可以使用其它类型的CRISPR/Cas系统(例如但不限于Cas12a、Cas3和其它直系同源物(例如金黄色葡萄球菌Cas9)或工程化Cas蛋白(增强的氨基酸球菌属Cas12a))来执行靶向单分子实时测序。

[0502] 在一个实施例中，可以使用不具有核酸酶活性的失活的Cas9(dCas9)来富集靶分子而无需裂解。例如，靶向DNA分子由包括生物素化dCas9和靶序列特异性gRNA的复合物结合。dCas9可能不会切割此类靶向DNA分子，因为dCas9缺乏核酸酶。通过使用链霉亲和素包被的磁珠，可以富集靶向DNA分子。

[0503] 在一个实施例中，可以在用Cas蛋白温育后使用核酸外切酶来消化DNA混合物。核酸外切酶可以降解未结合Cas蛋白的DNA分子，而核酸外切酶可以不降解或在降解与Cas蛋白结合的DNA分子方面效率较低。因此，有关由Cas蛋白结合的靶分子的信息可以在最终的测序结果中进一步富集。

[0504] 图96示出了组织和组织中Alu区域的甲基化水平的表格。许多组织示出甲基化水平在85-92％的范围内，包含88％到92％的范围。HCC肿瘤组织和胎盘组织示出了甲基化水平低于80％。如图96所示，示出了HCC肿瘤在我们的设计所靶向的Alu区域中经常被低甲基化。因此，本公开中呈现的Alu区域的甲基化测定可以用于在肿瘤进展或治疗期间使用从肿瘤活检或其它组织或细胞中提取的DNA来进行检测、分期和监测癌症。

[0505] 跨Alu区域的胎盘组织的低甲基化可以用于使用孕妇的血浆DNA执行无创性产前检测。例如，较高程度的低甲基化可能指示孕妇的胎儿DNA浓度较高。在另一个实例中，如果女性怀有染色体非整倍的胎儿，则通过此方法检测到的源自患病染色体的Alu片段的数量可能与怀有整倍体胎儿的女性在数量上有所不同(即增加或减少)。因此，如果胎儿具有21三体综合症，则与怀有整倍体胎儿的女性相比，通过此方法检测到的源自染色体21的Alu片段的数量可能会增加。另一方面，如果胎儿具有单体染色体，则与怀有整倍体胎儿的女性相比，通过此方法检测到的源自所述染色体的Alu片段的数量可能会减少。与未受影响的染色体相比，测定血浆中受影响的染色体(13、18或21)的额外低甲基化的表示可以用作区分怀有正常胎儿和异常胎儿的女性的分子指标。

[0506] 3.Cas9复合体针对不同类型癌症所靶向的Alu区域的甲基化分析

[0507] 即使靶向的Alu重复序列在不同组织中高度甲基化，假设不同的癌症类型将跨那些Alu重复序列含有不同的去甲基化谱式。在一个实施例中，可以根据本文公开的内容使用基于Cas9的靶向单分子实时测序来分析甲基化谱式以确定不同的癌症类型。

[0508] 图97示出了与不同癌症类型的Alu重复序列相关的甲基化信号的聚类分析。来自TCGA数据库(www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga)的癌症受试者使用微阵列技术(Infinium HumanMethylation450

BeadChip,Illumina公司)分析了CpG位点的甲基化状态。分析了微阵列芯片中存在并与CRISPR/Cas9复合物靶向的Alu区域重叠的3,024个CpG位点的甲基化状态。患者体内有许多源自所关注的Alu区域的CpG。每个CpG的甲基化水平通过微阵列(还被称为甲基化指数或β值)进行定量。基于患者中那些CpG位点的甲基化水平进行了分级聚类分析。因此,在那些CpG位点具有类似甲基化水平的患者会聚在一起,从而形成分化枝。不同患者之间甲基化谱式的类似性将通过聚类树状图中的高度值来指示。在此实例中,根据欧几里得(Euclidean)距离来计算高度。在其它实施例中,将使用其它距离度量,包含但不限于闵可夫斯基(Minkowski)距离、切比雪夫(Chebychev)距离、马哈拉诺比斯(Mahalanobism)距离、曼哈顿(Manhattan)距离、余弦距离、相关距离、斯皮尔曼(Spearman)距离、汉明(Hamming)距离、杰卡德(Jaccard)距离等。本文中使用的高度表示簇之间距离度量的值,其反映出簇之间的相关性。例如,如果观察到两个簇在高度x处合并,则表明这些集群之间的距离为x(例如,所有集群间患者之间的平均距离)。

[0509] 通过使用CpG位点的甲基化状态,患者根据聚类分析结果中的癌症类型被分为不同的独特的组。癌症类型包含膀胱尿路上皮癌(BLCA)、浸润性乳腺癌(BRCA)、卵巢浆液性囊腺癌(OV)、胰腺癌(PAAD)、HCC、肺腺癌(LUAD)、胃腺癌(STAD)、皮肤黑色素瘤(SKCM)和子宫癌肉瘤(UCS)。图中癌症类型之后的数字表示患者。因此,聚类表明,所选择的Alu重复序列中的甲基化信号为癌症类型的分类提供有用信息,包含图97中未示出的癌症类型。在一个实施例中,可以基于组织活检中的甲基化谱式来区分原发性和继发性肿瘤。

[0510] 4.子读段深度和长度截止值

[0511] 本节示出了子读段深度和/或长度截止值可以用于提高甲基化检测的准确性和/或效率。为了测试某些子读段深度或长度,可以修改文库构建。

[0512] 在Sequel II测序试剂盒2.0的基础上,分析了读段深度对测试数据集中整体甲基化水平定量的影响,所述测试数据集是从全基因组扩增或M.SsssI处理后的样本中产生的。研究了具有至少一定截止值的子读段覆盖的基因组位点,所述截止值例如但不限于≥1x、10x、20x、30x、40x、50x、60x、70x、80x、90x、100x等。

[0513] 图98A示出了在涉及全基因组扩增的测试数据集中读段深度对总体甲基化水平量化的影响。图98B示出了在涉及M.SsssI处理的测试数据集中读段深度对总体甲基化水平量化的影响。y轴示出了以百分比表示的总体甲基化水平。x轴示出了读段深度。虚线指示总体甲基化水平的预期值。

[0514] 如图98A所示,对于涉及全基因组扩增的数据集,总体甲基化在如但不限于1x、10x、20x、40x、50x等最初的几个截止值中下降,范围为5.7%到5.2%。甲基化水平在50x或更高的截止值下逐渐稳定在5%左右。

[0515] 另一方面,在图98B中,对于从M.SsssI处理后的样本中产生的数据集,总体甲基化在如但不限于1x、10x、20x、40x、50x等最初的几个截止值中增加,范围为70%到83%。甲基化水平在50x或更高的截止值下逐渐稳定在83%左右。

[0516] 在一个实施例中,可以调整子读段深度截止值,使得碱基修饰分析的性能适合不同的应用。在其它实施例中,可以使用较不严格的子读段深度截止值来获得更多适合下游分析的ZMW(即分子数)。在仍另一个实施例中,可以将由根据本公开的SMRT-seq测定的甲基化水平的读数校准为第二测量结果,例如但不限于BS-seq、数字微滴PCR(亚硫酸氢盐转化

的样本上)、甲基化-特异性PCR或甲基化胞嘧啶结合抗体或其它蛋白质。在另一个实施例中,通过对5mC保留的全基因组扩增后的DNA分子进行BS-seq、数字微滴PCR(亚硫酸氢盐转化的样本上)、甲基化特异性PCR或甲基-CpG结合结构域(MBD)蛋白质富集的基因组测序(MBD-seq)来获得第二测量结果。作为实例,可以通过DNA引发酶TthPrimPol、聚合酶phi29和DNMT1(DNA甲基转移酶1)来介导5mC保留的全基因组扩增。

[0517] 分析了不同子读段深度跨不同癌症类型和非肿瘤组织的甲基化水平。还将根据本公开通过SMRT-seq测定的甲基化水平与BS-seq测序结果进行了比较。使用Sequel II测序试剂盒2.0,获得了中值为4,300万的子读段(四分位距(IQR):3,000-5,200万),从而产生了与中值为460万个的与人类参考基因组(IQR:280-580万)比对的环形一致性序列(CCS)。在那些样本中,还对22个样本进行了良好建立的大规模并行亚硫酸氢盐测序(BS-seq)以测定甲基化谱式,从而为比较甲基化水平提供了第二测量结果。

[0518] 图99示出了由根据本公开的SMRT-seq(Sequel II测序试剂盒2.0)测定的和使用不同的子读段深度截止值的BS-seq测定的总体甲基化水平之间的比较。y轴示出了由SMRT-seq测定的百分比形式的甲基化水平。x轴示出了通过亚硫酸氢盐测序测定的百分比形式的甲基化水平。这些符号指示1x、10x和30x的不同子读段深度。三条对角线表示了不同子读段深度的拟合线。

[0519] 图99示出了当分析至少一次被子读段覆盖的基因组位点(即子读段深度截止值≥1x)时,通过根据本公开的SMRT-seq测定的CpG位点的甲基化水平与通过BS-seq测定的甲基化水平具有良好的相关性(r=0.8;P值<0.0001)。这些结果表明,本公开中呈现的实施例可以用于测量不同组织类型的甲基化水平,所述不同组织类型包含但不限于结肠直肠癌、结肠直肠组织、食道癌、食道组织、乳腺癌、非癌性乳腺癌组织、肾细胞癌、肾组织、肺癌和肺组织。还观察到,随着子读段深度截止值增加到10x和30x,这两个测量结果之间的相关性分别提高到0.87(P值<0.0001)和0.95(P值<0.0001)。在一些实施例中,子读段深度的增加或覆盖更多子读段的基因组区域的选择将提高根据本公开的基于SMRT-seq的甲基化测定的性能。

[0520] 图100是表格,其示出了子读段深度对SMRT-seq(Sequel II测序试剂盒2.0)和BS-seq的两个测量结果之间的甲基化水平的相关性的影响。第一列示出了子读段深度截止值。第二列示出了相关系数皮尔森氏r(Pearson's r)。第三列示出了与截止值相关联的CpG位点的数量,其中位点数量的范围在括号中示出。

[0521] 如图100所示,SMRT-seq与BS-seq的两个测量结果之间的甲基化水平的相关性根据不同的子读段深度截止值而不同。在一个实施例中,可以利用子读段深度截止值与两个测量结果之间的相关系数(例如,皮尔森相关系数)之间的关系来确定用于区分甲基化胞嘧啶与未甲基化胞嘧啶的子读段深度的最佳截止值。图100示出了在30x的子读段深度截止值下,(即≥30x),通过根据本公开的SMRT-seq测量的甲基化水平与通过BS-seq产生的结果具有最高的相关性(皮尔森氏r=0.952)。在其它实施例中,可以使用但不限于1x、10x、30x、40x、50x、60x、70x、80x、900x、100x、200x、300x、400x、500x、600x、700x、800x等的子读段深度截止值。

[0522] 用于甲基化分析的CpG位点的数量随着子读段深度截止值的增加而减少,如图100所示。对于100x的子读段深度截止值,与30x的子读段深度截止值相比(皮尔森氏r=

0.952)，观察到两个甲基化水平测量结果之间的相关性较低(皮尔森氏r＝0.875)。较高的子读段截止值的较低相关性可以归因于满足更严格的子读段深度截止值的CpG位点数量较少。在一个实施例中，可以考虑子读段深度的要求与可以用于甲基化分析的分子数量之间的权衡。例如，如果旨在扫描整个基因组的甲基化谱式，则更多的分子可能是期望的。如果使用靶向SMRT-seq聚焦于特定区域，则可能期望较高的子读段深度来获得所述区域的甲基化谱式。

[0523] 图101示出了由Sequel II测序试剂盒2.0产生的数据中关于片段长度的子读段深度分布。子读段深度在y轴上示出，并且DNA分子的长度在x轴上示出。DNA分子的长度根据环状一致性序列(CCS)的长度推导得出。

[0524] 由于子读段深度可能影响使用SMRT-seq数据的甲基化测定的性能并且子读段深度是被测序的DNA分子长度的函数，因此DNA分子的长度对于获得用于分析样本中甲基化谱式的最佳子读段深度可能是至关重要的。如图101所示，DNA越长，子读段深度就越小。例如，对于长度为1kb的分子群体，中值子读段深度为50x。对于长度为10kb的分子群体，中值子读段深度为15x。

[0525] 在一个实施例中，如图100所示，子读段深度的最佳截止值可以为至少30x，所述值产生最高的相关系数。为了进一步提高满足30x的最佳子读段深度截止值的分子的通量，可以使用子读段深度与DNA模板分子长度之间的关系。例如，在图101中，30x是长度为约4kb的分子的中值子读段深度。因此，可以在SMRT-seq文库制备之前将4kb DNA分子分级并且然后将测序限制为4kb DNA分子。在其它实施例中，可以使用用于DNA分子分级的其它长度截止值，包含但不限于100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、2kb、3kb、4kb、5kb、6kb、7kb、9kb、10kb、20kb、30kb、40kb、50kb、60kb、70kb、80kb、90kb、100kb、500kb、1Mb或不同的长度截止值组合。

[0526] 5.基于限制酶的靶向单分子实时测序

[0527] 本节描述了使用限制酶来提高修饰检测的实用性和/或通量和/或成本效益。用限制酶产生的DNA片段可以用于确定样本的来源。

[0528] a)用限制酶消化DNA分子

[0529] 在实施例中，可以在单分子实时测序(例如使用太平洋生物科学公司系统)之前使用一种或多种限制酶来消化DNA分子。因为限制酶识别位点的分布在人类基因组中是不均匀的，所以由限制酶消化的DNA可能会产生倾斜的长度分布。具有较多限制酶识别位点的基因组区域可以被消化成较小的片段，而具有较少限制酶识别位点的基因组区域可以被消化成较长的片段。在实施例中，根据长度范围，可以选择性地获得源自具有一个或多个限制酶的类似切割谱式的一个或多个区域的DNA分子。长度选择的期望长度范围可以通过一种或多种限制酶的计算机模拟切割分析来确定。可以使用计算机程序来确定参考基因组(例如人参考基因组)中感兴趣的限制酶识别位点的数量。根据那些识别位点，这种参考基因组在计算机模拟中被剪切成片段，这提供了感兴趣的基因组区域的尺寸信息。

[0530] 图126示出了使用DNA末端修复和A-加尾的基于MspI的靶向单分子实时测序的方法。在实施例中，如图126所示，可以使用识别5'C^CGG3'位点的MspI来消化生物体的DNA样本，例如但不限于人DNA样本。使具有5'CG突出端的经过消化的DNA片段经受长度选择，从而富集源自CpG岛的DNA分子。富含G和C残基(也被称为GC内含物)的基因组区域可以产生较短

的片段。因此，可以基于所关注区域的GC内含物来确定用于执行选择的片段长度的范围。本领域的技术人员可获得多种DNA片段长度选择工具，包含但不限于凝胶电泳、尺寸排阻电泳、毛细管电泳、色谱、质谱、过滤方法、基于沉淀的方法、微流体学和纳米流体。使长度分级的DNA分子经受DNA末端修复和A-加尾，使得期望的DNA产物可以与携带5'T突出端的发夹衔接子连接，从而形成环状DNA模板。

[0531] 在例如但不限于使用核酸外切酶(例如，核酸外切酶III和核酸外切酶VII)去除未经连接的衔接子、线性DNA和非完整环状DNA后，与发夹衔接子连接的DNA分子可以用于单分子实时测序以在测定如本文所公开的甲基化谱时确定IPD、PW和序列上下文。通过分析富含CpG的基因组区域，可以区分从不同组织或患有不同疾病和/或生理病状的组织或生物样本中获得的DNA，并通过由本公开的测序数据分析方法测定的甲基化谱对其进行分类。

[0532] 对于图126中涉及长度选择的步骤，在实施例中，期望的长度范围可以通过MspI的计算机模拟切割分析来确定。在人类参考中测定了总共2,286,541个MspI切割位点。人类参考基因组根据那些MspI切割位点在计算机模拟中被剪切成片段。获得了总共2,286,565个片段。通过所述片段的核苷酸总数测定了每个单独片段的长度。

[0533] 图127A和127B示出了MspI消化片段的长度分布。这些图的y轴是特定长度的片段的频率百分比。图127A的x轴的对数标度范围为50到500,000bp。图127B的x轴的线性标度范围为50到1,000bp。

[0534] 如图127A和127B所示，MspI消化的DNA分子具有倾斜的长度分布。MspI消化的片段的中值长度为404bp(IQR:98-1,411bp)。约53％的那些MspI消化的片段小于1kb。长度图谱中存在一系列可能由重复元素引起的尖峰。某些重复元件可能共享类似的MspI切割位点谱式，从而产生具有类似的片段长度的来源于MspI消化的一组分子。例如，具有最高频率的尖峰(即总共49,079个)对应于64bp的长度。其中有45,894(94％)个与Alu重复序列重叠。可以选择长度为64bp的DNA分子以富集源自Alu重复序列的DNA分子。数据表明，长度选择可以用于富集期望的DNA分子以进行根据本公开的下游甲基化分析。

[0535] 图128示出了具有某些选定长度范围内的DNA分子的数量的表格。第一列以碱基对示出了长度范围。第二列示出了长度范围内的分子相对于总片段的百分比。第三列示出了与CpG岛重叠的长度范围内的分子的数量。第四列示出了与CpG岛重叠的长度范围内的分子的百分比。第五列示出了被测序的CpG位点的数量。第六列示出了落入CpG岛内的CpG位点的数量。第七列示出了通过长度选择靶向的并落入在CpG岛内的CpG位点的百分比。如图128所示，从经受MspI消化的人类基因组中产生的DNA分子的量根据所讨论的不同长度范围而变化。与CpG岛重叠的DNA分子的数量随不同长度范围而变化。

[0536] 由于CCGG基序优先出现在CpG岛中，选择长度小于某个截止值的分子可以允许富集源自CpG岛的DNA分子。例如，对于50到200bp的长度范围，分子的数量为526,543，这占来源于经受MspI消化的人类基因组的总DNA片段的23.03％。在526,543个DNA分子中，有104,079个(19.76％)分子与CpG岛重叠。对于600到800bp的长度范围，分子的数量为133,927，这占来源于经受MspI消化的人类基因组的总DNA片段的5.86％。在133,927个分子中，有3,673个(2.74％)分子与CpG岛重叠。作为实例，可以选择50到200bp的长度来富集源自CpG岛的DNA片段。

[0537] 为了通过基于MspI的靶向单分子实时测序计算与CpG岛重叠的CpG位点的富集程

度,对通过超声处理剪切的DNA进行了模拟,基于正态分布模拟了从中值长度为200bp且标准偏差为20bp的ZMW产生的526,543个片段。只有0.88%的DNA分子与CpG岛重叠。总共71,495个CpG位点与CpG岛重叠。如图128所示,选择50到200bp范围内的MspI消化的片段将导致19.8%的片段与CpG岛重叠。因此,这些数据表明,与通过超声处理制备的DNA相比,通过MspI消化制备的DNA可能具有22.5倍的源自CpG岛的DNA片段的富集。此外,通过MspI消化分析了CpG岛中被富集的CpG位点。选择50到200bp范围内的MspI消化的片段可能产生885,041个与CpG岛重叠的CpG位点,这占所述长度范围内来自经过测序的片段的CpG位点总数的37.5%。与通过超声处理制备的DNA相比,与CpG岛重叠的CpG位点富集了12.3倍(即885,041/71,495)。基于图128所示的信息,可以选择合适的长度范围以包含CpG位点的期望数量和CpG岛内的CpG位点的期望倍数富集。

[0538] 图129是限制酶消化之后CpG岛内的CpG位点的覆盖百分比对DNA片段的长度的图。y轴示出了被具有给定长度的片段覆盖的CpG岛内CpG位点的百分比。x轴示出了限制酶消化后DNA片段的长度范围的上限。图129示出了通过扩大长度选择范围覆盖的CpG岛内CpG位点的百分比。在图129中,长度范围为从50bp到x轴所示的长度。在其它实施例中,长度范围的下限可以被定制,例如但不限于60bp、70bp、80bp、90bp、100bp、200bp、300bp、400bp和500bp。对于通过增加长度上限来扩大长度范围,可以观察到CpG岛内CpG位点的覆盖百分比逐渐增加并且在65%下达到稳定。有些CpG位点没有被覆盖,因为其位于50bp以下的DNA片段内,或者其位于超长分子内的片段内(例如,>100,000bp)。

[0539] 在一些实施例中,可以使用两种或更多种不同的限制酶(具有不同的限制位点)来分析DNA样本,以便增加CpG岛内CpG位点的覆盖率。可以在单独的反应中用不同的酶消化DNA样本,使得每个反应中仅存在一种限制酶。例如,识别CGˆCG位点的AccII可以用于优先在CpG岛上进行切割。在其它实施例中,可以使用具有CG二核苷酸作为识别位点的一部分的其它限制酶。在人类基因组内,有678,669个AccII切割位点。使用AccII限制酶对人类参考基因组执行了计算机模拟切割,并且总共获得了678,693个片段。然后对这些片段执行了计算机模拟长度选择,并根据上述针对MspI消化的方法计算了CpG岛内CpG位点的覆盖率百分比。随着长度选择范围的扩大,可以观察到CpG位点覆盖率百分比逐渐增加。覆盖率百分比在约50%下稳定。CpG位点的覆盖率在来自两个酶消化实验(即MspI消化)和AccII消化的组合数据内进一步增加。通过选择长度为50bp到400bp的DNA片段,可以覆盖80%的CpG岛内的CpG位点。此百分比高于通过两种酶中的任一种单独进行的消化实验的相应数值。通过使用其它限制酶分析DNA样本,可以进一步增加覆盖率。如果将DNA样本分为两个等分试样。一个等分试样用MspI消化,并且另一个用AccII消化。将两个经过消化的DNA样本以等摩尔混合在一起,并使用单分子实时测序以500万个ZMW进行测序。根据计算机模拟分析,83%的CpG岛内的CpG位点(即1,734,345个)将按照环状一致性序列进行至少4次测序。

[0540] 图130示出了不使用DNA末端修复和A-加尾的基于MspI的靶向单分子实时测序。在实施例中,可以在不进行DNA末端修复和A-加尾过程的情况下执行经过消化的DNA分子与发夹衔接子之间的连接。可以用携带5'Cg突出端的发夹衔接子直接连接携带5'CG突出端的经过消化的DNA分子,从而形成单分子实时测序的环状DNA模板。在清除未经连接的衔接子和自连接的衔接子二聚体之后,并且在一些实施例中在去除未经连接的衔接子、线性DNA和非完整环状DNA之后,与发夹衔接子连接的DNA分子可能适合于单分子实时测序,以获得IPD、

PW和序列上下文。根据本公开,将使用IPD、PW和序列上下文测定单个分子的甲基化谱。

[0541] 图131示出了衔接子自连接的可能性降低的基于MspI的靶向单分子实时测序。潜在的胞嘧啶碱基指示不具有5'磷酰基的碱基。在一些实施例中,为了最小化在衔接子连接过程中可能形成自连接的衔接子二聚体的可能性,可以使用去磷酸化的发夹衔接子将衔接子与那些MspI消化的DNA分子的连接。那些去磷酸化的发夹衔接子可能由于缺乏5'磷酰基而不能形成自连接的衔接子二聚体。连接后,使产物经受衔接子清除步骤以纯化与发夹衔接子连接的DNA分子。与可能携带缺口的发夹衔接子连接的DNA分子进一步经受了磷酸化作用(例如T4多核苷酸激酶)并通过DNA连接酶(例如T4 DNA连接酶)进行缺口密封。在实施例中,可以进一步去除未经连接的衔接子、线性DNA和非完整环状DNA。与发夹衔接子连接的DNA分子适用于单分子实时测序,以获得IPD、PW和序列上下文。根据本公开,将使用IPD、PW和序列上下文测定单个分子的甲基化谱。

[0542] 除了MspI之外,还可以使用如SmaI等具有识别位点CCCGGG的其它限制酶。

[0543] 在一些实施例中,期望的长度选择过程可以在DNA末端修复步骤之后执行。在一些实施例中,当确定了发夹衔接子对长度选择结果的影响时,可以在发夹衔接子连接之后执行期望的长度选择过程。在这些和其它实施例中,基于MspI的靶向单分子实时测序所涉及的程序步骤的顺序可以根据实验情况而改变。

[0544] 在实施例中,将使用基于凝胶电泳和/或基于磁珠的方法进行长度选择。在实施例中,限制酶可以包含但不限于BgIII、EcoRI、EcoRII、BamHI、HindIII、TaqI、NotI、HinFI、PvuII、Sau3AI、SmaI、HaeIII、HgaI、HpaII、AluI、EcoRV、EcoP15I、KpnI、PstI、SacI、SalI、ScaI、SpeI、SphI、StuI、XbaI和其组合。

[0545] b)用甲基化区分生物样本类型

[0546] 本节描述了使用通过限制酶消化产生的片段测定的甲基化谱来促进区分不同的生物样本。

[0547] 根据本公开的实施例,使用由基于MspI的单分子实时测序测定的甲基化谱评估了生物样本之间甲基化谱的差异。以胎盘组织DNA和血沉棕黄层DNA样本为例。以基于MspI的靶向单分子实时测序为基础,执行了计算机模拟以生成关于胎盘和血沉棕黄层DNA样本的数据。模拟基于动力学值,所述动力学值包含先前使用Sequel II测序试剂盒1.0通过SMRT测序胎盘组织DNA和血沉棕黄层DNA到全基因组覆盖生成的每个核苷酸的IPD和PW。然后,模拟了胎盘DNA和血沉棕黄层DNA样本经受MspI消化、随后使用50到200bp的长度范围进行基于凝胶的长度选择的条件。选定的DNA分子与发夹衔接子连接以形成环状DNA模板。对环状DNA模板进行单分子实时测序以获得有关IPD、PW和序列上下文的信息。

[0548] 假设有500,000个ZMW产生SMRT测序子读段,则那些子读段遵循如表1所示的长度范围为50到200bp内的MspI消化片段的基因组分布。对于胎盘和血沉棕黄层DNA样本,假设子读段深度均为30x。分别对胎盘DNA样本和血沉棕黄层DNA样本重复了10次模拟。因此,通过MspI消化的靶向单分子实时测序在计算机模拟中产生的数据集总共包括10个胎盘DNA样本和10个血沉棕黄层DNA样本。通过CNN进一步分析了数据集,从而根据本公开测定每个样本的甲基化谱。从CpG岛获得了中值为9,198的CpG位点(范围:5,497-13,928),这占经过测序的CpG位点总数的13.6%(范围:45,304-90,762)。根据本公开,通过CNN模型测定了每个分子中每个CpG位点的甲基化状态。

[0549] 图132是通过基于MspI的靶向单分子实时测序测定的胎盘样本与血沉棕黄层DNA样本之间的总体甲基化水平的图。y轴是百分比形式的甲基化水平。样本类型在x轴上列出。图132示出了与血沉棕黄层样本相比(中值:69.5%;范围:68.9%-70.4%),胎盘样本中的总体甲基化水平(中值:57.6%;范围:56.9%-59.1%)较低(P值<0.0001,曼-惠特尼U检验)。这些结果表明,通过基于MspI的单分子实时测序测定的甲基化谱可以用于基于甲基化差异区分组织样本或生物样本。因为这些数据示出,由于基于MspI的单分子实时测序检测到的甲基化差异,胎盘DNA可以与血沉棕黄层DNA区别开,因此可以将此方法应用于母本血浆中胎儿DNA分率的测量。可以使用甲基化来测量胎儿DNA分率,因为母本血浆或母本血清中的胎儿DNA来自胎盘,而样本中的其余DNA分子则主要来源于母本血沉棕黄层细胞。在实施例中,此技术将是用于区分不同组织或具有不同疾病和/或生理病状或生物样本的组织的有用工具。

[0550] 为了使用CpG岛的甲基化谱在胎盘DNA样本与血沉棕黄层DNA样本之间执行聚类分析,使用归类为甲基化的CpG位点在所述CpG岛的总CpG位点中的比例计算了CpG岛的DNA甲基化水平。出于说明的目的,使用了CpG岛区域的甲基化水平来执行聚类分析。

[0551] 图133示出了使用由基于MspI的靶向单分子实时测序测定的胎盘样本和血沉棕黄层样本的DNA甲基化谱的胎盘样本和血沉棕黄层样本的聚类分析。聚类树状图中的高度值指示了不同患者之间CpG岛的甲基化谱式的类似性。在此实例中,根据欧几里得距离来计算高度。在一个实施例中,可以使用高度截止值100将聚类树切割成两组,从而以100%的灵敏度和特异性区分胎盘样本和血沉棕黄层样本。在其它实施例中,可以使用其它高度截止值,包含但不限于50、60、70、80、90、120、130、140和150等。图133示出了使用通过基于MspI的单分子实时测序根据本公开测定的CpG岛的甲基化谱清晰地将10个胎盘DNA样本和10个血沉棕黄层DNA样本分别聚集成两组。

[0552] V.训练和检测方法

[0553] 本节示出了训练用于检测碱基修饰的机器学习模型和使用机器学习模型检测碱基修饰的示例方法。

[0554] A.模型训练

[0555] 图102示出了检测核酸分子中核苷酸修饰的示例方法1020。示例方法1020可以是训练用于检测修饰的模型的方法。修饰可以包含甲基化。甲基化可以包含本文所述的任何甲基化。修饰可以具有离散状态,如甲基化和未甲基化,并潜在地指定甲基化的类型。因此,核苷酸可能有两种以上的状态(分类)。

[0556] 在框1022处,接收多个第一数据结构。数据结构的各种实例在此处描述,例如在图4-16中描述。第一多个第一数据结构中的每个第一数据结构可以对应于在多个第一核酸分子的相应核酸分子中测序的核苷酸的相应窗口。与第一多个数据结构相关联的每个窗口可以包含4个或更多个连续核苷酸,包含5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个或更多个连续核苷酸。每个窗口可以具有相同数量的连续核苷酸。窗口可以重叠。每个窗口可以包含第一核酸分子的第一链上的核苷酸和第一核酸分子的第二链上的核苷酸。对于窗口内的每个核苷酸,第一数据结构还可包含链性质的值。链性质可以指示存在的核苷酸或第一链或第二链。窗口可以包含第二链中与第一链中对应位置处的核苷酸不互补的核苷酸。在一些实施例中,第二链上的所有核苷酸与第

一链上的核苷酸互补。在一些实施例中，每个窗口可以仅在第一核酸分子的一条链上包含核苷酸。

[0557] 第一核酸分子可以是环状DNA分子。可以通过使用Cas9复合物切割双链DNA分子形成经过切割的双链DNA分子来形成环状DNA分子。可以将发夹衔接子连接到经过切割的双链DNA分子的末端。在实施例中，可以切割并连接双链DNA分子的两端。例如，可以如图91所述进行切割、连接和随后的分析。

[0558] 第一多个第一数据结构可以包含5,000到10,000个、10,000到50,000个、50,000到100,000个、100,000到200,000个、200,000到500,000个、500,000到1,000,000个或1,000,000个或更多个第一数据结构。多个第一核酸分子可以包含至少1,000个、10,000个、50,000个、100,000个、500,000个、1,000,000个、5,000,000个或更多个核酸分子。作为另一个实例，可以产生至少10,000个、或50,000个、或100,000个、或500,000个、或1,000,000个、或5,000,000个序列读段。

[0559] 通过测量对应于核苷酸的信号中的脉冲对第一核酸分子中的每一个进行测序。信号可以是荧光信号或其它类型的光信号(例如，化学发光、光度学)。信号可以由核苷酸或与核苷酸相关联的标签产生。

[0560] 修饰在每个第一核酸分子的每个窗口中的靶位置处的核苷酸中具有已知的第一状态。第一状态可以是核苷酸中不存在修饰，或者可以是核苷酸中存在修饰。可以已知第一核酸分子中不存在修饰，或者可以对第一核酸分子进行处理使得不存在修饰。可以已知第一核酸分子中存在修饰，或者可以对第一核酸分子进行处理使得存在修饰。如果第一状态是不存在修饰，则每个第一核酸分子的每个窗口中可以不存在修饰并且仅在靶位置处存在修饰。已知的第一状态可以包含第一数据结构的第一部分的甲基化状态和第一数据结构的第二部分的未甲基化状态。

[0561] 靶位置可以是相应窗口的中心。对于跨越偶数个核苷酸的窗口，靶位置可以是紧邻窗口中心的上游或下游的位置。在一些实施例中，靶位置可以在相应窗口的任何其它位置，包含第一位置或最后位置。例如，如果窗口跨越一条链的n个核苷酸，则从第1位置到第n位置(上游或下游)，靶位置可以位于从第1位置到第n位置的任何位置。

[0562] 每个第一数据结构包含窗口内的性质值。性质可以针对窗口内的每个核苷酸。性质可以包含核苷酸的一致性。一致性可以包含碱基(例如，A、T、C或G)。性质还可以包含核苷酸相对于相应窗口内的靶位置的位置。例如，位置可以是相对于靶位置的核苷酸距离。当核苷酸在一个方向上与靶位置相距一个核苷酸时，位置可以是+1，并且当核苷酸在相反的方向上与靶位置相距一个核苷酸时，位置可以是-1。

[0563] 性质可以包含对应于核苷酸的脉冲宽度。脉冲宽度可以是脉冲最大值一半处的脉冲宽度。性质可以进一步包含脉冲间持续时间(IPD)，所述脉冲间持续时间表示对应于核苷酸的脉冲与对应于相邻核苷酸的脉冲之间的时间。脉冲间持续时间可以是与核苷酸相关联的脉冲的最大值与相邻核苷酸相关联的脉冲的最大值之间的时间。相邻核苷酸可以是邻近核苷酸。性质还可以包含与窗口内的每个核苷酸相对应的脉冲的高度。性质可以进一步包含链性质的值，所述值指示核苷酸是存在于第一核酸分子的第一链上还是第二链上。链的指示可以类似于图6中所示的矩阵。

[0564] 多个第一数据结构的每个数据结构可以排除IPD或宽度低于截止值的第一核酸分

子。例如，可以仅使用IPD值大于第10百分位(或第1、第5、第15、第20、第30、第40、第50、第60、第70、第80、第90或第95百分位)的第一核酸分子。百分位可以基于来自一个或多个参考样本中所有核酸分子的数据。宽度的截止值还可以对应于百分位。

[0565] 在框1024处，存储多个第一训练样本。每个第一训练样本包含第一多个第一数据结构之一和指示靶位置处的核苷酸修饰的第一状态的第一标记。

[0566] 在框1026处，接收第二多个第二数据结构。框1026可以是任选的。第二多个第二数据结构中的每个第二数据结构对应于在多个第二核酸分子的相应核酸分子中测序的核苷酸的相应窗口。第二多个核酸分子可以与多个第一核酸分子相同或不同。修饰在每个第二核酸分子的每个窗口内的靶位置处的核苷酸中具有已知的第二状态。第二状态是与第一状态不同的状态。例如，如果第一状态是存在修饰，则第二状态是不存在修饰，反之亦然。每个第二数据结构包含与第一多个第一数据结构相同性质的值。

[0567] 可以使用多重置换扩增来产生多个第一训练样本。在一些实施例中，可以通过使用一组核苷酸扩增第一多个核酸分子来产生多个第一训练样本。所述组核苷酸可以包含指定比率的第一类型的甲基化(例如，6mA或任何其它甲基化[例如，CpG])。相对于未甲基化核苷酸，指定比率可以包含1:10、1:100、1:1000、1:10000、1:100000或1:1000000。可以使用第一类型的未甲基化核苷酸的多重置换扩增来产生多个第二核酸分子。

[0568] 在框1028处，存储多个第二训练样本。框1028可以是任选的。每个第二训练样本包含第二多个第二数据结构之一和指示靶位置处核苷酸修饰的第二状态的第二标记。

[0569] 在框1029处，使用多个第一训练样本和任选地多个第二训练样本来训练模型。当向模型输入第一多个第一数据结构和任选地第二多个第二数据结构时，通过基于与第一标记和任选地第二标记的对应标记匹配或不匹配的模型的输出优化模型的参数来执行训练。模型的输出指定相应窗口中靶位置处的核苷酸是否具有修饰。方法可以仅包含多个第一训练样本，因为模型可以将离群值鉴定为处于不同于第一状态的状态。模型可以是统计模型，还被称为机器学习模型。

[0570] 在一些实施例中，模型的输出可以包含处于多种状态中的每种状态的概率。可以采用概率最高的状态作为状态。

[0571] 模型可以包含卷积神经网络(CNN)。CNN可以包含一组卷积过滤器，所述卷积过滤器被配置成对第一多个数据结构和任选地第二多个数据结构进行过滤。过滤器可以是本文所述的任何过滤器。每层的过滤器数量可以是10到20、20到30、30到40、40到50、50到60、60到70、70到80、80到90、90到100、100到150、150到200或更多。过滤器的内核尺寸可以是2、3、4、5、6、7、8、9、10、11、12、13、14、15、15到20、20到30、30到40或更多。CNN可以包含输入层，所述输入层被配置成接收经过过滤的第一多个数据结构和任选地经过过滤的第二多个数据结构。CNN还可以包含包括多个节点的多个隐藏层。多个隐藏层的第一层耦接到输入层。CNN可以进一步包含输出层，所述输出层耦接到多个隐藏层的最后一层并且被配置成输出输出数据结构。输出数据结构可以包含所述性质。

[0572] 模型可以包含监督学习模型。监督学习模型可以包含不同的方法和算法，其包含分析学习、人工神经网络、反向传播、增强(元算法)、贝叶斯统计、基于案例的推理、决策树学习、归纳逻辑编程、高斯过程回归、遗传编程、分组数据处理方法、核估计器、学习自动机、学习分类器系统、最小消息长度(决策树、决策图等)、多线性子空间学习、朴素贝叶斯分类

器、最大熵分类器、条件随机场、最近邻算法、可能近似正确学习(PAC)学习、链波下降规则、知识获取方法、符号机器学习算法、子符号机器学习算法、支持向量机、最小复杂度机器(MCM)、随机森林、分类器集合、序数分类、数据预处理、处理失衡的数据集、统计关系学习或Proaftn(一种多准则分类算法)。模型可以是线性回归、逻辑回归、深度递归神经网络(例如,长短期记忆,LSTM)、贝叶斯分类器、隐马尔可夫模型(HMM)、线性判别分析(LDA)、k均值聚类、基于密度的带噪声应用空间聚类(DBSCAN)、随机森林算法、支持向量机(SVM)或本文所述的任何模型。

[0573] 作为训练机器学习模型的组成部分,机器学习模型的参数(如权重、阈值,例如,可以用于神经网络中的激活函数等)可以基于训练样本(训练集)进行优化,以在对靶位置处核苷酸的修饰进行分类时提供优化的准确性。可以执行各种形式的优化,例如,反向传播、经验风险最小化和结构风险最小化。样本的验证集(数据结构和标记)可以用于验证模型的准确性。可以使用用于训练和验证的训练集的各个部分来执行交叉验证。模型可以包括多个子模型,由此提供组合模型。子模型可以是较弱的模型,所述较弱的模型一旦组合就可以提供更准确的最终模型。

[0574] 在一些实施例中,嵌合或杂合核酸分子可以用于验证模型。多个第一核酸分子中的至少一些各自包含对应于第一参考序列的第一部分和对应于第二参考序列的第二部分。第一参考序列可以来自与第二参考序列不同的染色体、组织(例如,肿瘤或非肿瘤)、生物体或物种。第一参考序列可以是人类,并且第二参考序列可以来自不同的动物。每个嵌合核酸分子可以包含对应于第一参考序列的第一部分和对应于第二参考序列的第二部分。第一部分可以具有第一甲基化谱式,并且第二部分可以具有第二甲基化谱式。第一部分可以用甲基化酶处理。第二部分可以不用甲基化酶处理,并且可以对应于第二参考序列的未甲基化部分。

[0575] B.修饰检测

[0576] 图103示出了用于检测核酸分子中核苷酸修饰的方法1030。修饰可以是图102的方法1020描述的任何修饰。

[0577] 在框1032处,接收输入数据结构。输入数据结构可以对应于在样本核酸分子中测序的核苷酸的窗口。可以通过测量光信号中的对应于核苷酸的脉冲来对样本核酸分子进行测序。窗口可以是在图102的框1022所描述的任何窗口,并且测序可以是图102的框1022所描述的任何测序。输入数据结构可以包含图102的框1022所描述的相同性质的值。方法1030可以包含对样本核酸分子进行测序。

[0578] 窗口内的核苷酸可以或可以不与参考基因组比对。可以使用环状一致性序列(CCS)测定窗口内的核苷酸,而无需将测序的核苷酸与参考基因组比对。每个窗口中的核苷酸可以由CCS鉴定而不是与参考基因组比对。在一些实施例中,可以在没有CCS并且没有将测序的核苷酸与参考基因组比对的情况下测定窗口。

[0579] 窗口内的核苷酸可以被富集或过滤。可以通过涉及Cas9的方法进行富集。Cas9方法可以包含使用Cas9复合物切割双链DNA分子以形成经过切割的双链DNA分子,并且将发夹衔接子连接到所述经过切割的双链DNA分子的末端上,类似于图91。可以通过选择长度在一定长度范围内的双链DNA分子来进行过滤。核苷酸可以来自这些双链DNA分子。可以使用保留分子的甲基化状态的其它方法(例如,甲基结合蛋白)。

[0580] 在框1034处，输入数据结构被输入到模型中。可以通过图102中的方法1020来训练模型。

[0581] 在一些实施例中，嵌合核酸分子可以用于验证模型。多个第一核酸分子中的至少一些各自包含对应于第一参考序列的第一部分和对应于与第一参考序列不相交的第二参考序列的第二部分。第一参考序列可以来自与第二参考序列不同的染色体、组织(例如，肿瘤或非肿瘤)、细胞器(例如线粒体、细胞核、叶绿体)、生物体(哺乳动物、病毒、细菌等)或物种。第一参考序列可以是人类，并且第二参考序列可以来自不同的动物。每个嵌合核酸分子可以包含对应于第一参考序列的第一部分和对应于第二参考序列的第二部分。第一部分可以具有第一甲基化谱式，并且第二部分可以具有第二甲基化谱式。第一部分可以用甲基化酶处理。第二部分可以不用甲基化酶处理，并且可以对应于第二参考序列的未甲基化部分。

[0582] 在框1036处，使用模型测定输入数据结构的窗口内的靶位置处的核苷酸中是否存在修饰。

[0583] 输入数据结构可以是多个输入数据结构中的一个输入数据结构。每个输入数据结构可以对应于在多个样本核酸分子的相应样本核酸分子中测序的核苷酸的相应窗口。多个样本核酸分子可以从受试者的生物样本中获得。生物样本可以是本文所述的任何生物样本。可以针对每个输入数据结构重复方法1030。方法可以包含接收多个输入数据结构。可以将多个输入数据结构输入到模型中。可以使用模型确定每个输入数据结构的相应窗口中的靶位置处的核苷酸中是否存在修饰。

[0584] 多个样本核酸分子中的每个样本核酸分子的长度可以大于截止长度。例如，截止长度可以是100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、2kb、3kb、4kb、5kb、6kb、7kb、9kb、10kb、20kb、30kb、40kb、50kb、60kb、70kb、80kb、90kb、100kb、500kb或1Mb。截止长度可能会导致子读段深度较高，其中任一种都可以增加修饰检测的准确性。在一些实施例中，方法可以包含在对DNA分子进行测序之前将DNA分子分级成一定长度。

[0585] 多个样本核酸分子可以与多个基因组区域比对。对于多个基因组区域中的每个基因组区域，许多样本核酸分子可以与基因组区域比对。样本核酸分子的数量可以大于截止数量。截止数量可以是子读段深度截止值。子读段深度截止数量可以是1x、10x、30x、40x、50x、60x、70x、80x、900x、100x、200x、300x、400x、500x、600x、700x或800x。可以确定子读段深度截止数量以提高或优化准确性。子读段深度截止数量可以与多个基因组区域的数量有关。例如，子读段深度截止数量较高，多个基因组区域的数量较低。

[0586] 可以确定一个或多个核苷酸处存在修饰。可以使用一个或多个核苷酸处的修饰的存在来确定病症的分类。病症的分类可以包含使用修饰的数量。可以将修饰的数量与阈值进行比较。可替代地或另外地，分类可以包含一个或多个修饰的位置。可以通过使核酸分子的序列读段与参考基因组比对来测定一种或多种修饰的位置。如果示出已知与病症相关的某些位置具有修饰，则可以确定病症。例如，可以将甲基化位点的谱式与病症的参考谱式进行比较，并且可以基于比较来确定病症。与参考谱式匹配或与参考谱式基本匹配(例如80％、90％或95％或更高)可以指示病症或很可能患有病症。病症可以是本文所述的癌症或任何病症(例如，妊娠相关病症、自身免疫性疾病)。

[0587] 可以分析统计上显著数量的核酸分子，以提供对病症、组织起源或临床相关DNA分率的准确测定。在一些实施例中，分析了至少1,000个核酸分子。在其它实施例中，可以分析

至少10,000个、或50,000个、或100,000个、或500,000个、或1,000,000个、或5,000,000个核酸分子。作为另一个实例，可以产生至少10,000个、或50,000个、或100,000个、或500,000个、或1,000,000个、或5,000,000个序列读段。

[0588] 方法可以包含确定病症的分类是受试者患有所述病症。使用修饰的数量和/或修饰的位点，分类可以包含病症的水平。

[0589] 可以使用一个或多个核苷酸处存在修饰来确定临床相关的DNA分率、胎儿甲基化谱、母本甲基化谱、印记基因区域的存在或起源组织(例如，来自含有不同细胞类型的混合物的样本)。临床相关的DNA分率包含但不限于胎儿DNA分率、肿瘤DNA分率(例如，来自含有肿瘤细胞和非肿瘤细胞混合物的样本)和移植DNA分率(例如，来自含有供体细胞和受体细胞混合物的样本)。

[0590] 方法可以进一步包含治疗病症。可以根据所测定的病症水平、经过鉴定的修饰和/或(例如，从癌症患者的循环中分离出的肿瘤细胞的)起源组织提供治疗。例如，可以用特定药物或化学疗法靶向经过鉴定的修饰。起源组织可以用于指导手术或任何其它形式的治疗。并且，病症水平可以用于测定任何类型的治疗的积极程度。

[0591] 实施例可以包含在测定患者的病状水平之后治疗患者的病症。治疗可以包含任何合适的疗法、药物、化学疗法、放射治疗或外科手术，包含本文提及的参考文献中所描述的任何治疗。参考文献中关于治疗的信息通过引用并入本文。

[0592] VI.单倍型分析

[0593] 在肿瘤组织样本中发现了两种单倍型之间的甲基化谱的差异。因此，单倍型之间的甲基化失衡可以用于确定癌症或其它病症水平的分类。单倍型的失衡还可以用于鉴定胎儿对单倍型的遗传。还可以通过分析单倍型之间的甲基化失衡来鉴定胎儿病症。细胞DNA可以用于分析单倍型的甲基化水平。

[0594] A.单倍型相关甲基化分析

[0595] 单分子实时测序技术允许鉴定单独的SNP。从单分子实时测序孔中产生的长读段(例如，多达若干千碱基)将允许通过利用每个一致性读段中存在的单倍型信息来定相基因组中的变体(Edge等人《基因组研究》2017;27:801-812;Wenger等人《自然生物技术》2019;37:1155-1162)。可以根据被CCS连接到相应单倍型上的等位基因的CpG位点的甲基化水平来分析单倍型的甲基化谱，如图77所展示的。此阶段性甲基化单倍型分析可以用于解决在不同的临床相关病状(如癌症)中两个同源染色体拷贝是共享类似的甲基化谱式还是共享不同的甲基化谱式的问题。在一个实施例中，单倍型甲基化将是由分配给所述单倍型的许多DNA片段贡献的总体甲基化水平。单倍型可以是长度不同的区块，所述长度包含但不限于50nt、100nt、200nt、300nt、400nt、500nt、1knt、2knt、3knt、4knt、5knt、10knt、20knt、30knt、40knt、50knt、100knt、200knt、300knt、400knt、500knt、1Mnt、2Mnt和3Mnt。

[0596] B.基于相对单倍型的甲基化失衡分析

[0597] 图104展示了基于相对单倍型的甲基化失衡分析。通过分析单分子实时测序结果来测定单倍型(即，Hap I和Hap II)。可以使用那些单倍型相关片段来测定与每个单倍型关联的甲基化谱式，所述单倍型相关片段的甲基化谱根据图77所述的方法测定。由此，可以将Hap I与Hap II之间的甲基化谱式进行比较。

[0598] 为了定量Hap I与Hap II之间的甲基化差异，计算了Hap I与Hap II之间的甲基化

水平差异(ΔF)。差异ΔF计算为：

[0599] $\Delta F = M_{HapI} - M_{HapI}$

[0600] 其中ΔF表示Hap I与Hap II之间的甲基化水平差异，并且M_{HapI}和M_{HapI}分别表示Hap I和Hap II的甲基化水平。ΔF的正值表明Hap I的DNA甲基化水平高于Hap II。

[0601] C.HCC肿瘤DNA的基于相对单倍型的甲基化失衡分析

[0602] 在一个实施例中，单倍型甲基化分析可以用于检测癌症基因组中的甲基化异常。例如，将分析基因组区域内两个单倍型之间的甲基化变化。基因组区域内的单倍型被定义为单倍型区块。单倍型区块可以被认为是染色体上已定相的一组等位基因。在一些实施例中，根据支持物理上连接在染色体上的两个等位基因的一组序列信息，将单倍型区块尽可能延长。对于案例3033，从邻近正常组织DNA的测序结果中获得了97,475个单倍型区块。单倍型区块的中值长度为2.8kb。25％的单倍型区块的长度大于8.2kb。单倍型区块的最大长度为282.2kb。数据集由通过Sequel II测序试剂盒1.0制备的DNA产生。

[0603] 出于说明的目的，使用了许多标准来鉴定潜在的单倍型区块，所述单倍型区块与邻近非肿瘤组织DNA相比在肿瘤DNA中展现出Hap I与Hap II之间的差异甲基化。所述标准是：(1)被分析的单倍型区块包含至少三个分别由三个测序孔产生的CCS序列；(2)邻近非肿瘤组织DNA中的Hap I与Hap II之间的甲基化水平的绝对差异小于5％；(3)肿瘤组织DNA中的Hap I与Hap II之间的甲基化水平的绝对差异大于30％。鉴定了满足上述标准的73个单倍型区块。

[0604] 图105A和105B是73个单倍型区块的表格，其示出了针对情况TBR3033的与邻近非肿瘤组织DNA相比HCC肿瘤DNA中的Hap I与Hap II之间的差异甲基化水平。第一列示出了与单倍型区块相关联的染色体。第二列示出了染色体内的单倍型区块的起始坐标。第三列示出了单倍型区块的结束坐标。第四列示出了单倍型区块的长度。第四列列出了单倍型区块id。第五列示出了与肿瘤组织邻近的非肿瘤组织中的Hap I的甲基化水平。第六列示出了非肿瘤组织中的Hap II的甲基化水平。第七列示出了肿瘤组织中的Hap I的甲基化水平。第八列示出了肿瘤组织中的Hap II的甲基化水平。

[0605] 与示出肿瘤组织DNA的单倍型之间的甲基化水平差异大于30％的73个单倍型区块相比，只有一个单倍型区块示出了非肿瘤组织DNA的差异大于30％而肿瘤组织DNA的差异小于5％。在一些实施例中，可以使用另一套标准来鉴定显示出差异甲基化的单倍型区块。可以使用其它最大和最小阈值差异。例如，最小阈值差异可以是10％、15％、20％、25％、30％、35％、40％、45％、50％或更大。作为实例，最大阈值差异可以是1％、5％、10％、15％、20％或30％。这些结果表明，单倍型之间甲基化差异的变化可以充当癌症诊断、检测、监测、预后和治疗指导的新生物标志物。

[0606] 在一些实施例中，当研究甲基化谱式时，长单倍型区块将在计算机模拟中被分割成更小的区块。

[0607] 对于案例3032，从邻近非肿瘤组织DNA的测序结果中获得了61,958个单倍型区块。单倍型区块的中值长度为9.3kb。25％的单倍型区块的长度大于27.6kb。单倍型区块的最大长度为717.8kb。作为说明，使用了上述三个相同标准来鉴定潜在的单倍型区块，所述单倍型区块与邻近肿瘤组织DNA相比在肿瘤DNA中展现出Hap I与Hap II之间的差异甲基化。鉴定了满足上述标准的20个单倍型区块。数据集由通过Sequel II测序试剂盒1.0制备的DNA

产生。

[0608] 图106是20个单倍型区块的表格，其示出了针对情况TBR3032的与邻近正常组织DNA相比肿瘤DNA中的Hap I与Hap II之间的差异甲基化水平。第一列示出了与单倍型区块相关联的染色体。第二列示出了染色体内的单倍型区块的起始坐标。第三列示出了单倍型区块的结束坐标。第四列示出了单倍型区块的长度。第四列列出了单倍型区块id。第五列示出了与肿瘤组织邻近的非肿瘤组织中的Hap I的甲基化水平。第六列示出了非肿瘤组织中的Hap II的甲基化水平。第七列示出了肿瘤组织中的Hap I的甲基化水平。第八列示出了肿瘤组织中的Hap II的甲基化水平。

[0609] 与图106中示出HCC肿瘤组织差异的20个单倍型区块相比，只有一个单倍型区块示出了非肿瘤组织中的差异大于30％而肿瘤组织中的差异小于5％。这些结果进一步表明，单倍型之间甲基化差异的变化将充当癌症诊断、检测、监测、预后和治疗指导的新生物标志物。对于其它实施例，可以使用其它标准来鉴定显示出差异甲基化的单倍型区块。

[0610] D.来自其它肿瘤类型的DNA的基于相对单倍型的甲基化失衡分析

[0611] 如上所述，对单倍型之间甲基化水平的分析揭示了与配对的邻近非肿瘤组织相比，HCC肿瘤组织含有更多显示甲基化失衡的单倍型区块。作为一个实例，针对示出肿瘤组织中的甲基化失衡的单倍型区块的标准是：(1)被分析的单倍型区块包含至少三个由三个测序孔产生的CCS序列；(2)基于历史数据，邻近非肿瘤组织DNA或正常组织DNA中的Hap I与Hap II之间的甲基化水平的绝对差异小于5％；(3)肿瘤组织DNA中的Hap I与Hap II之间的甲基化水平的绝对差异大于30％。纳入了标准(2)，因为示出甲基化水平中单倍型失衡的非肿瘤/正常组织可以指示印记区域而不是肿瘤区域。针对示出非肿瘤组织中的甲基化失衡的单倍型区块的标准是：(1)被分析的单倍型区块包含至少三个由三个测序孔产生的CCS序列；(2)基于历史数据，邻近非肿瘤组织DNA或正常组织DNA中的Hap I与Hap II之间的甲基化水平的绝对差异大于30％；(3)肿瘤组织DNA中的Hap I与Hap II之间的甲基化水平的绝对差异小于5％。

[0612] 在其它实施例中，可以使用其它标准。例如，为了鉴定失衡单倍型I型癌症基因组，在非肿瘤组织中，Hap I与Hap II之间的甲基化水平差异可以小于1％、5％、10％、20％、40％、50％或60％等，而在肿瘤组织中，Hap I和Hap II之间的甲基化水平差异可以大于1％、5％、10％、20％、40％、50％或60％等。为了鉴定失衡单倍型I型非癌症基因组，在非肿瘤组织中，Hap I与Hap II之间的甲基化水平差异可以大于1％、5％、10％、20％、40％、50％或60％等，而在肿瘤组织中，Hap I和Hap II之间的甲基化水平差异可以小于1％、5％、10％、20％、40％、50％或60％等。

[0613] 图107A是表格，其基于由Sequel II测序试剂盒2.0产生的数据总结了示出肿瘤与邻近非肿瘤组织之间的两种单倍型之间的甲基化失衡的单倍型区块的数量。第一列列出了组织类型。第二列列出了示出肿瘤组织中两种单倍型之间甲基化失衡的单倍型区块的数量。第三列列出了在配对的邻近非肿瘤组织中两种单倍型之间甲基化失衡的单倍型区块的数量。行示出了与配对的邻近非肿瘤组织相比具有更多的示出两种单倍型之间的甲基化失衡的单倍型区块的肿瘤组织。

[0614] 此分析中涉及的单倍型区块的中值长度为15.7kb(IQR:10.3-26.1kb)。包含肝脏的HCC结果，数据示出了7种组织类型的肿瘤组织含有更多的具有甲基化失衡的单倍型区

块。除肝脏外，其它组织包含结肠、乳房、肾、肺、前列腺和胃组织。因此，在一些实施例中，可以使用含有甲基化失衡的单倍型区块的数量来检测患者是否患有肿瘤或癌症。

[0615] 图107B是表格，其基于由Sequel II测序试剂盒2.0产生的数据总结了示出在不同肿瘤阶段的肿瘤组织中的两种单倍型之间的甲基化失衡的单倍型区块的数量。第一列示出了具有肿瘤的组织类型。第二列示出了在肿瘤组织中两个单倍型之间具有甲基化失衡的单倍型区块的数量。第三列使用恶性肿瘤的TNM分类列出了肿瘤分期信息。T3和T3a的肿瘤尺寸比T2大。

[0616] 所述表示出了更多单倍型区块，所述单倍型区块示出了乳腺和肾的较大肿瘤的甲基化失衡。例如，对于乳腺组织，被归类为肿瘤等级T3(TNM分期)、ER阳性并展现出ERBB2扩增的组织比被归类为肿瘤等级T2(TNM分期)、PR(孕酮受体)/ER(雌激素受体)阳性并且没有ERBB2扩增的组织的单倍型区块(18)具有更多的示出甲基化失衡的单倍型区块(57)。对于肾组织，被归类为肿瘤等级T3a的组织比被归类为肿瘤等级T2的组织的单倍型区块(0)具有更多的示出甲基化失衡的单倍型区块(68)。

[0617] 在一些实施例中，可以使用示出甲基化失衡的单倍型区块对肿瘤进行分类并与其临床行为(例如进展、预后或治疗反应)相关联。这些数据表明，基于单倍型的甲基化失衡的程度可以充当肿瘤的分类器，并且可以并入临床研究或试验或最终的临床服务中。肿瘤的分类可以包含尺寸和严重性。

[0618] E.母本血浆细胞游离DNA的基于单倍型的甲基化分析

[0619] 可以确定双亲或亲本之一的单倍型。单倍型方法可以包含长读段单分子测序、连接的短读段测序(例如10x基因组学)、长距离单分子PCR或群体推断。如果父本单倍型是已知的，则可以通过连接多个细胞游离DNA分子的甲基化谱来组合细胞游离胎儿DNA甲基化组，每个细胞游离DNA分子包含沿父本单倍型存在的至少一个父本特异性SNP等位基因。换句话说，父本单倍型被用作连接胎儿特异性读段序列的支架。

[0620] 图108展示了相对甲基化失衡的单倍型分析。如果母本单倍型是已知的，则可以使用两种单倍型(即Hap I和Hap II)之间的甲基化失衡来确定胎儿遗传的母本单倍型。如图108所示，使用单分子实时测序技术对来自孕妇的血浆DNA分子进行测序。可以根据本公开测定甲基化和等位基因信息。在一个实施例中，与致病基因连接的SNP被指定为Hap I。如果胎儿遗传了Hap I，则与携带Hap II等位基因的那些相比，母本血浆中将存在更多的携带Hap I等位基因的片段。与Hap II相比，来源于胎儿的DNA片段的低甲基化会降低Hap I的甲基化水平。作为结果，如果Hap I的甲基化水平低于Hap II的甲基化水平，则胎儿更可能遗传母本Hap I。否则，胎儿更可能遗传母本Hap II。在临床实践中，基于单倍型的甲基化失衡分析可以用于确定未出生的胎儿是否遗传了与遗传病症相关联的母本单倍型，例如但不限于单基因病症，包含脆性X综合征、肌肉营养不良、亨廷顿病或β地中海贫血。

[0621] F.示例病症分类方法

[0622] 图109示出了对具有第一单倍型和第二单倍型的生物体中的病症进行分类的示例方法1090。方法1090涉及比较两种单倍型之间的相对甲基化水平。

[0623] 在框1091处，分析来自生物样本的DNA分子以鉴定其在对应于生物体的参考基因组中的位置。DNA分子可以是细胞DNA分子。例如，可以对DNA分子进行测序以获得序列读段，并且可以将序列读段映射(比对)到参考基因组。如果生物体是人类，则参考基因组将是可

能来自特定亚群的参考人类基因组。作为另一个实例，可以用不同的探针(例如，按照PCR或其它扩增方法)分析DNA分子，其中每个探针对应一个基因组位置，所述基因组位置可以覆盖杂合子和一个或多个CpG位点，如下所述。

[0624] 进一步地，可以分析DNA分子以测定DNA分子的相应等位基因。例如，可以根据从测序获得的序列读段或与DNA分子杂合的特定探针来测定DNA分子的等位基因，其中这两种技术都可以提供序列读段(例如，当存在杂合时，可以将探针视为序列读段)。可以测定DNA分子在一个或多个位点(例如，CpG位点)的每个位点处的甲基化状态。

[0625] 在框1092处，鉴定第一染色体区域的第一部分的一个或多个杂合基因座。每个杂合基因座可以包含第一单倍型中对应的第一等位基因和第二单倍型中对应的第二等位基因。一个或多个杂合基因座可以是第一多个杂合基因座，其中第二多个杂合基因座可以对应于不同的染色体区域。

[0626] 在框1093处，鉴定多个DNA分子的第一集合。多个DNA分子中的每个DNA分子位于来自框1096的杂合基因座中的任一个杂合基因座处并且包含对应的第一等位基因，使得可以将DNA分子鉴定为对应于第一单倍型。DNA分子可能位于杂合基因座中的一个以上杂合基因座处，但通常读段将仅包含一个杂合基因座。DNA分子的第一集合中的每个DNA分子还包含N个基因组位点中的至少一个，其中基因组位点用于测量甲基化水平。N是整数，例如大于或等于1、2、3、4、5、10、20、50、100、200、500、1,000、2,000或5,000。因此，DNA分子的读段可以指示覆盖1个位点、2个位点等。1个基因组位点可以包含存在CpG核苷酸的位点。

[0627] 在框1094处，使用多个DNA分子的第一集合来测定第一单倍型的第一部分的第一甲基化水平。第一甲基化水平可以通过本文所述的任何方法测定。第一部分可以对应于单个位点或包含多个位点。第一单倍型的第一部分可以长于或等于1kb。例如，第一单倍型的第一部分可以长于或等于1kb、5kb、10kb、15kb或20kb。甲基化数据可以是来自细胞DNA的数据。

[0628] 在一些实施例中，可以针对第一单倍型的多个部分测定多个第一甲基化水平。对于第一单倍型的第一部分，每个部分的长度可以大于或等于5kb或具有本文公开的任何长度。

[0629] 在框1095处，鉴定多个DNA分子的第二集合。多个DNA分子中的每个DNA分子位于来自框1096的杂合基因座中的任一个杂合基因座处并且包含对应的第二等位基因，使得可以将DNA分子鉴定为对应于第二单倍型。DNA分子的第二集合中的每个DNA分子还包含N个基因组位点中的至少一个，其中基因组位点用于测量甲基化水平。

[0630] 在框1096处，使用多个DNA分子的第二集合来测定第二单倍型的第一部分的第二甲基化水平。第二甲基化水平可以通过本文所述的任何方法测定。第二单倍型的第一部分可以长于或等于1kb或第一单倍型的第一部分的任何长度。第一单倍型的第一部分可以与第二单倍型的第一部分互补。第一单倍型的第一部分和第二单倍型的第一部分可以形成环状DNA分子。可以使用来自环状DNA分子的数据来测定第一单倍型的第一部分的第一甲基化水平。例如，对环状DNA的分析可以包含图1、图2、图4、图5、图6、图7、图8、图50或图61所述的分析。

[0631] 可以通过使用Cas9复合物切割双链DNA分子形成经过切割的双链DNA分子来形成环状DNA分子。可以将发夹衔接子连接到经过切割的双链DNA分子的末端。在实施例中，可以

切割并连接双链DNA分子的两端。例如,可以如图91所述进行切割、连接和随后的分析。

[0632] 在一些实施例中,可以针对第二单倍型的多个部分测定多个第二甲基化水平。第二单倍型的多个部分中的每个部分可以与第一单倍型的多个部分中的一部分互补。

[0633] 在框1097处,使用第一甲基化水平和第二甲基化水平计算参数值。参数可以是分离值。分离值可以是两个甲基化水平之间的差或两个甲基化水平的比率。

[0634] 如果使用第二单倍型的多个部分,则对于第二单倍型的多个部分中的每个部分,可以使用第二单倍型部分的第二甲基化水平和使用第一单倍型的互补部分的第一甲基化水平来计算分离值。可以将分离值与截止值进行比较。

[0635] 截止值可以从不患有病症的组织中测定。参数可以是第二单倍型的分离值超过截止值的部分的数量。例如,第二单倍型的分离值超过截止值的部分的数量可以类似于图105A、图105B和图106中所示的差异大于30%的区域的数量。对于图105A、图105B和图106,分离值是比率,并且截止值是30%。在一些实施例中,可以从具有病症的组织中确定截止值。

[0636] 在另一个实例中,每个部分的分离值可以被聚合(例如,相加),这可以通过加权和或者相关分离值的函数和来完成。此类聚合可以提供参数值。

[0637] 在框1098处,可以将参数值与参考值进行比较。可以使用不患有病症的参考组织来确定参考值。参考值可以是分离值。例如,参考值可以表示两种单倍型的甲基化水平之间不应存在显著差异。例如,参考值可以是统计差异0或约1的比率。当使用多个部分时,参考值可以是健康生物体中两个单倍型示出超过截止值的分离值的多个部分。在一些实施例中,可以使用具有病症的参考组织来确定参考值。

[0638] 在框1099处,使用参数值与参考值的比较来确定生物体中病症的分类。如果参数的值超过参考值,则可以确定存在病症或很可能存在病症。所述病症可以包含癌症。癌症可以是本文所述的任何癌症。病症的分类可能是病症的可能性。病症的分类可以包含病症的严重性。例如,指示大量部分的单倍型失衡的较大参数值可以指示更严重形式的癌症。

[0639] 虽然图109描述的方法涉及病症的分类,但是可以使用类似的方法来确定可能由单倍型之间的甲基化水平失衡导致的任何病状或特性。例如,来自胎儿DNA的单倍型的甲基化水平可以低于来自母本DNA的单倍型的甲基化水平。甲基化水平可以用于将核酸归类为母本或胎儿。

[0640] 当病症是癌症时,肿瘤的不同染色体区域可能展现出此类甲基化的差异。根据受影响的区域,可以提供不同的治疗方法。进一步地,具有展现出此类甲基化的差异的不同区域的受试者可以具有不同的预后。

[0641] 具有足够间隔(例如,大于截止值)的染色体区域(部分)可以被鉴定为异常(或具有异常间隔)。可以将异常区域的谱式(潜在地解释了哪个单倍型高于另一个)与参考谱式(例如,根据患有癌症、潜在地是特定类型的癌症的受试者或健康受试者确定的)进行比较。如果两种谱式在阈值内(例如,小于不同的指定数量的区域/部分)与具有特定分类的参考谱式相同,则可以将受试者鉴定为具有所述病症的分类。这种分类可以包含印记病症,例如,如本文所述。

[0642] VII.杂合分子的单分子甲基化分析

[0643] 为了进一步评估本文公开的实施例在测定核酸的碱基修饰方面的性能和实用性,

人工创建了人类和小鼠杂合DNA片段,其中人组成部分是甲基化的并且小鼠组成部分是未甲基化的,反之亦然。测定杂合或嵌合DNA分子的接合点可以允许检测包含癌症的各种病症或疾病的基因融合。

[0644] A.创建人和小鼠杂合DNA片段的方法

[0645] 本节描述了创建杂合DNA片段,然后是用于测定片段的甲基化谱的程序。

[0646] 在一个实施例中,通过全基因组扩增来扩增人DNA,使得人类基因组中的原始甲基化特征将会被消除,因为全基因组扩增不会保留甲基化状态。可以使用耐外切核酸酶的硫代磷酸酯修饰的变性六聚体作为引物执行全基因组扩增,所述引物可以在基因组上随机结合,从而使聚合酶(例如Phi29 DNA聚合酶)无需热循环即可扩增DNA。经过扩增的DNA产物将是未甲基化的。经过扩增的人DNA分子用CpG甲基转移酶M.SssI进一步处理,所述M.SssI理论上将完全甲基化双链、未甲基化或半甲基化DNA中CpG上下文处的所有胞嘧啶。因此,此类用M.SssI处理的经过扩增的人DNA将变成甲基化DNA分子。

[0647] 相比之下,小鼠DNA经受了全基因组扩增,从而将产生未甲基化小鼠DNA片段。

[0648] 图110展示了产生人组成部分被甲基化而小鼠组成部分未被甲基化的人-小鼠杂合DNA片段。实心圆圈表示甲基化CpG位点。空心圆圈表示未甲基化CpG位点。具有斜条纹的粗条11010表示甲基化人组成部分。具有竖直条纹的粗条11020表示未甲基化小鼠组成部分。

[0649] 为了产生杂合的人-小鼠DNA分子,在一个实施例中,将经过扩增的全基因组和经过M.SssI处理的DNA分子用HindIII和NcoI进一步消化以产生粘性末端,以促进下游连接。在一个实施例中,将甲基化人DNA片段与未甲基化小鼠DNA片段以等摩尔比进一步混合。这种人-小鼠DNA混合物经受了连接过程,在一个实施例中,所述过程由DNA连接酶在20℃下介导15分钟。如图110所示,此连接反应将产生3种类型的所得分子,其包含人-小鼠杂合DNA分子(a:人-小鼠杂合片段);仅人DNA分子(b:人-人连接,和c:没有连接的人DNA);和仅小鼠DNA分子(d:小鼠-小鼠连接和e:没有连接的小鼠DNA)。连接后的DNA产物经受单分子实时测序。根据本文提供的公开内容分析测序结果以测定甲基化状态。

[0650] 图111展示了产生人组成部分被甲基化而小鼠组成部分未被甲基化的人-小鼠杂合DNA片段。实心圆圈表示甲基化CpG位点。空心圆圈表示未甲基化CpG位点。具有斜条纹的粗条11110代表甲基化的小鼠部分。具有竖直条纹的粗条11120表示未甲基化人组成部分。

[0651] 对于图111中的实施例,通过全基因组扩增来扩增小鼠DNA分子,使得小鼠基因组中的原始甲基化将会被消除。经过扩增的DNA产物将是未甲基化的。经过扩增的小鼠DNA将用M.SssI进一步处理。因此,此类用M.SssI处理的经过扩增的小鼠DNA将变成甲基化DNA分子。相反,人DNA片段经受了全基因组扩增,从而将获得未甲基化人片段。在一个实施例中,将甲基化人片段与未甲基化片段以等摩尔比进一步混合。这种人-小鼠DNA混合物经受了由DNA连接酶介导的连接过程。如图111所示,此连接反应将产生3种类型的所得分子,其包含人-小鼠杂合DNA分子(a:人-小鼠杂合片段);仅人DNA分子(b:人-人连接,和c:没有连接的人DNA);和仅小鼠DNA分子(d:小鼠-小鼠连接和e:没有连接的小鼠DNA)。连接后的DNA产物经受单分子实时测序。根据本文提供的公开内容分析测序结果以测定甲基化状态。

[0652] 根据图110所示的实施例,制备了人工DNA混合物(命名为样本MIX01),其包括人-小鼠杂合DNA分子、仅人DNA和仅小鼠DNA,其中人相关的DNA分子是甲基化的而小鼠DNA分子

是未甲基化的。对于样本MIX01,获得了1.66亿个子读段,其可以与人类或小鼠参考基因组比对,或与人类基因组和小鼠基因组部分比对。这些子读段从大约500万个太平洋生物科学公司的单分子实时(SMRT)测序孔中产生。单分子实时测序孔中的每个分子平均被测序32次(范围:1-881次)。

[0653] 为了测定杂合片段中的人DNA和小鼠DNA组成部分,首先通过组合来自孔中所有相关子读段的核苷酸信息来构建一致性序列。总共获得了样本MIX01的3,435,657个一致性序列。数据集由通过Sequel II测序试剂盒1.0制备的DNA产生。

[0654] 将一致性序列与包括人类和小鼠参考的参考基因组比对。获得了320万个比对的一致性序列。其中,有39.6%被归类为仅人DNA类型,有26.5%被归类为仅小鼠DNA类型,并且有30.2%被归类为人-小鼠杂合DNA。

[0655] 图112示出了连接之后DNA混合物(样本MIX01)中的DNA分子的长度分布。x轴示出了DNA分子的长度。y轴示出了与DNA分子的长度相关联的频率。如图112所示,人-小鼠杂合DNA分子具有较长的长度分布,这与人-小鼠杂合DNA分子是至少两种类型分子的组合的事实一致。

[0656] 图113展示了第一DNA(A)和第二DNA(B)结合在一起的接合点区域。可以用限制酶消化DNA(A)和DNA(B)。在一个实施例中,为了提高使用交错末端的连接效率,使用了分别识别A^AGCTT和C^CATGG位点的限制酶HindIII和NcoI,以在连接步骤之前消化人和小鼠DNA。然后可以连接DNA(A)和DNA(B)。在698,492个含有接合点区域的人-小鼠杂合DNA分子中,发现88%的人-小鼠杂合DNA分子携带A^AGCTT和C^CATGG的酶识别位点,这进一步表明人与小鼠DNA片段之间发生了连接。所述接合点区域被定义为第一DNA片段和第二DNA片段物理地连接在一起的区域或位点。因为接合点包含DNA(A)和DNA(B)两者一致性的序列,所以仅通过序列不能确定对应于接合点的一条链的部分是DNA(A)或DNA(B)的组成部分。分析对应于接合点的一条链的部分的甲基化谱式或密度可以用于确定所述部分是来自DNA(A)还是来自DNA(B)。作为实例,DNA(A)可以是病毒DNA,并且DNA(B)可以是人DNA。测定准确的接合点可以告知此类整合的DNA是否以及如何破坏蛋白质结构。

[0657] 图114展示了DNA混合物的甲基化分析。具有斜条纹的条11410指示在比对分析中观察到的接合点区域,所述接合点区域将在连接之前通过限制酶处理引入。“RE位点”表示限制酶(RE)识别位点。

[0658] 如图114所示,在一个实施例中,比对的一致性序列被分成如下三类:

[0659] (1)参考一个或多个比对标准,测序DNA仅与人类参考基因组比对,而不与小鼠参考基因组比对。在一个实施例中,一种比对标准可以被定义为但不限于测序DNA的100%、95%、90%、80%、70%、60%、50%、40%、30%或20%的连续核苷酸可以与人类参考比对。在一个实施例中,一种比对标准将是未与人类参考比对的测序片段的剩余组成部分不能与小鼠参考基因组比对。在一个实施例中,一种比对标准是测序DNA可以与参考人类基因组中的单个区域比对。在一个实施例中,比对可以是完美的。但在其它实施例中,比对可以适应核苷酸差异,包含插入、错配和缺失,条件是此类差异小于某些阈值,如但不限于比对序列长度的1%、2%、3%、4%、5%、10%、20%或30%。在另一个实施例中,可以在参考基因组中的多于一个位置进行比对。但在其它实施例中,与参考基因组中的一个或多个位点进行比对可以以概率方式(例如,指示错误比对的可能性)来陈述,并且概率测量结果可以用于后

续处理中。

[0660] (2)参考一个或多个比对标准,测序DNA仅与小鼠参考基因组比对,而不与人类参考基因组比对。在一个实施例中,一种比对标准可以被定义为但不限于测序DNA的100%、95%、90%、80%、70%、60%、50%、40%、30%或20%的连续核苷酸可以与小鼠参考比对。在一个实施例中,一种比对标准将是剩余组成部分不能与人类参考基因组比对。在一个实施例中,一种比对标准是测序DNA可以与参考小鼠基因组中的单个区域比对。在一个实施例中,比对可以是完美的。但在其它实施例中,比对可以适应核苷酸差异,包含插入、错配和缺失,条件是此类差异小于某些阈值,如但不限于比对序列长度的1%、2%、3%、4%、5%、10%、20%或30%。在另一个实施例中,可以在参考基因组中的多于一个位置进行比对。但在其它实施例中,与参考基因组中的一个或多个位点进行比对可以以概率方式(例如,指示错误比对的可能性)来陈述,并且概率测量结果可以用于后续处理中。

[0661] (3)测序DNA的一个组成部分与人类参考基因组唯一比对,而另一个组成部分与小鼠参考基因组唯一比对。在一个实施例中,如果在连接之前使用限制酶,则将在比对分析中观察到对应于限制酶切割位点的接合点区域。在一些实施例中,由于测序和比对误差,可能只能在某个区域内大概测定人与小鼠DNA部分之间的接合点区域。在一些实施例中,如果连接涉及没有切割限制酶的分子(例如,如果存在平端连接),则在人-小鼠杂合DNA片段的接合点区域中将观察不到限制酶识别位点。

[0662] 从对应于一致性序列的那些子读段中获得了CpG位点周围的脉冲间持续时间(IPD)、脉冲宽度(PW)和序列上下文。因此,可以根据本公开中呈现的实施例来测定每个DNA分子(包含仅人、仅小鼠和人-小鼠杂合DNA)的甲基化。

[0663] B.甲基化结果

[0664] 本节描述了杂合DNA片段的甲基化结果。甲基化密度可以用于鉴定杂合DNA片段的不同组成部分的起源。

[0665] 图115示出了样本MIX01中的CpG位点被甲基化的概率的箱线图。x轴示出了样本MIX01中存在的三种不同分子:仅人DNA、仅小鼠DNA和人-小鼠杂合DNA(包含人组成部分和小鼠组成部分)。y轴示出了特定单个DNA分子的CpG位点被甲基化的概率。此测定以人DNA甲基化程度更高而小鼠DNA未甲基化程度更高的方式执行。

[0666] 如图115所示,CpG位点在仅人DNA中被甲基化的概率(中值:0.66;范围:0-1)显著高于仅小鼠DNA的概率(中值:0.06;范围:0-1)(P值<0.0001)。这些结果与测定设计一致,其中人DNA由于CpG甲基转移酶M.SssI的处理而甲基化程度更高,而小鼠DNA的未甲基化程度更高,因为在全基因组扩增期间不能保留甲基化。此外,示出了人-小鼠杂合DNA分子中人DNA组成部分内的CpG位点被甲基化的概率(中值:0.69;范围:0-1)高于小鼠DNA组成部分内的CpG位点被甲基化的概率(中值:0.06;范围:0-1)(P值<0.0001)。这些数据指示,所公开的方法可以准确地测定DNA分子的甲基化状态以及DNA分子内的片段。

[0667] 甲基化概率是指基于所用的统计模型估计的单个分子内特定CpG位点的概率。概率为1指示,基于统计模型,100%的使用测得参数(包含IPD、PW和序列上下文)的CpG位点将是甲基化的。概率为0指示,基于统计模型,0%的使用测得参数(包含IPD、PW和序列上下文)的CpG位点将是甲基化的。换句话说,所有使用测得参数的CpG位点都将是未甲基化的。图115示出了甲基化概率的分布,其中仅人DNA和人组成部分的分布比小鼠对应物的分布宽。

使用亚硫酸氢盐测序测量类似样本的甲基化，以确认甲基化未完成，并且结果如下所示。图115示出了人DNA与小鼠DNA的甲基化之间的显著差异。

[0668] 根据图111所示的实施例，制备了人工DNA混合物(命名为样本MIX02)，其包括人-小鼠杂合DNA分子、仅人DNA和仅小鼠DNA，其中人组成部分是未甲基化的而小鼠组成部分是甲基化的。对于样本MIX02，获得了1.40亿个子读段，其可以与人类或小鼠参考基因组比对，或与人类基因组和小鼠基因组部分比对。这些子读段从大约500万个太平洋生物科学公司的单分子实时(SMRT)测序孔中产生。单分子实时测序孔中的每个分子平均被测序27次(范围:1-1028次)。

[0669] 还通过组合来自孔中所有相关子读段的核苷酸信息来构建一致性序列。总共获得了样本MIX02的3,265,487个一致性序列。使用BWA将一致性序列与包括人和小鼠参考的参考基因组比对(Li H等人,《生物信息学》2010;26(5):589-595)。获得了300万个比对的一致性序列。其中，有30.5％被归类为仅人DNA类型，有32.2％被归类为仅小鼠DNA类型，并且有33.8％被归类为人-小鼠杂合DNA。数据集由通过Sequel II测序试剂盒1.0制备的DNA产生。

[0670] 图116示出了样本MIX02的交叉连接之后DNA混合物中DNA分子的长度分布。x轴示出了DNA分子的长度。y轴示出了与DNA分子的长度相关联的频率。如图116所示，人-小鼠杂合DNA分子具有较长的长度分布，这与人-小鼠杂合DNA分子是通过连接多于一个分子而产生的事实一致。

