(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114286861 B
(45)授权公告日 2024.12.06

(21)申请号 202080051127.X
(22)申请日 2020.05.13
(65)同一申请的已公布的文献号
申请公布号 CN 114286861 A
(43)申请公布日 2022.04.05
(30)优先权数据
62/846,988 2019.05.13 US
(85)PCT国际申请进入国家阶段日
2022.01.13
(86)PCT国际申请的申请数据
PCT/US2020/032615 2020.05.13
(87)PCT国际申请的公布数据
WO2020/232081 EN 2020.11.19
(73)专利权人 LGC基因组学有限责任公司
地址 美国明尼苏达州
(72)发明人 莱昂德罗·戈米德·内维斯
(74)专利代理机构 北京德崇智捷知识产权代理有限公司 11467
专利代理师 贺征华
(51)Int.Cl.
C12N 15/00(2006.01)
C12N 15/09(2006.01)
C12N 15/10(2006.01)
C12N 15/11(2006.01)
(56)对比文件
CN 106661631 A,2017.05.10
CN 114207229 A,2022.03.18
Lulu Xu等.Recent advances in rolling circle amplification-based biosensing strategies-A review.《Analytica Chimica Acta》.2021,全文.
审查员 吴静
权利要求书3页 说明书15页
序列表3页 附图9页

(54)发明名称

捕获和分析靶基因组区域

(57)摘要

本公开内容涉及用于捕获靶基因组区域的材料和方法,包括使用内切酶在具有最少约10至约30个核苷酸的特定识别位点处裂解靶基因组区域,并通过与桥接寡核苷酸杂交来捕获靶基因组区域。本公开还涉及分析捕获的靶基因组区域。用于裂解的内切酶可以是特异性结合识别位点以引导裂解的可编程内切酶。可以优选地通过聚合酶链反应或滚环扩增来扩增捕获的靶基因组区域,并进行检测或测序。此外,本发明涉及用于实施本发明方法的试剂盒,其包含一种或多种设计用于裂解一个或多个靶基因组区域的内切酶和设计用于捕获一个或多个靶基因组区域的桥接寡。所述试剂盒可以针对特定的靶基因组区域进行定制。

CN 114286861 B

1.一种从遗传物质中捕获靶基因组区域的方法,所述方法包括以下步骤:

a)使用一种或多种具有第一识别位点和第二识别位点的内切酶从遗传物质裂解靶基因组区域,每个识别位点包含10至80个核苷酸的序列,其中所述识别位点位于靶基因组区域的两侧,

b)将裂解的遗传物质变性成单链形式,和

c)通过将所述靶基因组区域与桥接寡杂交,以单链形式捕获靶基因组区域,所述桥接寡包含在3'和5'端分别与单链形式的靶基因组区域的3'和5'端杂交的序列。

2.根据权利要求1所述的方法,其中所述第一识别位点和所述第二识别位点中的每一个包含10个至30个核苷酸的序列,并且所述第一和第二识别位点位于靶基因组区域的两侧。

3.根据权利要求1或2所述的方法,其中在特定识别位点裂解遗传物质的一种或多种内切酶是限制性内切酶或可编程内切酶。

4.根据权利要求3所述的方法,其中所述一种或多种限制性内切酶是大范围核酸酶。

5.根据权利要求3所述的方法,其中所述一种或多种可编程内切酶选自成簇规则间隔短回文重复序列(CRISPR)相关蛋白9内切酶(Cas9内切酶)、Cpf1、C2c1、C2c2、C2c、RNA或DNA引导的Argonaute蛋白质和结构引导的内切酶。

6.根据权利要求5所述的方法,其中所述可编程内切酶包括基于具有与所述第一识别位点互补的序列的第一引导分子在所述第一识别位点切割DNA的第一内切酶和基于具有与所述第二识别位点互补的序列的第二引导分子在所述第二识别位点切割DNA的第二可编程内切酶。

7.根据权利要求1所述的方法,其中所述桥接寡是单链寡核苷酸,其包含在3'和5'端分别与单链靶基因组区域的3'和5'端的序列杂交的序列。

8.根据权利要求1所述的方法,其中所述桥接寡是具有3'和5'端突出端的双链寡核苷酸,所述3'和5'端突出端分别与单链靶基因组区域的3'和5'端的序列杂交。

9.根据权利要求7所述的方法,其中所述桥接寡进一步包含对稀有切割限制性内切酶特异的限制位点、用于可编程性内切酶的裂解位点和/或引物结合序列。

10.根据权利要求7所述的方法,其中所述桥接寡被固定在固体底物上或被生物素化。

11.根据权利要求1所述的方法,其中第一和第二识别位点存在于遗传物质的同一条链上。

12.根据权利要求1所述的方法,其中第一和第二识别位点存在于遗传物质的不同链上。

13.根据权利要求1所述的方法,其中将裂解的遗传物质变性成单链形式包括使裂解的遗传物质经受75℃至115℃的温度。

14.根据权利要求1所述的方法,所述方法还包括分析所述靶基因组区域,包括:

d)连接与桥接寡杂交的单链靶基因组区域的游离端,以产生与所述桥接寡杂交的单链环状靶基因组区域,

e)可选地,降解未环化遗传物质,

f)可选地,通过核酸扩增来扩增所述靶基因组区域,以产生所述靶基因组区域的多个拷贝,和

g)分析扩增的靶基因组区域。

15.根据权利要求14所述的方法,包括通过外切酶的组合降解未环化遗传物质,所述外切酶从核酸暴露的3'或5'端降解核酸,从而仅留下与桥接寡杂交的环化靶基因组区域。

16.根据权利要求14所述的方法,包括通过聚合酶链反应(PCR)扩增所述靶基因组区域,以产生所述靶基因组区域的多个拷贝。

17.根据权利要求16所述的方法,其中所述PCR引物可以对目标区域特异,或者通过将它们设计成与桥接寡引入的共同区域结合而对所有目标通用。

18.根据权利要求14所述的方法,包括通过滚环扩增(RCA)反应扩增所述靶基因组区域,以产生所述靶基因组区域的多个连接拷贝。

19.根据权利要求18所述的方法,其中所述桥接寡用作所述RCA反应的引物,以产生单链形式的靶基因组区域的多个连接拷贝。

20.根据权利要求18所述的方法,其中所述桥接寡和/或一种或多种额外引物用于RCA反应,以产生双链形式的所述靶基因组区域的多个连接拷贝。

21.根据权利要求14所述的方法,其中所述分析包括检测所述靶基因组区域。

22.根据权利要求21所述的方法,其中检测所述靶基因组区域包括使所述靶基因组区域与特异性结合靶基因组区域的标记探针或结合扩增的目标分子的物质接触。

23.根据权利要求14所述的方法,其中所述分析包括对所述靶基因组区域进行测序。

24.根据权利要求23所述的方法,其中所述测序包括纳米孔测序、可逆染料终止子测序或单分子实时(SMRT)测序。

25.一种从遗传物质中捕获多个靶基因组区域的方法,包括:

a)使用多对内切酶从遗传物质裂解多个靶基因组区域,每对内切酶具有第一识别位点和第二识别位点,每个识别位点包含17到24个核苷酸的序列,其中每对识别位点位于来自多个靶基因组区域的一个靶基因组区域的两侧,

b)将裂解的遗传物质变性为单链形式,和

c)通过将多个靶基因组区域与多个桥接寡杂交,以单链形式捕获多个靶基因组区域,其中每个桥接寡包含在3'和5'端分别与来自单链形式的所述多个靶基因组区域的一个靶基因组区域的3'和5'端杂交的序列。

26.根据权利要求25所述的方法,还包括分析所述多个靶基因组区域,包括:

d)连接与多个桥接寡杂交的单链靶基因组区域的游离端以产生多个单链环状靶基因组区域,每个单链环状靶基因组区域包含与相应桥接寡杂交的来自多个靶基因组的一个靶基因组区域,

e)可选地,去除未环化的遗传物质,

f)可选地,通过核酸扩增来扩增所述多个靶基因组区域,以产生每个靶基因组区域的多个拷贝,以及

g)分析扩增的靶基因组区域。

27.根据权利要求25或26所述的方法,其中所述多个桥接寡被固定在固体底物上。

28.根据权利要求27所述的方法,包括通过核酸扩增来扩增多个靶基因组区域,以产生每个靶基因组区域的多个拷贝。

29.根据权利要求25所述的方法,包括通过聚合酶链反应(PCR)扩增所述靶基因组区

域,以产生所述靶基因组区域的多个拷贝。

30.根据权利要求29所述的方法,其中所述PCR引物可以对目标区域具有特异性,或者通过将它们设计为与所述桥接寡引入的共同区域结合而对所有目标通用。

31.根据权利要求25所述的方法,包括通过滚环扩增(RCA)反应扩增所述多个靶基因组区域中的每一个,以产生所述多个靶基因组区域中的每一个的多个连接拷贝。

32.根据权利要求31所述的方法,其中每个桥接寡用作RCA反应的引物,以产生单链形式的每个靶基因组区域的多个连接拷贝。

33.根据权利要求31所述的方法,其中所述桥接寡中的每一个和/或一个或多个另外的引物用于所述RCA反应,以产生双链形式的所述靶基因组区域中的每一个的多个连接拷贝。

34.根据权利要求26所述的方法,其中所述分析包括检测所述靶基因组区域。

35.根据权利要求34所述的方法,其中从多个靶基因组区域检测一个靶基因组区域包括:使所述靶基因组区域与特异性结合靶基因组区域的标记探针或结合扩增的目标分子的物质接触。

36.根据权利要求26所述的方法,其中所述分析包括对所述靶基因组区域进行测序。

37.根据权利要求36所述的方法,其中所述测序包括纳米孔测序、可逆染料终止子测序或单分子实时(SMRT)测序。

38.一种试剂盒,包括:

a)一种或多种引导分子,以将一种或多种具有第一识别位点和第二识别位点的内切酶引导至一个或多个基因组靶区域,其中每个识别位点包含10至80个核苷酸的序列,

b)在基因组靶区域切割DNA的一种或多种可编程内切酶,以及

c)设计用于捕获由所述一种或多种可编程内切酶产生的靶基因组区域的一种或多种桥接寡,其中所述桥接寡是单链寡核苷酸,其包含在3'和5'端分别与单链靶基因组区域的3'和5'端的序列杂交的序列。

39.根据权利要求38所述的试剂盒,还包括以下一种或多种:设计用于在PCR或RCA反应中扩增环化靶基因组区域的引物、连接酶、聚合酶和一种或多种稀有切割限制性内切酶。

40.根据权利要求38或39所述的试剂盒,其中所述试剂盒是针对一个或多个特定靶基因组区域而定制的。

41.根据权利要求38所述的试剂盒,其中所述多个桥接寡固定在固体底物上。

捕获和分析靶基因组区域

[0001] 相关申请的交叉引用

[0002] 本申请要求于2019年5月13日提交的美国临时申请序列号62/846,988的权益,其公开内容在此通过引用整体并入,包括所有图、表和氨基酸或核酸序列。

[0003] 本申请的序列表标记为“Seq-List.txt”,创建于2020年5月11日,大小为3KB。序列表的全部内容通过引用整体并入本文。

背景技术

[0004] 捕获和分析,例如靶基因组区域的测序,在从农艺学、分类学到医学的不同生物技术领域中极为重要。例如,对基因组中的目标区域进行测序在精准和个性化医学中很重要,在这种情况下,了解基因组特定区域的基因多态性可以与有关受试者健康的表型信息相关联,这反过来又可以决定治疗方案。此外,对基因组的目标区域进行测序在用于改善植物所需性状的农艺应用(例如种子和食物生产)中也很重要。

[0005] 存在多种分析靶基因组DNA或RNA序列的方法。例如,测序前的靶向扩增很常见,并且有多种方法可用于分离和/或扩增基因组的目标区域。作为一个共同主题,本领域已知的方法利用聚合酶链反应(PCR)、连接链反应或探针杂交的变体来分离和扩增靶基因组区域。例如,可以使用环介导等温扩增(LAMP)或链置换扩增来进行遗传物质等温检测的当前方法。然而,这些技术难以针对不同的目标进行优化,不适合分析长目标。此外,这些等温技术很难或不可能针对大量目标复用,并且不适合简便地对扩增产物进行测序。表征、特别是检测和测序DNA序列的较长片段,例如至少约1-50千碱基(kb),是特别令人感兴趣的。然而,用常规方法不容易实现这样的检测和测序。因此,需要改进的方法,用于表征(例如检测和测序)靶基因组区域,特别是检测和测序靶基因组区域的较长片段。

发明内容

[0006] 本发明的某些实施方式提供了用于捕获靶基因组区域和可选地进一步例如通过检测和/或测序进行分析的材料和方法。本文公开的材料和方法特别适用于分析含有至少约1kb、优选至少约10kb、甚至更优选至少约30kb或最优选至少约50kb的靶基因组区域。

