(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115698339 A
(43)申请公布日 2023.02.03

(21)申请号 202180038991.0
(22)申请日 2021.04.06
(30)优先权数据
63/006,556 2020.04.07 US
(85)PCT国际申请进入国家阶段日
2022.11.29
(86)PCT国际申请的申请数据
PCT/US2021/026043 2021.04.06
(87)PCT国际申请的公布数据
WO2021/207267 EN 2021.10.14
(71)申请人 私人基因诊断公司
地址 美国马里兰
(72)发明人 J·F·汤普森
(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 陈晶晶
(51)Int.Cl.
C12Q 1/6886(2006.01)
C12Q 1/6869(2006.01)
C12Q 1/6827(2006.01)
C12Q 1/6874(2006.01)
C40B 20/04(2006.01)

权利要求书6页 说明书25页
序列表3页 附图11页

(54)发明名称

不固定条形码

(57)摘要

本文提供了用于标记和分析包括具有预先确定数量的索引位置的索引条形码的核酸分子的寡核苷酸系统和组。本文还提供了用于标记和分析核酸分子的方法,以及使用本文所述的样品和分子条形码识别错误序列读段的方法。

CN 115698339 A

1.一种用于标记样品中的核酸分子的系统,其包含:

包括多个条形码的寡核苷酸组,每个条形码包含一段连续碱基,所述一段连续碱基包含:

(i)包括预先确定数量的样品索引位置的样品条形码,所述样品索引位置包含一个或多个特定核苷酸,其中所述样品索引位置的定位在样品之间变化;和

(ii)包括分子索引位置的分子条形码,所述分子索引位置包含与在所述样品索引位置处的核苷酸不同的核苷酸,其中所述样品索引位置散布在分子索引位置中。

2.如权利要求1所述的系统,其中所述样品条形码位置的预先确定数量在不同样品条形码之间变化。

3.如权利要求1所述的系统,其中所述条形码包含约10至约35个核苷酸。

4.如权利要求1所述的系统,其中所述条形码包含约12至约25个核苷酸。

5.如权利要求1所述的系统,其中所述样品条形码包括2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个样品索引位置或其组合。

6.如权利要求1所述的系统,其中所述样品条形码包括约4至约12个样品索引位置。

7.如权利要求1所述的系统,其中所述分子条形码包括约5至约25个分子索引位置。

8.如权利要求1所述的系统,其中所述分子条形码包括约5至约15个分子索引位置。

9.如权利要求1所述的系统,其中所述样品索引位置核苷酸和所述分子索引位置核苷酸选自:

(A)样品索引位置核苷酸为A,并且分子索引位置核苷酸为C、G、T或其组合;

(B)样品索引位置核苷酸为T,并且分子索引位置核苷酸为C、G、A或其组合;

(C)样品索引位置核苷酸为C,并且分子索引位置核苷酸为G、A、T或其组合;

(D)样品索引位置核苷酸为G,并且分子索引位置核苷酸为C、A、T或其组合;

(E)样品索引位置核苷酸为A、T或其组合,并且分子索引位置核苷酸为C、G或其组合;

(F)样品索引位置核苷酸为A、C或其组合,并且分子索引位置核苷酸为T、G或其组合;

(G)样品索引位置核苷酸为A、G或其组合,并且分子索引位置核苷酸为T、C或其组合;

(H)样品索引位置核苷酸为T、C或其组合,并且分子索引位置核苷酸为A、G或其组合;

(I)样品索引位置核苷酸为T、G或其组合,并且分子索引位置核苷酸为A、C或其组合;或者

(J)样品索引位置核苷酸为G、C或其组合,并且分子索引位置核苷酸为A、T或其组合。

10.如权利要求1所述的系统,其中每个条形码包括含有索引位置的一个或多个附加索引条形码。

11.如权利要求10所述的系统,其中所述一个或多个附加索引条形码是细胞条形码,提供未修复端的DNA长度的量度的条形码,或者细胞条形码和提供未修复端的DNA长度的量度的条形码二者。

12.如权利要求1所述的系统,其中所述寡核苷酸组中的每个寡核苷酸还包括非条形码位置,所述非条形码位置包括用于杂交的位点、用于序列引物结合的位点、用于扩增的位点、或其任何组合。

13.用于标记样品中的核酸分子的寡核苷酸组,其包括多个条形码,每个条形码包括:

(i)包括预先确定数量的样品索引位置的样品条形码,所述样品索引位置包含一个或

多个特定核苷酸,其中所述样品索引位置的定位在样品之间变化;和

(ii)包括分子索引位置的分子条形码,所述分子索引位置包含与在所述样品索引位置处的核苷酸不同的核苷酸,

其中样品索引位置和分子索引位置散布于一段连续碱基中。

14.如权利要求13所述的寡核苷酸组,其中所述样品条形码位置的预先确定数量在不同样品条形码之间变化。

15.如权利要求13所述的寡核苷酸组,其中所述条形码包含约10至约35个核苷酸。

16.如权利要求13所述的寡核苷酸组,其中所述条形码包含约12至约25个核苷酸。

17.如权利要求13所述的寡核苷酸组,其中所述样品条形码包括2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个样品索引位置或其组合。

18.如权利要求13所述的寡核苷酸组,其中所述样品条形码包括约4至约12个样品索引位置。

19.如权利要求13所述的寡核苷酸组,其中所述分子条形码包括约5至约25个分子索引位置。

20.如权利要求13所述的寡核苷酸组,其中所述分子条形码包括约5至约15个分子索引位置。

21.如权利要求13所述的寡核苷酸组,其中所述样品索引位置核苷酸和所述分子索引位置核苷酸选自:

(A)样品索引位置核苷酸为A,并且分子索引位置核苷酸为C、G、T或其组合;

(B)样品索引位置核苷酸为T,并且分子索引位置核苷酸为C、G、A或其组合;

(C)样品索引位置核苷酸为C,并且分子索引位置核苷酸为G、A、T或其组合;

(D)样品索引位置核苷酸为G,并且分子索引位置核苷酸为C、A、T或其组合;

(E)样品索引位置核苷酸为A、T或其组合,并且分子索引位置核苷酸为C、G或其组合;

(F)样品索引位置核苷酸为A、C或其组合,并且分子索引位置核苷酸为T、G或其组合;

(G)样品索引位置核苷酸为A、G或其组合,并且分子索引位置核苷酸为T、C或其组合;

(H)样品索引位置核苷酸为T、C或其组合,并且分子索引位置核苷酸为A、G或其组合;

(I)样品索引位置核苷酸为T、G或其组合,并且分子索引位置核苷酸为A、C或其组合;或者

(J)样品索引位置核苷酸为G、C或其组合,并且分子索引位置核苷酸为A、T或其组合。

22.如权利要求13所述的寡核苷酸组,其中每个条形码包括含有索引位置的一个或多个附加索引条形码。

23.如权利要求22所述的寡核苷酸组,其中所述一个或多个附加索引条形码是细胞条形码,提供未修复端的DNA长度的量度的条形码,或者细胞条形码和提供未修复端的DNA长度的量度的条形码二者。

24.如权利要求13所述的寡核苷酸组,其中所述寡核苷酸组中的每个寡核苷酸还包括非条形码位置,所述非条形码位置包括用于杂交的位点、用于序列引物结合的位点、用于扩增的位点、或其任何组合。

25.用于分析样品中的核酸分子的序列的方法,其包括:

(a)将多个寡核苷酸附接到所述核酸分子,其中每个寡核苷酸包括条形码,所述条形码

包括：

(i)包括预先确定数量的样品索引位置的样品条形码，所述样品索引位置包含一个或多个特定核苷酸，其中所述样品索引位置的定位在样品之间变化；和

(ii)包括分子索引位置的分子条形码，所述分子索引位置包含与在样品索引位置处的核苷酸不同的核苷酸，

其中所述样品索引位置和所述分子索引位置散布于一段连续碱基中；以及

(b)对所述核酸分子进行测序，其中序列读段包含条形码序列。

26.如权利要求25所述的方法，其还包括将包含相同样品条形码的寡核苷酸附接到样品中的核酸分子的每一端。

27.如权利要求25所述的方法，其中所述样品条形码位置的预先确定数量在不同样品条形码之间变化。

28.如权利要求25所述的方法，其中所述条形码包含约10至约35个核苷酸。

29.如权利要求25所述的方法，其中所述条形码包含约12至约25个核苷酸。

30.如权利要求25所述的方法，其中所述样品条形码包括2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个样品索引位置或其组合。

31.如权利要求25所述的方法，其中所述样品条形码包括约4至约12个样品索引位置。

32.如权利要求25所述的方法，其中所述分子条形码包括约5至约25个分子索引位置。

33.如权利要求25所述的方法，其中所述分子条形码包括约5至约15个分子索引位置。

34.如权利要求25所述的方法，其中所述样品索引位置核苷酸和所述分子索引位置核苷酸选自：

(A)样品索引位置核苷酸为A，并且分子索引位置核苷酸为C、G、T或其组合；

(B)样品索引位置核苷酸为T，并且分子索引位置核苷酸为C、G、A或其组合；

(C)样品索引位置核苷酸为C，并且分子索引位置核苷酸为G、A、T或其组合；

(D)样品索引位置核苷酸为G，并且分子索引位置核苷酸为C、A、T或其组合；

(E)样品索引位置核苷酸为A、T或其组合，并且分子索引位置核苷酸为C、G或其组合；

(F)样品索引位置核苷酸为A、C或其组合，并且分子索引位置核苷酸为T、G或其组合；

(G)样品索引位置核苷酸为A、G或其组合，并且分子索引位置核苷酸为T、C或其组合；

(H)样品索引位置核苷酸为T、C或其组合，并且分子索引位置核苷酸为A、G或其组合；

(I)样品索引位置核苷酸为T、G或其组合，并且分子索引位置核苷酸为A、C或其组合；或者

(J)样品索引位置核苷酸为G、C或其组合，并且分子索引位置核苷酸为A、T或其组合。

35.如权利要求25所述的方法，其中每个条形码包括含有索引位置的一个或多个附加索引条形码。

36.如权利要求35所述的方法，其中所述一个或多个附加索引条形码是细胞条形码，提供未修复端的DNA长度的量度的条形码，或者细胞条形码和提供未修复端的DNA长度的量度的条形码二者。

37.如权利要求25所述的方法，其还包括基于所述样品索引位置的定位将所述序列读段分配到样品家族。

38.如权利要求25所述的方法，其还包括基于分子索引位置的定位和在每个分子索引

位置处的核苷酸将所述序列读段分配到分子家族。

39.如权利要求25所述的方法,其还包括通过将序列读段中样品索引位置的数量和定位与样品索引位置的预先确定数量和定位比较来校正测序错误。

40.如权利要求25所述的方法,其还包括通过比较在序列读段两端处的样品条形码来校正测序错误。

41.如权利要求40所述的方法,其包括应用规则以将在所述序列读段的每一端处的不同样品条形码与允许的样品条形码比较。

42.如权利要求25所述的方法,其还包括应用一个或多个规则来(1)校正条形码内的错误,(2)校正在核酸分子每一端处的条形码之间的错误,(3)将序列读段多路分解到样品家族,(4)将序列读段分配到分子家族,或其任何组合。

43.如权利要求25所述的方法,其中每个寡核苷酸还包括非条形码位置,所述非条形码位置包括用于杂交的位点、用于序列引物结合的位点、用于扩增的位点、或其任何组合。

44.如权利要求25所述的方法,其还包括使用不同的基因组并测试每个寡核苷酸以灵敏地检测序列读段错误分配。

45.用于标记样品中的核酸分子的方法,其包括:

将包括条形码的多个寡核苷酸附接到核酸分子,每个条形码包括:

(i)包括预先确定数量的样品索引位置的样品条形码,所述样品索引位置包含一个或多个特定核苷酸,其中所述样品索引位置的定位在样品之间变化;和

(ii)包括分子索引位置的分子条形码,所述分子索引位置包含与在所述样品索引位置处的核苷酸不同的核苷酸,

其中所述样品索引位置和所述分子索引位置散布于一段连续碱基中。

46.如权利要求45所述的方法,其还包括将包含相同样品条形码的寡核苷酸附接到核酸分子的每一端。

47.如权利要求45所述的方法,其中所述样品条形码位置的预先确定数量在不同样品条形码之间变化。

48.如权利要求45所述的方法,其中所述条形码包含约10至约35个核苷酸。

49.如权利要求45所述的方法,其中所述条形码包含约12至约25个核苷酸。

50.如权利要求45所述的方法,其中所述样品条形码包括2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个样品索引位置。

51.如权利要求45所述的方法,其中所述样品条形码包括约4至约12个样品索引位置。

52.如权利要求45所述的方法,其中所述分子条形码包括约5至约25个分子索引位置。

53.如权利要求45所述的方法,其中所述分子条形码包括约5至约15个分子索引位置。

54.如权利要求45所述的方法,其中所述样品索引位置核苷酸和所述分子索引位置核苷酸选自:

(A)样品索引位置核苷酸为A,并且分子索引位置核苷酸为C、G、T或其组合;

(B)样品索引位置核苷酸为T,并且分子索引位置核苷酸为C、G、A或其组合;

(C)样品索引位置核苷酸为C,并且分子索引位置核苷酸为G、A、T或其组合;

(D)样品索引位置核苷酸为G,并且分子索引位置核苷酸为C、A、T或其组合;

(E)样品索引位置核苷酸为A、T或其组合,并且分子索引位置核苷酸为C、G或其组合;

(F)样品索引位置核苷酸为A、C或其组合，并且分子索引位置核苷酸为T、G或其组合；

(G)样品索引位置核苷酸为A、G或其组合，并且分子索引位置核苷酸为T、C或其组合；

(H)样品索引位置核苷酸为T、C或其组合，并且分子索引位置核苷酸为A、G或其组合；

(I)样品索引位置核苷酸为T、G或其组合，并且分子索引位置核苷酸为A、C或其组合；或者

(J)样品索引位置核苷酸为G、C或其组合，并且分子索引位置核苷酸为A、T或其组合。

55.如权利要求45所述的方法，其中每个条形码包括含有索引位置的一个或多个附加索引条形码。

56.如权利要求55所述的方法，其中所述一个或多个附加条形码是细胞条形码，提供未修复端的DNA长度的量度的条形码，或者细胞条形码和提供未修复端的DNA长度的量度的条形码二者。

57.如权利要求45所述的方法，其中每个寡核苷酸还包括非条形码位置，所述非条形码位置包括用于杂交的位点、用于序列引物结合的位点、用于扩增的位点、或其任何组合。

58.如权利要求25至44中任一项所述的方法，其还包括在没有多路分解的情况下存储核酸序列数据。

59.如权利要求58所述的方法，其中在没有多路分解的情况下存储核酸序列数据防止在没有多路分解密钥的情况下使用序列数据，并且防止未经授权使用数据。

60.用于识别错误序列读段的方法，其包括：

(a)将多个寡核苷酸附接到样品中的核酸分子，其中每个寡核苷酸包括条形码，所述条形码包括：

(i)包括预先确定数量的样品索引位置的样品条形码，所述样品索引位置包含一个或多个特定核苷酸，其中所述样品索引位置的定位在样品之间变化，并且其中将相同的样品条形码附接到样品中的核酸分子的每一端；和

(ii)包括分子索引位置的分子条形码，所述分子索引位置包含与在所述样品索引位置处的核苷酸不同的核苷酸，

其中所述样品索引位置和所述分子索引位置散布于一段连续碱基中；以及

(b)对核酸分子进行测序，其中序列读段包括条形码序列，

从而识别错误序列读段。

61.如权利要求60所述的方法，其中所述识别错误序列读段包括识别带有不相符的样品条形码的核酸分子。

62.如权利要求60所述的方法，其还包括通过比较在序列读段两端处的样品条形码来校正测序错误。

63.如权利要求61所述的方法，其还包括从序列读段和/或从分子家族中去除带有不相符的样品条形码的核酸分子。

64.如权利要求61所述的方法，其中所述识别带有不相符的样品条形码的核酸分子包括识别错误引发的核酸分子。

65.如权利要求64所述的方法，其中所述错误引发的核酸分子用适当的条形码来校正并用于提高序列质量。

66.如权利要求65所述的方法，其中将带有经校正的条形码的核酸分子分配到经校正

的读段家族。

67.如权利要求66所述的方法，其中将所述经校正的读段家族用于准确测定不同的覆盖度。

68.如权利要求67所述的方法，其中将不同的覆盖度测定用于评估核酸分子的文库。

69.如权利要求60所述的方法，其还包括基于分子索引位置的定位和在每个分子索引位置处的核苷酸将序列读段分配到分子家族。

70.如权利要求69所述的方法，其中识别错误序列读段包括识别被分配到多个分子家族的核酸分子。

71.如权利要求70所述的方法，其还包括从序列读段和/或从分子家族中去除被分配到多个分子家族的核酸分子。

不固定条形码

[0001] 相关申请的交叉引用

[0002] 本申请根据35U.S.C.§119(e)要求享有于2020年4月7日提交的美国临时申请号63/006,556的优先权。在先申请的公开被认为是本申请公开的一部分并通过引用并入。

