(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 112313241 A
(43)申请公布日 2021.02.02

(21)申请号 201980040610.5
(22)申请日 2019.04.17
(30)优先权数据
62/659,073 2018.04.17 US
62/767,633 2018.11.15 US
(85)PCT国际申请进入国家阶段日
2020.12.16
(86)PCT国际申请的申请数据
PCT/US2019/027788 2019.04.17
(87)PCT国际申请的公布数据
WO2019/204378 EN 2019.10.24
(71)申请人 总医院公司
地址 美国马萨诸塞州
(72)发明人 J·K·乔昂格 V·派特塔那雅克
K·佩特里 J·M·格尔克
K·E·佐佐木
(74)专利代理机构 北京林达刘知识产权代理事务所(普通合伙) 11277
代理人 刘新宇 李茂家
(51)Int.Cl.
C07H 21/04(2006.01)
C12N 9/22(2006.01)
C12Q 1/34(2006.01)
C12Q 1/6806(2006.01)
C12Q 1/6844(2006.01)

权利要求书4页 说明书21页 附图24页

(54)发明名称

核酸结合、修饰、和切割试剂的底物偏好和位点的灵敏体外试验

(57)摘要

用于进行高灵敏度体外试验来确定核酸结合、修饰和切割试剂的底物偏好和脱靶位点的方法和组合物。

CN 112313241 A

1.一种识别由酶切割、修饰或结合的双链DNA序列的方法,所述方法包括:

(i)提供已知序列的多个线性dsDNA寡核苷酸,每个寡核苷酸具有5'端和3'端,并且具有在所述寡核苷酸的3'和5'端二者处或二者附近的至少两个拷贝的特有标识符序列、和存在于多个寡核苷酸中的每一个寡核苷酸的5'和3'端的共有序列;

(ii)在足以发生切割、修饰或结合的条件下,在选自位点特异性核酸酶、DNA修饰蛋白和DNA结合结构域的酶的存在下孵育所述多个寡核苷酸;

(iii)选择并任选地富集被切割、修饰或结合的寡核苷酸;和

(iv)确定被切割、修饰或结合的所选的寡核苷酸的序列,从而识别由酶切割、修饰或结合的双链DNA序列。

2.一种识别由酶切割、修饰或结合的双链DNA序列的方法,所述方法包括:

(i)提供已知序列的初始多个线性dsDNA寡核苷酸,每个寡核苷酸具有5'端和3'端,并且具有在所述寡核苷酸的3'和5'端二者处或二者附近的两个拷贝的特有标识符序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列;

(ii)在足以发生切割、修饰或结合的条件下,在选自位点特异性核酸酶、修饰蛋白和DNA结合结构域的酶的存在下孵育所述多个寡核苷酸;

(iii)选择未被切割、修饰或结合的寡核苷酸;和

(iv)确定未被切割、修饰或结合的所选的寡核苷酸的序列;和

(v)将未被切割、修饰或结合的所选的寡核苷酸的序列与已知序列的初始多个预富集的线性dsDNA寡核苷酸的序列相比较;其中所述初始多个中的未选择的线性dsDNA寡核苷酸被识别为由酶切割、修饰或结合。

3.一种识别由碱基编辑酶(例如,将脱氧胞苷转换为脱氧尿苷的胞苷脱氨酶或将脱氧腺嘌呤转换为脱氧肌苷的腺嘌呤碱基编辑酶)修饰的双链DNA序列的方法,所述方法包括:

(i)提供已知序列的多个线性dsDNA寡核苷酸,每个寡核苷酸具有5'端和3'端,并且具有在所述寡核苷酸的3'和5'端二者处或二者附近的两个拷贝的特有标识符序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列;

(ii)在足以发生修饰的条件下,在碱基编辑酶的存在下孵育所述多个线性dsDNA寡核苷酸;

(iii)在DNA合成期间,由将被编辑的碱基对转换为规范碱基对的等同混合物的聚合酶(例如将dU:dG碱基对转换为dT:dA和dC:dG碱基对的等同混合物、或将dI:dT碱基对转换为dA:dT和dG:dC碱基对的等同混合物的尿嘧啶耐受聚合酶)扩增寡核苷酸(即,其中dATP核苷酸从dU对面并入或dCTP核苷酸从dI对面并入),使得将已经由所述碱基编辑酶修饰的寡核苷酸作为来自预处理文库的原始条码连接的序列和含有置换(例如dC->dT或dA->dG)的修饰的序列的混合物进行扩增;和

(iv)确定扩增的寡核苷酸的序列;

从而识别由所述碱基编辑酶修饰的双链DNA序列。

4.一种识别由胞苷脱氨酶碱基编辑酶修饰的双链DNA序列的方法,所述胞苷脱氨酶碱基编辑酶将胞苷转换为尿苷并在相对链上产生切口,所述方法包括:

(i)提供已知序列的多个线性dsDNA寡核苷酸,每个寡核苷酸具有5'端和3'端,并且具有在所述寡核苷酸的3'和5'端二者处或二者附近的两个拷贝的特有标识符序列、和存在于

多个寡核苷酸中的每一个寡核苷酸中的共有序列；

(ii)在足以发生修饰的条件下在碱基编辑酶的存在下孵育所述多个线性dsDNA寡核苷酸，然后在酶的存在下孵育所述多个线性dsDNA寡核苷酸来在具有尿苷核苷酸的位点处产生单链断裂(切口)，从而产生含有在相对链上的具有5’磷酸的两个切口的dsDNA寡核苷酸，从而产生突出端；

(iii)用自突出端产生5’磷酸化平末端的DNA聚合酶(例如T4 DNA聚合酶或Phusion DNA聚合酶或Phusion U DNA聚合酶)孵育所述dsDNA寡核苷酸；

(iv)用包含引物序列的双链DNA衔接子捕获所述磷酸化平末端；

(v)使用一种对所述衔接子具有特异性的引物和一种对所述共有序列主链具有特异性的引物扩增序列；

(vi)任选地通过在扩增之前或之后对较小的切割片段进行大小选择来进行额外选择；和

(iv)确定扩增的寡核苷酸的序列；

从而识别由碱基编辑酶修饰的双链DNA序列。

5.一种识别由腺嘌呤碱基编辑酶修饰的双链DNA序列的方法，所述腺嘌呤碱基编辑酶将脱氧腺嘌呤转换为脱氧肌苷并在相对链上产生切口，所述方法包括：

(i)提供已知序列的多个线性dsDNA寡核苷酸，每个寡核苷酸具有5’端和3’端，并且具有在所述寡核苷酸的3’和5’端二者处或二者附近的两个拷贝的特有标识符序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列；

(ii)在足以发生修饰的条件下在碱基编辑酶的存在下孵育所述多个线性dsDNA寡核苷酸，然后在核酸内切酶V酶的存在下孵育所述多个线性dsDNA寡核苷酸来在具有肌苷核苷酸的位点处产生单链断裂(切口)，从而产生含有在相对链上的具有5’磷酸的两个切口的dsDNA寡核苷酸，从而产生突出端；

(iii)用自所述突出端产生5’磷酸化平末端的DNA聚合酶(例如T4 DNA聚合酶或Phusion DNA聚合酶或Phusion U DNA聚合酶)孵育所述dsDNA寡核苷酸；

(iv)将所述磷酸化平末端与包含引物序列的双链DNA衔接子连接；

(v)使用一种对所述衔接子具有特异性的引物和一种对所述共有序列主链具有特异性的引物扩增序列；

(vi)任选地通过在扩增之前或之后对较小的切割片段进行大小选择来进行额外选择；和

(iv)确定扩增的寡核苷酸的序列；

从而识别由所述碱基编辑酶修饰的双链DNA序列。

6.一种识别由腺嘌呤碱基编辑酶或胞苷脱氨酶碱基编辑酶修饰的双链DNA序列的方法，所述腺嘌呤碱基编辑酶将脱氧腺嘌呤转换为脱氧肌苷并在相对链上产生切口，所述胞苷脱氨酶碱基编辑酶将胞苷转换为尿苷并在相对链上产生切口，所述方法包括：

(i)提供已知序列的多个线性dsDNA寡核苷酸，每个寡核苷酸具有5’端和3’端，并且具有在所述寡核苷酸的3’和5’端二者处或二者附近的两个拷贝的特有标识符序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列；

(ii)在来自极端嗜热古生菌(Thermococcus kodakarensis)的核酸内切酶MS

(TkoEndoMS)的存在下孵育所述多个线性dsDNA寡核苷酸来在底物DNA中的脱氨位点处诱导双链断裂(DSB)来产生以所述脱氨位点为中心的具有单链的、5个碱基对突出端的DNA片段;

(iii)用尿嘧啶DNA糖基化酶和核酸内切酶VIII处理所述DNA片段来从所述DNA片段的末端去除脱氧尿嘧啶碱基;

(iv)DNA片段的末端的末端修复和/或加A尾;

(v)将衔接子寡核苷酸(优选包含用于高通量测序的序列)连接至所述末端;和

(vi)对所述DNA片段进行测序。

7.一种在所选的gRNA或其它DNA结合结构域的存在下识别由催化失活的Cas9结合的双链DNA序列的方法,所述方法包括:

(i)提供已知序列的多个线性dsDNA寡核苷酸,每个寡核苷酸具有5'端和3'端,并且具有在所述寡核苷酸的3'和5'端二者处或二者附近的两个拷贝的特有标识符序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列;

(ii)在足以发生结合的条件下,在附着至磁珠(例如,共价结合或通过亲和手柄结合)的DNA结合结构域例如与sgRNA或其它DNA结合结构域复合的Cas9酶的存在下,孵育所述多个寡核苷酸;

(iii)通过一个或多个集合的磁珠的拉下,并在促进未结合分子解离至上清液的适当的缓冲液中洗涤,然后在促进任何结合的DNA的解离的适当的缓冲液中或在降解磁珠所结合的蛋白质并释放结合的DNA的含有例如蛋白酶K等蛋白酶的缓冲液中洗脱结合的DNA,来选择和任选地富集所结合的寡核苷酸;和

(iv)确定被切割的所选的寡核苷酸的序列,从而识别由DNA结合结构域结合的双链DNA序列。

8.根据权利要求1至7任一项所述的方法,其中所述线性dsDNA寡核苷酸包含:

(i)相对于识别的中靶位点具有一定数量以下的错配的参考基因组中的所有潜在脱靶序列的集合(类似于基因组DNA文库);

(ii)具有一定数量以下的错配的潜在脱靶位点的综合集合(类似于随机碱基置换文库);

(iii)存在于来自特定群体的变体基因组的集合中的潜在脱靶序列的文库(即,旨在反映存在于个体群体中的DNA序列变体的基因组DNA文库);或

(iv)潜在脱靶位点的其它相关特定的集合(例如,癌基因热点或来自肿瘤抑制基因的序列)。

9.根据权利要求1至7任一项所述的方法,其中:

首先将预富集的线性DNA文库成员例如在高密度寡核苷酸阵列上合成为单独的单链DNA序列;和

通过对所述共有序列进行引物作用而将所述单链DNA序列转换成双链DNA分子,任选地在自芯片释放之前或之后。

10.根据权利要求1至7任一项所述的方法,其中预富集的线性DNA文库成员表示1)相对于中靶位点具有一定数量以下的错配的参考基因组中的所有潜在脱靶序列的集合(类似于基因组DNA文库),2)具有一定数量以下的错配的潜在脱靶位点的综合集合(类似于随机碱基置换文库),3)存在于来自特定群体的变体基因组的集合中的潜在脱靶序列文库(即,旨

在反映存在于个体群体中的DNA序列变体的基因组DNA文库)，或4)潜在脱靶位点的其它相关特定集合(例如，癌基因热点或来自肿瘤抑制基因的序列)。

11.根据权利要求1至10任一项所述的方法，其中预富集的线性DNA文库成员包含1,000至10^{11}条不同的序列。

12.根据权利要求1至11任一项所述的方法，其中预富集的线性DNA文库成员包含长度为50至500bp的序列。

核酸结合、修饰、和切割试剂的底物偏好和位点的灵敏体外试验

[0001] 联邦赞助的研究或开发

[0002] 本发明借助美国政府支持在美国国防高级研究计划局(DARPA)授予的授权号HR0011-17-2-0042下做出。美国政府享有本发明的一定的权利。

技术领域

[0003] 本文提供用于进行高灵敏度体外试验以确定核酸结合、修饰和切割试剂的底物偏好和脱靶位点的方法和组合物。

背景技术

[0004] 脱靶活性是对于具有可定制化DNA结合活性的蛋白质(包括但不限于归巢核酸内切酶、锌指、转录激活因子样效应物(TALEs)和CRISPR-Cas9系统蛋白质)在临床、工业和研究环境中的安全或有效使用的主要挑战。

发明内容

[0005] 本文提供用于进行高灵敏度体外试验以确定核酸结合、修饰和切割试剂的底物偏好和脱靶位点的方法和组合物。

[0006] 本文提供用于识别由酶切割、修饰或结合的双链DNA序列的方法。所述方法包括(i)提供已知序列的多个线性dsDNA寡核苷酸,每个寡核苷酸具有5'端和3'端,并且具有在寡核苷酸的3'和5'端二者处或二者附近的至少两个拷贝的特有标识符序列(unique identifier sequence)、和存在于所述多个寡核苷酸中的每一个寡核苷酸的5'和3'端的共有序列;(ii)在足以发生切割、修饰或结合的条件下,在选自位点特异性核酸酶、DNA修饰蛋白和DNA结合结构域的酶的存在下孵育所述多个寡核苷酸;(iii)选择并任选地富集被切割、修饰或结合的寡核苷酸;和(iv)确定被切割、修饰或结合的所选的寡核苷酸的序列,从而识别由酶切割、修饰或结合的双链DNA序列。

[0007] 本文还提供用于识别由酶切割、修饰或结合的双链DNA序列的方法。所述方法包括(i)提供已知序列的初始多个线性dsDNA寡核苷酸,每个寡核苷酸具有5'端和3'端,并且在寡核苷酸的3'和5'端二者处或二者附近具有两个拷贝的特有标识符序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列;(ii)在足以发生切割、修饰或结合的条件下,在选自位点特异性核酸酶、修饰蛋白和DNA结合结构域的酶的存在下孵育所述多个寡核苷酸;(iii)选择未被切割、修饰或结合的寡核苷酸;和(iv)确定未被切割、修饰或结合的所选的寡核苷酸的序列;和(v)将未被切割、修饰或结合的所选的寡核苷酸的序列与已知序列的初始多个预富集的线性dsDNA寡核苷酸的序列相比较;其中所述初始多个中的未选择的线性dsDNA寡核苷酸被识别为由酶切割、修饰或结合。

[0008] 此外,本文提供用于识别由碱基编辑酶(例如,将脱氧胞苷转换为脱氧尿苷的胞苷脱氨酶或将脱氧腺嘌呤转换为脱氧肌苷的腺嘌呤碱基编辑酶)修饰的双链DNA序列的方法。

所述方法包括(i)提供已知序列的多个线性dsDNA寡核苷酸,每个寡核苷酸具有5'端和3'端,并且具有在寡核苷酸的3'和5'端二者处或二者附近的两个拷贝的特有标识符序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列;(ii)在足以发生修饰的条件下,在碱基编辑酶的存在下孵育多个线性dsDNA寡核苷酸;(iii)在DNA合成期间,由将被编辑的碱基对转换为规范碱基对的等同混合物(equal mixture)的聚合酶(例如将dU:dG碱基对转换为dT:dA和dC:dG碱基对的等同混合物、或将dI:dT碱基对转换为dA:dT和dG:dC碱基对的等同混合物的尿嘧啶耐受聚合酶)扩增寡核苷酸(即,其中dATP核苷酸从dU对面并入或dCTP核苷酸从dI对面并入),使得将已经由碱基编辑酶修饰的寡核苷酸作为来自预处理文库的原始条码连接的序列和含有置换(例如dC->dT或dA->dG)的修饰的序列的混合物进行扩增;和(iv)确定扩增的寡核苷酸的序列,从而识别由碱基编辑酶修饰的双链DNA序列。

[0009] 另外,本文提供用于识别由将胞苷转换为尿苷并在相对链上产生切口的胞苷脱氨酶碱基编辑酶修饰的双链DNA序列的方法。所述方法包括(i)提供已知序列的多个线性dsDNA寡核苷酸,每个寡核苷酸具有5'端和3'端,并且具有在寡核苷酸的3'和5'端二者处或二者附近的两个拷贝的特有标识符序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列;(ii)在足以发生修饰的条件下在碱基编辑酶的存在下孵育所述多个线性dsDNA寡核苷酸,然后在酶的存在下孵育所述多个线性dsDNA寡核苷酸以在具有尿苷核苷酸的位点处产生单链断裂(切口),从而产生含有在相对链上的具有5'磷酸的两个切口的dsDNA寡核苷酸,从而产生突出端;(iii)用自突出端产生5'磷酸化平末端的DNA聚合酶(例如,T4 DNA聚合酶或Phusion DNA聚合酶或Phusion U DNA聚合酶)孵育所述dsDNA寡核苷酸;(iv)用包含引物序列的双链DNA衔接子捕获磷酸化平末端;(v)使用一种对衔接子具有特异性的引物和一种对共有序列主链具有特异性的引物扩增序列;(vi)任选地通过在扩增之前或之后对较小的切割片段进行大小选择来进行额外选择;和(iv)确定扩增的寡核苷酸的序列,从而识别由碱基编辑酶修饰的双链DNA序列。

[0010] 此外,本文提供识别由将脱氧腺嘌呤转换为脱氧肌苷并在相对链上产生切口的腺嘌呤碱基编辑酶修饰的双链DNA序列的方法。所述方法包括(i)提供已知序列的多个线性dsDNA寡核苷酸,每个寡核苷酸具有5'端和3'端,并且具有在寡核苷酸的3'和5'端二者处或二者附近的两个拷贝的特有标识符序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列;(ii)在足以发生修饰的条件下在碱基编辑酶的存在下孵育所述多个线性dsDNA寡核苷酸,然后在核酸内切酶V酶的存在下孵育所述多个线性dsDNA寡核苷酸以在具有肌苷核苷酸的位点处产生单链断裂(切口),从而产生含有在相对链上的具有5'磷酸的两个切口的dsDNA寡核苷酸,从而产生突出端;(iii)用自突出端产生5'磷酸化平末端的DNA聚合酶(例如T4 DNA聚合酶或Phusion DNA聚合酶或Phusion U DNA聚合酶)孵育dsDNA寡核苷酸;(iv)将磷酸化平末端与包含引物序列的双链DNA衔接子连接;(v)使用一种对衔接子具有特异性的引物和一种对共有序列主链具有特异性的引物扩增序列;(vi)任选地通过在扩增之前或之后对较小的切割片段进行大小选择来进行额外选择;和(iv)确定扩增的寡核苷酸的序列;从而识别由碱基编辑酶修饰的双链DNA序列。

[0011] 一种识别由将脱氧腺嘌呤转换为脱氧肌苷并在相对链上产生切口的腺嘌呤碱基编辑酶或将胞苷转换为尿苷并在相对链上产生切口的胞苷脱氨酶碱基编辑酶修饰的双链DNA序列的方法,所述方法包括:(i)提供已知序列的多个线性dsDNA寡核苷酸,每个寡核苷

