(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116515797 A
(43)申请公布日 2023.08.01

(21)申请号 202310441977.6

(22)申请日 2020.02.14

(30)优先权数据
62/805,868 2019.02.14 US
62/805,878 2019.02.14 US
62/805,899 2019.02.14 US
62/874,414 2019.07.15 US

(62)分案原申请数据
202080028787.6 2020.02.14

(71)申请人 宏基因组学公司
地址 美国加利福尼亚州

(72)发明人 布莱恩·托马斯
克利斯多佛·布朗 罗斯·坎托尔
奥德拉·德沃托
克里斯蒂娜·布特弗尔德
利萨·亚历山大
丹妮拉·S·A·戈尔茨曼
詹森·刘

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
专利代理师 韦昌金 武晶晶

(51)Int.Cl.
C12N 9/22(2006.01)
C12N 15/113(2010.01)
C12N 15/10(2006.01)
C12N 15/90(2006.01)

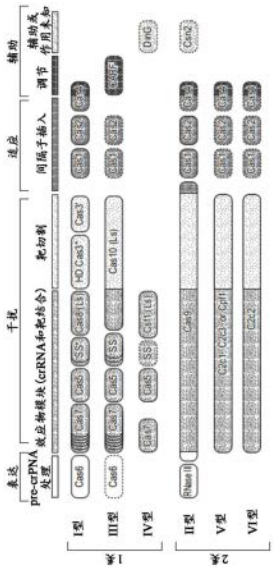
权利要求书2页 说明书87页 附图42页

(54)发明名称

具有RUVC结构域的酶

(57)摘要

本公开提供具有区别性结构域特征的核酸内切酶,以及使用此类酶或其变体的方法。



CN 116515797 A

1.一种工程改造的核酸酶组合物,其包含:

(a)包含RuvC_III结构域的核酸内切酶,所述RuvC_III结构域包含与SEQ ID NO 2242具有至少90%序列同一性的序列;和

(b)工程改造的指导核糖核酸结构,其被构造成与所述核酸内切酶形成复合物,所述复合物包含:

(i)指导核糖核酸序列,其被构造成与靶脱氧核糖核酸序列杂交;和

(ii)tracr核糖核酸序列,其被构造成与所述核酸内切酶结合,

其中所述核酸内切酶包含与SEQ ID NO 421具有至少90%序列同一性的序列。

2.如权利要求1所述的工程改造的核酸酶组合物,其中所述核酸内切酶进一步包含HNH结构域。

3.如权利要求1所述的工程改造的核酸酶组合物,其中所述tracr核糖核酸序列包含与来自SEQ ID NO 5495的约60至90个连续核苷酸具有至少80%序列同一性的序列。

4.如权利要求1所述的工程改造的核酸酶组合物,其中所述核酸内切酶被构造成与SEQ ID NO 5517的前间区序列邻近基序(PAM)序列结合。

5.如权利要求1所述的工程改造的核酸酶组合物,其中所述工程改造的指导核糖核酸结构包含至少两个核糖核酸多核苷酸。

6.如权利要求1所述的工程改造的核酸酶组合物,其中所述工程改造的指导核糖核酸结构包含一个包含所述指导核糖核酸序列和所述tracr核糖核酸序列的核糖核酸多核苷酸。

7.如权利要求1所述的工程改造的核酸酶组合物,其中所述指导核糖核酸序列与原核、细菌、古细菌、真核、真菌、植物、哺乳动物或人基因组序列互补。

8.如权利要求1所述的工程改造的核酸酶组合物,其中所述核酸内切酶包含一个或多个在所述核酸内切酶的N末端或C末端的近端的核定位序列(NLS)。

9.如权利要求8所述的工程改造的核酸酶组合物,其中所述一个或多个NLS包含SEQ ID NO 5597的序列。

10.如权利要求1所述的工程改造的核酸酶组合物,进一步包含:

单链或双链DNA修复模板,其从5'至3'包含:包含所述靶脱氧核糖核酸序列5'的至少20个核苷酸的序列的第一同源臂,至少10个核苷酸的合成DNA序列,以及包含所述靶脱氧核糖核酸序列3'的至少20个核苷酸的序列的第二同源臂。

11.如权利要求1所述的工程改造的核酸酶组合物,其中所述工程改造的核酸酶系统进一步包含Mg^{2+}源。

12.如权利要求1所述的工程改造的核酸酶组合物,其中所述核酸内切酶和所述tracr核糖核酸序列源自同一门内的不同细菌物种。

13.如权利要求1所述的工程改造的核酸酶组合物,其中所述指导RNA结构包含被预测为包含由杆和环组成的发夹的RNA序列,其中所述杆包含至少10个碱基配对的核糖核苷酸,以及所述环的4个碱基对内的不对称凸起。

14.如权利要求1所述的工程改造的核酸酶组合物,其中所述工程改造的指导核糖核酸结构的所述tracr核糖核酸序列包含发夹,所述发夹包含至少8个碱基配对的核糖核苷酸。

15.如权利要求1所述的工程改造的核酸酶组合物,其中所述工程改造的指导核糖核酸

结构的所述指导核糖核酸序列被预测为包含具有不间断碱基配对区的发夹,所述指导核糖核酸结构包含至少8个核苷酸的指导核糖核酸序列和至少8个核苷酸的tracr核糖核酸序列,并且其中所述tracr核糖核酸序列从5'至3'包含第一发夹和第二发夹,其中所述第一发夹具有比所述第二发夹更长的杆。

16.如权利要求1所述的工程改造的核酸酶组合物,其中所述核酸内切酶包含SEQ ID NO 421或与其具有至少90%同一性的保守取代序列。

具有RUVC结构域的酶

[0001] 本申请是申请日为2020年02月14日、申请号为202080028787.6、发明名称为“具有RUVC结构域的酶”的中国专利申请(其对应PCT申请的申请日为2020年02月14日、申请号为PCT/US2020/018432)的分案申请。

[0002] 交叉引用

[0003] 本申请要求于2019年2月14日提交的名称为“MG1 ENZYMES WITH RUVC DOMAINS”的美国临时申请第62/805,868号、于2019年7月15日提交的名称为“MG1 ENZYMES WITH RUVC DOMAINS”的美国临时申请第62/874,414号、于2019年2月14日提交的名称为“MG2 ENZYMES CONTAINING RUVC DOMAINS”的美国临时申请第62/805,878号以及于2019年2月14日提交的名称为“MG3ENZYMES WITH RUVC DOMAINS”的美国临时申请第62/805,899号的权益,其中每一个都通过引用完整并入本文。

背景技术

[0004] Cas酶及其相关的成簇规则间隔短回文重复(CRISPR)指导核糖核酸(RNA)似乎是原核免疫系统的普遍(约45%的细菌,约84%的古细菌)组分,用于保护此类微生物通过CRISPR-RNA指导的核酸切割来对抗非自身核酸,例如传染性病毒和质粒。虽然编码CRISPR RNA元件的脱氧核糖核酸(DNA)元件在结构和长度上可以相对保守,但它们的CRISPR相关(Cas)蛋白高度多样化,包含广泛多种核酸相互作用结构域。虽然早在1987年就已观察到CRISPR DNA元件,但CRISPR/Cas复合物的可编程核酸内切酶切割能力直到最近才被认识到,导致在各种DNA操作和基因编辑应用中使用重组CRISPR/Cas系统。

[0005] 序列表

[0006] 本申请包含序列表,所述序列表已以ASCII格式电子提交,并通过引用整体并入本文。创建于2020年2月13日的所述ASCII副本被命名为55921-703_601_SL.txt,并且大小是23,363,113字节。

发明内容

[0007] 在一些方面,本公开提供了一种工程改造的核酸酶系统,其包含:(a)包含RuvC_III结构域和HNH结构域的核酸内切酶,其中所述核酸内切酶源自未培养的微生物,其中所述核酸内切酶是2类II型Cas核酸内切酶;和(b)工程改造的指导核糖核酸结构,其被构造成与所述核酸内切酶形成复合物,所述复合物包含:(i)指导核糖核酸序列,其被构造成与靶脱氧核糖核酸序列杂交;和(ii)tracr核糖核酸序列,其被构造成与所述核酸内切酶结合。在一些实施方案中,RuvC_III结构域包含与SEQ ID NO:1827-3637中任一个具有至少70%、至少75%、至少80%或至少90%序列同一性的序列。

[0008] 在一些方面,本公开提供了一种工程改造的核酸酶系统,其包含:(a)包含与SEQ ID NO:1827-3637中任一个具有至少75%序列同一性的RuvC_III结构域的核酸内切酶;和(b)工程改造的指导核糖核酸结构,其被构造成与所述核酸内切酶形成复合物,所述复合物包含:(i)指导核糖核酸序列,其被构造成与靶脱氧核糖核酸序列杂交;和(ii)tracr核糖核

酸序列,其被构造成与所述核酸内切酶结合。

[0009] 在一些方面,本公开提供了一种工程改造的核酸酶系统,其包含:(a)被构造成与包含SEQ ID NO:5512-5537的前间区序列邻近基序(PAM)序列结合的核酸内切酶,其中所述核酸内切酶是2类II型Cas核酸内切酶;和(b)工程改造的指导核糖核酸结构,其被构造成与所述核酸内切酶形成复合物,所述复合物包含:(i)指导核糖核酸序列,其被构造成与靶脱氧核糖核酸序列杂交;和(ii)tracr核糖核酸序列,其被构造成与所述核酸内切酶结合。

[0010] 在一些实施方案中,所述核酸内切酶源自未培养的微生物。在一些实施方案中,所述核酸内切酶未被工程改造以结合不同的PAM序列。在一些实施方案中,所述核酸内切酶不是Cas9核酸内切酶、Cas14核酸内切酶、Cas12a核酸内切酶、Cas12b核酸内切酶、Cas12c核酸内切酶、Cas12d核酸内切酶、Cas12e核酸内切酶、Cas13a核酸内切酶、Cas13b核酸内切酶、Cas13c核酸内切酶或Cas13d核酸内切酶。在一些实施方案中,所述核酸内切酶与Cas9核酸内切酶具有小于80%的同一性。在一些实施方案中,所述核酸内切酶进一步包含HNH结构域。在一些实施方案中,所述tracr核糖核酸序列包含与选自SEQ ID NO:5476-5511和SEQ ID NO:5538中任一个的约60至90个连续核苷酸具有至少80%序列同一性的序列。

[0011] 在一些方面,本公开提供了一种工程改造的核酸酶系统,其包含:(a)工程改造的指导核糖核酸结构,其包含:(i)指导核糖核酸序列,其被构造成与靶脱氧核糖核酸序列杂交;和(ii)tracr核糖核酸序列,其被构造成与核酸内切酶结合,其中所述tracr核糖核酸序列包含与选自SEQ ID NO:5476-5511和SEQ ID NO:5538中任一个的约60至90个连续核苷酸具有至少80%序列同一性的序列;和(b)2类II型Cas核酸内切酶,其被构造成与所述工程改造的指导核糖核酸结合。在一些实施方案中,所述核酸内切酶被构造成与选自包含SEQ ID NO:5512-5537的组的前间区序列邻近基序(PAM)序列结合。

[0012] 在一些实施方案中,所述工程改造的指导核糖核酸结构包含至少两个核糖核酸多核苷酸。在一些实施方案中,所述工程改造的指导核糖核酸结构包含一个包含所述指导核糖核酸序列和所述tracr核糖核酸序列的核糖核酸多核苷酸。

[0013] 在一些实施方案中,指导核糖核酸序列与原核、细菌、真核、古细菌、真菌、植物、哺乳动物或人基因组序列互补。在一些实施方案中,指导核糖核酸序列的长度为15-24个核苷酸。在一些实施方案中,所述核酸内切酶包含一个或多个在所述核酸内切酶的N末端或C末端的近端的核定位序列(NLS)。在一些实施方案中,所述NLS包含选自SEQ ID NO:5597-5612的序列。

[0014] 在一些实施方案中,工程改造的核酸酶系统进一步包含单链或双链DNA修复模板,其从5'至3'包含:包含所述靶脱氧核糖核酸序列5'的至少20个核苷酸的序列的第一同源臂,至少10个核苷酸的合成DNA序列,以及包含所述靶序列3'的至少20个核苷酸的序列的第二同源臂。在一些实施方案中,第一同源臂或所述第二同源臂包含至少40、80、120、150、200、300、500或1,000个核苷酸的序列。

[0015] 在一些实施方案中,系统进一步包含Mg^{2+}源。

[0016] 在一些实施方案中,核酸内切酶和tracr核糖核酸序列源自同一门内的不同细菌物种。在一些实施方案中,核酸内切酶源自属于皮杆菌属的细菌。在一些实施方案中,所述核酸内切酶源自属于疣微菌门(Phylum Verrucomicrobia)、暂定异域菌门(Phylum Candidatus Peregrinibacteria)或暂定黑水仙菌门(Phylum Candidatus

Melainabacteria)的细菌。在一些实施方案中,核酸内切酶源自包含与SEQ ID NO:5592-5595中任一个具有至少90%同一性的16S rRNA基因的细菌。

[0017] 在一些实施方案中,HNH结构域包含与SEQ ID NO:5638-5460中任一个具有至少70%或至少80%同一性的序列。在一些实施方案中,核酸内切酶包含SEQ ID NO:1-1826或与其具有至少55%同一性的变体。在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:1827-1830或SEQ ID NO:1827-2140的序列至少70%、80%或90%相同的序列。

[0018] 在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:3638-3641或SEQ ID NO:3638-3954的序列至少70%、80%或90%相同的序列。在一些实施方案中,核酸内切酶包含至少1个、至少2个、至少3个、至少4个或至少5个选自SEQ ID NO:5615-5632的肽基序。在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:1-4或SEQ ID NO:1-319的序列至少70%、80%或90%相同的序列。

[0019] 在一些实施方案中,指导RNA结构包含与选自SEQ ID NO:5461-5464、SEQ ID NO:5476-5479或SEQ ID NO:5476-5489的序列至少70%、80%或90%相同的序列。在一些实施方案中,指导RNA结构包含被预测为包含由杆和环组成的发夹的RNA序列,所述杆包含至少10个、至少12个或至少14个碱基配对的核糖核苷酸,以及所述环的4个碱基对内的不对称凸起。

[0020] 在一些实施方案中,核酸内切酶被构造成与包含选自SEQ ID NO:5512-5515或SEQ ID NO:5527-5530的序列的PAM结合。

[0021] 在一些实施方案中:(a)所述核酸内切酶包含与SEQ ID NO:1827至少70%、至少80%或至少90%相同的序列;(b)所述指导RNA结构包含与SEQ ID NO:5461或SEQ ID NO:5476中的至少一个至少70%、至少80%或至少90%相同的序列;和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5512或SEQ ID NO:5527的PAM结合。在一些实施方案中:(a)所述核酸内切酶包含与SEQ ID NO:1828至少70%、至少80%或至少90%相同的序列;(b)所述指导RNA结构包含与SEQ ID NO:5462或SEQ ID NO:5477中的至少一个至少70%、至少80%或至少90%相同的序列;和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5513或SEQ ID NO:5528的PAM结合。在一些实施方案中:(a)所述核酸内切酶包含与SEQ ID NO:1829至少70%、至少80%或至少90%相同的序列;(b)所述指导RNA结构包含与SEQ ID NO:5463或SEQ ID NO:5478中的至少一个至少70%、至少80%或至少90%相同的序列;和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5514或SEQ ID NO:5529的PAM结合。在一些实施方案中:

[0022] (a)所述核酸内切酶包含与SEQ ID NO:1830至少70%、至少80%或至少90%相同的序列;(b)所述指导RNA结构包含与SEQ ID NO:5464或SEQ ID NO:5479中的至少一个至少70%、至少80%或至少90%相同的序列;和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5515或SEQ ID NO:5530的PAM结合。

[0023] 在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:2141-2142或SEQ ID NO:2141-2241序列至少70%、80%或90%相同的序列。在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:3955-3956或SEQ ID NO:3955-4055的序列至少70%、80%或90%相同的序列。在一些实施方案中,核酸内切酶包含选自SEQ ID NO:5632-5638的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:320-321或SEQ ID NO:320-420中的序列至少70%、80%或90%相同的序列。在一些实施方

案中,指导RNA结构包含与选自以下的序列至少70%、80%或90%相同的序列:SEQ ID NO:5465、SEQ ID NO:5490-5491或SEQ ID NO:5490-5494。在一些实施方案中,指导RNA结构包含含有发夹的tracr核糖核酸序列,所述发夹包含至少8个、至少10个或至少12个碱基配对的核糖核苷酸。在一些实施方案中,核酸内切酶被构造成与包含选自SEQ ID NO:5516和SEQ ID NO:5531的序列的PAM结合。在一些实施方案中:(a)所述核酸内切酶包含与SEQ ID NO:2141至少70%、80%或90%相同的序列;(b)所述指导RNA结构包含与

[0024] SEQ ID NO:5490至少70%、80%或90%相同的序列;和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5531的PAM结合。在一些实施方案中:(a)所述核酸内切酶包含与SEQ ID NO:2142至少70%、80%或90%相同的序列;(b)所述指导RNA结构包含与SEQ ID NO:5465或SEQ ID NO:5491至少70%、80%或90%相同的序列;和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5516的PAM结合。

[0025] 在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:2245-2246中的序列至少70%、80%或90%相同的序列。在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:4059-4060的序列至少70%、80%或90%相同的序列。在一些实施方案中,核酸内切酶包含选自

[0026] SEQ ID NO:5639-5648的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:424-425的序列至少70%、80%或90%相同的序列。在一些实施方案中,指导RNA结构包含与选自SEQ ID NO:5498-5499和SEQ IDNO:5539的序列至少70%、80%或90%相同的序列。在一些实施方案中,指导RNA结构包含被预测为包含具有不间断碱基配对区的发夹的指导核糖核酸序列,所述指导核糖核酸结构包含至少8个核苷酸的指导核糖核酸序列和至少8个核苷酸的tracr核糖核酸序列,并且其中所述tracr核糖核酸序列从5'至3'包含第一发夹和第二发夹,其中所述第一发夹具有比所述第二发夹更长的杆。

[0027] 在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:2242-2244或SEQ ID NO:2247-2249的序列至少70%、80%或90%相同的序列。在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:4056-4058和SEQ ID NO:4061-4063的序列至少70%、80%或90%相同的序列。在一些实施方案中,核酸内切酶包含选自SEQ ID NO:5639-5648的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。在一些实施方案中,核酸内切酶包含与选自SEQ ID NO:421-423或SEQ ID NO:426-428的序列至少70%、80%或90%相同的序列。在一些实施方案中,指导RNA结构包含与选自以下的序列至少70%、80%或90%相同的序列:SEQ ID NO:5466-5467、SEQ ID NO:5495-5497、SEQ ID NO:5500-5502和SEQ ID NO:5539。在一些实施方案中,指导RNA结构包含被预测为包含具有不间断碱基配对区的发夹的指导核糖核酸序列,所述指导核糖核酸结构包含至少8个核苷酸的指导核糖核酸序列和至少8个核苷酸的tracr核糖核酸序列,并且其中所述tracr核糖核酸序列从5'至3'包含第一发夹和第二发夹,其中所述第一发夹具有比所述第二发夹更长的杆。在一些实施方案中,核酸内切酶被构造成与包含选自SEQ ID NO:5517-5518或SEQ ID NO:5532-5534的序列的PAM结合。在一些实施方案中:(a)所述核酸内切酶包含与SEQ ID NO:2247至少70%、80%或90%相同的序列;(b)所述指导RNA结构包含与SEQ ID NO:5500至少70%、80%或90%相同的序列;和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5517或SEQ ID NO:5532的PAM结合。在一些实

施方案中：(a)所述核酸内切酶包含与SEQ ID NO:2248至少70％、80％或90％相同的序列；(b)所述指导RNA结构包含与SEQ ID NO:5501至少70％、80％或90％相同的序列；和

[0028] (c)所述核酸内切酶被构造成与包含SEQ ID NO:5518或SEQ ID NO:5533的PAM结合。在一些实施方案中：(a)所述核酸内切酶包含与SEQ ID NO:2249至少70％、80％或90％相同的序列；(b)所述指导RNA结构包含与SEQ ID NO:5502至少70％、80％或90％相同的序列；和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5534的PAM结合。

[0029] 在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:2253或SEQ ID NO:2253-2481的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:4067或SEQ ID NO:4067-4295的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含根据SEQ ID NO:5649的肽基序。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:432或SEQ ID NO:432-660中的序列至少70％、80％或90％相同的序列。在一些实施方案中，指导RNA结构包含与选自SEQ ID NO:ID NO:5468或SEQ ID NO:5503的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶被构造成与包含选自SEQ ID NO:5519的序列的PAM结合。在一些实施方案中：(a)所述核酸内切酶包含与SEQ ID NO:2253至少70％、80％或90％相同的序列；(b)所述指导RNA结构包含与SEQ ID NO:5468或SEQ ID NO:5503至少70％、80％或90％相同的序列；和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5519的PAM结合。

[0030] 在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:2482-

[0031] 2489中的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:4296-4303的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:661-668的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:2490-2498的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:4304-4312的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:669-677的序列至少70％、80％或90％相同的序列。在一些实施方案中，指导RNA结构包含与选自SEQ ID NO:5504的序列至少70％、80％或90％相同的序列。

[0032] 在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:2499或SEQ ID NO:2499-2750的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:4313或SEQ ID NO:4313-4564的序列至少70％、80％或90％相同的序列。在一些实施方案中，所述核酸内切酶包含选自SEQ ID NO:5650-5667的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:678或SEQ ID NO:678-

[0033] 929中的序列至少70％、80％或90％相同的序列。在一些实施方案中，指导RNA结构包含与SEQ ID NO:5469或SEQ ID NO:5505至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶被构造成与包含SEQ ID NO:5520或SEQ ID NO:5535的PAM结合。在一些实施方案中：(a)所述核酸内切酶包含与SEQ ID NO:2499至少70％、80％或90％相同的序列；(b)所述指导RNA结构包含与SEQ ID NO:5469或SEQ ID NO:5505至少70％、80％或90％相同的序列；和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5520或SEQ ID NO:5535的PAM结合。

[0034] 在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:2751或SEQ ID NO:2751-2913的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:4565或SEQ ID NO:4565-4727的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含选自SEQ ID NO:5668-5678的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:930或SEQ ID NO:930-1092中的序列至少70％、80％或90％相同的序列。在一些实施方案中，指导RNA结构包含与SEQ ID NO:5470或SEQ ID NO:5506至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶被构造成与包含选自SEQ ID NO:5521或SEQ ID NO:5536的序列的PAM结合。在一些实施方案中：(a)所述核酸内切酶包含与SEQ ID NO:2751至少70％、80％或90％相同的序列；(b)所述指导RNA结构包含与SEQ ID NO:5470或SEQ ID NO:5506至少70％、80％或90％相同的序列；和c)所述核酸内切酶被构造成与包含SEQ ID NO:5521或SEQ ID NO:5536的PAM结合。

[0035] 在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:2914或SEQ ID NO:2914-3174的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:4728或SEQ ID NO:4728-4988的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含选自SEQ ID NO:5676-5678的至少1个、至少2个或至少3个肽基序。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:1093或SEQ ID NO:1093-1353中的序列至少70％、80％或90％相同的序列。在一些实施方案中，指导RNA结构包含与选自以下的序列至少70％、80％或90％相同的序列：SEQ ID NO:5471、SEQ ID NO:5507和SEQ ID NO:5540-5542。在一些实施方案中，指导RNA结构包含被预测为包含至少两个发夹的tracr核糖核酸序列，所述发夹包含少于5个碱基配对的核糖核苷酸。在一些实施方案中，核酸内切酶被构造成与包含SEQ ID NO:5522的PAM结合。在一些实施方案中：(a)所述核酸内切酶包含与SEQ ID NO:2914至少70％、80％或90％相同的序列；(b)所述指导RNA结构包含与SEQ ID NO:5471或SEQ ID NO:5507至少70％、80％或90％相同的序列；和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5522的PAM结合。

[0036] 在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:3175或SEQ ID NO:3175-3330的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:4989或SEQ ID NO:4989-5146的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含选自SEQ ID NO:5679-5686的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:1354或SEQ ID NO:1354-1511中的序列至少70％、80％或90％相同的序列。在一些实施方案中，指导RNA结构包含与选自SEQ ID NO:5472或SEQ ID NO:5508的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶被构造成与包含选自SEQ ID NO:5523或SEQ ID NO:5537的序列的PAM结合。在一些实施方案中：(a)所述核酸内切酶包含与SEQ ID NO:3175至少70％、80％或90％相同的序列；(b)所述指导RNA结构包含与SEQ ID NO:5472或SEQ ID NO:5508至少70％、80％或90％相同的序列；和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5523或SEQ ID NO:5537的PAM结合。

[0037] 在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:3331或SEQ ID NO:3331-3474的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自

SEQ ID NO:5147或SEQ ID NO:5147-5290的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含选自SEQ ID NO:5674-5675和SEQ ID NO:5687-5693的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:1512或SEQ ID NO:1512-1655中的序列至少70％、80％或90％相同的序列。在一些实施方案中，指导RNA结构包含与选自SEQ ID NO:5473或SEQ ID NO:5509的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶被构造成与包含SEQ ID NO:5524的PAM结合。在一些实施方案中：(a)所述核酸内切酶包含与SEQ ID NO:3331至少70％、80％或90％相同的序列；(b)所述指导RNA结构包含与SEQ ID NO:5473或SEQ ID NO:5509至少70％、80％或90％相同的序列；和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5524的PAM结合。

[0038] 在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:3475或SEQ ID NO:3475-3568的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:5291或SEQ ID NO:5291-5389的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含选自SEQ ID NO:5694-5699的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:1656或SEQ ID NO:1656-1755的序列至少70％、80％或90％相同的序列。在一些实施方案中，指导RNA结构包含与SEQ ID NO:5474或SEQ ID NO:5510至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶被构造成与包含SEQ ID NO:5525的PAM结合。在一些实施方案中：(a)所述核酸内切酶包含与SEQ ID NO:3475至少70％、80％或90％相同的序列；

[0039] (b)所述指导RNA结构包含与SEQ ID NO:5474或SEQ ID NO:5510至少70％、80％或90％相同的序列；和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5525的PAM结合。

[0040] 在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:3569或SEQ ID NO:3569-3637的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:5390或SEQ ID NO:5390-5460的序列至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶包含选自SEQ ID NO:5700-5717的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。在一些实施方案中，核酸内切酶包含与选自SEQ ID NO:1756或SEQ ID NO:1756-1826中的序列至少70％、80％或90％相同的序列。在一些实施方案中，指导RNA结构包含与SEQ ID NO:5475或SEQ ID NO:5511至少70％、80％或90％相同的序列。在一些实施方案中，核酸内切酶被构造成与包含SEQ ID NO:5526的PAM结合。在一些实施方案中：(a)所述核酸内切酶包含与SEQ ID NO:3569至少70％、80％或90％相同的序列；(b)所述指导RNA结构包含与SEQ ID NO:5475或SEQ ID NO:5511至少70％、80％或90％相同的序列；和(c)所述核酸内切酶被构造成与包含SEQ ID NO:5526的PAM结合。在一些实施方案中，序列同一性由BLASTP、CLUSTALW、MUSCLE、MAFFT或Smith-Waterman同源性搜索算法确定。在一些实施方案中，序列同一性通过所述BLASTP同源性搜索算法来确定，所述BLASTP同源性搜索算法使用3的字长(W)、10的期望(E)的参数和设置缺口存在时罚分为11、延伸为1的BLOSUM62评分矩阵并使用条件型组成得分矩阵调整。

[0041] 在一些方面，本公开提供了一种工程改造的指导核糖核酸多核苷酸，其包含：(a)DNA靶向区段，其包含与靶DNA分子中的靶序列互补的核苷酸序列；和(b)蛋白质结合区段，其包含杂交形成双链RNA(dsRNA)双链体的两个互补核苷酸段，其中所述两个互补核苷酸段

通过居间核苷酸共价连接到彼此,以及其中所述工程改造的指导核糖核酸多核苷酸被构造成与包含与SEQ ID NO:1827-3637中任一个具有至少75%序列同一性的RuvC_III结构域的核酸内切酶形成复合物并将所述复合物靶向所述靶DNA分子的所述靶序列。在一些实施方案中,DNA靶向区段位于所述两个互补核苷酸段的5'。

[0042] 在一些实施方案中:(a)所述蛋白质结合区段包含与选自

[0043] SEQ ID NO:5476-5479或SEQ ID NO:5476-5489的序列具有至少70%、至少80%或至少90%同一性的序列;(b)所述蛋白质结合区段包含与选自(SEQ ID NO:5490-5491或SEQ ID NO:5490-5494)和SEQ ID NO:5538的序列具有至少70%、至少80%或至少90%同一性的序列;(c)所述蛋白质结合区段包含与选自SEQ ID NO:5498-5499的序列具有至少70%、至少80%或至少90%同一性的序列;(d)所述蛋白质结合区段包含与选自SEQ ID NO:5495-5497和SEQ ID NO:5500-5502的序列具有至少70%、至少80%或至少90%同一性的序列;(e)所述蛋白质结合区段包含与SEQ ID NO:5503具有至少70%、至少80%或至少90%同一性的序列;(f)所述蛋白质结合区段包含与SEQ ID NO:5504具有至少70%、至少80%或至少90%同一性的序列;(g)所述蛋白质结合区段包含与SEQ ID NO:5505具有至少70%、至少80%或至少90%同一性的序列;(h)所述蛋白质结合区段包含与SEQ ID NO:5506具有至少70%、至少80%或至少90%同一性的序列;(i)所述蛋白质结合区段包含与SEQ ID NO:5507具有至少70%、至少80%或至少90%同一性的序列;(j)所述蛋白质结合区段包含与SEQ ID NO:5508具有至少70%、至少80%或至少90%同一性的序列;(k)所述蛋白质结合区段包含与SEQ ID NO:5509具有至少70%、至少80%或至少90%同一性的序列;(l)所述蛋白质结合区段包含与SEQ ID NO:5510具有至少70%、至少80%或至少90%同一性的序列;或(m)所述蛋白质结合区段包含与SEQ ID NO:5511具有至少70%、至少80%或至少90%同一性的序列。

[0044] 在一些实施方案中:(a)所述指导核糖核酸多核苷酸包含含有发夹的RNA序列,所述发夹包含杆和环,其中所述杆包含至少10个、至少12个或至少14个碱基配对的核糖核苷酸和所述环的4个碱基对内的不对称凸起;(b)所述指导核糖核酸多核苷酸包含被预测包含发夹的tracr核糖核酸序列,所述发夹包含至少8个、至少10个或至少12个碱基配对的核糖核苷酸;(c)所述指导RNA结构包含被预测包含具有不间断碱基配对区的发夹的指导核糖核酸序列,所述指导核糖核酸结构包含至少8个核苷酸的指导核糖核酸序列和至少8个核苷酸的tracr核糖核酸序列,并且其中所述tracr核糖核酸序列从5'至3'包含第一发夹和第二发夹,所述第一发夹具有比所述第二发夹更长的杆;或(d)所述指导核糖核酸多核苷酸包含被预测包含至少两个发夹的tracr核糖核酸序列,所述发夹包含少于5个碱基配对的核糖核苷酸。

[0045] 在一些方面,本公开提供了一种脱氧核糖核酸多核苷酸,其编码本文所述的任何的工程改造的指导核糖核酸多核苷酸。

[0046] 在一些方面,本公开提供了一种包含优化用于在生物体中表达的工程改造的核酸序列的核酸,其中所述核酸编码包含RuvC_III结构域和HNH结构域的2类II型Cas核酸内切酶,并且其中所述核酸内切酶源自未培养的微生物。

[0047] 在一些方面,本公开提供了一种包含优化用于在生物体中表达的工程改造的核酸序列的核酸,其中所述核酸编码包含与SEQ ID NO:1827-3637中任一个具有至少70%序列

同一性的RuvC_III结构域的核酸内切酶。在一些实施方案中，核酸内切酶包含与SEQ ID NO:3638-5460中任一个具有至少70％或至少80％序列同一性的HNH结构域。在一些实施方案中，核酸内切酶包含SEQ ID NO:5572-5591或与其具有至少70％序列同一性的变体。在一些实施方案中，核酸内切酶包含对所述核酸内切酶的N末端或C末端的近端的一个或多个核定位序列(NLS)进行编码的序列。在一些实施方案中，NLS包含选自SEQ ID NO:5597-5612的序列。

[0048] 在一些实施方案中，生物体是原核生物体、细菌生物体、真核生物体、真菌生物体、植物生物体、哺乳动物生物体、啮齿动物生物体或人。在一些实施方案中，生物体是大肠杆菌，并且：(a)所述核酸序列与选自SEQ ID NO:5572-5575的序列具有至少70％、80％或90％的同一性；(b)所述核酸序列与选自SEQ ID NO:5576-5577的序列具有至少70％、80％或90％的同一性；(c)所述核酸序列与选自SEQ ID NO:5578-5580的序列具有至少70％、80％或90％的同一性；(d)所述核酸序列与SEQ ID NO:5581具有至少70％、80％或90％的同一性；(e)所述核酸序列与SEQ ID NO:5582具有至少70％、80％或90％的同一性；(f)所述核酸序列与SEQ ID NO:5583具有至少70％、80％或90％的同一性；(g)所述核酸序列与SEQ ID NO:5584具有至少70％、80％或90％的同一性；(h)所述核酸序列与SEQ ID NO:5585具有至少70％、80％或90％的同一性；(i)所述核酸序列与SEQ ID NO:5586具有至少70％、80％或90％的同一性；或(j)所述核酸序列与SEQ ID NO:5587具有至少70％、80％或90％的同一性。在一些实施方案中，生物体是人，并且：(a)所述核酸序列与SEQ ID NO:5588或SEQ ID NO:5589具有至少70％、80％或90％的同一性；或(b)所述核酸序列与SEQ ID NO:5590或SEQ ID NO:5591具有至少70％、80％或90％的同一性。

[0049] 在一些方面，本公开提供了一种载体，其包含编码包含RuvC_III结构域和HNH结构域的2类II型Cas核酸内切酶的核酸序列，其中所述核酸内切酶源自未培养的微生物。

[0050] 在一些方面，本公开提供了一种载体，其包含本文所述的任何核酸。在一些实施方案中，载体进一步包含编码工程改造的指导核糖核酸结构的核酸，所述工程改造的指导核糖核酸结构被构造成与所述核酸内切酶形成复合物，所述复合物包含：a)指导核糖核酸序列，其被构造成与靶脱氧核糖核酸序列杂交；和b)tracr核糖核酸序列，其被构造成与所述核酸内切酶结合。在一些实施方案中，载体是质粒、小环、CELiD、腺相关病毒(AAV)衍生的病毒体或慢病毒。

[0051] 在一些方面，本公开提供了一种细胞，其包含如本文所述的任何载体。

[0052] 在一些方面，本公开提供了一种制造核酸内切酶的方法，其包括培养如本文所述的任何细胞。

[0053] 在一些方面，本公开提供了一种用于对双链脱氧核糖核酸多核苷酸进行结合、切割、标记或修饰的方法，所述方法包括：(a)使所述双链脱氧核糖核酸多核苷酸与跟工程改造的指导核糖核酸结构复合的2类II型Cas核酸内切酶接触，所述工程改造的指导核糖核酸结构被构造成与所述核酸内切酶和所述双链脱氧核糖核酸多核苷酸结合；(b)其中所述双链脱氧核糖核酸多核苷酸包含前间区序列邻近基序(PAM)；以及(c)其中所述PAM包含选自SEQ ID NO:5512-5526或SEQ ID NO:5527-5537的序列。在一些实施方案中，双链脱氧核糖核酸多核苷酸包含第一链和第二链，所述第一链包含与所述工程改造的指导核糖核酸结构的序列互补的序列，所述第二链包含所述PAM。在一些实施方案中，PAM与跟所述工程改造的

指导核糖核酸结构的所述序列互补的所述序列的3'末端直接相邻。

[0054] 在一些实施方案中,2类II型Cas核酸内切酶不是Cas9核酸内切酶、Cas14核酸内切酶、Cas12a核酸内切酶、Cas12b核酸内切酶、Cas 12c核酸内切酶、Cas12d核酸内切酶、Cas12e核酸内切酶、Cas13a核酸内切酶、Cas13b核酸内切酶、Cas13c核酸内切酶或Cas 13d核酸内切酶。在一些实施方案中,2类II型Cas核酸内切酶源自未培养的微生物。在一些实施方案中,双链脱氧核糖核酸多核苷酸是真核、植物、真菌、哺乳动物、啮齿动物或人双链脱氧核糖核酸多核苷酸。

[0055] 在一些实施方案中:(a)所述PAM包含选自SEQ ID NO:5512-5515和SEQ ID NO:5527-5530的序列;(b)所述PAM包含SEQ ID NO:5516或SEQ ID NO:5531;(c)所述PAM包含SEQ ID NO:5539;(d)所述PAM包含SEQ ID NO:5517或SEQ ID NO:5518;(e)所述PAM包含SEQ ID NO:5519;(f)所述PAM包含SEQ ID NO:5520或SEQ ID NO:5535;(g)所述PAM包含SEQ ID NO:5521或SEQ ID NO:5536;(h)所述PAM包含SEQ ID NO:5522;(i)所述PAM包含SEQ ID NO:5523或SEQ ID NO:5537;(j)所述PAM包括SEQ ID NO:5524;(k)所述PAM包含SEQ ID NO:5525;或(l)所述PAM包含SEQ ID NO:5526。

[0056] 在一些方面,本公开提供了一种修饰靶核酸基因座的方法,所述方法包括将本文所述的任何工程改造的核酸酶系统递送至所述靶核酸基因座,其中所述核酸内切酶被构造成与所述工程改造的指导核糖核酸结构形成复合物,并且其中所述复合物被构造成使得在所述复合物与所述靶核酸基因座结合后,所述复合物修饰所述靶核酸基因座。在一些实施方案中,修饰所述靶核酸基因座包括对所述靶核酸基因座进行结合、切口、切割或标记。在一些实施方案中,靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。在一些实施方案中,靶核酸包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。在一些实施方案中,靶核酸基因座是在体外。在一些实施方案中,靶核酸基因座在细胞内。在一些实施方案中,细胞是原核细胞、细菌细胞、真核细胞、真菌细胞、植物细胞、动物细胞、哺乳动物细胞、啮齿动物细胞、灵长类细胞或人细胞。

[0057] 在一些实施方案中,将所述工程改造的核酸酶系统递送至所述靶核酸基因座包括递送如权利要求135-140任一项所述的核酸或如权利要求142-146任一项所述的载体。在一些实施方案中,将所述工程改造的核酸酶系统递送至所述靶核酸基因座包括递送包含编码所述核酸内切酶的开放阅读框的核酸。在一些实施方案中,核酸包含启动子,编码所述核酸内切酶的所述开放阅读框可操作地连接至所述启动子。在一些实施方案中,将所述工程改造的核酸酶系统递送至所述靶核酸基因座包括递送包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA。在一些实施方案中,将所述工程改造的核酸酶系统递送至所述靶核酸基因座包括递送翻译的多肽。在一些实施方案中,将所述工程改造的核酸酶系统递送至所述靶核酸基因座包括递送对与核糖核酸(RNA)pol III启动子可操作地连接的所述工程改造的指导核糖核酸结构进行编码的脱氧核糖核酸(DNA)。在一些实施方案中,核酸内切酶在所述靶基因座处或所述靶基因座的近端诱导单链断裂或双链断裂。

[0058] 从以下详细说明中,本公开的其他方面和优点对于本领域技术人员将变得显而易见,其中仅示出和描述了本公开的说明性实施方案。如将认识到的,本公开能够具有其他和不同的实施方案,并且所有这些在不脱离本公开的情况下其若干细节能够在各个明显方面进行修改。因此,应该认为附图和说明本质上都是示例性的,而不是限制性的。

[0059] 援引并入

[0060] 本说明书中提到的所有出版物、专利和专利申请均通过引用并入本文，其程度如同每个单独的出版物、专利或专利申请被明确地并单独地指出通过引用的方式并入一样。

附图说明

[0061] 本发明的新颖特征在所附的权利要求书中具体阐述。通过参考以下阐述其中利用了本发明的原理的说明性实施方案以及附图(在此也为“图”)的详细描述，可以更好地理解本发明的特征和优点，在附图中：

[0062] 图1描绘了不同类别和类型的CRISPR/Cas基因座的典型结构。

[0063] 图2描绘了天然2类/II型crRNA/tracrRNA对的架构，与其中两者连接的杂合sgRNA相比。

[0064] 图3描绘了显示编码来自MG1家族的酶的CRISPR基因座的结构的示意图。

[0065] 图4描绘了显示编码来自MG2家族的酶的CRISPR基因座的结构的示意图。

[0066] 图5描绘了显示编码来自MG3家族的酶的CRISPR基因座的结构的示意图。

[0067] 图6描绘了本公开的酶(MG1-1)与来自金黄色葡萄球菌的Cas9(SEQ ID NO:5613)的基于结构的比对。

[0068] 图7描绘了本公开的酶(MG2-1)与来自金黄色葡萄球菌的Cas9(SEQ ID NO:5613)的基于结构的比对。

[0069] 图8描绘了本公开的酶(MG3-1)与来自内氏放线菌的Cas9(SEQ ID NO:5614)的基于结构的比对。

[0070] 图9A、图9B、图9C、图9D、图9E、图9F、图9G和图9H描述了MG1家族酶MG1-1至MG1-6(SEQ ID NO:5、6、9、1、2和3)的基于结构的比对。

[0071] 图10描述了MG1-4与含有不同长度靶向序列的相应sgRNA复合物对DNA的体外切割。

[0072] 图11描绘了使用MG1-4连同其相应的sgRNA的大肠杆菌基因组DNA的细胞裂解。显示的是用MG1-4以及靶或非靶间隔子(顶部)转化的细胞稀释系列；底部分图显示定量的数据，其中左侧条表示非靶sgRNA，右侧条表示靶sgRNA。

[0073] 图12描绘了通过用实施例11中描述的MG1-4或MG1-6构建体连同其相应的sgRNA一起转染HEK细胞而产生的细胞插入缺失标记(indel)形成，所述sgRNA包含靶向人类基因组中不同位置的各种不同靶向序列。

[0074] 图13描述了MG3-6与含有不同长度靶向序列的相应sgRNA复合物对DNA的体外切割。

[0075] 图14描绘了使用MG3-7连同其相应的sgRNA，大肠杆菌基因组DNA的细胞裂解。显示的是用MG3-7以及靶或非靶间隔子(顶部)转化的细胞稀释系列；底部图显示定量的数据，其中左侧条表示非靶sgRNA，右侧条表示靶sgRNA。

[0076] 图15描绘了通过用实施例13中描述的MG3-7构建体连同其相应的sgRNA一起转染HEK细胞而产生的细胞插入缺失标记形成，所述sgRNA包含靶向人类基因组中不同位置的各种不同靶向序列。

[0077] 图16描述了MG15-1与其相应的sgRNA的复合物对DNA的体外切割，所述sgRNA含有

不同长度的靶向序列。

[0078] 图17、图18、图19和图20描绘了琼脂糖凝胶，其显示了在含有各种MG家族核酸酶及其相应tracrRNA或sgRNA的TXTL提取物存在下PAM载体文库切割的结果。

[0079] 图21、图22、图23、图24、图25和图26描绘了本文描述的MG酶的相应sgRNA的预测结构(例如，如实施例7中预测的)。

[0080] 图27、图28、图29、图30、图31、图32和图33描绘了如本文所述(例如，如实施例6中所述)通过NGS衍生的PAM序列的seqLogo表示。

[0081] 图34描绘了使用MG2-7连同其相应的sgRNA，大肠杆菌基因组DNA的细胞裂解。显示的是用MG2-7以及靶或非靶间隔子(顶部)转化的细胞稀释系列；底部图显示定量的数据，其中右侧条形表示非靶sgRNA，左侧条形表示靶sgRNA。

[0082] 图35描绘了使用MG14-1连同其相应的sgRNA，大肠杆菌基因组DNA的细胞裂解。显示的是用MG14-1以及靶或非靶间隔子(顶部)转化的细胞稀释系列；底部图显示定量的数据，其中右侧条形表示非靶sgRNA，左侧条形表示靶sgRNA。

[0083] 图36描绘了使用MG15-1连同其相应的sgRNA，大肠杆菌基因组DNA的细胞裂解。显示的是用MG15-1以及靶或非靶间隔子(顶部)转化的细胞稀释系列；底部图显示定量的数据，其中右侧条形表示非靶sgRNA，左侧条形表示靶sgRNA。

[0084] 序列表的简要说明

[0085] 随此提交的序列表提供了用于本公开的方法、组合物和系统中的示例性多核苷酸和多肽序列。以下是其中的序列的示例性描述。

[0086] MG1

[0087] SEQ ID NO:1-319显示了MG1核酸酶的全长肽序列。

[0088] SEQ ID NO:1827-2140显示了上述MG1核酸酶的RuvC_III结构域的肽序列。

[0089] SEQ ID NO:3638-3955显示了上述MG1核酸酶的HNH结构域的肽。

[0090] SEQ ID NO:5476-5479显示源自与上述MG1核酸酶相同基因座(例如，分别与SEQ ID NO:1-4相同基因座)的MG1 tracrRNA的核苷酸序列。

[0091] SEQ ID NO:5461-5464显示了经工程改造以与MG1核酸酶(例如，分别为SEQ ID NO:1-4)一起起作用的sgRNA的核苷酸序列，其中N表示靶向序列的核苷酸。

[0092] SEQ ID NO:5572-5575显示了大肠杆菌密码子优化的MG1家族酶的编码序列的核苷酸序列(SEQ ID NO:1-4)。

[0093] SEQ ID NO:5588-5589显示了人密码子优化的MG1家族酶的编码序列的核苷酸序列(SEQ ID NO:1和SEQ ID NO:3)。

[0094] SEQ ID NO:5616-5632显示了MG1家族酶的肽基序特征。

[0095] MG2

[0096] SEQ ID NO:320-420显示了MG2核酸酶的全长肽序列。

[0097] SEQ ID NO:2141-2241显示了上述MG2核酸酶的RuvC_III结构域的肽序列。

[0098] SEQ ID NO:3955-4055显示了上述MG2核酸酶的HNH结构域的肽。

[0099] SEQ ID NO:5490-5494显示了源自与上述MG2核酸酶相同的基因座(例如，分别与SEQ ID NO:320、321、323、325和326相同的基因座)的MG2 tracrRNA的核苷酸序列。

[0100] SEQ ID NO:5465显示了经工程改造以与MG2核酸酶一起起作用的sgRNA的核苷酸

序列(例如,上述SEQ ID NO:321)。

[0101] SEQ ID NO:5572-5575显示了大肠杆菌密码子优化的MG2家族酶的编码序列的核苷酸序列。

[0102] SEQ ID NO:5631-5638显示了MG2家族酶的肽序列特征。

[0103] MG3

[0104] SEQ ID NO:421-431显示了MG3核酸酶的全长肽序列。

[0105] SEQ ID NO:2242-2251显示了上述MG3核酸酶的RuvC_III结构域的肽序列。

[0106] SEQ ID NO:4056-4066显示了上述MG3核酸酶的HNH结构域的肽。

[0107] SEQ ID NO:5495-5502显示了源自与上述MG3核酸酶相同的基因座(例如,分别与SEQ ID NO:421-428相同的基因座)的MG3 tracrRNA的核苷酸序列。

