(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113227393 B
(45)授权公告日 2025.02.11

(21)申请号 201980087384.6
(22)申请日 2019.10.31
(65)同一申请的已公布的文献号
申请公布号 CN 113227393 A
(43)申请公布日 2021.08.06
(30)优先权数据
62/753,826 2018.10.31 US
(85)PCT国际申请进入国家阶段日
2021.06.30
(86)PCT国际申请的申请数据
PCT/US2019/059217 2019.10.31
(87)PCT国际申请的公布数据
WO2020/092807 EN 2020.05.07
(73)专利权人 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
奥斯卡·韦斯特桑 何宇鹏
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
专利代理师 张晨 郑霞
(51)Int.Cl.
*C12Q 1/6806*(2006.01)
*C12Q 1/6869*(2006.01)
*C12Q 1/6886*(2006.01)
(56)对比文件
WO 2008009365 A2,2008.01.24
WO 2017062970 A1,2017.04.13
审查员 陈晓真

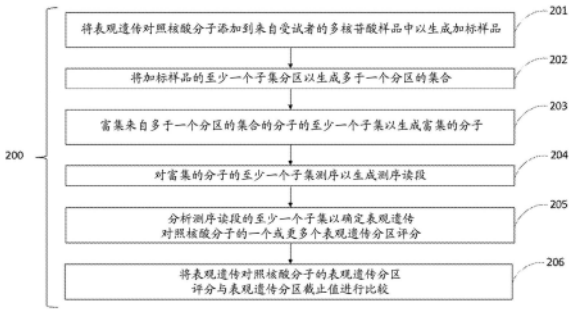
权利要求书5页 说明书52页 附图11页

(54)发明名称

# 用于校准表观遗传分区测定的方法、组合物和系统

(57)摘要

在一方面,一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:(a)向多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合,从而产生加标样品;(b)将加标样品的核酸分子分区成多于一个分区的集合;(c)富集来自多于一个分区的集合的分子的子集以生成富集的分子,其中富集的分子包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子;(d)对富集的分子测序以产生测序读段;(e)分析测序读段以生成表观遗传对照核酸分子的一个或更多个表观遗传分区评分;以及(f)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。




CN 113227393 B

1.一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,所述方法包括:

a.向所述多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合,从而产生加标样品,其中所述表观遗传对照核酸分子的集合是具有完全解析的基因组/表观遗传特征的核酸分子的集合;

b.基于表观遗传修饰的存在或不存在,和/或基于表观遗传修饰的程度将所述加标样品的至少一个子集的核酸分子分区成多于一个分区的集合,其中所述分区包括基于所述核酸分子对结合剂的不同结合亲和力对所述核酸分子进行分区,所述结合剂优先结合包含具有表观遗传修饰的核苷酸的核酸分子;

c.富集来自所述多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中所述富集的分子的集合包括一组表观遗传对照核酸分子和来自所述多核苷酸样品的一组核酸分子;

d.对所述富集的分子的集合的至少一个子集测序以产生测序读段的集合;

e.分析所述测序读段的集合的至少一个子集以生成所述表观遗传对照核酸分子的一个或更多个表观遗传分区评分;

f.将所述一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较;以及

g.将所述方法分类为(i)如果所述表观遗传对照核酸分子的一个或更多个表观遗传分区评分在对应的表观遗传分区截止值内,则所述方法成功;或者(ii)如果所述表观遗传对照核酸分子的一个或更多个表观遗传分区评分中的至少一个在对应的表观遗传分区截止值之外,则所述方法不成功。

2.一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,所述方法包括:

a.向所述多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合,从而产生加标样品,其中所述表观遗传对照核酸分子的集合是具有完全解析的基因组/表观遗传特征的核酸分子的集合;

b.基于表观遗传修饰的存在或不存在,和/或基于表观遗传修饰的程度将所述加标样品的至少一个子集的核酸分子分区成多于一个分区的集合,其中所述分区包括基于所述核酸分子对结合剂的不同结合亲和力对所述核酸分子进行分区,所述结合剂优先结合包含具有表观遗传修饰的核苷酸的核酸分子;

c.富集来自所述多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中所述富集的分子的集合包括一组表观遗传对照核酸分子和来自所述多核苷酸样品的一组核酸分子,其中所述来自所述多核苷酸样品的一组核酸分子包括内源对照分子的集合;

d.对所述富集的分子的集合的至少一个子集测序以产生测序读段的集合;

e.分析所述测序读段的集合的至少一个子集以生成所述表观遗传对照核酸分子和所述内源对照分子的集合的一个或更多个表观遗传分区评分;

f.将所述一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较,以及

g. 将所述方法分类为(i)如果所述表观遗传对照核酸分子的一个或更多个表观遗传分区评分和/或所述内源对照分子的集合的一个或更多个表观遗传分区评分在对应的表观遗传分区截止值内,则所述方法成功;或者(ii)如果所述表观遗传对照核酸分子的一个或更多个表观遗传分区评分和/或所述内源对照分子的集合的一个或更多个表观遗传分区评分中的至少一个在对应的表观遗传分区截止值之外,则所述方法不成功。

3.一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,所述方法包括:

a. 基于表观遗传修饰的存在或不存在,和/或基于表观遗传修饰的程度将来自所述多核苷酸样品的至少一个子集的核酸分子分区成多于一个分区的集合,其中所述分区包括基于所述核酸分子对结合剂的不同结合亲和力对所述核酸分子进行分区,所述结合剂优先结合包含具有表观遗传修饰的核苷酸的核酸分子;

b. 富集来自所述多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中所述富集的分子的集合包括来自所述多核苷酸样品的一组核酸分子,其中所述来自所述多核苷酸样品的一组核酸分子包括内源对照分子的集合;

c. 对所述富集的分子的集合的至少一个子集测序以产生测序读段的集合;

d. 分析所述测序读段的集合的至少一个子集以生成所述内源对照分子的集合的一个或更多个表观遗传分区评分;

e. 将所述一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较,以及

f. 将所述方法分类为(i)如果所述内源对照分子的集合的一个或更多个表观遗传分区评分在对应的表观遗传分区截止值内,则所述方法成功;或者(ii)如果所述内源对照分子的集合的一个或更多个表观遗传分区评分中的至少一个在对应的表观遗传分区截止值之外,则所述方法不成功。

4.根据权利要求1-3中任一项所述的方法,所述方法还包括用标签的集合对所述多于一个分区的集合中的分区的集合中的核酸分子加标签以产生加标签的核酸分子的群体,其中所述加标签的核酸分子包含一个或更多个标签。

5.根据权利要求4所述的方法,其中在所述多于一个分区的集合中的第一分区的集合中使用的标签的集合不同于在所述多于一个分区的集合中的第二分区的集合中使用的标签的集合。

6.根据权利要求5所述的方法,其中所述标签的集合通过将衔接子连接至所述核酸分子而附接至所述核酸分子,其中所述衔接子包含一个或更多个标签。

7.根据权利要求1或2所述的方法,其中所述表观遗传对照核酸分子的集合包含表观遗传对照核酸分子的两个或更多个子集,

其中所述表观遗传对照核酸分子的两个或更多个子集中的子集包含多于一个包含表观遗传修饰区域的表观遗传对照核酸分子。

8.根据权利要求7所述的方法,其中所述表观遗传对照核酸分子还包含标识符区域。

9.根据权利要求8所述的方法,其中所述标识符区域位于所述表观遗传对照核酸分子的表观遗传修饰区域的一侧或两侧。

10.根据权利要求7所述的方法,其中至少一个子集中的表观遗传对照核酸分子的表观

遗传修饰区域包含至少一个具有表观遗传修饰的核苷酸。

11.根据权利要求10所述的方法,其中所述子集包含具有相同数目的带有表观遗传修饰的核苷酸的表观遗传对照核酸分子。

12.根据权利要求10所述的方法,其中第一子集中具有表观遗传修饰的核苷酸的数目不同于第二子集中具有表观遗传修饰的核苷酸的数目。

13.根据权利要求8所述的方法,其中所述表观遗传对照核酸分子的标识符区域包含分子条形码。

14.根据权利要求7所述的方法,其中所述两个或更多个子集中的多于一个表观遗传对照核酸分子的表观遗传修饰区域包含相同的核酸序列。

15.根据权利要求8所述的方法,其中所述标识符区域还包含至少一个表观遗传状态条形码。

16.根据权利要求8所述的方法,其中所述标识符区域包含一个或更多个引物结合位点。

17.根据权利要求7所述的方法,其中第一子集中的多于一个表观遗传对照核酸分子的表观遗传修饰区域包含可与第二子集中的多于一个表观遗传对照核酸分子的表观遗传修饰区域的核酸序列区分的核酸序列。

18.根据权利要求7所述的方法,其中所述表观遗传修饰是DNA甲基化。

19.根据权利要求10所述的方法,其中具有表观遗传修饰的核苷酸包括甲基化核苷酸。

20.根据权利要求19所述的方法,其中所述甲基化核苷酸包括5-甲基胞嘧啶。

21.根据权利要求19所述的方法,其中所述甲基化核苷酸包括5-羟甲基胞嘧啶。

22.根据权利要求7所述的方法,其中表观遗传对照核酸分子的每个子集处于等摩尔浓度。

23.根据权利要求7所述的方法,其中表观遗传对照核酸分子的每个子集处于非等摩尔浓度。

24.根据权利要求19所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为0、1、2、3、4、5、6、7、8、9、10或11。

25.根据权利要求19所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少12。

26.根据权利要求19所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少15。

27.根据权利要求19所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少20。

28.根据权利要求19所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少25。

29.根据权利要求19所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少30。

30.根据权利要求19所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少40。

31.根据权利要求19所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸

分子中甲基化核苷酸的数目为至少50。

32.根据权利要求7所述的方法,其中所述表观遗传对照核酸分子包含对应于λ噬菌体DNA、人类基因组区域或两者的组合的序列。

33.根据权利要求1所述的方法,其中所述表观遗传状态是所述核酸分子的甲基化水平。

34.根据权利要求1或2所述的方法,其中所述多于一个分区的集合包括基于所述核酸分子的甲基化水平分区的所述加标样品的核酸分子。

35. 根据权利要求7所述的方法,其中所述表观遗传对照核酸分子的表观遗传修饰区域包含160 bp的长度。

36.根据权利要求1-3中任一项所述的方法,其中所述富集的分子的集合的测序通过核酸测序仪进行。

37.根据权利要求36所述的方法,其中所述核酸测序仪是下一代测序仪。

38.根据权利要求7所述的方法,其中所述表观遗传对照核酸分子的表观遗传修饰区域包含对应于非人类基因组的核酸序列。

39.根据权利要求1或2所述的方法,其中所述多核苷酸样品选自由DNA样品和RNA样品组成的组。

40.根据权利要求1或2所述的方法,其中所述多核苷酸样品选自由无细胞DNA样品和无细胞RNA样品组成的组。

41.根据权利要求1或2所述的方法,其中所述多核苷酸样品是无细胞DNA。

42.根据权利要求41所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为0、2、4、6、8、10、12或14。

43.根据权利要求41所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少16。

44.根据权利要求41所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少20。

45.根据权利要求41所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少30。

46.根据权利要求41所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少40。

47.根据权利要求41所述的方法,其中至少一个所述子集中的所述表观遗传对照核酸分子中甲基化核苷酸的数目为至少50。

48. 根据权利要求41所述的方法,其中所述无细胞DNA在1 ng和500 ng之间。

49.根据权利要求中1或2所述的方法,其中所述表观遗传对照核酸分子在1飞摩尔和200飞摩尔之间。

50.一种系统,所述系统包括控制器,所述控制器包括计算机可读介质或能够访问计算机可读介质,所述计算机可读介质包括非暂时性计算机可执行指令,当所述非暂时性计算机可执行指令由至少一个电子处理器执行时,至少进行:

a. 获得由核酸测序仪生成的样品的测序读段的集合,其中所述测序读段的集合包括从所述样品的多核苷酸生成的测序读段,并且其中测序之前根据权利要求1的步骤b和c对

所述样品进行分区和富集;

b. 分析所述测序读段的集合的至少一个子集以生成内源对照分子的一个或更多个表观遗传分区评分;和

c. 将所述一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较;以及

d.基于所述表观遗传分区评分的比较来生成分区方法的结果状态,其中所述分区方法的结果状态分类为(i)如果所述内源对照分子的一个或更多个表观遗传分区评分在对应的表观遗传分区截止值内,则所述分区方法成功;或者(ii)如果所述内源对照分子的一个或更多个表观遗传分区评分中的至少一个在对应的表观遗传分区截止值之外,则所述分区方法不成功。

51.根据权利要求50所述的系统,其中

所述样品是加标的样品,并且根据权利要求2的步骤b和c进行分区和富集;

步骤b额外地产生表观遗传对照核酸分子的一个或更多个表观遗传分区评分;并且

步骤d中的结果状态分类为(i)如果所述表观遗传对照核酸分子的一个或更多个表观遗传分区评分和/或所述内源对照分子的一个或更多个表观遗传分区评分在对应的表观遗传分区截止值内,则所述分区方法成功;或者(ii) 如果所述表观遗传对照核酸分子的一个或更多个表观遗传分区评分中的至少一个和/或所述内源对照分子的一个或更多个表观遗传分区评分中的至少一个在对应的表观遗传分区截止值之外,则所述分区方法不成功。

52. 根据权利要求50或51所述的系统,其中所述表观遗传分区评分为*0 CG评分*、*低评分*、*甲基-半*或*甲基-5*。

53.根据权利要求52所述的系统,其中所述表观遗传分区评分为

i) *0 CG评分*,其中所述*0 CG评分*的表观遗传分区截止值为0.01%、0.02%、0.05%、0.1%、0.2%、0.3%、0.4%、0.5%、0.6%、0.7%、0.8%、0.9%、1%、2%或5%;

ii) *低评分*,其中所述*低评分*的表观遗传分区截止值为0.1%、0.5%、1%、2%、3%、4%、5%或7%;

iii) *甲基-半*,其中所述*甲基-半*的表观遗传分区截止值为5个、10个、15个、20个、25个、30个、35个或40个mCG;或

iv)*甲基-5*,其中所述*甲基-5*的表观遗传分区截止值为5个、10个、20个、30个、40个或50个mCG。

54. 根据权利要求52所述的系统,其中所述*0 CG评分*的表观遗传分区截止值为至少5%。

55. 根据权利要求52所述的系统,其中所述*0 CG评分*的表观遗传分区截止值为至少10%。

56.根据权利要求52所述的系统,其中所述*低评分*的表观遗传分区截止值为至少10%。

# 用于校准表观遗传分区测定的方法、组合物和系统

[0001] 相关申请的交叉引用

[0002] 本申请要求2018年10月31日提交的美国临时专利申请第62/753,826号的优先权，其通过引用整体并入本文。

[0003] 背景

[0004] 癌症为全世界疾病的主要原因。每年，在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而死亡。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] 癌症可以由个体的正常细胞内的遗传变异的累积引起，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(indel)，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，癌症也可以根据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无需活检中鉴定可疑的癌细胞的情况下进行。然而，由于体液中核酸的量非常低且所存在的核酸在形式上是异质性(例如，RNA和DNA，单链和双链，以及复制后修饰和与蛋白诸如组蛋白缔合的各种状态)的事实，因此这样的液体活检测试是复杂的。

[0007] 概述

[0008] 在一方面，本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法，该方法包括：a)向多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合，从而产生加标(spiked-in)样品；b)将加标样品的至少一个子集的核酸分子分区成多于一个分区的集合；c)富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合，其中富集的分子的集合包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子；d)对富集的分子的集合的至少一个子集测序以产生测序读段的集合；e)分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子的一个或更多个表观遗传分区评分；以及f)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0009] 在另一方面，本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法，该方法包括：a)向多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合，从而产生加标样品；b)将加标样品的至少一个子集的核酸分子分区成多于一个分区的集合；c)富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合，其中富集的分子的集合包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子，其中来自多核苷酸样品的一组核酸分子包括内源对照分子的集合；d)对富集的分子的集合的至少一个子集测序以产生测序读段的集合；e)分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子的一个或更多个表观遗传分区评分和内源对照分子的集合的一个或更多个表观遗传分区评分；以及f)将一个或更多个表观遗传分区评

分与一个或更多个表观遗传分区截止值进行比较。

[0010] 在另一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:a)将来自多核苷酸样品的至少一个子集的核酸分子分区成多于一个分区的集合;b)富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中富集的分子的集合包括来自多核苷酸样品的一组核酸分子,其中来自多核苷酸样品的一组核酸分子包括内源对照分子的集合;c)对富集的分子的集合的至少一个子集测序以产生测序读段的集合;d)分析测序读段的集合的子集以生成内源对照分子的集合的一个或更多个表观遗传分区评分;以及e)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0011] 在另一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:a)向多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合,从而产生加标样品;b)将加标样品的至少一个子集的核酸分子分区成多于一个分区的集合;c)对分区的分子的至少一个子集测序以产生测序读段的集合;d)分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子的一个或更多个表观遗传分区评分;以及e)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。在一些实施方案中,该方法还包括,在测序步骤之前,富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中富集的分子的集合包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子。

[0012] 在另一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:a)向多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合,从而产生加标样品;b)将加标样品的至少一个子集的核酸分子分区成多于一个分区的集合;c)对分区的分子的至少一个子集测序以产生测序读段的集合;d)分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子的一个或更多个表观遗传分区评分和内源对照分子的集合的一个或更多个表观遗传分区评分;以及e)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。在一些实施方案中,该方法还包括,在测序之前,富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中富集的分子的集合包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子,其中来自多核苷酸样品的一组核酸分子包括内源对照分子的集合。

[0013] 在另一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:a)将来自多核苷酸样品的至少一个子集的分子分区成多于一个分区的集合;b)对富集的分子的集合的至少一个子集测序以产生测序读段的集合;d)分析测序读段的集合的子集以生成内源对照分子的集合的一个或更多个表观遗传分区评分;以及e)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。在一些实施方案中,该方法还包括,在测序之前,富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中富集的分子的集合包括来自多核苷酸样品的一组核酸分子,其中来自多核苷酸样品的一组核酸分子包括内源对照分子的集合。

[0014] 在另一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:a)向多核苷酸样品中的核酸分子添加表观遗传对照

核酸分子的集合,从而产生加标样品;b)将加标样品的至少一个子集的核酸分子分区成多于一个分区的集合;c)富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中富集的分子的集合包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子;以及d)对富集的分子的集合的至少一个子集测序以产生测序读段的集合。在一些实施方案中,该方法还包括,e)分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子的一个或更多个表观遗传分区评分;以及f)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0015] 在另一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:a)向多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合,从而产生加标样品;b)将加标样品的至少一个子集的核酸分子分区成多于一个分区的集合;c)富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中富集的分子的集合包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子,其中来自多核苷酸样品的一组核酸分子包括内源对照分子的集合;以及d)对富集的分子的集合的至少一个子集测序以产生测序读段的集合。在一些实施方案中,该方法还包括,e)分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子的一个或更多个表观遗传分区评分和内源对照分子的集合的一个或更多个表观遗传分区评分;以及f)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0016] 在另一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:a)将来自多核苷酸样品的至少一个子集的核酸分子分区成多于一个分区的集合;b)富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中富集的分子的集合包括来自多核苷酸样品的一组核酸分子,其中来自多核苷酸样品的一组核酸分子包括内源对照分子的集合;以及c)对富集的分子的集合的至少一个子集测序以产生测序读段的集合。在一些实施方案中,该方法还包括,d)分析测序读段的集合的子集以生成内源对照分子的集合的一个或更多个表观遗传分区评分;以及e)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0017] 在一些实施方案中,分析步骤包括估计在至少一个分区的集合中在给定表观遗传状态的表观遗传对照核酸分子和/或内源对照分子的数目/分数。

[0018] 在一些实施方案中,该方法还包括用标签的集合对多于一个分区的集合中的分区的集合中的核酸分子加标签以产生加标签的核酸分子的群体,其中加标签的核酸分子包含一个或更多个标签。在一些实施方案中,在多于一个分区的集合中的第一分区的集合中使用的标签(分子条形码)的集合不同于在多于一个分区的集合中的第二分区的集合中使用的标签(分子条形码)的集合。在一些实施方案中,标签的集合通过将衔接子连接至核酸分子而附接至核酸分子,其中衔接子包含一个或更多个标签(分子条形码)。所使用的标签(分子条形码)序列可以与分区的集合相关联,例如,在一个分区的集合中使用的标签(分子条形码)不在其他分区的集合中使用。

[0019] 在一些实施方案中,该方法还包括g)将分区方法分类为(i)如果表观遗传对照核酸分子的一个或更多个表观遗传分区评分和/或内源对照分子的集合的一个或更多个表观遗传分区评分中的每一个在对应的表观遗传分区截止值内,则该分区方法成功;或者(ii)如果表观遗传对照分子的一个或更多个表观遗传分区评分和/或内源对照分子的集合的一

个或更多个表观遗传分区评分中的至少一个在对应的表观遗传分区截止值之外,则该分区方法不成功。

[0020] 在一些实施方案中,表观遗传对照核酸分子的集合包括表观遗传对照核酸分子的两个或更多个子集,其中表观遗传对照核酸分子的两个或更多个子集中的子集包含多于一个包含表观遗传修饰区域的表观遗传对照核酸分子。

[0021] 在一些实施方案中,多于一个富集分子的测序通过核酸测序仪进行。在一些实施方案中,核酸测序仪是下一代测序仪。

[0022] 在另一方面,本公开内容提供了表观遗传对照核酸分子的集合,包括表观遗传对照核酸分子的两个或更多个子集,其中表观遗传对照核酸分子的两个或更多个子集中的子集包含多于一个包含表观遗传修饰区域的表观遗传对照核酸分子。

[0023] 在另一方面,本公开内容提供了核酸的群体,该核酸的群体包含:(i)表观遗传对照核酸分子的集合,其中表观遗传对照核酸分子的集合包括表观遗传对照核酸分子的两个或更多个子集,其中表观遗传对照核酸分子的两个或更多个子集中的子集包括多于一个包含表观遗传修饰区域的表观遗传对照核酸分子;和(ii)来自受试者的多核苷酸样品中的核酸分子的集合。

[0024] 在一些实施方案中,表观遗传对照核酸分子还包含标识符区域。在一些实施方案中,标识符区域位于表观遗传对照核酸分子的表观遗传修饰区域的一侧或两侧。

[0025] 在一些实施方案中,至少一个子集中的表观遗传对照核酸分子的表观遗传修饰区域包含至少一个具有表观遗传修饰的核苷酸。在一些实施方案中,该子集包含具有相同数目的带有表观遗传修饰的核苷酸的表观遗传对照核酸分子。在一些实施方案中,第一子集中具有表观遗传修饰的核苷酸的数目不同于第二子集中具有表观遗传修饰的核苷酸的数目。在一些实施方案中,具有表观遗传修饰的核苷酸包括甲基化核苷酸。在一些实施方案中,甲基化核苷酸包含5-甲基胞嘧啶。在一些实施方案中,甲基化核苷酸包含5-羟甲基胞嘧啶。

[0026] 在一些实施方案中,表观遗传对照核酸分子的标识符区域包含分子条形码。在一些实施方案中,标识符区域还包含至少一个表观遗传状态条形码。在一些实施方案中,标识符区域包含一个或更多个引物结合位点。

[0027] 在一些实施方案中,两个或更多个子集中的多于一个表观遗传对照核酸分子的表观遗传修饰区域包含相同的核酸序列。

[0028] 在一些实施方案中,第一子集中的多于一个表观遗传对照核酸分子的表观遗传修饰区域包含与第二子集中的多于一个表观遗传对照核酸分子的表观遗传修饰区域的核酸序列可区分的核酸序列。

[0029] 在一些实施方案中,表观遗传修饰是DNA甲基化。

[0030] 在一些实施方案中,表观遗传对照核酸分子的每个子集处于等摩尔浓度。在一些实施方案中,表观遗传对照核酸分子的每个子集处于非等摩尔浓度。

[0031] 在一些实施方案中,至少一个子集中的表观遗传对照核酸分子中甲基化核苷酸的数目为0、1、2、3、4、5、6、7、8、9、10、11、至少12、至少15、至少20、至少25、至少30、至少40或至少50。

[0032] 在一些实施方案中,表观遗传对照核酸分子包含对应于λ噬菌体DNA、人类基因组

区域或两者的组合的序列。

[0033] 在一些实施方案中,表观遗传状态是核酸分子的甲基化水平。在一些实施方案中,多于一个分区的集合包括基于核酸分子的甲基化水平分区的加标样品的核酸分子。

[0034] 在一些实施方案中,表观遗传对照核酸分子的表观遗传修饰区域包含约160bp的长度。

[0035] 在一些实施方案中,表观遗传对照核酸分子的表观遗传修饰区域包含对应于非人类基因组的核酸序列。

[0036] 在一些实施方案中,多核苷酸样品选自由DNA样品、RNA样品、多核苷酸样品、无细胞DNA样品和无细胞RNA样品组成的组。在一些实施方案中,多核苷酸样品选自由DNA样品、RNA样品、多核苷酸样品、无细胞DNA样品和无细胞RNA样品组成的组。在一些实施方案中,无细胞DNA在1ng和500ng之间。

[0037] 在一些实施方案中,表观遗传对照核酸分子在1飞摩尔和200飞摩尔之间。

[0038] 在一些实施方案中,分区包括基于核酸分子对结合剂的不同结合亲和力对核酸分子进行分区,所述结合剂优先结合包含具有表观遗传修饰的核苷酸的核酸分子。

[0039] 在另一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区方法的系统,该系统包括:通信接口,该通信接口通过通信网络接收由核酸测序仪生成的加标样品的测序读段的集合,其中测序读段的集合包括(i)从来源于样品的多核苷酸生成的测序读段的至少第一群体,其中来自第一群体的测序读段包括标签序列和衍生自来源于样品的多核苷酸的序列;和(ii)从表观遗传对照核酸分子生成的测序读段的至少第二群体,其中从第二群体生成的测序读段包含表观遗传修饰区域和任选地标识符区域;计算机,该计算机与通信接口通信,其中该计算机包括一个或更多个计算机处理器和计算机可读介质,该计算机可读介质包括机器可执行代码,该机器可执行代码在由一个或更多个计算机处理器执行后实现一种方法,该方法包括:通过通信网络接收来自核酸测序仪的测序读段的第一群体和第二群体的测序读段的集合;分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子和/或内源对照分子的一个或更多个表观遗传分区评分;以及将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0040] 在另一方面,本公开内容提供了一种系统,该系统包括控制器,该控制器包括计算机可读介质或能够访问计算机可读介质,该计算机可读介质包括非暂时性计算机可执行指令,当所述非暂时性计算机可执行指令由至少一个电子处理器执行时,至少进行:(a)获得由核酸测序仪生成的加标样品的测序读段的集合,其中加标样品包含样品多核苷酸和表观遗传对照核酸分子,并且测序读段的集合包含(i)从样品多核苷酸生成的测序读段的第一群体和(ii)从表观遗传对照核酸分子生成的测序读段的第二群体;(b)分析测序读段的集合的至少一个子集,以生成表观遗传对照核酸分子和/或内源对照分子的一个或更多个表观遗传分区评分;和(c)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0041] 在另一方面,本公开内容提供了一种系统,该系统包括控制器,该控制器包括计算机可读介质或能够访问计算机可读介质,该计算机可读介质包括非暂时性计算机可执行指令,当所述非暂时性计算机可执行指令由至少一个电子处理器执行时,至少进行:(a)获得由核酸测序仪生成的样品的测序读段的集合,其中测序读段的集合包括从样品多核苷酸生

成的测序读段;(b)分析测序读段的集合的至少一个子集以生成内源对照分子的一个或更多个表观遗传分区评分;和(c)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0042] 在一些实施方案中,该系统还包括g)基于表观遗传分区评分的比较来生成分区方法的结果状态。在一些实施方案中,分区方法的结果状态分类为(i)如果表观遗传对照核酸分子的一个或更多个表观遗传分区评分和/或内源对照分子的集合的一个或更多个表观遗传分区评分在对应的表观遗传分区截止值内,则该分区方法成功;或者(ii)如果表观遗传对照分子和/或内源对照分子的一个或更多个表观遗传分区评分中的至少一个在对应的表观遗传分区截止值之外,则该分区方法不成功。

[0043] 在一些实施方案中,表观遗传分区评分包括分区的集合中高甲基化(hypermethylated)表观遗传对照核酸分子和/或高甲基化对照分子的数目的分数或百分比。在一些实施方案中,表观遗传分区评分包括分区的集合中低甲基化(hypomethylated)表观遗传对照核酸分子和/或低甲基化对照分子的数目的分数或百分比。在一些实施方案中,分区的集合是高甲基化分区的集合。在一些实施方案中,分区的集合是低甲基化分区的集合。在一些实施方案中,表观遗传分区评分为OCG评分(OCG score)。在一些实施方案中,表观遗传分区评分为低评分(hypo score)。在一些实施方案中,表观遗传分区评分为甲基-半(methyl-half)。在一些实施方案中,表观遗传分区评分为甲基-5(methyl-5)。