[0007] 根据本文公开的方法,可基于在两个特定识别位点处的裂解来分离靶基因组区域,每个识别位点包含最少约10至约30个核苷酸。两个识别位点位于靶基因组区域的两侧。然后可以使用寡核苷酸(本文称为“桥接寡”)捕获裂解的靶基因组区域。桥接寡在3’和5’端具有序列,其分别与裂解的靶基因组区域的3’和5’端的序列杂交。然后捕获的靶基因组区域可以例如通过扩增和测序进行分析,例如,检测和/或测序。

[0008] 因此,本发明的某些实施方式提供了从遗传物质中捕获靶基因组区域的方法,包括:

[0009] a.使用一种或多种具有第一识别位点和第二识别位点的内切酶从遗传物质裂解靶基因组区域,每个识别位点包含最少约10至约30个核苷酸的序列,其中识别位点位于靶基因组区域的两侧,

[0010] b.将裂解的遗传物质变性成单链形式,和

[0011] c.通过将靶基因组区域与桥接寡杂交,捕获单链形式的靶基因组区域,该桥接寡核在3'和5'端包含序列,其分别与单链形式的靶基因组区域的3'和5'端杂交。

[0012] 捕获的靶基因组区域可以被进一步分析,例如检测或测序。这种分析可以包括以下步骤:

[0013] d.连接与桥接寡杂交的单链靶基因组区域的游离端,以产生与桥接寡杂交的单链环状靶基因组区域,

[0014] e.可选地,降解未环化遗传物质,

[0015] f.可选地,通过核酸扩增来扩增靶基因组区域,以产生靶基因组区域的多个拷贝,和

[0016] g.分析扩增的靶基因组区域。

[0017] 在优选的实施方式中,使用第一和第二可编程内切酶,例如成簇规则间隔短回文重复序列(CRISPR)相关蛋白9内切酶(Cas9内切酶),例如,包含具有与第一识别位点互补的序列的第一向导RNA(gRNA)的第一Cas9内切酶和包含具有与第二识别位点互补的序列的第二gRNA的第二Cas9内切酶,在识别位点处裂解靶基因组区域。

[0018] 在进一步的实施方式中,与桥接寡杂交的靶基因组区域通过扩增反应(例如等温扩增反应)、优选地通过聚合酶链反应或滚环扩增(RCA)反应进行扩增。当仅使用桥接寡作为扩增引物时,通过RCA产生单链扩增产物,其包含单链形式的靶基因组区域的连接拷贝。除了桥接寡之外,还可以在RCA反应中使用一种或多种引物以产生双链扩增产物,其包含双链形式的靶基因组区域的连接拷贝。也可以使用常规的PCR反应来扩增目标分子,使用结合到目标区域或由桥接寡引入的公共区域的适当的引物。

[0019] 可以使用本领域已知的技术检测扩增的靶基因组区域,例如使用与靶基因组区域内的序列互补的标记探针。还可以使用本领域已知的技术对扩增的靶基因组区域进行测序,例如使用纳米孔测序(Oxford Nanopore Technologies™)、可逆染料终止子测序(Illumina™)和单分子实时(SMRT)测序(PacBio™)。

[0020] 可以改进本文公开的材料和方法以捕获和可选地分析例如在复用反应中的多个靶基因组区域。在此类实施方式中,多对靶识别位点被设计成裂解多个靶基因组区域,并且这些多个靶基因组区域可以通过多个桥接寡被捕获,多个桥接寡中的每一个被专门设计用于捕获靶基因组区域。在某些实施方式中,多个桥接寡可以固定在固体底物例如切片上。可以使用本领域已知的技术检测或测序如此捕获的多个靶基因组区域。

[0021] 本发明的进一步实施方式还提供了用于实施本发明方法的试剂盒。本发明的试剂盒包含一种或多种设计用于裂解一个或多个靶基因组区域的内切酶和设计用于捕获一个或多个靶基因组区域的特定桥接寡。

[0022] 在某些实施方式中,本发明的试剂盒包含设计用于裂解一个或多个靶基因组区域的一种或多种DNA或RNA形式的引导分子,以及设计用于捕获一个或多个靶基因组区域的特定桥接寡。此类试剂盒还可包含一种或多种Cas9内切酶。

[0023] 试剂盒还可包含聚合酶、连接酶、引物和其他试剂,用于扩增捕获的一个或多个靶基因组区域。更进一步,试剂盒可以提供执行本发明方法的说明书。

附图说明

[0024] 专利或申请文件包含至少一张着色绘制的图。具有彩图的本专利或专利申请公开文本的副本,将在请求和支付必要费用后由专利局提供。

[0025] 图1,捕获靶基因组区域的方法的概述以及分析(即检测或测序)靶基因组区域的两个示例。

[0026] 图2,使用单分子实时(SMRT)测序(PacBio™)对捕获的靶基因组区域进行测序的示例。

[0027] 图3,使用设计用于捕获多个靶基因组区域的桥接寡的阵列检测多个靶基因组区域的示例。

[0028] 图4,通过PCR扩增环化目标分子的示例。示出的引物包含使用可逆染料终止子测序(Illumina™)进行测序所需的接头。A)过程概述。B)捕获序列并为引物提供着陆位点所需的桥接寡的结构。C)引物的结构,允许同时扩增分子、掺入独特的样品标识并掺入染料终止子测序所需的结构。

[0029] 图5,基于结合双链基因组的同一条链的两个gRNA设计识别位点和桥接寡的示例。

[0030] 图6,基于结合双链基因组的相反链的两个gRNA设计识别位点和桥接寡的示例。

[0031] 图7,另一个基于结合双链基因组的相反链的两个gRNA设计识别位点和桥接寡的示例。

[0032] 图8,另一个基于结合双链基因组的同一条链的两个gRNA设计识别位点和桥接寡的示例。

[0033] 图9,连接目标分子末端以形成环状单链结构的不同方法的示例。

[0034] 图10,由桥接寡与非互补5'端部分造成的RCA反应的示例。

[0035] 图11,利用单个可编程内切酶和从目标延伸的引物来形成环状单链分子和桥接寡的示例。

[0036] 序列简述

[0037] SEQ ID NO:1:根据本发明的方法研究的示例性区域。

[0038] SEQ ID NO:2:示例性的第一识别位点。

[0039] SEQ ID NO:3:示例性的第二识别位点。

[0040] SEQ ID NO:4:示例性的靶基因组区域。

[0041] SEQ ID NO:5:在SE QID NO:4的靶基因组区域的5'端的序列。

[0042] SEQ ID NO:6:在SEQ ID NO:4的靶基因组区域的3'端的序列

[0043] SEQ ID NO:7:设计用于捕获SEQ ID NO:4的靶基因组区域的桥接寡的5'端的序列。

[0044] SEQ ID NO:8:设计用于捕获SEQ ID NO:4的靶基因组区域的桥接寡的3'端的序列。

[0045] SEQ ID NO:9:设计用于捕获SEQ ID NO:4的靶基因组区域的桥接寡。

具体实施方式

[0046] 如本文所用,单数形式“一个”和“该”等也旨在包括复数形式,除非上下文另有明确指示。此外,就具体的说明书和/或权利要求中使用的术语“包括”、“具有”、“带有”或其变

体而言,此类术语旨在以与术语“包含”的方式是包容的。过渡性术语/短语(及其任何语法变体)“包括”、“基本上由……组成”和“由……组成”可以互换使用。

[0047] 短语“基本上由……组成”表示所描述的实施方式包括包含特定材料或步骤的实施方式以及不实质上影响所描述实施方式的基本和新颖特征的那些实施方式。

[0048] 术语“约”是指在由本领域普通技术人员确定的特定值的可接受误差范围内,这将部分取决于该值是如何测量或确定的,即,测量系统的限制。在使用术语“约”的多核苷酸长度的上下文中,这些多核苷酸包含所述数目的碱基或碱基对,在该值附近有0-10%的变化(X±10%)。

[0049] 在本公开中,范围以速记方式陈述,以避免必须详细阐述并描述该范围内的每个值。在适当的情况下,可以选择该范围内的任何合适的值作为该范围的上限值、下限值或终点。例如,0.1-1.0的范围代表0.1和1.0的终值,以及0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9的中间值,以及包含在0.1-1.0内的所有中间范围,例如0.2-0.5、0.2-0.8、0.7-1.0等。设想范围内具有至少两位有效数字的值,例如5-10的范围表示5.0和10.0以及5.00到10.00之间的所有值,包括终端值。当本文例如针对多核苷酸的大小使用范围时,范围的组合和子组合(例如,所公开范围内的亚范围)和其中的具体实施方式被明确包括。

[0050] 如本文所用,术语“生物体”包括病毒、细菌、真菌、植物和动物。生物体的其他实施例是本领域普通技术人员已知的,并且此类实施方式在本发明的范围内。本文所述的测定可用于分析从任何生物体获得的任何遗传物质。

[0051] 如本文所用,术语“基因组”、“基因组的”、“遗传物质”或它们的其他语法变体是指来自任何生物体的遗传物质。遗传物质可以是病毒基因组DNA或RNA、核遗传物质(例如基因组DNA)或存在于细胞器中的遗传物质(例如线粒体DNA或叶绿体DNA)。它还可以代表来自天然或人工混合物或几种生物体的遗传物质。

[0052] 本文使用的短语“长靶基因组区域”是指具有至少约1kb、优选至少约10kb、甚至更优选至少约30kb或最优选至少约50kb的靶基因组区域.

[0053] 本文公开的用于表征靶基因组区域、特别是长靶基因组区域的材料和方法解决了与用于捕获和检测靶基因组区域、特别是长靶基因组区域的常规方法相关的问题。

[0054] 在某些实施方式中,本发明提供用于从遗传物质中捕获靶基因组区域的方法。该方法包括以下步骤:

[0055] a.使用一种或多种具有第一识别位点和第二识别位点的内切酶从遗传物质裂解靶基因组区域,每个识别位点包含最少约10至约30个核苷酸的序列,其中识别位点位于靶基因组区域的两侧,

[0056] b.将裂解的遗传物质变性成单链形式,和

[0057] c.通过将靶基因组区域与桥接寡杂交,以单链形式捕获靶基因组区域,该桥接寡包含分别在3’和5’端与单链形式的靶基因组区域的3’和5’端杂交的序列。

[0058] 捕获的靶基因组区域可以被进一步分析,例如,检测或测序。这种分析可以包括以下步骤:

[0059] d.连接与桥接寡杂交的单链靶基因组区域的游离端,以产生与桥接寡杂交的单链环状靶基因组区域,

[0060] e.可选地,降解未环化遗传物质,

[0061] f.可选地,通过核酸扩增来扩增靶基因组区域,以产生靶基因组区域的多个拷贝,和

[0062] g.分析扩增的靶基因组区域。

[0063] 如本文所用,“靶基因组区域”是生物体基因组中的感兴趣区域。这种区域的两侧是第一识别位点和第二识别位点。第一和第二识别位点中的每一个都包含最少约10至约30个核苷酸的序列。

[0064] 基于靶基因组区域选择第一和第二识别位点。通常,选择位于靶基因组区域两侧的独特序列作为第一和第二识别位点。这些序列的独特性确保这些序列不太可能出现在基因组的其他地方,从而最小化非特异性裂解并避免捕获靶基因组区域以外的区域。此外,第一和第二识别位点的最小长度优选在约10和约30个核苷酸之间,这确保这些序列在基因组中很少出现,从而也最小化对靶基因组区域以外的区域的非特异性裂解和捕获。第一和第二识别位点的序列可以彼此相同或彼此不同。本领域普通技术人员可以基于靶基因组区域的序列和特定生物体的可用基因组序列,例如从基因组序列数据库确定第一和第二识别位点的合适序列。

[0065] 使用在特定识别位点处裂解遗传物质中的磷酸二酯键(切割)的一种或多种内切酶来进行识别位点处遗传物质的裂解。内切酶可以是限制性内切酶。切割包含最少约10至约30个核苷酸、例如最多约80个核苷酸的识别位点的某些限制性内切酶,包括大范围核酸酶,例如来自LAGLIDADG家族、GIY-YIG家族、HNH家族、His-Cys盒家族和PD-(D/E)XK家族的归位内切酶。切割包含最少约10至约30个核苷酸、例如最多约80个核苷酸的识别位点的限制性内切酶的其他示例是本领域已知的,并且此类实施方式在本发明的范围内。

[0066] 在优选的实施方式中,使用一种或多种可编程的内切酶进行识别位点的裂解。术语可编程内切酶用于描述不同类别的酶,这些酶可以靶向裂解DNA或RNA分子的特定区域。因此,可编程内切酶是可以设计或编程以裂解感兴趣的核苷酸序列的内切酶。例如,可编程内切酶可以包括靶识别部分和内切酶部分,其中共同的内切酶部分可以与任何靶识别部分组合以裂解感兴趣的核苷酸序列。