[0108] SEQ ID NO:5466-5467显示了经工程改造以与MG3核酸酶一起起作用的sgRNA的核苷酸序列(例如,SEQ ID NO:421-423)。

[0109] SEQ ID NO:5578-5580显示了大肠杆菌密码子优化的MG3家族酶的编码序列的核苷酸序列。

[0110] SEQ ID NO:5639-5648显示了MG3家族酶的肽序列特征。

[0111] MG4

[0112] SEQ ID NO:432-660显示了MG4核酸酶的全长肽序列。

[0113] SEQ ID NO:2253-2481显示了上述MG4核酸酶的RuvC_III结构域的肽序列。

[0114] SEQ ID NO:4067-4295显示了上述MG4核酸酶的HNH结构域的肽。

[0115] SEQ ID NO:5503显示了源自与上述MG4核酸酶相同的基因座的MG4tracrRNA的核苷酸序列。

[0116] SEQ ID NO:5468显示了经工程改造以与MG4核酸酶一起起作用的sgRNA的核苷酸序列。

[0117] SEQ ID NO:5649显示了MG4家族酶的肽序列特征。

[0118] MG6

[0119] SEQ ID NO:661-668显示了MG6核酸酶的全长肽序列。

[0120] SEQ ID NO:2482-2489显示了上述MG6核酸酶的RuvC_III结构域的肽序列。

[0121] SEQ ID NO:4296-4303显示了上述MG3核酸酶的HNH结构域的肽。

[0122] MG7

[0123] SEQ ID NO:669-677显示了MG7核酸酶的全长肽序列。

[0124] SEQ ID NO:2490-2498显示了上述MG7核酸酶的RuvC_III结构域的肽序列。

[0125] SEQ ID NO:4304-4312显示了上述MG3核酸酶的HNH结构域的肽。

[0126] SEQ ID NO:5504显示了源自与上述MG7核酸酶相同的基因座的MG7 tracrRNA的核苷酸序列。

[0127] MG14

[0128] SEQ ID NO:678-929显示了MG14核酸酶的全长肽序列。

[0129] SEQ ID NO:2499-2750显示了上述MG14核酸酶的RuvC_III结构域的肽序列。

[0130] SEQ ID NO:4313-4564显示了上述MG14核酸酶的HNH结构域的肽。

[0131] SEQ ID NO:5505显示了源自与上述MG14核酸酶相同的基因座的MG14 tracrRNA的

核苷酸序列。

[0132] SEQ ID NO:5581显示了大肠杆菌密码子优化的MG14家族酶的编码序列的核苷酸序列。

[0133] SEQ ID NO:5650-5667显示了MG14家族酶的肽序列特征。

[0134] MG15

[0135] SEQ ID NO:930-1092显示了MG15核酸酶的全长肽序列。

[0136] SEQ ID NO:2751-2913显示了上述MG15核酸酶的RuvC_III结构域的肽序列。

[0137] SEQ ID NO:4565-4727显示了上述MG15核酸酶的HNH结构域的肽。

[0138] SEQ ID NO:5506显示了源自与上述MG15核酸酶相同的基因座的MG15 tracrRNA的核苷酸序列。

[0139] SEQ ID NO:5470显示了经工程改造以与MG15核酸酶一起起作用的sgRNA的核苷酸序列。

[0140] SEQ ID NO:5582显示了大肠杆菌密码子优化的MG15家族酶的编码序列的核苷酸序列。

[0141] SEQ ID NO:5668-5675显示了MG15家族酶的肽序列特征。

[0142] MG16

[0143] SEQ ID NO:1093-1353显示了MG16核酸酶的全长肽序列。

[0144] SEQ ID NO:2914-3174显示了上述MG16核酸酶的RuvC_III结构域的肽序列。

[0145] SEQ ID NO:4728-4988显示了上述MG16核酸酶的HNH结构域的肽。

[0146] SEQ ID NO:5507显示了源自与上述MG3核酸酶相同的基因座的MG16 tracrRNA的核苷酸序列。

[0147] SEQ ID NO:5471显示了经工程改造以与MG16核酸酶一起起作用的sgRNA的核苷酸序列。

[0148] SEQ ID NO:5583显示了大肠杆菌密码子优化的MG16家族酶的编码序列的核苷酸序列。

[0149] SEQ ID NO:5676-5678显示了MG16家族酶的肽序列特征。

[0150] MG18

[0151] SEQ ID NO:1354-1511显示了MG18核酸酶的全长肽序列。

[0152] SEQ ID NO:3175-3330显示了上述MG18核酸酶的RuvC_III结构域的肽序列。

[0153] SEQ ID NO:4989-5146显示了上述MG18核酸酶的HNH结构域的肽。

[0154] SEQ ID NO:5508显示了源自与上述MG18核酸酶相同的基因座的MG18 tracrRNA的核苷酸序列。

[0155] SEQ ID NO:5472显示了经工程改造以与MG18核酸酶一起起作用的sgRNA的核苷酸序列。

[0156] SEQ ID NO:5584显示了大肠杆菌密码子优化的MG18家族酶的编码序列的核苷酸序列。

[0157] SEQ ID NO:5679-5686显示了MG18家族酶的肽序列特征。

[0158] MG21

[0159] SEQ ID NO:1512-1655显示了MG21核酸酶的全长肽序列。

[0160] SEQ ID NO:3331-3474显示了上述MG21核酸酶的RuvC_III结构域的肽序列。

[0161] SEQ ID NO:5147-5290显示了上述MG21核酸酶的HNH结构域的肽。

[0162] SEQ ID NO:5509显示了源自与上述MG21核酸酶相同的基因座的MG21 tracrRNA的核苷酸序列。

[0163] SEQ ID NO:5473显示了经工程改造以与MG21核酸酶一起起作用的sgRNA核苷酸序列。

[0164] SEQ ID NO:5585显示了大肠杆菌密码子优化的MG21家族酶的编码序列的核苷酸序列。

[0165] SEQ ID NO:5687-5692和5674-5675显示了MG21家族酶的肽序列特征。

[0166] MG22

[0167] SEQ ID NO:1656-1755显示了MG22核酸酶的全长肽序列。

[0168] SEQ ID NO:3475-3568显示了上述MG22核酸酶的RuvC_III结构域的肽序列。

[0169] SEQ ID NO:5291-5389显示了上述MG22核酸酶的HNH结构域的肽。

[0170] SEQ ID NO:5510显示了源自与上述MG22核酸酶相同的基因座的MG22 tracrRNA的核苷酸序列。

[0171] SEQ ID NO:5474显示了经工程改造以与MG22核酸酶一起起作用的sgRNA的核苷酸序列。

[0172] SEQ ID NO:5586显示了大肠杆菌密码子优化的MG22家族酶的编码序列的核苷酸序列。

[0173] SEQ ID NO:5694-5699显示了MG22家族酶的肽序列特征。

[0174] MG23

[0175] SEQ ID NO:1756-1826显示了MG23核酸酶的全长肽序列。

[0176] SEQ ID NO:3569-3637显示了上述MG23核酸酶的RuvC_III结构域的肽序列。

[0177] SEQ ID NO:5390-5460显示了上述MG23核酸酶的HNH结构域的肽。

[0178] SEQ ID NO:5511显示了源自与上述G23核酸酶相同的基因座的MG23 tracrRNA的核苷酸序列。

[0179] SEQ ID NO:5475显示了经工程改造以与MG23核酸酶一起起作用的sgRNA的核苷酸序列。

[0180] SEQ ID NO:5587显示了MG23家族酶的大肠杆菌密码子优化编码序列的核苷酸序列。

[0181] SEQ ID NO:5700-5717显示了MG23家族酶的肽序列特征。

具体实施方式

[0182] 虽然已经示出和在本文所述了本发明的各种实施方案,对本领域技术人员将是显而易见的是,仅通过举例的方式来提供这样的实施方案。在不脱离本发明的情况下,本领域技术人员可进行许多变型、改变和替代。应当理解,可以采用本文所述的本发明实施方案的各种替代方案。

[0183] 除非另有说明,否则本文公开的一些方法的实践使用免疫学、生物化学、化学、分子生物学、微生物学、细胞生物学、基因组学和重组DNA的技术。参见例如Sambrook和Green,

Molecular Cloning:A Laboratory Manual,第4版(2012);the series Current Protocols in Molecular Biology(F.M.Ausubel等人编辑);the series Methods In Enzymology(Academic Press,Inc.),PCR 2:A Practical Approach(M.J.MacPherson、B.D.Hames和G.R.Taylor编辑(1995)),Harlow和Lane编辑(1988)Antibodies,A Laboratory Manual,and Culture of Animal Cells:A Manual of Basic Technique and Specialized Applications,第6版(R.I.Freshney编辑(2010))(通过引用完全并入本文)。

[0184] 如本文所使用的,单数形式“一个/种(a/an)”和“所述(the)”也旨在包括复数形式,除非上下文另外明确指出。此外,如果在详细的说明书和/或权利要求中使用术语“包括”、“具有”、“带有”或其变体,则这些术语旨在以类似于术语“包含”的方式是包括性的。

[0185] 术语“约”或“大约”意指由本领域普通技术人员确定的特定值在可接受的误差范围之内,这将部分地取决于该值是怎样测定或确定的,即受到测量系统的限制。例如,根据本领域的实践,“约”可以意指在一个或超过一个标准差内。可替代地,“约”可表示给定值的多达20%、多达15%、多达10%、多达5%或多达1%的范围。

[0186] 如本文所用,“细胞”通常是指生物细胞。细胞可以是活生物体的基本结构、功能和/或生物学单位。细胞可以源自具有一个或多个细胞的任何生物体。一些非限制性示例包括:原核细胞、真核细胞、细菌细胞、古细菌细胞、单细胞真核生物的细胞、原生动物细胞、来自植物的细胞(例如来自植物作物、水果、蔬菜、谷物、大豆、玉米、玉蜀黍、小麦、种子、西红柿、稻、木薯、甘蔗、南瓜、干草、土豆、棉花、大麻、烟草、开花植物、松柏植物、裸子植物、蕨类、石松科、角苔纲、苔类植物、藓类的细胞)、藻类细胞(例如,布朗葡萄藻(Botryococcus braunii)、莱茵衣藻(Chlamydomonas reinhardtii)、Nannochloropsis gaditana、蛋白核小球藻(Chlorella pyrenoidosa)、展枝马尾藻(Sargassum patens)、微劳马尾藻(C.Agardh)等)、海藻(例如海带)、真菌细胞(例如酵母细胞、来自蘑菇的细胞)、动物细胞、来自无脊椎动物(例如果蝇、刺胞动物、棘皮动物、线虫等)的细胞、来自脊椎动物(例如鱼、两栖动物、爬行动物、鸟类、哺乳动物)的细胞,来自哺乳动物(例如猪、牛、山羊、绵羊、啮齿动物、大鼠、小鼠、非人灵长类动物、人等)的细胞,等等。有时细胞并非源自天然生物体(例如,细胞可以是合成的,有时称为人工细胞)。

[0187] 如本文所用,术语“核苷酸”通常是指碱基-糖-磷酸酯组合。核苷酸可包括合成的核苷酸。核苷酸可包括合成的核苷酸类似物。核苷酸可以是核酸序列的单体单元(例如,脱氧核糖核酸(DNA)和核糖核酸(RNA))。术语核苷酸可包括核糖核苷三磷酸、三磷酸腺苷(ATP)、尿苷三磷酸(UTP)、三磷酸胞嘧啶(CTP)、三磷酸鸟苷(GTP)和脱氧核糖核苷三磷酸,例如dATP、dCTP、dITP、dUTP、dGTP、dTTP或其衍生物。此类衍生物可包括例如[αS]dATP、7-脱氮-dGTP和7-脱氮-dATP,以及赋予含有它们的核酸分子核酸酶抗性的核苷酸衍生物。如本文所用,术语核苷酸可指双脱氧核糖核苷三磷酸(ddNTP)及其衍生物。双脱氧核糖核苷三磷酸的说明性示例可包括但不限于ddATP、ddCTP、ddGTP、ddITP和ddTTP。核苷酸可以是未标记的或可检测地标记的,例如使用包含光学可检测部分(例如,荧光团)的部分。也可以用量子点进行标记。可检测标记可以包括例如放射性同位素、荧光标记、化学发光标记、生物发光标记和酶标记。核苷酸的荧光标记可包括但不限于荧光素、5-羧基荧光素(FAM)、2’7’-二甲氧基-4’5-二氯-6-羧基荧光素(JOE)、罗丹明、6-羧基罗丹明(R6G)、N,N,N,’N’-四甲基-6-羧基罗丹明(TAMRA)、6-羧基-X-罗丹明(ROX)、4-(4’二甲基氨基苯基偶氮)苯甲酸

(DABCYL)、级联蓝、俄勒冈绿、德克萨斯红、花青和5-(2'-氨基乙基)氨基萘-1-磺酸(EDANS)。经荧光标记的核苷酸的具体示例可以包括[R6G]dUTP、[TAMRA]dUTP、

[0188] [R110]dCTP、[R6G]dCTP、[TAMRA]dCTP、[JOE]ddATP、

[0189] [R6G]ddATP、[FAM]ddCTP、[R110]ddCTP、[TAMRA]ddGTP、

[0190] [ROX]ddTTP、[dR6G]ddATP、[dR110]ddCTP、[dTAMRA]ddGTP和[dROX]ddTTP(可从Perkin Elmer,Foster City,Calif获得);FluoroLink脱氧核苷酸、FluoroLink Cy3-dCTP、FluoroLink Cy5-dCTP、FluoroLink FluorX-dCTP、FluoroLink Cy3-dUTP和FluoroLink Cy5-dUTP(可从Amersham,Arlington Heights,Ill.获得);荧光素-15-dATP、荧光素-12-dUTP、四甲基-罗丹明-6-dUTP、IR770-9-dATP、荧光素-12-ddUTP、荧光素-12-UTP和荧光素-15-2'-dATP(可从Boehringer Mannheim,Indianapolis,Ind.获得);和染色体标记的核苷酸、BODIPY-FL-14-UTP、BODIPY-FL-4-UTP、BODIPY-TMR-14-UTP、BODIPY-TMR-14-dUTP、BODIPY-TR-14-UTP、BODIPY-TR-14-dUTP、级联蓝-7-UTP、级联蓝-7-dUTP、荧光素-12-UTP、荧光素-12-dUTP、俄勒冈绿488-5-dUTP、罗丹明绿-5-UTP、罗丹明绿-5-dUTP、四甲基罗丹明-6-UTP、四甲基罗丹明-6-dUTP、德克萨斯红-5-UTP、德克萨斯红-5-dUTP和德克萨斯红-12-dUTP(可从Molecular Probes,Eugene,Oreg.获得)。核苷酸也可以通过化学修饰进行标记或标志。化学修饰的单核苷酸可以是生物素-dNTP。生物素化dNTP的一些非限制性示例可包括生物素-dATP(例如,bio-N6-ddATP、生物素-14-dATP)、生物素-dCTP(例如,生物素-11-dCTP、生物素-14-dCTP)和生物素-dUTP(例如,生物素-11-dUTP、生物素-16-dUTP、生物素-20-dUTP)。

[0191] 术语"多核苷酸"、"寡核苷酸"和"核酸"可互换使用,一般指任何长度的核苷酸的聚合形式,脱氧核糖核苷酸或核糖核苷酸或其类似物,无论是单链、双链或多链的形式。多核苷酸对于细胞可以是外源的或内源的。多核苷酸可以存在于无细胞环境中。多核苷酸可以是基因或其片段。多核苷酸可以是DNA。多核苷酸可以是RNA。多核苷酸可以具有任何三维结构并且可以执行任何功能。多核苷酸可包含一种或多种类似物(例如,改变的骨架、糖或核碱基)。如果存在,可以在聚合物组装之前或之后赋予对核苷酸结构的修饰。类似物的一些非限制性示例包括:5-溴尿嘧啶、肽核酸、xeno核酸、吗啉化合物、锁核酸、二醇核酸、苏糖核酸、双脱氧核苷酸、蛹虫草菌素、7-脱氮-GTP、荧光团(例如,连接至糖的罗丹明或荧光素)、含硫醇的核苷酸、生物素连接的核苷酸、荧光碱基类似物、CpG岛、甲基-7-鸟苷、甲基化的核苷酸、肌苷、硫尿苷、假尿苷、二氢尿苷、辫苷以及怀俄苷。多核苷酸的非限制性示例包括基因或基因片段的编码或非编码区、由连锁分析定义的基因座、外显子、内含子、信使RNA(mRNA)、转移RNA(tRNA)、核糖体RNA(rRNA)、短干扰RNA(siRNA)、短发夹RNA(shRNA)、微RNA(miRNA)、核酶、cDNA、重组多核苷酸、分支多核苷酸、质粒、载体、任何序列的分离DNA、任何序列的分离RNA、无细胞多核苷酸(包括无细胞DNA(cfDNA)和无细胞RNA(cfRNA))、核酸探针和引物。核苷酸的序列可以被非核苷酸组分中断。

[0192] 术语"转染"或"转染的"通常是指通过非病毒方法或基于病毒的方法将核酸引入细胞。核酸分子可以是编码完整蛋白质或其功能部分的基因序列。参见,例如,Sambrook等人,1989,Molecular Cloning:A Laboratory Manual,18.1-18.88.。

[0193] 术语"肽"、"多肽"和"蛋白质"在本文中可互换使用以一般指由一个或多个肽键连接的至少两个氨基酸残基的聚合物。该术语并不意味着特定长度的聚合物,也不旨在暗示

或区分肽是使用重组技术、化学或酶促合成产生的，还是天然存在的。该术语适用于天然存在的氨基酸聚合物以及包含至少一个经修饰氨基酸的氨基酸聚合物。在一些情况下，聚合物可以被非氨基酸中断。该术语包括任何长度的氨基酸链，包括全长蛋白质和具有或不具有二级和/或三级结构(例如结构域)的蛋白质。该术语还包括已被修饰的氨基酸聚合物，例如，通过二硫键形成、糖基化、脂化、乙酰化、磷酸化、氧化和任何其他操作(例如与标记组分缀合)。如本文所用，术语“氨基酸”通常指天然和非天然氨基酸，包括但不限于经修饰的氨基酸和氨基酸类似物。经修饰的氨基酸可以包括天然氨基酸和非天然氨基酸，它们已经被化学修饰以包括氨基酸上不天然存在的基团或化学部分。氨基酸类似物可以指氨基酸衍生物。术语“氨基酸”包括D-氨基酸和L-氨基酸。

[0194] 如本文所用，“非天然”通常可指在天然核酸或蛋白质中未发现的核酸或多肽序列。非天然可以指亲和标签。非天然可以指融合。非天然可以指包含突变、插入和/或缺失的天然存在的核酸或多肽序列。非天然序列可表现出和/或编码活性(例如，酶活性、甲基转移酶活性、乙酰转移酶活性、激酶活性、泛素化活性等)，所述活性也可由与非天然序列融合的核酸和/或多肽序列表现。非天然核酸或多肽序列可以通过基因工程与天然存在的核酸或多肽序列(或其变体)连接以产生嵌合核酸和/或编码嵌合核酸和/或多肽的多肽序列。

[0195] 如本文所用，术语“启动子”通常是指控制基因转录或表达的调控DNA区域，并且其可以位于RNA转录起始处的核苷酸或核苷酸区域附近或与之重叠。启动子可包含结合蛋白质因子的特定DNA序列，通常称为转录因子，其促进RNA聚合酶与导致基因转录的DNA的结合。“基础启动子”，也称为“核心启动子”，通常可以指包含促进可操作连接的多核苷酸转录表达的所有基础必需元件的启动子。真核基础启动子通常(但不一定)包含TATA盒和/或CAAT盒。

[0196] 如本文所用，术语“表达”通常指核酸序列或多核苷酸从DNA模板转录(例如转录为mRNA或其他RNA转录物)的过程和/或转录的mRNA随后被翻译成肽、多肽或蛋白质的过程。转录物和编码的多肽可统称为“基因产物”。如果多核苷酸源自基因组DNA，则表达可包括mRNA在真核细胞中的剪接。

[0197] 如本文所用，“可操作地连接”、“可操作的连接”、“可操作连接的”或其语法等价物通常是指遗传元件例如启动子、增强子、聚腺苷酸化序列等的并列，其中所述元件是处于一种允许他们以预期方式运作的关系中。例如，如果调节元件有助于启动编码序列的转录，则包含启动子和/或增强子序列的调节元件与编码区可操作地连接。调节元件和编码区之间可以存在中间残基，只要维持这种功能关系。

[0198] 如本文所用，“载体”通常是指包含多核苷酸或与多核苷酸缔合且可用于介导多核苷酸向细胞递送的大分子或大分子缔合。载体的示例包括质粒、病毒载体、脂质体和其他基因递送载体。载体通常包含遗传元件，例如调节元件，其与基因可操作地连接以促进基因在靶标中的表达。

[0199] 如本文所用，“表达盒”和“核酸盒”可互换使用，通常是指一起表达或可操作地连接用于表达的核酸序列或元件的组合。在一些情况下，表达盒是指调节元件和它们可操作地连接用于表达的一个或多个基因的组合。

[0200] DNA或蛋白质序列的“功能片段”通常是指保留与全长DNA或蛋白质序列的生物学活性基本相似的生物学活性(功能性或结构性)的片段。DNA序列的生物学活性可以是它以

已知归因于全长序列的方式影响表达的能力。

[0201] 如本文所用，“工程改造的”客体通常表示该客体已被人为干预修改。根据非限制性实例：核酸可以通过将其序列改变为自然界中不存在的序列来进行修饰；核酸可以通过将其连接到在自然界中不与之结合的核酸来进行修饰，从而使连接产物具有原始核酸中不存在的功能；使用自然界中不存在的序列可以在体外合成工程改造的核酸；蛋白质可以通过将其氨基酸序列改变为自然界中不存在的序列来进行修饰；工程改造的蛋白质可以获得新的功能或特性。“工程改造的”系统包括至少一个工程改造的组分。

[0202] 如本文所用，“合成的”和“人工的”可互换使用以指与天然存在的人蛋白质具有低序列同一性(例如，小于50％的序列同一性、小于25％的序列同一性、小于10％的序列同一性，小于5％的序列同一性，小于1％的序列同一性)的蛋白质或其结构域。例如，VPR和VP64结构域是合成的反式激活结构域。

[0203] 如本文所用，术语“tracrRNA”或“tracr序列”通常可以指与野生型示例性tracrRNA序列(例如，来自化脓性链球菌金黄色葡萄球菌等的tracrRNA或SEQ ID NO:5476-5511)具有至少约5％、10％、20％、30％、40％、50％、60％、70％、80％、90％、95％或100％序列同一性和/或序列相似性的核酸。tracrRNA可指与野生型示例性tracrRNA序列(例如，来自化脓性链球菌金黄色葡萄球菌等的tracrRNA)具有至多约5％、10％、20％、30％、40％、50％、60％、70％、80％、90％或100％序列同一性和/或序列相似性的核酸。tracrRNA可指tracrRNA的修饰形式，其可包含核苷酸变化，例如缺失、插入或置换、变体、突变或嵌合体。tracrRNA可以指与野生型示例性tracrRNA(例如，来自化脓性链球菌金黄色葡萄球菌等的tracrRNA)序列在一段至少6个连续核苷酸上具有至少约60％同一性的核酸。例如，tracrRNA序列可以与野生型示例性tracrRNA(例如，来自化脓性链球菌金黄色葡萄球菌等的tracrRNA)序列在一段至少6个连续核苷酸上至少约60％相同、至少约65％相同、至少约70％相同、至少约75％相同、至少约80％相同、至少约85％相同、至少约90％相同、至少约95％相同、至少约98％相同、至少约99％相同或100％相同。II型tracrRNA序列可以通过识别与相邻CRISPR阵列中部分重复序列互补的区域来预测基因组序列。

[0204] 如本文所用，“指导核酸”通常可指可与另一核酸杂交的核酸。指导核酸可以是RNA。指导核酸可以是DNA。指导核酸可以被编程以位点特异性地结合核酸序列。待靶向的核酸或靶核酸可包含核苷酸。指导核酸可包含核苷酸。靶核酸的一部分可以与指导核酸的一部分互补。双链靶多核苷酸的与指导核酸互补并杂交的链可称为互补链。双链靶多核苷酸的与互补链互补且因此可能不与指导核酸互补的链可称为非互补链。指导核酸可以包含多核苷酸链并且可以称为“单指导核酸”。指导核酸可以包含两条多核苷酸链并且可以称为“双指导核酸”。如果没有另外说明，术语“指导核酸”可以是包含性的，指单指导核酸和双指导核酸两者。指导核酸可包含可称为“核酸靶向区段”或“核酸靶向序列”的区段。核酸靶向区段可包含可称为“蛋白质结合区段”或“蛋白质结合序列”或“Cas蛋白质结合区段”的子区段。

[0205] 在两个或更多个核酸或多肽序列的上下文中，术语“序列同一性”或“同一性百分比”通常是指当在局部或全局比较窗口上进行比较和比对以获得最大对应时，如使用序列比较算法所测量，相同或具有特定百分比的相同氨基酸残基或核苷酸两个(例如，在成对比对中)或更多个(例如，在多序列比对中)序列。多肽序列的合适序列比较算法包括，例如

BLASTP(使用3的字长(W)、10的期望(E)的参数和BLOSUM62评分矩阵(设置缺口罚分(存在时)为11,延伸为1)并针对长度超过30个残基的多肽序列使用条件型组成得分矩阵调整);BLASTP,其使用2的字长(W)、1000000的期望(E)的参数和PAM30评分矩阵(对于少于30个残基的序列,设置缺口罚分针对开放缺口为9,针对延伸缺口为1)(这些是可从https://blast.ncbi.nlm.nih.gov获得的BLAST套件中BLASTP的默认参数);CLUSTALW,具有的参数为;Smith-Waterman同源搜索算法,具有以下参数:匹配为2,不匹配为-1,缺口为-1;带有默认参数的MUSCLE;MAFFT,具有以下参数:重树(retree)为2,最大迭代次数为1000;具有默认参数的Novafold;具有默认参数的HMMER hmmalign。

[0206] 如本文所用,术语"RuvC_III结构域"通常是指RuvC核酸内切酶结构域的第三个不连续区段(RuvC核酸酶结构域由三个不连续区段RuvC_I、RuvC_II和RuvC_III组成)。RuvC结构域或其区段通常可以通过与已知结构域序列的比对、与具有注释的结构域的蛋白质的结构比对或通过与基于已知结构域序列构建的隐马尔可夫模型(HMM)(例如,用于RuvC_III Pfam HMM PF18541)进行比较来鉴定。

[0207] 如本文所用,术语"HNH结构域"通常是指具有特征性组氨酸和天冬酰胺残基的所述核酸内切酶结构域。HNH结构域通常可以通过与已知结构域序列的比对、与具有注释结构域的蛋白质的结构比对或通过与基于已知域序列构建的隐马尔可夫模型(HMM)(例如,结构域HNH的Pfam HMM PF01844)进行比较来鉴定。

[0208] 概述

[0209] 具有独特功能和结构的新型Cas酶的发现可能提供进一步破坏脱氧核糖核酸(DNA)编辑技术、提高速度、特异性、功能性和易用性的潜力。相对于微生物中成簇规则间隔短回文重复(CRISPR)系统的预测普遍性和微生物物种的绝对多样性,文献中存在相对较少的功能经表征的CRISPR/Cas酶。这部分是因为在实验室条件下可能不容易培养大量微生物物种。来自代表大量微生物物种的自然环境生态位的宏基因组测序有可能大幅增加已知的新CRISPR/Cas系统的数量,并加速新寡核苷酸编辑功能的发现。2016年从天然微生物群落的宏基因组分析中发现了CasX/CasY CRISPR系统,证明了这种方法的成果最近示例。

[0210] CRISPR/Cas系统是RNA指导的核酸酶复合物,已被描述为用作微生物中的适应性免疫系统。在其自然环境中,CRISPR/Cas系统出现在CRISPR(成簇规则间隔短回文重复)操纵子或基因座中,通常包括两个部分:(i)由同样短的间隔序列隔开的一系列短重复序列(30-40bp),其编码基于RNA的靶向元件;和(ii)编码Cas的ORF,其编码由基于RNA的靶向元件引导的核酸酶多肽以及辅助蛋白/酶。特定靶核酸序列的有效核酸酶靶向通常需要(i)靶(靶种子)的前6-8个核酸与crRNA指导物之间的互补杂交;以及(ii)在靶种子限定的附近内存在前间区序列邻近基序(PAM)序列(PAM通常是宿主基因组中不常见的序列)。根据系统的确切功能和组织,CRISPR-Cas系统通常根据共享的功能特征和进化相似性分为2类、5种类型和16种亚型。

[0211] I类CRISPR-Cas系统具有大的多亚单位效应物复合物,包括I、III和IV型。

[0212] 就组分而言,I型CRISPR-Cas系统被认为具有中等复杂性。在I型CRISPR-Cas系统中,RNA靶向元件阵列被转录为长前体crRNA(pre-crRNA),其在重复元件处进行加工以释放短的、成熟的crRNA,当核酸靶后面是合适的称为前间区序列邻近基序(PAM)的短共有序列时crRNA可将核酸酶复合物引导至核酸靶。这种加工通过称为级联的大型核酸内切酶复合

物的内切核糖核酸酶亚基(Cas6)发生,该复合物还包含crRNA引导的核酸酶复合物的核酸酶(Cas3)蛋白质组分。CasI核酸酶主要作为DNA核酸酶起作用。

[0213] III型CRISPR系统的特征可能在于存在称为Cas10的中央核酸酶,以及包含Csm或Cmr蛋白亚基的重复相关神秘蛋白(RAMP)。与I型系统一样,成熟的crRNA是使用Cas6样酶从pre-crRNA加工而来。与I型和II型系统不同,III型系统似乎靶向并切割DNA-RNA双链体(例如用作RNA聚合酶模板的DNA链)。

[0214] IV型CRISPR-Cas系统具有效应物复合物,该复合物由以下组成:高度还原的大亚基核酸酶(csf1)、Cas5(csf3)和Cas7(csf2)组的RAMP蛋白的两个基因,在一些情况下,预测的小亚基的基因;此类系统常见于内源性质粒上。

[0215] II类CRISPR-Cas系统通常具有单多肽多结构域核酸酶效应物,并且包括II、V和VI型。

[0216] 就组分而言,II型CRISPR-Cas系统被认为是最简单的。在II型CRISPR-Cas系统中,将CRISPR阵列加工成成熟的crRNA不需要存在特殊的核酸内切酶亚基,而是需要具有与阵列重复序列互补的区域的小型反式编码crRNA(tracrRNA);tracrRNA与其相应的效应物核酸酶(例如,Cas9)和重复序列相互作用以形成前体dsRNA结构,该结构被内源性RNAse III切割以产生装载有tracrRNA和crRNA的成熟效应物酶。Cas II核酸酶被称为DNA核酸酶。2型效应物通常表现出由以下组成的结构:采用RNaseH折叠的RuvC样核酸内切酶结构域和插入RuvC样核酸酶结构域的折叠内的无关HNH核酸酶结构域。RuvC样结构域负责切割靶(例如,crRNA互补)DNA链,而HNH结构域负责切割置换的DNA链。

[0217] V型CRISPR-Cas系统的特征是核酸酶效应物(例如,Cas12)结构类似于II/效应物(包含RuvC样结构域)的结构。与II型相似,大多数(但不是全部)V型CRISPR系统使用tracrRNA将pre-crRNA加工成成熟的crRNA,然而,与需要RNAse III将pre-crRNA切割成多个crRNA的II型系统不同,V型系统能够使用效应物核酸酶本身来切割pre-crRNA。与II型CRISPR-Cas系统一样,V型CRISPR-Cas系统也被称为DNA核酸酶。与II型CRISPR-Cas系统不同,一些V型酶(例如,Cas12a)似乎具有稳健的单链非特异性脱氧核糖核酸酶活性,该活性由双链靶序列的第一crRNA定向切割激活。

[0218] VI型CRIPSR-Cas系统具有RNA指导的RNA核酸内切酶。VI型系统(例如,Cas13)的单个多肽效应物包含两个HEPN核糖核酸酶结构域,而不是RuvC样结构域。与II型和V型系统不同,VI型系统似乎也不需要tracrRNA来将pre-crRNA加工成crRNA。然而,与V型系统类似,一些VI型系统(例如,C2C2)似乎具有稳健的单链非特异性核酸酶(核糖核酸酶)活性,该活性由靶RNA的第一crRNA定向切割激活。

[0219] 由于其更简单的架构,II类CRISPR-Cas已被最广泛地用于工程改造和开发为设计者核酸酶/基因组编辑应用。

[0220] 可以在Jinek等人(Science.2012年8月17日;337

[0221] (6096):816-21,通过引用将其全部并入本文)中找到这种系统在体外使用的早期改适之一。Jinek研究首先描述了以下系统,该系统涉及(i)从化脓性链球菌SF370中分离出的重组表达的、纯化的全长Cas9(例如II类、II型Cas酶),(ii)纯化的成熟的约42nt crRNA,其带有与需要切割的靶DNA序列互补的约20nt 5'序列,然后是3'tracr结合序列(整个crRNA是从携带T7启动子序列的合成DNA模板体外转录);(iii)从携带T7启动子序列的合成

DNA模板体外转录的纯化的tracrRNA，和(iv)Mg2+。Jinek后来描述了一种改进的工程改造的系统，其中(ii)的crRNA通过接头(例如GAAA)连接到(iii)的5’末端，以形成能够通过自身将Cas9引导至靶的单一融合的合成指导RNA(sgRNA)(比较图2的上部和下部分图)。

[0222] Mali等人(Science.2013年2月15日;339(6121):823-

[0223] 826.)，其通过引用完全并入本文，后来通过提供编码以下的DNA载体将该系统适用于哺乳动物中：(i)ORF，其在合适的哺乳动物启动子下编码密码子优化的Cas9(例如，II类，II型Cas酶)，具有C末端核定位序列(例如，SV40 NLS)和合适的聚腺苷酸化信号(例如，TK pA信号)；和(ii)ORF，其在合适的聚合酶III启动子(例如U6启动子)下编码sgRNA(具有以G开头的5’序列，随后是连接至3’tracr结合序列的20nt互补靶核酸序列、接头和tracrRNA序列)。

[0224] MG1酶

[0225] 一方面，本公开提供包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下，核酸内切酶是Cas核酸内切酶。在一些情况下，核酸内切酶是II型、II类Cas核酸内切酶。核酸内切酶可包含RuvC_III结构域，其中所述RuvC_III结构域与SEQ ID NO:1827-2140中任一个具有至少约70％序列同一性。在一些情况下，核酸内切酶可包含RuvC_III结构域，其中该RuvC_III结构域与SEQ ID NO:1827-2140中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％的同一性。在一些情况下，核酸内切酶可包含RuvC_III结构域，其中与SEQ ID NO:1827-2140中任一个基本相同。核酸内切酶可包含与SEQ ID NO:1827-1831中任一个具有至少约70％序列同一性的RuvC_III结构域。在一些情况下，核酸内切酶可包含RuvC_III结构域，该RuvC_III结构域与SEQ ID NO:1827-1831中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％的同一性。在一些情况下，核酸内切酶可以包含与SEQ ID NO:1827-1831中任一个基本相同的RuvC_III结构域。在一些情况下，核酸内切酶可包含RuvC_III结构域，该RuvC_III结构域与SEQ ID NO:1827具有至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％的同一性。在一些情况下，核酸内切酶可包含与SEQ ID NO:1828具有至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％的同一性的RuvC_III结构域。在一些情况下，核酸内切酶可包含与SEQ ID NO:1829具有至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％的同一性的RuvC_III结构域。在一些情况下，核酸内切酶可包含与SEQ ID NO:1830具有至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至

少约99％的同一性的RuvC_III结构域。在一些情况下，核酸内切酶可以包含与SEQ ID NO:1831具有至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％的同一性的RuvC_III结构域。

[0226] 所述核酸内切酶可包含与SEQ ID NO:3638-3955中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:3638-3955中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:3638-3955中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:3638-3955中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:3638-3955中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:3638-3955中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:3638-3641中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:3638-3641中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:3638-3641中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:3638中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:3638中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:3638中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:3639中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:3639中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:3639中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:3640中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:3640中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:3640中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:3641中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含SEQ ID NO:3641中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:3641的任一个基本相同的HNH结构域。

[0227] 在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1-6或9-319中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可与SEQ ID NO:1-6或9-319中任一个基本相同。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1-4中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1-4中任一个基本相同。在一些情况下,所述核酸内切酶可以包含与SEQ ID NO:5615、5616或5617中任一个基本相同的肽基序。

[0228] 在一些情况下,核酸内切酶可包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:1-6或SEQ ID NO:9-319中任一个或的N末端或C末端,或附加到与SEQ ID NO:1-319中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可包含与SEQ ID NO:5593-5608中任一个具有至少约80%、至少约85%、至少约90%、至少约95%、至少约99%同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本上相同的序列。NLS可以包含下表1中的任何序列,或其组合:

[0229] 表1:可以与本公开的Cas效应物一起使用的示例NLS序列

[0230]

来源	NLS 氨基酸序列	SEQ ID NO:
SV40	PKKKRKV	5593
核质蛋白二分体 NLS	KRPAATKKAGQAKKKK	5594
c-myc NLS	PAAKRVKLD	5595
c-myc NLS	RQRRNELKRSP	5596
hRNPA1 M9 NLS	NQSSNFGPMKGGNFGGRSSGPYGGGGQYFAKPRNQGGY	5597

[0231]

输入蛋白-α IBB 结构域	RMRIZFKNKGKDTAELRRRRVEVS VELRKAKKDEQILKRRNV	5598
肌瘤 T 蛋白	VSRKRPRP	5599
肌瘤 T 蛋白	PPKKARED	5600
p53	PQPKKKPL	5601
小鼠 c-abl IV	SALIKKKKKMAP	5602
流感病毒 NS1	DRLRR	5603
流感病毒 NS1	PKQKKRK	5604
肝炎病毒δ抗原	RKLKKKIKKL	5605
小鼠 Mx1 蛋白	REKKKFLKRR	5606
人聚(ADP-核糖)聚合酶	KRKGDEVDGVDEVAKKKSKK	5607
类固醇激素受体(人)糖皮质激素	RKCLQAGMNLEARKTKK	5608

[0232] 在一些情况下,核酸内切酶可以是重组的(例如,通过合适的方法例如在大肠杆菌中表达然后表位标签纯化来克隆、表达和纯化)。在一些情况下,所述核酸内切酶可源自具有16S rRNA基因的细菌,所述16SrRNA基因与SEQ ID NO:5592-5595中任一个具有至少约90%同一性。所述核酸内切酶可以源自具有16S rRNA基因的物种,所述16S rRNA基因与SEQ ID NO:5592-5595中任一个具有至少约80%、至少约82%、至少约83%、至少约84%、至少约85%、至少约86%、至少约87%、至少约88%、至少约89%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约具有95%、至少约96%、至少约97%、至少约98%或至少约99%的同一性。所述核酸内切酶可以源自具有16SrRNA基因的物种,所述16S rRNA基因与SEQ ID NO:5592-5595中任一个基本相同。所述核酸内切酶可以来源于属于疣微菌门或暂定异域菌门的细菌。

[0233] 在一些情况下,序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数,并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11,延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0234] 在一些情况下,序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数,并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11,延伸为

1)并使用条件型组成得分矩阵调整)来确定。

[0235] 在一些情况下,tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA序列的至少约80%至至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5476-5489中任一个的至少约60-100(例如,至少约60、至具有少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80%的序列同一性。在一些情况下,tracrRNA可SEQ ID NO:5476-5489中任一个的至少约60-90(例如至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的同一性。在一些情况下,tracrRNA可以与SEQ ID NO:5476-5489中任一个的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸基本相同。tracrRNA可包含SEQ ID NO:5476-5489中任一个。

[0236] 在一些情况下,能够与所述核酸内切酶形成复合物的至少一种工程改造的合成的指导核糖核酸(sgRNA)可以包含与SEQ ID NO:5461-5464中任一个具有至少约80%同一性的序列。sgRNA可包含与SEQ ID NO:5461-5464中任一个具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的序列。sgRNA可包含与SEQ ID NO:5461-5464中任一个基本相同的序列。

[0237] 在一些情况下,上述系统可能包含两种不同的sgRNA,它们靶向第一区域和第二区域以在靶DNA基因座中切割,其中第二区域位于第一区域的3'。在一些情况下,上述系统可包含单链或双链DNA修复模板,其从5'到3'包含:第一同源臂(其包含第一区域5'的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列),至少约10个核苷酸的合成DNA序列,和第二同源臂(其包含第二区域3'的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0238] 另一方面,本公开提供了一种修饰靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物,并且在复合物与靶核酸基因座结合后,可修饰靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下,靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0239] 在靶核酸基因座可以在细胞内的情况下,酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供,所述RuvC_III结构域与SEQ ID NO:1827-2140中任一个具有至少约75%(例如,至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至

少约95％、至少约96％、至少约97％、至少约98％、至少约99％)同一性。包含编码所述核酸内切酶的开放阅读框的脱氧核糖核酸(DNA)可以包含与SEQ ID NO:5572-5575中任一个基本相同的序列或与SEQ ID NO:5572-5575中任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％的同一性的变体。在一些情况下,核酸包含启动子,编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供,所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下,生物体可以是真核生物。在一些情况下,生物体可以是真菌。在一些情况下,生物体可以是人。

[0240] 在一些情况下,本公开可提供包含本文公开的系统或本文所述的核酸的表达盒。在一些情况下,表达盒或核酸可以作为载体提供。在一些情况下,可以在细胞中提供表达盒、核酸或载体。在一些情况下,细胞是具有16S rRNA基因的细菌细胞,所述16S rRNA基因与SEQ ID NO:5592-5595中任一个具有至少约90％(例如,至少约99％)同一性。

[0241] MG2酶

[0242] 一方面,本公开提供包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下,核酸内切酶是Cas核酸内切酶。在一些情况下,核酸内切酶是II型、II类Cas核酸内切酶。核酸内切酶可包含RuvC_III结构域,其中所述RuvC_III结构域与SEQ ID NO:2141-2241中任一个具有至少约70％序列同一性。在一些情况下,核酸内切酶可包含RuvC_III结构域,其中所述RuvC_III结构域与SEQ ID NO:2141-2241中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％同一性。在一些情况下,核酸内切酶可以包含RuvC_III结构域,其中与SEQ ID NO:2141-2142中任一个基本相同。核酸内切酶可以包含与SEQ ID NO:2141-2142中任一个具有至少约70％序列同一性的RuvC_III结构域。在一些情况下,核酸内切酶可包含与SEQ ID NO:2141-2142中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％同一性的RuvC_III结构域。在一些情况下,核酸内切酶可以包含与SEQ ID NO:2141-2142中任一个基本相同的RuvC_III结构域。

[0243] 所述核酸内切酶可包含与SEQ ID NO:3955-4055中任一个具有至少约70％同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:3955-4055中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、或至少约99％

相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:3955-4055中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:3955-3956中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:3955-3956中任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:3955-3956中任一个基本相同的HNH结构域。

[0244] 在一些情况下,所述核酸内切酶可包含与SEQ ID NO:320-420中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可与SEQ ID NO:320-420中任一个基本相同。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:320-321中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:320-321中任一个基本相同。

[0245] 在一些情况下,所述核酸内切酶可包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:320-420中任一个或与SEQ ID NO:320-420中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可包含与SEQ ID NO:5593-5608中任一个具有至少约80%、至少约85%、至少约90%、至少约95%、至少约99%同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本上相同的序列。NLS可以包含表1中的任何序列或其组合:

[0246] 在一些情况下,序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数,并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11,延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0247] 在一些情况下,上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA),其能够与核酸内切酶形成复合物,带有与所期望的切割序列互补的5'靶向区域。在一些情况下,5'靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下,靶向区域的最5'核苷酸可以是G。在一些情况下,5'靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3'的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列,其前面是crRNA tracrRNA结合区3'的4-核苷酸接头。sgRNA可以包含,从5'到3':能够与细胞中的靶序列杂交的非天然指导核酸序列;和tracr序列。在一些情况下,非天然指导核酸序列和tracr序列共价连接。

[0248] 在一些情况下,tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA

序列的至少约80％至至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5490-5494中任一个的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85个或至少约90)个连续核苷酸具有至少约80％序列同一性。在一些情况下,tracrRNA可与SEQ ID NO:5490-5494中任一个的至少约60-90(例如至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性。在一些情况下,tracrRNA可以与SEQ ID NO:5490-5494中任一个的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸基本相同。tracrRNA可包含SEQ ID NO:5490-5494中任一个。

[0249] 在一些情况下,能够与所述核酸内切酶形成复合物的至少一种工程改造的合成的指导核糖核酸(sgRNA)可以包含与SEQ ID NO:5465具有至少约80％同一性的序列。sgRNA可包含与SEQ ID NO:5465具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的序列。sgRNA可包含与SEQ ID NO:5465基本相同的序列。

[0250] 在一些情况下,上述系统可包含两种不同的sgRNA,它们靶向第一区域和第二区域以在靶DNA基因座中切割,其中第二区域位于第一区域的3’。在一些情况下,上述系统可包含单链或双链DNA修复模板,其从5’到3’包含:第一同源臂(其包含第一区域5’的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列),至少约10个核苷酸的合成DNA序列,和第二同源臂(其包含第二区域3’的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0251] 另一方面,本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物,并且在复合物与目的靶核酸基因座结合后,可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下,靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0252] 在靶核酸基因座可以在细胞内的情况下,酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供,所述RuvC_III结构域与SEQ ID NO:2141-2241中任一个具有至少约75％(例如,至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％)同一性。包含编码所述核酸内切酶的开放阅读框的脱氧核糖核酸(DNA)可以包含与SEQ ID NO:5576-5577中任一个基本相同的序列或与SEQ ID NO:5576-5577中任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约

75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性变体。在一些情况下，核酸包含启动子，编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供，所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下，生物体可以是真核生物。在一些情况下，生物体可以是真菌。在一些情况下，生物体可以是人。

[0253] MG3酶

[0254] 在一个方面，本公开提供了一种包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下，所述核酸内切酶是Cas核酸内切酶。在一些情况下，所述核酸内切酶是II型、II类Cas核酸内切酶。所述核酸内切酶可包含RuvC_III结构域，其中所述RuvC_III结构域与SEQ ID NO:2242-2251中任一个具有至少约70％的序列同一性。在一些情况下，所述核酸内切酶可包含RuvC_III结构域，其中该RuvC_III结构域与SEQ ID NO:2242-2251中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％同一性。在一些情况下，所述核酸内切酶可包含RuvC_III结构域，其中与SEQ ID NO:2242-2251中任一个基本相同。所述核酸内切酶可包含与SEQ ID NO:2242-2244中任一个具有至少约70％序列同一性的RuvC_III结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:2242-2244中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、具有至少约99％同一性的RuvC_III结构域。在一些情况下，所述核酸内切酶可以包含与SEQ ID NO:2242-2244中任一个基本相同的RuvC_III结构域。

