(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 113366115 A
(43)申请公布日 2021.09.07

(21)申请号 202080011456.1

(22)申请日 2020.01.29

(30)优先权数据
62/798,378 2019.01.29 US
62/843,972 2019.05.06 US

(85)PCT国际申请进入国家阶段日
2021.07.29

(86)PCT国际申请的申请数据
PCT/IB2020/050715 2020.01.29

(87)PCT国际申请的公布数据
WO2020/157684 EN 2020.08.06

(71)申请人 深圳华大智造科技股份有限公司
地址 518083 广东省深圳市盐田区北山工业区综合楼及11栋2楼
申请人 深圳华大生命科学研究院

(72)发明人 布罗克·A·彼得斯 王欧
拉多吉·T·德尔马纳茨

(74)专利代理机构 上海胜康律师事务所 31263
代理人 李献忠 张华

(51)Int.Cl.
C12Q 1/6806(2018.01)
C12N 15/10(2006.01)
C40B 50/06(2006.01)

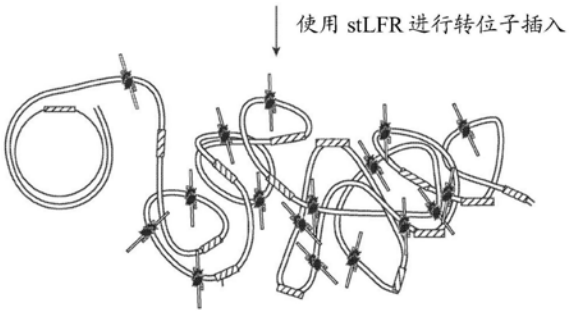
权利要求书8页 说明书37页 附图25页

(54)发明名称

高覆盖率STLFR

(57)摘要

本文描述的是高覆盖率单管长片段读取(stLFR)技术,其使用并对在共条形码化之前已被扩增的靶DNA片段执行stLFR,这为测序提供了更高的DNA量并增加了测序覆盖率。在一些实施方案中,本申请中描述的高覆盖率stLFR使用两轮stLFR。在一些实施方案中,用转位子使靶DNA片段转位,该转位子具有可用于对序列读段排序的特定位置条形码。



CN 113366115 A

1.一种用于制备用于确定靶核酸的序列的条形码化多核苷酸文库的方法,所述方法包括:

(a)提供源自所述靶核酸的片段,其中所述片段是双链或部分双链的;

(b)将交错的单链断裂引入至少一些双链片段中,由此产生多个第一复合物,其中每个第一复合物包含多个第一亚片段,以及

(c)将第一捕获寡核苷酸序列与所述第一亚片段中的至少一些进行关联,

其中每个第一捕获寡核苷酸序列包含第一条形码,并且任选地包含启动子序列或引物结合序列,以及

其中所述关联包括将(a)中的所述双链片段或(b)中的所述复合物与多个个体第一珠组合,其中每个个体第一珠包含固定在其上的多个第一捕获寡核苷酸,其中每个捕获寡核苷酸包含第一捕获寡核苷酸序列,其中固定在每个个体第一珠上的所述第一捕获寡核苷酸包含相同的第一捕获寡核苷酸序列,其中大多数不同的第一珠具有固定在其上的不同的第一捕获寡核苷酸,并且其中每个不同的第一捕获寡核苷酸序列包含不同的第一条形码,

由此提供条形码化第一亚片段;

(d)扩增条形码化第一亚片段的至少一部分以产生扩增的条形码化第一亚片段,其中所述扩增的条形码化第一亚片段是双链或部分双链的;

(e)将交错的单链断裂引入所述扩增的条形码化第一亚片段中的一些中以生成第二复合物,所述第二复合物各自包含多个第二亚片段;以及

(f)将第二捕获寡核苷酸序列与所述第二亚片段中的至少一些进行关联;

其中所述关联包括将(d)中所述扩增的条形码化第一亚片段或(e)中的所述第二复合物与多个个体第二珠组合,其中每个个体第二珠包含固定在其上的多个第二捕获寡核苷酸,其中每个第二捕获寡核苷酸包含第二捕获寡核苷酸序列,其中固定在每个个体第二珠上的所述第二捕获寡核苷酸包含相同的第二捕获寡核苷酸序列,其中大多数不同的第二珠具有固定在其上的不同的第二捕获寡核苷酸,并且其中每个不同的第二捕获寡核苷酸序列包含不同的第二条形码,

由此提供条形码化第二亚片段文库。

2.如权利要求1所述的方法,其中所述第一亚片段的平均长度在大小上是第二亚片段的平均长度的至少2倍。

3.如权利要求1所述的方法,其中步骤(c)在单一混合物中进行,其中所述单一混合物中第一珠的数量大于靶核酸片段的数量,并且其中每个第一珠包含固定在其上的所述第一捕获寡核苷酸的多个副本。

4.如权利要求3所述的方法,其中在步骤(b)中通过接合或通过合成将第一插入寡核苷酸添加到至少一些第一亚片段中,并且其中步骤(c)还包括:

(1)将所述第一捕获寡核苷酸接合至所述第一插入寡核苷酸,或

(2)将第一捕获寡核苷酸与所述第一插入寡核苷酸杂交,然后通过DNA聚合酶延伸所述插入寡核苷酸以并入第一条形码。

5.如权利要求1所述的方法,其中步骤(f)在单一混合物中进行,并且其中所述单一混合物中第二珠的数量大于所述扩增的条形码化第一亚片段的数量,其中每个第二珠包含固定在其上的所述第二捕获寡核苷酸的多个副本。

6.如权利要求5所述的方法,其中至少一些第二亚片段中的每个连接至第二插入寡核苷酸,并且其中步骤(e)还包括:

(1)将所述第二捕获寡核苷酸接合至所述第二插入寡核苷酸,或

(2)将所述第二捕获寡核苷酸与所述第二插入寡核苷酸杂交,然后通过DNA聚合酶延伸所述插入寡核苷酸以并入第二条形码。

7.如权利要求4所述的方法,其中通过转位引入所述第一插入寡核苷酸。

8.如权利要求6所述的方法,其中所述第二插入寡核苷酸通过转位引入。

9.如权利要求7所述的方法,其中步骤(c)还包括向反应中添加第一夹板寡核苷酸,其中所述第一插入寡核苷酸包含与所述第一夹板寡核苷酸的第一部分互补的第一杂交序列,并且其中所述捕获寡核苷酸包含与所述第一夹板寡核苷酸的第二部分互补的共同序列。

10.如权利要求8所述的方法,其中步骤(f)还包括向反应中添加第二分裂寡核苷酸,所述第二插入寡核苷酸包含与所述第二夹板寡核苷酸的第一部分互补的第二杂交序列,并且其中所述捕获寡核苷酸包含与第二夹板寡核苷酸的第二部分互补的共同序列。

11.如权利要求4所述的方法,其中所述第一插入寡核苷酸与互补寡核苷酸杂交以形成部分双链的第一插入寡核苷酸,并且

其中所述第一插入寡核苷酸通过3'分支接合而接合至所述断裂中的至少一些。

12.如权利要求1所述的方法,其中所述第一捕获寡核苷酸中的每个,或所述第二捕获寡核苷酸中的每个,或两者都包含独特分子标识符(UMI)。

13.如权利要求3所述的方法,其中所述方法:

每个第一珠包含所述第一捕获寡核苷酸的多个副本,其中将所述第一捕获寡核苷酸与互补寡核苷酸杂交以形成部分双链的第一捕获寡核苷酸,

其中所述方法步骤(b)包括:

(1)通过3'分支接合将所述第一捕获寡核苷酸接合至所述第一亚片段中的至少一些中的每个,或

(2)通过3'分支接合将所述互补寡核苷酸接合至所述第一亚片段中的至少一些中的每个,并且延伸所述互补寡核苷酸以并入所述第一条形码序列。

14.如权利要求1-13中任一项所述的方法,其中所述第一捕获寡核苷酸还包含启动子序列,并且其中在步骤(d)中所述扩增所述条形码化第一亚片段的至少一部分是通过以下进行的:

(1)转录所述条形码化第一复合物以生成RNA转录物,

(2)使用退火至所述启动子序列的引物逆转录所述RNA转录物以生成所述条形码化第一复合物的cDNA链,

(3)将所述cDNA环化以产生环化的cDNA链,

(4)通过滚环扩增将所述环化的cDNA链扩增,以及

(5)使用所述扩增的cDNA链作为模板来合成双链或部分双链的条形码化第一复合物。

15.如权利要求1-13中任一项所述的方法,其中所述扩增所述条形码化第一亚片段的至少一部分通过以下方式进行:

(1)从第一珠中释放所述条形码化第一亚片段,

(2)使已释放的条形码化第一亚片段变性以形成单链条形码化第一复合物,

(3)将所述单链条形码化第一亚片段环化，

(4)进行滚环扩增,以及

(5)使用所述扩增的单链条形码第一复合物作为模板来合成双链条形码化第一亚片段。

16.如权利要求14或15所述的方法,其中所述方法包括将权利要求14中的所述环化的cDNA链或权利要求15中的所述单链条形码化第一亚片段分级以选择大小在预定范围内的环。

17.如权利要求1-13中任一项所述的方法,其中所述扩增所述条形码化第一亚片段的至少一部分包括：

(1)使用与所述第一捕获寡核苷酸上的引物结合序列结合的引物来延伸所述条形码化第一亚片段，

(2)从所述第一珠中释放来自(1)的延伸的所述条形码化第一亚片段，

(3)使用单引物扩增将所述条形码化第一亚片段扩增约10-120倍,由此产生扩增的双链条形码化第一亚片段,以及

(4)将衔接子寡核苷酸接合至所述扩增的双链条形码化第一亚片段的末端。

18.如权利要求17所述的方法,其中至少一些第二亚片段中的每个连接至第二插入寡核苷酸,并且其中步骤(e)还包括：

(1)将所述第二捕获寡核苷酸接合至所述第二插入寡核苷酸,或

(2)将所述第二捕获寡核苷酸与所述第二插入寡核苷酸杂交,然后通过DNA聚合酶延伸所述插入寡核苷酸以并入第二条形码，

其中所述衔接子寡核苷酸具有与所述第二插入寡核苷酸相同的序列。

19.如前述权利要求中任一项所述的方法,其中所述靶核酸片段的大小在10kb-100M碱基的范围内。

20.如前述权利要求中任一项所述的方法,其中所述第一亚片段的大小在1kb-20kb的范围内。

21.一种用于分析转录物的方法,所述方法包括：

(a)在单一混合物中,将来自一个或多个细胞的mRNA与第一珠的群组组合,其中每个第一珠包含固定在其上的第一捕获寡核苷酸,其中所述第一捕获寡核苷酸包含共同引物序列、第一条形码序列、UMI和寡核苷酸dT序列,其中在所述单一混合物中第一珠的数量大于mRNA分子的数量，

(b)在包含三核苷酸GGG和所述共同引物序列的衔接子模板的存在下,逆转录捕获的RNA,以产生cDNA/mRNA杂交分子，

其中所述cDNA/mRNA杂交分子各自包含由捕获的RNA的逆转录产生的cDNA和所述第一捕获寡核苷酸，

(c)从所述第一珠中释放所述cDNA，

(d)扩增来自(c)的所述cDNA并产生双链或部分双链cDNA；

(e)将交错的单链断裂引入所述扩增的cDNA中的至少一些中以生成第二复合物,所述第二复合物各自包含多个第二亚片段,以及

(f)将第二捕获寡核苷酸引入所述多个第二亚片段中，

其中每个第二捕获寡核苷酸包含：

(1)任选地启动子序列或引物结合序列,以及

(2)第二条形码,其中固定在相同个体珠上的第二捕获寡核苷酸包含相同的第二条形码,并且大多数珠具有不同的第二条形码,由此提供条形码化第二亚片段。

22.如权利要求21所述的方法,其中所述单一混合物中的所述mRNA来自单一细胞。

23.如权利要求21所述的方法,其中在步骤(d)中所述扩增所述cDNA是通过使用与所述共同引物序列杂交的引物进行滚环扩增而进行的。

24.如权利要求21所述的方法,其中至少一些第二亚片段中的每个连接至第二插入寡核苷酸,并且其中步骤(f)还包括：

(1)将所述第二捕获寡核苷酸接合至所述第二插入寡核苷酸,或

(2)将所述第二捕获寡核苷酸与所述第二插入寡核苷酸杂交,然后通过DNA聚合酶延伸所述插入寡核苷酸以并入第二条形码。

25.如权利要求24所述的方法,其中在步骤(d)中所述扩增所述cDNA包括使用与所述共同引物序列杂交的引物通过单引物扩增来扩增所述cDNA,并将衔接子寡核苷酸接合至所述扩增的第一复合物的末端,其中所述衔接子寡核苷酸具有与所述插入寡核苷酸相同的序列。

26.如前述权利要求中任一项所述的方法,其中所述第一插入寡核苷酸中的每个包含第一位置条形码,其中不同的第一插入寡核苷酸包含不同的第一位置条形码,和/或

其中所述第二插入寡核苷酸中的每个包含第二位置条形码,其中不同的第二插入寡核苷酸包含不同的第二位置条形码。

27.如权利要求19所述的方法,其中一个或多个第一位置条形码与一个或多个第二位置条形码相同。

28.一种分析一个或多个靶区域的全长序列的方法,所述方法包括：

(a)扩增每个靶区域,

(b)将衔接子寡核苷酸接合至包含所述靶区域的扩增的核酸靶片段的两端,

(c)将交错的单链断裂引入来自(b)的所述扩增的核酸片段中的至少一些中,以产生多个第一复合物,所述第一复合物各自包含多个第一亚片段,

(d)将第一捕获寡核苷酸引入所述第一亚片段中的至少一些中,其中每个第一捕获寡核苷酸包含：

(1)任选地启动子序列或引物结合序列,以及

(2)第一条形码,其中固定在相同个体珠上的第一捕获寡核苷酸包含相同的第一条形码,并且大多数珠具有不同的第一条形码,由此提供条形码化第一亚片段。

29.如权利要求28所述的方法,

其中步骤(a)中扩增所述靶区域包括：

(1)用正向引物和反向引物扩增所述靶区域,由此产生扩增区域,所述正向引物和反向引物两者均包含对所述靶区域特异的序列,

其中正向引物包含共同序列和UMI,以及第一靶特异性序列,其中所述正向引物包含特殊碱基,其中所述特殊碱基不存在于天然DNA分子中并且所述特殊碱基可被试剂降解;

(2)通过添加所述试剂来降解过量的所述正向引物,

(3)使用与所述共同序列结合的引物和所述反向引物来扩增(i)中的所述扩增区域,由此产生包含所述UMI的进一步扩增的区域。

30.如权利要求29所述的方法,其中所述特殊碱基是尿嘧啶并且所述试剂能够特异性切割含尿嘧啶的寡核苷酸。

31.如权利要求29所述的方法,其中在步骤(a)中扩增所述靶区域还包括:

(4)接合两个寡核苷酸,每个与来自(3)的所述进一步扩增的区域的一端接合,其中所述两个寡核苷酸共享相同的共同序列,并且其中所述两个寡核苷酸包含不同的UMI,由此产生在两端具有共同序列的接合DNA产物,以及

(5)用与所述共同序列杂交的引物来扩增来自(4)的所述接合DNA产物,由此产生包含所述靶区域的扩增核酸靶片段。

32.如前述权利要求中任一项所述的方法,还包括将3'分支接合衔接子寡核苷酸接合至所述第二亚片段,其中所述接合衔接子寡核苷酸是3'分支接合,并且其中所述衔接子寡核苷酸包括第二PCR引物退火位点。

33.如权利要求25所述的方法,其中所述3'分支接合衔接子寡核苷酸是平端衔接子并且3'分支接合包括来自所述平端衔接子的5'磷酸酯与凹陷的3'羟基在所述第一片段的切口处的共价连接。

34.如权利要求25所述的方法,其中所述第一PCR引物退火位点和所述第二PCR引物退火位点具有不同的序列。

35.如权利要求25所述的方法,其中所述3'分支接合衔接子寡核苷酸包含条形码序列,任选地,该条形码序列是样本条形码序列。

36.如权利要求1-28中任一项所述的方法,其中当引入第一插入寡核苷酸时、引入所述第二插入寡核苷酸或两者均通过转位酶进行时,所述插入寡核苷酸转位酶保持与所述第一亚片段结合或插入寡核苷酸保持与所述第二亚片段结合,或两者兼有。

37.如权利要求29所述的方法,所述方法包括去除所述转位酶,由此分离个体第一亚片段、第二亚片段或两者。

38.如权利要求30所述的方法,还包括使用退火至所述第一和第二PCR引物退火位点的引物来扩增所述第一亚片段、第二亚片段或两者以产生扩增子。

39.如权利要求4所述的方法,其中所述第一插入寡核苷酸中的至少一些各自包含位置条形码,其中不同的第一插入寡核苷酸包含不同的位置条形码。

40.如权利要求6所述的方法,其中所述第二插入寡核苷酸中至少一些各自包含位置条形码,其中不同的第二插入寡核苷酸包含不同的位置条形码。

41.一种用于制备用于对靶核酸进行测序的测序文库的方法,所述方法包括:

(a)将衔接子和UMI的一个或多个副本接合至所述靶核酸的片段,

(b)将已与所述衔接子和所述UMI的所述一个或多个副本接合的所述片段变性以形成单链DNA分子,

(c)对所述单链分子中的至少一些进行滚环扩增以产生包含至少一个单链分子的大于5x副本的纳米球,

(d)将所述纳米球转化为双链或部分双链DNA分子,

(e)将交错的单链断裂引入(d)中的所述DNA分子中,由此生成第一复合物,所述第一复

合物各自包含多个第一亚片段，

(f)将第一捕获寡核苷酸引入第一亚片段中的至少一些中的每个中，其中每个第一捕获寡核苷酸包含：

(1)任选地启动子序列或引物结合序列，以及

(2)第一条形码，其中固定在相同个体珠上的第一捕获寡核苷酸包含相同的第一条形码，并且大多数珠具有不同的第一条形码，由此提供条形码化第一亚片段。

42.如权利要求41所述的方法，其中将至少一些第一亚片段中的每个连接至第一插入寡核苷酸，并且其中步骤(b)还包括：

(1)将所述第一捕获寡核苷酸接合至所述第一插入寡核苷酸，或

(2)将所述第一捕获寡核苷酸与所述第一插入寡核苷酸杂交，然后通过DNA聚合酶延伸所述插入寡核苷酸以并入第一条形码。

43.如权利要求42所述的方法，其中所述插入寡核苷酸中的每个包含位置条形码，其中不同的插入寡核苷酸包含不同的位置条形码。

44.一种将寡核苷酸插入靶核酸片段的方法，所述方法包括：

(a)将交错的单链断裂引入所述片段中，

(b)将来自(a)的所述片段与插入支架接触，其中所述衔接子锚定至所述支架并以预定间距分开，其中所述插入支架包括多个双链或部分双链衔接子和支架，

其中每个衔接子包括包含独特位置条形码的插入寡核苷酸，以及

其中所述接触导致所述多个插入寡核苷酸在所述单链断裂处被引入所述片段中，由此产生第一插入复合物，所述第一插入复合物各自包含多个第一亚片段。

45.如权利要求44所述的方法，其中所述方法还包括：

将所述支架与已插入所述靶核酸中的所述多个衔接子解离。

46.如权利要求44所述的方法，其中所述方法包括：

将多个支架与一些所述核酸片段中的每个接触，其中不同支架中的衔接子具有不同的支架条形码。

47.如权利要求44所述的方法，

其中所述支架是单链核酸分子，以及

其中每个衔接子还包括：

支架杂交序列，

其中所述转位子经由所述支架杂交序列与所述支架杂交，以及

其中所述支架杂交序列能被切割以使所述支架与所述衔接子解离，

其中所述插入寡核苷酸还包含在所述支架内由所有衔接子共享的支架条形码。

48.如权利要求44所述的方法，其中所述方法还包括：

(c)将(i)由(a)产生的所述第一插入复合物;以及(ii)第一珠的群组组合成单一混合物，其中每个第一珠包含固定在其上的第一捕获寡核苷酸的多个副本，所述第一捕获寡核苷酸包含第一条形码，其中固定在相同的个体第一珠上的所述第一捕获寡核苷酸包含相同的第一条形码并且大多数珠具有不同的第一条形码，

(d)对于多个所述第一亚片段中的每个，引入所述第一捕获寡核苷酸，由此产生条形码化第一亚片段，所述条形码化第一亚片段各自连接至所述第一条形码的副本。

49.如权利要求40所述的方法,其中所述方法还包括:

(e)扩增所述多个条形码化第一亚片段,

(f)将交错的单链断裂引入扩增的条形码化亚片段中,以及

(g)将来自步骤(f)的产物与第二插入支架接触,其中所述第二插入支架各自包含锚定至第二插入支架的多个第二衔接子,由此将所述第二衔接子上的第二插入寡核苷酸引入所述扩增的条形码化第一亚片段以产生第二插入复合物,所述第二插入复合物各自包含多个第二亚片段。

50.如权利要求49所述的方法,其中所述第二插入支架的所述支架为单链核酸分子,并且所述第二衔接子中的每个包括:

(1)第二支架杂交序列,

其中所述第二衔接子经由所述第二支架杂交序列与所述支架杂交,以及

其中所述第二插入支架杂交序列能被切割以使所述支架与所述第二衔接子解离;

(2)包含独特位置条形码的第二插入寡核苷酸,以及

在所述第二插入支架的所述支架内由所述第二衔接子共享的第二支架条形码。

51.如权利要求41所述的方法,其中所述第二插入复合物与第二珠的群组混合,其中每个珠包含固定在其上的第二捕获寡核苷酸,所述寡核苷酸包含第二条形码,其中固定在相同个体珠上的所述寡核苷酸包含相同的所述第二条形码并且大多数珠具有不同的第二条形码,

(g)对于所述第二插入复合物中的至少一些中的每个,引入包含第二条形码的所述第二捕获寡核苷酸的多个副本,其中所述多个副本来自单个珠,并由此产生多个条形码化第二亚片段,所述条形码化第二亚片段各自连接至所述第二条形码的至少一个副本。

52.如权利要求44-51所述的方法,其中所述方法还包括:

对所述多个第二亚片段进行测序以产生多个测序读段。

53.如权利要求44-52中任一项所述的方法,其中所述插入支架具有1-50kb的大小。

54.如权利要求44-53中任一项所述的方法,其中相邻的第一衔接子之间的预定间隔在3kb到5kb的范围内。

55.如权利要求44-54中任一项所述的方法,其中相邻的第二衔接子之间的预定间隔在200bp至1000bp的范围内。

56.根据权利要求44-55中任一项所述的方法,其中所述多个插入支架的长度总和等于或大于所述靶核酸的长度。

57.多个插入支架,

其中所述多个插入支架中的每个包含:

(1)多个衔接子,其中所述衔接子是双链或部分双链的,以及

(2)支架,并且所述衔接子锚固至所述支架并以预定间距分开,

其中对于每个插入支架,所述插入支架中的每个衔接子携带独特位置条形码和共同的支架条形码,以及

其中不同插入支架中的衔接子具有不同的支架条形码。

58.如权利要求57所述的插入支架,其中至少一个所述插入支架的大小等于1-10kb的多核苷酸的大小。

59.如权利要求57所述的插入支架,其中至少一个所述插入支架的大小等于10-50kb的多核苷酸的大小。

60.一种核酸复合物,其包含多个如权利要求57-59中任一项所述的插入支架和核酸片段,其中所述多个插入支架与所述靶核酸片段杂交。

61.一种在单个容器中的反应混合物,其中所述反应混合物包含多个如权利要求57-59中任一项所述的插入支架和多个源自靶核酸的片段。

高覆盖率STLFR

对相关申请的交叉引用

本申请要求2019年1月29日提交的美国临时申请号62/798,378和2019年5月6日提交的美国临时申请号62/843,972的优先权。这两个临时申请的内容通过引用并入本文以用于所有目的。

背景技术

[0001] 迄今为止,绝大多数个体全基因组序列缺乏关于作为同源染色体上的连续块传输的单碱基到多碱基变体的顺序的信息。最近开发了许多技术来实现这一点。大多数是基于共条形码化(co-barcoding)的过程(Peters等人,Frontiers in Genetics 5,466(2014)),即将相同的条形码(barcode)添加到单个长基因组DNA分子的亚片段中。测序后,条形码信息可用于确定哪些读段源自原始长DNA分子。该过程首先由Drmanac(WO2006/138284A2(2006))描述,并由Peters等人(Peters等人,Nature 487,190-195(2012))实施为384孔板测定,后来在Drmanac(US 2014/0323316)和Wang等人,BioRxiv,2018年6月29日,doi:https//doi.org/10.1101/357863中以单一反应容器格式实施。

发明内容

[0002] 一种用于制备用于确定靶核酸的序列的条形码化多核苷酸文库的方法,该方法包括:

(a)提供源自靶核酸的片段,其中片段是双链或部分双链的;

(b)将交错的单链断裂引入至少一些双链片段,由此产生多个第一复合物,其中每个第一复合物包含多个第一亚片段,以及

(c)将第一捕获寡核苷酸序列与第一亚片段中的至少一些进行关联,

其中每个第一捕获寡核苷酸序列包含第一条形码,并且任选地包含启动子序列或引物结合序列,以及

其中关联包括将(a)中的双链片段或(b)中的复合物与多个个体第一珠组合,其中每个个体第一珠包含固定在其上的多个第一捕获寡核苷酸,其中每个捕获寡核苷酸包含第一捕获寡核苷酸序列,其中固定在每个个体第一珠上的第一捕获寡核苷酸包含相同的第一捕获寡核苷酸序列,其中大多数不同的第一珠具有固定在其上的不同的第一捕获寡核苷酸,并且其中每个不同的第一捕获寡核苷酸序列包含不同的第一条形码,

由此提供条形码化第一亚片段;

(d)扩增条形码化第一亚片段的至少一部分以产生扩增的条形码化第一亚片段,其中扩增的条形码化第一亚片段是双链或部分双链的;

(e)将交错的单链断裂引入扩增的条形码化第一亚片段中的一些中以生成第二复合物,该第二复合物各自包含多个第二亚片段;以及

(f)将第二捕获寡核苷酸序列与第二亚片段中的至少一些进行关联;

其中关联包括将(d)中扩增的条形码化第一亚片段或(e)中的第二复合物与多个

个体第二珠组合，其中每个个体第二珠包含固定在其上的多个第二捕获寡核苷酸，其中每个第二捕获寡核苷酸包含第二捕获寡核苷酸序列，其中固定在每个个体第二珠上的第二捕获寡核苷酸包含相同的第二捕获寡核苷酸序列，其中大多数不同的第二珠具有固定在其上的不同的第二捕获寡核苷酸，并且其中每个不同的第二捕获寡核苷酸序列包含不同的第二条形码，

由此提供条形码化第二亚片段文库。

[0003] 一种用于制备用于确定靶核酸的序列的条形码化多核苷酸文库的方法，该方法包括：

(a)提供源自靶核酸的片段，其中片段是双链或部分双链的；

(b)将交错的单链断裂引入至少一些双链片段，由此产生多个第一复合物，其中每个第一复合物包含多个第一亚片段，以及

(c)将第一捕获寡核苷酸序列与第一亚片段中的至少一些进行关联，

其中每个第一捕获寡核苷酸序列包含第一条形码，并且任选地包含启动子序列或引物结合序列，以及

其中关联包括将(a)中的双链片段或(b)中的复合物与多个个体第一珠组合，其中每个个体第一珠包含固定在其上的多个第一捕获寡核苷酸，其中每个捕获寡核苷酸包含第一捕获寡核苷酸序列，其中固定在每个个体第一珠上的第一捕获寡核苷酸包含相同的第一捕获寡核苷酸序列，其中大多数不同的第一珠具有固定在其上的不同的第一捕获寡核苷酸，并且其中每个不同的第一捕获寡核苷酸序列包含不同的第一条形码，

由此提供条形码化第一亚片段；

(d)扩增条形码化第一亚片段的至少一部分以产生扩增的条形码化第一亚片段，其中扩增的条形码化第一亚片段是双链或部分双链的；

(e)将交错的单链断裂引入扩增的条形码化第一亚片段中的一些中以生成第二复合物，该第二复合物各自包含多个第二亚片段；以及

(f)将第二捕获寡核苷酸序列与第二亚片段中的至少一些进行关联；

其中关联包括将(d)中扩增的条形码化第一亚片段或(e)中的第二复合物与多个个体第二珠组合，其中每个个体第二珠包含固定在其上的多个第二捕获寡核苷酸，其中每个第二捕获寡核苷酸包含第二捕获寡核苷酸序列，其中固定在每个个体第二珠上的第二捕获寡核苷酸包含相同的第二捕获寡核苷酸序列，其中大多数不同的第二珠具有固定在其上的不同的第二捕获寡核苷酸，并且其中每个不同的第二捕获寡核苷酸序列包含不同的第二条形码，

