(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119585442 A
(43)申请公布日 2025.03.07

(21)申请号 202380049211.1

(22)申请日 2023.06.21

(30)优先权数据
2209189.6 2022.06.22 GB

(85)PCT国际申请进入国家阶段日
2024.12.23

(86)PCT国际申请的申请数据
PCT/EP2023/066881 2023.06.21

(87)PCT国际申请的公布数据
WO2023/247658 EN 2023.12.28

(71)申请人 断弦生物科学有限公司
地址 英国剑桥郡

(72)发明人 F·多布斯 S·利德
P·范艾伊克

(74)专利代理机构 中国专利代理(香港)有限公司 72001
专利代理师 肖传馨 蔡晓菡

(51)Int.Cl.
C12Q 1/6855(2006.01)

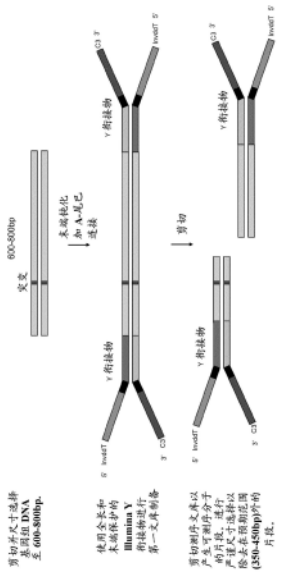
权利要求书3页 说明书35页
序列表（电子公布） 附图10页

(54)发明名称

用于核酸测序的方法和组合物

(57)摘要

在一个方面,本发明涉及文库制备方法和适合用在文库制备方法中的组合物。本发明还涉及测序方法和文库在测序中的用途。



CN 119585442 A

1.用于核酸测序的文库制备方法,所述方法包含:

a)提供多个核酸;

b)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物;

c)将所述多个核酸片段化;和

d)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件;

其中步骤b)和d)分开进行。

2.权利要求1所述的方法,其中在所述第一个衔接物连接步骤之后且在所述第二个衔接物连接步骤之前或作为所述第二个衔接物连接步骤的一部分,将所述多个核酸片段化。

3.权利要求1所述的方法,其中:

所述步骤依次并按照a)、b)、c)、d)的顺序进行;或

所述步骤依次并按照a)、d)、c)、b)的顺序进行;或

所述步骤按照步骤a)、步骤b)以及组合的步骤c)和d)的顺序进行;或

所述步骤按照步骤a)、步骤d)以及组合的步骤c)和b)的顺序进行。

4.权利要求1至3中的任一项所述的方法,其中所述非发夹衔接物包含与固定至基质的第一引物至少部分地互补的序列。

5.权利要求4所述的方法,其中所述与固定至基质的第一引物至少部分地互补的序列包含SEQ ID NO:1的至少5、10、15、16、17、18、19、20或所有21个碱基或SEQ ID NO:3的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基。

6.权利要求1至5中的任一项所述的方法,其中所述非发夹衔接物是Y-衔接物。

7.权利要求6所述的方法,其中所述Y-衔接物包含:

第一链,其包含与固定至基质的第一引物至少部分地互补的序列;和

第二链,其包含与第二引物的至少一个区域相同的序列。

8.权利要求7所述的方法,其中所述与第二引物的至少一个区域相同的序列包含SEQ ID NO:2的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基或SEQ ID NO:4的至少5、10、15、16、17、18、19或所有20个碱基。

9.任何前述权利要求所述的方法,其中所述非发夹衔接物是Y-衔接物,其包含:

第一链,其在5'至3'方向包含第一测序引物可以与其结合的第一杂交位点,以及与第一固定化引物至少部分地互补的序列;和

第二链,其在5'至3'方向包含与第二固定化引物的一个区域相同的序列,以及第二测序引物可以与其结合的第二杂交位点。

10.任何前述权利要求所述的方法,其中所述非发夹衔接物包含5'和/或3'保护特征。

11.权利要求10所述的方法,其中所述非发夹衔接物包含:第一链,其包含3'保护特征;和第二链,其包含5'保护特征。

12.任何前述权利要求所述的方法,其中所述非发夹衔接物是Y-衔接物,其包含:

第一链,其在5'至3'方向包含第一测序引物可以与其结合的第一杂交位点、与第一固定化引物至少部分地互补的序列,以及3'保护特征;和

第二链,其在5'至3'方向包含5'保护特征、与第二引物的至少一个区域相同的序列,以及第二测序引物可以与其结合的第二杂交位点。

13.任何前述权利要求所述的方法,其中所述多个核酸是DNA。

14.权利要求13所述的方法,其中所述多个核酸是基因组DNA(gDNA)。

15.任何前述权利要求所述的方法,其中所述方法包含:

e)在适合于所述第一固定化引物与互补核酸杂交的条件下,使所述多个核酸与包含第一固定化引物的基质接触;

其中所述非发夹衔接物包含与所述第一固定化引物至少部分地互补的序列。

16.权利要求15所述的方法,其中,在步骤e)之前不进行核酸扩增步骤。

17.权利要求15或权利要求16所述的方法,其中所述非发夹衔接物包含与第二引物的至少一个区域相同的序列并且所述第二引物固定化至所述基质。

18.权利要求17所述的方法,其中所述第一和第二固定化引物能够充当桥式扩增的正向引物和反向引物,且其中所述方法包含桥式扩增。

19.权利要求15至18中的任一项所述的方法,其中所述基质是流动池或珠子。

20.权利要求15至19中的任一项所述的方法,其中所述方法包含:

f)得到在步骤e)中与所述基质杂交的任何核酸的序列信息。

21.权利要求1至14中的任一项所述的方法,其中,在已经进行步骤a)、b)、c)和d)之后,所述方法进一步包含从所制备的文库得到序列信息。

22.通过权利要求1至14中的任一项所述的方法得到或可得到的核酸文库。

23.测序方法,其中所述方法包含得到在权利要求22所述的文库内的核酸的序列信息。

24.得到测序信息的方法,其中所述方法包含:

1)在适合于所述第一固定化引物与互补核酸杂交的条件下,使权利要求22所述的文库与包含第一固定化引物的基质接触;和

2)得到在步骤1)中与所述基质杂交的任何核酸的序列信息。

25.权利要求22所述的核酸文库或通过权利要求1至14中的任一项所述的方法得到或可得到的核酸文库在核酸测序方法中的用途。

26.用于核酸测序的文库制备方法,所述方法包含:

i)提供多个核酸;

ii)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物;和

iii)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件;

其中在所述文库制备过程中不扩增所述核酸。

27.权利要求26所述的方法,其中所述方法包含:

iv)在适合于所述第一固定化引物与互补核酸杂交的条件下,使所述多个核酸与包含第一固定化引物的基质接触;

其中所述非发夹衔接物包含与固定化至所述基质的第一引物至少部分地互补的序列;和

其中在步骤iv)之前不扩增所述核酸。

28.权利要求27所述的方法,其中所述非发夹衔接物包含与第二引物的至少一个区域相同的序列并且所述第二引物固定化至所述基质。

29.权利要求28所述的方法,其中所述第一和第二固定化引物能够充当桥式扩增的正

向引物和反向引物,且其中所述方法包含桥式扩增。

30.权利要求27至29中的任一项所述的方法,其中所述方法包含得到在步骤iv)中与所述基质杂交的任何核酸的序列信息。

31.权利要求26至30中的任一项所述的方法,其中步骤ii)和iii)分开进行,且其中在步骤ii)之后且在步骤iii)之前或在步骤iii)之后且在步骤ii)之前进行片段化步骤。

32.通过权利要求26所述的方法得到或可得到的核酸文库。

33.测序方法,其中所述方法包含得到在权利要求32所述的文库内的核酸的序列信息。

34.得到测序信息的方法,其中所述方法包含:

1)在适合于所述第一固定化引物与互补核酸杂交的条件下,使权利要求32所述的文库与包含第一固定化引物的基质接触;和

2)得到在步骤1)中与所述基质杂交的任何核酸的序列信息。

35.权利要求32所述的核酸文库或通过权利要求26所述的方法得到或可得到的核酸文库在核酸测序方法中的用途。

用于核酸测序的方法和组合物

技术领域

[0001] 在一个方面,本发明涉及文库制备方法和适合用于文库制备方法的组合物。本发明还涉及测序方法和文库在测序中的用途。

背景技术

[0002] 当前的核酸测序方法,诸如下一代测序(NGS),能够从极大样品中获取序列信息。例如,可能收集全基因组序列信息。但是,样品制备方法和测序方法都容易出错。对于需要在大样品中检测小变化的应用来说,这尤其成问题,例如检测基因组中的单个碱基对变化,因为即使非常低的错误率也可能影响结果。

[0003] 考虑到基因编辑技术诸如CRIPSR-Cas9等的越来越多应用,对检测罕见突变的需求日益增长。CRISPR基因组编辑使用合成的引导RNA将Cas9酶(充当基因剪刀的核酸酶)靶向到基因组中需要基因改变的特定位点。基因组编辑依赖于对这些位点的准确靶向以产生小的插入或删除,从而显现基因变化。但是,这涉及在DNA分子中引入DNA双链断裂(DSB)。这样的断裂可能对人体健康产生有害影响,并且可能造成癌症,这取决于其位置和细胞的修复它们的能力。该系统在其靶向方面是非常准确的,但是在编辑过程中,基因组中的次要的、所谓的脱靶位点也可能被无意地靶向。这些位置经常与靶序列相似,但相似性目前尚不完全清楚。实际上,仅基于引导RNA序列的生物信息学脱靶预测经常不够准确,以致于不能揭示所有通过实验检测到的脱靶位点。这对于改进引导设计和防止脱靶编辑是必要的。重要的是,注意到引导RNA的特异性具有高度可变性,这对于其在基因疗法中的安全使用具有重要意义。实际上,脱靶位点可以在整个基因组中发生断裂和/或突变,这对整个基因组编辑构成了重大且固有的风险。这又对当前传统的基于细胞的遗传毒性风险评估方法提出了严峻挑战,所述方法评估化学物质对基因组稳定性的影响。但是,基因组编辑使用了一类新型的靶向生物制剂,它提出了一种大海捞针类型的问题:当通过单独序列不可预测复杂基因组中的罕见脱靶编辑事件时,如何识别它们。CRISPR-Cas9基因组编辑已经加剧了脱靶问题,因为引入的脱靶现在非常罕见,以至于不可通过当前的基于细胞的方法检测它们。

[0004] 为了评估这些脱靶断裂的长期影响,重要的是测量由其准确修复决定的其突变结果。现有方法,诸如扩增子测序,已经报告了约$1/10^3$的突变检测限。因此,扩增子测序的灵敏度不足以检测在新的、低频率脱靶位点处的罕见突变。

[0005] Schmitt等人(PNAS,2012年9月4日,第109卷,第36期,第14508-14513页)公开了一种通过下一代测序来检测超罕见突变的方法。Kennedy,S.R.,等人(Detecting ultralow-frequency mutations by Duplex Sequencing.Nat Protoc,2014.9(11):第2586-606页)进行了进一步详细描述。Schmitt等人公开了对DNA双链体的两条链中的每一条独立地标记和测序。所公开的方法需要通过利用DNA聚合酶复制衔接物的一条链中的简并序列,将双链的、随机化的双链体标签序列附加到测序衔接物上。一种类似的方法包括Abascal等人(Somatic mutation landscapes at single-molecule resolution.Nature,593,405-410 2021)公开的NanoSeq,它描述了BotSeqS方法的优化版本,所述BotSeqS方法应用酶促片段

化和改良的末端修复程序来改进使用如上所述的UMI标签进行的错误校正测序。但是,对于较大的基因组来说,使用这些方法检测全基因组突变目前成本过高,因为这种方法需要高覆盖率(>10^4倍)(Kennedy,S.R.,等人.2014),从而限制它只能用于已经鉴定的脱靶的靶向测序,诸如通过DSB-检测方法。此外,在NanoSeq中使用的DNA片段化限制酶的选择仅提供了基因组的部分覆盖。已经提出使用超声处理和外切核酸酶钝化作为替代方案,但是这种方案会受到先前描述的相同过度覆盖要求的影响。

[0006] WO 2013/142389 A1公开了旨在使用双链体一致测序来降低大规模平行DNA测序的错误率的方法。WO 2013/142389 A1公开了通过将衔接物连接至DNA来形成文库,从而产生三种产物(被称作"产物I"、"产物II"和"产物III")。

[0007] 需要能够产生错误校正序列信息的其它方法。具体地,需要能够公正地且独立地确定在低频率脱靶位点处和整个基因组中接近背景水平的基因编辑诱导的突变的方法。

发明内容

[0008] 在一个方面,提供了用于核酸测序的文库制备方法,所述方法包含:

[0009] a)提供多个核酸;

[0010] b)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物;

[0011] c)将所述多个核酸片段化;和

[0012] d)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件;

[0013] 其中步骤b)和d)分开进行。

[0014] 在所述第一个衔接物连接步骤之后且在所述第二个衔接物连接步骤之前或作为所述第二个衔接物连接步骤的一部分,将所述多个核酸片段化。所述第一个连接步骤是待进行的步骤b)或步骤d)中的第一步。所述第二个连接步骤是待进行的步骤d)或步骤b)中的第二步。因此,所述第一个连接步骤是:i)步骤b),此时步骤d)是第二个连接步骤,或ii)步骤d),此时步骤b)是第二个连接步骤。

[0015] 所述步骤可以依次和按照a)、b)、c)、d)的顺序进行。所述步骤可以依次和按照a)、d)、c)、b)的顺序进行。所述步骤可以按照步骤a)、步骤b),以及组合的步骤c)和d)的顺序进行。所述步骤可以按照步骤a)、步骤d),以及组合的步骤c)和b)的顺序进行。

[0016] 所述非发夹衔接物可以包含与固定至基质的第一引物至少部分地互补的序列。所述与固定至基质的第一引物至少部分地互补的序列可以包含SEQ ID NO:1的至少5、10、15、16、17、18、19、20或所有21个碱基或SEQ ID NO:3的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基。

[0017] 所述非发夹衔接物可以是Y-衔接物。所述Y-衔接物可以包含:第一链,其包含与固定至基质的第一引物至少部分地互补的序列;和第二链,其包含与第二引物的至少一个区域相同的序列。所述与第二引物的至少一个区域相同的序列可以包含SEQ ID NO:2的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基或SEQ ID NO:4的至少5、10、15、16、17、18、19或所有20个碱基。所述非发夹衔接物是Y-衔接物,其包含:第一链,其在5'至3'方向包含第一测序引物可以与其结合的第一杂交位点,以及与第一固定化引物至少部分地互补的序列;和第二链,其在5'至3'方向包含与第二固定化引物的一个区域相同的序列,以及

第二测序引物可以与其结合的第二杂交位点。

[0018] 所述非发夹衔接物可以包含5’和/或3’保护特征。所述非发夹衔接物可以包含:第一链,其包含3’保护特征;和第二链,其包含5’保护特征。

[0019] 所述非发夹衔接物可以是Y-衔接物,其包含:第一链,其在5’至3’方向包含第一测序引物可以与其结合的第一杂交位点、与第一固定化引物至少部分地互补的序列,以及3’保护特征;和第二链,其在5’至3’方向包含5’保护特征、与第二引物的至少一个区域相同的序列,以及第二测序引物可以与其结合的第二杂交位点。

[0020] 所述多个核酸可以是DNA或基因组DNA(gDNA)。

[0021] 所述方法可以进一步包含:e)在适合于所述第一固定化引物与互补核酸杂交的条件下,使所述多个核酸与包含第一固定化引物的基质接触;其中所述非发夹衔接物包含与所述第一固定化引物至少部分地互补的序列。任选地,在步骤e)之前不进行核酸扩增步骤。所述基质可以是流动池或珠子。

[0022] 所述非发夹衔接物可以包含与第二引物的至少一个区域相同的序列并且所述第二引物固定化至所述基质。所述第一和第二固定化引物可以能够充当桥式扩增的正向引物和反向引物,且其中所述方法可以包含桥式扩增。

[0023] 所述方法可以进一步包含:f)得到在步骤e)中与所述基质杂交的任何核酸的序列信息。

[0024] 在已经进行步骤a)、b)、c)和d)之后,所述方法可以进一步包含从所制备的文库得到序列信息。

[0025] 在另一个方面,提供了通过本公开内容的方法得到或可得到的核酸文库。

[0026] 在另一个方面,提供了一种核酸文库,其包含具有连接到一端的非发夹衔接物和在另一端处的发夹的靶核酸,其中所述核酸文库包含小于99.9%、99%、95%、90%、80%、70%、60%、50%、40%、30%、20%、10%、5%、3%、1%、0.1%或0.01%(按质量计)或没有具有连接到两端的非发夹衔接物的靶核酸。

[0027] 在另一个方面,提供了一种测序方法,其中所述方法包含得到在本公开内容的文库内的核酸的序列信息。

[0028] 在另一个方面,提供了一种得到测序信息的方法,其中所述方法包含:

[0029] 1)在适合于所述第一固定化引物与互补核酸杂交的条件下,使本公开内容的文库与包含第一固定化引物的基质接触;和

[0030] 2)得到在步骤1)中与所述基质杂交的任何核酸的序列信息。

[0031] 在另一个方面,提供了本公开内容的核酸文库或通过本公开内容的方法得到或可得到的核酸文库在核酸测序方法中的用途。

[0032] 在另一个方面,提供了一种用于核酸测序的文库制备方法,所述方法包含:

[0033] i)提供多个核酸;

[0034] ii)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物;和

[0035] iii)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件;

[0036] 其中在所述文库制备过程中不扩增所述核酸。任选地,其中所述方法进一步包含:

[0037] iv)在适合于所述第一固定化引物与互补核酸杂交的条件下,使所述多个核酸与

包含第一固定化引物的基质接触;

[0038] 其中所述非发夹衔接物包含与固定化至所述基质的第一引物至少部分地互补的序列;和

[0039] 其中在步骤iv)之前不扩增所述核酸。

[0040] 所述非发夹衔接物可以包含与第二引物的至少一个区域相同的序列并且所述第二引物固定化至所述基质。所述第一和第二固定化引物可以能够充当桥式扩增的正向引物和反向引物,且其中所述方法可以包含桥式扩增。

[0041] 所述方法可以包含得到在步骤iv)中与所述基质杂交的任何核酸的序列信息。步骤ii)和iii)可以分开进行,且其中在步骤ii)之后且在步骤iii)之前或在步骤iii)之后且在步骤ii)之前可以进行片段化步骤。

[0042] 在另一个方面,提供了通过上述方法得到或可得到的核酸文库。在另一个方面,提供了一种测序方法,其中所述方法包含得到在所述文库内的核酸的序列信息。在另一个方面,提供了一种得到测序信息的方法,其中所述方法包含:1)在适合于所述第一固定化引物与互补核酸杂交的条件下,使所述文库与包含第一固定化引物的基质接触;和2)得到在步骤1)中与所述基质杂交的任何核酸的序列信息。还提供了所述核酸文库或通过所述方法得到或可得到的核酸文库在核酸测序方法中的用途。