[0067] 在一个实施方式中,可编程内切酶被引导RNA(gRNA)、引导DNA(gDNA)或引导分子与目标之间形成的结构靶向(Varshneyand Burgess,2016)。例如,Cas9是可编程的内切酶,因为它们通过在识别位点的特定位置产生双链断裂来裂解双链遗传物质(Jinek等,2012)。具有与识别位点序列互补的特定序列的gRNA将Cas9内切酶引导至识别位点。还表明gDNA或gRNA分子可用于靶向某些类型的argonaute蛋白质(Hegge等,2017年;Swarts等,2015年)。可编程内切酶的其他示例特别包括Cpf1、C2c1、C2c2、C2c、RNA或DNA引导的Argonaute蛋白质、结构引导的内切酶。

[0068] 新型蛋白质也可以被设计为基于识别在gDNA和目标序列之间形成的特定DNA结构(例如3’端错配)来裂解DNA(Xu等,2016)。通常,在本发明的方法中,使用两种Cas9内切酶从遗传物质裂解靶基因组区域,即,基于具有与第一识别位点互补的序列的第一gRNA在第一识别位点切割DNA的第一Cas9内切酶,以及基于具有与第二识别位点互补的序列的第二gRNA在第二识别位点切割DNA的第二Cas9内切酶。包含gRNA和Cas9内切酶成分的复合物被称为核糖核蛋白(RNP)复合物。在某些情况下,只需要一个RNP以裂解双链DNA的一条链,而双链DNA另一条链上的选择点是引物延伸反应的结果(图11)。

[0069] 可编程内切酶(例如Cas9)的使用提供了优于本领域已知的从目标区域产生、捕获或操纵环状分子的其他方法的显著改进(Dahl等,2007;Fredriksson等,2007)。首先,使用常见的限制性酶限制了可以分析的靶基因组区域,因为可能不存在在目标区域附近切割而不在目标区域内切割的限制性酶的组合。随着位点数量的增加,这尤其成问题。此外,即使存在这种限制性酶的组合,每次设想新目标时它们也可能会变化,从而使该过程很难扩展。相反,使用可编程内切酶允许在可扩展的过程中靶向几乎任何区域,因为独特的引导寡核苷酸的设计及其合成在本领域中是众所周知的。其次,如Fredriksson等(2007)所做的那样,使用PCR代替内切酶来提取要环化的位点会限制目标分子的大小,因为PCR往往最适合1Kb以下的片段。此外,PCR需要费力的优化,并且可能非常困难、成本高昂或无法同时对大量目标区域进行。

[0070] 本公开中第一和第二识别位点的序列从5'到3'方向表达。因此,gRNA的5'端的序列与识别位点的序列互补。除了与识别位点互补的序列外,Cas9-gRNA RNP复合物还识别约2到5个核苷酸(通常为3个核苷酸)的短保守序列基序,其位于目标DNA的非互补链附近并且定位到与识别位点互补的gRNA序列的3'端,被称为前间区序列相邻基序(PAM)。PAM对gRNA与识别位点的结合以及Cas9介导的目标遗传物质的裂解至关重要,尽管可以设想缺乏PAM位点的这种要求的工程化的Cas9酶。

[0071] PAM的一个实施例是序列NGG,但已知不同的Cas9酶需要不同的PAM位点,并且酶正在被修改以耐受可变的PAM位点(Hu等,2018年)。因此,在本文公开的材料和方法中有用的gRNA的某些实施方式中,设计识别位点,使得遗传物质上的识别位点的序列通过序列5'-NGG-3'紧跟在非互补链的3'侧。

[0072] 当gRNA与识别位点结合时,Cas9内切酶在朝向非互补链上NGG序列5'侧的三个核苷酸处(即从5'端开始并且朝向识别位点的3'端)在双链遗传物质中产生双链断裂,Cas9内切酶在第三个和第四个核苷酸之间产生双链断裂。

[0073] 一旦使用合适的内切酶裂解遗传物质,靶基因组区域和遗传物质的裂解片段的所得混合物被变性(例如通过将其置于变性条件下),以将其转化为单链形式。通常,使裂解的遗传物质经受变性条件包括在适当的化合物(如盐、二甲亚砜、氢氧化钠等)存在下并在适当的pH下使其经受适当的温度。也可以使用具有NaOH或高盐浓度的化学处理使DNA变性。在优选的实施方式中,裂解的遗传物质通过使其经受以下温度而变性:75℃至115℃,优选80℃至110℃,更优选85℃至105℃,甚至更优选90℃至100℃,最优选约95℃。本领域普通技术人员可以确定合适的变性条件,并且此类实施方式在本发明的范围内。

[0074] 为了捕获与遗传物质分离的靶基因组区域,将专门设计的桥接寡与裂解和变性的遗传物质接触。通常,桥接寡是单链寡核苷酸。桥接寡包含在3'和5'端分别与单链靶基因组区域的3'和5'端的序列杂交的序列。桥接寡也可以是具有分别与单链靶基因组区域的3'和5'端的序列杂交的3'和5'端突出端的双链寡核苷酸,可用于捕获目标DNA分子的两个链。可以通过本领域已知的不同方式保护桥接寡免于核酸酶降解,使得它们在用给定核酸酶处理后仍然存在于反应中,例如在其3'和/或5'端添加一个或多个硫代磷酸酯键,并分别在其3'和5'端添加反向dT和反向ddT。

[0075] 桥接寡有多种用途。桥接寡环化了由内切酶在目标上切割产生的正确靶基因组区域。由于在与内切酶的反应中预计会发生脱靶切割,因此桥接寡通过仅允许正确的分子被

环化,从而提供额外的特异性。此外,桥接寡本身可以作为后续扩增的引物。

[0076] 此外,可以设计桥接寡以提供额外的功能,例如制备用于测序的所得分子。例如,桥接寡可以固定在固体表面上以在切片上检测,或者可以生物素化,以使用链霉亲和素结合的珠子进行回收。桥接寡的5'和3'端也可以有额外的"尾部"序列,这些序列与目标区域不互补,创建可用于为分子连接附加功能的公共序列,例如提供位点用于PCR引物结合,添加生物素分子或本领域已知的其他改进(图10)。

[0077] 在与靶基因组区域杂交的两端的序列之间,桥接寡可以进一步包含序列,例如针对稀有切割限制性内切酶的限制位点、引物结合序列或可编程内切酶的目标。稀有切割限制位点或可编程内切酶的目标可用于从核酸扩增后产生的靶基因组区域的连接拷贝中裂解单个靶基因组区域。稀有切割限制性内切酶的非限制性实施例描述于PCT公开WO2009/079488中,其通过引用整体并入本文,特别是表1。

[0078] 如本文所用,"稀有切割限制性内切酶"是一种内切酶,其限制位点很少出现在遗传物质中。例如,对于人类基因组,稀有切割限制性内切酶是限制位点平均每50-100kb、优选地每100-200kb或更优选地每200-400kb、甚至更优选地每400-600Kb出现的内切酶。下表1给出了人类基因组的稀有切割限制性内切酶及其限制位点的示例:

[0079] 表1.人类稀有切割内切酶及其限制位点的示例。

[0080]

限制性酶	限制位点	人类基因组中的频率 (kb)
Not I	GCGGCCGC	1000
Xma III	CGGCCG	100
Sst II	CCGCGG	100
Sa lI	GTCGAC	100
Nru I	TCGCGA	300
Nhe I	GCTAGC	100

[0081] 其他的稀有切割内切酶描述于例如Restriction Endonucleases((Nucleic Acids and Molecular Biology),Pingoud(编者),Springer;第一版(2004))中。许多稀有切割内切酶也可商购获得,例如漫游类内切酶,其例如来自New England BioLabs(Beverly,MA)。稀有切割内切酶的更进一步的实施例是本领域已知的,并且这样的实施方式在本发明的范围内。

[0082] 桥接寡中的引物结合序列有助于在RCA反应中使用额外的引物,从而扩增靶基因组区域并通过原始分子的超支化以双链形式产生靶基因组区域的连接拷贝。

[0083] 第一和第二识别位点可以存在于遗传物质的相同链上或遗传物质的相反链上。

[0084] 图5提供了基于与双链基因组的同一条链结合的两个gRNA设计识别位点和桥接寡的示例。如图5所示,选择双链基因组DNA的底部链来设计识别位点。第一个gRNA与第一识别位点结合,Cas9内切酶在第一识别位点5'端下游三个核苷酸处切割基因组DNA。第二个gRNA与第二识别位点结合,Cas9内切酶在第二识别位点5'端下游三个核苷酸处切割基因组DNA。因此,靶基因组区域以双链形式从基因组DNA上裂解下来。该靶基因组区域在朝向第一识别

位点的末端具有除来自第一识别位点的前三个核苷酸之外的所有核苷酸,并且在朝向第二识别位点的末端具有来自第二识别位点的前三个核苷酸。该双链靶基因组区域可以转化为单链形式,并且可以设计一个或多个桥接寡来捕获一条或两条所述链。

[0085] 为了捕获图5中的底部链,桥接寡被设计为朝着3'端具有与第二识别位点的前三个核苷酸互补的序列,以及与存在于底部链上、超出并邻近第二识别位点的5'端的序列互补的其他核苷酸。该桥接寡朝着5'端具有与第一识别位点(除了前三个核苷酸)互补的序列,以及可选地,与存在于底部链上、超出并邻近第一识别位点的3'端的序列互补的其他核苷酸。

[0086] 为了捕获图5中的顶部链,桥接寡被设计为朝着3'端具有除前三个核苷酸之外的第一识别位点的序列,以及可选地,存在于底部链上、超出并邻近第一识别位点的3'端的序列。该桥接寡朝着5'端具有第二识别位点的前三个核苷酸的序列,以及存在于底部链上、超出并邻近第二识别位点的5'端的其他核苷酸。

[0087] 图6提供了一个基于结合双链基因组相反链的两个gRNA设计识别位点和桥接寡的示例。如图6所示,选择双链基因组DNA的底部链以设计第一识别位点,选择双链基因组DNA的顶部链以设计第二识别位点。第一个gRNA与底部链上的第一识别位点结合,Cas9内切酶在第一识别位点5'端下游三个核苷酸处切割基因组DNA。类似地,第二个gRNA与顶部链上的第二识别位点结合,Cas9内切酶在第二识别位点5'端下游的三个核苷酸处切割基因组DNA。因此,靶基因组区域以双链形式从基因组DNA上裂解下来。该靶基因组区域在朝向第一识别位点的末端具有除前三个核苷酸之外的第一识别位点的序列,并且在朝向第二识别位点的末端具有除前三个核苷酸之外的第二识别位点的序列。该双链靶基因组区域可以转化为单链形式,并且可以设计一个或多个桥接寡来捕获一条或两条所述链。

[0088] 为了捕获图6中的底部链,桥接寡被设计为朝向3'端具有除前三个核苷酸之外的第二识别位点的序列,以及可选地,存在于顶部链上、超出并邻近第二识别位点的3'端的序列。该桥接寡朝向5'端具有与除前三个核苷酸之外的第一识别位点互补的序列,以及可选地,与存在于底部链上、超出并邻近第一识别位点的3'端的序列互补的序列。

[0089] 为了捕获图6中的顶部链,桥接寡被设计为朝向3'端具有除前三个核苷酸之外的第一识别位点的序列,以及可选地,存在于底部链上、超出并邻近第一识别位点的3'端的序列。该桥接寡朝向5'端具有与除前三个核苷酸之外的第二识别位点互补的序列,以及可选地,与存在于顶部链上、超出和邻近第二识别位点的3'端的序列互补的序列。

[0090] 图7提供了另一个基于结合双链基因组的相反链的两个gRNA设计识别位点和桥接寡的示例。如图7所示,选择双链基因组DNA的顶部链以设计第一识别位点,选择双链基因组DNA的底部链以设计第二识别位点。第一个gRNA与顶部链上的第一识别位点结合,Cas9内切酶在第一识别位点5'端下游三个核苷酸处切割基因组DNA。类似地,第二个gRNA与底部链上的第二识别位点结合,Cas9内切酶在第二识别位点5'端下游三个核苷酸处切割基因组DNA。因此,靶基因组区域以双链形式从基因组DNA上裂解下来。该靶基因组区域在朝向第一识别位点的末端具有第一识别位点的前三个核苷酸,并且在朝向第二识别位点的末端具有第二识别位点的前三个核苷酸。该双链靶基因组区域可以转化为单链形式,并且可以设计一个或多个桥接寡来捕获一条或两条所述链。