[0003] 序列表的并入

[0004] 所附序列表中的材料在此通过引用并入本申请。所附序列表文本文件命名为PGDX3120-1WO_SL.txt,创建于2021年3月31日,大小为11kb。该文件可在使用Windows操作系统的计算机上利用Microsoft Word访问。

发明领域

[0005] 本发明总体涉及核酸序列,更具体地涉及用于标记和分析核酸分子的被称为条形码的序列。

[0006] 发明背景

[0007] 条形码经常被用于标记核酸,如被测序以鉴定其来源的DNA或RNA分子。可将条形码用于标记样品、细胞或DNA或RNA分子的其他来源。条形码可提供有关分子来源的信息,以及特定分子是否由于扩增而在一个池中被多次测序。通常,需要多条信息,例如样品和分子来源。来源越复杂,创建足够数量的条形码和/或条形码读段同时保证序列正确并避免来源的错误分配就越具有挑战性。具体地,例如,条形码数量不足以及难以校正复杂条形码中的序列错误,限制了对核酸分子如来自合并样品的核酸的基因组的分析。因此,需要新的条形码化核酸的系统和方法,其实现对核酸的多重基因组分析和改进的错误校正,以使由于条形码序列不确定性导致的错误分配和序列读段丢失最小化。

[0008] 发明概述

[0009] 本发明涉及用于标记和分析包括具有预先确定数量的索引位置的索引“条形码”的核酸分子。还提供了用于标记和分析核酸分子的方法。

[0010] 在一个实施方案中,本发明提供了用于标记样品中核酸分子的系统,其包括:包括多种条形码的寡核苷酸组,每种条形码包含一段连续碱基,该段连续碱基包括:(i)包括预先确定数量的样品索引位置的样品条形码,所述样品索引位置包含一个或多个特定核苷酸,其中样品索引位置的定位在样品之间变化;和(ii)包括分子索引位置的分子条形码,所述分子索引位置包含与样品索引位置处的核苷酸不同的核苷酸,其中样品索引位置散布在分子索引位置中。一方面,在本文提供的用于标记核酸分子的系统中,样品条形码位置的预先确定数量可在不同样品条形码之间变化。在一些方面,条形码包含约10至约35个核苷酸。在其他方面,条形码包含约12至约25个核苷酸。另一方面,样品条形码包含2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个样品索引位置或其组合。在一些方面,样品条形码包含约4至约12个样品索引位置。在其他方面,分子条形码包含约5至约25个分子索引位置。在不同方面,分子条形码包括约5至约15个分子索引位置。一方面,样品索引位置核苷酸和分子索引位置核苷酸选自以下:(A)样品索引位置核苷酸为A,并且分子索引位置核苷酸为C、G、T或其组合;(B)样品索引位置核苷酸为T,并且分子索引位置核苷酸为C、G、A或其组

合;(C)样品索引位置核苷酸为C,并且分子索引位置核苷酸是G、A、T或其组合;(D)样品索引位置核苷酸为G,并且分子索引位置核苷酸为C、A、T或其组合;(E)样品索引位置核苷酸为A、T或其组合,并且分子索引位置核苷酸为C、G或其组合;(F)样品索引位置核苷酸为A、C或其组合,并且分子索引位置核苷酸为T、G或其组合;(G)样品索引位置核苷酸为A、G或其组合,并且分子索引位置核苷酸为T、C或其组合;(H)样品索引位置核苷酸为T、C或其组合,并且分子索引位置核苷酸为A、G或其组合;(I)样品索引位置核苷酸为T、G或其组合,并且分子索引位置核苷酸为A、C或其组合;或者(J)样品索引位置核苷酸为G、C或其组合,并且分子索引位置核苷酸为A、T或其组合。在一些方面,每个条形码包括含有索引位置的一个或多个附加索引条形码。在许多方面,所述一个或多个附加索引条形码是细胞条形码、提供对未修复端DNA长度的量度的条形码、或者细胞条形码和提供对未修复端DNA长度的量度的条形码二者。在其他方面,寡核苷酸组中的每种寡核苷酸还包括非条形码位置,所述非条形码位置包括用于杂交的位点、用于序列引物结合的位点、用于扩增的位点、或其任何组合。

[0011] 在另一个实施方案中,本发明提供了用于标记样品中核酸分子的寡核苷酸组,其包括多种条形码,每种条形码包括:(i)包括预先确定数量的样品索引位置的样品条形码,所述样品索引位置包含一个或多个特定核苷酸,其中样品索引位置的定位在样品之间变化;和(ii)包括分子索引位置的分子条形码,所述分子索引位置包含与样品索引位置处的核苷酸不同的核苷酸,其中样品索引位置和分子索引位置散布在一段连续碱基中。一方面,预先确定的样品条形码位置的数量在不同样品条形码之间变化。在一些方面,条形码包含约10至约35个核苷酸。在其他方面,条形码包含约12至约25个核苷酸。另一方面,样品条形码包含2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个样品索引位置或其组合。在一些方面,样品条形码包括约4至约12个样品索引位置。一方面,分子条形码包括约5至约25个分子索引位置。在一些方面,分子条形码包括约5至约15个分子索引位置。在其他方面,样品索引位置核苷酸和分子索引位置核苷酸选自以下:(A)样品索引位置核苷酸为A,并且分子索引位置核苷酸为C、G、T或其组合;(B)样品索引位置核苷酸为T,并且分子索引位置核苷酸为C、G、A或其组合;(C)样品索引位置核苷酸为C,并且分子索引位置核苷酸是G、A、T或其组合;(D)样品索引位置核苷酸为G,并且分子索引位置核苷酸为C、A、T或其组合;(E)样品索引位置核苷酸为A、T或其组合,并且分子索引位置核苷酸为C、G或其组合;(F)样品索引位置核苷酸为A、C或其组合,并且分子索引位置核苷酸为T、G或其组合;(G)样品索引位置核苷酸为A、G或其组合,并且分子索引位置核苷酸为T、C或其组合;(H)样品索引位置核苷酸为T、C或其组合,并且分子索引位置核苷酸为A、G或其组合;(I)样品索引位置核苷酸为T、G或其组合,并且分子索引位置核苷酸为A、C或其组合;或者(J)样品索引位置核苷酸为G、C或其组合,并且分子索引位置核苷酸为A、T或其组合。在一些方面,每个条形码包括含有索引位置的一个或多个附加索引条形码。在许多方面,所述一个或多个附加索引条形码是细胞条形码、提供未修复端DNA长度的量度的条形码、或者细胞条形码和提供未修复端DNA长度的量度的条形码二者。在一些方面,寡核苷酸组中的每个寡核苷酸还包含非条形码位置,所述非条形码位置包括用于杂交的位点、用于序列引物结合的位点、用于扩增的位点、或其任何组合。

[0012] 在又一个实施方案中,本发明提供了用于分析样品中核酸分子的序列的方法,其包括:(a)将多个寡核苷酸附接到核酸分子,其中所述每个寡核苷酸包括条形码,所述条形

码包括:(i)包括预先确定数量的样品索引位置的样品条形码,所述样品索引位置包含一个或多个特定核苷酸,其中样品索引位置的定位在样品之间变化;和(ii)包括分子索引位置的分子条形码,所述分子索引位置包含与样品索引位置处的核苷酸不同的核苷酸,其中样品索引位置和分子索引位置散布于一段连续碱基中;以及(b)对核酸分子进行测序,其中序列读段包括条形码序列。一方面,本文提供的分析样品中核酸分子序列的方法可进一步包括将包括相同的样品条形码的寡核苷酸附接到该样品中核酸分子的每一端。另一方面,样品条形码位置的预先确定数量在不同样品条形码之间变化。在一些方面,条形码包含约10至约35个核苷酸。在其他方面,条形码包含约12至约25个核苷酸。在一些方面,样品条形码包括2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个样品索引位置或其组合。在其他方面,样品条形码包括约4至约12个样品索引位置。一方面,分子条形码包括约5至约25个分子索引位置。在一些方面,分子条形码包括约5至约15个分子索引位置。一方面,样品索引位置核苷酸和分子索引位置核苷酸选自以下:(A)样品索引位置核苷酸为A,并且分子索引位置核苷酸为C、G、T或其组合;(B)样品索引位置核苷酸为T,并且分子索引位置核苷酸为C、G、A或其组合;(C)样品索引位置核苷酸为C,并且分子索引位置核苷酸是G、A、T或其组合;(D)样品索引位置核苷酸为G,并且分子索引位置核苷酸为C、A、T或其组合;(E)样品索引位置核苷酸为A、T或其组合,并且分子索引位置核苷酸为C、G或其组合;(F)样品索引位置核苷酸为A、C或其组合,并且分子索引位置核苷酸为T、G或其组合;(G)样品索引位置核苷酸为A、G或其组合,并且分子索引位置核苷酸为T、C或其组合;(H)样品索引位置核苷酸为T、C或其组合,并且分子索引位置核苷酸为A、G或其组合;(I)样品索引位置核苷酸为T、G或其组合,并且分子索引位置核苷酸为A、C或其组合;或者(J)样品索引位置核苷酸为G、C或其组合,并且分子索引位置核苷酸为A、T或其组合。在其他方面,每个条形码包括含有索引位置的一个或多个附加索引条形码。在一些方面,所述一个或多个附加索引条形码是细胞条形码、提供未修复端DNA长度的量度的条形码、或者细胞条形码和提供未修复端DNA长度的量度的条形码二者。在一些方面,本文提供的分析样品中核酸分子的序列的方法还包括基于样品索引位置的定位将序列读段分配到样品家族。在其他方面,本文提供的分析样品中核酸分子的序列的方法还包括基于分子索引位置的定位和每个分子索引位置处的核苷酸将序列读段分配到分子家族。在一些方面,本文提供的分析样品中核酸分子的序列的方法还包括通过比较序列读段中的样品索引位置的数量和定位与预先确定的样品索引位置的数量和定位来校正测序错误。在其他方面,本文提供的分析样品中核酸分子的序列的方法还包括通过比较序列读段两端的样品条形码来校正测序错误。在一些方面,本文提供的分析样品中核酸分子的序列的方法还包括应用规则以将序列读段每一端的不同样品条形码与允许的样品条形码进行比较。在其他方面,本文提供的分析样品中核酸分子的序列的方法还包括应用一个或多个规则以(1)校正条形码内的错误,(2)校正核酸分子每一端的条形码之间的错误,(3)将序列读段多路分解到样品家族,(4)将序列读段分配到分子家族,或其任何组合。在一些方面,每个寡核苷酸还包括非条形码位置,所述非条形码位置包括杂交位点、序列引物结合位点、扩增位点、或其任何组合。在其他方面,本文提供的分析样品中核酸分子的序列的方法还包括使用不同基因组并测试每个寡核苷酸以灵敏地检测序列读段错误分配。在一些方面,本文提供的分析样品中核酸分子的序列的方法还包括在没有多路分解的情况下储存核酸序列数据。

[0013] 在一个实施方案中，本文提供了用于标记样品中核酸分子的方法，其包括：将多个寡核苷酸附接到包括条形码的核酸分子，每个条形码包括：(i)包括预先确定数量的样品索引位置的样品条形码，所述样品索引位置包含一个或多个特定核苷酸，其中样品索引位置的定位在样品之间变化；和(ii)包括分子索引位置的分子条形码，所述分子索引位置包含与样品索引位置处的核苷酸不同的核苷酸，其中样品索引位置和分子索引位置散布于一段连续碱基中。一方面，本文提供的标记样品中核酸分子的方法还可包括将包含相同的样品条形码的寡核苷酸附接到核酸分子的每一端。在一些方面，样品条形码位置的预先确定数量在不同样品条形码之间变化。在其他方面，条形码包含约10至约35个核苷酸。在不同方面，条形码包含约12至约25个核苷酸。在一些方面，样品条形码包括2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个样品索引位置。在其他方面，样品条形码包括约4至约12个样品索引位置。在不同方面，分子条形码包括约5至约25个分子索引位置。在一些方面，分子条形码包括约5至约15个分子索引位置。一方面，样品索引位置核苷酸和分子索引位置核苷酸选自：(A)样品索引位置核苷酸为A，并且分子索引位置核苷酸为C、G、T或其组合；(B)样品索引位置核苷酸为T，并且分子索引位置核苷酸为C、G、A或其组合；(C)样品索引位置核苷酸为C，并且分子索引位置核苷酸是G、A、T或其组合；(D)样品索引位置核苷酸为G，并且分子索引位置核苷酸为C、A、T或其组合；(E)样品索引位置核苷酸为A、T或其组合，并且分子索引位置核苷酸为C、G或其组合；(F)样品索引位置核苷酸为A、C或其组合，并且分子索引位置核苷酸为T、G或其组合；(G)样品索引位置核苷酸为A、G或其组合，并且分子索引位置核苷酸为T、C或其组合；(H)样品索引位置核苷酸为T、C或其组合，并且分子索引位置核苷酸为A、G或其组合；(I)样品索引位置核苷酸为T、G或其组合，并且分子索引位置核苷酸为A、C或其组合；或(J)样品索引位置核苷酸为G、C或其组合，并且分子索引位置核苷酸为A、T或其组合。在一些方面，每个条形码包括含有索引位置的一个或多个附加索引条形码。在多个方面，所述一个或多个附加条形码是细胞条形码、提供未修复端DNA长度的量度的条形码、或者细胞条形码和提供未修复端DNA长度的量度的条形码二者。在一些方面，每个寡核苷酸还包括非条形码位置，所述非条形码位置包括用于杂交的位点、用于序列引物结合的位点、用于扩增的位点、或其任何组合。在其他方面，本文提供的标记样品中核酸分子的方法还可包括测序标记的核酸分子。在一些方面，测序标记的核酸分子还包括在没有多路分解的情况下存储核酸序列数据。在多个方面，存储核酸序列数据而没有多路分解防止在没有多路分解密钥的情况下使用序列数据，并且防止未经授权使用数据。

[0014] 在另一个实施方案中，本发明提供了用于识别错误序列读段的方法，其包括：(a)将多个寡核苷酸附接于样品的核酸分子，其中每个寡核苷酸包括条形码，所述条形码包括：(i)包括预先确定数量的样品索引位置的样品条形码，所述样品索引位置包含一个或多个特定核苷酸，其中样品索引位置的定位在样品之间变化，并且其中将相同的样品条形码附接到该样品中核酸分子的每一端；和(ii)包括分子索引位置的分子条形码，所述分子索引位置包含与样品索引位置处的核苷酸不同的核苷酸，其中样品索引位置和分子索引位置散布于一段连续碱基中；以及(b)对核酸分子进行测序，其中序列读段包括条形码序列，从而识别错误序列读段。

[0015] 一方面，识别错误序列读段包括识别带有不相符的样品条形码的核酸分子。在一些方面，测序错误通过比较序列读段两端的样品条形码来进一步校正。在其他方面，进一步

从序列读段和/或从分子家族中去除带有不相符的样品条形码的核酸分子。另一方面，识别带有不相符样品条形码的核酸分子包括识别错误引发(misprimed)的核酸分子。在一些方面，错误引发的核酸分子用适当的条形码校正，并用于提高序列质量。在其他方面，将带有校正条形码的核酸分子分配到校正的读段家族。在多个方面，将校正的读段家族用于准确测定不同的覆盖度。在一些方面，将不同覆盖度测定用于评估核酸分子的文库。一方面，所述方法还包括基于分子索引位置的定位和每个分子索引位置处的核苷酸将序列读段分配到分子家族。在一些方面，识别错误序列读段包括识别被分配到多个分子家族的核酸分子。在其他方面，进一步从序列读段和/或从分子家族中去除被分配到多个分子家族的核酸分子。