[0043] 附图简述

[0044] 图1.“DEDUCE-seq”(通过无偏流动池富集进行的DuplEx测定和测序(DuplEx Determination by Unbiased flow Cell Enrichment and sequencing))的一个实施方案的示例性图示。

[0045] 图2.测序的示例性图示。由于DNA双链的连接,突变被读取两次。

[0046] 图3.与基质(在该情况下,流动池)结合的示例性图示。来自Y-衔接物连接的片段的DNA与流动池结合,并且仅发夹连接的片段被洗掉。

[0047] 图4.本公开内容的方法的一个实施方案的示例性概述。

[0048] 图5.试点-1的DEDUCE-seq文库制备;DNA尺寸分布和定量。1)左图-通过除去>300bp的DNA(灰色迹线),将基因组DNA尺寸选择为约100-500bp(黑色迹线)。2)中图-通过凝胶电泳测试连接的DNA(黑色迹线)和重新超声处理的DNA(灰色迹线)。3)右图-这里显示了DEDUCE-seq文库的超声处理后DNA(黑色迹线)和最终文库DNA。

[0049] 图6.来自试点-1的9.5M正向读取(左图)和反向读取(右图)的整个长度上的HP序列分布。

[0050] 图7.DEDUCE-seq文库不一致读取对。这里显示了来自DEDUCE-seq文库的不一致平行读取对的3个实例。最左边的读取对(F1F2,读取发生堆叠)由左对齐(读取F1,SAM标志67)和右对齐(读取F2,SAM标志131)组成,二者均映射到正向链(forward stand)。右读取对(R2R1)由左对齐(读取R2,SAM标志115)和右2(读取R1,SAM标志179)组成。未显示右侧第二个反向对齐的读取细节。

[0051] 图8.试点-2的DEDUCE-seq文库制备;DNA尺寸分布和定量。1)左图-通过除去>300bp的DNA(灰色迹线),将基因组DNA尺寸选择为约100-500bp(黑色迹线)。2)中图-通过凝胶电泳测试Y-衔接物连接的DNA(黑色迹线)和重新超声处理的DNA(灰色迹线)。3)右图-这里显示了DEDUCE-seq样品的最终文库DNA。

[0052] 图9.在试点-2中测序的DEDUCE-seq文库分子的读取R1和R2中的发夹序列分布。

具体实施方式

[0053] 本发明人在本文中提供了用于制备核酸文库的技术,所述核酸文库适合于产生错误校正测序数据。当对本公开内容的核酸文库测序时,提供核酸双链体的两条链的测序信息。这允许校正错误,包括通过文库制备引入的错误或作为测序结果出现的错误。

[0054] 在第一方面,提供了一种用于核酸测序的文库制备方法,所述方法包含:

[0055] a)提供多个核酸;

[0056] b)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物;

[0057] c)将所述多个核酸片段化;和

[0058] d)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件;

[0059] 其中步骤b)和d)分开进行。

[0060] 所述第一方面的方法的步骤b)和d)分开进行,并且因此所述非发夹衔接物和所述发夹衔接物不会作为同一反应的部分连接至所述核酸。换而言之,步骤b)和d)不同时进行。

[0061] 但是,如本文进一步讨论的,衔接物连接和片段化步骤可以同时、组合或并行进行。例如,标记化步骤可以用于连接衔接物并将多个核酸片段化。因此,步骤b)和c)、或步骤d)和c)可以同时、组合或并行进行。

[0062] 对于所有实施方案,在所述第一个衔接物连接步骤之后且在所述第二个衔接物连接步骤之前或作为所述第二个衔接物连接步骤的一部分,将所述多个核酸片段化。

[0063] 所述第一方面的方法的步骤b)、c)和d)可以依次进行。该方法的步骤可以但不必按照a)、b)、c)和然后d)的顺序执行。具体而言,步骤b)和d)可以交换,使得顺序可以是a)、d)、c)和然后b)。片段化步骤(步骤c))可以在一种类型的衔接物的连接与另一种类型的衔接物的连接之间进行。所述片段化步骤在衔接物的第二次连接时或之前进行。

[0064] 本文中使用的术语“依次”是指,所述步骤不是同时的。但是,依次步骤不必是连续的,并且可以在明确列举的步骤之间执行另外步骤。

[0065] 在其它实施方案中,所述步骤可以按如下顺序:步骤b)和然后步骤c)和d)并行、组合或同时。所述步骤可以按如下顺序:步骤d)和然后步骤c)和b)并行、组合或同时。

[0066] 由于步骤b)和d)分开进行的事实,所述非发夹衔接物和所述发夹衔接物不会同时或作为同一步骤的部分连接至所述核酸。对于其中形成发夹的实施方案,该过程不与所述非发夹衔接物的连接同时发生或作为同一步骤的部分发生。

[0067] 因此,在一个实施方案中,提供了一种用于核酸测序的文库制备方法,所述方法包含:

[0068] a)提供多个核酸;

[0069] b)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物以产生第一文库;

[0070] c)将所述第一文库片段化;和

[0071] d)在有利于连接的条件下将所述片段化的第一文库暴露于发夹衔接物以产生第二文库,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件以产生第二文库。所述第一文库的片段化和所述第二文库的产生可以是依次或同时步骤。

[0072] 在一个替代实施方案中,提供了一种用于核酸测序的文库制备方法,所述方法包含:

[0073] a)提供多个核酸;

[0074] b)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物以产生第一文库,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件以产生第一文库;

[0075] c)将所述第一文库片段化;和

[0076] d)在有利于连接的条件下将所述片段化的第一文库暴露于非发夹衔接物以产生第二文库。所述第一文库的片段化和所述第二文库的产生可以是依次或同时步骤。

[0077] 核酸文库是已经向其连接至少一种类型的衔接物的核酸的集合或多个核酸。

[0078] 由第一方面的方法提供的文库具有减少量的可测序核酸,这些核酸将产生仅与双链体的一条链相关的不可纠错的序列信息。这样的不希望的核酸包括包含例如连接到核酸两端的非发夹衔接物的那些。这是有利的,因为它消除了在测序之前或基于基质的步骤之前进行富集和/或扩增的需要。此外,文库的质量得到提高。无需在测序之前(例如在使文库与流动池等基质接触之前)进行扩增步骤即可从文库中得到序列信息的能力是有利的,因为这样的事步骤本身可能引入突变和偏差。在一些实施方案中,由第一方面的方法提供的文库具有显著减少的可测序核酸的存在,这些核酸将产生仅与双链体的一条链相关的不可纠错的序列信息,例如,足以使文库在基质上测序而无需在应用于基质之前进行富集或扩增的减少。在一些实施方案中,合乎需要的是,产生第一方面的文库,其不包含可测序的核酸,所述核酸会产生仅与双链体的一条链相关的不可纠错的序列信息。但是,它们的存在的任何减少(例如使得不再需要扩增)都是有利的。例如,在WO 2013/142389 A1中公开了导致含有不希望的产物的文库的现有技术方法。

[0079] 多个核酸的提供可以作为所述方法的第一步来执行。该步骤可以包含从样品中纯化核酸,诸如DNA。从样品中纯化或分离的核酸可能是基因组DNA(gDNA)。因此,多个核酸的提供可以是要测序的DNA或gDNA分子的提供,其可以被称作靶核酸。所述样品可以是生物样品,诸如从患者得到的样品或从生物细胞得到的样品。所述样品可以是组织样品、生物流体样品、细胞系或任何其它合适的样品。所述样品可以包含正常细胞、肿瘤细胞、恶性细胞或癌细胞。所述样品可以包含来自正常细胞、肿瘤细胞、恶性细胞或癌细胞的核酸。所述样品可以是肿瘤样品或者包含肿瘤细胞或癌细胞的组织的样品。所述样品可以是血液或血液部分,诸如血浆部分。所述样品可以是血液或血液部分,诸如血浆,其包含循环肿瘤DNA或疑似包含循环肿瘤DNA。所述样品可以包含循环肿瘤DNA或疑似包含循环肿瘤DNA。所述样品可以是血液或血液部分,诸如血浆,其包含循环胎儿DNA或疑似包含循环胎儿DNA。所述样品可以包含循环胎儿DNA或疑似包含循环胎儿DNA。所述样品可能已经进行过基因修饰或基因编辑。例如,所述样品可能已经进行过能够诱导DSB的编辑技术,例如CRISPR-Cas9、TALEN或其它核酸酶。因此,可以产生所述文库以允许检测由编辑技术诱导的脱靶突变。

[0080] 如果由于分离的核酸的性质而有必要,可以作为步骤a)的部分对核酸进行剪切或片段化。将核酸片段化的方法是本领域已知的,并且可以是例如机械剪切或酶剪切。所述片段化可以包含超声处理,例如,使用Bioruptor超声破碎器或Covaris超声破碎器。所述片段化可以是基于酶的,并且可以利用基于酶的试剂,所述试剂以时间依赖性方式剪切DNA以产生期望尺寸的片段。合适的商购可得的试剂包括NEBNext dsDNA片段化酶(NEB)。所述片段

化可以与所述第一个衔接物连接步骤同时进行,例如通过标记化。所述片段化可以包含使用核酸酶,例如,一种内切核酸酶、多种内切核酸酶、一种限制性酶或多种限制性酶。所述片段化可以包含使用核酸引导的内切核酸酶,诸如RNA引导的DNA内切核酸酶。所述片段化可以包含使用Cas蛋白或衍生物或变体。所述片段化可以包含使用Cas9、Cpf1、C2c2、C2c1、CasM、CasMini、逆转录子、原核argonaute、TALEN或大范围核酸酶。

[0081] 对于所有实施方案来说,可能不需要将片段化作为步骤a)的部分。例如,一些核酸源不需要片段化。例如,从血浆中得到的样品可能不需要片段化。可替换地,核酸可以含有双链断裂(DSB),其可能是天然存在的或诱导的,并且这样的样品可能不需要在步骤a)中片段化。在一些实施例中,衔接物可以直接连接到DSB。

[0082] 所述片段化可以产生具有特定尺寸或具有特定尺寸分布的核酸片段,并且随后可以进行尺寸选择步骤。其中步骤a)不包含片段化的方法还可以包含尺寸选择步骤。许多用于尺寸选择和/或净化步骤的系统或试剂是本领域已知的。例如,使用珠子进行尺寸选择以除去或选择特定尺寸的片段。所述珠子可以是固相可逆固定化(SPRI)珠子。商购可得的珠子包括“SPRIselect”(Beckman Coulter)或SPRI珠子(GC Biotech,CNGS-0005)。毛细管DNA电泳可以用于尺寸选择。毛细管DNA电泳也可以用于评估成功连接以及多余衔接物的除去。其它替代方案包括基于凝胶的电泳尺寸选择步骤或系统,例如包含使用琼脂糖凝胶或聚丙烯酰胺凝胶。合适的系统是商购可得的,诸如BluePippin系统(Sage Science)。用于尺寸选择和/或净化的系统的其它例子包括基于DNA提取柱的系统。

[0083] 所述方法可以包含除去其尺寸小于约100bp或小于约150bp的片段,和/或保留其尺寸大于约150bp的片段。

[0084] 在一些实施方案中,所得到的片段是100至1500bp、200至1300bp、300至1100bp、400至1000bp、500至900bp或600至800bp。在一个特定实施方案中,所述核酸是大约600至800bp的尺寸的片段。

[0085] 因此,在一个实施方案中,步骤a)可以是如下:

[0086] a)提供多个核酸;其中所述提供包含:

[0087] i)从样品分离多个核酸;

[0088] ii)将所述多个核酸片段化;和

[0089] iii)基于尺寸选择所述多个核酸的片段。

[0090] 可以处理所述片段化的核酸以适合用于衔接物连接。例如,可以向核酸添加一个结合特征或多个结合特征。所述结合特征可以包含5’特征和/或3’特征。所述结合特征可以是任何适合于促进衔接物连接的结合特征。

[0091] 例如,5’或3’结合特征可以包含以下之一:磷酸酯基团;三磷酸酯“T尾”,诸如脱氧胸苷三磷酸“T尾”;三磷酸酯“A-尾”,诸如脱氧腺苷三磷酸“A-尾”;至少一个随机N核苷酸,诸如多个N核苷酸,或任何其它已知的结合基团,以允许衔接物与核酸的连接。

[0092] 在一个特定实施方案中,所述片段化的核酸是末端钝化的且带有A尾。因此,5’磷酸酯和/或3’A尾可以添加到片段化的核酸中。

[0093] 因此,在一个实施方案中,步骤a)可以是如下:

[0094] a)提供多个核酸;其中所述提供包含:

[0095] i)从样品分离多个核酸;

[0096] 任选地ii)将所述多个核酸片段化;

[0097] 任选地iii)基于尺寸选择所述多个核酸的片段;和

[0098] iv)将5'和/或3'结合特征添加至所述多个核酸。

[0099] 步骤i)、ii)、iii)和iv)可以按照i)、ii)、iii)和然后iv)的顺序进行。但是,可以遵循任何顺序,其允许制备适合于本文公开的下游步骤的多个核酸。例如,顺序可以是i)、ii)、iv)和然后iii)。

[0100] 在特定的非限制性实施方案中,步骤a)可以是如下:

[0101] a)提供多个核酸;其中所述提供包含:

[0102] i)从样品分离多个核酸,其中所述多个核酸是gDNA;

[0103] ii)将所述分离的多个核酸片段化;

[0104] iii)基于尺寸选择所述多个核酸的片段;

[0105] iii)将所述选择的核酸末端钝化;和

[0106] iv)将A-尾添加至所述末端钝化的核酸。

[0107] 在另一个非限制性的实施方案中,步骤a)可以是如下:

[0108] a)提供多个核酸;其中所述提供包含:

[0109] i)从样品分离多个核酸,其中所述多个核酸是gDNA;

[0110] ii)将所述分离的多个核酸片段化;

[0111] iii)基于尺寸选择所述多个核酸的片段;

[0112] iii)将所述选择的核酸末端钝化和5'磷酸化;和

[0113] iv)将A-尾添加至所述末端钝化的核酸。

[0114] 在其它实施方案中,步骤a)包含核酸的片段化和衔接物的连接。例如,标记化步骤。所述连接的衔接物可以是所述非发夹衔接物或所述发夹衔接物,取决于执行所述步骤的顺序。

[0115] 步骤a)和步骤b)可以如下组合:

[0116] i)从样品分离多个核酸;和

[0117] ii)将非发夹衔接物片段化并连接至所述多个核酸;和

[0118] 任选地iii)基于尺寸选择所述多个核酸的片段。

[0119] 可替换地,步骤a)和步骤d)可以如下组合:

[0120] i)从样品分离多个核酸;和

[0121] ii)将发夹衔接物片段化并连接至所述多个核酸;和

[0122] 任选地iii)基于尺寸选择所述多个核酸的片段。

[0123] 在一个实施方案中,步骤a)和步骤b)如下组合:

[0124] 1)提供多个核酸;和

[0125] 2)在有利于连接和片段化的条件下将所述多个核酸暴露于非发夹衔接物。

[0126] 在另一个实施方案中,步骤a)和步骤d)如下组合:

[0127] 1)提供多个核酸;和

[0128] 2)在有利于连接和片段化的条件下将所述多个核酸暴露于发夹衔接物。

[0129] 在一些实施方案中,原位连接至少一种类型的衔接物。在这种情况下,步骤a)可以包含细胞或组织样品的透化。例如,步骤a)可以包含将样品暴露于透化剂。在连接衔接物之

后,可以从样品中分离核酸,诸如DNA或gDNA。在这些实施方案中,所述衔接物可以连接至DSB。所述DSB可以是天然存在的或诱导的。

[0130] 步骤b)包含在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物。因此,在一些实施方案中,所述非发夹衔接物将连接到核酸的可用的或未受保护的末端。在其中在步骤d)之前进行步骤b)的实施方案中,这将导致非发夹衔接物连接至所述多个核酸的至少一部分的两端。在其中在步骤d)之后进行步骤b)的实施方案中,这将导致非发夹衔接物连接至不存在发夹的核酸的末端。

[0131] 正如所讨论的,步骤b)可以与片段化分开进行或同时进行。在其中步骤b)与片段化同时进行的实施方案中,这可以是步骤a)的片段化,并且因此作为初始文库制备的部分,或者,如果步骤b)在步骤d)之后进行,则步骤b)可以与步骤c)组合(即,在所述第一个衔接物连接步骤之后发生片段化)。"非发夹衔接物"是不包含发夹环的衔接物。例如,所述非发夹衔接物将不包含单个核酸链,所述单个核酸链由于单个核酸链的一部分与相同单个核酸链的另一部分杂交而形成双链体。双链的非发夹衔接物可以包含两条单独的核酸链,其可以由于一条链的至少一部分与另一条链的至少一部分之间的杂交而形成双链体。

[0132] 在一些实施方案中,非发夹衔接物是或包含核酸。在一些实施方案中,所述非发夹衔接物是或包含DNA、RNA和/或异种核酸(XNA)。所述非发夹衔接物可以包含修饰的和/或未修饰的核苷酸。在一些实施方案中,所述非发夹衔接物是双链的。在一个特定实施方案中、所述非发夹衔接物包含双链DNA。

[0133] 所述非发夹衔接物可以包含能够通过杂交与固定至基质的引物结合的序列。例如,所述非发夹衔接物可以包含与固定至基质的引物至少部分地互补的序列。在一些实施例中,所述序列可以被称作流动池引物或珠子结合的引物的杂交位点。在这样的实施方案中,所述方法可以是用于核酸测序的文库制备方法,其中所述制备包含修饰核酸以适合用于结合至包含固定化引物的基质。所述互补区的长度可以是5、10、15、20、21、22、23、24或更多个碱基。可替换地,所述互补区可以包括5、10、15、20、21、22、23、24或更多个互补碱基。

[0134] 所述非发夹衔接物可以包含与第二引物的至少一部分或全部相同的序列。所述第二引物可以固定化至所述基质或可以是在溶液中。所述相同区域的长度可以是5、10、15、20、21、22、23、24或更多个碱基。所述第一引物和所述第二引物可以被构造为允许在所述基质上扩增核酸。

[0135] 在特定实施方案中,所述非发夹衔接物作为完整衔接物进行连接。因此,在这些实施方案中,不需要执行其它步骤来添加衔接物的特征。因此,所述非发夹衔接物可以作为完整衔接物连接至所述多个核酸,而无需聚合酶步骤或添加或填充任何核酸序列的步骤。具体而言,所述非发夹衔接物可以作为分子连接至所述多个核酸,所述分子包含可与所述基质杂交的序列和能够在所述基质上实现扩增的序列。

[0136] 在一个优选的实施方案中,所述非发夹衔接物是Y-衔接物。"Y-衔接物"包含两条仅部分地互补的链,使得所述Y-衔接物包含这样的部分:其包括两条非互补单链和一个双链互补部分(例如形成"Y"形状)。双链部分的末端可以连接到另一个核酸,并且由于单链部分,这可以导致一个序列连接到核酸的5'端,并且不同的序列连接到核酸的3'端。