[0255] 所述核酸内切酶可包含与SEQ ID NO:4056-4066中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:4056-4066中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4056-4066中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:4056-4058中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:4056-4058中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4056-4058中任一个基本相同的HNH结构域。

[0256] 在一些情况下，所述核酸内切酶可包含与SEQ ID NO:421-431中任一个具有至少

约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体。在一些情况下，所述核酸内切酶可以与SEQ ID NO:421-431中任一个基本相同。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:421-423中任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体。在一些情况下，所述核酸内切酶可以与SEQ ID NO:421-423中任一个基本相同。

[0257] 在一些情况下，所述核酸内切酶可以包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:421-431中任一个的N末端或C末端，或被附加到与SEQ ID NO:421-431任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可以包含与SEQ ID NO:5593-5608中任一个具有至少约80％、至少约85％、至少约90％、至少约95％、至少约99％同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本相同的序列。NLS可以包含表1中的任何序列或其组合：

[0258] 在一些情况下，序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数，并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11，延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0259] 在一些情况下，上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA)，其能够与核酸内切酶形成复合物，带有与所期望的切割序列互补的5’靶向区域。在一些情况下，5’靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下，靶向区域的最5’核苷酸可以是G。在一些情况下，5’靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3’的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列，其前面是crRNA tracrRNA结合区3’的4-核苷酸接头。sgRNA可以包含，从5’到3’：能够与细胞中的靶序列杂交的非天然指导核酸序列；和tracr序列。在一些情况下，非天然指导核酸序列和tracr序列共价连接。

[0260] 在一些情况下，tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA序列的至少约80％至至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5495-5502中任一个的至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85个或至少约90)个连续核苷酸具有至少约80％序列同一性。在一些情况下，tracrRNA可以与SEQ ID NO:5495-5502中任一个的至少约60-90(例如至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、

至少约98%或至少约99%同一性。在一些情况下，tracrRNA可以与SEQ ID NO:5495-5502中任一个的至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸基本相同。tracrRNA可包含SEQ ID NO:5495-5502中任一个。

[0261] 在一些情况下，能够与核酸内切酶形成复合物的至少一种工程改造的合成指导核糖核酸(sgRNA)可以包含与SEQ ID NO:5466-5467中任一个具有至少约80%同一性的序列。sgRNA可包含与SEQ ID NO:5466-5467中任一个具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的序列。sgRNA可以包含与SEQ ID NO:5466-5467中任一个基本上相同的序列。

[0262] 在一些情况下，上述系统可包含两种不同的sgRNA，它们靶向第一区域和第二区域以在靶DNA基因座中切割，其中第二区域位于第一区域的3’。在一些情况下，上述系统可包含单链或双链DNA修复模板，其从5’到3’包含：第一同源臂(其包含第一区域5’的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)，至少约10个核苷酸的合成DNA序列，和第二同源臂(其包含第二区域3’的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0263] 另一方面，本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物，并且在复合物与目的靶核酸基因座结合后，可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下，靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0264] 在靶核酸基因座可以在细胞内的情况下，所述酶可以作为包含编码具有RuvC_III结构域的酶的开放阅读框的核酸提供，该RuvC_III结构域与SEQ ID NO:2242-2251中任一个具有至少约75%(例如，至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%)同一性。包含编码所述核酸内切酶的开放阅读框的脱氧核糖核酸(DNA)可以包含与SEQ ID NO:5578-5580中任一个基本相同的序列或与SEQ ID NO:5578-5580中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下，所述核酸包含启动子，编码核酸内切酶的开放阅读框可操作地连接至该启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供，所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol

III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下,生物体可以是真核生物。在一些情况下,生物体可以是真菌。在一些情况下,生物体可以是人。

[0265] MG4酶

[0266] 一方面,本公开提供包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下,核酸内切酶是Cas核酸内切酶。在一些情况下,核酸内切酶是II型、II类Cas核酸内切酶。所述核酸内切酶可包含RuvC_III结构域,其中所述RuvC_III结构域与SEQ ID NO:2253-2481中任一个具有至少约70%的序列同一性。在一些情况下,所述核酸内切酶可包含RuvC_III结构域,其中该RuvC_III结构域与EQ ID NO:2253-2481中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%的同一性。在一些情况下,所述核酸内切酶可包含RuvC_III结构域,其中与SEQ ID NO:2253-2481中任一个基本相同。所述核酸内切酶可包含与SEQ ID NO:2253-2481中任一个具有至少约70%序列同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:2253-2481中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可以包含与SEQ ID NO:2253-2481中任一个基本相同的RuvC_III结构域。

[0267] 所述核酸内切酶可以包含与SEQ ID NO:4067-4295中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:4067-4295中任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4067-4295中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:4067-4295中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:4067-4295中任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4067-4295中任一个基本相同的HNH结构域。

[0268] 在一些情况下,所述核酸内切酶可包含与SEQ ID NO:432-660中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:432-660中任一个基本相同。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:432-660中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至

少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:432-660中任一个基本相同。

[0269] 在一些情况下,所述核酸内切酶可以包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:432-660中任一个的N末端或C末端,或被附加到与SEQ ID NO:432-660中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可以包含与SEQ ID NO:5593-5608中任一个具有至少80%、至少85%、至少90%、至少95%、至少99%同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本相同的序列。NLS可包括表1中的任何序列或其组合:

[0270] 在一些情况下,序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数,并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11,延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0271] 在一些情况下,上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA),其能够与核酸内切酶形成复合物,带有与所期望的切割序列互补的5'靶向区域。在一些情况下,5'靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下,靶向区域的最5'核苷酸可以是G。在一些情况下,5'靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3'的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列,其前面是crRNA tracrRNA结合区3'的4-核苷酸接头。sgRNA可以包含,从5'到3':能够与细胞中的靶序列杂交的非天然指导核酸序列;和tracr序列。在一些情况下,非天然指导核酸序列和tracr序列共价连接。

[0272] 在一些情况下,tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA序列的至少约80%至至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5503的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80%的同一性。在一些情况下,tracrRNA可SEQ ID NO:5503的至少约60-90(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的同一性。在一些情况下,tracrRNA可SEQ ID NO:5503的至少约60-90(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸基本相同。tracrRNA可包含SEQ ID NO:5503。

[0273] 在一些情况下,能够与所述核酸内切酶形成复合物的至少一种工程改造的合成的指导核糖核酸(sgRNA)可包含与SEQ ID NO:5468具有至少约80%同一性的序列。sgRNA可包含与SEQ ID NO:5468具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、

至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的序列。sgRNA可包含与SEQ ID NO:5468基本相同的序列。

[0274] 在一些情况下，上述系统可包含两种不同的sgRNA，它们靶向第一区域和第二区域以在靶DNA基因座中切割，其中第二区域位于第一区域的3'。在一些情况下，上述系统可包含单链或双链DNA修复模板，其从5'到3'包含：第一同源臂(其包含第一区域5'的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)，至少约10个核苷酸的合成DNA序列，和第二同源臂(其包含第二区域3'的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0275] 另一方面，本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物，并且在复合物与目的靶核酸基因座结合后，可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下，靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0276] 在靶核酸基因座可以在细胞内的情况下，酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供，所述RuvC_III结构域与SEQ ID NO:2253-2481中任一个具有至少约75%(例如，至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%)同一性。在一些情况下，核酸包含启动子，编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供，所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下，生物体可以是真核生物。在一些情况下，生物体可以是真菌。在一些情况下，生物体可以是人。

[0277] MG6酶

[0278] 一方面，本公开提供包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下，核酸内切酶是Cas核酸内切酶。在一些情况下，核酸内切酶是II型、II类Cas核酸内切酶。核酸内切酶可包含RuvC_III结构域，其中所述RuvC_III结构域与SEQ ID NO:2482-2489中任一个具有至少约70%序列同一性。在一些情况下，核酸内切酶可包含RuvC_III结构域，其中RuvC_III结构域与SEQ ID NO:2482-2489中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约

99％同一性。在一些情况下，核酸内切酶可包含RuvC_III结构域，其中该RuvC_III结构域与SEQ ID NO:2482-2489中任一个基本上相同。

[0279] 核酸内切酶可包含与SEQ ID NO:4296-4303中任一个具有至少约70％同一性的HNH结构域。在一些情况下，核酸内切酶可包含与SEQ ID NO:4296-4303中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。核酸内切酶可包含与SEQ ID NO:4056-4066中任一个基本上相同的HNH结构域。

[0280] 在一些情况下，核酸内切酶可包含与SEQ ID NO:661-668中任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体。在一些情况下，核酸内切酶可以与SEQ ID NO:661-668中任一个基本上相同。

[0281] 在一些情况下，所述核酸内切酶可包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:661-668中任一个的N末端或C末端或与SEQ ID NO:661-668具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可包含与SEQ ID NO:5593-5608中任一个具有至少约80％、至少约85％、至少约90％、至少约95％、至少约99％同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本上相同的序列。NLS可以包含表1中的任何序列，或其组合：

[0282] 在一些情况下，序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数，并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11，延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0283] 在一些情况下，上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA)，其能够与核酸内切酶形成复合物，带有与所期望的切割序列互补的5’靶向区域。在一些情况下，5’靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下，靶向区域的最5’核苷酸可以是G。在一些情况下，5’靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3’的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列，其前面是crRNA tracrRNA结合区3’的4-核苷酸接头。sgRNA可以包含，从5’到3’：能够与细胞中的靶序列杂交的非天然指导核酸序列；和tracr序列。在一些情况下，非天然指导核酸序列和tracr序列共价连接。

[0284] 在一些情况下，tracr序列可以具有特定序列。tracr序列可具有天然tracrRNA序列的至少约80％至至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。

[0285] 在一些情况下，上述系统可包含两种不同的指导RNA，它们靶向第一区域和第二区域以在靶DNA基因座中切割，其中第二区域位于第一区域的3’。在一些情况下，上述系统可

包含单链或双链DNA修复模板，其从5’到3’包含：第一同源臂(其包含第一区域5’的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)，至少约10个核苷酸的合成DNA序列，和第二同源臂(其包含第二区域3’的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0286] 另一方面，本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物，并且在复合物与目的靶核酸基因座结合后，可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下，靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0287] 在靶核酸基因座可以在细胞内的情况下，酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供，所述RuvC_III结构域与SEQ ID NO:2482-2489中任一个具有至少约75％(例如，至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％)同一性。在一些情况下，核酸包含启动子，编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供，所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下，生物体可以是真核生物。在一些情况下，生物体可以是真菌。在一些情况下，生物体可以是人。

[0288] MG7酶

[0289] 一方面，本公开提供包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下，核酸内切酶是Cas核酸内切酶。在一些情况下，核酸内切酶是II型、II类Cas核酸内切酶。核酸内切酶可包含RuvC_III结构域，其中所述RuvC_III结构域与SEQ ID NO:2490-2498中任一个具有至少约70％序列同一性。在一些情况下，核酸内切酶可包含RuvC_III结构域，其中该RuvC_III结构域与SEQ ID NO:2490-2498中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％同一性。在一些情况下，核酸内切酶可包含RuvC_III结构域，其中该RuvC_III结构域与SEQ ID NO:2490-2498中任一个基本相同。核酸内切酶可包含与SEQ ID NO:2490-2498中任一个具有至少约70％序列同一性的RuvC_III结构域。在一些情况下，核酸内切酶可包含RuvC_III结构域，其中该RuvC_III结构域与SEQ ID NO:2490-2498中任一个具有至少约

20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％同一性。在一些情况下，所述核酸内切酶可以包含与SEQ ID NO:2490-2498中任一个基本相同的RuvC_III结构域。

[0290] 核酸内切酶可包含与SEQ ID NO:4304-4312中任一个具有至少约70％同一性的HNH结构域。在一些情况下，核酸内切酶可包含与SEQ ID NO:4304-4312中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。核酸内切酶可包含与SEQ ID NO:4304-4312中任一个基本上相同的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:4304-4312中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的变体。所述核酸内切酶可以包含与SEQ ID NO:4304-4312中任一个基本相同的HNH结构域。

[0291] 在一些情况下，所述核酸内切酶可以包含与SEQ ID NO:669-677中任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体。在一些情况下，所述核酸内切酶可以与SEQ ID NO:669-677中任一个基本相同。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:669-677中任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体。在一些情况下，所述核酸内切酶可以与SEQ ID NO:669-677中任一个基本相同。

[0292] 在一些情况下，所述核酸内切酶可包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:669-677中任一个的N末端或C末端，或被附加到与SEQ ID NO:669-677中任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可包含与SEQ ID NO:5593-5608中任一个具有至少约80％、至少约85％、至少约90％、至少约95％、至少约99％同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本上相同的序列。NLS可以包含表1中的任何序列，或其组合：

[0293] 在一些情况下，序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数，并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11，延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0294] 在一些情况下，上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA)，其能够与核酸内切酶形成复合物，带有与所期望的切割序列互补的5’靶向区域。

在一些情况下,5’靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下,靶向区域的最5’核苷酸可以是G。在一些情况下,5’靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3’的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列,其前面是crRNA tracrRNA结合区3’的4-核苷酸接头。sgRNA可以包含,从5’到3’:能够与细胞中的靶序列杂交的非天然指导核酸序列;和tracr序列。在一些情况下,非天然指导核酸序列和tracr序列共价连接。

[0295] 在一些情况下,tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA序列的至少约80%至至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5504的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85个或至少约90)个连续核苷酸具有至少约80%序列同一性。在一些情况下,tracrRNA可与SEQ ID NO:5504的至少约60-90(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性。在一些情况下,tracrRNA可以与SEQ ID NO:5504的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸基本相同。tracrRNA可以包含SEQ ID NO:5504。

[0296] 在一些情况下,上述系统可包含两种不同的sgRNA,它们靶向第一区域和第二区域以在靶DNA基因座中切割,其中第二区域位于第一区域的3’。在一些情况下,上述系统可包含单链或双链DNA修复模板,其从5’到3’包含:第一同源臂(其包含第一区域5’的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列),至少约10个核苷酸的合成DNA序列,和第二同源臂(其包含第二区域3’的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0297] 另一方面,本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物,并且在复合物与目的靶核酸基因座结合后,可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下,靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0298] 在靶核酸基因座可以在细胞内的情况下,酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供,所述RuvC_III结构域与SEQ ID NO:2490-2498中任一个具有至少约75%(例如,至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%)同一性。在一些情况下,核酸

包含启动子，编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供，所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下，生物体可以是真核生物。在一些情况下，生物体可以是真菌。在一些情况下，生物体可以是人。

[0299] MG14酶

[0300] 一方面，本公开提供包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下，核酸内切酶是Cas核酸内切酶。在一些情况下，核酸内切酶是II型、II类Cas核酸内切酶。核酸内切酶可包含RuvC_III结构域，其中所述RuvC_III结构域与SEQ ID NO:2499-2750中任一个具有至少约70％序列同一性。在一些情况下，核酸内切酶可包含RuvC_III结构域，其中该RuvC_III结构域与SEQ ID NO:2499-2750中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％同一性。在一些情况下，核酸内切酶可包含RuvC_III结构域，其中该RuvC_III结构域与SEQ ID NO:2499-2750中任一个基本上相同。核酸内切酶可包含与SEQ ID NO:2499-2750中任一个具有至少约70％的序列同一性的RuvC_III结构域。在一些情况下，核酸内切酶可包含与SEQ ID NO:2499-2750中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％同一性的RuvC_III结构域。在一些情况下，核酸内切酶可包含与SEQ ID NO:2499-2750中任一个基本相同的RuvC_III结构域。

[0301] 核酸内切酶可包含与SEQ ID NO:4313-4564中任一个具有至少约70％同一性的HNH结构域。在一些情况下，核酸内切酶可包含与SEQ ID NO:4313-4564中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。核酸内切酶可包含与SEQ ID NO:4313-4564中任一个基本上相同的HNH结构域。核酸内切酶可包含与SEQ ID NO:4313-4564中任一个具有至少约70％同一性的HNH结构域。在一些情况下，核酸内切酶可包含与SEQ ID NO:4067-4295中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。核酸内切酶可包含与SEQ ID NO:4313-4564中任一个基本相同的HNH结构域。

[0302] 在一些情况下，所述核酸内切酶可以包含与SEQ ID NO:678-929中任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体。在一些情况下，所述核酸

内切酶可以与SEQ ID NO:678-929中任一个基本相同。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:678-929中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:678-929中任一个基本相同。

[0303] 在一些情况下,核酸内切酶可包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:678-929中任一个的N末端或C末端或与SEQ ID NO:678-929中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可包含与SEQ ID NO:5593-5608中任一个具有至少约80%、至少约85%、至少约90%、至少约95%、至少约99%同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本上相同的序列。NLS可以包含表1中的任何序列,或其组合:

[0304] 在一些情况下,序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数,并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11,延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0305] 在一些情况下,上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA),其能够与核酸内切酶形成复合物,带有与所期望的切割序列互补的5'靶向区域。在一些情况下,5'靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下,靶向区域的最5'核苷酸可以是G。在一些情况下,5'靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3'的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列,其前面是crRNA tracrRNA结合区3'的4-核苷酸接头。sgRNA可以包含,从5'到3':能够与细胞中的靶序列杂交的非天然指导核酸序列;和tracr序列。在一些情况下,非天然指导核酸序列和tracr序列共价连接。

[0306] 在一些情况下,tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA序列的至少约80%至至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可以与SEQ ID NO:5505的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80%的序列同一性。在一些情况下,tracrRNA可与SEQ ID NO:5505的至少约60-90(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的同一性。在一些情况下,tracrRNA可以与SEQ ID NO:5505的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸基本相同。tracrRNA可包含SEQ ID NO:5505。

[0307] 在一些情况下,能够与核酸内切酶形成复合物的至少一种工程改造的合成的指导核糖核酸(sgRNA)可以包含与SEQ ID NO:5469具有至少约80%同一性的序列。sgRNA可以包含与SEQ ID NO:5469具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的序列。sgRNA可包含与SEQ ID NO:5469基本相同的序列。

[0308] 在一些情况下,上述系统可包含两种不同的sgRNA,它们靶向第一区域和第二区域以在靶DNA基因座中切割,其中第二区域位于第一区域的3'。在一些情况下,上述系统可包含单链或双链DNA修复模板,其从5'到3'包含:第一同源臂(其包含第一区域5'的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列),至少约10个核苷酸的合成DNA序列,和第二同源臂(其包含第二区域3'的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0309] 另一方面,本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物,并且在复合物与目的靶核酸基因座结合后,可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下,靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0310] 在靶核酸基因座可以在细胞内的情况下,酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供,所述RuvC_III结构域与SEQ ID NO:2499-2750中任一个具有至少约75%(例如,至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%)同一性。包含编码所述核酸内切酶的开放阅读框的脱氧核糖核酸(DNA)可以包含与SEQ ID NO:5581基本相同的序列或与SEQ ID NO:5581具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,核酸包含启动子,编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供,所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下,生物体可以是真核生物。在一些情况下,生物体可以是真菌。在一些情况下,生物体可以是人。

[0311] MG15酶

[0312] 在一个方面,本公开提供了包含(a)核酸内切酶的工程改造的核酸酶系统。在一些

情况下,所述核酸内切酶是Cas核酸内切酶。在一些情况下,所述核酸内切酶是II型、II类Cas核酸内切酶。所述核酸内切酶可包含RuvC_III结构域,其中所述RuvC_III结构域与SEQ ID NO:2751-2913中任一个具有至少约70%的序列同一性。在一些情况下,所述核酸内切酶可包含RuvC_III结构域,其中该RuvC_III结构域与SEQ ID NO:2751-2913中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%同一性。在一些情况下,所述核酸内切酶可包含RuvC_III结构域,其中其与SEQ ID NO:2751-2913中任一个基本相同。所述核酸内切酶可包含与SEQ ID NO:2751-2913中任一个具有至少约70%序列同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:2751-2913中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、具有至少约99%同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可以包含与SEQ ID NO:2751-2913中任一个基本相同的RuvC_III结构域。

[0313] 所述核酸内切酶可包含与SEQ ID NO:4565-4727中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:4565-4727任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4565-4727中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:4565-4727中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:4565-4727任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4565-4727中任一个基本相同的HNH结构域。

[0314] 在一些情况下,所述核酸内切酶可包含与SEQ ID NO:930-1092中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:930-1092中任一个基本相同。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:930-1092中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:930-1092中任一个基本相同。

[0315] 在一些情况下,核酸内切酶可包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:930-1092中

任一个的N末端或C末端或与SEQ ID NO:930-1092中任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可包含与SEQ ID NO:5593-5608中任一个具有至少约80％、至少约85％、至少约90％、至少约95％、至少约99％同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本上相同的序列。NLS可以包含表1中的任何序列,或其组合:

[0316] 在一些情况下,序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数,并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11,延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0317] 在一些情况下,上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA),其能够与核酸内切酶形成复合物,带有与所期望的切割序列互补的5’靶向区域。在一些情况下,5’靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下,靶向区域的最5’核苷酸可以是G。在一些情况下,5’靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3’的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列,其前面是crRNA tracrRNA结合区3’的4-核苷酸接头。sgRNA可以包含,从5’到3’:能够与细胞中的靶序列杂交的非天然指导核酸序列;和tracr序列。在一些情况下,非天然指导核酸序列和tracr序列共价连接。

[0318] 在一些情况下,tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA序列的至少约80％至至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5506的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85,或至少约90)个连续核苷酸具有至少约80％的序列同一性。在一些情况下,tracrRNA可与SEQ ID NO:5506的至少约60-90(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性。在一些情况下,tracrRNA可以与SEQ ID NO:5506的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸基本相同。tracrRNA可以包含SEQ ID NO:5506。

[0319] 在一些情况下,能够与所述核酸内切酶形成复合物的至少一种工程改造的合成的指导核糖核酸(sgRNA)可以包含与SEQ ID NO:5470具有至少约80％同一性的序列。sgRNA可以包含与SEQ ID NO:5470具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的序列。sgRNA可包含与SEQ ID NO:5470基本相同的序列。

[0320] 在一些情况下,上述系统可包含两种不同的sgRNA,它们靶向第一区域和第二区域以在靶DNA基因座中切割,其中第二区域位于第一区域的3’。在一些情况下,上述系统可包含单链或双链DNA修复模板,其从5’到3’包含:第一同源臂(其包含第一区域5’的至少约20

(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列),至少约10个核苷酸的合成DNA序列,和第二同源臂(其包含第二区域3'的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0321] 另一方面,本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物,并且在复合物与目的靶核酸基因座结合后,可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下,靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0322] 在靶核酸基因座可以在细胞内的情况下,酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供,所述RuvC_III结构域与SEQ ID NO:2751-2913中任一个具有至少约75%(例如,至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%)同一性。包含编码所述核酸内切酶的开放阅读框的脱氧核糖核酸(DNA)可包含与SEQ ID NO:5582基本相同的序列或与SEQ ID NO:5582具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,核酸包含启动子,编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供,所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下,生物体可以是真核生物。在一些情况下,生物体可以是真菌。在一些情况下,生物体可以是人。

[0323] MG16酶

[0324] 一方面,本公开提供包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下,所述核酸内切酶是Cas核酸内切酶。在一些情况下,所述核酸内切酶是II型、II类Cas核酸内切酶。所述核酸内切酶可包含RuvC_III结构域,其中所述RuvC_III结构域与SEQ ID NO:2914-3174中任一个具有至少约70%的序列同一性。在一些情况下,所述核酸内切酶可包含RuvC_III结构域,其中该RuvC_III结构域与SEQ ID NO:2914-3174中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%同一性。在一些情况下,所述核酸内切酶可包含RuvC_III

结构域,其中与SEQ ID NO:2914-3174中任一个基本相同。所述核酸内切酶可包含与SEQ ID NO:2914-3174中任一个具有至少约70%序列同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:2914-3174中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可以包含与SEQ ID NO:2914-3174中任一个基本相同的RuvC_III结构域。

[0325] 所述核酸内切酶可包含与SEQ ID NO:4728-4988中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:4728-4988中任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4728-4988中任一个基本相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4728-4988中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:4728-4988中任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4728-4988中任一个基本相同的HNH结构域。

[0326] 在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1093-1353中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1093-1353中任一个基本相同。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1093-1353中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1093-1353中任一个基本相同。

[0327] 在一些情况下,所述核酸内切酶可以包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:1093-1353中任一个的N末端或C末端,或被附加到与SEQ ID NO:1093-1353中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可以包含与SEQ ID NO:5593-5608中任一个具有至少约80%、至少约85%、至少约90%、至少约95%、至少约99%同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本相同的序列。NLS可以包含表1中的任何序列或其组合:

[0328] 在一些情况下，序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数，并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11，延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0329] 在一些情况下，上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA)，其能够与核酸内切酶形成复合物，带有与所期望的切割序列互补的5’靶向区域。在一些情况下，5’靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下，靶向区域的最5’核苷酸可以是G。在一些情况下，5’靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3’的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列，其前面是crRNA tracrRNA结合区3’的4-核苷酸接头。sgRNA可以包含，从5’到3’：能够与细胞中的靶序列杂交的非天然指导核酸序列；和tracr序列。在一些情况下，非天然指导核酸序列和tracr序列共价连接。

[0330] 在一些情况下，tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA序列的至少约80％至至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5507的至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85个或至少约90)个连续核苷酸具有至少约80％序列同一性。在一些情况下，tracrRNA可与SEQ ID NO:5507的至少约60-90(例如至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％的同一性。在一些情况下，tracrRNA可以与SEQ ID NO:5507的至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸基本相同。tracrRNA可包含SEQ ID NO:5507。

[0331] 在一些情况下，能够与核酸内切酶形成复合物的至少一种工程改造的合成指导核糖核酸(sgRNA)可以包含与SEQ ID NO:5471具有至少约80％同一性的序列。sgRNA可包含与SEQ ID NO:5471具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的序列。sgRNA可以包含与SEQ ID NO:5471基本上相同的序列。

[0332] 在一些情况下，上述系统可包含两种不同的sgRNA，它们靶向第一区域和第二区域以在靶DNA基因座中切割，其中第二区域位于第一区域的3’。在一些情况下，上述系统可包含单链或双链DNA修复模板，其从5’到3’包含：第一同源臂(其包含第一区域5’的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)，至少约10个核苷酸的合成DNA序列，和第二同源臂(其包含第二区域3’的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0333] 另一方面，本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物，并且在复合物与目的靶核酸基因座结合后，可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述

系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下,靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0334] 在靶核酸基因座可以在细胞内的情况下,酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供,所述RuvC_III结构域与SEQ ID NO:2914-3174中任一个具有至少约75%(例如,至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%)同一性。包含编码所述核酸内切酶的开放阅读框的脱氧核糖核酸(DNA)可以包含与SEQ ID NO:5583基本相同的序列或与SEQ ID NO:5583具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,核酸包含启动子,编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供,所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下,生物体可以是真核生物。在一些情况下,生物体可以是真菌。在一些情况下,生物体可以是人。

[0335] MG18酶

[0336] 在一个方面,本公开提供了包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下,所述核酸内切酶是Cas核酸内切酶。在一些情况下,所述核酸内切酶是II型、II类Cas核酸内切酶。所述核酸内切酶可包含RuvC_III结构域,其中所述RuvC_III结构域与SEQ ID NO:3175-3300中任一个具有至少约70%的序列同一性。在一些情况下,所述核酸内切酶可包含RuvC_III结构域,其中该RuvC_III结构域与SEQ ID NO:3175-3300中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、在至少约98%、至少约99%同一性。在一些情况下,所述核酸内切酶可以包含RuvC_III结构域,其中与SEQ ID NO:3175-3300中任一个基本相同。所述核酸内切酶可以包含与SEQ ID NO:3175-3300中任一个具有至少约70%序列同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:3175-3300中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可以包含与

SEQ ID NO:3175-3300中任一个基本相同的RuvC_III结构域。

[0337] 所述核酸内切酶可包含与SEQ ID NO:4989-5146中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:4989-5146任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4989-5146中任一个基本相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4989-5146中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:4989-5146中任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:4989-5146中任一个基本相同的HNH结构域。

[0338] 在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1354-1511中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1354-1511中任一个基本相同。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1354-1511中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1354-1511中任一个基本相同。

[0339] 在一些情况下,所述核酸内切酶可以包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:1354-1511中任一个的N末端或C末端,或被附加到与SEQ ID NO:1354-1511中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可以包含与SEQ ID NO:5593-5608中任一个具有至少约80%、至少约85%、至少约90%、至少约95%、至少约99%同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本相同的序列。NLS可以包含表1中的任何序列或其组合:

[0340] 在一些情况下,序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数,并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11,延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0341] 在一些情况下,上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA),其能够与核酸内切酶形成复合物,带有与所期望的切割序列互补的5'靶向区域。在一些情况下,5'靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下,靶向区域

的最5’核苷酸可以是G。在一些情况下，5’靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3’的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列，其前面是crRNA tracrRNA结合区3’的4-核苷酸接头。sgRNA可以包含，从5’到3’：能够与细胞中的靶序列杂交的非天然指导核酸序列；和tracr序列。在一些情况下，非天然指导核酸序列和tracr序列共价连接。

[0342] 在一些情况下，tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA序列的至少约80％至至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5508的至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85，或至少约90)个连续核苷酸具有至少约80％序列同一性。在一些情况下，tracrRNA可与SEQ ID NO:5508的至少约60-90(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性。在一些情况下，tracrRNA可以与SEQ ID NO:5508的至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85，或至少约90)个连续核苷酸基本相同。tracrRNA可以包含SEQ ID NO:5508。

[0343] 在一些情况下，能够与核酸内切酶形成复合物的至少一种工程改造的合成的指导核糖核酸(sgRNA)可以包含与SEQ ID NO:5472具有至少约80％同一性的序列。sgRNA可以包含与SEQ ID NO:5472具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的序列。sgRNA可包含与SEQ ID NO:5472基本相同的序列。

[0344] 在一些情况下，上述系统可包含两种不同的sgRNA，它们靶向第一区域和第二区域以在靶DNA基因座中切割，其中第二区域位于第一区域的3’。在一些情况下，上述系统可包含单链或双链DNA修复模板，其从5’到3’包含：第一同源臂(其包含第一区域5’的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)，至少约10个核苷酸的合成DNA序列，和第二同源臂(其包含第二区域3’的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0345] 另一方面，本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物，并且在复合物与目的靶核酸基因座结合后，可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下，靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0346] 在靶核酸基因座可以在细胞内的情况下,酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供,所述RuvC_III结构域与SEQ ID NO:3175-3300中任一个具有至少约75%(例如,至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%)同一性。包含编码所述核酸内切酶的开放阅读框的脱氧核糖核酸(DNA)可以包含与SEQ ID NO:5584基本相同的序列或与SEQ ID NO:5584具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,核酸包含启动子,编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供,所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下,生物体可以是真核生物。在一些情况下,生物体可以是真菌。在一些情况下,生物体可以是人。

[0347] MG21酶

[0348] 在一个方面,本公开提供了包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下,所述核酸内切酶是Cas核酸内切酶。在一些情况下,所述核酸内切酶是II型、II类Cas核酸内切酶。所述核酸内切酶可包含RuvC_III结构域,其中所述RuvC_III结构域与SEQ ID NO:3331-3474中任一个具有至少约70%的序列同一性。在一些情况下,所述核酸内切酶可包含RuvC_III结构域,其中该RuvC_III结构域与SEQ ID NO:3331-3474中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%同一性。在一些情况下,所述核酸内切酶可以包含RuvC_III结构域,其中与SEQ ID NO:3331-3474中任一个基本相同。所述核酸内切酶可包含与SEQ ID NO:3331-3474中任一个具有至少约70%序列同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:3331-3474中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可以包含与SEQ ID NO:3331-3474中任一个基本相同的RuvC_III结构域。

[0349] 所述核酸内切酶可以包含与SEQ ID NO:5147-5290中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:5147-5290中任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:5147-5290中任一个基本相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:5147-5290中任一个具有至少约70%同

一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:5147-5290中任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:5147-5290中任一个基本相同的HNH结构域。

[0350] 在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1512-1655中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少具有约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1512-1655中任一个基本相同。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1512-1655中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1512-1655中任一个基本相同。

[0351] 在一些情况下,所述核酸内切酶可以包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:1512-1655中任一个的N末端或C末端,或被附加到与SEQ ID NO:1512-1655中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可以包含与SEQ ID NO:5593-5608中任一个具有至少约80%、至少约85%、至少约90%、至少约95%、至少约99%同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本相同的序列。NLS可以包含表1中的任何序列或其组合:

[0352] 在一些情况下,序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数,并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11,延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0353] 在一些情况下,上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA),其能够与核酸内切酶形成复合物,带有与所期望的切割序列互补的5'靶向区域。在一些情况下,5'靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下,靶向区域的最5'核苷酸可以是G。在一些情况下,5'靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3'的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列,其前面是crRNA tracrRNA结合区3'的4-核苷酸接头。sgRNA可以包含,从5'到3':能够与细胞中的靶序列杂交的非天然指导核酸序列;和tracr序列。在一些情况下,非天然指导核酸序列和tracr序列共价连接。

[0354] 在一些情况下,tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA

序列的至少约80％至至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5509的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸基本相同。在一些情况下,tracrRNA可与SEQ ID NO:5509的至少约60-90(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性。在一些情况下,tracrRNA可以与SEQ ID NO:5509的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸基本相同。tracrRNA可以包含SEQ ID NO:5509。

[0355] 在一些情况下,能够与所述核酸内切酶形成复合物的至少一种工程改造的合成的指导核糖核酸(sgRNA)可以包含与SEQ ID NO:5473具有至少约80％同一性的序列。sgRNA可以包含与SEQ ID NO:5473具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的序列。sgRNA可包含与SEQ ID NO:5473基本相同的序列。

[0356] 在一些情况下,上述系统可包含两种不同的sgRNA,它们靶向第一区域和第二区域以在靶DNA基因座中切割,其中第二区域位于第一区域的3’。在一些情况下,上述系统可包含单链或双链DNA修复模板,其从5’到3’包含:第一同源臂(其包含第一区域5’的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列),至少约10个核苷酸的合成DNA序列,和第二同源臂(其包含第二区域3’的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0357] 另一方面,本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物,并且在复合物与目的靶核酸基因座结合后,可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下,靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0358] 在靶核酸基因座可以在细胞内的情况下,酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供,所述RuvC_III结构域与SEQ ID NO:3331-3474中任一个具有至少约75％(例如,至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％)同一性。包含编码所述核酸内切酶的开放阅读框的脱氧核糖核酸(DNA)可以包含与SEQ ID NO:5585基本相同的序列或与SEQ ID NO:5585具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至

少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体。在一些情况下，核酸包含启动子，编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供，所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下，生物体可以是真核生物。在一些情况下，生物体可以是真菌。在一些情况下，生物体可以是人。

[0359] MG22酶

[0360] 在一个方面，本公开提供了包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下，所述核酸内切酶是Cas核酸内切酶。在一些情况下，所述核酸内切酶是II型、II类Cas核酸内切酶。所述核酸内切酶可包含RuvC_III结构域，其中所述RuvC_III结构域与SEQ ID NO:3475-3568中任一个具有至少约70％的序列同一性。在一些情况下，所述核酸内切酶可包含RuvC_III结构域，其中该RuvC_III结构域与SEQ ID NO:3475-3568中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％同一性。在一些情况下，所述核酸内切酶可包含RuvC_III结构域，其中与SEQ ID NO:3475-3568中任一个基本相同。所述核酸内切酶可包含与SEQ ID NO:3475-3568中任一个具有至少约70％序列同一性的RuvC_III结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:3475-3568中任一个具有至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、具有至少约99％同一性的RuvC_III结构域。在一些情况下，所述核酸内切酶可以包含与SEQ ID NO:3475-3568中任一个基本相同的RuvC_III结构域。

[0361] 所述核酸内切酶可以包含与SEQ ID NO:5291-5389中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:5291-5389中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:5291-5389中任一个基本相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:5291-5389中任一个具有至少约70％同一性的HNH结构域。在一些情况下，所述核酸内切酶可包含与SEQ ID NO:5291-5389中任一个至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:5291-5389中任一个基本相同的HNH结构域。

[0362] 在一些情况下，所述核酸内切酶可包含与SEQ ID NO:1656-1755中任一个具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至

少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1656-1755中任一个基本相同。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1656-1755中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1656-1755中任一个基本相同。

[0363] 在一些情况下,所述核酸内切酶可以包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:432-660中任一个的N末端或C末端,或被附加到与SEQ ID NO:1656-1755中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可以包含与SEQ ID NO:5593-5608中任一个具有至少约80%、至少约85%、至少约90%、至少约95%、至少约99%同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本相同的序列。NLS可以包含表1中的任何序列或其组合:

[0364] 在一些情况下,序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数,并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11,延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0365] 在一些情况下,上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA),其能够与核酸内切酶形成复合物,带有与所期望的切割序列互补的5'靶向区域。在一些情况下,5'靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下,靶向区域的最5'核苷酸可以是G。在一些情况下,5'靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3'的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列,其前面是crRNA tracrRNA结合区3'的4-核苷酸接头。sgRNA可以包含,从5'到3':能够与细胞中的靶序列杂交的非天然指导核酸序列;和tracr序列。在一些情况下,非天然指导核酸序列和tracr序列共价连接。

[0366] 在一些情况下,tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA序列的至少约80%至至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5510的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85,或至少约90)个连续核苷酸具有至少约80%的序列同一性。在一些情况下,tracrRNA可与SEQ ID NO:5510的至少约60-90(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少

约99％同一性。在一些情况下，tracrRNA可以与SEQ ID NO:5510的至少约60-100(例如，至少约60、至少约65、至少约70、至少约75、至少约80、至少约85，或至少约90)个连续核苷酸基本相同。tracrRNA可以包含SEQ ID NO:5510。

[0367] 在一些情况下，能够与所述核酸内切酶形成复合物的至少一种工程改造的合成的指导核糖核酸(sgRNA)可以包含与SEQ ID NO:5474具有至少约80％同一性的序列。sgRNA可以包含与SEQ ID NO:5474具有至少约80％、至少约85％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、具有至少约96％、至少约97％、至少约98％或至少约99％同一性的序列。sgRNA可包含与SEQ ID NO:5474基本相同的序列。

[0368] 在一些情况下，上述系统可包含两种不同的sgRNA，它们靶向第一区域和第二区域以在靶DNA基因座中切割，其中第二区域位于第一区域的3’。在一些情况下，上述系统可包含单链或双链DNA修复模板，其从5’到3’包含：第一同源臂(其包含第一区域5’的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)，至少约10个核苷酸的合成DNA序列，和第二同源臂(其包含第二区域3’的至少约20(例如，至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0369] 另一方面，本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物，并且在复合物与目的靶核酸基因座结合后，可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下，靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0370] 在靶核酸基因座可以在细胞内的情况下，酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供，所述RuvC_III结构域与SEQ ID NO:3475-3568中任一个具有至少约75％(例如，至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％)同一性。包含编码所述核酸内切酶的开放阅读框的脱氧核糖核酸(DNA)可以包含与SEQ ID NO:5586基本相同的序列或与SEQ ID NO:5586具有至少约30％、至少约35％、至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％同一性的变体。在一些情况下，核酸包含启动子，编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供，所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下，生物体可以是真核生物。在一

些情况下,生物体可以是真菌。在一些情况下,生物体可以是人。

[0371] MG23酶

[0372] 在一个方面,本公开提供了包含(a)核酸内切酶的工程改造的核酸酶系统。在一些情况下,所述核酸内切酶是Cas核酸内切酶。在一些情况下,所述核酸内切酶是II型、II类Cas核酸内切酶。所述核酸内切酶可包含RuvC_III结构域,其中所述RuvC_III结构域与SEQ ID NO:3569-3637中任一个具有至少约70%的序列同一性。在一些情况下,所述核酸内切酶可包含RuvC_III结构域,其中该RuvC_III结构域与SEQ ID NO:3569-3637中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%同一性。在一些情况下,所述核酸内切酶可包含RuvC_III结构域,其中与SEQ ID NO:3569-3637中任一个基本相同。所述核酸内切酶可以包含与SEQ ID NO:3569-3637中任一个具有至少约70%序列同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:3569-3637中任一个具有至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%同一性的RuvC_III结构域。在一些情况下,所述核酸内切酶可以包含与SEQ ID NO:3569-3637中任一个基本相同的RuvC_III结构域。

[0373] 所述核酸内切酶可包含与SEQ ID NO:5390-5460中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:5390-5460中任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:5390-5460中任一个基本相同的HNH结构域。所述核酸内切酶可包含与SEQ ID NO:5390-5460中任一个具有至少约70%同一性的HNH结构域。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:5390-5460任一个至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%相同的HNH结构域。所述核酸内切酶可以包含与SEQ ID NO:5390-5460中任一个基本相同的HNH结构域。

[0374] 在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1756-1826中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1756-1826中任一个基本相同。在一些情况下,所述核酸内切酶可包含与SEQ ID NO:1756-1826中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,所述核酸内切酶可以与SEQ ID NO:1756-1826中任一

个基本相同。

[0375] 在一些情况下,所述核酸内切酶可以包含具有一个或多个核定位序列(NLS)的变体。NLS可以在所述核酸内切酶的N末端或C末端的近端。NLS可以被附加到SEQ ID NO:1756-1826中任一个的N末端或C末端,或被附加到与SEQ ID NO:1756-1826中任一个具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体的N末端或C末端。NLS可以是SV40大T抗原NLS。NLS可以是c-myc NLS。NLS可以包含与SEQ ID NO:5593-5608中任一个具有至少约80%、至少约85%、至少约90%、至少约95%、至少约99%同一性的序列。NLS可以包含与SEQ ID NO:5593-5608中任一个基本相同的序列。NLS可以包含表1中的任何序列或其组合:

[0376] 在一些情况下,序列同一性可由BLASTP、CLUSTALW、MUSCLE、MAFFT、Novafold或Smith-Waterman同源性搜索算法确定。序列同一性可以通过BLASTP算法(其使用3的字长(W)、10的期望(E)的参数,并使用BLOSUM62评分矩阵(设置缺口(存在时)罚分为11,延伸为1)并使用条件型组成得分矩阵调整)来确定。

[0377] 在一些情况下,上述系统可包含(b)至少一种工程改造的合成指导核糖核酸(sgRNA),其能够与核酸内切酶形成复合物,带有与所期望的切割序列互补的5'靶向区域。在一些情况下,5'靶向区域可包含与核酸内切酶兼容的PAM序列。在一些情况下,靶向区域的最5'核苷酸可以是G。在一些情况下,5'靶向区域的长度可以是15-23个核苷酸。指导序列和tracr序列可以作为分开的核糖核酸(RNA)或单个核糖核酸(RNA)提供。指导RNA可包含靶向区域3'的crRNA tracrRNA结合序列。指导RNA可以包含tracrRNA序列,其前面是crRNA tracrRNA结合区3'的4-核苷酸接头。sgRNA可以包含,从5'到3':能够与细胞中的靶序列杂交的非天然指导核酸序列;和tracr序列。在一些情况下,非天然指导核酸序列和tracr序列共价连接。

[0378] 在一些情况下,tracr序列可以具有特定的序列。tracr序列可具有天然tracrRNA序列的至少约80%至至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸。tracr序列可与SEQ ID NO:5511的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80%的序列同一性。在一些情况下,tracrRNA可与SEQ ID NO:5511的至少约60-90(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85或至少约90)个连续核苷酸具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性。在一些情况下,tracrRNA可以与SEQ ID NO:5511的至少约60-100(例如,至少约60、至少约65、至少约70、至少约75、至少约80、至少约85,或至少约90)个连续核苷酸基本相同。tracrRNA可以包含SEQ ID NO:5511。

[0379] 在一些情况下,能够与所述核酸内切酶形成复合物的至少一种工程改造的合成的指导核糖核酸(sgRNA)可以包含与SEQ ID NO:5475具有至少约80%同一性的序列。sgRNA可以包含与SEQ ID NO:5475具有至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、具有至少约96%、至少约97%、至少约98%或

至少约99%同一性的序列。sgRNA可包含与SEQ ID NO:5475基本相同的序列。

[0380] 在一些情况下,上述系统可包含两种不同的sgRNA,它们靶向第一区域和第二区域以在靶DNA基因座中切割,其中第二区域位于第一区域的3'。在一些情况下,上述系统可包含单链或双链DNA修复模板,其从5'到3'包含:第一同源臂(其包含第一区域5'的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列),至少约10个核苷酸的合成DNA序列,和第二同源臂(其包含第二区域3'的至少约20(例如,至少约40、80、120、150、200、300、500或1kb)个核苷酸的序列)。

[0381] 另一方面,本公开提供了一种修饰目的靶核酸基因座的方法。所述方法可以包括将本文公开的任何非天然系统(包括本文公开的酶和至少一种合成的指导RNA(sgRNA))递送至靶核酸基因座。酶可与至少一种sgRNA形成复合物,并且在复合物与目的靶核酸基因座结合后,可修饰目的靶核酸基因座。将酶递送至所述基因座可包括用所述系统或编码所述系统的核酸转染细胞。将核酸酶递送至所述基因座可包括用所述系统或编码该系统的核酸电穿孔细胞。将核酸酶递送至所述基因座可以包括将系统在缓冲液中与包含目的基因座的核酸一起孵育。在一些情况下,靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。靶核酸基因座可包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。靶核酸基因座可以在细胞内。靶核酸基因座可以在体外。靶核酸基因座可以在真核细胞或原核细胞内。细胞可以是动物细胞、人细胞、细菌细胞、古细菌细胞或植物细胞。所述酶可以在目的靶基因座处或近端诱导单链或双链断裂。

[0382] 在靶核酸基因座可以在细胞内的情况下,酶可以作为含有编码具有RuvC_III结构域的酶的开放阅读框的核酸提供,所述RuvC_III结构域与SEQ ID NO:3569-3637中任一个具有至少约75%(例如,至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%)同一性。包含编码所述核酸内切酶的开放阅读框的脱氧核糖核酸(DNA)可以包含与SEQ ID NO:5587基本相同的序列或与SEQ ID NO:5587具有至少约30%、至少约35%、至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%同一性的变体。在一些情况下,核酸包含启动子,编码核酸内切酶的开放阅读框可操作地连接至所述启动子。启动子可以是CMV、EF1a、SV40、PGK1、Ubc、人β肌动蛋白、CAG、TRE或CaMKIIa启动子。核酸内切酶可以作为包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA提供。核酸内切酶可以作为翻译的多肽提供。至少一种工程改造的sgRNA可以作为脱氧核糖核酸(DNA)提供,所述脱氧核糖核酸(DNA)包含对与核糖核酸(RNA)pol III启动子可操作地连接的所述至少一种工程改造的sgRNA进行编码的基因序列。在一些情况下,生物体可以是真核生物。在一些情况下,生物体可以是真菌。在一些情况下,生物体可以是人。

[0383] 本发明提供了包括但不限于以下实施方式:

[0384] 1.一种工程改造的核酸酶系统,其包含:

[0385] (a)包含RuvC_III结构域和HNH结构域的核酸内切酶,其中所述核酸内切酶源自未培养的微生物,其中所述核酸内切酶是2类II型Cas核酸内切酶;和

[0386] (b)工程改造的指导核糖核酸结构,其被构造成与所述核酸内切酶形成复合物,所述复合物包含:

[0387] (i)指导核糖核酸序列,其被构造成与靶脱氧核糖核酸序列杂交;和

[0388] (ii)tracr核糖核酸序列,其被构造成与所述核酸内切酶结合。

[0389] 2.如实施方式1所述的工程改造的核酸酶系统,其中所述RuvC_III结构域包含与SEQ ID NO:1827-3637中任一个具有至少70%、至少75%、至少80%或至少90%序列同一性的序列。

[0390] 3.一种工程改造的核酸酶系统,其包含:

[0391] (a)包含与SEQ ID NO:1827-3637中任一个具有至少75%序列同一性的RuvC_III结构域的核酸内切酶;和

[0392] (b)工程改造的指导核糖核酸结构,其被构造成与所述核酸内切酶形成复合物,所述复合物包含:

[0393] (i)指导核糖核酸序列,其被构造成与靶脱氧核糖核酸序列杂交;和

[0394] (ii)tracr核糖核酸序列,其被构造成与所述核酸内切酶结合。

[0395] 4.一种工程改造的核酸酶系统,其包含:

[0396] (a)核酸内切酶,该核酸内切酶被构造成与包含SEQ ID NO:5512-5537的前间区序列邻近基序(PAM)序列结合,其中所述核酸内切酶是2类II型Cas核酸内切酶;和

[0397] (b)工程改造的指导核糖核酸结构,其被构造成与所述核酸内切酶形成复合物,所述复合物包含:

[0398] (i)指导核糖核酸序列,其被构造成与靶脱氧核糖核酸序列杂交;和

[0399] (ii)tracr核糖核酸序列,其被构造成与所述核酸内切酶结合。

[0400] 5.如实施方式4所述的工程改造的核酸酶系统,其中所述核酸内切酶源自未培养的微生物。

[0401] 6.如实施方式4-5中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶未被工程改造以结合不同的PAM序列。

[0402] 7.如实施方式4所述的工程改造的核酸酶系统,其中所述核酸内切酶不是Cas9核酸内切酶、Cas14核酸内切酶、Cas12a核酸内切酶、Cas12b核酸内切酶、Cas12c核酸内切酶、Cas12d核酸内切酶、Cas12e核酸内切酶、Cas13a核酸内切酶、Cas13b核酸内切酶、Cas13c核酸内切酶或Cas13d核酸内切酶。

[0403] 8.如实施方式4所述的工程改造的核酸酶系统,其中所述核酸内切酶与Cas9核酸内切酶具有小于80%的同一性。

[0404] 9.如实施方式3-8中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶进一步包含HNH结构域。

[0405] 10.如实施方式1-9中任一项所述的工程改造的核酸酶系统,其中所述tracr核糖核酸序列包含与选自SEQ ID NO:5476-5511和SEQ ID NO:5538中任一个的约60至90个连续核苷酸具有至少80%序列同一性的序列。

[0406] 11.一种工程改造的核酸酶系统,其包含:

[0407] (a)工程改造的指导核糖核酸结构,其包含:

[0408] (i)指导核糖核酸序列,其被构造成与靶脱氧核糖核酸序列杂交;和

[0409] (ii)tracr核糖核酸序列,其被构造成与核酸内切酶结合,

[0410] 其中所述tracr核糖核酸序列包含与选自SEQ ID NO:5476-5511和SEQ ID NO:

5538中任一个的约60至90个连续核苷酸具有至少80％序列同一性的序列;和

[0411] (b)2类II型Cas核酸内切酶,其被构造成与所述工程改造的指导核糖核酸结合。

[0412] 12.如实施方式1-3或11中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶被构造成与选自包含SEQ ID NO:5512-5537的组的前间区序列邻近基序(PAM)序列结合。

[0413] 13.如实施方式1-11中任一项所述的工程改造的核酸酶系统,其中所述工程改造的指导核糖核酸结构包含至少两个核糖核酸多核苷酸。

[0414] 14.如实施方式1-11中任一项所述的工程改造的核酸酶系统,其中所述工程改造的指导核糖核酸结构包含一个包含所述指导核糖核酸序列和所述tracr核糖核酸序列的核糖核酸多核苷酸。

[0415] 15.如实施方式1-14中任一项所述的工程改造的核酸酶系统,其中所述指导核糖核酸序列与原核、细菌、古细菌、真核、真菌、植物、哺乳动物或人基因组序列互补。

[0416] 16.如实施方式1-15中任一项所述的工程改造的核酸酶系统,其中所述指导核糖核酸序列的长度为15-24个核苷酸。

[0417] 17.如实施方式1-16中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含一个或多个在所述核酸内切酶的N末端或C末端的近端的核定位序列(NLS)。

[0418] 18.如实施方式1-17中任一项所述的工程改造的核酸酶系统,其中所述NLS包含选自SEQ ID NO:5597-5612的序列。

[0419] 19.如实施方式1-18中任一项所述的工程改造的核酸酶系统,其进一步包含单链或双链DNA修复模板,其从5’至3’包含:包含所述靶脱氧核糖核酸序列5’的至少20个核苷酸的序列的第一同源臂,至少10个核苷酸的合成DNA序列,以及包含所述靶序列3’的至少20个核苷酸的序列的第二同源臂。

[0420] 20.如实施方式19所述的工程改造的核酸酶系统,其中所述第一同源臂或所述第二同源臂包含至少40、80、120、150、200、300、500或1,000个核苷酸的序列。

[0421] 21.如实施方式1-20中任一项所述的工程改造的核酸酶系统,其中所述系统进一步包含Mg^{2+}源。

[0422] 22.如实施方式1-21中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶和所述tracr核糖核酸序列源自同一门内的不同细菌物种。

[0423] 23.如实施方式1-22中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶源自属于皮杆菌属的细菌。

[0424] 24.如实施方式1-22中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶源自属于疣微菌门、暂定异域菌门或暂定黑水仙菌门的细菌。

[0425] 25.如实施方式1-22中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶源自包含与SEQ ID NO:5592-5595中任一个具有至少90％同一性的16S rRNA基因的细菌。

[0426] 26.如实施方式1-25中任一项所述的工程改造的核酸酶系统,其中所述HNH结构域包含与SEQ ID NO:5638-5460中任一个具有至少70％或至少80％同一性的序列。

[0427] 27.如实施方式1-26中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含SEQ ID NO:1-1826或与其具有至少55％同一性的变体。

[0428] 28.如实施方式1-27中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:1827-1830或SEQ ID NO:1827-2140的序列至少70%、80%或90%相同的序列。

[0429] 29.如实施方式1-28中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:3638-3641或SEQ ID NO:3638-3954的序列至少70%、80%或90%相同的序列。

[0430] 30.如实施方式1-29中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含至少1个、至少2个、至少3个、至少4个或至少5个选自SEQ ID NO:5615-5632的肽基序。

[0431] 31.如实施方式1-30中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:1-4或SEQ ID NO:1-319的序列至少70%、80%或90%相同的序列。

[0432] 32.如实施方式1-31中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含与选自SEQ ID NO:5461-5464、SEQ ID NO:5476-5479或SEQ ID NO:5476-5489的序列至少70%、80%或90%相同的序列。

[0433] 33.如实施方式1-32中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含被预测为包含由杆和环组成的发夹的RNA序列,所述杆包含至少10个、至少12个或至少14个碱基配对的核糖核苷酸,以及所述环的4个碱基对内的不对称凸起。

[0434] 34.如实施方式1-33中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶被构造成与包含选自SEQ ID NO:5512-5515或SEQ ID NO:5527-5530的序列的PAM结合。

[0435] 35.如实施方式1-34中任一项所述的工程改造的核酸酶系统,其中:

[0436] a)所述核酸内切酶包含与SEQ ID NO:1827至少70%、至少80%或至少90%相同的序列;

[0437] b)所述指导RNA结构包含与SEQ ID NO:5461或SEQ ID NO:5476中的至少一个至少70%、至少80%或至少90%相同的序列;和

[0438] c)所述核酸内切酶被构造成与包含SEQ ID NO:5512或SEQ ID NO:5527的PAM结合。

[0439] 36.如实施方式1-34中任一项所述的工程改造的核酸酶系统,其中:

[0440] a)所述核酸内切酶包含与SEQ ID NO:1828至少70%、至少80%或至少90%相同的序列;

[0441] b)所述指导RNA结构包含与SEQ ID NO:5462或SEQ ID NO:5477中的至少一个至少70%、至少80%或至少90%相同的序列;和

[0442] c)所述核酸内切酶被构造成与包含SEQ ID NO:5513或SEQ ID NO:5528的PAM结合。

[0443] 37.如实施方式1-34中任一项所述的工程改造的核酸酶系统,其中:

[0444] a)所述核酸内切酶包含与SEQ ID NO:1829至少70%、至少80%或至少90%相同的序列;

[0445] b)所述指导RNA结构包含与SEQ ID NO:5463或SEQ ID NO:5478中的至少一个至少70%、至少80%或至少90%相同的序列;和

[0446] c)所述核酸内切酶被构造成与包含SEQ ID NO:5514或SEQ ID NO:5529的PAM结合。

[0447] 38.如实施方式1-34中任一项所述的工程改造的核酸酶系统,其中:

[0448] a)所述核酸内切酶包含与SEQ ID NO:1830至少70%、至少80%或至少90%相同的序列;

[0449] b)所述指导RNA结构包含与SEQ ID NO:5464或SEQ ID NO:5479中的至少一个至少70%、至少80%或至少90%相同的序列;和

[0450] c)所述核酸内切酶被构造成与包含SEQ ID NO:5515或SEQ ID NO:5530的PAM结合。

[0451] 39.如实施方式1-27中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:2141-2142或SEQ ID NO:2141-2241序列至少70%、80%或90%相同的序列。

[0452] 40.如实施方式1-27或实施方式39中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:3955-3956或SEQ ID NO:3955-4055的序列至少70%、80%或90%相同的序列。

[0453] 41.如实施方式1-27或实施方式39-40中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含选自SEQ ID NO:5632-5638的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。

[0454] 42.如实施方式1-27或实施方式39-41中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:320-321或SEQ ID NO:320-420中的序列至少70%、80%或90%相同的序列。

[0455] 43.如实施方式1-27或实施方式39-42中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含与选自以下的序列至少70%、80%或90%相同的序列:SEQ ID NO:5465、SEQ ID NO:5490-5491或SEQ ID NO:5490-5494。

[0456] 44.如实施方式1-27或实施方式39-43中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含含有发夹的tracr核糖核酸序列,所述发夹包含至少8个、至少10个或至少12个碱基配对的核糖核苷酸。

[0457] 45.如实施方式1-27或实施方式39-44中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶被构造成与包含选自SEQ ID NO:5516和SEQ ID NO:5531的序列的PAM结合。

[0458] 46.如实施方式1-27或实施方式39-45中任一项所述的工程改造的核酸酶系统,其中

[0459] a)所述核酸内切酶包含与SEQ ID NO:2141至少70%、80%或90%相同的序列;

[0460] b)所述指导RNA结构包含与SEQ ID NO:5490至少70%、80%或90%相同的序列;和

[0461] c)所述核酸内切酶被构造成与包含SEQ ID NO:5531的PAM结合。

[0462] 47.如实施方式1-27或实施方式39-45中任一项所述的工程改造的核酸酶系统,其中

[0463] a)所述核酸内切酶包含与SEQ ID NO:2142至少70%、80%或90%相同的序列;

[0464] b)所述指导RNA结构包含与SEQ ID NO:5465或SEQ ID NO:5491至少70%、80%或

90％相同的序列;和

[0465] c)所述核酸内切酶被构造成与包含SEQ ID NO:5516的PAM结合。

[0466] 48.如实施方式1-27中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:2245-2246中的序列至少70％、80％或90％相同的序列。

[0467] 49.如实施方式1-27或实施方式48中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:4059-4060的序列至少70％、80％或90％相同的序列。

[0468] 50.如实施方式1-27或实施方式48-49中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含选自SEQ ID NO:5639-5648的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。

[0469] 51.如实施方式1-27或实施方式48-50中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:424-425的序列至少70％、80％或90％相同的序列。

[0470] 52.如实施方式1-27或实施方式48-51中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含与选自SEQ ID NO:5498-5499和SEQ ID NO:5539的序列至少70％、80％或90％相同的序列。

[0471] 53.如实施方式1-27或实施方式48-52中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含被预测为包含具有不间断碱基配对区的发夹的指导核糖核酸序列,所述指导核糖核酸结构包含至少8个核苷酸的指导核糖核酸序列和至少8个核苷酸的tracr核糖核酸序列,并且其中所述tracr核糖核酸序列从5’至3’包含第一发夹和第二发夹,其中所述第一发夹具有比所述第二发夹更长的杆。

[0472] 54.如实施方式1-27中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:2242-2244或SEQ ID NO:2247-2249的序列至少70％、80％或90％相同的序列。

[0473] 55.如实施方式1-27或实施方式54中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:4056-4058和SEQ ID NO:4061-4063的序列至少70％、80％或90％相同的序列。

[0474] 56.如实施方式1-27或实施方式54-55中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含选自SEQ ID NO:5639-5648的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。

[0475] 57.如实施方式1-27或实施方式54-56中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:421-423或SEQ ID NO:426-428的序列至少70％、80％或90％相同的序列。

[0476] 58.如实施方式1-27或实施方式54-57中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含与选自以下的序列至少70％、80％或90％相同的序列:SEQ ID NO:5466-5467、SEQ ID NO:5495-5497、SEQ ID NO:5500-5502和SEQ ID NO:5539。

[0477] 59.如实施方式1-27或实施方式54-58中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含被预测为包含具有不间断碱基配对区的发夹的指导核糖核酸序列,所述指导核糖核酸结构包含至少8个核苷酸的指导核糖核酸序列和至少8个核苷酸的

tracr核糖核酸序列,并且其中所述tracr核糖核酸序列从5'至3'包含第一发夹和第二发夹,其中所述第一发夹具有比所述第二发夹更长的杆。

[0478] 60.如实施方式1-27或实施方式54-59中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶被构造成与包含选自SEQ ID NO:5517-5518或SEQ ID NO:5532-5534的序列的PAM结合。

[0479] 61.如实施方式1-27或实施方式54-60中任一项所述的工程改造的核酸酶系统,其中:

[0480] a)所述核酸内切酶包含与SEQ ID NO:2247至少70%、80%或90%相同的序列;

[0481] b)所述指导RNA结构包含与SEQ ID NO:5500至少70%、80%或90%相同的序列;和

[0482] c)所述核酸内切酶被构造成与包含SEQ ID NO:5517或SEQ ID NO:5532的PAM结合。

[0483] 62.如实施方式1-27或实施方式54-60中任一项所述的工程改造的核酸酶系统,其中:

[0484] a)所述核酸内切酶包含与SEQ ID NO:2248至少70%、80%或90%相同的序列;

[0485] b)所述指导RNA结构包含与SEQ ID NO:5501至少70%、80%或90%相同的序列;和

[0486] c)所述核酸内切酶被构造成与包含SEQ ID NO:5518或SEQ ID NO:5533的PAM结合。

[0487] 63.如实施方式1-27或实施方式54-60中任一项所述的工程改造的核酸酶系统,其中:

[0488] a)所述核酸内切酶包含与SEQ ID NO:2249至少70%、80%或90%相同的序列;

[0489] b)所述指导RNA结构包含与SEQ ID NO:5502至少70%、80%或90%相同的序列;和

[0490] c)所述核酸内切酶被构造成与包含SEQ ID NO:5534的PAM结合。

[0491] 64.如实施方式1-27中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:2253或SEQ ID NO:2253-2481的序列至少70%、80%或90%相同的序列。

[0492] 65.如实施方式1-27或实施方式64中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:4067或SEQ ID NO:4067-4295的序列至少70%、80%或90%相同的序列。

[0493] 66.如实施方式1-27或实施方式64-65中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含根据SEQ ID NO:5649的肽基序。

[0494] 67.如实施方式1-27或实施方式64-66中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:432或SEQ ID NO:432-660中的序列至少70%、80%或90%相同的序列。

[0495] 68.如实施方式1-27或实施方式64-67中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含与选自SEQ ID NO:ID NO:5468或SEQ ID NO:5503的序列至少70%、80%或90%相同的序列。

[0496] 69.如实施方式1-27或实施方式64-68中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶被构造成与包含选自SEQ ID NO:5519的序列的PAM结合。

[0497] 70.如实施方式1-27或实施方式64-69中任一项所述的工程改造的核酸酶系统,其

中：

[0498] a)所述核酸内切酶包含与SEQ ID NO:2253至少70％、80％或90％相同的序列；

[0499] b)所述指导RNA结构包含与SEQ ID NO:5468或SEQ ID NO:5503至少70％、80％或90％相同的序列；和

[0500] c)所述核酸内切酶被构造成与包含SEQ ID NO:5519的PAM结合。

[0501] 71.如实施方式1-27中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:2482-2489中的序列至少70％、80％或90％相同的序列。

[0502] 72.如实施方式1-27或实施方式71中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:4296-4303的序列至少70％、80％或90％相同的序列。

[0503] 73.如实施方式1-27或实施方式71-72中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:661-668的序列至少70％、80％或90％相同的序列。

[0504] 74.如实施方式1-27中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:2490-2498的序列至少70％、80％或90％相同的序列。

[0505] 75.如实施方式1-27或实施方式74中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:4304-4312的序列至少70％、80％或90％相同的序列。

[0506] 76.如实施方式1-27或实施方式74-75中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:669-677的序列至少70％、80％或90％相同的序列。

[0507] 77.如实施方式1-27或实施方式74-76中任一项所述的工程改造的核酸酶系统，其中所述指导RNA结构包含与选自SEQ ID NO:5504的序列至少70％、80％或90％相同的序列。

[0508] 78.如实施方式1-27中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:2499或SEQ ID NO:2499-2750的序列至少70％、80％或90％相同的序列。

[0509] 79.如实施方式1-27或实施方式78中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:4313或SEQ ID NO:4313-4564的序列至少70％、80％或90％相同的序列。

[0510] 80.如实施方式1-27或实施方式78-79中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含选自SEQ ID NO:5650-5667的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。

[0511] 81.如实施方式1-27或实施方式78-80中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:678或SEQ ID NO:678-929中的序列至少70％、80％或90％相同的序列。

[0512] 82.如实施方式1-27或实施方式78-81中任一项所述的工程改造的核酸酶系统，其中所述指导RNA结构包含与SEQ ID NO:5469或SEQ ID NO:5505至少70％、80％或90％相同的序列。

[0513] 83.如实施方式1-27或实施方式78-82中任一项所述的工程改造的核酸酶系统，其

中所述核酸内切酶被构造成与包含SEQ ID NO:5520或SEQ ID NO:5535的PAM结合。

[0514] 84.如实施方式1-27或实施方式78-83中任一项所述的工程改造的核酸酶系统,其中:

[0515] a)所述核酸内切酶包含与SEQ ID NO:2499至少70%、80%或90%相同的序列;

[0516] b)所述指导RNA结构包含与SEQ ID NO:5469或SEQ ID NO:5505至少70%、80%或90%相同的序列;和

[0517] c)所述核酸内切酶被构造成与包含SEQ ID NO:5520或SEQ ID NO:5535的PAM结合。

[0518] 85.如实施方式1-27中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:2751或SEQ ID NO:2751-2913的序列至少70%、80%或90%相同的序列。

[0519] 86.如实施方式1-27或实施方式85中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:4565或SEQ ID NO:4565-4727的序列至少70%、80%或90%相同的序列。

[0520] 87.如实施方式1-27或实施方式85-86中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含选自SEQ ID NO:5668-5678的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。

[0521] 88.如实施方式1-27或实施方式85-87中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:930或SEQ ID NO:930-1092中的序列至少70%、80%或90%相同的序列。

[0522] 89.如实施方式1-27或实施方式85-88中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含与SEQ ID NO:5470或SEQ ID NO:5506至少70%、80%或90%相同的序列。

[0523] 90.如实施方式1-27或实施方式85-89中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶被构造成与包含选自SEQ ID NO:5521或SEQ ID NO:5536的序列的PAM结合。

[0524] 91.如实施方式1-27或实施方式85-90中任一项所述的工程改造的核酸酶系统,其中:

[0525] a)所述核酸内切酶包含与SEQ ID NO:2751至少70%、80%或90%相同的序列;

[0526] b)所述指导RNA结构包含与SEQ ID NO:5470或SEQ ID NO:5506至少70%、80%或90%相同的序列;和

[0527] c)所述核酸内切酶被构造成与包含SEQ ID NO:5521或SEQ ID NO:5536的PAM结合。

[0528] 92.如实施方式1-27中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:2914或SEQ ID NO:2914-3174的序列至少70%、80%或90%相同的序列。

[0529] 93.如实施方式1-27或实施方式92中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:4728或SEQ ID NO:4728-4988的序列至少70%、80%或90%相同的序列。

[0530] 94.如实施方式1-27或实施方式92-93中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含选自SEQ ID NO:5676-5678的至少1个、至少2个或至少3个肽基序。

[0531] 95.如实施方式1-27或实施方式92-94中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:1093或SEQ ID NO:1093-1353中的序列至少70％、80％或90％相同的序列。

[0532] 96.如实施方式1-27或实施方式92-95中任一项所述的工程改造的核酸酶系统，其中所述指导RNA结构包含与选自以下的序列至少70％、80％或90％相同的序列：SEQ ID NO:5471、SEQ ID NO:5507和SEQ ID NO:5540-5542。

[0533] 97.如实施方式1-27或实施方式92-96中任一项所述的工程改造的核酸酶系统，其中所述指导RNA结构包含被预测为包含至少两个发夹的tracr核糖核酸序列，所述发夹包含少于5个碱基配对的核糖核苷酸。

[0534] 98.如实施方式1-27或实施方式92-97中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶被构造成与包含SEQ ID NO:5522的PAM结合。

[0535] 99.如实施方式1-27或实施方式92-98中任一项所述的工程改造的核酸酶系统，其中：

[0536] a)所述核酸内切酶包含与SEQ ID NO:2914至少70％、80％或90％相同的序列；

[0537] b)所述指导RNA结构包含与SEQ ID NO:5471或SEQ ID NO:5507至少70％、80％或90％相同的序列；和

[0538] c)所述核酸内切酶被构造成与包含SEQ ID NO:5522的PAM结合。

[0539] 100.如实施方式1-27中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:3175或SEQ ID NO:3175-3330的序列至少70％、80％或90％相同的序列。

[0540] 101.如实施方式1-27或实施方式100中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:4989或SEQ ID NO:4989-5146的序列至少70％、80％或90％相同的序列。

[0541] 102.如实施方式1-27或实施方式100-101中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含选自SEQ ID NO:5679-5686的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。

[0542] 103.如实施方式1-27或实施方式100-102中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:1354或SEQ ID NO:1354-1511中的序列至少70％、80％或90％相同的序列。

[0543] 104.如实施方式1-27或实施方式100-103中任一项所述的工程改造的核酸酶系统，其中所述指导RNA结构包含与选自SEQ ID NO:5472或SEQ ID NO:5508的序列至少70％、80％或90％相同的序列。

[0544] 105.如实施方式1-27或实施方式100-104中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶被构造成与包含选自SEQ ID NO:5523或SEQ ID NO:5537的序列的PAM结合。

[0545] 106.如实施方式1-27或实施方式100-105中任一项所述的工程改造的核酸酶系统，其中

[0546] a)所述核酸内切酶包含与SEQ ID NO:3175至少70%、80%或90%相同的序列;

[0547] b)所述指导RNA结构包含与SEQ ID NO:5472或SEQ ID NO:5508至少70%、80%或90%相同的序列;和

[0548] c)所述核酸内切酶被构造成与包含SEQ ID NO:5523或SEQ ID NO:5537的PAM结合。

[0549] 107.如实施方式1-27中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:3331或SEQ ID NO:3331-3474的序列至少70%、80%或90%相同的序列。

[0550] 108.如实施方式1-27或实施方式107中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:5147或SEQ ID NO:5147-5290的序列至少70%、80%或90%相同的序列。

[0551] 109.如实施方式1-27或实施方式107-108中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含选自SEQ ID NO:5674-5675和SEQ ID NO:5687-5693的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。

[0552] 110.如实施方式1-27或实施方式107-109中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:1512或SEQ ID NO:1512-1655中的序列至少70%、80%或90%相同的序列。

[0553] 111.如实施方式1-27或实施方式107-110中任一项所述的工程改造的核酸酶系统,其中所述指导RNA结构包含与选自SEQ ID NO:5473或SEQ ID NO:5509的序列至少70%、80%或90%相同的序列。

[0554] 112.如实施方式1-27或实施方式107-111中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶被构造成与包含SEQ ID NO:5524的PAM结合。

[0555] 113.如实施方式1-27或实施方式107-112中任一项所述的工程改造的核酸酶系统,其中:

[0556] a)所述核酸内切酶包含与SEQ ID NO:3331至少70%、80%或90%相同的序列;

[0557] b)所述指导RNA结构包含与SEQ ID NO:5473或SEQ ID NO:5509至少70%、80%或90%相同的序列;和

[0558] c)所述核酸内切酶被构造成与包含SEQ ID NO:5524的PAM结合。

[0559] 114.如实施方式1-27中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:3475或SEQ ID NO:3475-3568的序列至少70%、80%或90%相同的序列。

[0560] 115.如实施方式1-27或实施方式114中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:5291或SEQ ID NO:5291-5389的序列至少70%、80%或90%相同的序列。

[0561] 116.如实施方式1-27或实施方式114-115中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含选自SEQ ID NO:5694-5699的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。

[0562] 117.如实施方式1-27或实施方式114-116中任一项所述的工程改造的核酸酶系统,其中所述核酸内切酶包含与选自SEQ ID NO:1656或SEQ ID NO:1656-1755的序列至少

70％、80％或90％相同的序列。

[0563] 118.如实施方式1-27或实施方式114-117中任一项所述的工程改造的核酸酶系统，其中所述指导RNA结构包含与SEQ ID NO:5474或SEQ ID NO:5510至少70％、80％或90％相同的序列。

[0564] 119.如实施方式1-27或实施方式114-118中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶被构造成与包含SEQ ID NO:5525的PAM结合。

[0565] 120.如实施方式1-27或实施方式114-119中任一项所述的工程改造的核酸酶系统，其中：

[0566] a)所述核酸内切酶包含与SEQ ID NO:3475至少70％、80％或90％相同的序列；

[0567] b)所述指导RNA结构包含与SEQ ID NO:5474或SEQ ID NO:5510至少70％、80％或90％相同的序列；和

[0568] c)所述核酸内切酶被构造成与包含SEQ ID NO:5525的PAM结合。

[0569] 121.如实施方式1-27中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:3569或SEQ ID NO:3569-3637的序列至少70％、80％或90％相同的序列。

[0570] 122.如实施方式1-27或实施方式121中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:5390或SEQ ID NO:5390-5460的序列至少70％、80％或90％相同的序列。

[0571] 123.如实施方式1-27或实施方式121-122中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含选自SEQ ID NO:5700-5717的至少1个、至少2个、至少3个、至少4个或至少5个肽基序。

[0572] 124.如实施方式1-27或实施方式121-123中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶包含与选自SEQ ID NO:1756或SEQ ID NO:1756-1826中的序列至少70％、80％或90％相同的序列。

[0573] 125.如实施方式1-27或实施方式121-124中任一项所述的工程改造的核酸酶系统，其中所述指导RNA结构包含与SEQ ID NO:5475或SEQ ID NO:5511至少70％、80％或90％相同的序列。

[0574] 126.如实施方式1-27或实施方式121-125中任一项所述的工程改造的核酸酶系统，其中所述核酸内切酶被构造成与包含SEQ ID NO:5526的PAM结合。

[0575] 127.如实施方式1-27或实施方式121-126中任一项所述的工程改造的核酸酶系统，其中：

[0576] a)所述核酸内切酶包含与SEQ ID NO:3569至少70％、80％或90％相同的序列；

[0577] b)所述指导RNA结构包含与SEQ ID NO:5475或SEQ ID NO:5511至少70％、80％或90％相同的序列；和

[0578] c)所述核酸内切酶被构造成与包含SEQ ID NO:5526的PAM结合。

[0579] 128.如实施方式1-127中任一项所述的工程改造的核酸酶系统，其中所述序列同一性由BLASTP、CLUSTALW、MUSCLE、MAFFT或Smith-Waterman同源性搜索算法确定。

[0580] 129.如实施方式128所述的工程改造的核酸酶系统，其中所述序列同一性通过所述BLASTP同源性搜索算法来确定，所述BLASTP同源性搜索算法使用3的字长(W)、10的期望

(E)的参数和设置缺口存在时罚分为11、延伸为1的BLOSUM62评分矩阵并使用条件型组成得分矩阵调整。

[0581] 130.一种工程改造的指导核糖核酸多核苷酸,其包含:

[0582] a)DNA靶向区段,其包含与靶DNA分子中的靶序列互补的核苷酸序列;和

[0583] b)蛋白质结合区段,其包含杂交形成双链RNA(dsRNA)双链体的两个互补核苷酸段,

[0584] 其中所述两个互补核苷酸段通过居间核苷酸共价连接到彼此,以及

[0585] 其中所述工程改造的指导核糖核酸多核苷酸被构造成与包含与SEQ ID NO:1827-3637中任一个具有至少75%序列同一性的RuvC_III结构域的核酸内切酶形成复合物并将所述复合物靶向所述靶DNA分子的所述靶序列。

[0586] 131.如实施方式130所述的工程改造的指导核糖核酸多核苷酸,其中所述DNA靶向区段位于所述两个互补核苷酸段的5'。

[0587] 132.如实施方式130-131中任一项所述的工程改造的指导核糖核酸多核苷酸,其中:

[0588] a)所述蛋白质结合区段包含与选自SEQ ID NO:5476-5479或SEQ ID NO:5476-5489的序列具有至少70%、至少80%或至少90%同一性的序列;

[0589] b)所述蛋白质结合区段包含与选自(SEQ ID NO:5490-5491或SEQ ID NO:5490-5494)和SEQ ID NO:5538的序列具有至少70%、至少80%或至少90%同一性的序列;

[0590] c)所述蛋白质结合区段包含与选自SEQ ID NO:5498-5499的序列具有至少70%、至少80%或至少90%同一性的序列;

[0591] d)所述蛋白质结合区段包含与选自SEQ ID NO:5495-5497和SEQ ID NO:5500-5502的序列具有至少70%、至少80%或至少90%同一性的序列;

[0592] e)所述蛋白质结合区段包含与SEQ ID NO:5503具有至少70%、至少80%或至少90%同一性的序列;

[0593] f)所述蛋白质结合区段包含与SEQ ID NO:5504具有至少70%、至少80%或至少90%同一性的序列;

[0594] g)所述蛋白质结合区段包含与SEQ ID NO:5505具有至少70%、至少80%或至少90%同一性的序列;

[0595] h)所述蛋白质结合区段包含与SEQ ID NO:5506具有至少70%、至少80%或至少90%同一性的序列;

[0596] i)所述蛋白质结合区段包含与SEQ ID NO:5507具有至少70%、至少80%或至少90%同一性的序列;

[0597] j)所述蛋白质结合区段包含与SEQ ID NO:5508具有至少70%、至少80%或至少90%同一性的序列;

[0598] k)所述蛋白质结合区段包含与SEQ ID NO:5509具有至少70%、至少80%或至少90%同一性的序列;

[0599] l)所述蛋白质结合区段包含与SEQ ID NO:5510具有至少70%、至少80%或至少90%同一性的序列;或

[0600] m)所述蛋白质结合区段包含与SEQ ID NO:5511具有至少70%、至少80%或至少

90％同一性的序列。

[0601] 133.如实施方式130-132中任一项所述的工程改造指导核糖核酸多核苷酸,其中:

[0602] a)所述指导核糖核酸多核苷酸包含含有发夹的RNA序列,所述发夹包含杆和环,其中所述杆包含至少10个、至少12个或至少14个碱基配对的核糖核苷酸和所述环的4个碱基对内的不对称凸起;

[0603] b)所述指导核糖核酸多核苷酸包含被预测包含发夹的tracr核糖核酸序列,所述发夹包含至少8个、至少10个或至少12个碱基配对的核糖核苷酸;

[0604] c)所述指导RNA结构包含被预测包含具有不间断碱基配对区的发夹的指导核糖核酸序列,所述指导核糖核酸结构包含至少8个核苷酸的指导核糖核酸序列和至少8个核苷酸的tracr核糖核酸序列,并且其中所述tracr核糖核酸序列从5’至3’包含第一发夹和第二发夹,所述第一发夹具有比所述第二发夹更长的杆;或

[0605] d)所述指导核糖核酸多核苷酸包含被预测包含至少两个发夹的tracr核糖核酸序列,所述发夹包含少于5个碱基配对的核糖核苷酸。

[0606] 134.一种脱氧核糖核酸多核苷酸,其编码实施方式130-133中任一项所述的工程改造的指导核糖核酸多核苷酸。

[0607] 135.一种包含优化用于在生物体中表达的工程改造的核酸序列的核酸,其中所述核酸编码包含RuvC_III结构域和HNH结构域的2类II型Cas核酸内切酶,并且其中所述核酸内切酶源自未培养的微生物。

[0608] 136.一种包含优化用于在生物体中表达的工程改造的核酸序列的核酸,其中所述核酸编码包含与SEQ ID NO:1827-3637中任一个具有至少70％序列同一性的RuvC_III结构域的核酸内切酶。

[0609] 137.如实施方式135-136中任一项所述的核酸,其中所述核酸内切酶包含与SEQ ID NO:3638-5460中任一个具有至少70％或至少80％序列同一性的HNH结构域。

[0610] 138.如实施方式135-137中任一项所述的核酸,其中所述核酸内切酶包含SEQ ID NO:5572-5591或与其具有至少70％序列同一性的变体。

[0611] 139.如实施方式135-138中任一项所述的核酸,其中所述核酸内切酶包含对所述核酸内切酶的N末端或C末端的近端的一个或多个核定位序列(NLS)进行编码的序列。

[0612] 140.如实施方式139所述的核酸,其中所述NLS包含选自SEQ ID NO:5597-5612的序列。

[0613] 141.如实施方式135-140中任一项所述的核酸,其中所述生物体是原核生物体、细菌生物体、真核生物体、真菌生物体、植物生物体、哺乳动物生物体、啮齿动物生物体或人。

[0614] 142.如实施方式141所述的核酸,其中所述生物体是大肠杆菌,并且其中:

[0615] a)所述核酸序列与选自SEQ ID NO:5572-5575的序列具有至少70％、80％或90％的同一性;

[0616] b)所述核酸序列与选自SEQ ID NO:5576-5577的序列具有至少70％、80％或90％的同一性;

[0617] c)所述核酸序列与选自SEQ ID NO:5578-5580的序列具有至少70％、80％或90％的同一性;

[0618] d)所述核酸序列与SEQ ID NO:5581具有至少70％、80％或90％的同一性;

[0619] e)所述核酸序列与SEQ ID NO:5582具有至少70%、80%或90%的同一性;

[0620] f)所述核酸序列与SEQ ID NO:5583具有至少70%、80%或90%的同一性;

[0621] g)所述核酸序列与SEQ ID NO:5584具有至少70%、80%或90%的同一性;

[0622] h)所述核酸序列与SEQ ID NO:5585具有至少70%、80%或90%的同一性;

[0623] i)所述核酸序列与SEQ ID NO:5586具有至少70%、80%或90%的同一性;或

[0624] j)所述核酸序列与SEQ ID NO:5587具有至少70%、80%或90%的同一性。

[0625] 143.如实施方式141所述的核酸,其中所述生物体是人,并且其中

[0626] a)所述核酸序列与SEQ ID NO:5588或SEQ ID NO:5589具有至少70%、80%或90%的同一性;或

[0627] b)所述核酸序列与SEQ ID NO:5590或SEQ ID NO:5591具有至少70%、80%或90%的同一性。

[0628] 144.一种载体,其包含编码包含RuvC_III结构域和HNH结构域的2类II型Cas核酸内切酶的核酸序列,其中所述核酸内切酶源自未培养的微生物。

[0629] 145.一种载体,其包含如实施方式135-143中任一项所述的核酸。

[0630] 146.如实施方式144-145中任一项所述的载体,其进一步包含编码工程改造的指导核糖核酸结构的核酸,所述工程改造的指导核糖核酸结构被构造成与所述核酸内切酶形成复合物,所述复合物包含:

[0631] a)指导核糖核酸序列,其被构造成与靶脱氧核糖核酸序列杂交;和

[0632] b)tracr核糖核酸序列,其被构造成与所述核酸内切酶结合。

[0633] 147.如实施方式144-146中任一项所述的载体,其中所述载体是质粒、小环、CELiD、腺相关病毒(AAV)衍生的病毒体或慢病毒。

[0634] 148.一种细胞,其包含如实施方式144-147中任一项所述的载体。

[0635] 149.一种制造核酸内切酶的方法,其包括培养如实施方式146所述的细胞。

[0636] 150.一种用于对双链脱氧核糖核酸多核苷酸进行结合、切割、标记或修饰的方法,所述方法包括:

[0637] (a)使所述双链脱氧核糖核酸多核苷酸与跟工程改造的指导核糖核酸结构复合的2类II型Cas核酸内切酶接触,所述工程改造的指导核糖核酸结构被构造成与所述核酸内切酶和所述双链脱氧核糖核酸多核苷酸结合;

[0638] (b)其中所述双链脱氧核糖核酸多核苷酸包含前间区序列邻近基序(PAM);并且

[0639] (c)其中所述PAM包含选自SEQ ID NO:5512-5526或SEQ ID NO:5527-5537的序列。

[0640] 151.如实施方式149所述的方法,其中所述双链脱氧核糖核酸多核苷酸包含第一链和第二链,所述第一链包含与所述工程改造的指导核糖核酸结构的序列互补的序列,所述第二链包含所述PAM。

[0641] 152.如实施方式151所述的方法,其中所述PAM与跟所述工程改造的指导核糖核酸结构的所述序列互补的所述序列的3'末端直接相邻。

[0642] 153.如实施方式149-152中任一项所述的方法,其中所述2类II型Cas核酸内切酶不是Cas9核酸内切酶、Cas14核酸内切酶、Cas12a核酸内切酶、Cas12b核酸内切酶、Cas12c核酸内切酶、Cas12d核酸内切酶、Cas12e核酸内切酶、Cas13a核酸内切酶、Cas13b核酸内切酶、Cas13c核酸内切酶或Cas 13d核酸内切酶。

[0643] 154.如实施方式149-153中任一项所述的方法，其中所述2类II型Cas核酸内切酶源自未培养的微生物。

[0644] 155.如实施方式149-154中任一项所述的方法，其中所述双链脱氧核糖核酸多核苷酸是真核、植物、真菌、哺乳动物、啮齿动物或人双链脱氧核糖核酸多核苷酸。

[0645] 156.如实施方式149-155中任一项所述的方法，其中：

[0646] a)所述PAM包含选自SEQ ID NO:5512-5515和SEQ ID NO:5527-5530的序列；

[0647] b)所述PAM包含SEQ ID NO:5516或SEQ ID NO:5531；

[0648] c)所述PAM包含SEQ ID NO:5539；

[0649] d)所述PAM包含SEQ ID NO:5517或SEQ ID NO:5518；

[0650] e)所述PAM包含SEQ ID NO:5519；

[0651] f)所述PAM包含SEQ ID NO:5520或SEQ ID NO:5535；

[0652] g)所述PAM包含SEQ ID NO:5521或SEQ ID NO:5536；

[0653] h)所述PAM包含SEQ ID NO:5522；

[0654] i)所述PAM包含SEQ ID NO:5523或SEQ ID NO:5537；

[0655] j)所述PAM包括SEQ ID NO:5524；

[0656] k)所述PAM包含SEQ ID NO:5525；或

[0657] l)所述PAM包含SEQ ID NO:5526。

[0658] 157.一种修饰靶核酸基因座的方法，所述方法包括将如实施方式1-129中任一项所述的所述工程改造的核酸酶系统递送至所述靶核酸基因座，其中所述核酸内切酶被构造成与所述工程改造的指导核糖核酸结构形成复合物，并且其中所述复合物被构造成使得在所述复合物与所述靶核酸基因座结合后，所述复合物修饰所述靶核酸基因座。

[0659] 158.如实施方式156所述的方法，其中修饰所述靶核酸基因座包括对所述靶核酸基因座进行结合、切口、切割或标记。

[0660] 159.如实施方式156-158中任一项所述的方法，其中所述靶核酸基因座包含脱氧核糖核酸(DNA)或核糖核酸(RNA)。

[0661] 160.如实施方式159所述的方法，其中所述靶核酸包含基因组DNA、病毒DNA、病毒RNA或细菌DNA。

[0662] 161.如实施方式156-160中任一项所述的方法，其中所述靶核酸基因座是在体外。

[0663] 162.如实施方式156-160中任一项所述的方法，其中所述靶核酸基因座在细胞内。

[0664] 163.如实施方式162所述的方法，其中所述细胞是原核细胞、细菌细胞、真核细胞、真菌细胞、植物细胞、动物细胞、哺乳动物细胞、啮齿动物细胞、灵长类细胞或人细胞。

[0665] 164.如实施方式162-163任一项所述的方法，其中将所述工程改造的核酸酶系统递送至所述靶核酸基因座包括递送如实施方式135-140任一项所述的核酸或如实施方式142-146任一项所述的载体。

[0666] 165.如实施方式162-163中任一项所述的方法，其中将所述工程改造的核酸酶系统递送至所述靶核酸基因座包括递送包含编码所述核酸内切酶的开放阅读框的核酸。

[0667] 166.如实施方式164的方法，其中所述核酸包含启动子，编码所述核酸内切酶的所述开放阅读框可操作地连接至所述启动子。

[0668] 167.如实施方式162-163中任一项所述的方法，其中将所述工程改造的核酸酶系

统递送至所述靶核酸基因座包括递送包含编码所述核酸内切酶的所述开放阅读框的加帽mRNA。

[0669] 168.如实施方式162-163中任一项所述的方法,其中将所述工程改造的核酸酶系统递送至所述靶核酸基因座包括递送翻译的多肽。

[0670] 169.如实施方式162-163中任一项所述的方法,其中将所述工程改造的核酸酶系统递送至所述靶核酸基因座包括递送对与核糖核酸(RNA)pol III启动子可操作地连接的所述工程改造的指导核糖核酸结构进行编码的脱氧核糖核酸(DNA)。

[0671] 170.如实施方式156-169中任一项所述的方法,所述核酸内切酶在所述靶基因座处或所述靶基因座的近端诱导单链断裂或双链断裂。

[0672] 实施例

[0673] 实施例1.-新蛋白质的宏基因组分析

[0674] 从沉积物、土壤和动物中收集宏基因组样品。使用Zymobiomics DNA微量制备型试剂盒提取脱氧核糖核酸(DNA),并在Illumina HiSeq®2500上测序。在征得业主同意的情况下采集样品。来自公共来源的其他原始序列数据包括动物微生物组、沉积物、土壤、温泉、热液喷口、海洋、泥炭沼泽、永久冻土和污水序列。使用基于已知的Cas蛋白序列(包括II型Cas效应物蛋白)生成的隐马尔可夫模型搜索宏基因组序列数据。将通过搜索鉴定的新颖效应物蛋白与已知蛋白质进行比对,以鉴定潜在的活性位点。这种宏基因组工作流程产生了对本文所述的II类、II型CRISPR核酸内切酶的MG1、MG2、MG3、MG4、MG6、MG14、MG15、MG16、MG18、MG21、MG22和MG23家族的描述。

[0675] 实施例2A.-发现CRISPR系统的MG1家族

[0676] 对来自实施例1的宏基因组分析的数据的分析揭示了先前未描述的推定CRISPR系统的新簇,所述系统最初包括六个成员(MG1-1、MG1-2、MG1-3、MG1-4、MG1-5和MG1-6,分别记为SEQ ID NO:5、6、1、2和3)。该家族的特征是带有HNH和RuvC结构域的酶。该家族的RuvC结构域具有与先前描述的Cas9家族成员具有低同源性的RuvC_III部分。虽然最初的家族成员在它们之间具有最大56.8%的同一性,但所有6种酶都表现出RuvC结构域的不同RuvC_III部分,并带有RHHALDAMV(SEQ ID NO:5615)、KHHALDAMC(SEQ ID NO:5616)或KHHALDAIC(SEQ ID NO:5617)的共同基序。在其他描述的Cas9样酶中没有发现这些基序。这些新酶及其相关子结构域的相应蛋白质和核酸序列在序列表中提供。基于它们相对于其他基因的位置鉴定了推定的tracrRNA序列,并显示为SEQ ID NO:5476-5479。根据来自包含CRISPR系统的基因组箱的16S rRNA序列,酶系统似乎源自疣微菌门、暂定异域菌门或暂定黑水仙菌门。16S rRNA序列表示为SEQ ID NO:5592-5596。一起调用了Shmakov等人(Mol Cell.2015年11月5日;60(3):385-97)(其全部内容通过引用并入)描述的特征的CRISPR系统序列的详细结构域级比对在图9A、9B、9C、9D、9E、9F、9G和9H中描绘。MG1-1、1-2和1-3与另外的专有蛋白质数据集的比较揭示了具有相似结构的另外蛋白质序列,呈现为SEQ NO:7-319。这些MG1蛋白序列导致发现了如SEQ ID NO:5618-5632中所示的另外的MG1基序。

[0677] 实施例2B.-发现CRISPR系统的MG2家族

[0678] 对来自实施例1的宏基因组分析的数据的分析揭示了先前未描述的推定CRISPR系统的新簇,所述系统包括六个成员(MG2-1、MG2-2、MG2-3、MG2-5和MG2-6)。这些新酶及示例性子结构域的相应蛋白质和核酸序列呈现为SEQ ID NO:320、322-325。基于它们相对于其

他基因的位置,在操纵子中鉴定了推定的tracrRNA序列,并呈现为SEQ ID NOs:5490、5492-5494和5538。Shmakov等人(Mol Cell.2015年11月5日;60(3):385-97.)所述的这些序列与Cas9的详细结构域级比对在图7中描绘。

[0679] MG2-1、MG2-2、MG2-3、MG2-5和MG2-6与另外的专有蛋白质数据集的比较揭示了具有相似结构的另外蛋白质序列,呈现为SEQNO:321和326-420。通常在MG2家族成员中发现的基序呈现为SEQ ID NO:5631-5638。

[0680] 实施例2C.-发现CRISPR系统的MG3家族

[0681] 对来自实施例1的宏基因组分析的数据的分析揭示了先前未描述的推定CRISPR系统:MG3-1的新簇。这些新酶及其示例性子结构域的相应氨基酸序列呈现为SEQ ID NO:424、2245和4059。基于与操纵子中其他元件的接近度,鉴定了推定的含有tracrRNA的序列并作为SEQ ID NO:5498包括在内。该序列与来自内氏放线菌的Cas9的详细结构域级比对在图8中描绘。