由此提供条形码化第二亚片段文库。

[0004] 在一些方式中，第一亚片段的平均长度在大小上是第二亚片段的平均长度的至少2倍。

[0005] 在一些方式中，步骤(c)在单一混合物中进行，其中第一珠的数量大于单一混合物中靶核酸片段的数量，并且每个第一珠包含固定在其上的第一捕获寡核苷酸的多个副本。

[0006] 在一些方式中，在步骤(b)中通过接合(ligation)或通过合成将第一插入寡核苷酸添加到至少一些第一亚片段中，并且其中步骤(c)还包括：

(1)将第一捕获寡核苷酸接合至第一插入寡核苷酸，或

(2)将第一捕获寡核苷酸与第一插入寡核苷酸杂交,然后通过DNA聚合酶延伸插入寡核苷酸以并入第一条形码。

[0007] 在一些方式中,步骤(f)在单一混合物中进行,并且其中第二珠的数量大于单一混合物中扩增的条形码化第一亚片段的数量,其中每个第二珠包含固定在其上的第二捕获寡核苷酸的多个副本。

[0008] 在一些方式中,将至少一些第二亚片段中的每个连接(link)至第二插入寡核苷酸,并且其中步骤(e)还包括:

(1)将第二捕获寡核苷酸接合至第二插入寡核苷酸,或

(2)将第二捕获寡核苷酸与第二插入寡核苷酸杂交,然后通过DNA聚合酶延伸插入寡核苷酸以并入第二条形码。

[0009] 在一些方式中,通过转位引入第一插入寡核苷酸。

[0010] 在一些方式中,通过转位引入第二插入寡核苷酸。

[0011] 在一些方式中,步骤(c)还包括向反应中添加第一夹板寡核苷酸,其中第一插入寡核苷酸包含与第一夹板寡核苷酸的第一部分互补的第一杂交序列,并且其中捕获寡核苷酸包含与第一夹板寡核苷酸的第二部分互补的共同序列。

[0012] 在一些方式中,步骤(f)还包括向反应中添加第二夹板寡核苷酸,第二插入寡核苷酸包含与第二夹板寡核苷酸的第一部分互补的第二杂交序列,并且其中捕获寡核苷酸包含与第二夹板寡核苷酸的第二部分互补的共同序列。

[0013] 在一些方式中,将第一插入寡核苷酸与互补寡核苷酸杂交以形成部分双链的第一插入寡核苷酸,并且

其中通过3'分支接合将第一插入寡核苷酸接合至断裂中的至少一些。

[0014] 在一些方式中,第一捕获寡核苷酸中的每个或第二捕获寡核苷酸中的每个或两者都包含独特分子标识符(UMI)。

[0015] 在一些方式中,该方法:

每个第一珠包含第一捕获寡核苷酸的多个副本,其中将第一捕获寡核苷酸与互补寡核苷酸杂交以形成部分双链的第一捕获寡核苷酸,

其中方法步骤(b)包括:

通过3'分支接合将第一捕获寡核苷酸接合至第一亚片段中的至少一些中的每个,或

通过3'分支接合将互补寡核苷酸接合至第一亚片段中的至少一些中的每个,并且延伸互补寡核苷酸以并入第一条形码序列。

[0016] 如权利要求1-13中任一项所述的方法,第一捕获寡核苷酸还包含启动子序列,并且其中在步骤(d)中通过以下方式扩增条形码化第一亚片段的至少一部分:

(1)转录条形码化第一复合物以生成RNA转录物,

(2)使用退火到启动子序列的引物来逆转录RNA转录物以生成条形码化第一复合物的cDNA链,

(3)将cDNA环化以产生环化的cDNA链,

(4)通过滚环扩增将环化的cDNA链扩增,以及

(5)使用扩增的cDNA链作为模板来合成双链或部分双链的条形码化第一复合物。

[0017] 如权利要求1-13中任一项所述的方法,通过以下方式扩增条形码化第一亚片段的至少一部分:

(1)从第一珠中释放条形码化第一亚片段,

(2)使已释放的条形码化第一亚片段变性以形成单链条形码化第一复合物,

(3)将单链条形码化第一亚片段环化,

(4)进行滚环扩增,以及

(5)使用扩增的单链条形码第一复合物作为模板来合成双链条形码化第一亚片段。

[0018] 在一些方式中,该方法包括将权利要求14中的环化的cDNA链或权利要求15中的单链条形码化第一亚片段分级以选择大小在预定范围内的环。

[0019] 在一些方式中,扩增条形码化第一亚片段的至少一部分包括:

(1)使用结合到第一捕获寡核苷酸上的引物结合序列的引物来延伸条形码化第一亚片段,

(2)从第一珠中释放来自(1)的延伸的条形码化第一亚片段,

(3)使用单引物扩增将条形码化第一亚片段扩增约10-120倍,由此产生扩增的双链条形码化第一亚片段,以及

(4)将衔接子寡核苷酸接合至扩增的双链条形码化第一亚片段的末端。

在一些方式中,将至少一些第二亚片段中的每个连接至第二插入寡核苷酸,并且其中步骤(e)还包括:

(1)将第二捕获寡核苷酸接合至第二插入寡核苷酸,或

(2)将第二捕获寡核苷酸与第二插入寡核苷酸杂交,然后通过DNA聚合酶延伸插入寡核苷酸以并入第二条形码,

其中衔接子寡核苷酸具有与第二插入寡核苷酸相同的序列。

[0020] 如前述权利要求中任一项所述的方法,靶核酸片段的大小在10kb-100M(mega)碱基的范围内。

[0021] 如前述权利要求中任一项所述的方法,第一亚片段的大小在1kb-20kb的范围内。

[0022] 一种用于分析转录物的方法,该方法包括:

(a)在单一混合物中,将来自一个或多个细胞的mRNA与第一珠的群组组合,其中每个第一珠包含固定在其上的第一捕获寡核苷酸,其中第一捕获寡核苷酸包含共同引物序列、第一条形码序列、UMI和寡核苷酸dT序列,其中第一珠的数量大于单一混合物中mRNA分子的数量,

(b)在包含三核苷酸GGG和共同引物序列的衔接子模板的存在下,逆转录捕获的RNA,以产生cDNA/mRNA杂交分子,其中cDNA/mRNA杂交分子各自包含由捕获的RNA的逆转录产生的cDNA和第一捕获寡核苷酸,

(c)从第一珠中释放cDNA,

(d)扩增来自(c)的cDNA并产生双链或部分双链cDNA;

(e)将交错的单链断裂引入至少一些扩增的cDNA以生成第二复合物,该第二复合物各自包含多个第二亚片段,以及

(f)将第二捕获寡核苷酸引入多个第二亚片段,

其中每个第二捕获寡核苷酸包含：

(1)任选地启动子序列或引物结合序列，以及

(2)第二条形码，其中固定在相同个体珠上的第二捕获寡核苷酸包含相同的第二条形码，并且大多数珠具有不同的第二条形码，由此提供条形码化第二亚片段。

[0023] 在一些方式中，其中单一混合物中的mRNA来自单一细胞。

[0024] 在一些方式中，在步骤(d)中扩增cDNA是通过使用与共同引物序列杂交的引物进行滚环扩增。

[0025] 在一些方式中，将至少一些第二亚片段中的每个连接至第二插入寡核苷酸，并且其中步骤(f)还包括：

(1)将第二捕获寡核苷酸接合至第二插入寡核苷酸，或

(2)将第二捕获寡核苷酸与第二插入寡核苷酸杂交，然后通过DNA聚合酶延伸插入寡核苷酸以并入第二条形码。

[0026] 在一些方式中，在步骤(d)中扩增cDNA包括通过以下来扩增cDNA：使用与共同引物序列杂交的引物进行单引物扩增，并将衔接子寡核苷酸接合至扩增的第一复合物的末端，其中衔接子寡核苷酸与插入寡核苷酸具有相同的序列。

[0027] 如前述权利要求中任一项所述的方法，其中第一插入寡核苷酸中的每个包含第一位置条形码，其中不同的第一插入寡核苷酸包含不同的第一位置条形码，以及/或

其中第二插入寡核苷酸中的每个包含第二位置条形码，其中不同的第二插入寡核苷酸包含不同的第二位置条形码。

[0028] 在一些方式中，一个或多个第一位置条形码与一个或多个第二位置条形码相同。

[0029] 一种分析一个或多个靶区域的全长序列的方法，该方法包括：

(a)扩增每个靶区域，

(b)将衔接子寡核苷酸接合至包含靶区域的扩增的核酸靶片段的两端，

(c)将交错的单链断裂引入来自(b)的至少一些扩增的核酸片段中，以产生多个第一复合物，该第一复合物各自包含多个第一亚片段，

(d)将第一捕获寡核苷酸引入第一亚片段中的至少一些中，其中每个第一捕获寡核苷酸包含：

(1)任选地启动子序列或引物结合序列，以及

(2)第一条形码，其中固定在相同个体珠上的第一捕获寡核苷酸包含相同的第一条形码，并且大多数珠具有不同的第一条形码，由此提供条形码化第一亚片段。

[0030] 在一些方式中，

在步骤(a)中扩增靶区域包括：

[0031] 用正向引物和反向引物扩增靶区域，两者均包含对靶区域特异的序列，由此产生扩增区域，

其中正向引物包含共同序列和UMI，以及第一靶特异性序列，并且其中正向引物包含特殊碱基，其中特殊碱基不存在于天然DNA分子中并且特殊碱基可被试剂降解；

通过添加试剂降解过量的正向引物；

使用与共同序列和反向引物结合的引物扩增(i)中的扩增区域，由此产生包含UMI的进一步扩增的区域。

[0032] 在一些方式中,特殊碱基是尿嘧啶并且试剂能够特异性裂解含尿嘧啶的寡核苷酸。

[0033] 在一些方式中,步骤(a)中扩增靶区域还包括:

(4)接合两个寡核苷酸,每个接合至来自(3)的进一步扩增的区域的一端,其中两个寡核苷酸共享相同的共同序列,并且其中两个寡核苷酸包含不同的UMI,由此产生在两端具有共同序列的接合DNA产物,以及

(5)用与共同序列杂交的引物扩增来自(4)的接合DNA产物,由此产生包含靶区域的扩增核酸靶片段。

[0034] 如前述权利要求中任一项所述的方法,还包括将3'分支接合衔接子寡核苷酸接合至第二亚片段,其中接合衔接子寡核苷酸是3'分支接合,并且其中衔接子寡核苷酸包含第二PCR引物退火位点。

[0035] 在一些方式中,3'分支接合衔接子寡核苷酸是平端衔接子并且3'分支接合包括将来自平端衔接子的5'磷酸酯共价连接到第一片段的切口处的凹陷的3'羟基。

[0036] 在一些方式中,第一PCR引物退火位点和第二PCR引物退火位点具有不同的序列。

[0037] 在一些方式中,3'分支接合衔接子寡核苷酸包含条形码序列,其任选地是样本条形码序列。

[0038] 在一些方式中,引入第一插入寡核苷酸、引入第二插入寡核苷酸或两者是通过转位酶进行的,并且插入寡核苷酸转位酶保持与第一亚片段结合或插入寡核苷酸保持与第二亚片段结合,或两者。

[0039] 在一些方式中,该方法包括去除转位酶由此分离个体第一亚片段、第二亚片段或两者。

[0040] 在一些方式中,还包括使用退火到第一和第二PCR引物退火位点的引物扩增第一亚片段、第二亚片段或两者以产生扩增子。

[0041] 在一些方式中,第一插入寡核苷酸中的至少一些各自包含位置条形码,其中不同的第一插入寡核苷酸包含不同的位置条形码。

[0042] 在一些方式中,其中第二插入寡核苷酸中的至少一些各自包含位置条形码,其中不同的第二插入寡核苷酸包含不同的位置条形码。

[0043] 一种用于制备用于对靶核酸进行测序的测序文库的方法,该方法包括:

(a)将衔接子和一个或多个UMI副本接合至靶核酸的片段,

(b)将已与衔接子和一个或多个UMI副本接合的片段变性以形成单链DNA分子,

(c)对至少一些单链分子进行滚环扩增以产生包含至少一种单链分子的大于5x副本的纳米球,

(d)将纳米球转化为双链或部分双链DNA分子,

(e)将交错的单链断裂引入(d)中的DNA分子,由此生成第一复合物,该第一复合物各自包含多个第一亚片段,

(f)将第一捕获寡核苷酸引入第一亚片段中的至少一些中的每个中,其中每个第一捕获寡核苷酸包含:

(1)任选地启动子序列或引物结合序列,以及

(2)第一条形码,其中固定在相同个体珠上的第一捕获寡核苷酸包含相同的第一

条形码，并且大多数珠具有不同的第一条形码，由此提供条形码化第一亚片段。

在一些方式中，将第一亚片段中的至少一些中的每个连接至第一插入寡核苷酸，其中步骤(b)还包括：

(1)将第一捕获寡核苷酸接合至第一插入寡核苷酸，或

(2)将第一捕获寡核苷酸与第一插入寡核苷酸杂交，然后通过DNA聚合酶延伸插入寡核苷酸以并入第一条形码。

[0044] 在一些方式中，插入寡核苷酸中的每个包含位置条形码，其中不同的插入寡核苷酸包含不同的位置条形码。

[0045] 一种将寡核苷酸插入靶核酸片段的方法，该方法包括：

(a)将交错的单链断裂引入片段中，

(b)将来自(a)的片段与插入支架接触，其中衔接子锚定至支架并以预定间距分开，其中插入支架包括多个双链或部分双链衔接子和支架，

其中每个衔接子包括包含独特位置条形码的插入寡核苷酸，以及

其中接触导致多个插入寡核苷酸在单链断裂处被引入片段中，由此产生第一插入复合物，该第一插入复合物各自包含多个第一亚片段。

[0046] 在一些方式中，该方法还包括：

将支架与已插入靶核酸中的多个衔接子解离。

[0047] 在一些方式中，该方法包括：

将多个支架与一些核酸片段中的每个接触，其中不同支架中的衔接子具有不同的支架条形码。

[0048] 在一些方式中，

支架是单链核酸分子，并且

其中每个衔接子还包括：

支架杂交序列，

其中转位子经由支架杂交序列与支架杂交，以及

其中支架杂交序列可以被切割以使支架与衔接子解离，

其中插入寡核苷酸还包含支架内所有衔接子共享的支架条形码。

[0049] 在一些方式中，该方法还包括：

(c)将(i)由(a)产生的第一插入复合物与(ii)第一珠的群组组合成单一混合物，其中每个第一珠包含固定在其上的第一捕获寡核苷酸的多个副本，所述第一捕获寡核苷酸包含第一条形码，其中固定在相同的个体第一珠上的第一捕获寡核苷酸包含相同的第一条形码并且大多具有不同的第一条形码，

(d)对于多个第一亚片段中的每个，引入第一捕获寡核苷酸，由此产生条形码化第一亚片段，该条形码化第一亚片段各自连接至第一条形码的副本。

在一些方式中，该方法还包括：

(e)扩增多个条形码化第一亚片段，

(f)将交错的单链断裂引入扩增的条形码化亚片段中，以及

(g)将来自步骤(f)的产物与第二插入支架接触，其中第二插入支架各自包含锚定至第二插入支架的多个第二衔接子，由此将第二衔接子上的第二插入寡核苷酸引入扩增的

条形码化第一亚片段以产生第二插入复合物，该第二插入复合物各自包含多个第二亚片段。

[0050] 在一些方式中，第二插入支架的支架是单链核酸分子，并且每个第二衔接子包含：

(1)第二支架杂交序列，

其中第二衔接子经由第二支架杂交序列与支架杂交，以及

其中第二插入支架杂交序列可以被切割以使支架与第二衔接子解离；

(2)包含独特位置条形码的第二插入寡核苷酸，以及

由第二插入支架的支架内的第二衔接子共享的第二支架条形码。

[0051] 在一些方式中，将第二插入复合物与第二珠的群组混合，其中每个珠包含固定在其上的第二捕获寡核苷酸，所述寡核苷酸包含第二条形码，其中固定在相同个体珠上的寡核苷酸包含相同的第二条形码并且大多数珠具有不同的第二条形码，

(g)对于第二插入复合物中的至少一些中的每个，引入包含第二条形码的第二捕获寡核苷酸的多个副本，其中多个副本来自单个珠，并由此产生多个条形码化第二亚片段，该条形码化第二亚片段各自连接至第二条形码的至少一个副本。

[0052] 在一些方式中，该方法还包括：

对多个第二亚片段进行测序以产生多个测序读段。

[0053] 在一些方式中，插入支架具有1-50kb的大小。

[0054] 在一些方式中，相邻的第一衔接子之间的预定间隔在3kb到5kb的范围内。

[0055] 在一些方式中，相邻的第二衔接子之间的预定间隔在200bp到1000bp的范围内。

[0056] 在一些方式中，多个插入支架的长度总和等于或大于靶核酸的长度。

[0057] 多个插入支架，

其中多个插入支架中的每个包含：

(1)多个衔接子，其中衔接子是双链或部分双链的，以及

(2)支架，并且衔接子锚固至支架并以预定间距分开，

其中对于每个插入支架，插入支架中的每个衔接子带有独特位置条形码和共同的支架条形码，以及

其中不同插入支架中的衔接子具有不同的支架条形码。

[0058] 在一些方式中，至少一个插入支架的大小等于1-10kb的多核苷酸的大小。

[0059] 在一些方式中，至少一个插入支架的大小等于10-50kb的多核苷酸的大小。

[0060] 核酸复合物，其包含本文公开的多个插入支架，以及核酸片段，其中所述多个插入支架与靶核酸片段杂交。

[0061] 在单个容器中的反应混合物，其中反应混合物包含本文公开的多个插入支架和源自靶核酸的多个片段。

附图说明

[0062] 附图及其描述说明了本发明的示例性实施方案。本公开中提供的发明不限于这些附图中所示的实施方案。

[0063] 图1示出了由Tn5插入生成的基因组DNA片段的电泳。

[0064] 图2A示出了在包含固定的捕获寡核苷酸的珠上捕获DNA片段的示例性方法。包含

插入寡核苷酸(“第一插入寡核苷酸”)的转位子首先通过转位酶整合到双链或部分双链的靶核酸片段中以形成第一复合物。转位酶切开靶核酸但保持附着于靶核酸。所得片段(“第一片段”)与珠一起温育,每个珠包含多个捕获寡核苷酸(图2A中仅示出其中两个)。每个捕获寡核苷酸包含T7启动子或PCR引物结合序列①、独特条形码序列②(“组合珠条形码”)、独特分子标识符(UMI)③和共同的序列⑤。每个转位子包含杂交序列⑥、包含杂交序列⑥的插入寡核苷酸⑦。相同珠上的捕获寡核苷酸共享相同的独特条形码序列。夹板寡核苷酸④包含与共同序列⑤杂交的部分和与杂交序列⑥杂交的部分。一个转位子和侧翼转位子之间的第一复合物的部分称为第一亚片段⑧。如本文所用,“侧翼转位子”是指最接近参考转位子的转位子。如果参考转位子具有两个侧翼转位子(一个上游和一个下游),则两者都是侧翼转位子,也称为相邻转位子。如在本公开中使用的,如在“第一复合物”或“第二复合物”以及“第一插入复合物”或“第二插入复合物”中使用的复合物是指一系列单寡核苷酸链,其相互连接以形成具有交错的单链断裂的部分双链DNA结构。在一些情况下,这些单寡核苷酸链通过与这些链末端结合的酶(例如转位酶分子)相互连接,如图2A所示。在一些情况下,这些单寡核苷酸链通过重叠区域相互连接,如图28A所示(“I”)。因此,根据上下文,在一些实施方案中,第一复合物还可包含以规则间隔插入的酶(转位酶)和/或插入寡核苷酸,见图28A(“II”)。如本公开中所公开的,在“第一亚片段”或“第二亚片段”中使用的亚片段是指单寡核苷酸链,其是双链或部分双链DNA的一部分。图2B示出了在将捕获寡核苷酸与转位子中的插入寡核苷酸接合后形成的DNA分子。

[0065] 图3示出了本发明的说明性实施方案,其中捕获寡核苷酸被接合至插入的基因组片段并且片段通过切口平移延伸,使得一个捕获寡核苷酸被添加到每个第一亚片段⑧的每个末端。进行体外转录以便以线性方式扩增新形成的DNA片段。

[0066] 图4示出了本发明的说明性实施方案,其中用退火到T7启动子区域的引物对从图3中的方法步骤生成的转录物进行逆转录以产生cDNA。随后使用与共同序列结合的引物合成第二链以产生双链DNA分子。

[0067] 图5示出了作为图3中方式的替代方式,其中从图2B产生的片段通过切口平移延伸,并且延伸的片段从珠中释放。使用与PCR引物结合序列①结合的引物进行单引物长距离PCR以产生双链DNA。

[0068] 图6A和6B示出了一种方式,其中将衔接子寡核苷酸⑨分别接合至如图4和5所示的双链DNA分子的末端。

[0069] 图7A和7B示出了一种方式,其中插入寡核苷酸(“第二插入寡核苷酸”)通过转位子被分别引入与由图6A和6B中的步骤产生的衔接子寡核苷酸接合的双链DNA中。第二插入寡核苷酸具有与图6A和6B所示的衔接子寡核苷酸⑨相同的序列。

[0070] 图8示出了一种方式,其中全长mRNA被捕获在用独特珠条形码②、随机分子条形码(UMI)③、共同引物结合序列①和寡核苷酸dT(10)固定的珠上。

[0071] 图9示出了一种方式,其包含:经由通过逆转录酶添加的三核苷酸CCC尾和包含三核苷酸GGG的衔接子模板(12),将衔接子寡核苷酸(13)逆转录和并入到cDNA分子(11)中。

[0072] 图10示出了一种方式,其包含:从珠释放分子并用识别共同引物结合序列①的共同引物进行PCR。扩增倍数可以是10到100万或更多,具体取决于所要求的每个转录物的所需覆盖率。

[0073] 图11示出了一种方式,其中将衔接子寡核苷酸⑨接合至由图10所示的方法步骤产生的dsDNA分子的末端。

[0074] 图12示出了一种包含将插入寡核苷酸转位到图11所示的dsDNA分子中的方案。dsDNA分子在两端具有衔接子寡核苷酸。插入寡核苷酸具有与衔接子寡核苷酸⑨相同的序列。

[0075] 图13示出了一种方式,其中通过使用正向和反向引物扩增靶区域来产生扩增的PCR产物。正向引物包含与靶区域中的第一序列(“靶特异性序列1”)互补的靶特异性序列,而反向引物包含与靶区域中的第二序列(“靶特异性序列2”)互补的靶特异性序列。正向引物还包含UMI③和共同序列①。第一引物中的靶特异性序列包含多个尿嘧啶代替胸苷。

[0076] 图14示出了一种方式,其中用酶尿嘧啶特异性切除试剂(USER)处理反应混合物以去除过量的正向引物,并使用与共同序列①结合的正向引物和与靶区域中的第二序列结合的反向引物进一步扩增靶区域。

[0077] 图15示出了一种方式,其中将衔接子寡核苷酸⑨接合至图14中扩增产物的两端。

[0078] 图16示出了一种包括扩增靶区域的方式作为图13所示的方法的替代方式。图16中的方法使用图13中的正向引物和反向引物扩增靶区域。在有限数量的PCR循环中扩增靶区域。

[0079] 图17示出了一种方式,其包含将共同序列寡核苷酸(14)接合至由图16中的方法步骤产生的扩增产物的每个末端。每个共同序列寡核苷酸在5’端包含共同序列并且在3’端包含UMI。

[0080] 图18示出了包含用退火到共同序列的单引物扩增来自图17的接合产物的方式。扩增的接合产物在两端与衔接子寡核苷酸接合。

[0081] 图19示出了包含将插入寡核苷酸转位到从图18中的方法步骤获得的产物中的方式,该插入寡核苷酸具有与衔接子寡核苷酸相同的序列。

[0082] 图20示出了作为本发明的说明性实施方案的条形码序列组装。需要使用三种接合来生成～36亿不同的条形码。显示条形码组装每一步的预期序列。

[0083] 图21示出了条形码序列组装程序的说明性实施方案。

[0084] 图22A示出了作为本发明的说明性实施方案的环化单链DNA。图22B示出了对环进行滚环扩增(“RCR”)。图22C示出了将来自RCR反应的扩增产物转化为双链DNA或部分双链DNA(未示出)。在一些实施方案中,扩增产物还可包含分支结构(未示出)。图22D示出了转位子插入双链DNA。

[0085] 图23图示了包含使用插入支架S1,S2,...,Sn来使靶核酸转位的方式。衔接子(例如转位子元件)中的每个包括两个位置条形码(例如,1和2、3和4等)和支架独有的支架条形码(数字上方的垂直条是指位置条形码)。图23图示了,在重复区域中,个体转位子条形码可有助于对重复序列进行排序。1-50kb支架条形码可有助于处理更远距离的重复区域。相邻转位子和支架条形码之间的固定间距可有助于确定重复区域的大小。图23至25B所示的转位子也可以用任何双链或部分双链衔接子代替以执行本文公开的方法。

[0086] 图24示出了示例性转位子支架的构造。转位子经由支架杂交序列与支架杂交,从而固定转位子沿支架的顺序。

[0087] 图25A图示了具有位置条形码的转位子的第一轮插入。在该示例中,转位子产生了

3-5kb大小的分子。这些分子被捕获在stLFR珠上并进行共条形码化(未示出)、环化、RCR扩增,并转化为双链或部分双链DNA分子。

[0088] 图25B图示了在图25A中的第一轮插入之后执行第二轮插入的过程。在图25A所示的过程中产生的双链DNA分子可以使用以较短的间距将转位子锚定在其上的第二转位子支架再次转位。第二轮转位导致相邻转位子之间约为500bp。对由stLFR珠引入的条形码的分析可以产生来自第一轮(产生3-5kb的第一亚片段)和第二轮转位(产生约500bp的第二亚片段)的位置信息。此外,转位子的插入导致相邻亚片段之间有9bp的重叠,该信息还可用于对支架之间的亚片段进行排序。(我们也可以通过将衔接子接合至缺口DNA来执行此策略。我们不想只针对转位子插入。)

[0089] 图26示出了一种方式,其包含使用位置条形码层构建100kb基因组DNA分子的序列信息,并且每次插入转位子都会产生9bp的重叠。使用长度范围从10到50kb的转位子支架进行第一轮转位。使用长度范围从3到5kb的转位子支架进行第二轮转位。

[0090] 图27A和27B示出了一种方式,其中首先通过切口酶在靶片段中产生切口。使用Klenow片段或任何其他具有3'-5'核酸外切酶活性的酶扩大切口以形成单链缺口。得到的DNA片段通过3'分支接合而接合至珠上的部分双链捕获寡核苷酸衔接子。在图27A中,与捕获寡核苷酸互补的衔接子链被接合至DNA片段。在图27B中,包含捕获寡核苷酸的衔接子链被接合至DNA片段的3'末端。

[0091] 图28A和图28B示出了根据本公开的方案I的实施方案。图28A示出了第一轮stLFR,其中靶核酸被片段化以产生多个双链靶核酸片段。在一个方式("I")中,将交错的单链断裂引入片段中。在替代方式("II")中,通过转位酶/转位子引入交错的单链断裂(缺口)。尽管仅示出了从任一方式产生的一个第一复合物,但应理解以相同方式处理多个片段并将产生多个第一复合物。第一复合物各自包含多个第一亚片段。第一复合物与具有包含第一条形码的第一捕获寡核苷酸的多个副本的第一珠组合。如本公开所述,将第一条形码序列引入第一亚片段中的每个中以产生多个条形码化第一亚片段。将这些条形码化第一亚片段扩增,从而产生双链或部分双链的第一亚片段。然后将衔接子寡核苷酸添加到两端。图28B示出了第二轮stLFR,其中用携带第二插入寡核苷酸(插入寡核苷酸未在该图中示出)的转位子使双链第一亚片段转位。类似于图28A中的第一轮stLFR,该步骤也可以使用缺口酶和衔接子以引入插入寡核苷酸来进行。将各自包含多个第二亚片段的第二复合物(仅示出一个)与携带第二条形码的第二珠组合,这导致第二条形码序列被添加到第二亚片段,从而产生条形码化第二亚片段。

