(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112020554 A
(43)申请公布日 2020.12.01

(21)申请号 201980027641.7

(22)申请日 2019.02.22

(30)优先权数据
62/634257 2018.02.23 US
62/651991 2018.04.03 US

(85)PCT国际申请进入国家阶段日
2020.10.22

(86)PCT国际申请的申请数据
PCT/US2019/019086 2019.02.22

(87)PCT国际申请的公布数据
WO2019/165168 EN 2019.08.29

(71)申请人 先锋国际良种公司
地址 美国依阿华州

(72)发明人 侯正林 J·K·杨 G·加休纳斯 V·斯克斯尼斯

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 翟建伟 黄希贵

(51)Int.Cl.
C12N 9/22(2006.01)
C12N 15/10(2006.01)
C12N 15/113(2006.01)
C12N 15/82(2006.01)

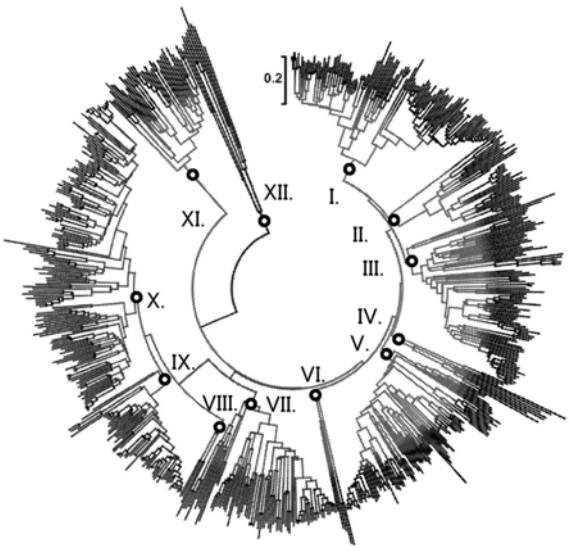
权利要求书12页 说明书138页
序列表（电子公布） 附图26页

(54)发明名称

新颖CAS9直系同源物

(57)摘要

提供了针对新颖Cas9直系同源物的组合物和方法，所述组合物和方法包括但不限于新颖指导多核苷酸/Cas9内切核酸酶复合物、单或双指导RNA、指导RNA元件和Cas9内切核酸酶。本公开还描述了用于在靶多核苷酸中产生双链断裂的方法，用于在细胞的基因组中在各种体内和体外条件下对靶序列进行基因组修饰、用于基因编辑、以及用于将目的多核苷酸插入细胞的基因组中的方法。还提供了具有通过本文描述的方法产生的经修饰的靶位点或改变的目的多核苷酸的核酸构建体和细胞。



CN 112020554 A

1.一种合成的组合物,所述合成的组合物包含异源组分和Cas内切核酸酶,其中所述Cas内切核酸酶包含至少一种选自下组的氨基酸特征,该组由以下组成:

(a)位置13处的异亮氨酸(I),

(b)位置21处的异亮氨酸(I),

(c)位置71处的亮氨酸(L),

(d)位置149处的亮氨酸(L),

(e)位置150处的丝氨酸(S),

(f)位置444处的亮氨酸(L),

(g)位置445处的苏氨酸(T),

(h)位置503处的脯氨酸(P),

(i)位置587处的F(苯丙氨酸),

(j)位置620处的A(丙氨酸),

(k)位置623处的L(亮氨酸),

(l)位置624处的T(苏氨酸),

(m)位置632处的I(异亮氨酸),

(n)位置692处的Q(谷氨酰胺),

(o)位置702处的L(亮氨酸),

(p)位置781处的I(异亮氨酸),

(q)位置810处的K(赖氨酸),

(r)位置908处的L(亮氨酸),

(s)位置931处的V(缬氨酸),

(t)位置933处的N/Q(天冬酰胺或谷氨酰胺),

(u)位置954处的K(赖氨酸),

(v)位置955处的V(缬氨酸),

(w)位置1000处的K(赖氨酸),

(x)位置1100处的V(缬氨酸),

(y)位置1232处的Y(酪氨酸),以及

(z)位置1236处的I(异亮氨酸);

其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的。

2.如权利要求1所述的合成的组合物,其中所述Cas内切核酸酶与选自由SEQ ID NO:86-170和511-1135组成的组的序列具有至少90%同一性。

3.如权利要求1所述的合成的组合物,其中所述Cas内切核酸酶包含与SEQ ID NO:1136-1730中任一个具有90%或更高同一性的结构域。

4.如权利要求1所述的合成的组合物,其中所述Cas内切核酸酶与异源多肽融合。

5.如权利要求4所述的合成的组合物,其中所述异源多肽包含核酸酶活性。

6.如权利要求4所述的合成的组合物,其中所述异源多肽是脱氨酶。

7.如权利要求1所述的合成的组合物,其进一步包含指导多核苷酸,所述多肽与所述指导多核苷酸形成复合物。

8.如权利要求2所述的合成的组合物,其中所述指导多核苷酸是单指导物,所述单指导

物包含选自由SEQ ID NO:426-510组成的组的序列。

9.如权利要求2所述的合成的组合物,其中所述指导多核苷酸包含tracrRNA,所述tracrRNA包含选自由SEQ ID NO:341-425组成的组的序列。

10.如权利要求2所述的合成的组合物,其中所述指导多核苷酸包含crRNA,所述crRNA包含选自由SEQ ID NO:171-255组成的组的序列。

11.如权利要求2所述的合成的组合物,其中所述指导多核苷酸包含反重复序列,所述反重复序列包含选自由SEQ ID NO:256-340组成的组的序列。

12.如权利要求2所述的合成的组合物,其中所述指导多核苷酸指导物包含DNA。

13.如权利要求1所述的合成的组合物,其与表4-83中列出的PAM共有序列选择性杂交。

14.一种Cas内切核酸酶或失活的Cas内切核酸酶,所述Cas内切核酸酶或失活的Cas内切核酸酶识别选自下组的PAM,该组由以下组成:NAR(G>A)WH(A>T>C)GN(C>T>R)、N(C>D)V(A>S)R(G>A)TTTN(T>V)、NV(A>G>C)TTTTT、NATTTTT、NN(H>G)AAAN(G>A>Y)N、N(T>V)NAAATN、NAV(A>G>C)TCNN、NN(A>S>T)NN(W>G>C)CCN(Y>R)、NNAH(T>M)ACN、NGTGANN、NARN(A>K>C)ATN、NV(G>A>C)RNTTN、NN(A>B)RN(A>G>T>C)CCN、NN(A>B)NN(T>V)CCH(A>Y)、NNN(H>G)NCDAA、NN(H>G)D(A>K)GGDN(A>B)、NNNNCCAG、NNNNCTAA、NNNNCVGANN、N(C>D)NNTCCN、NNNNCTA、NNNNCYAA、NAGRGNY、NNGH(W>C)AAA、NNGAAAN、NNAAAAA、NTGAR(G>A)N(A>Y>G)N(Y>R)、N(C>D)H(C>W)GH(Y>A)N(A>B)AN(A>T>S)、NNAAACN、NNGTAM(A>C)Y、NH(A>Y)ARNN(C>W>G)N、B(C>K)GGN(A>Y>G)N NN、N(T>C>R)AGAN(A>K>C)NN、NGGN(A>T>G>C)NNN、NGGD(A>T>G)TNN、NGGAN(T>A>C>G)NN、CGGWN(T>R>C)NN、NGGWGNN、N(B>A)GGNN(T>V)NN、NNGD(A>T>G)AY(T>C)N、N(T>V)H(T>C>A)AAAAN、NRTAANN、N(H>G)CAAH(Y>A)N(Y>R)N、NATAAN(A>T>S)N、NV(A>G>C)R(A>G)ACCN、CN(C>W>G)AV(A>S)GAC、NNRNCAC、N(A>B)GGD(W>G)D(G>W)NN、BGD(G>W)GTCN(A>K>C)、NAANACN、NRTHAN(A>B)N、BHN(H>G)NGN(T>M)H(Y>A)、NMRN(A>Y>G)AH(C>T>A)N、NNNCACN、NARN(T>A>S)ACN、NNNNATW、NGCNGCN、NNNCATN、NAGNGCN、NARN(T>M>G)CCN、NATCCTN、NRTAAN(T>A>S)N、N(C>T>G>A)AAD(A>G>T)CNN、NAAAGNN、NNGACNN、N(T>V)NTAAD(A>T>G)N、NNGAD(G>W)NN、NGGN(W>S)NNN、N(T>V)GGD(W>G)GNN、NGGD(A>T>G)N(T>M>G)NN、NNAAAGN、N(G>H)GGDN(T>M>G)NN、NNAGAAA、NN(T>M>G)AAAAA、N(C>D)N(C>W>G)GW(T>C)D(A>G>T)AA、NAAAAYN、NRGNNNN、NATGN(H>G)TN、NNDATTT和NATARCN(C>T>A>G)。

15.如权利要求1所述的合成的组合物,其是从表1中列出的生物体鉴定的。

16.如权利要求1所述的合成的组合物,其选自由SEQ ID NO:86-170组成的组。

17.如权利要求1所述的合成的组合物,其中靶细胞优化的多肽缺乏内切核酸酶活性。

18.如权利要求1所述的合成的组合物,其中靶细胞优化的多肽能够使单链靶多核苷酸产生切口。

19.如权利要求1所述的合成的组合物,其中靶细胞优化的多肽能够切割双链靶多核苷酸。

20.如权利要求1所述的合成的组合物,其进一步包含供体DNA分子。

21.如权利要求1所述的合成的组合物,其进一步包含修复模板DNA分子。

22.如权利要求1所述的合成的组合物,其中所述异源组合物选自由以下组成的组:异

源多核苷酸、异源多肽、粒子、固体基质、抗体、缓冲液组合物、Tris、EDTA、二硫苏糖醇(DTT)、磷酸盐缓冲盐水(PBS)、氯化钠、氯化镁、HEPES、甘油、牛血清白蛋白(BSA)、盐、乳化剂、洗涤剂、螯合剂、氧化还原剂、抗体、无核酸酶的水、粘度剂和组氨酸标签。

23.如权利要求22所述的合成的组合物,其进一步包含另外的异源组合物。

24.如权利要求1所述的合成的组合物,其进一步包含细胞。

25.如权利要求24所述的合成的组合物,其中所述细胞从选自下组的生物体获得或衍生,该组由以下组成:人、非人灵长类、哺乳动物、动物、古细菌、细菌、原生生物、真菌、昆虫、酵母、非常规酵母和植物。

26.如权利要求25所述的合成的组合物,其中所述植物细胞获得自或衍生自玉蜀黍、稻、高粱、黑麦、大麦、小麦、粟、燕麦、甘蔗、草坪草、柳枝稷、大豆、卡诺拉油菜、苜蓿、向日葵、棉花、烟草、花生、马铃薯、烟草、拟南芥属(Arabidopsis)、蔬菜或红花。

27.如权利要求25所述的合成的组合物,其中所述动物细胞选自下组,该组由以下组成:单倍体细胞、二倍体细胞、生殖细胞、神经元、肌肉细胞、内分泌或外分泌细胞、上皮细胞、肌肉细胞、肿瘤细胞、胚胎细胞、造血细胞、骨细胞、种质细胞、体细胞、干细胞、多能干细胞、诱导多能干细胞、祖细胞、减数分裂细胞和有丝分裂细胞。

28.一种多核苷酸,其编码如权利要求1所述的多肽。

29.如权利要求28所述的多核苷酸,其中所述多核苷酸包含在载体中,所述载体进一步包含至少一种异源多核苷酸。

30.一种试剂盒,其包含如权利要求1所述的合成的组合物或如权利要求28所述的多核苷酸。

31.如权利要求1所述的合成的组合物,其中所述多肽在液体制剂中。

32.如权利要求1所述的合成的组合物,其中所述多肽在冻干组合物中。

33.如权利要求1所述的合成的组合物,其中所述多肽在基本上无内毒素的制剂中。

34.如权利要求1所述的合成的组合物,其中所述多肽在具有以下pH的制剂中:1.0至14.0、2.0至13.0、3.0至12.0、4.0至11.0、5.0至10.0、6.0至9.0、7.0至8.0、4.5至6.5、5.5至7.5、或6.5至7.5。

35.如权利要求1所述的合成的组合物,其中所述多肽在以下温度储存或孵育:至少负200摄氏度、至少负150摄氏度、至少负135摄氏度、至少负90摄氏度、至少负80摄氏度、至少负20摄氏度、至少4摄氏度、至少17摄氏度、至少20摄氏度、至少25摄氏度、至少30摄氏度、至少35摄氏度、至少37摄氏度、至少39摄氏度、至少40摄氏度、至少45摄氏度、至少50摄氏度、至少55摄氏度、至少60摄氏度、至少65摄氏度、至少70摄氏度或大于70摄氏度。

36.如权利要求1所述的合成的组合物,其中所述多肽附接至固体基质。

37.如权利要求36所述的合成的组合物,其中所述固体基质是粒子。

38.一种检测靶多核苷酸序列的方法,所述方法包括:

(a)获得所述靶多核苷酸,

(b)在反应容器中组合Cas内切核酸酶、指导多核苷酸和所述靶多核苷酸,

(c)在至少10摄氏度的温度下孵育步骤(b)的组分至少1分钟,

(d)对反应混合物中的所得的一种或多种多核苷酸进行测序,并且

(e)表征由所述Cas内切核酸酶和所述指导多核苷酸鉴定的步骤(a)的靶多核苷酸的序

列；

(f)其中所述指导多核苷酸包含与所述靶多核苷酸的序列基本互补的多核苷酸序列；

其中所述Cas内切核酸酶包含至少一种选自下组的氨基酸特征，该组由以下组成：

(i)位置13处的异亮氨酸(I)，

(ii)位置21处的异亮氨酸(I)，

(iii)位置71处的亮氨酸(L)，

(iv)位置149处的亮氨酸(L)，

(v)位置150处的丝氨酸(S)，

(vi)位置444处的亮氨酸(L)，

(vii)位置445处的苏氨酸(T)，

(viii)位置503处的脯氨酸(P)，

(ix)位置587处的F(苯丙氨酸)，

(x)位置620处的A(丙氨酸)，

(xi)位置623处的L(亮氨酸)，

(xii)位置624处的T(苏氨酸)，

(xiii)位置632处的I(异亮氨酸)，

(xiv)位置692处的Q(谷氨酰胺)，

(xv)位置702处的L(亮氨酸)，

(xvi)位置781处的I(异亮氨酸)，

(xvii)位置810处的K(赖氨酸)，

(xviii)位置908处的L(亮氨酸)，

(xix)位置931处的V(缬氨酸)，

(xx)位置933处的N/Q(天冬酰胺或谷氨酰胺)，

(xxi)位置954处的K(赖氨酸)，

(xxii)位置955处的V(缬氨酸)，

(xxiii)位置1000处的K(赖氨酸)，

(xxiv)位置1100处的V(缬氨酸)，

(xxv)位置1232处的Y(酪氨酸)，以及

(xxvi)位置1236处的I(异亮氨酸)；

其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的。

39.一种将Cas内切核酸酶和指导多核苷酸复合物结合至靶多核苷酸的方法，所述方法包括：

(a)获得所述靶多核苷酸的序列，

(b)在反应容器中组合Cas内切核酸酶、指导多核苷酸和所述靶多核苷酸，

(c)在至少10摄氏度的温度下孵育步骤(b)的组分至少1分钟；

其中所述指导多核苷酸包含与所述靶多核苷酸的靶多核苷酸序列基本互补的多核苷酸序列；所述方法进一步包括检测与所述靶多核苷酸结合的所述Cas内切核酸酶和指导多核苷酸复合物；并且其中所述Cas内切核酸酶包含至少一种选自下组的氨基酸特征，该组由以下组成：

(i)位置13处的异亮氨酸(I)，

(ii)位置21处的异亮氨酸(I)，

(iii)位置71处的亮氨酸(L)，

(iv)位置149处的亮氨酸(L)，

(v)位置150处的丝氨酸(S)，

(vi)位置444处的亮氨酸(L)，

(vii)位置445处的苏氨酸(T)，

(viii)位置503处的脯氨酸(P)，

(ix)位置587处的F(苯丙氨酸)，

(x)位置620处的A(丙氨酸)，

(xi)位置623处的L(亮氨酸)，

(xii)位置624处的T(苏氨酸)，

(xiii)位置632处的I(异亮氨酸)，

(xiv)位置692处的Q(谷氨酰胺)，

(xv)位置702处的L(亮氨酸)，

(xvi)位置781处的I(异亮氨酸)，

(xvii)位置810处的K(赖氨酸)，

(xviii)位置908处的L(亮氨酸)，

(xix)位置931处的V(缬氨酸)，

(xx)位置933处的N/Q(天冬酰胺或谷氨酰胺)，

(xxi)位置954处的K(赖氨酸)，

(xxii)位置955处的V(缬氨酸)，

(xxiii)位置1000处的K(赖氨酸)，

(xxiv)位置1100处的V(缬氨酸)，

(xxv)位置1232处的Y(酪氨酸)，以及

(xxvi)位置1236处的I(异亮氨酸)；

其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的。

40.一种在靶多核苷酸中产生双链断裂的方法，所述方法包括：

(a)获得所述靶多核苷酸的序列，

(b)在反应容器中组合Cas内切核酸酶多肽、指导多核苷酸和所述靶多核苷酸，

(c)在至少10摄氏度的温度下孵育步骤(b)的组分至少1分钟；

其中所述指导多核苷酸包含与所述靶多核苷酸的靶多核苷酸序列基本互补的多核苷酸序列；所述方法进一步包括检测与所述靶多核苷酸结合的所述Cas内切核酸酶和指导多核苷酸复合物；并且其中所述Cas内切核酸酶包含至少一种选自下组的氨基酸特征，该组由以下组成：

(i)位置13处的异亮氨酸(I)，

(ii)位置21处的异亮氨酸(I)，

(iii)位置71处的亮氨酸(L)，

(iv)位置149处的亮氨酸(L)，

(v)位置150处的丝氨酸(S)，

(vi)位置444处的亮氨酸(L)，

(vii)位置445处的苏氨酸(T)，

(viii)位置503处的脯氨酸(P)，

(ix)位置587处的F(苯丙氨酸)，

(x)位置620处的A(丙氨酸)，

(xi)位置623处的L(亮氨酸)，

(xii)位置624处的T(苏氨酸)，

(xiii)位置632处的I(异亮氨酸)，

(xiv)位置692处的Q(谷氨酰胺)，

(xv)位置702处的L(亮氨酸)，

(xvi)位置781处的I(异亮氨酸)，

(xvii)位置810处的K(赖氨酸)，

(xviii)位置908处的L(亮氨酸)，

(xix)位置931处的V(缬氨酸)，

(xx)位置933处的N/Q(天冬酰胺或谷氨酰胺)，

(xxi)位置954处的K(赖氨酸)，

(xxii)位置955处的V(缬氨酸)，

(xxiii)位置1000处的K(赖氨酸)，

(xxiv)位置1100处的V(缬氨酸)，

(xxv)位置1232处的Y(酪氨酸)，以及

(xxvi)位置1236处的I(异亮氨酸)；

其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的。

41.如权利要求39或权利要求40所述的方法，其进一步包括至少一个另外的靶位点。

42.一种用于编辑细胞的基因组的方法，所述方法包括向所述细胞提供：

(a)至少一种Cas内切核酸酶，其包含至少一种选自下组的氨基酸特征，该组由以下组成：

(i)位置13处的异亮氨酸(I)，

(ii)位置21处的异亮氨酸(I)，

(iii)位置71处的亮氨酸(L)，

(iv)位置149处的亮氨酸(L)，

(v)位置150处的丝氨酸(S)，

(vi)位置444处的亮氨酸(L)，

(vii)位置445处的苏氨酸(T)，

(viii)位置503处的脯氨酸(P)，

(ix)位置587处的F(苯丙氨酸)，

(x)位置620处的A(丙氨酸)，

(xi)位置623处的L(亮氨酸)，

(xii)位置624处的T(苏氨酸)，

(xiii)位置632处的I(异亮氨酸)，

(xiv)位置692处的Q(谷氨酰胺)，

(xv)位置702处的L(亮氨酸)，

(xvi)位置781处的I(异亮氨酸)，

(xvii)位置810处的K(赖氨酸)，

(xviii)位置908处的L(亮氨酸)，

(xix)位置931处的V(缬氨酸)，

(xx)位置933处的N/Q(天冬酰胺或谷氨酰胺)，

(xxi)位置954处的K(赖氨酸)，

(xxii)位置955处的V(缬氨酸)，

(xxiii)位置1000处的K(赖氨酸)，

(xxiv)位置1100处的V(缬氨酸)，

(xxv)位置1232处的Y(酪氨酸)，以及

(xxvi)位置1236处的I(异亮氨酸)；

其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的；和

(b)指导多核苷酸，所述Cas内切核酸酶与所述指导多核苷酸形成复合物；

其中所述复合物能够识别、结合靶多核苷酸序列并任选地使靶多核苷酸序列产生切口或切割靶多核苷酸序列；并且鉴定在所述细胞的基因组DNA序列中具有修饰的至少一个细胞，其中所述修饰选自由以下组成的组：对现有核苷酸插入、缺失、取代以及添加或缔合原子或分子。

43.一种调节细胞中基因的表达的方法，所述方法包括向所述细胞提供：

(a)至少一种Cas内切核酸酶，其包含至少一种选自下组的氨基酸特征，该组由以下组成：

(i)位置13处的异亮氨酸(I)，

(ii)位置21处的异亮氨酸(I)，

(iii)位置71处的亮氨酸(L)，

(iv)位置149处的亮氨酸(L)，

(v)位置150处的丝氨酸(S)，

(vi)位置444处的亮氨酸(L)，

(vii)位置445处的苏氨酸(T)，

(viii)位置503处的脯氨酸(P)，

(ix)位置587处的F(苯丙氨酸)，

(x)位置620处的A(丙氨酸)，

(xi)位置623处的L(亮氨酸)，

(xii)位置624处的T(苏氨酸)，

(xiii)位置632处的I(异亮氨酸)，

(xiv)位置692处的Q(谷氨酰胺)，

(xv)位置702处的L(亮氨酸)，

(xvi)位置781处的I(异亮氨酸)，

(xvii)位置810处的K(赖氨酸)，

(xviii)位置908处的L(亮氨酸)，

(xix)位置931处的V(缬氨酸)，

(xx)位置933处的N/Q(天冬酰胺或谷氨酰胺)，

(xxi)位置954处的K(赖氨酸)，

(xxii)位置955处的V(缬氨酸)，

(xxiii)位置1000处的K(赖氨酸)，

(xxiv)位置1100处的V(缬氨酸)，

(xxv)位置1232处的Y(酪氨酸)，以及

(xxvi)位置1236处的I(异亮氨酸)；

其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的，和

(b)指导多核苷酸，所述Cas内切核酸酶与所述指导多核苷酸形成复合物；

其中所述复合物能够识别、结合所述细胞中的靶多核苷酸序列并任选地使所述细胞中的靶多核苷酸序列产生切口或切割所述细胞中的靶多核苷酸序列；以及

鉴定与未引入所述Cas内切核酸酶的细胞相比具有调节的基因表达的至少一个细胞。

44.如权利要求42或权利要求43所述的方法，其进一步包括向所述细胞提供供体DNA分子。

45.如权利要求42或权利要求43所述的方法，其进一步包括向所述细胞提供模板DNA分子。

46.如权利要求42或权利要求43所述的方法，其中所述方法赋予所述细胞或包含所述细胞的生物体益处。

47.如权利要求41所述的方法，其中所述益处选自由以下组成的组：改善的健康、改善的生长、改善的能育性、改善繁殖力、改善的环境耐受、改善的活力、改善的疾病抗性、改善的疾病耐受、改善的对异源分子的耐受、改善的适应性、改善的物理特征、更大的质量、增加的生化分子产生、减少的生化分子产生、基因的上调、基因的下调、生化途径的上调、生化途径的下调、细胞繁殖的刺激和细胞繁殖的抑制。

48.如权利要求42或权利要求43所述的方法，其中所述细胞与衍生所述Cas内切核酸酶的生物体是异源的，并且选自由以下组成的组：人、非人灵长类、哺乳动物、动物、古细菌、细菌、原生生物、真菌、昆虫、酵母、非常规酵母和植物细胞。

49.如权利要求48所述的方法，其中所述植物细胞获得自或衍生自玉蜀黍、稻、高粱、黑麦、大麦、小麦、粟、燕麦、甘蔗、草坪草、柳枝稷、大豆、卡诺拉油菜、苜蓿、向日葵、棉花、烟草、花生、马铃薯、烟草、拟南芥属、蔬菜或红花。

50.如权利要求48所述的方法，其中所述细胞是植物细胞，并且所述益处是调节包含所述细胞或其后代细胞的植物的具有农艺学意义的性状，所述具有农艺学意义的性状选自由以下组成的组：疾病抗性、干旱抗性、热耐性、寒耐性、盐耐性、金属耐性、除草剂耐性、改善的水分利用效率、改善的氮利用率、改善的固氮作用、有害生物抗性、食草动物抗性、病原体抗性、产率改善、健康增强、改善的能育性、活力改善、生长改善、光合能力改善、营养增强、改变的蛋白含量、改变的油含量、增加的生物量、增加的芽长度、增加的根长度、改善的根结构、代谢产物的调节、蛋白质组的调节、增加的种子重量、改变的种子碳水化合物组成、改变

的种子油组成、改变的种子蛋白组成、改变的种子营养物组成;如与不包含所述靶位点修饰的同系植物(isoline plant)相比,或与所述植物细胞中所述靶位点的修饰之前的植物相比。

51.如权利要求48所述的方法,其中所述动物细胞选自下组,该组由以下组成:单倍体细胞、二倍体细胞、生殖细胞、神经元、肌肉细胞、内分泌或外分泌细胞、上皮细胞、肌肉细胞、肿瘤细胞、胚胎细胞、造血细胞、骨细胞、种质细胞、体细胞、干细胞、多能干细胞、诱导多能干细胞、祖细胞、减数分裂细胞和有丝分裂细胞。

52.如权利要求48所述的方法,其中所述细胞是动物细胞并且所述益处是调节包含所述动物细胞或其后代细胞的生物体的具有生理学意义的表型,所述具有生理学意义的表型选自由以下组成的组:改善的健康、改善的营养状况、减少的疾病影响、疾病静止状态、疾病逆转、改善的能育性、改善的活力、改善的心智能力、改善的生物体生长、改善的增重、减重、内分泌系统的调节、外分泌系统的调节、减小的肿瘤大小、减小的肿瘤质量、刺激的细胞生长、降低的细胞生长、代谢产物的产生、激素的产生、免疫细胞的产生、以及刺激细胞产生。

53.一种编辑靶多核苷酸的至少一个碱基的方法,所述方法包括:

(a)使所述靶多核苷酸与以下接触:

i.脱氨酶,

ii.能够与表4-83中列出的PAM共有序列选择性杂交的Cas内切核酸酶,其中所述Cas内切核酸酶已被修饰为缺乏核酸酶活性,和

iii.与所述靶多核苷酸的序列具有互补性的指导多核苷酸,

其中所述Cas内切核酸酶和所述指导RNA形成识别并结合所述靶多核苷酸的复合物;并且

(b)检测在DNA靶位点处的至少一个修饰。

54.一种编辑靶多核苷酸的多个碱基的方法,所述方法包括:

(a)使所述靶多核苷酸与以下接触:

i.至少一种脱氨酶,

ii.多种Cas内切核酸酶,每种能够与表4-83中列出的PAM共有序列选择性杂交,其中所述Cas内切核酸酶已被修饰为缺乏核酸酶活性,和

iii.与所述靶多核苷酸的序列具有互补性的指导多核苷酸,

其中所述Cas内切核酸酶和所述指导RNA形成识别并结合所述靶多核苷酸的复合物;并且

(b)检测在DNA靶位点处的至少一个修饰。

55.一种优化Cas分子的活性的方法,所述方法包括将至少一个核苷酸修饰引入包含至少一种选自下组的氨基酸特征的序列,该组由以下组成:

(a)位置13处的异亮氨酸(I),

(b)位置21处的异亮氨酸(I),

(c)位置71处的亮氨酸(L),

(d)位置149处的亮氨酸(L),

(e)位置150处的丝氨酸(S),

(f)位置444处的亮氨酸(L),

(g)位置445处的苏氨酸(T)，

(h)位置503处的脯氨酸(P)，

(i)位置587处的F(苯丙氨酸)，

(j)位置620处的A(丙氨酸)，

(k)位置623处的L(亮氨酸)，

(l)位置624处的T(苏氨酸)，

(m)位置632处的I(异亮氨酸)，

(n)位置692处的Q(谷氨酰胺)，

(o)位置702处的L(亮氨酸)，

(p)位置781处的I(异亮氨酸)，

(q)位置810处的K(赖氨酸)，

(r)位置908处的L(亮氨酸)，

(s)位置931处的V(缬氨酸)，

(t)位置933处的N/Q(天冬酰胺或谷氨酰胺)，

(u)位置954处的K(赖氨酸)，

(v)位置955处的V(缬氨酸)，

(w)位置1000处的K(赖氨酸)，

(x)位置1100处的V(缬氨酸)，

(y)位置1232处的Y(酪氨酸)，以及

(z)位置1236处的I(异亮氨酸)；

其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的；

并且与核苷酸修饰之前的分子相比，鉴定至少一种改善的特征。

56.一种通过以下来优化Cas9分子的活性的方法：使亲本Cas9分子经历至少一轮随机蛋白改组，并选择具有至少一种不存在于所述亲本Cas9分子中的特征的所得分子；其中所述亲本Cas9分子包含至少一种选自下组的氨基酸特征，该组由以下组成：

(a)位置13处的异亮氨酸(I)，

(b)位置21处的异亮氨酸(I)，

(c)位置71处的亮氨酸(L)，

(d)位置149处的亮氨酸(L)，

(e)位置150处的丝氨酸(S)，

(f)位置444处的亮氨酸(L)，

(g)位置445处的苏氨酸(T)，

(h)位置503处的脯氨酸(P)，

(i)位置587处的F(苯丙氨酸)，

(j)位置620处的A(丙氨酸)，

(k)位置623处的L(亮氨酸)，

(l)位置624处的T(苏氨酸)，

(m)位置632处的I(异亮氨酸)，

(n)位置692处的Q(谷氨酰胺)，

(o)位置702处的L(亮氨酸)，

(p)位置781处的I(异亮氨酸)，

(q)位置810处的K(赖氨酸)，

(r)位置908处的L(亮氨酸)，

(s)位置931处的V(缬氨酸)，

(t)位置933处的N/Q(天冬酰胺或谷氨酰胺)，

(u)位置954处的K(赖氨酸)，

(v)位置955处的V(缬氨酸)，

(w)位置1000处的K(赖氨酸)，

(x)位置1100处的V(缬氨酸)，

(y)位置1232处的Y(酪氨酸)，以及

(z)位置1236处的I(异亮氨酸)；

其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的。

57.一种通过以下来优化Cas9分子的活性的方法：使亲本Cas9分子经历至少一轮非随机蛋白改组，并选择具有至少一种不存在于所述亲本Cas9分子中的特征的所得分子；其中所述亲本Cas9分子包含基序，所述基序选自由以下组成的组：包含至少一种选自下组的氨基酸特征，该组由以下组成：

(a)位置13处的异亮氨酸(I)，

(b)位置21处的异亮氨酸(I)，

(c)位置71处的亮氨酸(L)，

(d)位置149处的亮氨酸(L)，

(e)位置150处的丝氨酸(S)，

(f)位置444处的亮氨酸(L)，

(g)位置445处的苏氨酸(T)，

(h)位置503处的脯氨酸(P)，

(i)位置587处的F(苯丙氨酸)，

(j)位置620处的A(丙氨酸)，

(k)位置623处的L(亮氨酸)，

(l)位置624处的T(苏氨酸)，

(m)位置632处的I(异亮氨酸)，

(n)位置692处的Q(谷氨酰胺)，

(o)位置702处的L(亮氨酸)，

(p)位置781处的I(异亮氨酸)，

(q)位置810处的K(赖氨酸)，

(r)位置908处的L(亮氨酸)，

(s)位置931处的V(缬氨酸)，

(t)位置933处的N/Q(天冬酰胺或谷氨酰胺)，

(u)位置954处的K(赖氨酸)，

(v)位置955处的V(缬氨酸)，

(w)位置1000处的K(赖氨酸)，

(x)位置1100处的V(缬氨酸)，

(y)位置1232处的Y(酪氨酸)，以及

(z)位置1236处的I(异亮氨酸)；

其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的。

新颖CAS9直系同源物

[0001] 相关申请的交叉引用

[0002] 本申请要求于2018年2月23日提交的美国临时申请号62/634,257和2018年4月3日提交的美国临时申请号62/651,991的权益,其两者通过引用以全部内容结合在此。

技术领域

[0003] 本公开涉及分子生物学领域,尤其涉及具有指导多核苷酸/内切核酸酶系统的组合物,以及修饰多核苷酸序列(包括细胞基因组)的组合物和方法。

[0004] 以电子方式递交的序列表的引用

[0005] 序列表的正式副本作为ASCII格式的序列表(其文件名称是RTS26814AWOPCT_SequenceListing_ST25.txt,创建于2019年2月21日,并且大小为8,870,697字节)经由EFS-Web以电子方式提交,并且与说明书同时提交。所述ASCII格式的文档中包含的序列表是说明书的一部分,并且通过引用以全部内容结合在此。

背景技术

[0006] 重组DNA技术使得在靶基因组位置处插入DNA序列和/或修饰特定内源染色体序列成为可能。已经使用了采用位点特异性重组系统的位点特异性整合技术以及其他类型的重组技术来在各种生物体中产生目的基因的靶向插入。基因组编辑技术如锌指核酸酶(ZFN)、转录激活子样效应子核酸酶(TALEN)或归巢大范围核酸酶可以用于产生靶向基因组干扰,但这些系统倾向于具有低特异性并且使用需要对每个靶位点进行重新设计的核酸酶,这使得它们的制备成本高昂且耗时。

[0007] 已经鉴定了利用古细菌或细菌适应性免疫系统的较新技术,称为CRISPR(成簇的规律间隔的短回文重复序列(Clustered Regularly Interspaced Short Palindromic Repeats)),其包含效应子蛋白的不同结构域,所述效应子蛋白包含多种活性(DNA识别、结合和任选择地切割)。

[0008] 尽管已经鉴定和表征了这些系统中的一些,但仍需要鉴定新颖效应子和系统,以及证明在真核生物,特别是动植物中的活性,以实现内源和先前引入的异源多核苷酸的编辑以及体外多核苷酸结合和/或修饰。大多数CRISPR基因编辑几乎全部基于衍生自酿脓链球菌(Streptococcus pyogenes)的Cas9系统(Barrangou和Doudna,2016),所述系统通过识别靶多核苷酸上“NGG”的前间隔子邻近基序(PAM)序列,留下了平末端突出并实现基因编辑。期望具有不同生物物理和生物化学特征(包括不同的PAM识别序列)的Cas9蛋白的更大多样性。

发明内容

[0009] 提供了针对新颖Cas多核苷酸和cas多肽的组合物和方法。

[0010] 在一些方面,本发明提供了合成的组合物,所述合成的组合物包含异源组分和选自由以下组成的组的多核苷酸:与SEQ ID NO:1-85中任何一个的至少50、50至100、至少

100、100至150、至少150、150至200、至少200、200至250、至少250、250至300、至少300、300至350、至少350、350至400、至少400、400至450、至少500、500至550、至少550、550至600、至少600、600至650、至少650、650至700、至少700、700至750、至少750、750至800、至少800、800至850、至少850、850至900、至少900、900至950、至少950、950至1000、至少1000或甚至大于1000个连续核苷酸具有至少80％、80％至85％、至少85％、85％至90％、至少90％、90％至95％、至少95％、至少96％、至少97％、至少98％、至少99％、至少99.5％或大于99.5％同一性的多核苷酸,SEQ ID NO:1-85中任何一个的功能性变体,SEQ ID NO:1-85中任何一个的功能性片段,编码选自由SEQ ID NO:86-171和511-1135组成的组的Cas内切核酸酶的基因,编码识别表4-83中任何一个列出的PAM序列的Cas内切核酸酶的基因,编码Cas内切核酸酶的基因,所述Cas内切核酸酶鉴定自、衍生自或分离自选自下组的生物体,该组由以下组成:醋酸杆菌(Acetobacter aceti)、醋杆菌属物种(Acetobacter sp.)CAG:977、棕榈无胆甾原体(Acholeplasma palmae)、氨基酸球菌属物种(Acidaminococcus sp.)、肠氨基酸球菌(Acidaminococcus_intestini)_RyC-MR95)、解纤维热酸菌(Acidothermus cellulolyticus)、燕麦食酸菌(Acidovorax avenae)、Acidovorax ebreus、食酸菌属物种(Acidovorax sp.)MR-S7、荚膜放线杆菌(Actinobacillus capsulatus)、小放线杆菌(Actinobacillus minor)、琥珀酸放线杆菌(Actinobacillus succinogenes)、猪放线杆菌(Actinobacillus suis)、Actinomyces coleocanis、乔格放线菌(Actinomyces georgiae)、麦氏放线菌(Actinomyces meyeri)、内氏放线菌(Actinomyces naeslundii)、龋齿放线菌(Actinomyces odontolyticus)、放线菌属物种(Actinomyces sp.)ICM47、放线菌属物种口腔分类单元175(Actinomyces sp.oral taxon 175)、放线菌属物种口腔分类单元180(Actinomyces sp.oral taxon 180)、放线菌属物种口腔分类单元181(Actinomyces sp.oral taxon 181)、放线菌属物种口腔分类单元848(Actinomyces sp.oral taxon 848)、放线菌属物种(Actinomyces sp.)S6-Spd3、阿菲波菌属物种(Afipia sp.)P52-10、Akkermansia muciniphila、太平洋食烷菌(Alcanivorax pacificus)、Alicycliphilus、褐脂环酸芽孢杆菌(Alicyclobacillus hesperidum)、Aliiarcobacter faecis、Alistipes ihumii、Alistipes shahii、Alkaliflexus imshenetskii、谭氏拟普雷沃菌(Alloprevotella tannerae)、欧米克斯异斯卡多维亚氏菌(Alloscardovia omnicolens)、α变形杆菌(alpha proteobacterium)AAP38、α变形杆菌(alpha proteobacterium)AAP81b、四联气球菌(Anaerococcus tetradius)、Anaeromusa acidaminophila、厌氧芽孢杆菌物种(Anoxybacillus sp.)P3H1B、小棒水小杆菌(Aquabacterium parvum)、Asinibacterium sp.or53、霍洛芬施固氮螺菌(Azospirillum halopraeferens)、固氮螺菌属物种(Azospirillum sp.)B510、蜡样芽孢杆菌(Bacillus cereus)、胞毒芽孢杆菌(Bacillus cytotoxicus)、尼亚美芽孢杆菌(Bacillus niameyensis)、欧肯斯芽孢杆菌(Bacillus okhensis)、假性嗜水芽孢杆菌(Bacillus pseudalcaliphilus)、史密斯芽孢杆菌(Bacillus smithii)、细菌(bacterium)BRH_c32、细菌(bacterium)LF-3、细菌(bacterium)P3、拟杆菌目细菌CF(Bacteroidales bacterium)CF、拟杆菌属(Bacteroides)、嗜粪拟杆菌(Bacteroides coprophilus)、共产拟杆菌(Bacteroides coprosuis)、粪便拟杆菌(Bacteroides faecis)、福克萨斯拟杆菌(Bacteroides fluxus)、脆弱类杆菌(Bacteroides fragilis)、嗜果胶拟杆菌(Bacteroides pectinophilus)、Bacteroides

propionicifaciens、酿脓拟杆菌(Bacteroides pyogenes)、拟杆菌属物种(Bacteroides sp.)14(A)、Bacteroides timonensis、普通拟杆菌(Bacteroides vulgatus)、拟杆菌门口腔分类单元(Bacteroidetes oral taxon)274、Barnesiella viscericola、艾柯蛭弧菌(Bdellovibrio exovorus)、Belliella baltica、海藻百伯史坦菌(Bibersteinia trehalosi)、角形双歧杆菌(Biffdobacterium angulatum)、两岐双岐杆菌(Biffdobacterium bifidum)、邦比双岐杆菌(Biffdobacterium bombi)、双岐杆菌(Biffdobacterium callitrichos)、长双歧杆菌(Biffdobacterium longum)、墨西卡姆双岐杆菌(Biffdobacterium merycicum)、嗜热双歧杆菌(Biffdobacterium thermophilum)、苏密斯双岐杆菌(Biffdobacterium tsurumiense)、Blastopirellula marina、假兹鲍特菌(Bordetella pseudohinzii)、侧孢短芽孢杆菌(Brevibacillus laterosporus)、成团苔藓杆菌(Bryobacter aggregatus)、伯克氏菌目细菌(Burkholderiales bacterium)GJ-E10、亨氏丁酸弧菌(Butyrivibrio hungatei)、丁酸弧菌属物种(Butyrivibrio sp)AC2005、丁酸弧菌属物种(Butyrivibrio sp)NC3005、Caenispirillum salinarum、结肠弯曲菌(Campylobacter coli)、空肠弯曲菌(Campylobacter jejuni)、佩洛里迪斯弯曲菌(Campylobacter peloridis)、亚南极弯曲菌(Campylobacter subantarcticus)、候选门TA06细菌32111(candidate division TA06 bacterium 32_111)、Brocadia sinica候选种、Hepatoplasma crinochetorum Av候选种、Micropelagos thuwalensis候选种、Symbiothrix dinenymphae候选种、犬碳酸噬胞菌(Capnocytophaga canis)、希诺地米碳酸噬胞菌(Capnocytophaga cynodegmi)、黄褐碳酸噬胞菌(Capnocytophaga ochracea)、碳酸噬胞菌属物种(Capnocytophaga sp.)CM59、碳酸噬胞菌属物种口腔分类单元(Capnocytophaga sp.oral taxon)329、福蒂姆肉食杆菌(Carnobacterium funditum)、鸡肉杆菌(Carnobacterium gallinarum)、肉食杆菌属物种(Carnobacterium sp.)ZWU0011、Caviibacter abscessus、噬几丁质菌科细菌(Chitinophagaceae bacterium)PMP191F、沙眼衣原体(Chlamydia trachomatis)、绿菌门细菌(Chlorobi bacterium)NICIL-2、禽金黄杆菌(Chryseobacterium gallinarum)、产吲哚金黄杆菌(Chryseobacterium indologenes)、金黄杆菌属物种(Chryseobacterium sp)CF314、金黄杆菌属物种(Chryseobacterium sp)ERMR1:04、金黄杆菌属物种(Chryseobacterium sp)FH2、金黄杆菌属物种(Chryseobacterium sp)Hurlbut01、金黄杆菌属物种(Chryseobacterium sp)Leaf201、金黄杆菌属物种(Chryseobacterium sp)Leaf394、金黄杆菌属物种(Chryseobacterium sp)StRB126、金黄杆菌属物种(Chryseobacterium sp)YR485、特纳斯金黄杆菌(Chryseobacterium tenax)、Cloacibacillus evryensis、拜氏梭菌(Clostridium beijerinckii)、肉毒梭菌(Clostridium botulinum)、产气荚膜梭菌(Clostridium perfringens)、梭菌属物种(Clostridium sp.)CAG:230、梭菌属物种(Clostridium sp.)CAG:433、螺旋梭菌(Clostridium spiroforme)、柯林斯氏菌属物种(Collinsella sp.)CAG:289、丛毛单胞菌科细菌(Comamonadaceae bacterium)CCH4-C5、颗粒丛毛单胞菌(Comamonas granuli)、Coprobacter fastidiosus、Coprobacter secundus、猫粪球菌(Coprococcus catus)GD/7、红蝽菌目细菌(Coriobacteriales bacterium)DNF00809、小球科里氏杆菌(Coriobacterium glomerans)、小球科里氏杆菌(Coriobacterium_glomerans)_PW2、棒状杆菌属(Corynebacterium)、拥挤棒状杆菌