[0016] 附图简述

[0017] 图1示出传统产品条形码相比三种不固定DNA条形码的比较。

[0018] 图2A示出使用7/14标准的数字格式的16种样品条形码。

[0019] 图2B示出从数字格式至核苷酸格式的转换，7/14标准。

[0020] 图2C示出单个样品条形码从简并序列至实际序列的转换，7/20bp格式。

[0021] 图3A示出标准条形码。

[0022] 图3B示出不固定条形码。

[0023] 图4示出具有标准条形码的人工嵌合分子的生成。

[0024] 图5示出人类序列读段与标准条形码(左)和不固定条形码(右)的比对。

[0025] 图6示出基于连接步骤中接头丰度的错误引发的水平。

[0026] 图7示出基于接头浓度的错误引发率i7:i5的比率。

[0027] 图8示出分子条形码序列重复的频率。

[0028] 发明详述

[0029] 本发明基于以下发现，即可将基于核苷酸定位而非序列的条形码用于识别核酸分子和序列读段并将其分组。

[0030] 例如，基于核苷酸定位而非基于序列的条形码具有灵活性，因为可为一种索引生成相对少的数量的条形码，而为另一种索引生成非常高数量的条形码，或为按条形码的两种或更多种索引生成高数量的条形码。此外，具有预先确定的索引位置的条形码允许改进的错误校正方法。

[0031] 用于标记核酸的寡核苷酸系统和组

[0032] 在一个实施方案中，本发明提供了用于标记样品中的核酸分子的系统，其包括：包括多种条形码的寡核苷酸组，每种条形码包含一段连续碱基，所述一段连续碱基包括：(i)包括预先确定数量的样品索引位置的样品条形码，所述样品索引位置包含一个或多个特定核苷酸，其中样品索引位置的定位在样品之间变化；和(ii)包括分子索引位置的分子条形码，所述分子索引位置包含与样品索引位置处的核苷酸(一个或多个)不同的核苷酸，其中分子索引位置散布在样品索引位置中。

[0033] 用于标记样品中核酸分子的系统包括寡核苷酸组。如本文所用，“寡核苷酸组”意指可一起使用的寡核苷酸的组或集合。因此，可将本文提供的用于标记样品中核酸分子的系统中的寡核苷酸组一起使用来标记核酸。还可将寡核苷酸组的亚组用于标记样品中核酸分子的系统中。如本文所用，“寡核苷酸亚组”仅涉及用于标记样品中核酸的寡核苷酸组中

的一部分或一些寡核苷酸。因此，可将包括在寡核苷酸组中的寡核苷酸的全部或部分用于标记样品中的核酸。

[0034] 如本文所用，“标记核酸分子”意指修饰核酸分子用于例如检测、识别、分析或纯化。在一些方面，通过将一个或多个寡核苷酸附接到核酸分子来标记核酸。可将寡核苷酸附接到核酸分子的末端。在一些方面，将寡核苷酸附接到核酸分子的两端。在其他方面，附接到核酸分子末端的寡核苷酸的序列不同。在一些方面，附接到核酸分子末端的寡核苷酸的样品索引是相同的。在其他方面，附接到核酸分子末端的寡核苷酸的分子索引不同。

[0035] 包括例如DNA、RNA和核酸片段在内的任何核酸分子都可被标记。可被标记的DNA来源包括，例如染色体DNA、质粒DNA、cDNA、无细胞DNA(cfDNA)、循环肿瘤DNA(ctDNA)及其任何片段。例如，可将标记的核酸用于制备核酸文库。在一些方面，该文库是基因组文库。包括标记的核酸分子的文库可通过将本文提供的寡核苷酸组或亚组通过例如末端修复、A-加尾和接头连接而附接到核酸分子来制备。在一些方面，省略末端修复和A加尾，并包括与特定单个索引或索引组相关的可变末端，以确定核酸分子(如，DNA分子)的原始末端。经标记的核酸分子和经标记核酸分子的文库可通过例如测序来分析。可将任何合适的测序方法分析经标记的核酸分子。

[0036] 样品

[0037] 可使用本文提供的用于标记核酸的系统和寡核苷酸组来标记样品中的核酸。可被标记的核酸可存在于任何样品或任何类型的样品中。在一些方面，样品是血液、唾液、血浆、血清、尿液或其他生物流体。其他示例性生物流体包括浆膜液、淋巴、脑脊液、粘膜分泌物、阴道液、腹水、胸膜液、心包液、腹膜液和腹腔液。在其他方面，样品是组织样品。在一些方面，样品是细胞样品或单细胞。可使用新鲜样品或储存样品，包括例如储存的冷冻样品、福尔马林固定石蜡包埋(FFPE)样品和通过任何其他方法保存的样品。

[0038] 样品可来自正常或健康的受试者。样品也可来自患有疾病或病症的受试者。可使用本文提供的系统和寡核苷酸组标记来自患有任何疾病或病症的受试者的样品中的核酸。在一些方面，该疾病或病症是癌症。在一些方面，样品是来自患有癌症的受试者的液体样品。在其他方面，样品是来自患有癌症的受试者的组织样品。在一些方面，样品是来自患有癌症的受试者的细胞样品。在其他方面，样品是癌症样品。癌症样品可以是来自实体肿瘤或液体肿瘤的样品。癌症可以是肾脏癌、肾癌、泌尿膀胱癌、前列腺癌、子宫癌、乳腺癌、宫颈癌、卵巢癌、肺癌、结肠癌、直肠癌、口腔癌、咽癌、胰腺癌、甲状腺癌、黑素瘤、皮肤癌、头颈癌、脑癌、造血系统癌症、白血病、淋巴瘤、骨癌、肌肉癌、肉瘤、横纹肌肉瘤等。

[0039] 可在样品中标记核酸。还可在标记之前从样品中提取、分离或纯化核酸。可使用任何合适的提取、分离或纯化方法。示例方法包括酚-氯仿萃取、硫氰酸胍-酚氯仿提取、凝胶纯化以及柱和珠的使用。可将商用试剂盒用于核酸的提取、分离或纯化。

[0040] 条形码

[0041] 本文提供的用于标记样品中核酸分子的寡核苷酸组可包括多种条形码，每种条形码包括：(i)包括预先确定数量的样品索引位置的样品条形码，所述样品索引位置包含一种或多种特定核苷酸，其中样品索引位置的定位在样品之间变化；和(ii)包括分子索引位置的分子条形码，所述分子索引位置包含与样品索引位置处的核苷酸不同的核苷酸，其中样品索引位置和分子索引位置散布在一段连续碱基中。

[0042] 条形码索引位置可包含一段连续碱基。如本文所用，“连续碱基”意指序列中彼此相邻的碱基。在一些方面，一段连续碱基可包括条形码或索引位置以及非条形码或非索引位置。在其他方面，一段连续碱基可包括条形码或索引位置而没有非条形码或非索引位置。在一些方面，预先确定的样品条形码位置数量在不同样品条形码之间变化。

[0043] 条形码可包括任意数量的核苷酸。例如，条形码可包含约10至约35个核苷酸。又如，条形码可包含约12至约25个核苷酸。再如，条形码可包含约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30、约31、约32、约33、约34、约35、约36、约37、约38、约39、约40或更多个核苷酸。再如，条形码可包含至少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、至少21、至少22、至少23、至少24、至少25、至少26、至少27、至少28、至少29、至少30、至少31、至少32、至少33、至少34、至少35、至少36、至少37、至少38、至少39、至少40或更多个核苷酸。

[0044] 索引位置

[0045] 本文提供的条形码可包括一种或多种索引位置。示例性索引位置包括样品索引位置、分子索引位置、DNA末端索引位置和细胞索引位置。例如，条形码可包括样品索引位置、DNA末端索引位置和分子索引位置。条形码还可包括样品索引位置、分子索引位置、细胞索引位置、DNA末端索引位置、或其任意组合。

[0046] 如本文所用，术语“索引位置”意指可用于识别核酸分子的起源或来源的条形码内的核苷酸位置。因此，索引位置允许基于产生序列读段的核酸分子的起源或来源将由核酸分子生成的序列读段分配到类别或群组。例如，样品索引位置可用于识别核酸分子来源的样品，并允许将由核酸分子生成的序列读段按样品类别分组。因此，由同一样品的核酸分子生成的序列读段可分组在一起。又如，分子索引位置可用于识别产生序列读段的核酸分子。因此，分子索引位置可用于将由同一核酸分子生成的序列读段分组在一起。再如，细胞索引位置可用于识别核酸分子来源的细胞，并允许将由核酸分子生成的序列读段按细胞类别分组。因此，来自同一细胞的核酸分子的序列读段可分组在一起。

[0047] DNA末端索引位置可表示例如，未修复的DNA末端的长度。可制备能够与尚未修复的不同DNA分子连接的具有不同延伸的寡核苷酸。可对不同长度的突出端(overhang)进行索引，以识别未修复DNA分子中存在的突出端的长度。在一些方面，在癌症样品中识别未修复的DNA分子中存在的不同长度的突出端。在其他方面，识别未修复DNA分子中存在的不同长度的突出端来识别或检测癌症。寡核苷酸可具有任何长度的延伸，包括1个核苷酸、2个核苷酸、3个核苷酸、4个核苷酸、5个核苷酸、6个核苷酸、7个核苷酸、8个核苷酸、9个核苷酸、10个核苷酸、11个核苷酸、12个核苷酸、13个核苷酸、14个核苷酸、15个核苷酸、16个核苷酸、17个核苷酸、18个核苷酸、19个核苷酸、20个核苷酸或更多的延伸。寡核苷酸还可具有5’或3’延伸。

[0048] 本文提供的条形码可包括样品条形码。样品条形码可包括预先确定数量的样品索引位置。如本文所用，“预先确定数量的样品索引位置”意指可将特定数量的位置分配给样品索引来识别核酸分子来源的样品。预先确定的样品索引位置的数量可在样品之间变化。样品索引位置的定位也可在样品之间变化。在一些方面，预先确定的样品索引位置的数量和样品索引位置的定位可在样品之间变化。因此，核酸分子和核酸分子产生的序列读段的

样品来源可通过形成样品条形码的样品索引位置的数量、样品索引位置的定位或样品索引位置的数量和位置二者来识别。

[0049] 在一些实施方案中，由于样品索引位置的定位在样品之间变化，样品条形码可以是“不固定(floating)”或“数字化”条形码。如本文所用，“不固定条形码”或“数字化条形码”是指具有以下索引位置的条形码，其定位在群组或类别之间变化。包括可在群组或类别之间变化的索引位置的任何条形码，例如包括样品索引位置的样品条形码、包括分子索引位置的分子条形码、包括细胞索引位置的细胞条形码等，可以是不固定条形码。例如，除了如上所述可变化的样品索引位置的定位之外，分子条形码的分子索引位置的定位可在产生序列读段的不同核酸分子之间变化。又如，细胞条形码的细胞索引位置的定位可在获自来自不同细胞的核酸分子的序列读段之间变化。

[0050] 在一些方面，样品条形码中预先确定数量的样品索引位置包含一种或多种特定核苷酸，所述核苷酸定义其对应的索引类型。例如，在预先确定数量的样品索引位置中的一种或多种特定核苷酸可以是A、T、G或C。又如，在预先确定数量的样品索引位置中的一种或多种特定核苷酸可以是A和T、A和C、A和G、T和C、T和G、或G和C。

[0051] 在一些方面，样品条形码包括2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20或更多个样品索引位置或其组合。在一些方面，样品条形码包括约4至约12个样品索引位置。在其他方面，样品条形码包括约2、约3、约4、约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20或更多个样品索引位置或其组合。在一些方面，样品条形码包括至少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或更多个样品索引位置或其组合。

[0052] 本文提供的条形码可包括分子条形码。分子条形码可包括分子索引位置，其包含与样品索引位置处的核苷酸不同的核苷酸。例如，样品索引位置核苷酸和分子索引位置核苷酸可选自：(A)样品索引位置核苷酸为A，并且分子索引位置核苷酸为C、G、T或其组合；(B)样品索引位置核苷酸为T，并且分子索引位置核苷酸为C、G、A或其组合；(C)样品索引位置核苷酸为C，并且分子索引位置核苷酸为G、A、T或其组合；(D)样品索引位置核苷酸为G，并且分子索引位置核苷酸为C、A、T或其组合；(E)样品索引位置核苷酸为A、T或其组合，并且分子索引位置核苷酸为C、G或其组合；(F)样品索引位置核苷酸为A、C或其组合，并且分子索引位置核苷酸是T、G或其组合；(G)样品索引位置核苷酸为A、G或其组合，并且分子索引位置核苷酸为T、C或其组合；(H)样品索引位置核苷酸为T、C或其组合，并且分子索引位置核苷酸为A、G或其组合；(I)样品索引位置核苷酸为T、G或其组合，并且分子索引位置核苷酸为A、C或其组合；或(J)样品索引位置核苷酸为G、C或其组合，并且分子索引位置核苷酸为A、T或其组合。

[0053] 本文提供的样品条形码的样品索引位置可与分子索引位置穿插。因此，本文提供的条形码可包括不需要限制于核苷酸的特定连续段或块的样品索引位置和分子索引位置。例如，并非所有的样品索引位置都需要彼此相邻，也并非所有的分子索引位置都需要彼此相邻。样品索引位置和分子索引位置可以交替。任意数量的分子索引位置可位于样品索引位置之间。任意数量的分子索引位置可位于任意数量的样品索引位置之间。任意数量的分子索引位置和任意数量的非分子索引或其他索引位置的核苷酸可位于样品索引位置之间。任意数量的分子索引位置和任意数量的非分子索引或其他索引位置的核苷酸可位于任意

数量的样品索引位置之间。任意数量的非样品索引位置或非分子索引位置的核苷酸可位于样品索引位置和分子索引位置之间。

[0054] 一些样品索引位置可彼此相邻，而其他样品索引位置可定位于邻近条形码中的非样品索引位置的任何其他核苷酸。例如，样品索引位置和分子索引位置可以是不需要所有样品索引位置彼此相邻的任何配置。例如，样品索引位置和分子索引位置可以是不需要所有分子索引位置彼此相邻的任何配置。例如，样品索引位置和分子索引位置也可以是不需要所有样品索引位置和所有分子索引位置彼此相邻的任何配置。任何索引条形码的位置可以是不需要索引条形码的所有核苷酸彼此相邻的任何配置。示例性条形码索引包括样品条形码、分子条形码、细胞条形码等。

[0055] 本文提供的分子条形码可包括约5至约25个分子索引位置。在一些方面，本文提供的分子条形码包括约5至约15个分子索引位置。在其他方面，本文提供的分子条形码包括约2、约3、约4、约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30或更多个分子索引位置。在一些方面，本文提供的分子条形码包括至少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、至少21、至少22、至少23、至少24、至少25、至少26、至少27、至少28、至少29、至少30或更多个分子索引位置。在一些方面，本文提供的分子条形码包括约20个分子索引位置或少于约20个分子索引位置。

[0056] 本文提供的条形码可包括含有索引位置的一种或多种附加索引条形码。在一些方面，一种或多种附加索引条形码是细胞条形码。因此，本文提供的条形码可包括样品条形码、分子条形码、细胞条形码、提供未修复DNA末端长度的量度的条形码、任何其他索引条形码、或其任何组合。因此，本文提供的条形码可包括例如样品索引位置、分子索引位置和任何其他索引位置如细胞索引位置，这些位置散布在彼此中。本文提供的条形码的索引位置不需要限制于特定的连续核苷酸段或块。索引条形码和索引位置可以是不需要所有索引位置彼此相邻的任何配置。

[0057] 寡核苷酸组中的每种寡核苷酸还可包括非条形码位置。寡核苷酸中包括的非条形码位置可包括用于杂交的位点、用于扩增的位点、用于序列引物结合的位点、以及用于杂交、序列引物结合和扩增的位点。用于杂交、序列引物结合的位点和用于扩增的位点可包括约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30或更多个核苷酸。用于杂交的位点可包括例如探针结合位点。用于扩增的位点可包括例如引物结合位点。用于杂交、序列引物结合的位点和用于扩增的位点可以彼此不同。用于杂交、序列引物结合的位点和用于扩增的位点也可以重叠。用于杂交、序列引物结合的位点和用于扩增的位点可在任何程度上重叠。在一些方面，用于杂交、序列引物结合的位点和用于扩增的位点重叠约1、约2、约3、约4、约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30、或更多个核苷酸。在一些方面，用于杂交、序列引物结合的位点和用于扩增的位点完全重叠。在其他方面，用于杂交、序列引物结合的位点和用于扩增的位点没有重叠。

[0058] 用于分析核酸序列的方法

[0059] 在另一个实施方案中，本发明提供了分析样品中核酸分子的序列的方法。本文提供的分析核酸序列的方法可包括(a)将多个寡核苷酸附接到核酸分子，其中每个寡核苷酸包括条形码，所述条形码包括：(i)包括预先确定数量的样品索引位置的样品条形码，所述样品索引位置包含一个或多个特定核苷酸，其中样品索引位置的定位在样品之间变化；和(ii)包括分子索引位置的分子条形码，所述分子索引位置包含与样品索引位置处的核苷酸不同的核苷酸，其中样品索引位置和分子索引位置散布在一段连续碱基中；以及(b)对核酸分子进行测序，其中一些序列读段包括条形码序列。