[0671] 图117示出了样本MIX02中的CpG位点被甲基化的概率的箱线图。根据本文所述的方法测定甲基化状态。x轴示出了样本MIX01中存在的三种不同分子:仅人DNA、仅小鼠DNA和人-小鼠杂合DNA(包含人组成部分和小鼠组成部分)。y轴示出了CpG位点被甲基化的概率。此测定以人DNA是未甲基化的而小鼠DNA是甲基化的方式执行。

[0672] 如图117所示，CpG位点在仅人DNA中被甲基化的概率(中值:0.06;范围:0-1)显著低于仅小鼠DNA的概率(中值:0.93;范围:0-1)(P值<0.0001)。这些结果与测定设计一致，其中人DNA的未甲基化程度更高，因为在全基因组扩增期间不能保留甲基化，而小鼠DNA由于CpG甲基转移酶M.SssI的处理而甲基化程度更高。此外，示出了人-小鼠杂合DNA分子中人DNA组成部分内的CpG位点被甲基化的概率(中值:0.07;范围:0-1)低于小鼠DNA组成部分内的CpG位点被甲基化的概率(中值:0.93;范围:0-1)(P值<0.0001)。这些数据指示，所公开的方法可以准确地测定DNA分子的甲基化状态以及DNA分子内的片段。

[0673] 使用亚硫酸氢盐测序测量人-小鼠杂合片段的甲基化，所述人-小鼠杂合片段的甲基化谱式通过根据本公开的实施例的单分子实时测序来测定。通过超声处理将样本MIX01(人DNA是甲基化的而小鼠DNA是未甲基化的)和MIX02(人DNA是未甲基化的而小鼠DNA是甲基化的)剪切，从而形成DNA片段中值长度为196bp(四分位距:161-268)的混合物。然后在读段长度为300bp x2的MiSeq平台(Illumina)中执行配对末端亚硫酸氢盐测序(BS-Seq)。分别针对MIX01和MIX02获得了370万和290万经过测序的片段，所述片段与人或小鼠参考基因组比对，或与人基因组和小鼠基因组部分比对。对于MIX01，41.6％的比对片段被归类为仅人DNA，56.6％被归类为仅小鼠DNA，并且1.8％被归类为人-小鼠杂合DNA。对于MIX02，61.8％的比对片段被归类为仅人DNA，36.3％被归类为仅小鼠DNA，并且1.9％被归类为人-小鼠杂合DNA。在BS-Seq中测定为人-小鼠杂合DNA的测序片段的百分比(<2％)远低于在太平洋生物科学公司的测序结果中观察到的百分比(>30％)。值得注意的是，长片段(中值为

约2kb)是通过太平洋生物科学公司测序进行测序的,而长片段被剪切为适合于MiSeq的短片段(中值为约196bp)。这种剪切过程将极大地稀释人-小鼠杂合片段。

[0674] 图118示出了比较了通过亚硫酸氢盐测序和太平洋生物科学公司测序测定的MIXO1的甲基化的表格。表格的最左边部分示出了DNA的类型:1)仅人;2)仅小鼠;和3)人-小鼠杂合,其分为人组成部分和小鼠组成部分。表格的中间部分示出了亚硫酸氢盐测序的细节,包含CG位点的数量和甲基化密度。表格的最右边部分示出了太平洋生物科学公司测序的细节,包含CG位点的数量和甲基化密度。

[0675] 如图118所示,在亚硫酸氢盐测序和太平洋生物科学测序结果中,MIXO1的仅人DNA始终显示出高于仅小鼠DNA的甲基化密度。对于人-小鼠杂合片段,在亚硫酸氢盐测序结果中,人组成部分和小鼠组成部分的甲基化水平分别测定为46.8%和2.3%。这些结果证实了如根据本公开的太平洋生物科学公司测序所测定的,与小鼠组成部分相比,人组成部分的甲基化密度更高。对于太平洋生物科学公司测序,在人组成部分观察到57.4%的甲基化密度,并且在小鼠组成部分观察到12.1%的较低甲基化密度。这些结果表明,通过根据本公开的太平洋生物科学公司测序测定的甲基化是可行的。具体地,太平洋生物科学公司测序可以用于测定不同的甲基化密度,包含在某个区段的甲基化密度高于另一个区段的DNA中。观察到,通过根据本公开的太平洋生物科学公司测序测定的甲基化密度相对于亚硫酸氢盐测序更高。可以使用由这两种技术测定的结果之间的差异来调整这种估计,以便比较各个技术的结果。

[0676] 图119示出了比较了通过亚硫酸氢盐测序和太平洋生物科学公司测序测定的MIXO2的甲基化的表格。表格的最左边部分示出了DNA的类型:1)仅人;2)仅小鼠;和3)人-小鼠杂合,其分为人组成部分和小鼠组成部分。表格的中间部分示出了亚硫酸氢盐测序的细节,包含CG位点的数量和甲基化密度。表格的最右边部分示出了太平洋生物科学公司测序的细节,包含CG位点的数量和甲基化密度。

[0677] 如图119所示,在亚硫酸氢盐测序和太平洋生物科学测序结果中,MIXO2的仅人DNA始终显示出低于仅小鼠DNA的甲基化密度。对于人-小鼠杂合片段,在亚硫酸氢盐测序结果中,人组成部分和小鼠组成部分的甲基化水平分别测定为1.8%和67.4%。这些结果进一步证实了如根据本公开的太平洋生物科学公司测序所确定的,与小鼠组成部分相比,人组成部分的甲基化密度更低。对于太平洋生物科学公司测序,如通过根据本公开的太平洋生物科学公司测序所测定的,在人组成部分观察到13.1%的甲基化密度,并且在小鼠组成部分观察到72.2%的较高甲基化密度。还表明通过根据本公开的太平洋生物科学公司测序测定甲基化是可行的。具体地,太平洋生物科学公司测序可以用于测定不同的甲基化密度,包含在某个区段的甲基化密度低于另一个区段的DNA中。还观察到,通过根据本公开的太平洋生物科学公司测序测定的甲基化密度相对于亚硫酸氢盐测序更高。可以使用由这两种技术测定的结果之间的差异来调整这种估计,以便比较各个技术的结果。

[0678] 图120A示出了MIXO1的仅人和仅小鼠DNA的5Mb分类中的甲基化水平。图120B示出了MIXO2的仅人和仅小鼠DNA的5Mb分类中的甲基化水平。在两个图中,百分比形式的甲基化水平在y轴上示出。仅人DNA和仅小鼠DNA中的每一个的亚硫酸氢盐测序和太平洋生物科学公司测序在x轴上示出。

[0679] 发现通过根据本公开的太平洋生物科学公司测序测定的图120A和图120B中的结

果在样本MIX01和MIX02中跨各个分类系统地更高。

[0680] 图121A示出了MIX01的人-小鼠杂合DNA片段的人组成部分和小鼠组成部分的5Mb分类中的甲基化水平。图121B示出了MIX02的人-小鼠杂合DNA片段的人组成部分和小鼠组成部分的5Mb分类中的甲基化水平。在两个图中,百分比形式的甲基化水平在y轴上示出。人组成部分DNA和小鼠组成部分DNA中的每一个的亚硫酸氢盐测序和太平洋生物科学公司测序在x轴上示出。

[0681] 图121A和图121B都示出了与亚硫酸氢盐测序相比当使用太平洋生物科学公司测序时甲基化水平增加。此增加类似于图120A和图120B中用仅人DNA和仅小鼠DNA观察到的通过太平洋生物科学公司测序实现的甲基化水平增加。杂合片段的亚硫酸氢盐测序结果中存在的5Mb分类中甲基化水平可变性的增加可能是由于用于分析的CpG位点数量较少。

[0682] 图122A和122B是代表性图,其示出了单个人-小鼠杂合分子中的甲基化状态。图122A示出了样本MIX01中的人-小鼠杂合片段。图122B示出了样本MIX02中的人-小鼠杂合片段。实心圆圈指示甲基化位点,而空心圆圈指示未甲基化位点。根据本文所述的实施例测定这些片段中的甲基化状态。

[0683] 如图122A所示,来自样本MIX01的杂合分子的人组成部分被测定为甲基化程度更高。相比之下,小鼠DNA组成部分被测定为甲基化程度更低。相比之下,图122B示出来自样本MIX02的杂合分子的人组成部分被测定为甲基化程度更低,而小鼠DNA组成部分被测定为甲基化程度更高。

[0684] 这些结果表明,本公开中呈现的实施例允许通过分子的不同组成部分中的不同甲基化谱式来测定单个DNA分子中的甲基化变化。在一个实施例中,可以测量基因或其它基因组区域的甲基化状态,其中基因或基因组区域的不同组成部分将展现出不同的甲基化状态(例如,启动子对基因体)。在另一个实施例中,本文呈现的方法可以检测人-小鼠杂合片段,从而提供一种用于检测相对于参考基因组包含非连续片段的DNA分子(即嵌合分子)并分析其甲基化状态的通用方法。例如,可以使用此方法来分析(但不限于)基因融合、基因组重排、翻译、倒置、重复、结构变异、病毒DNA整合、减数分裂重组等。

[0685] 在一些实施例中,这些杂合片段可以在测序之前使用基于探针的杂合方法或CRISPR-Cas系统或用于靶DNA富集的变体方法进行富集。最近,据报道,来自蓝藻细菌(霍氏双岐藻(Scytonema hofmanni))的CRISPR相关的转座酶能够将DNA片段插入到所关注的靶位点附近的区域中(Strecker等人《科学(Science)》2019;365:48-53)。CRISPR相关的转座酶可以像Tn7介导的转座一样起作用。在一个实施例中,可以调整此CRISPR相关的转座酶,以在gRNA的指导下将例如标记有生物素的注释序列插入一个或多个所关注的基因组区域。可以使用涂覆有例如链霉亲和素的磁珠来捕获注释序列,由此根据本公开的实施例同时拉下靶向DNA序列以进行测序和甲基化分析。

[0686] 在一些实施例中,可以通过使用限制酶来富集片段,所述限制酶可以包含本文公开的任何限制酶。

[0687] C.示例嵌合分子检测方法

[0688] 图123示出了检测生物样本中的嵌合分子的方法1230。嵌合分子可以包含来自两种不同基因、染色体、细胞器(例如,线粒体、细胞核、叶绿体)、生物体(哺乳动物、细菌、病毒等)和/或物种的序列。方法1230可以应用于来自生物样本的多个DNA分子中的每个分子。在

一些实施例中，多个DNA分子可以是细胞DNA。在其它实施例中，多个DNA分子可以是来自孕妇血浆的细胞游离DNA分子。

[0689] 在框1232处，可以对DNA分子执行单分子测序，以获得在N个位点中的每个位点处提供甲基化状态的序列读段。N可以是5或更大，包含5到10、10到15、15到20或大于20。序列读段的甲基化状态可以形成甲基化谱式。DNA分子可以是多个DNA分子中的一个DNA分子，并且可以对多个DNA分子执行方法1230。甲基化谱式可以采取各种形式。例如，谱式可以是N个(例如，2个、3个、4个等)甲基化位点，随后是N个未甲基化位点，反之亦然。甲基化的这种变化可以指示接合点。甲基化的连续位点的数量可以与未甲基化的连续位点的数量不同。

[0690] 在框1234处，可以使甲基化谱式略过对应于嵌合分子的一种或多种参考谱式，所述嵌合分子具有来自参考人类基因组的两个组成部分的两个部分。参考谱式可以充当过滤器，以鉴定指示接合点的匹配谱式。可以追踪与参考谱式匹配的位点的数量，使得匹配位置对应于最大数量的匹配位点(即，甲基化状态与参考谱式匹配的数量)。参考人类基因组的两个组成部分可以是参考人类基因组的不连续组成部分。参考人类基因组的两个组成部分可以相隔超过1kb、5kb、10kb、100kb、1Mb、5Mb或10Mb。两个组成部分可能来自两个不同的染色体臂或染色体。一种或多种参考谱式可以包含甲基化状态与未甲基化状态之间的变化。

[0691] 在框1236处，可以鉴定甲基化谱式与一个或多个参考谱式的第一参考谱式之间的匹配位置。匹配位置可以鉴定序列读段中的参考人类基因组的两个组成部分之间的接合点。匹配位置可以对应于参考谱式与甲基化谱式之间的重叠函数中的最大值。重叠函数可以使用多个参考谱式，其中输出可能是聚集函数的最大值(即，每个参考谱式贡献一个输出值)或者是跨参考谱式鉴定的单个最大值。

[0692] 在框1238处，可以输出接合点作为嵌合分子中的基因融合的位置。可以将基因融合的位置与各种病症或疾病(包含癌症)的基因融合的参考位置进行比较。可以治疗生物体的病症或疾病，从所述生物体中获得生物样本。

[0693] 匹配位置可以输出到比对函数。可以完善基因融合的位置。完善基因融合的位置可以包含将序列读段的第一部分与参考人类基因组的第一组成部分比对。第一部分可以在接合点之前。完善基因融合的位置可以包含将序列读段的第二部分与参考人类基因组的第二组成部分比对。第二部分可以在接合点之后。参考人类基因组的第一组成部分可以与人类参考基因组的第二组成部分相距至少1kb。例如，参考人类基因组的第一组成部分可以与人类参考基因组的第二组成部分相距1.0到1.5kb、1.5到2.0kb、2.0到2.5kb、2.5到3.0kb、3到5kb或多于5kb。

[0694] 多个嵌合分子的接合点可以相互比较，以确认基因融合的位置。

[0695] VIII.结论

[0696] 开发了一种预测在单碱基分辨率下核酸的碱基修饰(例如甲基化)水平的高效方法。此新方法实施一种用于同时捕获被调查的碱基周围的聚合酶动力学、序列上下文和链信息的新方案。这种新的动力学转换使得动力学脉冲中发生的细微中断能够被鉴定和建模。与仅使用IPD的先前方法相比，本专利申请中提出的新方法大大提高了甲基化分析的分辨率和准确性。此新方案可以轻松地扩展用于其它目的，例如，检测5hmC(5-羟甲基胞嘧啶)、5fC(5-甲酰基胞嘧啶)、5caC(5-羧甲基胞嘧啶)、4mC(4-甲基胞嘧啶)、6mA(N6-甲基腺嘌呤)、8oxoG(7,8-二氢-8-氧鸟嘌呤)、8oxoA(7,8-二氢-8-氧鸟嘌呤)和其它形式的碱基修

饰以及DNA损伤。在另一个实施例中，此新方案(例如，类似于本申请中呈现的2-D数字矩阵的动力学转换)可以通过使用纳米孔测序系统用于碱基修饰分析。

[0697] 此甲基化检测实施方案可以用于来自不同来源的核酸样本(例如，细胞核酸)、来自环境采样的核酸(例如细胞污染物)、来自病原体的核酸(例如细菌和真菌)以及孕妇血浆中的cfDNA。这将为基因组研究和分子诊断开辟许多新的可能性，如无创产前检测、癌症检测和移植监测。对于基于cfDNA的无创产前诊断，这项新发明使得在诊断中同时使用每个分子的拷贝数异常、长度、突变、片段末端和碱基修饰而无需在测序前进行PCR和实验转化成为可能，从而增强了灵敏度。可以使用本文所述的方法检测单倍型之间甲基化水平的失衡。此类失衡可以指示DNA分子(例如，从病症中提取，如从癌症患者的血液中分离出的癌细胞)或病症的起源。

[0698] IX.示例系统

[0699] 图124展示了根据本发明的实施例的测量系统12400。所示系统包含样本12405，如样本架12410内的DNA分子，其中样本12405可以与测试剂12408接触以提供物理特性12415的信号。样本架的实例可以是流动槽，其包括分析器的探针和/或引物或液滴通过其移动的管(液滴包括于分析器中)。检测器12420检测来自样本的物理特性12415(例如，荧光强度、电压或电流)。检测器12402可以间隔地(例如，周期性间隔)进行测量以获得构成数据信号的数据点。在一个实施例中，模数转换器多次将来自检测器的模拟信号转换为数字形式。样本架12401和检测器12402可以形成测定装置，例如，根据本文所描述的实施例进行测序的测序装置。将数据信号12425从检测器12402发送到逻辑系统12403。数据信号12425可以存储在本地存储器12435、外部存储器12404或存储装置12445中。

[0700] 逻辑系统12403可以是或可以包含计算机系统、ASIC、微处理器等。其还可以包含显示器(例如，监视器、LED显示器等)和用户输入装置(例如，鼠标、键盘、按钮等)或与所述显示器和用户输入装置耦接。逻辑系统12403和其它组件可以是独立的或网络连接的计算机系统的一部分，或者逻辑系统可以直接附接到或结合在包含检测器12402和/或样本固持器12401的装置(例如，测序装置)中。逻辑系统12403还可以包含在处理器12405中执行的软件。逻辑系统12403可以包含存储用于控制系统12400以执行本文所描述的任何方法的指令的计算机可读介质。例如，逻辑系统12403可以向包含样本固持器12401的系统提供命令，使得测序或其它物理操作得以执行。可以按特定的顺序执行此类物理操作，例如，按特定的顺序添加和去除试剂。此类物理操作可以由机器人系统(例如，包含机械臂的机器人系统)执行，其可以用于获取样本并执行测定。

[0701] 本文提到的任何计算机系统均可以使用任何合适数量的子系统。此类子系统的实例如图125所示在计算机系统10中。在一些实施例中，计算机系统包含单个计算机设备，其中子系统可以是计算机设备的组件。在其它实施例中，计算机系统可以包含具有内部组件的多个计算机设备，每个计算机设备是子系统。计算机系统可以包含台式计算机和膝上型计算机、平板计算机、移动电话、其它移动装置和基于云的系统。

[0702] 图125中所示的子系统通过系统总线75互连。示出了另外的子系统，如打印机74、键盘78、一个或多个存储装置79、耦接到显示适配器82的监视器76(例如，显示屏，如LED)等。耦接到I/O控制器71的外围装置和输入/输出(I/O)装置可以通过任何数量的如输入/输出(I/O)端口77(例如，USB、FireWire®)等本领域已知的装置连接到计算机系统。例如，I/O

端口77或外部接口81(例如以太网、Wi-Fi等)可以用于将计算机系统10连接到广域网,如因特网、鼠标输入装置或扫描仪。通过系统总线75的互连允许中央处理器73与每个子系统连通并且控制来自系统存储器72或存储装置79(例如固定磁盘,如硬盘驱动器或光盘)的多个指令的执行,以及子系统之间的信息交换。系统存储器72和/或存储装置79可以体现为计算机可读媒体。另一种子系统是数据收集装置85,如相机、麦克风、加速计等。本文中提及的任何数据可以从一个组件输出到另一个组件且可以输出到用户。

[0703] 计算机系统可以包含例如通过外部接口81、通过内部接口或经由可以从一个组件连接到另一个组件并从另一个组件移除的可移除存储装置连接在一起的多个相同的组件或子系统。在一些实施例中,计算机系统、子系统或设备可以通过网络进行通信。在这种情况下,一台计算机可以被视为客户端,而另一台计算机可以被视为服务器,其中每台计算机可以是同一计算机系统的一部分。客户端和服务器可以各自包含多个系统、子系统或组件。

[0704] 实施例的各方面可以使用硬件电路系统(例如,专用集成电路或现场可编程门阵列)以控制逻辑的形式实施,和/或以模块化或集成方式使用具有一般可编程处理器的计算机软件实施。如本文所使用的,处理器可以包含单核处理器、同一集成芯片上的多核处理器,或者单个电路板上或联网的多个处理单元,以及专用硬件。基于本文提供的公开内容和教导,本领域普通技术人员将了解并且意识到使用硬件以及硬件和软件的组合来实现本发明的实施例的其它方式和/或方法。

[0705] 本申请中描述的任何软件组件或功能可以实施为由处理器使用任何合适的计算机语言,例如,Java、C、C++、C#、Objective-C、Swift或如Perl或Python等脚本语言使用例如,常规或面向对象的技术执行的软件代码。软件代码可以存储为计算机可读媒体上用于存储和/或传输的一系列指令或命令。合适的非暂时性计算机可读媒体可以包含随机存取存储器(RAM)、只读存储器(ROM)、如硬盘驱动器或软盘等磁媒体或如光盘(CD)或DVD(数字通用光盘)或蓝光光盘等光学媒体、闪存等。计算机可读介质可以是存储或传输装置的任何组合。

[0706] 也可以使用适于经由符合各种协议的有线、光学和/或无线网络(包含因特网)传输的载波信号来编码和传输此类程序。因此,计算机可读媒体可以使用以这类程序编码的数据信号产生。以程序代码编码的计算机可读媒体可以与兼容装置一起封装或与其它装置分开地提供(例如,通过因特网下载)。任何此类计算机可读媒体可以驻留在单个计算机产品(例如,硬盘驱动器、CD或整个计算机系统)上或内,并且可以存在于系统或网络内的不同计算机产品之上或之内。计算机系统可以包含监视器、打印机或其它合适的显示器,以便向用户提供本文提到的任何结果。

[0707] 本文描述的任何方法可以用包含一个或多个处理器的计算机系统完全或部分地执行,所述计算机系统可以被配置成执行步骤。因此,实施例可以针对被配置成执行本文描述的任何方法的步骤的计算机系统,所述计算机系统可能具有执行相应步骤或相应步骤组的不同组件。尽管以编号的步骤呈现,但是本文的方法步骤可以同时或在不同时间或以不同顺序执行。另外,这些步骤的部分可以与其它方法的其它步骤的部分一起使用。而且,步骤的全部或部分可以是任选的。另外,任何方法的任何步骤都可以用模块、单元、电路或用于执行这些步骤的其它系统的装置来执行。

[0708] 本申请还涉及以下实施方案:

[0709] 实施方案1.一种用于检测核酸分子中的核苷酸的修饰的方法,所述方法包括:

[0710] 接收输入数据结构,所述输入数据结构对应于在样本核酸分子中进行测序的核苷酸的窗口,其中通过测量对应于所述核苷酸的光信号脉冲对所述样本核酸分子进行测序,所述输入数据结构包括以下性质的值:

[0711] 对于所述窗口内的每个核苷酸:

[0712] 所述核苷酸的一致性,

[0713] 所述核苷酸的相对于相应窗口内的靶位置的位置,

[0714] 对应于所述核苷酸的所述脉冲的宽度,以及

[0715] 脉冲间持续时间,其表示对应于所述核苷酸的所述脉冲与对应于相邻核苷酸的脉冲之间的时间;

[0716] 将所述输入数据结构输入到模型中,所述模型通过以下进行训练:

[0717] 接收第一多个第一数据结构,所述第一多个数据结构中的每个第一数据结构对应于在多个第一核酸分子中的相应核酸分子中进行测序的核苷酸的相应窗口,其中通过测量对应于所述核苷酸的信号脉冲对所述第一核酸分子中的每一个进行测序,其中每个第一核酸分子的每个窗口中的靶位置处的核苷酸中的所述修饰具有已知的第一状态,每个第一数据结构包括性质与所述输入数据结构相同的值,

[0718] 存储多个第一训练样本,每个第一训练样本包含所述第一多个第一数据结构之一和指示所述靶位置处的所述核苷酸的所述第一状态的第一标记,以及

[0719] 当所述第一多个第一数据结构被输入到所述模型时,使用所述多个第一训练样本,基于所述模型的与所述第一标记的对应标记匹配或不匹配的输出来优化所述模型的参数,其中所述模型的输出指明相应窗口中的所述靶位置处的所述核苷酸是否具有所述修饰,

[0720] 使用所述模型确定所述输入数据结构中的窗口内的所述靶位置处的核苷酸中是否存在所述修饰。

[0721] 实施方案2.根据实施方案1所述的方法,其中:

[0722] 所述输入数据结构是多个输入数据结构中的一个输入数据结构,

[0723] 所述样本核酸分子是多个样本核酸分子中的一个样本核酸分子,

[0724] 所述多个样本核酸分子从受试者的生物样本中获得,并且

[0725] 每个输入数据结构对应于在所述多个样本核酸分子中的相应样本核酸分子中进行测序的核苷酸的相应窗口,并且

[0726] 所述方法进一步包括:

[0727] 接收所述多个输入数据结构,

[0728] 将所述多个输入数据结构输入到所述模型中,以及

[0729] 使用所述模型确定每个输入数据结构的相应窗口中的靶位置处的核苷酸中是否存在修饰。

[0730] 实施方案3.根据实施方案2所述的方法,其进一步包括:

[0731] 测定所述修饰存在于一个或多个核苷酸处,以及

[0732] 利用一个或多个核苷酸处的所述修饰的存在来确定病症的分类。

[0733] 实施方案4.根据实施方案3所述的方法,其中所述病症包括癌症。

[0734] 实施方案5.根据实施方案3所述的方法,其进一步包括:

[0735] 确定所述病症的分类是所述受试者患有所述病症,以及

[0736] 治疗所述受试者的所述病症。

[0737] 实施方案6.根据实施方案3所述的方法,其中确定所述病症的分类是利用修饰的数量或所述修饰的位点。

[0738] 实施方案7.根据实施方案2所述的方法,其进一步包括:

[0739] 确定所述修饰存在于一个或多个核苷酸处,以及

[0740] 利用一个或多个核苷酸处的所述修饰的存在来确定临床相关的DNA浓度分率、胎儿甲基化谱、母体甲基化谱、印记基因区域的存在或DNA来源的组织起源组织。

[0741] 实施方案8.根据实施方案2所述的方法,其中所述多个样本核酸分子中的每个样本核酸分子的长度尺寸大于截止长度尺寸。

[0742] 实施方案9.根据实施方案2所述的方法,其中:

[0743] 将所述多个样本核酸分子与多个基因组区域比对,

[0744] 对于所述多个基因组区域中的每个基因组区域:

[0745] 将许多样本核酸分子与所述基因组区域比对,

[0746] 样本核酸分子的数量大于截止数量。

[0747] 实施方案10.根据实施方案1所述的方法,其进一步包括对所述样本核酸分子进行测序。

[0748] 实施方案11.根据实施方案1所述的方法,其中所述模型包含机器学习模型、主成分分析主分量分析、卷积神经网络或逻辑回归。

[0749] 实施方案12.根据实施方案1所述的方法,其中:

[0750] 对应于所述输入数据结构的核苷酸的窗口包括所述样本核酸分子的第一链上的核苷酸和所述样本核酸分子的第二链上的核苷酸,并且

[0751] 对于所述窗口内的每个核苷酸,所述输入数据结构进一步包括链性质值,所述链性质指示所述核苷酸存在于所述第一链或所述第二链上。

[0752] 实施方案13.根据实施方案12所述的方法,其中所述样本核酸分子是通过以下形成的环状DNA分子:

[0753] 使用Cas9复合物切割双链DNA分子以形成经过切割的双链DNA分子,以及

[0754] 将发夹衔接子连接到所述经过切割的双链DNA分子的末端上。

[0755] 实施方案14.根据实施方案1所述的方法,其中使用环状一致性序列环状共有序列并且在无需将经过测序的核苷酸与参考基因组比对的情况下测定所述窗口内的所述核苷酸。

[0756] 实施方案15.根据实施方案1所述的方法,其中所述窗口内的每个核苷酸被富集或过滤。

[0757] 实施方案16.根据实施方案15所述的方法,其中所述窗口内的每个核苷酸通过以下进行富集:

[0758] 使用Cas9复合物切割双链DNA分子以形成经过切割的双链DNA分子,并且将发夹衔接子连接到所述经过切割的双链DNA分子的末端上,或者

[0759] 通过以下进行过滤:

[0760] 选择长度尺寸在一定长度尺寸范围内的双链DNA分子。

[0761] 实施方案17.根据实施方案1所述的方法,其中在无需使用环状一致性序列环状共有序列并且无需将经过测序的核苷酸与参考基因组比对的情况下测定所述窗口内的核苷酸。

[0762] 实施方案18.一种用于检测核酸分子中的核苷酸的修饰的方法,所述方法包括:

[0763] 接收第一多个第一数据结构,所述第一多个第一数据结构中的每个第一数据结构对应于在多个第一核酸分子中的相应核酸分子中进行测序的核苷酸的相应窗口,其中通过测量对应于所述核苷酸的光信号脉冲对所述第一核酸分子中的每一个进行测序,其中每个第一核酸分子的每个窗口中的靶位置处的核苷酸中的所述修饰具有已知的第一状态,每个第一数据结构包括以下性质的值:

[0764] 对于所述窗口内的每个核苷酸:

[0765] 所述核苷酸的一致性,

[0766] 所述核苷酸的相对于相应窗口内的所述靶位置的位置,

[0767] 对应于所述核苷酸的所述脉冲的宽度,以及

[0768] 脉冲间持续时间,其表示对应于所述核苷酸的所述脉冲与对应于相邻核苷酸的脉冲之间的时间,

[0769] 存储多个第一训练样本,每个第一训练样本包含所述第一多个第一数据结构之一和指示所述靶位置处的所述核苷酸的所述修饰的所述第一状态的第一标记,以及

[0770] 当所述第一多个第一数据结构被输入到模型时,通过基于所述模型的与所述第一标记的对应标记匹配或不匹配的输出来优化所述模型的参数,使用所述多个第一训练样本来训练所述模型,其中所述模型的输出指明相应窗口中的所述靶位置处的所述核苷酸是否具有所述修饰。

[0771] 实施方案19.根据实施方案18所述的方法,其进一步包括:

[0772] 接收第二多个第二数据结构,所述第二多个第二数据结构中的每个第二数据结构对应于在多个第二核酸分子中的相应核酸分子中进行测序的核苷酸的相应窗口,其中每个第二核酸分子的每个窗口内的靶位置处的核苷酸中的所述修饰具有已知的第二状态,每个第二数据结构包括性质与所述第一多个第一数据结构相同的值;

[0773] 存储多个第二训练样本,每个第二训练样本包含所述第二多个第二数据结构之一和指示所述靶位置处的所述核苷酸的所述第二状态的第二标记;

[0774] 其中训练:

[0775] 所述第一状态或所述第二状态是所述修饰存在,而另一种状态是所述修饰不存在,

[0776] 所述模型进一步包括当所述第二多个第二数据结构被输入到所述模型时,通过基于所述模型的与所述第二标记的对应标记匹配或不匹配的输出优化所述模型的参数来使用所述多个第二训练样本。

[0777] 实施方案20.根据实施方案19所述的方法,其中所述多个第一核酸分子与所述多个第二核酸分子相同。

[0778] 实施方案21.根据实施方案19所述的方法,其中所述修饰包括甲基化,其中所述多个第一核酸分子是使用多重置换扩增、通过第一类型的甲基化核苷酸产生的,并且其中所

述多个第二核酸分子是使用多重置换扩增、通过第一类型的非甲基化核苷酸产生的。

[0779] 实施方案22.根据实施方案18所述的方法,其中所述光信号是来自染料标记的核苷酸的荧光信号。

[0780] 实施方案23.根据实施方案18所述的方法,其中与所述第一多个数据结构相关联的每个窗口包括每个第一核酸分子的第一链上的4个连续核苷酸。

[0781] 实施方案24.根据实施方案23所述的方法,其中与所述第一多个数据结构相关联的所述窗口包括相同数量的连续核苷酸。

[0782] 实施方案25.根据实施方案18所述的方法,其中:

[0783] 与所述第一多个数据结构相关联的每个窗口包括所述第一核酸分子的第一链上的核苷酸和所述第一核酸分子的第二链上的核苷酸,并且

[0784] 对于所述窗口内的每个核苷酸,每个第一数据结构进一步包括链性质值,所述链性质指示所述核苷酸存在于所述第一链或所述第二链上。

[0785] 实施方案26.根据实施方案18所述的方法,其中所述相邻核苷酸是邻近核苷酸。

[0786] 实施方案27.根据实施方案18所述的方法,其中所述脉冲的宽度是所述脉冲在所述脉冲的最大值一半处的宽度。

[0787] 实施方案28.根据实施方案18所述的方法,其中所述脉冲间持续时间是与所述核苷酸相关联的脉冲的最大值与和所述相邻核苷酸相关联的脉冲的最大值之间的时间。

[0788] 实施方案29.根据实施方案18所述的方法,其中所述模型包括卷积神经网络,所述卷积神经网络包括:

[0789] 卷积过滤器集合,其被配置成对所述第一多个数据结构进行过滤,

[0790] 输入层,其被配置成接收经过过滤的第一多个数据结构,

[0791] 包含多个节点的多个隐藏层,所述多个隐藏层的第一层联接到所述输入层;以及

[0792] 输出层,其联接到所述多个隐藏层的最后一层并且被配置成将输出数据结构输出,所述输出数据结构包括所述性质。

[0793] 实施方案30.根据实施方案18所述的方法,其中所述修饰包括所述靶位置处的所述核苷酸的甲基化。

[0794] 实施方案31.根据实施方案30所述的方法,其中所述已知的第一状态包含所述第一数据结构的第一部分的甲基化状态和所述第一数据结构的第二部分的非甲基化状态。

[0795] 实施方案32.根据实施方案30所述的方法,其中所述甲基化包括4mC(N4-甲基胞嘧啶)、5mC(5-甲基胞嘧啶)、5hmC(5-羟甲基胞嘧啶)、5fC(5-甲酰基胞嘧啶)、5caC(5-羧基胞嘧啶)、1mA(N1-甲基腺嘌呤)、3mA(N3-甲基腺嘌呤)、6mA(N6-甲基腺嘌呤)、7mA(N7-甲基腺嘌呤)、3mC(N3-甲基胞嘧啶)、2mG(N2-甲基鸟嘌呤)、6mG(O6-甲基鸟嘌呤)、7mG(N7-甲基鸟嘌呤)、3mT(N3-甲基胸腺嘧啶)或4mT(O4-甲基胸腺嘧啶)。

[0796] 实施方案33.根据实施方案18所述的方法,其中所述修饰包括氧化变化。

[0797] 实施方案34.根据实施方案18所述的方法,其中每个数据结构进一步包括对应于所述窗口内的每个核苷酸的脉冲的高度值。

[0798] 实施方案35.根据实施方案18所述的方法,其中对应于所述核苷酸的所述光信号由所述核苷酸或与所述核苷酸相关联的标签产生。

[0799] 实施方案36.根据实施方案18所述的方法,其中每个靶位置是相应窗口的中心。

[0800] 实施方案37.根据实施方案18所述的方法,其中所述修饰不存在于每个第一核酸分子的每个窗口中。

[0801] 实施方案38.根据实施方案18所述的方法,其中:

[0802] 所述多个第一数据结构中的每个第一数据结构不包含脉冲间持续时间或脉冲宽度低于截止值的第一核酸分子。

[0803] 实施方案39.根据实施方案18所述的方法,其中:

[0804] 所述修饰包括甲基化,并且

[0805] 所述多个第一训练样本通过以下产生:

[0806] 使用核苷酸集合扩增多个核酸分子,其中所述核苷酸集合包含指定比率的6mA。

[0807] 实施方案40.根据实施方案39所述的方法,其中所述甲基化包括6mA(N6-甲基腺嘌呤)。

[0808] 实施方案41.根据实施方案1或实施方案18所述的方法,其中所述多个第一核酸分子中的至少一些第一核酸分子各自包含对应于第一参考序列的第一部分和对应于与所述第一参考序列不相交的第二参考序列的第二部分。

[0809] 实施方案42.根据实施方案1或实施方案18所述的方法,其进一步包括:

[0810] 使用多个嵌合核酸分子验证所述模型,每个嵌合核酸分子包含对应于第一参考序列的第一部分和对应于第二参考序列的第二部分,其中所述第一部分具有第一甲基化谱式,并且所述第二部分具有第二甲基化谱式。

[0811] 实施方案43.根据实施方案41或实施方案42所述的方法,其中所述第一部分用甲基化酶进行处理。

[0812] 实施方案44.根据实施方案43所述的方法,其中所述第二部分对应于所述第二参考序列的非甲基化部分。

[0813] 实施方案45.根据实施方案41或实施方案42所述的方法,其中所述第一参考序列是人,并且其中所述第二参考序列来自不同的动物。

[0814] 实施方案46.一种分析生物体的生物样本的方法,所述生物体在第一染色体区域中具有第一单倍型和第二单倍型,所述生物样本包含DNA分子,所述方法包括:

[0815] 分析来自所述生物样本的多个DNA分子,其中分析DNA分子包含:

[0816] 鉴定所述DNA分子在参考人类基因组中的位置;

[0817] 测定所述DNA分子的相应等位基因;以及

[0818] 确定所述DNA分子在一个或多个基因组位点处是否被甲基化;

[0819] 鉴定所述第一染色体区域的第一部分的一个或多个杂合基因座,每个杂合基因座包含所述第一单倍型中的对应的第一等位基因和所述第二单倍型中的对应的第二等位基因;

[0820] 鉴定所述多个DNA分子的第一集合,每个DNA分子:

[0821] 定位于所述一个或多个杂合基因座中的任一个杂合基因座处,

[0822] 包含所述杂合基因座的所述对应的第一等位基因,并且

[0823] 包含N个基因组位点中的至少一个基因组位点,N是大于或等于一的整数;

[0824] 使用所述多个DNA分子的所述第一集合测定所述第一单倍型的所述第一部分的第一甲基化水平;

[0825] 鉴定所述多个DNA分子的第二集合,每个DNA分子:

[0826] 定位于所述一个或多个杂合基因座中的任一个杂合基因座处,

[0827] 包含所述杂合基因座的所述对应的第二等位基因,并且

[0828] 包含所述N个基因组位点中的至少一个基因组位点;

[0829] 使用所述多个DNA分子的所述第二集合测定所述第二单倍型的所述第一部分的第二甲基化水平;

[0830] 使用所述第一甲基化水平和所述第二甲基化水平计算参数值;

[0831] 将所述参数值与参考值进行比较;以及

[0832] 使用所述参数值与所述参考值的比较来确定所述生物体中的病症的分类。

[0833] 实施方案47.根据实施方案46所述的方法,其中所述第一甲基化水平是使用所述多个DNA分子的所述第一集合的单链甲基化水平来测定的,并且其中所述第二甲基化水平是使用所述多个DNA分子的所述第二集合的单链甲基化水平来测定的。

[0834] 实施方案48.根据实施方案46所述的方法,其中所述第一甲基化水平是使用所述多个DNA分子的所述第一集合的单分子双链DNA甲基化水平来测定的,并且其中所述第二甲基化水平是使用所述多个DNA分子的所述第二集合的单分子双链DNA甲基化来测定的。

[0835] 实施方案49.根据实施方案46所述的方法,其中所述病症是癌症。

[0836] 实施方案50.根据实施方案46所述的方法,其中所述参数是分离值。

[0837] 实施方案51.根据实施方案46所述的方法,其进一步包括:

[0838] 测定所述第一单倍型的多个部分的多个第一甲基化水平,

[0839] 测定所述第二单倍型的多个部分的多个第二甲基化水平,所述第二单倍型的多个部分中的每个部分与所述第一单倍型的所述多个部分中的一部分互补,

[0840] 对于所述第二单倍型的所述多个部分中的每个部分:

[0841] 使用所述第二单倍型的所述部分的所述第二甲基化水平和所述第一单倍型的互补部分的所述第一甲基化水平计算分离值,并且

[0842] 将所述分离值与截止值进行比较,

[0843] 其中:

[0844] 所述第一单倍型的第一部分与所述第二单倍型的第一部分互补,并且

[0845] 所述参数包含所述第二单倍型的多个部分,其中所述分离值超过所述截止值。

[0846] 实施方案52.根据实施方案51所述的方法,其中所述截止值通过不患有所述病症的组织来确定。

[0847] 实施方案53.根据实施方案51所述的方法,其中所述第一单倍型的所述多个部分中的每个部分的长度大于或等于5kb。

[0848] 实施方案54.根据实施方案46所述的方法,其进一步包括:

[0849] 测定所述第一单倍型的多个部分的多个第一甲基化水平,

[0850] 测定所述第二单倍型的多个部分的多个第二甲基化水平,所述第二单倍型的所述多个部分中的每个部分与所述第一单倍型的所述多个部分中的一部分互补,

[0851] 对于所述第二单倍型的所述多个部分中的每个部分:

[0852] 使用所述第二单倍型的所述部分的所述第二甲基化水平和所述第一单倍型的互补部分的所述第一甲基化水平计算分离值,

[0853] 其中：

[0854] 所述第一单倍型的第一部分与所述第二单倍型单元型的第一部分互补，并且

[0855] 所述参数包含所述分离值的总和。

[0856] 实施方案55.根据实施方案46所述的方法，其进一步包括：

[0857] 测定所述第一单倍型的多个部分的多个第一甲基化水平，

[0858] 测定所述第二单倍型的多个部分的多个第二甲基化水平，所述第二单倍型的所述多个部分中的每个部分与所述第一单倍型的所述多个部分中的一部分互补，

[0859] 对于所述第二单倍型的所述多个部分中的每个部分：

[0860] 使用所述第二单倍型的所述部分的所述第二甲基化水平和所述第一单倍型的互补部分的所述第一甲基化水平计算分离值，并且

[0861] 将所述分离值与截止值进行比较，以鉴定所述部分是否在所述第一甲基化水平与所述第二甲基化水平之间存在异常分离，

[0862] 其中确定所述生物体中的所述病症的分类包含将具有异常分离的部分的谱式与参考谱式进行比较。

[0863] 实施方案56.根据实施方案46所述的方法，其中所述病症的分类是所述病症的概率。

[0864] 实施方案57.根据实施方案46所述的方法，其中：

[0865] 所述第一单倍型的所述第一部分和所述第二单倍型的所述第一部分形成环状DNA分子，并且

[0866] 测定所述第一单倍型的所述第一部分的所述第一甲基化水平包括使用来自所述环状DNA分子的数据。

[0867] 实施方案58.根据实施方案57所述的方法，其中所述环状DNA分子通过以下形成：

[0868] 使用Cas9复合物切割双链DNA分子以形成经过切割的双链DNA分子，以及

[0869] 将发夹衔接子连接到所述经过切割的双链DNA分子的末端上。

[0870] 实施方案59.根据实施方案46所述的方法，其中：

[0871] 所述第一单倍型的所述第一部分大于或等于1kb。

[0872] 实施方案60.根据实施方案46所述的方法，其中所述参考值是使用不患有所述病症的参考组织来确定的。

[0873] 实施方案61.根据实施方案46所述的方法，其中所述病症是印记基因失调性疾病印记病症。

[0874] 实施方案62.一种检测生物样本中的嵌合分子的方法，所述方法包括：

[0875] 对于来自所述生物样本的多个DNA分子中的每个DNA分子：

[0876] 对所述DNA分子执行单分子测序以获得序列读段，所述序列读段提供了N个位点中的每个位点处的甲基化状态，N为5或更大，其中所述序列读段的甲基化状态形成甲基化谱式；

[0877] 使所述甲基化谱式滑过对应于嵌合分子的一个或多个参考谱式，所述嵌合分子具有来自参考人类基因组的两个组成部分的两个部分，所述一个或多个参考谱式包含甲基化状态与非甲基化状态之间的变化；以及

[0878] 鉴定所述甲基化谱式与所述一个或多个参考谱式中的第一参考谱式之间的匹配

位置,所述匹配位置鉴定所述序列读段中的所述参考人类基因组的所述两个组成部分之间的接合点;以及

[0879] 输出所述接合点作为嵌合分子中的基因融合的位置。

[0880] 实施方案63.根据实施方案62所述的方法,其中所述匹配位置被输出到比对函数,所述方法进一步包括:

[0881] 通过以下完善所述基因融合的位置:

[0882] 将所述序列读段的第一部分与所述参考人类基因组的第一组成部分比对,所述第一部分在所述接合点之前;以及

[0883] 将所述序列读段的第二部分与所述参考人类基因组的第二组成部分比对,所述第二部分在所述接合点之后,其中所述参考人类基因组的所述第一组成部分与所述人类参考基因组的所述第二组成部分相距至少1kb。

[0884] 实施方案64.根据实施方案62所述的方法,其进一步包括比较所述嵌合分子彼此的接合点以确认所述基因融合的位置。

[0885] 实施方案65.一种计算机产品,其包括存储多个指令的非暂时性计算机可读介质,所述多个指令在被执行时控制计算机系统来执行根据前述实施方案中任一项所述的方法。

[0886] 实施方案66.一种系统,其包括:

[0887] 根据实施方案65所述的计算机产品;以及

[0888] 一个或多个处理器,用于执行存储在所述计算机可读介质上的指令。

[0889] 实施方案67.一种系统,其包括用于执行以上方法中的任何方法的装置。

[0890] 实施方案68.一种系统,其包括一个或多个处理器,所述一个或多个处理器被配置成执行以上方法中的任何方法。

[0891] 实施方案69.一种系统,其包括分别执行以上方法中的任何方法的步骤的模块。

[0892] 在不脱离本发明实施例的精神和范围的情况下,可以以任何合适的方式组合特定实施例的具体细节。然而,本发明的其它实施例可以针对涉及每个单独方面的特定实施例,或这些单独方面的特定组合。

[0893] 出于说明和描述的目的,已经呈现了本公开的示例实施例的以上描述。以上描述并非旨在穷举本发明或将本公开限制于所描述的精确形式,并且根据上述教导,许多修改和变化是可能的。

[0894] 除非特别指出相反的情况,否则对“一个(a)”、“一种(an)”或“所述(the)”的引用旨在表示“一个或多个”。除非特别指出相反的情况,否则“或”的使用旨在表示“包含性的或”,而非“排他性的或”。对“第一”组件的引用不一定要求提供第二组件。此外,除非明确说明,否则对“第一”或“第二”组件的引用并不将所引用的组件限制到特定位置。术语“基于”旨在表示“至少部分地基于”。

[0895] 本文所提及的所有专利、专利申请、出版物和描述出于所有目的通过引用整体并入本文。没有一项被承认为是现有技术。

[0896] 参考文献

[0897] Albert,T.J.等人(2007)通过微阵列杂合直接选择人类基因组基因座(Direct selection of human genomic loci by microarray hybridization)《自然方法(Nat.Methods)》,4,903-905。

[0898] Beckmann等人(2014)在低覆盖率和宏基因组学环境中检测表观遗传基序(Detecting epigenetic motifs in low coverage and metagenomics settings)《BMC生物信息学(BMC Bioinformatics)》,15(增刊9):S16。

[0899] Beaulaurier,J.等人(2019)使用现代测序技术破译细菌表观基因组(Deciphering bacterial epigenomes using modern sequencing technologies)《自然评论:遗传学(Nature Reviews Genetics)》,20:157-172。

[0900] Blow,M.J.等人(2016)原核生物的表观基因组景观(The Epigenomic Landscape of Prokaryotes)《公共科学图书馆:遗传学(PLOS Genet.)》,12,e1005854。

[0901] Breiman,L.(2001)随机森林(Random Forests)《机器学习(Mach.Learn.)》,45,5-32。

[0902] Chan,K.C.A.等人(2013)通过血浆DNA亚硫酸氢盐测序无创检测癌症相关的全基因组低甲基化和拷贝数异常(Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing)《美国国家科学院院刊(Proc.Natl.Acad.Sci.U.S.A.)》,110,18761-8。

[0903] Clark,T.A.等人(2013)通过Tet1氧化在单分子实时测序中增强的5-甲基胞嘧啶检测(Enhanced 5-methylcytosine detection in single-molecule,real-time sequencing via Tet1 oxidation)《BMC生物学(BMC Biol.)》,11,4。

[0904] Clark,T.A.等人(2012)使用单分子实时DNA测序表征DNA甲基转移酶特异性(Characterization of DNA methyltransferase specificities using single-molecule,real-time DNA sequencing)《核酸研究》,40:e29。

[0905] Eid,J.等人(2009)单聚合酶分子的实时DNA测序(Real-Time DNA Sequencing from Single Polymerase Molecules)《科学(Science)》323,133-138。

[0906] Feinberg,A.P.和Irizarry,R.A.(2010)作为发展、进化适应和疾病的驱动力的随机表观遗传变异(Stochastic epigenetic variation as a driving force of development,evolutionary adaptation,and disease)《美国国家科学院院刊》,107,1757-1764。

[0907] Feng,Z.等人(2013)通过模拟聚合酶动力学的序列上下文依赖性从SMRT测序数据中检测DNA修饰(Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic)《公共科学图书馆:计算生物学(PLoS Comput Biol.)》,9:e1002935。

[0908] Flusberg,B.A.等人(2010)单分子实时测序期间DNA甲基化的直接检测(Direct detection of DNA methylation during single-molecule,real-time sequencing)《自然方法》,7,461-465。

[0909] Frommer,M.等人(1992)在单个DNA链中产生阳性显示的5-甲基胞嘧啶残基的基因组测序方案(A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands)《美国国家科学院院刊》,89,1827-1831。

[0910] Gai,W.等人(2018)血浆中用于研究具有或不具有肝转移的大肠癌的肝特异性和结肠特异性DNA甲基化标志物(Liver-and colon-specific DNA methylation markers in

plasma for investigation of colorectal cancers with or without liver metastases)《临床化学(Clin.Chem.)》,64,1239-1249。

[0911] Gouil,Q.等人(2019)研究DNA甲基化的最新技术(Latest techniques to study DNA methylation)《生物化学论文(Essays Biochem.)》63(6):639-648。

[0912] Grunau,C.(2001)亚硫酸氢盐基因组测序:关键实验参数的系统研究(Bisulfite genomic sequencing:systematic investigation of critical experimental parameters)《核酸研究》,29,65e-65。

[0913] Herman,J.G.等人(1996)甲基化特异性PCR:用于CpG岛的甲基化状态的新颖PCR测定法(Methylation-specific PCR:a novel PCR assay for methylation status of CpG islands)《美国国家科学院院刊》,93,9821-9826。

[0914] Jiang,P.等人(2014)Methy-Pipe:用于全基因组亚硫酸氢盐测序数据分析的集成生物信息学管道(Methy-Pipe:An Integrated Bioinformatics Pipeline for Whole Genome Bisulfite Sequencing Data Analysis)《公共科学图书馆:综合》,9,e100360。

[0915] LeCun,Y.等人(1989)应用于手写邮政编码识别的反向传播(Backpropagation Applied to Handwritten Zip Code Recognition)《神经计算(Neural Comput.)》,1,541-551。

[0916] Lee,E.-J.等人(2011)通过溶液杂合选择和大规模并行测序进行的靶向亚硫酸氢盐测序(Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing)《核酸研究》,39,e127-e127。

[0917] Lehmann-Werman,R.等人(2016)使用循环DNA甲基化谱式鉴定组织特异性细胞死亡(Identification of tissue-specific cell death using methylation patterns of circulating DNA)《美国国家科学院院刊》,113,E1826-E1834。

[0918] Lister,R.等人(2009)碱基分辨率下的人DNA甲基化组示出广泛的表观基因组差异(Human DNA methylomes at base resolution show widespread epigenomic differences)《自然(Nature)》,462,315-322。

[0919] Liu,Q.等人(2019)通过关于牛津纳米孔测序数据的深度递归神经网络检测DNA碱基修饰(Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data)《自然通讯(Nature Commun.)》,10,2449。

[0920] Liu,Y.等人(2019)碱基分辨率下的5-甲基胞嘧啶和5-羟甲基胞嘧啶的无亚硫酸氢盐直接检测(Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution)《自然生物技术(Nat.Biotechnol.)》,37,424-429。

[0921] Lun,F.M.F.等人(2013)通过母本血浆DNA的全基因组亚硫酸氢盐测序进行的无创产前甲基化组分析(Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA)《临床化学》,59,1583-1594。

[0922] Nattestad,M.等人(2018)乳腺癌细胞系的长读段DNA和RNA测序揭示的复杂重排和癌基因扩增(Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line)《基因组研究》,28,1126-1135。

[0923] Ng,A.Y.(2004)特征选择、L_1对L_2正则化和旋转不变性(Feature selection,L_1vs.L_2regularization,and rotational invariance)第二十一届国际机器学习会议——ICML'04.ACM出版社,美国纽约,第78页。

[0924] Ni,P.等人(2019)深度信号:使用深度学习从纳米孔测序读段中检测DNA甲基化状态(DeepSignal:detecting DNA methylation state from Nanopore sequencing reads using deep-learning)《生物信息学》,35,4586-4595

[0925] Okou,D.T.等人(2007)高通量重新测序的基于微阵列的基因组选择(Microarray-based genomic selection for high-throughput resequencing)《自然方法》,4,907-909。

[0926] Olova,N.等人(2018)全基因组亚硫酸氢盐测序文库制备策略的比较鉴定了影响DNA甲基化数据的偏差来源(Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data)《基因组生物学》,19,33。

[0927] Robertson,K.D.(2005)DNA甲基化与人类疾病(DNA methylation and human disease)《自然评论:遗传学》,6,597-610。

[0928] Smith,Z.D.和Meissner,A.(2013)DNA甲基化:在哺乳动物发育中的作用(DNA methylation:roles in mammalian development)《自然评论:遗传学》,14,204-20。

[0929] Schadt,E.E.等人(2013)在第三代DNA测序数据中模拟动力学速率变化以检测对DNA碱基的假定修饰(Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases)《基因组研究》,23(1):129-41。

[0930] Sun,K.等人(2015)通过全基因组甲基化测序进行血浆DNA组织映射用于无创产前、癌症和移植评估(Plasma DNAtissue mapping by genome-wide methylation sequencing for noninvasive prenatal,cancer,and transplantation assessments)《美国国家科学院院刊》,112,E5503-E5512。

[0931] Suzuki,Y.等人(2016)AgIn:测量单个重复元素的CpG甲基化的景观(AgIn:measuring the landscape of CpG methylation of individual repetitive elements)《生物信息学》,32,2911-2919。

[0932] Watson,C.M.等人(2019)用于精确表征基因组重复的基于Cas9的富集和单分子测序(Cas9-based enrichment and single-molecule sequencing for precise characterization of genomic duplications)《实验室调查(Lab.Investig)》,100,135-146。

[0933] Zhang,W.等人(2015)使用甲基化标记、基因组位置和DNA调控元素预测全基因组DNA甲基化(Predicting genome-wide DNAmethylation using methylation marks,genomic position,and DNA regulatory elements)《基因组生物学》,16,14。

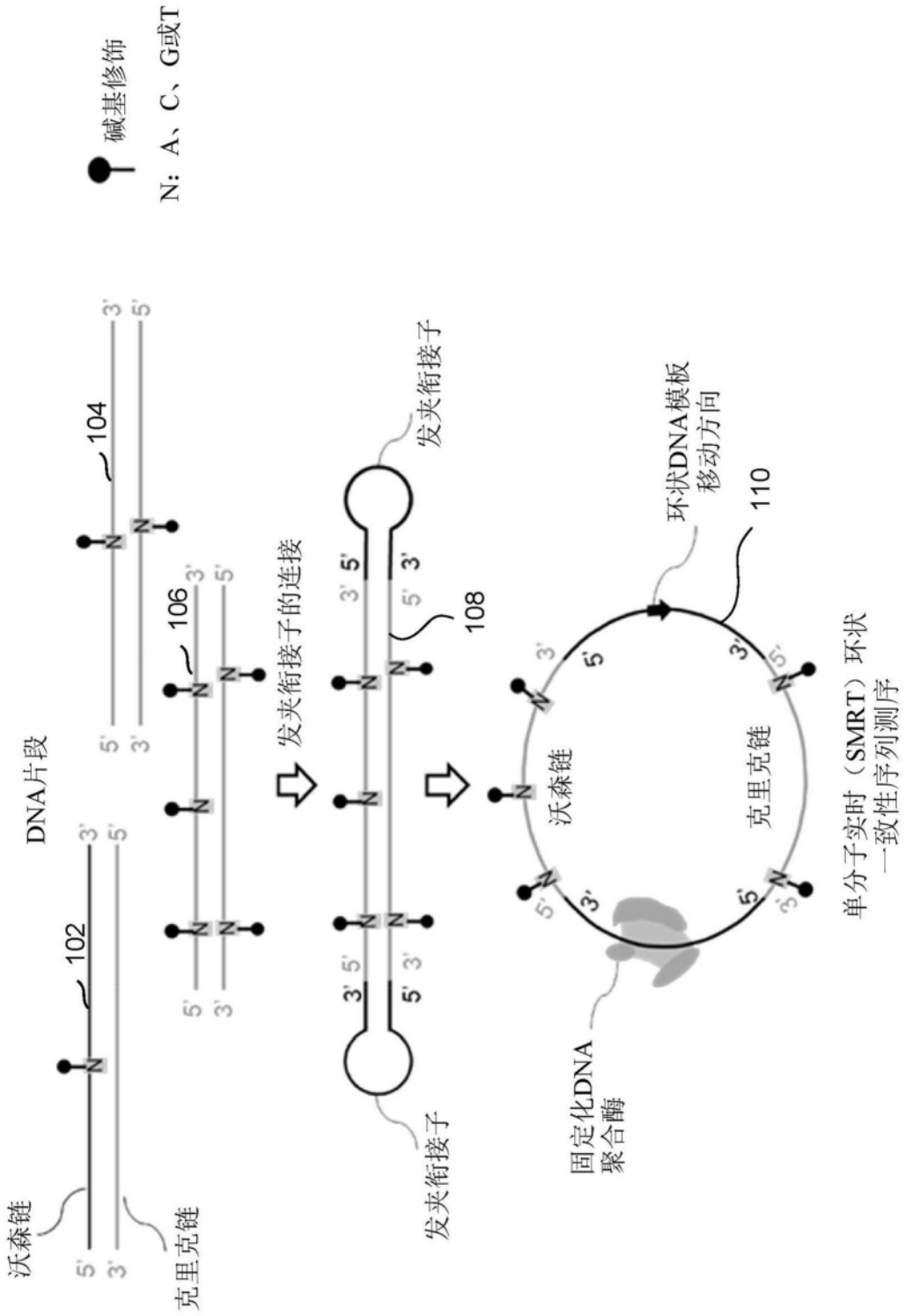

碱基修饰
N：A、C、G或T
DNA片段
102
104
106
沃森链
克里克链
发夹衔接子的连接
发夹衔接子
发夹衔接子
108
环状DNA模板
移动方向
110
固定化DNA
聚合酶
沃森链
克里克链
单分子实时（SMRT）环状
一致性序列测序

图1

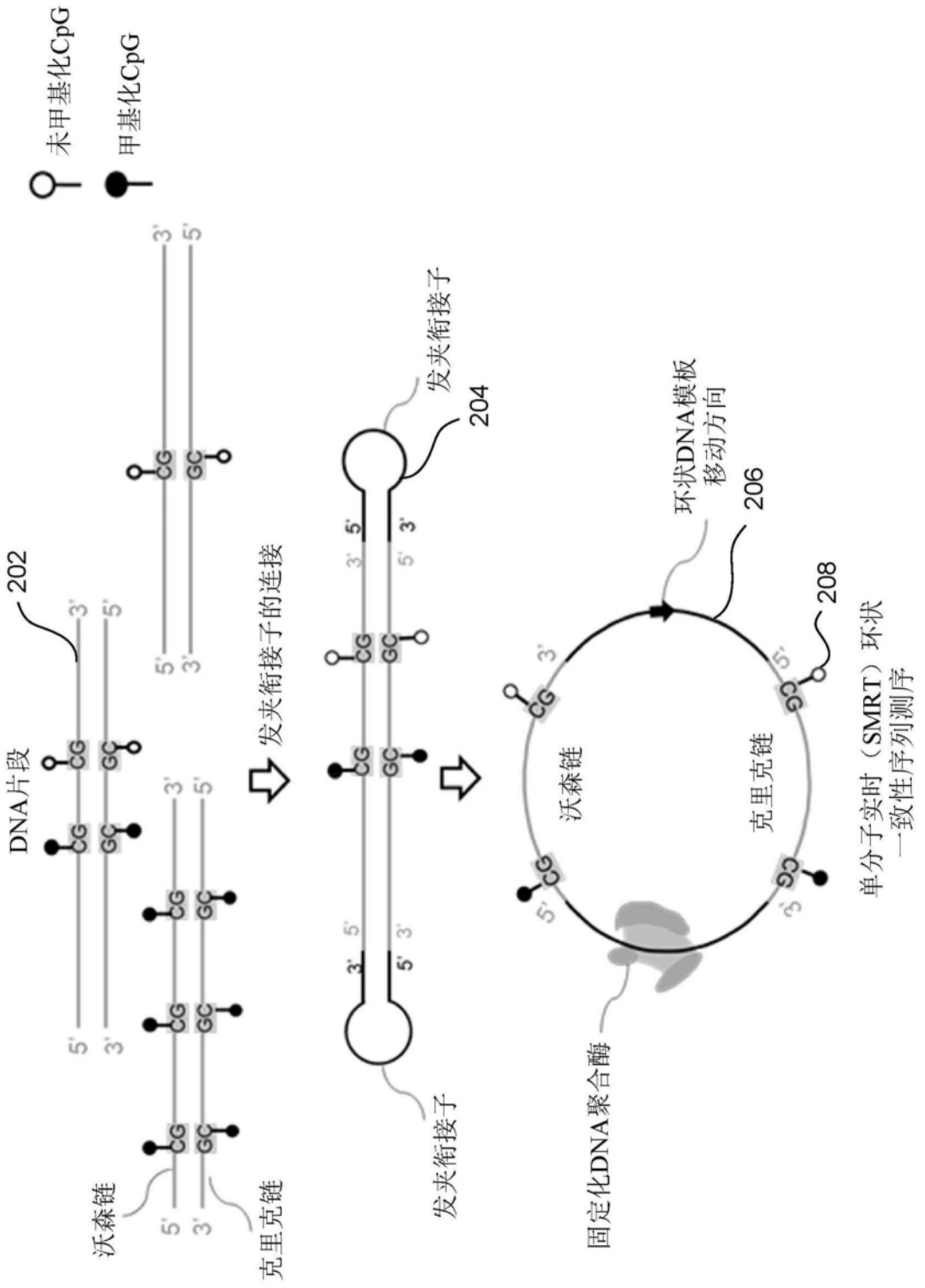

未甲基化CpG
甲基化CpG
DNA片段
202
发夹衔接子的连接
发夹衔接子
204
环状DNA模板
移动方向
206
208
沃森链
克里克链
固定化DNA聚合酶
单分子实时（SMRT）环状
一致性序列测序

图2

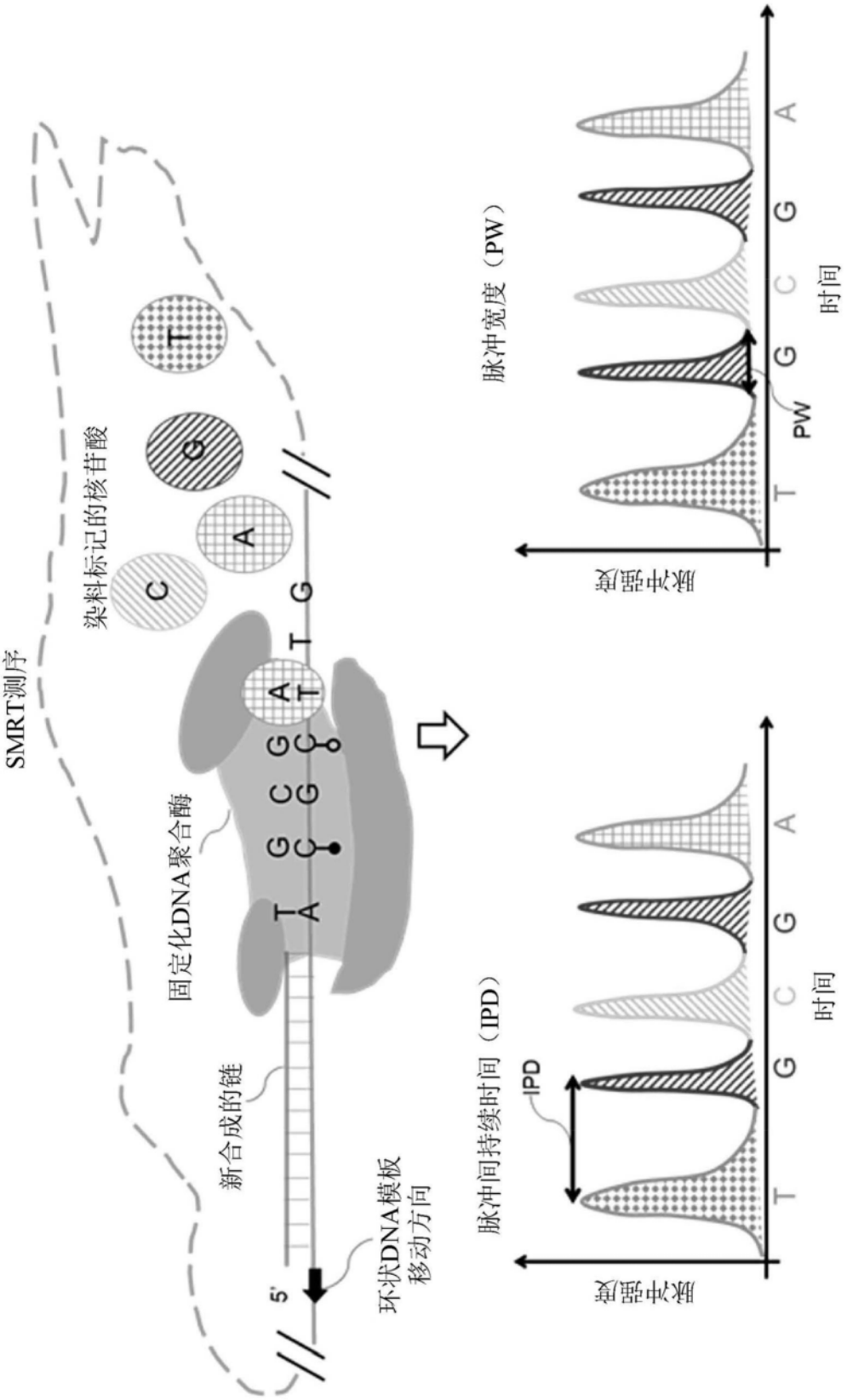

SMRT测序
染料标记的核苷酸
C
A
G
T
固定化DNA聚合酶
新合成的链
环状DNA模板
移动方向
5'
脉冲间持续时间（IPD）
IPD
脉冲强度
时间
脉冲宽度（PW）
PW

图3

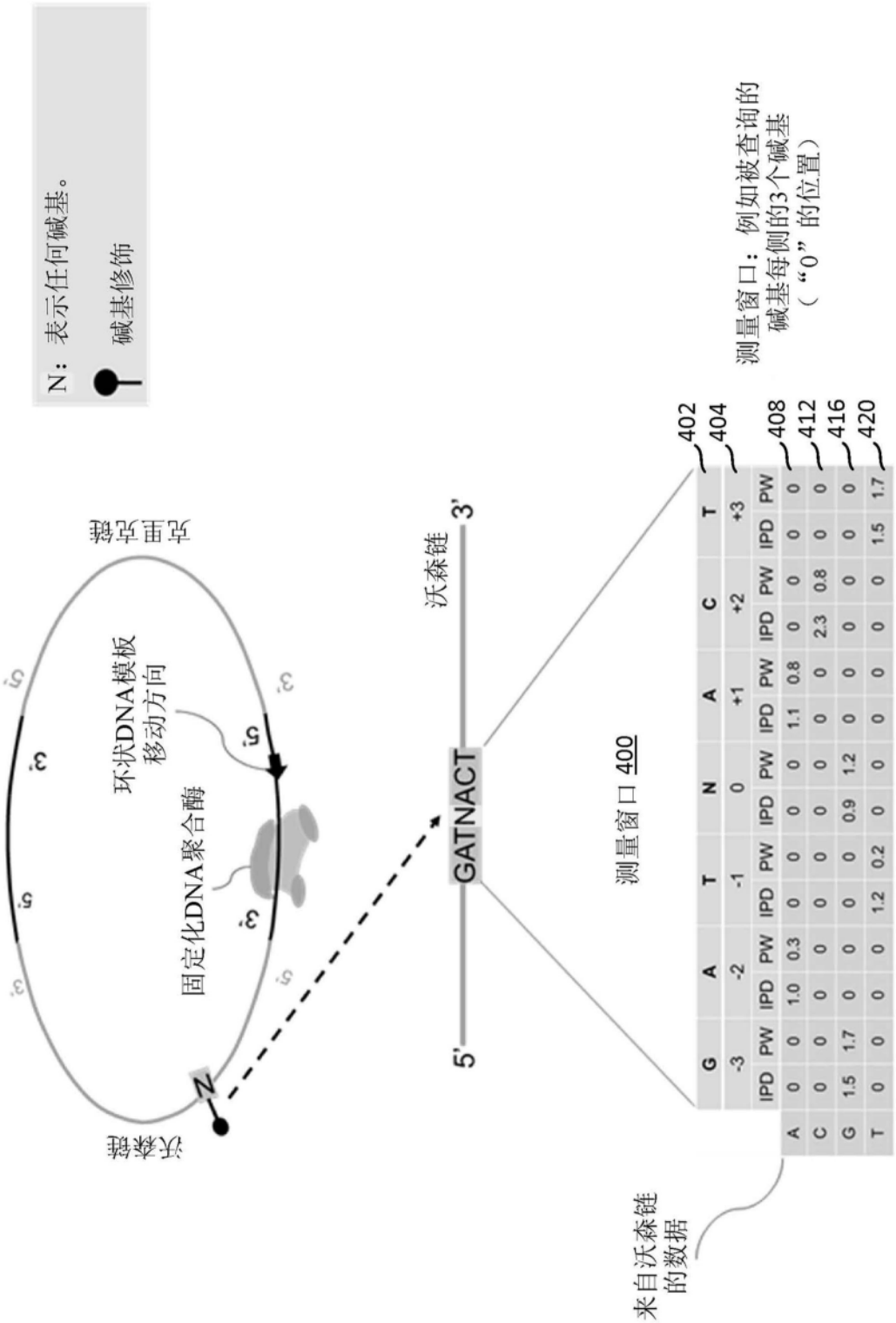
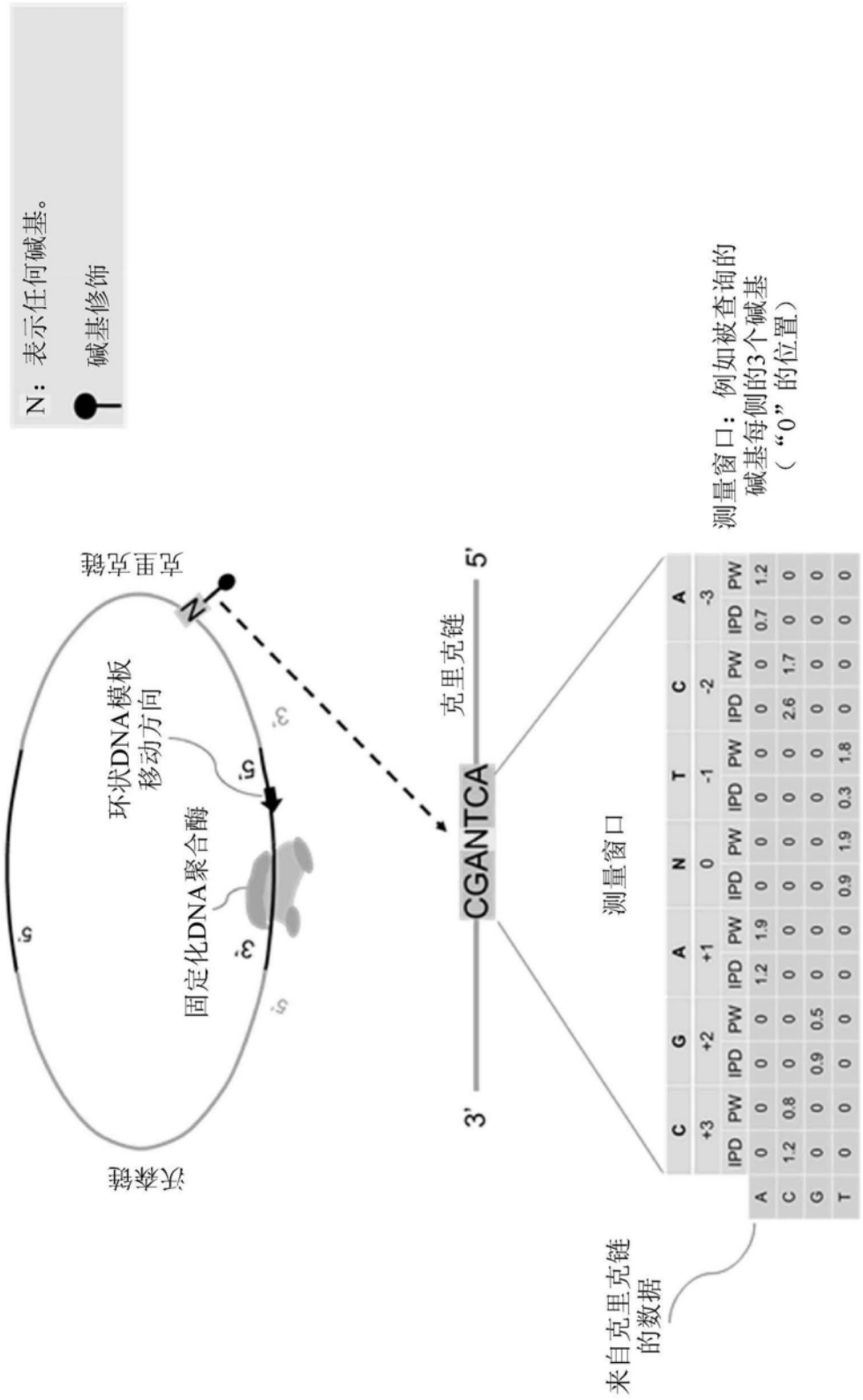

N：表示任何碱基。
碱基修饰
环状DNA模板
移动方向
固定化DNA聚合酶
克里克链
沃森链
沃森链
5'
3'
GATNACT
测量窗口 400
402
404
408
412
416
420
测量窗口：例如被查询的
碱基每侧的3个碱基
（“0”的位置）
来自沃森链
的数据

图4



	C		G		A		N		T		C		A	
	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0.7	1.2
C	1.2	0.8	0	0	0	0	0	0	0	0	2.6	1.7	0	0
G	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0.9	1.9	0.3	1.8	0	0	0	0

图5

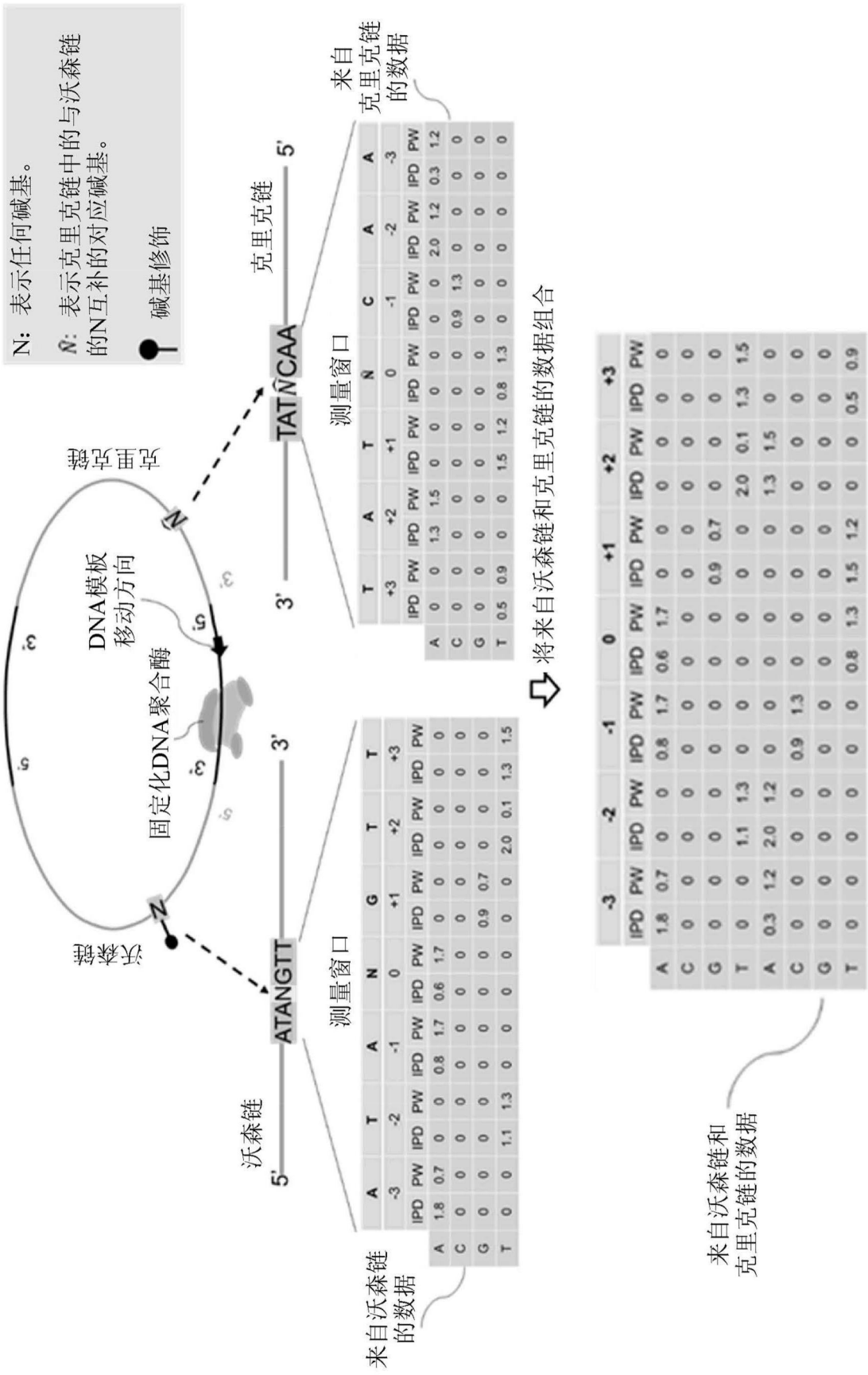

N：表示任何碱基。
Ñ：表示克里克链中的与沃森链的N互补的对应碱基。
碱基修饰
沃森链
克里克链
DNA模板移动方向
固定化DNA聚合酶
5'
3'
ATANGTT
TATÑCAA
测量窗口
来自沃森链的数据
来自克里克链的数据
将来自沃森链和克里克链的数据组合
来自沃森链和克里克链的数据

图6

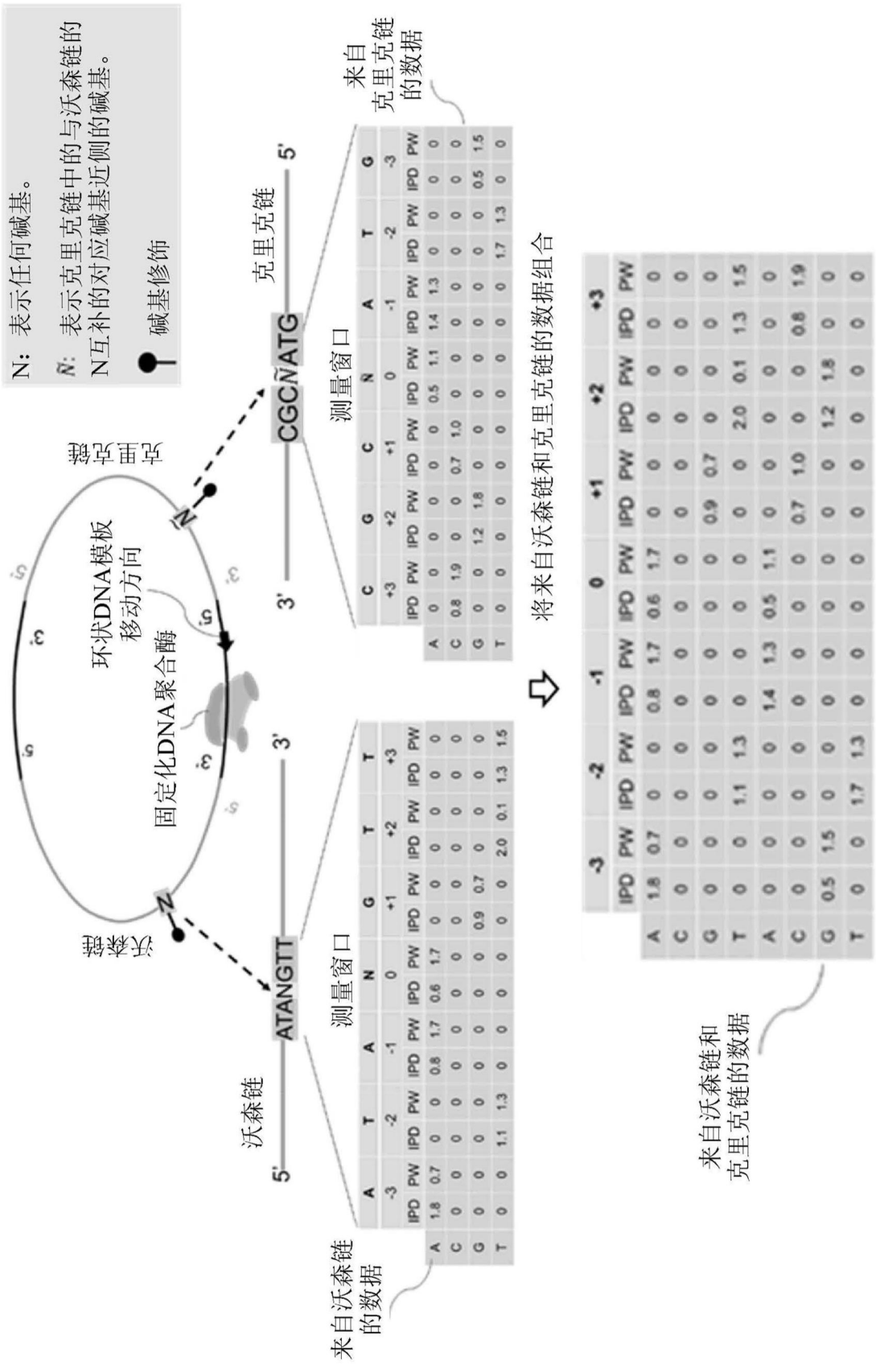

N: 表示任何碱基。
Ñ: 表示克里克链中的与沃森链的N互补的对应碱基近侧的碱基。
碱基修饰
环状DNA模板
移动方向
固定化DNA聚合酶
沃森链
克里克链
5'
3'
ATANGTT
CGCÑATG
测量窗口
来自沃森链的数据
来自克里克链的数据
将来自沃森链和克里克链的数据组合
来自沃森链和克里克链的数据