[0682] MG3-1与另外的专有蛋白质数据集的比较揭示了具有相似结构的另外蛋白质序列,呈现为SEQ NO:421-423、425-431。

[0683] 实施例2D.-发现CRISPR系统的MG4、7、14、15、16、18、21、22、23家族

[0684] 对来自实施例1的宏基因组分析的数据的分析揭示了先前未描述的推定CRISPR系统的新簇,所述系统包括9个家族,每个家族包括一名成员(MG 4-5、MG7-2、MG14-1、MG15-1、MG16-2、MG18-1、MG21-1、MG22-1、MG23-1)。这些新酶及其示例性子结构域的相应蛋白质和核酸序列呈现为SEQ ID NO:432、669、678、930、1093、1354、1512、1656、1756。基于与操纵子中其他元件的接近度,鉴定了每个家族的推定的含有tracr的序列。这些序列分别以SEQ ID NO:5503-5511提供在序列表中。

[0685] MG 4-5、MG7-2、MG14-1、MG15-1、MG16-2、MG18-1、MG21-1、MG22-1、MG23-1与另外的专有蛋白质数据集的比较揭示了具有相似结构的另外蛋白质序列,呈现为SEQ NO:433-660、670-677、679-929、931-1092、1094-1353、1355-1511、1513-1655、1657-1755和1757-1826。对于MG4,与这些CRISPR系统组的核酸酶共有的基序表示为SEQ ID NO:5649;对于MG14表示为SEQ ID NO:5650-5667;对于MG15表示为5668-5675;对于MG16表示为SEQ ID NO:5676-5678;对于MG18表示为SEQ ID NO:5679-5686;对于MG21表示为SEQ ID NO:5687-5693和SEQ ID NO:5674-5675;对于MG22表示为SEQ ID NO:5694-5699;对于MG23表示为SEQ ID NO:5700-5717。

[0686] 实施例3.-预示性——确定前间区序列邻近基序。

[0687] 实验按照Karvelis等人Methods.2017年5月15日;121-122:3-8(其全部内容通过引用并入本文)的任何实施例进行,以鉴定本文描述的新颖酶的前间区序列邻近基序(PAM)序列特异性,以允许最佳合成序列靶向。

[0688] 在一个实施例中(体内筛选),携带编码本文所述的任何酶和前间区序列靶向指导RNA的质粒的细胞用含有抗生素抗性基因和侧接随机化PAM序列的前间区序列的质粒文库共转化。含有功能性PAM的质粒被酶切割,导致细胞死亡。从存活细胞中分离出的抗酶切质粒池的深度测序显示了一组耗减的质粒,这些质粒含有功能性的允许切割的PAM。

[0689] 在另一个实施例(体外筛选)中,DNA质粒或串联重复形式的PAM文库受到体外或细胞裂解物内组装的RNP复合物(例如包括酶、tracrRNA和crRNA或酶和杂合sgRNA)的切割。由

成功切割事件产生的游离DNA末端由衔接子连接捕获，然后对PAM侧产物进行PCR扩增。对扩增的功能性PAM文库进行深度测序，并鉴定许可DNA切割的PAM。

[0690] 实施例4.-预示性——如本文所述的合成CRISPR系统在哺乳动物细胞中用于基因组编辑的用途

[0691] 制备了DNA/RNA序列，其编码(i)ORF，其在细胞相容性启动子下编码密码子优化的酶，具有细胞相容性C末端核定位序列(例如人细胞中的SV40 NLS)和合适的聚腺苷酸化信号(例如在人细胞的情况下TK pA信号)；和(ii)ORF，其在合适的聚合酶III启动子(例如哺乳动物细胞中的U6启动子)下编码sgRNA(具有以G开头的5'序列，随后是靶向基因组DNA的20nt互补靶向核酸序列，随后是通过实施例3鉴定的相应兼容PAM和3'tracr结合序列、接头和tracrRNA序列)。在一些实施方案中，这些序列在相同或不同的质粒载体上制备，通过合适的技术将质粒载体转染到真核细胞中。在一些实施方案中，这些序列被制备为分开的DNA序列，其被转染或显微注射到细胞中。在一些实施方案中，这些序列被制备为合成的RNA或体外转录的RNA，其被转染或显微注射到细胞中。在一些实施方案中，这些序列被翻译成蛋白质并被转染或显微注射到细胞中。

[0692] 无论选择哪种转染方法，(i)和(ii)被引入细胞。允许经过一段时间的孵育，以便酶和/或sgRNA可以转录和/或翻译成活性形式。孵育期后，分析(例如，通过测序)靶向序列附近的基因组DNA。作为酶介导的切割和非同源末端连接的结果，插入缺失标记被引入到靶向序列附近的基因组DNA中。

[0693] 在一些实施方案中，(i)和(ii)被引入具有第三修复核苷酸的细胞中，所述第三修复核苷酸编码位于切割位点侧翼的大小为25bp或更大的基因组区域，这将促进同源定向修复。包含在这些侧翼序列中的可能是单个碱基对突变、功能基因片段、用于表达的外源或天然基因，或组成生化途径的几个基因。

[0694] 实施例5.-预示性——如本文所述的合成CRISPR系统的体外用途

[0695] 将本文所述的任何酶克隆到含有纯化标签的合适的大肠杆菌表达质粒中，并在大肠杆菌中重组表达并使用重组标签进行纯化。通过合适的固相RNA合成方法合成RNA，所述RNA包含5'G后跟20nt靶向序列和PAM序列、兼容crRNA的tracrRNA结合区、GAAA接头和兼容的tracrRNA。重组酶和sgRNA在含有Mg2+(例如，20mM HEPES pH7.5、100mM KCl、5mM $MgCl_2$、1mM DTT、5％甘油)的合适裂解缓冲液中混合，反应通过引入包括与靶向序列和PAM序列互补的序列的靶DNA开始。DNA的切割通过合适的测定(例如，琼脂糖凝胶电泳，然后是溴化乙锭染色(或类似作用的DNA嵌入剂)和UV可视化)进行监测。

[0696] 实施例6.-(通用方案)本文描述的核酸内切酶的PAM序列鉴定/确认

[0697] PAM序列是通过测序质粒来确定的，所述质粒包含随机生成的PAM序列，这些序列可以被在基于大肠杆菌裂解物的表达系统(myTXTL，Arbor Biosciences)中表达的推定核酸内切酶切割。在该系统中，大肠杆菌密码子优化的核苷酸序列在T7启动子的控制下从PCR片段转录和翻译。在同一反应中转录具有T7启动子下的tracr序列和由T7启动子后跟重复-间隔子-重复序列组成的最小CRISPR阵列的第二PCR片段。在TXTL系统中成功表达核酸内切酶和tracr序列，然后进行CRISPR阵列处理，提供了活性体外CRISPR核酸酶复合物。

[0698] 包含与最小阵列中的间隔子序列匹配的间隔子序列随后是8N混合碱基(推定的PAM序列)的靶质粒的文库与TXTL反应的输出一起孵育。1-3小时后，停止反应并通过DNA净

化试剂盒(例如Zymo DCC、AMPure XP珠、QiaQuick等)回收DNA。将衔接子序列平端连接到具有已经被核酸内切酶切割的活性PAM序列的DNA上,而未切割的DNA无法连接。然后用对文库和衔接子序列特异的引物通过PCR扩增包含活性PAM序列的DNA片段。PCR扩增产物在凝胶上解析以鉴定与切割事件对应的扩增子。切割反应的扩增区段也被用作制备NGS文库的模板。对这个生成的文库(其是起始8N文库的一个子集)进行测序揭示了包含针对活性CRISPR复合物的正确PAM的序列。对于使用单个RNA构建体的PAM测试,重复相同的程序,不同之处在于将体外转录的RNA与质粒文库一起添加,并省略了tracr/最小CRISPR阵列模板。对于制备NGS文库情况下的核酸内切酶,seqLogo(参见例如Huber等人Nat Methods.2015年2月;12(2):115-21)表示被构建并显示在图27、图28、图29、图30、图31、图32、图33、图34和图35中。用于构建这些表示的seqLogo模块采用DNA序列基序(例如PAM序列)的位置权重矩阵,并绘制由Schneider和Stephens(参见例如Schneider等人Nucleic Acids Res.1990年10月25日;18(20):6097-100)引入的相应序列标志。在seqLogo表示中表示序列的字符已针对比对序列(例如PAM序列)中的每个位置堆叠在彼此的顶部。每个字母的高度与其出现的频率成正比,并且字母已经被分选,所以最常见的在最上面。

[0699] 实施例7.-(通用方案)tracrRNA和sgRNA结构的RNA折叠

[0700] 使用Andronescu等人Bioinformatics.2007年7月1日;23(13):i19-28(其全部内容通过引用并入)的方法计算了37℃下指导RNA序列的折叠结构。本文所述的示例性sgRNA的预测结构如图21、22、23、24、25和26所示。

[0701] 实施例8.-(通用方案)MG CRISPR复合物的体外切割效率

[0702] 在蛋白酶缺陷型大肠杆菌B菌株中,核酸内切酶从可诱导T7启动子表达为带有His标签的融合蛋白。表达带有His标签的蛋白的细胞通过超声裂解,带有His标签的蛋白通过Ni-NTA亲和层析在AKTA Avant FPLC(GE Lifescience)上的HisTrap FF柱(GE Lifescience)上纯化。洗脱液在丙烯酰胺凝胶(Bio-Rad)上通过SDS-PAGE解析并用InstantBlue Ultrafast考马斯亮蓝(Sigma-Aldrich)染色。使用ImageLab软件(Bio-Rad)对蛋白质条带进行密度测定来确定纯度。将纯化的核酸内切酶透析到由50mM Tris-HCl、300mM NaCl、1mM TCEP、5%甘油构成的储存缓冲液(pH7.5)中并储存于-80℃。

[0703] 通过DNA合成构建含有间隔子序列和PAM序列(例如如实施例6中确定的)的靶DNA。当PAM具有简并碱基时,选择具有代表性的单个PAM进行测试。靶DNA包含通过PCR扩增来自质粒的2200bp的线性DNA,其中PAM和间隔子距离一个末端700bp定位。成功的切割产生700和1500bp的片段。靶DNA、体外转录的单个RNA和纯化的重组蛋白在裂解缓冲液(10mM Tris、100mM NaCl、10mM $MgCl_2$)中与过量蛋白质和RNA混合并孵育5分钟至3小时,通常为1小时。通过加入RNAse A来终止反应并孵育60分钟。然后在1.2% TAE琼脂糖凝胶上解析反应,并在ImageLab软件中量化切割的靶DNA的分数。

[0704] 实施例9.-(通用方案)测试大肠杆菌中MG CRISPR复合物的基因组切割活性

[0705] 大肠杆菌缺乏有效修复双链DNA断裂的能力。因此,基因组DNA的切割可能是致命的事件。利用这一现象,通过在靶菌株中重组表达核酸内切酶和tracrRNA,在其基因组DNA中整合间隔子/靶和PAM序列,在大肠杆菌中测试了核酸内切酶活性。

[0706] 在该测定中,PAM序列对被测试的核酸内切酶具有特异性,如通过实施例6中所述的方法所确定的。基于tracrRNA的序列和预测结构确定sgRNA序列。从重复的5'末端开始,

选择8-12bp(通常为10bp)的重复-反重复配对。重复的剩余3'末端和tracrRNA的5'末端被替换为四环。通常,四环是GAAA,但也可以使用其他四环,特别是如果预测GAAA序列会干扰折叠。在这些情况下,使用了TTCG四环。

[0707] PAM序列整合到其基因组DNA中的工程改造的菌株用编码核酸内切酶的DNA进行转化。然后使转化体成为化学感受态,并用50ng的对靶序列特异("中靶")或对靶标非特异("非靶")的单指导RNA进行转化。热激后,在37℃下,在SOC中恢复转化2小时。然后通过在诱导培养基上生长的5倍稀释系列确定核酸酶效率。从稀释系列一式三份量化菌落。

[0708] 实施例10.-(通用方案)测试哺乳动物细胞中MG CRISPR复合物的基因组切割活性

[0709] 为了显示哺乳动物细胞中的靶向和切割活性,MG Cas效应物蛋白序列在两种哺乳动物表达载体中进行了测试:(a)一种带有C末端SV40NLS和2A-GFP标签,和(b)一种没有GFP标签,带有两个SV40 NLS序列,一个在N末端,一个在C末端。在一些情况下,编码核酸内切酶的核苷酸序列经过密码子优化以在哺乳动物细胞中表达。

[0710] 将带有靶向序列的相应单指导RNA序列(sgRNA)克隆到第二哺乳动物表达载体中。将这两种质粒共转染到HEK293T细胞中。将表达质粒和sgRNA靶向质粒共转染到HEK293T细胞中后72小时,提取DNA并用于制备NGS文库。NHEJ百分比是通过靶位点测序中的插入缺失标记来测量的,以证明该酶在哺乳动物细胞中的靶向效率。至少选择了10个不同的靶位点来测试每种蛋白质的活性。

[0711] 实施例11.-MG1家族成员的表征

[0712] PAM特异性,tracrRNA/sgRNA验证

[0713] 使用如实施例6中所述的myTXTL系统证实MG1家族核酸内切酶系统的靶向核酸内切酶活性。在该测定中,切割的靶质粒的PCR扩增产生在凝胶中迁移约170bp的产物,如图17-20所示。观察到MG1-4的扩增产物(双指导物:见凝胶1泳道3,单指导物:见凝胶6泳道2)、MG1-5的扩增产物(凝胶2泳道10)、MG1-6的扩增产物(双指导物:见凝胶5泳道6,单指导物:见凝胶6泳道5)和MG1-7的扩增产物(双指导物:见凝胶3泳道13,单指导物:见凝胶3泳道2)(分别为蛋白质SEQ ID NO:1-4)。对PCR产物的测序揭示了如表2所示的这些酶的活性PAM序列。

[0714] 表2:MG1酶的PAM序列特异性和相关数据

[0715]

酶	酶蛋白SEQ ID NO	天然的(双指导)PAM	PAM SEQ ID NO:	tracrRNA SEQ ID NO:	sgRNA SEQ ID NO:	合成的(单指导)PAM	合成的(单指导)PAM SEQ ID NO:
MG1-4	1	nRRRAA	5527	5476	5461	nRRR	5512
MG1-5	2	nnnnCC	5528	5477	5462	nnnnYY	5513
MG1-6	3	nnRRWC	5529	5478	5463	nnRRAY	5514

[0716]

MG1-7	4	nRRRAA	5530	5479	5464	nRRRAAG	5515

[0717] 基于tracrRNA的序列和预测结构设计合成的单指导RNA(sgRNA),并表示为SEQ ID NO:5461-5464。用sgRNA重复实施例6的PAM序列筛选。该实验的结果也列在表2中,揭示使用sgRNA时PAM特异性略有变化。

[0718] 体外靶向核酸内切酶活性

[0719] 使用实施例8的方法验证MG1-4核酸内切酶系统(蛋白质SEQ ID NO:1与sgRNA SEQ ID NO:5461)对具有PAM序列CAGGAAGG的靶DNA的体外活性。使用具有18-24nt的不同间隔子/靶向序列长度(替换序列的N)的上面报道的单指导序列(SEQ ID NO:5461)。结果如图10所示,其中左分图显示证明通过MG1-4结合具有不同靶向序列长度(18-24nt)的相应单指导sgRNA进行DNA切割的凝胶,右分图显示量化为条形图的相同数据。数据表明18-24个核苷酸的靶向序列与MG1-4/sgRNA系统一起起作用。

[0720] 细菌细胞中的靶向核酸内切酶活性

[0721] 使用如实施例9的PAM序列CAGGAAGG测试MG1-4核酸内切酶系统(蛋白质SEQ ID NO:1,sgRNA SEQ ID NO:5461)的体内活性。转化的大肠杆菌以系列稀释的方式铺板(plate),结果(左分图显示大肠杆菌系列稀释,右分图显示定量生长)如图11所示。与表达非靶sgRNA的大肠杆菌相比,表达中靶sgRNA的大肠杆菌的生长显著降低表明在大肠杆菌细胞中基因组DNA被核酸内切酶特异性切割。

[0722] 哺乳动物细胞中的靶向核酸内切酶活性

[0723] 实施例10的方法用于证明哺乳动物细胞中的靶向和切割活性。编码MG1-4(蛋白质SEQ ID NO:5527)和MG1-6(蛋白质SEQ ID NO:5529)序列的开放阅读框被克隆到2种哺乳动物表达载体中,一种具有C末端SV40 NLS和2A-GFP标签(大肠杆菌MG-BB),一种不具有GFP标签,具有2个NLS序列,一个在N端,一个在C端(大肠杆菌pMG5-BB)。对于MG1-6,开放阅读框还针对哺乳动物表达进行了密码子优化(SEQ ID NO:5589)并被克隆到2-NLS质粒骨架(MG-16hs)中。该实验的结果显示在图12中。将核酸内切酶表达载体与用于表达具有对所述核酸内切酶特异性的tracr序列和选自表3-4的指导sgRNA的sgRNA(例如,SEQ ID NO:5512或5515)的第二载体共转染到HEK293T细胞中。共转染后72小时,提取DNA并用于制备NGS文库。通过接近靶位点序列的内部缺失(NHEJ残余)的出现来检测切割活性。NHEJ百分比是通过靶位点测序中的插入缺失标记测量的,以证明该酶在哺乳动物细胞中的靶向效率,并示于图12中。

[0724] 表3:MG1-4哺乳动物靶向序列

[0725]

MG1-4靶ID	MG1-4靶向序列	靶向序列SEQ ID NO:	靶基因
1	aatatgtagctgtttgggaggt	5543	VEGFA
2	ctaggggcgctcggccaccac	5544	VEGFA
3	tggctaaagagggaatgggctt	5545	VEGFA
4	cacaccccggctctggctaaag	5546	VEGFA
5	tcggaggagccgtggtccgcgc	5547	VEGFA
6	gcggaccacggctcctccgaag	5548	VEGFA
7	gtacaaacggcagaagctggag	5549	EMX1
8	gaggaagggcctgagtccgagca	5550	EMX1
9	aaggcaaacatcctgataatgg	5551	载脂蛋白

[0726] 表4:MG1-6哺乳动物靶向序列

[0727]

MG1-6靶ID	MG1-6靶向序列	靶向序列SEQ ID NO:	靶基因
1	tctttagccagagccggggtgt	5552	VEGFA
2	tggacccctatttctgacctc	5553	VEGFA
3	atgggagcccttcttcttctgc	5554	EMX1
4	tgccacgaagcaggccaatggg	5555	EMX1
5	tggtgtctgtttgaggttgcta	5556	HBB-R01
6	gggcaggttggtatcaaggtta	5557	HBB-R01
7	aggtgctgacgtaggtagtgct	5558	FANCF
8	gccctacttccgctttcacctt	5559	FANCF
9	aatgtatgctggcttttaaggg	5560	IVS40
10	gctcctttggctagggaagtgt	5561	IVS40

[0728] 实施例12.-MG2家族成员的表征

[0729] PAM特异性,tracrRNA/sgRNA验证

[0730] 在如实施例6中所述的myTXTL系统中证实了MG2家族成员的靶向核酸内切酶活性。该测定的结果显示在图17-20中。在图17-20所示的测定中,成功切割文库的活性蛋白质会在凝胶中产生一条约170bp的条带。观察到MG2-1的扩增产物(见凝胶2泳道11和凝胶4泳道6)和MG2-7(见凝胶11泳道10)的扩增产物(分别为SEQ ID NO:320和321)。对PCR产物测序揭示了下表5中的活性PAM序列:

[0731] 表5:MG2酶的PAM序列特异性和相关数据

[0732]

酶	酶蛋白SEQ ID NO	天然的(双指导)PAM	PAM SEQ ID NO:	tracrRNA SEQ ID NO:	sgRNA SEQ ID NO:	合成的(单指导)PAM	合成的(单指导)PAM SEQ ID NO:
MG2-1	320	nRCGTA	5531	5490	N/A	N/A	N/A
MG2-7	321	N/A	N/A	5491	5465	NNNRTA	5516

[0733] 细菌细胞中的靶向核酸内切酶活性

[0734] 使用实施例9中描述的方法证实MG2-7核酸内切酶系统与sgRNA(所述核酸内切酶SEQ ID NO:321;sgRNA SEQ ID NO:5465)和AGCGTAAG PAM序列的体内活性。转化的大肠杆菌以系列稀释的方式铺板,结果(在左分图中显示了大肠杆菌系列稀释,在右分图中显示了定量的生长)显示于图34中。与表达非靶sgRNA的大肠杆菌相比,表达中靶sgRNA的大肠杆菌的生成显著降低表明大肠杆菌细胞中基因组DNA被MG1-4核酸内切酶特异性切割。

[0735] 实施例13.-MG3家族成员的表征

[0736] PAM特异性,tracrRNA/sgRNA验证

[0737] 使用如实施例6中所述的myTXTL系统使用tracr序列和CRISPR阵列证实MG3家族成员的靶向核酸内切酶活性。在该测定中,切割的靶质粒的PCR扩增产生在凝胶中迁移约170bp的产物,如图17-20所示。观察到MG3-6的扩增产物(双指导物:见凝胶2泳道8,单指导物:见凝胶3泳道3)、MG3-7的扩增产物(双指导物:见凝胶2泳道3,单指导物:见凝胶3泳道4)和MG3-8的扩增产物(双指导物:见凝胶9泳道5)(分别为蛋白质SEQ ID NO:421、422和423)。对PCR产物的测序揭示了如下表6中的活性PAM序列:

[0738] 表6:MG3酶的PAM序列特异性和相关数据

[0739]

酶	酶蛋白SEQ ID NO	天然的(双指导)PAM	PAM SEQ ID NO:	tracrRNA SEQ ID NO:	sgRNA SEQ ID NO:	合成的(单指导)PAM	合成的(单指导)PAM SEQ ID NO:
MG3-6	421	nnRGGTT	5532	5500	5466	nnGGG	5517
MG3-7	422	nnRnYAY	5533	5501	5467	nnGnTnT	5518
MG3-8	423	nnRGGTT	5534	5502	N/A	N/A	N/A

[0740] 基于tracrRNA的序列和预测结构设计合成的单指导RNA(sgRNA),并显示为SEQ ID NO:5466-5467。用sgRNA重复实施例6的PAM序列筛选。该实验的结果也示于表6中,表明使用sgRNA时PAM特异性略有变化。

[0741] 体外靶向核酸内切酶活性

[0742] 使用实施例8的方法用PAM序列GTGGGTTA验证MG3-6(核酸内切酶SEQ ID NO:421)的体外活性。使用具有18-24nt不同间隔子/靶向序列长度(替换序列的N)的上面报道的单指导序列(SEQ ID NO:5466)。结果如图13所示,其中上分图显示证明通过MG3-6结合具有不同靶向序列长度(18-24nt)的不同sgRNA进行DNA切割的凝胶,下分图显示量化为条形图的相同数据。数据表明18-24个核苷酸的靶向序列与MG3-6/sgRNA系统一起起作用。

[0743] 细菌细胞中的靶向核酸内切酶活性

[0744] 使用实施例9的方法用PAM序列TGGACCTG测试MG3-7核酸内切酶系统(蛋白质SEQ ID NO:422;sgRNA SEQ ID NO:5467)的体内活性。转化的大肠杆菌以系列稀释的方式铺板,结果(在上分图中显示大肠杆菌系列稀释,在下分图中显示定量生长)如图14所示。与表达非靶sgRNA的大肠杆菌相比,表达中靶sgRNA的大肠杆菌的生长显著降低揭示了基因组DNA

被MG3-7核酸内切酶系统特异性切割。

[0745] 哺乳动物细胞中的靶向核酸内切酶活性

[0746] 实施例10的方法用于证明哺乳动物细胞中的靶向和切割活性。将编码MG3-7(蛋白质SEQ ID NO:422)的开放阅读框克隆到2种哺乳动物表达载体中,一种具有C末端SV40 NLS和2A-GFP标签(大肠杆菌MG-BB),另一种不含GFP标签并含2个NLS列,一个在N末端,一个在C末端(大肠杆菌pMG5-BB)。将核酸内切酶表达载体与用于表达具有选自表7的指导序列的上述sgRNA的第二载体共转染至HEK293T细胞中。该实验的结果显示于图12中。共转染后72小时提取DNA并用于制备NGS文库。通过靶位点附近的内部缺失(NHEJ残余)的出现检测切割活性。结果示于图15中。

[0747] 在sgRNA质粒上编码的靶位点显示在下表7中。

[0748] 表7:MG3-7哺乳动物靶向序列

[0749]

MG3-7 靶 ID	MG3-7 靶向序列	靶向序列 SEQ ID NO:	靶基因
1	cccctatttctgacctcccaaa	5563	VEGFA
2	tgtggttccagaaccggaggac	5564	EMX1
3	ggccctgggcaggttggtatca	5565	HBB-R01
4	tccttaaacctgtcttgtaacc	5566	HBB-R01
5	ctgactcctgaggagaagtctg	5567	HBB-R01
6	tccgagcttctggcggtctcaa	5568	FANCF
7	tatcatttcgcggatgttccaa	5569	FANCF
8	tcgggcagagggtgcatcacct	5570	载脂蛋白
9	ataataagcagaacttttagtg	5571	纤维蛋白原
10	gttttctttagttattaatttc	5572	纤维蛋白原

[0750] 实施例13.-MG4家族成员的表征

[0751] PAM特异性,tracrRNA/sgRNA验证

[0752] 使用如实施例6中所述的myTXTL系统证实MG4家族核酸内切酶系统的靶向核酸内切酶活性。在该测定中,切割的靶质粒的PCR扩增产生在凝胶中迁移约170bp的产物,如图17-20所示。观察到MG4-2的扩增产物(双指导物:见凝胶2泳道9,单指导物:见凝胶10泳道7)(SEQ ID NO:432)。对PCR产物的测序揭示了如下表8中的活性PAM序列。

[0753] 表8:MG4酶的PAM序列特异性和相关数据

[0754]

酶	酶蛋白 SEQ ID NO	tracrRNA SEQ ID NO:	sgRNA SEQ ID NO:	合成的(单指导)PAM	合成的(单指导)PAM SEQ ID NO:
MG4-5	432	5503	5468	nnCCR	5519

[0755] 实施例14.-MG14家族成员的表征

[0756] PAM特异性,tracrRNA/sgRNA验证

[0757] 使用如实施例6中所述的myTXTL系统证实MG14家族成员的靶向核酸内切酶活性。在该测定中,切割的靶质粒的PCR扩增产生在凝胶中迁移约170bp的产物,如图17-20所示。观察到MG14-1的扩增产物(双指导物:见凝胶1泳道4,单指导物:见凝胶3泳道8)(SEQ ID NO:678)。对PCR产物测序揭示了下表9中所示的活性PAM序列特异性。

[0758]

酶	酶蛋白SEQ ID NO	确定的天然的(双指导)PAM	PAM SEQ ID NO:	tracrRNA SEQ ID NO:	sgRNA SEQ ID NO:	确定的合成的(单指导)PAM	合成的(单指导)PAM SEQ ID NO:
MG14-1	678	NNNNGGTA	5535	5505	5469	NNNGGRTA	5520

[0759] 细菌细胞中的靶向核酸内切酶活性

[0760] 使用实施例9中描述的方法证实MG14-1核酸内切酶系统与sgRNA(所述核酸内切酶SEQ ID NO:678;sgRNA SEQ ID NO:5469)和GGCGGGGA PAM序列的体内活性。转化的大肠杆菌以系列稀释的方式铺板,结果(在左分图中显示了大肠杆菌系列稀释,在右分图中显示了定量的生长)示于图35中。与表达非靶sgRNA的大肠杆菌相比,表达中靶sgRNA的大肠杆菌的生长显著降低表明在大肠杆菌细胞中基因组DNA被MG1-4核酸内切酶特异性切割。

[0761] 实施例15.-MG15家族成员的表征

[0762] PAM特异性,tracrRNA/sgRNA验证

[0763] 使用如实施例6中所述的myTXTL系统证实MG15家族成员的靶向核酸内切酶活性。在该测定中,切割的靶质粒的PCR扩增产生在凝胶中迁移约170bp的产物,如图17-20所示。观察到MG15-1的扩增产物(双指导物:见凝胶7泳道7,单指导物:见凝胶3泳道9)(SEQ ID NO:930)。对PCR产物测序揭示了下表10中详述的活性PAM序列特异性。

[0764] 表10:

[0765]

酶	酶蛋白SEQ ID NO	天然的(双指导)PAM	PAM SEQ ID NO:	tracrRNA SEQ ID NO:	sgRNA SEQ ID NO:	合成的(单指导)PAM	合成的(单指导)PAM SEQ ID NO:
MG15-1	930	nnnnC	5536	5506	5470	nnnnC	5521

[0766] 体外活性

[0767] 使用实施例8的方法用PAM序列GGGTCAAA测试MG15-1核酸内切酶系统(蛋白质SEQ ID NO:930;sgRNA SEQ ID NO:5470)的体外活性。使用以上报道的具有18-24nt不同间隔子/靶向序列长度(替换序列的N)的单指导序列(SEQ ID NO:5470)。结果如图16所示,其中上分图显示证明通过MG15-1结合具有不同靶向序列长度(18-24nt)的不同sgRNA进行DNA切

割的凝胶，下分图显示量化为条形图的相同数据。数据表明18-24个核苷酸的靶向序列与MG15-1/sgRNA系统一起起作用。

[0768] 细菌细胞中的靶向核酸内切酶活性

[0769] 使用实施例9中描述的方法证实MG15-1核酸内切酶系统与sgRNA(核酸内切酶SEQ ID NO:930;sgRNA SEQ ID NO:5470)和GGGTCAAA PAM序列的体内活性。转化的大肠杆菌以系列稀释的方式铺板，结果(在左分图中显示了大肠杆菌系列稀释，在右分图中显示了定量的生长)示于图35中。与表达非靶sgRNA的大肠杆菌相比，表达中靶sgRNA的大肠杆菌的生长显著降低表明在大肠杆菌细胞中基因组DNA被MG1-4核酸内切酶特异性切割。

[0770] 实施例16.-MG16家族成员的表征

[0771] PAM特异性，tracrRNA/sgRNA验证

[0772] 使用如实施例6中所述的myTXTL系统证实MG16家族成员的靶向核酸内切酶活性。在该测定中，切割的靶质粒的PCR扩增产生在凝胶中迁移约170bp的产物，如图17-20所示。观察到MG16-2的扩增产物(见凝胶11，泳道17)(SEQ ID NO:1093)。对PCR产物测序揭示了下表11中详述的活性PAM序列特异性。

[0773] 表11：

[0774]

酶	酶蛋白SEQ ID NO	sgRNA SEQ ID NO:	合成的(单指导)PAM	合成的(单指导)PAM SEQ ID NO:
MG16-2	1093	5471	nRTnCC	5522

[0775] 实施例17.-MG18家族成员的表征

[0776] PAM特异性，tracrRNA/sgRNA验证

[0777] 使用如实施例6中所述的myTXTL系统证实MG18家族成员的靶向核酸内切酶活性。在该测定中，切割的靶质粒的PCR扩增产生在凝胶中迁移约170bp的产物，如图17-20所示。观察到MG18-1的扩增产物(双指导物：见凝胶9泳道9，单指导物：见凝胶11泳道12)(SEQ ID NO:1354)。对PCR产物测序揭示了下表12中详述的活性PAM序列特异性。

[0778] 表12：

[0779]

酶	酶蛋白SEQ ID NO	天然的(双指导)PAM	PAM SEQ ID NO:	tracrRNA SEQ ID NO:	sgRNA SEQ ID NO:	合成的(单指导)PAM	合成的(单指导)PAM SEQ ID NO:
MG18-1	1354	nRWART	5537	5508	5472	nnnRRT	5523

[0780] 实施例18.-MG21家族成员的表征

[0781] PAM特异性，tracrRNA/sgRNA验证

[0782] 使用如实施例6中所述的myTXTL系统证实MG21家族的靶向核酸内切酶活性。在该测定中，切割的靶质粒的PCR扩增产生在凝胶中迁移约170bp的产物，如图17-20所示。观察到MG21-1的扩增产物(见凝胶11泳道2)(SEQ ID NO:1512)。对PCR产物测序揭示了下表13中

详述的活性PAM序列特异性。

[0783] 表13：

[0784]

酶	酶蛋白SEQ ID NO	sgRNA SEQ ID NO:	合成的（单指导）PAM	合成的（单指导）PAM SEQ ID NO:
MG21-1	1512	5473	nnRnR	5524

[0785] 实施例19.-MG22家族成员的表征

[0786] PAM特异性，tracrRNA/sgRNA验证

[0787] 使用如实施例6中所述的myTXTL系统证实MG22家族成员的靶向核酸内切酶活性。在该测定中，切割的靶质粒的PCR扩增产生在凝胶中迁移约170bp的产物，如图17-20所示。在图17-20所示的测定中，成功切割文库的活性蛋白质在凝胶中产生约170bp的条带。观察到MG22-1的扩增产物(见凝胶11泳道3)(蛋白质SEQ ID NO:1656)。对PCR产物测序揭示了下表14中详述的活性PAM序列特异性。

[0788] 表14：

[0789]

酶	酶蛋白SEQ ID NO	确定的天然的（双指导）PAM	PAM SEQ ID NO:	tracrRNA SEQ ID NO:	sgRNA SEQ ID NO:	确定的合成的（单指导）PAM	合成的（单指导）PAM SEQ ID NO:
MG22-1	1656	N/A	N/A	5510	5474	nnRCnT	5525

[0790] 实施例20.-MG23家族成员的表征

[0791] PAM特异性，tracrRNA/sgRNA验证

[0792] 使用如实施例6中所述的myTXTL系统证实MG23家族成员的靶向核酸内切酶活性。在该测定中，切割的靶质粒的PCR扩增产生在凝胶中迁移约170bp的产物，如图17-20所示。观察到MG23-1的扩增产物(见凝胶11泳道4)(SEQ ID NO:1756)。对PCR产物测序揭示了这些酶的活性PAM序列特异性，详述如下表15。

[0793] 表15：

[0794]

酶	酶蛋白SEQ ID NO	tracrRNA SEQ ID NO:	sgRNA SEQ ID NO:	合成的（单指导）PAM	合成的（单指导）PAM SEQ ID NO:
MG23-1	1756	5511	5475	nRRA	5526

[0795] 本公开的系统可用于各种应用，例如核酸编辑(例如，基因编辑)、与核酸分子的结合(例如，序列特异性结合)。例如，此类系统可用于解决(例如，去除或替换)可能导致受试者疾病的遗传突变；灭活基因以确定其在细胞中的功能；作为检测致病基因元件的诊断工具(例如通过逆转录病毒RNA或编码致病突变的扩增DNA序列的切割)；作为失活酶与探针结合以靶向和检测特定核苷酸序列(例如细菌中编码抗生素抗性的序列)；通过靶向病毒基因组使病毒失活或无法感染宿主细胞；添加基因或修改代谢途径以工程改造生物体以产生有价值的小分子、大分子或次级代谢物；建立基因驱动元件用于进化选择；作为生物传感器检测由外来小分子和核苷酸引起的细胞扰动。

[0796] 虽然已经示出并在本文所述了本发明的优选实施方案，对本领域技术人员将是显而易见的是，仅通过举例的方式来提供这样的实施方案。并不旨在将本发明限制为说明书中提供的具体实施例。虽然参考前述说明书已经描述了本发明，但是本文的实施方案的描述和说明并不意味着以限制性含义来解释。在不脱离本发明的情况下，本领域技术人员现在将清楚许多变型、改变和替代。此外，应当理解，本发明的所有方面不限于本文所述的具体描述、配置或相对比例，这取决于多种条件和变量。应当理解，在实践本发明时可以采用本文所述的本发明实施方案的各种替代方案。因此，预期本发明还应涵盖任何此类替代方案、修改、变型或等同物。所附权利要求旨在限定本发明内容的范围，并且由此覆盖这些权利要求的范围内的方法和结构及其等同物。

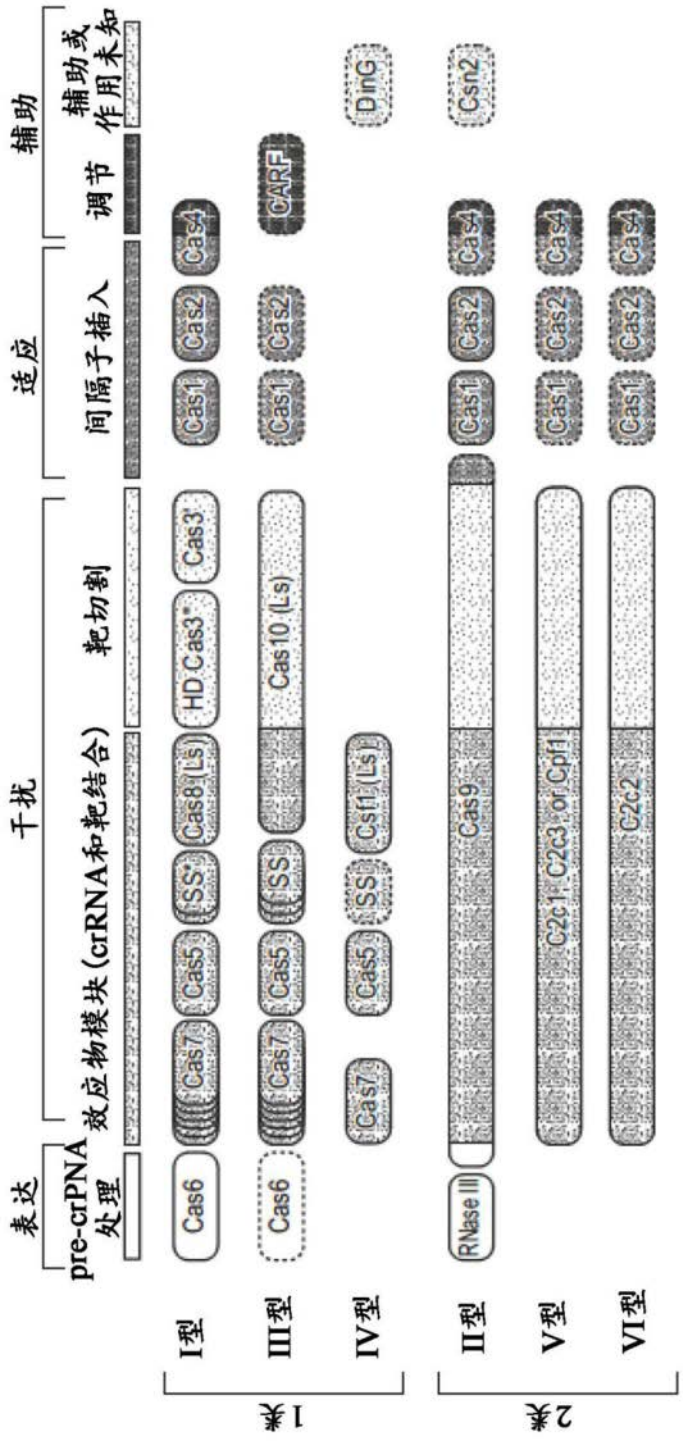

表达
干扰
适应
辅助
pre-crRNA处理
效应物模块（crRNA和靶结合）
靶切割
间隔子插入
调节
辅助或未知作用
I型
III型
IV型
II型
V型
VI型
1类
2类
Cas6
Cas7
Cas5
SS
Cas8 (Ls)
HD Cas3*
Cas3'
Cas10 (Ls)
Csf1 (Ls)
RNase III
Cas9
C2c1, C2c3, or Cpf1
C2c2
Cas1
Cas2
Cas4
CARF
DinG
Csn2

图1

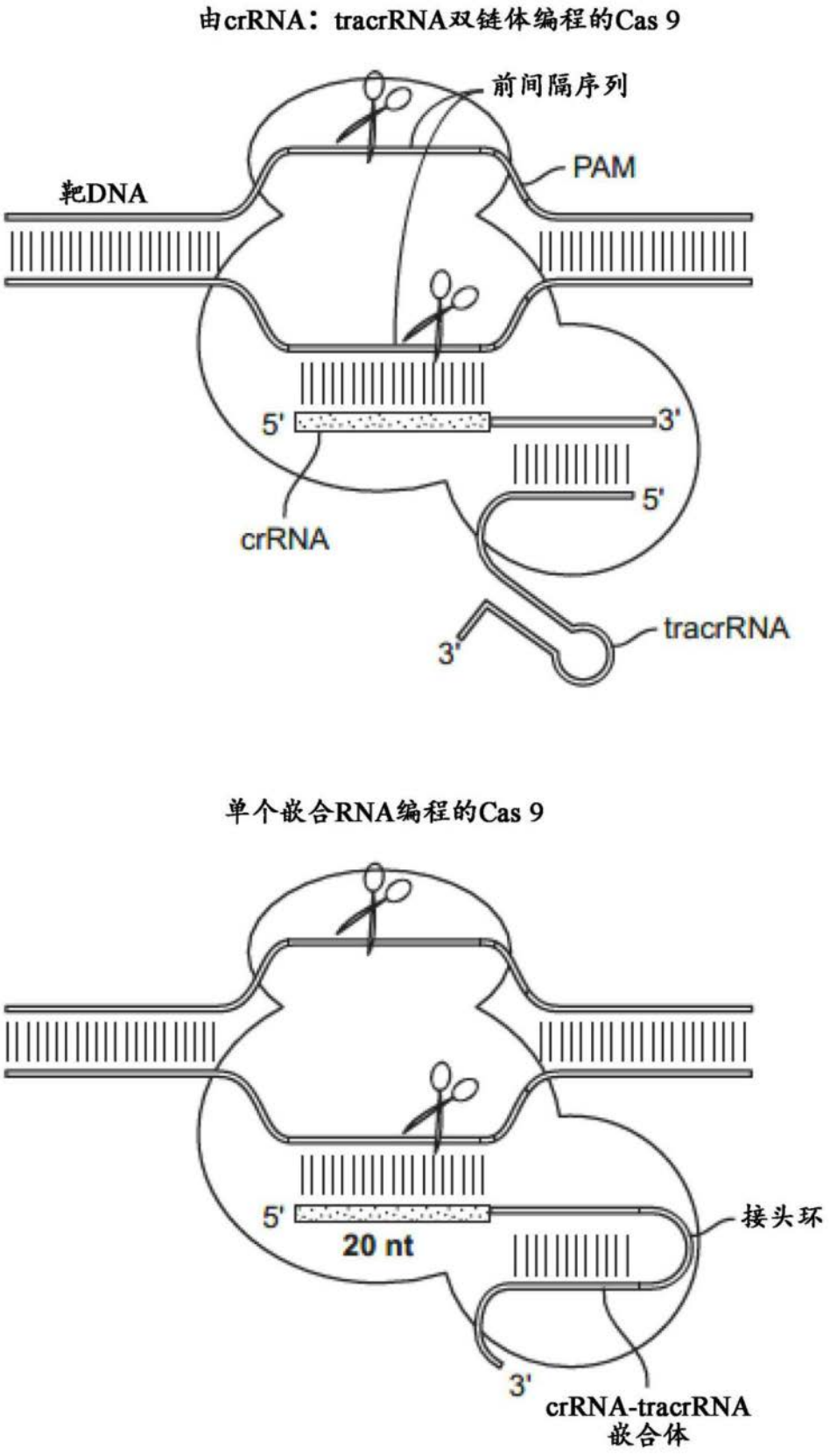

由crRNA：tracrRNA双链体编程的Cas 9
前间隔序列
PAM
靶DNA
5'
3'
5'
crRNA
3'
tracrRNA
单个嵌合RNA编程的Cas 9
5'
20 nt
接头环
3'
crRNA-tracrRNA
嵌合体

图2

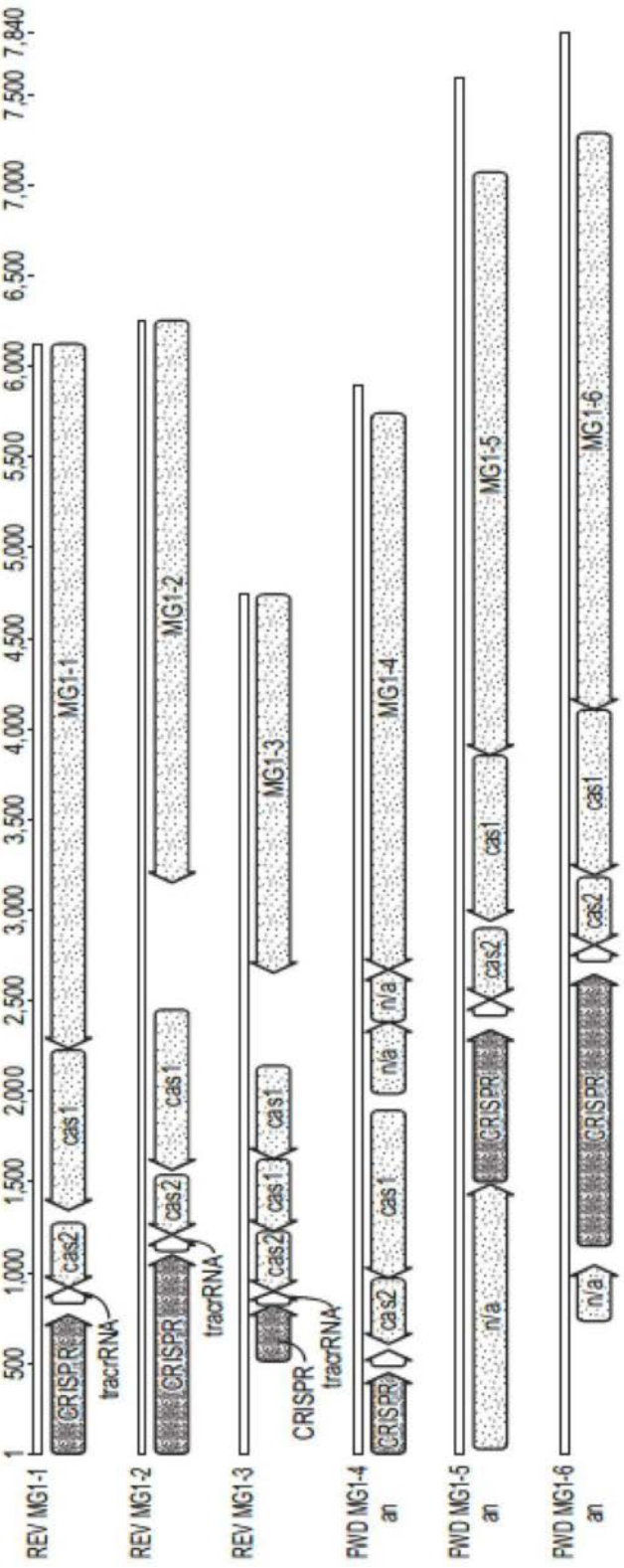

1
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
5,500
6,000
6,500
7,000
7,500
7,840
REV MG1-1
REV MG1-2
REV MG1-3
FWD MG1-4
an
FWD MG1-5
an
FWD MG1-6
an
CRISPR
cas2
cas1
tracrRNA
MG1-1
MG1-2
MG1-3
MG1-4
MG1-5
MG1-6
n/a