[0137] 例如,所述Y-衔接物可以包含第一核酸(例如DNA)链和第二核酸(例如DNA)链。在一个实施方案中,所述第一链在5'至3'方向包含与所述第二链互补的部分和不与所述第二

链互补的部分;并且所述第二链在5’至3’方向包含不与所述第一链互补的部分和与所述第一链互补的部分。

[0138] 在特定实施方案中,所述Y-衔接物作为完整衔接物进行连接。因此,在这些实施方案中,不需要执行其它步骤来添加Y-衔接物的特征。因此,所述Y-衔接物可以作为完整衔接物连接至所述多个核酸,而无需聚合酶步骤或添加或填充任何核酸序列的步骤。具体而言,所述Y-衔接物可以作为分子连接至所述多个核酸,所述分子包含可与所述基质杂交的序列和能够在所述基质上实现扩增的序列。

[0139] Y-衔接物是本领域已知的。例如,所述Y-衔接物可以是包含P5结合序列和P7结合序列的Illumina Y-衔接物。在一个实施方案中,所述Y-衔接物包含序列GTGTAGATCTCGGTGGTCGCCGTATCATT(SEQ ID NO:1)和/或序列CAAGCAGAAGACGGCATACGAGAT(SEQ ID NO:2)。在另一个实施方案中,所述Y-衔接物包含序列ATCTCGTATGCCGTCTTCTGCTTG(SEQ ID NO:3)和/或AATGATACGGCGACCACCGAGATCTACAC(SEQ ID NO:4)。在一个实施方案中,所述Y-衔接物包含SEQ ID NO:1的至少5、10、15、16、17、18、19、20或所有21个碱基。在一个实施方案中,所述Y-衔接物包含SEQ ID NO:2的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基。在一个实施方案中,所述Y-衔接物包含SEQ ID NO:3的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基。在一个实施方案中,所述Y-衔接物包含SEQ ID NO:4的至少5、10、15、16、17、18、19或所有20个碱基。在一个实施方案中,所述Y-衔接物包含SEQ ID NO:1的至少5、10、15、16、17、18、19、20或所有21个碱基和SEQ ID NO:2的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基。在一个实施方案中,所述Y-衔接物包含SEQ ID NO:3的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基和SEQ ID NO:4的至少5、10、15、16、17、18、19或所有20个碱基。所述Y-衔接物可以包含SEQ ID NO:1至4中的任一个的足够碱基以允许与互补引物杂交。

[0140] 所述Y-衔接物可以包含能够通过杂交与第一引物结合的序列和任选的能够通过杂交与第二引物结合的序列。所述第一引物和所述第二引物可以用于核酸的克隆扩增,例如,通过桥式扩增。所述Y-衔接物可以包含能够通过杂交与固定至基质的第一引物结合的序列以及与固定至基质的第二引物的至少一部分或全部相同的序列。例如,所述Y-衔接物可以包含与固定至基质的第一引物至少部分地互补的序列。与第一固定化引物至少部分地互补的序列以及与第二固定化引物的至少一部分相同的序列可以存在于所述Y-衔接物的不同链上,使得它们形成所述Y-衔接物的非互补部分的至少部分。在这些实施方案中,所述方法可以是用于核酸测序的文库制备方法,其中所述制备包含修饰核酸以适合用于结合至包含第一类固定化引物和第二类固定化引物的基质。在这样的实施方案中,所述第一固定化引物和所述Y-衔接物的互补部分以及所述第二固定化引物和所述Y-衔接物的相同部分可以适合用于执行靶核酸的桥式扩增。

[0141] 因此,在一个实施方案中,所述Y-衔接物包含:第一链,其包含与固定至基质的第一引物至少部分地互补的序列;和第二链,其包含与固定至基质的第二引物的至少一个区域相同的序列。在一个实施方案中,所述Y-衔接物包含:第一链,其包含SEQ ID NO:1的至少5、10、15、16、17、18、19、20或所有21个碱基;和第二链,其包含SEQ ID NO:2的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基。在一个实施方案中,所述Y-衔接物包含:第一链,其包含SEQ ID NO:3的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基;和第

二链,其包含SEQ ID NO:4的至少5、10、15、16、17、18、19或所有20个碱基。在一个实施方案中,所述Y-衔接物包含:第一链,其包含根据SEQ ID NO:1的序列;和第二链,其包含根据SEQ ID NO:2的序列。在一个实施方案中,所述Y-衔接物包含:第一链,其包含根据SEQ ID NO:3的序列;和第二链,其包含根据SEQ ID NO:4的序列。

[0142] 在其它实施方案中,所述Y-衔接物可以包含能够通过杂交与固定至基质的第一引物结合的序列以及与未固定至基质的第二引物的至少一部分相同的序列。在一个实施方案中,所述基质可以是珠子。

[0143] 所述非发夹衔接物可以包含测序引物可以与其结合的杂交位点。所述非发夹衔接物可以包含第一测序引物可以与其结合的第一杂交位点和第二测序引物可以与其结合的第二杂交位点。所述第一杂交位点和所述第二杂交位点可以存在于所述非发夹衔接物的不同链上。所述第一和第二杂交位点可以是至少部分地互补的。

[0144] 因此,在一个实施方案中,所述非发夹衔接物(例如Y-衔接物)可以包含:第一链,其包含第一测序引物可以与其结合的第一杂交位点;和第二链,其包含第二测序引物可以与其结合的第二杂交位点。

[0145] 合适的杂交位点的例子在本文中提供为SEQ ID NO:5-8。这些序列是纯粹示例性的。SEQ ID NO:5-8可以各自包含1至7、1至6、1至5、1至4、1至3、2或1个修饰诸如置换、删除或插入。在一个实施方案中,所述修饰是置换。但是,技术人员将理解,只要可以对用于测序的同源引物进行互补修饰,或者只要修饰不影响同源引物的杂交和功能,任何修饰都是可接受的。

[0146] 在其中所述非发夹衔接物包含能够通过杂交与固定至基质的引物结合的序列和测序引物可以与其结合的杂交位点的实施方案中,在连接以后,所述衔接物可以取向为使得能够通过杂交与固定至基质的引物结合的序列位于更靠近末端的位置,并且测序引物可以与其结合的杂交位点位于更靠近连接位点的位置。

[0147] 在一个特定实施方案中,所述非发夹衔接物是Y-衔接物,其包含:

[0148] 第一链,其在5'至3'方向包含第一测序引物可以与其结合的第一杂交位点,以及与第一固定化引物至少部分地互补的序列;和

[0149] 第二链,其在5'至3'方向包含与第二固定化引物相同的序列和第二测序引物可以与其结合的第二杂交位点。所述第一和第二杂交位点可以是至少部分地互补的。

[0150] 所述非发夹衔接物可以包含一个或多个5'和/或3'结合特征。所述结合特征可以是适合于促进衔接物的连接的任一种。例如,所述5'或3'结合特征可以包含以下之一:磷酸酯基团;三磷酸酯"T尾",诸如脱氧胸苷三磷酸"T尾";三磷酸酯"A-尾",诸如脱氧腺苷三磷酸"A-尾";至少一个随机N核苷酸,诸如多个N核苷酸,或任何其它已知的结合基团,以允许衔接物连接至核酸。在一个特定实施方案中,所述5'结合特征是磷酸酯基团且所述3'结合特征是T-尾。

[0151] 在一个特定实施方案中,所述非发夹衔接物(例如Y-衔接物)包含:第一链,其包含5'结合特征,例如磷酸酯基团;和第二链,其包含3'结合特征,例如T-尾。

[0152] 所述非发夹衔接物可以包含一个或多个5'和/或3'保护特征,特别是在其中在步骤d)之前执行步骤b)的实施方案中。所述保护特征可以是能够阻止另一个衔接物连接到受保护的衔接物的任何特征。例如,所述一个或多个保护特征可以阻止发夹衔接物与非发夹

衔接物的连接。所述非发夹衔接物可以包含两种不同的末端保护特征。并非所有实施方案都需要保护特征,例如,具有标记化特征的实施方案可能不需要保护特征的存在。

[0153] 在一个特定实施方案中,所述非发夹衔接物(例如Y-衔接物)包含:第一链,其包含与固定至基质的第一引物至少部分地互补的序列和3'保护特征;和第二链,其包含与固定至基质的第二引物的至少一个区域相同的序列和5'保护特征。

[0154] 所述5'和/或3'保护特征可以包含提供对以下任意一种或多种的抗性的特征:磷酸化活性、磷酸酶活性、末端转移酶活性、核酸杂交、内切核酸酶活性、外切核酸酶活性、连接酶活性、聚合酶活性和蛋白结合。这可以通过本领域技术人员已知的任何方式实现,诸如,但不限于硫代磷酸酯键、亚磷酰胺间隔物、磷酸酯基团、2'-O-甲基基团、反向脱氧和二脱氧-T修饰、锁核酸碱基、二脱氧核苷酸等。所述保护特征可以是C3间隔物亚磷酰胺(3SpC3)。这些特征提供的活性的例子如表1所示。

[0155] 在一个特定实施方案中,所述非发夹衔接物包含5'反向ddT和3'C3间隔物亚磷酰胺。在一个实施方案中,所述非发夹衔接物是包含5'反向ddT和3'C3间隔物亚磷酰胺的Y-衔接物。

[0156] 在一个特定实施方案中,所述非发夹衔接物(例如Y-衔接物)包含:第一链,其包含3'保护特征,例如C3间隔物亚磷酰胺;和第二链,其包含5'保护特征,例如反向ddT。

[0157] 在一个特定实施方案中,所述非发夹衔接物是Y-衔接物,其包含:第一链,其在5'至3'方向包含第一测序引物可以与其结合的第一杂交位点、与第一固定化引物至少部分地互补的序列,以及3'保护特征(例如C3间隔物亚磷酰胺);和第二链,其在5'至3'方向包含5'保护特征(例如反向ddT)、与第二固定化引物的至少一个区域相同的序列,以及第二测序引物可以与其结合的第二杂交位点。任选地,所述第一和第二杂交位点是至少部分地互补的。

[0158] 在一个特定实施方案中,所述非发夹衔接物是Y-衔接物,其包含:第一链,其在5'至3'方向包含5'结合特征(例如磷酸酯基团)、第一测序引物可以与其结合的第一杂交位点、与第一固定化引物至少部分地互补的序列,以及3'保护特征(例如C3间隔物亚磷酰胺);和第二链,其在5'至3'方向包含5'保护特征(例如反向ddT)、与第二固定化引物的至少一个区域相同的序列、第二测序引物可以与其结合的第二杂交位点和3'结合特征(例如T-尾)。任选地,所述第一和第二杂交位点是至少部分地互补的。

[0159] 所述非发夹衔接物可以任选地包含索引序列,其可以被称作条形码。所述索引序列可以允许鉴定来自特定样品的序列。例如,可以在测序之前将不同的样品合并,并且所述索引可以允许在以后鉴定序列所来源的样品。这可以被称作在测序后的去多路化。可以将索引序列定位以便在测序期间读取它,例如,它可以定位在测序引物的杂交位点的3'侧。所述索引序列可以是至少3、4、5、6、7、8、9、10、11、12、13、14、15、20个或更多个核苷酸长的序列。所述索引序列可以是至少3、4、5、6、7、8、9、10、12、13、14、15、20个或更多个核苷酸长的已知序列。所述索引序列可以是至少3、4、5、6、7、8、9、10、11、12、13、14、15、20个或更多个核苷酸长的随机序列。所述索引序列可以是简并或半简并序列。所述索引序列可以是5至10个碱基对长度。所述索引可以是5或7个核苷酸长。所述索引序列可以存在于衔接物的双链部分的两条链上,并且可以是互补的。所述非发夹衔接物可以包含两个索引用于双索引测序。

[0160] 所述非发夹衔接物可以任选地包含单分子标识符(SMI)。SMI的例子公开在WO2013/142389(通过引用并入本文)中。所述SMI可以允许鉴定已经衍生自单个母体分子的

扩增后核酸分子。所述SMI序列可以是双链的、互补SMI序列或单链SMI序列。所述SMI序列可以是简并的或半简并的,并且可以是随机简并序列。双链SMI序列可以包括第一个简并或半简并核苷酸n-聚体序列以及与所述第一个简并或半简并核苷酸n-聚体序列互补的第二个n-聚体序列,而单链SMI序列可以包括第一个简并或半简并核苷酸n-聚体序列。所述第一和/或第二简并或半简并核苷酸n-聚体序列可以是任何合适的长度以产生足够大数目的独特标签来标记来自DNA区段的一组剪切的DNA片段。每个n-聚体序列可以是大约3至20个核苷酸长度。因此,每个n-聚体序列可以是大约4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20个核苷酸长度。在一个实施方案中,所述SMI序列是随机简并核苷酸n-聚体序列,其为12个核苷酸长度。就本发明而言,包含SMI序列并不是必需的,因为在与基质结合之前不需要核酸扩增步骤。因此,在一些实施方案中,所述非发夹衔接物不包含SMI序列。

[0161] 所述Y-衔接物可以包含序列GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT(SEQ ID NO:5)、索引和SEQ ID NO:1,并且这些特征可以从5'至3'按所述顺序排列。所述索引可以是七个碱基长。所述Y-衔接物可以包含序列SEQ ID NO:2、索引和序列GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT(SEQ ID NO:6),并且这些特征可以从5'至3'按所述顺序排列。所述索引可以是五个碱基长。

[0162] 所述Y-衔接物可以包含序列GATCGGAAGAGCACACGTCTGAACTCCAGTCAC(SEQ ID NO:7)、索引和SEQ ID NO:3,并且这些特征可以从5'至3'按所述顺序排列。所述索引可以是七个碱基长。所述Y-衔接物可以包含SEQ ID NO:4、索引和ACACTCTTTCCCTACACGACGCTCTTCCGATCT(SEQ ID NO:8),并且这些特征可以从5'至3'按所述顺序排列。所述索引可以是五个碱基长。

[0163] 在一个特定实施方案中,所述非发夹衔接物是Y-衔接物,其包含:

[0164] 第一链,其在5'至3'方向包含SEQ ID NO:5、任选的索引和SEQ ID NO:1;和

[0165] 第二链,其在5'至3'方向包含SEQ ID NO:2、任选的索引和SEQ ID NO:6。SEQ ID NO:1、2、5和6可以各自包含1至7、1至6、1至5、1至4、1至3、2或1个修饰诸如置换、删除或插入。在一个实施方案中,所述修饰是置换。

[0166] 在一个特定实施方案中,所述非发夹衔接物是Y-衔接物,其包含:

[0167] 第一链,其在5'至3'方向包含SEQ ID NO:7、任选的索引和SEQ ID NO:3;和

[0168] 第二链,其在5'至3'方向包含SEQ ID NO:4、任选的索引和SEQ ID NO:8。SEQ ID NO:3、4、7和8可以各自包含1至7、1至6、1至5、1至4、1至3、2或1个修饰诸如置换、删除或插入。在一个实施方案中,所述修饰是置换。

[0169] 在有利于衔接物与多个核酸内的核酸连接的条件下,将非发夹衔接物提供给多个核酸。所述条件可以根据连接反应的性质和非发夹衔接物的结合特征以及多个核酸的结合特征而变化。例如,所述条件可以促进两个双链核酸之间的连接,其中每个包含5'磷酸酯,并且其中一个包含3'A-尾且另一个包含3'T-尾。将衔接物连接至核酸的其它合适方法以及必要的条件是本领域已知的。在衔接物连接之后可能包括纯化步骤。此步骤可以除去多余的衔接物分子。所述衔接物连接可以是通过也包含片段化(例如,标记化)的技术。

[0170] 在其中在步骤d)之前执行步骤b)的实施方案中,所述连接反应导致非发夹衔接物连接至多个核酸的至少一部分的两端。这可以被称作第一文库。在其中所述非发夹衔接物是Y-衔接物的实施方案中,所述第一文库包含要测序的核酸的片段,其中Y-衔接物连接至

所述片段的至少一部分的每个末端。因此,第一核酸序列可以连接至所述片段内的链的5'端,并且第二核酸序列可以连接至所述片段内的链的3'端。

[0171] 在其中在步骤d)之后执行步骤b)的实施方案中,所述连接反应导致非发夹衔接物连接至不存在发夹的核酸末端。因此,要测序的核酸的至少一部分在一端包含发夹而在另一端包含非发夹衔接物。这可以被称作第二文库。在其中所述非发夹衔接物是Y-衔接物的实施方案中,所述第二文库包含要测序的核酸的片段,其中所述片段的至少一部分包含连接到一端的Y-衔接物和连接在另一端的发夹。

[0172] 步骤c)包含将所述多个核酸片段化,并且可以在步骤b)或步骤d)之后进行。将步骤c)应用于第一文库,并在第二文库的形成之前或期间进行。

[0173] 将核酸片段化的方法是本领域已知的,并且可以是例如机械剪切或酶剪切。所述片段化可以包含超声处理,例如,使用Bioruptor超声破碎器或Covaris超声破碎器。所述片段化可以是基于酶的,并且可以利用基于酶的试剂,所述试剂以时间依赖性方式剪切DNA以产生期望尺寸的片段。合适的商购可得的试剂包括NEBNext dsDNA片段化酶(NEB)。所述片段化可以与第二个衔接物连接步骤同时进行,例如,通过标记化。所述片段化可能导致多个核酸中的双链断裂。所述片段化可能不是位点特异性的,并且因此可能诱导随机断裂,诸如随机双链断裂。所述片段化导致衔接物可以与其连接的双链断裂,任选地在末端修复或类似步骤之后。所述片段化可能导致衔接物可以与其连接的双链断裂,而无需利用一条链作为模板的先前的基于聚合酶的步骤。在一些实施方案中,所述片段化不包含使用位点特异性的切口酶。在一些实施方案中,所述片段化不包含使用位点特异性的切口酶来产生单链部分,其然后使用基于模板的聚合酶进行修复。

[0174] 所述片段化可以包含使用核酸酶,例如,一种内切核酸酶、多种内切核酸酶、一种限制性酶或多种限制性酶。所述片段化可以包含使用核酸引导的内切核酸酶,诸如RNA引导的DNA内切核酸酶。所述片段化可以包含使用Cas蛋白或衍生物或变体。所述片段化可以包含使用Cas9、Cpf1、C2c2、C2c1、CasM、CasMini、逆转录子、原核argonaute、TALEN或大范围核酸酶。

[0175] 所述片段化可以产生具有特定尺寸或具有特定尺寸分布的核酸片段,并且随后可以进行尺寸选择步骤。许多用于尺寸选择和/或净化步骤的系统或试剂是本领域已知的。例如,使用珠子进行尺寸选择以除去或选择特定尺寸的片段。所述珠子可以是固相可逆固定化(SPRI)珠子。商购可得的珠子包括"SPRIselect"(Beckman Coulter)或SPRI珠子(GC Biotech,CNGS-0005)。毛细管DNA电泳可以用于尺寸选择。毛细管DNA电泳也可以用于评估成功连接以及多余衔接物的除去。其它替代方案包括基于凝胶的电泳尺寸选择步骤或系统,例如包含使用琼脂糖凝胶或聚丙烯酰胺凝胶。合适的系统是商购可得的,诸如BluePippin系统(Sage Science)。用于尺寸选择和/或净化的系统的其它例子包括基于DNA提取柱的系统。

