(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 112823213 A
(43)申请公布日 2021.05.18

(21)申请号 202080005637.3
(22)申请日 2020.05.29
(30)优先权数据
62/855,795 2019.05.31 US
(85)PCT国际申请进入国家阶段日
2021.04.09
(86)PCT国际申请的申请数据
PCT/US2020/035380 2020.05.29
(87)PCT国际申请的公布数据
WO2020/243609 EN 2020.12.03
(71)申请人 福瑞诺姆控股公司
地址 美国加利福尼亚州
(72)发明人 埃里克·阿里亚兹 大卫·温伯格 格雷格·霍根 约翰·圣约翰 迈克尔·皮尔森
(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 武晶晶
(51)Int.Cl.
C12Q 1/6869(2018.01)

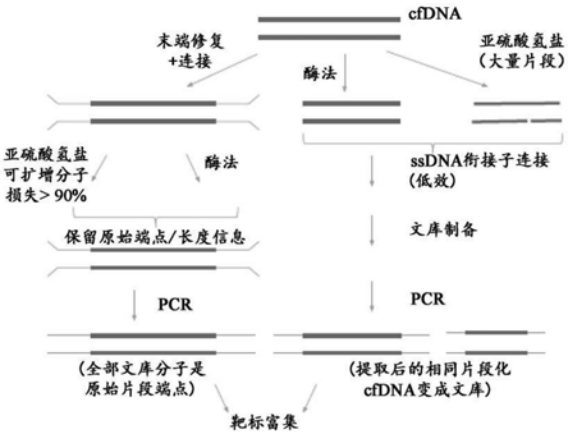
权利要求书5页 说明书38页 附图3页

(54)发明名称

用于甲基化核酸的高深度测序的方法和系统

(57)摘要

本文中提供的方法和系统通过改进核酸甲基化测序的质量和准确度及其检测疾病的用途而解决了目前基于亚硫酸氢盐的甲基化测序的局限性。包括用于甲基化测序的破坏性最小的转化方法以及特化UMI衔接子的方法提供了改进的测序文库和测序信息的质量。更高的准确度和更全面的甲基化状态信息允许产生更高质量的特征以供用于产生机器学习模型和分类器。



CN 112823213 A

1.一种对生物样品的核酸分子进行甲基化测序的方法,包括:

a)将包含唯一分子标识符的核酸衔接子连接到所述核酸分子,其中所述核酸分子包含未转化的核酸;

b)利用破坏性最小的转化方法将所述核酸分子中未甲基化的胞嘧啶转化成尿嘧啶,由此产生转化的核酸;

c)通过聚合酶链反应对所述转化的核酸进行扩增,由此产生扩增的转化的核酸;

d)用核酸探针探查所述扩增的转化的核酸,所述核酸探针与CpG或CH基因座的预鉴定的组合互补以富集与所述组合对应的序列,由此产生探查的转化的核酸;

e)以>100x的深度确定所述探查的转化的核酸的核酸序列;和

f)将所述探查的转化的核酸的所述核酸序列与CpG或CH基因座的所述预鉴定的组合的参考核酸序列进行比较,以确定所述生物样品的所述核酸分子的甲基化图谱。

2.根据权利要求1所述的方法,其中所述核酸分子是血浆cfDNA。

3.根据权利要求1所述的方法,其中所述破坏性最小的转化方法包括酶法转化、TAPS或CAPS。

4.根据权利要求1所述的方法,其中所述唯一分子标识符的长度为4bp至6bp并且具有5'胸苷突出端。

5.根据权利要求4所述的方法,其中所述核酸衔接子还包含唯一双重索引(UDI)序列。

6.根据权利要求5所述的方法,其中所述UDI序列的长度为4bp、5bp、6bp、7bp、8bp、9bp、10bp、11bp或12bp。

7.根据权利要求1所述的方法,其中扩增所述转化的核酸包括使用引物,所述引物包含唯一双重索引(UDI)序列。

8.根据权利要求1所述的方法,其中所述核酸衔接子是耐转化的衔接子,所述耐转化的衔接子包含鸟嘌呤、胸腺嘧啶、腺嘌呤和胞嘧啶碱基,并且不包含含有5mC的碱基和含有5hmC的碱基。

9.根据权利要求1所述的方法,其中所述核酸探针是未甲基化的核酸探针。

10.根据权利要求1所述的方法,其中所述核酸探针杂交至关注的靶区域,所述关注的靶区域是与所述参考核酸序列中的CpG位点处未甲基化的胞嘧啶一致的。

11.根据权利要求1所述的方法,其中所述核酸探针包含关注的靶区域,所述关注的靶区域是与所述参考核酸序列中的CpG位点处甲基化的胞嘧啶一致的。

12.根据权利要求1所述的方法,其中所述核酸探针是化学或酶法改变的甲基化的或未甲基化的核酸探针的混合物。

13.根据权利要求1所述的方法,其中所述探查的转化的核酸的CG背景中的一个或多个胞嘧啶被转化成胸腺嘧啶,并且所述探查的转化的核酸的CH背景中的所有胞嘧啶都被转化成胸腺嘧啶。

14.根据权利要求1所述的方法,其中所述未甲基化的胞嘧啶转化成尿嘧啶包括相继的TET/APOBEC酶法转化。

15.根据权利要求1所述的方法,其中所述未甲基化的胞嘧啶转化成尿嘧啶包括TAPS。

16.一种用于确定来自受试者的生物样品的核酸分子中的靶向甲基化模式的方法,包括:

a)将包含唯一分子标识符的核酸衔接子连接到所述核酸分子,其中所述核酸分子包含未转化的核酸;

b)将所述核酸分子中未甲基化的胞嘧啶酶法转化成尿嘧啶以产生转化的核酸;

c)通过聚合酶链反应将所述转化的核酸扩增;

d)用核酸探针探查所述转化的核酸,所述核酸探针与CpG或CH基因座的预鉴定的组合互补,以富集与CpG或CH基因座的所述预鉴定的组合对应的序列;

e)以>100x的深度确定所述转化的核酸的核酸序列;和

f)将所述转化的核酸的所述核酸序列与CpG或CH基因座的所述预鉴定的组合的参考核酸序列进行比较,以确定来自所述受试者的所述生物样品的所述核酸分子的所述靶向甲基化模式。

17.根据权利要求16所述的方法,其中确定所述转化的核酸的所述核酸序列包括类似双链体纠错。

18.根据权利要求16所述的方法,其中所述核酸衔接子是耐转化的衔接子,所述耐转化的衔接子包含鸟嘌呤、胸腺嘧啶、腺嘌呤和胞嘧啶碱基,并且不包含含有5mC的碱基和含有5hmC的碱基。

19.根据权利要求16所述的方法,其中CpG或CH基因座的所述预鉴定的组合包含与转录因子起始位点相关联的基因座。

20.根据权利要求16所述的方法,其中所述靶向甲基化模式包括半甲基化的CpG基因座。

21.一种用于确定来自受试者的无细胞DNA(cfDNA)样品的甲基化图谱的方法,包括:

a)将包含唯一分子标识符的核酸衔接子连接到所述cfDNA,其中所述cfDNA包含未转化的核酸;

b)将核酸分子中未甲基化的胞嘧啶酶法转化成尿嘧啶以产生转化的核酸;

c)通过聚合酶链反应将所述转化的核酸扩增;

d)用核酸探针探查所述转化的核酸,所述核酸探针与CpG或CH基因座的预鉴定的组合互补,以富集与CpG或CH基因座的所述预鉴定的组合对应的序列;

e)以>100x的深度确定所述转化的核酸的核酸序列;和

f)将所述转化的核酸的所述核酸序列与CpG或CH基因座的所述预鉴定的组合的参考核酸序列进行比较,以确定来自所述受试者的所述无细胞DNA(cfDNA)样品的所述甲基化图谱。

22.根据权利要求21所述的方法,其中所述核酸衔接子是耐转化的衔接子,所述耐转化的衔接子包含鸟嘌呤、胸腺嘧啶、腺嘌呤和胞嘧啶碱基,并且不包含含有5mC的碱基和含有5hmC的碱基。

23.根据权利要求21所述的方法,还包括:鉴定所述cfDNA样品的来源组织,鉴定所述cfDNA样品中的体细胞变体,推断所述cfDNA样品中的核小体定位,鉴定所述cfDNA样品中差别甲基化的区域,或鉴定所述cfDNA样品中的单元型块。

24.一种对生物样品的核酸分子进行甲基化测序的方法,所述方法包括:

a)从所述核酸分子的cfDNA片段制备甲基化测序文库,包括:

i)将双链体衔接子连接到所述cfDNA片段;

ii)将双链体唯一分子标识符连接到所述cfDNA片段;和

iii)利用破坏性最小的转化方法将所述cfDNA片段中未甲基化的胞嘧啶转化成尿嘧啶,由此从所述核酸分子的所述cfDNA制备所述甲基化测序文库;

b)使所述甲基化测序文库富集与CpG或CH基因座对应的序列,由此产生富集的甲基化测序文库;

c)利用单末端或成对末端读取以>100x的深度对所述富集的甲基化测序文库测序,由此产生单末端或成对末端读取的测序片段;

d)对于所述成对末端读取的每个测序片段,校正落在所述成对末端读取的重叠区域内的测序误差;

e)将测序片段折叠成链读取族,以校正由PCR和测序产生的误差;和

f)将所述链读取族折叠成双链体读取族,以鉴定所述核酸分子中对称CpG基因座的推断的甲基化状态中的甲基化偏差。

25.根据权利要求24所述的方法,其中所述破坏性最小的转化包括酶法转化、TAPS或CAPS。

26.一种产生分类器的方法,包括:

a)将包含唯一分子标识符的核酸衔接子与从健康受试者获得的生物样品和来自患有癌症的受试者的生物样品的核酸分子连接;

b)利用破坏性最小的转化方法将所述核酸分子中未甲基化的胞嘧啶转化成尿嘧啶,由此产生转化的核酸;

c)用聚合酶链反应将所述转化的核酸扩增,由此产生扩增的转化的核酸;

d)用核酸探针探查所述扩增的转化的核酸,所述核酸探针与CpG或CH基因座的预鉴定的组合互补以富集与所述组合对应的序列,由此产生探查的转化的核酸;

e)以>100x的深度确定所述探查的转化的核酸的核酸序列;

f)将所述探查的转化的核酸的所述核酸序列与CpG或CH基因座的所述预鉴定的组合的参考核酸序列进行比较,以获得代表来自所述健康受试者和来自所述患有癌症的受试者的甲基化图谱的输入特征的测量值集;和

g)训练机器学习模型以产生区分所述健康受试者与所述患有癌症的受试者的所述分类器。

27.根据权利要求26所述的方法,其中CpG或CH基因座的所述预鉴定的组合包含与转录起始位点相关联的基因座。

28.根据权利要求26所述的方法,还包括确定半甲基化的CpG或CH基因座。

29.根据权利要求26所述的方法,还包括鉴定所述核酸分子的来源组织。

30.根据权利要求26所述的方法,还包括鉴定所述核酸分子的基因组位置和片段长度。

31.根据权利要求26所述的方法,其中所述输入特征选自:CpG的碱基甲基化%、CHG的碱基甲基化%、CHH的碱基甲基化%、在区域中具有甲基化CpG的不同计数或比率的观察片段的计数或比率、转化效率、低甲基化块、CPG的甲基化水平、CHH的甲基化水平、CHG的甲基化水平、片段长度、片段中点、chrM的甲基化水平、LINE1的甲基化水平、ALU的甲基化水平、二核苷酸覆盖度、覆盖的均匀度、整体平均CpG覆盖度,以及在CpG岛、CGI架和CGI岸处的平均覆盖度。

32.一种区分健康个体群与患有癌症的个体的分类器,包括:代表来自健康受试者和患有所述癌症的受试者的甲基化测序数据的甲基化图谱的测量值集,

其中所述测量值用于产生与所述甲基化图谱的性质对应的特征集,其中将所述特征集输入到机器学习或统计模型,其中所述模型提供特征向量,所述特征向量用作区分所述健康个体群与患有所述癌症的个体的分类器。

33.一种用于检测受试者群中的癌症的方法,包括:

a)通过使用靶向的破坏性最小的转化甲基测序来测定来自受试者的生物样品的核酸以获得所述核酸的甲基化图谱;

b)通过将所述甲基化图谱输入到经训练的算法而将所述生物样品分类,所述经训练的算法将来自健康受试者和患有所述癌症的受试者的样品分类;和

c)将报告输出在计算机屏幕上,若所述经训练的算法以特定的置信水平将所述生物样品分类为所述癌症阴性,则所述报告将所述生物样品鉴定为所述癌症阴性。

34.根据权利要求33所述的方法,其中所述癌症是结肠直肠癌。

35.一种用于基于甲基化状态而将个体分类的系统,包含:

a)包含分类器的计算机可读介质产品,

其中所述分类器包括:代表来自健康受试者和患有癌症的受试者的甲基化测序数据的甲基化图谱的测量值集,其中所述测量值用于产生与来自健康受试者和患有所述癌症的受试者的所述甲基化图谱的性质对应的特征集,其中将所述特征输入到机器学习或统计模型,其中所述模型提供特征向量,所述特征向量用作区分健康个体群与患有所述癌症的个体的分类器;和

b)一个或多个处理器,用于执行所述计算机可读介质产品上存储的指令。

36.根据权利要求35所述的系统,其中所述系统包含分类电路,所述分类电路被配置为机器学习分类器,所述机器学习分类器选自:线性判别分析(LDA)分类器、二次判别分析(QDA)分类器、支持向量机(SVM)分类器、随机森林(RF)分类器、线性核支持向量机分类器、一阶多项式核支持向量机分类器、二阶多项式核支持向量机分类器、岭回归分类器、弹性网络算法分类器、序列最小优化算法分类器、朴素贝叶斯算法分类器,以及非负矩阵分解(NMF)预测算法分类器。

37.根据权利要求35所述的系统,其中所述系统包含用于执行上述方法中的任一种方法的设备。

38.根据权利要求35所述的系统,其中所述系统包含一个或多个处理器,所述处理器配置成执行上述方法中的任一种方法。

39.根据权利要求35所述的系统,其中所述系统包含模块,所述模块分别地执行上述方法中的任一种方法的步骤。

40.一种用于监测先前接受疾病治疗的受试者中的微小残留病状态的方法,包括:

a)确定在基线甲基化状态下从所述受试者获得的生物样品的基线甲基化图谱;

b)确定在所述基线甲基化状态之后的一个或多个预定时间点时从所述受试者获得的生物样品的测试甲基化图谱;和

c)确定与所述基线甲基化图谱相比所述测试甲基化图谱的变化,其中所述变化表明所述受试者的所述微小残留病状态的变化。

41.根据权利要求40所述的方法，其中所述微小残留病状态选自：对治疗的反应、肿瘤负荷、术后残留肿瘤、复发、二级筛查、初级筛查和癌症进展。

42.根据权利要求40所述的方法，其中所述疾病是结肠直肠癌。

43.一种用于检测肿瘤的试剂盒，包含用于实施上述方法的试剂，以及针对检测肿瘤信号的说明。

44.根据权利要求43所述的试剂盒，其中所述试剂选自引物集、PCR反应组分、测序试剂、破坏性最小的转化试剂和文库制备试剂。

用于甲基化核酸的高深度测序的方法和系统

交叉引用

[0001] 本申请要求2019年5月31日提交的美国临时申请62/855,795号的权益,该临时申请通过援引而整体地并入本文中。

援引并入

[0002] 在包括实施例的本说明书中提及的所有出版物、专利和专利申请通过援引而整体地并入本文中,如同每一篇出版物、专利或专利申请被明确且单独地表明通过援引并入。在矛盾的情况下,以包括本文中的任何定义的本申请为准。

背景技术

[0003] 由于DNA的稳定性以及DNA在正常分化和疾病例如癌症中的作用,DNA甲基化可以体现肿瘤特征和表型状态,并且因此有较高的潜力用于个性化药物。异常的DNA甲基化模式发生在癌症的发病机制早期,并且因此可以促进早期癌症检测。实际上,DNA甲基化异常是癌症的标志之一并且与癌症(从肿瘤启动至癌症进展和转移)的所有方面相关联。这些性质启发了将DNA甲基化模式用于癌症诊断的多种新方法。具体而言,无细胞DNA(cfDNA)是循环中存在的片段化DNA,并且片段化模式作为生物信号是有用的并且提供信息。相比之下,基因组DNA在体外被人工片段化用于文库制备,因此基因组DNA的片段化模式对于诊断方法而言并非如此重要。

[0004] DNA甲基化是对DNA的共价修饰并且是稳定的遗传性标志,其可以在抑制基因表达和调节染色质结构方面发挥重要作用。在人类中,DNA甲基化主要发生在CpG二核苷酸中的胞嘧啶残基处。不同于其他二核苷酸,CpG不均匀地分布在整个基因组并且可以集中在称为CpG岛的短CpG富集DNA区。一般而言,在基因组中大多数CpG位点被约70-75%甲基化。然而,甲基化模式随着细胞类型而不同,反映了它们在调节细胞类型特异性基因表达中的作用。以此方式,细胞的甲基化组可以将该细胞的终末分化状态程序化为例如神经元、肌肉细胞、免疫细胞等。

[0005] 此外,组织中的各种细胞亚型可以表现出不同的甲基化模式。在癌细胞中,CpG甲基化可能失调,并且甲基化模式的异常是肿瘤发生中存在的一些最早期事件。在给定的癌症类型中的甲基化图谱非常类似于癌症起源组织的甲基化图谱。因此,在cfDNA片段上异常的甲基化标志可以用于区分正常细胞与癌细胞,并且确定组织类型起源。一般而言,癌细胞中的整体CpG甲基化水平下降,但是在特定的基因座处,特定CpG位点的平均甲基化水平(或甲基化%)可以在癌细胞中和对应的正常细胞中有所不同。在正常细胞与病变细胞之间对甲基化的CpG(DMC;单个位点)或区域(DMR;局部区域中多于一个位点)进行区别性地图谱分析允许鉴定疾病的生物标志物。这种方法引导了SEPT9基因甲基化检测(Epi proColon)的开发,这是首个FDA批准的对结肠直肠癌(CRC)基于血液的诊断。

[0006] 亚硫酸氢盐转化或亚硫酸氢盐测序已变成DNA甲基化分析的广泛使用的方法。亚硫酸氢盐测序是将DNA甲基化定位到单个碱基的方便且有效的方法。遗憾地,亚硫酸氢盐转化对cfDNA而言是严苛且破坏性的过程,其导致样品DNA的>90%降解。构建亚硫酸氢盐测序

文库的两种主要方法是:(1)在文库构建之前对DNA进行亚硫酸氢盐转化,其需要建立单链DNA文库;和(2)在双链衔接子连接之后对DNA进行亚硫酸氢盐转化。这两种情况中的任一种都涉及DNA的严重降解,这可能是有问题的,特别是对于cfDNA而言,因为cfDNA以非常低浓度存在于血浆中并且是液体活检应用中的限制性资源。在ssDNA文库中,一些降解的cfDNA可以被保留在文库中,但是降解的片段上的端点信息丢失。这种文库限制了使用cfDNA端点或片段长度信息来研究DNA甲基化的能力。在dsDNA文库中,由亚硫酸氢盐切割的cfDNA插入片段从文库丢失,但是对于留存的cfDNA插入片段,端点信息得以保留。这需要过大的血液采集体积才能实现对基因组的高深度唯一性覆盖,或者局限于仅仅以低深度唯一性覆盖进行分析。

[0007] 下一代DNA测序的出现在临床医学和基础研究方面提供了进展。然而,虽然此技术具有在单次实验中产生DNA序列的数千亿核苷酸的能力,大约1%的出错率导致数亿测序错误。这些错误在一些应用中可能是容许的,但是对于遗传学上异质的混合物例如肿瘤或混合的微生物群体的"深度测序"而言变得极有问题。

[0008] 利用现有的方法,分析cfDNA的变体和cfDNA中的甲基化状态需要两种不同的测序测定和两种不同的cfDNA库。就血浆/cfDNA输入和相关成本而言这可能是成本高昂得难以承受的。此外,DNA被亚硫酸氢盐破坏可能(相对于酶法转化)降低变体判定(variant-calling)方法的灵敏度,该方法可以处理亚硫酸氢盐转化的DNA测序数据。因此,需要对cfDNA的甲基化进行分析的改进的方法以保持样品核酸的完整性并且能够实现在全基因组或靶向水平上甲基化状态分析的改进的准确度。

发明内容

[0009] 本文中提供的方法和系统通过改进核酸甲基化测序的质量和准确度及其检测疾病的用途来解决基于亚硫酸氢盐的甲基化测序的局限性。关于甲基化状态的更准确且完整的信息允许产生更高质量的特征以供用于产生机器学习模型和分类器。