[0060] 本文提供的分析核酸序列的方法可包括将多个寡核苷酸附接到核酸分子。可附接的多个寡核苷酸可包括寡核苷酸组。在一些方面，可附接的多个寡核苷酸包括寡核苷酸亚组。本文提供的任何寡核苷酸，包括寡核苷酸组和亚组，可用于本文提供的分析核酸分子或其片段的序列的方法。因此，可附接的多个寡核苷酸中的每个寡核苷酸可包括预先确定数量的样品索引位置，其包含一个或多个特定核苷酸。预先确定数量的样品索引位置的定位可在样品之间变化。多个寡核苷酸中的每个寡核苷酸还可包括含有分子索引位置的分子条形码。分子索引位置可包含与样品索引位置处的核苷酸不同的核苷酸。样品索引位置和分子索引位置可散布在一段连续碱基中。

[0061] 在其他方面，本文提供的分析核酸分子序列的方法包括将包括相同样品条形码的寡核苷酸附接到核酸分子的每一端。在一些方面，预先确定数量的样品条形码位置在不同样品条形码之间变化。在包括相同样品条形码的寡核苷酸中可能不存在连续的相同碱基段，因为包括在样品条形码中的核苷酸可与以下核苷酸穿插：包括在分子条形码中或构成分子索引位置的核苷酸，包括在细胞条形码中或构成细胞索引位置的核苷酸，包括在任何其他索引条形码中或构成任何其他索引位置的核苷酸，未包括在索引条形码中或未构成索引位置的核苷酸，或其任何组合。因此，在一些方面，附接到包括相同样品条形码的核酸分子的每一端的寡核苷酸不会交叉杂交，并且不会导致产生诸如嵌合分子的假产物(例如在扩增过程中)。在一些方面，本文提供的分析核酸分子序列的方法包括将包括不同样品条形码的寡核苷酸附接到核酸分子的每一端。

[0062] 一方面，本文提供的分析核酸分子序列的方法包括将包括相同分子条形码的寡核苷酸附接到核酸分子的每一端。在包括相同分子条形码的寡核苷酸中可能不存在连续的相同碱基段，因为包括在分子条形码中的核苷酸可与以下核苷酸穿插：包括在样品条形码中或构成样品索引位置的核苷酸，包括在细胞条形码中或构成细胞索引位置的核苷酸，包括在任何其他索引条形码中或构成任何其他索引位置的核苷酸，未包括在索引条形码中或未构成索引位置的核苷酸，或其任何组合。因此，在一些方面，附接到包括相同分子条形码的核酸分子的每一端的寡核苷酸不会交叉杂交，并且不会导致产生诸如嵌合分子的假产物(例如在扩增过程中)。在其他方面，本文提供的方法包括将包括不同分子条形码的寡核苷酸附接到核酸分子的每一端。

[0063] 在一些方面，本文提供的分析核酸分子序列的方法包括将包括相同样品条形码和相同分子条形码的寡核苷酸附接到核酸分子的每一端。在包括相同样品条形码和相同分子条形码的寡核苷酸中可能不存在连续的相同碱基段，因为包括在样品条形码和分子条形码中的核苷酸可与以下核苷酸穿插：包括在细胞条形码中或构成细胞索引位置的核苷酸，包括在任何其他索引条形码中或构成任何其他索引位置的核苷酸，未包括在索引条形码中或

未构成索引位置的核苷酸，或其任何组合。因此，在一些方面，附接到包括相同样品条形码和相同分子条形码的核酸分子的每一端的寡核苷酸不会交叉杂交，并且不会导致产生诸如嵌合分子的假产物(例如在扩增过程中)。在其他方面，本文提供的方法包括将包括不同样品条形码和不同分子条形码的寡核苷酸附接到核酸分子的每一端。

[0064] 在一些方面，本文提供的分析核酸分子序列的方法包括将寡核苷酸附接到样品中核酸分子的每一端，其中寡核苷酸包含相同样品条形码、相同分子条形码、相同细胞条形码、提供未修复DNA末端长度的量度的相同条形码、包含任何其他索引核苷酸的相同索引条形码、或其任何组合。包括样品条形码、分子条形码、细胞条形码、包括任何其他索引位置的核苷酸或索引条形码、或其任何组合的条形码中由于穿插的核苷酸可能不存在连续的相同碱基段。穿插的核苷酸可包括以下核苷酸：未包括在索引条形码中的核苷酸、未构成索引位置的核苷酸、或包括在核苷酸穿插的索引条形码以外的索引条形码中的核苷酸或构成核苷酸穿插的索引位置以外的索引位置的核苷酸。因此，可以防止扩增过程中交叉杂交和产生诸如嵌合分子的假产物。一方面，本文提供的方法包括将包括不同样品条形码、不同分子条形码、不同细胞条形码、包括任何其他索引核苷酸的不同索引条形码、或其任何组合的寡核苷酸附接到样品中核酸分子的每一端。

[0065] 任何合适的方法都可用于将包括条形码的寡核苷酸附接到核酸分子的末端。在多个方面，寡核苷酸是共价附接的。

[0066] 本文提供的分析核酸分子序列的方法中的条形码可包括任意数量的核苷酸。例如，本文提供的分析核酸分子序列的方法中的条形码可包括约10至约35个核苷酸。又如，本文提供的分析核酸分子序列的方法中的条形码可包括约12至约25个核苷酸。再如，本文提供的分析核酸分子序列的方法中的条形码可包括约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30、约31、约32、约33、约34、约35、约36、约37、约38、约39、约40或更多个核苷酸。又如，本文提供的分析核酸分子序列的方法中的条形码可包括至少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、至少21、至少22、至少23、至少24、至少25、，至少26、至少27、至少28、至少29、至少30、至少31、至少32、至少33、至少34、至少35、至少36、至少37、至少38、至少39、至少40或更多个核苷酸。

[0067] 本文提供的分析核酸分子序列的方法中的条形码可包括一个或多个索引位置。示例性索引位置包括样品索引位置、分子索引位置和细胞索引位置。例如，本文提供的分析核酸分子序列的方法中的条形码可包括样品索引位置和分子索引位置。本文提供的分析核酸分子序列的方法中的条形码还可包括样品索引位置、分子索引位置、细胞索引位置、提供未修复DNA末端长度的量度的索引位置、或其任意组合。

[0068] 本文提供的分析核酸分子序列的方法中的条形码可包括样品条形码。样品条形码可包括预先确定数量的样品索引位置。预先确定的样品索引位置的数量可在样品之间变化。样品索引位置的定位也可在样品之间变化。在一些方面，预先确定的样品索引位置的数量和样品索引位置的定位可在样品之间变化。因此，核酸分子和核酸分子产生的序列读段的样品来源可通过形成样品条形码的样品索引位置的数量、样品索引位置的定位、或者样品索引位置的数量和定位二者来识别。

[0069] 本文提供的分析核酸分子序列的方法中的样品条形码中预先确定数量的样品索引位置可包含一个或多个特定核苷酸。例如，在预先确定数量的样品索引位置中的一个或多个特定核苷酸可以是A、T、G或C。又如，在预先确定数量的样品索引位置中的一个或多个特定核苷酸可以是A和T、A和C、A和G、T和C、T和G、或G和C。

[0070] 在一些方面，本文提供的分析核酸分子序列的方法中的样品条形码包括2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20或更多个样品索引位置或其组合。在一些方面，本文提供的分析核酸分子序列的方法中的样品条形码包括约4至12个样品索引位置。在多个方面，本文提供的分析核酸分子序列的方法中的样品条形码包括约2、约3、约4、约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20或更多个样品索引位置或其组合。一方面，本文提供的分析核酸分子序列的方法中的样品条形码包括至少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、至少10个、至少11个、至少12个、至少13个、至少14个、至少15个、至少16个、至少17个、至少18个、至少19个、至少20个或更多个样品索引位置或其组合。

[0071] 本文提供的分析核酸分子序列的方法中的条形码可包括分子条形码。本文提供的分析核酸分子序列的方法中的分子条形码可包括分子索引位置，所述分子索引位置包含与样品索引位置处的核苷酸不同的核苷酸。例如，样品索引位置核苷酸和分子索引位置核苷酸可选自：(A)样品索引位置核苷酸为A，并且分子索引位置核苷酸为C、G、T或其组合；(B)样品索引位置核苷酸为T，并且分子索引位置核苷酸为C、G、A或其组合；(C)样品索引位置核苷酸为C，并且分子索引位置核苷酸为G、A、T或其组合；(D)样品索引位置核苷酸为G，并且分子索引位置核苷酸为C、A、T或其组合；(E)样品索引位置核苷酸为A、T或其组合，并且分子索引位置核苷酸为C、G或其组合；(F)样品索引位置核苷酸为A、C或其组合，并且分子索引位置核苷酸为T、G或其组合；(G)样品索引位置核苷酸为A、G或其组合，并且分子索引位置核苷酸为T、C或其组合；(H)样品索引位置核苷酸为T、C或其组合，并且分子索引位置核苷酸为A、G或其组合；(I)样品索引位置核苷酸为T、G或其组合，并且分子索引位置核苷酸为A、C或其组合；或(J)样品索引位置核苷酸为G、C或其组合，并且分子索引位置核苷酸为A、T或其组合。

[0072] 本文提供的分析核酸分子序列的方法中的样品条形码的样品索引位置可与分子索引位置穿插。因此，本文提供的分析核酸分子序列的方法中的条形码可包括不需要限制于特定的连续核苷酸段或块的样品索引位置和分子索引位置。例如，并非所有的样品索引位置都需要彼此相邻，也并非所有的分子索引位置都需要彼此相邻。样品索引位置和分子索引位置可以交替。任意数量的分子索引位置可位于样品索引位置之间。任意数量的分子索引位置可位于任意数量的样品索引位置之间。任意数量的分子索引位置和任意数量的非分子索引或其他索引位置的核苷酸可位于样品索引位置之间。任意数量的分子索引位置和任意数量的非分子索引或其他索引位置的核苷酸可位于任意数量的样品索引位置之间。任意数量的非样品索引位置或非分子索引位置的核苷酸可位于样品索引位置和分子索引位置之间。

[0073] 一些样品索引位置可彼此相邻，而其他样品索引位置可定位于邻近非样品索引位置的条形码中任何其他核苷酸。例如，样品索引位置和分子索引位置可以是不需要所有样品索引位置彼此相邻的任何配置。例如，样品索引位置和分子索引位置可以是不需要所有分子索引位置彼此相邻的任何配置。例如，样品索引位置和分子索引位置也可以是不需要

所有样品索引位置和所有分子索引位置彼此相邻的任何配置。任何索引条形码的位置可以是不需要索引条形码的所有核苷酸彼此相邻的任何配置。示例性条形码索引包括样品条形码、分子条形码、细胞条形码等。

[0074] 本文提供的分析核酸分子序列的方法中的分子条形码可包括约5至25个分子索引位置。一方面，本文提供的分析核酸分子序列的方法中的分子条形码包括约5至约15个分子索引位置。在一些方面，本文提供的分析核酸分子序列的方法中的分子条形码包括约2、约3、约4、约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30或更多个分子索引位置。在其他方面，本文提供的分析核酸分子序列的方法中的分子条形码包括至少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、至少21、至少22、至少23、至少24、至少25、至少26、至少27、至少28、至少29、至少30或更多个分子索引位置。

[0075] 本文提供的分析核酸分子序列的方法中的每个条形码可包括含有索引位置的一个或多个附加索引条形码。在一些方面，一个或多个附加索引条形码是细胞条形码。因此，本文提供的分析核酸分子序列的方法中的条形码可包括样品条形码、分子条形码、细胞条形码、任何其他索引条形码或其任何组合。因此，本文提供的分析核酸分子序列的方法中的条形码可包括例如样品索引位置、分子索引位置和任何其他索引位置如细胞索引位置，这些位置彼此穿插。本文提供的条形码的索引位置不需要限制于特定的连续核苷酸段或块。索引条形码和索引位置可以是不需要所有索引位置彼此相邻的任何配置。

[0076] 附接有本文提供的寡核苷酸的核酸分子可通过例如测序进行分析。获得的序列读段可包括条形码序列。任何合适的测序方法都可用于分析核酸分子。示例性测序方法包括例如下一代测序(NGS)。示例性NGS方法包括Roche454测序仪、Life Technologies SOLiD系统、Life Techologies Ion Torrent、BGI/MGI系统、Genapsys系统、以及Illumina系统，如Illuminia基因组分析仪II、Illumia MiSeq、Illum ina HiSeq、Elumina NextSeq和Illumna NovaSeq仪。可对每个核苷酸进行深度覆盖度的测序，包括，例如至少2倍覆盖度，至少10倍覆盖度；至少20倍的覆盖度；至少30倍的覆盖度；至少40倍覆盖度；至少50倍的覆盖度；至少60倍覆盖度；至少70倍覆盖度；至少80倍覆盖度；至少90倍覆盖度；至少100倍覆盖度；至少200倍的覆盖度；至少300倍覆盖度；至少400倍覆盖度；至少500倍覆盖度；至少600倍覆盖度；至少700倍覆盖度率；至少800倍覆盖度；至少900倍覆盖度；至少1,000倍覆盖度；至少2,000倍覆盖度；至少3,000倍覆盖度；至少4,000倍覆盖度；至少5,000倍覆盖度；至少6,000倍覆盖；至少7,000倍覆盖度；至少8,000倍覆盖度；至少9,000倍覆盖度；至少10,000倍覆盖度；至少15,000倍覆盖度；至少20,000倍覆盖度；以及其间的任意数值或范围。

[0077] 在一些方面，测序包括全基因组测序。在多个方面，测序包括外显子组测序或靶向面板(panel)。如本文所用，术语“外显子组测序”是指对基因组中基因的所有蛋白编码外显子进行测序。外显子组测序可包括靶富集方法，例如核酸的基于阵列的捕获和溶液中的捕获。靶向面板包括感兴趣区域的亚组，并且可包括蛋白编码区和非编码区两者。

[0078] 可使用本文提供的方法分析任何样品或样品类型中的核酸序列。在一些方面，样品是血液、唾液、血浆、血清、尿液或其他生物流体。其他示例性生物流体包括浆膜液、淋巴、脑脊液、粘膜分泌物、阴道液、腹水、胸膜液、心包液、腹膜液和腹腔液。在一些方面，样品是

组织样品。在其他方面，样品是细胞样品。可使用新鲜样品或储存的样品，包括例如储存的冷冻样品、福尔马林固定石蜡包埋(FFPE)样品以及通过任何其他方法保存的样品。

[0079] 样品可来自正常或健康受试者。样品也还可来自患有疾病或病症的受试者。可使用本文提供的方法分析来自患有任何疾病或病症的受试者的样品中的核酸序列。在一些方面，疾病或病症是癌症。在其他方面，样品是来自患有癌症的受试者的流体样品。在一些方面，样品是来自患有癌症的受试者的组织样品。在其他方面，样品是来自患有癌症的受试者的细胞样品。在一些方面，样品是癌症样品。癌症样品可以是来自实体肿瘤或液体肿瘤的样品。癌症可以是肾脏癌、肾癌、泌尿膀胱癌、前列腺癌、子宫癌、乳腺癌、宫颈癌、卵巢癌、肺癌、结肠癌、直肠癌、口腔癌、咽癌、胰腺癌、甲状腺癌、黑素瘤、皮肤癌、头颈癌、脑癌、造血系统癌、白血病、淋巴瘤、骨癌、肌肉癌、肉瘤、横纹肌肉瘤等。

[0080] 可在测序前从样品中提取、分离或纯化核酸。可使用任何合适的提取、分离或纯化方法。示例性方法包括酚-氯仿提取、硫氰酸胍-酚-氯仿萃取、凝胶纯化以及柱和珠的使用。商用试剂盒可用于核酸的提取、分离或纯化。

[0081] 本文提供的分析核酸分子序列的方法可包括核酸分子的测序文库。可制备附接有本文提供的寡核苷酸的核酸分子的文库。在一些方面，制备了基因组文库。在一些方面，通过扩增制备附接有本文提供的包括条形码的寡核苷酸的核酸分子或其片段的文库。可通过聚合酶链反应(PCR)来扩增包括附接本文提供的包括条形码的寡核苷酸的核酸分子和核酸分子片段。可对包括附接本文提供的包括条形码的寡核苷酸的核酸分子和核酸分子片段的扩增子进行测序。可将任何合适的测序方法用于测序附接有本文提供的包括条形码的寡核苷酸的核酸分子和核酸分子片段。