[0091] 为了捕获图7中的底部链,桥接寡被设计为朝向3'端具有与第二识别位点的前三

个核苷酸互补的序列,以及与存在于底部链上、超出并邻近第二识别位点的5'端的序列互补的其他核苷酸。该桥接寡朝向5'端具有第一识别位点的前三个核苷酸的序列,以及存在于顶部链上、超出并邻近第一识别位点5'端的其他核苷酸。

[0092] 为了捕获图7中的顶部链,桥接寡被设计为朝向3'端具有与第一识别位点的前三个核苷酸互补的序列,以及与存在于顶部链上、超出并邻近第一识别位点的5'端的序列互补的其他核苷酸。该桥接寡朝向5'端具有第二识别位点的前三个核苷酸的序列,以及存在于底部链上、超出并邻近第二识别位点5'端的其他核苷酸。

[0093] 图8提供了另一个基于结合双链基因组的同一条链的两个gRNA设计识别位点和桥接寡的示例。如图8所示,选择双链基因组DNA的顶部链以设计第一识别位点和第二识别位点。第一个gRNA与顶部链上的第一识别位点结合,Cas9内切酶在第一识别位点5'端下游三个核苷酸处切割基因组DNA。类似地,第二个gRNA与顶部链上的第二识别位点结合,Cas9内切酶在第二识别位点5'端下游三个核苷酸处切割基因组DNA。因此,靶基因组区域以双链形式从基因组DNA上裂解下来。该靶基因组区域在朝向第一识别位点的末端具有第一识别位点5'端的前三个核苷酸,在朝向第二识别位点的末端具有除第二识别位点的5'端前三个核苷酸之外的所有核苷酸。该双链靶基因组区域可以转化为单链形式,并且可以设计一个或多个桥接寡来捕获一条或两条所述链。

[0094] 为了捕获图8中的底部链,桥接寡被设计为朝向3'端具有除前三个核苷酸之外的第二识别位点的序列,以及可选地,存在于顶部链上、超出并邻近第二识别位点的3'端的序列。该桥接寡朝向5'端具有第一识别位点的前三个核苷酸的序列,以及存在于顶部链上、超出并邻近第一识别位点5'端的其他核苷酸。

[0095] 为了捕获图8中的顶部链,桥接寡被设计为朝向3'端具有与第一识别位点的前三个核苷酸互补的序列,以及与存在于顶部链上、超出并邻近第一识别位点的5'端的序列互补的其他核苷酸。该桥接寡朝向5'端具有与除前三个核苷酸之外的第二识别位点互补的序列,以及可选地,与存在于顶部链上、超出并邻近第二识别位点的3'端的序列互补的序列。

[0096] 图5-8中表示的并在前面段落中描述的桥接寡的序列不需要与相应序列完全互补或相同,只要桥接寡可以与靶基因组区域上的相应序列杂交即可。因此,可以允许一定程度的错配,并且这种变化在本发明的范围内。

[0097] 术语“与...杂交”表示两个序列彼此足够互补,以允许两个序列之间的杂交。彼此杂交的序列可以完全互补,但也可能在一定程度上具有错配。因此,桥接寡5'和3'端的序列可与靶基因组区域5'和3'端的相应序列有一些错配,只要桥接寡可以杂交并捕获靶基因组区域即可。根据杂交的严格度,两个互补序列之间约5%至约20%的错配将允许两个序列之间的杂交。通常,高严格度条件具有较高的温度和较低的盐浓度,而低严格度条件具有较低的温度和较高的盐浓度。杂交的高严格度条件是优选的,因此,桥接寡3'和5'端的序列优选分别与靶基因组区域的3'和5'端的序列完美互补.

[0098] 捕获的靶基因组区域可以被进一步分析,例如,检测或测序。这种分析可以包括以下步骤:

[0099] d.连接与桥接寡杂交的单链靶基因组区域的游离端,以产生与桥接寡杂交的单链环状靶基因组区域,

[0100] e.可选地,降解未环化遗传物质,

[0101] f.可选地,通过核酸扩增来扩增靶基因组区域,以产生靶基因组区域的多个拷贝,和

[0102] g.分析扩增的靶基因组区域。

[0103] 在一些实施方式中,使用连接酶来连接与桥接寡杂交的单链靶基因组区域的游离端,以产生与桥接寡杂交的单链环状靶基因组区域(图9A)。当桥接寡中存在超出与靶基因组区域末端互补的序列的额外序列时,连接单链靶基因组区域的游离端还包括核酸合成,从而使用适当的DNA聚合酶以填充靶基因组区域两端之间的间隙(图9B)。也可以通过将单链寡核苷酸与间隙中的桥接寡杂交并将所得分子连接在一起来填充目标末端之间的间隙(图9C)。适用于连接单链靶基因组游离端的连接酶是本领域已知的,包括T4DNA连接酶、Ampligase、T7 DNA连接酶和Taq DNA连接酶。

[0104] 在某些实施方式中,可将适当环化的靶序列转化为双链载体。这可以通过用没有链置换能力的DNA聚合酶和DNA连接酶补充反应来完成,这将填充合成与环化目标互补的第二条链(图4)。

[0105] 在某些实施方式中,一旦靶基因组区域被适当环化,剩余的遗传物质(例如未裂解的基因组DNA和脱靶基因组片段)被移除,例如被降解。这可以通过外切酶的组合来实现,所述外切酶从暴露的3'或5'末端降解核酸,无论是单链还是双链,DNA还是RNA。因此,用合适的外切酶处理只留下环化的目标分子,大大降低了后续步骤的难度。

[0106] 在某些实施方式中,通过外切酶处理与遗传物质的其余部分分离并与桥接寡杂交的适当环化的靶序列,可以通过核酸扩增进行扩增。在某些优选的实施方式中,可以通过PCR进行扩增,使用桥接寡中的序列来设计可以扩增环状分子并将它们转化回线性分子的引物。这特别适用于DNA的短片段(<10Kb),其中PCR最有效率。用于多个目标的桥接寡可以包含一个与目标不互补的公共区域,其用于利用一对用于所有平行片段的通用引物驱动PCR反应(图4)。在其他优选的实施方式中,这种扩增是等温扩增,例如滚环扩增(RCA)。等温扩增有利于长片段DNA的扩增,例如,可以产生靶基因组区域的多个拷贝,每个拷贝包含至少约10-50kb。

[0107] 在传统的RCA反应中,挂锁合成探针用于杂交靶核酸序列并形成一个环(Nilsson等,1994)。然后在RCA反应中扩增该环。在本发明中,通过RCA反应,扩增使用桥接寡捕获的环化靶序列。在这种RCA反应中,桥接寡作为RCA反应的引物,线性扩增靶序列,产生靶序列的长单链串联重复序列。然后可以将扩增的单链靶序列用于后续分析。扩增的靶序列也可以转化为双链形式,例如通过捕获两条目标链并使它们的RCA产物再退火。

[0108] 在一些实施方式中,可以通过超支化RCA使用一种或多种额外引物(除了桥接寡本身)以产生靶序列的双链串联重复序列来获得目标环化分子的指数扩增。桥接寡核苷酸以外的引物可以根据桥接寡的序列设计,也可以根据靶基因组区域中存在的序列设计,或由随机碱基组合组成以扩增环的多个位置。

[0109] 将单链的捕获的靶基因组区域转化为双链形式的其他方法是本领域已知的,并且此类实施方式在本发明的范围内。

[0110] RCA可以创建和扩增不同大小的靶基因组区域,从而绕过依赖于通过PCR扩增核酸并且因此对于长的靶基因组区域的扩增无效的常规方法的限制。本文公开的方法提供的一个重要优点是检测和分析长靶基因组区域,例如,包含至少约1kb,优选至少约15kb,甚至更

优选至少约30kb或者最优选地至少约50kb的基因组区域。

[0111] 在某些实施方式中,可以分析RCA反应的产物以检测或测序靶基因组区域。例如,可以基于反应的浊度、荧光检测或标记的分子信标来检测扩增产物。

[0112] 术语“标记”是指可通过光谱、光化学、生物化学、免疫化学、化学或其他物理方式检测的分子。例如,有用的标记包括荧光染料(荧光团)、荧光猝灭剂、发光剂、电子致密试剂、生物素、地高辛、^{32}P和其他同位素或其他可以例如通过掺入寡核苷酸而被检测到的分子。该术语包括标记试剂的组合,例如荧光团的组合,每个荧光团例如在特定波长或波长组合提供独特的可检测特征。

[0113] 示例性荧光团包括但不限于Alexa染料(例如,Alexa 350、Alexa 430、Alexa 488等)、AMCA、BODIPY 630/650、BODIPY 650/665、BODIPY-FL、BODIPY-R6G、BODIPY-TMR、BODIPY-TRX、瀑布蓝、Cy2、Cy3、Cy5、Cy5.5、Cy7、Cy7.5、Dylight染料(Dylight405、Dylight488、Dylight549、Dylight550、Dylight649、Dylight680、Dylight750、Dylight800)、6-FAM、荧光素、FITC、HEX、6-JOE、俄勒冈绿488、俄勒冈绿500、俄勒冈绿514、太平洋蓝、REG、罗丹明绿、罗丹明红、ROX、R-藻红蛋白(R-PE)、星光蓝染料(例如,星光蓝520、星光蓝700)、TAMRA、TET、四甲基罗丹明、德克萨斯红和TRITC。

[0114] 本发明的某些实施方式提供了对多于四个不同目标的并行检测。例如,不同的桥接寡可以固定在底物(例如切片)上,其中每个桥接寡以及因此每个靶基因组区域的坐标是已知的。使用这样的底物,可以在一个复用反应中检测和分析数十个、数百个或甚至数千个目标序列。

[0115] 在另外的实施方式中,对PCR或RCA反应的产物进行测序。多种测序方法可用于对PCR和RCA的产物进行测序,例如使用便携式Nanopore Minion™或台式机、Nanopore Promethion™、PacBio Sequel™或Illumina HiSeq™。测序步骤还可用于多个目标的复用检测和/或多态性检测。

[0116] 在某些实施方式中,可以制备适当环化的短目标序列用于PCR扩增后的测序(图4)。PCR引物可以对目标区域具有特异性,或者优选地,通过设计它们以扩增由桥接寡引入的公共区域(图4B-4C)而对所有目标通用。一旦扩增,可以按照特定制造商的建议将接头分子添加到所述分子中,使其与所使用的测序仪兼容。或者,具有与桥接寡不互补的“尾部”的引物可用于在PCR过程中结合必要的接头,使该过程更有效(图4A和4C)。这些引物还可能包含一个短的独特序列(4-16个核苷酸),在PCR期间添加该序列以将测序数据关联到相应的样本,通常称为索引或条形码(图4C)。

[0117] 在某些实施方式中,可以裂解连接拷贝以产生靶基因组区域的单个拷贝,例如,使用稀有切割限制性酶在通过桥接寡引入的限制位点处切割连接拷贝。这种裂解产生靶基因组区域的多个拷贝,每个拷贝都有粘性末端。

[0118] 粘性末端可用于将靶基因组区域与接头序列结合。例如,包含与稀有切割限制酶特异性的限制位点互补的突出端的接头可以与靶基因组区域的拷贝混合以产生双链DNA,该双链DNA包含两侧具有接头的靶基因组区域。

[0119] 本文所用的术语“接头”是指已知的核苷酸序列,其具有4到100个核苷酸,优选10到20个核苷酸,甚至更优选约15个核苷酸,这取决于所使用的测序技术。接头序列一旦并入靶基因组区域的扩增拷贝的末端,就可以促进靶基因组区域的测序,例如,通过提供引物的

结合位点。在一个实施方式中,使用双端测序对两侧具有接头的靶基因组区域进行测序。

[0120] 本文中使用的术语“双端测序”是指使用存在于双链多核苷酸的每个末端上的特异性引物结合位点对片段的两个末端进行测序的测序技术。双端测序生成高质量的测序数据,使用计算机软件程序进行比对,以生成两侧为两个引物结合位点的多核苷酸序列。从双链分子的两端进行测序可以获得来自双链分子两端的高质量数据,因为仅从分子的一端进行测序可能会导致测序质量随着执行更长的测序读数而下降。

[0121] 在双端测序中,在接头掺入结束时产生的双链多核苷酸使用特异性引物进行测序,该引物结合到两侧具有接头的双链靶基因组区域的两端。双端测序的一般描述和原理在“Illumina Sequencing Technology,Illumina,公开号770-2007-002”中提供,其内容通过引用整体并入本文。

[0122] 双端测序技术的非限制性实施例由Illumina MiSeq™、Illumina MiSeqDx™和Illumina MiSeqFGx™提供。可在本发明的测定中使用的双端测序技术的其他实施例是本领域已知的,并且此类实施方式在本发明的范围内。