[0010] 在第一方面中,提供了一种用于对核酸样品进行甲基化测序的方法,所述方法包括:

a)将包含唯一分子标识符的核酸衔接子连接到所述核酸分子,其中所述核酸分子包含未转化的核酸;

b)利用破坏性最小的转化方法将所述核酸分子中未甲基化的胞嘧啶转化成尿嘧啶,由此产生转化的核酸;

c)通过聚合酶链反应将所述转化的核酸扩增,由此产生扩增的转化的核酸;

d)用核酸探针探查所述扩增的转化的核酸,所述核酸探针与CpG或CH基因座(loci)的预鉴定的组合(pre-identified panel)互补以富集与所述组合对应的序列,由此产生探查的转化的核酸;

e)以>100x的深度确定所述探查的转化的核酸的核酸序列;和

f)将所述探查的转化的核酸的所述核酸序列与所述CpG或CH基因座的预鉴定的组合的参考核酸序列比较以确定生物样品的所述核酸分子的甲基化图谱。

[0011] 在一个实施方案中,所述核酸分子是血浆cfDNA。

[0012] 在一个实施方案中,所述破坏性最小的转化方法包括酶法转化、TAPS或CAPS。

[0013] 在一个实施方案中,所述唯一分子标识符的长度为4bp至6bp并且具有5'胸苷突出端。

[0014] 在一个实施方案中,所述核酸衔接子还包含唯一双重索引(UDI)序列。在一个实施方案中,所述UDI序列的长度为4bp、5bp、6bp、7bp、8bp、9bp、10bp、11bp或12bp。

[0015] 在一个实施方案中,扩增所述转化的核酸包括使用引物,所述引物包含唯一双重索引(UDI)序列。在一个实施方案中,所述UDI序列的长度为4bp、5bp、6bp、7bp、8bp、9bp、10bp、11bp或12bp。

[0016] 在一个实施方案中,所述核酸衔接子是耐转化的衔接子,所述耐转化的衔接子包含鸟嘌呤、胸腺嘧啶、腺嘌呤和胞嘧啶碱基,并且不包含含有5mC的碱基和含有5hmC的碱基。

[0017] 在一个实施方案中,所述核酸探针是未甲基化的核酸探针。

[0018] 在一个实施方案中,所述核酸探针杂交至关注的靶区域,所述靶区域是与所述参考核酸序列中的CpG位点处未甲基化的胞嘧啶一致的。

[0019] 在一个实施方案中,所述核酸探针包含关注的靶区域,所述靶区域是与所述参考核酸序列中的CpG位点处甲基化的胞嘧啶一致的。

[0020] 在一个实施方案中,所述核酸探针是化学或酶法改变的甲基化的或未甲基化的核酸探针的混合物。

[0021] 在一个实施方案中,所述探查的转化的核酸的CG背景中的一个或多个胞嘧啶被转化成胸腺嘧啶,并且所述探查的转化的核酸的CH背景中的所有胞嘧啶都被转化成胸腺嘧啶。

[0022] 在一个实施方案中,所述未甲基化的胞嘧啶转化成尿嘧啶包括相继的TET/APOBEC酶法转化。

[0023] 在一个实施方案中,所述未甲基化的胞嘧啶转化成尿嘧啶包括TAPS。

[0024] 在第二方面中,提供了一种用于确定来自受试者的生物样品的核酸分子中的靶向甲基化模式的方法,所述方法包括:

a)将包含唯一分子标识符的核酸衔接子连接到cfDNA,其中所述cfDNA包含未转化的核酸;

b)将核酸分子中未甲基化的胞嘧啶酶法转化成尿嘧啶以产生转化的核酸;

c)通过聚合酶链反应将所述转化的核酸扩增;

d)用核酸探针探查所述转化的核酸,所述核酸探针与CpG或CH基因座的预鉴定的组合互补以富集与CpG或CH基因座的所述预鉴定的组合对应的序列;

e)以>100x的深度确定所述转化的核酸的核酸序列;和

f)将所述转化的核酸的所述核酸序列与CpG或CH基因座的所述预鉴定的组合的参考核酸序列比较以确定来自所述受试者的无细胞DNA(cfDNA)样品的甲基化图谱。

[0025] 在一个实施方案中,确定所述转化的核酸的核酸序列包括类似双链体纠错(duplex-like error correction)。

[0026] 在一个实施方案中,所述核酸衔接子是耐转化的衔接子,所述耐转化的衔接子包含鸟嘌呤、胸腺嘧啶、腺嘌呤和胞嘧啶碱基,并且不包含含有5mC的碱基和含有5hmC的碱基。

[0027] 在一个实施方案中,CpG或CH基因座的所述预鉴定的组合包含与转录因子起始位点相关联的基因座。

[0028] 在一个实施方案中,所述靶向甲基化模式包括半甲基化的CpG基因座。

[0029] 在第三个方面中,提供了一种用于确定来自受试者的无细胞DNA(cfDNA)样品的甲基化图谱的方法,所述方法包括:

a)将包含唯一分子标识符的核酸衔接子连接到所述cfDNA,其中所述cfDNA包含未转化的核酸;

b)将核酸分子中未甲基化的胞嘧啶酶法转化成尿嘧啶以产生转化的核酸;

c)通过聚合酶链反应将所述转化的核酸扩增;

d)用核酸探针探查所述转化的核酸,所述核酸探针与CpG或CH基因座的预鉴定的组合互补以富集与CpG或CH基因座的所述预鉴定的组合对应的序列;

e)以>100x的深度确定所述转化的核酸的核酸序列;和

f)将所述转化的核酸的所述核酸序列与CpG或CH基因座的所述预鉴定的组合的参考核酸序列比较以确定来自所述受试者的所述无细胞DNA(cfDNA)样品的甲基化图谱。

[0030] 在一个实施方案中,所述核酸衔接子是耐转化的衔接子,所述耐转化的衔接子包含鸟嘌呤、胸腺嘧啶、腺嘌呤和胞嘧啶碱基,并且不包含含有5mC的碱基和含有5hmC的碱基。

[0031] 在一个实施方案中,所述唯一分子标识符的长度为4bp至6bp并且具有5'胸苷突出端。

[0032] 在一个实施方案中,所述核酸衔接子还包含唯一双重索引(UDI)序列。在一个实施方案中,所述UDI序列的长度为4bp、5bp、6bp、7bp、8bp、9bp、10bp、11bp或12bp。

[0033] 在一个实施方案中,所述方法还包括:鉴定所述cfDNA样品的来源组织,鉴定所述cfDNA样品中的体细胞变体,推断所述cfDNA样品中的核小体定位,鉴定所述cfDNA样品中差别甲基化的区域,或鉴定所述cfDNA样品中的单元型块。

[0034] 本文中还提供了一种通过双链测序(duplex sequencing)进行甲基化测序的方法。双链测序是基于标签的纠错方法,其可以改进测序准确度,例如,甲基化测序准确度。在此方法中,将衔接子连接到核酸模板上并且利用PCR进行扩增。在一个实施方案中,所述衔接子包含引物序列和随机12bp索引。深度测序提供来自每个唯一分子标签的共有序列信息。基于分子标签和测序引物,双链体序列可以被比对以确定DNA的真实序列。双链测序的优点包括非常低的误差率,以及对PCR扩增误差的检测和消除。在双链测序中,在添加衔接子之后,也不需要附加的文库制备步骤。

[0035] 在一个实施方案中,一种对生物样品的核酸分子进行甲基化测序的方法,所述方法包括:

a从所述核酸分子的cfDNA片段制备甲基化测序文库,包括:

i)将双链体衔接子连接到所述cfDNA片段;

ii)将双链体唯一分子标识符连接到所述cfDNA片段;和

iii)利用破坏性最小的转化方法将所述cfDNA片段中未甲基化的胞嘧啶转化成尿嘧啶,由此从所述核酸分子的所述cfDNA制备所述甲基化测序文库;

b)使所述甲基化测序文库富集与CpG或CH基因座对应的序列,由此产生富集的甲基化测序文库;

c)利用单末端或成对末端读取以>100x的深度对所述富集的甲基化测序文库测序,由此产生单末端或成对末端读取的测序片段;

d)对于所述成对末端读取的每个测序片段,校正落在所述成对末端读取的重叠区域内的测序误差;

e)将测序片段折叠成链读取族(stranded read family)以校正由PCR和测序产生的误差;和

f)将所述链读取族折叠成双链体读取族(duplex read family)以鉴定所述核酸分子中对称CpG基因座的推断的甲基化状态中的甲基化偏差。

[0036] 在一个实施方案中,所述破坏性最小的转化包括酶法转化、TAPS或CAPS。

[0037] 在第四个方面中,提供了一种用于产生分类器的方法,所述方法包括:

a)将包含唯一分子标识符的核酸衔接子与从健康受试者获得的生物样品和来自患有癌症的受试者的生物样品的核酸分子连接;

b)利用破坏性最小的转化方法将所述核酸分子中未甲基化的胞嘧啶转化成尿嘧啶,由此产生转化的核酸;

c)用聚合酶链反应将所述转化的核酸扩增,由此产生扩增的转化的核酸;

d)用核酸探针探查所述扩增的转化的核酸,所述核酸探针与CpG或CH基因座的预鉴定的组合互补以富集与所述组合对应的序列,由此产生探查的转化的核酸;

e)以>100x的深度确定所述探查的转化的核酸的核酸序列;

f)将所述探查的转化的核酸的所述核酸序列与CpG或CH基因座的所述预鉴定的组合的参考核酸序列比较以获得代表来自所述健康受试者和来自所述患有癌症的受试者的甲基化图谱的输入特征的测量值集;和

g)训练机器学习模型以产生区分所述健康受试者与所述患有癌症的受试者的所述分类器。

[0038] 在一个实施方案中,CpG或CH基因座的所述预鉴定的组合包含与转录起始位点相关联的基因座。

[0039] 在一个实施方案中,所述方法还包括确定半甲基化的CpG或CH基因座。

[0040] 在一个实施方案中,所述方法还包括鉴定所述核酸分子的来源组织。

[0041] 在一个实施方案中,所述方法还包括鉴定所述核酸分子的基因组位置和片段长度。

[0042] 在一个实施方案中,所述唯一分子标识符的长度为4bp至6bp并且具有5'胸苷突出端。

[0043] 在一个实施方案中,所述核酸衔接子还包含唯一双重索引(UDI)序列。在一个实施方案中,所述UDI序列的长度为4bp、5bp、6bp、7bp、8bp、9bp、10bp、11bp或12bp。

[0044] 在一个实施方案中,扩增所述转化的核酸包括使用引物,所述引物包含唯一双重索引(UDI)序列。在一个实施方案中,所述UDI序列的长度为4bp、5bp、6bp、7bp、8bp、9bp、10bp、11bp或12bp。

[0045] 在一个实施方案中,所述输入特征选自:CpG的碱基甲基化%、CHG的碱基甲基化%、CHH的碱基甲基化%、在区域中具有甲基化CpG的不同计数或比率的观察片段的计数或比率、转化效率(例如CHH的100平均甲基化%)、低甲基化块、CPG的甲基化水平、CHH的甲基化水平、CHG的甲基化水平、片段长度、片段中点、chrM的甲基化水平、LINE1的甲基化水平、ALU的甲基化水平、二核苷酸覆盖度(例如二核苷酸的标准化覆盖度)、覆盖的均匀度(例

如在1x和10x平均基因组覆盖度下的唯一CpG位点(例如对于S4运行))、整体平均CpG覆盖度(例如深度),以及在CpG岛、CGI架和CGI岸处的平均覆盖度。

[0046] 在第五个方面中,提供了区分健康个体群与患有癌症的个体的分类器,所述分类器包括:代表来自健康受试者和患有癌症的受试者的甲基化测序数据的甲基化图谱的测量值集,

其中所述测量值用于产生与所述甲基化图谱的性质对应的特征集,其中将所述特征集输入到机器学习或统计模型,其中所述模型提供特征向量,所述特征向量用作区分所述健康个体群与患有所述癌症的个体的分类器。

[0047] 在第六个方面中,提供了一种用于检测受试者群中的癌症的方法,所述方法包括:

a)通过使用靶向的破坏性最小的转化甲基测序来测定来自受试者的生物样品的核酸以获得所述核酸的甲基化图谱;

b)通过将所述甲基化图谱输入到经训练的算法而将所述生物样品分类,所述经训练的算法将来自健康受试者和患有所述癌症的受试者的样品分类;和

c)将报告输出在计算机屏幕上,若所述经训练的算法以特定的置信水平将所述生物样品分类为所述癌症阴性,则所述报告将所述生物样品鉴定为所述癌症阴性。

[0048] 在一个示例中,所述癌症是结肠直肠癌。

[0049] 在第七个方面,本公开内容提供了一种用于基于甲基化状态而将个体分类的系统,所述系统包含:

a)包含分类器的计算机可读介质产品,

其中所述分类器包括:代表来自健康受试者和患有癌症的受试者的甲基化测序数据的甲基化图谱的测量值集,其中所述测量值用于产生与来自健康受试者和患有所述癌症的受试者的所述甲基化图谱的性质对应的特征集,其中将所述特征输入到机器学习或统计模型,其中所述模型提供特征向量,所述特征向量用作区分健康个体群与患有所述癌症的个体的分类器;和

b)一个或多个处理器,用于执行所述计算机可读介质产品上存储的指令。

[0050] 在一个示例中,所述系统包含分类电路,所述分类电路被配置为机器学习分类器,所述机器学习分类器选自:线性判别分析(LDA)分类器、二次判别分析(QDA)分类器、支持向量机(SVM)分类器、随机森林(RF)分类器、线性核支持向量机分类器、一阶多项式核支持向量机分类器、二阶多项式核支持向量机分类器、岭回归分类器、弹性网络算法分类器、序列最小优化算法分类器、朴素贝叶斯算法分类器,以及非负矩阵分解(NMF)预测算法分类器。

[0051] 在一个实施方案中,所述系统包含用于执行上述方法中的任一种方法的设备。

[0052] 在一个实施方案中,所述系统包含一个或多个处理器,所述处理器配置成执行上述方法中的任一种方法。

[0053] 在一个实施方案中,所述系统包含模块,所述模块分别地执行上述方法中的任一种方法的步骤。

[0054] 在另一方面中,本公开内容提供了一种用于监测先前接受疾病治疗的受试者中的微小残留病状态的方法,所述方法包括:将本文中所述的甲基化图谱确定为基线甲基化状态并且重复分析以确定在一个或多个预定时间点时的甲基化图谱,其中相对于基线的变化表明所述受试者中在基线的所述微小残留病状态的变化。

[0055] 在另一方面中，本公开内容提供了一种用于监测先前接受疾病治疗的受试者中的微小残留病状态的方法，所述方法包括：

a)确定在基线甲基化状态下从所述受试者获得的生物样品的基线甲基化图谱；

b)确定在所述基线甲基化状态之后的一个或多个预定时间点时从所述受试者获得的生物样品的测试甲基化图谱；和

c)确定与所述基线甲基化图谱相比所述测试甲基化图谱的变化，其中所述变化表明所述受试者的所述微小残留病状态的变化。

[0056] 在一些实施方案中，所述疾病是癌症。在一些实施方案中，所述疾病是结肠直肠癌。

[0057] 在另一方面中，本公开内容提供了一种用于监测先前接受结肠直肠癌治疗的受试者中的微小残留病状态的方法，所述方法包括：检测来自所述受试者的生物样品中的甲基化片段，其中所述生物样品中的所述甲基化片段表明所述受试者中对于所述结肠直肠癌在基线的所述微小残留病状态的变化。

[0058] 在一些实施方案中，所述微小残留病状态选自：对治疗的反应、肿瘤负荷、术后残留肿瘤、复发、二级筛查、初级筛查和癌症进展。

[0059] 在另一方面中，提供了一种确定受试者对治疗的反应的方法。

[0060] 在另一方面中，提供了一种监测受试者中的肿瘤负荷的方法。

[0061] 在另一方面中，提供了一种检测受试者中的术后残留肿瘤的方法。

[0062] 在另一方面中，提供了一种检测受试者中的复发的方法。

[0063] 在另一方面中，提供了一种用作对受试者的二级筛查的方法。

[0064] 在另一方面中，提供了一种用作对受试者的初级筛查的方法。

[0065] 在另一方面中，提供了一种监测受试者中的癌症进展的方法。

[0066] 在另一方面中，本公开内容提供了一种用于检测肿瘤的试剂盒，所述试剂盒包含用于实施上述方法的试剂，以及针对检测肿瘤信号例如甲基化标志的说明。试剂可以包括：例如，引物集、PCR反应组分、测序试剂、破坏性最小的转化试剂和文库制备试剂。

附图说明

[0067] 图1提供了流程图，其显示常规的甲基亚硫酸氢盐转化和降解与本文中所述的保留片段长度信息的改良方法的比较。

[0068] 图2提供了示意图，其显示用于本文所述的方法中的交错衔接子。

[0069] 图3提供了示意图，其显示本文中所述的类似双链体纠错方法。

[0070] 图4的子图A提供了示意图，显示耐转化的衔接子的示例。子图B提供了对应的测序引物，其匹配与相容性PCR引物完全碱基配对的转化的衔接子序列。

[0071] 图5示出了计算机系统，其被编程或以其他方式配置为实施本文中提供的方法。

[0072] 图6提供了图，其显示示例性耐转化的衔接子/引物系统的测序文库收率。在转化的情况下，耐转化的衔接子与含有5mC的衔接子相比表现出更高更大的测序收率(sequencing yield)。

具体实施方式

[0073] 本文中提供了方法，该方法能够实现改进的文库制备和甲基化区域测序，以便对cfDNA进行甲基化图谱分析。该方法通过改进甲基化数据的覆盖度、覆盖的均匀性、分辨率和准确度而解决了对生物样品中核酸的常规甲基化测序和图谱分析的局限性，从而支持实际应用。从本文提供的方法获得的测序数据可用于利用甲基化图谱数据对个体群进行分类或分层的实际应用。对个体群的这种分类或分层可以包括：鉴定患有疾病的个体、对疾病进展分级，或对疾病的特定治疗作出反应。

I.定义

[0074] 除非上下文另有明确规定，本文中使用的单数术语例如“一种”、“一个”和“该”包括单数和复数的所指物。

[0075] 术语“血浆无细胞DNA”、“循环游离DNA”、“无细胞DNA”或“cfDNA”可以表示在血液的非细胞部分中循环的DNA分子。血液中的循环核酸由坏死或凋亡的细胞产生，并且在疾病例如癌症中观察到因凋亡所致的大幅升高水平的核酸。在癌症中，循环DNA带有该疾病的标志性病征，包括癌基因突变和微卫星改变。这些循环DNA可以被称为循环肿瘤DNA(ctDNA)。血浆中病毒基因组序列、DNA或RNA是疾病的潜在生物标志物。

[0076] 在一些实施方案中，血液的无细胞部分优选地是血清或血浆。本文中使用的生物样品的术语“无细胞部分”表示生物样品的基本上无细胞的部分。本文中使用的术语“基本上无细胞”可以表示来自生物样品的制备物包含少于约20,000个细胞/ml、少于约2,000个细胞/ml、少于约200个细胞/ml，或少于约20个细胞/ml。基因组DNA(gDNA)表示从白细胞释放的污染血液无细胞部分的非片段化DNA。为了减少污染样品的gDNA，可以实施高度控制的样品加工工作流程，并且可以对样本进行筛选以防gDNA的存在。

[0077] 本文中使用的状态或结果的术语“诊断(diagnose)”或“诊断(diagnosis)”包括预测或诊断状态或结果、确定对状态或结果的易感性、监测患者的治疗、诊断患者的治疗反应、状态或结果、进展和对特定治疗的反应的预后。

[0078] 本文中使用的术语“定位”表示核苷酸在核酸分子中的鉴定的链中的位置。

[0079] 本文中使用的术语“核酸”表示可以是单链(ss)或双链(ds)的DNA、RNA、DNA/RNA嵌合体或杂合体。核酸可以是基因组的，或衍生自真核或原核细胞的基因组，或是合成的、克隆的，扩增的或反转录的。在本方法和组合物的某些实施方案中，根据上下文的要求，核酸优选地表示基因组DNA。

[0080] 除非另有陈述，否则本文中使用的术语“修饰的胞嘧啶”表示5-甲基胞嘧啶(5mC)、5-羟甲基胞嘧啶(5hmC)、甲酰基修饰的胞嘧啶、羧基修饰的胞嘧啶、5-羧基胞嘧啶(5caC)，或由任何其他化学基团修饰的胞嘧啶。

[0081] 本文中使用的术语“甲基胞嘧啶双加氧酶”、“双加氧酶”或“加氧酶”表示将5mC转化成5hmC的酶。甲基胞嘧啶双加氧酶的非限制性示例包括TET1、TET2、TET3和Naeglaria TET。-TET2是将全部5mC中的至少90％、至少92％、至少94％、至少96％、至少98％或至少99％氧化的甲基胞嘧啶双加氧酶的示例。