图7

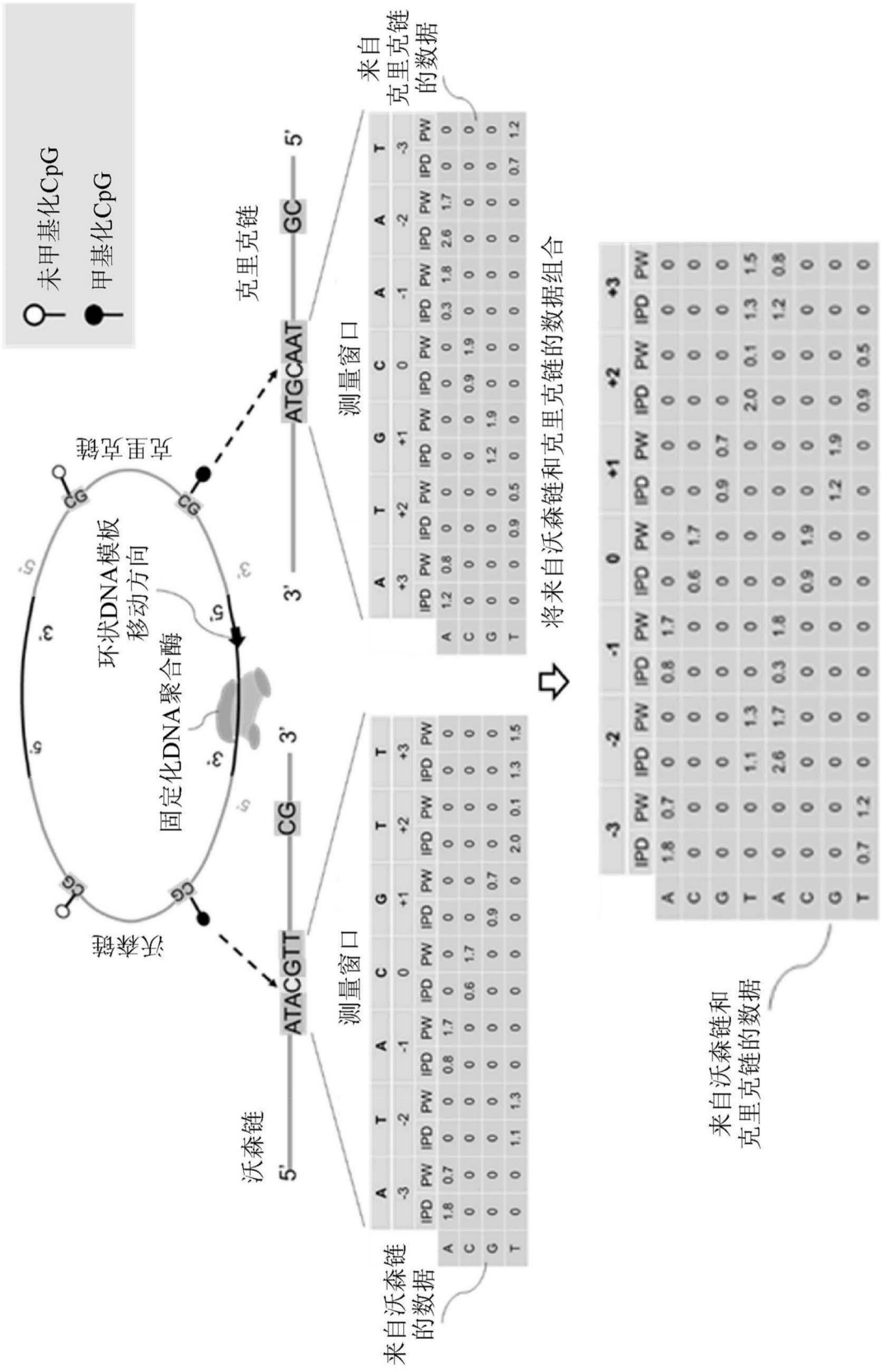

未甲基化CpG
甲基化CpG
沃森链
克里克链
环状DNA模板
移动方向
固定化DNA聚合酶
克里克链
沃森链
5'
ATACGTT
CG
3'
3'
ATGCAAT
GC
5'
测量窗口
测量窗口
来自沃森链的数据
来自克里克链的数据
将来自沃森链和克里克链的数据组合
来自沃森链和克里克链的数据

图8

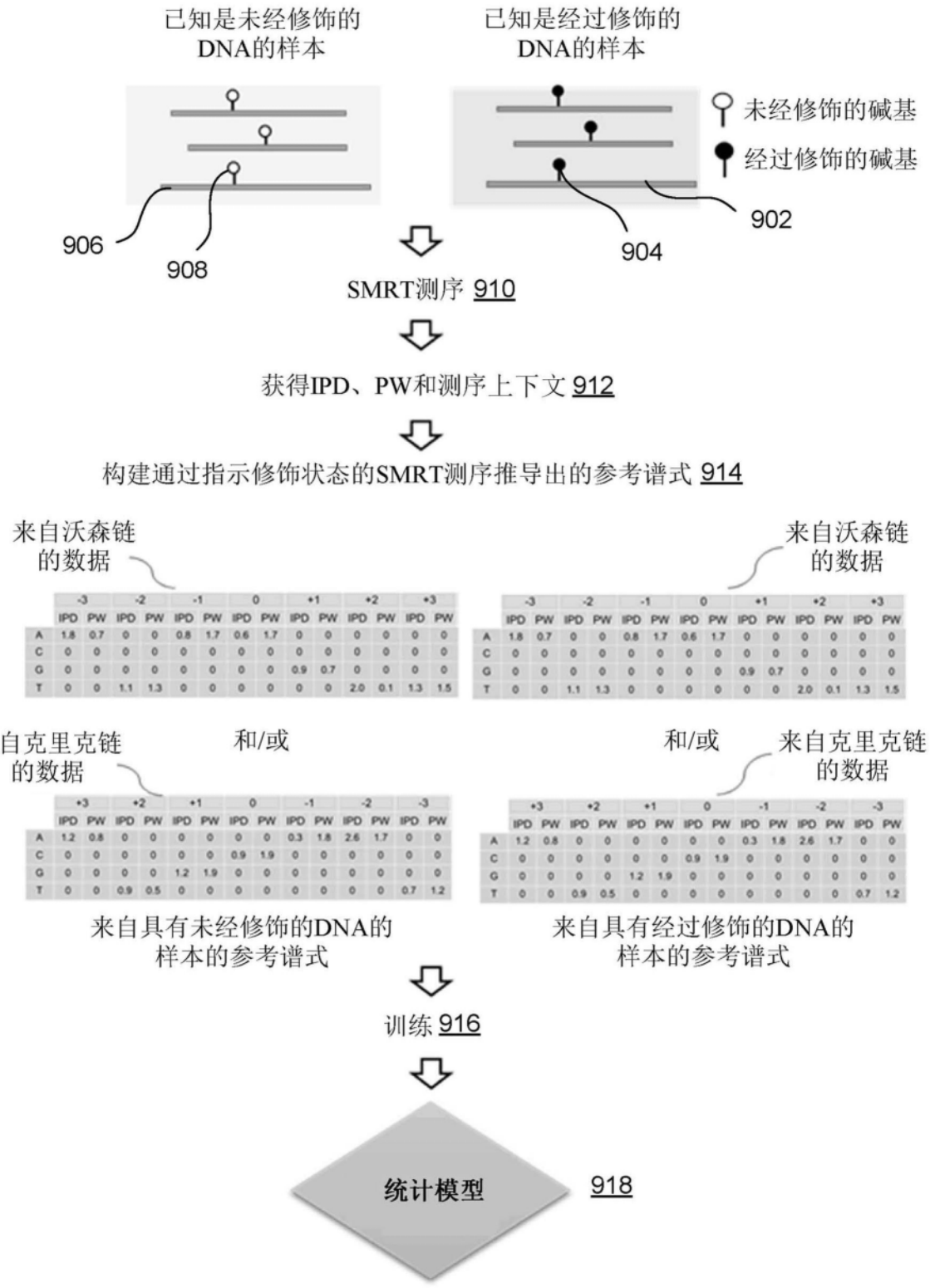
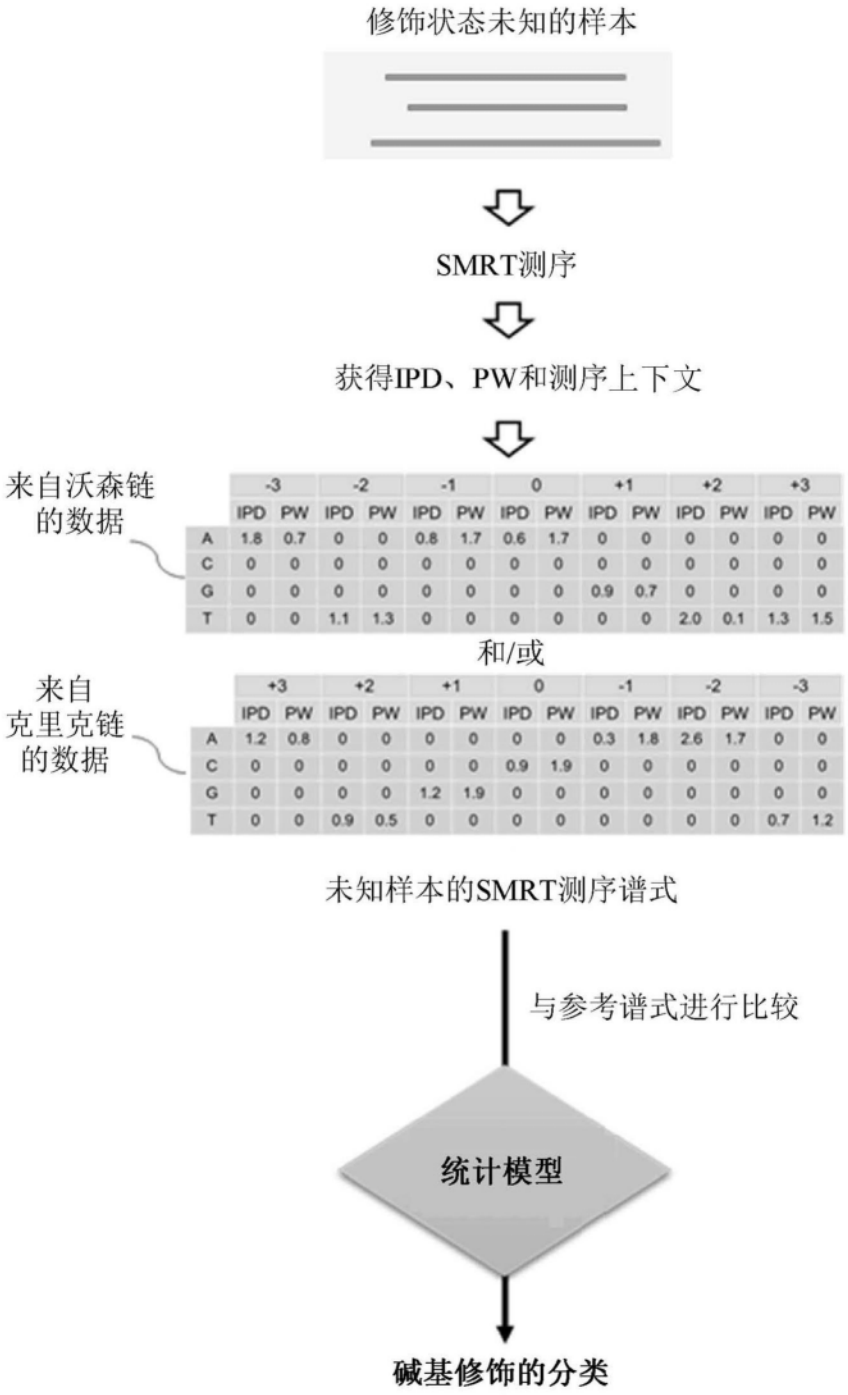

已知是未经修饰的DNA的样本
已知是经过修饰的DNA的样本
未经修饰的碱基
经过修饰的碱基
906
908
902
904
SMRT测序 910
获得IPD、PW和测序上下文 912
构建通过指示修饰状态的SMRT测序推导出的参考谱式 914
来自沃森链的数据
来自沃森链的数据
来自克里克链的数据
和/或
和/或
来自克里克链的数据
来自具有未经修饰的DNA的样本的参考谱式
来自具有经过修饰的DNA的样本的参考谱式
训练 916
统计模型
918

图9



	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5

	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.2	0.8	0	0	0	0	0	0	0.3	1.8	2.6	1.7	0	0
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0	0
T	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0.7	1.2

图10

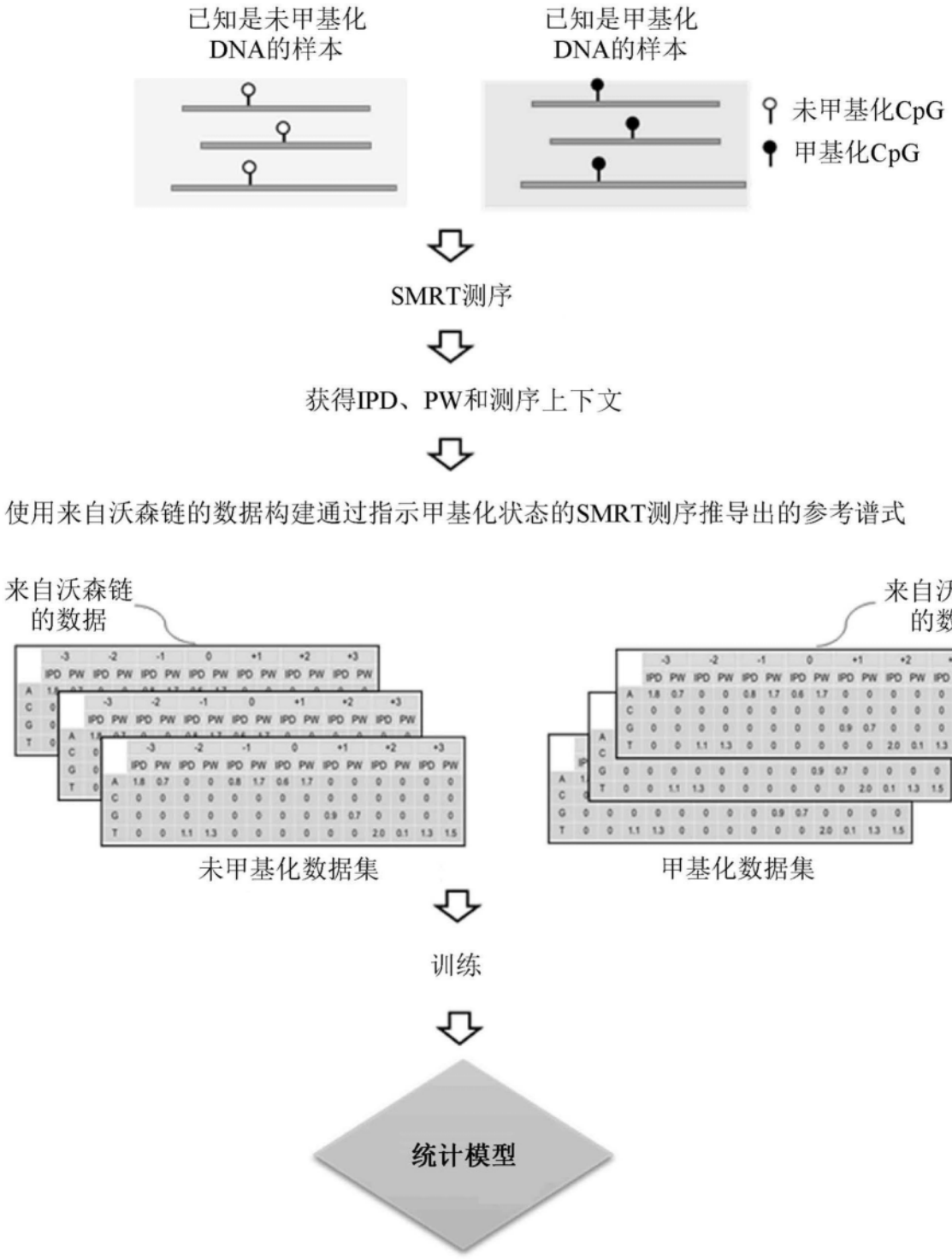
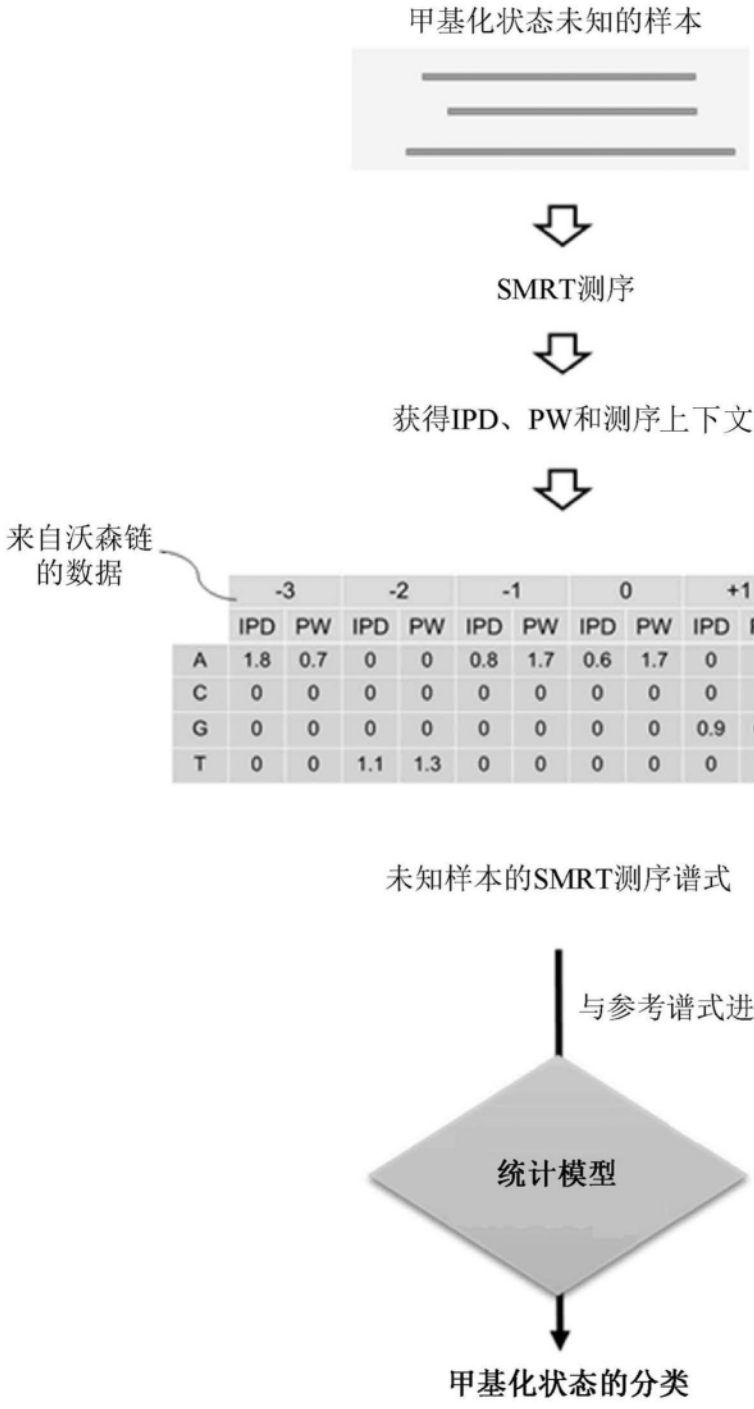

已知是未甲基化DNA的样本
已知是甲基化DNA的样本
未甲基化CpG
甲基化CpG
SMRT测序
获得IPD、PW和测序上下文
使用来自沃森链的数据构建通过指示甲基化状态的SMRT测序推导出的参考谱式
来自沃森链的数据
来自沃森链的数据
未甲基化数据集
甲基化数据集
训练
统计模型

图11



	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5

图12

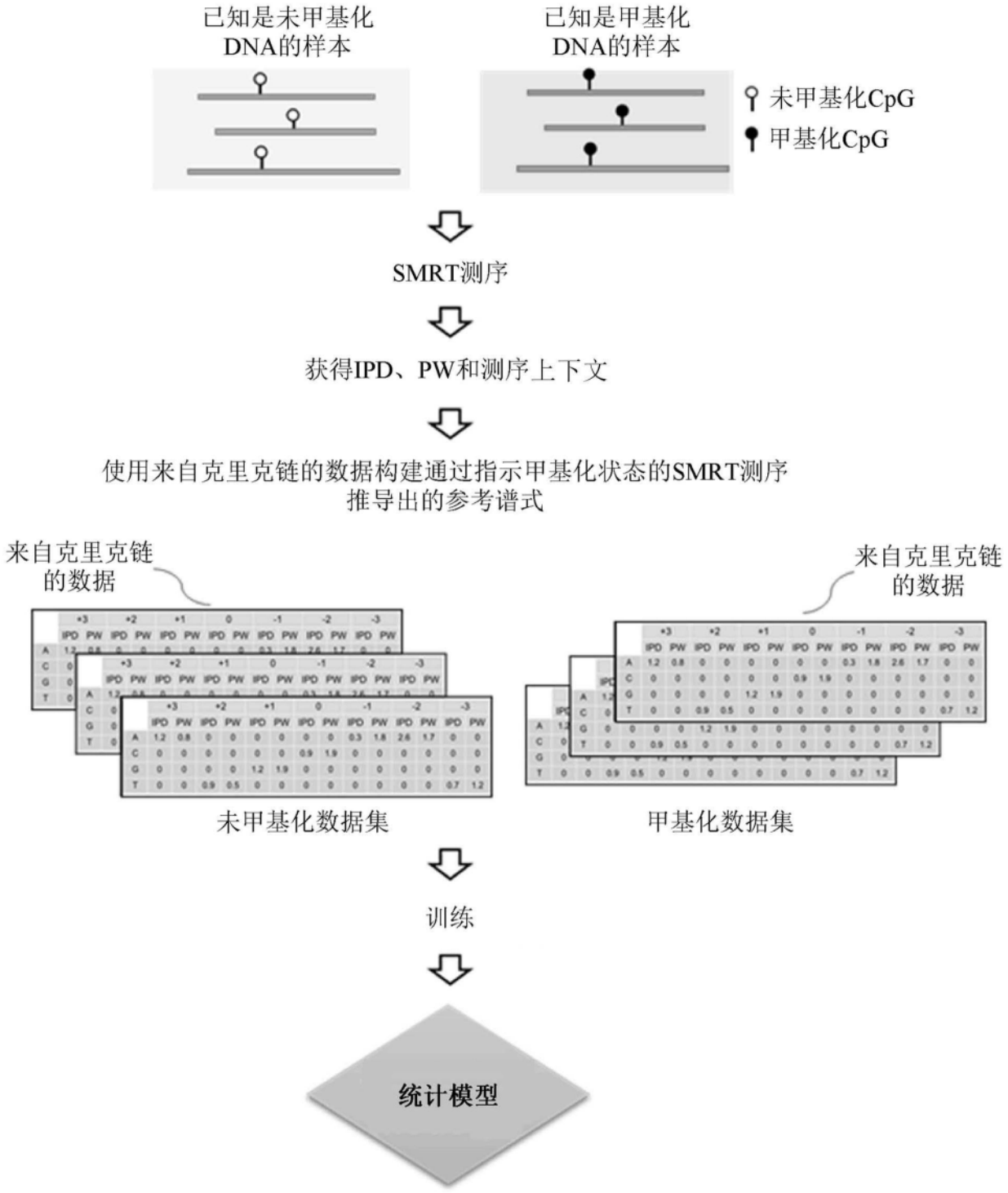
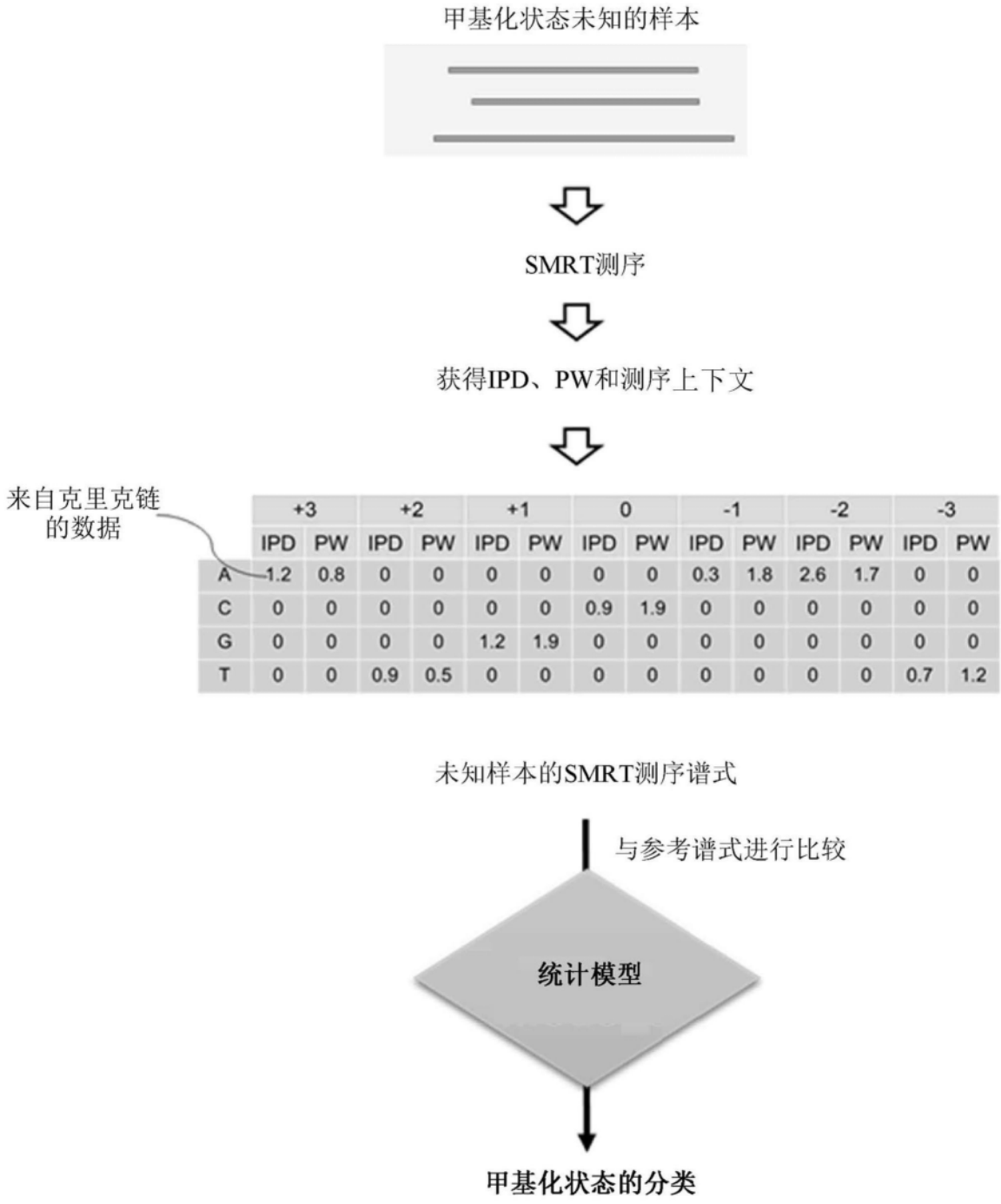

已知是未甲基化DNA的样本
已知是甲基化DNA的样本
未甲基化CpG
甲基化CpG
SMRT测序
获得IPD、PW和测序上下文
使用来自克里克链的数据构建通过指示甲基化状态的SMRT测序推导出的参考谱式
来自克里克链的数据
来自克里克链的数据
未甲基化数据集
甲基化数据集
训练
统计模型

图13



	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.2	0.8	0	0	0	0	0	0	0.3	1.8	2.6	1.7	0	0
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0	0
T	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0.7	1.2

图14

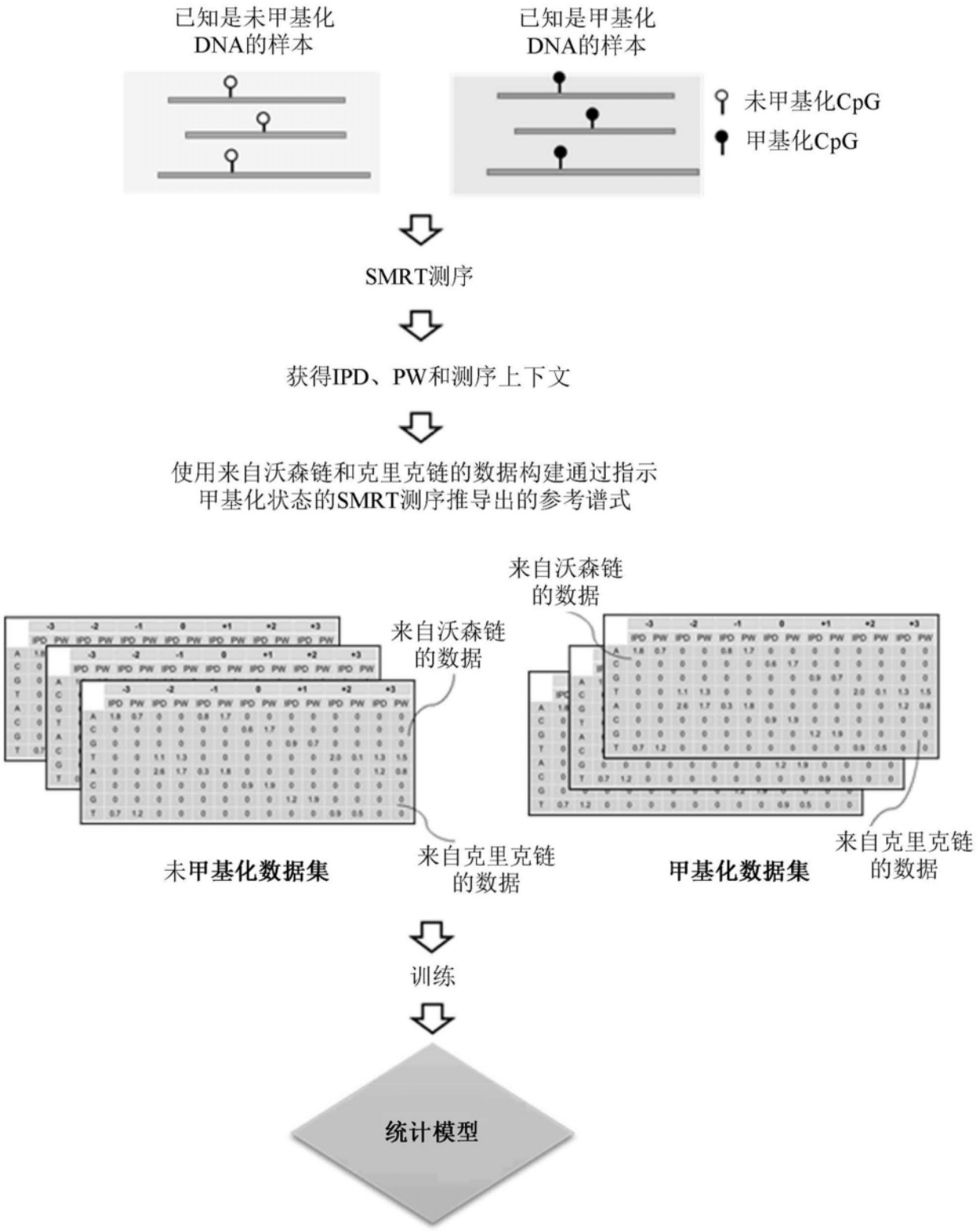
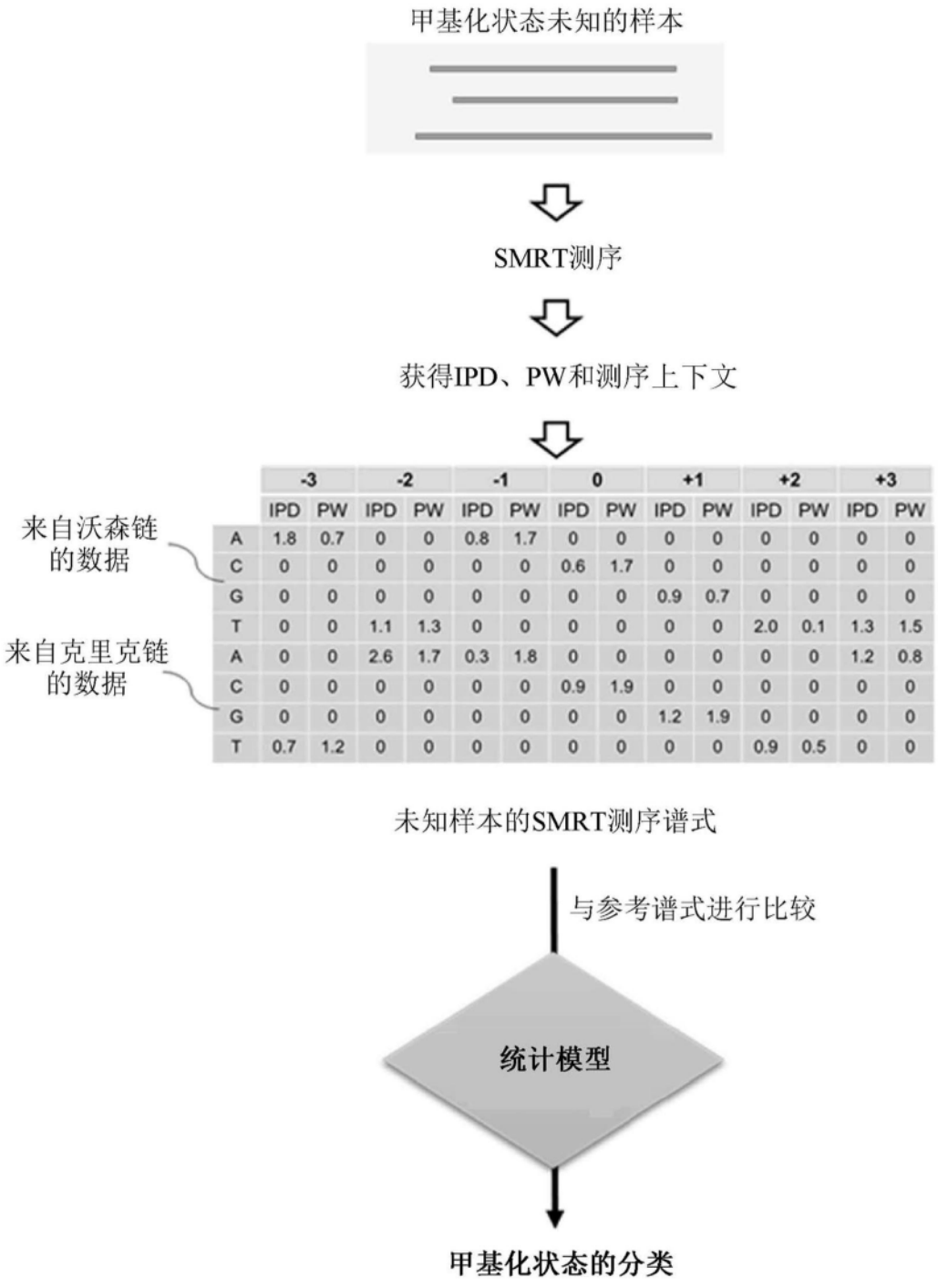

已知是未甲基化DNA的样本
已知是甲基化DNA的样本
未甲基化CpG
甲基化CpG
SMRT测序
获得IPD、PW和测序上下文
使用来自沃森链和克里克链的数据构建通过指示甲基化状态的SMRT测序推导出的参考谱式
来自沃森链的数据
来自沃森链的数据
来自克里克链的数据
来自克里克链的数据
未甲基化数据集
甲基化数据集
训练
统计模型

图15



	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0.6	1.7	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5
A	0	0	2.6	1.7	0.3	1.8	0	0	0	0	0	0	1.2	0.8
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	1.2	1.9	0	0	0	0
T	0.7	1.2	0	0	0	0	0	0	0	0	0.9	0.5	0	0

图16

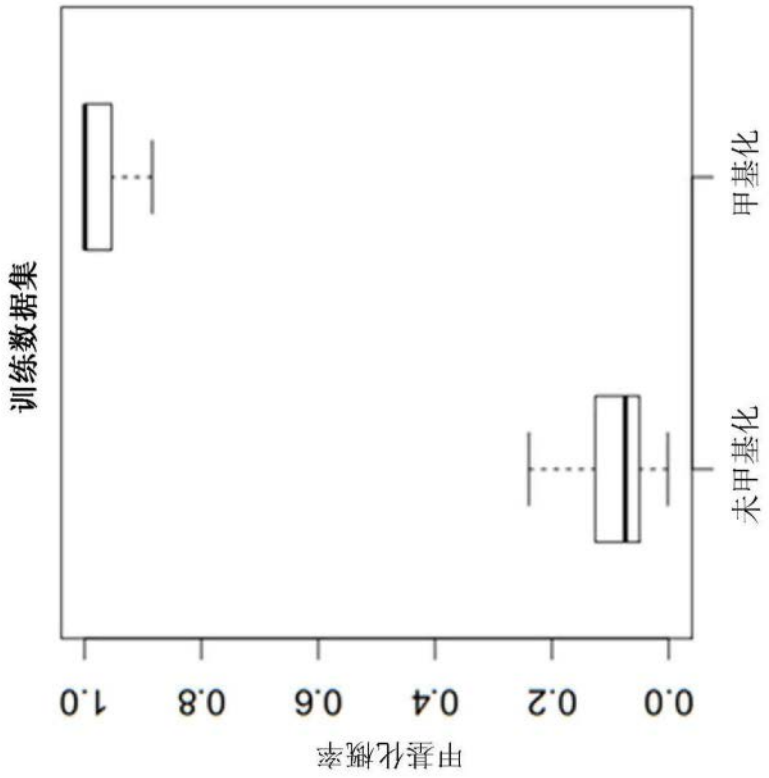

训练数据集
甲基化
未甲基化
0.0
0.2
0.4
0.6
0.8
1.0
甲基化概率

图17A

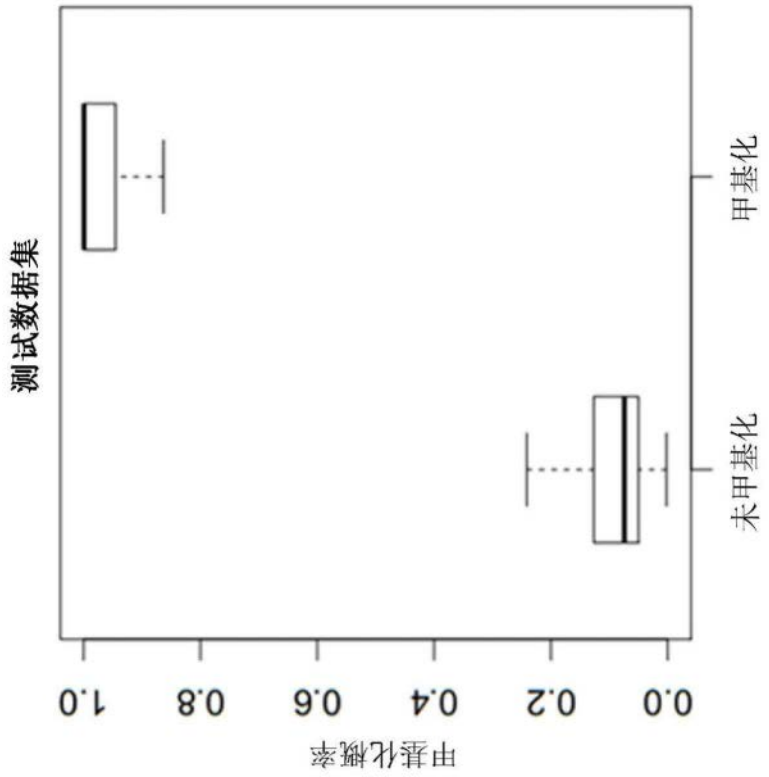

测试数据集
甲基化
未甲基化
0.0
0.2
0.4
0.6
0.8
1.0
甲基化概率

图17B

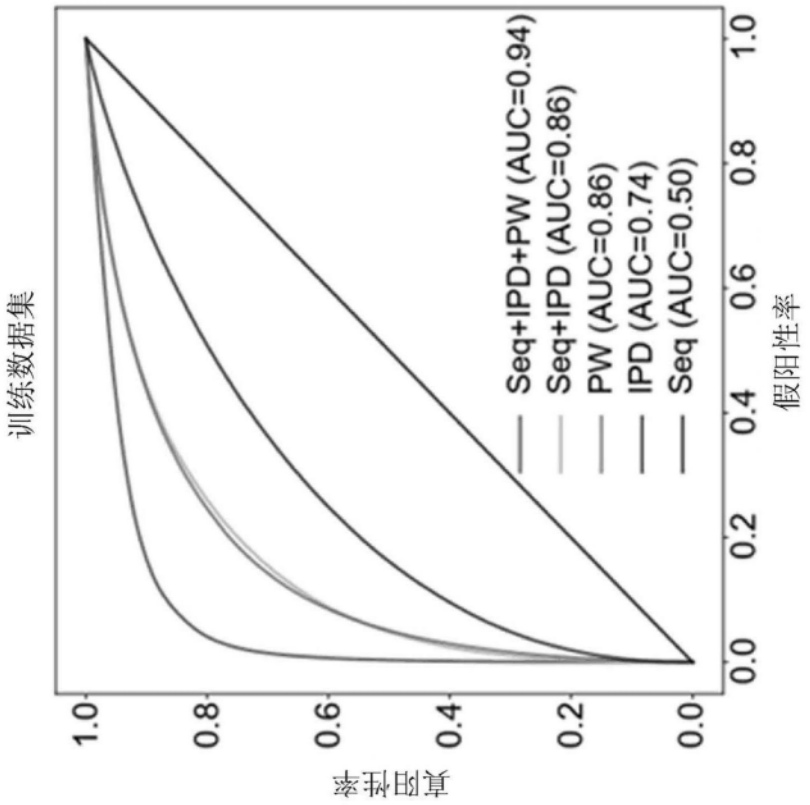

训练数据集
真阳性率
假阳性率
Seq+IPD+PW (AUC=0.94)
Seq+IPD (AUC=0.86)
PW (AUC=0.86)
IPD (AUC=0.74)
Seq (AUC=0.50)
0.0
0.2
0.4
0.6
0.8
1.0

图18A

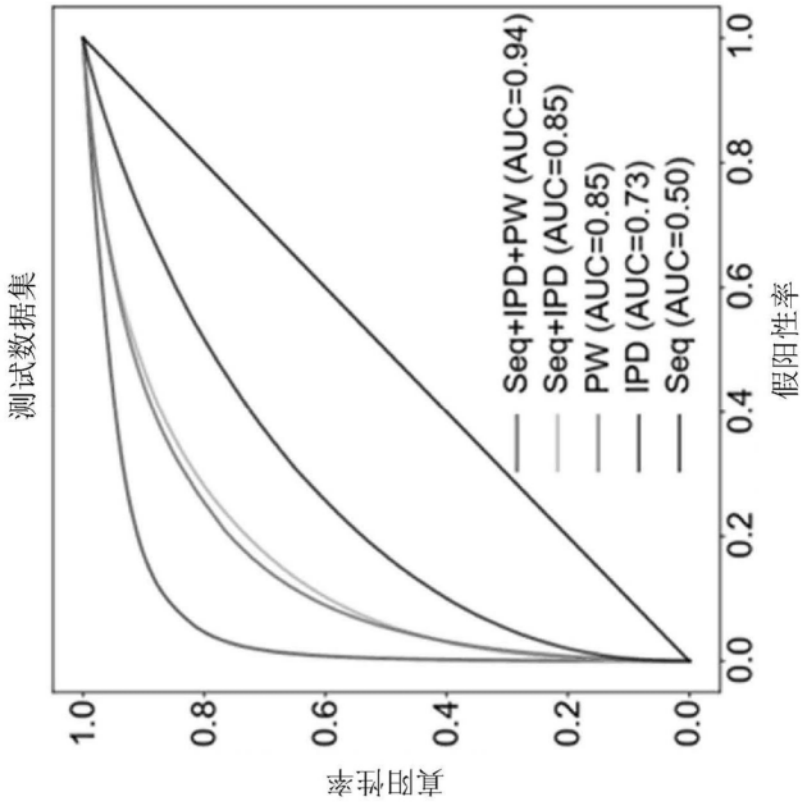

测试数据集
真阳性率
假阳性率
Seq+IPD+PW (AUC=0.94)
Seq+IPD (AUC=0.85)
PW (AUC=0.85)
IPD (AUC=0.73)
Seq (AUC=0.50)
0.0
0.2
0.4
0.6
0.8
1.0

图18B

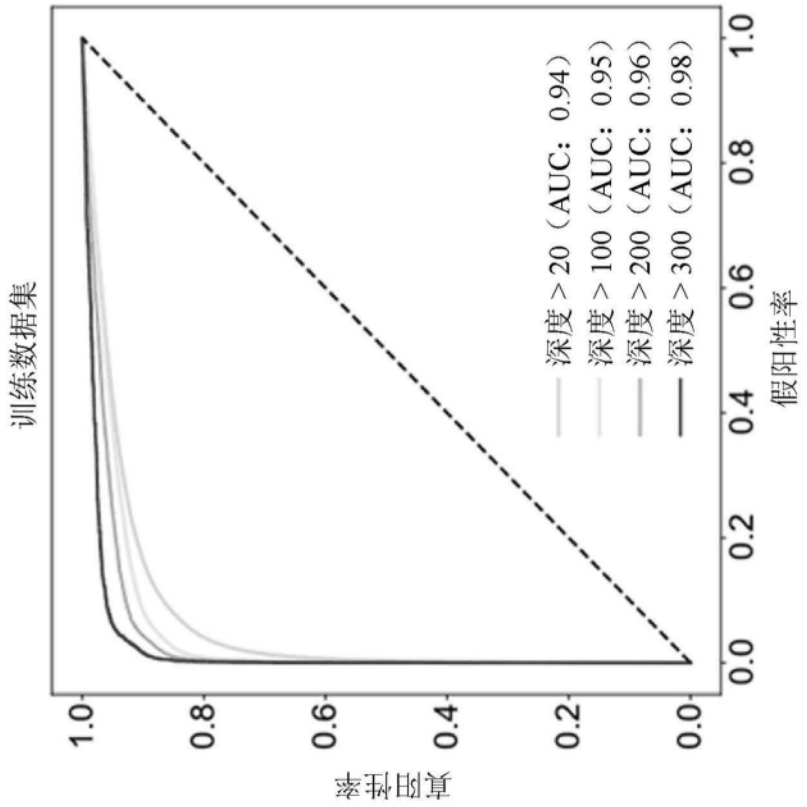

训练数据集
真阳性率
假阳性率
深度＞20（AUC：0.94）
深度＞100（AUC：0.95）
深度＞200（AUC：0.96）
深度＞300（AUC：0.98）
0.0
0.2
0.4
0.6
0.8
1.0

图19A

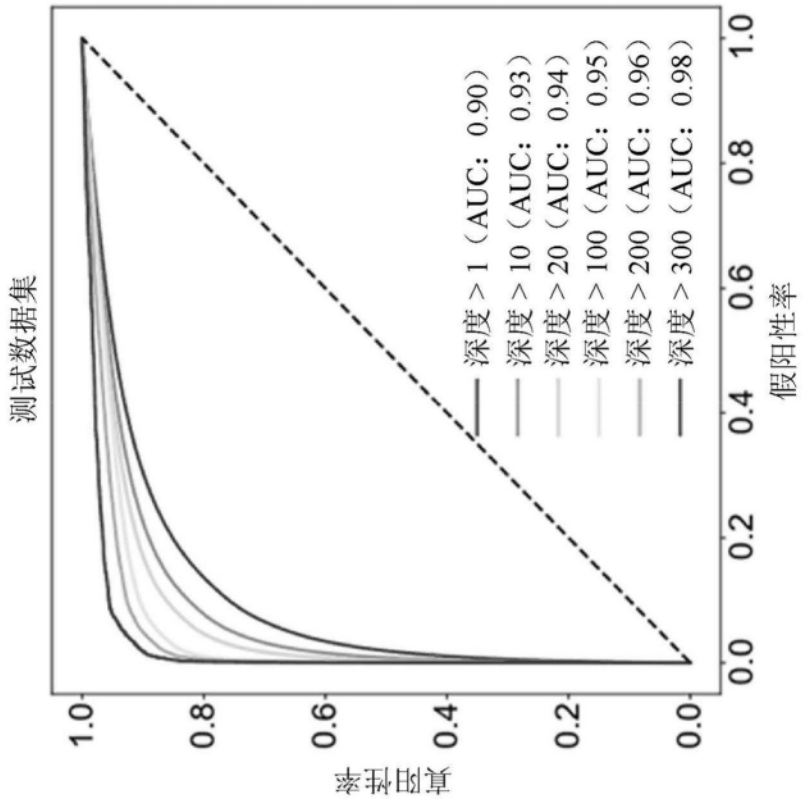

测试数据集
真阳性率
假阳性率
深度 > 1（AUC：0.90）
深度 > 10（AUC：0.93）
深度 > 20（AUC：0.94）
深度 > 100（AUC：0.95）
深度 > 200（AUC：0.96）
深度 > 300（AUC：0.98）
0.0
0.2
0.4
0.6
0.8
1.0

图19B

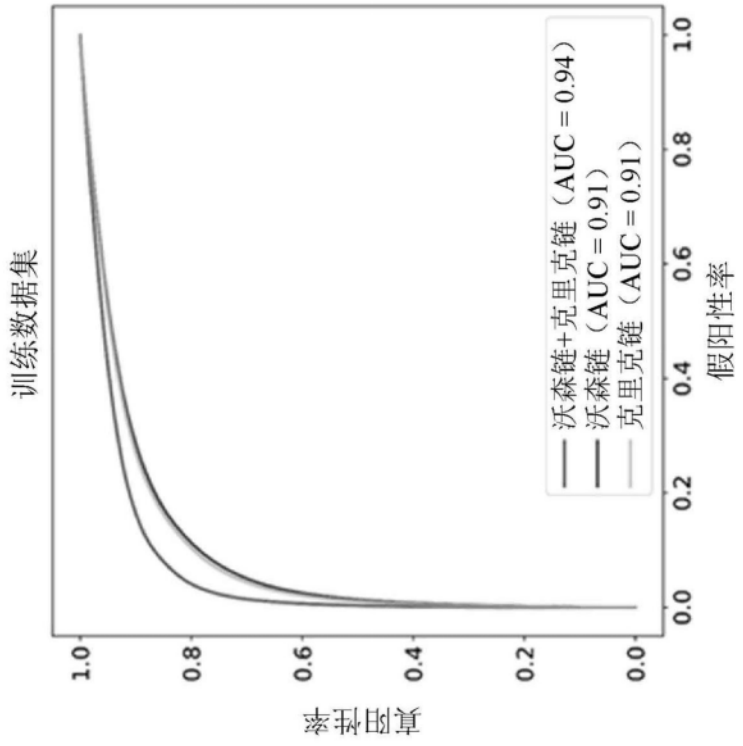

训练数据集
真阳性率
假阳性率
沃森链+克里克链（AUC = 0.94）
沃森链（AUC = 0.91）
克里克链（AUC = 0.91）
0.0
0.2
0.4
0.6
0.8
1.0

图20A

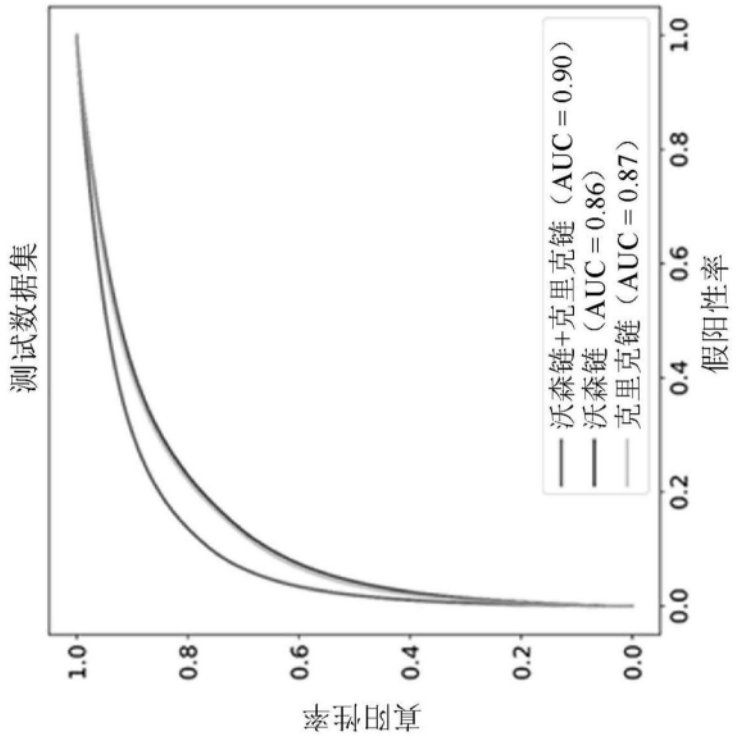

测试数据集
真阳性率
假阳性率
沃森链+克里克链（AUC = 0.90）
沃森链（AUC = 0.86）
克里克链（AUC = 0.87）
0.0
0.2
0.4
0.6
0.8
1.0

图20B

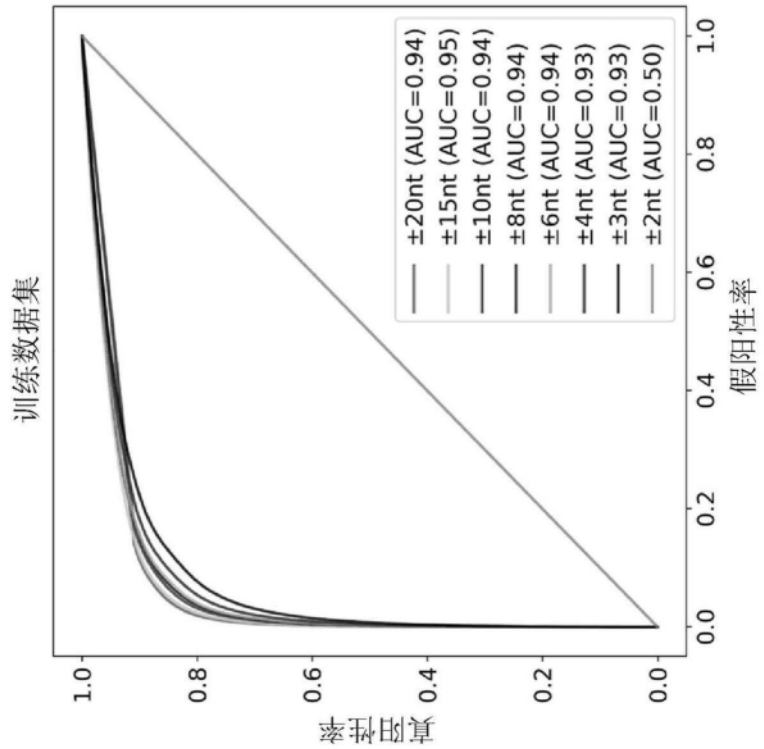

训练数据集
真阳性率
假阳性率
±20nt (AUC=0.94)
±15nt (AUC=0.95)
±10nt (AUC=0.94)
±8nt (AUC=0.94)
±6nt (AUC=0.94)
±4nt (AUC=0.93)
±3nt (AUC=0.93)
±2nt (AUC=0.50)
0.0
0.2
0.4
0.6
0.8
1.0

图21A

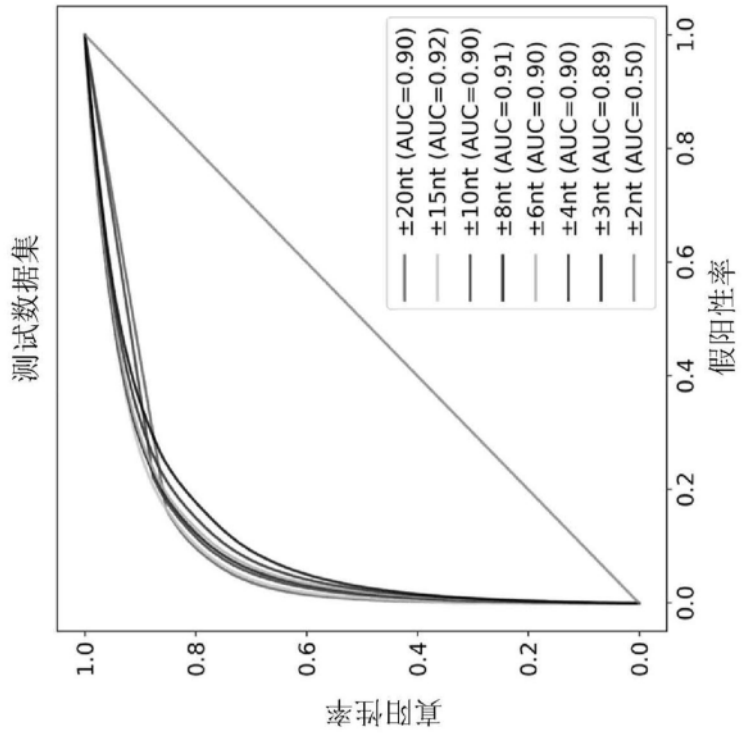

测试数据集
真阳性率
假阳性率
±20nt (AUC=0.90)
±15nt (AUC=0.92)
±10nt (AUC=0.90)
±8nt (AUC=0.91)
±6nt (AUC=0.90)
±4nt (AUC=0.90)
±3nt (AUC=0.89)
±2nt (AUC=0.50)
0.0
0.2
0.4
0.6
0.8
1.0

图21B

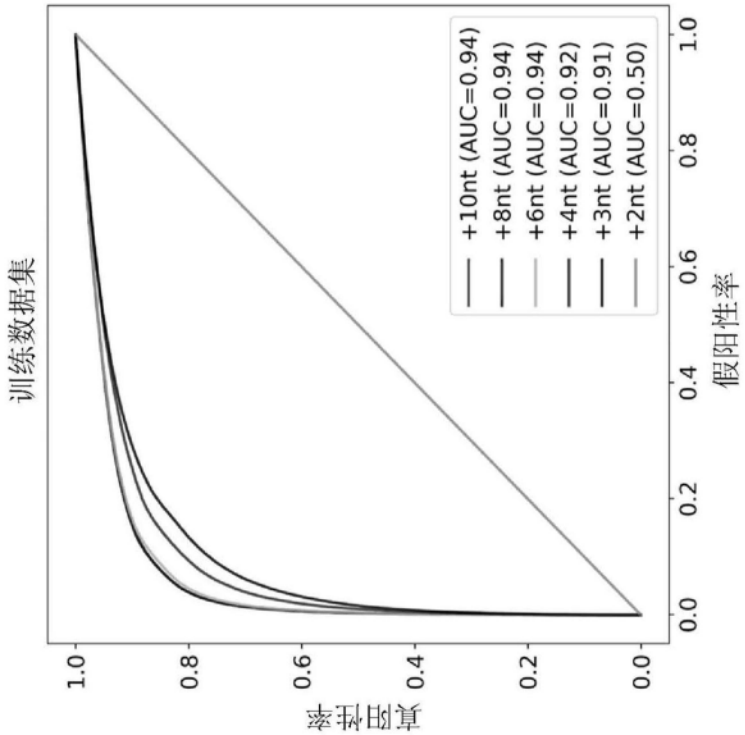

训练数据集
真阳性率
假阳性率
+10nt (AUC=0.94)
+8nt (AUC=0.94)
+6nt (AUC=0.94)
+4nt (AUC=0.92)
+3nt (AUC=0.91)
+2nt (AUC=0.50)
0.0
0.2
0.4
0.6
0.8
1.0

图22A

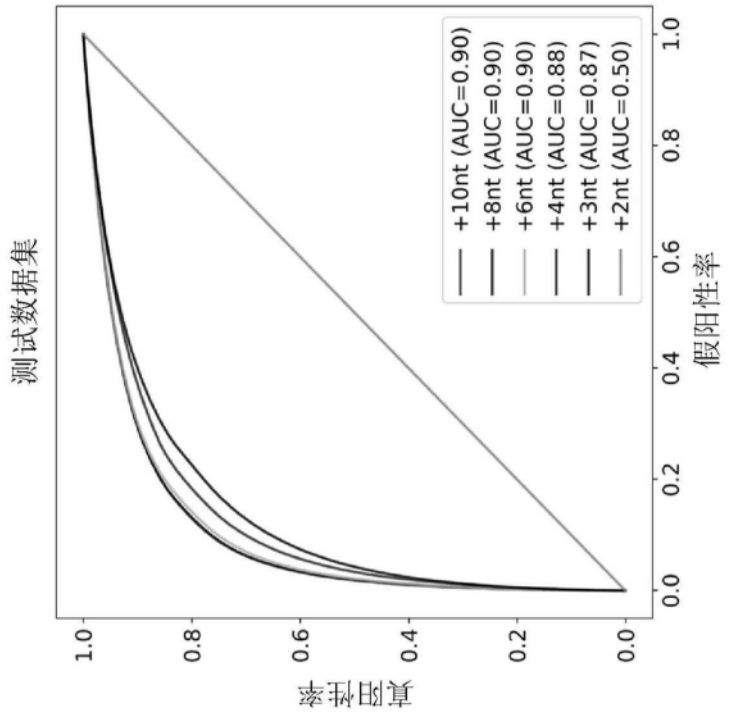

测试数据集
真阳性率
假阳性率
+10nt (AUC=0.90)
+8nt (AUC=0.90)
+6nt (AUC=0.90)
+4nt (AUC=0.88)
+3nt (AUC=0.87)
+2nt (AUC=0.50)
0.0
0.2
0.4
0.6
0.8
1.0

图22B

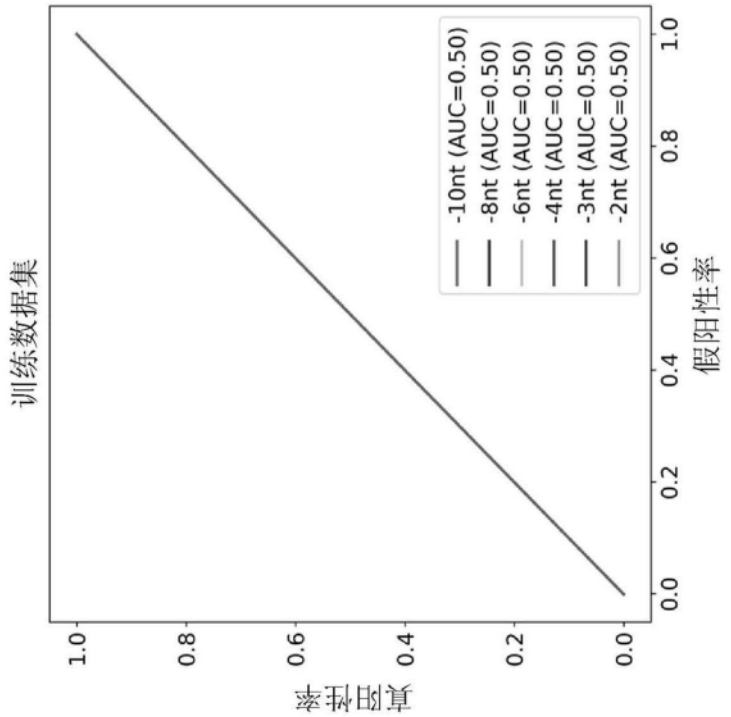

训练数据集
真阳性率
假阳性率
-10nt (AUC=0.50)
-8nt (AUC=0.50)
-6nt (AUC=0.50)
-4nt (AUC=0.50)
-3nt (AUC=0.50)
-2nt (AUC=0.50)
0.0
0.2
0.4
0.6
0.8
1.0

图23A

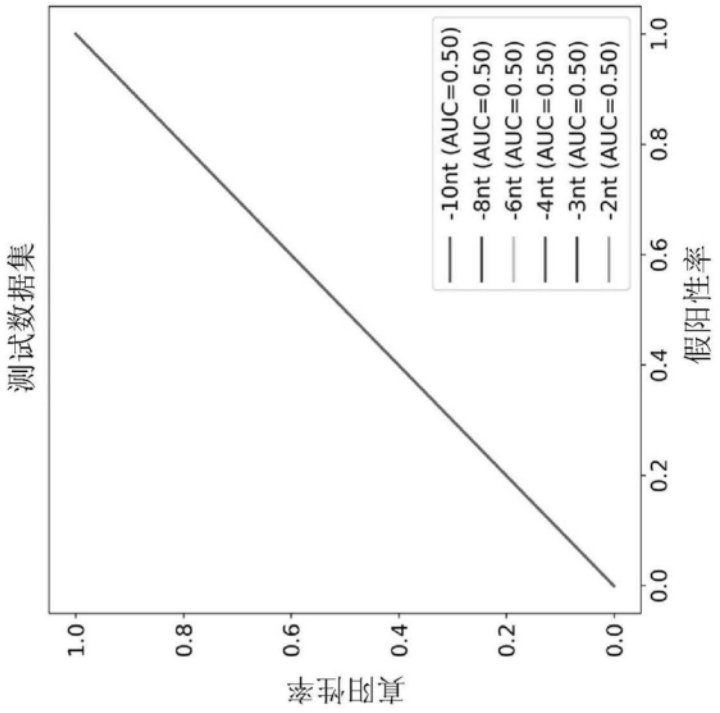

测试数据集
真阳性率
假阳性率
-10nt (AUC=0.50)
-8nt (AUC=0.50)
-6nt (AUC=0.50)
-4nt (AUC=0.50)
-3nt (AUC=0.50)
-2nt (AUC=0.50)
0.0
0.2
0.4
0.6
0.8
1.0

图23B

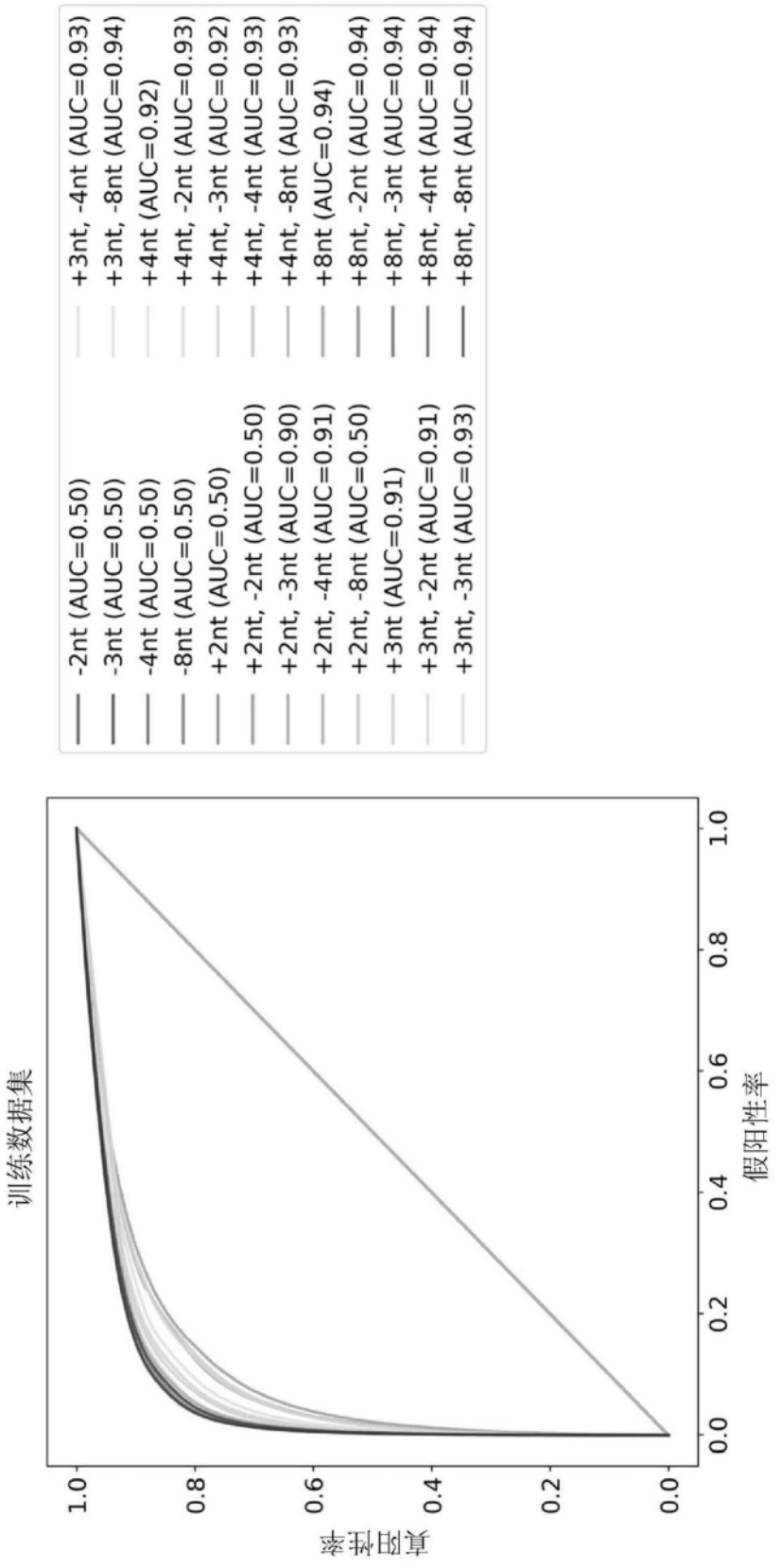

-2nt (AUC=0.50)
-3nt (AUC=0.50)
-4nt (AUC=0.50)
-8nt (AUC=0.50)
+2nt (AUC=0.50)
+2nt, -2nt (AUC=0.50)
+2nt, -3nt (AUC=0.90)
+2nt, -4nt (AUC=0.91)
+2nt, -8nt (AUC=0.50)
+3nt (AUC=0.91)
+3nt, -2nt (AUC=0.91)
+3nt, -3nt (AUC=0.93)
+3nt, -4nt (AUC=0.93)
+3nt, -8nt (AUC=0.94)
+4nt (AUC=0.92)
+4nt, -2nt (AUC=0.93)
+4nt, -3nt (AUC=0.92)
+4nt, -4nt (AUC=0.93)
+4nt, -8nt (AUC=0.93)
+8nt (AUC=0.94)
+8nt, -2nt (AUC=0.94)
+8nt, -3nt (AUC=0.94)
+8nt, -4nt (AUC=0.94)
+8nt, -8nt (AUC=0.94)
训练数据集
真阳性率
假阳性率
0.0
0.2
0.4
0.6
0.8
1.0

图24

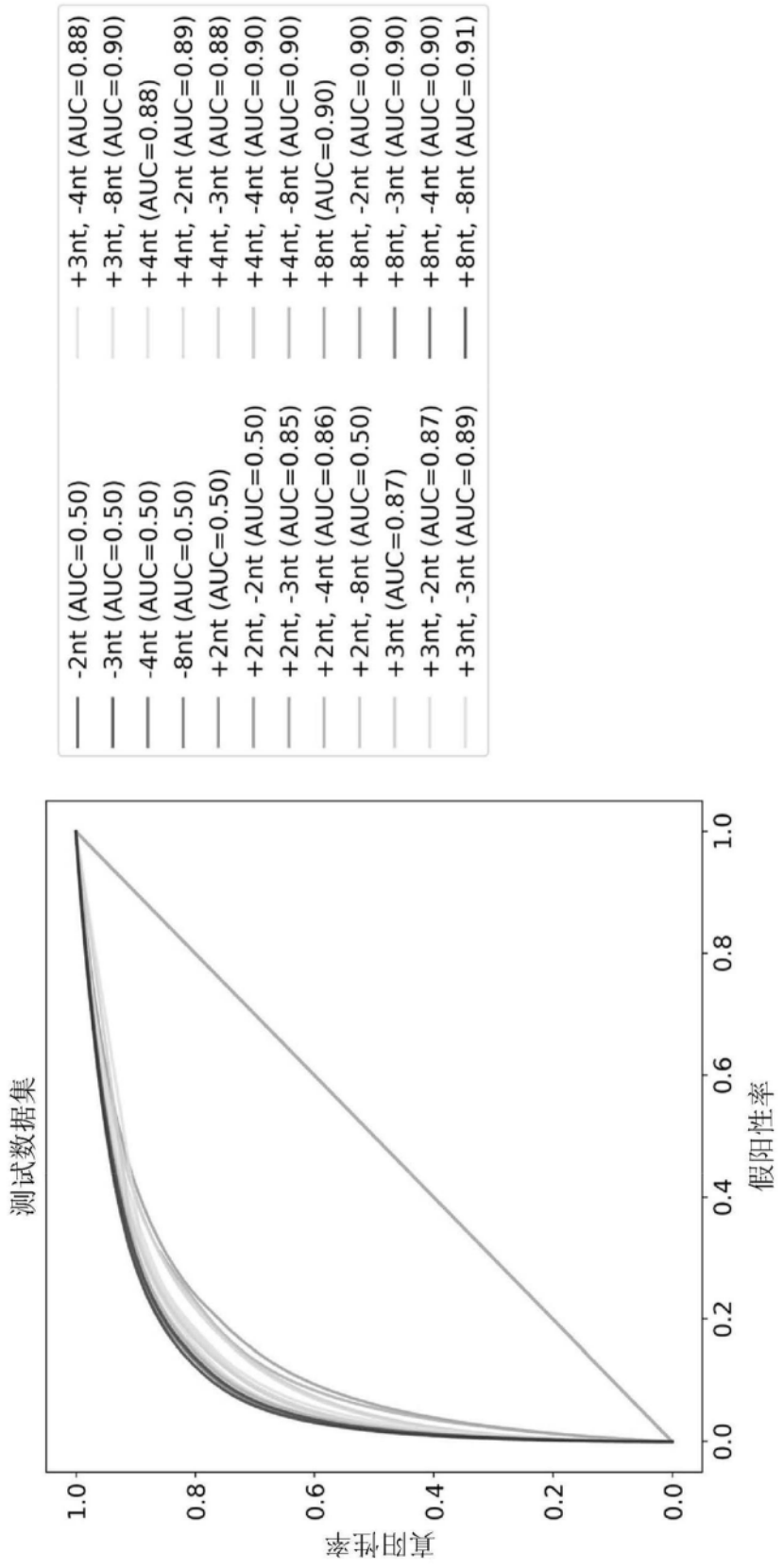

-2nt (AUC=0.50)
-3nt (AUC=0.50)
-4nt (AUC=0.50)
-8nt (AUC=0.50)
+2nt (AUC=0.50)
+2nt, -2nt (AUC=0.50)
+2nt, -3nt (AUC=0.85)
+2nt, -4nt (AUC=0.86)
+2nt, -8nt (AUC=0.50)
+3nt (AUC=0.87)
+3nt, -2nt (AUC=0.87)
+3nt, -3nt (AUC=0.89)
+3nt, -4nt (AUC=0.88)
+3nt, -8nt (AUC=0.90)
+4nt (AUC=0.88)
+4nt, -2nt (AUC=0.89)
+4nt, -3nt (AUC=0.88)
+4nt, -4nt (AUC=0.90)
+4nt, -8nt (AUC=0.90)
+8nt (AUC=0.90)
+8nt, -2nt (AUC=0.90)
+8nt, -3nt (AUC=0.90)
+8nt, -4nt (AUC=0.90)
+8nt, -8nt (AUC=0.91)
测试数据集
真阳性率
假阳性率
0.0
0.2
0.4
0.6
0.8
1.0

图25

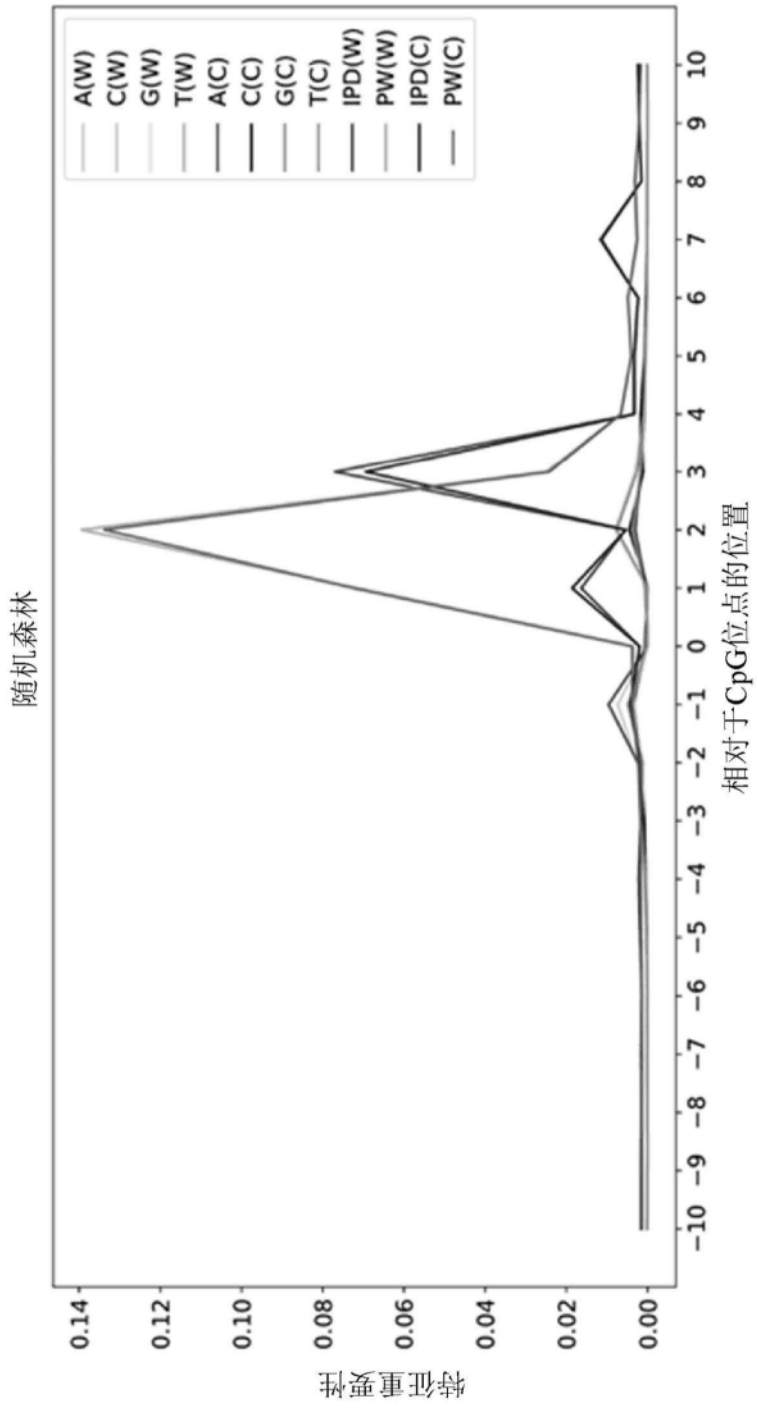

随机森林
特征重要性
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
相对于CpG位点的位置
−10 −9 −8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7 8 9 10
A(W)
C(W)
G(W)
T(W)
A(C)
C(C)
G(C)
T(C)
IPD(W)
PW(W)
IPD(C)
PW(C)

