(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113454218 A
(43)申请公布日 2021.09.28

(21)申请号 201980092697.0
(22)申请日 2019.12.20
(30)优先权数据
62/783,046 2018.12.20 US
(85)PCT国际申请进入国家阶段日
2021.08.20
(86)PCT国际申请的申请数据
PCT/US2019/068146 2019.12.20
(87)PCT国际申请的公布数据
WO2020/132628 EN 2020.06.25
(71)申请人 夸登特健康公司
地址 美国加利福尼亚州
(72)发明人 斯特凡尼·安·沃德·莫蒂默
达斯汀·霍华德·海特
劳拉·米歇尔·梅尔罗伊
(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 韦昌金 郑霞
(51)Int.Cl.
C12N 15/10(2006.01)
C12Q 1/6806(2006.01)

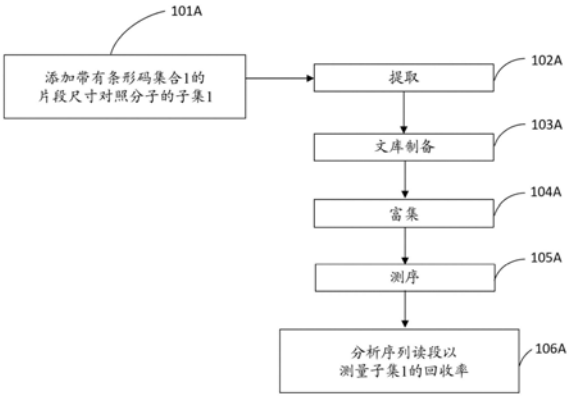
权利要求书10页 说明书48页 附图7页

(54)发明名称

用于改进核酸分子的回收的方法、组合物和系统

(57)摘要

在一方面，本公开内容提供了一种用于分析多核苷酸样品中的核酸分子的方法，该方法包括：(a)将片段尺寸对照分子的子集添加至核酸分子，从而产生第一加标样品；(b)从第一加标样品提取核酸；(c)处理提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将第一加标样品的至少一个子集分区、加标签和/或扩增；(d)富集经处理的样品的至少一个子集，从而产生富集的样品；(e)对富集的样品的至少一个子集测序以生成序列读段；(f)分析序列读段以生成片段尺寸对照分子的子集的片段尺寸评分；和(g)将片段尺寸评分与片段尺寸阈值进行比较。



CN 113454218 A

1.片段尺寸对照分子的集合,所述片段尺寸对照分子的集合包括预定的片段尺寸对照分子的至少一个子集,

其中所述预定的片段尺寸对照分子的至少一个子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。

2.根据权利要求1所述的片段尺寸对照分子的集合,其中所述至少一个子集包括至少一组片段尺寸对照分子。

3.根据权利要求2所述的片段尺寸对照分子的集合,其中一个组中的所述片段尺寸对照分子的片段尺寸区域具有相同的长度。

4.根据权利要求2或3所述的片段尺寸对照分子的集合,其中第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组片段尺寸对照分子中的片段尺寸区域的长度。

5.根据权利要求1所述的片段尺寸对照分子的集合,其中所述片段尺寸对照分子还包含标识符区域。

6.根据权利要求5所述的片段尺寸对照分子的集合,其中所述标识符区域位于所述片段尺寸区域的一侧或两侧。

7.根据权利要求5所述的片段尺寸对照分子的集合,其中所述标识符区域包含分子条形码。

8.根据上述权利要求中任一项所述的片段尺寸对照分子的集合,其中所述多于一个片段对照分子包含一个或更多个引物结合位点。

9.根据权利要求8所述的片段尺寸对照分子的集合,其中所述一个或更多个引物结合位点位于所述标识符区域中。

10.根据权利要求2-4中任一项所述的片段尺寸对照分子的集合,其中一个组中的所述片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。

11.根据权利要求2-9中任一项所述的片段尺寸对照分子的集合,其中一个组中的所述片段尺寸对照分子的片段尺寸区域包含至少两个种可区分的寡核苷酸序列。

12.根据上述权利要求中任一项所述的片段尺寸对照分子的集合,其中预定的片段尺寸对照分子的第一子集中的片段尺寸对照分子的片段尺寸区域包含可与预定的片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

13.根据上述权利要求中任一项所述的片段尺寸对照分子的集合,其中所述片段尺寸区域的长度为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、至少800bp、至少900bp或至少1000bp。

14.根据上述权利要求中任一项所述的片段尺寸对照分子的集合,其中所述片段尺寸区域的长度在10bp和1000bp之间。

15.根据上述权利要求中任一项所述的片段尺寸对照分子的集合,其中所述预定的片段尺寸对照分子的至少一个子集中的每个子集处于等摩尔浓度。

16.根据上述权利要求中任一项所述的片段尺寸对照分子的集合,其中所述预定的片段尺寸对照分子的至少一个子集中的每个子集处于非等摩尔浓度。

17.根据上述权利要求中任一项所述的片段尺寸对照分子的集合,其中所述至少一个

子集中的至少一组片段尺寸对照分子的每个组处于等摩尔浓度。

18.根据权利要求2-16中任一项所述的片段尺寸对照分子的集合，其中所述至少一个子集中的至少一组片段尺寸对照分子的每个组处于非等摩尔浓度。

19.一种核酸群体，所述核酸群体包括：

a.片段尺寸对照分子的集合，所述片段尺寸对照分子的集合包括预定的片段尺寸对照分子的至少一个子集，其中所述预定的片段尺寸对照分子的至少一个子集包括多于一个包含片段尺寸区域的片段尺寸对照分子；和

b.来自受试者的多核苷酸样品中的核酸分子的集合。

20.根据权利要求19所述的核酸群体，其中所述至少一个子集包括至少一组片段尺寸对照分子。

21.根据权利要求20所述的核酸群体，其中一个组中的所述片段尺寸对照分子的片段尺寸区域具有相同的长度。

22.根据权利要求20或21所述的核酸群体，其中第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组片段尺寸对照分子中的片段尺寸区域的长度。

23.根据权利要求19所述的核酸群体，其中所述片段尺寸对照分子还包含标识符区域。

24.根据权利要求23所述的核酸群体，其中所述标识符区域位于所述片段尺寸区域的一侧或两侧。

25.根据权利要求23所述的核酸群体，其中所述标识符区域包含分子条形码。

26.根据权利要求19-26中任一项所述的核酸群体，其中所述多于一个片段对照分子包含一个或更多个引物结合位点。

27.根据权利要求26所述的核酸群体，其中所述一个或更多个引物结合位点位于所述标识符区域中。

28.根据权利要求20-22中任一项所述的核酸群体，其中一个组中的所述片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。

29.根据权利要求20-27中任一项所述的片段尺寸对照分子的集合，其中一个组中的所述片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。

30.根据权利要求19-29中任一项所述的核酸群体，其中预定的片段尺寸对照分子的第一子集中的片段尺寸对照分子的片段尺寸区域包含可与预定的片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

31.根据权利要求19-30中任一项所述的核酸群体，其中所述片段尺寸区域的长度为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、至少800bp、至少900bp或至少1000bp。

32.根据权利要求19-31中任一项所述的核酸群体，其中所述片段尺寸区域的长度在10bp和1000bp之间。

33.根据权利要求19-32中任一项所述的核酸群体，其中所述预定的片段尺寸对照分子的至少一个子集中的每个子集处于等摩尔浓度。

34.根据权利要求19-33中任一项所述的核酸群体，其中所述预定的片段尺寸对照分子的至少一个子集中的每个子集处于非等摩尔浓度。

35.根据权利要求19-34中任一项所述的核酸群体,其中所述至少一个子集中的至少一组片段尺寸对照分子的每个组处于等摩尔浓度。

36.根据权利要求20-34中任一项所述的核酸群体,其中所述至少一个子集中的至少一组片段尺寸对照分子的每个组处于非等摩尔浓度。

37.一种用于分析多核苷酸样品中的核酸分子的方法,所述方法包括:

a)将片段尺寸对照分子的子集添加至所述多核苷酸样品中的核酸分子,从而产生第一加标样品;

b)从所述第一加标样品提取核酸;

c)处理提取的核酸的至少一个子集,从而产生经处理的样品,

其中所述处理包括将所述第一加标样品的至少一个子集分区、加标签和/或扩增;

d)富集所述经处理的样品的至少一个子集,从而产生富集的样品;

e)对所述富集的样品的至少一个子集测序以生成多于一个序列读段;和

f)分析所述多于一个序列读段以生成所述片段尺寸对照分子的子集的多于一个片段尺寸评分。

38.根据权利要求37所述的方法,所述方法还包括,在c)之前,添加片段尺寸对照分子的第二子集,从而产生第二加标样品。

39.根据权利要求38所述的方法,所述方法还包括,在d)之前,添加片段尺寸对照分子的第三子集,从而产生第三加标样品。

40.根据权利要求39所述的方法,所述方法还包括,在e)之前,添加片段尺寸对照分子的第四子集,从而产生第四加标样品。

41.一种用于分析多核苷酸样品中的核酸分子的方法,所述方法包括:

a.将片段尺寸对照分子的第一子集添加至所述多核苷酸样品中的核酸分子,从而产生第一加标样品;

b.从所述第一加标样品提取核酸;

c.将片段尺寸对照分子的第二子集添加至提取的核酸,从而产生第二加标样品;

d.处理所述第二加标样品的至少一个子集,从而产生经处理的样品,

其中所述处理包括将所述第二加标样品的所述至少一个子集分区、加标签和/或扩增;

e.将片段尺寸对照分子的第三子集添加至所述经处理的样品,从而产生第三加标样品;

f.富集所述第三加标样品的至少一个子集,从而产生富集的样品;

g.将片段尺寸对照分子的第四子集添加至所述富集的样品的至少一个子集,从而产生第四加标样品;

h.对所述第四加标样品测序以生成多于一个序列读段;和

i.分析所述多于一个序列读段以生成所述片段尺寸对照分子的第一子集、所述片段尺寸对照分子的第二子集、所述片段尺寸对照分子的第三子集和/或所述片段尺寸对照分子的第四子集的多于一个片段尺寸评分。

42.根据权利要求37或41所述的方法,所述方法还包括将所述多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较。

43.根据权利要求42所述的方法,所述方法还包括基于所述多于一个片段尺寸评分优

化所述多核苷酸样品中的核酸分子的分析。

44.根据权利要求42所述的方法，所述方法还包括使用所述多于一个片段尺寸评分来校正所述多核苷酸样品中的核酸分子的分析中的片段尺寸偏差。

45.根据权利要求42所述的方法，所述方法还包括(i)如果所述多于一个片段尺寸评分中的至少一个在所述多于一个片段尺寸阈值的相应片段尺寸阈值内，则将所述方法分类为成功；或者(ii)如果所述多于一个片段尺寸评分中的至少一个不在所述多于一个片段尺寸阈值的相应片段尺寸阈值内，则将所述方法分类为不成功。

46.一种检测第一样品被第二样品污染的方法，所述方法包括，

对于所述第一样品和所述第二样品中的每一个：

a.添加片段尺寸对照分子的子集以生成第一加标样品，其中添加至所述第一样品的片段尺寸对照分子的子集可以与添加至所述第二样品的片段尺寸对照分子的子集区分；

b.从所述第一加标样品提取核酸；

c.处理提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将所述第一加标样品的至少一个子集分区、加标签和/或扩增；

d.富集所述经处理的样品的至少一个子集，从而产生富集的样品；

e.对所述富集的样品的至少一个子集测序以生成多于一个序列读段；和

f.分析所述多于一个序列读段以生成所述片段尺寸对照分子的子集的一个或更多个污染评分。

47.根据权利要求46所述的方法，所述方法还包括，在c)之前，添加片段尺寸对照分子的第二子集，从而产生第二加标样品，其中添加至所述第一样品的片段尺寸对照分子的子集可以与添加至所述第二样品的片段尺寸对照分子的子集区分。

48.根据权利要求47所述的方法，所述方法还包括，在d)之前，添加片段尺寸对照分子的第三子集，从而产生第三加标样品，其中添加至所述第一样品的片段尺寸对照分子的子集可以与添加至所述第二样品的片段尺寸对照分子的子集区分。

49.根据权利要求48所述的方法，所述方法还包括，在e)之前，添加片段尺寸对照分子的第四子集，从而产生第四加标样品，其中添加至所述第一样品的片段尺寸对照分子的子集可以与添加至所述第二样品的片段尺寸对照分子的子集区分。

50.一种检测第一样品被第二样品污染的方法，所述方法包括，

对于所述第一样品和所述第二样品中的每一个：

a.添加片段尺寸对照分子的第一子集以生成第一加标样品，其中添加至所述第一样品的片段尺寸对照分子的第一子集可以与添加至所述第二样品的片段尺寸对照分子的第一子集区分；

b.从所述第一加标样品提取核酸；

c.将片段尺寸对照分子的第二子集添加至提取的核酸，从而产生第二加标样品，其中添加至所述第一样品的片段尺寸对照分子的第二子集可以与添加至所述第二样品的片段尺寸对照分子的第二子集区分；

d.处理所述提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将所述第一加标样品的至少一个子集分区、加标签和/或扩增；

e.将片段尺寸对照分子的第三子集添加至所述提取的核酸，从而产生第三加标样品，

其中添加至所述第一样品的片段尺寸对照分子的第三子集可以与添加至所述第二样品的片段尺寸对照分子的第三子集区分；

f.富集所述经处理的样品的至少一个子集；

g.将片段尺寸对照分子的第四子集添加至所述提取的核酸，从而产生第四加标样品，其中添加至所述第一样品的片段尺寸对照分子的第四子集可以与添加至所述第二样品的片段尺寸对照分子的第四子集区分；

h.对所述富集的样品的至少一个子集测序以生成多于一个序列读段；和

i.分析所述多于一个序列读段以生成所述片段尺寸对照分子的子集的一个或更多个污染评分。

51.根据权利要求46-50所述的方法，所述方法还包括将至少一个或更多个污染评分与至少一个或更多个污染阈值进行比较。

52.根据权利要求51所述的方法，所述方法还包括，(i)如果至少一个或更多个所述污染评分不在所述一个或更多个污染阈值的相应污染阈值内，则将所述第一样品分类为被所述第二样品污染；或者(ii)如果至少一个或更多个所述污染评分在所述一个或更多个污染阈值的相应污染阈值内，则将所述第一样品分类为没有被所述第二样品污染。

53.根据权利要求37-52中任一项所述的方法，其中所述片段尺寸对照分子的子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。

54.根据权利要求53所述的方法，其中所述片段尺寸对照分子还包含标识符区域。

55.根据权利要求53或54所述的方法，其中所述子集包括至少一组片段尺寸对照分子。

56.根据权利要求55所述的方法，其中一个组中的多于一个片段尺寸对照分子的片段尺寸区域具有相同的长度。

57.根据权利要求55所述的方法，其中第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组片段尺寸对照分子中的片段尺寸区域的长度。

58.根据权利要求54所述的方法，其中所述标识符区域位于所述片段尺寸区域的一侧或两侧。

59.根据权利要求54所述的方法，其中所述标识符区域包含分子条形码。

60.根据上述权利要求中任一项所述的方法，其中所述片段尺寸对照分子包含一个或更多个引物结合位点。

61.根据权利要求60所述的方法，其中所述引物结合位点位于所述标识符区域中。

62.根据权利要求37-61中任一项所述的方法，其中一个组中的所述片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。

63.根据权利要求55所述的方法，其中一个组中的所述片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。

64.根据权利要求37-63中任一项所述的方法，其中片段尺寸对照分子的第一子集中的片段尺寸对照分子的片段尺寸区域包含可与片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

65.根据权利要求37-64中任一项所述的方法，其中所述片段尺寸区域的长度为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、

至少800bp、至少900bp或至少1000bp。

66.根据权利要求37-64中任一项所述的方法，其中所述片段尺寸区域的长度在10bp和1000bp之间。

67.根据权利要求38-66中任一项所述的方法，其中所述片段尺寸对照分子的子集中的每一个处于等摩尔浓度。

68.根据权利要求38-66中任一项所述的方法，其中所述片段尺寸对照分子的子集中的每一个处于非等摩尔浓度。

69.根据权利要求37-68中任一项所述的方法，其中所述子集中的片段尺寸对照分子的组中的每一个处于等摩尔浓度。

70.根据权利要求37-68中任一项所述的方法，其中所述子集中的片段尺寸对照分子的组中的每一个处于非等摩尔浓度。

71.根据权利要求37或41所述的方法，其中所述分区包括将所述第二加标样品的所述至少一个子集的核酸分子分区成多于一个分区的集合。

72.根据权利要求71所述的方法，其中所述多于一个分区的集合包括基于所述第二加标样品的核酸分子的表观遗传修饰水平分区的所述第二加标样品的核酸分子。

73.根据权利要求37或41所述的方法，其中所述加标签包括将标签的集合附接至所述核酸以产生加标签的核酸的群体，其中所述加标签的核酸包含一个或更多个标签。

74.根据权利要求73所述的方法，其中在由所述分区产生的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。

75.根据权利要求74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而附接至所述核酸，其中所述衔接子包含一个或更多个标签。

76.一种用于产生多核苷酸样品的测序文库的方法，所述方法包括：

a)将片段尺寸对照分子的子集添加至所述样品，从而产生第一加标样品；

b)从所述第一加标样品提取核酸；

c)处理提取的核酸的至少一个子集，从而产生经处理的样品，

其中所述处理包括将所述第一加标样品的至少一个子集分区、加标签和/或扩增；和

d)富集所述经处理的样品的至少一个子集。

77.根据权利要求76所述的方法，所述方法还包括，在c)之前，添加片段尺寸对照分子的第二子集，从而产生第二加标样品。

78.根据权利要求77所述的方法，所述方法还包括，在d)之前，添加片段尺寸对照分子的第三子集，从而产生第三加标样品。

79.根据权利要求78所述的方法，所述方法还包括，e)添加片段尺寸对照分子的第四子集，从而产生第四加标样品。

80.根据权利要求37-79中任一项所述的方法，其中所述片段尺寸对照分子的量在1阿托摩尔和10皮摩尔之间。

81.根据权利要求37-80中任一项所述的方法，其中所述多核苷酸样品是无细胞多核苷酸样品。

82.根据权利要求81所述的方法，其中所述无细胞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。

83.根据权利要求81所述的方法,其中所述无细胞多核苷酸样品是无细胞DNA样品。

84.根据权利要求81所述的方法,其中所述无细胞DNA在1ng和500ng之间。

85.一种系统,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,进行包括以下的方法:

a.将片段尺寸对照分子的子集添加至多核苷酸样品中的核酸分子,从而产生第一加标样品;

b.从所述第一加标样品提取核酸;

c.处理提取的核酸的至少一个子集,从而产生经处理的样品,其中所述处理包括将所述第一加标样品的至少一个子集分区、加标签和/或扩增;

d.富集所述经处理的样品的至少一个子集,从而产生富集的样品;

e.对所述富集的样品的至少一个子集测序以生成多于一个序列读段;和

f.分析所述多于一个序列读段以生成所述片段尺寸对照分子的子集的多于一个片段尺寸评分。

86.根据权利要求85所述的系统,还包括,在c)之前,添加片段尺寸对照分子的第二子集,从而产生第二加标样品。

87.根据权利要求86所述的系统,还包括,在d)之前,添加片段尺寸对照分子的第三子集,从而产生第三加标样品。

88.根据权利要求87所述的系统,其中所述方法还包括,在e)之前,添加片段尺寸对照分子的第四子集,从而产生第四加标样品。

89.一种系统,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,进行包括以下的方法:

a.将片段尺寸对照分子的第一子集添加至所述多核苷酸样品中的核酸分子,从而产生第一加标样品;

b.从所述第一加标样品提取核酸;

c.将片段尺寸对照分子的第二子集添加至提取的核酸,从而产生第二加标样品;

d.处理所述第二加标样品的至少一个子集,从而产生经处理的样品,其中所述处理包括将所述第二加标样品的所述至少一个子集分区、加标签和/或扩增;

e.将片段尺寸对照分子的第三子集添加至所述经处理的样品,从而产生第三加标样品;

f.富集所述第三加标样品的至少一个子集,从而产生富集的样品;

g.将片段尺寸对照分子的第四子集添加至所述富集的样品的至少一个子集,从而产生第四加标样品;

h.对所述第四加标样品测序以生成多于一个序列读段;和

i.分析所述多于一个序列读段以生成所述片段尺寸对照分子的第一子集、所述片段尺寸对照分子的第二子集、所述片段尺寸对照分子的第三子集和/或所述片段尺寸对照分子的第四子集的多于一个片段尺寸评分。