[0082] 本文中使用的术语“耐转化的衔接子”或“耐转化的引物”分别表示用作衔接子或引物的核酸分子。在甲基化测序的转化反应期间，耐转化的衔接子或耐转化的引物仅仅掺入未修饰的碱基以允许完全碱基转化，而非掺入修饰的核苷酸碱基以防止碱基转化。在衔

接子/引物DNA序列中“未修饰的碱基”表示常规的鸟嘌呤、腺嘌呤、胞嘧啶和胸腺嘧啶。

[0083] 本文中使用的术语“胞苷脱氨酶”表示将胞嘧啶(C)脱氨而形成尿嘧啶(U)的酶。胞苷脱氨酶的非限制性示例包括胞苷脱氨酶的APOBEC家族，例如APOBEC3A。在任一实施方案中，本文中所述的胞苷脱氨酶可以具有与GenBank登记号AKE33285.1(其是人APOBEC3A的序列)的氨基酸序列至少90％一致(例如至少95％一致)的氨基酸序列。在一些实施方案中，本文中所述的胞苷脱氨酶以至少95％、98％或99％(优选至少99％)的效率将未修饰的胞嘧啶转化成尿嘧啶。

[0084] 本文中使用的术语“葡萄糖基转移酶”或“GT”表示催化β-D-葡萄糖基或α-D-葡萄糖基残基从UDP-葡萄糖转移至5hmC残基而形成5ghmC的酶。相对于将C或5mC转化成U，APOBEC可以以低比率将5hmC转化成U。GT的示例是T4-βGT(βGT)。在一个示例中，GT可以与双加氧酶同时使用。此组合确保阻止5hmC的脱氨化，从而少于5％、少于3％或少于1％的5hmC被脱氨酶转化为U。在另一个示例中，GT可以与双加氧酶一起用于含有DNA的同一反应混合物中，从而双加氧酶将5mC转化成5hmC和5caC，并且GT将任何残留5hmC转化成5ghmC以确保仅仅胞嘧啶被脱氨。

[0085] 本文中使用的核酸样品的“部分”和核酸样品的“等分”旨在表示相同意思并且可以可互换地使用。

[0086] 本文中使用的术语“比较”表示相对于彼此分析两个或更多个序列。在一些情况中，比较可以通过将两个或更多个序列互相比对来进行，从而将对应定位的核苷酸互相比对。

[0087] 本文中使用的术语“参考序列”表示被分析的片段的序列。参考序列可以从公共数据库获得，或者可以作为实验的部分被分别地测序。在一些情况中，参考序列可以是假设的，使得参考序列可以被计算地脱氨(即将C改变成U或T等)以允许进行序列比较。

[0088] 本文中使用的术语“G”、“A”、“T”、“U”、“C”、“5mC”、“5fC”、“c5aC”、“5hmC”和“5ghmC”表示分别包含鸟嘌呤(G)、腺嘌呤(A)、胸腺嘧啶(T)、尿嘧啶(U)、胞嘧啶(C)、5-甲基胞嘧啶、5-甲酰基胞嘧啶、5-羧基胞嘧啶(5caC)、5-羟甲基胞嘧啶和5-葡萄糖基羟甲基胞嘧啶的核苷酸。为了清楚，C、5fC、5caC、5mC和5ghmC各自是不同的部分。

[0089] 术语“微小残留病”或“MRD”表示在癌症治疗之后体内少量的癌细胞。可以执行MRD测试以确定癌症治疗是否有效并且指导进一步的治疗方案。各种度量可以可以用于评估MRD，包括但不限于：对治疗的反应、肿瘤负荷、术后残留肿瘤、复发、二级筛查、初级筛查和癌症进展。

[0090] 术语“下一代测序”或“NGS”一般应用于对大小少于1kb的基因组片段的文库测序。

[0001] 本文中使用的术语“健康”表示没有疾病的受试者或从其得到的样品。虽然健康是动态状态，但是该术语可以表示受试者的缺少提及的疾病状态(例如癌症)的病理学状态。在一个示例中，当涉及将患有癌症的受试者分类的甲基化图谱时，术语“健康”表示没有癌症例如CRC的个体。虽然在该受试者中可能存在其他疾病或健康状态，但是术语“健康”可以表示出于在患有疾病状态的受试者和没有疾病状态的受试者及从其得到的样品之间比较或分类的目的，缺少陈述的疾病。

[0091] 本文中使用的术语“阈值”一般表示为了在两群受试者之间进行判别、区别或区分而选择的值。在一些实施方案中，阈值在疾病(例如恶性)状态与非疾病(例如健康)状态之

间对甲基化状态进行判别。在一些实施方案中，阈值在疾病的阶段(例如1期、2期、3期或4期)之间进行判别。阈值可以根据讨论中的疾病进行设定，并且可以基于早期(例如训练集)的分析，或者根据具有已知特性(例如健康、疾病或疾病的阶段)的输入集通过计算进行确定。阈值还可以根据在特定位点甲基化的预测值针对基因区域进行设定。阈值可以对于每个甲基化位点有所不同，并且来自多个位点的数据可以在最终分析中合并。

[0092] 除非另有定义，本文中使用的所有技术和科学术语具有与本公开内容所属领域的普通技术人员通常理解的相同的含义。本文中描述了一些示例性方法和材料，但与本文中所述的那些类似或等同的任何方法和材料也都可以用于实践或测试本教导。

[0093] 对任何出版物的引用是针对其在本申请日之前的公开内容，并且不应被解释为承认本权利要求鉴于现有发明而没有资格早于这种出版物之前。此外，提供的出版物的日期可能不同于实际出版日期，实际出版日期可以独立地确认。

[0094] 本领域技术人员在阅读本公开内容后显而易见的是，本文中所述和所示的每个单独的实施方案具有分立的组分和特征，所述组分和特征可以容易地与其他若干实施方案中的任一实施方案的特征分离或组合，而不脱离本教导的范围或精神。任一描述的方法可以按照所述事件的顺序或按照逻辑上可能的任何其他顺序实施。

[0095] 本文中提及的所有专利和出版物，包括这些专利和出版物内公开的所有序列，均通过援引而明确地并入。

II.靶向甲基化测序

[0096] 在靶向甲基化测序方法中，对生物样品例如cfDNA中的靶向区域进行分析以确定靶基因序列的甲基化状态。在一些实施方案中，靶区域包含，或在严格条件下杂交至，关注的靶区域的邻接核苷酸，例如关注的靶区域的至少约16个邻接核苷酸。在不同的示例中，靶向测序可以利用杂交捕获和扩增子测序方法来完成。

A.杂交捕获

[0097] 本文中提供的杂交方法可以用于核酸杂交的各种形式中，例如溶液中杂交和固态载体上杂交(例如在膜、微阵列和细胞/组织载玻片上的RNA杂交、DNA杂交和原位杂交)。具体而言，该方法适合于溶液中杂交捕获，用于对在靶向下一代测序中使用的某些类型的基因组DNA序列(例如外显子)进行靶标富集。对于杂交捕获方法，对无细胞核酸样品进行文库制备。本文中使用的“文库制备”包括对无细胞DNA执行的末端修复、A-加尾、衔接子连接或任何其他制备以允许对DNA的后续测序。在某些示例中，制备的无细胞核酸文库序列包含连接到无细胞核酸样品分子上的衔接子、序列标签和索引条形码。各种商业上可获得的试剂盒可用于促进文库制备以供用于下一代测序方法。下一代测序文库构建可以包括利用一系列协调的酶反应来制备核酸靶标，以产生用于高通量测序的特定大小的DNA片段的随机集合。各种文库制备技术的进展和开发已经将下一代测序的应用拓展至例如转录物组学和表观遗传学领域。

[0098] 测序技术的改进已经导致文库制备的改变和改进。本文中使用的下一代测序文库制备试剂盒包括由Agilent、Bioo Scientific、Kapa Biosystems、New England Biolabs、Illumina,Life Technologies、Pacific Biosciences和Roche等公司开发的那些。

[0099] 在用于靶向捕获基因组合的各种示例中，各种文库制备试剂盒可以选自Nextera Flex(Illumina)、IonAmpliseq(Thermo Fisher Scientific)和Genexus(Thermo Fisher

Scientific)、Agilent ClearSeq(Illumina)、Agilent SureSelect Capture(Illumina)、Archer FusionPlex(Illumina)、BiooScientific NEXTflex(Illumina)、IDT xGen(Illumina)、Illumina TruSight(Illumina)、Nimblegene SeqCap(Illumina)和Qiagen GeneRead(Illumina)。

[0100] 在一些实施方案中,杂交捕获方法利用特异性探针对制备的文库序列执行。本文中使用的术语“特异性探针”可以表示对已知的甲基化位点具有特异性的探针。在一些实施方案中,特异性探针是基于将人类基因组用作参考序列并且将已知具有甲基化位点的特定基因组区域用作靶序列进行设计。具体地,已知具有甲基化位点的基因组区域可以包括以下中的至少一个:启动子区域、CpG岛区域、CGI岸区域和印记基因区域。因此,当通过使用一些实施方案的特异性探针来执行杂交捕获时,可以有效地捕获样品基因组中与靶序列互补的序列,例如样品基因组中已知具有甲基化位点的区域(其在本文中也被称为“特定的基因组区域”)。

[0101] 根据示例,本文中所述的甲基化的区域用于设计特异性探针。在一些实施方案中,特异性探针是利用商业上可获得的方法例如eArray系统进行设计。探针的长度可以足以以足够的特异性杂交至关注的甲基化的区域。在各种示例中,探针是10-聚体、11-聚体、12-聚体、13-聚体、14-聚体、15-聚体、16-聚体、17-聚体、18-聚体、19-聚体或20-聚体。

[0102] 甲基化分析的靶向区域可以通过利用数据库资源(例如基因本体论)而被筛查出。根据互补碱基配对原理,单链捕获探针可以互补地与单链靶序列组合,从而成功地捕获靶区域。在一些实施方案中,设计的探针可以被设计为固态捕获芯片(其中探针被固定在固态载体上)或被设计为液态捕获芯片(其中探针游离在液体中)。然而,由于限制性因素,例如探针长度、探针密度和高成本,固态捕获芯片极少被使用,而液态捕获芯片更经常被使用。

[0103] 在一些实施方案中,与正常序列(其中A、T、C和G碱基的平均含量各为25%)比较,GC-富集的序列(其中GC碱基的平均含量高于60%)由于C和G碱基的分子结构而可能导致捕获效率下降。对于关键研究区域,例如,CGI区域(CpG岛),可能需要增加量的探针来获得足够且准确的CGI数据。

B.基于扩增子的测序

[0104] 可以将转化的DNA的片段扩增。在一些实施方案中,用引物进行扩增,所述引物设计成与其中具有至少一个甲基化位点的甲基化转化的靶序列退火。甲基化测序转化导致未甲基化的胞嘧啶被转化成尿嘧啶,而5-甲基胞嘧啶不受影响。“转化的靶序列”可以表示这样的序列,其中已知为甲基化位点的胞嘧啶被固定为“C”(胞嘧啶),而已知为未甲基化的胞嘧啶被固定为“U”(尿嘧啶;为了引物设计的目的其可以被处理为“T”(胸腺嘧啶))。

[0105] 在各种示例中,DNA的来源是从全血、血浆、血清获得的无细胞DNA,或者从细胞或组织提取的基因组DNA。在一些实施方案中,扩增的片段的大小是约100至200个碱基对长度。在一些实施方案中,DNA来源是从细胞来源(例如组织、活检、细胞系)提取,并且扩增的片段是约100至350个碱基对长度。在一些实施方案中,扩增的片段包含至少一个20碱基对的序列,其包含至少一个、至少两个、至少三个或多于三个CpG二核苷酸。扩增可以根据本公开内容利用引物寡核苷酸集进行,并且可以使用热稳定性聚合酶。若干DNA区段的扩增可以在同一反应容器中同时进行。在一些实施方案中,两个或更多个片段被同时扩增。例如,扩增可以利用聚合酶链反应(PCR)进行。

[0106] 设计成靶向这种序列的引物可以表现出一定程度的向转化的甲基化序列的偏倚。在一些实施方案中,PCR引物被设计为对于靶向甲基化测序应用而言具有甲基化特异性。甲基化特异性引物可以在一些应用中允许更大的灵敏度。例如,引物可以被设计为包含辨别性核苷酸(对亚硫酸氢盐转化之后的甲基化序列具有特异性),该核苷酸被定位成在例如PCR应用中达到最佳辨别。辨别性核苷酸可以定位在3'末端或次末端位置。

[0107] 在一些实施方案中,引物被设计为使75至350bp长度的DNA片段扩增,该长度是循环DNA的一般大小范围。针对靶标大小优化引物设计可以增加本文中所述的方法的灵敏度。引物可以被设计成扩增约50至200、约75至150或约100或125bp长度的区域。

[0108] 在一个实施方案中,扩增步骤包括使用引物,所述引物包含唯一双重索引(UDI)序列。

[0109] 在一个实施方案中,所述UDI序列的长度为4bp、5bp、6bp、7bp、8bp、9bp、10bp、11bp或12bp。

[0110] 在一些实施方案中,在核酸序列内预选CpG位置的甲基化状态是利用甲基化特异性PCR(MSP)引物寡核苷酸通过基于扩增子的方法进行检测。甲基化特异性引物用于亚硫酸氢盐处理的DNA的扩增允许在甲基化核酸和未甲基化核酸之间进行区分。MSP引物对包含至少一个与转化的CpG二核苷酸杂交的引物。因此,所述引物的序列包含至少一个CpG、TpG或CpA二核苷酸。对非甲基化DNA具有特异性的MSP引物含有在CpG中C位置的3'位置处的"T"。因此,这些引物的碱基序列可以包括长度为至少18个核苷酸的序列,该序列杂交至预处理的核酸序列和与其互补的序列,并且碱基序列具有至少一个CpG、TpG或CpA二核苷酸。在所述方法的一些实施方案中,MSP引物具有2至5个CpG、TpG或CpA二核苷酸。在一些实施方案中,二核苷酸定位在引物的3'半部内,例如,对于具有18个碱基长度的引物,特定的二核苷酸定位在从序列的3'末端起的前9个碱基内。除了CpG、TpG或CpA二核苷酸之外,引物还可以包含若干甲基转化的碱基(例如胞嘧啶转化成胸腺嘧啶,或者在杂交链上鸟嘌呤转化成腺苷)。在一些实施方案中,引物被设计成具有不多于2个胞嘧啶和/或鸟嘌呤碱基。

[0111] 在一些实施方案中,利用多个引物对将每个区域扩增在节段中。在一些实施方案中,这些节段是非重叠的。节段可以紧密相邻或分隔开(例如分隔开多达10、20、30、40或50个bp)。因为靶区域(包括CpG岛、CpG岸和/或CpG架)通常长于75至150个bp,所以此示例允许给定靶区域的更多个(或所有)位点的甲基化状态得到评估。

[0112] 可以利用适合的工具,例如引物3、引物3Plus、引物-BLAST等设计引物用于靶区域。如所述,亚硫酸氢盐转化导致胞嘧啶转化成尿嘧啶并且5'-甲基-胞嘧啶转化成胸腺嘧啶。因此,引物定位或靶向可以利用亚硫酸氢盐转化的甲基化序列,这取决于所需的甲基化特异性的程度。

III.用于酶法甲基化测序的文库制备

[0113] 在第一个方面中,提供了制备测序文库的方法。本文中所述的方法提供了文库,所述文库对于下一代非甲基化和甲基化测序应用是可接受的,由此为来自单个样品的两种应用提供测序数据。所得的原始测序数据可以用于甲基化状态分析,以及更常规的cfDNA分析,例如拷贝数改变、种系变体检测、体细胞变体检测、核小体定位、转录因子图谱分析、染色质免疫沉淀等。

A.用于靶向测序应用的衔接子连接

[0114] 在一个方面中，本方法保留用于甲基化图谱分析的核酸序列的完整性和信息。在一个示例中，在酶法转化之前结合dsDNA衔接子连接能保留片段端点信息，同时为靶标富集提供最高可能性的文库复杂性(或直接用于全基因组测序)，由此提供更大的灵敏度来检测稀有事件，例如甲基化的ctDNA。在转化之前衔接子连接的优点和比较示于图1中。

[0115] 在一个示例中，将核酸衔接子与生物样品中核酸片段群的5'末端和3'末端连接以产生测序文库。在一个示例中，将核酸衔接子的集合与样品中的核酸片段连接，其中衔接子的集合包括4bp、5bp和6bp唯一分子标识符(UMI)序列的相等部分，其跟随着在最后位置(即3'末端)处的不变胸苷(T)以实现T/A突出端连接。因此，UMI定位成邻近于文库插入核酸。在测序期间，UMI还作为在5'末端的读取的一部分进行测序(可替代地，UMI在测序读取水平下与文库插入片段一致)。不变T在3个位置上交错以便在测序位置处保持碱基多样性。相比之下，利用具有不变胸苷的单一长度UMI导致在与不变胸苷对应的位置处的低复杂性测序，从而降低测序质量。每个UMI的前4个bp共同包含4-bp核心UMI序列集，其具有大于或等于2的编辑距离并且是核苷酸和颜色平衡的。尽管UMI序列的长度可变，但是使用单一长度的核心UMI促进生物信息学工具(其构建用于单一长度UMI以便UMI提取和去除重复)的使用。因此，4-bp核心序列用作识别序列，其向生物信息学工具提供信息以修整5、6或7个碱基(包括恒定T)，由此保持精确的cfDNA端点信息。交错衔接子的示意图示于图2中。UMI的使用允许在测序之后读取去除重复，单链纠错和双链体重建，由此允许使用读取的反向互补序列来增强误差校正，也称为双链纠错。在另一示例中，唯一双重索引(UDI)是附加的序列，其可以在文库制备期间添加到含有UMI的衔接子，以提供样品条码化和测序后样品的多路分解(de-multiplexing)。在各种示例中，UDI序列的长度为4bp、5bp、6bp、7bp、8bp或12bp。

[0116] 在各种实施方案中，核酸衔接子可以包含4bp至6bp长度的具有5'胸苷突出端的UMI。UMI被设计成非唯一(即抽取自特定的约束序列集)。

[0117] 在一个实施方案中，一些UMI包含一个或多个甲基胞嘧啶碱基。酶法甲基化转化反应(包括TET氧化和APOBEC脱氨)的效率可以基于UMI中与特定的约束的设计UMI序列集不匹配的部分通过UMI错配率进行评估。UMI错配率可以用作嵌入式质量控制度量来评估测序文库质量。此外，如果在生物信息学管路中需要完美的UMI匹配，则UMI错配率可以用作过滤器以去除由于不完全转化而可能质量较低的个别读取。

[0118] 在各种实施方案中，UMI错配率少于6％、少于5％、少于4％、少于3％或少于2％。

[0119] 在另一实施方案中，UMI包含一个或多个含有修饰的胞嘧啶，其可以用于监测酶活性。这些修饰的碱基的非限制性示例包括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。

B.用于DNA甲基化测序应用的酶法转化

[0120] Tet-辅助的吡啶硼烷测序(TAPS)是用于将核酸中胞嘧啶转化成尿嘧啶的破坏性最小的转化甲基化测序方法。这种无亚硫酸氢盐的方法允许DNA的最小降解，并且因此保留核酸分子的长度同时达到与亚硫酸氢钠测序类似的转化率。TAPS可以对胞嘧啶和鸟嘌呤碱基对产生更高的测序质量评分，并且可以提供对各种基因组特征例如CpG岛的更均匀的覆盖度。

[0121] 在TAPS中，十-十一易位(Tet1)酶将5mC和5hmC二者氧化成5caC。吡啶硼烷将5caC还原成二氢尿嘧啶，一种尿嘧啶衍生物，其在PCR之后被转化成胸腺嘧啶。TAPS可以以两种

其他方式进行:TAPSβ和化学辅助的吡啶硼烷测序(CAPS)。在TAPSβ中,β-葡萄糖基转移酶用于以葡萄糖标记5hmC来保护5hmC免于氧化和还原反应并且允许对5mC的特异性检测。在CAPS中,高钌酸钾充当Tet1的化学替代物并且特异性地氧化5hmC,因此允许直接检测。

[0122] 在一个示例中,将未修饰的C向U的酶法转化以及与文库插入片段一致的交错UMI衔接子联合用于对甲基化文库进行靶向测序。对于低深度测序应用,与亚硫酸氢盐转化测序相比,由于样品cfDNA没有同等程度地降解,所以此联合可以允许减少血浆的体积输入或cfDNA的质量输入。

[0123] 对于高深度测序应用,与亚硫酸氢盐转化测序相比,由于cfDNA没有同等程度地降解,所以可以从血浆或cfDNA的类似输入获得较高深度测序。

[0124] 在一个示例中,存在于衔接子核酸中的胞嘧啶被修饰而含有5-甲基或5-羟甲基以防止在衔接子中的C-向-T转化。

[0125] 此方法的一个优点在于,与进行亚硫酸氢盐转化而后ssDNA衔接子连接的方法相比,在转化之前的衔接子连接保持片段端点和长度信息。在连接衔接子之前核酸的显著降解可能导致片段端点信息和长度信息丢失。