具体实施方式

I.概述

[0092] 第一代stLFR使用微珠表面作为常规条形码化反应中使用的隔室(例如,384孔板的孔)的替代物。每个珠携带被转移到每个长DNA分子的亚片段中的独特条形码序列的许多副本。由于每个长DNA分子都携带相同条形码的多个副本,因此该过程被称为"共条形码化"。因此,第一代stLFR允许基于在单个管中生成的共条形码化亚片段的测序读段构建长DNA片段的序列信息。第一代stLFR被描述于共同未决的PCT专利公开号WO 2019/217452中,该PCT专利公开的全部内容通过引用并入本文。

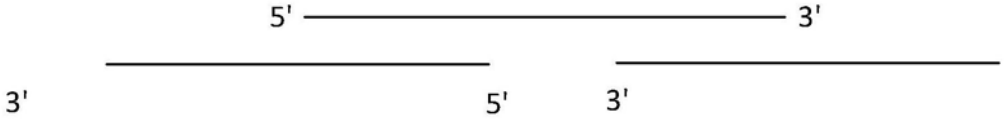
[0093] 这里我们描述了高覆盖率单管长片段读取(stLFR)技术,其使用多个条形码和UMI的组合以允许与第一代stLFR相比覆盖更多长DNA片段的序列构建。该技术对在共条形码化之前已被扩增的靶DNA片段执行stLFR,这为测序提供了更高的DNA量并增加了测序覆盖率。在一些实施方案中,本文所述的高覆盖率stLFR使用两轮stLFR。第一轮stLFR通过使用第一珠插入或接合靶DNA片段独有的第一条形码的许多副本来标记靶DNA片段(例如,长基因组DNA片段)。在一些实施方案中,UMI还与第一条形码一起被添加到靶DNA片段。添加不频繁,使得生成(位于两个相邻的第一条形码之间)的每个第一亚片段相对较大,例如10kb。然后扩增第一亚片段。在每个第一亚片段内,第二轮stLFR以高于第一轮标记的频率引入第二条形码以产生第二复合物。由于标记频率较高,位于相邻第二条形码之间的第二亚片段小于第一亚片段。第二亚片段的长度适于某些测序方法,例如约100-1000个碱基。可以基于相同的第二条形码组合来自第二亚片段的序列读段以构建每个第一亚片段的序列;并且可以将来自每个第一亚片段的构建序列组合起来构建原始的靶DNA分子,其可以是很长(例如约40kb-400kb)的基因组片段。由于这种额外的条形码化,这种改进的stLFR将覆盖率提高了至少10倍,并可用于构建非常长的靶核酸分子的序列。

[0094] 在一些实施方案中,高覆盖率stLFR可用于分析单个细胞的转录组。一种方式是,将细胞在合适的容器(例如,液滴或孔)中稀释,使得大多数隔室仅包含一个或零个细胞。细胞被裂解并且包含来自单个细胞的mRNA的细胞裂解物与具有独特第一条形码、UMI和dT寡核苷酸的珠混合以捕获来自单个细胞的mRNA。然后将mRNA逆转录以产生cDNA,并使用常规方法从cDNA中产生双链或部分双链DNA分子。然后对扩增的双链DNA进行stLFR,这添加了独特第二条形码的许多副本。在这些实施方案中,转录物的序列可以基于第一条形码分配给单个细胞(每个第一条形码对应于单个细胞)并且可以基于UMI将序列分配给每个mRNA。第二条形码用于通过组合具有相同第二条形码的序列读段从每个长转录物构建序列。

[0095] 在一些实施方案中,高覆盖率stLFR在进行stLFR之前使用靶特异性引物来扩增靶区域。在一种方式中,PCR用于扩增靶区域。PCR引物包含共同序列并且一个或两个PCR引物可以包含UMI。UMI的存在使用户能够将PCR过程产生的错误与序列本身的突变区分开来,从而提高变异调用的可信度。

II.定义

[0096] 术语“交错的单链断裂”是指引入DNA分子(双链或部分双链)的断裂(由核苷酸的切口、缺口和/或缺失产生),从而导致多个单链DNA分子。对于至少一些单链DNA分子,5’序列的一部分与另一个单链DNA分子的5’序列的至少一部分互补,并且3’末端的至少一部分与另一个单链DNA分子的3’序列的至少一部分互补,使得在杂交条件下多个单链DNA分子彼此杂交以形成核酸复合物。为了说明而非限制,下文说明了包含具有交错单链断裂的三个单链DNA分子的核酸复合物。应当理解,核酸复合物(或“复合物”)可以并且通常确实包含多于三个的单链DNA分子。



[0097] 一对单链DNA分子之间杂交区域的长度可以在例如20个碱基对到1000个碱基对或更多的范围内。非杂交区域(例如,上文上DNA分子的未杂交区域),如果有的话,可以在例如

下限为至少1个碱基、至少2个碱基、至少3个碱基、至少5个碱基、至少10个碱基或至少15个碱基，以及上限为20个碱基、100个碱基、1000个碱基或超过1000个碱基的范围内。在一些情况下，非杂交区域是零个碱基(例如，链被切开而没有核苷酸缺失)。术语“交错的双链断裂”具有相同的含义。

[0098] 术语“部分双链”是指两条DNA链彼此杂交并且一条链的至少一部分未与另一条链杂交。部分双链DNA的两条DNA链的长度可能不同，也可能相同。

[0099] 如本文所用，术语“转位子”是指被转位酶或整合酶识别并且能够通过转位酶插入到DNA分子中的核酸片段。

[0100] 如本文所用，术语“转位酶”具有其在本领域中的通常含义并且是指结合转位子末端并催化其插入多核苷酸中的酶(例如，通过剪切和粘贴机制或复制转位机制)。

[0101] 如本文所用，“独特分子标识符”(UMI)是指存在于DNA分子中的核苷酸序列，其可用于将个体DNA分子彼此区分开。见，例如，Kivioja,Nature Methods 9,72-74(2012)。UMI可以与它们关联的DNA序列一起测序，以识别来自相同源核酸的序列读段。术语“UMI”在本文中用于指代UMI的核苷酸序列和物理核苷酸，这将从上下文中显而易见。

[0102] UMI可以是随机的、伪随机的或部分随机的或非随机的核苷酸序列，其被插入到衔接子中或以其他方式并入待测序的源核酸(例如，DNA)分子中。在一些实现方式中，期望每个UMI唯一地识别样本中存在的任何给定的源DNA分子。

[0103] 如本文所用，术语“单管LFR”是指在例如美国专利公开2014/0323316中描述的过程，该美国专利公开的全部内容通过引用以其整体并入本文，其中，特别地，相同的独特条形码序列的多个副本与单个长核酸片段关联，例如通过接合关联。在单管LFR的典型实施方案中，长核酸分子以规则间隔用“插入寡核苷酸”标记。在一个实施方案中，通过一种或多种酶(例如转位酶、切口酶、接合酶)将插入寡核苷酸引入长核酸分子中。不同长核酸片段之间的条形码序列是不同的。因此，标记个体长核酸片段的过程可以在单个容器中方便地进行。此过程允许分析大量个体DNA片段，而无需在加标签步骤期间将片段分离到单独的管、容器、等分试样、孔或液滴中。

[0104] 如本文所用，“独特”条形码是指与个体珠关联并可用于区分个体珠的核苷酸序列。在各自都具有独特条形码的珠的群组中，与一个珠关联的条形码序列不同于该群组中至少90％珠、更常见是该群中至少99％珠，甚至更常见是该群体中至少99.5％珠，并且最常见是该群体中至少99.9％珠的条形码序列。

[0105] 如本文所用，除非另有说明或从上下文中清楚可见，否则有关核酸序列使用(例如，如在“相邻条形码”、“相邻转位子”、“相邻切口”、“相邻断裂”、“相邻插入寡核苷酸”等中使用)的术语“相邻”(与“侧翼”可互换使用)是指一系列空间分离的核酸序列中的两个最接近的核酸序列。例如，术语“相邻插入寡核苷酸”是指在stLFR过程期间并入靶核酸片段中的两个最接近的插入寡核苷酸，即，在这两个寡核苷酸之间没有插入寡核苷酸。

[0106] 有关两个或更多个核酸序列使用的术语“连接”(如在例如第一亚片段连接至第一条形码中使用的)是指两个核酸序列直接或间接相连。连接可以通过本领域熟知的多种方法(例如通过接合或通过合成)完成。

III.方法的示例性实施方案

[0107] 高覆盖率stLFR可以根据各种方案来执行。以下是方法的示例性实施方案和这些

实施方案的变形。受本公开指导的分子生物学和测序领域的技术人员将认识到可以并入以下方案中的许多变形。

[0108] 在单管LFR的典型实施方案中，将交错的单链断裂引入长核酸分子(例如，靶核酸的片段)并且将插入寡核苷酸在断裂处插入。在一个实施方案中，插入寡核苷酸和交错的单链断裂由一个或多个转位子引入。在另一个实施方案中，交错的单链断裂由缺口酶(例如切口酶、klenow)引入，并且插入寡核苷酸通过分支接合反应在交错的单链断裂处接合至片段。具有插入寡核苷酸的长核酸分子与珠接触，在该珠上固定有捕获寡核苷酸的多个副本，并且每个捕获寡核苷酸都包含该珠独有的条形码。条形码序列可以通过多种方式转移到长核酸片段上。在一个方式中，它通过插入寡核苷酸、夹板寡核苷酸和捕获寡核苷酸上的共同序列之间的杂交进行转移，其中夹板寡核苷酸的一部分与插入寡核苷酸的一部分杂交，而夹板寡核苷酸的另一部分与捕获寡核苷酸上的共同序列杂交。在另一方式中，通过捕获寡核苷酸和插入寡核苷酸之间的杂交将条形码序列转移到长核酸片段。然后将插入寡核苷酸延伸以产生条形码序列的副本。

方案I.用于以大覆盖率测序的长核酸片段的制备

[0109] 来自样本的基因组DNA或双链cDNA被片段化以生成靶核酸的片段，然后对这些片段中的至少一些进行第一轮stLFR。每个靶核酸片段的第一轮stLFR产生多个第一亚片段，这些第一亚片段中的至少一些中的每个被连接至第一条形码的副本(“条形码化第一亚片段”)。条形码化第一亚片段被扩增，并且对扩增产物进行第二轮stLFR。第二轮stLFR产生第二亚片段，这些第二亚片段中的至少一些中的每个被连接至第二条形码的副本(“条形码化第二亚片段”)。第二亚片段中的至少一个包含第二条形码和第一条形码。对第二亚片段进行测序，并且可以将来自具有相同第二条形码的第二亚片段的序列读段组装在一起以识别具有第一条形码的第一亚片段的序列。应当理解，该步骤中的序列读段将包括第二条形码序列。然后，可以将来自具有相同第一条形码的第一亚片段的序列组装在一起以产生靶核酸片段的序列信息。以这种方式，可以确定长DNA片段的序列。该实施方案的图示在图28A和28B中示出。尽管这里描述为分两步组装序列读段的过程，其中第一步组装基于第一条形码且第二步组装基于第二条形码，但应当理解，两个序列组装步骤可以同时和/或多次迭代进行。方法的说明性实施方案的细节描述如下。

1.第一轮stLFR中的第一插入

[0110] 在一个方式中，将靶核酸片段与转位子和转位酶一起温育。替代地，在相关方式中，靶核酸与缺口酶--在双链DNA中产生交错单链断裂的酶(例如，在没有游离核苷酸下提供的切口酶或DNA聚合酶)--以及接合酶和包含插入寡核苷酸的衔接子组合。这导致将插入寡核苷酸(例如，图2A和图2B中的⑦)引入靶核酸中。这些添加事件产生第一复合物，其包含插入到靶核酸片段中的插入寡核苷酸的多个副本。通过控制反应条件，例如捕获寡核苷酸、珠、靶核酸的比率，可以控制插入之间的距离，从而使插入序列以近似规则的间隔定位。

[0111] 由其生成第一复合物的靶核酸片段的长度可以在很宽的范围内变化。在一些实施方案中，靶核酸片段的长度可为10kb-100M碱基大小，例如10kb-50kb，20kb-100kb，20kb-300kb，100-200kb，100kb-500kb，或300kb-5000kb大小。如上所述，转位子和转位酶或缺口酶的浓度受到控制，使得插入寡核苷酸的添加很少发生，在相邻插入寡核苷酸之间留下间隔。靶核酸片段中的间隔(每个由两个相邻插入寡核苷酸定义)被称为第一亚片段(例如，图

2B中的⑧)。第一亚片段的长度通常在1kb-20kb(例如1kb-10kb,1kb-5kb,或3kb-15kb)的范围内。通过转位引入插入寡核苷酸在靶核酸片段中产生切口,并且转位酶保持结合并连接由插入事件产生的第一亚片段。在一些实施方案中,序列的第一插入发生在溶液中,即插入寡核苷酸不附着于任何固体支持物。在其他实施方案中,插入寡核苷酸附着于固体支持物,例如微米大小的磁珠。

2.第一捕获

[0112] 上面产生的第一亚片段可以使用其上固定有第一捕获寡核苷酸的许多副本的珠捕获。见图2A。每个第一捕获寡核苷酸可包含第一条形码和UMI。第一条形码是每个珠特有的,即同一珠的捕获寡核苷酸中的所有第一条形码都具有相同的序列。UMI是随机或半随机序列,其在同一珠上的第一捕获寡核苷酸之间的序列不同。第一捕获寡核苷酸还可包含T7启动子或PCR引物结合位点。在一个方式中,第一捕获寡核苷酸还包含第一共同序列(⑤),其与夹板寡核苷酸(④)的一部分杂交。夹板寡核苷酸包括可以与插入寡核苷酸中的杂交序列(⑥)杂交的另一部分。添加DNA接合酶以将捕获寡核苷酸接合至第一复合物中的插入寡核苷酸。这些步骤导致第一条形码转移到第一复合物中的第一亚片段并产生条形码化第一亚片段,该条形码化第一亚片段各自包含至少一个条形码序列(例如,一个或两个第一条形码序列)。在另一方式中,捕获寡核苷酸与连接至第一亚片段的插入寡核苷酸杂交。然后延伸插入寡核苷酸以产生条形码序列。

3.在一个步骤中进行第一插入和第一捕获

[0113] 替代地,上述步骤1和2可以合并为一个步骤。换句话说,将寡核苷酸插入到靶核酸的片段中和将第一亚片段捕获到珠上是同时发生的。在一个方式中,插入寡核苷酸与固定在珠上的捕获寡核苷酸杂交以形成适于分支接合的部分双链分子,并且插入寡核苷酸的引入(通过分支接合反应)发生在珠的表面上。见图27A。在另一方式中,插入寡核苷酸与固定在珠上的捕获寡核苷酸杂交以形成适于分支接合的部分双链分子,并且捕获寡核苷酸接合至第一亚片段。

[0114] 通常,在这种在一个步骤中进行插入和第一捕获的方式中,使用包含例如切口酶和Klenow的缺口酶将交错的单链断裂引入靶核酸的片段中,由此产生多个第一亚片段并且由单个核酸片段生成的所有第一亚片段统称为第一复合物。该方法还包括将第一复合物和第一珠群组组合在一起形成混合物(通常是单一混合物,例如单一管中的混合物)。通常单一混合物中第一珠的数量大于第一复合物的数量,使得每个第一复合物都被珠捕获。每个第一珠连接至与互补寡核苷酸杂交的第一捕获寡核苷酸。在一些实施方案中,第一捕获寡核苷酸包含(1)启动子序列或引物结合序列,(2)第一条形码,其中固定在相同个体珠上的第一捕获寡核苷酸包含相同的第一条形码,并且大多数珠具有不同的第一条形码。在一些实施方案中,第一捕获寡核苷酸的3'连接至珠。在一些实施方案中,第一捕获寡核苷酸的5'连接至珠。该方法还包括通过以下方式产生条形码化第一亚片段:(i)如果第一捕获寡核苷酸的3'连接到珠,则通过3'分支接合将第一捕获寡核苷酸接合至多个第一亚片段,或(ii)如果第一捕获寡核苷酸的5'连接至珠和互补寡核苷酸以并入第一条形码序列,则通过3'分支接合将互补寡核苷酸接合至多个第一亚片段。

[0115] 在一个步骤中将插入和捕获组合的另一方式中,将包含上述捕获寡核苷酸的转位子固定在各个珠上。在转位酶的存在下,靶核酸片段与珠组合。这允许将捕获寡核苷酸插入

到靶核酸片段中。该方式被描述于PCT公开WO 2014/145820中，相关公开内容通过引用并入本文。该方式可以应用于第一轮stLFR和/或第二轮stLFR两者。

4.释放

[0116] 在一个方式中，使用结合至PCR引物结合位点的引物通过切口平移来延伸条形码化第一亚片段，从而产生在两端侧接捕获寡核苷酸的第一复合物。任选地，将延伸的片段从珠中释放。从珠中释放可以通过降解第一珠或通过切割捕获寡核苷酸和珠之间的化学键来进行。在一些情况下，释放是通过使用EndoV酶从捕获寡核苷酸去除肌苷残基或通过尿嘧啶脱糖基酶和EndoIV/EndoVIII或其他具有类似功能的酶去除尿嘧啶核苷酸来实现的。在一些情况下，捕获寡核苷酸通过一个或多个二硫键与珠交联。在这种情况下，可以通过将珠暴露于还原剂(例如，二硫苏糖醇(DTT)或三(2-羧乙基)膦(TCEP))来实现该释放。然后通过SDS从第一靶DNA片段中去除转位酶，这产生更小的DNA片段，即第一亚片段。每个约为10kb-20kb。

[0117] 如果转位用于生成第一亚片段，也可以通过使转位酶变性来分离第一复合物中的条形码化第一亚片段。在使用缺口酶生成第一亚片段的情况下，每个第一复合物中的第一亚片段由于互补单链片段之间存在重叠而保持在一起。见图28A(“I”)。可以通过使第一复合物变性来分离第一亚片段。

5.扩增

[0118] 条形码化第一亚片段能够以多种方式扩增。在某些情况下，第一次扩增是线性扩增。在一些实施方案中，其中捕获寡核苷酸携带启动子(例如，T7启动子)，线性扩增通过与逆转录偶联的体外转录(图3)进行以产生第一链(图4)。之后可以进行第二链合成以生成双链DNA分子(未示出)。在一些实施方案中，条形码化第一亚片段的扩增包括(1)进行体外转录以生成RNA转录物，(2)使用退火到启动子序列的引物逆转录转录物以生成条形码化第一亚片段的cDNA链。任选地，如下文进一步描述的，将cDNA链环化并通过滚环扩增进一步扩增。可以使用cDNA链作为模板产生双链、条形码化第一片段。

[0119] 在一些实施方案中，使用单引物PCR进行扩增。在一些情况下，PCR引物可以是识别位于第一亚片段侧翼的捕获寡核苷酸中的引物结合位点的引物。单引物扩增通常仅进行几个循环，例如8个循环，以实现例如100x扩增。见图5。在一些实施方案中，扩增条形码化第一亚片段是通过使用与第一捕获寡核苷酸中的引物结合序列结合的引物延伸条形码化第一片段；从珠中释放延伸的条形码化第一亚片段，并通过单引物扩增将释放的条形码化第一亚片段扩增约80-120倍，由此产生扩增的条形码化第一亚片段。

6.接合衔接子寡核苷酸

[0120] 在一些情况下，在单引物扩增之后，衔接子寡核苷酸(⑨)被接合至扩增的条形码化第一亚片段的两端(图6A或图6B)。在一些实施方案中，使用这些方法中的任一种进行约3-100x(例如5-50倍、2-20倍、10-60倍)的扩增。衔接子寡核苷酸可以是双链或部分双链的DNA分子。

7.第二插入

[0121] 可以使用上文关于第一插入描述的方法来执行第二插入。在一些实施方案中，第二插入中使用的插入寡核苷酸(“第二插入寡核苷酸”)与第一插入中使用的插入寡核苷酸相同。在一些实施方案中，第二插入中使用的插入序列不同于第一插入中使用的序列。在一

些实施方案中，第二插入中使用的酶与第一插入中使用的酶相同。在一些实施方案中，第二插入中使用的酶与第一插入中使用的酶不同。在一些实施方案中，第一插入和第二插入通过转位进行。在一些实施方案中，第一位置通过运送进行，而第二插入使用如上所述的缺口酶(例如，与接合酶一起使用的切口酶)进行。在一些实施方案中，第一插入和第二插入两者均由包含切口酶和接合酶的酶进行。

[0122] 通常，第二插入的插入频率高于第一插入的插入频率，使得两个相邻的第二插入寡核苷酸之间的片段比第一插入产生的那些(即两个相邻的第一插入寡核苷酸之间的片段)短。在一些实施方案中，第二亚片段的长度通常为100-1000bp。

[0123] 更高的插入频率可以通过例如使用更高浓度的转位子和/或更高量的转位酶或更高量的缺口酶、接合酶和衔接子来实现。通常，增加插入酶的浓度导致更高的插入频率，这导致插入位点之间的DNA片段更小。如图1所示，在0.1pmol/10ng基因组DNA到1pmol/10ng基因组DNA范围内的不同浓度下使用的转位子和Tn5产生了不同大小的基因组片段：Tn5和转位子的浓度越高，插入寡核苷酸之间的片段越小。

[0124] 第二插入将第二插入寡核苷酸引入至少一些扩增的条形码化第一亚片段的每个中，由此产生多个第二亚片段，统称为“第二复合物”。每个第二复合物包含第二插入寡核苷酸的多个副本。见图12。在优选的实施方案中，第二插入寡核苷酸具有与衔接子寡核苷酸的序列相同的序列(图6A和6B中的⑨)。类似于第一插入，第二插入产生第二复合物，至少一些第二复合物的每个包含多个第二亚片段。每个第二亚片段位于一个第二插入寡核苷酸与其侧翼的第二插入寡核苷酸之间。在使用转位子转位酶的情况下，这些第二亚片段通过转位酶相互连接，该转位酶保持与第二亚片段的末端结合。在一些实施方案中，第二插入发生在溶液中，即插入的序列不附着于任何固体支持物。第二复合物的长度通常在1kb-20kb(例如1kb-10kb、1kb-5kb或3kb-15kb)的范围内。

8.第二捕获

[0125] 进行第二轮捕获，其中第二复合物以类似于第一轮捕获的方式被第二珠捕获。每个第二珠具有固定在其上的第二捕获寡核苷酸的许多副本。每个第二捕获寡核苷酸包含每个第二珠特有的第二条形码，即每个第二珠中的所有第二条形码都相同。第二捕获寡核苷酸还包含引物结合位点和第二共同序列。第二共同序列与第二夹板寡核苷酸的一部分杂交。第二夹板寡核苷酸包括可以与第二插入寡核苷酸中的杂交序列杂交的另一部分。可以添加接合酶以将单个珠的第二捕获寡核苷酸接合至第二复合物中的第二插入寡核苷酸。在另一方式中，捕获寡核苷酸与连接到第二亚片段的第二插入寡核苷酸杂交。然后延伸第二插入寡核苷酸以产生第二条形码序列。如本领域技术人员将理解的，描述的用于将第一条形码引入第一亚片段的任何方式可用于将第二条形码引入第二亚片段。这些步骤产生条形码化第二亚片段，该条形码化第二亚片段各自包含至少一个第二条形码序列(例如，一个或两个第二条形码序列)。

9.在一个步骤中进行第二插入和第二捕获

[0126] 在类似于在一个步骤中进行第一插入和第一捕获的方式中，第二插入和第二捕获(如上面的步骤7和8)也可以以这样的方式组合，使得第二插入寡核苷酸已经固定在珠上，并且第二插入寡核苷酸的添加和第二复合物的生成发生在珠的表面上。见图27A和27B。在该实施方案中，该方法包括将交错的单链断裂引入靶核酸的片段，由此产生第二复合物。每

个第二复合物包含多个第二亚片段并且每个第二亚片段位于两个相邻的单链断裂之间。该方法还包括将第二复合物与第二珠的群组组合成单一混合物。第二珠的数量大于单一混合物中第二复合物的数量，使得每个第一复合物都被珠捕获。每个第二珠包含固定在其上的双链或部分双链捕获寡核苷酸衔接子的多个副本。每个衔接子包含作为连接至珠的第二捕获寡核苷酸的一条链，以及与第二捕获寡核苷酸互补的一条链。在一些实施方案中，第二捕获寡核苷酸链包含(1)启动子序列或引物结合序列，(2)第二条形码，其中固定在相同个体珠上的第二捕获寡核苷酸包含相同的第二条形码，并且大部分珠具有不同的第二条形码。该方法还包括通过以下方式产生条形码化第二亚片段：(i)通过3'分支接合将第二捕获寡核苷酸衔接子接合至多个第二亚片段，或(ii)通过3'分支接合将与第二捕获寡核苷酸互补的链接合至多个第二亚片段，并延伸互补链以并入第二条形码序列。

10.分支接合

[0127] 任选地，第二捕获寡核苷酸包含第一PCR引物退火位点。任选地，3'分支接合衔接子寡核苷酸在切口处接合至第二亚片段，其中3'分支接合衔接子包含第二引物退火位点。分支接合被详细描述于Wang等人,BioRxiv,June 29,2018,doi:https//doi.org/10.1101/357863;以及PCT公开号WO 2019/217452;其相关公开内容通过引用并入本文。

11.制备第二亚片段用于测序

[0128] 通过去除结合到第二片段末端的酶(例如，使用SDS)或通过使第二复合物变性来制备个体第二亚片段。然后使用两个引物扩增个体第二亚片段，一个退火至第一PCR引物退火位点，而另一个退火至第二PCR引物退火位点。扩增片段可按如下所述进行测序。

[0129] 两个条形码和UMI的组合允许通过组合具有相同第二条形码的序列读段来构建中间片段，即第一亚片段。然后通过组合具有相同第一条形码的中间片段的序列，使用来自中间片段的序列来构建复合物靶核酸片段。这允许构建长靶核酸的序列。

12.滚环扩增

[0130] 如上文步骤4中所述，在一些实施方案中，可使用滚环扩增(“RCR”)进行扩增以实现无PCR的长片段读段分析。见图22A和22B。在一些实施方案中，RCR是可取的，因为它具有扩增长核酸片段(包括全长基因组)而不是小片段的优势。RCR通常比PCR(其中在PCR的早期循环中生成的错误会被带到后面的扩增循环中)具有更高的保真度。此外，RCR只需要一种引物，而PCR通常需要两种引物来进行扩增。此外，原始亚片段的副本作为单个串联体连接在一起，从而允许作为具有相同亚片段的许多副本的单个分子捕获到珠。

[0131] 滚环扩增需要环状模板。在一些实施方案中，如以上步骤4中所述，将条形码化第一亚片段变性为单链核酸分子。然后添加夹板寡核苷酸，然后在接合酶(例如T4或Taq接合酶)的存在下环化单链核酸。

[0132] 用于RCR的DNA聚合酶可以是具有链置换活性的任何DNA聚合酶，例如Phi29、Bst DNA聚合酶、DNA聚合酶I的Klenow片段和Deep-VentR NDA聚合酶(NEB#MO258)。已知这些DNA聚合酶具有不同强度的链置换活性。选择用于本发明的一种或多种合适的DNA聚合酶在本领域普通技术人员的能力范围内。

[0133] 在一些实施方案中，通过线性扩增来扩增条形码化第一亚片段，并且以类似方式环化扩增产物以用作RCR的环状模板。在一些实施方案中，RCR之前的线性扩增是2-10x(例如2-3x、2-5x或3-10x)扩增。这种扩增是需要的，因为它提供了覆盖几乎所有条形码化片段

的冗余，其中至少一个加标签的串联体结合至第二条形码化珠，尽管在环化串联体生成和被条形码化珠捕获的期间有大量DNA损失。如果初始条形码化效率为70-80％，这种组合，即线性扩增和RCR，允许初始长片段的超过90％的碱基由至少一条链表示。对于许多应用，超过70％或超过80％将是足够的。覆盖率越大，细胞数量越少。因此，该方法可用于分析少于50个、少于30个、少于20个、少于15个、少于10个或少于5个人类或其他细胞的细胞数量。