[0044] 在一些实施方案中,OCG评分的表观遗传分区截止值为0.01%、0.02%、0.05%、0.1%、0.2%、0.3%、0.4%、0.5%、0.6%、0.7%、0.8%、0.9%、1%、2%、5%、至少5%或至少10%。在一些实施方案中,低评分的表观遗传分区截止值为0.1%、0.5%、1%、2%、3%、4%、5%、7%或至少10%。在一些实施方案中,甲基-半的表观遗传分区截止值为5个、10个、15个、20个、25个、30个、35个或40个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值为5个、10个、20个、30个、40个或50个mCG。

[0045] 根据以下详细描述,本公开内容的另外的方面和优点对本领域技术人员而言将变得明显,详细描述中仅示出和描述了本公开内容的说明性实施方案。如将意识到的,本公开内容能够具有其他和不同的实施方案,并且其若干细节能够在各种明显的方面进行修改,所有这些都不偏离本公开内容。相应地,附图和描述应被认为本质上是说明性的而非限制性的。

[0046] 在一些实施方案中,本文公开的系统和/或方法的结果被用作输入以生成报告。报告可以是纸质或电子格式。例如,由本文公开的方法或系统所确定的关于核酸分子的分区的信息和/或源自核酸分子的分区的信息可以展示在这样的报告中。本文公开的方法或系统还可以包括将报告传送给第三方的步骤,第三方诸如是样品来源的受试者或健康护理从业者。

[0047] 本文公开的方法的各步骤,或由本文公开的系统进行的步骤,可以在相同时间或不同的时间和/或在同一地理位置或不同的地理位置例如国家进行。本文公开的方法的各步骤可以由同一人员或不同的人员进行。

[0048] 附图简述

[0049] 并入本说明书并构成其一部分的附图示出了某些实施方案,并与书面描述一起用于解释本文公开的方法、计算机可读介质和系统的某些原理。当结合附图阅读时,本文提供

的描述被更好地理解,附图以实例的方式而非限制的方式被包括在内。应当理解,除非上下文另有说明,否则在所有附图中,相同的附图标记表示相同的部件。还应当理解,一些或所有附图可以是出于说明目的的示意图,并不一定描绘所示元件的实际相对尺寸或位置。

[0050] 图1A和图1B是双链DNA中的完全甲基化(图1A)和半甲基化(图1B)CpG二联体(CpG dyad)的示意图。

[0051] 图2是根据本公开内容的一种实施方案的用于评估多核苷酸样品的分区的方法的流程图表示。

[0052] 图3是根据本公开内容的一种实施方案的用于评估多核苷酸样品的分区的方法的流程图表示。

[0053] 图4是根据本公开内容的一种实施方案的用于评估多核苷酸样品的分区的方法的流程图表示。

[0054] 图5是适用于本公开内容的一些实施方案的表观遗传对照核酸分子的示意图。

[0055] 图6是适用于本公开内容的一些实施方案的表观遗传对照核酸分子的示意图。

[0056] 图7是适用于本公开内容的一些实施方案的表观遗传对照核酸分子的示意图。

[0057] 图8是适用于本公开内容的一些实施方案的系统的实例的示意图。

[0058] 图9A、图9B和图9C是属于高分区的集合(图9A)、中等分区的集合(图9B)和低分区的集合(图9C)中的子集1、子集2、子集3、子集4、子集5和子集6的表观遗传对照核酸分子的表观遗传分区评分的图形表示。

[0059] 图10A和图10B是高分区的集合(图10A)和低分区的集合(图10B)中样品1的高甲基化对照分子的分数的图形表示。

[0060] 图11A和图11B是高分区的集合(图11A)和低分区的集合(图11B)中样品2的高甲基化对照分子的分数的图形表示。

[0061] 定义

[0062] 为了更容易地理解本公开内容,以下首先定义某些术语。以下术语和其他术语的另外定义可通过本说明书进行阐述。如果以下阐述的术语的定义与通过引用并入的申请或专利中的定义不一致,则本申请中阐述的定义应该用于理解该术语的含义。

[0063] 除非上下文另外清楚地指明,否则如本说明书和所附的权利要求书中使用的单数形式“一(a)”、“一(an)”和“所述/该(the)”包括复数指示物。因此,例如,提及“一种(a)方法”包括一种或更多种本文描述的类型的方法和/或步骤,和/或其对本领域普通技术人员而言在阅读本公开内容等后将变得明显。

[0064] 还应当理解的是,本文使用的术语仅为了描述特定实施方案的目的而非意图是限制性的。此外,除非另外定义,否则本文使用的所有技术术语和科学术语具有与本公开内容所属的领域的普通技术人员通常理解的相同含义。在描述和要求保护方法、计算机可读介质和系统方面,将根据下文阐述的定义使用以下术语及其语法变体。

[0065] 约:如本文使用的,应用于一个或更多个感兴趣的值或要素的“约”或“大约”是指与所述参考值或要素相似的值或要素。在某些实施方案中,术语“约”或“大约”是指在所述参考值或要素的任一方向上(大于或小于)落在所述参考值或要素的25%、20%、19%、18%、17%、16%、15%、14%、13%、12%、11%、10%、9%、8%、7%、6%、5%、4%、3%、2%、1%或更小以内的一系列值或要素,除非另有说明或从上下文中明显(除非这样的数字超过

可能的值或要素的100％)。

[0066] 衔接子:如本文使用的,“衔接子”是指通常是至少部分双链的并附接至给定样品核酸分子的任一末端或两个末端的短核酸(例如,长度小于约500个核苷酸、小于约100个核苷酸或小于约50个核苷酸)。衔接子可以包含核酸引物结合位点和/或测序引物结合位点,核酸引物结合位点允许两个末端处侧接衔接子的核酸分子的扩增,测序引物结合位点包括用于测序应用(诸如各种下一代测序(NGS)应用)的引物结合位点。衔接子还可以包含捕获探针(诸如附接至流动池支持物的寡核苷酸等)的结合位点。衔接子也可以包含本文所述的核酸标签。核酸标签通常相对于扩增引物结合位点和测序引物结合位点定位,使得核酸标签被包含在给定核酸分子的扩增子和序列读段中。相同或不同序列的衔接子可以连接至核酸分子的相应末端。在一些实施方案中,除了核酸标签不同的相同序列的衔接子被连接至核酸分子的相应末端。在一些实施方案中,衔接子是Y形衔接子,其中一个末端如本文所述是平末端或加尾的,用于连接也是平末端的或用一个或更多个互补核苷酸加尾的核酸分子。在又其他的示例性实施方案中,衔接子是钟形衔接子,其包含用于连接至待分析的核酸分子的平末端或加尾末端。衔接子的其他实例包括加T尾和加C尾的衔接子。

[0067] 扩增:如本文在核酸的上下文中使用的“扩增(amplify)”或“扩增(amplification)是指通常从少量的多核苷酸(例如,单个多核苷酸分子)开始产生多个拷贝的该多核苷酸或该多核苷酸的一部分,其中扩增产物或扩增子通常是可检测的。多核苷酸的扩增包括多种化学和酶促过程。扩增包括但不限于聚合酶链式反应(PCR)。

[0068] 条形码:如本文在核酸的上下文中使用的“条形码”或“分子条形码”是指包含可用作分子标识符的序列的核酸分子。例如,在下一代测序(NGS)文库制备期间,通常将个体“条形码”序列添加至DNA片段,以便在最终数据分析之前可以鉴定和分选测序读段。

[0069] 癌症类型:如本文使用的,“癌症类型”是指由例如组织病理学定义的癌症的类型或亚型。癌症类型可以通过任何常规标准来定义,诸如基于在特定组织中的发生(例如,血癌、中枢神经系统(CNS)癌、脑癌、肺癌(小细胞和非小细胞)、皮肤癌、鼻癌、喉癌、肝癌、骨癌、淋巴瘤、胰腺癌、肠癌、直肠癌、甲状腺癌、膀胱癌、肾癌、口癌、胃癌、乳腺癌、前列腺癌、卵巢癌、肺癌、小肠癌、软组织癌、神经内分泌癌、胃食管癌、头颈癌、妇科癌症、结肠直肠癌、尿路上皮癌、固态癌(solid state cancer)、异质性癌症(heterogeneous cancer)、同质性癌症(homogeneous cancer)),可以是原发性来源未知的或类似的,和/或可以具有相同细胞谱系(例如,癌、肉瘤、淋巴瘤、胆管癌、白血病、间皮瘤、黑素瘤或成胶质细胞瘤)和/或可以是显示出癌症标志物(诸如,但不限于Her2、CA15-3、CA19-9、CA-125、CEA、AFP、PSA、HCG、激素受体和NMP-22)的癌症。癌症也可以根据阶段(例如,阶段1、阶段2、阶段3或阶段4)以及是原发性还是继发性来分类。

[0070] 无细胞核酸:如本文使用的,“无细胞核酸”是指不包含在细胞内或本来不与细胞结合的核酸,或者在一些实施方案中,是指去除完整细胞后保留在样品中的核酸。无细胞核酸可以包括,例如,来源于来自受试者的体液(例如,血液、血浆、血清、尿液、脑脊髓液(CSF)等)的所有未被包封的核酸。无细胞核酸包括DNA(cfDNA)、RNA(cfRNA)及其杂合体,包括基因组DNA、线粒体DNA、循环DNA、siRNA、miRNA、循环RNA(cRNA)、tRNA、rRNA、小核仁RNA(snoRNA)、Piwi相互作用RNA(piRNA)、长非编码RNA(长ncRNA)和/或这些中的任一种的片段。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、凋亡等)释放到体液中。一些无细胞核酸是从癌细胞释放到体液中的,例如循环肿瘤DNA(ctDNA)。其他的是从健康细胞中释放的。ctDNA可以是未被包封的肿瘤来源的片段化DNA。无细胞核酸可以具有一种或更多种表观遗传修饰,例如,无细胞核酸可以被乙酰化、5-甲基化和/或羟基甲基化。

[0071] 细胞核酸:如本文使用的,“细胞核酸”是指至少在从受试者获取或收集样品时处于产生核酸的一个或更多个细胞内的核酸,即使作为特定分析过程的一部分,这些核酸随后被取出(例如,经由细胞裂解)。

[0072] 覆盖率:如本文使用的,术语“覆盖率”、“总分子计数”或“总等位基因计数”可互换使用。它们是指给定样品中在特定基因组位置处的DNA分子的总数。

[0073] CpG二联体:如本文使用的,术语“CpG二联体”是指双链DNA分子的有义链上的二核苷酸CpG(胞嘧啶-磷酸-鸟嘌呤(即,在核酸序列的5’→3’方向上,胞嘧啶后跟鸟嘌呤))及反义链上的其互补CpG(如图1所示)。

[0074] 脱氧核糖核酸或核糖核酸:如本文使用的,“脱氧核糖核酸“或“DNA”是指在糖部分的2’-位置处具有氢基团的天然或修饰的核苷酸。DNA通常包括包含以下四种类型的核苷酸碱基的核苷酸链:腺嘌呤(A)、胸腺嘧啶(T)、胞嘧啶(C)和鸟嘌呤(G)。如本文使用的,“核糖核酸”或”RNA”是指在糖部分的2’-位置处具有羟基基团的天然或修饰的核苷酸。RNA通常包括包含以下四种类型的核苷酸碱基的核苷酸链:A、尿嘧啶(U)、G和C。如本文使用的,术语“核苷酸”是指天然核苷酸或修饰的核苷酸。某些核苷酸对以互补方式彼此特异性结合(被称为互补碱基配对)。在DNA中,腺嘌呤(A)与胸腺嘧啶(T)配对并且胞嘧啶(C)与鸟嘌呤(G)配对。在RNA中,腺嘌呤(A)与尿嘧啶(U)配对并且胞嘧啶(C)与鸟嘌呤(G)配对。当第一条核酸链与由与第一条链的核苷酸互补的核苷酸构成的第二条核酸链结合时,两条链结合形成双链。如本文使用的,“核酸测序数据”、“核酸测序信息”、“序列信息”、“核酸序列”、“核苷酸序列”、“基因组序列”、“基因序列”、或“片段序列”、或“核酸测序读段”表示指示核酸诸如DNA或RNA的分子(例如,全基因组、全转录组、外显子组、寡核苷酸、多核苷酸或片段)中核苷酸碱基(例如,腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶或尿嘧啶)的顺序和身份的任何信息或数据。应当理解,本教导设想了使用所有可用的各种技术(technique)、平台或技术(technology)获得的序列信息,包括但不限于:毛细管电泳、微阵列、基于连接的系统、基于聚合酶的系统、基于杂交的系统、直接或间接的核苷酸鉴定系统、焦磷酸测序、基于离子或pH的检测系统以及基于电子信号的系统。

[0075] 内源对照分子:如本文所用,“内源对照分子”是指多核苷酸样品中对应于至少一个具有非可变表观遗传状态的人类基因组区域的核酸分子。在一些实施方案中,内源对照分子可以跨组织、受试者和癌症是一致地高度甲基化或低度甲基化的。在一些实施方案中,对应于具有一致地高度甲基化区域的人类基因组区域的内源对照分子可以被称为“高甲基化对照分子”。在一些实施方案中,对应于具有一致地低度甲基化区域的人类基因组区域的内源对照分子可以被称为“低甲基化对照分子”。

[0076] 表观遗传对照核酸分子:如本文使用的,“表观遗传对照核酸分子”是指添加到多核苷酸样品中以基于表观遗传修饰评价样品的分区的一组核酸分子。例如,表观遗传修饰可以是DNA甲基化,并且表观遗传对照核酸分子可以具有不同/可区分的甲基化水平。在一些实施方案中,表观遗传对照核酸分子包含表观遗传修饰区域和任选地标识符区域。在一

些实施方案中,表观遗传对照核酸分子包含表观遗传修饰区域和标识符区域。表观遗传对照核酸分子可以是合成的寡核苷酸。在一些实施方案中,表观遗传对照核酸分子可以具有非天然存在的核酸序列。在一些实施方案中,表观遗传对照核酸分子可以具有天然存在的核酸序列。在一些实施方案中,表观遗传对照核酸分子可以具有对应于非人类基因组的核酸序列。作为非限制性的实例,这些分子可以具有(i)对应于λ噬菌体DNA或人类基因组的区域的序列,(ii)非天然存在的序列,和/或(iii)(i)和(ii)的组合。

[0077] 表观遗传修饰:如本文使用的,"表观遗传修饰"是指核酸分子中核苷酸的碱基的修饰。修饰可以是核苷酸的碱基的化学修饰。在一些情况下,修饰可以是核苷酸的碱基的甲基化。例如,修饰可以是胞嘧啶的甲基化,产生5-甲基胞嘧啶。

[0078] 表观遗传修饰区域:如本文使用的,"表观遗传修饰区域"是指表观遗传对照核酸分子中代表表观遗传对照核酸分子的表观遗传修饰水平/程度的区域。在一些实施方案中,表观遗传修饰区域可以包含具有表观遗传修饰的核苷酸。在一些实施方案中,表观遗传修饰是DNA甲基化。在这些实施方案中,表观遗传对照核酸分子的表观遗传修饰区域可以具有甲基化的核苷酸。表观遗传修饰区域中甲基化核苷酸的数目在表观遗传对照核酸分子之间可以不同。在一些实施方案中,表观遗传对照核酸分子在表观遗传修饰区域中可以具有0个、1个、2个、3个、4个、5个、6个、7个、8个、9个、至少10个、至少15个、至少20个、至少30个、至少40个或至少50个甲基化核苷酸。表观遗传对照核酸分子可以根据表观遗传修饰区域中具有表观遗传修饰的核苷酸的数目分为子集。不同子集之间的表观遗传修饰区域可以具有相同的长度,例如约160bp。子集之间表观遗传修饰区域的长度可以不同。例如,表观遗传对照核酸分子可以根据表观遗传修饰区域中甲基化核苷酸的数目分为三个子集(子集A、子集B和子集C)。子集A、子集B和子集C可以具有在表观遗传修饰区域分别带有5个、10个和15个甲基化核苷酸的表观遗传对照核酸分子,并且子集A、子集B和子集C中表观遗传修饰区域的长度可以相同(例如160bp)或者可以不同—子集A、子集B和子集C分别为100bp、150bp和200bp。

[0079] 表观遗传分区评分:如本文使用的,"表观遗传分区评分"是指代表属于给定的分区的集合中特定表观遗传状态的核酸分子的分区的数值。在一些实施方案中,属于一种表观遗传状态的核酸分子的表观遗传分区评分是为每个分区的集合确定的。例如,可以确定属于特定表观遗传状态的表观遗传对照核酸分子和/或内源对照分子的表观遗传分区评分。表观遗传分区评分可以是属于特定表观遗传状态的核酸分子的数目(或统计学上估计的数目)的量度。表观遗传分区评分可以用分数或百分比来表示。表观遗传分区评分可以是属于被分区到至少一个分区的集合中的特定表观遗传状态的表观遗传对照核酸分子和/或内源对照分子的数目与属于其他剩余分区的集合中存在的该表观遗传状态的表观遗传对照核酸分子和/或内源对照分子的数目之比的量度。在一些实施方案中,表观遗传分区评分可以是属于被分区到至少一个分区的集合中的特定表观遗传状态的表观遗传对照核酸分子和/或内源对照分子的数目与所有分区的集合中属于该表观遗传状态的表观遗传对照核酸分子和/或内源对照分子的总数目之比的分数或百分比。在一些实施方案中,表观遗传分区评分是为每个分区的集合中的表观遗传对照核酸分子和/或内源对照分子的每种表观遗传状态确定的。在一些实施方案中,表观遗传分区评分是为一个或更多个分区的集合中具有一种或更多种特定表观遗传状态的表观遗传对照核酸分子和/或内源对照分子确定的。

在一些实施方案中,表观遗传分区评分是为特定分区的集合中具有特定表观遗传状态的表观遗传对照核酸分子和/或内源对照分子确定的。

[0080] 在一些实施方案中,表观遗传分区评分可以针对无CG('零'CG)的分子被分区到高分区的集合的效率。该评分可称为0CG评分。在一些实施方案中,可以用高分区的集合中无CG的分子的分数或百分比来表示0CG评分。在一些实施方案中,表观遗传分区评分可以是高分区的集合中具有以下至少一种的表观遗传对照核酸分子的分数和/或高甲基化对照分子的分数的量度:

[0081] (i)1个甲基CG(表观遗传分区评分可称为1CG评分),

[0082] (ii)2个甲基CG(表观遗传分区评分可称为2CG评分),

[0083] (iii)3个甲基CG(表观遗传分区评分可称为3CG评分),

[0084] (iv)4个甲基CG(表观遗传分区评分可称为4CG评分)和

[0085] (v)5个甲基CG(表观遗传分区评分可称为5CG评分)。

[0086] 在一些实施方案中,表观遗传分区评分可以针对低甲基化对照分子或低甲基化表观遗传对照核酸分子被分区到高甲基化分区的集合的效率。该评分可称为低评分。在一些实施方案中,低评分可以用高甲基化分区的集合中低甲基化对照分子或低甲基化表观遗传对照核酸分子的分数或百分比来表示。在一些实施方案中,表观遗传分区评分可以是低甲基化分区的集合中少于指定值例如5%的高甲基化对照分子和/或高甲基化表观遗传对照核酸分子所需的甲基化CG数目的量度。在低甲基化分区的集合中使用5%的高甲基化对照分子和/或高甲基化表观遗传对照核酸分子的实例中,即,表观遗传分区评分是低甲基化分区的集合中少于5%的高甲基化对照分子和/或高甲基化表观遗传对照核酸分子所需的甲基化CG数目的量度,为了方便起见,该评分可以被称为甲基-5。在一些实施方案中,表观遗传分区评分可以是高甲基化分区的集合中至少指定值例如50%的高甲基化对照分子和/或高甲基化表观遗传对照核酸分子所需的甲基化CG的数目的量度。在高甲基化分区的集合中使用50%的高甲基化对照分子和/或高甲基化表观遗传对照核酸分子的实例中,即,表观遗传分区评分是高甲基化分区的集合中至少50%的高甲基化对照分子和/或高甲基化表观遗传对照核酸分子所需的甲基化CG数目的量度,该评分可以被称为甲基-半。在不同的实施方案中,可以使用从0%到100%(不仅仅是5%和50%)的宽范围的不同值,并且可以使用涉及该指定值的对应的不同的方便名称。

[0087] 例如,使用表观遗传对照核酸分子的三个子集(子集A、子集B和子集C),并且每个子集的甲基化核苷酸数目不同。这三个子集中的表观遗传对照核酸分子可以根据它们与甲基结合蛋白的结合亲和力分为三个分区的集合—P1、P2和P3。对于每个子集,表观遗传分区评分是为每个分区的集合(P1、P2和P3)确定的,即属于子集A的表观遗传对照核酸分子将具有三个表观遗传分区评分—三个分区的集合P1、P2和P3各一个。同样,子集B和子集C每一个将具有三个表观遗传分区评分—三个分区的集合P1、P2和P3各一个。表观遗传分区评分也可以为内源对照分子确定。

[0088] 在另一种实施方案中,使用表观遗传对照核酸分子的三个子集(子集A、子集B和子集C),并且每个子集的甲基化核苷酸数目不同(即,每个子集具有不同的表观遗传状态)。这三个子集中的表观遗传对照核酸分子可以根据它们与甲基结合蛋白的结合亲和力分为三个分区的集合—P1、P2和P3。在该实施方案中,表观遗传评分仅为P1分区的集合中的子集A

分子确定。该表观遗传评分可以是P1分区的集合中的子集A分子相对于(在P1、P2和P3分区的集合中)子集A分子总数的分数或百分比的量度。

[0089] 表观遗传分区截止值:如本文使用的,“表观遗传分区截止值”是指用于评价特定分区的集合中属于特定表观遗传状态的核酸分子的分区的预定截止值或截止范围。在一些实施方案中,表观遗传分区截止值通过分析内部样品数据集(in-house sample dataset)来确定。每个分区的集合可以具有属于一种表观遗传状态的核酸分子的表观遗传分区截止值。如果属于一种或更多种表观遗传状态的表观遗传对照核酸分子的一个或更多个表观遗传分区评分(用于评价分区)在其对应的表观遗传分区截止值内,则分区方法是成功的。否则,分区方法是失败的。表观遗传分区截止值随核酸分子和分区的集合的表观遗传状态而不同,即,每种表观遗传状态将具有其自己的表观遗传分区截止值,并且每个分区的集合具有该表观遗传状态的单独的表观遗传分区截止值。截止值可以是百分比或评分,并且截止值可以是截止范围,而不是特定的截止值。例如,对于分区的集合P1、P2和P3,属于特定表观遗传状态的表观遗传对照核酸分子的表观遗传分区截止值可以分别在70%-79%之间、10%-15%之间和小于5%。如果属于该表观遗传状态的表观遗传对照核酸分子的表观遗传分区评分在对应的表观遗传分区截止值内,则分区方法是成功的。

[0090] 表观遗传状态:如本文使用的,“表观遗传状态”是指核酸分子的表观遗传修饰的水平/程度。例如,如果表观遗传修饰是DNA甲基化(或羟基甲基化),那么表观遗传状态可以指DNA碱基(例如胞嘧啶)上甲基化的存在或不存在,或者指核酸序列中甲基化的程度(例如,高度甲基化、低度甲基化、中等甲基化或未甲基化的核酸分子)。表观遗传状态也可以指带有表观遗传修饰的核苷酸的数目。例如,如果表观遗传修饰是DNA甲基化,那么表观遗传状态可以指核酸分子的甲基化核苷酸的数目。

[0091] 表观遗传状态条形码:如本文使用的,“表观遗传状态条形码”是指用于标识表观遗传对照核酸分子的表观遗传状态的核酸序列。标识可以通过使一个或多于一个特定表观遗传状态条形码和表观遗传对照核酸分子的表观遗传状态之间具有预定的相关性来实现。它可以指表观遗传对照核酸分子的表观遗传修饰区域中具有表观遗传修饰的核苷酸的数目。在一些实施方案中,表观遗传对照核酸分子的标识符区域包含至少一个表观遗传状态条形码。例如,如果表观遗传修饰是DNA甲基化,并且表观遗传对照核酸分子的一个子集具有5个甲基化核苷酸,那么该子集内的所有表观遗传对照核酸分子将具有相同的表观遗传状态条形码。在一些实施方案中,表观遗传状态条形码可用于标识表观遗传对照核酸分子的表观遗传修饰区域的表观遗传修饰水平/程度。表观遗传对照核酸分子可以根据表观遗传修饰区域中胞嘧啶或CpG核苷酸的数目分为子集。在一些实施方案中,在每个子集中,甲基化水平可以不同(例如,高度甲基化、中等甲基化、低度甲基化或未甲基化),并且每个甲基化水平可以具有单独的表观遗传状态条形码。例如,在子集A中,所有低度甲基化的表观遗传对照核酸分子将具有一种表观遗传状态条形码,例如ESB1,并且所有高度甲基化的表观遗传对照核酸分子将具有另一种表观遗传状态条形码,例如ESB3。在该实例中,表观遗传状态条形码用于标识甲基化的水平/程度。

[0092] 具有非可变表观遗传状态的人类基因组区域:如本文使用的,“具有非可变表观遗传状态的人类基因组区域”是指人类基因组中具有特定表观遗传状态的区域,并且该区域的表观遗传状态通常不变化/改变,并且总是保持相同或与不同受试者和/或不同类型的疾

病/疾病阶段保持一致。例如,具有非可变表观遗传状态的人类基因组区域可以主要是甲基化的或主要是未甲基化的。

[0093] 标识符区域:如本文使用的,“标识符区域”是指表观遗传对照核酸分子中用于区分一种表观遗传对照核酸分子与其他表观遗传对照核酸分子的区域。标识符区域可以具有分子条形码和/或表观遗传状态条形码。标识符区域可以存在于表观遗传修饰区域的一侧或两侧。分子条形码用作表观遗传对照核酸分子的标识符,而表观遗传状态条形码用作表观遗传对照核酸分子的表观遗传状态的标识符。标识符区域可以具有促进一个或更多个引物结合的另外区域(引物结合位点)。

[0094] 突变等位基因计数:如本文使用的,术语“突变等位基因计数”是指在特定基因组基因座处含有突变等位基因的DNA分子的数目。

[0095] 突变等位基因分数:如本文使用的,“突变等位基因分数”、“突变剂量”或“MAF”是指在给定样品中在给定基因组位置/基因座处含有等位基因改变或突变的核酸分子的分数。MAF通常表示为分数或百分比。例如,体细胞变异的MAF可以小于0.15。

[0096] 突变:如本文使用的,“突变”是指从已知的参考序列的变异,并且包括突变诸如,例如,单核苷酸变异(SNV)和插入或缺失(indel)。突变可以是种系突变或体细胞突变。在一些实施方案中,用于比较目的的参考序列是提供测试样品的受试者的物种的野生型基因组序列,通常是人类基因组。

[0097] 突变调用器(mutation caller):如本文使用的,“突变调用器”是指用于鉴定测试样品数据(例如,从受试者获得的序列信息)中的突变的算法(通常以软件或计算机实现的方式实现)。

[0098] 赘生物:如本文使用的,术语“赘生物”和“肿瘤”可互换使用。它们是指受试者的细胞的异常生长。赘生物或肿瘤可以是良性的、潜在恶性的或恶性的。恶性肿瘤是指癌症或癌性肿瘤。

[0099] 下一代测序:如本文使用的,“下一代测序”或“NGS”是指与传统的基于Sanger和毛细管电泳的方法相比具有增加的通量的测序技术,例如,具有一次产生数十万个相对较小的序列读段的能力。下一代测序技术的一些实例包括但不限于合成测序、连接测序和杂交测序。在一些实施方案中,下一代测序包括使用能够对单个分子进行测序的仪器。

[0100] 核酸标签:如本文使用的,“核酸标签”是指短核酸(例如,长度小于约500个核苷酸、约100个核苷酸、约50个核苷酸或约10个核苷酸),用于区分来自不同样品(例如,呈现为样品索引(sample index))的核酸或同一样品中不同类型的或经历不同处理的不同核酸分子(例如,呈现为分子条形码)。核酸标签包含预定的、固定的、非随机的、随机的或半随机的寡核苷酸序列。这样的核酸标签可用于标记不同的核酸分子或不同的核酸样品或子样品。核酸标签可以是单链的、双链的或至少部分双链的。核酸标签任选地具有相同的长度或不同的长度。核酸标签还可以包括具有一个或更多个平末端的双链分子,包括5’或3’单链区域(例如,突出端),和/或包括在特定分子内的其他位置处的一个或更多个其他单链区域。核酸标签可以被附接至其他核酸(例如,待被扩增和/或测序的样品核酸)的一端或两端。核酸标签可以被解码以揭示诸如特定核酸的样品来源、形式或加工的信息。例如,核酸标签也可以用于使包含带有不同分子条形码和/或样品索引的核酸的多个样品的汇集和/或并行处理成为可能,其中核酸随后通过检测(例如,读取)核酸标签被解卷积。核酸标签也可以被