[0126] 未修饰的C酶法转化成U对样品核酸片段而言严苛性较低,并且与亚硫酸氢盐转化方法相比可以导致更全面和均匀的覆盖。DNA的亚硫酸氢盐降解是不均一的,使得一些序列相对于其他序列被优先降解,包括CG二核苷酸,而其正是在甲基化测序中被查询的位点。因此,利用相同数量的唯一读取,与亚硫酸氢盐转化方法相比,酶方法在靶标富集应用中提供对CpG位点的更高覆盖率,以及捕获的读取的更高均匀性。此外,非亚硫酸氢盐方法(例如酶法和类似于TAPS的化学转化)提供对生物信号的提高分辨率,特别是在核酸序列中区别5mC和5hmC甲基化的能力提高。此信息和额外的分辨率可以在计算方法和其他方法中提供信息。

[0127] 在一些示例中,使DNA或条形码化的DNA经历酶反应(其将所述DNA或条形码化的DNA的胞嘧啶核碱基转化成尿嘧啶核碱基)包括“进行酶法转化”。

[0128] 在各种示例中,葡萄糖基化和氧化反应克服了观察到的5hmC和5mC被脱氨酶的内在脱氨。脱氨酶将5mC和未修饰的C转化成U,但是没有转化5ghmC和5caC。脱氨酶的非限制性示例包括APOBEC(载脂蛋白B mRNA编辑酶,类似于催化性多肽)。本文中所述的实施方案利用在胞嘧啶的葡萄糖基化、氧化和脱氨中基本上序列无偏性的酶。此外,这些实施方案在葡萄糖基化、氧化和脱氨反应期间对DNA的基本上不造成非特异性损伤。

[0129] 在一些实施方案中,葡萄糖基转移酶(GT),例如β-葡萄糖基转移酶(βGT),用于将葡萄糖共价连接到5hmC以保护该修饰的碱基免于脱氨。其他酶或化学反应可以用于修饰5hmC以达到相同效果。

[0130] 一般而言,在一个方面中,本文中提供的方法包括:(a)在反应混合物中用双加氧酶例如TET2和βGT处理核酸样品的等分(部分)以产生反应产物,其中基本上全部修饰的胞嘧啶(C)被氧化或者在5hmC的情况中被葡萄糖基化;和(b)将该反应产物用胞苷脱氨酶处理,从而将基本上全部未修饰的C转化成U。在全部这些示例和实施方案中使用的术语“修饰的”胞嘧啶表示5mC、5hmC、5ghmC、5fC和5caC中的一个或多个,其中5mC、5hmC和5fC的完全氧化产生5caC。βGT仅与5hmC反应。但是,在发生葡萄糖基化之前,一些5hmC可以被双加氧酶转化为5fC,而后转化为5caC。在双加氧酶的存在下,5mC大部分被完全氧化为5caC,但是可能

产生一些残留的5hmC。但是，残留的5hmC可以由βGT进行葡萄糖基化而防止5hmC的低脱氨率，否则可能降低甲基化测序的准确度。

[0131] 因此，所述方法通过在脱氨之前用双加氧酶处理核酸而在很大程度上区别未修饰的和修饰的胞嘧啶。但是，在基因组DNA中5mC的天然存在量可能大幅超过5hmC的量，其继而可能超过5fC和5caC的天然存在量。因此，修饰的胞嘧啶的天然存在量通常被视为5mC的天然存在量的近似值。

[0132] 在一个示例中，所述方法可以适用于执行5hmC测序。5hmC测序方法可以进一步包括：在不存在双加氧酶下用βGT处理核酸样品的等分，而后用胞苷脱氨酶处理以产生反应产物，其中等分中的基本上全部5hmC被葡萄糖基化，并且基本上全部未修饰的C和5mC被转化成U。在PCR扩增之后，U被转化成T，因此，当测序时胞嘧啶和5mC变得不可区别。可以将所得的反应产物测序并且与参考序列比较以将5hmC与C和5mC区分开。对这些部分的区分允许将这些修饰的核苷酸映射到参考序列，例如，来自数据库的参考序列或独立地确定的参考序列。

[0133] 在一些实施方案中，双加氧酶与βGT外加脱氨酶反应产物或其扩增产物可以被测序以确定哪些C被甲基化(其中可能包括较小部分的5hmC)以及哪些C未被修饰。在一些实施方案中，在没有双加氧酶的情况下，βGT外加脱氨酶反应产物或其扩增产物可以被测序以确定哪些C被羟甲基化以及哪些C未被羟甲基化。在一些实施方案中，在没有双加氧酶的情况下，βGT没有双加氧酶外加脱氨酶反应产物或其扩增产物可以被测序以确定哪些C被羟甲基化以及哪些C未被修饰。参考DNA可以通过对所得的反应产物测序而产生，所述反应产物是通过使核酸样品不与双加氧酶、βGT和脱氨酶中的任一种反应而产生的。可替代地，参考序列是已知的参考序列，例如来自序列的数据库。

[0134] 在一个实施方案中，双加氧酶与βGT外加脱氨酶反应产物的序列可以与参考序列比较。任选地，这也可以与βGT(无双加氧酶)外加脱氨酶反应产物的序列比较以确定核酸样品中的哪些胞嘧啶被甲基修饰，哪些被羟甲基修饰。

[0135] 在一个方面中，提供了一种用于对核酸样品进行靶向甲基化测序的方法，包括：

a)将包含唯一分子标识符的核酸衔接子连接到cfDNA，其中所述cfDNA包含未转化的核酸；

b)将核酸分子中未甲基化的胞嘧啶酶法转化成尿嘧啶以产生转化的核酸；

c)通过聚合酶链反应将所述转化的核酸扩增；

d)用核酸探针探查所述转化的核酸，所述核酸探针与CpG或CH基因座的预鉴定的组合互补以富集与CpG或CH基因座的所述预鉴定的组合对应的序列；

e)以>100x的深度确定所述转化的核酸的核酸序列；和

f)将所述转化的核酸的核酸序列与CpG或CH基因座的所述预鉴定的组合的参考核酸序列比较以确定来自受试者的无细胞DNA(cfDNA)样品的甲基化图谱。

[0136] 如果测试的转化的核酸序列是T，该T对应于在特定CpG基因座处的参考C，则在原始的测试核酸片段中该C未被甲基化。相比之下，如果测试的转化的核酸序列和参考序列在特定CpG基因座处都是C，则在原始的测试核酸片段中该C被甲基化。

[0137] 在一个示例中，以约50-500x、约25-1000x、约50-500x、约250-750x、约500-200x、约750-1500x，或约100-2000x的深度对转化的核酸分子的核酸序列进行测序。在一些实施

方案中,以>100x或>500x的深度对核酸序列进行测序。

[0138] 在一个示例中,以约500x、约1000x、约2000x、约3000x、约4000x、约5000x、约6000x、约7000x、约8000x、约9000x、约10000x或大于5000x的深度对转化的核酸分子的核酸序列进行测序。

[0139] 在一个示例中,以约300x唯一、约400x唯一、约500x唯一、约600x唯一、约700x唯一、约800x唯一、约900x唯一或约1000x唯一或大于500x唯一的深度对转化的核酸分子的核酸序列进行测序。

C.靶标富集测序应用

[0140] 还提供了用于在测序期间在靶标捕获应用中富集关注的甲基化区域的方法。应用靶标富集捕获组合与DNA甲基化文库的潜在问题是低比率的中靶读取/高比率的脱靶DNA片段捕获。对于组合中的每个区域,探针可以被设计为靶向从甲基化CpG衍生的DNA或靶向从未甲基化的CpG衍生的DNA。在任一探针类型中,沿着该区域的每个CpG位点被视为未甲基化的或甲基化的,根据适合于探针类型而定。探针可以与经过亚硫酸氢盐/酶法转化和PCR扩增后的文库分子杂交。然后,只有被探针捕获的文库分子被测序。由于只有少部分基因组被测序,所以此方法具有降低测序成本的优点。在一个示例中,约0.1%的基因组被测序。在一个示例中,约0.3%的基因组被测序。在一个示例中,约0.5%的基因组被测序。在一个示例中,约0.7%的基因组被测序。在一个示例中,约1%的基因组被测序。在其他示例中,约2%、约3%、约4%、约5%、约6%、约7%、约8%、约9%或约10%的基因组被测序。

[0141] 对亚硫酸氢盐和酶法转化的文库的靶标捕获富集方法可能发生显著的脱靶捕获率。脱靶捕获率部分是由于在与从甲基化CpG衍生的DNA杂交的两种探针类型中,不在CpG位点的全部胞嘧啶的C至T转化。探针中胞嘧啶含量降低导致序列复杂性减小,因此,探针与靶文库分子杂交的特异性减小。

[0142] 本文中使用的术语“甲基化的探针”和“未甲基化的探针”表示在转化后的核酸序列中分别用于与甲基化的和未甲基化的CpG杂交的探针。探针可以被设计为识别转化后的核酸序列。在转化后的甲基化CpG探针中,在转化后,C仍然为C。在转化后的未甲基化CpG探针中,在转化后,C被转化成T。在转化后的甲基化和未甲基化探针二者中,在转化后,非-CpG二核苷酸中的全部C被转化成T。

[0143] 甲基化的探针保留一些胞嘧啶(即在CpG位点中的胞嘧啶)。相比之下,在未甲基化的探针中全部胞嘧啶都被转化成胸腺嘧啶。未甲基化的探针与甲基化的探针相比复杂性较低,并且可能优先地贡献出脱靶捕获率。在一个示例中,将与从甲基化CpG衍生的DNA杂交的探针用于靶标富集方法。在一个示例中,将具有与靶标(该靶标与从甲基化CpG衍生的DNA杂交)基本上互补的序列的探针用于靶标富集方法。

[0144] 用于靶标富集的与从甲基化CpG衍生的DNA杂交的探针可以被选择来达成不同的方面。靶标捕获杂交反应发生在单一温度下。但是,与从甲基化CpG衍生的DNA杂交的探针的最佳解链温度(Tm)平均高于未设计成与从甲基化CpG衍生的DNA杂交的探针的Tm。

[0145] 胞嘧啶碱基配对涉及3个氢键,而胸腺嘧啶碱基配对涉及2个氢键。由于氢键减少,所以在探针中胞嘧啶转化为胸腺嘧啶降低该探针的Tm。由于甲基化的探针保留一些胞嘧啶而未甲基化的探针没有保留胞嘧啶,所以相对于匹配的未甲基化的探针,甲基化的探针将具有升高的Tm。随着区域中CpG位点的数量增大,甲基化的和未甲基化的探针之间解链温度

的差异也增大。与具有较低解链温度的探针相比，具有较高解链温度的探针可以更有效地与靶DNA片段杂交。杂交温度通常被选择为相对高以促进中靶捕获。但是，由于保留一些胞嘧啶而解链温度较高，因此在典型的杂交温度下，甲基化的探针将比未甲基化的探针更有效地杂交。较高的解链温度可以导致偏向靶标捕获杂交方法测得的CpG甲基化水平％高于对预捕获文库测序测得的水平。

[0146] 在一个示例中，将仅单一探针类型，即甲基化的或未甲基化的探针，用于杂交反应中以分别富集高甲基化的或低甲基化的文库分子。使用单一类型的甲基化或未甲基化的探针可以避免探针类型之间解链温度差异的问题。使用单一探针类型还可以促进更有效地捕获(或富集)相同的DNA片段类型。在一个示例中，相对于低甲基化的ROI，仅使用甲基化探针提供对高甲基化的ROI的优选结合。在另一示例中，仅使用未甲基化探针提供对未甲基化的ROI的富集。

[0147] 使用仅单一探针类型还允许使用更高的杂交温度来降低脱靶捕获而不影响甲基化ROI捕获与未甲基化ROI捕获的相对平衡。因此，探针组合可以基于期望设计用于富集高甲基化的或低甲基化的DNA片段。在一个示例中，在期望对高甲基化的DNA片段和低甲基化的DNA片段都进行定量的情况中，将两种平行但分开的杂交反应用于两种甲基化状态。

D.甲基化分析

[0148] 在各种示例中，当酶法甲基化测序完成时，测定用来分析生物样品中核酸的甲基化状态。在一个示例中，全基因组酶法甲基测序(“WG EM测序”)通过表征基因组中几乎每个胞苷核苷酸的DNA甲基化而提供高分辨率测序。其他靶向方法例如靶向酶法甲基测序(“TEM测序”)可以用于甲基化分析。

[0149] 在其他示例中，已常规用于亚硫酸氢盐转化的测定可以用于破坏性最小的转化方法，例如酶法转化、TAPS和CAPS。在各种示例中，用于甲基化分析的测定可以是：质谱法、甲基化-特异性PCR(MSP)、简化代表性亚硫酸氢盐测序(RRBS)、HELP测定、GLAD-PCR测定、ChIP-on-chip测定、限制性标记基因组扫描、甲基化DNA免疫沉淀(MeDIP)、亚硫酸氢盐处理的DNA的焦磷酸测序、分子断裂光(molecular break light)测定、甲基敏感性DNA印迹、高分辨率熔解分析(HRM或HRMA)、古老DNA甲基化重建，或甲基化敏感性单核苷酸引物延伸测定(msSNuPE)。

[0150] cfDNA的甲基化图谱可以通过应用序列比对方法来映射来自全基因组的甲基测序读取，或对人类参考基因组的靶向甲基测序进行鉴定。序列比对方法的非限制性示例包括：bwameth、bismark、Last、GSNAP、BSMAP、NovoAlign、Bison、宏基因组学系统发育(Metagenomic Phylogenetic)分析(例如，MetaPhlAn2)、BLAT、Burrows-Wheeler Aligner(BWA)、Bowtie、Bowtie2、Bfast、BioScope、CLC bio、Cloudburst、Eland/Eland2、GenomeMapper、GnuMap、Karma、MAQ、MOM、Mosaik、MrFAST/MrsFAST、PASS、PerM、RazerS、RMAP、SSAHA2、Segemehl、SeqMap、SHRiMP、Slider/SliderII、Srprism、Stampy、vmatch、ZOOM和SOAP/SOAP2比对工具。

E.利用基于双链体UMI-的甲基化共有判定的CpG纠错

[0151] 甲基化分析需要基于CpG背景内的‘C’在测序时是被读出为“C”’(甲基化的)还是“T”(未甲基化的)而对测序数据进行分析。然而，在测序时“T”可能因一些原因而出现在除了母DNA分子中存在未甲基化的CpG之外的一些位置。这些原因包括：测序期间引入的误差、

PCR期间引入的误差、在末端修复期间核苷酸填入、DNA损伤、CpG置换为另一二核苷酸的种系单核苷酸多态性(SNP)、CpG置换为另一二核苷酸的体细胞突变,和过度转化(即5mC被转化为T而不管是否有甲基化标志)。此外,在测序时"C"可能因一些原因而出现在除了母DNA分子中存在甲基化的CpG之外的一些位置。这些原因包括:测序期间引入的误差、PCR期间引入的误差、DNA损伤、和不完全转化(即未甲基化的C未被转化为T而不管是否缺少甲基化标志)。无法校正大多数的这些误差模式可能导致CpG甲基化状态的不准确读出,由此可能限制对要求极其准确地读出CpG甲基化状态的稀有事件(例如早期癌症中的ctDNA分子)的检测。此外,未考虑双链体信息的方法无法区分半甲基化的CpG位点与对称甲基化的CpG位点。这种信息可以用于解释甲基化信号的生物意义。例如,半甲基化直接鉴定从头甲基化事件,由此允许从头事件与维护因素之间的区分。

[0152] 双链测序方法通过解决这些普遍的误差而克服了测序准确度的局限性。例如,双链测序通过对DNA双链体的两条链中的每条链独立地加标签和测序而减少误差。因为两条链是互补的,所以真实的突变可以在两条链中在相同的位置处找到。类似地,因为CpG二核苷酸是对称的,完全甲基化的CpG基序在两条链中在相对的邻近位置处具有甲基化的胞嘧啶。相比之下,PCR或测序误差将导致仅一条链中的误差。此方法独特地利用了双链DNA中在链各处存在的冗余和额外信息,因此可以用于克服仅利用来自仅一条链的数据的方法的技术局限性。

[0153] 对于酶法甲基化测序,各个片段的APOBEC转化效率可以通过在CHH背景中的胞嘧啶被测序为胞嘧啶的数量进行评估。在100%效率的APOBEC反应中,在CHH背景中全部胞嘧啶都被转化成尿嘧啶并且被测序为胸腺嘧啶。相比之下,未有效地受到APOBEC酶作用(即未完全转化)的cfDNA片段可以在CHH背景中包含一个或多个未被转化为尿嘧啶的胞嘧啶被测序为胞嘧啶。CHH背景中未转化的胞嘧啶的数量可以用作过滤器来消除由于不完全转化而可能不可靠的噪音读取。

[0154] 作用于核酸的许多酶具有序列优选性,导致偏向于有效地受到酶作用的那些位点。实验数据可以用于鉴定各个酶的序列优选性。在各种实施方案中,然后,此数据可以用于掩蔽更可能被酶不完全转化的潜在位点。作为示例,相对于T在前的胞嘧啶,APOBEC A3A对A在前的胞嘧啶具有12倍辨别力。

[0155] 在一个示例中,提供了一种双链体甲基化共有判定的方法,该方法包括:

a)利用酶法转化从cfDNA制备甲基化测序文库,包括:

(i)将双链体衔接子连接到从生物样品获得的核酸片段;

(ii)进行靶标富集以实现对关注的先前鉴定的基因座的超高深度测序;

(iii)利用酶法转化从cfDNA制备甲基化测序文库(使得两条链都没有受损)并且在酶法转化之前连接双链体UMI(在酶法转化中涉及的变性步骤之前将双链体链标签化)

b)靶标富集以实现对关注的特定基因座的超高深度测序

c)利用单末端或成对末端读取来对富集的文库测序

d)对于成对末端读取的每个测序的片段,校正落在成对末端读取的重叠区域内的测序误差

e)将测序片段折叠成链读取族以校正由PCR和测序产生的误差;和

f)将链读取族折叠成双链体读取族以鉴定对称CpG的推断的甲基化状态的偏差。

[0156] 此方法的示意图示于图3中。

[0157] 利用双链体信息对甲基测序数据的CpG“纠错”提供了滤除噪音的优点，否则该噪音可能使得将甲基测序数据用作输入的分类器的灵敏度或特异性降低。因为在转化后核苷酸不平衡被引入序列中，在甲基测序的背景中使用交错长度的甲基化UMI的唯一UMI设计可以有助于提高测序准确度，通过帮助聚类识别(具体地利用平台例如NextSeq序列仪)而潜在地增大数据输出，并且降低对添加大量PhiX数据的依赖性(增大测序深度，由此降低相关成本)。不同于标准双链测序(其分析在碱基配对的核苷酸处变体判定的一致性)，基于CpG双链体的双链测序方法评估在链各处CpG甲基化的对称性(1-bp偏置)。在某些示例中，双链测序还可以允许区分SNP与未甲基化的CpG。酶法甲基测序方法相对于基于亚硫酸氢盐的方法的优点在于更高效率地捕获双链的序列。

F.在酶法甲基化测序中使用耐转化的衔接子进行基于双链体UMI的甲基化共有判定的CpG纠错

[0158] 在另一方面中，将耐转化的衔接子和引物用于甲基测序。用于鉴定碱基修饰的位置的测序方法，例如用于鉴定5mC的亚硫酸氢盐测序或酶法甲基化测序(EM测序)，通过化学或酶法改变每个未修饰的胞嘧啶碱基(C)来改变C的碱基配对性质而进行。例如，在EM测序过程期间，全部未修饰的C被APOBEC酶转化成尿嘧啶(U)并且随后被测序为胸腺嘧啶(T)。5mC碱基未被转化并且被测序为C。因为仅当DNA是单链时碱基才可以被转化，所以在C变为U的转化反应之前双链DNA必须被变性。

[0159] 当将双链测序与甲基化测序联合时可能产生的一个问题是PCR扩增和测序的减少。因为衔接子必须在DNA仍然是双链时(即在碱基转化之前)连接到DNA上，所以衔接子中的全部C将被转化为U，由此妨碍有效的PCR扩增和测序。

[0160] 此问题的解决方案是使用含有修饰的碱基(例如5mC、5hmC或其他C变体)的衔接子，所述修饰的碱基在脱氨反应期间不转化或转化的可能性较小。但是，含有修饰的碱基的寡核苷酸通常比仅含有标准碱基的寡核苷酸明显更昂贵。此外，此解决方案一般仅对亚硫酸氢盐甲基测序有效，其中5mC不能被转化。