13.无PCR长片段分析

[0134] 在一些实施方案中，用于测序分析的长核酸片段的制备是无PCR的，即仅通过RCR进行扩增。见图22A和22B。用于该分析的细胞数量可以变化很大，例如，5-50个细胞、3-20个细胞或10-100个细胞或更多。可以使用来自从大量(例如数千到数十亿个)细胞中分离的基因组DNA大型库的等量DNA。可以将细胞分配到单个管中，或者可以将单个细胞添加到微孔板的各个孔中以进行单细胞分析。细胞被裂解，并且DNA从细胞中释放出来。为此，可以如C.Chen等人，Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion(LIANTI).Science 356,189-194(2017)(其相关公开内容通过引用并入本文)中所述，进行裂解和蛋白酶处理。

[0135] 可以扩增来自细胞的基因组DNA，并且如果这些扩增产物是单链的(例如，通过滚环扩增产生)，则它们首先被转化为双链或部分双链的DNA分子。与上述靶核酸片段类似的这些DNA分子可以以与上述相似的方式进行第一插入、第一捕获等。如上所述，在一个方式中，将与转位酶偶联的转位子与扩增产物一起温育，使得将插入寡核苷酸插入到扩增产物中。由转位子引入的相邻插入寡核苷酸之间的间隔具有范围从1kb到50kb(例如从1kb到3kb，从1kb到5kb，从2kb到5kb，从3kb到6kb，从3kb到10kb，从5kb到20kb，或从10kb到50kb)的长度。如果扩增产物是单链DNA，则它们在转位前转化为双链DNA。见图22C和22D。替代地，可以在长DNA与条形码化珠相互作用之前或期间通过使用诸如切口/缺口之类的技术来制备用于共条形码化的DNA。结果是相邻插入寡核苷酸之间的间隔具有范围从300,000bp到3,000,000bp的长度。

[0136] 对于单细胞分析，具有独特条形码的转位子被分配到每个孔(或其他隔室)中，每个孔包含不超过一个细胞。以这种方式，来自各个细胞的每个DNA都用该细胞的特定条形码化转位子加标签。如上所述，在一些实施方案中，可以添加转位子插入的DNA，与平均直径范围从1到50um的珠一起温育。每个珠携带$1x10^3$-$1x10^9$个捕获寡核苷酸，它们如先前在例如US 2014/0323316中所述共享相同条形码。在一些实施方案中，可以使用1-$100x10^6$个珠。在通过与转位子和珠上的捕获寡核苷酸的杂交序列杂交来捕获靶核酸片段后，进行接合以便将条形码插入到靶核酸片段中。此时转位酶被破坏，并且由转位子插入产生的亚片段彼此分离，但仍连接至珠。通常，在此步骤中可以预期约50-80％的产率，这意味着将保留2.5-35个细胞等效DNA，并且原始长片段的75-91％用条形码化DNA表示。可以进行延伸步骤来复制在捕获片段3'侧上的转位子序列的一部分。如果进行单链环化，该序列将用作夹板寡核苷酸的杂交序列，否则其可用作用于延伸的引物。此外，转位子序列的一部分还可用于将衔接子接合至通过转位子插入产生的3'凹陷的末端以创建用于引物延伸和夹板寡核苷酸杂交的序列。此时，可以通过连续轮次的变性、引物的杂交和延伸对珠进行线性扩增，或者如果使用衔接子接合至每个片段的3'末端以添加额外的引物位点，则可以进行PCR。

[0137] 此外，可以通过首先通过例如使用EndoV酶从捕获寡核苷酸分子去除肌苷残基从

珠中释放片段来制备初始条形码化片段(也称为条形码化第一亚片段)以用于扩增(例如滚环扩增)。使用条形码化第一亚片段作为模板来产生双链或部分双链。接下来将dsDNA片段变性,添加夹板寡核苷酸,并使用T4或Taq接合酶进行ssDNA环化。见图22A。通过线性扩增生成的DNA也可用于环化。

[0138] 如果使用诸如Taq接合酶之类的热稳定接合酶,可以进行变性和接合的循环以增加产率(例如,增加到至少约70%)。我们预计这一步骤的产率约为70%,剩下1.75-24.5个DNA细胞。然而,DNA的两条链都可以被捕获,这表明每个片段的高达75%(0.7x0.7=49%的片段环化;0.51x0.51=～25%的片段不会由任何一条链表示)被至少一条DNA链覆盖。

[0139] 替代地,可以在环化之前进行2-3倍或2-5倍或3-10倍的线性扩增,这产生10-100x多联体。这提供了覆盖几乎所有条形码化靶核酸片段的冗余,其中至少一个加标签的多联体结合至第二条形码化珠,尽管在环化、多联体生成和将其与条形码化珠结合的步骤期间存在大量DNA损失。

[0140] 如果初始条形码化效率为70-80%,则该方式允许初始长片段的>90%的碱基由至少一条链表示。对于许多应用,高于70%或>80%的表示就足够了。覆盖率越大,序列确定所需的细胞数量越少。可以使用本文所述的方法从少于50个细胞,或<30、<20、<15、<10、<5个细胞(例如,人细胞)开始进行测序。

[0141] 使用上述任何这些扩增方法进行约3-100倍(优选10倍)的扩增。对于滚环扩增和线性扩增,通过与衔接子序列互补的引物退火并用DNA聚合酶进行延伸,可以将ssDNA片段转化为dsDNA。在某个实施方案中,DNA聚合酶具有切口平移能力。接合酶也可用于密封切口。随着所有分子转化为dsDNA,转位子可以以0.1-1.5kb(即相邻插入寡核苷酸之间的间隔长度的范围为0.1kb到1.5kb),优选0.3kb-1.5kb的频率再次插入。如上所述,这些转位子插入片段被捕获到条形码化珠上。

14.按大小进行环分级

[0142] 在一些实施方案中,有利的是在滚环扩增(RCA或RCR)之前按大小分离环(例如,通过环化第一亚片段形成的环,见上文)以实现对由第一条形码化产生的可变片段的相似读段覆盖率。通过充分的扩增,在准备测序之前对条形码化片段的大小选择可用于获得片段(从例如300到500碱基或400到600碱基)以实现更高效的测序。此外,制备阵列(通过DNB或簇)可以更有效地提供足够的DNA。对于15个细胞并以3kb片段覆盖每个细胞$6x10^9$个碱基,有$3x10^6$个片段各自都转化为～100kb串联体。因此,需要$30x10^6$-$100x10^6$或更多条形码化珠。

[0143] 在一些实施方案中,大小选择是通过对环分级来实现的。在一些实施方案中,通过杂交在固体支持物(例如珠)上捕获环,随后用链置换聚合酶控制引物延伸以从珠中置换环来进行环分级。通过杂交的捕获可以通过例如将环与其上固定有寡核苷酸的固体支持物混合来进行,其中寡核苷酸可以与环中的序列杂交。优选地,与固定化寡核苷酸杂交的环中的序列与延伸引物结合的环中的序列不重叠。延伸长度可以通过选择具有合适聚合速率或其他特性的一种(或多种)聚合酶,以及通过多种反应参数(包括但不限于反应温度、DNA聚合酶浓度、引物浓度和dNTP浓度)来控制。最佳条件可以凭经验确定。用于环分级的DNA聚合酶可以是具有如上所述链置换活性的任何DNA聚合酶。在一些实施方案中,DNA聚合酶是Phi29。在一些实施方案中,反应温度介于2℃和30℃之间,例如介于2℃和20℃之间,或介于

10℃和25℃之间。在一些实施方案中，可以使用仅允许预定义的延伸长度(例如300个碱基)的有限量的dNTP。在这种情况下，由于使用了几乎所有提供的dNTP，引物延伸反应实际上会在达到～300b后停止。

[0144] 置换的环被释放到反应混合物的上清液中。在引物延伸的相同条件(包括例如聚合酶速度、聚合酶浓度、dNTP浓度)下，需要更长的延伸反应来从固体支持物上置换较大的环，而不是较小的环。通过在聚合开始后以预定时间间隔收集上清液，可以收集具有所需大小的环。在一些实施方案中，在引物延伸开始后的2、3、4或5个不同时间点收集含有从固体支持物释放的环的上清液。

[0145] 然后通过滚环扩增来扩增具有所需大小的环，以产生具有靶DNA的相似副本数的多联体。通常，环越大，使用滚环扩增的时间越长。

[0146] 为了确保有效的环分级，需要确保环大致在同一时间开始滚动。确保环大致同时滚动的非限制性示例性方法包括使用高浓度聚合酶(例如，聚合酶相对于环至少过量10倍，例如聚合酶比环多10至100万倍的范围内)进行核苷酸聚合；将延伸引物与环杂交以形成杂交复合物，然后在固体支持物上捕获杂交复合物，将DNA聚合酶与环一起温育(例如，在EDTA的存在下)，然后在固体支持物上捕获环。

[0147] 本文公开的环分级也可用于许多其他应用，例如针对常规测序文库的大小选择(例如得到约400-600个碱基或500-1000个碱基或200-400个碱基或300-600个碱基的大小)，或用于测序(以获得相似的测序信号/强度)或其他目的的DNB的副本数的均衡化。

15.RCR产物的序列条形码

[0148] 首先在将ssDNA多联体转化为dsDNA时，除了合适的聚合酶之外，还可以使用接合酶以直接使用磷酸化引物或通过切口-平移-接合来密封切口。其次，每个衔接子(带有第一条形码)+DNA片段单位需要超过2X的测序覆盖率(优选3X或更多)，以确保在>90％或>95％的情况下对条形码进行测序。3X覆盖率还允许每个～3kb片段进行序列组装(“合成～3kb”；在某些情况下主要是一个或2-5个重叠群)。这将有助于从头组装。此外，能够从原始长DNA分子的～3kb片段中分离读段在从头组装中(尤其是在重复序列的区域中)是非常有效的。即使每～3kb片段的读段覆盖率较低，3kb区域中读段的定位也是非常有用的。这类似于先前描述的带有连接的条形码转位子的中间条形码化策略。

[0149] 使用每单位片段的3X覆盖率和15个细胞x 2条染色体(假设每条链中平均50％的DNA在第二个步骤中被条形码化，这相当于只有一条链)，总覆盖率为每人基因组～90X或300Gb。如果制作多联体的效率>90％，珠上的多联体的捕获效率为90％，并且测序覆盖率>90％，则环中的0.50个DNA x测序的环中的0.73个碱基＝原始长片段的每条链的37％或至少一条链中～60％被覆盖。

16.长DNA靶的群体测序

[0150] 一种或多种靶DNA片段(例如5-50kb、或10-50kb或10-300kb或30-300kb)可以从数百、数千或数百万个细胞或生物体(如细菌或真菌或其他)的基因组DNA中分离或富集。富集可以通过靶特异性远程PCR或通过杂交捕获或通过其他方法进行。制备的dsDNA将通过此处描述的双stLFR过程以进行成本有效的单长DNA分子测序。>80％或>90％或>95％的所有碱基将在数千到数百万个10kb+DNA分子中的每个中进行测序。

[0151] 应用可以包括全基因或基因簇的深度测序，以检测病原体种群中的癌症或其他体

细胞突变或变体(例如药物/抗生素抗性)或微生物种群中靶基因的多样性,例如肠道微生物组。

17.新型基因组文库

[0152] 使用本文公开的方法,可以得到基因组文库,其包含大量长基因组DNA片段(例如10-50kb、10-300kb、20-300kb、20-100kb)中的每个的>50%、70%、80%、90%、95%的碱基。这些长基因组DNA片段以属于长片段的较短亚片段(例如～1-10kb、～1-5kb、～3-15kb)的多个模板子集的形式在条形码模板中表示。这些新文库的数学表示如下所示。每个长片段都有一个独特条形码(例如,长片段1具有条形码1.0),其不同于其他长片段的条形码(例如,长片段2具有条形码2.0)。每个子集(例如"亚片段1.1子集")属于一个长片段(例如"长片段1"),该子集具有不同于属于相同长片段的其他子集(例如"亚片段1.2子集")的特定条形码(例如"条形码1.0"),并且来自每个此类子集的至少一个模板也具有长DNA的特定条形码(例如,亚片段1.1子集的至少一个模板具有条形码1.1和条形码1.0两者)。

18.新文库的数学表示:

[0153] 长片段1用条形码1.0共条形码化

亚片段1.1子集:多个模板带有条形码1.1+至少一个模板带有条形码1.1和条形码1.0

亚片段1.2子集:多个模板带有条形码1.2+至少一个模板带有条形码1.2和条形码1.0

亚片段1.n子集:多个模板带有条形码1.n+至少一个模板带有条形码为1.n和条形码1.0;n为约2至1000,或2-100,或3至10,或3-30、3-300或10-30、10-100。

[0154] 长片段2用条形码2.0共条形码化

亚片段2.1子集:多个模板带有条形码2.1+至少一个模板带有条形码2.1和条形码2.0

亚片段2.2子集:多个模板带有条形码2.2+至少一个模板带有条形码2.2和条形码2.0

亚片段2.n子集:多个模板带有条形码2.n+至少一个模板带有条形码为2.n和条形码2.0;n为约2至1000,或2-100,或3至10,或3-30、3-300或10-30、10-100。

……

长片段F用条形码f.0共条形码化亚片段f.1子集:多个模板带有条形码f.1+至少一个模板带有条形码f.1和条形码f.0亚片段f.2子集:多个模板带有条形码f.2+至少一个模板带有条形码f.2和条形码f.0。

……

亚片段f.n子集:多个模板带有条形码f.n+至少一个模板带有条形码f.n和条形码f.0;n为约2至1000,或2-100,或3至10,或3-30、3-300,或10-30、10-100;F是1k+,10k+,100k+,1M+,10m+。

[0155] 此类文库可以用于作为混合物的多个细胞或单个细胞。在一些实施方案中,此类文库由多个细胞制备,其中除了长DNA片段的条形码之外,每个细胞还具有指定的条形码。

方案II.单细胞转录组方法

[0156] 下面描述的该方案可用于分析细胞中的转录物。从起始细胞群中获得单个细胞。

在某些情况下，细胞在微孔板中被稀释。通过使用孔数大大超过细胞数的板，非常高比例的孔包含一个细胞或没有细胞。在这种情况下，以下方法可用于分析来自单个细胞的转录物。

1.mRNA捕获和逆转录

[0157] 裂解细胞并提取mRNA。mRNA在单一混合物中与第一珠的群组组合，每个珠具有固定在其上的多个第一捕获寡核苷酸。每个第一捕获寡核苷酸包含独特第一条形码序列、UMI、共同引物位点和用于捕获mRNA分子的polyA尾的寡聚-dT。见图8。

[0158] 然后将捕获的mRNA分子逆转录以产生cDNA，至少一些cDNA中的每个连接至第一珠上的第一捕获寡核苷酸以形成cDNA/mRNA杂交分子。见图9。在一些实施方案中，然后将如上产生的cDNA/mRNA杂交分子接合至包含GGG三核苷酸的衔接子，其与通过逆转录酶添加到cDNA 3'末端的CCC序列互补。见图9。衔接子还可以包含PCR引物结合序列，其可以与捕获寡核苷酸中的共同引物结合序列相同。在高浓度镁和/或锰离子等条件下，当逆转录酶到达RNA模板的5'末端时，可以将CCC三核苷酸添加到cDNA的3'末端。用于诱导添加CCC三寡核苷酸的方法是本领域众所周知的，例如Schmidt等人.Nucleic Acids.Res.27(21):e31(1999);Pinto et al.,Anal.Biochem.397(2):227-232(2010)。

2.释放和扩增

[0159] 任选地，从第一珠中释放出cDNA。在一些实施方案中，第一片段用识别第一捕获寡核苷酸上的共同引物结合序列以及衔接子上的引物结合序列(两者均整合到合成的cDNA中)的共同引物进行扩增。见图10。使用单引物扩增的优点是单引物扩增更有可能导致较短扩增产物的发夹形成。这些发夹结构不会被进一步扩增，不参与下游处理步骤。因此在此步骤中使用单引物PCR可以增加更长的扩增产物的数量。

[0160] 任选地，通过滚环扩增将分子环化和扩增。

3.将衔接子寡核苷酸接合至扩增的双链DNA

[0161] 扩增的双链DNA在两端与衔接子寡核苷酸接合。每个衔接子都是部分双链DNA。见图11。

4.按照方案I步骤6-8进行第二转位和单管LFR

[0162] 末端带有衔接子寡核苷酸的双链DNA用第二插入寡核苷酸转位，然后如方案I步骤6-8中所述与第二结合物一起温育。

[0163] 两个条形码(一个来自步骤1，另一个来自步骤4)和UMI的组合允许完整构建所有完整转录物并为其分配个体细胞。例如，当从单个细胞中提取的mRNA与第一珠一起温育时，每个第一条形码对应于个体细胞，并且每个UMI对应于该个体细胞中的个体转录物。

方案III.长扩增子组装方法

[0164] 方案III的方法可用于分析可能感兴趣的任何靶区域(例如16s或18s区域rRNA基因)的全长。一些实施方案在图13-19中示出。该测定可以多重进行以分析单个隔室中的多个靶区域。

方案IIIa.

1.靶特异性引物扩增

[0165] 一个或多个靶区域中的每个用与可以被酶降解的特殊碱基复合的正向引物，以及反向引物扩增有限数量的循环，例如两个循环。特殊碱基是一种不会出现在天然DNA中的碱基。酶可以通过识别特殊碱基来消化引物。在一个示例中，特殊碱基是尿嘧啶。正向引物附

加地包含5'处的第一共同序列和5'末端的UMI。正向和反向引物中的每个都包含靶特异性序列(图13中的"靶特异性序列1"和"靶特异性序列2"),这允许它们特异性地杂交并扩增靶DNA区域。

2.去除过量含UMI的正向引物

[0166] 然后用识别正向引物中的特殊碱基并将其降解的酶处理来自步骤1的反应混合物。见图13。在特殊碱基为尿嘧啶的情况下,可用USER消化过量的正向引物。由于正向引物包含UMI和靶特异性序列,因此去除过量的引物可以防止UMI被打乱并允许多重进行。

3.进一步扩增

[0167] 使用一种与第一共同序列杂交的引物和反向引物来扩增来自步骤2的包含靶区域的DNA以产生第一复合物、包含第一共同序列的双链DNA分子和靶UMI以产生进一步扩增的产物。见图14。

4.将衔接子寡核苷酸接合至进一步扩增的产物

[0168] 衔接子寡核苷酸被接合至进一步扩增的产物的末端。见图15。

5.如方案I步骤6-8中进行转位和单管LFR

[0169] 以类似于方案1、步骤6的方式通过转位将具有与衔接子寡核苷酸相同序列的插入寡核苷酸引入第一复合物中,并且如方案I、步骤7-8中所述进行StLFR。见图19。

方案IIIb.

1.靶特异性引物扩增

[0170] 使用靶特异性引物对将一个或多个靶DNA区域中的每个扩增有限数量的循环,例如两个循环。见图16。

2.添加UMI和常用序列

[0171] 将两种共同序列寡核苷酸,即第一和第二共同序列寡核苷酸接合至包含靶区域的扩增双链DNA(一个末端一种)。见图17。两种共同序列寡核苷酸中的每种都包含共同序列。在一些实施方案中,两种寡核苷酸共享相同的共同序列。共同序列寡核苷酸之一或两者可在3'处包含UMI。在一些实施方案中,两种寡核苷酸都包含UMI,并且两种共同序列寡核苷酸之间的UMI不同。然后使用退火至第一和第二共同序列寡核苷酸的引物对(正向和反向引物)扩增接合产物,该第一和第二共同序列寡核苷酸被接合至靶DNA区域以产生第一复合物。在一些实施方案中,正向和反向引物两者均具有与共同序列互补的序列。见图18。在一些实施方案中,共同序列寡核苷酸是部分双链寡核苷酸,其中UMI位于双链区域中并且共同序列位于共同序列寡核苷酸的单链区域中。在一些实施方案中,共同序列寡核苷酸是部分双链寡核苷酸,其中UMI和共同序列两者都位于单链区域中。

3.接合衔接子寡核苷酸

[0172] 将衔接子寡核苷酸接合至其侧翼是第一和第二寡核苷酸的扩增的靶区域的两端。见图18。

4.如方案I步骤6-8转位后进行单管LFR

[0173] 以类似于方案1、步骤6的方式通过转位将具有与衔接子寡核苷酸相同序列的插入寡核苷酸引入第一复合物中,并且如方案I、步骤7-8中所述进行StLFR。见图19。

[0174] 两个条形码和UMI的组合允许完整构建靶区域的全长序列。

方案IV.

[0175] 在一些实施方案中，将每个长的靶核酸分子(例如，约25kb至100kb)接合至衔接子和独特分子标识符(UMI)。UMI的长度足以将个体靶核酸彼此区分开来。靶核酸分子数越多，需要的UMI越长；例如，对于1000万个分子，UMI通常为11-20，例如11-18，例如12-15个碱基或更长。在某些情况下，UMI在并入衔接子之前制备成多个副本，以便增加具有包含UMI的至少一个序列读段的几率。因此，在一些实施方案中，将每个长的靶核酸分子接合至一个或多个UMI副本。

[0176] 在一些实施方案中，将插入了衔接子和UMI的靶核酸分子变性以形成单链核酸分子，然后环化。在一些实施方案中，将插入了衔接子和UMI的双链靶核酸分子环化，然后变性以形成环化的单链核酸，条件是在分子的一条链中存在切口。对环化单链核酸分子进行RCR以产生如上所述的DNA纳米球(“DNB”)。RCR反应时间可以变化，但通常反应时间的长度使得可以产生靶核酸的大于5x，例如大于10x，大于20x的副本。例如，对于长度为2kb-100kb的靶核酸分子，反应时间通常为约10-400分钟，这产生长度为约50kb至1Mb的DNB。因此，在一些实施方案中，RCR反应时间为约10-400分钟。然后通过退火与衔接子序列互补的引物将DNB转化为双链DNA，并使用DNA聚合酶进行延伸。在一些实施方案中，每个DNB被分段成多个片段，例如3-10个片段。DNB的分段可以(例如，由于DNB的长度)自发地发生，或者可以通过施加力来实现。使用转位酶将转位子插入双链DNA中，然后将转位的DNA捕获到珠上，如方案I，步骤1和2中所述。然后可以去除与双链DNA关联的转位酶，从而产生分离的DNA片段。

[0177] 然后对DNA片段进行测序以产生包含UMI序列的序列读段。UMI序列的存在允许具有相同UMI的多个DNB的序列关联，即使它们被捕获在不同的珠上。这可以有利地产生原始长靶核酸的接近于1X的序列覆盖率。

IV.方法的组分

1.样本

[0178] 可以从任何合适的来源获得含有靶核酸的样本。例如，样本可以从任何感兴趣的生物体获得或提供。此类生物体包括例如植物；动物(例如，哺乳动物，包括人类和非人类灵长类动物)；或病原体，例如细菌和病毒。在一些情况下，样本可以是或可以从感兴趣的此类生物体群体的细胞、组织或多核苷酸获得。作为另一个示例，样本可以是微生物组或微生物群。任选地，样本是环境样本，例如水、空气或土壤的样本。

来自感兴趣的生物体或感兴趣的此类生物体群体的样本可以包括但不限于体液(包括但不限于血液、尿液、血清、淋巴、唾液、肛门和阴道分泌物、汗液和精液)、细胞、组织的样本；活组织检查、研究样本(例如，核酸扩增反应(如PCR扩增反应)的产物)；纯化的样本，例如纯化的基因组DNA；RNA制剂；以及原始样本(细菌、病毒、基因组DNA等)。从生物体获得靶多核苷酸(例如基因组DNA)的方法是本领域公知的。

2.靶核酸

[0179] 如本文所用，术语“靶核酸”(或多核苷酸)或“感兴趣的核酸”是指适于通过本文所述的方法处理和测序的任何核酸(或多核苷酸)。核酸可以是单链或双链的并且可以包括DNA、RNA或其他已知的核酸。靶核酸可以是任何生物体的核酸，该生物体包括但不限于病毒、细菌、酵母、植物、鱼、爬行动物、两栖动物、鸟类和哺乳动物(包括但不限于小鼠、大鼠、狗、猫、山羊、绵羊、牛、马、猪、兔子、猴子和其他非人类灵长类动物和人类)。靶核酸可以从一个个体或从多个个体(即群体)获得。从中获得核酸的样本可以包含来自细胞或甚至生物

体的混合物的核酸，例如：包括人细胞和细菌细胞的人唾液样本；包括小鼠细胞和来自移植的人类肿瘤的细胞的小鼠异种移植物；等。靶核酸可以是未扩增的，或者它们可以通过本领域已知的任何合适的核酸扩增方法扩增。可以根据本领域已知的方法纯化靶核酸以去除细胞和亚细胞污染物(脂质、蛋白质、碳水化合物、除要测序的那些以外的核酸等)，或者它们可以是未纯化的，即，包括至少一些细胞和亚细胞污染物，包括但不限于被破坏以释放其核酸用于处理和测序的完整细胞。可以使用本领域已知的方法从任何合适的样本中获得靶核酸。此类样本包括但不限于生物样本，例如组织、分离的细胞或细胞培养物、体液(包括但不限于血液、尿液、血清、淋巴、唾液、肛门和阴道分泌物、汗液和精液)；以及环境样本，如空气、农业、水和土壤样本等。

[0180] 靶核酸可以是基因组DNA(例如，来自单个个体)、cDNA，以及/或可以是复杂的核酸，包括来自多个个体或基因组的核酸。复杂核酸的示例包括微生物组、孕妇血流中的循环胎儿细胞(见例如Kavanagh等人，J.Chromatol.B 878:1905-1911,2010)、来自癌症患者的血流中的循环肿瘤细胞(CTC)。在一个实施方案中，这种复杂核酸具有包含至少一个千兆碱基(Gb)的完整序列(二倍体人类基因组包含大约6Gb的序列)。

[0181] 在一些情况下，靶核酸或第一复合物是基因组片段。在一些实施方案中，基因组片段长于10kb，例如10-100kb、10-500kb、20-300kb、50-200kb、100-400kb，或长于500kb。在一些情况下，靶核酸或第一复合物的长度为5,000至100,000kb。在单一混合物中使用的DNA(例如，人类基因组DNA)的量可以<10ng、<3ng、<1ng、<0.3ng，或<0.1ng的DNA。在一些实施方案中，单一混合物中使用的DNA的量可以小于3,000x，例如小于900x、小于300x、小于100x或小于30x的单倍体DNA量。在一些实施方案中，单一混合物中使用的DNA的量可以是至少1x单倍体DNA，例如至少2x或至少10x单倍体DNA量。

[0182] 可以使用常规技术分离靶核酸，例如，如上文引用的Sambrook和Russell，Molecular Cloning:A Laboratory Manual中所公开的那样。在某些情况下，特别是如果在特定步骤中采用少量的核酸，每当仅少量的样本核酸是可用的并且存在通过非特异性结合(例如与容器壁等结合)而损耗的危险，提供载体DNA(例如无关的环状合成双链DNA)以与样本核酸混合并一起使用则是有利的。

[0183] 根据本发明的一些实施方案，通过任何已知方法，在纯化或不纯化的情况下从单个细胞或少量细胞获得基因组DNA或其他复杂核酸。

[0184] 本发明的方法需要长片段。可以通过任何已知方法从细胞中分离基因组DNA的长片段。例如，在Peters等人，Nature 487:190-195(2012)中描述了从人类细胞中分离长基因组DNA片段的方案。在一个实施方案中，将细胞裂解并且将完整的细胞核通过温和的离心步骤沉淀。然后通过蛋白酶K和RNase消化数小时释放基因组DNA。可以处理材料以降低剩余细胞废物的浓度，该处理例如通过透析一段时间(即从2到16个小时)和/或稀释进行。由于此类方法不需要采用许多破坏性过程(例如乙醇沉淀、离心和涡旋)，因此基因组核酸在很大程度上保持完整，从而产生大多数长度超过150千碱基的片段。在一些实施方案中，片段的长度为约5至约750千碱基。在进一步的实施方案中，片段的长度为约150至约600、约200至约500、约250至约400和约300至约350千碱基。可用于单倍体分析的最小片段是包含至少两个杂合子(hets)(大约2-5kb)的片段；虽然片段长度可能受到起始核酸制剂操作产生的剪切的限制，但是没有最大理论大小。

[0185] 在其他实施方案中，长DNA片段以使DNA对容器的剪切或吸收最小化的方式进行分离和操作，包括例如在琼脂糖凝胶塞或油中的琼脂糖中分离细胞，或使用特殊包被管和板。

[0186] 5'核酸外切酶的受控使用(在扩增之前或期间)可以促进来自单个细胞的原始DNA的多次复制，因而通过副本的复制使早期错误的传播最小化。

[0187] 可以通过将衔接子与单链引发突出端接合并使用衔接子特异性引物和phi29聚合酶从每个长片段制作两个副本来复制来自单个细胞的片段化DNA。这可以从单个细胞中生成相当于四个细胞的DNA。