[0176] 如果样品在第一衔接物连接之前或期间被片段化,则步骤c)可以包含选择大约为先前片段化步骤的尺寸的一半的片段。

[0177] 步骤c)可以包含除去其尺寸小于约100bp或小于约150bp的片段,和/或保留其尺寸大于约150bp的片段。

[0178] 在一些实施方案中,所得到的片段是100至700bp、150至650bp、200至600bp、250至

550bp、300至500bp或350至450bp。在一些实施方案中,所得到的片段是150至600bp、200至550bp、250至500bp、275至450bp或300至400bp。

[0179] 可以处理所述片段化的核酸以适合用于衔接物连接。例如,可以向核酸添加一个结合特征或多个结合特征。所述结合特征可以包含5'特征和/或3'特征。所述结合特征可以是适合于促进衔接物的连接的任一种,包括本文中公开的任何结合特征。如果执行步骤b)以产生第一文库,则由于保护特征在非发夹衔接物上的存在,因此可能不会向非发夹衔接物添加结合特征。

[0180] 在一个特定实施方案中,所述片段化的核酸是末端钝化的且带有A尾。因此,5'磷酸酯和/或3'A-尾可以添加到片段化的核酸中。

[0181] 因此,在一个实施方案中,步骤c)可以是如下:

[0182] c)将所述多个核酸片段化;并且进一步包含:

[0183] 任选地i)基于尺寸选择所述多个核酸的片段;和

[0184] ii)将5'和/或3'结合特征添加至所述多个核酸。

[0185] 步骤i)和ii)可以按照i)和然后ii)的顺序进行。但是,可以遵循任何顺序,其允许制备适合于本文公开的下游步骤的多个核酸。

[0186] 在特定的非限制性实施方案中,步骤c)可以是如下:

[0187] c)将所述多个核酸片段化;并且进一步包含:

[0188] i)基于尺寸选择所述多个核酸的片段;

[0189] ii)将所述选择的核酸末端钝化;和

[0190] iii)将A-尾添加至所述末端钝化的核酸。

[0191] 在其它实施方案中,步骤c)包含将核酸片段化和连接衔接物的步骤。例如,标记化步骤。所述连接的衔接物可以是所述非发夹衔接物或所述发夹衔接物,取决于执行所述步骤的顺序。

[0192] 因此,步骤c)和步骤b)可以如下组合:

[0193] i)将非发夹衔接物片段化并连接至所述多个核酸;和

[0194] 任选地ii)基于尺寸选择所述多个核酸的片段。

[0195] 可替换地,步骤c)和步骤d)可以如下组合:

[0196] i)将发夹衔接物片段化并连接至所述多个核酸;和

[0197] 任选地ii)基于尺寸选择所述多个核酸的片段。

[0198] 在其它实施方案中,步骤c)可以包含将识别位点插入所述片段化的核酸的标记化步骤。例如,能够形成发夹的酶(诸如端粒末端转移酶原)的识别位点。所述端粒末端转移酶原可以是TelN。

[0199] 在一些实施例中,步骤d)包含在有利于连接的条件下将所述多个核酸暴露于发夹衔接物。因此,所述发夹衔接物将连接至所述核酸的可用末端或未受保护的末端。在其中在步骤b)之前进行步骤d)的实施方案中,这将导致发夹衔接物连接至所述多个核酸的至少一部分的两端。在其中在步骤b)之后进行步骤d)的实施方案中,这将导致非发夹衔接物连接至发夹衔接物未与其连接的核酸末端。

[0200] 在其它实施例中,步骤d)包含将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件。例如,步骤d)可以包含使用能够在双链核酸分子中产生共价闭合末端的条件

或酶。合适的酶的一个例子是端粒末端转移酶原,诸如TelN。TelN识别序列可以存在于多个核酸中,或者可以已经引入多个核酸中。例如,可以将TelN识别序列作为片段化的部分通过标记化而引入。

[0201] 如所讨论的,步骤d)可以与片段化分开进行或同时进行。在其中步骤d)与片段化同时进行的实施方案中,这可以是步骤a)的片段化并因此作为初始文库制备的部分,或者,如果在步骤b)之后进行步骤d),则步骤d)可以与步骤c)组合(即,在所述第一个衔接物连接步骤之后发生片段化)。

[0202] "发夹"衔接物包含发夹环。发夹衔接物可以包含单个核酸链,其通过单个核酸链的一部分与相同单个核酸链的另一部分的杂交而形成双链体。发夹衔接物是本领域已知的。发夹衔接物可以被称作"U-衔接物"。

[0203] 仅在一端存在发夹的双链核酸能够变性以形成包含原始双链核酸的两条链的单位单链分子。

[0204] 在一些实施方案中,发夹衔接物是或包含核酸。在一些实施方案中,所述发夹衔接物是或包含DNA、RNA和/或XNA。所述发夹衔接物可以包含修饰的和/或未修饰的核苷酸。在一个特定实施方案中,所述发夹衔接物包含DNA。

[0205] 发夹衔接物的一个非限制性例子是:GGGCCTADDDDDDDDTAGGCCCT(SEQ ID NO:9),其中D是G、A或T(但不是C)。

[0206] 在有利于衔接物与多个核酸内的核酸连接的条件下,可以向多个核酸提供发夹衔接物。所述条件可以根据连接反应的性质和非发夹衔接物的结合特征以及多个核酸的结合特征而变化。例如,所述条件可以促进两个双链核酸之间的连接,其中每个包含5'磷酸酯,并且其中一个包含3'A-尾且另一个包含3'T-尾。将衔接物连接至核酸的其它合适方法以及必要的条件是本领域已知的。在衔接物连接之后可能包括纯化步骤。此步骤可以除去多余的衔接物分子。

[0207] 在其中在步骤b)之前进行步骤d)的实施方案中,步骤d)导致发夹衔接物连接至多个核酸的至少一部分的两端,或在多个核酸的至少一部分的两端处形成发夹。这可以被称作第一文库。在这样的实施方案中,所述方法可以包含除去任何线性核酸的步骤。该步骤可能导致仅保留在两端均存在发夹的核酸,其本质上是核酸环。

[0208] 在其中在步骤b)之后进行步骤d)的实施方案中,步骤d)导致发夹衔接物连接至非发夹衔接物未与其连接的核酸末端,或者在非发夹衔接物未与其连接的核酸末端处形成发夹。因此,要测序的核酸的至少一部分在一端包含发夹并且在另一端包含非发夹衔接物。这可以被称作第二文库。

[0209] 如所讨论的,步骤d)可以与片段化分开进行或同时进行。在其中步骤d)与片段化同时进行的实施方案中,这可以是步骤a)的片段化,并且因此作为初始文库制备的部分,或者,如果在步骤b)之后进行步骤d),则步骤d)可以与步骤c)组合(即,在所述第一个衔接物连接步骤之后发生片段化)。

[0210] 在一个实施方案中,提供了一种用于核酸测序的文库制备方法,所述方法包含按列举顺序的以下依次步骤:

[0211] 提供多个核酸;

[0212] 在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物;

[0213] 将所述多个核酸片段化;和

[0214] 在有利于连接的条件下将所述多个核酸暴露于发夹衔接物。

[0215] 在一个实施方案中,提供了一种用于核酸测序的文库制备方法,所述方法包含按列举顺序的以下依次步骤:

[0216] 提供多个核酸;

[0217] 在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物;

[0218] 将所述多个核酸片段化;和

[0219] 将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件。

[0220] 在另一个实施方案中,提供了一种用于核酸测序的文库制备方法,所述方法包含按列举顺序的以下依次步骤:

[0221] 提供多个核酸;

[0222] 在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物;和

[0223] 在有利于连接和片段化的条件下将所述多个核酸暴露于发夹衔接物。

[0224] 在一个实施方案中,提供了一种用于核酸测序的文库制备方法,所述方法包含按列举顺序的以下依次步骤:

[0225] 提供多个核酸;

[0226] 在有利于连接的条件下将所述多个核酸暴露于发夹衔接物;

[0227] 将所述多个核酸片段化;和

[0228] 在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物。

[0229] 在一个实施方案中,提供了一种用于核酸测序的文库制备方法,所述方法包含按列举顺序的以下依次步骤:

[0230] 提供多个核酸;

[0231] 将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件;

[0232] 将所述多个核酸片段化;和

[0233] 在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物。

[0234] 在另一个实施方案中,提供了一种用于核酸测序的文库制备方法,所述方法包含按列举顺序的以下依次步骤:

[0235] 提供多个核酸;

[0236] 在有利于连接的条件下将所述多个核酸暴露于发夹衔接物;和

[0237] 在有利于连接和片段化的条件下将所述多个核酸暴露于非发夹衔接物。

[0238] 在一些实施方案中,所述方法可以进一步包含,在适合于所述非发夹衔接物的部分与固定化引物的至少一部分杂交的条件下,使所述多个核酸(其在该阶段可以被称作第二文库)与包含固定化引物的基质接触。

[0239] 任何缺乏连接的非发夹衔接物的核酸都不会与流动池杂交。例如,在两端都包括发夹衔接物的任何核酸。

[0240] 所述基质可以是固体表面诸如流动池、珠子、载玻片或膜的表面。具体而言,所述基质可以是流动池。所述基质可以是图案化的或非图案化的流动池。所述基质可以包含玻璃、石英、二氧化硅、金属、陶瓷或塑料。所述基质表面可以包含聚丙烯酰胺基质或涂层。

[0241] 本文中使用的术语“流动池”旨在具有本领域中、特别是在合成测序领域中的普通

含义。示例性流动池包括、但不限于在核酸测序装置中使用的流动池,诸如用于由Illumina,Inc.(San Diego,Calif.)商业化的Genome Analyzer®、MiSeq®、NextSeq®、HiSeq®或NovaSeq®平台的流动池;或用于由Life Technologies(Carlsbad,Calif.)商业化的SOLiD™或Ion Torrent™测序平台的流动池。示例性流动池及其制造和使用方法也描述在例如WO2014/142841A1、美国专利申请公开号2010/0111768A1和美国专利号8,951,781中。

[0242] 所述基质可以包含固定化引物,例如,两种类型的引物,它们可以一起充当桥式扩增的正向引物和反向引物。固定至"基质"是指,即使在使双链核酸变性的条件下,引物也与基质结合。例如,所述引物可以共价结合至基质。所述引物定向使得5'端位于固定点近侧且3'端位于固定点远侧。这样的布置在本领域中是标准的。

[0243] 在一个特定实施方案中,所述基质可以包含第一和第二固定化引物。在一些实施方案中,所述固定化引物可以适合用于充当在桥式扩增过程中的引物。桥式扩增可以导致固定至基质的核酸的克隆扩增。所述非发夹衔接物可以是Y-衔接物,其包含与所述第一固定化引物互补的序列和与所述第二固定化引物相同的序列。因此,在这样的实施方案中,所述第二文库包含这样的核酸:其在一端包括发夹衔接物,并且在另一端包括与连接到一条链的第一固定化引物互补的序列以及与连接到另一条链的第二固定化引物相同的序列。与第一固定化引物互补的序列可以连接到核酸的3'端,并且与第二固定化引物相同的序列可以连接到核酸的5'端。

[0244] 在一些实施方案中,所述第二文库可以在接触基质之前变性,使得第二文库的核酸是单链的。在一些实施方案中,所述第二文库可以在变性条件下与基质接触,使得在文库内的核酸在接触时是单链的。

[0245] 在一个特定实施方案中,公开了一种用于核酸测序的文库制备方法,其中所述制备包含修饰核酸以适合用于结合至包含第一固定化引物的基质,所述方法包含:

[0246] a)提供多个核酸;

[0247] b)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物,其中所述非发夹衔接物包含与所述第一固定化引物互补的序列;

[0248] c)将所述多个核酸片段化;和

[0249] d)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件。在一个实施方案中,所述步骤可以按照a)、b)、c)和然后d)的顺序进行。这些步骤可以是依次的,或步骤c)和d)可以组合,例如,作为标记化步骤。步骤b)可以与更早的片段化步骤组合。在另一个实施方案中,所述步骤可以按照a)、d)、c)和然后b)的顺序进行。这些步骤可以是依次的,或步骤c)和b)可以组合,例如,作为标记化步骤。步骤d)可以与更早的片段化步骤组合。

[0250] 在上述的步骤之后,所述方法可以进一步包含:

[0251] e)在适合于所述第一固定化引物与互补核酸杂交的条件下,使所述多个核酸与所述基质接触。作为一个例子,所述基质可以是适合用于核酸测序的流动池。

[0252] 在特定实施方案中,在步骤e)之前不进行核酸扩增步骤,诸如PCR。例如,可以从组织样品开始并用样品中的gDNA片段通过与固定化引物杂交的连接的衔接物结合至测序流

动池结束,执行所述方法;其中,在此过程中不进行核酸扩增步骤,诸如PCR步骤。尽管可以包括PCR步骤以扩增靶标,但发明人惊奇地发现,这并不是本发明方法的要求。扩增步骤的排除可以有利地避免由于扩增对偏差的引入或对序列错误的引入。因此,本发明的排除扩增步骤的方法可以用于全基因组错误校正测序。

[0253] 因此,在一个实施方案中,公开了一种用于核酸测序的文库制备方法,其中所述制备包含修饰核酸以适合用于结合至包含第一固定化引物的基质,所述方法包含按列举顺序的下述步骤,其中步骤c)和d)可以组合:

[0254] a)提供多个核酸;

[0255] b)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物,其中所述非发夹衔接物包含与所述第一固定化引物互补的序列;

[0256] c)将所述多个核酸片段化;

[0257] d)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件;和

[0258] e)在适合于所述第一固定化引物与互补核酸杂交的条件下,使所述多个核酸与所述基质接触;

[0259] 其中在步骤e)之前不进行核酸扩增步骤,例如PCR。所述非发夹衔接物可以是本文中公开的任一种,诸如Y-衔接物。所述基质可以是本文中公开的任一种,诸如流动池。

[0260] 在一个替代实施方案中,公开了一种用于核酸测序的文库制备方法,其中所述制备包含修饰核酸以适合用于结合至包含第一固定化引物的基质,所述方法包含按列举顺序的下述步骤,其中步骤c)和d)可以组合:

[0261] a)提供多个核酸;

[0262] b)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件;

[0263] c)将所述多个核酸片段化;

[0264] d)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物,其中所述非发夹衔接物包含与所述第一固定化引物互补的序列;和

[0265] e)在适合于所述第一固定化引物与互补核酸杂交的条件下,使所述多个核酸与所述基质接触;

[0266] 其中在步骤e)之前不进行核酸扩增步骤,例如PCR。所述非发夹衔接物可以是本文中公开的任一种,诸如Y-衔接物。所述基质可以是本文中公开的任一种,诸如流动池。

[0267] 在步骤e)之后,所述方法可以进一步包含在适合于延伸固定化引物的条件下使任何杂交的核酸与聚合酶接触以合成核酸,其为与杂交的核酸互补的核苷酸链。然后可以扩增新形成的核酸。在一些实施方案中,用于扩增的引物也固定至基质,并且可以例如适用于桥式扩增。该过程是本领域已知的,并形成核酸的克隆簇。在其它实施例中,用于扩增的引物可以是在溶液中,例如对于其中基质是珠子的实施方案。然后可以按照通常的方法对扩增的核酸进行测序,例如通过合成测序。所述非发夹衔接物可以包含用于结合测序引物的位点以协助该过程。所述非发夹衔接物也可以包含索引。

[0268] 因此,在一个实施方案中,所述方法可以进一步包含:

[0269] f)得到在步骤e)中与所述基质杂交的任何核酸的序列信息。

[0270] 在其中不进行步骤e)的实施方案中,可以通过对第二文库进行测序来得到序列信息。

[0271] 包括得到序列信息的步骤的方法可以被称作核酸测序方法或错误校正核酸测序方法。这样的方法是“错误校正的”,因为序列信息来源于双链核酸的部分的两条链,并且因此,在提供核酸进行测序之后引入的任何错误都可以通过对比针对一条链得到的序列与针对另一条链得到的序列来校正。从本质上讲,原始核酸样品的每个部分被读取两次,并且每次读取都是一个独立序列,因此允许对仅存在于单次读取中的任何差异进行错误校正。

[0272] 在一个实施方案中,所述方法用于鉴定突变,并且所述方法包括将预期序列中在DNA分子的两条链上一致的任何变化鉴定为突变,并且如果变化在DNA分子的两条链上不一致,则不将预期序列中的任何变化鉴定为突变。这样的方法可以包括将序列读取与参考序列进行生物信息学比对,以鉴定与预期序列的偏差。所述参考序列可以是已知的序列例如人基因组,诸如在NCBI数据库中的人基因组参考序列Human Build 38补丁版本14(GRCh38.p14;Genome Reference Consortium)。

[0273] 在特定实施方案中,可以将所述方法应用于从样品得到的gDNA,并可以用于无偏全基因组错误校正测序。在一些实施方案中,可以采用所述方法来检测基因编辑技术的脱靶效应。例如,所述方法可以用于检测CRISPR-Cas9编辑、TALEN编辑或任何其它改变核酸序列的方法的脱靶效应。

[0274] 对核酸(诸如固定化的核酸簇)进行测序的方法是本领域已知的。在一些实施方案中,所述测序可能涉及使用一种测序引物或多种测序引物。例如,本文描述的非发夹衔接物的实施方案可以包含第一测序引物可以与其结合的第一杂交位点,并且步骤f)可以包含使用所述第一测序引物。在一些实施方案中,本文描述的非发夹衔接物也可以包含第二测序引物可以与其结合的第二杂交位点,并且步骤f)也可以包含使用第二测序引物。

[0275] 所述测序可以是下一代测序或可以是大规模平行测序。

[0276] 在一个特定实施方案中,提供了一种用于核酸测序的文库制备方法,其中所述制备包含修饰核酸以适合用于结合至包含固定化引物的基质,所述方法包含:

[0277] a)提供多个核酸;

[0278] b)在有利于连接的条件下将所述多个核酸暴露于Y-衔接物以产生第一文库;其中所述Y-衔接物包含:

[0279] 第一链,其包含与固定至基质的第一引物至少部分地互补的序列和任选的3’保护特征,和

[0280] 第二链,其包含与固定至基质的第二引物的至少一个区域相同的序列和任选的5’保护特征;

[0281] c)将所述第一文库片段化,并且进一步包含:

[0282] i)基于尺寸选择所述多个核酸的片段;和

[0283] d)在有利于连接的条件下将所选择的片段暴露于发夹衔接物以产生第二文库,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件。

[0284] 在另一个实施方案中,提供了一种用于核酸测序的文库制备方法,其中所述制备包含修饰核酸以适合用于结合至包含固定化引物的基质,所述方法包含:

[0285] a)提供多个核酸;

[0286] b)在有利于连接的条件下将所述多个核酸暴露于Y-衔接物以产生第一文库;其中所述Y-衔接物包含:

[0287] 第一链,其包含与固定至基质的第一引物至少部分地互补的序列和任选的3'保护特征,和

[0288] 第二链,其包含与固定至基质的第二引物的至少一个区域相同的序列和任选的5'保护特征;和

[0289] (组合的步骤)c)和d)在有利于连接和片段化的条件下将所述第一文库暴露于发夹衔接物以产生第二文库;任选地其中执行标记化。

[0290] 上述两个实施方案可以是从核酸得到序列信息的方法,其中所述方法进一步包含:

[0291] e)使所述第二文库变性以产生单链核酸,并在适合于所述第一固定化引物与互补核酸杂交的条件下使所述单链核酸与所述基质接触,并任选地通过桥式扩增产生固定化的核酸簇,其中所述第一和第二固定化引物充当桥式扩增的引物;和

[0292] f)得到在步骤e)中与所述基质杂交的任何核酸的序列信息。

[0293] 在第二方面,提供了通过本公开内容的第一方面的任意方法得到或可得到的核酸文库。所述第二方面的文库相对于本公开内容的第一方面被称作第二文库。

[0294] 所述第二方面的核酸文库包含需要其序列信息的核酸,其可以被称作靶核酸并且可以是衍生自样品的DNA(或衍生自所述DNA)。所述DNA可以衍生自哺乳动物或人样品。所述靶核酸可以是gDNA的片段或可以衍生自所述gDNA。

[0295] 所述文库包含靶核酸的部分,其在一端具有本文公开的连接的非发夹衔接物,并在另一端具有连接的发夹衔接物。与本发明文库的核酸连接的非发夹衔接物可以是本文公开的任一种。在一个实施方案中,靶核酸的部分在一端具有本文公开的连接的Y-衔接物,并在另一端具有本文公开的连接的发夹衔接物。所述Y-衔接物可以是包含P5结合序列和P7结合序列的Illumina Y-衔接物。

[0296] 本公开内容涵盖所述第二方面的文库,其已经变性以形成单链,使得形成发夹的部分在靶核酸的两条链之间形成接头,并且所述非发夹衔接物作为在5'末端的序列和在3'末端的序列存在。

[0297] 除了上述物质之外,所述第二方面的核酸文库可以包含在两端具有发夹的靶核酸。这样的物质不会结合基质,并因此不可测序。

[0298] 与现有技术相比,所述第二方面的核酸文库包含减少量的具有连接到两端的非发夹衔接物的靶核酸。在一些实施方案中,所述核酸文库不包含或不包含大量具有连接到两端的非发夹衔接物的靶核酸。这样的物质是可测序的,但不可错误校正,并且因此该物质的减少或避免允许提高测序准确性。此外,该物质的减少可以允许在没有先前扩增步骤的情况下对文库进行测序,例如,它可以允许在应用于基质之前没有扩增的情况下在基质上对文库进行测序。

[0299] 在实例中,所述文库可以包含小于99.9%、99%、95%、90%、80%、70%、60%、50%、40%、30%、20%、10%、5%、3%、1%、0.1%或0.01%(按质量计)的具有连接到两端的非发夹衔接物的靶核酸。

[0300] 因此,在一个特定实施方案中,公开了一种核酸文库,其包含具有连接到一端的非

发夹衔接物和在另一端处的发夹的靶核酸。所述核酸文库不包含、包含减少量的、或者不包含大量的具有连接到一端的非发夹衔接物和连接到另一端的非发夹衔接物的靶核酸。所述减少可以是相对于以相同方式制备的文库,但其中第一和第二衔接物连接步骤同时进行。

[0301] 具体而言,所述第二方面的核酸文库可以适合用于包含使所述文库与基质接触以使所述文库的部分结合至基质的测序方法。所述非发夹衔接物可以包含与固定至基质的第一引物至少部分地互补的序列。因此,在一个实施方案中,公开了一种核酸文库,其适合用于包含使所述文库与基质接触以使所述文库的部分结合至基质的测序方法,所述核酸文库包含:i)靶核酸,其具有连接到一端的非发夹衔接物和在另一端的发夹,其中所述非发夹衔接物包含与固定至基质的第一引物至少部分地互补的序列;和任选的ii)靶核酸,其具有在一端的发夹和在另一端的发夹。

[0302] 在第三方面,公开了一种测序方法,其中所述方法包含得到在本公开内容的第二方面的文库内的核酸的序列信息。

[0303] 在一个实施方案中,公开了一种得到测序信息的方法,其中所述方法包含:

[0304] 1)在适合于所述第一固定化引物与互补核酸杂交的条件下,使本公开内容的第二方面的文库与包含第一固定化引物的基质接触;和

[0305] 2)得到在步骤1)中与所述基质杂交的任何核酸的序列信息。

[0306] 所述第三方面的步骤1)具有与本公开内容的第一方面的步骤e)相同的特征。所述第三方面的步骤2)具有与本公开内容的第一方面的步骤f)相同的特征。

[0307] 在一个实施方案中,公开了一种得到测序信息的方法,其中所述方法包含:

[0308] 1)在适合于所述第一固定化引物与互补核酸杂交的条件下,使核酸文库与包含第一固定化引物的基质的接触;和

[0309] 2)得到在步骤1)中与所述基质杂交的任何核酸的序列信息;

[0310] 其中所述核酸文库通过包含以下步骤的方法制备或可得到:

[0311] a)提供多个核酸;

[0312] b)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物;

[0313] c)将所述多个核酸片段化;和

[0314] d)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件;

[0315] 其中步骤b)和d)分开进行。

[0316] 如本文公开的,发明人惊奇地发现,在与基质结合之前对样品中的核酸进行扩增并不是本发明方法的要求。

[0317] 因此,在第四方面,提供了一种用于核酸测序的文库制备方法,所述方法包含:

[0318] i)提供多个核酸;

[0319] ii)在有利于连接的条件下将所述多个核酸暴露于非发夹衔接物;和

[0320] iii)在有利于连接的条件下将所述多个核酸暴露于发夹衔接物,或将所述多个核酸暴露于能够在核酸分子末端处形成发夹的条件;

[0321] 其中在所述文库制备过程中不扩增所述核酸。

[0322] 与本公开内容的第一方面的步骤a)关联地公开的特征也适用于所述第四方面的步骤i)。

[0323] 所述第四方面的非发夹衔接物可以是为本公开内容的第一方面所公开的任一种。所述非发夹衔接物可以包括关于所述第一方面所公开的保护特征和/或结合特征。所述第四方面的发夹衔接物可以是本公开内容的第一方面所公开的任一种。所述能够在核酸分子末端处形成发夹的条件可以是本公开内容的第一方面所公开的任一种。

[0324] 根据第四方面,在所述文库制备过程中不扩增所述核酸。例如,不进行PCR步骤。

[0325] 在一个特定实施方案中,所述第四方面的步骤按照i)、ii)和然后iii)的顺序进行。在另一个实施方案中,所述第四方面的步骤按照i)、iii)和然后ii)的顺序进行。在一个特定实施方案中,步骤ii)和iii)分开进行并且在所述步骤之间包括片段化步骤。在另一个实施方案中,所述第二个连接步骤可以包含片段化,例如,它可以是标记化步骤。与本公开内容的第一方面的步骤c)关联地公开的特征也适用于所述第四方面的片段化步骤。

[0326] 由步骤i)、ii)和iii)产生的核酸文库可以被称作第二文库。可以从所述第二文库得到序列信息。在一个实施方案中,所述非发夹衔接物包含与固定至基质的第一引物至少部分地互补的序列,并且所述方法包含步骤iv),在适合于所述第一固定化引物与互补核酸杂交的条件下,使所述第二文库与包含第一固定化引物的基质接触。与本公开内容的第一方面的步骤e)关联地公开的特征也适用于所述第四方面的步骤iv)。与得到所述第一方面的序列信息关联地公开的特征也适用于所述第四方面。在这些实施方案中,在步骤iv)之前不进行核酸扩增步骤。

[0327] 在第五方面,提供了通过本公开内容的第四方面的任何方法得到或可得到的核酸文库。所述第五方面的文库相对于本公开内容的第一方面被称作第二文库。

[0328] 所述第五方面的文库不包含已经扩增的靶核酸,例如,未进行PCR反应的靶核酸。所述第五方面的文库的其余特征可以如关于本公开内容的第二方面所公开。

[0329] 在第六方面,公开了一种测序方法,其中所述方法包含得到在本公开内容的第五方面的文库内的核酸的序列信息。

[0330] 在一个实施方案中,公开了一种得到测序信息的方法,其中所述方法包含:

[0331] 1)在适合于所述第一固定化引物与互补核酸杂交的条件下,使本发明的第五方面的文库与包含第一固定化引物的基质接触;和

[0332] 2)得到在步骤1)中与所述基质杂交的任何核酸的序列信息。

[0333] 所述第六方面的步骤1)具有与本公开内容的第一方面的步骤e)相同的特征。所述第六方面的步骤2)具有与本公开内容的第一方面的步骤f)相同的特征。

[0334] 在第七方面,提供了一种非发夹衔接物,其包含:第一链,其在5'至3'方向包含与第一固定化引物至少部分地互补的序列和3'保护特征;和第二链,其在5'至3'方向包含5'保护特征和与第二引物的至少一个区域相同的序列。

[0335] 所述第七方面的非发夹衔接物可以包含如关于本公开内容的第一方面公开的与第一固定化引物至少部分地互补的任何序列。所述第七方面的非发夹衔接物可以包含如关于本公开内容的第一方面公开的与第二引物的至少一个区域相同的任何序列。与第一固定化引物至少部分地互补的序列可以是SEQ ID NO:1或SEQ ID NO:3。与第二固定化引物的至少一个区域相同的序列可以是SEQ ID NO:2或SEQ ID NO:4。在一个实施方案中,所述非发夹衔接物包含SEQ ID NO:1的至少5、10、15、16、17、18、19、20或所有21个碱基和/或包含SEQ ID NO:2的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基。在一个实施方案中,

所述非发夹衔接物包含SEQ ID NO:3的至少5、10、15、16、17、18、19、20、21、22、23或所有24个碱基和/或包含SEQ ID NO:4的至少5、10、15、16、17、18、19或所有20个碱基。所述非发夹衔接物可以包含SEQ ID NO:1至4中的任一个的足够碱基以允许与互补引物杂交。

[0336] 在一个特定实施方案中,所述非发夹衔接物是Y-衔接物,其包含:

[0337] 第一链,其在5’至3’方向包含SEQ ID NO:5、任选的索引、SEQ ID NO:1和3SpC3;和

[0338] 第二链,其在5’至3’方向包含5’阻断、SEQ ID NO:2、任选的索引和SEQ ID NO:6。SEQ ID NO:1、2、5和6可以各自包含1至7、1至6、1至5、1至4、1至3、2或1个修饰诸如置换、删除或插入。在一个实施方案中,所述修饰是置换。

[0339] 在一个特定实施方案中,所述非发夹衔接物是Y-衔接物,其包含:

[0340] 第一链,其在5’至3’方向包含SEQ ID NO:7、任选的索引、SEQ ID NO:3和3SpC3;和

[0341] 第二链,其在5’至3’方向包含5’阻断、SEQ ID NO:4、任选的索引和SEQ ID NO:8。SEQ ID NO:3、4、7和8可以各自包含1至7、1至6、1至5、1至4、1至3、2或1个修饰诸如置换、删除或插入。在一个实施方案中,所述修饰是置换。

[0342] 所述5’和3’保护特征可以是如关于本公开内容的第一方面公开的任一种。

[0343] 在一些实施方案中,非发夹衔接物是或包含核酸。在一些实施方案中,所述非发夹衔接物是或包含DNA、RNA和/或XNA。所述非发夹衔接物可以包含修饰的和/或未修饰的核苷酸。在一些实施方案中,所述非发夹衔接物是双链的。在一个特定实施方案中,所述非发夹衔接物包含双链DNA。所述非发夹衔接物可以是Y-衔接物。

[0344] 所述非发夹衔接物可以包含如关于本公开内容的第一方面公开的任何5’和/或任何3’结合特征。所述非发夹衔接物可以包含或可以不包含如关于本公开内容的第一方面公开的任何索引。

[0345] 在一个特定实施方案中,所述非发夹衔接物是Y-衔接物,其包含:第一链,其在5’至3’方向包含第一测序引物可以与其结合的第一杂交位点、与第一固定化引物至少部分地互补的序列,以及3’保护特征;和第二链,其在5’至3’方向包含5’保护特征、与第二固定化引物的至少一个区域相同的序列,以及第二测序引物可以与其结合的第二杂交位点。任选地,所述第一和第二杂交位点是至少部分地互补的。

[0346] 在本发明的第八方面,提供了一种试剂盒,其包含本公开内容的第七方面的非发夹衔接物和发夹衔接物。所述发夹衔接物可以是如关于本公开内容的第一方面公开的任一种。

[0347] 表1

[0348]

寡物修饰(P5 或 P7 衔接物)	活性
硫代磷酸酯键(寡物的 5'和 3'端)	耐受 5'>3'和 3'>5'外切核酸酶活性。
硫代磷酸酯键(遍布寡物)	耐受内切核酸酶活性。
C3 间隔物亚磷酰胺(3')	耐受外切核酸酶、连接酶、末端转移酶、5'>3'聚合酶活性。
3'磷酸酯基团	抑制一些 3'-外切核酸酶的降解，可以用于阻断 DNA 聚合酶延伸，并且耐受连接。
2'-O-甲基(2'OMe)	与未修饰的 DNA 相比，包括该修饰的 DNA 寡核苷酸对 DNA 酶的敏感性通常低 5-10 倍。
反向 dT (3')	在寡核苷酸的 3'端的反向 dT 导致 3'-3'连接，其将抑制 3'外切核酸酶的降解和 DNA 聚合酶的延伸。
5'反向 ddT	在寡核苷酸的 5'端的 5'反向 ddT 耐受连接并且可以提供对外切核酸酶的一些形式的抗性。
锁核酸碱基	增加寡物的熔化温度并且阻止假杂交，增加结合特异性。
二脱氧胞苷(ddC) (3')	3'链终止子，其阻止聚合酶的 3'延伸。

[0349]

DEDUCE-seq 衔接物:版本 1 - P7 在 3'端与流动池杂交

粗体加下划线的序列: Y 衔接物的互补区

	Y-衔接物		DNA 插入片段		发夹衔接物	
						DDD
5'	阻断 /A*ATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATC*T	3'	NNNNNNA	5'	P-GGGCCTA	D
3'	/3SpC3/GTTCGTCTTCTGCCGTATGCTCTA[i7]CACTGACCTCAAGTCTGCACA**CGAGAAGGCTAG**-P	5'	ANNNNNN	3'	T*CCCGGAT	D
						DDD

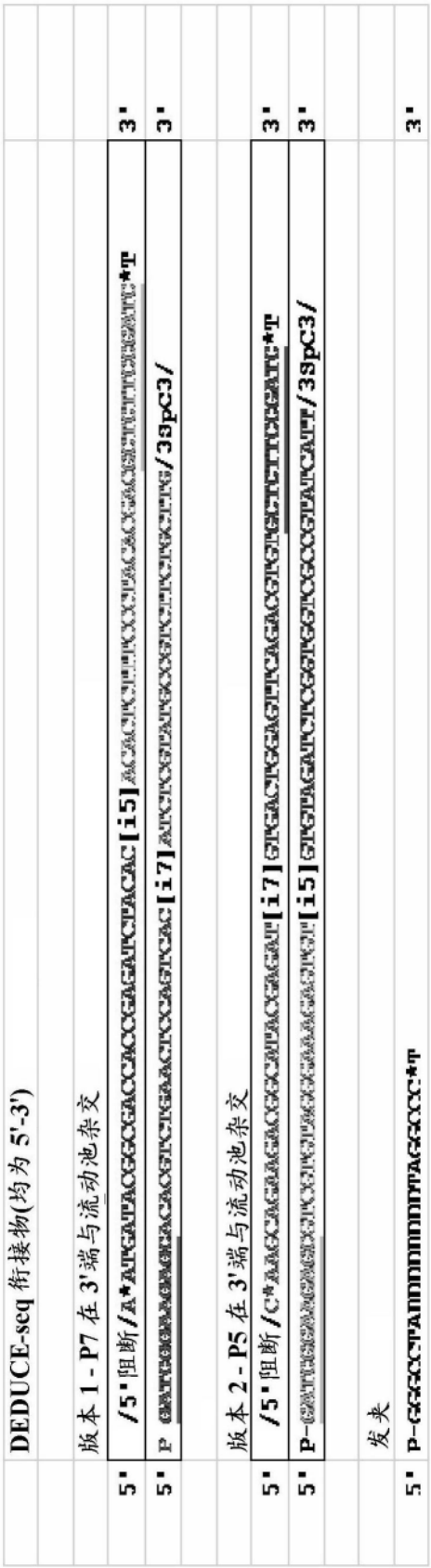
DEDUCE-seq 衔接物:版本 2 - P5 在 3'端与流动池杂交

	Y-衔接物		DNA 插入片段		发夹衔接物	
						DDD
5'	/5'阻断/C*AAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGT**GCTCTTCCGATC***T	3'	NNNNNNA	5'	P-GGGCCTA	D
3'	/3SpC3/TTACTATGCCGCTGGTGGCTCTAGATGTG[i5]TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAG-P	5'	ANNNNNN	3'	T*CCCGGAT	D
						DDD

[0350]

直链结构-一旦变性并且加载至 Illumina 流动池上													
DEDUCE-seq 衔接物:版本 1 - P7 在 3'端与流动池杂交													
5' P5 衔接物				DNA 插入片段	直链发夹衔接物			DNA 插入片段			3' P7 衔接物		
5'	阻断/i5/				NNNNNA	5'	P-GGGCCTAATTTT

[0351]



	DEDUCE-seq 衔接物(均为 5'-3')	
	版本 1 - P7 在 3'端与流动池杂交	
5'	/5'阻断/A*ATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATC*T	3'
5'	P GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG/3SpC3/	3'
	版本 2 - P5 在 3'端与流动池杂交	
5'	/5'阻断/C*AAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
5'	P-GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT[i5]GTGTAGATCTCGGTGGTCGCCGTATCATT/3SpC3/	3'
	发夹	
5'	P-GGGCCTAAAAAAAAAATAGGCCC*T	3'

[0352]

茎		全序列+潜在衍生物
	7 bp	5' P-GGGCCTADDDDDDDDTAGGCCC*T 3'
环		茎 环 茎 T-尾
	8 bp	
		DDD
		5' GGGCCTA D
		3' T*CCCGGAT D
		DDD
		D=G、A 或 T (不是 C)