90.根据权利要求85或89所述的系统，其中所述方法还包括将所述多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较。

91.根据权利要求90所述的系统，其中所述方法还包括基于所述多于一个片段尺寸评分优化所述多核苷酸样品中的核酸分子的分析。

92.根据权利要求90所述的系统，其中所述方法还包括使用所述多于一个片段尺寸评分来校正所述多核苷酸样品中的核酸分子的分析中的片段尺寸偏差。

93.根据权利要求90所述的系统，其中所述方法还包括(i)如果所述多于一个片段尺寸评分中的每一个在所述多于一个片段尺寸阈值的相应片段尺寸阈值内，则将所述方法分类为成功；或者(ii)如果所述多于一个片段尺寸评分中的至少一个不在所述多于一个片段尺寸阈值的相应片段尺寸阈值内，则将所述方法分类为不成功。

94.根据权利要求85或89所述的系统，其中所述方法还包括将所述多于一个片段尺寸评分中的至少一个与多于一个污染阈值中的至少一个进行比较。

95.根据权利要求94所述的方法，其中所述方法还包括，(i)如果所述多于一个片段尺寸评分中的至少一个不在所述多于一个污染阈值的相应污染阈值内，则将所述样品分类为被另一个样品污染；或者(ii)如果所述多于一个片段尺寸评分中的至少一个在所述多于一个污染阈值的相应污染阈值内，则将所述样品分类为没有被另一个样品污染。

96.根据权利要求85-95中任一项所述的系统，其中所述片段尺寸对照分子的子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。

97.根据权利要求96所述的系统，其中所述片段尺寸对照分子还包含标识符区域。

98.根据权利要求96所述的系统，其中所述片段尺寸对照分子的子集包括至少一组片段尺寸对照分子。

99.根据权利要求98所述的系统，其中一个组中的多于一个片段尺寸对照分子的片段尺寸区域具有相同的长度。

100.根据权利要求98所述的系统，其中第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组片段尺寸对照分子中的片段尺寸区域的长度。

101.根据权利要求97所述的系统，其中所述标识符区域位于所述片段尺寸区域的一侧或两侧。

102.根据权利要求97所述的系统，其中所述标识符区域包含分子条形码。

103.根据权利要求85-102中任一项所述的系统，其中所述片段尺寸对照分子包含一个或更多个引物结合位点。

104.根据权利要求103所述的系统，其中所述引物结合位点位于所述标识符区域中。

105.根据权利要求85-104中任一项所述的系统，其中一个组中的所述片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。

106.根据权利要求96所述的系统，其中一个组中的所述片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。

107.根据权利要求96所述的系统，其中片段尺寸对照分子的第一子集中的片段尺寸对照分子的片段尺寸区域包含可与片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

108.根据权利要求85-107中任一项所述的系统，其中所述片段尺寸区域的长度为至少

10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、至少800bp、至少900bp或至少1000bp。

109.根据权利要求85-107中任一项所述的系统，其中所述片段尺寸区域的长度在10bp和1000bp之间。

110.根据权利要求85-109中任一项所述的系统，其中所述片段尺寸对照分子的子集中的每一个处于等摩尔浓度。

111.根据权利要求85-109中任一项所述的系统，其中所述片段尺寸对照分子的子集中的每一个处于非等摩尔浓度。

112.根据权利要求85-111中任一项所述的系统，其中所述子集中的片段尺寸对照分子的组中的每一个处于非等摩尔浓度。

113.根据权利要求85-111中任一项所述的系统，其中所述子集中的片段尺寸对照分子的组中的每一个处于等摩尔浓度。

114.根据权利要求85或89所述的系统，其中所述分区包括将所述第二加标样品的所述至少一个子集的核酸分子分区成多于一个分区的集合。

115.根据权利要求114所述的系统，其中所述多于一个分区的集合包括基于所述第二加标样品的核酸分子的表观遗传修饰水平分区的所述第二加标样品的核酸分子。

116.根据权利要求85或89所述的系统，其中所述加标签包括将标签的集合附接至所述核酸以产生加标签的核酸的群体，其中所述加标签的核酸包含一个或更多个标签。

117.根据权利要求116所述的系统，其中在由所述分区产生的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。

118.根据权利要求116所述的系统，其中所述标签的集合通过将衔接子连接至所述核酸而附接至所述核酸，其中所述衔接子包含一个或更多个标签。

119.一种系统，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，进行包括以下的方法：

a.将片段尺寸对照分子的子集添加至多核苷酸样品，从而产生第一加标样品；

b.从所述第一加标样品提取核酸；

c.处理提取的核酸的至少一个子集，从而产生经处理的样品，

其中所述处理包括将所述第一加标样品的至少一个子集分区、加标签和/或扩增；和

d.富集所述经处理的样品的至少一个子集。

120.根据权利要求119所述的系统，其中所述方法还包括，在c)之前，添加片段尺寸对照分子的第二子集，从而产生第二加标样品。

121.根据权利要求120所述的系统，其中所述方法还包括，在d)之前，添加片段尺寸对照分子的第三子集，从而产生第三加标样品。

122.根据权利要求121所述的系统，其中所述方法还包括e)添加片段尺寸对照分子的第四子集，从而产生第四加标样品。

123.根据权利要求85-122中任一项所述的系统，其中所述片段尺寸对照分子的量在1

阿托摩尔和10皮摩尔之间。

124.根据权利要求85-123中任一项所述的系统,其中所述多核苷酸样品是无细胞多核苷酸样品。

125.根据权利要求123所述的系统,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。

126.根据权利要求123所述的系统,其中所述无细胞多核苷酸样品是无细胞DNA样品。

127.根据权利要求126所述的系统,其中所述无细胞DNA在1ng和500ng之间。

128.根据权利要求37-127中任一项所述的方法或系统,所述方法或系统还包括生成报告,所述报告任选地包括关于所述核酸分子的分析的信息和/或源自所述核酸分子的分析的信息。

129.根据权利要求128所述的方法或系统,所述方法或系统还包括将所述报告传送给第三方,所述第三方诸如是样品来源的受试者或健康护理从业者。

130.根据上述权利要求中任一项所述的方法或系统,其中所述片段尺寸对照分子是合成的分子。

131.根据上述权利要求中任一项所述的方法或系统,其中所述片段尺寸对照分子通过PCR扩增以扩增子生成。

用于改进核酸分子的回收的方法、组合物和系统

[0001] 交叉引用

[0002] 本申请要求于2018年12月20日提交的美国临时申请第62/783,046号的权益和优先权，该申请通过引用以其整体并入本文。

[0003] 背景

[0004] 当前的癌症诊断测定无细胞核酸(例如，无细胞DNA或无细胞RNA)的方法集中于检测肿瘤相关的体细胞变异，包括单核苷酸变异(SNV)、拷贝数变异(CNV)、融合和插入缺失(indel)(即，插入或缺失)，它们都是用于液体活组织检查的主流靶。越来越多的证据表明，无细胞DNA的尺寸分布和片段化模式可以提供关于无细胞DNA来源和疾病水平的信息。无细胞DNA的尺寸分布和片段化模式，当与体细胞突变调用组合时，可以产生比单独由任一种方法可得的肿瘤状态评估更全面的肿瘤状态评估。

[0005] 概述

[0006] 本公开内容提供了用于使用片段尺寸分子分析核酸的方法、组合物和系统。

[0007] 在一方面，本公开内容提供了一种用于分析多核苷酸样品中的核酸分子的方法，该方法包括：a)将片段尺寸对照分子的子集添加至无细胞多核苷酸样品中的核酸分子，从而产生第一加标(spike-in)样品；b)从第一加标样品提取核酸；c)处理提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将第一加标样品的至少一个子集分区、加标签和/或扩增；d)富集经处理的样品的至少一个子集，从而产生富集的样品；e)对富集的样品的至少一个子集测序以生成多于一个序列读段；和f)分析多于一个序列读段以生成片段尺寸对照分子的子集的多于一个片段尺寸评分。在一些实施方案中，该方法还包括在c)之前，添加片段尺寸对照分子的第二子集，从而产生第二加标样品。在一些实施方案中，该方法还包括在d)之前，添加片段尺寸对照分子的第三子集，从而产生第三加标样品。在一些实施方案中，该方法包括在e)之前添加片段尺寸对照分子的第四子集，从而产生第四加标样品。

[0008] 在一些实施方案中，多核苷酸样品是无细胞多核苷酸样品。在一些实施方案中，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。在一些实施方案中，无细胞多核苷酸的样品是无细胞DNA。在一些实施方案中，无细胞DNA在1ng和500ng之间。在一些实施方案中，片段尺寸对照分子的量在1阿托摩尔(attomole)和10皮摩尔之间。

[0009] 在另一方面，本公开内容提供了片段尺寸对照分子的集合，该片段尺寸对照分子的集合包括预定的片段尺寸对照分子的至少一个子集，其中预定的片段尺寸对照分子的至少一个子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。

[0010] 在一些实施方案中，至少一个子集包括至少一组片段尺寸对照分子。

[0011] 在一些实施方案中，片段尺寸对照分子的子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。在一些实施方案中，子集包括至少一组片段尺寸对照分子。

[0012] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域具有相同的长

度。在一些实施方案中，第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组片段尺寸对照分子中的片段尺寸区域的长度。

[0013] 在一些实施方案中，标识符区域位于片段尺寸区域的一侧或两侧。在一些实施方案中，标识符区域包含分子条形码。

[0014] 在一些实施方案中，多于一个片段对照分子包含一个或更多个引物结合位点。在一些实施方案中，一个或更多个引物结合位点位于标识符区域中。

[0015] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。在一些实施方案中，预定的片段尺寸对照分子的第一子集中的片段尺寸对照分子的片段尺寸区域包含可与预定的片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

[0016] 在一些实施方案中，片段尺寸区域的长度为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、至少800bp、至少900bp或至少1000bp。在一些实施方案中，片段尺寸区域的长度在10bp和1000bp之间。

[0017] 在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于等摩尔浓度。在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于非等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于非等摩尔浓度。

[0018] 在一些实施方案中，片段尺寸对照分子的子集中的每一个处于等摩尔浓度。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于非等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于非等摩尔浓度。

[0019] 在另一方面，本公开内容提供了一种用于评价无细胞多核苷酸样品中的核酸分子的分析中的片段尺寸偏差的方法，该方法包括：a)将片段尺寸对照分子的第一子集添加至多核苷酸样品中的核酸分子，从而产生第一加标样品；b)从第一加标样品提取核酸；c)将片段尺寸对照分子的第二子集添加至提取的核酸，从而产生第二加标样品；d)处理第二加标样品的至少一个子集，从而产生经处理的样品，其中所述处理包括将第二加标样品的所述至少一个子集分区、加标签和/或扩增；e)将片段尺寸对照分子的第三子集添加至经处理的样品，从而产生第三加标样品；f)富集第三加标样品的至少一个子集，从而产生富集的样品；g)将片段尺寸对照分子的第四子集添加至富集的样品的至少一个子集，从而产生第四加标样品；h)对第四加标样品测序以生成多于一个序列读段；和i)分析多于一个序列读段以生成片段尺寸对照分子的第一子集、片段尺寸对照分子的第二子集、片段尺寸对照分子的第三子集和/或片段尺寸对照分子的第四子集的多于一个片段尺寸评分。在一些实施方案中，该方法还包括将多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较。在一些实施方案中，该方法还包括基于多于一个片段尺寸评分优化多核苷酸样品中的核酸分子的分析。在一些实施方案中，该方法还包括使用多于一个片段尺寸评分来校正无细胞多核苷酸样品中的核酸分子的分析中的片段尺寸偏差。在一些实施方案中，该方法还包括(i)如果

多于一个片段尺寸评分中的至少一个在多于一个片段尺寸阈值的相应片段尺寸阈值内，则将该方法分类为成功；或者(ii)如果多于一个片段尺寸评分中的至少一个不在多于一个片段尺寸阈值的相应片段尺寸阈值内，则将该方法分类为不成功。

[0020] 在另一方面，本公开内容提供了检测第一样品被第二样品污染的方法，该方法包括，对于第一样品和第二样品中的每一个：(a)添加片段尺寸对照分子的子集以生成第一加标样品，其中添加至第一样品的片段尺寸对照分子的子集可以与添加至第二样品的片段尺寸对照分子的子集区分；(b)从第一加标样品提取核酸；(c)处理提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将第一加标样品的至少一个子集分区、加标签和/或扩增；(d)富集经处理的样品的至少一个子集；(e)对富集的样品的至少一个子集测序以生成多于一个序列读段；和(f)分析多于一个序列读段以生成片段尺寸对照分子的子集的一个或更多个污染评分。在一些实施方案中，该方法还包括，在c)之前，添加片段尺寸对照分子的第二子集，从而产生第二加标样品，其中添加至第一样品的片段尺寸对照分子的子集可以与添加至第二样品的片段尺寸对照分子的子集区分。在一些实施方案中，该方法还包括，在d)之前，添加片段尺寸对照分子的第三子集，从而产生第三加标样品，其中添加至第一样品的片段尺寸对照分子的子集可以与添加至第二样品的片段尺寸对照分子的子集区分。在一些实施方案中，该方法还包括，在e)之前，添加片段尺寸对照分子的第四子集，从而产生第四加标样品，其中添加至第一样品的片段尺寸对照分子的子集可以与添加至第二样品的片段尺寸对照分子的子集区分。

[0021] 在另一方面，本公开内容提供了检测第一样品被第二样品污染的方法，该方法包括，对于第一样品和第二样品中的每一个：(a)添加片段尺寸对照分子的第一子集以生成第一加标样品，其中添加至第一样品的片段尺寸对照分子的第一子集可以与添加至第二样品的片段尺寸对照分子的第一子集区分；(b)从第一加标样品提取核酸；(c)将片段尺寸对照分子的第二子集添加至提取的核酸，从而产生第二加标样品，其中添加至第一样品的片段尺寸对照分子的第二子集可以与添加至第二样品的片段尺寸对照分子的第二子集区分；(d)处理提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将第一加标样品的至少一个子集分区、加标签和/或扩增；(e)将片段尺寸对照分子的第三子集添加至提取的核酸，从而产生第三加标样品，其中添加至第一样品的片段尺寸对照分子的第三子集可以与添加至第二样品的片段尺寸对照分子的第三子集区分；(f)富集经处理的样品的至少一个子集；(g)将片段尺寸对照分子的第四子集添加至提取的核酸，从而产生第四加标样品，其中添加至第一样品的片段尺寸对照分子的第四子集可以与添加至第二样品的片段尺寸对照分子的第四子集区分；(h)对富集的样品的至少一个子集测序以生成多于一个序列读段；和(i)分析多于一个序列读段以生成片段尺寸对照分子的子集的一个或更多个污染评分。

[0022] 在一些实施方案中，该方法还包括将至少一个或更多个污染评分与至少一个或更多个污染阈值进行比较。在一些实施方案中，该方法还包括，(i)如果至少一个或更多个污染评分不在一个或更多个污染阈值的相应污染阈值内，则将第一样品分类为被第二样品污染；或者(ii)如果至少一个或更多个污染评分在一个或更多个污染阈值的相应污染阈值内，则将第一样品分类为没有被第二样品污染。

[0023] 在一些实施方案中，分区包括将第二加标样品的所述至少一个子集的核酸分子分

区成多于一个分区的集合。在一些实施方案中，多于一个分区的集合包括基于第二加标样品的核酸分子的表观遗传修饰水平分区的第二加标样品的核酸分子。

[0024] 在一些实施方案中，加标签包括将标签的集合附接至核酸以产生加标签的核酸的群体，其中加标签的核酸包含一个或更多个标签。在一些实施方案中，在由分区产生的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。在一些实施方案中，标签的集合通过将衔接子连接至核酸而附接至该核酸，其中衔接子包含一个或更多个标签。

[0025] 在一些实施方案中，多核苷酸样品是无细胞多核苷酸样品。在一些实施方案中，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。在一些实施方案中，无细胞多核苷酸的样品是无细胞DNA。在一些实施方案中，无细胞DNA在1ng和500ng之间。在一些实施方案中，片段尺寸对照分子的量在1阿托摩尔和10皮摩尔之间。

[0026] 在另一方面，本公开内容提供了片段尺寸对照分子的集合，该片段尺寸对照分子的集合包括预定的片段尺寸对照分子的至少一个子集，其中预定的片段尺寸对照分子的至少一个子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。

[0027] 在一些实施方案中，至少一个子集包括至少一组片段尺寸对照分子。

[0028] 在一些实施方案中，片段尺寸对照分子的子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。在一些实施方案中，子集包括至少一组片段尺寸对照分子。

[0029] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域具有相同的长度。在一些实施方案中，第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组片段尺寸对照分子中的片段尺寸区域的长度。

[0030] 在一些实施方案中，标识符区域位于片段尺寸区域的一侧或两侧。在一些实施方案中，标识符区域包含分子条形码。

[0031] 在一些实施方案中，多于一个片段对照分子包含一个或更多个引物结合位点。在一些实施方案中，一个或更多个引物结合位点位于标识符区域中。

[0032] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。在一些实施方案中，预定的片段尺寸对照分子的第一子集中的片段尺寸对照分子的片段尺寸区域包含可与预定的片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

[0033] 在一些实施方案中，片段尺寸区域的长度为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、至少800bp、至少900bp或至少1000bp。在一些实施方案中，片段尺寸区域的长度在10bp和1000bp之间。

[0034] 在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于等摩尔浓度。在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于非等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分

子的每个组处于非等摩尔浓度。

[0035] 在一些实施方案中，片段尺寸对照分子的子集中的每一个处于等摩尔浓度。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于非等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于非等摩尔浓度。

[0036] 在另一方面，本公开内容提供了一种用于产生无细胞多核苷酸样品的测序文库的方法，该方法包括：a)将片段尺寸对照分子的子集添加至样品，从而产生第一加标样品；b)从第一加标样品提取核酸；c)处理提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将第一加标样品的至少一个子集分区、加标签和/或扩增；和d)富集经处理的样品的至少一个子集。在一些实施方案中，该方法还包括，在c)之前，添加片段尺寸对照分子的第二子集，从而产生第二加标样品。在一些实施方案中，该方法还包括，在d)之前，添加片段尺寸对照分子的第三子集，从而产生第三加标样品。在一些实施方案中，该方法还包括e)添加片段尺寸对照分子的第四子集，从而产生第四加标样品。

[0037] 在一些实施方案中，至少一个子集包括至少一组片段尺寸对照分子。

[0038] 在一些实施方案中，片段尺寸对照分子的子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。在一些实施方案中，子集包括至少一组片段尺寸对照分子。

[0039] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域具有相同的长度。在一些实施方案中，第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组片段尺寸对照分子中的片段尺寸区域的长度。

[0040] 在一些实施方案中，标识符区域位于片段尺寸区域的一侧或两侧。在一些实施方案中，标识符区域包含分子条形码。

[0041] 在一些实施方案中，多于一个片段对照分子包含一个或更多个引物结合位点。在一些实施方案中，一个或更多个引物结合位点位于标识符区域中。

[0042] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。在一些实施方案中，预定的片段尺寸对照分子的第一子集中的片段尺寸对照分子的片段尺寸区域包含可与预定的片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

[0043] 在一些实施方案中，片段尺寸区域的长度为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp对、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp或至少500bp。在一些实施方案中，片段尺寸区域的长度在10bp和1000bp之间。

[0044] 在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于等摩尔浓度。在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于非等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于非等摩尔浓度。

[0045] 在又另一方面，本公开内容提供了核酸的群体，该核酸的群体包含：(a)片段尺寸

对照分子的集合，该片段尺寸对照分子的集合包括预定的片段尺寸对照分子的至少一个子集，其中预定的片段尺寸对照分子的至少一个子集包括多于一个包含片段尺寸区域的片段尺寸对照分子；和(b)来自受试者的多核苷酸样品中的核酸分子的集合。在一些实施方案中，至少一个子集包括至少一组片段尺寸对照分子。在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域具有相同的长度。在一些实施方案中，第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组片段尺寸对照分子中的片段尺寸区域的长度。在一些实施方案中，片段尺寸对照分子还包含标识符区域。在一些实施方案中，标识符区域位于片段尺寸区域的一侧或两侧。在一些实施方案中，标识符区域包含分子条形码。在一些实施方案中，多于一个片段对照分子包含一个或更多个引物结合位点。在一些实施方案中，一个或更多个引物结合位点位于标识符区域中。在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。在一些实施方案中，预定的片段尺寸对照分子的第一子集中的片段尺寸对照分子的片段尺寸区域包含可与预定的片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。在一些实施方案中，片段尺寸区域的长度为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、至少800bp、至少900bp或至少1000bp。在一些实施方案中，片段尺寸区域的长度在10bp和1000bp之间。在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于非等摩尔浓度。