[0188] 根据本发明的一个实施方案，从比测序所需的更多的长片段开始以实现足够的序列覆盖率并且仅用有限数量的包含标签的序列，或标签组装—其包括一个标签序列的许多(或许数百个)副本—对长片段的仅一部分加标签以增加对长片段进行独特加标签的可能性。缺少提供引物结合或捕获寡核苷酸结合的引入序列的未加标签的亚片段可能在下游处理中被消除。此类标签组装包括，例如，通过滚环复制(DNA纳米球)产生的含标签序列的端到端多联体、其上附着有含标签序列的许多副本的珠，或其他实施方案。

[0189] 根据另一个实施方案，为了在样本具有少量细胞(例如，来自微活检或循环肿瘤或胎儿细胞的例如1、2、3、4、5、10、10、15、20、30、4、050或100个细胞)的情况下获得统一的基因组覆盖率，所有从细胞获得的长片段都被加标签。

3.转位

[0190] 任何合适的转位子/转位酶或转位子/整合酶系统均可用于本公开的方法中。示例包括体外Mu转位(Haapa等人，Nucl.Acids Res.，27:2777-2784，1999；Savilahti等人，EMBO J.14:4893-4903，1995)；Tyl(Devine和Boeke，Nucl.Acids Res.，22:3765-3772，1994；国际专利申请WO 95/23875)；Tn7(Craig，Curr.Topics Microbiol.Immunol.204:27-48，1996)；Tn 10和IS 10(Kleckner等人，Curr.Top.Microbiol.Immunol.204:49-82，1996)；Mariner(Lampe等人，EMBO J.15:5470-5479，1996)；Tcl(Vos等人，Genes Dev.，10:755-761，1996)；Tn5(Park等人，Taehan Misaengmul Hakhoechi 27:381-389，1992)；P元素(Kaufman和Rio，Cell 69:27-39，1992)；Tn3(Ichikawa和Ohtsubo，J.Biol.Chem.265:18829-18832，1990)；细菌插入寡核苷酸(Ohtsubo和Sekine，Curr.Top.Microbiol.Immunol.，204:1-26，1996)；逆转录病毒(Varmus和Brown，“Retroviruses，”in Mobile DNA，Berg和Howe，编辑，American Society for Microbiology，Washington，DC，pp.53-108，1989)；以及酵母逆转录转位子(Boeke，“Transposable elements in Saccharomyces cerevisiae，”in Mobile DNA，Berg和Howe，编辑，American Society for Microbiology，Washington，DC，pp.53-108，1989)。其他已知的转位子包括但不限于Tn5，AC7，Tn5SEQ1，Tn916，Tn951，Tn1721，Tn2410，Tn1681，Tn1，Tn2，Tn4，Tn6，Tn9，Tn30，Tn101，Tn903，Tn501，Tn1000(γ6)，Tn1681，Tn2901，AC转位子，Mp转位子，Spm转位子，En转位子，Dotted转位子，Ds转位子，dSpm转位子和I转位子。可以使用转位子末端和/或转位酶的修饰形式，例如，如Nextera™技术(威斯康星州麦迪逊的Epicentre Biotechnologies)中的修饰的Tn5转位酶。见US20180016571 A1，相关公开内容通过引用并入本文。

[0191] 许多转位酶识别不同的插入寡核苷酸，因此应理解，基于转位酶的载体将包含由也在基于转位酶的载体中发现的特定转位酶识别的插入寡核苷酸。可以修改和使用(包括)来自基于真核转位子的载体的转位酶和插入寡核苷酸。然而，基于非真核转位子的元素降

低了受体生物体(例如人类受试者)中的真核转位酶将识别包围转基因的原核插入寡核苷酸的可能性。

[0192] 在使用时,转位子与DNA的长片段(双链或部分双链)组合,并且转位酶的添加导致插入寡核苷酸转位到长片段中。转位子携带其序列与靶序列互补的两个转位子末端。WO 2014/145820中描述了以不同方式发生转位的方法。转位子包括插入寡核苷酸,其包含用于杂交的单链区域("杂交序列")以及被转位酶识别并能够进行转位反应的双链镶嵌序列。转位酶这种酶具有在转位事件后保持与基因组DNA结合的特性,从而有效地使转位子整合的长基因组DNA分子保持完整,该分子可以通过来自同一珠的条形码加标签。条形码化后,转位酶可以通过例如用SDS处理来去除,这产生分离的核酸片段。

[0193] 转位酶催化切除的转位子随机插入DNA靶。在剪切和粘贴转位期间,转位酶在靶DNA中产生随机交错的单链断裂,并将转移的转位子链的3'末端共价连接到靶DNA的5'末端。转位酶/转位子复合物在转位子插入靶核酸的点处插入任意DNA序列。在某些情况下,使用随机插入靶核酸序列的转位子。已经描述了几种转位子并在体外转位系统中使用。例如,在Nextera™技术(Nature Methods 6,2009年11月;威斯康星州麦迪逊的Epicentre Biotechnologies)中,插入不需要整个复合物;自由转位子末端足以整合。当使用游离转位子末端时,靶DNA被片段化,并且转位子末端寡核苷酸的转移链共价连接至靶片段的5'末端。转位子末端可以通过添加所需序列(例如PCR引物结合位点、条形码/标签等)来修饰。可以通过改变转位酶和转位子末端的数量来控制片段的大小分布。带有附加序列的转位子末端产生可用于高通量测序的DNA文库。

[0194] 转位事件的频率与转位子和转位酶的浓度呈正相关,即,为了生成大的基因组片段,如在第一轮转位事件中,应使用较低浓度的转位子和转位酶;并且为了生成较小的基因组片段,如在第二轮转位中,使用较高浓度的转位子和转位酶。

[0195] 转位子中的插入寡核苷酸可包含标签或序列,其可用于与其他核酸杂交以供进一步分析。在一些实施方案中,插入寡核苷酸包含可与夹板寡核苷酸杂交的序列,该夹板寡核苷酸还与附着于珠的捕获寡核苷酸杂交,如下所述。

4.使用切口酶创建切口

[0196] 该方法中的各种实施方案需要在核酸片段被珠条形码化之前在核酸片段中引入切口。作为通过转位位引入切口的替代方式之一,切口酶可用于在核酸片段中产生切口。在一些实施方案中,切口酶在其识别序列处切割DNA。在一些实施方案中,切口酶在随机位置处切割DNA。切口酶的非限制性示例包括Nb.BsrDI,Nb.BsmI,Nt.BbvCI,Nb.Bbv,Nb.BtsI或Nt.BstNBI等。切口酶产生的切口被酶(如klenow片段)扩大并且所产生的片段通过接合酶接合以捕获固定在珠上的寡核苷酸衔接子。捕获寡核苷酸衔接子可以是双链或部分双链的。随着时间的推移,切割继续,并且打开更多的缺口以便更多的衔接子接合到缺口中。在一些实施方案中,切口酶以低浓度使用,Klenow片段以中等浓度使用,并且接合酶以高浓度使用以允许将切口低开口到缺口和快速接合,这将DNA片段锁定到珠。

[0197] 珠上的每个捕获寡核苷酸衔接子在一端通过一条或两条链附着于珠上,并且在另一端具有平端。衔接子的一条链包含捕获寡核苷酸,而另一条链与捕获寡核苷酸互补。捕获寡核苷酸可以包含本申请中描述的各种组分,例如引物结合序列、启动子序列、条形码和/或UMI。在一些实施方案中,捕获寡核苷酸的5'末端附着于珠上,并且其互补链的5'可以通

过3'分支接合而接合至DNA片段的3'末端(图27A)。在一些实施方案中,捕获寡核苷酸的3'末端附着于珠上,并且捕获寡核苷酸链的5'可以通过3'分支接合而接合至DNA片段的3'(图27B)。

4.条形码

[0198] 根据一个实施方案,使用具有两个、三个或更多个片段(其中一个片段例如是条形码序列)的包含条形码的序列。例如,引入的序列可以包括已知序列的一个或多个区域和用作条形码或标签的简并序列的一个或多个区域。已知序列(B)可以包括例如PCR引物结合位点、转位子末端、限制性内切核酸酶识别序列(例如稀有切割酶的位点,例如Not I、Sac II、Mlu I、BssH II等),或其他序列。用作标签的简并序列(N)足够长以提供等于或优选大于待分析的靶核酸片段数量的不同序列标签群。

[0199] 根据一个实施方案,包含条形码的序列包含任何选定长度的已知序列的一个区域。根据另一个实施方案,包含条形码的序列包含选定长度的已知序列的两个区域,它们位于选定长度的简并序列的区域的侧翼,即$B_nN_nB_n$,其中N可以具有足以为靶核酸的长片段加标签的任何长度,包括但不限于N=10、11、12、13、14、15、16、17、18、19或20,并且B可以具有容纳所需序列(例如转位子末端、引物结合位点等)的任何长度,此类实施方案可以是$B_{20}N_{15}B_{20}$。

[0200] 在一个实施方案中,两段或三段设计用于用来为长片段加标签的条形码。这种设计通过允许通过将不同的条形码片段接合在一起以形成完整的条形码片段或通过使用片段作为寡核苷酸合成中的试剂来生成组合条形码片段,允许更广泛的可能的条形码。这种组合设计提供了更多可能的条形码,同时减少了需要生成的全尺寸条形码的数量。在进一步的实施方案中,每个长片段的独特识别是用8-12个碱基对(或更长)的条形码实现的。

[0201] 在一个实施方案中,使用两个不同的条形码段。A和B段很容易修改为每个包含不同的半条形码序列,以产生数千种组合。在进一步的实施方案中,条形码序列被并入相同的衔接子上。这可以通过将B衔接子分成两部分来实现,每一部分都有由用于接合的公共重叠序列分隔的半条形码序列。两个标签组分各自具有4-6个碱基。8-碱基(2x4个碱基)标签集能够为65,000个序列独特地加标签。2x5个碱基和2x6个碱基标签两者都可包括使用简并碱基(即"通配符")来实现最佳解码效率。

[0202] 在进一步的实施方案中,每个序列的独特识别是用8-12个碱基对纠错条形码实现的。条形码可以具有一定长度,作为说明而非限制地,该长度为5-20个信息碱基,通常8-16个信息碱基。

5.条形码化珠

[0203] 第一珠和第二珠由具有固定在珠上的条形码的寡核苷酸进行条形码化。每个珠都携带被转移到每个长DNA分子的亚片段中的独特条形码序列的许多副本。

[0204] 所使用的珠可以具有范围为1-20um(替代地为2-8um、3-6um或1-3um,例如约2.8um)的直径。例如,珠上条形码化寡核苷酸的间距可以是至少1、至少2、至少3、至少4、至少5、至少6或至少7nm。在一些实施方案中,间隔小于10nm(例如,5-10nm)、小于15nm、小于20nm、小于30nm、小于40nm或小于50nm。在一些实施方案中,每个混合物使用的不同条形码的数量可以是>1M、>10M、>30M、>100M、>300M或>1B。如下所讨论的,可以产生非常大量的条形码以用于本发明,例如,使用本文所述的方法。在一些实施方案中,每个混合物使用的不

同条形码的数量可以是>1M、>10M、>30M、>100M、>300M或>1B，并且它们是从至少10倍大的多样性库(例如从珠上>10M、>0.1B、0.3B、>0.5B、>1B、>3B、>10B不同的条形码)取样的。在一些实施方案中，每个珠的条形码数量在100k到10M之间，例如在200k和1M，在300k到800k之间，或约400k。

[0205] 在一些实施方案中，条形码区域的长度为约3-15个核苷酸，例如长度为5-12、8-12或10个核苷酸。在一些情况下，条形码区域的每个条形码的长度为约3-12个核苷酸，或长度为3-5个核苷酸。因此，条形码，无论是样本条形码、细胞条形码还是其他条形码，其长度可以是3、4、5、6、7、8、9、10、11、12、13、14、15或更多个核苷酸。在一个特定的示例中，每个条形码区域包括三个条形码，每个条形码由10个碱基组成，并且三个条形码由共同序列的6个碱基分隔开。

[0206] 条形码化珠被转移到靶序列。在一些实施方案中，转移通过将捕获衔接子的3'末端接合至由桥或夹板(术语可互换使用，夹板寡核苷酸的一个示例如图2中的④所示)寡核苷酸(其第一区域与捕获衔接子互补并且第二区域与杂交序列互补)介导的转位子插入的杂交序列的5'末端而定期发生。收集珠并破坏DNA/转位酶复合物，从而产生小于1kb大小的亚片段。在一些实施方案中，转移通过与靶DNA的定向接合而发生。

[0207] 在一些实施方案中，使用三组双链条形码DNA分子通过基于分裂和汇集接合的策略构建条形码化珠。在一些实施方案中，每组双链条形码DNA分子由10个碱基对组成，并且三组的核酸序列不同。PCT公开号WO 2019/217452(其公开内容通过引用以其整体并入本文)中描述了分裂和合并接合以产生条形码化珠的示例性方法。图20和21还图示了分裂和汇集法的方法。包含PCR引物退火位点的常见衔接子序列通过5'双生物素接头附着于Dynabeads™M-280链霉亲和素(ThermoFisher,Waltham,MA)磁珠。通过Integrated DNA Technologies(Coralville,IA)构建了包含重叠序列区域的三组1,536个条形码寡核苷酸。接合在384孔板的15μL反应液中进行，该反应液含有50mM Tris-HCl(pH 7.5)、10mM $MgCl_2$、1mM ATP、2.5％PEG-8000、571个单位的T4接合酶、580pmol条形码寡核苷酸和650万M-280珠。接合反应在室温下在旋转器上温育1小时。在接合之间，通过离心将珠汇集到单个容器中，使用磁铁收集到容器的一侧，并用高盐洗涤缓冲液(50mM Tris-HCl(pH 7.5)、500mM NaCl、0.1mM EDTA和0.05％吐温、20)洗涤一次并用低盐洗涤缓冲液(50mM Tris-HCl(pH 7.5)、150mM NaCl和0.05％吐温20)洗涤两次。将珠重新悬浮在1X接合缓冲液中并分布在384孔板中，并重复接合步骤。

[0208] 本文中提及的某些“条形码”是“三部分条形码”。三部分指的是它们的结构和/或它们的合成。如图20所示，三部分条形码可以通过较短(例如，4-20个核苷酸)序列的连续接合来合成。在一个实施方案中，较短的条形码长度为10个碱基。如该图所示，示例性结构包括CS1-BC1-CS2-BC2-CS3-BC3-CS4，其中CS是存在于所有捕获衔接子上的恒定序列，并且BC序列是不同的10个碱基条形码，如这里所讨论的那样。三部分条形码可以使用部分双链寡核苷酸构建，其中退火至较短的寡核苷酸的结构CSa-BC-CSb是如图所示的BC(即BC')的互补物。

[0209] 在一个方面，本发明提供了一种组合物，该组合物包含附着有包含克隆条形码的捕获寡核苷酸的珠，其中该组合物包含超过30亿个不同的条形码并且其中该条形码是具有结构5'-CS1-BC1-CS2-BC2-CS3-BC3-CS4的三部分条形码。在一些实施方案中，CS1和CS4比

CS2和CS3长。在一些实施方案中,CS2和CS3为4-20个碱基,CS1和CS4为5或10至40(例如20-30)个碱基,并且BC序列的长度为4-20个碱基(例如,10个碱基)。在一些实施方案中,CS4与夹板寡核苷酸互补。在一些实施方案中,组合物包含桥寡核苷酸。在一些实施方案中,组合物包含桥寡核苷酸、包含如上所讨论的三部分条形码的珠和包含具有与桥寡核苷酸互补的区域的杂交序列的基因组DNA。

[0210] 诸如珠或与标签的多个副本关联的其他支持物之类的克隆条形码的另一种来源可以通过乳液PCR或CPG(受控孔玻璃)或其他粒子与制备的自适应条形码的副本的化学合成来制备。可以通过已知方法在油包水(w/o)乳液中在珠上对含有标签的DNA序列群组进行PCR扩增。见例如Tawfik和Griffiths Nature Biotechnology 16:652-656(1998);Dressman等人,Proc.Natl.Acad.Sci.USA 100:8817-8820,2003;以及Shendure等人,Science 309:1728-1732(2005)。这导致每个珠上每一个包含标签的序列都有许多副本。

[0211] 用于制备克隆条形码的来源的另一种方法是在“混合和分裂”组合过程中通过微珠或CPG上的寡核苷酸合成。使用这个过程,可以创建一组珠,每个珠都有一个条形码的一群副本。例如,要使得全部为$B_{20}N15B_{20}$,其中大约10亿个中的每个在100个珠中的每个珠上以～1000+个副本表示,平均而言,可以从～1000亿个珠开始,在它们全部上合成B_{20}共同序列(衔接子),然后将它们分裂为1024个合成柱以在每个中制备不同的5聚体,然后将它们混合,然后将它们再次分裂为1024个柱并制备额外的5聚体,然后再次重复以完成N15,然后将它们混合并在一个大柱中合成最后一个B_{20}作为第二衔接子。因此,在3050次合成中,可以使用～10000亿个珠(1^{12}个珠)制备与一个大型仿真PCR反应相同的“克隆样”条形码集,因为只有十分之一的珠具有起始模板(其他9个将没有)以防止每个珠有两个带有不同条形码的模板。

[0212] 条形码序列组装的示例性过程在图21中示出。

6.UMI

[0213] 在各种实施方案中,独特分子标识符(UMI)用于将个体DNA分子彼此区分开。例如,UMI用于在固定在第一珠上的捕获寡核苷酸之间进行区分(例如,方案I步骤2)。生成衔接子的集合,每个衔接子都有一个UMI,这些衔接子连接到要测序的片段或其他源DNA分子上,已测序的个体分子各自都具有有助于将其与所有其他片段区分开来的UMI。在此类实现方式中,可使用大量不同(例如,数千至数百万个)的UMI来独特地识别样本中的DNA片段。

[0214] UMI的长度足以确保每一个源DNA分子的独特性。在一些实施方案中,独特分子标识符的长度为约3-12个核苷酸,或长度为3-5个核苷酸。在一些情况下,每个独特分子标识符的长度为约3-12个核苷酸,或长度为3-5个核苷酸。因此,独特分子标识符的长度可以是3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18或更多个核苷酸。

7.扩增

[0215] 在一些实施方案中,在方法步骤中产生的亚片段或片段被扩增。此类扩增方法包括但不限于:多重置换扩增(MDA)、聚合酶链反应(PCR)、接合链反应(有时称为寡核苷酸接合酶扩增OLA)、循环探针技术(CPT)、链置换分析(SDA)、转录介导扩增(TMA)、基于核酸序列的扩增(NASBA)、滚环扩增(RCR)(用于环化片段)和侵入性切割技术。可以在片段化之后或在本文概述的任何步骤之前或之后进行扩增。

8.逆转录

[0216] 逆转录由逆转录酶介导,其使用RNA作为模板来合成与RNA模板互补的cDNA。逆转录酶通常具有RNA依赖性DNA聚合酶活性和RNase H活性。在某些情况下,逆转录酶还具有DNA依赖性DNA聚合酶活性。逆转录通常包括多个步骤:1.在退火的引物的存在下,逆转录酶与RNA模板结合并启动反应;2.RNA依赖性DNA聚合酶活性合成互补的DNA链,从而并入dNTP;3.RNase H活性降解DNA:RNA复合物的RNA模板;4.DNA依赖性DNA聚合酶活性(如果存在)识别单链cDNA为模板,将RNA片段用作引物,并且合成第二链cDNA以形成双链cDNA。

[0217] 逆转录酶的非限制性示例包括;来自人类免疫缺陷病毒1型的HIV-1逆转录酶;来自莫洛尼鼠白血病病毒的M-MLV逆转录酶;来自禽成髓细胞瘤病毒的AMV逆转录酶;维持真核染色体的端粒的端粒酶逆转录酶。

9.3'分支接合

[0218] 3'分支接合包括来自平端衔接子(供体DNA)的5'磷酸在3'凹陷链、缺口或切口处共价连接到双链DNA受体的3'羟基端。与传统的DNA接合相比,3'分支接合不需要互补碱基配对。3'分支接合被描述于PCT公开号WO 2019/217452;US专利公开US2018/0044668和国际申请WO 2016/037418,US专利公开2018/0044667,以及Wang等人,2018年6月29日,http://dx.doi.org/10.1101/357863(其全部通过引用并入以用于所有目的)中。使用该方法,理论上可对捕获的基因组分子的所有亚片段进行扩增和测序。因此,3'分支接合具有广泛的分子应用,包括,例如,在NGS文库制备期间将衔接子连接到DNA或RNA。

[0219] 此外,该接合步骤能够将样本条形码放置在基因组序列附近以进行多重采样(sampling multiplexing)。使用这些衔接子进行样本条形码化的益处在于条形码可以放置在基因组DNA附近,使得可以使用相同的引物对条形码和基因组DNA进行测序,并且不需要额外的测序引物来读取条形码。样本条形码化允许在序列之前汇集来自多个样本的制备物,并通过条形码进行区分。3'分支接合衔接子可以在96、384或1536板格式中合成,其中每个孔包含携带相同条形码的衔接子的许多副本,并且孔之间的每个条形码都不同。在珠上捕获后,这些衔接子可用于96、384或1536板格式的接合。

[0220] 在该接合步骤之后,进行PCR并且文库准备好进入任何标准的下一代测序(NGS)工作流程。应当理解,可以使用与捕获寡核苷酸或其互补物上的位点杂交的第一引物(见图W1A)和与3'分支接合衔接子或其互补物上的位点杂交的第二引物来进行PCR(或其他扩增)。在BGISEQ-500的情况下,文库如先前所述被环化(17)。从单链环制备DNA纳米球并将其加载到图案化的纳米阵列上(17)。然后对这些纳米阵列在BGISEQ-500上进行基于组合探针-锚合成(cPAS)的测序(18-20)。测序后,提取条形码序列。通过独特条形码映射读段数据显示,大多数带有相同条形码的读段都聚集在与文库制备期间使用的DNA长度相对应的基因组的区域中(图1B)。实施例1和2中描述了此方法的详细说明以及用于制备珠的方案。

10.使用包含位置条形码的转位子进行转位

[0221] 在一些情况下,根据上述任何实施方案插入靶核酸中的转位子可携带独特位置条形码。此位置条形码表示每个转位子相对于其他转位子的位置。此位置条形码可用于组装包含序列重复的区域的序列读段。如图23例示了包含四个序列区域I、II、III和IV的基因组区域。在该区域中,I、II和IV是重复的,III与I、II或IV仅相差一个核苷酸。在没有位置条形码的情况下,将难以确定特定的序列读段属于区域I、II、III还是IV。使用位置条形码可以避免这个问题,通过将序列读段与区域I、II、III或IV所独有的特定位置条形码关联,可以

将序列读段分配给特定区域。因此，使用携带位置条形码的转位子可以方便地对重复序列进行排序，并且与通过珠引入的条形码一起，可以有效地确定长距离重复区域的序列。

[0222] 可以通过将一个或多个转位子支架与靶核酸分子接触来实现将带有位置条形码的转位子插入靶核酸(例如，基因组DNA分子)中。转位子支架以预定间距保持带有独特位置条形码的转位子，所述间距决定了转位子被插入靶核酸分子后的间隔大小。下面对这些方法的实施方案作进一步描述。

支架

[0223] 如本文所述，插入支架是包括支架和锚定在其上的衔接子的复合结构。衔接子是双链或部分双链的。支架可以是能将衔接子保持在所需间距的任何合适的材料。支架的非限制性示例包括核酸、蛋白质、碳水化合物分子和可溶于水溶液的其他长化学结构。在一些实施方案中，支架是单链核酸分子，例如衍生自已知质粒(如PUC 19)的单链DNA。支架核酸分子的大小可以变化。在一些实施方案中，支架的大小可以在1至50kb(例如1至30kb、2-30kb或5至10kb)的范围内。衔接子包含可以插入到靶DNA片段中的位置条形码和/或支架条形码。在一些实施方案中，插入是通过3'分支接合进行的。衔接子可以是转位子。衔接子包含位置条形码和/或支架条形码，其通过转位插入到DNA片段中。然而，本领域技术人员将理解，可以生产具有各种位置条形码和/或支架条形码的任何双链或部分双链衔接子，并且这些衔接子可以以与转位子相同的方式锚定到支架。如上所述，这些衔接子可以通过例如3'分支接合插入到靶DNA片段中。

转位子的序列特征

转位子中的支架杂交序列

[0224] 在一些实施方案中，在每个支架中使用的衔接子(例如转位子)包含特定支架杂交序列，即与支架中的不同区域互补的序列，使得当衔接子(例如转位子)与支架混合时，衔接子(例如转位子)经由这些杂交序列以所需间距锚定到支架上。支架杂交序列被设计为可切割的，使得在通过转位并入靶核酸后，杂交序列可以被切割以使衔接子(例如转位子)与支架解离。在支架杂交序列裂解后，支架条形码和位置条形码保留在插入靶核酸分子中的衔接子(例如转位子)中。然后可以通过例如洗涤从反应中除去已解离的支架。

支架条形码

[0225] 衔接子(例如转位子)可以包含每个转位子支架特有的支架条形码。相同转位子支架中的衔接子(例如转位子)共享相同的支架条形码，并且来自不同转位子支架的衔接子(例如转位子)具有不同的支架条形码。这些支架条形码允许将对应于相同支架的序列信息分组。图23示出了许多转位子支架S1,S2,...,Sn的示例性设计。S1支架中的所有衔接子(例如转位子)共享相同的S1支架条形码，并且S2支架中的所有衔接子(例如转位子)共享相同的S2支架等。用于不同转位子支架的支架条形码S1,S2,S3,...,Sn都是可区分的。

位置条形码

[0226] 相同支架中的每个转位子可携带独特位置条形码，并且相同支架内的不同衔接子(例如转位子)携带不同的位置条形码。图23示出了包括包含位置条形码的衔接子(例如转位子)的转位子支架的示例。例如，由支架S1保持的衔接子(例如转位子)具有相同的S1支架条形码，但具有不同的位置条形码。1和2是第一转位子上的位置条形码，3和4是第二转位子上的位置条形码，5和6是第三转位子上的位置条形码，7和8是第四转位子上的位置条形码

等。

[0227] 在某些情况下,相同转位子的两个插入寡核苷酸可具有相同的位置条形码;例如,图24中的位置条形码1和2可以是相同的。在某些情况下,相同转位子的两个插入寡核苷酸可具有不同的条形码序列;例如,位置条形码1和2可以是不同的。

生产插入支架

[0228] 具有上述一种或多种序列特征(例如位置条形码、支架条形码和/或支架杂交序列)的衔接子(例如转位子)可以经由适于支架的化学物锚定在支架上。在一些情况下,支架是单链核酸,并且衔接子(例如转位子)经由衔接子(例如转位子)中包含的支架杂交序列在特定位置处与支架杂交。在一些情况下,单链核酸支架的序列是已知的,使得衔接子(例如转位子)的支架杂交序列可以设计成确保衔接子(例如转位子)以所需频率锚定在所需位置。换句话说,个体衔接子(例如转位子)的位置和相邻衔接子(例如转位子)之间的间距两者都可以预先确定。在某些情况下,沿着支架以范围从300个碱基到5kb(例如400个碱基到4kb、500个碱基到2kb或约500个碱基到1kb)的间距锚定衔接子(例如转位子)。相同支架中相邻衔接子(例如转位子)之间的间距不必相同,尽管在一些实施方案中,相同支架中相邻衔接子(例如转位子)之间的间距是相同的。每个支架的衔接子(例如转位子)的数量也可有所不同,具体取决于支架的大小和衔接子(例如转位子)之间所需的间距。在一些情况下,每个支架的衔接子(例如转位子)的数量可以在3到20(例如4到15,或5到10)的范围内。图24示出了示例性转位子支架,其中六个衔接子(例如转位子)与其序列已知的单链DNA分子杂交。

[0229] 在某些实施方案中,支架是经修饰以含有化学修饰位点的蛋白质或底物,该位点可用于将衔接子(例如转位子)共价或非共价地连接至底物。类似于其中支架是核酸分子的转位子支架,这些类型的转位子支架也可以设计成具有以所需间距在所需位置处锚定到支架上的衔接子(例如转位子)。

用转位子支架进行转位

[0230] 用于使每个靶核酸转位的转位子支架的数量可以变化。其可以基于靶核酸的大小和每个支架的大小来确定以确保靶核酸的足够覆盖率。在一些实施方案中,使用足够数量的支架使得支架的总长度等于或大于靶核酸的长度。