图3

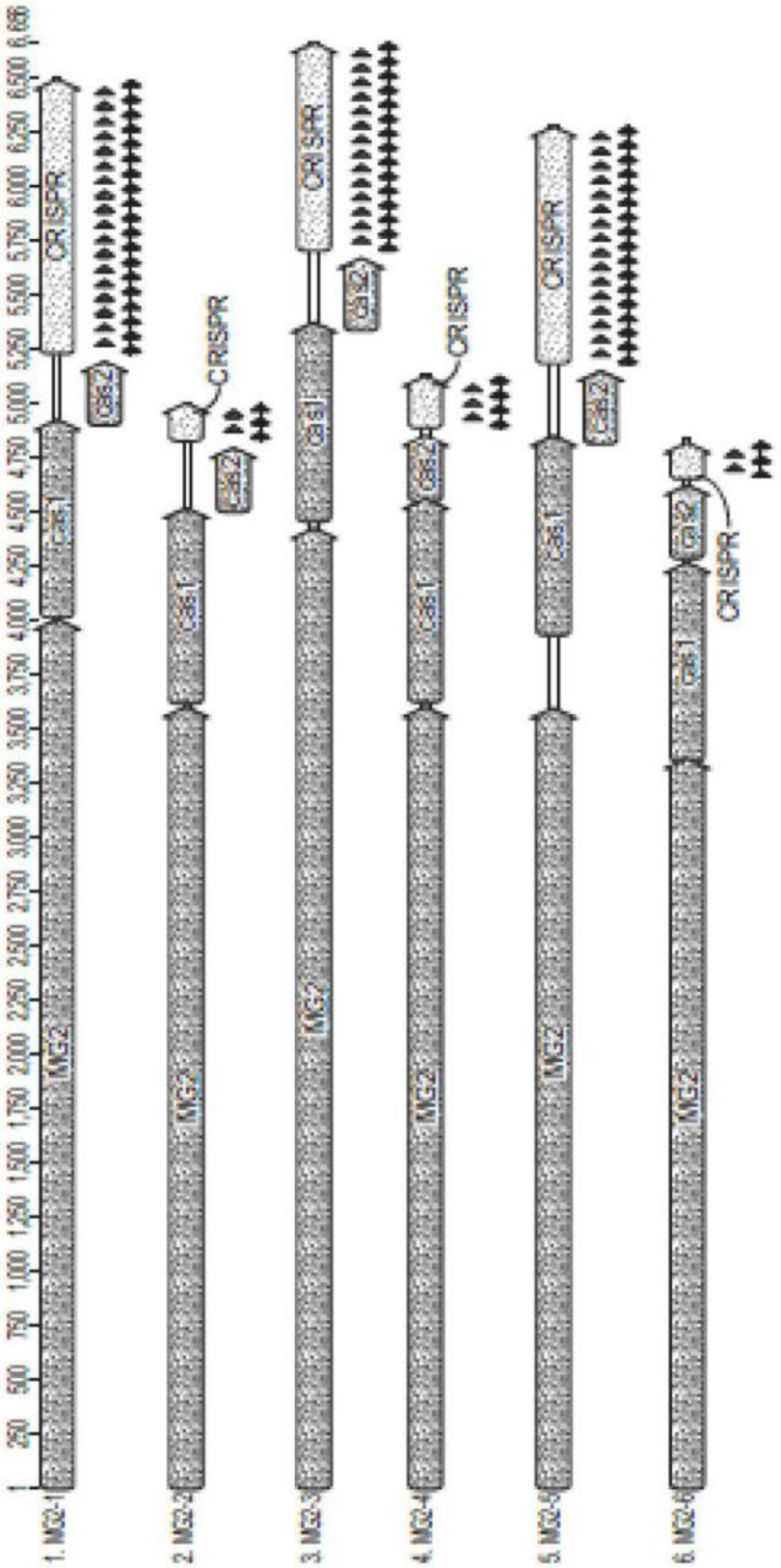

1 250 500 750 1,000 1,250 1,500 1,750 2,000 2,250 2,500 2,750 3,000 3,250 3,500 3,750 4,000 4,250 4,500 4,750 5,000 5,250 5,500 5,750 6,000 6,250 6,500 6,686
CRISPR
cas2
cas1
MG2
1. MG2-1
2. MG2-2
3. MG2-3
4. MG2-4
5. MG2-5
6. MG2-6

图4

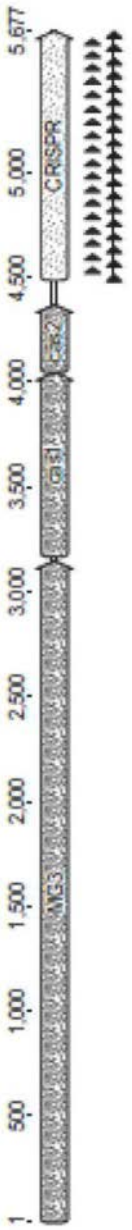

5,677
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
1
CRISPR
ES2
ORF1
XM3

图5

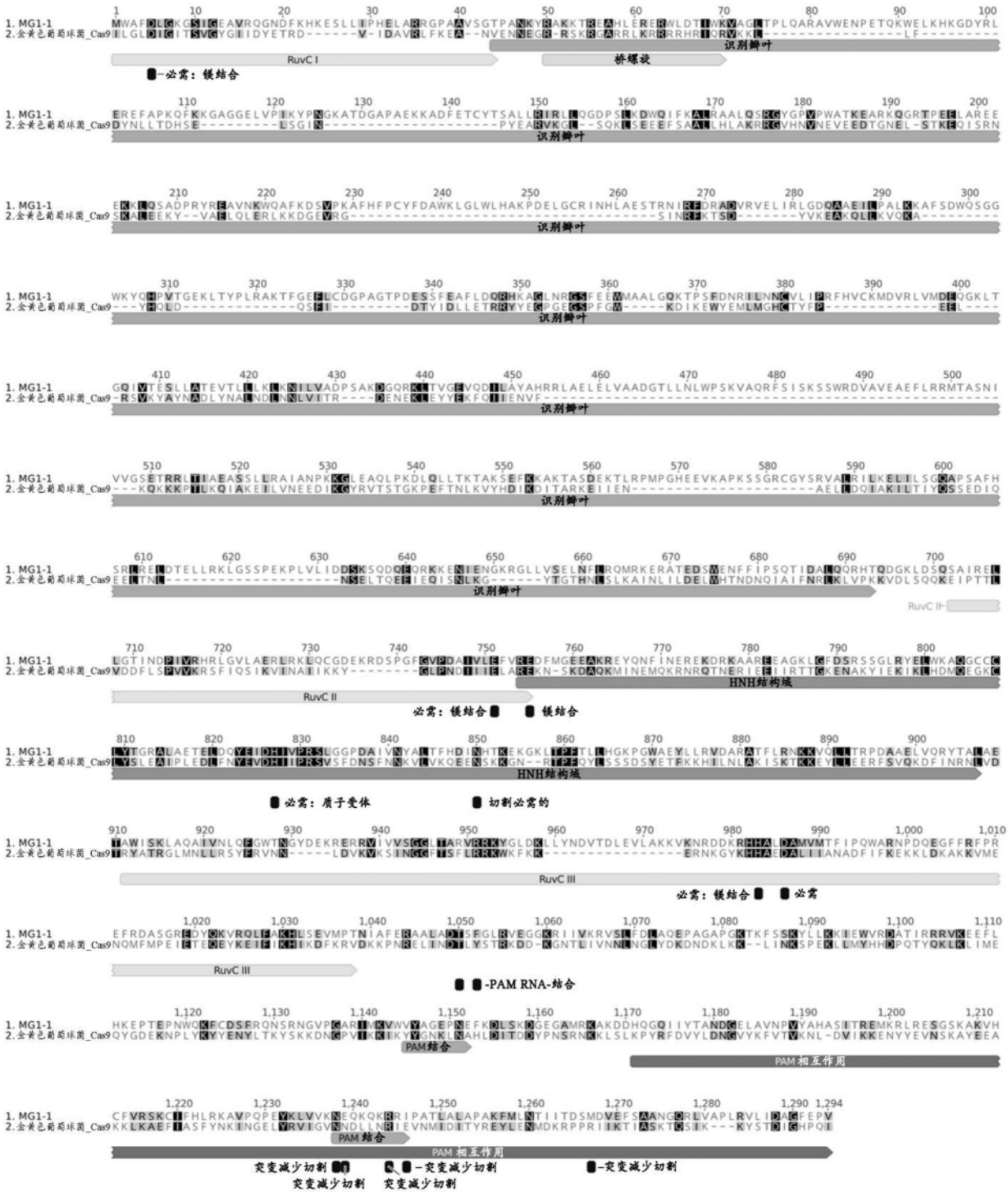

1. MG1-1
2.金黄色葡萄球菌_Cas9
RuvC I
桥螺旋
识别鳞叶
-必需：镁结合
RuvC II
HNH结构域
必需：镁结合
镁结合
RuvC III
-PAM RNA-结合
PAM结合
PAM 相互作用
必需：质子受体
切割必需的
突变减少切割
-突变减少切割

图6

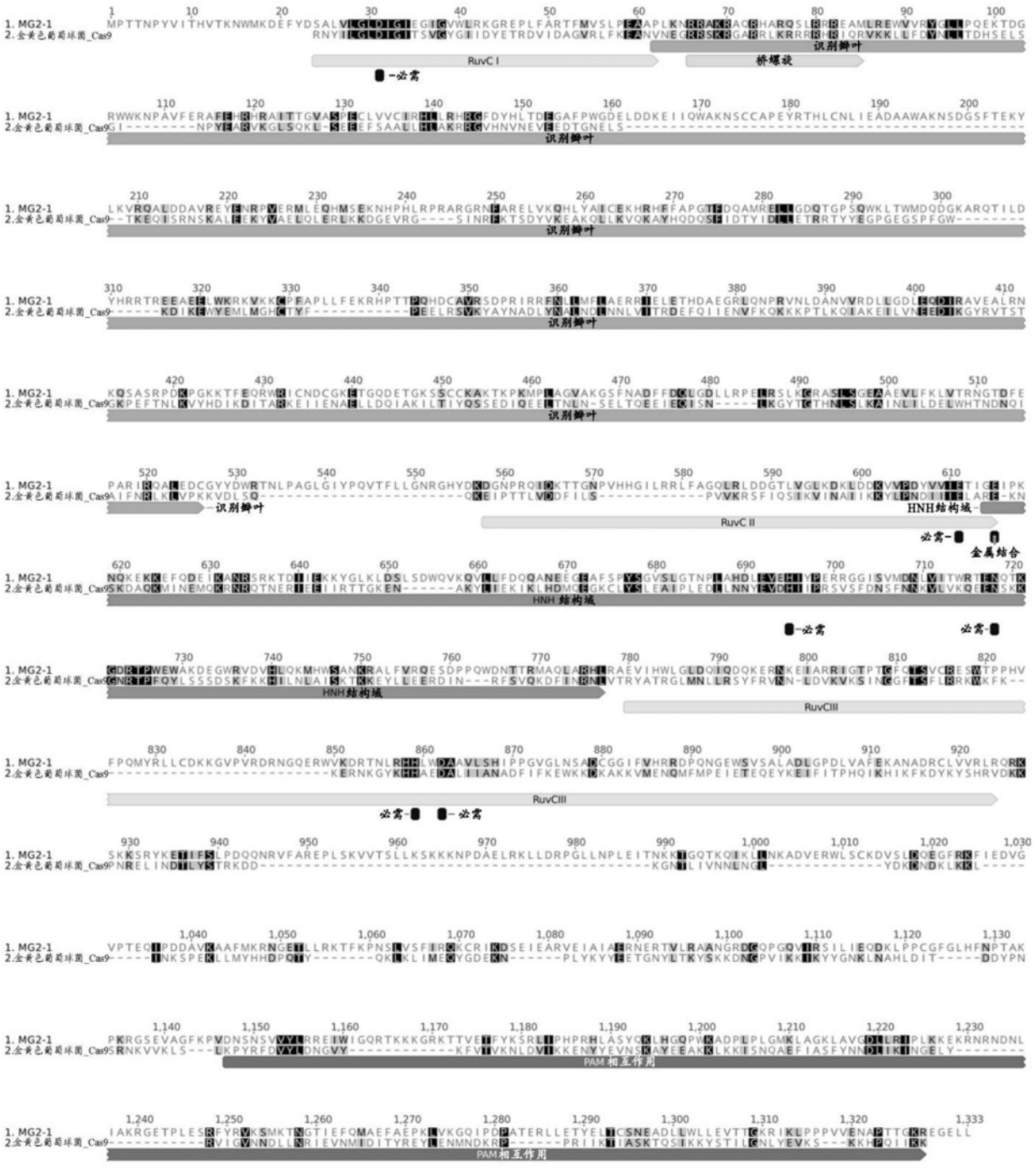

1. MG2-1
2.金黄色葡萄球菌_Cas9
识别鳍叶
RuvC I
桥螺旋
-必需
RuvC II
HNH结构域
金属结合
RuvCIII
PAM相互作用

图7

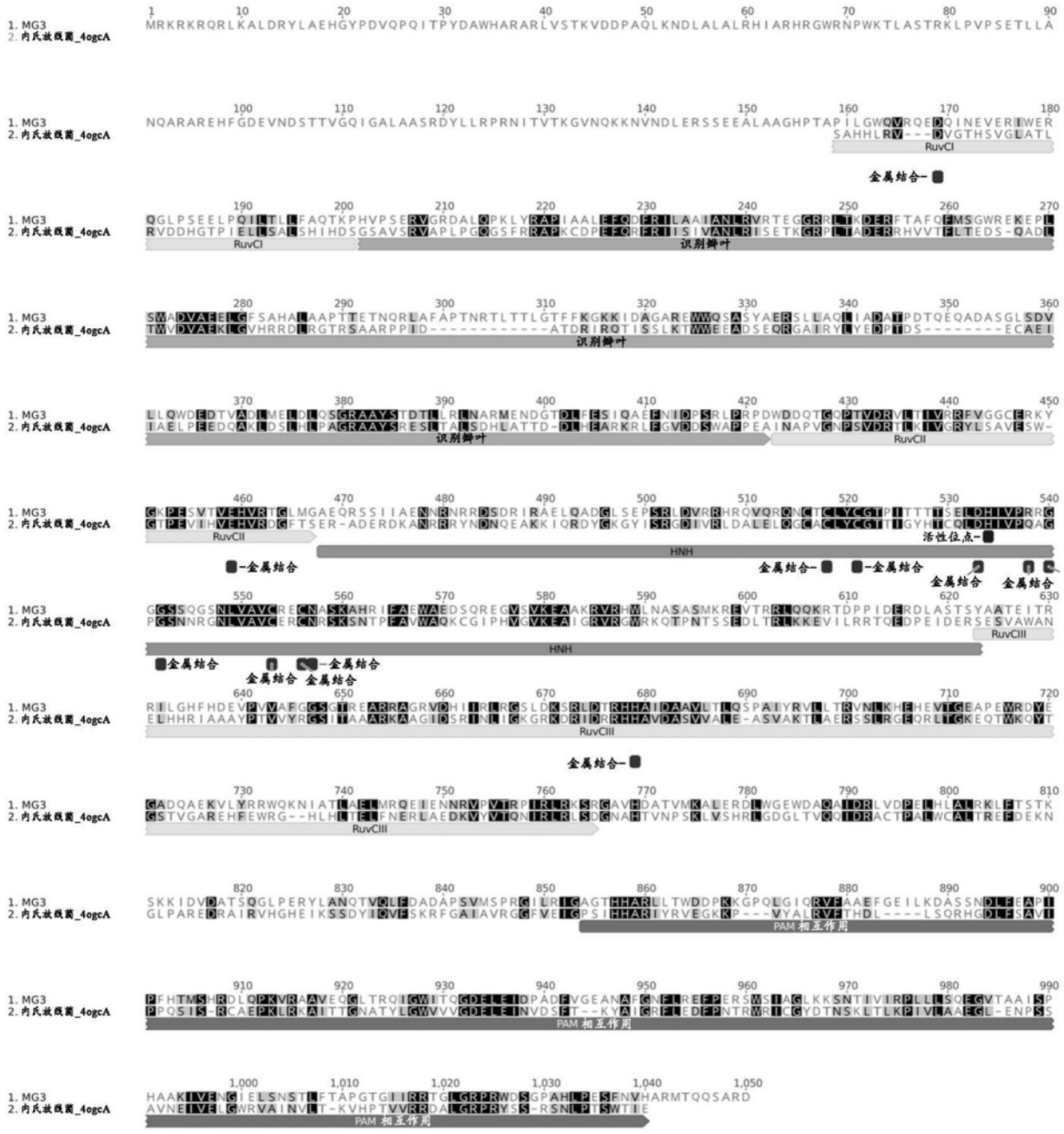

1. MG3
2. 内氏放线菌_4ogcA
RuvCI
识别鳍叶
金属结合-
RuvCII
HNH
活性位点-
RuvCIII
PAM 相互作用

图8

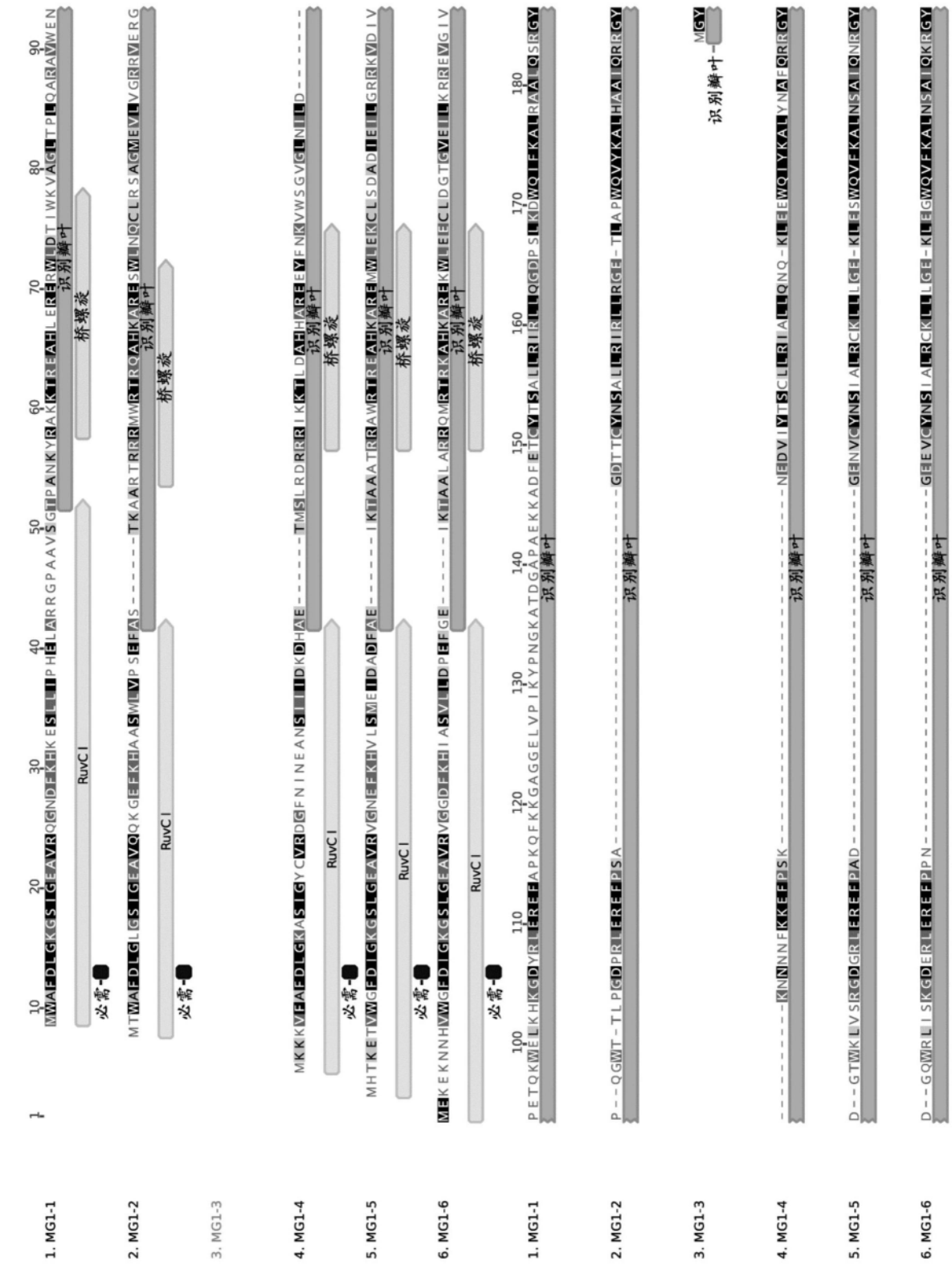


图9A

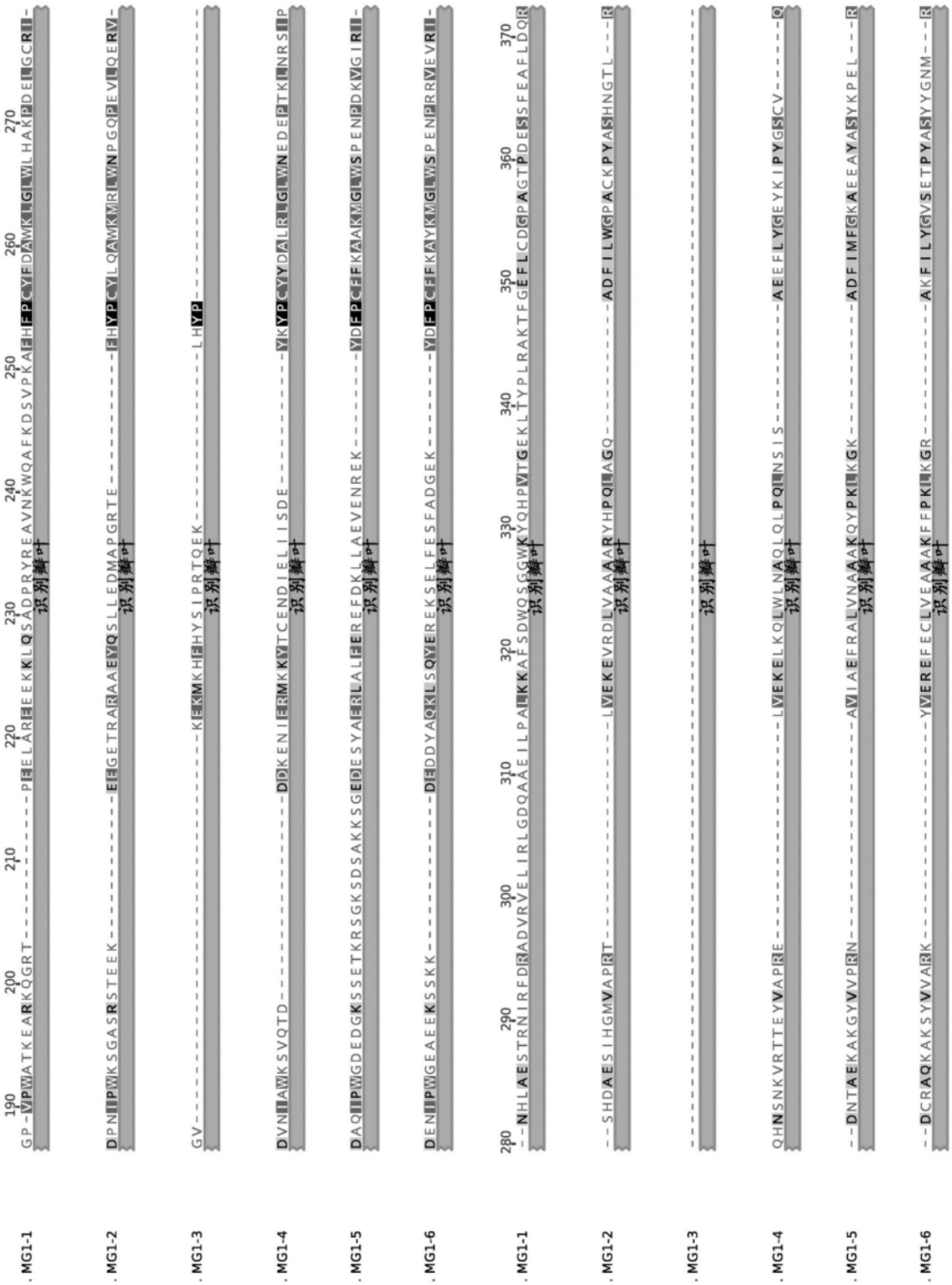


图9B

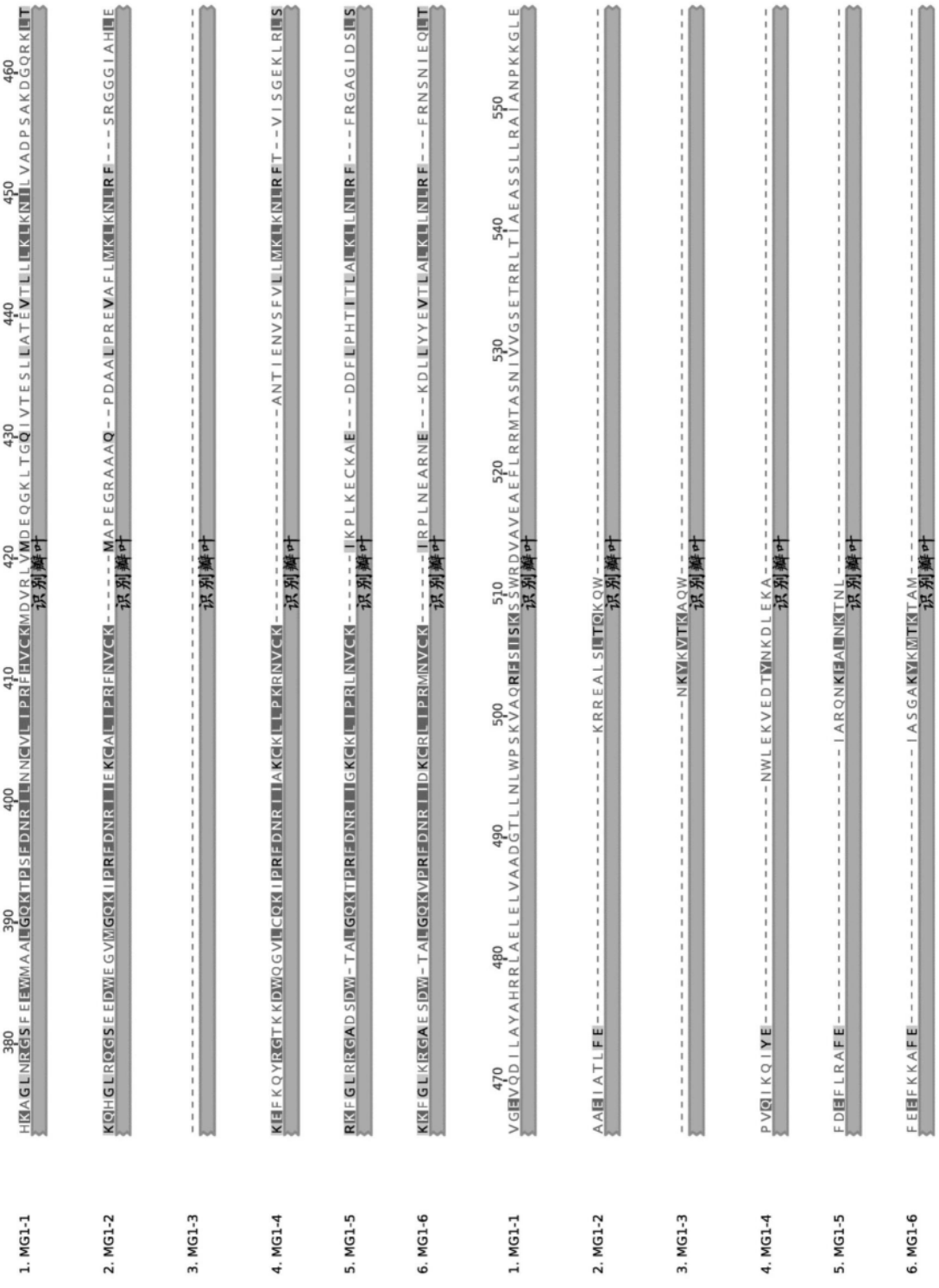

380
390
400
410
420
430
440
450
460
1. MG1-1
2. MG1-2
3. MG1-3
4. MG1-4
5. MG1-5
6. MG1-6
470
480
490
500
510
520
530
540
550
1. MG1-1
2. MG1-2
3. MG1-3
4. MG1-4
5. MG1-5
6. MG1-6

图9C

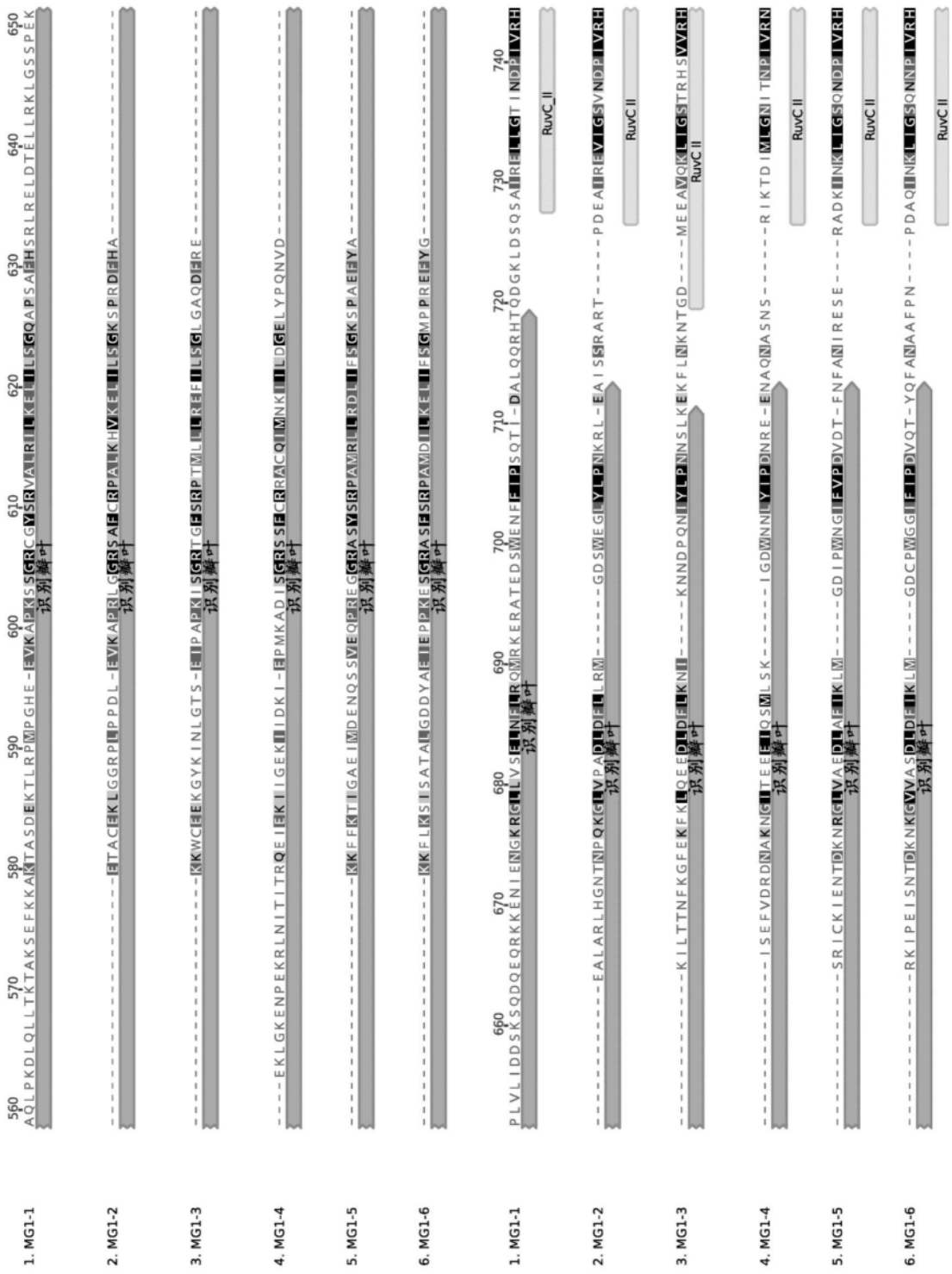


图9D



图9E


1. MG1-1
2. MG1-2
3. MG1-3
4. MG1-4
5. MG1-5
6. MG1-6
HNH结构域
RuvC_III
RuvCIII
必需
必需

图9F

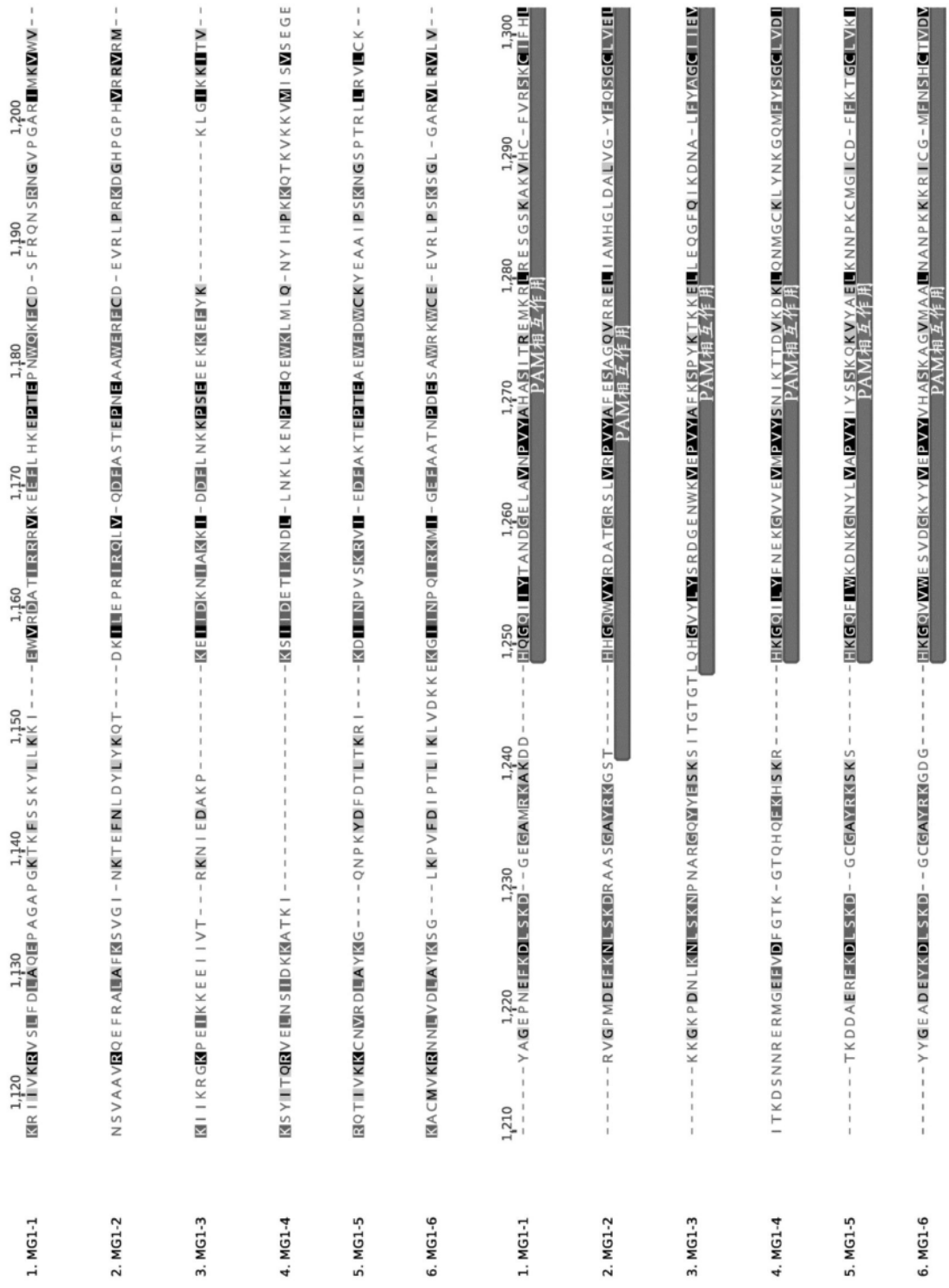


图9G

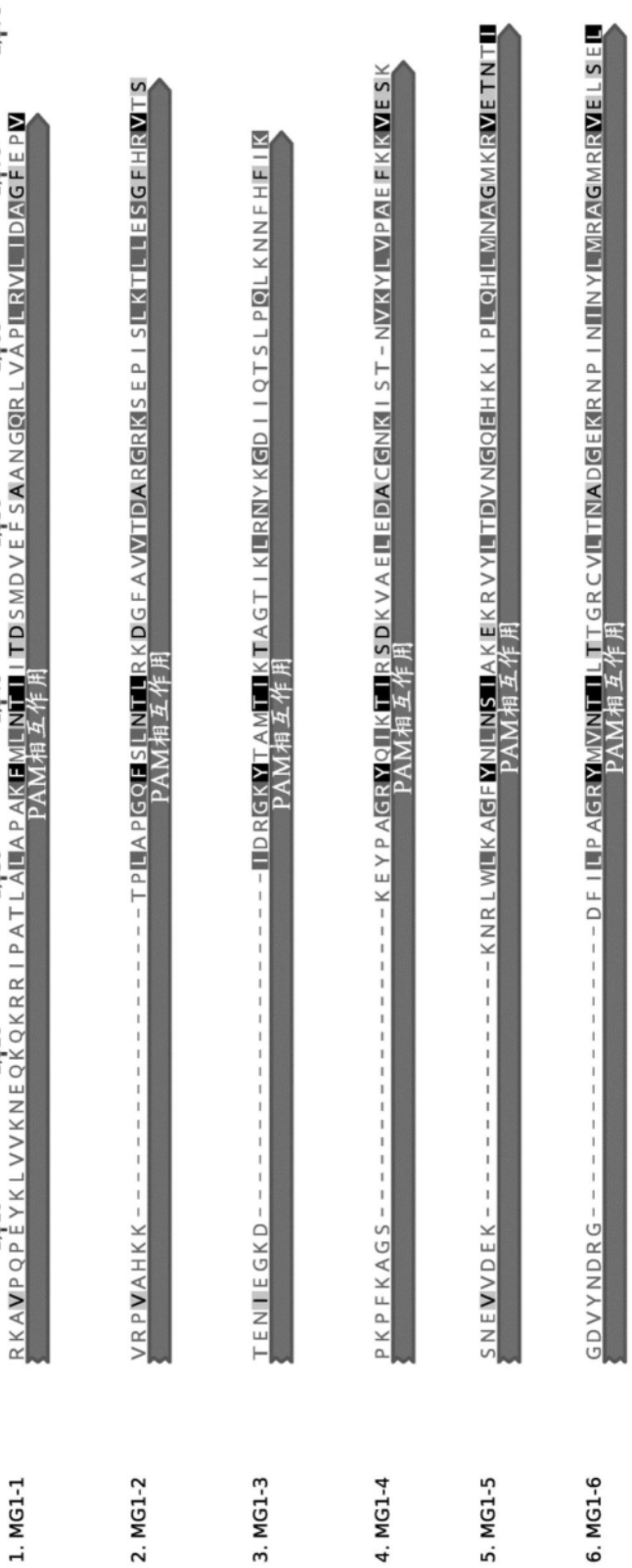
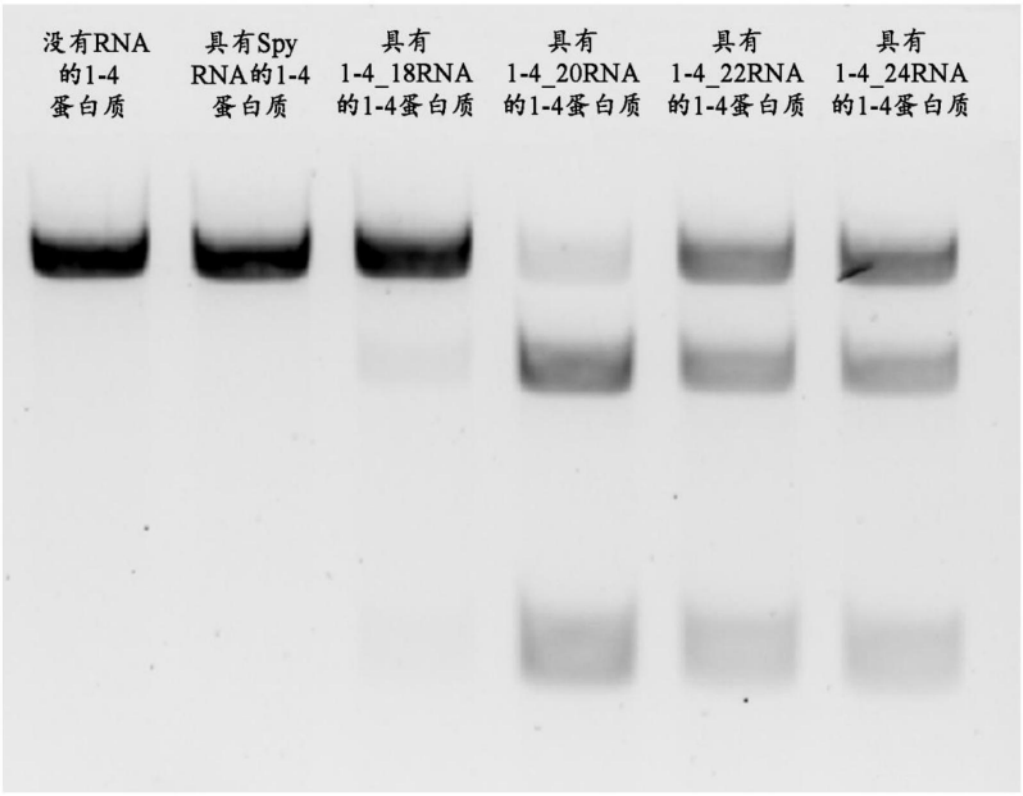

1,310
1,320
1,330
1,340
1,350
1,360
1,370
1,378
RKAVPQPEYKLVVKNEQKQKRRIPATLALAPAKFMLNTIITDSMDVEFSAANGQRLVAPLRVLIDAGFEPV
PAM相互作用
VRPVAHKK----------TPLAPGQFSLNTLRKDGFAVVTDARGRKSEPISLKTLLESGFHRVTS
PAM相互作用
TENIEGKD----------IDRGKYTAMTIKTAGTIKLRNYKGDIIQTSLPQLKNNFHFIK
PAM相互作用
PKPFKAGS----------KEYPAGRYQTKTIRSDKVAELEDACGNKISTNVKYLVPAEFKKVESK
PAM相互作用
SNEVVDEK----------KNRLWLKAGFYNLNSIAKEKRVYLTDVNGQEHKKIPLQHLMNAGMKRVETNTI
PAM相互作用
GDVYNDRG----------DFILPAGRYMVNTILTTGRCVLTNADGEKRNPINITNYLMRAGMRRVELSEL
PAM相互作用
1. MG1-1
2. MG1-2
3. MG1-3
4. MG1-4
5. MG1-5
6. MG1-6

图9H


没有RNA的1-4蛋白质
具有Spy RNA的1-4蛋白质
具有1-4_18RNA的1-4蛋白质
具有1-4_20RNA的1-4蛋白质
具有1-4_22RNA的1-4蛋白质
具有1-4_24RNA的1-4蛋白质

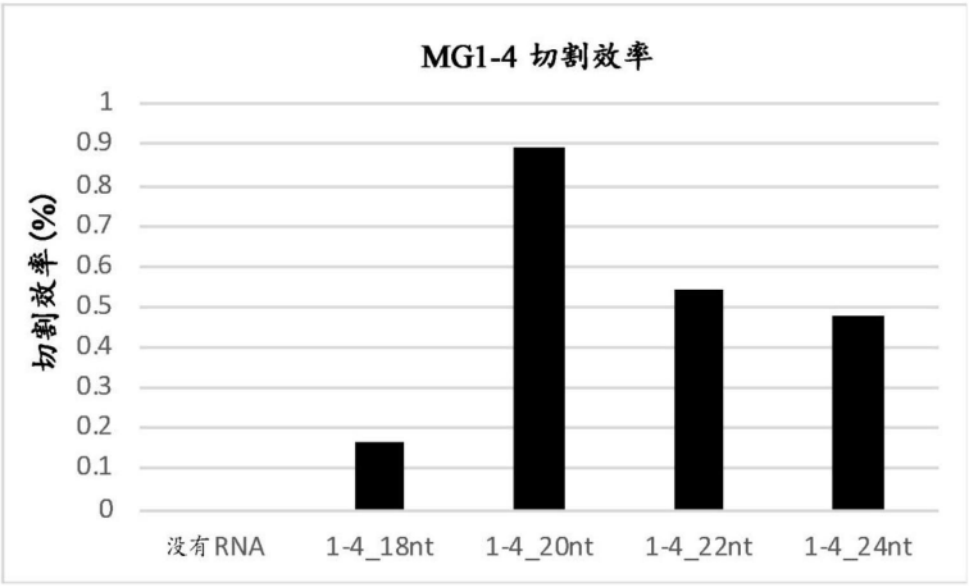

MG1-4 切割效率
切割效率(%)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
没有RNA
1-4_18nt
1-4_20nt
1-4_22nt
1-4_24nt