[0046] 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，进行包括以下的方法：a)将片段尺寸对照分子的子集添加至多核苷酸样品中的核酸分子，从而产生第一加标样品；b)从第一加标样品提取核酸；c)处理提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将第一加标样品的至少一个子集分区、加标签和/或扩增；d)富集经处理的样品的至少一个子集，从而产生富集的样品；e)对富集的样品的至少一个子集测序以生成多于一个序列读段；和f)分析多于一个序列读段以生成片段尺寸对照分子的子集的多于一个片段尺寸评分。在一些实施方案中，该方法还包括在c)之前，添加片段尺寸对照分子的第二子集，从而产生第二加标样品。在一些实施方案中，该方法还包括在d)之前，添加片段尺寸对照分子的第三子集，从而产生第三加标样品。在一些实施方案中，该方法还包括在e)之前，添加片段尺寸对照分子的第四子集，从而产生第四加标样品。

[0047] 在一些实施方案中，多核苷酸样品是无细胞多核苷酸样品。在一些实施方案中，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。在一些实施方案中，无细胞多核苷酸的样品是无细胞DNA。在一些实施方案中，无细胞DNA在1ng和500ng之间。在一些实施方案中，片段尺寸对照分子的量在1阿托摩尔和10皮摩尔之间。

[0048] 在另一方面，本公开内容提供了片段尺寸对照分子的集合，该片段尺寸对照分子

的集合包括预定的片段尺寸对照分子的至少一个子集，其中预定的片段尺寸对照分子的至少一个子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。

[0049] 在一些实施方案中，至少一个子集包括至少一组片段尺寸对照分子。

[0050] 在一些实施方案中，片段尺寸对照分子的子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。在一些实施方案中，子集包括至少一组片段尺寸对照分子。

[0051] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域具有相同的长度。在一些实施方案中，第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组片段尺寸对照分子中的片段尺寸区域的长度。

[0052] 在一些实施方案中，标识符区域位于片段尺寸区域的一侧或两侧。在一些实施方案中，标识符区域包含分子条形码。

[0053] 在一些实施方案中，多于一个片段对照分子包含一个或更多个引物结合位点。在一些实施方案中，一个或更多个引物结合位点位于标识符区域中。

[0054] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。在一些实施方案中，预定的片段尺寸对照分子的第一子集中的片段尺寸对照分子的片段尺寸区域包含可与预定的片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

[0055] 在一些实施方案中，片段尺寸区域的长度为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、至少800bp、至少900bp或至少1000bp。在一些实施方案中，片段尺寸区域的长度在10bp和1000bp之间。

[0056] 在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于等摩尔浓度。在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于非等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于非等摩尔浓度。

[0057] 在一些实施方案中，片段尺寸对照分子的子集中的每一个处于等摩尔浓度。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于非等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于非等摩尔浓度。

[0058] 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，进行包括以下的方法：a)将片段尺寸对照分子的第一子集添加至多核苷酸样品中的核酸分子，从而产生第一加标样品；b)从第一加标样品提取核酸；c)将片段尺寸对照分子的第二子集添加至提取的核酸，从而产生第二加标样品；d)处理第二加标样品的至少一个子集，从而产生经处理的样品，其中所述处理包括将第二加标样品的所述至少一个子集分区、加标签和/或扩增；e)将

片段尺寸对照分子的第三子集添加至经处理的样品，从而产生第三加标样品；f)富集第三加标样品的至少一个子集，从而产生富集的样品；g)将片段尺寸对照分子的第四子集添加至富集的样品的至少一个子集，从而产生第四加标样品；h)对第四加标样品测序以生成多于一个序列读段；和i)分析多于一个序列读段以生成片段尺寸对照分子的第一子集、片段尺寸对照分子的第二子集、片段尺寸对照分子的第三子集和/或片段尺寸对照分子的第四子集的多于一个片段尺寸评分。在一些实施方案中，该方法还包括将多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较。在一些实施方案中，该方法还包括基于多于一个片段尺寸评分优化多核苷酸样品中的核酸分子的分析。在一些实施方案中，该方法还包括使用多于一个片段尺寸评分来校正多核苷酸样品中的核酸分子的分析中的片段尺寸偏差。在一些实施方案中，该方法还包括(i)如果多于一个片段尺寸评分中的每一个在多于一个片段尺寸阈值的相应片段尺寸阈值内，则将该方法分类为成功；或者(ii)如果多于一个片段尺寸评分中的至少一个不在多于一个片段尺寸阈值的相应片段尺寸阈值内，则将该方法分类为不成功。在一些实施方案中，该方法还包括将多于一个片段尺寸评分中的至少一个与多于一个污染阈值中的至少一个进行比较。在一些实施方案中，该方法还包括(i)如果多于一个片段尺寸评分中的至少一个不在多于一个污染阈值的相应污染阈值内，则将样品分类为被另一个样品污染；或者(ii)如果多于一个片段尺寸评分中的至少一个在多于一个污染阈值的相应污染阈值内，则将样品分类为没有被另一个样品污染。

[0059] 在一些实施方案中，多核苷酸样品是无细胞多核苷酸样品。在一些实施方案中，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。在一些实施方案中，无细胞多核苷酸的样品是无细胞DNA。在一些实施方案中，无细胞DNA在1ng和500ng之间。在一些实施方案中，片段尺寸对照分子的量在1阿托摩尔和10皮摩尔之间。

[0060] 在另一方面，本公开内容提供了片段尺寸对照分子的集合，该片段尺寸对照分子的集合包括预定的片段尺寸对照分子的至少一个子集，其中预定的片段尺寸对照分子的至少一个子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。

[0061] 在一些实施方案中，至少一个子集包括至少一组片段尺寸对照分子。

[0062] 在一些实施方案中，片段尺寸对照分子的子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。在一些实施方案中，子集包括至少一组片段尺寸对照分子。

[0063] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域具有相同的长度。在一些实施方案中，第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组片段尺寸对照分子中的片段尺寸区域的长度。

[0064] 在一些实施方案中，标识符区域位于片段尺寸区域的一侧或两侧。在一些实施方案中，标识符区域包含分子条形码。

[0065] 在一些实施方案中，多于一个片段对照分子包含一个或更多个引物结合位点。在一些实施方案中，一个或更多个引物结合位点位于标识符区域中。

[0066] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。在一些实施方案中，预定的片段尺寸对照分子的第一子集中

的片段尺寸对照分子的片段尺寸区域包含可与预定的片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

[0067] 在一些实施方案中，片段尺寸区域的长度为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、至少800bp、至少900bp或至少1000bp。在一些实施方案中，片段尺寸区域的长度在10bp和1000bp之间。

[0068] 在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于等摩尔浓度。在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于非等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于非等摩尔浓度。

[0069] 在一些实施方案中，片段尺寸对照分子的子集中的每一个处于等摩尔浓度。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于非等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于非等摩尔浓度。

[0070] 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，进行包括以下的方法：a)将片段尺寸对照分子的子集添加至样品，从而产生第一加标样品；b)从第一加标样品提取核酸；c)处理提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将第一加标样品的至少一个子集分区、加标签和/或扩增；和d)富集经处理的样品的至少一个子集。在一些实施方案中，该方法还包括在c)之前，添加片段尺寸对照分子的第二子集，从而产生第二加标样品。在一些实施方案中，该方法还包括在d)之前，添加片段尺寸对照分子的第三子集，从而产生第三加标样品。在一些实施方案中，该方法还包括e)添加片段尺寸对照分子的第四子集，从而产生第四加标样品。

[0071] 在一些实施方案中，多核苷酸样品是无细胞多核苷酸样品。在一些实施方案中，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。在一些实施方案中，无细胞多核苷酸的样品是无细胞DNA。在一些实施方案中，无细胞DNA在1ng和500ng之间。在一些实施方案中，片段尺寸对照分子的量在1阿托摩尔和10皮摩尔之间。

[0072] 在另一方面，本公开内容提供了片段尺寸对照分子的集合，该片段尺寸对照分子的集合包括预定的片段尺寸对照分子的至少一个子集，其中预定的片段尺寸对照分子的至少一个子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。

[0073] 在一些实施方案中，至少一个子集包括至少一组片段尺寸对照分子。

[0074] 在一些实施方案中，片段尺寸对照分子的子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子还包含标识符区域。在一些实施方案中，子集包括至少一组片段尺寸对照分子。

[0075] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域具有相同的长度。在一些实施方案中，第一组片段尺寸对照分子中的片段尺寸区域的长度不同于第二组

片段尺寸对照分子中的片段尺寸区域的长度。

[0076] 在一些实施方案中，标识符区域位于片段尺寸区域的一侧或两侧。在一些实施方案中，标识符区域包含分子条形码。

[0077] 在一些实施方案中，多于一个片段对照分子包含一个或更多个引物结合位点。在一些实施方案中，一个或更多个引物结合位点位于标识符区域中。

[0078] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。在一些实施方案中，预定的片段尺寸对照分子的第一子集中的片段尺寸对照分子的片段尺寸区域包含可与预定的片段尺寸对照分子的第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

[0079] 在一些实施方案中，片段尺寸区域的长度为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、至少800bp、至少900bp或至少1000bp。在一些实施方案中，片段尺寸区域的长度在10bp和1000bp之间。

[0080] 在一些实施方案中，片段尺寸对照分子是合成的分子。在一些实施方案中，片段尺寸对照分子通过PCR扩增以扩增子生成。

[0081] 在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于等摩尔浓度。在一些实施方案中，预定的片段尺寸对照分子的至少一个子集中的每个子集处于非等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于等摩尔浓度。在一些实施方案中，至少一个子集中的至少一组片段尺寸对照分子的每个组处于非等摩尔浓度。

[0082] 在一些实施方案中，片段尺寸对照分子的子集中的每一个处于等摩尔浓度。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于非等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于非等摩尔浓度。

[0083] 本公开内容还提供了用于实施任何上述方法的试剂盒。一种示例性试剂盒包含：(a)片段尺寸对照分子的集合，该片段尺寸对照分子的集合包括预定的片段尺寸对照分子的至少一个子集，其中预定的片段尺寸对照分子的至少一个子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。

[0084] 在一些实施方案中，该方法或系统还包括生成报告，该报告任选地包括关于核酸分子的分析的信息和/或源自核酸分子的分析的信息。

[0085] 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析的信息和/或源自核酸分子的分析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。

[0086] 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。

[0087] 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。

[0088] 附图简述

[0089] 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。

[0090] 图1A是根据本公开内容的一种实施方案，用于分析多核苷酸样品中的核酸分子的方法的示意图。图1B是根据图1A所示实施方案，用于分析多核苷酸样品中的核酸分子的方法的流程图。

[0091] 图2A是根据本公开内容的一种实施方案，用于分析多核苷酸样品中的核酸分子的方法的示意图。图2B是根据图2A所示实施方案，用于分析多核苷酸样品中的核酸分子的方法的流程图。

[0092] 图3是适用于本公开内容的一些实施方案的片段尺寸对照分子的示意图。

[0093] 图4是适用于本公开内容的一些实施方案的片段尺寸对照分子的示意图。

[0094] 图5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。

[0095] 定义

[0096] 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。如果下文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。

[0097] 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和/或在阅读本公开内容等后将变得明显的类型的方法和/或步骤。

[0098] 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。

[0099] 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％)。

[0100] 衔接子：如本文使用的，“衔接子”是指通常是至少部分双链的短核酸(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)，并且衔接子可以附接至特定样品核酸分子的任一末端或两个末端。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特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或用一个或更多个互补核苷酸加尾的核酸分子，并且Y形衔接子的另一个末端包含不杂交形成双链的非互补序列。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。

[0101] 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)。

[0102] 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。

[0103] 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、单链的或其杂合体。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。

[0104] 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,通过细胞裂解)。

[0105] 污染:如本文使用的,术语“污染”或“样品污染”是指一种样品被另一种样品的任何化学污染或数字污染。污染可以由各种来源引起,诸如但不限于:样品之间液体的物理残留(例如,通过样品制备或测序系统的移液、自动化液体处理、操作扩增的材料);解复用伪像(demultiplexing artifacts)(例如,混淆具有有限成对汉明距离的样品索引的碱基调用错误;混淆具有有限成对编辑距离的样品索引的插入/缺失)和试剂杂质(例如,样品索引寡核苷酸被含有另一种样品索引的寡核苷酸污染(通过残留或合成错误))。

[0106] 污染评分:如本文使用的,“污染评分”是指第一样品中代表添加至第二样品的片段尺寸对照分子的存在的评分。在一些实施方案中,在一个样品的一个子集中使用的子集标识符条形码可以不同于在同一样品的其他子集和其他样品的子集中使用的子集标识符。在这些实施方案中,根据片段尺寸对照分子中存在的子集标识符条形码的序列,可以鉴定属于不同的第二样品的片段尺寸对照分子的存在。在一些实施方案中,污染评分可以是对子集中使用的片段尺寸对照分子的每个类型/组特定的(即,子集中使用的片段尺寸对照分子的不同长度/组的单独污染评分),或者污染评分可以是代表片段尺寸对照分子的不同长度/组的总评分。在一些实施方案中,污染评分可以基于属于不同的第二样品的片段尺寸对照分子的数目来估计。在一些实施方案中,污染评分可以基于属于不同的第二样品的片段尺寸对照分子的测序读段数目来估计。在一些实施方案中,样品的污染评分可以基于属于其他样品的片段尺寸对照分子的测序读段数目相对于添加至该样品的片段尺寸对照分子的测序读段数目的分数或百分比来估计。在一些实施方案中,样品的污染评分可以基于属于其他样品的片段尺寸对照分子的分子数目相对于添加至该样品的片段尺寸对照分子的分子数目的分数或百分比来估计。在这些实施方案中,添加至该分子的片段尺寸对照分子可以根据片段尺寸对照分子的子集标识符条形码来鉴定。

[0107] 污染阈值:如本文使用的,“污染阈值”是指用于评价样品中的核酸分子的分析中的样品污染的预定的阈值或范围。这些阈值也可用于在任何步骤(诸如提取、文库制备、富集、洗涤/清洁和测序)中优化测定。在一些实施方案中,污染阈值可以是对子集中使用的片段尺寸对照分子的每个类型/组特定的(即,子集中使用的片段尺寸对照分子的不同长度/组的单独污染阈值)。在一些实施方案中,污染阈值可以是子集中片段尺寸对照分子的不同长度/组的总阈值,或者是在测定中添加的一个或更多个子集中使用的片段尺寸对照分子的总阈值。在一些实施方案中,子集中的每个组具有特定的污染阈值。例如,片段尺寸对照分子的集合包括两个子集—子集1和子集2。如果每个子集包括两个组—G11和G12(对于子集1)以及G21和G22(对于子集2),则基于片段尺寸对照分子的回收率,这四个组中的每个组可以具有污染评分(S)—G11的S11、G12的S12、G21的S21和G22的S22。子集中的每个组可以具有单独的预定的污染阈值。在该实例中,T11、T12、T21和T22分别是G11、G12、G21和G22的污染阈值。阈值可以是百分比或分数,并且阈值可以是阈值范围,而不是特定的阈值。为了使该方法被认为是成功的,这些污染评分中的至少一个应该在其相应的污染阈值内。在一些实施方案中,如果多于一个污染评分中的每一个在相应的污染阈值内,则该方法被分类为成功进行。

[0108] 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。它们是指特定样品中在特定基因组位置处的DNA分子的总数。

[0109] 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。

[0110] DNA序列:如本文使用的,“DNA序列”或“序列”是指“原始序列读段”和/或“共有序列”。原始序列读段是DNA测序仪的输出,并且通常包括相同亲本分子的冗余序列,例如扩增后。“共有序列”是源自意图代表原始亲本分子的序列的亲本分子的冗余序列的序列。共有序列可以通过投票(voting)(其中每种主要核苷酸,例如,在序列中的特定碱基位置处最常观察到的核苷酸是共有核苷酸)或其他方法(诸如与参考基因组进行比较)来产生。共有序列可以通过用独特或非独特的分子标签对原始亲本分子加标签来产生,这允许通过追溯标签和/或使用序列读段内部信息来追溯子序列(例如,扩增后)。加标签或条形码化以及标签或条形码的使用的实例在例如美国专利公布第2015/0368708号、第2015/0299812号、第2016/0040229号和第2016/0046986号中提供,其中每一项通过引用整体并入本文。

[0111] 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碱基的修饰,其影响特定核酸序列的调节和/或基因表达。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。

[0112] 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。

[0113] 富集的样品:如本文使用的,“富集的样品”是指已经针对特定感兴趣区域富集的

样品。可以通过选择性地扩增感兴趣区域或通过使用可与感兴趣核酸分子杂交的双链DNA/RNA探针(例如来自Twist Biosciences的探针)或单链DNA/RNA探针(例如SureSelect®探针,Agilent Technol)来富集样品。在一些实施方案中,富集的样品是指被富集的经处理的样品的子集,其中被富集的经处理的样品的子集包含来自无细胞多核苷酸样品和来自片段尺寸对照分子的第一和/或第二子集的核酸分子。在其他实施方案中,富集的样品是指被富集的第三加标样品的子集,其中被富集的第三加标样品的子集包含来自无细胞多核苷酸样品和来自片段尺寸对照分子的第一、第二和/或第三子集的核酸分子。

[0114] 第一加标样品:如本文使用的,"第一加标样品(first spike-in sample)"是其中片段尺寸对照分子的第一子集已被添加至来自受试者的无细胞多核苷酸样品的样品。

[0115] 第四加标样品:如本文使用的,"第四加标样品"是指其中片段尺寸对照分子的第四子集已经被添加至富集的样品的子集的样品。

[0116] 片段尺寸偏差:如本文使用的,"片段尺寸偏差"是指在测定中对被分析的核酸分子的片段长度或尺寸的任何人为偏差。伪象偏差(artefact bias)可以由操作者的处理或测定中使用的试剂和/或步骤引起。这种偏差不包括在健康受试者和患病受试者之间观察到的关于片段尺寸的真实生物学偏差。测定可包括一个或更多个步骤,诸如但不限于液体处理、提取、文库制备、富集、洗涤/清洁和测序。在一些实施方案中,不同样品之间可以存在不同的片段尺寸偏差。基于这些步骤中每一步骤的反应条件和程序,属于特定片段尺寸的核酸分子的回收率可能有偏差。在片段组分析(fragmentome analysis)中,对特定尺寸的cfDNA分子进行精确的定量测量可用于估计无细胞DNA的片段长度分布和片段化模式。通过估计测定中的片段尺寸偏差,人们可以校正由测定工作流程中每个步骤引入的片段尺寸偏差(人为偏差),并更佳地估计cfDNA分子的原始片段尺寸分布(其反映真实生物学偏差)。

[0117] 片段尺寸对照分子:如本文使用的,"片段尺寸对照分子"是指添加至多核苷酸样品以评价和/或优化样品中核酸分子的分析的一组核酸分子。片段尺寸对照分子可以具有两个区域—片段尺寸区域和标识符区域。在一些实施方案中,片段尺寸对照分子仅由片段尺寸区域组成。在一些实施方案中,片段尺寸对照分子可以具有片段尺寸区域和标识符区域两者。片段尺寸对照分子的集合可以被分为片段尺寸对照分子的子集,并且片段尺寸对照分子的每个子集可以在测定中的一个或更多个不同步骤中被添加,以便在任何步骤(多核苷酸样品的提取、文库制备、富集和/或测序)中评价片段长度分布、QC指标和/或样品污染。在一些实施方案中,片段尺寸对照分子可用于估计样品中多核苷酸的原始末端在整个测定工作流程中良好保存的程度。这些片段尺寸对照分子的长度是预定的,并且可以模拟无细胞DNA的天然尺寸分布。片段尺寸对照分子的每个子集可以基于片段尺寸区域的长度进一步分为片段尺寸对照分子的组,并且每个组可以具有不同长度的片段尺寸区域。例如,片段尺寸对照分子的子集可以基于片段尺寸区域的长度分为三组—第一组可以具有120bp长度的片段尺寸区域,第二组可以具有160bp长度的片段尺寸区域,并且第三组可以具有320bp长度的片段尺寸区域。在一些实施方案中,片段尺寸对照分子可以是合成的分子—即,可以在体外使用或不使用酶来合成。在一些实施方案中,片段尺寸对照分子可以具有非天然存在的核酸序列。在一些实施方案中,片段尺寸对照分子可以具有天然存在的核酸序列。在一些实施方案中,通过扩增任何基因组、质粒或载体或其一部分中的特定区域(即,具有特定长度)而生成的扩增子可以用作片段尺寸对照分子。在一些实施方案中,质粒、载体

