(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110325650 A
(43)申请公布日 2019.10.11

(21)申请号 201780087130.5

(22)申请日 2017.12.22

(30)优先权数据
62/438,240 2016.12.22 US
62/512,936 2017.05.31 US
62/550,540 2017.08.25 US

(85)PCT国际申请进入国家阶段日
2019.08.21

(86)PCT国际申请的申请数据
PCT/US2017/068329 2017.12.22

(87)PCT国际申请的公布数据
WO2018/119452 EN 2018.06.28

(71)申请人 夸登特健康公司
地址 美国加利福尼亚州

(72)发明人 安德鲁·肯尼迪
斯特凡尼·安·沃德·莫蒂默
埃尔米·埃尔图凯
阿米尔阿里·塔拉萨兹
戴安娜·阿布杜伊瓦
马修·舒尔茨

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 刘晓杰

(51)Int.Cl.
C12Q 1/6806(2006.01)
C12Q 1/6827(2006.01)

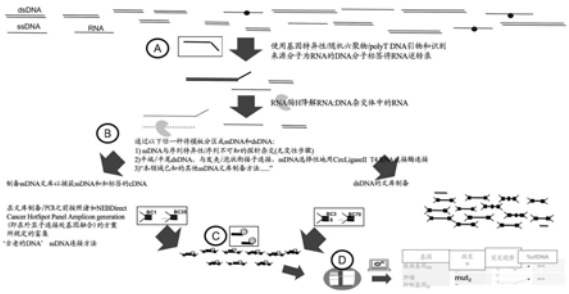
权利要求书3页 说明书71页 附图39页

(54)发明名称

用于分析核酸分子的方法和系统

(57)摘要

本公开内容提供了用于处理包含不同形式(例如,RNA和DNA,单链或双链)和/或修饰程度(例如,胞嘧啶甲基化,与蛋白质缔合)的核酸群体的方法。这些方法适应样品中核酸的多种形式和/或修饰,使得可以获得针对多种形式的序列信息。所述方法经过处理和分析仍保持多种形式或修饰状态的身份,使得序列分析可以与表观遗传分析组合。



CN 110325650 A

1.一种分析核酸群体的方法,所述核酸群体包含选自双链DNA、单链DNA和单链RNA的至少两种形式的核酸,其中所述至少两种形式中的每一种包含多于一个分子,所述方法包括:

(a)将所述形式的核酸中的至少一种与至少一种标签核酸连接,以将所述形式彼此区分开,

(b)扩增所述形式的核酸,其中的至少一种被连接至至少一种核酸标签,其中核酸和连接的核酸标签被扩增,以产生扩增的核酸,其中从所述至少一种形式扩增的核酸被加标签;

(c)测定所述扩增的核酸的序列数据,所述扩增的核酸中的至少一些被加标签;其中所述测定获得足以解码所述扩增的核酸的所述标签核酸分子的序列信息,以揭示所述群体中为已测定了针对其的序列数据的连接至所述标签核酸分子的所述扩增的核酸提供原始模板的核酸的形式。

2.如权利要求1所述的方法,还包括以下步骤:解码所述扩增的核酸的所述标签核酸分子,以揭示为已测定了针对其的序列数据的连接至所述标签核酸分子的所述扩增的核酸提供原始模板的核酸的形式。

3.如权利要求1或2所述的方法,还包括相对于一种或更多种其它形式富集所述形式中的至少一种。

4.如权利要求1或2所述的方法,其中所述群体中的每种形式的核酸的至少70%的分子在步骤(b)中被扩增。

5.如权利要求1或2所述的方法,其中所述群体中存在至少三种形式的核酸,并且所述形式中的至少两种连接至不同的标签核酸形式,所述不同的标签核酸形式将所述三种形式中的每一种彼此区分开。

6.如权利要求5所述的方法,其中所述群体中的所述至少三种形式的核酸中的每一种连接至不同的标签。

7.如权利要求1或2所述的方法,其中相同形式的每个分子连接至包含相同识别标签的标签。

8.如权利要求1或2所述的方法,其中相同形式的分子连接至不同类型的标签。

9.如权利要求1或2所述的方法,其中步骤(a)包括:使所述群体经历用加标签的引物的逆转录,其中所述加标签的引物被掺入从所述群体中的RNA产生的cDNA中。

10.如权利要求9所述的方法,其中所述逆转录是序列特异性的。

11.如权利要求9所述的方法,其中所述逆转录是随机的。

12.如权利要求9所述的方法,还包括降解与所述cDNA成双链的RNA。

13.如权利要求5所述的方法,还包括分离单链DNA和双链DNA,和将核酸标签连接至所述双链DNA。

14.如权利要求13所述的方法,其中所述单链DNA通过与一种或更多种捕获探针杂交被分离。

15.如权利要求5所述的方法,还包括用环连接酶使单链DNA环化,和将核酸标签连接至所述双链DNA。

16.如权利要求1所述的方法,包括在测定之前,合并包含不同形式的核酸的加标签的核酸。

17.如权利要求1-16所述的方法,其中所述核酸群体来自体液样品。

18.如权利要求17所述的方法,其中所述体液样品是血液、血清或血浆。

19.如权利要求1或2所述的方法,其中所述核酸群体是无细胞核酸群体。

20.如权利要求18所述的方法,其中所述体液样品来自疑似具有癌症的受试者。

21.如权利要求1-20所述的方法,其中所述序列数据指示体细胞变体或生殖系变体的存在。

22.如权利要求1-21所述的方法,其中所述序列数据指示拷贝数变异的存在。

23.如权利要求1-22所述的方法,其中所述序列数据指示单核苷酸变异(SNV)、插入缺失或基因融合的存在。

24.一种分析包含具有不同修饰程度的核酸的核酸群体的方法,包括:

使所述核酸群体与优先结合带有所述修饰的核酸的剂接触,

分离与所述剂结合的第一核酸池和未与所述剂结合的第二核酸池,其中所述第一核酸池对于所述修饰是呈现过度的,并且所述第二池中的核酸对于所述修饰是呈现不足的;

将所述第一池和/或所述第二池中的核酸连接至区分所述第一池和所述第二池中的核酸的一个或更多个核酸标签,以产生加标签的核酸的群体;

扩增标记的核酸,其中所述核酸和连接的标签被扩增;

测定所述扩增的核酸和连接的标签的序列数据;其中所述测定获得用于解码所述标签的序列数据,以揭示已测定了针对其的序列数据的核酸是从所述第一池中的模板扩增的还是从所述第二池中的模板扩增的。

25.如权利要求24所述的方法,还包括以下步骤:解码所述标签以揭示已测定了针对其的序列数据的核酸是从所述第一池中的模板扩增的还是从所述第二池中的模板扩增的。

26.如权利要求25或26所述的方法,其中所述修饰是将核酸与蛋白质结合。

27.如权利要求25或26所述的方法,其中所述蛋白质是组蛋白或转录因子。

28.如权利要求25或26所述的方法,其中所述修饰是对核苷酸的复制后修饰。

29.如权利要求27所述的方法,其中所述复制后修饰是5-甲基-胞嘧啶,并且所述剂与核酸的结合程度随着所述核酸中5-甲基-胞嘧啶的程度而增加。

30.如权利要求27所述的方法,其中所述复制后修饰是5-羟甲基-胞嘧啶,并且所述剂与核酸的结合程度随着所述核酸中5-羟甲基-胞嘧啶的程度而增加。

31.如权利要求27所述的方法,其中所述复制后修饰是5-甲酰基-胞嘧啶或5-羧基-胞嘧啶,并且所述剂的结合程度随着所述核酸中5-甲酰基-胞嘧啶或5-羧基-胞嘧啶的程度而增加。

32.如权利要求25或26所述的方法,还包括洗涤与所述剂结合的核酸,和将洗涤物收集为第三池,所述第三池包括相对于所述第一池和所述第二池具有中等程度的复制后修饰的核酸。

33.如权利要求25或26所述的方法,包括在测定之前,合并来自所述第一池和所述第二池的加标签的核酸。

34.如权利要求25或26所述的方法,其中所述剂是5-甲基-结合结构域磁珠。

35.如权利要求24-34所述的方法,其中所述核酸群体来自体液样品。

36.如权利要求35所述的方法,其中所述体液样品是血液、血清或血浆。

37.如权利要求25或26所述的方法,其中所述核酸群体是无细胞核酸群体。

38.如权利要求35所述的方法,其中所述体液样品来自疑似具有癌症的受试者。

39.如权利要求25-38所述的方法,其中所述序列数据指示体细胞变体或生殖系变体的存在。

40.如权利要求25-39所述的方法,其中所述序列数据指示拷贝数变异的存在。

41.如25-39中任一项所述的方法,其中所述序列数据指示单核苷酸变异(SNV)、插入缺失或基因融合的存在。

iv)核酸,其中所述核酸和连接的标签被扩增;并且用NGS仪器测定加分子标签的分区;

v)软件模块,用于生成用于解码所述标签的序列数据;和

vi)软件模块,用于分析所述序列数据以解码所述标签,以揭示已测定了针对其的序列数据的所述核酸是从所述第一池中的模板扩增的还是从所述第二池中的模板扩增的。

用于分析核酸分子的方法和系统

[0001] 相关专利申请的引用

[0002] 本申请要求2016年12月22日提交的美国临时专利申请62/438,240、2017年5月31日提交的美国临时专利申请62/512,936和2017年8月25日提交的美国临时专利申请62/550,540的优先权日的权益,所述美国临时专利申请全部通过引用以其整体并入本文。

[0003] 背景

[0004] 癌症为全世界疾病的主要原因。每年,世界各地有数千万人被诊断为患有癌症,并且多于一半的人最终因其死亡。在许多国家,癌症列为继心血管疾病之后第二大最常见的死亡原因。早期检测与许多癌症的改善结果相关。

[0005] 癌症可能是由个体的正常细胞内的遗传变异的累积引起的,其中至少一些遗传变异导致细胞分裂调节不当。这样的变异通常包括拷贝数变异(CNV)、单核苷酸变异(SNV)、基因融合、插入和/或缺失(插入缺失),表观遗传变异包括胞嘧啶的5-甲基化(5-甲基胞嘧啶)以及DNA与染色质和转录因子的缔合。

[0006] 癌症通常通过肿瘤活检,然后分析细胞、标志物或从细胞提取的DNA来检测。但是最近已经提出,癌症也可以从体液诸如血液或尿液中的无细胞核酸检测。这样的测试具有这样的优点,即它们是非侵入性的并且可以在活检中进行而不鉴定可疑的癌细胞。然而,由于体液中核酸的量非常低,并且所存在的核酸在形式(例如,RNA和DNA,单链和双链,以及复制后修饰和与蛋白质诸如组蛋白缔合的各种状态)上是异质的这一事实,这样的测试是复杂的。

[0007] 增加液体活检测定的灵敏度,同时减少过程中循环中的核酸(原始材料)或数据的损失是期望的。

[0008] 概述

[0009] 本公开内容提供了用于分析核酸群体的方法、组合物和系统,所述核酸群体包含选自双链DNA、单链DNA和单链RNA的至少两种形式的核酸。在一些实施方案中,所述方法包括(a)将这些形式的核酸中的至少一种与至少一种标签核酸连接,以将这些形式彼此区分开,(b)扩增这些形式的核酸,其中的至少一种被连接至至少一种核酸标签,其中核酸和连接的核酸标签(如果存在的话)被扩增,以产生扩增的核酸,其中从所述至少一种形式扩增的核酸被加标签;(c)测定扩增的核酸的序列数据,其中至少一些被加标签;和(d)解码扩增的核酸的标签核酸分子,以揭示群体中为已测定了针对其的序列数据的连接至标签核酸分子的扩增的核酸提供原始模板的核酸的形式。

[0010] 在一些实施方案中,所述方法还包括相对于一种或更多种其它形式富集这些形式中的至少一种。在一些实施方案中,群体中的每种形式的核酸的至少70%的分子在步骤(b)中被扩增。在一些实施方案中,群体中存在至少三种形式的核酸,并且所述形式中的至少两种连接至不同的标签核酸形式,所述不同的标签核酸形式将三种形式中的每一种彼此区分开。在一些实施方案中,群体中的至少三种形式的核酸中的每一种连接至不同的标签。在一些实施方案中,相同形式的每个分子连接至包含相同识别信息标签的标签(例如,具有相同序列或包含相同序列的标签)。在一些实施方案中,相同形式的分子连接至不同类型的标

签。在一些实施方案中，步骤(a)包括：使群体经历用加标签的引物的逆转录，其中加标签的引物被掺入从群体中的RNA产生的cDNA中。在一些实施方案中，逆转录是序列特异性的。在一些实施方案中，逆转录是随机的。在一些实施方案中，所述方法还包括降解与cDNA成双链的RNA。在一些实施方案中，所述方法还包括分离单链DNA和双链DNA，并且将核酸标签连接至双链DNA。在一些实施方案中，单链DNA通过与一种或更多种捕获探针杂交来分离。在一些实施方案中，所述方法还包括使用对单链核酸起作用的连接酶用单链标签对单链DNA差异化地加标签，以及使用对双链核酸起作用的连接酶用双链衔接子对双链DNA差异化地加标签。在一些实施方案中，所述方法还包括在测定之前，合并包含不同形式的核酸的加标签的核酸。在一些实施方案中，所述方法还包括在单独的测定中单独地分析分区的DNA的池。测定可以是相同的、基本上相似的、等同的或不同的。

[0011] 在任一种上文的方法中，序列数据可以指示体细胞或生殖系变体、或拷贝数变异或单核苷酸变异、或插入缺失或基因融合的存在。

[0012] 本公开内容还提供了一种分析包含具有不同修饰程度的核酸的核酸群体的方法。在一些情况下，本公开内容提供了用于筛选与疾病相关的特征(例如，5’甲基胞嘧啶)的方法。所述方法包括使核酸群体与优先结合带有修饰的核酸的剂(诸如甲基结合结构域或蛋白质)接触；分离与所述剂结合的第一核酸池和未与所述剂结合的第二核酸池，其中第一核酸池对于修饰是呈现过度的(overrepresented)，并且第二池中的核酸对于修饰是呈现不足的(underrepresented)；将第一池和/或第二池中的核酸连接至一个或更多个核酸标签，所述核酸标签区分第一池和第二池中的核酸以产生加标签的核酸的群体；扩增加标签的核酸，其中核酸和连接的标签被扩增；测定扩增的核酸和连接的标签的序列数据；解码标签以揭示已测定了针对其的序列数据的核酸是从第一池中的模板扩增的还是从第二池中的模板扩增的。

[0013] 在一些实施方案中，修饰是将核酸与蛋白质结合。在一些实施方案中，蛋白质是组蛋白或转录因子。在一些实施方案中，核酸修饰是对核苷酸的复制后修饰。在一些实施方案中，复制后修饰是5-甲基胞嘧啶，并且捕获剂与核酸的结合程度随着核酸中5-甲基胞嘧啶的程度而增加。在一些实施方案中，复制后修饰是5-羟甲基胞嘧啶，并且剂与核酸的结合程度随着核酸中5-羟甲基胞嘧啶的程度而增加。在一些实施方案中，复制后修饰是5-甲酰基胞嘧啶或5-羧基胞嘧啶，并且剂的结合程度随着核酸中5-甲酰基胞嘧啶或5-羧基胞嘧啶的程度而增加。在一些实施方案中，复制后修饰是N^6-甲基腺嘌呤。在一些实施方案中，所述方法还包括洗涤与剂结合的核酸，并且将洗涤物(wash)收集为第三池，所述第三池包括相对于第一池和第二池具有中等程度的复制后修饰的核酸。一些方法还包括，在测定之前，合并来自第一池和第二池的加标签的核酸。在一些实施方案中，剂包括甲基结合结构域或甲基-CpG-结合结构域(MBD)。MBD可以是蛋白质、抗体或能够特异性结合感兴趣的修饰的任何其他剂。优选地，MBD还包括磁珠、链霉亲和素或用于进行亲和分离步骤的其他结合结构域。

[0014] 本公开内容还提供了一种用于分析核酸群体的方法，其中至少一些核酸包括一个或更多个修饰的胞嘧啶残基。所述方法包括将捕获部分，例如生物素，连接至群体中的核酸，以用作用于扩增的模板；进行扩增反应以从模板产生扩增产物；分离与捕获部分连接的模板和扩增产物；通过亚硫酸氢盐测序测定与捕获部分连接的模板的序列数据；以及测定扩增产物的序列数据。

[0015] 在一些实施方案中，捕获部分包含生物素。在一些实施方案中，分离通过使模板与链霉亲和素珠接触来进行。在一些实施方案中，修饰的胞嘧啶残基是5-甲基胞嘧啶、5-羟甲基胞嘧啶、5-甲酰基胞嘧啶或5-羧基胞嘧啶。在一些实施方案中，捕获部分包含连接至包含一个或更多个修饰的残基的核酸标签的生物素。在一些实施方案中，捕获部分经由可裂解的连接体连接至群体中的核酸。在一些实施方案中，可裂解的连接体是可光裂解的连接体。在一些实施方案中，可裂解的连接体包括尿嘧啶核苷酸。

[0016] 本公开内容还提供了一种分析包含具有不同程度的5-甲基胞嘧啶的核酸的核酸群体的方法。所述方法包括(a)使核酸群体与优先结合5-甲基化核酸的剂接触；(b)分离与剂结合的第一核酸池和未与剂结合的第二核酸池，其中第一核酸池对于5-甲基胞嘧啶是呈现过度的，并且第二池中的核酸对于5-甲基化是呈现不足的；(c)将第一池和/或第二池中的核酸连接至一个或更多个区分第一池和第二池中的核酸的核酸标签，其中连接至第一池中的核酸的核酸标签包括捕获部分(例如生物素)；(d)扩增标记的核酸，其中核酸和连接的标签被扩增；(e)分离带有捕获部分的扩增的核酸和不带有捕获部分的扩增的核酸；以及(f)测定分离的、扩增的核酸的序列数据。

[0017] 本公开内容还提供了一种分析包含具有不同修饰程度的核酸的核酸群体的方法，包括：使群体中的核酸与衔接子接触以产生侧翼为包含引物结合位点的衔接子的核酸群体；从引物结合位点引发衔接子，扩增侧翼为衔接子的核酸；使扩增的核酸与优先结合带有修饰的核酸的剂接触；分离与剂结合的第一核酸池和未与剂结合的第二核酸池，其中第一核酸池对于修饰是呈现过度的，并且第二池中的核酸对于修饰是呈现不足的；对第一池和第二池中的核酸进行第二扩增步骤；以及测定第一池和第二池中的扩增的核酸的序列数据。每个池的扩增可以分别在不同的反应容器中发生。使用池特异性标签允许在测序之前对扩增子的后续合并。

[0018] 本公开内容还提供了一种分析核酸群体的方法，其中至少一些核酸包括一个或更多个修饰的胞嘧啶残基，所述方法包括使核酸群体与包含引物结合位点的衔接子接触以形成侧翼为衔接子的核酸，所述引物结合位点包含至少一个修饰的胞嘧啶；从核酸侧翼的衔接子中的引物结合位点引发，扩增侧翼为衔接子的核酸；将扩增的核酸分成第一等分试样和第二等分试样；测定第一等分试样的核酸的序列数据；使第二等分试样的核酸与亚硫酸氢盐接触，使未修饰的胞嘧啶(C)转化为尿嘧啶(U)；从核酸侧翼的引物结合位点引发，扩增由亚硫酸氢盐处理产生的核酸，其中通过亚硫酸氢盐处理引入的U被转化为T；测定来自第二等分试样的扩增的核酸的序列数据；比较第一等分试样和第二等分试样中的核酸的序列数据，以识别核酸群体中的哪些核苷酸是修饰的胞嘧啶。

[0019] 在任一种上文的方法中，核酸群体可以来自体液样品，诸如血液、血清或血浆。在一些实施方案中，核酸群体是无细胞核酸群体。在一些实施方案中，体液样品来自疑似具有癌症的受试者。

[0020] 在一个方面中，本文提供了一种分析核酸群体的方法，所述核酸群体包含选自双链DNA、单链DNA和单链RNA的至少两种形式的核酸，其中至少两种形式中的每一种包含多于一个分子，所述方法包括：将这些形式的核酸中的至少一种与至少一种标签核酸连接，以将这些形式彼此区分开；扩增这些形式的核酸，其中的至少一种连接至至少一种核酸标签，其中核酸和连接的核酸标签被扩增，以产生扩增的核酸，其中从所述至少一种形式扩增的核

酸被加标签;测定扩增的核酸的序列数据,其中的至少一些被加标签;其中所述测定获得了足以解码扩增的核酸的标签核酸分子的序列信息,以揭示群体中为已测定了针对其的序列数据的连接至标签核酸分子的扩增的核酸提供原始模板的核酸的形式。在一个实施方案中,所述方法还包括以下步骤:解码扩增的核酸的标签核酸分子,以揭示群体中为已测定了针对其的序列数据的连接至标签核酸分子的扩增的核酸提供原始模板的核酸的形式。在另一个实施方案中,所述方法还包括相对于一种或更多种其它形式富集这些形式中的至少一种。在另一个实施方案中,群体中的每种形式的核酸的至少70%的分子被扩增。在另一个实施方案中,群体中存在至少三种形式的核酸,并且所述形式中的至少两种连接至不同的标签核酸形式,所述不同的标签核酸形式将三种形式中的每一种彼此区分开。在另一个实施方案中,群体中的至少三种形式的核酸中的每一种连接至不同的标签。在另一个实施方案中,相同形式的每个分子连接至包含相同标签信息的标签。在另一个实施方案中,相同形式的分子连接至不同类型的标签。在另一个实施方案中,所述方法还包括使群体经历用加标签的引物的逆转录,其中加标签的引物被掺入从群体中的RNA产生的cDNA中。在另一个实施方案中,逆转录是序列特异性的。在另一个实施方案中,其中逆转录是随机的。在另一个实施方案中,所述方法还包括降解与cDNA成双链的RNA。在另一个实施方案中,所述方法还包括分离单链DNA和双链DNA,和将核酸标签连接至双链DNA。在另一个实施方案中,单链DNA通过与一种或更多种捕获探针杂交来分离。在另一个实施方案中,所述方法还包括用环连接酶(circligase)使单链DNA环化,和将核酸标签连接到双链DNA。在另一个实施方案中,所述方法包括在测定之前,合并包含不同形式的核酸的加标签的核酸。在另一个实施方案中,核酸群体来自体液样品。在另一个实施方案中,体液样品是血液、血清或血浆。在另一个实施方案中,核酸群体是无细胞核酸群体。在另一个实施方案中,体液样品来自疑似具有癌症的受试者。在另一个实施方案中,序列数据指示体细胞变体或生殖系变体的存在。在另一个实施方案中,序列数据指示拷贝数变异的存在。在另一个实施方案中,序列数据指示单核苷酸变异(SNV)、插入缺失或基因融合的存在。在另一个实施方案中,序列数据指示单核苷酸变异(SNV)、插入缺失或基因融合的存在。

[0021] 在另一个方面中,本文提供了分析核酸群体的方法,所述核酸群体包含具有不同修饰程度的核酸,所述方法包括:使核酸群体与优先结合带有修饰的核酸的剂接触;分离与剂结合的第一核酸池和未与剂结合的第二核酸池,其中第一核酸池对于修饰是呈现过度的,并且第二池中的核酸对于修饰是呈现不足的;将第一池和/或第二池中的核酸连接至一个或更多个核酸标签,所述核酸标签区分第一池和第二池中的核酸,以产生加标签的核酸的群体;扩增标记的核酸,其中核酸和连接的标签被扩增;以及测定扩增的核酸和连接的标签的序列数据;其中所述测定获得用于解码标签的序列数据以揭示已测定了针对其的序列数据的核酸是从第一池中的模板扩增的还是从第二池中的模板扩增的。在一个实施方案中,所述方法包括以下步骤:解码标签以揭示已测定了针对其的序列数据的核酸是从第一池中的模板扩增的还是从第二池中的模板扩增的。在另一个实施方案中,修饰是将核酸与蛋白质结合。在另一个实施方案中,蛋白质是组蛋白或转录因子。在另一个实施方案中,修饰是对核苷酸的复制后修饰。在另一个实施方案中,复制后修饰是5-甲基-胞嘧啶,并且剂与核酸的结合程度随着核酸中5-甲基-胞嘧啶的程度而增加。在另一个实施方案中,复制后修饰是5-羟甲基-胞嘧啶,并且剂与核酸的结合程度随着核酸中5-羟甲基-胞嘧啶的程度而

增加。在另一个实施方案中，复制后修饰是5-甲酰基-胞嘧啶或5-羧基-胞嘧啶，并且剂的结合程度随着核酸中5-甲酰基-胞嘧啶或5-羧基-胞嘧啶的程度而增加。在另一个实施方案中，所述方法还包括洗涤与剂结合的核酸，并且将洗涤物收集为第三池，所述第三池包括相对于第一池和第二池具有中等程度的复制后修饰的核酸。在另一个实施方案中，所述方法包括，在测定之前，合并来自第一池和第二池的加标签的核酸。在另一个实施方案中，剂是5-甲基-结合结构域磁珠。在另一个实施方案中，核酸群体来自体液样品。在另一个实施方案中，体液样品是血液、血清或血浆。在另一个实施方案中，核酸群体是无细胞核酸群体。在另一个实施方案中，体液样品来自疑似具有癌症的受试者。在另一个实施方案中，序列数据指示体细胞变体或生殖系变体的存在。在另一个实施方案中，序列数据指示拷贝数变异的存在。在另一个实施方案中，序列数据指示单核苷酸变异(SNV)、插入缺失或基因融合的存在。

[0022] 在另一个方面中，本文提供了一种分析核酸群体的方法，其中至少一些核酸包括一个或更多个修饰的胞嘧啶残基，所述方法包括将捕获部分连接至群体中的核酸，该核酸用作用于扩增的模板；进行扩增反应以从模板产生扩增产物；分离连接至捕获标签的模板和扩增产物；通过亚硫酸氢盐测序测定连接至捕获标签的模板的序列数据；以及测定扩增产物的序列数据。在一个实施方案中，捕获部分包含生物素。在另一个实施方案中，分离通过使模板与链霉亲和素珠接触来进行。在另一个实施方案中，修饰的胞嘧啶残基是5-甲基-胞嘧啶、5-羟甲基胞嘧啶、5-甲酰基胞嘧啶或5-羧基胞嘧啶。在另一个实施方案中，捕获部分包含连接至包含一个或更多个修饰的残基的核酸标签的生物素。在另一个实施方案中，捕获部分经由可裂解的连接体连接至群体中的核酸。在另一个实施方案中，可裂解的连接体是可光裂解的连接体。在另一个实施方案中，可裂解的连接体包括尿嘧啶核苷酸。在另一个实施方案中，核酸群体来自体液样品。在另一个实施方案中，体液样品是血液、血清或血浆。在另一个实施方案中，核酸群体是无细胞核酸群体。在另一个实施方案中，体液样品来自疑似具有癌症的受试者。在另一个实施方案中，序列数据指示体细胞变体或生殖系变体的存在。在另一个实施方案中，序列数据指示拷贝数变异的存在。在另一个实施方案中，序列数据指示单核苷酸变异(SNV)、插入缺失或基因融合的存在。

[0023] 在另一个方面中，本文提供了一种分析核酸群体的方法，所述核酸群体包含具有不同5-甲基化程度的核酸，所述方法包括：使核酸群体与优先结合5-甲基化核酸的剂接触；分离与剂结合的第一核酸池和未与剂结合的第二核酸池，其中第一核酸池对于5-甲基化是呈现过度的，并且第二池中的核酸对于5-甲基化是呈现不足的；将第一池和/或第二池中的核酸连接至一个或更多个区分第一池和第二池中的核酸的核酸标签，其中连接至第一池中的核酸的核酸标签包括捕获部分(例如生物素)；扩增标记的核酸，其中核酸和连接的标签被扩增；分离带有捕获部分的扩增的核酸和不带有捕获部分的扩增的核酸；以及测定分离的、扩增的核酸的序列数据。

[0024] 在另一个方面中，本文提供了一种分析包含具有不同修饰程度的核酸的核酸群体的方法，所述方法包括：使群体中的核酸与衔接子接触以产生侧翼为包含引物结合位点的衔接子的核酸群体；从引物结合位点引发，扩增侧翼为衔接子的核酸；使扩增的核酸与优先结合带有修饰的核酸的剂接触；分离与剂结合的第一核酸池和未与剂结合的第二核酸池，其中第一核酸池对于修饰是呈现过度的，并且第二池中的核酸对于修饰是呈现不足的；对

第一池和第二池中的加标签的核酸进行并行扩增;以及测定第一池和第二池中的扩增的核酸的序列数据。在另一个实施方案中,衔接子是发夹状衔接子。

[0025] 在另一个方面中,本文提供了一种分析核酸群体的方法,其中至少一些核酸包括一个或更多个修饰的胞嘧啶残基,所述方法包括使核酸群体与包含引物结合位点的衔接子接触以形成侧翼为衔接子的核酸,所述引物结合位点包含修饰的胞嘧啶;从核酸侧翼的衔接子中的引物结合位点引发,扩增侧翼为衔接子的核酸;将扩增的核酸分成第一等分试样和第二等分试样;测定第一等分试样的核酸的序列数据;使第二等分试样的核酸与亚硫酸氢盐接触,这使未修饰的C转化为U;从核酸侧翼的引物结合位点引发,扩增由亚硫酸氢盐处理产生的核酸,其中通过亚硫酸氢盐处理引入的U被转化为T;以及测定来自第二等分试样的扩增的核酸的序列数据;其中所述测定产生可以用于比较第一等分试样和第二等分试样中的核酸的序列数据以识别核酸群体中的哪些核苷酸是修饰的胞嘧啶的序列数据。在一个实施方案中,所述方法包括比较第一等分试样和第二等分试样中的核酸的序列数据,以识别核酸群体中的哪些核苷酸是修饰的胞嘧啶。在另一个实施方案中,衔接子是发夹状衔接子。

[0026] 在另一个方面中,本文提供了一种方法,包括:对来自人类样品的DNA分子物理地分级分离以生成两个或更多个分区;将差异化的分子标签和支持NGS(NGS-enabling)的衔接子应用于两个或更多个分区中的每一个以生成加分子标签的分区;在NGS仪器上测定加分子标签的分区以生成用于将样品解卷积成被差异化分区的分子的序列数据。在一个实施方案中,所述方法还包括通过将样品解卷积成被差异化分区的分子来分析序列数据。在另一个实施方案中,DNA分子来自提取的血浆。在另一个实施方案中,物理分级分离包括基于不同的甲基化程度对分子分级分离。在另一个实施方案中,不同的甲基化程度包括超甲基化和低甲基化。在另一个实施方案中,物理分级分离包括用甲基结合结构域蛋白(“MBD”)珠-分级分离,以形成不同的甲基化程度的层。在另一个实施方案中,差异化分子标签是对应于MBD分区的不同的分子标签组。在另一个实施方案中,物理分级分离包括使用免疫沉淀分离DNA分子。在另一个实施方案中,所述方法还包括重新组合所生成的加分子标签的级分中的两种或更多种加分子标签的级分。在另一个实施方案中,所述方法还包括富集重新组合的加分子标签的级分或组。在另一个实施方案中,一个或更多个特征是甲基化。在另一个实施方案中,分级分离包括使用包含甲基结合结构域的蛋白质分离甲基化的核酸和非甲基化的核酸,以生成包含不同甲基化程度的核酸分子的组。在另一个实施方案中,组中的一个包含超甲基化的DNA。在另一个实施方案中,至少一个组的特征在于甲基化程度。在另一个实施方案中,分级分离包括分离蛋白质结合的核酸。在另一个实施方案中,分离包括免疫沉淀。

[0027] 在另一个方面中,本文提供了一种用于通过NGS对MBD-珠分级分离的文库进行分子标签识别的方法,包括:使用甲基结合结构域蛋白质-珠纯化试剂盒对提取的DNA样品进行物理分级分离,保留所有洗脱物用于下游处理;将差异化的分子标签和支持NGS的衔接子序列并行应用于每个级分或组;重新组合所有加分子标签的级分或组,并且使用衔接子特异性DNA引物序列进行后续扩增;(d)对重新组合和扩增的总文库进行富集/杂交,靶向感兴趣的基因组区域;重新扩增富集的总DNA文库,附以样品标签;以及合并不同的样品,并且在NGS仪器上对其进行多重测定;其中由所述仪器产生的NGS序列数据提供了用于识别独特分

子的分子标签的序列,以及用于将样品解卷积成被差异化MBD分区的分子的序列数据。在一个实施方案中,所述方法包括用用于识别独特分子的分子标签对NGS数据进行分析,以及将样品解卷积成被差异化MBD分区的分子。在另一个实施方案中,分级分离包括物理分级分离。在另一个实施方案中,基于选自由以下组成的组的一个或更多个特征对核酸分子的群体分区:甲基化状态、糖基化状态、组蛋白修饰、长度和起始/终止位置。在另一个实施方案中,所述方法还包括合并核酸分子。在另一个实施方案中,分级分离包括基于单核小体特征谱的差异的分级分离。在另一个实施方案中,分级分离能够为至少一组核酸分子生成当与正常相比时不同的单核小体特征谱。在另一个实施方案中,所述方法还包括基于不同的特征分级分离至少一组核酸分子。在另一个实施方案中,分析包括将对应于第一组核酸分子的第一特征与对应于第二组核酸分子的第二特征在一个或更多个基因座处进行比较。在另一个实施方案中,核酸分子是循环肿瘤DNA。在另一个实施方案中,核酸分子是无细胞DNA(“cfDNA”)。在另一个实施方案中,标签用于区分同一样品中的不同分子。在另一个实施方案中,一个或更多个特征是癌症标志物。

[0028] 在另一个方面中,本文提供了一种方法,包括:提供从受试者的身体样品获得的核酸分子的群体;基于一个或更多个特征对核酸分子的群体分级分离以生成多于一组的核酸分子,基于一个或更多个特征将多于一组中的核酸分子差异化地加标签以将多于一组中的每一组中的核酸分子彼此区分开;对多于一组的核酸分子测序以生成序列读段;包含足够的数据以针对多于一组的核酸分子中的每一组生成关于核小体定位、核小体修饰或结合DNA-蛋白质相互作用的相关信息。在一个实施方案中,所述方法还包括分析序列读段以针对多于一组的核酸分子中的每一组生成关于核小体定位、核小体修饰或结合DNA-蛋白质相互作用的相关信息。在另一个实施方案中,所述方法还包括使用经训练的分类器以基于一个或更多个特征对受试者分类。在另一个实施方案中,一个或更多个特征包括映射读段的定量特征。在另一个实施方案中,分级分离包括物理分级分离。在另一个实施方案中,所述方法还包括合并核酸分子。在另一个实施方案中,分级分离包括基于单核小体特征谱的差异的分级分离。在另一个实施方案中,分级分离能够为至少一组核酸分子生成当与正常相比时不同的单核小体特征谱。在另一个实施方案中,所述方法还包括基于不同的特征对至少一组核酸分子分级分离。在另一个实施方案中,分析包括将对应于第一组核酸分子的第一特征与对应于第二组核酸分子的第二特征在一个或更多个基因座处进行比较。在另一个实施方案中,分析包括分析组的一个或更多个特征相对于正常样品在一个或更多个基因座处的特征。在另一个实施方案中,一个或更多个特征选自由以下组成的组:在参考序列上的一个碱基位置处的碱基调用频率、映射到参考序列上的一个碱基或序列的分子的数目、具有映射到参考序列上的一个碱基位置的起始位点的分子的数目和具有映射到参考序列上的一个碱基位置的终止位点的分子的数目,以及映射到参考序列上的一个基因座的分子的长度。在另一个实施方案中,所述方法还包括使用经训练的分类器基于一个或更多个特征对受试者分类。在另一个实施方案中,经训练的分类器将一个或更多个特征分类为与受试者中的组织相关。在另一个实施方案中,经训练的分类器将一个或更多个特征分类为与受试者中的癌症类型相关。在另一个实施方案中,一个或更多个特征指示基因表达或疾病状态。在另一个实施方案中,核酸分子是循环肿瘤DNA。在另一个实施方案中,核酸分子是无细胞DNA(“cfDNA”)。在另一个实施方案中,标签用于区分同一样品中的不同分子。在另一个实

施方案中,一个或更多个特征是癌症标志物。

[0029] 在另一个方面中,本文提供了一种方法,包括:提供从受试者的身体样品获得的核酸分子的群体;基于甲基化状态对核酸分子的群体分级分离以生成多于一组的核酸分子;基于一个或更多个特征将多于一组中的核酸分子差异化地加标签以将多于一组中的每一组中的核酸分子彼此区分开;对多于一组的核酸分子测序以生成序列读段;以及分析序列读段以检测多于一组的核酸分子中的一组核酸分子中的一个或更多个特征,其中一个或更多个特征指示核小体定位、核小体修饰或DNA-蛋白质相互作用。在另一个实施方案中,所述方法还包括使用经训练的分类器以基于一个或更多个特征对受试者分类。在另一个实施方案中,一个或更多个特征包括映射读段的定量特征。在另一个实施方案中,分级分离包括物理分级分离。在另一个实施方案中,所述方法还包括合并核酸分子。在另一个实施方案中,分级分离包括基于单核小体特征谱的差异的分级分离。在另一个实施方案中,分级分离能够为至少一组核酸分子生成当与正常相比时不同的单核小体特征谱。在另一个实施方案中,所述方法还包括基于不同的特征分级分离至少一组核酸分子。在另一个实施方案中,分析包括将对应于第一组核酸分子的第一特征与对应于第二组核酸分子的第二特征在一个或更多个基因座处进行比较。在另一个实施方案中,分析包括分析组的一个或更多个特征相对于正常样品在一个或更多个基因座处的特征。在另一个实施方案中,一个或更多个特征选自由以下组成的组:参考序列上的一个碱基位置处的碱基调用频率、映射到参考序列上的一个碱基或序列的分子的数目、具有映射到参考序列上的一个碱基位置的起始位点的分子的数目和具有映射到参考序列上的一个碱基位置的终止位点的分子的数目,以及映射到参考序列上的基因座的分子的长度。在另一个实施方案中,所述方法还包括使用经训练的分类器基于一个或更多个特征对受试者分类。在另一个实施方案中,经训练的分类器将一个或更多个特征分类为与受试者中的组织相关。在另一个实施方案中,经训练的分类器将一个或更多个特征分类为与受试者中的癌症类型相关。在另一个实施方案中,一个或更多个特征指示基因表达或疾病状态。在另一个实施方案中,核酸分子是循环肿瘤DNA。在另一个实施方案中,核酸分子是无细胞DNA(“cfDNA”)。在另一个实施方案中,标签用于区分同一样品中的不同分子。在另一个实施方案中,一个或更多个特征是癌症标志物。

[0030] 在另一个方面中,本文提供了一种方法,包括:提供从受试者的身体样品获得的核酸分子的群体;对核酸分子的群体分级分离以生成包含蛋白质结合的无细胞核酸的多于一组的核酸分子;基于一个或更多个特征将多于一组中的核酸分子差异化地加标签以将多于一组中的每一组中的核酸分子彼此区分开;以及对多于一组的核酸分子测序以生成序列读段;其中所获得的序列信息足以将序列读段映射到参考序列上的一个或更多个基因座;并且足以分析序列读段以检测多于一组的核酸分子中的一组的一个或更多个特征,其中一个或更多个特征指示核小体定位、核小体修饰或DNA-蛋白质相互作用。在一个实施方案中,所述方法还包括将序列读段映射到参考序列上的一个或更多个基因座;以及分析序列读段以检测多于一组的核酸分子中的一组的一个或更多个特征,其中一个或更多个特征指示核小体定位、核小体修饰或DNA-蛋白质相互作用。在另一个实施方案中,所述方法还包括使用经训练的分类器基于一个或更多个特征对受试者分类。在另一个实施方案中,一个或更多个特征包括映射读段的定量特征。在另一个实施方案中,分级分离包括物理分级分离。在另一个实施方案中,基于选自由以下组成的组的一个或更多个特征对核酸分子的群体分区:甲

基化状态、糖基化状态、组蛋白修饰、长度和起始/终止位置。在另一个实施方案中，所述方法还包括合并核酸分子。在另一个实施方案中，一个或更多个特征是甲基化。在另一个实施方案中，分级分离包括使用包含甲基结合结构域的蛋白质分离甲基化的核酸和非甲基化的核酸，以生成包含不同甲基化程度的核酸分子的组。在另一个实施方案中，组中的一个包含超甲基化的DNA。在另一个实施方案中，至少一个组的特征在于甲基化程度。在另一个实施方案中，分级分离包括分离单链DNA分子和/或双链DNA分子。在另一个实施方案中，双链DNA分子使用发夹状衔接子来分离。在另一个实施方案中，分级分离包括分离蛋白质结合的核酸。在另一个实施方案中，分级分离包括基于单核小体特征谱的差异的分级分离。在另一个实施方案中，分级分离能够为至少一组核酸分子生成当与正常相比时不同的单核小体特征谱。在另一个实施方案中，分离包括免疫沉淀。在另一个实施方案中，所述方法还包括基于不同的特征分级分离至少一组核酸分子。在另一个实施方案中，分析包括将对应于第一组核酸分子的第一特征与对应于第二组核酸分子的第二特征在一个或更多个基因座处进行比较。在另一个实施方案中，分析包括分析组的一个或更多个特征相对于正常样品在一个或更多个基因座处的特征。在另一个实施方案中，一个或更多个特征选自由以下组成的组：在参考序列上的一个碱基位置处的碱基调用频率、映射到参考序列上的一个碱基或序列的分子的数目、具有映射到参考序列上的一个碱基位置的起始位点的分子的数目和具有映射到参考序列上的一个碱基位置的终止位点的分子的数目，以及映射到参考序列上的一个基因座的分子的长度。在另一个实施方案中，所述方法还包括使用经训练的分类器基于一个或更多个特征对受试者分类。在另一个实施方案中，经训练的分类器将一个或更多个特征分类为与受试者中的组织相关。在另一个实施方案中，经训练的分类器将一个或更多个特征分类为与受试者中的癌症类型相关联。在另一个实施方案中，一个或更多个特征指示基因表达或疾病状态。在另一个实施方案中，核酸分子是循环肿瘤DNA。在另一个实施方案中，核酸分子是无细胞DNA(“cfDNA”)。在另一个实施方案中，标签用于区分同一样品中的不同分子。

[0031] 在另一个方面中，本文提供了一种方法，包括：提供从受试者的身体样品获得的核酸分子的群体；基于一个或更多个特征对核酸分子的群体分级分离以生成多于一组的核酸分子；基于一个或更多个特征将多于一组中的核酸分子差异化地加标签以将多于一组中的每一组中的核酸分子彼此区分开；对多于一组的核酸分子测序以生成序列读段；其中所获得的序列信息足以将序列读段映射到参考序列上的一个或更多个基因座；以及分析序列读段以检测多于一组的核酸分子中的一组的一个或更多个特征，其中一个或更多个特征在来自多于一组的序列读段的池中不能检测。在一个实施方案中，所述方法还包括将序列读段映射到参考序列上的一个或更多个基因座；以及分析序列读段以检测多于一组的核酸分子中的一组的一个或更多个特征，其中一个或更多个特征在来自多于一组的序列读段的池中不能检测。在另一个实施方案中，分级分离包括物理分级分离。

[0032] 在另一个方面中，本文提供了一种方法，包括：提供从受试者的身体样品获得的核酸分子的群体；基于一个或更多个特征对核酸分子的群体分级分离以生成多于一组的核酸分子，其中多于一组中的每一组的核酸分子包含不同的标识物；合并多于一组的核酸分子；对合并的多于一组的核酸分子测序以生成多于一组序列读段；以及基于标识物对序列读段分级分离。

[0033] 在另一个方面中，本文提供了一种组合物，所述组合物包含含有差异化地加标签的核酸分子的核酸分子的池，其中所述池包含多于一组核酸分子，所述多于一组核酸分子基于选自由以下组成的组的一个或更多个特征被差异化地加标签：甲基化状态、糖基化状态、组蛋白修饰、长度和起始/终止位置，其中所述池来源于生物样品。在一个实施方案中，多于一组是2组、3组、4组、5组或多于5组中的任一种。

[0034] 在另一个方面中，本文提供了一种方法，包括：将核酸分子的群体分级分离成多于一组，所述多于一组包括特征不同的核酸；用一组标签对多于一组中的每一组中的核酸加标签，所述一组标签区分多于一组中的每一组中的核酸，以产生加标签的核酸的群体，其中加标签的核酸中的每一个包含一个或更多个标签；对加标签的核酸的群体测序以生成序列读段；使用一个或更多个标签将每组序列读段分组；以及分析序列读段以检测至少一个组相对于正常样品或分类器的信号。在一个实施方案中，所述方法还包括针对另一组或全基因组序列将至少一个组的信号归一化。

[0035] 在另一个方面中，本文提供了一种方法，包括：提供来自生物样品的无细胞DNA的群体；基于相比于非癌细胞以不同水平存在于来源于癌细胞的无细胞DNA的特征，对无细胞DNA的群体分级分离，从而生成无细胞DNA的亚群；扩增无细胞DNA的亚群中的至少一个；以及对扩增的无细胞DNA的亚群中的至少一个测序。在一个实施方案中，特征是：无细胞DNA的甲基化水平；无细胞DNA的糖基化水平；无细胞DNA片段的长度；或无细胞DNA中单链断裂的存在。

[0036] 在另一个方面中，本文提供了一种方法，包括：提供来自生物样品的无细胞DNA的群体；基于无细胞DNA的甲基化水平对无细胞DNA的群体分级分离，从而生成无细胞DNA的亚群；扩增无细胞DNA的亚群中的至少一个；以及对扩增的无细胞DNA的亚群中的至少一个测序。

[0037] 在另一个方面中，本文提供了一种用于确定无细胞DNA的甲基化状态的方法，包括：提供来自生物样品的无细胞DNA的群体；基于无细胞DNA的甲基化水平对无细胞DNA的群体分级分离，从而生成无细胞DNA的亚群；对无细胞DNA的至少一个亚群测序，从而生成序列读段；以及，根据相应序列读段出现在其中的亚群，为每个无细胞DNA指定甲基化状态。

[0038] 在另一个方面中，本文提供了一种对受试者分类的方法，其中所述方法包括：提供来自受试者的生物样品的无细胞DNA的群体；基于无细胞DNA的甲基化水平对无细胞DNA的群体分级分离，从而生成无细胞DNA的亚群；对无细胞DNA的亚群测序，从而生成序列读段；以及，使用经训练的分类器根据序列读段在哪个亚群中出现对受试者分类。在另一个实施方案中，无细胞DNA的群体通过提供健康状态和患病状态之间的信号差异的一个或更多个特征来分级分离。在另一个实施方案中，基于无细胞DNA的甲基化水平对无细胞DNA的群体分级分离。在另一个实施方案中，确定无细胞DNA的片段化模式还包括分析映射到参考基因组中的每个碱基位置的序列读段的数目。在另一个实施方案中，所述方法还包括通过分析映射到参考基因组中的每个碱基位置的序列读段的数目来确定每个亚群中的无细胞DNA的片段化模式。

[0039] 在另一个方面中，本文提供了一种用于分析无细胞DNA的片段化模式的方法，包括：提供来自生物样品的无细胞DNA的群体；对无细胞DNA的群体分级分离，从而生成无细胞DNA的亚群；对无细胞DNA的至少一个亚群测序，从而生成序列读段；将序列读段与参考基因

组比对;以及,通过分析以下的任一数目来确定每个亚群中的无细胞DNA的片段化模式:映射到参考基因组中的每个碱基位置的每个序列读段的长度;作为序列读段的长度的函数的映射到参考基因组中的碱基位置的序列读段的数目;在参考基因组中的每个碱基位置起始的序列读段的数目;或在参考基因组中的每个碱基位置终止的序列读段的数目。在另一个实施方案中,一个或更多个特征包括选自由以下组成的组的化学修饰:甲基化、羟甲基化、甲酰化、乙酰化和糖基化。

[0040] 本文描述的任一方法,,其中DNA:珠的比为1:100。

[0041] 本文描述的任一方法,其中DNA:珠的比为1:50。

[0042] 本文描述的任一方法,其中DNA:珠的比为1:20。

[0043] 在一个方面中,本文提供了在循环肿瘤DNA(ctDNA)的分析期间使用基于DNA甲基化程度的物理分级分离来确定基因表达或疾病状态。

[0044] 在一个方面中,本文提供了使用提供正常状态和患病状态之间的信号差异的特征在ctDNA分析期间对ctDNA物理地分区。

[0045] 在一个方面中,本文提供了使用提供正常状态和患病状态之间的信号差异的特征对ctDNA物理地分区。

[0046] 在一个方面中,本文提供了使用提供正常状态和患病状态之间的信号差异的特征在测序和任选的下游分析之前对ctDNA物理地分区。

[0047] 在一个方面中,本文提供了使用提供正常状态和患病状态之间的信号差异的特征对ctDNA物理地分区以便对其差异化标记/加标签。在一个实施方案中,差异片段化模式指示基因表达或疾病状态。在另一个实施方案中,差异片段化模式的特征在于相对于正常的选自由以下组成的组的一个或更多个差异:映射到参考基因组中的每个碱基位置的每个序列读段的长度;作为序列读段的长度的函数的映射到参考基因组中的碱基位置的序列读段的数目;在参考基因组中的每个碱基位置处起始的序列读段的数目;以及在参考基因组中的每个碱基位置处终止的序列读段的数目。

[0048] 在一个方面中,本文提供了在ctDNA分析期间使用基于差异片段化模式的分级分离。在一个实施方案中,差异片段化模式指示基因表达或疾病状态。在另一个实施方案中,差异片段化模式的特征在于相对于正常的选自由以下组成的组的一个或更多个差异:映射到参考基因组中的每个碱基位置的每个序列读段的长度;作为序列读段的长度的函数的映射到参考基因组中的碱基位置的序列读段的数目;在参考基因组中的每个碱基位置处起始的序列读段的数目;以及在参考基因组中的每个碱基位置处终止的序列读段的数目。

