(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117545855 A
(43)申请公布日 2024.02.09

(21)申请号 202280037190.7
(22)申请日 2022.04.12
(30)优先权数据
63/173,728 2021.04.12 US
(85)PCT国际申请进入国家阶段日
2023.11.23
(86)PCT国际申请的申请数据
PCT/CN2022/086260 2022.04.12
(87)PCT国际申请的公布数据
WO2022/218290 EN 2022.10.20
(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 邓佳恩
(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 洪欣 伊硕
(51)Int.Cl.
C12Q 1/6869(2018.01)
G16B 20/00(2019.01)
G16B 30/00(2019.01)
G16B 50/00(2019.01)

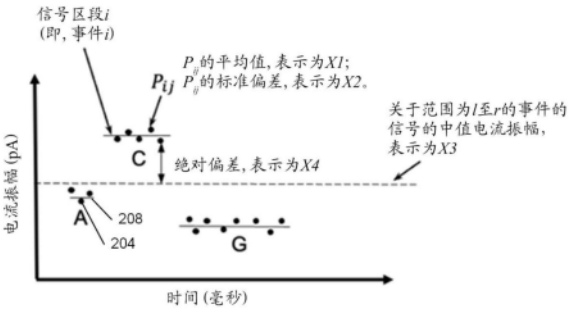
权利要求书5页 说明书29页 附图23页

(54)发明名称

使用电信号的碱基修饰分析

(57)摘要

本文中描述使用电信号和其他数据确定碱基修饰的系统和方法。实施方案可利用从与测序相关的电信号获得的特征,诸如使用纳米孔获得的那些特征,所述特征受各种碱基修饰影响,以及确定甲基化状态的目标位置的周围窗口中核苷酸的身份。其他特征可包含对应于该核苷酸的区段电信号的统计值的向量和核酸分子的区域中的窗口中的电信号的统计值。所检测的碱基修饰可用于对生物样本进行额外分析。



CN 117545855 A

1.一种用于检测核酸分子中核苷酸的修饰的方法,所述方法包括:

接收输入数据结构,所述输入数据结构对应于样本核酸分子中测序的核苷酸的窗口,其中所述样本核酸分子通过测量对应于所述核苷酸的电信号来测序,所述输入数据结构包括以下特性的值:

对于所述窗口内的每个核苷酸:

所述核苷酸的身份,

所述核苷酸相对于各个窗口内的目标位置的位置,和

包括所述电信号的对应于所述核苷酸的区段的第一区段统计值的向量;

将所述输入数据结构输入至模型中,所述模型通过以下操作进行训练:

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多个第一核酸分子的各个核酸分子中测序的核苷酸的各个窗口,其中所述第一核酸分子中的每一者通过测量对应于所述核苷酸的所述电信号来测序,其中所述修饰在每个第一核酸分子的每个窗口中的目标位置处的核苷酸中具有已知的第一状态,每个第一数据结构包括与所述输入数据结构相同特性的值,

储存多个第一训练样本,每个样本包含所述第一多个第一数据结构之一和指示所述目标位置处的核苷酸的第一状态的第一标记,和

当将所述第一多个第一数据结构输入至所述模型时,基于所述模型的匹配或不匹配所述第一标记的相应标记的输出,使用所述多个第一训练样本来使所述模型的参数优化,其中所述模型的输出指定所述各个窗口中所述目标位置处的核苷酸是否具有所述修饰,

使用所述模型确定所述修饰是否存在于所述输入数据结构中的所述窗口内的所述目标位置处的核苷酸中。

2.如权利要求1所述的方法,其中所述第一区段统计值表示所述电信号的对应于所述核苷酸的所述区段的平均值。

3.如权利要求1所述的方法,其中所述第一区段统计值表示所述电信号的对应于所述核苷酸的所述区段的所述电信号的变化。

4.如权利要求1所述的方法,其中所述第一区段统计值表示所述电信号的对应于所述核苷酸的所述区段的平均值的归一化值。

5.如权利要求1、2或4中任一项所述的方法,其中所述向量包括表示所述电信号的对应于所述核苷酸的所述区段的变化的第二区段统计值。

6.如权利要求1、2或3中任一项所述的方法,其中所述向量包括表示所述电信号的对应于所述核苷酸的所述区段的平均值的归一化值的第二区段统计值。

7.如权利要求2所述的方法,其中:

所述向量包括表示所述电信号的对应于所述核苷酸的所述区段的变化的第二区段统计值,且

所述向量包括表示所述第一区段统计值的归一化值的第三区段统计值。

8.如前述权利要求中任一项所述的方法,其中所述输入数据结构包括所述核酸分子的等于或大于所述窗口的区域中的所述电信号的第一区统计值的值。

9.如权利要求8所述的方法,其中所述第一区统计值表示所述区域中的所述电信号的平均值或中值。

10.如权利要求8所述的方法,其中所述第一区统计值表示相对于所述区域中的所述电信号的所述平均值或中值的所述电信号的变化绝对值的中值或平均值。

11.如权利要求9所述的方法,其中所述输入数据结构进一步包括第二区统计值,所述第二区统计值表示相对于所述区域中的所述电信号的所述平均值或中值的所述电信号的变化绝对值的中值或平均值。

12.如权利要求8至11中任一项所述的方法,其中所述区域在所述样本核酸分子的一个股上。

13.如权利要求8至12中任一项所述的方法,其中所述区域为所述样本核酸分子或包括至少5、10、15、20、25、30、50、100、200、300、400、500或1k、5k、10k、50k或1M个核苷酸。

14.如权利要求8至13中任一项所述的方法,其中所述区域以所述核苷酸为中心。

15.如前述权利要求中任一项所述的方法,其中所述窗口包括所述样本核酸分子的两个股上的核苷酸。

16.如前述权利要求中任一项所述的方法,其中所述修饰为甲基化或氧化。

17.如前述权利要求中任一项所述的方法,其中所述电信号为电流、电压、电阻、电感、电容或阻抗。

18.如前述权利要求中任一项所述的方法,其进一步包括使用纳米孔对所述样本核酸分子进行测序。

19.如权利要求1所述的方法,其中:

所述修饰为甲基化,且

所述样本核酸分子为游离的且获自怀有胎儿的女性个体的生物样本,

所述方法进一步包括:

使用所述目标位置处的核苷酸的修饰状态,和任选地所述样本核酸分子的一个或多个其他核苷酸的修饰状态来确定所述样本核酸分子为胎儿来源还是母体来源,其中所述修饰状态为所述修饰是否存在。

20.如权利要求19所述的方法,其中确定所述样本核酸分子为胎儿来源还是母体来源包括:

使用所述一个或多个核苷酸的所述修饰状态确定所述样本核酸分子的甲基化水平;和

将所述样本核酸分子的所述甲基化水平与参考值进行比较。

21.如权利要求20所述的方法,其中所述参考值由一个或多个母体核酸分子的甲基化水平确定。

22.如权利要求20所述的方法,其中:

将所述样本核酸分子的所述甲基化水平与所述参考值进行比较包括确定所述样本核酸分子的所述甲基化水平低于所述参考值,和

确定所述样本核酸分子为胎儿来源还是母体来源包括使用所述比较确定所述样本核酸分子为胎儿来源。

23.如权利要求19所述的方法,其进一步包括:

将所述样本核酸分子鉴定为与预定基因组区域对齐。

24.如权利要求19所述的方法,其中:

所述样本核酸分子为多个样本核酸分子中的一个样本核酸分子,

所述方法进一步包括：

使用所述修饰状态确定所述多个样本核酸分子中的每一个为胎儿来源还是母体来源，和

使用对所述多个样本核酸分子的胎儿或母体来源的确定来确定胎儿分数。

25.如权利要求1所述的方法，其中：

所述修饰为甲基化，

所述样本核酸分子为游离的且获自怀有胎儿的女性个体的生物样本，且

所述样本核酸分子为多个样本核酸分子中的一个样本核酸分子，

所述方法进一步包括：

将所述多个样本核酸分子鉴定为与胎儿基因组的区域对齐，

确定所述多个样本核酸分子中的每一个样本核酸分子的一个或多个核苷酸的修饰状态，

使用所述多个样本核酸分子中的每一个样本核酸分子的所述一个或多个核苷酸的所述修饰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所述胎儿基因组的所述区域中是否存在拷贝数畸变。

26.一种用于检测核酸分子中核苷酸的修饰的方法，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多个第一核酸分子中的各个核酸分子中测序的核苷酸的各个窗口，其中所述第一核酸分子中的每一者通过测量对应于所述核苷酸的电信号来测序，其中所述修饰在每个第一核酸分子的每个窗口中目标位置处的核苷酸中具有已知的第一状态，每个第一数据结构包括以下特性的值：

对于所述窗口内的每个核苷酸：

所述核苷酸的身份，

所述核苷酸相对于各个窗口内的目标位置的位置，和

包括所述电信号的对应于所述核苷酸的区段的第一区段统计值的向量；

储存多个第一训练样本，每个样本包含所述第一多个第一数据结构之一和指示所述目标位置处的核苷酸的修饰的第一状态的第一标记；和

当将所述第一多个第一数据结构输入至模型时，基于所述模型的匹配或不匹配所述第一标记的相应标记的输出，使用所述多个第一训练样本使所述模型的参数优化而训练所述模型，其中所述模型的输出指定所述各个窗口中所述目标位置处的核苷酸是否具有所述修饰。

27.如权利要求26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多个第二核酸分子的各个核酸分子中测序的核苷酸的各个窗口，其中所述修饰在每个第二核酸分子的每个窗口内的目标位置处的核苷酸中具有已知的第二状态，每个第二数据结构包括与所述第一多个第一数据结构相同的特性的值；

储存多个第二训练样本，每个样本包含所述第二多个第二数据结构之一和指示所述目标位置处的核苷酸的第二状态的第二标记；

其中训练：

所述第一状态或所述第二状态为存在所述修饰,且另一状态为不存在所述修饰,

所述模型进一步包括当将所述第二多个第二数据结构输入至所述模型时,基于所述模型的匹配或不匹配所述第二标记的相应标记的输出,使用所述多个第二训练样本使所述模型的参数优化。

28.如权利要求27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。

29.如权利要求26所述的方法,其中:

与所述第一多个第一数据结构相关的每个窗口包括所述第一核酸分子的第一股上的核苷酸和所述第一核酸分子的第二股上的核苷酸,和

每个第一数据结构进一步包括对于所述窗口内的每个核苷酸的股特性的值,所述股特性指示所述核苷酸存在于所述第一股或所述第二股上。

30.如权利要求26所述的方法,其中所述修饰包括所述目标位置处的核苷酸的甲基化。

31.如权利要求30所述的方法,其中所述已知的第一状态包括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未甲基化状态。

32.如权利要求26所述的方法,其中所述第一区段统计值表示所述电信号的对应于所述核苷酸的所述区段的平均值。

33.如权利要求26所述的方法,其中所述第一区段统计值表示所述电信号的对应于所述核苷酸的所述区段的所述电信号的变化。

34.如权利要求26所述的方法,其中所述第一区段统计值表示所述电信号的对应于所述核苷酸的所述区段的平均值的归一化值。

35.如权利要求26、32或34中任一项所述的方法,其中所述向量包括表示所述电信号的对应于所述核苷酸的所述区段的变化的第二区段统计值。

36.如权利要求26、32或33中任一项所述的方法,其中所述向量包括表示所述电信号的对应于所述核苷酸的所述区段的平均值的归一化值的第二区段统计值。

37.如权利要求32所述的方法,其中:

所述向量包括表示所述电信号的对应于所述核苷酸的所述区段的变化的第二区段统计值,且

所述向量包括表示所述第一区段统计值的归一化值的第三区段统计值。

38.如权利要求26至37中任一项所述的方法,其中每个第一数据结构包括所述各个核酸分子的等于或大于所述窗口的区域中的所述电信号的第一区统计值的值。

39.如权利要求38所述的方法,其中所述第一区统计值表示所述区域中的所述电信号的平均值或中值。

40.如权利要求38所述的方法,其中所述第一区统计值表示相对于所述区域中的所述电信号的所述平均值或中值的所述电信号的变化绝对值的中值或平均值。

41.如权利要求39所述的方法,其中所述第一数据结构进一步包括第二区统计值,所述第二区统计值表示相对于所述区域中的所述电信号的所述平均值或中值的所述电信号的变化绝对值的中值或平均值。

42.如权利要求38至41中任一项所述的方法,其中所述区域在所述各个核酸分子的一个股上。

43.如权利要求38至45中任一项所述的方法,其中所述区域为所述各个核酸分子或包括至少5、10、15、20、25、30、50、100、200、300、400、500或1k、5k、10k、50k或1M个核苷酸。

44.如权利要求38至43中任一项所述的方法,其中所述区域以所述核苷酸为中心。

45.如权利要求26至44中任一项所述的方法,其中所述窗口包括所述各个核酸分子的两个股上的核苷酸。

46.一种计算机产品,其包括储存多个指令的非暂时性计算机可读介质,所述多个指令在执行时控制计算机系统以执行如前述权利要求中任一项所述的方法。

47.一种系统,其包括:

权利要求46所述的计算机产品;和

一个或多个处理器,其用于执行储存于所述计算机可读介质上的指令。

48.一种系统,其包括用于执行上述方法中的任一个的装置。

49.一种系统,其包括经配置以执行上述方法中的任一个的一个或多个处理器。

50.一种系统,其包括分别执行上述方法中的任一个的步骤的模块。

使用电信号的碱基修饰分析

相关申请案的交叉引用

[0001] 本申请要求2021年4月12日提交的美国临时专利申请63/173,728的优先权益,其以全文引用的方式并入本文中且用于所有目的。

背景技术

[0002] 核酸中碱基修饰的存在在包括病毒、细菌、植物、真菌、线虫、昆虫和脊椎动物(例如人类)等的不同生物体中各不相同。最常见的碱基修饰为将甲基添加至不同位置的不同DNA碱基,即所谓的甲基化。在胞嘧啶、腺嘌呤、胸腺嘧啶和鸟嘌呤上均已发现甲基化,诸如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)、N6-甲基腺嘌呤(6mA)、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为最常见的碱基甲基化类型,其次为鸟嘌呤(即在CpG情况下)。

[0003] DNA甲基化对哺乳动物的发育至关重要,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、维持有丝分裂过程中基因组的稳定性和调控亲源基因组印记方面具有显著的作用。

[0004] DNA甲基化在启动子和增强子的沉默中以协调的方式发挥着许多重要作用(Robertson,2005;Smith和Meissner,2013)。已发现许多人类疾病与DNA甲基化的畸变有关,包括但不限于印记病症(例如贝克威思-威德曼综合征(Beckwith-Wiedemann syndrome)和普瑞德威利综合征(Prader-Willi syndrome))、重复不稳定性疾病(例如X脆折综合征)、自体免疫性病症(例如全身性红斑狼疮)、代谢障碍(例如I型和II型糖尿病)、神经病症、衰老等。

[0005] 准确测量DNA分子上的甲基化修饰将具有许多临床意义。一种广泛使用的测量DNA甲基化的方法为经由使用亚硫酸氢盐测序(BS-seq)(Lister等人,2009;Frommer等人,1992)。在此方法中,DNA样本首先用亚硫酸氢盐处理,将未甲基化的胞嘧啶(即C)转化为尿嘧啶。相反,甲基化的胞嘧啶保持不变。随后通过DNA测序分析亚硫酸氢盐修饰的DNA。在另一种方法中,在亚硫酸氢盐转化之后,接着使用可区分具有不同甲基化谱的经亚硫酸氢盐转化的DNA的引物对经修饰的DNA进行聚合酶链式反应(PCR)扩增(Herman等人,1996)。后一种方法称为甲基化特异性PCR。

[0006] 此类基于亚硫酸氢盐的方法的一个缺点为,据报导亚硫酸氢盐转化步骤会显著降解大多数经处理的DNA(Grunau,2001)。另一个缺点为亚硫酸氢盐转化步骤会产生强烈的CG偏差(Olova等人,2018),导致具有异质甲基化状态的DNA混合物典型的信噪比降低。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是对长DNA分子进行测序的理想方法。

[0007] 正在持续努力以实现对核酸的碱基修饰的无亚硫酸氢盐确定。然而,很少有商业上可行的工具能够达到与亚硫酸氢盐测序相当的灵敏度和特异度水平。纳米孔测序为一种

不需要对样本进行化学标记的具有吸引力的测序类型。用纳米孔测序检测碱基修饰可为成本相对较低的且高效的。

[0008] 因此,需要用纳米孔测序来确定碱基修饰。在本公开内容中,我们描述了处理通过具有高灵敏度和特异性的纳米孔测序所产生的电流信号以用于碱基修饰确定的新的方法和系统。

发明内容

[0009] 所描述的实施方案允许在没有模板DNA预处理(诸如酶促和/或化学转化,或蛋白质和/或抗体结合)的情况下确定核酸中的碱基修饰,诸如5mC。本公开内容中存在的实施方案可用于检测不同类型的碱基修饰,例如,包括但不限于4mC、5hmC、5fC、5caC、1mA、3mA、6mA、7mA、3mC、2mG、6mG、7mG、3mT、4mT等。此类实施方案可利用从与测序相关的电信号获得的特征(诸如使用纳米孔获得的那些特征,所述特征受各种碱基修饰影响),以及确定甲基化状态的目标位置的周围窗口中核苷酸的身份。核苷酸的原始电信号也可与核苷酸上游或下游的核苷酸有关。可以使用合适的技术将原始电信号分配给不同的核苷酸。

[0010] 本发明的实施方案可与纳米孔测序一起使用。纳米孔测序系统的一个实例为由牛津纳米孔科技有限公司(Oxford Nanopore Technologies)商业化的系统。方法可使用利用纳米孔测量的电信号。方法可使用核苷酸的身份、核苷酸相对于目标位置的位置、包含对应于该核苷酸的区段电信号的统计值的向量和核酸分子的区域中的窗口中的电信号的统计值。

[0011] 我们开发的方法可充当检测生物样本中碱基修饰的工具,以评估样本中的甲基化谱,用于各种目的,包括但不限于研究和诊断目的。检测到的甲基化谱可用于不同的分析。甲基化谱可用于检测DNA的来源(例如母体或胎儿、组织、细菌)。检测组织中的异常甲基化谱有助于鉴别个体的发育病症和其他病症。

[0012] 可参考以下详细描述和附图来获得对本发明的实施方案的性质和优势的较佳理解。

附图的简要说明

[0013] 图1示出纳米孔测序。

[0014] 图2示出根据本发明的实施方案的不同信号特征。

[0015] 图3示出根据本发明的实施方案的电流信号分段和信号特征向量的建构。

[0016] 图4为根据本发明的实施方案的每个核苷酸穿过纳米孔的事件长度(即,持续时间)的分布图。

[0017] 图5示出根据本发明的实施方案的使用包括电流模式、测序位置和测序背景(sequencing context)的整合式表示矩阵的5mC检测的原理。

[0018] 图6示出根据本发明的实施方案的使用包括电流模式、测序位置和基于双股DNA的两个股的测序背景的整合式表示矩阵的碱基修饰检测的原理。

[0019] 图7展示根据本发明的实施方案的核尺寸对碱基修饰分析的性能的影响。

[0020] 图8展示根据本发明的实施方案的关于甲基化检测的用于训练和测试的测序分子数目。

[0021] 图9A至图9D为根据本发明的实施方案的使用IPM-CNN和IPM-RNN方法的WGADNA与

M.SssI处理过的DNA数据集之间的CpG甲基化概率的盒状图。

[0022] 图10A和图10B展示根据本发明的实施方案的训练数据集和测试数据集的接受者操作特征(ROC)曲线。

[0023] 图11为根据本发明的实施方案的用于甲基化分析的不同工具的性能的表。

[0024] 图12为根据本发明的实施方案的检测核酸分子中核苷酸的修饰的方法的流程图。

[0025] 图13为根据本发明的实施方案的检测核酸分子中核苷酸的修饰的方法的流程图。

[0026] 图14示出根据本发明的实施方案的测量系统。

[0027] 图15展示可与根据本发明的实施方案的系统和方法一起使用的实例计算机系统的框图。

[0028] 图16展示根据本发明的实施方案的不同参数组合对ROC曲线下面积(AUC)的影响的图。

[0029] 图17展示根据本发明的实施方案的窗口大小对AUC的影响的图。

[0030] 图18示出根据本发明的实施方案的使用包括电流模式、测序位置和测序背景的整合式表示矩阵的6mA检测的原理。

[0031] 图19展示根据本发明的实施方案的6mA检测的AUC的图。

[0032] 图20为针对根据本发明的实施方案的源自血沉棕黄层(buffy coat)和NPC肿瘤样本的DNA,通过IPM-RNN模型确定的单分子甲基化水平的比较。

[0033] 图21展示根据本发明的实施方案的单分子甲基化模式的实例。

[0034] 图22为根据本发明的实施方案的母体特异性和胎儿特异性游离DNA分子的单分子甲基化水平的图。

[0035] 图23为根据本发明的实施方案的使用由IPM-CNN模型确定的甲基化模式确定游离DNA分子的胎儿和母体来源的ROC曲线。

术语

[0036] “组织”对应于一组细胞,其共同归类为一个功能单元。可在单一组织中找到超过一种类型的细胞。不同类型的组织可由不同类型的细胞(例如肝细胞、肺泡细胞或血细胞)组成,但也可对应于来自不同生物体的组织(母亲与胎儿;接受移植的个体的组织;经微生物或病毒感染的生物体的组织)或健康细胞与肿瘤细胞。“参考组织”可对应于用于确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用于确定该组织类型的组织特异性甲基化水平。

[0037] “生物样本”是指取自人类个体的任何细胞样本。生物样本可为组织活检、细针抽吸物或血细胞。样本也可为获自孕妇的游离样本,例如血浆或血清或尿液。在各种实施方案中,已富集游离DNA的来自孕妇的生物样本(例如经由离心方案获得的血浆样本)中的大多数DNA可为游离的,例如大于50%、60%、70%、80%、90%、95%或99%的DNA可为游离的。离心方案可包含例如3,000g×10分钟获得流体部分,和以例如30,000g再离心10分钟以移除残余细胞。在某些实施方案中,在3,000g离心步骤之后,可接着对流体部分进行过滤(例如使用孔径(直径)为5μm或更小的过滤器)。

[0038] “序列读数”是指自核酸分子的任何部分或全部测序的一串核苷酸。举例而言,序列读数可为自核酸片段测序的短核苷酸串(例如20至150个)、在核酸片段的一端或两端的短核苷酸串或存在于生物样本中的整个核酸片段的测序。序列读数可以多种方式获得,例

如使用测序技术或使用探针,例如杂交数组或捕获探针;或扩增技术,诸如聚合酶链式反应(PCR)或使用单一引物的线性扩增或等温扩增。

[0039] “位点”(也称作“基因组位点”)对应于单一位点,其可为单一碱基位置或相关碱基位置群,例如CpG位点或相关碱基位置的较大群。“基因座”可对应于包含多个位点的区域。基因座可仅包含一个位点,这将使得基因座在该情形下等同于一个位点。

[0040] “甲基化状态”是指给定位点处的甲基化状态。举例而言,位点可为甲基化的、未甲基化的或在一些情况下不能确定。

[0041] 各基因组位点(例如CpG位点)的“甲基化指数”可指在该位点处显示甲基化的DNA片段(例如,如由序列读数或探针确定)相对于涵盖该位点的读数总数的比例。“读数”可对应于获自DNA片段的信息(例如,位点处的甲基化状态)。读数可使用优先杂交至在一个或多个位点处具有特定甲基化状态的DNA片段的试剂(例如引物或探针)来获得。通常,所述试剂是在用视DNA分子的甲基化状态而有差异地修饰或有差异地辨识DNA分子的方法处理后施用,该方法例如为亚硫酸氢盐转化、或甲基化敏感限制酶、或甲基化结合蛋白、或抗甲基胞嘧啶抗体、或辨识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(例如,来自美国太平洋生物科学公司(Pacific Biosciences))和纳米孔测序(例如来自牛津纳米孔科技有限公司))。

[0042] 区域的“甲基化密度”可指显示甲基化的区域内的位点处的读数数目除以覆盖该区域中的位点的读数总数。所述位点可具有特定特性,例如为CpG位点。因此,区域的“CpG甲基化密度”可指显示CpG甲基化的读数数目除以覆盖该区域中的CpG位点(例如特定CpG位点、CpG岛或较大区域内的CpG位点)的读数总数。例如,人类基因组中各100kb区段(bin)的甲基化密度可自亚硫酸氢盐处理之后在CpG位点处未转化的胞嘧啶(其对应于甲基化胞嘧啶)的总数确定为映射至100kb区域的序列读数所覆盖的所有CpG位点的比例。也可针对其他区段大小,例如500bp、5kb、10kb、50kb或1Mb等执行此分析。区域可为整个基因组或染色体或染色体的一部分(例如染色体臂)。替代地,甲基化密度可在无亚硫酸氢盐转化的情况下使用纳米孔测序使用本公开内容所描述的实施方案来确定。当区域仅包含CpG位点时,CpG位点的甲基化指数与区域的甲基化密度相同。“甲基化胞嘧啶的比例”可指相对于所分析的胞嘧啶残基,即包含该区域中除CpG情形之外的胞嘧啶的总数而言显示为甲基化(例如在亚硫酸氢盐转化之后未经转化)的胞嘧啶位点“C”数目。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和在一个或多个位点处甲基化的分子(例如胞嘧啶)比例为“甲基化水平”的实例。除亚硫酸氢盐转化以外,可使用本领域技术人员已知的其他方法来查询DNA分子的甲基化状态,包括但不限于对甲基化状态敏感的酶(例如甲基化敏感限制酶)、甲基化结合蛋白、使用对甲基化状态敏感的平台进行的单分子测序(例如纳米孔测序(Schreiber等人,《国家科学院院刊(Proc Natl Acad Sci)》2013;110:18910-18915)和通过单分子实时测序(例如来自美国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))。