[0123] 在某些实施方式中,靶基因组区域的裂解拷贝的粘性末端可用于将靶基因组区域与发夹接头结合。例如,包含与稀有切割限制性酶特异性的限制位点互补的突出端的发夹接头可与靶基因组区域的拷贝混合,以产生包含两侧具有发夹接头的靶基因组区域的双链DNA。

[0124] 如本文所用,短语“发夹接头”是指包含双链主干和单链发夹环的多核苷酸。发夹接头的单链发夹环区域可为测序提供引物结合位点。因此,一旦发夹接头与靶基因组序列的两个粘性末端杂交,它就会产生一个双链DNA模板,该模板在两端由发夹环封端的双链区域中包含靶基因组区域。此类模板可用于通过单分子实时(SMRT)测序(PacBio™)对靶基因组区域进行测序。

[0125] “Pacific Biosciences(2018),公开号BR108-100318”中提供了SMRT测序的描述和原理,其内容通过引用整体并入本文。

[0126] 在进一步的实施方式中,纳米孔技术可用于对靶基因组区域进行测序。在某些此类实施方式中,如例如“Nanopore Technology Brochure,Oxford Nanopore Technologies(2019)”和“Nanopore Product Brochure,Oxford Nanopore Technologies(2018)”中所述,处理靶基因组区域的拷贝以对靶基因组区域进行测序。这些手册的内容通过引用整体并入本文。

[0127] 在某些实施方式中,多个靶基因组区域被捕获并可选地进一步分析,例如检测或测序。在此类实施方式中,为每个靶基因组区域设计一对gRNA。为了设计用于多个靶基因组区域的多个gRNA,选择gRNA的序列,使得用于一个靶基因组区域的Cas9内切酶不会破坏其他靶基因组区域。基于相关生物体的已知基因组序列,本领域普通技术人员可以确定识别位点的适当组合,以特异性裂解和分离多个靶基因组区域。还可以利用简并或摆动碱基设计和合成gRNA,以允许与多个或不确定的位置杂交。例如,这样做可以补偿gRNA目标位点中的已知多态性,添加多个简并碱基以随机杂交到基因组的多个位置,以及设计与基于遗传密码简并性编码已知氨基酸序列的未知基因杂交的gRNA。在这种情况下,如果需要,桥接寡也可以适于包含简并碱基。

[0128] 因此,本发明的某些实施方式提供了一种用于从遗传物质中捕获多个靶基因组区

域的方法。该方法包括以下步骤：

[0129] a.使用多对内切酶从遗传物质裂解多个靶基因组区域，每对内切酶包含第一识别位点和第二识别位点，每个识别位点包括最少约17至约24个核苷酸的序列，其中每对识别位点位于来自多个靶基因组区域的靶基因组区域的两侧，

[0130] b.将裂解的遗传物质变性成单链形式，和

[0131] c.通过将靶基因组区域与多个桥接寡杂交，以单链形式捕获多个靶基因组区域，其中每个桥接寡包含分别在3'和5'端与来自单链形式的多个靶基因组区域的靶基因组区域的3'和5'端杂交的序列。

[0132] 上述捕获靶基因组区域的方面，例如设计识别位点、用于裂解遗传物质的内切酶和设计用于捕获裂解的靶基因组区域的桥接寡，也适用于捕获多个靶基因组区域的本方法。

[0133] 在捕获多个靶基因组区域的一个实施方式中，提供了一种底物，其具有在特定已知位置与该底物结合的多个桥接寡。用多对内切酶裂解的遗传物质被转化为单链形式，并且单链形式的多个靶基因组区域在适当条件下与底物接触，以允许与底物结合的桥接寡核苷酸与多个靶基因组区域杂交。一旦靶基因组区域与相应的桥接寡杂交并环化，就可以进一步分析存在于特定位置的靶基因组区域。在一些实施方式中，捕获的靶基因组区域使用RCA扩增并进一步检测，例如使用激光激发和发射检测方法。

[0134] 在某些实施方式中，多个靶基因组区域被进一步分析，例如，检测或测序。这种分析可以包括以下步骤：

[0135] d.连接与多个桥接寡杂交的单链靶基因组区域的游离端以产生多个单链环状遗传物质，每个包含来自多个靶基因组区域的与相应的桥接寡杂交的靶基因组区域，

[0136] e.可选地，从底物中去除未环化遗传物质，

[0137] f.可选地，通过核酸扩增来扩增多个靶基因组区域，以产生每个靶基因组区域的多个拷贝，和

[0138] g.分析扩增的多个靶基因组区域。

[0139] 上述分析靶基因组区域的方面，例如连接单链靶基因组区域的游离端、检测靶基因组区域或对靶基因组区域进行测序，也适用于分析多个靶基因组区域的本方法。

[0140] 在某些实施方式中，可以通过用外切酶降解或通过将其从底物洗掉来去除未环化遗传物质。

[0141] 本发明的进一步实施方式提供了用于进行本发明的测定的试剂盒。本发明的试剂盒可以包含进行本发明的测定所必需的特定内切酶、设计用于靶向一个或多个靶基因组区域的特定引导分子、设计用于处理从测定中获得的测序数据的计算机软件程序以及任选的提供执行测定的说明的材料。在一个实施方式中，本发明的试剂盒包括：

[0142] a)一对或多对引导分子，

[0143] b)一种或多种桥接寡，其设计用于捕获一个或多个靶基因组区域。

[0144] 该试剂盒还可包含一种或多种可编程的内切酶。

[0145] 此外，试剂盒可包含用于环化靶基因组区域的PCR或RCA反应的引物、DNA连接酶、聚合酶和其他用于PCR或RCA的试剂、限制性内切酶(例如裂解靶基因组区域的连接拷贝的稀有切割剂)和测序试剂。

[0146] 在某些实施方式中,本发明的试剂盒可以针对一个或多个特定靶基因组区域进行定制。例如,用户可以提供一个或多个靶基因组区域的序列,并且可以生产试剂盒来针对所述一个或多个目标序列进行本发明的测定。

[0147] 在此提及或引用的所有专利、专利申请、临时申请和出版物通过引用整体并入,包括所有附图和表格,只要它们与本说明书的明确教导不矛盾。

[0148] 以下是说明实施本发明的程序的实施例。这些实施例不应被解释为限制。除非另有说明,所有百分比均按重量计,所有溶剂混合物比例均按体积计。

[0149] 实施例1-设计识别位点并捕获靶基因组区域

[0150] 本实施例描述了根据本文公开的材料和方法用于捕获和分析靶基因组区域的典型程序。例如,在本实施例中确定了以下序列用于分析。

[0151] 5'ACCCACTGTTGAGAGTCAGTGGCAAGAGAAGTCTCGTCTTTTACGTCCCGTATCAAAATGAGTGTACAATACAT[A/G]TCTATGTGCGAGTGAGA|GCACGGGTCTGGGCCTGGAGAGTGGACCACCTACAATTGGCCATTTCGGTTTGCGGAAGCTGTCAAGTCAACGCGAGTCCTAG[G/A]ATCTCATAGTCTTCGCATTAACCCGTATTAAGTGGACTCGCCTACAGTTTGTCTTATGCTAGCAACCCAGGCA[T/C]AGTCTGTACGCGCGATCGTCCACTGG|TAGTGGGGTCCTTCCTGGAGTGT[C/G]GTATGGGCCATCCGGTCTACCTTACTAACTTAGGCTTTAAGCGCATTTCTATGTGCGTGAGGTGTCGCATTCATACTTAGTCTGGTCCTAAGTCTGTCACC 3'(SEQ ID NO:1)。

[0152] SEQ ID NO:1的序列提供了非互补链的序列,即gRNA与相反链结合。因此,识别位点存在于相反的链上,对应于识别位点的区域为加粗斜体,NGG PAM基序带下划线,Cas9内切酶裂解的位点用竖线表示。第一个gRNA被设计成结合具有与TCTATGTGCGAGTGAGAGCA(SEQ ID NO:2)的序列反向互补的序列的第一识别位点,并且第二个gRNA被设计成结合具有与GCGCGATCGTCCACTGGTAG(SEQ ID NO:3)序列反向互补的序列的第二识别位点。这两个识别位点位于一个基因组序列的两侧,该基因组序列具有某种单核苷酸多态性(SNP),其由实心括号中的核苷酸表示。Cas9介导的该序列的裂解将产生184bp的序列。对该片段进行测序将提供有关该区域内存在的SNP的信息。

[0153] Cas9介导的含有SEQ ID NO:1序列的遗传物质的裂解将产生以下片段,其是靶基因组区域。

[0154] 5'GCACGGGTCTGGGCCTGGAGAGTGGACCACCTACAATTGGCCATTTCGGTTTGCGGAAGCTGTCAAGTCAACGCGAGTCCTAG[G/A]ATCTCATAGTCTTCGCATTAACCCGTATTAAGTGGACTCGCCTACAGTTTGTCTTATGCTAGCAACCCAGGCA[T/C]AGTCTGTACGCGCGATCGTCCACTGG 3'(SEQ ID NO:4)。

[0155] 为了创建桥接寡以捕获SEQ ID NO:4的靶基因组区域,从靶基因组区域的末端选择以下序列,如以上复制的序列中下划线所示:

[0156] 5'GCACGGGTCTGGGCCTGGAGAGTGGACCAC(SEQ ID NO:5)

[0157] 5'GCATAGTCTGTACGCGCGATCGTCCACTGG(SEQ ID NO:6)。

[0158] 为桥接寡的末端产生与这些序列互补的序列,以捕获SEQ ID NO:4的靶基因组区域。

[0159] 5'GTGGTCCACTCTCCAGGCCCAGACCCGTGC3'(SEQ ID NO:7)

[0160] 5'CCAGTGGACGATCGCGCGTACAGACTATGC3'(SEQ ID NO:8)

[0161] SEQ ID NO:7和8连接以产生以下序列的桥接寡:

[0162] 5'GTGGTCCACTCTCCAGGCCCAGACCCGTGCCCAGTGGACGATCGCGCGTACAGACTATGC3'

(SEQIDNO:9)。

[0163] 应当理解,本文所述的实施例和实施方式仅用于说明目的,本领域技术人员将根据其进行各种修改或变化,这些修改或变化将被包括在本申请的精神和范围内和所附权利要求的范围内。此外,本文公开的任何发明或其实施方式的任何要素或限定可以与本文公开的任何和/或所有其他要素或限定(单独或以任何组合)或任何其他发明或其实施方式组合,并且所有此类组合均是在本发明的范围内考虑到的,但不限制本发明的范围。

[0164] 参考文献

[0165] Dahl,F.,Stenberg,J.,Fredriksson,S.,Welch,K.,Zhang,M.,Nilsson,M.,et al.(2007).Multigene amplification and massively parallel sequencing for cancer mutation discovery.Proc.Natl.Acad.Sci.U.S.A.104,9387-92.doi:10.1073/pnas.0702165104.

[0166] Fredriksson,S.,Banér,J.,Dahl,F.,Chu,A.,Ji,H.,Welch,K.,et al.(2007).Multiplex amplification of all coding sequences within 10cancer genes by Gene-Collector.Nucleic Acids Res.35,e47.doi:10.1093/nar/gkm078.

[0167] Hegge,J.W.,Swarts,D.C.,and van der Oost,J.(2017).Prokaryotic Argonaute proteins:novel genome-editing tools?Nat.Rev.Microbiol.16,5-11.doi:10.1038/nrmicro.2017.73.

[0168] Hu,J.H.,Miller,S.M.,Geurts,M.H.,Tang,W.,Chen,L.,Sun,N.,et al.(2018).Evolved Cas9 variants with broad PAM compatibility and high DNA specificity.Nature 556,57-63.doi:10.1038/nature26155.

[0169] Jinek,M.,Chylinski,K.,Fonfara,I.,Hauer,M.,Doudna,J.A.,and Charpentier,E.(2012).A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.Science 337,816-21.doi:10.1126/science.1225829.

[0170] Nilsson,M.,Malmgren,H.,Samiotaki,M.,Kwiatkowski,M.,Chowdhary,B.P.,and Landegren,U.(1994).Padlock probes:circularizing oligonucleotides for localized DNA detection.Science 265,2085-8.Available at:http://www.ncbi.nlm.nih.gov/pubmed/7522346[Accessed May 1,2018].

[0171] Swarts,D.C.,Hegge,J.W.,Hinojo,I.,Shiimori,M.,Ellis,M.A.,Dumrongkulraksa,J.,et al.(2015).Argonaute of the archaeon Pyrococcus furiosus is a DNA-guided nuclease that targets cognate DNA.Nucleic Acids Res.43,5120-5129.doi:10.1093/nar/gkv415.

[0172] Varshney,G.K.,and Burgess,S.M.(2016).DNA-guided genome editing using structure-guided endonucleases.Genome Biol.17,187.doi:10.1186/s13059-016-1055-4.