称为标识符(例如分子标识符、样品标识符)。另外或可选地,核酸标签可以被用作分子标识符(例如,用于区分同一样品或子样品中的不同分子或不同亲本分子的扩增子)。这包括,例如,对给定样品中的不同的核酸分子独特地加标签,或对这样的分子非独特地加标签。在非独特加标签应用的情况下,可以使用有限数目的标签(即分子条形码)对每个核酸分子加标签,使得不同分子可以基于其内源序列信息(例如,其映射至所选择的参考基因组的起始和/或终止位置、序列的一端或两端的子序列和/或序列的长度)与至少一个分子条形码的组合而被区分。通常,使用足够数目的不同分子条形码,使得任何两个分子可具有相同的内源序列信息(例如,起始和/或终止位置、序列的一个或两个末端的子序列和/或长度)以及还具有相同的分子条形码的概率较低(例如,小于约10%、小于约5%、小于约1%或小于约0.1%的概率)。

[0101] 分区:如本文使用的,"分区"和"表观遗传分区"可互换使用。它是指基于核酸分子的特征(例如表观遗传修饰的水平/程度)将核酸分子分离或分级。分区可以是分子的物理分区。分区可以包括基于表观遗传修饰的水平(即表观遗传状态)将核酸分子分成组或集合。例如,核酸分子可以基于核酸分子的甲基化水平来分区。在一些实施方案中,用于分区的方法和系统可见于PCT专利申请第PCT/US2017/068329号中,该申请通过引用以其整体并入。

[0102] 分区的集合:如本文使用的,"分区的集合"是指基于核酸分子对结合剂的不同结合亲和力而分区成集合/组的核酸分子的集合。结合剂优先地结合包含具有表观遗传修饰的核苷酸的核酸分子。例如,如果表观遗传修饰是甲基化,则结合剂可以是甲基结合结构域(MBD)蛋白。在一些实施方案中,分区的集合可以包含属于特定表观遗传修饰水平/程度(即,表观遗传状态)的核酸分子。例如,核酸分子可以被分区成三个集合:一个集合为高度甲基化的核酸分子(或高甲基化核酸分子),其可以被称为高甲基化分区的集合或高分区的集合;另一个集合为低度甲基化的核酸分子(或低甲基化核酸分子),其可以被称为低甲基化分区的集合或低分区的集合;以及第三集合为中等甲基化的核酸分子,其可以被称为中等甲基化分区的集合或中等分区的集合。在另一种实例中,核酸分子可以基于具有表观遗传修饰的核苷酸数目来分区——一个分区的集合可以具有带有9个甲基化核苷酸的核酸分子,并且另一个分区的集合可以具有未甲基化的核酸分子(零甲基化核苷酸)。

[0103] 多核苷酸:如本文使用的,"多核苷酸"、"核酸"、"核酸分子"或"寡核苷酸"是指通过核苷间键连接的核苷(包括脱氧核糖核苷、核糖核苷或其类似物)的线性聚合物。通常,多核苷酸包含至少三个核苷。寡核苷酸的大小的范围通常从几个单体单元例如3-4个到几百个单体单元。每当多核苷酸以一串字母诸如"ATGCCTG"表示时,应当理解,这些核苷酸从左到右是5'→3'的顺序,并且在DNA的情况下,"A"表示脱氧腺苷,"C"表示脱氧胞苷,"G"表示脱氧鸟苷,并且"T"表示脱氧胸苷,除非另外说明。如本领域标准的,字母A、C、G和T可以用于指碱基本身、核苷或包含这些碱基的核苷酸。

[0104] 参考序列:如本文使用的,"参考序列"是指用于与实验确定的序列进行比较的目的的已知序列。例如,已知序列可以是整个基因组、染色体或其任何区段。参考序列通常包括至少约20个、至少约50个、至少约100个、至少约200个、至少约250个、至少约300个、至少约350个、至少约400个、至少约450个、至少约500个、至少约1000个或多于1000个核苷酸。参考序列可以与基因组或染色体的单个连续序列对齐,或者可以包括与基因组或染色体的不

同区域对齐的非连续区段。参考序列的实例包括,例如,人类基因组,诸如,hG19和hG38。

[0105] 样品:本文使用的,“样品”意指能够通过本文公开的方法和/或系统分析的任何东西。

[0106] 测序:如本文使用的,“测序”是指用于确定生物分子例如核酸诸如DNA或RNA的序列(例如,单体单元的身份和顺序)的若干种技术中的任一种。测序方法的实例包括但不限于靶向测序、单分子实时测序、外显子或外显子组测序、内含子测序、基于电子显微术的测序、panel测序、晶体管介导的测序、直接测序、随机鸟枪法测序、Sanger双脱氧终止测序、全基因组测序、杂交测序、焦磷酸测序、毛细管电泳、凝胶电泳、双链体测序、循环测序、单碱基延伸测序、固相测序、高通量测序、大规模并行信号测序(massively parallel signature sequencing)、乳液PCR、低变性温度共扩增PCR(COLD-PCR)、多重PCR、可逆染料终止子测序、配对末端测序、near-term测序、外切核酸酶测序、连接测序、短读段测序、单分子测序、合成测序、实时测序、反向终止子测序、纳米孔测序、454测序、Solexa基因组分析仪测序、SOLiD™测序、MS-PET测序及其组合。在一些实施方案中,测序可以通过诸如,例如可从Illumina,Inc.、Pacific Biosciences,Inc.或Applied Biosystems/Thermo Fisher Scientific以及许多其他的商业可获得的基因分析仪进行。

[0107] 序列信息:如本文在核酸聚合物的上下文中使用的“序列信息”意指该聚合物中单体单元(例如,核苷酸等)的顺序和身份。

[0108] 体细胞突变:如本文使用的,术语“体细胞突变”或“体细胞变异”可互换使用。它们是指受孕后发生的基因组中的突变。体细胞突变可以发生在除生殖细胞外的任何身体细胞中,并且因此不会传给后代。

[0109] 加标样品:如本文使用的,“加标样品(spiked-in sample)”是其中表观遗传对照核酸分子被添加到来自受试者的多核苷酸样品中的样品。

[0110] 受试者:如本文使用的,“受试者”是指动物,诸如哺乳动物物种(例如,人类),或禽类(例如,鸟类)物种,或其他生物体,诸如植物。更特别地,受试者可以是脊椎动物,例如,哺乳动物,诸如小鼠、灵长类动物、猿或人类。动物包括农场动物(例如,生产用牛(production cattle)、奶牛、家禽、马、猪等)、运动动物和伴侣动物(例如,宠物或辅助动物)。受试者可以是健康的个体,患有或被怀疑患有一种疾病或有患该疾病倾向的个体,或需要治疗或被怀疑需要治疗的个体。术语“个体”或“患者”与“受试者”意图是可互换的。

[0111] 例如,受试者可以是已经被诊断患有癌症、将要接受癌症治疗和/或已经接受至少一种癌症治疗的个体。受试者可以处于癌症缓解中。作为另一个实例,受试者可以是被诊断患有自身免疫疾病的个体。作为另一个实例,受试者可以是妊娠或计划妊娠的女性个体,其可能已经被诊断患有或被怀疑患有一种疾病,例如癌症、自身免疫疾病。

[0112] 详细描述

[0113] I.综述

[0114] 基于基因组/表观遗传分区的方法可以允许在一次测定中进行多分析物、同时检测信号。然而,基于分区的分析物的检测信号可能具有较差的分辨率,并且受到改变信号灵敏度和特异性的可变测定条件的影响。期望增加液体活检测定的灵敏度,同时减少过程中循环核酸(原始材料)或数据的损失。还期望通过使用本文所述的一种或更多种对照来控制测定可变性,从而提供跨不同实验比较结果的能力。

[0115] 本公开内容提供了用于校准表观遗传分区测定的方法和组合物。本发明包括使用具有完全解析的基因组/表观遗传特征(例如,合成的寡核苷酸双链体中甲基化胞嘧啶的离散数目)的表观遗传对照核酸分子的集合作为对照或参考,以提高被分析样品的信号灵敏度和特异性。这些分子可用于评价样品中核酸分子基于表观遗传修饰的分区,并且还可用于确定样品中核酸分子的表观遗传状态。

[0116] 核酸分子,诸如无细胞多核苷酸,可以基于表观遗传特征(诸如甲基化)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。例如,DNA分子在其表观遗传修饰的程度上可以不同。修饰的程度可以指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高甲基化的。形式可以通过特征的组合,例如单链未甲基化的或双链甲基化的来表征。基于一个特征或特征的组合对分子的分级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于一种形式和/或修饰,使得可以获得针对多于一种形式的序列信息。所述方法经过处理和分析仍保持初始的多于一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。一些方法包括分离、加标签和随后汇集不同的形式或修饰状态,减少分析样品中存在的多于一种形式所需的处理步骤的数目。对样品中的多于一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(当非常低的核酸总量可用时,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,遗传信息和表观遗传信息)可以彼此关联,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。

[0117] 核酸分子的特征可以是修饰,修饰可以包括各种化学修饰(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甲基化包括将甲基基团添加至CpG位点(胞嘧啶-磷酸-鸟嘌呤位点(即,在核酸序列的5'→3'方向胞嘧啶后跟鸟嘌呤))处的胞嘧啶。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$N^6$-甲基腺嘌呤方面。在一些实施方案中,DNA甲基化是5-甲基化(对胞嘧啶的6-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-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位点处,例如,甲基化可以发生在CpA、CpT或CpC位点处。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以破坏表观遗传调节。表观遗传调节中的破坏,例如抑制,可以引起疾病,诸如癌症。启动子的DNA甲基化可以指示癌症。

[0118] CpG二联体是双链DNA分子的有义链上的二核苷酸CpG(胞嘧啶-磷酸-鸟嘌呤,即,在核酸序列的5'→3'方向上,胞嘧啶后跟鸟嘌呤)及反义链上的其互补CpG。CpG二联体可以是完全甲基化或半甲基化的。图1是双链DNA中完全甲基化和半甲基化的CpG二联体的示意图。图1A示出了完全甲基化的CpG二联体103,其中链101和链102两者上的CpG二联体的胞嘧

啶核苷酸都是甲基化的(M-甲基胞嘧啶;G-鸟嘌呤)。图1B示出了半甲基化的CpG二联体104,其中一条链101上的CpG二联体的胞嘧啶核苷酸是甲基化的,而互补链102上的CpG二联体的胞嘧啶核苷酸是未甲基化的(C-未甲基化的胞嘧啶;G-鸟嘌呤)。

[0119] CpG二核苷酸在正常人类基因组中是未被充分代表的(underrepresented),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。

[0120] 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TSS处的CpG岛中的DNA甲基化的局部增加。这种高甲基化可以与涉及的基因的转录能力的异常损失相关,并且与引起基因表达改变的点突变和缺失至少同样频繁地发生。DNA甲基化谱分析(profile)可以用于检测基因组中具有不同甲基化程度的区域(“差异性甲基化区域”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疾病)的扰动。

[0121] 甲基化谱分析可以包括确定遍及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,每个分子甲基化核苷酸的相对数目)对分子分区和测序之后,可以将不同分区中的分子的序列映射至参考基因组。这可以显示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。以此方式,与个体分子形成对比,基因组区域在其甲基化程度上可以不同。除了甲基化之外,其他表观遗传修饰可以类似地进行谱分析。

[0122] 样品中的核酸分子可以基于一个或更多个特征来分级或分区。将样品中的核酸分子分区可以增加罕见的信号。例如,存在于高甲基化DNA中但较少(或不)存在于低甲基化DNA中的遗传变异可以通过将样品分区成高甲基化和低甲基化的核酸分子来更容易地检测。通过分析样品的多于一个级分,可以进行单个分子的多维分析,并且因此可以实现更高的灵敏度。分级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。分级可以包括基于基因组特征(诸如表观遗传修饰)存在的程度将核酸分子物理地分区成分区组。可以基于指示差异性基因表达或疾病状态的特征,将样品分级或分区成一个或更多个组分区。样品在核酸分析期间可以基于提供正常状态和患病状态之间的信号差异的特征或特征组合来分级,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”)。

[0123] 本公开内容提供了用于评估或评价核酸分子的分区以及确定核酸分子的表观遗传状态(例如甲基化状态)和表观遗传修饰的核苷酸数目(例如甲基化核苷酸数目)的方法、组合物和系统。该方法可包括基于一种或多于一种表观遗传修饰将核酸分子分区成不同的分区的集合,随后测序(单独或一起)并分析每个分区中的核酸分子。在一些实施方案中,核酸的分区被针对特定的靶基因组区域进行富集。在一些实施方案中,核酸分子的分区在富集之前和/或之后被扩增。在一些实施方案中,富集可以在用分子条形码对分区的集合进行差异性加标签(differentially tagged)并重新组合为差异性加标签的分区的集合的混合物之后进行。这些方法可用于各种应用,诸如疾病的预后、诊断和/或监测。在一些实施方案中,疾病是癌症。

[0124] 核酸分子的分区方法可以通过使用表观遗传对照核酸分子来评价。表观遗传对照核酸分子是可以具有表观遗传修饰的核苷酸的合成的核酸分子。在一些实施方案中,表观

遗传对照核酸分子可以包含具有不同表观遗传状态的核酸分子。表观遗传状态可以指核酸分子的表观遗传修饰的水平/程度。例如,如果表观遗传修饰是DNA甲基化,那么表观遗传状态可以指高度甲基化、低度甲基化或中等甲基化的核酸分子。表观遗传状态也可以指带有表观遗传修饰的核苷酸的数目。例如,如果表观遗传修饰是DNA甲基化,那么表观遗传状态可以指核酸分子的甲基化核苷酸的数目。表观遗传修饰可以是不改变核酸分子的序列和/或碱基配对特异性的核苷酸的碱基的任何修饰。修饰可以是核苷酸的碱基的化学修饰。在一些情况下,修饰可以是核苷酸的碱基的甲基化。例如,修饰可以是胞嘧啶的甲基化,产生5-甲基胞嘧啶。

[0125] 在一些实施方案中,表观遗传对照核酸分子是合成的分子,表观遗传对照核酸分子的序列以及表观遗传对照核酸分子中表观遗传修饰的核苷酸的位置和数目在分析之前是已知的。因此,通过将表观遗传对照核酸分子添加到多核苷酸样品中,并通过追溯分区的集合中的表观遗传对照核酸分子,人们可以分析表观遗传对照核酸分子的分区的有效性。

[0126] 因此,在一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:(a)向多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合,从而产生加标样品;(b)将加标样品的至少一个子集的核酸分子分区成多于一个分区的集合;(c)富集来自多于一个分区的集合的分子的子集以生成富集的分子的集合,其中富集的分子的集合包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子;(d)对富集的分子的集合测序以产生测序读段的集合;(e)分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子的一个或更多个表观遗传分区评分;以及(f)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。在这些实施方案中,样品的核酸分子和表观遗传对照核酸分子的分区必须同时进行。在一些实施方案中,分析步骤包括估计在至少一个分区的集合中在给定表观遗传状态的表观遗传对照核酸分子的数目/分数。

[0127] 图2示出了用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法200的示例性实施方案。在201,将表观遗传对照核酸分子添加到其分区将被评价的样品中以生成加标样品。

[0128] 在一些实施方案中,表观遗传对照核酸分子可以包含具有不同水平的表观遗传状态(即,不同数目的表观遗传修饰的核苷酸)的一个或更多个核酸分子子集。在一些实施方案中,表观遗传对照核酸分子可以包含具有不同序列和/或不同长度的核酸分子。在其他实施方案中,表观遗传对照核酸分子可以包含具有相同序列或相同长度的核酸分子。

[0129] 在202,基于分子的表观遗传状态,将包括表观遗传对照核酸分子和来自多核苷酸样品的核酸分子两者的加标样品的至少一个子集的核酸分子分区或分级成多于一个分区的集合。分区可以基于表观遗传修饰的存在或不存在和/或可以基于表观遗传修饰的程度。表观遗传修饰的实例包括但不限于甲基化的存在或不存在、甲基化水平和甲基化类型(5'胞嘧啶)。在一些实施方案中,表观遗传修饰可以是DNA甲基化。在那些实施方案中,基于不同的甲基化水平(不同数目的甲基化核苷酸)将加标样品的分子分区。在一些实施方案中,加标样品可以被分区成两个或更多个分区的集合(例如,至少3个、4个、5个、6个或7个分区的集合)。在一些实施方案中,分区是基于核酸分子对结合剂的不同结合亲和力。结合剂的实例包括但不限于甲基结合结构域((MBD)和甲基结合蛋白(MBP)。本文设想的MBP的实例包

括,但不限于:

[0130] (a)相比于结合未修饰的胞嘧啶,优先结合5-甲基-胞嘧啶的蛋白MeCP2;

[0131] (b)相比于结合未修饰的胞嘧啶,优先结合5-羟甲基-胞嘧啶的RPL26、PRP8和DNA错配修复蛋白MHS6;

[0132] (c)相比于结合未修饰的胞嘧啶,优先结合5-甲酰基-胞嘧啶的FOXKI、FOXK2、FOXP1、FOXP4和FOXI3(Iurlaro等人,Genome Biol.14,R119(2013));和

[0133] (d)对一个或更多个甲基化核苷酸碱基特异的抗体。

[0134] 对于一些亲和剂和修饰,尽管与剂的结合可以取决于核酸是否带有修饰而以基本上全或无的方式发生,但是分离可以是一定程度的。在这样的实施方案中,与修饰未被充分代表的核酸相比,修饰被过度代表的核酸(nucleic acids overrepresented in a modification)与剂以更大的程度与剂结合。可替选地,具有修饰的核酸可以以全或无的方式结合。但是然后,各种水平的修饰可以从结合剂顺序洗脱。

[0135] 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,可以使用甲基结合结构域蛋白(例如MethylMiner甲基化DNA富集试剂盒(ThermoFisher Scientific))将所有甲基化片段与未甲基化的片段分区。随后,另外的分区可以包括通过调整含有甲基结合结构域和结合片段的溶液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。

[0136] 在一些实施方案中,分区包括基于核酸分子对结合剂的不同结合亲和力对核酸分子进行分区,所述结合剂优先结合包含具有表观遗传修饰的核苷酸的核酸分子。

[0137] 在一些实施方案中,分区的集合代表具有不同程度的修饰(过度代表性(over representative)或代表性不足(under representative)的修饰)的核酸。过度代表性和代表性不足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,如果样品中的核酸分子中5-甲基胞嘧啶核苷酸的中位数为2,则包含多于两个5-甲基胞嘧啶残基的核酸分子的该修饰是过度代表性的,而具有1个或0个5-甲基胞嘧啶残基的核酸是代表性不足的。亲和分离的作用是将结合相中修饰被过度代表的核酸与非结合相(即,溶液中)中修饰未被充分代表的的核酸分区。结合相的核酸可以在后续处理之前洗脱。

[0138] 当使用MethylMiner甲基化DNA富集试剂盒(ThermoFisher Scientific)时,可以使用顺序洗脱将不同水平的甲基化分区。例如,可以通过使核酸群体与来自试剂盒的附接至磁珠的MBD接触,将低甲基化分区(无甲基化)与甲基化分区分离。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,以洗脱具有各种甲基化水平的核酸。例如,第一组甲基化核酸可以在约150mM或约160mM或更高,例如至少150mM、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,磁性分离再次用于将较高水平的甲基化核酸与具有较低甲基化水平的核酸分离。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)和高甲基化分区(代表高甲基化水平)。

[0139] 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。这样的核酸可以富集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)程度的修饰的核酸。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。

[0140] 核酸分子的分区可以通过对分区的核酸分子测序或通过数字液滴PCR(ddPCR)或定量PCR(qPCR)来分析。在分析分区之前,可以富集分区的集合中的核酸分子,使得来自感兴趣的核酸分子的信号可以增加,并从而提高灵敏度。在203,富集多于一个分区的集合中的核酸分子的至少一个子集,使得表观遗传对照核酸分子和来自属于感兴趣区域的多核苷酸样品的核酸分子被富集。

[0141] 在一些实施方案中,在富集之前,对多于一个分区的集合中的每一个进行差异性加标签。然后将加标签的分区的集合汇集在一起,用于集体样品制备和/或测序。分区的集合的差异性加标签有助于保持追溯属于特定分区的集合的核酸分子。标签通常作为衔接子的组件提供。不同分区的集合中的核酸分子接收可以将一个分区的集合的成员与另一个分区的集合的成员区分开的不同的标签。与同一分区的集合的核酸分子相连的标签可以彼此相同或不同。但是如果彼此不同,则标签序列的一部分可以是共有的,以便将它们所附接的分子鉴定为特定的分区的集合。例如,如果加标样品的分子被分区成两个分区的集合—P1和P2,那么P1中的分子可以用A1、A2、A3等加标签,并且P2中的分子可以用B1、B2、B3等加标签。这样的加标签系统允许区分分区的集合和区分分区的集合内的分子。

[0142] 在204,富集的分子的至少一个子集被测序。获得的序列信息包括核酸分子的序列和附接至核酸分子的标签。根据附接至核酸分子的标签的序列,人们可以将标签与核酸分子的分区的集合相关联。序列信息用于鉴定表观遗传对照核酸分子及其对应的分区的集合。该信息用于分析表观遗传对照核酸分子的分区。在205,确定属于一个或更多个分区的集合中的一种或更多种表观遗传状态的表观遗传对照核酸分子的一个或更多个表观遗传分区评分。在一些实施方案中,分区方法的灵敏度和/或特异性可以通过表观遗传分区评分来评估。表观遗传分区评分是代表属于特定表观遗传状态的核酸分子的分区的评分。属于一种表观遗传状态的核酸分子的表观遗传分区评分是为每个分区的集合确定的。例如,可以确定属于特定表观遗传状态的表观遗传对照核酸分子的表观遗传分区评分。表观遗传分区评分可以是属于特定表观遗传状态的核酸分子的数目(或统计学上估计的数目)的量度。表观遗传分区评分可以用分数或百分比来表示。表观遗传分区评分可以是属于被分区到至少一个分区的集合中的特定表观遗传状态的表观遗传对照核酸分子的数目与属于其他剩余分区的集合中存在的该表观遗传状态的表观遗传对照核酸分子的数目之比的量度。在一些实施方案中,表观遗传分区评分可以是属于被分区到至少一个分区的集合中的特定表观遗传状态的表观遗传对照核酸分子的数目与所有分区的集合中属于该表观遗传状态的表观遗传对照核酸分子的总数目之比的分数或百分比。在一些实施方案中,表观遗传分区评分是为每个分区的集合中的表观遗传对照核酸分子的每种表观遗传状态确定的。在一些实施方案中,表观遗传分区评分是为在一个或更多个分区的集合中具有一种或更多种特定表观遗传状态的表观遗传对照核酸分子确定的。在一些实施方案中,表观遗传分区评分是为特定分区的集合中具有特定表观遗传状态的表观遗传对照核酸分子确定的。

[0143] 在一些实施方案中,表观遗传分区评分可以针对无CG('零'CG)的分子被分区到高分区的集合的效率。该评分可称为0CG评分。在一些实施方案中,可以用高分区的集合中无CG的分子的分数或百分比来表示0CG评分。在一些实施方案中,表观遗传分区评分可以是高甲基化分区的集合(即,高度甲基化的分区的集合)中具有以下至少一种的表观遗传对照核酸分子的分数和/或高甲基化对照分子的分数的量度:

[0144] (i)1个甲基CG(表观遗传分区评分可称为1CG评分),

[0145] (ii)2个甲基CG(表观遗传分区评分可称为2CG评分),

[0146] (iii)3个甲基CG(表观遗传分区评分可称为3CG评分),

[0147] (iv)4个甲基CG(表观遗传分区评分可称为4CG评分)和

[0148] (v)5个甲基CG(表观遗传分区评分可称为5CG评分)。

[0149] 在一些实施方案中,表观遗传分区评分可以针对低甲基化(即,低度甲基化)表观遗传对照核酸分子被分区到高甲基化分区的集合的效率。该评分可称为低评分。在一些实施方案中,低评分可以用高甲基化分区的集合中低甲基化表观遗传对照核酸分子的分数或百分比来表示。在一些实施方案中,表观遗传分区评分可以是低甲基化分区的集合中少于5%的高甲基化表观遗传对照核酸分子所需的甲基化CG数目的量度。该评分可称为甲基-5。在一些实施方案中,表观遗传分区评分可以是高甲基化分区的集合中至少50%的高甲基化表观遗传对照核酸分子所需的甲基化CG的数目的量度。该评分可称为甲基-半。

[0150] 例如,使用表观遗传对照核酸分子的三个子集(子集A、子集B和子集C),并且每个子集的甲基化核苷酸数目不同。这三个子集中的表观遗传对照核酸分子可以根据它们与甲基结合蛋白的结合亲和力分为三个分区的集合—P1、P2和P3。对于每个子集,表观遗传分区评分是为每个分区的集合(P1、P2和P3)确定的,即属于子集A的表观遗传对照核酸分子将具有三个表观遗传分区评分—三个分区的集合P1、P2和P3各一个。同样,子集B和子集C每一个将具有三个表观遗传分区评分—三个分区的集合P1、P2和P3各一个。表观遗传分区评分也可以为内源对照分子确定。

[0151] 在另一种实施方案中,使用表观遗传对照核酸分子的三个子集(子集A、子集B和子集C),并且每个子集的甲基化核苷酸数目不同(即,每个子集具有不同的表观遗传状态)。这三个子集中的表观遗传对照核酸分子可以根据它们与甲基结合蛋白的结合亲和力分为三个分区的集合—P1、P2和P3。在该实施方案中,表观遗传评分仅为P1分区的集合中的子集A分子确定。该表观遗传评分可以是P1分区的集合中的子集A分子相对于(在P1、P2和P3分区的集合中)子集A分子总数的分数或百分比的量度。

[0152] 表观遗传分区评分可以是0-1之间(以分数表示)或0%-100%之间(以百分比表示)的任何值或范围。在一些实施方案中,表观遗传分区评分可以以甲基化CG的数目(例如,甲基-半和甲基-5)表示。

[0153] 在206,将表观遗传对照核酸分子的表观遗传分区评分与表观遗传分区截止值(预定截止值)进行比较以评价分区方法。表观遗传分区截止值是用于评价属于特定表观遗传状态的核酸分子的分区的预定截止值或截止范围,并且每个分区的集合具有属于一种表观遗传状态的核酸分子的表观遗传分区截止值。表观遗传分区截止值随核酸分子和分区的集合的表观遗传状态而不同,即,每种表观遗传状态将具有其自己的表观遗传分区截止值,并且每个分区的集合具有该表观遗传状态的单独的表观遗传分区截止值。截止值可以是百分比或评分,并且截止值可以是截止范围,而不是特定的截止值。例如,对于分区的集合P1、P2和P3,属于特定表观遗传状态的表观遗传对照核酸分子的表观遗传分区截止值可以分别在70%-79%之间、10%-15%之间和小于5%。如果属于该表观遗传状态的表观遗传对照核酸分子的表观遗传分区评分在对应的表观遗传分区截止值内,则分区方法是成功的。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.01%、0.02%、0.05%、0.1%、0.2%、

0.3%、0.4%、0.5%、0.6%、0.7%、0.8%、0.9%、1%、2%、5%、至少5%或至少10%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.01%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.02%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.03%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.04%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.05%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.1%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.2%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.3%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.4%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.5%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.6%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.7%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.8%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.9%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是1%。

[0154] 在一些实施方案中,低评分的表观遗传分区截止值可以是0.1%、0.5%、1%、2%、3%、4%、5%、7%或至少10%。在一些实施方案中,低评分的表观遗传分区截止值可以是0.1%。在一些实施方案中,低评分的表观遗传分区截止值可以是0.5%。在一些实施方案中,低评分的表观遗传分区截止值可以是1%。在一些实施方案中,低评分的表观遗传分区截止值可以是2%。在一些实施方案中,低评分的表观遗传分区截止值可以是3%。在一些实施方案中,低评分的表观遗传分区截止值可以是4%。在一些实施方案中,低评分的表观遗传分区截止值可以是5%。

[0155] 在一些实施方案中,截止值可以以甲基化CG的数目表示(例如,以甲基-5和甲基-半)。在一些实施方案中,甲基-5的表观遗传分区截止值可以是5个、10个、20个、30个、40个或50个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是5个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是10个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是20个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是30个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是40个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是50个mCG。

[0156] 在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是5个、10个、15个、20个、25个、30个、35个或40个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是5个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是10个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是15个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是20个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是25个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是30个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是35个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是40个mCG。

[0157] 在一些实施方案中,如果一个或更多个分区的集合中属于一种或更多种表观遗传状态的表观遗传对照核酸分子的一个或更多个表观遗传分区评分在对应的表观遗传分区截止值内,则分区方法可被分类为成功。否则,如果表观遗传分区评分在截止值以外,则所有分区的集合可被分类为不成功的分区方法。例如,有两个表观遗传对照核酸分子的子集—子集A和子集B,并且每个子集的表观遗传修饰程度不同(即,每个子集中的表观遗传状