或任何基因组可以使用限制酶来消化，并且消化产物可以用作片段尺寸对照分子。在一些实施方案中，片段尺寸对照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，片段尺寸对照分子可以包含非天然存在的核苷酸类似物。

[0118] 片段尺寸区域：如本文使用的，术语“片段尺寸区域”是指片段尺寸对照分子中代表片段尺寸对照分子长度的区域。每组片段尺寸对照分子的片段尺寸区域的长度不同。例如，片段尺寸对照分子的子集可以基于片段尺寸区域的长度分为三组—第一组可以具有120bp长度的片段尺寸区域，第二组可以具有160bp长度的片段尺寸区域，并且第三组可以具有320bp长度的片段尺寸区域。片段尺寸区域的长度可以在10bp和1000bp之间。

[0119] 片段尺寸评分：如本文使用的，“片段尺寸评分”是指代表属于特定组和特定子集的片段尺寸对照分子的回收率的评分。片段尺寸对照分子的身份和片段尺寸对照分子所属的子集通过使用条形码来保持。在一些实施方案中，片段尺寸评分可以基于属于特定组和特定子集的片段尺寸对照分子的数目来估计。在一些实施方案中，片段尺寸评分可以基于属于特定组和特定子集的片段尺寸对照分子的测序读段的数目来估计。在一些实施方案中，可以为特定子集中的特定组估计片段尺寸评分。在一些实施方案中，可以为每个子集中的每个组估计片段尺寸评分。在一些实施方案中，片段尺寸评分可以是在特定子集中使用的片段尺寸对照分子的总评分(即，代表子集中所有片段尺寸对照分子的单个评分)，或者是在测定的一个或更多个步骤中添加的片段尺寸对照分子的总评分(即，代表在测定中添加的一个或更多个子集中的片段尺寸对照分子的单个评分)。在一些实施方案中，片段尺寸评分可以被测量为在特定步骤添加的片段尺寸对照分子的子集的量与在不同步骤添加的片段尺寸对照分子的另一子集的量的差异，并且在一些实施方案中，该量可以是质量(例如，fg、pg、ng、μg)、测序读段数目或片段尺寸分子(属于一个子集或属于特定组)的数目。在一些实施方案中，片段尺寸评分可以被测量为(属于一个子集或属于特定组)的片段尺寸对照分子的回收率(分数或百分比)，其中回收率可以利用添加至样品的特定子集中的片段尺寸对照分子的相对丰度的正交测量结果(orthogonal measurement)来计算。在一些实施方案中，正交测量结果可以从凝胶电泳、qPCR、ddPCR或无PCR测序获得。在一些实施方案中，片段尺寸评分可以被测量为具有特定长度的片段尺寸对照分子的量与具有不同长度的片段尺寸对照分子的量的差异。

[0120] 片段尺寸阈值：如本文使用的，“片段尺寸阈值”是指用于评价样品中的核酸分子的分析中的片段尺寸偏差的预定的阈值或范围。这些阈值也可用于在任何步骤(诸如提取、文库制备、富集、洗涤/清洁和测序)中校正任何片段长度分布。在一些实施方案中，可以为特定子集中的特定组估计片段尺寸阈值。在一些实施方案中，子集中的每个组具有特定的片段尺寸阈值。在一些实施方案中，可以为每个子集中的每个组估计片段尺寸阈值。在一些实施方案中，片段尺寸阈值可以是在特定子集中使用的片段尺寸对照分子的总阈值(即，子集中所有片段尺寸对照分子的单个阈值)，或者是在测定的一个或更多个步骤中添加的片段尺寸对照分子的总阈值(即，在测定中添加的一个或更多个子集中的片段尺寸对照分子的单个阈值)。例如，片段尺寸对照分子的集合包括两个子集—子集1和子集2。如果每个子集包括两个组—G11和G12(对于子集1)以及G21和G22(对于子集2)，则基于片段尺寸对照分子的回收率，这四个组中的每个组可以具有片段尺寸评分(S)—G11的S11、G12的S12、G21的

S21和G22的S22。子集中的每个组可以具有单独的预定的片段尺寸阈值。在该实例中，T11、T12、T21和T22分别是G11、G12、G21和G22的片段尺寸阈值。阈值可以是百分比或分数，并且阈值可以是阈值范围，而不是特定的阈值。为了使该方法被认为是成功的，这些片段尺寸评分中的至少一个应该在其相应的片段尺寸阈值内。在一些实施方案中，如果多于一个片段尺寸评分中的每一个在相应的片段尺寸阈值内，则该方法被分类为成功进行。

[0121] 基因组区域：如本文使用的，“基因组区域”是指基因组(例如，整个基因组、染色体、基因或外显子)的任何区域(例如，碱基对位置的范围)。基因组区域可以是连续或不连续的区域。“遗传基因座”(或“基因座”)可以是基因组区域的一部分或全部(例如，基因、基因的一部分或基因的单个核苷酸)。

[0122] 标识符区域：如本文使用的，“标识符区域”是指片段尺寸对照分子中用于区分一种片段尺寸对照分子与其他片段尺寸对照分子的区域。标识符区域也用于区分来自一个子集的片段尺寸对照分子与来自另一个子集的片段尺寸对照分子。标识符区域可以具有分子条形码。标识符区域可以存在于片段尺寸区域的一侧或两侧。标识符区域可以是连续的或非连续的。分子条形码用作片段尺寸对照分子的标识符。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)。在一些实施方案中，标识符区域还可以具有另外的流动池结合位点，诸如P5和P7，其允许片段尺寸对照分子附接至下一代测序仪(例如，Illumina测序仪)的流动池表面。在一些实施方案中，标识符区域包含条形码，条形码用作(i)分子标识符条形码，例如，用于鉴定每个片段尺寸对照分子并区分一种片段尺寸对照分子与另一种片段尺寸对照分子的条形码，和(ii)子集标识符条形码，例如，用于鉴定片段尺寸对照分子所属子集(即，片段尺寸对照分子属于子集1还是子集2)的条形码。对于一个子集中的所有片段尺寸对照分子，子集标识符条形码可以是相同的，并且一个子集的子集标识符条形码可以不同于其他子集的子集标识符条形码。在一些实施方案中，一个样品中一个子集的子集标识符条形码可以不同于其他样品中相应子集的子集标识符条形码。例如，如果片段尺寸对照分子用于评价样品的污染，则一个子集的子集标识符条形码不同于在同一流动池/批次内的所有其它样品中使用的相应子集的子集标识符条形码。在一些实施方案中，标识符区域可以包含用于区分一个样品的片段尺寸对照分子与其他样品的片段尺寸对照分子的样品索引以及用于附接至测序仪的流动池的序列。例如，在富集步骤之后添加的片段尺寸对照分子的子集的标识符区域可以包含样品索引序列。在一些实施方案中，标识符区域可以通过连接被附接至片段尺寸区域。在一些实施方案中，标识符区域可以通过PCR被添加至片段尺寸区域。

[0123] 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。

[0124] 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。

[0125] 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。

[0126] 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、形式或处理的信息。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(分子条形码)也可以称为标识符(例如分子标识符、样品标识符)。另外地或可选地，核酸标签可以用作分子标识符(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，对特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)。

[0127] 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。

[0128] 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)。

[0129] 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。寡核苷酸的尺寸范围通常从若干单体单元例如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。字母A、C、G和T可以用于指碱基本身、核苷或包含碱基的核苷酸。

[0130] 处理：如本文使用的，“处理”是指用于生成适于测序的核酸文库的一组步骤。这组步骤可以包括但不限于核酸的分区、末端修复、添加测序衔接子、加标签和/或PCR扩增。

[0131] 经处理的样品：如本文使用的，“经处理的样品”是指已如本文别处所述进行处理的样品。在一些实施方案中，经处理的样品可以指从被处理的第一加标样品提取的核酸的子集，其中从第一加标样品中提取的核酸的子集包含来自无细胞多核苷酸样品和来自片段尺寸对照分子的第一子集的核酸分子。在其他实施方案中，经处理的样品是指被处理的第二加标样品的子集，其中被处理的第二加标样品的子集包含来自无细胞多核苷酸样品和来自片段尺寸对照分子的第一子集和第二子集的核酸分子。

[0132] 定量量度：如本文使用的，“定量量度”是指绝对量度或相对量度。定量量度可以是但不限于数字、统计量度(例如，频率、平均值、中位值、标准差或分位数)或程度或相对量(例如，高、中和低)。定量量度可以是两个定量量度的比率。定量量度可以是定量量度的线性组合。定量度量可以是归一化的度量。

[0133] 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。在一些实施方案中，参考序列可以是至少约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。在一些实施方案中，参考序列可以是完整基因组。参考序列的实例包括，例如，人类基因组，诸如，hG19和hG38。

[0134] 样品：如本文使用的，“样品”意指能够通过本文公开的方法和/或系统分析的任何事物。

[0135] 第二加标样品：如本文使用的，“第二加标样品”是其中片段尺寸对照分子的第二子集已被添加至从第一加标样品提取的核酸中的样品，其中从第一加标样品提取的核酸包含来自无细胞多核苷酸样品和来自片段尺寸对照分子的第一子集的核酸分子。

[0136] 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、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。

[0137] 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。

[0138] 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。

[0139] 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。

[0140] 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疾病，例如癌症、自身免疫疾病。

[0141] 第三加标样品：如本文使用的，“第三加标样品”是其中片段尺寸对照分子的第三子集已被添加至经处理的样品中的样品。

[0142] 详细描述

[0143] I.综述

[0144] 循环无细胞DNA中的片段尺寸分布和片段化模式可以提供关于无细胞DNA来源以及疾病水平的信息。然而，通过各种提取过程和用于读出或分析(例如，用于下一代测序的文库制备)片段长度和模式的方法，偏差可被引入片段尺寸分布和片段化模式中。因此，有可用于了解偏差来源的对照(偏差发生在过程的哪些步骤)以及使用这些对照来校正偏差或对样品间和批次水平的片段尺寸偏差进行归一化是至关重要的。

[0145] 本公开内容提供了用于将片段长度回收率归一化；优化测定以回收全部/大部分所需尺寸的片段并作为QC指标的方法、组合物和系统。这样的方法可以包括使用片段尺寸对照分子作为对照，其可以在不同的步骤，诸如从提取到测序，添加至无细胞多核苷酸样品。这些对照可以具有预定的片段长度，以模拟无细胞DNA的天然尺寸分布。在不同步骤添加的片段尺寸对照分子的身份可以通过使用不同的条形码来保持，以便人们可以保持追溯在特定步骤的片段尺寸偏差。这些对照分子可用于许多应用，诸如但不限于以下应用：(i)监测片段尺寸偏差，(ii)优化过程以减少片段尺寸偏差，(iii)通过分析片段尺寸对照分子的回收率对过程期间引入的片段尺寸偏差进行校正或归一化，(iv)作为基于片段尺寸对照分子的回收率来评估测定性能的QC指标，(v)作为分析任何样品污染的QC指标，和(vi)优化

测定以使任何污染最小化。

[0146] 癌症的形成和进展可由脱氧核糖核酸(DNA)的遗传修饰和表观遗传修饰两者引起。本公开内容提供了分析DNA诸如无细胞DNA(cfDNA)的表观遗传修饰的方法。这样的表观遗传学分析包括可通过测量片段长度分布的变化或片段终点映射至基因组座位的频率的变化来辨别的甲基化和DNA片段模式。这样的“片段组(fragmentome)”分析可单独使用或与现有技术组合使用,以确定疾病或状况的存在或不存在、诊断的疾病或状况的预后、诊断的疾病或状况的治疗处理或疾病或状况的预测治疗结果。这样的疾病或状况的一种实例是癌症。

[0147] 循环无细胞DNA(cfDNA)可以主要是从死亡组织细胞脱落到体液诸如外周血(血浆或血清)中的短DNA片段(例如,长度为约100个至400个碱基对,模式为约165bp)。除了癌症相关的遗传变异外,对cfDNA的分析还可以揭示死亡细胞的吞噬清除的表观遗传学足迹和特征,死亡细胞的吞噬清除可以导致当前恶性肿瘤(例如肿瘤)及其微环境组分的聚集核小体占据情况。

[0148] 可能对血浆片段组信号(例如,从分析cfDNA片段获得的信号)有贡献的组分或因素包括(i)细胞死亡类型和在DNA分解期间相关的染色质浓缩事件,(ii)清除机制,其可能涉及由受试者的免疫系统调节的各种类型的吞噬机制,(iii)血液组成的非恶性变化,其可能受循环中细胞类型的潜在组合影响,(iv)特定类型的器官或组织中非恶性细胞死亡的多于一种来源或原因,和(v)癌症中细胞类型的异质性,因为恶性实体瘤包括肿瘤相关的正常细胞、上皮细胞和基质细胞、免疫细胞和血管细胞,所有这些细胞中的任何一种可能对cfDNA样品(例如,其可以从受试者的体液获得)有贡献并在cfDNA样品中得以代表。

[0149] 组蛋白保护复合物形式的无细胞DNA可以由各种宿主细胞(包括中性粒细胞、巨噬细胞、嗜酸性粒细胞以及肿瘤细胞)释放。循环DNA通常具有短的半衰期(例如,约10分钟至15分钟),并且肝通常是从血液循环去除循环DNA片段的主要器官。循环中cfDNA的积累可以由细胞死亡和/或活化的增加、cfDNA清除的受损和/或内源性DNA酶水平的降低引起。在受试者血流中循环的无细胞DNA通常可被包装成膜包被结构(例如,凋亡小体)或与生物聚合物(例如,组蛋白或结合DNA的血浆蛋白)的复合物。可以分析DNA片段化和随后转运的过程对由片段组分析检测到的无细胞DNA信号特征的影响。

[0150] 在细胞核(例如,人类的细胞核)中,DNA通常存在于核小体中,核小体被组织成包含约145个碱基对(bp)的DNA缠绕在核心组蛋白八聚体周围的结构。DNA和核小体之间的静电和氢键相互作用可以导致DNA在蛋白表面发生能量上不利的弯曲。这样的弯曲可以在空间上抑制其他DNA结合蛋白,并且因此可以用于调节细胞核中对DNA的接近。核小体在细胞中的定位可以动态波动(例如,随着时间的推移以及不同的细胞状态和条件),例如,DNA解缠绕和重新缠绕。由于片段组信号可以反映源自受核小体单位影响的构型的核小体保护的DNA片段,因此核小体的稳定性和动态可以影响这样的片段组信号。这些核小体动态可以源于各种因素,诸如:(i)依赖于ATP的重塑复合物,其可以利用ATP水解的能量使核小体滑动,并从染色质纤维交换或驱逐组蛋白,(ii)组蛋白变体,其可以具有不同于典型组蛋白的特性,并在染色质纤维中产生定位的特定结构域,(iii)组蛋白伴侣,其可以控制游离组蛋白的供应,并在组蛋白沉积和驱逐中与染色质重塑物(chromatin remodeler)合作,(iv)组蛋白的翻译后修饰(PTM)(例如,乙酰化、甲基化、磷酸化和泛素化),其可直接或间接影响染色

质结构，和(v)转录因子和RNA聚合酶的活化转录。

[0151] 因此，cfDNA的片段化信号或模式可以指示源于与遍及基因组的染色质组织异质性相关的多种事件的聚集cfDNA信号。这样的染色质组织可取决于诸如整体细胞身份、代谢状态、区域调节状态、死亡细胞中的局部基因活性和DNA清除机制等因素而异。此外，无细胞DNA片段组信号可以仅部分地归因于贡献细胞的潜在染色质结构。这样的cfDNA片段组信号可以指示细胞死亡和保护DNA免受酶消化期间染色质压缩的更复杂足迹。因此，由于在细胞死亡或碎片运输的不同阶段核小体稳定性、构象和组成的变化，特定细胞类型或细胞谱系类型特异的染色质图谱可以仅部分地对DNA可及性的内在异质性有贡献。因此，一些核小体可以变得优先存在或不存在于无细胞DNA中(例如，可能存在影响cfDNA清除和释放到血液循环中的过滤机制)，这可以取决于诸如死亡和细胞尸体清除的模式和机制等因素。

[0152] 在细胞过程诸如细胞凋亡和坏死期间，由于细胞核DNA片段化，细胞中可以生成片段组信号，并作为cfDNA释放到血液循环中。这样的片段化可能由于在不同细胞阶段不同的核酸酶作用于DNA而产生，导致序列特异性的DNA裂解模式，这种模式可以在cfDNA片段组信号中进行分析。这样的裂解模式分类可以是细胞环境(例如，肿瘤微环境、炎症、疾病状态、肿瘤发生等)的临床相关标志物。

[0153] 本公开内容提供了用于评价和校正多核苷酸样品(在一些实施方案中，多核苷酸可以是无细胞多核苷酸)的核酸分子的分析中的片段尺寸偏差的方法、组合物和系统。这些方法可用于各种应用，诸如疾病的预后、诊断和/或监测。

[0154] 多核苷酸样品中的核酸分子的分析可以通过测量片段尺寸对照分子的回收率来优化和校正片段尺寸偏差。片段尺寸对照分子可以是具有预定片段长度的合成的核酸分子。片段尺寸对照分子的集合可以包括包含片段尺寸对照分子的至少一个子集的核酸分子，该片段尺寸对照分子的至少一个子集在特定步骤被添加至多核苷酸样品。在一些实施方案中，片段尺寸对照分子的集合可以包括片段尺寸对照分子的两个或更多个子集，并且每个子集在测定的不同步骤被添加。每个子集包括一组或更多组片段尺寸对照分子，并且每个组具有不同的片段长度或尺寸。例如，如果在必须分析片段尺寸偏差的测定中有三个步骤，则片段尺寸对照分子的集合可以包括片段尺寸对照分子的三个子集(S1、S2和S3)，并且在每个步骤之前将每个子集添加至样品(即，S1在步骤1之前添加，S2在步骤2之前添加，并且S3在步骤3之前添加)。

[0155] 在一些实施方案中，片段尺寸对照分子可以是合成的核酸分子。在一些实施方案中，通过扩增任何基因组、质粒或载体或其一部分中的特定区域(即，具有特定长度)而生成的扩增子可以用作片段尺寸对照分子。在分析之前，片段尺寸对照分子的序列和长度可以是已经已知的。在一些实施方案中，片段尺寸对照分子被设计成使得这些分子不形成任何二级结构。在一些实施方案中，片段尺寸对照分子的序列被设计成使得它们不与任何人类基因组区域重叠。因此，在一些实施方案中，通过将片段尺寸对照分子添加至多核苷酸样品，并通过追溯子集中的片段尺寸对照分子，人们可以分析片段尺寸对照分子的回收率，并且从而估计片段尺寸偏差。

[0156] 因此，在一方面，本公开内容提供了一种分析多核苷酸样品中的核酸分子的方法，该方法包括：(a)将片段尺寸对照分子的子集添加至多核苷酸样品中的核酸分子，从而产生第一加标样品；(b)从第一加标样品提取核酸；(c)处理提取的核酸的至少一个子集，从而产

生经处理的样品，其中所述处理包括将第一加标样品的至少一个子集分区、加标签和/或扩增；(d)富集经处理的样品的至少一个子集，从而产生富集的样品；(e)对富集的样品的至少一个子集测序以生成多于一个序列读段；和(f)分析多于一个序列读段以生成片段尺寸对照分子的子集的多于一个片段尺寸评分。在一些实施方案中，多核苷酸样品可以是无细胞多核苷酸(例如，无细胞DNA)的样品。