[0049] 在一个方面中,本文提供了使用差异片段化模式对ctDNA分区。在一个实施方案中,差异片段化模式指示基因表达或疾病状态。在另一个实施方案中,差异片段化模式的特征在于相对于正常的选自由以下组成的组的一个或更多个差异:映射到参考基因组中的每个碱基位置的每个序列读段的长度;作为序列读段的长度的函数的映射到参考基因组中的碱基位置的序列读段的数目;在参考基因组中的每个碱基位置处起始的序列读段的数目;以及在参考基因组中的每个碱基位置处终止的序列读段的数目。

[0050] 在一个方面中,本文提供了在测序和任选的下游分析之前使用差异片段化模式对ctDNA分区。在一个实施方案中,差异片段化模式指示基因表达或疾病状态。在另一个实施方案中,差异片段化模式的特征在于相对于正常的选自由以下组成的组的一个或更多个差

异:映射到参考基因组中的每个碱基位置的每个序列读段的长度;作为序列读段的长度的函数的映射到参考基因组中的碱基位置的序列读段的数目;在参考基因组中的每个碱基位置处起始的序列读段的数目;以及在参考基因组中的每个碱基位置处终止的序列读段的数目。

[0051] 在一个方面中,本文提供了使用差异片段化模式对ctDNA分区以便对其差异化标记/加标签。

[0052] 在一个方面中,本文提供了对由分子结合结构域(MBD)-珠分区的DNA分子使用差异化地加分子标签以形成不同的DNA甲基化程度的层,然后通过下一代测序(NGS)来定量所述不同的DNA甲基化程度。

[0053] 在一个方面中,本文提供了一种分析核酸群体的方法,所述核酸群体包含选自双链DNA、单链DNA和单链RNA的至少两种形式的核酸,其中至少两种形式中的每一种包含多于一个分子,所述方法包括:将这些形式的核酸中的至少一种与至少一种标签核酸连接,以将这些形式彼此区分开;扩增这些形式的核酸,其中的至少一种连接至至少一种核酸标签,其中核酸和连接的核酸标签被扩增以产生扩增的核酸,其中从至少一种形式扩增的核酸被加标签;以及对已经连接至标签的多于一个的扩增的核酸测序,其中序列数据足以被解码以揭示群体中的核酸在连接至至少一种标签之前的形式。在一个实施方案中,分子标签包括一个或多于一个核酸条形码。在另一个实施方案中,一组中任何两个条形码的组合具有与任何其他组中任何两个条形码的组合相比不同的组合序列。

[0054] 在另一个方面中,本文提供了加标签的核酸分子的池,池中的每个核酸分子包括选自多于一个标签组中的一个标签组的分子标签,每个标签组包含多于一个不同的标签,其中任何一组中的标签不同于任何其他组中的标签,并且其中每个标签组包含(i)指示其所附接的分子或该分子所源自的亲本分子的特征的信息,以及(ii)单独地或与来自其所附接的分子的信息组合地,将其所附接的分子与用来自相同标签组的标签加标签的其他分子独特地区分开的信息。在一个实施方案中,分子标签包括附接在分子的相对端的两个核酸条形码。在另一个实施方案中,条形码的长度在10个和30个核苷酸之间。

[0055] 在另一个方面中,本文提供了一种系统,包括:核酸测序仪;包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器的数字处理设备;以及通信连接核酸测序仪和数字处理设备的数据链路;其中数字处理设备还包括可被执行以创建用于分析核酸群体的应用程序的指令,所述核酸群体包括选自双链DNA、单链DNA和单链RNA的至少两种形式的核酸,至少两种形式中的每一种包含多于一个分子,所述应用程序包括:软件模块,其经由数据链路从核酸测序仪接收序列数据、扩增的核酸的序列数据,扩增的核酸中的至少一些被加标签;所述序列数据通过以下生成:将所述形式的核酸中的至少一种与至少一种加标签的核酸连接以将所述形式彼此区分开,扩增所述形式的核酸,其中的至少一种被连接至至少一种核酸标签,其中核酸和连接的核酸标签被扩增以产生扩增的核酸,其中从至少一种形式扩增的核酸被加标签;以及软件模块,其通过获得足以解码扩增的核酸的加标签的核酸分子的序列信息来测定扩增的核酸的序列数据,以揭示群体中为已测定了针对其的序列数据的连接至标签核酸分子的扩增的核酸提供原始模板的核酸的形式。在一个实施方案中,应用程序还包括解码扩增的核酸的加标签的核酸分子的软件模块,以揭示群体中为已测定了针对其的序列数据的连接至标签核酸分子的扩增的核酸提供原始模板的核酸的

形式。在另一个实施方案中，应用程序还包括经由通信网络传输测定结果的软件模块。

[0056] 在另一个方面中，本文提供了一种系统，包括：下一代测序(NGS)仪器；包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器的数字处理设备；以及通信连接NGS仪器和数字处理设备的数据链路；其中数字处理设备还包括可被执行以创建应用程序的指令，所述应用程序包括：用于经由数据链路从NGS仪器接收序列数据的软件模块，所述序列数据通过以下来生成：对来自人类样品的DNA分子物理分级分离以生成两个或更多个分区，将差异化分子标签和支持NGS的衔接子应用于两个或更多个分区中的每一个以生成加分子标签的分区，以及用NGS仪器测定加分子标签的分区；用于生成用于将样品解卷积成被差异化分区的分子的序列数据的软件模块；以及用于通过将样品解卷积成被差异化分区的分子来分析序列数据的软件模块。在一个实施方案中，应用程序还包括经由通信网络传输测定结果的软件模块。

[0057] 在另一个方面中，本文提供了一种系统，其包括：下一代测序(NGS)仪器；包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器的数字处理设备；以及通信连接NGS仪器和数字处理设备的数据链路；其中数字处理设备还包括可由至少一个处理器执行以创建用于MBD-珠分级分离文库的分子标签识别的应用程序的指令，所述应用程序包括：被配置为经由数据链路从NGS仪器接收序列数据的软件模块，所述序列数据通过以下生成：使用甲基结合结构域蛋白质珠纯化试剂盒对提取的DNA样品进行物理分级分离，保留所有洗脱物用于下游处理；差异化分子标签和支持NGS的衔接子序列对每个级分或组的并行应用；重新组合所有加分子标签的级分或组，并且随后使用衔接子特异性DNA引物序列进行扩增；对重新组合并扩增的总文库进行富集/杂交，靶向感兴趣的基因组区域；重新扩增富集的总DNA文库，附以样品标签；合并不同的样品，和在NGS仪器上对其进行多重测定；其中由所述仪器产生的NGS序列数据提供了用于识别独特分子的分子标签的序列，以及用于将样品解卷积成被差异化MBD分区的分子的序列数据；以及被配置为通过使用分子标签以识别独特的分子并且将样品解卷积成被差异化MBD-分区的分子来进行序列数据的分析的软件模块。在一个实施方案中，应用程序还包括被配置为经由通信网络传输分析结果的软件模块。

[0058] 上文提供的概述是实施方案的示例性列表，而不意图是实施方案的完整列表。

[0059] 通过引用并入

[0060] 本说明书中提及的所有出版物、专利和专利申请通过引用并入本文，其程度如同每一个单独的出版物、专利或专利申请被具体和单独地指明通过引用并入的相同程度。

[0061] 附图简述

[0062] 图1示出了用于对RNA、单链DNA和双链DNA分区的示例性方案。

[0063] 图2示出了用于对RNA、单链DNA和双链DNA分区的另外的示例性方案。

[0064] 图3示出了用于分析含有不同程度的5-甲基胞嘧啶呈现的DNA的方案。

[0065] 图4示出了用于对甲基化DNA进行亚硫酸氢盐测序的方案。

[0066] 图5示出了用于分析含有不同程度的5-甲基胞嘧啶呈现的DNA的另外的方案。

[0067] 图6示出了用于对甲基化DNA进行亚硫酸氢盐测序的另外的方案。

[0068] 图7示出了差异化加标签的概述。

[0069] 图8示出了分区方法的概述。

[0070] 图9示出了方法的概述。

[0071] 图10示出了对分级分离的核酸分子使用片段组(fragmentomic)数据分析的实例。基因组位置示出在X轴上,片段长度示出在Y轴上并且覆盖度或拷贝示出在Z轴上,且显示出升高的低甲基化或超甲基化的相应区域。

[0072] 图11示出了正常样品和肺癌样品的甲基化特征分析。

[0073] 图12A、图12B和图12C示出了使用全基因组测序的甲基化特征分析。图12A示出了沿X轴的转录起始位点(TSS)的600bp区域的位置和沿Y轴的超甲基化位点的频率。图12B示出了沿X轴的转录起始位点(TSS)的600bp区域的位置和沿Y轴的低甲基化位点的频率。图12C示出了X轴上的超甲基化百分比和Y轴上的片段长度。

[0074] 图13A和图13B示出了MOB3A和WDR88的甲基化特征分析。图13A在X轴上示出了MOB3A基因在基因组中的位置,并且来自不同分级分离的组的核酸分子的片段长度通过单独的行显示。分级分离的组包括用于比较的超甲基化组、低甲基化组、超甲基化与低甲基化的混合组(hyper+hypo)和未分级分离的组(无MBD)。

[0075] 图14A和图14B示出了分级分离的组和未分级分离的组的甲基化特征分析。图14A分别在X轴和Y轴上示出了具有对来自未分级分离组(无MBD)和来自分级分离后的混合分区的覆盖度的热图。

[0076] 图15示出了分级分离的样品和未分级分离的样品的核小体组织方式。

[0077] 图16示出了MBD信号的验证。

[0078] 图17显示了关于输入基因组区域与所有被假定受多个基因组区域调节的基因的TSS关联的统计数据。X轴指示到TSS的距离,以千碱基(kb)计,而Y轴指示区域-基因关联,以百分比(%)计。在图中的每个条形上面,列出了被计数的项目的绝对数目。由暗条形表示的最显著的(foreground)基因组区域选自由亮条指示的背景基因组区域的超集。背景基因组区域是选自基因组中的所有重复元件的已被指派功能性作用的重复元件。

[0079] 图18A和图18B示出了AP3D1基因的甲基化特征分析。图18A示出了X轴上AP3D1基因的基因组位置和对来自以单独的行指示的不同组的核酸分子的读段的覆盖度。这些组包括分级分离的组,诸如超甲基化组、低甲基化组和未分级分离的组(无MBD),以用于比较。TSS被示为热图中间的垂直线,箭头指示转录的方向。图18B在X轴上示出了超甲基化百分比并在Y轴上示出了片段长度。例如,在图18B中,未分级分离的核酸样品中的甲基化百分比可以是约65%,如红色虚线所示的。

[0080] 图19A和图19B示出了DNMT1基因的甲基化特征分析。图19A示出了X轴上DNMT1基因的基因组位置和对来自以单独的行指示的不同组的核酸分子的读段的覆盖度。这些组包括分级分离的组,诸如超甲基化组、低甲基化组和未分级分离的组(无MBD),以用于比较。TSS被示为热图中间的垂直线,箭头指示转录的方向。图19B在X轴上示出了超甲基化百分比并在Y轴上示出了片段长度。

[0081] 图20示出了基于核酸分子的链型的分级分离的程序。

[0082] 图21示出了核酸分子被分级分离成ssDNA和dsDNA。X轴示出了具有不同输入DNA(200ng和500ng)的两个样品的两次技术重复。Y轴示出了使用定量PCR扩增的中靶分子的拷贝数。该图示出了对每组分级分离的cfDNA中靶序列的定量确定。

[0083] 图22示出了在将核酸分子分级分离成ssDNA和dsDNA后的PCR收率。X轴示出了以两

次技术重复的cfDNA输入(200ng和500ng),而Y轴示出了PCR收率,以pmol计。

[0084] 图23使用全基因组测序示出了启动子区域的甲基化特征分析。

[0085] 图24提供了使用甲基结合结构域蛋白(MBD-分区)对被分区或分级分离的核酸分子加标签的策略的三个实例。

[0086] 图25A和图25B示出了在靶向测序测定中对于MBD样品和无MBD样品的覆盖度之间的比较。

[0087] 图26A和图26B示出了使用15ng的cfDNA输入和两个临床样品(PowerpoolV1和PowerpoolV2),对组中基因的覆盖度。

[0088] 图27A和图27B示出了使用150ng的cfDNA输入和两个临床样品(PowerpoolV1和PowerpoolV2),对组中的基因的覆盖度。

[0089] 图28A、图28B和图28C示出了使用15ng的cfDNA输入,检测组中基因的变体或突变的特异性和灵敏度。

[0090] 图29A、图29B和图29C示出了使用150ng的cfDNA输入,检测组中的基因的变体或突变的特异性和灵敏度。

[0091] 图30示出了通过全基因组亚硫酸氢盐测序(WGBS)和MBD分区测量的平均甲基化水平之间的相关性。

[0092] 图31A和图31B示出了使用MBD分区(Y轴)和使用全基因组亚硫酸氢盐测序测定(WGBS,X轴)检测甲基化DNA的灵敏度(图31A)和特异性(图31B)。

[0093] 图32示出了数字处理设备的实施方案。

[0094] 图33示出了应用程序供应系统的实施方案。

[0095] 图34示出了采用基于云的架构的应用程序供应系统的实施方案。

[0096] 详细描述

[0097] 如本文使用的术语“无细胞DNA”和“无细胞DNA群体”指的是最初发现于大型复杂生物有机体例如哺乳动物的一个或多个细胞中的DNA,并且从所述细胞释放到有机体中存在的液体流体中,例如血浆、淋巴、脑脊液、尿液,其中DNA可以通过获得流体样品来获得而不需要进行体外细胞裂解步骤。

[0098] 一般性

[0099] 本公开内容提供了许多方法、试剂、组合物和系统,其用于分析复杂基因组材料,同时减少或消除最初存在于复杂基因组材料中的分子特征(例如表观遗传或其他类型的结构)信息的损失。在一些实施方案中,分子标签可以用于追踪不同形式的核酸,并且对这样的不同形式计数,以用于确定遗传修饰(例如,SNV、插入缺失、基因融合和拷贝数变异)的目的。在一些实施方案中,本文描述的方法用于检测、分析或监测受试者中的状况诸如癌症或胎儿状态。在一些实施方案中,受试者是未孕的。

[0100] 本公开内容提供了用于处理包含不同形式的核酸群体的方法。如本文使用的,不同形式的核酸具有不同的特征。例如,且非限制性地,RNA和DNA是基于糖的身份而不同的形式。单链(ss)核酸和双链(ds)核酸的链的数目不同。核酸分子可以基于表观遗传特征诸如5-甲基胞嘧啶或与蛋白质诸如组蛋白的缔合而不同。核酸可以具有不同的核苷酸序列,例如特定的基因或遗传基因座。特征在程度方面可以不同。

[0101] 例如,DNA分子在其表观遗传修饰的程度上可以不同。修饰程度可以指的是分子已

经经历的修饰事件的数目，诸如甲基化基团的数目(甲基化程度)或其他表观遗传变化的数目。例如，甲基化的DNA可以是低甲基化的或超甲基化的。形式可以通过特征的组合来表征，例如单链未甲基化的或双链甲基化的。基于一个特征或特征的组合的分子的分级分离对于单个分子的多维分析可以是有用的。这些方法适应样品中核酸的多种形式和/或修饰，使得可以获得针对多种形式的序列信息。所述方法经过处理和分析仍保持初始的多种形式或修饰状态的身份，使得核碱基序列的分析可以与表观遗传分析组合。一些方法涉及分离、加标签和随后合并不同的形式或修饰状态，减少分析样品中存在的多种形式所需的处理步骤的数目。对样品中的多种形式的核酸的分析提供了更多的信息，部分是因为存在更多待分析的分子(当非常低总量的核酸可用时，这可能是重要的)，而且还因为不同的形式或修饰状态可以提供不同的信息(例如，突变可能仅存在于RNA中)，并且因为不同类型的信息(例如，遗传和表观遗传)可能相互关联，从而产生更大的准确度、确定性，或导致发现新的与医学状况的关联。

[0102] CpG二核苷酸在正常人类基因组中是呈现不足的，其中大部分CpG二核苷酸序列是转录惰性的(例如染色体的近着丝粒部分和重复元件中的DNA异染色质区域)并且是甲基化的。然而，许多CpG岛被保护免受这样的甲基化，尤其是在转录起始位点(TSS)周围。

[0103] 癌症可以通过表观遗传变异诸如甲基化来指示。癌症中甲基化变化的实例包括参与正常生长控制、DNA修复、细胞周期调节和/或细胞分化的基因的转录起始位点(TSS)处的CpG岛中的DNA甲基化的局部增加。该超甲基化可以与参与的基因的转录能力的异常损失有关，并且至少与导致基因表达改变的点突变和缺失一样频繁地发生。DNA甲基化特征分析可以用于检测基因组具有不同甲基化程度的区域(“差异化的甲基化区域”或“DMR”)，这些区域在发育期间改变或受到疾病例如癌症或任何癌症相关的疾病的扰动。癌细胞的基因组在上文的DNA甲基化模式方面具有不平衡，并且因此在DNA的功能包装方面也具有不平衡。因此，染色质组织的异常与甲基化变化相关联，并且当联合分析时，可以有助于增强癌症特征分析。结合MBD分区与片段组数据，诸如片段映射的起始和终止位置(与核小体位置相关)、片段长度和相关的核小体占据，可以用于超甲基化研究中的染色质结构分析，目的是提高生物标志物检测率。

[0104] 甲基化特征分析可涉及确定跨基因组的不同区域的甲基化模式。例如，在基于甲基化程度(例如，每个分子甲基化位点的相对数目)对分子分区和测序之后，可将不同分区中的分子的序列映射到参考基因组。这可以示出基因组中与其他区域相比甲基化更高或甲基化不太高的区域。以该方式，与单个分子相比，基因组区域在其甲基化程度上可以不同。

[0105] 核酸分子的特征可以是修饰，其可以包括各种化学修饰或蛋白质修饰(即表观遗传修饰)。化学修饰的非限制性实例可以包括但不限于共价DNA修饰，包括DNA甲基化。在一些实施方案中，DNA甲基化包括在CpG(在核酸序列中胞嘧啶后跟鸟嘌呤)位点将甲基基团添加至胞嘧啶。在一些实施方案中，DNA甲基化包括将甲基基团添加至腺嘌呤，例如在N^6-甲基腺嘌呤中。在一些实施方案中，DNA甲基化是5-甲基化(对胞嘧啶的6元碳环的第5个碳的修饰)。在一些实施方案中，5-甲基化包括将甲基基团添加至胞嘧啶的5C位置，以产生5-甲基胞嘧啶(m5c)。在一些实施方案中，甲基化包括m5c的衍生物。m5c的衍生物包括但不限于5-羟甲基胞嘧啶(5-hmC)、5-甲酰基胞嘧啶(5-fC)和5-羧基胞嘧啶(5-caC)。在一些实施方案中，DNA甲基化是3C甲基化(对胞嘧啶的6元碳环的第3个碳的修饰)。在一些实施方案中，3C

甲基化包括将甲基基团添加至胞嘧啶的3C位置,以生成3-甲基胞嘧啶(3mC)。甲基化还可以发生在非CpG位点,例如,甲基化可以发生在CpA、CpT或CpC位点。DNA甲基化可以改变甲基化DNA区域的活性。例如,当启动子区域中的DNA被甲基化时,基因的转录可以被抑制。DNA甲基化对正常发育至关重要,并且甲基化的异常可能破坏表观遗传调节。表观遗传调节中的破坏,例如抑制,可能导致疾病,诸如癌症。DNA中启动子甲基化可以指示癌症。

[0106] 蛋白质修饰包括结合染色质组分,特别是组蛋白,包括其修饰形式,以及结合其他蛋白质,诸如参与复制或转录的蛋白质。本公开内容提供了处理和分析具有不同修饰程度的核酸的方法,使得它们最初修饰的性质与核酸标签相关联,并且可以在分析核酸时通过对标签测序来解码。样品核酸修饰的遗传变异然后可以与原始样品中核酸的修饰程度(表观遗传变异)相关联。

[0107] 如本文使用的,术语“分级分离”和“分区”指的是基于不同特征分离分子。样品中的核酸分子可以基于一个或更多个特征来分级分离。分级分离可以包括基于基因组特征的存在或不存在将核酸分子物理地分区成子集或组。分级分离可以包括基于基因组特征存在的程度将核酸分子物理地分区到分区组。基于指示差异化基因表达或疾病状态的特征,可以将样品分级分离或分区到一个或更多个组分区。样品在核酸分析期间可以基于在正常状态和患病状态之间提供信号差异的特征或其组合来分级分离,所述核酸例如无细胞DNA(“cfDNA”)、非cfDNA、肿瘤DNA、循环肿瘤DNA(“ctDNA”)和无细胞核酸(“cfNA”)。

[0108] 本公开内容提供了用于有效分析核酸分子的方法和系统。所述方法可以包括基于一个或多于一个特征将核酸分子分级分离到不同的分区,然后测序(单独地或一起)并分析每个分区中的核酸分子。在一些情况下,核酸分子的分区在测序之前和/或之后被扩增。所述方法可以用于各种应用,诸如预后、诊断和/或用于疾病监测。

[0109] 核酸分子可以通过一个或更多个特征中的任一种来表征。核酸分子的特征可以包括链型、蛋白质结合区域、核酸长度、起始/终止位置、化学修饰或蛋白质修饰。核酸分子的链型可以包括单链分子(例如ssDNA或RNA)或双链分子(例如dsDNA)。

[0110] 核酸分子的基因组特征可以是修饰,其可以包括各种化学修饰。作为非限制性实例,化学修饰可以包括共价DNA修饰,诸如DNA甲基化(5mC)、羟甲基化(5hmC)、甲酰基甲基化(5fC)、羧基甲基化(5CaC)、N^6-甲基腺嘌呤或糖基化。DNA甲基化包括将甲基基团添加至DNA(例如CpG),并且可以改变甲基化DNA区域的表达。例如,当启动子区域中的DNA被甲基化时,基因的转录可以被抑制。DNA甲基化对正常发育至关重要,并且甲基化的异常可能破坏表观遗传调节。表观遗传调节中的破坏,例如抑制,可能导致疾病,诸如癌症。DNA中启动子甲基化可以指示癌症。

[0111] 作为非限制性实例,涉及分区单链RNA和/或DNA以及双链DNA以表征样品的方法的益处包括:

[0112] 1.除dsDNA外,来自ssDNA和RNA分子的对SNV、CNV和插入缺失判别的额外支持;

[0113] 2.与DNA相比,RNA中基因融合的识别(靶向)更容易,因为内含子DNA中可变的断点在RNA中产生确定的外显子-外显子连接;

[0114] 3.信使RNA(mRNA)、微RNA(miRNA)和长非编码RNA(lncRNA)的识别或差异化表达水平可以是许多疾病状态的特征。确认和额外支持循环肿瘤DNA(ctDNA)群体与来自白细胞的健康的无细胞DNA(cfDNA)相比在核小体定位变化方面存在的表达特征在癌症的早期检测

中可能是重要的。此外,白细胞来源的cfDNA和cfRNA表达变化也可以指示对疾病的免疫应答。

[0115] 4.不稳定分子的证据。捕获较短的循环肿瘤DNA(ctDNA)—对无细胞DNA的研究发现肿瘤DNA(ctDNA)的长度可显著短于正常DNA。一些证据指示,这些较短的序列是不稳定的,并且可以作为ssDNA存在。这些也可以提供关于ctDNA与cfDNA相比转录因子结合变化的信息,这在癌症的早期检测中可能是重要的。类似地,cfDNA也可以指示疾病应答;以及

[0116] 5.捕获可能临床上相关且包含单链"空位"区域的受损/降解的DNA。

[0117] 分析样品中多种形式的核酸可以通过例如在测序之前对不同形式的核酸差异化地加标签和/或对不同形式的核酸分区来进行。

[0118] II.对样品中的不同核酸形式差异化地加标签

[0119] 核酸样品,诸如体液中的无细胞核酸,通常含有多种形式的核酸,包括单链DNA和双链DNA以及单链RNA。因为这样的样品中核酸的总量可能很低,并且因为具有不同特征和/或修饰的不同形式的核酸可能产生关于样品的不同信息,本文提供了分析2种、3种或所有这样的形式的方法。

[0120] 如果至少一些步骤可以并行进行,则对多种形式的制备和分析会更有效。如果在处理后特定核酸的序列信息可以与样品中核酸的原始形式相关联,则从这样的样品确定的信息是最有益的(informative)。例如,如果处理后在特定核酸中确定了SNV,则可以确定该核酸是来源于原始样品中的RNA、单链DNA还是双链DNA。

[0121] 样品中不同形式的核酸的识别可以通过在所述形式以模糊其原始形式的方式,诸如通过第二链合成或扩增,被改变之前对样品中的不同形式的核酸进行差异化加标签来实现。因此,在包括多种形式的核酸中,至少一种形式连接至核酸标签,以将其与样品中存在的一种或更多种其他形式区分开。在含有三种形式的核酸诸如单链DNA、单链RNA和双链DNA的样品中,这三种形式可以通过对至少两种形式差异化地加标签或通过差异化地标记所有三种形式来区分。连接至相同形式的核酸分子的标签可以彼此相同或不同。但是,如果彼此不同,在一些实施方案中,这些标签可以具有它们的共同编码的一部分,以便将它们所附接的分子识别为特定的形式。例如,特定形式的核酸分子可以带有形式A1、A2、A3、A4等的编码,以及不同形式B1、B2、B3、B4等的编码。这样的编码系统允许区分形式和形式中的分子。图24(下文描述的)提供了用于对具有不同特征,例如使用甲基结合结构域蛋白质确定的甲基化程度的核酸分子差异化地加标签的示例性策略。

[0122] 在用核酸标签对样品中的一种、一些或所有形式的核酸进行差异化标记后,可以扩增这些形式,使得核酸标签与原始样品中的形式一起被扩增。然后扩增的核酸可以经历序列分析,以读取样品中原始核酸的部分或全部序列以及连接的核酸标签的序列。然后标签的序列可以被解码以指示原始样品中核酸的形式。然后可以比较不同形式的序列,以观察遗传变异是主要还是仅在某些形式的核酸中发现,还是独立于原始形式以大约相同的频率出现。对不同形式差异化加标签后的一些或所有步骤,特别是扩增和测序,可以用合并的不同形式的核酸来进行。这样的方法优选地导致样品中存在的两种、三种或更多种形式的核酸分子的至少40%、50%、60%、70%、80%、90%或95%的扩增和测序。

[0123] 双链核酸可以通过连接到至少部分双链的衔接子来差异化地标记。通常,双链核酸在两端连接到这样的衔接子。这样的衔接子中的任一种或两种可以包括核酸标签。如果

两种各自具有标签的衔接子连接至核酸的相应末端，则标签组合可以用作标识物。单链DNA或RNA分子不会在很大程度上连接到衔接子的双链末端，并且因此不会接收核酸标签。双链衔接子可以是完全双链的或部分双链的，如Y形衔接子或发夹状衔接子的情况。示例性的Y形衔接子的序列在下文示出。

[0124] 通用衔接子

[0125] SEQ ID No.1

[0126]

5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3' (SEQ ID NO:1)

[0127] 衔接子标签

[0128] SEQ ID No.2:

[0129]

5'GATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG-3' (SEQ ID NO:2)

[0130] 这些衔接子序列的截短形式已经由Rohland等人，Genome Res.2012年5月;22(5):939-946描述。

[0131] 因为Y形衔接子具有单链末端，如果将要执行后续步骤分离单链样品核酸和其他样品核酸，可能需要避免(例如，通过用不结合Y形衔接子的探针来分离单链DNA)或保护Y形衔接子。

[0132] RNA分子可以用核酸标签来差异化地标记，因为它们是样品中唯一可以被RNA依赖性DNA聚合酶的逆转录酶作用的分子形式。核酸标签可以作为用于引发逆转录的引物的5'标签引入。逆转录可以是随机的或序列特异性的。在逆转录后，原始RNA链可以降解，然后合成第二互补DNA链。如果需要，现在双链的DNA可以被平端化，并且以与样品中已经存在的双链DNA分子相似的方式连接到衔接子。可替代地，RNA/DNA杂交分子可以直接连接到衔接子。

[0133] 单链DNA分子可以通过用分子内连接酶处理来从双链DNA分子分级分离。在一些实施方案中，分子内连接酶是CircLigase™ ssDNA连接酶，用于用3'标签对ssDNA差异化地加标签。在用分子内连接酶处理之前，ssDNA在5'末端去磷酸化，以防止ssDNA环化。在一种情况下，用于将标签附接至单链DNA的连接酶是CircLigase™ ssDNA连接酶。CircLigase™ ssDNA连接酶是一种热稳定的ATP依赖性连接酶。第二链合成可以通过若干机制发生，包括(例如，用T4RNA连接酶在一端将单链DNA连接至提供引物结合位点的寡核苷酸，将单链DNA与互补寡核苷酸杂交，互补寡核苷酸用作基于它们杂交的模板序列延伸的引物，或者与随机寡核苷酸杂交，随机寡核苷酸同样用作基于它们杂交的模板序列延伸的引物。一种方法使用单链连接酶以将具有可延伸3'末端的寡核苷酸附加到单链DNA文库成员(参见Gansauge&Meyer,Nature Protocols 8,737(2013))。使用衔接子作为引物结合位点填充第二DNA链。然后进行5'DNA磷酸化步骤和标准(dsDNA)连接以将衔接子添加至文库分子的5'末端。

[0134] 在另一种方法中，商业上可获得的NEBDirect方法的步骤可以被包括在所述方法中，将单链DNA分子与序列特异性引物杂交，以用于第二链合成，随后进行末端修复，并连接

到侧翼衔接子(参见neb.com/nebnext-direct/nebnext-direct-for-target-enrichment)。含有引物的第二DNA链被降解,因此其未被测序。另一种方法使用在其5'末端具有衔接子序列和在3'末端具有随机碱基的随机引物。通常存在6个随机碱基,但长度可以在4个和9个碱基之间。该方法特别适于RNA-seq或亚硫酸氢盐测序的低输入/单细胞扩增(Smallwood等人,Nat.Methods 2014年8月;11(8):817-820)。

[0135] 由于省略了杂交前的标准变性步骤,ssDNA可以被核酸(NA)探针选择性地捕获。ssDNA-探针杂交体可以通过常规方法(例如生物素化的DNA/RNA探针,由链霉亲和素-珠磁体捕获)从cfNA(无细胞核酸)群体分离。探针序列可以是靶特异性的,并且与具有dsDNA工作流程的组、该工作流程的子集相同,或者不同(例如,靶向外显子-外显子连接处的RNA融合、"热点"DNA序列)。所有单链核酸(ssNA)可以在该步骤中通过利用具有"通用核苷酸碱基"诸如脱氧肌苷、3-硝基吡咯和5-硝基吲哚的探针以序列不可知的方式被捕获。

[0136] 图1示出了用于分离核酸形式的示例性方案。图的上部部分示出了包括双链DNA、单链DNA和单链RNA的样品。RNA用带有5'RNA识别核酸标签的序列特异性的或随机的多聚T引物逆转录。在合成互补DNA链后,通过选择性杂交,用RNA酶H或NaOH或核糖体耗尽来降解RNA模板。然后在不使样品变性的情况下,用捕获探针(其可以是序列特异性的或序列不可知的)处理样品。这些探针与单链分子杂交,从样品去除单链分子。然后,在该实例中,样品中的双链DNA分子被平端化并且连接至包括核酸标签的衔接子。在该实例中,衔接子是Y形的,并且Y的双链臂部分连接到DNA分子。同时,分离的单链核酸通过如上文所讨论的包括标签附接的DNA方案或NEBdirect方案来处理。

[0137] 图2示出了另外的示例性方案,以包括双链DNA、单链DNA和单链RNA的样品起始,具有简化的工作流程,最显著的是避免了5'DNA磷酸化步骤。样品中的双链DNA首先连接到包括核酸标签的发夹状衔接子。然后样品被5'DNA去磷酸化,并且然后RNA被转化为cDNA并且还连接到不同的标签。然后,与图1类似地处理单链DNA。在一些实施方案中,发夹状衔接子可以在文库扩增之前被裂解成两条链。

[0138] 图7图示出了差异化加标签的一个实施方案。在步骤701中,获得核酸群体。核酸可以是循环核酸(cNA),诸如来自液体活检样品(血清、血浆或血液)。在步骤702中,对第一形式的核酸差异化地加标签,以形成第一、加标签的核酸形式和第二、未加标签的核酸形式的混合物(703)。随后,在步骤704中,第二形式的核酸(或残留核酸)用不同的标签加标签。在步骤704之前,上文的方法可以包括两个或更多个不同的差异化加标签的步骤(702)。在对群体中的两种或更多种形式的核酸加标签之后,在一些实施方案中,不同的形式可以被分区。如果将不同的形式分区,则可以在测序之前将差异化加标签的核酸合并在一起或分开测序。优选地,对不同形式的核酸差异化加标签发生在一个管或反应体积中,并且对加标签的分子整体测序(而不分区)。从测序获得的读段可以用于将对来源于不同核酸形式以及共同核酸样品的读段进行的分析。

[0139] 在一些实施方案中,被差异化加标签的第一形式的核酸是dsDNA,并且差异化加标签通过附接至包含第一组标签的dsDNA双链衔接子来进行。然后,ssDNA(残留核酸)用不同的一组标签(第二组标签)来加标签。

[0140] 在一些实施方案中,被差异化加标签的第一形式的核酸是来自开放染色质区域的DNA,并且加标签通过使核酸群体与Tn5介导的转座酶活性接触来进行。

[0141] 在一些实施方案中，被差异化加标签的第一形式的核酸是双链核酸，并且加标签通过将发夹状衔接子附接至双链核酸来进行。

[0142] III.分区具有不同修饰程度的核酸

[0143] 在本文描述的某些实施方案中，在加标签和测序之前，不同形式的核酸群体可以基于核酸的一个或更多个特征来分区。通过对异质核酸群体分区，人们可以增加稀有信号，例如通过富集在群体的一个级分(或分区)中更普遍的稀有核酸分子。例如，在RNA中存在但在DNA中较少(或没有)的遗传变异可以通过将RNA与DNA分区来检测。类似地，在超甲基化DNA中存在但在低甲基化DNA中较少(或没有)的遗传变异可以通过将样品分区为超甲基化核酸分子和低甲基化核酸分子来更容易地检测。通过分析样品的多个级分，可以对单个分子进行多维分析，并且因此可以获得更大的灵敏度。

[0144] 在一些情况下，异质核酸样品被分区为两个或更多个分区(例如，至少3个、4个、5个、6个或7个分区)。在一些实施方案中，每个分区被差异化地加标签。然后将加标签的分区合并在一起，以用于共同样品制备和/或测序。分区-加标签-合并步骤可以发生多于一次，其中每一轮分区基于不同的特征(本文提供的实例)发生，并且使用区别于其他分区和分区装置的差异化加标签来加标签。

[0145] 可以用于分区的特征的实例包括序列长度、甲基化水平、核小体结合、序列错配、免疫沉淀和/或结合DNA的蛋白质。产生的分区可以包括一种或更多种下列的核酸形式：核糖核酸(RNA)、单链DNA(ssDNA)、双链DNA(dsDNA)、较短的DNA片段和较长的DNA片段。在一些实施方案中，异质的核酸群体被分区为与核小体相关的核酸分子和不含核小体的核酸分子。可替代地或另外地，异质的核酸群体被分区为RNA和DNA。可替代地或另外地，异质的核酸群体可以被分区为单链DNA(ssDNA)和双链DNA(dsDNA)。可替代地或另外地，异质的核酸群体可以被分区为具有一个或更多个表观遗传修饰的核酸和不具有一个或更多个表观遗传修饰的核酸。表观遗传修饰的实例包括甲基化的存在或不存在；甲基化水平、甲基化的类型(5'胞嘧啶)；以及与一种或更多种蛋白质诸如组蛋白的缔合和缔合水平。可替代地或另外地，异质的核酸群体可以基于核酸长度(例如，多达160bp的分子和具有大于160bp的长度的分子)来分区。

[0146] 在一些情况下，每个分区(代表不同的核酸形式)被差异化地标记，并且在测序之前将分区合并在一起。在其他情况下，不同形式被分开地测序。

[0147] 图8图示出了本公开内容的一个实施方案。不同核酸的群体(801)被分区为(802)两个或更多个不同的分区(803a、803b)。每个分区(803a、803b)代表不同的核酸形式。每个分区被区别地加标签(804)。在测序(808)之前，将加标签的核酸合并在一起(807)。读段用计算机模拟分析。标签用于分选来自不同分区的读段。检测遗传变体的分析可以在分区-分区水平以及全核酸群体水平上进行。例如，分析可以包括计算机模拟分析以确定每个分区中核酸的遗传变体，诸如CNV、SNV、插入缺失、融合。在一些情况下，计算机模拟分析可以包括确定染色质结构。例如，序列读段的覆盖度或拷贝数可以用于确定染色质中核小体定位。较高的覆盖度可能与基因组区域中较高的核小体占据相关，而较低的覆盖度可能与较低的核小体占据或核小体耗尽区域(NDR)相关。

[0148] 样品可以包括修饰方面不同的核酸，所述修饰包括对核苷酸的复制后修饰和与一种或更多种蛋白质的通常是非共价的结合。

[0149] 在一个实施方案中，核酸群体是从疑似具有癌症或先前诊断患有癌症的受试者的血清、血浆或血液样品获得的核酸群体。核酸包括具有不同甲基化水平的核酸。甲基化可以由任一种或更多种复制后修饰或转录修饰发生。复制后修饰包括对核苷酸胞嘧啶，特别地5-甲基胞嘧啶、5-羟甲基胞嘧啶、5-甲酰基胞嘧啶和5-羧基胞嘧啶的修饰。

[0150] 核酸的分区通过使核酸与甲基化结合蛋白(“MBP”)的甲基化结合结构域(“MBD”)接触来进行。MBD结合5-甲基胞嘧啶(5mC)。MBD经由生物素连接体与顺磁性珠诸如Dynabeads®M-280链霉亲和素偶联。分区为具有不同甲基化程度的级分可以通过增加NaCl浓度来洗脱级分来进行。

[0151] 一般来说，洗脱是每个分子甲基化位点数目的函数，在增加的盐浓度的情况下，分子具有更多的甲基化洗脱。为了基于甲基化程度将DNA洗脱到不同的群体中，人们可以使用一系列增加NaCl浓度的洗脱缓冲液。盐浓度可以在从约100mM至约2500mM NaCl的范围内。在一个实施方案中，该过程导致三(3)个分区。分子与在第一盐浓度且包括含有甲基结合结构域的分子的溶液接触，该分子可以附接至捕获部分，诸如链霉亲和素。在第一盐浓度时，分子群体将结合MBD，而群体将保持未结合。未结合的群体可以被分离为“低甲基化的”群体。例如，代表低甲基化的DNA形式的第一分区是在低盐浓度，例如160nM保持未结合的分区。代表中等甲基化的DNA的第二分区使用中间盐浓度，例如100mM和2000mM之间的浓度来洗脱。这也从样品分离。代表超甲基化的DNA形式的第三分区使用高盐浓度，例如至少约2000nM来洗脱。

[0152] 每个分区被差异化地加标签。标签可以是包含指示与标签缔合的分子的特征的信息的分子，诸如核酸。例如，分子可以带有样品标签(其将一个样品中的分子与不同样品中的分子区分开)、分区标签(其将一个分区中的分子与不同分区中的分子区分开)或分子标签(其将不同分子彼此区分开)(在加独特和非独特的标签的情形两者下)。在某些实施方案中，标签可以包括一个条形码或条形码的组合。如本文使用的，术语“条形码”指的是具有特定核苷酸序列的核酸分子，或指的是核苷酸序列本身，这取决于上下文。条形码可以具有例如10个和100个之间的核苷酸。根据特定目的的需要，条形码的集合可以具有简并序列，或者可以具有具有特定汉明距离的序列。因此，例如，样品指数、分区指数或分子指数可以包括一个条形码或两个条形码的组合，每个条形码附接至分子的不同末端。

[0153] 标签可以用于标记单个多核苷酸群体分区，以便将一个标签(或多个标签)与特定分区相关联。在一些实施方案中，单个标签可以用于标记特定的分区。在一些实施方案中，多个不同的标签可以用于标记特定的分区。在采用多个不同的标签来标记特定分区的实施方案中，用于标记一个分区的标签组可以容易地与用于标记其他分区的标签组区分开。在一些实施方案中，标签可以具有另外的功能，例如标签可以用于索引样品来源或用作独特的分子标识物(其可以用于通过区分测序错误和突变来改进测序数据的质量)。类似地，在一些实施方案中，标签可以具有另外的功能，例如标签可以用于索引样品来源或用作非独特的分子标识物(其可以用于通过区分测序错误和突变来改进测序数据的质量)。

[0154] 在一个实施方案中，分区加标签包括用样品标签的等价物对每个分区中的分子加标签。在重新组合分区和对分子测序后，样品标签识别来源分区。在另一个实施方案中，不同的分区用例如包括一对条形码的不同的一组分子标签来加标签。以该方式，每个分子条形码指示来源分区，也可用于区分分区内的分子。例如，第一组的35个条形码可以用于对第

一分区中的分子加标签，而第二组的35个条形码可以用于对第二分区中的分子加标签。

[0155] 虽然标签可以附接至已经基于一个或更多个特征分区的分子，但是文库中最终加标签的分子可能不再具有该特征。例如，虽然单链DNA分子可以被分区和加标签，但文库中的最终加标签的分子可能是双链的。类似地，虽然RNA可能经历分区，但在最终的文库中，来源于这些RNA分子的加标签的分子可能是DNA。因此，附接至文库中的分子的标签通常指示最终的加标签的分子来源的“亲本分子”的特征，而不一定指示加标签的分子本身的特征。

[0156] 例如，条形码1、2、3、4等用于对第一分区中的分子加标签和标记；条形码A、B、C、D等用于对第二分区中的分子加标签和标记；并且条形码a、b、c、d等用于对第三分区中的分子加标签和标记。差异化加标签的分区可以在测序之前被合并。差异化加标签的分区可以分开地测序或一起同时测序，例如在Illumina测序仪的同一流动池中。

[0157] 在测序之后，分析读段，检测遗传变体可以在分区-分区水平以及全核酸群体水平上进行。标签用于分选来自不同分区的读段。分析可以包括使用序列信息、基因组坐标长度和覆盖度或拷贝数确定遗传变体和染色质结构的计算机模拟分析。较高的覆盖度可能与基因组区域中较高的核小体占据相关，而较低的覆盖度可能与较低的核小体占据或核小体耗尽区域(NDR)相关。

[0158] 在一些实施方案中，原始群体中的核酸可以是单链的和/或双链的DNA和/或RNA。基于单链型与双链型的分区可以通过例如使用对ssDNA分区的标记的捕获探针和使用分区dsDNA的双链衔接子来实现。基于RNA与DNA组成的分区包括但不限于使用分区dsDNA的双链衔接子，和使用分区RNA的带有或不带有捕获探针的逆转录。

[0159] 亲和剂可以是具有期望的特异性的抗体、天然结合配偶体或其变体(Bock等人，Nat Biotech 28:1106-1114(2010)；Song等人，Nat Biotech 29,68-72(2011))，或例如通过噬菌体展示选择的对给定靶具有特异性的人工肽。

[0160] 本文设想的捕获部分的实例包括甲基结合结构域(MBD)和甲基结合蛋白质(MBP)。本文设想的MBP的实例包括，但不限于：

[0161] (a)相比于结合未修饰的胞嘧啶，MeCP2是优先结合5-甲基-胞嘧啶的蛋白质。

[0162] (b)相比于结合未修饰的胞嘧啶，RPL26、PRP8和DNA错配修复蛋白质MHS6优先结合5-羟甲基-胞嘧啶。

[0163] (c)相比于结合未修饰的胞嘧啶，FOXK1、FOXK2、FOXP1、FOXP4和FOXI3优先结合5-甲酰基-胞嘧啶(Iurlaro等人，Genome Biol.14,R119(2013))。

[0164] (d)对一个或更多个甲基化核苷酸碱基特异的抗体。

[0165] 同样地，可以使用组蛋白结合蛋白进行对不同形式的核酸的分区，这可以分离与组蛋白结合的核酸和游离核酸或未结合的核酸。可以用于本文公开的方法的组蛋白结合蛋白的实例包括RBBP4、RbAp48和SANT结构域肽。

[0166] 对于某些亲和剂和修饰，尽管取决于核酸是否带有修饰，其与剂的结合可以以基本上全部的方式发生或完全不发生，但是分离可以是一定程度的。在这样的情况下，修饰方面呈现过度的核酸与剂以比修饰方面呈现不足的核酸更大的程度结合。可替代地，具有修饰的核酸可以以完全的方式结合或完全不结合。但是，不同的修饰水平可以从结合剂顺序洗脱。

[0167] 例如，在一些实施方案中，分区可以是二元的或者基于修饰的程度/水平。例如，所

有甲基化片段可以使用甲基结合结构域蛋白(例如MethylMinder甲基化DNA富集试剂盒(ThermoFisher Scientific))从未甲基化的片段分区。随后,另外的分区可以涉及通过调节具有甲基结合结构域和结合片段的溶液中的盐浓度来洗脱具有不同甲基化水平的片段。随着盐浓度增加,具有更大甲基化水平的片段被洗脱。

[0168] 在一些情况下,最终的分区代表具有不同修饰程度(修饰的过度呈现或不足呈现)的核酸。过度呈现和不足呈现可以由核酸带有的修饰的数目相对于群体中每条链的修饰的中位数来定义。例如,如果样品中的核酸中5-甲基胞嘧啶残基的中位数为2,则包含多于两个5-甲基胞嘧啶残基的核酸在该修饰中是过度呈现的,而具有1个或0个5-甲基胞嘧啶残基的核酸是不足呈现的。亲和分离的效果是在结合相中富集在修饰中过度呈现的核酸和在未结合相中(即溶液中)富集在修饰中不足呈现的核酸。结合相中的核酸可以在后续处理之前洗脱。

[0169] 当使用MethylMiner甲基化DNA富集试剂盒(ThermoFisher Scientific)时,可以使用顺序洗脱来分区不同的甲基化水平。例如,通过使核酸群体与来自试剂盒的附接至磁珠的MBD接触,可以将低甲基化分区(无甲基化)与甲基化分区分开。珠用于从非甲基化核酸中分离出甲基化核酸。随后,顺序进行一个或更多个洗脱步骤,以洗脱具有不同甲基化水平的核酸。例如,第一组甲基化核酸可以在160mM或更高,例如至少200mM、300mM、400mM、500mM、600mM、700mM、800mM、900mM、1000mM或2000mM的盐浓度洗脱。在这样的甲基化核酸被洗脱后,磁性分离再次用于分离较高水平的甲基化核酸和具有较低甲基化水平的核酸。洗脱和磁性分离步骤本身可以重复进行以产生各种分区,诸如低甲基化分区(代表无甲基化)、甲基化分区(代表低甲基化水平)和超甲基化分区(代表高甲基化水平)。

[0170] 在一些方法中,与用于亲和分离的剂结合的核酸经历洗涤步骤。洗涤步骤洗去与亲和剂弱结合的核酸。这样的核酸可以在具有接近平均值或中值(即,在样品与剂初始接触时保持与固相结合的核酸和不与固相结合的核酸之间的中间值)的程度的修饰的核酸中富集。

[0171] 亲和分离导致具有不同修饰程度的核酸的至少两个,和有时三个或更多个分区。尽管分区仍然是分开的,但至少一个分区和通常两个或三个(或更多个)分区的核酸连接至核酸标签,核酸标签通常作为衔接子的组分被提供,其中不同分区中的核酸接收将一个分区的成员与另一个分区的成员区分开的不同的标签。连接至同一分区的核酸分子的标签可以彼此相同或不同。但是,如果彼此不同,标签可以具有它们的共同编码的一部分,以便将它们所附接的分子识别为特定分区。

[0172] 图3示出了示例性方案。样品包括具有不同甲基化程度的核酸,其中一些也具有遗传变异。样品与连接至亲和试剂的磁珠接触,相比于结合胞嘧啶,该亲和试剂优先结合5-甲基胞嘧啶。亲和纯化导致核酸的两个分区。图左侧的分区代表与亲和试剂结合的核酸,并且富集5-甲基胞嘧啶过度呈现的核酸。右侧的分区代表不与亲和试剂结合的核酸,并且富集缺乏5-甲基胞嘧啶或5-甲基胞嘧啶呈现不足的核酸。然后将两个分区附接至包括不同核酸标签的Y形衔接子并扩增。然后测定扩增的核酸的序列数据,样品核酸的序列指示遗传变异,并且标签序列指示样品核酸被分区为哪个分区,从而指示修饰的程度。

[0173] 图24提供了MBD分区和加标签方法的说明性实例。在工作流程(1)中,一组分子标签(例如,35x35标签)可以在分区之前应用于整个样品。在该实例中,在分区之后,对于超甲

基化形式和低甲基化形式的每个分区中的分子任选地被扩增，并且然后独立地测序。在工作流程(2)中，样品中的分子例如基于甲基化特征被分区。对每个分区单独地加标签、扩增和测序。在工作流程(3)中，多于一个样品中的每一个样品中的分子经历分区，用分区特异性标签加标签，合并并扩增。然后给每个样品中的分子提供样品标签，以对它们所来源的样品解卷积。

[0174] 在一些实施方案中，核酸分子可以基于与特定蛋白质或其片段结合的核酸分子和不与该特定蛋白质或其片段结合的核酸分子被分级分离成不同的分区。核酸分子可以基于DNA-蛋白质结合来分级分离。蛋白质-DNA复合物可以基于蛋白质的特定性质来分级分离。这样的性质的实例包括各种表位、修饰(例如组蛋白甲基化或乙酰化)或酶促活性。可以结合DNA并用作用于分级分离的基础的蛋白质的实例可以包括但不限于蛋白质A和蛋白质G。任何合适的方法可以用于基于蛋白质结合区域来分级分离核酸分子。用于基于蛋白质结合区域来分级分离核酸分子的方法的实例包括但不限于SDS-PAGE、染色质免疫沉淀(ChIP)、肝素色谱法和不对称场流动分级分离法(AF4)。