态不同)。这些表观遗传对照核酸分子可以分区成两个分区的集合—P1和P2。对于属于子集A的分子,将基于它们的分区来确定两个表观遗传分区评分(例如S1和S2),每个分区的集合P1和P2一个。同样,对于属于子集B的分子,将确定两个表观遗传分区评分(例如S3和S4),一个用于P1,并且一个用于P2。具有特定表观遗传状态的分子的每个子集对于每个分区的集合将具有预定的表观遗传分区截止值。在该实例中,子集A的表观遗传对照核酸分子将具有两个表观遗传分区截止值C1和C2(对于两个分区的集合P1和P2),并且同样,子集B的表观遗传对照核酸分子将具有两个表观遗传分区截止值C3和C4。将两个子集的表观遗传分区评分与其对应的表观遗传分区截止值进行比较。在该实例中,只有当所有四个表观遗传分区评分都在其对应的表观遗传分区截止值内(即,在该实例中,S1<C1且S2<C2且S3<C3且S4<C4)时,分区方法才被认为是成功的。否则,如果表观遗传分区评分在截止值以外,则所有分区的集合可被分类为不成功的分区方法。

[0158] 在另一种实施方案中,使用表观遗传对照核酸分子的三个子集(子集A、子集B和子集C),并且每个子集的甲基化核苷酸数目不同(即,每个子集具有不同的表观遗传状态)。这三个子集中的表观遗传对照核酸分子可以根据它们与甲基结合蛋白的结合亲和力分为三个分区的集合—P1、P2和P3。在该实施方案中,表观遗传评分仅为P1分区的集合中的子集A分子确定。该表观遗传评分可以是P1分区的集合中的子集A分子相对于(在P1、P2和P3分区的集合中)子集A分子总数的分数或百分比的量度。如果该表观遗传分区评分在其对应的表观遗传分区截止值内,则分区方法被分类为成功。否则,分区方法被分类为不成功。

[0159] 在另一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:(a)向多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合,从而产生加标样品;(b)将加标样品的至少一个子集的核酸分子分区成多于一个分区的集合;(c)富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中富集的分子的集合包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子,其中来自多核苷酸样品的一组核酸分子包括内源对照分子的集合;(d)对富集的分子的集合的至少一个子集测序以产生测序读段的集合;(e)分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子的一个或更多个表观遗传分区评分和内源对照分子的集合的一个或更多个表观遗传分区评分;以及(f)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。在这些实施方案中,样品的核酸分子和表观遗传对照核酸分子的分区必须同时进行。在一些实施方案中,分析步骤包括估计在至少一个分区的集合中在给定表观遗传状态的表观遗传对照核酸分子和/或内源对照分子的数目/分数。

[0160] 图3示出了用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法300的示例性实施方案。在该实施方案中,分析表观遗传对照核酸分子和内源对照分子两者的分区以评价分区方法。人类基因组中存在具有特定表观遗传状态的区域,并且该区域的表观遗传状态通常不变化/改变,并且总是保持相同/与不同受试者和/或不同类型的疾病/疾病阶段保持一致。多核苷酸样品中对应于这种具有非可变表观遗传状态的人类基因组区域的核酸分子被称为内源对照分子。在301,将表观遗传对照核酸分子添加到其分区将被评价的多核苷酸样品中以生成加标样品。

[0161] 在一些实施方案中,表观遗传对照核酸分子可以包含具有不同水平的表观遗传状

态(即,不同数目的表观遗传修饰的核苷酸)的一个或更多个核酸分子子集。在一些实施方案中,表观遗传对照核酸分子可以包含具有不同序列和/或不同长度的核酸分子。在其他实施方案中,表观遗传对照核酸分子可以包含具有相同序列或相同长度的核酸分子。

[0162] 在302,基于分子的表观遗传状态,将包括表观遗传对照核酸分子和来自多核苷酸样品的核酸分子两者的加标样品的至少一个子集的核酸分子分区或分级成多于一个分区的集合。分区可以基于表观遗传修饰的存在或不存在和/或可以基于表观遗传修饰的程度。表观遗传修饰的实例包括但不限于甲基化的存在或不存在、甲基化水平和甲基化类型(5'胞嘧啶)。在一些实施方案中,表观遗传修饰可以是DNA甲基化。在那些实施方案中,基于不同的甲基化水平(不同数目的甲基化核苷酸)将加标样品的分子分区。在一些实施方案中,加标样品可以被分区成两个或更多个分区的集合(例如,至少3个、4个、5个、6个或7个分区的集合)。在一些实施方案中,分区是基于核酸分子对结合剂的不同结合亲和力。

[0163] 核酸分子的分区可以通过对分区的核酸分子测序,通过数字液滴PCR(ddPCR)或定量PCR(qPCR)来分析。在分析分区之前,可以富集分区的集合中的核酸分子,使得来自感兴趣的核酸分子的信号可以增加,并从而提高灵敏度。在303,富集多于一个分区的集合中的核酸分子的至少一个子集,使得表观遗传对照核酸分子、内源对照分子(来自多核苷酸样品)和来自属于感兴趣区域的多核苷酸样品的其他核酸分子被富集。

[0164] 在一些实施方案中,在富集之前,对多于一个分区的集合中的每一个进行差异性加标签。然后将加标签的分区的集合汇集在一起,用于集体样品制备和/或测序。分区的集合的差异性加标签有助于保持追溯属于特定分区的集合的核酸分子。标签通常作为衔接子的组件提供。不同分区的集合中的核酸分子接收可以将一个分区的集合的成员与另一个分区的集合的成员区分开的不同的标签。与同一分区的集合的核酸分子相连的标签可以彼此相同或不同。但是如果彼此不同,则标签序列的一部分可以是共有的,以便将它们所附接的分子鉴定为特定的分区的集合。

[0165] 在304,富集的分子的至少一个子集被测序。获得的序列信息包括核酸分子的序列和附接至核酸分子的标签。根据附接至核酸分子的标签的序列,人们可以将标签与核酸分子的分区的集合相关联。序列信息用于鉴定表观遗传对照核酸分子和内源对照分子及它们对应的分区的集合。该信息用于分析表观遗传对照核酸分子和内源对照分子的分区。在305,确定属于一个或更多个分区的集合中的一种或更多种表观遗传状态的表观遗传对照核酸分子和内源对照分子的一个或更多个表观遗传分区评分。在一些实施方案中,分区方法的灵敏度和/或特异性可以通过表观遗传分区评分来评估。表观遗传分区评分是代表属于特定表观遗传状态的核酸分子的分区的评分。属于一种表观遗传状态的核酸分子的表观遗传分区评分是为每个分区的集合确定的。例如,可以确定属于特定表观遗传状态的表观遗传对照核酸分子和内源对照分子的表观遗传分区评分。表观遗传分区评分可以是属于特定表观遗传状态的核酸分子的数目(或统计学上估计的数目)的量度。表观遗传分区评分可以用分数或百分比来表示。表观遗传分区评分可以是以下的量度:(i)对于表观遗传对照核酸分子:属于被分区到至少一个分区的集合中的特定表观遗传状态的表观遗传对照核酸分子的数量与属于其他剩余分区的集合中存在的该表观遗传状态的表观遗传对照核酸分子的数量之比,和(ii)对于内源对照分子:属于被分区到至少一个分区的集合中的特定表观遗传状态的内源对照分子的数量与属于其他剩余分区的集合中存在的该表观遗传状态的

内源对照分子的数量之比。在一些实施方案中,表观遗传分区评分可以是:(i)对于表观遗传对照核酸分子:属于被分区到至少一个分区的集合中的特定表观遗传状态的表观遗传对照核酸分子的数量与所有分区的集合中属于该表观遗传状态的表观遗传对照核酸分子的总数量之比的分数或百分比,和(ii)对于内源对照分子:属于被分区到至少一个分区的集合中的特定表观遗传状态的内源对照分子的数量与所有分区的集合中属于该表观遗传状态的内源对照分子的总数量之比的分数或百分比。在一些实施方案中,表观遗传分区评分是为每个分区的集合中的表观遗传对照核酸分子和内源对照分子的每种表观遗传状态确定的。在一些实施方案中,表观遗传分区评分是为一个或更多个分区的集合中具有一种或更多种特定表观遗传状态的表观遗传对照核酸分子和内源对照分子确定的。在一些实施方案中,表观遗传分区评分是为特定分区的集合中具有特定表观遗传状态的表观遗传对照核酸分子和内源对照分子确定的。

[0166] 在一些实施方案中,表观遗传分区评分可以针对无CG('零'CG)的分子被分区到高分区的集合的效率。该评分可称为0CG评分。在一些实施方案中,可以用高分区的集合中无CG的分子的分数或百分比来表示0CG评分。在一些实施方案中,表观遗传分区评分可以是高甲基化分区的集合(即,高度甲基化的分区的集合)中具有以下至少一种的表观遗传对照核酸分子的分数和/或高甲基化对照分子的分数的量度:

[0167] (vi)1个甲基CG(表观遗传分区评分可称为1CG评分),

[0168] (vii)2个甲基CG(表观遗传分区评分可称为2CG评分),

[0169] (viii)3个甲基CG(表观遗传分区评分可称为3CG评分),

[0170] (ix)4个甲基CG(表观遗传分区评分可称为4CG评分)和

[0171] (x)5个甲基CG(表观遗传分区评分可称为5CG评分)。

[0172] 在一些实施方案中,表观遗传分区评分可以针对低甲基化对照分子或低甲基化表观遗传对照核酸分子被分区到高甲基化分区的集合的效率。该评分可称为低评分。在一些实施方案中,低评分可以用高甲基化分区的集合中低甲基化对照分子或低甲基化表观遗传对照核酸分子的分数或百分比来表示。在一些实施方案中,表观遗传分区评分可以是低甲基化分区的集合中少于5%的高甲基化对照分子和/或高甲基化表观遗传对照核酸分子所需的甲基化CG数目的量度。该评分可称为甲基-5。在一些实施方案中,表观遗传分区评分可以是高甲基化分区的集合中至少50%的高甲基化对照分子和/或高甲基化表观遗传对照核酸分子所需的甲基化CG的数目的量度。该评分可称为甲基-半。

[0173] 例如,使用表观遗传对照核酸分子的三个子集(子集A、子集B和子集C),并且每个子集的甲基化核苷酸数目不同。这三个子集中的表观遗传对照核酸分子可以根据它们与甲基结合蛋白的结合亲和力分为三个分区的集合—P1、P2和P3。对于每个子集,表观遗传分区评分是为每个分区的集合(P1、P2和P3)确定的,即属于子集A的表观遗传对照核酸分子将具有三个表观遗传分区评分—三个分区的集合P1、P2和P3各一个。同样,子集B和子集C每一个将具有三个表观遗传分区评分—三个分区的集合P1、P2和P3各一个。表观遗传分区评分也可以为内源对照分子确定。

[0174] 在另一种实施方案中,使用表观遗传对照核酸分子的三个子集(子集A、子集B和子集C),并且每个子集的甲基化核苷酸数目不同(即,每个子集具有不同的表观遗传状态)。这三个子集中的表观遗传对照核酸分子可以根据它们与甲基结合蛋白的结合亲和力分为三

个分区的集合—P1、P2和P3。在该实施方案中,表观遗传评分仅为P1分区的集合中的子集A分子确定。该表观遗传评分可以是P1分区的集合中的子集A分子相对于(在P1、P2和P3分区的集合中)子集A分子总数的分数或百分比的量度。

[0175] 表观遗传分区评分可以是0-1之间(以分数表示)或0%-100%之间(以百分比表示)的任何值或范围。在一些实施方案中,表观遗传分区评分可以以甲基化CG的数目表示(例如,以甲基-半和甲基-5)。

[0176] 在306,将表观遗传对照核酸分子和内源对照分子的表观遗传分区评分与其对应的表观遗传分区截止值(预定截止值)进行比较以评价分区方法。表观遗传分区截止值是用于评价属于特定表观遗传状态的核酸分子的分区的预定截止值或截止范围,并且每个分区的集合具有属于一种表观遗传状态的核酸分子的表观遗传分区截止值。表观遗传分区截止值随核酸分子和分区的集合的表观遗传状态而不同,即,每种表观遗传状态将具有其自己的表观遗传分区截止值,并且每个分区的集合具有该表观遗传状态的单独的表观遗传分区截止值。截止值可以是百分比或评分,并且截止值可以是截止范围,而不是特定的截止值。例如,对于分区的集合P1、P2和P3,属于特定表观遗传状态的表观遗传对照核酸分子的表观遗传分区截止值可以分别在70%-79%之间、10%-15%之间和小于5%。如果属于该表观遗传状态的表观遗传对照核酸分子的表观遗传分区评分在对应的表观遗传分区截止值内,则分区方法是成功的。

[0177] 在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.01%、0.02%、0.05%、0.1%、0.2%、0.3%、0.4%、0.5%、0.6%、0.7%、0.8%、0.9%、1%、2%、5%、至少5%或至少10%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.01%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.02%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.03%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.04%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.05%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.1%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.2%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.3%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.4%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.5%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.6%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.7%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.8%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.9%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是1%。

[0178] 在一些实施方案中,低评分的表观遗传分区截止值可以是0.1%、0.5%、1%、2%、3%、4%、5%、7%或至少10%。在一些实施方案中,低评分的表观遗传分区截止值可以是0.1%。在一些实施方案中,低评分的表观遗传分区截止值可以是0.5%。在一些实施方案中,低评分的表观遗传分区截止值可以是1%。在一些实施方案中,低评分的表观遗传分区截止值可以是2%。在一些实施方案中,低评分的表观遗传分区截止值可以是3%。在一些实施方案中,低评分的表观遗传分区截止值可以是4%。在一些实施方案中,低评分的表观遗传分区截止值可以是5%。

[0179] 在一些实施方案中,甲基-5的表观遗传分区截止值可以是5个、10个、20个、30个、

40个或50个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是5个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是10个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是20个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是30个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是40个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是50个mCG。

[0180] 在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是5个、10个、15个、20个、25个、30个、35个或40个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是5个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是10个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是15个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是20个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是25个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是30个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是35个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是40个mCG。

[0181] 在一些实施方案中,如果一个或更多个分区的集合中属于一种或更多种表观遗传状态的表观遗传对照核酸分子和内源对照分子的一个或更多个表观遗传分区评分在对应的表观遗传分区截止值内,则分区方法可被分类为成功。否则,分区方法可被分类为不成功。

[0182] 在另一方面,本公开内容提供了一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法,该方法包括:(a)将多核苷酸样品的至少一个子集的核酸分子分区成多于一个分区的集合;(b)富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中富集的分子的集合包括来自多核苷酸样品的一组核酸分子,其中来自无细胞多核苷酸样品的一组核酸分子包括内源对照分子的集合;(c)对富集的分子的集合的至少一个子集测序以产生测序读段的集合;(d)分析测序读段的集合的至少一个子集以生成内源对照分子的集合的一个或更多个表观遗传分区评分;以及(e)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。在这些实施方案中,样品的核酸分子和表观遗传对照核酸分子的分区必须同时进行。在一些实施方案中,分析步骤包括估计在至少一个分区的集合中在给定表观遗传状态的内源对照分子的数目/分数。

[0183] 图4示出了用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区的方法400的示例性实施方案。在该实施方案中,分析多核苷酸样品中内源对照分子的分区以评价分区方法。人类基因组中存在具有特定表观遗传状态的区域,并且该区域的表观遗传状态通常不变化/改变,并且总是保持相同/与不同受试者和/或不同类型的疾病/疾病阶段保持一致。多核苷酸样品中对应于这种具有非可变表观遗传状态的人类基因组区域的核酸分子被称为内源对照分子。在401,考虑来自受试者的多核苷酸样品。在402,基于分子的表观遗传状态,将多核苷酸样品的至少一个子集的核酸分子分区或分级成多于一个分区的集合。分区可以基于表观遗传修饰的存在或不存在和/或可以基于表观遗传修饰的程度。表观遗传修饰的实例可包括但不限于甲基化的存在或不存在、甲基化水平和甲基化类型(5'胞嘧啶)。在一些实施方案中,表观遗传修饰可以是DNA甲基化。在这些实施方案中,基于不同的甲基化水平(不同数目的甲基化核苷酸)将样品的分子分区。在一些实施方案中,样品可以

被分区成两个或更多个分区的集合(例如,至少3个、4个、5个、6个或7个分区的集合)。在一些实施方案中,分区是基于核酸分子对结合剂的不同结合亲和力。

[0184] 核酸分子的分区可以通过对分区的核酸分子测序或通过数字液滴PCR(ddPCR)来分析。在分析分区之前,可以富集分区的集合中的核酸分子,使得来自感兴趣的核酸分子的信号可以增加,并从而提高灵敏度。在403,富集多于一个分区的集合中的核酸分子的至少一个子集,使得内源对照分子(来自多核苷酸样品)和来自属于感兴趣区域的多核苷酸样品的核酸分子被富集。

[0185] 在一些实施方案中,在富集之前,对多于一个分区的集合中的每一个进行差异性加标签。然后将加标签的分区的集合汇集在一起,用于集体样品制备和/或测序。分区的集合的差异性加标签有助于保持追溯属于特定分区的集合的核酸分子。标签通常作为衔接子的组件提供。不同分区的集合中的核酸分子接收可以将一个分区的集合的成员与另一个分区的集合的成员区分开的不同的标签。与同一分区的集合的核酸分子相连的标签可以彼此相同或不同。但是如果彼此不同,则标签序列的一部分可以是共有的,以便将它们所附接的分子鉴定为特定的分区的集合。

[0186] 在404,富集的分子的至少一个子集被测序。获得的序列信息包括核酸分子的序列和附接至核酸分子的标签。根据附接至核酸分子的标签的序列,人们可以将标签与核酸分子的分区的集合相关联。序列信息用于鉴定内源对照分子及其对应的分区的集合。该信息用于分析内源对照分子的分区。在405,确定属于一个或更多个分区的集合的内源对照分子的一个或更多个表观遗传分区评分。在一些实施方案中,分区方法的灵敏度和/或特异性可以通过表观遗传分区评分来评估。表观遗传分区评分是代表属于特定表观遗传状态的核酸分子的分区的评分。在一些实施方案中,属于一种表观遗传状态的核酸分子的表观遗传分区评分是为每个分区的集合确定的。例如,可以确定属于特定表观遗传状态的内源对照分子的表观遗传分区评分。表观遗传分区评分可以是属于特定表观遗传状态的核酸分子的数目(或统计学上估计的数目)的量度。表观遗传分区评分可以用分数或百分比来表示。表观遗传分区评分可以是属于被分区到至少一个分区的集合中的特定表观遗传状态的内源对照分子的数量与属于其他剩余分区的集合中存在的该表观遗传状态的内源对照分子的数量之比的量度。在一些实施方案中,表观遗传分区评分可以是属于被分区到至少一个分区的集合中的特定表观遗传状态的内源对照分子的数量与所有分区的集合中属于该表观遗传状态的内源对照分子的总数量之比的分数或百分比。在一些实施方案中,表观遗传分区评分是为每个分区的集合中的内源对照分子的每种表观遗传状态确定的。在一些实施方案中,表观遗传分区评分是为在一个或更多个分区的集合中具有一种或更多种特定表观遗传状态的内源对照分子确定的。在一些实施方案中,表观遗传分区评分是为在特定分区的集合中具有特定表观遗传状态的内源对照分子确定的。

[0187] 在一些实施方案中,表观遗传分区评分可以针对无CG('零'CG)的分子被分区到高分区的集合的效率。该评分可称为0CG评分。在一些实施方案中,可以用高分区的集合中无CG的分子的分数或百分比来表示0CG评分。在一些实施方案中,表观遗传分区评分可以是高甲基化分区的集合(即,高度甲基化的分区的集合)中具有以下至少一种的高甲基化对照分子的分数的量度:

[0188] (xi)1个甲基CG(表观遗传分区评分可称为1CG评分),

[0189] (xii)2个甲基CG(表观遗传分区评分可称为2CG评分),

[0190] (xiii)3个甲基CG(表观遗传分区评分可称为3CG评分),

[0191] (xiv)4个甲基CG(表观遗传分区评分可称为4CG评分)和

[0192] (xv)5个甲基CG(表观遗传分区评分可称为5CG评分)。

[0193] 在一些实施方案中,表观遗传分区评分可以针对低甲基化对照分子被分区到高甲基化分区的集合的效率。该评分可称为低评分。在一些实施方案中,低评分可以用高甲基化分区的集合中低甲基化对照分子的分数或百分比来表示。在一些实施方案中,表观遗传分区评分可以是低甲基化分区的集合中少于5%的高甲基化对照分子所需的甲基化CG数目的量度。该评分可称为甲基-5。在一些实施方案中,表观遗传分区评分可以是高甲基化分区的集合中至少50%的高甲基化对照分子所需的甲基化CG的数目的量度。该评分可称为甲基-半。

[0194] 例如,分析内源对照分子的两个子集(子集A和子集B),并且每个子集的甲基化水平/程度不同(即,每个子集具有不同的表观遗传状态)。这两个子集中的内源对照分子可以根据它们与甲基结合蛋白的结合亲和力分为三个分区的集合—P1、P2和P3。对于每个子集,表观遗传分区评分是为每个分区的集合(P1、P2和P3)确定的,即属于子集A的表观遗传对照核酸分子将具有三个表观遗传分区评分—三个分区的集合P1、P2和P3各一个。同样,子集B将具有三个表观遗传分区评分—三个分区的集合P1、P2和P3各一个。

[0195] 在另一种实施方案中,分析内源对照分子的三个子集(子集A、子集B和子集C),并且每个子集的甲基化水平/程度不同(即,每个子集具有不同的表观遗传状态)。这三个子集中的内源对照分子可以根据它们与甲基结合蛋白的结合亲和力分为三个分区的集合—P1、P2和P3。在该实施方案中,表观遗传评分仅为P1分区的集合中的子集A的内源分子确定。该表观遗传评分可以是P1分区的集合中子集A的内源对照分子相对于(在P1、P2和P3分区的集合中)子集A内源对照分子总数的分数或百分比的量度。

[0196] 表观遗传分区评分可以是0-1之间(以分数表示)或0%-100%之间(以百分比表示)的任何值或范围。在一些实施方案中,表观遗传分区评分可以以甲基化CG的数目表示(例如,以甲基-半和甲基-5)。

[0197] 在406,将内源对照分子的表观遗传分区评分与其对应的表观遗传分区截止值(预定截止值)进行比较以评价分区方法。表观遗传分区截止值是用于评价属于特定表观遗传状态的核酸分子的分区的预定截止值或截止范围,并且每个分区的集合具有属于一种表观遗传状态的核酸分子的表观遗传分区截止值。表观遗传分区截止值随核酸分子和分区的集合的表观遗传状态而不同,即,每种表观遗传状态将具有其自己的表观遗传分区截止值,并且每个分区的集合具有该表观遗传状态的单独的表观遗传分区截止值。截止值可以是百分比或评分,并且截止值可以是截止范围,而不是特定的截止值。例如,对于分区的集合P1、P2和P3,属于特定表观遗传状态的内源对照分子的表观遗传分区截止值可以分别在70%-79%之间、10%-15%之间和小于5%。如果属于该表观遗传状态的内源对照分子的表观遗传分区评分在对应的表观遗传分区截止值内,则分区方法是成功的。

[0198] 在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.01%、0.02%、0.05%、0.1%、0.2%、0.3%、0.4%、0.5%、0.6%、0.7%、0.8%、0.9%、1%、2%、5%、至少5%或至少10%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.01%。在一些

实施方案中,0CG评分的表观遗传分区截止值可以是0.02%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.03%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.04%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.05%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.1%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.2%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.3%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.4%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.5%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.6%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.7%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.8%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是0.9%。在一些实施方案中,0CG评分的表观遗传分区截止值可以是1%。

[0199] 在一些实施方案中,低评分的表观遗传分区截止值可以是0.1%、0.5%、1%、2%、3%、4%、5%、7%或至少10%。在一些实施方案中,低评分的表观遗传分区截止值可以是0.1%。在一些实施方案中,低评分的表观遗传分区截止值可以是0.5%。在一些实施方案中,低评分的表观遗传分区截止值可以是1%。在一些实施方案中,低评分的表观遗传分区截止值可以是2%。在一些实施方案中,低评分的表观遗传分区截止值可以是3%。在一些实施方案中,低评分的表观遗传分区截止值可以是4%。在一些实施方案中,低评分的表观遗传分区截止值可以是5%。

[0200] 在一些实施方案中,甲基-5的表观遗传分区截止值可以是5个、10个、20个、30个、40个或50个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是5个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是10个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是20个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是30个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是40个mCG。在一些实施方案中,甲基-5的表观遗传分区截止值可以是50个mCG。

[0201] 在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是5个、10个、15个、20个、25个、30个、35个或40个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是5个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是10个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是15个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是20个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是25个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是30个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是35个mCG。在一些实施方案中,甲基-半评分的表观遗传分区截止值可以是40个mCG。

[0202] 在一些实施方案中,如果一个或更多个分区的集合中属于一种或更多种表观遗传状态的内源对照分子的一个或更多个表观遗传分区评分在对应的表观遗传分区截止值内,则分区方法可被分类为成功。否则,分区方法可被分类为不成功。

[0203] 在另一方面,本公开内容提供了一种用于确定多核苷酸样品中核酸分子的表观遗传状态的方法,该方法包括:(a)向多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合,从而产生加标样品;(b)将加标样品的至少一个子集的核酸分子分区成多于一个分区的集合;(c)富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的

集合,其中富集的分子的集合包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子;(d)对富集的分子的集合的至少一个子集测序以产生测序读段的集合;(e)分析测序读段的集合的至少一个子集以生成多于一个分区的集合中在不同表观遗传状态的表观遗传对照核酸分子的多于一个分区谱;以及(f)使用表观遗传对照核酸分子的多于一个分区谱来估计样品的核酸分子的表观遗传状态的概率。在这些实施方案中,样品的核酸分子和表观遗传对照核酸分子的分区必须同时进行。

[0204] 在一些实施方案中,分析步骤包括确定在多于一个分区的集合中每种表观遗传状态的表观遗传对照核酸分子的数目或分数。分区谱可以指在两个或更多个分区的集合中在每种表观遗传状态的表观遗传对照核酸分子的分数/数目的表示。在一些实施方案中,分区谱还包括关于表观遗传对照核酸分子中带有表观遗传修饰的核苷酸的数目、表观遗传对照核酸分子中带有表观遗传修饰的核苷酸的位置和/或表观遗传对照核酸分子的序列组成的信息。这种分区谱可用于估计样品中核酸分子的表观遗传状态的概率。在一些实施方案中,如果表观遗传修饰是甲基化,那么分区谱可用于估计样品的核酸分子的甲基化状态的概率(即,甲基化的水平/程度或甲基化核苷酸的数目)。

[0205] 在另一方面,本公开内容提供了一种用于确定多核苷酸样品中核酸分子的表观遗传状态的方法,该方法包括:(a)将来自样品的至少一个子集的核酸分子分区成多于一个分区的集合;(b)富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子的集合,其中富集的分子的集合包括来自多核苷酸样品的一组核酸分子,其中来自多核苷酸样品的一组核酸分子包括内源对照分子的集合;(c)对富集的分子的集合的至少一个子集测序以产生测序读段的集合;(d)分析测序读段的集合的至少一个子集以生成多于一个分区的集合中在不同表观遗传状态的内源对照分子的多于一个分区谱;以及(e)使用内源对照分子的多于一个分区谱来估计核酸分子的表观遗传状态的概率。

[0206] 在一些实施方案中,分析步骤包括确定多于一个分区的集合中每种表观遗传状态的内源对照分子的数目。分区谱可以指在两个或更多个分区的集合中在每种表观遗传状态的内源对照分子的分数/数目的表示。在一些实施方案中,分区谱还包括关于表观遗传对照核酸分子中带有表观遗传修饰的核苷酸的数目、表观遗传对照核酸分子中带有表观遗传修饰的核苷酸的位置和/或表观遗传对照核酸分子的序列组成的信息。在一些实施方案中,内源对照分子中甲基化CpG的数目基于先前的实验数据和/或根据文献来确定。这种分区谱可用于估计样品中核酸分子的表观遗传状态的概率。在一些实施方案中,如果表观遗传修饰是甲基化,那么分区谱可用于估计样品的核酸分子的甲基化状态的概率(即,甲基化的水平/程度或甲基化核苷酸的数目)。

[0207] 在一些实施方案中,内源对照分子(例如,高甲基化对照分子和低甲基化对照分子)可用于估计样品的核酸分子的甲基化状态。如果有三个分区的集合—P1、P2和P3,则可以基于三个分区的集合中每一个的高甲基化对照分子的分数和高甲基化对照分子中存在的甲基化CpG的数目,为P1、P2和P3生成高甲基化对照分子的分区谱。同样,对于低甲基化对照分子,可以基于三个分区的集合中每一个的低甲基化对照分子的分数和低甲基化对照分子中存在的未甲基化CpG的数目,为P1、P2和P3生成低甲基化对照分子的分区谱。在一些实施方案中,在使用内源对照分子的情况下,内源对照分子中甲基化CpG的数目基于先前的实验数据和/或根据文献来确定。这六个分区谱可用于估计在特定区域样品的核酸分子中存