[0043] “甲基化组”提供基因组中的多个位点或基因座处的DNA甲基化的量的量度。甲基化组可对应于所有基因组、基因组的相当大部分或基因组的一个或多个相对小的部分。

[0044] “妊娠血浆甲基化组”为自妊娠动物(例如人类)的血浆或血清确定的甲基化组。妊娠血浆甲基化组为游离甲基化组的实例,因为血浆和血清包含游离DNA。妊娠血浆甲基化组

也为混合甲基化组的实例,因为其为来自体内不同器官或组织或细胞的DNA的混合物。在一个实施方案中,此类细胞为造血细胞,包括但不限于红血球系(即红血球)、骨髓系(例如嗜中性粒细胞和其前体)和巨核细胞系的细胞。在妊娠期,血浆甲基化组可含有来自胎儿和母亲的甲基化组信息。"细胞甲基化组"对应于自患者的细胞(例如血细胞)确定的甲基化组。血细胞的甲基化组称为血细胞甲基化组。

[0045] "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包括但不限于CpG位点的甲基化指数、区域中的CpG位点的甲基化密度(简称MD)、CpG位点在连续区域上的分布、含有超过一个CpG位点的区域内的各个别CpG位点的甲基化模式或水平,和非CpG甲基化。在一个实施方案中,甲基化谱可包含超过一种类型的碱基(例如胞嘧啶或腺嘌呤)的甲基化或非甲基化模式。基因组的相当大部分的甲基化谱可视为等同于甲基化组。哺乳动物基因组中的"DNA甲基化"通常指将甲基添加至CpG二核苷酸当中的胞嘧啶残基的5'碳(即5-甲基胞嘧啶)。DNA甲基化可在例如CHG和CHH的其他情形下发生于胞嘧啶中,其中H为腺嘌呤、胞嘧啶或胸腺嘧啶。胞嘧啶甲基化也可呈5-羟甲基胞嘧啶形式。还已经报导非胞嘧啶甲基化,诸如N^6-甲基腺嘌呤。

[0046] "甲基化模式"是指甲基化和非甲基化碱基的次序。举例而言,甲基化模式可为单个DNA股、单个双股DNA分子或另一类型的核酸分子上的甲基化碱基的次序。作为一实例,三个连续CpG位点可具有以下甲基化模式中的任一者:UUU、MMM、UMM、UMU、UUM、MUM、MUU或MMU,其中"U"指示未甲基化位点且"M"指示甲基化位点。当将此概念扩展至包括但不限于甲基化的碱基修饰时,将使用术语"修饰模式",其是指经修饰和未经修饰碱基的次序。举例而言,修饰模式可为单个DNA股、单个双股DNA分子或另一类型的核酸分子上的经修饰碱基的次序。作为一实例,三个连续潜在地可修饰位点可具有以下修饰模式中的任一者:UUU、MMM、UMM、UMU、UUM、MUM、MUU或MMU,其中"U"指示未经修饰位点且"M"指示经修饰位点。不基于甲基化的碱基修饰的一个实例为诸如于8-侧氧基-鸟嘌呤中的氧化变化。

[0047] 术语"高甲基化"和"低甲基化"可指单个DNA分子的甲基化密度,如通过其单分子甲基化水平所测量,例如分子内的甲基化碱基或核苷酸的数目除以该分子内的可甲基化碱基或核苷酸的总数。高甲基化分子为其中单分子甲基化水平等于或高于阈值的分子,该阈值可根据不同应用而界定。阈值可为5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或95%。低甲基化分子为其中单分子甲基化水平等于或低于阈值的分子,该阈值可根据不同应用而界定且可根据不同应用而变化。阈值可为5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或95%。

[0048] 术语"高甲基化"和"低甲基化"也可指DNA分子群体的甲基化水平,如通过这些分子的多分子甲基化水平所测量。高甲基化分子群体为其中多分子甲基化水平等于或高于阈值的分子群体,该阈值可根据不同应用而界定且可根据不同应用而变化。阈值可为5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或95%。低甲基化分子群体为其中多分子甲基化水平等于或低于阈值的分子群体,该阈值可根据不同应用而界定。阈值可为5%、10%、20%、30%、40%、50%、60%、70%、80%、90%和95%。在一个实施方案中,可将分子群体与一个或多个经选择的基因组区域进行比对。在一个实施方案中,一个或多个经选择的基因组区域可与诸如遗传病症、印记病症、表观遗传病症、代谢病症或神经病症的疾病相关。一个或多个经选择的基因组区域的长度可为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。

[0049] 如本文所使用的术语“分类”是指与样本的特定特性相关的任何数字或其他字符。举例而言,“+”符号(或词语“阳性”)可表示将样本分类为具有缺失或扩增。分类可为二元的(例如阳性或阴性)或具有更多分类水平(例如1至10或0至1的标度)。

[0050] 术语“截止值”和“阈值”是指操作中所使用的预定数值。举例而言,截止值大小可指一种大小,大于此大小则排除该片段。阈值可为高于或低于特定分类适用的值。在这些情形中的任一者下均可使用这些术语中的任一者。截止值或阈值可为表示特定分类或在两种或更多种分类之间进行辨别的“参考值”或源自该参考值。如技术人员应了解,此类参考值可以各种方式确定。例如,可针对具有不同已知分类的两个不同个体群组确定度量,且可选择参考值作为一个分类的代表(例如平均值)或介于度量的两个集群之间的值(例如经选择以获得所需的灵敏度和特异性)。作为另一实例,参考值可基于样本的统计分析或模拟来确定。

[0051] “病理等级”(或病症等级)可指与生物体相关的病理的量、水平或严重性,其可经由对其细胞的分析来测量。病理的另一实例为移植器官的排斥。其他例示性病理可包含基因组印记病症、自体免疫攻击(例如损害肾脏的狼疮性肾炎或损害神经系统的多发性硬化症)、发炎性疾病(例如肝炎)、纤维化过程(例如肝硬化)、脂肪浸润(例如脂肪性肝病)、退行性过程(例如阿尔茨海默氏病(Alzheimer’s disease))和缺血性组织损伤(例如心肌梗塞或中风)。个体的健康状态可视为无病理的分类。

[0052] “妊娠相关病症”包含以母体和/或胎儿组织中基因的相对表现水平异常为特征的任何病症。这些病症包括但不限于子痫前症、宫内发育迟缓、侵入性胎盘形成、早产、新生儿溶血性疾病、胎盘功能不全、胎儿水肿、胎儿畸形、HELLP(溶血、肝酶升高和血小板计数低)综合征、全身性红斑狼疮(SLE)和母亲的其他免疫性疾病。在一些实施方案中,妊娠相关病症为与妊娠期间的生理或形态异常相关的任何病状。

[0053] 缩写“bp”是指碱基对。在一些情况下,“bp”可用于表示DNA片段的长度,即使DNA片段可为单股的且不包含碱基对。在单股DNA的情形下,“bp”可解释为提供核苷酸的长度。

[0054] 缩写“nt”是指核苷酸。在一些情况下,“nt”可用于表示以碱基为单位的单股DNA长度。此外,“nt”可用于表示相对位置,诸如所分析的基因座的上游或下游。在关于技术概念化、数据显示、处理和分析的一些情形下,“nt”和“bp”可互换使用。

[0055] 术语“序列上下文(sequence context)”可指一段DNA中的碱基组成(A、C、G或T)和碱基顺序。此段DNA可围绕进行碱基修饰分析或作为碱基修饰分析的目标的碱基。举例而言,序列上下文可指进行碱基修饰分析的碱基的上游和/或下游的碱基。

[0056] 术语“机器学习模型”可包含基于使用样本数据(例如训练数据)对测试数据作出预测的模型,且因此可包含监督式学习。机器学习模型常常使用计算机或处理器来研发。机器学习模型可包括统计模型。

[0057] 术语“数据分析框架”可包括可将数据视为输入且随后输出所预测结果的算法和/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、其他人工智能模型和其组合。

[0058] 术语“实时测序”可指涉及在测序所涉及的过程期间进行数据收集或监测的技术。

举例而言,实时测序可涉及当核苷酸股易位纳米孔时对通过该纳米孔的离子电流进行电信号监测。

[0059] 术语“电信号”可指传达信息的电压或电流。电信号可以多种规律和/或不规律的信号波形类型和/或形状,诸如方形波、矩形波、三角形波、锯齿形波形,或多种脉冲和尖峰来表示。电信号可包含电压或电流随时间推移的变化的视觉表示。可在特定时间(例如,毫秒)对电信号的测量进行采样。举例而言,以1kHz、2kHz、3kHz、4kHz、5kHz、10kHz、20kHz、30kHz、40kHz、50kHz、100kHz等的频率对电流进行采样。

[0060] 术语“信号区段”或“区段”可指与对特定核苷酸进行测序相关的电信号的迹线的一部分。该区段可对应于由纳米孔测序中的碱基识别确定的核苷酸。该区段可涵盖迹线的某一持续时间。不同区段可具有不同的持续时间。各区段可不重叠。在一些实施方案中,电信号幅度可在区段中具有一定的变化。举例而言,电信号幅度可在该区段中的平均值或中值电信号幅度的5%、10%、20%、30%或40%内。

[0061] 术语“约(about/approximately)”可意指在如通过本领域技术人员所确定的特定值的可接受误差范围内,其将部分地视该值如何被测量或确定,即测量系统的限制而定。举例而言,根据本领域中的实践,“约”可意指在1或大于1个标准偏差内。可替代地,“约”可意指给定值的至多20%、至多10%、至多5%或至多1%的范围。可替代地,尤其关于生物系统或方法,术语“约”可意指在值的一定数量级内、在5倍内且更佳地在2倍内。当特定值描述于本申请案和申请专利范围中时,除非另外说明,否则应假定术语“约”意指在特定值的可接受误差范围内。术语“约”可具有如本领域技术人员通常所理解的含义。术语“约”可指±10%。术语“约”可指±5%。

详细描述

[0062] 需要使用纳米孔测序检测碱基修饰(例如甲基化)的准确和有效的方法。调研性研究已研究使用由纳米孔测序产生的电信号分析DNA甲基化的可行性(Simpson等人,《自然方法学(Nat Methods)》2017;14:407-410;Liu等人,《自然通讯(Nat Commun.)》2019;10:2449;Ni等人,《生物信息(Bioinformatics)》2019;35:4586-4595)。5-甲基胞嘧啶(5mC)的报导性能在许多验证研究中为次优的。举例而言,当基于样本NA12878分析H.sapiens R9.4 1D数据时,使用名为DeepSignal的计算工具进行5mC检测的灵敏度据报导为79%,特异性为88%,(Ni等人,《生物信息》2019;35:4586-4595)。如果旨在实现较高的特异性(例如>95%),则预期灵敏度将进一步恶化。对于称为nanopolish的另一工具(Liu等人,《自然通讯》2019;10:2449),当分析相同的数据集时,灵敏度仅为0.61,特异性为0.46。nanopolish软件是基于具有以下假设的隐藏式马可夫模型(hidden Markov model):(1)DNA序列中的6-核苷酸寡聚物(即6-单体单元)的电信号遵循高斯分布(Gaussian distribution);(2)特定碱基的甲基化状态(甲基化或未甲基化)仅取决于前一碱基的甲基化状态的概率;(3)输出仅取决于产生电流信号的甲基化状态而不取决于任何其他甲基化状态或任何其他电流信号的特定电流水平的概率。那些假设在纳米孔测序期间产生的真实电流信号中可能不正确,因此会导致较低的灵敏度和特异性。

[0063] 用于基于牛津纳米孔测序进行DNA甲基化分析的名为DeepMod的最新计算工具尝试使用双向递归神经网络(RNN)。然而,此类方法的设计旨在通过利用电信号合计测序读数的预测结果来测量基因组位置中的甲基化水平,因此不具有分析单分子水平下的甲基化模

式的能力。另外,整个数据集(包含大肠杆菌(Escherichia coli)、莱茵衣藻(Chlamydomonas reinhardtii)和智人(Homo sapiens))的中值测序深度为约33×。在许多商业应用中,将需要较低的测序深度以节省经济成本和分析时间。尚不清楚DeepMod软件是否能够以实际上有意义的准确性分析单分子水平下的甲基化模式。

[0064] 在一项研究中,Yuen等人系统地衡量用于由纳米孔测序进行CpG甲基化检测的工具,且得出结论:大多数工具展示高分散性和与每个CpG位点的预期甲基化百分比的低一致性(Yuen等人,bioRxiv.2020;doi:doi.org/10.1101/2020.10.14.340315)。

[0065] Tse等人报导了使用来自太平洋生物科学公司(PacBio)的单分子实时测序(SMRT-seq),DNA聚合酶的动力学特征,包括通过在DNA聚合期间并入荧光团标记的核苷酸所产生的光信号,诸如脉冲间隔持续时间(IPD)和脉冲宽度(PW),可用于基于使用卷积类神经网络分析由超过一个碱基组成的测量窗口来区分甲基化和未甲基化CpG位点(Tse等人,《美国国家科学院院刊》2021;118:e2019768118;美国专利第11,091,794号)。此类测量窗口将IPD和PW分组成不同的测序背景和测序位置。然而,纳米孔测序使用完全不同的测序机制,视由穿过纳米孔的双股DNA的一个股所引起的电流信号而定。此类原始电信号视穿过纳米孔的不同核苷酸而变化,且特定核苷酸的电信号将受该核苷酸附近的上游和下游核苷酸影响。因此,不同核苷酸将具有检测到的不同长度的电信号迹线,且甚至相同的核苷酸将具有不同长度的电信号迹线。当分析与穿过纳米孔的特定核苷酸或超过一个核苷酸相关的电信号时,在各碱基上检测到的电信号迹线的长度随时间推移为不固定的。相比之下,使用PacBio SMRT-seq进行5mC检测的前述研究是基于两个与各核苷酸的光信号相关的固定测量,即IPD和PW(Tse等人,《美国国家科学院院刊》2021;118:e2019768118)。因此,Tse等人的研究中提出的训练模型(Tse等人,《美国国家科学院院刊》2021;118:e2019768118)不适用于此类通过纳米孔测序产生的电信号。

[0066] 本文所描述的实施方案使用从纳米孔测序获得的电信号来检测核苷酸修饰。核苷酸修饰可包含本文所描述的任何甲基化。从纳米孔测序获得的信息可包含核苷酸的身份、核苷酸相对于目标位置的位置、包含对应于该核苷酸的区段电信号的统计值的向量和核酸分子的区域中的窗口中的电信号的统计值。

[0067] 本公开内容中提供的实施方案可用于从获自生物体的细胞样本(例如,细胞株、实体器官、实体组织、经由内窥镜检获得的样本、绒毛膜样本)获得的DNA。本公开内容中的实施方案也可用于从环境(例如,细菌、细胞污染物)、食品(例如肉)获得的细胞样本。本公开内容提供的实施方案也可用于从孕妇获得的血浆或血清。在一些实施方案中,本公开内容中提供的方法也可在首先例如使用杂交探针(Albert等人,2007;Okou等人,2007;Lee等人,2011),或基于物理分离(例如基于大小等)的方法或在限制酶消化(例如MspI)后,或基于Cas9的富集(Watson等人,2019)富集基因组碎片的步骤之后应用。尽管本发明不需要酶促或化学转化来起作用,但在某些实施方案中,可包括此类转化步骤以进一步增强本发明的性能。

[0068] 本公开内容的实施方案改良纳米孔测序以能够准确且有效地检测经修饰的碱基。可直接检测碱基修饰。实施方案可避免可能无法保留所有修饰信息以供检测的酶促或化学转化。另外,某些酶促或化学转化可能与某些类型的修饰不兼容。本公开内容的实施方案也可避免通过PCR扩增,其可能不会将碱基修饰信息转移至PCR产物。另外,DNA的两个股可一

起测序,从而使一个股的序列与其互补序列配对至另一个股。相比之下,PCR扩增会分开双股DNA的两个股,因此难以对两个组成股的序列进行此类组合分析。

[0069] 此外,相比于其他测序技术,纳米孔测序更具有成本效益和便携性。举例而言,纳米孔测序系统Oxford Nanopore Technologies MinION™为约5,000USD,而基于光信号的测序系统PacBio SMRT™ Sequel II系统为约500,000至700,000USD。纳米孔测序速度为约450个核苷酸/秒,而PacBio SMRT™测序为约5个核苷酸/秒。因此,在相同的时间段内,纳米孔测序可获得比基于光信号的测序系统更多的数据。

[0070] 在有或没有酶促或化学转化的情况下确定的甲基化谱可用于分析生物样本。在一个实施方案中,甲基化谱可用于检测细胞DNA的来源(例如母体或胎儿、组织或病毒)。检测组织中的异常甲基化谱有助于鉴别个体的发育障碍。单分子中的甲基化模式可鉴别嵌合(例如在病毒与人类之间)和杂合DNA(例如在天然基因组中正常未融合的两个基因之间);或在两个物种之间(例如经由基因或基因组操纵)。

I.纳米孔测序原理

[0071] 单分子测序技术的一实例为纳米孔测序(牛津纳米孔科技有限公司)。图1展示用于DNA分子(例如DNA分子104)的纳米孔测序的原理。当单DNA分子穿过具有纳米大小的孔隙时,由离子电流流动跨过膜所引起的电信号模式用于确定核酸的序列。此类孔隙可例如但不限于由蛋白质(例如α溶血素、气单胞菌溶素(aerolysin)和包皮垢分枝杆菌孔蛋白A(Mycobacterium smegmatis porin A,MspA))或合成材料(诸如硅或石墨烯)产生(Magi等人,《生物咨询学简报(Brief Bioinform.)》2018;19:1256-1272)。

[0072] 在一个实施方案中,双股DNA分子会经历末端修复过程。此过程将DNA转化至钝端DNA,接着添加促进测序适配子连接的A尾端。各自携载马达蛋白的测序适配子(即,马达适配子)(例如,马达蛋白108)连接至DNA分子的两端。测序过程是在马达蛋白(例如,马达蛋白112)松解双股DNA时开始,使得第一股能够穿过纳米孔。当DNA股穿过纳米孔116时,传感器(例如电极)根据序列背景以及相关碱基修饰(称作一维(1D)读数))测量随时间推移(毫秒,ms)的离子电流变化(以皮安(pA)为单位)。曲线图120展示实例电流信号与时间。在另一实施方案中,发夹序列适配子将用于将第一股和其互补股共价栓系在一起以形成双股DNA分子。因此,在测序期间,测序双股DNA分子的一个股,接着测序互补股(称作$1D^2$或二维(2D)读数),可潜在地改良测序准确性。在又一实施方案中,通过蛋白质栓系的双股DNA分子的一个末端将增加在完成测序同一分子的第一股之后测序互补股的可能性,从而产生$1D^2$读数。

[0073] 原始信号(例如曲线图120中的电流)用于碱基识别和碱基修饰分析。在一些实施方案中,藉助于机器学习方法,例如但不限于递归神经网络(RNN)、卷积类神经网络(CNN)、隐藏式马可夫模型(HMM)或其一个或多个组合实施碱基识别和碱基修饰分析。

[0074] 在一个实施方案中,我们研发出一种处理通过纳米孔测序产生的电流信号的新颖方法,且分析经处理的信号以基于卷积类神经网络(CNN)或递归神经网络(RNN)确定单分子水平下的DNA甲基化。

II.电流信号分析

[0075] 可分析来自纳米孔测序的电流信号以鉴别碱基修饰。然而,图1中描述的机器学习方法不仅仅使用使用纳米孔获得的原始电流的输入。本文所描述的实施方案使用电流的部分的一个或多个统计值。这些一个或多个统计值的向量可与对应于核苷酸的窗口的其他信

息(包含核苷酸的身份和核苷酸的位置)组合。核苷酸的位置可相对于窗口内的目标位置,其中目标位置为检测到修饰或缺失的位置。可包含核苷酸的窗口的信息以及核酸分子的区域中的电信号的统计值以形成输入数据结构。在这些输入数据结构上训练的模型可用于检测碱基修饰。

A.电流向量参数

[0076] 对于穿过纳米孔的核苷酸股,我们将检测N个事件(即,与鉴别出的不同核苷酸相关的信号区段)。在一个实施方案中,一个事件对应于在碱基识别期间鉴别出的一个核苷酸,其中在特定单位时间(例如,毫秒)采样一系列电信号。在一个实例中,以4kHz的频率对电流进行采样(Rang等人,《基因组生物学(Genome Biol.)》2018;19:90)。在另一实施方案中,一个事件对应于在碱基识别期间鉴别出的超过一个核苷酸,其中以特定时间速率采样一系列电信号。

[0077] 图2展示电流信号的曲线图。y轴为以皮安为单位的电流振幅。x轴为以毫秒为单位的时间。圆点(例如圆点204)展示个别信号测量。通过相邻圆点的线(例如线208)指示具有与核苷酸相关的信号测量的信号区段(例如线208的A)。对于事件i,假设存在m_i电流信号,关于事件i的电流信号j的振幅由P_{ij}表示。在一个实施方案中,对于一个核苷酸,包含X1、X2、X3、X4和X5的信号特征向量用于表征与该核苷酸相关的电信号的模式。X1、X2和X3的定义在图2中示出。X1为P_{ij}的平均值。X2为P_{ij}的标准偏差。X3为P_{ij}的中值。X4为电流相对于X3的绝对偏差的中值(在图2中仅标记一个绝对偏差)。X5为X1相对于电流信号的平均值的差除以标准偏差。X5可视为一个区段的电流信号的z-评分。

[0078] 在一个实施方案中,P_{ij}可为归一化信号。归一化可涉及通过使用与部分或整个核苷酸股相关的最小值和最大值从初始范围重新调整电流信号以使得经归一化的信号值在0与1的范围内。归一化可涉及重新调整电流信号以使得经归一化的信号值的平均值为0且标准偏差为1。归一化可涉及通过使用与部分或整个核苷酸股相关的中值和偏差来重新调整电流信号。

[0079] X1和X2表示与事件i相关的P_{ij}的平均值和标准偏差。

[0080] X1通过以下定义:

$$X1 = \frac{\sum_{j=1}^{j=m_i} P_{ij}}{m_i}$$

[0081] X2通过以下定义:

$$X2 = \sqrt{\frac{\sum_{j=1}^{j=m_i} (P_{ij} - X1)^2}{m_i - 1}}$$

[0082] X3通过以下界定:

X3=中值(P_{ij}),

其中i在l至r范围内,包含查询碱基修饰分析的碱基周围的事件(例如CpG位点处的甲基化)。变量l和r表示一系列事件(对应于核苷酸序列)的窗口的左右侧。l与r之间的核苷酸序列应通常比下文论述的电流信号模式的整合式表示矩阵(称为IPM)长。对于给定事件i,j在1至m_i的范围内。X3可为用于确定所有区段的中值电流信号。X3对于所有区段可为相同的值,因为X3系使用不止一个区段的电流确定的。在一些实施方案中,X3可用于特定窗

口。在其他实施方案中,X3可为跨越多个窗口的中值。

[0083] X4通过以下定义:

X4=中值(|P_{ij}-X3|),

其中|·|表示绝对值;且i在1至r范围内,包含查询碱基修饰分析的碱基周围的事件(例如,CpG位点处的甲基化)。对于给定i,j在1至m_i的范围内。X4可为用于确定所有区段的电流信号的绝对偏差的中值。X4可使用不止一个区段的电流(例如使用所有采样的电流值)来计算且因此对于所有区段可为相同的值。

[0084] X5通过以下定义:

$$X5 = \frac{X1-\mu}{\sigma},$$

其中$\mu = \frac{\sum_{i=l}^{i=r}\sum_{j=1}^{j=m_r} P_{ij}}{M-1}$且$\sigma = \sqrt{\frac{\sum_{i=l}^{i=r}\sum_{j=1}^{j=m_r}(P_{ij}-\mu)^2}{M-1}}$

其中i在1至r范围内,包含查询碱基修饰分析的碱基周围的事件(例如CpG位点处的甲基化)。对于给定i,j在1至m_i的范围内。M为针对在1至r范围内的事件采样的电流信号的总数目。与多个电流信号相关且用于确定X3的区域大小可为DNA片段的大小。举例而言,若DNA片段为500bp,则区域的大小为500。若片段为300bp,则区域的大小为300。在一些实施方案中,将DNA片段进一步划分成较小子片段以确定X3可能为有用的。用于确定X3的区域大小可为5nt、10nt、20nt、30nt、40nt、50nt、60nt、70nt、90nt、100nt、200nt、300nt、400nt、500nt、600nt、800nt、900nt、1kb、2kb、3kb、4kb、5kb、10kb、50kb等。