(Corynebacterium accolens)、卡泼西斯棒状杆菌(Corynebacterium camporealensis)、卡皮姆棒状杆菌(Corynebacterium caspium)、白喉棒状杆菌(Corynebacterium diphtheriae)、假棒状杆菌(Corynebacterium falsenii)、乳酸棒状杆菌(Corynebacterium lactis)、假白喉棒状杆菌(Corynebacterium pseudodiphtheriticum)、维他密斯棒状杆菌(Corynebacterium vitaeruminis)、Croceitalea dokdonensis、噬纤维菌目细菌(Cytophagales bacterium)B6、脱氮脱氯菌(Dechloromonas denitrificans)、Defluviimonas、Demequina sediminicola、白蚁脱硫弧菌(Desulfovibrio termitidis)、戴沃斯菌属物种(Devosia sp.)Root635、Dielma fastidiosa、Dinoroseobacter shibae、Dorea longicatena、Dysgonomonas sp.HGC4、埃格特菌属物种(Eggerthella sp.)YY7918、埃格特菌属物种(Eggerrhella_sp.)_YY7918、Eggerthellaceae细菌AT8、按蚊脓毒性菌(Elizabethkingia anophelis)、脑膜败血病脓毒性菌(Elizabethkingia meningoseptica)、Elusimicrobium minutum、短稳杆菌(Empedobacter brevis)、假稳杆菌(Empedobacter falsenii)、Endomicrobium proavitum、犬肠球菌(Enterococcus canis)、盲肠肠球菌(Enterococcus cecorum)、殊异肠球菌(Enterococcus dispar)、粪肠球菌(Enterococcus faecalis)、粪肠球菌(Enterococcus faecalis)OG1RF、屎肠球菌(Enterococcus faecium)、海氏肠球菌(Enterococcus hirae)、意大利肠球菌(Enterococcus italicus)、马赛肠球菌(Enterococcus massiliensis)、蒙氏肠球菌(Enterococcus mundtii)、菲欧卡拉肠球菌(Enterococcus phoeniculicola)、假禽肠球菌(Enterococcus pseudoavium)、泰国肠球菌(Enterococcus thailandicus)、环境宏基因组(Environmental metagenome)、细长真杆菌(eubacterium dolichum)、细枝真杆菌(Eubacterium ramulus)、直肠真杆菌(Eubacterium rectale)、真杆菌属物种(Eubacterium sp.)、真杆菌属物种(Eubacterium sp.)CAG:251、凸腹真杆菌(Eubacterium ventriosum)、尤氏真杆菌玛格丽特亚种(Eubacterium yurii subsp.margaretiae)ATCC 43715、人费克蓝姆菌(Facklamia hominis)、产琥珀酸丝状杆菌(Fibrobacter succinogenes)、龈沟产线菌(Filifactor alocis)、大芬戈尔德菌(Finegoldia magna)、大芬戈尔德菌(Finegoldia_magna)_ATCC_29328、厚壁菌门细菌(Firmicutes bacterium)M10-2、阿维恩斯黄杆菌(Flavobacterium akiainvivens)、嗜分支黄杆菌(Flavobacterium branchiophilum)、柱状黄杆菌(Flavobacterium columnare)、大田黄杆菌(Flavobacterium daejeonense)、线状黄杆菌(Flavobacterium filum)、冷黄杆菌(Flavobacterium frigidarium)、嗜冷黄杆菌(Flavobacterium psychrophilum)、黄杆菌属物种(Flavobacterium sp.)83、黄杆菌属物种(Flavobacterium sp.)ACAM 123、黄杆菌属物种(Flavobacterium sp.)TAB 87、赛恩斯黄杆菌(Flavobacterium suncheonense)、Fluviicola taffensis、西班牙弗朗西斯氏菌(Francisella hispaniensis)、费城弗朗西斯氏菌(Francisella philomiragia)、土拉弗朗西斯氏菌(Francisella tularensis)、飞科纳斯弗朗西斯氏菌(Fructobacillus ficulneus)、果糖弗朗西斯氏菌(Fructobacillus fructosus)、弗朗西斯氏菌属物种(Fructobacillus sp.)EFB-N1、坏死梭杆菌(Fusobacterium necrophorum)、具核梭杆菌(Fusobacterium nucleatum)、牙周梭杆菌(Fusobacterium periodonticum)、Galbibacter marinus、卡氏杆菌(Gallibacterium anatis)、γ变形杆菌(gamma proteobacterium)HdN1、γ变形杆菌

(gamma proteobacterium)HTCC5015、道加德纳菌(Gardnerella vaginalis)、伯格孪生球菌(Gemella bergeri)、串孔孪生球菌(Gemella cuniculi)、溶血孪生球菌(Gemella haemolysans)、土芽孢杆菌属物种(Geobacillus sp.)血格鲁比卡菌(Globicatella sanguinis)、嗜重氮葡糖醋杆菌(Gluconacetobacter diazotrophicus)、Gordonibacter pamelaeae、粒子链菌属(Granulicatella)、嗜血杆菌属(Haemophilus)、副流感嗜血杆菌(Haemophilus parainfluenzae)、唾液嗜血杆菌(Haemophilus sputorum)、苏西创伤球菌(Helcococcus sueciensis)、森鼠螺杆菌(Helicobacter apodemus)、加拿大螺杆菌(Helicobacter canadensis)、同性恋螺杆菌(Helicobacter cinaedi)、芬纳尔螺杆菌(Helicobacter fennelliae)、鼠型螺杆菌(Helicobacter muridarum)、雪貂螺旋杆菌(Helicobacter mustelae)、帕美提斯螺杆菌(Helicobacter pametensis)、啮齿类螺杆菌(Helicobacter rodentium)、泰罗尼斯螺杆菌(Helicobacter typhlonius)、Hugenholtzia roseola、生丝单胞菌属(Hyphomonas)、Ignavibacterium album、营养泥杆菌(Ilyobacter polytropus)、Indibacter alkaliphilus、Jejuia pallidilutea、Jeotgalibaca dankookensis、Joostella marina、Kandleria vitulina、金格金氏杆菌(Kingella kingae)、Kiritimatiella glycovorans、Kordia algicida、Kordia jejudonensis、Kurthia huakuii、牛毛形杆菌(Lachnobacterium bovis)、经产妇毛螺菌属(Lachnospira multipara)、毛螺菌科细菌(Lachnospiraceae bacterium)AC2029、毛螺菌科细菌(Lachnospiraceae bacterium)MA2020、毛螺菌科细菌(Lachnospiraceae bacterium)NK4A179、Lacinutrix jangbogonensis、乳杆菌属(Lactobacillus)、阿法尼乳杆菌(Lactobacillus acidifarinae)、活泼乳杆菌(Lactobacillus agilis)、动物乳杆菌(Lactobacillus animalis)、动物乳杆菌(Lactobacillus animalis)KCTC 3501、阿蒂尼乳杆菌(Lactobacillus apodemi)、短乳杆菌(Lactobacillus brevis)、布氏乳杆菌(Lactobacillus buchneri)、可可乳杆菌(Lactobacillus cacaonum)、干酪乳杆菌(Lactobacillus casei)、西堤乳杆菌(Lactobacillus ceti)、西堤乳杆菌(Lactobacillus ceti)DSM 22408、复合乳杆菌(Lactobacillus composti)、凹乳杆菌(Lactobacillus concavus)、棒状乳杆菌(Lactobacillus coryniformis)、弯曲乳酸杆菌(Lactobacillus curvatus)、德氏乳酸杆菌(Lactobacillus delbrueckii)、地里润斯乳杆菌(Lactobacillus diolivorans)、香肠乳杆菌(Lactobacillus farciminis)、发酵乳杆菌(Lactobacillus fermentum)、花乳杆菌(Lactobacillus floricola)、多花乳杆菌(Lactobacillus florum)、福西斯乳杆菌(Lactobacillus fuchuensis)、福赛斯乳杆菌(Lactobacillus futsaii)、胃乳杆菌(Lactobacillus gastricus)、大猩猩乳杆菌(Lactobacillus gorillae)、匍匐乳酸菌(Lactobacillus graminis)、汉莫斯乳杆菌(Lactobacillus hammesii)、黑龙江乳杆菌(Lactobacillus heilongjiangensis)、大麦乳杆菌(Lactobacillus hordei)、惰性乳杆菌(Lactobacillus iners)、詹氏乳杆菌(Lactobacillus jensenii)、开菲尔乳杆菌(Lactobacillus kefiri)、坤可乳杆菌(Lactobacillus kunkeei)、林氏乳杆菌(Lactobacillus lindneri)、马里乳杆菌(Lactobacillus mali)、梅氏乳杆菌(Lactobacillus melliventris)、米德斯乳杆菌(Lactobacillus mindensis)、粘膜乳杆菌(Lactobacillus mucosae)、那慕斯乳杆菌(Lactobacillus namurensis)、诺德斯乳杆菌(Lactobacillus nodensis)、寡发酵乳杆菌

(Lactobacillus oligofermentans)、欧克斯乳杆菌(Lactobacillus otakiensis)、欧真斯乳杆菌(Lactobacillus ozensis)、副干酪乳杆菌(Lactobacillus paracasei)、副胶原乳杆菌(Lactobacillus paracollinoides)、副食乳杆菌(Lactobacillus paragasseri)、戊糖乳杆菌(Lactobacillus pentosus)、植物乳杆菌(Lactobacillus plantarum)、皮塔西乳杆菌(Lactobacillus psittaci)、瑞尼尼乳杆菌(Lactobacillus rennini)、罗伊氏乳杆菌(Lactobacillus reuteri)、鼠李糖乳杆菌(Lactobacillus rhamnosus)、罗斯乳杆菌(Lactobacillus rossiae)、瘤乳杆菌(Lactobacillus ruminis)、塞纳斯乳杆菌(Lactobacillus saerimneri)、清酒乳杆菌(Lactobacillus sakei)、唾液乳杆菌(Lactobacillus salivarius)、旧金山乳杆菌(Lactobacillus sanfranciscensis)、三维瑞乳杆菌(Lactobacillus saniviri)、森足克乳杆菌(Lactobacillus senmaizukei)、深圳乳杆菌(Lactobacillus shenzhenensis)、乳杆菌属物种(Lactobacillus sp.)、乳酸菌属属物种(Lactobacillus sp.)wkB8、图特提乳杆菌(Lactobacillus tucceti)、文德斯乳杆菌(Lactobacillus versmoldensis)、沃森斯乳杆菌(Lactobacillus wasatchensis)、酶乳杆菌(Lactobacillus zymae)、鼠李糖乳杆菌(Lactobacillus_rhamnosus)_LOCK900、Lagierella massiliensis、Lawsonella clevelandensis、嗜肺军团菌(Legionella pneumophila)、柔毛藻口腔分类单元(Leptotrichia sp.oral taxon)215、葛迪度穆明串珠菌(Leuconostoc gelidum)、Limnohabitans planktonicus、费曼氏李斯特菌(Listeria fleischmannii)、绵羊李斯特菌(Listeria ivanovii)、单核细胞增多性李斯特氏菌(Listeria monocytogenes)、单核细胞增多性李斯特氏菌(Listeria monocytogenes)Lm_1880、斯氏李斯特氏菌(Listeria seeligeri)、Lunatimonas lonarensis、Lutibacter profundi、曼氏杆菌属(Mannheimia)、马恩西斯曼氏杆菌(Mannheimia massilioguelmaensis)、曼氏杆菌属物种(Mannheimia sp.)USDA-ARS-USMARC-1261、Massilibacterium senegalense、巨球形菌属物种(Megasphaera sp.)UPII 135-E、中慢生根瘤菌属物种(Mesorhizobium sp.)、中慢生根瘤菌属物种(Mesorhizobium sp.)LC103、甲基孢囊菌属物种(Methylocystis sp.)ATCC 49242、嗜甲基菌属物种(Methylophilus sp.)5、嗜甲基菌属物种(Methylophilus sp.)OH31、甲基弯曲菌属(Methylosinus)、Methylovulum miyakonense、克氏动弯杆菌(Mobiluncus curtisii)、Mucilaginibacter paludis、Mucinivorans hirudinis、Mucispirillum schaedleri、精氨酸支原体(Mycoplasma arginini)、犬枝原体(Mycoplasma canis)、殊异支原体(Mycoplasma dispar)、败血支原体(Mycoplasma gallisepticum)、猪滑液支原体(Mycoplasma hyosynoviae)、移动支原体(Mycoplasma mobile)、绵羊肺炎支原体(Mycoplasma ovipneumoniae)、滑液囊支原体(Mycoplasma synoviae)、鸡毒支原体(Mycoplasma_gallisepticum)_CA06、气味香味菌(Myroides odoratus)、Necropsobacter massiliensis、北极奈瑟氏菌(Neisseria arctica)、杆菌状奈瑟氏菌(Neisseria bacilliformis)、脑膜炎奈瑟氏菌(Neisseria meningitidis)、奈瑟氏菌属物种(Neisseria sp.)、奈瑟氏菌属物种(Neisseria sp.)74A18、瓦茨瓦尔奈瑟氏菌(Neisseria wadsworthii)、Niabella soli、Nitratifractor salsuginis、亚硝化单胞菌属物种(Nitrosomonas sp.)AL212、新鞘脂菌属物种(Novosphingobium sp.)MD-1、Oceanivirga salmonicida、曼西斯大洋芽胞杆菌(Oceanobacillus manasiensis)、Odoribacter

laneus、北原葡萄球菌17330(Oenococcus kitaharae DSM 17330)、尿道寡源杆菌(Oligella urethralis)、多发寡源杆菌(Olsenella profusa)、寡源杆菌属物种(Olsenella sp.)DNF00959、尤里寡源杆菌(Olsenella uli)、鼻气管炎鸟细菌(Ornithobacterium rhinotracheale)、Ottowia属物种口腔分类单元(Ottowia sp.oral taxon)894、Pannonibacter phragmitetus、Parabacteroides johnsonii DSM 18315、Parabacteroides sp.、Parabacteroides sp.D26、Parasutterella excrementihominis、Parvibaculum lavamentivorans、微单胞菌属物种(Parvimonas sp.)KA00067、禽多杀性巴氏杆菌(Pasteurella multocida)、乳酸片球菌(Pediococcus acidilactici)、有害片球菌(Pediococcus damnosus)、意外片球菌(Pediococcus inopinatus)、小片球菌(Pediococcus parvulus)、戊糖片球菌(Pediococcus pentosaceus)、斯提西片球菌(Pediococcus stilesii)、顾替科斯土地杆菌(Pedobacter glucosidilyticus)、Pelomonas sp.Root1237、杜德尼嗜胨菌(Peptoniphilus duerdenii)、肥胖嗜胨菌(Peptoniphilus obesi)、嗜胨菌属物种口腔分类单元(Peptoniphilus sp.oral taxon)386、厌氧菌胃链球菌(Peptostreptococcus anaerobius)CAG:621、赛特斯考拉杆菌(Phascolarctobacterium succinatutens)、南极动球菌(Planococcus antarcticus)、口腔卟啉单胞菌(Porphyromonas catoniae)、牙龈卟啉单胞菌(Porphyromonas gingivalis)、萨姆依卟啉单胞菌(Porphyromonas somerae)、卟啉单胞菌属口腔分类单元278(Porphyromonas sp.oral taxon 278)、羊普雷沃菌(Prevotella amnii)、树蛙普雷沃菌(Prevotella aurantiaca)、巴尼普雷沃菌(Prevotella baroniae)、二路普雷沃菌(Prevotella bivia)、口颊普雷沃菌(Prevotella buccalis)、人体普雷沃菌(Prevotella corporis)、栖牙普雷沃菌(Prevotella denticola)、解糖胨普雷沃菌(Prevotella disiens)、栖组织普雷沃菌(Prevotella histicola)、中间普雷沃菌(Prevotella intermedia)、洛氏普雷沃菌(Prevotella loescheii)、产黑色普雷沃菌(Prevotella melaninogenica)、纳西斯普雷沃菌(Prevotella nanceiensis)、变黑普雷沃菌(Prevotella nigrescens)、口腔普雷沃菌(Prevotella oralis)、皮迪思普雷沃菌(Prevotella pleuritidis)、栖瘤胃普雷沃菌(Prevotella ruminicola)、解糖普雷沃菌(Prevotella saccharolytica)、普雷沃菌属物种(Prevotella sp.)C561、普雷沃菌属物种(Prevotella sp.)DNF00663、普雷沃菌属物种(Prevotella sp.)HJM029、普雷沃菌属物种(Prevotella sp.)HUN102、普雷沃菌属物种(Prevotella sp.)MSX73、普雷沃菌属物种口腔分类单元306(Prevotella sp.oral taxon 306)、普雷沃菌属物种口腔分类单元317(Prevotella sp.oral taxon 317)、普雷沃菌属物种(Prevotella sp.)P5-119、斯瑞尔普雷沃菌(Prevotella stercorea)、Propionimicrobium lymphophilum、Pseudaminobacter salicylatoxidans、铜绿假单胞菌(Pseudomonas aeruginosa)、亚麻假单胞菌(Pseudomonas lini)、扭曲冷弯曲菌(Psychroflexus torquis)、冷蛇菌属物种(Psychroserpens sp.)Hel_I_66、青罗尔斯通氏菌(Ralstonia solanacearum)、红杆菌科细菌(Rhodobacteraceae bacterium)HLUCCA08、红杆菌科细菌(Rhodobacteraceae bacterium)HLUCCA12、深红红螺菌(Rhodospirillum rubrum)、小红卵菌属物种(Rhodovulum sp.)PH10、鸭疫里默氏杆菌(Riemerella anatipestifer)、小梭文肯菌(Rikenella microfusus)、理研菌科物种(Rikenellaceae sp.)、Rodentibacter

pneumotropicus、肠罗氏菌(Roseburia intestinalis)、罗氏菌属物种(Roseburia sp.)CAG:197、艾瑞尔罗思氏菌(Rothia aeria)、龋齿罗思氏菌(Rothia dentocariosa)、粘滑罗思氏菌(Rothia mucilaginosa)、Rubritepida flocculans、Rugosibacter aromaticivorans、Ruminiclostridium cellulolyticum、白色瘤胃球菌(Ruminococcus albus)、黄瘤胃球菌(Ruminococcus flavefaciens)、乳酸瘤胃球菌(Ruminococcus lactaris)、Saccharibacter sp.AM169、Salegentibacter sp.Hel_I_6、Salinispira pacifica、Salinivirga cyanobacteriivorans、科瑞栖盐水芽胞杆菌(Salsuginibacillus kocurii)、Scardovia inopinata、Scardovia wiggsiae、Schleiferia thermophila、Sedimenticola thiotaurini、Sediminibacterium sp.C3、Sharpea azabuensis、Shimia marina、米氏西蒙斯氏菌(Simonsiella muelleri)、Skermanella aerolata、Solobacterium moorei、Sphaerochaeta globosa、食醇鞘氨醇杆菌(Sphingobacterium spiritivorum)、巴蒂瑞鞘脂菌属(Sphingobium baderi)、鞘脂菌属物种(Sphingobium sp.)AP49、鞘脂菌属物种(Sphingobium sp.)C100、鞘脂单胞菌属(Sphingomonas)、长白鞘脂单胞菌(Sphingomonas changbaiensis)、珊尼鞘脂单胞菌(Sphingomonas sanxanigenens)、鞘脂单胞菌属物种(Sphingomonas sp.)Leaf412、鞘脂单胞菌属物种(Sphingomonas sp.)MM-1、鞘脂单胞菌属物种(Sphingomonas sp.)SRS2、阿皮斯鞘脂单胞菌(Spiroplasma apis)、滨海鞘脂单胞菌(Spiroplasma litorale)、托卡姆鞘脂单胞菌(Spiroplasma turonicum)、生孢噬纤维菌(Sporocytophaga myxococcoides)、维尼芽孢乳杆菌(Sporolactobacillus vineae)、阿涅蒂斯葡萄球菌(Staphylococcus agnetis)、溶血性葡萄球菌(Staphylococcus haemolyticus)、人葡萄球菌(Staphylococcus hominis)、路邓葡萄球菌(Staphylococcus lugdunensis)、微小葡萄球菌(Staphylococcus microti)、巴氏葡萄球菌(Staphylococcus pasteuri)、中间葡萄球菌(Staphylococcus pseudintermedius)、施氏葡萄球菌(Staphylococcus schleiferi)、模仿葡萄球菌(Staphylococcus simulans)、葡萄球菌属物种(Staphylococcus sp.)CAG:324、猫链杆菌(Streptobacillus felis)、念珠状链杆菌(Streptobacillus moniliformis)、链球菌属(Streptococcus)、无乳链球菌(Streptococcus agalactiae)、咽峡炎链球菌(Streptococcus anginosus)、狗链球菌(Streptococcus canis)、星座链球菌(Streptococcus constellatus)、停乳链球菌(Streptococcus dysgalactiae)、马链球菌(Streptococcus equi)、马肠链球菌(Streptococcus equinus)、解没食子酸链球菌(Streptococcus gallolyticus)、格氏链球菌(Streptococcus gordonii)、苦丁链球菌(Streptococcus henryi)、婴儿链球菌(Streptococcus infantarius)、海豚链球菌(Streptococcus iniae)、猕猴链球菌(Streptococcus macacae)、马克顿链球菌(Streptococcus macedonicus)、哺乳链球菌(Streptococcus marimammalium)、马赛链球菌(Streptococcus massiliensis)、缓症链球菌(Streptococcus mitis)、变形链球菌(Streptococcus mutans)、口腔链球菌(Streptococcus oralis)、口腔链球菌提格里斯亚种(Streptococcus oralis subsp.tigurinus)AZ_3a、奥西尼链球菌(Streptococcus orisasini)、奥拉提链球菌(Streptococcus orisratti)、羊链球菌(Streptococcus ovis)、副血链球菌(Streptococcus parasanguinis)、普洛托姆链球菌(Streptococcus plurextorum)、假肺炎链球菌(Streptococcus pseudopneumoniae)、假猪链球菌

(Streptococcus pseudoporcinus)、酿脓链球菌(Streptococcus pyogenes)、鼠链球菌(Streptococcus ratti)、血链球菌(Streptococcus sanguinis)、中华链球菌(Streptococcus sinensis)、远缘链球菌(Streptococcus sobrinus)、链球菌属物种(Streptococcus sp.)C150、链球菌属物种(Streptococcus sp.)C300、链球菌属物种(Streptococcus sp.)HSISB1、链球菌属物种(Streptococcus sp.)I-G2、猪链球菌(Streptococcus suis)、嗜热链球菌(Streptococcus thermophilus)、瓦拉尼链球菌(Streptococcus varani)、无乳链球菌(Streptococcus_agalactiae)_NEM316、停乳链球菌似马亚种(Streptococcus_dysgalactiae_subsp._equisimilis)_AC-2713、解没食子酸链球菌解没食子酸亚种(Streptococcus_gallolyticus_subsp._gallolyticus)_ATCC_43143、格氏链球菌卡尔斯株系CH1亚株系(Streptococcus_gordonii_str._Challis_substr._CH1)、变形链球菌(Streptococcus_mutans)_GS-5、唾液链球菌(Streptococcus_salivarius)_JIM8777、猪链球菌(Streptococcus_suis)_D9、嗜热链球菌(Streptococcus_thermophilus)_LMG_18311、Subdoligranulum sp.4_3_54A2FAA、东克拉亚硫酸杆菌(Sulfitobacter donghicola)、Sulfuritalea hydrogenivorans、硫磺单胞菌属物种(Sulfurospirillum sp.)、硫磺单胞菌属物种(Sulfurospirillum sp.)SCADC、Sulfurovum lithotrophicum、沃德西斯萨特氏菌(Sutterella wadsworthensis)、Tamlana sedimentorum、福赛斯坦纳菌(Tannerella forsythia)、海洋黄杆菌(Tenacibaculum maritimum)、特达瑞斯热硫杆状菌(Thermithiobacillus tepidarius)、Thermophagus xiamenensis、Thioalkalivibrio、Tissierellia细菌KA00581、Tissierellia细菌S5-A11、运动替斯崔纳菌(Tistrella mobilis)、齿垢密螺旋体(Treponema denticola)、嗜麦芽糖密螺旋体(Treponema maltophilum)、足螺旋体(Treponema pedis)、恶臭螺旋体(Treponema putidum)、索氏螺旋体(Treponema socranskii)、齿密螺旋体(Treponema_denticola)_ATCC_35405、Turicibacter sp.、未培养白蚁1组细菌(uncultured Termite group 1 bacterium)、嗜热球形脲芽胞杆菌(Ureibacillus thermosphaericus)、Urinacoccus massiliensis、非典型韦荣球菌(Veillonella atypica)、麦格纳韦荣球菌(Veillonella magna)、小韦荣球菌(Veillonella parvula)、小韦荣球菌(Veillonella parvula)ATCC 17745、韦荣球菌属物种(Veillonella sp.)6_1_27、韦荣球菌属物种(Veillonella sp.)AS16、韦荣球菌属物种(Veillonella sp.)CAG:933、韦荣球菌属物种(Veillonella sp.)DNF00869、韦荣球菌属物种(Veillonella sp.)DorA_A_3_16_22、Verminephrobacter aporrectodeae、Verminephrobacter eiseniae、疣微菌门细菌(Verrucomicrobia bacterium)IMCC2613、塞内加尔枝芽孢杆菌(Virgibacillus senegalensis)、马赛威克斯菌(Weeksella massiliensis)、有毒威克斯菌(Weeksella virosa)、耐盐威克斯菌(Weissella halotolerans)、坎氏威克斯菌(Weissella kandleri)、产琥珀酸沃廉菌(Wolinella succinogenes)、马里蒂马木洞菌(Woodsholea maritima)、Yoonia vestfoldensis、和暗王祖农菌(Zunongwangia profunda)。

[0011] 在一些方面,本发明提供了合成的组合物,所述合成的组合物包含异源组分和选自由以下组成的组的多肽:与SEQ ID NO:86-171和511-1135中任何一个的至少50、50至100、至少100、100至150、至少150、150至200、至少200、200至250、至少250、250至300、至少300、300至350、至少350、350至400、至少400、400至450、至少500、500至550、至少550、550至

600、至少600、600至650、至少650、650至700、至少700、700至750、至少750、750至800、至少800、800至850、至少850、850至900、至少900、900至950、至少950、950至1000、至少1000或甚至大于1000个连续氨基酸具有至少80％、80％至85％、至少85％、85％至90％、至少90％、90％至95％、至少95％、至少96％、至少97％、至少98％、至少99％、至少99.5％或大于99.5％同一性的多肽;SEQ ID NO:86-171和511-1135中任何一个的功能性变体;SEQ ID NO:86-171和511-1135中任何一个的功能性片段;Cas内切核酸酶,其由选自由SEQ ID NO:1-85组成组的多核苷酸编码;Cas内切核酸酶,其识别表4-83中任何一个列出的PAM序列;Cas内切核酸酶,其识别选自下组的PAM序列,该组由以下组成:NAR(G＞A)WH(A＞T＞C)GN(C＞T＞R)、N(C＞D)V(A＞S)R(G＞A)TTTN(T＞V)、NV(A＞G＞C)TTTTT、NATTTTT、NN(H＞G)AAAN(G＞A＞Y)N、N(T＞V)NAAATN、NAV(A＞G＞C)TCNN、NN(A＞S＞T)NN(W＞G＞C)CCN(Y＞R)、NNAH(T＞M)ACN、NGTGANN、NARN(A＞K＞C)ATN、NV(G＞A＞C)RNTTN、NN(A＞B)RN(A＞G＞T＞C)CCN、NN(A＞B)NN(T＞V)CCH(A＞Y)、NNN(H＞G)NCDAA、NN(H＞G)D(A＞K)GGDN(A＞B)、NNNNCCAG、NNNNCTAA、NNNNCVGANN、N(C＞D)NNTCCN、NNNNCTA、NNNNCYAA、NAGRGNY、NNGH(W＞C)AAA、NNGAAAN、NNAAAAA、NTGAR(G＞A)N(A＞Y＞G)N(Y＞R)、N(C＞D)H(C＞W)GH(Y＞A)N(A＞B)AN(A＞T＞S)、NNAAACN、NNGTAM(A＞C)Y、NH(A＞Y)ARNN(C＞W＞G)N、B(C＞K)GGN(A＞Y＞G)N NN、N(T＞C＞R)AGAN(A＞K＞C)NN、NGGN(A＞T＞G＞C)NNN、NGGD(A＞T＞G)TNN、NGGAN(T＞A＞C＞G)NN、CGGWN(T＞R＞C)NN、NGGWGNN、N(B＞A)GGNN(T＞V)NN、NNGD(A＞T＞G)AY(T＞C)N、N(T＞V)H(T＞C＞A)AAAAN、NRTAANN、N(H＞G)CAAH(Y＞A)N(Y＞R)N、NATAAN(A＞T＞S)N、NV(A＞G＞C)R(A＞G)ACCN、CN(C＞W＞G)AV(A＞S)GAC、NNRNCAC、N(A＞B)GGD(W＞G)D(G＞W)NN、BGD(G＞W)GTCN(A＞K＞C)、NAANACN、NRTHAN(A＞B)N、BHN(H＞G)NGN(T＞M)H(Y＞A)、NMRN(A＞Y＞G)AH(C＞T＞A)N、NNNCACN、NARN(T＞A＞S)ACN、NNNNATW、NGCNGCN、NNNCATN、NAGNGCN、NARN(T＞M＞G)CCN、NATCCTN、NRTAAN(T＞A＞S)N、N(C＞T＞G＞A)AAD(A＞G＞T)CNN、NAAAGNN、NNGACNN、N(T＞V)NTAAD(A＞T＞G)N、NNGAD(G＞W)NN、NGGN(W＞S)NNN、N(T＞V)GGD(W＞G)GNN、NGGD(A＞T＞G)N(T＞M＞G)NN、NNAAAGN、N(G＞H)GGDN(T＞M＞G)NN、NNAGAAA、NN(T＞M＞G)AAAAA、N(C＞D)N(C＞W＞G)GW(T＞C)D(A＞G＞T)AA、NAAAAYN、NRGNNNN、NATGN(H＞G)TN、NNDATTT和NATARCN(C＞T＞A＞G);Cas内切核酸酶,其能够识别长度为一、二、三、四、五、六、七、八、九或十个核苷酸的PAM序列;Cas内切核酸酶,其包含与SEQ ID NO:1136-1730中的任何一个具有至少80％、80％至85％、至少85％、85％至90％、至少90％、90％至95％、至少95％、至少96％、至少97％、至少98％、至少99％、至少99.5％、或大于99.5％同一性的结构域;Cas内切核酸酶,其具有以下活性得分(根据与实例9的方法相同或相似的方法)或表86A的氨基酸表的位置得分的总和:至少1.0、1.0至2.0、至少2.0、2.0至3.0、至少3.0、3.0至4.0、至少4.0、4.0至5.0、至少5.0、5.0至6.0、至少6.0、6.0至7.0、至少7.0、7.0至8.0、至少8.0、8.0至9.0、至少9.0、9.0至10.0、至少10.0或甚至大于10.0;Cas内切核酸酶,其包含与SEQ ID NO:1125的相对序列位置编号的比对相比,表86B中鉴定的一、二、三、四、五、六、七、八、九、十、十一、十二、十三、十四、十五、十六、十七、十八、十九、二十、二十一、二十二、二十三、二十四、二十五或二十六个特征氨基酸;以及Cas内切核酸酶,所述Cas内切核酸酶能够与包含SEQ ID NO:426-510、341-425、141-255或256-340中任一个的指导物多核苷酸形成复合物。在一些方面,Cas9多核苷酸具有多个先前列出的特征。

[0012] 在一些方面,本发明提供了能够与Cas内切核酸酶形成复合物以识别、结合并任选

地切口或切割靶多核苷酸的一种或多种指导多核苷酸和/或种或多种组分。在一些方面，指导多核苷酸包含与SEQ ID NO:426-510、341-425、171-255或256-340中的任何一个具有至少80％、80％至85％、至少85％、85％至90％、至少90％、90％至95％、至少95％、至少96％、至少97％、至少98％、至少99％、至少99.5％或大于99.5％同一性的序列。

[0013] 在一些方面，本发明提供了Cas内切核酸酶，其能够在双链靶多核苷酸中产生单链断裂或切口。在一些方面，Cas内切核酸酶能够产生粘性末端突出双链断裂。在一些方面，Cas内切核酸酶能够产生平末端双链断裂。

[0014] 在一些方面，所述异源组分选自由以下组成的组：细胞、异源多核苷酸、供体DNA分子、修复模板多核苷酸、异源多肽、脱氨酶、异源核酸酶、粒子、固体基质、抗体、缓冲液组合物、Tris、EDTA、二硫苏糖醇(DTT)、磷酸盐缓冲盐水(PBS)、氯化钠、氯化镁、HEPES、甘油、牛血清白蛋白(BSA)、盐、乳化剂、洗涤剂、螯合剂、氧化还原剂、抗体、无核酸酶的水、粘度剂和组氨酸标签。在一些方面，所述异源多肽包含核酸酶结构域、转录激活子结构域、转录阻遏子结构域、表观遗传修饰结构域、切割结构域、核定位信号、细胞穿透性结构域、脱氨酶结构域、碱基编辑结构域、易位结构域、标志物和转基因。在一些方面，所述异源多核苷酸选自由以下组成的组：指导多核苷酸、嵌合指导多核苷酸、化学修饰的指导多核苷酸、同时DNA和RNA两者的指导多核苷酸、非编码表达元件、基因、标志物和编码多个组氨酸残基的多核苷酸。在一些方面，所述合成的组合物包含至少两个、至少三个、至少四个、至少五个或甚至大于五个异源组分。在一些方面，存在多个不同的异源组分。在一些方面，存在多个相同类型的异源组分。在一些方面，存在多个相同的异源组分。

[0015] 在一些方面，所述合成的组合物的pH为1.0至14.0、2.0至13.0、3.0至12.0、4.0至11.0、5.0至10.0、6.0至9.0、7.0至8.0、4.5至6.5、5.5至7.5、或6.5至7.5。在一些方面，Cas9直系同源物在以下pH具有最佳活性：1.0至14.0、2.0至13.0、3.0至12.0、4.0至11.0、5.0至10.0、6.0至9.0、7.0至8.0、4.5至6.5、5.5至7.5、或6.5至7.5。

[0016] 在一些方面，所述Cas9直系同源物在以下温度具有最佳活性：0摄氏度至100摄氏度、至少0摄氏度至10摄氏度、至少10摄氏度至20摄氏度、至少20摄氏度至25摄氏度、至少25摄氏度至30摄氏度、至少30摄氏度至40摄氏度、至少40摄氏度至50摄氏度、至少50摄氏度至60摄氏度、至少60摄氏度至70摄氏度、至少70摄氏度至80摄氏度、至少80摄氏度至90摄氏度、至少90摄氏度至100摄氏度、或大于100摄氏度。

[0017] 在一些方面，所述合成的组合物在以下温度储存或孵育：至少负200摄氏度、至少负150摄氏度、至少负135摄氏度、至少负90摄氏度、至少负80摄氏度、至少负20摄氏度、至少4摄氏度、至少17摄氏度、至少25摄氏度、至少30摄氏度、至少35摄氏度、至少37摄氏度、至少39摄氏度、或大于39摄氏度。

[0018] 在一些方面，所述合成的组合物中的任何都可以处于基本上无核酸酶的环境中。在一些方面，所述合成的组合物中的任何都可以处于基本上无内毒素的环境中。在一些方面，所述合成的组合物中的任何都可以处于基本上无核酸酶且无内毒素的环境中。在一些方面，所述合成的组合物中的任何都可以被冻干。在一些方面，所述合成的组合物中的任何都可以存在于水溶液中。在一些方面，所述合成的组合物中的任何都可以存在于非水溶液中。

[0019] 在一方面，本发明提供了一种通过以下来调节Cas9直系同源物/指导多核苷酸复

合物与其野生型活性相比的靶多核苷酸特异性的方法:改变选自由以下组成的组的参数:指导多核苷酸长度、指导多核苷酸组成、PAM识别序列的长度、PAM识别序列的组成以及Cas9分子与靶多核苷酸主链的亲和力;并评估具有改变的参数的复合物的靶多核苷酸特异性,并将其与具有野生型参数的复合物的活性进行比较。在一些实施例中,靶多核苷酸特异性可以用更长的PAM识别序列来增加。在一些实施例中,靶多核苷酸特异性可以用更短的PAM识别序列来降低。在一些实施例中,可以通过工程改造非天然存在的PAM识别序列来调节靶多核苷酸特异性。

[0020] 一方面,本发明提供了一种通过以下来优化Cas9分子的活性的方法:使亲本Cas9分子经历至少一轮随机蛋白改组或分子进化,并选择具有至少一种不存在于亲本Cas9分子中的特征的所得分子。在一些实施例中,可以执行多轮。

[0021] 一方面,本发明提供了一种通过以下来优化Cas9分子的活性的方法:使亲本Cas9分子经历至少一轮非随机蛋白改组或分子进化,并选择具有至少一种不存在于亲本Cas9分子中的特征的所得分子。在一些实施例中,可以执行多轮。

[0022] 一方面,本发明提供了(使用本文提供的任何组合物或衍生自本文提供的组合物的任何组合物或用本文提供的任何方法鉴定的任何组合物)实现靶多核苷酸的单链切口或双链断裂的方法,修饰分离的多核苷酸或基因组多核苷酸的方法,体外多核苷酸修饰的方法,体内多核苷酸修饰的方法,编辑多核苷酸的一个或多个碱基的方法,调节细胞中内源或转基因多核苷酸表达的方法,或赋予已经引入了所述组合物的细胞、组织或生物体益处的方法。

[0023] 本文提供的基因组修饰方法包括至少一个核苷酸的插入、至少一个核苷酸的缺失、至少一个核苷酸的修饰、至少一个核苷酸的交换、至少一个核苷酸的化学改变、至少一个核苷酸的脱氨基,或前述的任何组合。

[0024] 在一些方面,已经修饰了Cas内切核酸酶以改变其野生型活性、以更高频率地切割靶多核苷酸、以更低频率地切割多核苷酸、或降低或消除核酸酶活性。

[0025] 在一些方面,Cas内切核酸酶与另一种多肽结合以产生融合蛋白,例如与脱氨酶或异源核酸酶。

[0026] 在本文提供的方法或组合物的任何方面,细胞可以选自由以下组成的组:人、非人灵长类、哺乳动物、动物、古细菌、细菌、原生生物、真菌、昆虫、酵母、非常规酵母和植物细胞。在一些实施例中,细胞与Cas9内切核酸酶从其衍生的生物是异源的。在一些实施例中,细胞是选自由单子叶植物和双子叶植物细胞组成的组的植物细胞。在一些实施例中,细胞是选自由以下组成的组的植物细胞:玉蜀黍、水稻、高粱、黑麦、大麦、小麦、粟、燕麦、甘蔗、草坪草、柳枝稷、大豆、卡诺拉油菜、苜蓿、向日葵、棉花、烟草、花生、马铃薯、烟草、拟南芥、蔬菜和红花细胞。在一些实施例中,细胞是动物细胞,任选地是哺乳动物细胞,任选地是灵长类细胞,或任选地是人细胞,所述人细胞选自由以下组成的组:单倍体细胞、二倍体细胞、生殖细胞、神经元、肌肉细胞、内分泌或外分泌细胞、上皮细胞、肌肉细胞、肿瘤细胞、胚胎细胞、造血细胞、骨细胞、种质细胞、体细胞、干细胞、多能干细胞、诱导多能干细胞、祖细胞、减数分裂细胞和有丝分裂细胞。

[0027] 在任何方面,由于本文提供的组合物或方法,使所述细胞、或包含所述细胞的生物、或细胞的后续世代或衍生自所述细胞的生物体受益。在一些实施例中,通过将所述细

胞、或包含所述细胞的生物、或细胞的后续世代或衍生自所述细胞的生物体与未进行本文提供的方法或不包含本文提供的至少一种组合物的同系细胞进行比较来确定益处。在一些实施例中，由于多核苷酸修饰、缺失或插入而提供益处。在一些实施例中，所述益处选自由以下组成的组：改善的健康、改善的生长、改善的能育性、改善繁殖力、改善的环境耐受、改善的活力、改善的疾病抗性、改善的疾病耐受、改善的对异源分子的耐受、改善的适应性、改善的物理特征、更大的质量、增加的生化分子产生、减少的生化分子产生、基因的上调、基因的下调、生化途径的上调、生化途径的下调、细胞繁殖的刺激和细胞繁殖的抑制，如与不包含或不衍生自含有所述供体多核苷酸的细胞的同系植物(isoline plant)相比。在一些实施例中，所述靶位点的修饰导致包含或衍生自所述细胞或其后代细胞的植物的具有农艺学意义的性状的调节，所述具有农艺学意义的性状选自由以下组成的组：疾病抗性、干旱抗性、热耐性、寒耐性、盐耐性、金属耐性、除草剂耐性、改善的水分利用效率、改善的氮利用率、改善的固氮作用、有害生物抗性、食草动物抗性、病原体抗性、产率改善、健康增强、改善的能育性、活力改善、生长改善、光合能力改善、营养增强、改变的蛋白含量、改变的油含量、增加的生物量、增加的芽长度、增加的根长度、改善的根结构、代谢产物的调节、蛋白质组的调节、增加的种子重量、改变的种子碳水化合物组成、改变的种子油组成、改变的种子蛋白组成、改变的种子营养物组成；如与不包含或不衍生自含有所述供体多核苷酸的细胞的同系植物相比。在一些实施例中，所述细胞是动物细胞，其中所述靶位点的修饰导致包含所述动物细胞或其后代细胞的生物的具有生理学意义的表型的调节，所述具有生理学意义的表型选自由以下组成的组：改善的健康、改善的营养状况、减少的疾病影响、疾病静止状态、疾病逆转、改善的能育性、改善的活力、改善的心智能力、改善的生物体生长、改善的增重、减重、内分泌系统的调节、外分泌系统的调节、减小的肿瘤大小、减小的肿瘤质量、刺激的细胞生长、降低的细胞生长、代谢产物的产生、激素的产生、免疫细胞的产生、以及刺激细胞产生。

[0028] 附图和序列表的说明

[0029] 根据下列的详细描述和附图以及序列表，可以更全面地理解本公开，所述详细描述和附图以及序列表形成本申请的一部分。这些序列描述以及所附序列表遵守如37C.F.R.§§1.821和1.825所列出的管理专利申请中核苷酸和氨基酸序列公开内容的规则。这些序列描述包含如在37C.F.R.§§1.821和1.825中所定义的用于氨基酸的三字母代码，将其通过引用结合在此。

附图说明

[0030] 图1是产生的用于鉴定实例1中所述的12个进化枝的系统发生图的图形表示。

[0031] 图2描绘了针对实例1中描述的12个进化枝中的每个进化枝的一些Cas9直系同源物鉴定的指导RNA分子的二级结构图。

[0032] 图3显示了针对实例1中描述的12个进化枝的每个进化枝的Cas9直系同源物中的一些而确定的共有PAM序列，如表4-83中详述。

[0033] 图4显示了组I Cas9直系同源物(SEQ ID NO:58、62、64、63、65、71、69、74、66、67、70、72、73、68、83、79、82、76、78、80、81、77和75)的共有序列，所述组I Cas9直系同源物与金黄色葡萄球菌Cas9结构PDB ID 5CZZ_A(“Crystal structure of Staphylococcus aureus

Cas9[金黄色葡萄球菌Cas9的晶体结构]",Nishimasu,H.,Cong,L.,Yan,W.X.,Ran,F.A.,Zetsche,B.,Li,Y.,Kurabayashi,A.,Ishitani,R.,Zhang,F.,Nureki,O.,(2015)Cell[细胞]162:1113-1126)比对。绝对保守的残基以粗体加下划线的文本(X)描绘。

[0034] 图5显示了组III Cas9直系同源序列(SEQ ID NO:51、52、53、54、55、56、57、59、84、85、86、87、88、89、90、91、92、93、94、95、96和97)的共有序列,所述组III Cas9直系同源序列与酿脓链球菌血清型M1结构PDB ID 4UN3_B("Structural Basis of Pam-Dependent Target DNA Recognition by the Cas9 Endonuclease[Cas9内切核酸酶对Pam依赖性靶DNA识别的结构基础]",Anders,C.,Niewoehner,O.,Duerst,A.,Jinek,M.,(2014)Nature[自然]513:569-573)比对。绝对保守的残基以粗体加下划线的文本(X)描绘。

[0035] 图6显示了组IV Cas9直系同源物(SEQ ID NO:98和99)的共有序列,所述组IV Cas9直系同源物与内氏放线菌结构PDB ID 4OGE_A("Structures of Cas9 endonucleases reveal RNA-mediated conformational activation[Cas9内切核酸酶的结构揭示了RNA介导的构象激活]",Jinek,M.,Jiang,F.,Taylor,D.W.,Sternberg,S.H.,Kaya,E.,Ma,E.,Anders,C.,Hauer,M.,Zhou,K.,Lin,S.,Kaplan,M.,Iavarone,A.T.,Charpentier,E.,Nogales,E.,Doudna,J.A.,(2014)Science[科学]343:1247997-1247997)比对。绝对保守的残基以粗体加下划线的文本(X)描绘。

[0036] 图7显示了实例9中所述的用于测试用Cas9切割后的HDR频率的实验方法:图7A描绘了经由荧光报告子的重复区域的HDR,图7B描绘了与Cas9一起引入的修复模板。

[0037] 图8显示了通过两种不同方法(IVT和RNP)对选择的Cas9直系同源物的WebLogo比较。用纯化的核糖核蛋白(RNP)以几种不同的浓度证实了IVT方法结果。