在的甲基化水平/程度或甲基化核苷酸数目的概率。

[0208] 在一些实施方案中,内源对照核酸分子(例如,高度甲基化和低度甲基化的表观遗传对照核酸分子)可用于估计样品的核酸分子的甲基化状态。如果有三个分区的集合—P1、P2和P3,则可以基于三个分区的集合中每一个的高度甲基化表观遗传对照核酸分子的分数和高度甲基化表观遗传对照核酸分子中存在的甲基化CpG的数目,为P1、P2和P3生成高度甲基化表观遗传对照核酸分子的分区谱。同样,对于低度甲基化表观遗传对照核酸分子,可以基于三个分区的集合中每一个的低度甲基化表观遗传对照核酸分子的分数和低度甲基化表观遗传对照核酸分子中存在的未甲基化CpG的数目,为P1、P2和P3生成低度甲基化表观遗传对照核酸分子的分区谱。这六个分区谱可用于估计在特定区域样品的核酸分子中存在的甲基化水平/程度或甲基化核苷酸数目的概率。

[0209] II.表观遗传对照核酸分子

[0210] 表观遗传对照核酸分子用作对照或参考分子,以评价基于表观遗传修饰的样品中核酸分子的分区。这些表观遗传对照核酸分子也可用于确定样品中核酸分子的表观遗传状态。例如,表观遗传修饰可以是DNA甲基化,并且表观遗传对照核酸分子可以具有不同/可区分的甲基化水平。表观遗传对照核酸分子可以是合成的寡核苷酸。在一些实施方案中,表观遗传对照核酸分子可以具有非天然存在的核酸序列。在一些实施方案中,表观遗传对照核酸分子可以具有天然存在的核酸序列。在一些实施方案中,表观遗传对照核酸分子可以具有对应于非人类基因组的核酸序列。例如,这些分子可以具有(i)对应于λ噬菌体DNA或人类基因组区域的序列,(ii)非天然存在的序列,和/或(iii)(i)和(ii)的组合。此外,表观遗传对照核酸分子可以被分成子集,并且每个子集可以具有代表表观遗传修饰程度的特定数目的核苷酸,并且该数目不同于每个其他子集中代表表观遗传修饰程度的核苷酸数目。

[0211] 在另一方面,本公开内容提供了表观遗传对照核酸分子的集合,包括表观遗传对照核酸分子的一个或更多个子集,其中每个子集包含多于一个表观遗传对照核酸分子,并且每个表观遗传对照核酸分子包含表观遗传修饰区域。表观遗传修饰区域是表观遗传对照核酸分子中代表表观遗传对照核酸分子的表观遗传状态的区域。表观遗传状态是核酸分子的表观遗传修饰的水平/程度。例如,如果表观遗传修饰是DNA甲基化,那么表观遗传状态可以指高度甲基化、低度甲基化或中等甲基化的核酸分子。表观遗传状态也可以指带有表观遗传修饰的核苷酸的数目。例如,如果表观遗传修饰是DNA甲基化,那么表观遗传状态可以指核酸分子的甲基化核苷酸的数目。

[0212] 在一些实施方案中,表观遗传对照核酸分子包含以下的至少一种:(i)表观遗传修饰区域和(ii)标识符区域。在一些实施方案中,表观遗传修饰区域包含具有表观遗传修饰的核苷酸。在一些实施方案中,表观遗传修饰是DNA甲基化。在这些实施方案中,表观遗传对照核酸分子的表观遗传修饰区域可以具有甲基化的核苷酸。表观遗传修饰区域中甲基化核苷酸的数目在表观遗传对照核酸分子之间可以不同。在一些实施方案中,表观遗传对照核酸分子在表观遗传修饰区域中可以具有0个、1个、2个、3个、4个、5个、6个、7个、8个、9个、至少10个、至少15个、至少20个、至少30个、至少40个或至少50个甲基化核苷酸。表观遗传对照核酸分子可以根据表观遗传修饰区域中的表观遗传状态(即,具有表观遗传修饰的核苷酸数目/表观遗传修饰水平)分成子集。不同子集之间的表观遗传修饰区域可以具有相同的长度,例如约160bp。子集之间表观遗传修饰区域的长度可以不同。例如,表观遗传对照核酸分

子可以根据表观遗传修饰区域中甲基化核苷酸的数目分为三个子集(子集A、子集B和子集C)。子集A、子集B和子集C可以具有在表观遗传修饰区域分别带有5个、10个和15个甲基化核苷酸的表观遗传对照核酸分子,并且子集A、子集B和子集C中表观遗传修饰区域的长度可以相同(例如160bp)或者可以不同—子集A、子集B和子集C分别为100bp、150bp和200bp。

[0213] 在某些实施方案中,表观遗传对照核酸分子可以被分成子集,其中每个子集代表表观遗传修饰的程度,并且每个子集内的多核苷酸数目不同于每个其他子集中的核苷酸数目。在一些实施方案中,子集中甲基化核苷酸的数目为0、1、2、3、4、5、6、7、8、9、10、11、至少12、至少15、至少20、至少25、至少30、至少40或至少50。在一些实施方案中,至少一个子集中的表观遗传对照核酸分子的表观遗传修饰区域包含至少一个具有表观遗传修饰的核苷酸。在一些实施方案中,表观遗传对照核酸分子的至少一个子集可以在表观遗传对照核酸分子的表观遗传修饰区域中包含没有任何表观遗传修饰的核苷酸(即,表观遗传上未修饰的核苷酸)。

[0214] 在一些实施方案中,子集中每个表观遗传对照核酸分子的表观遗传修饰区域包含相同数目的具有表观遗传修饰的核苷酸。在一些实施方案中,第一子集中具有表观遗传修饰的核苷酸的数目不同于第二子集中具有表观遗传修饰的核苷酸的数目。在一些实施方案中,一个或更多个子集中的多于一个表观遗传对照核酸分子的表观遗传修饰区域包含相同的核酸序列。在一些实施方案中,第一子集中的多于一个表观遗传对照核酸分子的表观遗传修饰区域包含与第二子集中的多于一个表观遗传对照核酸分子的表观遗传修饰区域的核酸序列可区分的核酸序列。

[0215] 在一些实施方案中,一个或更多个子集中的表观遗传对照核酸分子的表观遗传修饰区域可以具有相同的长度和具有相同的序列组成,但是在一个或更多个子集的每一个中,具有表观遗传修饰的核苷酸的数目可以不同。在一些实施方案中,一个或更多个子集中的表观遗传对照核酸分子的表观遗传修饰区域可以具有相同的长度和具有相同数目的带有表观遗传修饰的核苷酸,但是在一个或更多个子集的每一个中,带有表观遗传修饰的核苷酸的位置可以不同。在一些实施方案中,一个或更多个子集中的表观遗传对照核酸分子的表观遗传修饰区域可以具有相同的长度,具有相同数目的带有表观遗传修饰的核苷酸,并且带有表观遗传修饰的核苷酸的位置可以相同,但是在一个或更多个子集的每一个中,具有表观遗传修饰的核苷酸的任一侧上的相邻核苷酸可以不同。

[0216] 在一些实施方案中,表观遗传对照核酸分子的每个子集处于等摩尔浓度。在一些实施方案中,表观遗传对照核酸分子的每个子集处于非等摩尔浓度。在一些实施方案中,表观遗传修饰是DNA甲基化。在一些实施方案中,具有表观遗传修饰的核苷酸包括甲基化核苷酸。在一些实施方案中,甲基化核苷酸包含5-甲基胞嘧啶。在一些实施方案中,甲基化核苷酸包含5-羟甲基胞嘧啶。在一些实施方案中,甲基化核苷酸包含$N^6$-甲基腺嘌呤。

[0217] 在一些实施方案中,表观遗传对照核酸分子还包含标识符区域。标识符区域是表观遗传对照核酸分子中用于区分一种表观遗传对照核酸分子与其他表观遗传对照核酸分子的区域。标识符区域可以具有分子条形码和/或表观遗传状态条形码。标识符区域可以存在于表观遗传修饰区域的一侧或两侧。分子条形码用作表观遗传对照核酸分子的标识符,而表观遗传状态条形码用作表观遗传对照核酸分子的表观遗传状态的标识符。表观遗传状态条形码是用于标识表观遗传对照核酸分子的表观遗传状态的一种类型的条形码(核酸序

列)。在一些实施方案中,表观遗传状态条形码可标识(通过预定的相关性)表观遗传对照核酸分子的表观遗传修饰区域中具有表观遗传修饰的核苷酸的数目。在一些实施方案中,表观遗传状态条形码可标识表观遗传对照核酸分子的表观遗传修饰区域中的表观遗传修饰水平。在一些实施方案中,表观遗传对照核酸分子的标识符区域包含表观遗传状态条形码。例如,如果表观遗传修饰是DNA甲基化,并且表观遗传对照核酸分子的一个子集具有5个甲基化核苷酸,那么该子集内的所有表观遗传对照核酸分子将具有相同的表观遗传状态条形码。在一些实施方案中,表观遗传状态条形码可用于标识表观遗传对照核酸分子的表观遗传修饰区域的表观遗传修饰水平/程度。表观遗传对照核酸分子可以根据表观遗传修饰区域中胞嘧啶或CpG核苷酸的数目分为子集。在一些实施方案中,在每个子集中,甲基化水平可以不同(例如,高度甲基化、中等甲基化和低度甲基化),并且每个甲基化水平可以具有单独的表观遗传状态条形码。例如,在子集A中,所有低度甲基化的表观遗传对照核酸分子将具有一种表观遗传状态条形码,例如ESB1,并且所有高度甲基化的表观遗传对照核酸分子将具有另一种表观遗传状态条形码,例如ESB3。在该实例中,表观遗传状态条形码用于标识甲基化的水平/程度。标识符区域中的分子条形码可以是独特条形码(每个分子具有一个独特条形码)或非独特的条形码。分子条形码可以具有2个和50个核苷酸之间的任何长度。在一些实施方案中,分子条形码可以是至少2个、至少3个、至少4个、至少5个、至少6个、至少7个、至少8个、至少9个或至少10个核苷酸。在一些实施方案中,表观遗传状态条形码可以是至少2个、至少3个、至少4个、至少5个、至少6个、至少7个或至少8个核苷酸。

[0218] 图5是适用于本公开内容的一些实施方案的表观遗传对照核酸分子的示意图。在此描述的表观遗传对照核酸分子具有的长度与被测定样品的长度相似,并且所有子集具有相同的序列组成,以减少任何序列特异性分区效应。在图5中,例如,表观遗传对照核酸分子已被分成四个子集—子集1、子集2、子集3和子集4。图5中的表观遗传对照核酸分子是双链DNA分子。为了说明的目的,在图中仅示出了每个子集中的表观遗传对照核酸分子的一种代表。在该实施方案中,表观遗传对照核酸分子的表观遗传修饰区域的序列在所有子集中是相同的。所有四个子集中的表观遗传对照核酸分子的表观遗传修饰区域具有5个CpG二联体。双链DNA序列中的‘---’区域代表除CpG二联体以外的任何其他序列,并且M代表5-甲基胞嘧啶,C代表胞嘧啶,并且G代表鸟嘌呤。在图5中,一个子集中的表观遗传对照核酸分子的表观遗传状态(甲基化水平)不同于其他子集的表观遗传状态。子集1具有零个甲基化CpG二联体,子集2具有1个完全甲基化CpG二联体,子集3具有3个完全甲基化CpG二联体,并且子集4具有5个完全甲基化CpG二联体。在该实施方案中,标识符区域位于表观遗传修饰区域的两侧。位于两侧的标识符区域具有表观遗传状态条形码(ESB),而分子条形码(MB)仅位于一侧。分子条形码被用作表观遗传对照核酸分子的标识符,并且每种表观遗传对照核酸分子具有独特的分子条形码(即,分子1具有MB1,分子2具有MB2,分子3具有MB3,等等)。表观遗传状态条形码可以用作表观遗传对照核酸分子的表观遗传状态的标识符。在此,表观遗传状态条形码用于标识表观遗传对照核酸分子中完全甲基化的CpG二联体的数目。子集1的所有表观遗传对照核酸分子具有零个甲基化CpG二联体,因此子集1的所有表观遗传对照核酸分子具有相同的表观遗传状态条形码—ESB1。同样,子集2、子集3和子集4的所有表观遗传对照核酸分子分别具有1个、3个和5个完全甲基化的CpG二联体。因此,子集2、子集3和子集4的所有表观遗传对照核酸分子分别具有表观遗传状态条形码ESB2、ESB3和ESB4。在该实例中,

相同的表观遗传状态条形码位于表观遗传修饰区域的两侧。

[0219] 在一些实施方案中,分子条形码可以位于表观遗传修饰区域的一侧或两侧。在一些实施方案中,表观遗传状态条形码可以位于表观遗传修饰区域的一侧或两侧。在一些实施方案中,位于表观遗传修饰区域两侧的表观遗传状态条形码可以相同或不同和/或可以随机附接。

[0220] 在一些实施方案中,标识符区域可以具有促进一个或更多个引物结合的另外区域(引物结合位点)。在一些实施方案中,一个子集中的标识符区域的引物结合位点不同于其他子集中的引物结合位点。在一些实施方案中,如果在一个子集内,表观遗传对照核酸分子具有不同的表观遗传状态,那么对于分子内的每种表观遗传状态,引物结合位点可以不同,即,每种独特的表观遗传状态具有独特的引物结合位点。在一些实施方案中,这些引物结合位点用于分析表观遗传对照核酸分子的分区。在一些实施方案中,代替通过测序分析表观遗传对照核酸分子的分区,表观遗传对照核酸分子的分区可以通过使用结合这些引物结合位点的引物的数字液滴PCR(ddPCR)来分析。

[0221] 在一些实施方案中,表观遗传对照核酸分子可以被分成子集,使得每个子集中的表观遗传对照核酸分子具有相同序列,但是每个子集中的表观遗传对照核酸分子的表观遗传状态可以不同。

[0222] 图6是可适用于本公开内容的某些实施方案的表观遗传对照核酸分子的示意图。本文描述的表观遗传对照核酸分子还可以考虑核酸分子分区期间序列组成和CpG二联体/完全甲基化CpG二联体的数目的影响。在图6中,例如,表观遗传对照核酸分子已被分成三个子集—子集1、子集2和子集3。图6中的表观遗传对照核酸分子是双链DNA分子。为了说明的目的,在图中仅示出了每个子集中每种表观遗传状态的表观遗传对照核酸分子的一种代表。在该实施方案中,子集1、子集2和子集3中的表观遗传对照核酸分子的表观遗传修饰区域具有不同的长度。子集1、子集2和子集3中的表观遗传对照核酸分子的表观遗传修饰区域分别具有1个、3个和5个CpG二联体。双链DNA序列中的‘---’区域代表除CpG二联体以外的任何其他序列,并且M代表5-甲基胞嘧啶,C代表胞嘧啶,并且G代表鸟嘌呤。在图6中,在每个子集中,表观遗传对照核酸分子处于不同的表观遗传状态—例如,低度甲基化状态、中等甲基化状态和高度甲基化状态。子集1的表观遗传对照核酸分子处于两种不同的表观遗传状态—低度甲基化(零个甲基化CpG二联体)和高度甲基化(1个完全甲基化CpG二联体)。子集2的表观遗传对照核酸分子处于三种不同的表观遗传状态—低度甲基化(零个甲基化CpG二联体)、中等甲基化(1个完全甲基化CpG二联体)和高度甲基化(3个完全甲基化CpG二联体)。子集3的表观遗传对照核酸分子处于三种不同的表观遗传状态—低度甲基化(1个完全甲基化CpG二联体)、中等甲基化(3个完全甲基化CpG二联体)和高度甲基化(5个完全甲基化CpG二联体)。在此,标识符区域位于表观遗传修饰区域的两侧。位于两侧的标识符区域具有表观遗传状态条形码(ESB)和分子条形码(MB)。分子条形码被用作表观遗传对照核酸分子的标识符,并且每种表观遗传对照核酸分子具有独特的分子条形码(即,分子1具有MB1,分子2具有MB2,分子3具有MB3,等等)。表观遗传状态条形码用作表观遗传对照核酸分子的表观遗传状态的标识符。在此,表观遗传状态条形码用于标识表观遗传对照核酸分子的甲基化程度/水平,即,低度甲基化状态、中等甲基化状态或高度甲基化状态。子集1、子集2和子集3中所有低度甲基化表观遗传对照核酸分子具有相同的表观遗传状态条形码—ESB1。子集2和

子集3具有中等甲基化表观遗传对照核酸分子,并且所有这些分子具有相同的表观遗传状态条形码—ESB2(子集1不具有中等甲基化状态,因此所有表观遗传对照核酸分子将都不具有ESB2表观遗传状态条形码)。因此,根据表观遗传对照核酸分子的序列和表观遗传状态条形码的序列,可以鉴定该表观遗传对照核酸分子的表观遗传状态和该表观遗传核酸分子所属的子集。

[0223] 此外,标识符区域可以具有引物结合位点。不同的引物结合位点可用于区分每个子集内和子集间的不同表观遗传状态。例如,子集1中的低度甲基化表观遗传对照核酸分子可以在表观遗传修饰区域的任一侧具有引物结合位点—Pr1和Pr2。子集1中的高度甲基化表观遗传对照核酸分子可以在表观遗传修饰区域的任一侧具有引物结合位点—Pr3和Pr4。同样,在子集2中,低度、中等和高度甲基化表观遗传对照核酸分子可以在表观遗传修饰区域的任一侧分别具有引物结合位点Pr5和Pr6、Pr7和Pr8以及Pr9和Pr10。类似地,在子集3中,低度、中等和高度甲基化表观遗传对照核酸分子可以在表观遗传修饰区域的任一侧分别具有引物结合位点Pr11和Pr12、Pr13和Pr14以及Pr15和Pr16。此外,根据用于不同子集中不同表观遗传状态分子的不同引物集合,人们可以通过ddPCR或定量PCR(qPCR)来估计特定子集中属于特定表观遗传状态的表观遗传对照核酸分子的数目的量度。在该实施方案中,根据表观遗传状态条形码序列和表观遗传修饰区域的序列,可以确定表观遗传修饰区域中的CpG二联体的数目和表观遗传修饰区域中的完全甲基化CpG二联体的数目。

[0224] 图7是适用于本公开内容的一些实施方案的表观遗传对照核酸分子的示意图。本文描述的表观遗传对照核酸分子可以考虑核酸分子分区期间完全甲基化CpG二联体的位置特异性效应。在图7中,表观遗传对照核酸分子被分成五个子集。表观遗传对照核酸分子的表观遗传修饰区域的序列长度和序列组成在所有子集中是相同的。每个子集具有两个完全甲基化CpG二联体,但是两个完全甲基化CpG二联体的位置随子集而变化(即,两个完全甲基化CpG二联体之间的距离随子集而变化)。在子集1中,两个完全甲基化CpG二联体相距较远,而在子集4中,两个完全甲基化CpG二联体彼此非常接近。在此,标识符区域位于表观遗传修饰区域的两侧。位于两侧的标识符区域具有表观遗传状态条形码(ESB)和分子条形码(MB)。分子条形码被用作单独的表观遗传对照核酸分子的标识符,并且每种表观遗传对照核酸分子具有独特的分子条形码,即,分子1具有MB1,分子2具有MB2,分子3具有MB3,等等。基于完全甲基化CpG二联体位置的影响,这些子集将具有不同的结合亲和力。在此,表观遗传状态条形码可用于标识完全甲基化CpG二联体的位置。子集1的所有表观遗传对照核酸分子在相同位置处具有两个完全甲基化CpG二联体,因此子集1的表观遗传对照核酸分子具有相同的表观遗传状态条形码—ESB1。同样,子集2、子集3、子集4和子集5的所有表观遗传对照核酸分子分别具有表观遗传状态条形码ESB2、ESB3和ESB4。在该实例中,相同的表观遗传状态条形码位于表观遗传修饰区域的两侧。

[0225] 在另一方面,本公开内容提供了核酸的群体,该核酸的群体包含:表观遗传对照核酸分子的集合,其中表观遗传对照核酸分子的集合包括表观遗传对照核酸分子的一个或更多个子集,其中每个子集包含多于一个表观遗传对照核酸分子,并且每个表观遗传对照核酸分子包含表观遗传修饰区域;和来自受试者的多核苷酸样品中的核酸分子的集合。

[0226] 在一些实施方案中,表观遗传对照核酸分子可以具有(i)对应于λ噬菌体DNA或人类基因组的区域的序列,(ii)非天然存在的序列,和/或(iii)(i)和(ii)的组合。在一些实

施方案中,表观遗传对照核酸分子可以包含非天然存在的序列。

[0227] 在一些实施方案中,多核苷酸样品是DNA样品、RNA样品、无细胞多核苷酸样品、无细胞DNA样品或无细胞RNA样品。在一些实施方案中,多核苷酸样品是无细胞DNA样品。

[0228] 在一些实施方案中,无细胞DNA为至少1ng、至少5ng、至少10ng、至少15ng、至少20ng、至少30ng、至少50ng、至少75ng、至少100ng、至少150ng、至少200ng、至少250ng、至少300ng、至少350ng、至少400ng、至少450ng或至少500ng。

[0229] 在一些实施方案中,表观遗传对照核酸分子的量为至少1飞摩尔、至少2飞摩尔、至少5飞摩尔、至少10飞摩尔、至少15飞摩尔、至少20飞摩尔、至少50飞摩尔、至少75飞摩尔、至少100飞摩尔、至少125飞摩尔、至少150飞摩尔或至少200飞摩尔。

[0230] III.方法的一般特征

[0231] A.样品

[0232] 样品可以是从受试者分离的任何生物样品。样品可以包括身体组织、全血、血小板、血清、血浆、粪便、红细胞、白细胞(white blood cell)或白细胞(leucocyte)、内皮细胞、组织活检(例如,来自已知或疑似的实体瘤的活检)、脑脊液、滑液、淋巴液、腹水、间质液或细胞外液(例如,来自细胞间隙的流体)、齿龈液、龈沟液、骨髓、胸腔积液、脑脊液、唾液、粘液、痰、精液、汗液和尿液。样品可以是体液,诸如血液及其级分,以及尿液。这样的样品可以包括从肿瘤脱落的核酸。核酸可以包括DNA和RNA,并且可以呈双链形式和单链形式。样品可以呈最初从受试者分离的形式,或者可以经过进一步处理以去除或添加组分,诸如细胞,相对于另一种组分富集一种组分,或者将一种形式的核酸转化为另一种,诸如将RNA转化为DNA或将单链核酸转化为双链的。因此,例如,用于分析的体液可以是含有无细胞核酸例如无细胞DNA(cfDNA)的血浆或血清。

[0233] 在一些实施方案中,取自受试者的体液的样品体积取决于期望的测序区域的读段深度。体积的实例为约0.4-40毫升(mL)、约5-20mL、约10-20mL。例如,体积可以为约0.5mL、约1mL、约5mL、约10mL、约20mL、约30mL、约40mL或更多。取样血浆的体积通常在约5mL至约20mL之间。

[0234] 样品可以包含不同量的核酸。通常,特定样品中核酸的量等同于多个基因组当量。例如,约30纳克(ng)DNA的样品可以包含约10,000($10^4$)个单倍体人类基因组当量,而在cfDNA的情况下,可以包含约2000亿($2x\ 10^{11}$)个单独的多核苷酸分子。类似地,约100ng DNA的样品可以包含约30,000个单倍体人类基因组当量,而在cfDNA的情况下,可以包含约6000亿个单独的分子。

[0235] 在一些实施方案中,样品包含来自不同来源,例如,来自细胞来源和来自无细胞来源(例如,血液样品等)的核酸。通常,样品包括携带突变的核酸。例如,样品任选地包含携带种系突变和/或体细胞突变的DNA。通常,样品包含携带癌症相关突变(例如,癌症相关的体细胞突变)的DNA。

[0236] 扩增前样品中无细胞核酸的示例性量的范围通常为从约1飞克(fg)至约1微克(μg),例如,约1皮克(pg)至约200纳克(ng)、约1ng至约100ng、约10ng至约1000ng。在一些实施方案中,样品包括最多约600ng、最多约500ng、最多约400ng、最多约300ng、最多约200ng、最多约100ng、最多约50ng或最多约20ng的无细胞核酸分子。任选地,该量为至少约1fg、至少约10fg、至少约100fg、至少约1pg、至少约10pg、至少约100pg、至少约1ng、至少约10ng、至少

约100ng、至少约150ng或至少约200ng的无细胞核酸分子。在一些实施方案中,该量为最多约1fg、约10fg、约100fg、约1pg、约10pg、约100pg、约1ng、约10ng、约100ng、约150ng或约200ng的无细胞核酸分子。在一些实施方案中,方法包括从样品获得约1fg至约200ng之间的无细胞核酸分子。

[0237] 无细胞核酸通常具有长度约100个核苷酸和长度约500个核苷酸之间的大小分布,长度约110个核苷酸至长度约230个核苷酸之间的分子代表样品中约90%的分子,模式为长度约168个核苷酸(在来自人类受试者的样品中),并且第二次要峰的长度在约240个核苷酸至约440个核苷酸之间的范围内。在一些实施方案中,无细胞核酸的长度为从约160个核苷酸至约180个核苷酸,或长度为从约320个核苷酸至约360个核苷酸,或长度为从约440个核苷酸至约480个核苷酸。

[0238] 在一些实施方案中,通过分区步骤(partitioning step)从体液分离无细胞核酸,在该分区步骤中,在溶液中存在的无细胞核酸与体液中的完整细胞和其他不可溶性组分被分开。在一些实施方案中,分区包括诸如离心或过滤的技术。可选地,体液中的细胞可以被裂解,并且无细胞核酸和细胞核酸可以一起处理。通常,在添加缓冲液和洗涤步骤后,可以用例如醇来沉淀无细胞核酸。在一些实施方案中,使用另外的清洁(clean up)步骤诸如基于二氧化硅的柱来去除污染物或盐。例如,任选地在整个反应中添加非特异性批量(bulk)载体核酸以对示例性程序的诸如收率的多个方面进行优化。在这样的处理后,样品通常包含各种形式的核酸,包括双链DNA、单链DNA和/或单链RNA。任选地,单链DNA和/或单链RNA被转化成双链形式,使得它们被包括在随后的处理和分析步骤中。

[0239] B.加标签

[0240] 在一些实施方案中,核酸分子(来自多核苷酸样品)可以用样品索引和/或分子条形码(通常称为“标签”)进行加标签。标签可以通过化学合成、连接(例如,平末端连接或粘末端连接)或重叠延伸聚合酶链式反应(PCR)等方法掺入到衔接子中或以其他方式连接至衔接子。这样的衔接子可以最终连接至靶核酸分子。在其他实施方案中,通常应用一轮或更多轮扩增循环(例如,PCR扩增)来使用常规核酸扩增方法将样品索引引入核酸分子。扩增可以在一种或更多种反应混合物中进行(例如,阵列中的多于一个微孔)。分子条形码和/或样品索引可以同时引入或以任何顺序引入。在一些实施方案中,在进行序列捕获步骤之前和/或之后引入分子条形码和/或样品索引。在一些实施方案中,在探针捕获之前仅引入分子条形码,并且在进行序列捕获步骤之后引入样品索引。在一些实施方案中,在进行基于探针的捕获步骤之前,引入分子条形码和样品索引两者。在一些实施方案中,在进行序列捕获步骤之后引入样品索引。在一些实施方案中,通过衔接子经由连接(例如,平末端连接或粘末端连接)将分子条形码掺入样品中的核酸分子(例如cfDNA分子)。在一些实施方案中,通过重叠延伸聚合酶链式反应(PCR)将样品索引掺入样品中的核酸分子(例如cfDNA分子)。通常,序列捕获方案包括引入与靶核酸序列互补的单链核酸分子,所述靶核酸序列例如基因组区域的编码序列,并且这样的区域的突变与癌症类型相关。

[0241] 在一些实施方案中,标签可以位于样品核酸分子的一个末端或两个末端。在一些实施方案中,标签是预定或随机或半随机序列的寡核苷酸。在一些实施方案中,标签的长度可以小于约500个、200个、100个、50个、20个、10个、9个、8个、7个、6个、5个、4个、3个、2个或1个核苷酸。标签可以随机或非随机地连接至样品核酸。

[0242] 在一些实施方案中,每个样品被用样品索引或样品索引的组合独特地加标签。在一些实施方案中,样品或子样品的每个核酸分子被用分子条形码或分子条形码的组合独特地加标签。在其他实施方案中,可以使用多于一个条形码,使得分子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(例如,非独特分子条形码)。在这些实施方案中,分子条形码通常与个体分子附接(例如,通过连接),使得分子条形码和可以与其附接的序列的组合产生可以被单独地追溯的独特序列。检测非独特地加标签的分子条形码与内源序列信息(例如,对应于样品中原始核酸分子序列的开始(起始)和/或结束(终止)基因组座位(location)/位置(position),在一个或两个末端处的序列读段的子序列、序列读段的长度和/或样品中原始核酸分子的长度)的组合,通常允许将独特身份分配给特定分子。在一些实施方案中,检测非独特地加标签的分子条形码与内源序列信息(例如,序列读段与参考序列对齐的开始(起始)和/或结束(终止)区域、在一个或两个末端处的序列读段的子序列、序列读段的长度和/或样品中原始核酸分子的长度)的组合,通常允许将独特身份分配给特定分子。在一些实施方案中,开始区域包括测序读段的基因组起始位置,在该位置处测序读段的5'末端被确定为开始与参考序列对齐,而结束区域包括测序读段的基因组终止位置,在该位置处测序读段的3'末端被确定为终止与参考序列对齐。在一些实施方案中,开始区域包括与参考序列对齐的测序读段的5'末端处的前1个、前2个、前5个、前10个、前15个、前20个、前25个、前30个或至少前30个碱基位置。在一些实施方案中,结束区域包括与参考序列对齐的测序读段的3'末端的最后1个、最后2个、最后5个、最后10个、最后15个、最后20个、最后25个、最后30个或至少最后30个碱基位置。