[0173] Xu,S.,Cao,S.,Zou,B.,Yue,Y.,Gu,C.,Chen,X.,et al.(2016).An alternative novel tool for DNA editing without target sequence limitation:the structure-guided nuclease.Genome Biol.17,186.doi:10.1186/s13059-016-1038-5.

[0001] 序列表

[0002] <110> 快速基因组学有限责任公司

[0003] <120> 捕获和分析靶基因组区域

[0004] <130> RAP.101XPCT

[0005] <150> US 62/846,988

[0006] <151> 2019-05-13

[0007] <160> 9

[0008] <170> PatentIn版本3.5

[0009] <210> 1

[0010] <211> 401

[0011] <212> DNA

[0012] <213> 人工序列

[0013] <220>

[0014] <223> 合成

[0015] <220>

[0016] <221> 混合特征

[0017] <222> (75)..(75)

[0018] <223> n = Ala or Gly

[0019] <220>

[0020] <221> 混合特征

[0021] <222> (176)..(176)

[0022] <223> n = Gly or Ala

[0023] <220>

[0024] <221> 混合特征

[0025] <222> (250)..(250)

[0026] <223> n = Thr or Cys

[0027] <220>

[0028] <221> 混合特征

[0029] <222> (300)..(300)

[0030] <223> n = Cys or Gly

[0031] <400> 1

[0032] acccactgtt gagagtcagt ggcaagagaa gtctcgtctt ttacgtcccg tatcaaaatg 60

[0033] agtgtacaat acatntctat gtgcgagtga gagcacgggt ctgggcctgg agagtggacc 120

[0034] acctacaatt ggccatttcg gtttgcggaa gctgtcaagt caacgcgagt cctagnatct 180

[0035] catagtcttc gcattaaccc gtattaagtg gactcgccta cagtttgtct tatgctagca 240

[0036] acccaggcan agtctgtacg cgcgatcgtc cactggtagt ggggtccttc ctggagtgtn 300

[0037] gtatgggcca tccggtctac cttactaact taggctttaa gcgcatttct atgtgcgtga 360

[0038] ggtgtcgcat tcatacttag tctggtccta agtctgtcac c 401

[0039] <210> 2

[0040] <211> 20

[0041] <212> DNA

[0042] <213> 人工序列

[0043] <220>

[0044] <223> 合成

[0045] <400> 2

[0046] tctatgtgcg agtgagagca 20

[0047] <210> 3

[0048] <211> 20

[0049] <212> DNA

[0050] <213> 人工序列

[0051] <220>

[0052] <223> 合成

[0053] <400> 3

[0054] gcgcgatcgt ccactggtag 20

[0055] <210> 4

[0056] <211> 184

[0057] <212> DNA

[0058] <213> 人工序列

[0059] <220>

[0060] <223> 合成

[0061] <220>

[0062] <221> 混合特征

[0063] <222> (84)..(84)

[0064] <223> n = Gly or Ala

[0065] <220>

[0066] <221> 混合特征

[0067] <222> (158)..(158)

[0068] <223> n = Thr or Cys

[0069] <400> 4

[0070] gcacgggtct gggcctggag agtggaccac ctacaattgg ccatttcggt ttgcggaagc 60

[0071] tgtcaagtca acgcgagtcc tagnatctca tagtcttcgc attaacccgt attaagtgga 120

[0072] ctcgcctaca gtttgtctta tgctagcaac ccaggcanag tctgtacgcg cgatcgtcca 180

[0073] ctgg 184

[0074] <210> 5

[0075] <211> 30

[0076] <212> DNA

[0077] <213> 人工序列

[0078] <220>

[0079] <223> 合成

[0080] <400> 5

[0081] gcacgggtct gggcctggag agtggaccac 30

[0082] <210> 6

[0083] <211> 30

[0084] <212> DNA

[0085] <213> 人工序列

[0086] <220>

[0087] <223> 合成

[0088] <400> 6

[0089] gcatagtctg tacgcgcgat cgtccactgg 30

[0090] <210> 7

[0091] <211> 30

[0092] <212> DNA

[0093] <213> 人工序列

[0094] <220>

[0095] <223> 合成

[0096] <400> 7

[0097] gtggtccact ctccaggccc agacccgtgc 30

[0098] <210> 8

[0099] <211> 30

[0100] <212> DNA

[0101] <213> 人工序列

[0102] <220>

[0103] <223> 合成

[0104] <400> 8

[0105] ccagtggacg atcgcgcgta cagactatgc 30

[0106] <210> 9

[0107] <211> 60

[0108] <212> DNA

[0109] <213> 人工序列

[0110] <220>

[0111] <223> 合成

[0112] <400> 9

[0113] gtggtccact ctccaggccc agacccgtgc ccagtggacg atcgcgcgta cagactatgc 60

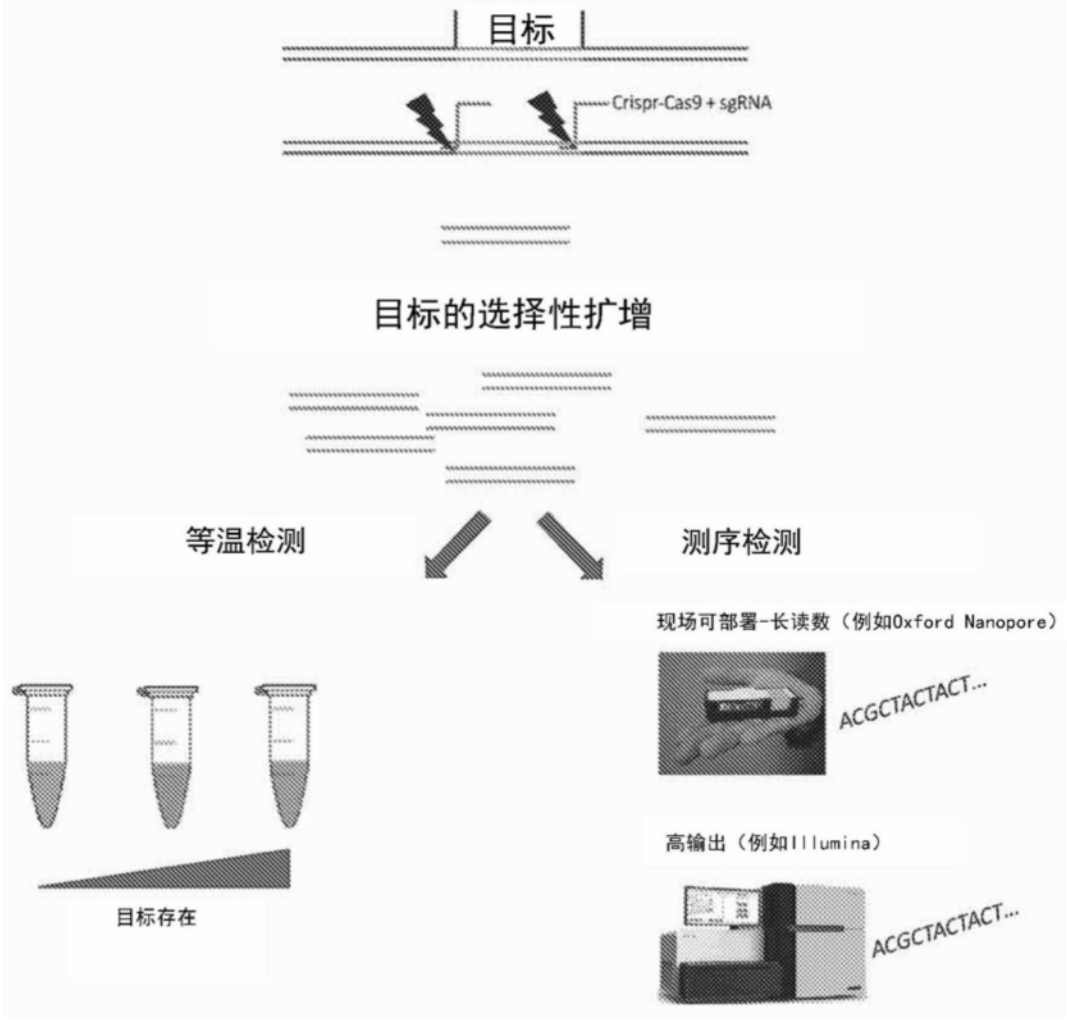

目标
Crispr-Cas9 + sgRNA
目标的选择性扩增
等温检测
测序检测
现场可部署-长读数（例如Oxford Nanopore）
ACGCTACTACT...
高输出（例如Illumina）
ACGCTACTACT...
目标存在

图1

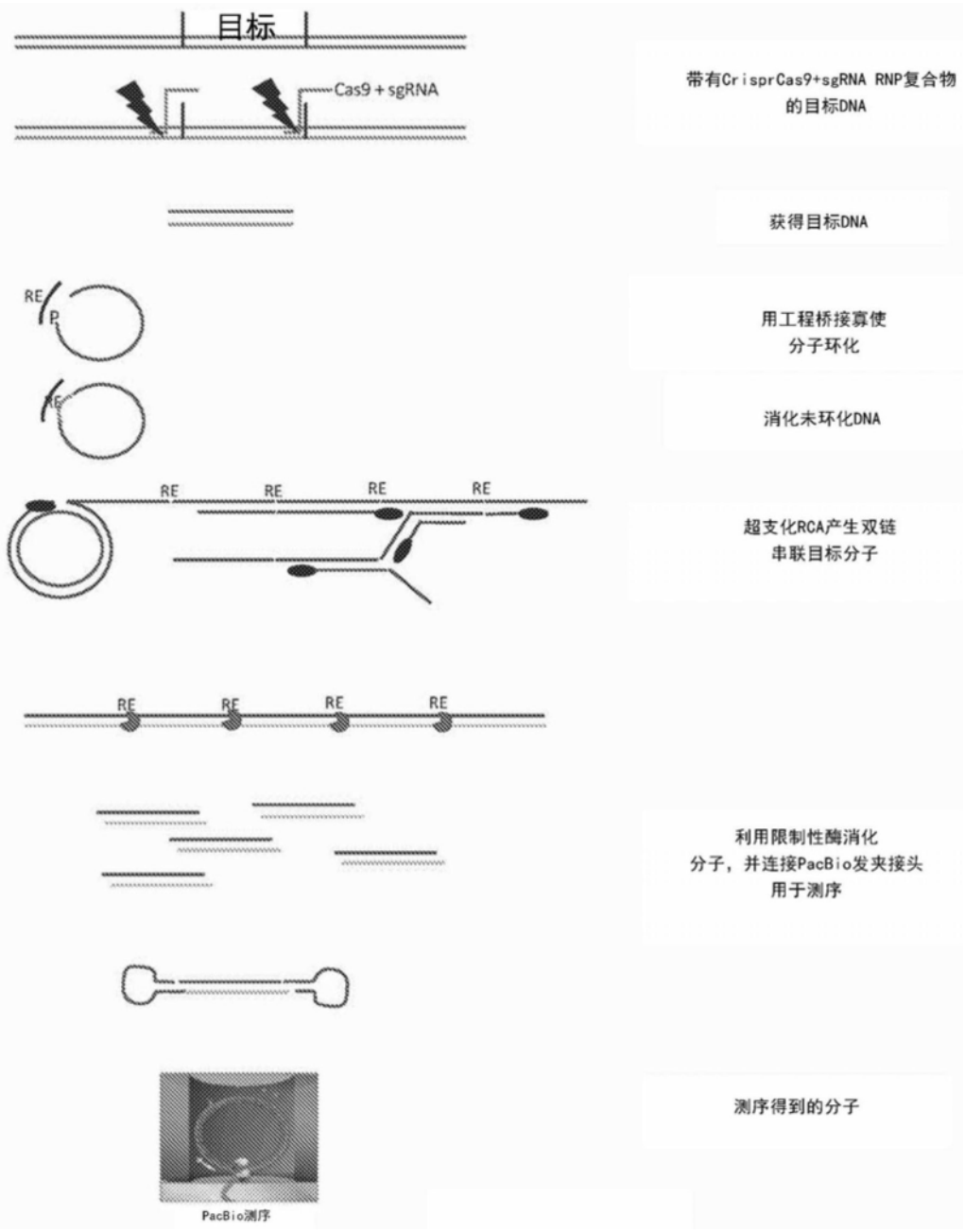

目标
Cas9+sgRNA
带有CrisprCas9+sgRNA RNP复合物
的目标DNA
获得目标DNA
RE
P
用工程桥接寡使
分子环化
消化未环化DNA
RE
RE
RE
RE
超支化RCA产生双链
串联目标分子
RE
RE
RE
RE
利用限制性酶消化
分子，并连接PacBio发夹接头
用于测序
测序得到的分子
PacBio测序