图26

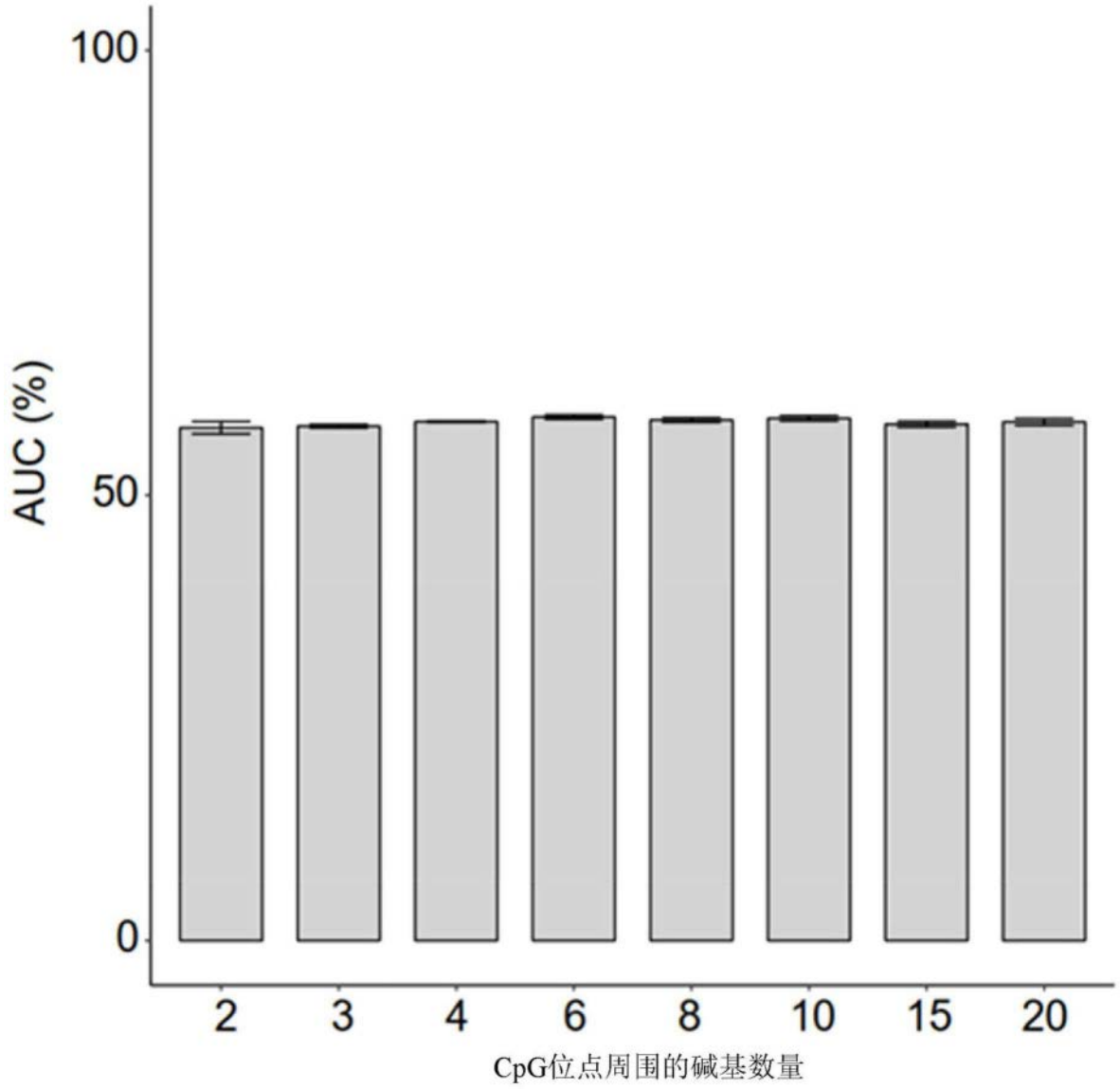


图27

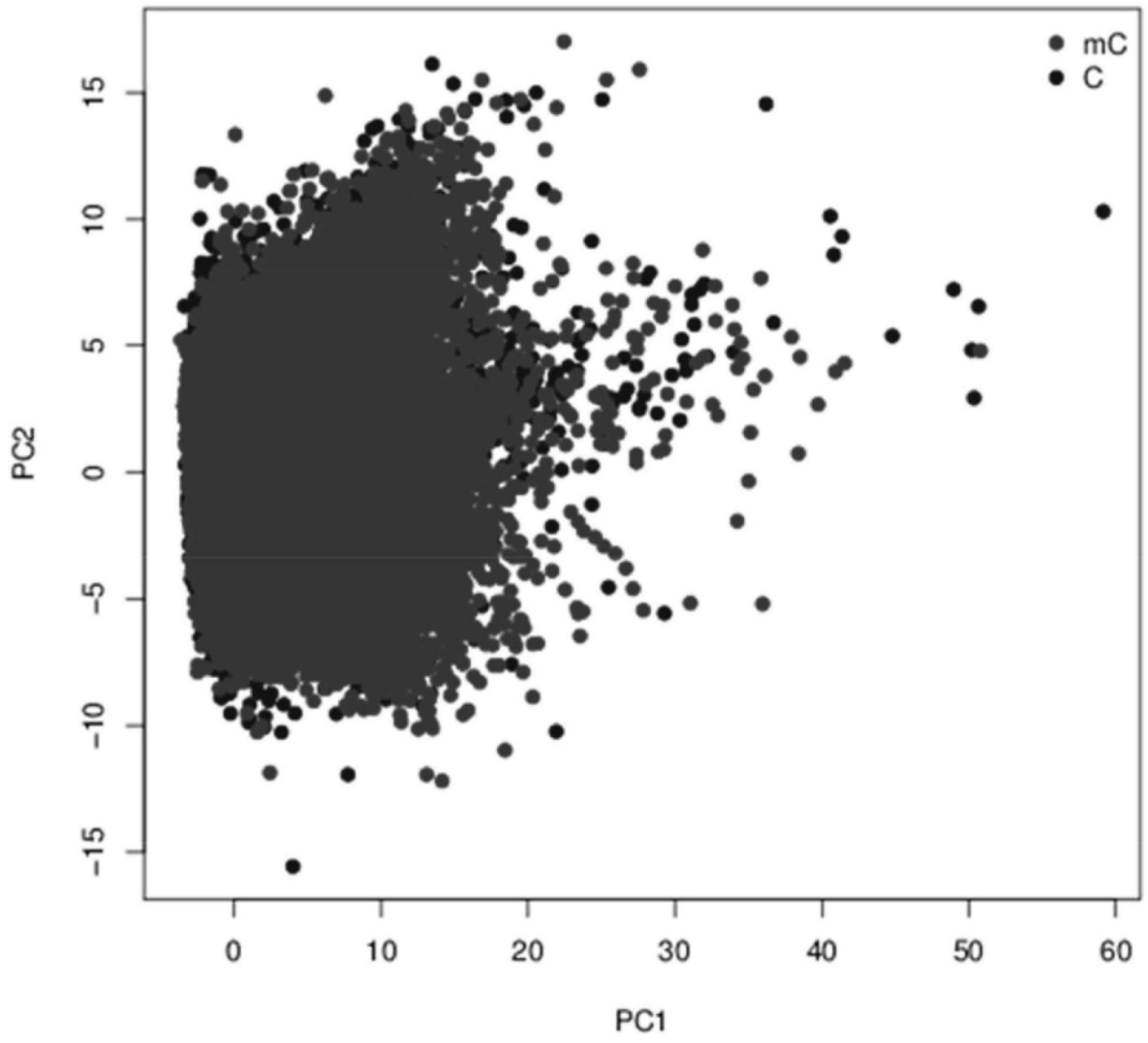


图28

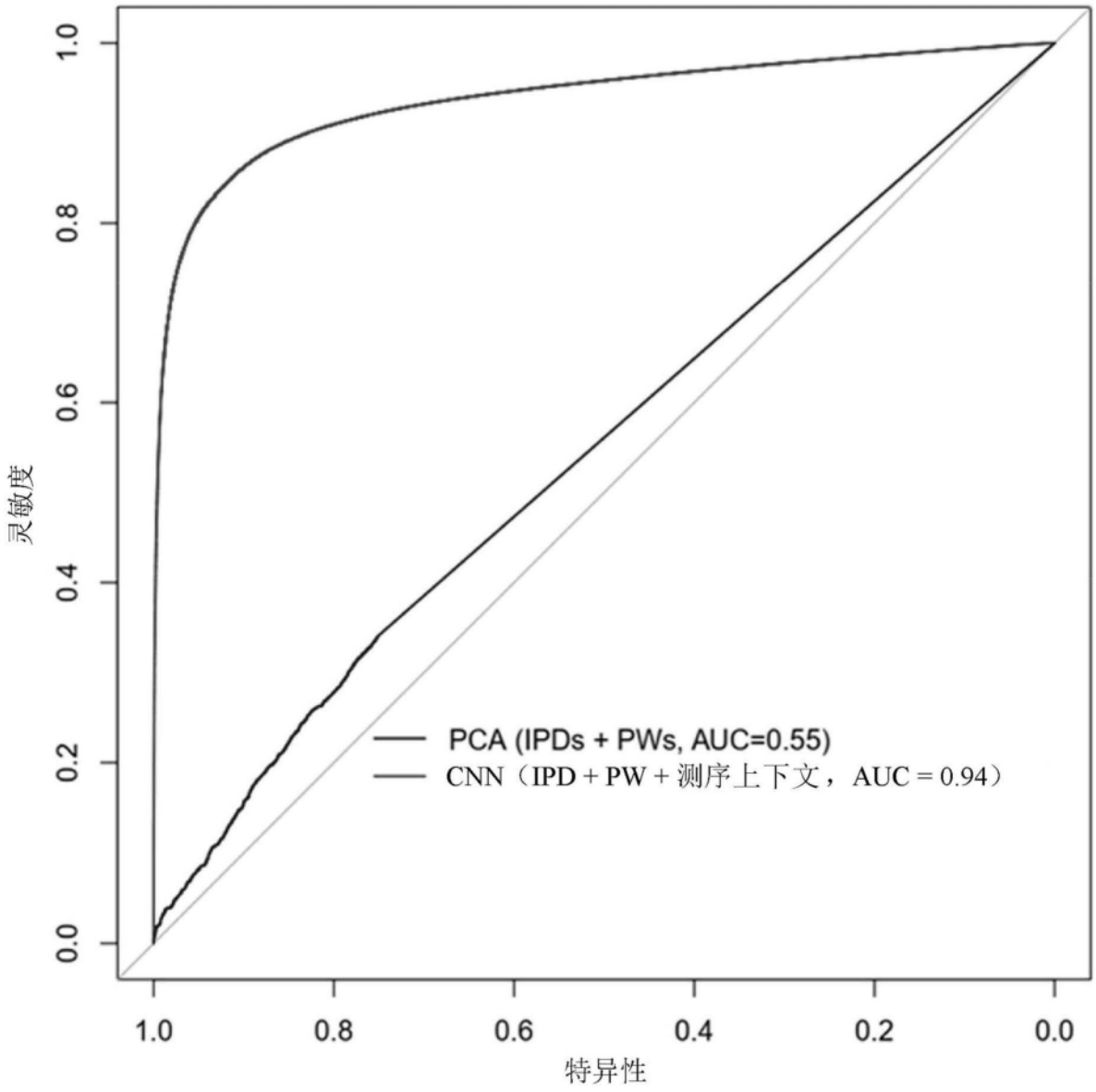

灵敏度
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
特异性
PCA (IPDs + PWs, AUC=0.55)
CNN（IPD + PW + 测序上下文，AUC = 0.94）

图29

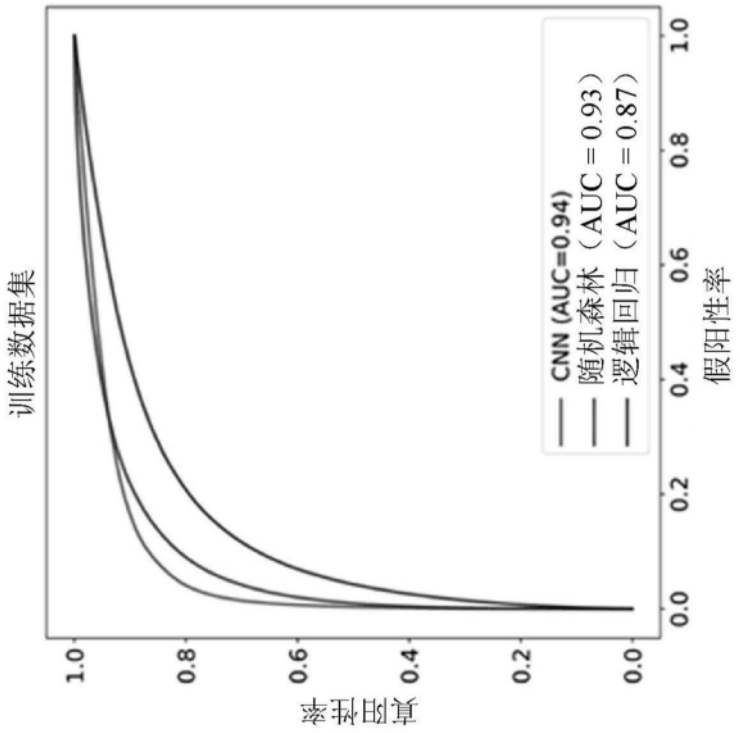

训练数据集
真阳性率
假阳性率
CNN (AUC=0.94)
随机森林（AUC = 0.93）
逻辑回归（AUC = 0.87）
0.0
0.2
0.4
0.6
0.8
1.0

图30A

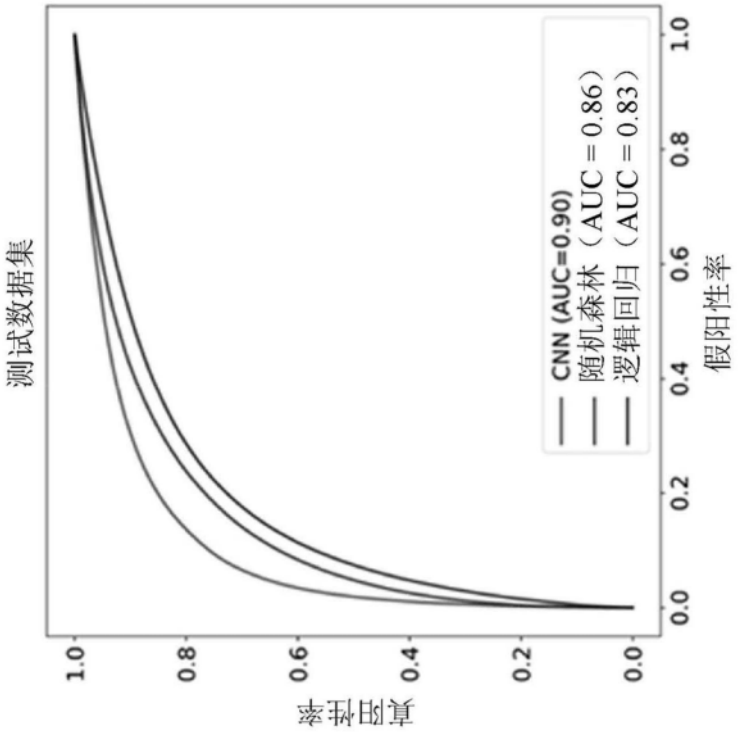

测试数据集
真阳性率
假阳性率
CNN (AUC=0.90)
随机森林（AUC = 0.86）
逻辑回归（AUC = 0.83）
0.0
0.2
0.4
0.6
0.8
1.0

图30B

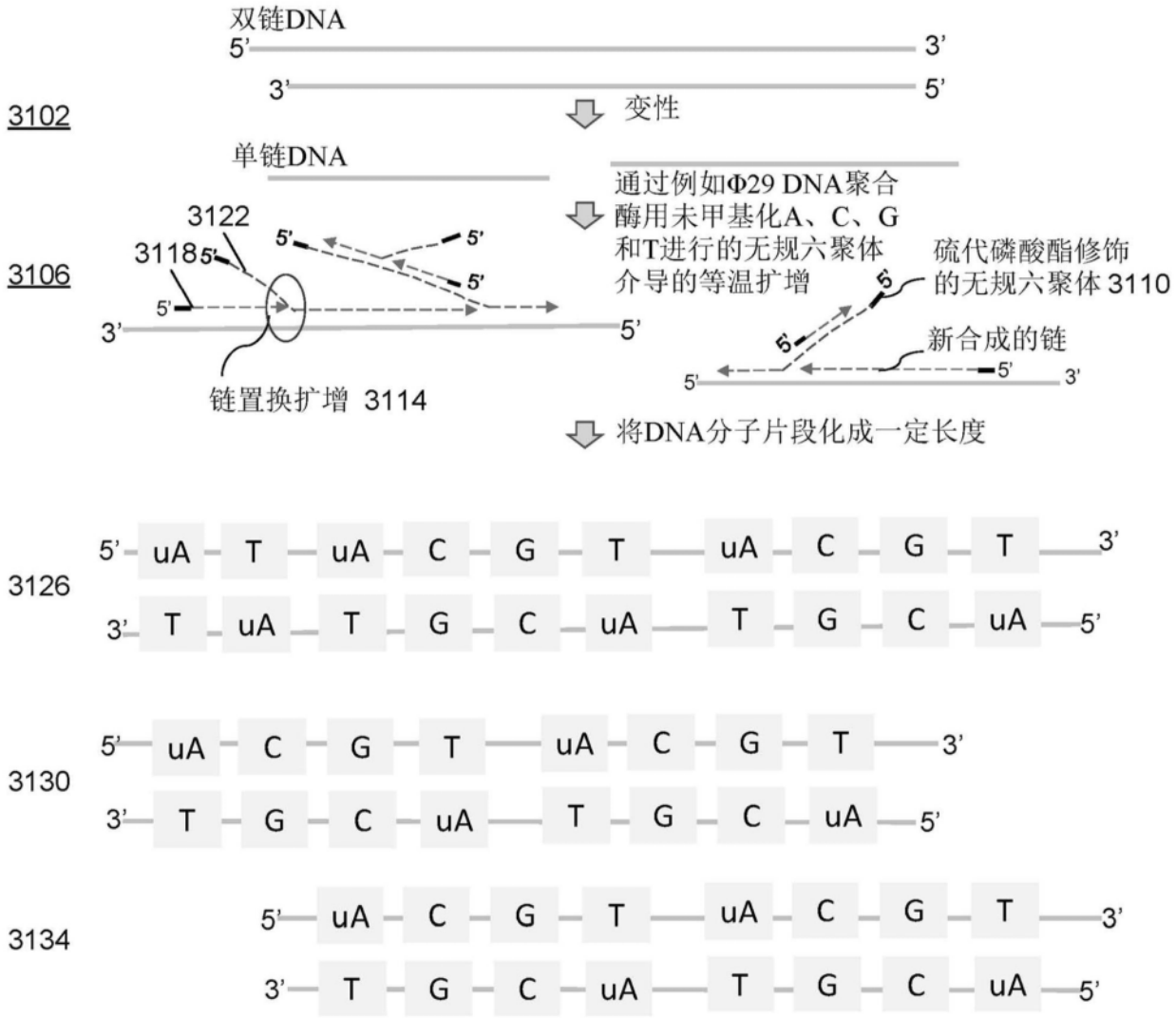
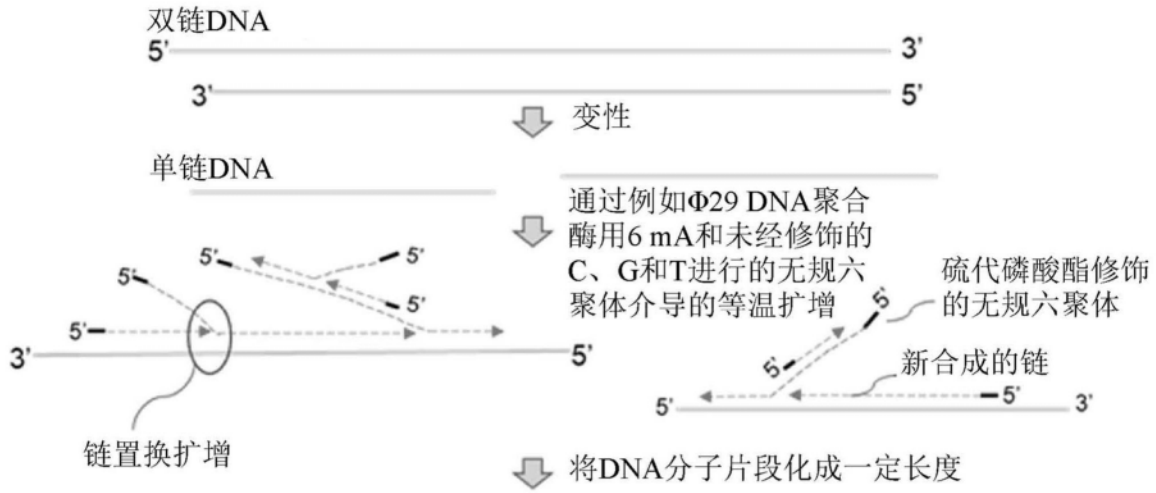
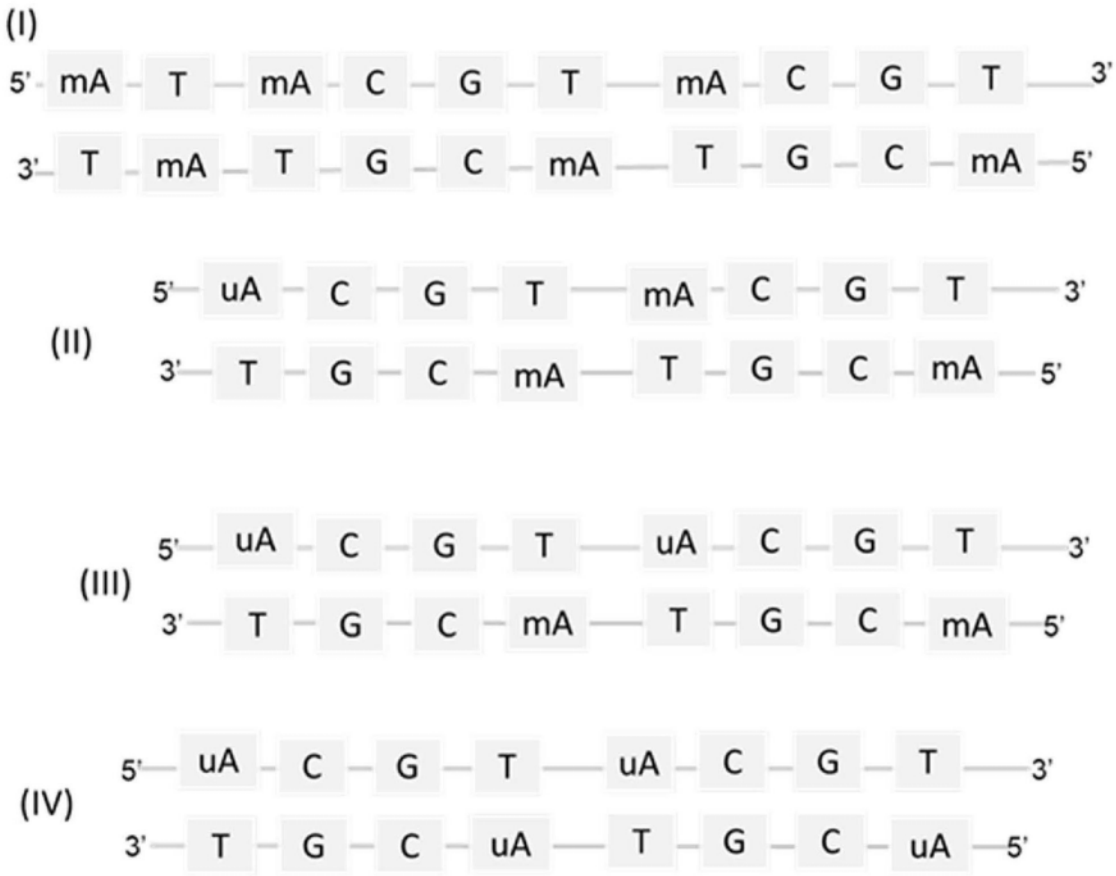

双链DNA
5'
3'
3'
5'
3102
变性
单链DNA
通过例如Φ29 DNA聚合
酶用未甲基化A、C、G
和T进行的无规六聚体
介导的等温扩增
硫代磷酸酯修饰
的无规六聚体 3110
3122
3118
3106
5'
3'
链置换扩增 3114
新合成的链
将DNA分子片段化成一定长度
3126
uA T uA C G T uA C G T
T uA T G C uA T G C uA
3130
uA C G T uA C G T
T G C uA T G C uA
3134
uA C G T uA C G T
T G C uA T G C uA
全基因组扩增的DNA产物

图31A





全基因组扩增的DNA产物

图31B

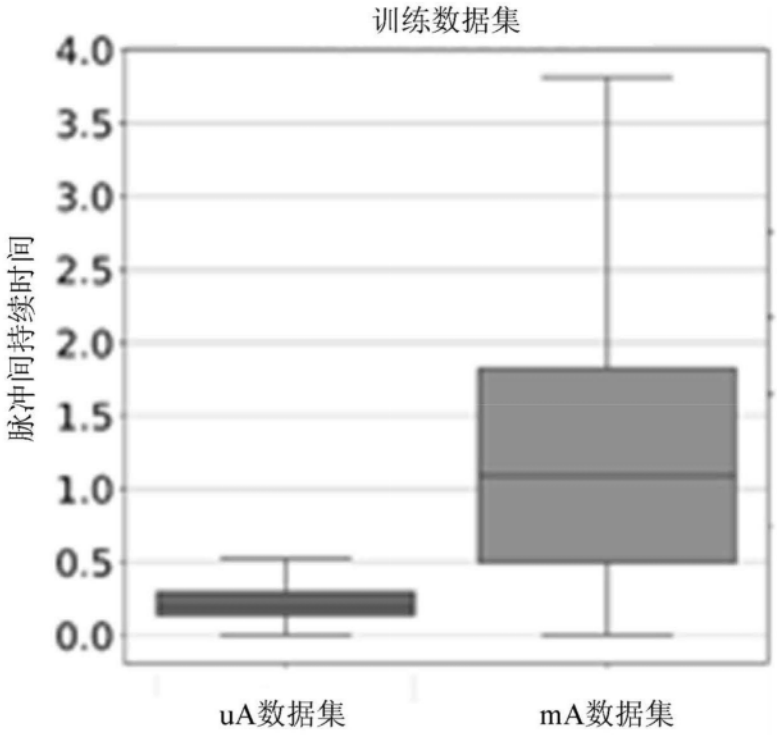

训练数据集
脉冲间持续时间
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
uA数据集
mA数据集

图32A

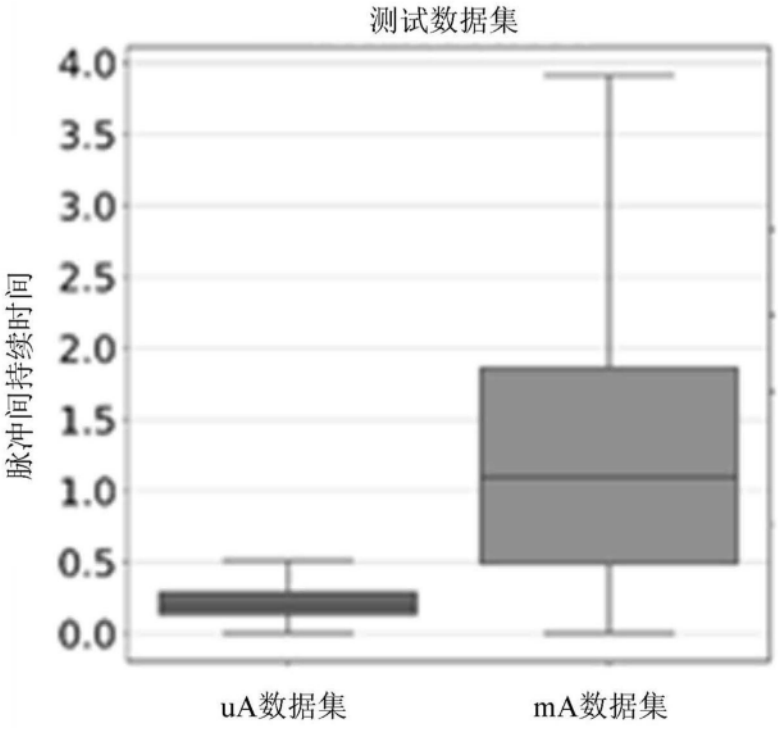

测试数据集
脉冲间持续时间
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
uA数据集
mA数据集

图32B

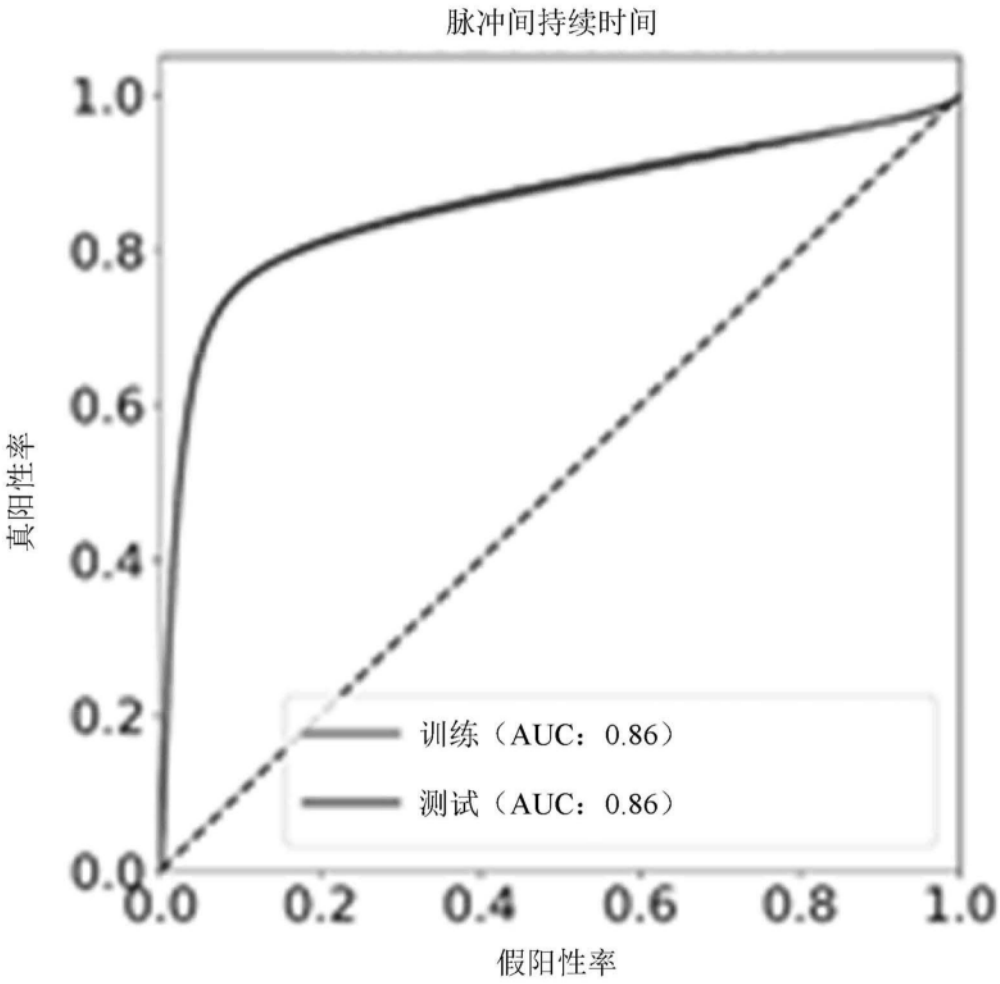

脉冲间持续时间
真阳性率
假阳性率
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
训练（AUC：0.86）
测试（AUC：0.86）

图32C

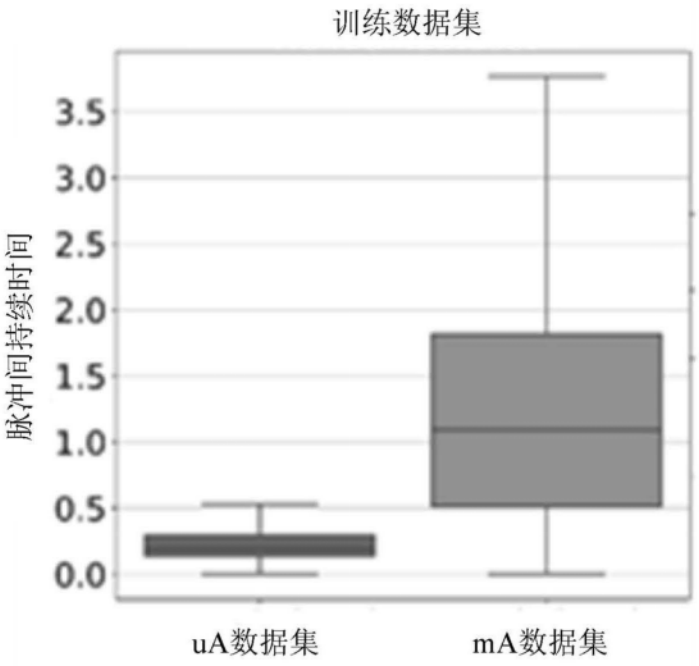

训练数据集
脉冲间持续时间
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
uA数据集
mA数据集

图33A

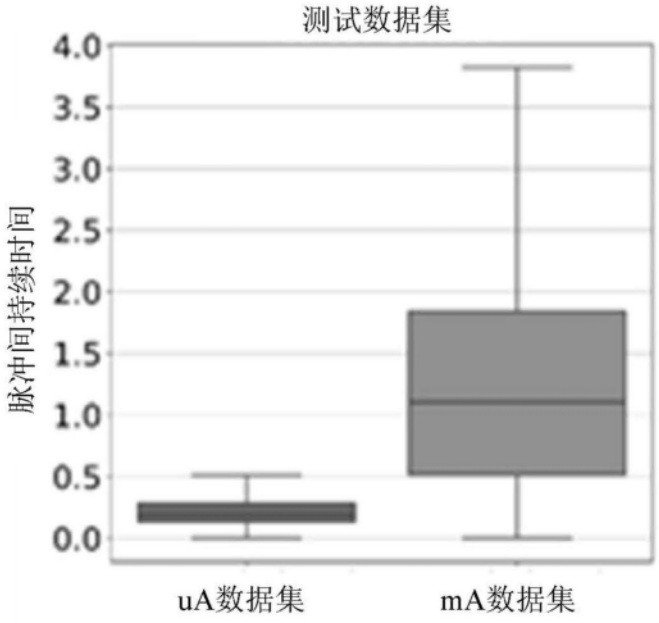

测试数据集
脉冲间持续时间
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
uA数据集
mA数据集

图33B

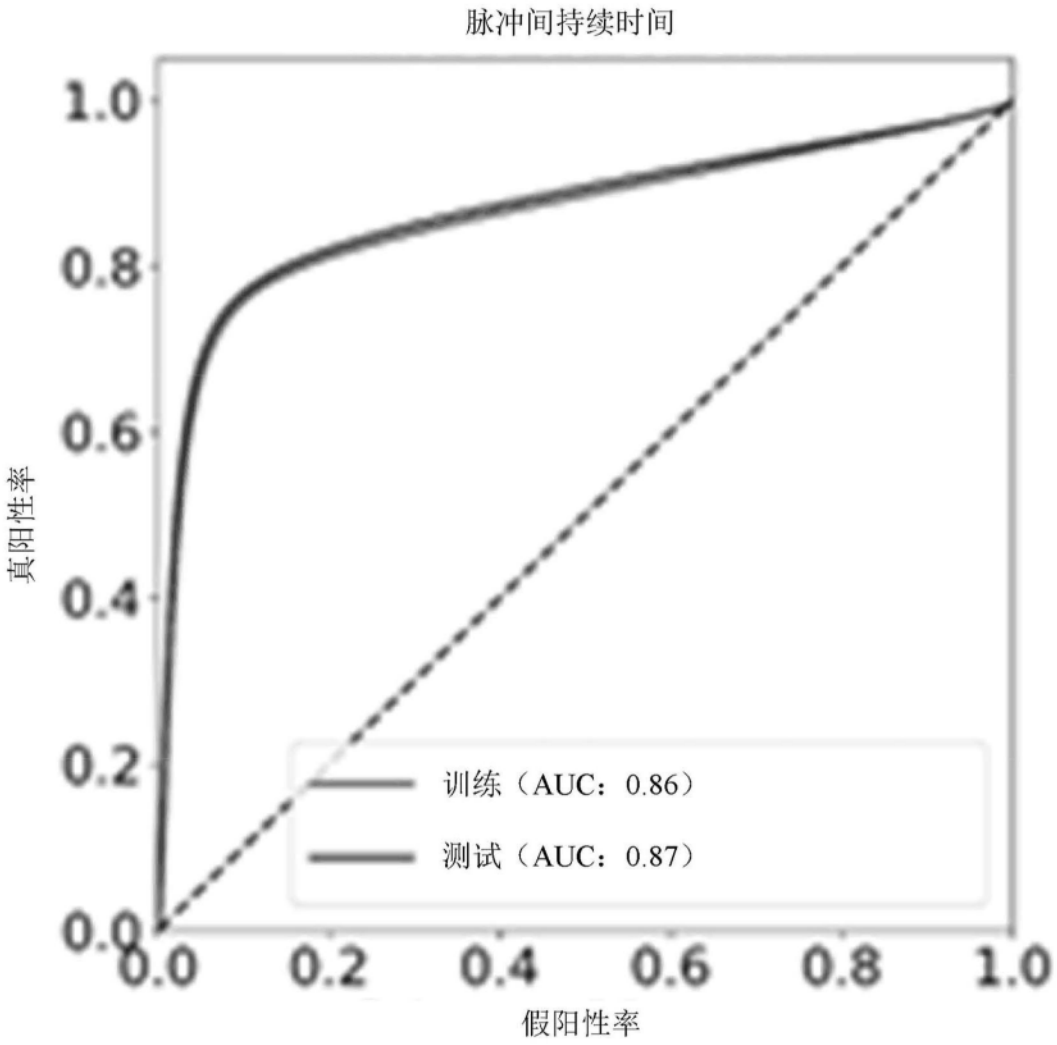

脉冲间持续时间
1.0
0.8
0.6
0.4
0.2
0.0
真阳性率
0.0
0.2
0.4
0.6
0.8
1.0
假阳性率
训练（AUC：0.86）
测试（AUC：0.87）

图33C

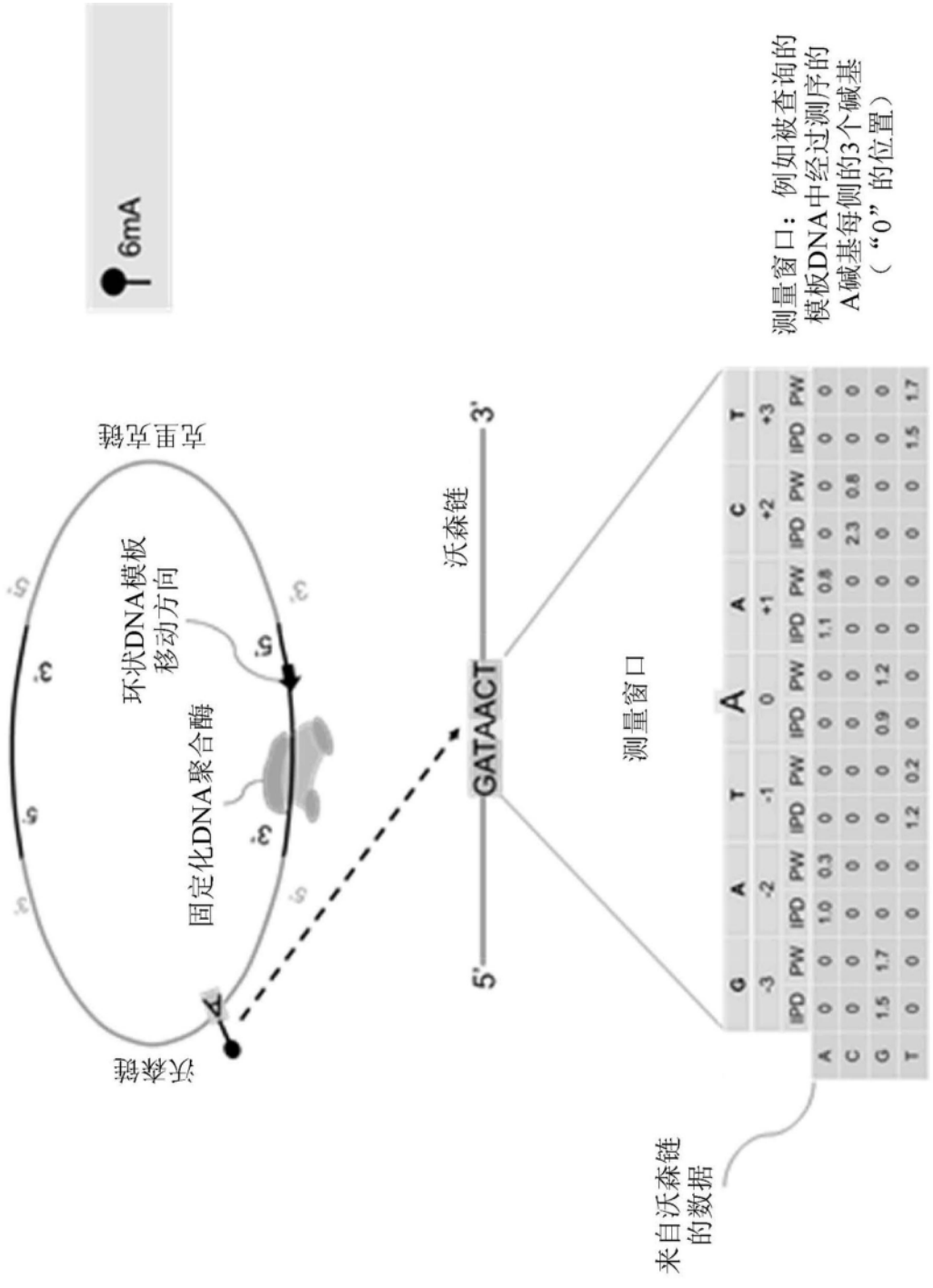
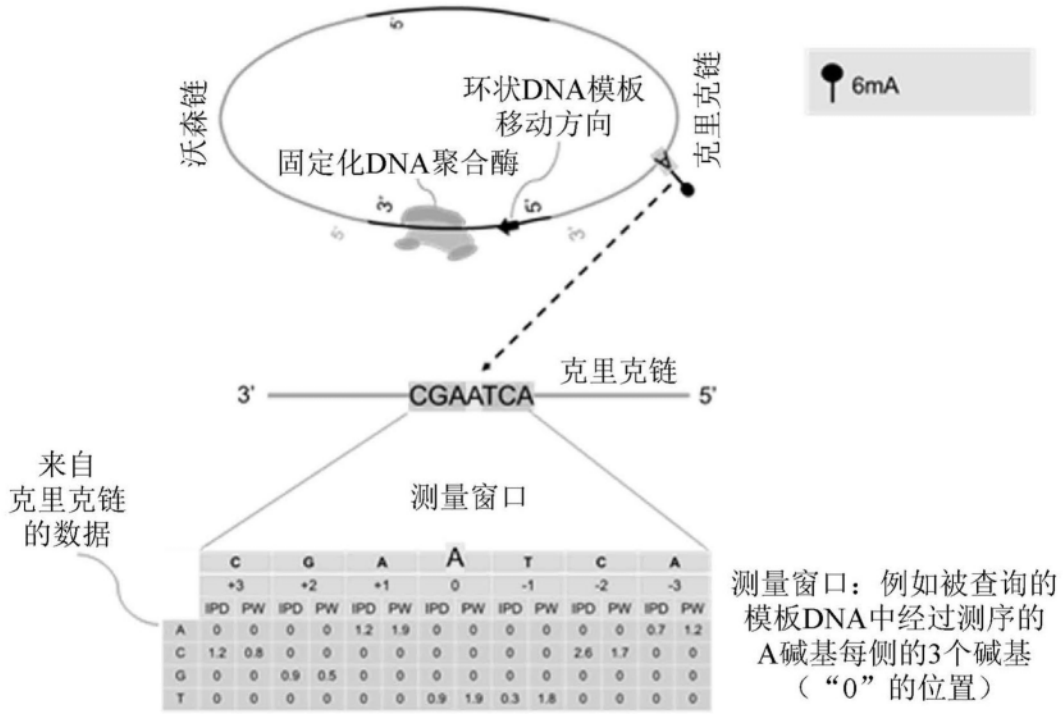

6mA
环状DNA模板
移动方向
固定化DNA聚合酶
克里克链
沃森链
沃森链
5'
3'
GATAACT
测量窗口
测量窗口：例如被查询的
模板DNA中经过测序的
A碱基每侧的3个碱基
（“0”的位置）
来自沃森链
的数据

图34



	C		G		A		A		T		C		A	
	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0.7	1.2
C	1.2	0.8	0	0	0	0	0	0	0	0	2.6	1.7	0	0
G	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0.9	1.9	0.3	1.8	0	0	0	0

图35

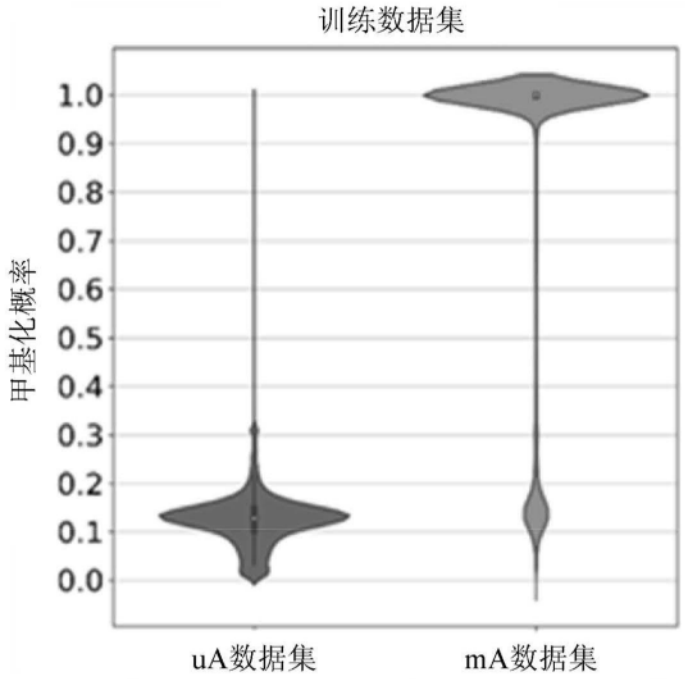


图36A

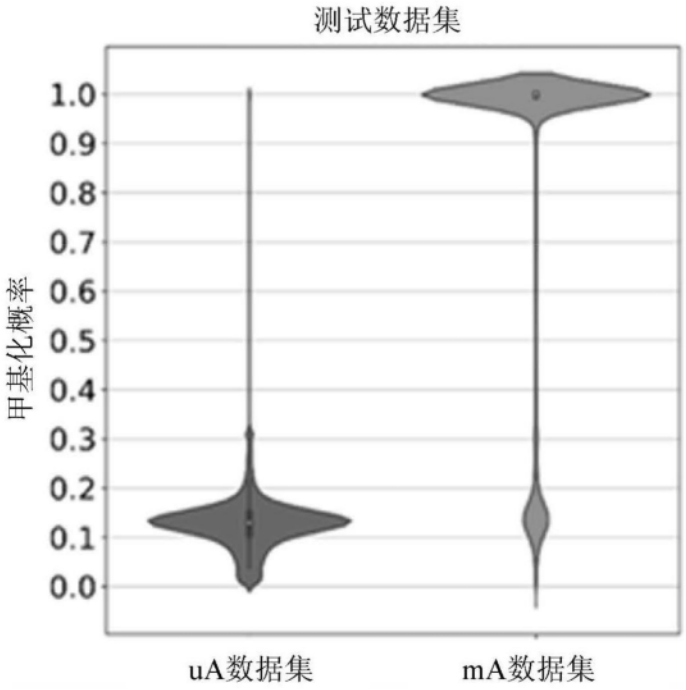


图36B

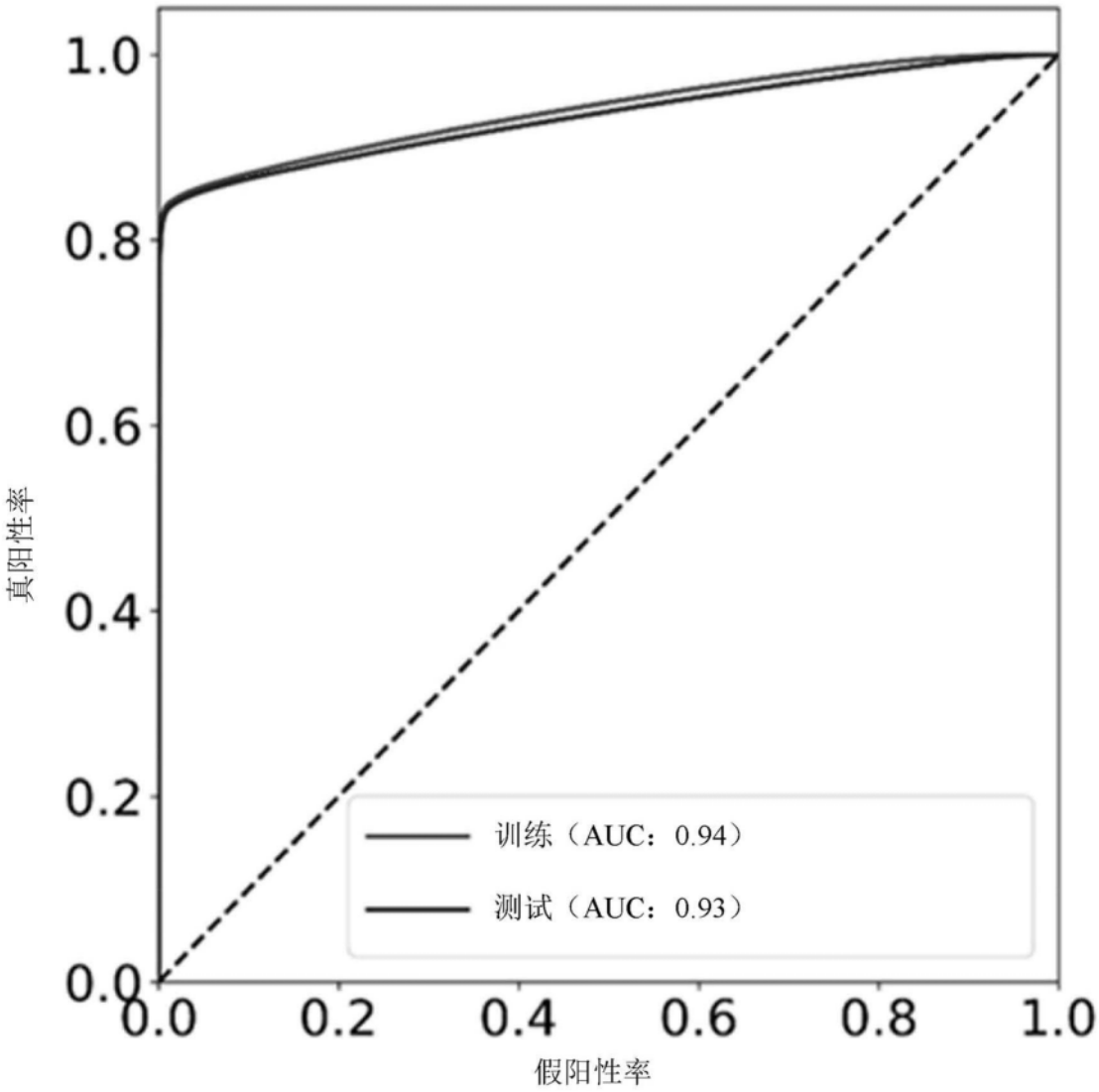


图37

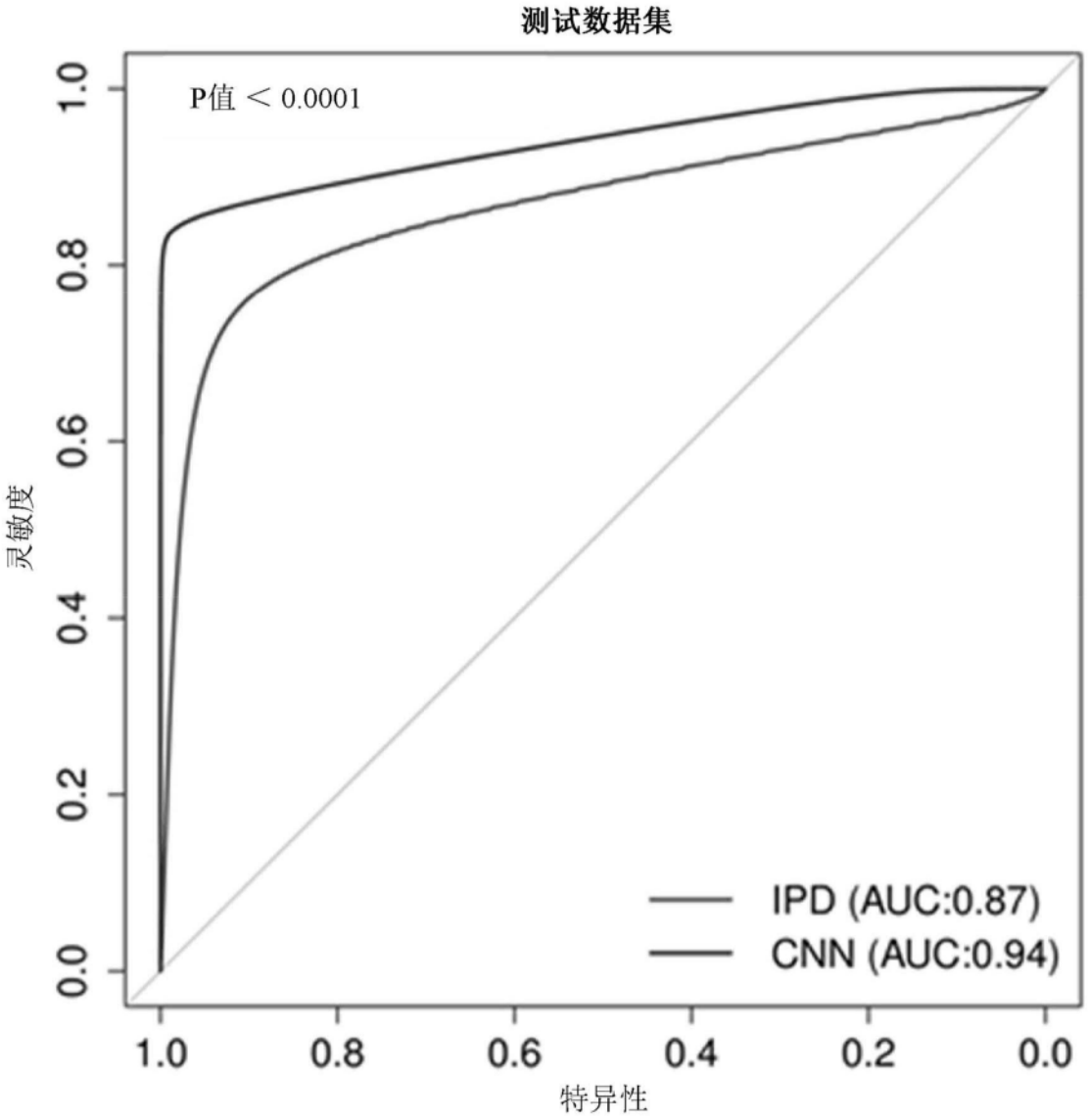


图38

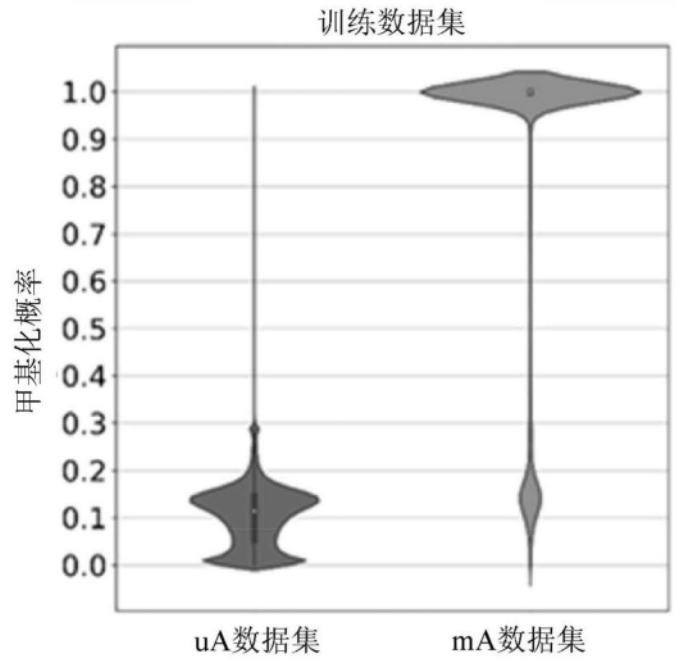

训练数据集
甲基化概率
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
uA数据集
mA数据集

图39A

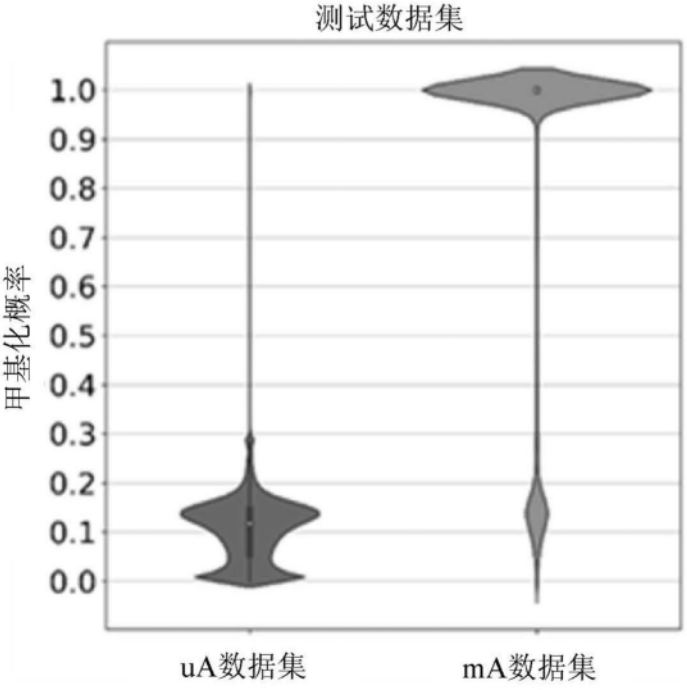

测试数据集
甲基化概率
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
uA数据集
mA数据集

图39B

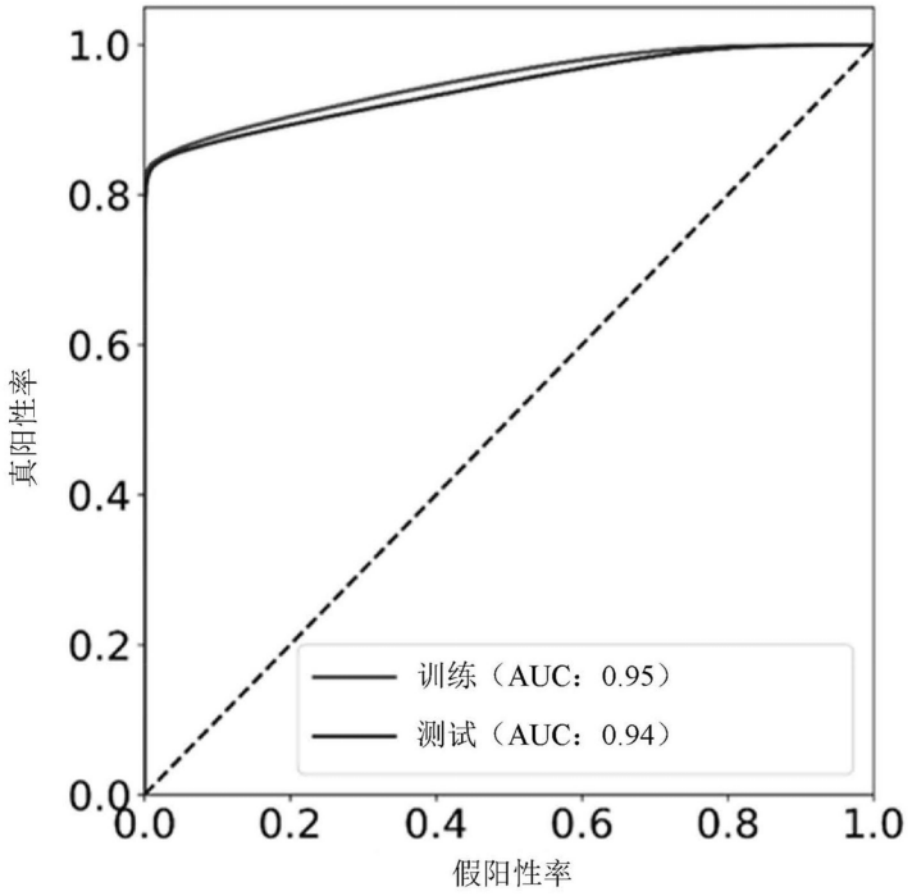


图40

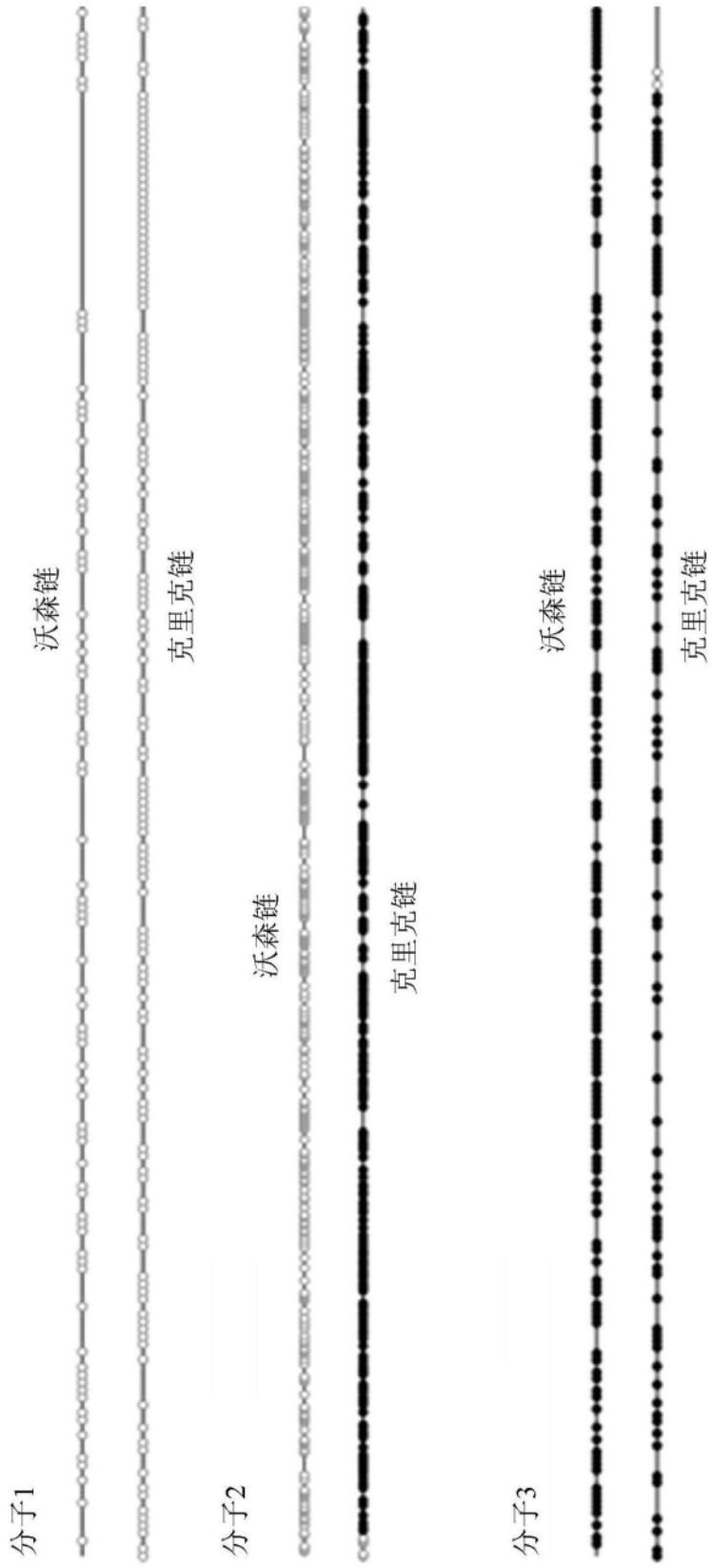


图41

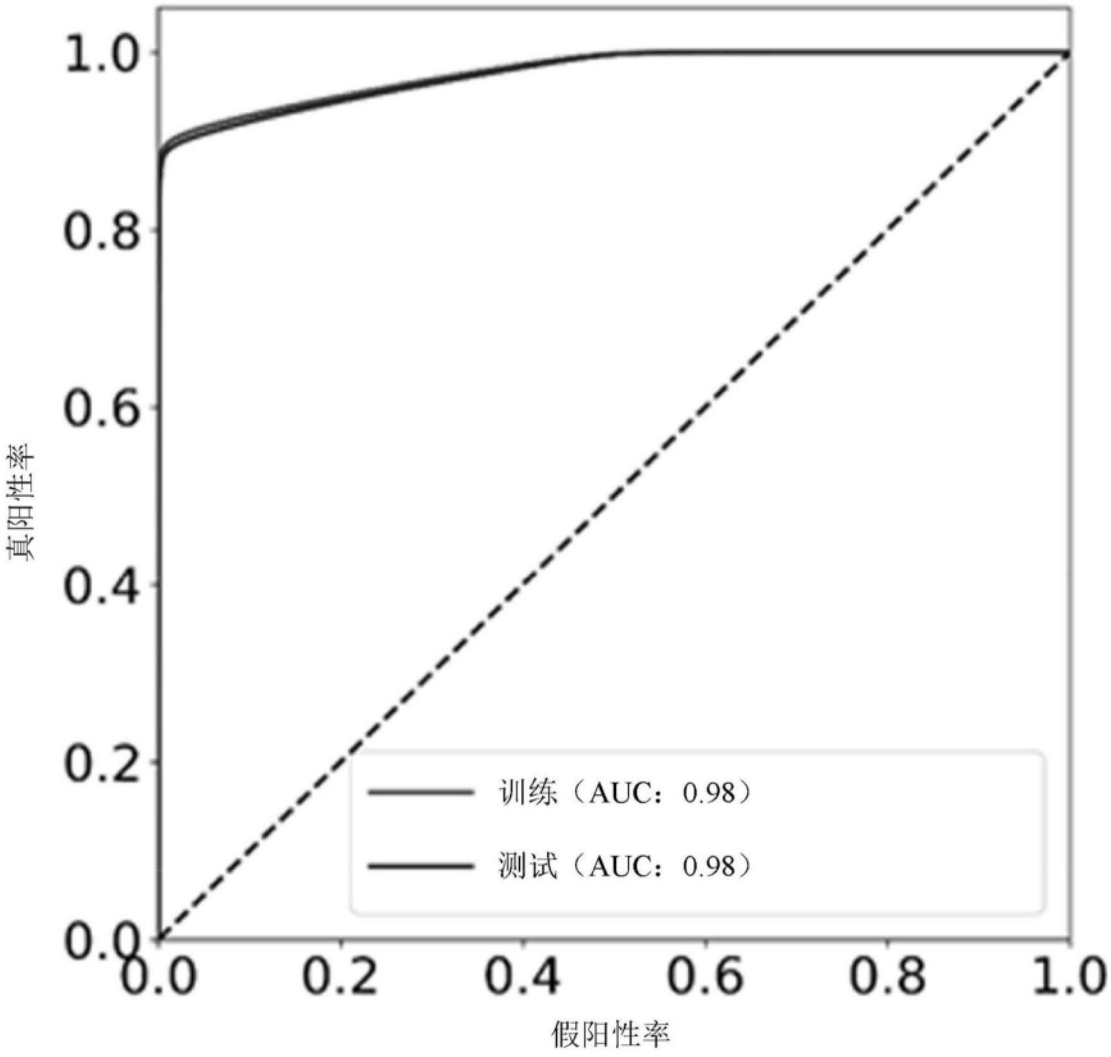


图42

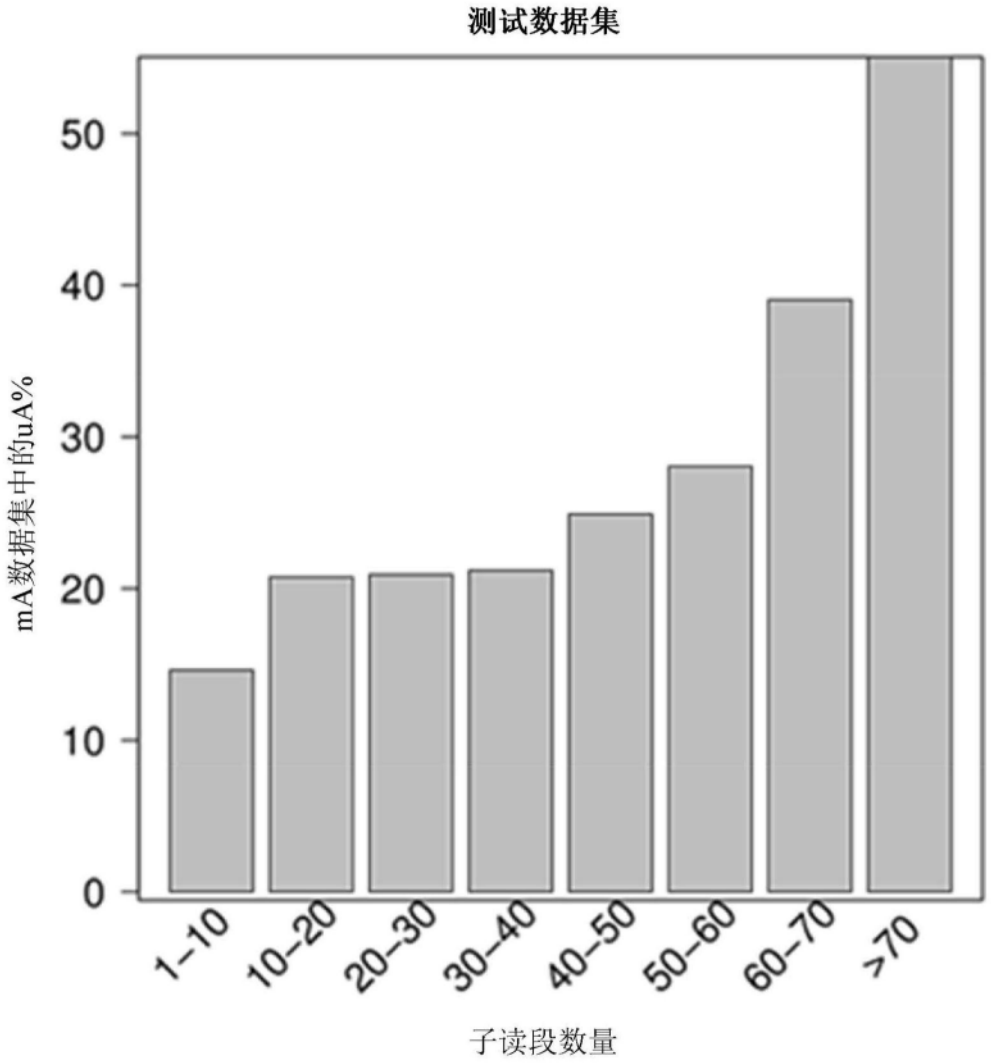

测试数据集
mA数据集中的uA%
0
10
20
30
40
50
1-10
10-20
20-30
30-40
40-50
50-60
60-70
>70
子读段数量

图43

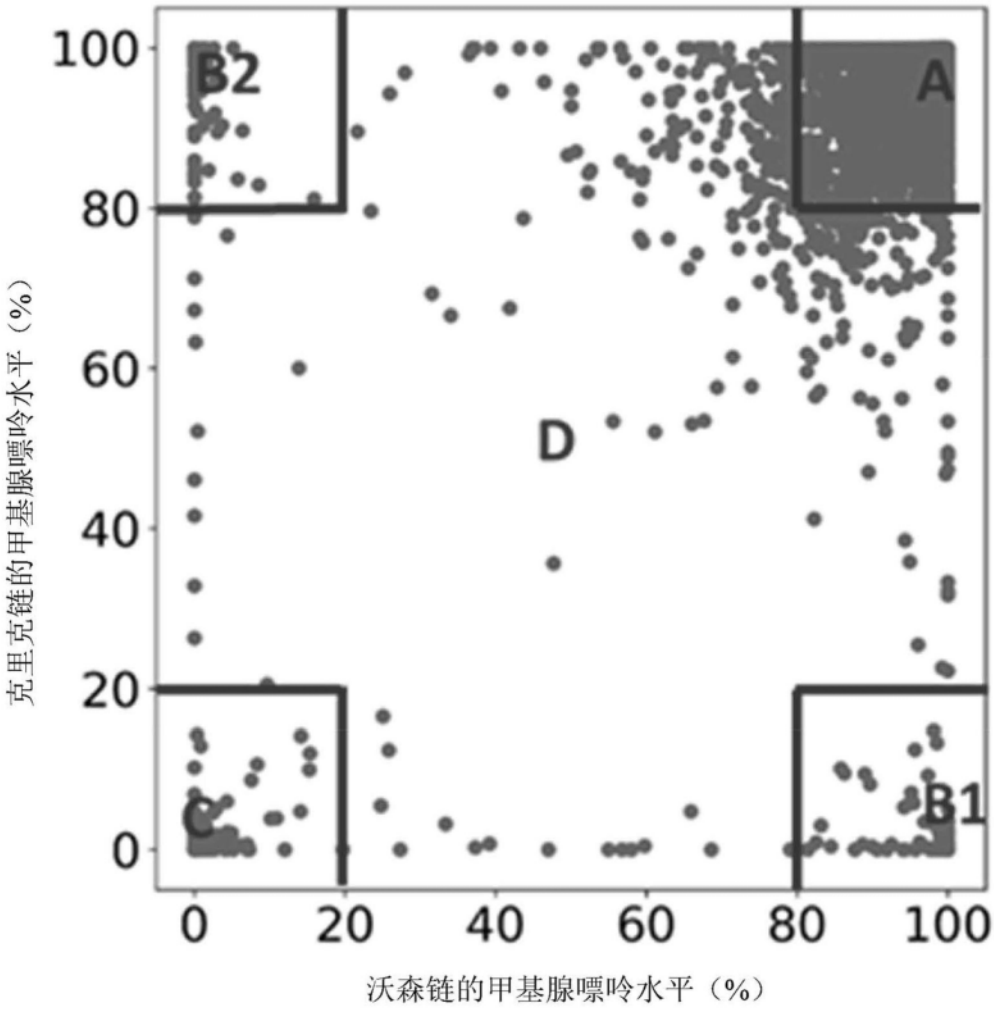


图44

类别	训练数据集	测试数据集
完全未甲基化	283 (7.0%)	276 (7.0%)
半甲基化	401 (10.0%)	389 (9.8%)
完全甲基化	3194 (79.4%)	3142 (79.4%)
交织状甲基化谱式	145 (3.6%)	148 (3.7%)

图45

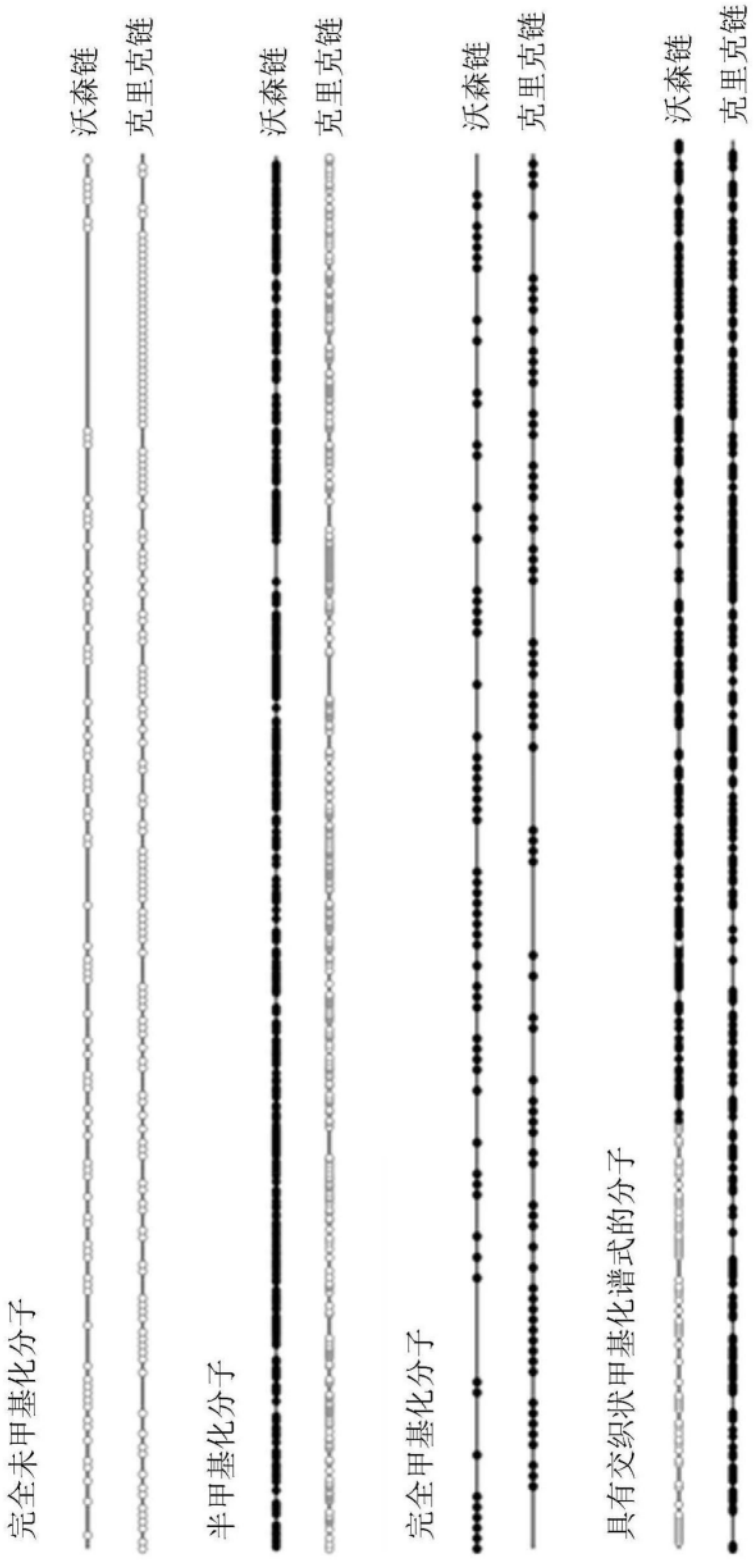

完全未甲基化分子
沃森链
克里克链
半甲基化分子
沃森链
克里克链
完全甲基化分子
沃森链
克里克链
具有交织状甲基化谱式的分子
沃森链
克里克链

图46

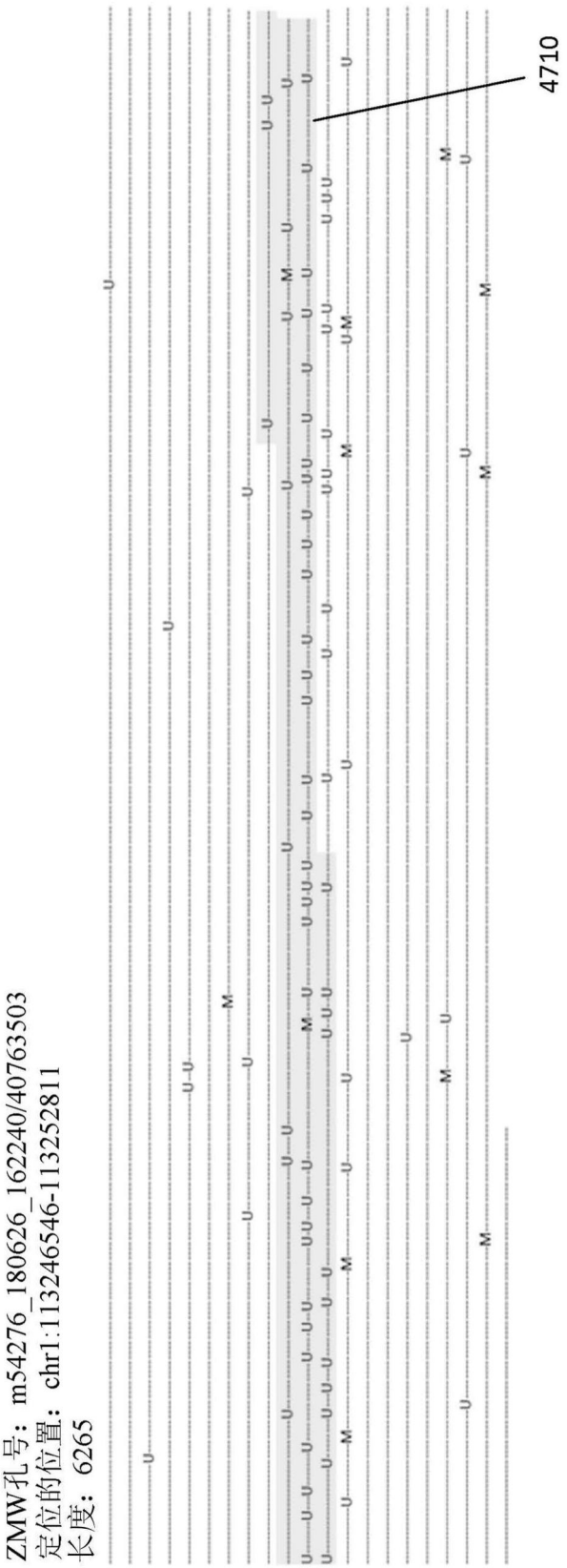

ZMW孔号：m54276_180626_162240/40763503
定位的位置：chr1:113246546-113252811
长度：6265
4710

图47

染色体	开始	结束	印记基因名称	CpG岛长度	通过PacBio SMRT测序进行测序的分子和根据本公开中存在的实施例测定的甲基化状态	分子的甲基化判读
chr11	2013333	2013617	H19	284	-U----------M----------M-M----------U---U----------M----------M---- --U----------M----------M----------M---- M----------M-M----U----------[T]----------M----------M---- --M--	甲基化
chr11	2019565	2019863	H19	298	-M-M----------M----------M----------[C]----M-M-M----M---------- ----------M----------M----M-M----------M----------M---- M----M-M--M----------M----M----------M-M----------M---- ----------	甲基化
chr11	32460586	32461004	WT1-AS/WT1	418	-U--U----------U-M------[C]----------U----------U---U---------- U----------U---U-U----------U---M--M----------U---------- ----U-U-U-----U-----U----------U---U----------U-U----------U-- ----M	非甲基化
chr14	101192851	101193499	DLK1	648	-U----------U----------U----------M--M----------U----------U---------- -U------U----------U-U--U--U-U----------U----------U- --U-U----------U----------U----U-U--U----------U-U-U----U---- M---	非甲基化
chr14	101201559	101201763	DLK1	204	-M--M----------U----------M-----M---M----------[T]---------- ----------M---M----------M---M----------M---------- ----------M----------M-------M-	甲基化
chr14	101292863	101293101	MEG3	238	----------M----------U----------U----------M- ---M----------M-M-------M-	甲基化
chr15	25981176	25981392	ATP10A	216	----*----------M----------M----------M----------M---------- M-M----------M----------[T]----M-M----------U-	甲基化
chr2	80531367	80531719	LRRTM1	352	--*---[G]----------U---U-------M-------M----------U---U-- ---U-----U----------U----------U----------M----------U---- --U-U----------M----------U----------U--U----------U-U---- --	非甲基化
chr7	79082174	79082427	MAGI2	253	----------*-U----------[A]-M-U----------U---U---U-------U--- U----------U----------*-------	非甲基化