[0161] 不同于亚硫酸氢盐测序，EM测序过程需要附加的酶法步骤，其需要防止5mC或5hmC被APOBEC转化为U。此步骤使用Tet酶来氧化5mC或5hmC碱基，或使用βGT来葡萄糖基化5hmC，由此保护5hmC免于转化。如果此步骤不完全有效，则衔接子中的一些5mC或5hmC将被转化为尿嘧啶，导致文库失去复杂性和测序质量下降。Tet氧化反应对反应条件非常敏感，并且可能导致可变的测序文库质量。

[0162] 为了改进EM测序双链测序的稳健性和氧化效率(并且降低目前的经济负担)，可以使用仅含有未修饰的碱基的耐转化的衔接子。未修饰的碱基表示常规碱基，即在没有修饰的情况下的鸟嘌呤、胞嘧啶、腺嘌呤和胸腺嘧啶。与利用修饰的碱基例如5mC或5hmC来限制衔接子分子的总体转化的常规方法相反，此方法允许衔接子中总体胞嘧啶转化而提供提高的效率和测序质量。耐转化的衔接子的示例示于图4子图A中。

[0163] 用这些耐转化的衔接子产生的测序文库可以用与原始衔接子序列匹配的PCR和测序引物集进行扩增和测序。在转化后，测序文库可以利用与转化的衔接子序列匹配的PCR和测序引物进行扩增和测序，图4子图B。

G.在酶法转化期间使用内部过程对照

[0164] 对于靶向酶法甲基化测序，合成的内部过程对照(IPC)可以在酶法甲基化转化期间用于监测氧化和脱氨反应。

[0165] 在各种实施方案中，IPC可以在C前后两个碱基的窗口(NNCNN)中包含所有256种可能的胞嘧啶背景。

[0166] 在各种实施方案中，IPC是通过PCR合成的双链体，其包含100％未修饰的C、100％甲基化的C或100％羟基化的C(或C的另一种修饰)。就此而言，可以监测IPC的转化或保护效率。在一些实施方案中，该转化或保护效率可以通过测序或定量的PCR进行监测。

H.半甲基化分析

[0167] 在另一示例中，在甲基测序中使用UMI允许对半甲基化的误差校正和分析/去除。可替代地，在另一示例中，链特异性甲基化测序实现对半甲基化DNA的鉴定。CpG/CpG二联体的甲基化状态通常是一致的，即完全甲基化的或完全未甲基化的。但是，甲基化不一致(即半甲基化的)的CpG二联体通常以低至中等的频率发生，除了在DNA复制期间经历短暂地转录沉默或再激活的区域。这些半甲基化的二联体提供额外的信息，该信息可以告知用于对群体分层的分类器。对半甲基化的二联体的识别提供更全面的甲基测序图谱并且提供在产生分类器期间除去或包括此信息的选项。

[0168] 酶法甲基测序方法的另一优点是更好地区分甲基化的C与未转化的C。通过酶法转化来保留片段完整性和长度允许使用双链体UMI甲基测序来提高确定核酸分子的真实甲基化状态的准确度。此方法可以解决在例如提取(DNA损伤)、文库制备(末端修复填入)、酶法转化(转化不足或过度转化)、PCR(碱基掺入误差)和测序(碱基判定误差)期间引入可能的误差。通过利用这些基于甲基化的表观基因序列差异，提高甲基化状态确定的准确度改进了特征化和用于群分层的分类器产生。在一个示例中，衔接子的方向性质用于鉴定dsDNA片段是源自顶部链还是底部链(基于读取1映射到哪个基因组链)，这示意性地示于图3中。此方法不同于依赖于索引条形码来纠错的方法。

I.鉴定体细胞变体

[0169] 在各种示例中，酶法转化的DNA用于推断基因组中C残基的甲基化状态。但是，因为DNA的酶法转化将未甲基化的C残基转化为U残基并且没有将其他化学变化引入DNA中，所以与参考或查询序列中的C或T碱基不对应的体细胞变体也可以在转化的DNA中进行鉴定。这些体细胞变体可以利用设计用于未转化的DNA的现有方法(包括纠错方法例如双链测序)进行鉴定。此外，与参考或查询序列中的C或T碱基对应的体细胞变体可以利用双链测序，基于体细胞变体应在双链体DNA分子的双链中的相同位置处被找到而甲基化相关的模式不应如此(即因为观察到C和T碱基没有相互碱基配对)的这一预期而从甲基化相关的测序模式区分出来。总体地，此差异实现在EM测序中鉴定CpG位点的甲基化状态和体细胞变体二者。

J.推断核小体定位

[0170] 与侧面的DNA相比，在跨过核小体的DNA中可以大量富集CpG位点的胞嘧啶的甲基化。因此，CpG甲基化模式还可以用于用机器学习方法推断核小体定位。EM测序数据集还可以根据用于WGS的相同方法进行分析，来产生特征以供输入机器学习方法和模型中而不管甲基化转化如何。随后，5mC模式可以用于预测核小体定位，其可以有助于推断疾病和癌症的基因表达和/或分类。在另一示例中，特征可以从甲基化状态和核小体定位信息的组合获得。

[0171] 用于甲基化分析的度量包括但不限于:M偏倚(CpG、CHG、CHH的碱基甲基化%)、转化效率(例如CHH的100平均甲基化%)、低甲基化块、甲基化水平(例如CPG、CHH、CHG、chrM、LINE1或ALU的总体平均甲基化)、二核苷酸覆盖度(二核苷酸的标准化覆盖度)、覆盖的均匀度(例如在1x和10x平均基因组覆盖度下的唯一CpG位点(例如对于S4运行))、整体平均CpG覆盖度(深度),以及在CpG岛、CGI架和CGI岸处的平均覆盖度。在一个示例中,基于双链体的CpG甲基化判定的输出用作此分析的输入。在一个示例中,片段端点和长度信息用作分析的特征输入。这些度量可以用作特征输入用于机器学习方法和模型。

[0172] 在另一方面中,本公开内容提供了一种方法,包括:(a)提供来自受试者的包含cfDNA的生物样品;(b)使cfDNA经受足以任选富集样品中的甲基化cfDNA的条件;(c)和将cfDNA的未甲基化的胞嘧啶核碱基酶法转化为尿嘧啶核碱基;(d)对cfDNA测序,由此产生序列读取;(e)计算机处理读取序列以(i)基于尿嘧啶核碱基的存在而确定cfDNA的甲基化程度;和(ii)对cfDNA的至少部分降解建模,由此产生降解参数;和(f)利用降解参数和甲基化程度来确定基因序列特征。

[0173] 在一些示例中,cfDNA的测序包括基于未转化的胞嘧啶核碱基与转化的胞嘧啶核碱基的比率而确定DNA的甲基化程度。在一些示例中,转化的胞嘧啶核碱基被检测为尿嘧啶核碱基。在一些示例中,尿嘧啶核碱基被观察为读取序列中的胸腺嘧啶核碱基。

[0174] 在一些示例中,产生降解参数包括使用贝叶斯模型。在一些示例中,贝叶斯模型是基于链偏倚性或酶法转化或过度转化。在一些示例中,对读取序列的计算机处理包括在成对HMM或Naive贝叶斯模型的框架下使用降解参数。

K.分析差别甲基化区域(DMR)

[0175] 在一个示例中,甲基化分析是差别甲基化区域(DMR)的分析。DMR用于相对于基因组的区域量化CpG甲基化。区域根据发现进行动态地分配。来自不同类别的多个样品可以被分析,并且在不同分类之间最大差别甲基化的区域可以被鉴定。可以选择差别地甲基化的区域子集并且用于分类。在区域中捕获的CpG的数量可以用于分析。在一个示例中,基于双链体的CpG甲基化判定的输出用作此分析的输入。区域可以是大小可变的。在一个示例中,执行预发现过程,使得将多个CpG位点打包在一起作为区域。在一个示例中,DMR用作输入特征用于机器学习方法和模型。

L.甲基化单元型块和甲基化单元型负荷

[0176] 在一个示例中,将单元型块测定应用于样品。甲基化单元型块的鉴定有助于异质组织样品的解卷积和从血浆DNA映射肿瘤来源组织。紧密偶联的CpG位点,称为甲基化单元型块(MHB),可以在WGBS数据中进行鉴定。称为甲基化单元型负荷(MHL)的度量用于在块水平下进行组织特异性甲基化分析。此方法提供可用于异质样品的解卷积的信息块。此方法可用于定量估计肿瘤负荷并将来源组织映射在循环cfDNA中。在一个示例中,基于双链体的CpG甲基化判定的输出用作此分析的输入。在一个示例中,单元型块用作输入特征用于机器学习方法和模型。

M.用于鉴定来源细胞类型的靶向甲基化判定分析

[0177] 在一个方面中,方法用于靶向甲基化判定来基于甲基化模式鉴定cfDNA分子的来源细胞类型。该方法提供在个体读取序列上多个邻近CpG位点的共同甲基化状态的概率模型,以便将DNA甲基化的弥漫性质探索用于信号放大。该模型针对每种细胞类型形成读取序

列的概率,而后针对全体细胞类型形成混合模型并拟合到该模型。

[0178] 传统的DNA甲基化分析集中在细胞群的各个CpG位点的甲基化率(β-值)以表明其中CpG位点被甲基化的细胞的比例。这种群体平均的测量通常灵敏度不足以捕获仅影响小部分cfDNA的异常甲基化信号。但是,基于DNA甲基化的弥漫性质,疾病特异性cfDNA读取可以计算地区别于正常cfDNA读取。

[0179] 另外,鉴于DNA甲基化的弥漫性质,多个邻近CpG位点的共同甲基化状态可以用于容易地区分癌症特异性cfDNA读取与正常cfDNA读取。在给定读取中全部CpG位点的甲基化平均值(表示α-值)提供异常甲基化的cfDNA与正常cfDNA(α肿瘤＝0％和α正常＝100％)之间的差异(0和1)。甲基化α-值用于估计在读取中全部CpG位点的共同概率是否遵循疾病的DNA甲基化标志。此方法可以灵敏地从血浆中的全部cfDNA鉴别出原始cfDNA的多种细胞类型。

[0180] 在各种示例中,比对工具用于将读取与参考基因组比对并且判定甲基化的胞嘧啶。将PCR重复物除去并且针对每个CpG位点量化甲基化的和未甲基化的胞嘧啶的数量。CpG簇的甲基化水平被计算为甲基化的胞嘧啶的数量与该簇内胞嘧啶的总数量之间的比率。此WGBS数据处理过程计算出用于鉴定甲基化标志物的正常血浆样品中CpG簇的平均甲基化水平。当血浆cfDNA样品用作测试数据时,将与标志物组合的区域比对的各个读取序列的全部CpG位点的共同甲基化状态提取出,而后输入到机器学习模型中。在此方法中,基于双链体的CpG甲基化判定用作输入特征以供甲基化状态分析和特征产生。为了改进具有高覆盖的cfDNA甲基化数据的输入数据质量,可以将覆盖<2、<3或<4个CpG位点的读取滤除。

[0181] 本文中所述的甲基化测序方法改进测序读取质量,例如,通过降低PCR误差和偏倚并且减少亚硫酸氢盐转化时出现的DNA降解。在一个示例中,甲基化测序数据用于对重叠区域建模。在一个示例中,机器学习模型可以确定鉴定的甲基化DNA区域的来源细胞类型。

[0182] 在各种示例中,模型可以分类多于两种来源细胞类型。在其他示例中,模型可以将序列分类成3、4、5、6、7、8、9、10、15、20、50、75、100或多于100种不同的细胞类型。

N.DNA羟甲基化分析

[0183] 在本发明的一个方面中,5hmC测序的实现可以是通过在衔接子连接步骤中对衔接子核酸的羟甲基化进行取代,并随后仅使用βGT将葡萄糖缀合至测试核酸文库插入片段中的5hmC残基,而不使用双加氧酶和βGT来缀合5mC和5hmC。当将所得的测序数据与参考基因组比较时,在测试序列中显示对应的C的参考中每个C位置被解释为羟甲基化的C,并且在测试序列中显示为T的参考中的每个C被解释为未修饰的C或甲基化的C。由此,用于羟甲基化分析的数据解释是与甲基化分析相同的。

[0184] 在本发明的一个方面中,甲基化和羟甲基化测序文库可以被比较以说明在单个核苷酸分辨率下每个胞嘧啶修饰(例如5m或5mC)的水平。

[0185] 在本发明的一个方面中,因为羟甲基化状态读出是与甲基化状态相同的,所以用于甲基化测序数据的所有分析方法都可以应用于羟甲基化测序数据。

IV.计算机系统和机器学习方法

A.样品特征

[0186] 如本文所使用的,当术语“特征”涉及机器学习和模式识别时,它指的可以是被观察的现象的单个可测量的属性或特性。特征通常是数字特征,但结构特征如字符串和图形

被用于句法模式识别。“特征”的概念与统计技术如线性回归中使用的解释变量的概念有关。

[0187] 在一个实施方案中，特征被输入到特征矩阵中以用于机器学习分析。

[0188] 对于多个测定，系统鉴定特征集以输入至机器学习模型。系统对每种分子类别进行测定，并根据测量值形成特征向量。系统将特征向量输入到机器学习模型中，并获得生物样品是否具有指定属性的输出分类。

[0189] 在一个实施方案中，机器学习模型输出分类器，所述分类器区分在个体的群体或群体的特征中的个体或特征的两个组或类别。在一个实施方案中，分类器是经训练的机器学习分类器。

[0190] 在一个实施方案中，对癌症组织中的信息基因座或生物标志物的特征进行测定，以形成图谱。接收者操作特性(ROC)曲线可用于绘制在区分两个群体(例如，对治疗剂有反应和无反应的个体)时特定特征(例如，本文所述的任何生物标志物和/或任何其他生物医学信息项目)的性能。通常，基于单个特征的值以升序对整个群体(例如，病例和对照)中的特征数据进行排序。

[0191] 在一些实施方案中，病症是晚期腺瘤(AA)、结肠直肠癌(colorectal cancer，CRC)、结直肠癌(colorectal carcinoma)或炎性肠病。

[0192] 术语“输入特征”或“特征”表示由模型用来预测样品的输出分类(标记)的变量，该输出分类是例如状况、序列含量(例如突变)、建议的数据收集操作或建议的治疗。变量的值可以针对样品进行确定并且用于确定分类。基因数据的输入特征的示例包括：比对的变量(涉及序列数据(例如读取序列)与基因组比对)，以及非比对的变量(例如涉及读取序列的序列含量、蛋白质或自身抗体的测量、或在基因组区域的平均甲基化水平)。

[0193] 变量的值可以针对样品进行确定并且用于确定分类。基因数据的输入特征的示例包括：比对的变量(涉及序列数据(例如读取序列)与基因组比对)，以及非比对的变量(例如涉及读取序列的序列含量、蛋白质或自身抗体的测量、或在基因组区域的平均甲基化水平)。在各种示例中，基因特征，例如，V-作图量度、FREE-C、在转录起始位点上的cfDNA测量，以及在cfDNA片段上的DNA甲基化水平，用作机器学习方法和模型的输入特征。

[0194] 在一个示例中，测序信息包括关于多种基因特征的信息，例如，但不限于：转录起始位点、转录因子结合位点、染色质打开和闭合状态、核小体定位或占位及诸如此类。

B.数据分析

[0195] 在一些实施方案中，本公开内容提供了在软件应用、计算硬件或两者中实现数据分析的系统、方法或套件。在各种实施方案中，分析应用或系统至少包括数据接收模块、数据预处理模块、数据分析模块(其可以对一种或多种类型的基因组数据进行操作)、数据解释模块或数据可视化模块。在一个实施方案中，数据接收模块可以包括将实验室硬件或仪器与处理实验室数据的计算机系统连接的计算机系统。在一个实施方案中，数据预处理模块可以包括硬件系统或计算机软件，其对数据执行操作以备用于分析。可以在预处理模块中应用于数据的操作示例包括仿射变换、去噪操作、数据清理、重新格式化或二次采样。可以被专门化以用于分析来自一种或多种基因组材料的基因组数据的数据分析模块，例如，可以获取已组装的基因组序列并进行概率和统计分析，以鉴定与疾病、病理、状态、风险、病况或表型相关的异常模式。数据解释模块可以使用例如从统计学、数学或生物学中得出的

分析方法，以支持对所鉴定的异常模式与健康状况、功能状态、预后或风险之间的关系的理解。数据可视化模块可以使用数学建模、计算机图形学或渲染的方法来创建数据的视觉表示，该视觉表示可以促进对结果的理解或解释。

[0196] 在各种实施方案中，应用机器学习方法来区分样品群体中的样品。在一个实施方案中，应用机器学习方法来区分健康样品和晚期腺瘤样品。

[0197] 在一个实施方案中，用于训练基于甲基化的预测引擎的一个或多个机器学习操作包括以下中一项或多项：广义线性模型、广义加性模型、非参数回归运算、随机森林分类器、空间回归运算、贝叶斯回归模型、时间序列分析、贝叶斯网络、高斯网络、决策树学习运算、人工神经网络、递归神经网络、强化学习运算、线性/非线性回归运算、支持向量机、聚类运算和遗传算法运算。

[0198] 在各种实施方案中，计算机处理方法选自逻辑回归、多元线性回归(MLR)、降维、偏最小二乘(PLS)回归、主成分回归、自动编码器、变分自动编码器、奇异值分解、傅立叶基数(Fourier base)、小波、判别分析、支持向量机、决策树、分类和回归树(CART)、基于树的方法、随机森林、梯度提升树、逻辑回归、矩阵分解、多维缩放(MDS)、降维方法、t-分布随机邻域嵌入(t-SNE)、多层感知器(MLP)、网络聚类、神经模糊和人工神经网络。

[0199] 在一些实施方案中，本文中公开的方法可以包括对来自一个个体或来自多个个体的样品的核酸测序数据进行计算分析。分析可以鉴定从序列数据推断的变体，从而基于概率建模、统计建模、机理建模、网络建模或统计推断来鉴定序列变体。分析方法的非限制性示例包括：主成分分析、自动编码器、奇异值分解、傅里叶基数、小波、判别分析、回归、支持向量机、基于树的方法、网络、矩阵分解和聚类。变体的非限制性示例包括种系变异或体细胞突变。在一些实施方案中，变体可以表示已知的变体。已知的变体可以在文献中科学地确认或报告。在一些实施方案中，变体可以表示与生物变化相关联的推定变体。生物变化可以是已知或未知的。在一些实施方案中，推定变体可能报告在文献中，但是尚未在生物学上被确认。

[0200] 或者，推定变体未报告在文献中，但是可以基于本文中公开的计算分析进行推断。在一些实施方案中，种系变体可以表示诱导天然或正常变异的核酸。

[0201] 天然或正常变异可以包括，例如，皮肤颜色、毛发颜色和正常重量。在一些实施方案中，体细胞突变可以表示诱导获得性或异常变异的核酸。获得性或异常变异可以包括，例如，癌症、肥胖、病况、症状、疾病和病症。在一些实施方案中，分析可以包括在种系变体之间进行辨别。种系变体可以包括，例如，个体变体和体细胞突变。在一些实施方案中，鉴定的变体可以由临床医生或其他保健专业人员用来改善医疗方法、诊断的准确度和降低成本。

[0202] 本文中还提供了改进的方法和计算系统或软件介质，其可以在通过扩增和/或测序技术引入的核酸序列误差、体细胞突变和种系变体之间进行区分。提供的方法可以包括同时对来自从患者获得的所有样品的比对测序数据的变体进行判定和评分。

[0203] 还可以使用从除了患者以外的受试者获得的样品。其他样品也可以从先前接受测序或靶向测序(即靶向再测序)分析的受试者收集。本文中公开的方法、计算系统或软件介质可以改进变异或突变(例如种系或体细胞的变异或突变，包括拷贝数变异、单核苷酸变异、插入/缺失、基因融合)的鉴定和准确度，并且通过减少假阳性和假阴性鉴定的数量而降低检测限。

C.分类器产生

[0204] 在一方面,本系统和方法提供了基于从来自cfDNA的生物样品的甲基化序列分析得到的特征信息而产生。分类器构成了预测引擎的一部分,该预测引擎可根据在生物样品(例如cfDNA)中鉴定出的甲基化序列特征来区分群体中的组。

[0205] 在一个实施方案中,通过将甲基化信息的相似部分格式化为统一格式和统一标度来标准化甲基化信息而创建分类器;将标准化的甲基化信息存储在列式数据库中;通过将一个或多个机器学习操作应用于存储的标准化的甲基化信息来训练甲基化预测引擎,该甲基化预测引擎针对特定群体映射一个或多个特征的组合;将甲基化预测引擎应用于所访问的字段信息,以鉴定与组相关的甲基化;并将个体分类到一个组中。

[0206] 特异性可以定义为在没有疾病的个体中阴性测试的概率。特异性等于测试为阴性的无疾病的人数除以无疾病个体的总数。

[0207] 在各种实施方案中,模型、分类器或预测性测试具有至少40%、至少45%、至少50%、至少55%、至少60%、至少65%、至少70%、至少75%、至少80%、至少85%、至少90%、至少95%或至少99%的特异性。