酸具有5’端和3’端，并且具有在所述寡核苷酸的3’和5’端二者处或二者附近的两个拷贝的特有识别序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列；(ii)在来自极端嗜热古生菌(Thermococcus kodakarensis)的核酸内切酶MS(TkoEndoMS)的存在下孵育多个线性dsDNA寡核苷酸以在底物DNA中的脱氨位点处诱导双链断裂(DSB)，以产生以脱氨位点为中心的具有单链的、5个碱基对突出端的DNA片段；(iii)用尿嘧啶DNA糖基化酶和核酸内切酶VIII处理DNA片段以从DNA片段的末端去除脱氧尿嘧啶碱基；(iv)DNA片段的末端的末端修复和/或加A尾；(v)将衔接子寡核苷酸(优选包含用于高通量测序的序列)连接至末端；和(vi)对DNA片段进行测序。

[0012] 另外，本文提供在所选的gRNA或其它DNA结合结构域的存在下识别由催化失活的Cas9结合的双链DNA序列的方法。所述方法包括(i)提供已知序列的多个线性dsDNA寡核苷酸，每个寡核苷酸具有5’端和3’端，并且具有在所述寡核苷酸的3’和5’端二者处或二者附近的两个拷贝的特有识别序列、和存在于多个寡核苷酸中的每一个寡核苷酸中的共有序列；(ii)在足以发生结合的条件下，在附着至磁珠(例如，共价结合或通过亲和手柄(affinity handle)结合)的DNA结合结构域例如与sgRNA或其它DNA结合结构域复合的Cas9酶的存在下，孵育所述多个寡核苷酸；(iii)通过一个或多个集合的磁珠的拉下，并在促进未结合分子解离至上清液的适当的缓冲液中洗涤，然后在促进任何结合的DNA的解离的适当的缓冲液中或在降解磁珠所结合的蛋白质并释放结合的DNA的含有例如蛋白酶K等蛋白酶的缓冲液中洗脱结合的DNA，来选择和任选地富集所结合的寡核苷酸；和(iv)确定被切割的所选的寡核苷酸的序列，从而识别由DNA结合结构域结合的双链DNA序列。

[0013] 在一些实施方案中，本文所述方法中使用的线性dsDNA寡核苷酸包含(i)相对于识别的中靶位点具有一定数量以下的错配的参考基因组中的所有潜在脱靶序列的集合(类似于基因组DNA文库)；(ii)具有一定数量以下的错配的潜在脱靶位点的综合集合(类似于随机碱基置换文库)；(iii)存在于来自特定群体的变体基因组的集合的潜在脱靶序列的文库(即，旨在反映存在于个体群体中的DNA序列变体的基因组DNA文库)；或(iv)潜在脱靶位点的其它相关特定集合(例如，癌基因热点或来自肿瘤抑制基因的序列)。

[0014] 在一些实施方案中，预富集的线性DNA文库成员首先例如在高密度寡核苷酸阵列上合成为单独的单链DNA序列；和通过对所述共有序列进行引物作用而将所述单链DNA序列转换成双链DNA分子，任选地在自芯片释放之前或之后。

[0015] 在一些实施方案中，预富集的线性DNA文库成员表示1)相对于中靶位点具有一定数量以下的错配的参考基因组中的所有潜在脱靶序列的集合(类似于基因组DNA文库)，2)具有一定数量以下的错配的潜在脱靶位点的综合集合(类似于随机碱基置换文库)，3)存在于来自特定群体的变体基因组的集合的潜在脱靶序列的文库(即，旨在反映存在于个体群体中的DNA序列变体的基因组DNA文库)，或4)潜在脱靶位点的其它相关特定集合(例如，癌基因热点或来自肿瘤抑制基因的序列)。

[0016] 在一些实施方案中，预富集的线性DNA文库成员包含1,000、2500、5000或10,000以上且10^6、10^7、10^8、10^9、10^{10}、或10^{11}以下条不同的序列，例如10-100k条不同的序列。

[0017] 在一些实施方案中，预富集的线性DNA文库成员包含长度为50至500、例如100至400、例如150至300bp、例如长度为200至280bp的序列。

[0018] 除非另有定义，否则本文中使用的所有技术和科学术语具有与本发明所属技术领

域的普通技术人员通常所理解的相同含义。本文描述了用于本发明的方法和材料;也可以使用本领域已知的其它合适的方法和材料。材料、方法和实施例仅是说明性的而不旨在进行限制。本文提及的所有出版物、专利申请、专利、序列、数据库条目和其它参考文献通过引用将其全部内容并入本文。在有冲突的情况下,本说明书包括定义将占主导。

[0019] 本发明的其它特征和优点将从以下详细描述和附图、以及从权利要求中显而易见。

附图说明

[0020] 图1.碱基置换文库(SEQ ID NO:1-6、1、和7-11)和基因组DNA文库(SEQ ID NO:1和12-16)之间的文库复杂度的差异的说明。对于基因组文库,中靶位点(在此实施例中为红色框)与其它大约30亿个基因组序列之间保持非常小的相似性。对于碱基置换文库,预选文库中富集了不是必须存在于基因组中但与预定靶位点相似的位点。置换由小写字母表示。

[0021] 图2.示例性方法的说明概述。通过在高密度寡核苷酸阵列上合成而生成潜在DNA底物的用户确定的集合(通常但不限于10,000至100,000个序列),然后将其制成双链。使用的序列的集合的三个潜在实例包括在人类基因组中具有六个以下的错配、在人类外显子组中具有八个以下的错配的所有序列、或具有三个以下的错配的所有可能的DNA序列的集合。然后可将双链DNA文库用于序列修饰的筛选、序列缺失的筛选、或序列修饰和/或切割的选择。黑线表示存在于文库的每个成员中的恒定序列并用于扩增步骤和生物信息处理中的引物结合位点。

[0022] 图3.可以选择靶位点文库进行切割(策略1)、或筛选靶位点文库以消除切割的序列(策略2)。黑线表示存在于文库的每个成员中的恒定序列和用于扩增步骤和生物信息处理中的引物结合位点。显示SEQ ID NO.:17-22。

[0023] 图4.通过从随机碱基置换文库中选择来富集Cas9切割位点。预定靶位点(SEQ ID NO:23)列于热图下方。每个黑色框表示在与以下列出的靶位点核苷酸相对应的位置处的特定核苷酸(在左侧表示)的丰度,黑色代表丰度最大的核苷酸/每个位置和白色表示无丰度。

[0024] 图5.通过筛选随机碱基置换文库和筛选基因组DNA文库来识别中靶序列。在底物谱文库和基因组激发文库(genome-inspired library)中的底物具有较少突变数(由较小m数表示,其中(m_d_i)->m=突变数,i=插入数,d=缺失数),其中,XmO表示不具有任何插入的m个错配,RNAmd表示在位点中的其余碱基对处具有m个错配的长度为d的靶位点缺失,DNAmi表示在位点中的其余碱基对处具有m个错配的长度为I的靶位点插入)。

[0025] 图6.显示代表性基因组DNA寡核苷酸文库的组成。根据错配和凸起(bulge)的数量列出了各基因组位点的数量。除非另有说明,否则这些文库将用于后续附图中概述的实验。

[0026] 图7.文库表征。在寡核苷酸合成、文库扩增和Illumina测序之后,显示了每个文库的一致性度量(uniformity metric)和缺失百分比(drop out percentage)。第90个百分位数(percentile)测序计数是指按照增加的读段排序时,第90个百分位数中获得的文库成员的测序读段数。90/10比率是第90个百分位数文库成员除以第10个百分位数文库成员的测序读段数的比率,并且是文库一致性的度量。缺失是指在测序的扩增的文库中未表示的序列的数量。

[0027] 图8.已知GUIDE-seq位点的富集。显示使用本文描述的方法(其实例称为ONE-seq

法)的代表性切割选择的分簇散点图。每个圆圈表示针对单个文库成员的针对给定的指导RNA选择(列在顶部)归一化至中靶序列的聚集读段计数。黑色星表示中靶文库成员。实心圆圈表示由公开的GUIDE-seq实验识别的位点。无公开的RNF2 GUIDE-seq位点。

[0028] 图9.高度富集的CIRCLE-seq位点的富集。显示使用ONE-seq法的代表性切割选择的分簇散点图。每个圆圈表示针对单个文库成员的针对给定的指导RNA选择(列在顶部)归一化至中靶序列的聚集读段计数。黑色星表示中靶文库成员。实心圆圈表示具有由公开的CIRCLE-seq实验识别的>100个读段计数的位点。无公开的RNF2 GUIDE-seq位点。

[0029] 图10.中等富集的CIRCLE-seq位点的富集。显示使用ONE-seq法的代表性切割选择的分簇散点图。每个圆圈表示针对单个文库成员的针对给定的指导RNA选择(列在顶部)归一化至中靶序列的聚集读段计数。黑色星表示中靶文库成员。实心圆圈表示具有由公开的CIRCLE-seq实验识别的10-99个读段计数的位点。

[0030] 图11.低富集的CIRCLE-seq位点的富集。显示使用ONE-seq法的代表性切割选择的分簇散点图。每个圆圈表示针对单个文库成员的针对给定的指导RNA选择(列在顶部)归一化至中靶序列的聚集读段计数。黑色星表示中靶文库成员。实心圆圈表示具有由公开的CIRCLE-seq实验识别的1-9个读段计数的位点。

[0031] 图12.维恩图显示使用中靶ONE-seq聚集读段计数的1％的截止值,识别六个SpCas9:sgRNA的62个高度富集(>100个读段)的CIRCLE-seq位点(较浅的圆圈)中的全部60个的示例性方法(较深的圆圈)。不超过1％ONE-seq截止值的2个位点不一定表示真正的脱靶序列,并且在CIRCLE-seq方法中可能是假阳性。CIRCLE-seq无法识别这些指导RNA的478个ONE-seq识别位点。

[0032] 图13.通过ONE-seq而不是GUIDE-seq或CIRCLE-seq识别的三个FANCF脱靶位点的验证结果。靶向扩增子测序是对自HEK293T细胞由ONE-seq识别的五个最高度富集的新型脱靶候选中的三个进行的,针对SpCas9:FANCF sgRNA构建体的表达的最高十分位分选HEK293T细胞。在左侧,显示包含插入缺失(已编辑的)的序列读段的总数和参考读段的总数,并显示编辑百分比和未编辑的候选脱靶序列。右侧显示来自三个单独分选和对照(未处理的)实验的单独数据。(SEQ ID NO:24-39按顺序出现)

[0033] 图14.变体文库中富集得分的再现性。在EMX1基因组脱靶文库和EMX1基因组变体脱靶文库上进行ONE-seq的选择。显示由两个文库共享的文库成员的富集得分(相对于中靶序列)。重叠的线对应于来自两个选择的相等的富集得分。

[0034] 图15.ONE-seq识别存在于群体中但不存在于参考基因组中的候选脱靶位点。显示了识别自参考基因组的脱靶候选和含有在1000个基因组群体中发现的SNP的成对脱靶候选的归一化聚集读段计数(其中1.0为中靶位点)。用实心圆圈表示存在于群体的>40％中的变体。重叠的线对应于来自两个成对文库成员的相等的富集得分。

[0035] 图16.碱基编辑器筛选策略。将针对EMX1靶位点设计的随机碱基置换文库与BE1在体外孵育,并通过具有Kapa HiFi Uracil+DNA聚合酶的PCR进行扩增,其在DNA合成期间将U:G碱基对转换为T:A和C:G碱基对的等同混合物(将dATP核苷酸从dU对面并入)。因此,在测序时,任何可由BE1修饰的文库成员都将作为来自预处理文库的原始条码连接的序列和含有C->T置换(和其它罕见置换)的修饰的序列的混合物进行测序。

[0036] 图17.碱基编辑器筛选显示含有NGG的位点的富集,并显示在靶位点(SEQ ID NO:

23)的PAM近端处的高特异性和在PAM远端处的低特异性。热图以与图4相同的方式解释。

[0037] 图18.BE3选择策略。在该策略中，靶位点文库暴露于BE3酶，并通过在具有尿苷核苷酸(通过USER)和切口(通过BE3)的位点处形成的双链断裂而富集修饰的成员。

[0038] 图19.由ONE-seq的BE3脱靶位点的富集。显示针对基因组DNA激发文库上的八个ONE-seq选择归一化的聚集读段计数(其中1.0对应于中靶位点)。仅显示得分为0.01以上(中靶富集的1%)的位点。黑色星表示中靶文库成员。实心黑色圆圈表示与Digenome-seq相比新验证的脱靶位点(ABE位点18除外，其未经Digenome-seq试验)。空心黑色圆圈表示Digenome-seq候选位点。

[0039] 图20.新识别的和验证的BE3脱靶位点。与未处理的对照相比，显示来自表达指定的BE3:sgRNA复合体的HEK293T细胞的基因组DNA的靶向扩增子测序数据。实验一式三份进行。与Digenome-seq相比，仅显示28个新识别的和验证的BE3脱靶位点。

[0040] 图21.ABE选择策略。在该实施例中，使用腺嘌呤碱基编辑器(ABE)。ABE在DNA中产生A->I变化。除了使用不同的核酸内切酶即核酸内切酶V在DNA中的脱氧肌苷位点处产生切口以外，此处我们用于确定ABE的脱靶位点的方法与实施例3中所使用的方法相似。

[0041] 图22.碱基置换文库上的ABE选择。热图以与图4相同的方式解释。来自选择的数据表明NGG PAM的富集，也表明相对于图4中的SpCas9切割选择和图8中的BE3选择，富集在5位处含有A的序列，表明在ABE的某些编辑窗口中需要A来显示活性。(SEQ ID NO:23)

[0042] 图23.由ONE-seq的ABE7.10脱靶位点的富集。显示基因组DNA激发文库中八个ONE-seq选择的归一化聚集读段计数(其中1.0对应于中靶位点)。黑色星表示中靶文库成员。实心圆圈表示已验证的脱靶位点。空心黑色圆圈表示在验证研究中测序的脱靶候选。

[0043] 图24.验证的ABE脱靶位点。与未处理的对照相比，显示来自表达指定的ABEmax:sgRNA复合体的HEK293T细胞的基因组DNA的靶向扩增子测序的数据。实验一式三份进行。

[0044] 图25.来自表明TkoEndoMS的核酸内切酶活性对体外G:U DNA错配的特异性的实验的毛细管电泳数据。用纯化的BE蛋白质和可变sgRNA孵育800个碱基对PCR扩增子2小时以诱导位点特异性脱氨作用。纯化后，用纯化的TkoEndoMS蛋白质孵育脱氨的PCR扩增子7分钟以诱导G:U错配处的双链断裂。然后通过毛细管电泳按大小分离DNA并成像。

[0045] 图26.通过拉下富集结合位点的概述。在该方法中，用潜在脱靶位点的文库孵育dCas9包被的磁珠。未结合的文库成员被洗涤至上清液中，并且结合的文库成员通过用蛋白酶K消化磁珠结合的蛋白质而洗脱。所得的洗脱文库可以扩增并进行额外轮(round)的拉下、或进行通过高通量测序的分析。

[0046] 图27.拉下条件可以区分中靶位点和脱靶位点。用dCas9:EMX1 sgRNA包被的磁珠对不同长度的三个双链DNA的混合物进行结合位点拉下。中靶位点存在于280个碱基对DNA上，和两个脱靶位点(OT2或OT4)之一存在于220个碱基对DNA上。混合物中也存在既不含有中靶位点也不含有脱靶位点的第三200bp(“随机”)DNA。将富集的DNA在QIAxcel上跑胶。泳道A3表示大小梯状条带(size ladder)。泳道A1显示280个碱基对中靶位点的选择性拉下，但无OT2位点或200bp DNA的拉下。泳道A2显示，OT4位点在所述方法中仍可被结合，表明用于拉下的条件可以富集可由dCas9:EMX1 sgRNA结合的脱靶序列。(SEQ ID NO:23、40-41依次显示)

[0047] 图28.通过拉下富集基因组DNA激发文库。在FANCF文库中在50ug/ml肝素的存在下

进行的拉下导致中靶位点(黑色星)富集最丰富的拉下后文库成员。

[0048] 图29.I-PpoI的选择后文库组成。显示具有预期的I-PpoI靶位点的至少1%的归一化读段计数的序列的序列标志(sequence logo)。自5'至3'端的位点中的位置显示在横轴上,字母的堆叠的高度表示各个位置的信息含量(以位为单位)。各个单独的核苷酸的高度强调该核苷酸对该位置的信息含量的相对贡献。位置2、13和14是最高说明的(信息含量最多)。位置15是最少说明的(信息含量最低)。

具体实施方式

[0049] 理解DNA结合结构域的中靶和脱靶活性的体外/生化策略通常分为两种类型(图1):在第一种类型中,对相关特定系统中的DNA序列的集合(例如,人类基因组)查找脱靶切割事件。利用此策略的方法的实例包括CIRCLE-seq(Tsai等人,Nat Meth.14:607(2017)]、SITE-seq(Cameron等人.Nat Meth.14:600(2017))、和Digenome-seq(Kim等人.Nat Meth.12:237(2015))。这些全基因组方法的范围限于在用于研究的特定基因组DNA中存在的脱靶位点。与此相反,在策略的第二种类型中,通过全面查找其中某些碱基位置被所有潜在可选碱基随机置换(而不是像第一种策略的情况中那样只限于有限的碱基置换的集合)的结合位点的文库而以更无偏见的方式分析DNA结合/修饰蛋白的底物偏好。已经对各种核酸酶(ZFN(Pattanayak等人.Nat Meth.8:765(2011)]、TALEN(Guilinger等人.Nat Meth.11:429(2014))、和CRISPR-Cas9(Pattanayak等人.Nat Biotech.31:839(2013)))在体外进行"基因组DNA"和"随机碱基置换文库"方法,并提供对这些核酸酶的生化功能和特异性的见解。

[0050] 两种研究脱靶活性的策略都具有局限性,影响它们识别真正的脱靶位点的能力。在全基因组选择中,必须从数十亿个未切割的其它位点的背景中富集数十至数百个切割的脱靶位点(人类基因组的长度约为30亿个碱基对,因此含有约60亿个待分析的位点)。例如,由于富集方法和测序结果中的噪音,CIRCLE-seq法仅限于相对于中靶位点具有不超过六个错配的位点的检测,其仅表示分析中存在的基因组材料的约0.002%。尽管某些方法,例如Digenome-seq,依赖于核酸酶处理的DNA文库的大量的过度测序,但诸如CIRCLE-seq和GUIDE-seq等方法通常并入编辑的序列的富集步骤。该富集步骤可以在细胞(GUIDE-seq)中或体外(CIRCLE-seq)进行。尽管CIRCLE-seq法比用于脱靶筛选的其它方法实质上更灵敏,但CIRCLE-seq法对于各个试验样品需要基因组DNA的非常大的输入(25μg)。