[0353] *硫代磷酸酯键耐受外切核酸酶活性

[0354] /3SpC3/C3间隔物亚磷酰胺(共价阻断)耐受外切核酸酶、连接酶和5'>3'聚合酶活性

[0355] -P 5'磷酸酯基团促进5'连接

[0356] 索引 Illumina测序索引序列实现混合的测序文库的去多路化

[0357] *T 具有硫代磷酸酯键的3'脱氧胸苷三磷酸"T尾"提供基质用于连接至"A-尾"DNA片段

[0358] 第36至40页上的序列是:

[0359] Y-衔接物版本1(按顺序):SEQ ID NO:4、8、9、7和3。

[0360] Y-衔接物版本2(按顺序):SEQ ID NO:2、6、9、5和1。

[0361] 发夹:SEQ ID NO:9

[0362] 实施例

[0363] 实施例1-DEDUCE-Seq

[0364] 通过无偏流动池富集进行的DuplEx测定和测序(DEDUCE-seq)使用全长Y-衔接物来构建Illumina测序所需的所有必要DNA元件,而第二发夹衔接物将锁定寡物并连接两条链,从而保留直链分子的双链体信息(图1)。与使用基于标签的系统保留双链体信息的Duplex-seq方法[Kennedy,S.R.,等人,Detecting ultralow-frequency mutations by Duplex Sequencing.Nat Protoc,2014.9(11):第2586-606页]一样,DEDUCE-seq利用DNA的互补性质来区分真正的突变和测序错误。但是,DEDUCE-seq通过将DNA双链体的两条链物理连接到单个可测序的DNA分子中来实现这一点。此外,当前测序仪的总体碱基调用准确度在过去十年中提高了至少一个数量级(达到$1/10^3$),从而改善了变异被调用的理论极限。通过使用来自两条DNA链的冗余信息,只有真正的突变或变异才会在双链体的两条链上被调用,而技术错误可以被消除,因为它们只存在于两个读取中的仅一个中(图1)。该策略已经被证实会大大提高超罕见突变(低至$1x10^{-6}$)的准确检测[Salk,J.J.和S.R.Kennedy,Next-Generation Genotoxicology:Using Modern Sequencing Technologies to Assess Somatic Mutagenesis and Cancer Risk.Environ Mol Mutagen,2020.61(1):第135-151页]。但是,Duplex-seq依赖于独特分子标识符(UMI)和PCR扩增来实现这一点[Kennedy,S.R.,等人,Detecting ultralow-frequency mutations by Duplex Sequencing.Nat Protoc,2014.9(11):第2586-606页]。因此,该方法并非无需PCR,这大大增加了测序成本,并限制了其在靶向测序中的应用。但是,DEDUCE-seq在设计上不使用PCR,而是使用

Illumina流动池进行DNA富集,并且被设计成能够有成本效益地检测全基因组范围内的罕见突变。

[0365] 首先,发明人将使用DEDUCE-seq来检测先前进行的同源酵母实验中的突变。在该项目中,对多个用紫外线照射处理的酵母菌株进行了突变调查,之后将细胞繁殖约1,200代以积累突变。使用传统的WGS和变异调用来测量在这些实验中得到的突变。因此,小尺寸酵母基因组的这种遗留数据允许DEDUCE-seq可以作为体内突变检测的基准。

[0366] 接下来,将DEDUCE-seq应用于检测通过INDUCE-seq发现的新脱靶位点处的突变(WO2022/038291A1)。CRISPR基因组编辑项目可用于检测人细胞中的突变。使用INDUCE-seq,发明人已经发现了非常严格的引导RNA和靶向性很差的RNA的新型脱靶,这允许DEDUCE-seq测定该范围的两端,以评价人细胞中高频率和低频率脱靶断裂位点的突变诱导。

[0367] 方法-初步设计

[0368] 发明人将利用在酵母中进行的突变调查所得到的基因组DNA(见上文)。为了建立文库制备,将基因组DNA片段化成约600-800bp的尺寸。第一次连接使用全长Y-衔接物来构建测序所需的所有衔接物组件(图1)。接下来,纯化DNA以除去多余的衔接物DNA,并进行第二轮片段化至约200-300bp。DNA尺寸选择、成功连接和衔接物DNA的除去都将使用毛细管DNA电泳进行评估。

[0369] 对于第二次连接,发明人将使用发夹衔接物以物理方式连接DNA的互补链,并将双链体信息锁定到单个可测序的分子中(见图1)。对该文库DNA的尺寸选择和纯化还将除去多余的发夹衔接物DNA。为了验证成功连接,将使用qPCR来量化样品的可测序DNA级分。这将揭示图1中所示的衔接物的设计产生了功能性的、可测序的分子。这些试点实验将首先进行小规模的双端测序。这将确认成功的连接策略并揭示文库的总体质量。为此,发明人将首先对未经处理的细胞(低突变率)和高度致突变的细胞(高突变率)的DNA进行测序。这些实验的初步数据将用于开始开发利用双链体信息的数据分析流程。接下来,试点DEDUCE-seq实验将被扩大到更多的样品和更高的覆盖率(约100倍),其中使用高容量测序平台(MiSeq v3或NextSeq 550)来检测来自(i)未处理的野生型细胞、(ii)紫外线照射的野生型细胞和(iii)具有已知突变表型的细胞的早期和晚期酵母中的突变。最后,在这些试点实验之后,发明人将应用DEDUCE-seq来检测来自上述先前进行的诱变项目的大量酵母样品中的突变。现在可以与以约10-25倍覆盖率执行的原始WGS相比,使用由此产生的数据来评估DEDUCE-seq的性能。一旦建立,所述方法将用于检测来自CRISPR-Cas9编辑的基因组DNA的全基因组突变,并通过INDUCE-seq测量低频率和高频率脱靶。

[0370] 方法-针对试点1和试点2

[0371] 基因组DNA输入

[0372] 为了产生DEDUCE-seq文库,使用片段化的基因组酵母DNA作为输入。将基因组DNA样品解冻并在自动电泳系统(Agilent TapeStation 2100,高灵敏度D1000 screentape)上运行以评估尺寸分布和质量。接下来,使用高灵敏度试剂盒(Qubit™dsDNA HS Assay Kit)使用Qubit-2(ThermoFisher)对DNA进行定量,并在无核酸酶水(NFW)中归一化至200-250ng/50μL。

[0373] DEDUCE-seq文库制备

[0374] 使用单侧尺寸选择法制备基因组DNA。首先,用0.6×(v/v)SPRI珠子(CleanNGS,GCBiotech)除去大于300bp的片段,从而将溶液中的目的DNA保持在100到500bp。在第二个纯化步骤中,添加SPRI珠子至1.8×(v/v)的最终浓度,并将DNA洗脱至25μL NFW的最终体积。接下来,使用NEBNext®Ultra™II末端修复/dA-加尾模块(E7546L,New England Biolabs)将DNA平端化并加A尾,最终体积为30μL,准备好使用NEBNext®Ultra™II连接模块(E7595L,New England Biolabs)进行连接。对于第一次连接,试点-1使用1.25μL 7.5μM的全长Y-衔接物(P5-P7),而试点-2使用1.25μL 7.5μM的发夹衔接物。纯化总DNA,并使用1.8×(v/v)SPRI珠子除去剩余的衔接物,此后在100μL NFW中洗脱DNA,准备好进行超声处理。使用Bioruptor(Diagenode)对连接的DNA进行重新超声处理60个循环(30秒开/关,高输出)。为了准备第二轮末端制备和连接,使用1.8×(v/v)SPRI珠子纯化DNA,并在25μL NFW中洗脱以减少适合NEBNext Ultra II模块的体积。使用NEBNext®Ultra™II末端修复/dA-加尾模块(E7546L,New England Biolabs)将重新超声处理的DNA平端化并加A尾,然后使用NEBNext®Ultra™II连接模块(E7595L,New England Biolabs)进行连接,如上所述。在试点-1中使用1.25μL7.5μM发夹衔接物,在试点-2中,在第二次连接中使用1.25μL 7.5μM的全长Y-衔接物(P5-P7)。在第二次和最终连接以后,使用1.8×(v/v)SPRI珠子纯化DNA,并在28μL NFW中洗脱。使用qPCR对最终文库进行定量,并在自动电泳系统(Agilent TapeStation 2100、高灵敏度D1000 screentape)上进行试验,以评估尺寸分布和质量。在整个方案过程中,使用电泳测量文库DNA的尺寸和质量,以评估衔接物除去、重新超声处理和最终文库。

[0375] DEDUCE-seq文库定量

[0376] 将最终DEDUCE-seq文库DNA在稀释缓冲液(10mM Tris-HCl,pH 8.0,0.05%吐温-20)中稀释50倍,并使用20μL的最终PCR反应体积对4μL稀释的文库DNA进行一式三份qPCR(KAPA文库定量试剂盒Illumina®平台)。按照供应商指南的建议,使用循环方案扩增文库DNA,并使用提供的工具进行量化,以得到未稀释的文库浓度(μM)。

[0377] DEDUCE-seq文库的测序

[0378] 将最终的DNA文库汇集到相关位置,并使用SpeedVac将最终体积减少至40μL。在加载到测序流动池上之前,根据以下改进的变性方案制备DEDUCE-seq文库:将最终文库(40μL)与40μL新鲜稀释的0.2N NaOH在室温混合5分钟以使DNA变性。接下来,使用40μL的200mM Tris-HCl(pH7)中和溶液。所得的变性文库(120μL)与1179μL预冷的HT1和1μL变性并稀释的PhiX对照(20pM)互补。将1.3mL的该混合物加载到NextSeq筒上进行完整测序。

[0379] 测序数据处理

[0380] 使用Illumina的在线基础空间实用程序或离线序列分析查看器(SAV,Illumina)评估测序运行。读取通过过滤器、碱基调用质量(Q30)和簇密度被用作第一次通过质量控制。然后检索去多路化的数据以准备进行下游分析,如下所述。

[0381] 二次数据分析

[0382] 从基础空间下载去多路化的测序数据作为FASTQ文件。使用trim_galore(v0.6.7),对读取进行质量调整,并使用标准参数对衔接物进行修剪。使用FASTQC对修剪后的和未修剪的数据进行质量检查。为了检索包含读取的HP,使用标准命令行工具GNU grep(3.7)和AWK(1.3.4)来查询数据。

[0383] 首先使用Seqkit fq2fa将FASTQ文件转换为FASTA格式,然后应用定位来计算读取1和2中发夹序列的确切位置,其中使用以下命令:

[0384] seqkit fq2fa-j$threads$Read1-o$Read1.fa.gz

[0385] seqkit locate-j$threads-i-d-P-p AGGGCCTANNNNNNNNTAGGCCC$Read1.fa.gz>$Reads1_HP-locate.tsv

[0386] 使用bowtie2(2.5.1)进行DEDUCE-seq数据的比对,其中使用默认参数进行探索性分析,比对一致读取对和不一致DEDUCE-seq读取,方式如下:

[0387] #默认一致比对

[0388] bowtie2-p$threads-x$refseq-1$mate1 -2$mate2

[0389] #DEDUCE-seq不一致比对

[0390] bowtie2-p$threads--ff--no-mix--dovetail-x$refseq-1$mate1 -2$mate2

[0391] 使用samtools(1.6)删除相关的未映射读取和二次比对:

[0392] samtools view-Shu-f 3-F 256-@$threads$input

[0393] 使用samtools(1.6)将比对的数据转换为BAM文件,并使用Intergrated Genomics Viewer(2.14.1)(Robinson等人.2011)进行可视化。

[0394] 实施例2-DEDUCE-Seq试点1和试点2

[0395] 此处描述的试点研究旨在建立DEDUCE-seq文库的核心元件并确定最有效的连接策略。因此,我们产生了DEDUCE-seq文库,其中首先连接发夹并随后连接Y-衔接物(试点-1),反之亦然(试点-2)。在这些研究中,基因组酵母DNA被用于产生DEDUCE-seq文库。该DNA先前曾在一项旨在检测同源酵母菌株中紫外光照射诱导的突变的研究中被用于测量突变(Nandi等人.2018),并提供了已知来源和已知突变负担的合适基因组DNA来源。这些样品以约200-300bp的片段化DNA形式储存,并归一化为4-5ng/μL。首先,对于试点-1,将2个样品平行处理并进行右侧尺寸选择以除去>300bp的较大DNA片段(显示了一个代表性样品的数据)。起始DNA范围从100至500bp(图5,左图)。

[0396] 试点1-首先连接DEDUCE-seq发夹并随后连接Y-衔接物-文库构建

[0397] 接下来,将DNA钝化,加A-尾,并使用NEBNext Ultra II试剂盒连接发夹衔接物。对连接的DNA进行纯化,并在TapeStation上进行检查以确认发夹衔接物DNA的除去(图5,中图,黑色迹线)。对DNA的重新超声处理会导致尺寸分布的偏移,以约200bp为中心,范围从75至500bp(图5,中图,灰色迹线)。对第二个Y-衔接物重复末端制备和连接过程,得到图5所示的最终纯化的文库(右图,灰色迹线)。最后的连接不会导致尺寸分布的重大偏移。残留的Y-衔接物在50bp处即可检测到,并且在Y-衔接物连接后在900bp左右检测到高分子量片段(图5,右图)。重要的是,这是已知的Y-衔接物连接伪像,在先前使用粗短Y-衔接物作为Duplex-seq方法论手稿的部分进行展示(Kennedy等人.2014)。该大级分可以安全地忽略。

[0398] 最终的文库含有分子的混合物,其中一部分由功能性DEDUCE-seq HP-Y-衔接物连接的片段组成。由于全长P5-P7杂合Y-衔接物的存在,DEDUCE-seq文库分子含有定量所需的组分引物结合位点。因此,我们应用qPCR来量化最终文库制备中的可测序分子,并分别测量了样品1和2的1.1和2.0nM文库浓度。重要的是,图5(右图)中显示的高分子量伪像对qPCR读数没有贡献。在这些样品中检测不到具有过高熔化温度的分子(数据未显示)。在这些浓度,预计样品1和样品2的最终文库每1μL未稀释的文库分别含有1.9亿到3.55亿个可测序分子。

因此,对样品2的1μL文库在NextSeq500高输出2x150bp流动池上进行测序,得到了2.4亿个读数,其中96%通过了过滤,Q30得分为93%。成功的测序表明,试点-1的HP-Y-衔接物连接策略产生了功能性的、可测序的分子,所述分子可以通过qPCR进行量化并加载到测序仪上以产生与qPCR读数成比例的簇和读取。

[0399] 试点1-首先连接DEDUCE-seq发夹并然后连接Y-衔接物-数据分析

[0400] DEDUCE-seq文库的设计是非标准的,并且预计会导致在正向-正向(F1F2)或反向-反向(R2R1)方向的不一致读取对,并非所有比对器都接受其为合法输出。类似地,根据插入片段长度和修剪,从该文库可以产生燕尾读取,并且并非被所有对齐器接受。此外,我们从初步比对实验中发现,大多数读取被比对为一致对(F1R2或R1F2),并推测这些来自双Y-衔接物连接产物,其被预测为少数DEDUCE-seq输出(数据未显示)。因此,我们首先通过以无偏方式搜索包含发夹的读取来评估DEDUCE-seq测序读取的组成。以编程方式从读取1和读取2中检索包含发夹的读取,分别从总共237M读取中返回33和44M读取(14-18.4%)(表2)。如果基因组DNA插入片段尺寸大于150bp,则预计并非每个双链体分子都会在读取中包含HP序列。有趣的是,从这次搜索中得到的基本位置信息显示,大多数读取在读取开头包含HP序列,而约7.5M读取在读取中间某处包含HP序列。但是,我们很少在读取的最末端发现HP序列(表2)。

[0401] 表2

[0402]

	读取 1	读取 2
没有发夹的读取	204,467,929	193,705,101
有发夹的读取	33,297,970	44,060,798
在读取开始	25,611,951	36,501,713
在读取中间	7,648,190	7,523,125
在读取末端	37,829	35,960
总计	**237,765,899**	**237,765,899**

[0403] 利用该信息,我们提取了所有包含R1和R2中的HP序列的读取对,以阐明这些分子的构象。使用这个9.5M读取对的列表,我们计算了每个读取中发夹序列的精确位置,并绘制了发夹位置作为读取长度(<151bp)的函数的分布。

[0404] 我们发现,对于这类读取,大约一半在读取的开始处包含发夹DNA,如表2所示。在剩余的读取中,HP序列均匀分布在读取长度上。接下来,我们收集了包含HP序列的9.5M读取对(4%),并使用bowtie2将它们与参考基因组比对。Bowtie2可以容纳不一致的读取对,从而使我们能够比对约200-800个读取并在基因组浏览器中检查它们(图7)。这揭示了所有这些读取的正确双平行方向(F1F2和R1R2),正如从DEDUCE-seq文库设计所预期的。这证实了包含HP的读取可以使用bowtie2进行比对并且处于正确的方向,从而进一步揭示了定义这些读取对的相关SAM标志(F1F2,SAM标志67-133和R2R1,SAM标志115-179)。

[0405] 利用该信息,我们使用上面描述的bowtie2的特定配置比对了230M读取的整个数据集,并使用相应的SAM标志过滤掉正确比对的DEDUCE-seq读取。这产生了具有预期的DEDUCE-seq-特异性方向的约120K读取(图7和表3),从而表明发夹和Y-衔接物连接的组合

产生了一个功能文库。表4总结了此比对的结果。

[0406] 表3:DEDUCE-seq文库不一致读取对

	左对齐	右对齐	左对齐	右对齐
旗帜	67	131	115	179
映射质量	40	40	42	42
CIGAR	92M	133M	79M	79M
映射配偶	是	是	是	是
位置	对中的第一个	对中的第二个	对中的第一个	对中的第二个
对方向	F1F2	F1F2	R2R1	R1R2

[0407]

[0408] 表4

	读取		旗帜	描述
读取 1	197,911,359	197,8520,805	77	配对|未映射|未映射配偶|第 1 个
读取 2	197,911,359	197,8520,805	141	配对|未映射|未映射配偶|第 2 个
与自身适当地配对并且	181,108	38,869	67	配对|正向|在适当对中映射|第 1 个
映射配偶				
		38,869	131	配对|正向|在适当对中映射|第 2 个
		43,857	115	配对|反向|在适当对中映射|第 1 个
		43,857	179	配对|反向|在适当对中映射|第 2 个
		3,555	65	配对|正向|第 1 个
		3,555	129	配对|正向|第 2 个
		4,273	113	配对|反向|第 1 个
		4,273	177	配对|反向|第 2 个

[0409]

[0410]

[0411] 以这种方式通过bowtie2强制比对DEDUCE-seq数据会导致>190M读取对未映射,这通常(基于SAM标志)会导致一致的对(数据未显示)。相反,适当地配对的和映射的读取落入正确类别的不一致平行读取(F1F2和R2R1),总计180K(表2)。