[0157] 在一些实施方案中，该方法还包括将多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较。在一些实施方案中，该方法可用于基于多于一个片段尺寸评分优化多核苷酸样品中的核酸分子的分析。在一些实施方案中，该方法可用于使用多于一个片段尺寸评分来校正多核苷酸样品中的核酸分子的分析中的片段尺寸偏差。在一些实施方案中，该方法可以用作质量控制(QC)指标，其中(i)如果多于一个片段尺寸评分中的至少一个在多于一个片段尺寸阈值的相应片段尺寸阈值内，则将该方法分类为成功；或者(ii)如果多于一个片段尺寸评分中的至少一个不在多于一个片段尺寸阈值的相应片段尺寸阈值内，则将该方法分类为不成功。

[0158] 图1A是根据本公开内容的一种实施方案，用于分析无细胞多核苷酸样品中的核酸分子的方法的示意图。在一些实施方案中，片段尺寸对照分子的集合可以包括预定的片段尺寸对照分子的至少一个子集。在一些实施方案中，片段尺寸对照分子的集合可以包括片段尺寸对照分子的一个子集。在101A，将片段尺寸对照分子的子集(子集1)添加至可以分析其片段尺寸偏差的多核苷酸的样品，以便在提取多核苷酸之前生成第一加标样品。在一些实施方案中，片段尺寸对照分子被用一个或更多个标签或分子条形码加标签，这些标签或分子条形码可以帮助鉴定子集中的每个单独的片段尺寸对照分子(分子标识符)以及它所属的子集，即，子集1(子集标识符)。

[0159] 在一些实施方案中，片段尺寸对照分子的子集可以包括一组或更多组片段尺寸对照分子。在一些实施方案中，一组或更多组片段尺寸对照分子可以包含不同长度和/或不同序列的核酸分子。在一些实施方案中，一组片段尺寸对照分子可以包含相同长度且相同序列的片段尺寸对照分子。在一些实施方案中，每个组可以包含相同长度但不同序列的片段尺寸对照分子。

[0160] 在102A，提取第一加标样品的核酸分子。在103A，处理提取的核酸分子以生成适于测序的核酸分子文库，包括诸如但不限于以下的步骤：末端修复、添加测序衔接子、加标签和/或PCR扩增。在一些实施方案中，处理提取的核酸分子以生成核酸分子文库包括诸如但不限于以下的步骤：分区、末端修复、添加测序衔接子、加标签、洗涤/清洁和/或PCR扩增。在104A，富集经处理的样品中的核酸中的(i)无细胞多核苷酸样品中属于特定感兴趣区域的核酸分子和(ii)子集1的片段尺寸对照分子。在105A，对富集的样品测序，从而可以分析子集1中片段尺寸对照分子的回收率。在106A，分析从测序仪生成的序列读段，以测量子集1片段尺寸对照分子的回收率。在一些实施方案中，可以利用添加至样品的特定子集中的片段尺寸对照分子的相对丰度的正交测量结果来计算片段尺寸对照分子的回收率。在一些实施方案中，正交测量结果可以从凝胶电泳、qPCR、ddPCR或无PCR测序获得。

[0161] 图1B示出了用于分析多核苷酸样品(例如，样品可以是无细胞DNA样品)中的核酸分子的方法100B的示例性实施方案。在101B，将片段尺寸对照分子的子集(子集1)添加至多核苷酸样品，以便在提取步骤之前生成第一加标样品。将片段尺寸对照分子添加至样品，以

便监测可能在任何步骤中引入的片段尺寸偏差。在一些实施方案中，使用片段尺寸对照分子估计的片段尺寸偏差可用于优化样品中多核苷酸的计算的片段长度分布。在一些实施方案中，片段尺寸对照分子的集合可以包括预定的片段尺寸对照分子的至少一个子集。在一些实施方案中，片段尺寸对照分子的集合可以包括片段尺寸对照分子的一个子集(例如，子集1)。在一些实施方案中，每个子集包括至少一组片段尺寸对照分子。在一些实施方案中，每个组包含相同长度且相同序列的片段尺寸对照分子。在一些实施方案中，一个组中的片段尺寸对照分子的长度不同于其他组中的片段尺寸对照分子的长度。在一些实施方案中，每个组包含相同长度但不同序列的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子被用一个或更多个标签或分子条形码加标签，这些标签或分子条形码可以帮助鉴定子集中的每个单独的片段尺寸对照分子(分子标识符)以及它所属的子集，即，子集1(子集标识符)。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于等摩尔浓度。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于非等摩尔浓度。

[0162] 在102B，提取第一加标样品的核酸分子。在103B，处理提取的核酸以生成经处理的样品，其中经处理的样品包含(i)无细胞多核苷酸样品中的核酸分子和(ii)子集1的片段尺寸对照分子。

[0163] 核酸分子在103B被处理以生成适于测序(例如，通过下一代测序仪)的核酸文库。处理可以包括诸如但不限于以下的步骤：末端修复、添加测序衔接子、加标签、洗涤/清洁和/或核酸扩增。在一些实施方案中，处理可以包括诸如但不限于以下的步骤：分区、末端修复、添加测序衔接子、加标签、洗涤/清洁和/或核酸扩增。

[0164] 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化学修饰(例如，甲基化)的核苷酸的核酸分子。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：

[0165] (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；

[0166] (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；

[0167] (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FOXK1、FOXK2、FOXP1、FOXP4和FOXI3(例如，如Iurlaro等人，Genome Biol.14,R119(2013)描述的，在此将其通过引用以其整体并入)；和

[0168] (d)对一个或更多个甲基化核苷酸碱基特异的抗体。

[0169] 分区可以指基于核酸分子的特征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(例如，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，在此将该申请通过引用以其整体并入。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)将核酸分区。在一些实施方案中，可以将核酸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在该修饰中是过度代表性的，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。亲和分离的作用是富集结合相中修饰被过度代表的核酸和非结合相(即，溶液中)中修饰未被充分代表的核酸。结合相的核酸可以在后续处理之前洗脱。在一些实施方案中，多于一个分区的集合中的每一个被差异性加标签(differentially tagged)。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、富集和/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。标签可以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中的分子。

[0170] 在104B，经处理的样品被富集以生成富集的样品。富集的样品包含(i)无细胞多核苷酸样品中属于特定感兴趣区域的核酸分子和(ii)子集1的片段尺寸对照分子。在105B，对富集的样品测序以生成多于一个序列读段，从而可以分析子集1中片段尺寸对照分子的回收率，以便计算片段尺寸偏差。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。根据附接至片段尺寸对照分子的标签的序列，人们可以将标签与单个片段尺寸对照分子和片段尺寸对照分子所属的子集相关联。该信息用于分析子集中片段尺寸对照分子的回收率。

[0171] 在106B，分析序列读段以生成片段尺寸对照分子的片段尺寸评分。片段尺寸评分代表属于特定组和特定子集的片段尺寸对照分子的回收率。片段尺寸对照分子的身份和片段尺寸对照分子所属的子集通过使用标签或条形码来保持。在一些实施方案中，片段尺寸评分可以基于属于特定组和特定子集的片段尺寸对照分子的数目来估计。在一些实施方案中，片段尺寸评分可以基于属于特定组和特定子集的片段尺寸对照分子的测序读段的数目来估计。在一些实施方案中，可以为特定子集中的特定组估计片段尺寸评分。在一些实施方案中，可以为每个子集中的每个组估计片段尺寸评分。在一些实施方案中，片段尺寸评分可以是在特定子集中使用的片段尺寸对照分子的总评分(即，代表子集中所有片段尺寸对照分子的单个评分)，或者是在测定的一个或更多个步骤中添加的片段尺寸对照分子的总评分(即，代表在测定中添加的一个或更多个子集中的片段尺寸对照分子的单个评分)。在一些实施方案中，片段尺寸评分可以被测量为在特定步骤添加的片段尺寸对照分子的子集的量与在不同步骤添加的片段尺寸对照分子的另一子集的量的差异，并且在一些实施方案中，该量可以是质量(例如，fg、pg、ng、μg)、测序读段数目或片段尺寸分子(属于一个子集或属于特定组)的数目。在一些实施方案中，片段尺寸评分可以被测量为(属于一个子集或属于特定组)的片段尺寸对照分子的回收率(分数或百分比)，其中回收率可以利用添加至样品的特定子集中的片段尺寸对照分子的相对丰度的正交测量结果来计算。在一些实施方案中，正交测量结果可以从凝胶电泳、qPCR、ddPCR或无PCR测序获得。在一些实施方案中，片段尺寸评分可以被测量为具有特定长度的片段尺寸对照分子的量与具有不同长度的片段尺

寸对照分子的量的差异。

[0172] 在107B,将片段尺寸评分与相应的片段尺寸阈值进行比较,以便确定测定中的片段尺寸偏差。片段尺寸阈值是用于评价或优化样品中的核酸分子的分析中的片段尺寸偏差的预定的阈值或范围。这些阈值也可用于在任何步骤(诸如提取、文库制备、富集、洗涤/清洁和测序)中校正任何片段长度分布。在一些实施方案中,可以为特定子集中的特定组估计片段尺寸阈值。在一些实施方案中,子集中的每个组具有特定的片段尺寸阈值。在一些实施方案中,可以为每个子集中的每个组估计片段尺寸阈值。在一些实施方案中,片段尺寸阈值可以是在特定子集中使用的片段尺寸对照分子的总阈值(即,子集中所有片段尺寸对照分子的单个阈值),或者是在测定的一个或更多个步骤中添加的片段尺寸对照分子的总阈值(即,在测定中添加的一个或更多个子集中的片段尺寸对照分子的单个阈值)。例如,片段尺寸对照分子的集合包括两个子集—子集1和子集2。如果每个子集包括两个组—G11和G12(对于子集1)以及G21和G22(对于子集2),则基于片段尺寸对照分子的回收率,这四个组中的每个组可以具有片段尺寸评分(S)—G11的S11、G12的S12、G21的S21和G22的S22。子集中的每个组可以具有单独的预定的片段尺寸阈值。在该实例中,T11、T12、T21和T22分别是G11、G12、G21和G22的片段尺寸阈值。阈值可以是百分比或分数,并且阈值可以是阈值范围,而不是特定的阈值。为了使该方法被认为是成功的,这些片段尺寸评分中的至少一个应该在其相应的片段尺寸阈值内。在一些实施方案中,如果多于一个片段尺寸评分中的每一个在相应的片段尺寸阈值内,则该方法被分类为成功进行。在一些实施方案中,片段尺寸阈值可以是百分比或分数。在一些实施方案中,片段尺寸阈值可以是阈值范围,而不是特定的阈值。

[0173] 在一些实施方案中,在103B之前,可以将片段尺寸对照分子的第二子集(子集2)添加至提取的核酸以生成第二加标样品。在这些实施方案中,处理第二加标样品以生成经处理的样品。经处理的样品包含(i)无细胞多核苷酸样品中的核酸分子和(ii)子集1和子集2的片段尺寸对照分子。在一些实施方案中,在104B之前,可以将片段尺寸对照分子的第三子集(子集3)添加至经处理的样品以生成第三加标样品。在这些实施方案中,富集第三加标样品以生成富集的样品。富集的样品包含(i)无细胞多核苷酸样品中属于特定感兴趣区域的核酸分子和(ii)子集1、子集2和子集3的片段尺寸对照分子。在一些实施方案中,在105B之前,可以将片段尺寸对照分子的第四子集(子集4)添加至富集的样品以生成第四加标样品。第四加标样品包含(i)无细胞多核苷酸样品中属于特定感兴趣区域的核酸分子和(ii)子集1、子集2、子集3和子集4的片段尺寸对照分子。在这些实施方案中,对第四加标样品测序以生成多于一个多个序列读段。

[0174] 在另一方面,本公开内容提供了一种用于分析多核苷酸样品中的核酸分子的方法,该方法包括:(a)将片段尺寸对照分子的第一子集添加至多核苷酸样品中的核酸分子,从而产生第一加标样品;(b)从第一加标样品提取核酸;(c)将片段尺寸对照分子的第二子集添加至提取的核酸,从而产生第二加标样品;(d)处理第二加标样品的至少一个子集,从而产生经处理的样品,其中所述处理包括将第二加标样品的所述至少一个子集分区、加标签和/或扩增;(e)将片段尺寸对照分子的第三子集添加至经处理的样品,从而产生第三加标样品;(f)富集第三加标样品的至少一个子集,从而产生富集的样品;(g)将片段尺寸对照分子的第四子集添加至富集的样品的至少一个子集,从而产生第四加标样品;(h)对第四加

标样品测序以生成多于一个序列读段;(i)分析多于一个序列读段以生成片段尺寸对照分子的第一子集、片段尺寸对照分子的第二子集、片段尺寸对照分子的第三子集和/或片段尺寸对照分子的第四子集的多于一个片段尺寸评分;和(j)将多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较。

[0175] 在一些实施方案中,该方法还包括将多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较。在一些实施方案中,该方法可用于基于多于一个片段尺寸评分优化多核苷酸样品中的核酸分子的分析。在一些实施方案中,该方法可用于使用多于一个片段尺寸评分来校正多核苷酸样品中的核酸分子的分析中的片段尺寸偏差。在一些实施方案中,该方法可以用作质量控制(QC)指标,其中(i)如果多于一个片段尺寸评分中的至少一个在多于一个片段尺寸阈值的相应片段尺寸阈值内,则将该方法分类为成功;或者(ii)如果多于一个片段尺寸评分中的至少一个不在多于一个片段尺寸阈值的相应片段尺寸阈值内,则将该方法分类为不成功。在一些实施方案中,多核苷酸样品可以是无细胞多核苷酸(例如,无细胞DNA)的样品。

[0176] 图2A是根据本公开内容的一种实施方案,用于分析多核苷酸(例如,无细胞多核苷酸)样品中的核酸分子的方法的示意图。在一些实施方案中,片段尺寸对照分子的集合可以包括预定的片段尺寸对照分子的至少一个子集。在201A,将片段尺寸对照分子的第一子集(子集1)添加至必须分析其片段尺寸偏差的多核苷酸的样品,以便在提取多核苷酸之前生成第一加标样品。在一些实施方案中,片段尺寸对照分子被用一个或更多个标签或条形码加标签,这些标签或条形码可以帮助鉴定子集中的每个单独的片段尺寸对照分子(分子标识符)以及它所属的子集,即子集1(子集标识符)。

[0177] 在一些实施方案中,片段尺寸对照分子的子集可以包括一组或更多组片段尺寸对照分子。在一些实施方案中,一组或更多组片段尺寸对照分子可以包含不同长度和/或不同序列的核酸分子。在一些实施方案中,一组片段尺寸对照分子可以包含相同长度且相同序列的片段尺寸对照分子。在一些实施方案中,每个组可以包含相同长度但不同序列的片段尺寸对照分子。

[0178] 在202A,提取第一加标样品的无细胞核酸分子。在203A,将片段尺寸对照分子的第二子集(子集2)添加至提取的核酸分子以生成第二加标样品。在204A,处理第二加标样品以生成适于测序的核酸分子文库,包括诸如但不限于以下的步骤:末端修复、添加测序衔接子、加标签、洗涤/清洁和/或PCR扩增。在一些实施方案中,处理第一加标样品以生成核酸分子文库包括诸如但不限于以下的步骤:分区、末端修复、添加测序衔接子、加标签、洗涤/清洁和/或PCR扩增。在205A,将片段尺寸对照分子的第三子集(子集3)添加至经处理的样品以生成第三加标样品。在206A,富集第三加标样品中的核酸中的(i)无细胞多核苷酸样品中属于特定感兴趣区域的核酸分子和(ii)子集1、子集2和子集3的片段尺寸对照分子。在207A,将片段尺寸对照分子的第四子集(子集4)添加至富集的样品以生成第四加标样品。在208A,对第四加标样品测序,从而可以分析子集1、子集2、子集3和子集4中片段尺寸对照分子的回收率,以便确定片段尺寸偏差。在209A,分析从测序仪生成的序列读段,以测量子集1、子集2、子集3和子集4片段尺寸对照分子的回收率。

[0179] 图2B示出了用于分析多核苷酸(例如,无细胞多核苷酸)样品中的核酸分子的方法200B的示例性实施方案。在201B,将片段尺寸对照分子的子集(子集1)添加至无细胞多核苷

酸样品，以便在提取步骤之前生成第一加标样品。将片段尺寸对照分子添加至样品，以便监测和/或校正在任何步骤中引入的片段尺寸偏差。在一些实施方案中，片段尺寸对照分子的集合可以包括预定的片段尺寸对照分子的至少一个子集。在一些实施方案中，片段尺寸对照分子的集合可以包括片段尺寸对照分子的一个子集(例如，子集1)。在一些实施方案中，每个子集包括至少一组片段尺寸对照分子。在一些实施方案中，每个组包含相同长度且相同序列的片段尺寸对照分子。在一些实施方案中，一个组中的片段尺寸对照分子的长度不同于其他组中的片段尺寸对照分子的长度。在一些实施方案中，每个组包含相同长度但不同序列的片段尺寸对照分子。在一些实施方案中，片段尺寸对照分子被用一个或更多个标签或条形码加标签，这些标签或条形码可以帮助鉴定子集中的每个单独的片段尺寸对照分子(分子标识符)以及它所属的子集，即，子集1(子集标识符)。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于等摩尔浓度。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于非等摩尔浓度。

[0180] 在202B，提取第一加标样品的核酸分子。在203B，将片段尺寸对照分子的第二子集(子集2)添加至提取的核酸分子以生成第二加标样品。在一些实施方案中，通过使用不同的标签，即子集1中的片段尺寸对照分子的标签(子集标识符)不同于子集2中的片段尺寸对照分子的标签，区分第一子集中的片段尺寸对照分子与第二子集中的片段尺寸对照分子。例如，子集1中的所有片段尺寸对照分子可以具有子集标识符标签(S1)，并且子集2中的所有片段尺寸对照分子可以具有子集标识符标签(S2)。除了子集标识符标签，每个片段尺寸对照分子包含用于识别每个单独片段尺寸对照分子的分子标识符标签。

[0181] 在204B，处理第二加标样品以生成经处理的样品，其中经处理的样品包含(i)无细胞多核苷酸样品中的核酸分子和(ii)子集1和子集2的片段尺寸对照分子。第二加标样品在204B被处理以生成适于测序(例如，通过下一代测序仪)的核酸文库。处理可以包括诸如但不限于以下的步骤：末端修复、添加测序衔接子、加标签、洗涤/清洁和/或核酸扩增。

[0182] 在一些实施方案中，处理可以包括诸如但不限于以下的步骤：分区、末端修复、添加测序衔接子、加标签、洗涤/清洁和/或核酸扩增。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化学修饰(例如，甲基化)的核苷酸的核酸分子。在一些实施方案中，可以将核酸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，多于一个分区的集合中的每一个被差异性加标签，使得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的集合。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、富集和/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。在一些实施方案中，标签可以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。

[0183] 在一些实施方案中，加标签包括将标签的集合附接至核酸以产生加标签的核酸的群体，其中加标签的核酸包含一个或更多个标签。在一些实施方案中，标签的集合通过将衔接子连接至核酸而附接至该核酸，其中衔接子包含一个或更多个标签。

[0184] 在205B,将片段尺寸对照分子的第三子集(子集3)添加至经处理的样品以生成第三加标样品。在206B,富集第三加标样品以生成富集的样品。富集的样品包含(i)无细胞多核苷酸样品中属于特定感兴趣区域的核酸分子和(ii)子集1、子集2和子集3的片段尺寸对照分子。在207B,将片段尺寸对照分子的第四子集(子集4)添加至富集的样品以生成第四加标样品。在208B,对第四加标样品测序以生成多于一个序列读段,从而可以分析子集1、子集2、子集3和子集4中片段尺寸对照分子的回收率,以便计算片段尺寸偏差。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。根据附接至片段尺寸对照分子的标签的序列,人们可以将标签与单个片段尺寸对照分子和该分子所属的子集相关联。该信息用于分析子集中片段尺寸对照分子的回收率。