[0051] 无偏见的碱基置换文库中的体外选择受限于文库大小(可以实际试验的序列的集合)。例如,SpCas9靶位点含有22个潜在指定的碱基对(20个来自与指导RNA的杂交和两个来自PAM序列)。为了分析在所有位置具有碱基置换的所有可能组合的所有潜在靶位点,将需要产生和查找至少4^{22}～10^{13}个独特的DNA分子,两者使用当前技术都不可能实现。例如,文库构建方法当前限于产生10^{11}-10^{12}个独特的DNA分子。此外,即使文库构建方法得到改善,对10^{12}个DNA分子进行测序是不可行的。为了克服这一限制,传统上使用掺杂的寡核苷酸合成来生成具有遵循二项分布的碱基置换的位点的文库,使得中靶位点以比具有单个突变的文库中的各个变异位点更多的拷贝存在,其各自以比具有双变异位点的文库中的各个变异位点更多的拷贝存在,依此类推。因此,由这些随机碱基置换文库进行的选择受限于以下事实:1)无法生成完全无偏见的文库(即,它们严重偏向预期的中靶位点序列)和2)无法生成

一致表示潜在序列空间的文库。此外，使用来自特定文库试验的输出来预测或识别基因组序列中的脱靶位点通常需要外推法(Sander等人.Nucleic Acids Res.41:e181(2013))，这是因为无法保证所有相关的基因组序列都包括在预选择(限于10^{12}条序列，其对应于六个或七个置换)或选择后文库(由测序能力限于10^{7-8}条序列)中。

[0052] 识别DNA结合、修饰或切割位点的方法

[0053] 本文中，我们提供了改进的方法(图2)，其可以识别DNA修饰蛋白质/蛋白质复合体的中靶和脱靶结合、修饰、或切割位点(包括但不限于：融合至效应子结构域的dCas9、基于Cas9的碱基编辑器、或活性Cas9蛋白)，并且其克服“基因组DNA”和“随机碱基置换文库”两种方法的弊端。使用该方法，可以通过高密度寡核苷酸合成来生成由特定用户指定的序列组成的线性DNA的预富集文库，然后查找可以由序列特异性蛋白质或蛋白质复合体结合、修饰或切割的那些序列。最低限度地，该方法允许识别序列，所述序列是其修饰作用可导致核酸的序列修饰、结合或切割的任何试剂的潜在底物。

[0054] 预富集的线性DNA文库成员最初在高密度寡核苷酸阵列上合成为单个单链DNA序列，每个序列具有特有的标识符/条码，其在寡核苷酸的两侧均存在/复制(图2)。合成的寡核苷酸自芯片释放并通过针对存在于在芯片上合成的所有DNA分子中的共有序列进行引物作用而转换为双链DNA分子。然后将该合并的文库与目标的位点特异性核酸酶、修饰蛋白、或DNA结合结构域一起孵育，并以选择形式富集切割、修饰或结合的序列(参见实施例1、3和4)或筛选修饰(参见实施例2)。然后，可以从最初位于这些位点两侧和现在分成两个分子的相同条码中的一者来重建切割位点的DNA序列。

[0055] 可以指定合成的分子来表示：1)在相对于中靶位点具有一定数量以下的错配的参考基因组中的所有潜在脱靶序列的集合(类似于基因组DNA文库)，2)具有一定数量以下的错配的潜在脱靶位点的综合集合(类似于随机碱基置换文库)，3)存在于来自特定群体的变体基因组的集合的潜在脱靶序列的文库(即，旨在反映存在于个体群体中的DNA序列变体的基因组DNA文库)，或4)潜在脱靶位点的其它相关特定集合(例如，癌基因热点或来自肿瘤抑制基因的序列)。该策略对于构建这些文库具有重要优势。对于随机碱基置换文库，确定的置换数量内的所有序列都可以等同地表示并且可以使用当前的下一代测序方法容易地采样。对于基因组或外显子组DNA文库，仅包括最可能相关的位点(例如，具有六个或更少的置换的所有潜在的脱靶位点)，其消除由不是底物的位点的约99.998％造成的噪音。重要的是，由于此方法导致DNA(或RNA)结合蛋白的潜在脱靶位点的富集的集合的双链DNA(或RNA)文库的生成，其可用于确定不仅是核酸酶、还可以是结合或修饰核酸的其它蛋白质的特异性和脱靶位点，所述其它蛋白质包括但不限于可定制化碱基编辑酶(Komor等人.Nature 533:420,2016,Gaudelli等人.Nature.551:464(2017))、转录激活子(Mali等人,Nat Biotech.31:833(2013)、Chavez等人.Nat Meth.12:326(2015))、转录抑制子(Bikard等人,Nucleic Acids Res,41:7429(2013)、Thakore等人.Nat Meth.12:1143(2015))、和表观基因组编辑酶(Zentner和Henikoff.Nat Biotech.33:606(2015)中的综述)。

[0056] 确定在特定识别位点两侧的相同条码的能力表示超越先前体外分析方法(US专利9,322,006、US专利9,163,284)的重大进步，这是因为文库成员的序列被编码在DNA池的各个单独成员的至少三个位置上。当试图确定其中靶序列被修饰的DNA修饰活性(例如碱基编辑)时，这种信息的冗余特别有利。即使文库成员本身的实际DNA序列被修饰，也可以从两侧

的条码中包含的信息内容中获得原始序列信息。两个条码与识别位点中的信息的冗余还允许在每文库成员以单个拷贝存在的潜在切割位点上进行核酸内切酶切割选择(或成对碱基修饰+切割选择),与多个拷贝相反(美国专利9,322,006,美国专利9,163,284)。如果没有当前的条码策略,则在识别位点内被切割的文库成员序列不可重组,因为切割在空间上将切割位点的两侧分开(上图,右下,蓝色区域)。

[0057] 实施例

[0058] 在以下实施例中进一步描述本发明,这些实施例不限制权利要求中描述的本发明的范围。

[0059] 用于以下实施例的靶位点:

[0060]

靶名称	序列 (5' -> 3')	SEQ ID NO:
EMX1	GAGTCCGAGCAGAAGAAGAAGGG	22
RNF2	GTCATCTTAGTCATTACCTGAGG	42
FANCF	GGAATCCCTTCTGCAGCACCTGG	43
HBB	TTGCCCCACAGGGCAGTAACGG	44
HEK2 (HEK293_2)	GAACACAAAGCATAGACTGCGGG	45
HEK3 (HEK293_3)	GGCCCAGACTGAGCACGTGATGG	46
HEK4 (HEK293_4)	GGCACTGCGGCTGGAGGTGGGGG	47
ABE14 (ABE_14)	GGCTAAAGACCATAGACTGTGGG	48
ABE16 (ABE_16)	GGGAATAAATCATAGAATCCTGG	49
ABE18 (ABE_18)	ACACACACACTTAGAATCTGTGG	50
VEGFA3 (VEGFA_3)	GGTGAGTGAGTGTGTGCGTGTGG	51

[0061] 实施例1:由SpCas9和SpCas9-HF1的DNA切割选择

[0062] 在该实施例中,对由针对人EMX1基因的中靶位点设计的指导RNA(gRNA)(以下称为EMX1 gRNA和EMX1靶位点)工程化的SpCas9核酸酶而设计的随机碱基置换文库和来自人参考基因组的潜在EMX1 gRNA脱靶位点的文库进行选择以用于由SpCas9或SpCas9-HF1切割。

[0063] 在该实施例中(图3),可以采用选择(策略1)和筛选(策略2)两者。在策略1中,合并的文库包含约50,000个条码的文库成员(含有在EMX1 SpCas9中靶位点的三个错配内的所有可能序列的随机碱基置换文库,或者含有来自EMX1SpCas9中靶位点的六个错配内的hg19人参考基因组的所有可能序列的基因组DNA激发文库-有关文库的详情,参见方法部分。)

[0064] 使用策略1在SpCas9:sgRNA:DNA文库(EMX1靶位点)的比例为1:1:1的随机碱基置换文库中进行的选择表明可被切割的序列的富集(图4)。靶位点中的位置在横轴上(具有以下列出的中靶碱基)。文库中可能的碱基(置换或中靶)表示在纵轴上。合并数据并总结至热图中,其中较深的黑色矩形表示含有来自纵轴的对应的碱基的选择后文库中的较大比例的位点。作为原理的证明,该热图与先前的研究一致,所述研究表明NGG PAM序列的N未被指定并且其在靶位点的PAM远端的特异性低于其在PAM近端的特异性。

[0065] 使用策略2对随机碱基置换文库进行的筛选产生了相似的结果(图5)。在底物谱文库和基因组激发文库中的底物具有较少突变数(由较小m数表示,其中(m_d_i)->m=突变数,i=插入数,d=缺失数),其中Xm0表示不具有任何插入的m个错配,RNAmd表示在位点中的其余碱基对处具有m个错配的长度为d的靶位点缺失,DNAmi表示在位点中的其余碱基对

处具有m个错配的长度为I的靶位点插入)。

[0066] 基因组文库通常由相对于中靶序列具有零至六个错配的hg19参考人类基因组中的所有潜在脱靶位点组成,四个以下的错配结合一个或两个核苷酸的DNA凸起,和四个以下的错配结合一个核苷酸的RNA凸起,和三个以下的错配和两个核苷酸的RNA凸起(图6)。对预选择文库测序以评价质量度量(图7),表明低缺失率(0.20%以下)和高一致性(90/10比率>-2)。据我们所知,这些度量尚未针对其它特异性方法进行计算,因此不可直接比较。

[0067] 使用策略1(称为ONE-seq)和基因组DNA激发文库对具有相对少的预期的脱靶序列的六个非混杂指导RNA(HBB、RNF2、HEK2、HEK3、FANCF和EMX1)进行选择。中靶序列(图8,黑色星)是所试验的六个非混杂指导RNA中成千上万个中最丰富的或前3丰富的文库成员。总结六个非混杂指导RNA,ONE-seq富集所有163个GUIDE-seq识别的脱靶位点(图8,实心圆圈),选择后的读段计数范围为中靶序列的11%至120%。该方法还富集高度富集的CIRCLE-seq位点(此处定义为具有>100个序列读段的那些,图9),并适当地富集较小程度的中等富集的CIRCLE-seq位点(10-99个读段,图10),或低富集的(1-9个读段,图11)。如果此处所述的ONE-seq方法中的中靶富集的截止值为1%,则ONE-seq识别62个高度富集的CIRCLE-seq位点中的60个(图12),而CIRCLE-seq无法识别478个高度富集的ONE-seq候选。值得注意的是,未由ONE-seq高度富集的两个高度富集的CIRCLE-seq位点可能是CIRCLE-seq方法的假阳性。通过在SpCas9:FANCF sgRNA表达的最高十分位数中分选HEK293T细胞,表明新的且未由GUIDE-seq或CIRCLE-seq识别的ONE-seq位点的验证(图13)。这些结果表明,此处描述的方法至少与现有方法一样灵敏,并且可能更灵敏。

[0068] 另外,该方法可以推广至核酸序列的任何文库/特定的集合。例如,使用来自1000个基因组项目的公开数据,在EMX1基因组脱靶位点文库中进行ONE-seq选择,所述文库说明群体规模上天然存在的序列变异。在此示例性文库中,包括了来自参考hg19人类基因组集合的所有序列,这些序列在原始EMX1文库中(图6)并且在1000个基因组数据库中含有SNP。此外,还包括了含有SNP的序列作为其他文库成员,以说明个体可能具有参考基因组中未含有的脱靶序列的可能性。在该含有SNP的EMX1文库中进行的ONE-seq切割选择提供存在于参考hg19基因组中的脱靶候选的再现富集(图14)。变体文库中的ONE-seq切割选择表明对存在于EMX1指导RNA的脱靶位点的候选集合的群体中的成千上万个变体的评价,识别区别地富集的数个(图15,黑色圆圈)。实施例2:具有BE1的碱基编辑器筛选

[0069] 在该实施例中(图16),筛选策略用于识别由BE1酶产生的碱基修饰(Komor等人.Nature 533:420,2016),该修饰在DNA的特定窗口中典型地产生C->U改变。

[0070] 将遵循上述方案的具有BE1的碱基编辑器筛选应用于EMX1靶位点并且底物谱文库产生预期的耐受的脱靶位点的谱的富集(图17)。

[0071] 实施例3:具有BE3的碱基编辑器选择

[0072] 在该实施例中(图18),使用选择策略来富集由BE3酶修饰的位点。可由BE3酶识别的文库成员(Komor等人.Nature 533:420,2016)应表现C->U修饰和相对链上的切口两者。USER酶(NEB)用于通过由切口代替dU核苷酸来实现作为BE3底物的文库成员的双链切割。因此,所得的修饰的文库成员将含有在相对链上的具有5'磷酸的两个切口,并与可钝化这些DNA突出端的DNA聚合酶(例如:T4 DNA聚合酶或Phusion DNA聚合酶)一起孵育。使用一种对衔接子具有特异性的引物和一种对文库骨架具有特异性的引物的扩增/选择之前,用双链

DNA衔接子捕获所得的磷酸化平末端(如实施例1)。通过在扩增之前或之后对较小的切割片段进行大小选择,可以获得额外的选择严格性。

[0073] 使用这种方法,我们对八个靶位点用基因组DNA激发文库检查BE3靶向,包括先前由Digenome-seq试验的所有七个BE3靶(Kim等人.Nat.Biotech.35:475,2017)。ONE-seq选择结果显示,对于所有八个选择,预期靶位点对成千上万个文库成员的前13个的富集(图19,黑色星)。在八个选择中的三个,预期的靶位点是最丰富的位点。由Digenome-seq先前验证的所有42个脱靶位点存在于富集的选择后文库中(图19,空心黑色圆圈),并且42个中的40个位于各个选择的前61个位点之中。为了进一步验证我们的ONE-seq结果,我们自人HEK293T细胞中扩增并测序来自每个选择的约20-40个高名次位点。我们的结果表明,有28个经验证的BE3脱靶位点未由Digenome-seq识别为候选(图19,实心黑色圆圈)。28个新位点中的6个在细胞中具有大于1%的编辑百分比(图20),高达23.9%,表明ONE-seq不仅检测弱脱靶位点还检测高频脱靶位点的更高水平的灵敏度。

[0074] 实施例4:具有ABE的碱基编辑器选择

[0075] 在该实施例中(图21),我们使用腺嘌呤碱基编辑器(ABE;Gaudelli等人.Nature.551:464(2017))与EMX1 gRNA和碱基置换谱文库和基因组DNA文库进行选择。ABE是融合至可催化脱氧腺苷至脱氧肌苷的转换的蛋白质结构域的sgRNA指导的Cas9切口酶。在该实施例中,预选择文库的双链切割以两个步骤完成(图21)。首先,用ABE酶和指导RNA孵育导致与指导RNA杂交的识别的文库成员的链的切口形成。其次,利用在文库成员中对于可由ABE活性引起的脱氧肌苷的3'产生切口的酶,即核酸内切酶V的随后孵育,导致在非杂交DNA链上形成切口,导致具有突出端的双链断裂。随后用DNA聚合酶填充双链断裂导致平末端的形成,可以根据实施例1中对Cas9核酸酶的描述选择平末端。

[0076] 碱基置换文库的选择表明具有NGG的底物的富集。此外,正如预期的那样,该实验(图22)表明在靶位点的位置5处具有A的底物的富集(其中1是距PAM最远的碱基对),反映ABE对修饰的偏好为A与PAM的距离比典型EMX1靶位点中存在的更远。值得注意的是,在选择后文库中的100个最丰富的序列中,95个在位置5处具有A。这些结果表明我们的策略对富集和识别ABE的脱靶位点起效。

[0077] 表1.在ABE选择中在位置5处具有A的序列的富集。

[0078]

选择后文库成员的前五个核苷酸	在前 100 个最丰富的选择后文库成员中观察到的次数
GAGTA	83
AAGTA	12
GAGTC (典型的前五个核苷酸)	3
GAAGT	1
GGAGT	1

[0079] 我们还对EMX1基因组DNA文库(表2)进行了上述选择,其表明EMX1中靶位点的富集(突出显示;第96个最丰富的选择后文库序列)和具有最高脱靶识别性的EMX1脱靶位点(粗体和星号;第9个最丰富的选择后文库序列)。

[0080] 表2.选择后文库中前96个最丰富的位点,用于在潜在EMX1脱靶位点的基因组DNA文库中进行ABE选择。

[0081]

染色体	位置	靶	SEQ ID NO:
chr4	33321459	GTACAGGAGCAGGAGAAGAATGG	52
chr17	72740376	CAAACGGAGCAGAAGAAGAAAGG	53
chr10	58848711	GAGCACGAGCAAGAGAAGAAGGG	54
chr10	128080178	GAGTACAAGCAGATGAAAAACGG	55
chr6	99699155	GAGTTAGAGCAGAGGAAGAGAGG	56
chr7	141972555	AAGTCCGGGCAAAAGAGGAAAGG	57
chr19	24250496	GAGTCCAAGCAGTAGAGGAAGGG	58
chr11	111680799	CAGTAGTGAGCAGAAGAAGATAGG	59
chr5	**45359060***	**GAGTTAGAGCAGAAGAAGAAAGG**	60
chr7	17446431	GTCCAAGAGCAGGAGAAGAAGGG	61
chr12	106646073	AAGTCCATGCAGAAGAGGAAGGG	62
chr15	22366604	GGAGTAGAGCAGAGGAAGAAGGG	63
chr10	109561613	GGAACTGAGCAAAAGAAGATAGG	64
chr11	62365266	GAATCCAAGCAGAAGAAGAGAAG	65
chr2	21489994	GCGACAGAGCAGAAGAAGAAGGG	66
chr1	234492858	GAAGTAGAGCAGAAGAAGAAGCG	67
chr2	218378101	GAGTCTAAGCAGGAGAATAAAGG	68
chr18	32722283	TGTCCAGAGCAGATGAAGAATGG	69
chr22	22762518	GAACATGAGCAGAAGAAGAGGAG	70
chr11	34538379	AGGCCAGAGCAAAAGAAGAGAGG	71
chr11	106142352	GTACAAGAGCAGGAGAAGAAGGG	72
chr15	91761953	GAGTCAGGGCAGAAGAAGAAAAT	73
chr4	87256685	GAGTAAGAGAAGAAGAAGAAGGG	74
chr4	21141327	AAGCCCGAGCAGAAGAAGTTGAG	75
chr8	128801241	GAGTCCTAGCAGGAGAAGAAGAG	76
chr7	106584579	GAGGGGAGCAAAAGAAGGAGGG	77
chr1	117139004	CAGGGAGAGCAAAAGAAGAGAGG	78
chr1	231750724	GAGTCAGAGCAAAAGAAGTAGTG	79
chr15	44109746	GAGTCTAAGCAGAAGAAGAAGAG	80
chr21	23586410	CAGGGAGAAGAAGAAGAAGGG	81
chr7	2127682	GAGTTAGAGAAGAAGAAGACTGG	82
chr10	98718174	ACAATCGAGCAGCAGAAGAATGG	83
chr1	221020698	GAGTAGGAGCAGATGAAGAGAGG	84
chr9	115729750	CAGTATGAGCAAAAGAAGAAAGA	85
chr11	102753237	GAGTCCATACAGAGGAAGAAAAG	86

[0082]