[0412] 试点-2DEDUCE-seq首先连接Y-衔接物、再连接发夹:文库构建

[0413] 与试点-1类似,试点-2的DEDUCE-seq文库衍生自相同的基因组DNA。在本例中,对来自4个独立样品的总共250ng DNA进行尺寸选择以除去DNA的大片段(>500bp)并准备用于连接。在第一轮中,将全长Y-衔接物连接到DNA上。在纯化并除去未连接的Y-衔接物以后,将DNA重新超声处理60个循环并纯化。将DNA通过另一轮末端制备和连接进行处理以连接发夹衔接物,此后使用qPCR纯化和定量DNA。这些样品的最终文库浓度是在2.8至8.3pM范围内。

[0414] 因此,通过先连接Y-衔接物、再连接发夹衔接物来制备DEDUCE-seq文库会导致可测序分子的产量比在试点-1中执行的反向顺序低约3个数量级。这表明,Y-或发夹衔接物之间的连接效率是不同的,并且连接顺序会影响DEDUCE-seq文库的产量。在试点-2中制备的样品的估计测序读取是在11至35M范围内。因此,我们将4个样品合并在一起,得到总共73M

预测读取,并在NextSeq 500高输出2x150bp流动池上对该池进行测序。此次测序运行产生了3800万个读取,其中91%通过了过滤,Q30得分为91%。重要的是,这里产生的DEDUCE-seq文库的测序输出低于从qPCR定量的预期。以这种方式构建的DEDUCE-seq文库的原始设计有利于流动池富集正确连接的Y-DNA-发夹产物,同时将未连接的片段锁定在双发夹连接环中,从而使这些分子变得惰性。从理论上讲,这将改善在流动池上的可测序分子的选择性富集。

[0415] 试点-2DEDUCE-seq首先连接Y-衔接物、再连接发夹:数据分析

[0416] 基于来自试点-1的结果,我们使用上面描述的相同方案,并整理包含发夹的读取来计算每个读取中发夹序列的位置。

[0417] 表5

[0418]

	读取 1	读取 2
没有发夹的读取	38,737,660	38,750,812
有发夹的读取	474,893	422,286
在读取开始	64	70
在读取中间	466,201	414,949
在读取末端	8,628	7,267
总计	**38,856,384**	**38,856,384**

[0419] 总的来说,这从总共38M读取对返回了含有发夹序列的R1和R2的420-470K读取(1.1-1.2%)。组合其中两个读取都包含HP序列的对,留下200K读取对(约0.5%),其中23K与参考基因组对齐(约0.05%)。发夹位置的分布如图9所示。其次连接发夹衔接物会导致大多数HP序列按预期定位到读取1或读取2的3'或末端。

[0420] 重要的是,在读取1和2中存在非编码发夹序列的情况下进行此比对,而该序列在酵母参考基因组中不存在,因此不会干扰比对器。从这些读取中修剪发夹DNA会改善比对(数据未显示)。

[0421] 结论

[0422] 综合考虑,DEDUCE-seq的Y-和HP-衔接物连接的两个方向,产生符合DEDUCE-seq设计的平行双链体分子。

[0423] 像在试点-2中那样,首先连接Y-衔接物、再连接发夹可能是充分利用流动池从惰性双发夹分子中富集正确形成的Y-HP分子的优选选项。但是,与在试点-1中应用的策略相比,该文库策略效率较低。在试点-1中,文库的总产量(nM)高于试点-2(pM)。

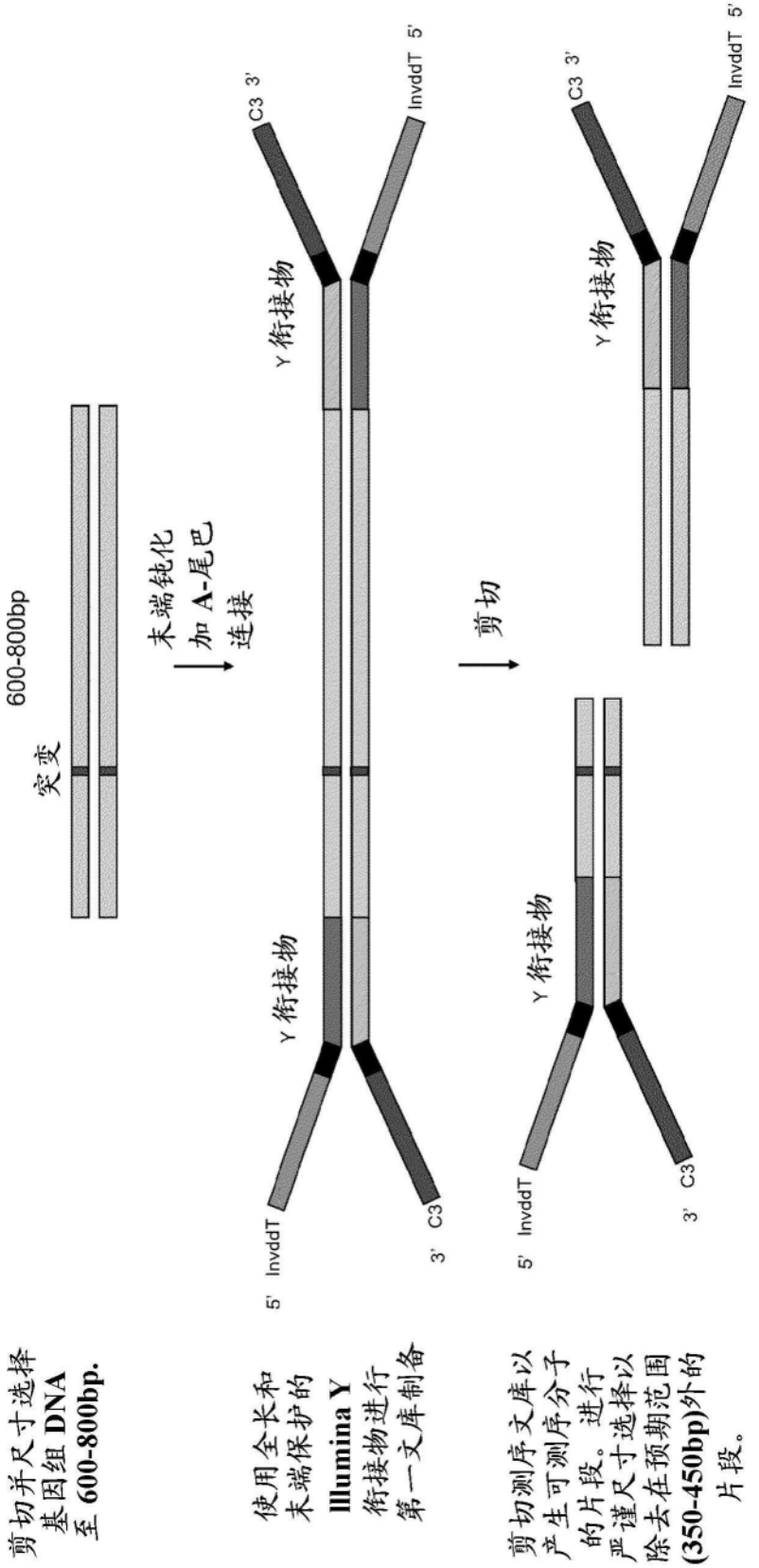

剪切并尺寸选择
基因组 DNA
至 600-800bp.
600-800bp
突变
末端钝化
加 A-尾巴
连接
Y 衔接物
C3 3'
InvddT 5'
使用全长和
末端保护的
Illumina Y
衔接物进行
第一文库制备
5' InvddT
3' C3
剪切
剪切测序文库以
产生可测序分子
的片段。进行
严谨尺寸选择以
除去在预期范围
(350-450bp)外的
片段。

图1

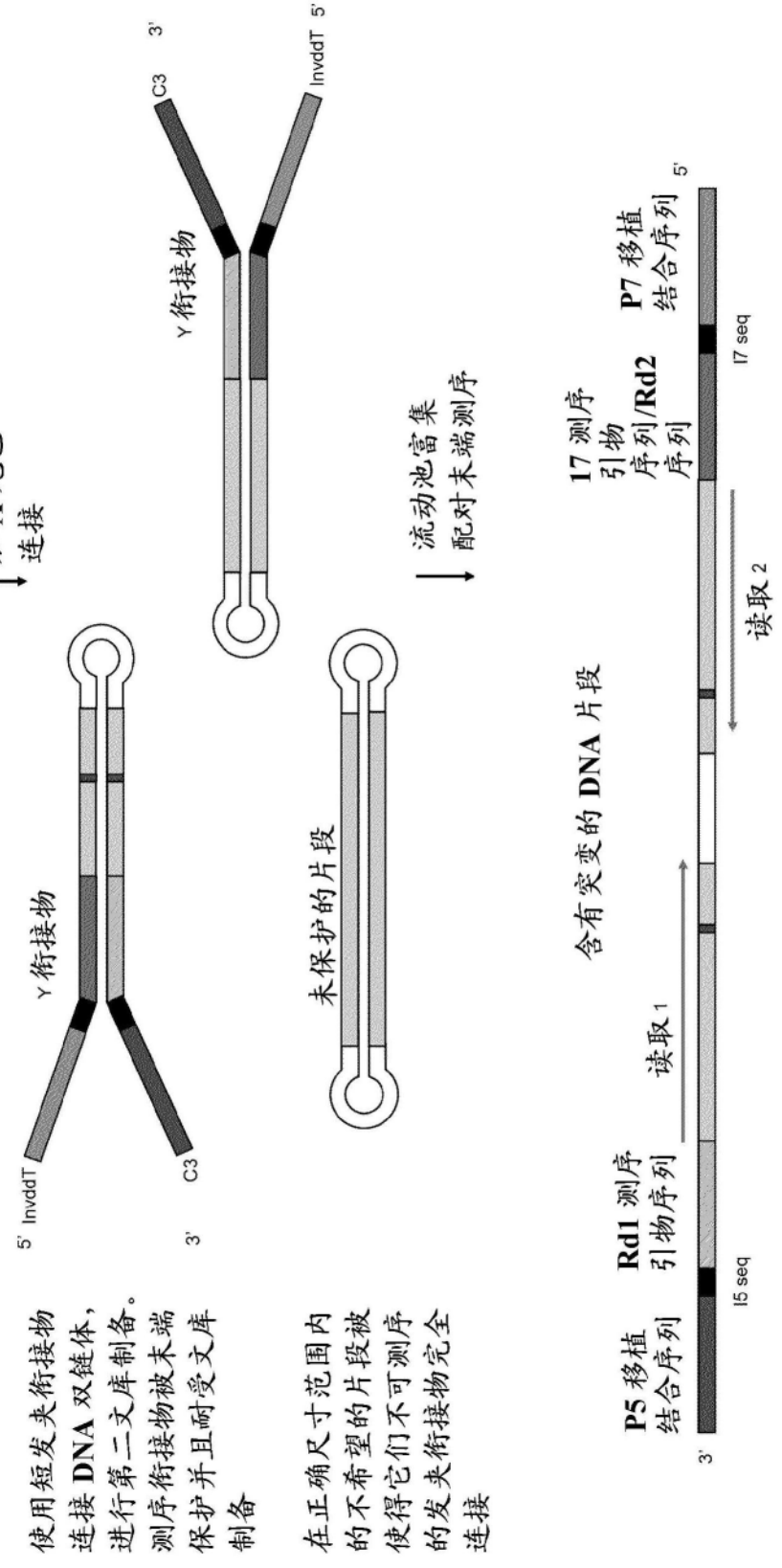

末端钝化
加 A-尾巴
连接
Y 衔接物
5' InvddT
3' C3
使用短发夹衔接物
连接 DNA 双链体，
进行第二文库制备。
测序衔接物被末端
保护并且耐受文库
制备
未保护的片段
在正确尺寸范围内
的不希望的片段被
使得它们不可测序
的发夹衔接物完全
连接
流动池富集
配对末端测序
含有突变的 DNA 片段
P5 移植
结合序列
Rd1 测序
引物序列
读取 1
I5 seq
17 测序
引物
序列/Rd2
序列
P7 移植
结合序列
读取 2
I7 seq

图1(续)

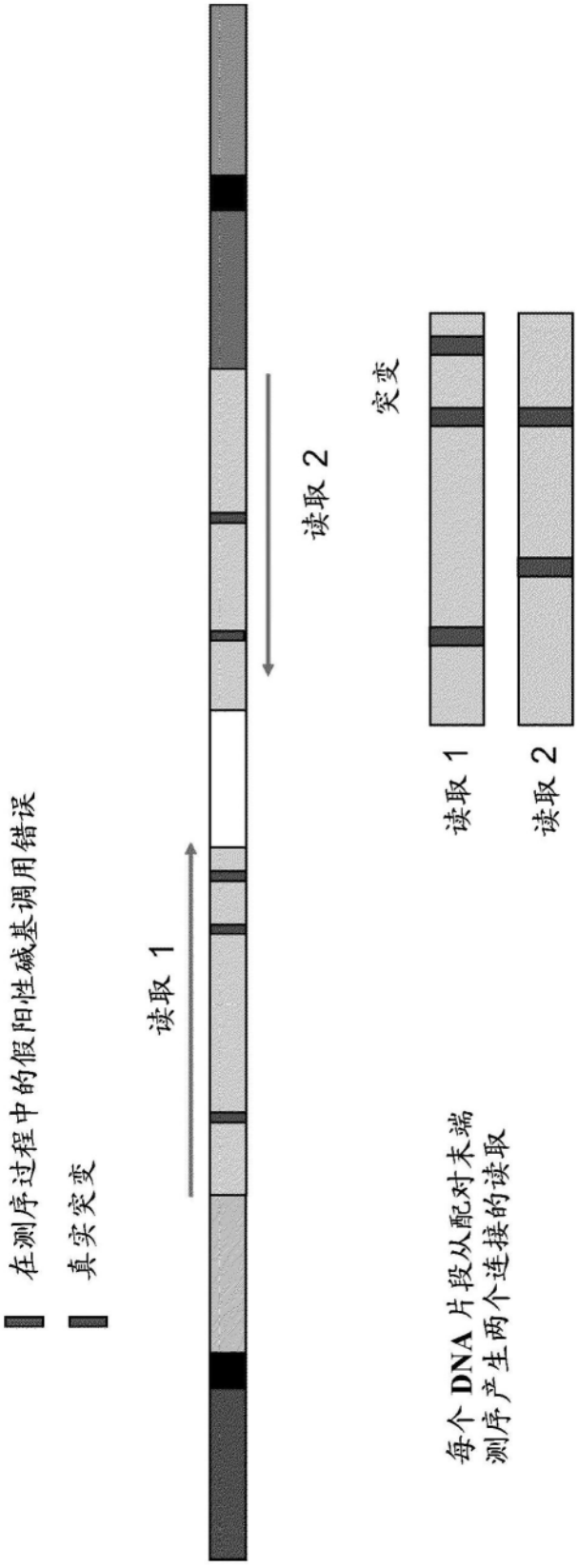

在测序过程中的假阳性碱基调用错误
真实突变
读取 1
读取 2
突变
读取 1
读取 2
每个 DNA 片段从配对末端
测序产生两个连接的读取

图2

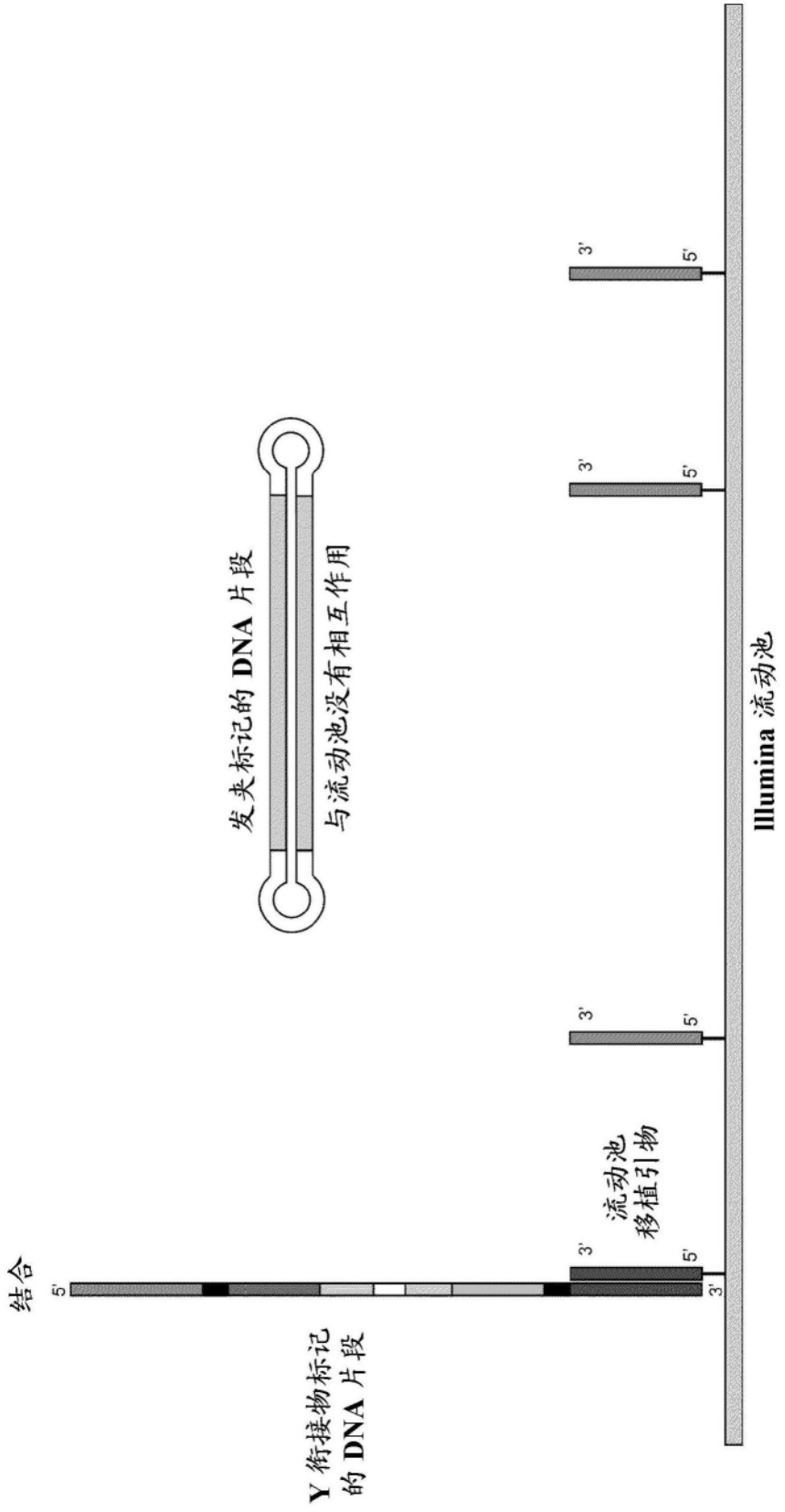
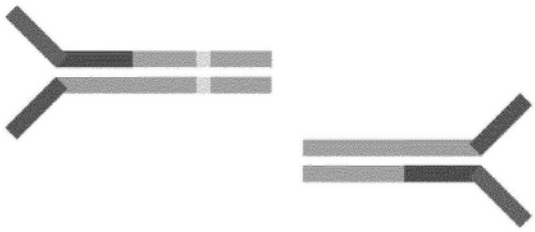