[0038] 图9显示了将其中过量回收Illumina序列(导致相比阴性对照的读段覆盖的峰或尖)的原间隔子-衔接子连接位置表示为切割位置,其中数值结果作为衔接子连接的读段的分率。图9A显示了进化枝I、II、III和V的选择的序列的结果,图9B显示了进化枝VI、VII、VIII和IX的选择的序列的结果。图9C显示了进化枝X、XI和XII的选择的序列的结果。

[0039] 图10A显示了那些在前间隔子位置而不是紧接3之后产生显性切割的Cas9蛋白然后通过捕获由切割、末端修复、3'腺嘌呤添加和切割的文库靶的前间隔子侧的衔接子连接产生的切割产物来重新检查。

[0040] 图10B显示了显示出粘性末端切割的选择的Cas9直系同源物中的八种的切割的位置和类型(基于针对切割的原间隔子和PAM切割侧比较的所得频率,考虑到T4 DNA聚合酶末端修复)。

[0041] 图11显示了在五种不同缓冲液组合物中用两种不同长度的间隔子(20个核苷酸和24个核苷酸)测试的Cas9直系同源物中的一些的体外切割数据。

[0042] 图12显示使用酿脓链球菌sgRNA的选择的Cas9直系同源物的体外切割数据。

[0043] 图13显示了Cas9直系同源物之一ID46的体外切割活性相比于温度,显示了宽范围的温度活性,其中最佳活性在约15摄氏度至约60摄氏度,间隔子核苷酸长度为24个核苷酸;以及窄的活性窗口,其中最大温度约45摄氏度,间隔长度为20个核苷酸。

[0044] 图14显示了在代表性数量的Cas9直系同源物情况下,转化后两天玉蜀黍细胞中平均NHEJ频率。

[0045] 图15显示了由选择的Cas9直系同源物产生的20个不同突变体中的预期剪切位点。

图15A显示了ID33的结果，并且图15B显示了ID64的结果。

[0046] 图16显示了与用酿脓链球菌Cas9修饰的对照植物相比，玉蜀黍T0植物中跨三个不同靶位点(MS45、MS26和LIG)的两种不同Cas9直系同源物(ID33和ID64)的结果。

[0047] 图17显示了与酿脓链球菌Cas9的活性相比，用重组构建体(所述重组构建体包含编码相应Cas9直系同源物的DNA序列)转化的细胞中，选择的Cas9直系同源物在HEK细胞WTAP基因座处的结果。

[0048] 图18显示了与酿脓链球菌Cas9的活性相比，用重组构建体(所述重组构建体包含编码相应Cas9直系同源物的DNA序列)转化的细胞中，选择的Cas9直系同源物在HEK细胞RunX1基因座处的结果。

[0049] 图19显示了由选择的Cas9直系同源物产生的20个不同突变体中的预期剪切位点。图19A显示了在玉蜀黍细胞中ID46的结果，并且图19B显示了在玉蜀黍细胞中ID56的结果。

[0050] 图20显示了与酿脓链球菌Cas9的活性相比，用核糖核蛋白(所述核糖核蛋白包含各自Cas9直系同源物及其适当指导RNA)转化的细胞中，选择的Cas9直系同源物在HEK细胞WTAP基因座处的结果。

[0051] 序列

[0052] SEQ ID NO:1-85是分别编码Cas9直系同源物序列SEQ ID 86-170的多核苷酸序列，其中Cas9直系同源物序列ID号、来源生物和系统进化枝描述在表1中。

[0053] SEQ ID NO:86-170和511-1135是编码图1中所示的Cas9直系同源物的多肽序列。

[0054] SEQ ID NO:171-255分别是对应于SEQ ID 86-170的Cas9直系同源物的crRNA重复序列。

[0055] SEQ ID NO:256-340分别是对应于SEQ ID 86-170的Cas9直系同源物的反重复序列。

[0056] SEQ ID NO:341-425是分别对应于SEQ ID 86-170的Cas9直系同源物的3’tracrRNA序列。

[0057] SEQ ID NO:426-510是分别对应于SEQ ID 86-170的Cas9直系同源物的sgRNA序列的CER结构域。

[0058] SEQ ID NO:1136-1220是表2B中列出的Cas9直系同源物ID号的REC结构域的蛋白序列。

[0059] SEQ ID NO:1221-1305是表2B中列出的Cas9直系同源物ID号的RUVC1结构域的蛋白序列。

[0060] SEQ ID NO:1306-1390是表2B中列出的Cas9直系同源物ID号的RUVC2结构域的蛋白序列。

[0061] SEQ ID NO:1391-1475是表2B中列出的Cas9直系同源物ID号的RUVC3结构域的蛋白序列。

[0062] SEQ ID NO:1476-1560是表2B中列出的Cas9直系同源物ID号的HNH结构域的蛋白序列。

[0063] SEQ ID NO:1561-1645是表2B中列出的Cas9直系同源物ID号的WED结构域的蛋白序列。

[0064] SEQ ID NO:1646-1730是表2B中列出的Cas9直系同源物ID号的PI结构域的蛋白序

列。

[0065] SEQ ID NO:1731是衔接子A1的DNA序列。

[0066] SEQ ID NO:1732是衔接子A2的DNA序列。

[0067] SEQ ID NO:1733是R0引物的DNA序列。

[0068] SEQ ID NO:1734是C0引物的DNA序列。

[0069] SEQ ID NO:1735是F1引物的DNA序列。

[0070] SEQ ID NO:1736是R1引物的DNA序列。

[0071] SEQ ID NO:1737是5′末端桥扩增序列的DNA序列。

[0072] SEQ ID NO:1738是3′末端桥扩增序列的DNA序列。

[0073] SEQ ID NO:1739是F2引物的DNA序列。

[0074] SEQ ID NO:1740是R2引物的DNA序列。

[0075] SEQ ID NO:1741是C1引物的DNA序列。

[0076] SEQ ID NO:1742是序列产物的DNA序列。

[0077] SEQ ID NO:1743是衔接子和靶的DNA序列。

[0078] SEQ ID NO:1744是PAM上游5′序列的DNA序列。

[0079] SEQ ID NO:1746是ID33 WT切割模式的DNA靶序列。

[0080] SEQ ID NO:1747-1766是ID33的前20个靶序列切割模式。

[0081] SEQ ID NO:1767是ID64 WT切割模式的DNA靶序列。

[0082] SEQ ID NO:1768-1787是ID64的前20个靶序列切割模式。

[0083] SEQ ID NO:1788是ID46 WT切割模式的DNA靶序列。

[0084] SEQ ID NO:1789-1808是ID46的前20个靶序列切割模式。

[0085] SEQ ID NO:1809是ID56 WT切割模式的DNA靶序列。

[0086] SEQ ID NO:1810-1829是ID56的前20个靶序列切割模式。

具体实施方式

[0087] 提供了用于新颖Cas9系统和包含这样的系统的元件的组合物，所述组合物包括但不限于新颖的指导多核苷酸/Cas内切核酸酶复合物、单指导RNA、指导RNA元件和Cas9内切核酸酶。本公开进一步包括用于细胞基因组中的靶序列的基因组修饰、用于基因编辑、以及用于将目的多核苷酸插入细胞基因组中的组合物和方法。

[0088] 还提供了用于直接递送内切核酸酶、Cas蛋白、指导RNA和指导RNA/内切核酸酶复合物的组合物和方法。本公开进一步包括用于细胞基因组中的靶序列的基因组修饰、用于基因编辑、以及用于将目的多核苷酸插入细胞基因组中的组合物和方法。

[0089] 还提供了用于体外表征和修饰分离的多核苷酸的组合物和方法。

[0090] 鉴于II型CRISPR-Cas系统的多样性(Fonfara等人.(2014)Nucleic Acids Res.[核酸研究]42:2577-2590)，合理的是许多Cas9内切核酸酶和同源指导RNA可能具有不同于先前描述的或表征的唯一的序列识别和酶特性。例如，切割活性和特异性可能被增强或前间隔子邻近基序(PAM)序列可能是不同的，导致增加的基因组靶位点密度。为了利用这一巨大的未开发的多样性并扩展可用于基因组靶向的Cas9内切核酸酶和关联指导RNA的储库，需要为每个新系统建立Cas9靶位识别的两个组分，PAM序列和指导RNA(双链CRISPR RNA

(crRNA)和反式激活CRISPR RNA(tracrRNA)或crRNA和tracrRNA的嵌合融合体(单指导RNA(sgRNA))。

[0091] 如本文所述,通过搜索由微生物基因组组成的内部先锋-杜邦(Pioneer-DuPont)数据库,鉴定了来自未表征的CRISPR-Cas系统的CRISPR-Cas基因座(包括Cas9基因和可读框、CRISPR阵列和反重复序列)。本文所述的Cas9内切核酸酶可以通过本领域已知的方法表达和纯化。如本文所述,可以推导所有CRISPR-Cas系统的tracrRNA的转录方向,并且是针对本文所述的每种新的不同的CRISPR-Cas内切核酸酶鉴定sgRNA及其组分(可变靶向结构域(VT))、crRNA重复序列、环、反重复序列和3' tracrRNA的实例。

[0092] 除非另有指定,否则权利要求书和说明书中使用的术语如下文阐述定义。必须注意,除非上下文另外清楚地指明,否则如本说明书及所附权利要求书中所用,单数形式"一个/一种(a/an)"和"该(the)"包括复数指示物。

[0093] 定义

[0094] 如本文所用,"核酸"意指多核苷酸,并且包括脱氧核糖核苷酸或核糖核苷酸碱基的单链或双链聚合物。核酸还可以包括片段和修饰的核苷酸。因此,术语"多核苷酸"、"核酸序列"、"核苷酸序列"和"核酸片段"可互换使用以表示单链或双链的RNA和/或DNA和/或RNA-DNA的聚合物,任选地包含合成的、非天然存在的或改变的核苷酸碱基。核苷酸(通常发现处于其5′-单磷酸形式)可以通过其单字母名称表示如下:"A"用于腺苷或脱氧腺苷(分别针对RNA或DNA),"C"用于胞嘧啶或脱氧胞嘧啶,"G"用于鸟苷或脱氧鸟苷,"U"用于尿苷,"T"用于脱氧胸苷,"R"用于嘌呤(A或G),"Y"用于嘧啶(C或T),"K"用于G或T,"H"用于A或C或T,"I"用于肌苷,并且"N"用于任何核苷酸。

[0095] 术语"基因组"当应用于原核或真核细胞或生物体细胞时不仅涵盖在细胞核内发现的染色体DNA,还涵盖在细胞的亚细胞组分(例如线粒体、或质体)内发现的细胞器DNA。

[0096] "可读框"缩写为ORF。

[0097] 术语"选择性地杂交"或"选择性杂交"包括参考在严格的杂交条件下将核酸序列杂交到特定的核酸靶序列上,相比其杂交到非靶核酸序列和基本上排除非靶核酸,该杂交达到可检测地更大程度(例如,至少为背景值的2倍)。选择性杂交序列典型地彼此具有约至少80%序列同一性、或90%序列同一性、高达并且包括100%序列同一性(即,完全互补)。

[0098] 术语"严格条件"或"严格杂交条件"包括提及在体外杂交测定中多核苷酸/探针将与其靶序列选择性杂交的条件。严格条件是序列依赖性的,并且在不同情况下将有所不同。通过控制杂交条件和/或洗涤条件的严格性,可以鉴定与多核苷酸/探针100%互补的靶序列(同源探测)。可替代地,可以调节严格条件以允许序列中的一些错配,以便检测到更低程度的相似性(异源探测)。通常,多核苷酸/探针的长度为少于约1000个核苷酸、少于500个核苷酸、少于100个核苷酸、少于90个核苷酸、少于80个核苷酸、少于70个核苷酸、少于60个核苷酸、少于50个核苷酸、少于40个核苷酸、少于30个核苷酸、少于20个核苷酸、10个核苷酸或甚至少于10个核苷酸。典型地,严格条件将是以下条件:在pH 7.0至8.3下盐浓度为小于约1.5M Na离子、典型地约0.01至1.0M Na离子浓度(或其他一种或多种盐),并且对于短多核苷酸/探针(例如,10至50个核苷酸)为至少30℃,并且对于长多核苷酸/探针(例如,大于50个核苷酸)为至少60℃。添加去稳定剂如甲酰胺也可以实现严格条件。示例性低严格条件包括在37℃下与30%至35%甲酰胺、1M NaCl、1%SDS(十二烷基硫酸钠)的缓冲溶液杂交,并

且在50℃至55℃下在1X至2X SSC(20X SSC＝3.0M NaCl/0.3M柠檬酸三钠)中洗涤。示例性中严格条件包括在37℃下在40％至45％甲酰胺、1M NaCl、1％SDS中杂交，并且在55℃至60℃下在0.5X至1X SSC中洗涤。示例性高严格条件包括在37℃下在50％甲酰胺、1M NaCl、1％SDS中杂交，并且在60℃至65℃下在0.1X SSC中洗涤。

[0099] “同源”意指DNA序列是相似的。例如，在供体DNA上发现的“与基因组区域同源的区域”是与细胞或生物体基因组中给定的“基因组序列”具有类似序列的DNA的区域。同源的区域可以具有足以促进在切割的靶位点处的同源重组的任何长度。例如，同源的区域的长度可以包括至少5-10、5-15、5-20、5-25、5-30、5-35、5-40、5-45、5-50、5-55、5-60、5-65、5-70、5-75、5-80、5-85、5-90、5-95、5-100、5-200、5-300、5-400、5-500、5-600、5-700、5-800、5-900、5-1000、5-1100、5-1200、5-1300、5-1400、5-1500、5-1600、5-1700、5-1800、5-1900、5-2000、5-2100、5-2200、5-2300、5-2400、5-2500、5-2600、5-2700、5-2800、5-2900、5-3000、5-3100或更多个碱基，这样使得同源的区域具有充足同源性，从而经历与相应的基因组区域的同源重组。“足够的相似性”指示两个多核苷酸序列具有足够的结构等同性以充当同源重组反应的底物。结构等同性包括每个多核苷酸片段的总长度以及多核苷酸的序列相似性。序列相似性可以通过在序列的整个长度上的百分比序列同一性和/或通过包含局部相似性(例如具有100％序列同一性的连续核苷酸)的保守区域以及在序列长度的一部分上的百分比序列同一性来描述。

[0100] 如本文所用，“基因组区域”是存在于靶位点任一例上的细胞的基因组中的染色体的区段，或者可替代地，还包含靶位点的一部分。基因组区域可以包含至少5-10、5-15、5-20、5-25、5-30、5-35、5-40、5-45、5-50、5-55、5-60、5-65、5-70、5-75、5-80、5-85、5-90、5-95、5-100、5-200、5-300、5-400、5-500、5-600、5-700、5-800、5-900、5-1000、5-1100、5-1200、5-1300、5-1400、5-1500、5-1600、5-1700、5-1800、5-1900、5-2000、5-2100、5-2200、5-2300、5-2400、5-2500、5-2600、5-2700、5-2800。5-2900、5-3000、5-3100或更多个碱基，这样使得基因组区域具有足够的相似性以与相应的同源区域进行同源重组。

[0101] 如本文所用，“同源重组(HR)”包括在同源的位点处的两个DNA分子之间的DNA片段的交换。同源重组的频率受多个因素影响。不同的生物体相对于同源重组的量和同源与非同源重组的相对比例而变化。通常，同源区域的长度会影响同源重组事件的频率：同源区域越长，频率越高。为观察同源重组而需要的同源区域的长度也是随物种而异的。在许多情况下，已经利用了至少5kb的同源性，但已经观察到具有仅25-50bp的同源性的同源重组。参见，例如，Singer等人，(1982)Cell[细胞]31:25-33;Shen和Huang，(1986)Genetics[遗传学]112:441-57;Watt等人，(1985)Proc.Natl.Acad.Sci.USA[美国国家科学院院刊]82:4768-72,Sugawara和Haber，(1992)Mol Cell Biol[分子细胞生物学]12:563-75,Rubnitz和Subramani，(1984)Mol Cell Biol[分子细胞生物学]4:2253-8;Ayares等人，(1986)Proc.Natl.Acad.Sci.USA[美国国家科学院院刊]83:5199-203;Liskay等人，(1987)Genetics[遗传学]115:161-7。

[0102] 在核酸的或多肽的序列的上下文中，“序列同一性”或“同一性”是指在两个序列中的核酸碱基或氨基酸残基当在指定的比较窗口上比对最大对应度时是相同的。

[0103] “序列同一性的百分比”是指通过在比较窗口上比较两个最佳比对的序列所确定的值，其中与参考序列(其不包含添加或缺失)比较两个序列的最佳比对时，该多核苷酸或

多肽序列在比较窗口中的部分可以包含添加或缺失(即空位)。通过以下方式计算所述百分比:确定在两个序列中出现相同核酸碱基或氨基酸残基的位置的数目以产生匹配位置的数目,将匹配位置的数目除以比较窗口中的位置的总数目,然后将所述结果乘以100以产生序列同一性的百分比。百分比序列同一性的有用实例包括但不限于50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%、100%或从50至100%的任何增量或分数百分比。可以使用本文描述的任何程序确定这些同一性。

[0104] 序列比对和百分比同一性或相似性计算可以使用设计用于检测同源序列的多种比较方法来确定,这些方法包括但不限于LASERGENE生物信息计算包(DNASTAR公司(DNASTAR Inc.),麦迪逊(Madison),威斯康星州)的MegAlign™程序。在此申请的上下文中,应当理解的是,在使用序列分析软件来分析的情况下,分析的结果将基于参考的程序的"默认值",除非另有说明。如本文所用,"默认值"将意指当第一次初始化时,最初加载该软件的任何一组值或参数。

[0105] "Clustal V比对方法"对应于标记为Clustal V的比对方法(由Higgins和Sharp,(1989)CABIOS 5:151-153;Higgins等人,(1992)Comput Appl Biosci[生物学中的计算机应用]8:189-191描述),并见于LASERGENE生物信息学计算套件的MegAlign™程序(DNASTAR公司,威斯康辛州麦迪逊)。对于多重比对,默认值对应于空位罚分(GAP PENALTY)=10和空位长度罚分(GAP LENGTH PENALTY)=10。使用Clustal方法进行逐对比对和蛋白序列的百分比同一性计算的默认参数为KTUPLE=1、空位罚分=3、窗口(WINDOW)=5、以及存储的对角线(DIAGONALS SAVED)=5。对于核酸,这些参数是KTUPLE=2、空位罚分=5、窗口=4、并且存储的对角线=4。使用Clustal V程序比对序列后,可能通过查看同一程序中的"序列距离"表来获得"百分比同一性"。"Clustal W比对方法"对应于标记为Clustal W的比对方法(由Higgins和Sharp,(1989)CABIOS 5:151-153;Higgins等人,(1992)Comput Appl Biosci[生物学中的计算机应用]8:189-191描述),并见于LASERGENE生物信息学计算套件的MegAlign™v6.1程序(DNASTAR公司,威斯康辛州麦迪逊)。用于多重比对的默认参数(空位罚分=10、空位长度罚分=0.2、延迟发散序列(Delay Divergen Seqs,%)=30、DNA转换权重=0.5、蛋白权重矩阵=Gonnet系列、DNA权重矩阵=IUB)。除非另有说明,本文中提供的序列同一性/相似性值是指使用GAP版本10(GCG,Accelrys公司,圣迭戈,加利福尼亚州)使用以下参数获得的值:核苷酸序列的%同一性和%相似性使用空位创建罚分权重为50、空位长度延伸罚分权重为3、以及nwsgapdna.cmp打分矩阵;氨基酸序列的%同一性和%相似性使用空位创建罚分权重为8、空位长度延伸罚分为2、以及BLOSUM62打分矩阵(Henikoff和Henikoff,(1989)Proc.Natl.Acad.Sci.USA[美国国家科学院院刊]89:10915)。GAP使用Needleman和Wunsch(1970)J Mol Biol[分子生物学杂志]48:443-53的算法来找到使匹配数目最大化并且使空位数目最小化的两个完整序列的比对。GAP考虑所有可能的比对和空位位置,并且使用匹配碱基的单位中的空位产生罚分和空位延伸罚分,产生具有最大数目的匹配碱基和最少的空位的比对。"BLAST"是美国国家生物技术信息中心(National Center for Biotechnology Information,NCBI)提供的用于寻找生物序列之间的相似性的区域的搜索算法。该程序将核苷酸或者蛋白序列与序列数据库比较,并计算匹配的统计显著性以鉴定出与查询序列具有足够的相似性的序列,这样使得相似性不会被预测为已经随机发生。BLAST报告鉴定的序列和它们与查询序列的局部比对。本领域技术人员很清楚地

理解，许多水平的序列同一性在鉴定来自其他物种的多肽或修饰的天然的或合成的多肽中是有用的，其中这样的多肽具有相同或相似的功能或活性。百分比同一性的有用实例包括但不限于50％、55％、60％、65％、70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、100％或从50至100％的任何增量或分数百分比。实际上，在描述本公开中，从50％至100％的任何氨基酸同一性会是有用的，如51％、52％、53％、54％、55％、56％、57％、58％、59％、60％、61％、62％、63％、64％、65％、66％、67％、68％、69％、70％、71％、72％、73％、74％、75％、76％、77％、78％、79％、80％、81％、82％、83％、84％、85％、86％、87％、88％、89％、90％、91％、92％、93％、94％、95％、96％、97％、98％或99％。

[0106] 多核苷酸和多肽序列、其变体、以及这些序列的结构关系，可用术语“同源性”、“同源的”、“基本上相同的”、“基本上类似的”、以及“基本上相应”来描述，这些术语在本文中可互换使用。这些是指多肽或核酸序列，其中在一个或多个氨基酸或核苷酸碱基上的变化不影响分子的功能，如介导基因表达或产生某种表型的能力。这些术语还指相对于初始未修饰的核酸，基本上不改变所得核酸的功能特性的核酸序列的一个或多个修饰。这些修饰包括核酸片段中一个或多个核苷酸的缺失、取代和/或插入，或原子或分子与多核苷酸中现有核苷酸的缔合(例如但不限于：一个甲基的共价添加，或与金属离子的离子相互作用)。所涵盖的基本上类似的核酸序列可以通过这些核酸序列与本文所示例的序列杂交，或与本文所公开的并且与任何本文所公开的核酸序列在功能上等价的核苷酸序列的任何部分杂交(在中严格条件下，例如0.5X SSC，0.1％SDS，60℃)的能力来定义。可以调整严格条件以筛选适度类似的片段(如来自远缘生物体的同源序列)，至高度类似的片段(如复制来自近缘生物体的功能性酶的基因)。杂交后的洗涤决定了严格条件。

[0107] “厘摩”(cM)或“图距单位”是两个多核苷酸序列、连锁的基因、标志物、靶位点、基因座或它们的任何配对之间的距离，其中1％的减数分裂的产物是重组的。因此，一厘摩与等于两个连锁的基因、标志物、靶位点、基因座或它们的任何配对之间的1％平均重组频率的距离相当。

[0108] “分离的”或“纯化的”核酸分子、多核苷酸、多肽或蛋白或其生物活性部分是基本上或本质上不含与如在其天然存在的环境中发现的多核苷酸或蛋白正常相伴或相互作用的组分。因此，分离的或纯化的多核苷酸或多肽或蛋白当通过重组技术产生时基本上不含其他细胞物质或培养基，或者当化学合成时基本上不含化学前体或其他化学品。最佳地，“分离的”多核苷酸不含在从其衍生出该多核苷酸的生物体的基因组DNA中天然地在该多核苷酸侧翼的序列(即，位于该多核苷酸的5′和3′末端的序列)(最佳地是蛋白编码序列)。例如，在不同实施例中，该分离的多核苷酸可以包含小于约5kb、4kb、3kb、2kb、1kb、0.5kb或0.1kb的核苷酸序列，在该多核苷酸从其衍生出的细胞的基因组DNA中，该核苷酸序列天然地位于该多核苷酸的侧翼。分离的多核苷酸可从它们天然存在于其中的细胞纯化。技术人员已知的常规核酸纯化方法可用于获得分离的多核苷酸。该术语也涵盖重组多核苷酸和化学合成的多核苷酸。

[0109] 术语“片段”是指一组连续的多核苷酸或多肽。在一个实施例中，片段是2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20或大于20个连续的多核苷酸。在一个实施例中，片段是2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20或大于20个连续的多肽。片段可能表现出或可能不会表现出在所述片段的长度上共享一定百分比同一性的序列

的功能。

[0110] 术语“在功能上等价的片段”和“功能等价片段”在本文中可互换使用。这些术语是指分离的核酸片段或多肽的显示出与其衍生自的较长序列相同的活性或功能的一部分或子序列。在一个实例中，无论片段是否编码活性蛋白，该片段都保留改变基因表达或产生某种表型的能力。例如，片段可用于设计基因以在修饰的植物中产生所希望的表型。可以将基因设计为用于在抑制中使用，无论该基因是否编码活性酶，通过以相对于启动子序列的有义或反义取向连接其核酸片段。

[0111] “基因”包括表达功能性分子(诸如但不限于，特定蛋白)的核酸片段，包括在编码序列之前(5′非编码序列)和之后(3′非编码序列)的调节序列。“天然基因”是指在其天然内源位置中发现的具有其自身调节序列的基因。

[0112] 术语“内源”是指天然存在于细胞或生物体中的序列或其他分子。在一个方面，内源多核苷酸通常存在于其所来源的细胞的基因组中；也就是说，不是异源的。

[0113] “等位基因”是占据染色体上给定基因座的基因的若干种替代形式中的一种。当染色体上在给定基因座处存在的所有等位基因都相同时，该植物在该基因座处是纯合的。如果染色体上在给定基因座处存在的等位基因不同，则该植物在该基因座处是杂合的。

[0114] “编码序列”是指可被转录成RNA分子并任选地进一步翻译成多肽的多核苷酸序列。“调节序列”是指位于编码序列的上游(5′非编码序列)、内部或下游(3′非编码序列)的核苷酸序列，并且其影响相关的编码序列的转录、RNA加工或稳定性、或翻译。调节序列包括但不限于：启动子、翻译前导序列、5′非翻译序列、3′非翻译序列、内含子、聚腺苷酸化靶序列、RNA加工位点、效应子结合位点、和茎环结构。

[0115] “突变基因”是通过人为干预已经改变的基因。这样的“突变基因”具有通过至少一个核苷酸添加、缺失或取代而与相应的非突变基因的序列不同的序列。在本公开的某些实施例中，该突变的基因包含由如本文公开的指导多核苷酸/Cas内切核酸酶系统引起的改变。突变的植物是包含突变基因的植物。

[0116] 如本文所用，术语“靶向突变”是通过使用本领域技术人员已知的任何方法(包括涉及如本文公开的受指导的Cas内切核酸酶系统的方法)改变靶基因内的靶序列而产生的基因(称为靶基因)包括天然基因中的突变。

[0117] 术语“敲除”、“基因敲除”和“基因的敲除”在本文中可互换使用。敲除表示已经通过用Cas蛋白进行靶向使得细胞的DNA序列部分或完全无效；例如，这样的DNA序列在敲除之前可能已编码氨基酸序列，或可能已具有调节功能(例如，启动子)。

[0118] 术语“敲入”、“基因敲入”、“基因插入”和“基因的敲入”在本文中可互换使用。敲入代表通过用Cas蛋白(例如通过同源重组(HR)，其中还使用适合的供体DNA多核苷酸)靶向在细胞中的特异性DNA序列处进行的DNA序列的替换或插入。敲入的实例是异源氨基酸编码序列在基因的编码区中的特异性插入，或转录调节元件在遗传基因座中的特异性插入。

[0119] “结构域”意指核苷酸(可以为RNA、DNA和/或RNA-DNA组合序列)或氨基酸的连续延伸。

[0120] 术语“保守结构域”或“基序”是指沿进化相关蛋白的比对序列在特定位置处保守的一组多核苷酸或氨基酸。虽然同源蛋白之间在其他位置处的氨基酸可以发生变化，但在特定位置处高度保守的氨基酸表明对蛋白的结构、稳定性或活性来说是必需的氨基酸。因

为它们通过蛋白同系物家族的比对序列中的高度保守性而被鉴定，所以它们可以用作标识符或“特征”，以确定具有新确定的序列的蛋白是否属于先前鉴定的蛋白家族。

[0121] “密码子修饰的基因”或“密码子偏好的基因”或“密码子优化的基因”是其密码子使用的频率被设计为模拟宿主细胞的偏好的密码子使用的频率的基因。

[0122] “优化的”多核苷酸是已经过优化以改善特定异源宿主细胞中的表达或功能的序列。

[0123] “植物优化的核苷酸序列”是为在植物中表达或功能(特别是为了在植物中增加的表达)而优化的核苷酸序列。植物优化的核苷酸序列包括密码子优化的基因。可以使用一个或多个植物偏好的密码子来改善表达，通过修饰编码蛋白(诸如像本文公开的Cas内切核酸酶)的核苷酸序列，来合成植物偏好的核苷酸序列。参见，例如，Campbell和Gowri(1990)Plant Physiol.[植物生理学]92:1-11对宿主偏好的密码子使用的讨论。

[0124] “启动子”是参与RNA聚合酶和其他蛋白的识别和结合以起始转录的DNA区域。启动子序列由近端元件和较远端上游元件组成，后一元件通常称为增强子。“增强子”是可以刺激启动子活性的DNA序列，并且可以是该启动子的固有元件或被插入以增强启动子的水平或组织特异性的异源元件。启动子可以全部来源于天然基因，或者由来源于在自然界存在的不同启动子的不同元件构成，和/或包含合成的DNA区段。本领域技术人员应当理解，不同的启动子可能引导基因在不同组织或细胞类型中、或在不同发育阶段、或者响应于不同环境条件的表达。进一步认识到，由于在大多数情况下调节序列的确切边界尚未完全限定，一些变异的DNA片段可能具有相同的启动子活性。

[0125] 在多数情况下引起基因在大多数细胞型中表达的启动子通常称为“组成型启动子”。术语“诱导型启动子”是指对内源或外源刺激的存在，例如通过化学化合物(化学诱导剂)响应，或对环境、激素、化学品、和/或发育信号响应，选择性表达编码序列或功能RNA的启动子。诱导型或调节型启动子包括例如通过光、热、胁迫、水淹或干旱、盐胁迫、渗透胁迫、植物激素、伤口或化学品(如乙醇、脱落酸(ABA)、茉莉酮酸酯、水杨酸或安全剂)诱导或调节的启动子。

[0126] “翻译前导序列”是指位于基因的启动子序列和编码序列之间的多核苷酸序列。翻译前导序列存在于翻译起始序列的mRNA上游。翻译前导序列可以影响初级转录物对mRNA的加工、mRNA稳定性、或翻译效率。已经描述了翻译前导序列的实例(例如，Turner和Foster，(1995)Mol Biotechnol[分子生物技术]3:225-236)。

[0127] “3′非编码序列”、“转录终止子”、或“终止序列”是指位于编码序列的下游的DNA序列，并且包括聚腺苷酸化识别序列和编码能够影响mRNA加工或基因表达的调节信号的其他序列。聚腺苷酸化信号通常表征为影响聚腺苷酸片添加到mRNA前体的3′末端。由Ingelbrecht等人，(1989)Plant Cell[植物细胞]1:671-680示例了不同的3′非编码序列的用途。

[0128] “RNA转录物”是指由DNA序列的RNA聚合酶催化的转录产生的产物。当RNA转录物是DNA序列的完全互补拷贝时，RNA转录物被称为初级转录物或前mRNA。当RNA转录物是源自初级转录物前mRNA的转录后加工的RNA序列时，RNA转录物被称为成熟RNA或mRNA。“信使RNA”或“mRNA”是指不含内含子并且可以被细胞翻译成蛋白的RNA。“cDNA”是指与mRNA模板互补并且使用逆转录酶从mRNA模板合成的DNA。cDNA可以是单链的或者可以使用DNA聚合成酶I

的Klenow片段转化成双链形式。“有义”RNA是指包含mRNA并且可以在细胞内或体外翻译成蛋白的RNA转录物。“反义RNA”是指与靶初级转录物或mRNA的全部或部分互补、并且阻断靶基因的表达的RNA转录物(参见,例如美国专利号5,107,065)。反义RNA可与特定基因转录物的任何部分,即5′非编码序列、3′非编码序列、内含子或编码序列互补。“功能性RNA”是指反义RNA、核糖酶RNA、或可以不进行翻译而仍对细胞过程具有作用的其他RNA。术语“互补序列”和“反向互补序列”在本文中关于mRNA转录物可互换使用,并且意在限定信使的反义RNA。

[0129] 术语“基因组”意指存在于生物体或病毒或细胞器的每个细胞中的遗传物质的全部互补序列(基因和非编码序列);和/或从一个亲本遗传为(单倍体)单位的完整染色体组。

[0130] 术语“可操作地连接”是指单个核酸片段上的核酸序列的关联,这样使得其中一个核酸序列的功能被另一个核酸序列调节。例如,当启动子能够调节编码序列的表达(即,该编码序列在启动子的转录控制下)时,启动子与该编码序列可操作地连接。编码序列可以在有义或反义取向上可操作地连接到调节序列。在另一个实例中,互补的RNA区域可以直接或间接与靶mRNA的5′、或靶mRNA的3′可操作地连接、或在靶mRNA内,或第一互补区是5′且其互补序列是靶mRNA的3′。

[0131] 通常,“宿主”是指已引入异源组分(多核苷酸、多肽、其他分子、细胞)的生物体或细胞。如本文所用,“宿主细胞”是指体内或体外的真核细胞、原核细胞(例如,细菌或古细菌细胞),或来自作为单细胞实体培养的多细胞生物体的细胞(例如,细胞系),其中已引入异源多核苷酸或多肽。在一些实施例中,所述细胞选自下组,所述组由以下组成:原始细胞、细菌细胞、真核细胞、真核单细胞生物体、体细胞、生殖细胞、干细胞、植物细胞、藻类细胞、动物细胞、无脊椎动物细胞、脊椎动物细胞、鱼类细胞、青蛙细胞、鸟类细胞、昆虫细胞、哺乳动物细胞、猪细胞、牛细胞、山羊细胞、绵羊细胞、啮齿动物细胞、大鼠细胞、小鼠细胞、非人类的灵长类动物细胞和人类细胞。在一些情况下,该细胞是体外细胞。在一些情况下,该细胞是体内细胞。

[0132] 术语“重组”是指例如通过化学合成或者通过基因工程技术操纵分离的核酸区段来将两个原本分开的序列区段进行人工组合。

[0133] 术语“质粒”、“载体”和“盒”是指线性或环状染色体外元件,其通常携带非细胞中心代谢的一部分的基因,并且通常呈双链DNA的形式。这样的元件可以是衍生自任何来源的、单链或双链DNA或RNA的、处于直链或环状形式的自主复制序列、基因组整合序列、噬菌体、或核苷酸序列,其中许多核苷酸序列已经被连接或重组成能够将目的多核苷酸引入细胞中的独特构造。“转化盒”是指包含基因并具有促进特定宿主细胞转化的基因之外的元件的特定载体。“表达盒”是指包含基因并具有允许在宿主中表达该基因的基因之外的元件的特定载体。

[0134] 术语“重组DNA分子”、“重组DNA构建体”、“表达构建体”、“构建体”、和“重组构建体”在本文中可互换使用。重组DNA构建体包含核酸序列,例如在自然界中未全部一起发现的调节序列和编码序列的人工组合。例如,重组DNA构建体可以包含衍生自不同来源的调节序列和编码序列,或者包含衍生自相同来源但以不同于天然发生的方式排列的调节序列和编码序列。这种构建体可以单独使用或可以与载体结合使用。如果使用载体,则载体的选择取决于如本领域技术人员熟知的将用于将载体引入宿主细胞的方法。例如,可以使用质粒

载体。技术人员充分了解必须存在于载体上以便成功转化，选择和繁殖宿主细胞的遗传元件。本领域技术人员还将认识到，不同的独立转化事件可能导致不同的表达水平和模式(Jones等人，(1985)EMBO J[欧洲分子生物学组织杂志]4:2411-2418;De Almeida等人，(1989)Mol Gen Genetics[分子遗传学和普通遗传学]218:78-86)，因此典型地筛选多个事件，以获得显示所希望的表达水平和模式的品系。此类筛选可以是完成的标准分子生物学测定、生物化学测定以及其他测定，这些测定包括DNA的印迹分析、mRNA表达的Northern分析、PCR、实时定量PCR(qPCR)、逆转录PCR(RT-PCR)、蛋白表达的免疫印迹分析、酶测定或活性测定、和/或表型分析。

[0135] 术语“异源”是指特定多核苷酸或多肽序列的原始环境、位置或组成与其当前环境、位置或组成之间的差异。非限制性实例包括分类学衍生的差异(例如，如果从玉蜀黍(Zea mays)获得的多核苷酸序列插入到水稻(Oryza sativa)植物的基因组或玉蜀黍的不同变种或栽培品种的基因组中，则该多核苷酸序列是异源的;或从细菌获得的多核苷酸被引入植物的细胞中，则该多核苷酸序列是异源的)或序列的差异(例如从玉蜀黍获得的多核苷酸序列被分离、修饰并重新引入玉蜀黍植物中)。如本文所用，关于序列的“异源”可以指该序列源于不同物种、变种、外来物种，或者，如果源于相同物种的话，则是通过蓄意人为干预从其在组合物和/或基因组基因座中的天然形式进行实质性修饰得到的序列。例如，可操作地连接至异源多核苷酸的启动子来自与从其衍生该多核苷酸的物种不同的物种，或者，如果来自相同/类似的物种，那么一方或双方基本上由它们的原来形式和/或基因组基因座修饰得到，或者该启动子不是被可操作地连接的多核苷酸的天然启动子。可替代地，本文提供的一个或多个调节区域和/或多核苷酸可以是整体地合成的。

[0136] 如本文所用，术语“表达”是指处于前体抑或成熟形式的功能性终产物(例如，mRNA、指导RNA或蛋白)的产生。

[0137] “成熟”蛋白是指翻译后加工的多肽(即，从其中已经去除存在于初级翻译产物中的任何前肽(pre-peptide)或原肽(propeptide)的一种多肽)。

[0138] “前体”蛋白是指mRNA的翻译的初级产物(即，仍存在前肽或原肽)。前肽或原肽可以是但不限于细胞内定位信号。

[0139] “CRISPR”(成簇的规律间隔的短回文重复序列(Clustered Regularly Interspaced Short Palindromic Repeats))基因座是指DNA切割系统的某些遗传基因座编码组分，例如，被细菌和古细菌细胞用来破坏外源DNA的那些(Horvath和Barrangou，2010，Science[科学]327:167-170;2007年3月1日公开的WO 2007/025097)。CRISPR基因座可以由CRISPR阵列组成，包含由短的可变DNA序列(称为‘间隔子’)分开的短的正向重复序列(CRISPR重复序列)，其可以是侧翼不同Cas(CRISPR相关的)基因。

[0140] 如本文所用，“效应子”或“效应子蛋白”是具有包括识别、结合和/或切割多核苷酸靶或使多核苷酸靶产生切口的活性的蛋白。CRISPR系统的“效应子复合物”包括参与crRNA及靶识别和结合的Cas蛋白。一些组分Cas蛋白可以另外包含参与靶多核苷酸切割的结构域。

[0141] 术语“Cas蛋白”是指由Cas(CRISPR-相关的)基因编码的多肽。Cas蛋白包括但不限于：本文公开的新型Cas9直系同源物、Cas9蛋白、Cpf1(Cas12)蛋白、C2c1蛋白、C2c2蛋白、C2c3蛋白、Cas3、Cas3-HD、Cas5、Cas7、Cas8、Cas10或这些的组合或复合物。当与适合的多核

苷酸组分复合时，Cas蛋白可以是能够识别、结合特定DNA靶序列的全部或部分、并任选地使特定DNA靶序列的全部或部分产生切口或切割特定DNA靶序列的全部或部分的“Cas内切核酸酶”。本文描述的Cas内切核酸酶包含一个或多个核酸酶结构域。Cas蛋白被进一步定义为天然Cas蛋白的功能性片段或功能性变体，或与天然Cas蛋白的至少50个、50至100个、至少100个、100至150个、至少150个、150至200个、至少200个、200至250个、至少250个、250至300个、至少300个、300至350个、至少350个、350至400个、至少400个、400至450个、至少500个或大于500个连续氨基酸具有至少50％、50％至55％、至少55％、55％至60％、至少60％、60％至65％、至少65％、65％至70％、至少70％、70％至75％、至少75％、75％至80％、至少80％、80％至85％、至少85％、85％至90％、至少90％、90％至95％、至少95％、95％至96％、至少96％、96％至97％、至少97％、97％至98％、至少98％、98％至99％、至少99％、99％至100％或100％的序列同一性并且保留至少部分活性的蛋白。

[0142] Cas内切核酸酶的“功能性片段”、“功能上等效的片段”和“功能等效片段”在本文中可互换地使用，并且指本公开的Cas内切核酸酶的一部分或子序列，其中保留识别、结合靶位点并任选地使靶位点产生切口或切割(引入单链或双链断裂)靶位点的能力。该Cas内切核酸酶的部分或子序列可以包含具有其任何一个结构域的完整或部分(功能性)肽，诸如但不限于HD结构域的完整或功能性部分、解旋酶结构域的完整或功能性部分、内切核酸酶结构域的完整或功能性部分、与PAM相互作用的结构域的完整或功能性部分、楔入结构域的完整或功能性部分、RuvC结构域的完整或功能部分、锌指结构域的完整或功能性部分或Cas蛋白的完整或功能部分(如但不限于Cas9、Cpf1、Cas5、Cas5d、Cas7、Cas8b1、Cas1、Cas2、Cas4或Cas9直系同源物)。

[0143] 术语Cas内切核酸酶的“功能性变体”、“功能上等同的变体”和“功能等同变体”或Cas内切核酸酶，包括本文所述的Cas9直系同源物，在本文中可互换使用，并且是指本文所公开的Cas内切核酸酶的变体，其中保留了识别、结合以及任选地解旋、切口或切割全部或部分的靶序列的能力。

[0144] 在一些方面，功能性片段或功能性变体保留与其所衍生自的亲本分子大约相同的水平和类型(例如靶多核苷酸识别、结合和切割)的活性。在一些方面，功能性片段或功能性变体显示出与其所衍生自的亲本分子相同类型的活性(例如，增加的靶多核苷酸识别特异性)。在一些方面，功能性片段或功能性变体显示出与其所衍生自的亲本分子相同类型的活性降低(例如，较低的靶多核苷酸结合亲和力)。在一些方面，功能性片段或功能性变体显示出作为其所衍生自的亲本分子的部分活性(例如，多核苷酸识别和结合，但非切割)。在一些方面，功能性片段或功能性变体显示出与其所衍生自的亲本分子不同的活性类型(例如，在靶多核苷酸上产生单链切口相比于双链断裂)。根据从业者的需要，可以选择活性类型或水平的任何相似性或差异作为所希望的结果。

[0145] Cas内切核酸酶还可包括多功能Cas内切核酸酶。术语“多功能Cas内切核酸酶”和“多功能Cas内切核酸酶多肽”在本文中可互换使用，并且包括提及具有Cas内切核酸酶功能(包含至少一个可用作Cas内切核酸酶的蛋白结构域)和至少另一种功能的单个多肽，该至少另一种功能诸如但不限于，形成级联的功能(至少包括可与其他蛋白形成级联的第二蛋白结构域)。在一个方面，该多功能Cas内切核酸酶包含相对于Cas内切核酸酶的那些典型结构域的至少一个另外的蛋白结构域(在内部上游(5′)或下游(3′)，或在内部5′和3′两处，或

其任何组合)。

[0146] 术语“cascade”和“cascade复合物”在本文中可互换使用,并且包括提及可与多核苷酸组装形成多核苷酸-蛋白复合物(PNP)的多亚基蛋白复合物。cascade是一种依赖于多核苷酸的PNP,以实现复合物组装和稳定性以及鉴定靶核酸序列。cascade用作监视复合物,其发现并任选地结合与指导多核苷酸的可变靶向结构域互补的靶核酸。

[0147] 术语“切割就绪的Cascade”、“crCascade”、“切割就绪的Cascade复合物”、“crCascade复合物”、“切割就绪的Cascade系统”、“CRC”和“crCascade系统”在本文中可互换使用,并包括提及可以与多核苷酸组装形成多核苷酸-蛋白复合物(PNP)的多亚基蛋白复合物,其中cascade蛋白之一是Cas内切核酸酶,所述Cas内切核酸酶能够识别、结合靶序列的全部或部分、并任选地使靶序列的全部或部分解旋、使靶序列的全部或部分产生切口或切割靶序列的全部或部分。