染色体	位置	靶	SEQ ID NO:
chr1	48581991	GAATGAGCAAAAGAAGAAAGC	87
chr12	73504668	GAGTTAGAGCAGAAAAAAAATGG	88
chr1	184236226	AATACAGAGCAGAAGAAGAATGG	89
chr11	119322554	TAGTGAGCAGAAGAAGAGAGA	90
chr1	151027591	TTCTCCAAGCAGAAGAAGAAGAG	91
chr11	68772640	GAGTCCATACAGGAGAAGAAAGA	92
chr2	9821536	AGGTGGGAGCAGAAGAAGAAGGG	93
chr2	54284994	AAGGCAGAGCAGAGGAAGAGAGG	94
chr1	99102020	GAGGCACAAGCAAAAGAAGAAAAG	95
chr19	1438808	GAAGTAGAGCAGAAGAAGAAGCG	96
chr2	**73160981**	**GAGTCCGAGCAGAAGAAGAAGGG**	22

[0083] 突出显示的两个序列是星号标注的最活跃的切割脱靶位点(chr5:45359060)和中靶位点(chr2:73160981)。由于在编辑窗口中的更有利的位置存在A,因此可以预期脱靶位点将在选择中更丰富。

[0084] 我们还对设计用于识别六个指导RNA的脱靶序列的基因组DNA文库另外进行上述选择(图23)。将修饰的ONE-seq选择方案应用于六个ABE靶显示对于试验的五个非混杂指导,预期的中靶位点富集选择后文库的前3个(HEK4是已知的混杂指导RNA)。通过来自人HEK293T细胞的DNA的扩增子测序来验证,该人HEK293T细胞分别由合适的最高候选位点的ABE7.10:sgRNA对(每个选择中各约20个)转染,从而在六个靶位点中识别总计12个确证的细胞脱靶位点。该集合包括对由ONE-seq和EndoV-seq(Liang等人Nature Communications.10:67(2019))或Digenome-seq(Kim等人Nature Biotechnology.37:430(2019))试验的两个指导RNA识别的三个验证的脱靶位点、以及未被这些方法的任一种识别为潜在候选的九个新验证的脱靶位点。12个位点中的9个具有低于1%的脱靶修饰率或仅显示单个核苷酸置换的证据,这两者之一可能是由于测序误差引起的,尽管对测序读段进行了严格的质量过滤(成对读段中的所有位置都必须具有>Phred 30的质量得分)并一式三份地进行验证。为了提高我们对这些位点是真正的脱靶位点的信心,我们用由表达ABEmax、ABE7.10的密码子优化版本和GFP的质粒转染的细胞进行第二轮验证实验,并将细胞分选以富集GFP的最高十分位数,并因此表达ABE(图24)。分选后立即进行基因组DNA提取,无需进一步扩展。在分选的验证集合中,中靶修饰的频率范围与未分选的验证集合中的31%-56%相比,为61%-94%。未分选的验证集合中的所有12个脱靶在分选的验证集合中以较高的频率被修饰,确认它们是真正的脱靶,并且在小于1%的修饰频率下识别另外五个脱靶位点。一个相对于中靶位点含有单个错配的ABE_14脱靶位点在分选验证中,DNA的85%被修饰(未分选验证中为18%),表明一些ABE脱靶位点可以以高频率被修饰。

[0085] 实施例5:使用在已被修饰的位置处产生双键断裂的酶的具有ABE或BE3的碱基编辑器选择

[0086] 在该实施例中,通过TkoEndoMS蛋白的作用,可以使含有脱氧肌苷的修饰的文库成员具有平的、双链末端(Ishino等人,Nucleic Acids Res.44:2977(2016))。TkoEndoMS可用于在由ABE的dA->dI编辑所导致的dI:dT碱基对处产生双链断裂。如果使用没有切口活性的碱基编辑酶,则将具有双链断裂的DNA进行与实施例1相同的下游步骤,将衔接子与磷酸化

的平末端DNA连接。如果使用具有切口活性的碱基编辑酶，由产生平末端的DNA聚合酶(例如T4或Phusion)进行的末端修饰，例如实施例4和5，用于允许切割文库成员的两侧的富集。

[0087] 我们已经表明TkoEndoMS还可在由dC->dU编辑(在该实施例中由BE1)导致的dG:dU错配的碱基对处产生双链断裂，表明其对BE1、BE3和其它在DNA结合后引起dC->dU改变的酶的额外适用性(参见USSN 62/571,222和图25)。这强烈表明，我们可以在我们合成的DNA位点文库上使用TkoEndoMS以识别由诱导dC至dU编辑的各种碱基编辑器导致的脱靶碱基编辑，无论是否还存在Cas9诱导的切口。

[0088] 实施例6:由拉下的DNA结合位点的富集

[0089] SELEX(由指数富集的配体的选择性进化)已被用于确定DNA结合结构域的DNA结合特异性(最初由Oliphant等人，Mol Cell Biol.9:2944,1989)。在SELEX方法中，对随机化的DNA序列的文库进行多轮拉下，并用目标的固定化DNA结合结构域进行富集，以识别初始池中可与目标DNA结合的序列。SELEX方法已被应用于ZFN(Perez等人，Nat Biotech.26:808(2008))和TALEN(Miller等人Nat Biotech.29:143(2011))的锌指和TALE部分，然而，没有关于Cas9蛋白的SELEX研究的报道。我们推测SELEX对Cas蛋白的研究是困难的，因为需要从大文库中选择性富集22个碱基对的靶位点，该文库必须含有>10^{13}个特有分子或最少10^{12}个分子，对应于20个碱基对的靶位点，如果NGG PAM是固定的。

[0090] 在该实施例中，我们利用我们的预选择文库预富集最可能由给定的Cas9:sgRNA复合体(或具有可预测结合基序的其它DNA结合结构域)结合的位点的优势。我们通过在预富集的文库中进行连续轮的DNA拉下实验来评价Cas9 DNA的结合偏好和特异性(图26)。这是通过将灭活的Cas9(dCas9)栓至磁珠来实现的。为了使dCas9与磁珠化学结合，我们采用带有称为SNAP标签的Cas9蛋白。具有SNAP标签的蛋白可以与如磁珠等载有苄基鸟嘌呤的底物分子共价结合。我们设想用磁珠结合的SNAP标记的dCas9孵育任一类型的寡核苷酸文库，并通过磁珠捕获结合序列并洗去未结合的序列来富集对Cas9具有高结合亲和力的DNA底物。通过扩增洗脱的文库成员并将所得的富集的DNA文库用作基于磁珠的选择的起始文库，可以在多个循环中重复此过程。使用该方法，在单个循环中，我们已证实与脱靶位点相比，可导致中靶位点的由dCas9:EMX1 sgRNA的选择性拉下的条件(图27)。此外，FANCF基因组DNA激发文库的拉下导致相对于其它位点的中靶位点的最大富集(图28)。Cas9结合的中靶位点的详细知识对于机械研究改善基因工程化Cas9变体例如高保真Cas9的性能特别有价值。重要的是，具有有限的效应子结构域和DNA结合结构域的相互依赖性的Cas9融合蛋白(或具有其它DNA结合结构域的融合蛋白)的脱靶模式可以主要由融合蛋白的DNA结合特性来确定。因此，对预富集的寡核苷酸文库进行DNA拉下实验可以有助于我们对融合蛋白脱靶分布的理解。此外，通过在具有有限复杂度的文库中进行Cas9(或其它DNA结合蛋白结构域)的DNA结合研究，可以获得具有小背景噪声的高质量结合数据，这些数据随后可用于推断和预测更复杂文库例如细胞的基因组的结合。

[0091] 实施例7:归巢核酸内切酶选择

[0092] 归巢核酸内切酶，例如I-PpoI，代表一组天然存在的核酸酶，其碱基识别基序比大多数限制酶更长。尽管归巢核酸内切酶(也称为大范围核酸酶(meganuclease))不具有可容易地被再工程化的特异性，如果它们靶向目标的基因组序列，它们可具有研究、商业或临床用途。在这里，我们显示我们可适应我们的体外选择以分析I-PpoI归巢核酸内切酶的特异

性谱。我们生成了潜在I-PpoI脱靶的无偏见的文库，包括具有3个以下的错配和单个DNA/RNA凸起的所有位点。I-PpoI文库含有15533个成员。I-PpoI选择富集15533个文库成员中的501个(表3)，而预期的中靶位点排名接近选择的顶部(15533个中的28个)。具有一个错配或一个插入的序列是最富集的文库成员。在得分最高的I-PpoI脱靶候选中的错配位置的分析表明，识别基序中的某些位置对I-PpoI切割比其它位置更重要(图29)。特别地，位置2、13和14似乎是高度保守的，并且对于I-PpoI介导的DNA切割最重要。体外选择对归巢核酸内切酶的适应性表明，这些选择可广泛用于分析多种核酸酶的脱靶谱，包括被切割以显示粘性末端的那些(如I-PpoI和Cas12a)。I-PpoI留下4bp的3'突出端，其为一种DNA末端构造，已知降低通过例如GUIDE-seq或CIRCLE-seq等现有方法来进行的脱靶检测的效率。因此，我们证实体外选择可用于分析诱导交错DNA断裂的核酸酶。

[0093] 表3.I-PpoI的选择后文库对无偏见DNA文库的前30个最富集的位点。

[0094]

比对	靶	#	发现序列_切割的	发现序列_切割的_rmv
1_0_1	CTATCTTAAGGTAGTC	97.	1507	1459
1_0_1	ACTCTCTTAAGGTAGC	98.	1329	1294
1_0_1	CTATCTTAAGGTAGCC	99.	1264	1235
3_0_0	CTACCTTAAGGTAGT	100.	1100	1071
3_0_0	CTACCTTAAGGGAGC	101.	1017	989
2_0_0	CTATCTTAAGGGAGC	102.	967	951
2_0_0	CTCCCTTAAGGGAGC	103.	960	923
1_0_1	CTATCTTAAGGTAGGC	104.	947	919
1_0_1	CTCTCTTAAGGGAGCC	105.	920	896
1_0_1	CTCTCTTAAGGTAGCT	106.	913	883
2_0_0	CTCCCTTAAGGTAGT	107.	885	866
0_0_1	CTCTCTTAAGGTAGTC	108.	865	842
1_0_1	CTCTCTTAAGATAGCC	109.	858	836
2_0_0	CTACCTTAAGGTAGC	110.	829	799
1_0_1	CTCCCTTAAGGTAGTC	111.	781	765
1_0_1	CTCTCATAAGGTAGTC	112.	744	724
1_0_1	CTCTCATAAGGTAGCC	113.	744	722
1_0_1	CTCTGTTAAGGTAGTC	114.	729	710
3_0_0	CTCCCTTAAGAGAGC	115.	732	702
1_0_1	CTCCCTTAAGGTAGCC	116.	713	694
1_0_1	CTCCCTTAAGGTAGAC	117.	709	689
1_0_0	CTCTCTTAAGGTAGT	118.	679	670
2_0_0	CTATCTTAAGGTAGT	119.	687	670
1_0_0	CTCTCTTAAGGGAGC	120.	673	653
3_0_0	CTATCTTAAGGGAGT	121.	651	639
1_0_1	CTCTGTTAAGGTAGCC	122.	650	636
0_0_1	CTCTCTTAAGGTAGGC	123.	650	633
0_0_0	CTCTCTTAAGGTAGC	124.	643	628
1_0_0	CTATCTTAAGGTAGC	125.	629	620
2_0_0	CTCTCTTAAGAGAGC	126.	634	620

[0095] #,SEQ ID NO:

[0096] 最紧密配对的脱靶候选富集至选择的顶部。然而,选择表明I-PpoI脱靶候选大量存在。

[0097] 方法:库生成

[0098] 高密度芯片阵列上的寡核苷酸文库合成购自Agilent。

[0099] 底物谱文库:

[0100] 1)开发寡核苷酸主链,该主链具有50%的GC含量并且没有潜在的典型PAM序列(针对化脓链球菌(S.pyogenes)Cas9的NGG)。

[0101] 2)生成13-14个碱基对条码,这些条码与所有其它条码至少相差两个置换,具有40-60%GC,并且不含有任何针对最小无偏见文库的典型PAM序列:

[0102] 3)对SpCas9靶位点的置换、插入和缺失的所有可能组合生成潜在脱靶位点(这可为可变的):

[0103]

置换	单碱基对缺失	单碱基对插入
<=3	0	0
<=1	1	0
0	2	0
<=1	0	1

[0104] 4)将对所有i脱靶位点(I约50,000)的条码/潜在脱靶位点结合至主链中:

[0105] GACGTTCTCACAGCAATTCGTACAGTCGACGTCGATTCGTGCT(条码$_i$)TT TGACATTCTGCAATTGCACACAGCGT(潜在_脱靶_位点$_i$)TGCAGACTG TAAGTATGTATGCTTCGCGCAGTGCGACTTCGCAGCGCATCACTTCA(条码$_i$)AGTAGCTGCGAGTCTTACAGCATTGC (SEQ ID NO:127)

[0106] 基因组激发文库:

[0107] 1)根据下表(这些参数可变化)由CasOffFinder生成潜在脱靶位点并添加20-113bp(这可为可变的)的基因组侧翼序列

[0108]

置换	单碱基对缺失	单碱基对插入
<=6	0	0
<=4	<=2	0
<=3	0	<=2
4	0	1

[0109] 对于EMX1位点,此处为表现给出的上述参数的序列数的实例。

[0110]

错配	插入(DNA 凸起)长度	缺失(RNA 凸起长度)	序列的#
0	0	0	1
2	0	0	1
3	0	0	25
4	0	0	378
5	0	0	3903
6	0	0	30213
1	0	2	1
2	1	0	6
2	2	0	7
2	0	1	17
2	0	2	161

[0111]

3	1	0	130
3	2	0	126
3	0	1	566
3	0	2	7579
4	1	0	2214
4	2	0	1942
4	0	1	8279
		总计	55549

[0112] 2)将对所有i脱靶位点(i约50,000)的条码/潜在脱靶位点结合至主链中,关于具有最大基因组侧翼环境的最小无偏见文库:

[0113] GACGTTCTCACAGCAATTCGT(条码$_i$)(侧翼基因组环境$_i$)(潜在_脱靶_位点$_i$)(侧翼_基因组_环境$_i$)(条码$_i$)TGCGAGTCTTACAGCATTGC (SEQ ID NO:128)

[0114] 随着侧翼基因组环境的变化,可以增加恒定的主链序列。

[0115] 例如,在两侧都有10bp的基因组侧翼序列:

[0116] GACGTTCTCACAGCAATTCGTACAGTCGACGTCGATTCGTGCT(条码$_i$)TTTGACATTCTGCAATGT(侧翼_基因组_环境$_i$)(潜在_脱靶_位点$_i$)(侧翼_基因组_环境$_i$)(AAGTATGTATGCTTCGCGCAGTGCGACTTCGCAGCGCATCACTTCA(条码$_i$)AGTAGCTGCGAGTCTTACAGCATTGC (SEQ ID NO:129)

[0117] 其它文库生成策略:

[0118] -将基于群体的SNP并入基因组序列

[0119] -仅基于编码DNA序列生成文库

[0120] -生成癌基因热点或肿瘤抑制基因的位点的文库

[0121] 以下是使用利用上述原理构建的脱靶文库的方法的实例。

[0122] 用于切割的文库成员的体外选择的方法

[0123] 1.文库扩增

[0124] 我们使用与在所有文库成员中发现的恒定侧翼区结合的引物扩增寡核苷酸文库。

这些引物含有引入额外长度和特有分子标识符的5’端突出端。使用采用2μl的5nM输入文库的以下方案扩增文库。

[0125]

SV (2l 的 5nM 输入文库)	2
Thermopol 缓冲液	5
Taq Polym.	0.25
dNTP 10mM	1
KP_延伸_新_fw*	1
KP_延伸_新_rev*	1
H2O	39.75
RV	50
PCR 程序	
循环	12
ID 95	30
D 95	20
A 50	15
E 68	1
FE 68	30min

[0126] SV-样品体积

[0127] RV-反应体积

[0128] *KP_延伸_新_fw,引物序列：

[0129] GCTGACTAGACACTGCTATCACACTCTCTCANNNNNNNNAGACGTTCTCACAGCAATTCG (SEQ ID NO:130)

[0130] *KP_延伸_新_rev,引物序列：

[0131] GCGTAATCACTGATGCTTCGTAAATGAGACANNNNNNNNTGCAATGCTGTAAGACTCGCA (SEQ ID NO:131)

[0132] 2.DNA纯化：

[0133] 根据制造商的方案,由AMPure磁珠以0.9X的样品:磁珠比率进行DNA纯化。

[0134] 3.酶孵育：

[0135] 在不同的酶浓度和孵育时间下,用目标蛋白质孵育300ng的芯片合成的文库。在大多数情况下(Cas9、Cas9HF、BE3、ABE),足以对蛋白质、sgRNA和DNA底物的摩尔比分别为10:1:1的300ng的寡核苷酸文库在活性缓冲液中进行1-2h的酶孵育。取决于特定的蛋白质功能,这些参数可能需要优化。

[0136] 4.任选的DNA切口：

[0137] 取决于所分析的蛋白,酶孵育不会导致DNA双链断裂(DSB)的产生。在BE3和ABE的情况中,两种酶仅在DNA的链上形成切口,而对其它进行碱基编辑。通过分别对于BE3和ABE采用USER酶或核酸内切酶V,可以将该DNA切口转换为交错的DSB(参见图8和9)。为实现此目的,用USER酶或核酸内切酶V在其各自的活性缓冲液中在37℃下孵育来自步骤4的磁珠纯化的DNA。

[0138] 5.DNA纯化：

[0139] 根据制造商的方案,由AMPure磁珠以1.5X的样品:磁珠比率进行DNA纯化。

[0140] 6.任选的DNA钝化:

[0141] 如果需要额外的切口步骤(5),则通过由Phusion聚合酶在72℃下孵育20分钟然后冷却至4℃来钝化交错的DSB。

[0142] 7.DNA纯化:

[0143] 根据制造商的方案,由AMPure磁珠以1.5X的样品:磁珠比率进行DNA纯化。

[0144] 8.衔接子连接:

[0145] 然后,将半功能性Y形衔接子连接至来自步骤7的钝化的DNA。为实现此目的,我们提供超过文库片段10倍摩尔过量的衔接子,并使用NEB快速连接试剂盒进行连接,在25℃下孵育反应10分钟。

[0146] 9.凝胶纯化:

[0147] 然后,我们通过采用2.5%琼脂糖凝胶对连接反应进行凝胶纯化。电泳在120伏下进行1小时。1小时后,在180bp片段大小周围切下含有泳道的样品,并根据制造商的方案使用Qiagen凝胶提取试剂盒提取DNA。

[0148] 10.PCR-扩增:

[0149] 随后将来自步骤9的洗脱液用作两个PCR反应的输入,这两个PCR反应扩增切下的文库成员的前间隔序列相邻位点和PAM相邻位点。用于该PCR的引物含有其后可用于附加Illumina测序条码的5'突出端。任选地,可以进行QPCR以确定所需的最小PCR循环数。使用以下参数进行PCR:

[0150]

样品体积	6
Phusion 高保真缓冲液 5X	10
Phusion 聚合酶	0.5
dNTP 10mM	1
引物 A	2.5
引物 B	2.5
H2O	27.5
PCR 程序	
循环	25-35
ID 98	30
D 98	10
A 65	20
E 72	5
FE 72	5 min

[0151] 11.DNA纯化:

[0152] 根据制造商的方案,由AMPure磁珠以1.5X的样品:磁珠比率进行DNA纯化。

[0153] 12.使用毛细管电泳进行质量控制:

[0154] 通过毛细管电泳检查PCR产物来进行质量控制。

[0155] 13.基于PCR的NGS文库制备:

[0156] 通过由含有Illumina测序衔接子的引物进行PCR来将测序衔接子附加至来自步骤12的PCR产物中。使用以下参数进行PCR:

[0157]

样品体积	总计 50ng
Phusion 高保真缓冲液 5X	10
Phusion 聚合酶	0.5
dNTP 10mM	1
IndexPrimerA	2.5
IndexPrimerB	2.5
H2O	Ad 50
PCR 程序	
循环	10
ID 98	30
D 98	10
A 65	30
E 72	35
FE 72	10 min

[0158] 14.DNA纯化：

[0159] 根据制造商的方案，由AMPure磁珠以1.5X的样品:磁珠比率进行DNA纯化。

[0160] 15.Illumina测序仪上的下一代测序：

[0161] 通过数字微滴式PCR定量来自步骤14的DNA文库，并根据制造商的方案在Illumina测序仪上测序。

[0162] 用于通过拉下的DNA结合位点的富集的方法

[0163] 1)重悬快照捕获磁珠(Snap Capture Bead,NEB)

[0164] 2)将80uL的磁珠移液至新的1.5mL的Eppendorf管中

[0165] 3)将管置于磁性颗粒分离器中并丢弃上清液

[0166] 4)添加1mL的固定缓冲液(20mM HEPES,150mM NaCl,0.5％Tween20、1mM DTT,pH 6.5)并缓慢涡旋

[0167] 5)将管置于磁性颗粒分离器中并丢弃上清液

[0168] 6)制备蛋白质:将Engen Spy dCas9(SNAP-tag)(NEB)(4.5uL的20uM每个拉下反应)添加至500uL的固定缓冲液中

[0169] 7)将稀释的蛋白添加至磁珠中并通过移液混合均匀

[0170] 8)在室温下振荡孵育1小时

[0171] 9)将管置于磁性颗粒分离器中并丢弃上清液

[0172] 10)洗涤磁珠。添加1mL的固定缓冲液，移液混合均匀，然后将管置于磁性颗粒分离器中并丢弃上清液

[0173] 11)重复步骤10两次，更多为总计3次洗涤。用含有10ug/mL的肝素的固定缓冲液进行最后一次清洗

[0174] 12)每次拉下时，将磁珠重悬于45uL的固定缓冲液中

[0175] 13)混合以下成分：

[0176]

组分	用于1个拉下反应的量
水	在添加包括0.9pmol的文库的所有后，添加足以使最终体积为60uL的量
10X固定缓冲液+100ug/mL肝素	6uL
gRNA	3500ng
Engen Spy dCas9 (SNAP-tag) +磁珠	45uL

[0177] 14)在25摄氏度下孵育10分钟

[0178] 15)添加0.9pmol的文库

[0179] 16)在37摄氏度下孵育30分钟

[0180] 17)将管置于磁珠分离器上并丢弃上清液

[0181] 18)用含有10ug/mL的肝素的200uL的固定缓冲液洗涤磁珠5次

[0182] 19)添加50uL的水和2uL的蛋白酶K并在摇动的同时在室温下孵育10分钟

[0183] 20)用DNA纯化磁珠(例如Ampure)清理拉下产物并用10uL的0.1X缓冲液EB(QIAgen)洗脱

[0184] 其它实施方式

[0185] 应当理解,尽管已经结合本发明的详细描述描述了本发明,但前述描述旨在说明而不是限制本发明的范围,本发明的范围由所附权利要求的范围限定。其它方面、优点和修改在所附权利要求的范围内。

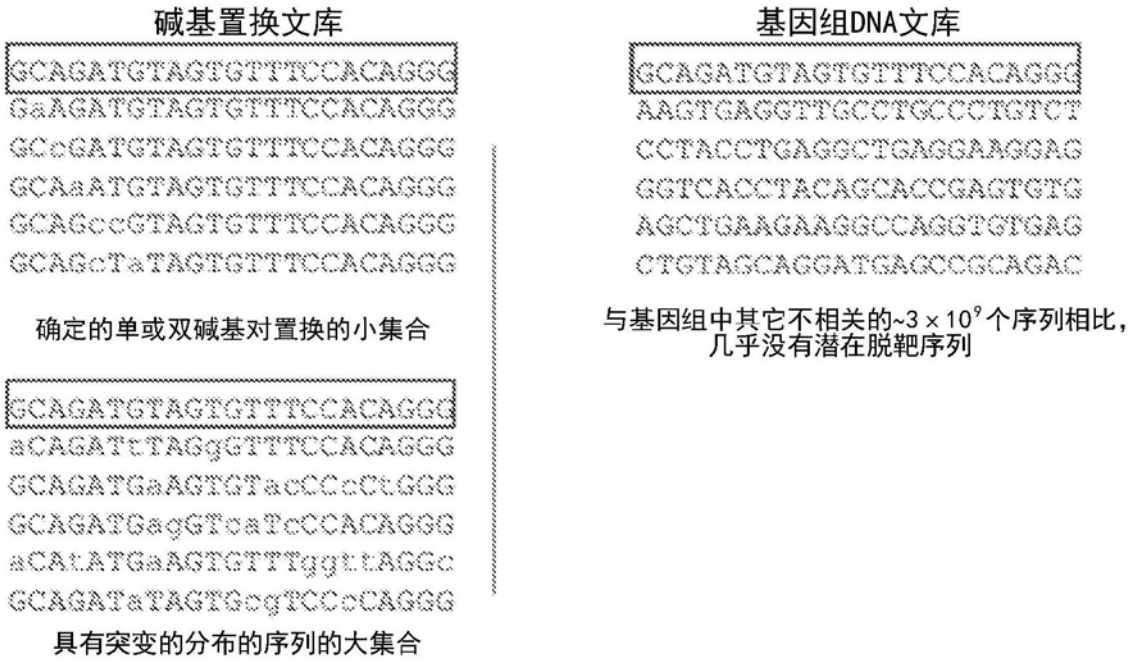
碱基置换文库
GCAGATGTAGTGTTTCCACAGGG
GaAGATGTAGTGTTTCCACAGGG
GCcGATGTAGTGTTTCCACAGGG
GCAaATGTAGTGTTTCCACAGGG
GCAGccGTAGTGTTTCCACAGGG
GCAGcTaTAGTGTTTCCACAGGG
确定的单或双碱基对置换的小集合
GCAGATGTAGTGTTTCCACAGGG
aCAGATtTAGgGTTTCCACAGGG
GCAGATGaAGTGTaCCcCtGGG
GCAGATGagGTcaTcCCACAGGG
aCAtATGaAGTGTTTggtttAGGc
GCAGATaTAGTGcgTCCcCAGGG
具有突变的分布的序列的大集合
基因组DNA文库
GCAGATGTAGTGTTTCCACAGGG
AAGTGAGGTTGCCTGCCCTGTCT
CCTACCTGAGGCTGAGGAAGGAG
GGTCACCTACAGCACCGAGTGTG
AGCTGAAGAAGGCCAGGTGTGAG
CTGTAGCAGGATGAGCCGCAGAC
与基因组中其它不相关的~3×10^9个序列相比，
几乎没有潜在脱靶序列

图1

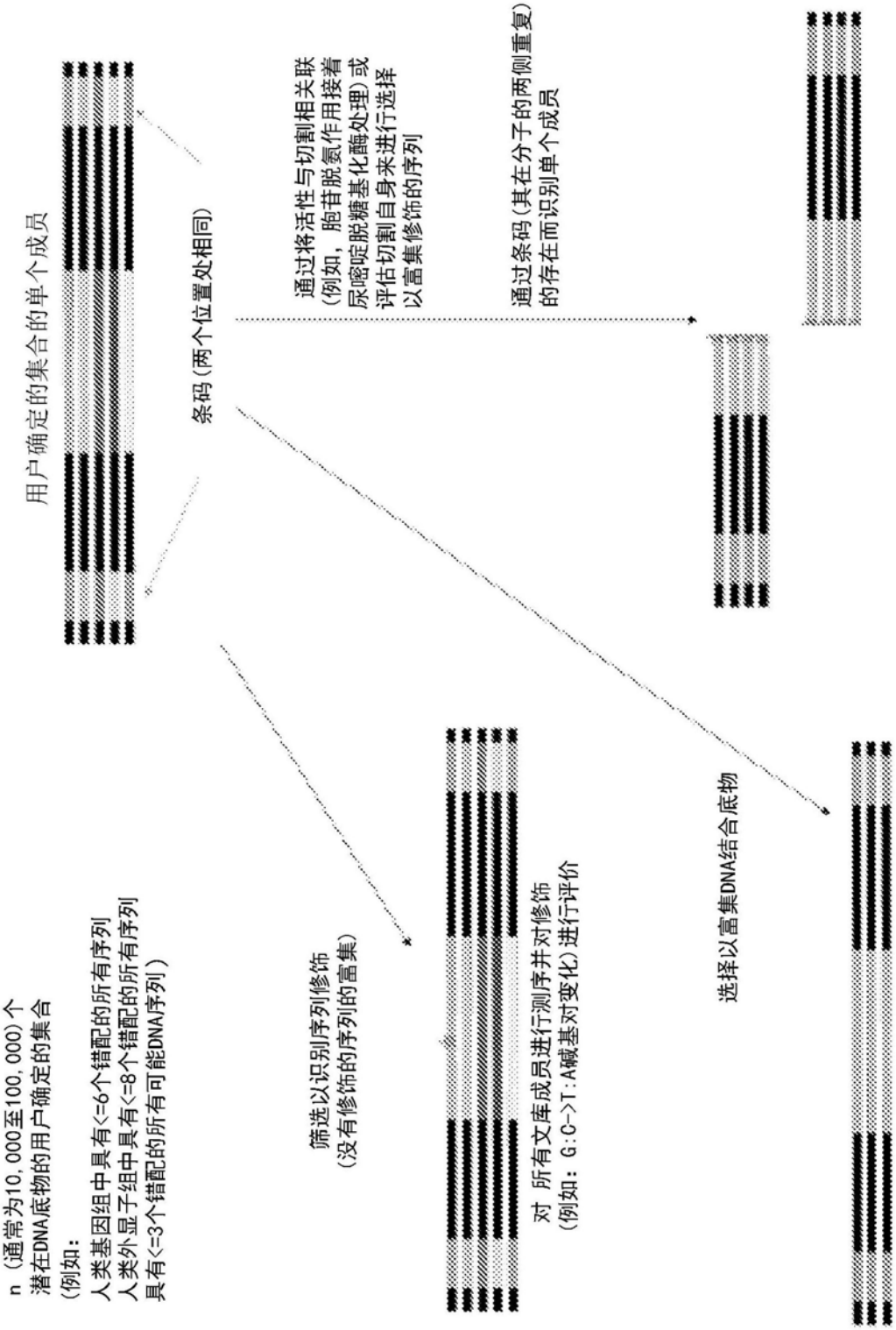
n（通常为10,000至100,000）个
潜在DNA底物的用户确定的集合
（例如：
人类基因组中具有<=6个错配的所有序列
人类外显子组中具有<=8个错配的所有序列
具有<=3个错配的所有可能DNA序列）
用户确定的集合的单个成员
条码（两个位置处相同）
通过将活性与切割相关联
（例如，胞苷脱氨作用接着
尿嘧啶脱糖基化酶处理）或
评估切割自身来进行选择
以富集修饰的序列
通过条码（其在分子的两侧重复）
的存在而识别单个成员
筛选以识别序列修饰
（没有修饰的序列的富集）
对 所有文库成员进行测序并对修饰
（例如：G:C->T:A碱基对变化）进行评价
选择以富集DNA结合底物

图2

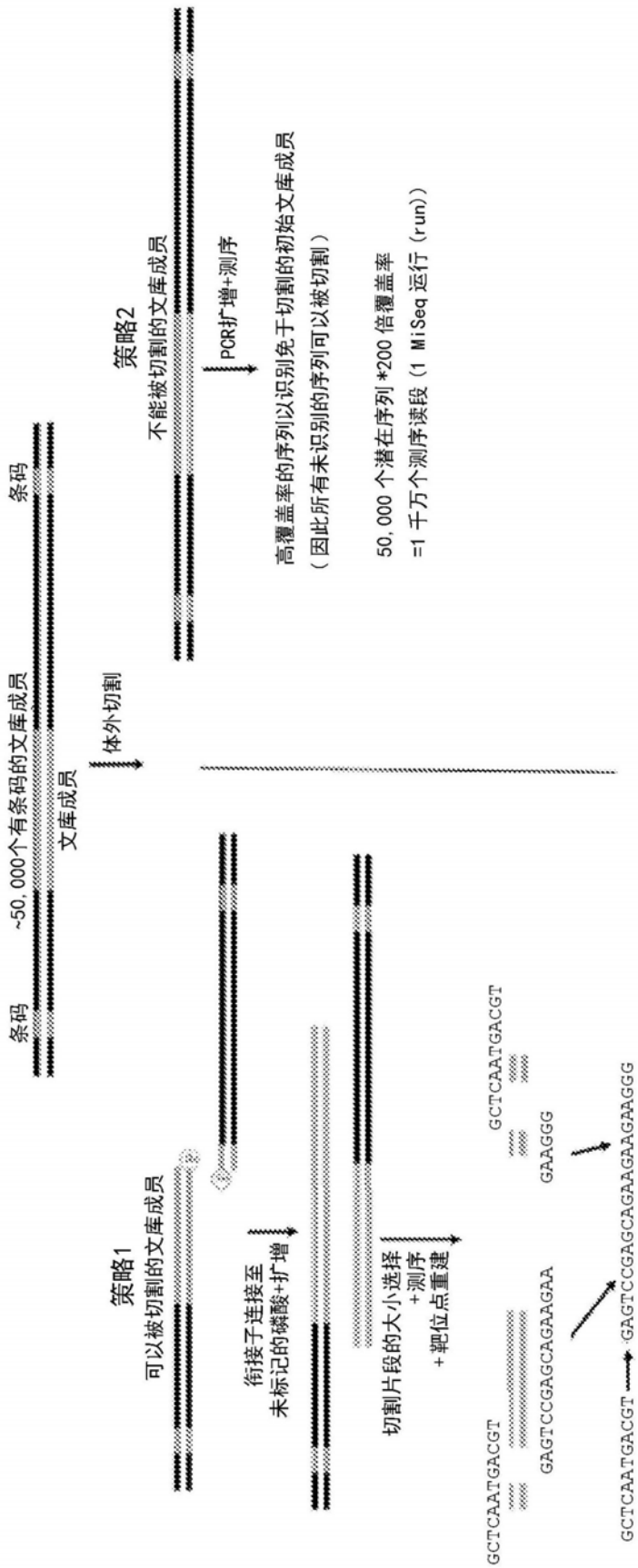
条码
~50,000个有条码的文库成员
条码
文库成员
体外切割
策略2
不能被切割的文库成员
PCR扩增+测序
高覆盖率的序列以识别免于切割的初始文库成员
（因此所有未识别的序列可以被切割）
50,000 个潜在序列 *200 倍覆盖率
=1 千万个测序读段（1 MiSeq 运行（run））
策略1
可以被切割的文库成员
衔接子连接至
未标记的磷酸+扩增
切割片段的大小选择
+测序
+靶位点重建
GCTCAATGACGT
GAGTCCGAGCAGAAGAA
GCTCAATGACGT
GAAGGG
GCTCAATGACGT
GAGTCCGAGCAGAAGAAGAAGGG
初始文库成员的识别

图3

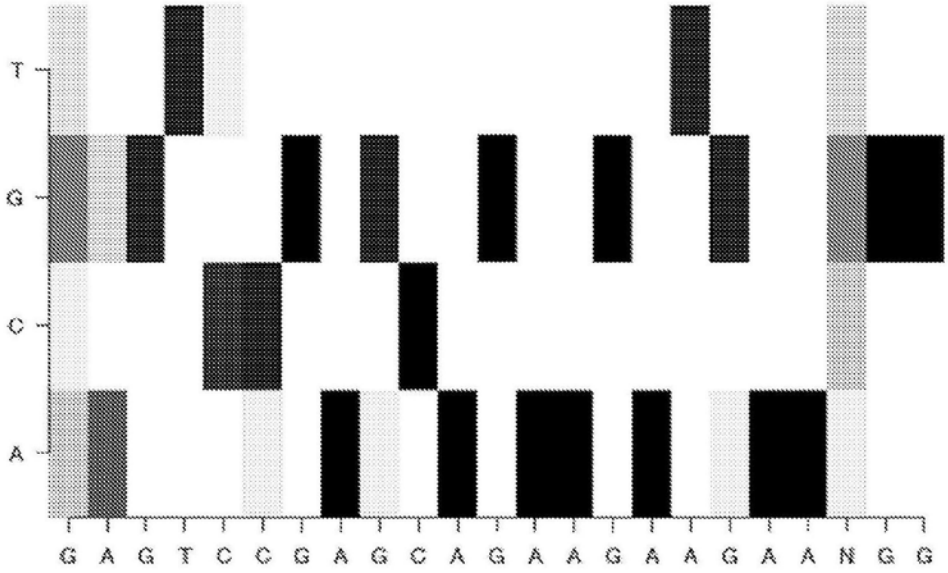

图4

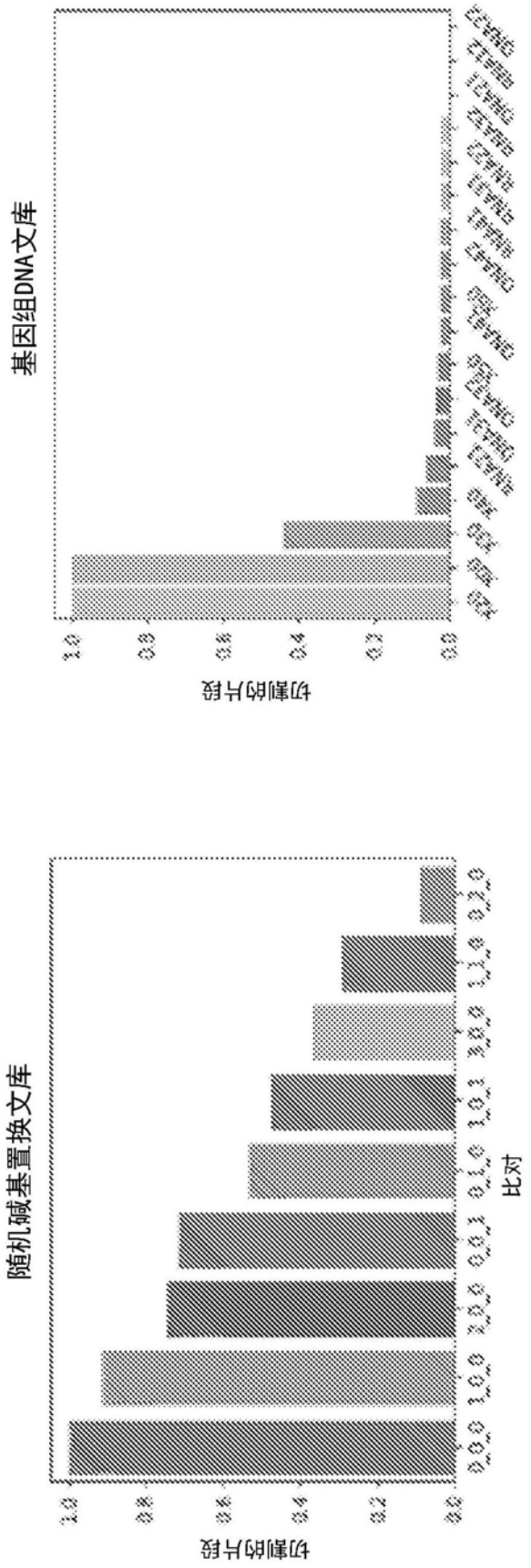

图5

凸起类型	凸起大小	错配	ABE14	ABE16	ABE18	EMX1	FANCF	HBB	HEK2	HEK3	HEK4	RNF2	VEGFA3
无凸起	0	0	1	1	1	1	1	1	1	1	1	1	1
	0	1	1	0	1	0	0	0	0	0	0	0	1
	0	2	1	0	4	1	2	4	1	1	9	0	35
	0	3	9	20	73	25	31	51	14	14	116	10	1041
	0	4	151	259	814	398	421	610	225	141	1149	199	22465
	0	5	1713	3072	6528	4013	2778	5757	2904	1706	8446	1868	0
	0	6	14601	30685	44576	30914	19151	44090	24602	14119	47599	16426	0
DNA	1	1	0	0	1	0	0	0	0	0	0	0	1
	2	1	0	0	0	0	0	0	0	1	1	0	0
	1	2	1	3	20	6	2	1	5	0	13	2	61
	2	2	1	2	0	7	4	15	6	6	7	6	0
	1	3	49	74	327	133	75	106	128	65	356	67	0
	2	3	53	92	48	127	92	258	138	60	242	74	0
	1	4	745	1284	4013	2283	1062	2006	2176	1001	5625	1167	0
	2	4	853	1675	1193	1988	1645	3752	2512	1114	3365	1151	0
RNA	1	1	0	0	0	0	1	0	0	0	3	0	12
	2	1	0	0	0	1	4	0	3	3	7	0	0
	1	2	8	14	55	17	12	11	22	30	74	5	338
	2	2	59	55	19	164	63	121	87	42	243	59	0
	1	3	244	348	920	576	223	380	485	297	1074	316	0
	2	3	790	1777	931	7804	1431	2883	1496	949	3777	1382	0
	1	4	4242	5024	10263	8520	3779	7248	8149	4227	13197	4589	0
		总计	23522	44385	69787	56978	30777	67294	42954	23777	85304	27322	23955