图2

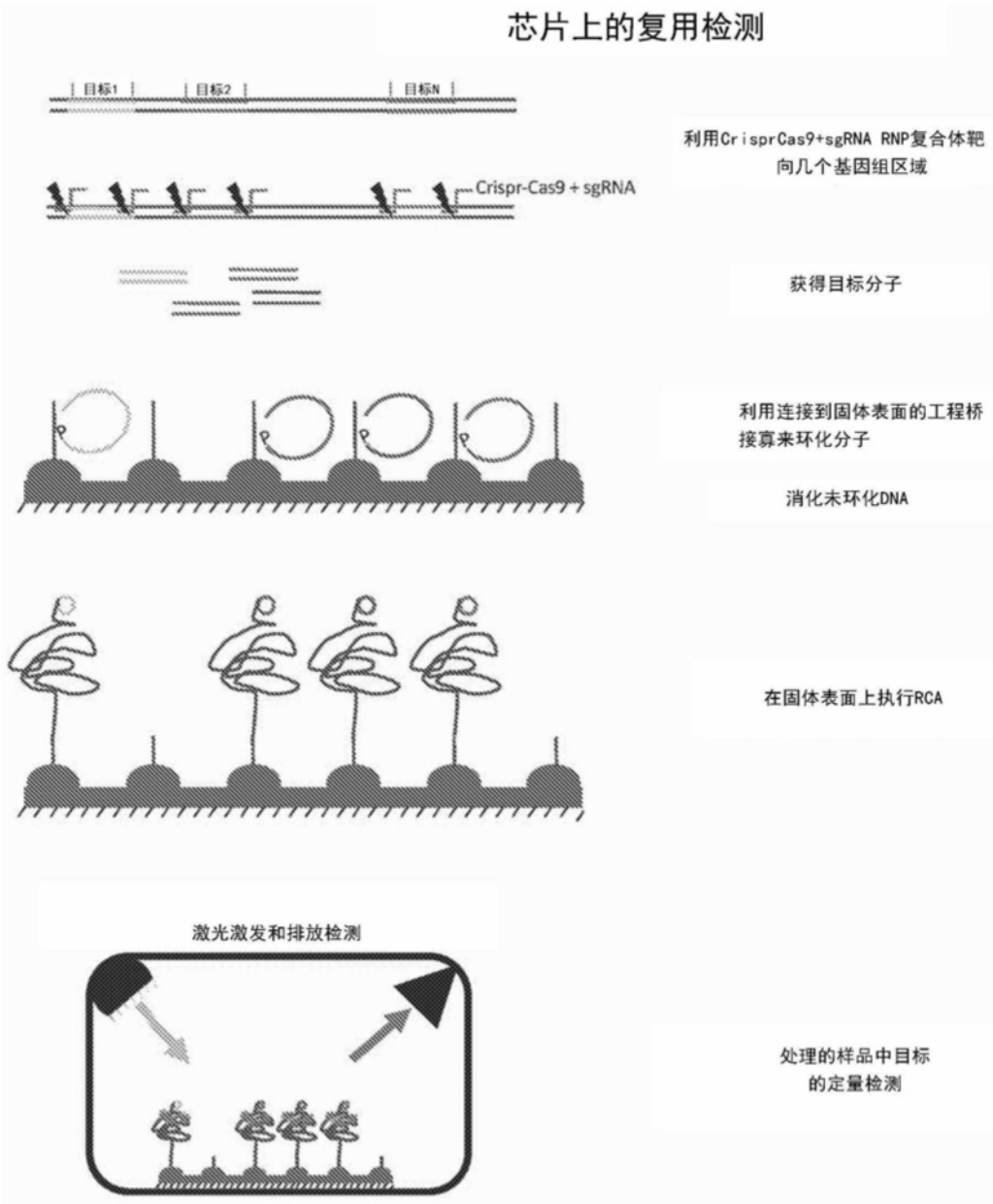

芯片上的复用检测
目标1
目标2
目标N
利用CrisprCas9+sgRNA RNP复合体靶向几个基因组区域
Crispr-Cas9 + sgRNA
获得目标分子
利用连接到固体表面的工程桥接募来环化分子
消化未环化DNA
在固体表面上执行RCA
激光激发和排放检测
处理的样品中目标的定量检测

图3

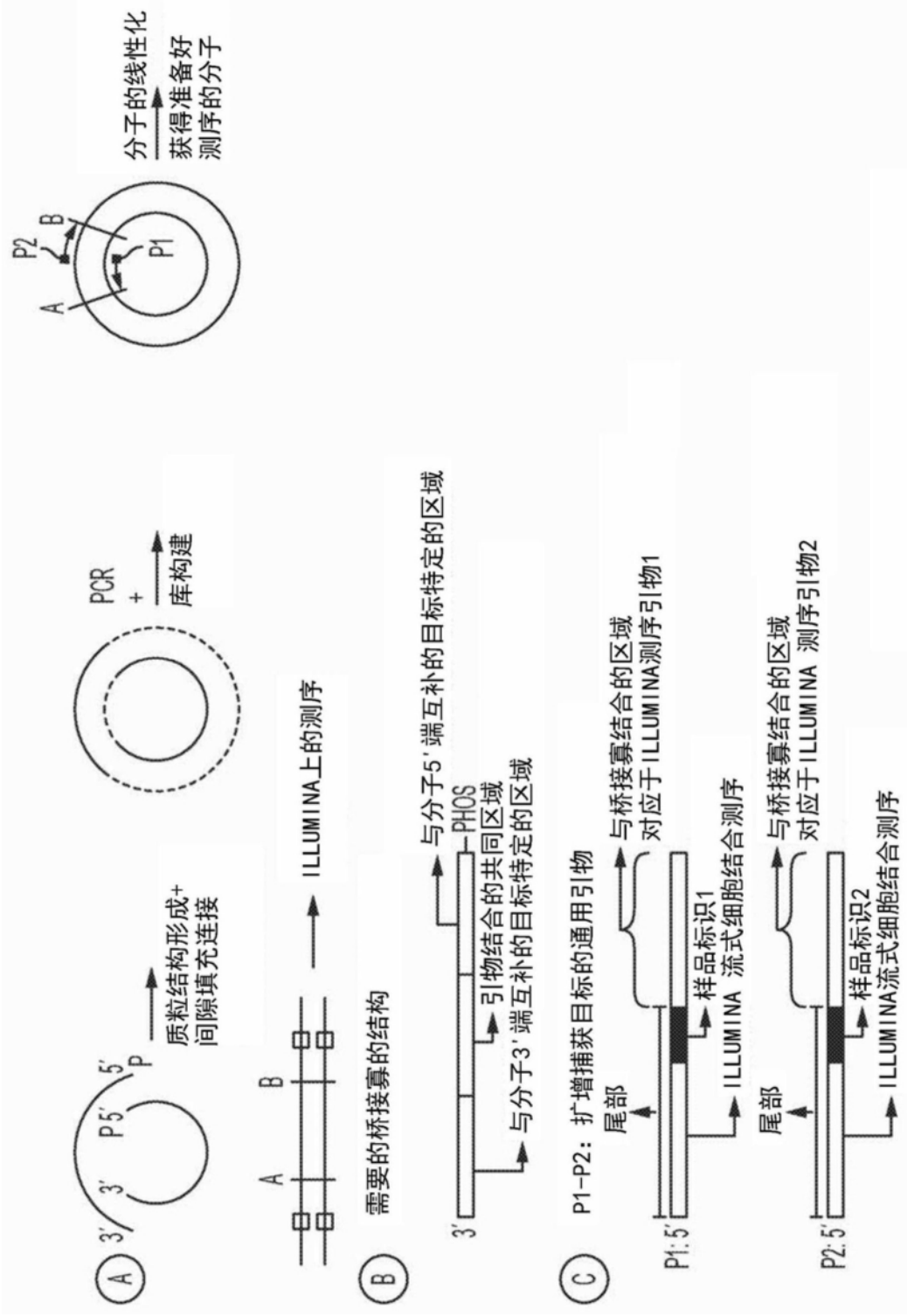

A
3'
P5'
5'
P
质粒结构形成+
间隙填充连接
PCR
+
库构建
分子的线性化
获得准备好
测序的分子
P2
B
A
P1
B
需要的桥接寡的结构
A
B
ILLUMINA上的测序
与分子5'端互补的目标特定的区域
PHOS
引物结合的共同区域
与分子3'端互补的目标特定的区域
3'
C
P1-P2：扩增捕获目标的通用引物
尾部
与桥接寡结合的区域
对应于ILLUMINA测序引物1
P1: 5'
样品标识1
ILLUMINA 流式细胞结合测序
尾部
与桥接寡结合的区域
对应于ILLUMINA 测序引物2
P2: 5'
样品标识2
ILLUMINA流式细胞结合测序