[0175] IV.核酸的5-甲基胞嘧啶模式的确定

[0176] 基于亚硫酸氢盐的测序及其变化形式提供了一种确定核酸的甲基化模式的方法。在一些实施方案中，确定甲基化模式包括区分5-甲基胞嘧啶(5mC)与非甲基化胞嘧啶。在一些实施方案中，确定甲基化模式包括区分N^6-甲基腺嘌呤与非甲基化腺嘌呤。在一些实施方案中，确定甲基化模式包括区分5-羟甲基胞嘧啶(5hmC)、5-甲酰基胞嘧啶(5fC)和5-羧基胞嘧啶(5caC)与非甲基化胞嘧啶。亚硫酸氢盐测序的实例包括但不限于氧化亚硫酸氢盐测序(OX-BS-seq)、Tet辅助亚硫酸氢盐测序(TAB-seq)和还原亚硫酸氢盐测序(redBS-seq)。

[0177] 氧化亚硫酸氢盐测序(OX-BS-seq)用于区分5mC和5hmC，通过首先将5hmC转化为5fC，并且然后如先前描述的进行亚硫酸氢盐测序。Tet辅助亚硫酸氢盐测序(TAB-seq)也可以用于区分5mc和5hmC。在TAB-seq中，5hmC受糖基化保护。然后在进行亚硫酸氢盐测序之前，使用Tet酶将5mC转化为5caC，如先前描述的。还原亚硫酸氢盐测序用于区分5fC与修饰的胞嘧啶。

[0178] 通常，在亚硫酸氢盐测序中，核酸样品被分成两个等分试样，并且一个等分试样用亚硫酸氢盐处理。亚硫酸氢盐将天然胞嘧啶和某些修饰的胞嘧啶核苷酸(例如5-甲酰基胞嘧啶或5-羧基胞嘧啶)转化为尿嘧啶，而其他修饰的胞嘧啶(例如5-甲基胞嘧啶、5-羟甲基胞嘧啶)不被转化。来自两个等分试样的分子的核酸序列的比较指示哪些胞嘧啶被转化为尿嘧啶，而哪些没有被转化为尿嘧啶。因此，可以确定被修饰的和未被修饰的胞嘧啶。最初将样品分成两个等分试样对于仅含有少量核酸和/或包括异质细胞/组织来源诸如含有无细胞DNA的体液的样品是不利的。

[0179] 本公开内容提供了允许亚硫酸氢盐测序及其变化形式的方法。这些方法通过将群体中的核酸连接至捕获部分，即可以被捕获或固定的标记物来工作。捕获部分包括但不限于生物素、亲和素、链霉亲和素、包含特定核苷酸序列的核酸、抗体识别的半抗原和可磁性吸引的颗粒。提取部分可以是结合对的成员，诸如生物素/链霉亲和素或半抗原/抗体。在一些实施方案中，附接至分析物的捕获部分被附接至可分离部分的其结合对捕获，所述可分离部分诸如可磁性吸引的颗粒或可以通过离心沉降的大颗粒。捕获部分可以是允许带有捕获部分的核酸与缺乏捕获部分的核酸亲和分离的任何类型的分子。示例性的捕获部分是生

物素，其允许通过结合连接至或可连接至固相的链霉亲和素或寡核苷酸来亲和分离，这允许通过结合连接至或可连接至固相的互补寡核苷酸来亲和分离。在捕获部分与样品核酸连接后，样品核酸用作用于扩增的模板。在扩增后，原始模板保持与捕获部分连接，但扩增子不与捕获部分连接。

[0180] 捕获部分可以作为衔接子的组分连接至样品核酸，衔接子也可以提供扩增和/或测序引物结合位点。在一些方法中，样品核酸在两个末端处连接至衔接子，其中两个衔接子带有捕获部分。优选地，衔接子中的任何胞嘧啶残基被修饰，诸如被5-甲基胞嘧啶修饰，以保护免受亚硫酸氢盐的作用。在一些情况下，捕获部分通过可裂解的连接体(例如，可光裂解的脱硫生物素-TEG或被USER™酶可裂解的尿嘧啶残基，Chem.Commun.(Camb).2015年2月21日;51(15):3266-3269)连接至原始模板，在这种情况下，如果需要，可以去除捕获部分。

[0181] 将扩增子变性并与用于捕获标签的亲和试剂接触。原始模板结合亲和试剂，而扩增产生的核酸分子不结合。因此，原始模板可以与扩增产生的核酸分子分离。

[0182] 在分离或分区后，相应的核酸群体(即原始模板和扩增产物)可以经历亚硫酸氢盐处理，其中原始模板群体接收亚硫酸氢盐处理，而扩增产物不接收亚硫酸氢盐处理。可替代地，扩增产物可以经历亚硫酸氢盐处理，而原始模板群体不经历亚硫酸氢盐处理。在这样的处理后，相应的群体可以被扩增(在原始模板群体的情况下，这将尿嘧啶转化为胸腺嘧啶)。群体也可以经历生物素探针杂交以用于富集。然后分析相应的群体并比较序列以确定哪些胞嘧啶在最初是5-甲基化的(或5-羟甲基化的)。检测模板群体中的T核苷酸(对应于被转化为尿嘧啶的未甲基化的胞嘧啶)和在扩增的群体的相应位置处的C核苷酸指示未修饰的C。在原始模板和扩增的群体的相应位置处C的存在指示原始样品中修饰的C。

[0183] 在一些实施方案中，一种方法使用顺序DNA-seq和亚硫酸氢盐-seq(BIS-seq)NGS文库制备加分子标签的DNA文库(参见图4)。该过程通过标记衔接子(例如生物素)、全文库的DNA-seq扩增、亲本分子回收(例如链霉亲和素珠下拉)、亚硫酸氢盐转化和BIS-seq来进行。在一些实施方案中，所述方法通过有和没有亚硫酸氢盐处理的亲本文库分子的顺序NGS准备的扩增，以单碱基分辨识别5-甲基胞嘧啶。这可以通过用在两条衔接子链之一上的标记物(例如生物素)修饰在BIS-seq中使用的5-甲基化的NGS衔接子(定向衔接子;Y形/叉形，用5-甲基胞嘧啶替代)来实现。样品DNA分子是连接的衔接子，并被扩增(例如，通过PCR)。由于仅亲本分子将具有标记的衔接子末端，因此它们可以通过标记特异性捕获方法(例如链霉亲和素磁珠)从其扩增的子代选择性地回收。由于亲本分子保留5-甲基化标志物，捕获文库上的亚硫酸氢盐转化将在BIS-seq后产生单碱基分辨率5-甲基化状态，将分子信息保留到相应的DNA-seq。在一些实施方案中，亚硫酸氢盐处理的文库可以在通过在标准多重NGS工作流程中添加样品标签DNA序列来富集/NGS之前与未处理的文库组合。与BIS-seq工作流程一样，生物信息学分析可以针对基因组比对和5-甲基化碱基识别来进行。总之，该方法提供了在文库扩增后选择性地回收携带5-甲基胞嘧啶标志物的亲本、连接的分子的能力，从而允许亚硫酸氢盐转化的DNA的并行处理。这克服了亚硫酸氢盐处理对从工作流程提取的DNA-seq信息的质量/灵敏度的破坏性质。用该方法，回收的连接的、亲本DNA分子(经由标记的衔接子)允许扩增完整的DNA文库，并且并行应用引起表观遗传DNA修饰的处理。本公开内容讨论了使用BIS-seq方法识别胞嘧啶5-甲基化(5-甲基胞嘧啶)，但这不应该是限制性的。BIS-seq的变化形式已经被开发出来以识别羟甲基化胞嘧啶(5hmC;OX-BS-seq、TAB-seq)、

甲酰基胞嘧啶(5fC;redBS-seq)和羧基胞嘧啶。这些方法可以用本文描述的顺序/并行文库制备来实现。

[0184] 修饰的核酸分析的可替代的方法

[0185] 本公开内容提供了用于分析修饰的核酸(例如甲基化的、与组蛋白连接的以及上文讨论的其他修饰)的可替代的方法。在一些这样的方法中,取决于修饰的程度,在对带有不同程度修饰(例如,每个核酸分子0个、1个、2个、3个、4个、5个或更多个甲基基团)的核酸群体进行分级分离之前,使该群体与衔接子接触。衔接子附接至群体中的核酸分子的一个末端或两个末端。优选地,衔接子包括足够数目的不同标签,使得标签组合的数目导致具有相同起点和终点的两个核酸接收相同标签组合的概率较低,例如95%、99%或99.9%。在附接衔接子后,核酸从结合衔接子内的引物结合位点的引物扩增。衔接子,无论是否带有相同或不同的标签,都可以包括相同或不同的引物结合位点,但优选地衔接子包括相同的引物结合位点。在扩增后,核酸与优选地结合带有修饰的核酸的剂(诸如先前描述的这样的剂)接触。核酸被分成至少两个分区,至少两个分区的差异在于带有修饰的核酸对剂的结合程度不同。例如,如果剂对带有修饰的核酸具有亲和力,则在修饰方面过度呈现的核酸(与群体中的中值代表相比)优先结合剂,而对于修饰呈现不足的核酸不结合剂或更容易从剂洗脱。在分离后,不同的分区然后可以经历另外的处理步骤,这通常包括并行但单独的另外的扩增和序列分析。然后可以比较来自不同分区的序列数据。

[0186] 用于进行这样的分离的示例性方案在图5中示出。核酸在两个末端与包括引物结合位点和标签的Y形衔接子连接。分子被扩增。扩增的分子然后通过与优先结合5-甲基胞嘧啶的抗体接触来分级分离,以产生两个分区。一个分区包括缺乏甲基化的原始分子和具有损失的甲基化的扩增拷贝。另一个分区包括具有甲基化的原始DNA分子。然后对这两个分区分别进行处理和测序,进一步扩增甲基化分区。然后可以比较两个分区的序列数据。在该实例中,标签不用于区分甲基化DNA和未甲基化DNA,而是用来区分这些分区中的不同的分子,以便人们可以确定具有相同起点和终点的读段是基于相同的还是不同的分子。

[0187] 本公开内容提供了用于分析核酸群体的另外的方法,其中至少一些核酸包括一个或更多个修饰的胞嘧啶残基,诸如5-甲基胞嘧啶和先前描述的任何其他修饰。在这些方法中,核酸群体与包括一个或更多个在5C位置处修饰的胞嘧啶残基诸如5-甲基胞嘧啶的衔接子接触。优选地,这样的衔接子中的所有胞嘧啶残基也被修饰,或者衔接子的引物结合区域中的所有这样的胞嘧啶被修饰。衔接子附接至群体中的核酸分子的两个末端。优选地,衔接子包括足够数目的不同标签,使得标签组合的数目导致具有相同起点和终点的两个核酸接收相同标签组合的概率较低,例如95%、99%或99.9%。这样的衔接子中的引物结合位点可以相同或不同,但优选地相同。在附接衔接子后,核酸从结合衔接子的引物结合位点的引物扩增。扩增的核酸被分成第一等分试样和第二等分试样。在有或没有另外的处理的情况下,测定第一等分试样的序列数据。因此,确定第一等分试样中的分子的序列数据,而不论核酸分子的初始甲基化状态。第二等分试样中的核酸分子用亚硫酸氢盐处理。该处理将未修饰的胞嘧啶转化为尿嘧啶。然后亚硫酸氢盐处理的核酸经历扩增,该扩增由引物引发至连接至核酸的衔接子的原始引物结合位点。现在仅最初连接至衔接子的核酸分子(不同于其扩增产物)是可扩增的,因为这些核酸在衔接子的引物结合位点保留胞嘧啶,而扩增产物失去了这些胞嘧啶残基的甲基化,这些胞嘧啶残基在亚硫酸氢盐处理中已经经历转化为尿嘧

啶。因此，仅群体中的原始分子经历扩增，至少其中一些是甲基化的。在扩增后，这些核酸经历序列分析。比较从第一等分试样和第二等分试样确定的序列可以指示除其他以外，核酸群体中的哪些胞嘧啶经历甲基化。

[0188] 用于该分析的示例性方案在图6中示出。甲基化DNA在两个末端连接至包括引物结合位点和标签的Y形衔接子。衔接子中的胞嘧啶是5-甲基化的。引物的甲基化用于在后续的亚硫酸氢盐步骤中保护引物结合位点。在附接衔接子后，DNA分子被扩增。扩增产物被分成两个等分试样，以用于有亚硫酸氢盐处理和没有亚硫酸氢盐处理的测序。未经历亚硫酸氢盐测序的等分试样可以在有或没有另外的处理的情况下经历序列分析。另一个等分试样用亚硫酸氢盐处理，这将未甲基化的胞嘧啶转化为尿嘧啶。只有受胞嘧啶甲基化保护的引物结合位点，当与对原始引物结合位点特异的引物接触时，可以支持扩增。因此，仅原始分子而不是来自第一扩增的拷贝经历另外的扩增。然后另外的扩增的分子经历序列分析。然后可以比较来自两个等分试样的序列。如图5，衔接子中的核酸标签不用于区分甲基化DNA和未甲基化DNA，而是用于区分同一分区内的核酸分子。

[0189] V.方法的一般特征

[0190] 1.样品

[0191] 样品可以是从受试者分离的任何生物样品。样品可以是身体样品。样品可以包括身体组织，诸如已知或怀疑的实体瘤、全血、血小板、血清、血浆、粪便、红细胞、白血细胞或白细胞、内皮细胞、组织活检、脑脊液、滑液、淋巴液、腹水、组织间隙液或细胞外液、细胞之间空间中的流体，包括龈沟液、骨髓、胸膜渗出物、脑脊液、唾液、粘液、痰、精液、汗液、尿液。样品优选地为体液，特别地血液及其级分，以及尿液。样品可以呈最初从受试者分离出来的形式，或者可以已经经历另外的处理以去除或添加组分，诸如细胞，或相对于另一种组分富集一种组分。因此，用于分析的优选的体液是含有无细胞核酸的血浆或血清。样品可以从受试者分离或获得，并且被运送到样品分析场所。样品可以在期望的温度例如室温、4℃、-20℃和/或-80℃保存和运输。样品可以在样品分析的场所从受试者分离或获得。受试者可以是人、哺乳动物、动物、伴侣动物、服务型动物或宠物。受试者可以具有癌症。受试者可以不具有癌症或可检测的癌症症状。受试者可能已经用一种或更多种癌症疗法，例如化疗、抗体、疫苗或生物制剂中的任一种或更多种治疗。受试者可能处于缓解。受试者可能被诊断为易患癌症或任何癌症相关的遗传突变/障碍，或者可能不被诊断为易患癌症或任何癌症相关的遗传突变/障碍。

[0192] 血浆的体积可以取决于测序区域所需的读段深度。示例性体积为0.4ml-40ml、5ml-20ml、10ml-20ml。例如，体积可以是0.5mL、1mL、5mL、10mL、20mL、30mL或40mL。取样的血浆的体积可以是5mL至20mL。

[0193] 样品可以包含不同量的核酸，该量包括基因组当量。例如，约30ng DNA的样品可以含有约10,000(10^4)个单倍体人类基因组当量，并且在cfDNA的情况下，含有约2,000亿($2x10^{11}$)个单个多核苷酸分子。类似地，约100ng DNA的样品可以含有约30,000个单倍体人类基因组当量，并且在cfDNA的情况下，含有约6,000亿个单个分子。

[0194] 样品可以包含来自不同来源的核酸，例如来自同一受试者的细胞和无细胞的核酸，来自不同受试者的细胞和无细胞的核酸。样品可以包含携带突变的核酸。例如，样品可以包含携带生殖系突变和/或体细胞突变的DNA。生殖系突变指的是存在于受试者的生殖系

DNA中的突变。体细胞突变指的是源自受试者的体细胞例如癌细胞的突变。样品可以包含携带癌症相关突变(例如,癌症相关的体细胞突变)的DNA。样品可以包含表观遗传变体(即化学修饰或蛋白质修饰),其中表观遗传变体与遗传变体诸如癌症相关突变的存在相关。在一些实施方案中,样品包含与遗传变体的存在相关的表观遗传变体,其中样品不包含遗传变体。

[0195] 在扩增之前样品中的无细胞核酸的示例性量在从约1fg至约1μg,例如1pg至200ng、1ng至100ng、10ng至1000ng的范围。例如,无细胞核酸分子的量可以多达约600ng、多达约500ng、多达约400ng、多达约300ng、多达约200ng、多达约100ng、多达约50ng或多达约20ng。无细胞核酸分子的量可以是至少1fg、至少10fg、至少100fg、至少1pg、至少10pg、至少100pg、至少1ng、至少10ng、至少100ng、至少150ng或至少200ng。无细胞核酸分子的量可以多达1飞克(fg)、10fg、100fg、1皮克(pg)、10pg、100pg、1ng、10ng、100ng、150ng或200ng。所述方法可以包括获得1飞克(fg)至200ng。

[0196] 无细胞核酸是不包含在细胞内或不以其他方式与细胞结合的核酸,或者换句话说,是在去除完整细胞后保留在样品中的核酸。无细胞核酸包括DNA、RNA及其杂交体,包括基因组DNA、线粒体DNA、siRNA、miRNA、循环RNA(cRNA)、tRNA、rRNA、核小RNA(snoRNA)、piwi-相互作用RNA(piRNA)、长非编码RNA(长ncRNA)或这些的任一种的片段。无细胞核酸可以是双链的、单链的或其杂交体。无细胞核酸可以通过分泌或细胞死亡过程,例如细胞坏死和凋亡,释放到体液中。一些无细胞核酸从癌细胞,例如循环肿瘤DNA(ctDNA)释放到体液中。其他从健康细胞释放。在一些实施方案中,cfDNA是无细胞胎儿DNA(cffDNA)。在一些实施方案中,无细胞核酸由肿瘤细胞产生。在一些实施方案中,无细胞核酸由肿瘤细胞和非肿瘤细胞的混合物产生。

[0197] 无细胞核酸具有约100个-500个核苷酸的示例性大小分布,110个至约230个核苷酸的分子代表约90%的分子,模式为约168个核苷酸,并且第二个次要峰在240个至440个核苷酸的范围内。

[0198] 无细胞核酸可以通过分级分离或分区步骤从体液分离,在分级分离或分区步骤中,如溶液中发现的,无细胞核酸从完整的细胞和体液的其他不可溶的组分分离。分区可以包括诸如离心或过滤的技术。可替代地,体液中的细胞可以裂解,并且一起处理无细胞核酸和细胞核酸。通常,在添加缓冲液和洗涤步骤后,核酸可以用醇沉淀。可以使用进一步的清洁步骤诸如基于二氧化硅的柱以去除污染物或盐。可以在整个反应中添加非特异性批量载体核酸,诸如C_0t-1DNA,或用于亚硫酸氢盐测序、杂交和/或连接的DNA或蛋白质,以优化该程序的某些方面诸如收率。

[0199] 在这样的处理后,样品可以包含各种形式的核酸,包括双链DNA、单链DNA和单链RNA。在一些实施方案中,单链DNA和RNA可以转化为双链形式,因此它们被包括在后续的处理和分析步骤中。

[0200] 2.将DNA分子连接至衔接子

[0201] 样品中的双链DNA分子和被转化为双链DNA分子的单链RNA或DNA分子可以在一个末端或两个末端处连接至衔接子。通常,双链分子在所有四种标准核苷酸的存在下,通过用具有5'-3'聚合酶和3'-5'核酸外切酶(或校正功能)的聚合酶处理而被平端化。Klenow大片段和T4聚合酶是合适的聚合酶的实例。平端的DNA分子可以与至少部分地双链衔接子(例

如,Y形衔接子或钟形衔接子)连接。可替代地,互补核苷酸可以被添加至样品核酸和衔接子的平端,以便于连接。本文设想的是平端连接和粘端连接两者。在平端连接中,核酸分子和衔接子标签两者都具有平端。在粘端连接中,通常,核酸分子带有“A”突出端,而衔接子带有“T”突出端。

[0202] 3.扩增

[0203] 侧翼为衔接子的样品核酸可以通过PCR和其他扩增方法来扩增。扩增通常是通过引物与待扩增的DNA分子侧翼的衔接子中的引物结合位点结合而引发的。扩增方法可以涉及由热循环导致的变性、退火和延伸的循环,或者可以是等温的,如在转录介导的扩增中。其他扩增方法包括连接酶链式反应、链置换扩增、基于核酸序列的扩增和基于自我维持序列的复制。

[0204] 优选地,本方法用T尾和C尾衔接子进行dsDNA“T/A连接”,这导致在连接至衔接子之前至少50%、60%、70%或80%的双链核酸扩增。优选地,本方法相对于单独用T尾衔接子进行的对照方法,扩增的分子的量或数目增加了至少10%、15%或20%。

[0205] 4.标签

[0206] 包含条形码的标签可以被掺入衔接子或以其他方式连接到衔接子。标签可以通过连接、重叠延伸PCR和其他方法掺入。

[0207] 加分子标签策略

[0208] 加分子标签指的是一种加标签实践,其允许人们区分序列读段所来源的分子。加标签策略可以分为加独特标签策略和加非独特标签策略。在加独特标签中,样品中的所有或基本上所有分子都带有不同的标签,使得可以基于单独的标签信息将读段指定给原始分子。在这样的方法中使用的标签有时被称为“独特标签”。在加非独特标签中,同一样品中的不同分子可以带有相同的标签,使得除了标签信息之外,其他信息用于将序列读段指定给原始分子。这样的信息可以包括起始和终止坐标、分子映射到的坐标、单独的起始或终止坐标等。在这样的方法中使用的标签有时被称为“非独特标签”。因此,没有必要对样品中的每个分子独特地加标签。对样品中落入可识别类别的分子独特地加标签就足够了。因此,不同可识别家族中的分子可以带有相同的标签,而不会丢失关于加标签的分子的身份的信息。

[0209] 在加非独特标签的某些实施方案中,所使用的不同标签的数目可以足以使得特定组的所有分子带有不同标签的可能性非常高(例如,至少99%、至少99.9%、至少99.99%或至少99.999%)。应注意,当条形码用作标签时,以及当条形码被例如随机地附接至分子的两端时,条形码的组合一起可以构成标签。该数目,就术语而言,是落入判别的分子数目的函数。例如,类别可以是所有映射到参考基因组上的相同起始-终止位置的分子。类别可以是跨越特定遗传基因座,例如,特定碱基或特定区域(例如,多达100个碱基或基因或基因外显子)映射的所有分子。在某些实施方案中,用于独特地识别一类中的多个分子z的不同标签的数目可以在2*z、3*z、4*z、5*z、6*z、7*z、8*z、9*z、10*z、11*z、12*z、13*z、14*z、15*z、16*z、17*z、18*z、19*z、20*z或100*z中的任一个(例如,下限)和100,000*z、10,000*z、1000*z或100*z中的任一个(例如,上限)之间。

[0210] 例如,在约5ng至30ng的无细胞DNA的样品中,人们期望约3000个分子映射到特定的核苷酸坐标,并且具有任何起始坐标的约3个和10个之间的分子共享相同的终止坐标。因此,约50个至约50,000个不同的标签(例如,约6个和220个之间的条形码组合)足以独特地

对所有这样的分子加标签。为了独特地对跨一个核苷酸坐标映射的所有3000个分子加标签，将需要约100万至约2000万个不同的标签。

[0211] 通常，反应中独特的标签条形码或非独特的标签条形码的指定遵循由美国专利申请20010053519、20030152490、20110160078和美国专利第6,582,908号和美国专利第7,537,898号和美国专利第9,598,731号描述的方法和系统。标签可以随机或非随机地连接至样品核酸。

[0212] 在一些实施方案中，对加载到微孔板后的加标签的核酸测序。微孔板可以具有96个、384个或1536个微孔。在一些情况下，它们以独特标签与微孔的预期比率引入。例如，可以加载独特标签使得每基因组样品加载多于约1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、20个、50个、100个、500个、1000个、5000个、10000个、50,000个、100,000个、500,000个、1,000,000个、10,000,000个、50,000,000个或1,000,000,000个独特标签。在一些情况下，可以加载独特标签使得每基因组样品加载少于约2个、3个、4个、5个、6个、7个、8个、9个、10个、20个、50个、100个、500个、1000个、5000个、10000个、50,000个、100,000个、500,000个、1,000,000个、10,000,000个、50,000,000个或1,000,000,000个独特标签。在一些情况下，每样品基因组加载的独特标签的平均数目少于或大于每基因组样品的约1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、20个、50个、100个、500个、1000个、5000个、10000个、50,000个、100,000个、500,000个、1,000,000个、10,000,000个、50,000,000个或1,000,000,000个独特标签。

[0213] 一种优选的格式使用连接到靶核酸两端的20个-50个不同的标签条形码。例如，35个不同的标签条形码连接到靶分子的两端，产生35×35排列，对于35个标签条形码，这等于1225。这样的标签的数目是足够的，使得具有相同起点和终点的不同分子具有接收不同标签组合的高概率(例如，至少94％、99.5％、99.99％、99.999％)。其他条形码组合包括10和500之间的任何数字，例如，约15x15、约35x35、约75x75、约100x100、约250x250、约500x500。

[0214] 在一些情况下，独特标签可以是预定的或者是随机的或半随机的序列寡核苷酸。在其他情况下，可以使用多于一个条形码使得条形码在所述多于一个条形码中相对于彼此不必是独特的。在该实例中，条形码可以与个体分子连接，使得条形码和可以与其连接的序列的组合产生可以被单独地追溯的独特序列。如本文描述的，非独特条形码的检测与在序列读段的开始(起始)和结束(终止)部分的序列数据组合可以允许将独特的身份指定至特定分子。个体序列读段的长度或碱基对的数目也可以用于将独特身份指定至这样的分子。如本文描述的，来自已经指定了独特身份的核酸单链的片段可以从而允许随后识别来自亲本链的片段。

[0215] 5.靶富集

[0216] 在某些实施方案中，样品中的核酸可以经历靶富集，其中具有靶序列的分子被捕获以用于后续分析。靶富集可以包括使用诱饵组，该诱饵组包括用捕获部分诸如生物素标记的寡核苷酸诱饵。探针可以具有被选择平铺在一组区域，诸如基因上的序列。在一些实施方案中，对于更具体期望的感兴趣的序列，诱饵组可以具有更高的相对浓度。这样的诱饵组在允许靶分子与诱饵杂交的条件下与样品组合。然后，使用捕获部分分离捕获的分子，例如，基于珠的链霉亲和素的生物素捕获部分。例如，在2017年2月7日提交的USSN 15/426,668(美国专利9,850,523，2017年12月26日公布)中进一步描述了这样的方法。

[0217] 6.测序

[0218] 在进行或不进行先前的扩增的情况下,侧翼为衔接子的样品核酸可以经历测序。测序方法包括例如,Sanger测序、高通量测序、焦磷酸测序、合成测序、单分子测序、纳米孔测序、半导体测序、连接测序、杂交测序、RNA-Seq(Illumina)、数字基因表达(Helicos)、下一代测序(NGS)、单分子合成测序(SMSS)(Helicos)、大规模并行测序、克隆单分子阵列(Solexa)、鸟枪法测序、Ion Torrent、Oxford Nanopore、Roche Genia、Maxim-Gilbert测序、引物步移、使用PacBio、SOLiD、Ion Torrent或Nanopore平台的测序。测序反应可以在多种样品处理单元中进行,多种样品处理单元可以是多通路、多通道、多孔或基本上同时处理多个样品组的其他装置。样品处理单元还可以包括多个样品室,以便能够同时处理多个运行。

[0219] 测序反应可以对一种或更多种形式的核酸进行,其中至少一种已知含有癌症或其他疾病的标志物。测序反应也可以对样品中存在的任何核酸片段进行。测序反应可以为基因组提供至少5%、10%、15%、20%、25%、30%、40%、50%、60%、70%、80%、90%、95%、99%、99.9%或100%的序列覆盖度。在其他情况下,基因组的序列覆盖度可以小于5%、10%、15%、20%、25%、30%、40%、50%、60%、70%、80%、90%、95%、99%、99.9%或100%。序列覆盖度可以对至少5个、10个、20个、70个、100个、200个或500个不同基因进行,或对至多5000个、2500个、1000个、500个或100个不同基因进行。

[0220] 同时测序反应可以使用多重测序进行。在一些情况下,可以用至少1000个、2000个、3000个、4000个、5000个、6000个、7000个、8000个、9000个、10000个、50000个、100,000个测序反应对无细胞核酸测序。在其他情况下,可以用少于1000个、2000个、3000个、4000个、5000个、6000个、7000个、8000个、9000个、10000个、50000个、100,000个测序反应对无细胞核酸测序。测序反应可以顺序地或同时地进行。随后的数据分析可以对所有或部分测序反应进行。在一些情况下,可以对至少1000个、2000个、3000个、4000个、5000个、6000个、7000个、8000个、9000个、10000个、50000个、100,000个测序反应进行数据分析。在其他情况下,可以对少于1000个、2000个、3000个、4000个、5000个、6000个、7000个、8000个、9000个、10000个、50000个、100,000个测序反应进行数据分析。示例性读段深度是每基因座(碱基)1000个-50000个读段。

[0221] 7.分析

[0222] 本方法可以用于诊断受试者中状况特别是癌症的存在,以表征状况(例如,对癌症分期或确定癌症的异质性),监测对状况的治疗的响应,产生发展状况或状况后续进程的预后风险。本公开内容还可以用于确定特定治疗选项的功效。如果治疗成功,则成功的治疗选项可以增加在受试者的血液中检测到的拷贝数变异或稀有突变的量,因为更多的癌症可能死亡并使DNA脱落。在其他实例中,这可能不会发生。在另一个实例中,也许某些治疗选项可能与癌症随时间推移的遗传特征谱相关。这种相关性可以用于选择疗法。另外,如果观察到癌症在治疗之后缓解,则本方法可以用于监测剩余的疾病或疾病的复发。

[0223] 可以被检测到的癌症的类型和数目可以包括血癌、脑癌、肺癌、皮肤癌、鼻癌、喉癌、肝癌、骨癌、淋巴瘤、胰腺癌、皮肤癌、肠癌、直肠癌、甲状腺癌、膀胱癌、肾癌、口腔癌、胃癌、实体瘤(solid state tumors)、异质肿瘤、均质肿瘤等。癌症的类型和/或阶段可以由遗传变异检测到,所述遗传变异包括突变、稀有突变、插入缺失、拷贝数变异、颠换、易位、倒

位、缺失、非整倍性、部分非整倍性、多倍性、染色体不稳定性、染色体结构改变、基因融合、染色体融合、基因截短、基因扩增、基因复制、染色体损伤、DNA损伤、核酸化学修饰的异常改变、表观遗传模式的异常改变、以及核酸5-甲基胞嘧啶的异常改变。

[0224] 遗传数据还可以用于表征特定形式的癌症。癌症在组成和分期两方面经常是异质的。遗传特征谱数据可以允许表征癌症的具体亚型，该表征在该具体亚型的诊断或治疗中可能是重要的。该信息还可以为受试者或从业者提供关于癌症的具体类型的预后的线索，并且允许受试者或从业者根据疾病的进展调节治疗选项。一些癌症可以进展变成更具侵袭性和遗传上不稳定的。其他癌症可以保持为良性的、非活动的、或休眠的。本公开内容的系统和方法可以用于确定疾病进展。

[0225] 本分析还可用于确定特定治疗选项的功效。如果治疗成功，则成功的治疗选项可以增加在受试者的血液中检测到的拷贝数变异或稀有突变的量，因为更多的癌症可能死亡并使DNA脱落。在其他实例中，这可能不会发生。在另一个实例中，也许某些治疗选项可能与癌症随时间推移的遗传特征谱相关。这种相关性可以用于选择疗法。另外，如果观察到癌症在治疗之后缓解，则本方法可以用于监测剩余的疾病或疾病的复发。

[0226] 本方法还可以用于检测除癌症以外的状况中的遗传变异。当存在某些疾病后，免疫细胞，诸如B细胞，可以经历快速克隆扩增。克隆扩增可以使用拷贝数变异检测来监测并且可以监测某些免疫状态。在该实例中，拷贝数变异分析可以随时间进行，以产生特定疾病可能如何进展的谱。拷贝数变异或甚至稀有突变检测可以用于确定病原体群体在感染过程期间是如何改变的。这在慢性感染诸如HIV/AIDS或肝炎感染期间可能特别重要，病毒可以藉以在感染过程期间改变生命周期状态和/或突变为毒力更强的形式。当免疫细胞试图破坏移植的组织时，本方法可以用于确定或谱系分析宿主体的排斥活动，以监测移植的组织的状态以及改变治疗过程或预防排斥。

[0227] 此外，本公开内容的方法可以用于表征受试者中的异常状况的异质性。这样的方法可以包括例如生成源自受试者的细胞外多核苷酸的遗传特征谱，其中所述遗传特征谱包括由拷贝数变异和稀有突变分析得到的多于一个数据。在一些实施方案中，异常状况是癌症。在一些实施方案中，异常状况可以是导致异质基因组群体的状况。在癌症的实例中，已知一些肿瘤包含处于癌症的不同阶段的肿瘤细胞。在其他实例中，异质性可以包括多个疾病病灶。再次，在癌症的实例中，可以存在多个肿瘤病灶，或许其中一个或更多个病灶是已从原发部位扩散的转移的结果。

[0228] 本方法可以用于生成为由异质性疾病中的不同细胞得到的遗传信息的总和的指纹图谱或数据集或对其进行特征分析。该数据集可以包含单独的或组合的拷贝数变异和突变分析。

[0229] 本方法可以用于诊断、预后、监测或观察癌症或其他疾病。在一些实施方案中，本文的方法不涉及诊断、预后或监测胎儿，并且因此不涉及非侵入性产前测试。在其他实施方案中，这些方法可以用于妊娠的受试者，以诊断、预后、监测或观察未出生的受试者的癌症或其他疾病，未出生的受试者的DNA和其他多核苷酸可以与母体分子共循环。

[0230] 用于通过NGS对MBD-珠分区文库进行分子标签识别的示例性方法如下：

[0231] 1.使用甲基结合结构域蛋白质-珠纯化试剂盒对提取的DNA样品(例如，从人类样品提取的血浆DNA)进行物理分区，节省了来自该过程的所有洗脱物以用于下游处理。

[0232] 2.将差异化分子标签和支持NGS的衔接子序列并行应用于每个分区。例如,超甲基化分区、残留甲基化(“洗涤”)分区和低甲基化分区与带有分子标签的NGS衔接子连接。

[0233] 3.重新组合所有加分子标签的分区,并且随后使用衔接子特异性DNA引物序列进行扩增。

[0234] 4.重新组合和扩增的总文库的富集/杂交,靶向感兴趣的基因组区域(例如,癌症特异性遗传变体和差异化甲基化区域)。

[0235] 5.重新扩增富集的总DNA文库,附加样品标签。合并不同的样品,并在NGS仪器上进行多重测定。

[0236] 6.用用于识别独特分子的分子标签对NGS数据进行生物信息学分析,以及将样品解卷积成被差异化MBD分区的分子。该分析可以产生关于基因组区域的相对5-甲基胞嘧啶的信息,同时进行标准基因测序/变异检测。

[0237] VI.实践本公开内容的模式

[0238] 本公开内容提供了一种方法,所述方法包括将无细胞核酸(cfNA)群体分区成共享一个或更多个相似特征的分区。

[0239] 可以进行本公开内容的方法来分区单链核酸(ssNA;ssDNA、RNA)和dsDNA,dsDNA分子藉以通过标准文库制备来制备,而ssNA在辅助文库制备工作流程中制备,该工作流程将ssNA转化成适合富集、测序(例如NGS)和分析的形式,同时保留关于来源生物分子类型(即,RNA、ssDNA、dsDNA)的信息。

[0240] 含cfNA的文库制备的方法可以包括(a)将RNA转化为可识别的ssDNA,和(b)将ssDNA和dsDNA分子分区以用于进行并行NGS文库制备,(c)随后(任选的)进行靶富集,(d)NGS和下游数据分析以识别具有序列的分子类型(参见图1)。

[0241] 在一些实施方案中,在RNA分子加标签、特异性连接、cDNA转化和NGS文库制备之前可以进行cfNA群体的dsDNA特异性NGS衔接子连接。如图2中示出的,可以将同时测序方法应用于cfNA样品,其中将dsDNA,然后RNA,顺序连接以用于NGS文库建立,而不进行分区。

[0242] 在一些实施方案中,平台连接使用Y形或“叉形”衔接子,其产生具有ssDNA 5’和3’末端的连接的ds-cfDNA分子。在同时测序或传统的ssDNA文库制备方法中,这些末端可以被RNA连接酶(或Circligase™II)错误地连接。通过将Y形衔接子的末端改变为“发夹”或“泡”,连接的cf-dsDNA分子不再具有ssDNA末端,并且不是用于同时测序/传统DNA文库制备中随后的ssNA连接的底物。因此,将NGS衔接子改造为不含游离ssDNA末端,能够进行除dsDNA工作流程之外的RNA和ssDNA文库制备,而不对分子类型进行分区。

[0243] 本公开内容的方法可以对cfNA群体以逆转录酶进行,使用具有分子标签尾的基因特异性/随机/聚T DNA引物,随后通过RNA酶H或NaOH水解去除RNA,产生加标签的ssDNA(cDNA)来取代每个RNA分子。本领域技术人员已知的另外的方法可以用于去除不需要的RNA序列,诸如通过选择性杂交的核糖体RNA耗尽。

[0244] ssDNA可以通过省略杂交前的标准变性步骤被NA探针选择性地捕获。ssDNA探针杂交体可以通过本领域已知的方法(例如生物素化的DNA/RNA探针,被链霉亲和素珠磁体捕获)从cfNA群体分离。探针序列可以是靶特异性的,并且与具有dsDNA工作流程的组、该工作流程的子集相同,或者不同(例如,靶向外显子-外显子连接处的RNA融合、“热点”DNA序列)。此外,所有ssNA可以在该步骤中通过利用具有“通用核苷酸碱基”诸如脱氧肌苷、3-硝基吡

咯和5-硝基吲哚的探针以序列不可知的方式被捕获。

[0245] 除了由DNA测序识别的遗传变异诸如SNV、插入缺失、基因融合和CNV之外，表观遗传变异(诸如5-甲基胞嘧啶、组蛋白甲基化、核小体定位以及微小非编码RNA表达和长非编码RNA表达)可以导致疾病进展诸如癌症或参与疾病进展诸如癌症。表观遗传标志物的高通量测量需要复杂的分子生物学技术，专门针对每种类型的表观遗传标志物来开发。因此，表观遗传测序项目通常与DNA(遗传)测序并行，并且需要大量的输入。换句话说，多分析物生物标志物检测伴随样品破坏。

[0246] 无细胞DNA的遗传(DNA)测序和表观遗传测序两者对非侵入性产前测试(NIPT)和癌症监测/检测具有诊断价值。在这两种应用中，遗传物质的量是有限的并且识别稀有的分子事件是最重要的。因此，利用当前的方法，进行表观遗传测序导致检测遗传变体的灵敏度降低，因为每种类型的标志物都需要一个专用样品。

[0247] 本公开内容提供了获得关于DNA 5-甲基胞嘧啶的表观遗传过程的信息的方法，但是针对5-甲基胞嘧啶概述的“用分子标签分区”的方法也可以应用于其他表观遗传机制。类似地，如本公开内容中关于5-甲基胞嘧啶(5mC)识别概述的，标记和回收NGS衔接子连接的亲本DNA分子，也可以用于识别其他表观遗传DNA修饰标志物(例如分别羟甲基化、甲酰基化和羧基化的5hmC、5fC和5caC)。

[0248] 关于5-甲基胞嘧啶，亚硫酸氢盐测序是最流行的方法，能够以单碱基分辨率解析5-甲基胞嘧啶碱基。该方法包括对所有胞嘧啶碱基起作用的化学处理(亚硫酸氢盐)，将它们转化为尿嘧啶，除非它们是5-甲基化的或5-羟甲基化的。亚硫酸氢盐处理后的测序将导致5-甲基化胞嘧啶和5-羟甲基化胞嘧啶残基被检测为胞嘧啶，而未甲基化的胞嘧啶、5-甲酰基甲基化胞嘧啶和5-羧甲基化胞嘧啶被检测为胸腺嘧啶。先前描述的亚硫酸氢盐测序的变化可以进一步区分5mC、5hmC、5fC和5caC。该方法的主要局限是大部分遗传物质丢失了。严苛的亚硫酸氢盐处理降解了<99％的输入DNA，从而降低样品的分子复杂性和可达到的检测限值。目前的分子生物学DNA扩增技术(例如PCR、LAMP、RCA)对胞嘧啶的5-甲基化状态是不可知的，并且因此，5-甲基化标志物随扩增丢失。这在液体活检应用中是非常不期望的。此外，用亚硫酸氢盐转化的DNA文库检测体细胞变体变得更具挑战性(例如，将C->T SNV与未甲基化的胞嘧啶区分开)。因此，亚硫酸氢盐处理的DNA不用于液体活检应用中的遗传变体检测。进行5-甲基胞嘧啶分析和对DNA的遗传变体判别需要对样品进行分隔，这降低了每个工作流程中检测的输入/灵敏度，并且防止在单个分子上识别5-甲基胞嘧啶信息和遗传变体。

[0249] 在某些实施方案中，核酸基于甲基化差异来分区。核酸的“超甲基化”和“低甲基化”形式可以被定义为分别高于和低于特定甲基化程度的分子，所述特定甲基化程度由所使用的特定分区方法区分。例如，分区方法可以选择具有至少2个、至少3个、至少4个、至少5个或至少6个甲基化核苷酸的分子。甲基化程度指的是核酸片段中甲基化核苷酸的数目。通过捕获与甲基结合结构域(MBD)蛋白或其片段或变体结合的分子，可以实现对在DNA样品中相对“超甲基化”的DNA分子的识别。MBD也可以被称为甲基-CpG结合结构域。MBD蛋白可以与磁珠复合。在一些实施方案中，与MBD结合的蛋白质是MECP2、MBD1、MBD2、MBD3、MBD4或其片段或变体。尽管该方法没有直接指出5-甲基化位点(没有亚硫酸氢盐转化)，但是对重叠的超甲基化片段的生物信息学分析可以解析5-甲基胞嘧啶的特定位点。该方法的主要缺点

是,通过仅对超甲基化分区测序,大多数未甲基化(按重量计～80%-97%)的人类基因组未被测序,这阻止/限制了遗传变体(例如,SNV、插入缺失和CNV)的识别,因为这些是低覆盖度区域或根本不存在于超甲基化分区中。

[0250] 本公开内容提供了用于获得5-甲基胞嘧啶数据和用于获得测序数据的方法,所述测序数据用于检测同一低输入样品(例如,液体活检工作流程)中的稀有遗传变体。例如,包括MBD分级分离和加标签的方法对样品中的核酸是非破坏性的,并且在扩增后保持基因组的复杂性。此外,分级分离-加标签方法(例如,MBD分级分离和加标签)可以重新组合差异化分区的核酸分子,以确保基因组复杂性的保持,并使得能够进行多分析物生物标志物检测(遗传变体和表观遗传变体)。相比之下,其他方法可能对样品中的核酸分子是破坏性的。这些其他方法可以包括亚硫酸氢盐测序、甲基敏感限制性内切酶消化以及在仅分析一种级分或一组核酸分子(例如,超甲基化核酸分子)的情况下的MBD富集。例如,亚硫酸氢盐测序对核酸分子产生物理损伤。甲基敏感限制性内切酶消化通过破坏未甲基化的级分,仅留下完整的甲基化核酸,降低了基因组的复杂性。在仅分析MBD结合的核酸分子的情况下,MBD富集可以类似地用于仅分离样品中的核酸的单一级分。仅分析核酸分子的单一级分的方法破坏了关于存在于非富集部分的核酸分子的信息。

[0251] 本文提供的获得5-甲基胞嘧啶数据(或其他甲基化状态数据)的方法可以与上文描述的用于获得单链核酸和双链核酸信息的方法组合实践。在一些实施方案中,本文的方法通过对已经被MBD珠分区成不同甲基化程度的DNA分子差异化加标签来定量超甲基化DNA的%。(参见图3)。在该方法中,可以回收来自MBD分区方案的所有洗脱物,并且用对应于它们的MBD分区的不同组的分子标签制备NGS文库。因此,MBD分区过程减少了典型亚硫酸氢盐处理中存在的物质损失。由于连接的分区可以在扩增/富集/NGS之前重新组合,因此DNA测序工作流程存在最小的缺陷。MBD结合双链DNA(dsDNA),并且因此,MBD分区保留了样品DNA的双链性质,允许通过灵敏的DNA测序方法进行双链分子加标签。

[0252] 在MBD分区的分子标签NGS工作流程中,分子标签可以用于两个目的—从样品识别独特的DNA分子(通过标签和基因组起始/终止坐标的组合),以及指示分子的相对5-甲基胞嘧啶水平。分子标签可以用于识别和计数独特的核酸分子。该信息可以用于计算扩增不平衡。分子标签可以允许待辨别的样品的原始复杂性。即使存在不均匀扩增,加分子标签可以用于识别和计数样品中的核酸分子。

[0253] 上文的方法描述了按照5-甲基胞嘧啶的程度进行的物理分区、差异化分子标签的应用、任选的文库重新组合、富集、NGS和每个分子来源分区的生物信息学解卷积,与用于遗传测序/变体检测的DNA-seq同时进行。该方法可扩展到通过用保持DNA分子的双链性质的不同的DNA结合元件和蛋白质结合元件取代甲基化结合蛋白(MBD)分区来表征其他表观遗传相互作用。例如,在各种免疫沉淀方案中使用的针对组蛋白、修饰的组蛋白和转录因子的抗体可以取代MBD分区,以通过使用不同组的分子标签产生与样品中的每个DNA分子相关的关于核小体定位、核小体修饰和转录因子结合的相关信息。

[0254] 数据分析

[0255] 对液体活检物中的癌症甲基化分析面临的主要挑战是细胞类型异质性。除了固有的和证据充分的癌症异质性外,血浆中的无细胞DNA代表混合细胞死亡类型,其不是主要癌症相关的。例如,细胞死亡可以在无恶性病变器官、生理造血谱系中。除了该复杂性之外,甚

至基质组分中的非癌细胞非常不同，例如血管和淋巴内皮细胞和周细胞、免疫细胞诸如巨噬细胞、白细胞和淋巴细胞、基质成纤维细胞、肌成纤维细胞、肌上皮细胞以及脂肪细胞、内分泌细胞、神经细胞和其他具有不同发育起源的细胞和组织元件。因此，在一些实施方案中，当分析和解释来自液体活检物的结果时，对细胞类型组成的变化进行调整。

[0256] 分析线路可以包括以下步骤：

[0257] a)解析核小体占据

[0258] b)定位二分体，指定严格性

[0259] c)跨全基因组在单个基因组元件中拟合高斯混合模型

[0260] d)在基因水平解卷积细胞谱系

[0261] 作为说明性实例，可以在来自单个分区的样品中单独地确定cfDNA片段起始富集谱。例如，分区的样品可以包含超甲基化DNA、低甲基化DNA或中等甲基化DNA。所确定的cfDNA片段起始富集谱可以用于建立在相关调节元件，例如TSS、增强子区域、远端基因间元件中的核小体占据。对每个分区，可以确定占据峰，例如二分体，并且可以指定它们的严格性。与健康血浆样品中观察到的细胞状态相关的规范谱可以通过确定cfDNA片段起始富集谱并将二分体定位在大型非恶性对照(例如，来自健康个体或多于一个健康个体的样品)来建立。对于任何样品，高斯混合模型可以使用如上文定义的规范谱来拟合，以产生对应于在分区的样品中观察到的恶性(非规范)染色质状态的剩余占据，从而确定非规范cfDNA片段峰和谱。非规范cfDNA片段峰和谱可能与每个分区的样品中的癌症中的恶性染色质状态相关。通过甲基化的生物调节可以由单个CpG或一组彼此非常接近的CpG介导。因此，DNA甲基化的区域分析为甲基化数据提供了更全面和系统的观点。通常，甲基化信息汇总在平铺窗口或一组预定义区域(启动子、CpG岛、内含子等)。

[0262] 核小体组织可以通过两个独立的度量，诸如核小体占据和核小体定位来确定。核小体占据可以被理解为核小体存在于细胞群体中的特定基因组区域的概率。核小体占据可以在基于测序的实验中测量为覆盖度(映射到基因组区域的对齐的测序读段的数目)。核小体定位可以是核小体参考点(例如二分体)相对于周围坐标位于特定基因组坐标的概率。如图9中示出的，良好的核小体定位可以从生物学上解释为每次出现在同一基因组坐标的核小体二分体。不良定位可以被解释为核小体二分体占据了整个核小体的相同常规足迹内的一系列位置。在一个实例中，来自8名患有肺癌的受试者的样品用于确定二分体中心。确定核小体定位和核小体占据。例如，当覆盖度为>0.5分位数(Qu)且峰宽为<0.5Qu时，可以指示高占据和良好定位。在一些情况下，可将分级分离的样品(诸如超甲基化级分/低甲基化级分)的二分体中心之间的距离与未分级分离的样品(无MBD)比较。在一些情况下，二分体中心以及相邻染色质结构可以通过将二分体中心指定到跨基因组具有高于5％的占据覆盖度的所有峰来解析。占据覆盖度可以是15％、20％、25％或30％。可以使用机器学习方法通过确定峰位置、宽度、长度、中心和宽度分辨率来指定占据覆盖度。这为血浆DNA的染色质结构提供了经验分辨率。

[0263] 序列读段的覆盖度的增加可能与更大的核小体占据相关。此外，核小体占据可能与核小体耗尽区域(NDR)负相关。核小体占据的增加可以指示改变的染色质结构，诸如更紧密的染色质。紧密的染色质可以指示基因表达的下调，这可能干扰正常细胞功能行使。正常细胞功能行使的干扰可以作为疾病诸如癌症的指征。

[0264] 无细胞DNA包括来自异质细胞群体(例如濒死、恶性、非恶性等)的信号。异质细胞群体可以具有具有多种染色质状态的核酸。在一些情况下,多种染色质状态可以包括核小体占据的不同状态,诸如良好定位或分散(“模糊的(fuzzy)”)的核小体。良好定位的核小体示出更大的覆盖度,而失真的核小体示出更低覆盖度的序列读段。基于序列读段的覆盖度,可以解析染色质上的核小体占据。