[0208] 灵敏度可以定义为在患有疾病的个体中阳性测试的概率。灵敏度等于测试为阳性的患病个体数除以患病个体的总数。

[0209] 在各种实施方案中,模型、分类器或预测性测试具有至少40%、至少45%、至少50%、至少55%、至少60%、至少65%、至少70%、至少75%、至少80%、至少85%、至少90%、至少95%或至少99%的灵敏度。

[0210] 在一个实施方案中,所述组选自健康(无症状)、炎性肠病、AA或CRC。

D.数字处理装置

[0211] 在一些实施方案中,本文所述的主题可以包括数字处理装置或其用途。在一些实施方案中,数字处理装置可以包括执行该装置的功能的一个或多个硬件中央处理单元(CPU)、图形处理单元(GPU)或张量处理单元(TPU)。在一些实施方案中,数字处理装置可以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,数字处理装置可以任选地连接到计算机网络。在一些实施方案中,数字处理装置可以任选地连接到互联网,使得其可以访问万维网。在一些实施方案中,数字处理装置可以任选地连接到云计算基础设施。在一些实施方案中,数字处理装置可以任选地连接到内联网。在一些实施方案中,数字处理装置可以任选地连接到数据存储装置。

[0212] 合适的数字处理装置的非限制性示例包括服务器计算机、台式计算机、膝上型计算机、笔记本电脑、小型笔记本电脑、上网本电脑、上网板电脑(netpad computer)、机顶计算机、手持计算机、互联网电器、移动智能手机和平板电脑。合适的平板计算机可以包括,例如,具有小册子、平板和可转换配置的那些计算机。

[0213] 在一些实施方案中,数字处理装置可以包括被配置为执行可执行指令的操作系统。例如,操作系统可以包括软件(包括程序和数据),其管理装置的硬件并提供用于执行应用程序的服务。操作系统的非限制性示例包括Ubuntu、FreeBSD、OpenBSD、NetBSD®、Linux、Apple®Mac OS X Server®、Oracle® Solaris®、Windows Server®和Novell® NetWare®。合适的个人计算机操作系统的非限制性示例包括

Microsoft® Windows®、Apple®Mac OS X®、UNIX®和类似UNIX的操作系统(如GNU/Linux®)。在一些实施方案中,操作系统可以由云计算提供,并且云计算资源可以由一个或多个服务提供商提供。

[0214] 在一些实施方案中,装置可以包括存储和/或存储器装置。该存储和/或存储器装置可以是一个或多个用于临时或永久地存储数据或程序的物理装置。在一些实施方案中,装置可以是易失性存储器,并且需要电源来维持存储的信息。在一些实施方案中,装置可以是非易失性存储器,并且在数字处理装置未通电时保留所存储的信息。在一些实施方案中,非易失性存储器可以包括闪存。在一些实施方案中,非易失性存储器可以包括动态随机存取存储器(DRAM)。在一些实施方案中,非易失性存储器可以包括铁电随机存取存储器(FRAM)。在一些实施方案中,非易失性存储器可以包括相变随机存取存储器(PRAM)。在一些实施方案中,装置可以是存储装置,包括例如,CD-ROM、DVD、闪存装置、磁盘驱动器、磁带驱动器、光盘驱动器和基于云计算的存储装置。在一些实施方案中,存储和/或存储器装置可以是如本文公开的那些装置的组合。

[0215] 在一些实施方案中,数字处理装置可以包括显示器,用于将视觉信息发送给用户。在一些实施方案中,显示器可以是阴极射线管(CRT)。在一些实施方案中,显示器可以是液晶显示器(LCD)。

在一些实施方案中,显示器可以是薄膜晶体管液晶显示器(TFT-LCD)。在一些实施方案中,显示器可以是有机发光二极管(OLED)显示器。

在一些实施方案中,OLED显示器可以是无源矩阵OLED(PMOLED)或有源矩阵OLED(AMOLED)显示器。在一些实施方案中,显示器可以是等离子显示器。在一些实施方案中,显示器可以是视频投影仪。

在一些实施方案中,显示器可以是如本文公开的那些装置的组合。

[0216] 在一些实施方案中,数字处理装置可以包括输入装置以从用户接收信息。在一些实施方案中,输入装置可以是键盘。在一些实施方案中,输入装置可以是定点装置,包括例如鼠标、轨迹球、轨迹板、操纵杆、游戏控制器或指示笔。在一些实施方案中,输入装置可以是触摸屏或多点触摸屏。在一些实施方案中,输入装置可以是用于捕获语音或其他声音输入的麦克风。在一些实施方案中,输入装置可以是摄像机,用于捕获运动或视觉输入。在一些实施方案中,输入装置可以是如本文公开的那些装置的组合。

E.非暂时性计算机可读存储介质

[0217] 在一些实施方案中,本文公开的主题可以包括一种或多种用程序编码的非暂时性计算机可读存储介质,所述程序包括可由作为任选联网的数字处理装置的一部分的操作系统执行的指令。在一些实施方案中,计算机可读存储介质可以是数字处理装置的有形部件。在一些实施方案中,计算机可读存储介质可以任选地从数字处理装置移除。在一些实施方案中,计算机可读存储介质可以包括,例如,CD-ROM、DVD、闪存装置、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。在一些实施方案中,程序和指令可以被永久、基本上永久、半永久或非暂时地编码在介质上。

F.计算机系统

[0218] 本公开内容提供了被编程为实现本公开内容的计算机系统。图5示出了计算机系统501,该计算机系统501被编程或以其他方式配置为存储、处理、鉴定或解释患者数据、生

物数据、生物序列或参考序列。计算机系统501可以处理本公开内容的患者数据、生物数据、生物序列或参考序列的各个方面。计算机系统501可以是用户的电子装置或相对于电子装置远程定位的计算机系统。电子装置可以是移动电子装置。

[0219] 计算机系统501包括中央处理单元(CPU,在本文也称为“处理器”和“计算机处理器”)505,其可以是单核或多核处理器,或者是用于并行处理的多个处理器。计算机系统501还包括存储器或存储器位置510(例如,随机存取存储器、只读存储器、闪存)、电子存储单元515(例如,硬盘)、用于与一个或多个其他系统进行通信的通信接口520(例如,网络适配器),以及外围装置525,如高速缓存、其他存储器、数据存储装置和/或电子显示适配器。存储器510、存储单元515、接口520和外围装置525通过诸如主板的通信总线(实线)与CPU 505通信。存储单元515可以是用于存储数据的数据存储单元(或数据存储库)。计算机系统501可以借助于通信接口520可操作地耦合到计算机网络(“网络”)530。网络530可以是因特网、互联网和/或外联网,或者内联网和/或与因特网通信的外联网。在一些实施方案中,网络530是电信和/或数据网络。网络530可以包括一个或多个计算机服务器,其可以实现分布式计算,如云计算。在一些实施方案中,借助于计算机系统501,网络530可以实现对等网络,该对等网络可以使耦合到计算机系统501的装置能够表现为客户端或服务器。

[0220] CPU 505可以执行一系列机器可读指令,这些机器可读指令可以体现在程序或软件中。指令可以存储在存储器位置,如存储器510中。指令可以被定向到CPU 505,其可以随后对CPU 505进行编程或以其他方式配置CPU 505以实现本公开内容的方法。CPU 505执行的操作实施方案可以包括获取、解码、执行和写回。

[0221] CPU 505可以是电路(例如集成电路)的一部分。系统501的一个或多个其他部件可以被包括在电路中。在一些实施方案中,该电路是专用集成电路(ASIC)。

[0222] 存储单元515可以存储文件,如驱动程序、库和保存的程序。存储单元515可以存储用户数据,例如,用户偏好和用户程序。在一些实施方案中,计算机系统501可以包括在计算机系统501外部,例如位于通过内联网或因特网与计算机系统501通信的远程服务器上的一个或多个附加数据存储单元。

[0223] 计算机系统501可以通过网络530与一个或多个远程计算机系统通信。例如,计算机系统501可以与用户的远程计算机系统通信。远程计算机系统的示例包括个人计算机(例如,便携式PC)、平板电脑或平板计算机(例如,Apple®iPad、Samsung®Galaxy Tab)、电话、智能电话(例如,Apple®iPhone、支持Android的装置、Blackberry®)或个人数字助理。用户可以通过网络530访问计算机系统501。

[0224] 本文所述的方法可以通过存储在计算机系统501的电子存储位置,例如,存储器510或电子存储单元515上的机器(例如,计算机处理器)可执行代码来实现。机器可执行或机器可读代码可以以软件的形式提供。在使用过程中,代码可以由处理器505执行。在一些实施方案中,可以从存储单元515检索代码并将其存储在存储器510中,以供处理器505随时访问。在一些实施方案中,可以不包括电子存储单元515,并且将机器可执行指令存储在存储器510中。

[0225] 代码可以被预编译和配置,以用于具有适合于执行代码的处理器的机器,或者可以在运行时期间被解释或编译。可以以如下编程语言提供代码:所述编程语言可以被选择

为使代码能够以预编译、解释或即时编译的方式执行。

[0226] 本文提供的系统和方法的方面，如计算机系统501，可以在编程中得到实施。技术的各个方面可以被认为“产品”或“制品”，通常呈在机器可读介质的类型上承载或以在机器可读介质的类型实施的机器(或处理器)可执行代码和/或关联数据的形式。机器可执行代码可以被存储在电子存储单元，例如存储器(例如，只读存储器、随机存取存储器、闪存)或硬盘上。“存储”类型的介质可以包括计算机、处理器等的任何或所有有形存储器，或其相关模块，例如可以在任何时候为软件编程提供非暂时性存储的各种半导体存储器、磁带驱动器、磁盘驱动器等。软件的全部或部分有时可以通过因特网或其他各种电信网络进行通信。例如，这种通信可以使得软件能够从一个计算机或处理器加载到另一个计算机或处理器，例如，从管理服务器或主机加载到应用服务器的计算机平台。因此，可以承载软件元件的另一种介质类型包括光波、电波和电磁波，例如通过有线和光学陆线网络以及经各种空中链路、在本地装置之间的物理接口之间所使用的光波、电波和电磁波。携带这种波的物理元件，例如有线或无线链路、光链路等也可以被视为承载软件的介质。如本文所用，除非限于非暂时性的、有形的“存储”介质，否则诸如计算机或机器“可读介质”的术语是指参与向处理器提供指令以供执行的任何介质。

[0227] 因此，机器可读介质，例如计算机可执行代码，可以采取多种形式，包括但不限于有形存储介质、载波介质或物理传输介质。非易失性存储介质包括，例如，光盘或磁盘，例如任何计算机中的任何存储装置等，其例如可用于实现附图所示的数据库等。易失性存储介质包括动态存储器，例如此类计算机平台的主存储器。有形的传输介质包括同轴电缆；铜线和光纤，包括构成计算机系统内总线的电线。载波传输介质可以采用电信号或电磁信号的形式，也可以采用声波或光波的形式，如在射频(RF)和红外(IR)数据通信期间产生的那些。因此，计算机可读介质的常见形式包括例如：软盘、柔性盘、硬盘、磁带、任何其他磁介质、CD-ROM、DVD或DVD-ROM、任何其他光学介质、穿孔卡纸带、带孔图案的任何其他物理存储介质、RAM、ROM、PROM和EPROM、FLASH-EPROM、任何其他存储芯片或盒，传输数据或指令的载波、传输此类载波的线缆或链路，或计算机可以从中读取编程代码和/或数据的任何其他介质。计算机可读介质的许多这些形式可以涉及将一个或多个指令的一个或多个序列载送给处理器以供执行。

[0228] 计算机系统501可以包括电子显示器135或与电子显示器135通信，该电子显示器135包括用户界面(UI)540，用于提供，例如，核酸序列、富集的核酸样品、表达谱和表达谱分析。UI的示例包括，但不限于，图形用户界面(GUI)和基于web的用户界面。

[0229] 本公开内容的方法和系统可以通过一种或多种算法来实现。算法可以在由中央处理单元505执行时，通过软件来实现。例如，该算法可以探测多个调控元件、对核酸样品进行测序、富集核酸样品、确定核酸样品的表达谱、分析核酸样品的表达谱，以及存档或传播表达谱分析的结果。

[0230] 在一些实施方案中，本文公开的主题可以包括至少一个计算机程序或该计算机程序的用途。计算机程序可以执行一系列指令，这些指令可以在数字处理装置的CPU、GPU或TPU中执行，并被编写为执行指定任务。计算机可读指令可以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，例如函数、对象、应用程序编程接口(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可以用各种语言的各种版本来编写计算机

程序。

[0231] 计算机可读指令的功能可以根据需要在各种环境中进行组合或分布。在一些实施方案中，计算机程序可以包括一个指令序列。在一些实施方案中，计算机程序可以包括多个指令序列。在一些实施方案中，可以从一个位置提供计算机程序。在一些实施方案中，可以从多个位置提供计算机程序。在一些实施方案中，计算机程序可以包括一个或多个软件模块。在一些实施方案中，计算机程序可以部分或全部包括一个或多个web应用程序、一个或多个移动应用程序、一个或多个独立运行的应用程序、一个或多个web浏览器插件、扩展、加载项或附加项或其组合。

[0232] 在一些实施方案中，计算机处理可以是统计学、数学、生物学或其任意组合的方法。在一些实施方案中，计算机处理方法包括降维法，包括例如，逻辑回归、降维、主成分分析、自动编码器、奇异值分解、傅立叶基数、奇异值分解、小波、判别分析、支持向量机、基于树的方法、随机森林、梯度提升树、逻辑回归、矩阵分解、网络聚类和神经网络。

[0233] 在一些实施方案中，计算机处理方法是监督机器学习方法，包括，例如，回归、支持向量机、基于树的方法和网络。

[0234] 在一些实施方案中，计算机处理方法是无监督的机器学习方法，包括，例如，聚类、网络、主成分分析和矩阵分解。

G.数据库

[0235] 在一些实施方案中，本文公开的主题可以包括一个或多个数据库，或者所述一个或多个数据库用于存储患者数据、生物数据、生物序列或参考序列的用途。参考序列可以从数据库获得。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可以适合存储和检索序列信息。在一些实施方案中，合适的数据库可以包括，例如，关系数据库、非关系数据库、对象定向性数据库、对象数据库、实体关系模型数据库、关联数据库和XML数据库。在一些实施方案中，数据库可以是基于互联网的。在一些实施方案中，数据库可以是基于web的。在一些实施方案中，数据库可以是基于云计算的。在一些实施方案中，数据库可以基于一个或多个本地计算机存储装置。

V.癌症诊断和检测

[0236] 本文中所述的经训练的机器学习方法、模型和判别分类器可用于各种医学应用，包括癌症检测、诊断和治疗反应性。因为模型采用个体元数据和分析物得到的特征进行了训练，因此应用可以被定制成将群中的个体分层并且相应地指导治疗决策。

A.诊断

[0237] 本文提供的方法和系统可以使用基于人工智能的方法来执行预测分析，以分析从受试者(患者)获取的数据，以产生患有癌症(例如，CRC)的受试者的诊断输出。例如，该应用可以将预测算法应用于所获取的数据，以产生患有癌症的受试者的诊断。预测算法可以包括基于人工智能的预测器，例如基于机器学习的预测器，其配置为处理所获取的数据以产生患有癌症的受试者的诊断。

[0238] 机器学习预测器可以使用数据集(例如，通过对一组或多组患有癌症的患者群组的个体生物样品进行多分析物测定而产生的数据集)作为输入，并使用受试者的已知诊断(例如，分期和/或肿瘤分级)结果作为机器学习预测器的输出来训练。

[0239] 训练数据集(例如，通过对个体的生物样品进行多分析物测定而产生的数据集)可

以从，例如，具有共同特征(特征)和结果(标记)的一组或多组受试者中产生。训练数据集可以包括一组特征和对应于与诊断有关的特征的标记。特征可以包括特性，例如cfDNA测定测量的某些范围或类别，例如从健康和疾病样品获得的生物样品中的cDNA片段与参考基因组的一组箱(基因组窗口)的每一个重叠或落入其中的计数。例如，在给定时间点从给定受试者收集的一组特征可以共同作为诊断标志，其可以指示在给定时间点已鉴定出受试者的癌症。特性还可以包括指示受试者的诊断结果的标记，如针对一种或多种癌症。

[0240] 标记可以包括结果，例如，受试者的已知诊断结果(例如，分期和/或肿瘤分级)。结果可包括与受试者癌症相关的特性。例如，特性可以指示受试者患有一种或多种癌症。

[0241] 训练集(例如，训练数据集)可以通过对与一组或多组受试者(例如，患有或不患有一种或多种癌症的患者的回顾性和/或前瞻性群组)相对应的一组数据进行随机采样来选择。或者，训练集(例如，训练数据集)可以通过对与一组或多组受试者(例如，患有或不患有一种或多种癌症的患者的回顾性和/或前瞻性群组)相对应的一组数据进行比例采样来选择。训练集可以在与一组或多组受试者(例如，来自不同临床站点或试验的患者)相对应的数据集之间平衡。可以训练机器学习预测器，直到满足用于准确性或性能的某些预定条件，例如具有与诊断准确性测量相对应的最小期望值。例如，诊断准确性量度可以对应于对受试者中一种或多种癌症的诊断、分期或肿瘤分级的预测。

[0242] 诊断准确度量度的示例可以包括灵敏度、特异性、阳性预测值(PPV)、阴性预测值(NPV)、准确性以及与检测或预测癌症(例如，结直肠癌)的诊断准确性相对应的ROC曲线的曲线下面积(AUC)。

[0243] 在另一方面中，本公开内容提供了一种用于鉴定受试者中的癌症的方法，所述方法包括：(a)提供包含来自所述受试者的无细胞核酸(cfNA)分子的生物样品；(b)对来自所述受试者的所述cfNA分子进行甲基化测序以产生多个cfNA测序读取；(c)将所述多个cfNA测序读取与参考基因组比对；(d)在所述参考基因组的第一多个基因组区域中的每一个处产生所述多个cfNA测序读取的定量测量以产生第一cfNA特征集，其中所述参考基因组的所述第一多个基因组区域包含至少约10个不同区域，所述至少约10个不同区域中的每一个；和(e)将经训练的算法应用于所述第一cfNA特征集以产生所述受试者患有所述癌症的可能性。

[0244] 例如，这种预定条件可以是预测癌症(例如结肠直肠癌、乳癌、胰腺癌或肝癌)的灵敏度包括例如至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％的值。

[0245] 作为另一示例，这种预定条件可以是预测癌症(例如结肠直肠癌、乳癌、胰腺癌或肝癌)的特异性包括例如至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％的值。

[0246] 作为另一示例，这种预定条件可以是预测癌症(例如结肠直肠癌、乳癌、胰腺癌或肝癌)的阳性预测值(PPV)包括例如至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％的值。

[0247] 作为另一示例，这种预定条件可以是预测癌症(例如结肠直肠癌、乳癌、胰腺癌或肝癌)的阴性预测值(NPV)包括例如至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％的值。

[0248] 作为另一示例，这种预定条件可以是预测癌症(例如结肠直肠癌、乳癌、胰腺癌或肝癌)的ROC曲线的AUC包括例如至少约0.50、至少约0.55、至少约0.60、至少约0.65、至少约0.70、至少约0.75、至少约0.80、至少约0.85、至少约0.90、至少约0.95、至少约0.96、至少约0.97、至少约0.98或至少约0.99的值。

[0249] 在上述方面的任一方面的一些示例中，一种方法还包括监测受试者中的疾病的进展，其中所述监测至少部分地基于基因序列特征。在一些示例中，所述疾病是癌症。

[0250] 在上述方面的任一方面的一些示例中，一种方法还包括确定受试者中的癌症的来源组织，其中所述确定至少部分地基于基因序列特征。

[0251] 在上述方面的任一方面的一些示例中，一种方法还包括估计受试者中的肿瘤负荷，其中所述估计至少部分地基于基因序列特征。

B.治疗反应性

[0252] 本文中所述的预测性分类器、系统和方法可用于对个体群进行分类以用于多种临床应用(例如基于对个体的生物样品执行的多分析物测定)。这种临床应用的示例包括：检测早期癌症、诊断癌症、将癌症分类为疾病的具体阶段，或确定对用于治疗癌症的治疗剂的反应性或耐受性。

[0253] 本文中所述的方法和系统可应用于各种癌症类型，类似于分级和分期，并且不限于单一癌症疾病类型。因此，分析物和测定的组合可以用于本系统和方法中以预测在不同组织中不同癌症类型对癌症疗法的反应性并且基于治疗反应性对个体进行分类。在一个示例中，本文中所述的分类器将个体组分层为有治疗反应者和无治疗反应者。

[0254] 本公开内容还提供一种用于确定关注的病况或疾病的药物靶标(例如对于具体的类别而言是相关/重要的基因)的方法，包括对于从个体获得的样品评估至少一个基因的基因表达水平；使用邻域分析例程来确定与样品的分类相关的基因，由此查明与分类相关的一个或多个药物靶标。