[0185] 在209B,分析序列读段以生成片段尺寸对照分子的片段尺寸评分。片段尺寸评分代表属于特定组和特定子集的片段尺寸对照分子的回收率。在一些实施方案中,可以利用添加至样品的特定子集中的片段尺寸对照分子的相对丰度的正交测量结果来计算片段尺寸对照分子的回收率。在一些实施方案中,正交测量结果可以从凝胶电泳、qPCR、ddPCR或无PCR测序获得。片段尺寸对照分子的身份以及片段尺寸对照分子所属的组和子集通过使用标签或条形码来保持。在一些实施方案中,片段尺寸评分可以基于属于特定组和特定子集的片段尺寸对照分子的数目来估计。在一些实施方案中,片段尺寸评分可以基于属于特定组和特定子集的片段尺寸对照分子的测序读段的数目来估计。在一些实施方案中,可以为特定子集中的特定组估计片段尺寸评分。在一些实施方案中,可以为每个子集中的每个组估计片段尺寸评分。在一些实施方案中,片段尺寸评分可以是在特定子集中使用的片段尺寸对照分子的总评分(即,代表子集中的片段尺寸对照分子的单个评分),或者是在测定的一个或更多个步骤中添加的片段尺寸对照分子的总评分(即,代表在测定中添加的一个或更多个子集中的片段尺寸对照分子的单个评分)。在一些实施方案中,片段尺寸评分可以被测量为在特定步骤添加的片段尺寸对照分子的子集的量与在不同步骤添加的片段尺寸对照分子的另一子集的量的差异,并且在一些实施方案中,该量可以是质量(例如,fg、pg、ng、μg)、测序读段数目或片段尺寸分子(属于一个子集或属于特定组)的数目。在一些实施方案中,片段尺寸评分可以被测量为(属于一个子集或属于特定组)的片段尺寸对照分子的回收率(分数或百分比),其中回收率可以利用添加至样品的特定子集中的片段尺寸对照分子的相对丰度的正交测量结果来计算。在一些实施方案中,正交测量结果可以从凝胶电泳、qPCR、ddPCR或无PCR测序获得。在一些实施方案中,片段尺寸评分可以被测量为具有特定长度的片段尺寸对照分子的量与具有不同长度的片段尺寸对照分子的量的差异。

[0186] 在210B,将片段尺寸评分与相应的片段尺寸阈值进行比较,以便确定测定中的片段尺寸偏差。片段尺寸阈值是用于评价或优化样品中的核酸分子的分析中的片段尺寸偏差的预定的阈值或范围。这些阈值也可用于在任何步骤(诸如提取、文库制备、富集、洗涤/清洁和测序)中校正任何片段长度分布。在一些实施方案中,可以为特定子集中的特定组估计片段尺寸阈值。在一些实施方案中,子集中的每个组具有特定的片段尺寸阈值。在一些实施方案中,可以为每个子集中的每个组估计片段尺寸阈值。在一些实施方案中,片段尺寸阈值可以是在特定子集中使用的片段尺寸对照分子的总阈值(即,子集中的片段尺寸对照分子的单个阈值),或者是在测定的一个或更多个步骤中添加的片段尺寸对照分子的总阈值(即,在测定中添加的一个或更多个子集中的片段尺寸对照分子的单个阈值)。例如,片段尺

寸对照分子的集合包括两个子集—子集1和子集2。如果每个子集包括两个组—G11和G12(对于子集1)以及G21和G22(对于子集2),则基于片段尺寸对照分子的回收率,这四个组中的每个组可以具有片段尺寸评分(S)—G11的S11、G12的S12、G21的S21和G22的S22。子集中的每个组可以具有单独的预定的片段尺寸阈值。在该实例中,T11、T12、T21和T22分别是G11、G12、G21和G22的片段尺寸阈值。阈值可以是百分比或分数,并且阈值可以是阈值范围,而不是特定的阈值。为了使该方法被认为是成功的,这些片段尺寸评分中的至少一个应该在其相应的片段尺寸阈值内。在一些实施方案中,如果多于一个片段尺寸评分中的每一个在相应的片段尺寸阈值内,则该方法被分类为成功进行。在一些实施方案中,片段尺寸阈值可以是百分比或分数。在一些实施方案中,片段尺寸阈值可以是阈值范围,而不是特定的阈值。

[0187] 在一些实施方案中,使用片段尺寸对照分子估计的核酸分子的片段尺寸偏差的分析可用于优化样品中的多核苷酸的计算的片段长度分布。在一些实施方案中,多于一个片段尺寸评分用于校正核酸分子的分析中的片段尺寸偏差。在一些实施方案中,校正片段尺寸偏差包括使用从子集中的组的片段尺寸评分与片段尺寸阈值的比率导出的定量量度和/或使用从子集中的两个组的片段尺寸评分的比率和/或两个子集的片段尺寸评分的比率导出的定量量度。在一些实施方案中,(i)如果多于一个片段尺寸评分中的至少一个在多于一个片段尺寸阈值的相应片段尺寸阈值内,则将该方法分类为成功;或者(ii)如果多于一个片段尺寸评分中的至少一个不在多于一个片段尺寸阈值的相应片段尺寸阈值内,则将该方法分类为不成功。

[0188] 在一些实施方案中,如果所有子集中所有组的片段尺寸对照分子的片段尺寸评分在所有组的相应片段尺寸阈值内,则分析核酸分子的方法可被分类为成功。否则,如果任何一个片段尺寸评分在其相应的片段尺寸阈值之外,则分析核酸分子的方法可被分类为不成功。例如,片段尺寸对照分子的集合包括片段尺寸对照分子的两个子集—子集1和子集2。如果每个子集包括两个组—G11和G12(对于子集1)以及G21和G22(对于子集2),则基于片段尺寸对照分子的回收率,这四个组中的每个组可以具有片段尺寸评分(S)—G11的S11、G12的S12、G21的S21和G22的S22。子集中的每个组具有单独的片段尺寸阈值。在该实例中,T11、T12、T21和T22分别是G11、G12、G21和G22的片段尺寸阈值。如果S11<T11,S12<T12,S21<T21,并且S22<T22,则该方法可被分类为成功。如果任何一个片段尺寸评分在其对应的片段尺寸阈值之外(即,不在其内),则该方法可被分类为不成功。

[0189] 在一些实施方案中,片段尺寸对照分子可用于单个样品中,以校正由技术(人为偏差)而不是真实的生物学原因导致的观察到的片段长度分布中的片段尺寸偏差。例如,由于源自肿瘤细胞的cfDNA分子通常表现出比源自造血细胞的片段更短的平均长度,因此,片段长度分布的汇总统计数据,包括但不限于平均值、中位值、模式和IQR,可以单独或与其他特征组合用作检测恶性肿瘤的特征。然而,这些汇总统计数据可以受技术因素的影响—在装运、样品和液体处理以及任何步骤(例如,提取、分区、加标签、扩增、洗涤/清洁、富集和测序)期间引入的人为偏差,并且这些影响可以混淆恶性肿瘤的检测。为了避免这种潜在的混淆来源,可以基于不同长度的片段尺寸对照分子的相对回收率来调整观察到的片段长度分布。例如,如果特定长度的片段尺寸对照分子的预期回收率为75%,但是在特定样品中该特定长度的片段尺寸对照分子的观察回收率仅为25%,则对应于该片段尺寸对照分子长度的

样品多核苷酸的片段长度分布的组分密度可以增加三倍以校正低回收率。相反，如果特定长度的片段尺寸对照分子的预期回收率为60％，但是在特定样品中该特定长度的片段尺寸对照分子的观察回收率为80％，则对应于该片段尺寸对照分子长度的样品多核苷酸的片段长度分布密度可以减少25％。

[0190] 对观察到的片段长度分布的调整可以发生在回收的片段尺寸对照分子的长度周围的窗中。例如，如果片段尺寸对照分子具有200个碱基对的长度，则可以调整长度在180bp和220bp之间或170bp和230bp之间的样品多核苷酸的观察到的长度分布密度。窗不必是对称的，并且对于落入窗内的每个多核苷酸长度，所应用的调整不必是相同的。例如，所应用的调整可以使用以片段尺寸对照分子的长度为中心的高斯核(Gaussian kernel)，而不是均匀核(uniform kernel)。

[0191] 不同长度的片段尺寸对照分子的预期回收率可以通过进行一个或更多个对照实验来确定。可以为任何特定长度的片段尺寸对照分子的回收率设定片段尺寸阈值。如果片段尺寸对照分子的观察到的回收率低于片段尺寸阈值或不在片段尺寸阈值范围内，则可认为样品的质量控制失败。对于不同长度的片段尺寸对照分子，这些片段尺寸阈值可以不同。

[0192] 在另一方面，本公开内容提供了一种用于产生无细胞多核苷酸样品的测序文库的方法，该方法包括：(a)将片段尺寸对照分子的子集添加至样品，从而产生第一加标样品；(b)从第一加标样品提取核酸；(c)处理提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将第一加标样品的至少一个子集分区、加标签和/或扩增；和(d)富集经处理的样品的至少一个子集。在一些实施方案中，该方法还包括，在处理之前，添加片段尺寸对照分子的第二子集，从而产生第二加标样品。在一些实施方案中，该方法还包括，在富集之前，添加片段尺寸对照分子的第三子集，从而产生第三加标样品。在一些实施方案中，该方法还包括，e)添加片段尺寸对照分子的第四子集，从而产生第四加标样品。

[0193] 在另一方面，本公开内容提供了一种用于检测第一样品被第二样品污染的方法，该方法包括，对于每个第一样品和第二样品：(a)添加片段尺寸对照分子的子集以生成第一加标样品，其中添加至第一样品的片段尺寸对照分子的子集可以与添加至第二样品的片段尺寸对照分子的子集区分；(b)从第一加标样品提取核酸；(c)处理提取的核酸的至少一个子集，从而产生经处理的样品，其中所述处理包括将第一加标样品的至少一个子集分区、加标签和/或扩增；(d)富集经处理的样品的至少一个子集；(e)对富集的样品的至少一个子集测序以生成多于一个序列读段；和(f)分析多于一个序列读段以生成片段尺寸对照分子的子集的一个或更多个污染评分。在一些实施方案中，该方法还包括，在处理之前，添加片段尺寸对照分子的第二子集，从而产生第二加标样品，其中添加至第一样品的片段尺寸对照分子的子集可以与添加至第二样品的片段尺寸对照分子的子集区分。在一些实施方案中，该方法还包括，在富集之前，添加片段尺寸对照分子的第三子集，从而产生第三加标样品，其中添加至第一样品的片段尺寸对照分子的子集可以与添加至第二样品的片段尺寸对照分子的子集区分。在一些实施方案中，该方法还包括，在测序之前，添加片段尺寸对照分子的第四子集，从而产生第四加标样品，其中添加至第一样品的片段尺寸对照分子的子集可以与添加至第二样品的片段尺寸对照分子的子集区分。在一些实施方案中，通过在第一样品中使用不同于其他样品中使用的子集标识符条形码的子集标识符条形码，可以区分添加至第一样品的片段尺寸对照分子的子集与添加至第二样品的片段尺寸对照分子的子集。

[0194] 污染评分是指第一样品中代表添加至第二样品的片段尺寸对照分子的存在的评分。在一些实施方案中，在一个样品的一个子集中使用的子集标识符条形码可以不同于在其他样品的其他子集中使用的子集标识符。在这些实施方案中，根据片段尺寸对照分子中存在的子集标识符条形码的序列，可以鉴定属于不同的第二样品的片段尺寸对照分子的存在。在一些实施方案中，污染评分可以是对子集中使用的片段尺寸对照分子的每个类型/组特定的(即，子集中使用的片段尺寸对照分子的不同长度/组的单独污染评分)，或者污染评分可以是代表片段尺寸对照分子的不同长度/组的总评分。在一些实施方案中，污染评分可以基于属于不同的第二样品的片段尺寸对照分子的数目来估计。在一些实施方案中，污染评分可以基于属于不同的第二样品的片段尺寸对照分子的测序读段数目来估计。在一些实施方案中，污染评分可以基于属于不同的第二样品的片段尺寸对照分子的测序读段数目来估计。在一些实施方案中，样品的污染评分可以基于属于其他样品的片段尺寸对照分子的测序读段数目相对于添加至该样品的片段尺寸对照分子的测序读段数目的分数或百分比来估计。在一些实施方案中，样品的污染评分可以基于属于其他样品的片段尺寸对照分子的分子数目相对于添加至该样品的片段尺寸对照分子的分子数目的分数或百分比来估计。在这些实施方案中，添加至该分子的片段尺寸对照分子可以根据片段尺寸对照分子的子集标识符条形码来鉴定。

[0195] 在一些实施方案中，该方法还包括将至少一个或更多个污染评分与至少一个或更多个污染阈值进行比较。污染阈值是指用于评价样品中的核酸分子的分析中的样品污染的预定的阈值或范围。这些阈值也可用于在任何步骤(诸如提取、文库制备、富集、洗涤/清洁和测序)中优化测定。在一些实施方案中，污染阈值可以是对子集中使用的片段尺寸对照分子的每个类型/组特定的(即，子集中使用的片段尺寸对照分子的不同长度/组的单独污染阈值)。在一些实施方案中，污染阈值可以是子集中片段尺寸对照分子的不同长度/组的总阈值，或者是在测定中添加的一个或更多个子集中使用的片段尺寸对照分子的总阈值。在一些实施方案中，子集中的每个组具有特定的污染阈值。例如，片段尺寸对照分子的集合包括两个子集—子集1和子集2。如果每个子集包括两个组—G11和G12(对于子集1)以及G21和G22(对于子集2)，则基于片段尺寸对照分子的回收率，这四个组中的每个组可以具有污染评分(S)—G11的S11、G12的S12、G21的S21和G22的S22。子集中的每个组可以具有单独的预定的污染阈值。在该实例中，T11、T12、T21和T22分别是G11、G12、G21和G22的污染阈值。阈值可以是百分比或分数，并且阈值可以是阈值范围，而不是特定的阈值。为了使该方法被认为是成功的，这些污染评分中的至少一个应该在其相应的污染阈值内。在一些实施方案中，如果多于一个污染评分中的每一个在相应的污染阈值内，则该方法被分类为成功进行。在一些实施方案中，该方法还包括，(i)如果至少一个或更多个污染评分不在一个或更多个污染阈值的相应污染阈值内，则将第一样品分类为被第二样品污染；或者(ii)如果至少一个或更多个污染评分在一个或更多个污染阈值的相应污染阈值内，则将第一样品分类为没有被第二样品污染。

[0196] II.片段尺寸对照分子

[0197] 片段尺寸对照分子是添加至多核苷酸样品以评价和/或优化样品中核酸分子的分析的核酸分子。片段尺寸对照分子可以具有两个区域—片段尺寸区域和标识符区域。片段尺寸对照分子的集合可以被分为片段尺寸对照分子的子集，并且片段尺寸对照分子的每个

子集可以在测定中的一个或更多个不同步骤中被添加,以便在任何步骤(无细胞多核苷酸样品的提取、文库制备、富集和/或测序)中评价片段长度分布、QC指标和/或样品污染。这些片段尺寸对照分子的长度是预定的,并且在一些实施方案中,可以模拟无细胞DNA的天然尺寸分布。片段尺寸对照分子的每个子集可以基于片段尺寸区域的长度进一步分为片段尺寸对照分子的组,并且每个组可以具有不同长度的片段尺寸区域。例如,片段尺寸对照分子的子集可以基于片段尺寸区域的长度分为三组—第一组可以具有120bp长度的片段尺寸区域,第二组可以具有160bp长度的片段尺寸区域,并且第三组可以具有320bp长度的片段尺寸区域。在一些实施方案中,片段尺寸对照分子可以是合成的寡核苷酸。在一些实施方案中,片段尺寸对照分子可以具有非天然存在的核酸序列。在一些实施方案中,片段尺寸对照分子可以具有天然存在的核酸序列。在一些实施方案中,通过扩增任何基因组、质粒或载体或其一部分中的特定区域(即,具有特定长度)而生成的扩增子可以用作片段尺寸对照分子。在一些实施方案中,片段尺寸对照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,片段尺寸对照分子可以包含非天然存在的核苷酸类似物。

[0198] 在另一方面,本公开内容提供了片段尺寸对照分子的集合,该片段尺寸对照分子的集合包括预定的片段尺寸对照分子的至少一个子集,其中预定的片段尺寸对照分子的至少一个子集包括多于一个包含片段尺寸区域的片段尺寸对照分子。在一些实施方案中,片段尺寸对照分子还包含标识符区域。片段尺寸区域是片段尺寸对照分子中代表片段尺寸对照分子长度的区域。在一些实施方案中,至少一个子集包括至少一组片段尺寸对照分子。在一些实施方案中,一个组中的片段尺寸对照分子的片段尺寸区域具有相同的长度。每组片段尺寸对照分子的片段尺寸区域的长度可以不同。例如,片段尺寸对照分子的子集可以基于片段尺寸区域的长度分为三组—第一组可以具有120bp长度的片段尺寸区域,第二组可以具有160bp长度的片段尺寸区域,并且第三组可以具有320bp长度的片段尺寸区域。片段尺寸区域的长度可以在10bp和1000bp之间。在一些实施方案中,一个组中的片段尺寸区域的长度不同于其他组中的片段尺寸区域的长度。

[0199] 标识符区域是片段尺寸对照分子中用于区分一种片段尺寸对照分子与其他片段尺寸对照分子的区域。标识符区域也用于区分来自一个子集的片段尺寸对照分子与来自另一个子集的片段尺寸对照分子。在一些实施方案中,标识符区域位于片段尺寸区域的一侧或两侧。在一些实施方案中,标识符区域包含分子条形码。分子条形码用作片段尺寸对照分子的标识符。标识符区域可以存在于片段尺寸区域的一侧或两侧。分子条形码用作片段尺寸对照分子的标识符。在一些实施方案中,标识符区域包含条形码,条形码用作(i)分子标识符条形码,例如,用于鉴定每个片段尺寸对照分子并区分一种片段尺寸对照分子与另一种片段尺寸对照分子的条形码,和(ii)子集标识符条形码,例如,用于鉴定片段尺寸对照分子所属子集(即,片段尺寸对照分子属于子集1还是子集2)的条形码。对于一个子集中的所有片段尺寸对照分子,子集标识符条形码可以是相同的,并且一个子集的子集标识符条形码可以不同于其他子集的子集标识符条形码。例如,子集1中的所有片段尺寸对照分子可以具有子集标识符标签S1,并且子集2中的所有片段尺寸对照分子可以具有子集标识符标签S2。在一些实施方案中,一个样品中一个子集的子集标识符条形码可以不同于其他样品中

相应子集的子集标识符条形码。例如，如果片段尺寸对照分子用于评价样品的污染，则一个子集的子集标识符条形码不同于在同一流动池/批次内的所有其它样品中使用的相应子集的子集标识符条形码。在一些实施方案中，标识符区域可以包含用于区分一个样品的片段尺寸对照分子与其他样品的片段尺寸对照分子的样品索引。例如，在富集步骤之后添加的片段尺寸对照分子的子集的标识符区域可以包含样品索引序列。在一些实施方案中，标识符区域可以通过连接被附接至片段尺寸区域。

[0200] 在一些实施方案中，片段尺寸对照分子包含一个或更多个引物结合位点。在一些实施方案中，引物结合位点位于标识符区域中。在一些实施方案中，标识符区域还可以具有另外的流动池结合位点，诸如P5和P7，其允许片段尺寸对照分子附接至下一代测序仪(例如，Illumina测序仪)的流动池表面。

[0201] 在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含相同的寡核苷酸序列。在一些实施方案中，一个组中的片段尺寸对照分子的片段尺寸区域包含至少两种可区分的寡核苷酸序列。在一些实施方案中，第一子集中的片段尺寸对照分子的片段尺寸区域包含可与第二子集中的片段尺寸对照分子的片段尺寸区域的寡核苷酸序列区分的寡核苷酸序列。

[0202] 在一些实施方案中，片段尺寸区域的长度可以为至少10bp、至少50bp、至少60bp、至少70bp、至少80bp、至少90bp、至少100bp、至少120bp、至少150bp、至少200bp、至少250bp、至少300bp、至少400bp、至少500bp、至少600bp、至少700bp、至少800bp、至少900bp或至少1000bp。在一些实施方案中，片段尺寸区域的长度可以在10bp和1000bp之间。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于等摩尔浓度。在一些实施方案中，片段尺寸对照分子的子集中的每一个处于非等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于等摩尔浓度。在一些实施方案中，子集中的片段尺寸对照分子的组中的每一个处于非等摩尔浓度。