[0085] X1和X2可用于反映在事件i内的信号变化,表示各核苷酸的电信号的局部模式。X3、X4和X5可用于反映事件i相对于在1至r范围内的其他周围事件的信号变化。在一些实施方案中,周围事件可为查询碱基修饰分析的碱基的X-nt的上游和Y-nt的下游。X可包括但不限于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可包括但不限于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。在一个实施方案中,周围事件可为穿过纳米孔的整个核苷酸股。

B.单股分析

[0086] 图3展示电流信号的曲线图。y轴为以皮安为单位的电流振幅。x轴为以毫秒为单位的时间。迹线304为随时间推移的电流振幅。信号区段(例如区段308)为迹线304的与核苷酸相关的部分。电流变化将视穿过纳米孔的不同核苷酸而变化。纳米孔测序中的碱基识别通常依赖于将电流信号转化成不同的局部静止状态(即,事件)。将电流信号转化成不同事件的过程称为电信号分段。离子电流变化包括但不限于对应于信号区段中的一个或多个核苷酸的事件振幅(例如以皮安,pA为单位测量)、离子电流的方向、对应于信号区段中的一个或多个核苷酸的电流事件的持续时间、离子电流的变化率和不同信号区段之间的相对振幅。振幅可指电流的强度或量值且不一定暗示交流电。使用例如命名Tombo的软件将那些电流事件分配至不同碱基(Stoiber等人,bioRxiv.2016;doi.org/10.1101/094672)。一个核苷

酸将与一系列具有不同振幅的事件相关。此类工具(Tombo)试图测试分配至两个样本之间的基因组碱基的纳米孔信号的差异以基于曼-惠特尼U-测试(Mann-Whitney U-test)来推断此类碱基经修饰或未经修饰(Stoiber等人,bioRxiv.2016;doi.org/10.1101/094672)。此工具(Tombo)不考虑上游和下游信号以及序列背景,且不能分析单分子水平下的甲基化模式,因为来自不同序列读数的所有信号会汇聚至基因组碱基中。已比较Tombo的性能与诸如Nanopolish和DeepSginal的其他工具的那些性能(Yuen等人,bioRxiv.2020;doi:doi.org/10.1101/2020.10.14.340315)。

[0087] 在一个实施方案中,为表征信号区段内与核苷酸相关的电流模式,计算该信号区段内的事件的那些电流振幅的平均值(X1)和标准偏差(X2)。确定与整个分子相关的事件的电流振幅的中值(X3)和与整个分子相关的事件的电流振幅的中值绝对偏差(X4)。通过下式确定信号区段的归一化信号(X5):

$$X5 = \frac{X1-\mu}{\sigma},$$

其中X1为该信号区段内与所讨论的核苷酸相关的事件的那些电流振幅的平均值;μ为所研究的整个分子内的事件的那些电流振幅的平均值;σ为所研究的整个分子内的事件的那些电流振幅的标准偏差。在一个实施方案中,可在移除最大值和最小的指定小百分比之后得到平均值和标准偏差。

[0088] 对于一个核苷酸,信号特征向量,包含X1、X2、X3、X4和X5,用于反映与该核苷酸相关的电信号的模式。举例而言,区段308可具有[X1,X2,X3,X4,X5]的信号特征向量。

[0089] X1和X2表示在信号区段i内的事件的电流振幅的平均值和标准偏差。X3表示与整个分子相关的事件的电流振幅的中值。X4表示与整个分子相关的事件的电流振幅的中值绝对偏差。X5表示信号区段i的归一化信号。

[0090] 图4为信号区段的长度的频率的曲线图。与核苷酸相关的电流事件的长度(即,以毫秒为单位的持续时间)是在x轴上。长度的频率展示于y轴上。图4展示与核苷酸相关的各信号区段的长度为可变的,其中中值为9(范围:1至3540)。

[0091] 碱基修饰将影响与其上游和下游核苷酸相关的电信号。在本公开内容中,我们共同地利用与用于碱基修饰分析的核苷酸相关的电流信号、与所关注的核苷酸附近的核苷酸相关的电流信号以及测序背景,以便改良性能。CpG位点处的DNA甲基化(即,胞嘧啶的第5个碳处的甲基化)为脊椎动物基因组中最常见的碱基甲基化类型。对CpG位点处的DNA甲基化的分析用作本公开内容的说明性实例。

[0092] 图5展示使用经由纳米孔测序的一个股的电流信号来确定甲基化的方法。在框504处,提供双股DNA分子。在框508处,使双股DNA分子与适用于纳米孔测序的测序适配子连接。在框512处,进行纳米孔测序。单双股分子的一个股移动通过内嵌于膜中的孔隙,从而改变流动通过纳米孔的离子电流信号。在框516处,获得电流信号。可例如通过跨电极来测量离子电流信号。

[0093] 将通过使用例如Tombo的分段步骤处理电流信号(Stoiber等人,bioRxiv.2016;doi.org/10.1101/094672)。这些分段式电事件将分配至不同核苷酸。在框520处,建构整合式表示矩阵(IPM)。IPM为电流信号模式的矩阵,其包含每个碱基的电流信号、测序背景和跨越用于碱基修饰分析的基因座附近或周围的一系列核苷酸的测序位置信息。在一个实施方

案中,与核苷酸相关的分段式电事件通过信号特征向量,即,[X1,X2,X3,X4,X5]描述。CpG位点内的胞嘧啶和例如该胞嘧啶的上游和下游10-nt(即,例如总共21nt)以及多个信号特征向量用于形成电流信号模式的IPM。出于说明的目的,5'-T[CCATGC]CATCGTC[GATGCA]G-3'的21-nt序列用作一实例,得到IPM 524。为简单起见,省略括号中的碱基(由"…"表示)。对于与腺嘌呤的碱基("A")对应的-2位置,与"A"相关的信号特征向量,[X1=1.7,X2=0.29,X3=24.2,X4=436,X5=-0.3]填写在"-2"行与"A"列之间的对应单元格中。相同行中的其他单元格填写为"0"。使用相同的规则填写与21-nt序列上下文相关的每个核苷酸的剩余信号特征向量,由此形成21-nt IPM。因此,此IPM将同时对电流信号模式、测序背景、测序位置以及随时间推移而改变的模式进行编码。源自甲基化和未甲基化DNA数据集的多个IPM用于训练CNN或RNN模型,该模型随后将用于确定测试样本中CpG位点处的甲基化状态。

[0094] 框528展示CNN分析。对于CNN分析,将IPM馈入输入层中,接着为卷积层和输出层的处理。CpG的甲基化概率(即,输出甲基化评分,在0至1的范围内)是基于输出层中的sigmoid函数来确定。此方法称为IPM-CNN。在一个实施方案中,甲基化CpG位点(M.SssI处理过的DNA)和未甲基化CpG位点(全基因组扩增(WGA)的DNA)的IPM用于训练CNN模型。自经M.Sss处理的DNA获得的数据集中CpG位点的甲基化目标值定义为"1",而自WGA DNA获得的数据集中CpG位点的甲基化目标值定义为"0"。通过经由迭代地更新模型参数使由sigmoid函数计算的输出评分与所要目标输出(二进制值:0或1)之间的总预测误差最小化来获得IPM-CNN的最佳参数。总预测误差是通过深度学习算法(keras.io/)中的sigmoid交叉熵损失函数来确定。从训练数据集得知的模型参数用于分析测试数据集中的甲基化状态,输出表明CpG位点被甲基化的可能性的概率性评分(即甲基化概率)。在一个实施方案中,CNN模型利用四个二维(2D)卷积层,各自具有32、64、128、256个核尺寸为25的过滤器。那些卷积层使用矫正线性单元(ReLU)的激活函数。随后应用批次归一化层。进一步增加一个扁平化层,接着丢弃速率为0.5的丢弃层,且接着为全连接层,该全连接层包括200个使用ReLU激活函数的神经元。最终应用具有一个神经元的输出层,利用sigmoid激活函数得到CpG位点甲基化的概率评分(即甲基化概率)。CNN模型的程序是基于Keras深度学习框架(https://keras.io/)实施。

[0095] 框532展示RNN分析。对于RNN分析,将IPM馈入输入层中,接着为长短期记忆(LSTM)层和输出层的处理。CpG的甲基化概率(在0至1的范围内)是基于输出层中的sigmoid函数来确定。此方法称为IPM-RNN。使用与IPM-RNN中使用的训练程序类似的训练程序,通过经由迭代地更新模型参数而使由sigmoid函数计算的输出评分与所要目标输出(二进制值:0或1)之间的总预测误差最小化来获得IPM-RNN的最佳参数。从训练数据集得知的模型参数用于分析测试数据集中的甲基化状态,输出表明CpG位点被甲基化的可能性的概率性评分(即甲基化概率)。在一个实施方案中,将具有LSTM单元的RNN模型与两个全连接隐藏层一起使用,该两个全连接隐藏层各自具有256个隐藏节点。最后一层之后为具有丢弃速率0.2的丢弃层。最终应用具有一个神经元的输出层,利用sigmoid激活函数得到CpG位点甲基化的概率性评分(即甲基化概率)。CNN模型的程序是基于Keras深度学习框架(keras.io/)实施。

C.双股分析

[0096] 图6展示使用两个DNA股的电流信号经由纳米孔测序确定甲基化的方法。在一个实施方案中,当双股DNA分子以第二核苷酸股(称为互补股或克里克股(Crick strand))将在第一核苷酸股(称为沃森股(Watson strand))完成穿过纳米孔之后紧接着穿过同一纳米孔

的方式测序时,可获得此类双股DNA分子的两个核苷酸股的电流信号。用于在同一纳米孔对双股DNA的两个核苷酸股进行依序测序的此类技术称为$1D^2$或2D测序。在框604处,提供双股DNA分子。在框608处,使双股DNA分子与适用于纳米孔测序的测序适配子连接。在框612处,使单双股分子的一个股移动通过内嵌于膜中的孔隙,接着使互补股移动通过该孔隙。在框616处,获得每个双股DNA分子的两个股的电流信号。可通过跨电极来测量离子电流信号。所获得的电流信号用于推导DNA分子的核苷酸信息,该信息是使用Guppy(牛津纳米孔科技有限公司(Oxford Nanopore Technologies Ltd))进行测序(即,碱基识别)。在一些实施方案中,可使用其他碱基识别工具,包括但不限于Albacore(nanoporetech.com/)、WaveNano(Wang等人,《定量生物学(Quantitative Biology.)》,2018;6:359-368)、Chiron(Teng等人,《大数据科学(GigaScience.)》2018;7:giy037)、Flappie(github.com/nanoporetech/flappie)、Scrappie(github.com/nanoporetech/scrappie)等。

[0097] 以特定时间速率(例如毫秒)采样的电流信号将分配至所检测的不同核苷酸用于碱基修饰分析。将通过使用例如Tombo的分段步骤处理电流信号(Stoiber等人,bioRxiv.2016;doi.org/10.1101/094672)。这些分段式电事件将分配至不同核苷酸。在框620处,建构包含每个双股DNA分子的两个股的整合式表示矩阵(IPM)。在一个实施方案中,与核苷酸相关的分段式电事件通过信号特征向量,即,[X1,X2,X3,X4,X5]描述。获得互补股的对应碱基的信号特征向量,即[X1',X2',X3',X4',X5']。CpG位点内的胞嘧啶和例如该胞嘧啶的上游和下游10-nt(即,例如总共21nt)以及多个信号特征向量用于形成电流信号模式的IPM。获得相同的双股DNA分子的互补股中的对应碱基的IPM。合并自沃森股和克里克股得到的IPM,进而形成具有较高维度的新颖IPM矩阵以用于碱基修饰分析。

[0098] 在一些实施方案中,可使用其他计算工具将电流信号指配至不同核苷酸,包含NanoMod(Liu等人,《英国医学委员会基因组学(BMC Genomics.)》2019;20:78)、Albacore(nanoporetech.com/)、Chiron(Teng等人,《大数据科学(GigaScience.)》2018;7:giy037)、Nanopolish(Simpson等人,《自然方法学(Nat Methods.)》2017;13:407-410)、Scrappie(https://github.com/nanoporetech/scrappie)、UNCALLED(Kovaka等人,《自然生物技术(Nat Biotechnol.)》2020;doi:10.1038/s41587-020-0731-9)等。这些计算工具和针对双股分析描述的其他技术可用于单股分析。

[0099] 出于说明的目的,5'-T[CCATGC]CATCGTC[GATGCA]G-3'的21-nt序列作为一实例用作IPM 624的基础。IPM 624可类似于IPM524,但包含沃森股和克里克股两者。为简单起见,省略括号中的碱基(由"…"表示)。对于与沃森股中的腺嘌呤的碱基("A")对应的-2位置,与"A"相关的信号特征向量,即[X1=1.7,X2=0.29,X3=436,X4=24.2,X5=-0.3]填写在由"沃森股"指示的区域中的"-2"行与"A"列之间的对应单元格中。对于其在互补股(即克里克股)中的对应碱基"T",与"T"相关的信号特征向量,[X1'=-1.9,X2'=0.23,X3'=24.2,X4'=436,X5'=-1.4]填写在由"克里克股"指示的区域中的"-2"行与"T"列之间的对应单元格中。相同行中的其他单元格填写为"0"。在一些实施方案中,可改变信号特征向量中要素的次序。举例而言,可使用[X2,X1,X3,X4,X5]、[X2,X3,X4,X5,X1]、[X1,X3,X5,X4,X2]或其他组合。在一些实施方案中,信号特征向量的大小可不局限于5。举例而言,通过增加更多处理的电信号特征或原始电信号,信号特征向量的大小可包括但不限于6、7、8、9、10、15、20、30、40、50、100等。通过编辑或删除信号特征向量中的一些特征,信号特征向量的大小可包括但

不限于1、2、3、4。

[0100] 使用相同的规则填写与21-nt序列上下文相关的每个核苷酸的剩余信号特征向量,由此形成21-nt IPM。因此,此IPM将同时对电流信号模式、测序背景、测序位置以及随时间推移而改变的模式进行编码。源自甲基化和未甲基化DNA数据集的多个IPM用于训练CNN或RNN模型,该模型随后将用于确定测试样本中CpG位点处的甲基化状态。

[0101] 框628展示CNN分析。在实施方案中,CNN模型利用四个二维(2D)卷积层,各自具有32、64、128、256个核尺寸为1×25的过滤器。那些卷积层使用矫正线性单元(ReLU)的激活函数。随后应用批次归一化层。进一步增加一个扁平化层,接着丢弃速率为0.5的丢弃层,且接着为全连接层,该全连接层包括200个使用ReLU激活函数的神经元。最终应用具有一个神经元的输出层,利用sigmoid激活函数得到CpG位点甲基化的概率性评分(即甲基化概率)。CNN模型的程序是基于Keras深度学习框架(keras.io/)实施。在一些实施方案中,可改变核尺寸n×m,其中"n"可包括但不限于1、2、3、4、5、10、15、20、30、35、40、45、50、100等,且"m"可包括但不限于1、2、3、4、5、10、15、20、30、35、40、45、50、100等。

[0102] 图7为核尺寸对碱基修饰分析的性能的影响的表。第一列展示不同的核尺寸。第二列展示训练数据集的AUC(ROC[接受者操作特征]曲线下面积)。第三列展示测试数据集的AUC。图7展示一系列核尺寸,诸如1×5、1×10、1×15、1×20和1×25将在区分甲基化CpG位点和未甲基化CpG位点中提供相当的性能,如通过分别为0.96、0.96、0.97、0.96和0.96的AUC所指示。

[0103] 框632展示RNN分析。在实施方案中,将具有LSTM单元的RNN模型与两个全连接隐藏层一起使用,该两个全连接隐藏层各自具有256个隐藏节点。LSTM隐藏单元的电流输出是通过电流输入和储存于LSTM单元中的先前信息来确定。作为一个实例,与21-nt IPM的第一列指示的位置相关的信号特征向量[X1,X2,X3,X4,X5]被视为在特定时间步长下的LSTM单元的输入X_t。正向LSTM RNN将基于如下的运算根据时间步长递归地计算隐藏层H(Gers等人,《IEEE神经网络汇刊(IEEE Transactions on Neural Networks)》2001;12:1333-1340):

$$A_{t,F}=\mathrm{sigmoid}(W_{xa,F}X_{t,F}+W_{ha,F}H_{t-1,F}+W_{ca,F}\odot C_{t-1,F}+b_{a,F}),$$
$$F_{t,F}=\mathrm{sigmoid}(W_{xf,F}X_{t,F}+W_{hf,F}H_{t-1,F}+W_{cf,F}\odot C_{t-1,F}+b_{f,F}),$$
$$C_{t,F}=F_{t,F}\odot C_{t-1,F}+A_{t,F}\odot \tanh(W_{xc,F}X_{t,F}+W_{hc,F}\odot C_{t,F}+b_{c,F}),$$
$$O_{t,F}=\mathrm{sigmoid}(W_{xo,F}X_{t,F}+W_{ho,F}H_{t-1,F}+W_{co,F}\odot C_{t,F}+b_{o,F}),$$
$$H_{t,F}=O_{t,F}\odot \tanh(C_{t,F})。$$

[0104] 反向LSTM RNN将基于如下的运算根据时间步长递归地计算隐藏层H(Gers等人,《IEEE神经网络汇刊》2001;12:1333-1340):

$$A_{t,B}=\mathrm{sigmoid}(W_{xa,B}X_{t,B}+W_{ha,B}H_{t-1,B}+W_{ca,B}\odot C_{t-1,B}+b_{a,B}),$$
$$F_{t,B}=\mathrm{sigmoid}(W_{xf,B}X_{t,B}+W_{hf,B}H_{t-1,B}+W_{cf,B}\odot C_{t-1,B}+b_{f,B}),$$
$$C_{t,B}=F_{t,F}\odot C_{t-1,B}+A_{t,B}\odot \tanh(W_{xc,F}x_{t,B}+W_{hc,B}\odot C_{t,B}+b_{c,B}),$$
$$O_{t,B}=\mathrm{sigmoid}(W_{xo,B}X_{t,B}+W_{ho,B}H_{t-1,B}+W_{co,B}\odot C_{t,B}+b_{o,B}),$$
$$H_{t,B}=O_{t,B}\odot \tanh(C_{t,B})。$$

其中W和b为权重和偏差;X为输入向量;A为输入门的激活向量;F为遗忘门的sigmoid函数;C为单元状态;O为输出门的sigmoid函数且H为LSTM隐藏单元的输出。

[0105] 将正向和反向LSTM RNN单元的输出合并。

$Z_t = H_{t,F} \oplus H_{t,B}$ 。

[0106] LSTM RNN输出的最后一层之后为具有丢弃速率0.2的丢弃层。最终应用具有一个神经元的输出层,利用sigmoid激活函数得到CpG位点甲基化的概率性评分(即甲基化概率)。CNN模型的程序是基于Keras深度学习框架(keras.io/)实施。

D.参数分析

[0107] 分析不同电流向量参数和不同窗口大小对AUC(ROC[接受者操作特征]曲线下面积)的影响。我们根据本公开内容提供的实施方案基于IPM-CNN模型分析在使用IPM中的不同参数的情况下的区分能力。为此目的,分别分析来自WGA DNA和M.SssI处理过的DNA数据集的8,282个分子(38,238个CpG位点)和8,247个分子(39,708个CpG位点)。

[0108] 图16展示不同参数组合对AUC的影响的曲线图。电流向量参数的不同组合在x轴上,且AUC在y轴上。图16展示使用IPM中但不限于X1、X2、X3、X4和X5的不同参数组合会产生CpG甲基化分析的不同性能。举例而言,使用IPM中的X1产生0.954的AUC,而IPM中X1和X2的组合产生0.893的AUC。IPM中X1、X2和X3的组合使AUC提高至0.963。IPM中X1、X2、X3和X4的组合使AUC进一步提高至0.978,接着在此实例中使用X1、X2、X3、X4和X5的情况下使性能平稳在0.977的AUC。因此,在一些实施方案中,IPM中的不同参数组合将允许确定在区分甲基化和未甲基化CpG位点中的所需性能。

[0109] 测试单独地而非组合地使用X1、X2、X3、X4和X5。单独地使用X1、X2、X3、X4和X5的结果分别为0.95、0.92、0.98、0.88和0.95的AUC。X3(即,区域中的P_{ij}的中值)得到0.98的高AUC。高AUC可至少部分为完整片段水平上的甲基化差异的结果。所使用的数据集涉及WGA(完全未甲基化)和M.SssI(完全甲基化)。然而,实际上片段将不为完全甲基化或完全未甲基化的。对于并非完全甲基化或完全未甲基化的样本,单独使用X3可不会产生高AUC。

[0110] 图17展示窗口大小对AUC的影响的曲线图。x轴展示以核苷酸为单位的窗口大小。y轴展示AUC。IPM中使用的核苷酸数目(又称窗口大小)将捕获在纳米孔测序期间产生的电流信号的不同信息内容,且可影响甲基化分析的性能。图17展示使用IPM-CNN模型区分甲基化和未甲基化CpG位点的性能呈现:随着IPM中使用的核苷酸数目从1nt增加至10nt,AUC从0.715逐步增加至0.969。在此实例中,在7nt的窗口大小处达到性能平稳。因此,在一些实施方案中,调节IPM的窗口大小将允许确定在区分甲基化和未甲基化CpG位点中的所需性能。

[0111] 实施方案可不需要使用产生最高AUC的电流向量参数或窗口大小的组合。较低AUC对于某些用途可能足够,或较高AUC可不值得与额外参数相关的额外计算和储存成本。此外,可调节不同参数以实现期望AUC、特异性和/或灵敏度。举例而言,较大窗口大小可用于补偿较少使用X1、X2、X3、X4和X5中的参数。

E.6mA修饰的检测

[0112] 为确定电流信号分析对除5mC以外的修饰的适用性,使用电流信号分析来检测N6-甲基腺嘌呤(6mA)。

[0113] 图18展示使用经由纳米孔测序的一个股的电流信号来确定6mA甲基化的方法。图18类似于展示用于确定5mC甲基化的方法的图5。在框1804处,提供双股DNA分子。在框1808处,使双股DNA分子与适用于纳米孔测序的测序适配子连接。在框1812处,进行纳米孔测序。在框1816处,获得电流信号。在框1820处,建构整合式表示矩阵(IPM)。框1804至1820可与框504至520相同。

[0114] 出于确定6mA甲基化的说明目的,5'-G[TACCCG]GGTACTG[TCTAGA]G-3'的21-nt序列作为一实例用作IPM的基础,以进行甲基化分析的核苷酸A(例如对应于0位置)为中心。IPM 1824展示使用21-nt序列的结果。为简单起见,省略括号中的碱基(由"…"表示)。对于与一个股中的腺嘌呤的碱基("A")对应的0位置,与"A"相关的信号特征向量(即,[X1=0.39,X2=0.04,X3=389,X4=46.3,X5=0.32])填写在矩阵的"0"行与"A"列之间的对应单元格。相同行中的其他单元格填写为"0"。在一些实施方案中,可改变信号特征向量中要素的次序。举例而言,可使用[X2,X1,X3,X4,X5]、[X2,X3,X4,X5,X1]、[X1,X3,X5,X4,X2]或其他组合。在一些实施方案中,信号特征向量的大小可不仅为5。举例而言,通过增加更多处理的电信号特征或原始电信号,信号特征向量的大小可包括但不限于6、7、8、9、10、15、20、30、40、50、100等。通过编辑或删除信号特征向量中的一些特征,信号特征向量的大小可包括但不限于1、2、3或4。

[0115] 使用相同的规则填写与21-nt序列上下文相关的每个核苷酸的剩余信号特征向量,由此形成21-nt IPM。因此,此IPM将同时对电流信号模式、测序背景、测序位置以及随时间推移而改变的模式进行编码。源自与核苷酸A相关联的甲基化和未甲基化DNA数据集的多个IPM用于训练CNN或RNN模型,该模型随后将用于确定测试样本中A位点处的甲基化状态。框1828展示CNN分析,且框1832展示RNN分析。这些框可与框528和532相同。

[0116] 为测试上文示出的我们的方法(IPM-CNN或IPM-RNN)是否能够确定腺嘌呤甲基化(6mA),我们下载包括来自先前研究(Rand等人,《自然方法》2017;14:411-413)的pUC19质粒DNA的纳米孔测序结果的两个公共数据集。第一数据集(6mA数据集)是由在含有dam和dcm甲基转移酶两者的大肠杆菌(E.coli)中生长的pUC19质粒DNA产生,其中所有GATC基序经推测为A位点均甲基化。第二数据集(uA数据集)是由用未经修饰的核苷酸进行PCR扩增的DNA产生,其中所有A位点经推测为未甲基化。在训练程序中,我们使用IPM-CNN模型分析来自6mA数据集的2052个含有GATC基序的分子和来自uA数据集的2081个分子。

[0117] 图19展示使用IPM-CNN模型得到的AUC。x轴展示特异性。y轴展示灵敏度。线1904展示训练数据集的结果。训练数据集的AUC为0.94。在训练程序中,我们将训练的IPM-CNN模型应用于来自6mA数据集的522个含有GATC基序的分子和来自uA数据集的481个分子。测试数据集的AUC为0.92。另外,当使用IPM-RNN模型时,对于训练和测试数据集两者均得到0.89的AUC。这些数据表明IPM-CNN和IPM-RNN可允许区分6mA位点与未甲基化A位点。