[0148] 术语“5′-帽”和“7-甲基鸟苷酸(m7G)帽”在本文中可互换使用。7-甲基鸟苷酸残基位于真核生物中信使RNA(mRNA)的5′末端。在真核生物中,RNA聚合酶II(Pol II)转录mRNA。信使RNA加帽通常如下:用RNA末端磷酸酶去除mRNA转录物的最末端5’磷酸根基团,留下两个末端磷酸根。用鸟苷酸转移酶将一磷酸鸟苷(GMP)添加至转录物的末端磷酸根,在转录物末端处留下5′-5′三磷酸连接的鸟嘌呤。最后,此末端鸟嘌呤的7-氮被甲基转移酶甲基化。

[0149] 术语“不具有5′-帽”等在本文中用于指具有例如5′-羟基基团而不是5′-帽的RNA。例如,此类RNA可以被称为“未带帽的RNA”。因为5′-带帽的RNA有核输出的倾向,转录以后未带帽的RNA可以更好地积累在细胞核中。本文中的一种或多种RNA组分是未带帽的。

[0150] 如本文所用,术语“指导多核苷酸”涉及可以与Cas内切核酸酶(包括本文所述的Cas内切核酸酶)形成复合物,并且使得该Cas内切核酸酶能够识别、任选地结合并任选地切割DNA靶位点的多核苷酸序列。指导多核苷酸序列可以是RNA序列、DNA序列或其组合(RNA-DNA组合序列)。

[0151] 术语指导RNA、crRNA或tracrRNA的“功能性片段”、“功能上等效的片段”和“功能等效片段”在本文中可互换地使用,并且分别指本公开的指导RNA、crRNA或tracrRNA的一部分或子序列,其中分别保留用作指导RNA、crRNA或tracrRNA的能力。

[0152] 术语指导RNA、crRNA或tracrRNA(分别地)的“功能性变体”、“功能上等效的变体”和“功能等效变体”在本文中可互换地使用,并且分别指本公开的指导RNA、crRNA或tracrRNA的变体,其中分别保留用作指导RNA、crRNA或tracrRNA的能力。

[0153] 术语“单指导RNA”和“sgRNA”在本文中可互换使用,并涉及两个RNA分子的合成融合,其中包含可变靶向结构域(与tracrRNA杂交的tracr配对序列连接)的crRNA(CRISPR RNA)与tracrRNA(反式激活CRISPR RNA)融合。单指导RNA可以包含可与II型Cas内切核酸酶形成复合物的II型CRISPR/Cas系统的crRNA或crRNA片段和tracrRNA或tracrRNA片段,其中所述指导RNA/Cas内切核酸酶复合物可以将Cas内切核酸酶引导至DNA靶位点,使得Cas内切核酸酶能够识别、任选地结合DNA靶位点、并任选地使DNA靶位点产生切口或切割(引入单链或双链断裂)DNA靶位点。

[0154] 术语“可变靶向结构域”或“VT结构域”在本文中可互换使用,并且包括可以与双链DNA靶位点的一条链(核苷酸序列)杂交(互补)的核苷酸序列。第一核苷酸序列结构域(VT结构域)与靶序列之间的互补百分比可以为至少50%、51%、52%、53%、54%、55%、56%、

57％、58％、59％、60％、61％、62％、63％、63％、65％、66％、67％、68％、69％、70％、71％、72％、73％、74％、75％、76％、77％、78％、79％、80％、81％、82％、83％、84％、85％、86％、87％、88％、89％、90％、91％、92％、93％、94％、95％、96％、97％、98％、99％或100％。可变靶向结构域可以是至少12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29或30个核苷酸长度。在一些实施例中，可变靶向结构域包含12至30个核苷酸的连续延伸。可变靶向域可以由DNA序列、RNA序列、修饰的DNA序列、修饰的RNA序列或其任何组合构成。

[0155] 术语(指导多核苷酸的)“Cas内切核酸酶识别结构域”或“CER结构域”在本文中可互换地使用，并且包括与Cas内切核酸酶多肽相互作用的核苷酸序列。CER结构域包含(反式作用)tracr核苷酸伴侣序列，随后是tracr核苷酸序列。CER结构域可以由DNA序列、RNA序列、修饰的DNA序列、修饰的RNA序列(参见，例如，2015年2月26日公开的US 20150059010 A1)或其任何组合构成。

[0156] 如本文所用，术语“指导多核苷酸/Cas内切核酸酶复合物”、“指导多核苷酸/Cas内切核酸酶系统”、“指导多核苷酸/Cas复合物”、“指导多核苷酸/Cas系统”和“指导Cas系统”、“多核苷酸指导的内切核酸酶”、“PGEN”在本文中可互换使用，并且是指能够形成复合物的至少一种指导多核苷酸和至少一种Cas内切核酸酶，其中所述指导多核苷酸/Cas内切核酸酶复合物可以将Cas内切核酸酶引导至DNA靶位点，使Cas内切核酸酶能够对DNA靶位点进行识别、结合、并且任选地产生切口或进行切割(引入单链或双链断裂)。本文中的指导多核苷酸/Cas内切核酸酶复合物可包含一种或多种Cas蛋白和任何已知的CRISPR系统的一个或多个合适的多核苷酸组分(Horvath和Barrangou，2010，Science[科学]327:167-170；Makarova等人，2015，Nature Reviews Microbiology[自然微生物学综述]卷13:1-15；Zetsche等人，2015，Cell[细胞]163，1-13；Shmakov等人，2015，Molecular Cell[分子细胞]60，1-13)。

[0157] 术语“指导RNA/Cas内切核酸酶复合物”、“指导RNA/Cas内切核酸酶系统”、“指导RNA/Cas复合物”、“指导RNA/Cas系统”、“gRNA/Cas复合物”、“gRNA/Cas系统”、“RNA指导的内切核酸酶”，“RGEN”在本文中可互换地使用并且指至少一种RNA组分和至少一种能够形成复合物的Cas内切核酸酶，其中所述指导RNA/Cas内切核酸酶复合物可以将Cas内切核酸酶引导至DNA靶位点，使Cas内切核酸酶能够识别、结合DNA靶位点并任选地使DNA靶位点产生切口或切割(引入单链或双链断裂)DNA靶位点。在一些方面，提供这些组分作为Cas内切核酸酶蛋白和指导RNA的核糖核蛋白复合物(“RNP”)。

[0158] 术语“靶位点”、“靶序列”、“靶位点序列”、“靶DNA”、“靶基因座”、“基因组靶位点”、“基因组靶序列”、“基因组靶基因座”和“前间隔子”在本文中可互换地使用，并且是指多核苷酸序列，例如，但不限于，在细胞的染色体、附加体、基因座或基因组中的任何其他DNA分子(包括染色体DNA、叶绿体DNA、线粒体DNA、质粒DNA)上的核苷酸序列，在这些序列处指导多核苷酸/Cas内切核酸酶复合物可以进行识别、结合并任选地产生切口或进行切割。靶位点可以是细胞的基因组中的内源位点，或者可替代地，靶位点对于该细胞可以是异源的并且从而不是天然存在于细胞的基因组中，或者与在自然界发生的位置相比，可以在异质基因组位置中找到靶位点。如本文所用，术语“内源靶序列”和“天然靶序列”在本文中可互换使用，是指对细胞基因组来说是内源的或天然的、并且位于细胞的基因组中该靶序列的内源或天然位置处的靶序列。“人工靶位点”或“人工靶序列”在本文中可互换使用，并且是指

已经引入细胞的基因组中的靶序列。这样的人工靶序列可以在序列上与细胞的基因组中的内源或天然靶序列相同，但是位于细胞的基因组中的不同位置(即，非内源的或非天然的位置)处。

[0159] 本文中的“前间隔子邻近基序”(PAM)指与由本文所述的指导多核苷酸/Cas内切核酸酶系统识别的(靶向的)靶序列(前间隔子序列)邻近的短核苷酸序列。在一些方面，如果靶DNA序列与PAM序列不相邻或不邻近，则Cas内切核酸酶可能无法成功识别该靶DNA序列。在一些方面，该PAM在靶序列(例如，Cas12a)之前。在一些方面，该PAM在靶序列(例如，酿脓链球菌Cas9)之后。本文中的PAM的序列和长度可以取决于所使用的Cas蛋白或Cas蛋白复合物而不同。所述PAM序列可以是任何长度，但典型地是1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个核苷酸长度。

[0160] “改变的靶位点”、“改变的靶序列”、“经修饰的靶位点”、“经修饰的靶序列”在本文中可互换使用，并且是指如本文公开的靶序列，当与非改变的靶序列相比时，所述靶序列包括至少一个改变。此类“改变”包括，例如：(i)至少一个核苷酸的替代、(ii)至少一个核苷酸的缺失、(iii)至少一个核苷酸的插入、或(iv)(i)-(iii)的任何组合。

[0161] “经修饰的核苷酸”或“经编辑的核苷酸”是指当与其非修饰的核苷酸序列相比时，包含至少一个改变的目的核苷酸序列。此类“改变”包括，例如：(i)至少一个核苷酸的替代、(ii)至少一个核苷酸的缺失、(iii)至少一个核苷酸的插入、或(iv)(i)-(iii)的任何组合。

[0162] 用于“修饰靶位点”和“改变靶位点”的方法在本文中可互换使用，并且是指用于产生改变的靶位点的方法。

[0163] 如本文所用，“供体DNA”是DNA构建体，其包括待插入到Cas内切核酸酶的靶位点的目的多核苷酸。

[0164] 术语“多核苷酸修饰模板”包括，当与待编辑的核苷酸序列相比时，包含至少一个核苷酸修饰的多核苷酸。核苷酸修饰可以是至少一个核苷酸取代、添加或缺失。任选地，多核苷酸修饰模板可以进一步包含位于至少一个核苷酸修饰侧翼的同源核苷酸序列，其中侧翼同源核苷酸序列为待编辑的希望的核苷酸序列提供了充足同源性。

[0165] 本文的术语“植物优化的Cas内切核酸酶”是指由已经针对在植物细胞或植物中表达进行优化的核苷酸序列编码的Cas蛋白，包括多功能Cas蛋白。

[0166] “编码Cas内切核酸酶的植物优化的核苷酸序列”、“编码Cas内切核酸酶的植物优化的构建体”和“编码Cas内切核酸酶的植物优化的多核苷酸”在本文中可互换使用，并且是指编码Cas蛋白、或其变体或功能性片段的核苷酸序列，已经针对在植物细胞或植物中表达对其进行优化。

[0167] 术语“植物”一般包括整株植物、植物器官、植物组织、种子、植物细胞、种子和植物的后代。植物细胞包括但不限于得自下列物质的细胞：种子、悬浮培养物、胚、分生区域、愈伤组织、叶、根、芽、配子体、孢子体、花粉和小孢子。“植物元件”意在指整个植物或植物组分，可以包括分化和/或未分化的组织，例如但不限于植物组织、部分和细胞类型。在一个实施例中，植物元件是以下之一：整株植物、幼苗、分生组织、基本组织、维管组织、皮膜组织、种子、叶、根、芽、茎、花、果实、匍匐茎、鳞茎、块茎、球茎、无性末梢枝、芽、幼芽、肿瘤组织，以及细胞和培养物的各种形式(例如，单细胞、原生质体、胚胎和愈伤组织)。术语“植物器官”是指植物组织或构成植物的形态上和功能上不同部分的一组组织。如本文所用，“植物元

件”是植物的“部分”的同义词，是指植物的任何部分，并且可以包括不同的组织和/或器官，并且可以在全文中与术语“组织”互换使用。类似地，“植物繁殖元件”意在一般性地指能够通过该植物的有性或无性繁殖而创造其他植物的任何植物部分，例如但不限于：种子、幼苗、根、芽、切条、接穗、嫁接苗、匍匐茎、鳞茎、块茎、球茎、无性末梢枝或幼芽。植物元件可以存在于植物中或植物器官、组织培养物或细胞培养物中。

[0168] “后代”包括植物的任何后续世代。

[0169] 如本文使用，术语“植物部分”是指植物细胞、植物原生质体、可再生植物的植物细胞组织培养物、植物愈伤组织、植物块和在植物或植物部分(如胚、花粉、胚珠、种子、叶、花、枝、果、核、穗、穗轴、壳、茎、根、根尖、花药等)中完好的植物细胞，连同这些部分自身。籽粒意指由商业种植者出于栽培或繁殖物种之外的目的所生产的成熟种子。这些再生植物的后代、变体和突变体也包括在本发明的范围内，条件是这些部分包含经引入的多核苷酸。

[0170] 术语“单子叶植物的”或“单子叶植物”是指被子植物的亚类，也称为“单子叶植物纲”，其种子典型地仅包含一个胚叶或子叶。该术语包括对整个植物、植物元件、植物器官(例如，叶、茎、根等)、种子、植物细胞及其后代的指代。

[0171] 术语“双子叶植物的”或“双子叶植物”是指被子植物的亚类，也称为“双子叶植物纲”，其种子典型地包含两个胚叶或子叶。该术语包括对整个植物、植物元件、植物器官(例如，叶、茎、根等)、种子、植物细胞及其后代的指代。

[0172] 如本文使用，“雄性不育植物”是不产生有活力的或在其他情况下能够受精的雄配子的植物。如本文使用，“雌性不育植物”是不产生有活力的或在其他情况下能够受精的雌配子的植物。应当认识到雄性不育植物和雌性不育植物可以分别是雌性可育的和雄性可育的。应当进一步认识到，雄性可育(但雌性不育)植物当与雌性可育植物杂交时可以产生有活力的后代，并且雌性可育(但雄性不育)植物当与雄性可育植物杂交时可以产生有活力的后代。

[0173] 本文中术语“非常规酵母”是指不是酵母属(例如，酿酒酵母)或裂殖酵母属酵母物种的任何酵母。(参见“Non-Conventional Yeasts in Genetics,Biochemistry and Biotechnology:Practical Protocols[遗传学、生物化学和生物技术中的非常规酵母菌：实践方案]”,K.Wolf,K.D.Breunig,G.Barth编辑,Springer-Verlag,Berlin,Germany[德国柏林施普林格出版社],2003)。

[0174] 在本公开的上下文中，术语“杂交的”或“杂交”(cross或crossing)是指经由授粉将配子融合从而产生后代(即，细胞、种子、或植物)。该术语涵盖有性杂交(一株植物被另一株植物授粉)和自交(自体授粉，即当花粉和胚珠(或小孢子和大孢子)是来自同一植物或基因相同的植物时)。

[0175] 术语“渗入”是指基因座的期望等位基因从一种遗传背景传递到另一种遗传背景的现象。例如，可以经由两个亲本植物之间的有性杂交将指定基因座处的所希望的等位基因的渗入传递给至少一个后代植物，其中至少一个亲本植物在其基因组内具有所希望的等位基因。可替代地，例如等位基因的传递可以通过两个供体基因组之间的重组而发生，例如在融合原生质体中，其中至少其中一个供体原生质体在其基因组中具有所希望的等位基因。所希望的等位基因可以是，例如转基因、修饰的(突变的或编辑的)天然等位基因、或标志物或QTL的选择的等位基因。

[0176] 术语“同系”是一个比较术语，指遗传上相同但处理方法不同的生物体。在一个实例中，可以将两个遗传上相同的玉蜀黍植物胚胎分成两个不同的组，一个组接受处理(如引入CRISPR-Cas效应子内切核酸酶)，而一个组作为对照不接受这种处理。因此，两组之间的任何表型差异都可能仅归因于该处理，而不是归因于该植物的内源基因组成的任何固有性。

[0177] “引入”旨在意指以这样一种方式将多核苷酸或多肽或多核苷酸-蛋白复合物提供于靶，如细胞或生物体中，以致于这一种或多种组分得以进入该生物体的细胞的内部或进入细胞自身。

[0178] “目的多核苷酸”包括编码改善作物的合意性的蛋白或多肽的任何核苷酸序列。目的多核苷酸：包括但不限于，编码对农艺学、除草剂-抗性、杀昆虫抗性、疾病抗性、线虫抗性、除草剂抗性、微生物抗性、真菌抗性、病毒抗性、能育性或不育性、籽粒特征、商业产品、表型标志物而言重要的或任何其他具有重要农艺学或商业意义的性状的多核苷酸。目的多核苷酸可以另外以有义或反义取向加以利用。此外，可以一起或“堆叠”利用多于一个目的多核苷酸以提供额外的益处。

[0179] “复杂性状基因”座包括具有彼此遗传连锁的多个转基因的基因组基因座。

[0180] 本文的组合物和方法可以为植物提供改善的“农艺性状”或“具有农艺学重要性的性状”或“具有农艺学意义的性状”，这些性状可以包括但不限于以下：与不包含衍生自本文方法和组合物的修饰的同系植物相比的抗病性、耐旱性、耐热性、耐寒性、耐盐性、金属耐性、除草剂耐性、改善的水分利用效率、改善的氮利用率、改善的固氮作用、有害生物抗性、食草动物抗性、病原抗性、产量改善、健康增强、活力改善、生长改善、光合能力改善、营养增强、改变的蛋白含量、改变的油含量、生物量增加、芽长度增加、根长度增加、根结构改善、代谢产物的调节、蛋白质组的调节、种子重量的增加、改变的种子碳水化合物组成、改变的种子油组成、改变的种子蛋白组成、改变的种子营养成分。

[0181] “农艺性状潜力”意在指植物元件在其生命周期中的某个时刻表现出一种表型(优选地为一种改善的农艺性状)的能力，或将所述表型传递至在同一种植物中与其关联的另一种植物元件的能力。

[0182] 如本文所用，术语“减少”、“较少”、“较慢”和“增加”、“较快”、“增强”、“更大”是指与未修饰的植物元件或产生的植物相比，经修饰的植物元件或产生的植物的特征降低或增加。例如，特征的降低可以是低于未处理的对照至少1％、至少2％、至少3％、至少4％、至少5％、5％至10％、至少10％、10％至20％、至少15％、至少20％、20％至30％、至少25％、至少30％、30％至40％、至少35％、至少40％、40％至50％、至少45％、至少50％、50％至60％、至少60％、60％至70％、70％至80％、至少75％、至少80％、80％至90％、至少90％、90％至100％、至少100％、100％和200％、至少200％、至少300％、至少400％或更多，增加可以是高于未处理的对照至少1％、至少2％、至少3％、至少4％、至少5％、5％至10％、至少10％、10％至20％、至少15％、至少20％、20％至30％、至少25％、至少30％、30％至40％、至少35％、至少40％、40％至50％、至少45％、至少50％、50％至60％、至少60％、60％至70％、70％至80％、至少75％、至少80％、80％至90％、至少90％、90％至100％、至少100％、100％和200％、至少200％、至少300％、至少400％或更多。

[0183] 如本文所用，当提到序列位置时，术语“之前”是指一个序列在另一序列上游或5′

处出现。

[0184] 缩写的含义如下:“sec”意指秒、“min”意指分钟、“h”意指小时、“d”意指天、“uL”意指微升、“mL”意指毫升、“L”意指升、“uM”意指微摩尔、“mM”意指毫摩尔、“M”意指摩尔、“mmol”意指毫摩尔、“umole”或“umole”微摩尔、“g”意指克、“ug”或“ug”意指微克、“ng”意指纳克、“U”意指单位、“bp”意指碱基对、以及“kb”意指千碱基。

[0185] CRISPR-Cas系统的分类

[0186] CRISPR-Cas系统已根据组分的序列和结构分析进行了分类。已经描述了多种CRISPR/Cas系统,包括具有多亚基效应子复合物的1类系统(包括I型、III型和IV型),以及具有单蛋白效应子的2类系统(包括II型、V型和VI型)(Makarova等人,2015,Nature Reviews Microbiology[自然微生物学综述]卷13:1-15;Zetsche等人,2015,Cell[细胞]163,1-13;Shmakov等人,2015,Molecular Cell[分子细胞学]60,1-13;Haft等人,2005,Computational Biology,PLoS Comput Biol[美国科学公共图书馆计算生物学]1(6):e60;以及Koonin等人,2017,Curr Opinion Microbiology[微生物学新见]37:67-78)。

[0187] CRISPR-Cas系统至少包含一种CRISPR RNA(crRNA)分子和至少一种与CRISPR相关的(Cas)蛋白,以形成crRNA核糖核蛋白(crRNP)效应复合物。CRISPR-Cas基因座包含一系列相同的重复序列,这些重复序列散布有编码crRNA组分的DNA靶向间隔子以及编码Cas蛋白组分的cas基因的操纵子样单元。产生的核糖核蛋白复合物称为级联,它以序列特异性方式识别多核苷酸(Jore等人,Nature Structural&Molecular Biology[自然结构与分子生物学]18,529-536(2011))。该crRNA通过与互补DNA链形成碱基对,同时置换非互补链形成所谓的R环,从而充当效应子(蛋白或复合物)与双链DNA序列进行序列特异性结合的指导RNA。(Jore等人,2011.Nature Structural&Molecular Biology[自然结构与分子生物学]18,529-536)。

[0188] Cas内切核酸酶由单个CRISPR RNA(crRNA)指导,通过直接RNA-DNA碱基配对来识别紧邻前间隔子邻近基序(PAM)的DNA靶位点(Jore,M.M.等人,2011,Nat.Struct.Mol.Biol.[自然结构分子生物学]18:529-536,Westra,E.R.等人,2012,Molecular Cell[分子细胞学]46:595-605,以及Sinkunas,T.等人,2013,EMBO J.[欧洲分子生物学学会杂志]32:385-394)。1类CRISPR-Cas系统包括I型、III型和IV型。I类系统的特征是存在效应内切核酸酶复合物而不是单个蛋白。2类CRISPR-Cas系统包括II型、V型和VI型。2类系统的特征是存在单个Cas蛋白,而不是效应子模块内切核酸酶复合物。II型和V型Cas蛋白包含采用RNA酶H折叠的RuvC样内切核酸酶结构域。

[0189] 2类II型CRISPR/Cas系统采用crRNA和tracrRNA(反式激活CRISPR RNA)将Cas内切核酸酶指导到其DNA靶上。该crRNA包含与双链DNA靶的一条链互补的间隔子区域和与tracrRNA(反式激活CRISPR RNA)碱基配对的区域,该tracrRNA形成引导Cas内切核酸酶切割DNA靶的RNA双链体。对于酿脓链球菌Cas9内切核酸酶,该切割留下平末端。II型CRISR-Cas基因座可以编码tracrRNA,该tracrRNA与重复序列在对应的CRISPR阵列内部分互补,并且可以包含其他蛋白。

[0190] Cas内切核酸酶CRISPR-Cas系统组分

[0191] Cas内切核酸酶和效应子

[0192] 内切核酸酶是在多核苷酸链内切割磷酸二酯键的酶。内切核酸酶的实例包括限制

性内切核酸酶，大范围核酸酶，TAL效应子核酸酶(TALEN)，锌指核酸酶和Cas(CRISPR-associated)效应子内切核酸酶。

[0193] Cas内切核酸酶(作为单一效应子蛋白或与其他组分的效应子复合物)在靶序列处解开DNA双链体并任选地切割至少一条DNA链，如通过由与Cas内切核酸酶复合的多核苷酸(例如但不限于crRNA或指导RNA)识别靶序列所介导的。如果正确的前间隔子邻近基序(PAM)位于或相邻于DNA靶序列的3′末端，则通过Cas内切核酸酶对靶序列进行的此类识别和切割典型地会发生。可替代地，本文中的Cas内切核酸酶可能缺乏DNA切割或切口活性，但是当与合适的RNA组分复合时，仍然可以特异性结合DNA靶序列。(还参见2015年3月19日公开的美国专利申请US 20150082478和2015年2月26日公开的US 20150059010)。

[0194] 已描述的Cas内切核酸酶包括但不限于，例如：Cas3(1类I型系统的特征)、Cas9(2类II型系统的特征)和Cas12(Cpf1)(2类V型系统的特征)。

[0195] Cas9(以前称为Cas5、Csn1、或Csx12)是与cr核苷酸和tracr核苷酸或与单指导多核苷酸形成复合物的Cas内切核酸酶，其用于特异性识别和切割DNA靶序列的全部或部分。规范的Cas9识别靶dsDNA上的3’GC富集PAM序列(典型地包含NGG基序)。本文所述的Cas9直系同源物可以识别另外的PAM序列，并用于以不同的识别序列特异性来修饰靶位点。

[0196] Cas9蛋白包含RuvC核酸酶，以及与RuvC-II结果域相邻的HNH(H-N-H)核酸酶。RuvC核酸酶和HNH核酸酶各自可以在靶序列处切割单个DNA链(两个结构域的协同作用导致DNA双链切割，而一个结构域的活性导致切口)。通常，RuvC结构域包含亚结构域I、II和III，其中亚结构域I位于Cas9的N末端附近，并且亚结构域II和III位于蛋白的中间，即位于HNH结构域的侧翼(Hsu等人，2013，Cell[细胞]157：1262-1278)。Cas9内切核酸酶通常来源于II型CRISPR系统，该系统包括利用与至少一种多核苷酸组分复合的Cas9内切核酸酶的DNA切割系统。例如，Cas9可以与CRISPR RNA(crRNA)和反式激活CRISPR RNA(tracrRNA)复合。在另一个实例中，Cas9可以与单指导RNA复合(Makarova等人，2015，Nature Reviews Microbiology[自然综述微生物学]第13卷：1-15)。

[0197] Cas内切核酸酶和效应子蛋白可以用于靶向的基因组编辑(经由单个和多个双链断裂和缺口)和靶向的基因组调节(经由将表观遗传效应子结构域系链到Cas蛋白或sgRNA)。Cas内切核酸酶还可以被工程化作为RNA指导的重组酶起作用，并且经由RNA系链可以充当用于组装多蛋白和核酸复合物的支架(Mali等人，2013Nature Methods[自然方法]第10卷：957-963)。

[0198] 本文所述的Cas9直系同源物进一步包含内切核酸酶活性。

[0199] Cas9直系同源蛋白进一步被定义为天然Cas9直系同源蛋白的功能性片段或功能性变体，或与SEQ ID NO：86-170和511-1135中任何一个的至少50、50至100、至少100、100至150、至少150、150至200、至少200、200至250、至少250、250至300、至少300、300至350、至少350、350至400、至少400、400至450、至少500、500至550、至少550、550至600、至少600、600至650、至少650、650至700、至少700、700至750、至少750、750至800、至少800、800至850、至少850、850至900、至少900、900至950、至少950、950至1000、至少1000或甚至大于1000个连续氨基酸具有至少50％、50％至55％、至少55％、55％至60％、至少60％、60％至65％、至少65％、65％至70％、至少70％、70％至75％、至少75％、75％至80％、至少80％、80％至85％、至少85％、85％至90％、至少90％、90％至95％、至少95％、95％至96％、至少96％、96％至

97％、至少97％、97％至98％、至少98％、98％至99％、至少99％、99％至100％或100％序列同一性的蛋白，并且保留SEQ ID NO:86-170和511-1135中任何一个的天然全长Cas9直系同源蛋白的至少部分活性。

[0200] 在一些方面，Cas9直系同源物可以包含选自下组的多肽，该组由以下组成：与SEQ ID NO:86-171和511-1135中任何一个的至少50、50至100、至少100、100至150、至少150、150至200、至少200、200至250、至少250、250至300、至少300、300至350、至少350、350至400、至少400、400至450、至少500、500至550、至少550、550至600、至少600、600至650、至少650、650至700、至少700、700至750、至少750、750至800、至少800、800至850、至少850、850至900、至少900、900至950、至少950、950至1000、至少1000或甚至大于1000个连续氨基酸具有至少80％、80％至85％、至少85％、85％至90％、至少90％、90％至95％、至少95％、至少96％、至少97％、至少98％、至少99％、至少99.5％或大于99.5％同一性的多肽；SEQ ID NO:86-171和511-1135中任何一个的功能性变体；SEQ ID NO:86-171和511-1135中任何一个的功能性片段；Cas内切核酸酶，其由选自由SEQ ID NO:1-85组成组的多核苷酸编码；Cas内切核酸酶，其识别表4-83中任何一个列出的PAM序列；Cas内切核酸酶，其识别选自下组的PAM序列，该组由以下组成：NAR(G＞A)WH(A＞T＞C)GN(C＞T＞R)、N(C＞D)V(A＞S)R(G＞A)TTTN(T＞V)、NV(A＞G＞C)TTTTT、NATTTTT、NN(H＞G)AAAN(G＞A＞Y)N、N(T＞V)NAAATN、NAV(A＞G＞C)TCNN、NN(A＞S＞T)NN(W＞G＞C)CCN(Y＞R)、NNAH(T＞M)ACN、NGTGANN、NARN(A＞K＞C)ATN、NV(G＞A＞C)RNTTN、NN(A＞B)RN(A＞G＞T＞C)CCN、NN(A＞B)NN(T＞V)CCH(A＞Y)、NNN(H＞G)NCDAA、NN(H＞G)D(A＞K)GGDN(A＞B)、NNNNCCAG、NNNNCTAA、NNNNCVGANN、N (C＞D)NNTCCN、NNNNCTA、NNNNCYAA、NAGRGNY、NNGH(W＞C)AAA、NNGAAAN、NNAAAAA、NTGAR(G＞A)N(A＞Y＞G)N(Y＞R)、N(C＞D)H(C＞W)GH(Y＞A)N(A＞B)AN(A＞T＞S)、NNAAACN、NNGTAM(A＞C)Y、NH(A＞Y)ARNN(C＞W＞G)N、B(C＞K)GGN(A＞Y＞G)N NN、N(T＞C＞R)AGAN(A＞K＞C)NN、NGGN(A＞T＞G＞C)NNN、NGGD(A＞T＞G)TNN、NGGAN(T＞A＞C＞G)NN、CGGWN(T＞R＞C)NN、NGGWGNN、N(B＞A)GGNN(T＞V)NN、NNGD(A＞T＞G)AY(T＞C)N、N(T＞V)H(T＞C＞A)AAAAN、NRTAANN、N(H＞G)CAAH(Y＞A)N(Y＞R)N、NATAAN(A＞T＞S)N、NV(A＞G＞C)R(A＞G)ACCN、CN(C＞W＞G)AV(A＞S)GAC、NNRNCAC、N(A＞B)GGD(W＞G)D(G＞W)NN、BGD(G＞W)GTCN(A＞K＞C)、NAANACN、NRTHAN(A＞B)N、BHN(H＞G)NGN(T＞M)H(Y＞A)、NMRN(A＞Y＞G)AH(C＞T＞A)N、NNNCACN、NARN(T＞A＞S)ACN、NNNNATW、NGCNGCN、NNNCATN、NAGNGCN、NARN(T＞M＞G)CCN、NATCCTN、NRTAAN(T＞A＞S)N、N(C＞T＞G＞A)AAD (A＞G＞T)CNN、NAAAGNN、NNGACNN、N(T＞V)NTAAD(A＞T＞G)N、NNGAD(G＞W)NN、NGGN(W＞S)NNN、N(T＞V)GGD(W＞G)GNN、NGGD(A＞T＞G)N(T＞M＞G)NN、NNAAAGN、N(G＞H)GGDN(T＞M＞G)NN、NNAGAAA、NN(T＞M＞G)AAAAA、N(C＞D)N(C＞W＞G)GW(T＞C)D(A＞G＞T)AA、NAAAAYN、NRGNNNN、NATGN(H＞G)TN、NNDATTT和NATARCN(C＞T＞A＞G)；Cas内切核酸酶，其能够识别长度为一、二、三、四、五、六、七、八、九或十个核苷酸的PAM序列；Cas内切核酸酶，其包含与SEQ ID NO:1136-1730中的任何一个具有至少80％、80％至85％、至少85％、85％至90％、至少90％、90％至95％、至少95％、至少96％、至少97％、至少98％、至少99％、至少99.5％、或大于99.5％同一性的结构域；Cas内切核酸酶，其具有以下活性得分(根据与实例9的方法相同或相似的方法)或表86A的氨基酸表的位置得分的总和：至少1.0、1.0至2.0、至少2.0、2.0至3.0、至少3.0、3.0至4.0、至少4.0、4.0至5.0、至少5.0、5.0至6.0、至少6.0、6.0至7.0、至少7.0、7.0至8.0、至少8.0、8.0至9.0、

至少9.0、9.0至10.0、至少10.0或甚至大于10.0;Cas内切核酸酶,其包含与SEQ ID NO:1125的相对序列位置编号的比对相比,表86B中鉴定的一、二、三、四、五、六、七、八、九、十、十一、十二、十三、十四、十五、十六、十七、十八、十九、二十、二十一、二十二、二十三、二十四、二十五或二十六个特征氨基酸;以及Cas内切核酸酶,所述Cas内切核酸酶能够与包含SEQ ID NO:426-510、341-425、141-255或256-340中任一个的指导物多核苷酸形成复合物。在一些方面,Cas9多核苷酸具有多个先前列出的特征。

[0201] 本文公开的Cas9直系同源物或cas9直系同源物可进一步包含异源组分。在一些方面,所述异源组分选自由以下组成的组:异源多核苷酸、异源多肽、粒子、固体基质和组氨酸标签。在一些方面,所述异源多核苷酸是指导多核苷酸,或编码与其可操作地连接的标志物或纯化标签、或异源非编码调节元件的多核苷酸。

[0202] 在一些方面,编码Cas9内切核酸酶直系同源物的多核苷酸包含在重组载体内,所述重组载体可进一步包含另外的组分,例如但不限于异源启动子或其他非编码调节元件。

[0203] 用于本公开方法的Cas9直系同源内切核酸酶、效应子蛋白或其功能性片段可从天然来源或重组来源中分离,在重组来源中,遗传修饰的宿主细胞被修饰以表达编码所述蛋白的核酸序列。可替代地,Cas9直系同源蛋白可以是使用无细胞蛋白表达系统产生的,或是合成产生的。Cas内切核酸酶可以被分离并引入异源细胞,或者可以从其天然形式进行修饰,以表现出与其天然来源不同的活性类型或大小。此类修饰包括但不限于:片段、变体、取代、缺失和插入。

[0204] Cas9直系同源物的片段和变体可以通过如定点诱变和合成构建等方法来获得。测量内切核酸酶活性的方法是本领域众所周知的,例如但不限于,2013年11月7日公开的WO 2013166113、2016年11月24日公开的WO 2016186953和2016年11月24日公开的WO 2016186946。

[0205] Cas9直系同源物可以包含Cas多肽的经修饰的形式。Cas多肽的经修饰的形式可包括降低Cas蛋白的天然存在的核酸酶活性的氨基酸改变(例如,缺失、插入或取代)。例如,在一些情况下,该Cas蛋白的修饰形式具有低于50%、低于40%、低于30%、低于20%、低于10%、低于5%、或低于1%的相应的野生型Cas多肽(2014年3月6日公开的US 20140068797)的核酸酶活性。在某些情况下,Cas多肽的修饰形式没有实质的核酸酶活性,被称为催化"失活的Cas"或"失活的Cas(dCas)"。失活的Cas/失活的Cas包括失活Cas内切核酸酶(dCas)。可以将无催化失活的Cas内切核酸酶与异源序列融合,以诱导或修饰活性。

[0206] Cas9直系同源物可以是包含一个或多个异源蛋白结构域(例如除Cas蛋白之外的1、2、3或更多个结构域)的融合蛋白的一部分。这样的融合蛋白可以包含任何另外的蛋白序列,以及任选地在任何两个结构域之间(例如在Cas和第一异源结构域之间)的连接体序列。可以与本文中的Cas蛋白融合的蛋白结构域的实例包括但不限于表位标签(例如,组氨酸[His]、V5、FLAG、流感血球凝集素[HA]、myc、VSV-G、硫氧还蛋白[Trx]);报告子(例如谷胱甘肽-5-转移酶[GST]、辣根过氧化物酶[HRP]、氯霉素乙酰转移酶[CAT]、β-半乳糖苷酶、β-葡萄糖醛酸酶[GUS]、荧光素酶、绿色荧光蛋白[GFP]、HcRed、DsRed、青色荧光蛋白[CFP]、黄色荧光蛋白[YFP]、蓝色荧光蛋白[BFP]);以及具有一个或多个以下活性的结构域:甲基化酶活性、脱甲基酶活性、转录激活活性(例如,VP16或VP64)、转录抑制活性、转录释放因子活性、组蛋白修饰活性、RNA切割活性和核酸结合活性。Cas9直系同源物还可以与结合DNA分子

或其他分子的蛋白融合，例如麦芽糖结合蛋白(MBP)、S-标签、Lex A DNA结合结构域(DBD)、GAL4A DNA结合结构域和单纯疱疹病毒(HSV)VP16。

[0207] 可以将催化活性和/或失活的Cas9直系同源物融合至异源序列(2014年3月6日公开的US 20140068797)。适合的融合配偶体包括，但不限于提供活性的多肽，该活性通过直接作用于靶DNA上或与该靶DNA相关的多肽(例如，组蛋白或其他DNA-结合蛋白)上间接地增加转录。另外的适合的融合配偶体包括，但不限于提供甲基转移酶活性、脱甲基酶活性、乙酰基转移酶活性、脱乙酰基酶活性、激酶活性、磷酸酶活性、泛素连接酶活性、去泛素化酶活性、腺苷酸化活性、去腺苷酸化活性、苏素化活性、去苏素化活性、核糖基化活性、去核糖基化活性、豆蔻酰化活性，或去豆蔻酰化活性的多肽。此外适合的融合配偶体包括，但不限于直接提供靶核酸的增加的转录的多肽(例如，募集转录激活因子、小分子/药物-应答性转录调节因子等的转录激活因子或其片段，蛋白或其片段)。还可以将催化失活的Cas融合到FokI核酸酶从而产生双链断裂(Guilinger等人Nature Biotechnology[自然生物技术]，第32卷，第6期，2014年6月)。在一些方面，Cas9直系同源物是融合蛋白，其进一步包含核酸酶结构域、转录激活子结构域、转录阻遏子结构域、表观遗传修饰结构域、切割结构域、核定位信号、细胞穿透结构域、易位结构域、标志物、或与靶多核苷酸序列或从其获得或衍生出所述靶多核苷酸序列的细胞异源的转基因。在一些方面，核酸酶融合蛋白包含Clo51或Fok1。

[0208] 本文所述的Cas9直系同源物可以通过本领域已知的方法表达和纯化，例如如2016年11月24日公开的WO/2016/186953中所述。

[0209] Cas内切核酸酶可包含异源核定位序列(NLS)。例如，本文中的异源NLS氨基酸序列可能具有足够的强度来驱动在本文的酵母细胞细胞核中可检测的量的Cas蛋白的积累。NLS可以包含碱性、带正电荷的残基(例如赖氨酸和/或精氨酸)的一个(单分型)或多个(例如，二分型)短序列(例如，2至20个残基)，并且可以位于Cas氨基酸序列中的任何地方，但使得其暴露于蛋白表面上。例如，NLS可以可操作地连接到本文中的Cas蛋白的N-末端或C-末端。两个或更多个NLS序列可以连接到Cas蛋白，例如在Cas蛋白的N-末端和C-末端两者。Cas内切核酸酶基因可以可操作地连接至Cas密码子区域上游的SV40核靶向信号和Cas密码子区域下游的二分型VirD2核定位信号(Tinland等人，(1992)Proc.Natl.Acad.Sci.USA[美国国家科学院院刊]89:7442-6)。本文中适合的NLS序列的非限制性实例包括在美国专利号6,660,830和7,309,576中公开的那些。

[0210] 可以通过本领域已知的任何方法从天然的或亲本的Cas内切核酸酶分子产生人工(非天然存在的)Cas内切核酸酶。在一些方面，这是通过诱变编码内切核酸酶蛋白的基因来实现的。在一些方面，诱变是通过选自下组的方法实现，该组由以下组成：使用作用于内切核酸酶基因的双链断裂诱导剂；辐射诱变；化学诱变；编码内切核酸酶的基因中至少一个多核苷酸的添加、缺失、取代、插入或改变；或氨基酸的一个或多个密码子的取代。在一些方面，可以采用内切核酸酶分子的定向进化来优化Cas内切核酸酶的表达或活性，并且可以通过本领域已知的随机或非随机蛋白改组方法来实现。

[0211] 前间隔子邻近基序(PAM)

[0212] 本文中的“前间隔子邻近基序”(PAM)是指与由指导多核苷酸/Cas内切核酸酶系统可以识别的(靶向的)靶序列(前间隔子)相邻的短核苷酸序列。在一些方面，如果靶DNA序列与PAM序列不相邻或不邻近，则Cas内切核酸酶可能无法成功识别该靶DNA序列。在一些方

面，该PAM在靶序列(例如，Cas12a)之前。在一些方面，该PAM在靶序列(例如，酿脓链球菌Cas9)之后。本文中的PAM的序列和长度可以取决于所使用的Cas蛋白或Cas蛋白复合物而不同。所述PAM序列可以是任何长度，但典型地是1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个核苷酸长度。

[0213] “随机的PAM”和“随机的前间隔子邻近基序”在本文中可互换地使用，并且意指邻近由指导多核苷酸/Cas内切核酸酶系统识别(靶向)的靶序列(前间隔子)的随机DNA序列。随机的PAM序列可以是任何长度，但典型地是1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个核苷酸长度。随机的核苷酸包括核苷酸A、C、G或T中的任一者。

[0214] 迄今为止，已经描述了可以识别特定PAM序列(2016年11月24日公开的WO 2016186953、2016年11月24日公开的WO 2016186946和Zetsche B等人2015.Cell[细胞]163，1013)并在特定位置切割靶DNA的许多Cas内切核酸酶。应当理解的是，基于本文所述的使用新颖的受指导的Cas系统的方法和实施例，现在本领域技术人员可以定制这些方法，使得它们可以利用任何受指导的内切核酸酶系统。

[0215] 表4-50中描述了与本发明的一些Cas9直系同源物相对应的PAM序列。

[0216] 指导多核苷酸

[0217] 指导多核苷酸使得Cas内切核酸酶能够进行靶识别、结合和任选地切割，并且可以是单分子或双分子。指导多核苷酸序列可以是RNA序列、DNA序列或其组合(RNA-DNA组合序列)。任选地，指导多核苷酸可以包含至少一种核苷酸、磷酸二酯键或连接修饰，例如但不限于锁核酸(LNA)、5-甲基dC、2，6-二氨基嘌呤、2′-氟代A、2′-氟代U、2′-O-甲基RNA、硫代磷酸酯键、与胆固醇分子的连接、与聚乙二醇分子的连接、与间隔子18(六乙二醇链)分子的连接、或导致环化的5′至3′共价连接。仅包含核糖核酸的指导多核苷酸也称为“指导RNA”或“gRNA”(2015年3月19日公开的US 20150082478和2015年2月26日公开的US 20150059010)。指导多核苷酸可以被工程改造或合成。

[0218] 指导多核苷酸包括嵌合的非天然存在的指导RNA，所述指导RNA包含在自然界中未一起发现的区域(即，它们彼此是异源的)。例如，嵌合的非天然存在的指导RNA包含可与靶DNA中的核苷酸序列杂交的第一核苷酸序列结构域(称为可变靶向结构域或VT结构域)，所述第一核苷酸序列结构域与可识别Cas内切核酸酶的第二核苷酸序列连接，使得所述第一和第二核苷酸序列在自然界中未被发现连接在一起。

[0219] 指导多核苷酸可以是包含cr核苷酸序列和tracr核苷酸序列的双分子(也称为双链体指导多核苷酸)。cr核苷酸包括可以与靶DNA中的核苷酸序列杂交的第一核苷酸序列区域(称为可变靶向结构域或VT结构域)和作为Cas内切核酸酶识别(CER)域的一部分的第二核苷酸序列(也称为tracr配对序列)。tracr配对序列可以沿互补区域与tracr核苷酸杂交，并一起形成Cas内切核酸酶识别结构域或CER结构域。CER结构域能够与Cas内切核酸酶多肽相互作用。双链体指导多核苷酸的cr核苷酸和tracr核苷酸可以是RNA、DNA和/或RNA-DNA组合序列。

[0220] 在一些实施例中，双链体指导多核苷酸的cr核苷酸分子被称为“crDNA”(当由DNA核苷酸的连续延伸构成时)或“crRNA”(当由RNA核苷酸的连续延伸构成时)或“crDNA-RNA”(当由DNA和RNA核苷酸的组合构成时)。cr核苷酸可以包含在细菌和古细菌中天然存在的crRNA的片段。可以存在于本文披露的cr核苷酸中的、细菌和古细菌中天然存在的crRNA片

段的大小可以是但不限于2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20或更多个核苷酸。

[0221] 在5′-至-3′方向上，tracrRNA(反式激活CRISPR RNA)包含(i)与CRISPR II型crRNA的重复区退火的“反重复”序列和(ii)含茎环的部分(Deltcheva等人，Nature[自然]471:602-607)。双链体指导多核苷酸可以与Cas内切核酸酶形成复合物，其中所述指导多核苷酸/Cas内切核酸酶复合物(还称为指导多核苷酸/Cas内切核酸酶系统)可以将Cas内切核酸酶引导至基因组靶位点，使所述Cas内切核酸酶能够识别、结合靶位点、并任选地使靶位点产生切口或切割(引入单链或双链断裂)靶位点。(2015年3月19日公开的US 20150082478和2015年2月26日公开的US 20150059010)。在一些实施例中，tracr核苷酸被称为“tracrRNA”(当由RNA核苷酸的连续延伸构成时)或“tracrDNA”(当由DNA核苷酸的连续延伸构成时)或“tracrDNA-RNA”(当由DNA和RNA核苷酸的组合构成时)。