[0082] 本文提供的分析样品中核酸分子序列的方法可进一步包括将序列读段分配到群组或类别。例如，可基于样品索引位置的定位和数量将序列读段分配到样品家族。因此，可将产生序列读段的核酸分子分配到该核酸分子来源的样品。在一些方面，样品索引位置的数量可用于错误校正。还可基于分子索引位置的定位和每个分子索引位置处的核苷酸将序列读段分配到分子家族。分子索引位置的数量和定位也可用于将序列读段分配到分子家族。因此，可将序列读段分配到产生序列读段的核酸分子。在一些方面，分子索引位置的数量可用于错误校正。又如，可基于细胞索引位置，例如定位、数量和每个细胞索引位置处的核苷酸及其组合，将序列读段分配到细胞家族。因此，可将序列读段和产生序列读段的核酸分子分配给起源的细胞。一方面，细胞索引位置的数量可用于错误校正。可根据本文提供的寡核苷酸和寡核苷酸组的条形码中包括的索引位置进行序列读段的任何分配。

[0083] 本文提供的分析样品中核酸分子序列的方法可进一步包括校正测序错误。错误来源可包括例如合成错误、测序假象或扩增步骤期间的聚合酶滑动。测序错误可通过将序列读段中的样品索引位置的数量和定位与样品索引位置的预先确定的数量和定位进行比较来校正。

[0084] 测序错误还可通过比较序列读段两端的样品条形码来校正。可应用规则比较序列读段每一端的不同样品条形码与允许的样品条形码。一方面，可应用规则比较测序读段两端的不同样品条形码，其中将包括相同样品条形码的寡核苷酸附接到核酸分子或其片段的每一端。在一些方面，可应用规则比较序列读段两端的不同样品条形码，其中将包括不同样品条码的寡核苷酸附接到核酸分子或其片段的每一端。在其他方面，本文提供的分析核酸

分子序列的方法包括使用不同的基因组并测试每种寡核苷酸以灵敏地检测读段错误分配。

[0085] 分析样品中核酸分子序列的方法可进一步包括应用一个或多个规则以(1)校正条形码内的错误,(2)校正核酸分子每一端的条形码之间的错误,(3)将序列读段多路分解到样品家族,(4)将序列读段分配到分子家族,或其任何组合。如本文所用,“多路分解(demultiplexing)”意指将序列读段分配到组或类别,如样品家族或样品来源(其中多个样品已被合并进行测序),例如,分子家族、细胞家族或任何其他期望的组或组的组合。

[0086] 本文提供的分析样品中核酸分子序列的方法中的寡核苷酸组中的每个寡核苷酸可进一步包括非条形码位置。寡核苷酸中包括的非条形码位置可包括用于杂交的位点、用于扩增的位点、用于序列引物结合的位点、以及用于杂交、序列引物结合和扩增的位点。用于杂交、序列引物结合的位点和用于扩增的位点可包括约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30或更多个核苷酸。用于杂交的位点可包括例如探针结合位点。用于扩增的位点可包括例如引物结合位点。用于杂交、序列引物结合的位点和用于扩增的位点可能彼此不同。用于杂交、序列引物结合的位点和用于扩增的位点也可能重叠。用于杂交、序列引物结合的位点和用于扩增的位点可在任何程度上重叠。在一些方面,用于杂交、序列引物结合的位点和用于扩增的位点重叠约1、约2、约3、约4、约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30或更多个核苷酸。在其他方面,用于杂交、序列引物结合的位点和用于扩增的位点完全重叠。一方面,用于杂交、序列引物结合的位点和用于扩增的位点不存在重叠。

[0087] 本文提供的分析核酸序列的方法可进一步包括存储核酸序列数据而无需多路分解。例如,多路分解密钥可用于例如将序列数据分配给测序读段的组。存储核酸序列数据而无需多路分解可保护序列数据。例如,存储核酸序列数据可防止没有正确多路分解密钥的个人使用序列数据,从而防止未经授权使用数据。

[0088] 用于标记核酸分子的方法

[0089] 在一个实施方案中,本发明提供了用于标记样品中核酸分子的方法,其包括:将包括条形码的多个寡核苷酸附接到核酸分子,每个条形码包括:(i)包括预先确定数量的样品索引位置的样品条形码,所述样品索引位置包含一个或多个特定核苷酸,其中样品索引位置的定位在样品之间变化;和(ii)包括分子索引位置的分子条形码,所述分子索引位置包含与样品索引位置处的核苷酸不同的核苷酸,其中样品索引位置和分子索引位置散布在一段连续碱基中。

[0090] 本文提供的任何寡核苷酸,包括寡核苷酸组和亚组,可用于在本文提供的标记核酸分子的方法中标记核酸分子或其片段。一方面,本文提供的方法包括将包含相同样品条形码的寡核苷酸附接到核酸分子的每一端。在一些方面,本文提供的方法包括将包含不同样品条形码的寡核苷酸附接到核酸分子的每一端。在其他方面,样品条形码位置的预先确定数量在不同样品条形码中变化。

[0091] 任何合适的方法都可用于将包含一个或多个条形码的寡核苷酸附接到核酸分子的末端。在一些方面,寡核苷酸是共价附接的。

[0092] 任何样品中的核酸均可使用本文提供的方法来标记。可被标记的核酸可存在于任

何样品或任何类型的样品中。在一些方面，样品是血液、唾液、血浆、血清、尿液或其他生物流体。其他示例性生物流体包括浆膜液、淋巴、脑脊液、粘膜分泌物、阴道液、腹水、胸膜液、心包液、腹膜液和腹腔液。在一些方面，样品是组织样品。在其他方面，样品是细胞样品。可使用新鲜样品或储存的样品，包括例如储存的冷冻样品、福尔马林固定石蜡包埋(FFPE)样品以及通过任何其他方法保存的样品。

[0093] 样品可来自正常或健康的受试者。样品也可来自患有疾病或病症的受试者。来自患有任何疾病或病症的受试者的样品中的核酸可使用本文提供的方法来标记。一方面，疾病或病症是癌症。在一些方面，样品是来自患有癌症的受试者的流体样品。在其他方面，样品是来自患有癌症的受试者的组织样品。在一些方面，样品是来自患有癌症的受试者的细胞样品。在其他方面，样品是癌症样品。癌症样品可以是来自实体肿瘤或液体肿瘤的样品。癌症可以是肾脏癌、肾癌、泌尿膀胱癌、前列腺癌、子宫癌、乳腺癌、宫颈癌、卵巢癌、肺癌、结肠癌、直肠癌、口腔癌、咽癌、胰腺癌、甲状腺癌、黑素瘤、皮肤癌、头颈癌、脑癌、造血系统癌、白血病、淋巴瘤、骨癌、肌肉癌、肉瘤、横纹肌肉瘤等。

[0094] 核酸可在样品中被标记。也可在标记前从样品中提取、分离或纯化核酸。可使用任何合适的提取、分离或纯化方法。示例性方法包括酚-氯仿提取、硫氰酸胍-酚-氯仿萃取、凝胶纯化以及柱和珠的使用。商用试剂盒可用于核酸的提取、分离或纯化。

[0095] 标记的核酸可用于制备例如核酸的文库。在一些方面，该文库是基因组文库。包括经标记的核酸分子的文库可通过将本文提供的寡核苷酸组或亚组通过末端修复、A-加尾和接头连接等附接到核酸分子或其片段来制备。在一些方面，省略末端修复和A-加尾，并包括与特定单个索引或索引组相关的可变末端来确定核酸分子如DNA分子的原始末端。标记的核酸分子及其片段以及标记的核酸分子及其片段的文库可通过例如测序来分析。任何合适的测序方法都可用于分析被标记的核酸分子。测序方法还可包括存储核酸序列数据而无需多路分解。例如，多路分解密钥可用于将序列数据分配给测序读段的组。存储核酸序列数据而无需多路分解可以保护序列数据。例如，存储核酸序列数据可防止没有正确多路分解密钥的个人使用序列数据，从而防止未经授权使用数据。

[0096] 本文提供的标记核酸分子的方法中的条形码可包含任意数量的核苷酸。例如，条形码可包含约10至约35个核苷酸。又如，条形码可包含约12至约25个核苷酸。再如，条形码可包含约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30、约31、约32、约33、约34、约35、约36、约37、约38、约39、约40或更多个核苷酸。又如，条形码可包含至少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、至少21、至少22、至少23、至少24、至少25、至少26、至少27、至少28、至少29、至少30、至少31、至少32、至少33、至少34、至少35、至少36、至少37、至少38、至少39、至少40或更多个核苷酸。

[0097] 本文提供的标记核酸分子的方法中的条形码可包括一个或多个索引位置。示例性索引位置包括样品索引位置、分子索引位置、DNA末端索引位置和细胞索引位置。例如，条形码可包括样品索引位置和分子索引位置。条形码还可包括样品索引位置、分子索引位置、细胞索引位置、DNA末端索引位置或其任意组合。

[0098] 本文提供的标记核酸分子的方法中的条形码可包括样品条形码。样品条形码可包

括预先确定数量的样品索引位置。预先确定的样品索引位置的数量可在样品之间变化。样品索引位置的定位也可在样品之间变化。在一些方面，预先确定的样品索引位置的数量和样品索引位置的定位可在样品之间变化。因此，核酸分子和核酸分子产生的序列读段的样品来源可通过形成样品条形码的样品索引位置的数量、样品索引位置的定位、或样品索引位置的数量和定位二者来识别。

[0099] 本文提供的标记核酸分子的方法中的样品条形码中的预先确定数量的样品索引位置可包含一个或多个特定核苷酸。例如，预先确定数量的样品索引位置中的一个或多个特定核苷酸可以是A、T、G或C。再如，预先确定数量的样品索引位置中的一个或多个特定核苷酸可以是A和T、A和C、A和G、T和C、T和G、或G和C。

[0100] 在一些方面，本文提供的标记核酸分子的方法中的样品条形码包括2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20或更多个样品索引位置或其组合。在其他方面，本文提供的标记核酸分子的方法中的样品条形码包括约4至约12个样品索引位置。在一些方面，本文提供的标记核酸分子的方法中的样品条形码包括约2、约3、约4、约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20或更多个样品索引位置或其组合。在其他方面，本文提供的标记核酸分子的方法中的样品条形码包括至少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或更多个样品索引位置或其组合。

[0101] 本文提供的标记核酸分子的方法中的条形码可包括分子条形码。分子条形码可包括分子索引位置，其包含与样品索引位置处的核苷酸不同的核苷酸。例如，样品索引位置核苷酸和分子索引位置核苷酸可选自：(A)样品索引位置核苷酸为A，并且分子索引位置核苷酸为C、G、T或其组合；(B)样品索引位置核苷酸为T，并且分子索引位置核苷酸为C、G、A或其组合；(C)样品索引位置核苷酸为C，并且分子索引位置核苷酸为G、A、T或其组合；(D)样品索引位置核苷酸为G，并且分子索引位置核苷酸为C、A、T或其组合；(E)样品索引位置核苷酸为A、T或其组合，并且分子索引位置核苷酸为C、G或其组合；(F)样品索引位置核苷酸为A、C或其组合，并且分子索引位置核苷酸为T、G或其组合；(G)样品索引位置核苷酸为A、G或其组合，并且分子索引位置核苷酸为T、C或其组合；(H)样品索引位置核苷酸为T、C或其组合，并且分子索引位置核苷酸为A、G或其组合；(I)样品索引位置核苷酸为T、G或其组合，并且分子索引位置核苷酸为A、C或其组合；或(J)样品索引位置核苷酸为G、C或其组合，并且分子索引位置核苷酸为A、T或其组合。

[0102] 本文提供的标记核酸分子的方法中的样品条形码的样品索引位置可与分子索引位置穿插。因此，本文提供的标记核酸分子的方法中的条形码可包括不需要限制于特定的连续核苷酸段或块的样品索引位置和分子索引位置。例如，并非所有的样品索引位置都需要彼此相邻，也并非所有的分子索引位置都需要彼此相邻。样品索引位置和分子索引位置可以交替。任意数量的分子索引位置可位于样品索引位置之间。任意数量的分子索引位置可位于任意数量的样品索引位置之间。任意数量的分子索引位置和任意数量的非分子索引或其他索引位置的核苷酸可位于样品索引位置之间。任意数量的分子索引位置和任意数量的非分子索引或其他索引位置的核苷酸可位于任意数量的样品索引位置之间。任意数量的非样品索引位置或非分子索引位置的核苷酸可位于样品索引位置和分子索引位置之间。

[0103] 一些样品索引位置可以彼此相邻，而其他样品索引位置可定位于邻近非样品索引

位置的条形码中的任何其他核苷酸。例如，样品索引位置和分子索引位置可以是不需要所有样品索引位置彼此相邻的任何配置。例如，样品索引位置和分子索引位置可以是不需要所有分子索引位置彼此相邻的任何配置。例如，样品索引位置和分子索引位置也可以是不需要所有样品索引位置和所有分子索引位置彼此相邻的任何配置。任何索引条形码的位置可以是不需要索引条形码的所有核苷酸彼此相邻的任何配置。示例性条形码索引包括样品条形码、分子条形码、细胞条形码、DNA末端索引位置等。

[0104] 本文提供的标记核酸分子的方法中的分子条形码可包括约5至约25个分子索引位置。在一些方面，本文提供的标记核酸分子的方法中的分子条形码包括约5至约15个分子索引位置。在其他方面，本文提供的标记核酸分子的方法中的分子条形码包括约2、约3、约4、约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30、或更多个分子索引位置。在多个方面，本文提供的标记核酸分子的方法中的分子条形码包括至少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、至少21、至少22、至少23、至少24、至少25、至少26、至少27、至少28、至少29、至少30或更多个分子索引位置。

[0105] 本文提供的标记核酸分子的方法中的条形码可包括含有索引位置的一个或多个附加索引条形码。在一些方面，一个或多个附加索引条形码是细胞条形码。在其他方面，一个或多个附加索引条形码是提供未修复DNA末端长度的量度的条形码。因此，本文提供的标记核酸分子的方法中的条形码可包括样品条形码、分子条形码、细胞条形码、提供未修复DNA末端长度的量度的条形码、任何其他索引条形码或其任何组合。因此，本文提供的标记核酸分子的方法中的条形码可包括例如样品索引位置、分子索引位置和任何其他索引位置如细胞索引位置，这些位置彼此穿插。本文提供的标记核酸分子的方法中的条形码的索引位置不需要限制于特定的连续核苷酸段或块。索引条形码和索引位置可以是不需要所有索引位置彼此相邻的任何配置。

[0106] 本文提供的标记样品中核酸分子的方法中的寡核苷酸组中的每个寡核苷酸可进一步包括非条形码位置。包括于寡核苷酸中的非条形码位置可包括用于杂交的位点、用于扩增的位点、用于序列引物结合的位点、以及用于杂交、序列底物结合和扩增的位点。用于杂交、序列引物结合的位点和用于扩增的位点可包括约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30或更多个核苷酸。用于杂交的位点可包括例如探针结合位点。用于扩增的位点可包括例如引物结合位点。用于杂交、序列引物结合的位点和用于扩增的位点可以彼此不同。用于杂交、序列引物结合的位点和用于扩增的位点也可以重叠。用于杂交、序列引物结合的位点和用于扩增的位点可以在任何程度上重叠。在一些方面，用于杂交、序列引物结合的位点和用于扩增的位点重叠约1、约2、约3、约4、约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约26、约27、约28、约29、约30、或更多个核苷酸。在一些方面，用于杂交、序列引物结合的位点和用于扩增的位点完全重叠。在其他方面，用于杂交、序列引物结合的位点和用于扩增的位点没有重叠。

[0107] 用于识别错误序列读段的方法

[0108] 在一个实施方案中，本发明提供了用于识别错误序列读段的方法，其包括：(a)将多个寡核苷酸附接到样品的核酸分子，其中每个寡核苷酸包括条形码，所述条形码包括：(i)包括预先确定数量的样品索引位置的样品条形码，所述样品索引位置包含一个或多个特定核苷酸，其中样品索引位置的定位在样品之间变化，并且其中将相同的样品条形码附接到该样品中的核酸分子的每一端；和(ii)包括分子索引位置的分子条形码，所述分子索引位置包含与样品索引位置处的核苷酸不同的核苷酸，其中样品索引位置和分子索引位置散布在一段连续碱基中；以及(b)对核酸分子测序，其中序列读段包括条形码序列，从而识别错误序列读段。