图6

测序计数

sgRNA	第90个百分位数	第10个百分位数	90/10比率	缺失	总计	缺失百分比
EMX1	23	10	2.3	41	51743	0.08%
FANCF	42	21	2	15	29207	0.05%
HBB	56	28	2	1	63627	0.002%
HEK2	38	18	2.1	29	39533	0.07%
HEK3	59	29	2.0	4	22733	0.02%
HEK4	57	23	2.5	157	78205	0.20%
RNF2	69	34	2.0	15	26102	0.06%
ABE14	45	20	2.3	0	22483	0.00%
ABE16	47	21	2.2	0	39232	0.00%
ABE18	30	12	2.5	3	66580	0.00%
VEGFA3	72	30	2.4	6	22364	0.03%
平均	48.9	22.4	2.2	24.6	41982.6	0.05%

图7

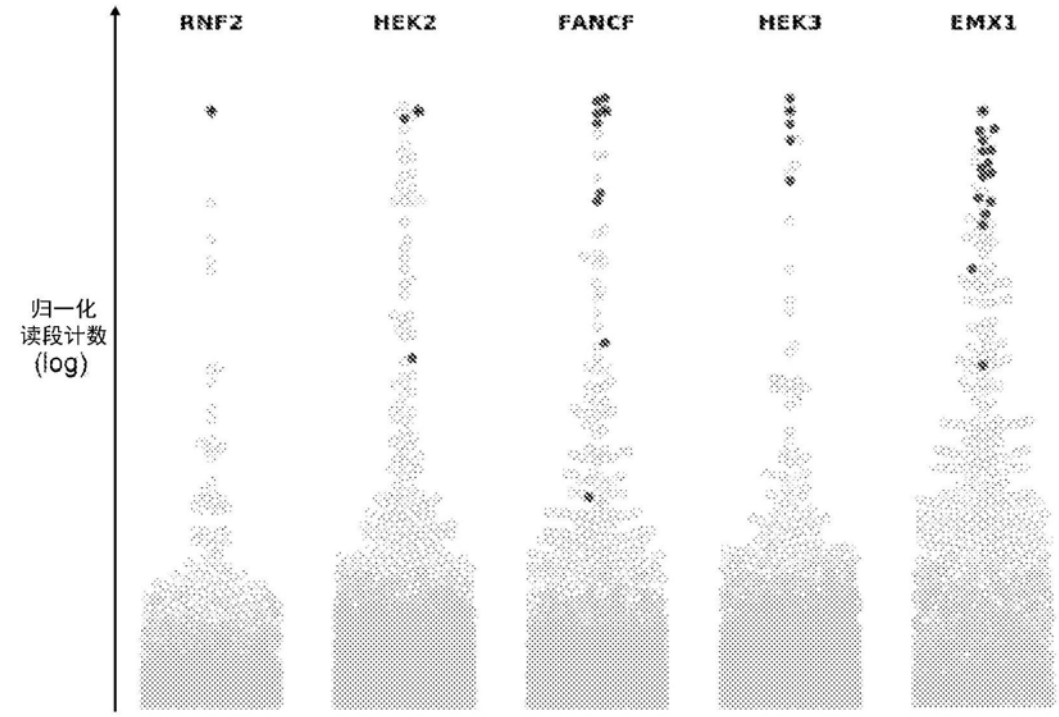
RNF2
HEK2
FANCF
HEK3
EMX1
归一化
读段计数
(log)

图8

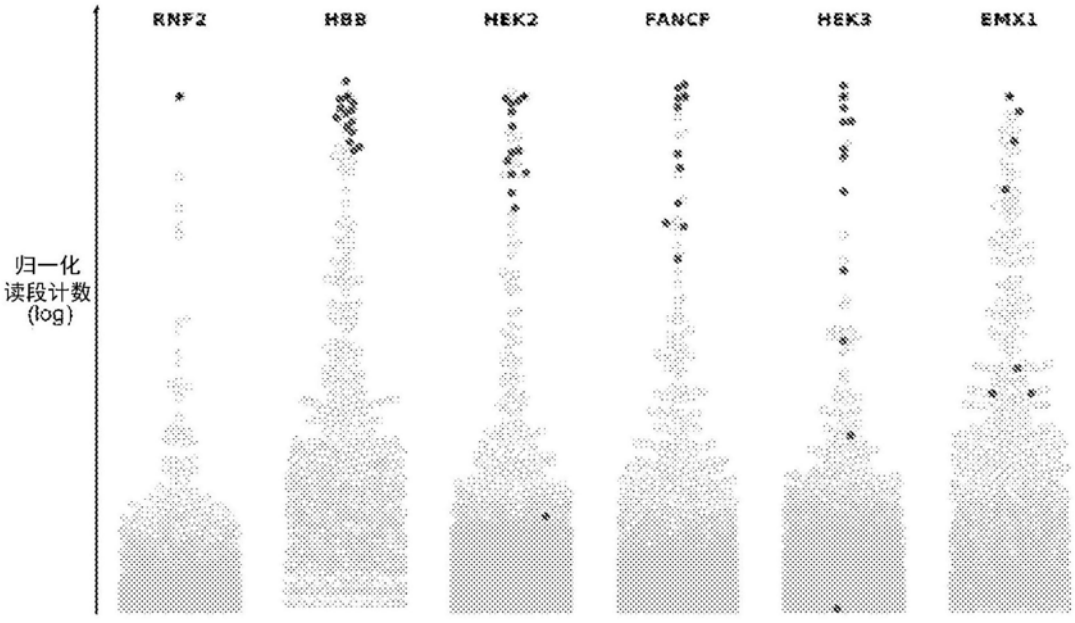
RNF2
HBB
HEK2
FANCF
HEK3
EMX1
归一化
读段计数
(log)

图9

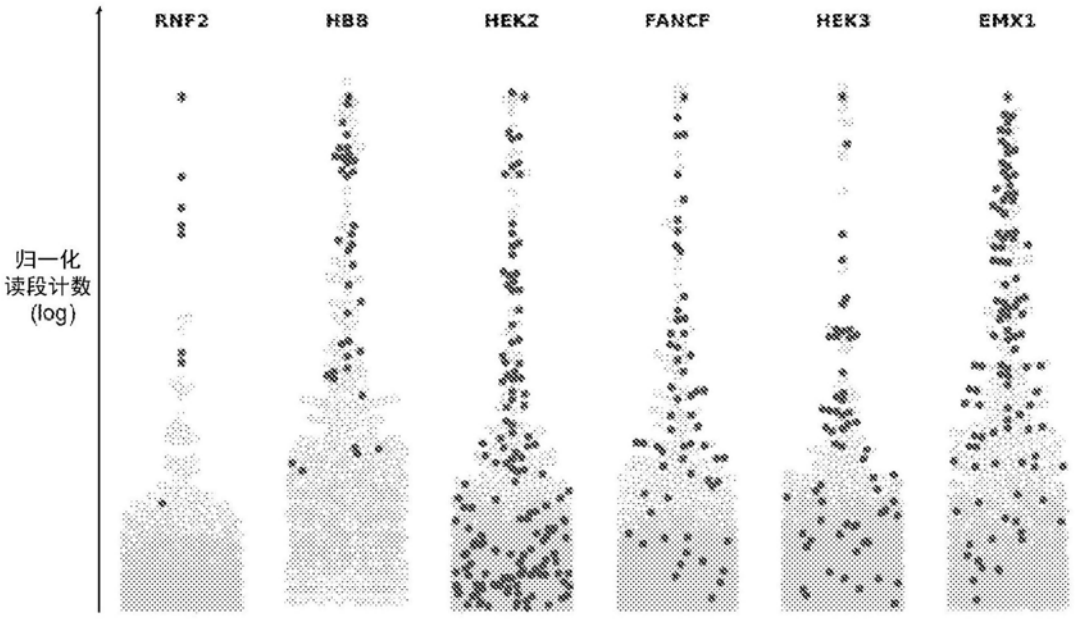
RNF2
HBB
HEK2
FANCF
HEK3
EMX1
归一化
读段计数
(log)

图10

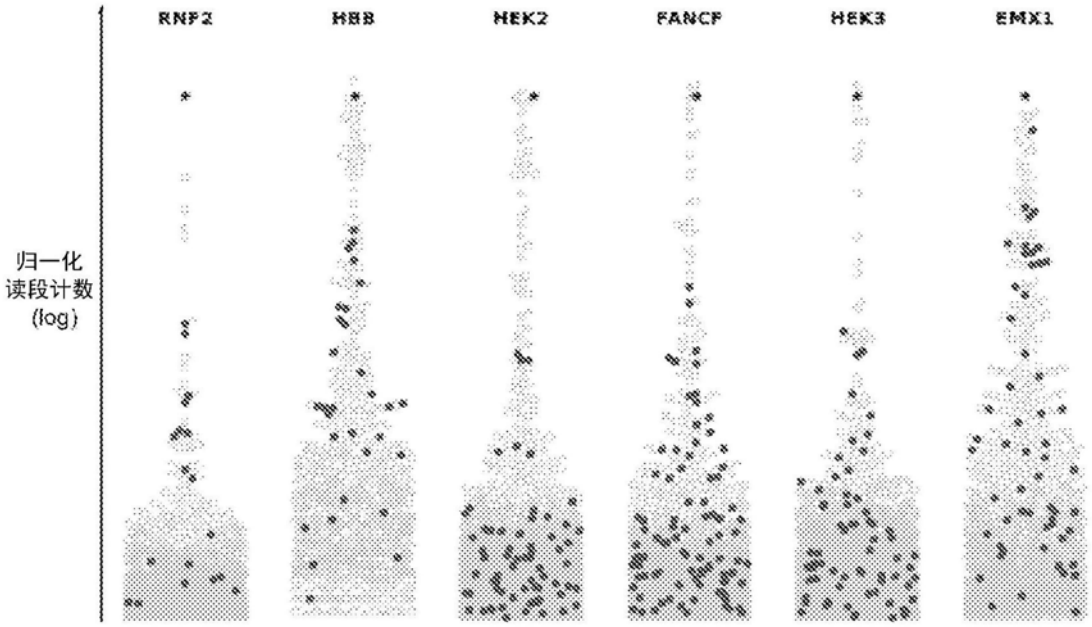
RNF2
HBB
HEK2
FANCF
HEK3
EMX1
归一化
读段计数
(log)

图11

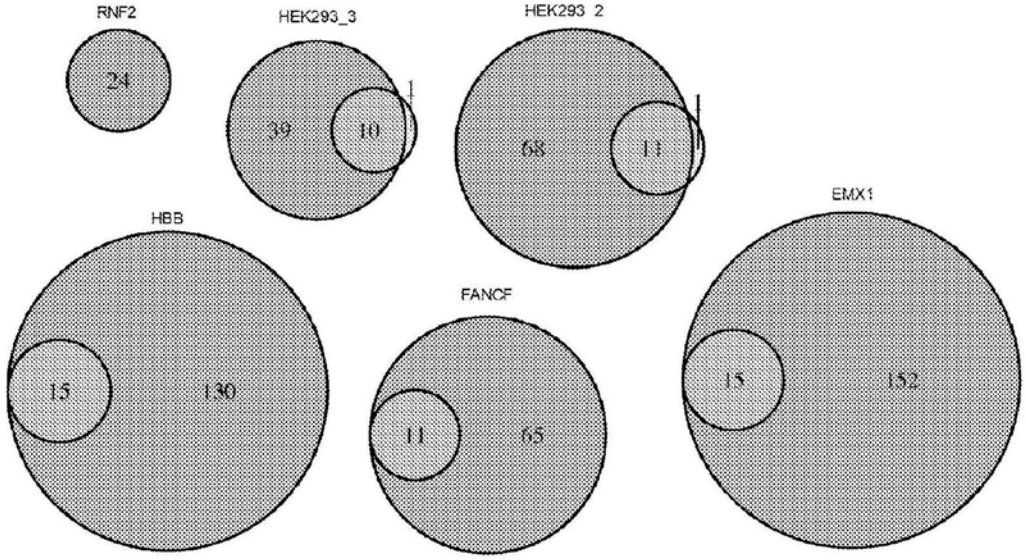
RNF2
24
HEK293_3
39
10
1
HEK293 2
68
11
1
EMX1
15
152
HBB
15
130
FANCF
11
65

图12

处理的			对照			处理的和分选的重复				对照重复		
编辑百分比	总编辑的	总参考	总编辑的	总参考	参考脱靶位点	编辑的位点	处理的#1	处理的#2	处理的#3	对照#1	对照#2	对照#3
					GGAAAgCCCT TCTGCAGCA							
0.01%	36	333903	0	330492	CCAGa	GGAAGCCCTTCT------ACCAGA	0	8	0	0	0	0
						GGAAGCCCTTCTGAAGCACACCAGA	0	7	0	0	0	0
						GGAAGCCCTTCTGAAGCCACCAGA	0	14	7	0	0	0
					GGAATCCCT TCTGCAGCA							
0.01%	81	648232	0	641626	taGtG	GGAATCCCTTCTGCAGCCAGCATAGTG	21	0	0	0	0	0
						GGAATCCCTTCTGCAGCGCATAGTG	0	0	15	0	0	0
						GGAATCCCTTCTGCA---TAGTG	45	0	0	0	0	0
					GGAAaCCCt TCTGCAGCA							
0.03%	275	785676	0	730960	CCAGc	GGAAACCCCTCTGCA---CCAGC	0	0	11	0	0	0
						GGAAACCCCTCTGCAGCAACCAGC	0	0	15	0	0	0
						GGAAACCCCTCTGCAGCAGCAATGGCAACAAC	0	0	9	0	0	0
						GTTGGCAAACTATTAACTGGCGAACTACTTAA						
						GGAAACCCCTCTGCAGCCACCAGC	0	84	56	0	0	0
						GGAAACCCCTCTGCAGC-CCAGC	72	0	0	0	0	0
						GGAAACCCCTCTGCAGCGCCACCAGC	0	0	10	0	0	0
						GGAAACCCCTCTGCAGCGGTGAAAGAAAGAAT						
						CGTCAACTTTGGCCATCTCATTGCTACCAGC	0	18	0	0	0	0

图13

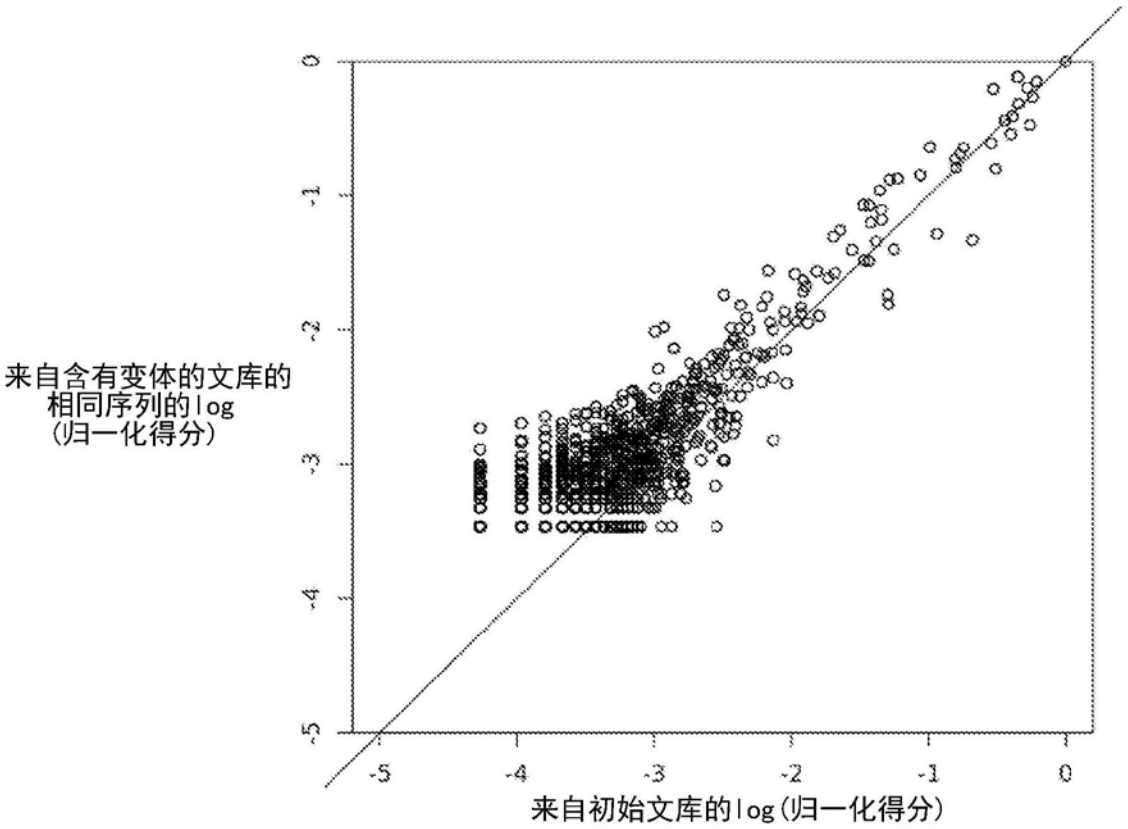
来自含有变体的文库的
相同序列的log
(归一化得分)
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
来自初始文库的log(归一化得分)