[0265] “解卷积”可以指的是分解相互重叠的无细胞DNA片段占据峰的过程,从而提取关于“被隐藏的峰”的信息。核小体占据峰的解卷积可以通过MBD分区来实现。将核酸分区成超甲基化分区和低甲基化分区可以产生两个不同的峰,峰1和峰2。然而,当核酸没有被分级分离时,可以获得一个连续的峰,并且将恶性肿瘤相关的峰1从非恶性峰2中解卷积可能是不可行的。

[0266] 二分体可以是DNA被核小体中心所占据的区域。二分体可以位于分区的样品中。在一些情况下,核酸被分区成超甲基化级分和低甲基化级分。二分体定位(positioning)或定位(localization)可以使用无参考方法或基于参考的方法来进行。无参考方法可以包括超分区和低分区的计算机模拟组合,以确定潜在的二分体位置,从而确定二分体图。在一些情况下,来自超甲基化分区和低甲基化分区的测序数据被组合以确定核小体占据,并且在分区之间进行比较,例如,组合来自所有分区的信号并检测占据峰,然后比较那些在超甲基化分区对比低甲基化分区中观察到的位置。基于参考的方法可以包括对分区的独立分析。例如,确定超甲基化级分和低甲基化级分的核小体占据。第一个实验中每个分区的核小体占据可以用于后续实验中的相应分区,其中相同的部分1在一大组样品上独立完成(标准WGS就足够了,因为不使用基于分区的信息,并且信息被组合以提高峰分辨率),并且单独的分区(或两者)可与针对其被存储为“参考”的占据峰的映射比较。

[0267] 基于片段组数据的片段组特征

[0268] 例如,在2017年7月6日提交的美国公布2016/0201142(Lo)、WO 2016/015058(Shendure)和PCT/US17/40986(“Methods For Fragmentome Profiling Of Cell-Free Nucleic Acids”)中描述了检查片段组数据的方法,所有这些文献均通过引用并入本文。片段组数据指的是通过分析核酸片段获得的序列数据。例如,序列数据可以包括片段长度(以碱基对计)、基因组坐标(例如参考基因组上的起始和终止位置)、覆盖度(例如拷贝数)或序列信息(例如碱基A、G、C、T)。片段组数据指的是片段起始和终止的序列信息,以及对应于在血液或血浆中观察到的无细胞DNA的受保护的内容物的富集的无细胞DNA中的相关占据。

[0269] 例如,人们可以确定样品中具有映射到整个基因组或其靶部分的特定核苷酸坐标的中心点的cfDNA分子的数目。在健康个体中,这通常将产生波形图,其中图的峰代表核小体位置(例如,在转化为cfDNA期间细胞DNA没有裂解),并且谷代表核小体间位置(例如,在核小体间位置,许多分子被裂解,并且因此很少分子位于中心)。峰之间的距离代表核小体二分体。在恶性细胞中,核小体的位置可以例如作为甲基化的函数移动。在该情况下,人们预期图中峰和谷的位置移动。通过基于不同特征将分子分区,并检查每个分区的片段分布,可以更容易地检测到这样的移动。片段数据可以在一个或多于一个的多个维度上进一步分析。例如,在任何坐标,映射到其的分子的数目可以基于片段大小来进一步区分。在基于这样的数据的图中,第三个“Z”维度代表片段大小。因此,例如,在二维图中,X轴代表基因组坐标,并且Y轴代表映射到坐标的分子的数目。在三维图中,X轴代表基因组坐标,Z轴代表片段

长度,并且Y轴代表映射到坐标的每个大小的分子的数目。这样的三维图可以表示为二维热图,其中,X轴和Z轴以二维显示,而Y轴上的值由例如颜色强度(例如,较暗表示较大值)或颜色的“热度”(例如,蓝色表示较低值,而红色表示较高值)表示。可以挖掘这样的数据来确定被检查的状态的核小体位置模式特征,诸如癌症的存在或不存在、癌症的类型、转移程度等。

[0270] 个体群组可能都具有共享的特征。该共享的特征可以选自由以下组成的组:肿瘤类型、炎症状况、凋亡状况、坏死状况、肿瘤复发和对治疗的耐受性。在一些情况下,群组包括具有特定类型的癌症(例如,乳腺癌、结肠直肠癌、胰腺癌、前列腺癌、黑素瘤、肺癌或肝癌)的个体。为了获得癌症的核小体特征,患有癌症的个体提供血液样品。从血液样品获得无细胞DNA。对无细胞DNA测序(在从基因组选择性富集或不富集一组区域的情况下)。来自测序反应的呈序列读段的形式的序列信息被映射到人类参考基因组。在一些实施方案中,分子在映射操作之前或之后瓦解成独特的分子读段。

[0271] 由于给定样品中的无细胞DNA片段代表了从其产生无细胞DNA的细胞的混合物,因此每种细胞类型的不同核小体占据可能导致对代表给定无细胞DNA样品的数学模型的贡献。例如,片段长度的分布可能由于不同细胞类型之间或肿瘤细胞与非肿瘤细胞之间不同的核小体保护而引起。该方法可以用于基于序列数据的单参数、多参数和/或统计分析来开发一组临床上有用的评估。

[0272] 样品中的核酸分子可以基于一种或更多种特征来分级分离。分级分离可以包括基于基因组特征的存在或不存在将核酸分子物理地分区成子集或组。分级分离可以包括基于基因组特征存在的程度将核酸分子物理地分区成组。基于指示差异基因表达或疾病状态的特征,可以将样品分级分离或分区成一个或更多个组。样品可以基于在核酸分析期间提供正常状态和患病状态之间的信号差异的特征来分级分离,所述核酸例如cfDNA、非cfDNA、肿瘤DNA、循环肿瘤DNA(ctDNA)。

[0273] 片段组数据可以用于推断遗传变体。遗传变体包括拷贝数变异(CNV)、插入和/或缺失(插入缺失)、单核苷酸变异(SNV)和/或基因融合。片段组数据可以用于推断表观遗传变体,诸如指示癌症的变体。可以确定每个分级分离组或分区组和/或未分级分离的核酸中的一个或更多个遗传变体。分级分离或分区可以基于各种特征中的至少一种来进行,各种特征包括但不限于核酸的甲基化状态、大小、长度和转录结合。在分级分离组或分区组中确定的遗传变体可以相互比较和/或与可能具有或可能不具有相同特征的未分级分离的核酸比较。分级分离或分区的核酸可以重新组合,并且片段组数据可以与未分级分离的核酸和/或不具有与分级分离的或分区的核酸相同特征的核酸进行比较,以确定遗传变体的存在。

[0274] 模型可以用于组配置中,以选择性富集区域(例如片段组谱相关区域),并且确保跨越特定突变的大量读段,也可以考虑重要的染色质中心事件如转录起始位点(TSS)、启动子区域、连接位点和内含子区域。

[0275] 在一个实例中,在内含子和外显子的连接(或边界)处或连接(或边界)附近发现片段组谱的差异。一个或更多个体细胞突变的识别可以与一个或更多个多参数或单参数模型相关联,以揭示cfDNA片段分布处的基因组位置。该相关性分析可以揭示一个或更多个内含子-外显子连接,在这些连接中片段组特征谱破坏是最明显的。

[0276] 作为另一个实例,可以观察到样品在远离TSS的区域中的超甲基化。在距TSS 0kb

和5kb、5kb和50kb、和/或50kb和500kb之间的距离内可以观察到超甲基化区域的富集。在距TSS 5kb和50kb之间可以观察到超甲基化区域的富集。在距TSS少于5kb、10kb、15kb、20kb、25kb、30kb、35kb、40kb、50kb、100kb、200kb、300kb、400kb和/或500kb可以观察到超甲基化区域的富集。在距TSS多于5kb、10kb、15kb、20kb、25kb、30kb、35kb、40kb、50kb、100kb、200kb、300kb、400kb和/或500kb可以观察到超甲基化区域的富集。超甲基化的位置和富集可以在从健康或正常受试者获得的DNA(正常DNA)和从患病受试者获得的DNA之间变化。例如,来自疑似具有或具有肺癌的受试者的DNA(肺癌DNA)可以示出距TSS中规范位置距离最远的超甲基化的富集,并且超甲基化级分中良好定位的核小体占据启动子区域附近(图17)。例如,来自肺癌患者的未分级分离的核酸(无MBD)用于测序。基于片段组数据,诸如基因组位置,确定核小体二分体中心的序列读段。进一步基于片段组数据,进一步分析具有小于或等于5%的覆盖度或小于或等于95%的覆盖度的序列读段。基因注释工具,诸如注释工具的基因组区域富集(Genomic Regions Enrichment of Annotations Tool,GREAT)被用来基于附近的基因为一组基因组区域指定功能。确定序列读段和其推定调节的基因之间的距离(图17)。这些距离被分成四个独立的箱元:一个从0kb到5kb,另一个从5kb到50kb,第三个从50kb到500kb,以及超过500kb的所有关联的最后一个箱元。为了精确起见,箱元是[0,5kb]、[5kb,50kb]、[50kb,500kb]、[500kb,无穷大]。在图中,精确到0(即在TSS上)的所有关联在[-5kb,0]和[0,5kb]箱元之间平均分隔。使用该方法,在背景基因组区域(例如所有核小体)和最明显的基因组区域(例如甲基化核小体)两者中远离TSS的区域中观察到样品中的超甲基化。例如,在[5kb,50kb]箱元之间观察到超甲基化区域的富集。

[0277] 片段组特征可能有助于确定核小体占据、核小体定位、RNA聚合酶II暂停、细胞死亡特异性DNA酶超敏反应和细胞死亡期间染色质浓缩。这样的特征还可以提供对细胞碎片清除和运输的洞察。例如,细胞碎片清除可以包括由凋亡死亡的细胞中的半胱天冬酶激活的DNA酶(CAD)进行的DNA片段化,而且还可以在死亡细胞被吞噬后由溶酶体DNA酶II进行,产生不同的裂解图。

[0278] 基因组分区图可以通过在基因组范围内识别与染色质的前述性质相关的恶性状况对比非恶性状况中的差异染色质状态,经由聚集重要窗口进入感兴趣的区域来构建。这样的感兴趣的区域通常被称为基因组分区图。

[0279] 基于甲基化状态的分级分离

[0280] 样品中的核酸分子可以基于5-甲基胞嘧啶的特征来分级分离。DNA可以在胞嘧啶处甲基化,诸如在CpG二核苷酸区域。DNA甲基化和组蛋白复合物可能影响DNA包装成染色质以及基因表达的表观遗传调节。表观遗传改变在各种疾病中可能发挥着重要的作用,诸如在癌症进展的所有阶段,原发性或早期癌症的开始,复发或转移癌。例如,正常低甲基化区域,诸如参与正常生长、DNA修复、细胞周期调节和细胞分化的基因的转录起始位点(TSS)的超甲基化,可以指示癌症。超甲基化可以通过抑制转录来改变基因表达。在一些情况下,超甲基化可以降低和/或抑制基因表达。例如,超甲基化可以降低和/或抑制致癌基因阻遏物的表达。在一些情况下,超甲基化可以增加和/或促进基因表达。例如,抑制剂的超甲基化可以导致增加和/或促进的下游响应物的基因表达,例如通常被抑制剂抑制的致癌基因。

[0281] 基于DNA甲基化状态,样品中的核酸分子可以被分级分离成不同的组,使得可以使用实验程序富集具有类似甲基化状态的核酸分子。例如,甲基结合结构域(MBD)蛋白可以用

于亲和纯化具有类似甲基化状态的核酸分子，所述甲基化状态例如超甲基化、低甲基化和残留甲基化。在另一个实例中，对5-甲基-胞嘧啶特异的抗体可以用于免疫沉淀具有类似甲基化水平的核酸分子。在另一个实例中，基于亚硫酸氢盐的方法可以用于选择性地富集高度甲基化的核酸分子。在又另一个实例中，甲基化敏感的限制性内切酶可以用于选择性地富集高度甲基化的核酸分子。

[0282] 在使用特征之一分级分离后，可以对每组中的核酸分子测序，以生成序列读段。序列读段可以被映射到参考基因组。映射可以生成序列信息。可以分析序列信息以确定遗传变异，包括例如单核苷酸变异、拷贝数变异、插入缺失或融合。在使用本文公开的方法测定无细胞DNA的情况下，可以生成片段组数据，片段组数据可以在分级分离的核酸分子的各组之间变化。片段组数据可以包括基因组坐标、大小、覆盖度或序列信息。本公开内容提供了用于将片段组数据与来自每个分区的序列读段整合的方法。这样的整合可以用于准确且快速地检测指示疾病状态的生物标志物。

[0283] 本文描述的方法可以用于基于片段组数据计算机模拟富集核酸分子。例如，来自肺癌患者的未分级分离的核酸分子(无MBD)可以用于测序。在另一个实例中，可以基于单独的单核小体或双核小体谱的差异或联合其他特征诸如大小和/或甲基化状态来实现分级分离。单核小体谱可以指的是包裹单个核小体所需的近似长度(例如，约146bp)的片段的覆盖度或计数。双核小体谱可以指的是包裹单个核小体两次所需的近似长度(例如，约292bp)的片段的覆盖度或计数。

[0284] 数据分析

[0285] 在某些实施方案中，来自不同类别的受试者的数据，例如癌症/无癌症、癌症类型1/癌症类型2，可以用于训练机器学习算法，以将样品分类为属于这些类别之一。如本文使用的术语“机器学习算法”指的是由计算机执行的算法，其自动化分析模型构建，例如用于聚类、分类或模式识别。机器学习算法可以是有监督或无监督的。学习算法包括例如人工神经网络(例如反向传播网络)、判别分析(例如贝叶斯分类器或费歇尔分析)、支持向量机、决策树(例如递归分区过程，诸如CART分类和回归树)、随机森林、线性分类器(例如多元线性回归(MLR)、偏最小二乘(PLS)回归和主成分回归(PCR))、层级聚类和聚类分析。机器学习算法学习的数据集可以被称为“训练数据”。

[0286] 如本文使用的术语“分类器”指的是算法计算机代码，其接收测试数据作为输入，并且产生输入数据的分类作为输出，该分类属于一个或另一个类别。

[0287] 如本文使用的术语“数据集”指的是表征系统元件的值的集合。系统可以是例如来自生物样品的cfDNA。这样的系统的元件可以是遗传基因座。数据集(或“数据集”)的实例包括指示选自以下特征的定量量度的值：(i)映射到遗传基因座的DNA序列，(ii)在遗传基因座起始的DNA序列，(iii)在遗传位点终止的DNA序列；(iv)DNA序列的双核小体保护或单核小体保护；(v)位于参考基因组的内含子或外显子的DNA序列；(vi)具有一个或更多个特征的DNA序列的大小分布；(vii)具有一个或更多个特征的DNA序列的长度分布，等等。

[0288] 如本文使用的术语“值”指的是数据集中的条目，其可以是表征该值所指的特征的任何事物。这包括但不限于数字、单词或短语、符号(例如+或-)或程度。

[0289] 数字处理设备

[0290] 在一些实施方案中，本文公开的方法利用数字处理设备。在另外的实施方案中，数

字处理设备包括执行设备功能的一个或更多个硬件中央处理单元(CPU)或通用图形处理单元(GPGPU)。在仍另外的实施方案中,数字处理设备还包括被配置为执行可执行指令的操作系统。在一些实施方案中,数字处理设备任选地被连接到计算机网络。在另外的实施方案中,数字处理设备任选地被连接到因特网,使得它访问万维网。在仍另外的实施方案中,数字处理设备任选地被连接到云计算基础设施。在其他实施方案中,数字处理设备任选地被连接到内联网。在其他实施方案中,数字处理设备任选地被连接到数据存储设备。

[0291] 根据本文的描述,通过非限制性实例,合适的数字处理设备包括服务器计算机、台式计算机、膝上型计算机、笔记本计算机、手持计算机、互联网设备、移动智能手机和平板计算机。

[0292] 在一些实施方案中,数字处理设备包括被配置为执行可执行指令的操作系统。操作系统是例如软件,包括程序和数据,它管理设备的硬件并为应用程序的执行提供服务。本领域技术人员将认识到,通过非限制性实例,合适的服务器操作系统包括FreeBSD、OpenBSD、NetBSD®、Linux、Apple® Mac OS X Server®、Oracle® Solaris®、Windows Server®和Novell® NetWare®。本领域技术人员将认识到,通过非限制性实例,合适的个人计算机操作系统包括Microsoft® Windows®、Apple® Mac OS X®、UNIX®和UNIX-类操作系统,诸如GNU/Linux®。在一些实施方案中,操作系统由云计算提供。本领域技术人员还将认识到,通过非限制性实例,合适的移动智能手机操作系统包括Nokia® Symbian®OS、Apple® iOS®、Research In Motion® BlackBerry OS®、Google® Android®、Microsoft® Windows Phone®OS、Microsoft® Windows Mobile®OS、Linux®和Palm® WebOS®。

[0293] 在一些实施方案中,所述设备包括存储设备和/或存储器设备。存储设备和/或存储器设备是用于临时或永久存储数据或程序的一个或更多个物理设备。在一些实施方案中,所述设备是易失性存储器,并且需要电力来维持存储的信息。在一些实施方案中,所述设备是非易失性存储器,并且当数字处理设备未通电时保留存储的信息。在另外的实施方案中,非易失性存储器包括闪存。在一些实施方案中,非易失性存储器包括动态随机存取存储器(DRAM)。在一些实施方案中,非易失性存储器包括铁电随机存取存储器(FRAM)。在一些实施方案中,非易失性存储器包括相变随机存取存储器(PRAM)。在其他实施方案中,所述设备是存储设备,通过非限制性实例,包括CD-ROM、DVD、闪存设备、磁盘驱动器、磁带驱动器、光盘驱动器和基于云计算的存储器。在另外的实施方案中,存储和/或存储设备是诸如本文公开的设备的组合。

[0294] 在一些实施方案中,数字处理设备包括向用户发送视觉信息的显示器。在一些实施方案中,显示器是液晶显示器(LCD)。在另外的实施方案中,显示器是薄膜晶体管液晶显示器(TFT-LCD)。在一些实施方案中,显示器是有机发光二极管(OLED)显示器。在各种另外的实施方案中,在OLED显示器上是无源矩阵OLED(PMOLED)显示器或有源矩阵OLED(AMOLED)显示器。在一些实施方案中,显示器是等离子体显示器。在其他实施方案中,显示器是视频投影仪。在又其他实施方案中,显示器是与数字处理设备通信的头戴式显示器,诸如VR头机。在另外的实施方案中,通过非限制性实例,合适的VR头机包括HTC Vive、Oculus Rift、

Samsung Gear VR、Microsoft HoloLens、Razer OSVR、FOVE VR、Zeiss VR One、Avegant Glyph、Freefly VR头机等。在仍另外的实施方案中，显示器是诸如本文公开的设备的组合。

[0295] 在一些实施方案中，数字处理设备包括从用户接收信息的输入设备。在一些实施方案中，输入设备是键盘。在一些实施方案中，输入设备是指示设备，通过非限制性实例，包括鼠标、轨迹球、轨迹垫、操纵杆、游戏控制器或输入笔。在一些实施方案中，输入设备是触摸屏或多触摸屏。在其他实施方案中，输入设备是捕捉语音或其他声音输入的麦克风。在其他实施方案中，输入设备是捕捉运动或视觉输入的摄像机或其他传感器。在另外的实施方案中，输入设备是Kinect、Leap Motion等。在仍另外的实施方案中，输入设备是诸如本文公开的设备的组合。

[0296] 参考图32，在特定的实施方案中，示例性的数字处理设备101被编程或以其他方式配置成分析、测定、解码和/或解卷积序列和/或标签数据。在实施方案中，数字处理设备101包括中央处理单元(CPU，在本文中也为“处理器”和“计算机处理器”)105，其可以是单核或多核处理器或用于并行处理的多于一个处理器。数字处理设备101还包括存储器或存储器位置110(例如，随机存取存储器、只读存储器、闪速存储器)、电子存储单元115(例如，硬盘)、用于与一个或更多个其他系统通信的通信界面120(例如，网络适配器)和外围设备125，诸如高速缓冲存储器、其他存储器、数据存储和/或电子显示适配器。存储器110、储存单元115、界面120和外围设备125与CPU 105通过通信总线(实线)，诸如主板(motherboard)通信。存储单元115可以是用于存储数据的数据存储单元(或数据储存库)。数字处理设备101可以借助于通信界面120被可操作地耦合至计算机网络(“网络”)130。网络130可以是因特网(Internet)、互联网(internet)和/或外联网、或与因特网通信的内联网和/或外联网。在一些情况下，网络130是电信和/或数据网络。网络130可以包括一个或更多个计算机服务器，这可以支持分布式计算，诸如云计算。在一些情况下，借助于设备101，网络130可以实现对等网络(peer-to-peer network)，其可以使耦合至设备101的设备能够作为客户端或服务器运行。

[0297] 继续参考图32，CPU 105可以执行一系列的机器可读指令，该机器可读指令可以以程序或软件来体现。指令可以被存储于存储器位置，诸如存储器110中。指令可以被导向CPU 105，其可以随后编程或以其他方式配置CPU 105，以实现本公开内容的方法。由CPU 105进行的操作的实例可以包括读取、解码、执行和写回。CPU 105可以是电路诸如集成电路的一部分。设备101的一个或更多个其他组件可以被包含在该电路中。在一些情况下，电路为专用集成电路(ASIC)或现场可编程门阵列(FPGA)。

[0298] 继续参考图32，存储单元115可以存储文件，诸如驱动程序、库和保存的程序。存储单元115可以存储用户数据，例如，用户偏好和用户程序。在一些情况下，数字处理设备101可以包括一个或更多个另外的数据存储单元，该数据存储单元在外部，诸如位于通过内联网或因特网通信的远程服务器上。

[0299] 继续参考图32，数字处理设备101可以与一个或更多个远程计算机系统通过网络130通信。例如，设备101可以与用户的远程计算机系统通信。远程计算机系统的实例包括个人计算机(例如便携式PC)、板型或平板PC(例如Apple® iPad、Samsung® Galaxy Tab和Microsoft® Surface®)和智能手机(例如Apple® iPhone或Android支持的设备)。

[0300] 如本文描述的方法可以通过机器(例如,计算机处理器)可执行代码的方式至少部分地实现,该机器可执行代码被存储在数字处理设备101的电子存储位置,诸如例如存储器110或电子存储单元115上。机器可执行代码或机器可读代码可以以软件的形式提供。在使用期间,代码可以由处理器105执行。在一些情况下,代码可以从存储单元115检索并存储在存储器110上,以用于由处理器105迅速访问。在一些情况下,可以排除电子存储单元115,而将机器可执行指令存储于存储器110中。

[0301] 非瞬时性计算机可读存储介质

[0302] 在一些实施方案中,本文公开的方法利用一个或更多个非瞬时性计算机可读存储介质,该非瞬时性计算机可读存储介质用包括由任选地联网的数字处理设备的操作系统可执行的指令的程序编码。在另外的实施方案中,计算机可读存储介质是数字处理设备的有形组件。在仍另外的实施方案中,计算机可读存储介质任选地从数字处理设备可移除。在一些实施方案中,通过非限制性实例,计算机可读存储介质包括CD-ROM、DVD、闪存设备、固态存储器、磁盘驱动器、磁带驱动器、光盘驱动器、云计算系统和服务等。在一些情况下,程序和指令在介质上被永久地、基本上永久地、半永久地或非瞬时性地编码。

[0303] 可执行指令

[0304] 在一些实施方案中,本文公开的方法利用由数字处理设备可执行的以至少一个计算机程序形式的指令。例如,计算机程序包括一系列指令,在数字处理设备的CPU中可执行,被写入以执行指定的任务。计算机可读指令可以被实现为执行特定任务或实现特定抽象数据类型的程序模块,诸如函数、对象、应用编程接口(API)、数据结构等。根据本文提供的公开内容,本领域技术人员将认识到,计算机程序可以以各种语言的各种版本来编写。

[0305] 计算机可读指令的功能可以根据需要在各种环境中组合或分布。在一些实施方案中,计算机程序包括一系列指令。在一些实施方案中,计算机程序包括多于一个的系列的指令。在一些实施方案中,从一个位置提供计算机程序。在其他实施方案中,从多于一个位置提供计算机程序。在各种实施方案中,计算机程序包括一个或更多个软件模块。在各种实施方案中,计算机程序部分地或全部地包括一个或更多个网络应用程序、一个或更多个移动应用程序、一个或更多个独立应用程序、一个或更多个网络浏览器插件、扩展、内插附件(add-in)或附加组件(add-on)或其组合。

[0306] 网络应用程序

[0307] 在一些实施方案中,计算机程序包括网络应用程序。根据本文提供的公开内容,本领域技术人员将认识到,在各种实施方案中,网络应用程序利用一个或更多个软件框架和一个或更多个数据库系统。在一些实施方案中,在软件框架诸如Microsoft®.NET或Ruby on Rails(RoR)上创建网络应用程序。在一些实施方案中,网络应用程序利用一个或更多个数据库系统,通过非限制性实例,包括关系数据库系统、非关系数据库系统、面向对象数据库系统、关联数据库系统和XML数据库系统。在另外的实施方案中,通过非限制性实例,合适的关系数据库系统包括Microsoft®SQL服务器、mySQL™和Oracle®。本领域技术人员还将认识到,在各种实施方案中,网络应用程序是以一种或更多种语言的一种或更多种版本编写的。网络应用程序可以以一种或更多种标记语言、表示定义语言、客户端脚本语言、服务器端编码语言、数据库查询语言或其组合来编写。在一些实施方案中,网络应用程序在某种程度上以标记语言,诸如超文本标记语言(HTML)、可扩展超文本标记语言(XHTML)或可扩展

标记语言(XML)来编写。在一些实施方案中,网络应用程序在某种程度上以表示定义语言诸如层叠样式表(CSS)来编写。在一些实施方案中,网络应用程序在某种程度上以客户端脚本语言,诸如异步Javascript和XML(AJAX)、Flash®Actionscript、Javascript或Silverlight®来编写。在一些实施方案中,网络应用程序在某种程度上以服务器端编码语言,诸如活动服务器页面(ASP)、ColdFusion®、Perl、Java™、JavaServer Pages(JSP)、超文本预处理器(PHP)、Python™、Ruby、Tcl、Smalltalk、WebDNA®或Groovy来编写。在一些实施方案中,网络应用程序在某种程度上以数据库查询语言,诸如结构化查询语言(SQL)来编写。在一些实施方案中,网络应用程序集成了企业服务器产品,诸如IBM®Lotus Domino®。在一些实施方案中,网络应用程序包括媒体播放器元件。在各种另外的实施方案中,媒体播放器元件利用许多合适的多媒体技术中的一种或更多种,通过非限制性实例,包括Adobe® Flash®、HTML 5、Apple® QuickTime®、Microsoft® Silverlight®、Java™和Unity®。

[0308] 参考图33,在特定的实施方案中,应用程序供应系统包括由关系数据库管理系统(RDBMS)210访问的一个或更多个数据库200。合适的RDBMS包括Firebird、MySQL、PostgreSQL、SQLite、Oracle Database、Microsoft SQL Server、IBM DB2、IBM Informix、SAP Sybase、Teradata等。在该实施方案中,应用程序供应系统还包括一个或更多个应用程序服务器220(诸如Java服务器、.NET服务器、PHP服务器等)和一个或更多个网络服务器230(诸如Apache、IIS、GWS等)。网络服务器任选地经由app应用编程接口(API)240经由网络诸如因特网展示一个或更多个网络服务,系统提供基于浏览器的和/或移动本地用户接口。

[0309] 参考图34,在特定的实施方案中,应用程序供应系统可替代地具有分布式的基于云的架构300,并且包括弹性负载平衡的自动缩放的网络服务器资源310和应用服务器资源320以及同步复制的数据库330。

[0310] 移动应用程序

[0311] 在一些实施方案中,计算机程序包括提供给移动数字处理设备的移动应用程序。在一些实施方案中,移动应用程序在制造时被提供给移动数字处理设备。在其他实施方案中,移动应用程序经由本文描述的计算机网络被提供给移动数字处理设备。

[0312] 根据本文提供的公开内容,通过本领域技术人员已知的技术,使用本领域已知的硬件、语言和开发环境来创建移动应用程序。本领域技术人员将认识到,移动应用程序是以若干语言编写的。通过非限制性实例,合适的编程语言包括C、C++、C#、Objective-C、Java™、Javascript、Pascal、Object Pascal、Python™、Ruby、VB.NET、WML、和带有或没有CSS的XHTML/HTML,或其组合。

[0313] 合适的移动应用程序开发环境从若干来源可获得。通过非限制性实例,商业上可获得的开发环境包括AirplaySDK、alcheMo、Appcelerator®、Celsius、Bedrock、Flash Lite、.NET紧凑框架、Rhomobile和WorkLight移动平台。其他开发环境也是免费获得的,通过非限制性实例,包括Lazarus、MobiFlex、MoSync和PhoneGap。此外,移动设备制造商发布软件开发工具包,通过非限制性实例,包括iPhone和iPad(iOS)SDK、Android™SDK、BlackBerry®SDK、BREW SDK、Palm®OS SDK、Symbian SDK、webOS SDK和Windows®

Mobile SDK。

[0314] 本领域技术人员将认识到，若干商业论坛可用于移动应用程序的发布，通过非限制性实例，包括Apple®App Store、Google®Play、Chrome WebStore、BlackBerry®App World、用于Palm设备的App Store、用于webOS的App Catalog、用于Mobile的Windows® Marketplace、用于Nokia®设备的Ovi Store、Samsung®Apps和Nintendo®DSi Shop。

[0315] 独立应用程序

[0316] 在一些实施方案中，计算机程序包括独立应用程序，其是作为独立计算机进程运行的程序，而不是现有进程的附加组件，例如，不是插件。本领域技术人员将认识到，独立应用程序经常是被编译的。编译器是将以编程语言编写的源代码转换成二进制目标代码，诸如汇编语言或机器代码的计算机程序。通过非限制性实例，合适的编译编程语言包括C、C++、Objective-C、COBOL、Delphi、Eiffel、Java™、Lisp、Python™、Visual Basic和VB.NET或其组合。编译经常被执行，至少部分是为了创建可执行程序。在一些实施方案中，计算机程序包括一个或更多个可执行的编译应用程序。

[0317] 软件模块

[0318] 在一些实施方案中，本文公开的方法利用软件、服务器和/或数据库模块。根据本文提供的公开内容，软件模块通过本领域技术人员已知的技术使用本领域已知的机器、软件和语言来创建。本文公开的软件模块以多种方式实现。在各种实施方案中，软件模块包括文件、代码段、编程对象、编程结构或其组合。在另外的各种实施方案中，软件模块包括多于一个文件、多于一个代码段、多于一个编程对象、多于一个编程结构或其组合。在各种实施方案中，通过非限制性实例，一个或更多个软件模块包括网络应用程序、移动应用程序和独立应用程序。在一些实施方案中，软件模块在一个计算机程序或应用程序中。在其他实施方案中，软件模块在多于一个计算机程序或应用程序中。在一些实施方案中，软件模块托管在一台机器上。在其他实施方案中，软件模块托管在多于一台机器上。在另外的实施方案中，软件模块托管在云计算平台上。在一些实施方案中，软件模块托管在一个位置的一台或更多台机器上。在其他实施方案中，软件模块托管在多于一个位置的一台或更多台机器上。

[0319] 数据库

[0320] 在一些实施方案中，本文公开的方法利用一个或更多个数据库。根据本文提供的公开内容，本领域技术人员将认识到，许多数据库适合于存储和检索患者、序列、标签、编码/解码、遗传变体和疾病信息。在各种实施方案中，通过非限制性实例，合适的数据库包括关系数据库、非关系数据库、面向对象数据库、对象数据库、实体关系模型数据库、关联数据库和XML数据库。另外的非限制性实例包括SQL、PostgreSQL、MySQL、Oracle、DB2和Sybase。在一些实施方案中，数据库是基于因特网的。在另外的实施方案中，数据库是基于网络的。在仍另外的实施方案中，数据库是基于云计算的。在其他实施方案中，数据库基于一个或更多个本地计算机存储设备。

[0321] 在一个方面中，本文提供了一种包括计算机的系统，该计算机包括处理器和计算机存储器，其中计算机与通信网络通信，并且其中计算机存储器包括代码，当代码被处理器执行时，(1)从通信网络接收序列数据到计算机存储器中；(2)使用本文描述的方法，确定序列数据中的遗传变体代表生殖系突变体还是体细胞突变体；以及(3)通过通信网络报告该

确定。

[0322] 通信网络可以是连接到因特网的任何可用的网络。通信网络可以利用例如高速传输网络，包括但不限于电力线宽带(BPL)、电缆调制解调器、数字用户线路(DSL)、光纤、卫星和无线。

[0323] 在一个方面中，本文提供了一种系统，包括：局域网；一个或更多个DNA测序仪，其包括被配置为存储DNA序列数据的被连接到局域网的计算机存储器；生物信息学计算机，其包括计算机存储器和处理器，该计算机连接到局域网；其中计算机还包括代码，当被执行时，代码拷贝存储在DNA测序仪上的DNA序列数据，将拷贝的数据写入生物信息学计算机中的存储器，并执行如本文描述的步骤。

[0324] 本文还提供了用于实现所描述的方法的许多系统。在一些实施方案中，所述系统包括核酸测序仪，包括下一代DNA测序仪，该测序仪与数字处理设备数据通信，其中当测序仪从已经由本主题的方法分区和加标签的分区和加标签的DNA序列获得DNA序列信息时，由数字处理设备上的一个或多个软件模块接收的数据由测序仪生成。如果系统组件之间存在合适的数据通信，则测序仪和数字处理设备不需要彼此靠近，并且在一些实施方案中，可以分开很大的物理距离。下文描述的具体的系统实施方案是本发明提供的更大种类的系统的实例。本领域技术人员将理解，本文描述的包括数据分析步骤的方法可以通过本文公开的系统容易地实现，其中数字处理设备上的一个或多个软件模块用于分析通过对由本主题的方法产生的加标签的核酸群体测序获得的序列数据。

[0325] 一个实施方案是一种系统，包括：

[0326] 核酸测序仪；数字处理设备，其包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器；以及

[0327] 通信连接核酸测序仪和数字处理设备的数据链路；其中数字处理设备还包括可被执行以创建用于分析核酸群体的应用程序的指令，所述核酸群体包括选自双链DNA、单链DNA和单链RNA的至少两种形式的核酸，至少两种形式中的每一种包含多于一个分子，所述应用程序包括：(i)软件模块，其经由数据链路从核酸测序仪接收序列数据、所述扩增的核酸的序列数据，扩增的核酸中的至少一些被加标签；所述序列数据通过以下生成：将这些形式的核酸中的至少一种与至少一种加标签的核酸连接以将所述形式彼此区分开，扩增这些形式的核酸，其中的至少一种被连接至至少一种核酸标签，其中核酸和连接的核酸标签被扩增以产生扩增的核酸，其中从至少一种形式扩增的核酸被加标签；以及(ii)软件模块，其通过获得足以解码扩增的核酸的加标签的核酸分子的序列信息来测定扩增的核酸的序列数据，以揭示群体中为已测定了针对其的序列数据的连接至标签核酸分子的扩增的核酸提供原始模板的的核酸的形式。在另一个实施方案中，所述系统还包括解码扩增的核酸的加标签的核酸分子的软件模块，以揭示群体中为已测定了针对其的序列数据的连接至标签核酸分子的扩增的核酸提供原始模板的核酸的形式。在系统的其他另一个实施方案中，应用程序还包括经由通信网络传输测定的结果的软件模块。

[0328] 另一个实施方案是一种系统，包括：下一代测序(NGS)仪器；

[0329] 数字处理设备，其包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器；以及通信连接NGS仪器和数字处理设备的数据链路；其中数字处理设备还包括可被执行以创建应用程序的指令，所述应用程序包括：(i)软件模块，其用于经由数据链路从NGS

仪器接收序列数据,所述序列数据通过以下生成:对来自人类样品的DNA分子物理分级分离以生成两个或更多个分区,将差异化分子标签和支持NGS的衔接子应用于两个或更多个分区中的每一个以生成加分子标签的分区,以及用NGS仪器测定加分子标签的分区;(ii)软件模块,其用于生成用于将样品解卷积成被差异化分区的分子的序列数据;以及(iii)软件模块,其用于通过将样品解卷积成被差异化分区的分子来分析序列数据。在系统的其他另一个实施方案中,所述系统还包括经由通信网络传输测定的结果的软件模块。

[0330] 另一个实施方案是一种系统,包括:下一代测序(NGS)仪器;数字处理设备,其包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器;以及

[0331] 通信连接NGS仪器和数字处理设备的数据链路;

[0332] 其中数字处理设备还包括可由至少一个处理器执行以创建用于分子标签识别MBD-珠分级分离文库的应用程序的指令,所述应用程序包括:软件模块,其被配置为经由数据链路从NGS仪器接收序列数据,所述序列数据通过以下生成:使用甲基结合结构域蛋白质-珠纯化试剂盒对提取的DNA样品进行物理分级分离,保留所有洗脱物用于下游处理;进行差异化分子标签和支持NGS的衔接子序列对每个级分或组的并行应用;重新组合所有加分子标签的级分或组,并且随后使用衔接子特异性DNA引物序列进行扩增;对重新组合和扩增的总文库进行富集/杂交,靶向感兴趣的基因组区域;重新扩增富集的总DNA文库,附以样品标签;合并不同的样品;和在NGS仪器上对其进行多重测定;其中由所述仪器产生的NGS序列数据提供了用于识别独特分子的分子标签的序列,以及用于将样品解卷积成被差异化MBD分区的分子的序列数据;以及软件模块,其被配置为通过使用分子标签识别独特的分子并且将样品解卷积成被差异化MBD-分区的分子来进行序列数据的分析。另一个实施方案是一种系统,其中应用程序还包括经由通信网络传输分析的结果的软件模块。

[0333] 另一个实施方案是一种系统,包括:(a)下一代测序(NGS)仪器;(b)数字处理设备,其包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器;以及(c)通信连接NGS仪器和数字处理设备的数据链路;

[0334] 其中数字处理设备还包括可被执行以创建应用程序的指令,所述应用程序包括:i)软件模块,其用于经由数据链路从NGS仪器接收序列数据,所生成的序列数据加载有标记的核酸,所述标记的核酸通过以下来制备:使核酸群体与优先结合带有修饰的核酸的剂接触,分离与剂结合的第一核酸池和与剂未结合的第二核酸池,其中第一核酸池对于修饰是呈现过度的,并且第二池中的核酸对于修饰是呈现不足的;将第一池和/或第二池中的核酸连接至一个或更多个核酸标签,所述核酸标签区分第一池和第二池中的核酸,以产生加标签的核酸的群体;扩增标记的核酸,其中核酸和连接的标签被扩增;并且用NGS仪器测定加分子标签的分区;ii)软件模块,其用于生成解码标签的序列数据;以及iii)软件模块,其用于分析序列数据以解码标签以揭示已测定了针对其的序列数据的核酸是从第一池还是第二池中的模板扩增的。另一个实施方案是一种系统,所述系统还包括经由通信网络传输测定的结果的软件模块。

[0335] VII.实施例:

[0336] 实施例1:用于基于甲基结合结构域(MBD)分级分离的实验程序

[0337] 样品收集

[0338] 来自患有肺癌(例如NSCLC)的受试者的样品,诸如血液、血清或血浆从示出通过

GUARDANT360™测定法确定的高循环肿瘤DNA(ctDNA)含量的Guardant Health repository中选择。如先前描述的,来自健康正常供体的无细胞DNA(cfDNA)从血液分离的血浆提取(Lanman等人,Analytical and clinical validation of a digital sequencing panel for quantitative,highly accurate evaluation of cell-free circulating tumor DNA,PLoS ONE 10(10):e0140712(2015))。

[0339] cfDNA提取

[0340] 样品经历蛋白酶K消化。用异丙醇沉淀DNA。将DNA捕获在DNA纯化柱(例如,QIAamp DNA血液迷你试剂盒)上,并且洗脱在100μl溶液中。用Ampure SPRI磁珠捕获剂(PEG/盐)选择低于500bp的DNA。将产生的产物悬浮在30μl H_2O中。检查尺寸分布(主峰=166个核苷酸;次峰=330个核苷酸)并定量。通常,5ng的提取的DNA含有约1700的单倍体基因组当量("HGE")。DNA的量和HGE之间的一般关联如下列出:3pg DNA=1HGE;3ng DNA=1KHGE;3ng DNA=1M HGE;10pg DNA=3HE;10ngDNA=3K HGE;10ng DNA=3M HGE。

[0341] DNA分级分离

[0342] 将DNA分级分离成多个级分(或分区)。使用MethylMiner™亲和富集方案(Thermo Fisher Scientific,目录#ME10025)将cfDNA(10ng-150ng)分级分离成超甲基化级分、中等甲基化级分和低甲基化级分,不同之处在于使用300mM NaCl孵育和洗涤缓冲液对反应条件进行修改,并且按比例缩小一微克DNA输入的方案用于亚微克DNA输入。

[0343] 珠制备

[0344] 洗涤Dynabeads®M-280链霉亲和素:Dynabeads®M-280链霉亲和素在与MBD-生物素蛋白偶联之前,使用含有300mM NaCl的洗涤缓冲液洗涤。将Dynabeads®M-280链霉亲和素的原液重悬以获得均匀的悬浮液。对于每微克的输入DNA,将10μl的珠添加至1.7-ml无DNA酶的微量离心管中。用1X结合/洗涤缓冲液使珠的体积达到100μl。将管放置在磁架上1分钟,以在去除和丢弃液体之前将所有珠浓缩在管的内壁上。将管从磁架取出,并且添加等体积(例如,约100-250μl)的1×结合/洗涤缓冲液以重悬珠。将重悬的珠再次浓缩并洗涤,然后进行MBD-生物素蛋白与珠的偶联。

[0345] 将Dynabeads®M-280链霉亲和素与MBD-生物素蛋白偶联:对于每微克的输入DNA,将7μl(3.5μg)的MBD-生物素蛋白添加至1.7-ml无DNA酶的微量离心管中。用含有300mM NaCl的1X结合/洗涤缓冲液使珠的体积达到100μl。稀释MBD-生物素蛋白,并从最初的珠洗涤液转移到重悬的珠的管中。将珠-蛋白质混合物在室温在旋转混合器上混合1小时,然后进行洗涤MBD-珠。

[0346] 洗涤MBD-珠:通过将管放置在磁架上1分钟,浓缩管中的MBD-珠。去除并丢弃液体。将珠用100-250μl的含有300mM NaCl的1X结合/洗涤缓冲液重悬,并且在室温在旋转混合器上混合5分钟。按上文描述的将珠再浓缩、洗涤和重悬两次。然后将管放置在磁架上1分钟,并且小心地去除并丢弃液体。将珠用100-250μl的含有300mM NaCl的1X结合/洗涤缓冲液重悬,然后捕获甲基化DNA。

[0347] 捕获MBD-珠上的片段化的甲基化DNA,并且用片段化的DNA孵育MBD-珠:一般来说,输入DNA可以在5ng-1μg的范围内。对照反应通常使用1μg的K-562DNA。向洁净的1.7-ml无DNA酶的微量离心管中添加20μl的5X含有300mM NaCl的洗涤/结合缓冲液。将片段化的样品

DNA，例如5ng-1μg，添加至管中，并且用无DNA酶的水使最终体积达到100μl。将DNA/缓冲液混合物转移到含有MBD-珠的管中，并且在室温在旋转混合器上混合1小时。可替代地，混合物可以在4℃混合过夜。

[0348] 从珠溶液收集未捕获的DNA：未捕获的/非甲基化的DNA从DNA和MBD-珠的混合物收集。将含有DNA和MBD-珠的混合物的管放置在磁架上1分钟以浓缩珠，并且取出上清液并保存在洁净的无DNA酶的微量离心管中。该保存的上清液是未捕获的DNA上清液，并且可以储存在冰上。在旋转混合器上将珠用200μl的含有300mM NaCl的1x结合/洗涤缓冲液洗涤，持续3分钟。按上文描述的将珠浓缩，并且按上文描述的将含有未捕获的/非甲基化的/低甲基化的DNA的上清液取出、保存并储存在冰上。将珠洗涤、混合、浓缩，取出上清液并再保存一次以收集两次洗涤级分。将每次洗涤级分储存在冰上。可将这些洗涤级分合并在一起并相应地标记。

[0349] 洗脱捕获的DNA：使用含有2000mM NaCl的洗脱缓冲液洗脱捕获的DNA。将珠重悬在200μl的洗脱缓冲液(2000mM NaCl)中。将珠在旋转混合器上孵育3分钟，并放置在磁架上1分钟以浓缩所有珠，并且将含有捕获的/超甲基化的DNA的液体取出并保存在洁净的无DNA酶的微量离心管中。将保存的捕获的/甲基化的DNA的第一级分储存在冰上。将珠重悬并再次孵育，并且将含有捕获的/甲基化的DNA的液体取出并保存在第二个洁净的管中。将第一次收集和第二次收集的捕获的/甲基化的DNA合并并储存在冰上。

[0350] 用于分析的甲基化分级分离的DNA的制备：将分区的cfDNA，超甲基化DNA、中等甲基化DNA和非甲基化DNA例如通过SPRI珠清除法(Ampure XP,Beckman Coulter)来纯化，随后准备用于连接(使用NEBNext®Ultra™End Repair/dA-Tailing模块)，然后与含有非随机分子条形码的修饰的Y形dsDNA衔接子连接，如Lanman等人，2015描述的。超甲基化cfDNA分区、中等甲基化cfDNA分区和低甲基化cfDNA分区分别与11个、12个和12个不同的非随机分子条形码化衔接子连接。将每个样品的连接的、分区的cfDNA分子用SPRI珠(Ampure XP)再次纯化，然后与所有衔接子连接的分子通用的寡核苷酸(NEBNext Ultra II™Q5主混合物)重新组合到PCR反应中，一起扩增来自一个样品的所有cfDNA分子。将扩增的DNA文库再次使用SPRI珠(Ampure XP)纯化，为使用标准准备技术进行靶富集或全基因组测序(WGS)做准备。

[0351] 靶捕获和富集：可以使用商业上可获得的方案，例如用于Illumina多重测序的SureSelect^{XT}靶富集系统来富集DNA样品。

[0352] 实施例3：CDKN2A的甲基化特征分析

[0353] DNA甲基化特征分析联合片段组数据用于捕获CDKN2A基因中差异化地甲基的化区域(DMR)。CDKN2A基因是肿瘤抑制基因，其编码参与细胞周期调节的p16INK4A和p14ARF蛋白。使用MBD亲和纯化将cfDNA样品分级分离成低甲基化分区和超甲基化分区。在分级分离后，对每组中的核酸分子测序，以生成序列读段。当被映射到参考基因组时，序列读段提供片段组数据，然后该片段组数据与来自每个分级分离的分区的序列读段组合(图10)。CDKN2A基因示出低甲基化分区的覆盖度与超甲基化分区相比的总体增加。

[0354] 实施例4：正常样品和肺癌样品的甲基化特征谱

[0355] 如图11中示出的，将MBD分区方法应用于来自健康供体的四个cfDNA样品(正常13893、正常13959、正常13961、正常13962)和来自具有高ctDNA百分比的肺癌患者的两个

cfDNA样品(肺A1345402、肺A0516902),具有不同的输入量(10ng-150ng cfDNA)和重复(例如,3次重复)。按跨组中被靶向的所有基因组基因座的超甲基化DNA的百分比将样品分级聚类。超甲基化DNA的百分比可以通过将超甲基化无细胞DNA片段的数目除以跨所有分区观察到的无细胞DNA片段的总数来确定。组是一个定制的基因的组,其覆盖约30kb的基因组区域。该组对检测不同癌症,诸如肺癌、结肠直肠癌等也具有较高的灵敏度。来自健康供体的样品与来自肺癌患者的样品被单独聚类。各个肺癌样品具有不同的甲基化谱,这些甲基化谱被进一步单独聚类(即每个肺癌样品的重复品被正确地识别并分组在一起)。参见,例如WO 2017/181146,2017年10月19日。

[0356] 实施例5:使用全基因组测序的甲基化特征分析

[0357] 将DNA甲基化特征分析与片段组数据整合,以确定临床样品中异常的片段化模式,并且因此确定改变的染色质结构(图12A、图12B和图12C)。核酸分子来源于肺癌患者。使用MBD亲和纯化将核酸分子分级分离成低甲基化分区和超甲基化分区。在分级分离后,对每个分区中的核酸分子测序,以生成序列读段。当被映射到参考基因组时,序列读段提供片段组数据。将片段组数据,诸如基因组位置、片段长度和覆盖度,与来自每个分区的序列读段组合。如图12A和图12B中示出的,转录起始位点(TSS)的600bp区域在X轴上,并且频率或覆盖度在Y轴上指示。图12C示出了与X轴上的总片段和Y轴上的频率相比,超甲基化片段的分数。例如,在图12C中,总片段中的超甲基化片段的分数为约0.2(即约20%)。

[0358] 实施例6:MOB3A和WDR88的甲基化特征分析

[0359] 将DNA甲基化特征分析与片段组数据整合以确定表观遗传调节的差异(图13A和图13B)。使用MBD亲和纯化将核酸分子分级分离成低甲基化分区和超甲基化分区。在分级分离后,对每个分区中的核酸分子测序,以生成序列读段。这些序列读段当被映射到参考基因组时提供片段组数据。片段组数据,诸如基因组位置和覆盖度,与来自每个分级分离的组的序列读段组合。

[0360] MOB3A基因可能具有未知的生化功能,并且可能牵涉维持肿瘤生长和增殖。如图13A中的热图示出了在来自健康个体的样品中,超甲基化在与TSS的起始位点附近的覆盖度与低甲基化相比更高。该实施例提供了将分级分离的组与片段组数据组合应用,以检测可以指示癌症的基因的TSS中的标志物。这些数据示出,超甲基化和低甲基化的分级分离的组(或分区)为辨别跨一个基因组区域诸如TSS的甲基化状态提供了更好的分辨。如上文描述的,分级分离的组的覆盖度示出了跨TSS的甲基化状态的差异。该实施例提供了对核酸分子分级分离的应用,以提供对跨一个基因的甲基化状态的更好分辨。