[0109] 如本文所用，术语“错误序列读段”意指可通过本文所述方法识别的任何测序错误。

[0110] 一方面，识别错误序列读段包括识别带有不相符的样品条形码的核酸分子。

[0111] 本文所述方法依赖于将相同的样品条形码附接到核酸分子的每一端。术语“不相符的样品条形码”是指由于制备用于测序的核酸过程中发生错误，核酸分子附接的条形码在该核酸分子每一端不同的情况。这可能导致分子家族中的错误分配，从而可干扰序列读段的正确分析。

[0112] 在一些方面，通过比较序列读段两端的样品条形码进一步校正测序错误。在其他方面，带有不相符样品条形码的核酸分子进一步从序列读段和/或从分子家族中去除。

[0113] 另一方面，识别带有不相符样品条形码的核酸分子包括识别错误引发的核酸分子。

[0114] 如本文所用，“错误引发的核酸分子”可指包含多对分子条形码的核酸分子。在这种情况下，分子数量可能被错误地夸大，和/或错误的样品可能被分配给不正确的分子读段，这可能会对读段变体的频率和/或身份产生负面影响。这两种情况都导致结果的分析和临床解释出现问题。

[0115] 在一些方面，错误引发的核酸分子用适当的条形码校正，并用于提高序列质量。在其他方面，具有校正条形码的核酸分子被分配到校正的读段家族。

[0116] 在多个方面，将校正的读段家族用于准确测定不同的覆盖度。在一些方面，将不同的覆盖度测定用于评估核酸分子的文库。

[0117] 一方面，该方法还包括基于分子索引位置的定位和每个分子索引位置处的核苷酸将序列读段分配到分子家族。在一些方面，识别错误的序列读段包括识别被分配到多个分子家族的核酸分子。在其他方面，被分配到多个分子家族的核酸分子进一步从序列读段和/或从分子家族中去除。

[0118] 如本文使用的，单数形式“一个”、“一种”和“所述/该”包括复数引用，除非上下文另有明确规定。因此，例如，提及“所述/该方法”包括本文所述类型的一个(种)或多个(种)方法和/或步骤，这对于本领域技术人员在阅读本公开等之后将变得显而易见。

[0119] 除非另有定义，本文使用的所有技术和科学术语具有与本发明所属领域的技术人员通常理解的含义相同的含义。

[0120] 当提及可测量的值如数量、时间持续时间等时，本文使用的“约”意指包含从指定值变化±20％或±10％、或±5％、或甚至±1％的变量，因为这类变量适用于所公开的组合物或实施所公开的方法。

[0121] 如本文所用，术语“核酸”是指任何脱氧核糖核酸(DNA)分子、核糖核酸分子或核酸类似物。DNA或RNA分子可以是双链或单链，并且可以是任何大小。示例性核酸包括但不限于染色体DNA、质粒DNA、cDNA、无细胞DNA(cfDNA)、循环肿瘤DNA(ctDNA)、mRNA、tRNA、rRNA、siRNA、微RNA(miRNA或miR)、hnRNA。示例性核酸类似物包括肽核酸、吗啉代核酸和锁核酸、二醇核酸和苏糖核酸。如本文所用，术语“核酸分子”意指包括例如核酸分子的片段以及任何全长或非片段化核酸分子。

[0122] 如本文所用，术语“核苷酸”包括核糖核酸和脱氧核糖核酸的单个单位及核苷和核苷酸类似物，以及修饰的核苷酸如标记的核苷酸。此外，“核苷酸”包括非天然存在的类似物结构，如其中糖、磷酸根和/或碱基单位缺失或被其他化学结构取代的那些结构。因此，术语“核苷酸”包括单个肽核酸(PNA)(Nielsen等人，Bioconjug.Chem.1994;5(1):3-7)和锁核酸(LNA)(Braasch and Corey，Chem.Biol.2001;8(1):1-7)单位以及其他类似单位。

[0123] 如本文所用，术语“受试者”是指对其实施本文所公开方法的任何个体或患者。术语“受试者”可与术语“个体”或“患者”互换使用。受试者可以是人类，但是受试者可以是动物，如本领域技术人员所理解的。因此，其他动物，包括哺乳动物，如啮齿动物(包括小鼠、大鼠、仓鼠和豚鼠)、猫、狗、兔，农场动物(包括牛、马、山羊、绵羊、猪等)，以及灵长类动物(包括猴子、黑猩猩、猩猩和大猩猩)都包括在受试者的定义中。受试者还可以是植物或微生物。

[0124] 如本文所用，术语“处理”、“治疗”、“疗法”和“治疗的”等是指获得期望的药理学和/或生理学效果，包括但不限于缓解、延缓或减缓进展，减轻影响或症状，预防发病，抑制、改善疾病或病症的发病，获得关于疾病、病症或医疗状况的有益或期望结果，如治疗益处和/或预防益处。本文所用的“治疗”涵盖对哺乳动物(尤其是人类)疾病的任何治疗，并且包括：(a)防止疾病在可能易患疾病或有获得疾病风险但尚未被诊断为患有疾病的受试者中发生；(b)抑制疾病，即阻止其发展；以及(c)缓解疾病，即导致疾病消退。治疗益处包括根除或改善正在治疗的潜在病症。此外，治疗益处随着根除或改善与潜在病症相关的一种或多种生理症状而获得，这样观察到受试者的改善，尽管受试者可能仍然患有该潜在的病症。在某些情况下，为了预防益处，对处于患特定疾病风险的受试者或对报告一种或多种疾病生理症状的受试者进行治疗，即使可能尚未对该疾病进行诊断。本公开的方法可用于任何哺乳动物或其他动物。在某些情况下，治疗可导致症状的减少或停止。预防作用包括延迟或消除疾病或病况的出现，延迟或消除疾病或病况的症状的发作，减缓、停止或逆转疾病或病况的进展，或其任何组合。

实施例

[0125] 实施例1

[0126] 本实施例描述了用于多索引样品的不固定/数字化条形码的设计。

[0127] 不固定或数字化条形码的给定位置处的核苷酸存在或缺失提供了信息内容，类似于消费品条形码(UPC)(图1)。对于不同的索引，核苷酸或“条”移动或浮动到不同位置，并且那些新的位置表示备选的索引。随着可用序列位置增加，可能的条形码的数量迅速增加。未用作主索引的位置可用作次级索引或附加索引。还可以包括在诸如单细胞测序的方法中有用的额外索引水平。例如，对于单细胞测序，可使得样品索引、细胞索引和分子索引均在单个条形码中。根据创建条形码的条件选择，可供使用不同数量的主条形码和次级条形码，并

且可根据需要调整错误检测和错误校正的强度。

[0128] 样品中不同分子的数量通常非常高,每个样品有数百万或更多的分子被测序。由于分子数量如此之多,通常不可能为每种分子条形码合成和纯化单个寡核苷酸。在多个位置的简并核苷酸通常用于提供区分不同分子所需的多样性。通常,将确定的样品条形码和随机化的分子条形码相互分离进行分析。利用不固定/数字化条形码系统,在一个区域内混合多种类型的条形码。

[0129] 与标准的固定长度条形码相比,这代表了根本不同的样品索引方法,该方法使用基于位置的方法,其中序列不直接与参考进行比较。样品条形码的定位随样品而变化,并且该定位用于识别样品家族。利用标准条形码时,将序列相互比较,并将完全或接近完全的序列同一性分组为一个样品家族。对于不固定/数字化条形码,序列不直接相互比较,而是用于以数字+/-方式标记定位。+/-定位数据随后用于区分样品,与传统产品条形码相似(图1)。在图1所示的示例中,具有核苷酸“A”的任何位置都是样品条形码的一部分,而任何其他核苷酸都是分子条形码的一部分。无论何时对“A”进行测序,都会记录其位置并用于确定样品家族。

[0130] 新型条形码基于多种要求而设计,包括以下,例如:(1)应有足够的独特条形码,以容纳任何运行中的样品和分子数量;(2)每个分子读段不同端的组合样品/分子条形码应当不同,但样品条形码是可预测的,以检测高容量测序仪的索引跳跃;(3)条形码在碱基组成中不应包含影响序列质量的大量多核苷酸重复或极端值;(4)分子索引应是高度可变的,以区分所有可能的分子;以及(5)样品条形码设计应与寡核苷酸合成的可行数量兼容。

[0131] 新设计的不固定或数字化条形码符合上述标准。这种新的条形码设计能够在相对较短的序列中并入所有这些特征,其已与例如NextSeq和NovaSeq Illumina测序仪兼容。相同或类似的设计可与其他测序系统兼容。

[0132] 新的不固定/数字化条形码在邻近位置使样品和分子条形码混合,并使用位置信息而不是直接序列比较来分配样品家族。将任何给定位置的核苷酸序列用于确定该位置是否应当被指定为样品或分子位置。该位置信息随后用于确定条形码和分配样品家族。如果样品条形码定位的数量与预期数量或位置不匹配,则可丢弃分子或尝试校正条形码。这些条形码的设计允许灵活分配条形码和类别,从而可用于多种应用,包括测序运行中的多路复用样品或需要将读段分配到特定样品和细胞的单细胞方法。

[0133] 许多条形码配置都是可能的。作为许多可能性的一个例子,样品索引可以始终是核苷酸“A”,而分子索引可以是任何其他核苷酸(C、G、T)。使用IUPAC命名法,C、G或T由符号“B”表示,并且A、C或G由符号“V”表示。可以这种方式潜在使用的序列示例如图2A至图2C所示。

[0134] 对于给定的位置数量(n),具有的可能条形码的数量可由下式计算:

[0135] $Cr=n!/r!(n-r)!$

[0136] 其中n是可能的位置数量,r是待填充的位置数量。不同序列大小可能的最大数量示于表1。

[0137] 表1可能的条形码组合

[0138]

混合序列的长度	不同条形码的最大数量
4	6

6	20
8	70
10	252
12	924
14	3432
16	12,870
18	48,620
20	184,756
n	Cr＝n!/r!(n-r)!

[0139] 在每个位置，二项选择确定该位置是用作分子索引还是样品索引位置。如果序列与样品索引序列(如，A)匹配，则它是样品条形码的一部分。如果不匹配(如C、G或T)，则它是简并分子索引的一部分。在图2C所示的例子中，在每20nt段内，多达7个位置被分配给样品索引位置，并且13或更多个位置被三倍简并，使得每个样品条形码20nt段3^13或1,594,323倍简并。因为每个分子都有两个这样的条形码，任何单个分子都可以是1,594,323^2或2.5万亿倍简并。

[0140] 如图3A所示，许多类型的标准接头具有位于不同接头寡核苷酸上的简并分子条形码和固定样品条形码(参见SEQ ID NO:1和2)。如图3B所示，两者混合的不固定条形码的情况并非如此(参见SEQ ID NO:5和6)。

[0141] 错误校正以及样品和分子条形码的模式可采取多种形式。在某些情况下，例如体细胞变体的测序，重要的是读段不被错误分配。因此，具有稳健的错误检测和校正是重要的。例如，如果存在固定数量的样品条形码位置，则匹配该数量提供一种类型的质量检查。如果条形码不是所选长度，则该特定分子中一定存在测序错误。可以基于预期的条形码来校正错误，或者可能需要从总体结果中消除序列以避免错误分配。可选择地，可以使用可变数量的样品条形码位置，但以可基于允许的模式检测并固定任何单个测序错误的方式生成它们。在这种情况下，每个样品条形码与所有其他样品条形码的差异为至少两个或至少三个或更多个变化。在其他情况下，偶尔的错误分配可能不是重要问题，更重要的是提供最大数量的条形码。这将防止某些类型的错误检测/校正，但仍允许比较同一分子的两端的条形码。

[0142] 除了代表样品条形码的单核苷酸外，其他变异也是可能的。例如，样品(或细胞)条形码可以用固定的A或T表示，并且分子条形码可以用简并的G/C表示。这种配置生成多很多的样品/细胞条形码和更少的分子条形码。改变样品/分子条形码位置的数量和简并性，允许优化当前应用中二者的数量。

[0143] 不固定或数字化条形码系统允许在相同核酸分子的两端放置相同的样品条形码。利用传统的DNA条形码，在相同分子的两端不能使用相同的样品条形码。如果将相同的标准样品条形码放置在同一分子的两端，不同的分子可能会交叉杂交，从而在扩增过程中产生人工嵌合分子的风险很高。在分子两端有相同的条形码序列的情况下，这两个最3'的区域可以杂交并产生部分重复的分子。由于标准样品条形码可能在被扩增的样品中出现数百万次，嵌合分子形成的可能性很高(参见图4和SEQ ID NO:7和8)。不固定条形码的情况并非如此，因为即使使用相同的样品条形码，也不会有长的连续相同碱基段。由于不固定接头的样

品条形码只有很短的同源性区域，几乎没有非特异性相互作用和嵌合体形成的风险。因此，可将相同的样品条形码放置在相同分子的两端，以便比较两个条形码中另一个的错误。如果没有发现错误，则可自信地分配样品。如果这两个条形码不相同，可将它们与允许的条形码列表进行比较，并相应地校正。每个索引使用的条形码数量决定了错误可被校正的程度。

[0144] 因此，在相同分子的两端放置相同的样品条形码且嵌合体形成的风险低，这种能力提供了简单但强大的错误校正潜力。人们简单地比较分子每一端的样品条形码来验证同一性。如果相同，可将分子放置在合适的样品家族中。如果它们不匹配，可将两者与允许的样品条形码组进行比较，并可能校正错误的条形码。这种方法提供了确保将分子分配到合适的样品家族且读段丢失最小的强大方法。样品条形码校正的例子如表2所示。条形码之间的编辑距离将决定如何校正条形码，且在编辑距离较大时校正条形码并保持读段的能力更强。

[0145] 相同分子不同末端的样品条形码缺乏一致性为样品制备中有问题的过程提供了证据。通过监测嵌合分子的频率，如不匹配的样品条形码所证明的，可对文库制备和测序方法进行改进。

[0146] 如果特定的分子条形码与多个不同的分子条形码匹配，并且不匹配的数量表明它不是由简单的测序错误引起的，则表明一个或多个分子读段不匹配。可将分子对的相对频率用于确定哪一种是优势种类，可以原样使用，哪一中可能是假象，需要校正或去除。一对样品中i5和i7接头如何分布的分解参见表3。正确和可校正的条形码可以简单直接的方式使用，而如果要挽救读段，错误引发的分子需要更复杂的分析。在不知道哪些读段是错误引发的情况下，错误信息可能会被纳入分析。知道哪里发生错误引发允许正确处理序列读段。只有当错误引发处于足够低的水平能够可靠地检测到时，才能校正错误引发。

[0147] 如图6所示，连接步骤中过多的接头可能会导致重大问题，这是当通过PCR引物(如，SEQ ID NO:3和4)延伸剩余的接头并随后在扩增后期使用时。在0.2μM及更低时，错误引发水平相对较低，而在0.5μM及更高时，错误引发实质性增加。

[0148] 表2.编辑距离＝2的来自相同分子读段的样品条形码的校正

[0149]

i7距离	i5距离	模式匹配?	片段分配
0	0	是	i7/i5
1	0	是	i7/i5
1	1	否	无
n/a	0	n/a	i5
n/a	1	n/a	无
0	1	否	无
n/a	n/a	n/a	无

[0150] 表3.对一对样品的i5和i7接头的分布

[0151]

	编辑距离	状态	样品1	样品2
i7	0	正确	13.3%	10.5%

[0152]

	1,2,3	可校正的	1.5%	0.8%
	>3	错误引发的错误	85.2%	88.7%
i5	0	正确	90.5%	87.8%
	1,2,3	可校正的	0.5%	1.0%
	>3	错误引发的错误	9.0%	11.2%

[0153] 总之,设计新的不固定或数字化条形码的方法的根本区别是使用核苷酸位置而不是特定的核苷酸序列作为条形码。这个主题有多种可能的变化,允许在条形码数量和错误校正方法方面具有灵活性。新条形码的一些益处包括:(1)改进NGS读段分配到样品和分子家族;(2)减少复杂样品的寡聚体合成/纯化次数;(3)减少简并区中有问题的均聚物和富含GC段的数量。

[0154] 实施例2

[0155] 本实施例描述了使用样品测试不固定条形码。

[0156] 为了测试不固定条形码,设计了以最大灵敏度检测读段不匹配的实验。使用标准的文库制备方案。未观察到标准条形码和不固定条形码之间产率的显著差异。

[0157] 为了检测错误分配,制备了三份样品,并用标准条形码和不固定条形码并行测序。使用不同的条形码制备每份样品。这三种样品是使用用于人类DNA的靶向面板捕获的人类DNA,和来自大肠杆菌(E.coli)和拟南芥(Arabidopsis thaliana)的基因组DNA,这些DNA已被修剪但未被选择性捕获。所有六个样品在相同的NextSeq测序运行设置上运行进行20nt索引测序。然后,将所得到的读段多路分解两次,一次使用标准条形码,一次使用不固定条形码。然后分别分析这些读段,以确定读段与哪个基因组比对。对于人类比对序列,初始算法与标准比对一样好或更好,其中与被分配到大肠杆菌和拟南芥的条形码比对的读段不到0.002%,如图5所示。较低的脱靶读段映射导致较低的读段分配的错误率。