[0243] 个体序列读段的长度或碱基对数目也任选地用于将独特身份分配至给定的分子。如本文描述的,来自已经被分配了独特身份的核酸单链的片段可以从而允许随后鉴定来自亲本链和/或互补链的片段。

[0244] 在一些实施方案中,以一组标识符(例如,独特或非独特的分子条形码的组合)与样品中分子的预期比率引入分子条形码。一种示例形式使用连接至靶分子两端的约2个至约1,000,000个不同的分子条形码序列、或约5个至约150个不同的分子条形码序列、或约20个至约50个不同的分子条形码序列。可选地,可以使用约25个至约1,000,000个不同的分子条形码序列。例如,可以使用20-50×20-50个分子条形码序列(即,20-50个不同的分子条形码序列之一可以附接至靶分子的每一端)。这种数目的标识符通常足以使具有相同起点和终点的不同分子具有接收不同的标识符组合的高概率(例如,至少94%、99.5%、99.99%或99.999%)。在一些实施方案中,约80%、约90%、约95%或约99%的分子具有相同的分子条形码组合。

[0245] 在一些实施方案中,反应中独特或非独特的分子条形码的分配使用例如美国专利申请第20010053519、第20030152490号和第20110160078号、以及美国专利第6,582,908号、第7,537,898号、第9,598,731号和第9,902,992号中描述的方法和系统来进行,在此将它们中的每一个通过引用以其整体并入。可选地,在一些实施方案中,样品的不同核酸分子可以仅使用内源序列信息(例如,起始和/或终止位置、序列一端或两端的子序列和/或长度)来鉴定。

[0246] 表观遗传状态条形码(ESB)是一种附接至表观遗传对照核酸分子的表观遗传修饰区域的标签。ESB可以用作表观遗传对照核酸分子的表观遗传状态的标识符。它可以指表观

遗传对照核酸分子的表观遗传修饰区域中具有表观遗传修饰的核苷酸的数目。在一些实施方案中,表观遗传对照核酸分子的标识符区域包含至少一个表观遗传状态条形码。在一些实施方案中,ESB是表观遗传对照核酸分子的标识符区域的一部分。例如,如果表观遗传修饰是DNA甲基化,并且表观遗传对照核酸分子的一个子集具有5个甲基化核苷酸,那么该子集内的所有表观遗传对照核酸分子将具有相同的表观遗传状态条形码。在一些实施方案中,表观遗传状态条形码可用于标识表观遗传对照核酸分子的表观遗传修饰区域的表观遗传修饰水平/程度。表观遗传对照核酸分子可以根据表观遗传修饰区域中胞嘧啶或CpG核苷酸的数目分为子集。在一些实施方案中,在每个子集中,甲基化水平可以不同(例如,高度甲基化、中等甲基化和低度甲基化),并且每个甲基化水平可以具有单独的表观遗传状态条形码。例如,在子集A中,所有低度甲基化的表观遗传对照核酸分子将具有一种表观遗传状态条形码,例如ESB1,并且所有高度甲基化的表观遗传对照核酸分子将具有另一种表观遗传状态条形码,例如ESB3。在该实例中,表观遗传状态条形码用于标识甲基化的水平/程度。

[0247] 在一些实施方案中,反应中独特或非独特的分子条形码的分配使用例如美国专利申请第20010053519、第20030152490号和第20110160078号、以及美国专利第6,582,908号、第7,537,898号、第9,598,731号和第9,902,992号中描述的方法和系统来进行,在此将它们中的每一个通过引用以其整体并入。

[0248] C.扩增

[0249] 样品核酸可以侧接衔接子,并且使用与待扩增的DNA分子侧翼的衔接子中的引物结合位点结合的核酸引物通过PCR和其它扩增方法来扩增。在一些实施方案中,扩增方法包括由热循环产生的延伸、变性和退火的循环,或者可以是等温的,例如,在转录介导的扩增中。可以任选地利用的扩增方法的其他实例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身持续序列的复制(self-sustained sequence-based replication)。

[0250] 通常,扩增反应生成多于一个非独特或独特地加标签的核酸扩增子,其分子条形码和样品索引的大小范围为约150个核苷酸(nt)至约700nt、250nt至约350nt或约320nt至约550nt。在一些实施方案中,扩增子具有约180nt的大小。在一些实施方案中,扩增子具有约200nt的大小。

[0251] D.富集

[0252] 在一些实施方案中,在对核酸测序之前富集序列。任选地对特定靶区域进行富集或对(“靶序列”)进行非特异性地富集。在一些实施方案中,感兴趣的靶区域可以用针对一种或更多种诱饵集组(one or more bait set panels)选择的核酸捕获探针(“诱饵(bait)”)使用差异性平铺和捕获方案来富集。差异性平铺和捕获方案通常使用不同相对浓度的诱饵集以在遍及与诱饵相关的基因组区域中差异性平铺(例如,以不同的“分辨率”),经受一组限制(例如,测序仪限制,诸如测序载量、每种诱饵的效用等),并以下游测序所需的水平捕获靶核酸。这些感兴趣的靶基因组区域任选地包括天然核苷酸序列或核酸构建体的合成核苷酸序列。在一些实施方案中,具有针对一个或更多个感兴趣区域的探针的生物素标记的珠可以用于捕获靶序列,并且任选地随后扩增这些区域,以富集感兴趣区域。

[0253] 序列捕获通常包括使用与靶核酸序列杂交的寡核苷酸探针。在一些实施方案中,探针设定策略包括将探针平铺在感兴趣的区域内。这样的探针的长度可以为,例如,约60个至约120个核苷酸。该集可以具有约2X、3X、4X、5X、6X、7X、8X、9X、10X、15X、20X、50X或多于

50X的深度(例如,覆盖深度)。序列捕获的有效性通常部分地取决于靶分子中与探针序列互补(或几乎互补)的序列的长度。

[0254] E.测序

[0255] 通常对任选地侧翼有衔接子、有或没有预先扩增的样品核酸进行测序。任选地利用的测序方法或商业上可得的形式包括,例如,Sanger测序、高通量测序、焦磷酸测序、合成测序、单分子测序、基于纳米孔的测序、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纳米孔、Roche Genia、Maxim-Gilbert测序、引物步移、使用PacBio、SOLiD、Ion Torrent或纳米孔平台测序。测序反应可以在多种样品处理单元中进行,所述单元可以包括多行道(multiple lane)、多通道、多孔或基本上同时处理多个样品集的其他装置。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。

[0256] 可以对一种或更多种已知包含癌症或其他疾病的标志物的核酸片段类型或区域进行测序反应。也可以对样品中存在的任何核酸片段进行测序反应。可以对基因组的至少约5%、10%、15%、20%、25%、30%、40%、50%、60%、70%、80%、90%、95%、99%、99.9%或100%进行测序反应。在其他情况下,可以对基因组的少于约5%、10%、15%、20%、25%、30%、40%、50%、60%、70%、80%、90%、95%、99%、99.9%或100%进行测序反应。

[0257] 可以使用多重测序技术进行同时测序反应。在一些实施方案中,用至少约1000个、2000个、3000个、4000个、5000个、6000个、7000个、8000个、9000个、10000个、50000个或100,000个测序反应对无细胞多核苷酸进行测序。在其他实施方案中,用少于约1000个个、2000个、3000个、4000个、5000个、6000个、7000个、8000个、9000个、10000个、50000个或100,000个测序反应对无细胞多核苷酸进行测序。测序反应通常顺序性地进行或同时进行。随后的数据分析通常对全部或部分的测序反应进行。在一些实施方案中,对至少约1000个、2000个、3000个、4000个、5000个、6000个、7000个、8000个、9000个、10000个、50000个或100,000个测序反应进行数据分析。在其他实施方案中,对少于约1000个、2000个、3000个、4000个、5000个、6000个、7000个、8000个、9000个、10000个、50000个或100,000个测序反应进行数据分析。读段深度的一种实例是每个基因座(例如,碱基位置)约1000个至约50000个读段。

[0258] F.分析

[0259] 测序可以产生多于一个测序读段或读段(sequencing reads or reads)。测序读段或读段可以包括长度小于约150个碱基或长度小于约90个碱基的核苷酸序列的数据。在一些实施方案中,读段的长度在约80个碱基和约90个碱基之间,例如,约85个碱基。在一些实施方案中,本公开内容的方法被应用于非常短的读段,例如,长度小于约50个碱基或约30个碱基。测序读段数据可以包括序列数据以及元信息。序列读段数据可以以任何合适的文件格式存储,包括例如,VCF文件、FASTA文件或FASTQ文件。

[0260] FASTA可以指用于检索序列数据库的计算机程序,并且名称FASTA也可以指标准文件格式。例如,FASTA由例如Pearson&Lipman,1988,Improved tools for biological sequence comparison,PNAS 85:2444-2448描述,在此将其通过引用以其整体并入。FASTA格式的序列以单行描述开始,随后为序列数据行。描述行通过第一列中的大于(“>”)符号与序列数据区分开。“>”符号后面的词是序列的标识符,并且该行的其余部分是描述(都是任

选的)。在“>”和标识符的第一个字母之间不应当有空格。建议文本的所有行少于80个字符。如果出现以“>”开头的另一行,则序列结束;这指示另一个序列的开始。

[0261] FASTQ格式是基于文本的格式,用于存储生物序列(通常是核苷酸序列)及其对应的质量评分。它与FASTA格式相似,但是在序列数据之后具有质量评分。为简洁起见,序列字母和质量评分都使用单个ASCII字符编码。FASTQ格式是用于存储高通量测序仪器诸如Illumina Genome Analyzer的输出结果的约定俗成的标准,例如Cock等人(“The Sanger FASTQ file format for sequences with quality scores,and the Solexa/Illumina FASTQ variants,”Nucleic Acids Res 38(6):1767-1771,2009)所描述的,在此将其通过引用以其整体并入。

[0262] 对于FASTA和FASTQ文件,元信息包括描述行但不包括序列数据行。在一些实施方案中,对于FASTQ文件,元信息包括质量评分。对于FASTA和FASTQ文件,序列数据在描述行之后开始,并且通常使用一些任选地带有“-”的IUPAC模糊代码的子集呈现。在一种实施方案中,序列数据可以使用A、T、C、G和N字符,任选地根据需要包括“-”或者包括U(例如,以表示空位或尿嘧啶)。

[0263] 在一些实施方案中,至少一个主序列读段文件和输出文件被存储为纯文本文件(例如,使用诸如ASCII、ISO/IEC 646、EBCDIC、UTF-8或UTF-16的编码)。本公开内容提供的计算机系统可以包括能够打开纯文本文件的文本编辑器程序。文本编辑器程序可以指能够在计算机屏幕上呈现文本文件(诸如纯文本文件)的内容、允许人员编辑文本(例如使用显示器、键盘和鼠标)的计算机程序。文本编辑器的实例包括但不限于Microsoft Word、emacs、pico、vi、BBEdit和TextWrangler。文本编辑器程序可以能够以人类可读格式在计算机屏幕上显示纯文本文件,显示元信息和序列读段(例如,不是二进制编码而是使用字母数字字符,因为它们可以用于打印或人类书写)。

[0264] 虽然已经参照FASTA或FASTQ文件讨论了方法,但是本公开内容的方法和系统可以用于压缩任何合适的序列文件格式,包括例如Variant Call Format(VCF)格式的文件。典型的VCF文件可以包括标题部分和数据部分。标题包含任何数目的元信息行,每行都以字符‘##’开始,以及以单个‘#’字符开始的TAB分隔字段定义行。字段定义行命名了八个必填列,而主体部分包含填充了这些字段定义行定义的列的数据行。VCF格式由例如Danecek等人(“The variant call format and VCFtools,”Bioinformatics27(15):2156-2158,2011)描述,在此将其通过引用以其整体并入。标题部分可以被视为要写入压缩文件的元信息,并且数据部分可以被视为行,其中每一行只有在为独特的情况下才会被存储在主文件中。

[0265] 一些实施方案提供了测序读段的装配。例如,在通过比对的装配中,将测序读段彼此比对或与参考序列比对。通过比对每个读段,继而与参考基因组比对,所有读段被按照关于彼此的关系定位以创建装配体。另外,将测序读段与参考序列比对或映射至参考序列也可以用于鉴定测序读段中的变异序列。鉴定变异序列可以与本文描述的方法和系统组合使用,以进一步帮助疾病或状况的诊断或预后或用于指导治疗决定。

[0266] 在一些实施方案中,任何或全部步骤是自动化的。可选地,本公开内容的方法可以全部或部分地在一个或更多个专用程序中实现,例如每一个任选地以编译语言诸如C++写入,然后以二进制编译和分发。本公开内容的方法可以全部或部分地作为现有序列分析平台内的模块或通过调用现有序列分析平台内的功能而实现。在一些实施方案中,本公开内

容的方法包括响应于单个启动队列(例如,源自人类活动、另一个计算机程序或机器的触发事件中的一个事件或事件组合)而都被自动调用的多个步骤。因此,本公开内容提供了其中任何步骤或步骤的任何组合可以响应于队列而自动发生的方法。“自动地”通常意指不介入人类输入、影响或交互(例如,仅响应于原来的或预先排队的人类活动)。

[0267] 本公开内容的方法还可以包括多种形式的输出,所述多种形式的输出包括对受试者的核酸样品的准确和灵敏的解释。检索的输出可以以计算机文件的格式提供。在一些实施方案中,输出是FASTA文件、FASTQ文件或VCF文件。输出可以被处理以产生含有序列数据诸如与参考基因组的序列比对的核酸序列的文本文件或XML文件。在其他实施方案中,处理产生包含坐标或描述受试者核酸中相对于参考基因组的一个或更多个突变的字串的输出。比对字串可以包括Simple UnGapped Alignment Report(SUGAR)、Verbose Useful Labeled Gapped Alignment Report(VALGAR)和Compact Idiosyncratic Gapped Alignment Report(CIGAR)(例如,Ning等人,Genome Research 11(10):1725-9,2001描述的,在此将其通过引用以其整体并入)。这些字串可以例如在来自European Bioinformatics Institute(Hinxton,UK)的Exonerate序列比对软件中实现。

[0268] 在一些实施方案中,产生包含CIGAR字串的序列比对—诸如,例如序列比对图(SAM)或二元比对图(BAM)文件(SAM格式在例如Li等人,“The Sequence Alignment/Map format and SAMtools,”Bioinformatics,25(16):2078-9,2009中描述,在此将其通过引用以其整体并入)。在一些实施方案中,CIGAR显示或包括每行一个空位的比对。CIGAR是一种报告为CIGAR字串的压缩的成对比对格式。CIGAR字串可以用于呈现长的(例如,基因组)成对比对。CIGAR字串可以在SAM格式中使用以表示读段与参考基因组序列的比对。

[0269] CIGAR字串可以遵循建立的基序。每个字符前面是数字,给出事件的碱基计数。使用的字符可以包括M、I、D、N和S(M=匹配;I=插入;D=缺失;N=空位;S=取代)。CIGAR字串定义匹配/不匹配和缺失(或空位)的序列。例如,CIGAR字串2MD3M2D2M可以指示,比对包含2个匹配、1个缺失(为了节省一些空间省略数字1)、3个匹配、2个缺失和2个匹配。

[0270] 在一些实施方案中,通过在一端或两端具有单链突出端的双链核酸上酶促形成平末端来制备用于测序的核酸群体。在这些实施方案中,在核苷酸(例如,A、C、G和T或U)存在的情况下,通常用具有5’-3’DNA聚合酶活性和3’-5’核酸外切酶活性的酶处理该群体。可以任选地使用的酶或其催化片段的实例包括Klenow大片段和T4聚合酶。在5’突出端处,酶通常延伸相对链上凹陷的3’端,直到它与5’端齐平以产生平末端。在3’突出端处,酶通常从3’端消化,达到相对链的5’端并且有时超过相对链的5’端。如果该消化行进超过了相对链的5’端,则缺口可以通过具有与对5’突出端使用的具有相同的聚合酶活性的酶填补。双链核酸上平末端的形成有利于例如衔接子的附接和随后的扩增。

[0271] 在一些实施方案中,核酸群体经受另外的处理,诸如将单链核酸转化为双链核酸和/或将RNA转化为DNA(例如,互补DNA或cDNA)。这些形式的核酸还任选地与衔接子连接并扩增。

[0272] 在具有或没有预先扩增的情况下,经受上文描述的形成平末端的处理的核酸以及任选地样品中的其它核酸,可以被测序以产生测序的核酸。测序的核酸可以指核酸的序列(例如,序列信息)或其序列已被确定的核酸。可以进行测序,以便从样品中个体核酸分子的扩增产物的共有序列直接或间接地提供样品中个体核酸分子的序列数据。

[0273] 在一些实施方案中,样品中具有单链突出端的双链核酸在平末端形成后,在两端处被与包含条形码的衔接子连接,并且测序确定了核酸序列以及通过衔接子引入的直线连接的(in-line)条形码。平末端DNA分子任选地与至少部分双链的衔接子(例如,Y形或钟形衔接子)的平末端连接。可选地,样品核酸和衔接子的平末端可以用互补核苷酸加尾以促进连接(例如,粘末端连接)。

[0274] 通常使核酸样品与足够数目的衔接子接触,使得相同核酸的任何两个拷贝从连接在两端的衔接子接收相同衔接子条形码组合的概率较低(例如,小于约1%或0.1%)。以这种方式使用衔接子可以允许对在参考核酸上具有相同的起点和终点并且被连接至相同条形码组合的核酸序列家族的鉴定。这样的家族可以代表扩增前的样品中的核酸的扩增产物序列。可以对家族成员的序列进行汇编,以获得原始样品中的核酸分子的共有核苷酸或完整的共有序列,所述核酸分子通过平末端形成和衔接子附接被修饰。换言之,占据样品中核酸的特定位置的核苷酸可以被确定为占据家族成员序列中对应位置的核苷酸的共有核苷酸。家族可以包括双链核酸的一条链或两条链的序列。如果家族的成员包括来自双链核酸的两条链的序列,则为了对序列汇编以获得共有核苷酸或序列的目的,一条链的序列可以被转化为它们的互补序列。一些家族仅包含单个成员序列。在该情况下,该序列可以被视为扩增前样品中的核酸的序列。可选地,仅有单个成员序列的家族可以从随后的分析中排除。

[0275] 通过将测序的核酸与参考序列进行比较,可以确定测序的核酸中的核苷酸变异(例如,SNV或插入和/或缺失)。参考序列通常是已知序列,例如,来自受试者的已知的全部或部分的基因组序列(例如,人类受试者的全基因组序列)。参考序列可以是,例如,hG19或hG38。如上文描述的,测序的核酸可以代表直接确定的样品中的核酸的序列或这种核酸的扩增产物的共有序列。可以在参考序列上的一个或更多个指定位置处进行比较。当相应的序列被最大程度地比对时,可以鉴定测序的核酸的子集,该子集包括与参考序列的指定位置对应的位置。在这样的子集中,可以确定哪些(如果有的话)测序的核酸在指定位置处包含核苷酸变异,以及任选地哪些(如果有的话)包含参考核苷酸(例如,与参考序列中的相同)。如果包含核苷酸变异的子集中的测序的核酸的数目超过选定的阈值,则变异核苷酸可以在指定位置被调用。阈值可以是简单的数字,诸如包含核苷酸变异的子集中的至少1个、2个、3个、4个、5个、6个、7个、8个、9个或10个测序的核酸,或者阈值可以是包含核苷酸变异的子集中的测序的核酸的比率,诸如至少0.5、1、2、3、4、5、10、15或20,以及其他可能性。可以对参考序列中感兴趣的任何指定位置重复比较。有时可以对占据参考序列上至少约20个、100个、200个或300个连续位置,例如,约20-500个或约50-300个连续位置的指定位置进行比较。

[0276] 关于核酸测序的另外的细节,包括本文描述的形式和应用,还提供在以下文献中:例如,Levy等人,Annual Review of Genomics and Human Genetics,17:95-115(2016);Liu等人,J.of Biomedicine and Biotechnology,Volume 2012,Article ID 251364:1-11(2012);Voelkerding等人,Clinical Chem.,55:641-658(2009);MacLean等人,Nature Rev.Microbiol.,7:287-296(2009),Astier等人,J Am Chem Soc.,128(5):1705-10(2006);美国专利第6,210,891号、美国专利第6,258,568号、美国专利第6,833,246号、美国专利第7,115,400号、美国专利第6,969,488号、美国专利第5,912,148号、美国专利第6,130,073号、美国专利第7,169,560号、美国专利第7,282,337号、美国专利第7,482,120号、

美国专利第7,501,245号、美国专利第6,818,395号、美国专利第6,911,345号、美国专利第7,501,245号、美国专利第7,329,492号、美国专利第7,170,050号、美国专利第7,302,146号、美国专利第7,313,308号和美国专利第7,476,503号,在此将这些中的每一个通过引用以其整体并入。

[0277] IV.计算机系统

[0278] 本公开内容的方法可以使用或借助于计算机系统来实现。例如,这样的方法可以用计算机处理器执行,所述方法可以包括(a)向多核苷酸样品中的核酸分子添加表观遗传对照核酸分子的集合,从而产生加标样品;(b)将加标样品的核酸分子分区成多于一个分区的集合;(c)富集来自多于一个分区的集合的分子的子集以生成多于一个富集的分子,其中多于一个富集的分子包括一组表观遗传对照核酸分子和来自多核苷酸样品的一组核酸分子;(d)对多于一个富集的分子测序以产生多于一个测序读段;(e)分析多于一个测序读段以生成表观遗传对照核酸分子的多于一个表观遗传分区评分;以及(f)将多于一个表观遗传分区评分与多于一个表观遗传分区截止值进行比较。在该实施方案中,该系统包括用于添加表观遗传对照核酸分子、分区、富集和测序的组件。

[0279] 在另一种实施方案中,一种用于评价多核苷酸样品中核酸分子基于表观遗传状态的分区方法的系统,该系统包括:通信接口,该通信接口通过通信网络接收由核酸测序仪生成的加标样品的测序读段的集合,其中测序读段的集合包括(i)从来源于样品的多核苷酸生成的测序读段的至少第一群体,其中来自第一群体的测序读段包括标签序列和衍生自来源于样品的多核苷酸的序列;和(ii)从表观遗传对照核酸分子生成的测序读段的至少第二群体,其中从第二群体生成的测序读段包含表观遗传修饰区域和任选地标识符区域;计算机,该计算机与通信接口通信,其中该计算机包括一个或更多个计算机处理器和计算机可读介质,该计算机可读介质包括机器可执行代码,该机器可执行代码在由一个或更多个计算机处理器执行后实现一种方法,该方法包括:(i)通过通信网络接收来自核酸测序仪的测序读段的第一群体和第二群体的测序读段的集合;(ii)分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子和/或内源对照分子的一个或更多个表观遗传分区评分;以及(iii)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0280] 在另一种实施方案中,一种系统,该系统包括控制器,该控制器包括计算机可读介质或能够访问计算机可读介质,该计算机可读介质包括非暂时性计算机可执行指令,当所述非暂时性计算机可执行指令由至少一个电子处理器执行时,至少进行:(a)获得由核酸测序仪生成的加标样品的测序读段的集合,其中加标样品包含样品多核苷酸和表观遗传对照核酸分子,并且测序读段的集合包含(i)从样品多核苷酸生成的测序读段的第一群体和(ii)从表观遗传对照核酸分子生成的测序读段的第二群体;(b)分析测序读段的集合的至少一个子集,以生成表观遗传对照核酸分子和/或内源对照分子的一个或更多个表观遗传分区评分;和(c)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0281] 在另一种实施方案中,一种系统,该系统包括控制器,该控制器包括计算机可读介质或能够访问计算机可读介质,该计算机可读介质包括非暂时性计算机可执行指令,当所述非暂时性计算机可执行指令由至少一个电子处理器执行时,至少进行:(a)获得由核酸测

序仪生成的样品的测序读段的集合,其中测序读段的集合包括从样品多核苷酸生成的测序读段;(b)分析测序读段的集合的至少一个子集以生成内源对照分子的一个或更多个表观遗传分区评分;和(c)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0282] 在一些实施方案中,该系统还包括g)基于表观遗传分区评分的比较来生成分区方法的结果状态。在一些实施方案中,分区方法的结果状态分类为(i)如果表观遗传对照核酸分子的一个或更多个表观遗传分区评分和/或内源对照分子的集合的一个或更多个表观遗传分区评分在对应的表观遗传分区截止值内,则该分区方法成功;或者(ii)如果表观遗传对照分子和/或内源对照分子的一个或更多个表观遗传分区评分中的至少一个在对应的表观遗传分区截止值之外,则该分区方法不成功。

[0283] 图8示出了被编程或以其他方式配置成实现本公开内容的方法的计算机系统801。计算机系统801可以控制样品制备、测序和/或分析的各方面。在一些实例中,计算机系统801被配置成执行样品制备和样品分析,包括核酸测序。

[0284] 计算机系统801包括中央处理单元(CPU,本文中也被为“处理器”和“计算机处理器”)805,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。计算机系统801还包括存储器或存储器位置810(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815(例如,硬盘)、用于与一个或更多个其他系统进行通信的通信接口820(例如,网络适配器)和外围设备825,诸如高速缓冲存储器(cache)、其他存储器、数据存储和/或电子显示适配器。存储器810、储存单元815、接口820和外围设备825与CPU 805通过通信网络或总线(实线路),诸如主板(motherboard)通信。存储单元815可以是用于存储数据的数据存储单元(或数据储存库)。计算机系统801可以借助于通信接口820被可操作地耦合至计算机网络830。计算机网络830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,计算机网络830为电信和/或数据网络。计算机网络830可以包括一个或更多个计算机服务器,这可以启动分布式计算,诸如云计算。在一些情况下,借助于计算机系统801,计算机网络830可以实现对等网络(peer-to-peer network),其可以启动耦合至计算机系统801的设备作为客户端或服务器运行。

[0285] CPU 805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810中。由CPU 805进行的操作的实例可以包括读取、解码、执行和写回。

[0286] 存储单元815可以存储文件,诸如驱动程序、库和保存的程序。存储单元815可以存储用户生成的程序和记录的会话以及与程序相关的输出。存储单元815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,计算机系统801可以包括一个或更多个另外的数据存储单元,该另外的数据存储单元在计算机系统801的外部,诸如位于通过内联网或因特网与计算机系统801通信的远程服务器上。可以使用例如通信网络或物理数据传输器(例如,使用硬盘驱动器、拇指驱动器或其他数据存储机制)将数据从一个位置传输到另一个位置。

[0287] 计算机系统801可以与一个或更多个远程计算机系统通过网络830进行通信。对于实施方案,计算机系统801可以与用户(例如,操作者)的远程计算机系统进行通信。远程计算机系统的实例包括个人计算机(例如,便携式PC)、板式(slate)或平板PC(例如,Apple®

iPad、Samsung®Galaxy Tab)、电话、智能电话(例如,Apple®iPhone、Android支持的设备、Blackberry®)或个人数字助手。用户可以经由网络830访问计算机系统801。

[0288] 如本文描述的方法可以通过机器(例如,计算机处理器)可执行代码的方式实现,该机器可执行代码被存储在计算机系统801的电子存储位置,诸如,例如存储器810或电子存储单元815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805执行。在一些情况下,代码可以从存储单元815检索并存储在存储器810上,以便于处理器805即时访问。在一些情况下,可以不包括电子存储单元815,而将机器可执行指令存储于存储器810上。

[0289] 在一方面,本公开内容提供了非瞬时性计算机可读介质,所述非瞬时性计算机可读介质包含计算机可执行指令,当所述计算机可执行指令被至少一个电子处理器执行时,执行包括以下步骤的方法:(a)获得由核酸测序仪生成的测序读段的集合;(b)分析测序读段的集合的至少一个子集以生成表观遗传对照核酸分子的一个或更多个表观遗传分区评分;以及(c)将一个或更多个表观遗传分区评分与一个或更多个表观遗传分区截止值进行比较。

[0290] 代码可以被预编译并配置成用于与具有适于执行该代码的处理器的机器一起使用或可以在运行时间期间被编译。代码可以以编程语言的形式提供,该编程语言可以被选择使得代码能够以预编译的或按原来编译(as-compiled)的方式被执行。

[0291] 本文所提供的系统和方法的各方面,诸如计算机系统701,可以体现在编程中。技术的各方面可以被认为是通常被携带在一种类型的机器可读介质或以一种类型的机器可读介质体现的机器(或处理器)可执行代码和/或相关数据的形式的“产品”或“制品(articles of manufacture)”。机器可执行代码可以被存储于电子存储单元诸如存储器(例如,只读存储器、随机存取存储器、闪速存储器)或硬盘上。“存储”型介质可以包括计算机、处理器等或其相关模块,诸如多种半导体存储器、磁带驱动器、磁盘驱动器等的任一种或全部有形存储器,其可以在任何时间为软件编程提供非暂时性存储。