图4

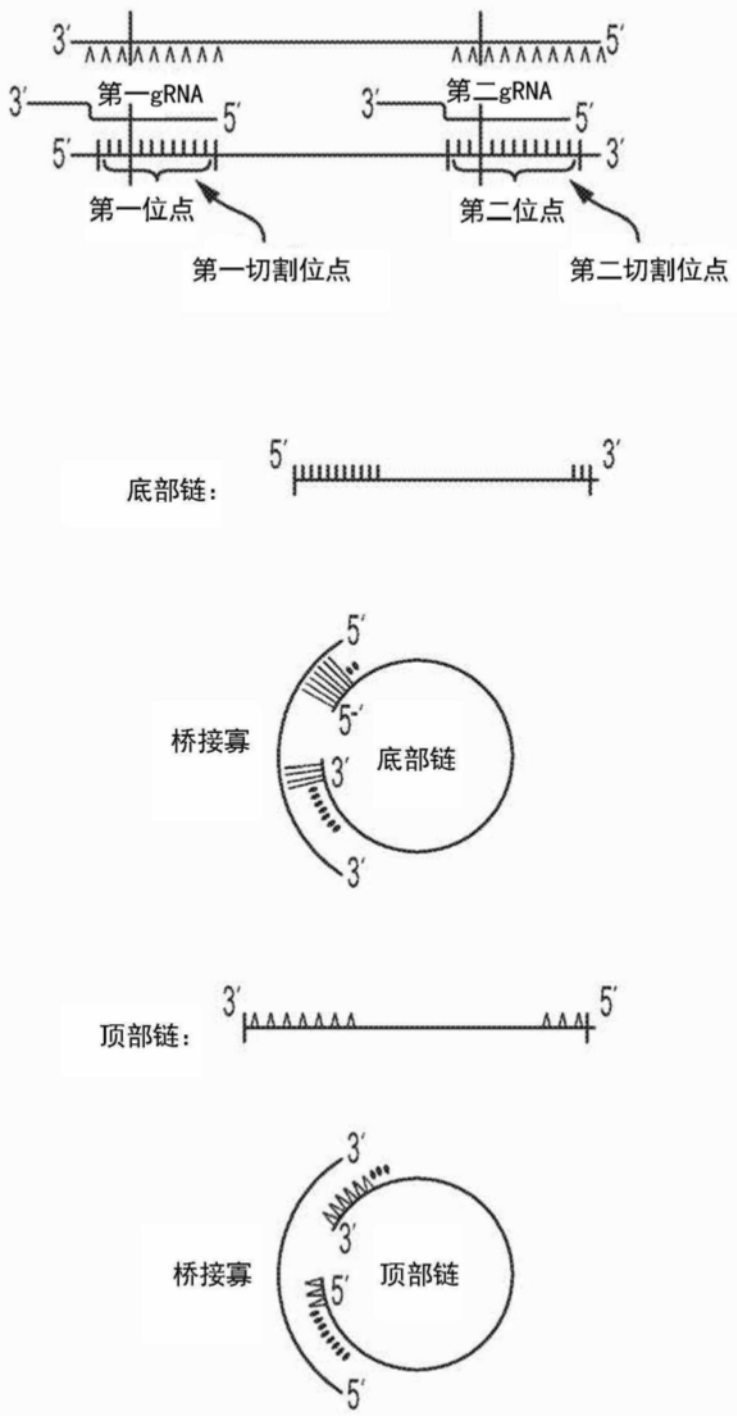

3′
5′
第一gRNA
第二gRNA
第一位点
第二位点
第一切割位点
第二切割位点
底部链：
桥接寡
底部链
顶部链：
桥接寡
顶部链

图5

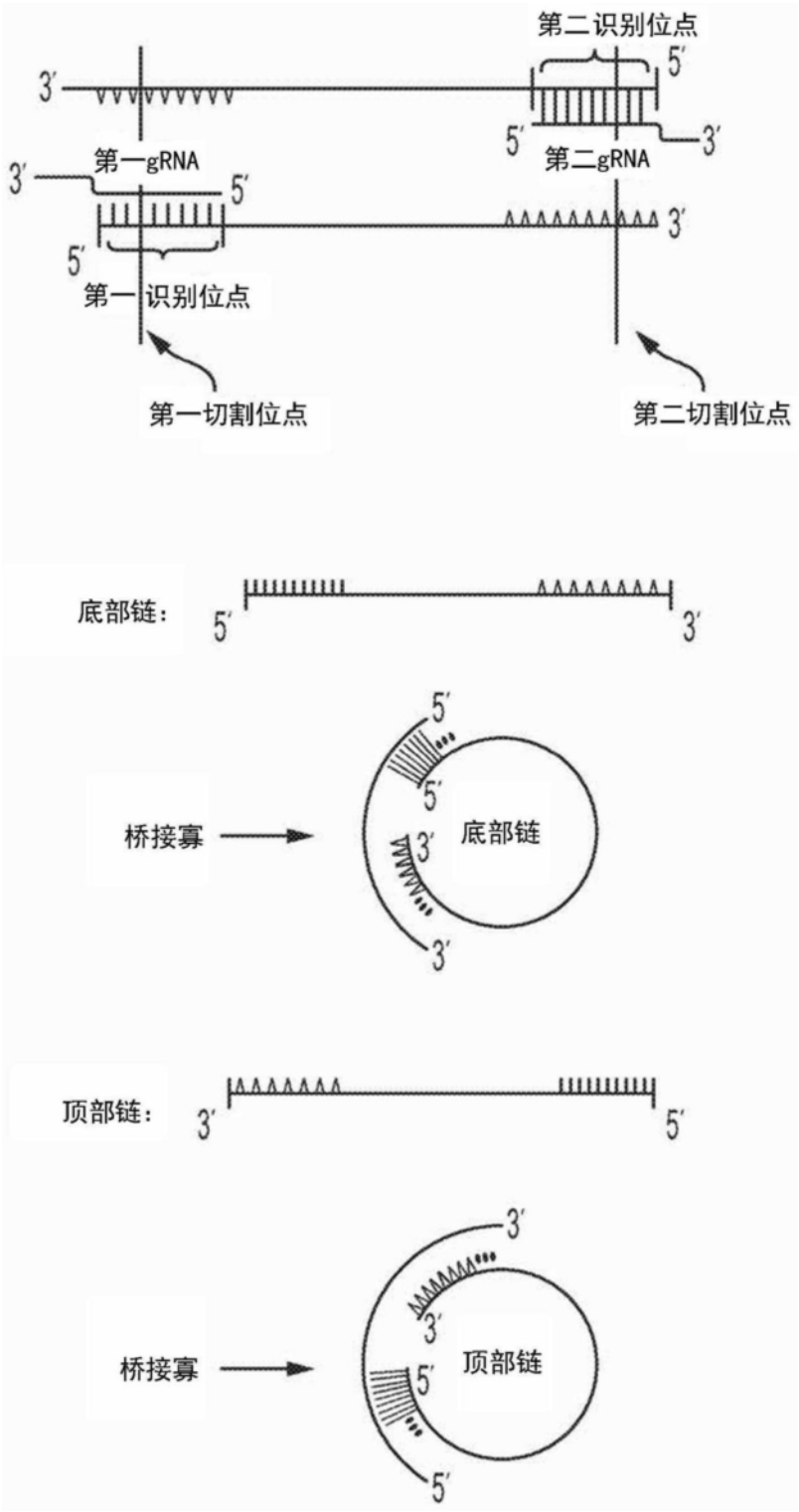

第二识别位点
3′
5′
5′
3′
第一gRNA
第二gRNA
3′
5′
3′
5′
第一识别位点
第一切割位点
第二切割位点
底部链：
5′
3′
5′
5′
桥接寡
3′
底部链
3′
顶部链：
3′
5′
3′
3′
桥接寡
5′
顶部链
5′

图6

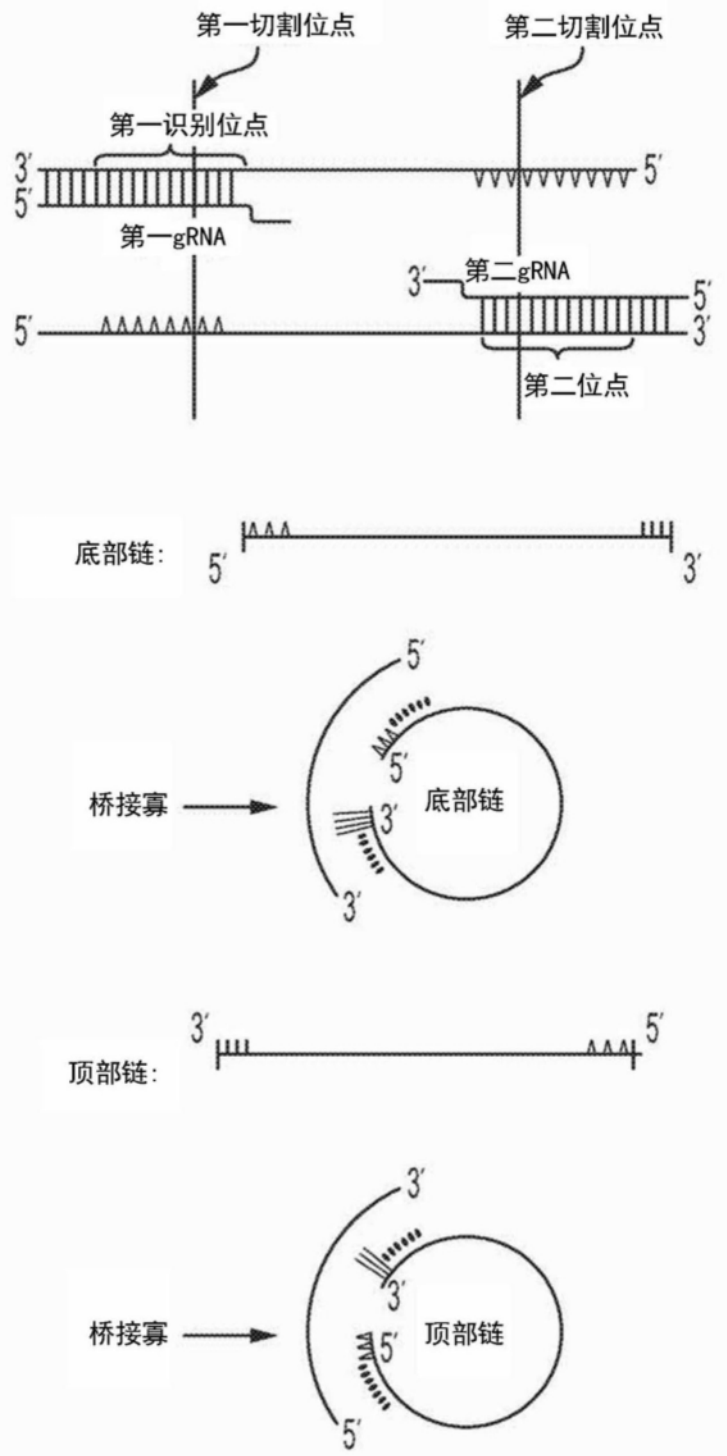

第一切割位点
第二切割位点
第一识别位点
3′
5′
5′
第一gRNA
3′
第二gRNA
5′
5′
3′
第二位点
底部链:
5′
3′
5′
5′
桥接寡
3′
底部链
3′
3′
5′
顶部链:
3′
3′
桥接寡
5′
顶部链
5′

图7

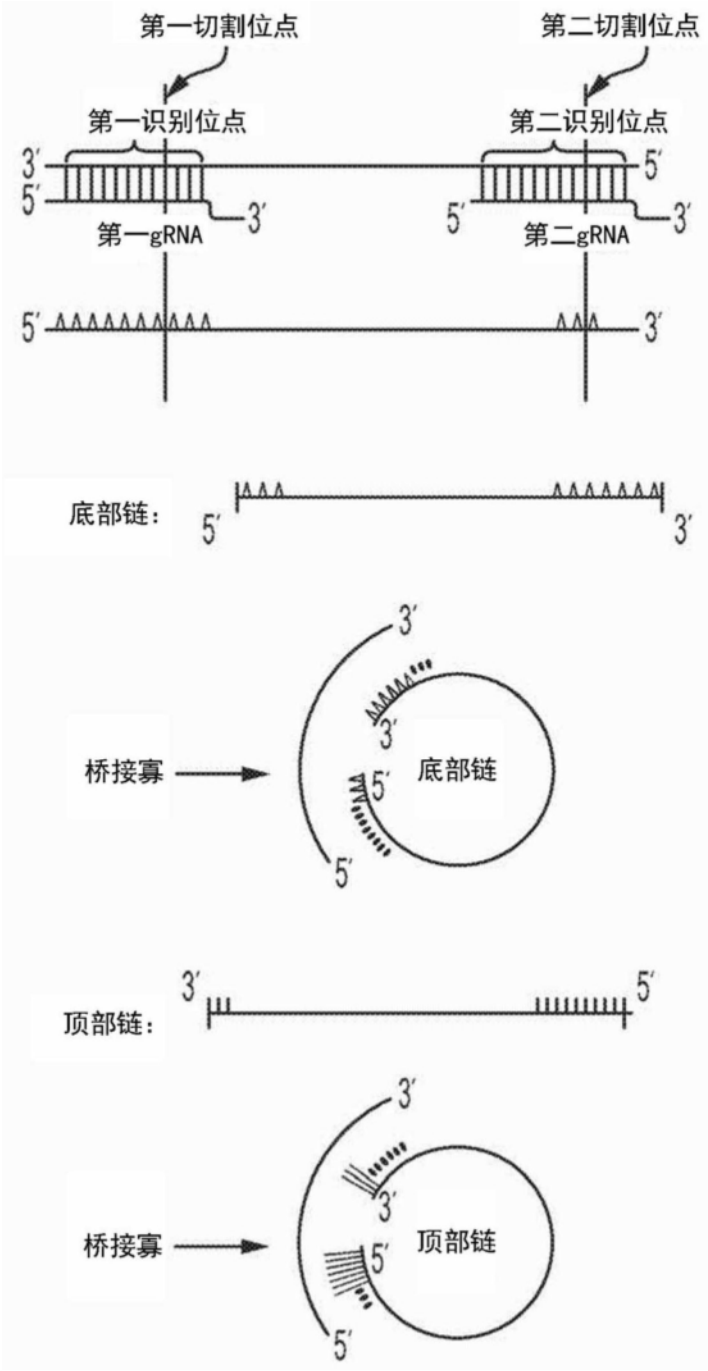

第一切割位点
第二切割位点
第一识别位点
第二识别位点
3′
5′
第一gRNA
第二gRNA
底部链：
桥接寡
底部链
顶部链：
顶部链

图8

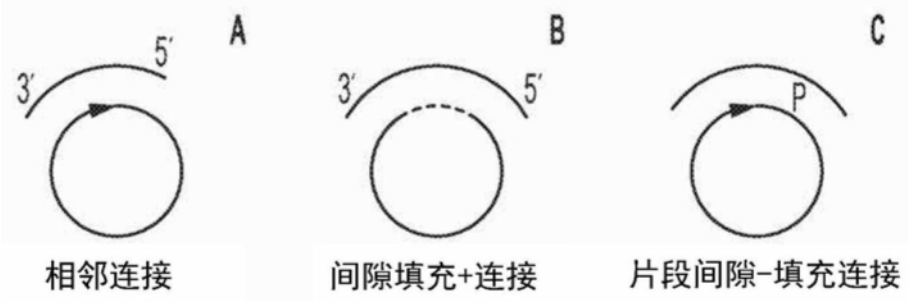

A
B
C
5′
3′
3′
5′
P
相邻连接
间隙填充+连接
片段间隙-填充连接

图9

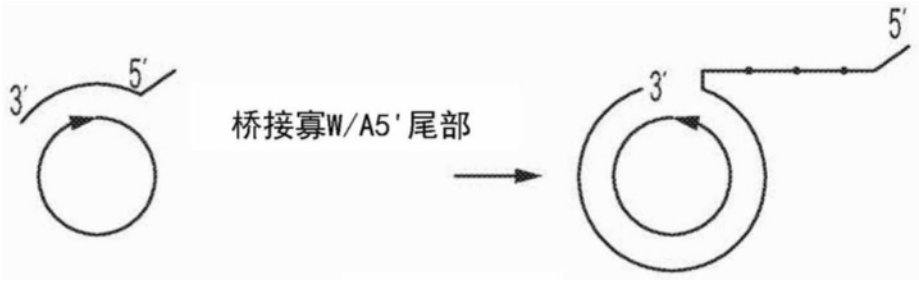

3′
5′
桥接寡W/A5′尾部
3′
5′

图10


Cas9
Cas9+引物延伸

图11