[0222] 在一个实施例中，指导RNA/Cas内切核酸酶复合物的RNA是包含双链体crRNA-tracrRNA的双链体化的RNA。

[0223] 在一个方面，所述指导多核苷酸是能够形成文所述的PGEN的指导多核苷酸，其中所述指导多核苷酸包含与靶DNA中的核苷酸序列互补的第一核苷酸序列结构域和与所述Cas内切核酸酶多肽相互作用的第二核苷酸序列结构域。

[0224] 在一个方面，所述指导多核苷酸是本文所述的针对多核苷酸，其中所述第一核苷酸序列和所述第二核苷酸序列结构域选自由以下组成的组:DNA序列、RNA序列及其组合。

[0225] 在一个方面，所述指导多核苷酸是本文所述的指导多核苷酸，其中所述第一核苷酸序列和所述第二核苷酸序列结构域选自由以下组成的组:增强稳定性的RNA主链修饰，增强稳定性的DNA主链修饰及其组合(参见Kanasty等人，2013，Common RNA-backbone modifications[常见RNA主链修饰]，Nature Materials[自然材料]12:976-977;2015年3月19日公开的US 20150082478和2015年2月26日公开的US 20150059010)

[0226] 所述指导RNA包括双分子，所述双分子包含与至少一个tracrRNA连接的嵌合的非天然存在的crRNA。嵌合的非天然存在的crRNA包括包含在自然界中不一起发现的区域(即，它们彼此异源)的crRNA。例如，crRNA包含可与靶DNA中的核苷酸序列杂交的第一核苷酸序列结构域(称为可变靶向结构域或VT结构域)，所述第一核苷酸序列结构域与第二核苷酸序列(也称为tracr配对序列)连接，使得所述第一和第二序列在自然界中未被发现连接在一起。

[0227] 指导多核苷酸也可以是包含连接至tracr核苷酸序列的cr核苷酸序列的单分子(也称为单指导多核苷酸)。单指导多核苷酸包含可以与靶DNA中的核苷酸序列杂交的第一核苷酸序列结构域(称为可变靶向(Variable Targeting)结构域或VT结构域)和与Cas内切核酸酶多肽相互作用的Cas内切核酸酶识别(Cas endonuclease recognition)结构域(CER结构域)。

[0228] 术语“可变靶向结构域”或“VT结构域”在本文中可互换使用，并且包括可以与双链DNA靶位点的一条链(核苷酸序列)杂交(互补)的核苷酸序列。第一核苷酸序列结构域(VT结构域)与靶序列之间的互补％可以为至少50％、51％、52％、53％、54％、55％、56％、57％、58％、59％、60％、61％、62％、63％、63％、65％、66％、67％、68％、69％、70％、71％、72％、73％、74％、75％、76％、77％、78％、79％、80％、81％、82％、83％、84％、85％、86％、87％、

88％、89％、90％、91％、92％、93％、94％、95％、96％、97％、98％、99％或100％。可变靶向结构域可以是至少12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29或30个核苷酸长度。

[0229] 单指导多核苷酸的VT结构域和/或CER结构域可以包含RNA序列、DNA序列或RNA-DNA组合序列。由来自cr核苷酸和tracr核苷酸的序列构成的单指导多核苷酸可以被称为“单指导RNA”(当由RNA核苷酸的连续延伸构成时)或“单指导DNA”(当由DNA核苷酸的连续延伸构成时)或“单指导RNA-DNA”(当由RNA和DNA核苷酸的组合构成时)。单指导多核苷酸可以与Cas内切核酸酶形成复合物,其中所述指导多核苷酸/Cas内切核酸酶复合物(还称为指导多核苷酸/Cas内切核酸酶系统)可以将Cas内切核酸酶引导至基因组靶位点,使所述Cas内切核酸酶能够识别、结合靶位点、并任选地使靶位点产生切口或切割(引入单链或双链断裂)靶位点。(2015年3月19日公开的US 20150082478和2015年2月26日公开的US 20150059010)。

[0230] 嵌合的非天然存在的单指导RNA(sgRNA)包括包含在自然界中不一起发现的区域(即,它们彼此异源)的sgRNA。例如,sgRNA包含可与靶DNA中的核苷酸序列杂交的第一核苷酸序列结构域(称为可变靶向结构域或VT结构域),所述第一核苷酸序列结构域与在自然界中未被发现连接在一起的第二核苷酸序列(也称为tracr配对序列)连接。

[0231] 连接单指导多核苷酸的cr核苷酸和tracr核苷酸的核苷酸序列可以包含RNA序列、DNA序列或RNA-DNA组合序列。在一个实施例中,连接单指导多核苷酸的cr核苷酸和tracr核苷酸的核苷酸序列(也称为“环”)可以是至少3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99或100个核苷酸的长度。在另一个实施例中,连接单指导多核苷酸的cr核苷酸和tracr核苷酸的核苷酸序列可以包括四环序列,如但不限于GAAA四环序列。

[0232] 指导多核苷酸可以通过本领域已知的任何方法产生,包括化学合成指导多核苷酸(例如但不限于Hendel等人2015,Nature Biotechnology[自然生物技术]33,985-989)、体外产生的指导多核苷酸、和/或自剪接指导RNA(例如但不限于Xie等人2015,PNAS[美国国家科学院院刊]112:3570-3575)。

[0233] 在真核细胞中表达RNA组分(例如gRNA)用于进行Cas9介导的DNA靶向的方法已经使用RNA聚合酶III(Pol III)启动子,其允许具有精确定义的未修饰的5′-和3′-末端的RNA转录(DiCarlo等人,Nucleic Acids Res.[核酸研究]41:4336-4343;Ma等人,Mol.Ther.Nucleic Acids[分子治疗-核酸]3:e161)。此策略已经成功应用于若干不同物种(包括玉蜀黍和大豆)的细胞中(2015年3月19日公开的US 20150082478)。已经描述了用于表达并不具有5′-帽的RNA组分的方法(2016年2月18日公开的WO 2016/025131)。

[0234] 单指导RNA(sgRNA)分子可以包含VT结构域。

[0235] 单指导RNA(sgRNA)分子可包含crRNA重复序列。在一些方面,crRNA重复序列选自由以下组成的组:SEQ ID NO:171-255。

[0236] 单指导RNA(sgRNA)分子可以包含环。

[0237] 单指导RNA(sgRNA)分子可包含反重复序列。在一些方面,反重复序列选自由以下

组成的组:SEQ ID NO:256-340。

[0238] 单指导RNA(sgRNA)分子可以包含3'tracrRNA。在一些方面,3'tracrRNA选自由以下组成的组:SEQ ID NO:341-425。

[0239] 术语"单指导RNA"和"sgRNA"在本文中可互换使用,并涉及两个RNA分子的合成融合,其中包含可变靶向结构域(与tracrRNA杂交的tracr配对序列连接)的crRNA(CRISPR RNA)与tracrRNA(反式激活CRISPR RNA)融合。单指导RNA可以包含可与II型Cas9内切核酸酶形成复合物的II型CRISPR/Cas9系统的crRNA或crRNA片段和tracrRNA或tracrRNA片段,其中所述指导RNA/Cas9内切核酸酶复合物可以将Cas9内切核酸酶引导至DNA靶位点,使得Cas9内切核酸酶能够识别、结合DNA靶位点、并任选地使DNA靶位点产生切口或切割(引入单链或双链断裂)DNA靶位点。

[0240] 在一些方面,sgRNA选自由以下组成的组:SEQ ID NO:426-510。

[0241] 可以通过用可与任何期望的靶序列杂交的随机核苷酸改变本文所述的任何指导多核苷酸的可变靶向结构域(VT)来设计靶向生物体基因组中的靶位点的单指导RNA。

[0242] 在一些实施例中,主题核酸(例如,指导多核苷酸,包含编码指导多核苷酸的核苷酸序列的核酸;编码本公开的Cas9内切核酸酶的核酸;crRNA或编码crRNA的核苷酸,tracrRNA或编码tracrRNA的核苷酸,编码VT结构域的核苷酸,编码CER结构域的核苷酸等)包含提供另外的所需特征(例如,经修饰或调节的稳定性;亚细胞靶向性;追踪例如荧光标记物;蛋白或蛋白复合物的结合位点;等)的修饰或序列。指导多核苷酸、VT结构域和/或CER结构域的核苷酸序列修饰可以选自但不限于由以下各项组成的组:5′帽、3′聚腺苷酸尾、核糖开关序列、稳定性控制序列、形成dsRNA双链体的序列、将指导多核苷酸靶向亚细胞位置的修饰或序列、提供跟踪的修饰或序列、提供蛋白结合位点的修饰或序列、锁核酸(LNA)、5-甲基dC核苷酸、2,6-二氨基嘌呤核苷酸、2′-氟代A核苷酸、2′-氟代U核苷酸、2′-O-甲基RNA核苷酸、硫代磷酸酯键、与胆固醇分子的连接、与聚乙二醇分子的连接、与间隔子18分子的连接、5′至3′共价连接、或其任何组合。这些修饰可以产生至少一个另外的有益特征,其中该另外的有益特征选自由以下组成的组:修改的或调节的稳定性、亚细胞靶向、跟踪、荧光标记、用于蛋白或蛋白复合物的结合位点、对互补靶序列的修改的结合亲和力、修改的细胞降解抗性和增加的细胞通透性。

[0243] 本公开的指导多核苷酸的功能性变体可以包括修饰的指导多核苷酸,其中修饰包括:在单指导RNA中,添加、去除、或以其他方式改变环和/或发夹。

[0244] 本公开的指导多核苷酸的功能性变体可以包括修饰的指导多核苷酸,其中修饰包括:在核苷酸序列中的一个或多个经修饰的多核苷酸,其中所述一个或多个经修饰的多核苷酸包括至少一个非天然存在的核苷酸、核苷酸模拟物(如在2014年3月6日公开的美国申请US 2014/0068797中描述)、或其类似物,或其中所述一个或多个经修饰的核苷酸选自由以下组成的组:2′-O-甲基类似物、2′-氟类似物2-氨基嘌呤、5-溴-尿苷、假尿苷、和7-甲基鸟苷。

[0245] 在一个方面,指导RNA的功能性变体可以形成指导RNA/Cas9内切核酸酶复合物,所述复合物可以对靶序列进行识别、结合、并且任选地产生切口或进行切割。

[0246] 指导多核苷酸/Cas内切核酸酶复合物

[0247] 本文所述的指导多核苷酸/Cas内切核酸酶复合物能够识别、结合靶序列的全部或

部分并任选地使靶序列的全部或部分产生切口、解旋或切割靶序列的全部或部分。

[0248] 可以切割DNA靶序列的两条链的指导多核苷酸/Cas内切核酸酶复合物通常包含具有处于功能状态的所有其内切核酸酶结构域的Cas蛋白(例如野生型内切核酸酶结构域或其变体在每个内切核酸酶结构域中保留一些或全部活性)。因此,在Cas蛋白的每个内切核酸酶结构域中保留一些或全部活性的野生型Cas蛋白(例如,本文披露的Cas蛋白)或其变体是可以切割DNA靶序列的两条链的Cas内切核酸酶的合适实例。

[0249] 可以切割DNA靶序列的一条链的指导多核苷酸/Cas内切核酸酶酶复合物可以在本文中表征为具有切口酶活性(例如,部分切割能力)。Cas切口酶通常包含一个功能性内切核酸酶结构域,该结构域允许Cas仅切割DNA靶序列的一条链(即,形成切口)。例如,Cas切口酶可以包含(i)突变的、功能失调的RuvC结构域和(ii)功能性HNH结构域(例如野生型HNH结构域)。作为另一个实例,Cas切口酶可以包含(i)功能性RuvC结构域(例如野生型RuvC结构域)和(ii)突变的功能失调的HNH结构域。在2014年7月3日公开的US 20140189896中公开了适用于本文的Cas切口酶的非限制性实例。可以使用一对Cas切口酶来增加DNA靶向的特异性。一般来说,这可以通过提供两个Cas切口酶来进行,这两个Cas切口酶通过与具有不同引导序列的RNA组分缔合,在希望靶向的区域的相反链上在DNA序列附近进行靶向和切口。每个DNA链的这样的附近切割产生双链断裂(即,具有单链突出端的DSB),其然后被识别为非同源末端连接(NHEJ)(倾向于产生导致突变的不完美修复)或同源重组(HR)的底物。在这些实施例中的每个切口可以彼此隔开例如至少5、5至10、至少10、10至15、至少15、15至20、至少20、20至30、至少30、30至40、至少40、40至50、至少50、50至60、至少60、60至70、至少70、70至80、至少80、80至90、至少90、90至100或100或更多(或5至100的任何数字)个碱基。本文中的一种或两种Cas切口酶蛋白可以用于Cas切口酶对。例如,可以使用具有突变的RuvC结构域但具有功能性HNH结构域的Cas切口酶(即,Cas HNH+/RuvC-)(例如,酿脓链球菌Cas HNH+/RuvC-)。通过使用本文中的合适的RNA组分(具有将每个切口酶靶向每个特异性DNA位点的指导RNA序列),将每个Cas切口酶(例如,Cas HNH+/RuvC-)引导到彼此邻近(分离多达100个碱基对)的特定的DNA位点。

[0250] 在某些实施例中指导多核苷酸/Cas内切核酸酶复合物可以结合DNA靶位点序列,但不切割在靶位点序列处的任何链。这样的复合物可以包含其中所有核酸酶结构域都是突变的、功能失调的Cas蛋白。例如,可以结合到DNA靶位点序列但在靶位点序列处不切割任何链的Cas蛋白可以包含突变的、功能失调的RuvC结构域和突变的、功能失调的HNH结构域。结合但不切割靶DNA序列的本文中的Cas蛋白可以用于调节基因表达,例如,在该情况下,Cas蛋白可以与转录因子(或其部分)融合(例如阻遏子或激活子,例如本文披露的那些中的任一种)。

[0251] 在本公开的一个实施例中,指导多核苷酸/Cas内切核酸酶复合物是包含至少一种指导多核苷酸和至少一种Cas内切核酸酶多肽的指导多核苷酸/Cas内切核酸酶复合物(PGEN)。在一些方面,所述Cas内切核酸酶多肽包含另一Cas蛋白的至少一个蛋白亚基或其功能性片段,其中所述指导多核苷酸是嵌合的非天然存在的指导多核苷酸,其中所述指导多核苷酸/Cas内切核酸酶复合物能够识别、结合靶序列的全部或部分并任选地使靶序列的全部或部分产生切口、解旋或切割靶序列的全部或部分。

[0252] 在一些方面,PGEN是核糖核蛋白复合物(RNP),其中将Cas9直系同源物作为蛋白提

供,并且将指导多核苷酸作为核糖核苷酸提供。

[0253] Cas内切核酸酶蛋白可以是本文公开的Cas9直系同源物。

[0254] 在本公开的一个实施例中,指导多核苷酸/Cas效应子复合物是包含至少一种指导多核苷酸和Cas9直系同源内切核酸酶的指导多核苷酸/Cas内切核酸酶复合物(PGEN),其中所述指导多核苷酸/Cas内切核酸酶复合物能够识别、结合靶序列的全部或部分并任选地使靶序列的全部或部分产生切口、解旋或切割靶序列的全部或部分。

[0255] PGEN可以是指导多核苷酸/Cas内切核酸酶复合物,其中所述Cas内切核酸酶进一步包含另外的Cas蛋白的至少一个蛋白亚基或其功能性片段的一个拷贝或多个拷贝。

[0256] 一方面,本文所述的指导多核苷酸/Cas内切核酸酶复合物(PGEN)是PGEN,其中所述Cas内切核酸酶共价或非共价连接到至少一个Cas蛋白亚基或其功能性片段。PGEN可以是指导多核苷酸/Cas内切核酸酶复合物,其中所述Cas内切核酸酶多肽共价或非共价连接,或组装成Cas蛋白的至少一个蛋白亚基(选自由Cas1蛋白亚基、Cas2蛋白亚基、Cas4蛋白亚基及其任何组合组成的组)或其功能性片段的一个或多个拷贝,在一些方面中有效地形成了切割就绪的Cascade。PGEN可以是指导多核苷酸/Cas内切核酸酶复合物,其中所述Cas内切核酸酶共价或非共价连接或组装至选自由Cas1、Cas2和Cas4组成的组的Cas蛋白的至少两个不同的蛋白亚基。PGEN可以是指导多核苷酸/Cas内切核酸酶复合物,其中所述Cas内切核酸酶共价或非共价连接至选自由Cas1、Cas2和Cas4以及其任何组合组成的组的Cas蛋白的至少三个不同的蛋白亚基或其功能性片段。

[0257] 指导多核苷酸/Cas内切核酸酶复合物的任何组分、指导多核苷酸/Cas内切核酸酶复合物自身、连同一个或多个多核苷酸修饰模板和/或一个或多个DNA供体,可以通过本领域已知的任何方法,被引入到异源细胞或生物中。

[0258] 指导RNA/Cas9内切核酸酶系统的一些用途包括但不限于修饰或替代目的核苷酸序列(例如调节元件)、目的多核苷酸的插入、基因敲除、基因敲入、剪接位点的修饰和/或引入替代的剪接位点、编码目的蛋白的核苷酸序列的修饰、氨基酸和/或蛋白融合、以及通过将反向重复表达为目的基因的基因沉默。

[0259] 用于细胞转化的重组构建体

[0260] 可以将本文公开的公开的指导多核苷酸、Cas内切核酸酶、多核苷酸修饰模板、供体DNA、指导多核苷酸/Cas内切核酸酶系统以及其任意一种组合(任选地进一步包含一个或多个目的多核苷酸)引入细胞中。细胞包括但不限于人类、非人类、动物、细菌、真菌、昆虫、酵母、非常规酵母和植物细胞,以及通过本文所述的方法产生的植物和种子。

[0261] 本文使用的标准重组DNA和分子克隆技术是在本领域熟知的,并且更全面地描述于Sambrook等人,Molecular Cloning:A Laboratory Manual[分子克隆:实验室手册];Cold Spring Harbor Laboratory:Cold Spring Harbor,NY[冷泉港实验室:冷泉港,纽约州](1989)中。转化方法是本领域技术人员熟知的并且在下文中进行了描述。

[0262] 载体和构建体包括环状质粒和包含目的多核苷酸的线状多核苷酸,以及任选地包括接头、衔接子、用于调节或分析的其他组分。在一些实例中,识别位点和/或靶位点可以包含在内含子、编码序列、5′UTR、3′UTR、和/或调节区内。

[0263] 在原核和真核细胞中表达和利用新颖CRISPR-Cas系统的组分

[0264] 本发明还提供了用于在原核或真核细胞/生物体中表达指导RNA/Cas系统的表达

构建体，所述指导RNA/Cas系统能够识别、结合靶序列的全部或部分并任选地使靶序列的全部或部分产生切口、解旋或切割靶序列的全部或部分。

[0265] 在一个实施例中，本发明的表达构建体包含与编码Cas基因的核苷酸序列(或优化的序列，包括本文所述的Cas内切核酸酶基因)可操作地连接的启动子和与本发明的指导RNA可操作地连接的启动子。该启动子能够驱动在原核或真核细胞/生物中可操作地连接的核苷酸序列的表达。

[0266] 指导多核苷酸、VT结构域和/或CER结构域的核苷酸序列修饰可以选自但不限于由以下各项组成的组：5′帽、3′聚腺苷酸尾、核糖开关序列、稳定性控制序列、形成dsRNA双链体的序列、将指导多核苷酸靶向亚细胞位置的修饰或序列、提供跟踪的修饰或序列、提供蛋白结合位点的修饰或序列、锁核酸(LNA)、5-甲基dC核苷酸、2,6-二氨基嘌呤核苷酸、2′-氟代A核苷酸、2′-氟代U核苷酸、2′-O-甲基RNA核苷酸、硫代磷酸酯键、与胆固醇分子的连接、与聚乙二醇分子的连接、与间隔子18分子的连接、5′至3′共价连接、或其任何组合。这些修饰可以产生至少一个另外的有益特征，其中该另外的有益特征选自由以下组成的组：修改的或调节的稳定性、亚细胞靶向、跟踪、荧光标记、用于蛋白或蛋白复合物的结合位点、对互补靶序列的修改的结合亲和力、修改的细胞降解抗性和增加的细胞通透性。

[0267] 在真核细胞中表达RNA组分(例如gRNA)用于进行Cas9介导的DNA靶向的方法已经使用RNA聚合酶III(Pol III)启动子，其允许具有精确定义的未修饰的5′-和3′-末端的RNA转录(DiCarlo等人，Nucleic Acids Res.[核酸研究]41:4336-4343;Ma等人，Mol.Ther.Nucleic Acids[分子治疗-核酸]3:e161)。此策略已经成功应用于若干不同物种(包括玉蜀黍和大豆)的细胞中(2015年3月19日公开的US 20150082478)。已经描述了用于表达并不具有5′帽的RNA组分的方法(2016年2月18日公开的WO 2016/025131)。

[0268] 可以采用不同方法和组合物来获得细胞或生物体，所述细胞或生物体具有插入针对Cas内切核酸酶的靶位点中的目的多核苷酸。此类方法可以采用同源重组(HR)以提供目的多核苷酸在靶位点处的整合。在本文所述的一种方法中，经由供体DNA构建体，将目的多核苷酸引入生物体细胞。

[0269] 供体DNA构建体进一步包含位于目的多核苷酸侧翼的同源的第一区域和第二区域。供体DNA的同源的第一区域和第二区域分别与存在于细胞或生物体基因组的靶位点中或位于所述靶位点侧翼的第一和第二基因组区域共享同源性。

[0270] 供体DNA可以与指导多核苷酸进行系链。系链的供体DNA可以允许共定位靶和供体DNA，可用于基因组编辑、基因插入和靶向的基因组调节，并且还可以用于靶向有丝分裂后期细胞，在这些细胞中内源HR机制的功能预计会大大降低(Mali等人，2013Nature Methods[自然方法]第10卷:957-963)。

[0271] 由靶和供体多核苷酸共享的同源性或序列同一性的量可以变化，并且包括总长度和/或在约1-20bp、20-50bp、50-100bp、75-150bp、100-250bp、150-300bp、200-400bp、250-500bp、300-600bp、350-750bp、400-800bp、450-900bp、500-1000bp、600-1250bp、700-1500bp、800-1750bp、900-2000bp、1-2.5kb、1.5-3kb、2-4kb、2.5-5kb、3-6kb、3.5-7kb、4-8kb、5-10kb，或多达并包括靶位点的总长度的范围内具有单位整数值的区域。这些范围包括所述范围内的每个整数，例如1-20bp的范围包括1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19和20bp。同源性的量也可以通过在两个多核苷酸的完整比对长度上的百分

比序列同一性来描述，其包括约至少50％、55％、60％、65％、70％、71％、72％、73％、74％、75％、76％、77％、78％、79％、80％、81％、82％、83％、84％、85％、86％、87％、88％、89％、90％、91％、92％、93％、94％、95％、96％、97％、98％、98％至99％、99％、99％至100％或100％的百分比序列同一性。足够的同源性包括多核苷酸长度、总体百分比序列同一性，和任选地连续核苷酸的保守区域或局部百分比序列同一性的任何组合，例如，足够的同源性可以被描述为与靶基因座的区域具有至少80％序列同一性的75-150bp的区域。还可以通过用来在高严格条件下特异性杂交的两个多核苷酸的预测能力来描述足够的同源性，参见例如Sambrook等人，(1989)Molecular Cloning:A Laboratory Manual[分子克隆:实验室手册](Cold Spring Harbor Laboratory Press,NY[纽约冷泉港实验室出版社])；Current Protocols in Molecular Biology[分子生物学现代方案]，Ausubel等人，编辑(1994)Current Protocols[实验室指南](Greene Publishing Associates,Inc.[格林出版合伙公司]和John Wiley&Sons,Inc.[约翰威利父子公司])；以及Tijssen(1993)Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes[生物化学和分子生物学中的实验室技术--与核酸探针杂交](Elsevier[爱思唯尔出版社]，纽约)。

[0272] 在给定的基因组区域和在供体DNA上发现的相应的同源的区域之间的结构相似性可以是允许同源重组发生的任何程度的序列同一性。例如，由供体DNA的“同源的区域”和生物体基因组的“基因组区域”共享的同源性或序列同一性的量可以是至少50％、55％、60％、65％、70％、75％、80％、81％、82％、83％、84％、85％、86％、87％、88％、89％、90％、91％、92％、93％、94％、95％、96％、97％、98％、99％或100％序列同一性，这样使得序列进行同源重组。

[0273] 供体DNA上的同源的区域可以与靶位点侧翼的任何序列具有同源性。虽然在一些情况下，同源的区域与紧邻靶位点侧翼的基因组序列共享显著的序列同源性，但是应当认识到同源的区域可以被设计为与可能更靠近靶位点的5′或3′的区域具有足够的同源性。同源的区域还可以与靶位点的片段以及下游基因组区域具有同源性。

[0274] 在一个实施例中，第一同源的区域进一步包含靶位点中的第一片段，并且第二同源的区域包含靶位点中的第二片段，其中第一片段和第二片段不同。

[0275] 目的多核苷酸

[0276] 在本文中进一步描述了目的多核苷酸，并且包括反映涉及作物发育的那些的商业市场和利益的多核苷酸。目的作物和市场发生变化，以及随着发展中国家打开国际市场，新作物和技术也将出现。此外，随着我们对农艺学性状和特征(例如产量和杂种优势增加)的理解逐渐深入，对用于基因工程的基因的选择将会相应变化。

[0277] 目的多核苷酸的一般类别包括，例如涉及信息的那些目的基因(例如锌指)，涉及通讯的那些基因(例如激酶)，以及涉及管家的那些基因(例如热休克蛋白)。更具体的目的多核苷酸包括但不限于:涉及作物产量、籽粒质量、作物营养成分、淀粉和碳水化合物质量和数量的基因、连同及影响籽粒大小、蔗糖载量、蛋白量和数量、固氮和/或氮利用、脂肪酸和油组成的那些基因、编码赋予对非生物胁迫(例如干旱、氮、温度、盐度、毒性金属、或痕量元素)的抗性的蛋白，或赋予对毒素(例如杀有害生物剂和除草剂)的抗性的那些蛋白的基因、编码赋予对生物胁迫(例如真菌、病毒、细菌、昆虫和线虫的攻击以及与这些生物体相关

的疾病的发展)的抗性的蛋白的基因。

[0278] 除了使用传统的育种方法之外,还可通过遗传方式改变农艺学上重要的性状(例如油、淀粉、和蛋白含量)。修饰包括增加油酸、饱和及不饱和油的含量、增加赖氨酸和硫的水平、提供必需氨基酸、以及还有对淀粉的修饰。在美国专利号5,703,049、5,885,801、5,885,802和5,990,389中描述了戈多硫蛋白(hordothionin)的蛋白修饰。

[0279] 目的多核苷酸序列可以编码涉及提供疾病或有害生物抗性的蛋白。"疾病抗性"或"有害生物抗性"意在是植物避免为植物-病原体相互作用后果的有害症状的发生。有害生物抗性基因可以编码对严重影响产量的有害生物的抗性,这些有害生物例如根虫、切根虫、欧洲玉蜀黍螟等。疾病抗性基因和抗昆虫基因,例如用于抗细菌保护的溶菌酶或天蚕杀菌肽,或用于抗真菌保护的蛋白,例如防御素、葡聚糖酶、或几丁质酶,或用于控制线虫或昆虫的苏云金芽孢杆菌内毒素、蛋白酶抑制剂、胶原酶、凝集素、或糖苷酶,均是有用的基因产物的实例。编码疾病抗性性状的基因包括解毒基因,例如抗伏马毒素(美国专利号5,792,931);无毒力(avr)和疾病抗性(R)基因(Jones等人(1994)Science[科学]266:789;Martin等人(1993)Science[科学]262:1432;和Mindrinos等人(1994)Cell[细胞]78:1089);等。抗昆虫基因可以编码对严重影响产量的有害生物的抗性,这些有害生物例如根虫、切根虫、欧洲玉蜀黍螟等。此类基因包括,例如,苏云金芽孢杆菌毒性蛋白基因(美国专利号5,366,892;5,747,450;5,736,514;5,723,756;5,593,881;和Geiser等人(1986)Gene[基因]48:109);等。

[0280] "除草剂抗性蛋白"或由"除草剂抗性编码核酸分子"表达生成的蛋白包括这样的蛋白,其赋予细胞与未表达该蛋白的细胞相比耐受更高浓度除草剂的能力,或赋予细胞与未表达该蛋白的细胞相比对某种浓度的除草剂耐受更长时段的能力。除草剂抗性性状可通过如下基因引入进植物中:编码对起到抑制乙酰乳酸合酶(ALS,也称为乙酰羟基酸合酶,AHAS)的作用的除草剂(特别是磺酰脲(sulfonylurea)(UK:磺酰脲(sulphonylurea))类除草剂)的抗性的基因、编码对起到抑制谷氨酰胺合酶的作用的除草剂(例如草丁膦或basta)的抗性的基因(例如bar基因)、编码对草甘膦的抗性的基因(例如EPSP合酶基因和GAT基因)、编码对HPPD抑制剂的抗性的基因(例如HPPD基因)或本领域已知的其他此类基因。参见例如美国专利号7,626,077、5,310,667、5,866,775、6,225,114、6,248,876、7,169,970、6,867,293和9,187,762。bar基因编码对除草剂basta的抗性,nptII基因编码对抗生素卡那霉素和遗传霉素的抗性,以及ALS-基因突变体编码对除草剂氯磺隆的抗性。

[0281] 此外,认识到目的多核苷酸还可以包括与针对目的所靶向的基因序列的信使RNA(mRNA)的至少一部分互补的反义序列。构建反义核苷酸以与相应的mRNA杂交。可以对该反义序列作出修饰,只要该序列与相应的mRNA杂交并干扰相应的mRNA的表达。在该方式中,可以使用与相应的反义序列具有70%、80%、或85%序列同一性的反义构建体。此外,反义核苷酸的部分可以用来破坏该靶基因的表达。通常,可以使用至少50个核苷酸、100个核苷酸、200个核苷酸、或更多个核苷酸的序列。

[0282] 此外,目的多核苷酸还可以按有义取向来使用从而抑制植物中内源基因的表达。以有义取向使用多核苷酸用于抑制植物中基因表达的方法是本领域已知的。这些方法通常涉及用包含启动子的DNA构建体的转化植物,该启动子可操作地连接到至少一部分的对应于该内源基因的转录物的核苷酸序列上,驱动在植物中的表达。通常,此类核苷酸序列与内

源基因的转录物的序列具有实质性的序列同一性，通常大于约65％序列同一性、约85％序列同一性、或大于约95％序列同一性。参见美国专利号5,283,184和5,034,323。

[0283] 目的多核苷酸还可以是表型标志物。表型标志物是可筛选或可选择标志物，其包括视觉标志物和可选择标志物，无论它是阳性还是阴性可选择标志物。可以使用任何表型标志物。具体地，可选择或可筛选标志物包含允许人们通常在特定条件下鉴定或选择包含它的分子或细胞或对其进行选择的DNA区段。这些标志物可以编码活性，例如但不限于RNA、肽或蛋白的产生，或可以提供RNA、肽、蛋白、无机和有机化合物或组合物等的结合位点。

[0284] 可选择标志物的实例包括但不限于包含限制酶位点的DNA区段；编码提供对包括抗生素在内的其他毒性化合物的抗性的产物的DNA区段，该抗生素例如是大观霉素、氨苄青霉素、卡那霉素、四环素、巴斯塔(Basta)、新霉素磷酸转移酶II(NEO)和潮霉素磷酸转移酶(HPT)；编码另外在受体细胞中缺少的产物的DNA区段(例如，tRNA基因、营养缺陷型标志物)；编码可以容易地鉴定的产物的DNA区段(例如，表型标志物例如β-半乳糖苷酶、GUS；荧光蛋白，例如绿色荧光蛋白(GFP)、青色荧光蛋白(CFP)、黄色荧光蛋白(YFP)、红色荧光蛋白(RFP)和细胞表面蛋白)；产生用于PCR的新引物位点(例如，以前未并列的两个DNA序列的并列)，包含通过限制性内切核酸酶或其他DNA修饰酶、化学品等不起作用或起作用的DNA序列；并且包含允许其鉴定的特异性修饰(例如，甲基化)所需的DNA序列。

[0285] 另外的可选择标志物包括赋予除草剂化合物(例如磺酰脲、草胺磷、溴草腈、咪唑啉酮和2,4-二氯苯氧基乙酸酯(2,4-D))抗性的基因。参见例如，用于对磺酰脲、咪唑啉酮、三唑并嘧啶磺酰胺、嘧啶水杨酸和磺酰基氨基羰基-三唑啉酮(Shaner和Singh，1997，Herbicide Activity:Toxicol Biochem Mol Biol[除草剂活性:毒理学，生物化学，分子生物学]69-110)；草甘膦抗性5-烯醇丙酮莽草酸-3-磷酸(EPSPS)(Saroha等人，1998，J.Plant Biochemistry&Biotechnology[植物生物化学&生物技术杂志]卷7:65-72)的抗性的乙酰乳酸合酶(ALS)；

[0286] 目的多核苷酸包括与其他性状(例如但不限于除草剂抗性或本文描述的任何其他性状)组合堆叠或使用的基因。目的多核苷酸和/或性状可以在复杂性状基因座中堆叠在一起，如2013年10月3日公开的US 20130263324和2013年8月1日公开的WO/2013/112686中所述。

[0287] 目的多肽包括由本文描述的目的多核苷酸编码的蛋白或多肽。

[0288] 进一步提供了用于鉴定至少一个植物细胞的方法，该植物细胞在其基因组中包含在靶位点处整合的目的多核苷酸。可以使用多种方法来鉴定在靶位点处或靶位点附近插入到基因组中的那些植物细胞。此类方法可被认为是直接分析靶序列以检测靶序列中的任何变化，包括但不限于PCR方法、测序方法、核酸酶消化、DNA印迹法、及其任何组合。参见例如，2009年5月21日公开的US 20090133152。所述方法还包括从植物细胞重新获得包含整合至其基因组中的目的多核苷酸的植物。所述植物可以是不育的或可育的。应当认识到，可以提供任何目的多核苷酸，将该多核苷酸在靶位点处整合到植物的基因组中，并在植物中表达。

[0289] 用于在植物中表达的序列的优化

[0290] 本领域中可获得用于合成植物偏好性基因的方法。参见，例如，美国专利号5,380,831和5,436,391，以及Murray等人(1989)Nucleic Acids Res.[核酸研究]17:477-498。已知另外的序列修饰以增强在植物宿主中的基因表达。例如，这些序列修饰包括消除:编码假

多聚腺苷酸化信号的一个或多个序列、一个或多个外显子-内含子剪接位点信号、一个或多个转座子样重复、以及其他可能对基因表达有害的此类良好表征的序列。可以将序列的G-C含量调节至通过参考宿主植物细胞中表达的已知基因而计算出的给定植物宿主的平均水平。当可能时，修饰序列以避免出现一个或多个预测的发夹二级mRNA结构。因此，本公开的“植物优化的核苷酸序列”包括一个或多个此类序列修饰。

[0291] 表达元件

[0292] 可以将本文的编码Cas蛋白或其他CRISPR系统组分的任何多核苷酸功能性连接至异源表达元件，以促进宿主细胞中的转录或调节。此类表达元件包括但不限于：启动子、前导子、内含子和终止子。表达元件可以是“最小的”-意指源自天然来源的较短序列，其仍充当表达调节子或修饰子起作用。可替代地，表达元件可以是“优化的”-意指其多核苷酸序列已经从其天然状态改变，以便在特定宿主细胞中以更期望的特征起作用。可替代地，表达元件可以是“合成的”-意指其是用计算机设计的并且被合成用于在宿主细胞中使用。合成的表达元件可以是完全合成的或部分合成的(包含天然存在的多核苷酸序列的片段)。

[0293] 已经显示某些启动子能够以比其他启动子更高的速率引导RNA合成。这些被称为“强启动子”。已经显示某些其他启动子仅以较高的水平在特定类型的细胞或组织中指导RNA合成，并且如果所述启动子优选在某些组织中而且还以降低的水平在其他组织中指导RNA合成则通常将其称为“组织特异性启动子”或“组织偏好性启动子”。

[0294] 植物启动子包括能够在植物细胞中起始转录的启动子。关于植物启动子的综述，参见Potenza等人，2004 In vitro Cell Dev Biol[体外细胞与发育生物学]40:1-22；Porto等人，2014，Molecular Biotechnology[分子生物技术](2014)，56(1)，38-49。

[0295] 组成型启动子包括，例如，核心CaMV 35S启动子(Odell等人，(1985)Nature[自然]313:810-2)；稻肌动蛋白(McElroy等人，(1990)Plant Cell[植物细胞]2:163-71)；泛素(Christensen等人，(1989)Plant Mol Biol[植物分子生物学]12:619-32；ALS启动子(美国专利号5,659,026)等。

[0296] 组织偏好性启动子可以用于靶向特定植物组织内的增强的表达。组织偏好性启动子包括，例如，2013年7月11日公开的WO 2013103367，Kawamata等人，(1997)Plant Cell Physiol[植物细胞生理学]38:792-803；Hansen等人，(1997)Mol Gen Genet[分子和普通遗传学]254:337-43；Russell等人，(1997)Transgenic Res[转基因研究]6:157-68；Rinehart等人，(1996)Plant Physiol[植物生理学]112:1331-41；Van Camp等人，(1996)Plant Physiol.[植物生理学]112:525-35；Canevascini等人，(1996)Plant Physiol.[植物生理学]112:513-524；Lam，(1994)Results Probl Cell Differ[细胞分化中的结果与问题]20:181-96；以及Guevara-Garcia等人，(1993)Plant J.[植物杂志]4:495-505。叶偏好性启动子包括，例如，Yamamoto等人，(1997)Plant J[植物杂志]12:255-65；Kwon等人，(1994)Plant Physiol[植物生理学]105:357-67；Yamamoto等人，(1994)Plant Cell Physiol[植物细胞生理学]35:773-8；Gotor等人，(1993)Plant J[植物杂志]3:509-18；Orozco等人，(1993)Plant Mol Biol[植物分子生物学]23:1129-38；Matsuoka等人，(1993)Proc.Natl.Acad.Sci.USA[美国国家科学院院刊]90:9586-90；Simpson等人，(1958)EMBO J[欧洲分子生物学学会杂志]4:2723-9；Timko等人，(1988)Nature[自然]318:57-8。根偏好性启动子包括，例如，Hire等人，(1992)Plant Mol Biol[植物分子生物学]20:207-18(大豆

根特异性谷氨酰胺合酶基因);Miao等人,(1991)Plant Cell[植物细胞]3:11-22(胞质谷氨酰胺合酶(GS));Keller和Baumgartner,(1991)Plant Cell[植物细胞]3:1051-61(法国菜豆的GRP 1.8基因中的根特异性控制元件);Sanger等人,(1990)Plant Mol Biol[植物分子生物学]14:433-43(根癌农杆菌(A.tumefaciens)的甘露氨酸合酶(MAS)的根特异性启动子);Bogusz等人,(1990)Plant Cell[植物细胞]2:633-41(从榆科糙叶山黄麻(Parasponia andersonii)和山黄麻(Trema tomentosa)分离的根特异性启动子);Leach和Aoyagi,(1991)Plant Sci[植物科学]79:69-76(发根农杆菌(A.rhizogenes)rolC和rolD根诱导型基因);Teeri等人,(1989)EMBO J[欧洲分子生物学学会杂志]8:343-50(农杆菌伤口诱导的TR1′和TR2′基因);VfENOD-GRP3基因启动子(Kuster等人,(1995)Plant Mol Biol[植物分子生物学]29:759-72);以及rolB启动子(Capana等人,(1994)Plant Mol Biol[植物分子生物学]25:681-91);菜豆球蛋白基因(Murai等人,(1983)Science[科学]23:476-82;Sengopta-Gopalen等人,(1988)Proc.Natl.Acad.Sci.USA[美国国家科学院院刊]82:3320-4)。还参见美国专利号5,837,876;5,750,386;5,633,363;5,459,252;5,401,836;5,110,732和5,023,179。

[0297] 种子偏好性启动子包括在种子发育期间有活性的种子特异性启动子以及在种子发芽期间有活性的种子发芽性启动子两者。参见Thompson等人,(1989)BioEssays[生物学分析]10:108。种子偏好性启动子包括但不限于Cim1(细胞分裂素诱导的信号);cZ19B1(玉蜀黍19kDa玉米蛋白);和milps(肌醇-1-磷酸盐合酶);以及例如,在2000年3月2日公开的WO 2000011177和美国专利6,225,529中公开的那些。对于双子叶植物,种子偏好性启动子包括但不限于:菜豆β-菜豆素、油菜籽蛋白、β-伴大豆球蛋白、大豆凝集素、十字花科蛋白等。对于单子叶植物,种子偏好性启动子包括但不限于玉蜀黍15kDa玉蜀黍蛋白、22kDa玉蜀黍蛋白、27kDaγ玉蜀黍蛋白、蜡质、收缩素1、收缩素2、球蛋白1、油质蛋白和nuc1。还参见2000年3月9日公开的WO 2000012733,其中公开了来自END1和END2基因的种子偏好性启动子。

[0298] 可以使用化学诱导型(调节型)启动子以通过应用外源化学调节剂来调节原核和真核细胞或生物体中的基因表达。在应用化学品诱导基因表达的情况下启动子可以是化学品诱导型启动子,或者在应用化学品阻抑基因表达的情况下启动子可以是化学品阻抑型启动子。化学品诱导型启动子包括但不限于:由苯磺酰胺除草剂安全剂激活的玉蜀黍In2-2启动子(De Veylder等人,(1997)Plant Cell Physiol[植物细胞生理学]38:568-77)、由用作出苗前除草剂的疏水性亲电子化合物激活的玉蜀黍GST启动子(GST-II-27,1993年1月21日公开的WO 1993001294)、以及由水杨酸激活的烟草PR-1a启动子(Ono等人,(2004)Biosci Biotechnol Biochem[生物科学生物技术生物化学]68:803-7)。其他化学品调节型启动子包括类固醇反应启动子(参见,例如,糖皮质激素诱导型启动子(Schena等人,(1991)Proc.Natl.Acad.Sci.USA[美国国家科学院院刊]88:10421-5;McNellis等人,(1998)Plant J[植物杂志]14:247-257);四环素诱导型启动子和四环素阻抑型启动子(Gatz等人,(1991)Mol Gen Genet[分子和普通遗传学]227:229-37;美国专利号5,814,618和5,789,156)。

[0299] 在被病原体感染后诱导的病原体诱导型启动子包括但不限于调节PR蛋白、SAR蛋白、β-1,3-葡聚糖酶、几丁质酶等的表达的启动子。

[0300] 胁迫诱导型启动子包括RD29A启动子(Kasuga等人(1999)Nature Biotechnol[自然生物技术].17:287-91)。本领域技术人员熟悉模拟胁迫条件(如干旱、渗透胁迫、盐胁迫、

和温度胁迫)并评价植物的胁迫耐受性的规程,所述植物已经遭受了模拟的或天然存在的胁迫条件。

[0301] 在植物细胞中有用的诱导型启动子的另一个实例是ZmCAS1启动子,描述于2013年11月21日公开的US 20130312137中。

[0302] 不断发现在植物细胞中有用的不同类型的新启动子;许多实例可以在Okamuro和Goldberg,(1989)The Biochemistry of Plants[植物生物化学],第115卷,Stumpf和Conn编辑(纽约,纽约州:学术出版社)1-82页的汇编中发现。

[0303] 用新颖CRISPR-Cas系统组分修饰基因组

[0304] 如本文描述,受指导的Cas内切核酸酶可以识别、结合DNA靶序列,并且引入单链(切口)或双链断裂。一旦在DNA中诱导单链断裂或双链断裂,则细胞的DNA修复机制被激活来修复断裂。易错DNA修复机制可以在双链断裂位点处产生突变。用来将断裂的末端结合在一起的最常见的修复机制是非同源末端连接(NHEJ)途径(Bleuyard等人,(2006)DNA Repair[DNA修复]5:1-12)。染色体的结构完整性典型地通过修复来保存,但是缺失、插入或其他重排(如染色体易位)是可能的(Siebert和Puchta,2002Plant Cell[植物细胞]14:1121-31;Pacher等人,2007Genetics[遗传学]175:21-9)。