[0118] 在实施方案中,用于人类或非人类DNA的6mA确定的训练数据集可基于分别使用6mA核苷酸和未甲基化A核苷酸进行PCR扩增来建构。在几个PCR周期之后,大部分DNA分子将携载6mA核苷酸以用于由6mA核苷酸进行扩增的DNA产生的数据集,而大部分DNA分子将携载未甲基化A核苷酸以用于由未甲基化A核苷酸进行扩增的DNA产生的数据集。此两种类型的数据集可用于训练CNN和/或RNN模型以确定测试样本中A核苷酸的甲基化状态。

[0119] 使用电流信号分析检测除5mC之外的6mA证实此分析适用于其他甲基化类型。因此,这些方法应准确地检测本文所描述的其他甲基化。

F.人类个体的非肿瘤与肿瘤组织之间的CpG甲基化分析

[0120] 通过使用本文所描述的实施方案确定的位点的甲基化可用于区分不同类型的组织。使用根据本公开内容的实施方案的IPM-RNN模型,我们分析源自鼻咽癌(NPC)肿瘤和血沉棕黄层样本的细胞DNA分子的甲基化模式。为此目的,我们使用来自NPC肿瘤的147个分

子,其中中值大小为4,406bp(四分位数范围(IQR):1,962至8,128bp)且中值为32个CpG/分子(IQR:13至61)。我们分析来自血沉棕黄层的另外147个分子,其中中值大小为6,823bp(四分位数范围(IQR):2,515至9,304bp)且中值为49个CpG/分子(IQR:23至118)。

[0121] 图20展示来自血沉棕黄层样本和NPC肿瘤组织样本的DNA分子的比较图。x轴展示组织类型。y轴展示呈百分比形式的甲基化水平。发现血沉棕黄层中的单分子甲基化水平(即,分子中确定为甲基化的CpG位点的百分比)(中值:74.8%;IQR:71.1%至80.1%)显著高于NPC肿瘤中的单分子甲基化水平(中值:50;IQR:45.7至53.1)(P值<0.0001,威尔卡森秩和检定(Wilcoxon rank-sum test))。源自肿瘤组织的DNA分子呈现为低甲基化,其与基于短读数亚硫酸氢盐测序的先前结论一致(Chan等人,《美国国家科学院院刊》2013;110:18761-8)。然而,本文所述的新颖纳米孔测序技术允许对几乎整个长DNA分子进行测序,且分析DNA分子的甲基化模式。举例而言,纳米孔测序可分析大小大于600bp的DNA分子,其不能通过短读数测序平台(例如Illumina)进行查询。

[0122] 图21示出肿瘤DNA分子和血沉棕黄层DNA分子中的甲基化模式。实心黑色圆(例如圆2104)指示甲基化CpG位点。空心圆(例如圆2108)指示未甲基化CpG位点。圆展示CpG位点相对于所分析的DNA分子的5'端的相对位置(即,图中DNA分子的左侧更接近5'端)。如图21中所示,相比于源自血沉棕黄层样本的那些DNA分子,源自肿瘤组织的DNA分子倾向于在分子中携载更多未甲基化CpG位点。仅5.4%的来自血沉棕黄层样本的分子具有<50%的单分子甲基化水平和2,091bp的中值长度。相比之下,39.5%的来自NPC肿瘤组织的分子具有<50%的单分子甲基化水平和2,924bp的中值长度。DNA分子的长度在897bp至10,424bp范围内。

[0123] 这些数据展示本文所描述的用于检测甲基化的纳米孔测序技术可用于单分子甲基化模式分析以区分来自组织活检体样本的各DNA分子的组织来源(例如非肿瘤DNA与肿瘤DNA分子)。组织活检的单分子甲基化模式分析将允许检查肿瘤级别或亚型、监测癌症或其他疾病的治疗、评估器官异常(例如肾脏衰竭)等。

G.胎儿与母体DNA分子之间的分析

[0124] 通过使用本文所描述的实施方案确定的位点的甲基化可用于区分胎儿与母体DNA分子。根据IPM-CNN模型,我们通过1,262个胎儿特异性游离DNA分子(中值大小:530bp;IQR:361至779bp)和6,108个母体特异性游离DNA分子(中值大小:668bp;IQR:448至1,089bp)的至少5个CpG位点,利用母体血沉棕黄层与胎盘组织之间的SNP信息来确定单分子甲基化模式,所述分子获自妊娠三个月的孕妇。此孕妇的血浆DNA中的胎儿DNA分数为26.0%。

[0125] 图22展示母体特异性与胎儿特异性DNA分子之间的单分子甲基化水平。x轴展示游离DNA分子的类别:母体特异性或胎儿特异性。y轴展示呈百分比形式的单分子甲基化水平。单血浆DNA分子的中值甲基化水平(即,分子中确定为甲基化的CpG位点的百分比)对于胎儿特异性游离DNA分子为66.6%(IQR:28.5%至86.6%),其显著低于母体特异性游离DNA分子的中值甲基化水平(中值:78.5%;IQR:50%至93.7%)(P值:<0.0001,曼-惠特尼U测试)。结果表明使用游离DNA分子的甲基化信息允许区分各血浆DNA分子的母体和胎儿来源。

[0126] 另外,通过比较由IPM-CNN模型确定的甲基化模式与如2021年2月5日申请的美国专利申请第17/168,950号中所描述的血沉棕黄层和胎盘组织的各别参考甲基化模式,可得到0.87的AUC,以用于区分孕妇中胎儿和母体来源的血浆DNA分子。

[0127] 图23展示基于由IPM-CNN模型确定的甲基化模式对孕妇中的游离DNA分子进行胎儿和母体来源分析的ROC曲线。x轴为特异性,且y轴为灵敏度。

III.用于评估基于IPM的甲基化确定的数据集

[0128] 未甲基化数据集含有经由全基因组扩增(WGA)制备的经扩增DNA的测序结果(表示为WGA DNA数据集)。使用WGA中的未经修饰的核苷酸得到几乎不含碱基修饰的扩增DNA(除了少量输入基因组DNA以外)。甲基化数据集含有在测序之前通过M.SssI(CpG甲基转移酶,从含有来自螺原体属菌株MQ1的甲基转移酶基因的大肠杆菌菌株分离,将使双股DNA中的所有CpG位点甲基化)处理的DNA的测序结果(表示为M.SssI处理过的DNA数据集)。M.SssI甲基转移酶致使CpG位点甲基化。

[0129] 为制备WGA DNA数据集,通过将反应混合物(含有phi29反应缓冲液和dNTP)在95℃下的加热块中孵育5分钟接着冷却至4℃,将核酸外切酶抗性随机引物预退火至1ng的DNA模板。接着将phi29聚合酶添加至反应混合物中且在30℃下孵育4小时。DNA用Ampure XP珠粒纯化且用Qubit荧光计定量。通常,200ng DNA可获自20μl反应物。

[0130] 为制备M.SssI处理过的DNA数据集,在WGA之后,将一半DNA用M.SssI酶处理。将甲基转移酶反应缓冲液、S-腺苷甲硫胺酸(SAM)和M.SssI与DNA混合,且在37℃下孵育2小时。通过在65℃下加热20分钟使反应停止。连接测序试剂盒(SQK-LSK109)(牛津纳米孔)用于文库制备。用NEBNext FFPE DNA修复混合物以及NEBNext Ultra II末端修复/dA-加尾模块处理DNA。在Ampure XP珠粒清除之后,通过添加适配子混合物、连接缓冲液和NEBNext Quick T4 DNA连接酶将测序适配子连接至经修复的DNA。经连接的DNA用Ampure XP珠粒清洁且用短片段缓冲液洗涤。将文库再悬浮于洗脱缓冲液中。R9.4.1流通池用于对WGA(样本_01)和经M.SssI处理(样本_02)文库中的每一者进行测序。流通池首先用含有冲洗系链液(Flush Tether)和冲洗缓冲液的流通池预处理混合物进行预处理(primed)。接着通过混合测序缓冲液、负载珠粒和DNA文库来制备负载文库的混合物。以逐滴方式将负载文库的混合物添加至流通池样本口中。将负载的流通池插入PromethION中的狭槽中且使用默认参数测序64小时。

[0131] 针对样本_01和样本_02,我们分别获得1560和1530万纳米孔测序读数,其中1380(88.7%)和1380(90.7%)万读数可通过使用Minimap2(Li H,《生物信息(Bioinformatics)》2018;34(18):3094-3100)与人类参考基因组(UCSC hg19)对准。样本_01和样本_02的中值读数长度分别为510nt(四分位数范围(IQR):333至778nt)和606nt(IQR:382至911nt)。在一些实施方案中,BLASR(Mark J Chaisson等人,《BMC生物信息(BMC Bioinformatics)》2012;13:238)、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》2002;12(4):656-664)、BWA(Li H等人,《生物信息》2010;26(5):589-595)、NGMLR(Sedlazeck FJ等人,《自然方法》2018;15(6):461-468)和LAST(Kielbasa SM等人,《基因组研究》2011;21(3):487-493)可用于将经测序读数与参考基因组进行比对。

[0132] 图8为展示基于IPM用于训练和测试CNN和RNN模型的测序分子的数目的表。第一列为数据集。M.SssI处理过的DNA为甲基化DNA数据集,且WGA DNA为未甲基化DNA数据集。第二列为用于训练的分子数目和CpG位点数目。第三列为用于测试的分子数目和CpG位点数目。对于训练数据集,我们随机使用分别来自M.SssI处理过的DNA(甲基化DNA)和WGADNA(未甲

基化DNA)的7,989和8,052个测序分子。此训练数据集包括38,470个甲基化CpG位点和37,150个未甲基化CpG位点。对于测试数据集,我们随机使用分别来自M.SssI处理过的DNA(甲基化DNA)和WGADNA(未甲基化DNA)的4,826和5,041个测序分子。此训练数据集包括9,716个甲基化CpG位点和11,444个未甲基化CpG位点。

[0133] 图9A至图9D为使用IPM-CNN和IPM-RNN方法的WGA DNA与M.SssI处理过的DNA数据集之间的CpG的甲基化概率的盒状图。图具有在x轴上的数据集。甲基化概率在y轴上。图9A和图9B展示使用IPM-CNN分析的结果。图9A展示对训练数据集的IPM-CNN分析,其中M.SssI处理过的DNA数据集中CpG的甲基化概率(中值:0.99;IQR:0.987至0.999)显著高于WGA DNA数据集中的甲基化概率(中值:0.03;IQR:0.001至0.15)(P值＜0.0001,曼-惠特尼U测试)。图9B展示对测试数据集的IPM-CNN分析,其也展示WGA(中值:0.4;IQR:0.002至0.18)与M.SssI处理过的DNA数据集(中值:0.99;IQR:0.980至0.999)之间的CpG的甲基化概率的显著差异(P值＜0.0001,曼-惠特尼U测试)。

[0134] 图9C和图9D展示使用IPM-RNN分析的结果。图9C展示对训练数据集的IPM-RNN分析,其中M.SssI处理过的DNA数据集中CpG的甲基化概率(中值:0.994;IQR:0.92至0.99)显著高于WGA DNA数据集中的甲基化概率(中值:0.079;IQR:0.059至0.118)(P值＜0.0001,曼-惠特尼U测试)。图9D展示对测试数据集的IPM-RNN分析,其也展示WGA(中值:0.077;IQR:0.057至0.115)与M.SssI处理过的DNA数据集(中值:0.994;IQR:0.919至0.999)之间的CpG的甲基化概率的显著差异(P值＜0.0001,曼-惠特尼U测试)。这些结果表明,根据本公开内容提供的实施方案使用由纳米孔测序产生的电信号确定CpG位点的甲基化状态是可行的。在一个实施方案中,0.5的甲基化概率阈值可用于确定CpG位点的甲基化状态。在使用此阈值的情况下,对于IPM-CNN分析,DNA甲基化检测的特异性和灵敏度对于训练数据集分别为96％和91％,且对于测试数据集分别为93％和88％。对于IPM-RNN分析,DNA甲基化检测的特异性和灵敏度对于训练和测试数据集两者分别为97％和88％。在一些实施方案中,可根据各种应用调节甲基化概率的阈值。

[0135] 图10A和图10B展示接受者操作特征(ROC)曲线分析。特异性展示于x轴上。灵敏度展示于y轴上。图10A展示训练数据集的结果。图10B展示测试数据集的结果。IPM-CNN结果以线1004和1008展示。IPM-RNN结果以线1012和1016展示。DeepMod(Liu等人,《自然通讯》2019;10:2449)结果以线1020和1024展示。Nanopolish(Liu等人,《自然通讯》2019;10:2449)结果以线1028和1032展示。基于IPM的CNN和RNN分析为训练和测试数据集两者供应良好性能,其中ROC曲线下面积(AUC)不小于0.95。相比于DeepMod(0.83)和nanopolish(0.91),基于IPM的CNN和RNN模型在测试数据集中产生ROC曲线下面积(AUC)为0.95和0.97的更佳性能。发现基于IPM的RNN或CNN与其他包含DeepMod和nanopolish的工具的所有比较的P值(DeLong测试)＜0.0001。这些结果表明IPM-CNN和IPM-RNN在DNA甲基化分析方面优于其他工具。

[0136] 图11为针对不同分析的给定特异性的灵敏度的表。第一列展示分析类型。第二列展示灵敏度。第三列展示特异性。图11展示在给定特异性下,IPM-CNN和IPM-RNN分析实现高得多的灵敏度。举例而言,在90％的特异性下,IPM-CNN和IPM-RNN分别分析实现90％和93％的灵敏度,而DeepMod和nanopolish方法分别实现仅53％和74％的灵敏度。在95％的特异性下,IPM-CNN和IPM-RNN分析分别实现86％和90％的灵敏度,而DeepMod和nanopolish方法仅

分别实现38％和55％的灵敏度。在99％的特异性下,IPM-CNN和IPM-RNN分析分别实现70％和83％的灵敏度,而DeepMod和nanopolish分别实现仅13％和16％的灵敏度。这些结果进一步证实,序列区段的电流信号模式的整合式表示矩阵将大大地提高DNA甲基化确定的准确性。具体而言,IPM-RNN在那些方法中产生最佳的性能。

[0137] 在一些实施方案中,对于IPM,经历碱基修饰分析的碱基周围的DNA链段的长度可为对称或不对称的。举例而言,该碱基的上游X-nt和下游Y-nt可用于碱基修饰分析。X可包括但不限于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可包括但不限于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。X和Y可相同或不同。

[0138] 在一些实施方案中,核酸中的碱基修饰将根据本公开内容中的实施方案在不同生物体中进行分析,所述生物体包括病毒、细菌、植物、真菌、线虫、昆虫和脊椎动物(例如人类)等。最常见的碱基修饰为将甲基添加至不同位置的不同DNA碱基中,即所谓的甲基化。在胞嘧啶、腺嘌呤、胸腺嘧啶和鸟嘌呤上均已发现甲基化,诸如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)、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)和4mT(O4-甲基胸腺嘧啶)。

[0139] 在一些实施方案中,可通过不同的统计和/或数学模型分析电流信号模式的整合式表示矩阵,所述模型包括但不限于线性回归、逻辑回归、深度递归神经网络(例如长短期记忆,LSTM)、贝氏分类(Bayes classifier)、隐藏式马可夫模型(HMM)、线性判别分析(LDA)、k均值聚类、具有噪声的基于密度的空间聚类应用(DBSCAN)、随机森林算法和支持向量机(SVM)。在又一实施方案中,自然语言处理将应用于电信号分析以进行碱基修饰分析。

[0140] 在一些实施方案中,可使用不同类型的纳米孔,包括但不限于生物纳米孔,诸如蛋白质α-溶血素和其通过蛋白质工程化技术的变化形式、由程序化细菌产生的孔蛋白、由合成材料、、石墨烯制成的固态纳米孔等。

[0141] 在实施方案中,这些方法可用于通过参考诸如人类参考基因组(hg19)的参考基因组设计引导RNA,例如长散布核组件(LINE)重复序列来靶向大量共享同源序列的长DNA分子。在一个实例中,此类分析可用于分析孕妇的母体血浆中的循环游离DNA,以检测胎儿非整倍体(Kinde等人《公共科学图书馆·综合(PLOS One)》2012;7(7):e41162。在实施方案中,去活化的或'死亡的'Cas9(dCas9)和其相关单引导RNA(sgRNA)可用于在不切割双股DNA分子的情况下富集靶向的长DNA。举例而言,sgRNA的3'端可经设计以携带额外通用短序列。可使用与该通用短序列互补的经生物素标记的单股寡核苷酸以捕获dCas9所结合的那些目标长DNA分子。在另一实施方案中,可使用经生物素标记的dCas9蛋白或sgRNA或两者以促进富集。

[0142] 在实施方案中,可执行尺寸选择以在对所关注的一个或多个特定基因组区域无限制的情况下使用包括但不限于化学方法、物理方法、酶促方法、基于凝胶的方法和基于磁珠

的方法或合并远不止所述途径的方法的途径富集长DNA片段。

IV.实例方法

[0143] 此部分展示使用机器学习模型检测碱基修饰和训练用于检测碱基修饰的机器学习模型的实例方法。

A.修饰的检测

[0144] 图12为与检测核酸分子中核苷酸的修饰相关的例示性方法1200的流程图。修饰可包含本文所描述的任何甲基化或任何氧化。氧化可为8-侧氧基-鸟嘌呤。在一些实施方案中,图12的一个或多个过程框可通过系统(例如测量系统1400)执行。在一些实施方案中,图12的一个或多个过程框可由与系统分离或包含该系统的另一装置或装置群组执行。另外或可替代地,图12的一个或多个程序框可通过测量系统1400的一个或多个组件执行,诸如检测器1420、逻辑系统1430、局部存储器1435、外部存储器1440、存储装置1445和/或处理器1450。

[0145] 在框1210处,接收输入数据结构。输入数据结构可对应于样本核酸分子中测序的核苷酸的窗口。通过测量对应于核苷酸的电信号来对样本核酸分子进行测序。电信号可为电流、电压、电阻、电感、电容或阻抗。可通过使用纳米孔进行测序。方法1200可进一步包括使用纳米孔对样本核酸进行测序。纳米孔可为本文所描述的任何纳米孔。

[0146] 输入数据结构可包括若干特性的值。针对窗口内的每个核苷酸的特性可包括核苷酸的身份、核苷酸相对于各个窗口内的目标位置的位置和包含电信号的对应于核苷酸的区段的第一区段统计值的向量。特性可包括核酸分子的等于或大于窗口的区域中电信号的第一区统计值。举例而言,输入数据结构可包括整合式表示矩阵[IPM]。

[0147] 核苷酸的身份可为碱基(例如A、T、C或G)。可经由利用纳米孔测序的碱基识别技术来确定碱基。碱基识别技术可使电信号的区段与核苷酸相关联。核苷酸的位置可为相对于目标位置的核苷酸距离。举例而言,当核苷酸在一个方向上距离目标位置一个核苷酸时,位置可为+1,且当核苷酸在相反方向上距离目标位置一个核苷酸时,位置可为-1。

[0148] 第一区段统计值可表示电信号的对应于核苷酸的区段的平均值。在一些实施方案中,第一区段统计值可表示电信号的对应于核苷酸的区段的电信号变化(例如标准偏差)。在实施方案中,第一区段统计值可表示电信号的对应于核苷酸的区段的平均值的归一化值。归一化可包括重新调整以使得第一区段统计值在某一范围(例如0至1的范围)内。归一化可包括使用部分或所有核苷酸股的中值、平均值和/或偏差。归一化可为本文所描述的任何归一化,包括z-评分(例如X5)。

[0149] 向量可包括第二区段统计值,其表示电信号的对应于核苷酸的区段的变化。向量可包括第三区段统计值,其表示第一区段统计值的归一化值。向量可包括本文所描述的变量X1、X2和X5的任何组合。

[0150] 第一区统计值可表示该区域中电信号的平均值或中值。举例而言,第一区统计值可为X3。在实施方案中,第一区统计值可表示电信号相对于该区域中的电信号的平均值或中值的变化的绝对值的中值或平均值。变化可为标准偏差。举例而言,第一区统计值可为X4。在一些实施方案中,第一区统计值可为可选的。

[0151] 输入数据结构可进一步包括第二区统计值,其表示电信号相对于该区域中的电信号的平均值或中值的变化的绝对值的中值或平均值。举例而言,第二区统计值可为X4。

[0152] 对于窗口内的不同核苷酸,第一区统计值可为相同值。对于窗口内的不同核苷酸,第二区统计值可为相同值。因此,第一区统计值和第二区统计值可视为与具有第一区段统计值和/或第二区段统计值的向量不同。替代地,对于每个核苷酸,向量也可包括第一区统计值和/或第二区统计值可包含在向量中,即使所述值在核苷酸之间为相同的。在IPM 524和IPM 624中示出重复所述区域统计值的途径。

[0153] 该区域可在样本核酸分子的一个股上。在一些实施方案中,该区域可在样本核酸分子的两个股上。窗口可包含样本核酸分子的两个股上的核苷酸。该区域可为样本核酸分子。该区域可包含至少5、10、15、20、25、30、50、100、200、300、400、500、1k、5k、10k、50k或1M个核苷酸。在一些实施方案中,该区域可少于50、100、200、300、400、500、1k、5k、10k、50k或1M个核苷酸。该区域可以目标位置处的核苷酸为中心。

[0154] 核苷酸的窗口可以目标位置处的核苷酸为中心。在一些实施方案中,窗口可不以目标位置处的核苷酸为中心。窗口可包含目标位置处的核苷酸的上游X-nt和下游Y-nt。X可包括但不限于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可包括但不限于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。窗口中核苷酸的最小数目可为2、3、4、5、6、7、8、9、10、20、30、40、50、100、200,或大于目标位置的上游和下游的核苷酸数目中任一者的和的数目。窗口可类似于图5中展示和描述的窗口。

[0155] 窗口可包含核酸分子的两个股,类似于图6所描述的技术。

[0156] 在框1220处,将输入数据结构输入至模型中。通过接收第一多个第一数据结构来训练模型。第一多个数据结构的每个第一数据结构对应于多个第一核酸分子的各个核酸分子中测序的核苷酸的各个窗口。通过测量对应于核苷酸的电信号来对第一核酸分子中的每一者进行测序。修饰在每个第一核酸分子的每个窗口中目标位置处的核苷酸中具有已知的第一状态。每个第一数据结构包含与输入数据结构相同的特性的值。模型可为本文所描述的任何机器学习模型。

[0157] 通过储存多个第一训练样本来进一步训练模型。每个第一训练样本包含第一多个第一数据结构中的一者和指示目标位置处的核苷酸的第一状态的第一标记。另外,当将第一多个第一数据结构输入至模型时,通过基于模型的匹配或不匹配第一标记的相应标记的输出使用多个第一训练样本使模型的参数优化而训练模型。模型的输出指定在各个窗口中目标位置处的核苷酸是否具有修饰。训练可如稍后图13所描述来进行。

[0158] 在框1230处,使用该模型确定修饰是否存在于输入数据结构中窗口内的目标位置处的核苷酸中。

[0159] 修饰状态可用于进一步分析中。在从孕妇获得的样本中,在本公开内容中的实施方案可用于基于甲基化状态确定血浆DNA分子的胎儿或母体来源。可通过具有比参考值更高或更低的甲基化水平的基因组区域确定母体或胎儿来源。在实施方案中,从孕妇获得的样本可为游离的,例如血浆或血清。在一些实施方案中,样本核酸分子可鉴别为与预定基因组区域对准。可已知预定基因组区域在胎儿或母体基因组中为高甲基化或低甲基化的。该

方法可包括使用目标位置处的核苷酸的修饰状态和视情况样本核酸分子的一个或多个其他核苷酸的修饰状态来确定样本核酸为胎儿来源还是母体来源。

[0160] 确定样本核酸分子为胎儿来源还是母体来源可包含使用一个或多个核苷酸的甲基化状态来确定样本核酸分子的甲基化水平。可将样本核酸分子的甲基化水平与参考值进行比较。参考值可由一个或多个母体核酸分子的甲基化水平来确定。将样本核酸分子的甲基化水平与参考值进行比较可包含确定样本核酸分子的甲基化水平低于参考值。确定样本核酸分子为胎儿来源还是母体来源可包含使用该比较确定样本核酸分子为胎儿来源。

[0161] 在一些实施方案中,样本核酸分子可为多个样本核酸分子中的一个样本核酸分子。该方法可进一步包括使用甲基化状态确定多个样本核酸分子中的每一者为胎儿来源还是母体来源。可使用对多个样本核酸分子的胎儿或母体来源的确定来确定胎儿分数。

[0162] 在一些实施方案中,修饰状态可用于确定拷贝数畸变是否存在于一区域中。修饰可为甲基化。样本核酸分子可为游离的且获自怀有胎儿的女性个体的生物样本。样本核酸分子可为多个样本核酸分子中的一个样本核酸分子。该方法可进一步包括将多个样本核酸分子鉴别为与胎儿基因组的区域对准。可确定多个样本核酸分子中的每个样本核酸分子的一个或多个核苷酸的修饰状态。可使用多个样本核酸分子中的每个样本核酸分子的一个或多个核苷酸的甲基化状态确定该区域的甲基化水平。该方法可进一步包括使用甲基化水平确定拷贝数畸变是否存在于胎儿基因组的区域中。该区域可为染色体,且该方法可进一步包括确定存在拷贝数畸变和确定胎儿具有染色体非整倍体。