[0361] WDR88基因可能牵涉细胞周期调节、凋亡和自噬。热图示出了在来自健康个体的样品中,超甲基化在与TSS的起始位点附近的覆盖度与低甲基化相比更高(图13B)。此外,图13B示出了,超甲基化和低甲基化的分级分离的组(或分区)为辨别跨一个基因组区域诸如TSS的甲基化状态提供了更好的分辨。如上文描述的,分级分离的组的覆盖度示出了跨TSS的甲基化状态的差异。该实施例提供了对核酸分子分级分离的应用,以提供对跨一个基因的甲基化状态的更好分辨。

[0362] 实施例7:重组的分区和未分级分离的样品的甲基化特征分析

[0363] 图14A分别在X轴和Y轴上示出了具有对来自未分级分离的组(无MBD)和MBD-亲和分区之后的重组的分区(总MBD)的覆盖度的热图。在分区为超甲基化分区和低甲基化分区

后,这些分区以计算机模拟重组,以形成“超+低”或“总MBD”。热图示出了对无MBD和总MBD的覆盖度之间的线性相关性。线性相关性指示相似的覆盖度,并且可以提供对跨一个基因组基因座的甲基化状态的相似分辨。无MBD和/或总MBD获得的分辨率水平可能不足以区分分辨基因座之间的甲基化状态的差异,表明了基于MBD-亲和进行分区的预料不到的优势。

[0364] 图14B示出了总MBD的MVA绘制热图。x轴示出了总MBD(重组的超甲基化分区和低甲基化分区)中的平均片段为(a+b)/2,其中a＝总MBD并且b＝无MBD。

[0365] 实施例8:重组的分区(总MBD)和未分级分离的样品之间的核小体组织方式

[0366] 如图15中示出的,在X轴上绘制了跨一个基因组区域的总MBD(计算机模拟重组的超甲基化分区和低甲基化分区)和无MBD(未分级分离的)样品的核小体占据中心之间的距离差异。在Y轴上(如以“密度”指示的)绘制了跨一个基因组区域的总MBD和无MBD样品的核小体占据中心之间的距离分布的差异。总MBD样品通过计算机模拟重新组合超甲基化分区和低甲基化分区来制备。这些结果表明,MBD分区不影响核小体占据。

[0367] 实施例9:MBD信号的验证

[0368] MBD分区的样品用于辨别健康样品和癌症样品的核小体占据。在该实施例中,获得了来自六名肺癌患者和三名无恶性疾病的健康成人的血液样品。从这些样品提取无细胞核酸,并且使用MBD-亲和纯化将其分区成超甲基化分区和低甲基化分区。使用全基因组测序对核酸样品测序。确定对于每个分区和对于所有样品的超甲基化片段的百分比。图16示出了来自肺癌患者(从上部第1行和第2行)和来自健康成人(第3行和第4行)的超甲基化分区和低甲基化分区中的MBD信号。如图16中示出的,与来自健康个体的超甲基化分区相比,来自肺癌患者的无细胞DNA片段在超甲基化分区(LungSigHyper)中示出基因内区域远端的富集。此外,前5%最高百分比的超甲基化峰(LungSigHyper)和低甲基化峰(LungSigHypo)中的特征分布示出了除外显子1之外的所有外显子的中低甲基化峰的显著富集(图16,第1行和第2行)。

[0369] 实施例10:AP3D1基因的甲基化特征分析

[0370] 本文描述的方法用于肺癌的预后。在一项实验中,使用MBD-亲和纯化将来自肺癌患者的具有核酸分子的样品分级分离成低甲基化分区和超甲基化分区。作为对照,一个样品未被分区(无MBD)。使用全基因组测序对样品测序。

[0371] AP3D1基因可以编码可能牵涉细胞器运输的AP3复合亚基δ-1。热图示出了,超甲基化分区在TSS附近的覆盖度与低甲基化分区和/或无MBD相比更大(图18A)。超甲基化分区示出了比无MBD组更强和/或更局部的覆盖度。如热图中示出的,超甲基化分区在TSS附近具有非常局部化的强覆盖度,而无MBD组跨基因组区域具有相似的覆盖度。还确定了平均超甲基化百分比,如图18B中的红线所示。该实施例可以提供对核酸分子分级分离的应用,以提供对跨一个基因的甲基化状态的更好分辨。这些结果示出了AP3D1基因是超甲基化的,尤其是在TSS附近(图18A),并且AP31基因是超甲基化的(>60%,如图18B中示出的)。AP3D1基因的去调控(Deregulation)可能参与了导致肺癌。因此,该实施例通过监测个体的甲基化谱可以提供该方法在肺癌的预后中的应用。

[0372] 实施例11:DNMT1基因的甲基化特征分析

[0373] 在另一个实施例中,检查DNMT1基因的甲基化特征分析。DNMT1基因编码催化甲基基团转移到DNA中的特定CpG二核苷酸的酶。DNMT1牵涉DNA甲基化的维持,以确保遗传的表

观遗传模式的复制的保真度。异常甲基化模式可能与癌症和发育异常有关。

[0374] 示出了关于TSS的超甲基化、低甲基化和无MBD的热图(图19A)。超甲基化分区示出了比无MBD组更强和/或更局部的覆盖度。超甲基化分区在TSS附近具有局部的和更强的覆盖度,而无MBD组具有跨该基因的相似的覆盖度。还确定了平均超甲基化百分比,如图19B中的红线所示的为约75%。这些结果示出了DNMT1基因是超甲基化的,尤其是在TSS附近(图19A),并且DNMT1基因是超甲基化的(约75%,如图19B中示出的)。异常甲基化模式连同染色质结构的变化可能导致DNMT1的去调控,这可能参与了导致肺癌。因此,该实施例通过监测个体的甲基化谱可以提供该方法在肺癌的预后中的应用。该实施例还可以提供对核酸分子分级分离的应用,以提供对跨一个基因的甲基化状态的更好分辨。

[0375] 实施例12:修饰的组蛋白的分级分离

[0376] 该实施例展示了使用修饰的组蛋白方式的分区。基于组蛋白修饰来对DNA分区。简言之,将琼脂糖珠用BSA封闭,并且在洗涤后,将珠与抗H3K9me3和H4K20me3的抗体(Millipore,Temecula,CA,USA)在4℃预孵育4小时。随后,将200μl的血浆稀释到800μl的分区稀释缓冲液中,并且然后添加至与抗体预孵育的沉淀的琼脂糖珠中。在4℃孵育过夜后,用低盐、高盐、LiCl和Tris/EDTA缓冲液洗涤珠。最后,通过在65℃孵育珠来洗脱染色质,并且通过用蛋白酶K处理来去除蛋白质。然后使用合适的纯化试剂盒纯化分区的DNA,并在-20℃储存。

[0377] 实施例13:基于蛋白质结合区域的分级分离

[0378] 该实施例展示了使用蛋白质结合区域的分区方法。基于与蛋白质A结合的差异来对DNA分区。样品中的核酸分子也可以基于蛋白质结合区域被分级分离。例如,核酸分子可以基于与特定蛋白质结合的核酸分子和不与该特定蛋白质结合的核酸分子被分级分离成不同的组。核酸分子可以基于DNA-蛋白质结合被分级分离。蛋白质-DNA复合物可以基于蛋白质的特定性质被分级分离。这样的性质的实例包括不同表位、修饰(例如组蛋白甲基化或乙酰化)或酶促活性。可以结合DNA并用作用于分级分离的基础的蛋白质的实例可以包括例如蛋白质A或蛋白质G。实验程序,诸如染色质-免疫-沉淀用于基于蛋白质A结合区域来对核酸分子分级分离。

[0379] 实施例14:基于羟甲基化的分级分离

[0380] 该实施例展示了使用修饰的组蛋白方式的分区。基于羟甲基化来对DNA分区。简言之,5-hmC修饰的碱基在体外糖基化。5-hmC的特异性糖基化通过遵循来自Zymo Research的高活性5-hmC糖基转移酶(zymoresearch.com/epigenetics/dna-hydroxymethylation/5-hmc-glucosyltransferase)的方案来实现。J-结合蛋白-1(JBP-1)以高亲和力特异地结合糖基化的DNA,允许通过基于JBP-1的富集确定5-hmC水平。此外,5-hmC的糖基化改变了若干限制性内切酶对DNA的消化,并且因此5-hmC-糖基化的DNA的消化模式可以用于评估DNA羟甲基化状态。

[0381] 实施例15:基于核酸分子的链型的分级分离

[0382] 将样品中的核酸分子基于链型来分级分离。例如,将ssDNA和dsDNA分级分离成两组。这些组单独地或同时地经历测序测定。通过在分级分离期间不使样品经历变性步骤对具有ssDNA和dsDNA两者的核酸样品进行分级分离。变性步骤将dsDNA转化为ssDNA,并且不允许基于链型的对核酸分子的分级分离。

[0383] 实施例16:用改进的预扩增靶捕获方案(NEBNext Direct)对ssDNA和dsDNA进行分子分区

[0384] 一种新型的杂交捕获方法,其中将预扩增杂交捕获靶测序方案(例如,NEBNext Direct HotSpot Cancer Panel)应用于无细胞DNA(cfDNA)样品,而没有DNA变性,捕获了ssDNA分子(图18)。将含有dsDNA分子的未结合的级分分离,变性至ssDNA并应用于捕获方案。

[0385] 所使用的预扩增杂交捕获测序方案是NEBNext Direct HotSpot Cancer Panel,包含来自50个基因的190个常见癌症靶标的诱饵,涵盖约40kb的序列,并且包含超过18,000个COSMIC特征(NEBNext Direct HotSpot Cancer Panel;neb.com/products/e7000-nebnext-direct-cancer-hotspot-panel)。简言之,NEBNext Direct靶富集方法快速将DNA样品与生物素化的寡核苷酸诱饵杂交,寡核苷酸诱饵界定每个感兴趣的靶的3'末端。诱饵-靶杂交体与链霉亲和素珠结合,并且使用酶促反应去除3'脱靶序列。后续的文库制备将靶转化为Illumina兼容的文库,该文库包括分子标签和样品条形码。试剂盒的使用允许通过使DNA样品变性来捕获样品中的所有ssDNA和dsDNA分子,然后与诱饵杂交。

[0386] 使含有ss-cfDNA和ds-cfDNA的cfDNA样品经历靶捕获方案,省略了前期的dsDNA变性步骤。通过NEBNext方案为NGS准备捕获的ssDNA分子(图20中的左栏),同时将来自捕获物的上清液应用于第二靶捕获方案,具有标准的前期dsDNA变性步骤并且随后准备用于NGS(图20中的右栏)。从血浆提取的cfDNA通过基于电泳的测量来定量。将等同于200ng或500ng的样品体积应用于NEBNExt Direct HotSpot Cancer Panel测定,省略DNA变性步骤,使得仅ssDNA分子与诱饵杂交。保留含有dsDNA分子和未被靶向的ssDNA分子的捕获物的上清液,并进行第二次靶捕获(图20)。ssDNA文库和dsDNA文库两者被单独地制备以用于NGS,所述文库具有在下游生物信息学分析中被识别的独特的样品条形码标签。在Illumina NextSeq 500(2x 75配对末端)上对ssDNA和dsDNA制备的文库测序,并且计算中靶分子(对应于40kb诱饵)的总数(图1)。

[0387] 使用上文描述的方法,将具有单链无细胞DNA(ss-cfDNA)和双链无细胞DNA(ds-cfDNA)两者的无细胞DNA(cfDNA)样品分别分级分离成ss-cfDNA组和ds-cfDNA组(图20)。在两个测序样品中,ssDNA文库包含～80％的dsDNA(中靶分子,第一个200ng和第二个500ngcfDNA输入)。第二个200ng cfDNA未能产生ssDNA文库和dsDNA文库,并且指示在ssDNA/dsDNA分区过程上游的样品处理中的可能的错误,而第一个500ng cfDNA输入仅产生显著的dsDNA文库,这表明cfDNA样品中的ssDNA和dsDNA的相对量是可变的。按照来自Broad Institute的Picard包(Picard metrics;broadinstitute.github.io/picard/picard-metric-definitions.html)所定义的计算中靶分子。该实验的PCR收率在图20中示出。对于所有四个样品,相对收率,ssDNA的PCR收率/dsDNA的PCR收率,被确定为在20％和75％之间。

[0388] 实施例17:用基于MBD的甲基化分区方法保持了敏感的体细胞突变检测

[0389] 样品收集和合并

[0390] 样品从示出了高cfDNA收率的Guardant Health repository中选择。临床样品通过混合96个等体积的样品来制备。这作为用于突变检测的测定灵敏度的测试材料,因为池含有来自参考基因组的<0.02％至100％的突变。制备了两种不同的临床样品(PowerpoolV1和PowerpoolV2)和独特的组分样品(component sample)。

[0391] DNA分区

[0392] 将Powerpool cfDNA分区成多个级分。使用MethylMiner™亲和富集方案(Thermo Fisher Scientific,目录#ME10025)将cfDNA(15ng或150ng)分区成超甲基化级分、中等甲基化级分和低甲基化级分,不同之处在于使用300mM NaCl孵育和洗涤缓冲液对反应条件进行了修改,并且按比例缩小一微克DNA输入的方案用于亚微克DNA输入。

[0393] 珠制备

[0394] 洗涤Dynabeads®M-280链霉亲和素

[0395] Dynabeads®M-280链霉亲和素在与MBD-生物素蛋白偶联之前,使用1X结合/洗涤缓冲液(含有160mM NaCl)洗涤。简言之,将Dynabeads®M-280链霉亲和素的原液重悬以获得均匀的悬浮液。对于每微克的输入DNA,将10μl的珠添加至1.7-ml无DNA酶的微量离心管中。用1X结合/洗涤缓冲液使珠的体积达到100μl。将管放置在磁架上1分钟,以在去除和丢弃液体之前将所有珠浓缩在管的内壁上。将管从磁架取出,并且添加等体积(例如,约100-250μl)的1X结合/洗涤缓冲液以重悬珠。将重悬的珠再次浓缩并洗涤,然后进行MBD-生物素蛋白与珠的偶联。

[0396] 将**Dynabeads®**M-280链霉亲和素与MBD-生物素蛋白偶联

[0397] 对于每微克的输入DNA,将7μl(3.5μg)的MBD-生物素蛋白添加至1.7-ml无DNA酶的微量离心管中。用含有300mM NaCl的1X结合/洗涤缓冲液使珠的体积达到100μl。稀释MBD-生物素蛋白,并从最初的珠洗涤液转移到重悬的珠的管中。将珠-蛋白质混合物在室温在旋转混合器上混合1小时,然后进行洗涤MBD-珠。

[0398] 洗涤MBD-珠

[0399] 通过将MBD-珠放置在磁架上1分钟来浓缩含有MBD-珠的管。去除并丢弃液体。将珠用100-250μl的含有160mM NaCl的1X结合/洗涤缓冲液重悬,并且在室温在旋转混合器上混合持续5分钟。按上文描述的将珠再浓缩、洗涤和重悬两次。然后将管放置在磁架上1分钟,并且小心地去除并丢弃液体。重悬珠,每μl所使用的链霉亲和素珠用10μl1X DNA捕获缓冲液(含有300mM NaCl)。

[0400] 捕获MBD-珠上的片段化的甲基化DNA

[0401] 用片段化的DNA孵育MBD-珠

[0402] 一般来说,输入DNA可以在5ng-1μg的范围内。对照反应通常使用1μg的K-562DNA。向洁净的1.7-ml无DNA酶的微量离心管或PCR管添加片段化的样品DNA,例如5ng-1μg,与等体积的2xDNA捕获缓冲液(含有300mM NaCl),并且用1xDNA捕获缓冲液使最终体积达到100μl或200μl。将DNA/缓冲液混合物转移到含有MBD-珠的管中,并且在室温在旋转混合器上混合1小时。可替代地,混合物可以在4℃混合过夜。

[0403] 从珠溶液收集未捕获的DNA

[0404] 未捕获的/非甲基化的DNA从DNA和MBD-珠的混合物收集。简言之,将含有DNA和MBD-珠的混合物的管放置在磁架上1分钟以浓缩所有珠,并且取出上清液并保存在洁净的无DNA酶的微量离心管中。保存的上清液是未捕获的DNA上清液/非甲基化的DNA级分,并且可以储存在冰上。在旋转混合器上将珠用200μl的含有300mM NaCl的1X DNA捕获缓冲液洗涤,持续3分钟。按上文描述的将珠浓缩,并且按上文描述的将含有未捕获的/非甲基化的/

低甲基化的DNA的上清液取出、保存并储存在冰上。将珠洗涤、混合、浓缩，取出上清液并再保存一次以收集两次洗涤级分。将每次洗涤级分储存在冰上。可将这些洗涤级分合并在一起并相应地标记。

[0405] 洗脱捕获的DNA

[0406] 使用含有2000mM NaCl的洗脱缓冲液洗脱捕获的DNA。将珠重悬在200μl的洗脱缓冲液(2000mM NaCl)中。将珠在旋转混合器上孵育3分钟，并放置在磁架上1分钟以浓缩所有珠，并且将含有捕获的/超甲基化的DNA的液体取出并保存在洁净的无DNA酶的微量离心管中。将保存的第一次级分储存在冰上。将珠重悬并再次孵育，并且将含有捕获的/甲基化的DNA的液体取出并保存在第二个洁净的管中。将第一次收集和第二次收集的捕获的/超甲基化的DNA合并并储存在冰上。可替代地，可以进行NaCl浓度递增的多次洗脱，以进一步将DNA分区成具有递增的DNA甲基化的级分。

[0407] 用于分析的甲基化分级分离的DNA的制备

[0408] 将分区的cfDNA，甲基化DNA、中等甲基化DNA和非甲基化DNA例如通过SPRI珠清除法(Ampure XP，Beckman Coulter)来纯化，随后准备用于连接(使用NEBNext® Ultra™End Repair/dA-Tailing模块)，然后与含有非随机分子条形码的修饰的Y形dsDNA衔接子连接，如Lanman等人，2015描述的。超甲基化cfDNA分区、中等甲基化cfDNA分区和低甲基化cfDNA分区分别与11个、12个和12个不同的非随机分子条形码化衔接子连接。将每个样品的连接的、分区的cfDNA分子用SPRI珠(Ampure XP)再次纯化，然后与所有衔接子连接的分子通用的寡核苷酸(NEBNext Ultra II™Q5主混合物)重新组合到PCR反应中，一起扩增来自一个样品的所有cfDNA分子。将扩增的DNA文库使用SPRI珠(Ampure XP)再次纯化，为通过杂交捕获的靶富集(Agilent SureSelect 30kb panel；'panel')做准备。

[0409] 用于分析的未分区的DNA的制备

[0410] 制备用于连接的Powerpool cfDNA(10ng或150ng)(使用NEBNext®Ultra™End Repair/dA-Tailing模块)，然后与含有非随机分子条形码的修饰的Y形dsDNA衔接子连接，如Lanman等人，2015描述的。将cfDNA与35个不同的、非随机的分子条形码化衔接子连接。将每个样品的连接的cfDNA分子用SPRI珠(Ampure XP)再次纯化，然后与所有衔接子连接的分子通用的寡核苷酸(NEBNext Ultra II™Q5主混合物)放置到PCR反应中，一起扩增来自一个样品的所有cfDNA分子。将扩增的DNA文库使用SPRI珠(Ampure XP)再次纯化，为通过杂交捕获的靶富集(Agilent SureSelect 30kb panel；'panel')做准备。

[0411] 本公开内容提供了用于处理包含不同形式(例如，RNA和DNA，单链的或双链的)和/或修饰程度(例如，胞嘧啶甲基化，与蛋白质缔合)的核酸群体的方法。这些方法适应样品中核酸的多种形式和/或修饰，使得可以获得针对多种形式的序列信息。所述方法经过处理和分析仍保持多种形式或修饰状态的身份，使得序列分析可以与表观遗传分析组合。

[0412] 数据分析

[0413] 将来自不同样品的DNA文库合并并在Illumina HiSeq2500，2x150配对末端测序仪上测序。生物信息学处理按照Lanman等人，2015和其他地方描述的标准GUARDANT360™方案进行。对于MBD分区的样品，另外使用分子条形码来识别其中DNA被分级分离(超甲基化、中等甲基化和低甲基化)的MBD分区。在被组靶向的每个基因组基因座处，对齐的超甲基化、中等甲基化和低甲基化的分子被总计。%超甲基化被定义为在给定的基因座，跨该基因座的

被超甲基化的的总分子的分数。对于MBD分区的DNA样品和未分区的DNA样品两者，在靶向区域，使用拥有知识产权的Guardant Health变体判别软件(variant calling software)来判别来自参考基因组的突变等位基因分数(MAF)。

[0414] 实施例18:靶向测序测定中MBD样品和无MBD样品的覆盖度之间的比较

[0415] 在该实施例中，样品按实施例17中描述的处理。不同的cfDNA临床样品(PowerpoolV1和PowerpoolV2)分别一式三份地在有和没有MBD分区('MBD'和'无MBD')的靶向测序测定中测定。对于在15ng(图25A)和150ng(图25B)测定输入的PowerpoolV1，比较了MBD和无MBD中来自该组的基因在每个靶向基因组位置测序的独特分子。组是一个定制的基因的组，其覆盖约30kb基因组区域。该组对检测不同癌症，诸如肺癌、结肠直肠癌等也具有较高的灵敏度。图25A和图25B示出了在应用MBD分区的情况下，对靶向测序测定中的分子的高效回收被保持。powerpoolV1(a)15ng和(b)150ng输入的靶向测序测定中的分子计数在MBD分区(Y轴)或无MBD分区的情况下运行。在MBD和无MBD分子计数或覆盖度之间观察到线性相关性，指示MBD分区并不影响测定的回收。

[0416] 比较了无MBD样品和MBD样品之间来自该组的基因的分子计数或覆盖度。使用从两个临床样品(图26A-PowerpoolV1和图26B-PowerpoolV2)提取的15ng输入cfDNA或使用从两个临床样品(图27A-PowerpoolV1;图27B-PowerpoolV2)提取的150ng输入cfDNA来制备MBD样品和无MBD样品。左侧的图的X轴代表分子计数或覆盖度，中间的图的X轴代表用两个配对末端读段确认的突变体(双链重叠;DSO)，并且右侧的图的X轴代表对两个DNA链测序的分子计数(双链支持;DS)。对于分子计数、DSO和DS，MBD样品和无MBD样品之间的强相关性示出，与无MBD相比，MBD样品可以捕获大多数分子(图26A中～94％，图26B和图27A中～80％-85％，以及图27B中～90％)。跨该组没有分子覆盖度的位置偏差，也没有其他重要的变体判别度量(variant calling metrics)(DSO,DS)。

[0417] 实施例19:MBD样品和无MBD样品的变体检测的灵敏度和特异性

[0418] 在该实施例中，样品按实施例17中描述的处理。为了测量对变体或突变检测的灵敏度和特异性的影响，使用15ng输入cfDNA针对组中的基因比较了MBD(Y-轴)样品和无-MBD(X-轴)样品之间的突变等位基因分数(MAF)。将不同的MAF范围例如，0-100％(图28A)、0-5％(图28B)和0-0.5％(图28C)绘制在X轴上。MAF值来自MBD和无MBD的一式三份样品。针对MBD样品确定的MAF与针对无MBD样品确定的MAF一致。对于具有15ng输入(图28A;0-100％)和处于检测下限(图28B;0-5％)的PowerpoolV1，MBD和无MBD之间的MAF示出了线性相关性。MBD和无MBD之间的MAF在检测限以下不良好相关(图28C;0-0.5％MAF)。类似地，在来自PowerpoolV1的150ng cfDNA输入的情况下，MBD样品和无MBD样品示出了MAF方面的一致(图29A和图29B)，但是在0-0.5％范围内不存在强一致(图29C)。

[0419] 实施例20:使用全基因组测序对启动子区域的甲基化特征分析

[0420] 分子分区的样品可以增强基因组结构的分析，诸如无细胞DNA片段占据和癌症的检测。例如，在分析通常经由甲基化驱动的基因沉默而被癌症靶向的肿瘤抑制基因的启动子区域时，可以通过考虑无细胞DNA片段占据来检测转录相关的超甲基化事件。人们可以共同检查不同MBD分区中的无细胞DNA片段占据信号和超甲基化分数，以验证MBD驱动的发现癌症样品中的转录相关的超甲基化事件和基因沉默的可行性。

[0421] 作为说明性实例，人们可以使用公众可获得的基因编码(gencode)(v26lift37)数

据来产生可用的无恶性疾病的健康成人群组中的所有基因编码基因的TSS区域中的超甲基化百分比(超甲基化分区中的片段的数目/所有MBD分区中的片段的总数)。无细胞DNA片段占据信号可以遍布无恶性疾病的健康成人群组聚集。所有TSS可以基于MBD分区测定中观察到的超甲基化级分的百分比来分箱元。可以检查每个箱元中的无MBD WSG群组中的片段占据。图23示出了基因表达和甲基化状态的相关性。示出了启动子特征谱的WGS占据与MBD甲基化的百分比。如图23看到的,在TSS附近,低甲基化DNA(0-0.1%超甲基化)具有低的片段占据覆盖度,而超甲基化DNA(10-50%超甲基化或>50%超甲基化)在TSS附近具有高的片段占据覆盖度和明显的NDR。在一些情况下,低甲基化DNA的片段占据覆盖度被用来将序列深度和/或序列的可映射性归一化。超甲基化核酸片段或低甲基化核酸片段的百分比可以通过将超甲基化无细胞片段或低甲基化无细胞片段的数目除以遍布所有分区观察到的无细胞DNA片段的总数来确定。

[0422] 实施例21:MBD样品和全基因组亚硫酸氢盐测序(WGBS)样品的甲基化水平的比较

[0423] 为了评估通过使用MBD方案制备的不同分区中的片段的甲基化水平,使用良好表征的样品,NA12878(catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=GM12878)。将样品分区成超甲基化分区、低甲基化分区和中等甲基化分区,然后按实施例1中描述的计算机模拟重新组合这些分区(MBD样品)。将MBD样品与公众可获得的利用全基因组亚硫酸氢盐测序(WGBS)的标准甲基化数据集(basespace.illumina.com/datacentral(HiSeq 4000:TruSeq DNA Methylation(NA12878,1x151)比较。WGBS询问单个胞嘧啶的甲基化状态。图31示出了在160bp窗口中通过WGBS(X轴)和MBD(Y轴)测量的平均甲基化水平的相关性。MBD甲基化水平通过将落入该窗口中的超甲基化分区的读段的数目除以超甲基化分区和低甲基化分区中的读段的总数来计算。WGBS甲基化水平通过将窗口中的甲基化碱基的数目除以甲基化和未甲基化碱基的数目来计算。该实验以若干不同的珠比率运行,这影响甲基化片段的分区。较少的珠将超甲基化分区局限于高度甲基化的片段(即,使测定对甲基化更特异),而较多的珠减少了片段进入超分区所需的甲基化的量(即,使测定对甲基化更敏感)。经验地,发现1:50的输入DNA:珠的比在分区的片段与其甲基化水平之间相关。这些结果指示,MBD分区确实准确地反映了样品根本的甲基化状态。

[0424] 在该分析中,确定了片段中CG位点的数目对该片段的分区的影响。公众用标准甲基化数据集(NA12878;与先前分析中相同)可获得的指示高度超甲基化或低甲基化(全基因组亚硫酸氢盐测序甲基化水平>90%或<10%,如先前的分析中计算的)的片段被选择用于分析。这些片段按照它们所包含的CG位点的数目来分层。具有3个或更多个CG位点的高度甲基化的片段最终进入超甲基化分区,指示该测定对少量甲基化敏感(图31A)。相反,不管片段中CG位点的数目如何,没有甲基化的片段主要被分区到低甲基化分区,指示测定具有高特异性程度(图31B)。

[0425] 虽然在本文中已经示出和描述了本公开内容的优选的实施方案,但是对于本领域技术人员明显的是,这样的实施方案仅以示例的方式提供。本领域技术人员现将想到不偏离本公开内容的许多改变、变化和替换。应当理解,在实践本公开内容时可以采用本文描述的本公开内容的实施方案的各种替代方案。以下权利要求意图界定本公开内容的范围,并且从而涵盖在这些权利要求范围内的方法和结构及其等同物。

[0426] 本发明的一些实施方案。

[0427] 下文提供了以专利权利要求形式提供的本发明的一些实施方案。

[0428] 1.一种分析核酸群体的方法,所述核酸群体包含选自双链DNA、单链DNA和单链RNA的至少两种形式的核酸,其中所述至少两种形式中的每一种包含多于一个分子,所述方法包括:

[0429] (a)将所述形式的核酸中至少一种与至少一种标签核酸连接,以将所述形式彼此区分开,

[0430] (b)扩增所述形式的核酸,其中的至少一种被连接至至少一种核酸标签,其中所述核酸和连接的核酸标签被扩增,以产生扩增的核酸,其中从所述至少一种形式扩增的核酸被加标签;

[0431] (c)测定所述扩增的核酸的序列数据,所述扩增的核酸中的至少一些被加标签;其中所述测定获得足以解码所述扩增的核酸的所述标签核酸分子的序列信息,以揭示所述群体中为已测定了针对其的序列数据的连接至所述标签核酸分子的所述扩增的核酸提供原始模板的核酸的形式。

[0432] 1A.如权利要求1所述的方法,还包括以下步骤:解码所述扩增的核酸的所述标签核酸分子,以揭示为已测定了针对其序列数据的连接至所述标签核酸分子的所述扩增的核酸提供原始模板的群体中的核酸的形式。

[0433] 2.如权利要求1所述的方法,还包括相对于一种或更多种其它形式富集所述形式中的至少一种。

[0434] 3.如权利要求1所述的方法,其中所述群体中的每种形式的核酸的至少70%的分子在步骤(b)中被扩增。

[0435] 4.如权利要求1所述的方法,其中所述群体中存在至少三种形式的核酸,并且所述形式中的至少两种连接至不同的标签核酸形式,所述不同的标签核酸形式将所述三种形式中的每一种彼此区分开。

[0436] 5.如权利要求4所述的方法,其中所述群体中的所述至少三种形式的核酸中的每一种连接至不同的标签。

[0437] 6.如权利要求1所述的方法,其中相同形式的每个分子连接至包含相同识别信息标签的标签。

[0438] 7.如权利要求1所述的方法,其中相同形式的分子连接至不同类型的标签。

[0439] 8.如权利要求1所述的方法,其中步骤(a)包括:使所述群体经历用加标签的引物的逆转录,其中所述加标签的引物被掺入从所述群体中的RNA产生的cDNA中。

[0440] 9.如权利要求8所述的方法,其中所述逆转录是序列特异性的。

[0441] 10.如权利要求8所述的方法,其中所述逆转录是随机的。

[0442] 11.如权利要求8所述的方法,还包括降解与所述cDNA成双链的RNA。

[0443] 12.如权利要求4所述的方法,还包括分离单链DNA和双链DNA,和将核酸标签连接至所述双链DNA。

[0444] 13.如权利要求12所述的方法,其中所述单链DNA通过与一种或更多种捕获探针杂交来分离。

[0445] 14.如权利要求4所述的方法,还包括用环连接酶使单链DNA环化,和将核酸标签连接至所述双链DNA。

[0446] 15.如权利要求1所述的方法，包括在测定之前，合并包含不同形式的核酸的加标签的核酸。

[0447] 16.如任一前述权利要求所述的方法，其中所述核酸群体来自体液样品。

[0448] 17.如权利要求16所述的方法，其中所述体液样品是血液、血清或血浆。

[0449] 18.如权利要求1所述的方法，其中所述核酸群体是无细胞核酸群体。

[0450] 19.如权利要求17所述的方法，其中所述体液样品来自疑似具有癌症的受试者。

[0451] 20.如权利要求1-19所述的方法，其中所述序列数据指示体细胞变体或生殖系变体的存在。

[0452] 21.如权利要求1-20所述的方法，其中所述序列数据指示拷贝数变异的存在。

[0453] 22.如权利要求1-21所述的方法，其中所述序列数据指示单核苷酸变异(SNV)、插入缺失或基因融合的存在。

[0454] 23.一种分析包含具有不同修饰程度的核酸的核酸群体的方法，包括：

[0455] 使所述核酸群体与优先结合带有所述修饰的核酸的剂接触，

[0456] 分离与所述剂结合的第一核酸池和未与所述剂结合的第二核酸池，其中所述第一核酸池对于所述修饰是呈现过度的，并且所述第二池中的核酸对于所述修饰是呈现不足的；

[0457] 将所述第一池和/或所述第二池中的核酸连接至区分所述第一池和所述第二池中的核酸的一个或更多个核酸标签，以产生加标签的核酸的群体；

[0458] 扩增标记的核酸，其中所述核酸和连接的标签被扩增；

[0459] 测定扩增的核酸和连接的标签的序列数据；其中所述测定获得用于解码所述标签的序列数据，以揭示已测定了针对其的序列数据的核酸是从所述第一池中的模板扩增的还是从所述第二池中的模板扩增的。

[0460] 23A.如权利要求23所述的方法，包括以下步骤：解码所述标签以揭示已测定了针对其的序列数据的核酸是从所述第一池中的模板扩增的还是所述第二池中的模板扩增的。

[0461] 24.如权利要求23所述的方法，其中所述修饰是将核酸与蛋白质结合。

[0462] 25.如权利要求23所述的方法，其中所述蛋白质是组蛋白或转录因子。

[0463] 26.如权利要求23所述的方法，其中所述修饰是对核苷酸的复制后修饰。

[0464] 27.如权利要求26所述的方法，其中所述复制后修饰是5-甲基-胞嘧啶，并且所述剂与核酸的结合程度随着所述核酸中5-甲基-胞嘧啶的程度而增加。

[0465] 28.如权利要求26所述的方法，其中所述复制后修饰是5-羟甲基-胞嘧啶，并且所述剂与核酸的结合程度随着所述核酸中5-羟甲基-胞嘧啶的程度而增加。

[0466] 29.如权利要求26所述的方法，其中所述复制后修饰是5-甲酰基-胞嘧啶或5-羧基-胞嘧啶，并且所述剂的结合程度随着所述核酸中5-甲酰基-胞嘧啶或5-羧基-胞嘧啶的程度而增加。

[0467] 30.如权利要求23所述的方法，还包括洗涤与所述剂结合的核酸，并且将洗涤物收集为第三池，所述第三池包括相对于所述第一池和所述第二池具有中等程度的复制后修饰的核酸。

[0468] 31.如权利要求23所述的方法，包括在测定之前，合并来自所述第一池和所述第二池的加标签的核酸。

[0469] 32.如权利要求23所述的方法，其中所述剂是5-甲基-结合结构域磁珠。

[0470] 33.如任一前述权利要求所述的方法，其中所述核酸群体来自体液样品。

[0471] 34.如权利要求33所述的方法，其中所述体液样品是血液、血清或血浆。

[0472] 35.如权利要求23所述的方法，其中所述核酸群体是无细胞核酸群体。

[0473] 36.如权利要求33所述的方法，其中所述体液样品来自疑似具有癌症的受试者。

[0474] 37.如权利要求23-36所述的方法，其中所述序列数据指示体细胞变体或生殖系变体的存在。

[0475] 38.如权利要求23-37所述的方法，其中所述序列数据指示拷贝数变异的存在。

[0476] 39.如权利要求23-38中任一项所述的方法，其中所述序列数据指示单核苷酸变异(SNV)、插入缺失或基因融合的存在。

[0477] 40.一种分析核酸群体的方法，其中至少一些核酸包括一个或更多个修饰的胞嘧啶残基，所述方法包括

[0478] 将捕获部分连接至所述群体中的核酸，所述核酸用作用于扩增的模板；

[0479] 进行扩增反应以从所述模板产生扩增产物；

[0480] 分离连接至捕获标签的模板和扩增产物；

[0481] 通过亚硫酸氢盐测序测定连接至捕获标签的模板的序列数据；和

[0482] 测定所述扩增产物的序列数据。

[0483] 41.如权利要求40所述的方法，其中所述捕获部分包括生物素。

[0484] 42.如权利要求41所述的方法，其中所述分离通过使模板与链霉亲和素珠接触来进行。

[0485] 43.如权利要求40所述的方法，其中修饰的胞嘧啶残基是5-甲基-胞嘧啶、5-羟甲基胞嘧啶、5-甲酰基胞嘧啶或5-羧基胞嘧啶。

[0486] 44.如权利要求40所述的方法，其中所述捕获部分包含生物素，所述生物素连接至包含一个或更多个修饰的残基的核酸标签。

[0487] 45.如权利要求40所述的方法，其中所述捕获部分经由可裂解的连接体连接至所述群体中的核酸。

[0488] 46.如权利要求45所述的方法，其中所述可裂解的连接体是可光裂解的连接体。

[0489] 47.如权利要求45所述的方法，其中所述可裂解的连接体包含尿嘧啶核苷酸。

[0490] 48.如任一前述权利要求所述的方法，其中所述核酸群体来自体液样品。

[0491] 49.如权利要求48所述的方法，其中所述体液样品是血液、血清或血浆。

[0492] 50.如权利要求40所述的方法，其中所述核酸群体是无细胞核酸群体。

[0493] 51.如权利要求48所述的方法，其中所述体液样品来自疑似具有癌症的受试者。

[0494] 52.如任一前述权利要求所述的方法，其中所述序列数据指示体细胞变体或生殖系变体的存在。

[0495] 53.如任一前述权利要求所述的方法，其中所述序列数据指示拷贝数变异的存在。

[0496] 54.如任一前述权利要求所述的方法，其中所述序列数据指示单核苷酸变异(SNV)、插入缺失或基因融合的存在。

[0497] 55.一种分析包含具有不同程度的5-甲基化的核酸的核酸群体的方法，包括：

[0498] (a)使所述核酸群体与优先结合5-甲基化的核酸的剂接触；

[0499] (b)分离与所述剂结合的第一核酸池和未与所述剂结合的第二核酸池,其中所述第一核酸池对于5-甲基化是呈现过度的,并且所述第二池中的核酸对于5-甲基化是呈现不足的;

[0500] (c)将所述第一池和/或所述第二池中的核酸连接至区分所述第一池和所述第二池中的核酸的一个或更多个核酸标签,其中连接至所述第一池中的核酸的核酸标签包括捕获部分(例如,生物素);

[0501] (d)扩增标记的核酸,其中所述核酸和连接的标签被扩增;

[0502] (e)分离带有所述捕获部分的扩增的核酸和不带有所述捕获部分的扩增的核酸;和

[0503] (f)测定分离的、扩增的核酸的序列数据。

[0504] 56.一种分析包含具有不同修饰程度的核酸的核酸群体的方法,包括:

[0505] 使所述群体中的核酸与衔接子接触,以产生侧翼为包含引物结合位点的衔接子的核酸群体;

[0506] 从所述引物结合位点引发衔接子,扩增侧翼为所述衔接子的核酸;

[0507] 使扩增的核酸与优先结合带有修饰的核酸的剂接触,

[0508] 分离与所述剂结合的第一核酸池和未与所述剂结合的第二核酸池,其中所述第一核酸池对于所述修饰是呈现过度的,并且所述第二池中的核酸对于所述修饰是呈现不足的;

[0509] 对所述第一池和所述第二池中的加标签的核酸进行并行扩增;

[0510] 测定所述第一池和所述第二池的扩增的核酸的序列数据。

[0511] 57.一种分析核酸群体的方法,其中至少一些核酸包括一个或更多个修饰的胞嘧啶残基,所述方法包括

[0512] 使所述核酸群体与包含引物结合位点的衔接子接触以形成侧翼为衔接子的核酸,所述引物结合位点包含修饰的胞嘧啶;

[0513] 从核酸侧翼的衔接子中的引物结合位点引发所述衔接子,扩增侧翼为所述衔接子的核酸;

[0514] 将扩增的核酸分成第一等分试样和第二等分试样;

[0515] 测定所述第一等分试样的核酸的序列数据;

[0516] 使所述第二等分试样的核酸与亚硫酸氢盐接触,这将未修饰的C转化为U;

[0517] 从核酸侧翼的引物结合位点引发,扩增由亚硫酸氢盐处理产生的核酸,其中通过亚硫酸氢盐处理引入的U被转化为T;

[0518] 测定来自所述第二等分试样的扩增的核酸的序列数据;

[0519] 比较所述第一等分试样和所述第二等分试样中的核酸的序列数据,以鉴定所述核酸群体中的哪些核苷酸是修饰的胞嘧啶。

[0520] 58.如权利要求56或57所述的方法,其中所述衔接子是发夹状衔接子。

[0521] 59.一种方法,包括:

[0522] (a)从人类样品物理地分级分离DNA分子,以生成两个或更多个分区;

[0523] (b)将差异化分子标签和支持NGS的衔接子应用于两个或更多个分区中的每一个,以生成加分子标签的分区;

[0524] (c)在NGS仪器上测定加分子标签的分区,以生成用于将所述样品解卷积成被差异化分区的分子的序列数据。

[0525] 60.如权利要求59所述的方法,还包括通过将所述样品解卷积成被差异化分区的分子来分析所述序列数据。

[0526] 61.如权利要求59所述的方法,其中所述DNA分子来自提取的血浆。

[0527] 62.如权利要求59所述的方法,其中物理分级分离包括基于不同的甲基化程度来对分子分级分离。

[0528] 63.如权利要求61所述的方法,其中不同的甲基化程度包括超甲基化和低甲基化。

[0529] 64.如权利要求59所述的方法,其中物理分级分离包括用甲基结合结构域蛋白质("MBD")-珠分级分离,以形成不同的甲基化程度的层。

[0530] 65.如权利要求59所述的方法,其中所述差异化分子标签是对应于MBD分区的不同组的分子标签。

[0531] 66.如权利要求59所述的方法,其中所述物理分级分离包括使用免疫沉淀分离DNA分子。

[0532] 67.如权利要求59所述的方法,还包括重新组合所生成的加分子标签的级分中的两种或更多种加分子标签的级分。

[0533] 68.如权利要求66所述的方法,还包括富集重新组合的加分子标签的级分或组。

[0534] 69.一种通过NGS对MBD-珠分级分离的文库进行分子标签识别的方法,包括:

[0535] (a)使用甲基结合结构域蛋白质-珠纯化试剂盒对提取的DNA样品进行物理分级分离,保留所有洗脱物用于下游处理;

[0536] (b)将差异化分子标签和支持NGS的衔接子序列并行应用于每个级分或组;

[0537] (c)重新组合所有加分子标签的级分或组,并且随后使用衔接子特异性DNA引物序列进行扩增;

[0538] (d)对重新组合并扩增的总文库进行富集/杂交,靶向感兴趣的基因组区域;

[0539] (e)重新扩增富集的总DNA文库,附以样品标签;

[0540] (f)合并不同的样品,并且在NGS仪器上对其进行多重测定;其中由仪器产生的NGS序列数据提供了用于识别独特分子的分子标签的序列,以及用于将所述样品解卷积成被差异化MBD分区的分子的序列数据。

[0541] 69A.如权利要求69所述的方法,包括用用于识别独特分子的分子标签对NGS数据进行分析,以及将所述样品解卷积成被差异化MBD分区的分子。

[0542] 70.一种方法,包括:

[0543] (a)提供从受试者的身体样品获得的核酸分子的群体;

[0544] (b)基于一个或更多个特征对所述核酸分子的群体分级分离,以生成多于一组的核酸分子,

[0545] (c)基于所述一个或更多个特征,对所述多于一组中的核酸分子差异化地加标签,以将所述多于一组中的每一组中的核酸分子彼此区分开;

[0546] (d)对所述多于一组的核酸分子测序以生成序列读段;包含足够的数据以针对多于一组的核酸分子中的每一组生成关于核小体定位、核小体修饰或DNA-蛋白质结合相互作用的相关信息。

[0547] 70A.如权利要求70所述的方法,还包括分析所述序列读段以针对多于一组的核酸分子中的每一组生成关于核小体定位、核小体修饰或DNA-蛋白质结合相互作用的相关信息。

[0548] 71.一种方法,包括:

[0549] (a)提供从受试者的身体样品获得的核酸分子的群体;

[0550] (b)基于甲基化状态对所述核酸分子的群体分级分离,以生成多于一组的核酸分子;

[0551] (c)基于所述一个或更多个特征,对所述多于一组中的核酸分子差异化地加标签,以将所述多于一组中的每一组中的核酸分子彼此区分开;

[0552] (d)对所述多于一组的核酸分子测序以生成序列读段;其中所述序列读段足以检测多于一组的核酸分子中的一组核酸分子中的一个或更多个特征,其中所述一个或更多个特征指示核小体定位、核小体修饰或DNA-蛋白质相互作用。

[0553] 71A.如权利要求71所述的方法,包括分析所述序列读段以检测所述多于一组的核酸分子中的一组核酸分子中的一个或更多个特征,其中所述一个或更多个特征指示核小体定位、核小体修饰或DNA-蛋白质相互作用。

[0554] 72.一种方法,包括:

[0555] 提供从受试者的身体样品获得的核酸分子的群体;

[0556] (a)对所述核酸分子的群体分级分离以生成多于一组的包含蛋白质结合的无细胞核酸的核酸分子;

[0557] (b)基于所述一个或更多个特征,对所述多于一组中的核酸分子差异化地加标签,以将所述多于一组中的每一组中的核酸分子彼此区分开;

[0558] (c)对所述多于一组的核酸分子测序以生成序列读段;其中所获得的序列信息足以将所述序列读段映射到参考序列上的一个或更多个基因座;并且足以用于分析所述序列读段以检测所述多于一组的核酸分子中的一组核酸分子中的一个或更多个特征,其中所述一个或更多个特征指示核小体定位、核小体修饰或DNA-蛋白质相互作用。

[0559] 72A.如权利要求72所述的方法,还包括将所述序列读段映射到参考序列上的一个或更多个基因座;以及分析所述序列读段以检测所述多于一组的核酸分子中的一组中的一个或更多个特征,其中所述一个或更多个特征指示核小体定位、核小体修饰或DNA-蛋白质相互作用。

[0560] 73.一种方法,包括:

[0561] 提供从受试者的身体样品获得的核酸分子的群体;

[0562] (a)基于一个或更多个特征对所述核酸分子的群体分级分离,以生成多于一组的核酸分子;

[0563] (b)基于所述一个或更多个特征,对所述多于一组中的核酸分子差异化地加标签,以将所述多于一组中的每一组中的核酸分子彼此区分开;

[0564] (c)对所述多于一组的核酸分子测序以生成序列读段,其中所获得的序列信息足以

[0565] 将所述序列读段映射到参考序列上的一个或更多个基因座;以及

[0566] 分析所述序列读段以检测所述多于一组的核酸分子中的一组中的一个或更多个

特征,其中所述一个或更多个特征在来自所述多于一组的核酸分子的序列读段的池中不能检测。

[0567] 73A.如权利要求73所述的方法,还包括将所述序列读段映射到参考序列上的一个或更多个基因座;以及分析所述序列读段以检测所述多于一组的核酸分子中的一组中的一个或更多个特征,其中所述一个或更多个特征在来自所述多于一组的序列读段的池中不能检测。

[0568] 74.如权利要求70-72中任一项所述的方法,其中所述一个或更多个特征包括映射读段的定量特征。

[0569] 75.如权利要求69-73中任一项所述的方法,其中所述分级分离包括物理分级分离。

[0570] 76.如权利要求69或72所述的方法,其中基于选自由以下组成的组的一个或更多个特征对所述核酸分子的群体进行分区:甲基化状态、糖基化状态、组蛋白修饰、长度和起始/终止位置。

[0571] 77.如权利要求69-72中任一项所述的方法,还包括合并(b)的核酸分子。

[0572] 78.如权利要求69或72所述的方法,其中所述一个或更多个特征是甲基化。

[0573] 79.如权利要求77所述的方法,其中分级分离包括使用包含甲基结合结构域的蛋白质分离甲基化的核酸和非甲基化的核酸,以生成包含不同甲基化程度的核酸分子的组。

[0574] 80.如权利要求78所述的方法,其中所述组中的一组包括超甲基化的DNA。

[0575] 81.如权利要求78所述的方法,其中至少一个组的特征在于甲基化程度。

[0576] 82.如权利要求72所述的方法,其中分级分离包括分离单链DNA分子和/或双链DNA分子。

[0577] 83.如权利要求81所述的方法,其中所述双链DNA分子使用发夹状衔接子来分离。

[0578] 84.如权利要求69或72所述的方法,其中分级分离包括分离蛋白质结合的核酸。

[0579] 85.如权利要求69-72中任一项所述的方法,其中分级分离包括基于单核小体特征谱的差异的分级分离。

[0580] 86.如权利要求69-72中任一项所述的方法,其中当与正常相比时,分级分离能够为至少一组核酸分子生成不同的单核小体特征谱。

[0581] 87.如权利要求85所述的方法,其中所述分离包括免疫沉淀。

[0582] 88.如权利要求69-72中任一项所述的方法,还包括基于不同的特征对至少一组核酸分子分级分离。

[0583] 89.如权利要求69-72中任一项所述的方法,其中分析包括将对应于第一组核酸分子的第一特征与对应于第二组核酸分子的第二特征在一个或更多个基因座处进行比较。

[0584] 90.如权利要求70-72中任一项所述的方法,其中分析包括分析组的一个或更多个特征相对于正常样品在一个或更多个基因座处的特征。

[0585] 91.如权利要求70-72中任一项所述的方法,其中所述一个或更多个特征选自由以下组成的组:在参考序列上的一个碱基位置处的碱基调用频率、映射到参考序列上的一个碱基或序列的分子的数目、具有映射到参考序列上的一个碱基位置的起始位点的分子的数目和具有映射到参考序列上的一个碱基位置的终止位点的分子的数目,以及映射到参考序列上的一个基因座的分子的长度。