[0292] 软件的全部或一部分有时可以通过互联网或多种其他通信网络进行通信。例如,此类通信可以使得将软件从一个计算机或处理器加载到另一个计算机或处理器中,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中。因此,可以携带软件元件的另一类型的介质包括诸如那些在本地设备之间跨物理界面、通过有线和光纤陆线网络以及在多种空中链路(air-link)上使用的光波、电波和电磁波。携带此类波的物理元件,诸如有线或无线链路、光链路等,也可被认为是携带软件的介质。如本文使用的,除非被限制为非暂时性的、有形的“存储”介质,否则术语诸如计算机或机器“可读介质”是指参与将指令提供至处理器以便执行的任何介质。

[0293] 因此,机器可读介质,诸如计算机可执行代码,可以采取许多形式,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,诸如附图中示出的在任何计算机等中的任何存储设备,诸如可以用于实现数据库等。易失性存储介质包括动态存储器,诸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的总线的导线。载波传输介质可采取电信号或电磁信号或者声波或光波的形式,诸如在射频(RF)和红外(IR)数据通信期间生成的那些。因此,计算机可读介质的常见形式包括例如:软盘(floppy disk)、软磁盘(flexible disk)、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、穿孔卡片、纸带、具有孔模式的任何其他物理存储介质、RAM、ROM、PROM和EPROM、FLASH-EPROM、任何其他存储器芯片或盒、传输数据或指令的载波、传输此类载波的缆线或链路,或者计算机可以从其读取编程代码和/或数据的任何其他介质。这些形式的计算机可读介质中的许多可以参与将一串或更多串的一个或更多个指令运送至处理器以便执行。

[0294] 计算机系统801可以包括电子显示器或与之通信,该电子显示器包括用户界面(UI),以便提供例如样品分析的一个或更多个结果。UI的实例包括但不限于图形用户界面(GUI)和基于网络的用户界面。

[0295] 关于计算机系统和网络、数据库以及计算机程序产品的另外的细节还提供于以下文献中:例如,Peterson,Computer Networks:A Systems Approach,Morgan Kaufmann,第5版(2011);Kurose,Computer Networking:ATop-Down Approach,Pearson,第7版(2016),Elmasri,Fundamentals of Database Systems,Addison Wesley,第6版(2010),Coronel,Database Systems:Design,Implementation,&Management,Cengage Learning,第11版(2014),Tucker,Programming Languages,McGraw-Hill Science/Engineering/Math,第2版(2006);和Rhoton,Cloud Computing Architected:Solution Design Handbook,Recursive Press(2011),将这些中的每一项在此通过引用以其整体并入。

[0296] V.应用

[0297] A.癌症和其他疾病

[0298] 在一些实施方案中,本文公开的方法和系统可以基于将核酸变异分类为体细胞来源或种系来源而用于鉴定定制或靶向的疗法以治疗患者的特定疾病或状况。通常,所考虑的疾病是一种类型的癌症。这样的癌症的非限制性实例包括胆道癌、膀胱癌、移行细胞癌、尿路上皮癌、脑癌、神经胶质瘤、星形细胞瘤、乳腺癌、化生癌、宫颈癌、宫颈鳞状细胞癌、直肠癌、结肠直肠癌、结肠癌、遗传性非息肉性结肠直肠癌、结肠腺癌、胃肠间质瘤(GIST)、子宫内膜癌、子宫内膜间质肉瘤、食管癌、食管鳞状细胞癌、食管腺癌、眼黑素瘤、葡萄膜黑素瘤、胆囊癌、胆囊腺癌、肾细胞癌、透明细胞肾细胞癌(clear cell renal cell carcinoma)、移行细胞癌、尿路上皮癌、肾母细胞瘤、白血病、急性淋巴细胞白血病(ALL)、急性髓性白血病(AML)、慢性淋巴细胞白血病(CLL)、慢性髓性白血病(CML)、慢性粒单核细胞白血病(CMML)、肝癌(liver cancer)、肝癌(liver carcinoma)、肝细胞瘤、肝细胞癌、胆管癌、肝母细胞瘤、肺癌、非小细胞肺癌(NSCLC)、间皮瘤、B细胞淋巴瘤、非霍奇金淋巴瘤、弥漫性大B细胞淋巴瘤、套细胞淋巴瘤、T细胞淋巴瘤、非霍奇金淋巴瘤、前体T淋巴母细胞淋巴瘤/白血病、外周T细胞淋巴瘤、多发骨髓瘤、鼻咽癌(NPC)、神经母细胞瘤、口咽癌、口腔鳞状细胞癌、骨肉瘤、卵巢癌、胰腺癌、胰腺导管腺癌、假乳头状肿瘤、泡细胞癌。前列腺癌、前列腺腺癌、皮肤癌、黑素瘤、恶性黑素瘤、皮肤黑素瘤、小肠癌、胃癌(stomach cancer)、胃癌(gastric carcinoma)、胃肠间质瘤(GIST)、子宫癌或子宫肉瘤。

[0299] 任选地使用本文公开的方法和系统评估的其它基于遗传的疾病、紊乱或状况的非限制性实例包括软骨发育不全、α-1抗胰蛋白酶缺乏症、抗磷脂综合征、孤独症、常染色体显性多囊肾病、夏科-马里-图思病(CMT)、猫叫综合征、克罗恩病、囊性纤维化、Dercum病、唐氏综合征、Duane综合征、杜兴氏肌营养不良症、因子V Leiden易栓症、家族性高胆固醇血症、家族性地中海热、脆性X综合征、戈谢病、血色素沉着病、血友病、全前脑畸形、亨廷顿病、克

兰费尔特综合征、马方综合征、强直性肌营养不良、神经纤维瘤病、努南综合征、成骨不全、帕金森病、苯丙酮尿症、Poland异常、卟啉症、早老症、视网膜色素变性、重症联合免疫缺陷病(scid)、镰状细胞病、脊髓性肌萎缩症、泰-萨克斯病、地中海贫血、三甲基胺尿症、特纳综合征、颚心脸综合征(velocardiofacial syndrome)、WAGR综合征、威尔逊病等。

[0300] B.治疗和相关管理

[0301] 在某些实施方案中,本文公开的方法涉及鉴于核酸变异为体细胞来源或种系来源的状态,鉴定定制疗法并向患者施用定制疗法。在一些实施方案中,基本上任何癌症疗法(例如,手术疗法、放射疗法、化疗疗法和/或类似疗法)都可以被包括为这些方法的一部分。通常,定制疗法包括至少一种免疫疗法(或免疫治疗剂)。免疫疗法通常是指增强针对特定癌症类型的免疫应答的方法。在某些实施方案中,免疫疗法是指增强针对肿瘤或癌症的T细胞应答的方法。

[0302] 在某些实施方案中,来自受试者的样品的核酸变异为体细胞来源或种系来源的状态可以与来自参考群体的比较用结果的数据库进行比较,以鉴定用于该受试者的定制或靶向疗法。通常,参考群体包括与测试的受试者具有相同癌症或疾病类型的患者和/或正在接受或已经接受与测试的受试者相同的疗法的患者。当核酸变体和比较结果满足某些分类标准(例如,基本或近似匹配)时,可以鉴定定制或靶向治疗(或多种治疗)。

[0303] 在某些实施方案中,本文描述的定制疗法通常为胃肠外(例如,静脉内或皮下)施用。包含免疫治疗剂的药物组合物通常被静脉内施用。某些治疗剂是口服施用的。然而,定制疗法(例如,免疫治疗剂等)也可以通过本领域已知的任何方法被施用,包括例如,含服、舌下、直肠、阴道、尿道内、局部(topical)、眼内、鼻内和/或耳内,所述施用可以包括片剂、胶囊、颗粒、水性悬浮液、凝胶、喷雾剂、栓剂、油膏(salve)、软膏(ointment)等。

## 实施例

[0304] 实施例1:评价无细胞DNA样品的分区

[0305] 在此分析了来自患者的无细胞DNA样品。通过将无细胞DNA样品与表观遗传对照核酸分子的集合组合来产生加标样品。在本实施例中,表观遗传对照核酸分子是双链DNA分子,并且表观遗传对照核酸分子的集合是表观遗传对照核酸分子的6个不同子集(子集1至子集6)的汇集(pool)。子集1、子集2、子集3、子集4、子集5和子集6包含在表观遗传修饰区域中具有0个、1个、3个、5个、7个和9个甲基化胞嘧啶(5-甲基胞嘧啶)的表观遗传对照核酸分子。表观遗传对照核酸分子在表观遗传修饰区域的一个末端处具有分子条形码,并且表观遗传状态条形码存在于表观遗传修饰区域的两个末端处。在此使用的分子条形码是独特分子条形码,即,每种表观遗传对照核酸分子具有不同的分子条形码。

[0306] 然后,将该加标样品与甲基结合结构域(MBD)缓冲液和与MBD蛋白缀合的磁珠组合,并过夜孵育。在该孵育期间,甲基化的DNA(如果存在于无细胞DNA样品中)和甲基化的表观遗传对照核酸分子被MBD蛋白结合。用包含递增浓度的盐的缓冲液将未甲基化或甲基化程度较低的DNA从珠洗掉。最后,使用高盐缓冲液从MBD蛋白洗掉高度甲基化的DNA。这些洗涤产生甲基化程度递增的DNA的三个分区(三个分区的集合—低、中等和高)。存在于分区的集合中的分区的DNA包括来自无细胞DNA样品和表观遗传对照核酸分子的DNA。将三个分区的集合中的分区的DNA进行清洗,以去除盐,并浓缩,以准备用于文库制备的酶促步骤。

[0307] 将分区的集合中的DNA浓缩后,将分区的DNA的末端突出物延伸,并且将腺苷残基添加至片段的3’末端。将每个片段的5’末端磷酸化。这些修饰使得分区的DNA可连接。添加DNA连接酶和衔接子,以将每个分区的DNA分子在两个末端处与衔接子连接。这些衔接子包含非独特条形码,并且每个分区的集合与具有非独特条形码的衔接子连接,所述非独特条形码与其他分区的集合中使用的衔接子中的条形码可区分。连接后,将3个分区的集合汇集在一起,并通过PCR扩增。

[0308] PCR后,扩增的DNA在富集前被再次清洗和浓缩。浓缩后,将扩增的DNA与盐缓冲液和靶向特定的感兴趣区域以及表观遗传对照核酸分子的生物素化的RNA探针组合,并将该混合物过夜孵育。生物素化的RNA探针被链霉亲和素磁珠捕获,并通过一系列盐洗涤与未捕获的扩增DNA分离,从而富集样品。富集后,通过PCR扩增将样品索引掺入富集的分子。PCR扩增后,将来自不同样品(一批内)的扩增分子汇集在一起,并使用Illumina NovaSeq测序仪进行测序。

[0309] 然后使用生物信息学工具/算法分析由测序仪生成的序列读段,以生成属于三个分区的集合的每一个中存在的每个子集的表观遗传对照核酸分子的表观遗传分区评分。图9A示出了本实施例中描述的属于高分区的集合中的六个子集(子集1、子集2、子集3、子集4、子集5和子集6)的每一个的表观遗传对照核酸分子的表观遗传分区评分的曲线图。图9B示出了属于中等分区的集合中的六个子集的每一个的表观遗传对照核酸分子的表观遗传分区评分的曲线图。图9C示出了属于低分区的集合中的六个子集的每一个的表观遗传对照核酸分子的表观遗传分区评分的曲线图。如图9所示,在高分区的集合、中等分区的集合和低分区的集合中子集1的表观遗传分区评分分别为约0.1%、0.3%和99.6%。在高分区的集合、中等分区的集合和低分区的集合中,子集1的预定表观遗传分区截止值分别为≤0.3%、≤0.5%和≥97%。在此,表观遗传分区评分用百分比表示。将每个分区的集合中子集1的表观遗传分区评分与子集1的对应表观遗传截止值进行比较—即,将高分区的集合中子集1的表观遗传分区评分(0.1%)与高分区的集合中子集1的表观遗传分区截止值(≤0.3%)进行比较。高分区的集合中子集1的表观遗传分区评分(0.1%)在高分区的集合中子集1的表观遗传分区截止值(≤0.3%)内。同样,将中等分区的集合和低分区的集合中子集1的表观遗传分区评分与中等分区的集合和低分区的集合中子集1的相应表观遗传分区截止值进行比较。类似地,将子集2、子集3、子集4、子集5和子集6的表观遗传分区评分与所有三个分区的集合中相应的表观遗传分区截止值进行比较。因此,我们总共有18个表观遗传分区评分(对于三个分区集合中的六个子集,6×3=18),并且将这些表观遗传分区评分中的每一个与对应的表观遗传分区截止值进行比较。发现所有18个表观遗传分区评分都在相应的表观遗传分区截止值内。因此,对在此分析的无细胞DNA样品进行的分区方法被分类为成功。

[0310] 实施例2:评价无细胞DNA样品的分区

[0311] 在此分析了来自一组患者的一组无细胞DNA样品。在本实施例中,未使用表观遗传对照核酸分子。替代地,使用无细胞DNA样品中的内源对照分子来评价无细胞DNA样品的分区。将每个患者的无细胞DNA样品与甲基结合结构域(MBD)缓冲液和与MBD蛋白缀合的磁珠组合,并过夜孵育。在该孵育期间,甲基化的DNA(如果存在于无细胞DNA样品中)和甲基化的内源对照分子被MBD蛋白结合。用包含递增浓度的盐的缓冲液将未甲基化或甲基化程度较低的DNA从珠洗掉。最后,使用高盐缓冲液从MBD蛋白洗掉高度甲基化的DNA。这些洗涤产生

甲基化程度递增的DNA的三个分区(三个分区的集合—低、中等和高)。存在于分区的集合中的分区的DNA包括来自无细胞DNA样品和内源对照分子的DNA。将三个分区的集合中的分区的DNA进行清洗,以去除盐,并浓缩,以准备用于文库制备的酶促步骤。

[0312] 将分区的集合中的DNA浓缩后,将分区的DNA的末端突出物延伸,并且将腺苷残基添加至片段的3’末端。将每个片段的5’末端磷酸化。这些修饰使得分区的DNA可连接。添加DNA连接酶和衔接子,以将每个分区的DNA分子在两个末端处与衔接子连接。这些衔接子包含非独特条形码,并且每个分区的集合与具有非独特条形码的衔接子连接,所述非独特条形码与其他分区的集合中使用的衔接子中的条形码可区分。连接后,将3个分区的集合汇集在一起,并通过PCR扩增。

[0313] PCR后,扩增的DNA在富集前被再次清洗和浓缩。浓缩后,将扩增的DNA与盐缓冲液和靶向特定的感兴趣区域以及内源对照分子的生物素化的RNA探针组合,并将该混合物过夜孵育。生物素化的RNA探针被链霉亲和素磁珠捕获,并通过一系列盐洗涤与未捕获的扩增DNA分离,从而富集样品。富集后,通过PCR扩增将样品索引掺入富集的分子。PCR扩增后,将来自不同样品(一批内)的扩增分子汇集在一起,并使用Illumina NovaSeq测序仪进行测序。

[0314] 然后使用生物信息学工具/算法分析由测序仪生成的序列读段,以生成内源对照分子的一个或更多个表观遗传分区评分。在本实施例中,甲基-半和甲基-5被用作表观遗传分区评分。图10A示出了高分区的集合中样品1的高甲基化对照分子的分数的曲线图,并且样品1的甲基-半评分为11。图10B示出了低分区的集合中样品1的高甲基化分子的分数的曲线图,并且样品1的甲基-5评分为13。图11A示出了高分区的集合中样品2的高甲基化对照分子的分数的曲线图,并且样品2的甲基-半评分为13。图11B示出了低分区的集合中样品2的高甲基化分子的分数的曲线图,并且样品2的甲基-5评分不能被确定(如图11B所示)。在本实施例中,甲基-半和甲基-5的表观遗传分区截止值分别为15个和20个甲基化CG。样品1的甲基-半和甲基-5评分在对应的表观遗传分区截止值内。但是,对于样品2,甲基-半评分在其对应的表观遗传分区截止值内,但是甲基-5评分不在其对应的表观遗传分区截止值内。因此,样品1的分区方法被分类为成功,而样品2的分区方法被分类为不成功。

[0315] 虽然本文已经示出和描述了本发明的优选实施方案,但对于本领域技术人员将明显的是,此类实施方案仅通过示例的方式提供。并不意图本发明限于本说明书中提供的特定实例。虽然已参考以上提及的说明书描述了本发明,但本文实施方案的描述和说明并不意图以限制性的意义来解释。本领域技术人员现将想到不偏离本发明的许多变化、改变和替换。此外,应当理解,本发明的所有方面并不限于本文根据各种条件和变量阐述的具体描述、配置或相对比例。应当理解,在实践本发明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选择。因此设想本公开内容还应涵盖任何此类替代选择、修改、变化或等同物。所附权利要求意图限定本发明的范围,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。

[0316] 虽然为了清楚与理解的目的,已经通过图示和实例的方式对前述公开内容进行了一些详细描述,但是本领域普通技术人员通过阅读本公开内容将会清楚,在不偏离本公开内容的真实范围的情况下,可以进行形式和细节上的多种改变,并且可以在所附权利要求书的范围内实施。例如,所有方法、系统、计算机可读介质和/或组件特征、步骤、元件或其他

方面都可以以多种组合来使用。

[0317] 本文引用的所有专利、专利申请、网站、其他出版物或文件、登录号等都为了所有目的被通过引用以其整体并入，其程度如同每个单独的项目都被具体且单独地指示通过引用如此并入一样。如果一个序列的不同版本在不同时间与一个登记号相关联，则意指在本申请的实际提交日期与该登记号相关联的版本。如果适用的话，有效提交日期意指真实提交日期或提及该登记号的优先权申请的提交日期中较早的一个。同样，如果出版物、网站等的不同版本在不同时间发布，则意指在本申请的实际提交日期最近发布的版本，除非另有指示。

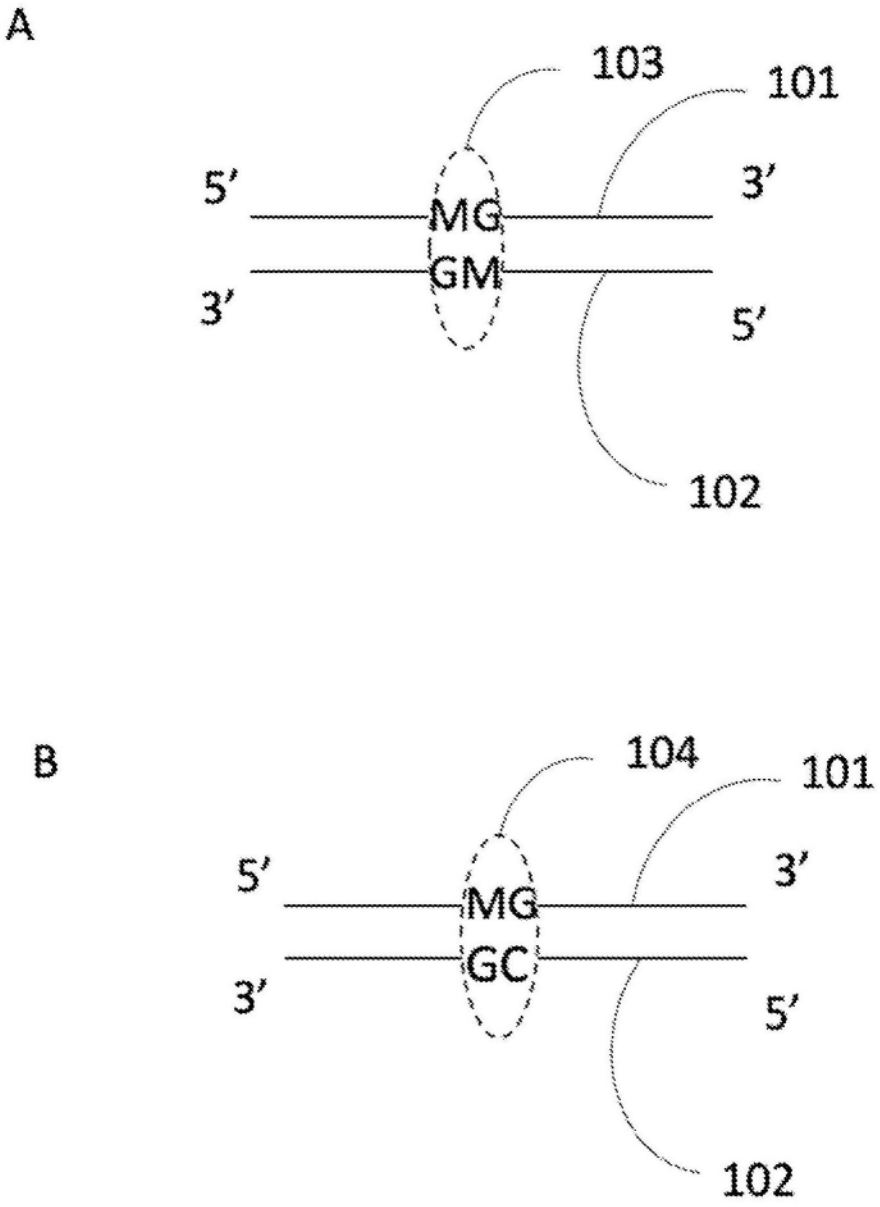



图1

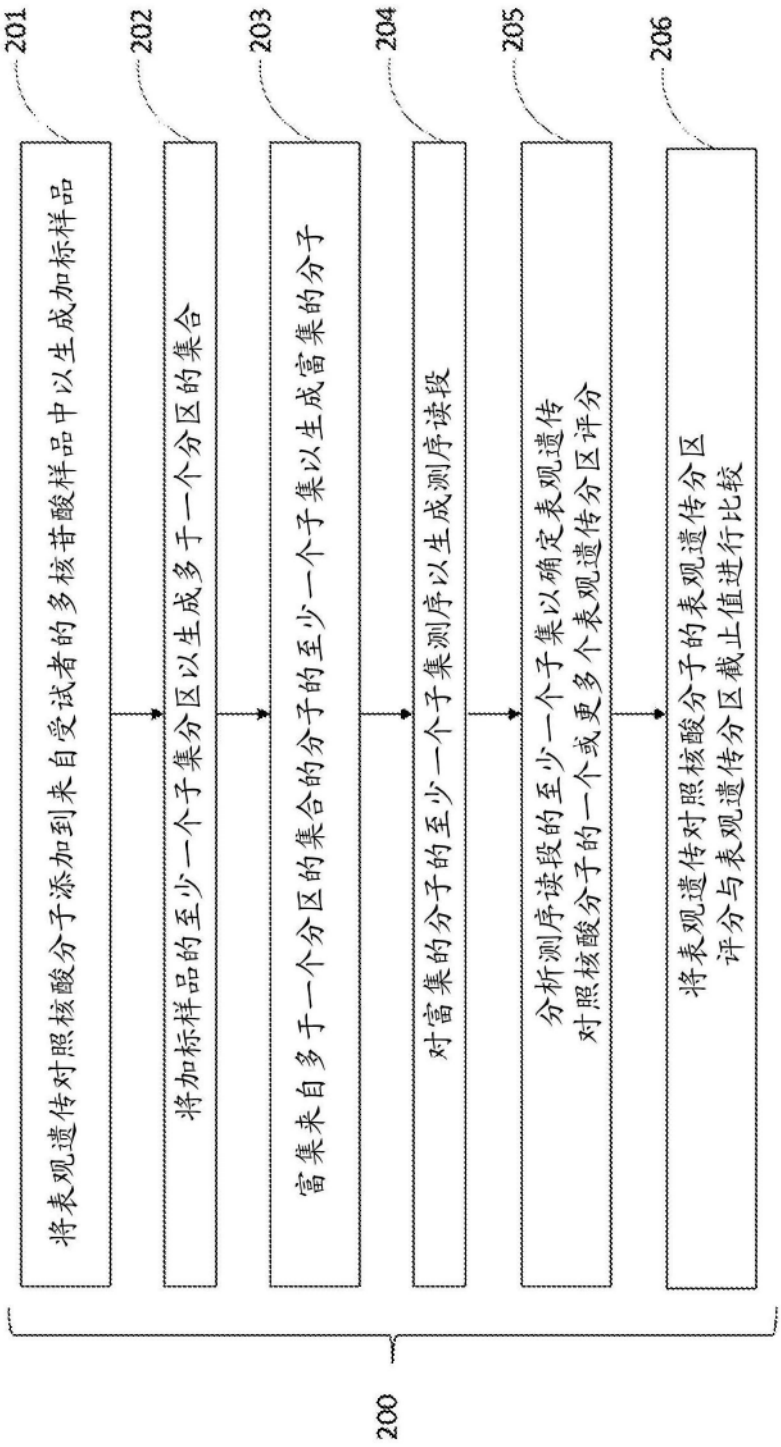

201
将表观遗传对照核酸分子添加到来自受试者的多核苷酸样品中以生成加标样品
202
将加标样品的至少一个子集分区以生成多于一个分区的集合
203
富集来自多于一个分区的集合的分子的至少一个子集以生成富集的分子
204
对富集的分子的至少一个子集测序以生成测序读段
205
分析测序读段的至少一个子集以确定表观遗传
对照核酸分子的一个或更多个表观遗传分区评分
206
将表观遗传对照核酸分子的表观遗传分区
评分与表观遗传分区截止值进行比较
200

图2

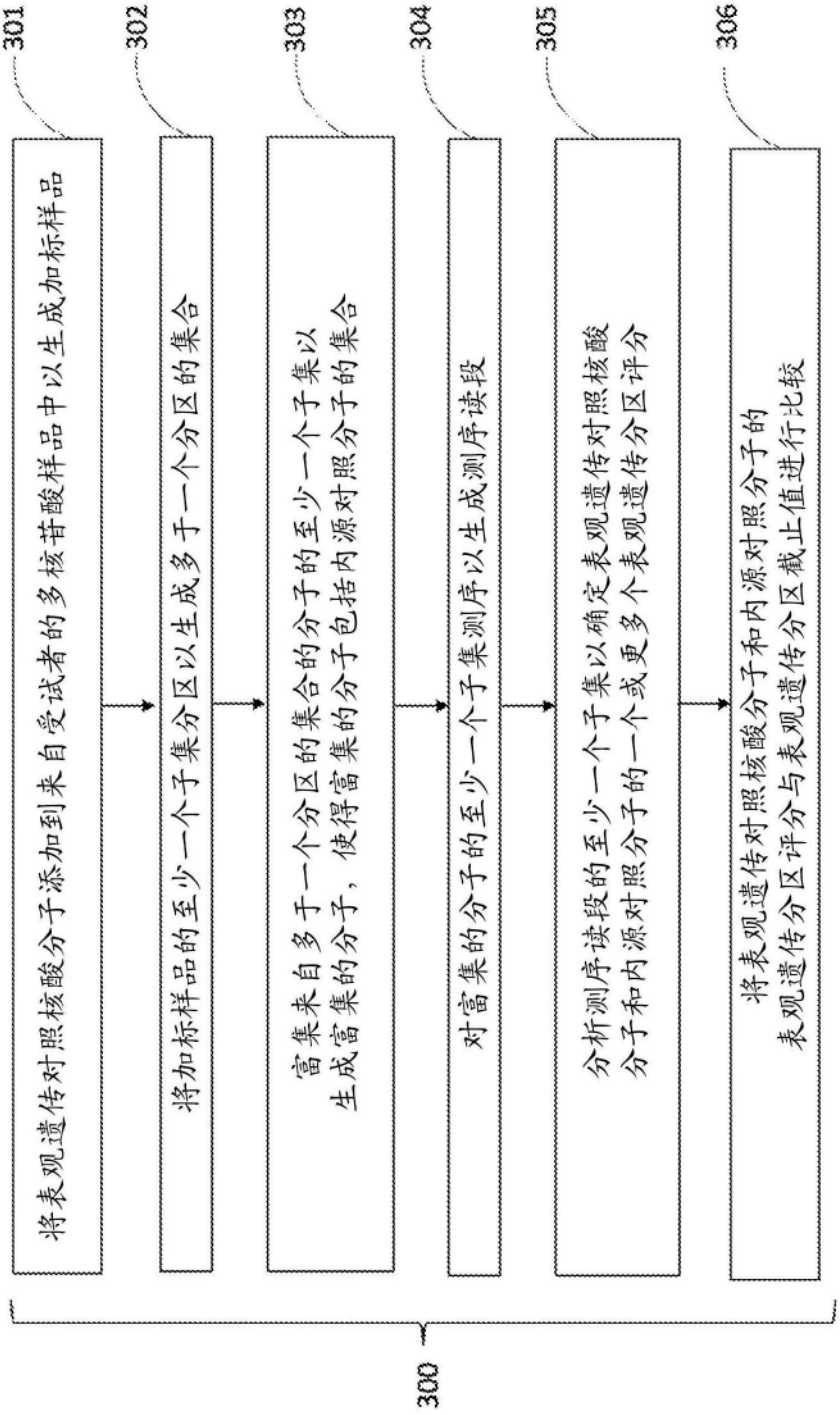

300
301
将表观遗传对照核酸分子添加到来自受试者的多核苷酸样品中以生成加标样品
302
将加标样品的至少一个子集分区以生成多于一个分区的集合
303
富集来自多于一个分区的集合的分子的至少一个子集以
生成富集的分子，使得富集的分子包括内源对照分子的集合
304
对富集的分子的至少一个子集测序以生成测序读段
305
分析测序读段的至少一个子集以确定表观遗传对照核酸
分子和内源对照分子的一个或更多个表观遗传分区评分
306
将表观遗传对照核酸分子和内源对照分子的
表观遗传分区评分与表观遗传分区截止值进行比较