[0231] 在转位酶的存在下,在适于转位的条件下,转位子支架可以与靶核酸(基因组DNA)混合。然后将衔接子(例如转位子)插入靶核酸中。见图25A。如上所述,在转位之后,并入的衔接子(例如转位子)的支架杂交序列可被切割以去除支架。以这种方式,当衔接子(例如转位子)并入靶核酸时,它们在支架中的位置信息得以保留。

带有位置条形码的衔接子(例如转位子)

[0232] 带有位置条形码和支架条形码的衔接子(例如转位子)可用于上述任何方案中。在一些情况下,使用具有位置条形码的衔接子(例如转位子)进行一轮转位。如图25A中例示的,在转位之后,核酸片段使用stLFR珠被捕获并如上所述进行条形码化。然后去除转位酶以释放由转位酶保持在一起的个体片段。然后可以使用本领域公知的方法扩增片段。在某些情况下,使用RCR将这些片段环化和扩增。在一些实施方案中,使用stLFR珠捕获片段并如上所述进行共条形码化。然后去除转位酶并通过RCR扩增个体片段。

[0233] 在一些实施方案中,可以使用具有位置条形码的衔接子(例如转位子)进行两轮转位。通常,在第一轮转位后,将由相邻衔接子(例如转位子)定义的片段(第一亚片段)扩增并

对扩增的第一亚片段进行第二轮转位。根据扩增方法，扩增的第一亚片段可以是双链或单链的。如果扩增的第一亚片段是双链DNA，则可以直接对它们进行第二轮转位处理。如果扩增的第一亚片段是单链DNA，则在进行第二轮转位之前将它们转化为双链DNA。第二轮转位产生第二复合物，然后可以在stLFR珠上捕获第二复合物并其进行共条形码化。去除转位酶后，可以对亚片段进行处理以进行测序。

[0234] 图25A和25B示出了本发明一示例性实施方案，其中使用具有位置条形码的衔接子(例如转位子)进行两轮转位。图25A示出了第一轮转位导致形成3-5kb大小的分子(第一亚片段)，其在stLFR珠上捕获、共条形码化、环化并通过RCR扩增。扩增产物被转化为双链DNA。图25B示出了对从第一轮转位产生的3-5kb双链DNA进行第二轮转位。这在相邻的衔接子(例如转位子)之间产生500bp的片段(第二亚片段)，并且可以对每个500bp片段进行测序以生成序列读段。

[0235] 在一些实施方案中，第二轮转位中使用的衔接子(例如转位子)中的位置条形码不同于第一轮转位中使用的衔接子(例如转位子)中的位置条形码。在一些实施方案中，第二轮转位中使用的衔接子(例如转位子)中的位置条形码与第一轮转位中使用的衔接子(例如转位子)中的一些位置条形码相同。

[0236] 在第一轮和第二轮转位中，每个转位子的插入导致相邻亚片段之间有9bp的重叠，如图26所示。如下进一步所述，该信息可用于对支架之间的亚片段进行排序。

使用位置条形码和支架条形码进行从头测序

[0237] 本节中描述的使用携带支架条形码和位置条形码的转位子方法提供关于来自测序片段的每个序列读段的相对位置的重要信息，并提高长靶核酸分子的从头测序的准确性和效率。位置条形码可用于对每个支架内的序列进行排序。此外，每个转位子插入导致相邻亚片段(例如，从第一轮转位产生的第一亚片段和从第二轮转位产生的第二亚片段)之间的重叠。例如，Tn5的转位通常导致9bp的序列重叠。在典型的基因组DNA片段的背景下，重叠的序列是独一无二的，并且可以很容易地在测序结果中定位该重叠。与支架条形码组合的序列重叠可用于对与来自不同支架的转位子关联的序列进行排序。例如，共享序列重叠但携带不同支架条形码的两个序列的结果表明这两个序列与来自两个相邻支架的两个转位子关联，并且这两个转位子位于其各自支架的末端。

[0238] 在使用具有上述特征的转位子进行两轮转位并且每轮转位之后还进行stLFR珠捕获和共条形码化的情况下，对所得第二亚片段(例如图25B中的500bp片段)的测序可以在很多层上生成丰富的序列信息。例如，这些第二亚片段可以包含：1)第一支架条形码，2)第一位置条形码，3)由第一次转位产生的9bp重叠，4)第二支架条形码，5)第二位置条形码，6)第一珠条形码，7)第二珠条形码，8)由第二转位产生的9bp重叠，或其组合。所有这些特征都可以在序列读段中检测到，并且这些特征的任何组合都可以大大提高从头测序的准确性和效率。

[0239] 图26图示了100kb基因组DNA分子的从头测序的示例。进行两轮转位，并且每轮后的转位片段使用stLFR珠捕获并进行条形码化。对第二轮转位和捕获后得到的基因组片段进行测序。根据第二转位子上的位置条形码和支架条形码对来自第二亚片段的序列读段进行排序，并且由转位产生了重叠以构建每个第一亚片段的序列。然后在第一轮转位中使用转位子中的位置条形码对多个第一亚片段的序列信息进行排序以构建第一复合物的序列，

例如100kb基因组DNA分子。

11.测序

[0240] 本文所述的方法可以用作使用本领域已知的任何测序方法对二倍体基因组进行测序的预处理步骤,该测序方法包括例如但不限于基于聚合酶的边合成边测序(例如,HiSeq 2500system,Illumina,San Diego,CA),基于接合的测序(例如,SOLiD 5500,Life Technologies Corporation,Carlsbad,CA),离子半导体测序(例如,Ion PGM or Ion Proton sequencers,Life Technologies Corporation,Carlsbad,CA),零-模式波导(例如,PacBio RS sequencer,Pacific Biosciences,Menlo Park,CA),纳米孔测序(例如,英国牛津的Oxford Nanopore Technologies Ltd.),焦磷酸测序(例如,454Life Sciences,Branford,CT),或其他测序技术。这些测序技术中一些是短读段技术,但其他技术产生较长的读段,例如GS FLX+(454Life Sciences;高达1000bp)、PacBio RS(Pacific Biosciences;大约1000bp)和纳米孔测序(Oxford Nanopore Technologies Ltd.;100kb)。对于单倍体定相,较长的读段是有利的,从而需要较少的计算,尽管它们往往具有更高的错误率并且在这种长读段中的错误可能需要在单倍体定相之前根据本文所述的方法进行识别和纠正。

[0241] 根据一个实施方案,使用组合探针-锚连接(cPAL)进行测序,如例如在US 20140051588、U.S.20130124100(两者通过引用以其整体并入本文以用于所有目的)中所描述的那样。

[0242] 用于调用与参考多核苷酸序列相比的多核苷酸序列中的变异和用于多核苷酸序列组装(或重新组装)的示例性方法,例如被提供在美国专利公开号2011-0004413(申请号12/770,089,其全部内容通过引用以其整体并入本文以用于所有目的)中。也见Drmanac等人,Science 327,78-81,2010。同样通过引用以其整体并入并用于所有目的的是共同未决的题为“Identification Of DNA Fragments And Structural Variations”的相关申请号61/623,876;作为美国专利公开2013-0096841公布的申请号13/649,966;以及作为美国专利公开2013/0124100公开的题为“Processing and Analysis of Complex Nucleic Acid Sequence Data”的申请号13/447,087。

12.组合物

[0243] 还提供了包含多个插入支架和靶核酸片段的核酸复合物。多个插入支架与靶核酸片段杂交并且多个插入支架中的每个包含多个双链或部分双链衔接子和支架。衔接子被锚固在支架上并以预定间距分开。每个插入支架中的每个衔接子都携带独特位置条形码,并且相同插入支架中的所有衔接子共享一个共同的支架条形码,并且不同插入支架中的衔接子具有不同的支架条形码。

在一些方式中,一个或多个衔接子是转位子。

[0244] 还提供了一种组合物,其包括:(1)包含附着于其上的捕获寡核苷酸的克隆副本的珠群组;和(2)以上一种或多种插入支架或一种或多种核酸复合物。每个珠包含带有相同条形码的多个捕获寡核苷酸,并且该群组中不同的珠包含不同的条形码。

[0245] 还提供了在单个容器中的反应混合物,其中该反应混合物包含本文公开的多个插入支架和多个核酸片段。反应混合物可以还包含以下一种或多种:

i)核酸外切酶;

ii)DNA聚合酶；

iii)尿嘧啶-DNA糖基化酶；

iv)接合酶；

iv)3'分支接合衔接子,以及

v)转位酶。

[0246] 还提供了包含上述多个核酸复合物的阵列。

***

[0247] 虽然已经参考特定方面和实施方案公开了本发明,但很明显的是,本领域的其他技术人员可以设计本发明的其他实施方案和变形而不脱离本发明的真实精神和范围。

[0248] 出于在美国的所有目的,在本公开中引用的每一个出版物和专利文件都通过引用并入本文,就如同每个此类出版物或文件被具体地和单独地指示为通过引用并入本文。出版物和专利文件的引用并不表示任何此类文件是相关的现有技术,也不构成对其内容或日期的承认。

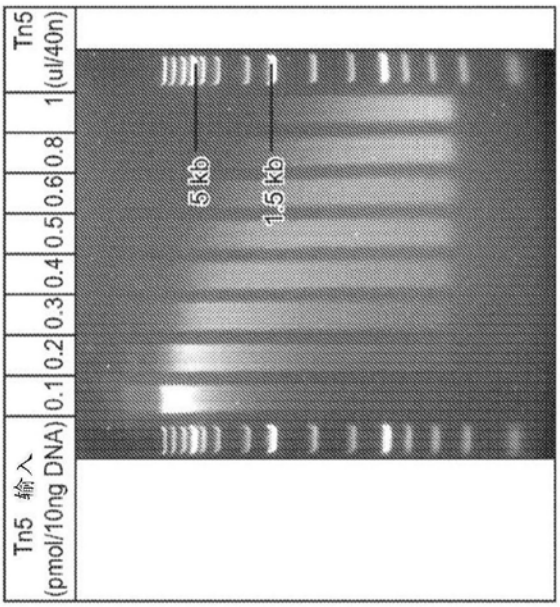


图1

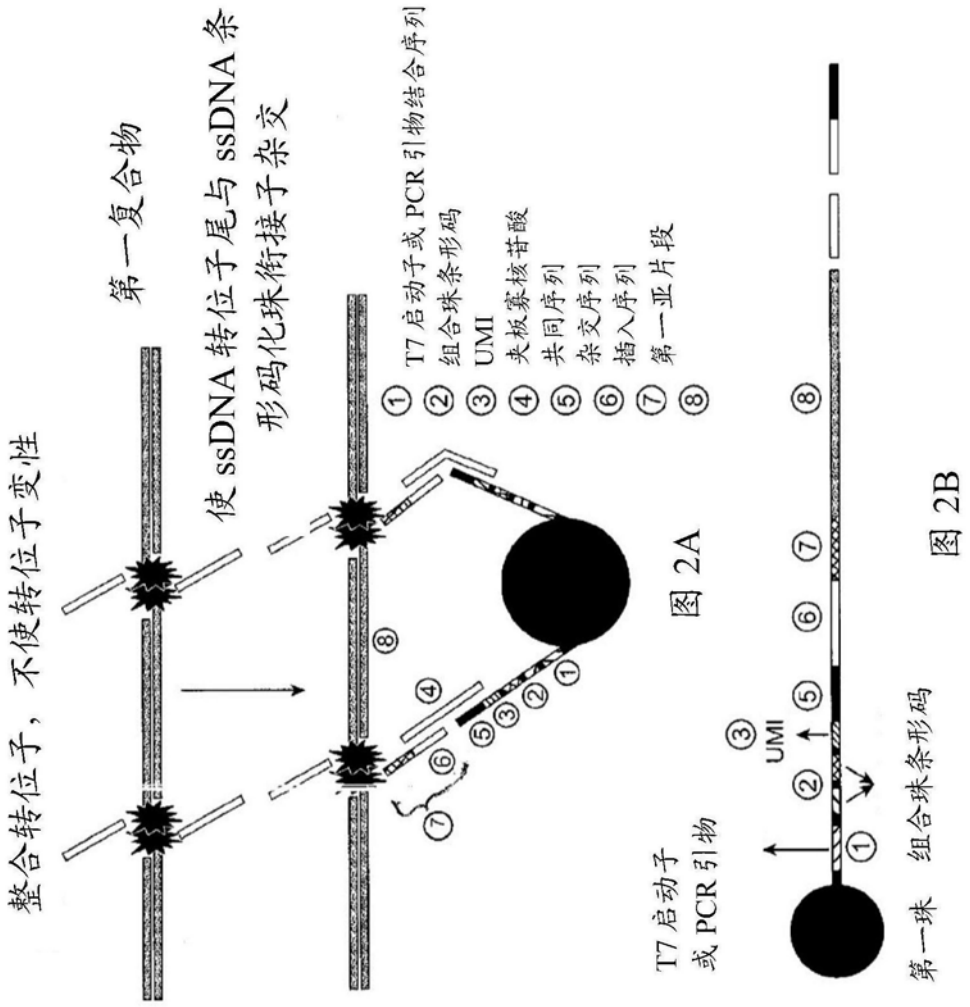

整合转位子，不使转位子变性
第一复合物
使 ssDNA 转位子尾与 ssDNA 条形码化珠衔接子杂交
① T7 启动子或 PCR 引物结合序列
② 组合珠条形码
③ UMI
④ 夹板寡核苷酸
⑤ 共同序列
⑥ 杂交序列
⑦ 插入序列
⑧ 第一亚片段
T7 启动子
或 PCR 引物
UMI
第一珠 组合珠条形码

图 2A

图 2B

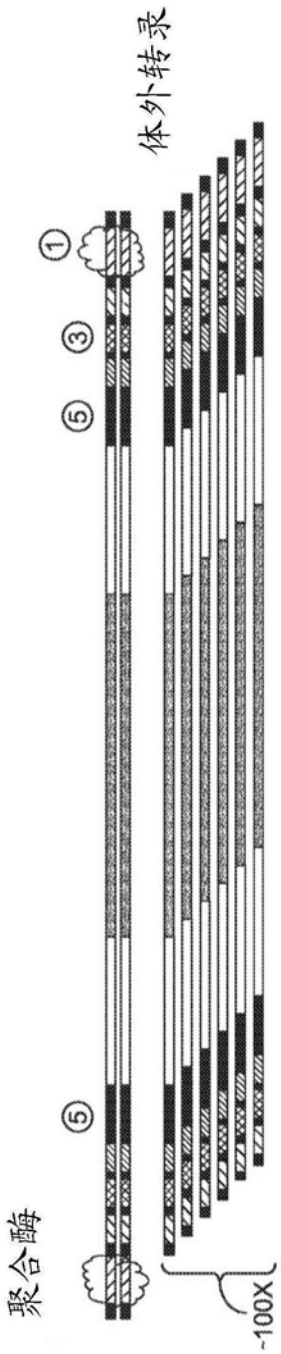


图 3

逆转录以合成
cDNA

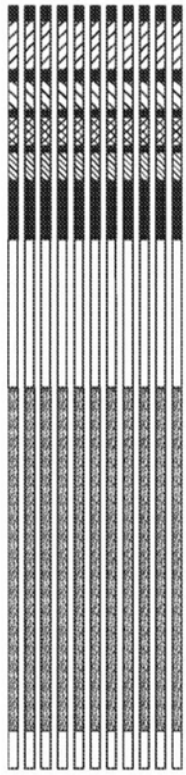


图 4

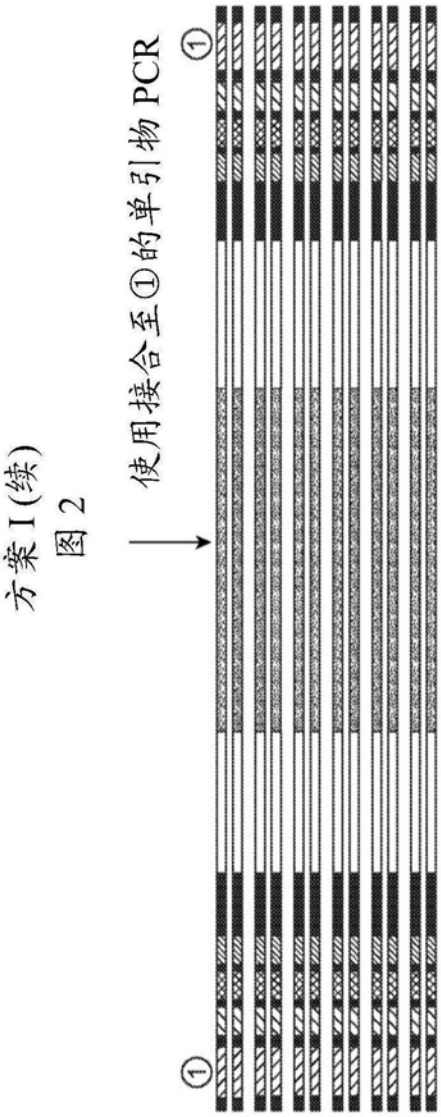
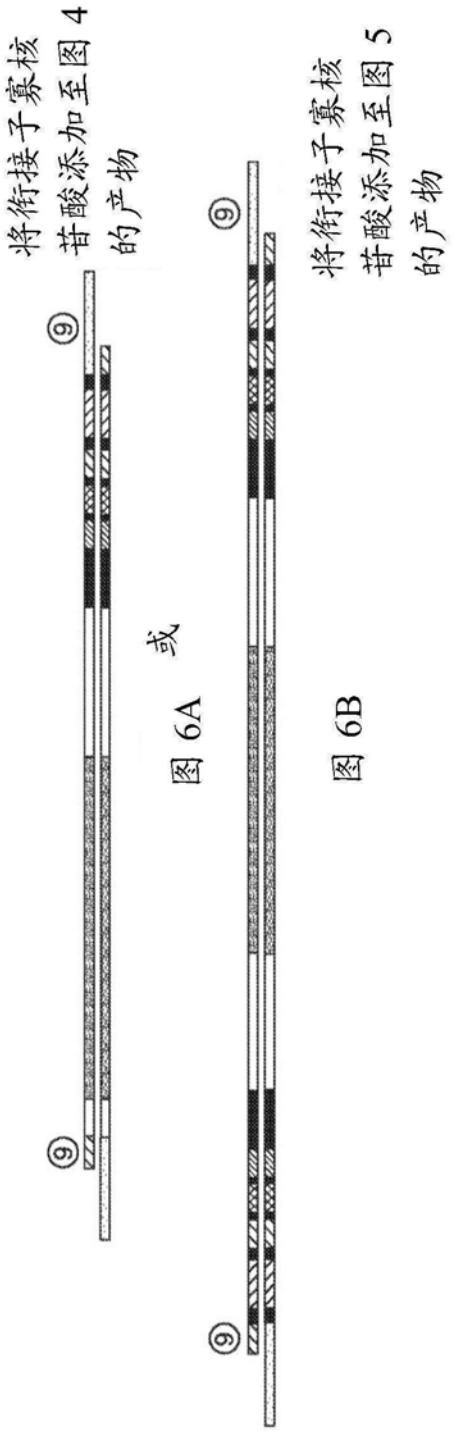
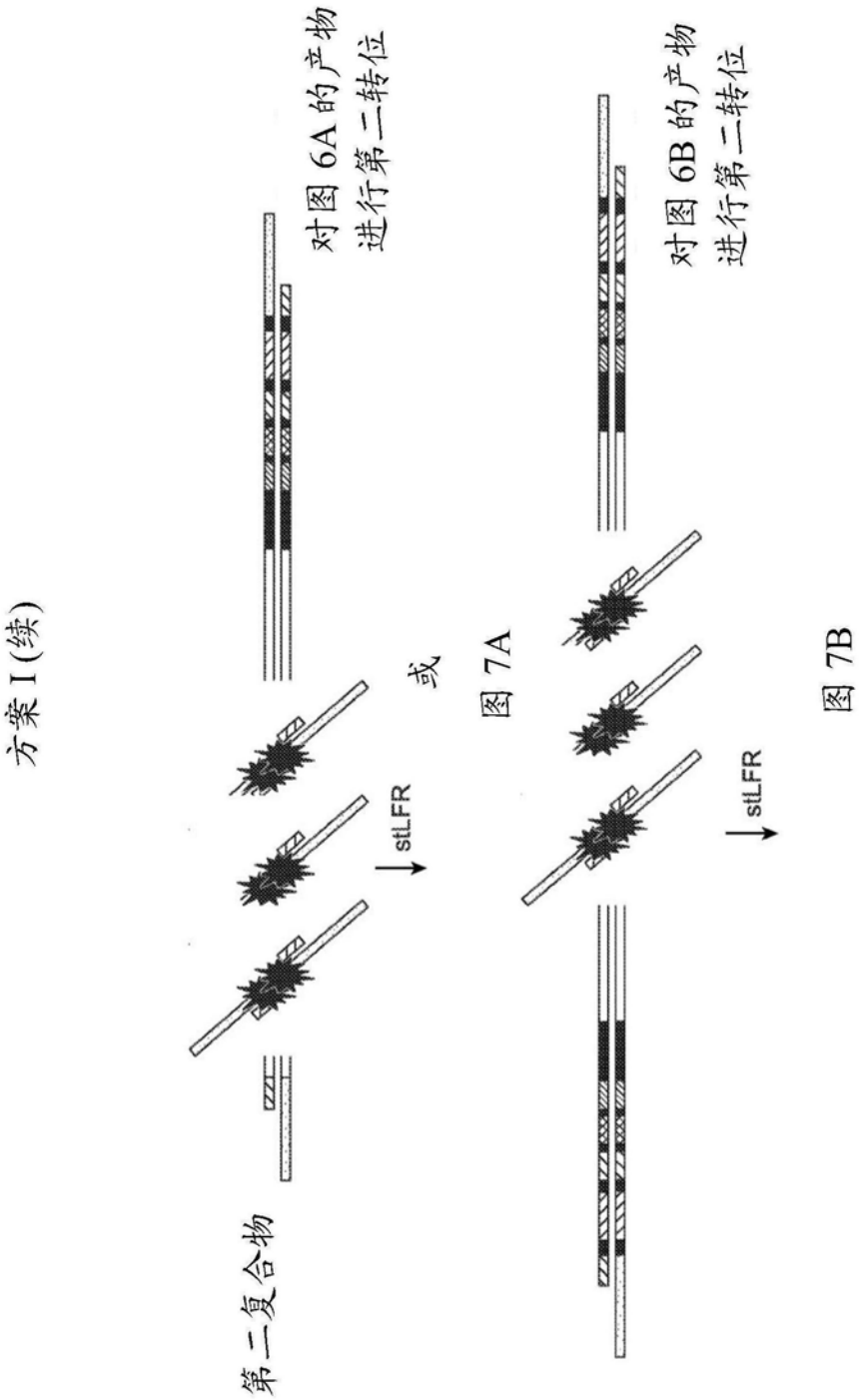


图5


将衔接子寡核
苷酸添加至图 4
的产物
⑨
⑨
或
图 6A
将衔接子寡核
苷酸添加至图 5
的产物
⑨
⑨
图 6B


方案 I (续)
第二复合物
stLFR
对图 6A 的产物
进行第二转位
或
图 7A
stLFR
对图 6B 的产物
进行第二转位
图 7B

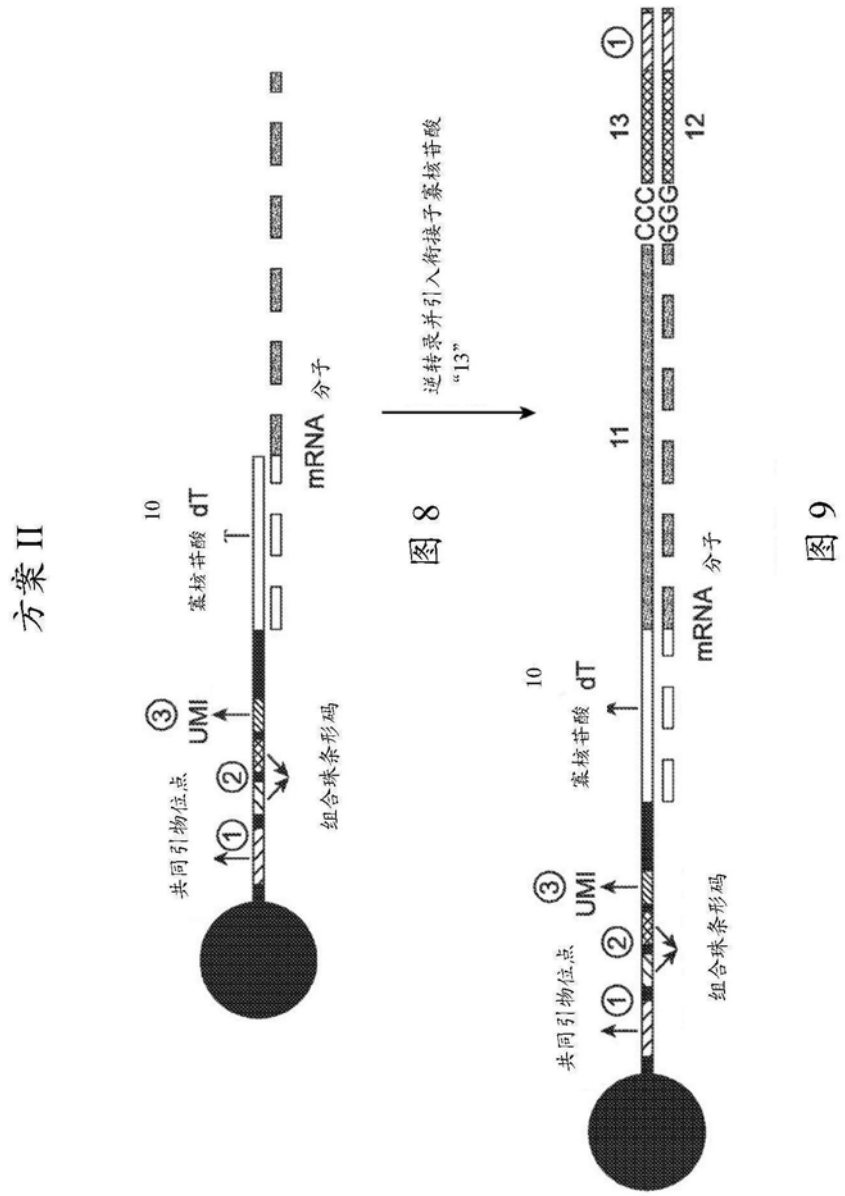

方案II
10
寡核苷酸 dT
T
mRNA 分子
③
UMI
②
①
共同引物位点
组合珠条形码
逆转录并引入衔接子寡核苷酸
“13”
11
13
12
CCC
GGG