[0158] 这些数据表明,与标准条形码相比,不固定或数字化条形码表现良好。实验室方案的优化,包括例如改变阻断剂,以及软件/算法,包括例如用于多路分解的软件、错误校正和创建读段家族,将进一步改进使用不固定或数字化条形码进行序列分析所获得的结果。此外,不固定或数字化条形码可用于其中多种索引有用的多种应用,例如在单细胞分析和系统中标记细胞,其中例如将一种、两种、三种或更多种索引用于标记分子、细胞和/或样品性质并分组到相应类别。

[0159] 总之,新的不固定或数字化条形码系统为分析提供了多种优势,例如灵活性、寡聚体合成成本更低以及易于错误校正的方法,这些都预料不到且令人惊讶地改进了当前错误校正的方法,从而更好地例如将读段分配到正确样品和分子家族。

[0160] 实施例3

[0161] 本实施例描述了如何使用不固定条形码来识别和从样品中去除被错误分配的分子读段。

[0162] 由于样品条形码在每个分子的两端编码,可对条形码进行比较以用于错误校正和确认由多个样品产生非期望的嵌合分子未在较大程度上发生。如图6所示,即使使用标准条件,嵌合分子的形成也可能是重大问题。该问题可表现为以下形式,即相同分子得到多对分

子条形码并人为地夸大分子数量或错误样品被分配给分子读段，导致变体的错误频率或身份。这两种情况都导致可能影响临床结果解释的分析问题。

[0163] 文库制备中扩增引物的绝对浓度和相对浓度导致条形码的效率和准确性发生变化。接头的初始浓度越高，连接的效率越高，可以回收的样品比例越大。不幸地，过量接头可能导致扩增问题，其中接头被扩增或用作引物时添加的条形码在扩增过程而不仅是连接阶段被添加(图7)。如果在扩增过程中添加新的样品条形码，读段将被分配到错误样品，并且变体的频率或存在变得不太准确。如果在扩增过程中添加新的分子条形码，则每个分子都有多对条形码，以致分子多样性将被高估，并且那些读段的错误校正变得更难或更不可能。使用标准条形码，甚至无法测量这些问题的程度。使用不固定条形码，很容易检测到这类问题，然后可以改进方法以优化准确性。

[0164] 实施例4

[0165] 分子条形码是随机的，但由于其散布在样品条形码中，它不包含可能导致问题的长段完全随机的碱基。完全随机的条形码可能是100％ GC，而20nt总序列必须包括样品条形码，其可以是全部A或全部T，因此设置GC含量的上限，通常为65％。这也防止了长的均聚物。已表明完全随机条形码具有以数百个拷贝出现的某些序列，而大多数序列仅出现几次。[Kinde I,Wu J,Papadopoulos N,Kinzler KW,Vogelstein B.Detection and quantification of rare mutations with massively parallel sequencing.Proc Natl Acad Sci U S A.2011 Jun 7;108(23):9530-5.doi:10.1073/pnas.1105422108.Epub 2011 May 17.PMID:21586637;PMCID:PMC3111315.]这些分子条形码的更均匀含量如图8所示，其中很少有条形码的比例明显过高。

[0166] 尽管已参考上述实施例描述了本发明，但应理解，修改和变化包括在本发明的精神和范围内。因此，本发明仅受以下权利要求的限制。

序列表

<110> PERSONAL GENOME DIAGNOSTICS INC.

THOMPSON, JOHN F.

<120> 不固定条形码

<130> PGDX3120-1WO

<140>

<141>

<150> 63/006,556

<151> 2020-04-07

<160> 8

<170> PatentIn version 3.5

<210> 1

<211> 72

<212> DNA

<213> 人工序列

<220>

<223> 人工序列描述: 合成

寡核苷酸

<220>

<221> modified_base

<222> (30)..(39)

<223> a, c, t, g, 未知或其他

<400> 1

aatgatacgg cgaccaccga gatctacacn nnnnnnnnna cactctttcc ctacacgacg 60

ctcttccgat ct 72

<210> 2

<211> 65

<212> DNA

<213> 人工序列

<220>

<223> 人工序列描述: 合成

寡核苷酸

<400> 2

gatcggaaga gcacacgtct gaactccagt cactggagta gatctcgtat gccgtcttct 60

gcttg 65

<210> 3

<211> 28

<212> DNA

<213> 人工序列
<220>
<223> 人工序列描述: 合成
寡核苷酸
<400> 3
aatgatacgg cgaccaccga gatctaca 28
<210> 4
<211> 24
<212> DNA
<213> 人工序列
<220>
<223> 人工序列描述: 合成
寡核苷酸
<400> 4
caagcagaag acggcatacg agat 24
<210> 5
<211> 82
<212> DNA
<213> 人工序列
<220>
<223> 人工序列描述: 合成
寡核苷酸
<400> 5
aatgatacgg cgaccaccga gatctacacb bbbbabbabb abababaaba cactctttcc 60
ctacacgacg ctcttccgat ct 82
<210> 6
<211> 77
<212> DNA
<213> 人工序列
<220>
<223> 人工序列描述: 合成
寡核苷酸
<400> 6
gatcggaaga gcacacgtct gaactccagt cacbbbbbab babbababab aabatctcgt 60
atgccgtctt ctgcttg 77
<210> 7
<211> 168
<212> DNA
<213> 人工序列

<220>
<223> 人工序列描述: 合成
多核苷酸
<220>
<221> modified_base
<222> (83)..(90)
<223> a, c, t, g, 未知或其他
<400> 7
aatgatacgg cgaccaccga gatctacaca aaaaaaaaaa aaaaaaaaaa cactctttcc 60
ctacacgacg ctcttccgat ctnnnnnnnn agatcggaag agcacacgtc tgaactccag 120
tcacaaaaaa aaaaaaaaaa aaaaatctcg tatgccgtct tctgcttg 168
<210> 8
<211> 169
<212> DNA
<213> 人工序列
<220>
<223> 人工序列描述: 合成
多核苷酸
<220>
<221> modified_base
<222> (79)..(87)
<223> a, c, t, g, 未知或其他
<400> 8
caagcagaag acggcatacg agatttttt ttttttttt ttttgtgact ggagttcaga 60
cgtgtgctct tccgatctnn nnnnnnnaga tcggaagagc gtcgtgtagg gaaagagtgt 120
ttttttttt ttttttttt gtgtagatct cggtggtcgc cgtatcatt 169

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
A A A A A A A A A A A A A A A A A A A A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
A A A A A A A A A A A A A A A A A A A A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
A A A A A A A A A A A A A A A A A A A A

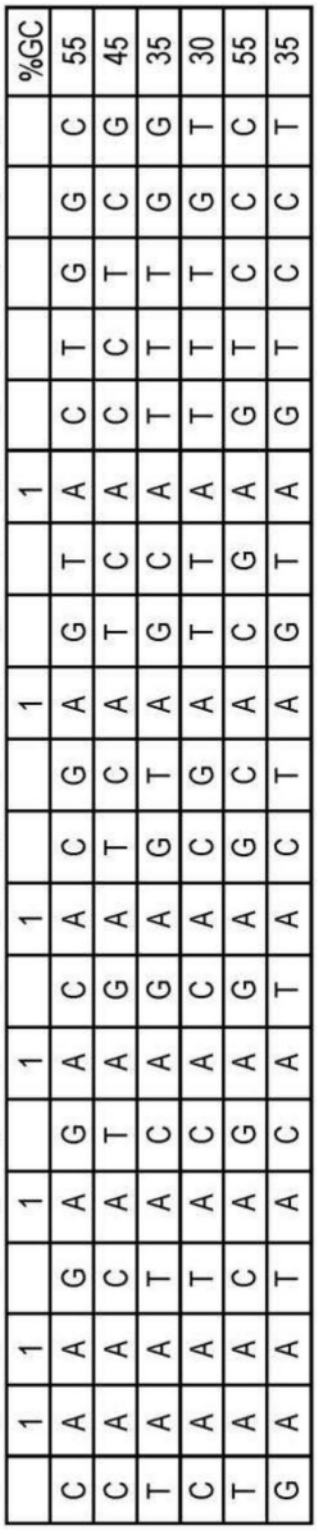
图1

#	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

图2A

#	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
1	A	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A
2	A	A	B	A	B	A	B	A	B	A	B	A	B	A	B	B
3	A	A	B	A	B	A	B	A	B	A	B	A	B	A	B	B
4	A	A	B	B	A	A	B	B	A	A	B	B	A	A	B	A
5	A	A	B	B	A	B	A	A	B	B	A	A	B	B	A	A
6	A	A	B	B	A	B	A	B	A	B	A	B	A	B	A	A
7	A	A	B	B	A	B	A	B	A	B	A	B	A	B	A	B
8	A	B	A	A	B	B	A	A	B	B	A	A	B	B	A	A
9	A	B	A	A	B	A	B	B	A	A	B	B	A	A	B	B
10	A	B	A	A	B	B	A	B	A	B	A	B	A	B	A	B
11	A	B	A	A	B	B	A	B	A	B	A	B	A	B	A	A
12	A	B	A	B	A	A	B	B	A	A	B	B	A	B	A	A
13	A	B	A	B	A	B	A	A	B	B	A	B	A	B	B	A
14	A	B	A	B	A	B	A	B	A	A	B	A	B	A	B	A
15	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	A
16	A	B	B	A	B	A	A	B	B	A	A	B	B	A	A	B

图2B



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	%GC
C	A	A	G	A	G	A	C	A	C	G	A	G	T	A	C	T	G	G	C	55
C	A	A	C	A	T	A	G	A	T	C	A	T	C	A	C	C	T	C	G	45
T	A	A	T	A	C	A	G	A	G	T	A	G	C	A	T	T	T	G	G	35
C	A	A	T	A	C	A	C	A	C	G	A	T	T	A	T	T	T	G	T	30
T	A	A	C	A	G	A	G	A	G	C	A	C	G	A	G	T	C	C	C	55
G	A	A	T	A	C	A	T	A	C	T	A	G	T	A	G	T	C	C	T	35

图2C

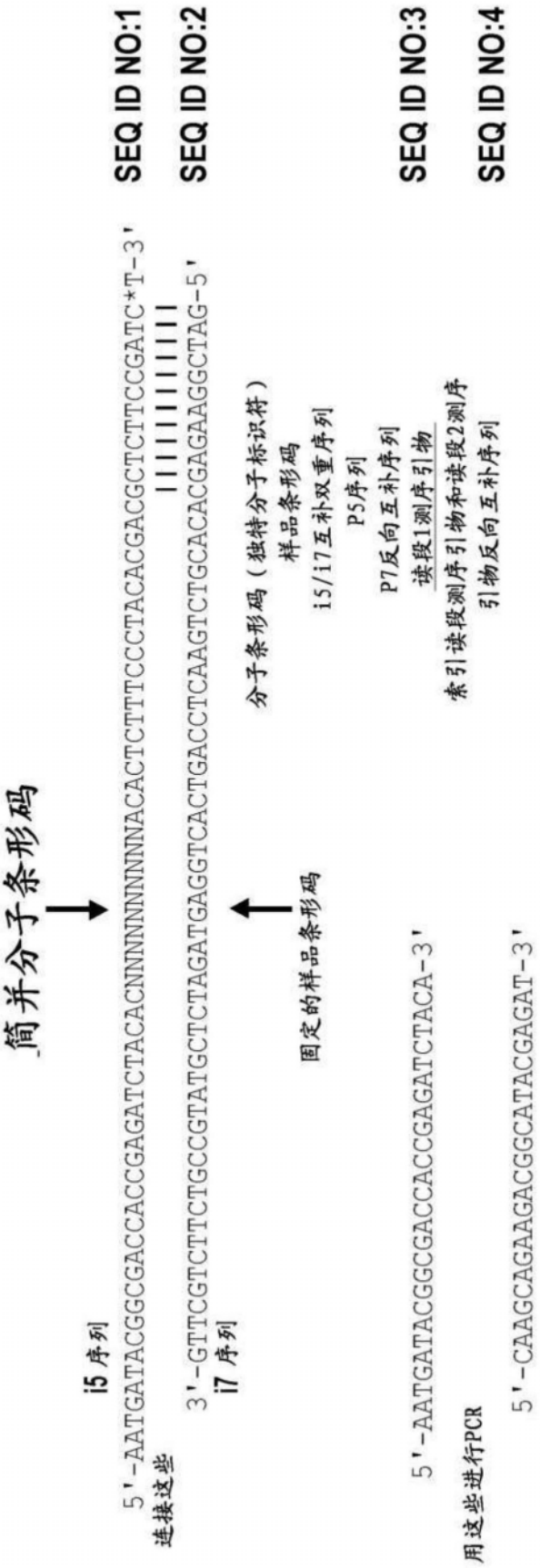

简并分子条形码
i5 序列
5'-AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'
连接这些
3'-GTTCGTCTTCTGCCGTATGCTCTAGATGAGGTCACTGACCTCAAGTCTGCACACGAGAAGGCTAG-5'
i7 序列
分子条形码（独特分子标识符）
样品条形码
固定的样品条形码
i5/i7互补双重序列
P5序列
P7反向互补序列
读段1测序引物
索引读段测序引物和读段2测序
引物反向互补序列
5'-AATGATACGGCGACCACCGAGATCTACA-3'
用这些进行PCR
5'-CAAGCAGAAGACGGCATACGAGAT-3'
SEQ ID NO:1
SEQ ID NO:2
SEQ ID NO:3
SEQ ID NO:4

图3A

固定的样品/简并分子条形码
连接这些
5'- i5 序列
AATGATACGGCGACCACCGAGATCTACACBBBBABBBABBBABABABBABA
SEQ ID NO:5
3'- GTTCGTCTTCTCTGCCGTATGCTCTABAABABABABBABBBABBBBBB
SEQ ID NO:6
i7 序列
固定的样品/简并分子条形码
组合的样品/分子条形码
i5/i7互补双重序列
P5序列
P7反向互补序列
读段1测序引物
索引读段测序引物和读段2测序
引物反向互补序列
5'-AATGATACGGCGACCACCGAGATCTACA-3'
SEQ ID NO:3
用这些进行PCR
5'-CAAGCAGAAGACGGCATACGAGAT-3'
SEQ ID NO:4