[0203] 图3是适用于本公开内容的一些实施方案的片段尺寸对照分子的示意图。这里描述的片段尺寸对照分子的集合具有的长度类似于无细胞DNA的天然尺寸分布的长度。在图3中，例如，片段尺寸对照分子的集合已被分类成两个子集—子集1和子集2。图3中的片段尺寸对照分子是双链DNA分子。为了说明的目的，仅示出了双链片段尺寸对照分子的一条链。在该实施方案中，每个子集被分成三组—组1、组2和组3。在图3中，每个组中所有片段尺寸对照分子的片段尺寸区域具有相同的序列，并且一个组中片段尺寸区域的长度不同于其他组中片段尺寸区域的长度。在图3中，“------”区域代表片段尺寸对照分子的片段尺寸区域。在该实施方案中，标识符区域存在于片段尺寸区域的两侧。标识符区域在片段尺寸区域的两个末端处具有引物结合位点。此外，这里，子集1中的组1的片段尺寸区域的序列与子集2中的组1的片段尺寸区域的序列相同。同样，子集1中的组2和组3的片段尺寸区域的序列分别与子集2中的组2和组3的片段尺寸区域的序列相同。

[0204] 片段尺寸区域两侧的标识符区域具有分子标识符条形码(MB)，而子集标识符条形码(S)仅位于一侧。分子条形码用作单独的片段尺寸对照分子的标识符，并且每个片段尺寸对照分子具有独特的分子条形码(即，分子1具有MB1&MB2，分子2具有MB3&MB4，分子3具有MB5&MB6，等等)。子集标识符条形码可用作片段尺寸对照分子所属子集的标识符。这里，子集1和子集2的所有片段尺寸对照分子分别具有子集标识符条形码S1和S2。在该实例中，子

集标识符条形码位于片段尺寸区域的一侧。

[0205] 在一些实施方案中，分子条形码可以位于片段尺寸区域的一侧或两侧。在一些实施方案中，子集标识符条形码可以位于片段尺寸区域的一侧或两侧。

[0206] 在该实施方案中，标识符区域在片段尺寸区域的两侧具有引物结合位点。这里，对于子集1，组1、组2和组3的片段尺寸对照分子在片段尺寸区域的两侧分别具有Pr1&Pr2、Pr3&Pr4和Pr5&Pr6引物结合位点。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，这些引物结合位点用于分析片段尺寸对照分子的回收率。在一些实施方案中，代替通过测序分析片段尺寸对照分子的回收率，片段尺寸对照分子的回收率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)、定量qPCR或凝胶电泳来分析。

[0207] 图4是适用于本公开内容的一些实施方案的片段尺寸对照分子的示意图。在图4中，例如，片段尺寸对照分子的集合已被分类成两个子集—子集1和子集2。图4中的片段尺寸对照分子是双链DNA分子。为了说明的目的，仅示出了双链片段尺寸对照分子的一条链。在该实施方案中，每个子集被分成三组—组1、组2和组3。在图4中，每个组中所有片段尺寸对照分子的片段尺寸区域具有相同的长度，但是一个组中的片段尺寸区域的序列可以不同。在图4中，“------”和“～～～～”区域代表片段尺寸对照分子的片段尺寸区域。“----”和“～～～～”区域示出，每个组包含两种不同的序列，但它们具有相同的长度。一个组中的片段尺寸区域的长度不同于其他组中的片段尺寸区域的长度。在该实施方案中，标识符区域存在于片段尺寸区域的两侧。此外，这里，子集1中的组1的片段尺寸区域的两种序列与子集2中的组1的片段尺寸区域的两种序列相同。同样，子集1中的组2和组3的片段尺寸区域的两种序列分别与子集2中的组2和组3的片段尺寸区域的两种序列相同。

[0208] 两侧的标识符区域具有分子标识符条形码(MB)，而子集标识符条形码(S)仅位于一侧。分子条形码用作单独的片段尺寸对照分子的标识符，并且每个片段尺寸对照分子具有独特的分子条形码(即，分子1具有MB1&MB2，分子2具有MB3&MB4，分子3具有MB5&MB6，等等)。子集标识符条形码可用作片段尺寸对照分子所属子集的标识符。这里，子集1和子集2的所有片段尺寸对照分子分别具有子集标识符条形码S1和S2。在该实例中，子集标识符条形码位于片段尺寸区域的一侧。

[0209] 在一些实施方案中，分子条形码可以位于片段尺寸区域的一侧或两侧。在一些实施方案中，子集标识符条形码可以位于片段尺寸区域的一侧或两侧。

[0210] 在一些实施方案中，片段尺寸对照分子可以具有非天然存在的核酸序列。在一些实施方案中，片段尺寸对照分子可以具有天然存在的核酸序列。在一些实施方案中，片段尺寸对照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，片段尺寸对照分子可以包含非天然存在的核苷酸类似物。

[0211] 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。

[0212] 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。

[0213] 在一些实施方案中，片段尺寸对照分子的量为至少1阿托摩尔、至少2阿托摩尔、至少5阿托摩尔、至少10阿托摩尔、至少15阿托摩尔、至少20阿托摩尔、至少50阿托摩尔、至少75阿托摩尔、至少100阿托摩尔、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、至少300飞摩尔、至少400飞摩尔、至少500飞摩尔、至少600飞摩尔、至少700飞摩尔、至少800飞摩尔、至少900飞摩尔、至少1皮摩尔、至少2皮摩尔、至少5皮摩尔或至少10皮摩尔。在一些实施方案中，片段尺寸对照分子的量可以在1阿托摩尔和10皮摩尔之间。

[0214] 本公开内容的另外的实施方案包括添加了尺寸片段对照分子的组合物。例如，添加了尺寸片段对照分子的无细胞DNA样品是本公开内容的一种实施方案。类似地，在实施本发明主题的方法期间产生的包含尺寸片段对照分子的一种或更多种不同子集的许多组合物被认为是本公开内容的实施方案。

[0215] III.方法的一般特征

[0216] A.样品

[0217] 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。

[0218] 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或更多毫升。取样血浆的体积通常在约5mL至约20mL之间。

[0219] 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($2x10^{11}$)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。

[0220] 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。

[0221] 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(μ

g)，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。

[0222] 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，其中众数为长度约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。

[0223] 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。

[0224] B.加标签

[0225] 在一些实施方案中，核酸分子(来自多核苷酸样品和片段尺寸对照分子)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，分子条形码通过连接(例如，钝末端连接或粘末端连接)通过衔接子掺入到样品中的核酸分子(例如，cfDNA分子)中。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。

[0226] 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。

[0227] 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。

[0228] 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指定至特定分子。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。

[0229] 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。

[0230] 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。

[0231] 子集标识符条形码可以是片段尺寸对照分子的标识符区域的一部分。子集标识符条形码用于鉴定片段尺寸对照分子所属的子集(即,片段尺寸对照分子属于子集1还是子集2)。对于一个子集中的所有片段尺寸对照分子,子集标识符条形码可以是相同的,并且一个子集的子集标识符条形码可以不同于同一样品中和/或同一批次中也使用的其他样品中的其他子集的子集标识符条形码。在一些实施方案中,一个样品中一个子集的子集标识符条形码可以不同于其他样品中相应子集的子集标识符条形码。例如,如果片段尺寸对照分子用于评价样品的污染,则一个子集的子集标识符条形码不同于在同一流动池/批次内的所有其它样品中使用的相应子集的子集标识符条形码。在一些实施方案中,标识符区域可以包含用于区分一个样品的片段尺寸对照分子与其他样品的片段尺寸对照分子的样品索引。例如,在富集步骤之后添加的片段尺寸对照分子的子集的标识符区域可以包含样品索引序列。

[0232] 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。

[0233] C.扩增

[0234] 样品核酸和片段尺寸对照分子可以侧接衔接子,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主维持序列的复制(self-sustained sequence-based replication)。

[0235] 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。

[0236] D.富集

[0237] 在一些实施方案中,在对核酸测序之前富集序列。任选地进行对特定靶区域的富集或进行非特异性地富集(“靶序列”)。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。

[0238] 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。

[0239] E.测序

[0240] 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和片段尺寸对照分子进行测序。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。

[0241] 可以对一种或更多种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。

[0242] 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、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。

[0243] F.分析

[0244] 测序可以生成多于一个序列读段(sequence reads)或读段(reads)。序列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。序列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。

[0245] 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。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可有空格。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。

[0246] 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。

[0247] 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)。

[0248] 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)。

[0249] 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,”Bioinformatics 27(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可以被存储在主文件中。

[0250] 一些实施方案提供了序列读段的装配。例如,在通过比对的装配中,将序列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,将序列读段与参考序列比对或映射到参考序列也可以用于鉴定序列读段中的变异序列。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。

[0251] 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)。

[0252] 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。

[0253] 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。

[0254] 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串定义匹配和/或不匹配和缺失(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。

[0255] 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。

[0256] 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。

[0257] 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。

[0258] 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)。

[0259] 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。

[0260] 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对齐时,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。

[0261] 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。

[0262] IV.计算机系统

[0263] 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，这样的方法可以使用计算机处理器进行，所述方法可以包括：(a)将片段尺寸对照分子的第一子集添加至无细胞多核苷酸样品中的核酸分子，从而产生第一加标样品；(b)从第一加标样品提取核酸；(c)将片段尺寸对照分子的第二子集添加至提取的核酸，从而产生第二加标样品；(d)处理第二加标样品的至少一个子集，从而产生经处理的样品，其中所述处理包括将第二加标样品的所述至少一个子集分区、加标签和/或扩增；(e)将片段尺寸对照分子的第三子集添加至经处理的样品，从而产生第三加标样品；(f)富集第三加标样品的至少一个子集，从而产生富集的样品；(g)将片段尺寸对照分子的第四子集添加至富集的样品的至少一个子集，从而产生第四加标样品；(h)对第四加标样品测序以生成多于一个序列读段；(i)分析多于一个序列读段以生成片段尺寸对照分子的多于一个片段尺寸评分；和(j)将多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较。在该实施方案中，该系统包括用于添加片段尺寸对照分子、分区、扩增、富集和测序的组件。

[0264] 图5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501。计算机系统501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501被配置成执行样品制备和样品分析，包括核酸测序。

[0265] 计算机系统501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”)505，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501还包括存储器或存储器位置510(例如，随机存取存储器、只读存储器、闪速存储器)、电子存储单元515(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520(例如，网络适配器)和外围设备525，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510、储存单元515、接口520和外围设备525与CPU 505通过通信网络或总线(实线路)，诸如主板(motherboard)通信。存储单元515可以是用于存储数据的数据存储单元(或数据储存库)。计算机系统501可以借助于通信接口520被可操作地耦合至计算机网络530。计算机网络530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530为电信和/或数据网络。计算机网络530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501，计算机网络530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501的设备作为客户端或服务器运行。

[0266] CPU 505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510中。由CPU 505进行的操作的实例可以包括读取、解码、执行和写回。

[0267] 存储单元515可以存储文件，诸如驱动程序、库和保存的程序。存储单元515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515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501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501的外部，诸如位于通过内联网或因特网与计算机系统501通信的远程服务器上。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。

[0268] 计算机系统501可以与一个或更多个远程计算机系统通过网络530进行通信。对于实施方案，计算机系统501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。用户可以通过网络530访问计算机系统501。

[0269] 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501的电子存储位置，诸如，例如存储器510或电子存储单元515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505执行。在一些情况下，代码可以从存储单元515检索并存储在存储器510上，以便于处理器505即时访问。在一些情况下，可以不包括电子存储单元515，而将机器可执行指令存储于存储器510上。

[0270] 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，执行包括以下的方法的至少一部分：(a)将片段尺寸对照分子的第一子集添加至无细胞多核苷酸样品中的核酸分子，从而产生第一加标样品；(b)从第一加标样品提取核酸；(c)将片段尺寸对照分子的第二子集添加至核酸，从而产生第二加标样品；(d)处理第二加标样品的至少一个子集，从而产生经处理的样品，其中所述处理包括将第二加标样品的所述至少一个子集分区、加标签和/或扩增；(e)将片段尺寸对照分子的第三子集添加至经处理的样品，从而产生第三加标样品；(f)富集第三加标样品的至少一个子集，从而产生富集的样品；(g)将片段尺寸对照分子的第四子集添加至富集的样品的至少一个子集，从而产生第四加标样品；(h)对第四加标样品测序以生成多于一个序列读段；(i)分析多于一个序列读段以生成片段尺寸对照分子的多于一个片段尺寸评分；和(j)将多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较。

[0271] 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。

[0272] 本文所提供的系统和方法的各方面，诸如计算机系统501，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。

[0273] 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。

[0274] 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。

[0275] 计算机系统501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。

[0276] 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。

[0277] V.应用

[0278] A.癌症和其他疾病

[0279] 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。

[0280] 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。

[0281] B.治疗和相关管理

[0282] 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。

[0283] 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受试者相同治疗的患者。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)。

[0284] 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。含有免疫治疗剂的药物组合物通常静脉内施用。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)也可以通过以下方法施用,诸如例如,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。

[0285] 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。

[0286] 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。

[0287] 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。

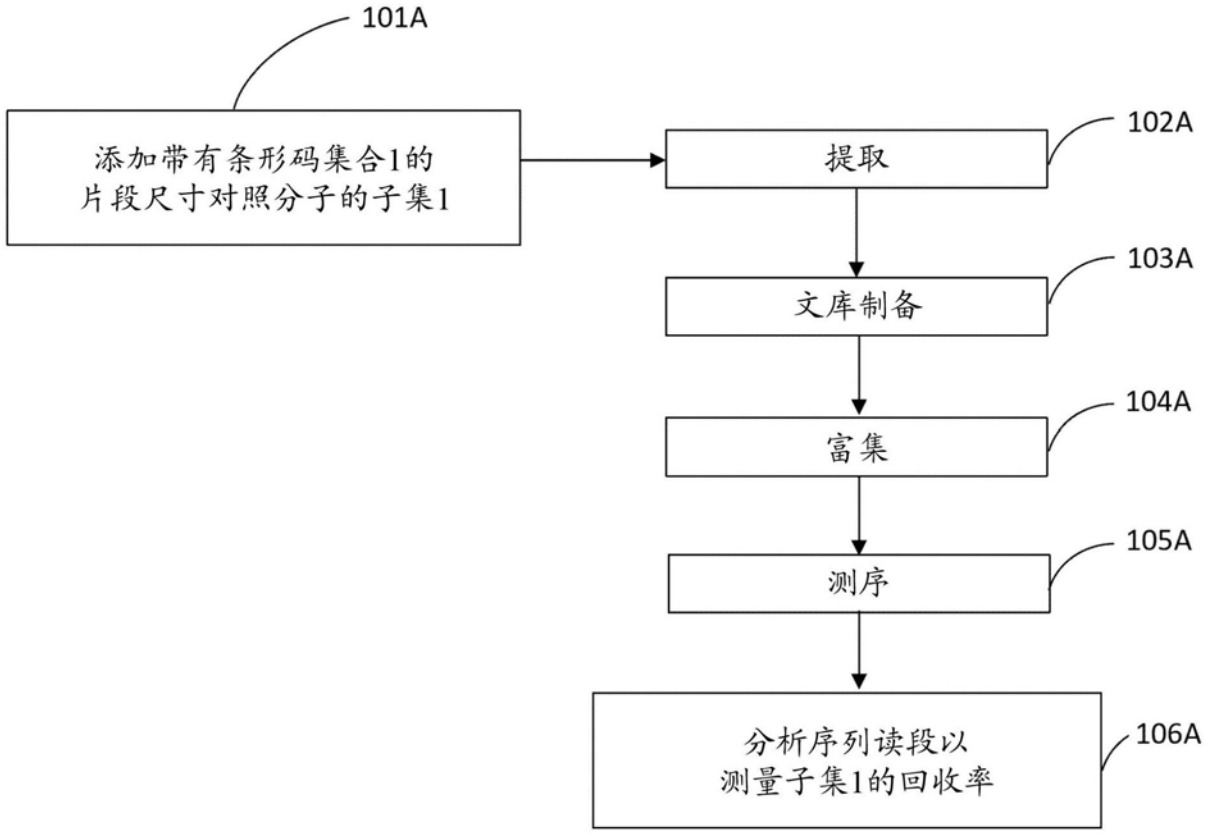

101A
添加带有条形码集合1的
片段尺寸对照分子的子集1
102A
提取
103A
文库制备
104A
富集
105A
测序
106A
分析序列读段以
测量子集1的回收率

图1A

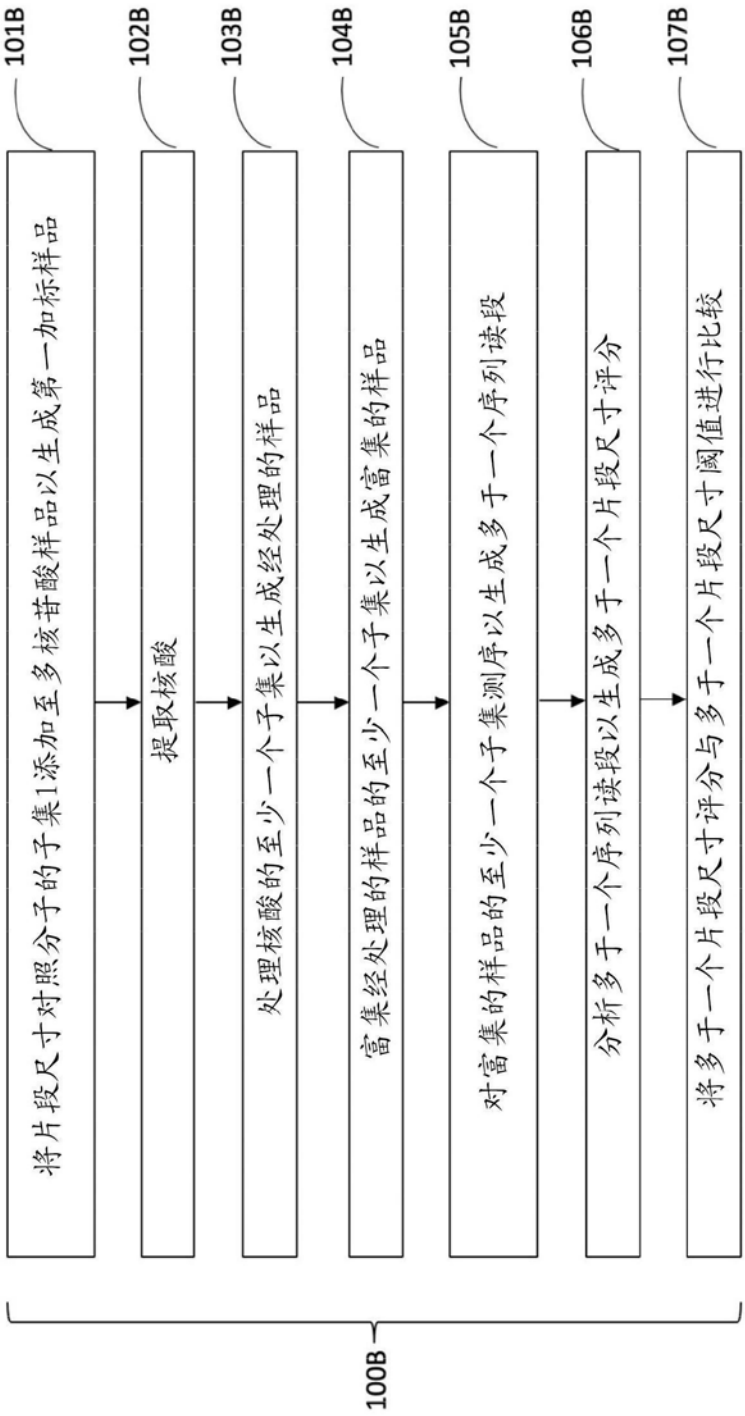

101B
将片段尺寸对照分子的子集1添加至多核苷酸样品以生成第一加标样品
102B
提取核酸
103B
处理核酸的至少一个子集以生成经处理的样品
104B
富集经处理的样品的至少一个子集以生成富集的样品
105B
对富集的样品的至少一个子集测序以生成多于一个序列读段
106B
分析多于一个序列读段以生成多于一个片段尺寸评分
107B
将多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较
100B