结合
5'
3'
Y 衔接物标记
的 DNA 片段
流动池
移植引物
发夹标记的 DNA 片段
与流动池没有相互作用
Illumina 流动池

图3

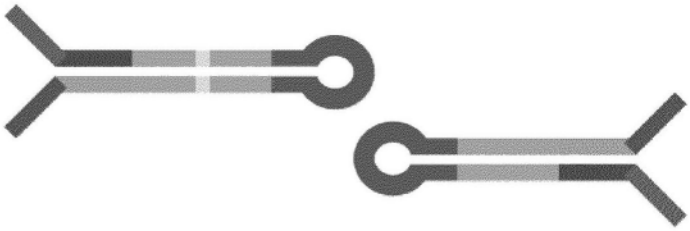
1)在 DNA 的两条链中存在的突变。



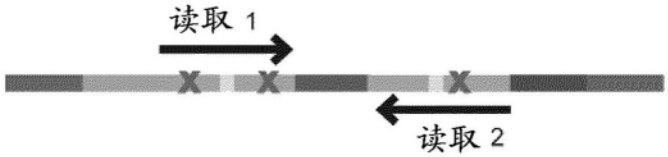
2)第一个无 PCR 文库制备物标记所有用于测序的 DNA 链。
通过片段化来暴露末端。



3)通过发夹接头的连接而物理连接的 DNA 双链体。



4)通过 Illumina 测序来读取连接的双链体片段。
读取 1 和读取 2 揭示了相同 DNA 双链体的两条链的序列。



5)突变存在于两个读取中的相同位置。
测序错误可辨别，因为它们仅发生在一个读取中。

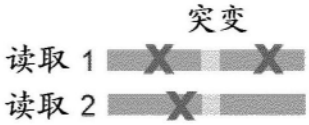


图4

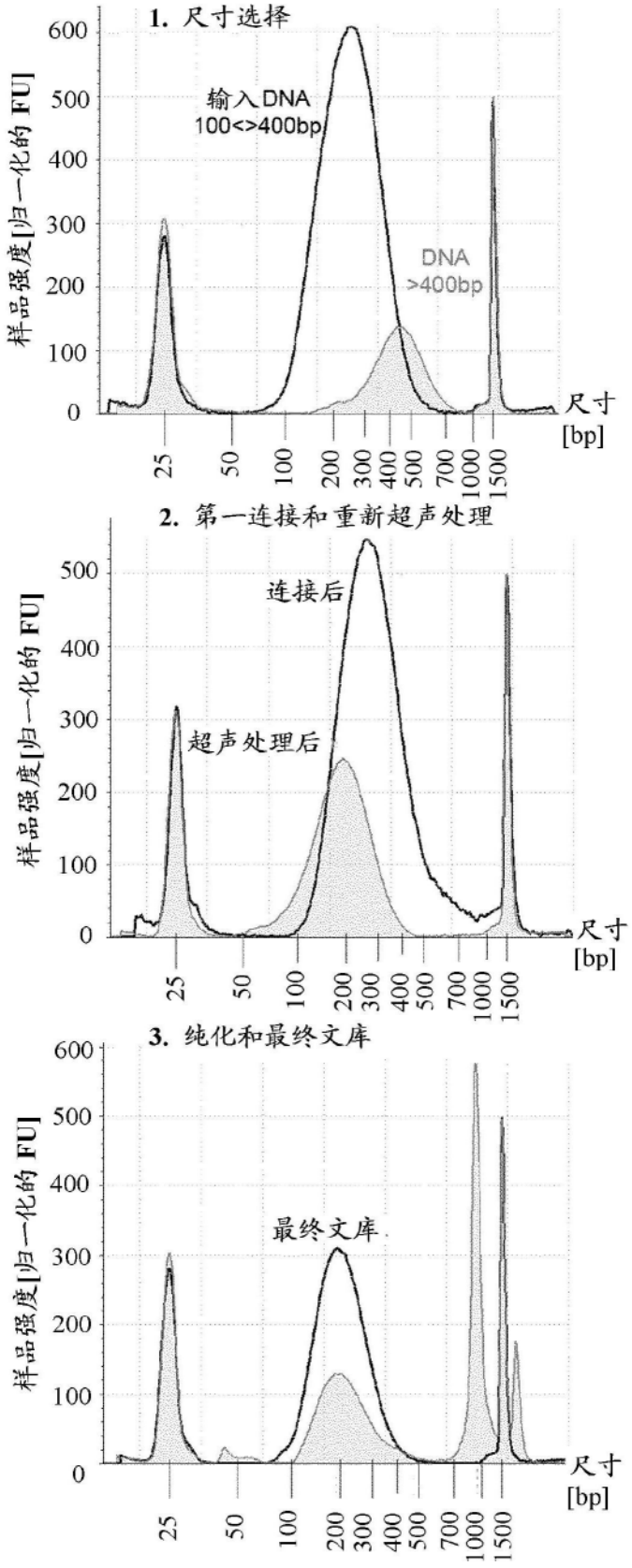

1. 尺寸选择
输入DNA
100<>400bp
DNA
>400bp
样品强度[归一化的 FU]
尺寸
[bp]
2. 第一连接和重新超声处理
连接后
超声处理后
3. 纯化和最终文库
最终文库

图5

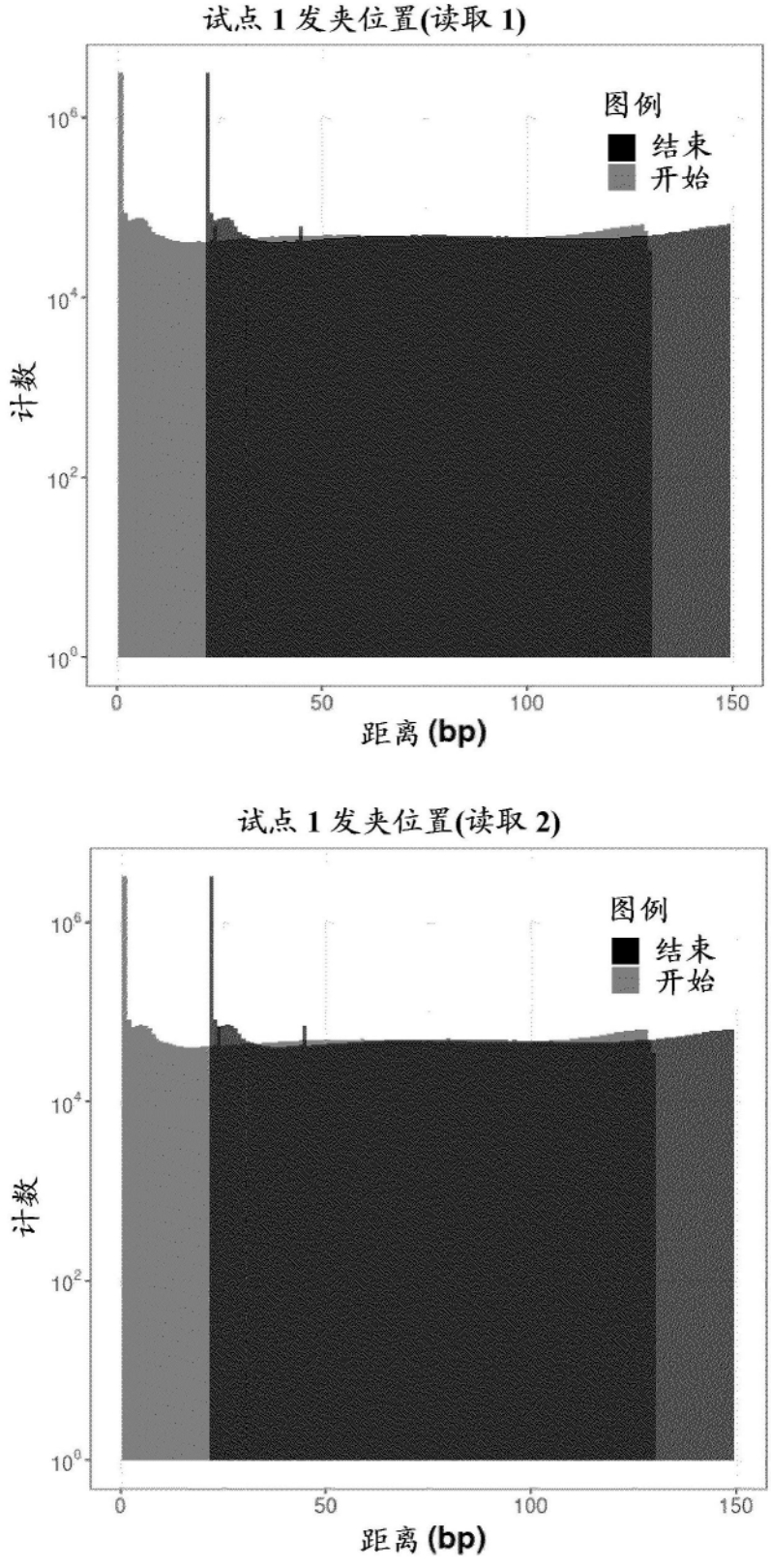

试点 1 发夹位置(读取 1)
图例
结束
开始
计数
10^6
10^4
10^2
10^0
0
50
100
150
距离 (bp)
试点 1 发夹位置(读取 2)
图例
结束
开始
计数
10^6
10^4
10^2
10^0
0
50
100
150
距离 (bp)

图6

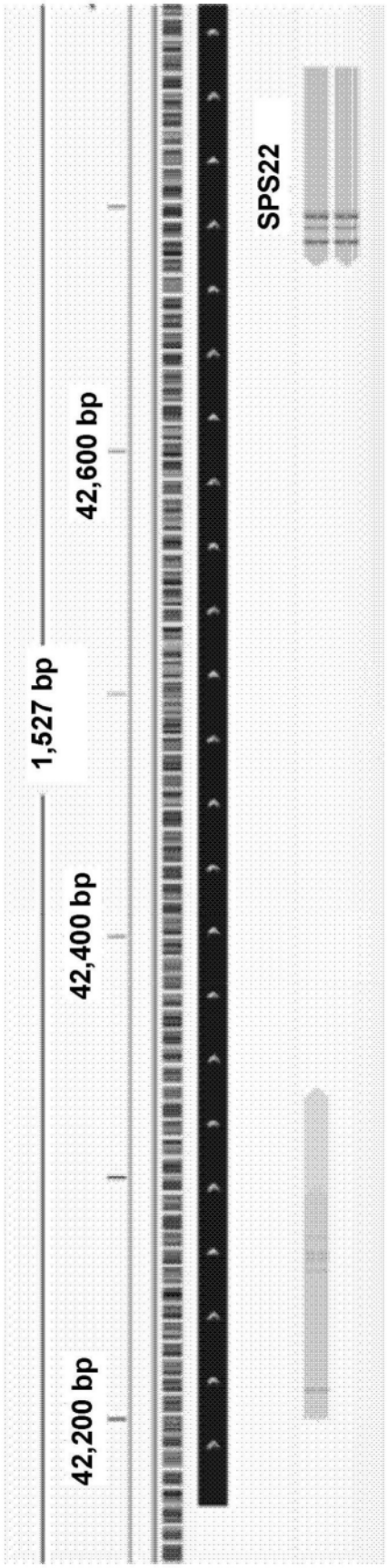


图7

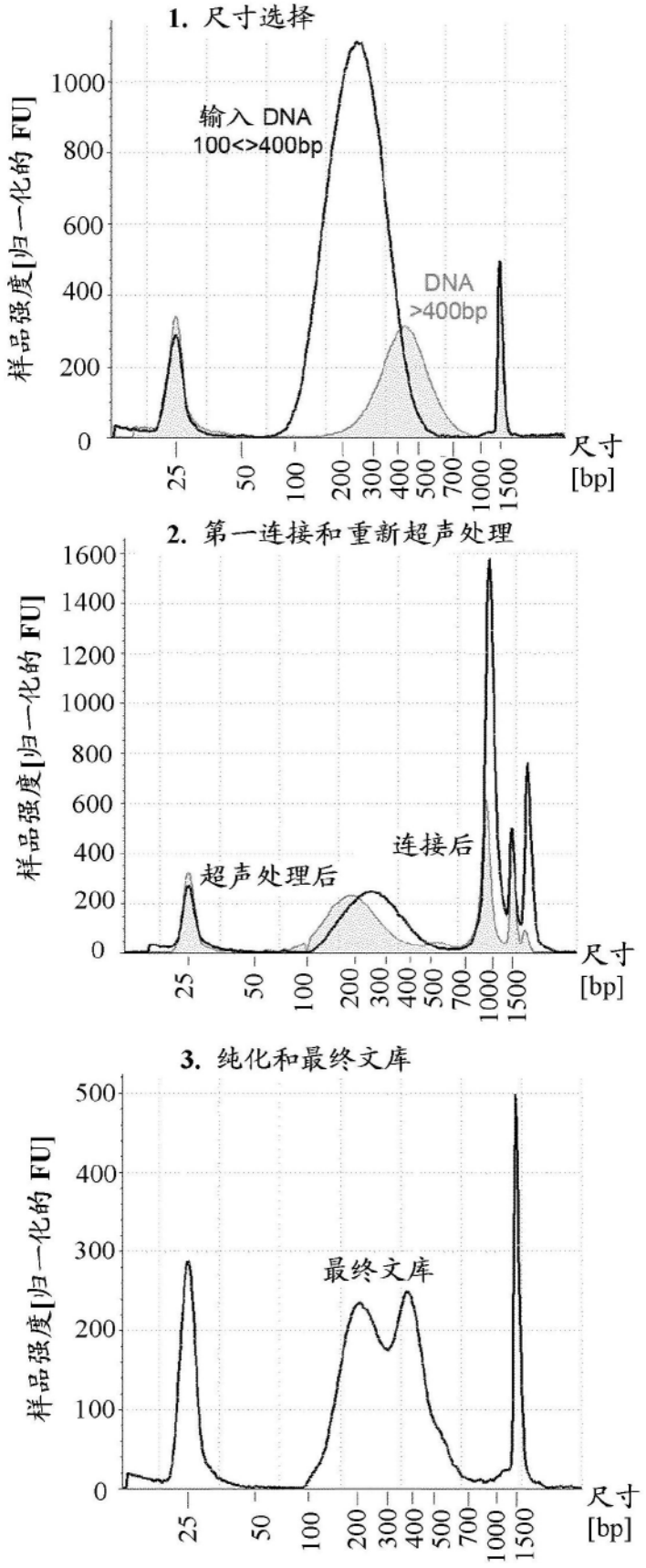

1. 尺寸选择
样品强度[归一化的 FU]
输入 DNA
100<>400bp
DNA
>400bp
尺寸
[bp]
2. 第一连接和重新超声处理
超声处理后
连接后
3. 纯化和最终文库
最终文库

图8

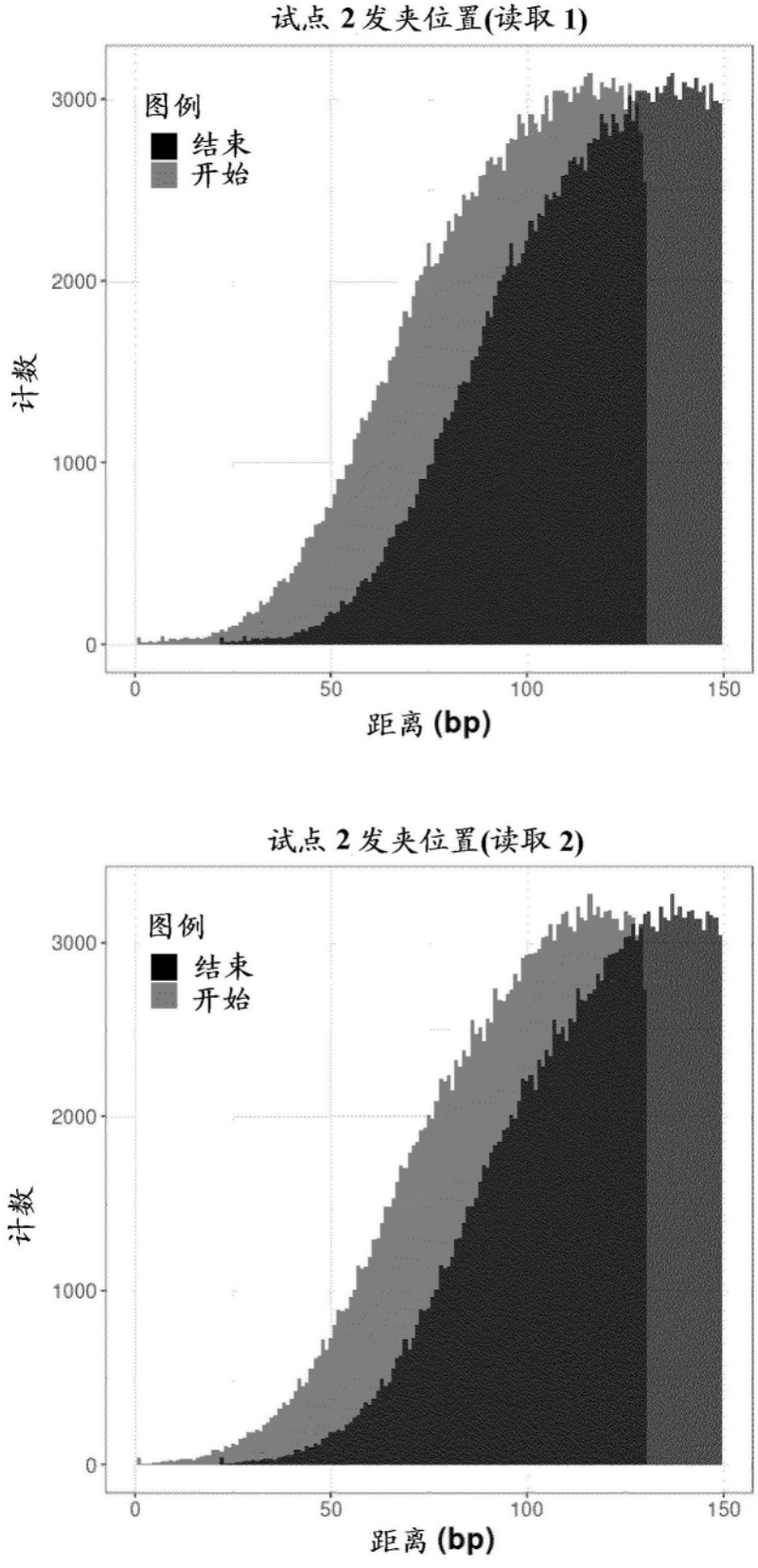

试点 2 发夹位置(读取 1)
图例
结束
开始
计数
3000
2000
1000
0
0
50
100
150
距离 (bp)
试点 2 发夹位置(读取 2)
图例
结束
开始
计数
3000
2000
1000
0
0
50
100
150
距离 (bp)

图9