图 8

图 9

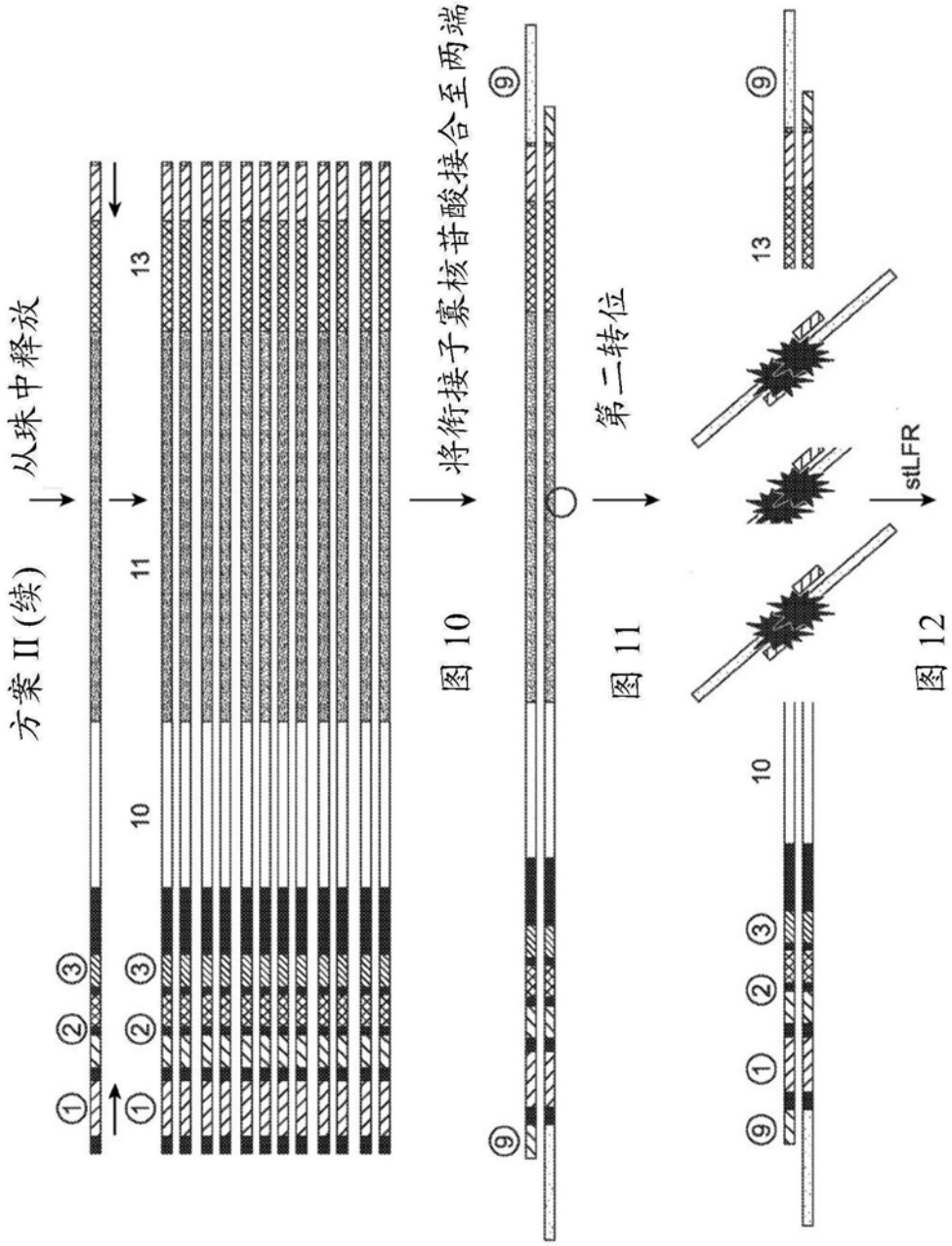


图 10 图 11 图 12

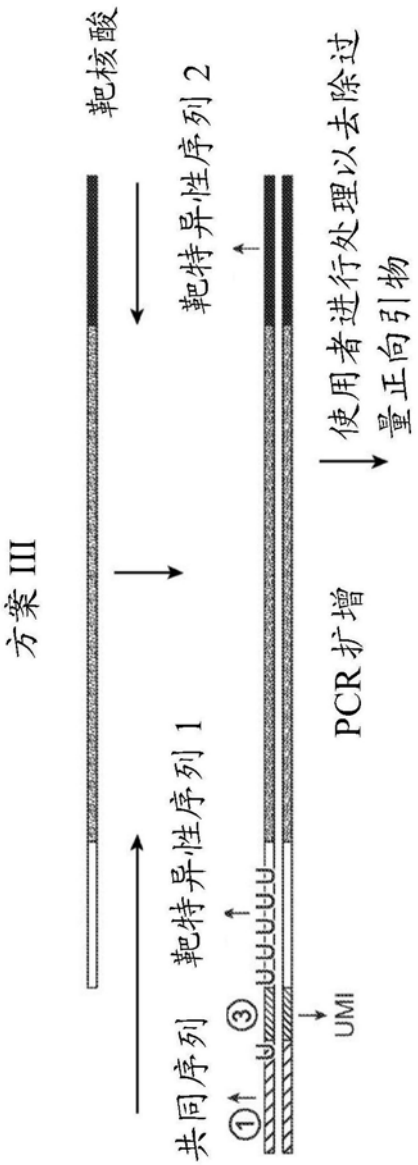

方案 III
靶核酸
靶特异性序列 2
靶特异性序列 1
共同序列
①
③
UMI
PCR 扩增
使用者进行处理以去除过量正向引物

图13

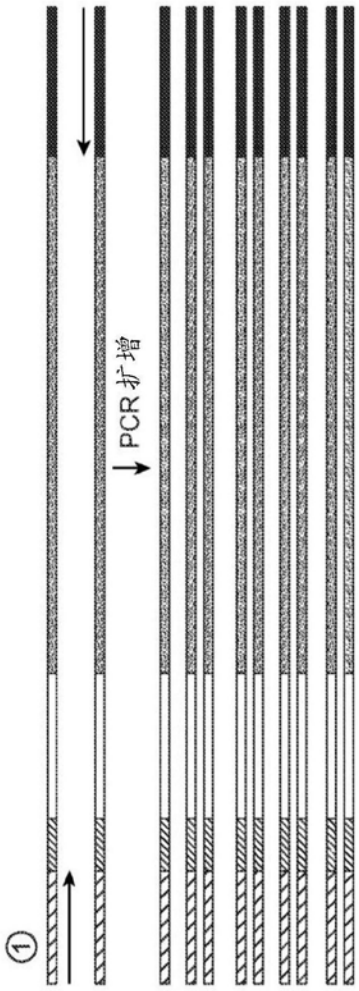
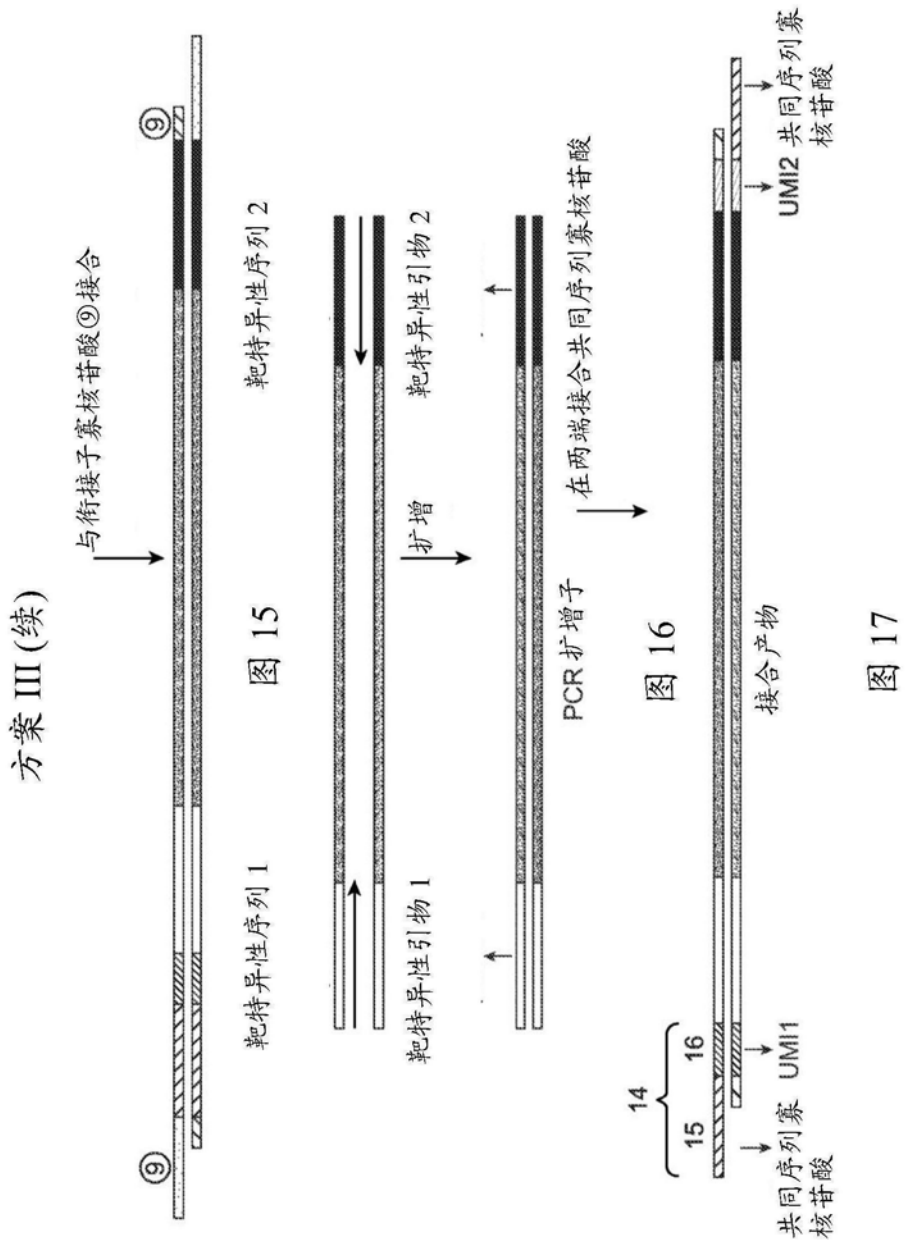


图14


方案 III (续)
与衔接子寡核苷酸⑨接合
⑨
⑨
靶特异性序列 1
靶特异性序列 2
图 15
扩增
靶特异性引物 1
靶特异性引物 2
PCR 扩增子
在两端接合共同序列寡核苷酸
图 16
14
15
16
共同序列寡核苷酸
UMI1
接合产物
UMI2
共同序列寡核苷酸
图 17

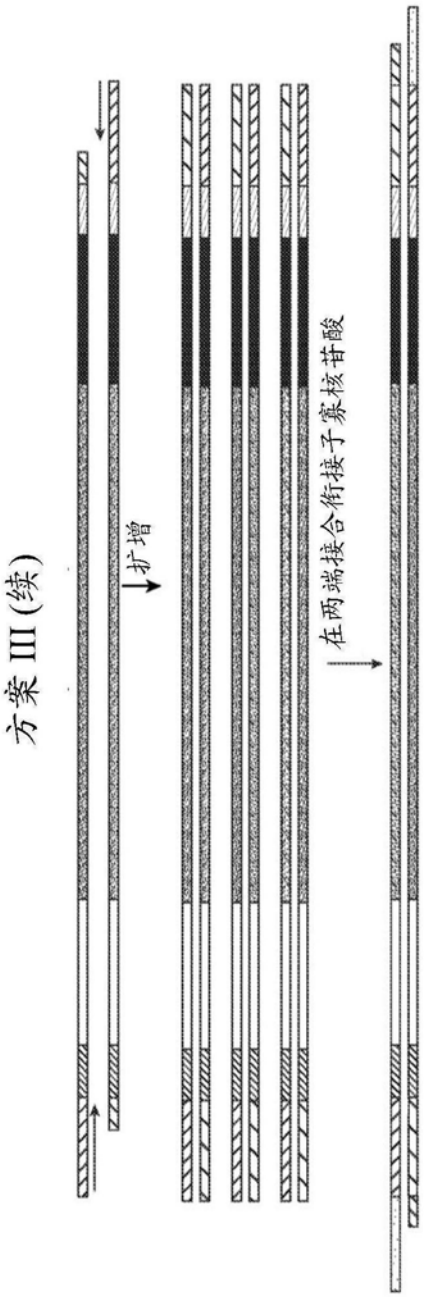

方案 III (续)
扩增
在两端接合衔接子寡核苷酸

图 18

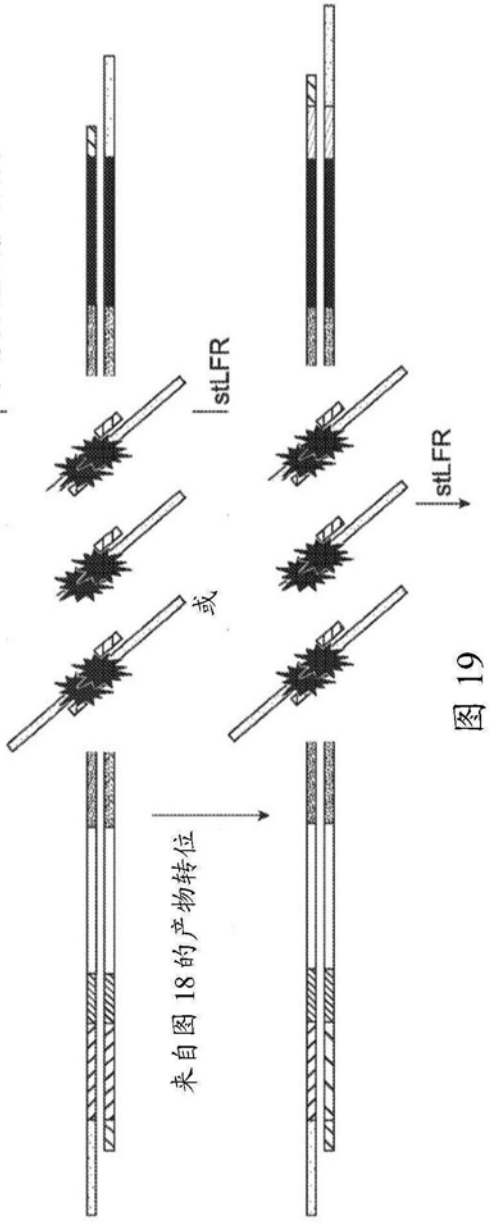

来自图 15 的产物转位
stLFR
或
来自图 18 的产物转位
stLFR

图 19

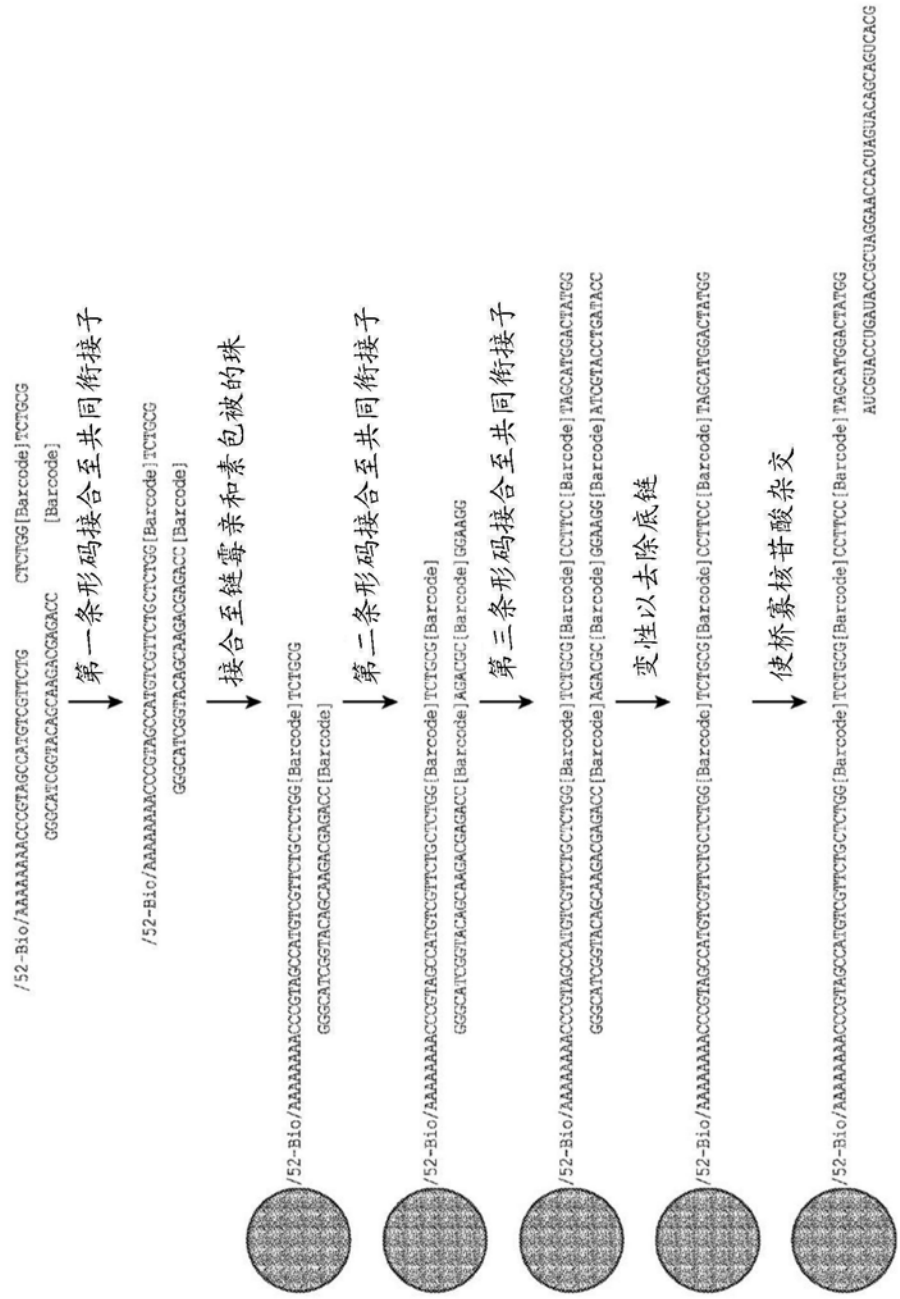

/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTG CTCTGG[Barcode]TCTGCG
GGGCATCGGTACAGCAAGACGAGACC [Barcode]
第一条形码接合至共同衔接子
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG[Barcode]TCTGCG
GGGCATCGGTACAGCAAGACGAGACC[Barcode]
接合至链霉亲和素包被的珠
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG[Barcode]TCTGCG
GGGCATCGGTACAGCAAGACGAGACC[Barcode]
第二条形码接合至共同衔接子
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG[Barcode]TCTGCG[Barcode]
GGGCATCGGTACAGCAAGACGAGACC[Barcode]AGACGC[Barcode]GGAAGG
第三条形码接合至共同衔接子
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG[Barcode]TCTGCG[Barcode]CCTTCC[Barcode]TAGCATGGACTATGG
GGGCATCGGTACAGCAAGACGAGACC[Barcode]AGACGC[Barcode]GGAAGG[Barcode]ATCGTACCTGATACC
变性以去除底链
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG[Barcode]TCTGCG[Barcode]CCTTCC[Barcode]TAGCATGGACTATGG
使桥寡核苷酸杂交
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG[Barcode]TCTGCG[Barcode]CCTTCC[Barcode]TAGCATGGACTATGG
AUCGUACCUGAUACCGCUAGGAACCACUAGUACAGCAGUCACG

图20

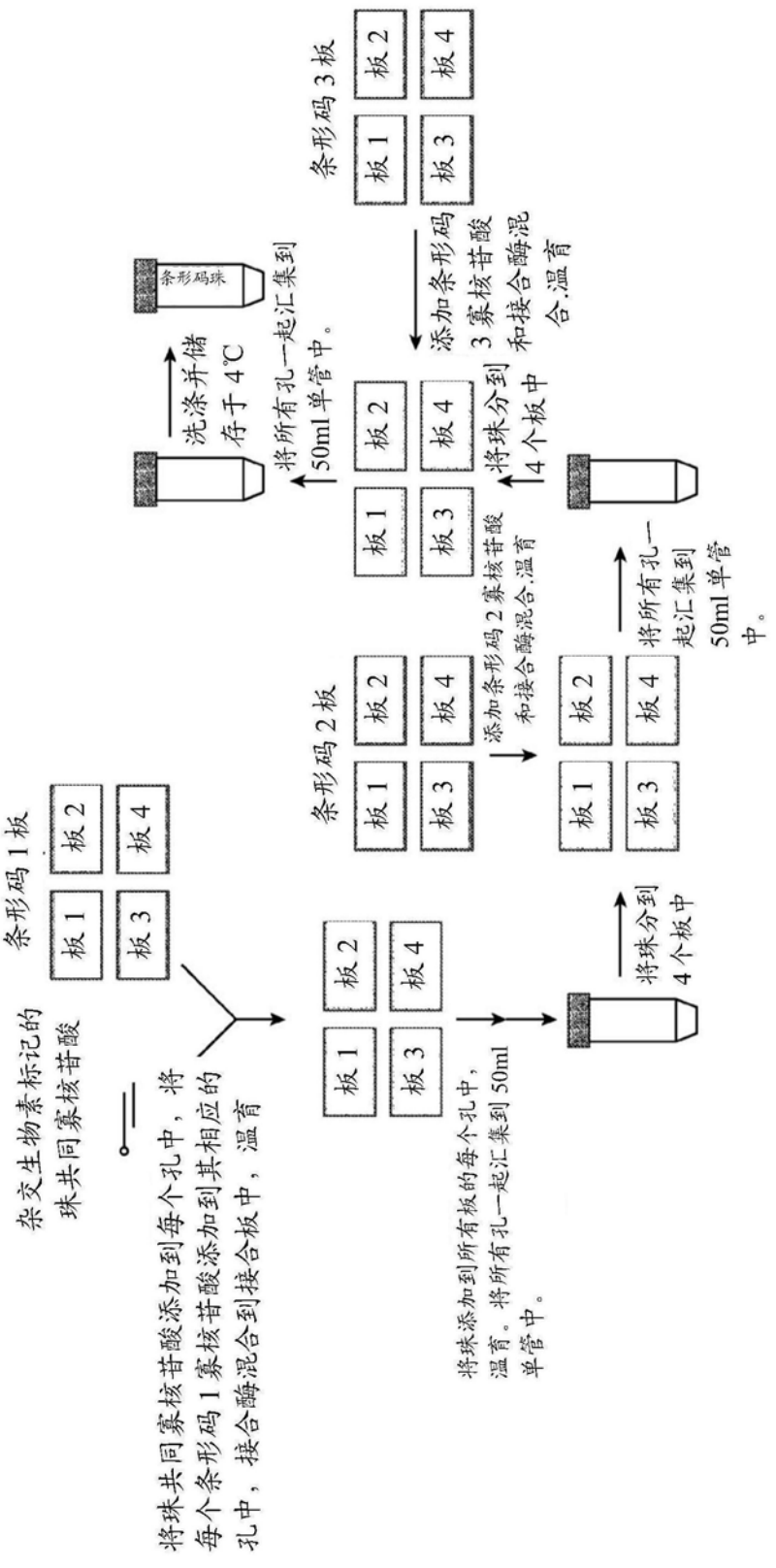

杂交生物素标记的
珠共同寡核苷酸
将珠共同寡核苷酸添加到每个孔中，将
每个条形码1寡核苷酸添加到其相应的
孔中，接合酶混合到接合板中，温育
条形码1板
板1
板2
板3
板4
将珠添加到所有板的每个孔中，
温育。将所有孔一起汇集到50ml
单管中。
将珠分到
4个板中
条形码2板
添加条形码2寡核苷酸
和接合酶混合,温育
将所有孔一
起汇集到
50ml单管
中。
将珠分到
4个板中
条形码3板
添加条形码
3寡核苷酸
和接合酶混
合,温育
将所有孔一起汇集到
50ml单管中。
洗涤并储
存于4℃
条形码珠

图21

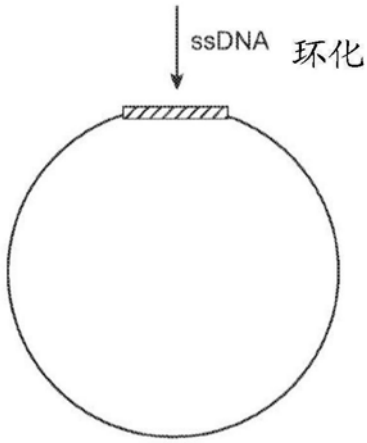


图22A

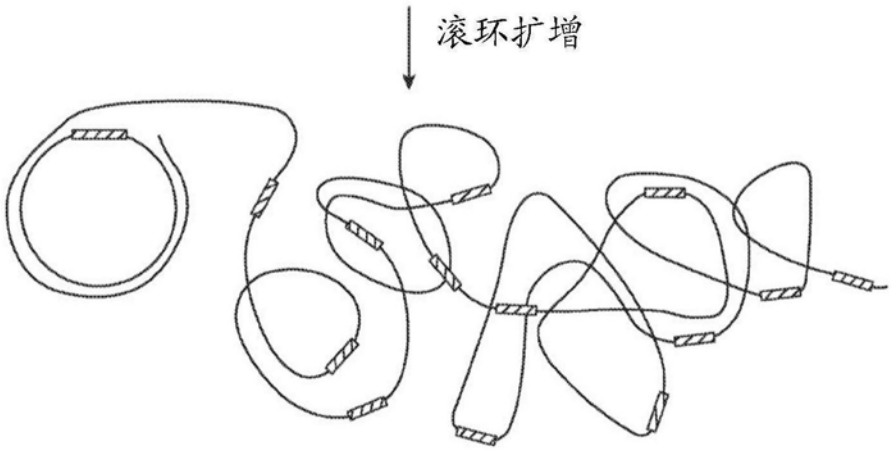


图22B

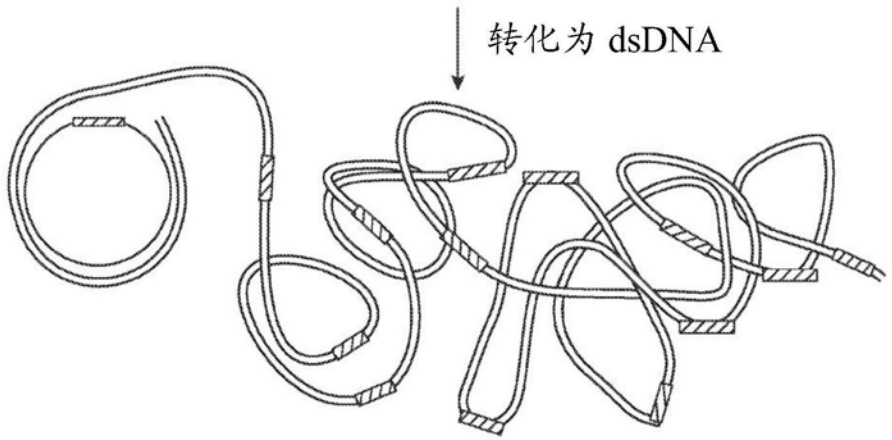


图22C

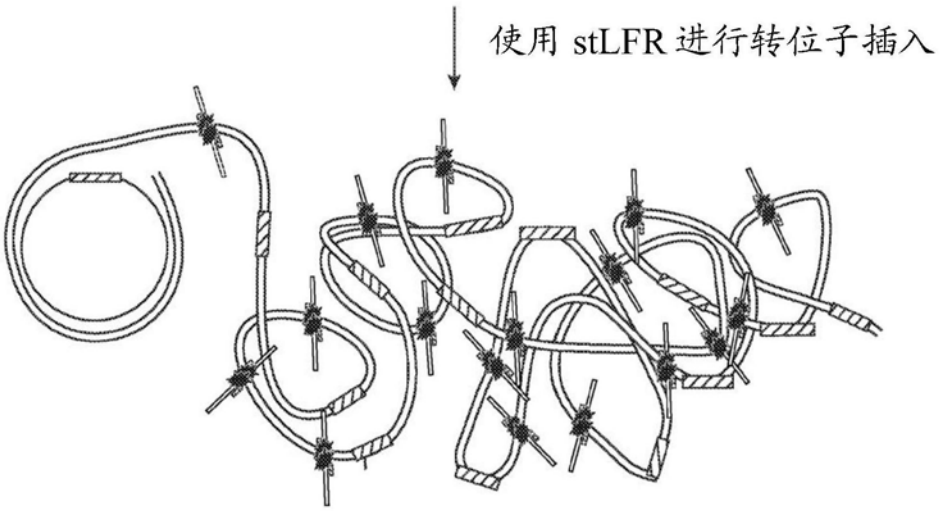


图22D

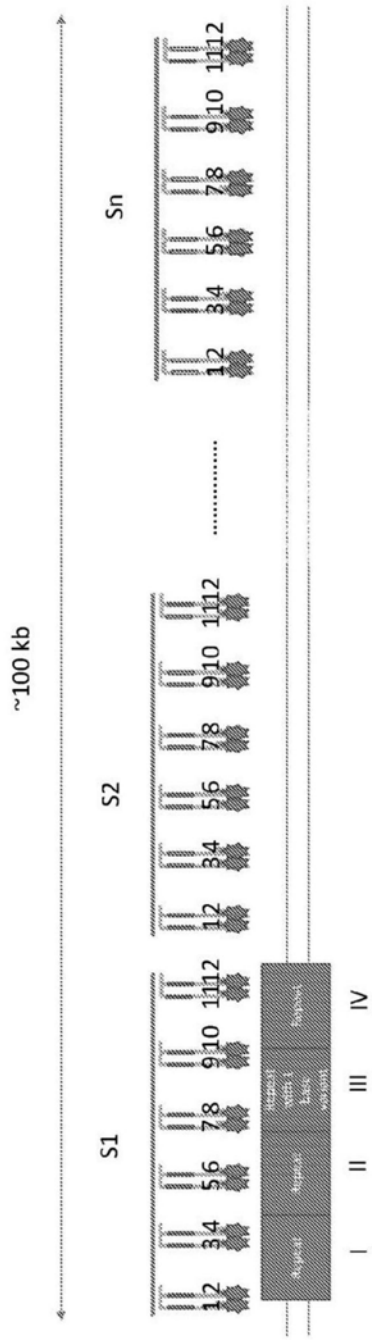


图23

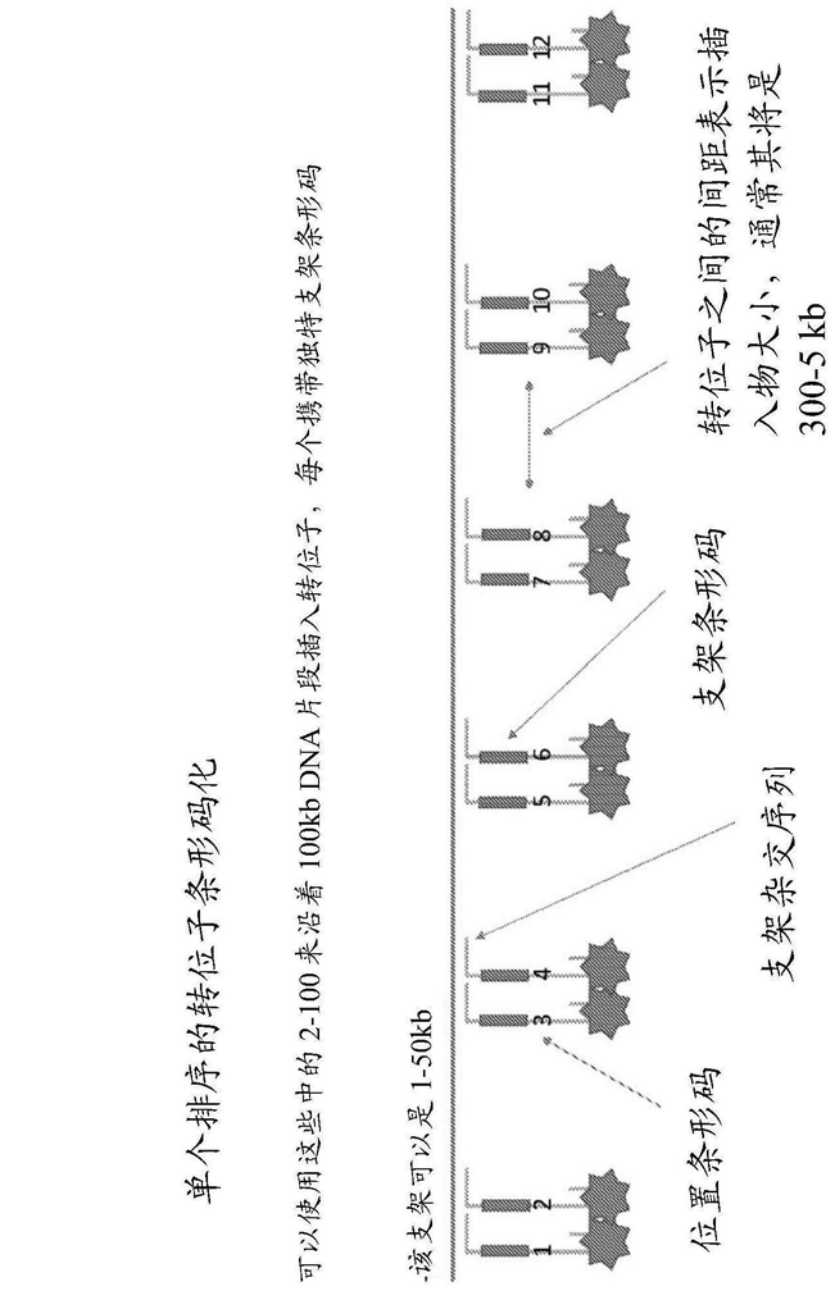

单个排序的转位子条形码化
可以使用这些中的 2-100 来沿着 100kb DNA 片段插入转位子，每个携带独特支架条形码
该支架可以是 1-50kb
1
2
3
4
5
6
7
8
9
10
11
12
位置条形码
支架条形码
支架杂交序列
转位子之间的间距表示插入物大小，通常其将是 300-5 kb