[0255] 本公开内容还提供了一种用于确定设计为治疗疾病类别的药物的功效的方法，包括：从患有所述疾病类别的个体获得样品；使所述样品经受所述药物；对暴露于药物的样品评估至少一个基因的基因表达水平；和使用构建有加权表决方案的计算机模型，以便根据所述样品相对于所述模型的相对基因表达水平而将暴露于药物的样品分类至疾病的类别。

[0256] 本公开内容还提供了一种用于确定设计为治疗疾病类别的药物的功效的方法，其中个体已经接受所述药物，所述方法包括：从接受所述药物的所述个体获得样品；对所述样品评估至少一个基因的基因表达水平；和使用构建有加权表决方案的模型而将所述样品分类至疾病的类别，包括与所述模型的基因表达水平相比评价所述样品的基因表达水平。

[0257] 又一种应用是一种确定个体是否属于某表型类别(例如智力、对治疗的反应、寿命长度、病毒感染的可能性或肥胖)的方法，所述方法包括：从所述个体获得样品；对所述样品评估至少一个基因的基因表达水平；和使用构建有加权表决方案的模型而将所述样品分类至疾病的类别，包括与所述模型的基因表达水平相比评价所述样品的基因表达水平。

[0258] 生物标志物可以用于预测患有结肠癌的患者的预后。将患者分类为高风险(预后不良)或低风险(预后良好)的能力可以实现针对这些患者选择适当的疗法。例如,高风险患者可能得益于激进疗法,而该激进疗法可能对低风险患者没有显著优势。

[0259] 预测性生物标志物可以通过鉴定对特定的癌症疗法可能是“特殊反应者”的患者子集或者可能得益于替代治疗方式的个体来指导治疗决策。

[0260] 在一个方面中,本文中所述的系统和方法涉及基于对化疗剂治疗癌症的治疗反应性而对群体进行分类,所述化疗剂的类别为DNA损伤剂、DNA修复靶向疗法、DNA损伤信号传导的抑制剂、DNA损伤诱导的细胞周期停滞的抑制剂,以及对间接导致DNA损伤的过程的抑制,但不限于这些类别。这些化疗剂中的每种都可以被视为“DNA损伤治疗剂”。

[0261] 患者的分析物数据被分类在高风险患者组和低风险患者组,例如临床复发的高风险或低风险患者,并且结果可以用于确定治疗的过程。例如,确定为高风险患者的患者可以在手术后接受辅助化疗的治疗。对于被视为低风险患者的患者,可以在手术后不给予辅助化疗。因此,在某些方面,本公开内容提供了一种用于制备表明复发风险的结肠癌基因表达图谱的方法。

[0262] 在各种示例中,本文中所述的分类器将个体群在有治疗反应者和无治疗反应者之间进行分层。

[0263] 在各种示例中,治疗选自:烷化剂、植物生物碱、抗肿瘤抗生素、抗代谢物、拓扑异构酶抑制剂、类视黄醇、检查点抑制剂疗法和VEGF抑制剂。

[0264] 群体可以分层为有治疗反应者和无治疗反应者的治疗的示例包括但不限于化疗剂和抗体疗法,所述化疗剂包括:索拉非尼(sorafenb)、瑞戈菲尼、伊马替尼、艾瑞布林、吉西他滨、卡培他滨、帕唑帕尼、拉帕替尼、达拉菲尼、马来酸舒尼替尼、克唑替尼、依维莫司、torisirolimus、西罗莫司、阿西替尼、吉非替尼、anastrole、比卡鲁胺、氟维司群、雷替曲塞、培美曲塞、乙酸戈舍瑞林、厄洛替尼、维莫非尼、visiodegib、柠檬酸他莫昔芬、紫杉醇、多西他赛、卡巴他赛、奥沙利铂、阿柏西普、贝伐珠单抗、曲司珠单抗、培妥珠单抗、帕木单抗、紫杉烷、博来霉素、美法仑、白花丹素、camptosar、丝裂霉素-C、米托蒽醌、聚(苯乙烯马来酸)-缀合的新制癌菌素(SMANCS)、多柔比星、聚乙二醇多柔比星、FOLFORI、5-氟尿嘧啶、替莫唑胺、帕瑞肽、替加氟、吉美拉西、奥替拉西、伊曲康唑、硼替佐米、来那度胺、伊立替康、表柔比星、罗米地辛、瑞米诺司他、他喹莫德、瑞法替尼、拉帕替尼、Tyverb®、Arenegyr、NGR-TNF、帕瑞肽、Signifor®、ticilimumab、曲美木单抗、兰索拉唑、PrevOnco®、ABT-869、利尼伐尼(linifanib)、vorolanib、tivantinib、Tarceva®、厄洛替尼、Stivarga®、瑞戈菲尼、氟代索拉非尼、布立尼布、脂质体多柔比星、乐伐替尼、雷莫芦单抗、peretinoin、Ruchiko、muparfostat、Teysuno®、替加氟、吉美拉西、奥替拉西和奥兰替尼;所述抗体疗法包括:阿仑珠单抗、阿特珠单抗、伊匹单抗、纳武单抗、奥法木单抗、帕博利珠单抗或利妥昔单抗。

[0265] 在其他示例中,群体可以被分层为对检查点抑制剂疗法(例如与PD-1或CTLA4结合的化合物)有反应者和无反应者。

[0266] 在其他示例中,群体可以被分层为对与VEGF途径靶标结合的抗VEGF疗法有反应者和无反应者。

VI.适应症

[0267] 在一些示例中，生物状况可以包括疾病。在一些示例中，生物状况可以是疾病的阶段。在一些示例中，生物状况可以是生物状态的逐渐变化。在一些示例中，生物状况可以是治疗效果。在一些示例中，生物状况可以是药物效果。在一些示例中，生物状况可以是手术效果。在一些示例中，生物状况可以是生活方式改善之后的生物状态。生活方式改善的非限制性示例包括膳食变化、吸烟改变和睡眠模式改变。在一些示例中，生物状况是未知的。本文中所述的分析可以包括机器学习以推断未知的生物状况或解释未知的生物状况。

[0268] 在一个示例中，本系统和方法具体地可用于与结肠癌(在结肠(大肠的最长部分)的组织中形成的癌)相关的应用。大多数结肠癌是腺癌(起始于组成线内部器官并且具有腺样性质的细胞的癌)。癌症进展是以体内癌症的阶段或范围为特征。分期通常是基于肿瘤的尺寸、淋巴结是否包含癌，以及癌是否已经从原始位点扩展到身体其他部分。结肠癌的阶段包括I期、II期、III期和IV期。除非另有明确说明，术语“结肠癌”表示处于0期、I期、II期(包括IIA或IIB期)、III期(包括IIIA、IIIB或IIIC期)或IV期的结肠癌。在本文中的一些示例中，结肠癌来自任何阶段。在一个示例中，结肠癌是I期结肠直肠癌。在一个示例中，结肠癌是II期结肠直肠癌。在一个示例中，结肠癌是III期结肠直肠癌。在一个示例中，结肠癌是IV期结肠直肠癌。

[0269] 可以通过本公开内容的方法推断的状况包括：例如，癌症、肠相关疾病、免疫介导的炎性疾病、神经病、肾病、产前疾病和代谢性疾病。

[0270] 在一些示例中，本公开内容的方法可以用于诊断癌症。癌症的非限制性示例包括：腺瘤(腺瘤性息肉)、无蒂锯齿状腺瘤(sessile serrated adenoma，SSA)、晚期腺瘤、结肠直肠发育异常、结肠直肠腺瘤、结肠直肠癌、结肠癌、直肠癌、结直肠癌、结肠直肠腺癌、类癌瘤、胃肠类癌瘤、胃肠基质瘤(GIST)、淋巴瘤和肉瘤。

[0271] 可以由本公开的方法和系统推断的癌症的非限制性示例包括急性淋巴母细胞白血病(ALL)、急性髓细胞性白血病(AML)、肾上腺皮质癌、卡波济肉瘤、肛门癌、基底细胞癌、胆管癌、膀胱癌、骨癌、骨肉瘤、恶性纤维组织细胞瘤、脑干神经胶质瘤、脑癌、颅咽管瘤、成室管膜细胞瘤、室管膜瘤、成神经管细胞瘤、髓上皮瘤(medulloeptithelioma)、松果体实质肿瘤、乳腺癌、支气管肿瘤、伯基特淋巴瘤、非霍奇金淋巴瘤、类癌肿瘤、宫颈癌、脊索瘤、慢性淋巴细胞性白血病(CLL)、慢性髓性白血病(CML)、结肠癌、结直肠癌、皮肤T细胞淋巴瘤、原位导管癌、子宫内膜癌、食道癌、尤因肉瘤、眼癌、眼内黑色素瘤、视网膜母细胞瘤、纤维组织细胞瘤、胆囊癌、胃癌、神经胶质瘤、毛细胞白血病、头颈癌、心脏癌、肝细胞(肝)癌、霍奇金淋巴瘤、下咽癌、肾癌、喉癌、唇癌、口腔癌、肺癌、非小细胞癌、小细胞癌、黑色素瘤、口腔癌、骨髓增生异常综合症、多发性骨髓瘤、髓母细胞瘤、鼻腔癌、鼻旁窦癌、成神经细胞瘤、鼻咽癌、口腔癌、口咽癌、骨肉瘤、卵巢癌、胰腺癌、乳头瘤病、副神经节瘤、甲状旁腺癌、阴茎癌、咽癌、垂体瘤、浆细胞赘生物、前列腺癌、直肠癌、肾细胞癌、横纹肌肉瘤、唾液腺癌、Sezary综合征、皮肤癌、小肠癌、软组织肉瘤、鳞状细胞癌、睾丸癌、喉癌、胸腺瘤、甲状腺癌、尿道癌、子宫癌、子宫肉瘤、阴道癌、外阴癌、瓦尔登斯特伦巨球蛋白血症和维尔姆斯瘤。

[0272] 可以通过本公开的方法和系统推断的肠相关疾病的非限制性示例包括：克罗恩病、结肠炎、溃疡性结肠炎(UC)、炎性肠病(IBD)、肠易激综合征(IBS)和乳糜泻。在一些示例中，所述疾病是炎性肠病、结肠炎、溃疡性结肠炎、克罗恩病、微小性结肠炎、胶原性结肠炎、

淋巴细胞性结肠炎、改道性结肠炎、白塞病和未定类结肠炎。

[0273] 可以通过本公开的方法和系统推断的免疫介导的炎性疾病的非限制性示例包括：银屑病、结节病、类风湿关节炎、哮喘、鼻炎(枯草热)、食物变态反应、湿疹、狼疮、多发性硬化、纤维肌痛、1型糖尿病和莱姆病。可以通过本公开的方法和系统推断的神经病的非限制性示例包括：帕金森病、亨廷顿病、多发性硬化、阿尔茨海默病、卒中、癫痫、神经变性和神经病变。可以通过本公开的方法和系统推断的肾病的非限制性示例包括间质性肾炎、急性肾衰和肾病。可以通过本公开的方法和系统推断的产前疾病的非限制性示例包括：唐氏综合征、非整倍性、脊柱裂、三体性、爱德华兹综合征、畸胎、骶尾部畸胎瘤(SCT)、巨脑室、肾不发育、囊性纤维化和胎儿水肿。可以通过本公开的方法和系统推断的代谢性疾病的非限制性示例包括：胱氨酸病、法布里病、戈谢病、莱施-奈恩综合征、尼曼-匹克病、苯丙酮尿症、蓬佩病、泰-萨克斯病。

[0274] 具体示例的特定细节可以以任何适合的方式组合而不脱离本公开内容的发明示例的精神和范围。然而，本发明的其他示例可以涉及与每个方面相关的特定示例，或者这些单独的方面的特定组合。本文中提及的所有专利、专利申请、出版物和说明书出于所有目的通过援引而整体地并入。

VII.试剂盒

[0275] 本公开内容提供了用于鉴定或监测受试者的癌症的试剂盒。试剂盒可以包含探针，用于鉴定受试者的无细胞生物样品中多个癌症相关的基因组基因座各自的序列的定量测量(例如表明存在、不存在或相对量)。无细胞生物样品中多个癌症相关的基因组基因座各自的序列的定量测量(例如表明存在、不存在或相对量)可以表明一种或多种癌症。探针可以对于无细胞生物样品中多个癌症相关的基因组基因座处的序列是选择性的。试剂盒可以包含说明以便将探针用于处理无细胞生物样品，从而产生数据集，所述数据集表明受试者的无细胞生物样品中多个癌症相关的基因组基因座各自的序列的定量测量(例如表明存在、不存在或相对量)。在一个实施方案中，试剂盒包含：引物集、PCR反应组分、测序试剂、破坏性最小的转化试剂和文库制备试剂。

[0276] 试剂盒中的探针可以对于无细胞生物样品中多个癌症相关的基因组基因座处的序列是选择性的。试剂盒中的探针可以被配置成选择性地富集与多个癌症相关的基因组基因座对应的核酸(例如RNA或DNA)分子。试剂盒中的探针可以是核酸引物。试剂盒中的探针可以具有与多个癌症相关的基因组基因座或基因组区域中的一个或多个的核酸序列互补的序列。多个癌症相关的基因组基因座或基因组区域可以包含被鉴定用于靶向甲基化测序的至少2、至少3、至少4、至少5、至少6、至少7、至少8、至少9、至少10、至少11、至少12、至少13、至少14、至少15、至少16、至少17、至少18、至少19、至少20或更多个不同癌症相关的基因组基因座或基因组区域。

[0277] 试剂盒中的说明可以包含利用探针测定无细胞生物样品的说明，所述探针对于无细胞生物样品中多个癌症相关的基因组基因座处的序列是选择性的。这些探针可以是核酸分子(例如RNA或DNA)，其具有与多个癌症相关的基因组基因座中的一个或多个的核酸序列(例如RNA或DNA)互补的序列。这些核酸分子可以是引物或富集序列。测定无细胞生物样品的说明可以包括对阵列杂交、聚合酶链反应(PCR)或核酸测序(例如DNA测序或RNA测序)的说明以处理无细胞生物样品而产生数据集，所述数据集表明无细胞生物样品中多个癌症相

关的基因组基因座各自的序列的定量测量(例如表明存在、不存在或相对量)。无细胞生物样品中多个癌症相关的基因组基因座各自的序列的定量测量(例如表明存在、不存在或相对量)可以表明一种或多种癌症。

[0278] 试剂盒中的说明可以包含对测量和解释测定读出的说明,该读出可以在多个癌症相关的基因组基因座或基因组区域中的一个或多个处进行量化而产生数据集,所述数据集表明无细胞生物样品中多个癌症相关的基因组基因座各自的序列的定量测量(例如表明存在、不存在或相对量)。例如,与多个癌症相关的基因组基因座对应的阵列杂交或聚合酶链反应(PCR)的量化可以产生数据集,所述数据集表明无细胞生物样品中多个癌症相关的基因组基因座各自的序列的定量测量(例如表明存在、不存在或相对量)。测定读出可以包括:定量PCR(qPCR)值、数字PCR(dPCR)值、数字液滴PCR(ddPCR)值、荧光值等,或其标准化值。

实施例

实施例1:靶向EM测序文库制备和分类器产生

[0279] 起始材料:10-200ng的双链DNA。

1.DNA制备

[0280] 在氧化之前,将EDTA从DNA除去并且该DNA样品具有29μl的最终体积。对照DNA用于评估氧化和脱氨。为了在Illumina平台上测序,关于用法推荐,参考了酶法甲基测序试剂盒手册(NEB#E7120)。

2.衔接子连接

3.5-甲基胞嘧啶和5-羟甲基胞嘧啶的氧化

[0281] 制备了TET2缓冲剂。然后将TET2反应剂添加到具有TET2反应缓冲剂补剂的一试管,随后充分混合。在冰上,将TET2反应缓冲剂、氧化补剂、氧化促进剂和TET2酶直接添加到DNA样品。然后将该混合物通过涡旋进行充分混合。在简短地离心之后,将稀释的Fe(II)溶液添加到该混合物。然后将该混合物通过涡旋或通过上下移液而进行充分混合,并且简短地离心。然后将混合物在热循环仪中在37℃孵育1小时。随后将样品转移至冰,之后用1μl终止试剂(黄色)处理。然后将混合物通过涡旋或通过上下移液至少10次而进行充分混合,并且简短地离心。最后,将混合物在热循环仪中在37℃孵育30分钟而后在4℃孵育。

4.清理TET2转化的DNA

[0282] 将样品纯化珠通过涡旋进行重新混悬。接着,将NEBNext样品纯化珠添加到每个样品,随后通过上下移液进行充分混合。将样品在工作台上在室温孵育至少5分钟。然后将试管抵靠适当的磁性支架放置以便使珠从上清液分离。在5分钟后(或者当溶液澄清时),将上清液小心地移出以避免扰动包含DNA靶标的珠,并且弃去。在磁性支架上,将新鲜制备的80%乙醇添加到每个试管。在将上清液小心地移出和弃去之前,将样品在室温下孵育30秒。重复清洗一次,总计两次清洗。在第二次清洗之后,将所有可见的液体用p10移液器吸头移除。然后使珠风干2分钟,同时使试管在磁性支架上保持盖子打开。然后将试管从磁性支架移除。将DNA用洗脱缓冲剂从珠洗脱。将洗脱缓冲剂添加到每个试管并且通过上下移液10次进行充分混合。然后将样品在室温下孵育至少1分钟。若需要,在将试管放回磁性支架上之前,将样品快速离心以从试管侧面收集液体。然后将试管放回磁性支架上。在3分钟之后(或每当溶液澄清时),将上清液中洗脱的DNA转移至新的PCR试管。

5.DNA的变性

[0283] 在胞嘧啶脱氨之前,将DNA利用甲酰胺或0.1N氢氧化钠进行变性。

6.胞嘧啶的脱氨

[0284] 在冰上,将APOBEC反应缓冲剂、BSA和APOBEC添加到变性的DNA。然后将该混合物通过涡旋或通过上下移液至少10次而进行充分混合,而后简短地离心。然后将混合物在热循环仪中在37℃孵育3小时,而后在4℃孵育。

7.清理脱氨的DNA

[0285] 将样品纯化珠通过涡旋进行重新混悬。接着,将100μl重新混悬的NEBNext样品纯化珠添加到每个样品,随后通过上下移液至少10次进行充分混合。在最后混合期间,将所有液体小心地排出吸头。然后将样品在工作台上在室温孵育至少5分钟。在5分钟后(或者当溶液澄清时),将上清液小心地移出并且弃去。在磁性支架上,将新鲜制备的80%乙醇添加到试管。在将上清液小心地移出和弃去之前,将样品在室温下孵育30秒。重复清洗一次,总计两次清洗。然后使珠风干90秒,同时使试管在磁性支架上保持盖子打开。然后将DNA靶标用洗脱缓冲剂从珠洗脱。将洗脱缓冲剂添加到每个试管并且通过上下移液10次而进行充分混合。将样品在室温下孵育至少1分钟。若需要,在将试管放回磁性支架上之前,将样品快速离心以从试管侧面收集液体。然后将试管放回磁性支架上。在3分钟之后(或每当溶液澄清时),将上清液中洗脱的DNA靶标转移至新的PCR试管。

8.多路复用扩增和靶向甲基化分类

[0286] 原始数据文件用于以常规工具进行比对和甲基化判定,从而允许对基因组的预鉴定区域进行靶向甲基化分析。对酶法转化的DNA进行全基因组扩增。利用5'-生物素化捕获探针对酶法转化的文库进行靶标富集以特异性地捕捉包含靶CpG位点的预鉴定的DNA片段。杂交选择是利用Illumina TruSightVR快速捕获试剂盒进行。在杂交步骤中使用捕获靶缓冲剂3(Illumina)而不使用富集杂交缓冲剂。在杂交之后,将捕获的DNA片段在14个PCR循环下进行扩增。将靶标捕获文库在Illumina HiSeqVR 2500测序仪上在快速运行模式下对4至5个样品利用2x100次循环运行进行测序。将10%PhiX掺加到酶法测序文库中,从而提高碱基多样性,测序质量更好。

[0284] 利用常规方法将FASTQ文件映射到参考基因组,并且对甲基化评分进行计算以供疾病分类。将包括与健康、疾病、疾病状态和治疗反应性相关联的CpG位点集的特征化数据输入机器学习模型中以鉴定将群中的个体分层的分类器。

实施例2:用耐转化的测序衔接子/引物系统进行靶向EM测序

[0287] 将已知序列的鉴定的衔接子与样品中含有未知序列的DNA分子的末端连接。然后,将衔接子用于利用与已知衔接子对应的单个引物集来PCR扩增不同分子的整个文库。在后续测序反应期间,还将连接的衔接子序列用作测序引物的结合位点。为了利用由双链测序提供的数据,将含有唯一分子标识符(UMI)的部分双链的衔接子连接到双链DNA。