[0305] DNA双链断裂似乎是刺激同源重组途径的有效因子(Puchta等人,(1995)Plant Mol Biol[植物分子生物学]28:281-92;Tzfira和White,(2005)Trends Biotechnol[生物技术趋势]23:567-9;Puchta,(2005)J Exp Bot[实验植物学杂志]56:1-14)。使用DNA断裂剂,在植物中的人工构建的同源DNA重复序列之间观察到同源重组的两倍至九倍的增加(Puchta等人,(1995)Plant Mol Biol[植物分子生物学]28:281-92)。在玉蜀黍原生质体中,用线性DNA分子进行的实验证实了在质粒之间增强的同源重组(Lyznik等人,(1991)Mol Gen Genet[分子和普通遗传学]230:209-18)。

[0306] 同源-定向修复(HDR)是在细胞中用来修复双链DNA和单链DNA断裂的机制。同源-定向修复包括同源重组(HR)和单链退火(SSA)(Lieber.2010 Annu.Rev.Biochem[生物化学年鉴].79:181-211)。HDR的最常见形式称为同源重组(HR),其在供体和受体DNA之间具有最长的序列同源性要求。HDR的其他形式包括单链退火(SSA)和断裂诱导的复制,并且这些需要相对于HR更短的序列同源性。缺口(单链断裂)处的同源-定向修复可以经由与在双链断裂处的HDR不同的机制发生(Davis和Maizels.PNAS[美国国家科学院院刊](0027-8424),111(10),第E924-E932页)。

[0307] 原核和真核细胞或生物细胞的基因组的改变,例如通过同源重组(HR),对于基因工程而言的有力工具。已经证明了在植物中(Halfter等人,(1992)Mol Gen Genet[分子和普通遗传学]231:186-93)和昆虫中(Dray和Gloor,1997,Genetics[遗传学]147:689-99)的同源重组。在其他生物体中也可以实现同源重组。例如,在寄生原生动物利什曼原虫中,至少需要150-200bp的同源性进行同源重组(Papadopoulou和Dumas,(1997)Nucleic Acids Res[核酸研究]25:4278-86)。在丝状真菌构巢曲霉中,已经用仅50bp侧翼同源性实现基因替代(Chaveroche等人,(2000)Nucleic Acids Res[核酸研究]28:e97)。在纤毛虫嗜热四膜虫中也已经证明了靶向基因替代(Gaertig等人,(1994)Nucleic Acids Res[核酸研究]22:5391-8)。在哺乳动物中,使用可以在培养基中生长、转化、选择、和引入小鼠胚胎中的多能胚胎干细胞系(ES),同源重组在小鼠中已经是最成功的(Watson等人,(1992)Recombinant

DNA[重组DNA],第2版,Scientific American Books distributed by WH Freeman&Co.[由WH Freeman&Co.公司发行的科学美国人图书])。

[0308] 基因靶向

[0309] 本文描述的指导多核苷酸/Cas系统可以用于基因靶向。

[0310] 通常,可以通过在具有与合适的多核苷酸组分缔合的Cas蛋白的细胞中的特异性多核苷酸序列处切割一条或两条链来进行DNA靶向。一旦在DNA中诱导单链断裂或双链断裂,则细胞的DNA修复机制被激活来经由会导致靶位点处的修饰的非同源末端连接(NHEJ)、或同源定向修复(HDR)过程修复断裂。

[0311] 靶位点处的DNA序列的长度可以变化,并且包括例如为至少12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30个或多于30个核苷酸长度的靶位点。还有可能靶位点可以是回文的,即,一条链上的序列与在互补链上以相反方向的读取相同。切口/切割位点可以在靶序列内,或者切口/切割位点可以在靶序列之外。在另一种变异中,切割可以发生在彼此正好相对的核苷酸位置处,以产生平端切割,或者在其他情况下,切口可以交错以产生单链突出端,也称为"粘性末端"或"交错末端",其可以是5′突出端或3′突出端。还可以使用基因组靶位点的活性变体。此类活性变体可以包含与给定靶位点至少65%、70%、75%、80%、85%、90%、91%、92%、93%、94%、95%、96%、97%、98%、99%或更高的序列同一性,其中所述活性变体保留生物活性,因此能够被Cas内切核酸酶识别和切割。

[0312] 测量由内切核酸酶引起的靶位点的单链或双链断裂的测定是本领域已知的,并且通常测量试剂在包含识别位点的DNA底物上的总体活性和特异性。

[0313] 本文的靶向方法能以例如在该方法中靶向两个或更多个DNA靶位点的这样的方式进行。这种方法可以任选地被表征为多重方法。在某些实施例中,可以同时靶向两个、三个、四个、五个、六个、七个、八个、九个、十个或更多个靶位点。多路复用方法典型地通过本文的靶向方法进行,其中提供了多个不同的RNA组分,每一个被设计成将指导多核苷酸/Cas内切核酸酶复合物引导到唯一的DNA靶位点。

[0314] 基因编辑

[0315] 编辑组合有DSB和修饰模板的基因组序列的过程通常包括:向宿主细胞引入DSB诱导剂或编码DSB诱导剂的核酸(识别染色体序列中的靶序列并且能够诱导基因组序列中的DSB),和与待编辑的核苷酸序列相比时包含至少一个核苷酸变化的至少一个多核苷酸修饰模板。多核苷酸修饰模板还可以包含侧翼于所述至少一个核苷酸变化的核苷酸序列,其中侧翼序列与侧翼于DSB的染色体区域基本同源。已经在例如以下中描述了使用DSB诱导剂(如Cas-gRNA复合物)的基因组编辑:2015年3月19日公开的US 20150082478,2015年2月26日公开的WO 2015026886,2016年1月14日公开的WO 2016007347,以及于2016年2月18日公开的WO/2016/025131。

[0316] 已经描述了指导RNA/Cas内切核酸酶系统的一些用途(参见例如:2015年3月19日公开的US 20150082478 A1,2015年2月26日公开的WO 2015026886和2015年2月26日公开的US 20150059010)并且包括但不限于修饰或取代目的核苷酸序列(如调节元件)、目的多核苷酸插入、基因敲除、基因敲入、剪接位点的修饰和/或引入交替剪接位点、编码目的蛋白的核苷酸序列的修饰、氨基酸和/或蛋白融合物、以及通过在目的基因中表达反向重复序列引起的基因沉默。

[0317] 可以按不同方式改变蛋白,这些方式包括氨基酸取代、缺失、截短、和插入。用于此类操作的方法通常是已知的。例如,可以通过在DNA中的突变制备一种或多种蛋白的氨基酸序列变体。用于诱变和核苷酸序列改变的方法包括,例如,Kunkel,(1985)Proc.Nat/.Acad.Sci.USA[美国国家科学院院刊]82:488-92;Kunkel等人,(1987)Meth Enzymol[酶学方法]154:367-82;美国专利号4,873,192;Walker和Gaastra编辑(1983)Techniques in Molecular Biology[分子生物学技术](MacMillan Publishing Company,New York[麦克米伦出版公司,纽约]),以及其中所引用的文献。发现关于不太可能影响蛋白生物学活性的氨基酸取代的引导,例如,在Dayhoff等人,(1978)Atlas of Protein Sequence and Structure[蛋白序列和结构图谱集](Natl Biomed Res Found,Washington,D.C.[国家生物医学研究基金会,美国华盛顿哥伦比亚特区])的模型中。保守取代,例如将一个氨基酸与具有相似特性的另一个氨基酸交换,会是优选的。未预期保守缺失、插入、和氨基酸取代会产生在蛋白特征中的根本变化,并且可以通过常规筛选测定来评价任何取代、缺失、插入、或其组合的作用。对双链-断裂-诱导活性的测定是已知的,并且通常测量试剂对包含靶位点的DNA底物的总体活性和特异性。

[0318] 本文描述了用Cas(CRISPR Associated)内切核酸酶进行基因组编辑的方法。在对指导RNA(或指导多核苷酸)和PAM序列进行表征后,包含Cas内切核酸酶和指导RNA(或指导多核苷酸)的核糖核蛋白(RNP)复合物可用于修饰靶多核苷酸,所述靶多核苷酸包括但不限于:其他生物(包括植物)中的合成DNA、分离的基因组DNA或染色体DNA。为了促进最佳表达和核定位(对于真核细胞),可以对包含Cas内切核酸酶的基因进行优化,然后通过本领域已知的方法将其作为DNA表达盒递送至细胞中。也可以将必需包含活性RNP的组分作为RNA(具有或不具有保护RNA免于降解的修饰)或作为有帽或无帽的mRNA(Zhang,Y.等人,2016,Nat.Commun.[自然通讯]7:12617)或Cas蛋白指导多核苷酸复合物(公开于2017年4月27日的WO 2017070032)、或其任何组合递送。另外,复合物的一个或多个部分可以从DNA构建体表达,而将其他组分作为RNA(具有或不具有保护RNA免于降解的修饰)或以带帽或不带帽的mRNA(Zhang等人2016Nat.Commun.[自然通讯]7:12617)或Cas蛋白指导多核苷酸复合物(公开于2017年4月27日的WO 2017070032)或其任何组合递送。为了体内产生crRNA,tRNA衍生的元件也可以用于募集内源RNA酶以将crRNA转录物切割成能够将复合物引导至其DNA靶位点的成熟形式,例如,如2017年6月22日公开的WO 2017105991中所述。此外,可以通过改变切割结构域中的关键催化残基来使Cas内切核酸酶的切割活性失活(Sinkunas,T.等人,2013,EMBO J[欧洲分子生物学学会杂志].32:385-394),从而产生受RNA指导的解旋酶,其可用于增强同源定向修复,诱导转录激活或重塑局部DNA结构。而且,Cas切割和解旋酶结构域的活性可以都被敲除并与其他DNA剪切、DNA切口、DNA结合、转录激活、转录阻遏、DNA重塑、DNA脱氨、DNA解旋、DNA重组增强、DNA整合、DNA倒置和DNA修复剂组合使用。

[0319] 可以如2016年11月24日公开的WO 2016186946和2016年11月24日公开的WO 2016186953中所述推导用于CRISPR-Cas系统(如果存在的话)和CRISPR-Cas系统的其他组分(例如可变靶向结构域、crRNA重复序列、环、反重复序列)的tracrRNA的转录方向。

[0320] 如本文所述,一旦建立了适当的指导RNA要求,就可以检查本文公开的每个新系统的PAM偏好。如果切割性RNP复合物(包含Cas内切核酸酶和指导多核苷酸)导致随机PAM文库的降解,则可以通过诱变关键残基或通过在无ATP的情况下组装反应使活性无效,从而将复

合物转化为切口酶，如先前所述(Sinkunas，T.等人，2013，EMBO J.[欧洲分子生物学学会杂志]32:385-394)。可以利用由两个前间隔子靶隔开的PAM随机化的两个区域来生成双链DNA断裂，所述双链DNA断裂可以被捕获并测序以检查支持复合物切割的PAM序列。

[0321] 在一个实施例中，本发明描述了用于修饰细胞的基因组中靶位点的方法，所述方法包括将至少一种本文所述的PGEN引入细胞，并鉴定在所述靶处具有修饰的至少一个细胞，其中所述靶位点处的修饰选自下组，该组由以下组成：(i)至少一个核苷酸的替代、(ii)至少一个核苷酸的缺失、(iii)至少一个核苷酸的插入、和(iv)(i)-(iii)的任何组合。

[0322] 待编辑的核苷酸可以位于由Cas内切核酸酶识别和切割的靶位点的内部或外部。在一个实施例中，该至少一个核苷酸修饰不是由Cas内切核酸酶识别和切割的靶位点上的修饰。在另一个实施例中，所述待编辑的至少一个核苷酸和基因组靶位点之间有至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、30、40、50、100、200、300、400、500、600、700、900或1000个核苷酸。

[0323] 可以通过插入缺失(通过NHEJ在靶DNA序列中插入或缺失核苷酸碱基)，或通过特异性去除在靶向位点处或其附近处降低或完全破坏序列功能的序列来产生敲除。

[0324] 指导多核苷酸/Cas内切核酸酶诱导的靶向突变可以发生在位于由Cas内切核酸酶识别和切割的基因组靶位点内部或外部的核苷酸序列中。

[0325] 用于编辑细胞的基因组中的核苷酸序列的方法可以是通过恢复无功能基因产物的功能而不使用外源可选择标志物的方法。

[0326] 在一个实施例中，本发明描述了用于修饰细胞的基因组中的靶位点的方法，所述方法包括将至少一种本文所述的PGEN和至少一种供体DNA引入细胞中，其中所述供体DNA包含目的多核苷酸，并且任选地，所述方法进一步包括鉴定至少一个将所述目的多核苷酸整合到所述靶位点中或附近的细胞。

[0327] 在一方面，本文公开的方法可采用同源重组(HR)以在靶位点处提供目的多核苷酸的整合。

[0328] 可以采用多种方法和组合物来产生具有通过本文所述的CRISPR-Cas系统组分的活性插入靶位点的目的多核苷酸的细胞或生物。在本文所述的一种方法中，经由供体DNA构建体，将目的多核苷酸引入生物体细胞。如本文所用，“供体DNA”是DNA构建体，其包括待插入到Cas内切核酸酶的靶位点的目的多核苷酸。供体DNA构建体进一步包含位于目的多核苷酸侧翼的同源的第一区域和第二区域。供体DNA的同源的第一区域和第二区域分别与存在于细胞或生物体基因组的靶位点中或位于所述靶位点侧翼的第一和第二基因组区域共享同源性。

[0329] 供体DNA可以与指导多核苷酸进行系链。系链的供体DNA可以允许共定位靶和供体DNA，可用于基因组编辑、基因插入和靶向的基因组调节，并且还可以用于靶向有丝分裂后期细胞，在这些细胞中内源HR机制的功能预计会大大降低(Mali等人，2013Nature Methods[自然方法]第10卷:957-963)。

[0330] 由靶和供体多核苷酸共享的同源性或序列同一性的量可以变化，并且包括总长度和/或在约1-20bp、20-50bp、50-100bp、75-150bp、100-250bp、150-300bp、200-400bp、250-500bp、300-600bp、350-750bp、400-800bp、450-900bp、500-1000bp、600-1250bp、700-1500bp、800-1750bp、900-2000bp、1-2.5kb、1.5-3kb、2-4kb、2.5-5kb、3-6kb、3.5-7kb、4-

8kb、5-10kb，或多达并包括靶位点的总长度的范围内具有单位整数值的区域。这些范围包括所述范围内的每个整数，例如1-20bp的范围包括1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19和20bp。同源性的量也可以通过在两个多核苷酸的完整比对长度上的百分比序列同一性来描述，其包括约至少50％、55％、60％、65％、70％、71％、72％、73％、74％、75％、76％、77％、78％、79％、80％、81％、82％、83％、84％、85％、86％、87％、88％、89％、90％、91％、92％、93％、94％、95％、96％、97％、98％、99％或100％的百分比序列同一性。足够的同源性包括多核苷酸长度、总体百分比序列同一性，和任选地连续核苷酸的保守区域或局部百分比序列同一性的任何组合，例如，足够的同源性可以被描述为与靶基因座的区域具有至少80％序列同一性的75-150bp的区域。还可以通过用来在高严格条件下特异性杂交的两个多核苷酸的预测能力来描述足够的同源性，参见例如Sambrook等人，(1989)Molecular Cloning:A Laboratory Manual[分子克隆:实验室手册](Cold Spring Harbor Laboratory Press,NY[纽约冷泉港实验室出版社]);Current Protocols in Molecular Biology[分子生物学现代方案],Ausubel等人，编辑(1994)Current Protocols[实验室指南](Greene Publishing Associates,Inc.[格林出版合伙公司]和John Wiley&Sons,Inc.[约翰威利父子公司]);以及Tijssen(1993)Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes[生物化学和分子生物学中的实验室技术--与核酸探针杂交](Elsevier[爱思唯尔出版社],纽约)。

[0331] 还可以将附加体DNA分子连接至双链断裂中，例如，将T-DNA整合至染色体双链断裂中(Chilton和Que，(2003)Plant Physiol[植物生理学]133:956-65;Salomon和Puchta，(1998)EMBO J.[欧洲分子生物学学会杂志]17:6086-95)。一旦双链断裂周围的序列被改变，例如被涉及双链断裂的成熟的外切核酸酶活性改变，则基因转换途径可以恢复原始结构，如果有同源序列的话，例如非分裂的体细胞中的同源染色体，或DNA复制后的姊妹染色单体(Molinier等人，(2004)Plant Cell[植物细胞]16:342-52)。异位的和/或表观遗传的DNA序列还可以充当用于同源重组的DNA修复模板(Puchta，(1999)Genetics[遗传学]152:1173-81)。

[0332] 在一个实施例中，本公开包括用于编辑细胞的基因组中的核苷酸序列的方法，所述方法包括引入至少一种本文所述的PGEN、和多核苷酸修饰模板，其中所述多核苷酸修饰模板包含所述核苷酸序列的至少一个核苷酸修饰，并且所述方法任选地进一步包括选择至少一个包含经编辑的核苷酸序列的细胞。

[0333] 指导多核苷酸/Cas内切核酸酶系统可以与至少一个多核苷酸修饰模板组合使用以允许编辑(修饰)目的基因组核苷酸序列。(还参见2015年3月19日公开的US 20150082478和2015年2月26日公开的WO 2015026886)。

[0334] 目的多核苷酸和/或性状可以在复杂性状基因座中堆叠在一起，如在2012年9月27日公开的WO 2012129373和2013年8月1日公开的WO 2013112686中所述。本文所述的指导多核苷酸/Cas内切核酸酶系统提供了用来产生双链断裂并允许将性状在复杂性状基因座中堆叠的有效系统。

[0335] 如本文所述的介导基因靶向的指导多核苷酸/Cas系统可以在以下方法中使用，所述方法用于以类似于2012年9月27日公开的WO 2012129373中公开的方式引导异源基因插入和/或产生包含多个异源基因的复杂性状基因座，其中使用如本文公开的指导多核苷酸/

Cas系统来代替使用双链断裂诱导剂引入目的基因。通过将独立的转基因插入在彼此的0.1、0.2、0.3、0.4、0.5、1.0、2、或甚至5厘摩(cM)内，这些转基因可以作为单个遗传基因座进行育种(例如，参见2013年10月3日公开的US 20130263324或2013年3月14日公开的WO 2012129373)。在选择包含转基因的植物后，可以将包含(至少)一个转基因的植物进行杂交从而形成包含全部两个转基因的F1。在来自这些F1(F2或BC1)的后代中，1/500的后代将具有重组在相同的染色体上的两个不同的转基因。然后，可以将复合物基因座繁育为具有全部两个转基因性状的单遗传基因座。可以重复该过程以堆叠尽可能多的性状。

[0336] 已经描述了指导RNA/Cas内切核酸酶系统的进一步用途(参见例如:2015年3月19日公开的US 20150082478,2015年2月26日公开的WO 2015026886,2015年2月26日公开的US 20150059010,2016年1月14日公开的WO 2016007347,和2016年2月18日公开的PCT申请WO 2016025131)并包括但不限于修饰或取代目的核苷酸序列(如调节元件)、目的多核苷酸插入、基因敲除、基因敲入、剪接位点的修饰和/或引入交替剪接位点、编码目的蛋白的核苷酸序列的修饰、氨基酸和/或蛋白融合物、以及通过在目的基因中表达反向重复序列引起的基因沉默。

[0337] 可以评估本文描述的基因编辑组合物和方法产生的特征。可以鉴定与目的表型或性状相关的染色体区间。本领域熟知的多种方法可用于鉴定染色体区间。此类染色体区间的边界扩展到涵盖将与控制目的性状的基因连锁的标志物。换句话说，扩展染色体区间，这样使得位于区间内的任何标志物(包括限定区间的边界的末端标志物)可以用作特定性状的标志物。在一个实施例中，染色体区间包含至少一个QTL，并且此外，确实可以包含多于一个QTL。相同区间中非常接近的多个QTL可以搅乱特定标志物与特定QTL的关联，因为一个标志物可显示与多于一个QTL连锁。相反地，例如如果非常接近的两个标志物显示与期望表型性状共分离，则有时分不清楚是否那些标志物中的每一个鉴定相同QTL或两个不同的QTL。术语“数量性状座位”或“QTL”是指在至少一种遗传背景下(例如在至少一个育种群体中)，与数量表型性状的差异表达关联的DNA区域。QTL的区域涵盖或紧密地连锁于影响所考虑的性状的一个或多个基因。“QTL的等位基因”可以包含在连续的基因组区域或连锁群中的多个基因或其他遗传因子，例如单倍型。QTL的等位基因可以表示在指定窗口内的单倍型，其中所述窗口是可以用一组的一个或多个多态性标志物定义和追踪的连续的基因组区域。单倍型可以指定被窗口内的每一标志物的等位基因的独特指纹定义。

[0338] 除了双链断裂诱导剂，还可以实现位点特异性碱基转化，以工程化一个或多个核苷酸变化，从而在基因组中创建一个或多个编辑。这些包括例如，由C·G至T·A或A·T至G·C碱基编辑脱氨酶介导的位点特异性碱基编辑(Gaudelli等人，Programmable base editing of A·T to G·C in genomic DNA without DNA cleavage[在无DNA切割时基因组DNA中A·T至G·C的可编程碱基编辑]."Nature[自然](2017);Nishida等人“Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems[使用杂交体原核和脊椎动物适应性免疫系统进行靶向核苷酸编辑]."Science[科学]353(6305)(2016);Komor等人“Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage[在无双链DNA切割时基因组DNA中靶碱基的可编程编辑]."Nature[自然]533(7603)(2016):420-4)。与胞苷脱氨酶或腺嘌呤脱氨酶蛋白融合的催化“死亡”或失活Cas9(dCas9)(例如本文公开的Cas9直向同源物的催化失活的

“死亡”形式)成为特异性的碱基编辑器,其可以改变DNA碱基而不会诱导DNA断裂。碱基编辑器转换C-＞T(或在相反链上,G-＞A)或腺嘌呤碱基编辑器将腺嘌呤转换为肌苷,从而在gRNA指定的编辑窗口内导致A-＞G变化。

[0339] 将CRISPR-Cas系统组分引入细胞

[0340] 本文描述的方法不取决于用于将序列引入生物体或细胞中的具体方法,只要多核苷酸或多肽进入生物体的至少一个细胞的内部即可。引入包括提到将核酸合并到真核细胞或原核细胞中,其中核酸可以被并入细胞的基因组中,并且包括提到核酸、蛋白或多核苷酸-蛋白复合物(PGEN、RGEN)被瞬时(直接)提供至细胞中。

[0341] 用于将多核苷酸或多肽或多核苷酸-蛋白复合物引入细胞或生物体的方法是本领域已知的,并且包括但不限于显微注射、电穿孔、稳定转化方法、瞬时转化方法、弹道粒子加速(粒子轰击)、晶须介导的转化、农杆菌介导的转化、直接基因转移、病毒介导的引入、转染、转导、细胞穿透肽、介孔二氧化硅纳米粒子(MSN)-介导的直接蛋白递送、局部应用、有性杂交、有性育种、及其任何组合。

[0342] 例如,指导多核苷酸(指导RNA,cr核苷酸+tracr核苷酸,指导DNA和/或指导RNA-DNA分子)可以作为单链或双链多核苷酸分子直接引入细胞(瞬时地)。指导RNA(或crRNA+tracrRNA)还可以通过引入包含编码指导RNA(或crRNA+tracrRNA)的异源核酸片段的重组DNA分子被间接引入细胞中,所述指导RNA与能够在所述细胞中转录所述指导RNA(或crRNA+tracrRNA)的特异性启动子可操作地连接。特异性启动子可以是但不限于RNA聚合酶III启动子,其允许具有精确定义的未修饰的5′-和3′-末端的RNA转录(Ma等人,2014,Mol.Ther.Nucleic Acids[分子治疗-核酸]3:e161;DiCarlo等人,2013,Nucleic Acids Res.[核酸研究]41:4336-4343;2015年2月26日公开的WO 2015026887)。可以使用能够在细胞中转录指导RNA的任何启动子,并且这些启动子包括可操作地连接到编码指导RNA的核苷酸序列的热休克/热可诱导的启动子。

[0343] 本文中的Cas内切核酸酶,例如本文所述的Cas内切核酸酶可以通过直接引入Cas多肽本身(称为Cas内切核酸酶的直接递送)、编码Cas蛋白的mRNA和/或指导多核苷酸/Cas内切核酸酶复合物本身,使用本领域已知的任何方法而导入细胞。Cas内切核酸酶也可以通过引入编码Cas内切核酸酶的重组DNA分子间接引入细胞。使用本领域已知的任何方法,可以瞬时地将内切核酸酶引入细胞中,或可以将内切核酸酶并入宿主细胞的基因组中。可以用如在2016年5月12日公开的WO 2016073433中描述的细胞穿透肽(CPP),促进内切核酸酶和/或指导的多核苷酸摄取进入细胞。可以使用能够在细胞中表达Cas内切核酸酶的任何启动子,并且这些启动子包括可操作地连接到编码Cas内切核酸酶的核苷酸序列的热休克/热可诱导的启动子。

[0344] 将多核苷酸修饰模板直接递送到植物细胞中可以通过粒子介导递送来实现,并且任何其他直接递送方法,例如但不限于聚乙二醇(PEG)介导的原生质体转染、晶须介导的转化、电穿孔、粒子轰击、细胞穿透肽或介孔二氧化硅纳米粒子(MSN)介导的直接蛋白递送可以成功地用于在真核细胞(例如植物细胞)中递送多核苷酸修饰模板。

[0345] 可以通过本领域已知的任何手段引入供体DNA。可以通过本领域已知的任何转化方法(包括,例如农杆菌介导的转化或生物射弹粒子轰击)提供供体DNA。供体DNA可以瞬时地存在于细胞中,或可以经由病毒复制子引入。在Cas内切核酸酶和靶位点的存在下,供体

DNA被插入到转化植物的基因组中。

[0346] 受指导的Cas系统组分中的任何一个的直接递送可以伴随着可以促进接受指导多核苷酸/Cas内切核酸酶复合物组分的细胞的富集和/或可视化的其他mRNA的直接递送(共递送)。例如,指导多核苷酸/Cas内切核酸酶组分(和/或指导多核苷酸/Cas内切核酸酶复合物本身)与编码表型标志物(例如但不限于转录激活剂如CRC(Bruce等人2000 The Plant Cell[植物细胞]12:65-79)的mRNA直接共递送可通过恢复无功能基因产物的功能而不使用外源性可选择标志物来实现细胞的选择和富集,如在2017年4月27日公开的WO 2017070032中所述。

[0347] 将本文所述的指导RNA/Cas内切核酸酶复合物引入细胞中包括将所述复合物的各组分单独地或组合地引入细胞中,并且直接地(作为RNA(对于指导物)和蛋白(对于Cas内切核酸酶和Cas蛋白亚基或其功能性片段)直接递送)或经由表达这些组分(指导RNA、Cas内切核酸酶、Cas蛋白亚基或其功能性片段)的重组构建体引入。将指导RNA/Cas内切核酸酶复合物(RGEN)引入细胞中包括将该指导RNA/Cas内切核酸酶复合物作为核糖核苷酸-蛋白引入细胞中。可以将该核糖核苷酸-蛋白在引入如本文所述的细胞中之前进行组装。包含指导RNA/Cas内切核酸酶核糖核苷酸蛋白(至少一种Cas内切核酸酶、至少一种指导RNA、至少一种Cas蛋白亚基)的组分可在体外组装或在引入细胞(靶向用于如本文所述基因组修饰)之前通过本领域已知的任何方法组装。

[0348] 植物细胞与人类和动物细胞的不同之处在于,植物细胞含有植物细胞壁,其可以作为RGEN核糖核蛋白的直接递送和/或RGEN组分的直接递送的屏障。

[0349] 可以通过粒子介导的递送(粒子轰击)实现将RGEN核糖核蛋白直接递送到植物细胞中。基于本文所述的实验,技术人员现在可以预想任何其他直接递送方法(例如但不限于聚乙二醇(PEG)介导的对原生质体的转染、电穿孔、细胞穿透肽或介孔二氧化硅纳米粒子(MSN)介导的直接蛋白递送)都可以成功用于将RGEN核糖核蛋白递送到植物细胞中。

[0350] RGEN核糖核蛋白的直接递送允许在细胞的基因组中的靶位点进行基因组编辑,其后可以迅速降解复合物,并且仅允许细胞中短暂存在该复合物。RGEN复合物的这种短暂存在可能导致脱靶效应降低。相比之下,经由质粒DNA序列递送RGEN组分(指导RNA、Cas内切核酸酶)可以导致RGEN从这些质粒的恒定表达,该恒定表达可以加强脱靶效应(Cradick,T.J.等人(2013)Nucleic Acids Res[核酸研究]41:9584-9592;Fu,Y等人(2014)Nat.Biotechnol.[自然生物技术]31:822-826)。

[0351] 直接递送可以通过将指导RNA/Cas内切核酸酶复合物(RGEN)的任何一种组分(例如至少一种向导RNA、至少一种Cas蛋白和至少一种Cas蛋白)与包含微粒子(例如但不限于金粒子、钨粒子和碳化硅晶须粒子)的粒子递送基质组合来实现(还参见2017年4月27日公开的WO 2017070032)。

[0352] 在一个方面,指导多核苷酸/Cas内切核酸酶复合物是复合物,其中形成所述指导RNA/Cas内切核酸酶复合物的指导RNA和Cas内切核酸酶蛋白分别作为RNA和蛋白引入细胞。

[0353] 在一个方面,指导多核苷酸/Cas内切核酸酶复合物是复合物,其中形成所述指导RNA/Cas内切核酸酶复合物的指导RNA和Cas内切核酸酶蛋白和Cas蛋白的至少一个蛋白亚基分别作为RNA和蛋白引入细胞。

[0354] 在一个方面,指导多核苷酸/Cas内切核酸酶复合物是复合物,其中形成所述指导

RNA/Cas内切核酸酶复合物(切割就绪的cascade)的指导RNA和Cas内切核酸酶蛋白和Cascade的至少一个蛋白亚基在体外预组装并作为核糖核苷酸-蛋白复合物引入细胞。

[0355] 用于在真核细胞例如植物或植物细胞中引入多核苷酸、多肽或多核苷酸-蛋白复合物(PGEN,RGEN)的方案是已知的并且包括显微注射(Crossway等人,(1986)Biotechniques[生物技术]4:320-34和美国专利号6,300,543);分生组织转化(美国专利号5,736,369);电穿孔(Riggs等人,(1986)Proc.Natl.Acad.Sci.USA[美国国家科学院院刊]83:5602-6);农杆菌介导的转化(美国专利号5,563,055和5,981,840);晶须介导的转化(Ainley等人2013,Plant Biotechnology Journal[植物生物技术杂志]11:1126-1134;Shaheen A.和M.Arshad 2011 Properties and Applications of Silicon Carbide[碳化硅的特性和应用](2011),345-358,编辑:Gerhardt,Rosario.,出版商:印天科技公司(InTech),里耶卡(Rijeka),克罗地亚(Croatia),代码:69PQBP;ISBN:978-953-307-201-2);直接基因转移(Paszkowski等人,(1984)EMBO J[欧洲分子生物学学会杂志]3:2717-22);以及弹道粒子加速(美国专利号4,945,050;5,879,918;5,886,244;5,932,782;Tomes等人,(1995)“Direct DNA Transfer into Intact Plant Cells via Microprojectile Bombardment”[经由微粒轰击将DNA直接转移到完整植物细胞中]在Plant Cell,Tissue,and Organ Culture:Fundamental Methods[植物细胞、组织和器官培养:基本方法],编辑Gamborg和Phillips(Springer-Verlag,Berlin[柏林施普林格出版社);McCabe等人(1988)Biotechnology[生物技术]6:923-6;Weissinger等人,(1988)Ann Rev Genet[遗传学年鉴]22:421-77;Sanford等人,(1987)Particulate Science and Technology[微粒科学与技术]5:27-37(洋葱);Christou等人,(1988)Plant Physiol[植物生理学]87:671-4(大豆);Finer和McMullen,(1991)In vitro Cell Dev Biol[体外细胞与发育生物学]27P:175-82(大豆);Singh等人,(1998)Theor Appl Genet[理论与应用遗传学]96:319-24(大豆);Datta等人,(1990)Biotechnology[生物技术]8:736-40(稻);Klein等人,(1988)Proc.Natl.Acad.Sci.USA[美国国家科学院院刊]85:4305-9(玉蜀黍);Klein等人,(1988)Biotechnology[生物技术]6:559-63(玉蜀黍);美国专利号5,240,855;5,322,783和5,324,646;Klein等人,(1988)Plant Physiol[植物生理学]91:440-4(玉蜀黍);Fromm等人,(1990)Biotechnology[生物技术]8:833-9(玉蜀黍);Hooykaas-Van Slogteren等人,(1984)Nature[自然]311:763-4;美国专利号5,736,369(谷类);Bytebier等人,(1987)Proc.Natl.Acad.Sci.USA[美国国家科学院院刊]84:5345-9(百合科(Liliaceae));De Wet等人,(1985)在The Experimental Manipulation of Ovule Tissues[胚珠组织的实验操作]中,编辑Chapman等人,(Longman,New York[纽约朗文出版社]),第197-209页(花粉);Kaeppler等人,(1990)Plant Cell Rep[植物细胞报告]9:415-8)以及Kaeppler等人,(1992)Theor Appl Genet[理论与应用遗传学]84:560-6(晶须介导的转化);D′Halluin等人,(1992)Plant Cell[植物细胞]4:1495-505(电穿孔);Li等人,(1993)Plant Cell Rep[植物细胞报告]12:250-5;Christou和Ford(1995)Annals Botany[植物学年鉴]75:407-13(稻)以及Osjoda等人,(1996)Nat Biotechnol自然生物技术]14:745-50(经由根癌农杆菌转化的玉蜀黍)。

[0356] 可替代地,可以通过使细胞或生物体与病毒或病毒核酸接触来将多核苷酸引入植物或植物细胞中。通常,此类方法涉及将多核苷酸掺入病毒DNA或RNA分子内。在一些实例

中，可以最初将目的多肽作为病毒多聚蛋白的一部分合成，然后将合成的多肽在体内或在体外通过蛋白水解加工从而产生所希望的重组蛋白。用于将多核苷酸引入植物，并且表达在其中编码的蛋白(涉及病毒DNA或RNA分子)的方法是已知的，参见例如，美国专利号5,889,191、5,889,190、5,866,785、5,589,367、以及5,316,931。

[0357] 可以使用多种瞬时转化方法，将多核苷酸或重组DNA构建体提供至或引入原核和真核细胞或生物体中。这种瞬时转化法包括但不限于将多核苷酸构建体直接引入植物中。

[0358] 可以通过任何方法将核酸和蛋白提供给细胞，所述方法包括使用分子来促进受指导的Cas系统(蛋白和/或核酸)的任何或所有组分(例如细胞穿透肽和纳米载剂)的摄取的方法。还参见2011年2月10日公开的US 20110035836和2015年1月7日公开的EP 2821486 A1。

[0359] 可以使用将多核苷酸引入原核和真核细胞或生物体或植物部分的其他方法，包括质体转化方法，以及用于将多核苷酸引入来自幼苗或成熟种子的组织中的方法。

[0360] “稳定转化”旨在表示经引入生物体中的核苷酸构建体合并到该生物体的基因组中，并且能够被其后代遗传。“瞬时转化”旨在表示将多核苷酸引入该生物体中并且不合并到该生物体的基因组中，或者将多肽引入生物体中。瞬时转化表明所引入的组合物仅在生物体中暂时表达或存在。

[0361] 可以使用多种方法来鉴定在靶位点处或靶位点附近具有改变的基因组的那些细胞，而不使用可筛选标志物表型。此类方法可被认为是直接分析靶序列以检测靶序列中的任何变化，包括但不限于PCR方法、测序方法、核酸酶消化、DNA印迹法、及其任何组合。

[0362] 可以将本文公开的多核苷酸和多肽引入细胞中。细胞包括但不限于人类、非人类、动物、哺乳动物、细菌、原生生物、真菌、昆虫、酵母、非常规酵母和植物细胞，以及通过本文所述的方法产生的植物和种子。在一些方面，生物体的细胞是生殖细胞、体细胞、减数分裂细胞、有丝分裂细胞、干细胞或多能干细胞。

[0363] 细胞和植物

[0364] 可以将本文公开的多核苷酸和多肽引入植物细胞中。植物细胞包括通过本文所述方法产生的植物和种子。任何植物(包括单子叶植物和双子叶植物以及植物元件)都可以与本文所述的组合物和方法一起使用。

[0365] 所公开的新颖Cas9直系同源物可以用于以各种方式编辑植物细胞的基因组。在一方面，可能需要缺失一个或多个核苷酸。在另一方面，可能期望插入一个或多个核苷酸。在一方面，可能期望替换一个或多个核苷酸。在另一方面，可能期望通过与另一原子或分子的共价或非共价相互作用来修饰一个或多个核苷酸。在一些方面，细胞是二倍体。在一些方面，细胞是单倍体。

[0366] 通过Cas9直系同源物的基因组修饰可用于在靶生物体上实现基因型和/或表型改变。这种改变优选与目的性状或农艺学上重要的特征的改善、内源缺陷的校正或某种类型的表达标志物的表达有关。在一些方面，目的性状或农艺学上重要的特征与植物的整体健康、适应性或能育性、植物产物的产量、植物的生态适应性或植物的环境稳定性有关。在一些方面，目的性状或农艺学上重要的特征选自由以下组成的组：农艺学、除草剂抗性、昆虫抗性、疾病抗性、线虫抗性、微生物抗性、真菌抗性、病毒抗性、能育性或不育性、籽粒特征，商业产物产生。在一些方面，目的性状或农艺学上重要的特征选自由以下组成的组：如与不

包含衍生自本文方法和组合物的修饰的同系植物相比的疾病抗性、干旱抗性、热耐性、寒耐性、盐耐性、金属耐性、除草剂耐性、改善的水分利用效率、改善的氮利用率、改善的固氮作用、有害生物抗性、食草动物抗性、病原体抗性、产率改善、健康增强、活力改善、生长改善、光合能力改善、营养增强、改变的蛋白含量、改变的淀粉含量、改变的碳水化合物含量、改变的糖含量、改变的纤维含量、改变的油含量、增加的生物量、增加的芽长度、增加的根长度、改善的根结构、代谢产物的调节、蛋白质组的调节、增加的种子重量、改变的种子碳水化合物组成、改变的种子油组成、改变的种子蛋白组成、改变的种子营养物组成。

[0367] 可以使用的单子叶植物的实例包括但不限于,玉蜀黍(玉蜀黍(Zea mays))、稻(水稻(Oryza sativa))、黑麦(黑麦(Secale cereale))、高粱(双色高粱(Sorghum bicolor)、高粱(Sorghum vulgare))、粟(例如,珍珠粟、御谷(Pennisetum glaucum))、黍稷(粟米(Panicum miliaceum))、谷子(谷子(Setaria italica))、穆子(龙爪稷(Eleusine coracana))、小麦(小麦属物种,例如小麦(Triticum aestivum)、一粒小麦(Triticum monococcum))、甘蔗(甘蔗属物种(Saccharum spp.))、燕麦(燕麦属(Avena))、大麦(大麦属(Hordeum))、柳枝稷(柳枝黍(Panicum virgatum))、菠萝(菠萝(Ananas comosus))、香蕉(香蕉属物种(Musa spp.))、棕榈、观赏植物、草坪草、以及其他草。

[0368] 可以使用的双子叶植物的实例包括但不限于大豆(大豆(Glycine max))、芸苔属物种(例如但不限于:油菜或卡诺拉油菜)(欧洲油菜(Brassica napus)和白菜型油菜(B.campestris)、芜菁(Brassica rapa)、芥菜(Brassica.juncea))、苜蓿(紫花苜蓿(Medicago sativa)、烟草(烟草(Nicotiana tabacum))、拟南芥属(Arabidopsis)(拟南芥(A.thaliana))、向日葵(向日葵(Helianthus annuus))、棉花(木本棉(Gossypium arboreum)、海岛棉(Gossypium barbadense))、和花生(花生(Arachis hypogaea))、番茄(番茄(Solanum lycopersicum))、马铃薯(马铃薯(Solanum tuberosum))等。

[0369] 可以使用的另外的植物包括红花(safflower、Carthamus tinctorius)、甘薯(番薯(Ipomoea batatas)),木薯(cassava,Manihot esculenta),咖啡(咖啡属物种(Coffea spp.)),椰子(coconut,Cocos nucifera),柑橘树(柑橘属物种(Citrus spp.)),可可(cocoa,Theobroma cacao),茶树(tea,Camellia sinensis),香蕉(芭蕉属物种(Musa spp.)),鳄梨(avocado,Persea americana),无花果(fig或(Ficus casica)),番石榴(guava,Psidium guajava),芒果(mango,Mangifera indica),橄榄(olive,Olea europaea),木瓜(番木瓜(Carica papaya)),腰果(cashew,Anacardium occidentale),澳洲坚果(macadamia,Macadamia integrifolia),巴旦杏(almond,Prunus amygdalus),甜菜(sugar beets,Beta vulgaris),蔬菜,观赏植物和针叶树。

[0370] 可以使用的蔬菜包括番茄(Lycopersicon esculentum)、莴苣(例如,莴苣(Lactuca sativa))、青豆(菜豆(Phaseolus vulgaris))、利马豆(lima bean,Phaseolus limensis)、豌豆(香豌豆属物种(Lathyrus spp.))和黄瓜属的成员诸如黄瓜(cucumber,C.sativus)、香瓜(cantaloupe,C.cantalupensis)和甜瓜(musk melon,C.melo)。观赏植物包括杜鹃(杜鹃花属物种(Rhododendron spp.))、八仙花(Macrophylla hydrangea)、朱槿(Hibiscus rosasanensis)、玫瑰(蔷薇属物种(Rosa spp.))、郁金香(郁金香属物种(Tulipa spp.))、水仙(水仙属物种(Narcissus spp.))、矮牵牛(Petunia hybrida)、康乃馨(Dianthus caryophyllus)、一品红(Euphorbia pulcherrima)和菊花。

[0371] 可以使用的针叶树包括松树,如火炬松(loblolly pine,Pinus taeda)、湿地松(slash pine,Pinus elliotii)、西黄松(ponderosa pine,Pinus ponderosa)、黑松(lodgepole pine,Pinus contorta)和辐射松(Monterey pine,Pinus radiata);花旗松(Douglasfir,Pseudotsuga menziesii);西方铁杉(Western hemlock,Tsuga canadensis);北美云杉(Sitka spruce,Picea glauca);红杉(redwood,Sequoia sempervirens);枞树(true firs),如银杉(胶冷杉(Abies amabilis))和胶枞(香脂冷杉(Abies balsamea));以及雪松,如西方红雪松(Thuja plicata)和阿拉斯加黄雪松(Chamaecyparis nootkatensis)。

[0372] 在本公开的某些实施例中,可育植物是产生活雄配子和雌配子并且是自身可育的植物。这样的自体受精的植物可以产生后代植物,而没有来自任何其他植物的配子及其中所含的遗传物质的贡献。本公开的其他实施例可以涉及使用非自身可育的植物,因为该植物不产生有活力的或在其他情况下能够受精的雄配子或雌配子或二者。

[0373] 本公开可用于包含一个或多个引入性状或经编辑的基因组的植物的育种。

[0374] 如下描述两个性状如何以彼此之间例如5cM的遗传距离堆叠到基因组中的非限制性实例:将包含整合到基因组窗口内的第一DSB靶位点中且不具有第一目的基因组基因座的第一转基因靶位点的第一植物与第二转基因植物杂交,所述第二转基因植物在基因组窗口内的不同基因组插入位点处包含目的基因组基因座,并且所述第二植物不包含所述第一转基因靶位点。来自该杂交的约5%的植物后代将基因组窗口内具有整合到第一DSB靶位点中的第一转基因靶位点和整合在不同基因组插入位点处的第一目的基因组基因座。在定义的基因组窗口中具有两个位点的后代植物可以进一步与第三转基因植物杂交,所述第三转基因植物在定义的基因组窗口内包含整合到第二DSB靶位点中的第二转基因靶位点、和/或第二目的基因组基因座并且缺乏所述第一转基因靶位点和所述第一目的基因组基因座。然后选择具有在基因组窗口内的不同基因组插入位点处整合的第一转基因靶位点、第一目标基因组基因座和第二目的基因组基因座的后代。这样的方法可用于产生包含复杂性状基因座的植物,所述复杂性状基因座具有至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、19、19、20、21、22、23、24、25、26、27、28、29、30、31或更多个整合到DSB靶位点中的转基因靶位点和/或整合在基因组窗口内的不同位点的目的基因组基因座。以这种方式,可以产生各种复杂性状基因座。