[0586] 92.如权利要求70-72中任一项所述的方法,还包括(f)使用经训练的分类器基于所述一个或更多个特征对所述受试者分类。

[0587] 93.如权利要求91所述的方法,其中经训练的分类器将所述一个或更多个特征分类为与所述受试者中的组织相关。

[0588] 94.如权利要求91所述的方法,其中经训练的分类器将所述一个或更多个特征分类为与所述受试者中的癌症类型相关。

[0589] 95.如权利要求70-72所述的方法,其中所述一个或更多个特征指示基因表达或疾病状态。

[0590] 96.如权利要求69-72中任一项所述的方法,其中所述核酸分子是循环肿瘤DNA。

[0591] 97.如权利要求69-72中任一项所述的方法,其中所述核酸分子是无细胞DNA(“cfDNA”)。

[0592] 98.如权利要求69-71中任一项所述的方法,其中所述一个或更多个特征是癌症标志物。

[0593] 99.如权利要求69-72中任一项所述的方法,其中所述标签用于区分同一样品中的不同分子。

[0594] 100.一种方法,包括:

[0595] (a)提供从受试者的身体样品获得的核酸分子的群体;

[0596] (b)基于一个或更多个特征对所述核酸分子的群体分级分离以生成多于一组的核酸分子,其中所述多于一组中的每一组的核酸分子包含不同的标识物;

[0597] (c)合并所述多于一组的核酸分子;

[0598] (d)对合并的多于一组的核酸分子测序以生成多于一组的序列读段;和

[0599] (e)基于所述标识物对所述序列读段分级分离。

[0600] 101.一种组合物,所述组合物包含含有差异化地加标签的核酸分子的核酸分子池,其中所述池包含多于一组的核酸分子,所述多于一组的核酸分子基于选自由以下组成的组的一个或更多个特征被差异化地加标签:甲基化状态、糖基化状态、组蛋白修饰、长度和起始/终止位置,其中所述池来源于生物样品。

[0601] 102.如权利要求101所述的组合物,其中所述多于一组是2组、3组、4组、5组或多于5组中的任一种。

[0602] 103.一种方法,包括:

[0603] (a)将核酸分子的群体分级分离成多于一组,所述多于一组包含特征不同的核酸;

[0604] (b)用一组标签对所述多于一组中的每一组中的核酸加标签,所述一组标签区分所述多于一组中的每一组中的核酸以产生加标签的核酸的群体,其中每个加标签的核酸包含一个或更多个标签;

[0605] (c)对加标签的核酸的群体测序以生成序列读段,其中所述序列读段允许使用一个或更多个标签来将每组序列读段分组;以及分析所述序列读段以检测所述组中的至少一个组相对于正常样品或分类器的信号。

[0606] 103A.如权利要求103所述的方法,还包括使用所述一个或更多个标签来将每组序列读段分组;和

[0607] 分析所述序列读段以检测所述组中的至少一组相对于正常样品或分类器的信号。

[0608] 104.如权利要求102所述的方法，还包括将至少一组中的信号针对另一组或全基因组序列归一化。

[0609] 105.一种方法，包括：

[0610] i.提供来自生物样品的无细胞DNA的群体；

[0611] ii.基于相比于非癌细胞以不同水平存在于来源于癌细胞的无细胞DNA中的特征，对所述无细胞DNA的群体分级分离，从而生成无细胞DNA的亚群；

[0612] iii.扩增无细胞DNA亚群中的至少一个；和

[0613] iv.对扩增的无细胞DNA亚群中的至少一个测序。

[0614] 106.如权利要求104所述的方法，其中所述特征为：

[0615] i.所述无细胞DNA的甲基化水平；

[0616] ii.所述无细胞DNA的糖基化水平；

[0617] iii.所述无细胞DNA片段的长度；或

[0618] iv.所述无细胞DNA中单链断裂的存在。

[0619] 107.一种方法，包括：

[0620] i.提供来自生物样品的无细胞DNA的群体；

[0621] ii.基于所述无细胞DNA的甲基化水平对所述无细胞DNA的群体分级分离，从而生成无细胞DNA的亚群；

[0622] iii.扩增无细胞DNA亚群中的至少一个；和

[0623] iv.对扩增的无细胞DNA亚群中的至少一个测序。

[0624] 108.一种用于确定无细胞DNA的甲基化状态的方法，包括：

[0625] i.提供来自生物样品的无细胞DNA的群体；

[0626] ii.基于所述无细胞DNA的甲基化水平对所述无细胞DNA的群体分级分离，从而生成无细胞DNA的亚群；

[0627] iii.对无细胞DNA的至少一个亚群测序，从而生成序列读段；

[0628] iv.根据相应序列读段出现在其中的亚群，为每个无细胞DNA指定甲基化状态。

[0629] 109.一种对受试者分类的方法，其中所述方法包括：

[0630] i.提供来自所述受试者的生物样品的无细胞DNA的群体；

[0631] ii.基于所述无细胞DNA的甲基化水平对所述无细胞DNA的群体分级分离，从而生成无细胞DNA的亚群；

[0632] iii.对无细胞DNA的亚群测序，从而生成序列读段；和

[0633] iv.根据哪个序列读段在哪个亚群中出现，使用经训练的分类器对所述受试者分类。

[0634] 110.一种用于分析无细胞DNA的片段化模式的方法，包括：

[0635] i.提供来自生物样品的无细胞DNA的群体；

[0636] ii.对所述无细胞DNA的群体分级分离，从而生成无细胞DNA的亚群；

[0637] iii.对无细胞DNA的至少一个亚群测序，从而生成序列读段；

[0638] iv.将所述序列读段与参考基因组对齐；和

[0639] v.通过分析以下任一数目确定每个亚群中所述无细胞DNA的片段化模式：

[0640] a.映射到参考基因组中的每个碱基位置的每个序列读段的长度；

[0641] b.按照所述序列读段的长度的函数,映射到参考基因组中的碱基位置的序列读段的数目;

[0642] c.在参考基因组中的每个碱基位置起始的序列读段的数目;或

[0643] d.在参考基因组中的每个碱基位置终止的序列读段的数目。

[0644] 111.如权利要求109所述的方法,其中所述无细胞DNA的群体通过提供健康状态和患病状态之间的信号差异的一个或更多个特征被分级分离。

[0645] 112.如权利要求110所述的方法,其中所述一个或更多个特征包括选自由以下组成的组的化学修饰:甲基化、羟甲基化、甲酰化、乙酰化和糖基化。

[0646] 113.如前述权利要求中任一项所述的方法,其中DNA:珠的比为1:100。

[0647] 114.如前述权利要求中任一项所述的方法,其中DNA:珠的比为1:50。

[0648] 115.如前述权利要求中任一项所述的方法,其中DNA:珠的比为1:20。

[0649] 116.如权利要求109所述的方法,其中基于所述无细胞DNA的甲基化水平对所述无细胞DNA的群体分级分离。

[0650] 117.如权利要求109所述的方法,其中确定所述无细胞DNA的片段化模式还包括分析映射到参考基因组中的每个碱基位置的序列读段的数目。

[0651] 118.如权利要求109所述的方法,还包括通过分析映射到参考基因组中的每个碱基位置的序列读段的数目来确定每个亚群中的无细胞DNA的片段化模式。

[0652] 119.基于DNA甲基化程度的物理分级分离在循环肿瘤DNA(ctDNA)的分析期间对于确定基因表达或疾病状态的用途。

[0653] 120.提供正常状态和患病状态之间的信号差异的特征对于在ctDNA分析期间对ctDNA物理地分区的用途。

[0654] 121.提供正常状态和患病状态之间的信号差异的特征对于对ctDNA物理地分区的用途。

[0655] 122.提供正常状态和患病状态之间的信号差异的特征对于在测序和任选的下游分析之前对ctDNA物理地分区的用途。

[0656] 123.提供正常状态和患病状态之间的信号差异的特征对于对ctDNA物理地分区以便对其差异化标记/加标签的用途。

[0657] 124.在ctDNA分析期间基于差异片段化模式的分级分离的用途。

[0658] 125.差异片段化模式对于对ctDNA分区的用途。

[0659] 126.差异片段化模式对于在测序和任选的下游分析之前对ctDNA分区的用途。

[0660] 127.差异片段化模式对于对ctDNA分区以便对其差异化标记/加标签的用途。

[0661] 128.如权利要求123-126所述的用途,其中所述差异片段化模式指示基因表达或疾病状态。

[0662] 129.如权利要求123-126所述的用途,其中所述差异片段化模式的特征在于相对于正常的选自由以下组成的组的一个或更多个差异:

[0663] (a)映射到参考基因组中的每个碱基位置的每个序列读段的长度;

[0664] (b)作为所述序列读段的长度的函数的映射到参考基因组中的碱基位置的序列读段的数目;

[0665] (c)在参考基因组中的每个碱基位置处起始的序列读段的数目;和

[0666] (d)在参考基因组中的每个碱基位置处终止的序列读段的数目。

[0667] 130.对由分子结合结构域(MBD)-珠分区的DNA分子差异化地加分子标签对于形成不同的DNA甲基化程度的层的用途,所述不同的DNA甲基化程度然后通过下一代测序(NGS)被定量。

[0668] 131.一种分析核酸群体的方法,所述核酸群体包含选自双链DNA、单链DNA和单链RNA的至少两种形式的核酸,其中所述至少两种形式中的每一种包含多于一个分子,所述方法包括:

[0669] (a)将至少一种所述形式的核酸与至少一种标签核酸连接,以将所述形式彼此区分开,

[0670] (b)扩增所述形式的核酸,其中的至少一种被连接至至少一种核酸标签,其中所述核酸和连接的核酸标签被扩增以产生扩增的核酸,其中从所述至少一种形式扩增的核酸被加标签;

[0671] (c)对已经连接至标签的多于一个扩增的核酸测序,其中所述序列数据足以被解码以揭示所述群体中的核酸在连接至至少一个标签之前的形式。

[0672] 132.一种加标签的核酸分子的池,所述池中的每个核酸分子包括选自多于一个标签组中的一个标签组的分子标签,每个标签组包含多于一个不同的标签,其中任何一组中的标签不同于任何其他组中的标签,并且其中每个标签组包含(i)指示其所附接的分子或该分子所源自的亲本分子的特征的信息,以及(ii)单独地或与来自其所附接的分子的信息组合地,将其所附接的分子与用来自相同标签组的标签加标签的其他分子独特地区分开的信息。

[0673] 133.如权利要求132所述的加标签的核酸分子的池,其中所述分子标签包括一个或多于一个核酸条形码。

[0674] 134.如权利要求133所述的加标签的核酸分子的池,其中所述分子标签包括附接在所述分子的相对端的两个核酸条形码。

[0675] 135.如权利要求134所述的加标签的核酸分子的池,其中一组中任何两个条形码的组合具有与任何其他组中任何两个条形码的组合相比不同的组合序列。

[0676] 136.如权利要求133所述的加标签的核酸分子的池,其中所述条形码的长度在10个和30个核苷酸之间。

[0677] 137.如权利要求132所述的加标签的核酸分子的池,其中每个标签组包括多于一个不同的标签,所述多于一个不同的标签足以独特地对分子加标签,所述分子被所述标签组加标签,并且具有相同的起始-终止坐标或具有相同的核苷酸序列或映射到相同的基因组坐标。

[0678] 138.如权利要求132所述的加标签的核酸分子的池,其中所述多于一个标签组是2个、3个、4个、5个、6个或多于6个。

[0679] 139.如权利要求132所述的加标签的核酸分子的池,其中所述池包括具有用来自一个标签组的标签加标签的DNA序列的分子,以及具有用来自另一标签组的标签加标签的cDNA序列的分子。

[0680] 140.如权利要求132所述的加标签的核酸分子的池,其中由所述标签组指示的分子的特征包括以下中的一种或更多种:DNA、RNA、单链的、双链的、甲基化的、非甲基化的、甲

基化程度或上述的组合。

[0681] 141.一种系统,包括:

[0682] 核酸测序仪;

[0683] 数字处理设备,其包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器;以及

[0684] 数据链路,其通信连接所述核酸测序仪和所述数字处理设备;

[0685] 其中所述数字处理设备还包括可被执行以创建用于分析核酸群体的应用程序的指令,所述核酸群体包括选自双链DNA、单链DNA和单链RNA的至少两种形式的核酸,所述至少两种形式中的每一种包含多于一个分子,所述应用程序包括:

[0686] 软件模块,其经由所述数据链路从所述核酸测序仪接收序列数据、扩增的核酸的序列数据,所述扩增的核酸中的至少一些被加标签;所述序列数据通过以下生成:将所述形式的核酸中的至少一种与至少一种加标签的核酸连接以将所述形式彼此区分开,扩增所述形式的核酸,其中的至少一种被连接至至少一种核酸标签,其中核酸和连接的核酸标签被扩增以产生扩增的核酸,其中从所述至少一种形式扩增的核酸被加标签;以及

[0687] 软件模块,其通过获得足以解码扩增的核酸的加标签的核酸分子的序列信息来测定扩增的核酸的序列数据,以揭示所述群体中为已测定了针对其的序列数据的连接至所述标签核酸分子的扩增的核酸提供原始模板的核酸的形式。

[0688] 142.如权利要求141所述的系统,其中所述应用程序还包括解码所述扩增的核酸中的所述加标签的核酸分子的软件模块,以揭示所述群体中为已测定了针对其的序列数据的连接至所述标签核酸分子的所述扩增的核酸提供原始模板的核酸的形式。

[0689] 143.如权利要求141所述的系统,其中所述应用程序还包括经由通信网络传输所述测定的结果的软件模块。

[0690] 144.一种系统,包括:

[0691] 下一代测序(NGS)仪器;

[0692] 数字处理设备,其包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器;以及

[0693] 数据链路,其通信连接所述NGS仪器和所述数字处理设备;

[0694] 其中所述数字处理设备还包括可被执行以创建应用程序的指令,所述应用程序包括:

[0695] 软件模块,用于经由所述数据链路从所述NGS仪器接收序列数据,所述序列数据通过以下生成:对来自人类样品的DNA分子物理地分级分离以生成两个或更多个分区,将差异化分子标签和支持NGS的衔接子应用于两个或更多个分区中的每一个以生成加分子标签的分区,以及用所述NGS仪器测定加分子标签的分区;

[0696] 软件模块,用于生成用于将所述样品解卷积成被差异化分区的分子的序列数据;以及

[0697] 软件模块,用于通过将所述样品解卷积成被差异化分区的分子来分析所述序列数据。

[0698] 145.如权利要求144所述的系统,其中所述应用程序还包括经由通信网络传输所述测定的结果的软件模块。

[0699] 146.一种系统,包括:

[0700] 下一代测序(NGS)仪器;

[0701] 数字处理设备,其包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器;以及

[0702] 数据链路,其通信连接所述NGS仪器和所述数字处理设备;

[0703] 其中所述数字处理设备还包括可由所述至少一个处理器执行以创建用于MBD-珠分级分离文库的分子标签识别的应用程序的指令,所述应用程序包括:

[0704] 软件模块,被配置为经由所述数据链路从所述NGS仪器接收序列数据,所述序列数据通过以下生成:使用甲基结合结构域蛋白质-珠纯化试剂盒对提取的DNA样品进行物理分级分离,保留所有洗脱物用于下游处理;进行差异化分子标签和支持NGS的衔接子序列对每个级分或组的并行应用;重新组合所有加分子标签的级分或组,并且随后使用衔接子特异性DNA引物序列进行扩增;对重新组合和扩增的总文库进行富集/杂交,靶向感兴趣的基因组区域;重新扩增富集的总DNA文库,附以样品标签;合并不同的样品,和在NGS仪器上对其进行多重测定;其中由所述仪器产生的NGS序列数据提供了用于识别独特分子的分子标签的序列,以及用于将所述样品解卷积成被差异化MBD分区的分子的序列数据;以及

[0705] 软件模块,被配置为通过使用分子标签识别独特的分子并且将所述样品解卷积成被差异化MBD-分区的分子来进行序列数据的分析。

[0706] 147.如权利要求146所述的系统,其中所述应用程序还包括被配置为经由通信网络传输所述分析的结果的软件模块。

[0707] 148.一种系统,包括:

[0708] a)下一代测序(NGS)仪器

[0709] b)数字处理设备,其包括至少一个处理器、被配置为执行可执行指令的操作系统、和存储器;以及

[0710] c)数据链路,其通信连接所述NGS仪器和所述数字处理设备;

[0711] 其中所述数字处理设备还包括可被执行以创建应用程序的指令,所述应用程序包括:

[0712] i)软件模块,其用于经由所述数据链路从所述NGS仪器接收序列数据,所生成的序列数据加载有标记的核酸,所述标记的核酸通过以下来制备:使所述核酸群体与优先结合带有修饰的核酸的剂接触,分离与所述剂结合的第一核酸池和未与所述剂结合的第二核酸池,其中所述第一核酸池对于所述修饰是呈现过度的,并且所述第二池中的核酸对于所述修饰是呈现不足的;将所述第一池和/或所述第二池中的核酸连接至区分所述第一池和所述第二池中的核酸的一个或更多个核酸标签,以产生加标签的核酸的群体;扩增标记的核酸,其中所述核酸和连接的标签被扩增;以及用所述NGS仪器测定加分子标签的分区;

[0713] ii)软件模块,其用于生成用于解码所述标签的序列数据;和

[0714] iii)软件模块,其用于分析所述序列数据以解码所述标签,以揭示已测定了针对其的序列数据的核酸是从所述第一池中的模板扩增的还是从所述第二池中的模板扩增的。

[0715] 149.如权利要求148所述的系统,还包括经由通信网络传输所述测定的结果的软件模块。

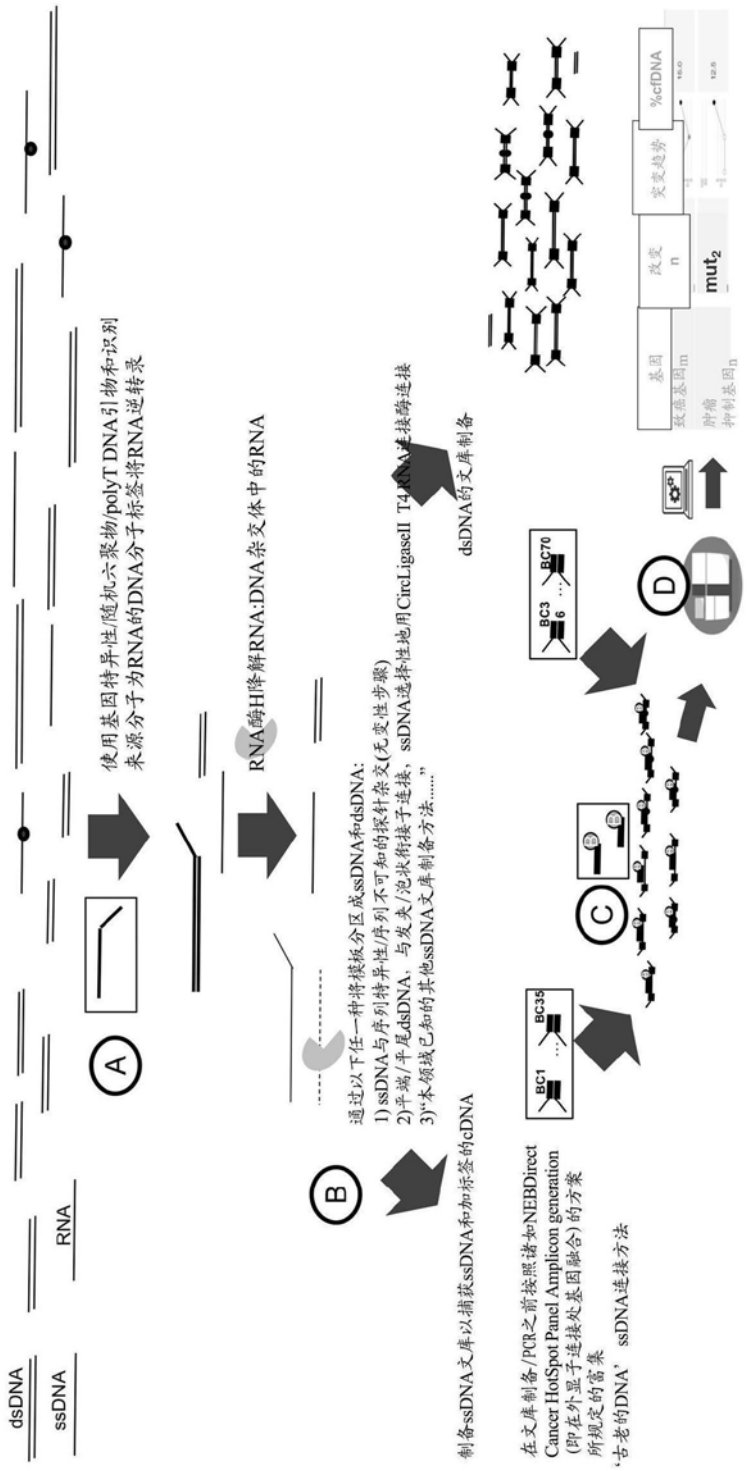

dsDNA
ssDNA
RNA
A
使用基因特异性/随机六聚物/polyT DNA引物和识别来源分子为RNA的DNA分子标签将RNA逆转录
RNA酶H降解RNA:DNA杂交体中的RNA
B
通过以下任一种将模板分区成ssDNA和dsDNA:
1) ssDNA与序列特异性/序列不可知的探针杂交(无变性步骤)
2)平端/半尾dsDNA，与发夫/泡状衔接子连接，ssDNA选择性地用CircLigaseII T4 RNA连接酶连接
3)"本领域已知的其他ssDNA文库制备方法....."
dsDNA的文库制备
制备ssDNA文库以捕获ssDNA和加标签的cDNA
BC1
BC35
BC3
BC70
C
D
在文库制备/PCR之前按照诸如NEBDirect Cancer HotSpot Panel Amplicon generation (即在外显子连接处基因融合)的方案所规定的富集
'古老的DNA' ssDNA连接方法
基因
致癌基因m
肿瘤
抑制基因n
改变
n
mut_2
变异趋势
%cfDNA

图1

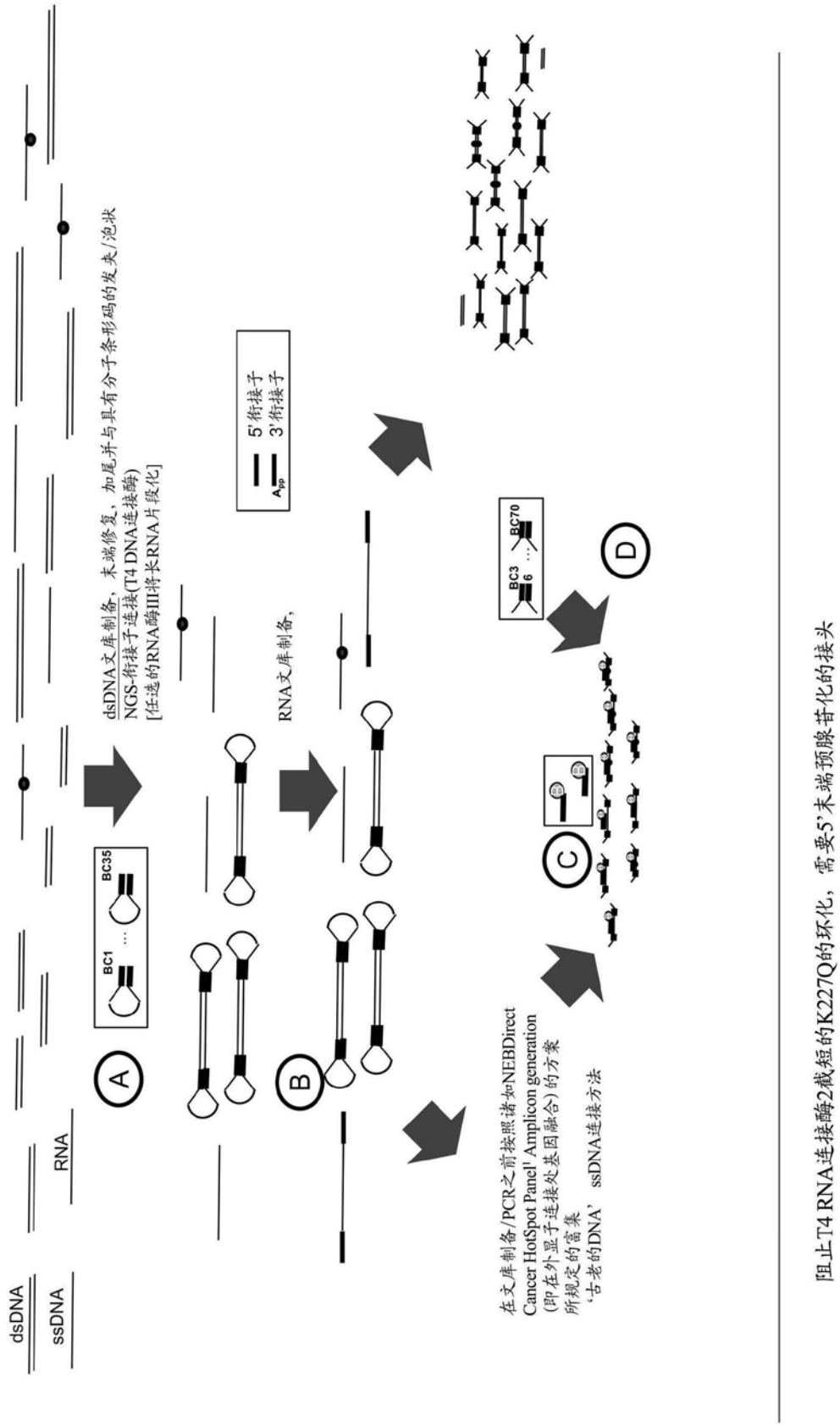

dsDNA
ssDNA
RNA
A
BC1
BC35
dsDNA文库制备，末端修复，加尾并与具有分子条形码的发夹/泡状
NGS-衔接子连接(T4 DNA连接酶)
[任选的RNA酶III将长RNA片段化]
5'衔接子
3'衔接子
B
RNA文库制备，
C
BC3
6
BC70
D
在文库制备/PCR之前按照诸如NEBDirect
Cancer HotSpot Panel' Amplicon generation
(即在外显子连接处基因融合)的方案
所规定的富集
'古老的DNA' ssDNA连接方法
阻止T4 RNA连接酶2截短的K227Q的环化，需要5'末端预腺苷化的接头

图2

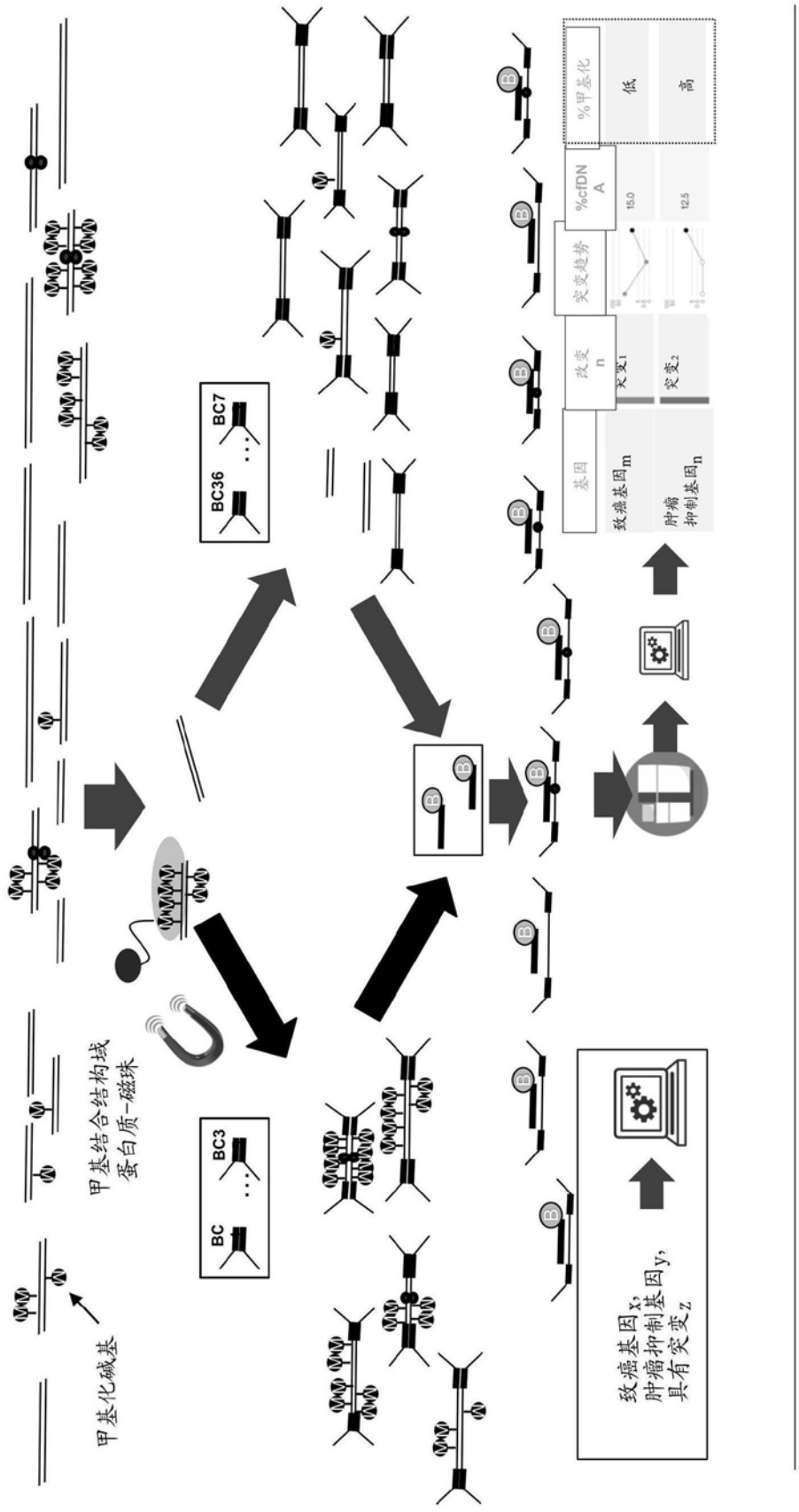

甲基化碱基
甲基结合结构域
蛋白质-磁珠
BC
BC3
BC36
BC7
致癌基因x,
肿瘤抑制基因y,
具有突变z
基因
改变
n
突变趋势
%cfDN
A
%甲基化
低
高
致癌基因m
肿瘤
抑制基因n

图3

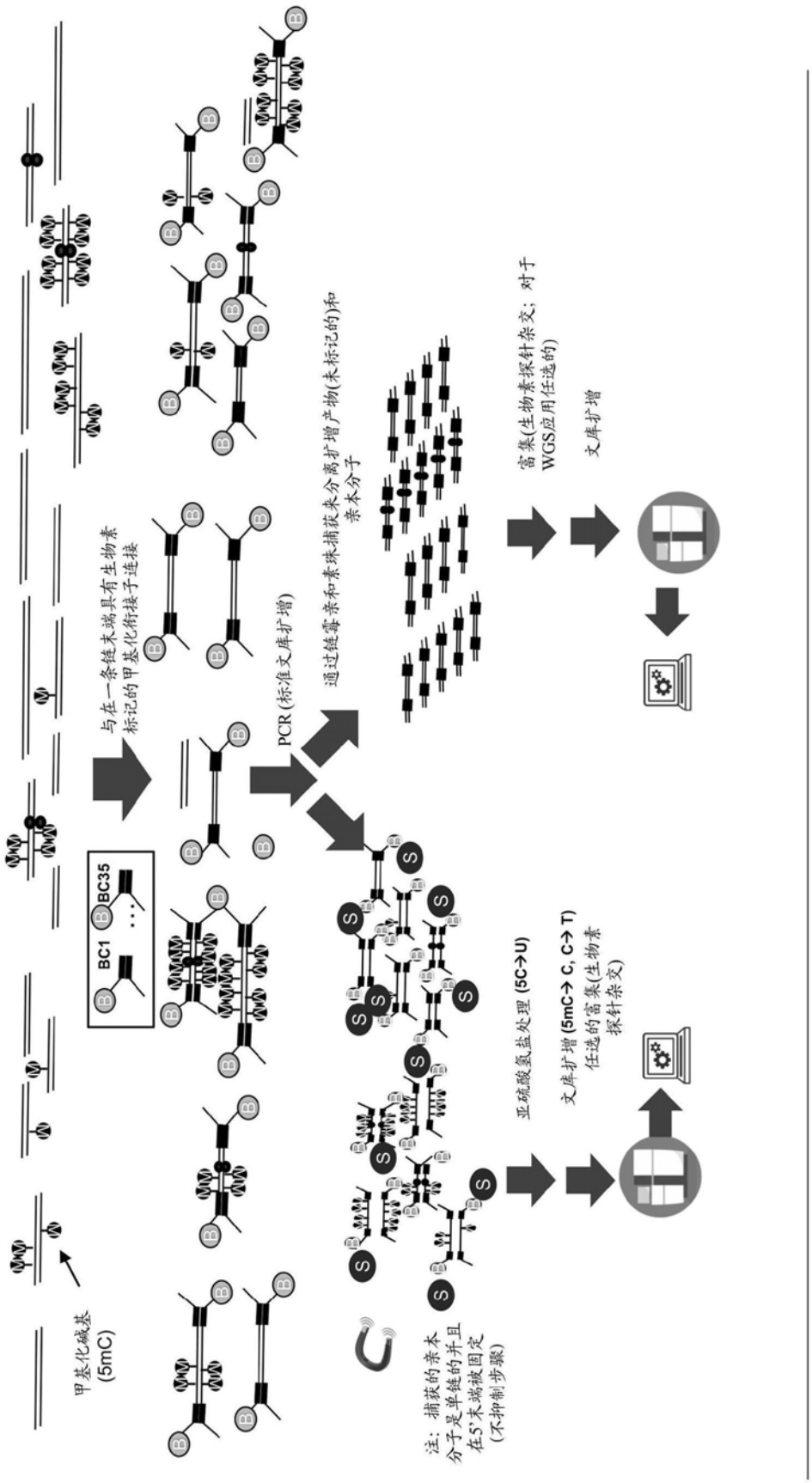

甲基化碱基
(5mC)
BC1
BC35
与在一条链末端具有生物素
标记的甲基化衔接子连接
PCR (标准文库扩增)
通过链霉亲和素珠捕获来分离扩增产物(未标记的)和
亲本分子
注：捕获的亲本
分子是单链的并且
在5'末端被固定
(不抑制步骤)
亚硫酸氢盐处理 (5C→U)
文库扩增 (5mC→ C, C→ T)
任选的富集(生物素
探针杂交)
富集(生物素探针杂交；对于
WGS应用任选的)
文库扩增

图4

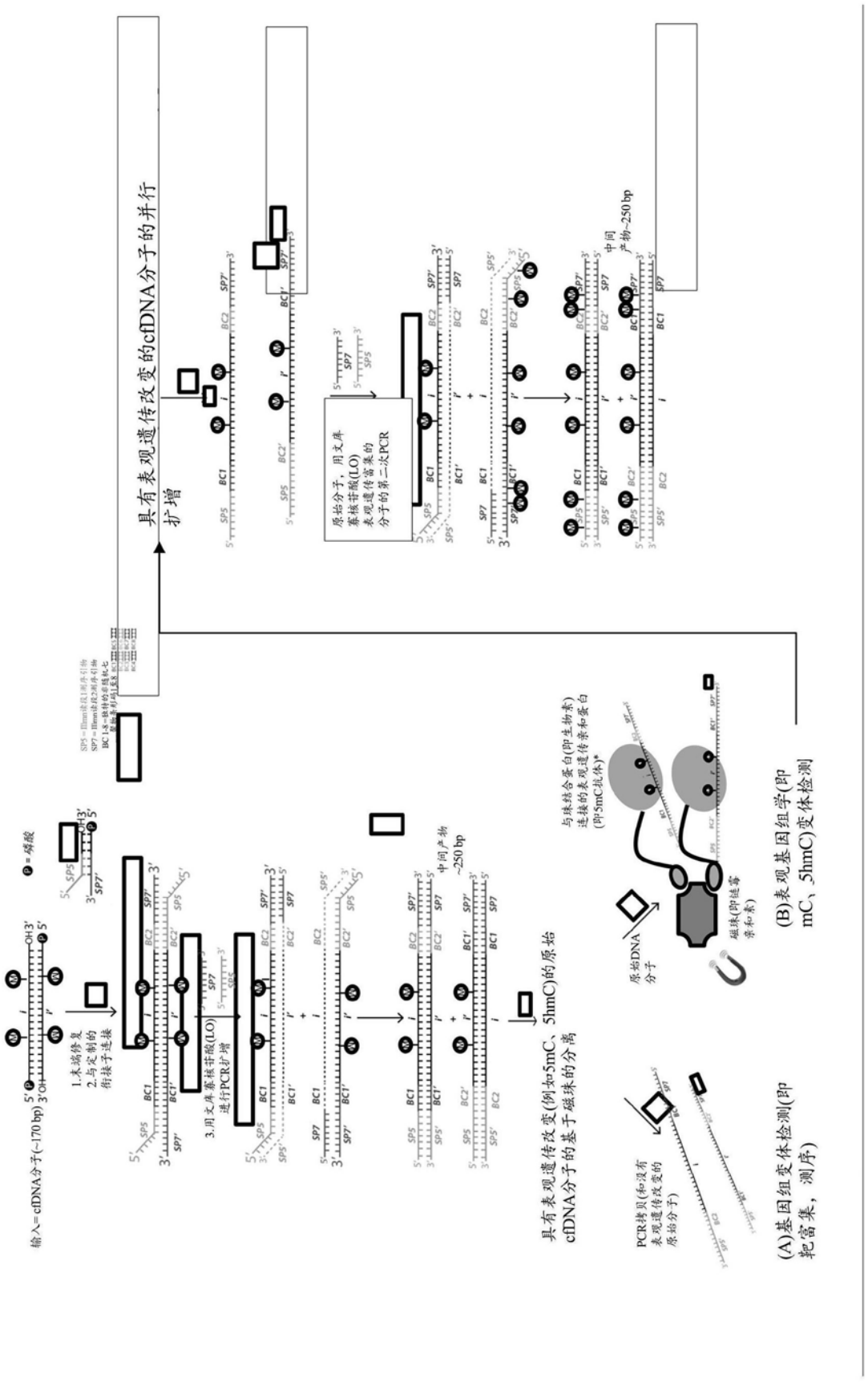

具有表观遗传改变的cfDNA分子的并行扩增
输入= cfDNA分子(~170 bp)
1.末端修复
2.与定制的衔接子连接
3.用文库靶核苷酸(LO)进行PCR扩增
中间产物~250 bp
具有表观遗传改变(例如5mC、5hmC)的原始cfDNA分子的基于磁珠的分离
原始分子，用文库靶核苷酸(LO)表观遗传富集的分子的第二次PCR
中间产物~250 bp
与珠结合蛋白(即生物素)连接的表观遗传亲和蛋白(即5mC抗体)*
原始DNA分子
磁珠(即链霉亲和素)
PCR拷贝(和没有表观遗传改变的原始分子)
(A)基因组变体检测(即靶富集，测序)
(B)表观基因组学(即mC、5hmC)变体检测

图5

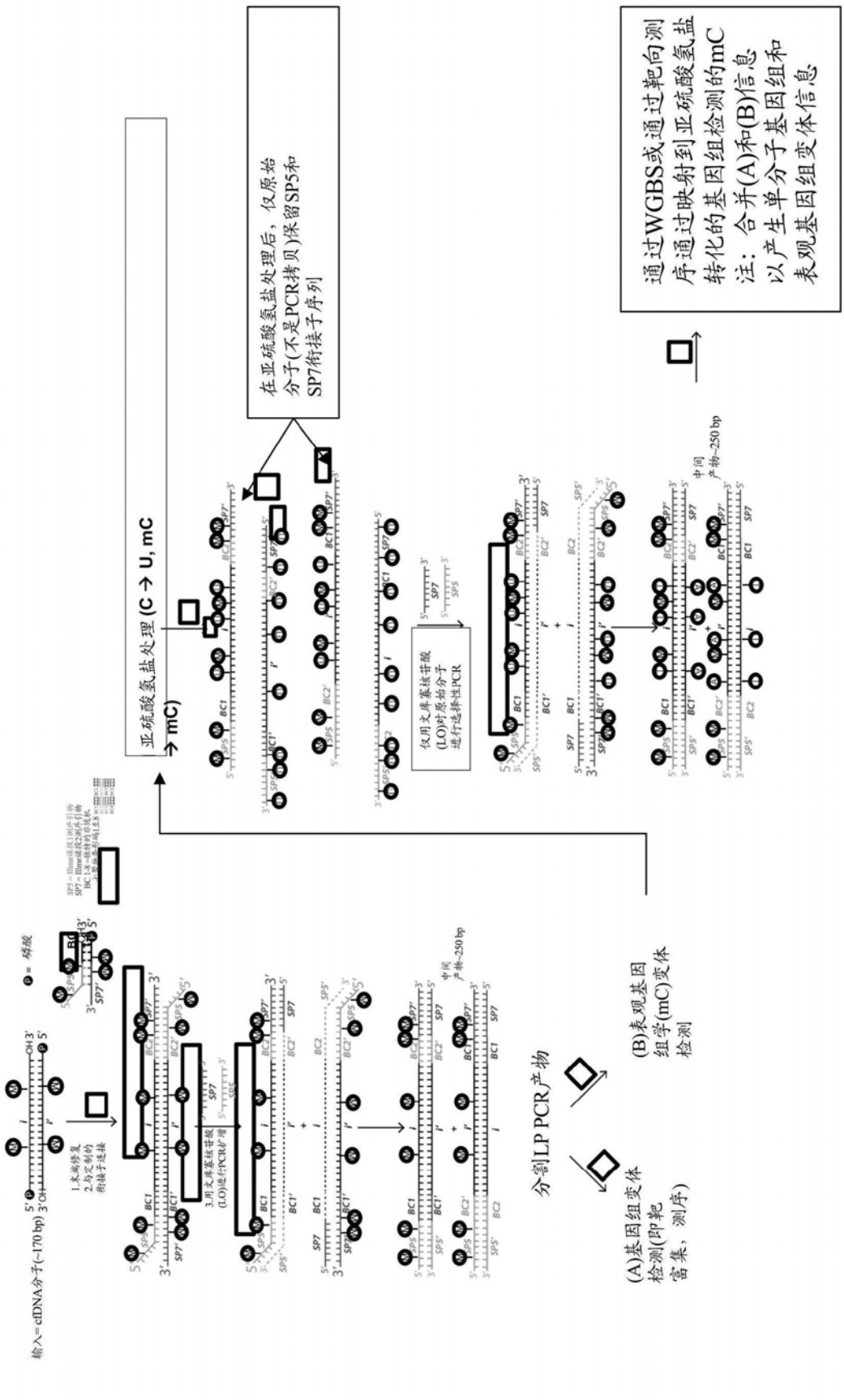

输入= cfDNA分子(~170 bp)
1.末端修复
2.分子制约
衔接子连接
3.用文库基因带酸
(LO)进行PCR扩增
亚硫酸氢盐处理 (C → U, mC)
→ mC)
在亚硫酸氢盐处理后，仅原始
分子(不是PCR拷贝)保留SP5和
SP7衔接子序列
仅用文库基因带酸
(LO)对原始分子
进行选择性PCR
中间
产物~250 bp
分割LP PCR产物
(A)基因组变体
检测(印靶
富集，测序)
(B)表观基因
组学(mC)变体
检测
通过WGBS或通过靶向测
序通过映射到亚硫酸氢盐
转化的基因组检测的mC
注：合并(A)和(B)信息
以产生单分子基因组和
表观基因组变体信息

图6

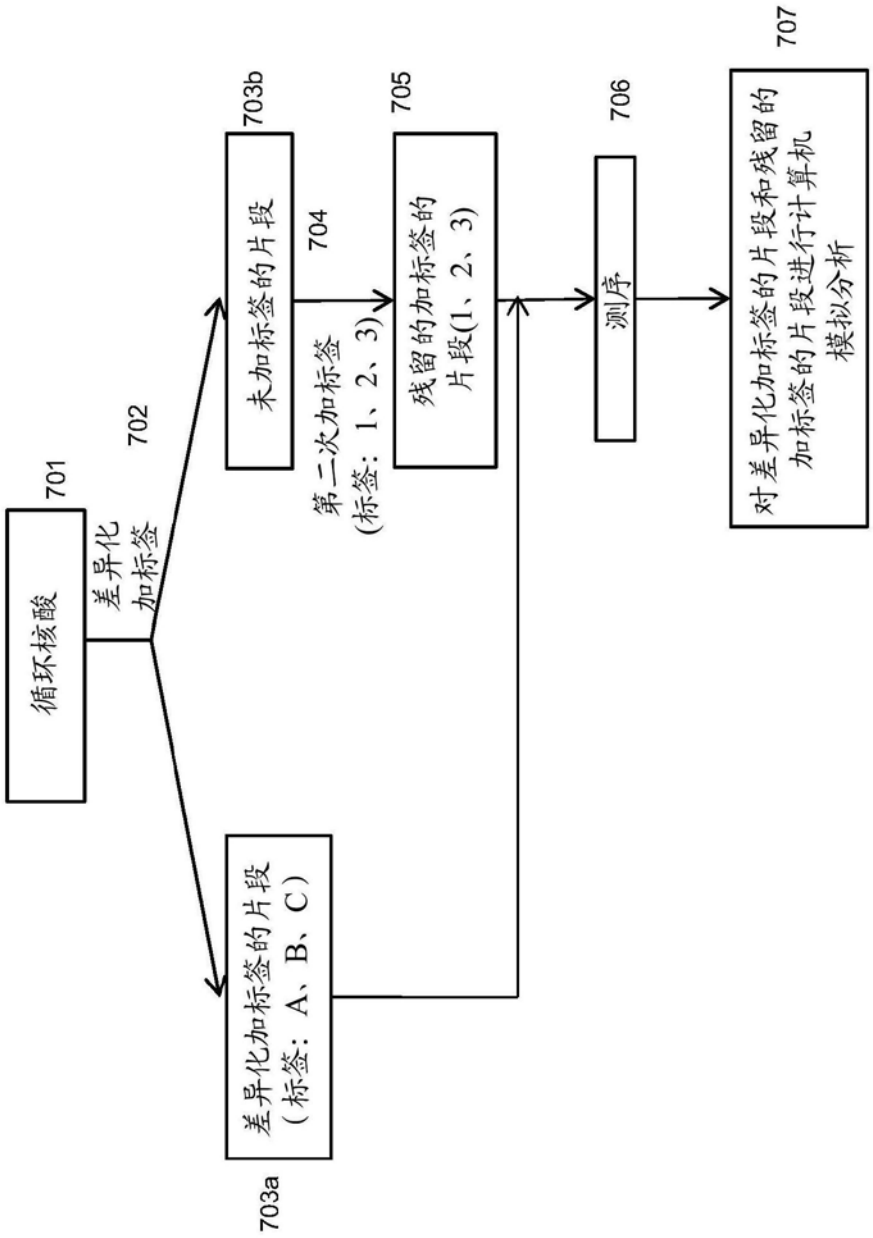

循环核酸
701
差异化加标签
702
703a
差异化加标签的片段（标签：A、B、C）
703b
未加标签的片段
704
第二次加标签(标签：1、2、3)
705
残留的加标签的片段(1、2、3)
706
测序
707
对差异化加标签的片段和残留的加标签的片段进行计算机模拟分析

图7

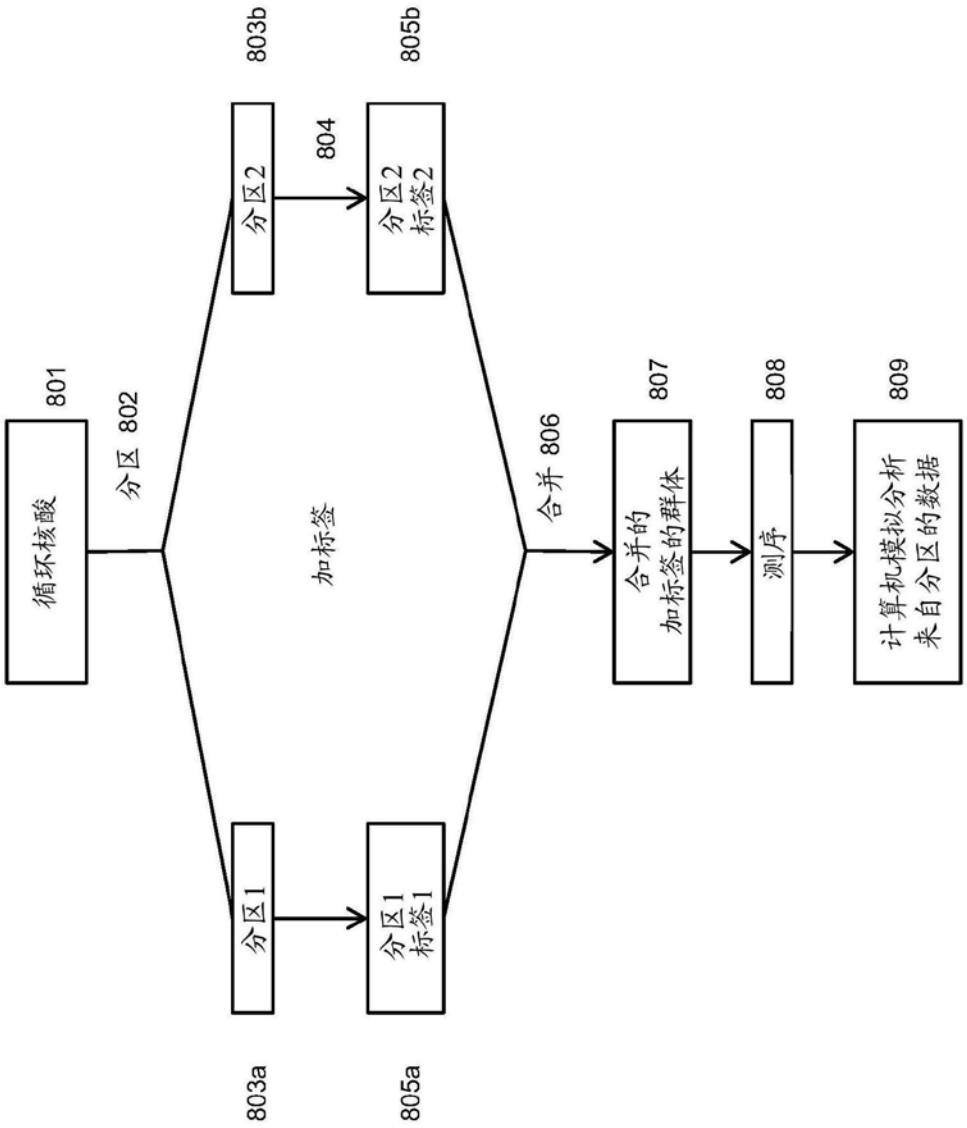

803b
805b
分区2
804
分区2
标签2
801
分区 802
循环核酸
加标签
合并 806
807
808
809
合并的
加标签的群体
测序
计算机模拟分析
来自分区的数据
分区1
分区1
标签1
803a
805a