图24

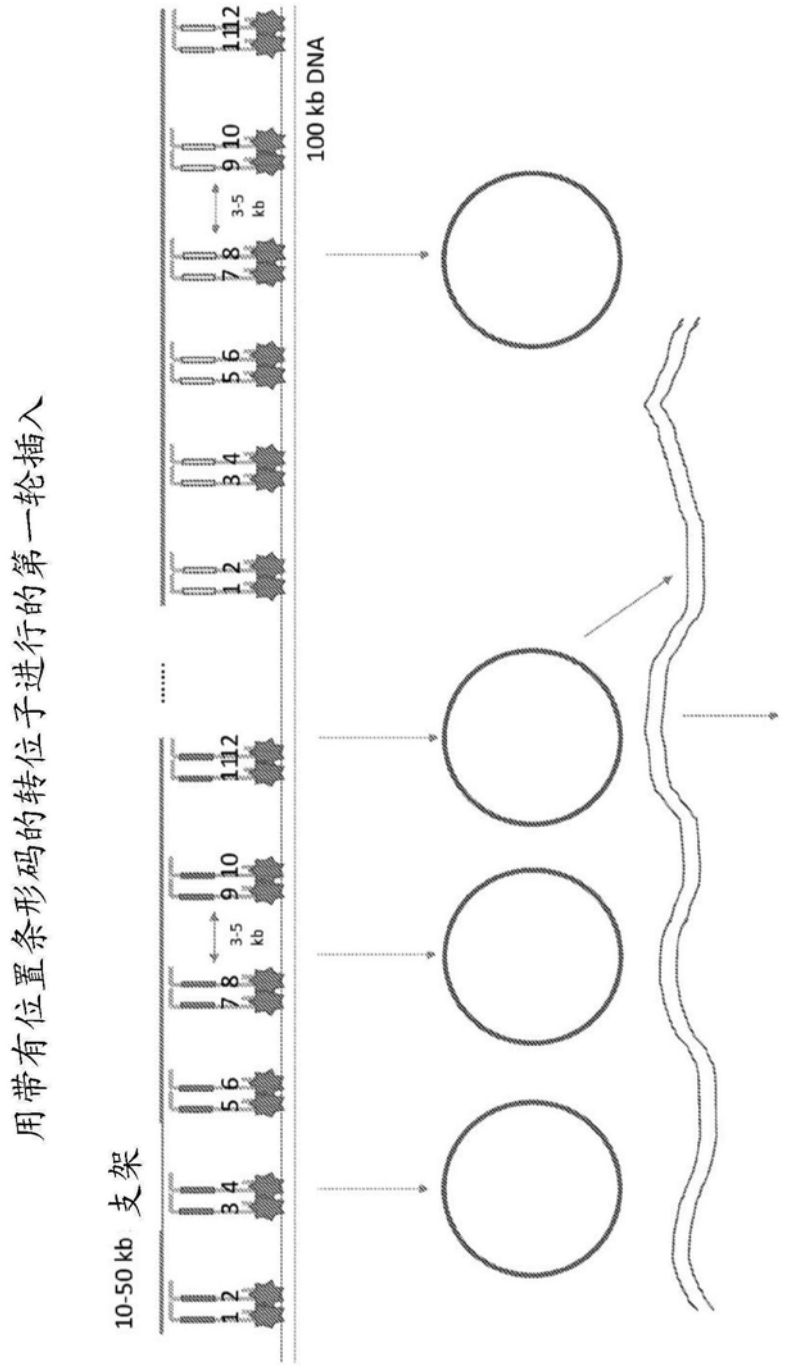


图25A

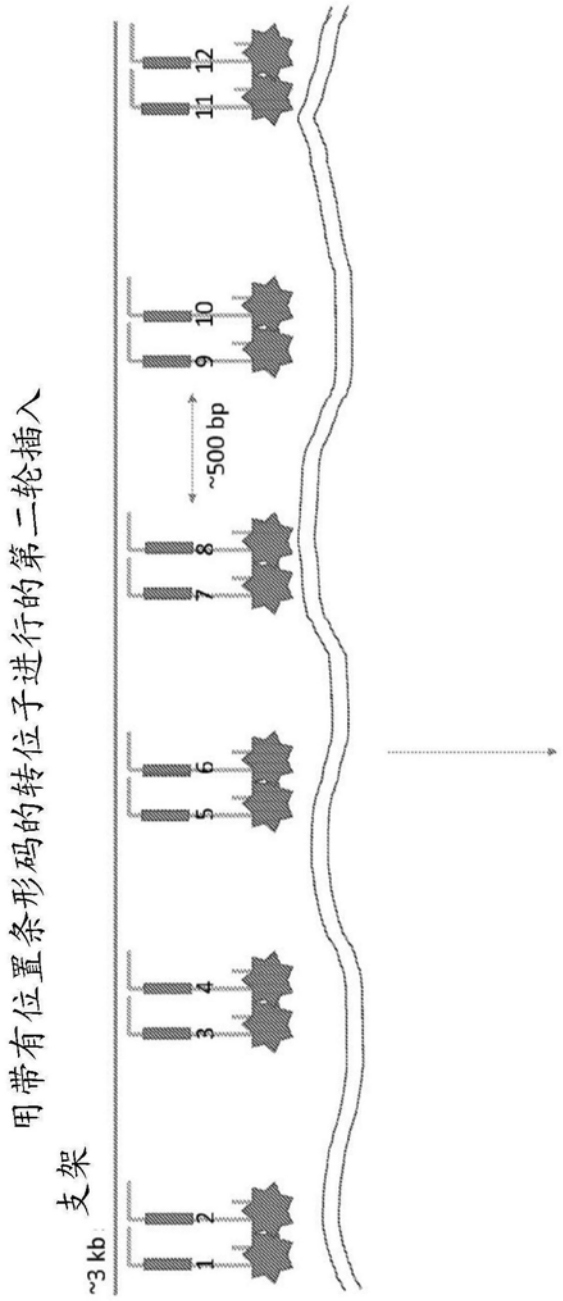


图25B

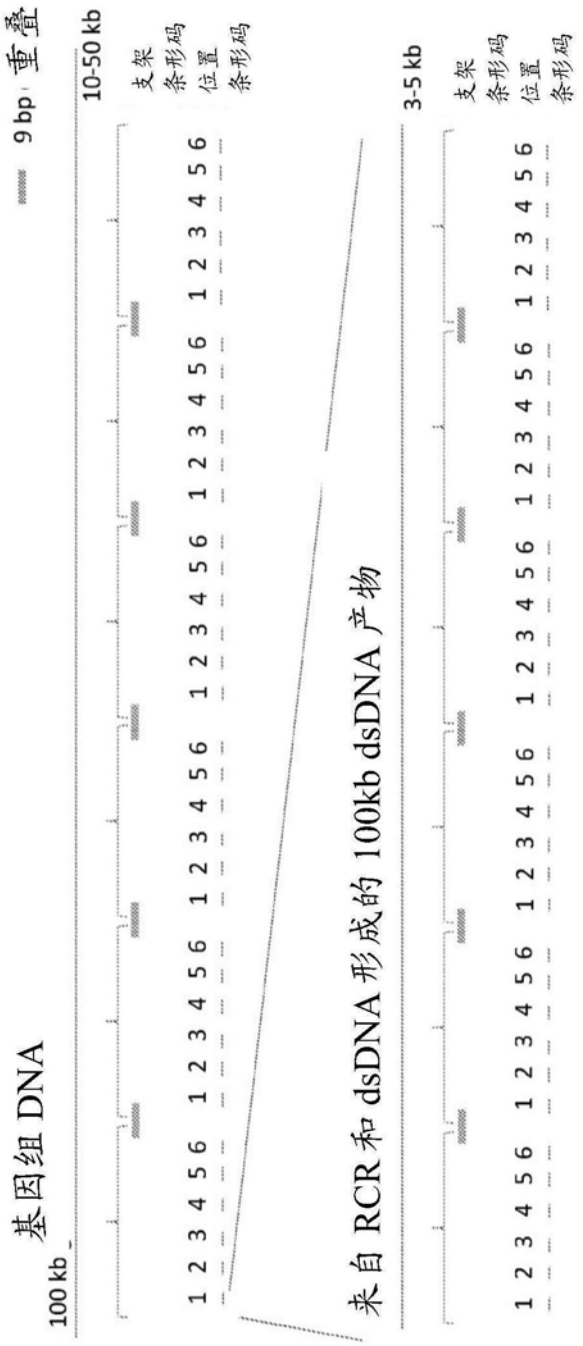

基因组 DNA
100 kb
9 bp 重叠
10-50 kb
3-5 kb
支架
条形码
位置
条形码
1 2 3 4 5 6
来自 RCR 和 dsDNA 形成的 100kb dsDNA 产物

图26

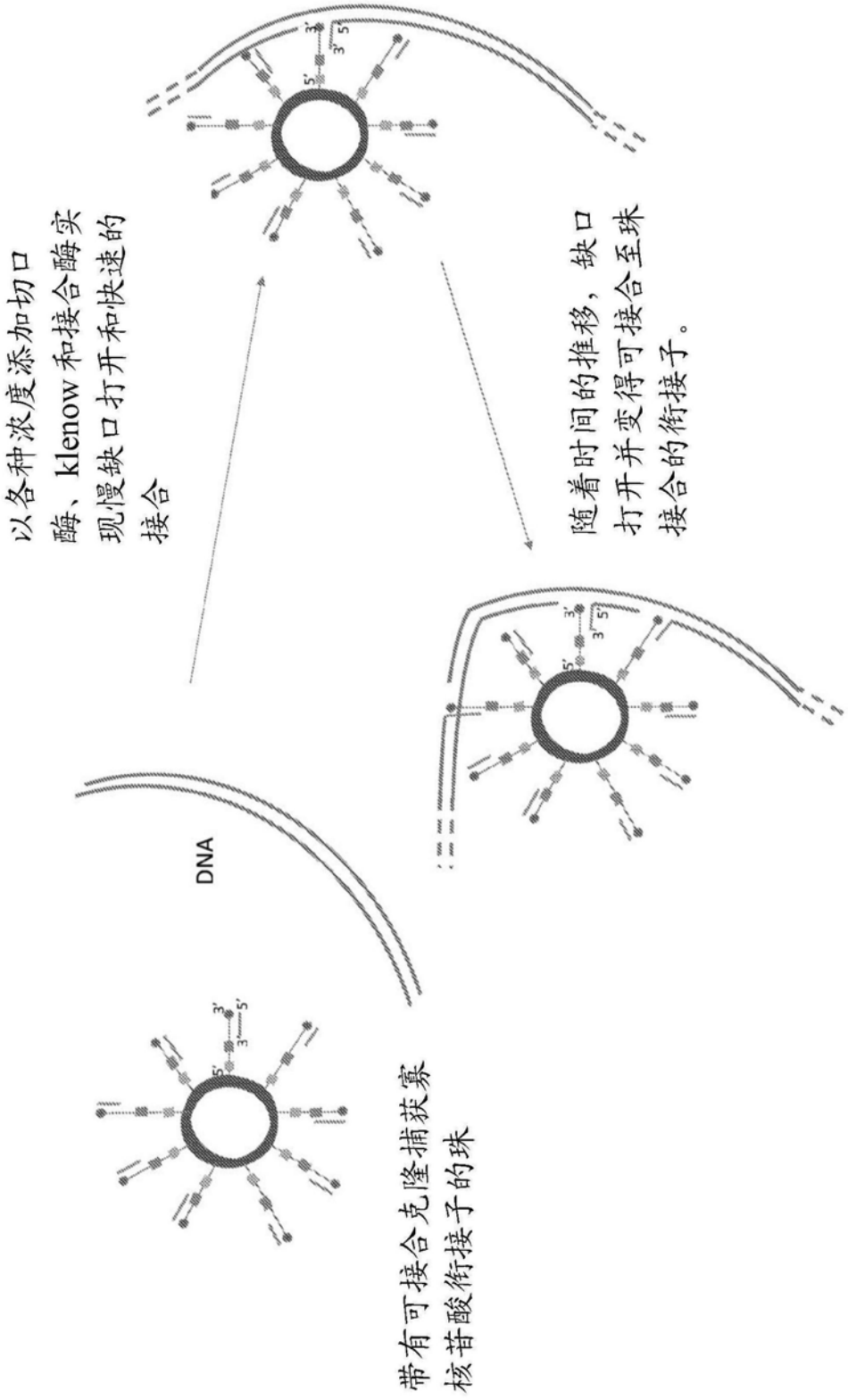

带有可接合克隆捕获寡核苷酸衔接子的珠
DNA
以各种浓度添加切口酶、klenow和接合酶实现慢缺口打开和快速的接合
随着时间的推移，缺口打开并变得可接合至珠接合的衔接子。

图27A

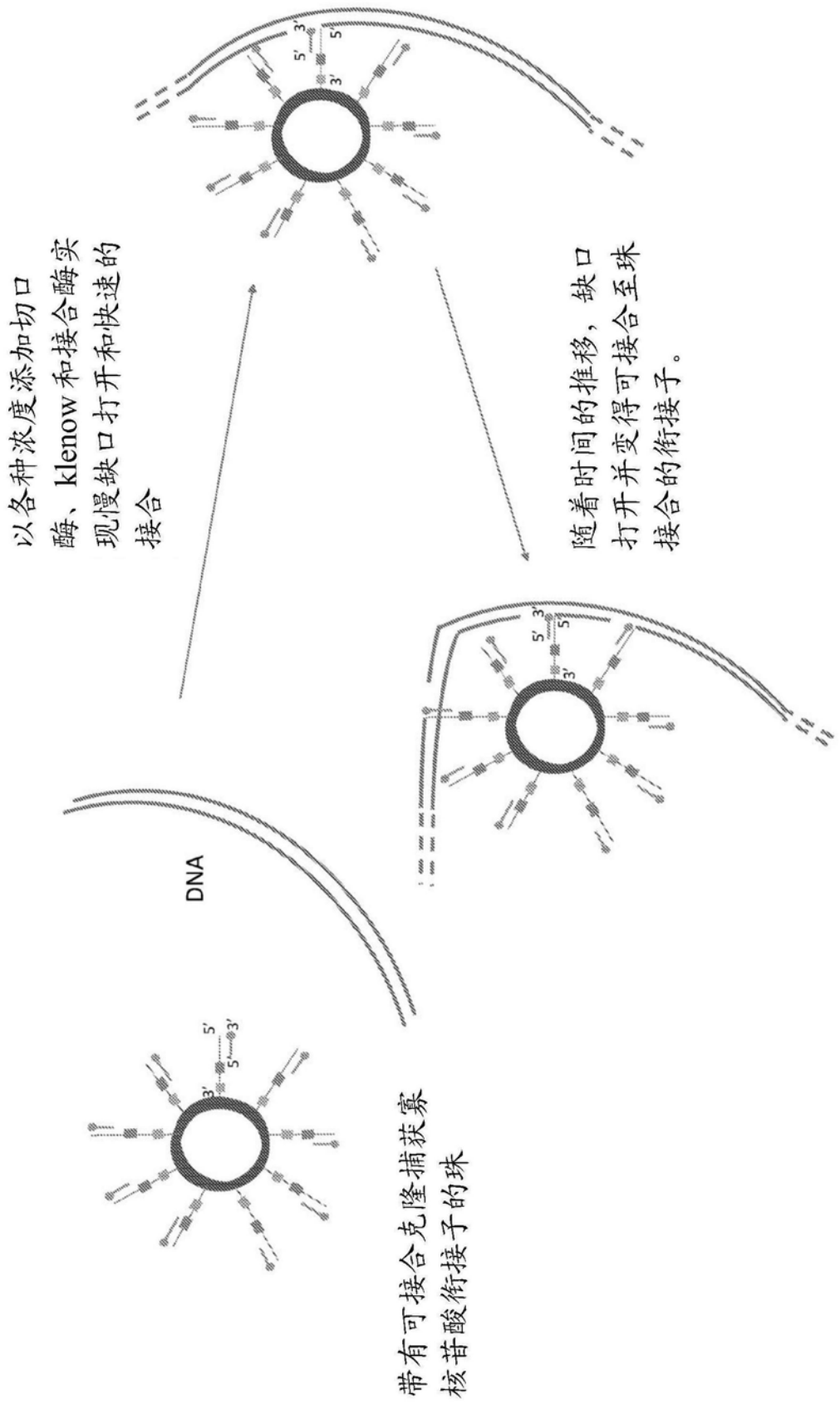

带有可接合克隆捕获寡
核苷酸衔接子的珠
DNA
以各种浓度添加切口
酶、klenow和接合酶实
现慢缺口打开和快速的
接合
随着时间的推移，缺口
打开并变得可接合至珠
接合的衔接子。

图27B

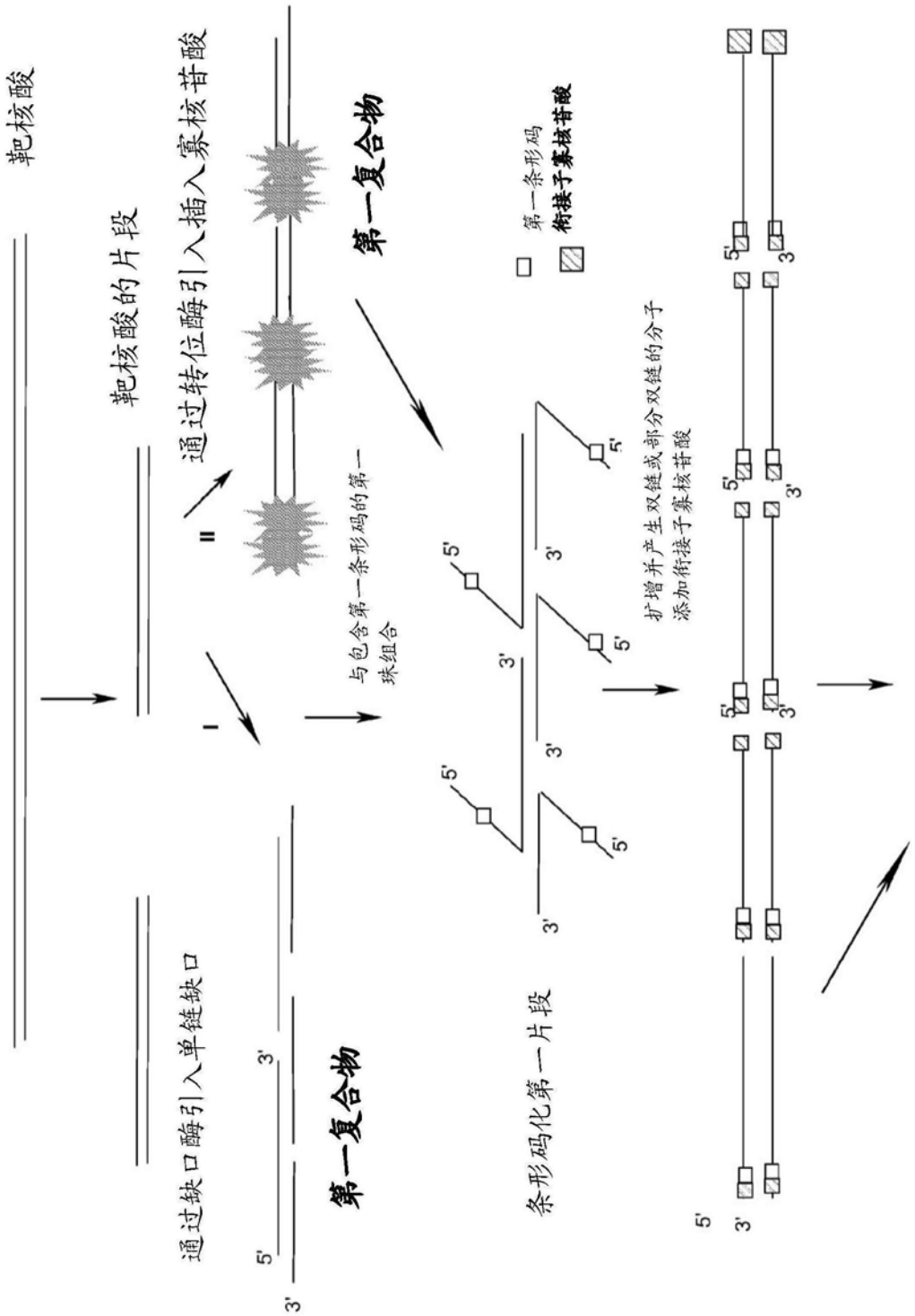

靶核酸
靶核酸的片段
通过缺口酶引入单链缺口
通过转位酶引入插入寡核苷酸
第一复合物
与包含第一条形码的第一珠组合
第一复合物
条形码化第一片段
扩增并产生双链或部分双链的分子
添加衔接子寡核苷酸
第一条形码
衔接子寡核苷酸

图28A

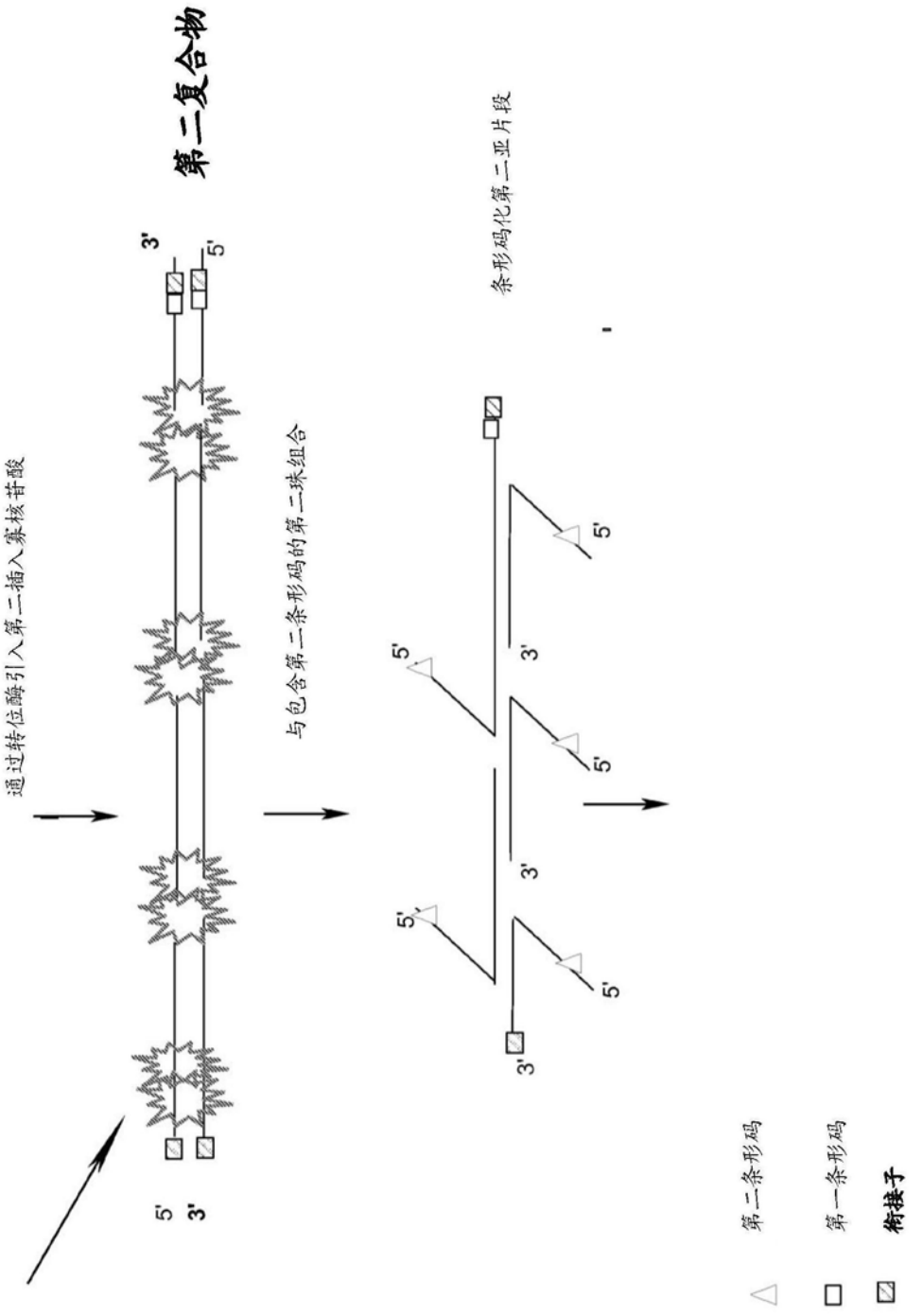

第二复合物
3'
5'
5'
3'
通过转位酶引入第二插入寡核苷酸
与包含第二条形码的第二珠组合
条形码化第二亚片段
5'
3'
5'
5'
3'
5'
5'
3'
5'
第二条形码
第一条形码
衔接子

图28B