图3B

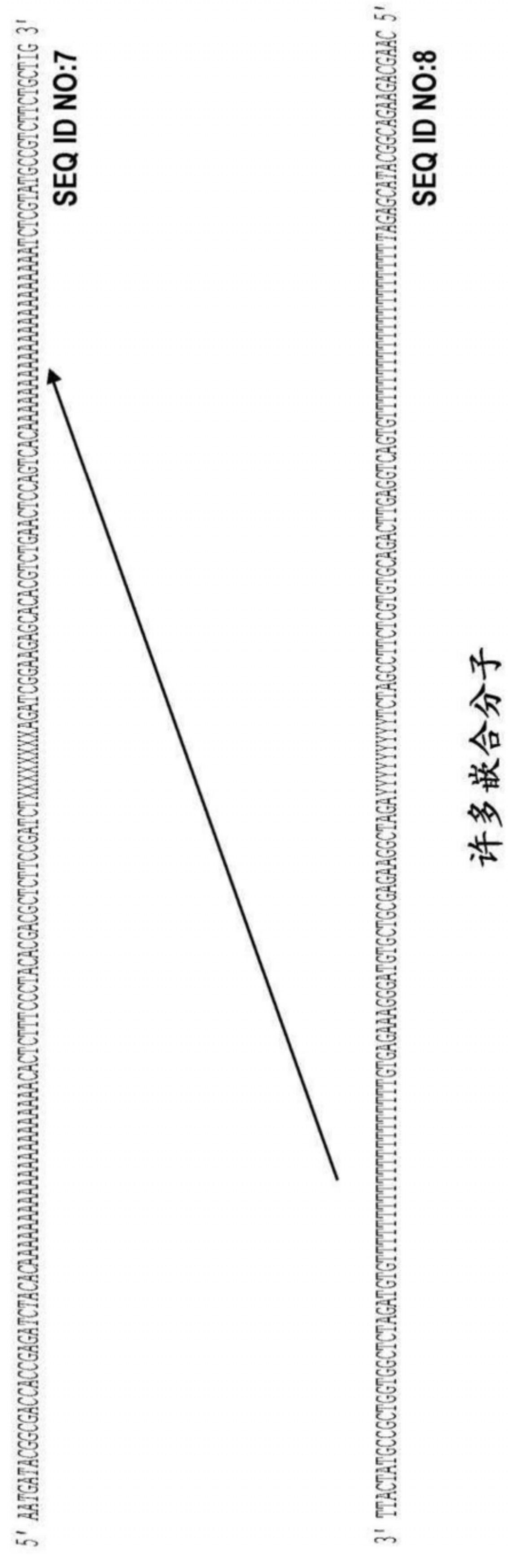


图4

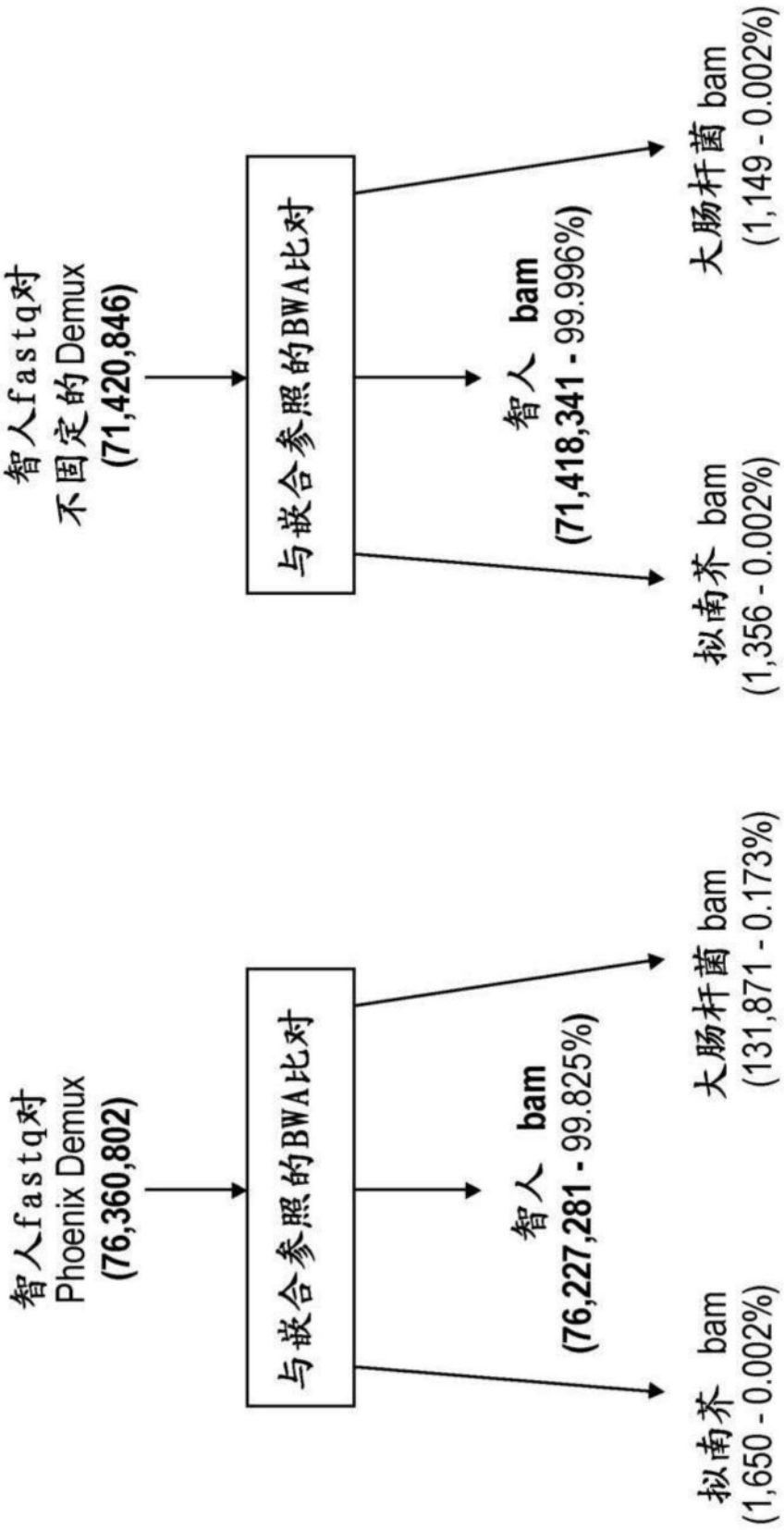

智人fastq对
不固定的Demux
(71,420,846)
与嵌合参照的BWA比对
智人 bam
(71,418,341 - 99.996%)
大肠杆菌 bam
(1,149 - 0.002%)
拟南芥 bam
(1,356 - 0.002%)
智人fastq对
Phoenix Demux
(76,360,802)
与嵌合参照的BWA比对
智人 bam
(76,227,281 - 99.825%)
大肠杆菌 bam
(131,871 - 0.173%)
拟南芥 bam
(1,650 - 0.002%)

图5

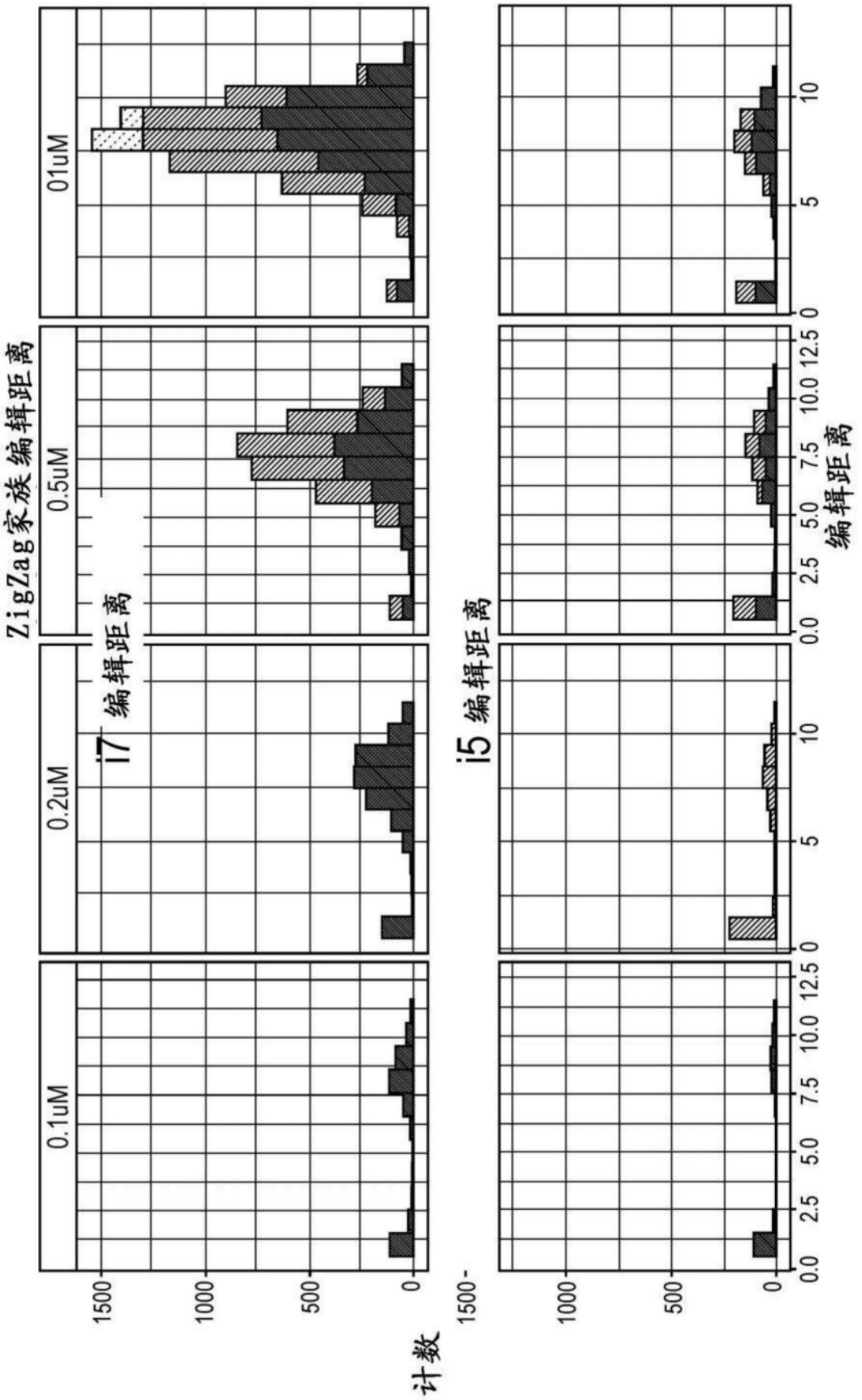

ZigZag家族 编辑距离
0.1uM
0.2uM
0.5uM
01uM
i7 编辑距离
i5 编辑距离
计数
1500
1000
500
0
0.0
2.5
5.0
7.5
10.0
12.5
编辑距离

图6

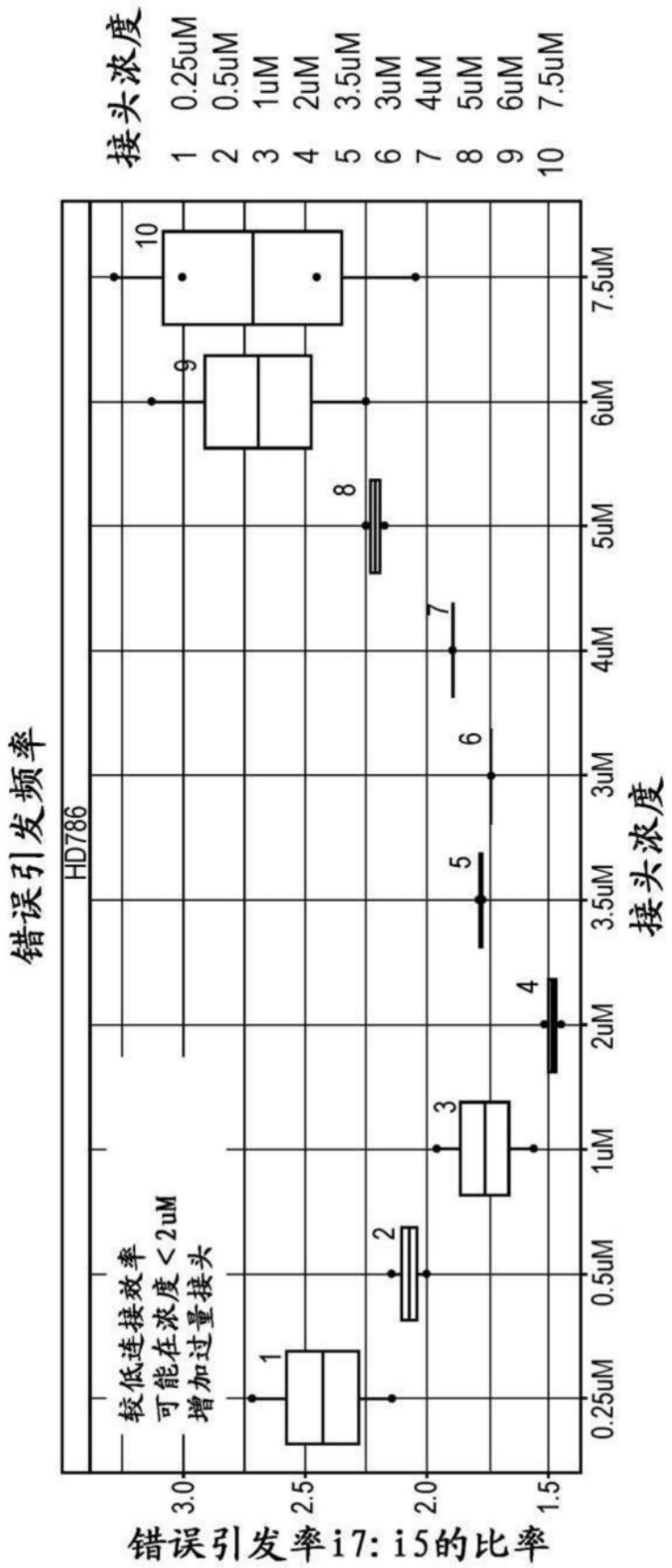

接头浓度
1 0.25uM
2 0.5uM
3 1uM
4 2uM
5 3.5uM
6 3uM
7 4uM
8 5uM
9 6uM
10 7.5uM
HD786
较低连接效率
可能在浓度<2uM
增加过量接头
错误引发频率
错误引发率i7:i5的比率
接头浓度
0.25uM
0.5uM
1uM
2uM
3.5uM
3uM
4uM
5uM
6uM
7.5uM
3.0
2.5
2.0
1.5

图7

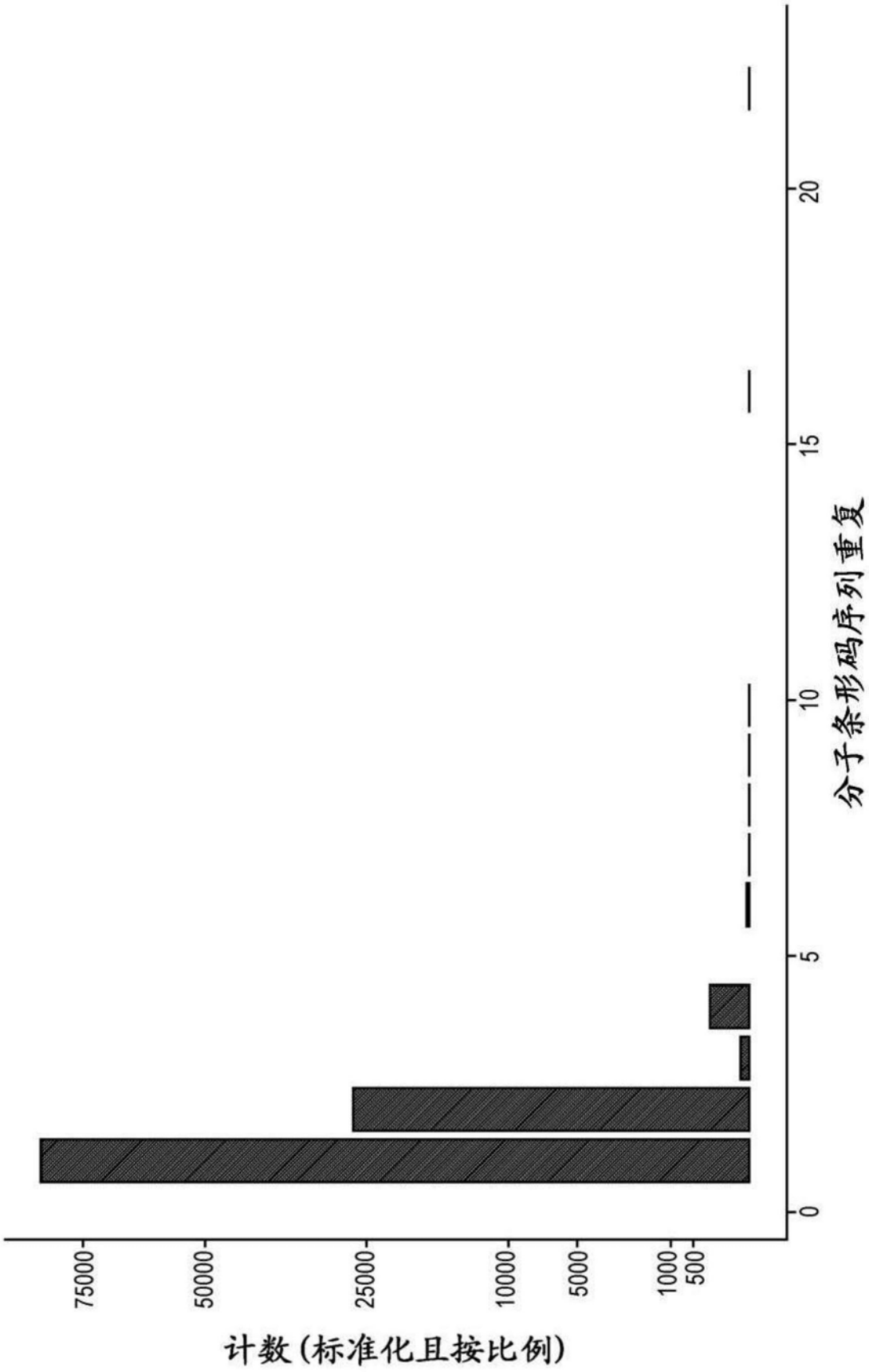

分子条形码序列重复
0
5
10
15
20
计数(标准化且按比例)
75000
50000
25000
10000
5000
1000
500

图8