图48

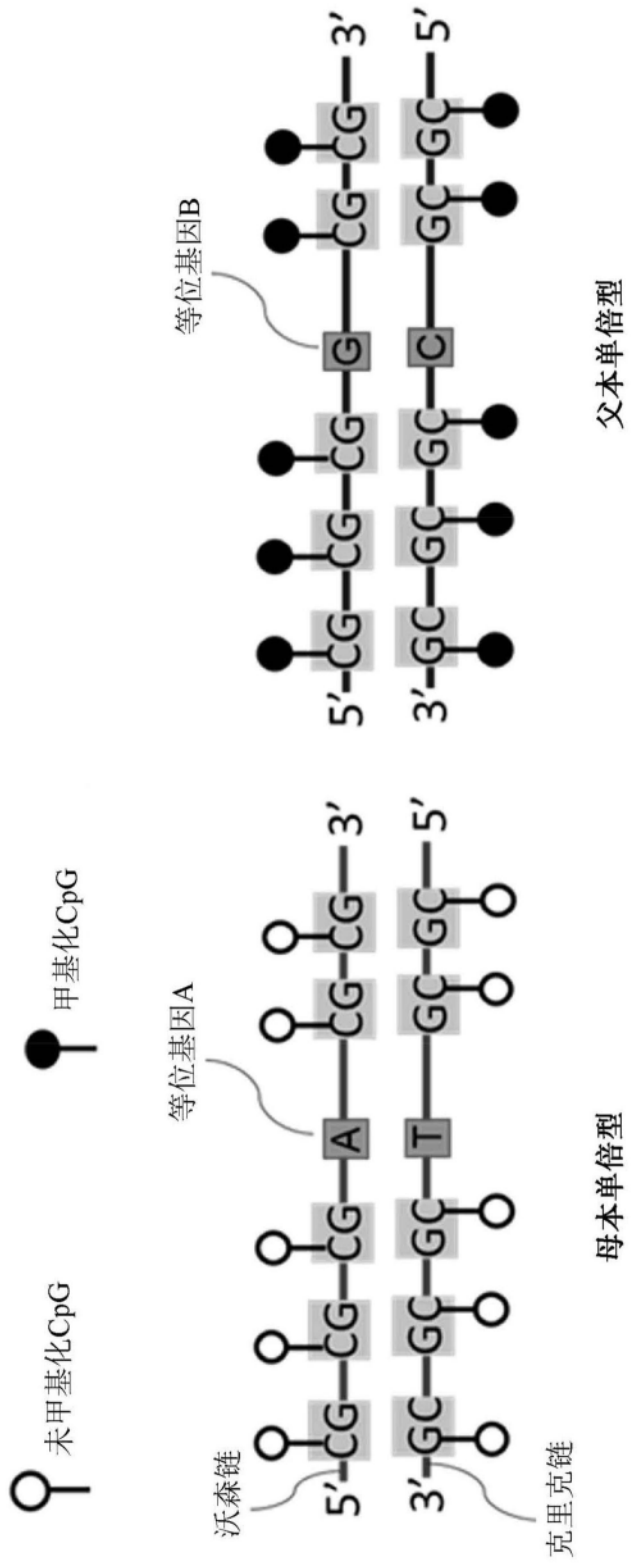

父本印记区域中存在的甲基化谱式
甲基化CpG
未甲基化CpG
等位基因A
等位基因B
沃森链
克里克链
5′
3′
A
T
G
C
母本单倍型
父本单倍型

图49

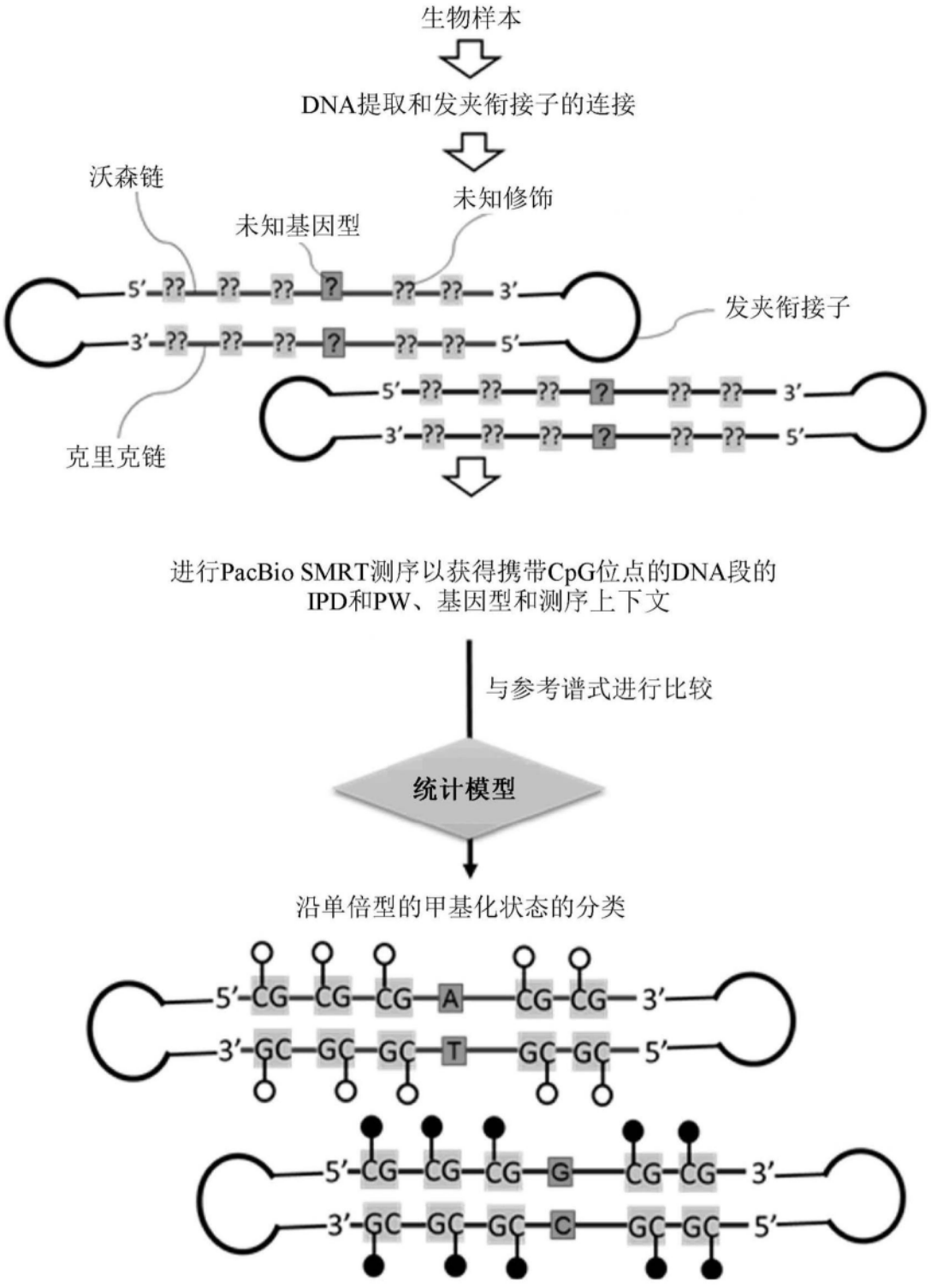

生物样本
DNA提取和发夹衔接子的连接
沃森链
未知修饰
未知基因型
发夹衔接子
克里克链
进行PacBio SMRT测序以获得携带CpG位点的DNA段的
IPD和PW、基因型和测序上下文
与参考谱式进行比较
统计模型
沿单倍型的甲基化状态的分类

图50

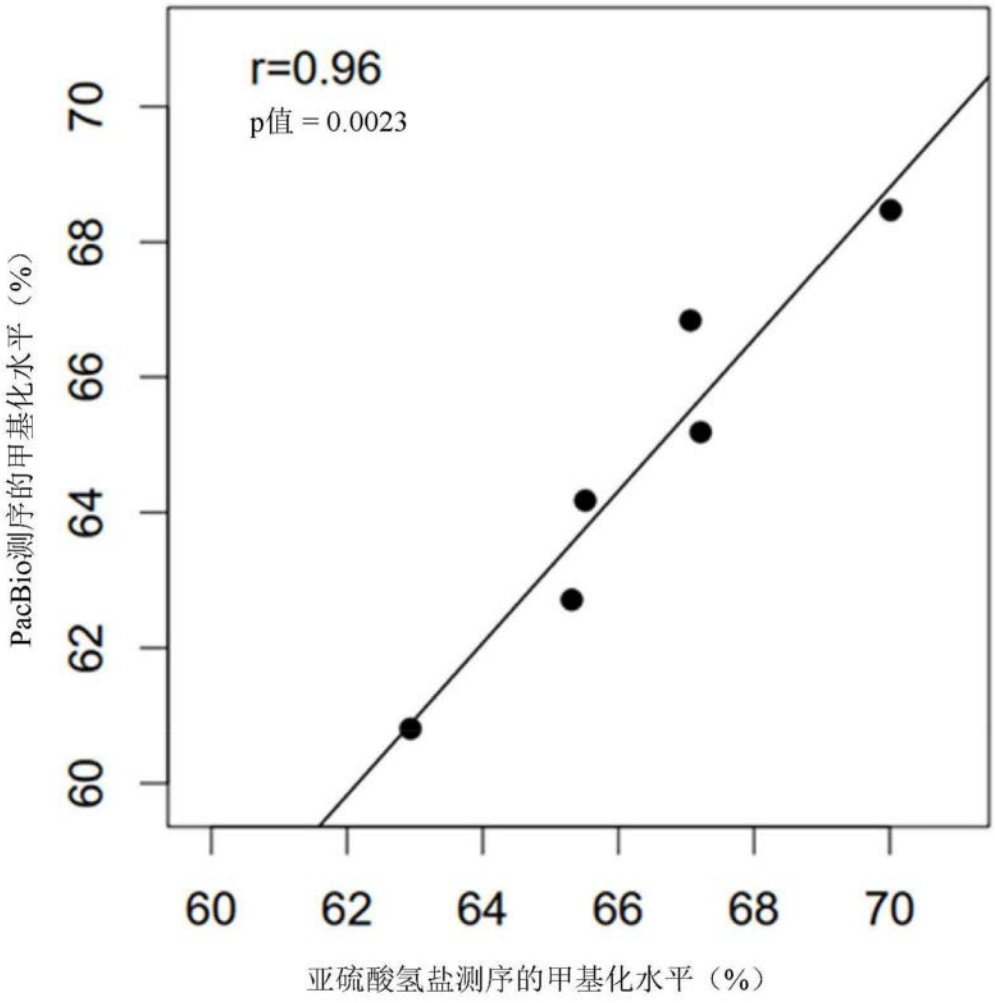

r=0.96
p值 = 0.0023
PacBio测序的甲基化水平（%）
60
62
64
66
68
70
亚硫酸氢盐测序的甲基化水平（%）

图51

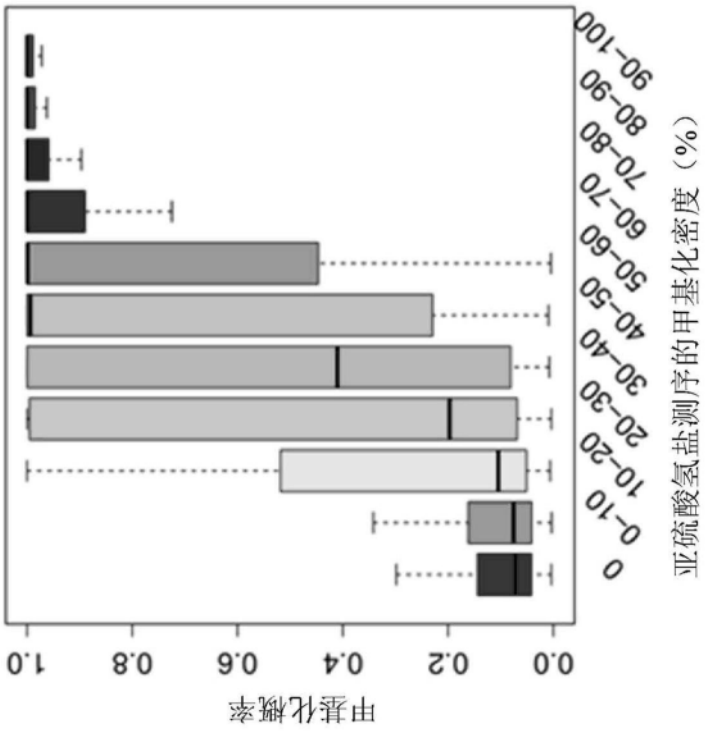

甲基化概率
0.0
0.2
0.4
0.6
0.8
1.0
亚硫酸氢盐测序的甲基化密度（%）
0
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100

图52A

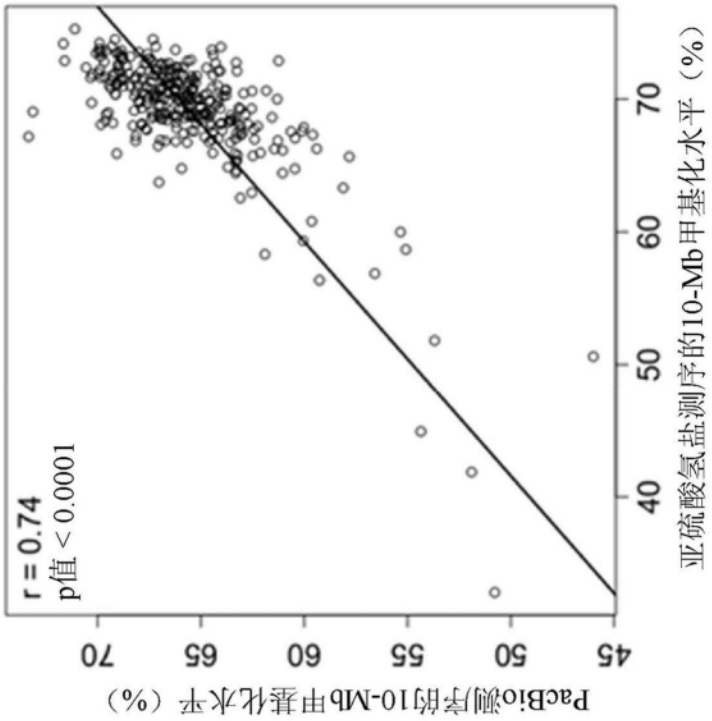

PacBio测序的10-Mb甲基化水平（%）
45
50
55
60
65
70
亚硫酸氢盐测序的10-Mb甲基化水平（%）
40
50
60
70
r = 0.74
p值 < 0.0001

图52B

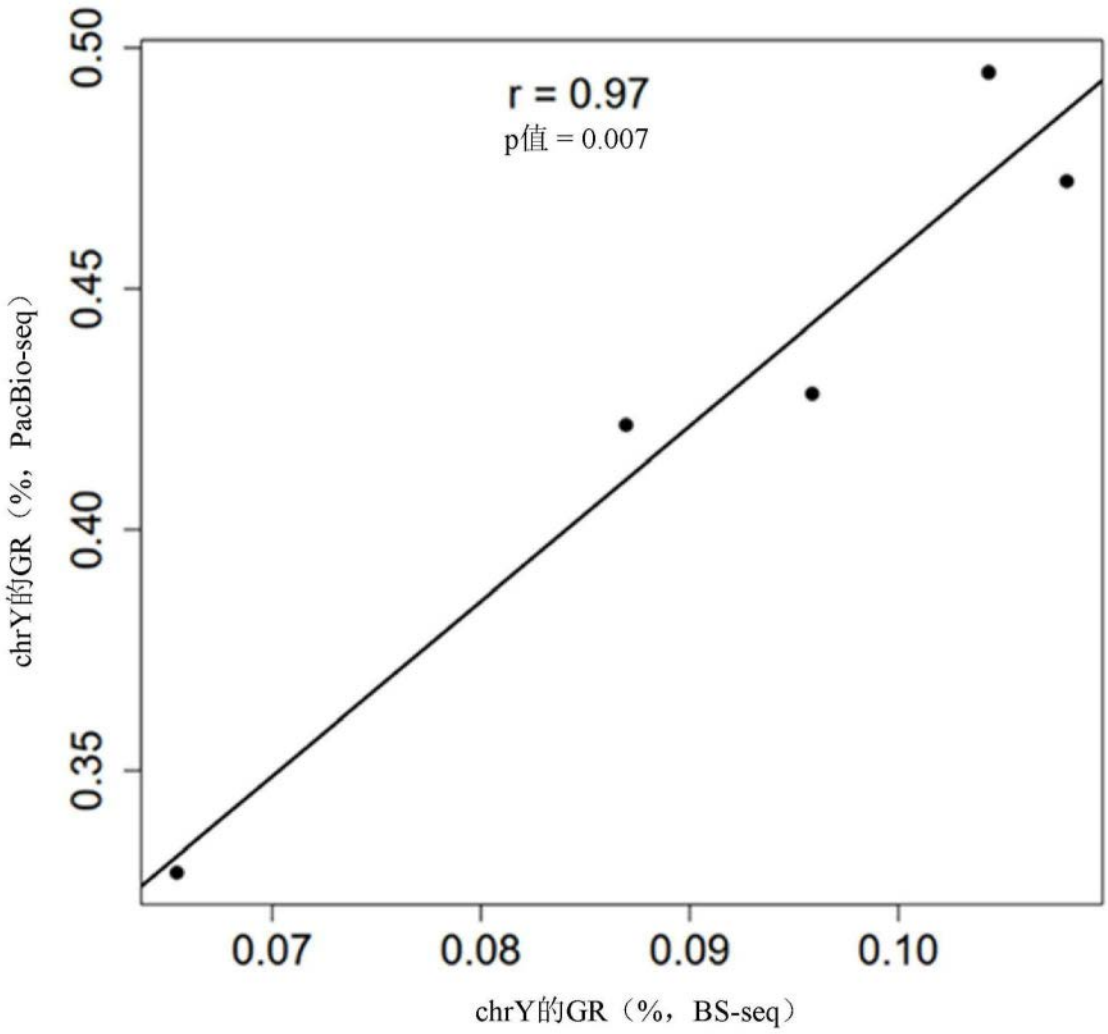

r = 0.97
p值 = 0.007
0.50
0.45
0.40
0.35
chrY的GR（%，PacBio-seq）
0.07
0.08
0.09
0.10
chrY的GR（%，BS-seq）

图53

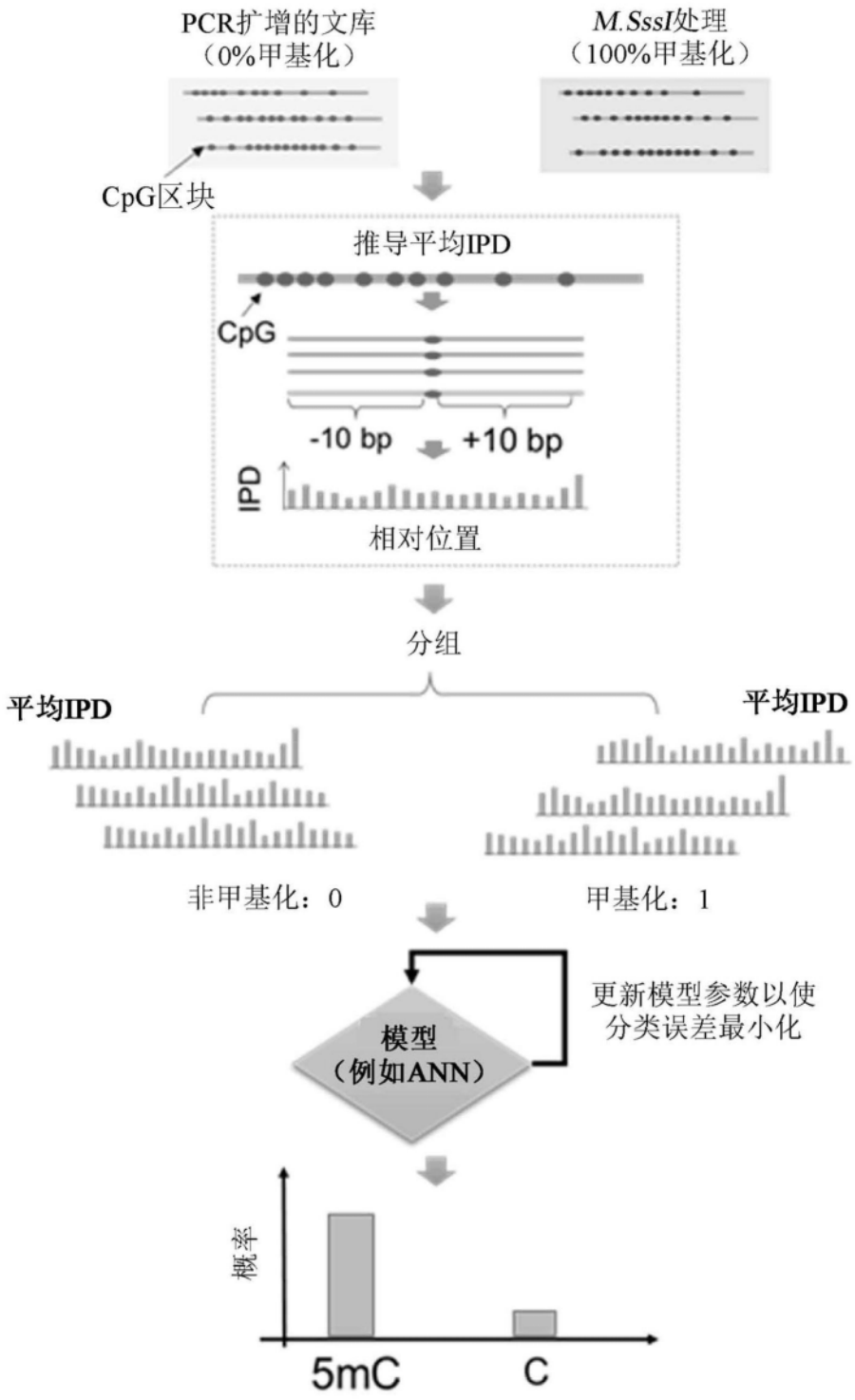

PCR扩增的文库
（0%甲基化）
M.SssI处理
（100%甲基化）
CpG区块
推导平均IPD
CpG
-10 bp
+10 bp
IPD
相对位置
分组
平均IPD
平均IPD
非甲基化：0
甲基化：1
更新模型参数以使
分类误差最小化
模型
（例如ANN）
概率
5mC
C

图54

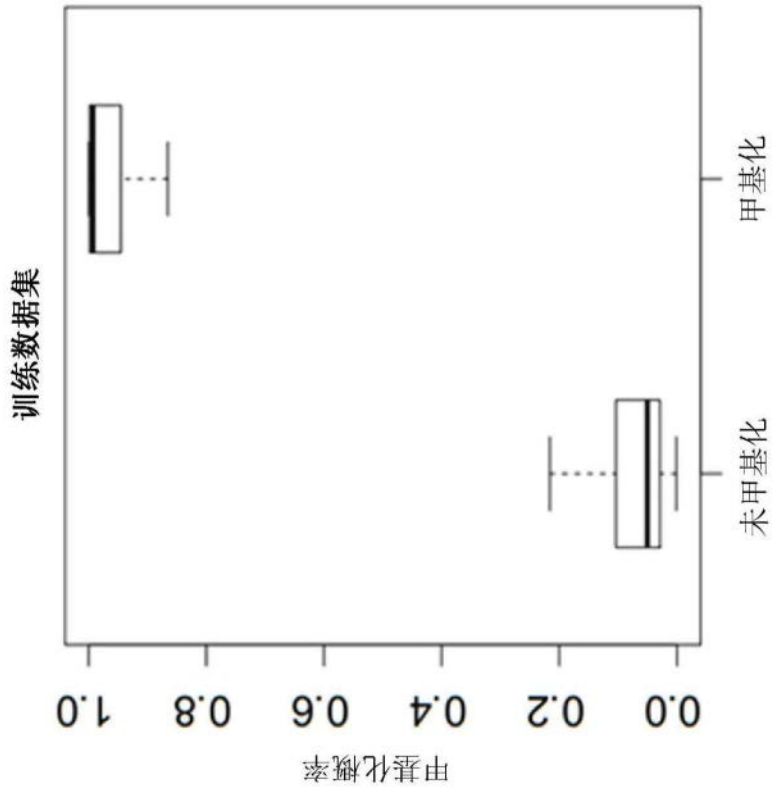

训练数据集
甲基化
未甲基化
0.0
0.2
0.4
0.6
0.8
1.0
甲基化概率

图55A

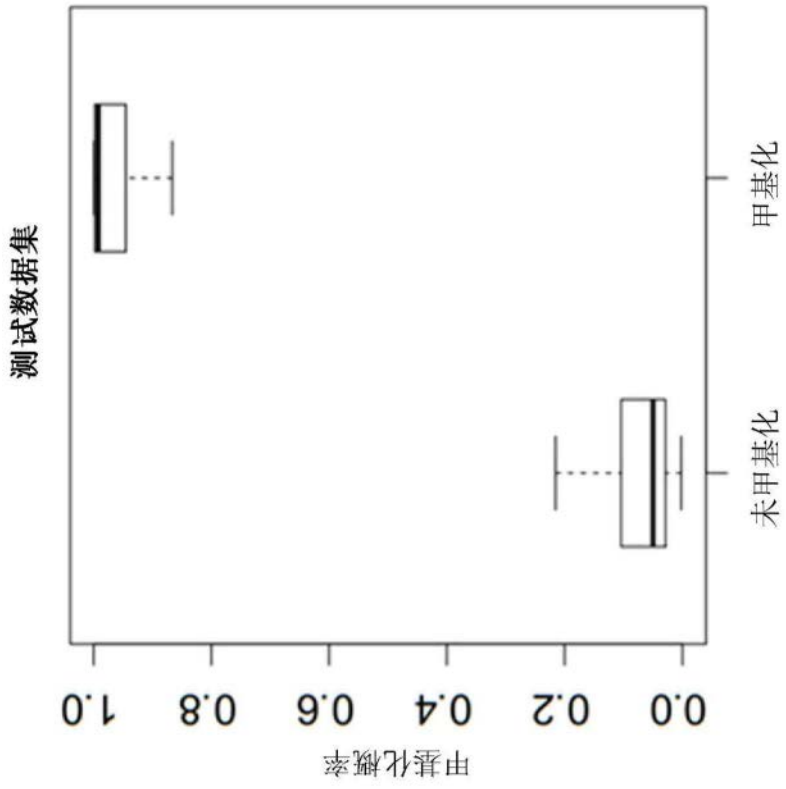

测试数据集
甲基化
未甲基化
0.0
0.2
0.4
0.6
0.8
1.0
甲基化概率

图55B

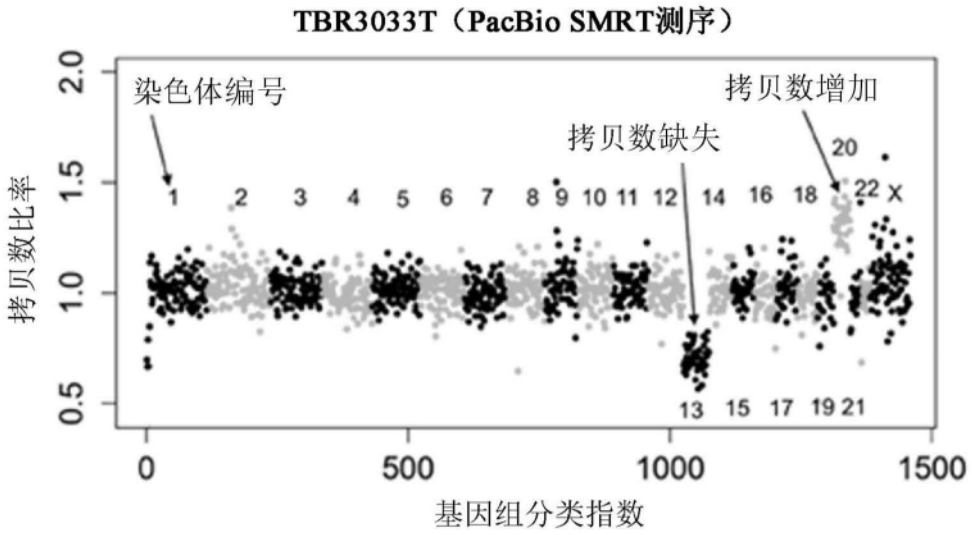

TBR3033T（PacBio SMRT测序）
染色体编号
拷贝数增加
拷贝数缺失
拷贝数比率
基因组分类指数
0.5
1.0
1.5
2.0
0
500
1000
1500
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

图56A

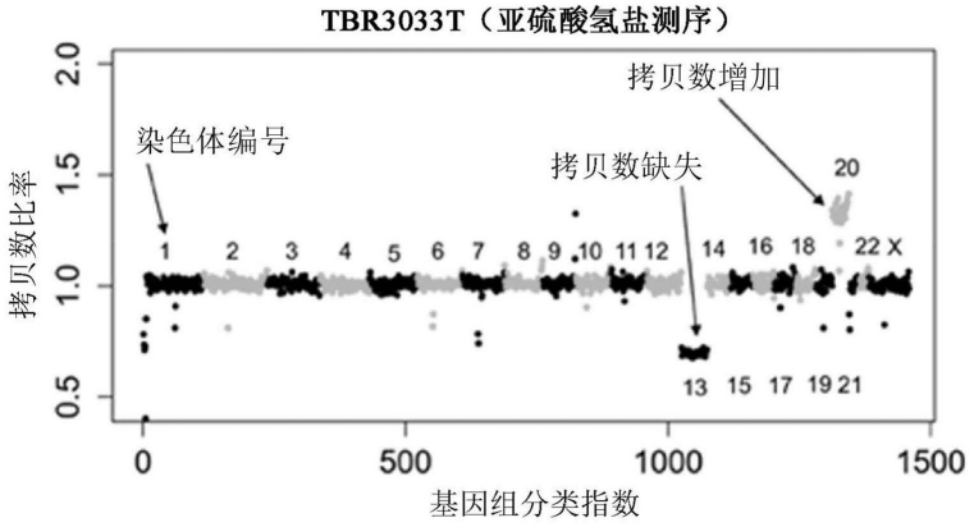

TBR3033T（亚硫酸氢盐测序）
染色体编号
拷贝数增加
拷贝数缺失
拷贝数比率
基因组分类指数
0.5
1.0
1.5
2.0
0
500
1000
1500
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

图56B

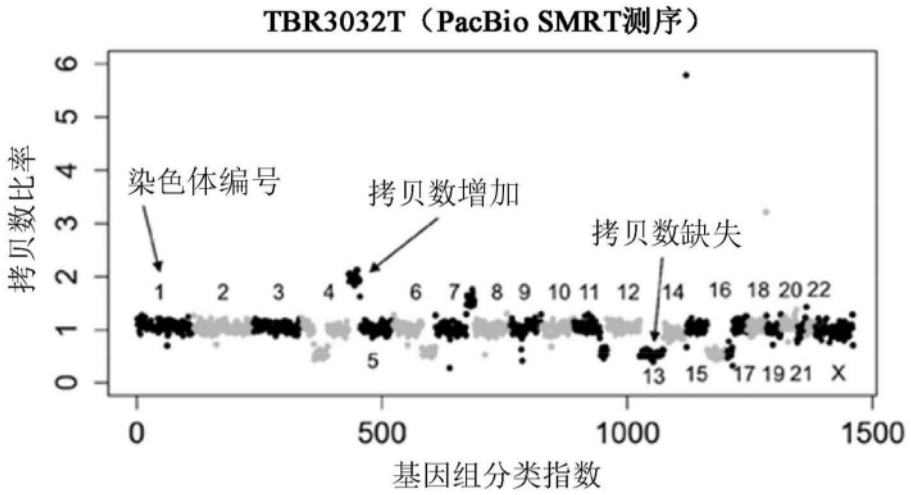

TBR3032T（PacBio SMRT测序）
拷贝数比率
染色体编号
拷贝数增加
拷贝数缺失
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
0 1 2 3 4 5 6
0 500 1000 1500
基因组分类指数

图57A

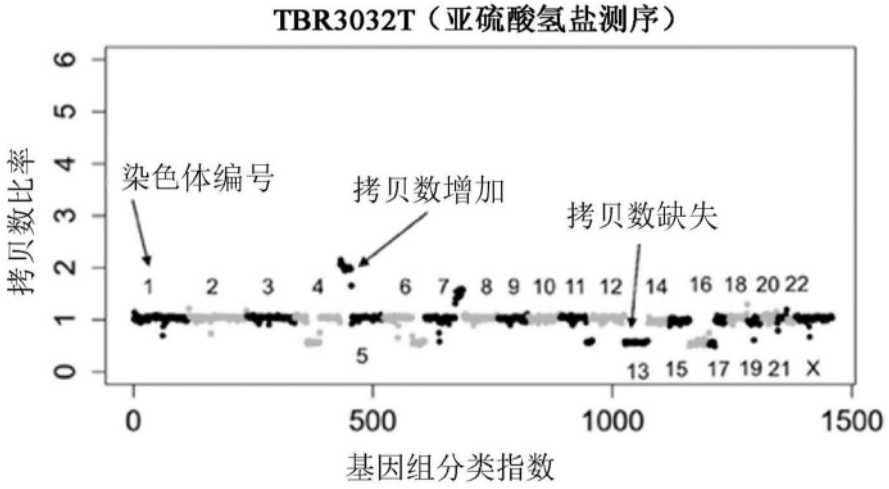

TBR3032T（亚硫酸氢盐测序）
拷贝数比率
染色体编号
拷贝数增加
拷贝数缺失
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
0 1 2 3 4 5 6
0 500 1000 1500
基因组分类指数

图57B

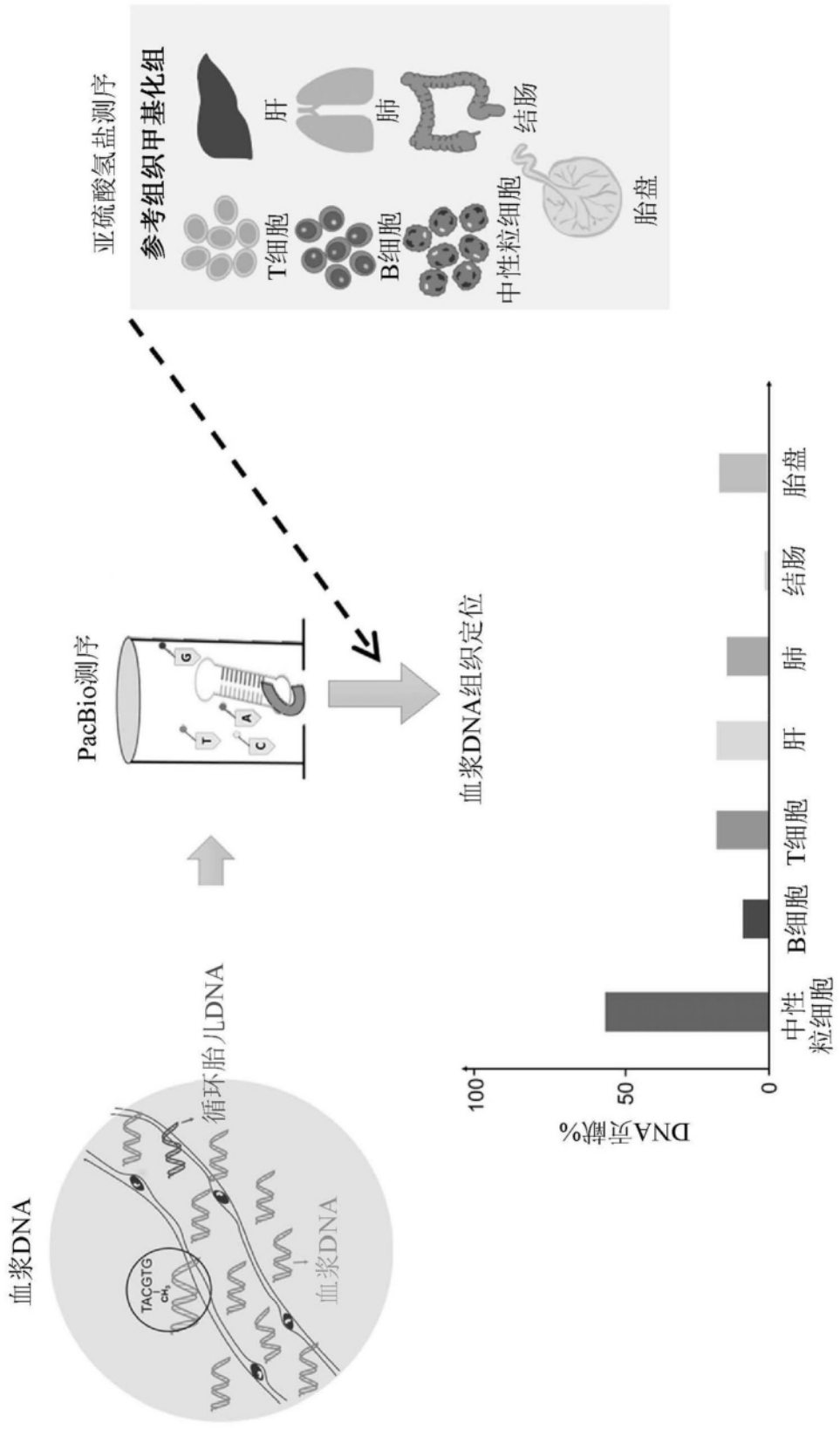

血浆DNA
TACGTG
CH3
循环胎儿DNA
血浆DNA
PacBio测序
G
T
A
C
亚硫酸氢盐测序
参考组织甲基化组
T细胞
B细胞
中性粒细胞
肝
肺
结肠
胎盘
血浆DNA组织定位
DNA贡献%
100
50
0
中性粒细胞
B细胞
T细胞
肝
肺
结肠
胎盘

图58

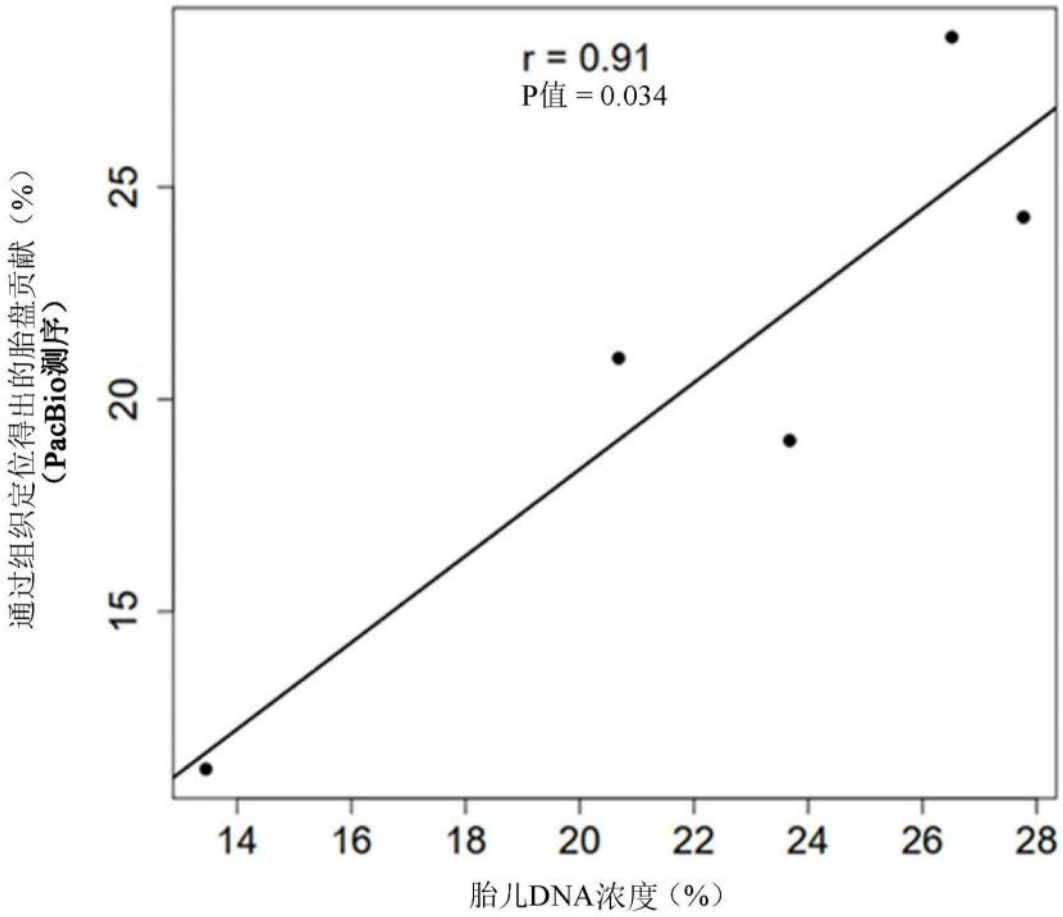

r = 0.91
P值 = 0.034
通过组织定位得出的胎盘贡献（%）
（PacBio测序）
15
20
25
14
16
18
20
22
24
26
28
胎儿DNA浓度（%）

图59

分组	样本	子读段总数	定位的子读段	子读段可定位性（%）	每个SMRT孔的平均子读段深度（x）	SMRT孔的数量	可定位孔	可定位孔率（%）
母本血沉棕黄层	M13153w	39,006,460	30,673,525	78.6	13.4	3,157,310	2,295,002	72.7
胎盘	N13153	23,013,428	16,374,758	71.2	10.4	2,393,400	1,573,540	65.7
HCC组织	TBR3032T	20,164,513	15,232,744	75.5	13.1	1,742,990	1,147,985	64.8
	TBR3033T	22,639,692	17,479,024	77.2	8.1	2,832,627	2,157,196	76.2
邻近正常组织	TBR3033N	73,118,110	56,446,202	77.2	12.6	6,881,142	4,471,370	65.0
	TBR3032N	76,852,680	60,145,452	78.3	12.8	6,000,227	4,702,130	78.4
血沉棕黄层（健康对照受试者）	M1	44,777,423	28,325,587	63.3	7.7	7,316,000	3,659,996	50.0
	F2	49,840,758	32,994,645	66.2	8.6	7,215,112	3,823,329	53.0
	F1	40,012,804	24,717,289	61.8	6.5	7,301,768	3,800,392	52.0
	M2	152,530,411	88,596,520	58.1	7.7	21,794,606	11,563,500	53.1
HCC细胞系	HepG2	47,308,982	34,581,721	73.1	7.3	6,220,000	4,750,581	76.4

图60

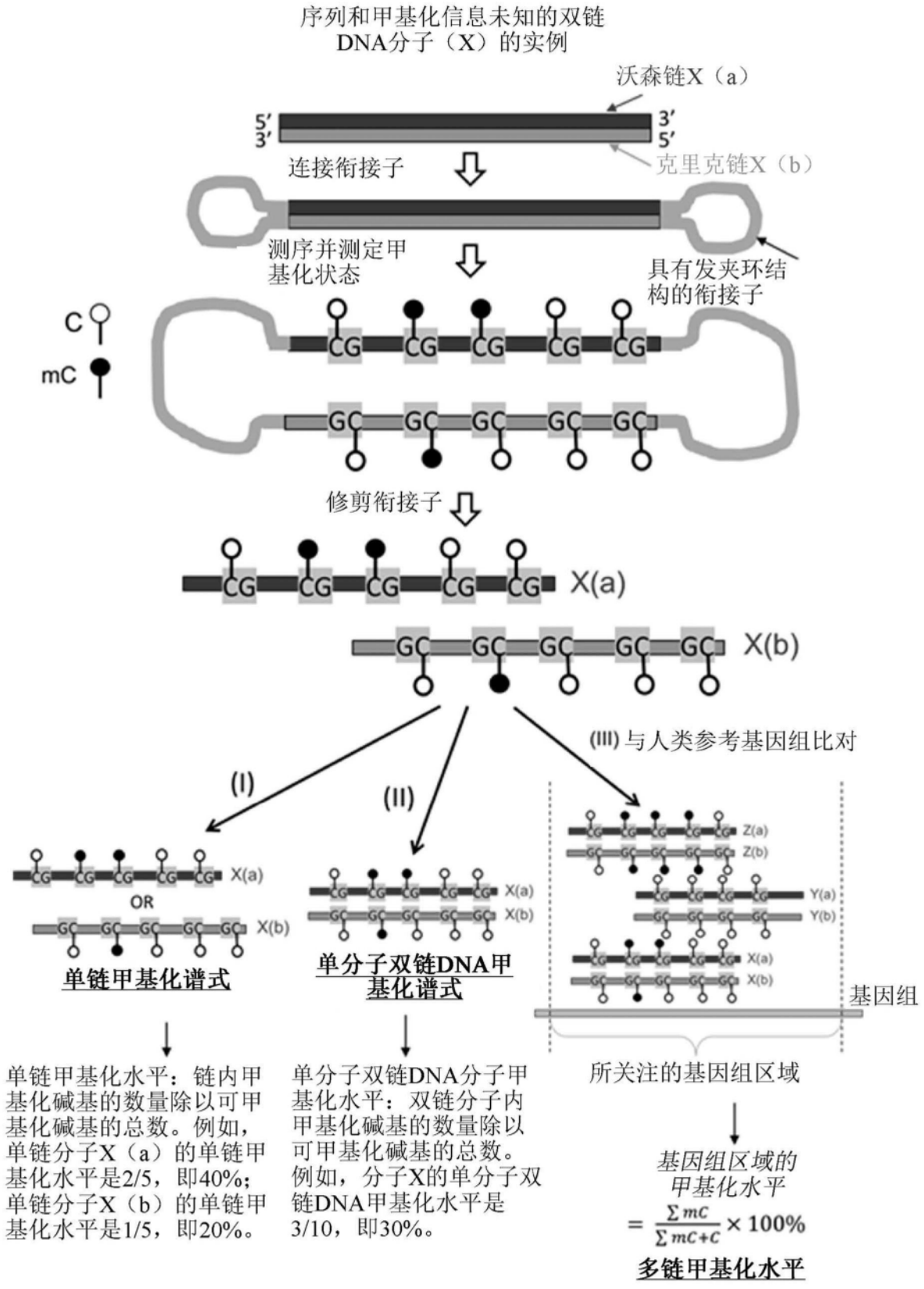

序列和甲基化信息未知的双链
DNA分子（X）的实例
沃森链X（a）
克里克链X（b）
连接衔接子
测序并测定甲
基化状态
具有发夹环结
构的衔接子
C
mC
修剪衔接子
X(a)
X(b)
(I)
(II)
(III) 与人类参考基因组比对
OR
单链甲基化谱式
单分子双链DNA甲
基化谱式
Z(a)
Z(b)
Y(a)
Y(b)
基因组
所关注的基因组区域
单链甲基化水平：链内甲基化碱基的数量除以可甲基化碱基的总数。例如，单链分子X（a）的单链甲基化水平是2/5，即40%；单链分子X（b）的单链甲基化水平是1/5，即20%。
单分子双链DNA分子甲基化水平：双链分子内甲基化碱基的数量除以可甲基化碱基的总数。例如，分子X的单分子双链DNA甲基化水平是3/10，即30%。
基因组区域的
甲基化水平
$= \frac{\sum mC}{\sum mC+C} \times 100\%$
多链甲基化水平

图61

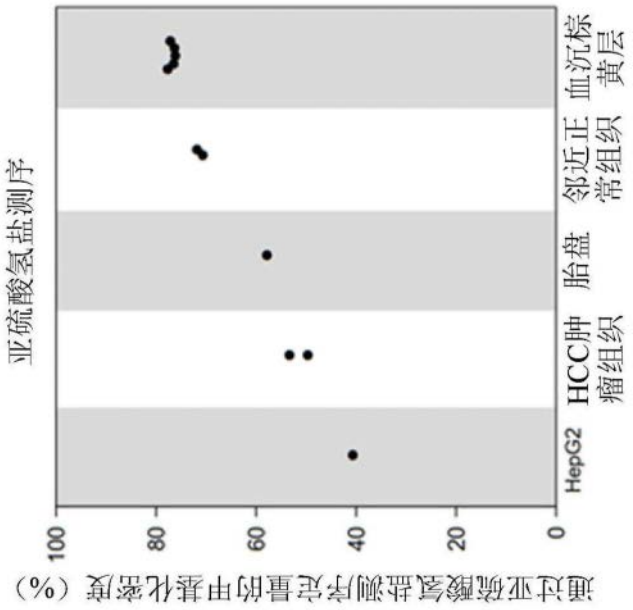

亚硫酸氢盐测序
通过亚硫酸氢盐测序定量的甲基化密度（%）
100
80
60
40
20
0
HepG2
HCC肿瘤组织
胎盘
邻近正常组织
血沉棕黄层

图62A

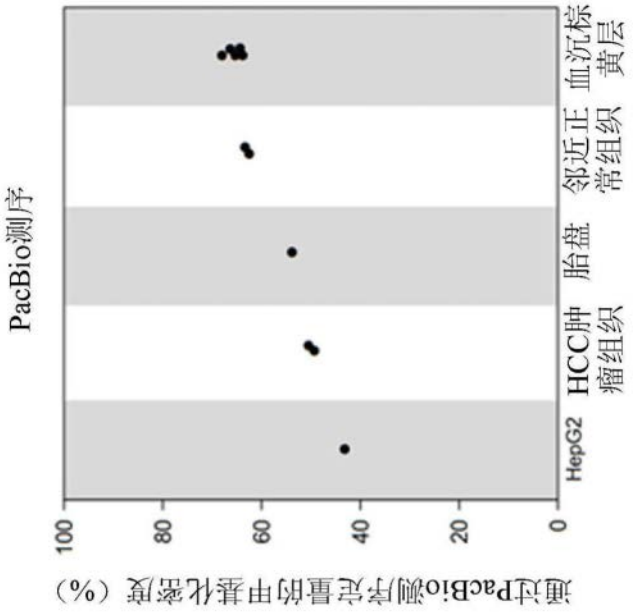

PacBio测序
通过PacBio测序定量的甲基化密度（%）
100
80
60
40
20
0
HepG2
HCC肿瘤组织
胎盘
邻近正常组织
血沉棕黄层

图62B

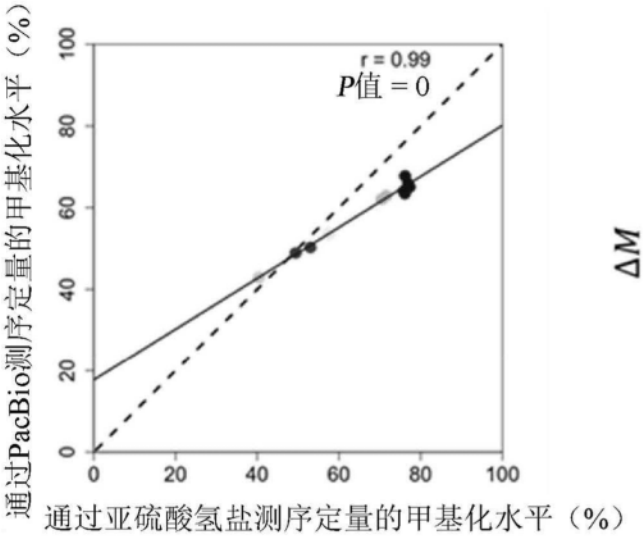


图63A

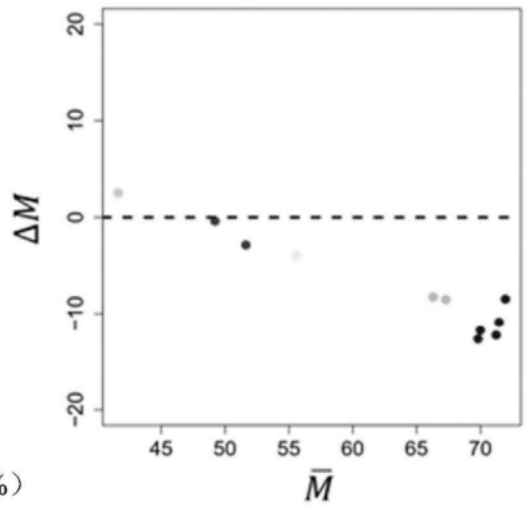


图63B

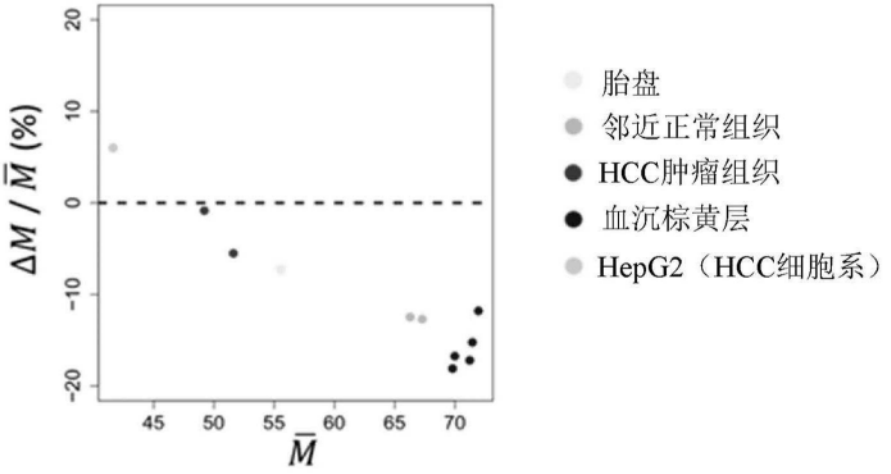


图63C

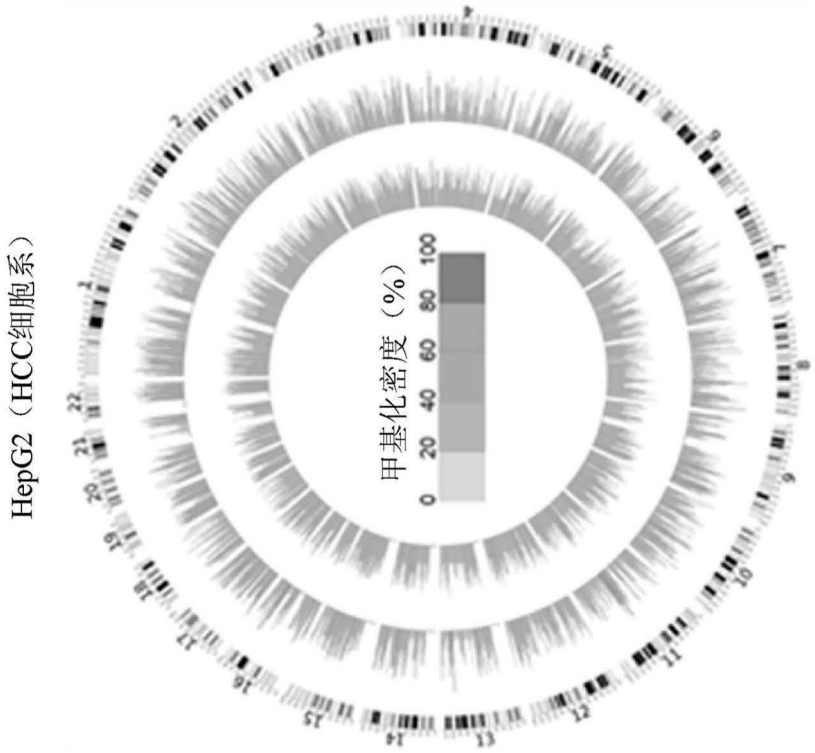

HepG2（HCC细胞系）
甲基化密度（%）
0
20
40
60
80
100

图64A

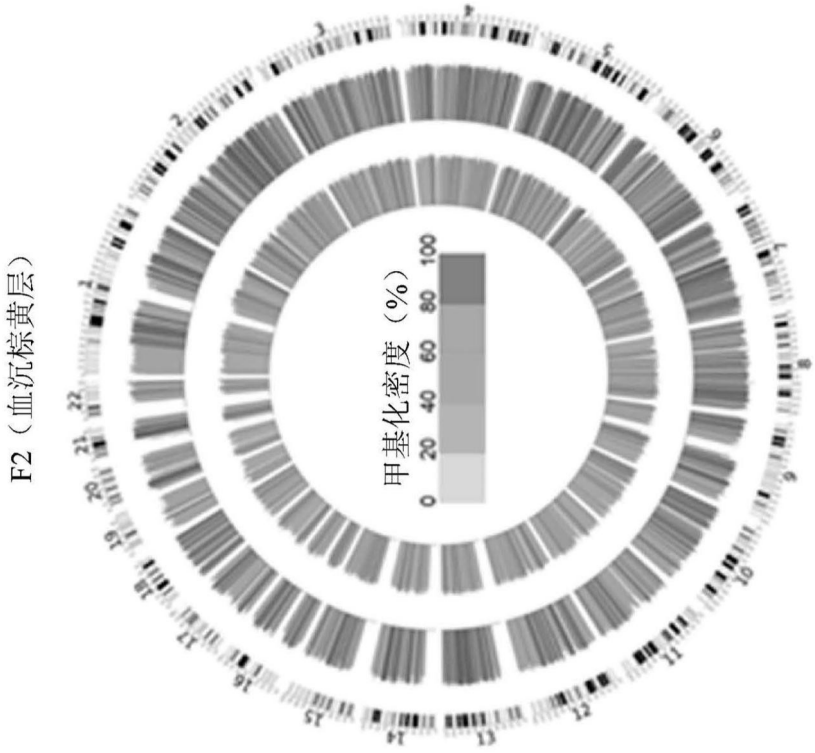

F2（血沉棕黄层）
甲基化密度（%）
0
20
40
60
80
100

图64B

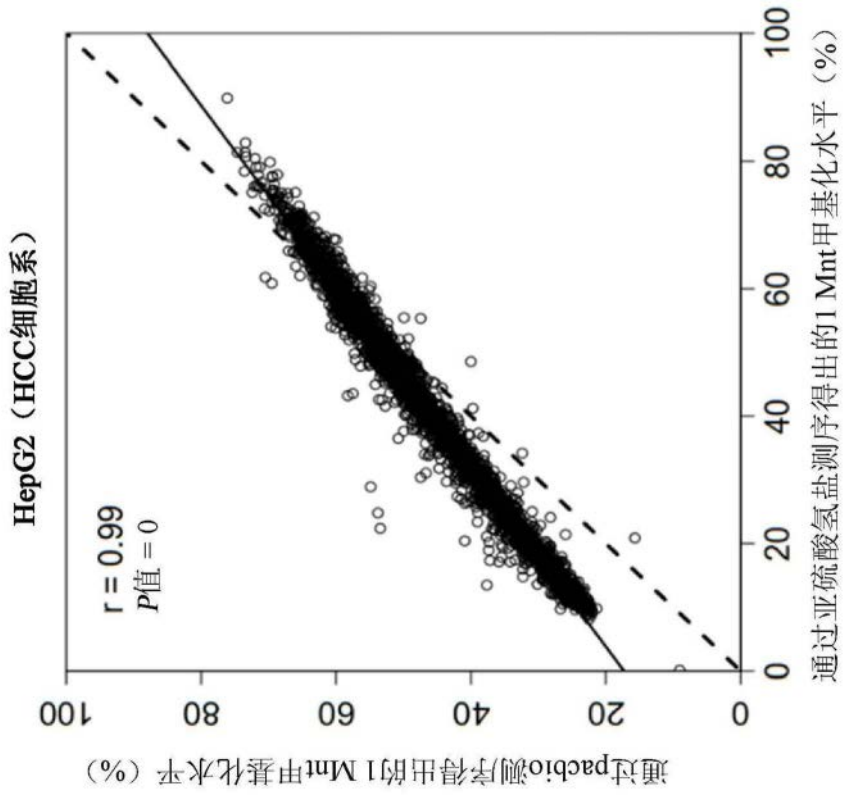

HepG2（HCC细胞系）
r = 0.99
P值 = 0
通过pacbio测序得出的1 Mnt甲基化水平（%）
通过亚硫酸氢盐测序得出的1 Mnt甲基化水平（%）
0
20
40
60
80
100

图65A

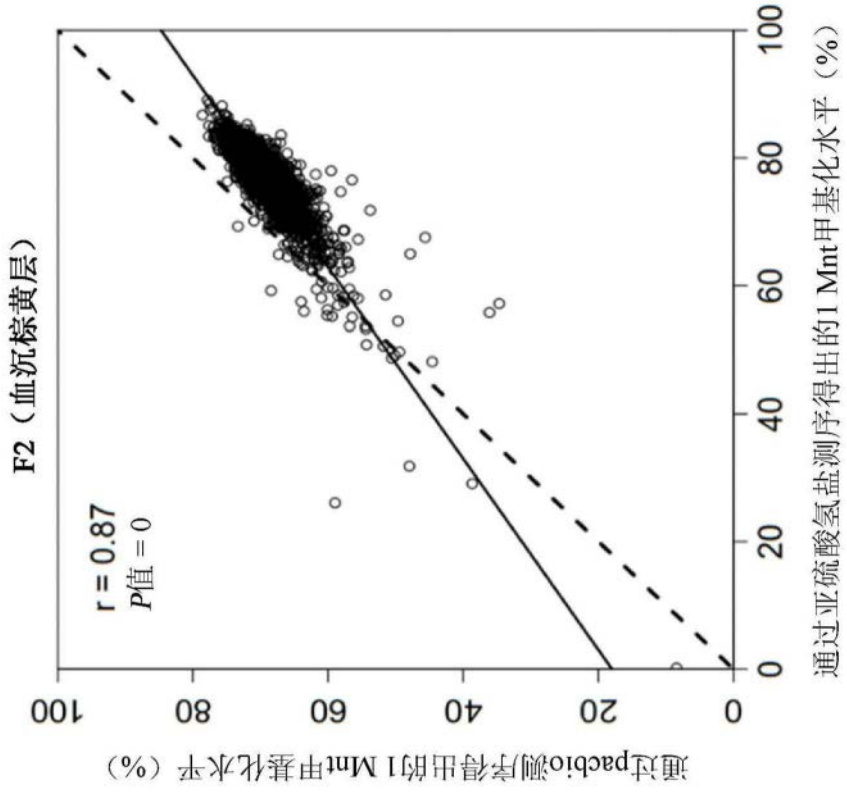

F2（血沉棕黄层）
r = 0.87
P值 = 0
通过pacbio测序得出的1 Mnt甲基化水平（%）
通过亚硫酸氢盐测序得出的1 Mnt甲基化水平（%）
0
20
40
60
80
100

图65B

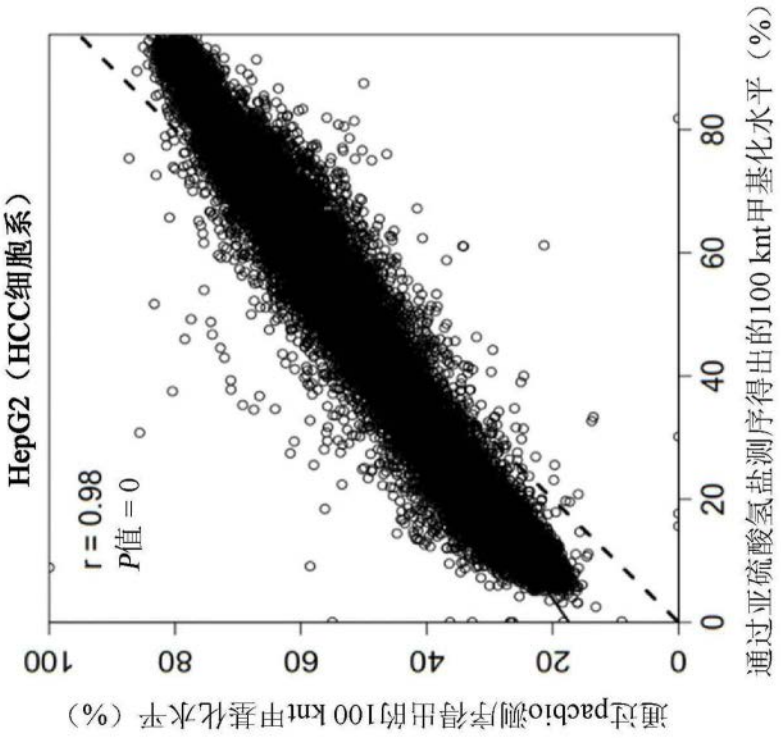

HepG2（HCC细胞系）
r = 0.98
P值 = 0
通过pacbio测序得出的100 kmt甲基化水平（%）
通过亚硫酸氢盐测序得出的100 kmt甲基化水平（%）
0
20
40
60
80
100

图66A

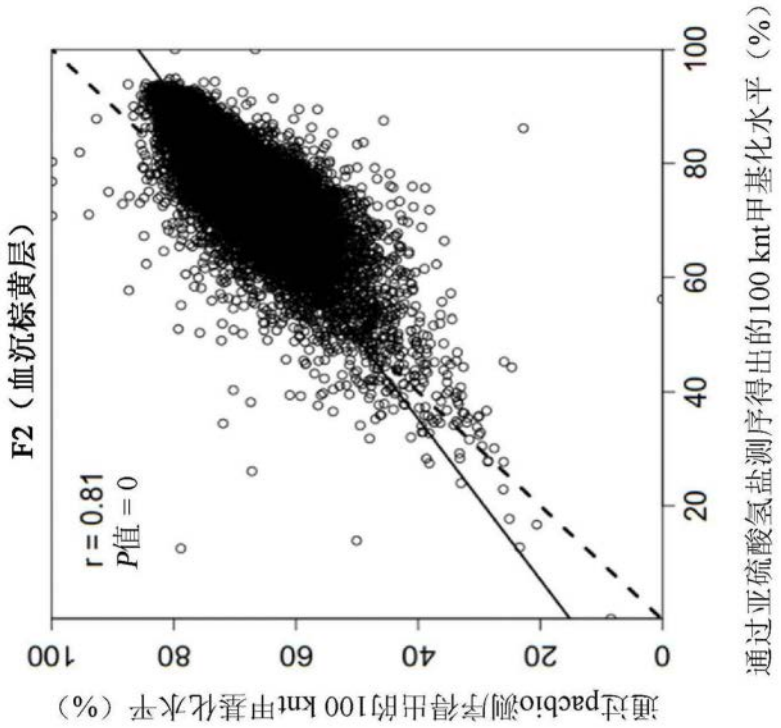

F2（血沉棕黄层）
r = 0.81
P值 = 0
通过pacbio测序得出的100 kmt甲基化水平（%）
通过亚硫酸氢盐测序得出的100 kmt甲基化水平（%）
0
20
40
60
80
100

图66B

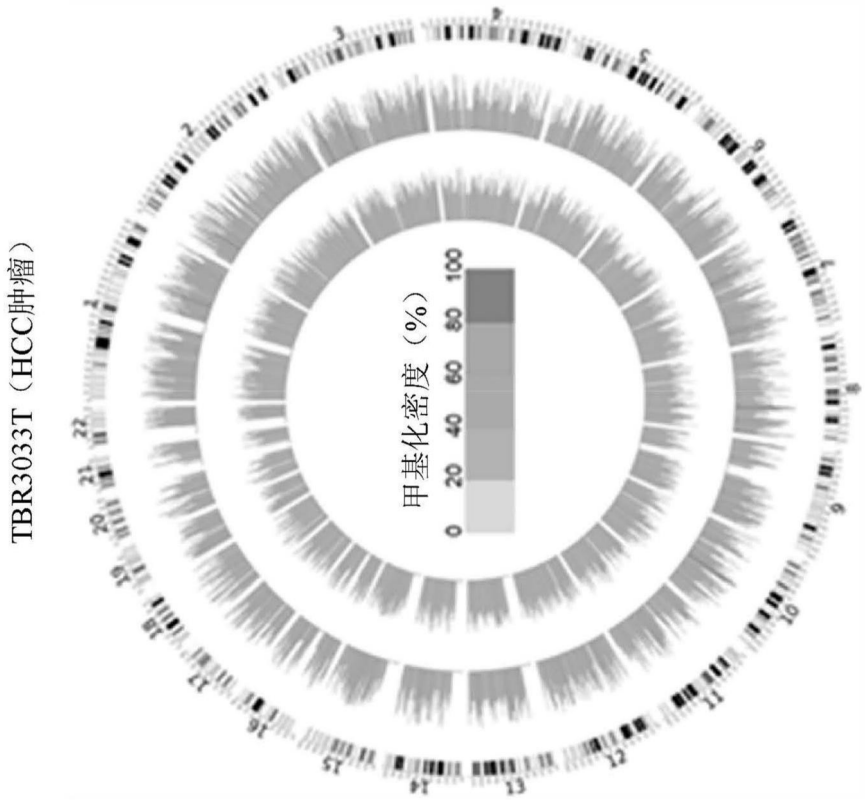

TBR3033T（HCC肿瘤）
甲基化密度（%）
0
20
40
60
80
100

图67A

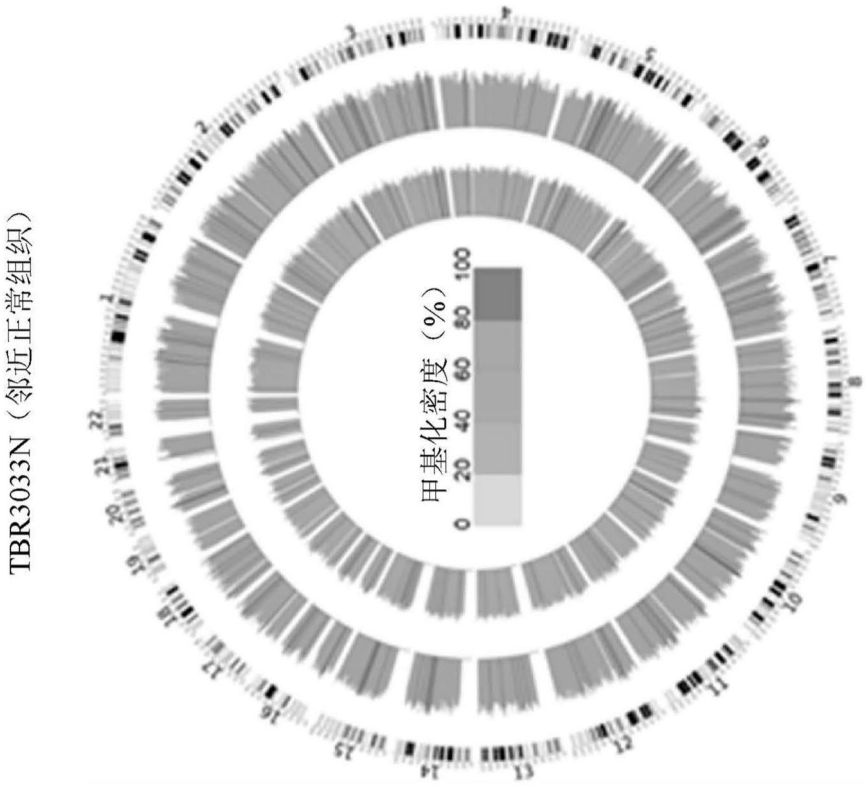

TBR3033N（邻近正常组织）
甲基化密度（%）
0
20
40
60
80
100

图67B

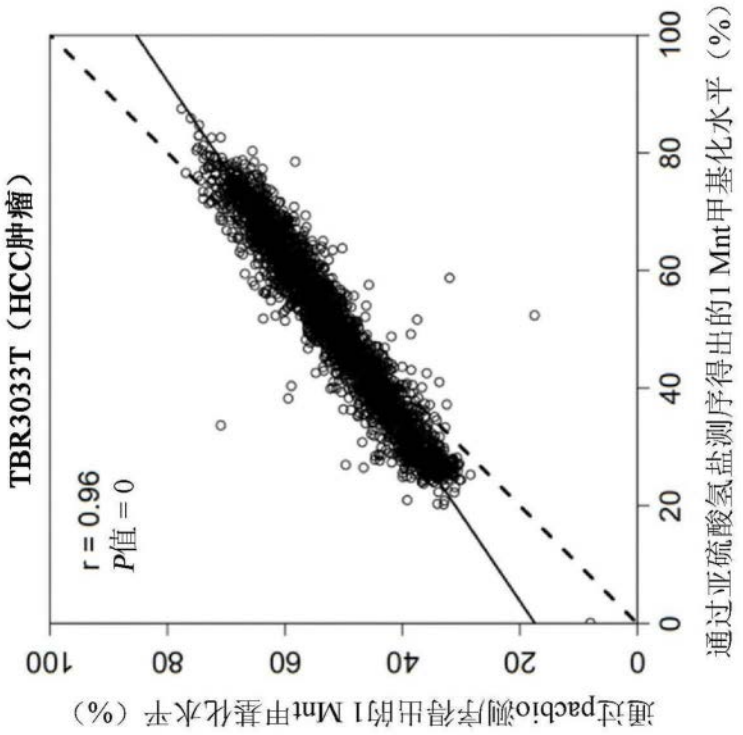

TBR3033T（HCC肿瘤）
r = 0.96
P值 = 0
通过亚硫酸氢盐测序得出的1 Mnt甲基化水平（%）
通过pacbio测序得出的1 Mnt甲基化水平（%）
0
20
40
60
80
100

图68A

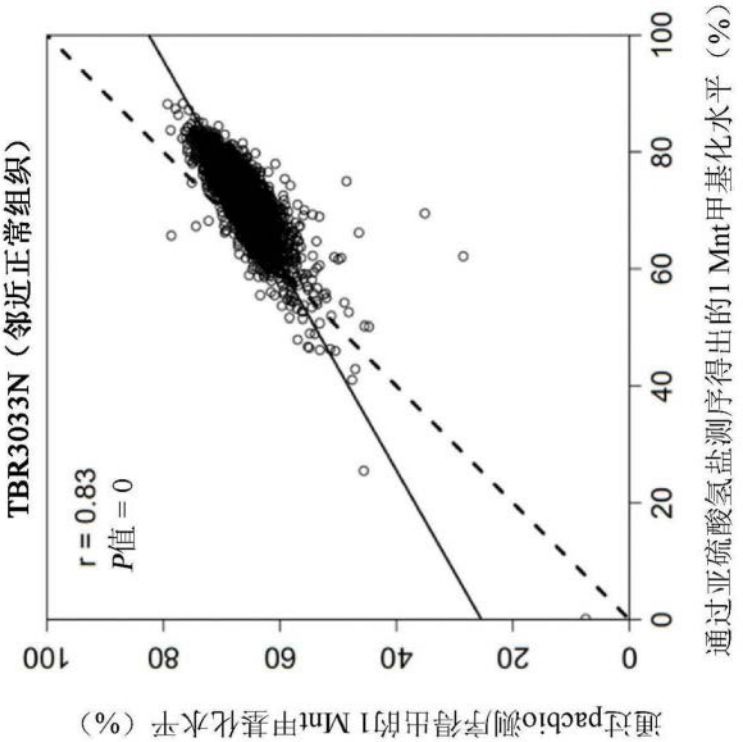

TBR3033N（邻近正常组织）
r = 0.83
P值 = 0
通过亚硫酸氢盐测序得出的1 Mnt甲基化水平（%）
通过pacbio测序得出的1 Mnt甲基化水平（%）
0
20
40
60
80
100

图68B

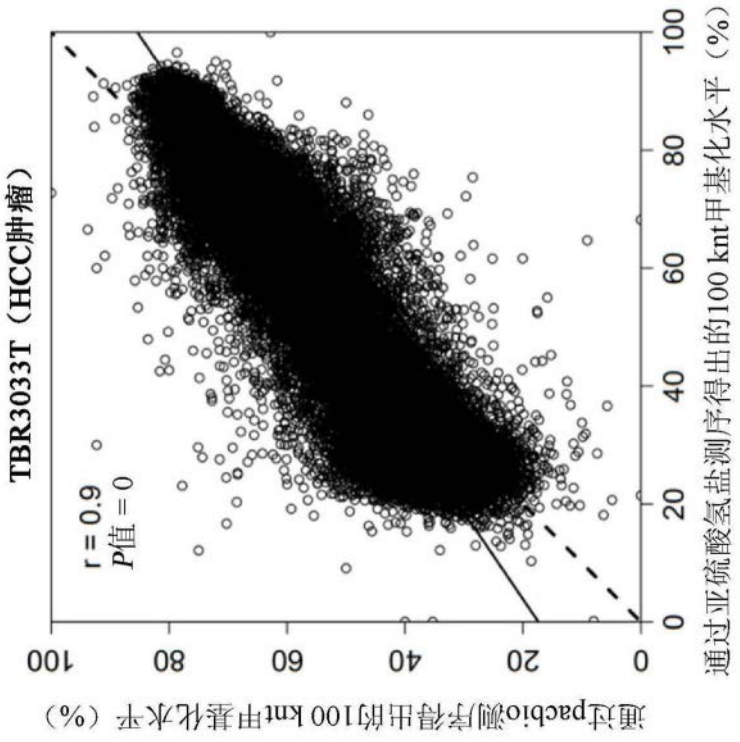

TBR3033T（HCC肿瘤）
r = 0.9
P值 = 0
通过pacbio测序得出的100 kmt甲基化水平（%）
通过亚硫酸氢盐测序得出的100 kmt甲基化水平（%）
0
20
40
60
80
100

图69A

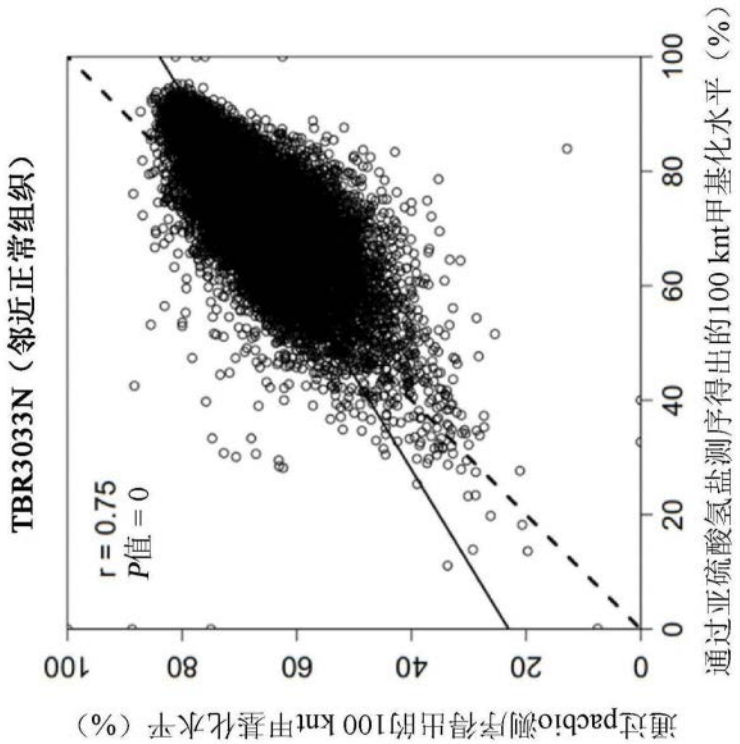

TBR3033N（邻近正常组织）
r = 0.75
P值 = 0
通过pacbio测序得出的100 kmt甲基化水平（%）
通过亚硫酸氢盐测序得出的100 kmt甲基化水平（%）
0
20
40
60
80
100