[0288] 为了改进EM测序双链测序的稳健性、氧化效率(和降低成本),可将耐转化的衔接子用于提高测序文库质量的一致性。耐转化的衔接子仅包含未修饰的碱基并且允许衔接子的完全碱基转化。耐转化的衔接子的示例示于图4子图A中。

[0289] 在无转化的情况下,用这些耐转化的衔接子产生的测序文库可以利用与原始衔接子序列匹配的PCR和测序引物集进行扩增和测序。在有转化的情况下,测序文库可以利用与转化的衔接子序列匹配的PCR和测序引物进行扩增和测序,如图4板块B中所示。

[0290] 耐转化的衔接子、PCR引物和测序引物的功能性示例集已经被测试(图6)。用耐转化的衔接子或含有5mC的衔接子产生的文库的测序文库收率。未执行Tet介导的氧化步骤，所以全部C和5mC易于发生C向U转化。在无转化的情况下含有5mC的衔接子系统更有效，而耐转化的衔接子系统需要转化，以便耐转化的衔接子可以利用转化特异性PCR引物进行扩增。这些转化特异性PCR引物的DNA序列列于表1中。

表1

SEQ ID NO:	寡聚体名称	DNA 序列
1	衔接子顶部（UMI GTCT）	TGAGGAATGAGCACGTACTGTCTT
2	衔接子底部（UMI GTCT）	/5'-磷酸/AGACAGTACGTGCTCATTGATAGAGTG
3	索引 PCR 引物，具有转化 F（索引 GACACAGT）	AATGATACGGCGACCACCGAGATCTACACGACACAGTACA CTCTTTCCCTACACGACGTTGGGTGAGGAATGAGTATGTATT
4	索引 PCR 引物，具有转化 R（索引 GACACAGT）	CAAGCAGAAGACGGCATACGAGATGACACAGTGTGACTGG AGTTCAGACGTGTCCCACTCTATCAATAAACACATACT
5	测序引物，具有转化 r1	TCCCTACACGACGTTGGGTGAGGAATGAGTATGTATT
6	测序引物，具有转化 r2	GTTCAGACGTGTCCCACTCTATCAATAAACACATACT
7	测序引物，具有转化 i1	TTATTGATAGAGTGGGACACGTCTGAACTCCAGTCAC
	测序引物，具有转化 i2	CTCATTCCTCACCCAACGTCGTGTAGGGAAAGAGTGT
8	索引 PCR 引物，无转化 F（索引 GACACAGT）	AATGATACGGCGACCACCGAGATCTACACGACACAGTAC ACTCTTTCCCTACACGACGTTGGGTGAGGAATGAGCACGTACT
9	索引 PCR 引物，无转化 R（索引 GACACAGT）	CAAGCAGAAGACGGCATACGAGATGACACAGTGTGACTGG AGTTCAGACGTGTCCCACTCTATCAATGAGCACGTACT
10	测序引物，无转化 r1	TCCCTACACGACGTTGGGTGAGGAATGAGCACGTACT
11	测序引物，无转化 r2	GTTCAGACGTGTCCCACTCTATCAATGAGCACGTACT
12	测序引物，无转化 i1	TCATTGATAGAGTGGGACACGTCTGAACTCCAGTCAC
13	测序引物，无转化 i2	CTCATTCCTCACCCAACGTCGTGTAGGGAAAGAGTGT

[0291] 虽然本文中已经示出和描述了本发明的优选实施方案，但是对于本领域技术人员显而易见的是这些实施方案仅仅作为示例而提供。本发明并非意在受限于说明书内提供的具体示例。虽然已经参考上述说明书对本发明进行描述，但是本文中对实施方案的描述和说明并非意在以限制含义进行解释。本领域技术人员在不脱离本发明的情况下现将想到多种变化、改变和替代。此外，应当理解，本发明的所有方面不限于本文所阐述的具体描述、配

置或相对比例，它们依赖于各种条件和变量。应理解，本文中所述的发明的实施方案的各种替代方案可以用于实施本发明。因此，考虑到本发明还应覆盖任何这种替代方案、修改、改变或等同方案。以下权利要求旨在定义本发明的范围并且由此覆盖这些权利要求范围内的方法和结构及其等同方案。

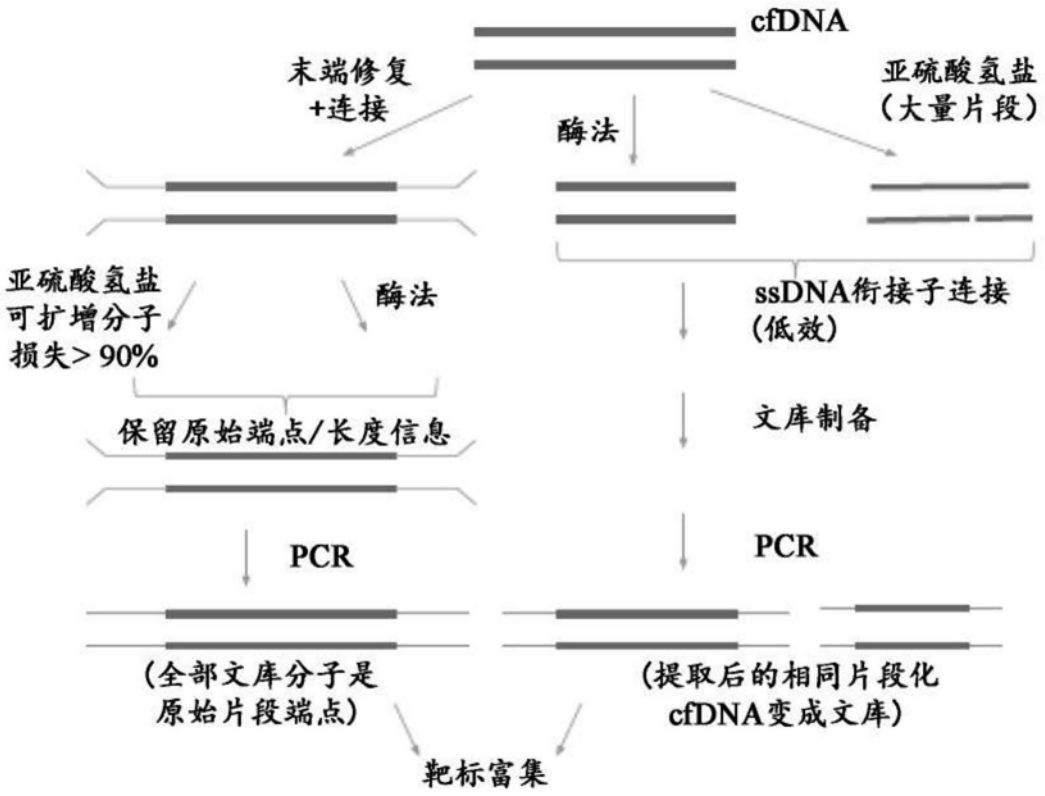

cfDNA
末端修复
+连接
酶法
亚硫酸氢盐
(大量片段)
亚硫酸氢盐
可扩增分子
损失>90%
酶法
ssDNA衔接子连接
(低效)
保留原始端点/长度信息
文库制备
PCR
PCR
(全部文库分子是
原始片段端点)
(提取后的相同片段化
cfDNA变成文库)
靶标富集

图1

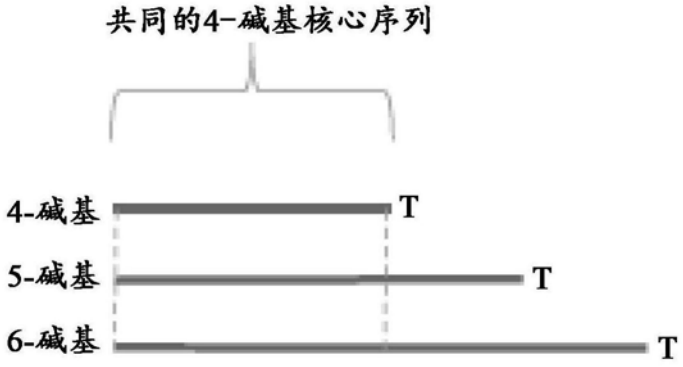

共同的4-碱基核心序列
4-碱基
T
5-碱基
T
6-碱基
T

图2

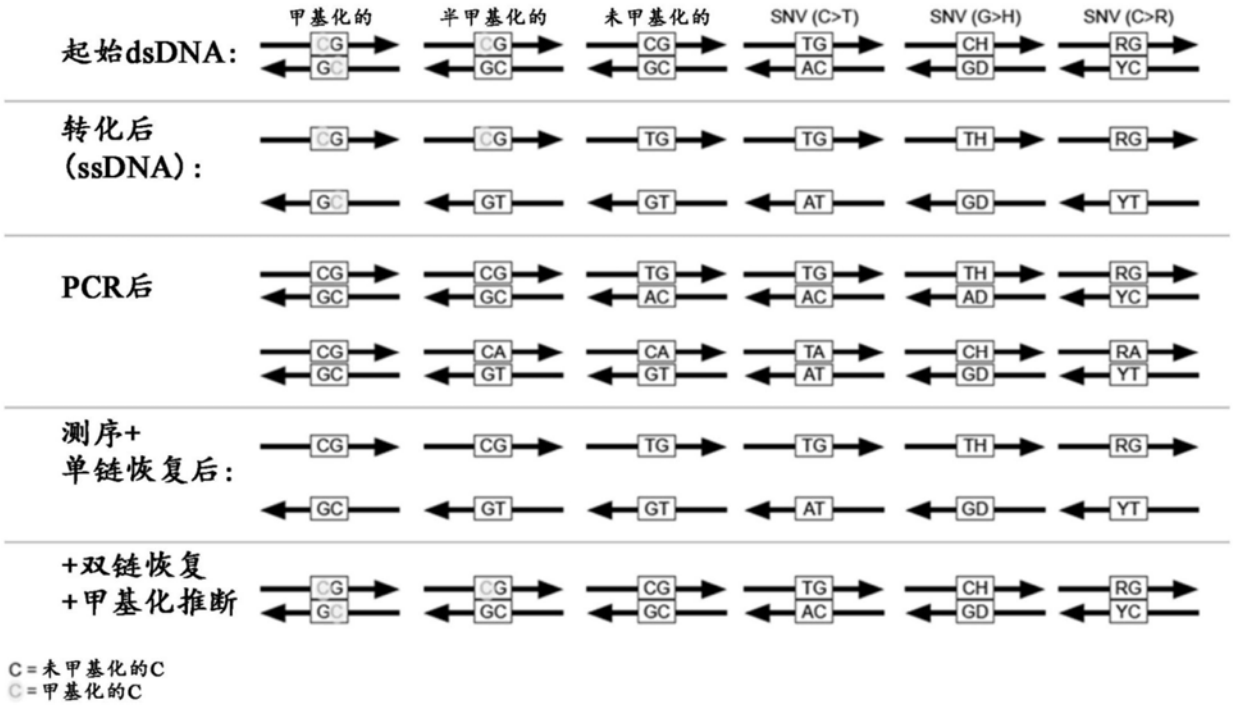

甲基化的
半甲基化的
未甲基化的
SNV (C>T)
SNV (G>H)
SNV (C>R)
起始dsDNA:
转化后
(ssDNA):
PCR后
测序+
单链恢复后:
+双链恢复
+甲基化推断
C=未甲基化的C
C=甲基化的C

图3

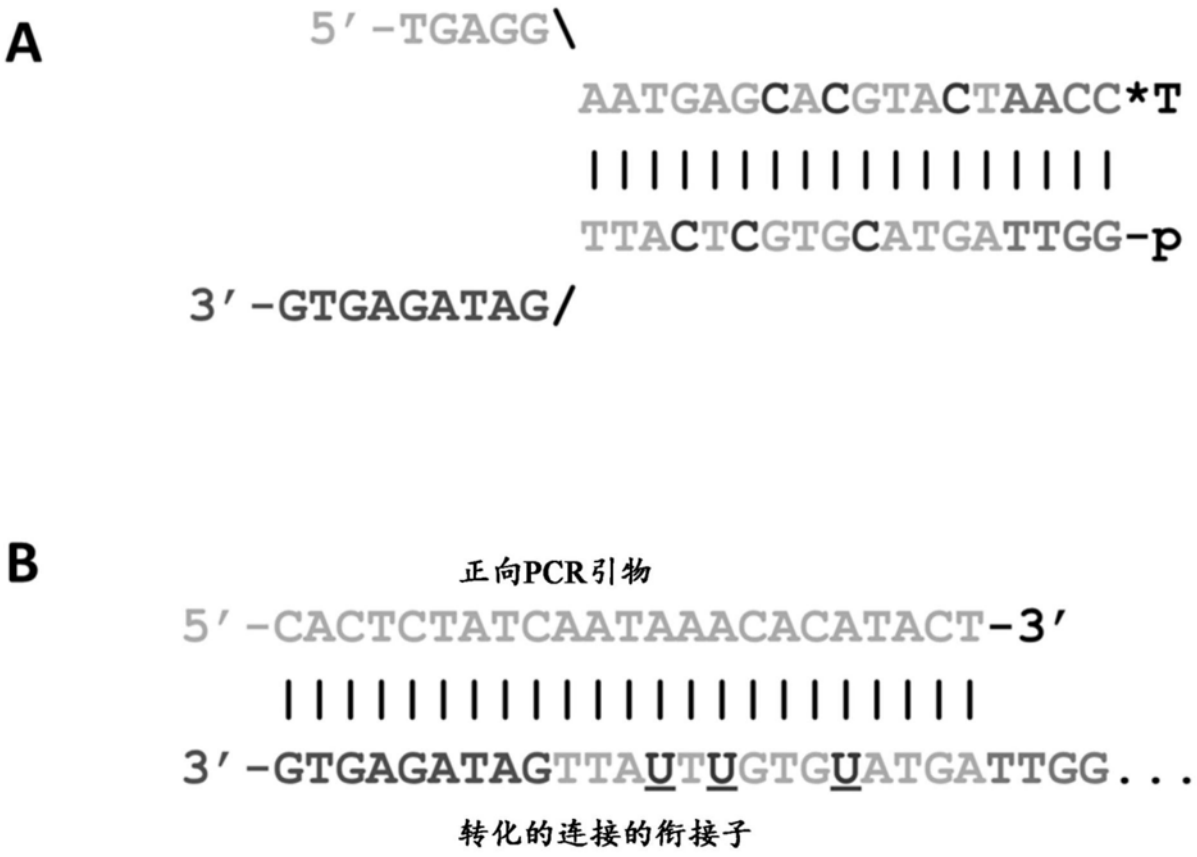

A
5′-TGAGG\
AATGAGCACGTACTAACC*T
TTACTCGTGCATGATTGG-p
3′-GTGAGATAG/
B
正向PCR引物
5′-CACTCTATCAATAAACACATACT-3′
3′-GTGAGATAGTTAUTUGTGUATGATTGG...
转化的连接的衔接子

图4

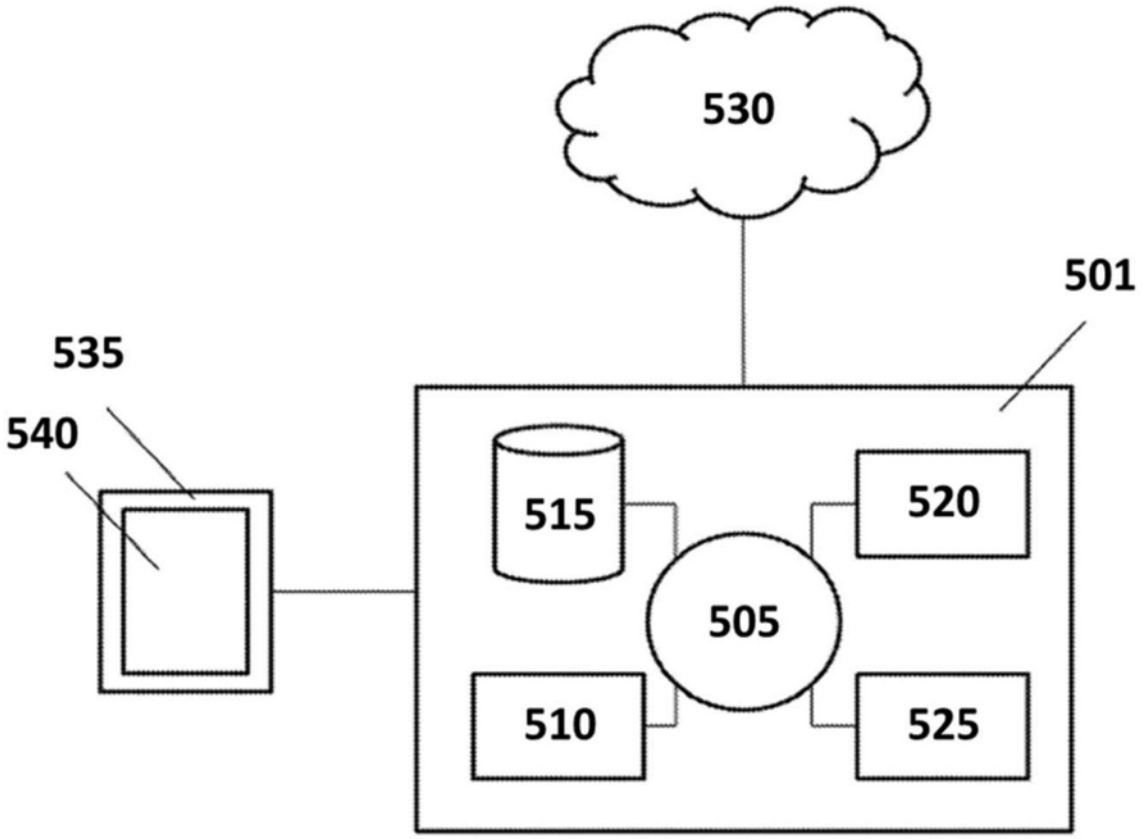

530
501
535
540
515
520
505
510
525

图5

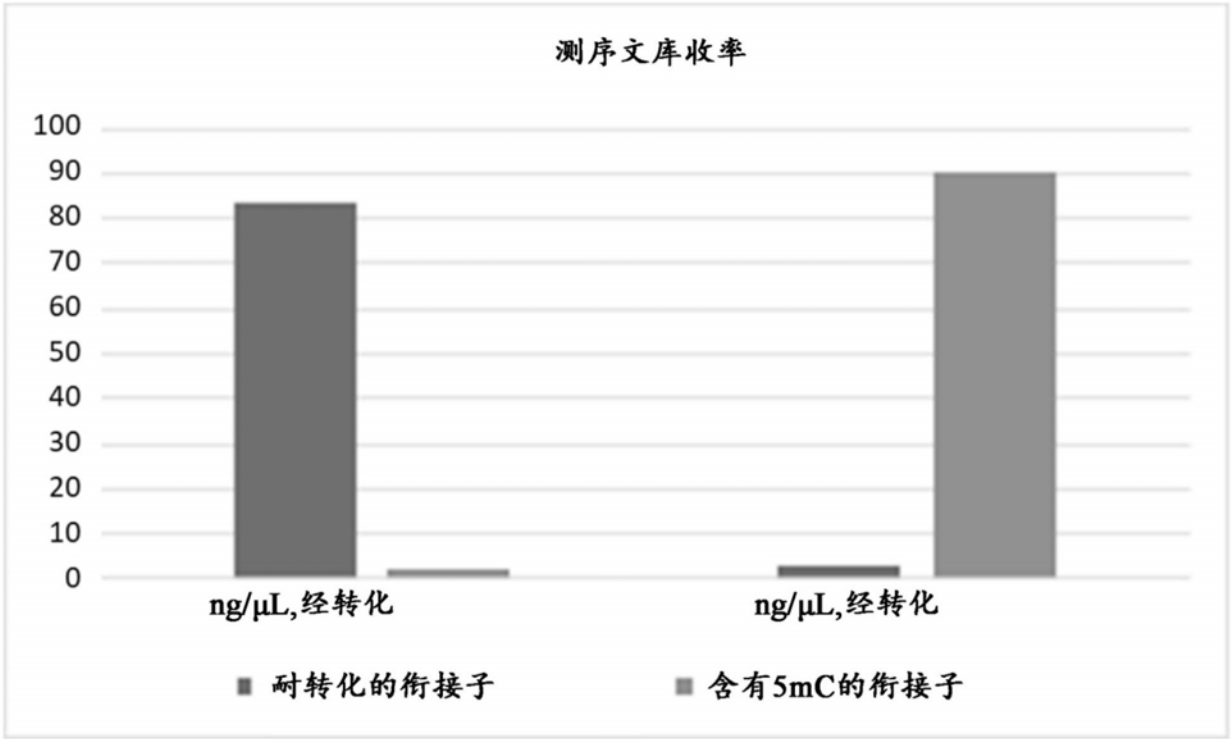

测序文库收率
100
90
80
70
60
50
40
30
20
10
0
ng/μL,经转化
ng/μL,经转化
耐转化的衔接子
含有5mC的衔接子

图6