图10

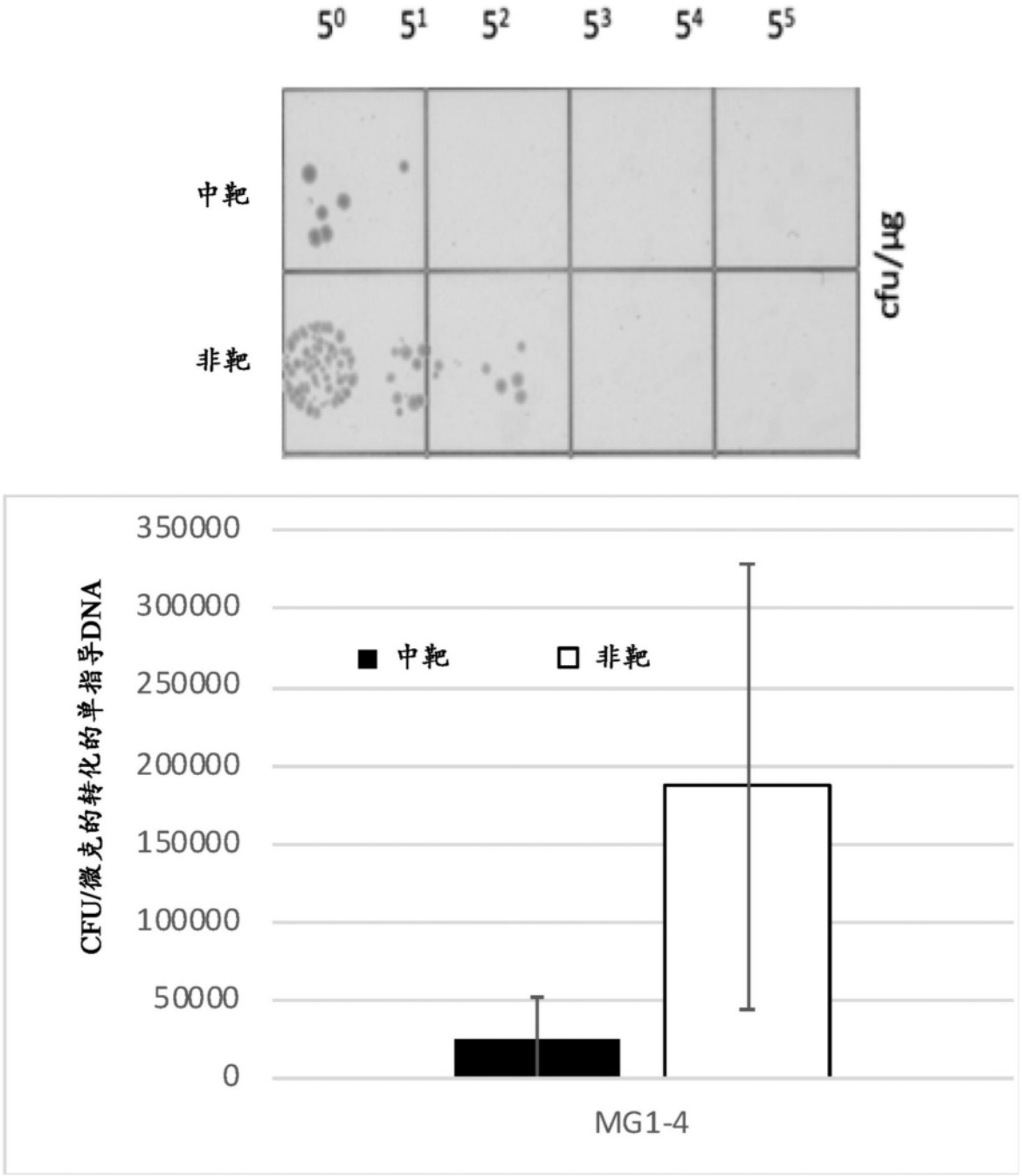
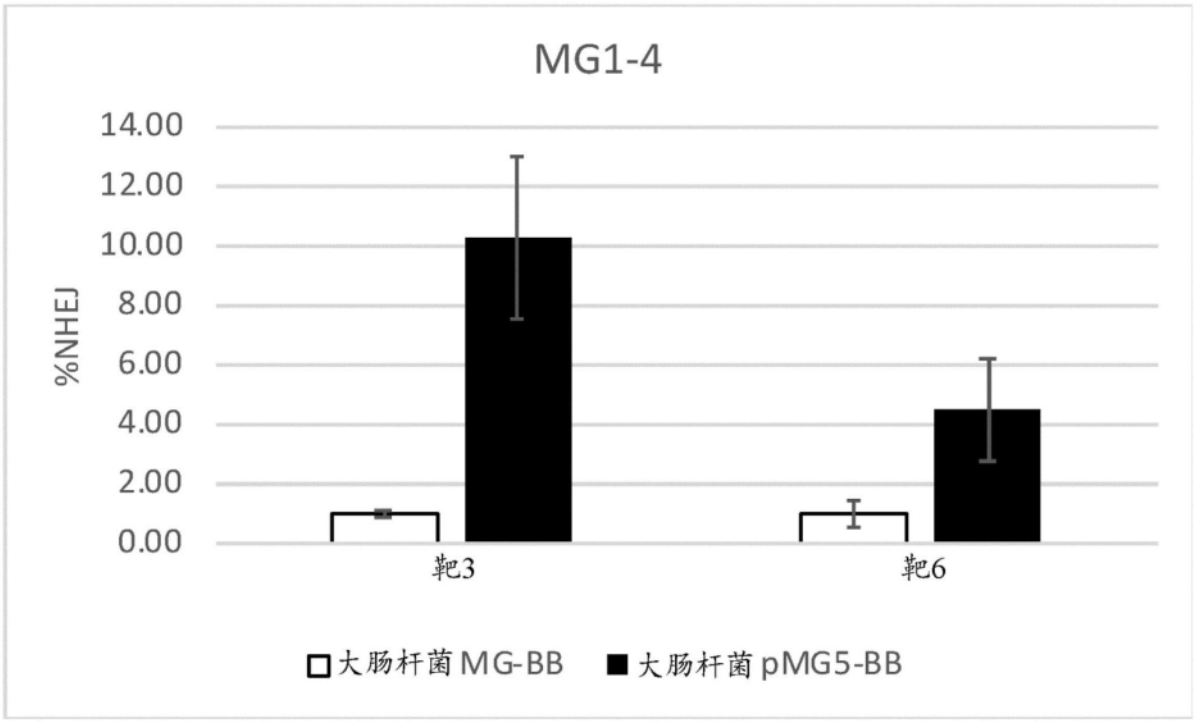

5⁰ 5¹ 5² 5³ 5⁴ 5⁵
中靶
非靶
cfu/μg
350000
300000
250000
200000
150000
100000
50000
0
CFU/微克的转化的单指导DNA
中靶
非靶
MG1-4

图11


MG1-4
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
%NHEJ
靶3
靶6
大肠杆菌 MG-BB
大肠杆菌 pMG5-BB

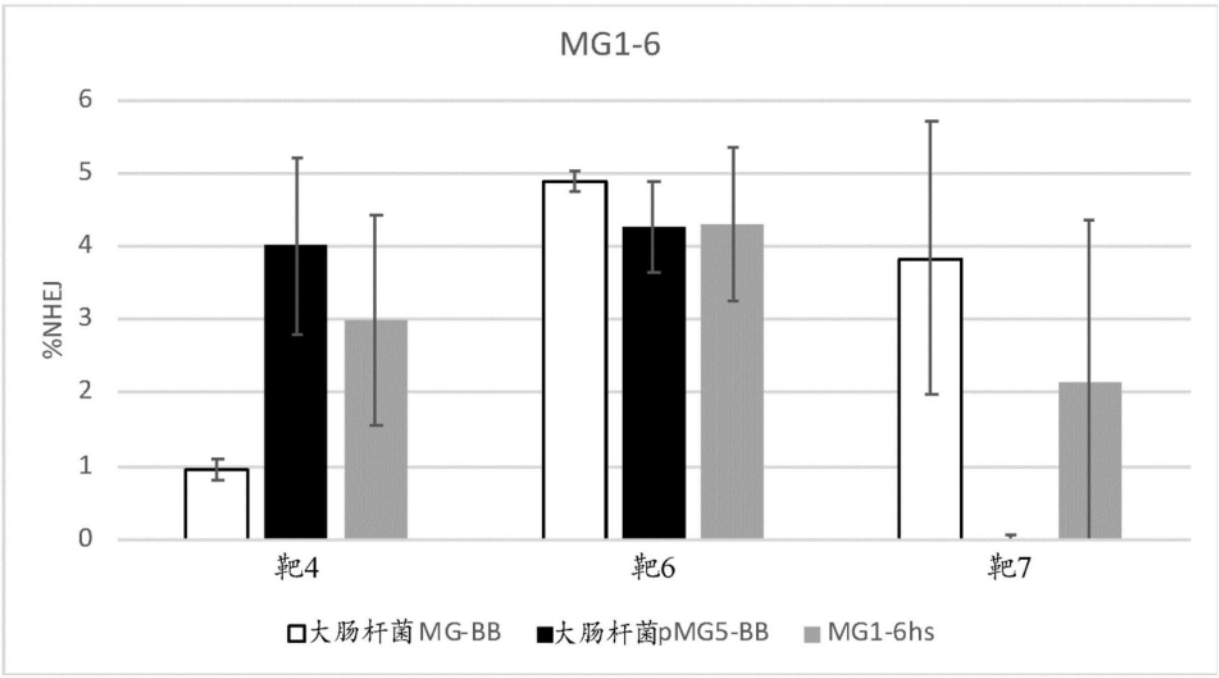
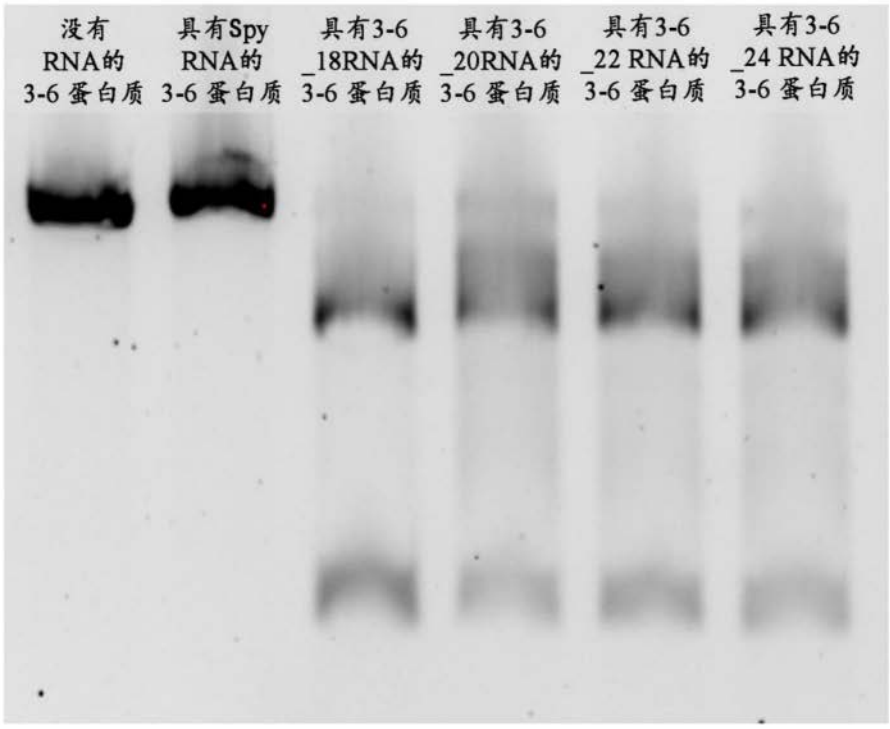

MG1-6
6
5
4
3
2
1
0
%NHEJ
靶4
靶6
靶7
大肠杆菌MG-BB
大肠杆菌pMG5-BB
MG1-6hs

图12


没有RNA的3-6蛋白质
具有Spy RNA的3-6蛋白质
具有3-6_18RNA的3-6蛋白质
具有3-6_20RNA的3-6蛋白质
具有3-6_22 RNA的3-6蛋白质
具有3-6_24 RNA的3-6蛋白质

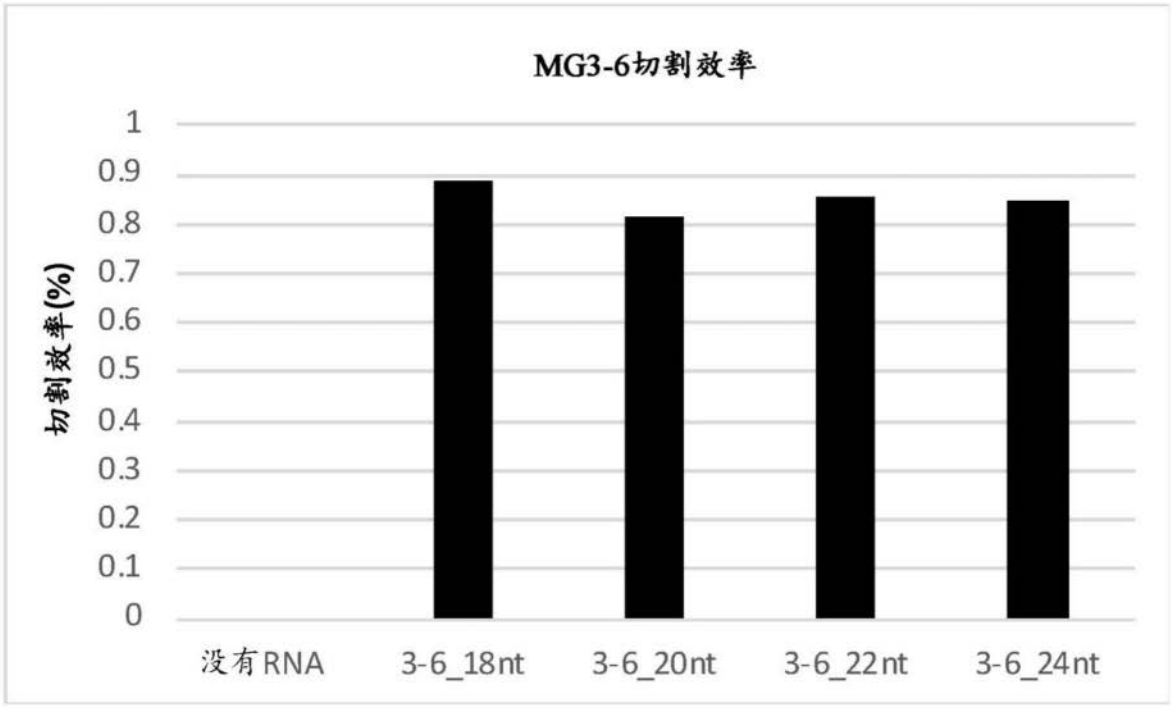

MG3-6切割效率
切割效率(%)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
没有RNA
3-6_18nt
3-6_20nt
3-6_22nt
3-6_24nt

图13

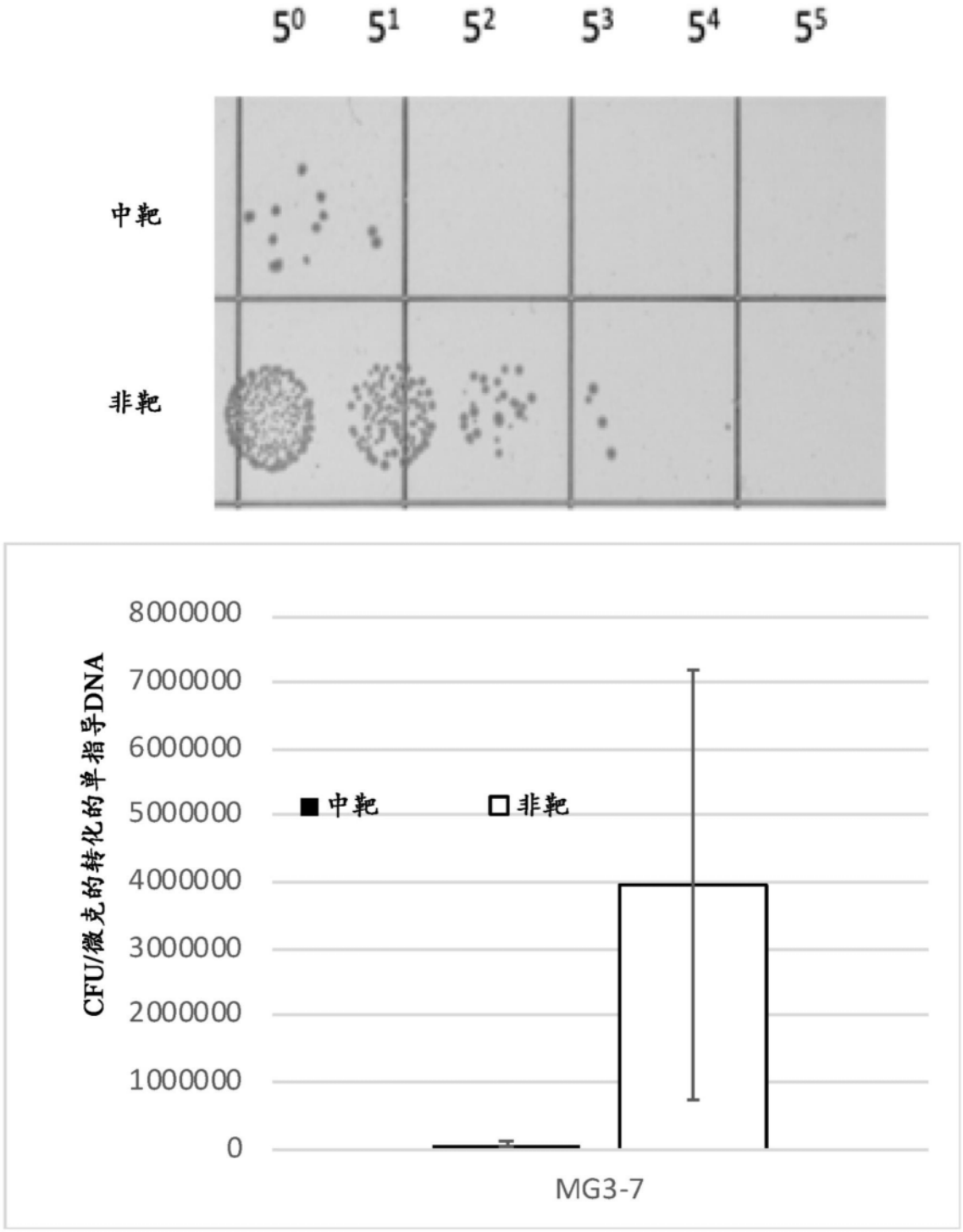

5⁰
5¹
5²
5³
5⁴
5⁵
中靶
非靶
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
CFU/微克的转化的单指导DNA
中靶
非靶
MG3-7

图14

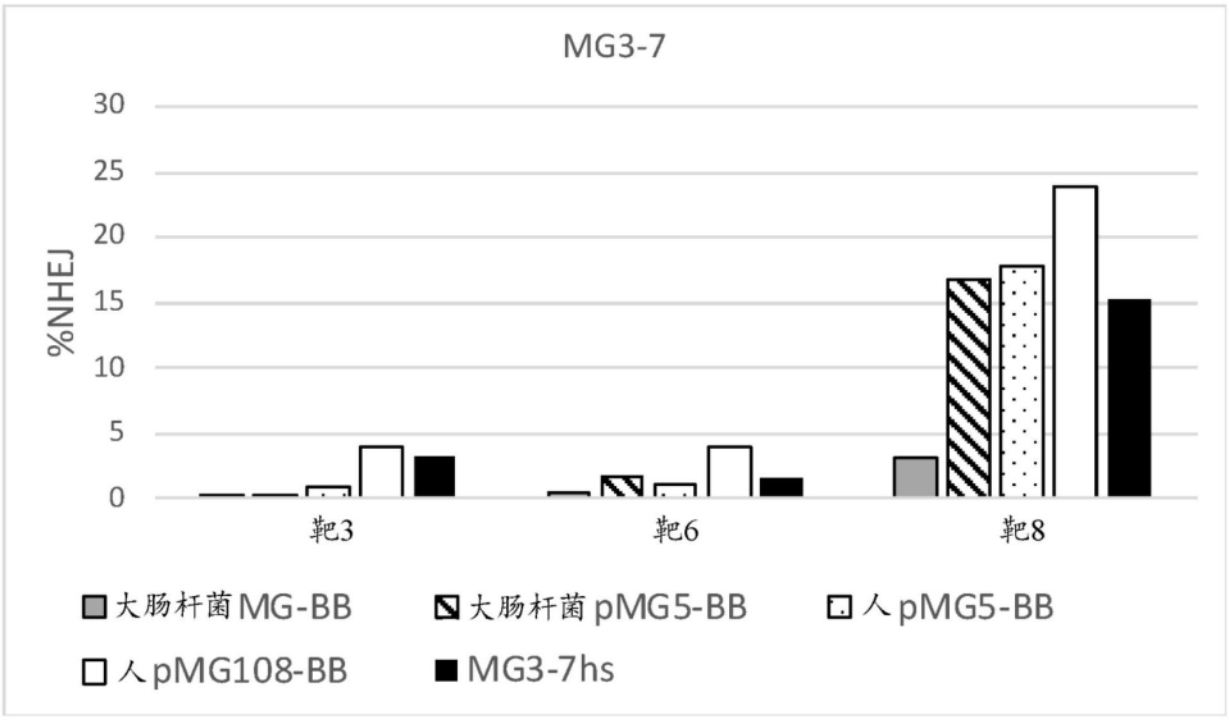
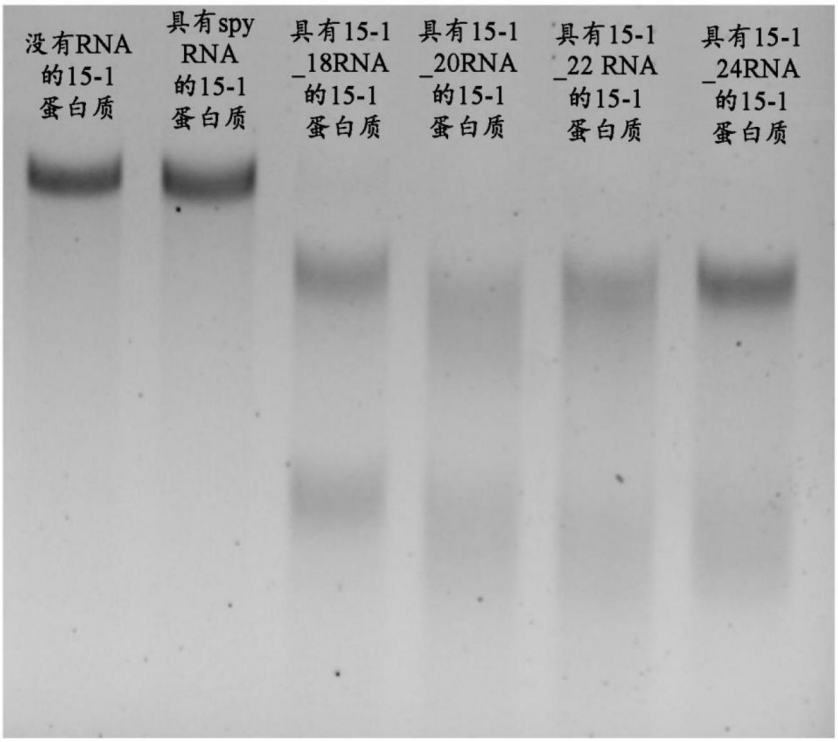

MG3-7
30
25
20
15
10
5
0
%NHEJ
靶3
靶6
靶8
大肠杆菌MG-BB
大肠杆菌pMG5-BB
人pMG5-BB
人pMG108-BB
MG3-7hs

图15


没有RNA的15-1蛋白质
具有spy RNA的15-1蛋白质
具有15-1_18RNA的15-1蛋白质
具有15-1_20RNA的15-1蛋白质
具有15-1_22 RNA的15-1蛋白质
具有15-1_24RNA的15-1蛋白质

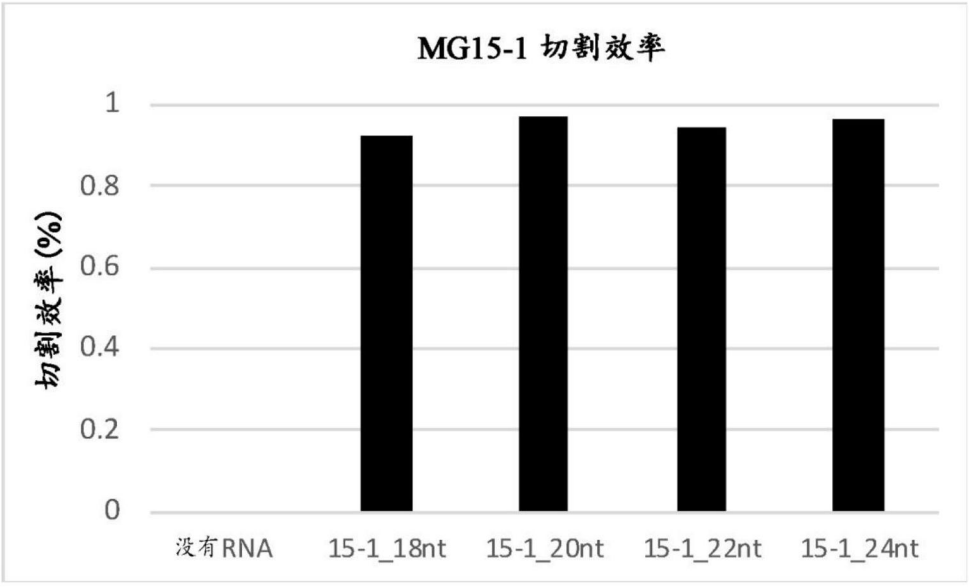
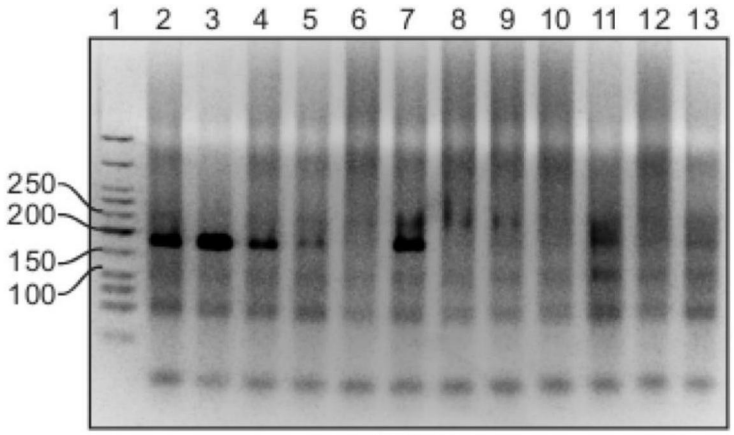
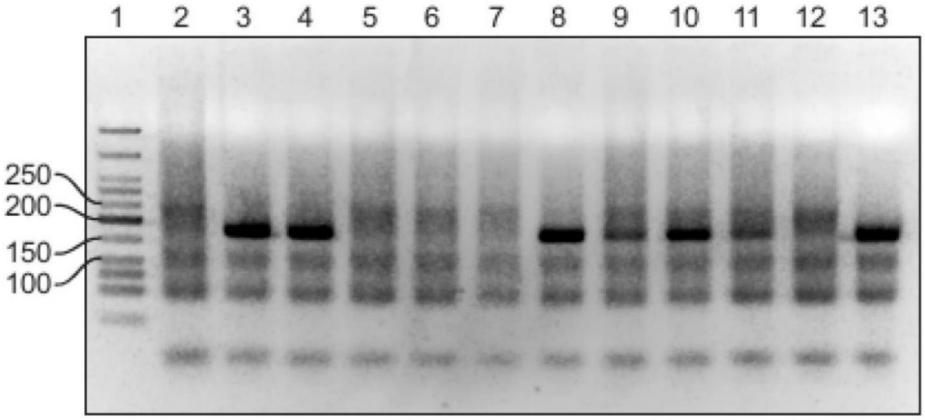

MG15-1 切割效率
切割效率(%)
1
0.8
0.6
0.4
0.2
0
没有RNA
15-1_18nt
15-1_20nt
15-1_22nt
15-1_24nt