[0163] 可确定修饰存在于一个或多个核苷酸处。可使用在一个或多个核苷酸处的修饰的存在来确定病症的分类。病症的分类可包含使用修饰的数目。可将修饰的数目与阈值进行比较。替代或另外地,分类可包含一个或多个修饰的位置。一个或多个修饰的位置可通过将核酸分子的序列读数与参考基因组比对来确定。若已知与病症相关的某些位置显示为具有修饰,则可确定病症。举例而言,甲基化位点的模式可与病症的参考模式进行比较,且可基于该比较确定病症。与参考模式的匹配或与参考模式的实质性匹配(例如,80%、90%或95%或更高)可指示病症或病症的可能性较高。病症可为任何妊娠相关的病症(例如子痫前症、宫内发育迟缓、侵入性胎盘形成和早产)。

[0164] 可分析统计学上显著数目个核酸分子以便提供对一个或多个怀孕个体中的病症、组织来源或临床相关的DNA分数的准确确定。在一些实施方案中,分析至少1,000个核酸分子。在其他实施方案中,可分析至少10,000或50,000或100,000或500,000或1,000,000或5,000,000个核酸分子。作为另一实例,可产生至少10,000或50,000或100,000或500,000或1,000,000或5,000,000个序列读数。

[0165] 该方法可包括确定病症的分类为个体患有该病症。分类可包含使用修饰的数目和/或修饰的位点的病症等级。

[0166] 可使用一个或多个核苷酸处的修饰的存在确定胎儿DNA分数、胎儿甲基化谱、母体甲基化谱、印记基因区域的存在。

[0167] 方法1200可包括额外的实施方案,诸如任何单一实施方案或下文描述和/或结合本文中在别处描述的一个或多个其他方法的实施方案的任何组合。

[0168] 尽管图12展示方法1200的实例框,但在一些实施方案中,相比于图12中所描绘的那些框,方法1200可包含额外的框、更少的框、不同的框或以不同方式配置的框。另外或替

代地,可并行地执行方法1200的框中的两者或更多者。

B.模型训练

[0169] 图13展示检测核酸分子中核苷酸的修饰的例示性方法1300。例示性方法1300可为训练用于检测修饰的模型的方法。该修饰可包含甲基化。甲基化可包含本文所描述的任何甲基化。该修饰可具有离散状态,诸如甲基化和未甲基化,且可能指定甲基化的类型。因此,核苷酸可能有多于两种状态(分类)。图13中的训练可与图12的方法1200一起使用。

[0170] 在框1310处,接收多个第一数据结构。本文描述数据结构的各种实例,例如在图5和图6中。第一多个第一数据结构中的每个第一数据结构可对应于多个第一核酸分子的各个核酸分子中测序的核苷酸的各个窗口。与第一多个数据结构相关的每个窗口可包含4个或更多个连续核苷酸,包含5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21或更多个连续核苷酸。每个窗口可具有相同数目的连续核苷酸。窗口可为重叠的。每个窗口可包含第一核酸分子的第一股上的核苷酸和第一核酸分子的第二股上的核苷酸。第一数据结构也可包含窗口内的每个核苷酸的股特性的值。股特性可指示核苷酸存在于第一股或第二股。窗口可包含第二股中与第一股中对应位置的核苷酸不互补的核苷酸。在一些实施方案中,第二股上的所有核苷酸均与第一股上的核苷酸互补。在一些实施方案中,每个窗口可包含第一核酸分子的仅一股上的核苷酸。

[0171] 第一多个第一数据结构可包含5,000至10,000、10,000至50,000、50,000至100,000、100,000至200,000、200,000至500,000、500,000至1,000,000或1,000,000或更多个第一数据结构。多个第一核酸分子可包含至少1,000、10,000、50,000、100,000、500,000、1,000,000、5,000,000或更多个核酸分子。作为另一实例,可产生至少10,000或50,000或100,000或500,000或1,000,000或5,000,000个序列读数。

[0172] 通过测量与核苷酸对应的电信号来对第一核酸分子中的每一者进行测序。电信号可来自纳米孔测序。

[0173] 修饰在每个第一核酸分子的每个窗口中目标位置处的核苷酸中具有已知的第一状态。第一状态可为核苷酸中不存在修饰,或可为核苷酸中存在修饰。可已知第一核酸分子中不存在修饰,或可对第一核酸分子进行处理以使得修饰不存在。可已知第一核酸分子中存在修饰,或可对第一核酸分子进行处理以使得修饰存在。若第一状态为不存在修饰,则修饰可在每个第一核酸分子的每个窗口中不存在,而非仅在目标位置不存在。已知的第一状态可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。可经由使用亚硫酸氢盐测序或使用单分子实时测序的光信号的技术来确定已知的甲基化第一状态。

[0174] 目标位置可为各个窗口的中心。对于具有跨越偶数个核苷酸的窗口,目标位置可为紧靠窗口中心的上游或紧靠下游的位置。在一些实施方案中,目标位置可在各个窗口的任何其他位置,包含第一位置或最后位置。举例而言,若窗口跨越一个股的n个核苷酸,自第1位至第n位(上游或下游),则目标位置可在第1位至第n位的任何位置。

[0175] 每个第一数据结构包含窗口内的特性的值。特性可为框1210处描述的特性中的任一者。

[0176] 在框1320处,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中的一者和指示目标位置处的核苷酸的修饰的第一状态的第一标记。

[0177] 在框1330处,接收第二多个第二数据结构。框1330为任选的。第二多个第二数据结构中的每个第二数据结构对应于多个第二核酸分子中的各个核酸分子中测序的核苷酸的各个窗口。第二多个核酸分子可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目标位置处的核苷酸中具有已知的第二状态。第二状态为与第一状态不同的状态。举例而言,若第一状态为存在修饰,则第二状态为不存在修饰,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的特性的值。

[0178] 在框1340处,储存多个第二训练样本。框1340为任选的。每个第二训练样本包含第二多个第二数据结构中的一者和指示目标位置处的核苷酸的修饰的第二状态的第二标记。

[0179] 在框1350处,使用多个第一训练样本和任选的多个第二训练样本训练模型。当将第一多个第一数据结构和任选的第二多个第二数据结构输入至模型时,通过基于模型的匹配或不匹配第一标记和任选的第二标记的相应标记的输出使模型的参数优化来进行训练。模型的输出指定在各个窗口中目标位置处的核苷酸是否具有修饰。该方法可仅包含多个第一训练样本,因为模型可将离群值鉴别为与第一状态不同的状态。该模型可为统计模型,也称为机器学习模型。

[0180] 在一些实施方案中,模型的输出可包含处于多个状态中的每一者的概率。可将具有最高概率的状态视为状态。

[0181] 该模型可包含卷积神经网络(CNN)。CNN可包含一组卷积过滤器,其经配置以过滤第一多个数据结构和任选的第二多个数据结构。过滤器可为本文所描述的任何过滤器。每层的过滤器的数目可为10至20、20至30、30至40、40至50、50至60、60至70、70至80、80至90、90至100、100至150、150至200或更多。过滤器的核尺寸可为2、3、4、5、6、7、8、9、10、11、12、13、14、15、15至20、20至30、30至40或更多。CNN可包含经配置以接收经过滤的第一多个数据结构和任选的经过滤的第二多个数据结构的输入层。CNN也可包含多个隐藏层,其包含多个节点。多个隐藏层中的第一层耦合至输入层。CNN可进一步包含输出层,其耦合至多个隐藏层的最后一层且经配置以输出输出数据结构。输出数据结构可包含所述特性。

[0182] 该模型可包含递归类神经网络(RNN)。RNN模型包含多个与测量窗口中的多个核苷酸相关联的长短期记忆(LSTM)单元。LSTM单元的数目可等于测量窗口中核苷酸的数目。在一些实施方案中,LSTM单元的数目可少于测量窗口中核苷酸的数目。LSTM单元的数目可为但不限于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、30、40、50、100、200、300、400、500、1,000、2,000、3,000、4,000、5,000、10,000、50,000等。一个LSTM单元可将与电流信号特征相关的信息传输至下一个LSTM单元,该信息将经历多轮线性或非线性变换。此类跨越LSTM单元的信息传输通常以顺序方式(例如根据时间步长)组构。此类跨越LSTM单元的信息传输可为双向的(即,包含时间顺序和备用时间顺序)。每个LSTM单元包含可程序化运算,诸如遗忘门、输入门、单元状态和输出门。经由那些运算,一个LSTM可确定来自前一时间步长的电流信号信息是否为记住的或不相关的且可被遗忘(遗忘门)。一个LSTM单元尝试学习自输入达至此单元(输入门)的新信息。该单元将更新的信息自当前时间步长传递至下一个时间步长(输出门)。本文中的单元状态携载该信息以及所有时间步长。可使用LSTM单元的多个层。LSTM层的数目可为1、2、3、4、5、6、7、8、9、10、15、20、30等。可使用各层之间的全连接。sigmoid函数通常用作输入门、输出门和遗忘门的门函数(gating function)。sigmoid函数的输出值可在0与1之间,从而确定没有信息流动通过所述门或信

息完全流动通过所述门。双曲正切激活函数(又称Tanh)可用作输出激活函数,其处理来自输出门的信息值以形成值在-1与1之间的新信息,该信息可传递至下一个LSTM单元。在一些实施方案中,可使用其他激活函数,包括但不限于二进制阶梯函数、线性激活函数、sigmoid函数、矫正线性单元等。由LSTM的最终层产生的值可传递至输出层(即,密集层,具有一定数目的神经元)上,其中每个神经元均为全连接。密集层中的神经元数目可为但不限于2、3、4、5、6、7、8、9、10、20、30、40、50、100、200、300、400、500、1000、2000个等。可使用多个密集层,包括但不限于1、2、3、4、5、6、7、8、9、10、20、30、40、50、100、5000、1000个等。输出层可输出甲基化评分,例如基于sigmoid激活函数或SoftMax激活函数,其可用于对甲基化状态进行分类。举例而言,若甲基化评分大于0.5,则确定碱基为甲基化。否则,确定碱基为未甲基化。在一些实施方案中,用于对甲基化状态进行分类的阈值可为但不限于至少0.1、0.2、0.3、0.4、0.6、0.7、0.8、0.9等。在一些实施方案中,可丢弃模型中的一些神经元以使过度拟合问题最小化。丢弃的神经元百分比可为但不限于1％、5％、10％、15％、20％、25％、30％、40％、50％、60％、70％等,其可根据不同层而不同。

[0183] 该模型可包含监督式学习模型。监督式学习模型可包含不同的方法和算法,包含分析学习、人工神经网络、后向传播、提升(boosting)(元算法)、贝氏统计、案例式推理、确定树学习、归纳逻辑程序设计、高斯过程回归(Gaussian process regression)、基因程序设计、数据分组处理方法、核估计法(kernel estimator)、学习自动机、学习分类系统、最小讯息长度(确定树、决策图等)、多线性子空间学习、单纯贝氏分类(naive Bayes classifier)、最大熵分类、条件随机场、最近相邻算法、可能近似正确学习(PAC)学习、涟波下降规则(rippledown rule)、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、分类集成、有序分类、数据预处理、处理不平衡数据集、统计关系学习或Proaftn(一种多准则分类算法)。模型可线性回归、逻辑回归、深度递归类神经网络(例如长短期内存,LSTM)、贝氏分类器、隐藏式马可夫模型(HMM)、线性判别分析(LDA)、k均值聚类、具有噪声的基于密度的空间聚类应用(DBSCAN)、随机森林算法、支持向量机(SVM)或本文所描述的任何模型。

[0184] 作为训练机器学习模型的一部分,机器学习模型的参数(诸如权重、阈值,例如可用于神经网络中的激活函数等)可基于训练样本(训练集)而经优化,以提供对目标位置处的核苷酸的修饰进行分类的优化准确度。可进行各种形式的优化,例如反向传播、经验风险最小化和结构风险最小化。可使用验证样本集(数据结构和标记)来验证模型的准确度。可使用训练集中用于训练和验证的各个部分来进行交叉验证。该模型可包括多个子模型,从而提供集合模型。子模型可为较弱的模型,一旦组合就提供更准确的最终模型。

V.例示性系统

[0185] 图14示出根据本发明的一实施方案的测量系统1400。如图所示的系统包含在样本架1410内的样本1405,诸如DNA分子,其中样本1405可与测定1408接触,以提供物理特征1415的信号。样本架的一实例可为包含测定的探针和/或引物的流通池或液滴藉以移动的套管(在包含液滴的测定的情况下)。通过检测器1420检测样本的物理特征1415(例如,荧光强度、电压或电流)。检测器1420可按时间间隔(例如,周期性间隔)进行测量,以获得构成数据信号的数据点。在一个实施方案中,模拟数字转换器在多个时间将来自检测器的模拟信号转换成数字形式。样品架1410和检测器1420可形成测定装置,例如根据本文所描述的实

施方案进行测序的测序装置。数据信号1425从检测器1420发送至逻辑系统1430。数据信号1425可储存于局部存储器1435、外部存储器1440或存储装置1445中。

[0186] 逻辑系统1430可为或可包含计算机系统、ASIC、微处理器等。其也可包含显示器(例如监测器、LED显示器等)和用户输入装置(例如鼠标、键盘、按钮等)。逻辑系统1430和其他组件可为独立的或网络连接的计算机系统的一部分,或其可直接连接至或并入包含检测器1420和/或样品架1410的装置(例如测序装置)中。逻辑系统1430也可包含在处理器1450中执行的软件。逻辑系统1430可包含计算机可读介质,其储存用于控制系统1400执行本文所描述的方法中的任一者的指令。举例而言,逻辑系统1430可向包含样品架1410的系统提供命令,使得执行测序或其他物理操作。此类物理操作可以特定次序进行,例如在试剂以特定次序添加和移除的情况下。此类物理操作可由可用于获得样本且执行分析的例如包含机械臂的机器人系统执行。

[0187] 本文所提及的任一种计算机系统可利用任何适合数目个子系统。此类子系统的实例展示于图15中的计算机系统10中。在一些实施方案中,计算机系统包含单一计算机设备,其中子系统可为计算机设备的组件。在其他实施方案中,计算机系统可包含具有内部组件的多个计算机设备,其各自为一个子系统。计算机系统可包含桌上型和膝上型计算机、平板计算机、移动电话、其他行动装置和基于云端的系统。

[0188] 图15中所示的子系统经由系统总线75互连。展示额外子系统,诸如打印机74、键盘78、一个或多个存储装置79、与显示器适配器82耦接的监测器76(例如显示屏幕,诸如LED)和其他装置。耦接至输入/输出(I/O)控制器71的周边装置和I/O装置可通过此项技术中已知的多种手段(诸如输入/输出(I/O)埠77(例如,USB、Lightning、Thunderbolt™))连接至计算机系统。举例而言,I/O端口77或外部接口81(例如以太网络(Ethernet)、Wi-Fi等)可用于将计算机系统10连接至广域网(诸如因特网)、鼠标输入设备或扫描仪。经由系统总线75互连允许中央处理器73与各子系统通信且控制系统内存72或存储装置79(例如,固定磁盘,诸如硬盘机,或光盘)执行多个指令,以及子系统之间的信息交换。系统内存72和/或一个或多个存储装置79可实施为计算机可读介质。另一子系统为数据收集装置85,诸如照相机、麦克风、加速计和其类似物。本文中所提和的数据中的任一者可自一个组件输出至另一组件且可输出至用户。

[0189] 计算机系统可包含多个相同组件或子系统,例如通过外部接口81、通过内部接口或经由可移式存储装置连接在一起,所述可移式存储装置可自一个组件连接至另一组件且移除。在一些实施方案中,计算机系统、子系统或设备可经由网络通信。在此类情况下,一台计算机可视为客户端,且另一台计算机视为服务器,其中各计算机可为同一计算机系统的一部分客户端和服务器各自可包含多个系统、子系统或组件。

[0190] 实施方案的方面可以控制逻辑形式使用硬件电路(例如特殊应用集成电路或场域可程序化门阵列)和/或使用具有大体上可程序化处理器的计算机软件以模块化或整合式方式来实施。如本文所使用,处理器可包含单核处理器、同一个积体芯片上的多核处理器或单一电路板或网络硬件以及专用硬件上的多个处理单元。基于本文所提供的揭示内容和教示内容,本领域中的一般熟习此项技术者将知道和了解使用硬件和硬件与软件的组合来实施本发明的实施方案的其他方式和/或方法。

[0191] 本申请案中所描述的任何软件组件或功能可使用例如习知或面向对象技术,以软

件程序代码形式实施,该软件程序代码是由使用任何适合计算机语言(诸如Java、C、C++、C#、Objective-C、Swift)或脚本处理语言(诸如Perl或Python)的处理器执行。软件程序代码可以一系列指令或命令形式储存于计算机可读取媒体上以进行储存和/或传输。适合的非暂时性计算机可读介质可包含随机存取内存(RAM)、只读存储器(ROM)、磁性媒体(诸如硬盘机或软盘驱动器)或光学媒体,诸如光盘(CD)或数字化通用光盘(DVD)或蓝光盘、闪存和其类似者。计算机可读介质可为此类储存或传输装置的任何组合。

[0192] 此类程序也可使用适用于经由符合多种协议的有线、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。因此,计算机可读取媒体可使用以此类程序编码的数据信号建立。以程序代码编码的计算机可读介质可与兼容装置一起封装或与其他装置分开提供(例如经由因特网下载)。任何此类计算机可读介质可存在于单一计算机产品(例如硬盘机、CD或整个计算机系统)上或其内部,且可存在于系统或网络内的不同计算机产品上或其内部。计算机系统可包含用于向使用者提供本文所提和的任何结果的监测器、打印机、或其他适合的显示器。

[0193] 本文中所描述的方法中的任一者可完全或部分地使用计算机系统来执行,该计算机系统包含可经配置以执行步骤的一个或多个处理器。因此,实施方案可针对经配置以执行本文所描述的任何方法的步骤的计算机系统,潜在地使用不同组件执行各别步骤或各别步骤群组。尽管以带编号的步骤形式呈现,但本文中的方法的步骤可同时或在不同时间或以不同顺序执行。另外,这些步骤的一部分可与其他方法的其他步骤的部分一起使用。另外,可视情况选用步骤的全部或部分。此外,任何方法的任何步骤可使用用于执行这些步骤的系统的模块、单元、电路或其他构件来执行。

[0194] 可在不脱离本发明的实施方案的精神和范畴的情况下以任何合适方式组合特定实施方案的特定细节。然而,本发明的其他实施方案可针对与各个别方面或这些个别方面的特定组合相关的特定实施方案。

[0195] 已出于说明和描述的目的呈现本公开内容的例示性实施方案的上述描述。其并不意欲为详尽的或将本公开内容限于所描述的精确形式,且鉴于以上教示,许多修改和变化为可能的。

[0196] 除非相反地特定指示,否则“一(a/an)”或“所述(the)”的叙述欲意谓“一个或多个(种)”。除非相反地特定指示,否则“或”的使用欲意谓“包括性的或”,而非“互斥性的或”。提和“第一”组件不一定需要提供第二组件。此外,除非明确陈述,否则提和“第一”或“第二”组件不会将所提和组件限制于特定位置。术语“基于”意指“至少部分地基于”。

[0197] 出于所有目的,本文所提和的所有专利、专利申请案、公开案和描述均以全文引用的方式并入。不承认任一者为先前技术。

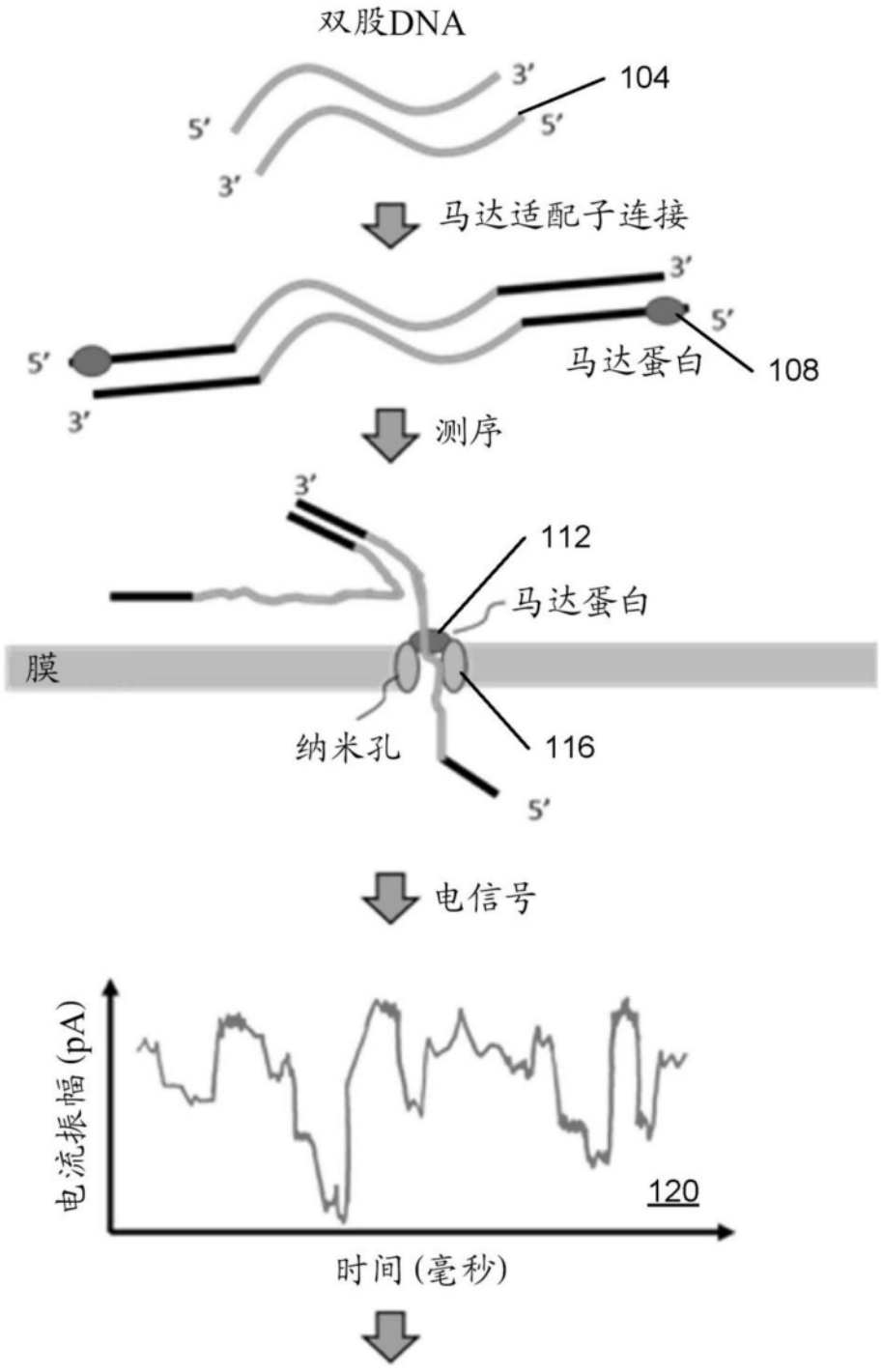

双股DNA
3′
104
5′
5′
3′
马达适配子连接
3′
5′
5′
马达蛋白
108
3′
测序
3′
112
马达蛋白
膜
纳米孔
116
5′
电信号
电流振幅(pA)
120
时间(毫秒)
1)碱基识别
2)碱基修饰分析

图1

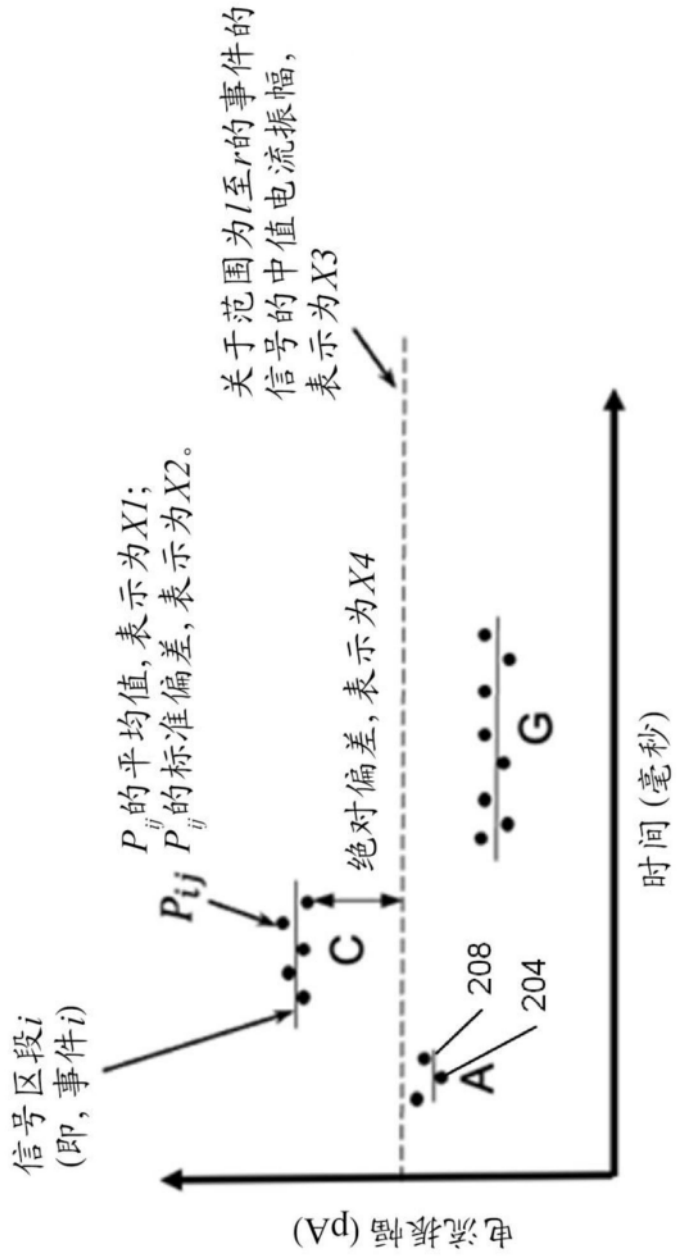

信号区段i
(即,事件i)
P_{ij}
P_{ij} 的平均值,表示为X1;
P_{ij} 的标准偏差,表示为X2。
关于范围为1至r的事件的
信号的中值电流振幅,
表示为X3
绝对偏差,表示为X4
C
208
204
A
G
时间(毫秒)
电流振幅(pA)