图69B

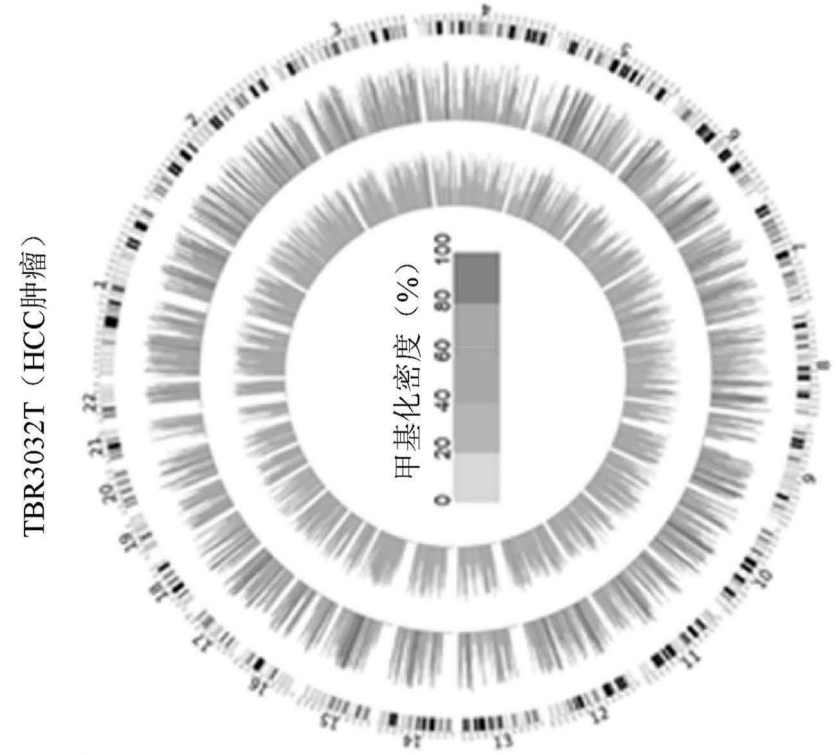

TBR3032T（HCC肿瘤）
甲基化密度（%）
0
20
40
60
80
100

图70A

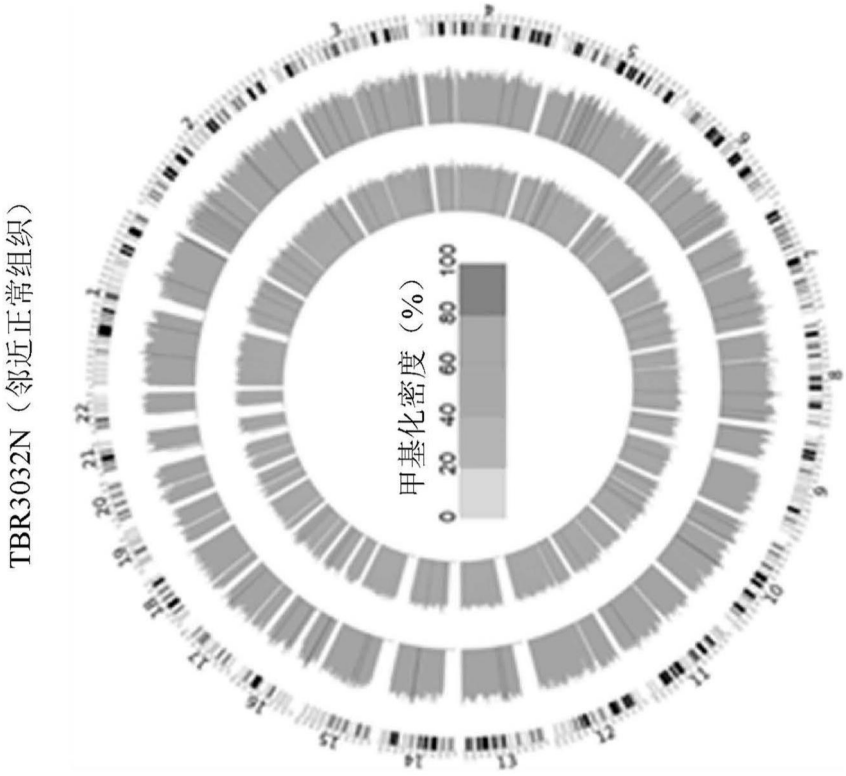

TBR3032N（邻近正常组织）
甲基化密度（%）
0
20
40
60
80
100

图70B

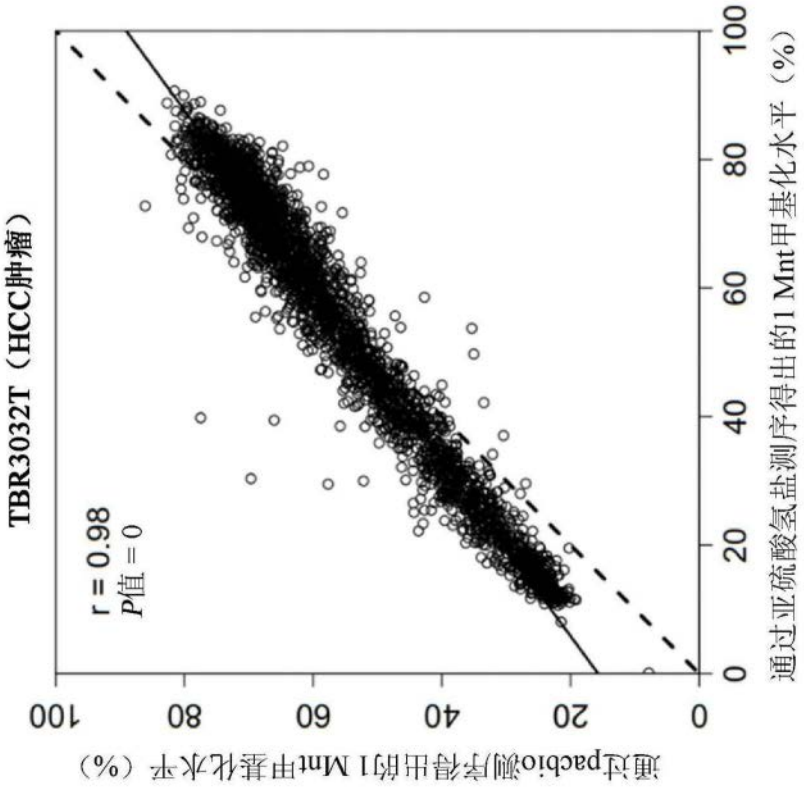

TBR3032T（HCC肿瘤）
r = 0.98
P值 = 0
通过pacbio测序得出的1 Mnt甲基化水平（%）
通过亚硫酸氢盐测序得出的1 Mnt甲基化水平（%）
0
20
40
60
80
100

图71A

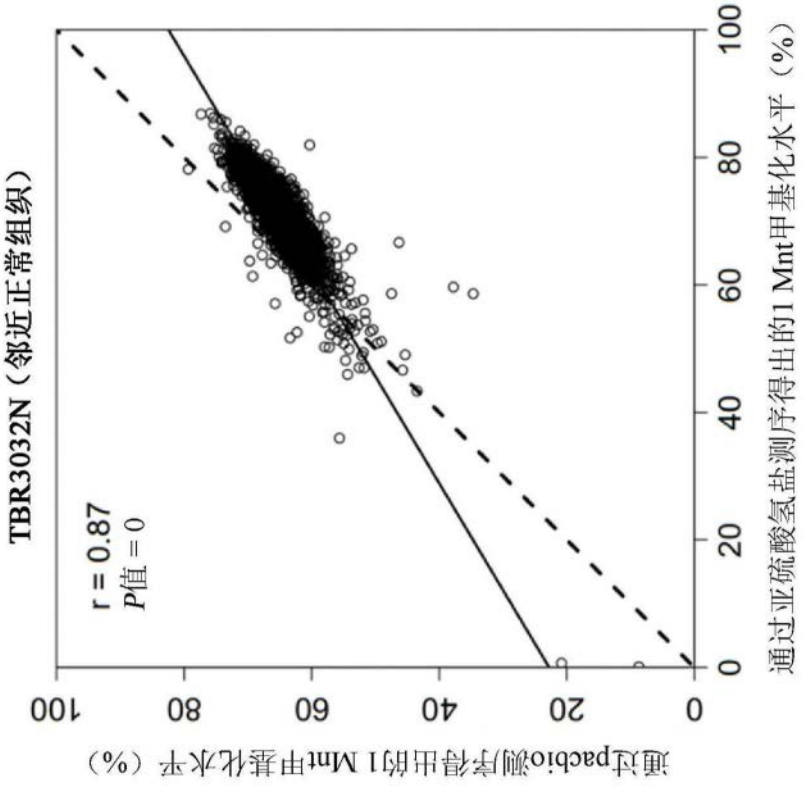

TBR3032N（邻近正常组织）
r = 0.87
P值 = 0
通过pacbio测序得出的1 Mnt甲基化水平（%）
通过亚硫酸氢盐测序得出的1 Mnt甲基化水平（%）
0
20
40
60
80
100

图71B

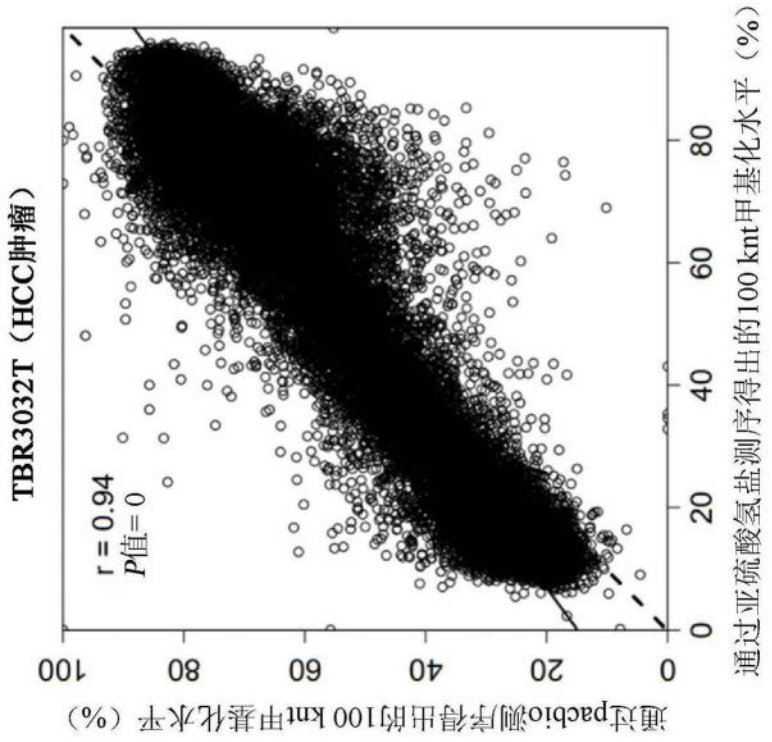

TBR3032T（HCC肿瘤）
r = 0.94
P值= 0
通过pacbio测序得出的100 knt甲基化水平（%）
通过亚硫酸氢盐测序得出的100 knt甲基化水平（%）
0
20
40
60
80
100

图72A

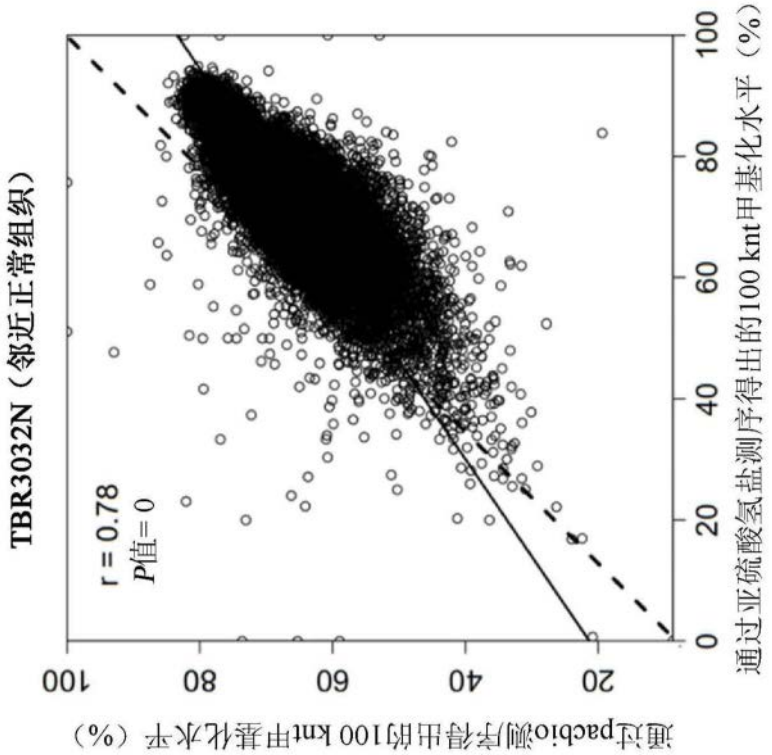

TBR3032N（邻近正常组织）
r = 0.78
P值= 0
通过pacbio测序得出的100 knt甲基化水平（%）
通过亚硫酸氢盐测序得出的100 knt甲基化水平（%）
0
20
40
60
80
100

图72B

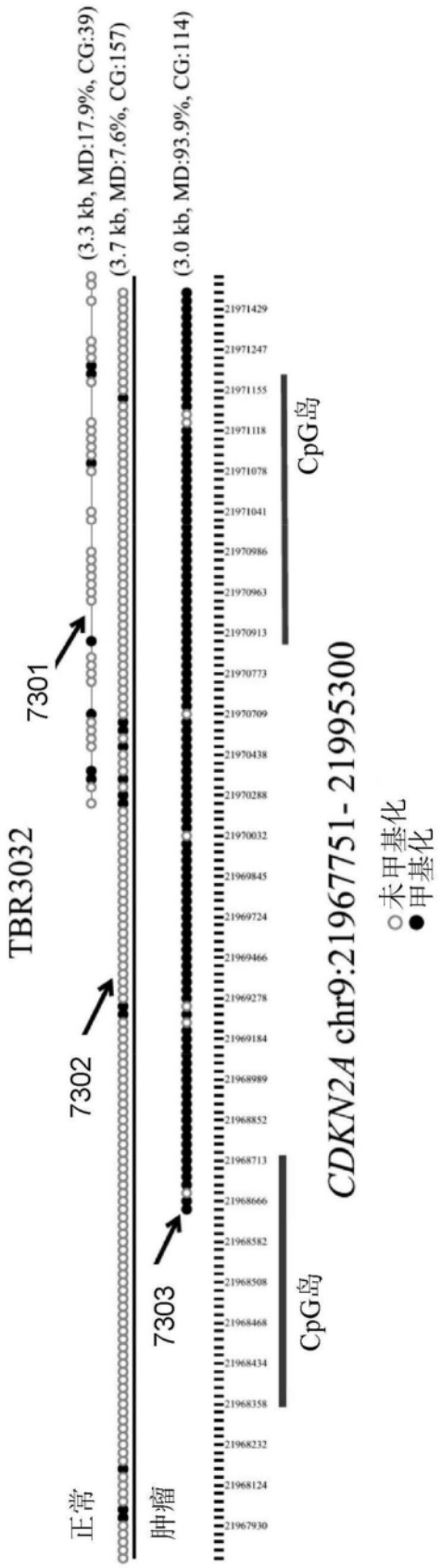

TBR3032
7301
7302
7303
(3.3 kb, MD:17.9%, CG:39)
(3.7 kb, MD:7.6%, CG:157)
(3.0 kb, MD:93.9%, CG:114)
正常
肿瘤
CpG岛
CpG岛
CDKN2A chr9:21967751- 21995300
未甲基化
甲基化

图73

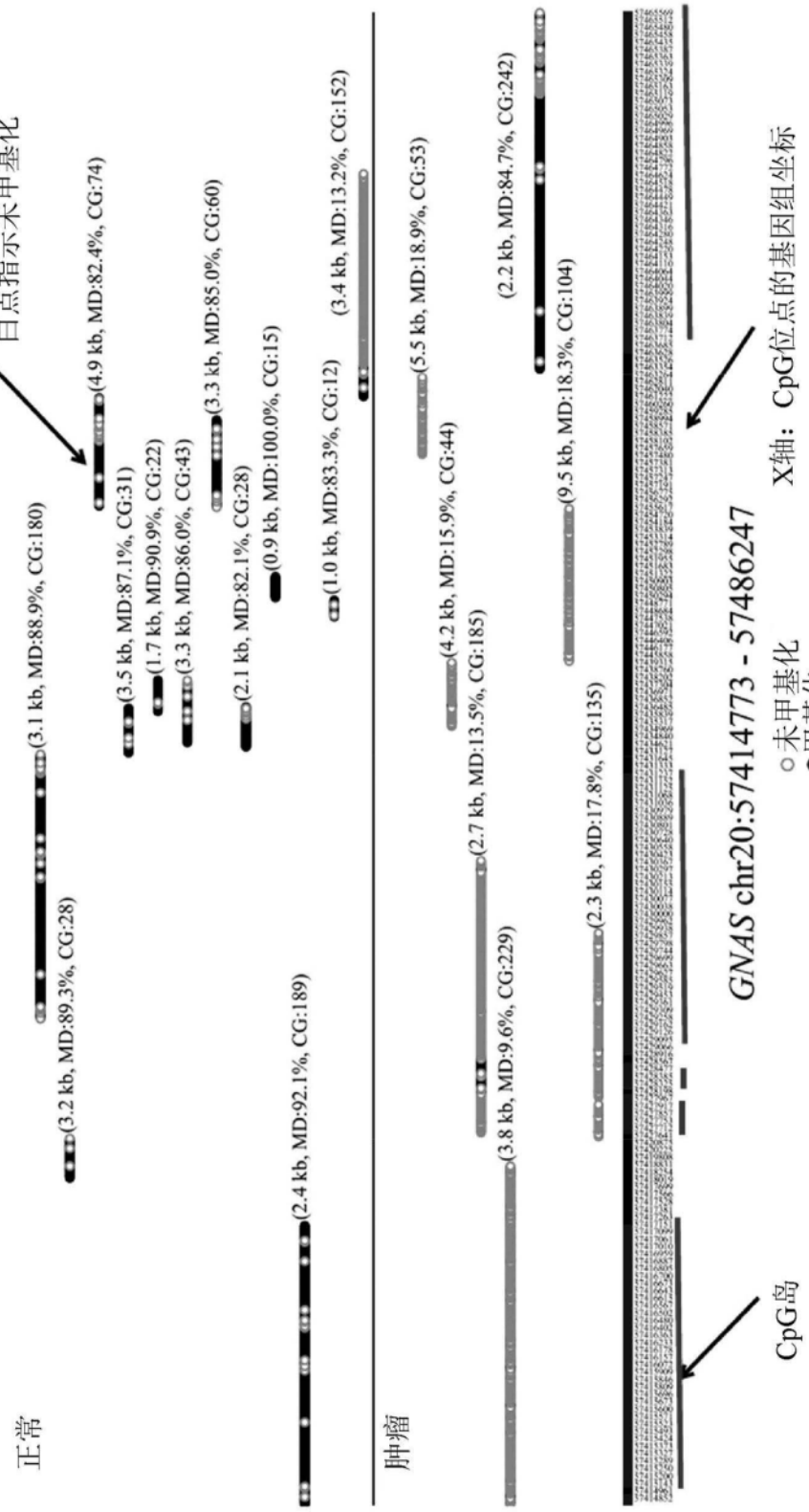

DNA片段
黑点指示甲基化
白点指示未甲基化
正常
肿瘤
(3.1 kb, MD:88.9%, CG:180)
(3.2 kb, MD:89.3%, CG:28)
(2.4 kb, MD:92.1%, CG:189)
(4.9 kb, MD:82.4%, CG:74)
(3.5 kb, MD:87.1%, CG:31)
(1.7 kb, MD:90.9%, CG:22)
(3.3 kb, MD:86.0%, CG:43)
(2.1 kb, MD:82.1%, CG:28)
(0.9 kb, MD:100.0%, CG:15)
(1.0 kb, MD:83.3%, CG:12)
(3.4 kb, MD:13.2%, CG:152)
(5.5 kb, MD:18.9%, CG:53)
(4.2 kb, MD:15.9%, CG:44)
(2.7 kb, MD:13.5%, CG:185)
(3.8 kb, MD:9.6%, CG:229)
(2.3 kb, MD:17.8%, CG:135)
(2.2 kb, MD:84.7%, CG:242)
(9.5 kb, MD:18.3%, CG:104)
GNAS chr20:57414773 - 57486247
○未甲基化
●甲基化
X轴： CpG位点的基因组坐标
CpG岛

图74A

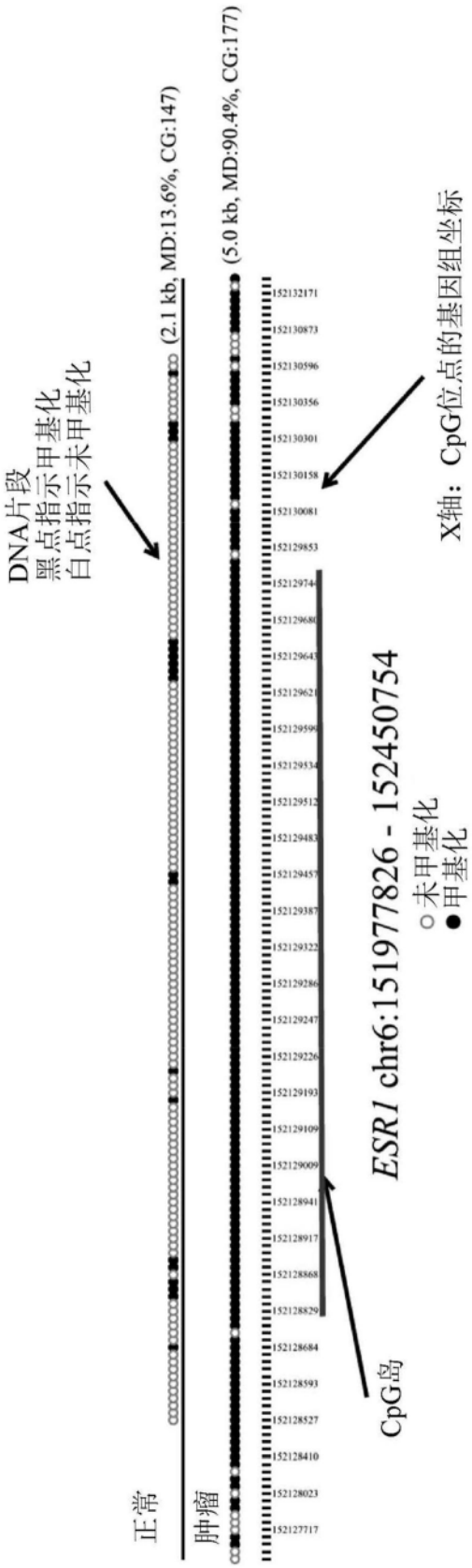

DNA片段
黑点指示甲基化
白点指示未甲基化
(2.1 kb, MD:13.6%, CG:147)
(5.0 kb, MD:90.4%, CG:177)
X轴：CpG位点的基因组坐标
ESR1 chr6:151977826 - 152450754
○未甲基化
●甲基化
CpG岛
正常
肿瘤

图74B

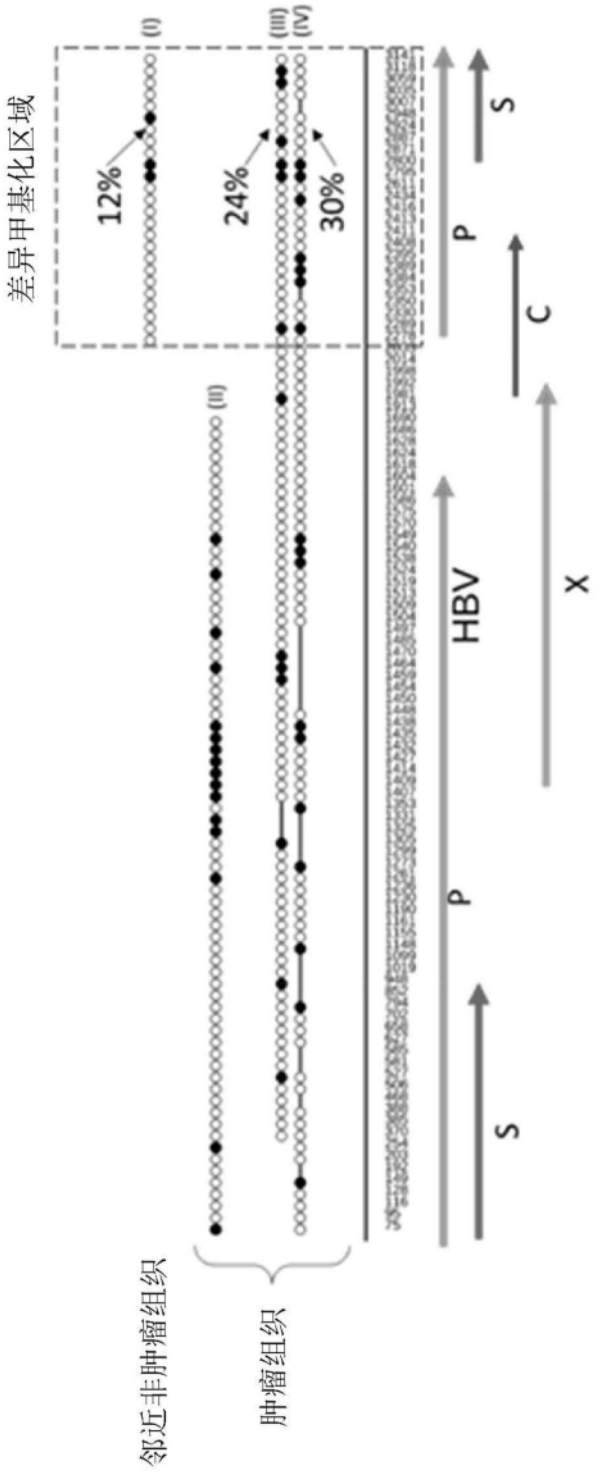


图75

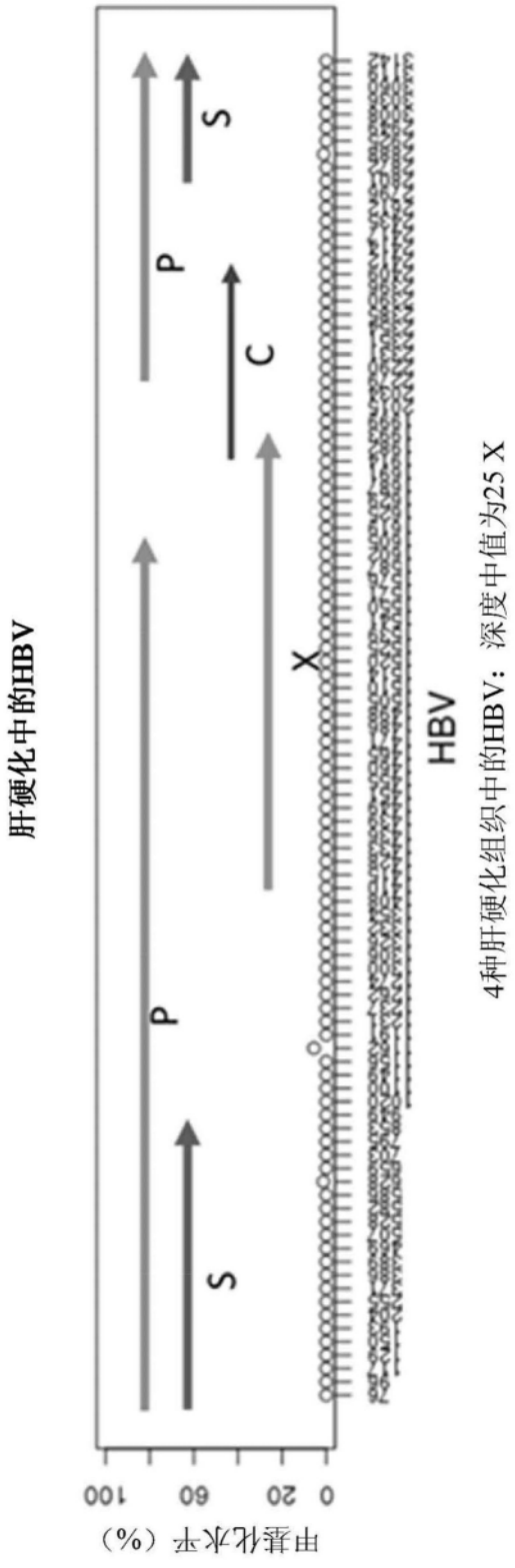

肝硬化中的HBV
HBV
4种肝硬化组织中的HBV：深度中值为25 X
S
P
X
C
P
S
0
20
60
100
甲基化水平（%）

图76A

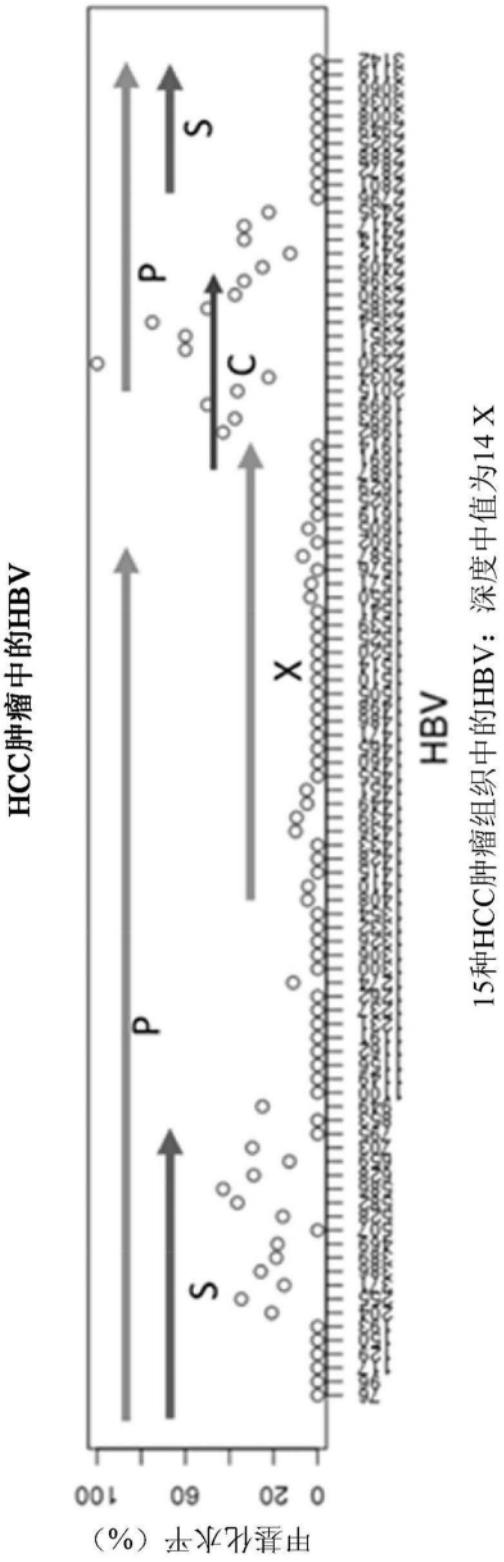

HCC肿瘤中的HBV
HBV
15种HCC肿瘤组织中的HBV：深度中值为14 X
甲基化水平（%）
0
20
60
100
S
P
X
C
P
S

图76B

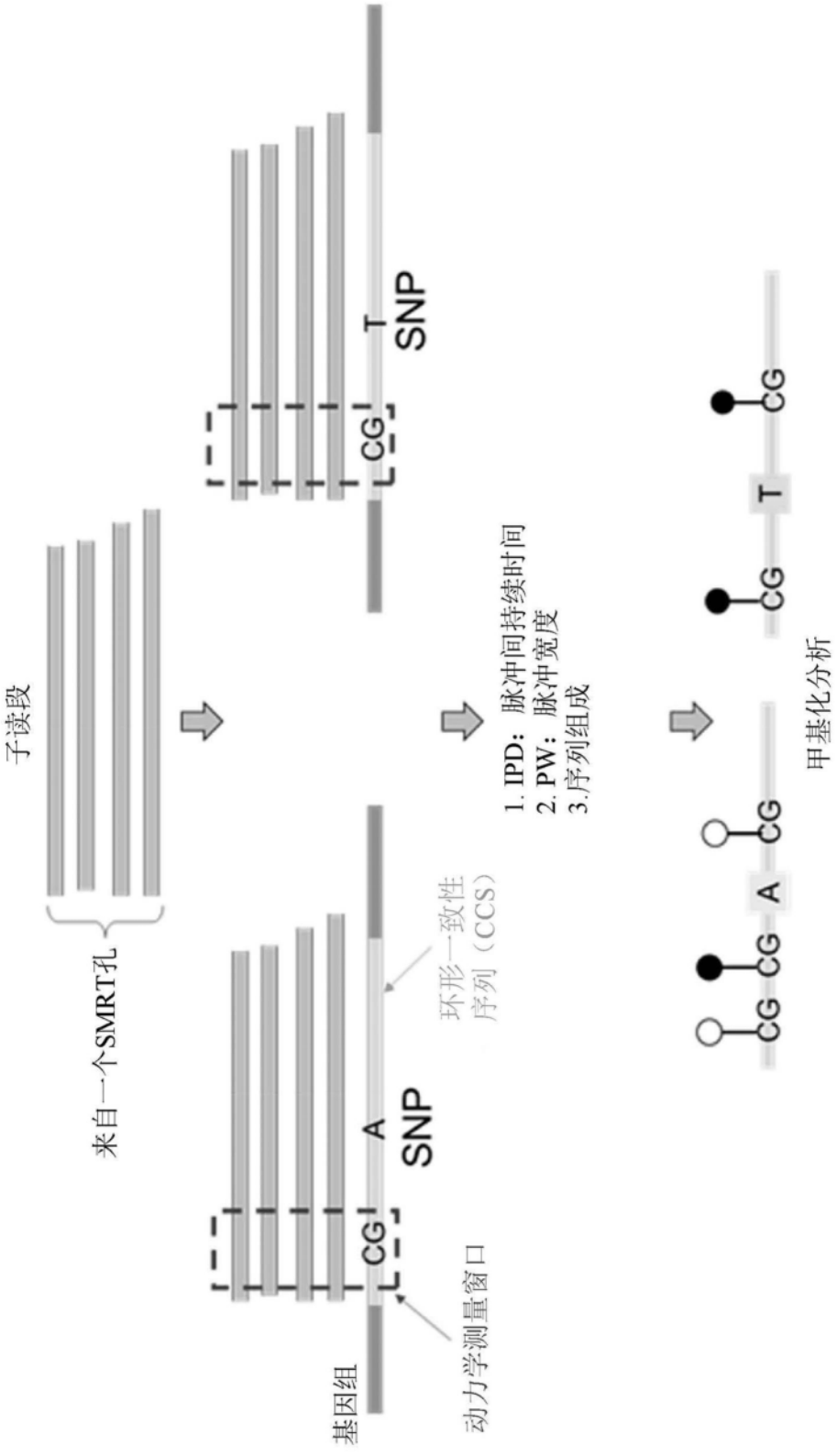

子读段
来自一个SMRT孔
基因组
CG
A
SNP
环形一致性序列（CCS）
动力学测量窗口
1. IPD：脉冲间持续时间
2. PW：脉冲宽度
3.序列组成
CG
T
SNP
CG
A
CG
CG
CG
T
CG
甲基化分析

图77

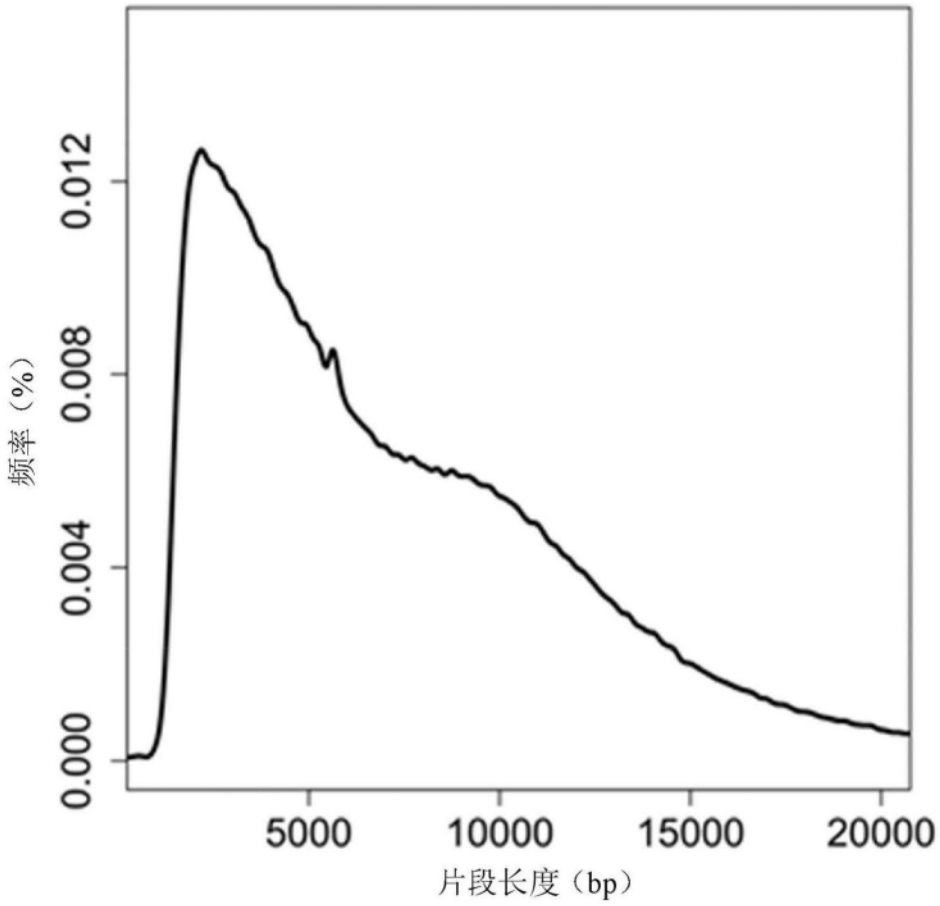


图78

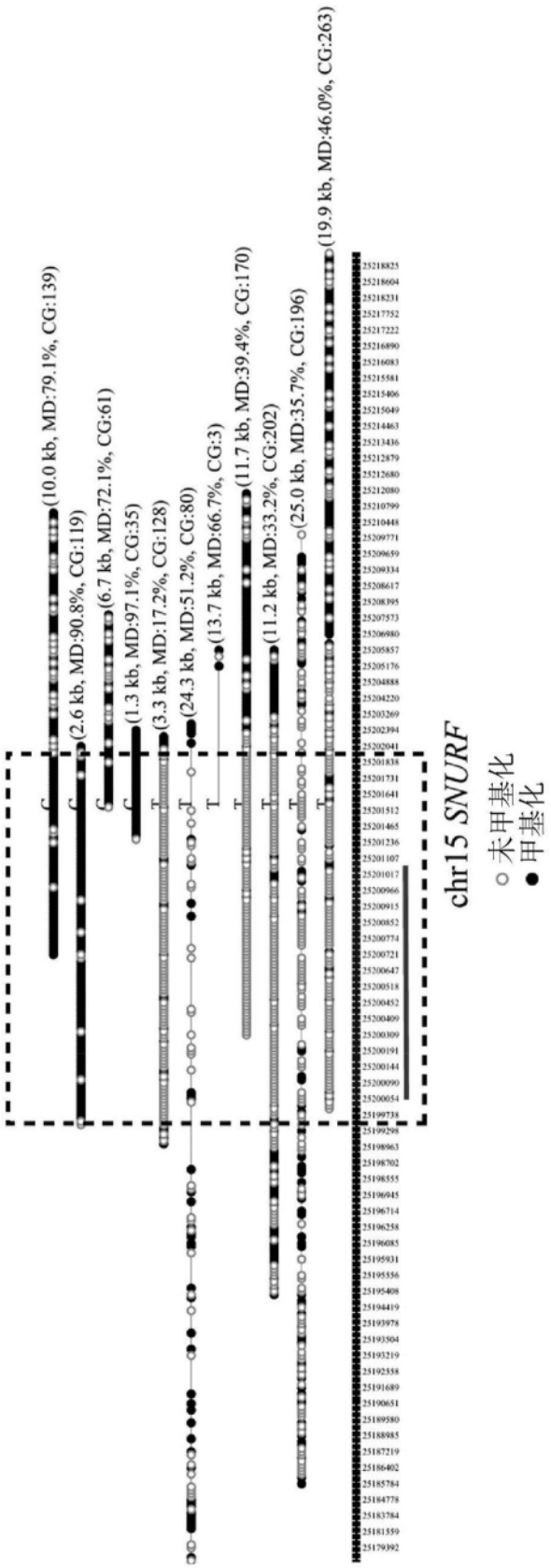

(10.0 kb, MD:79.1%, CG:139)
(2.6 kb, MD:90.8%, CG:119)
(6.7 kb, MD:72.1%, CG:61)
(1.3 kb, MD:97.1%, CG:35)
(3.3 kb, MD:17.2%, CG:128)
(24.3 kb, MD:51.2%, CG:80)
(13.7 kb, MD:66.7%, CG:3)
(11.7 kb, MD:39.4%, CG:170)
(11.2 kb, MD:33.2%, CG:202)
(25.0 kb, MD:35.7%, CG:196)
(19.9 kb, MD:46.0%, CG:263)
chr15 SNURF
未甲基化
甲基化

图79A

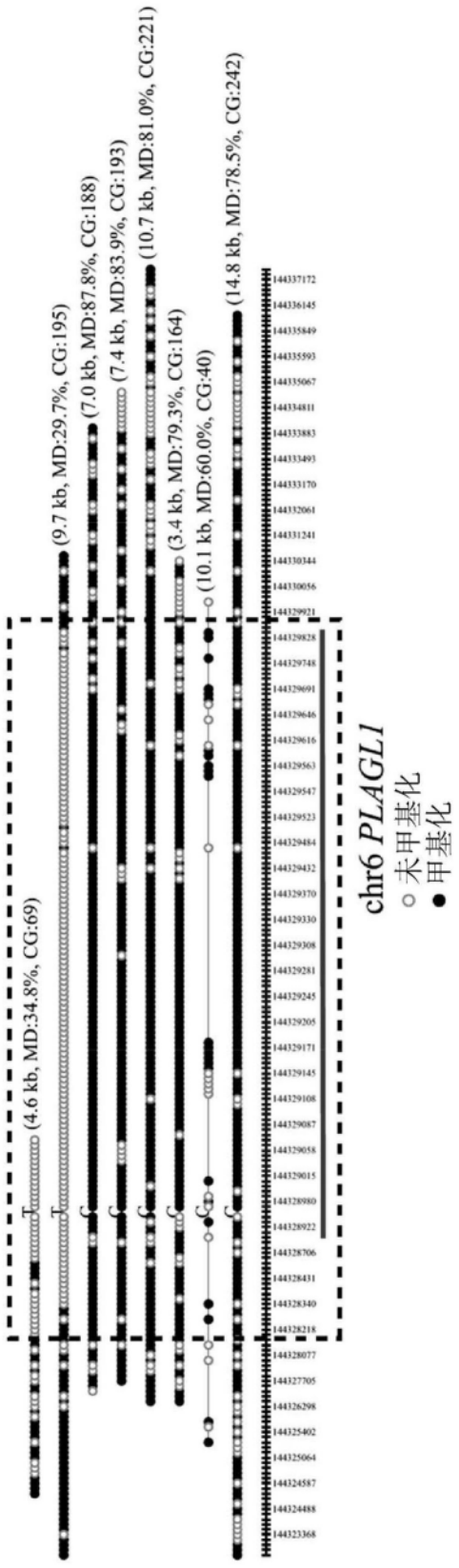

(4.6 kb, MD:34.8%, CG:69)
(9.7 kb, MD:29.7%, CG:195)
(7.0 kb, MD:87.8%, CG:188)
(7.4 kb, MD:83.9%, CG:193)
(10.7 kb, MD:81.0%, CG:221)
(3.4 kb, MD:79.3%, CG:164)
(10.1 kb, MD:60.0%, CG:40)
(14.8 kb, MD:78.5%, CG:242)
chr6 PLAGL1
未甲基化
甲基化

图79B

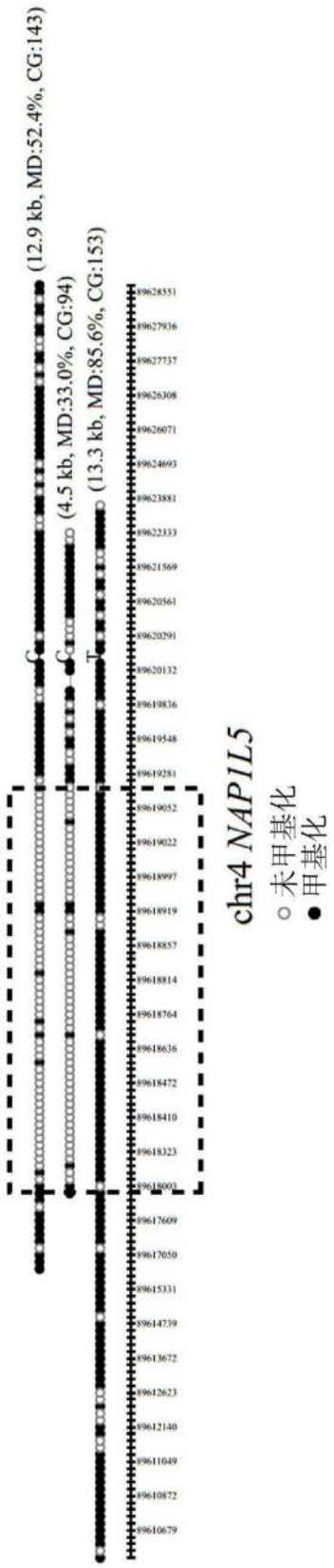

(12.9 kb, MD:52.4%, CG:143)
(4.5 kb, MD:33.0%, CG:94)
(13.3 kb, MD:85.6%, CG:153)
C
C
T
chr4 NAP1L5
未甲基化
甲基化

图79C

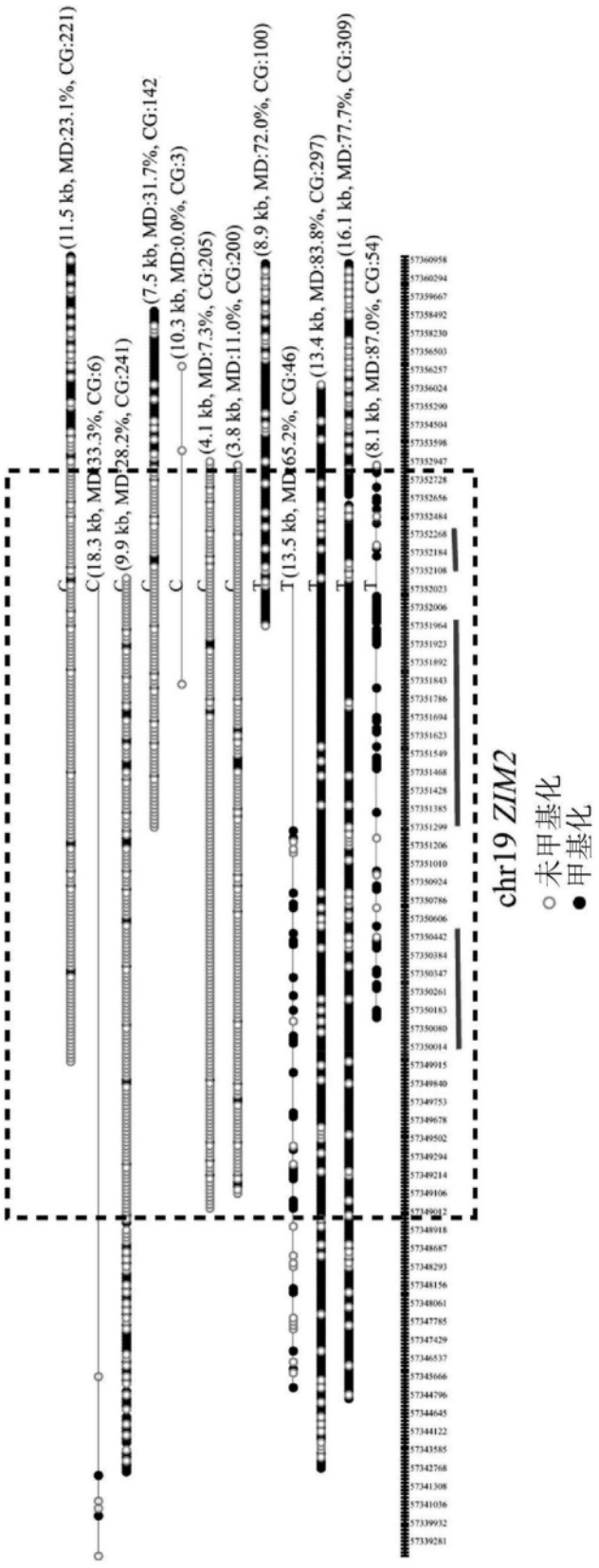

(11.5 kb, MD:23.1%, CG:221)
(18.3 kb, MD:33.3%, CG:6)
(9.9 kb, MD:28.2%, CG:241)
(7.5 kb, MD:31.7%, CG:142
(10.3 kb, MD:0.0%, CG:3)
(4.1 kb, MD:7.3%, CG:205)
(3.8 kb, MD:11.0%, CG:200)
(8.9 kb, MD:72.0%, CG:100)
(13.5 kb, MD:65.2%, CG:46)
(13.4 kb, MD:83.8%, CG:297)
(16.1 kb, MD:77.7%, CG:309)
(8.1 kb, MD:87.0%, CG:54)
chr19 ZIM2
未甲基化
甲基化

图79D

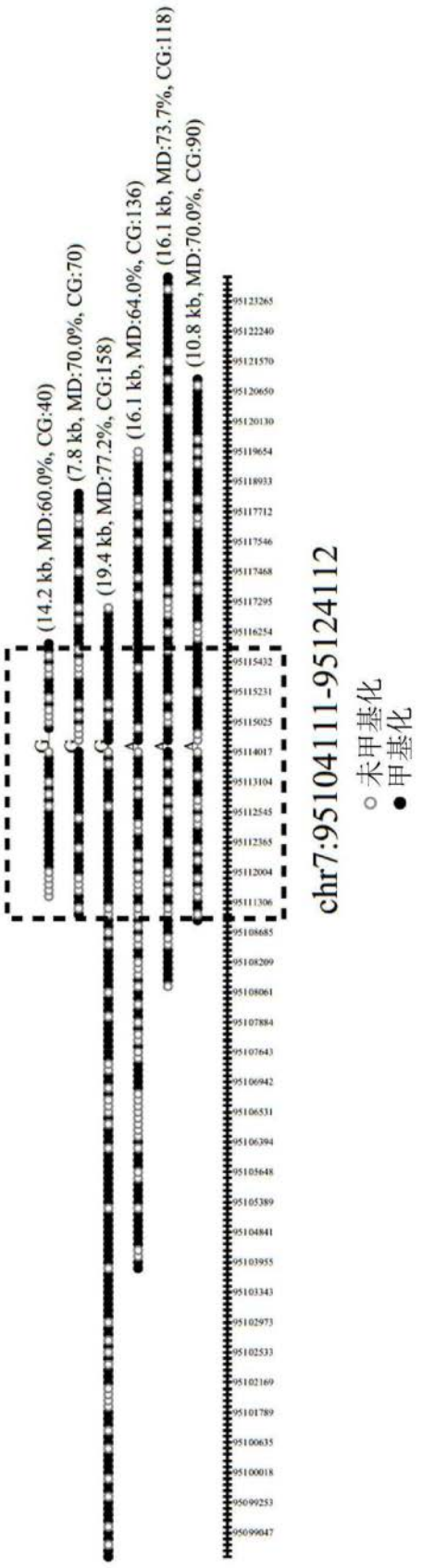

(14.2 kb, MD:60.0%, CG:40)
(7.8 kb, MD:70.0%, CG:70)
(19.4 kb, MD:77.2%, CG:158)
(16.1 kb, MD:64.0%, CG:136)
(16.1 kb, MD:73.7%, CG:118)
(10.8 kb, MD:70.0%, CG:90)
chr7:95104111-95124112
未甲基化
甲基化

图80A

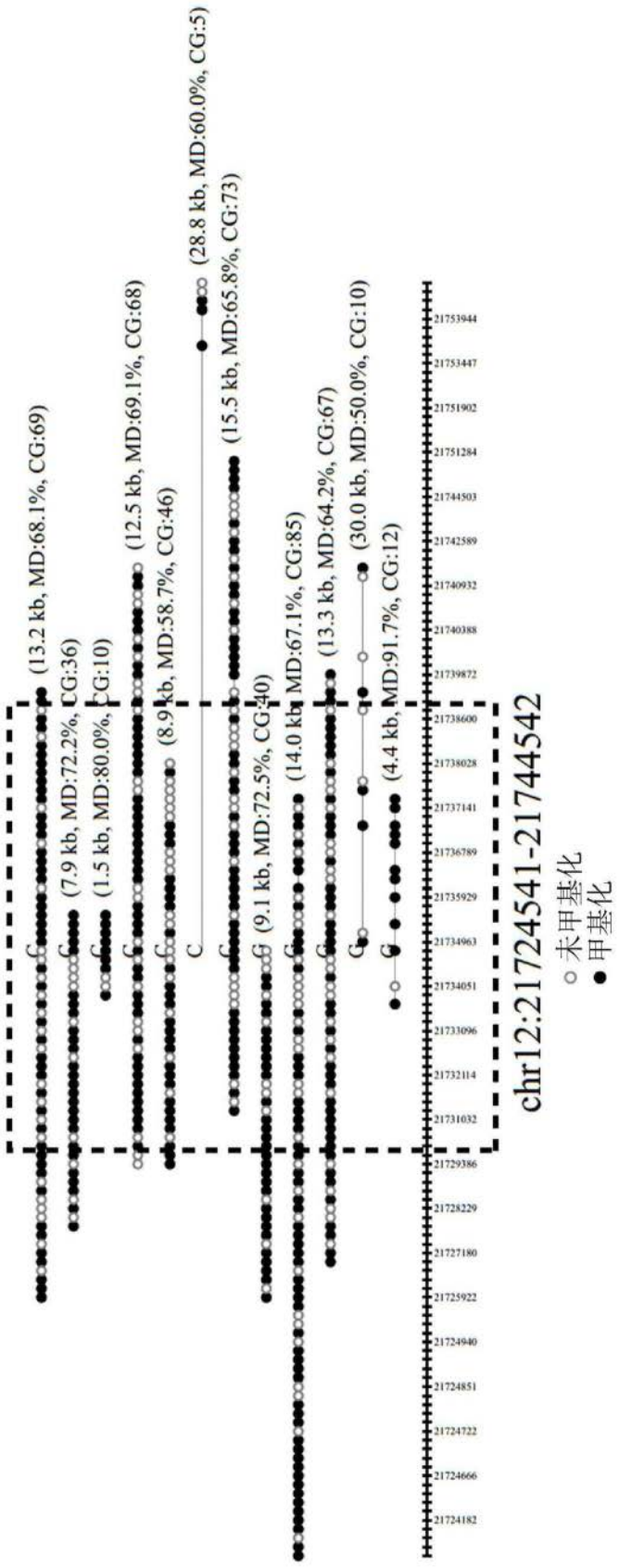

(13.2 kb, MD:68.1%, CG:69)
(7.9 kb, MD:72.2%, CG:36)
(1.5 kb, MD:80.0%, CG:10)
(12.5 kb, MD:69.1%, CG:68)
(8.9 kb, MD:58.7%, CG:46)
(28.8 kb, MD:60.0%, CG:5)
(15.5 kb, MD:65.8%, CG:73)
(9.1 kb, MD:72.5%, CG:40)
(14.0 kb, MD:67.1%, CG:85)
(13.3 kb, MD:64.2%, CG:67)
(30.0 kb, MD:50.0%, CG:10)
(4.4 kb, MD:91.7%, CG:12)
chr12:21724541-21744542
未甲基化
甲基化

图80B

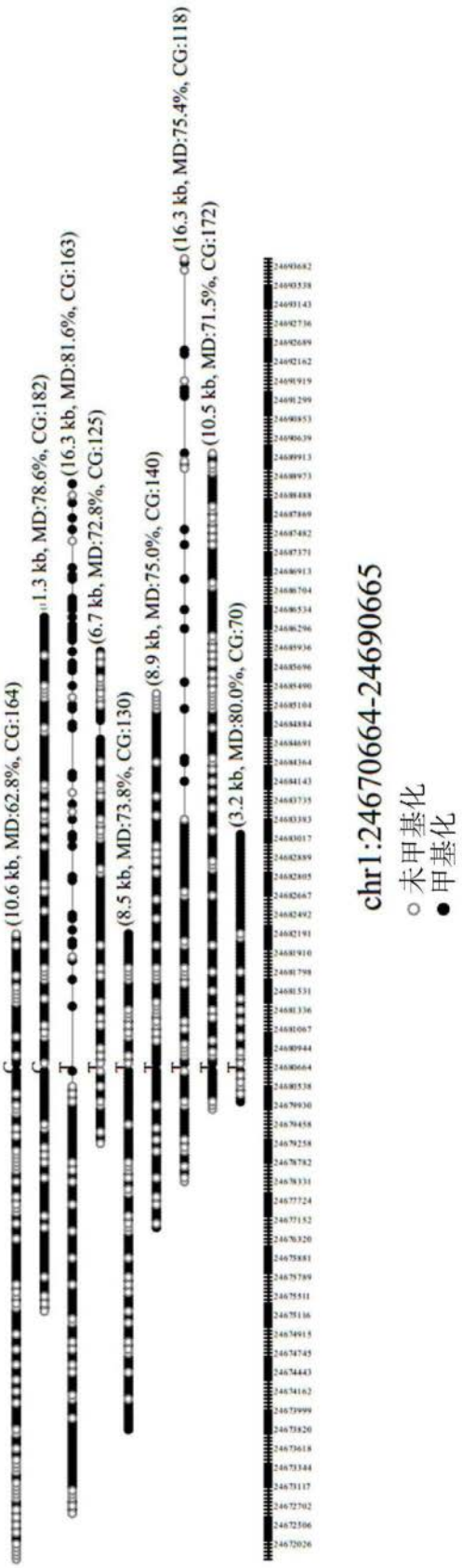

(10.6 kb, MD:62.8%, CG:164)
(1.3 kb, MD:78.6%, CG:182)
(16.3 kb, MD:81.6%, CG:163)
(6.7 kb, MD:72.8%, CG:125)
(8.5 kb, MD:73.8%, CG:130)
(8.9 kb, MD:75.0%, CG:140)
(16.3 kb, MD:75.4%, CG:118)
(10.5 kb, MD:71.5%, CG:172)
(3.2 kb, MD:80.0%, CG:70)
chr1:24670664-24690665
未甲基化
甲基化

图80C

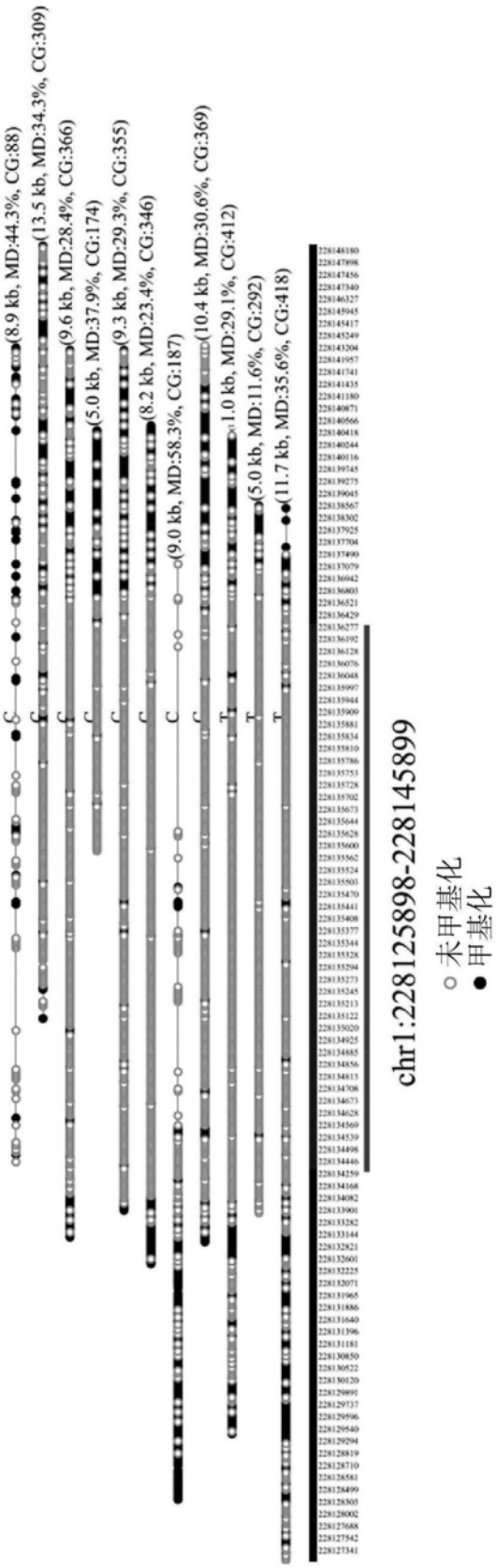

(8.9 kb, MD:44.3%, CG:88)
(13.5 kb, MD:34.3%, CG:309)
(9.6 kb, MD:28.4%, CG:366)
(5.0 kb, MD:37.9%, CG:174)
(9.3 kb, MD:29.3%, CG:355)
(8.2 kb, MD:23.4%, CG:346)
(9.0 kb, MD:58.3%, CG:187)
(10.4 kb, MD:30.6%, CG:369)
(1.0 kb, MD:29.1%, CG:412)
(5.0 kb, MD:11.6%, CG:292)
(11.7 kb, MD:35.6%, CG:418)
chr1:228125898-228145899
未甲基化
甲基化

图80D

	基因	等位基因1	等位基因2	甲基化水平（%）	
				等位基因1	等位基因2
印记基因	*SNURF*	T	C	15.73	89.37
	PLAGL1	T	C	7.56	89.41
	NAP1L5	C	T	12.5	91.07
	ZIM2	C	T	13	84.64
随机选择区域	区域01	G	A	71.79	69.17
	区域02	T	G	63.22	65.22
	区域03	C	T	73.33	74.9
	区域04	C	T	10.83	8.51

图81

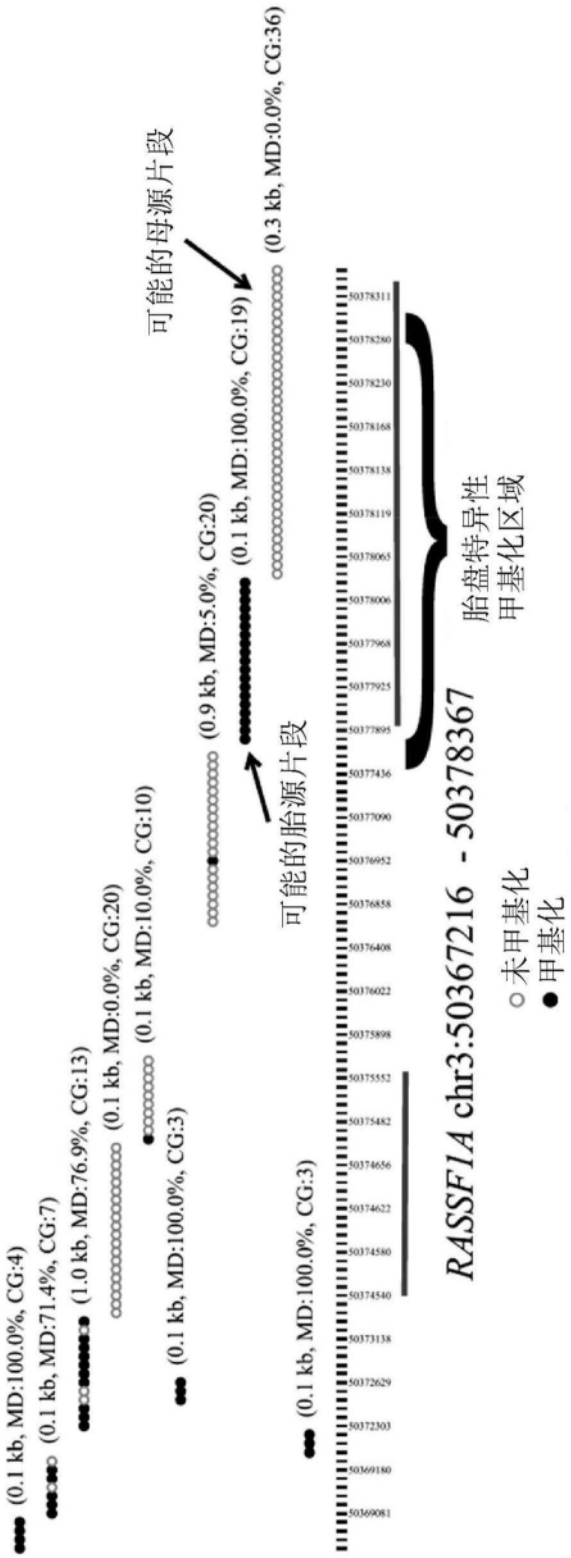

(0.7 kb, MD:20.0%, CG:15)
(0.1 kb, MD:0.0%, CG:2)
(0.1 kb, MD:100.0%, CG:4)
(0.1 kb, MD:71.4%, CG:7)
(1.0 kb, MD:76.9%, CG:13)
(0.1 kb, MD:0.0%, CG:20)
(0.1 kb, MD:10.0%, CG:10)
(0.1 kb, MD:100.0%, CG:3)
(0.9 kb, MD:5.0%, CG:20)
(0.1 kb, MD:100.0%, CG:19)
(0.3 kb, MD:0.0%, CG:36)
可能的母源片段
可能的胎源片段
(0.1 kb, MD:100.0%, CG:3)
RASSF1A chr3:50367216 - 50378367
胎盘特异性甲基化区域
○未甲基化
●甲基化

图82

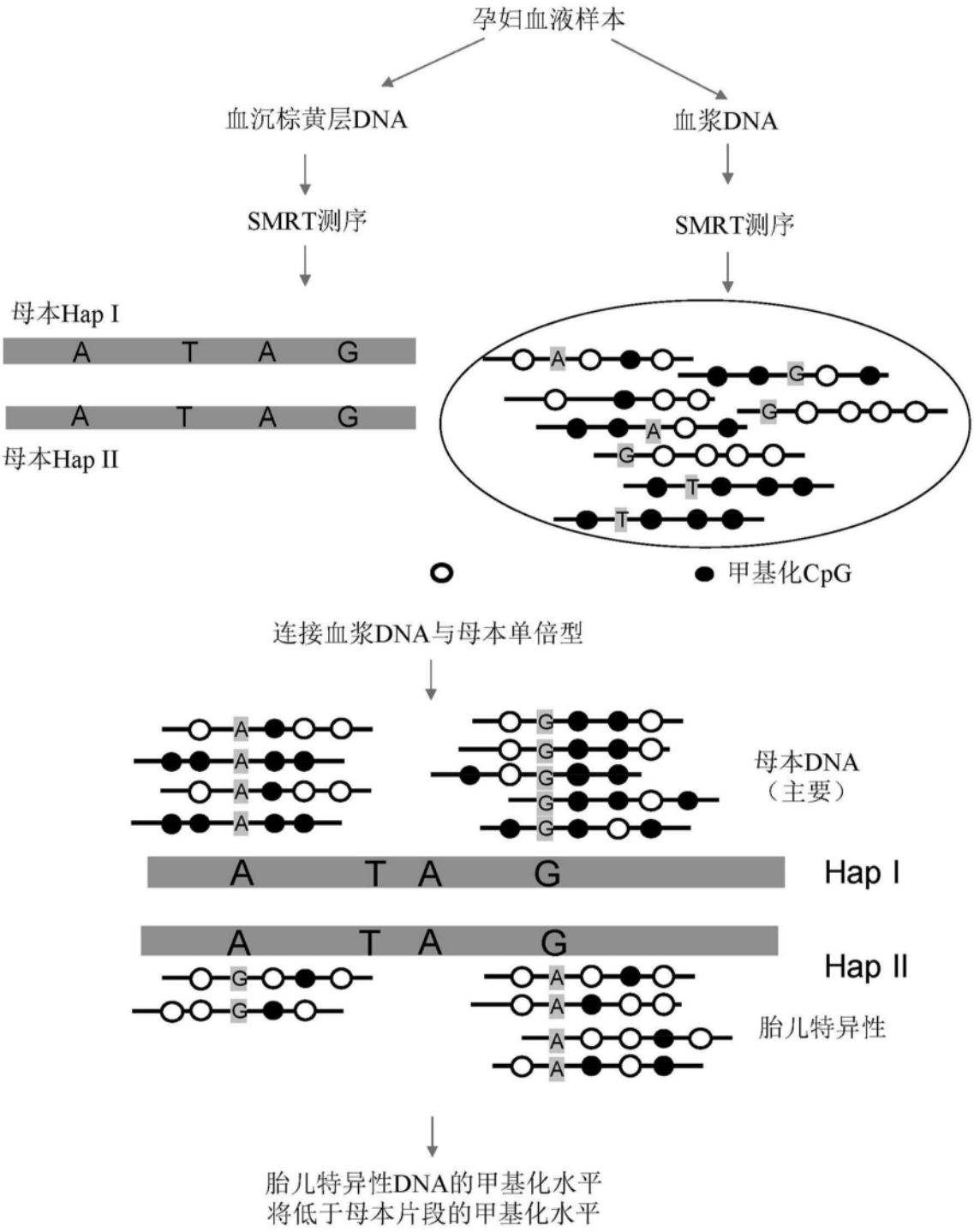

孕妇血液样本
血沉棕黄层DNA
血浆DNA
SMRT测序
SMRT测序
母本Hap I
A T A G
A T A G
母本Hap II
甲基化CpG
连接血浆DNA与母本单倍型
母本DNA
（主要）
A T A G
Hap I
A T A G
Hap II
胎儿特异性
胎儿特异性DNA的甲基化水平
将低于母本片段的甲基化水平

图83

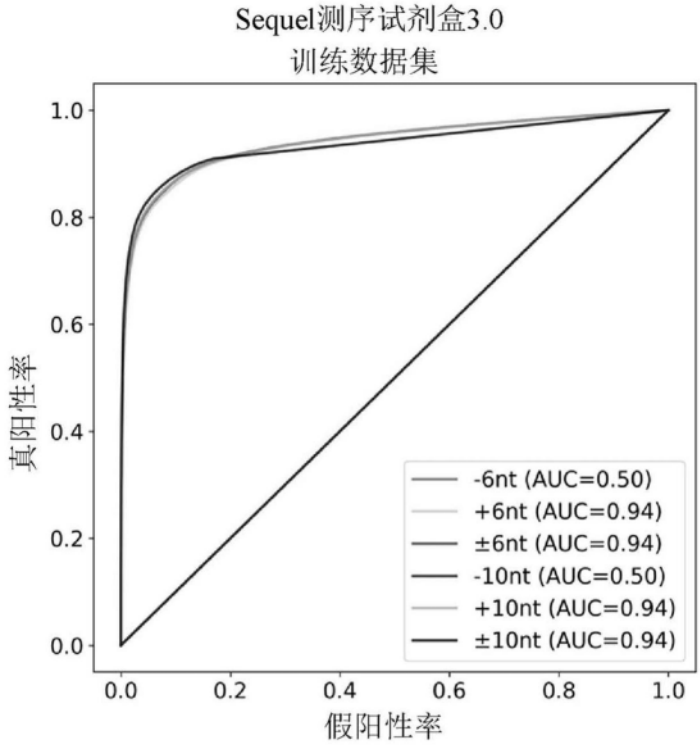

Sequel测序试剂盒3.0
训练数据集
真阳性率
假阳性率
-6nt (AUC=0.50)
+6nt (AUC=0.94)
±6nt (AUC=0.94)
-10nt (AUC=0.50)
+10nt (AUC=0.94)
±10nt (AUC=0.94)

图84A

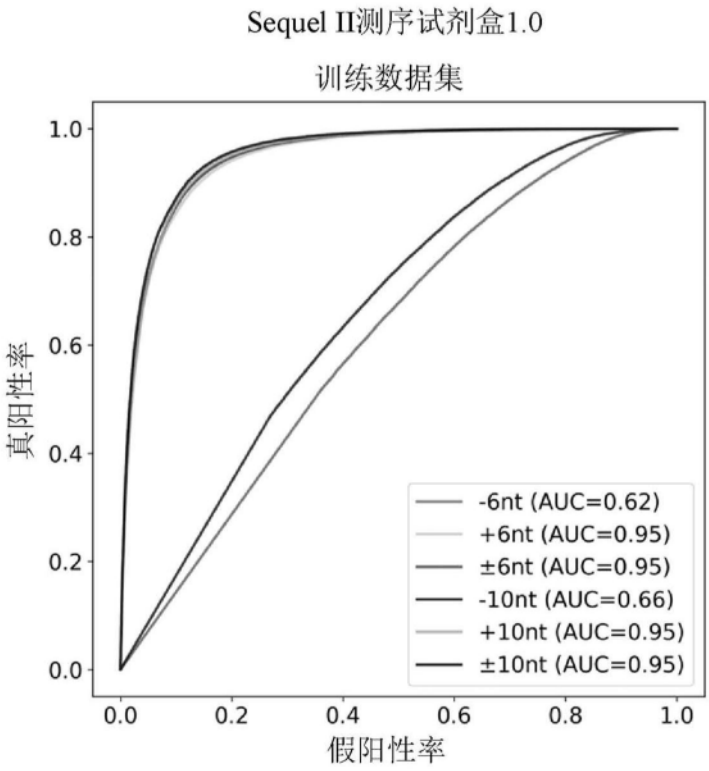

Sequel II测序试剂盒1.0
训练数据集
真阳性率
假阳性率
-6nt (AUC=0.62)
+6nt (AUC=0.95)
±6nt (AUC=0.95)
-10nt (AUC=0.66)
+10nt (AUC=0.95)
±10nt (AUC=0.95)

图84B

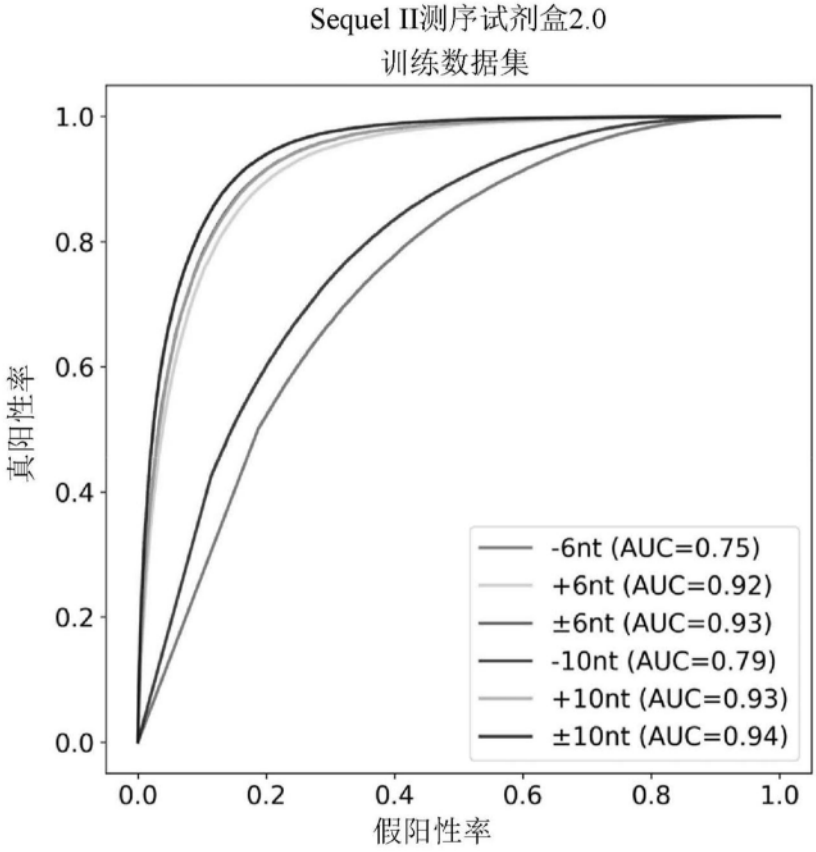

Sequel II测序试剂盒2.0
训练数据集
真阳性率
假阳性率
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.75)
+6nt (AUC=0.92)
±6nt (AUC=0.93)
-10nt (AUC=0.79)
+10nt (AUC=0.93)
±10nt (AUC=0.94)

图84C

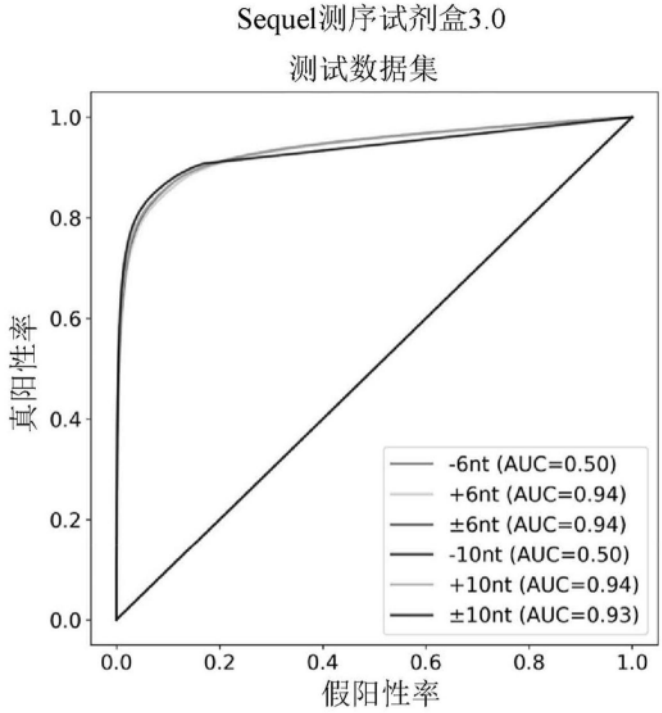

Sequel测序试剂盒3.0
测试数据集
真阳性率
假阳性率
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.50)
+6nt (AUC=0.94)
±6nt (AUC=0.94)
-10nt (AUC=0.50)
+10nt (AUC=0.94)
±10nt (AUC=0.93)

图85A

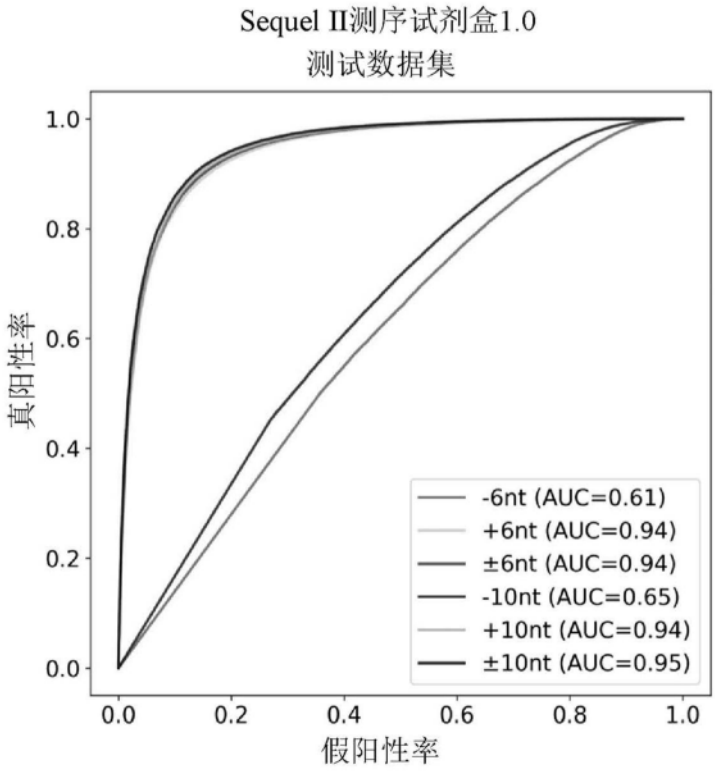

Sequel II测序试剂盒1.0
测试数据集
真阳性率
假阳性率
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.61)
+6nt (AUC=0.94)
±6nt (AUC=0.94)
-10nt (AUC=0.65)
+10nt (AUC=0.94)
±10nt (AUC=0.95)