图14

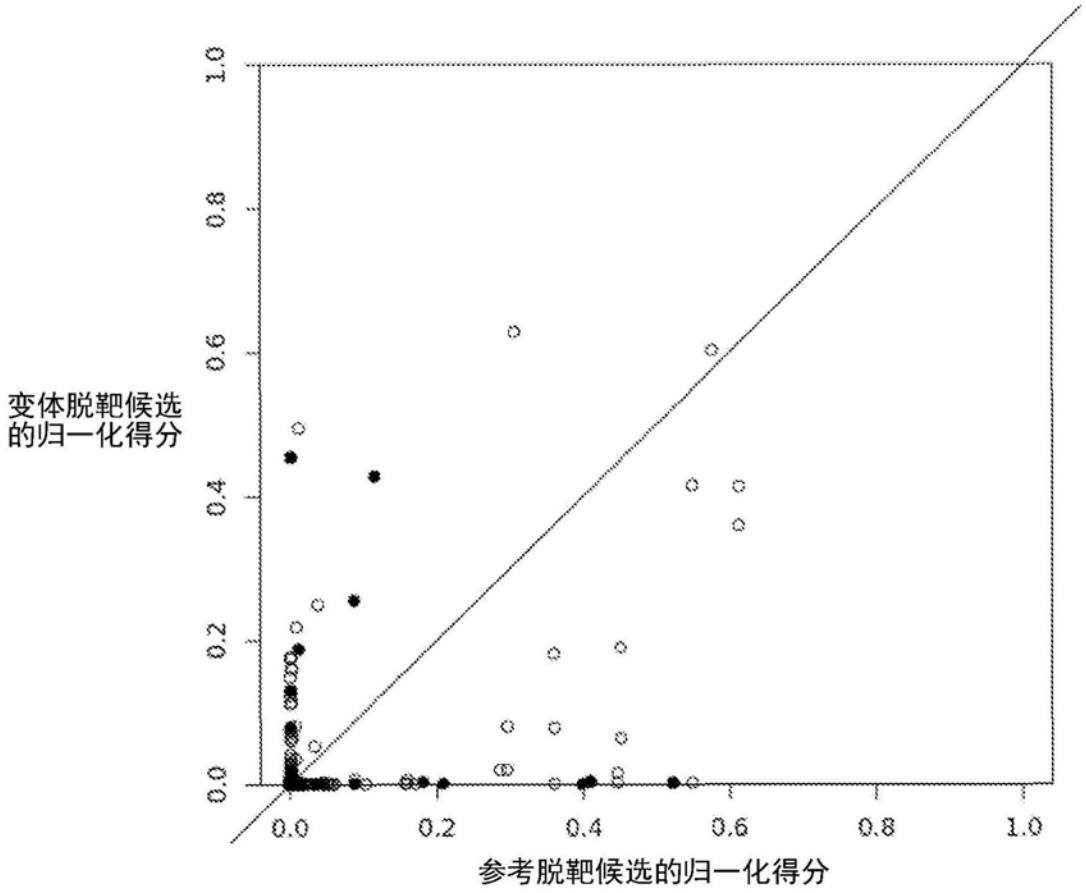
变体脱靶候选
的归一化得分
0.0
0.2
0.4
0.6
0.8
1.0
参考脱靶候选的归一化得分

图15

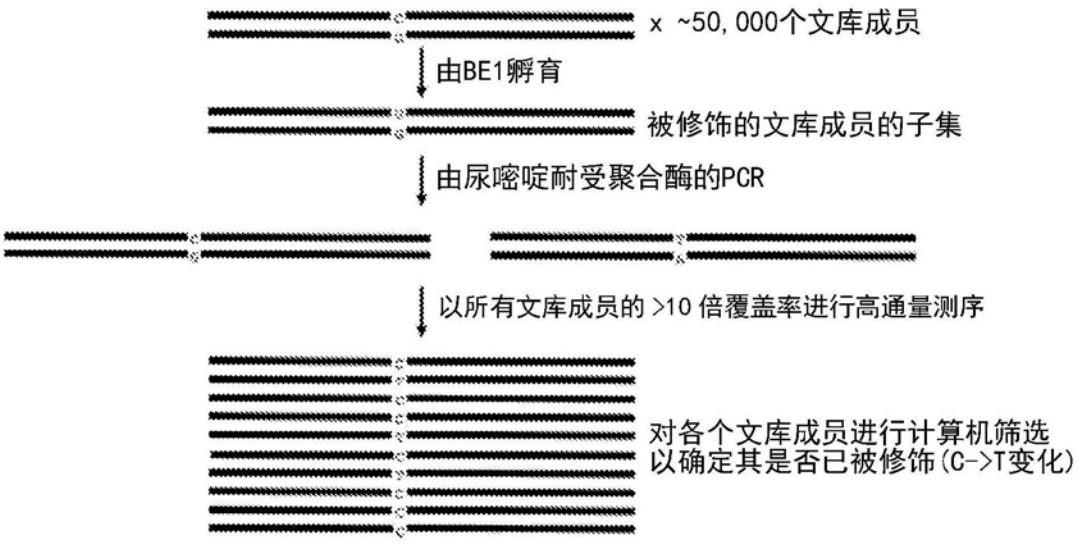
x ~50,000个文库成员
由BE1孵育
被修饰的文库成员的子集
由尿嘧啶耐受聚合酶的PCR
以所有文库成员的 >10 倍覆盖率进行高通量测序
对各个文库成员进行计算机筛选
以确定其是否已被修饰(C->T变化)

图16

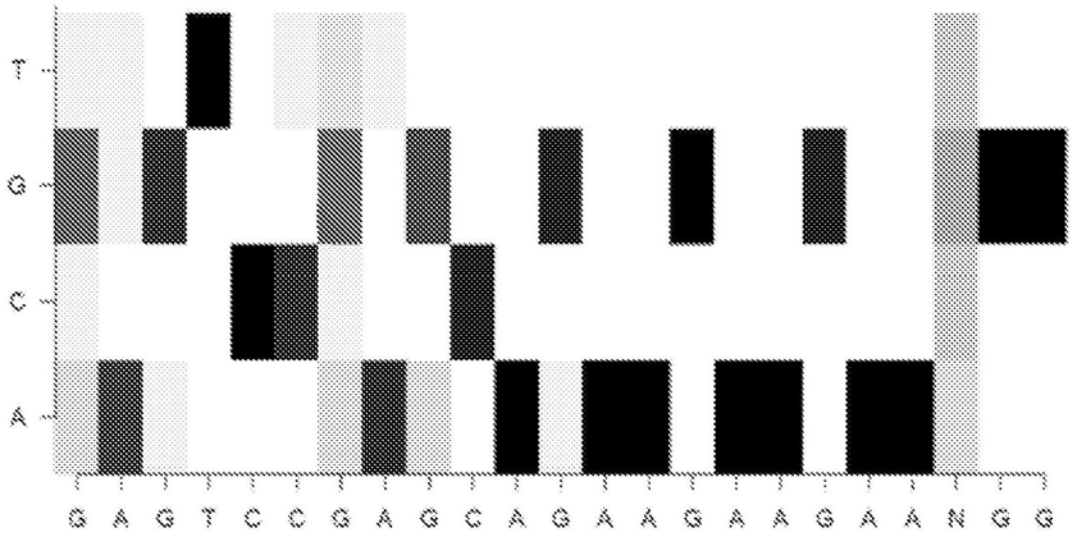
T
G
C
A
G A G T C C G A G C A G A A G A A G A A N G G

图17

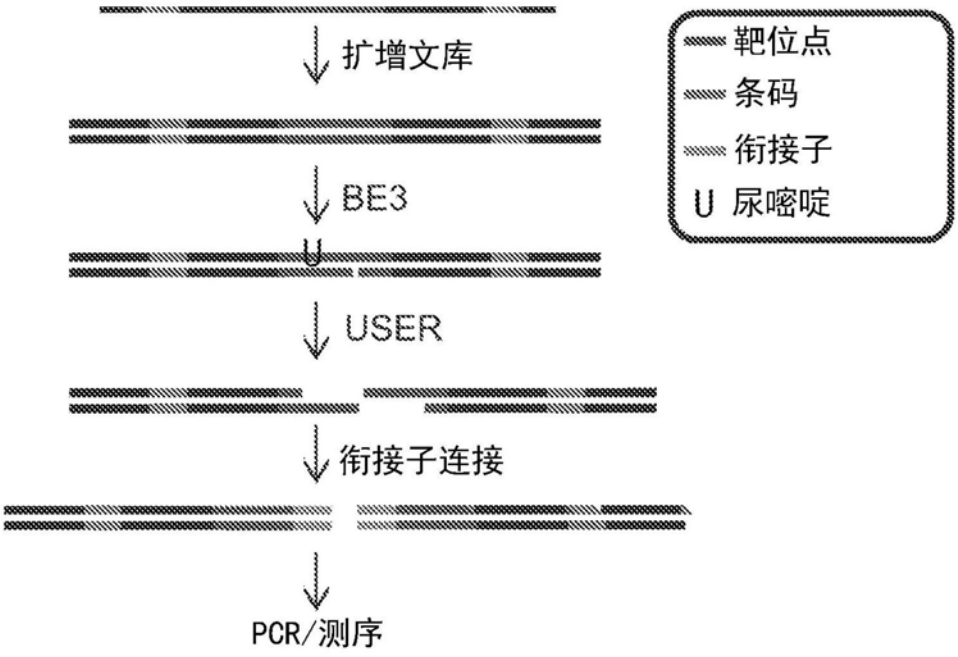
扩增文库
BE3
U
USER
衔接子连接
PCR/测序
靶位点
条码
衔接子
U 尿嘧啶

图18

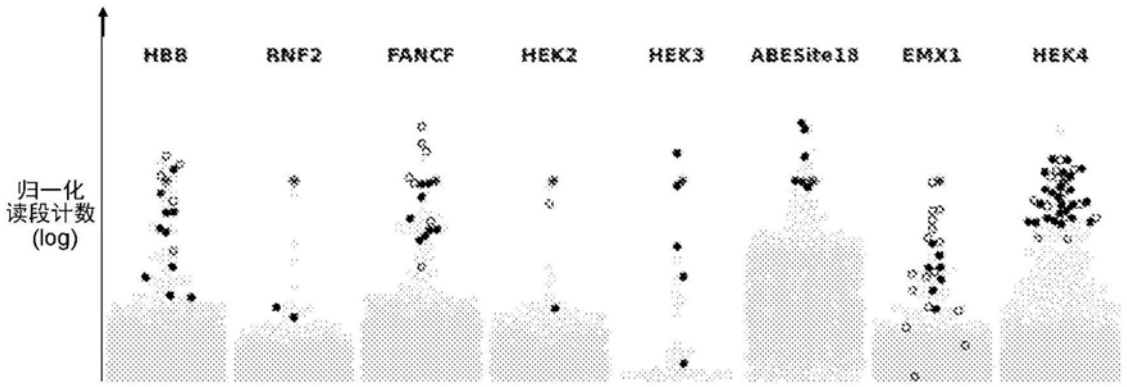

图19

sgRNA	染色体	位置		脱靶序列	处理的			对照		
					Rep1	Rep2	Rep3	Rep1	Rep2	Rep3
RNF2	chr5	92036959	OT1		0.788	0.802	0.721	0.019	0.049	0.03
RNF2	chr6	90393053	OT2		0.224	0.321	0.255	0.0	0.0	0.0
HEK4	chr13	88900085	OT1		21.596	25.676	24.391	0.075	0.077	0.053
HEK4	chr16	49798946	OT2		9.635	10.056	9.803	0.051	0.031	0.008
HEK4	chr3	51725445	OT3		6.937	8.122	8.848	0.009	0.011	0.062
HEK4	chr5	1832931	OT4		0.858	1.174	1.438	0.012	0.026	0.0
HEK4	chr9	5556615	OT5		0.351	0.385	0.373	0.047	0.014	0.03
HEK3	chr5	120833004	OT1		0.253	0.359	0.284	0.0	0.0	0.0
HEK3	chr16	88931545	OT2		0.035	0.051	0.029	0.0	0.0	0.0
HEK2	chr15	93557672	OT1		0.139	0.107	0.142	0.0	0.0	0.0
HBB	chr17	66624238	OT1		8.637	10.328	8.147	0.053	0.06	0.091
HBB	chr2	121715223	OT2		0.569	0.755	0.611	0.0	0.0	0.0
HBB	chr19	923893	OT3		0.281	0.441	0.742	0.0	0.0	0.0
HBB	chr15	46598112	OT4		0.201	0.261	0.141	0.0	0.0	0.0
HBB	chr22	17230606	OT5		0.137	0.153	0.091	0.0	0.0	0.0
HBB	chr14	36889531	OT6		0.126	0.117	0.124	0.0	0.0	0.0
HBB	chr12	27234748	OT7		0.119	0.17	0.129	0.0	0.0	0.0
HBB	chr1	17346684	OT8		0.049	0.03	0.049	0.0	0.0	0.0
HBB	chrX	10360008	OT9		0.034	0.027	0.078	0.0	0.0	0.0
FANCF	chr22	45871974	OT1		0.273	0.275	0.353	0.0	0.0	0.0
FANCF	chr6	41457558	OT2		0.046	0.066	0.05	0.0	0.0	0.0
FANCF	chr16	28615183	OT3		0.124	0.134	0.087	0.0	0.0	0.0
FANCF	chr17	39675782	OT4		0.082	0.057	0.178	0.0	0.004	0.0
EMX1	chr3	95690179	OT1		2.515	2.733	2.243	0.011	0.032	0.022
EMX1	chr1	151027591	OT2		0.369	0.521	0.442	0.0	0.0	0.0
EMX1	chr15	100405507	OT3		0.292	0.255	0.313	0.01	0.015	0.022
EMX1	chr5	64109055	OT4		0.204	0.266	0.219	0.0	0.005	0.0
EMX1	chr14	75723901	OT5		0.299	0.204	0.214	0.0	0.0	0.0

图20

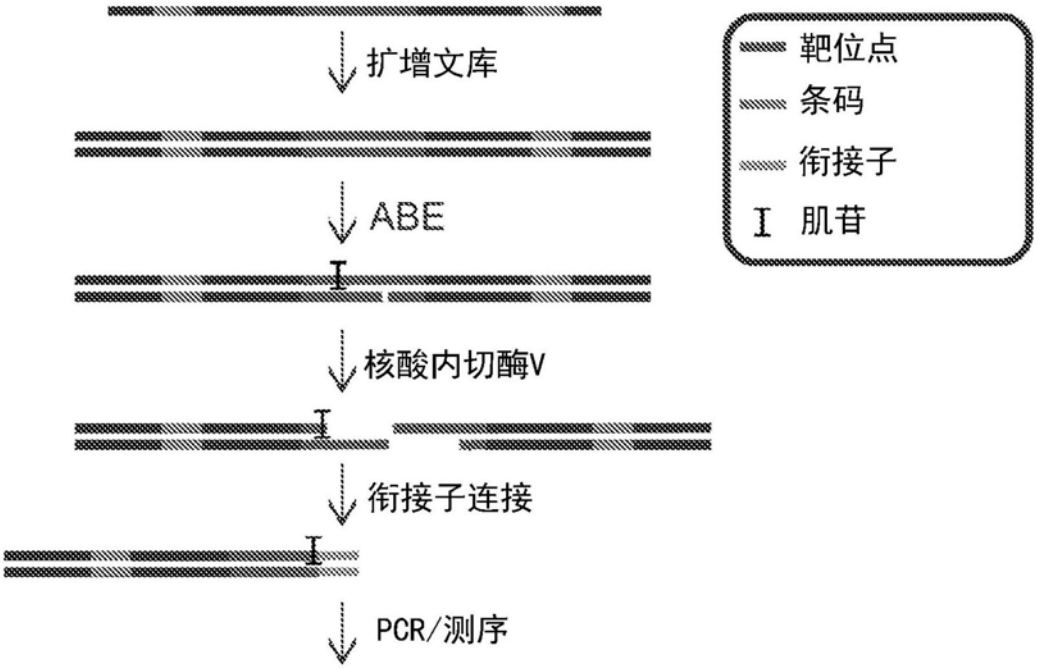
扩增文库
ABE
核酸内切酶V
衔接子连接
PCR/测序
靶位点
条码
衔接子
肌苷

图21

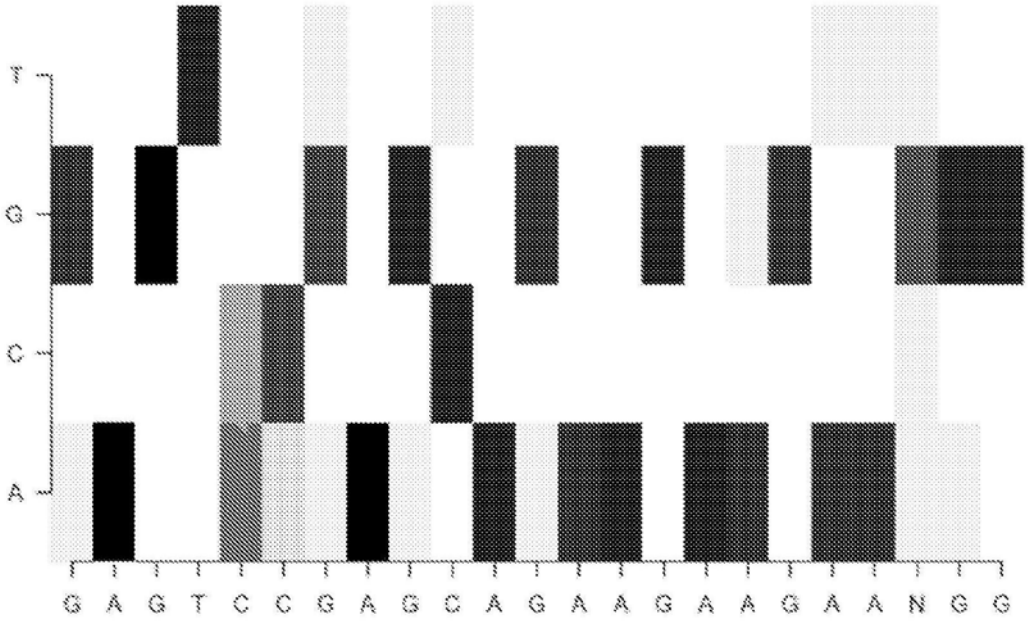
T
G
C
A
G A G T C C G A G C A G A A G A A G A A N G G