图2

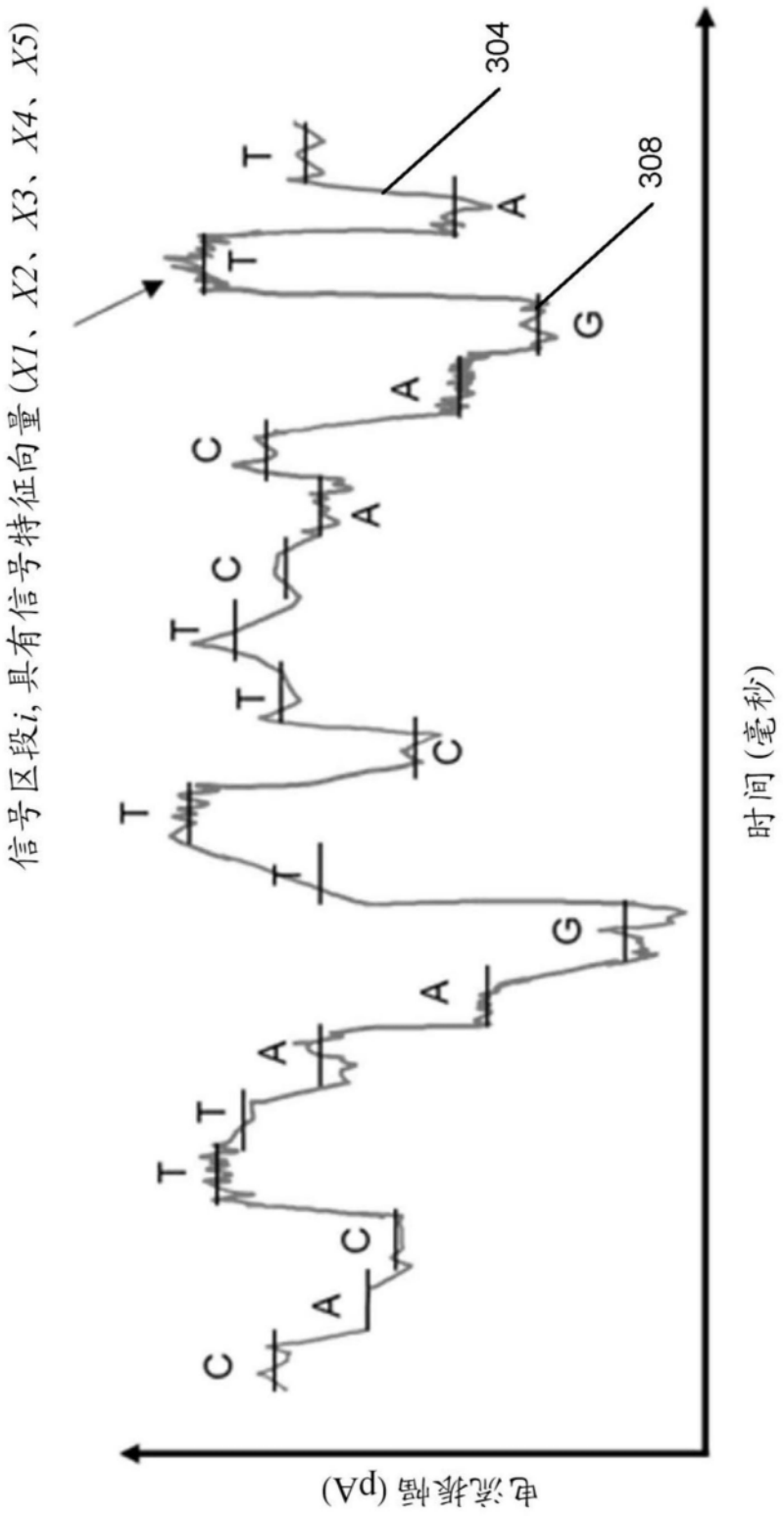

信号区段i，具有信号特征向量(X1、X2、X3、X4、X5)
304
308
C
A
C
T
T
A
T
T
C
A
T
C
T
C
A
A
G
T
T
C
A
T
T
A
G
电流振幅 (pA)
时间 (毫秒)

图3

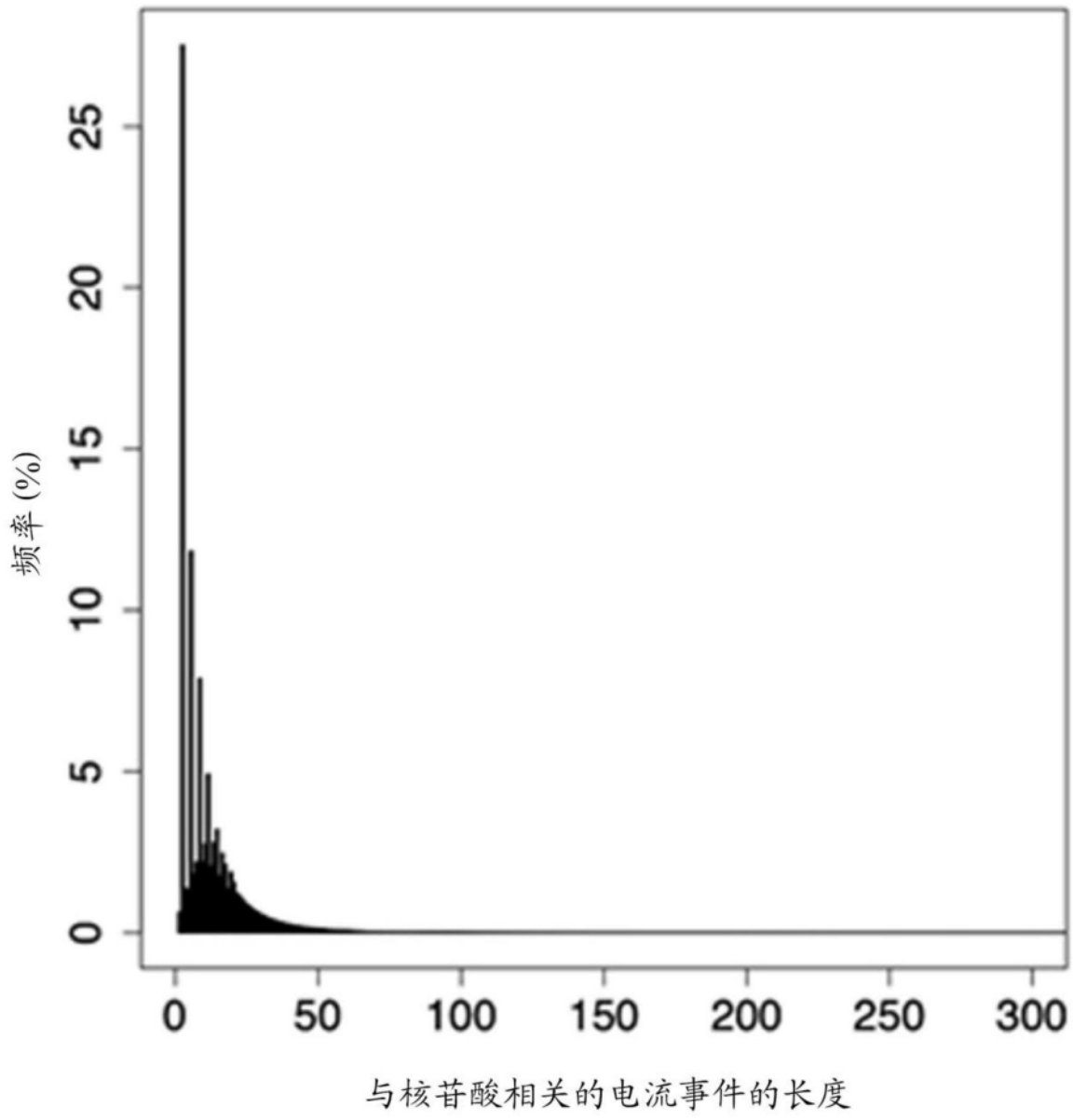


图4

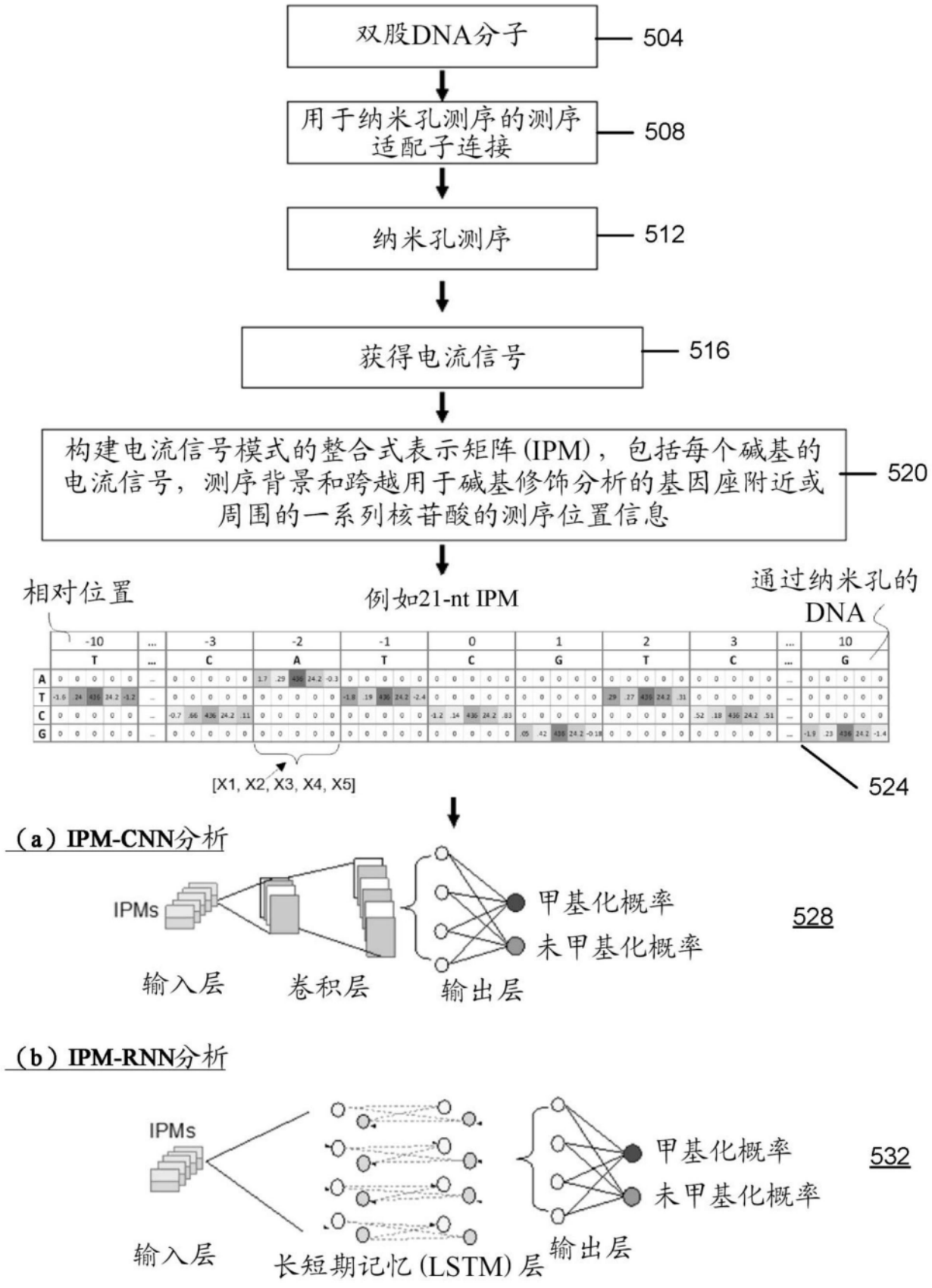

双股DNA分子
504
用于纳米孔测序的测序适配子连接
508
纳米孔测序
512
获得电流信号
516
构建电流信号模式的整合式表示矩阵(IPM)，包括每个碱基的电流信号，测序背景和跨越用于碱基修饰分析的基因座附近或周围的一系列核苷酸的测序位置信息
520
相对位置
例如21-nt IPM
通过纳米孔的DNA
[X1, X2, X3, X4, X5]
524
(a) IPM-CNN分析
IPMs
输入层
卷积层
输出层
甲基化概率
未甲基化概率
528
(b) IPM-RNN分析
IPMs
输入层
长短期记忆(LSTM)层
输出层
甲基化概率
未甲基化概率
532

图5

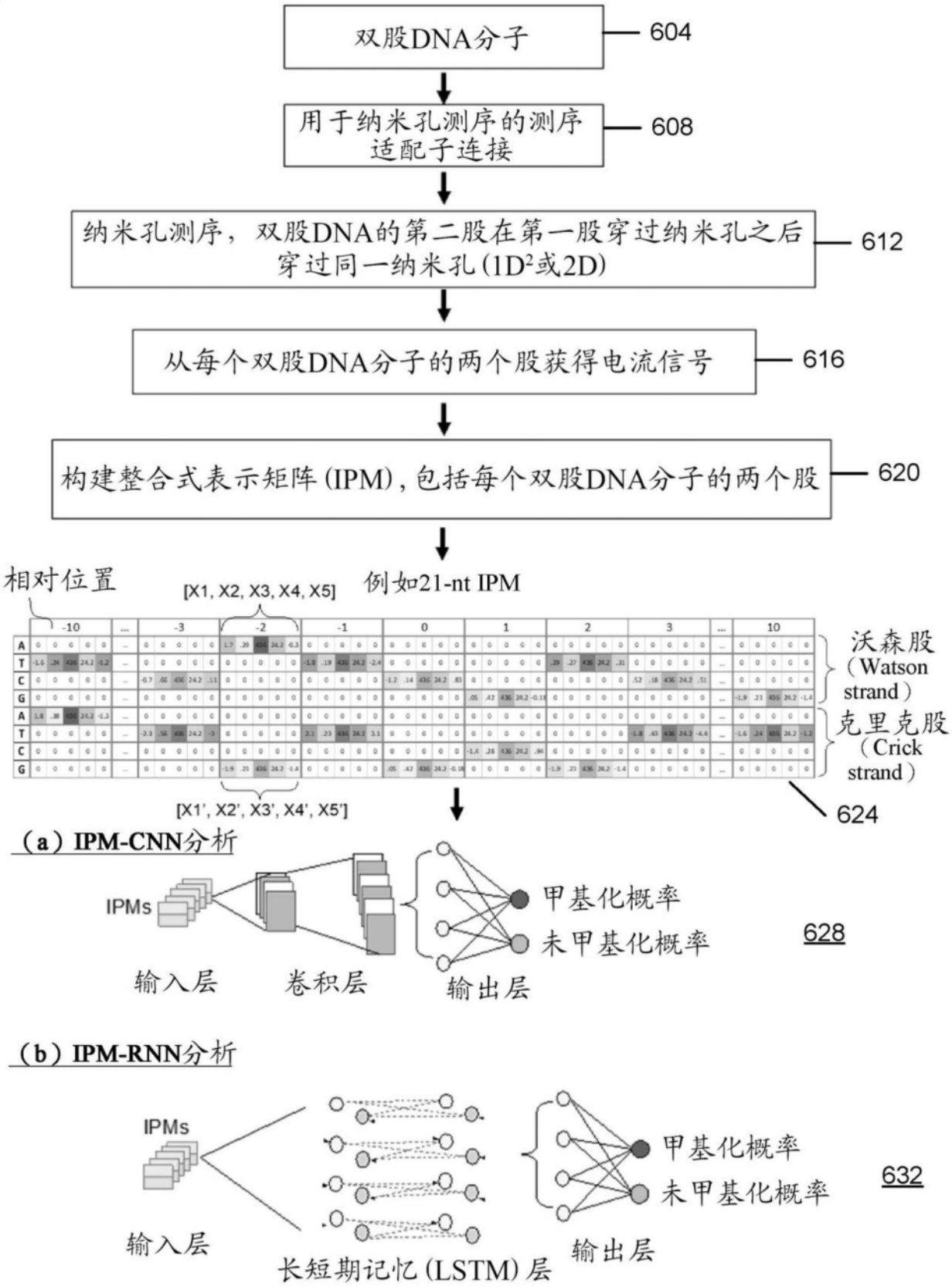

双股DNA分子
604
用于纳米孔测序的测序适配子连接
608
纳米孔测序，双股DNA的第二股在第一股穿过纳米孔之后穿过同一纳米孔(1D²或2D)
612
从每个双股DNA分子的两个股获得电流信号
616
构建整合式表示矩阵(IPM)，包括每个双股DNA分子的两个股
620
相对位置
[X1, X2, X3, X4, X5]
例如21-nt IPM
沃森股（Watson strand）
克里克股（Crick strand）
[X1', X2', X3', X4', X5']
624
（a）IPM-CNN分析
IPMs
输入层
卷积层
输出层
甲基化概率
未甲基化概率
628
（b）IPM-RNN分析
IPMs
输入层
长短期记忆(LSTM)层
输出层
甲基化概率
未甲基化概率
632

图6

核尺寸	AUC	
	训练数据集	测试数据集
1x5	0.98	0.96
1x10	0.98	0.96
1x15	0.97	0.97
1x20	0.98	0.96
1x25	0.98	0.96
1x30	0.97	0.94

图7

数据集	分子数目(CpG位点数目)	
	训练	测试
M.SssI处理过的DNA	7,989 (38,470)	4,826 (9,716)
WGA DNA	8,052 (37,150)	5,041 (11,444)

图8

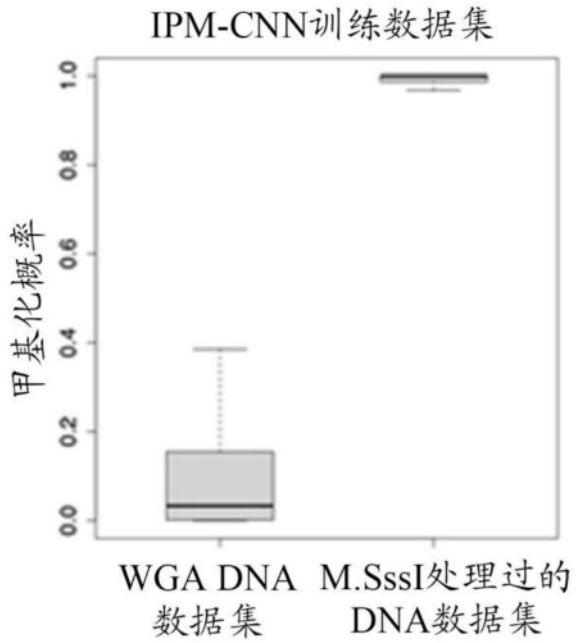

IPM-CNN训练数据集
甲基化概率
0.0
0.2
0.4
0.6
0.8
1.0
WGA DNA 数据集
M.SssI处理过的DNA数据集

图9A

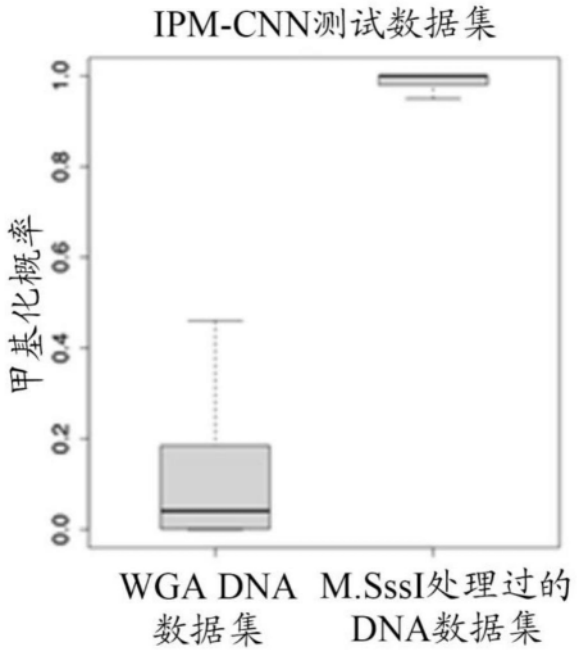

IPM-CNN测试数据集
甲基化概率
0.0
0.2
0.4
0.6
0.8
1.0
WGA DNA 数据集
M.SssI处理过的DNA数据集

图9B

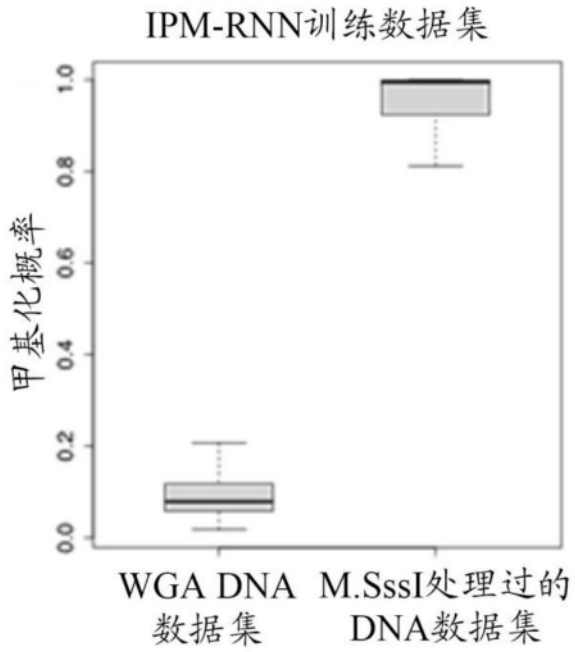

IPM-RNN训练数据集
甲基化概率
0.0
0.2
0.4
0.6
0.8
1.0
WGA DNA 数据集
M.SssI处理过的DNA数据集

图9C

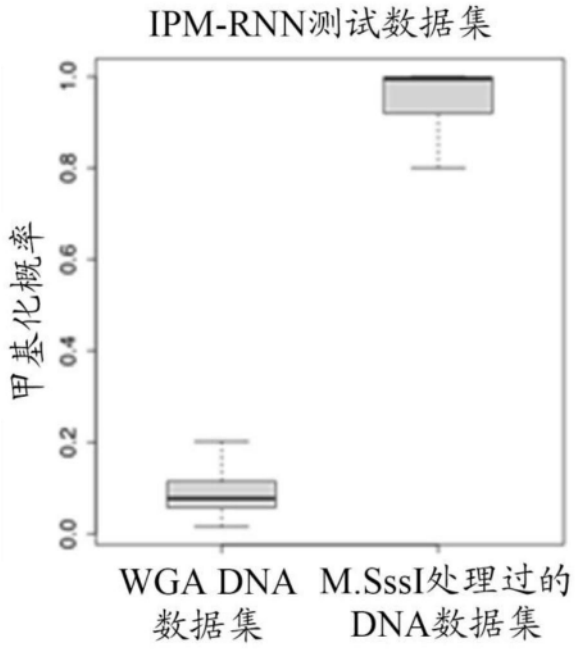

IPM-RNN测试数据集
甲基化概率
0.0
0.2
0.4
0.6
0.8
1.0
WGA DNA 数据集
M.SssI处理过的DNA数据集

图9D

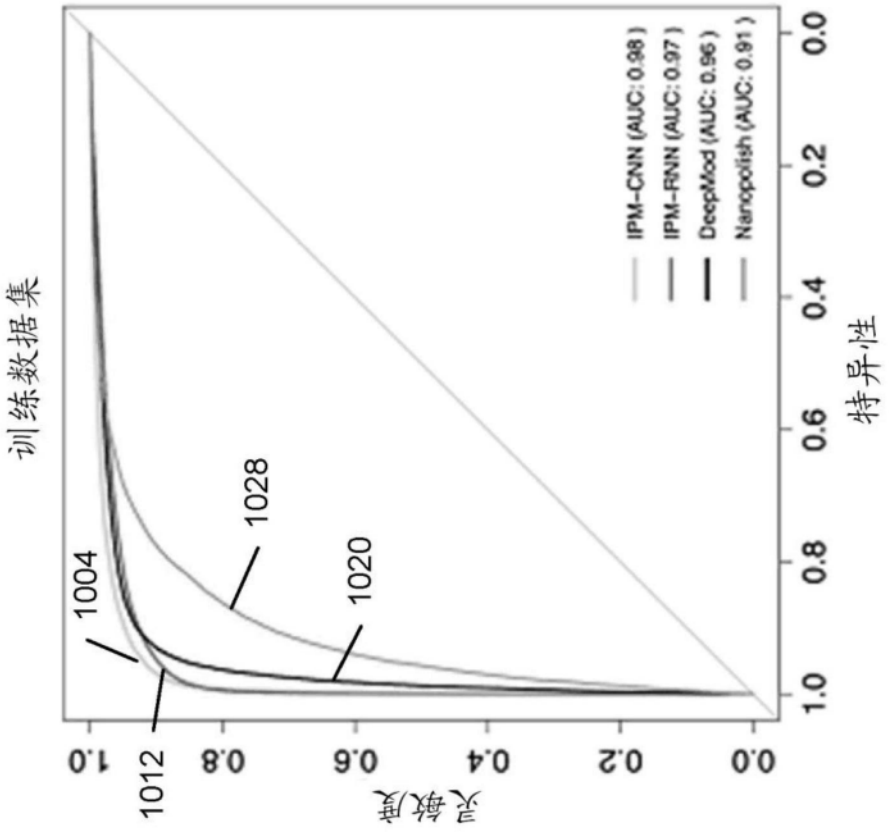

训练数据集
灵敏度
特异性
1.0
0.8
0.6
0.4
0.2
0.0
1004
1012
1028
1020
IPM-CNN (AUC: 0.98)
IPM-RNN (AUC: 0.97)
DeepMod (AUC: 0.96)
Nanopolish (AUC: 0.91)