图3

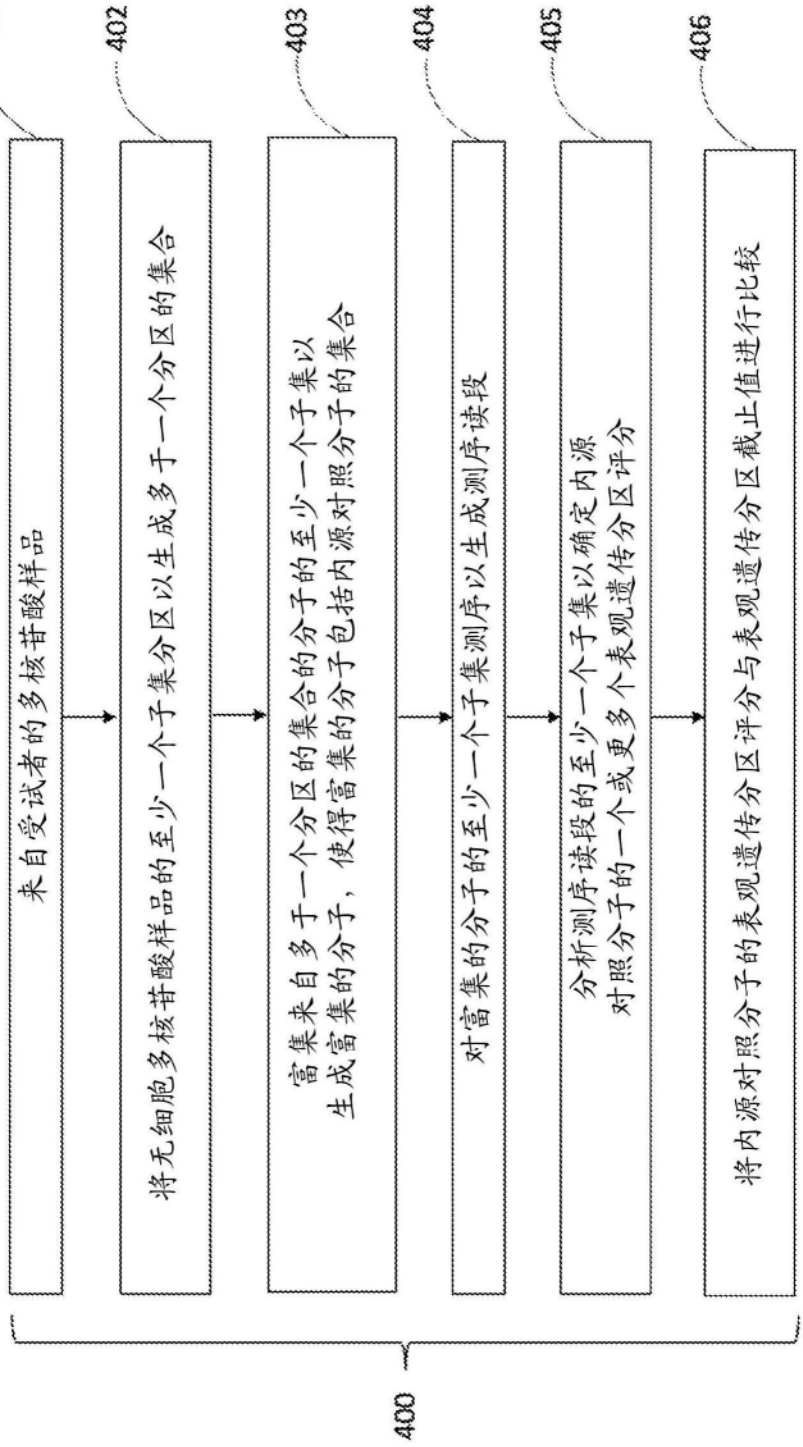

401
来自受试者的多核苷酸样品
402
将无细胞多核苷酸样品的至少一个子集分区以生成多于一个分区的集合
403
富集来自多于一个分区的集合的分子的至少一个子集以
生成富集的分子，使得富集的分子包括内源对照分子的集合
404
对富集的分子的至少一个子集测序以生成测序读段
405
分析测序读段的至少一个子集以确定内源
对照分子的一个或更多个表观遗传分区评分
406
将内源对照分子的表观遗传分区评分与表观遗传分区截止值进行比较
400

图4

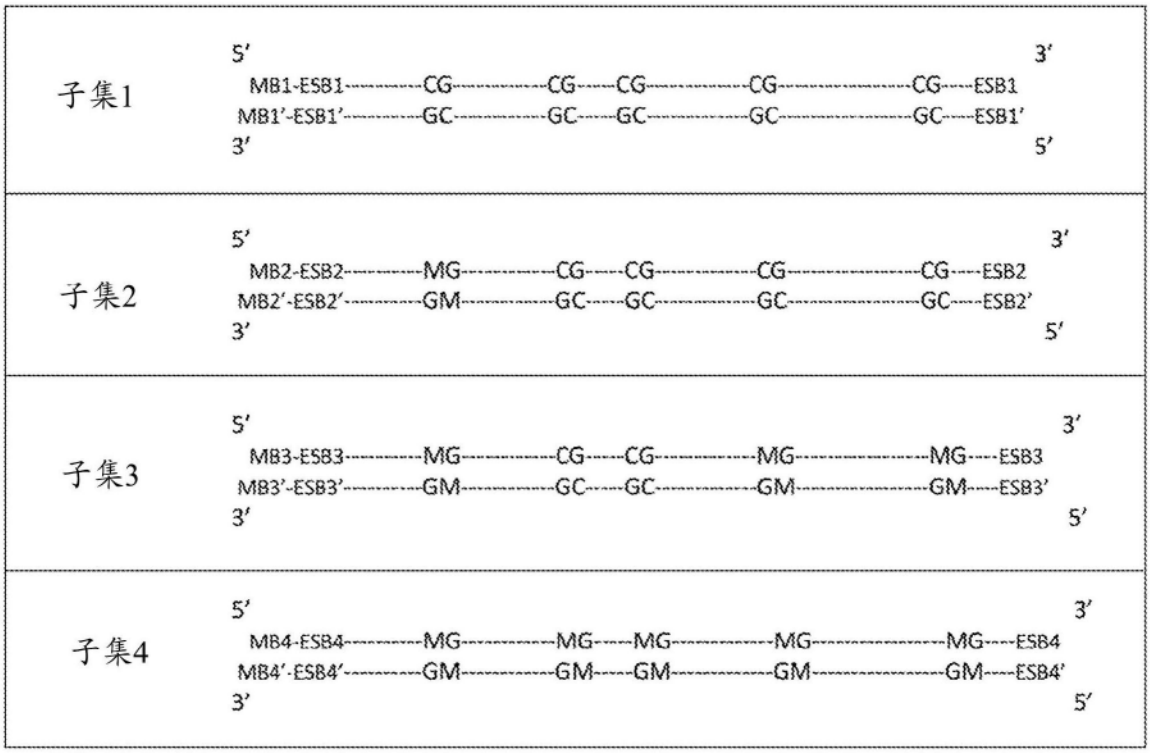

子集1
5′ 3′
MB1-ESB1----------CG----------CG-----CG----------CG----------CG----ESB1
MB1′-ESB1′----------GC----------GC-----GC----------GC----------GC----ESB1′
3′ 5′
子集2
5′ 3′
MB2-ESB2----------MG----------CG-----CG----------CG----------CG----ESB2
MB2′-ESB2′----------GM----------GC-----GC----------GC----------GC----ESB2′
3′ 5′
子集3
5′ 3′
MB3-ESB3----------MG----------CG-----CG----------MG----------MG----ESB3
MB3′-ESB3′----------GM----------GC-----GC----------GM----------GM----ESB3′
3′ 5′
子集4
5′ 3′
MB4-ESB4----------MG----------MG-----MG----------MG----------MG----ESB4
MB4′-ESB4′----------GM----------GM-----GM----------GM----------GM----ESB4′
3′ 5′

图5

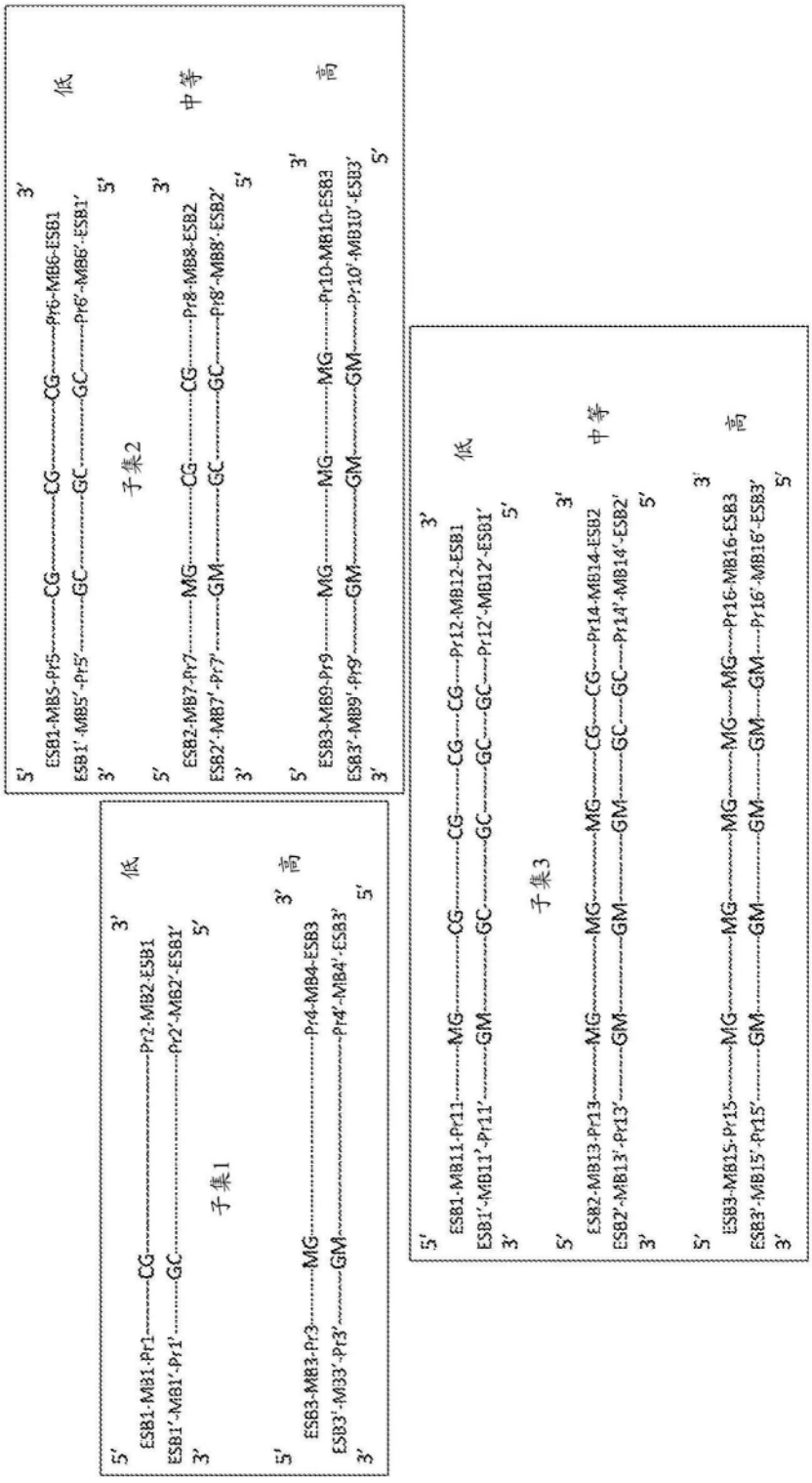



图6

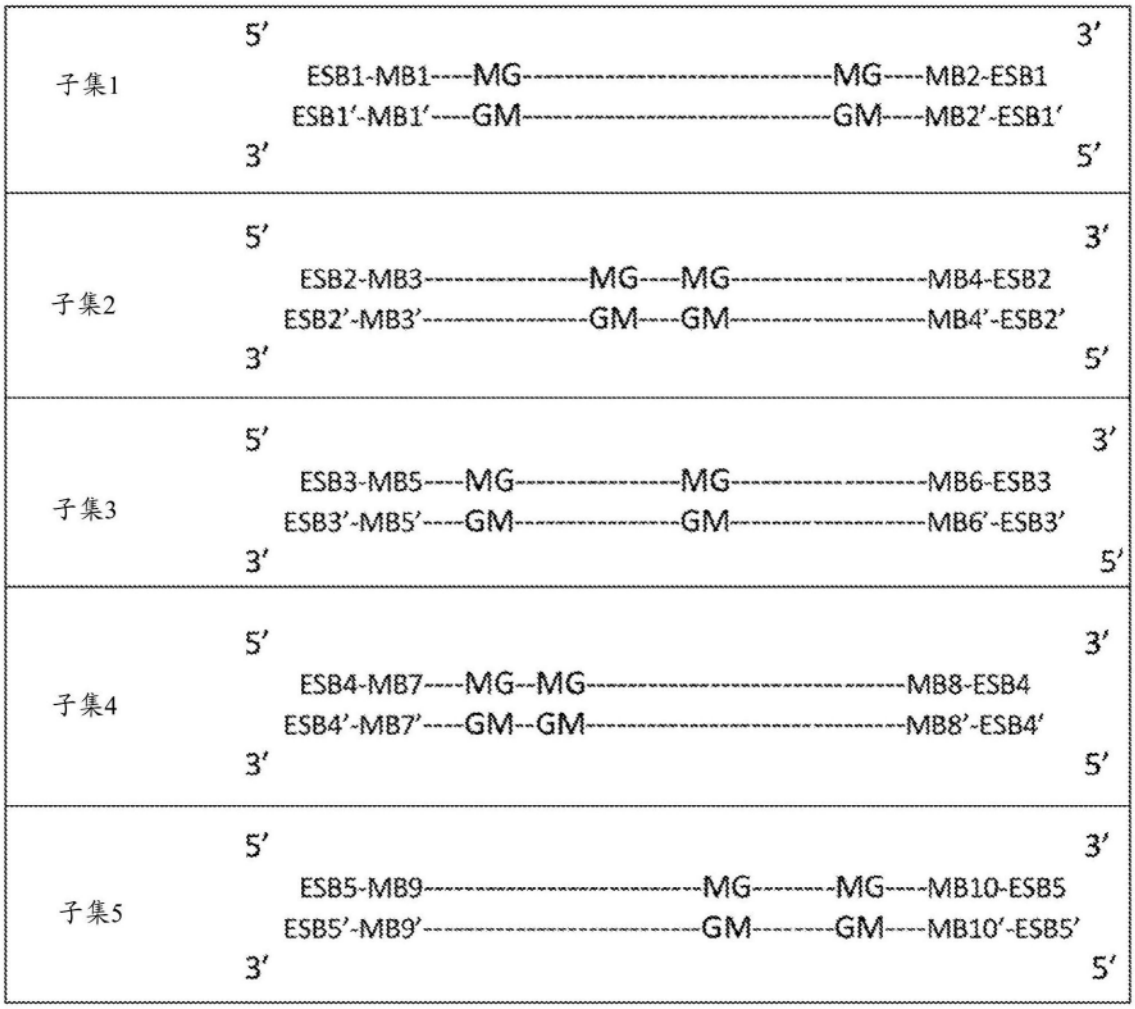

子集1
5′
3′
ESB1-MB1----MG----------MG----MB2-ESB1
ESB1′-MB1′----GM----------GM----MB2′-ESB1′
3′
5′
子集2
5′
3′
ESB2-MB3----------MG----MG----------MB4-ESB2
ESB2′-MB3′----------GM----GM----------MB4′-ESB2′
3′
5′
子集3
5′
3′
ESB3-MB5----MG----------MG----------MB6-ESB3
ESB3′-MB5′----GM----------GM----------MB6′-ESB3′
3′
5′
子集4
5′
3′
ESB4-MB7----MG--MG----------MB8-ESB4
ESB4′-MB7′----GM--GM----------MB8′-ESB4′
3′
5′
子集5
5′
3′
ESB5-MB9----------MG--------MG----MB10-ESB5
ESB5′-MB9′----------GM--------GM----MB10′-ESB5′
3′
5′

图7

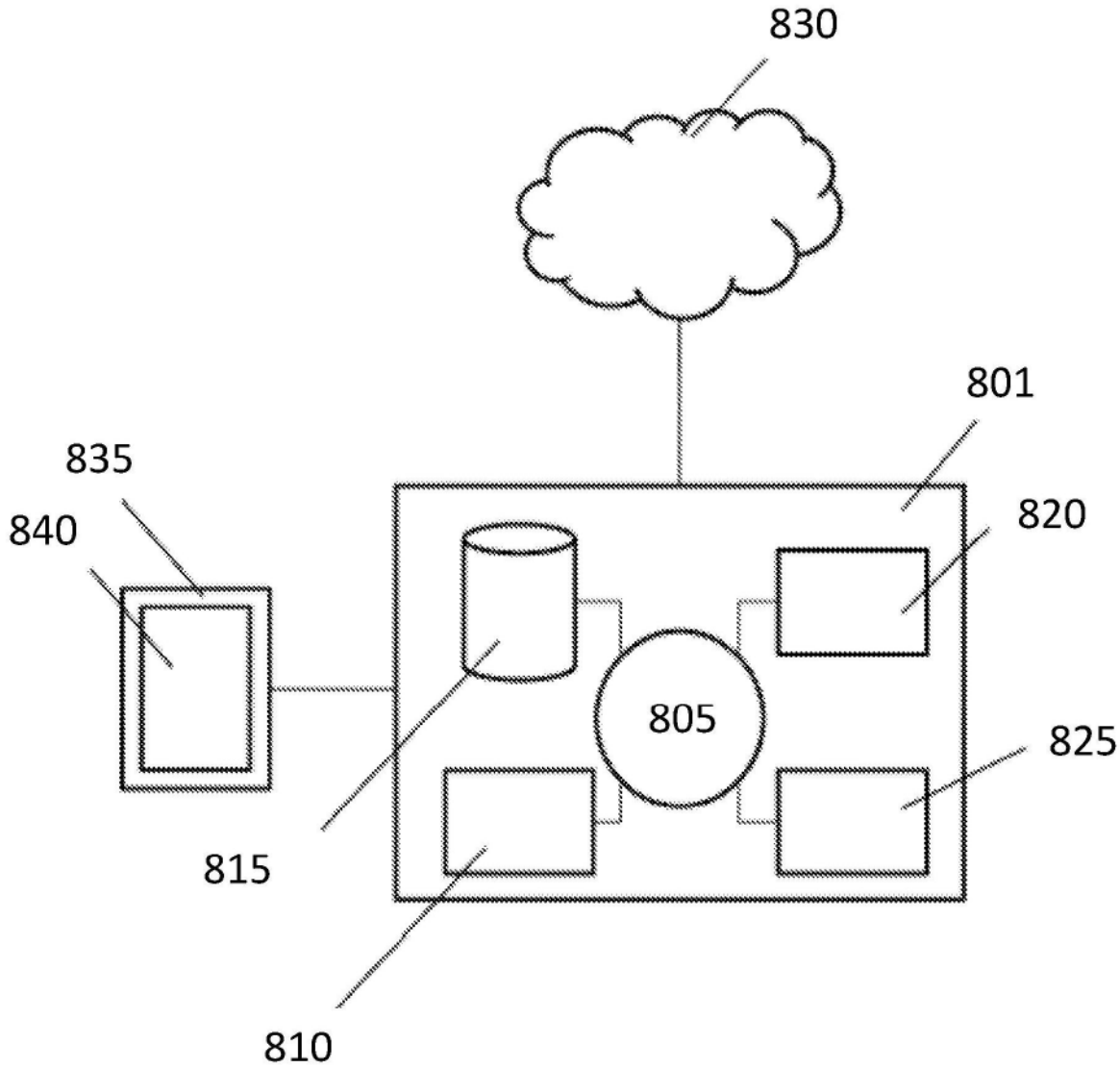



图8

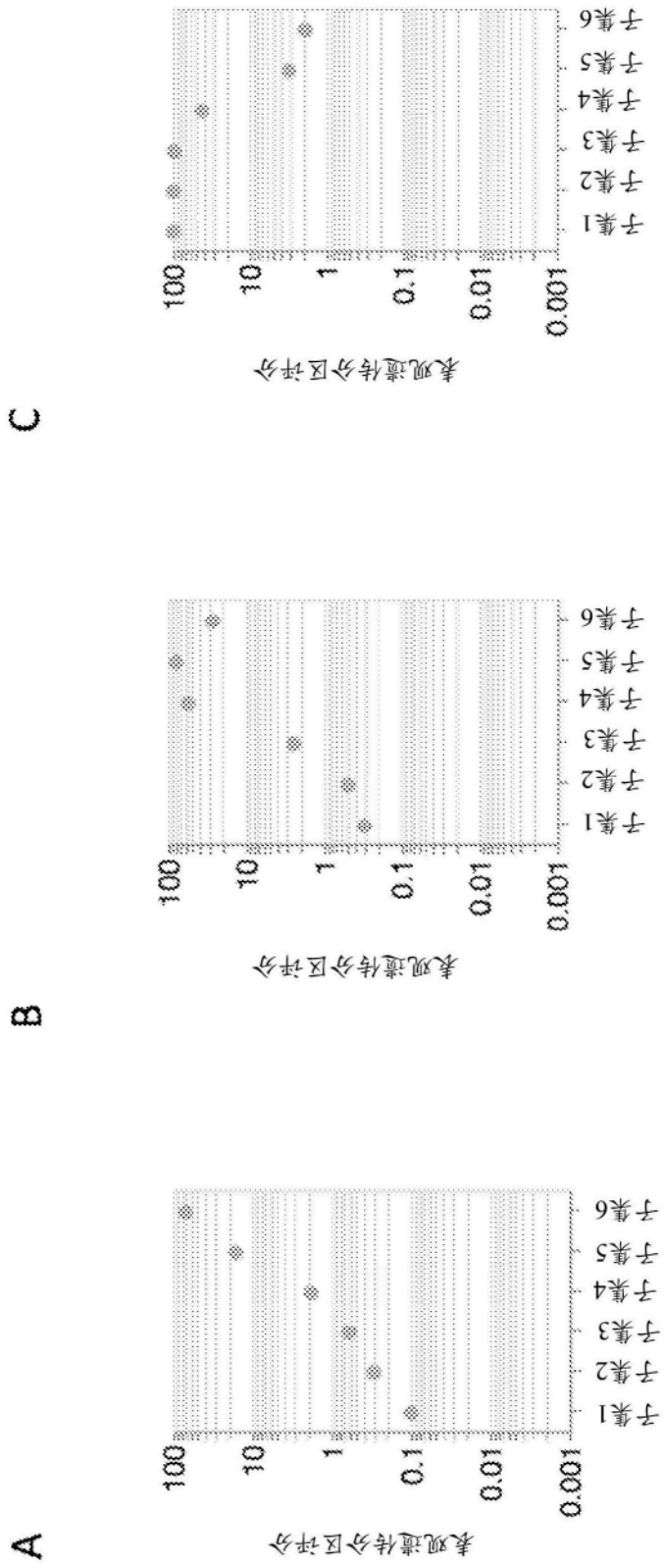
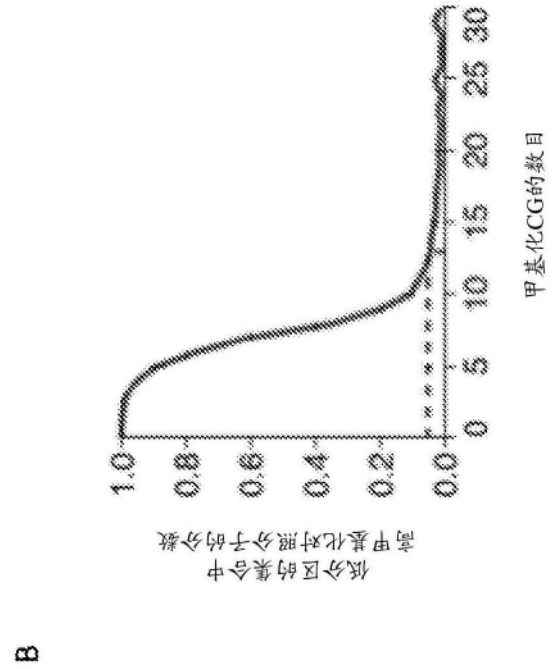



图9


B
低分区的集合中
高甲基化对照分子的分数
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
甲基化CG的数目

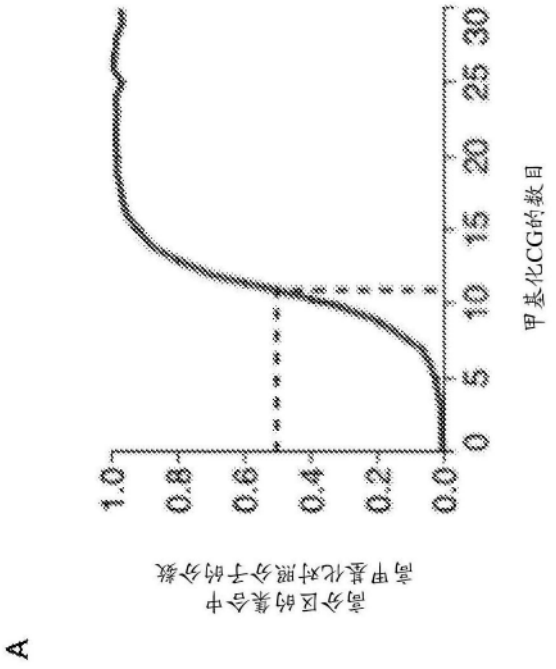
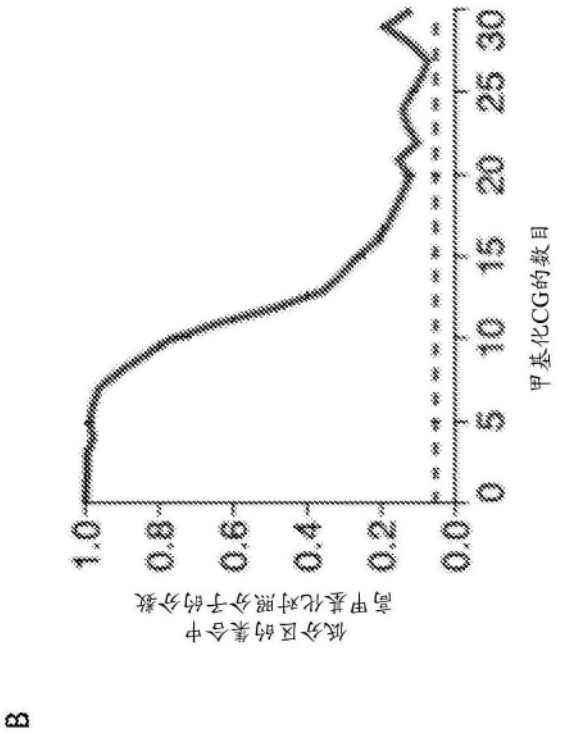

A
高分区的集合中
高甲基化对照分子的分数
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
甲基化CG的数目

图10


B
低分区的集合中
高甲基化对照分子的分数
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
甲基化CG的数目

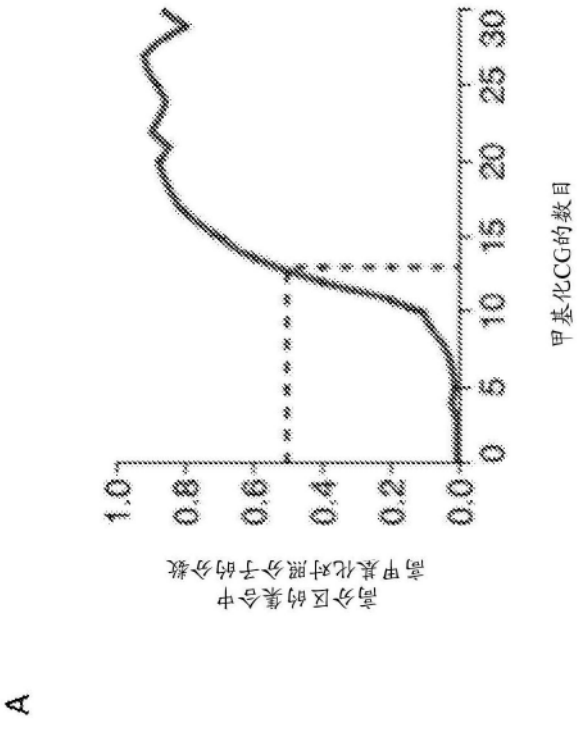

A
高分区的集合中
高甲基化对照分子的分数
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
甲基化CG的数目

图11