图85B

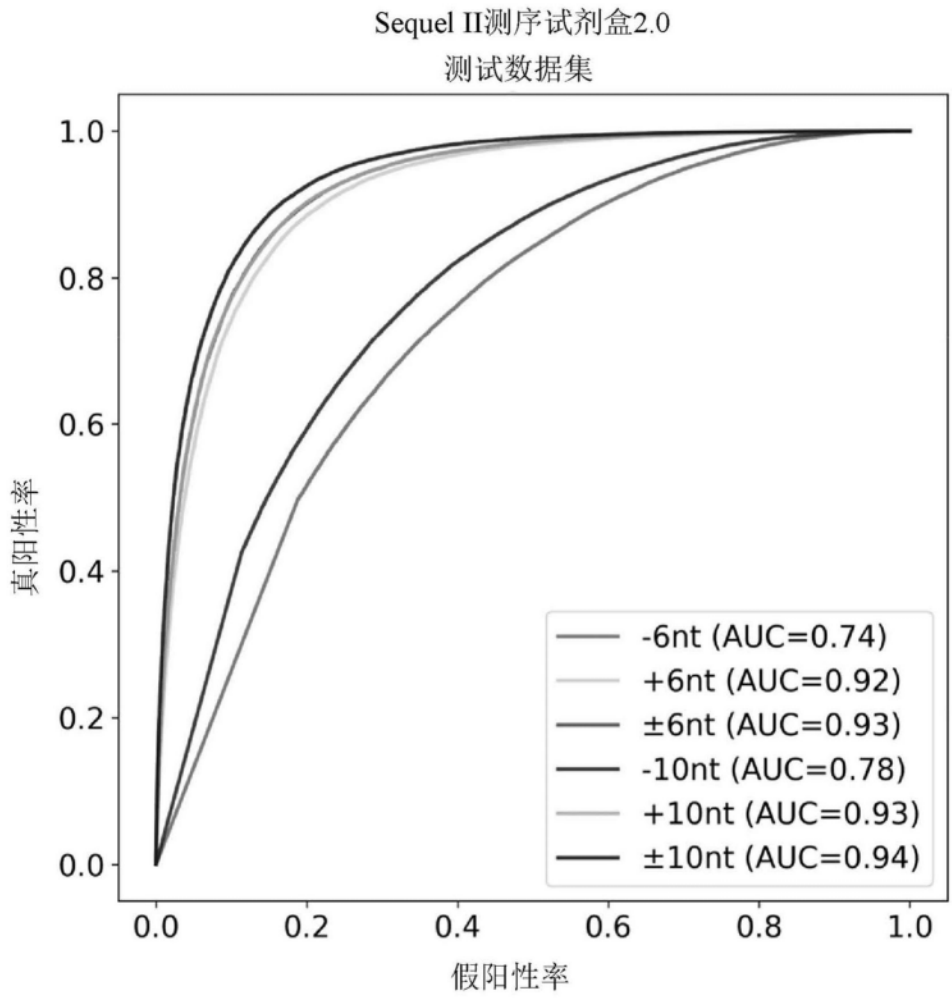

Sequel II测序试剂盒2.0
测试数据集
真阳性率
假阳性率
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.74)
+6nt (AUC=0.92)
±6nt (AUC=0.93)
-10nt (AUC=0.78)
+10nt (AUC=0.93)
±10nt (AUC=0.94)

图85C

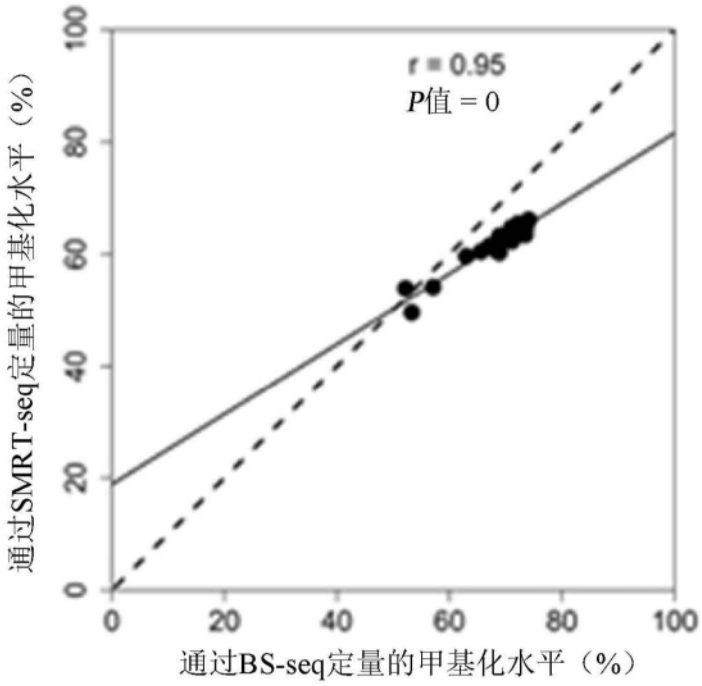

r = 0.95
P值 = 0
通过SMRT-seq定量的甲基化水平（%）
通过BS-seq定量的甲基化水平（%）
0
20
40
60
80
100

图86A

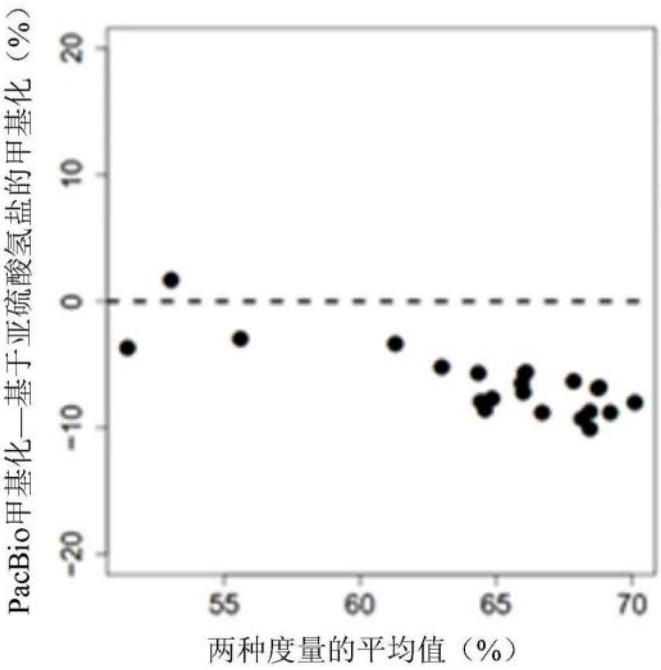

PacBio甲基化—基于亚硫酸氢盐的甲基化（%）
两种度量的平均值（%）
-20
-10
0
10
20
55
60
65
70

图86B

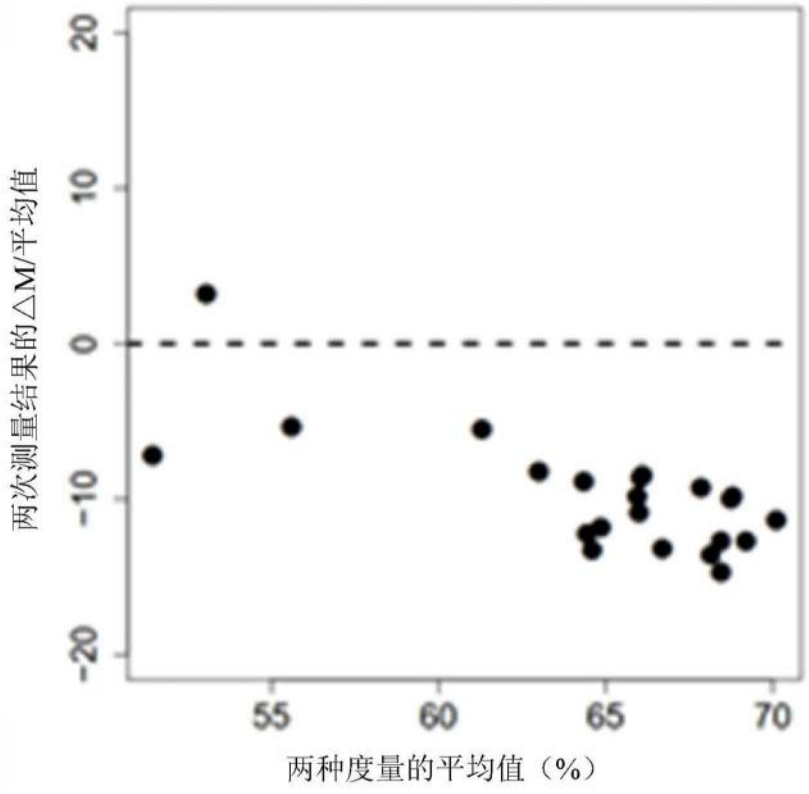

两次测量结果的△M/平均值
20
10
0
-10
-20
55
60
65
70
两种度量的平均值（%）

图86C

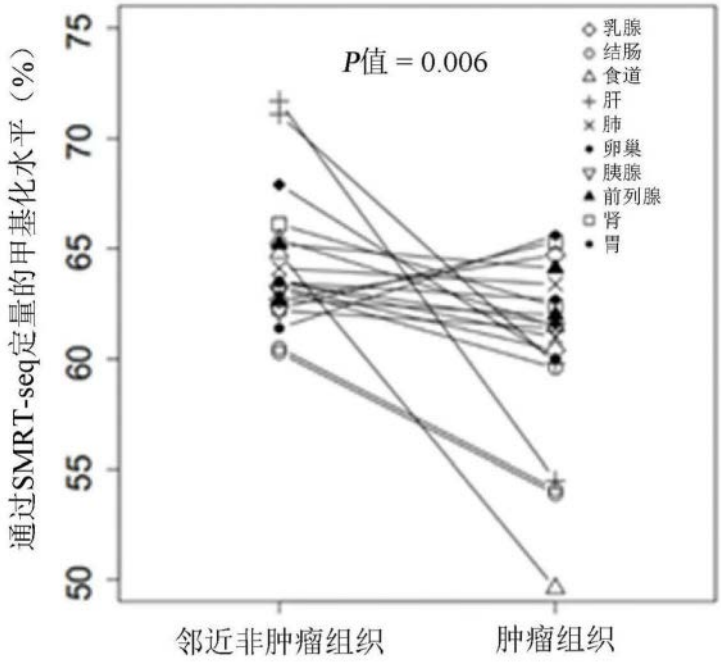

通过SMRT-seq定量的甲基化水平（%）
75
70
65
60
55
50
P值 = 0.006
乳腺
结肠
食道
肝
肺
卵巢
胰腺
前列腺
肾
胃
邻近非肿瘤组织
肿瘤组织

图87A

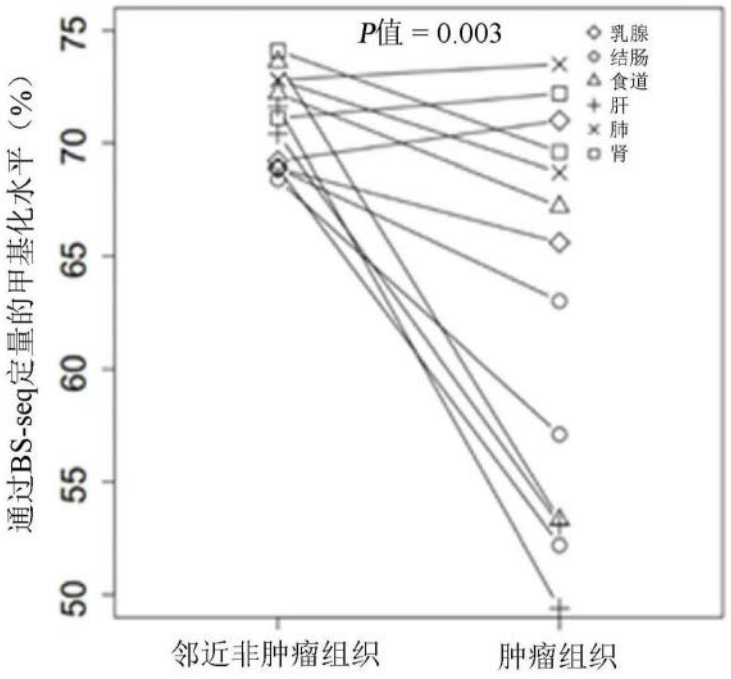

P值 = 0.003
乳腺
结肠
食道
肝
肺
肾
通过BS-seq定量的甲基化水平（%）
75
70
65
60
55
50
邻近非肿瘤组织
肿瘤组织

图87B

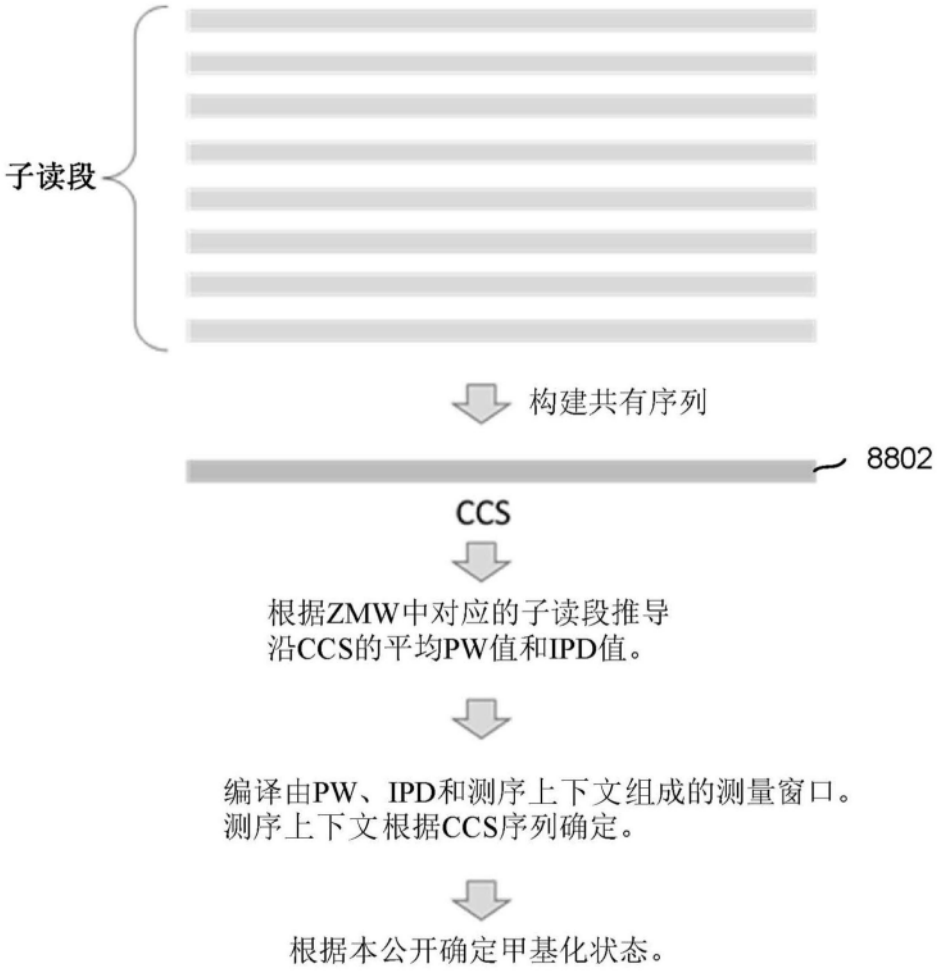

子读段
构建共有序列
8802
CCS
根据ZMW中对应的子读段推导
沿CCS的平均PW值和IPD值。
编译由PW、IPD和测序上下文组成的测量窗口。
测序上下文根据CCS序列确定。
根据本公开确定甲基化状态。

图88

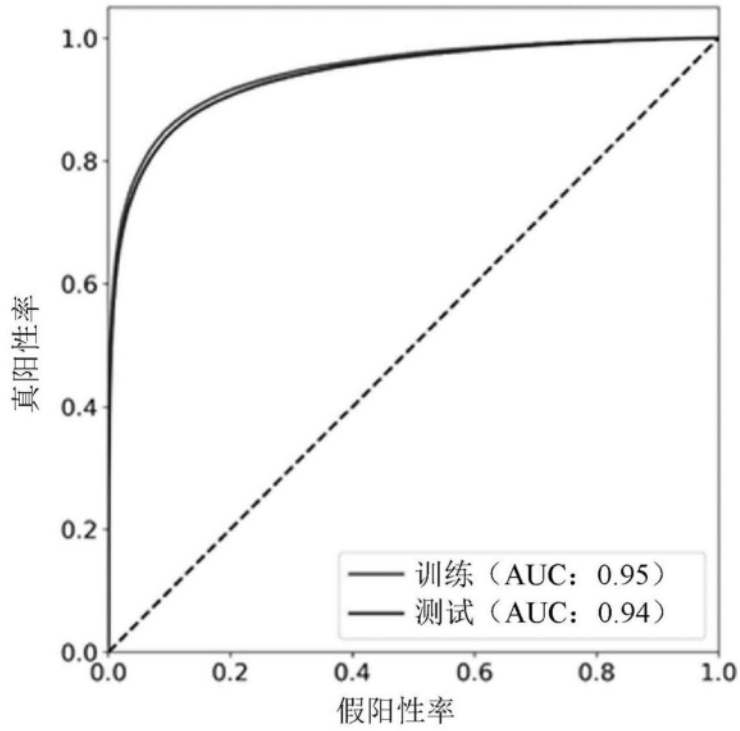

1.0
0.8
0.6
0.4
0.2
0.0
真阳性率
0.0
0.2
0.4
0.6
0.8
1.0
假阳性率
训练（AUC：0.95）
测试（AUC：0.94）

图89

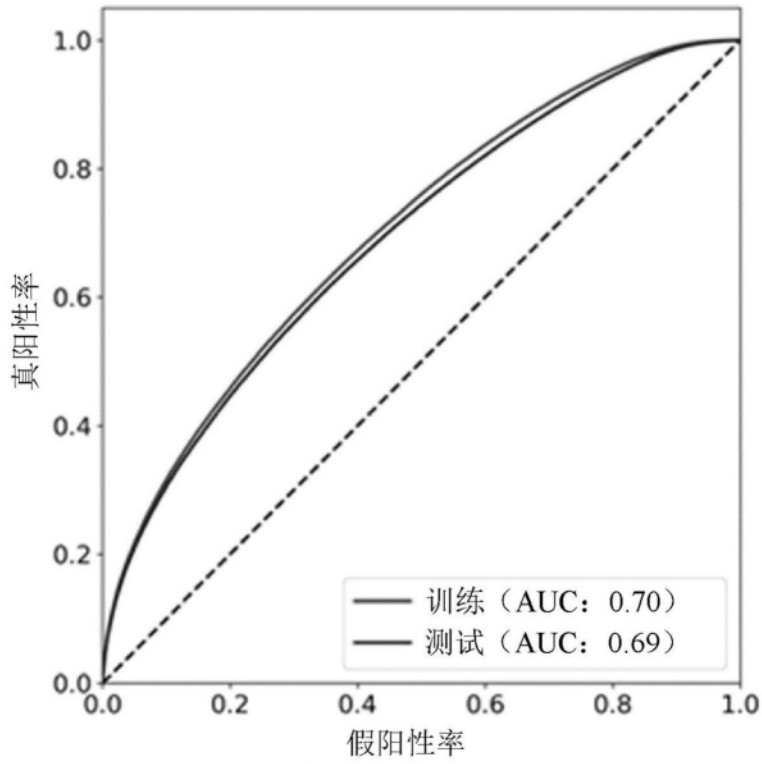

1.0
0.8
0.6
0.4
0.2
0.0
真阳性率
0.0
0.2
0.4
0.6
0.8
1.0
假阳性率
训练（AUC：0.70）
测试（AUC：0.69）

图90

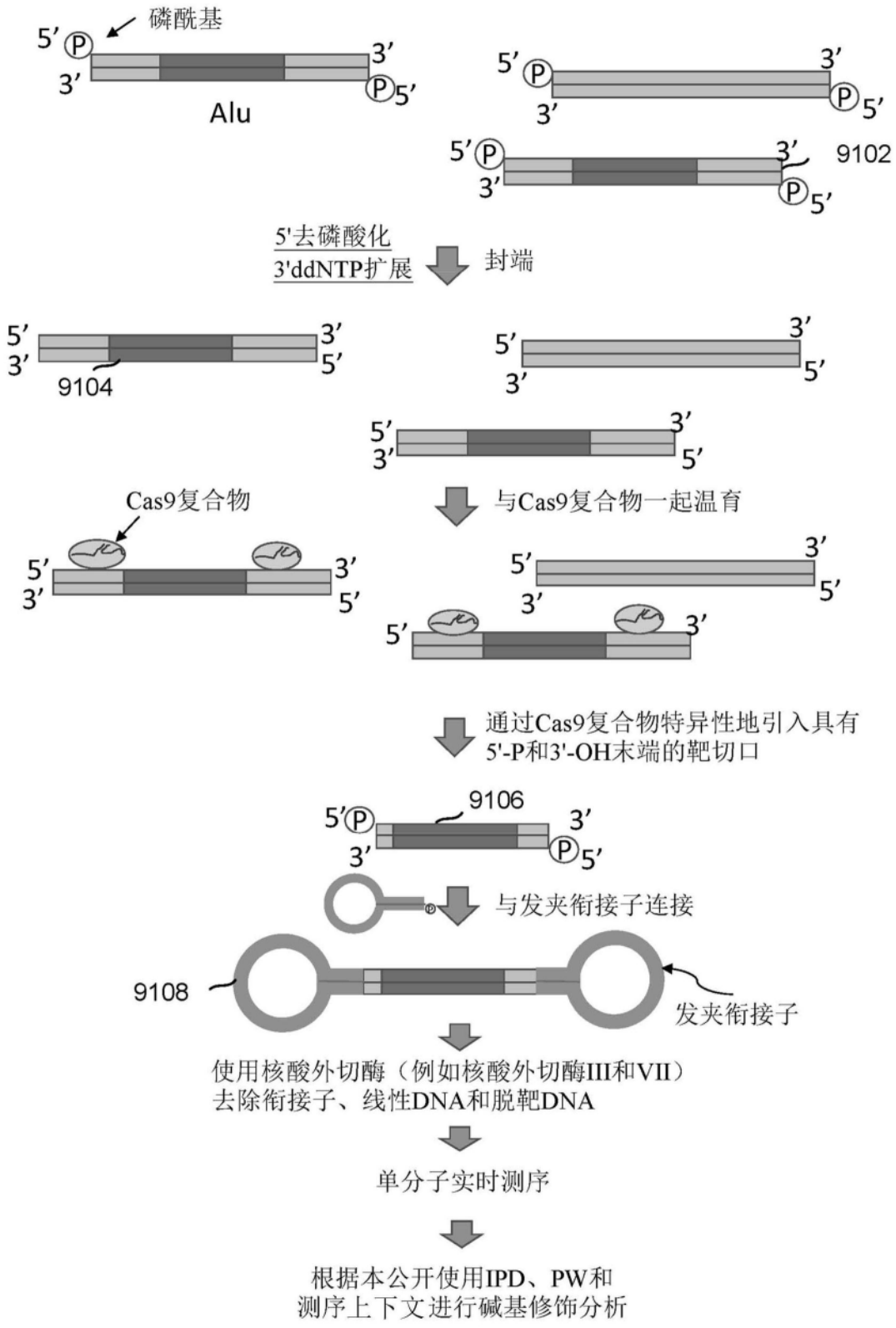

磷酰基
5′
3′
Alu
9102
5'去磷酸化
3'ddNTP扩展
封端
9104
Cas9复合物
与Cas9复合物一起温育
通过Cas9复合物特异性地引入具有
5'-P和3'-OH末端的靶切口
9106
与发夹衔接子连接
9108
发夹衔接子
使用核酸外切酶（例如核酸外切酶III和VII）
去除衔接子、线性DNA和脱靶DNA
单分子实时测序
根据本公开使用IPD、PW和
测序上下文进行碱基修饰分析

图91

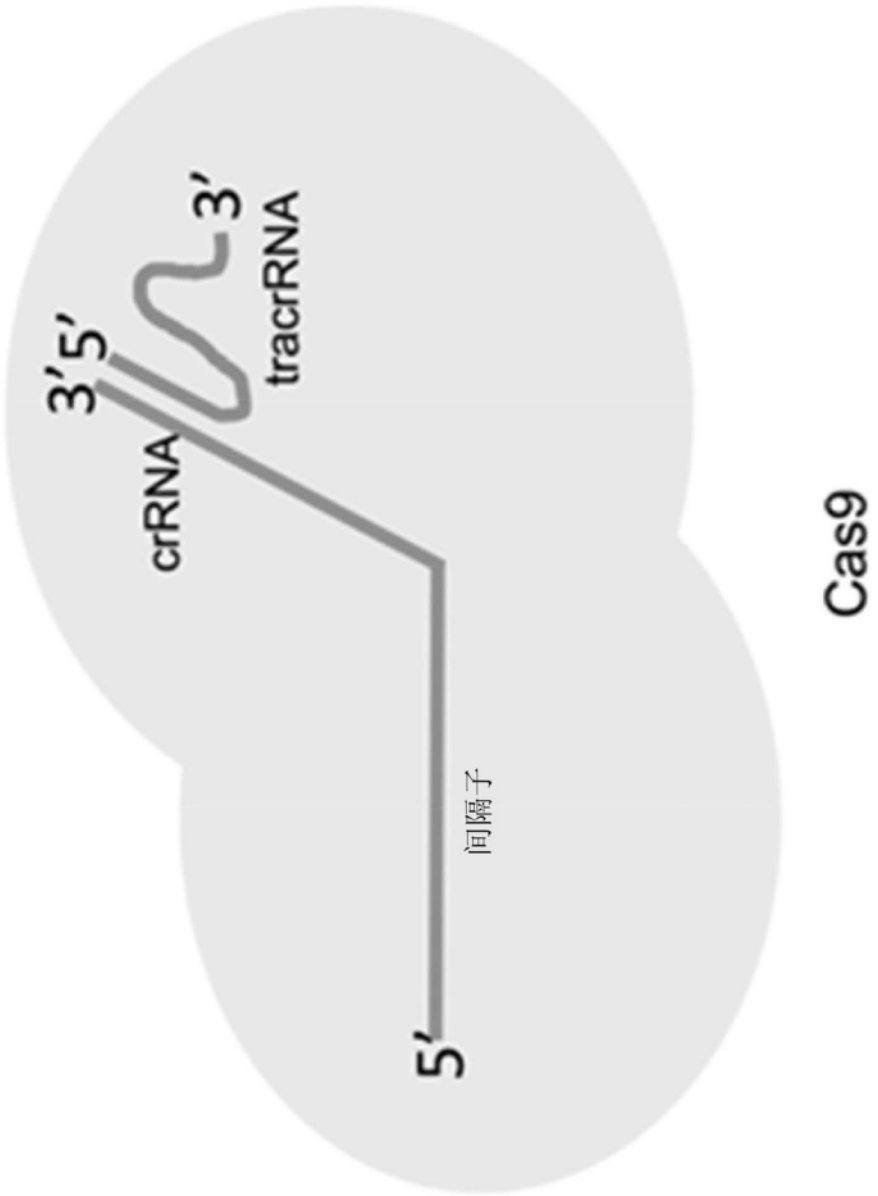


图92

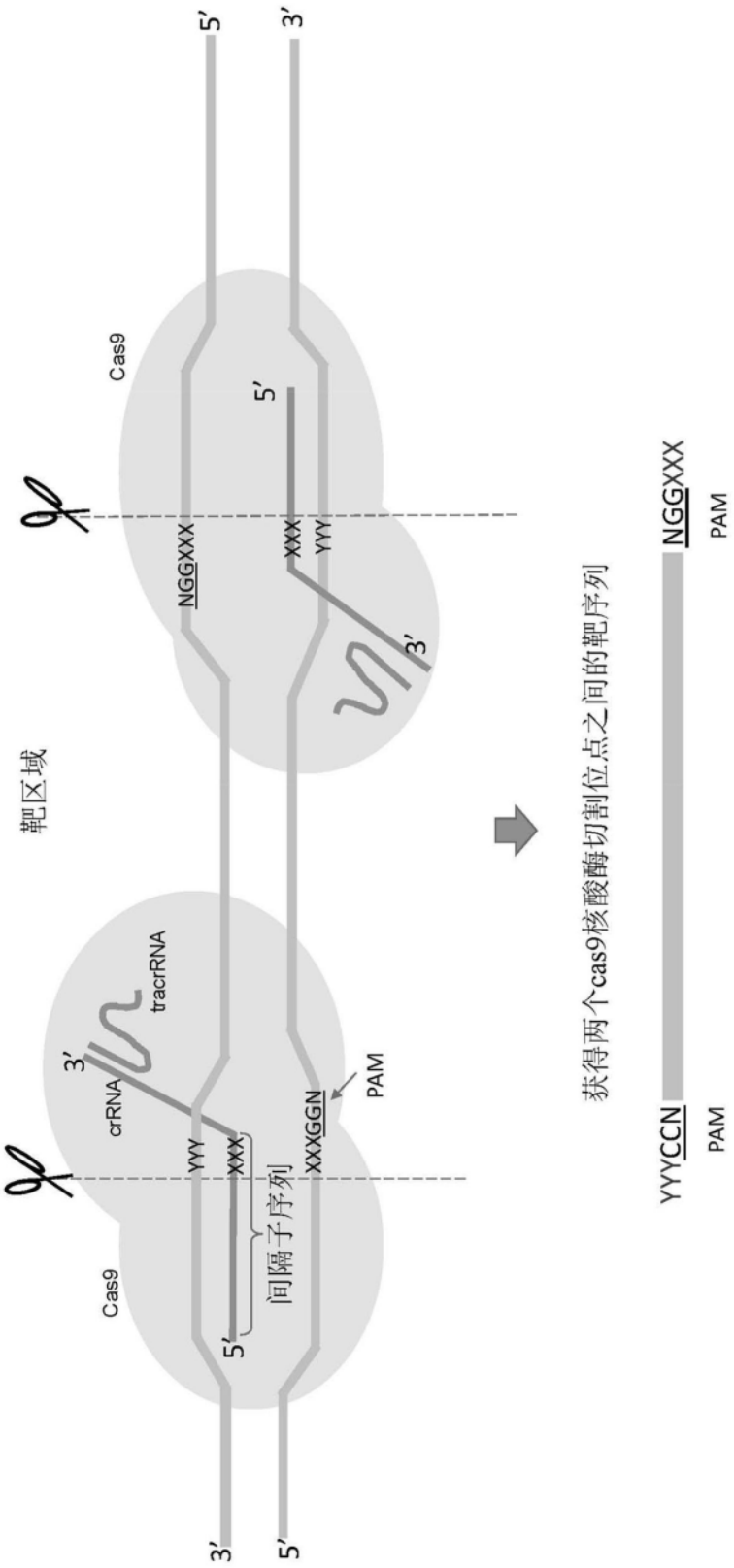

5′
3′
Cas9
NGGXXX
XXX
YYY
5′
3′
靶区域
tracrRNA
3′
crRNA
YYY
XXX
XXXGGN
PAM
间隔子序列
Cas9
5′
3′
获得两个cas9核酸酶切割位点之间的靶序列
NGGXXX
PAM
YYYCCN
PAM

图93

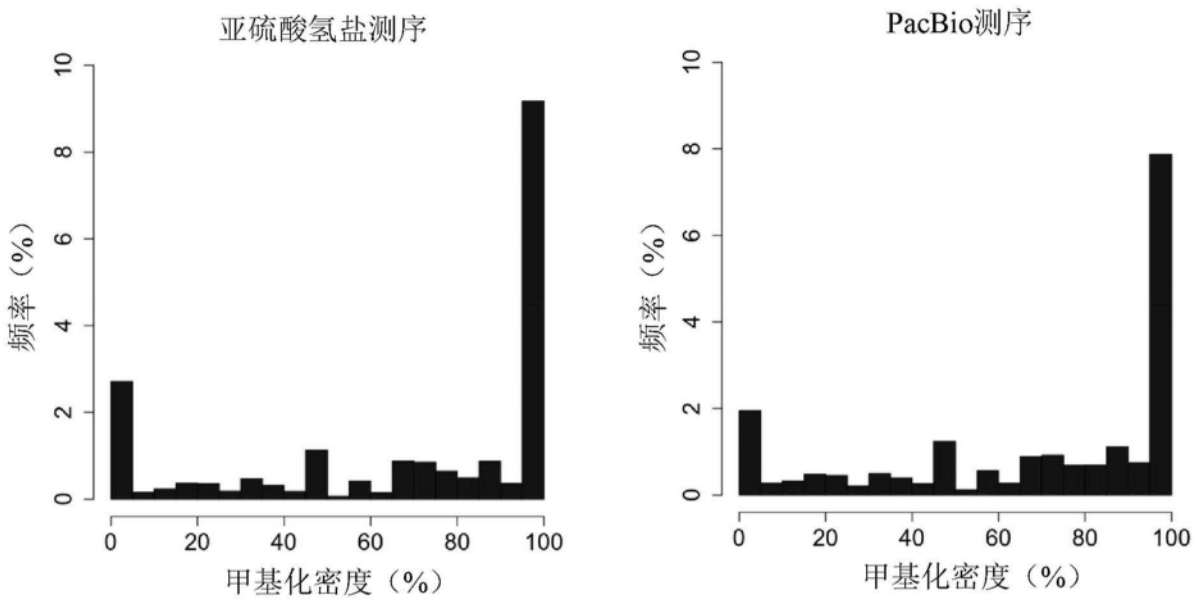

亚硫酸氢盐测序
频率（%）
甲基化密度（%）
PacBio测序
频率（%）
甲基化密度（%）

图94

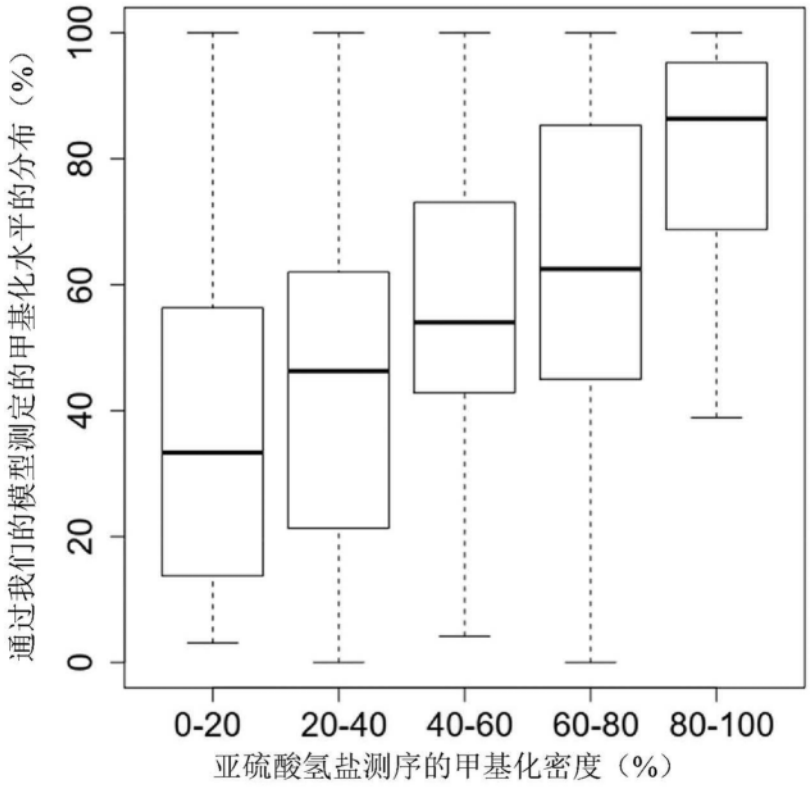

通过我们的模型测定的甲基化水平的分布（%）
0-20
20-40
40-60
60-80
80-100
亚硫酸氢盐测序的甲基化密度（%）

图95

组织	Alu甲基化水平（%）
血沉棕黄层	89.54
肝	88.18
结肠	89.56
肺	91.52
小肠	86.56
肾上腺	89.07
脂肪	91.44
胰腺	85.82
脑	91.79
HCC	76.74
胎盘	73.04

图96

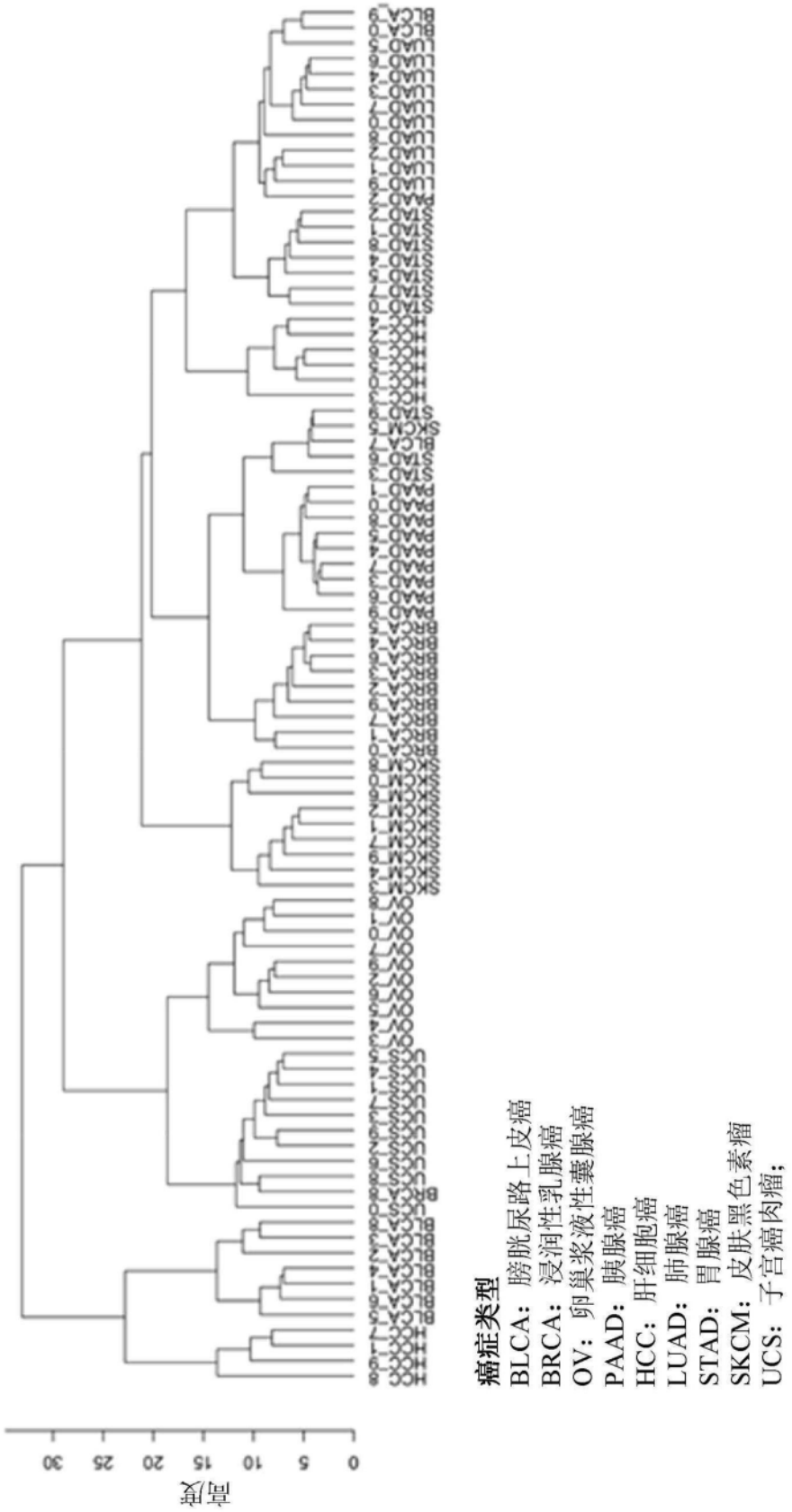

高度
0
5
10
15
20
25
30
癌症类型
BLCA：膀胱尿路上皮癌
BRCA：浸润性乳腺癌
OV：卵巢浆液性囊腺癌
PAAD：胰腺癌
HCC：肝细胞癌
LUAD：肺腺癌
STAD：胃腺癌
SKCM：皮肤黑色素瘤
UCS：子宫癌肉瘤；

图97

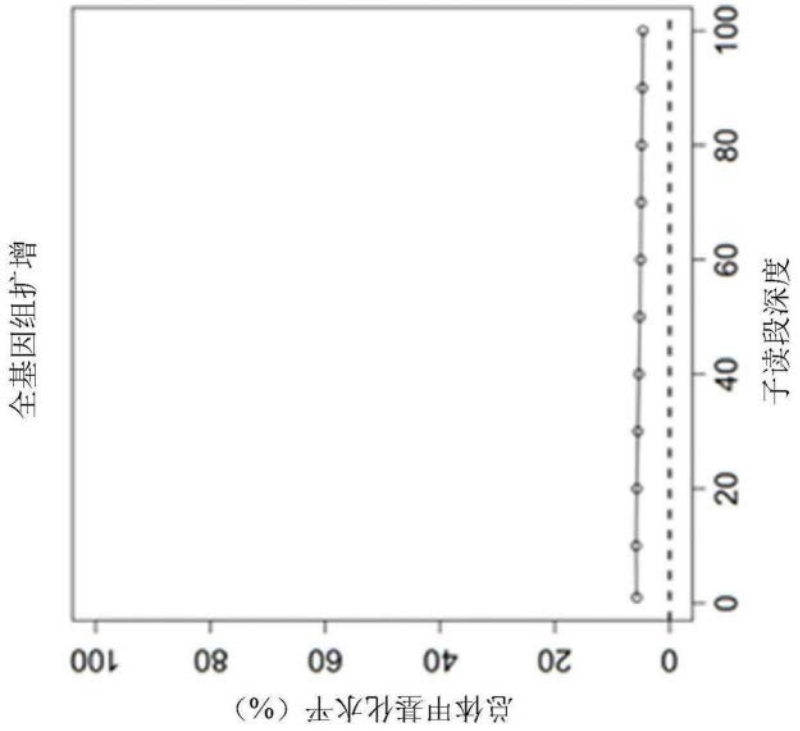

全基因组扩增
子读段深度
总体甲基化水平（%）
0
20
40
60
80
100

图98A

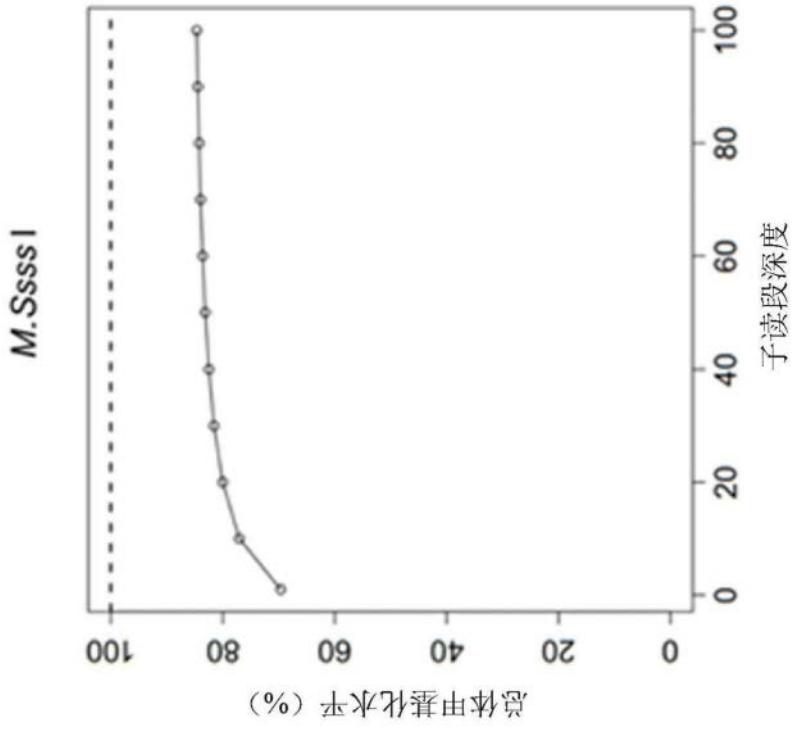

M.Ssss I
子读段深度
总体甲基化水平（%）
0
20
40
60
80
100

图98B

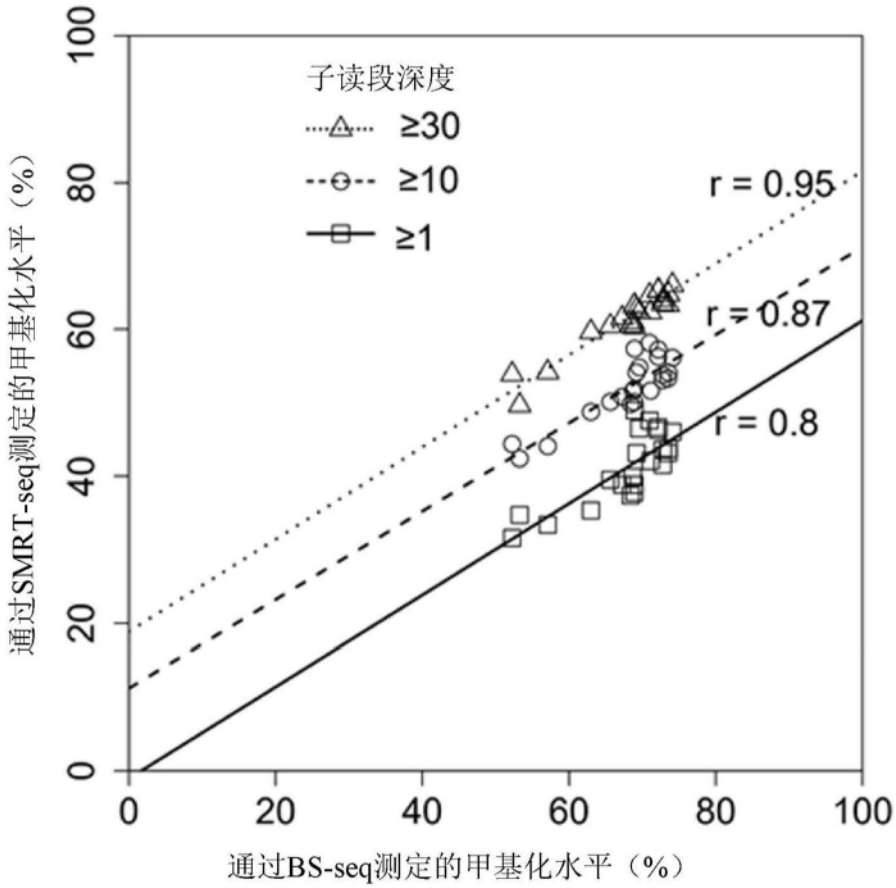


图99

子读段深度 截止值≥	皮尔森氏r （SMRT-seq对BS-seq）	CpG位点的数量
1	0.797	25,606,068 (23,949,832-27,008,582)
10	0.873	21,668,418 (18,263,886-23,515,147)
20	0.933	14,276,212 (10,526,406-16,736,887)
30	0.952	6,736,890 (4,255,452-10,449,814)
40	0.948	3,420,790 (2,232,511-5,792,825)
50	0.941	1,684,871 (1,278,475-3,055,876)
60	0.929	911,961 (707,295-1,581,313)
70	0.917	532,422 (350,001-866,045)
80	0.907	284,375 (177,698-534,540)
90	0.906	150,974 (98,000-333,933)
100	0.875	89,788 (58,552-182,861)

图100

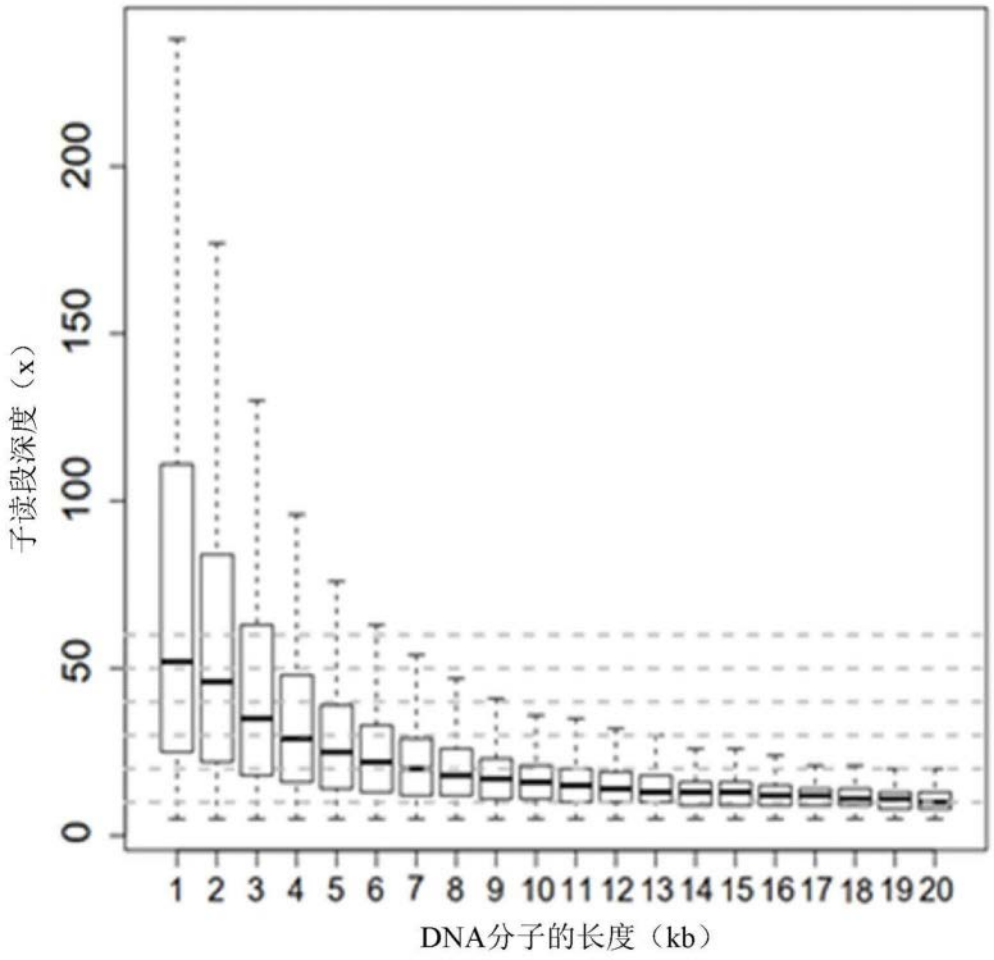

子读段深度（x）
0
50
100
150
200
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
DNA分子的长度（kb）

图101

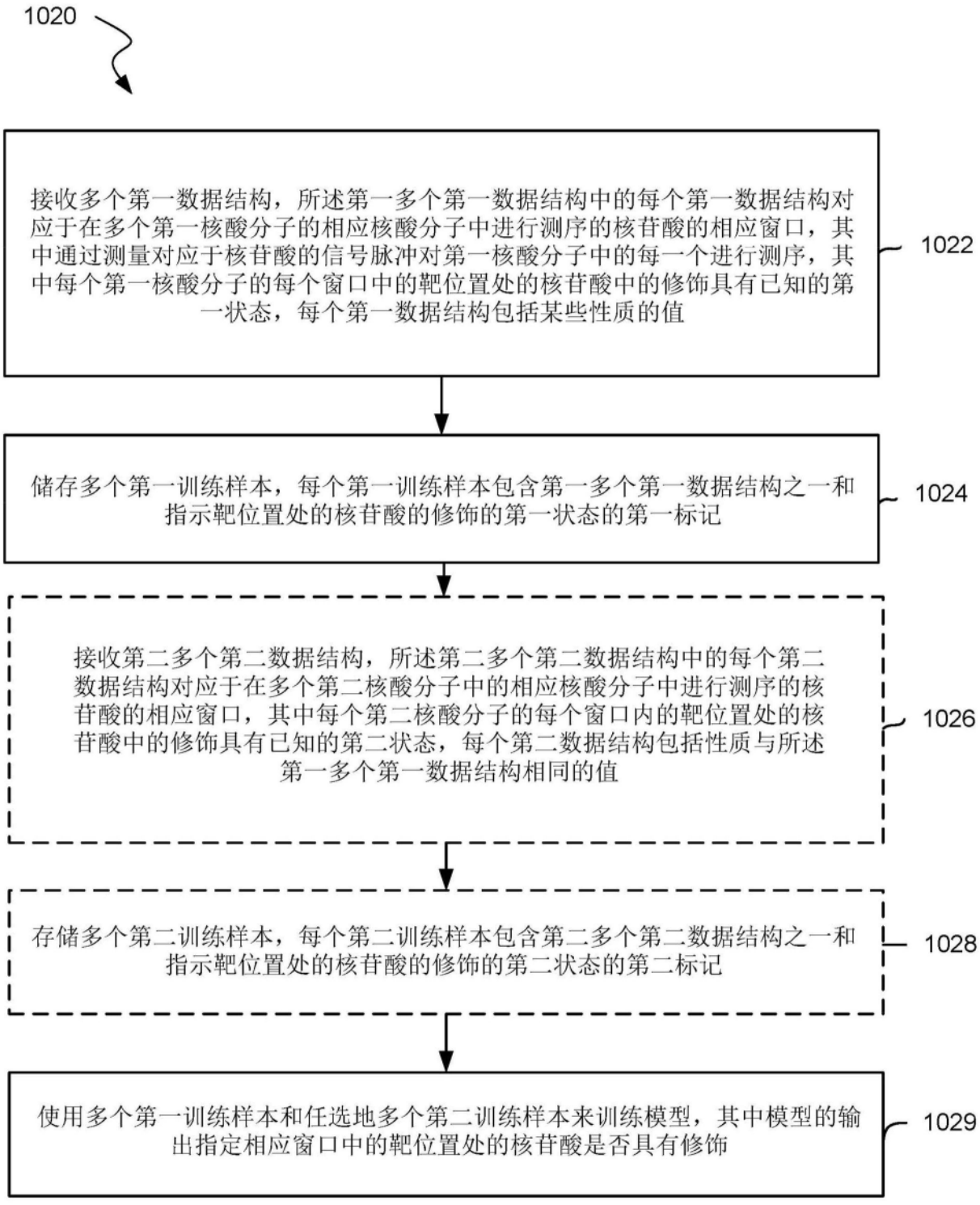

1020
接收多个第一数据结构，所述第一多个第一数据结构中的每个第一数据结构对应于在多个第一核酸分子的相应核酸分子中进行测序的核苷酸的相应窗口，其中通过测量对应于核苷酸的信号脉冲对第一核酸分子中的每一个进行测序，其中每个第一核酸分子的每个窗口中的靶位置处的核苷酸中的修饰具有已知的第一状态，每个第一数据结构包括某些性质的值
1022
储存多个第一训练样本，每个第一训练样本包含第一多个第一数据结构之一和指示靶位置处的核苷酸的修饰的第一状态的第一标记
1024
接收第二多个第二数据结构，所述第二多个第二数据结构中的每个第二数据结构对应于在多个第二核酸分子中的相应核酸分子中进行测序的核苷酸的相应窗口，其中每个第二核酸分子的每个窗口内的靶位置处的核苷酸中的修饰具有已知的第二状态，每个第二数据结构包括性质与所述第一多个第一数据结构相同的值
1026
存储多个第二训练样本，每个第二训练样本包含第二多个第二数据结构之一和指示靶位置处的核苷酸的修饰的第二状态的第二标记
1028
使用多个第一训练样本和任选地多个第二训练样本来训练模型，其中模型的输出指定相应窗口中的靶位置处的核苷酸是否具有修饰
1029

图102

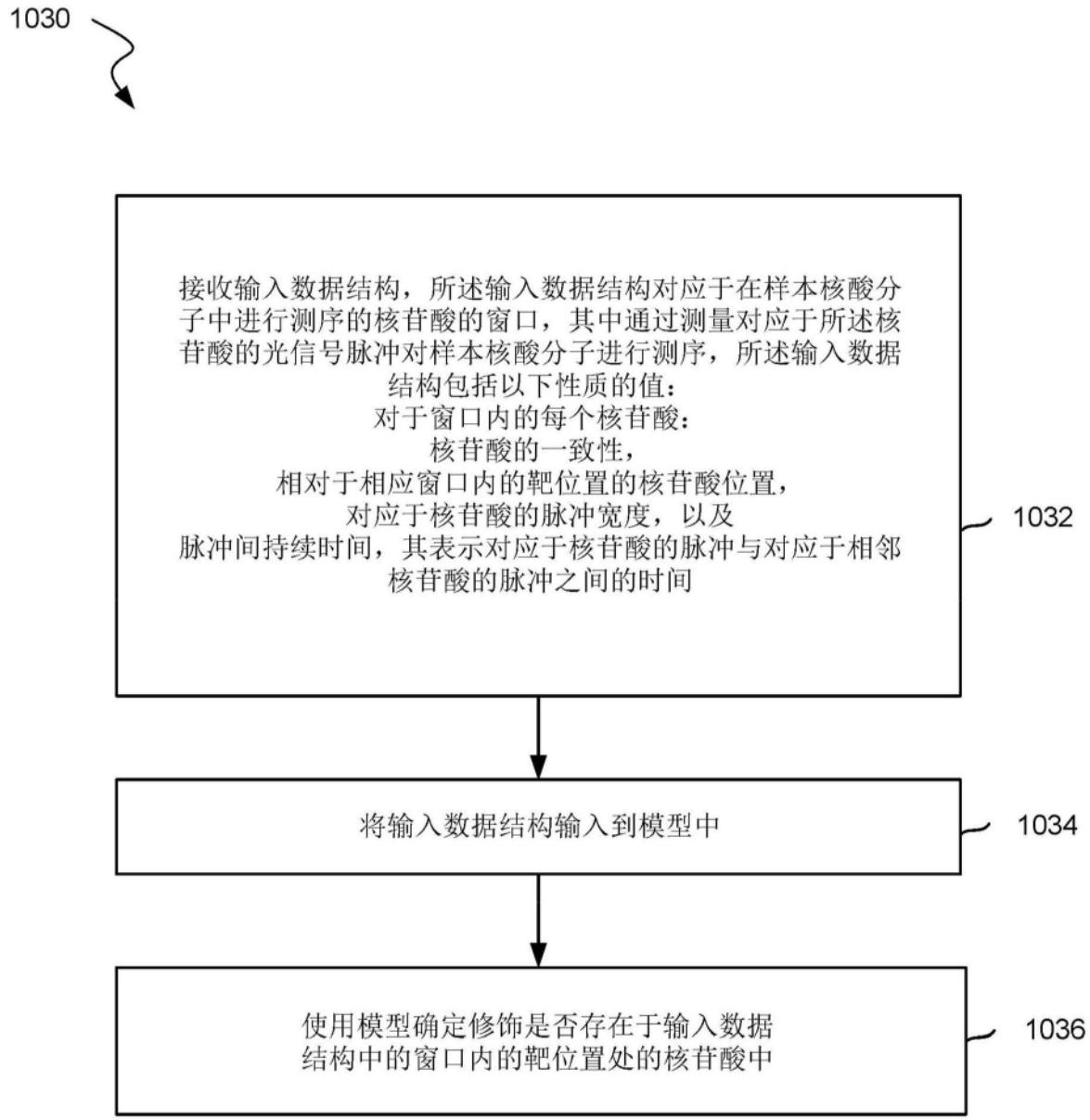

1030
接收输入数据结构，所述输入数据结构对应于在样本核酸分子中进行测序的核苷酸的窗口，其中通过测量对应于所述核苷酸的光信号脉冲对样本核酸分子进行测序，所述输入数据结构包括以下性质的值：
对于窗口内的每个核苷酸：
核苷酸的一致性，
相对于相应窗口内的靶位置的核苷酸位置，
对应于核苷酸的脉冲宽度，以及
脉冲间持续时间，其表示对应于核苷酸的脉冲与对应于相邻核苷酸的脉冲之间的时间
1032
将输入数据结构输入到模型中
1034
使用模型确定修饰是否存在于输入数据结构中的窗口内的靶位置处的核苷酸中
1036

图103

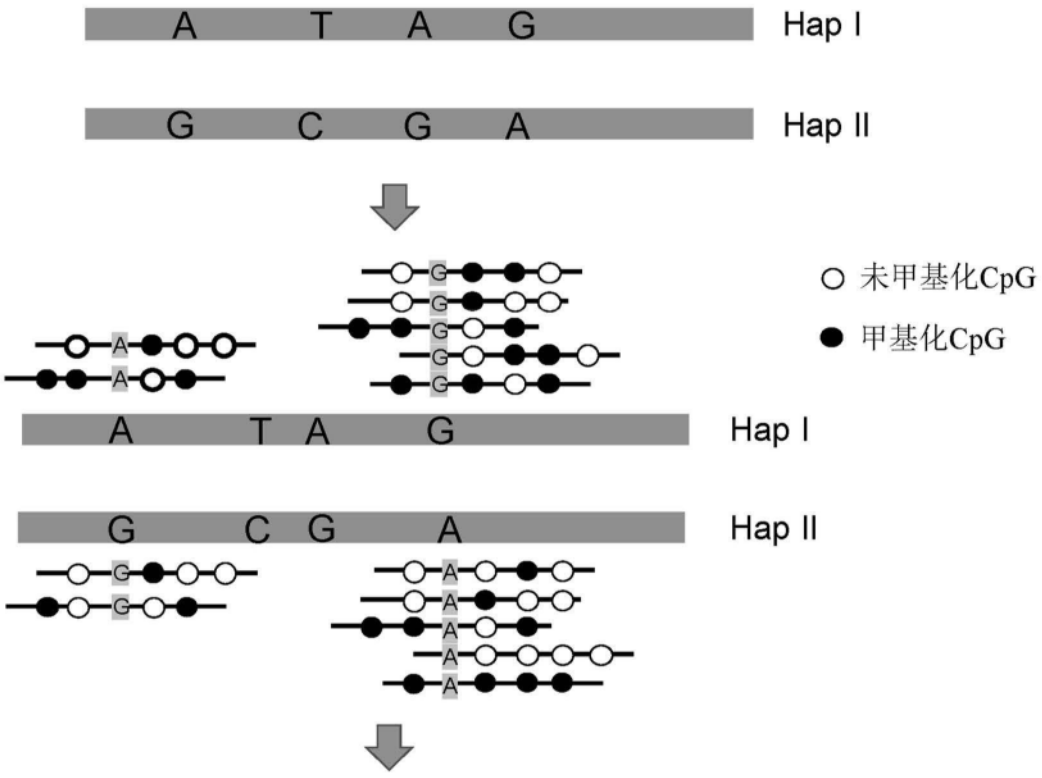

A T A G
Hap I
G C G A
Hap II
未甲基化CpG
甲基化CpG
A T A G
Hap I
G C G A
Hap II
基于相对单倍型的甲基化失衡指数
例如Hap I的甲基化水平—Hap II的甲基化水平

图104

Chr	开始	结束	长度	单倍型嵌段id	PacBio测序			
					邻近非肿瘤组织的甲基化水平		肿瘤组织的甲基化水平	
					Hap I	Hap II	Hap I	Hap II
chr1	56312395	56347696	35301	hap1927	68.2	67.4	60.3	23.5
chr1	194413819	194424806	10987	hap5953	52.8	49.5	48.8	9.3
chr1	220674478	220699011	24533	hap6863	63.0	64.5	50.4	17.3
chr10	113088792	113124248	35456	hap11838	62.7	63.4	38.1	5.7
chr11	5482746	5498801	16055	hap12904	70.3	75.0	16.3	51.7
chr11	42819351	42852772	33421	hap14385	54.6	54.9	65.3	17.8
chr11	57983961	58051078	67117	hap14930	67.3	66.4	58.2	18.6
chr11	60174708	60204209	29501	hap14990	58.4	59.8	49.6	10.8
chr12	128079419	128114656	35237	hap22249	60.0	58.3	12.1	45.2
chr15	20480575	20533464	52889	hap29631	64.7	69.1	27.7	59.3
chr15	94902853	94946231	43378	hap32161	74.1	74.5	74.9	15.8
chr15	96526684	96549225	22541	hap32221	70.8	68.8	28.9	64.4
chr16	31595372	31613277	17905	hap33499	55.9	59.3	46.3	14.4
chr16	80151778	80182097	30319	hap34821	71.1	71.0	11.5	51.8
chr16	82519715	82554191	34476	hap34920	71.3	66.5	47.4	13.0
chr17	21668593	21685572	16979	hap36049	50.3	47.8	67.4	19.6
chr17	44999177	45012087	12910	hap36640	47.1	45.2	81.6	35.1
chr17	69911623	69926625	15002	hap37435	67.3	63.0	37.8	5.2
chr18	11441122	11458521	17399	hap38335	65.5	66.8	65.9	22.4
chr18	23405569	23423387	17818	hap38673	66.3	61.7	3.3	48.1
chr18	68887284	68925031	37747	hap40390	63.0	61.0	22.0	53.4
chr18	69487809	69505470	17661	hap40414	74.5	74.1	33.3	72.2
chr2	41480394	41514135	33741	hap43972	54.0	54.0	14.9	77.8
chr2	114171214	114182880	11666	hap46226	72.4	68.8	79.7	16.7
chr2	123762541	123797629	35088	hap46589	66.7	68.1	24.0	54.5
chr2	125236882	125241950	5068	hap46673	58.9	59.2	10.7	46.4
chr2	130016110	130040331	24221	hap46835	54.6	50.8	5.6	41.6
chr2	137757638	137783716	26078	hap47090	61.8	61.4	13.5	69.2
chr2	144128597	144160845	32248	hap47343	65.8	66.6	9.3	50.3
chr20	15736792	15753459	16667	hap51505	78.9	74.3	45.8	77.3
chr20	26167979	26177235	9256	hap51868	55.0	52.2	38.5	68.6
chr20	44255808	44264190	8382	hap52246	57.4	56.1	9.7	50.6
chr20	59518410	59559273	40863	hap52761	61.0	62.4	30.0	72.8
chr21	21402034	21424129	22095	hap53197	63.5	67.3	25.0	75.5
chr21	24750027	24768793	18766	hap53333	68.2	64.6	3.4	38.9
chr21	26666833	26701575	34742	hap53418	62.1	66.5	47.6	16.7
chr3	2364024	2387896	23872	hap55539	67.4	67.8	54.9	10.9
chr3	21036965	21049451	12486	hap56223	54.8	51.4	53.1	21.1
chr3	56011690	56046642	34952	hap57346	64.2	61.2	71.2	22.6

图105A

chr3	73330942	73371216	40274	hap57939	60.9	62.9	9.4	42.9
chr3	106372440	106401301	28861	hap59077	67.8	67.9	13.8	53.2
chr3	107772994	107807482	34488	hap59122	69.6	73.5	30.4	66.4
chr3	116742501	116776747	34246	hap59493	64.3	69.1	14.1	51.6
chr3	171076306	171100102	23796	hap61495	68.0	66.0	80.6	48.8
chr3	193058272	193080344	22072	hap62231	65.5	64.7	54.6	20.0
chr4	30411613	30432317	20704	hap63589	59.3	60.6	53.4	14.6
chr4	31304718	31338193	33475	hap63633	60.2	60.0	7.2	55.0
chr4	92003467	92030505	27038	hap65794	65.3	65.1	54.1	21.7
chr4	155224697	155250915	26218	hap68104	60.5	57.5	57.3	25.0
chr5	2281802	2299281	17479	hap69632	71.5	66.9	69.9	6.6
chr5	4624948	4664704	39756	hap69739	62.8	61.0	14.0	52.0
chr5	89593236	89606080	12844	hap72628	76.6	74.0	20.3	78.4
chr5	119214026	119233058	19032	hap73698	62.8	61.2	57.6	13.1
chr5	119940397	119972658	32261	hap73720	59.1	54.7	53.8	12.2
chr5	132859668	132877415	17747	hap74150	62.5	66.6	59.5	28.3
chr6	26914610	26936918	22308	hap76887	41.9	40.9	71.9	32.6
chr6	66879106	66957243	78137	hap78266	61.6	59.6	25.4	62.0
chr6	77349083	77377529	28446	hap78674	64.5	66.4	27.0	62.9
chr6	159738794	159751033	12239	hap81616	79.6	79.0	21.2	59.8
chr7	26585255	26641907	56652	hap83161	66.2	64.7	49.4	13.3
chr7	48214640	48248036	33396	hap84003	76.0	76.7	78.0	32.3
chr7	88558182	88575482	17300	hap85335	63.8	59.6	63.8	22.9
chr7	96588562	96607580	19018	hap85620	60.4	63.1	19.7	50.0
chr7	122942180	122956897	14717	hap86454	42.3	39.0	19.2	50.0
chr7	132321970	132344802	22832	hap86807	61.4	60.7	52.5	11.5
chr7	153296219	153302441	6222	hap87487	48.7	53.7	64.4	19.3
chr7	156356247	156371897	15650	hap87631	74.9	71.6	87.5	56.6
chr7	159091986	159119486	27500	hap87738	54.0	49.1	52.0	13.2
chr8	51530582	51550889	20307	hap89477	66.4	65.7	68.0	19.9
chr8	63513932	63537543	23611	hap89942	62.0	63.3	11.6	48.4
chr8	72373321	72398122	24801	hap90226	58.0	54.9	71.6	32.0
chr8	94100451	94141855	41404	hap90991	65.2	65.7	36.2	68.7
chr8	109300499	109326404	25905	hap91510	63.6	67.7	29.5	65.8

图105B

Chr	开始	结束	长度	单倍型嵌段id	PacBio测序			
					邻近非肿瘤组织的甲基化水平		肿瘤组织的甲基化水平	
					Hap I	Hap II	Hap I	Hap II
chr9	27803548	27888202	84654	hap58508	64.2	60.9	20.6	75.4
chr6	242149	386636	144487	hap47880	62.3	63.3	77.4	32.2
chr5	28219159	28302858	83699	hap44666	59.3	58.0	16.8	58.2
chr5	18119943	18153743	33800	hap44475	61.6	65.0	53.2	21.7
chr7	24906307	25046195	139888	hap52069	69.3	68.7	44.0	76.2
chr15	27689897	27752573	62676	hap18337	65.9	61.9	64.8	20.5
chr12	42183870	42212433	28563	hap12045	63.5	68.4	19.4	51.2
chr21	9825597	9935752	110155	hap34175	54.3	53.5	60.9	29.1
chr2	118813055	118893366	80311	hap30060	62.6	62.3	77.0	38.6
chr6	90307702	90344869	37167	hap49779	69.1	66.4	84.7	53.9
chr7	107932914	108049376	116462	hap53838	67.2	62.9	43.8	76.4
chr7	137039327	137160933	121606	hap54447	59.5	60.9	22.9	72.0
chr17	21193754	21254930	61176	hap22633	59.2	54.3	69.7	31.6
chr12	11473697	11644714	171017	hap11451	62.8	66.4	35.5	75.9
chr5	129212299	129353349	141050	hap46632	50.9	54.5	45.5	14.0
chr11	93910738	94028887	118149	hap10288	67.6	63.6	36.6	74.2
chr3	131707434	132003636	296202	hap38642	57.8	55.9	17.9	60.2
chr3	43024004	43161785	137781	hap36769	69.1	66.5	46.1	80.2
chr3	190403156	190606658	203502	hap39947	60.9	61.6	36.9	72.7
chr15	40218970	40279780	60810	hap18606	53.4	57.5	79.1	47.4

图106

组织类型	表示肿瘤组织中的两种单倍型之间甲基化失衡的单倍型区块的数量	表示成对的邻近非肿瘤组织中的两种单倍型之间甲基化失衡的单倍型区块的数量
结肠	92	47
乳腺	57	13
肾	68	18
肺	31	21
胎盘	26	19
胃	2	0

图107A

组织类型	表示肿瘤组织中的两种单倍型之间甲基化失衡的单倍型区块的数量	可用的肿瘤分期信息（TNM）
乳腺	18	T2
	57	T3
肾	68	T3a
	0	T2

图107B

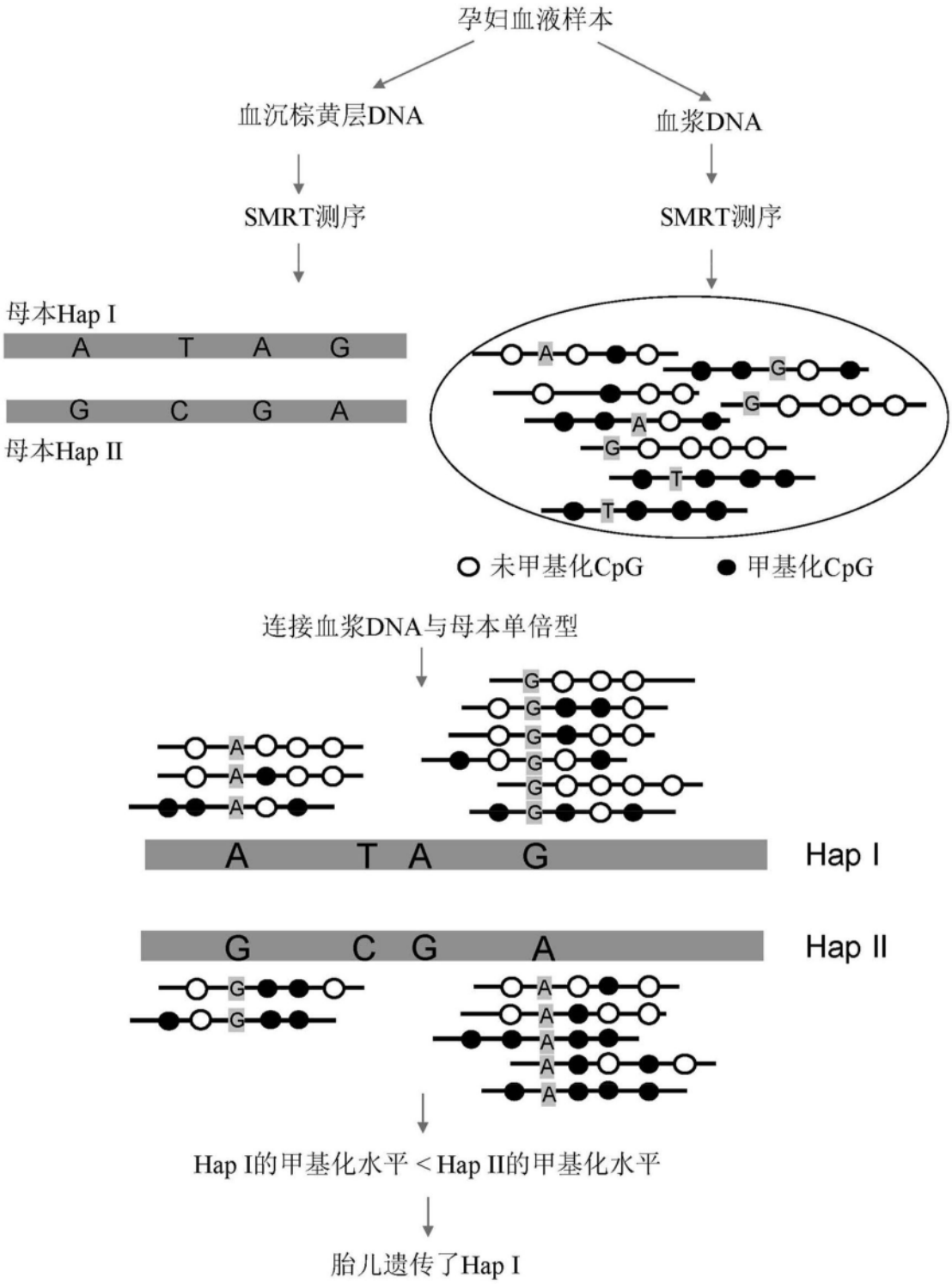

孕妇血液样本
血沉棕黄层DNA
血浆DNA
SMRT测序
SMRT测序
母本Hap I
A T A G
G C G A
母本Hap II
未甲基化CpG
甲基化CpG
连接血浆DNA与母本单倍型
A T A G
Hap I
G C G A
Hap II
Hap I的甲基化水平 < Hap II的甲基化水平
胎儿遗传了Hap I

图108

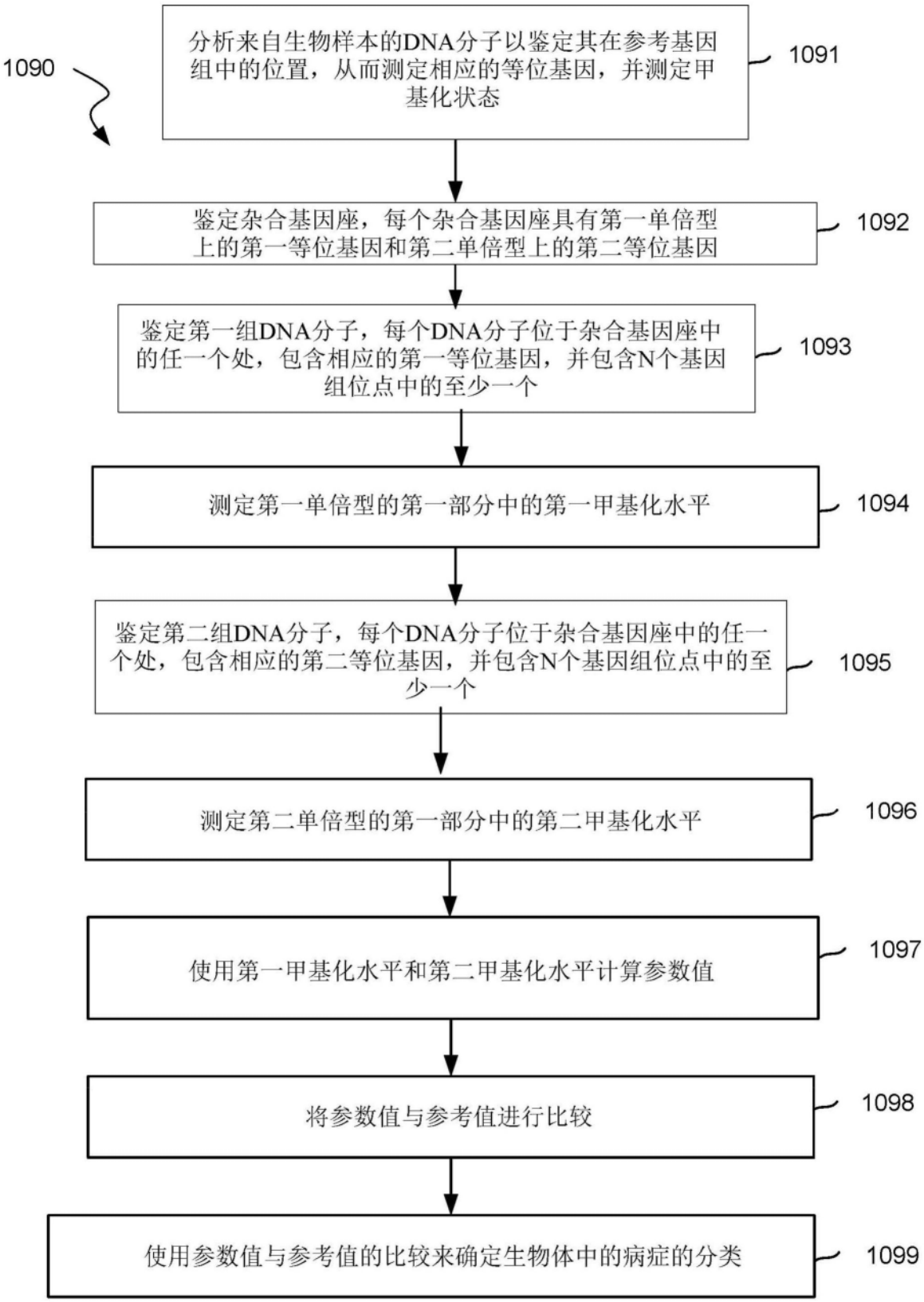

1090
分析来自生物样本的DNA分子以鉴定其在参考基因组中的位置，从而测定相应的等位基因，并测定甲基化状态
1091
鉴定杂合基因座，每个杂合基因座具有第一单倍型上的第一等位基因和第二单倍型上的第二等位基因
1092
鉴定第一组DNA分子，每个DNA分子位于杂合基因座中的任一个处，包含相应的第一等位基因，并包含N个基因组位点中的至少一个
1093
测定第一单倍型的第一部分中的第一甲基化水平
1094
鉴定第二组DNA分子，每个DNA分子位于杂合基因座中的任一个处，包含相应的第二等位基因，并包含N个基因组位点中的至少一个
1095
测定第二单倍型的第一部分中的第二甲基化水平
1096
使用第一甲基化水平和第二甲基化水平计算参数值
1097
将参数值与参考值进行比较
1098
使用参数值与参考值的比较来确定生物体中的病症的分类
1099

图109

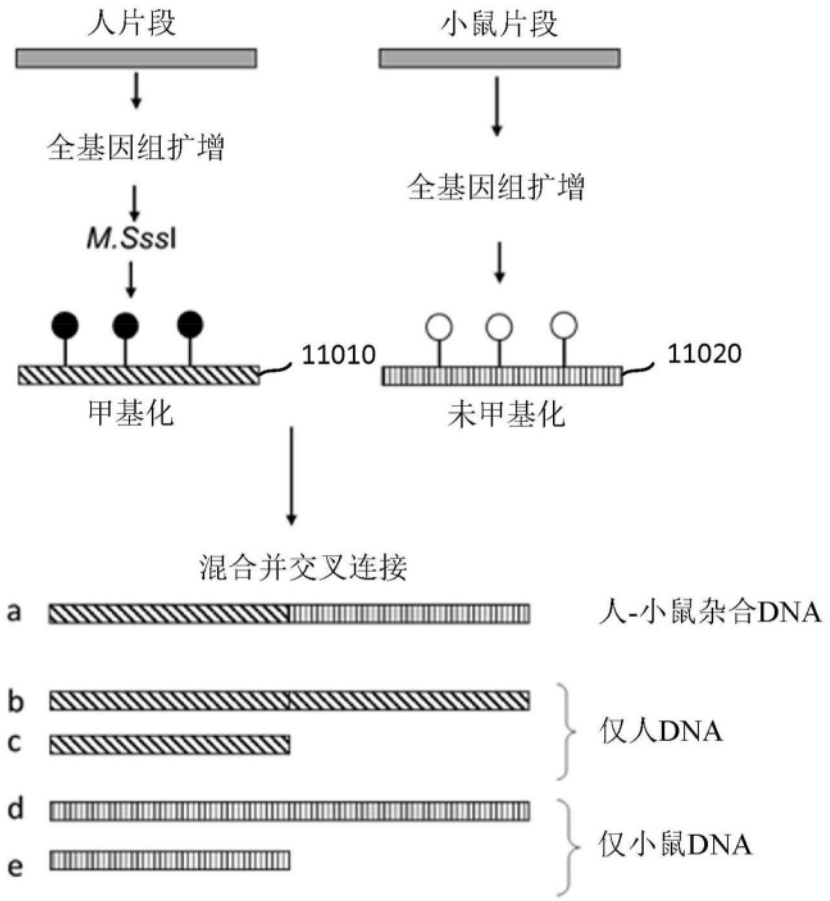

人片段
小鼠片段
全基因组扩增
全基因组扩增
M.SssI
11010
11020
甲基化
未甲基化
混合并交叉连接
a
人-小鼠杂合DNA
b
c
仅人DNA
d
e
仅小鼠DNA