图1B

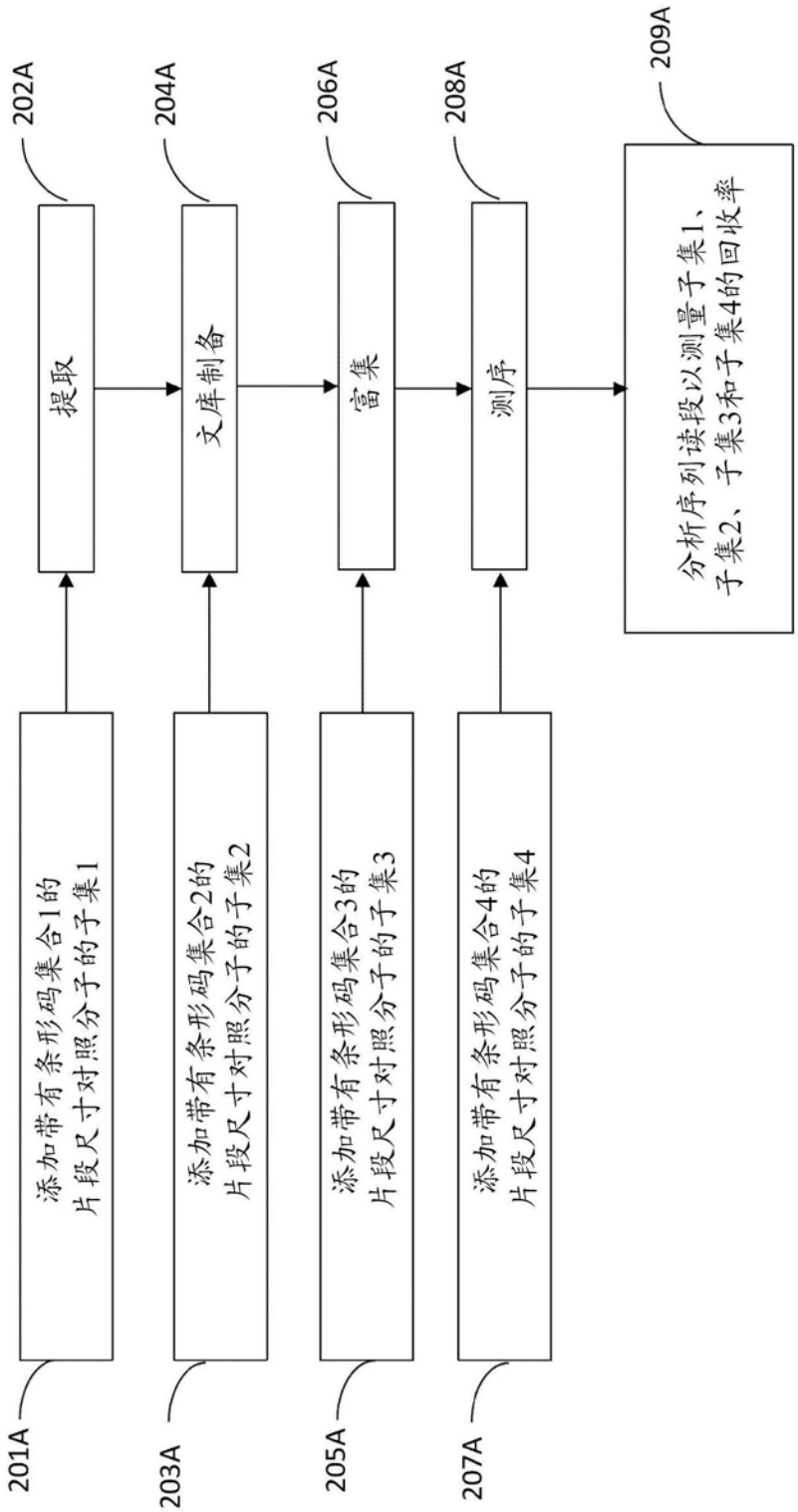

202A
204A
206A
208A
209A
提取
文库制备
富集
测序
分析序列读段以测量子集1、子集2、子集3和子集4的回收率
添加带有条形码集合1的片段尺寸对照分子的子集1
添加带有条形码集合2的片段尺寸对照分子的子集2
添加带有条形码集合3的片段尺寸对照分子的子集3
添加带有条形码集合4的片段尺寸对照分子的子集4
201A
203A
205A
207A

图2A

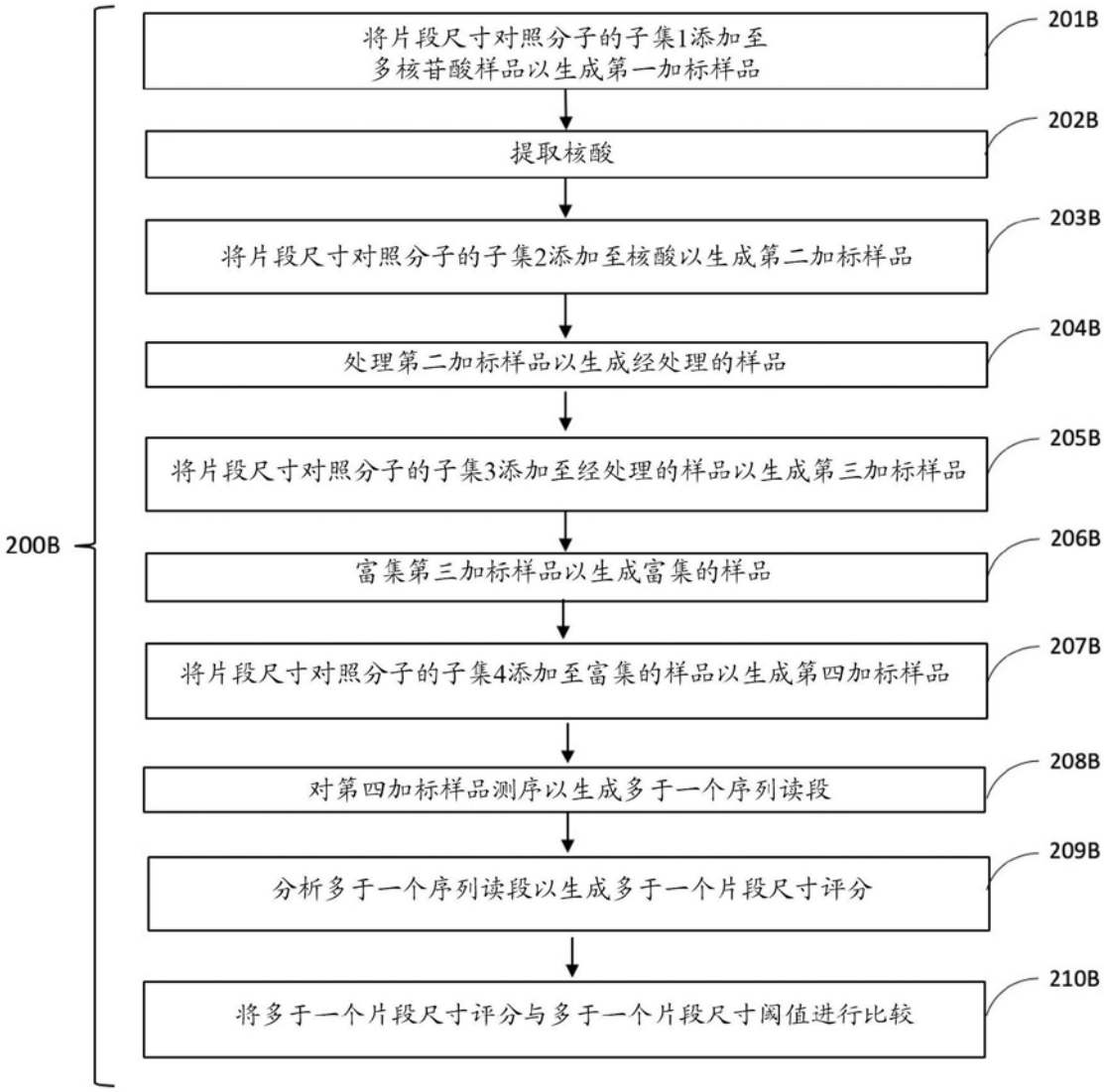

200B
将片段尺寸对照分子的子集1添加至多核苷酸样品以生成第一加标样品
201B
提取核酸
202B
将片段尺寸对照分子的子集2添加至核酸以生成第二加标样品
203B
处理第二加标样品以生成经处理的样品
204B
将片段尺寸对照分子的子集3添加至经处理的样品以生成第三加标样品
205B
富集第三加标样品以生成富集的样品
206B
将片段尺寸对照分子的子集4添加至富集的样品以生成第四加标样品
207B
对第四加标样品测序以生成多于一个序列读段
208B
分析多于一个序列读段以生成多于一个片段尺寸评分
209B
将多于一个片段尺寸评分与多于一个片段尺寸阈值进行比较
210B

图2B

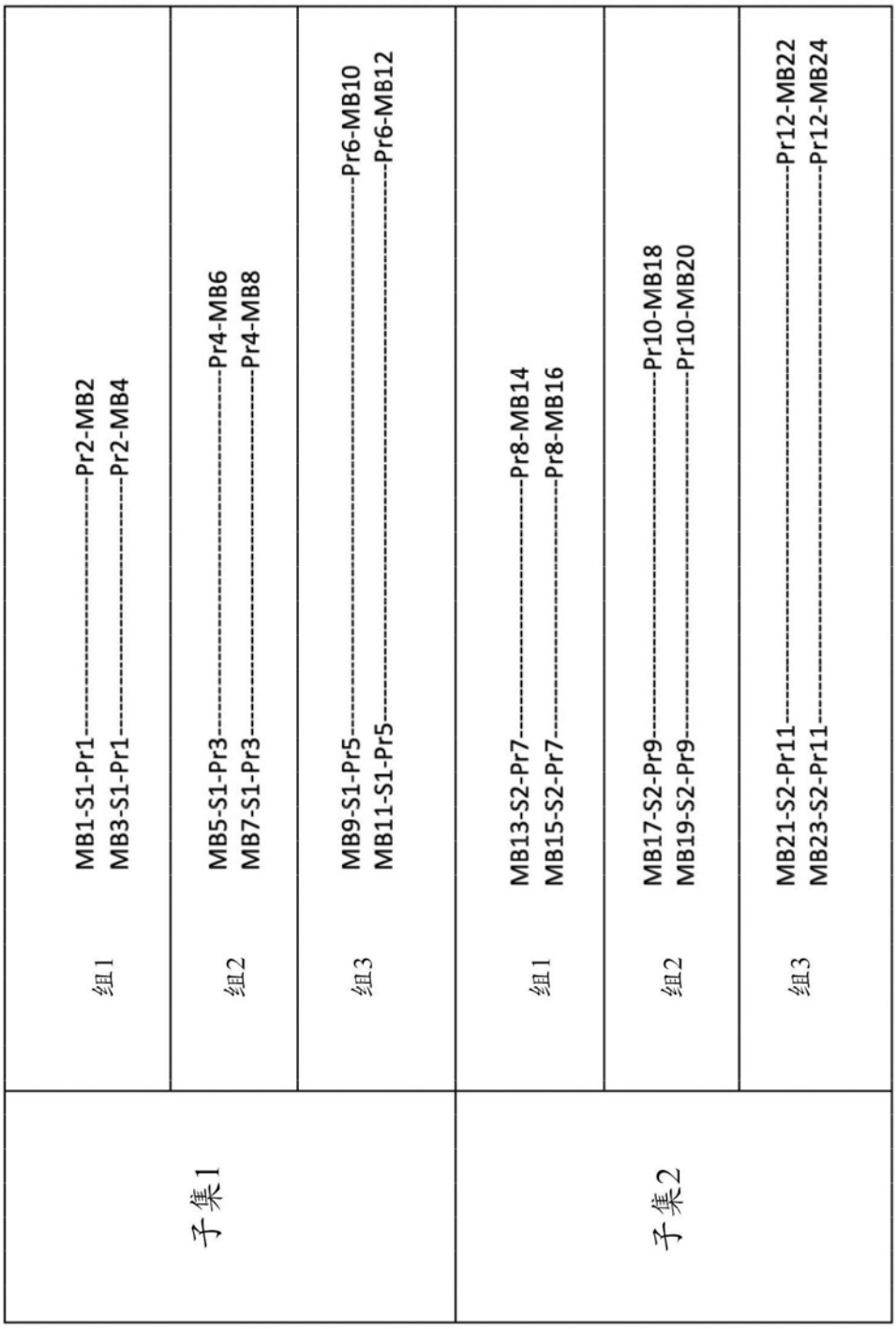
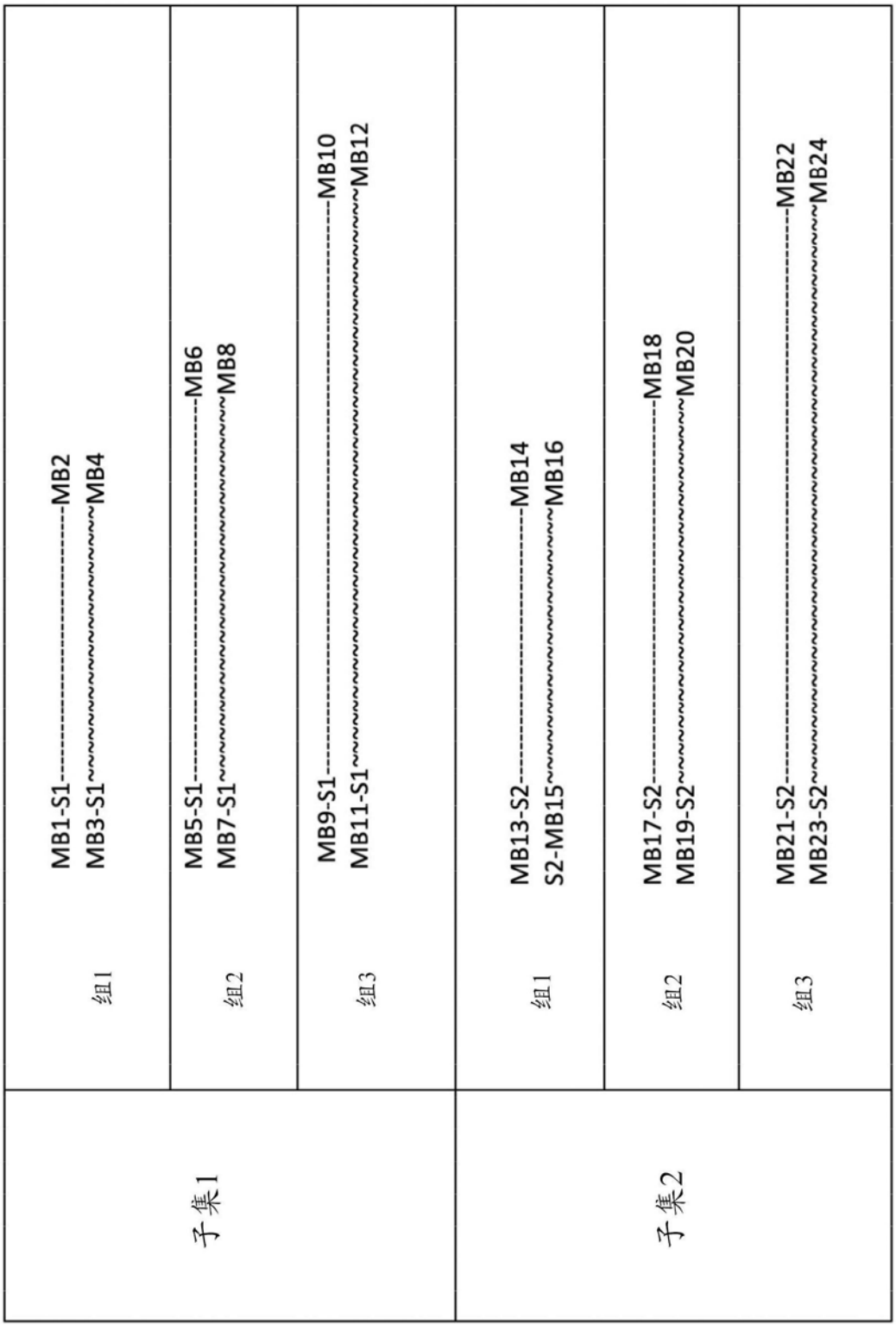

子集1
组1
MB1-S1-Pr1----------Pr2-MB2
MB3-S1-Pr1----------Pr2-MB4
组2
MB5-S1-Pr3----------Pr4-MB6
MB7-S1-Pr3----------Pr4-MB8
组3
MB9-S1-Pr5----------Pr6-MB10
MB11-S1-Pr5----------Pr6-MB12
子集2
组1
MB13-S2-Pr7----------Pr8-MB14
MB15-S2-Pr7----------Pr8-MB16
组2
MB17-S2-Pr9----------Pr10-MB18
MB19-S2-Pr9----------Pr10-MB20
组3
MB21-S2-Pr11----------Pr12-MB22
MB23-S2-Pr11----------Pr12-MB24

图3


子集1
子集2
组1
组2
组3
组1
组2
组3
MB1-S1----------MB2
MB3-S1~~~

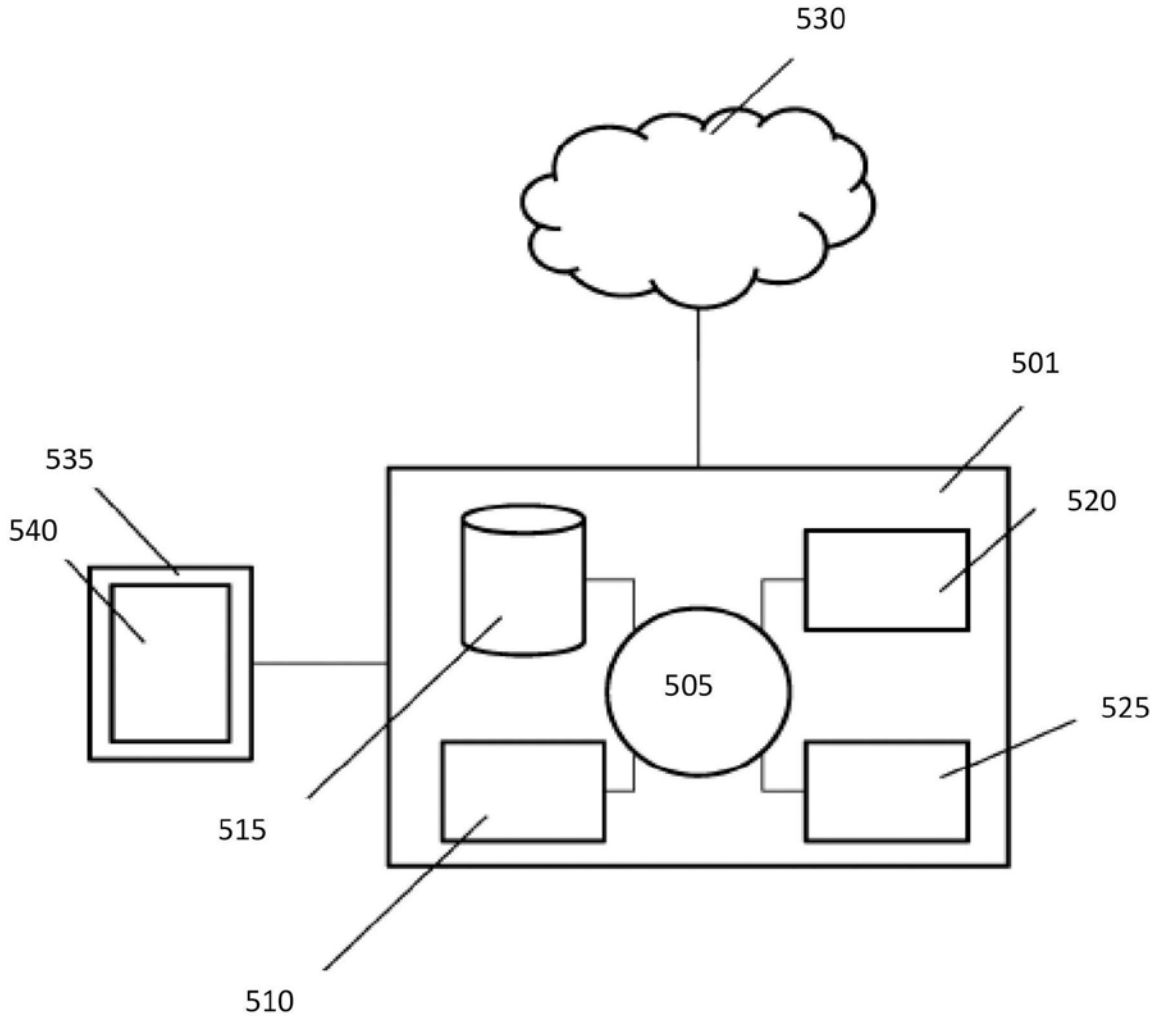


图5