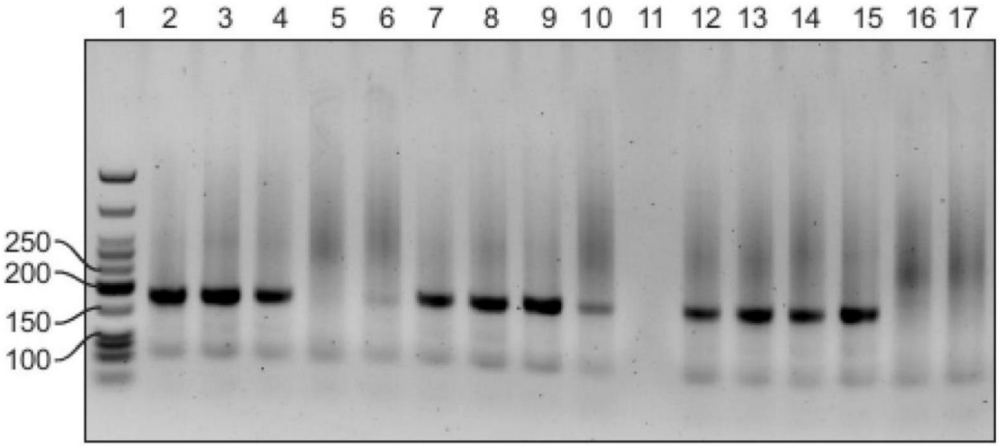
图16



凝胶1

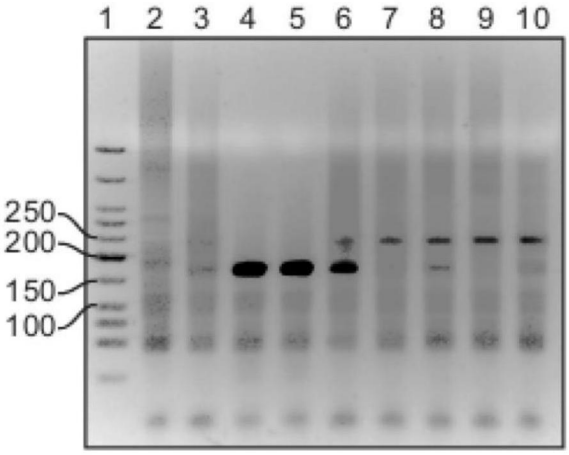


凝胶2

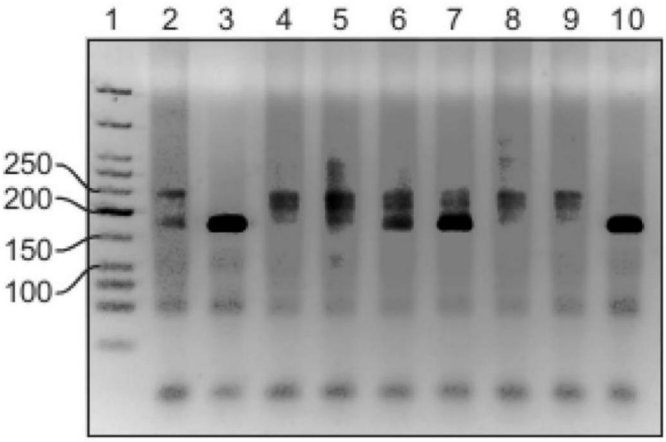


凝胶3

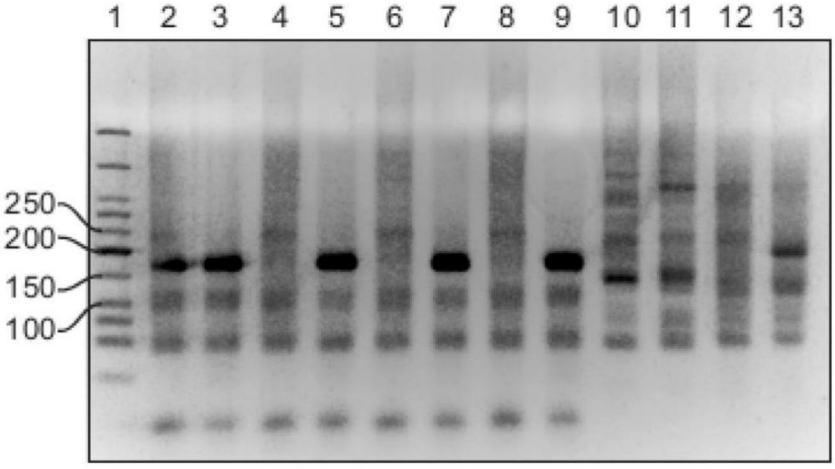
图17



凝胶4

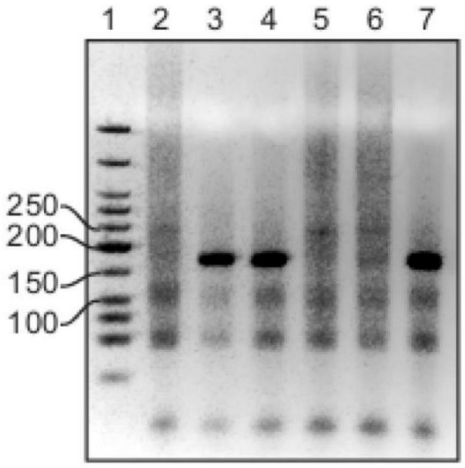


凝胶5

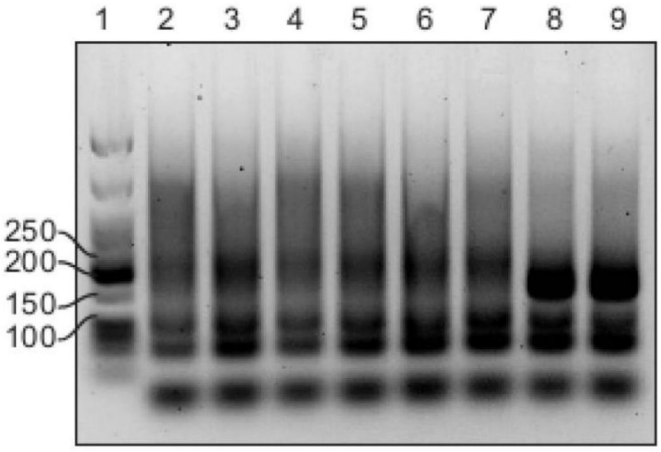


凝胶6

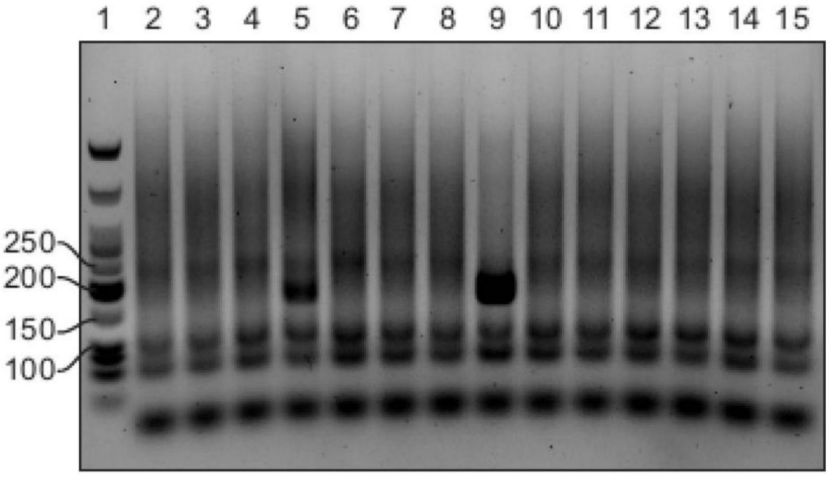
图18



凝胶7

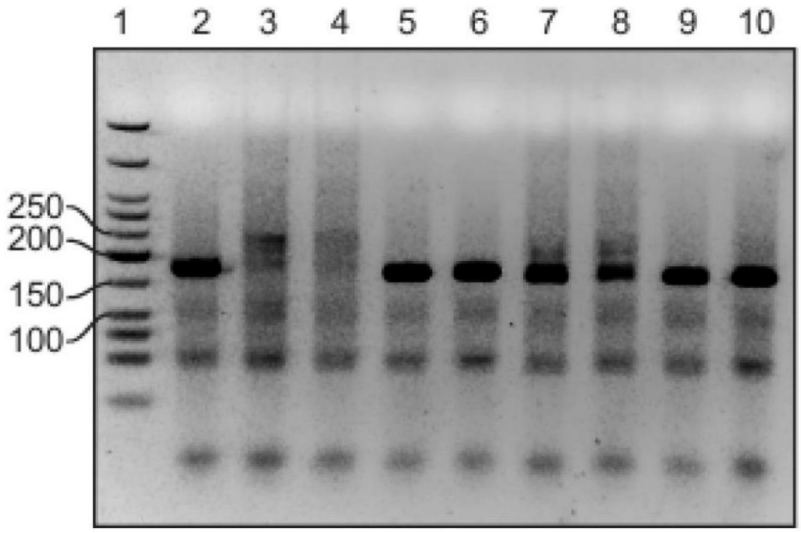


凝胶8

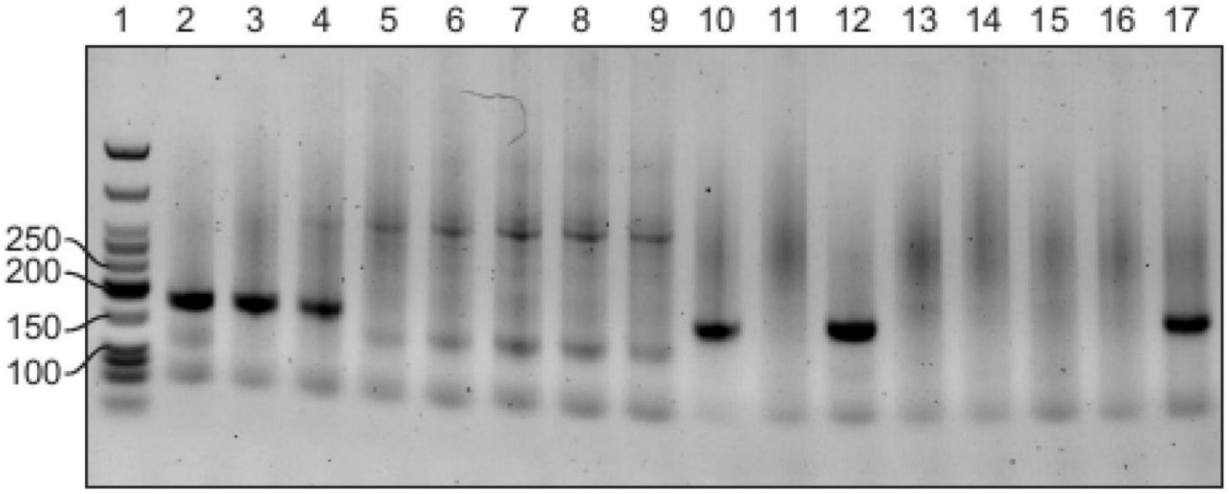


凝胶9

图19



凝胶10



凝胶11

图20

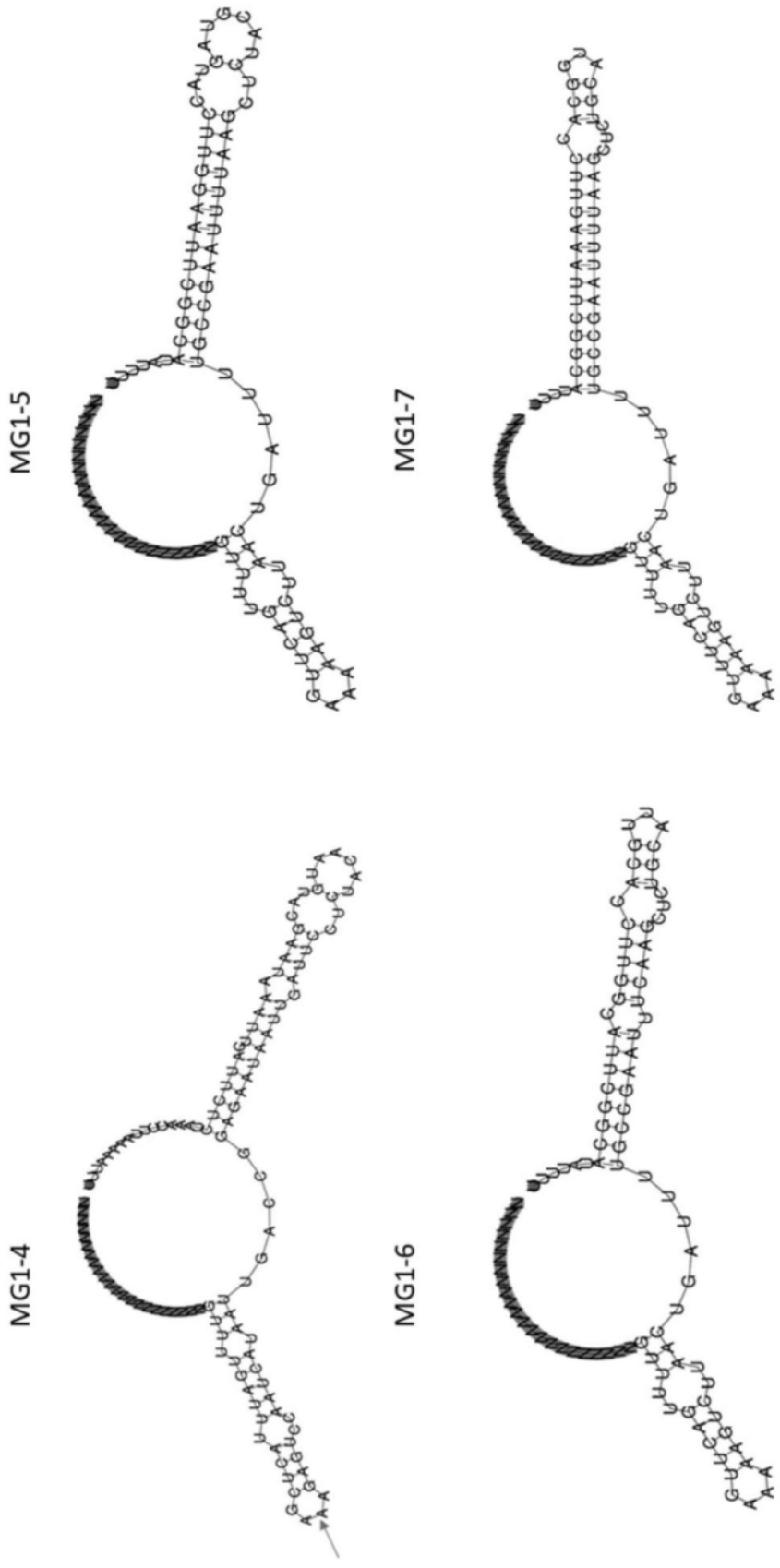


图21

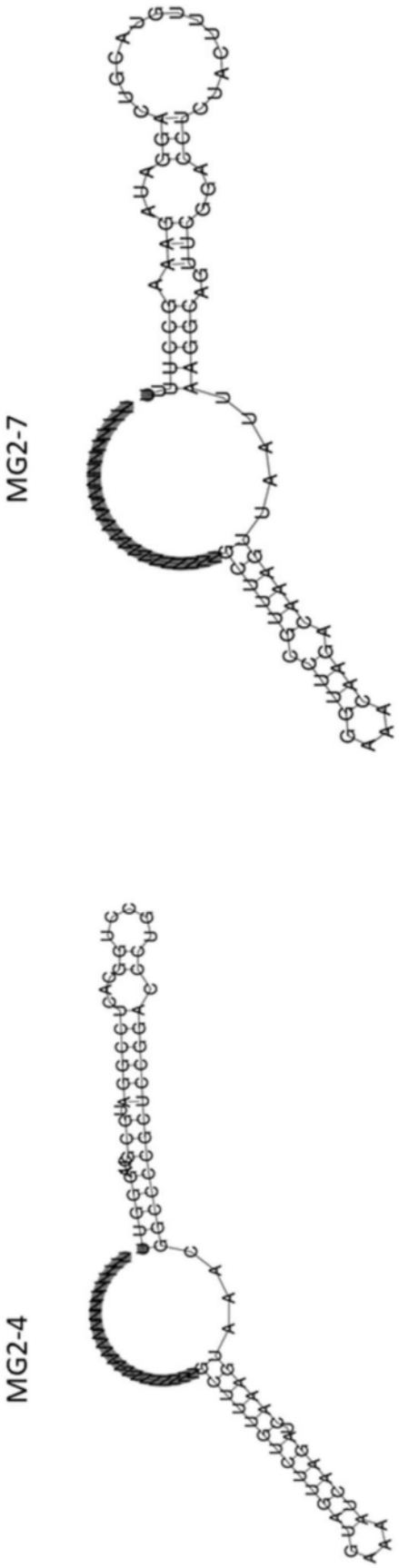


图22

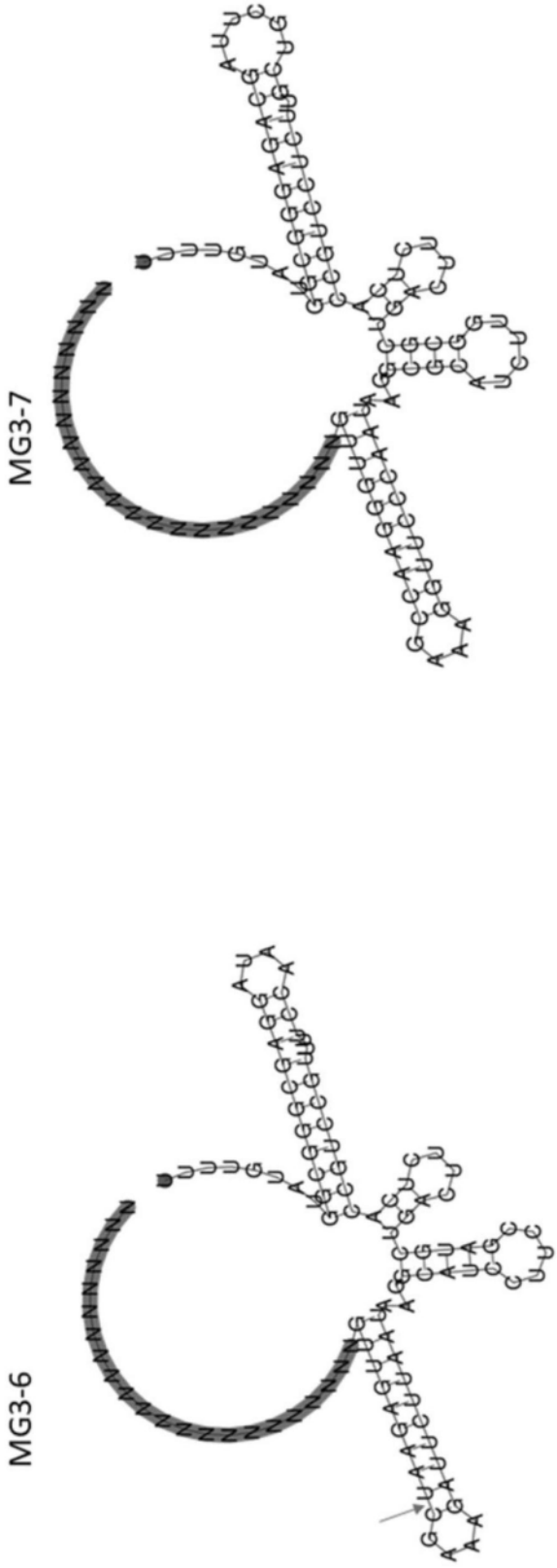


图23



图24

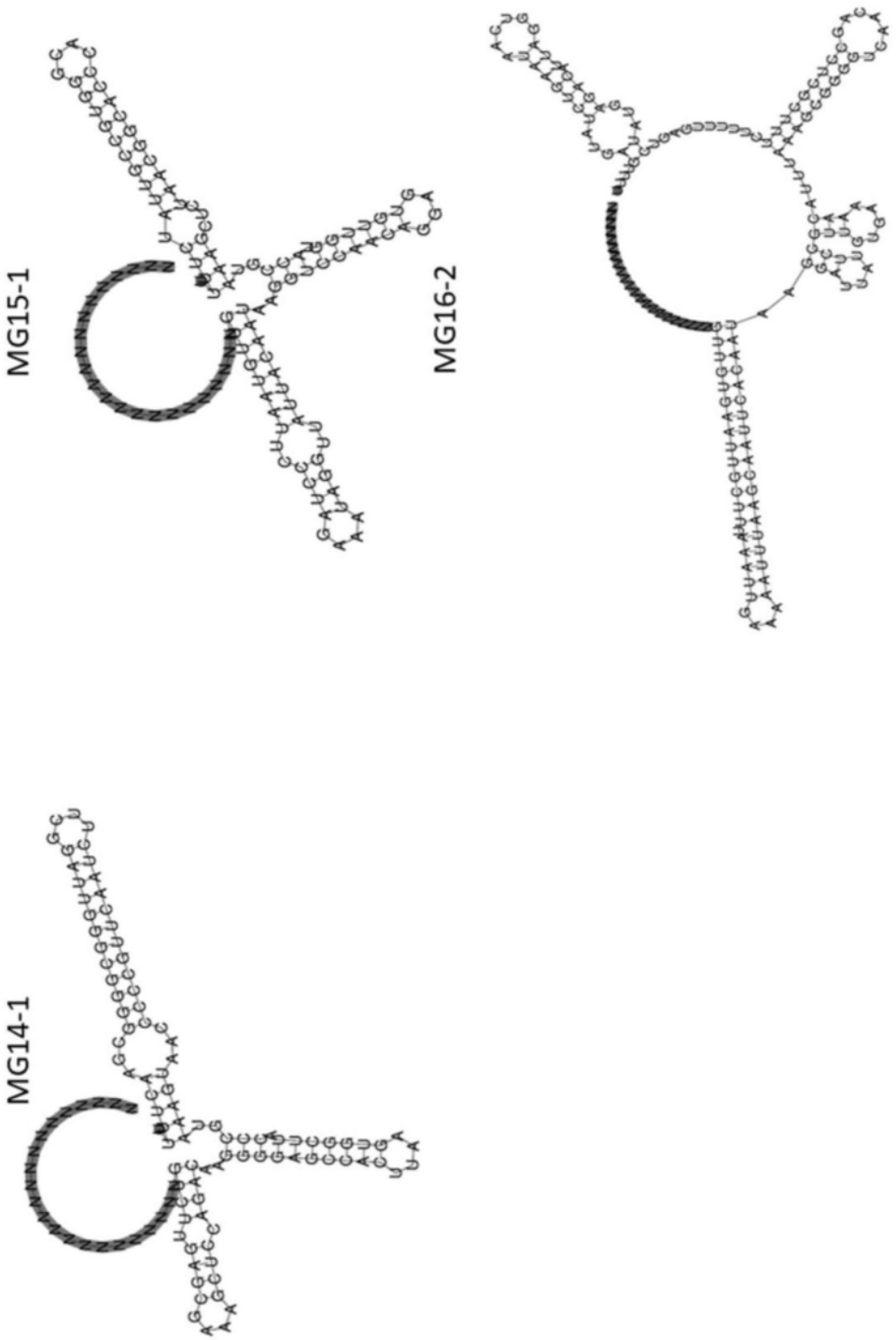


图25

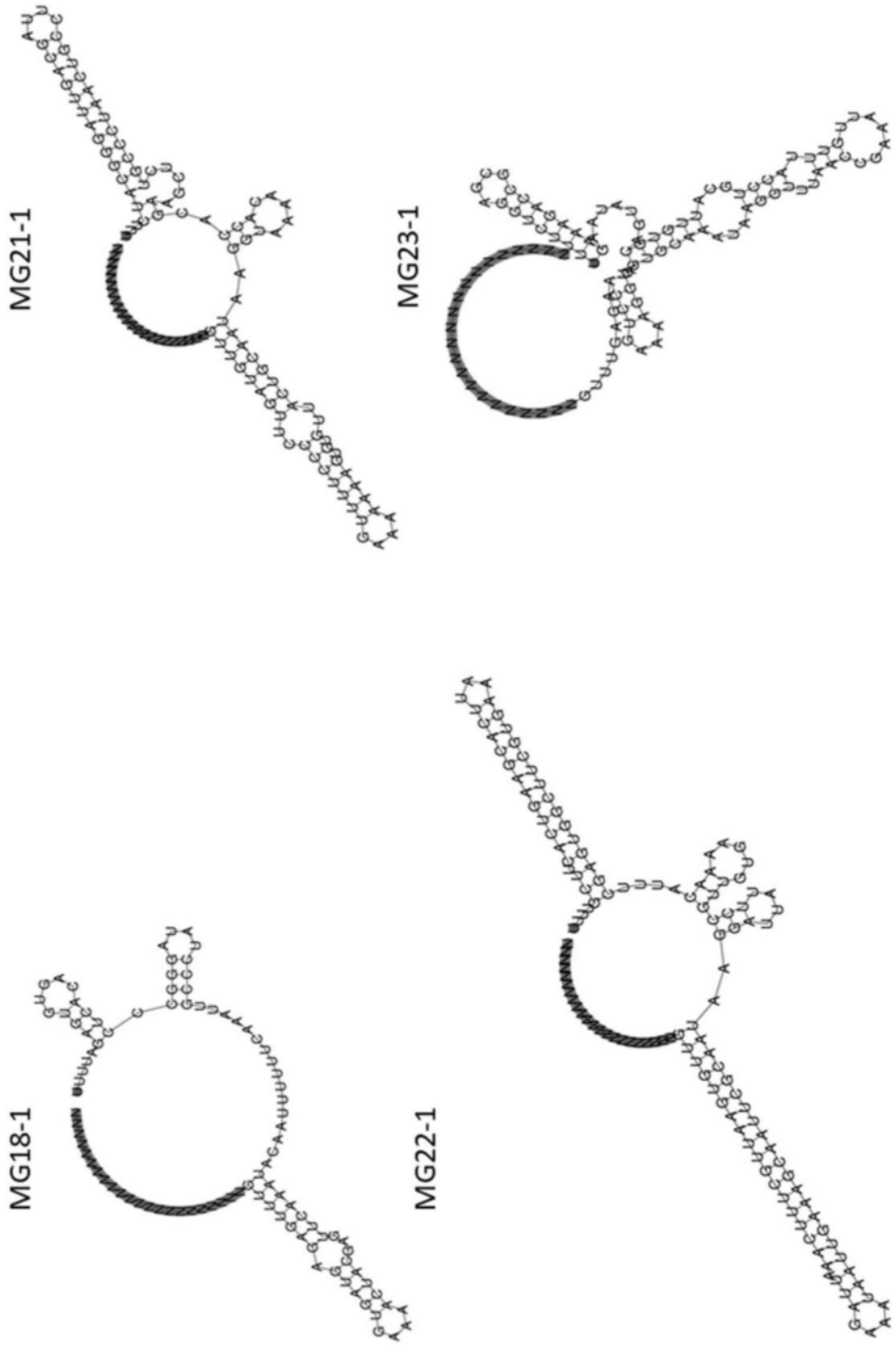


图26

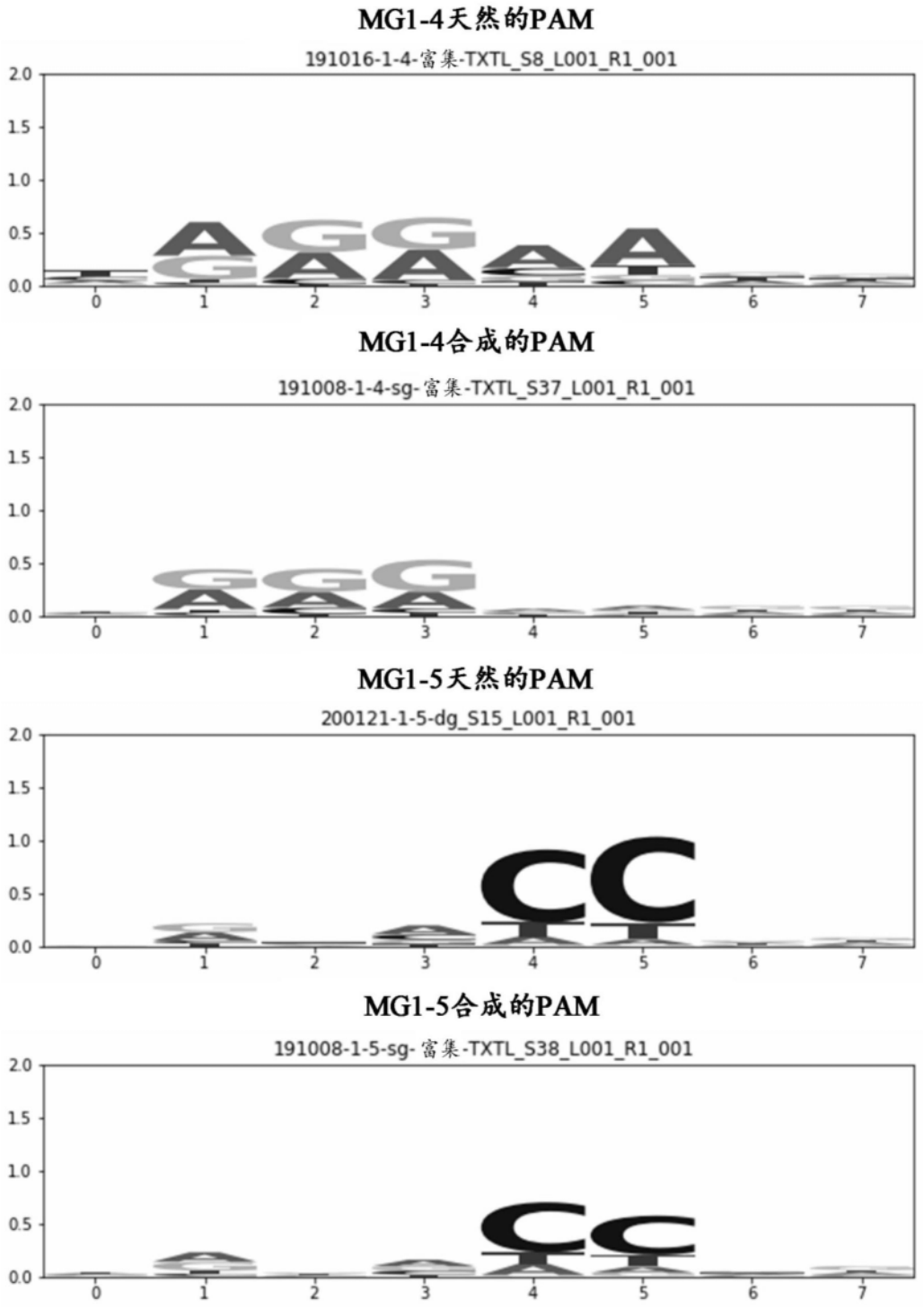

MG1-4天然的PAM
191016-1-4-富集-TXTL_S8_L001_R1_001
MG1-4合成的PAM
191008-1-4-sg-富集-TXTL_S37_L001_R1_001
MG1-5天然的PAM
200121-1-5-dg_S15_L001_R1_001
MG1-5合成的PAM
191008-1-5-sg-富集-TXTL_S38_L001_R1_001

图27

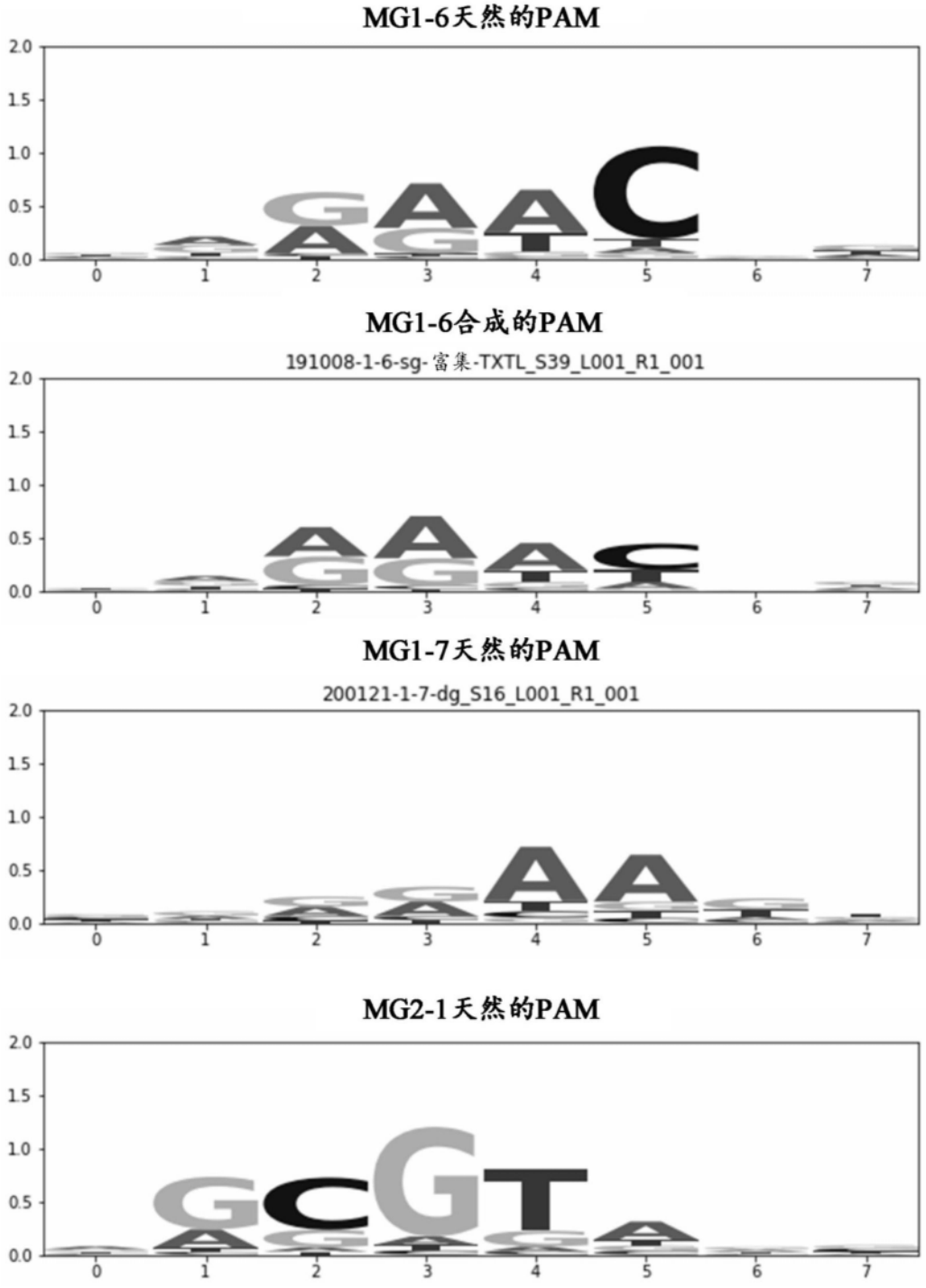
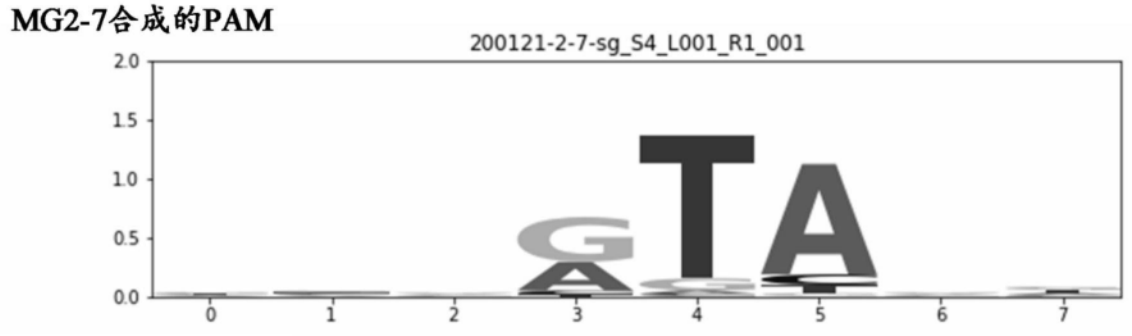
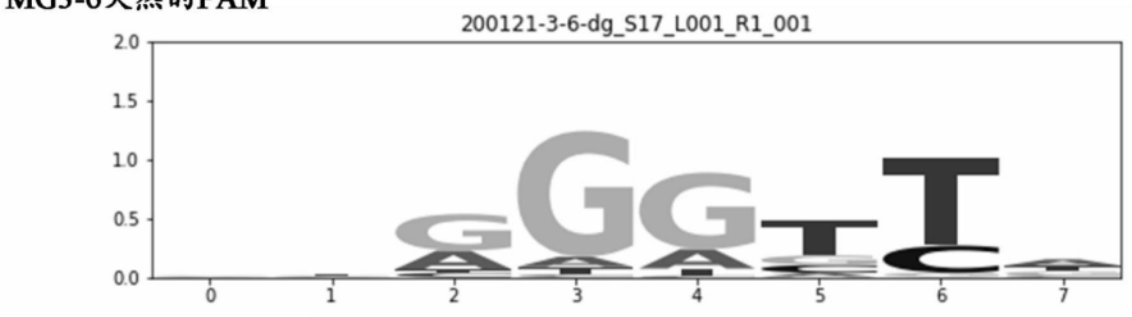

MG1-6天然的PAM
MG1-6合成的PAM
191008-1-6-sg-富集-TXTL_S39_L001_R1_001
MG1-7天然的PAM
200121-1-7-dg_S16_L001_R1_001
MG2-1天然的PAM

图28

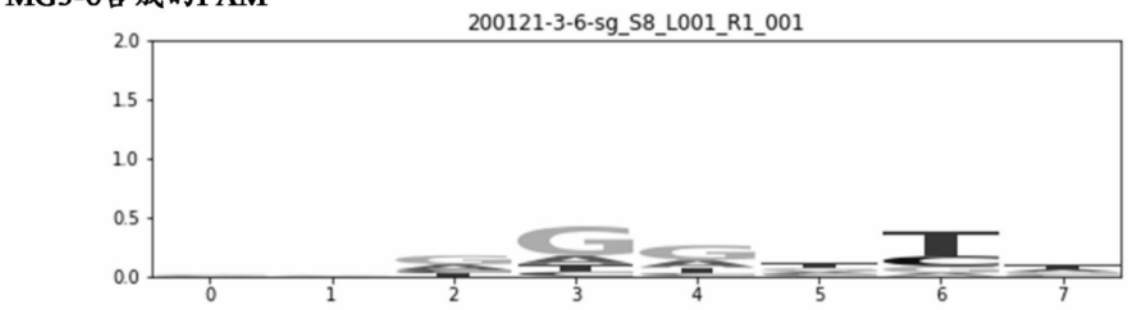
MG2-7合成的PAM



MG3-6天然的PAM



MG3-6合成的PAM



MG3-7天然的PAM

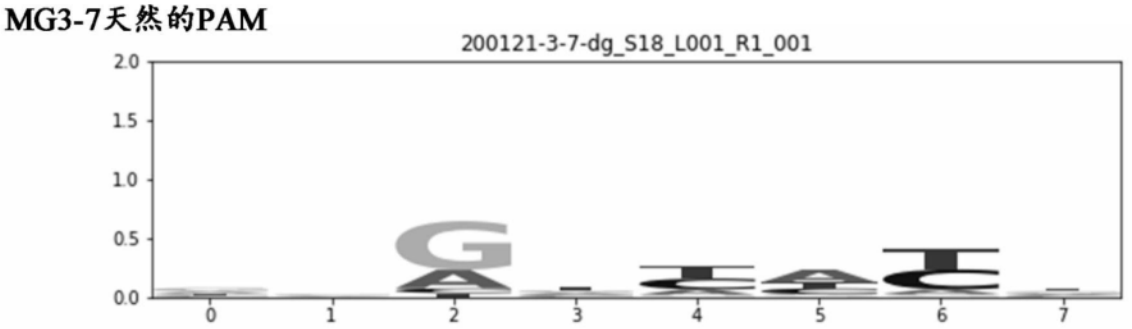


图29

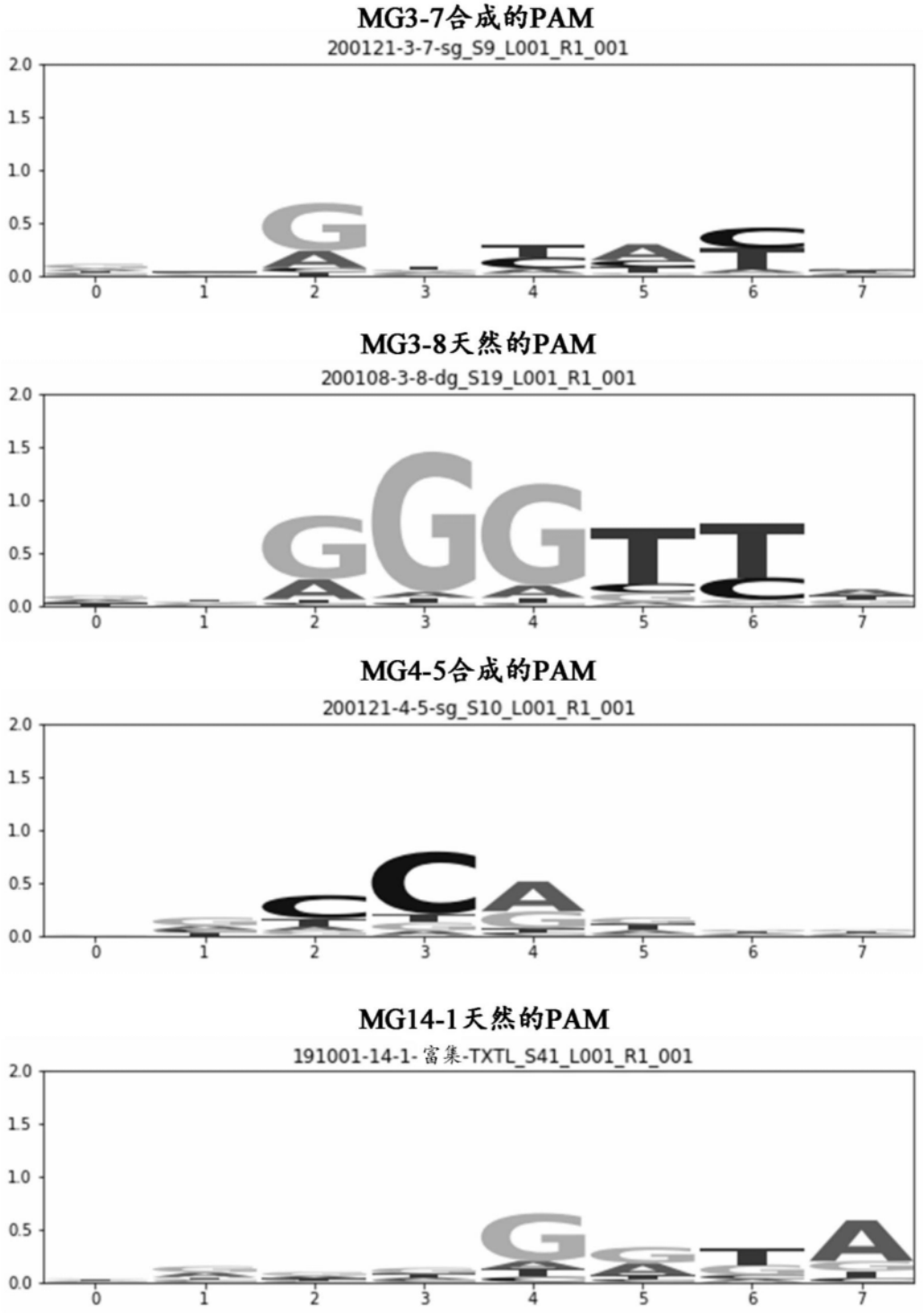

MG3-7合成的PAM
200121-3-7-sg_S9_L001_R1_001
MG3-8天然的PAM
200108-3-8-dg_S19_L001_R1_001
MG4-5合成的PAM
200121-4-5-sg_S10_L001_R1_001
MG14-1天然的PAM
191001-14-1-富集-TXTL_S41_L001_R1_001

图30

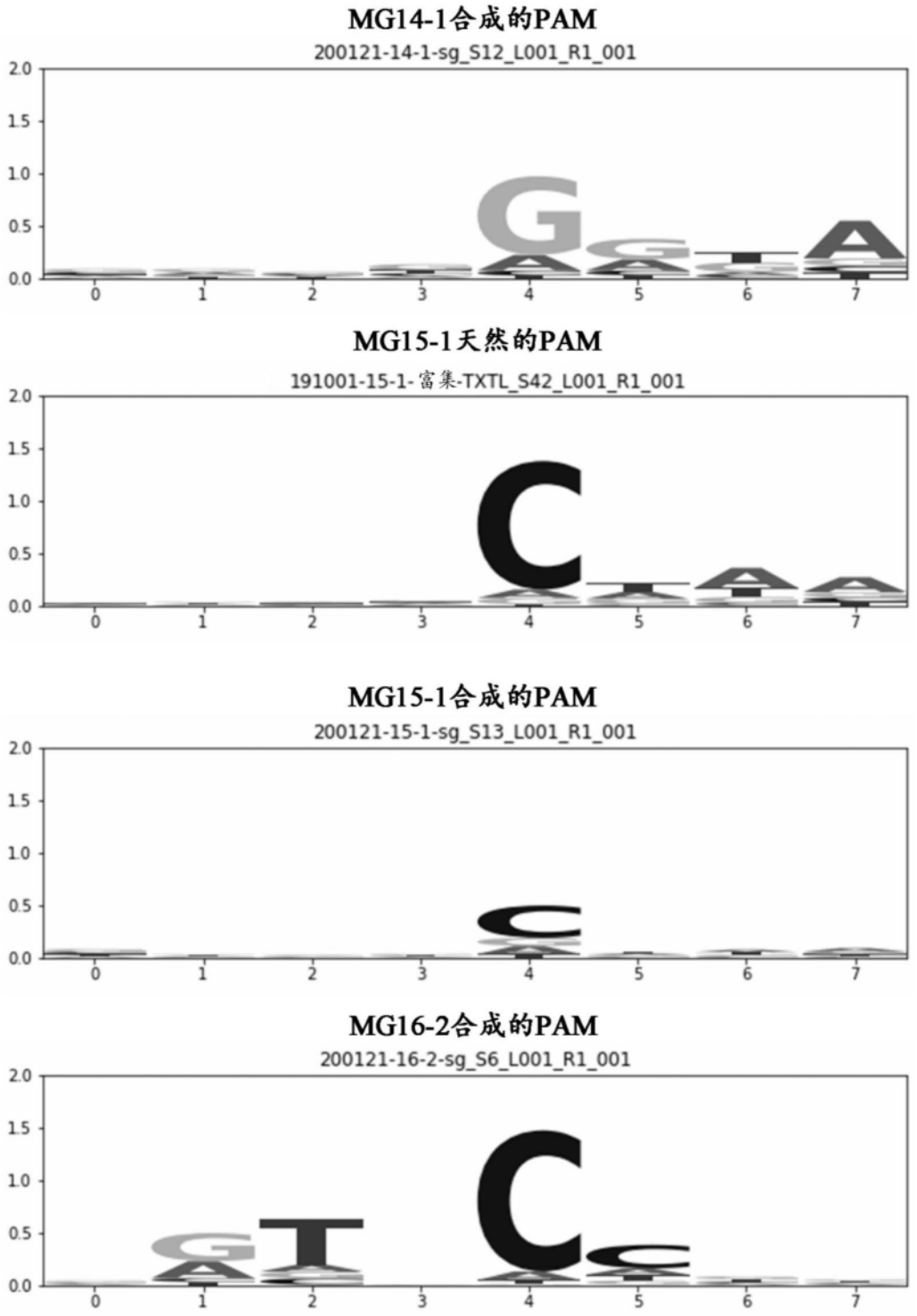
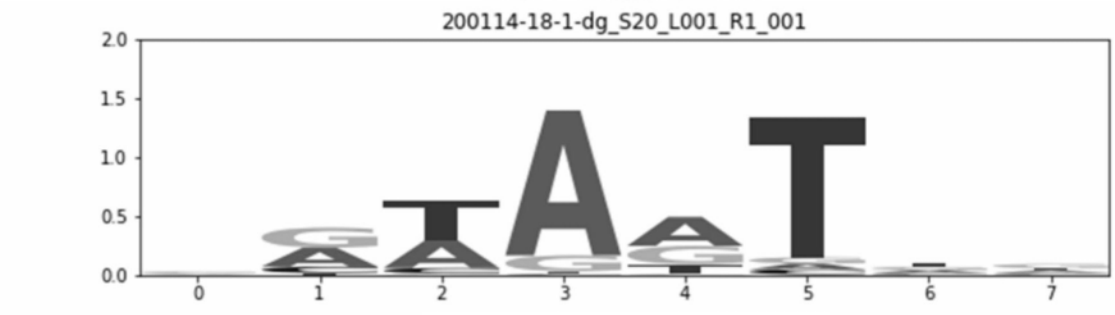
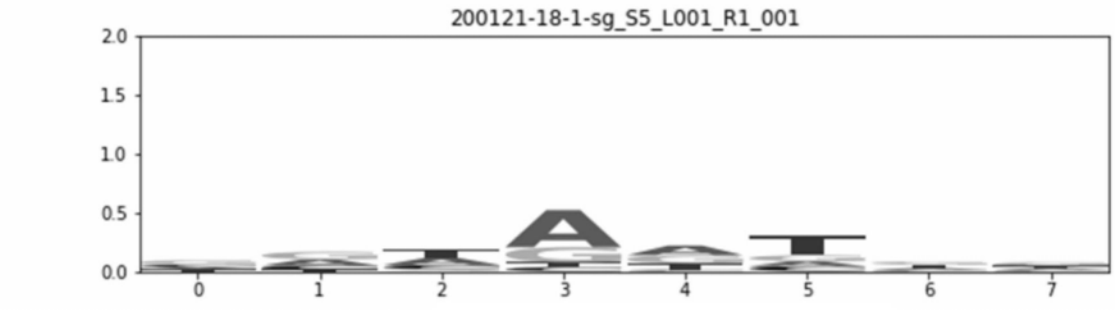

MG14-1合成的PAM
200121-14-1-sg_S12_L001_R1_001
MG15-1天然的PAM
191001-15-1-富集-TXTL_S42_L001_R1_001
MG15-1合成的PAM
200121-15-1-sg_S13_L001_R1_001
MG16-2合成的PAM
200121-16-2-sg_S6_L001_R1_001

图31

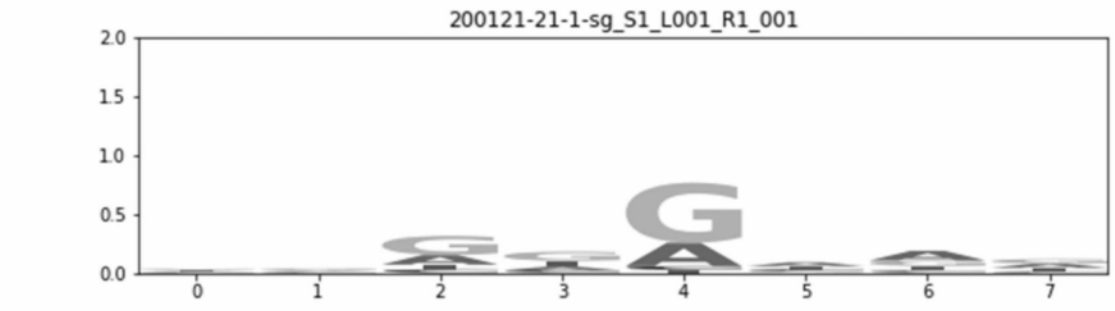
MG18-1天然的PAM



MG18-1合成的PAM



MG21-1合成的PAM



MG22-1合成的PAM

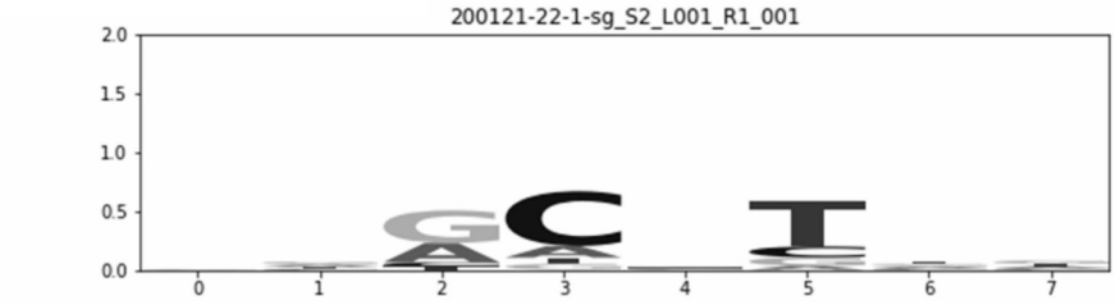


图32

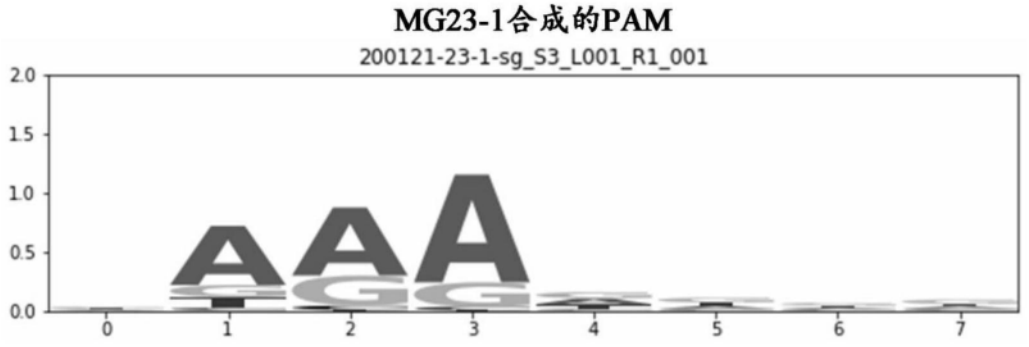

MG23-1合成的PAM
200121-23-1-sg_S3_L001_R1_001
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
7

图33

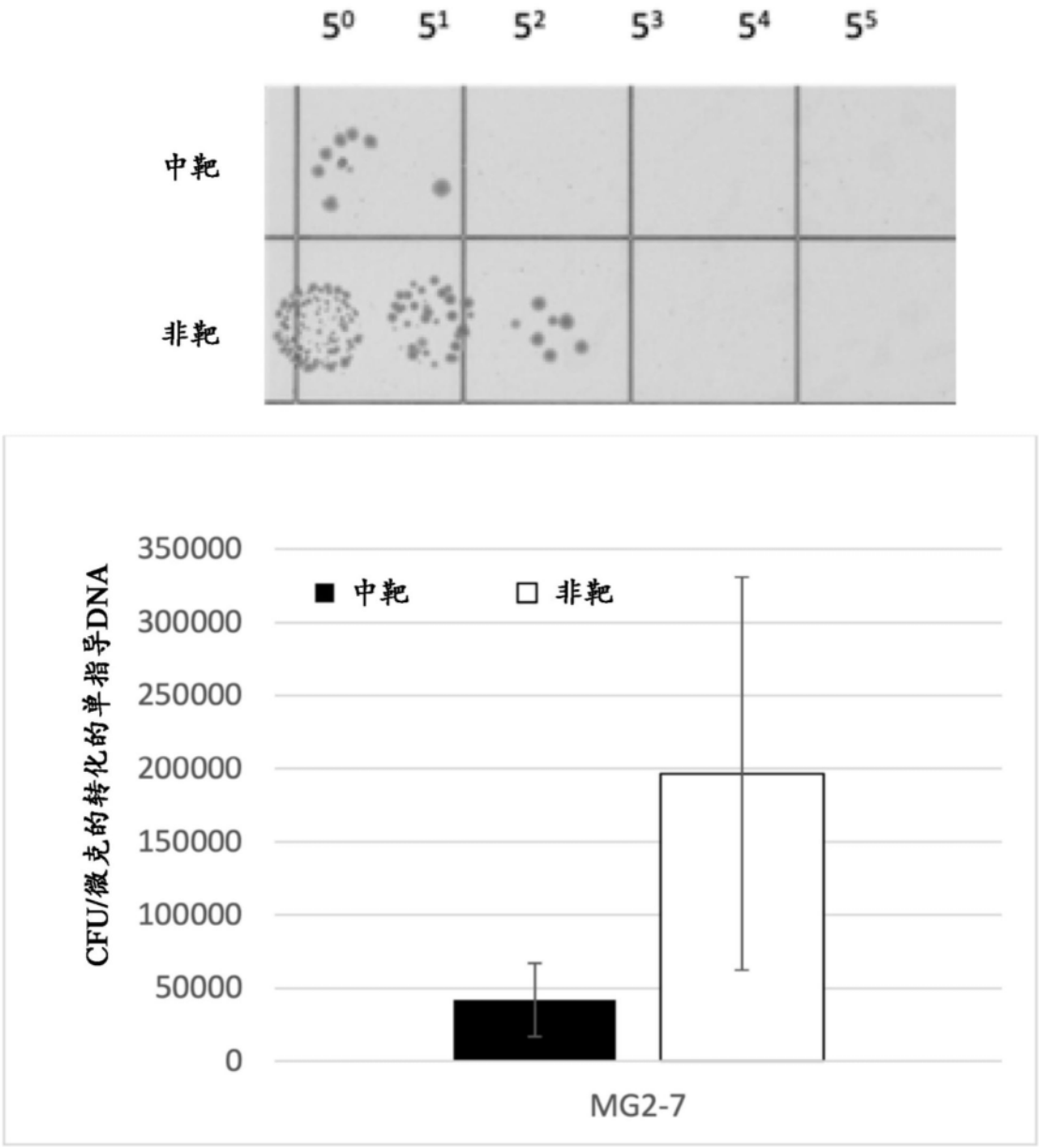
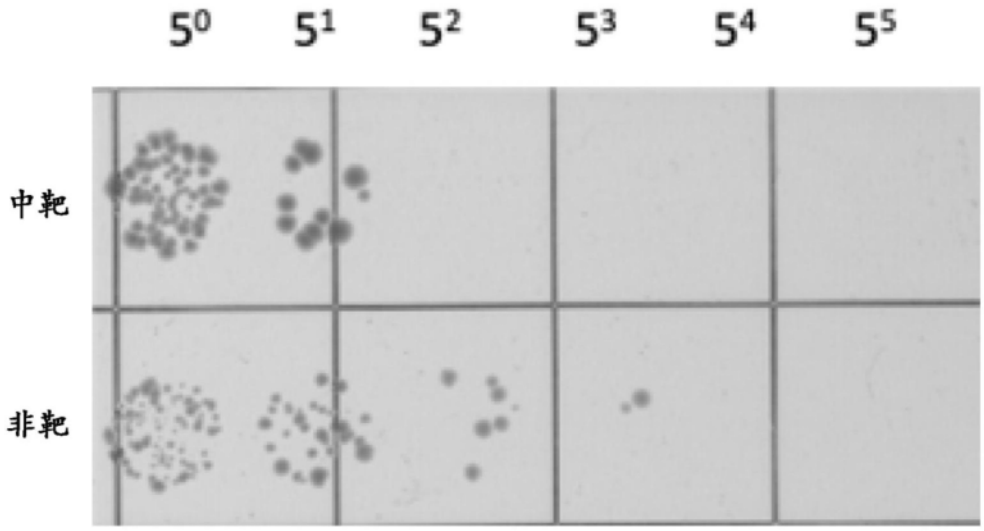

5^0
5^1
5^2
5^3
5^4
5^5
中靶
非靶
CFU/微克的转化的单指导DNA
350000
300000
250000
200000
150000
100000
50000
0
中靶
非靶
MG2-7

图34


5^0
5^1
5^2
5^3
5^4
5^5
中靶
非靶

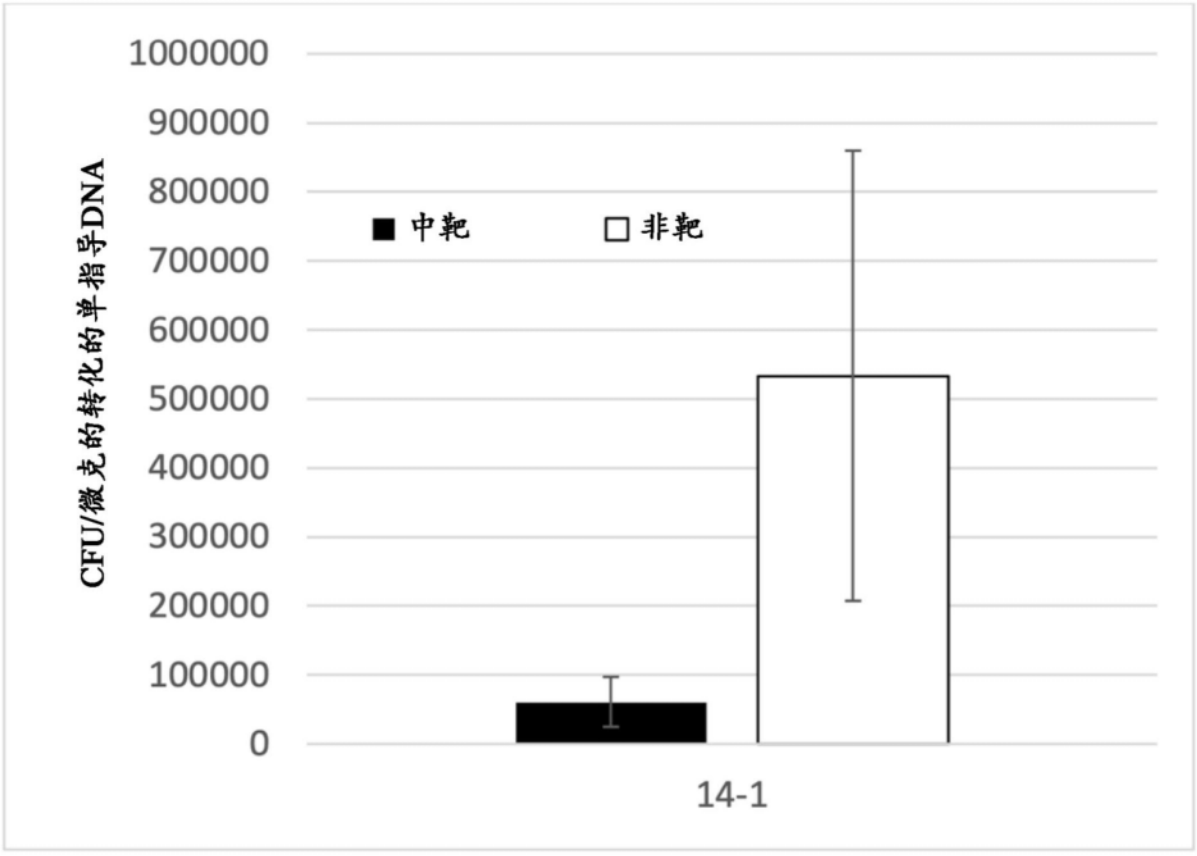
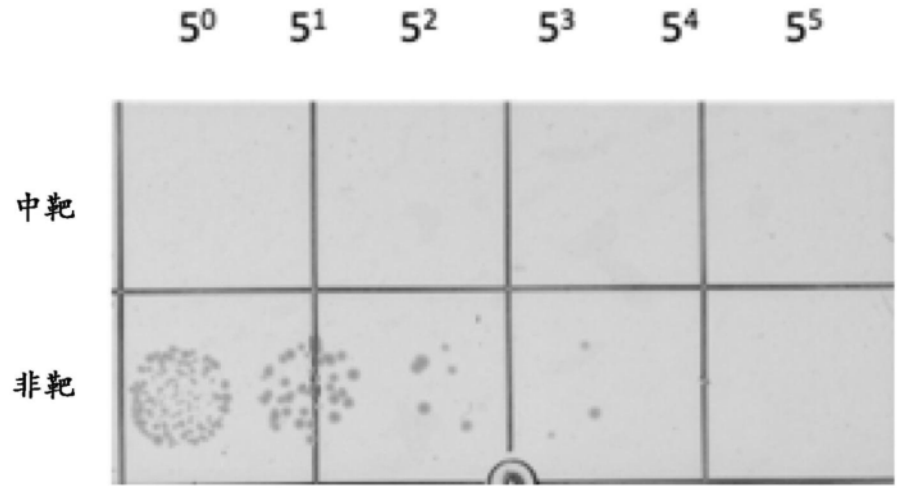

1000000
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
CFU/微克的转化的单指导DNA
中靶
非靶
14-1

图35


5^0
5^1
5^2
5^3
5^4
5^5
中靶
非靶

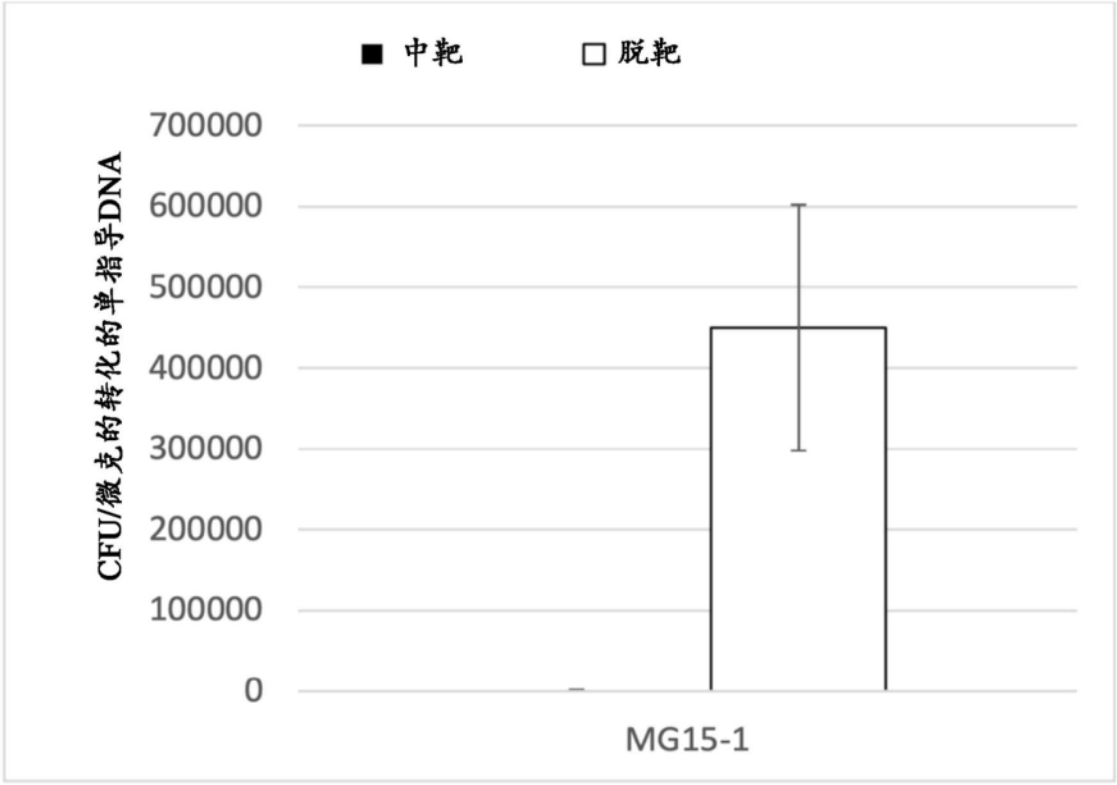

中靶
脱靶
CFU/微克的转化的单指导DNA
700000
600000
500000
400000
300000
200000
100000
0
MG15-1

图36