图22

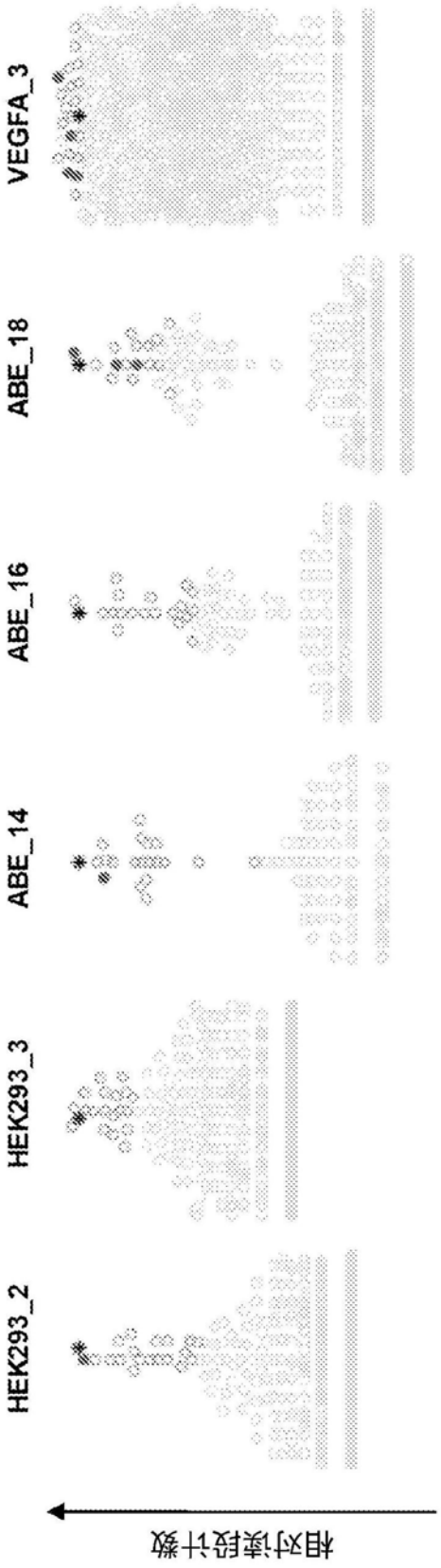

图23

sgRNA	染色体	位置		脱靶序列	处理的 Rep1	处理的 Rep2	处理的 Rep3	对照 Rep1	对照 Rep2	对照 Rep3
VEGFA3	chr2	177463419	OT1		31.0	55.1	62.1	0.007	0.009	0.0
VEGFA3	chr6	143890809	OT2		4.5	14.8	17.5	0.0	0.007	0.0
VEGFA3	chr18	74103157	OT3		3.9	11.5	13.9	0.014	0.019	0.037
VEGFA3	chr6	157078320	OT4		2.4	6.7	8.1	0.01	0.004	0.016
VEGFA3	chr15	79723835	OT5		0.3	0.8	1.2	0.0	0.038	0.035
VEGFA3	chrX	56327299	OT6		0.2	0.9	0.8	0.0	0.01	0.007
VEGFA3	chr2	74655952	OT7		0.1	0.6	0.8	0.004	0.0	0.007
HEK3	chr9	110184619	中靶		61.448	61.844	60.878	0.018	0.012	0.023
HEK2	chr5	87240596	中靶		93.236	92.812	94.386	0.0	0.002	0.009
HEK2	chr4	90522166	OT1		2.173	2.743	3.03	0.0	0.0	0.0
HEK2	chr1	167742852	OT2		0.282	0.335	0.141	0.0	0.006	0.009
HEK2	chr19	35505469	OT3		0.208	0.187	0.438	0.006	0.003	0.004
ABE18	chr1	184944031	中靶		90.763	91.615	90.981	0.006	0.024	0.016
ABE18	chr3	193547407	OT1		36.171	38.892	36.704	0.026	0.034	0.045
ABE18	chr6	34069282	OT2		20.327	18.275	19.021	0.03	0.015	0.021
ABE18	chr13	27250848	OT3		4.969	5.429	4.354	0.035	0.03	0.009
ABE16	chr1	179795817	中靶		92.874	94.161	93.633	0.009	0.004	0.016
ABE16	chr12	125515810	OT1		0.641	0.777	0.761	0.0	0.0	0.0
ABE14	chr1	154283563	中靶		86.31	86.226	86.1	0.018	0.02	0.021
ABE14	chr1	154292470	OT1		85.764	85.801	84.709	0.008	0.025	0.012
ABE14	chr7	131015337	OT2		0.996	0.953	0.716	0.006	0.006	0.012
ABE14	chr13	103838107	OT3		0.146	0.167	0.134	0.02	0.02	0.014
ABE14	chr1	127611759	OT4		0.076	0.165	0.031	0.0	0.0	0.0

图24

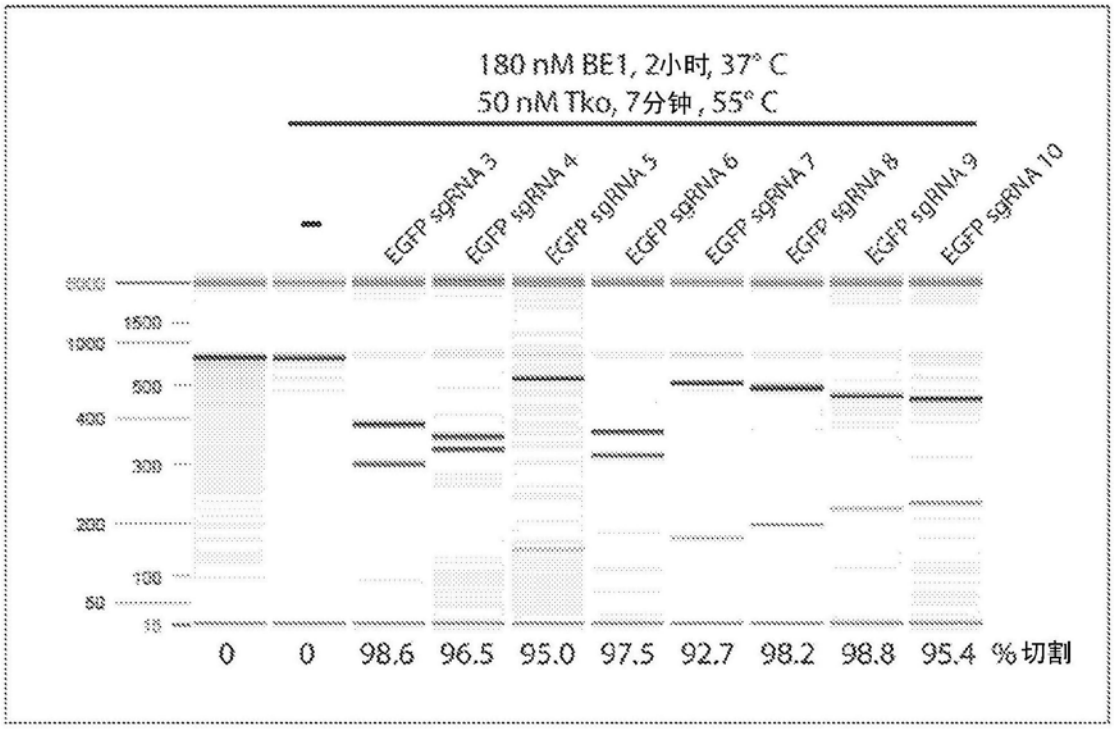
180 nM BE1, 2小时, 37° C
50 nM Tko, 7分钟, 55° C
—
EGFP sgRNA 3
EGFP sgRNA 4
EGFP sgRNA 5
EGFP sgRNA 6
EGFP sgRNA 7
EGFP sgRNA 8
EGFP sgRNA 9
EGFP sgRNA 10
1500
1000
800
400
300
200
100
50
15
0 0 98.6 96.5 95.0 97.5 92.7 98.2 98.8 95.4 %切割

图25

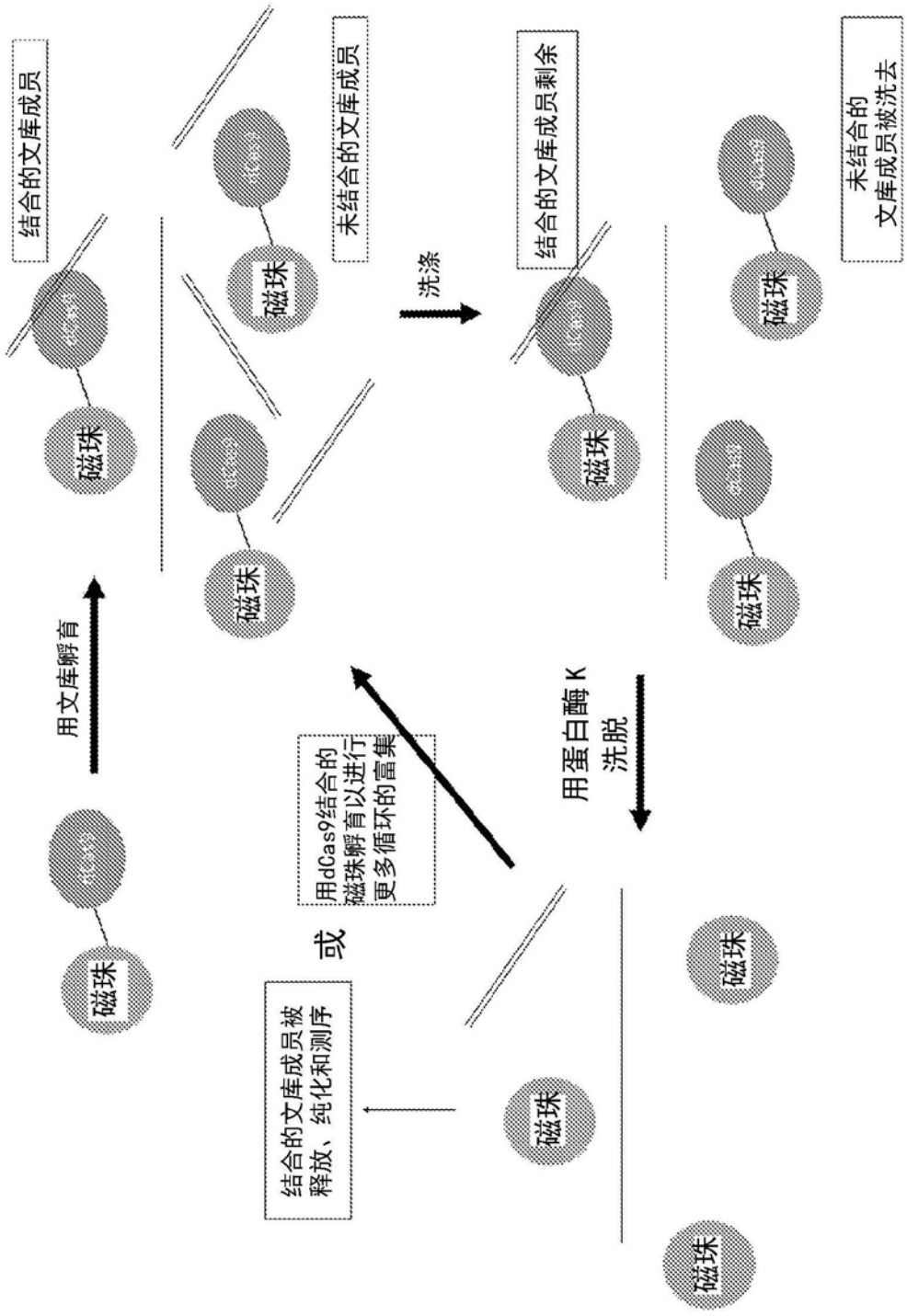
结合的文库成员
磁珠
用文库孵育
未结合的文库成员
磁珠
洗涤
结合的文库成员剩余
磁珠
未结合的
文库成员被洗去
磁珠
用dCas9结合的
磁珠孵育以进行
更多循环的富集
或
用蛋白酶 K
洗脱
结合的文库成员被
释放、纯化和测序
磁珠
磁珠
磁珠

图26

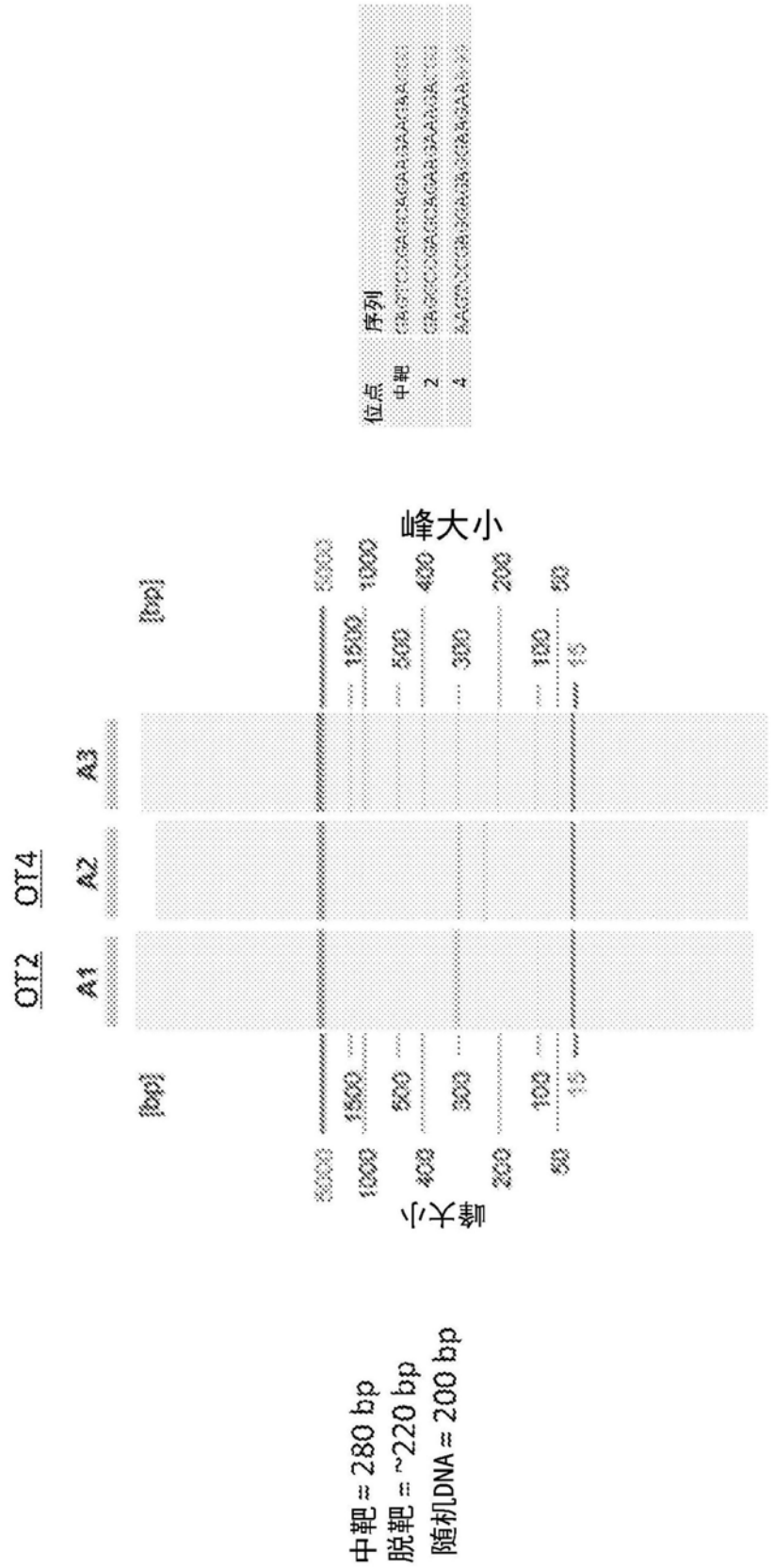
位点 序列
中靶
2
4
峰大小
(bp)
5000
1500
1000
500
400
300
200
100
50
15
OT2
OT4
A1
A2
A3
中靶 = 280 bp
脱靶 = ~220 bp
随机DNA = 200 bp

图27

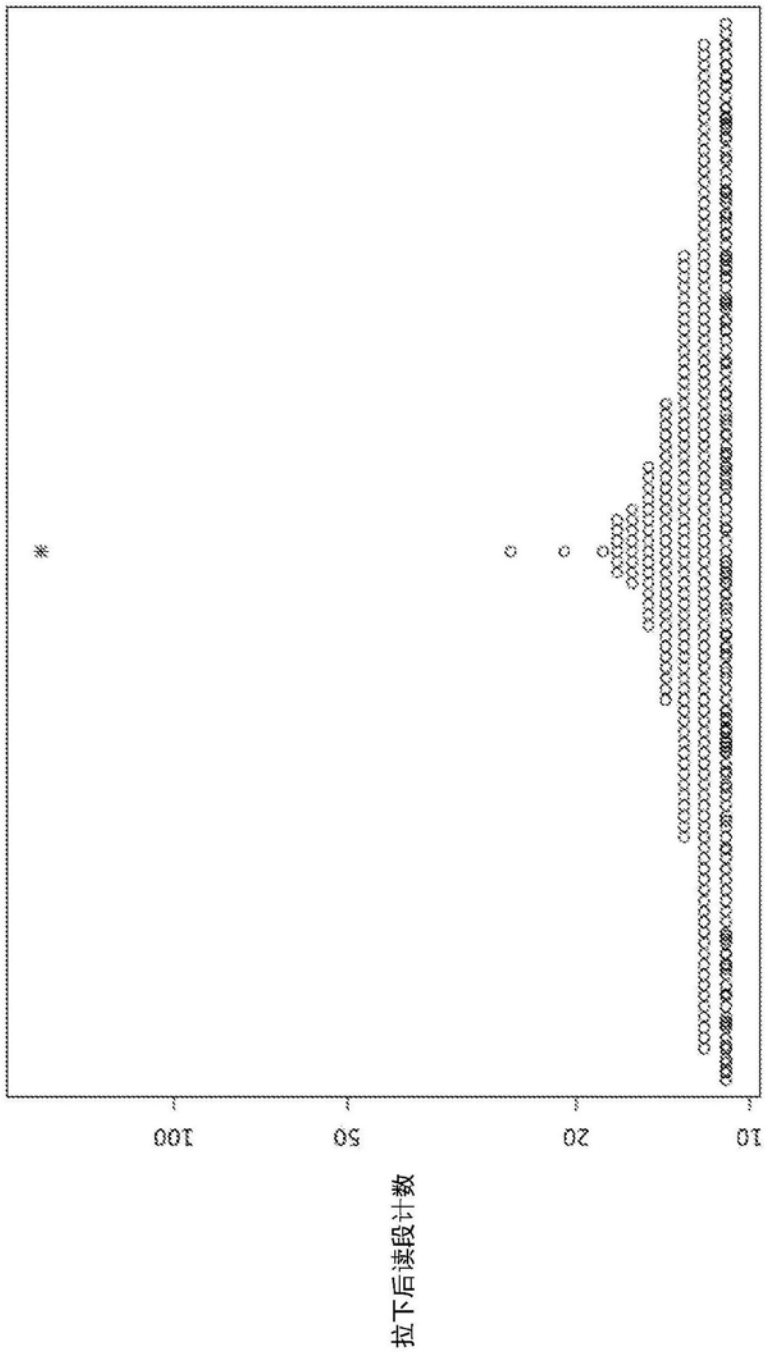

图28

图29