图8

占据: 高 低 高 低
定位: 良好 良好 差 差
单细胞中的核小体
核小体
DNA
测序信号
1. 使用8个无MBD肺癌样品来识别二分体中心
2. 分离具有良好定位的二分体和高占据的规范的核小体
（覆盖度>.5Qu & 峰宽<.5Qu）
3. 将二分体中心定位在组合的MBD样品
4. 比较无MBD与MBD对齐的样品来计算二分体中心距离

图9

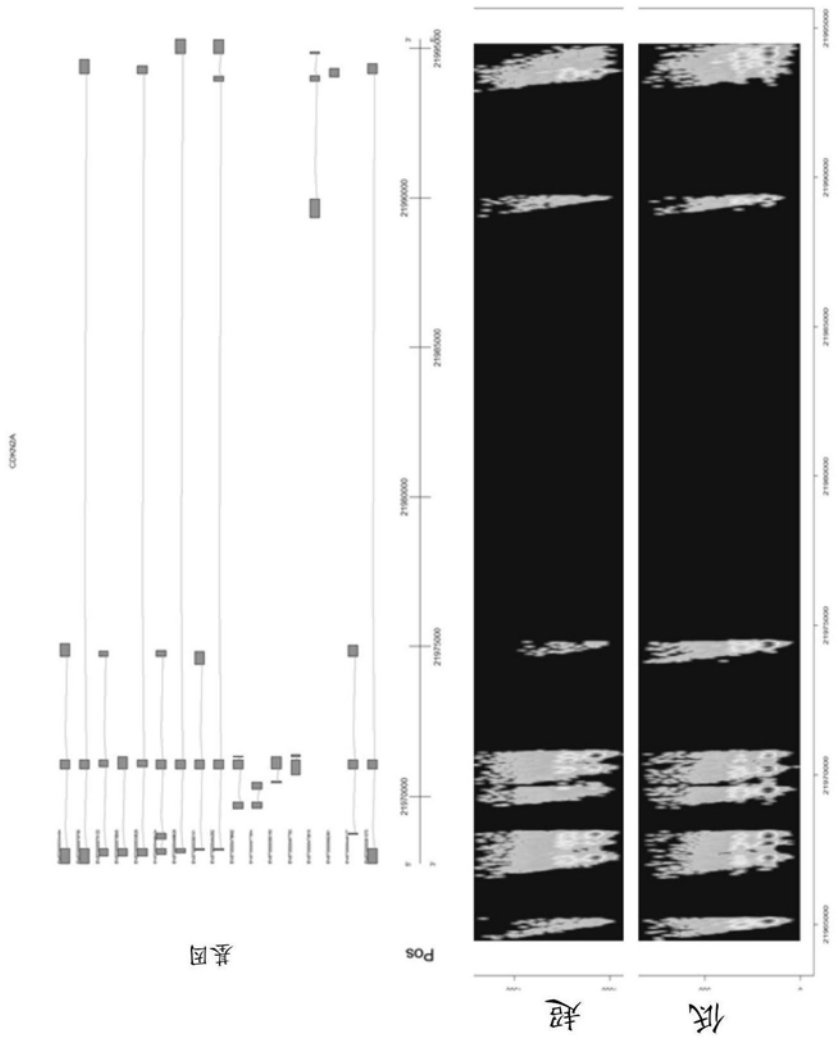


图10

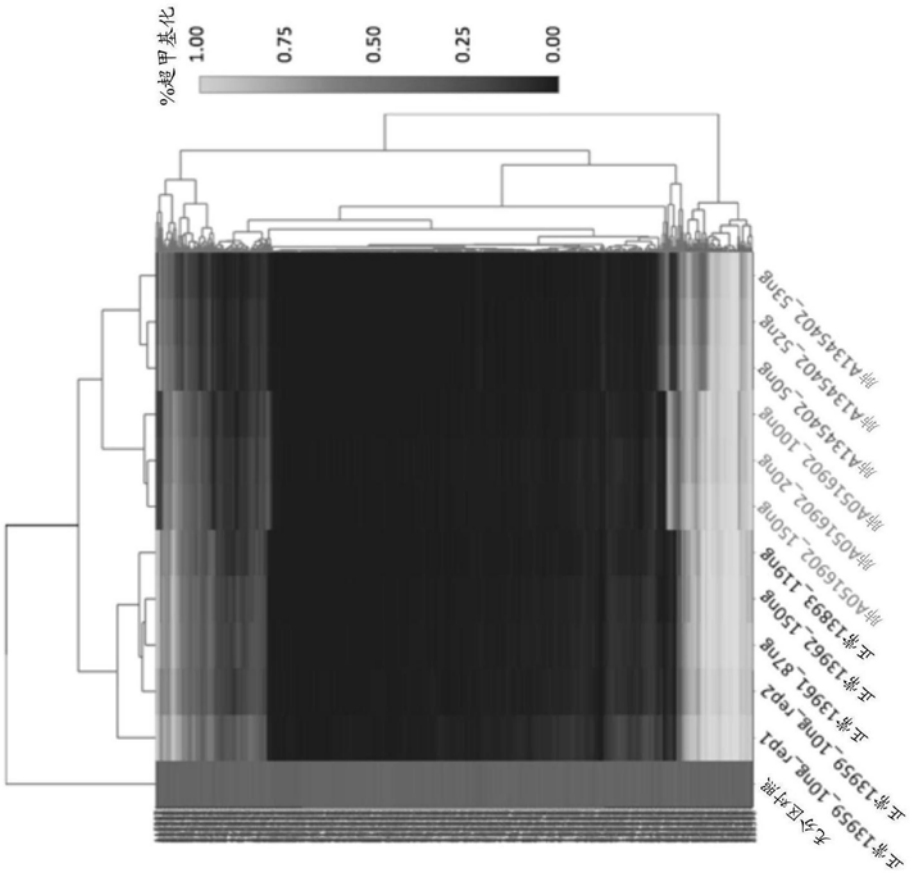

%超甲基化
1.00
0.75
0.50
0.25
0.00
肿A13454O2_53nE
肿A13454O2_52nE
肿A13454O2_50nE
肿A13454O2_100nE
肿A0S16902_20nE
肿A0S16902_15OnE
肿A0S16902_119nE
正常13893_15OnE
正常13962_87nE
正常13961_repZ
正常13959_10nE_repl
正常13959_10nE
无分化肿瘤

图11

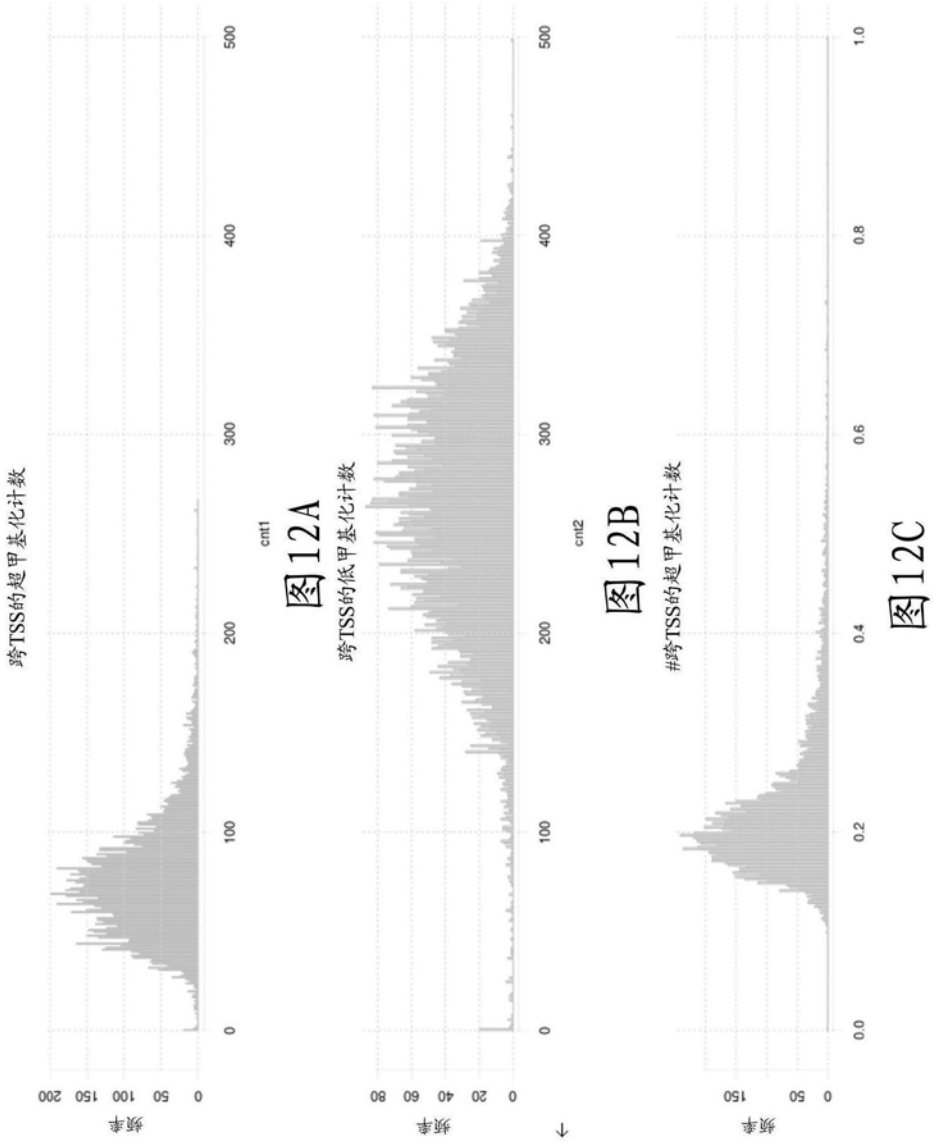


图12A

图12B

图12C

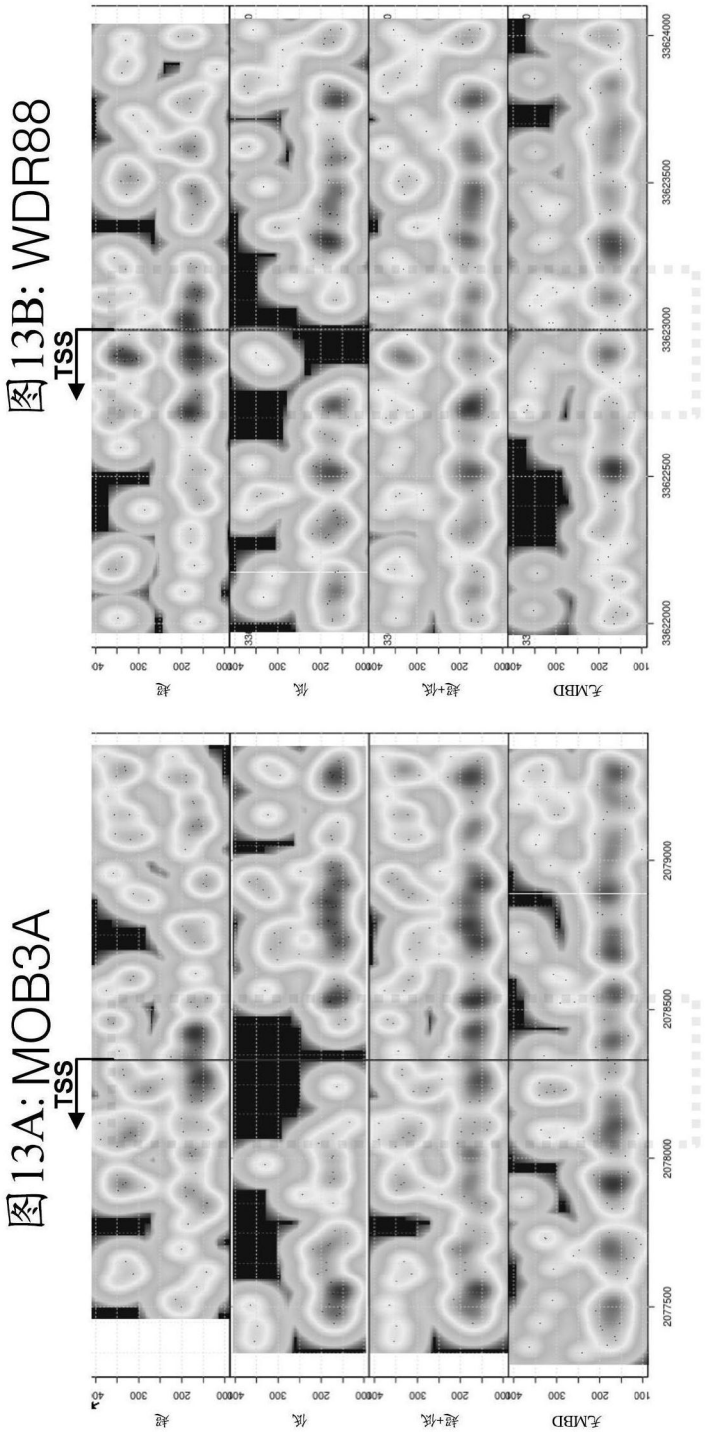


图13B: WDR88
TSS
33624000
33623500
33623000
33622500
33622000
ChIP
Input
ChIP+Input
无MBD
图13A: MOB3A
TSS
2079000
2078500
2078000
2077500
ChIP
Input
ChIP+Input
无MBD

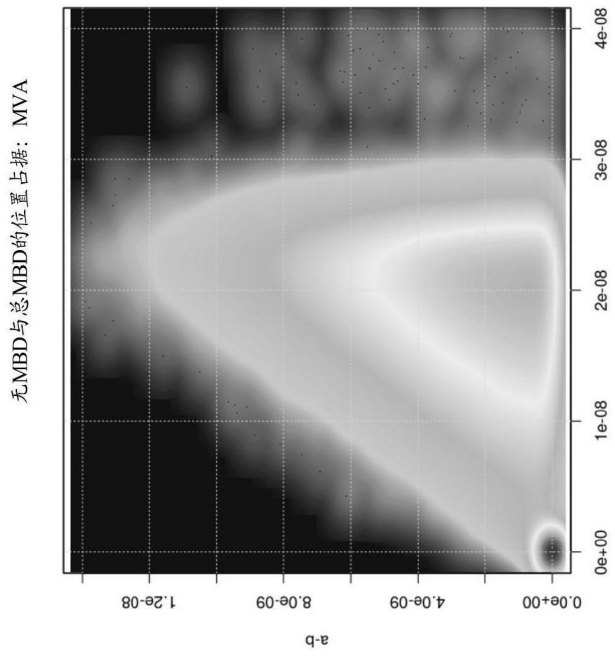

无MBD与总MBD的位置占据：MVA
0e+00
1e-08
2e-08
3e-08
4e-08
(a+b)/2
0.0e+00
4.0e-09
8.0e-09
1.2e-08
a-b

图14B

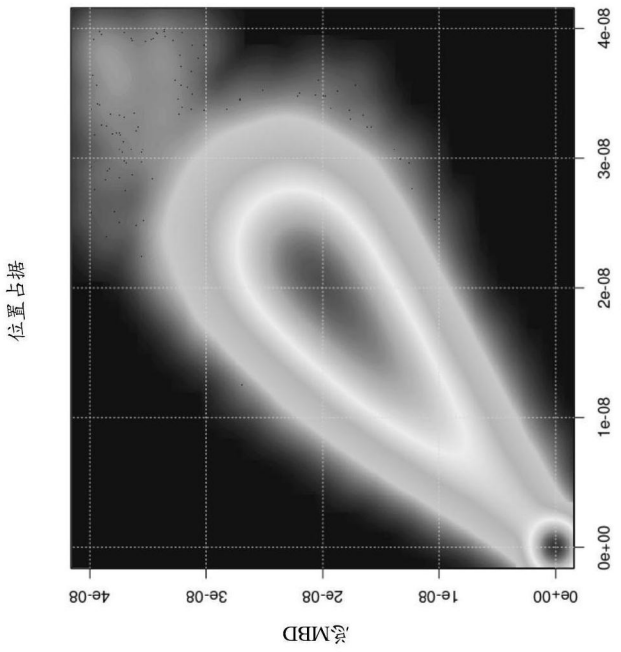

位置占据
0e+00
1e-08
2e-08
3e-08
4e-08
无MBD
0e+00
1e-08
2e-08
3e-08
4e-08
总MBD

图14A

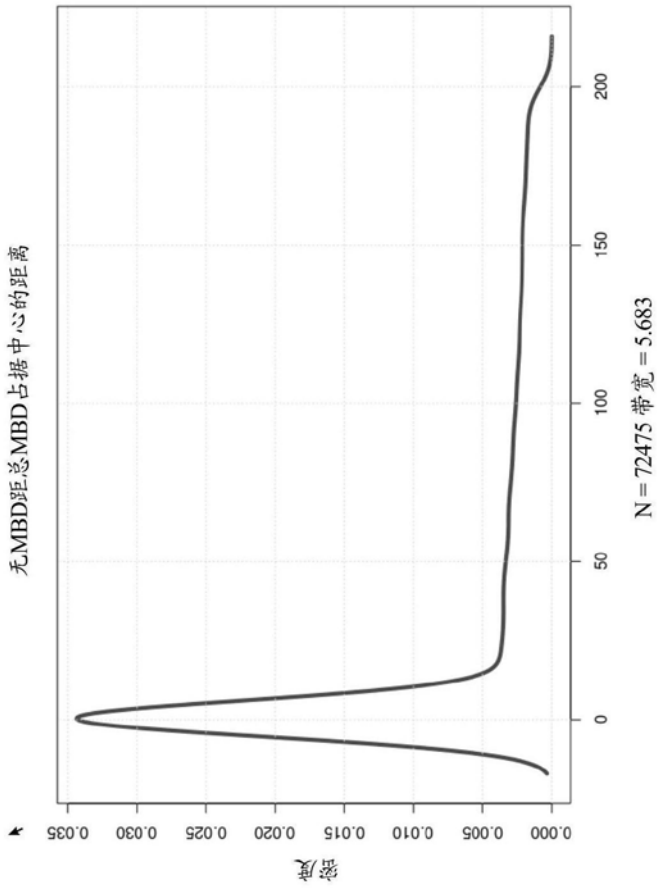

无MBD距总MBD占据中心的距离
密度
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0
50
100
150
200
N = 72475 带宽 = 5.683

图15

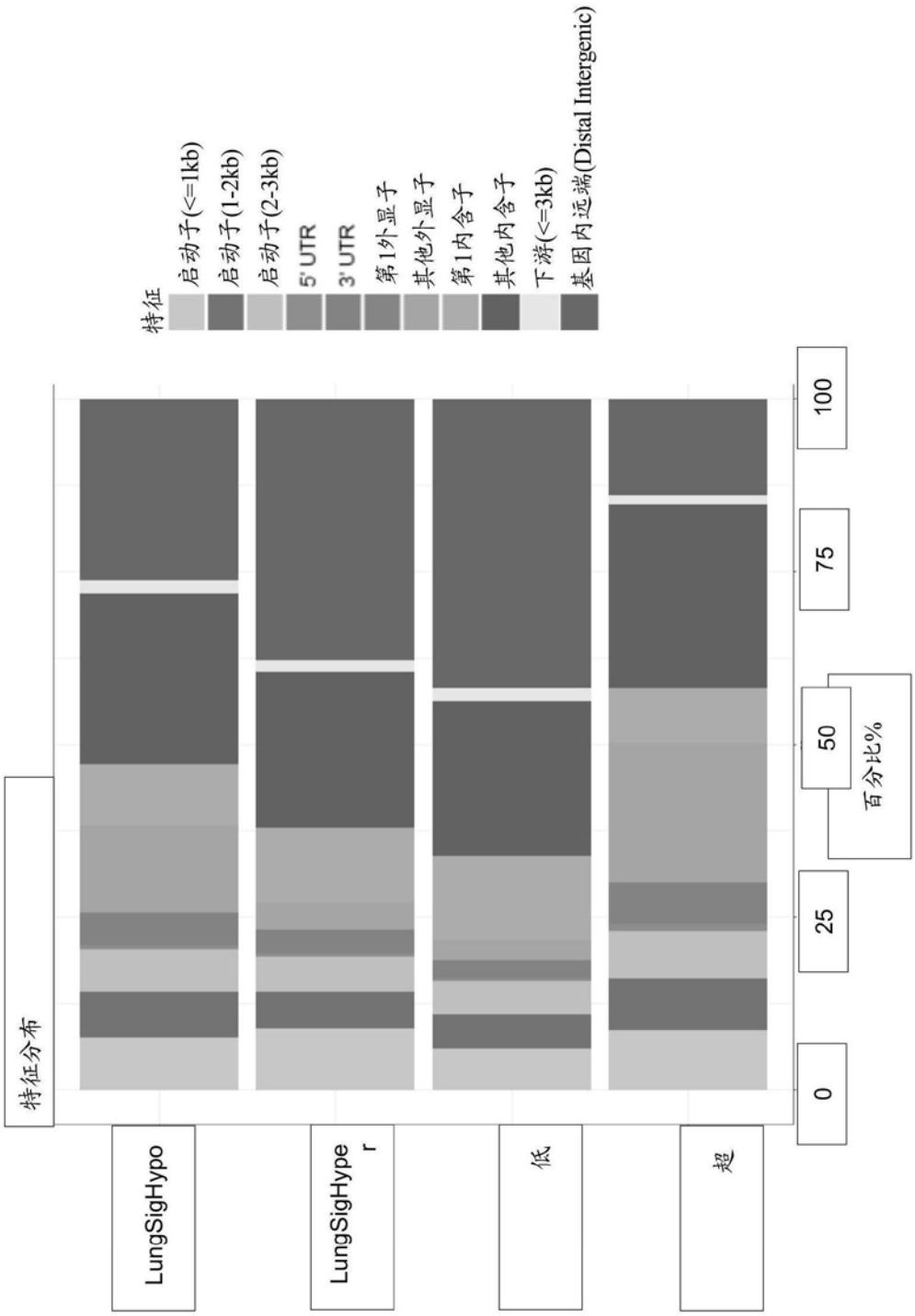

特征
启动子(<=1kb)
启动子(1-2kb)
启动子(2-3kb)
5' UTR
3' UTR
第1外显子
其他外显子
第1内含子
其他内含子
下游(<=3kb)
基因内远端(Distal Intergenic)
特征分布
LungSigHypo
LungSigHyper
低
超
0
25
50
75
100
百分比%

图16

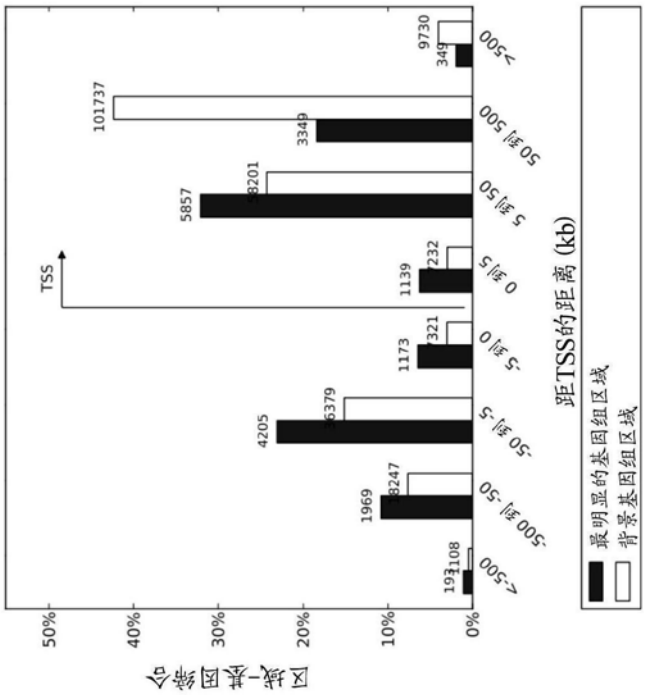
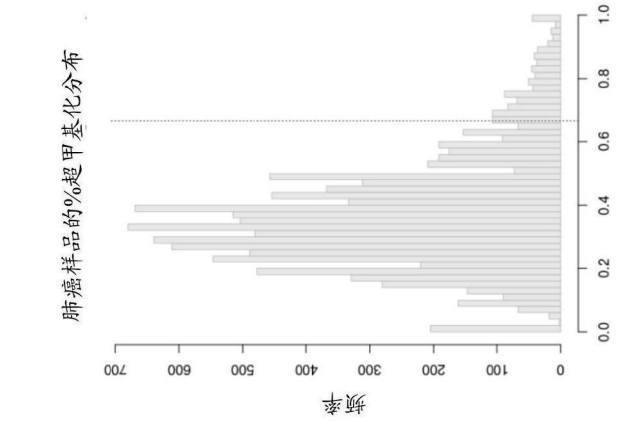
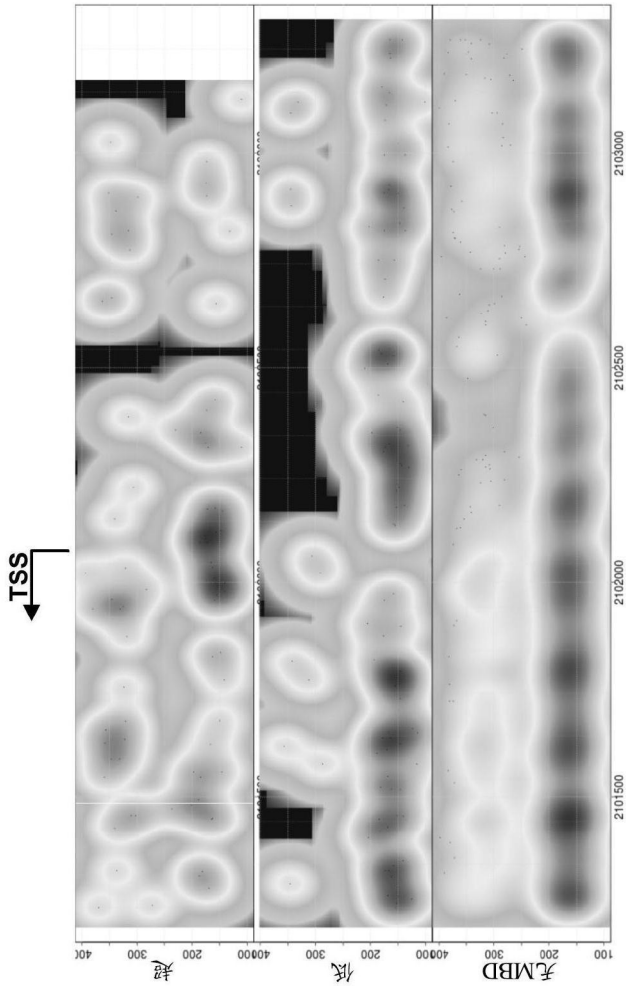

区域-基因综合
50%
40%
30%
20%
10%
0%
TSS
-7500
-5000 到 -50
-50 到 -5
-5 到 0
0 到 5
5 到 50
50 到 500
500 到 7500
9730
349
101737
3349
8201
5857
232
1139
321
1173
6379
4205
3247
1969
193
1108
距TSS的距离 (kb)
最明显的基因组区域
背景基因组区域

图17



AP3D1

图18B



AP3D1

图18A

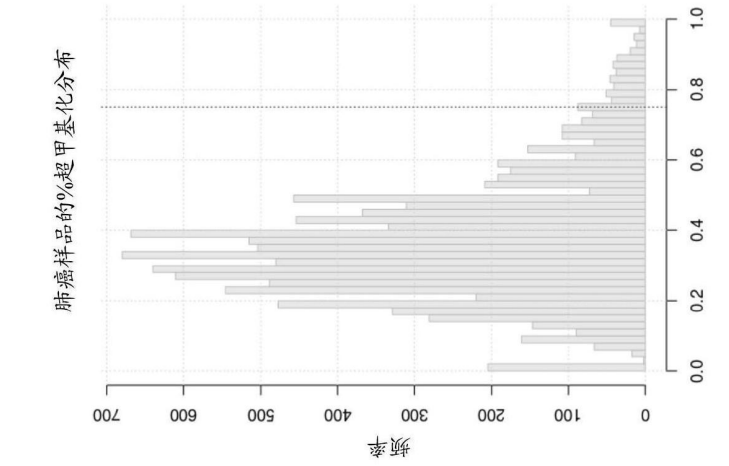

肺癌样品的%超甲基化分布
频率
0
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0
DNMT1

图19B

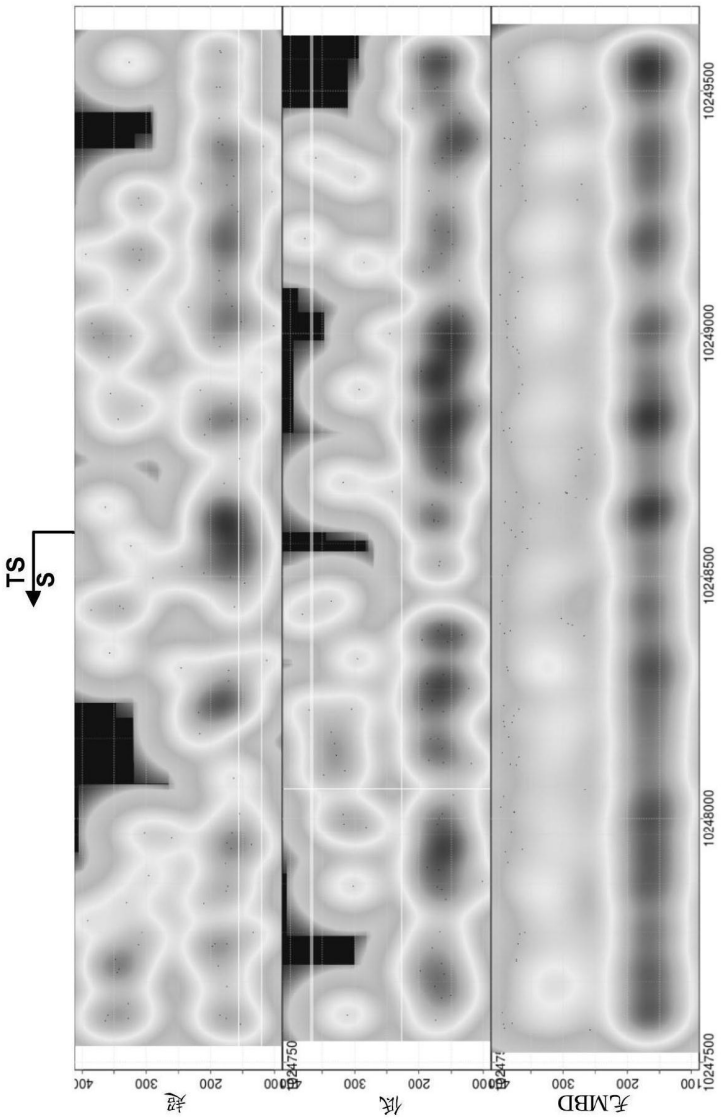

TS
S
无MBD
低
超
10247500
10248000
10248500
10249000
10249500
DNMT1

图19A

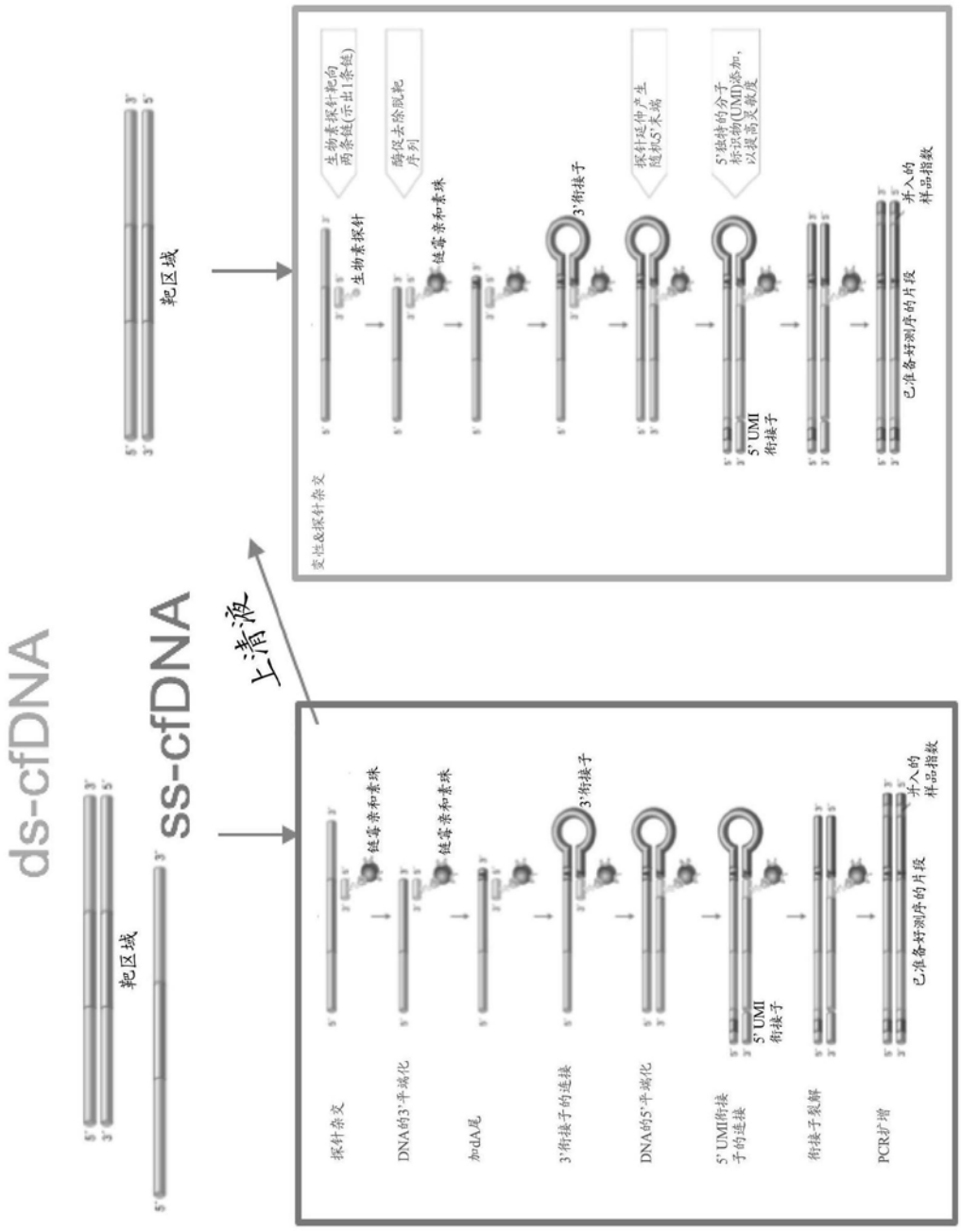

ds-cfDNA
靶区域
SS-cfDNA
靶区域
上清液
探针杂交
链霉亲和素珠
DNA的3'平端化
链霉亲和素珠
加dA尾
3'衔接子的连接
3'衔接子
DNA的5'平端化
5' UMI衔接子的连接
5' UMI
衔接子
衔接子裂解
PCR扩增
已准备好测序的片段
导入的样品指数
变性及探针杂交
生物素探针
生物素探针靶向两条链(示出1条链)
链霉亲和素珠
酶促去除靶序列
3'衔接子
探针延伸产生随机5'末端
5' UMI
衔接子
5'连接的分子标识物(UMI)添加，以提高灵敏度
已准备好测序的片段
导入的样品指数

图20

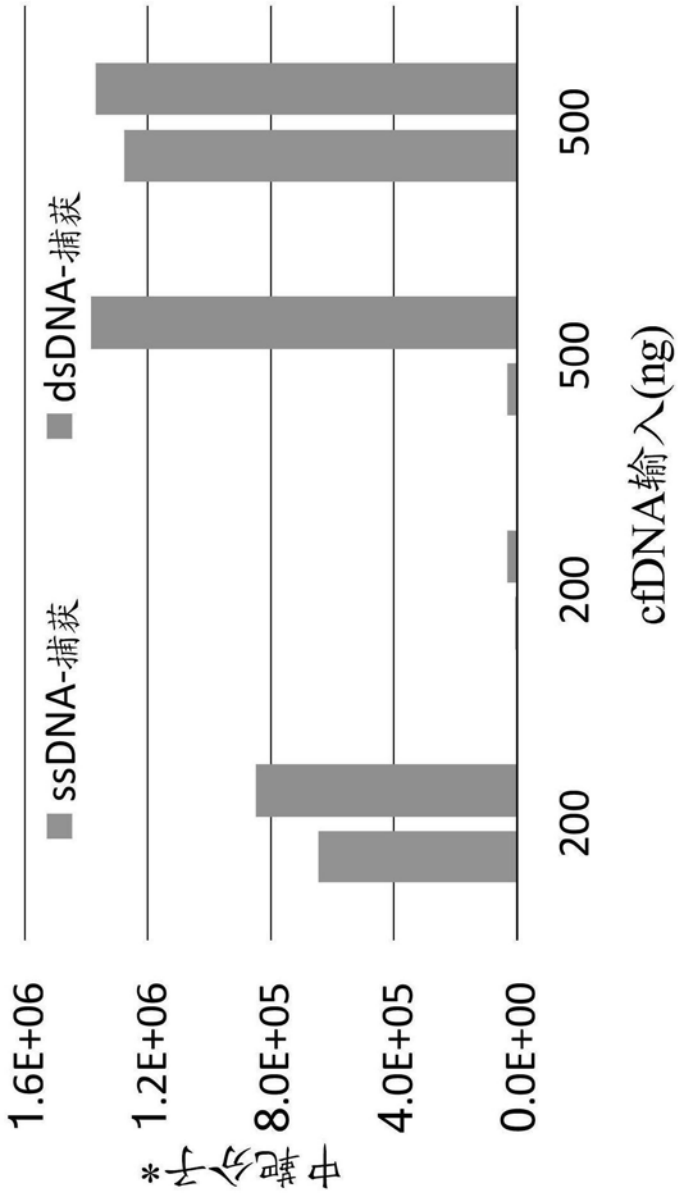
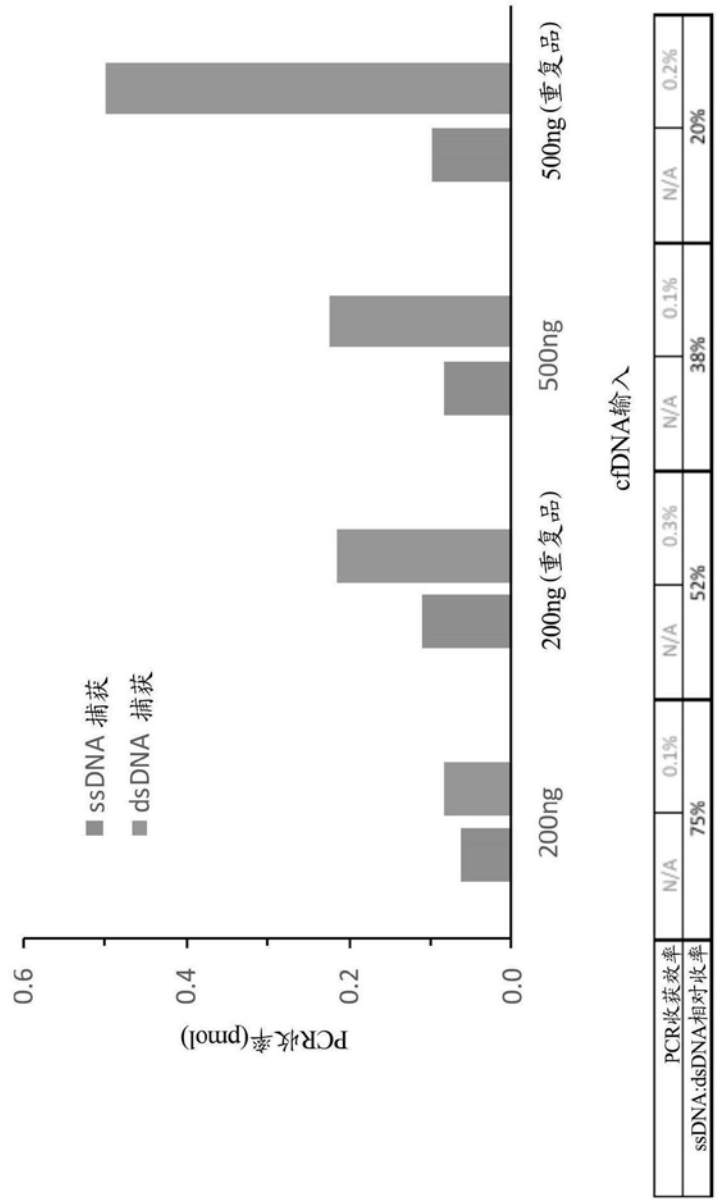

ssDNA-捕获
dsDNA-捕获
1.6E+06
1.2E+06
8.0E+05
4.0E+05
0.0E+00
200
200
500
500
cfDNA输入(ng)

图21



PCR收获效率	N/A	0.1%	N/A	0.3%	N/A	0.1%	N/A	0.2%
ssDNA:dsDNA相对收率	75%		52%		38%		20%	

图22

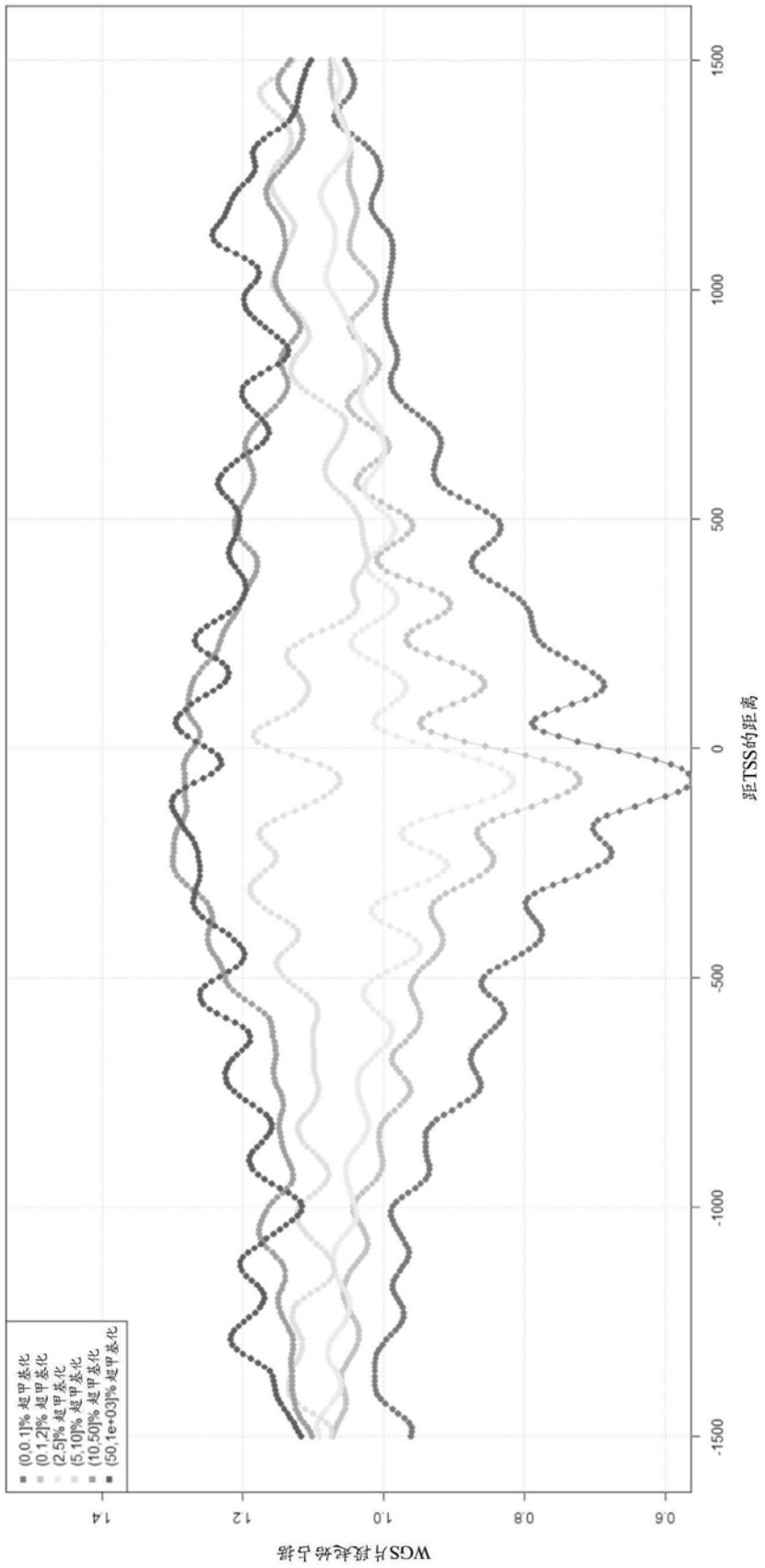

WGS片段起始占据
0.6
0.8
1.0
1.2
1.4
距TSS的距离
-1500
-1000
-500
0
500
1000
1500
(0,0.1]% 超甲基化
(0.1,2]% 超甲基化
(2,5]% 超甲基化
(5,10]% 超甲基化
(10,50]% 超甲基化
(50,1e+03]% 超甲基化

图23

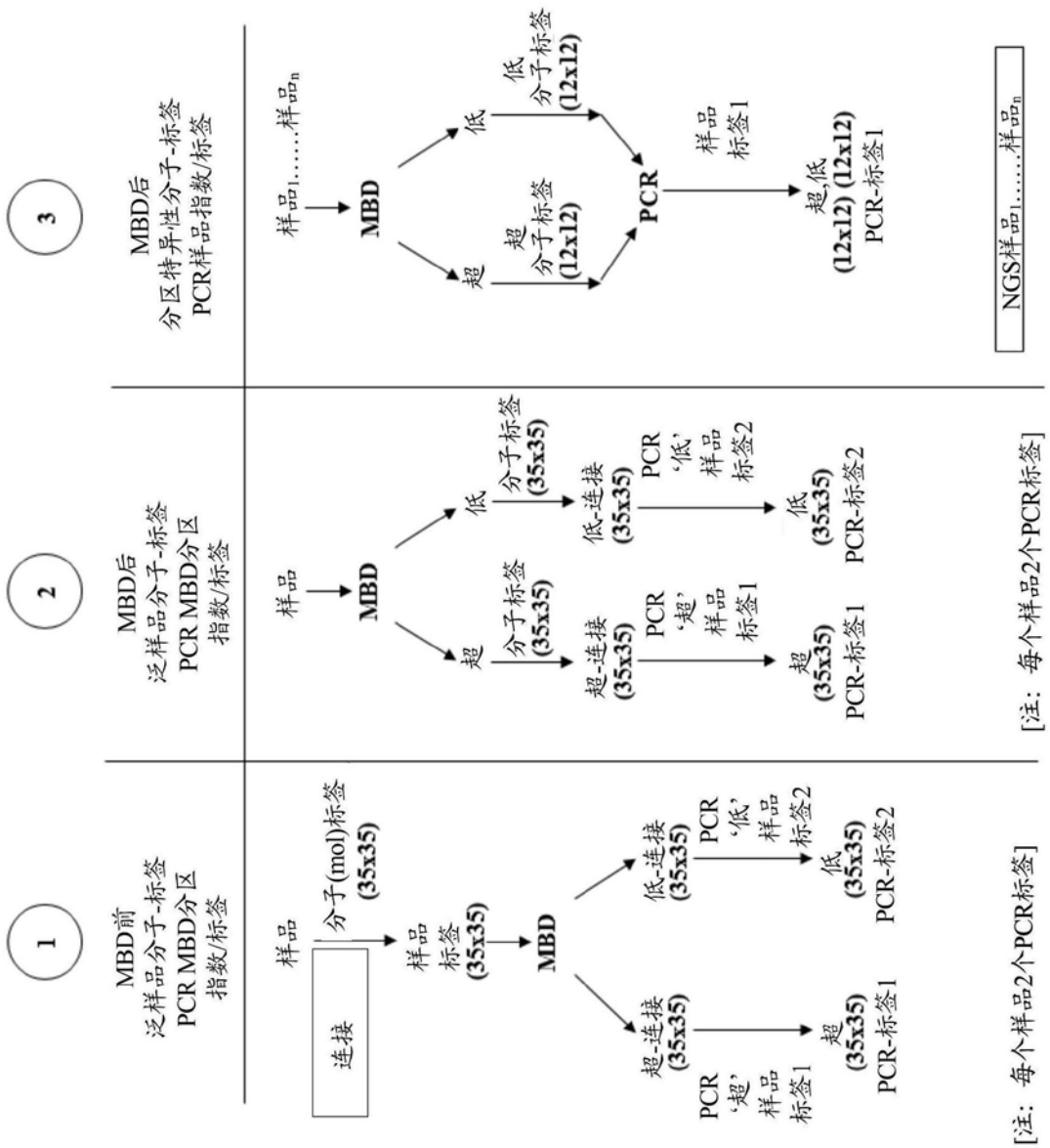

1
MBD前
泛样品分子-标签
PCR MBD分区
指数/标签
样品
连接
分子(mol)标签
(35x35)
样品
标签
(35x35)
MBD
超-连接
(35x35)
PCR
‘超’
样品
标签1
超
(35x35)
PCR-标签1
低-连接
(35x35)
PCR
‘低’
样品
标签2
低
(35x35)
PCR-标签2
[注：每个样品2个PCR标签]
2
MBD后
泛样品分子-标签
PCR MBD分区
指数/标签
样品
MBD
超
分子标签
(35x35)
超-连接
(35x35)
PCR
‘超’
样品
标签1
超
(35x35)
PCR-标签1
低
分子标签
(35x35)
低-连接
(35x35)
PCR
‘低’
样品
标签2
低
(35x35)
PCR-标签2
[注：每个样品2个PCR标签]
3
MBD后
分区特异性分子-标签
PCR样品指数/标签
样品1.....样品n
MBD
超
分子标签
(12x12)
低
分子标签
(12x12)
PCR
样品
标签1
超,低
(12x12) (12x12)
PCR-标签1
NGS样品1.....样品n