[0375] 细胞与动物

[0376] 可以将本文公开的多核苷酸和多肽引入动物细胞中。动物细胞可以包括但不限于:以下门的生物体,所述门包括脊索动物门、节肢动物门、软体动物门、环节动物门、腔肠动物门或棘皮动物门;以下纲的生物体,所述纲包括哺乳动物、昆虫、鸟、两栖动物、爬行动物或鱼。在一些方面,所述动物是人类、小鼠、秀丽隐杆线虫(C.elegans)、大鼠、果蝇(果蝇属物种(Drosophila spp.))、斑马鱼、鸡、狗、猫、豚鼠、仓鼠、鸡、日本稻鱼、海七鳃鳗、河豚、树蛙(例如非洲爪蟾属物种(Xenopus spp.))、猴或黑猩猩。预期的特定细胞类型包括单倍体细胞、二倍体细胞、生殖细胞、神经元、肌肉细胞、内分泌或外分泌细胞、上皮细胞、肌肉细胞、肿瘤细胞、胚胎细胞、造血细胞、骨细胞、种质细胞、体细胞、干细胞、多能干细胞、诱导多能干细胞、祖细胞、减数分裂细胞和有丝分裂细胞。在一些方面,可以使用来自生物体的多个细胞。

[0377] 所公开的新颖的Cas9直系同源物可以用于以各种方式编辑动物细胞的基因组。在一方面，可能需要缺失一个或多个核苷酸。在另一方面，可能期望插入一个或多个核苷酸。在一方面，可能期望替换一个或多个核苷酸。在另一方面，可能期望通过与另一原子或分子的共价或非共价相互作用来修饰一个或多个核苷酸。

[0378] 通过Cas9直系同源物的基因组修饰可用于在靶生物体上实现基因型和/或表型改变。这种改变优选与目的表型或生理学上重要的特征的改善、内源缺陷的校正或某种类型的表达标志物的表达有关。在一些方面，目的表型或生理学上重要的特征与以下有关：动物的整体健康、适应性或能育性、动物的生态适应性或动物与环境中其他生物体的关系或相互作用。在一些方面，有意义的表型或生理学上重要的特征选自由以下组成的组：改善的总体健康、疾病逆转、疾病修饰、疾病稳定、疾病预防、寄生虫感染的治疗、病毒感染的治疗、逆转录病毒感染的治疗、细菌感染的治疗、神经障碍(例如但不限于：多发性硬化)的治疗、内源遗传缺陷(例如但不限于：代谢障碍、软骨病、α-1抗胰蛋白酶缺乏症、抗磷脂综合征、自闭症、常染色体显性多囊肾病、巴斯综合症(Barth syndrome)、乳腺癌、夏科-马里-图思病(Charcot-Marie-Tooth)、结肠癌、猫叫综合征(Cri du chat)、克罗恩病、囊性纤维化、痛性脂肪病(Dercum Disease)、唐氏综合征(Down Syndrome)、杜安氏综合征(Duane Syndrome)、杜兴氏肌营养不良(Duchenne Muscular Dystrophy)、V因子莱顿易栓症(Factor V Leiden Thrombophilia)、家族性高胆固醇血症、家族性地中海热、脆性X综合征、戈谢病(Gaucher Disease)、血色素沉着症、血友病、前脑无裂畸形、亨廷顿病、克兰费尔特综合征(Klinefelter syndrome)、马凡综合征(Marfan syndrome)、肌强直性营养不良、神经纤维瘤病、努南综合征(Noonan Syndrome)、成骨不全症、帕金森病、苯酮尿症、波兰得异常(Poland Anomaly)、卟啉症、早衰症、前列腺癌、视网膜色素变性、严重合并免疫缺陷(SCID)、镰状细胞病、皮肤癌、脊髓性肌萎缩症、黑朦性痴呆(Tay-Sachs)、地中海贫血、三甲基胺尿症、特纳综合征(Turner Syndrome)、腭心面综合征(Velocardiofacial Syndrome)、WAGR综合征和威尔逊病(Wilson Disease))的校正、先天性免疫障碍(例如但不限于：免疫球蛋白亚类缺陷)的治疗、获得性免疫障碍(例如但不限于：AIDS和其他与HIV相关的障碍)的治疗、癌症的治疗以及包括罕见或“孤儿”病症在内的疾病的治疗，这些通过其他方法无法找到有效的治疗选择。

[0379] 使用本文公开的组合物或方法进行了遗传修饰的细胞可以出于诸如基因疗法等目的移植到受试者，例如用于治疗疾病或作为抗病毒、抗病原体或抗癌治疗剂，用于农业中生产遗传修饰的生物体或用于生物学研究。

[0380] 体外多核苷酸检测、结合和修饰

[0381] 在一些方面，本文公开的组合物可以进一步用作用于(在一些方面与一种或多种分离的多核苷酸序列一起)体外方法的组合物。所述一种或多种分离的多核苷酸序列可以包含一种或多种用于修饰的靶序列。在一些方面，所述一种或多种分离的多核苷酸序列可以是基因组DNA、PCR产物或合成的寡核苷酸。

[0382] 组合物

[0383] 靶序列的修饰可以是以下形式：核苷酸插入、核苷酸缺失、核苷酸取代、向现有核苷酸添加原子分子、核苷酸修饰或异源多核苷酸或多肽与所述靶序列的结合。一个或多个核苷酸的插入可通过在反应混合物中包含供体多核苷酸来完成：将所述供体多核苷酸插入

由所述Cas9直系同源物多肽产生的双链断裂中。插入可以经由非同源末端连接或经由同源重组。

[0384] 在一方面，靶多核苷酸的序列在修饰之前是已知的，并且与由Cas9直系同源物处理产生的一种或多种多核苷酸的一种或多种序列进行比较。在一方面，靶多核苷酸的序列在修饰之前是未知的，并且Cas9直系同源物处理被用作确定所述靶多核苷酸的序列的方法的一部分。

[0385] 用Cas9直系同源物进行的多核苷酸修饰可通过使用从Cas基因座鉴定的全长多肽，或从Cas基因座鉴定的多肽的片段、修饰或变体完成。在一些方面，所述Cas9直系同源物获自或衍生自表1中所列的生物体。在一些方面，所述Cas9直系同源物是与SEQ ID NO:86-170或511-1135中的任一个具有至少80％同一性的多肽。在一些方面，所述Cas9直系同源物是SEQ ID NO:86-170或511-1135中任一个的功能性变体。在一些方面，所述Cas9直系同源物是SEQ ID NO:86-170或511-1135中任一个的功能性片段。在一些方面，所述Cas9直系同源物是由选自由以下组成的组的多核苷酸编码的Cas9多肽:SEQ ID NO:86-170或511-1135。在一些方面，所述Cas9直系同源物是识别表4-83中任一个所列的PAM序列的Cas9多肽。在一些方面，所述Cas9直系同源物是从序列表中所列生物体中鉴定的Cas9多肽。

[0386] 在一些方面，Cas9直系同源物作为cas9多核苷酸提供。在一些方面，所述cas9多核苷酸选自由以下组成的组:SEQ ID NO:1-85，或是与SEQ ID NO:1-85中的任何一个具有至少80％、85％、90％、95％、97％、为99％或100％的序列。

[0387] 在一些方面，Cas9直系同源物可以选自由以下组成的组:未经修饰的野生型Cas9直系同源物;功能性Cas9直系同源物变体;功能性Cas9直系同源物片段;包含活性或失活的Cas9直系同源物的融合蛋白;Cas9直系同源物，其在C末端上或在N末端上或在N和C末端两者上进一步包含一个或多个核定位序列(NLS);生物素化的Cas9直系同源物;Cas9直系同源物切口酶;Cas9直系同源物内切核酸酶;进一步包含组氨酸标签的Cas9直系同源物;和上述任何两者或更多的混合物。

[0388] 在一些方面，Cas9直系同源物是融合蛋白，其进一步包含核酸酶结构域、转录激活子结构域、转录阻遏子结构域、表观遗传修饰结构域、切割结构域、核定位信号、细胞穿透结构域、易位结构域、标志物、或与靶多核苷酸序列或从其获得或衍生出所述靶多核苷酸序列的细胞异源的转基因。

[0389] 在一些方面，期望多个Cas9直系同源物。在一些方面，所述多个可以包含衍生自不同来源生物体或衍生自相同生物内的不同基因座的Cas9直系同源物。在一些方面，所述多个可以包含对靶多核苷酸具有不同结合特异性的Cas9直系同源物。在一些方面，所述多个可以包含具有不同切割效率的Cas9直系同源物。在一些方面，所述多个可以包含具有不同PAM特异性的Cas9直系同源物。在一些方面，所述多个可以包含具有不同分子组成(即多核苷酸cas9直系同源物和多肽Cas9直系同源物)的直系同源物。

[0390] 指导多核苷酸可以提供为单指导RNA(sgRNA)、包含tracrRNA的嵌合分子、包含crRNA的嵌合分子、嵌合RNA-DNA分子、DNA分子或包含一个或多个化学修饰的核苷酸的多核苷酸。

[0391] Cas9直系同源物和/或指导多核苷酸的储存条件包括温度、物质状态和时间的参数。在一些方面，Cas9直系同源物和/或指导多核苷酸在约-80摄氏度、约-20摄氏度、约4摄

氏度、约20-25摄氏度或约37摄氏度下储存。在一些方面，Cas9直系同源物和/或指导多核苷酸以液体、冷冻液体或冻干粉的形式存储。在一些方面，Cas9直系同源物和/或指导多核苷酸稳定至少一天、至少一周、至少一个月、至少一年或甚至大于一年。

[0392] 反应的任何或所有可能的多核苷酸组分(例如，指导多核苷酸，供体多核苷酸，任选地cas9多核苷酸)可以提供为载体、构建体、线性化或环化质粒的一部分或作为嵌合分子的一部分。每种组分可以单独或一起提供给反应混合物。在一些方面，一种或多种多核苷酸组分可操作地连接至调节其表达的异源非编码调节元件。

[0393] 用于修饰靶多核苷酸的方法包括将最少的元件组合到反应混合物中，所述反应混合物包含：Cas9直系同源物(或如上所述的变体、片段或其他相关分子)、指导多核苷酸(其包含与靶多核苷酸的靶多核苷酸序列基本互补或选择性杂交的序列)、以及用于修饰的靶多核苷酸。在一些方面，Cas9直系同源物作为多肽提供。在一些方面，Cas9直系同源物作为cas9直系同源物多核苷酸提供。在一些方面，所述指导多核苷酸提供为RNA分子、DNA分子、RNA:DNA杂合体或包含化学修饰的核苷酸的多核苷酸分子。

[0394] 可以针对稳定性、功效或其他参数优化组分中任何一种的储存缓冲液、或反应混合物。储存缓冲液或反应混合物的另外的组分可包括缓冲液组合物、Tris、EDTA、二硫苏糖醇(DTT)、磷酸盐缓冲盐水(PBS)、氯化钠、氯化镁、HEPES、甘油、BSA、盐、乳化剂、洗涤剂、螯合剂、氧化还原剂、抗体、无核酸酶的水、蛋白酶和/或粘度剂。在一些方面，所述储存缓冲液或反应混合物还包含具有以下组分中的至少一种的缓冲溶液：HEPES、MgCl2、NaCl、EDTA、蛋白酶、蛋白酶K、甘油、无核酸酶的水。

[0395] 孵育条件将根据所期望的结果而变化。温度优选为至少10摄氏度、10至15、至少15、15至17、至少17、17至20、至少20、20至22、至少22、22至25、至少25、25至27、至少27、27至30、至少30、30至32、至少32、32至35、至少35、至少36、至少37、至少38、至少39、至少40或甚至大于40摄氏度。孵育时间为至少1分钟、至少2分钟、至少3分钟、至少4分钟、至少5分钟、至少6分钟、至少7分钟、至少8分钟、至少9分钟、至少10分钟、或甚至大于10分钟。

[0396] 孵育之前、期间或之后，反应混合物中一种或多种多核苷酸的一种或多种序列可以通过本领域已知的任何方法来确定。一方面，可以通过在与Cas9直系同源物结合之前，将从反应混合物中纯化的一种或多种多核苷酸的一种或多种序列与靶多核苷酸的序列进行比较来确定靶多核苷酸的修饰。

[0397] 试剂盒中可包含可用于体外或体内多核苷酸检测、结合和/或修饰的本文公开的组合物中的任何一种或多种。试剂盒包含Cas9直系同源物或编码这样的Cas9直系同源物或多核苷酸cas9直系同源物，以及任选地进一步包含能够有效储存的缓冲液组分，以及一种或多种另外的组合物，所述一种或多种另外的组合物能够将所述Cas9直系同源物或cas9直系同源物引入异源多核苷酸，其中所述Cas9直系同源物或cas9直系同源物能够实现对所述异源多核苷酸的至少一个核苷酸的修饰、添加、缺失或取代。在另一方面，本文公开的Cas9直系同源物可用于从混合池富集一种或多种多核苷酸靶序列。在另一方面，可以将本文公开的Cas9直系同源物固定在基质上，以用于体外靶多核苷酸检测、结合和/或修饰。

[0398] 检测方法

[0399] 检测与靶多核苷酸结合的Cas9:指导多核苷酸复合物的方法可以包括本领域中任何已知的方法，包括但不限于显微镜检查、色谱分离、电泳、免疫沉淀、过滤、纳米孔分离、微

阵列以及下文所述的那些。

[0400] DNA电泳迁移率变动分析(EMSA):研究与已知DNA寡核苷酸探针结合的蛋白,并评估相互作用的特异性。所述技术基于以下原理:当进行聚丙烯酰胺或琼脂糖凝胶电泳时,蛋白-DNA复合物的迁移速度比游离DNA分子慢。由于DNA迁移的速度在蛋白结合后被阻滞,因此所述测定也称为凝胶阻滞测定。将蛋白特异性抗体添加到结合组分中会产生更大的复合物(抗体-蛋白-DNA),所述复合物在电泳过程中迁移甚至更慢,这被称为超变动并且可用于确认蛋白身份。

[0401] DNA下拉测定使用标记有高亲和力标签(例如生物素)的DNA探针,所述标签允许回收或固定探针。可以将DNA探针与来自EMSA中使用的类似的反应中细胞裂解物的蛋白复合并且然后用于使用琼脂糖或磁珠进行纯化。然后从DNA洗脱蛋白,并通过蛋白印迹检测或通过质谱鉴定。可替代地,可以用亲和标签标记蛋白,或者可以使用针对目的蛋白的抗体分离DNA-蛋白复合物(类似于超变动测定)。在这种情况下,通过DNA印迹或PCR分析检测与蛋白结合的未知DNA序列。

[0402] 报告子测定提供目的启动子翻译活性的实时体内读出。报告基因是靶启动子DNA序列和报告基因DNA序列(所述报告基因DNA序列由研究者定制并且编码具有可检测特性的蛋白,例如萤火虫/雷尼利亚萤光素酶或碱性磷酸酶)的融合体。这些基因仅在目的启动子被激活时才产生酶。酶继而催化底物以产生可以通过光谱仪器检测到的光或颜色变化。来自报告基因的信号用作对于由同一启动子驱动的内源蛋白的翻译而言的间接决定因素。

[0403] 微孔板捕获和检测测定使用固定化的DNA探针来捕获特异性蛋白-DNA相互作用,并确认蛋白身份和与靶特异性抗体的相对含量。通常,DNA探针固定在包被链霉亲和素的96或384孔微孔板的表面上。制备并添加细胞提取物以使结合蛋白结合至寡核苷酸。然后去除提取物,并且每个孔洗涤几次以去除非特异性结合的蛋白。最后,使用经标记用于检测的特异性抗体检测蛋白。该方法非常灵敏,能检测低于0.2pg靶蛋白/孔。该方法也可用于标记有其他标签(例如可以固定在包被胺反应性表面化学物质的微板上的伯胺)的寡核苷酸。

[0404] DNA足迹法是获得有关蛋白-DNA复合物中各个核苷酸甚至是活细胞内部详细信息的最广泛使用的方法之一。在这样的实验中,使用化学药品或酶来修饰或消化DNA分子。当序列特异性蛋白与DNA结合时,它们可以保护结合位点不被修饰或消化。这随后可以通过变性凝胶电泳来可视化,其中未保护的DNA或多或少地被随机切割。因此,它表现为条带的“阶梯”,并且受蛋白保护的位点没有相应的条带,并看起来像条带图案中的足迹。通过在蛋白-DNA结合位点鉴定出特定的核苷,在这里留下足迹。

[0405] 显微镜技术包括光学、荧光、电子和原子力显微镜(AFM)。

[0406] 染色质免疫沉淀分析(ChIP)使蛋白与它们的DNA靶共价结合,然后将它们解连接并分别表征。

[0407] 通过指数富集(SELEX)进行配体的系统进化将靶蛋白暴露于寡核苷酸的随机文库。那些结合的基因通过PCR分离和扩增。

[0408] 非限制性方面

[0409] 方面1:一种合成的组合物,其包含选自下组的cas9多核苷酸,该组由以下组成:(a)与SEQ ID NO:86-170或511-1135中的任何一个具有至少80%同一性的多核苷酸,(b)SEQ ID NO:86-170或511-1135中的任何一个的功能性变体,(c)SEQ ID NO:86-170或511-

1135中的任何一个的功能性片段，(d)编码Cas9多肽的cas9基因，所述Cas9多肽选自由以下组成的组：SEQ ID NO:86-170，(e)编码Cas9多肽的cas9基因，所述Cas9多肽识别表4-83中的任何一个列出的PAM序列，和(f)从表1中列出的生物体鉴定的cas9基因；和异源组分。

[0410] 方面2：一种合成的组合物，其包含选自下组的Cas9多肽，该组由以下组成：(a)与SEQ ID NO:86-170或511-1135中的任何一个具有至少80％同一性的多肽，(b)SEQ ID NO:86-170或511-1135中的任何一个的功能性变体，(c)SEQ ID NO:86-170的任何一个的功能性片段，(d)由多核苷酸编码的选自由以下组成的组的Cas9多肽：SEQ ID NO:86-170或511-1135，(e)识别表4-83中的任何一个列出的PAM序列的Cas9多肽，和(f)从表1中或序列表中列出的生物体鉴定的Cas9多肽；和异源组分。

[0411] 方面3：一种失活的Cas9多肽，其中所述失活的Cas9多肽能够与靶多核苷酸结合，但是缺乏至少一个负责核苷酸切割的结构域。

[0412] 方面4：一种包含Cas9多肽和异源多肽的合成的融合蛋白，其中所述Cas9多肽选自由以下组成的组：

[0413] 方面5：一种包含单指导RNA的合成的组合物，所述单指导RNA选自由以下组成的组：(a)与SEQ ID NO:426-510中的任何一个具有至少80％同一性的多核苷酸：(b)SEQ ID NO:426-510中的任何一个的功能性变体，(c)SEQ ID NO:426-510中的任何一个的功能性片段，和(d)从表1中列出的生物体鉴定或衍生的单指导RNA分子；和异源组分。

[0414] 方面6：一种包含tracrRNA的合成的组合物，所述tracrRNA选自由以下组成的组：(a)与SEQ ID NO:341-425中的任何一个具有至少80％同一性的多核苷酸：(b)SEQ ID NO:341-425中的任何一个的功能性变体，(c)SEQ ID NO:341-425中的任何一个的功能性片段，和(d)从表1中列出的生物体鉴定的tracrRNA分子；和异源组分。

[0415] 方面7：一种包含crRNA重复序列的合成的组合物，所述crRNA重复序列选自由以下组成的组：(a)与SEQ ID NO:171-255中的任何一个具有至少80％同一性的多核苷酸：(b)SEQ ID NO:171-255中的任何一个的功能性变体，(c)SEQ ID NO:171-255中的任何一个的功能性片段，和(d)从表1中列出的生物体鉴定的crRNA重复序列分子；和异源组分。

[0416] 方面8：一种包含反重复序列的合成的组合物，所述反重复序列选自由以下组成的组：(a)与SEQ ID NO:256-340中的任何一个具有至少80％同一性的多核苷酸：(b)SEQ ID NO:256-340中的任何一个的功能性变体，(c)SEQ ID NO:256-340中的任何一个的功能性片段，和和(d)从表1中列出的生物体鉴定的反重复序列分子；和异源组分。

[0417] 方面9：一种合成的组合物，所述合成的组合物包含与由SEQ ID NO:86-170组成的组的多肽具有至少80％同一性的多肽，以及选自由以下组成的组的多核苷酸：(a)与选自由SEQ ID NO:171-255组成的组的多核苷酸具有至少80％同一性的多核苷酸，(b)与选自由SEQ ID NO:341-425组成的组的多核苷酸具有至少80％同一性的多核苷酸，和(c)与选自由SEQ ID NO:426-510组成的组的多核苷酸具有至少80％同一性的多核苷酸；其中所述合成的组合物进一步包含异源组分。

[0418] 方面10：一种包含指导多核苷酸和Cas9直系同源物的合成的组合物，其中所述Cas9直系同源物选自由以下组成的组：(a)如方面3所述的失活的Cas9多肽，(b)与以下SEQ ID NO:86-170或511-1135中的任何一个具有至少80％同一性的多肽，(c)SEQ ID NO:86-170或511-1135的任何一个的功能性变体，(d)SEQ ID NO:86-170或511-1135中任何一个的

功能性片段，(e)识别表4-83中的任何一个列出的PAM序列的Cas9多肽，(f)从表1中列出的生物体鉴定的Cas9多肽，(g)选自由SEQ ID NO:86-170或511-1135组成的组的cas9多核苷酸，以及(h)编码(a)至(f)的任何多肽的cas9多核苷酸；并且所述指导多核苷酸选自由以下组成的组：(i)与选自由SEQ ID NO:426-510组成的组的序列具有至少80％同一性的单指导RNA，(j)包含SEQ ID NO:426-510的功能性片段的单指导RNA，(k)包含SEQ ID NO:426-510的功能性变体的单指导RNA，(l)包含与tracrRNA连接的嵌合非天然存在的crRNA的单指导RNA，其中所述tracrRNA包含选自下组的核苷酸序列，该组由以下组成:SEQ ID NO:341-425,SEQ ID NO:341-425的功能性片段，和SEQ ID NO:341-425的功能性变体，(m)单导RNA包含与tracrRNA连接的嵌合非天然存在的crRNA，其中所述嵌合非天然存在的crRNA包含选自下组的核苷酸序列，该组由以下组成:SEQ ID NO:171-255,SEQ ID NO:171-255的功能性片段，和SEQ ID NO:171-255的功能性变体，(n)指导RNA，其是包含嵌合非天然存在的crRNA和tracrRNA的双链体分子，其中所述嵌合非天然存在的crRNA包含能够与所述靶序列杂交的可变靶向结构域，其中所述tracrRNA包含选自下组的核苷酸序列，该组由以下组成:SEQ ID NO:341-425,SEQ ID NO:341-425的功能性片段，和SEQ ID NO:341-425的功能性变体，其中所述嵌合非天然存在的crRNA包含能够与所述靶序列杂交的可变靶向结构域，(o)指导RNA，其是包含嵌合非天然存在的crRNA和tracrRNA的双链体分子，其中所述嵌合非天然存在的crRNA包含选自下组的核苷酸序列，该组由以下组成:SEQ ID NO:171-255,SEQ ID NO:171-255的功能性片段，和SEQ ID NO:171-255的功能性变体，其中所述嵌合非天然存在的crRNA包含能够与所述靶序列杂交的可变靶向结构域，(p)包含DNA和RNA两者的多核苷酸，(q)包含至少一个化学修饰的核苷酸的多核苷酸，和(r)编码(h)至(n)的任何RNA分子的DNA分子；其中所述指导多核苷酸和所述Cas9直系同源物能够形成复合物，所述复合物能够识别、结合靶多核苷酸序列并任选地使靶多核苷酸序列产生切口或切割靶多核苷酸序列；进一步包含至少一种异源组分。

[0419] 方面11:如方面10所述的指导多核苷酸/Cas9内切核酸酶复合物，其中所述靶多核苷酸序列位于细胞的基因组中。

[0420] 方面12:如方面10所述的指导多核苷酸/Cas9内切核酸酶复合物，其中所述靶多核苷酸序列是从基因组环境中分离的。

[0421] 方面13:如方面10所述的指导多核苷酸/Cas9内切核酸酶复合物，其中所述靶多核苷酸序列是合成的。

[0422] 方面14:如方面1-10中任一项所述的合成的组合物，其中所述异源组分选自由以下组成的组:异源多核苷酸、异源多肽、粒子、固体基质、抗体、缓冲液组合物、Tris、EDTA、二硫苏糖醇(DTT)、磷酸盐缓冲盐水(PBS)、氯化钠、氯化镁、HEPES、甘油、牛血清白蛋白(BSA)、盐、乳化剂、洗涤剂、螯合剂、氧化还原剂、抗体、无核酸酶的水、粘度剂和组氨酸标签。

[0423] 方面15:如方面14所述的合成的组合物，其中所述异源多肽包含核酸酶结构域、转录激活子结构域、转录阻遏子结构域、表观遗传修饰结构域、切割结构域、核定位信号、细胞穿透性结构域、脱氨酶结构域、碱基编辑结构域、易位结构域、标志物和转基因。

[0424] 方面16:方面14的合成的组合物，其中所述异源多核苷酸选自:指导多核苷酸、嵌合指导多核苷酸、化学修饰的指导多核苷酸、同时DNA和RNA两者的指导多核苷酸、非编码表达元件、基因、标志物和编码多个组氨酸残基的多核苷酸。

[0425] 方面17:如方面14所述的合成的组合物,其包含至少两种不同的所述异源组分。

[0426] 方面18:如方面14所述的合成的组合物,其中pH为1.0至14.0、2.0至13.0、3.0至12.0、4.0至11.0、5.0至10.0、6.0至9.0、7.0至8.0、4.5至6.5、5.5至7.5、或6.5至7.5。

[0427] 方面19:如方面14所述的合成的组合物,其中所述Cas9直系同源物在以下pH具有最佳活性:1.0至14.0、2.0至13.0、3.0至12.0、4.0至11.0、5.0至10.0、6.0至9.0、7.0至8.0、4.5至6.5、5.5至7.5、或6.5至7.5。

[0428] 方面20:如方面14所述的合成的组合物,其中所述Cas9直系同源物在以下温度具有最佳活性:0摄氏度至100摄氏度、至少0摄氏度至10摄氏度、至少10摄氏度至20摄氏度、至少20摄氏度至25摄氏度、至少25摄氏度至30摄氏度、至少30摄氏度至40摄氏度、至少40摄氏度至50摄氏度、至少50摄氏度至60摄氏度、至少60摄氏度至70摄氏度、至少70摄氏度至80摄氏度、至少80摄氏度至90摄氏度、至少90摄氏度至100摄氏度、或100摄氏度。

[0429] 方面21:如方面14所述的合成的组合物,其在以下温度储存或孵育:至少负200摄氏度、至少负150摄氏度、至少负135摄氏度、至少负90摄氏度、至少负80摄氏度、至少负20摄氏度、至少4摄氏度、至少17摄氏度、至少25摄氏度、至少30摄氏度、至少35摄氏度、至少37摄氏度、至少39摄氏度、或大于39摄氏度。

[0430] 方面22:一种基本上无核酸酶、无内毒素的组合物,所述组合物包含如方面1-10中任一项所述的合成的组合物。

[0431] 方面23:一种冻干组合物,其包含如方面10或方面15所述的合成的组合物。

[0432] 方面24:一种细胞,其包含如方面1-10中任一项所述的合成的组合物。

[0433] 方面25:一种如方面23所述的细胞的后代细胞,其中与亲本细胞的靶多核苷酸位点相比,所述后代细胞包含其基因组的至少一个修饰。

[0434] 方面26:如方面24所述的细胞,所述细胞选自由以下组成的组:人、非人灵长类、哺乳动物、动物、古细菌、细菌、原生生物、真菌、昆虫、酵母、非常规酵母和植物。

[0435] 方面27:如方面26所述的人细胞,其中所述人细胞选自由以下组成的组:单倍体细胞、二倍体细胞、生殖细胞、神经元、肌肉细胞、内分泌或外分泌细胞、上皮细胞、肌肉细胞、肿瘤细胞、胚胎细胞、造血细胞、骨细胞、种质细胞、体细胞、干细胞、多能干细胞、诱导多能干细胞、祖细胞、减数分裂细胞和有丝分裂细胞。

[0436] 方面28:如方面26所述的植物细胞,其中所述植物细胞选自由以下组成的组:单子叶植物和双子叶植物的细胞。

[0437] 方面29:如方面26所述的植物细胞,其中所述植物细胞选自由以下组成的组:玉蜀黍、稻、高粱、黑麦、大麦、小麦、粟、燕麦、甘蔗、草坪草、柳枝稷、大豆、卡诺拉油菜、苜蓿、向日葵、棉花、烟草、花生、马铃薯、烟草、拟南芥属、蔬菜和红花细胞。

[0438] 方面30:如方面2所述的合成的组合物,其中所述Cas9内切核酸酶已被修饰为缺乏至少一个核酸酶结构域。

[0439] 方面31:如方面2所述的合成的组合物,其中所述Cas9内切核酸酶已被修饰为缺乏内切核酸酶活性。

[0440] 方面32:一种试剂盒,其包含如方面23所述的冻干组合物或如方面22所述的合成的组合物。

[0441] 方面33:一种检测靶多核苷酸序列的体外方法,所述方法包括:(a)获得所述靶多

核苷酸，(b)在反应容器中组合Cas9直系同源物多肽、指导多核苷酸和所述靶多核苷酸，(c)在至少10摄氏度的温度下孵育步骤(b)的组分至少1分钟，(d)对反应混合物中的所得的一种或多种多核苷酸进行测序，并且(e)表征由所述Cas9直系同源物多肽和所述指导多核苷酸鉴定的步骤(a)的靶多核苷酸的序列；其中所述指导多核苷酸包含与所述靶多核苷酸的序列基本互补的多核苷酸序列。

[0442] 方面34：一种将Cas9直系同源物和指导多核苷酸复合物结合至靶多核苷酸的体外方法，所述方法包括：(a)获得所述靶多核苷酸的序列，(b)在反应容器中组合Cas9直系同源物多肽、指导多核苷酸和所述靶多核苷酸，(c)在至少10摄氏度的温度下孵育步骤(b)的组分至少1分钟；其中所述指导多核苷酸包含与所述靶多核苷酸的靶多核苷酸序列基本互补的多核苷酸序列；进一步包括检测与所述靶多核苷酸结合的所述Cas9直系同源物和指导多核苷酸复合物。

[0443] 方面35：如方面34所述的方法，其中所述Cas9直系同源物进一步包含可检测的融合蛋白结构域、组氨酸标签或化学标志物。

[0444] 方面36：如方面34的方法，其中检测与所述靶多核苷酸结合的所述Cas9直系同源物和引导多核苷酸复合物进一步包括以下步骤，所述步骤包括酶联免疫吸附测定、放射免疫测定、亲和色谱、尺寸排阻色谱、离子交换色谱、疏水相互作用色谱、电泳迁移率变动测定、染色质免疫沉淀测定、酵母单杂交系统、细菌单杂交系统、X射线晶体学、下拉测定、报告子测定、标志物表达测定、微孔板捕获测定和DNA足迹。

[0445] 方面37：一种修饰靶多核苷酸的体外方法，所述方法包括：(a)获得所述靶多核苷酸的序列，(b)在反应容器中组合Cas9直系同源物多肽、指导多核苷酸和所述靶多核苷酸，(c)在至少10摄氏度的温度下孵育步骤(b)的组分至少1分钟，(d)对反应混合物中的所得的一种或多种多核苷酸进行测序，并且(e)与步骤(a)中获得的靶多核苷酸的序列相比，鉴定所述所得的一种或多种多核苷酸的至少一个序列修饰；其中所述指导多核苷酸包含与所述靶多核苷酸的靶多核苷酸序列基本互补的多核苷酸序列。

[0446] 方面38：如方面33、34或37中任一项所述的方法，其中所述靶多核苷酸是在步骤(c)的孵育之前从宿主生物体获得或衍生，并且在步骤(c)的孵育之后重新引入相同的宿主生物体中。

[0447] 方面39：如方面33、34或37中任一项所述的方法，其中所述Cas9直系同源物多肽粘附至固体基质。

[0448] 方面40：如方面33、34或37中任一项所述的方法，其中所述Cas9直系同源物多肽是核酸酶、切口酶，或缺乏核酸酶或切口酶活性。

[0449] 方面41：如方面33所述的方法，其中所述靶多核苷酸是在步骤(c)的孵育之前从宿主生物体获得或衍生，并且在步骤(c)的孵育之后引入不同的生物体中。

[0450] 方面42：如方面33所述的方法，其中所述Cas9直系同源物多肽选自由以下组成的组：未经修饰的野生型Cas9直系同源物；功能性Cas9直系同源物变体；功能性Cas9直系同源物片段；包含活性或失活的Cas9直系同源物的融合蛋白；Cas9直系同源物，其在C末端上或在N末端上或在N和C末端两者上进一步包含一个或多个核定位序列(NLS)；生物素化的Cas9直系同源物；Cas9直系同源物切口酶；Cas9直系同源物内切核酸酶；进一步包含组氨酸标签的Cas9直系同源物；多种Cas9直系同源物；和上述任何两者或更多的混合物。

[0451] 方面43:如方面33所述的方法,其中所述Cas9直系同源物多肽选自由以下组成的组:(a)与SEQ ID NO:86-170中的任何一个具有至少80%同一性的多肽,(b)SEQ ID NO:86-170中的任何一个的功能性变体,(c)SEQ ID NO:86-170的任何一个的功能性片段,(d)由多核苷酸编码的选自由以下组成的组的Cas9多肽:SEQ ID NO:86-170或511-1135,(e)识别表4-83中的任何一个列出的PAM序列的Cas9多肽,和(f)从表1中列出的生物体鉴定的Cas9多肽。

[0452] 方面44:如方面33所述的方法,其进一步包括选自下组的组合物,该组由以下组成:200mM HEPES、50mM MgCl2、1M NaCl和1mM EDTA、蛋白酶、蛋白酶K和无核酸酶的水。

[0453] 方面45:如方面33所述的方法,其中所述修饰选自由以下组成的组:对现有核苷酸插入、缺失、取代以及添加或缔合原子或分子。

[0454] 方面46:如方面33所述的方法,其进一步包括供体多核苷酸,其中所述供体多核苷酸插入由所述Cas9直系同源物多肽产生的双链断裂中。

[0455] 方面47:一种修饰靶多核苷酸序列的体内方法,所述方法包括向细胞提供组合物,所述组合物包含如方面1-10中任一项所述的合成的组合物,

[0456] 其中所述细胞在其基因组中包含能够被所述组合物识别、结合并切割的多核苷酸序列。

[0457] 方面48:一种修饰细胞的基因组中的靶位点的方法,所述方法包括向所述细胞提供至少一种选自下组的Cas9直系同源物,该组由以下组成:(a)如方面3所述的失活的Cas9多肽,(b)与以下SEQ ID NO:86-170中的任何一个具有至少80%同一性的多肽,(c)SEQ ID NO:86-170的任何一个的功能性变体,(d)SEQ ID NO:86-170中任何一个的功能性片段,(e)识别表4-83中的任何一个列出的PAM序列的Cas9多肽,(f)从表1中列出的生物体鉴定的Cas9多肽,(g)选自由SEQ ID NO:86-170或511-1135组成的组的由cas9多核苷酸编码的Cas9多肽,以及(h)编码(a)至(g)的任何多肽的Cas9多肽;并且所述指导多核苷酸选自由以下组成的组:(i)与选自由SEQ ID NO:426-510组成的组的序列具有至少80%同一性的单指导RNA,(j)包含SEQ ID NO:426-510的功能性片段的单指导RNA,(k)包含SEQ ID NO:426-510的功能性变体的单指导RNA,(l)包含与tracrRNA连接的嵌合非天然存在的crRNA的单指导RNA,其中所述tracrRNA包含选自下组的核苷酸序列,该组由以下组成:SEQ ID NO:341-425,SEQ ID NO:341-425的功能性片段,和SEQ ID NO:341-425的功能性变体,(m)单导RNA包含与tracrRNA连接的嵌合非天然存在的crRNA,其中所述嵌合非天然存在的crRNA包含选自下组的核苷酸序列,该组由以下组成:SEQ ID NO:171-255,SEQ ID NO:171-255的功能性片段,和SEQ ID NO:171-255的功能性变体,(n)指导RNA,其是包含嵌合非天然存在的crRNA和tracrRNA的双链体分子,其中所述嵌合非天然存在的crRNA包含能够与所述靶序列杂交的片段,其中所述tracrRNA包含选自下组的核苷酸序列,该组由以下组成:SEQ ID NO:341-425,SEQ ID NO:341-425的功能性片段,和SEQ ID NO:341-425的功能性变体,(o)指导RNA,其是包含嵌合非天然存在的crRNA和tracrRNA的双链体分子,其中所述嵌合非天然存在的crRNA包含选自下组的核苷酸序列,该组由以下组成:SEQ ID NO:171-255,SEQ ID NO:171-255的功能性片段,和SEQ ID NO:171-255的功能性变体,其中所述嵌合非天然存在的crRNA包含能够与所述靶序列杂交的可变靶向结构域,(p)包含DNA和RNA两者的多核苷酸,(q)包含至少一个化学修饰的核苷酸的多核苷酸,和(r)能够转录成(i)至(q)的任何RNA分子的

DNA分子;其中所述指导多核苷酸和所述Cas9直系同源物能够形成复合物,所述复合物能够识别、结合并任选地使靶多核苷酸序列产生切口或切割靶多核苷酸序列;并鉴定至少一个细胞,所述至少一个细胞在所述细胞的靶位点处具有修饰,其中所述靶位点处的修饰选自由以下组成的组:(i)至少一个核苷酸的替代、(ii)至少一个核苷酸的缺失、(iii)至少一个核苷酸的插入、和(iv)至少一个核苷酸的修饰,和(v)(i)-(iv)的任何组合。

[0458] 方面49:如方面48所述的方法,所述方法包括向所述细胞提供多种Cas9多肽,其各自识别表4-83中任一个列出的不同PAM序列。

[0459] 方面50:如方面48所述的方法,其中将Cas9直系同源物的浓度以小于100微摩尔的浓度提供给所述细胞。

[0460] 方面51:如方面48所述的方法,其进一步包括向所述细胞提供多核苷酸修饰模板,其中与所述细胞的靶核苷酸序列相比,所述多核苷酸修饰模板包含至少一个核苷酸修饰。

[0461] 方面52:如方面49所述的方法,其中所述供体DNA包含目的多核苷酸。

[0462] 方面53:如方面52所述的方法,其进一步包括鉴定至少一个将所述目的多核苷酸整合到所述靶位点中或附近的细胞。

[0463] 方面54:如方面52所述的方法,其中所述目的多核苷酸赋予所述细胞或包含所述细胞的生物体益处。

[0464] 方面55:如方面54所述的方法,其中所述多核苷酸修饰或益处被赋予给所述细胞的或包含所述细胞的所述生物体的后续世代。

[0465] 方面56:如方面54或方面55所述的方法,其中所述益处选自由以下组成的组:改善的健康、改善的生长、改善的能育性、改善繁殖力、改善的环境耐受、改善的活力、改善的疾病抗性、改善的疾病耐受、改善的对异源分子的耐受、改善的适应性、改善的物理特征、更大的质量、增加的生化分子产生、减少的生化分子产生、基因的上调、基因的下调、生化途径的上调、生化途径的下调、细胞繁殖的刺激和细胞繁殖的抑制。

[0466] 方面57:如方面51-56中任一项所述的方法,其中所述细胞选自由以下组成的组:人、非人灵长类、哺乳动物、动物、古细菌、细菌、原生生物、真菌、昆虫、酵母、非常规酵母和植物细胞。

[0467] 方面58:如方面51-56中任一项所述的方法,其中所述细胞与衍生所述Cas9直系同源物的生物体是异源的。

[0468] 方面59:如方面57所述的方法,其中所述植物细胞选自由以下组成的组:单子叶植物和双子叶植物的细胞。

[0469] 方面60:如方面57所述的方法,其中所述植物细胞选自由以下组成的组:玉蜀黍、稻、高粱、黑麦、大麦、小麦、粟、燕麦、甘蔗、草坪草、柳枝稷、大豆、卡诺拉油菜、苜蓿、向日葵、棉花、烟草、花生、马铃薯、烟草、拟南芥属、蔬菜和红花细胞。

[0470] 方面61:如方面51-56中任一项所述的方法,其中所述细胞是植物细胞,并且其中所述靶位点的修饰导致包含所述细胞或其后代细胞的植物的具有农艺学意义的性状的调节,所述具有农艺学意义的性状选自由以下组成的组:疾病抗性、干旱抗性、热耐性、寒耐性、盐耐性、金属耐性、除草剂耐性、改善的水分利用效率、改善的氮利用率、改善的固氮作用、有害生物抗性、食草动物抗性、病原体抗性、产率改善、健康增强、改善的能育性、活力改善、生长改善、光合能力改善、营养增强、改变的蛋白含量、改变的油含量、增加的生物量、增

加的芽长度、增加的根长度、改善的根结构、代谢产物的调节、蛋白质组的调节、增加的种子重量、改变的种子碳水化合物组成、改变的种子油组成、改变的种子蛋白组成、改变的种子营养物组成;如与不包含所述靶位点修饰的同系植物(isoline plant)相比,或与所述植物细胞中所述靶位点的修饰之前的植物相比。

[0471] 方面62:如方面57所述的方法,其中所述人细胞选自由以下组成的组:单倍体细胞、二倍体细胞、生殖细胞、神经元、肌肉细胞、内分泌或外分泌细胞、上皮细胞、肌肉细胞、肿瘤细胞、胚胎细胞、造血细胞、骨细胞、种质细胞、体细胞、干细胞、多能干细胞、诱导多能干细胞、祖细胞、减数分裂细胞和有丝分裂细胞。

[0472] 方面63:如方面51-56中任一项所述的方法,其中所述细胞是动物细胞,并且其中所述靶位点的修饰导致包含所述动物细胞或其后代细胞的生物体的具有生理学意义的表型的调节,所述具有生理学意义的表型选自由以下组成的组:改善的健康、改善的营养状况、减少的疾病影响、疾病静止状态、疾病逆转、改善的能育性、改善的活力、改善的心智能力、改善的生物体生长、改善的增重、减重、内分泌系统的调节、外分泌系统的调节、减小的肿瘤大小、减小的肿瘤质量、刺激的细胞生长、降低的细胞生长、代谢产物的产生、激素的产生、免疫细胞的产生、刺激细胞产生。

[0473] 方面64:如方面50所述的方法,其中所述动物细胞是人细胞。

[0474] 方面65:一种包含经修饰的靶位点的植物,其中所述植物来源于包含经修饰的靶位点的植物细胞,所述经修饰的靶位点通过如方面51-56中任一项所述的方法产生。

[0475] 方面66:一种包含经编辑的核苷酸的植物,其中所述植物来源于包含经编辑的核苷酸的植物细胞,所述经编辑的核苷酸通过如方面49所述的方法产生。

[0476] 方面67:一种编辑多个多核苷酸靶序列的方法,所述方法包括向所述多个多核苷酸靶序列提供多种Cas9多肽,每种识别表4-83中任一个列出的不同PAM序列。

[0477] 方面68:一种通过以下来调节Cas9直系同源物/指导多核苷酸复合物与其野生型活性相比的靶多核苷酸特异性的方法:改变选自由以下组成的组的参数:(a)指导多核苷酸的长度,(b)指导多核苷酸的组成,(c)PAM序列的长度,(d)PAM序列的组成,以及(e)Cas9分子与靶多核苷酸主链的亲和力;并评估具有改变的参数的复合物的靶多核苷酸特异性,并将其与具有野生型参数的复合物的活性进行比较。

[0478] 方面69:一种优化Cas9分子的活性的方法,所述方法包括将至少一个核苷酸修饰引入选自由SEQ ID NO:86-170组成的组的序列中,并鉴定与SEQ ID NO:86-170相比的至少一种改善的特征。

[0479] 方面70:一种通过以下来优化Cas9分子的活性的方法:使亲本Cas9分子经历至少一轮随机蛋白改组,并选择具有至少一种不存在于所述亲本Cas9分子中的特征的所得分子。

[0480] 方面71:一种通过以下来优化Cas9分子的活性的方法:使亲本Cas9分子经历至少一轮非随机蛋白改组,并选择具有至少一种不存在于所述亲本Cas9分子中的特征的所得分子。

[0481] 方面72:一种合成的组合物,其包含Cas9直系同源内切核酸酶和能够与靶多核苷酸的PAM共有序列选择性杂交的异源多核苷酸,其中所述PAM共有序列的长度为至少3个核苷酸、至少4个核苷酸、至少5个核苷酸、至少6个核苷酸、至少7个核苷酸或大于7个核苷酸。

[0482] 方面73:一种实现靶多核苷酸的单链缺口或双链断裂的方法,其中所述靶多核苷酸包含能够被指导多核苷酸识别的PAM共有序列,所述方法包括将所述指导多核苷酸和Cas9直系同源物引至所述靶多核苷酸,其中所述单链缺口或双链断裂发生在所述靶多核苷酸内。