图10A

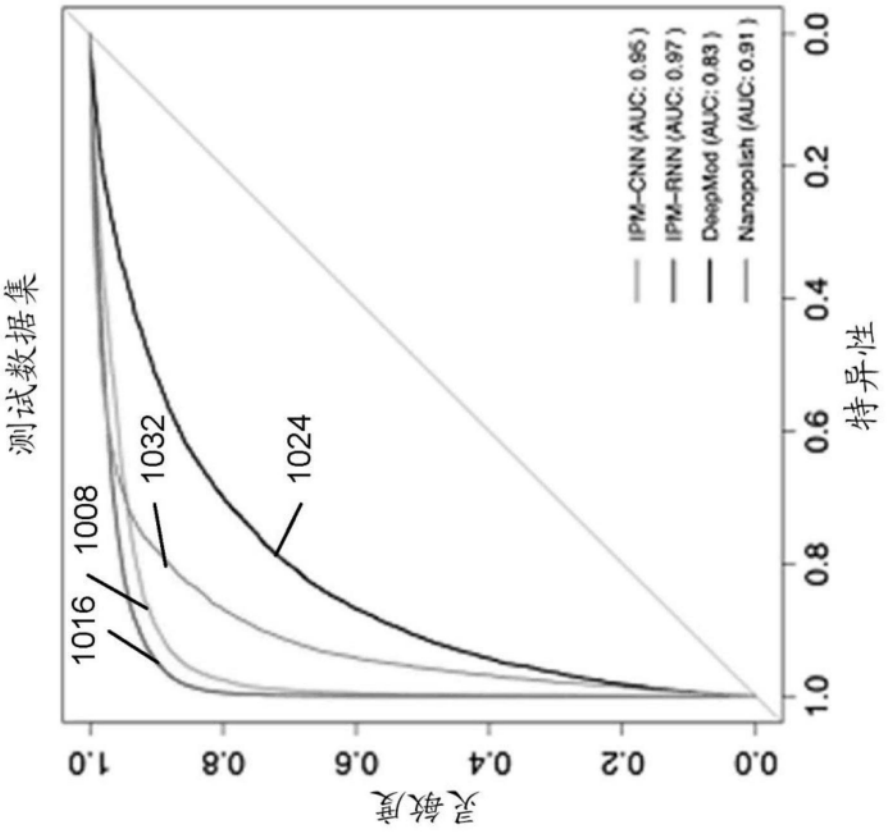


图10B

	灵敏度	特异性
IPM-CNN	**90%**	90%
IPM-RNN	**93%**	
DeepMod	53%	
nanopolish	74%	
IPM-CNN	**86%**	95%
IPM-RNN	**90%**	
DeepMod	38%	
nanopolish	55%	
IPM-CNN	**70%**	99%
IPM-RNN	**83%**	
DeepMod	13%	
nanopolish	16%	

图11

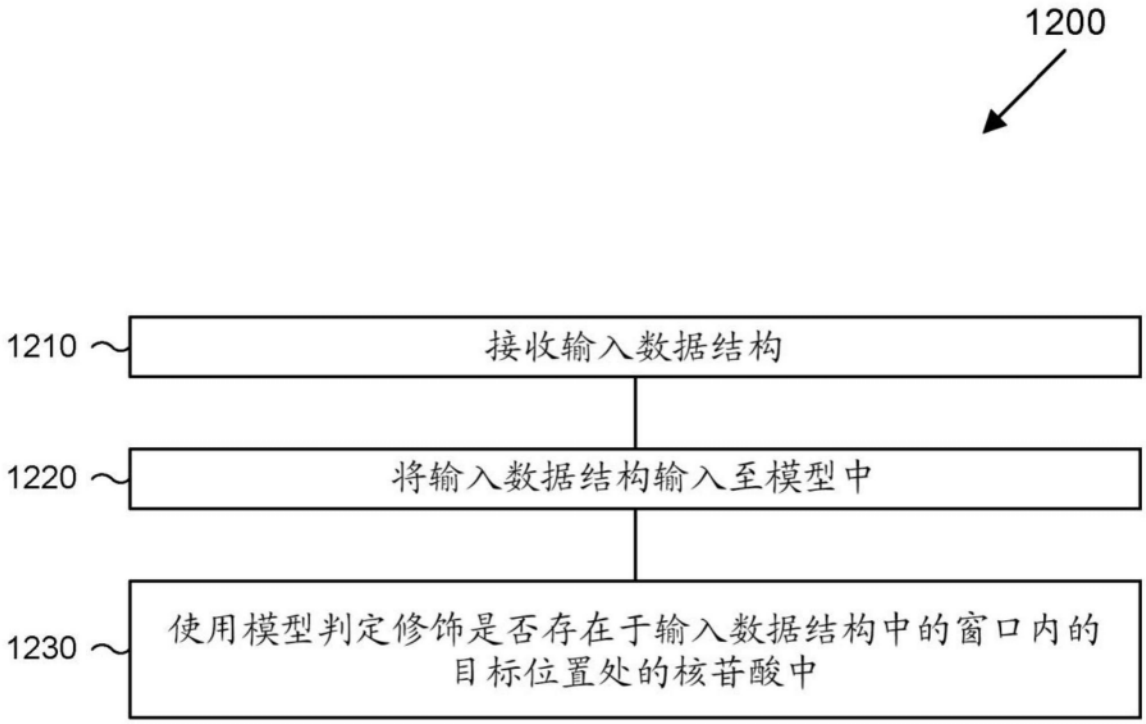

1200
1210
接收输入数据结构
1220
将输入数据结构输入至模型中
1230
使用模型判定修饰是否存在于输入数据结构中的窗口内的目标位置处的核苷酸中

图12

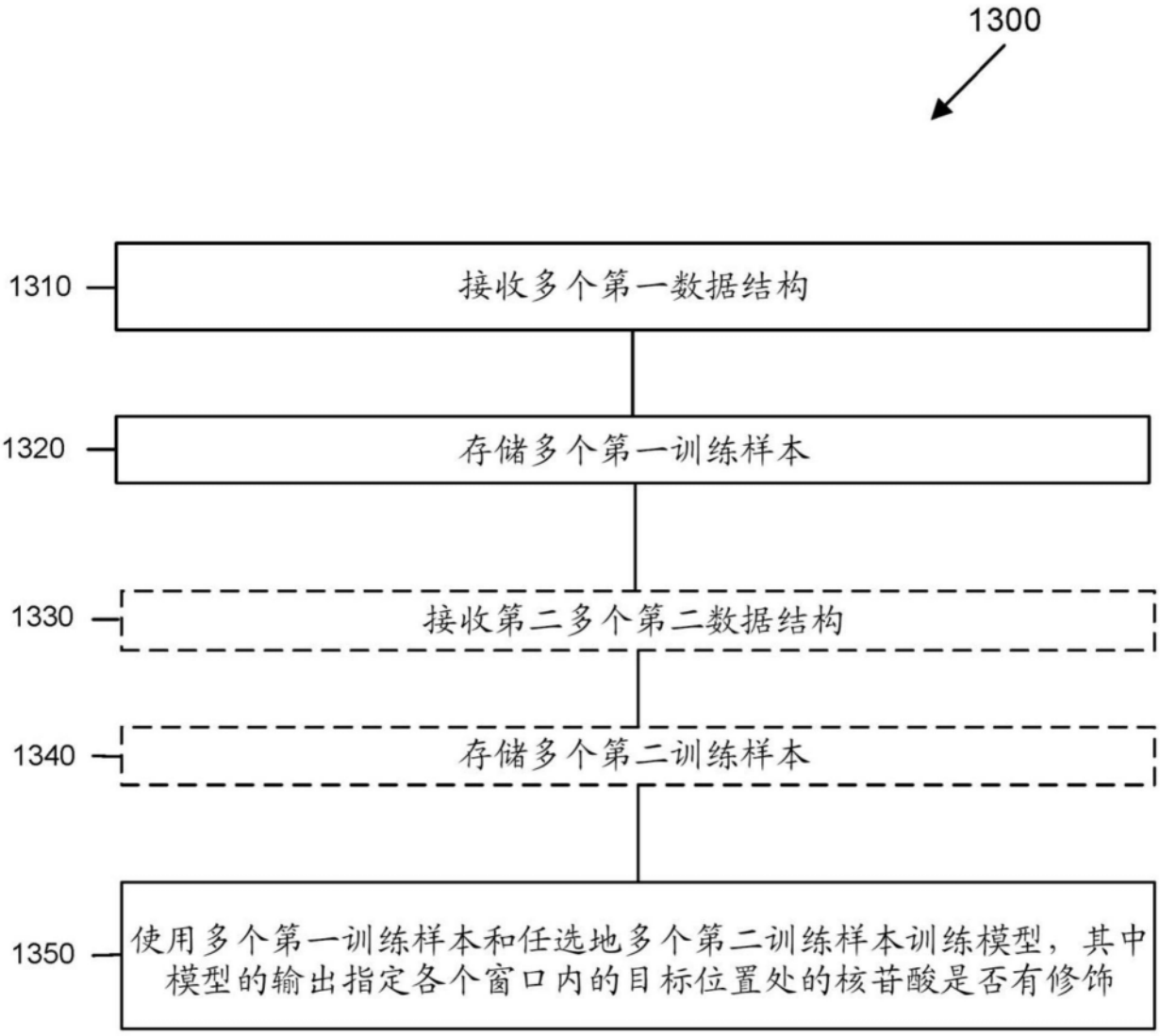

1300
1310
接收多个第一数据结构
1320
存储多个第一训练样本
1330
接收第二多个第二数据结构
1340
存储多个第二训练样本
1350
使用多个第一训练样本和任选地多个第二训练样本训练模型，其中模型的输出指定各个窗口内的目标位置处的核苷酸是否有修饰

图13

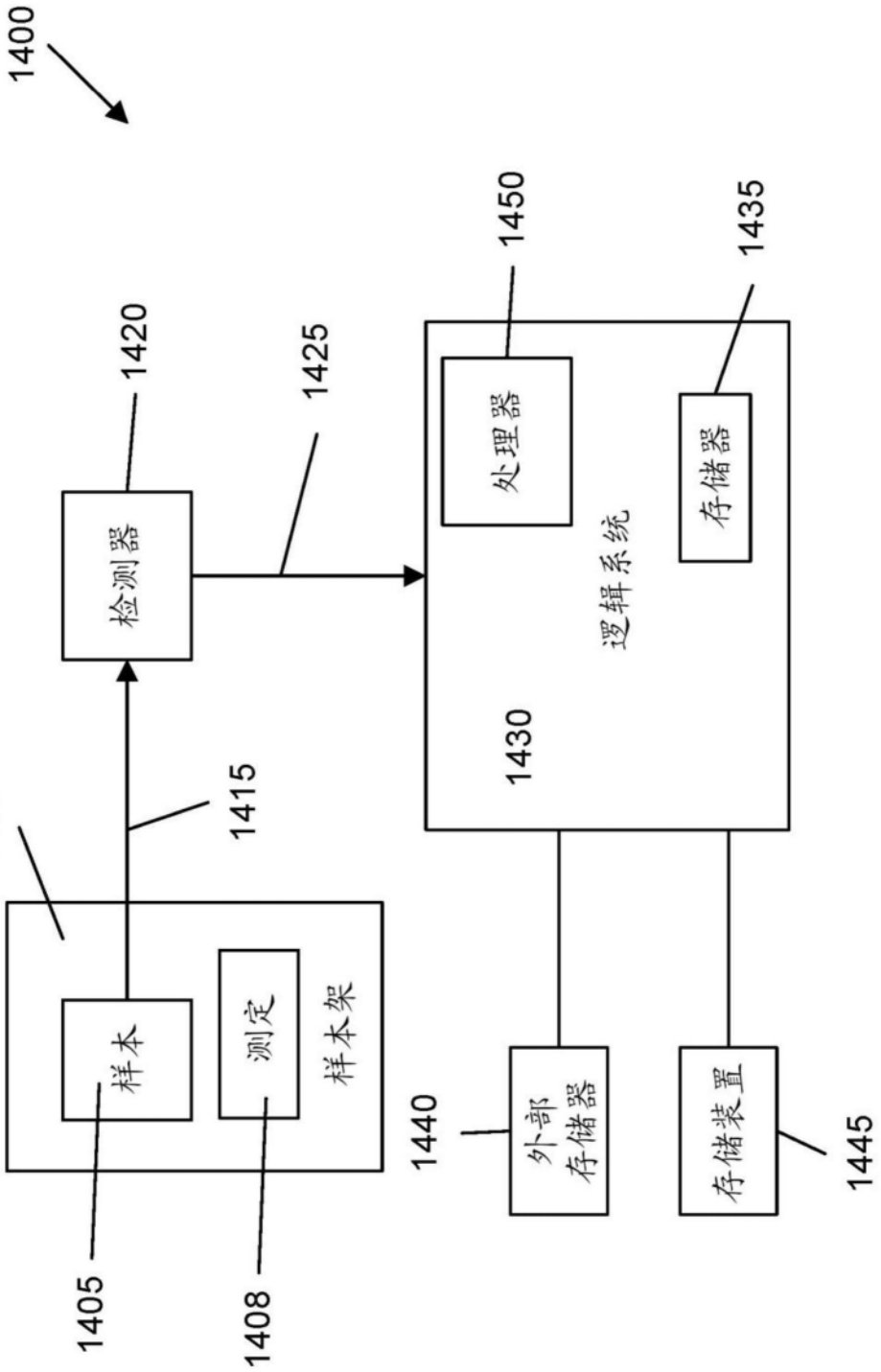

1400
1450
1435
1420
1425
处理器
存储器
检测器
逻辑系统
1430
1415
1410
样本
测定
样本架
1440
外部存储器
存储装置
1445
1405
1408

图14

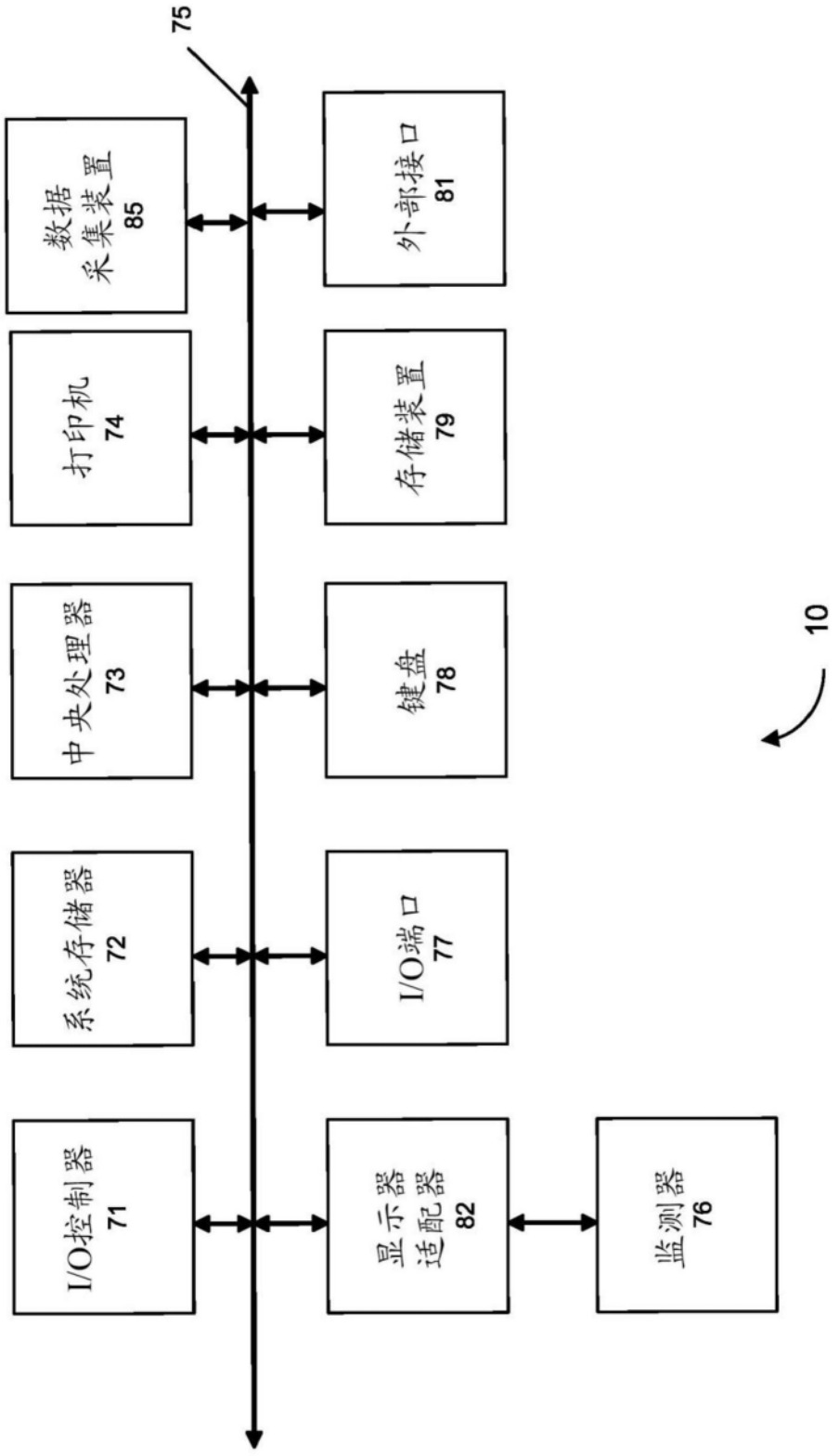

75
数据采集装置 85
外部接口 81
打印机 74
存储装置 79
中央处理器 73
键盘 78
系统存储器 72
I/O端口 77
I/O控制器 71
显示器适配器 82
监测器 76
10

图15

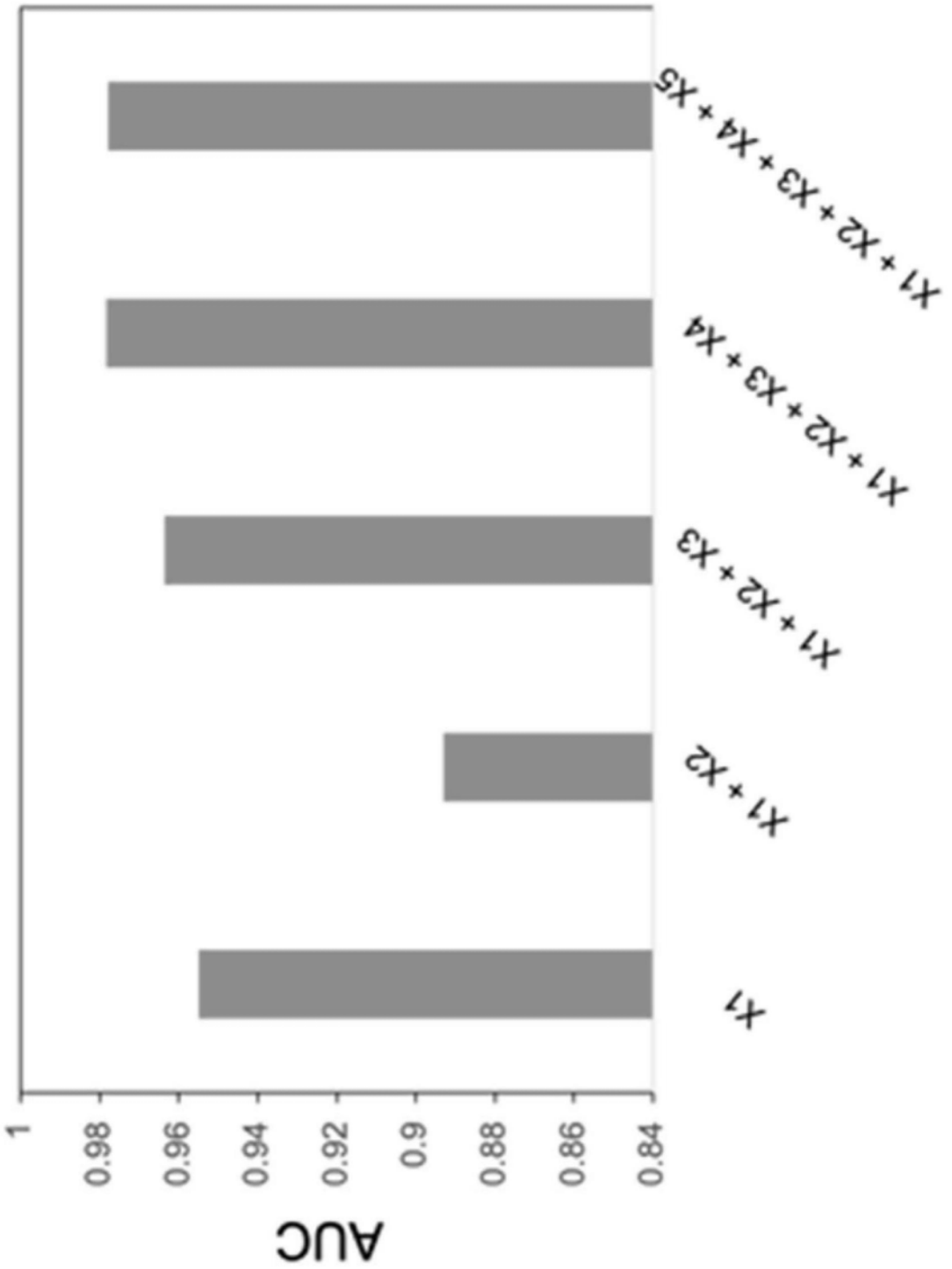

AUC
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
X7
X7×X2
X7×X2×X3
X7×X2×X3×X4
X7×X2×X3×X4×X5