图24

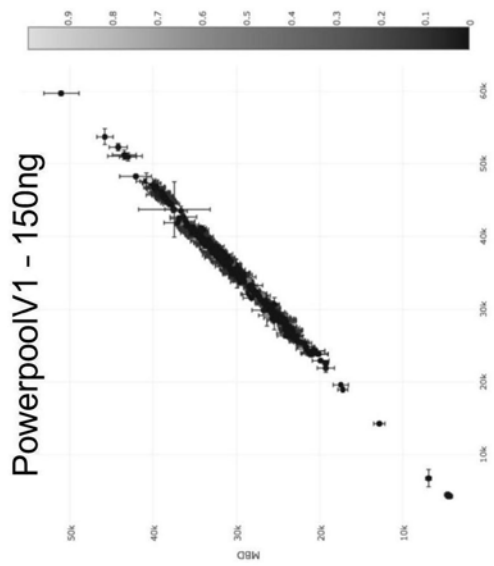


图25B

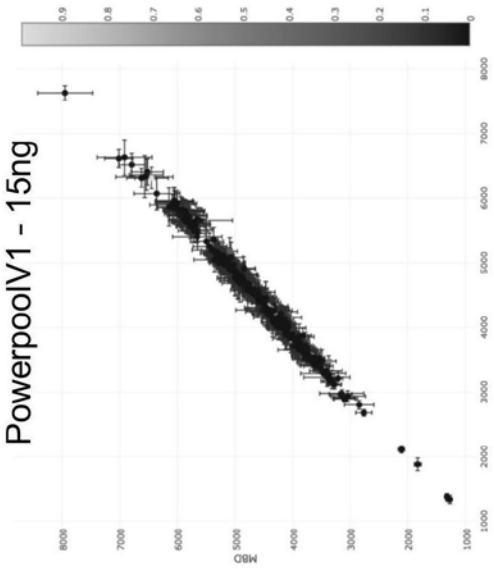


图25A

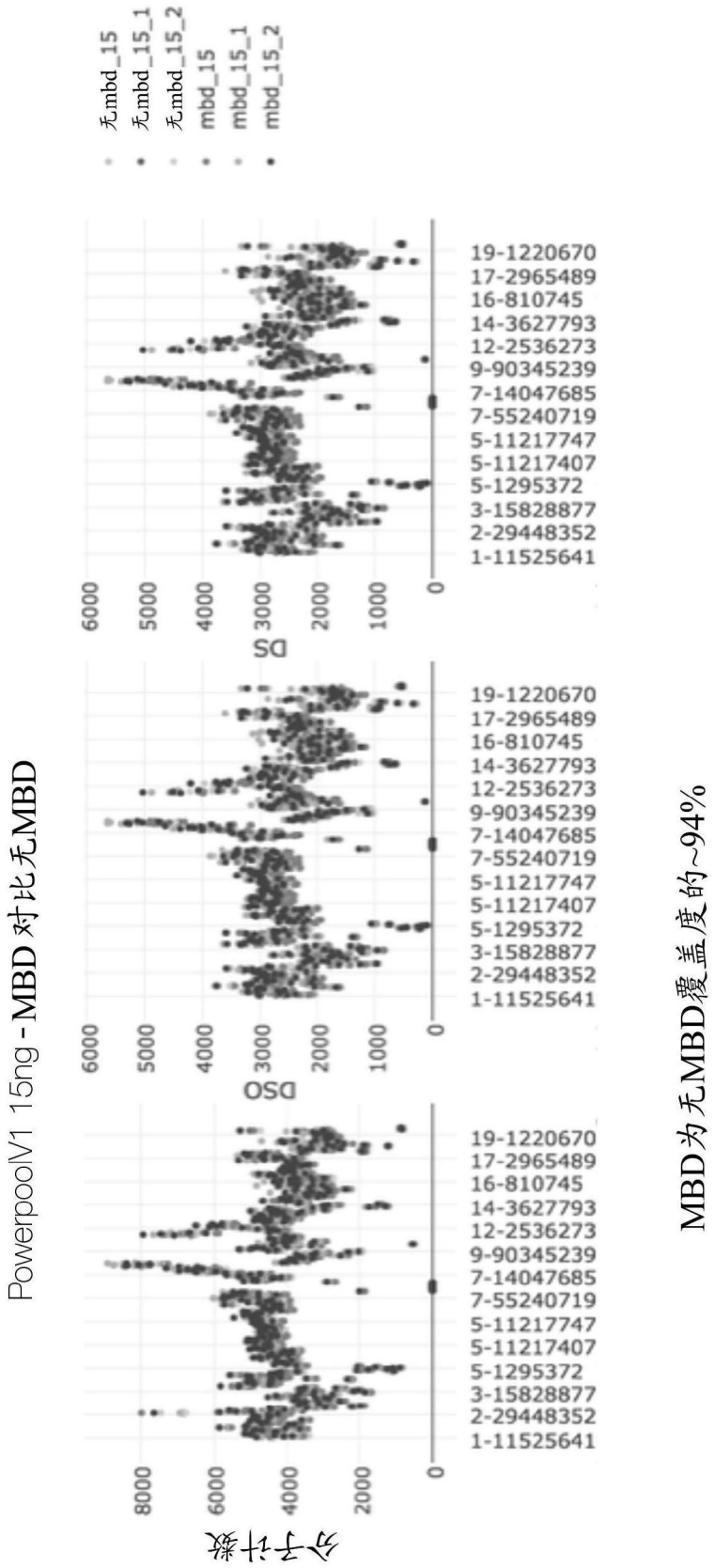

PowerpoolV1 15ng - MBD 对比无MBD
无mbd_15
无mbd_15_1
无mbd_15_2
mbd_15
mbd_15_1
mbd_15_2
DS
DSO
分子计数
19-1220670
17-2965489
16-810745
14-3627793
12-2536273
9-90345239
7-14047685
7-55240719
5-11217747
5-11217407
5-1295372
3-15828877
2-29448352
1-11525641
MBD为无MBD覆盖度的~94%

图26A

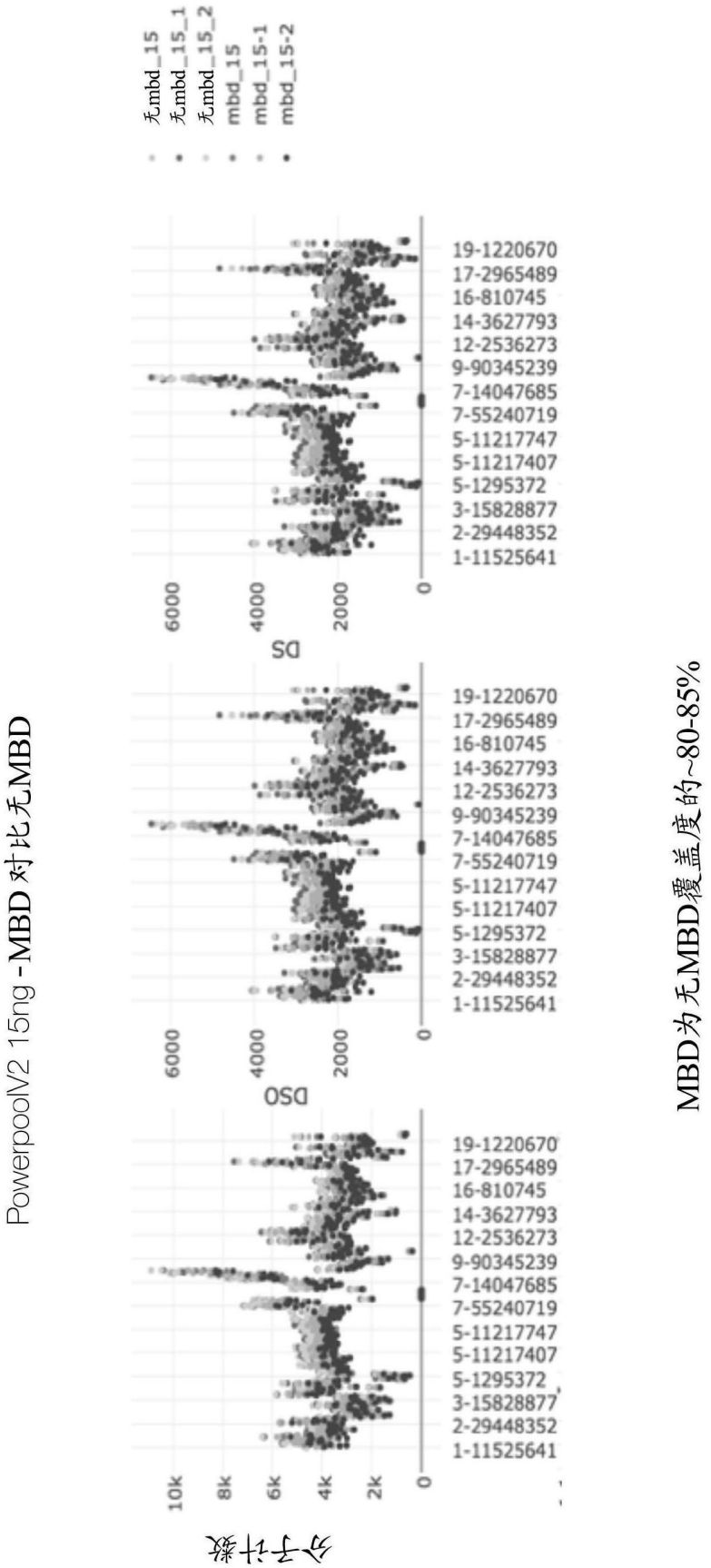

PowerpoolV2 15ng -MBD 对比无MBD
MBD为无MBD覆盖度的~80-85%
无mbd_15
无mbd_15_1
无mbd_15_2
mbd_15
mbd_15-1
mbd_15-2
DS
DSO
分子计数
19-1220670
17-2965489
16-810745
14-3627793
12-2536273
9-90345239
7-14047685
7-55240719
5-11217747
5-11217407
5-1295372
3-15828877
2-29448352
1-11525641
6000
4000
2000
0
10k
8k
6k
4k
2k

图26B

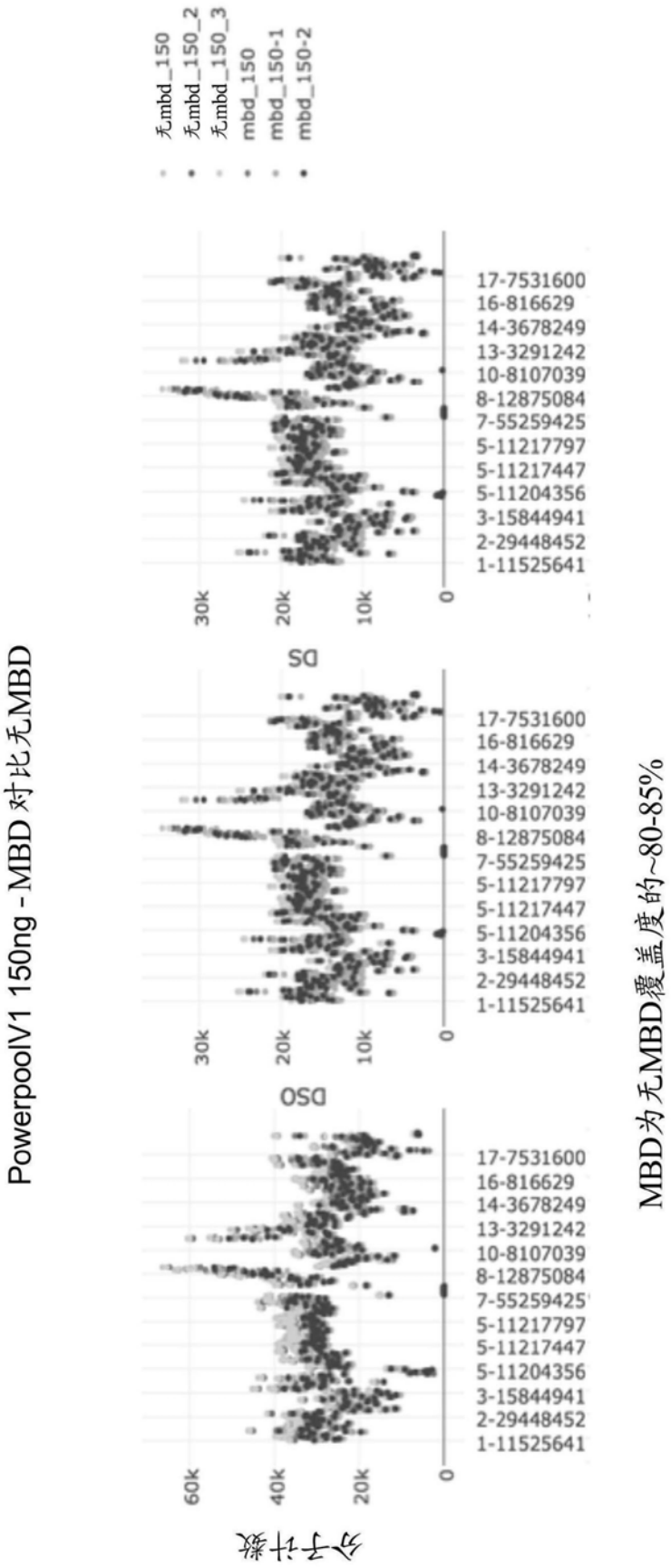

PowerpoolV1 150ng - MBD 对比无MBD
MBD为无MBD覆盖度的~80-85%
分子计数
无mbd_150
无mbd_150_2
无mbd_150_3
mbd_150
mbd_150-1
mbd_150-2
17-7531600
16-816629
14-3678249
13-3291242
10-8107039
8-12875084
7-55259425
5-11217797
5-11217447
5-11204356
3-15844941
2-29448452
1-11525641
60k
40k
20k
0
30k
20k
10k
0

图27A

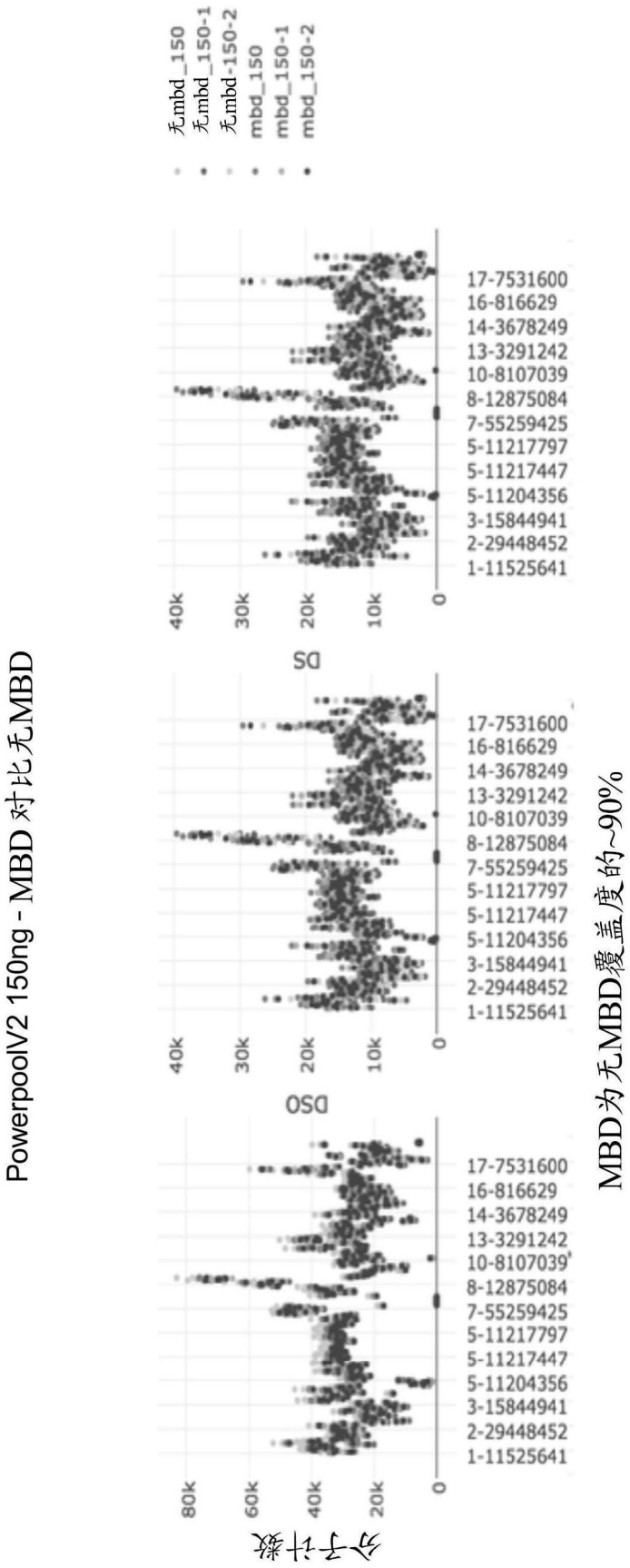

PowerpoolV2 150ng - MBD 对比无MBD
无mbd_150
无mbd_150-1
无mbd-150-2
mbd_150
mbd_150-1
mbd_150-2
DS
DSO
17-7531600
16-816629
14-3678249
13-3291242
10-8107039
8-12875084
7-55259425
5-11217797
5-11217447
5-11204356
3-15844941
2-29448452
1-11525641
40k
30k
20k
10k
0
80k
60k
40k
20k
分子计数
MBD为无MBD覆盖度的~90%

图27B

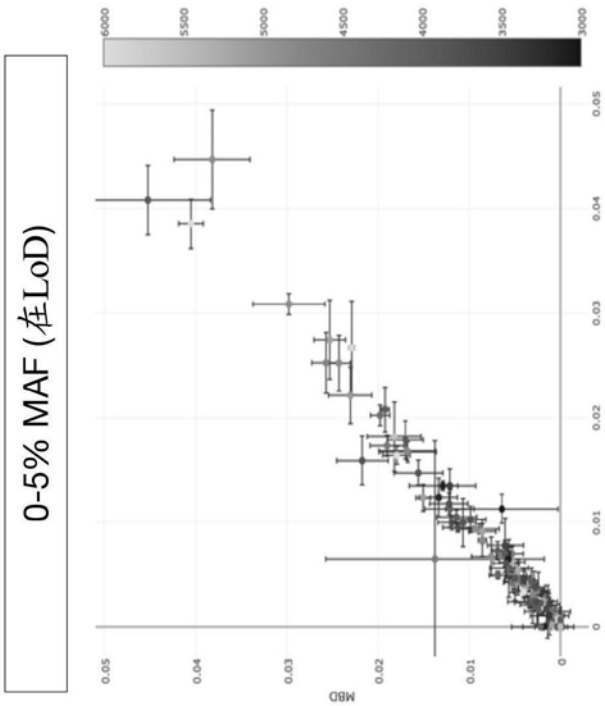


图28B

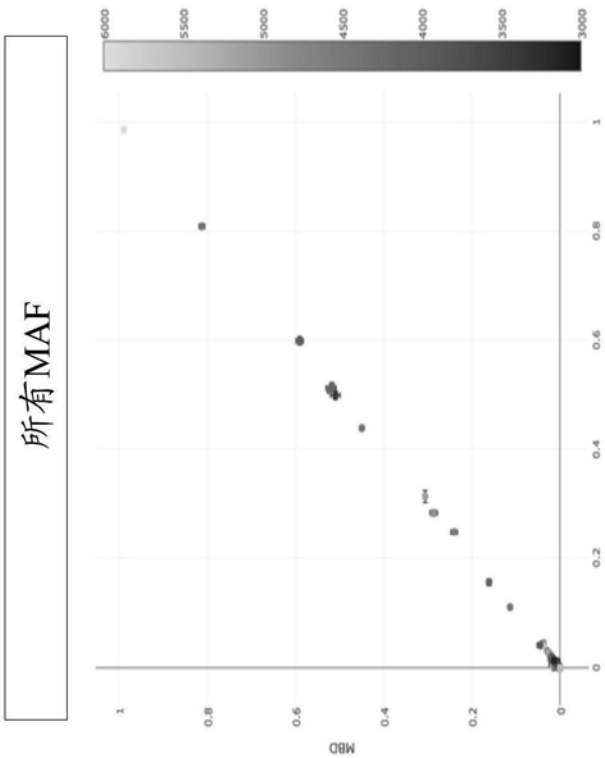


图28A

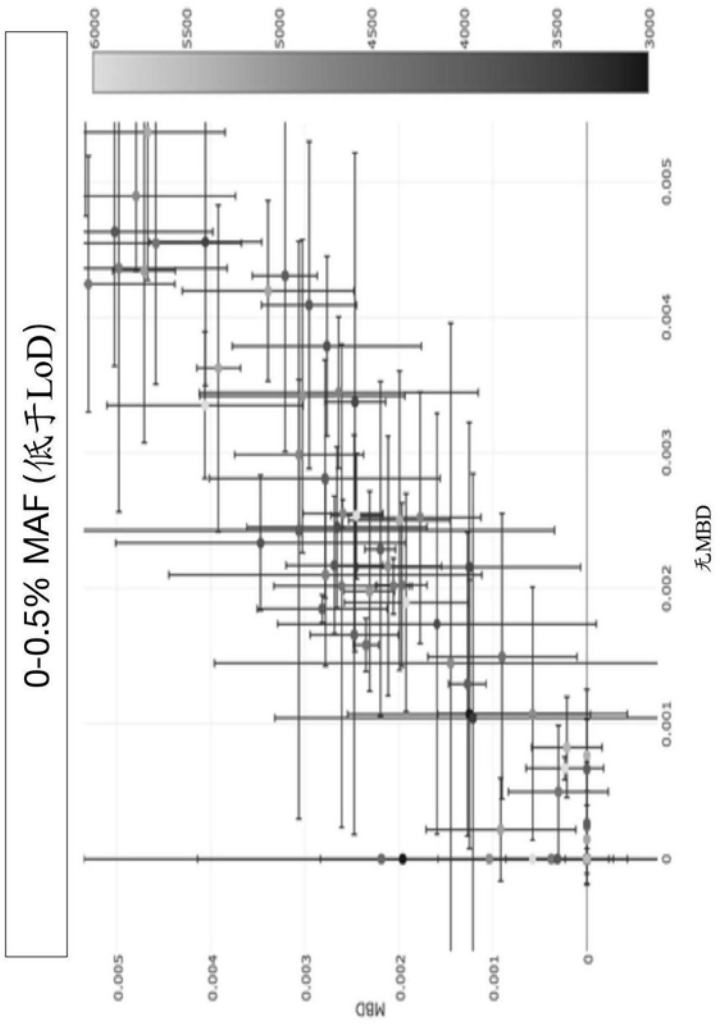

0-0.5% MAF (低于LoD)
6000
5500
5000
4500
4000
3500
3000
0.005
0.004
0.003
0.002
0.001
0
无MBD
MBD

图28C

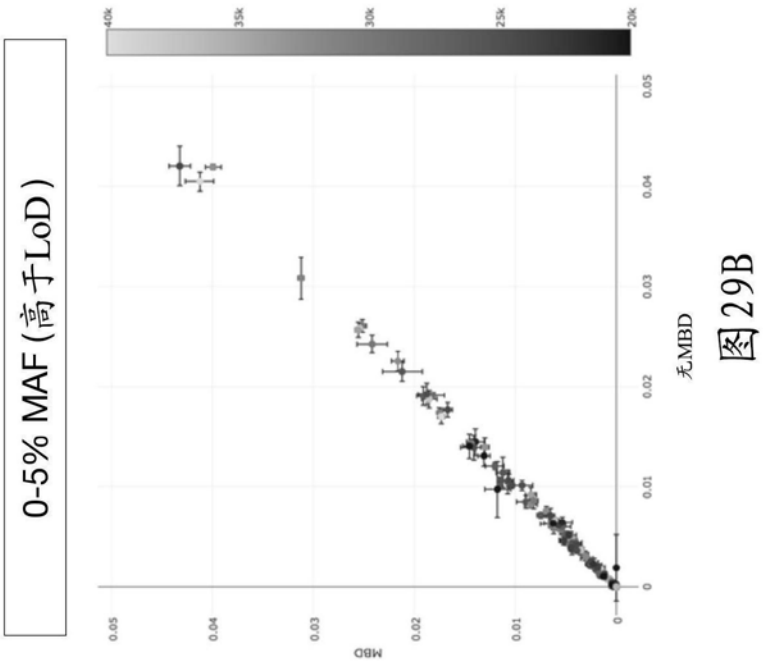

0-5% MAF (高于LoD)
MBD
无MBD
40k
35k
30k
25k
20k
0
0.01
0.02
0.03
0.04
0.05

图29B

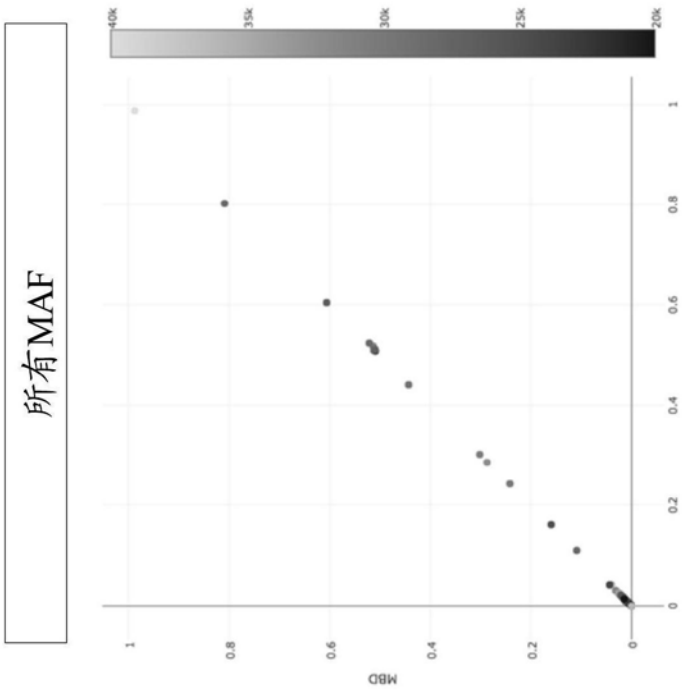

所有MAF
MBD
无MBD
40k
35k
30k
25k
20k
0
0.2
0.4
0.6
0.8
1

图29A

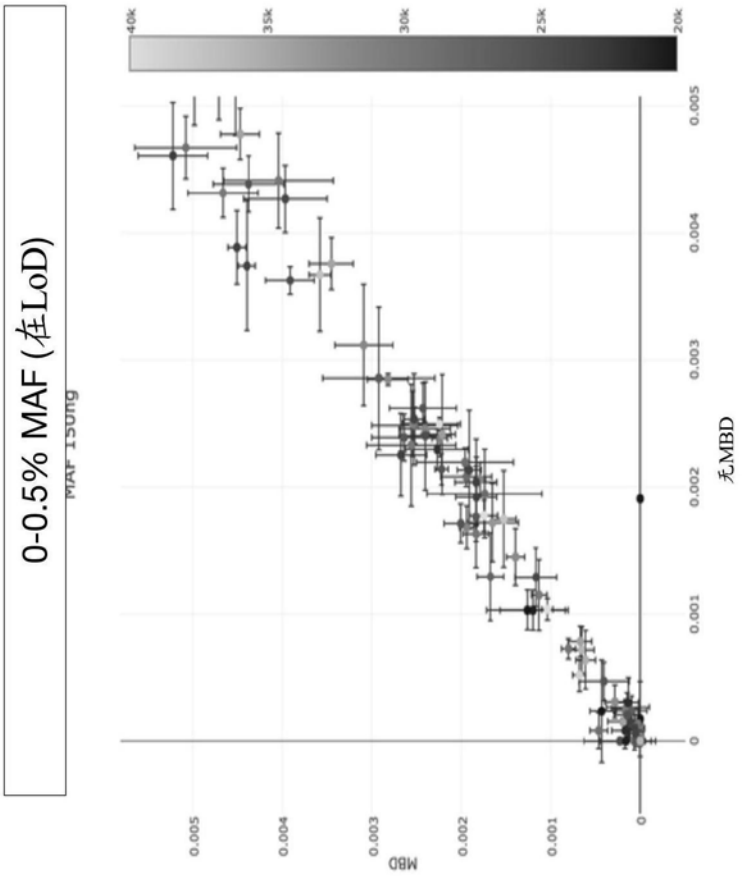

0-0.5% MAF (在LoD)
MAF 150ng
40k
35k
30k
25k
20k
0.005
0.004
0.003
0.002
0.001
0
MBD
无MBD

图29C

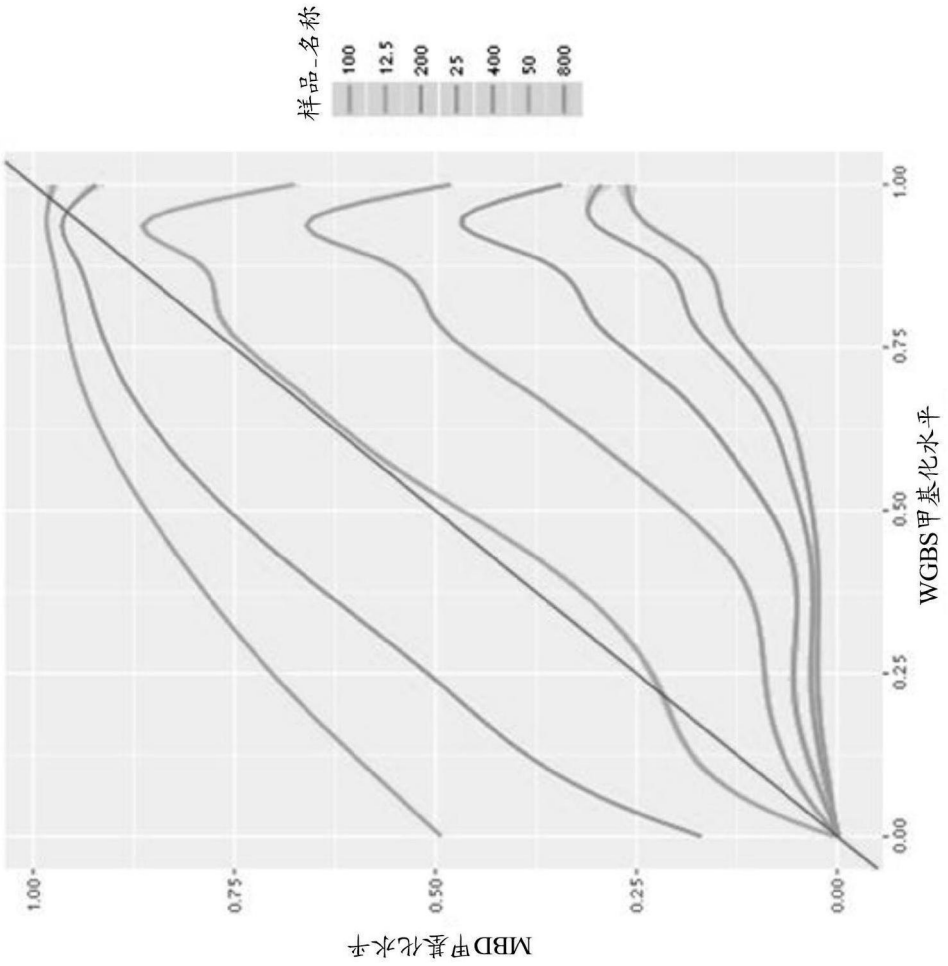

样品_名称
100
12.5
200
25
400
50
800
1.00
0.75
0.50
0.25
0.00
WGBS甲基化水平
MBD甲基化水平

图30

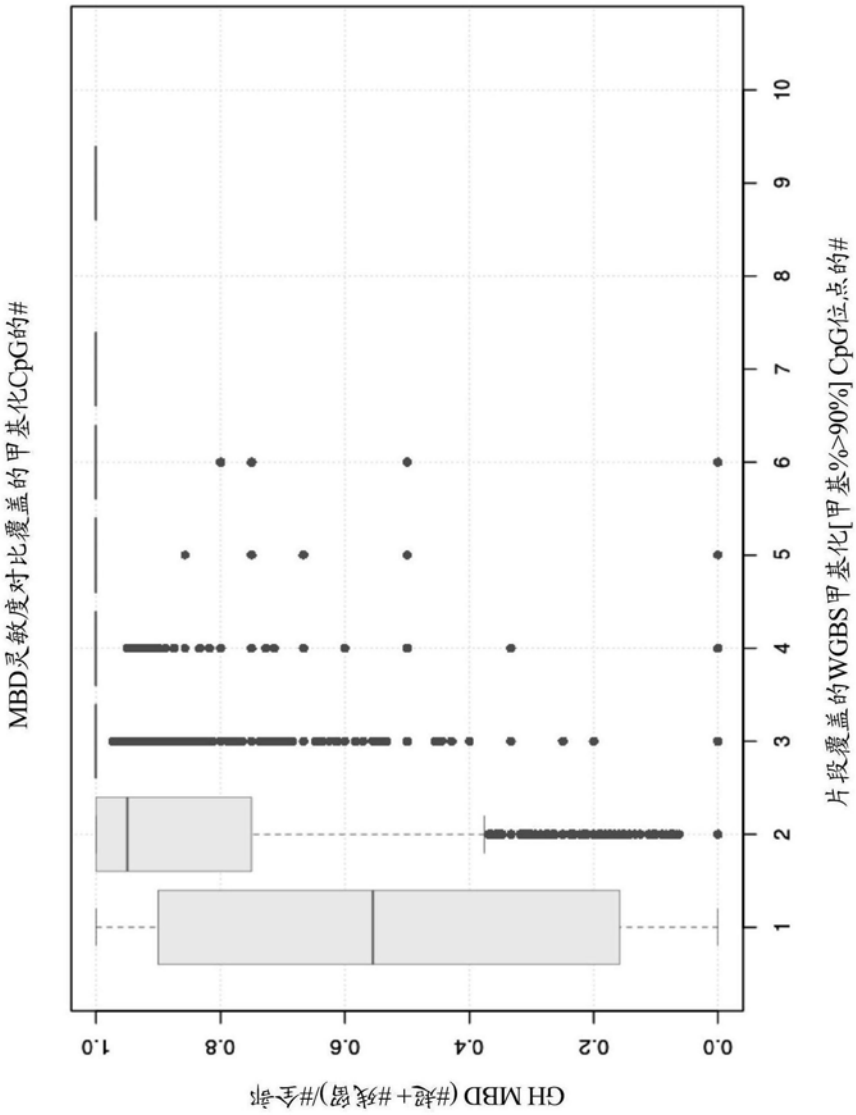

MBD灵敏度对比覆盖的甲基化CpG的#
GH MBD (#起+ #残留)/#全部
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
片段覆盖的WGBS甲基化[甲基%>90%] CpG位点的#

图31A

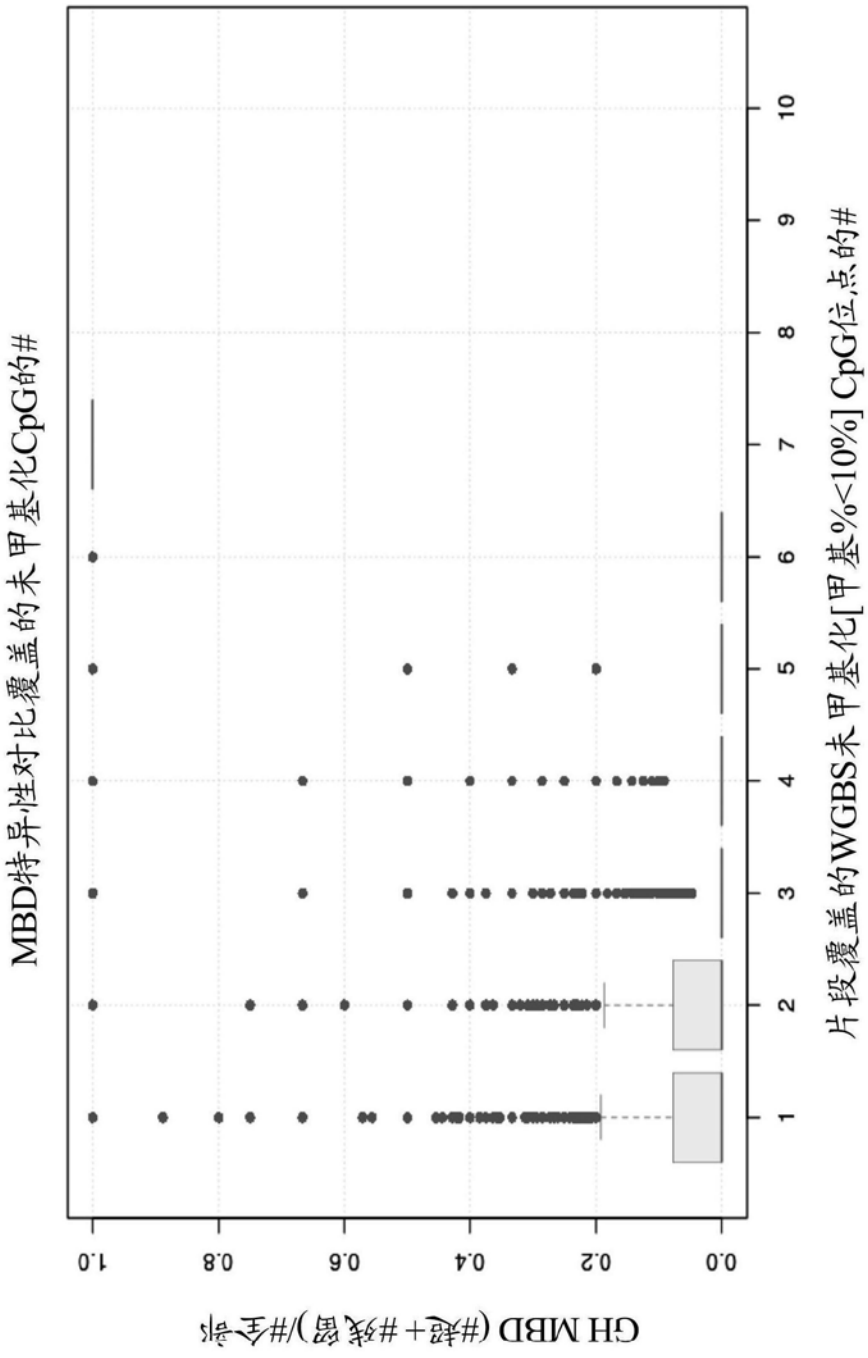

GH MBD (#超+ #残留)/#全部
0.0
0.2
0.4
0.6
0.8
1.0
MBD特异性对比覆盖的未甲基化CpG的#
1
2
3
4
5
6
7
8
9
10
片段覆盖的WGBS未甲基化[甲基%<10%] CpG位点的#

图31B

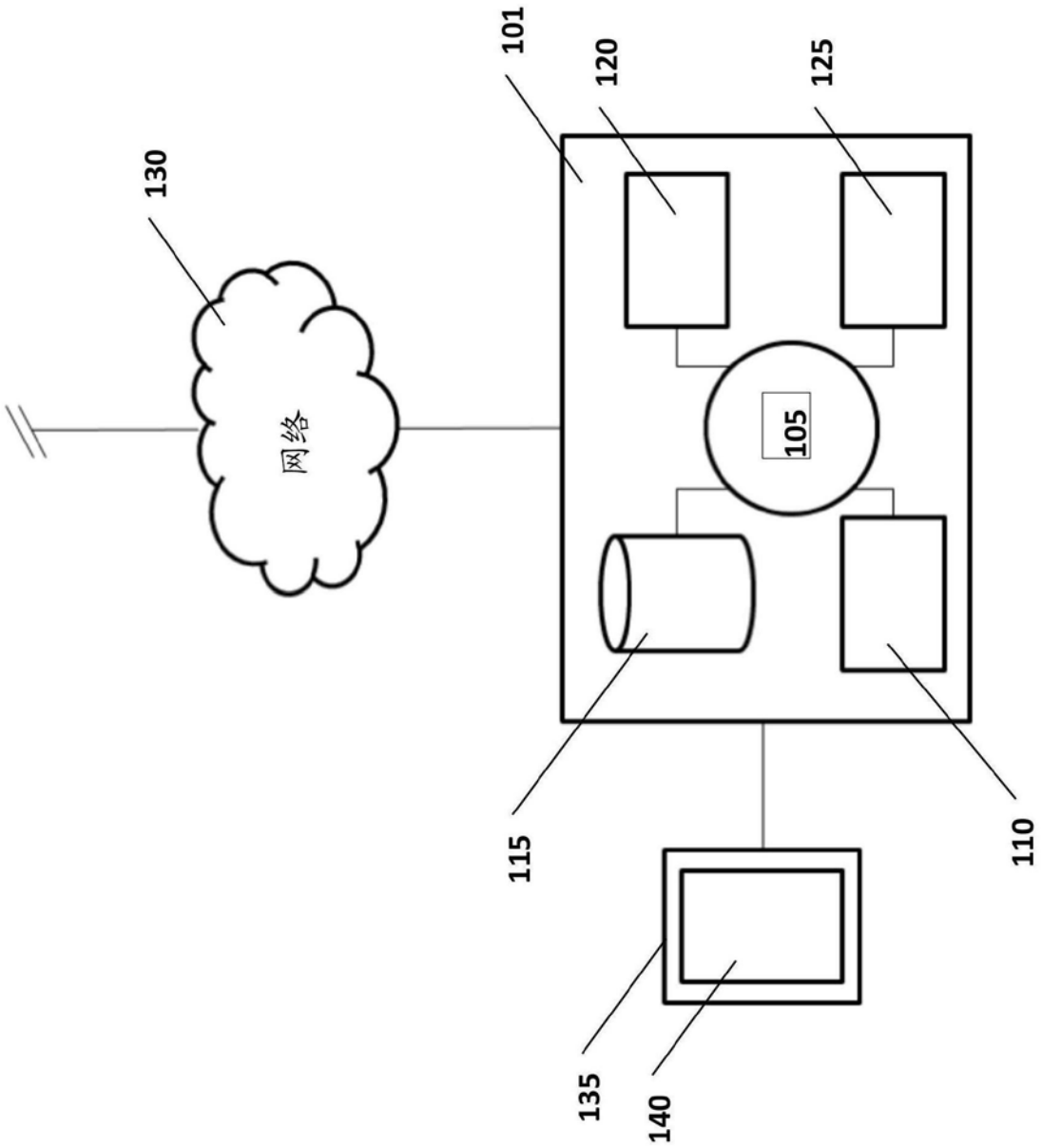


图32

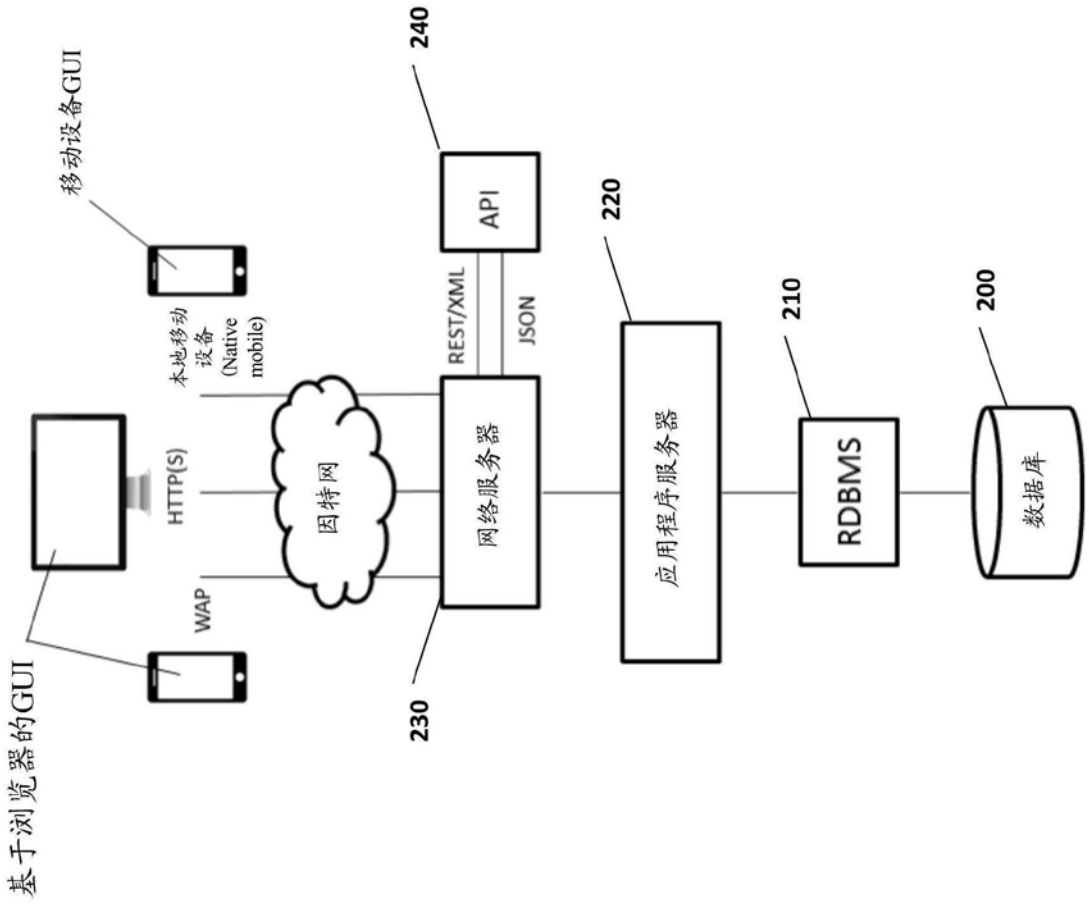

移动设备GUI
基于浏览器的GUI
本地移动设备 (Native mobile)
HTTP(S)
WAP
因特网
网络服务器
API
REST/XML
JSON
应用程序服务器
RDBMS
数据库
240
230
220
210
200

图33

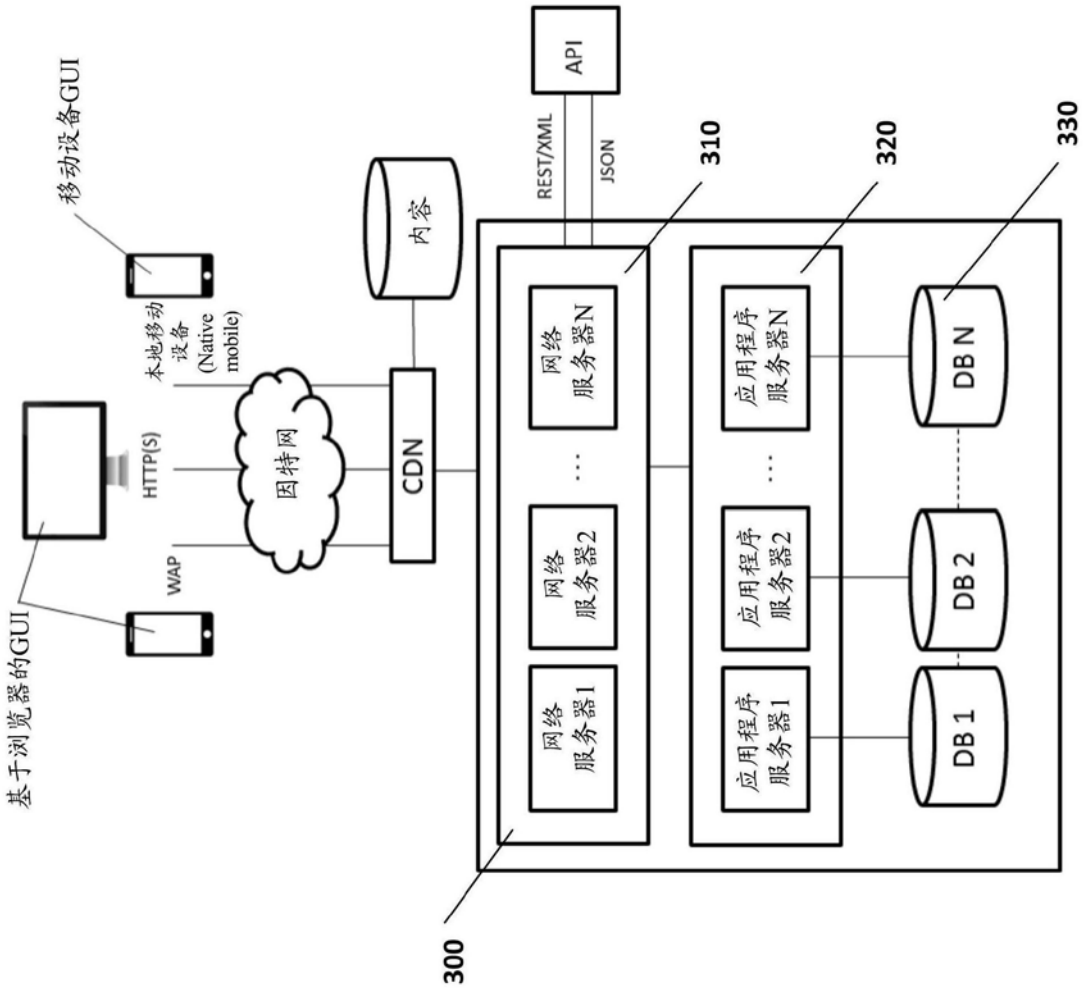

移动设备GUI
本地移动设备(Native mobile)
HTTP(S)
WAP
基于浏览器的GUI
因特网
CDN
内容
API
REST/XML
JSON
310
320
330
300
网络服务器N
网络服务器2
网络服务器1
应用程序服务器N
应用程序服务器2
应用程序服务器1
DB N
DB 2
DB 1

图34