图110

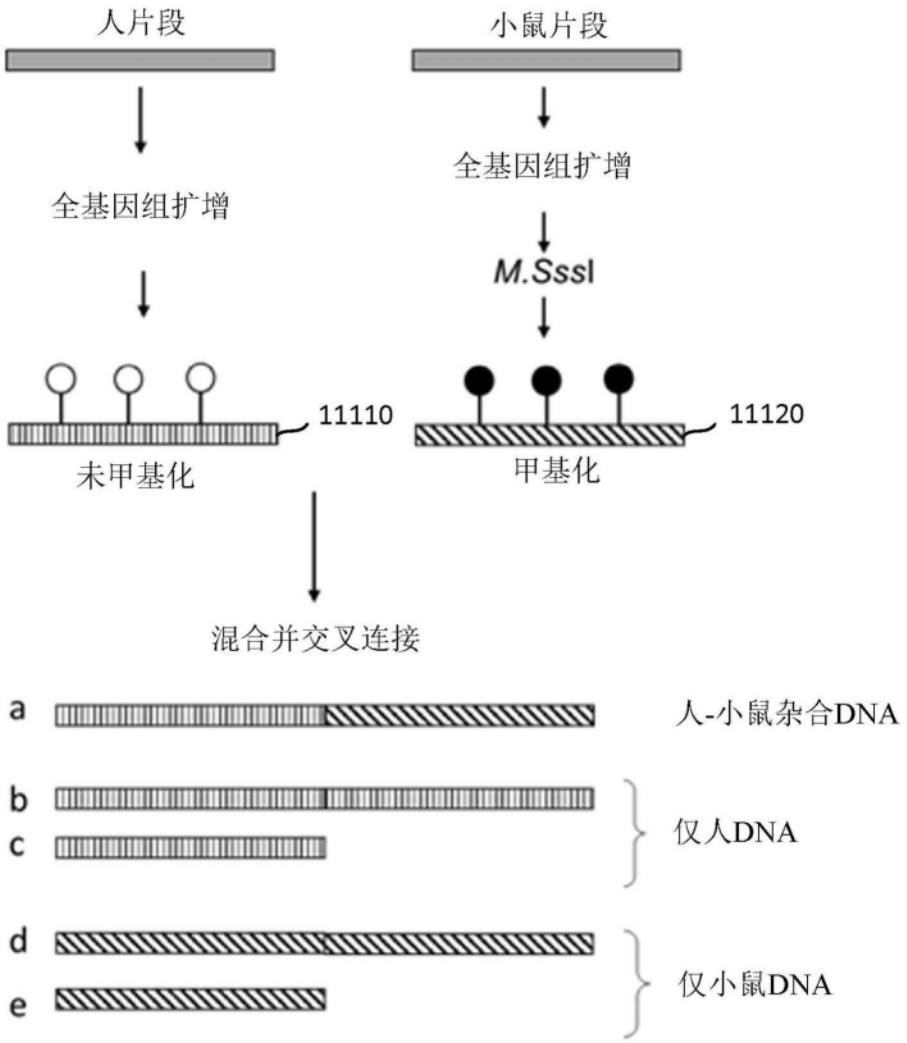

人片段
小鼠片段
全基因组扩增
全基因组扩增
M.SssI
11110
11120
未甲基化
甲基化
混合并交叉连接
a
人-小鼠杂合DNA
b
c
仅人DNA
d
e
仅小鼠DNA

图111

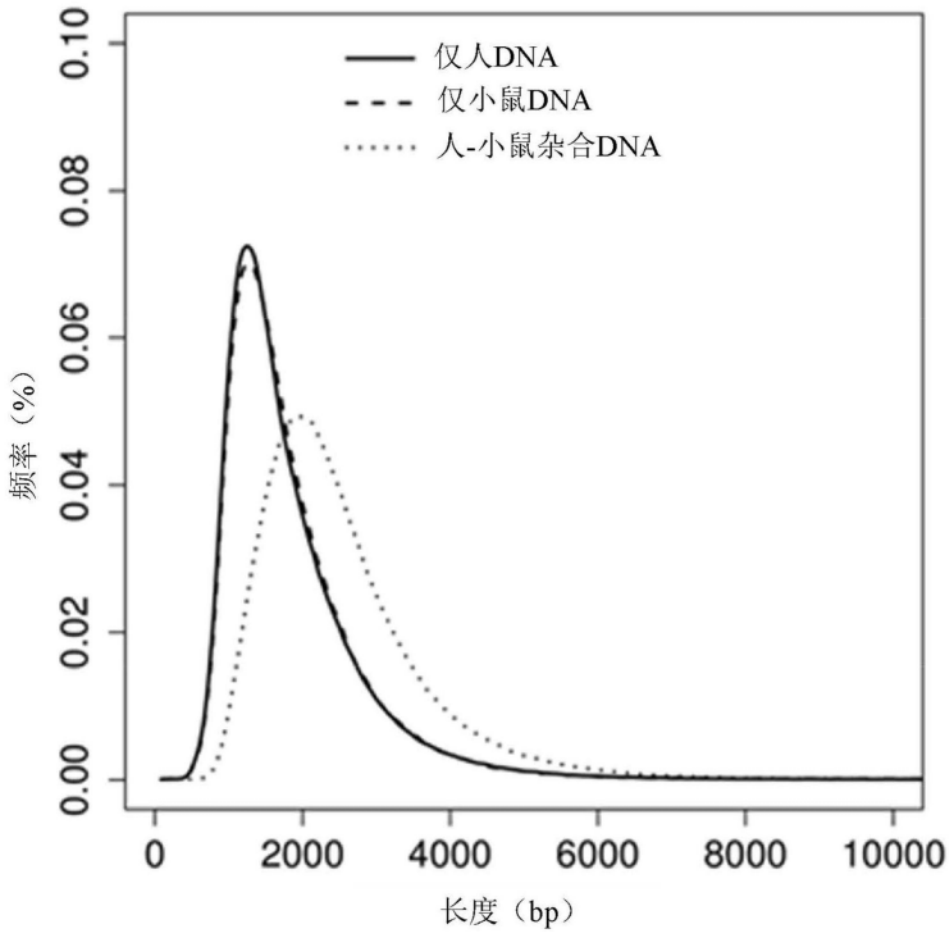

仅人DNA
仅小鼠DNA
人-小鼠杂合DNA
频率（%）
0.00
0.02
0.04
0.06
0.08
0.10
0
2000
4000
6000
8000
10000
长度（bp）

图112

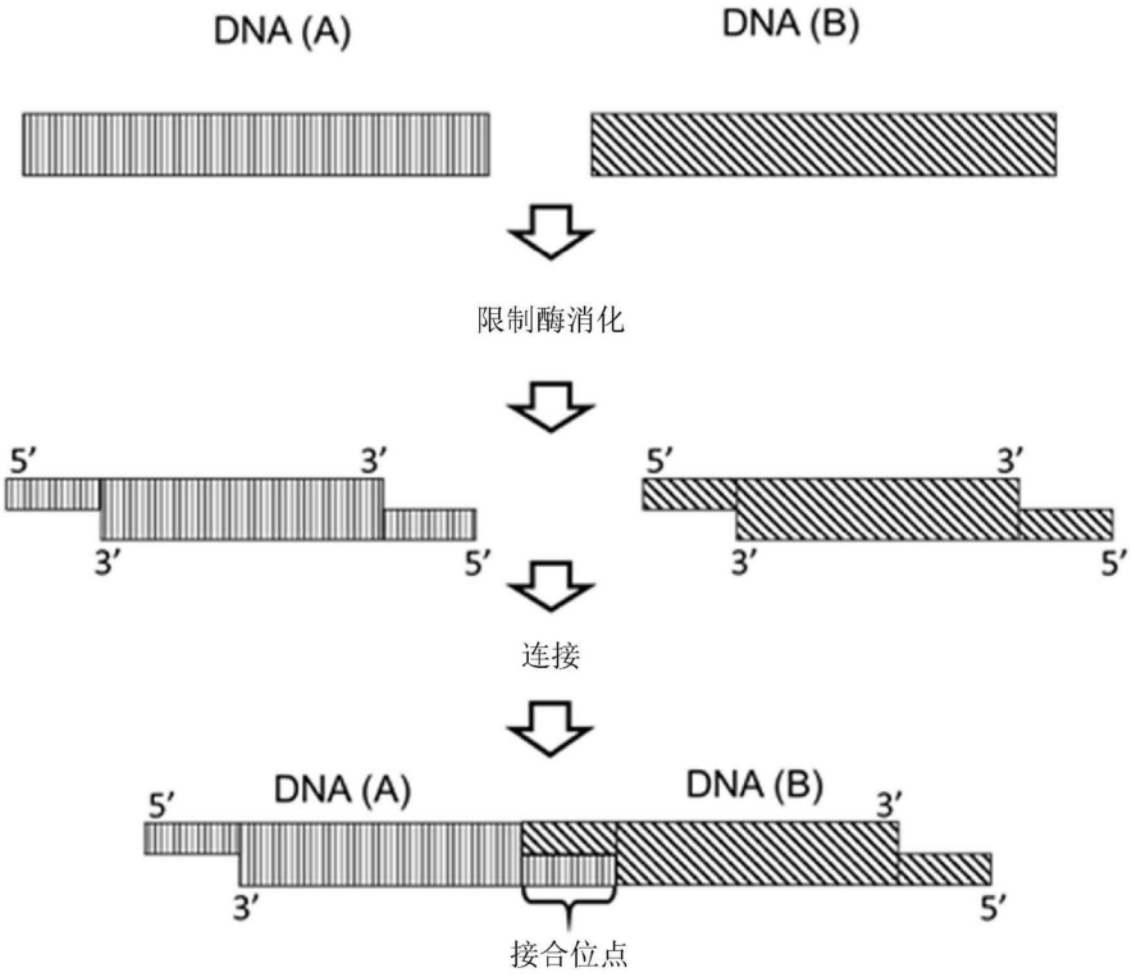


图113

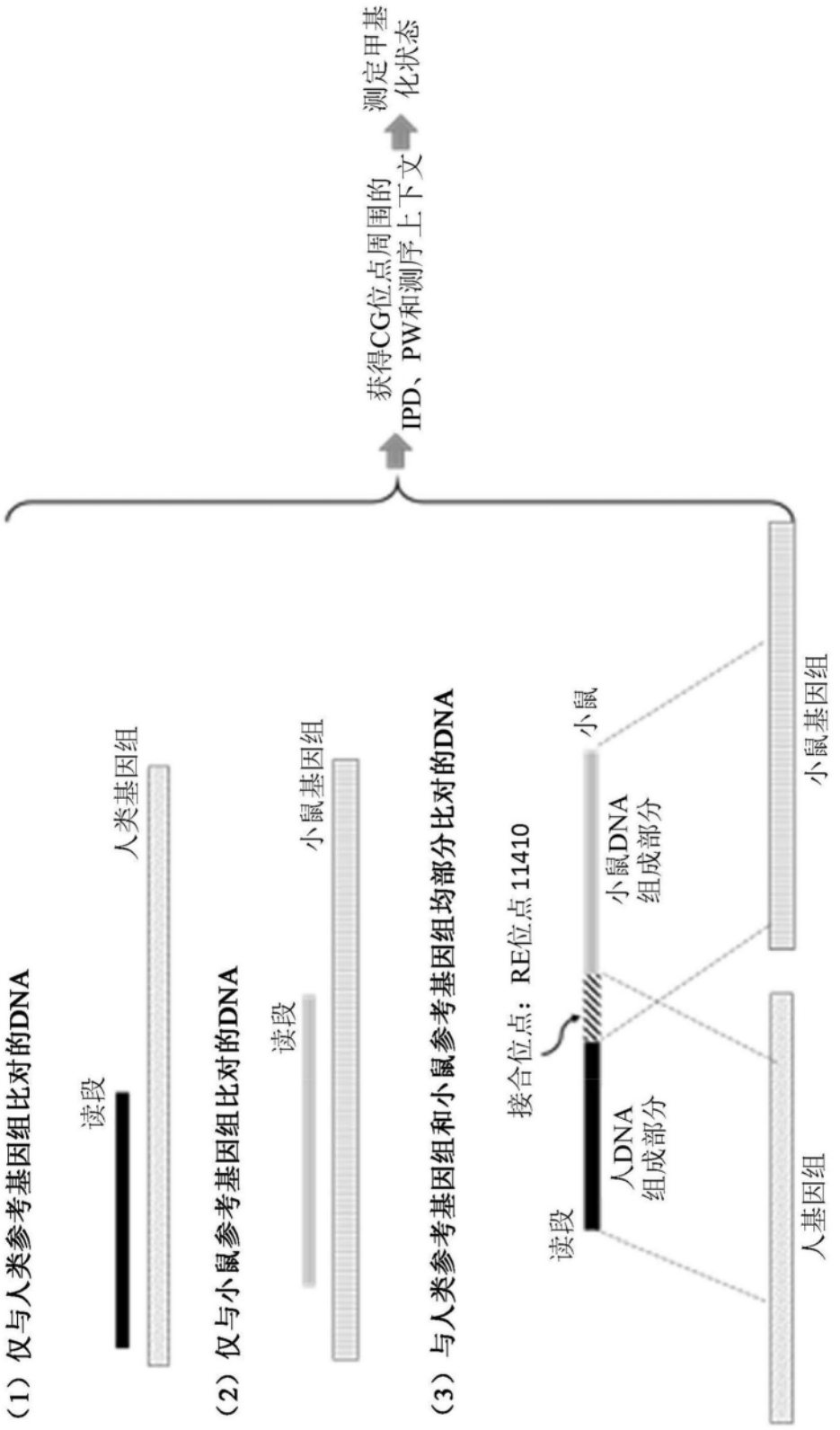

（1）仅与人类参考基因组比对的DNA
读段
人类基因组
（2）仅与小鼠参考基因组比对的DNA
读段
小鼠基因组
（3）与人类参考基因组和小鼠参考基因组均部分比对的DNA
接合位点：RE位点11410
读段
人DNA
组成部分
小鼠
小鼠DNA
组成部分
人基因组
小鼠基因组
获得CG位点周围的
IPD、PW和测序上下文
测定甲基
化状态

图114

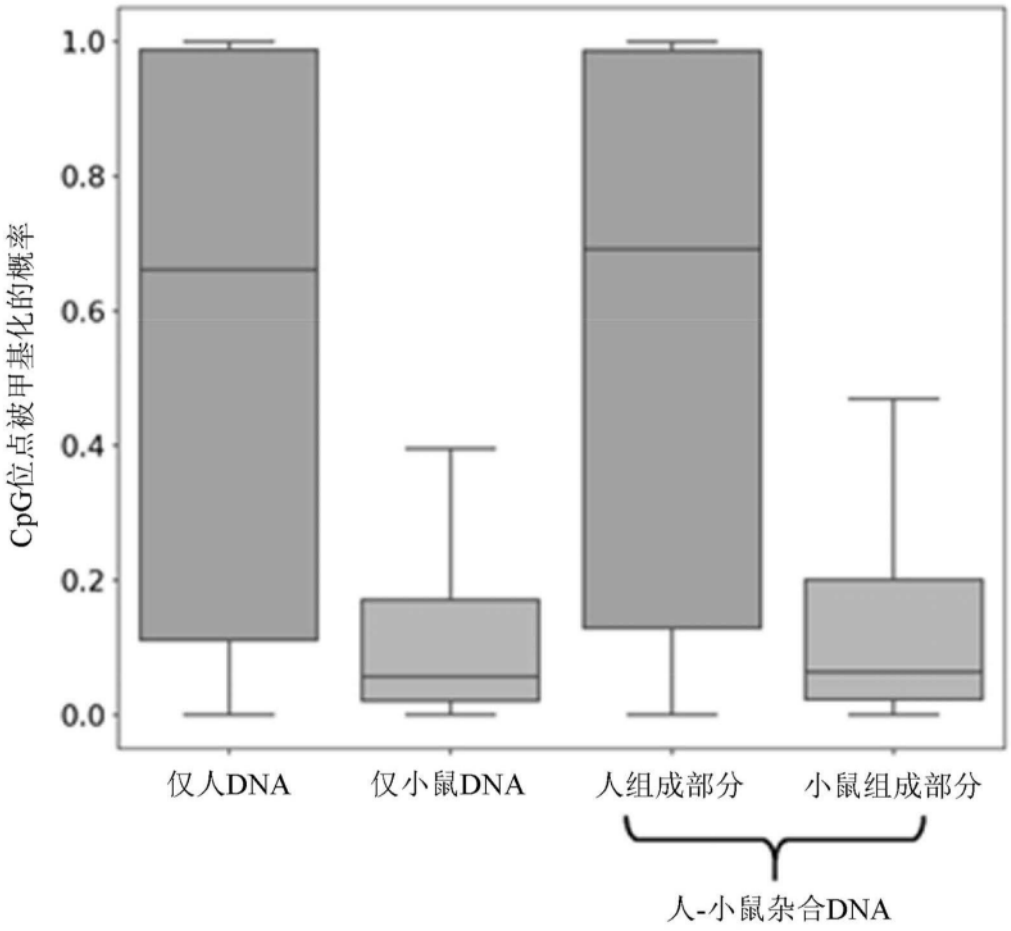

1.0
0.8
0.6
0.4
0.2
0.0
CpG位点被甲基化的概率
仅人DNA
仅小鼠DNA
人组成部分
小鼠组成部分
人-小鼠杂合DNA

图115

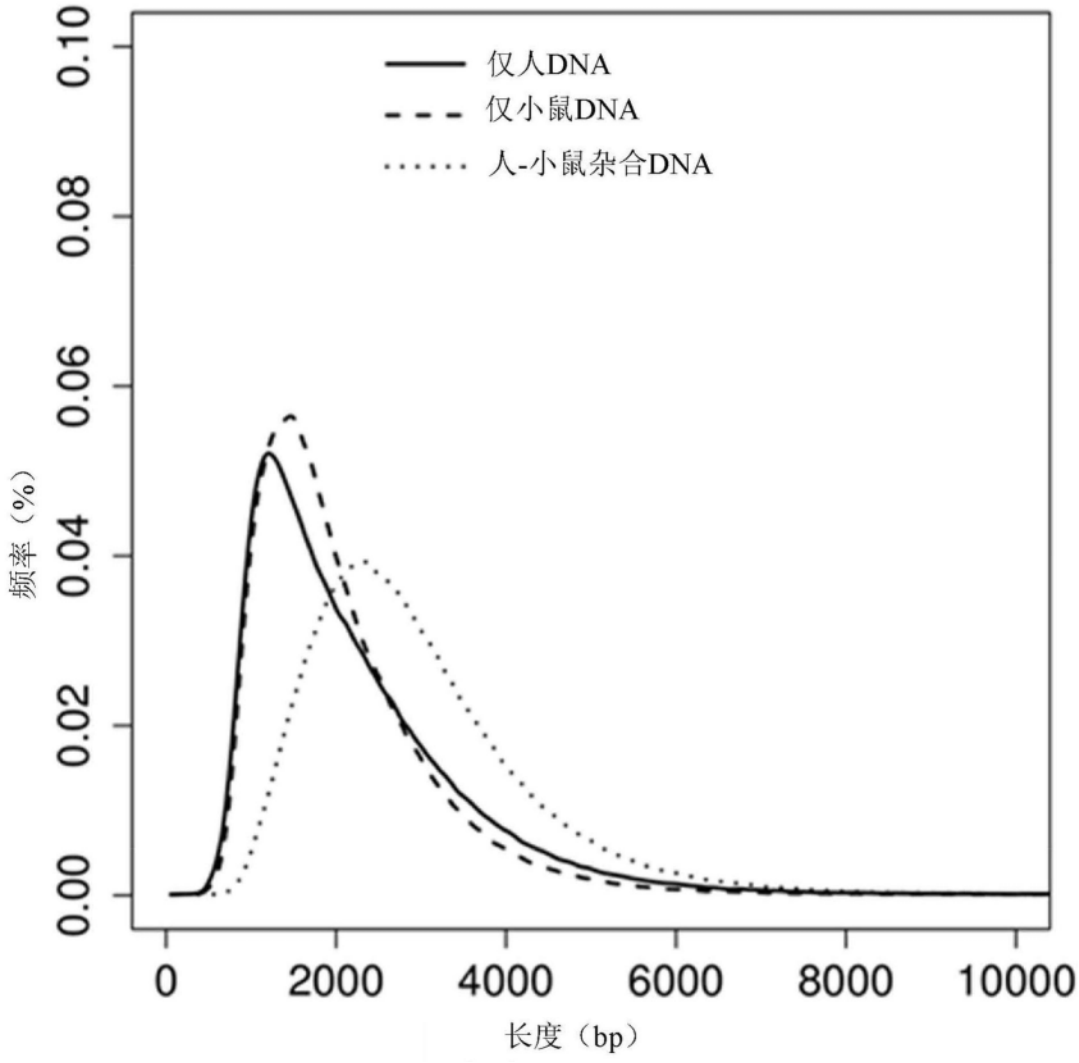

仅人DNA
仅小鼠DNA
人-小鼠杂合DNA
频率（%）
0.00
0.02
0.04
0.06
0.08
0.10
0
2000
4000
6000
8000
10000
长度（bp）

图116

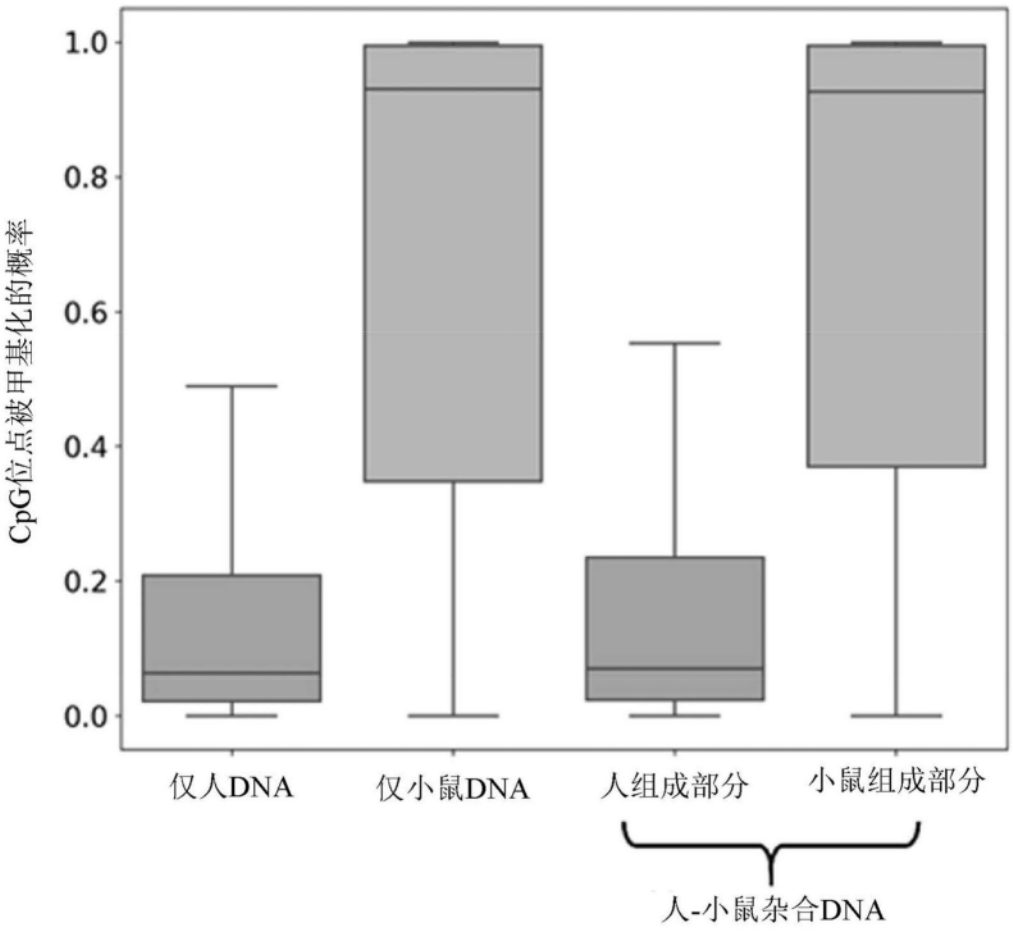
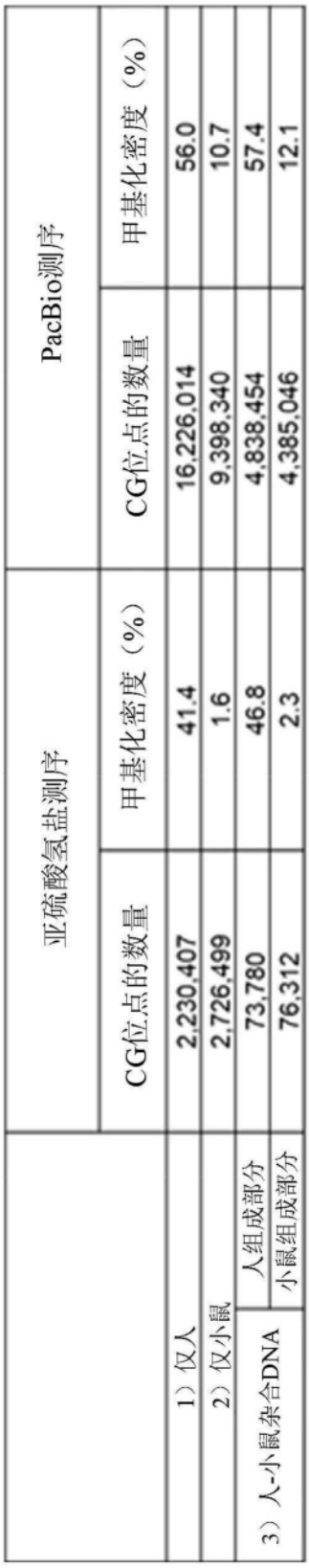
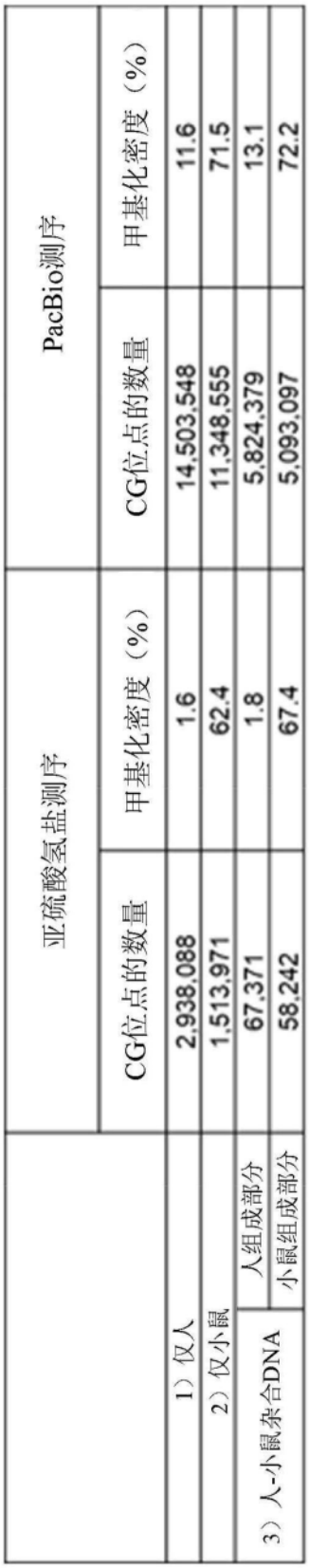

1.0
0.8
0.6
0.4
0.2
0.0
CpG位点被甲基化的概率
仅人DNA
仅小鼠DNA
人组成部分
小鼠组成部分
人-小鼠杂合DNA

图117



		亚硫酸氢盐测序		PacBio测序	
		CG位点的数量	甲基化密度（%）	CG位点的数量	甲基化密度（%）
1）仅人		2,230,407	41.4	16,226,014	56.0
2）仅小鼠		2,726,499	1.6	9,398,340	10.7
3）人-小鼠杂合DNA	人组成部分	73,780	46.8	4,838,454	57.4
	小鼠组成部分	76,312	2.3	4,385,046	12.1

图118



		亚硫酸氢盐测序		PacBio测序	
		CG位点的数量	甲基化密度（%）	CG位点的数量	甲基化密度（%）
1）仅人		2,938,088	1.6	14,503,548	11.6
2）仅小鼠		1,513,971	62.4	11,348,555	71.5
3）人-小鼠杂合DNA	人组成部分	67,371	1.8	5,824,379	13.1
	小鼠组成部分	58,242	67.4	5,093,097	72.2

图119

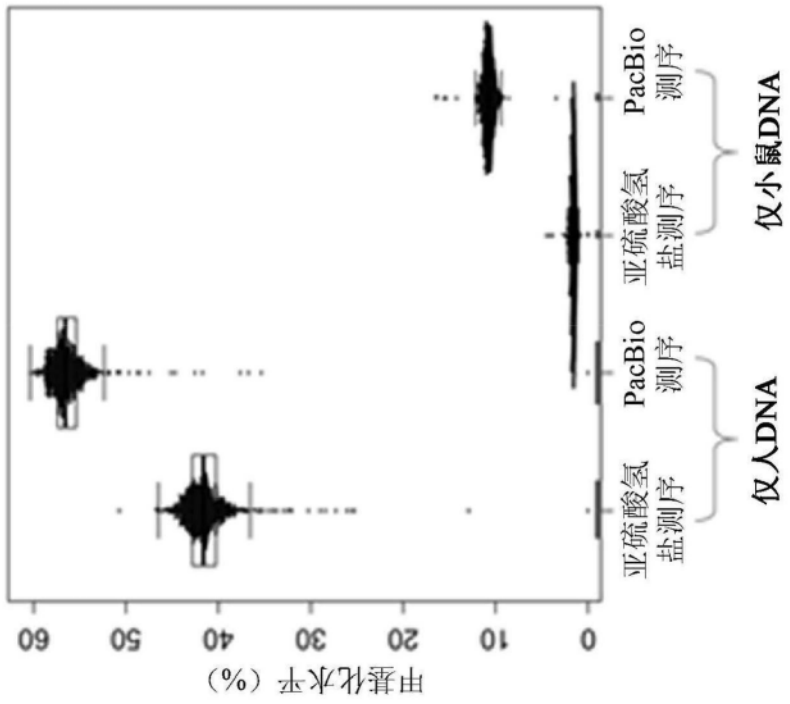

甲基化水平（%）
0
10
20
30
40
50
60
亚硫酸氢盐测序
PacBio测序
仅人DNA
亚硫酸氢盐测序
PacBio测序
仅小鼠DNA

图120A

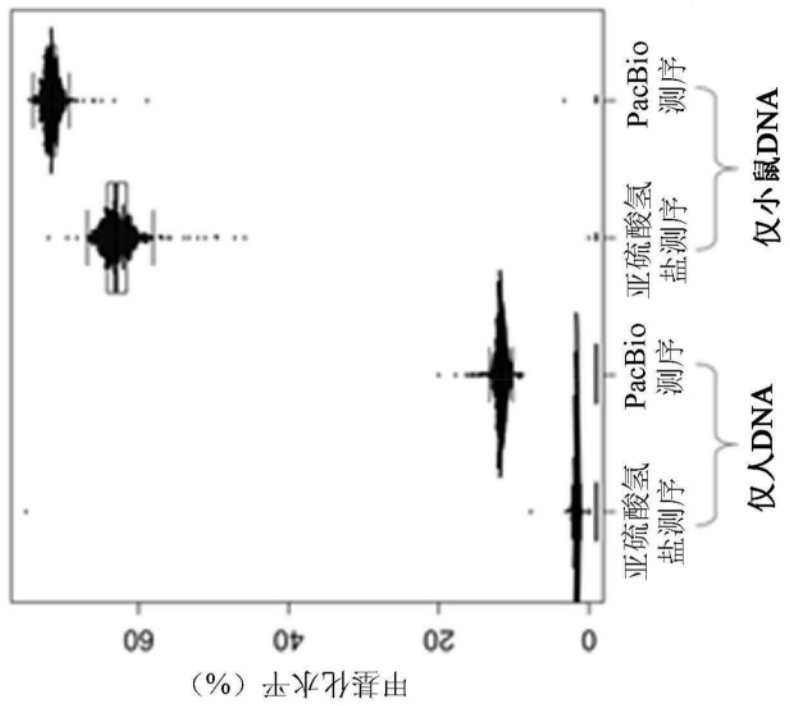

甲基化水平（%）
0
20
40
60
亚硫酸氢盐测序
PacBio测序
仅人DNA
亚硫酸氢盐测序
PacBio测序
仅小鼠DNA

图120B

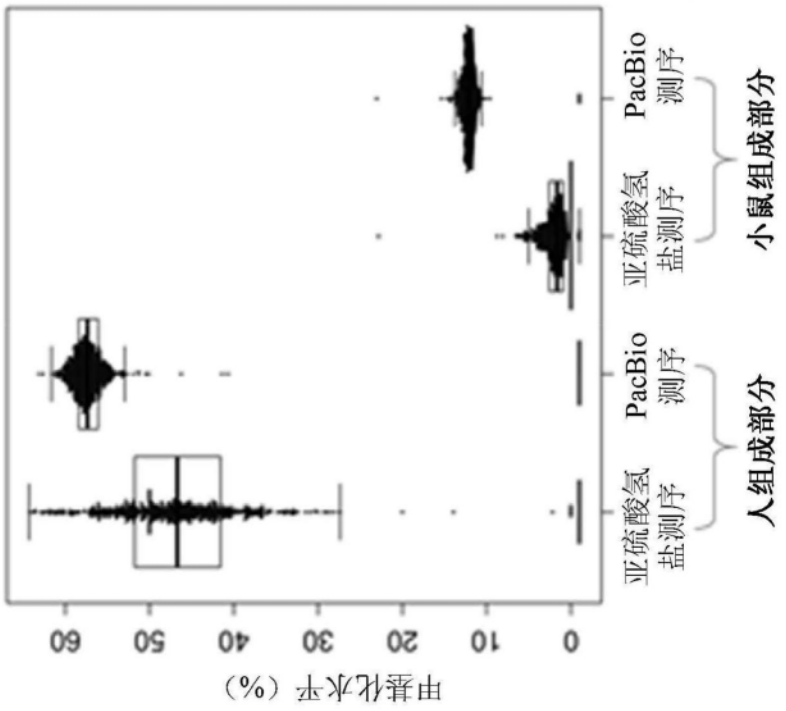

甲基化水平（%）
0
10
20
30
40
50
60
亚硫酸氢盐测序
PacBio测序
亚硫酸氢盐测序
PacBio测序
人组成部分
小鼠组成部分

图121A

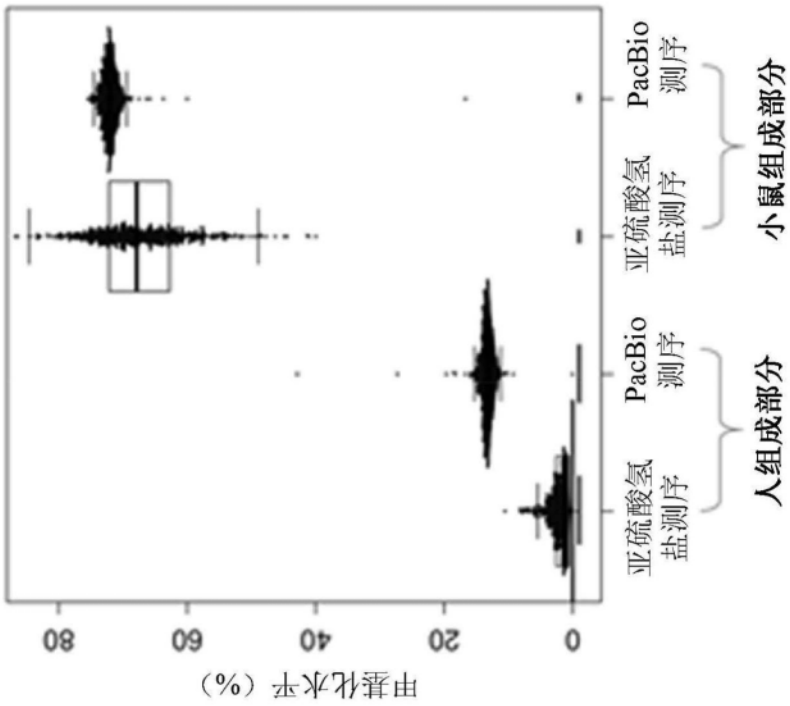

甲基化水平（%）
0
20
40
60
80
亚硫酸氢盐测序
PacBio测序
亚硫酸氢盐测序
PacBio测序
人组成部分
小鼠组成部分

图121B

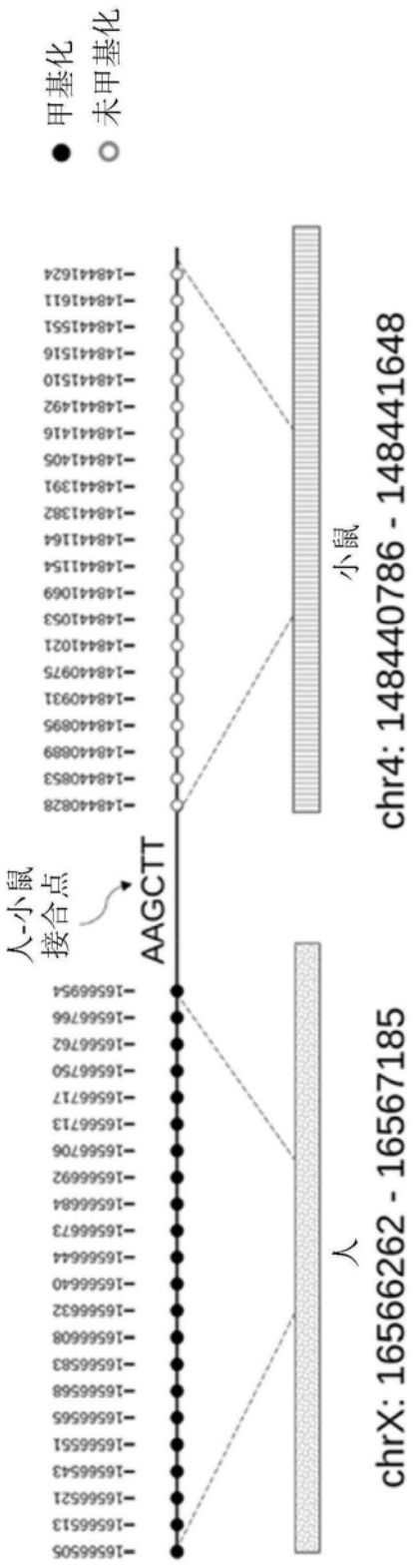

甲基化
未甲基化
小鼠
chr4: 148440786 - 148441648
人
chrX: 16566262 - 16567185
人-小鼠
接合点
AAGCTT

图122A

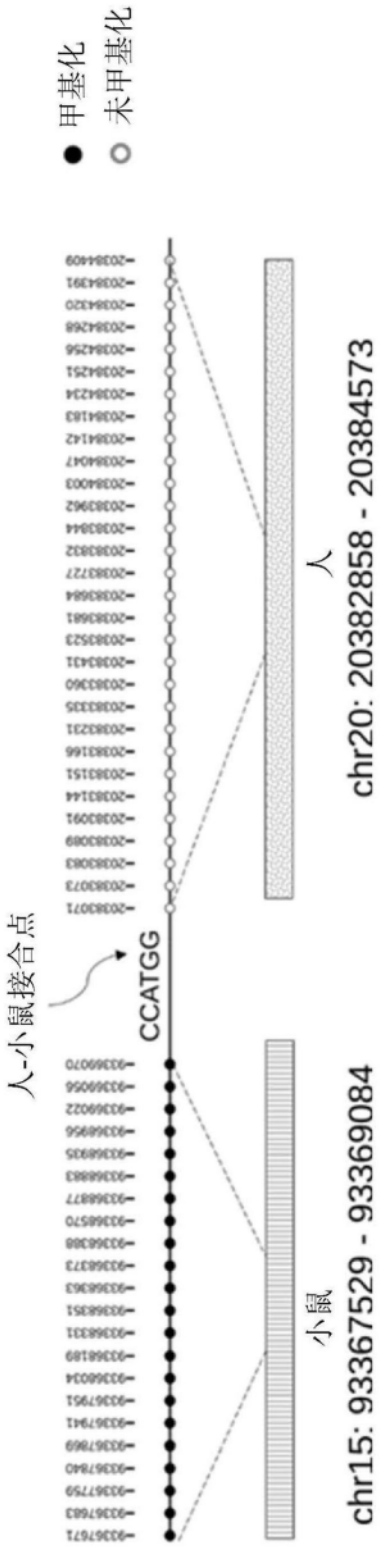

甲基化
未甲基化
人-小鼠接合点
CCATGG
人
chr20: 20382858 - 20384573
小鼠
chr15: 93367529 - 93369084

图122B

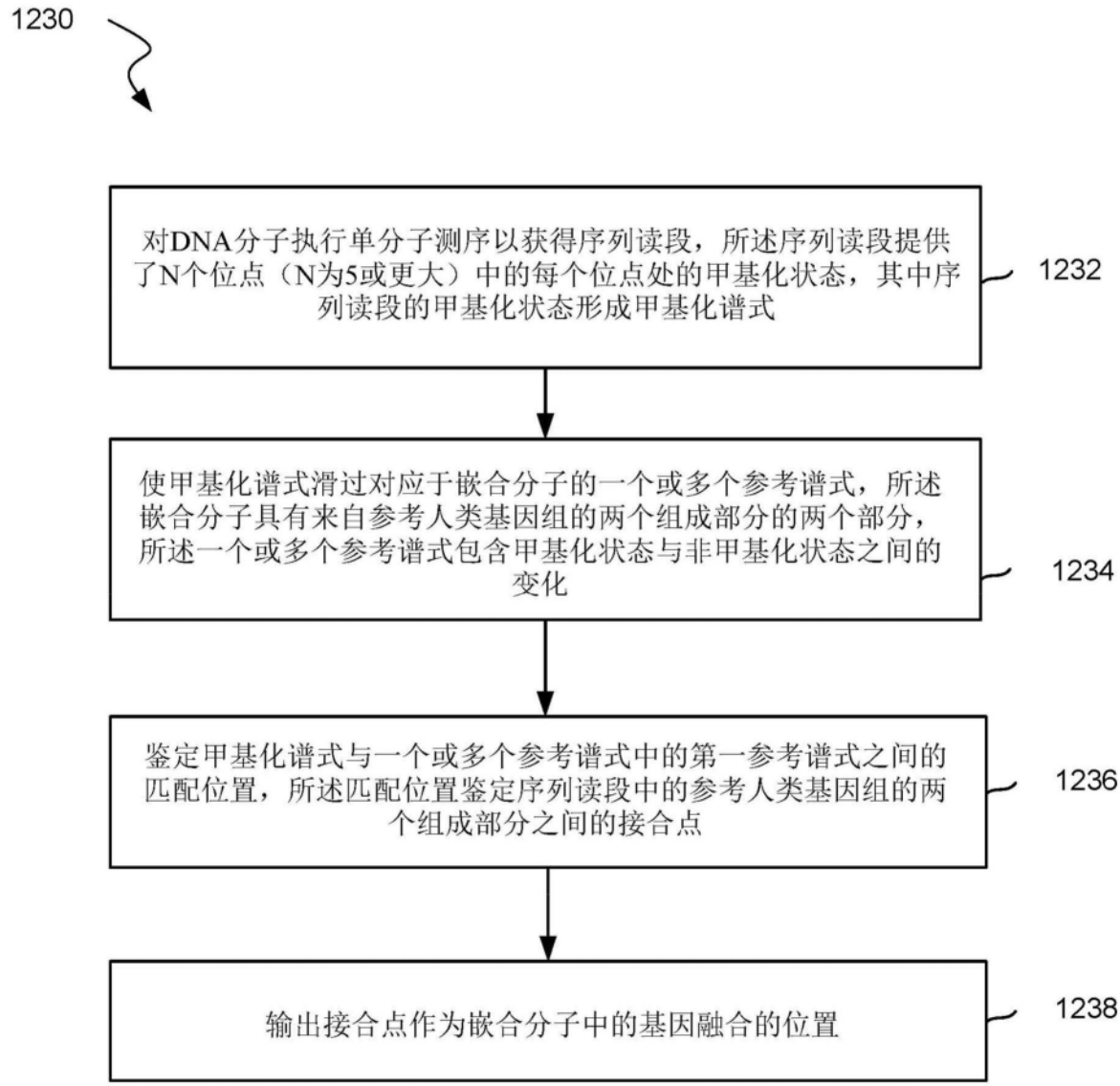

1230
对DNA分子执行单分子测序以获得序列读段，所述序列读段提供了N个位点（N为5或更大）中的每个位点处的甲基化状态，其中序列读段的甲基化状态形成甲基化谱式
1232
使甲基化谱式滑过对应于嵌合分子的一个或多个参考谱式，所述嵌合分子具有来自参考人类基因组的两个组成部分的两个部分，所述一个或多个参考谱式包含甲基化状态与非甲基化状态之间的变化
1234
鉴定甲基化谱式与一个或多个参考谱式中的第一参考谱式之间的匹配位置，所述匹配位置鉴定序列读段中的参考人类基因组的两个组成部分之间的接合点
1236
输出接合点作为嵌合分子中的基因融合的位置
1238

图123

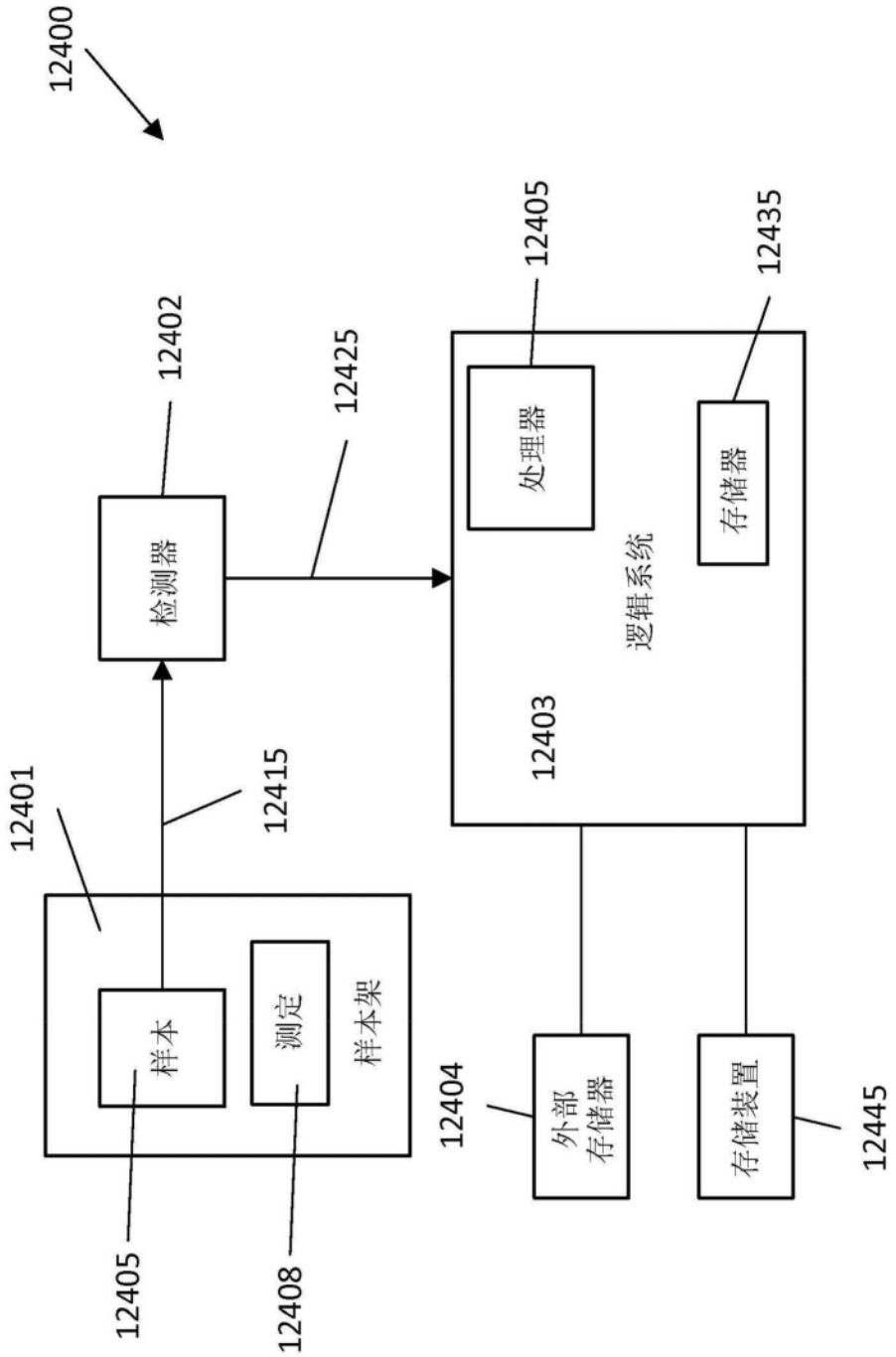

12400
12402
12425
12405
12435
检测器
处理器
存储器
逻辑系统
12403
12401
12415
样本
测定
样本架
12405
12408
12404
外部
存储器
存储装置
12445

图124

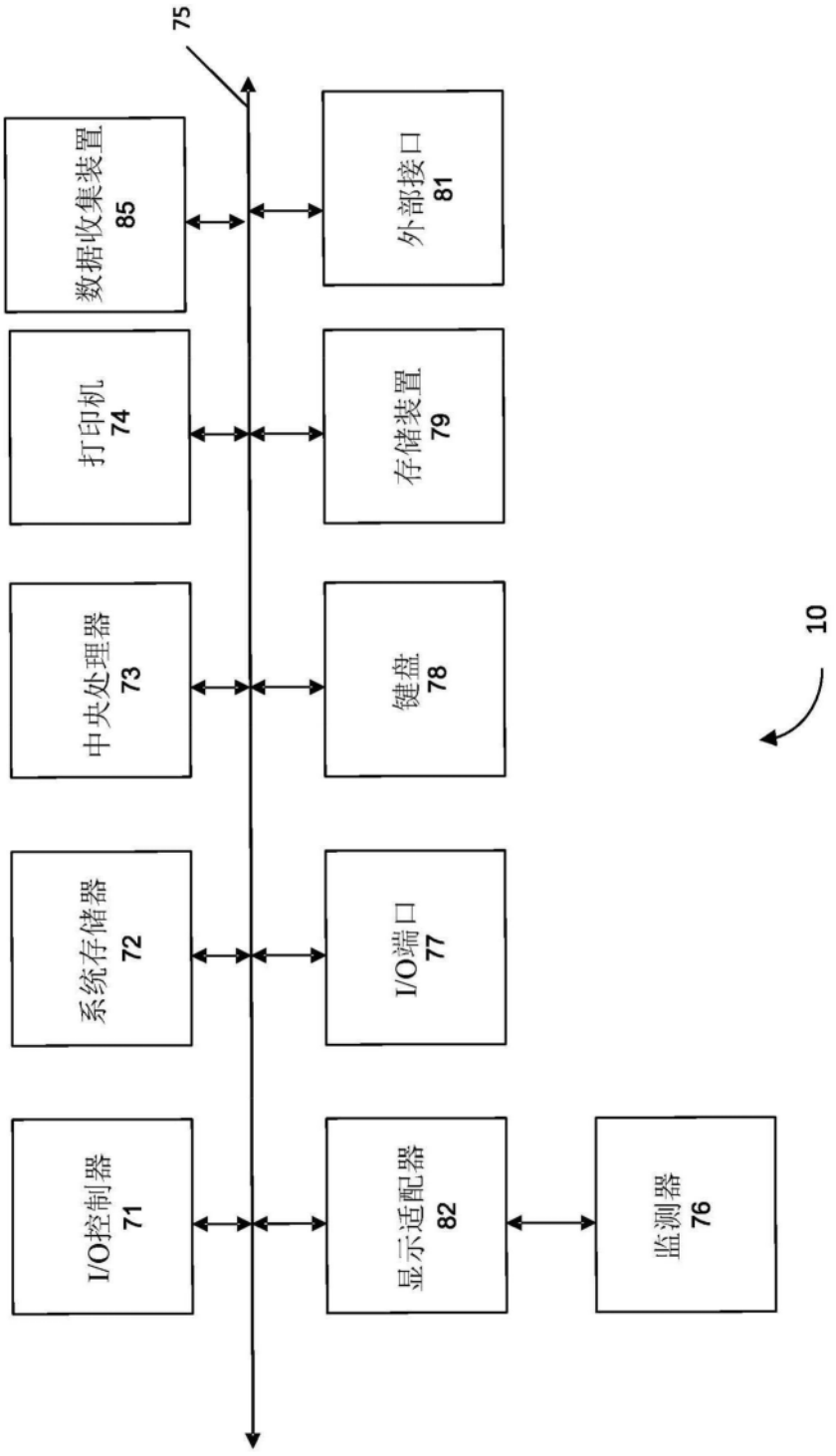

75
数据收集装置
85
外部接口
81
打印机
74
存储装置
79
中央处理器
73
键盘
78
10
系统存储器
72
I/O端口
77
I/O控制器
71
显示适配器
82
监测器
76

图125

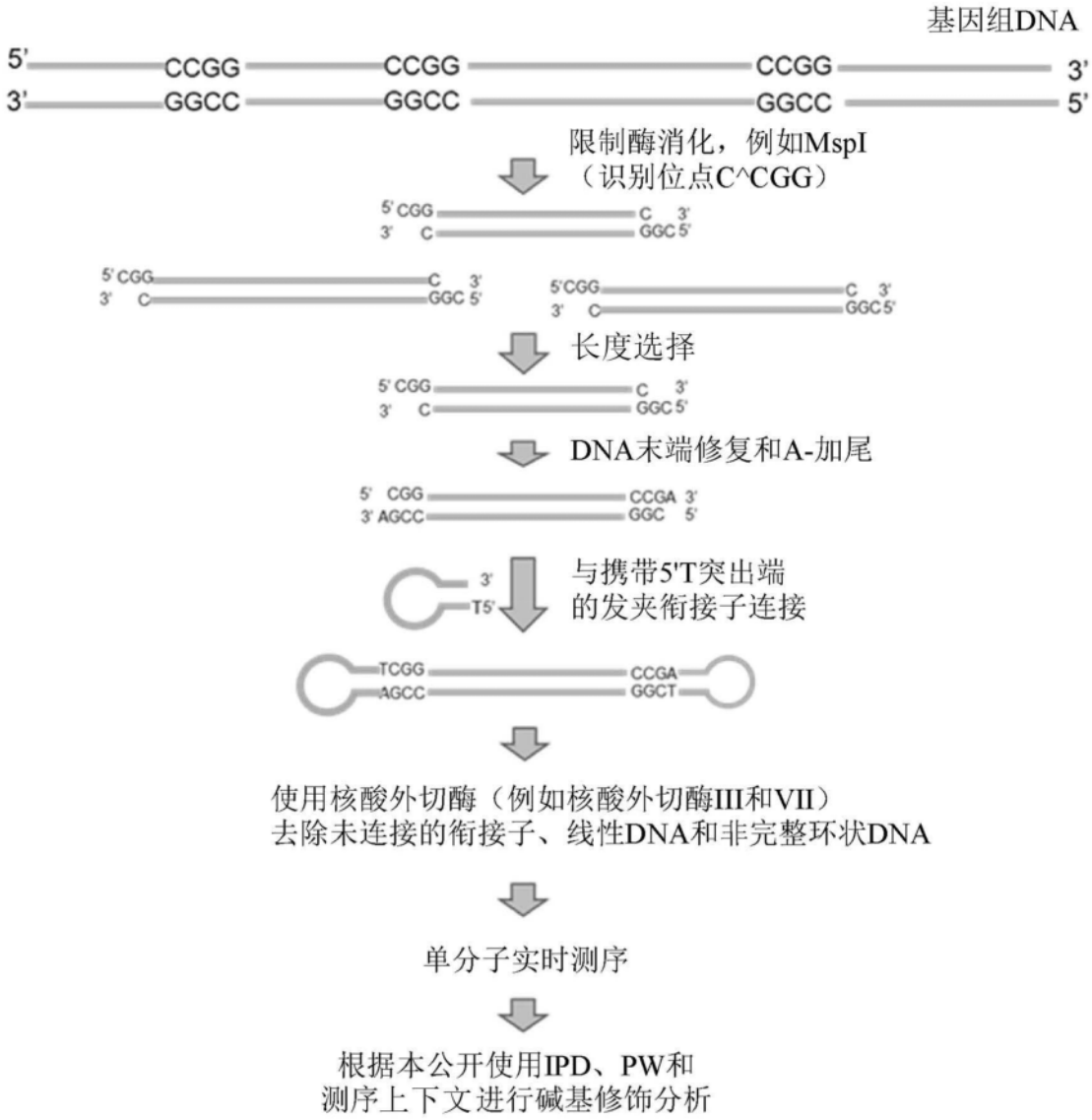

基因组DNA
5'
CCGG
CCGG
CCGG
3'
3'
GGCC
GGCC
GGCC
5'
限制酶消化，例如MspI
（识别位点C^CGG）
长度选择
DNA末端修复和A-加尾
与携带5'T突出端
的发夹衔接子连接
使用核酸外切酶（例如核酸外切酶III和VII）
去除未连接的衔接子、线性DNA和非完整环状DNA
单分子实时测序
根据本公开使用IPD、PW和
测序上下文进行碱基修饰分析

图126

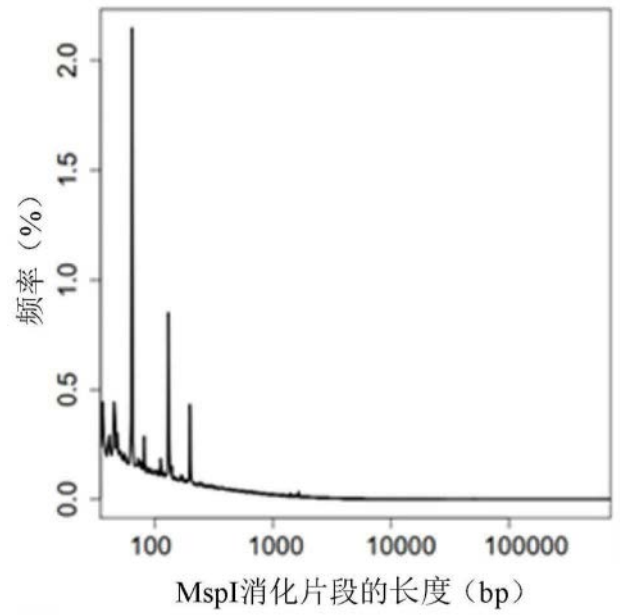

频率（%）
0.0
0.5
1.0
1.5
2.0
100
1000
10000
100000
MspI消化片段的长度（bp）

图127A

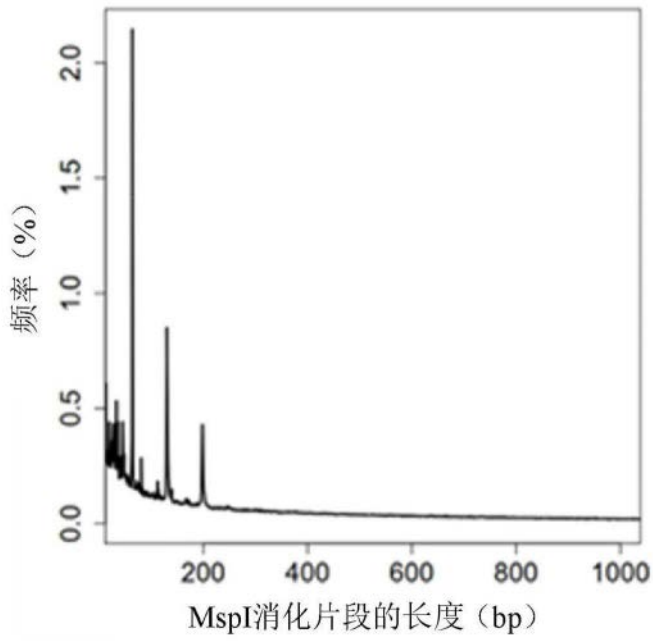

频率（%）
0.0
0.5
1.0
1.5
2.0
200
400
600
800
1000
MspI消化片段的长度（bp）

图127B

长度范围（bp）	分子数量	一定长度范围内的分子片段总数的百分比（%）	一定长度范围内的与CpG岛重叠的分子的数量	一定长度范围内的与CpG岛重叠的分子的百分比（%）	被测序的CpG位点的数量	落入CpG岛内的CpG位点的数量	通过尺寸选择靶向的和落入CpG岛内的CpG位点的百分比（%）
50-200	526,543	23.03	104,059	19.76	2,358,020	885,041	37.53
200-400	269,562	11.79	23,927	8.88	1,781,556	353,087	19.82
400-600	177,776	7.77	7,369	4.15	1,468,561	107,130	7.29
600-800	133,927	5.86	3,673	2.74	1,326,544	48,851	3.68
800-1000	104,976	4.59	2,168	2.07	1,193,233	25,821	2.16
1000-2000	311,596	13.63	4,596	1.47	4,610,504	58,288	1.26
2000-3000	149,468	6.54	1,771	1.18	3,036,951	25,106	0.83
3000-4000	86,760	3.79	809	0.93	2,165,171	10,785	0.50
5000-6000	36,931	1.62	266	0.72	1,242,712	3,412	0.27
6000-7000	25,027	1.09	202	0.81	947,874	3,354	0.35
7000-8000	17,597	0.77	86	0.49	736,830	791	0.11
8000-9000	12,658	0.55	76	0.60	583,680	993	0.17
9000-10000	9,184	0.40	48	0.52	461,935	591	0.13
10000-15000	20,790	0.91	97	0.47	1,255,731	2,003	0.16
15000-20000	5,111	0.22	16	0.31	414,400	163	0.04
20000-25000	1,441	0.06	6	0.42	147,731	34	0.02

图128

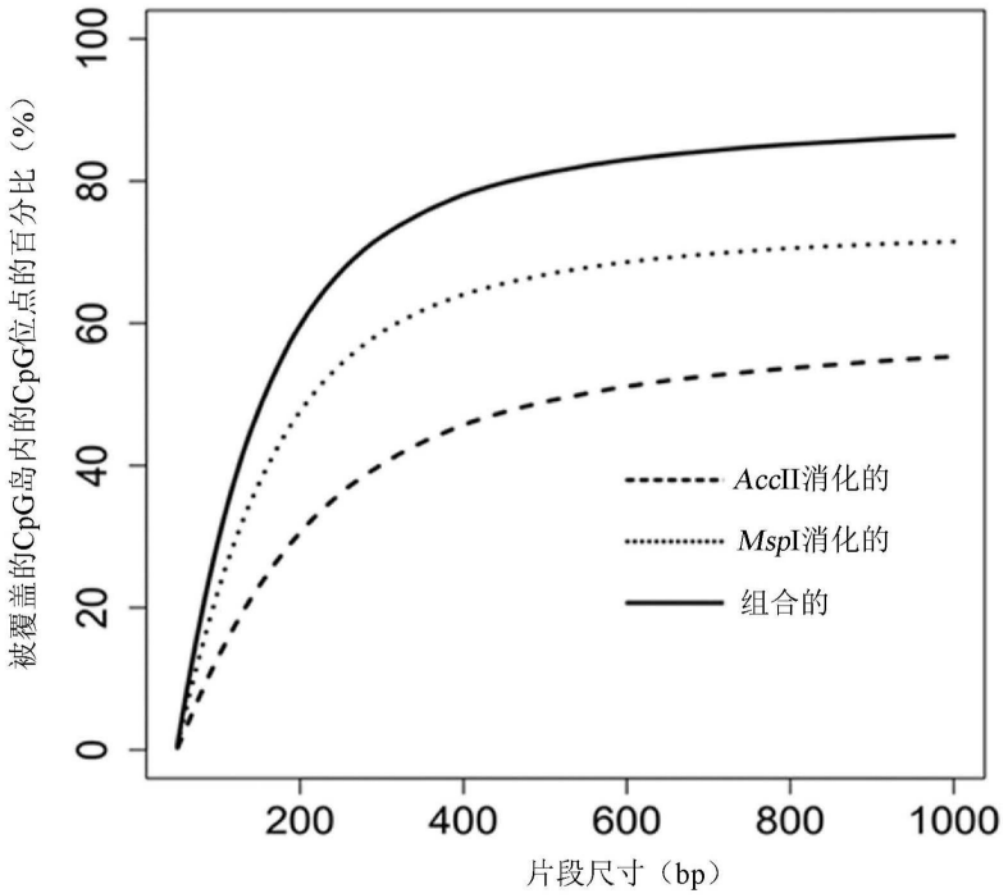

被覆盖的CpG岛内的CpG位点的百分比（%）
0
20
40
60
80
100
200
400
600
800
1000
片段尺寸（bp）
AccII消化的
MspI消化的
组合的

图129

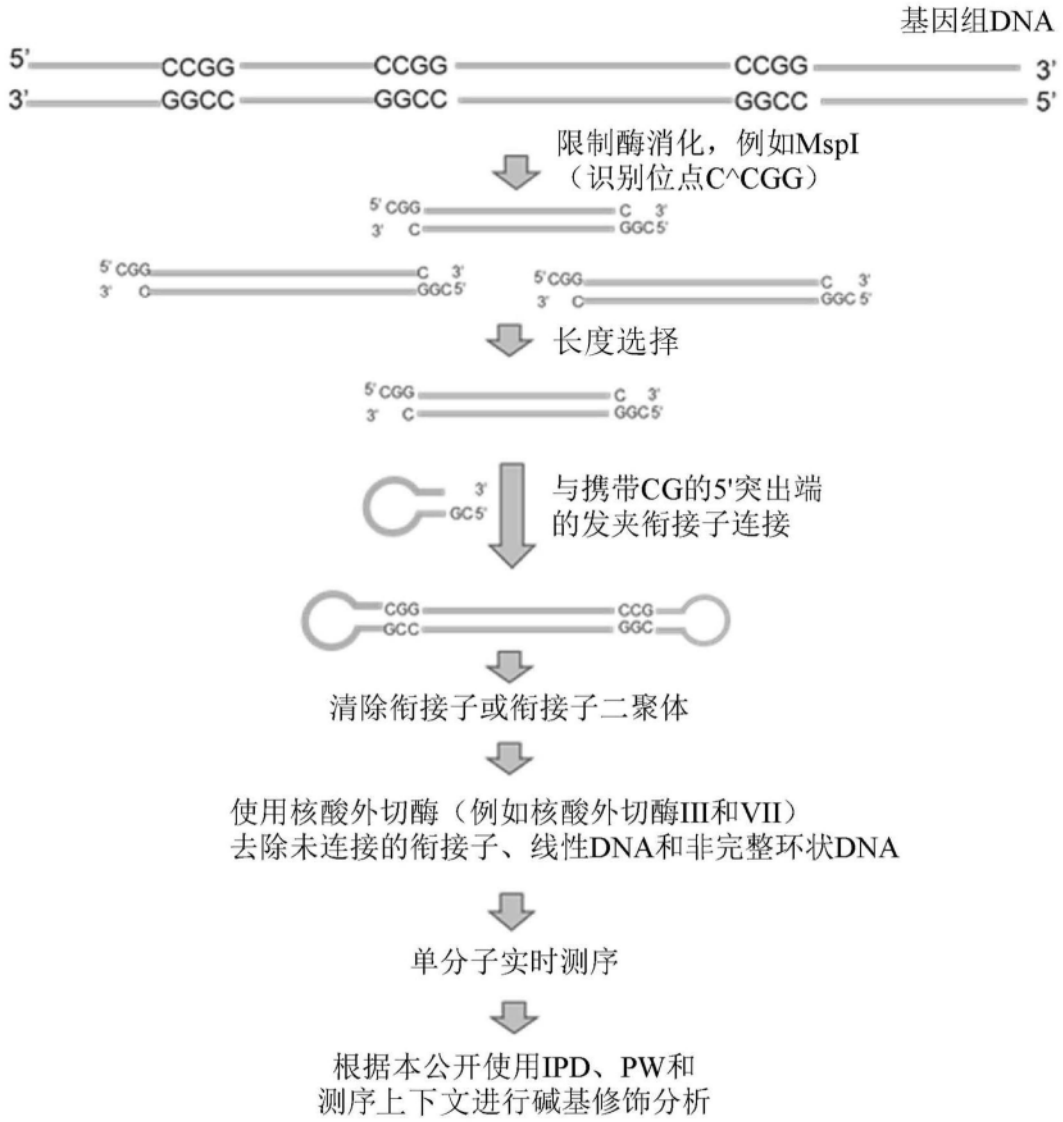

基因组DNA
5'
CCGG
CCGG
CCGG
3'
3'
GGCC
GGCC
GGCC
5'
限制酶消化，例如MspI
（识别位点C^CGG）
长度选择
与携带CG的5'突出端
的发夹衔接子连接
清除衔接子或衔接子二聚体
使用核酸外切酶（例如核酸外切酶III和VII）
去除未连接的衔接子、线性DNA和非完整环状DNA
单分子实时测序
根据本公开使用IPD、PW和
测序上下文进行碱基修饰分析

图130

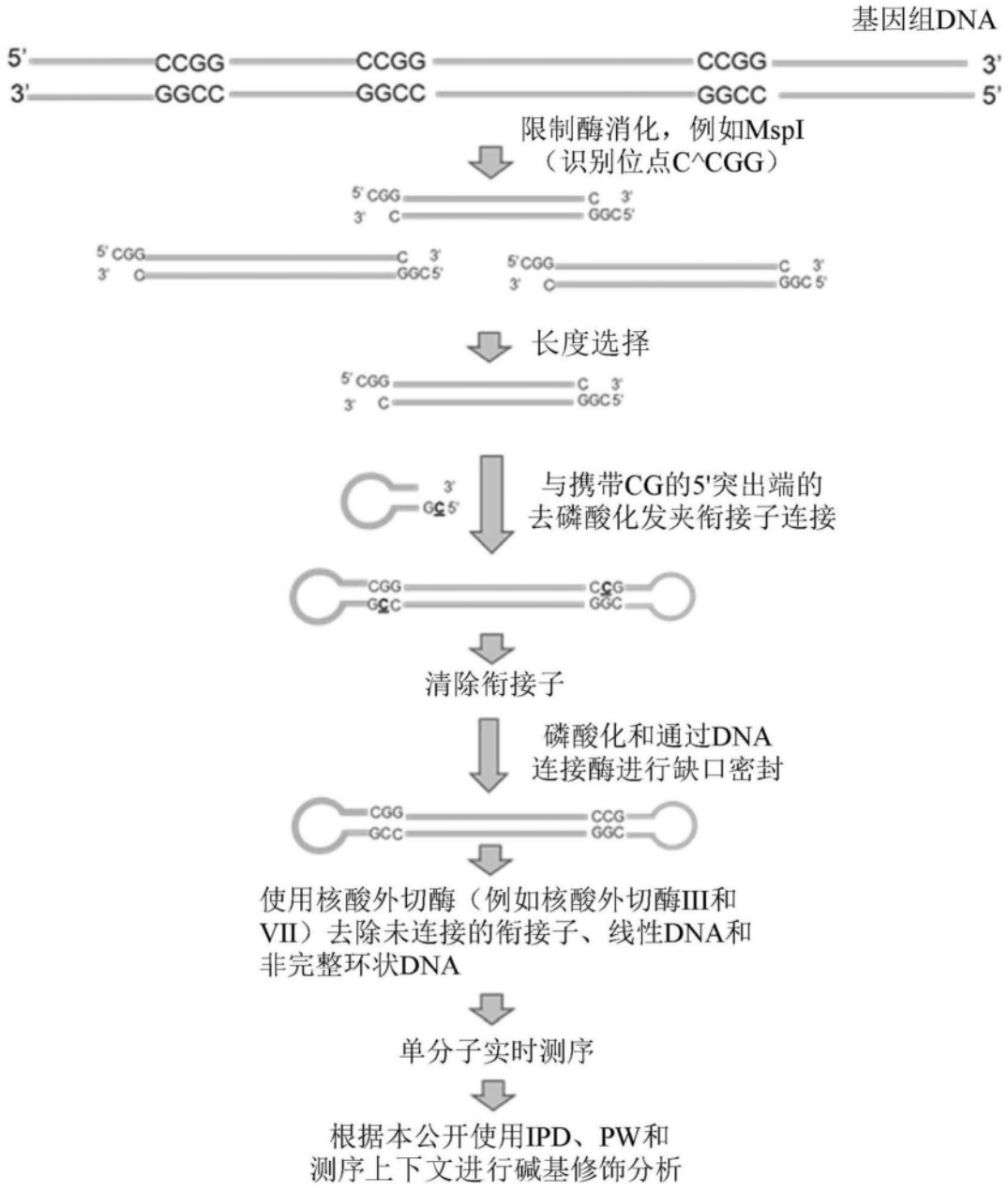

基因组DNA
5' CCGG CCGG CCGG 3'
3' GGCC GGCC GGCC 5'
限制酶消化，例如MspI
（识别位点C^CGG）
长度选择
与携带CG的5'突出端的
去磷酸化发夹衔接子连接
清除衔接子
磷酸化和通过DNA
连接酶进行缺口密封
使用核酸外切酶（例如核酸外切酶III和
VII）去除未连接的衔接子、线性DNA和
非完整环状DNA
单分子实时测序
根据本公开使用IPD、PW和
测序上下文进行碱基修饰分析

图131

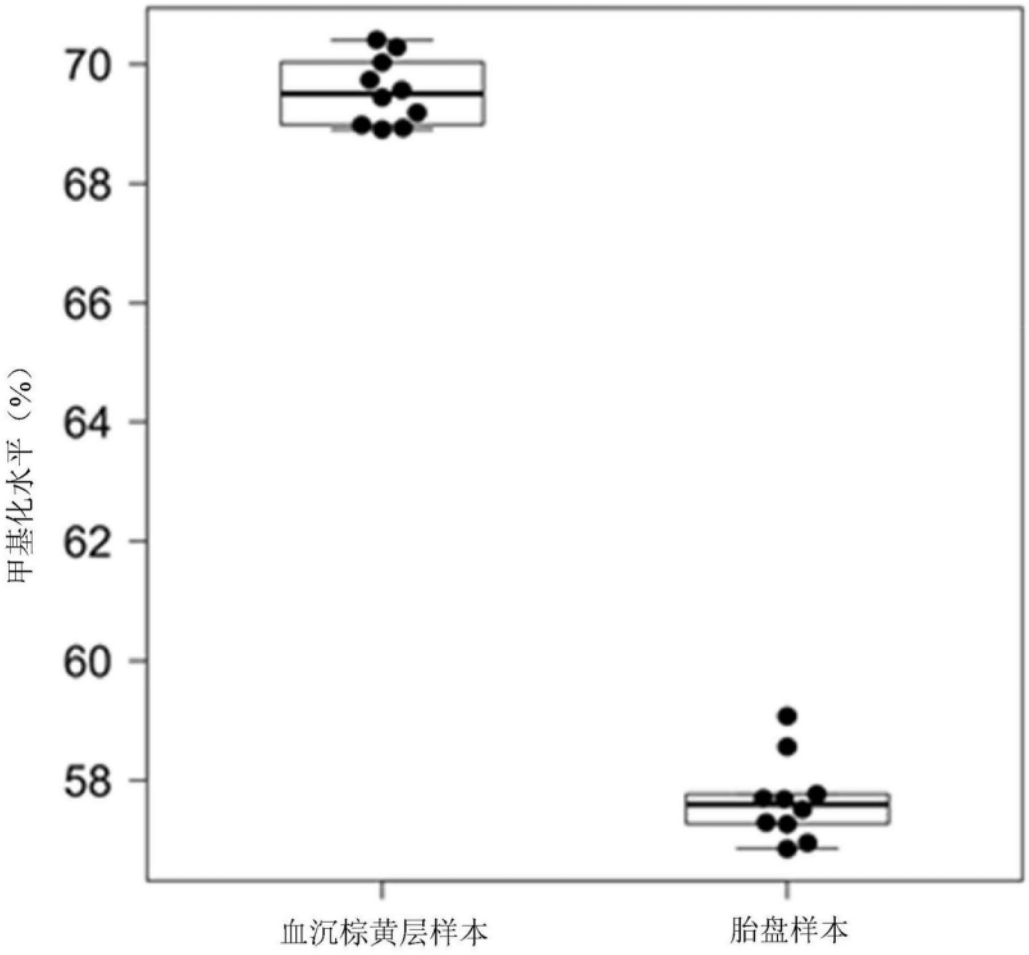


图132

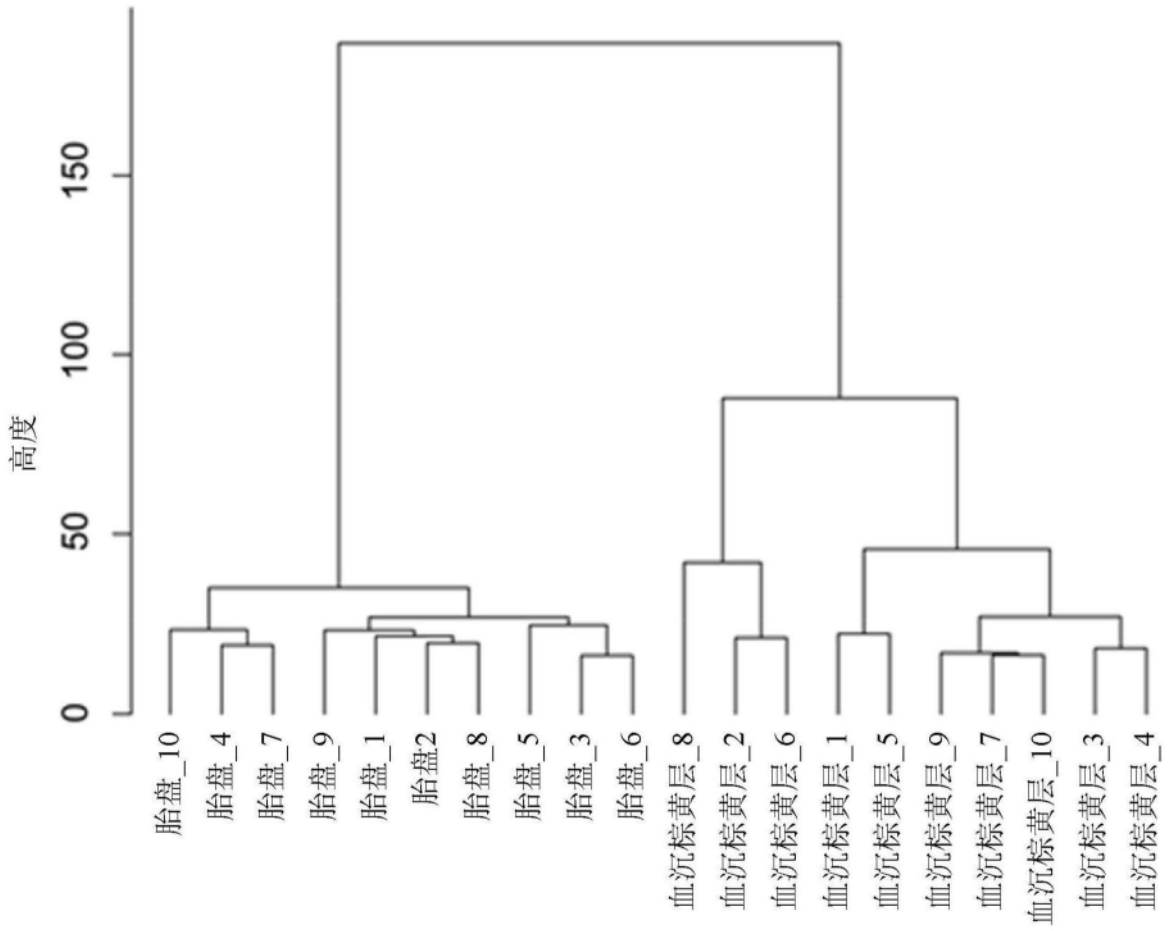

高度
0
50
100
150
胎盘_10
胎盘_4
胎盘_7
胎盘_9
胎盘_1
胎盘2
胎盘_8
胎盘_5
胎盘_3
胎盘_6
血沉棕黄层_8
血沉棕黄层_2
血沉棕黄层_6
血沉棕黄层_1
血沉棕黄层_5
血沉棕黄层_9
血沉棕黄层_7
血沉棕黄层_10
血沉棕黄层_3
血沉棕黄层_4

图133