图16

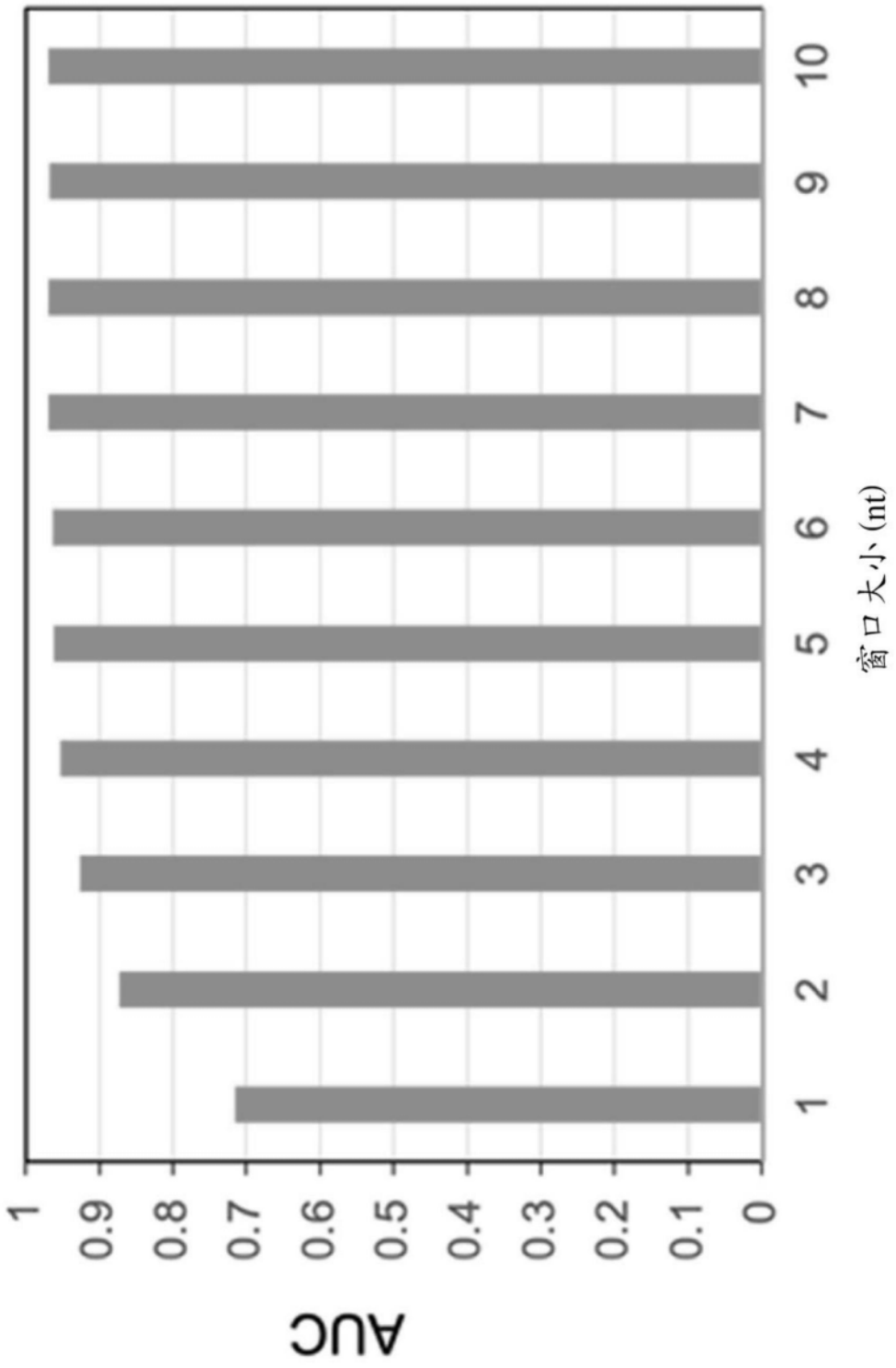


图17

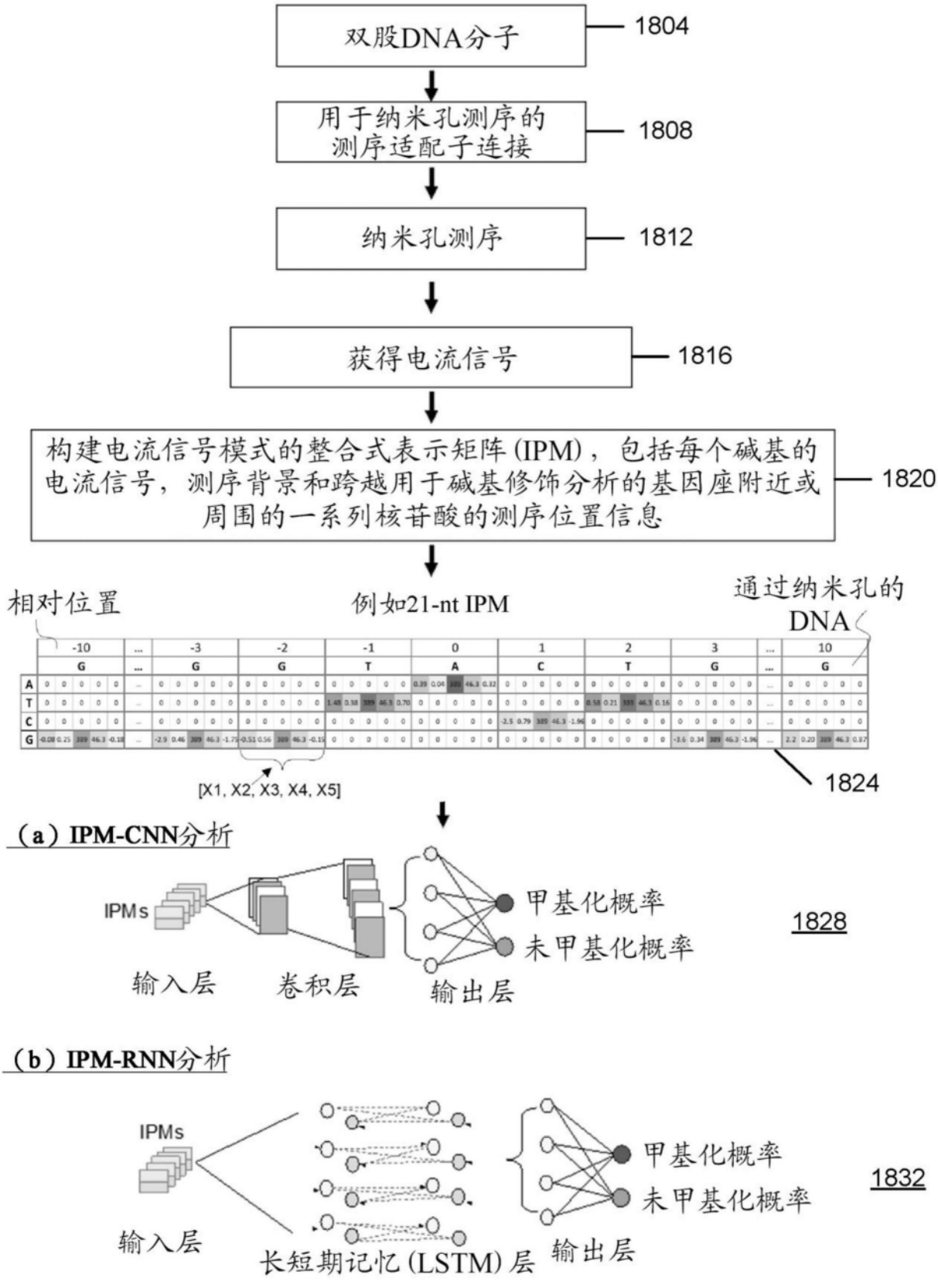

双股DNA分子
1804
用于纳米孔测序的测序适配子连接
1808
纳米孔测序
1812
获得电流信号
1816
构建电流信号模式的整合式表示矩阵(IPM)，包括每个碱基的电流信号，测序背景和跨越用于碱基修饰分析的基因座附近或周围的一系列核苷酸的测序位置信息
1820
相对位置
例如21-nt IPM
通过纳米孔的DNA
-10 ... -3 -2 -1 0 1 2 3 ... 10
G ... G G T A C T G ... G
A
T
C
G
[X1, X2, X3, X4, X5]
1824
(a) IPM-CNN分析
IPMs
输入层
卷积层
输出层
甲基化概率
未甲基化概率
1828
(b) IPM-RNN分析
IPMs
输入层
长短期记忆(LSTM)层
输出层
甲基化概率
未甲基化概率
1832

图18

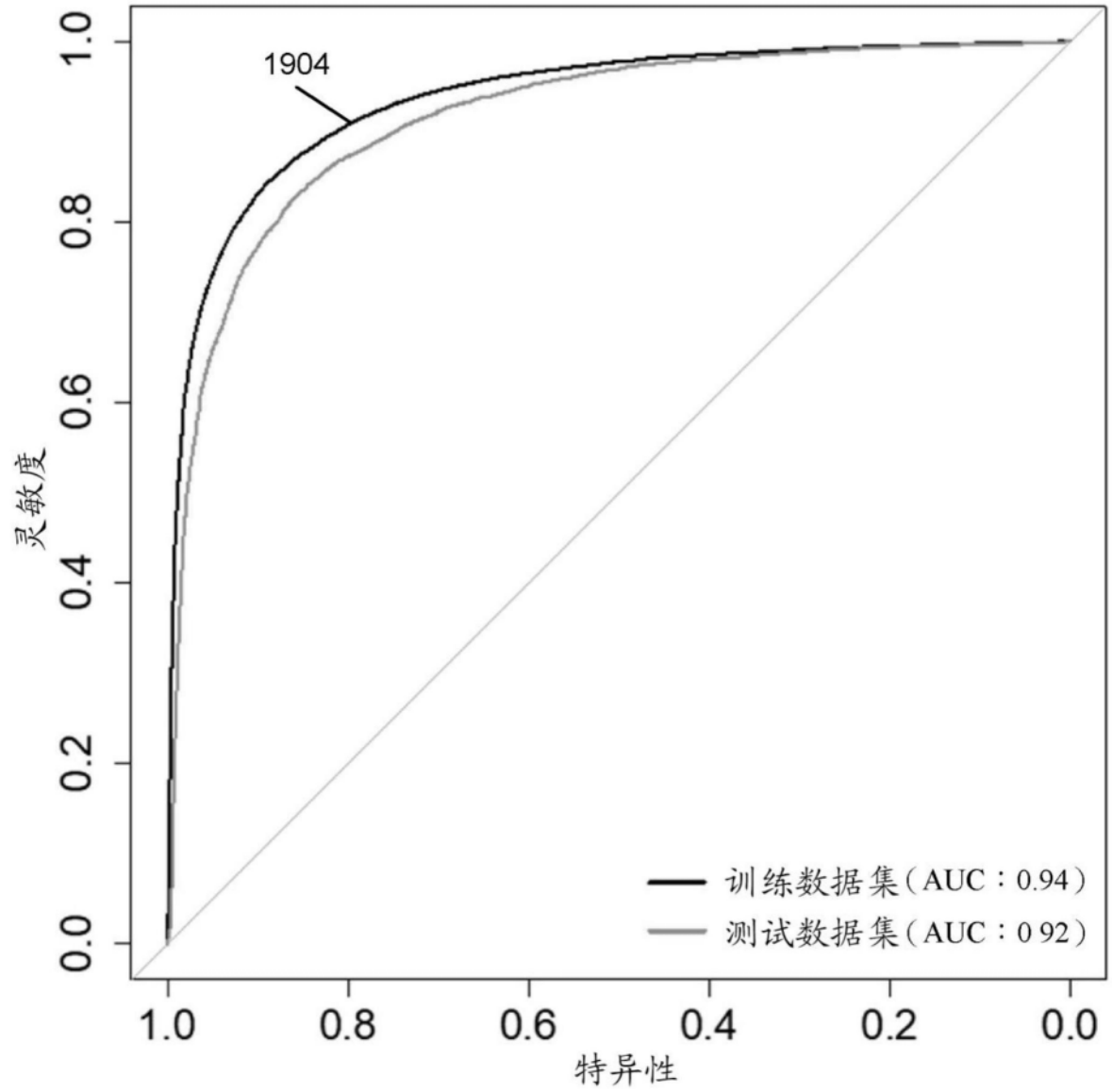


图19

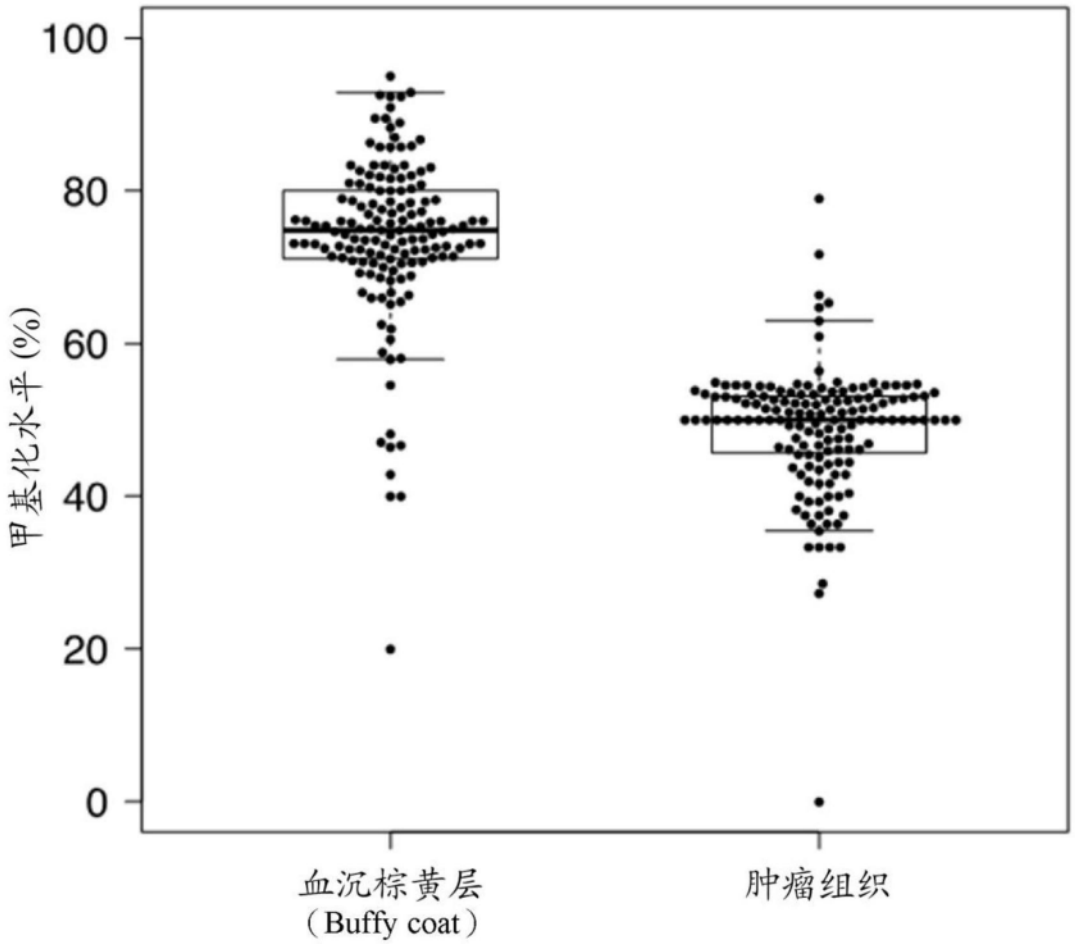


图20

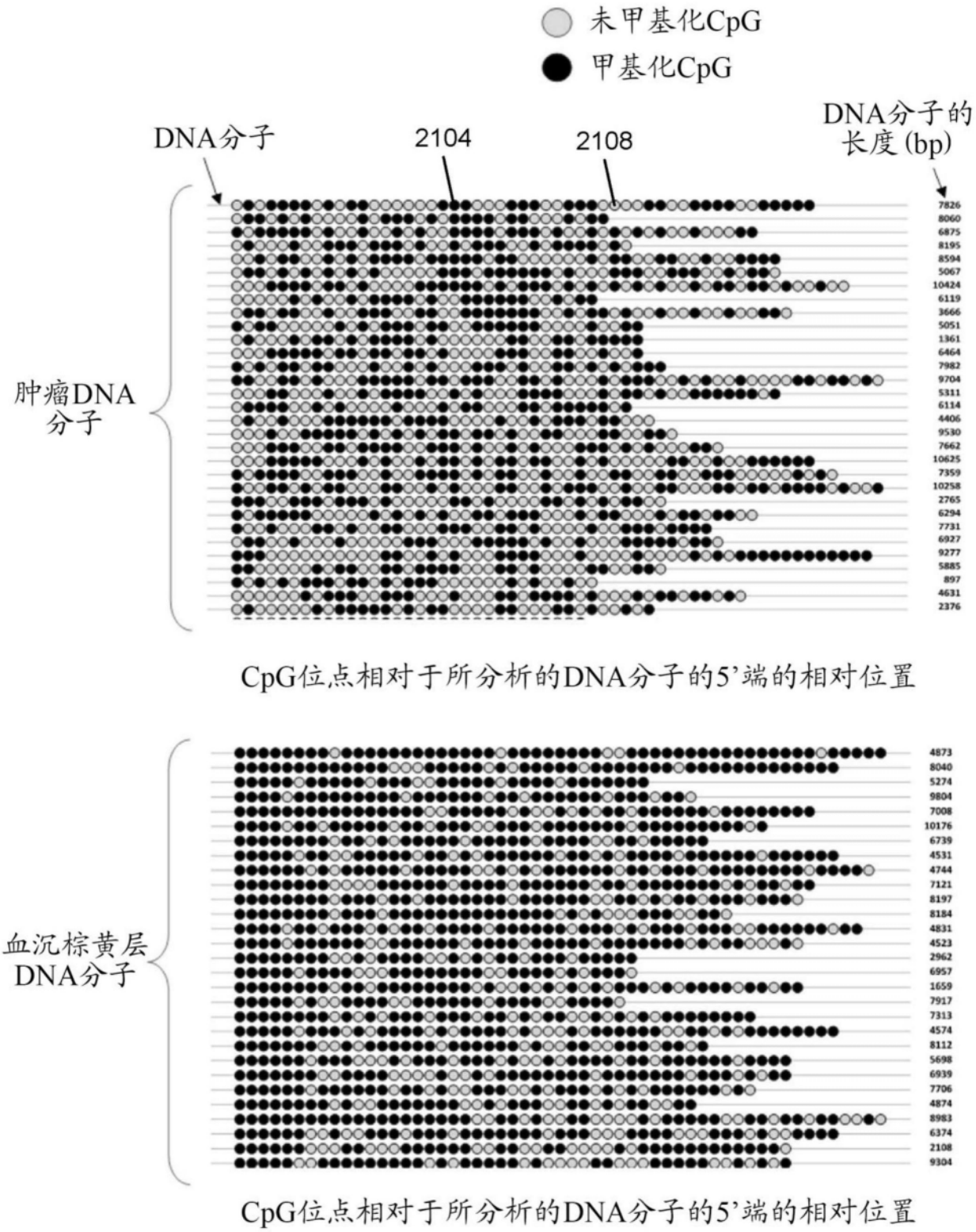

未甲基化CpG
甲基化CpG
DNA分子
2104
2108
DNA分子的长度(bp)
肿瘤DNA分子
7826
8060
6875
8195
8594
5067
10424
6119
3666
5051
1361
6464
7982
9704
5311
6114
4406
9530
7662
10625
7359
10258
2765
6294
7731
6927
9277
5885
897
4631
2376
CpG位点相对于所分析的DNA分子的5'端的相对位置
血沉棕黄层DNA分子
4873
8040
5274
9804
7008
10176
6739
4531
4744
7121
8197
8184
4831
4523
2962
6957
1659
7917
7313
4574
8112
5698
6939
7706
4874
8983
6374
2108
9304
CpG位点相对于所分析的DNA分子的5'端的相对位置

图21

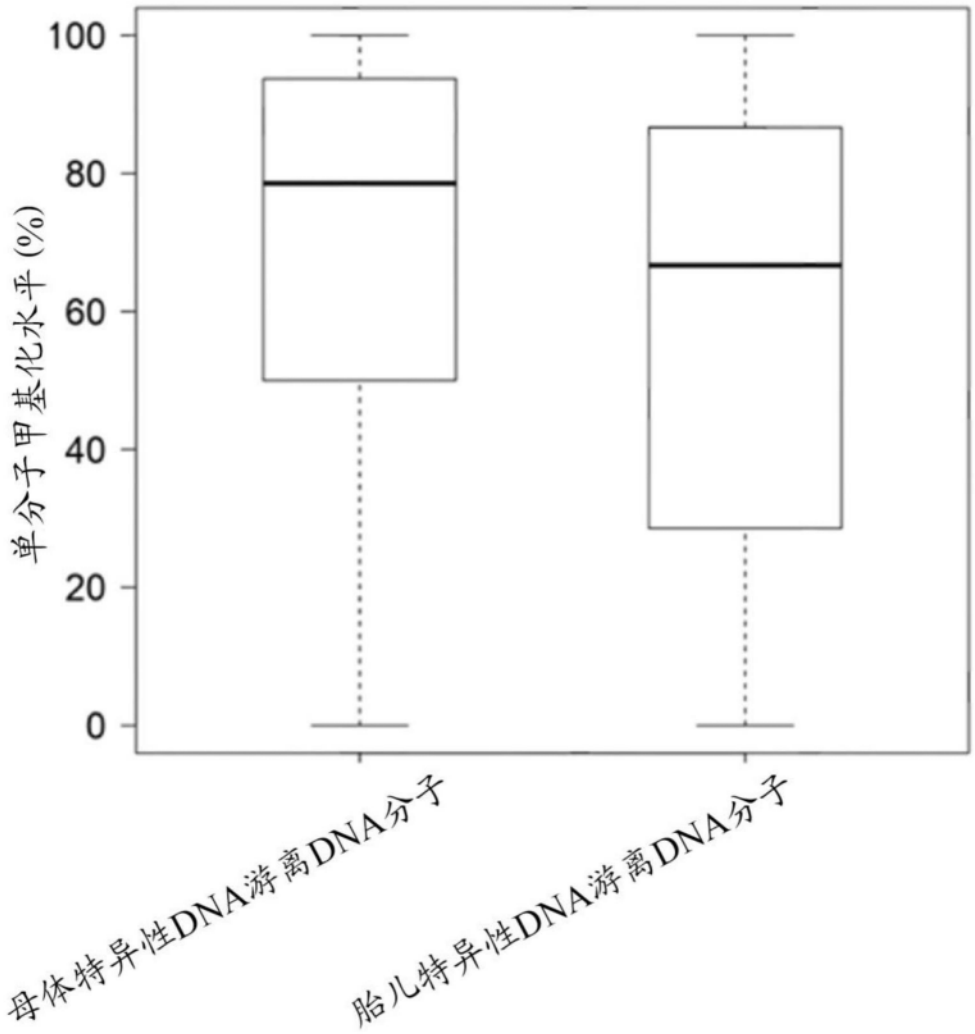

100
80
60
40
20
0
单分子甲基化水平 (%)
母体特异性DNA游离DNA分子
胎儿特异性DNA游离DNA分子

图22

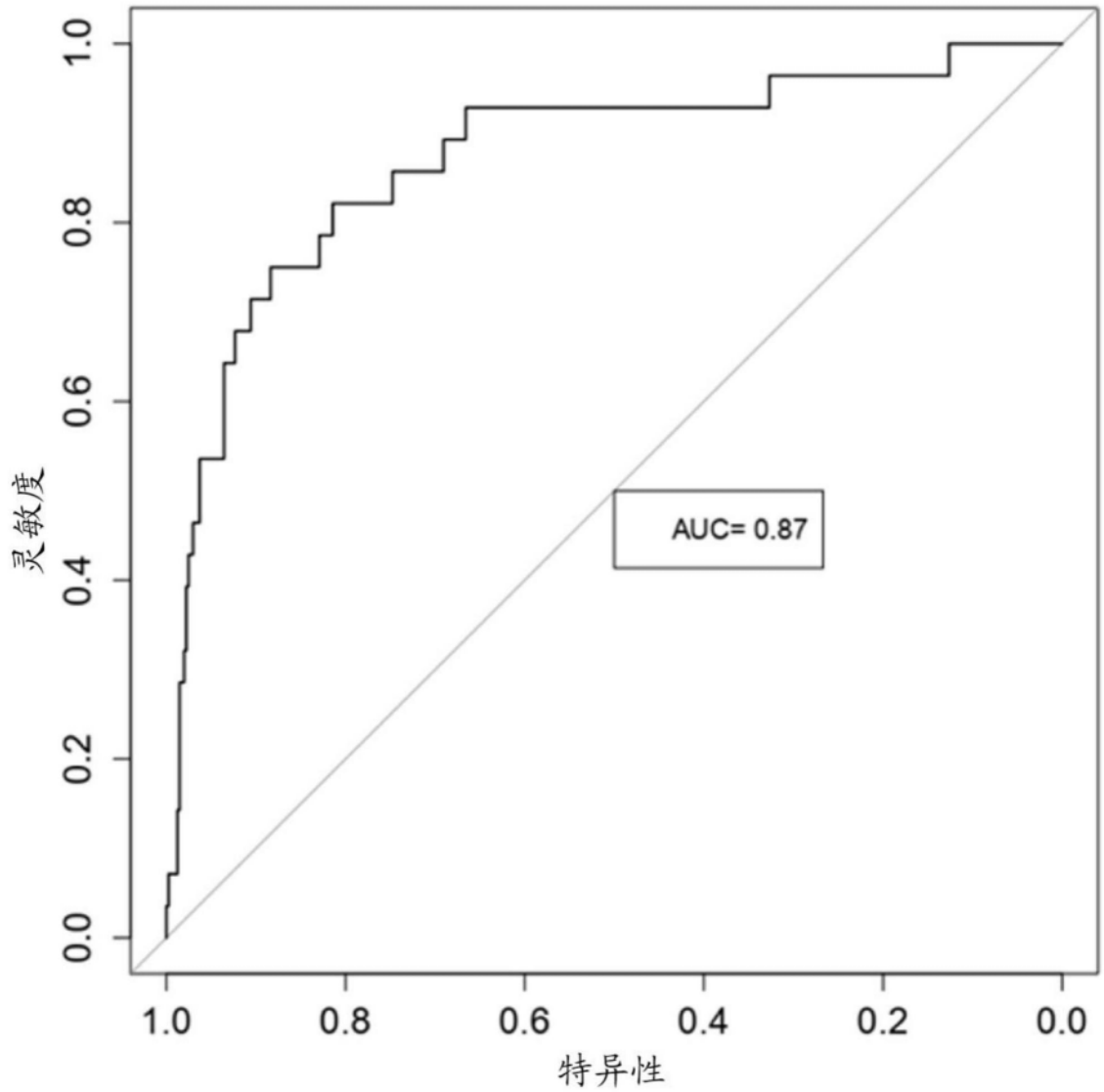


图23