[0483] 方面74:一种合成的组合物,其包含Cas9直系同源内切核酸酶和能够与PAM共有核苷酸序列选择性杂交的异源多核苷酸,所述PAM共有核苷酸序列选自由以下组成的组:(a)AAA、(b)AAAA、(c)AAAAA、(d)AAAC、(e)AAAT、(f)AGA、(g)AGRG、(h)AHAC、(i)ANGG、(j)ARHHG、(k)ARNAT、(l)ATAA、(m)ATTTTT、(n)BAVMAR、(o)BGGAT、(p)CAA、(q)CAHGGDD(r)CC、(s)CCA、(t)CCH、(u)CDA、(v)CNA、(w)CNAVGAC、(x)CNG、(y)CT(z)CTA、(aa)CVG、(bb)DGGD(cc)GAAA、(dd)GG、(ee)GGAH、(ff)GGDG、(gg)GGN、(hh)GHAAA、(ii)GNA、(jj)GNAC、(kk)GNAY、(ll)GNG、(mm)GTAMY、(nn)GTGA、(oo)HAR(pp)NDGGD(qq)RNCAC、(rr)RTAA(ss)TC、(tt)TGAR、(uu)TTTTT、(vv)VNCC、(ww)VRACC、(xx)VRNTT和(yy)VRTTT;其中A=腺嘌呤,C=胞嘧啶,G=鸟嘌呤,T=胸腺嘧啶,R=A或G,Y=C或T,S=G或C,W=A或T,K=G或T,M=A或C,B=C或G或T,D=A或G或T,H=A或C或T,V=A或C或G,以及N=任意碱基;任选地,其中任何核苷酸可以在所述PAM共有核苷酸序列的侧翼。

[0484] 方面75:一种合成的组合物,所述合成的组合物包含异源组分和Cas内切核酸酶,其中所述Cas内切核酸酶包含至少一种选自下组的氨基酸特征,该组由以下组成:(a)位置13处的异亮氨酸(I),(b)位置21处的异亮氨酸(I),(c)位置71处的亮氨酸(L),(d)位置149处的亮氨酸(L),(e)位置150处的丝氨酸(S),(f)位置444处的亮氨酸(L),(g)位置445处的苏氨酸(T),(h)位置503处的脯氨酸(P),(i)位置587处的F(苯丙氨酸),(j)位置620处的A(丙氨酸),(k)位置623处的L(亮氨酸),(l)位置624处的T(苏氨酸),(m)位置632处的I(异亮氨酸),(n)位置692处的Q(谷氨酰胺),(o)位置702处的L(亮氨酸),(p)位置781处的I(异亮氨酸),(q)位置810处的K(赖氨酸),(r)位置908处的L(亮氨酸),(s)位置931处的V(缬氨酸),(t)位置933处的N/Q(天冬酰胺或谷氨酰胺),(u)位置954处的K(赖氨酸),(v)位置955处的V(缬氨酸),(w)位置1000处的K(赖氨酸),(x)位置1100处的V(缬氨酸),(y)位置1232处的Y(酪氨酸),以及(z)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQID NO:1125的序列比对确定的。

[0485] 方面76:如方面1所述的合成的组合物,其中所述Cas内切核酸酶与选自由SEQID NO:86-170和511-1135组成的组的序列具有至少90%同一性。

[0486] 方面77:如方面1所述的合成的组合物,其中根据表86A的氨基酸位置得分计算,所述Cas内切核酸酶的总得分大于3.14。

[0487] 方面78:如方面1所述的合成的组合物,其中所述Cas内切核酸酶已被修饰。

[0488] 方面79:如方面4所述的合成的组合物,其中所述Cas内切核酸酶已被修饰为缺乏内切核酸酶活性。

[0489] 方面80:如方面4所述的合成的组合物,其中所述Cas内切核酸酶已被修饰为使所述靶多核苷酸的单链产生切口。

[0490] 方面81:如方面4所述的合成的组合物,其中所述Cas内切核酸酶已被修饰以进一步包含异源核酸酶结构域,转录激活子结构域、转录阻遏子结构域、表观遗传修饰结构域、切割结构域、核定位信号、细胞穿透性结构域、脱氨酶结构域、碱基编辑结构域或易位结构

域。

[0491] 方面82:一种多核苷酸,其编码如方面1所述的多肽。

[0492] 方面83:一种质粒,其包含如方面8所述的多核苷酸。

[0493] 方面84:如方面9所述的质粒,其进一步包含与编码所述Cas内切核酸酶的多核苷酸可操作地连接的表达元件。

[0494] 方面85:如方面9所述的质粒,其进一步包含编码可选择标志物或转基因的基因。

[0495] 方面86:如方面1所述的合成的组合物,其中所述异源组分选自由以下组成的组:异源多核苷酸、异源多肽、粒子、固体基质、抗体、Tris、EDTA、二硫苏糖醇(DTT)、磷酸盐缓冲盐水(PBS)、氯化钠、氯化镁、HEPES、甘油、牛血清白蛋白(BSA)、盐、乳化剂、洗涤剂、螯合剂、蛋白酶、蛋白酶K、氧化还原剂、抗体、无核酸酶的水、粘度剂和组氨酸标签。

[0496] 方面87:如方面1所述的合成的组合物,其中所述Cas内切核酸酶在液体制剂中。

[0497] 方面88:如方面1所述的合成的组合物,其中所述Cas内切核酸酶在冻干制剂中。

[0498] 方面89:如方面1所述的合成的组合物,其中所述Cas内切核酸酶在基本上无内毒素的制剂中。

[0499] 方面90:如方面1所述的合成的组合物,其中所述Cas内切核酸酶在具有以下pH的制剂中:1.0至14.0、2.0至13.0、3.0至12.0、4.0至11.0、5.0至10.0、6.0至9.0、7.0至8.0、4.5至6.5、5.5至7.5、或6.5至7.5。

[0500] 方面91:如方面1所述的合成的组合物,其中所述Cas内切核酸酶在以下温度储存或孵育:至少负200摄氏度、至少负150摄氏度、至少负135摄氏度、至少负90摄氏度、至少负80摄氏度、至少负20摄氏度、至少4摄氏度、至少17摄氏度、至少20摄氏度、至少25摄氏度、至少30摄氏度、至少35摄氏度、至少37摄氏度、至少39摄氏度、至少40摄氏度、至少45摄氏度、至少50摄氏度、至少55摄氏度、至少60摄氏度、至少65摄氏度、至少70摄氏度或大于70摄氏度。

[0501] 方面92:如方面1所述的合成的组合物,其中所述Cas内切核酸酶附接至固体基质。

[0502] 方面93:如方面1所述的合成的组合物,其中所述固体基质是粒子。

[0503] 方面94:一种试剂盒,其包含如方面1所述的合成的组合物。

[0504] 方面95:如方面1所述的合成的组合物,其进一步包含指导多核苷酸。

[0505] 方面96:如方面1所述的合成的组合物,其进一步包含异源细胞。

[0506] 方面97:如方面22所述的合成的组合物,其中所述细胞获自真核、原核、植物或动物生物体。

[0507] 方面98:一种在靶多核苷酸中产生双链断裂的方法,所述方法包括使所述靶多核苷酸与以下接触:与所述靶核苷酸具有互补性的指导多核苷酸以及选自下组的Cas内切核酸酶,该组由以下组成:(a)多肽,其包含至少一种选自下组的氨基酸特征,该组有以下组成:(i)位置13处的异亮氨酸(I),(ii)位置21处的异亮氨酸(I),(iii)位置71处的亮氨酸(L),(iV)位置149处的亮氨酸(L),(v)位置150处的丝氨酸(S),(vi)位置444处的亮氨酸(L),(vii)位置445处的苏氨酸(T),(viii)位置503处的脯氨酸(P),(ix)位置587处的F(苯丙氨酸),(x)位置620处的A(丙氨酸),(xi)位置623处的L(亮氨酸),(xii)位置624处的T(苏氨酸),(xiii)位置632处的I(异亮氨酸),(xiv)位置692处的Q(谷氨酰胺),(xv)位置702处的L(亮氨酸),(xvi)位置781处的I(异亮氨酸),(xvii)位置810处的K(赖氨酸),(xviii)位

置908处的L(亮氨酸),(xix)位置931处的V(缬氨酸),(xx)位置933处的N/Q(天冬酰胺或谷氨酰胺),(xxi)位置954处的K(赖氨酸),(xxii)位置955处的V(缬氨酸),(xxiii)位置1000处的K(赖氨酸),(xxiv)位置1100处的V(缬氨酸),(xxv)位置1232处的Y(酪氨酸),以及(xxvi)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的;以及(b)多肽,其包含与选自由SEQ ID NO:1136-1730组成的组的序列至少90%相同的结构域;其中所述Cas内切核酸酶和所述指导RNA形成识别、结合并切割所述靶多核苷酸的复合物。

[0508] 方面99:如方面24所述的方法,其中所述多肽与SEQ ID NO:86-170和511-1135中任何一个具有至少90%的同一性。

[0509] 方面100:如方面24所述的方法,其中所述双链断裂包含粘性末端突出。

[0510] 方面101:如方面25所述的方法,其中所述Cas内切核酸酶包含与选自由SEQ ID NO:46、68、63、70、102、108、119和131组成的组的序列至少80%相同的多肽。

[0511] 方面102:如方面24所述的方法,其中所述双链断裂包含平末端。

[0512] 方面103:如方面25所述的方法,其中所述Cas内切核酸酶包含与选自下组的序列至少80%相同的氨基酸序列,该组由以下组成:SEQ ID NO:33、50、56、64、79、2、3、4、5、6、8、9、12、13、16、17、18、19、27、28、29、30、32、35、41、44、47、48、51、52、60、61、65、66、67、71、77、78、80、81、85、87、94和97。

[0513] 方面104:一种修饰DNA靶位点的方法,所述方法包括:(a)使包含所述DNA靶位点的多核苷酸与Cas内切核酸酶接触,所述Cas内切核酸酶包含选自下组的多肽,该组由以下组成:(i)多肽,其包含至少一种选自下组的氨基酸特征,该组有以下组成:(1)位置13处的异亮氨酸(I),(2)位置21处的异亮氨酸(I),(3)位置71处的亮氨酸(L),(4)位置149处的亮氨酸(L),(5)位置150处的丝氨酸(S),(6)位置444处的亮氨酸(L),(7)位置445处的苏氨酸(T),(8)位置503处的脯氨酸(P),(9)位置587处的F(苯丙氨酸),(10)位置620处的A(丙氨酸),(11)位置623处的L(亮氨酸),(12)位置624处的T(苏氨酸),(13)位置632处的I(异亮氨酸),(14)位置692处的Q(谷氨酰胺),(15)位置702处的L(亮氨酸),(16)位置781处的I(异亮氨酸),(17)位置810处的K(赖氨酸),(18)位置908处的L(亮氨酸),(19)位置931处的V(缬氨酸),(20)位置933处的N/Q(天冬酰胺或谷氨酰胺),(21)位置954处的K(赖氨酸),(22)位置955处的V(缬氨酸),(23)位置1000处的K(赖氨酸),(24)位置1100处的V(缬氨酸),(25)位置1232处的Y(酪氨酸),以及(26)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的;以及(ii)多肽,其包含与选自由SEQ ID NO:1136-1730组成的组的序列至少90%相同的结构域;以及(b)与所述DNA靶位点内或附近的序列具有互补性的指导多核苷酸,其中所述Cas内切核酸酶和所述指导RNA形成识别、结合所述DNA靶位点并使所述DNA靶位点产生切口或切割所述DNA靶位点的复合物;并且(c)检测在所述DNA靶位点处的至少一个修饰。

[0514] 方面105:如方面30所述的方法,其中所述Cas内切核酸酶是与SEQ ID NO:86-170和511-1135中任何一个具有至少90%同一性的多肽。

[0515] 方面106:如方面30所述的方法,其还包括在步骤(a)中引入供体DNA分子,其中所述供体DNA分子被整合到所述靶位点中。

[0516] 方面107:如方面30所述的方法,其进一步包括在步骤(a)中引入模板DNA分子,其

中所述模板DNA分子引导所述切割位点的修复结果。

[0517] 方面108:一种编辑靶多核苷酸的至少一个碱基的方法,所述方法包括:(a)使所述靶多核苷酸与以下接触:(i)脱氨酶,(ii)Cas内切核酸酶,其包含与SEQ ID NO:1136-1730中的任何一个具有至少90%同一性的多肽,其中所述Cas内切核酸酶已被修饰为缺乏核酸酶活性,以及(iii)指导多核苷酸,其与所述靶多核苷酸的序列具有互补性,其中所述Cas内切核酸酶和所述指导RNA形成识别并结合所述靶多核苷酸的复合物;并且(b)检测在DNA靶位点处的至少一个修饰。

[0518] 方面109:如方面34所述的方法,其中所述Cas内切核酸酶已被修饰为缺乏内切核酸酶活性。

[0519] 方面110:一种修饰细胞的基因组的方法,所述方法包括:

[0520] (a)将与细胞中的DNA靶位点中或附近的序列具有互补性的指导多核苷酸以及包含选自下组的多肽的异源Cas内切核酸酶引入所述细胞中,该组由以下组成:(i)多肽,其包含至少一种选自下组的氨基酸特征,该组有以下组成:位置13处的异亮氨酸(I),位置21处的异亮氨酸(I),位置71处的亮氨酸(L),位置149处的亮氨酸(L),位置150处的丝氨酸(S),位置444处的亮氨酸(L),位置445处的苏氨酸(T),位置503处的脯氨酸(P),位置587处的F(苯丙氨酸),位置620处的A(丙氨酸),位置623处的L(亮氨酸),位置624处的T(苏氨酸),位置632处的I(异亮氨酸),位置692处的Q(谷氨酰胺),位置702处的L(亮氨酸),位置781处的I(异亮氨酸),位置810处的K(赖氨酸),位置908处的L(亮氨酸),位置931处的V(缬氨酸),位置933处的N/Q(天冬酰胺或谷氨酰胺),位置954处的K(赖氨酸),位置955处的V(缬氨酸),位置1000处的K(赖氨酸),位置1100处的V(缬氨酸),位置1232处的Y(酪氨酸),以及位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的;以及(ii)多肽,其包含与选自由SEQ ID NO:1136-1730组成的组的序列至少90%相同的结构域;并且其中所述Cas内切核酸酶和所述指导RNA形成识别、结合并使所述所述DNA靶位点产生切口或切割所述DNA靶位点的复合物;并且(b)与未引入所述Cas内切核酸酶和指导多核苷酸的同系细胞相比,鉴定至少一个修饰。

[0521] 方面111:如方面35所述的方法,其进一步包括在步骤(a)中引入异源多核苷酸,其中所述异源多核苷酸是供体DNA或模板DNA。

[0522] 方面112:如方面35所述的方法,其中在步骤(a)之前将所述细胞从来源生物体移出,并在步骤(a)之后重新引入所述来源生物体中或引入新的生物体中。

[0523] 方面113:如方面35所述的方法,其中将所述细胞置于支持生长的培养基中,并从所述细胞再生组织或生物体

[0524] 方面114:如方面35所述的方法,其中修饰所述细胞的基因组的方法导致对从所述细胞获得或衍生的生物体的益处。

[0525] 方面115:如方面35所述的方法,其中所述细胞选自由以下组成的组:人、非人灵长类、哺乳动物、动物、古细菌、细菌、原生生物、真菌、昆虫、酵母、非常规酵母和植物细胞。

[0526] 方面116:如方面40所述的方法,其中所述生物体是植物。

[0527] 方面117:如方面42所述的方法,其中所述植物选自由以下组成的组:玉蜀黍、稻、高粱、黑麦、大麦、小麦、粟、燕麦、甘蔗、草坪草、柳枝稷、大豆、卡诺拉油菜、苜蓿、向日葵、棉花、烟草、花生、马铃薯、烟草、拟南芥属、蔬菜和红花。

[0528] 方面118:如方面42所述的方法,其中所述益处选自由以下组成的组:疾病抗性、干旱抗性、热耐性、寒耐性、盐耐性、金属耐性、除草剂耐性、改善的水分利用效率、改善的氮利用率、改善的固氮作用、有害生物抗性、食草动物抗性、病原体抗性、产率改善、健康增强、改善的能育性、活力改善、生长改善、光合能力改善、营养增强、改变的蛋白含量、改变的油含量、增加的生物量、增加的芽长度、增加的根长度、改善的根结构、代谢产物的调节、蛋白质组的调节、增加的种子重量、改变的种子碳水化合物组成、改变的种子油组成、改变的种子蛋白组成、改变的种子营养物组成;如与不包含所述靶位点修饰的同系植物(isoline plant)相比,或与所述植物细胞中所述靶位点的修饰之前的植物相比。

[0529] 方面119:如方面40所述的方法,其中所述生物体是动物。

[0530] 方面120:如方面45所述的方法,其中所述动物是人。

[0531] 方面121:如方面45所述的方法,其中所述动物细胞选自由以下组成的组:单倍体细胞、二倍体细胞、生殖细胞、神经元、肌肉细胞、内分泌或外分泌细胞、上皮细胞、肌肉细胞、肾细胞、肿瘤细胞、胚胎细胞、造血细胞、骨细胞、种质细胞、体细胞、干细胞、多能干细胞、诱导多能干细胞、祖细胞、减数分裂细胞和有丝分裂细胞。

[0532] 方面122:如方面45所述的方法,其中所述靶位点的修饰导致包含所述动物细胞或其后代细胞的生物体的具有生理学意义的表型的调节,所述具有生理学意义的表型选自由以下组成的组:改善的健康、改善的营养状况、减少的疾病影响、疾病静止状态、疾病逆转、改善的能育性、改善的活力、改善的心智能力、改善的生物体生长、改善的增重、减重、内分泌系统的调节、外分泌系统的调节、减小的肿瘤大小、减小的肿瘤质量、刺激的细胞生长、降低的细胞生长、代谢产物的产生、激素的产生、免疫细胞的产生、以及刺激细胞产生。

[0533] 方面123:一种Cas内切核酸酶,所述Cas内切核酸酶识别选自下组的PAM,该组由以下组成:NAR(G>A)WH(A>T>C)GN(C>T>R)、N(C>D)V(A>S)R(G>A)TTTN(T>V)、NV(A>G>C)TTTTT、NATTTTT、NN(H>G)AAAN(G>A>Y)N、N(T>V)NAAATN、NAV(A>G>C)TCNN、NN(A>S>T)NN(W>G>C)CCN(Y>R)、NNAH(T>M)ACN、NGTGANN、NARN(A>K>C)ATN、NV(G>A>C)RNTTN、NN(A>B)RN(A>G>T>C)CCN、NN(A>B)NN(T>V)CCH(A>Y)、NNN(H>G)NCDAA、NN(H>G)D(A>K)GGDN(A>B)、NNNNCCAG、NNNNCTAA、NNNNCVGANN、N(C>D)NNTCCN、NNNNCTA、NNNNCYAA、NAGRGNY、NNGH(W>C)AAA、NNGAAAN、NNAAAAA、NTGAR(G>A)N(A>Y>G)N(Y>R)、N(C>D)H(C>W)GH(Y>A)N(A>B)AN(A>T>S)、NNAAACN、NNGTAM(A>C)Y、NH(A>Y)ARNN(C>W>G)N、B(C>K)GGN(A>Y>G)N NN、N(T>C>R)AGAN(A>K>C)NN、NGGN(A>T>G>C)NNN、NGGD(A>T>G)TNN、NGGAN(T>A>C>G)NN、CGGWN(T>R>C)NN、NGGWGNN、N(B>A)GGNN(T>V)NN、NNGD(A>T>G)AY(T>C)N、N(T>V)H(T>C>A)AAAAN、NRTAANN、N(H>G)CAAH(Y>A)N(Y>R)N、NATAAN(A>T>S)N、NV(A>G>C)R(A>G)ACCN、CN(C>W>G)AV(A>S)GAC、NNRNCAC、N(A>B)GGD(W>G)D(G>W)NN、BGD(G>W)GTCN(A>K>C)、NAANACN、NRTHAN(A>B)N、BHN(H>G)NGN(T>M)H(Y>A)、NMRN(A>Y>G)AH(C>T>A)N、NNNCACN、NARN(T>A>S)ACN、NNNNATW、NGCNGCN、NNNCATN、NAGNGCN、NARN(T>M>G)CCN、NATCCTN、NRTAAN(T>A>S)N、N(C>T>G>A)AAD(A>G>T)CNN、NAAAGNN、NNGACNN、N(T>V)NTAAD(A>T>G)N、NNGAD(G>W)NN、NGGN(W>S)NNN、N(T>V)GGD(W>G)GNN、NGGD(A>T>G)N(T>M>G)NN、NNAAAGN、N(G>H)GGDN(T>M>G)NN、NNAGAAA、NN(T>M>G)AAAAA、N(C>D)N(C>W>G)GW(T>C)D(A>G>T)AA、NAAAAYN、NRGNNNN、NATGN(H>G)TN、NNDATTT和NATARCN(C>T>A>G)。

[0534] 方面124:一种合成的组合物,所述合成的组合物包含异源组分和Cas内切核酸酶,其中所述Cas内切核酸酶包含至少一种选自下组的氨基酸特征,该组由以下组成:(a)位置13处的异亮氨酸(I),(b)位置21处的异亮氨酸(I),(c)位置71处的亮氨酸(L),(d)位置149处的亮氨酸(L),(e)位置150处的丝氨酸(S),(f)位置444处的亮氨酸(L),(g)位置445处的苏氨酸(T),(h)位置503处的脯氨酸(P),(i)位置587处的F(苯丙氨酸),(j)位置620处的A(丙氨酸),(k)位置623处的L(亮氨酸),(l)位置624处的T(苏氨酸),(m)位置632处的I(异亮氨酸),(n)位置692处的Q(谷氨酰胺),(o)位置702处的L(亮氨酸),(p)位置781处的I(异亮氨酸),(q)位置810处的K(赖氨酸),(r)位置908处的L(亮氨酸),(s)位置931处的V(缬氨酸),(t)位置933处的N/Q(天冬酰胺或谷氨酰胺),(u)位置954处的K(赖氨酸),(v)位置955处的V(缬氨酸),(w)位置1000处的K(赖氨酸),(x)位置1100处的V(缬氨酸),(y)位置1232处的Y(酪氨酸),以及(z)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQID NO:1125的序列比对确定的。

[0535] 方面125:如方面1所述的合成的组合物,其中所述Cas内切核酸酶与选自由SEQID NO:86-170和511-1135组成的组的序列具有至少90%同一性。

[0536] 方面126:如方面1所述的合成的组合物,其中所述Cas内切核酸酶包含与SEQID NO:1136-1730中任一个具有90%或更高同一性的结构域。

[0537] 方面127:如方面1所述的合成的组合物,其中所述Cas内切核酸酶与异源多肽融合。

[0538] 方面128:如方面4所述的合成的组合物,其中所述异源多肽包含核酸酶活性。

[0539] 方面129:如方面4所述的合成的组合物,其中所述异源多肽是脱氨酶。

[0540] 方面130:如方面1所述的合成的组合物,其进一步包含指导多核苷酸,所述多肽与所述指导多核苷酸形成复合物。

[0541] 方面131:如方面2所述的合成的组合物,其中所述指导多核苷酸是单指导物,所述单指导物包含选自由SEQID NO:426-510组成的组的序列。

[0542] 方面132:如方面2所述的合成的组合物,其中所述指导多核苷酸包含tracrRNA,所述tracrRNA包含选自由SEQID NO:341-425组成的组的序列。

[0543] 方面133:如方面2所述的合成的组合物,其中所述指导多核苷酸包含crRNA,所述crRNA包含选自由SEQID NO:171-255组成的组的序列。

[0544] 方面134:如方面2所述的合成的组合物,其中所述指导多核苷酸包含反重复序列,所述反重复序列包含选自由SEQID NO:256-340组成的组的序列。

[0545] 方面135:如方面2所述的合成的组合物,其中所述指导多核苷酸指导物包含DNA。

[0546] 方面136:如方面1所述的合成的组合物,其与表4-83中列出的PAM共有序列选择性杂交。

[0547] 方面137:一种Cas内切核酸酶或失活的Cas内切核酸酶,所述Cas内切核酸酶或失活的Cas内切核酸酶识别选自下组的PAM,该组由以下组成:NAR(G>A)WH(A>T>C)GN(C>T>R)、N(C>D)V(A>S)R(G>A)TTTN(T>V)、NV(A>G>C)TTTTT、NATTTTT、NN(H>G)AAAN(G>A>Y)N、N(T>V)NAAATN、NAV(A>G>C)TCNN、NN(A>S>T)NN(W>G>C)CCN(Y>R)、NNAH(T>M)ACN、NGTGANN、NARN(A>K>C)ATN、NV(G>A>C)RNTTN、NN(A>B)RN(A>G>T>C)CCN、NN(A>B)NN(T>V)CCH(A>Y)、NNN(H>G)NCDAA、NN(H>G)D(A>K)GGDN(A>B)、

NNNNCCAG、NNNNCTAA、NNNNCVGANN、N(C＞D)NNTCCN、NNNNCTA、NNNNCYAA、NAGRGNY、NNGH(W＞C)AAA、NNGAAAN、NNAAAAA、NTGAR(G＞A)N(A＞Y＞G)N(Y＞R)、N(C＞D)H(C＞W)GH(Y＞A)N(A＞B)AN(A＞T＞S)、NNAAACN、NNGTAM(A＞C)Y、NH(A＞Y)ARNN(C＞W＞G)N、B(C＞K)GGN(A＞Y＞G)N NN、N(T＞C＞R)AGAN(A＞K＞C)NN、NGGN(A＞T＞G＞C)NNN、NGGD(A＞T＞G)TNN、NGGAN(T＞A＞C＞G)NN、CGGWN(T＞R＞C)NN、NGGWGNN、N(B＞A)GGNN(T＞V)NN、NNGD(A＞T＞G)AY(T＞C)N、N(T＞V)H(T＞C＞A)AAAAN、NRTAANN、N(H＞G)CAAH(Y＞A)N(Y＞R)N、NATAAN(A＞T＞S)N、NV(A＞G＞C)R(A＞G)ACCN、CN(C＞W＞G)AV(A＞S)GAC、NNRNCAC、N(A＞B)GGD(W＞G)D(G＞W)NN、BGD(G＞W)GTCN(A＞K＞C)、NAANACN、NRTHAN(A＞B)N、BHN(H＞G)NGN(T＞M)H(Y＞A)、NMRN(A＞Y＞G)AH(C＞T＞A)N、NNNCACN、NARN(T＞A＞S)ACN、NNNNATW、NGCNGCN、NNNCATN、NAGNGCN、NARN(T＞M＞G)CCN、NATCCTN、NRTAAN(T＞A＞S)N、N(C＞T＞G＞A)AAD(A＞G＞T)CNN、NAAAGNN、NNGACNN、N(T＞V)NTAAD(A＞T＞G)N、NNGAD(G＞W)NN、NGGN(W＞S)NNN、N(T＞V)GGD(W＞G)GNN、NGGD(A＞T＞G)N(T＞M＞G)NN、NNAAAGN、N(G＞H)GGDN(T＞M＞G)NN、NNAGAAA、NN(T＞M＞G)AAAAA、N(C＞D)N(C＞W＞G)GW(T＞C)D(A＞G＞T)AA、NAAAAYN、NRGNNNN、NATGN(H＞G)TN、NNDATTT和NATARCN(C＞T＞A＞G)。

[0548] 方面138:如方面1所述的合成的组合物,其是从表1中列出的生物体鉴定。

[0549] 方面139:如方面1所述的合成的组合物,其选自由SEQ ID NO:86-170组成的组。

[0550] 方面140:如方面1所述的合成的组合物,其中靶细胞优化的多肽缺乏内切核酸酶活性。

[0551] 方面141:如方面1所述的合成的组合物,其中靶细胞优化的多肽能够使单链靶多核苷酸产生切口。

[0552] 方面142:如方面1所述的合成的组合物,其中靶细胞优化的多肽能够切割双链靶多核苷酸。

[0553] 方面143:如方面1所述的合成的组合物,其进一步包含供体DNA分子。

[0554] 方面144:如方面1所述的合成的组合物,其进一步包含修复模板DNA分子。

[0555] 方面145:如方面1所述的合成的组合物,其中所述异源组合物选自由以下组成的组:异源多核苷酸、异源多肽、粒子、固体基质、抗体、缓冲液组合物、Tris、EDTA、二硫苏糖醇(DTT)、磷酸盐缓冲盐水(PBS)、氯化钠、氯化镁、HEPES、甘油、牛血清白蛋白(BSA)、盐、乳化剂、洗涤剂、螯合剂、氧化还原剂、抗体、无核酸酶的水、粘度剂和组氨酸标签。

[0556] 方面146:如方面19所述的合成的组合物,其进一步包含另外的异源组合物。

[0557] 方面147:如方面1所述的合成的组合物,其进一步包含细胞。

[0558] 方面148:如方面21所述的合成的组合物,其中所述细胞从选自下组的生物体获得或衍生,该组由以下组成:人、非人灵长类、哺乳动物、动物、古细菌、细菌、原生生物、真菌、昆虫、酵母、非常规酵母和植物。

[0559] 方面149:如方面22所述的合成的组合物,其中所述植物细胞获得自或衍生自玉蜀黍、稻、高粱、黑麦、大麦、小麦、粟、燕麦、甘蔗、草坪草、柳枝稷、大豆、卡诺拉油菜、苜蓿、向日葵、棉花、烟草、花生、马铃薯、烟草、拟南芥属、蔬菜或红花。

[0560] 方面150:如方面22所述的合成的组合物,其中所述动物细胞选自由以下组成的组:单倍体细胞、二倍体细胞、生殖细胞、神经元、肌肉细胞、内分泌或外分泌细胞、上皮细胞、肌肉细胞、肿瘤细胞、胚胎细胞、造血细胞、骨细胞、种质细胞、体细胞、干细胞、多能干细

胞、诱导多能干细胞、祖细胞、减数分裂细胞和有丝分裂细胞。

[0561] 方面151:一种多核苷酸,其编码如方面1所述的多肽。

[0562] 方面152:如方面25所述的多核苷酸,其中所述多核苷酸包含在载体中,所述载体进一步包含至少一种异源多核苷酸。

[0563] 方面153:一种试剂盒,其包含方面1所述的合成的组合物或如方面25所述的多核苷酸。

[0564] 方面154:如方面1所述的合成的组合物,其中所述多肽在液体制剂中。

[0565] 方面155:如方面1所述的合成的组合物,其中所述多肽在冻干组合物中。

[0566] 方面156:如方面1所述的合成的组合物,其中所述多肽在基本上无内毒素的制剂中。

[0567] 方面157:如方面1所述的合成的组合物,其中所述多肽在具有以下pH的制剂中:1.0至14.0、2.0至13.0、3.0至12.0、4.0至11.0、5.0至10.0、6.0至9.0、7.0至8.0、4.5至6.5、5.5至7.5、或6.5至7.5。

[0568] 方面158:如方面1所述的合成的组合物,其中所述多肽在以下温度储存或孵育:至少负200摄氏度、至少负150摄氏度、至少负135摄氏度、至少负90摄氏度、至少负80摄氏度、至少负20摄氏度、至少4摄氏度、至少17摄氏度、至少20摄氏度、至少25摄氏度、至少30摄氏度、至少35摄氏度、至少37摄氏度、至少39摄氏度、至少40摄氏度、至少45摄氏度、至少50摄氏度、至少55摄氏度、至少60摄氏度、至少65摄氏度、至少70摄氏度或大于70摄氏度。

[0569] 方面159:如方面1所述的合成的组合物,其中所述多肽附接至固体基质。

[0570] 方面160:如方面33所述的合成的组合物,其中所述固体基质是粒子。

[0571] 方面161:一种检测靶多核苷酸序列的方法,所述方法包括:(a)获得所述靶多核苷酸,(b)在反应容器中组合Cas内切核酸酶、指导多核苷酸和所述靶多核苷酸,(c)在至少10摄氏度的温度下孵育步骤(b)的组分至少1分钟,(d)对反应混合物中的所得的一种或多种多核苷酸进行测序,并且(e)表征由所述Cas内切核酸酶和所述指导多核苷酸鉴定的步骤(a)的靶多核苷酸的序列;(f)其中所述指导多核苷酸包含与所述靶多核苷酸的序列基本互补的多核苷酸序列;其中所述Cas内切核酸酶包含至少一种选自下组的氨基酸特征,该组由以下组成:(a)位置13处的异亮氨酸(I),(b)位置21处的异亮氨酸(I),(c)位置71处的亮氨酸(L),(d)位置149处的亮氨酸(L),(e)位置150处的丝氨酸(S),(f)位置444处的亮氨酸(L),(g)位置445处的苏氨酸(T),(h)位置503处的脯氨酸(P),(i)位置587处的F(苯丙氨酸),(j)位置620处的A(丙氨酸),(k)位置623处的L(亮氨酸),(l)位置624处的T(苏氨酸),(m)位置632处的I(异亮氨酸),(n)位置692处的Q(谷氨酰胺),(o)位置702处的L(亮氨酸),(p)位置781处的I(异亮氨酸),(q)位置810处的K(赖氨酸),(r)位置908处的L(亮氨酸),(s)位置931处的V(缬氨酸),(t)位置933处的N/Q(天冬酰胺或谷氨酰胺),(u)位置954处的K(赖氨酸),(v)位置955处的V(缬氨酸),(w)位置1000处的K(赖氨酸),(x)位置1100处的V(缬氨酸),(y)位置1232处的Y(酪氨酸),以及(z)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQID NO:1125的序列比对确定的。

[0572] 方面162:一种将Cas内切核酸酶和指导多核苷酸复合物结合至靶多核苷酸的方法,所述方法包括:(a)获得所述靶多核苷酸的序列,(b)在反应容器中组合Cas内切核酸酶、指导多核苷酸和所述靶多核苷酸,(c)在至少10摄氏度的温度下孵育步骤(b)的组分至少1

分钟;其中所述指导多核苷酸包含与所述靶多核苷酸的靶多核苷酸序列基本互补的多核苷酸序列;进一步包括检测与所述靶多核苷酸结合的所述Cas内切核酸酶和指导多核苷酸复合物;并且其中所述Cas内切核酸酶包含至少一种选自下组的氨基酸特征,该组由以下组成:(a)位置13处的异亮氨酸(I),(b)位置21处的异亮氨酸(I),(c)位置71处的亮氨酸(L),(d)位置149处的亮氨酸(L),(e)位置150处的丝氨酸(S),(f)位置444处的亮氨酸(L),(g)位置445处的苏氨酸(T),(h)位置503处的脯氨酸(P),(i)位置587处的F(苯丙氨酸),(j)位置620处的A(丙氨酸),(k)位置623处的L(亮氨酸),(l)位置624处的T(苏氨酸),(m)位置632处的I(异亮氨酸),(n)位置692处的Q(谷氨酰胺),(o)位置702处的L(亮氨酸),(p)位置781处的I(异亮氨酸),(q)位置810处的K(赖氨酸),(r)位置908处的L(亮氨酸),(s)位置931处的V(缬氨酸),(t)位置933处的N/Q(天冬酰胺或谷氨酰胺),(u)位置954处的K(赖氨酸),(v)位置955处的V(缬氨酸),(w)位置1000处的K(赖氨酸),(x)位置1100处的V(缬氨酸),(y)位置1232处的Y(酪氨酸),以及(z)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQID NO:1125的序列比对确定的。

[0573] 方面163:一种在靶多核苷酸中产生双链断裂的方法,所述方法包括:(d)获得所述靶多核苷酸的序列,(e)在反应容器中组合Cas内切核酸酶多肽、指导多核苷酸和所述靶多核苷酸,(f)在至少10摄氏度的温度下孵育步骤(b)的组分至少1分钟;其中所述指导多核苷酸包含与所述靶多核苷酸的靶多核苷酸序列基本互补的多核苷酸序列;进一步包括检测与所述靶多核苷酸结合的所述Cas内切核酸酶和指导多核苷酸复合物;并且其中所述Cas内切核酸酶包含至少一种选自下组的氨基酸特征,该组由以下组成:(a)位置13处的异亮氨酸(I),(b)位置21处的异亮氨酸(I),(c)位置71处的亮氨酸(L),(d)位置149处的亮氨酸(L),(e)位置150处的丝氨酸(S),(f)位置444处的亮氨酸(L),(g)位置445处的苏氨酸(T),(h)位置503处的脯氨酸(P),(i)位置587处的F(苯丙氨酸),(j)位置620处的A(丙氨酸),(k)位置623处的L(亮氨酸),(l)位置624处的T(苏氨酸),(m)位置632处的I(异亮氨酸),(n)位置692处的Q(谷氨酰胺),(o)位置702处的L(亮氨酸),(p)位置781处的I(异亮氨酸),(q)位置810处的K(赖氨酸),(r)位置908处的L(亮氨酸),(s)位置931处的V(缬氨酸),(t)位置933处的N/Q(天冬酰胺或谷氨酰胺),(u)位置954处的K(赖氨酸),(v)位置955处的V(缬氨酸),(w)位置1000处的K(赖氨酸),(x)位置1100处的V(缬氨酸),(y)位置1232处的Y(酪氨酸),以及(z)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQID NO:1125的序列比对确定的。

[0574] 方面164:如方面36或方面37所述的方法,其进一步包括至少一个另外的靶位点。

[0575] 方面165:一种用于编辑细胞的基因组的方法,所述方法包括向所述细胞提供:(a)至少一种Cas内切核酸酶,其包含至少一种选自下组的氨基酸特征,该组由以下组成:(i)位置13处的异亮氨酸(I),(ii)位置21处的异亮氨酸(I),(iii)位置71处的亮氨酸(L),(iv)位置149处的亮氨酸(L),(v)位置150处的丝氨酸(S),(vi)位置444处的亮氨酸(L),(vii)位置445处的苏氨酸(T),(viii)位置503处的脯氨酸(P),(ix)位置587处的F(苯丙氨酸),(x)位置620处的A(丙氨酸),(xi)位置623处的L(亮氨酸),(xii)位置624处的T(苏氨酸),(xiii)位置632处的I(异亮氨酸),(xiv)位置692处的Q(谷氨酰胺),(xv)位置702处的L(亮氨酸),(xvi)位置781处的I(异亮氨酸),(xvii)位置810处的K(赖氨酸),(xviii)位置908处的L(亮氨酸),(xix)位置931处的V(缬氨酸),(xx)位置933处的N/Q(天冬酰胺或谷氨酰胺),(xxi)位置954处的K(赖氨酸),(xxii)位置955处的V(缬氨酸),(xxiii)位置1000处的K(赖氨酸),

(xxiv)位置1100处的V(缬氨酸),(xxv)位置1232处的Y(酪氨酸),以及(xxvi)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的;和(b)指导多核苷酸,所述Cas内切核酸酶与所述指导多核苷酸形成复合物;其中所述复合物能够识别、结合靶多核苷酸序列并任选地使靶多核苷酸序列产生切口或切割靶多核苷酸序列;并且鉴定在所述细胞的基因组DNA序列中具有修饰的至少一个细胞,其中所述修饰选自由以下组成的组:对现有核苷酸插入、缺失、取代以及添加或缔合原子或分子。

[0576] 方面166:一种调节细胞中基因的表达的方法,所述方法包括向所述细胞提供:(a)至少一种Cas内切核酸酶,其包含至少一种选自下组的氨基酸特征,该组由以下组成:(i)位置13处的异亮氨酸(I),(ii)位置21处的异亮氨酸(I),(iii)位置71处的亮氨酸(L),(iv)位置149处的亮氨酸(L),(v)位置150处的丝氨酸(S),(vi)位置444处的亮氨酸(L),(vii)位置445处的苏氨酸(T),(viii)位置503处的脯氨酸(P),(ix)位置587处的F(苯丙氨酸),(x)位置620处的A(丙氨酸),(xi)位置623处的L(亮氨酸),(xii)位置624处的T(苏氨酸),(xiii)位置632处的I(异亮氨酸),(xiv)位置692处的Q(谷氨酰胺),(xv)位置702处的L(亮氨酸),(xvi)位置781处的I(异亮氨酸),(xvii)位置810处的K(赖氨酸),(xviii)位置908处的L(亮氨酸),(xix)位置931处的V(缬氨酸),(xx)位置933处的N/Q(天冬酰胺或谷氨酰胺),(xxi)位置954处的K(赖氨酸),(xxii)位置955处的V(缬氨酸),(xxiii)位置1000处的K(赖氨酸),(xxiv)位置1100处的V(缬氨酸),(xxv)位置1232处的Y(酪氨酸),以及(xxvi)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的,和(b)指导多核苷酸,所述Cas内切核酸酶与所述指导多核苷酸形成复合物;其中所述复合物能够识别、结合所述细胞中的靶多核苷酸序列并任选地使所述细胞中的靶多核苷酸序列产生切口或切割所述细胞中的靶多核苷酸序列;并且鉴定与未引入所述Cas内切核酸酶的细胞相比具有调节的基因表达的至少一个细胞。

[0577] 方面167:如方面39或方面40所述的方法,其进一步包括向所述细胞提供供体DNA分子。

[0578] 方面168:如方面39或方面40所述的方法,其进一步包括向所述细胞提供模板DNA分子。

[0579] 方面169:如方面39或方面40所述的方法,其中所述方法赋予所述细胞或包含所述细胞的生物体益处。

[0580] 方面170:如方面41所述的方法,其中所述益处选自由以下组成的组:改善的健康、改善的生长、改善的能育性、改善繁殖力、改善的环境耐受、改善的活力、改善的疾病抗性、改善的疾病耐受、改善的对异源分子的耐受、改善的适应性、改善的物理特征、更大的质量、增加的生化分子产生、减少的生化分子产生、基因的上调、基因的下调、生化途径的上调、生化途径的下调、细胞繁殖的刺激和细胞繁殖的抑制。

[0581] 方面171:如方面39或方面40所述的方法,其中所述细胞与衍生所述Cas内切核酸酶的生物体是异源的,并且选自由以下组成的组:人、非人灵长类、哺乳动物、动物、古细菌、细菌、原生生物、真菌、昆虫、酵母、非常规酵母和植物细胞。

[0582] 方面172:如方面45所述的方法,其中所述植物细胞获得自或衍生自玉蜀黍、稻、高粱、黑麦、大麦、小麦、粟、燕麦、甘蔗、草坪草、柳枝稷、大豆、卡诺拉油菜、苜蓿、向日葵、棉花、烟草、花生、马铃薯、烟草、拟南芥属、蔬菜或红花。

[0583] 方面173:如方面45所述的方法,其中所述细胞是植物细胞,并且所述益处是调节包含所述细胞或其后代细胞的植物的具有农艺学意义的性状,所述具有农艺学意义的性状选自由以下组成的组:疾病抗性、干旱抗性、热耐性、寒耐性、盐耐性、金属耐性、除草剂耐性、改善的水分利用效率、改善的氮利用率、改善的固氮作用、有害生物抗性、食草动物抗性、病原体抗性、产率改善、健康增强、改善的能育性、活力改善、生长改善、光合能力改善、营养增强、改变的蛋白含量、改变的油含量、增加的生物量、增加的芽长度、增加的根长度、改善的根结构、代谢产物的调节、蛋白质组的调节、增加的种子重量、改变的种子碳水化合物组成、改变的种子油组成、改变的种子蛋白组成、改变的种子营养物组成;如与不包含所述靶位点修饰的同系植物(isoline plant)相比,或与所述植物细胞中所述靶位点的修饰之前的植物相比。

[0584] 方面174:如方面45所述的方法,其中所述动物细胞选自由以下组成的组:单倍体细胞、二倍体细胞、生殖细胞、神经元、肌肉细胞、内分泌或外分泌细胞、上皮细胞、肌肉细胞、肿瘤细胞、胚胎细胞、造血细胞、骨细胞、种质细胞、体细胞、干细胞、多能干细胞、诱导多能干细胞、祖细胞、减数分裂细胞和有丝分裂细胞。

[0585] 方面175:如方面45所述的方法,其中所述细胞是动物细胞并且所述益处是调节包含所述动物细胞或其后代细胞的生物体的具有生理学意义的表型,所述具有生理学意义的表型选自由以下组成的组:改善的健康、改善的营养状况、减少的疾病影响、疾病静止状态、疾病逆转、改善的能育性、改善的活力、改善的心智能力、改善的生物体生长、改善的增重、减重、内分泌系统的调节、外分泌系统的调节、减小的肿瘤大小、减小的肿瘤质量、刺激的细胞生长、降低的细胞生长、代谢产物的产生、激素的产生、免疫细胞的产生、以及刺激细胞产生。

[0586] 方面176:一种编辑靶多核苷酸的至少一个碱基的方法,所述方法包括:(a)使所述靶多核苷酸与以下接触:i.脱氨酶,ii.Cas内切核酸酶,其能够与表4-83中列出的PAM共有序列选择性杂交,其中所述Cas内切核酸酶已被修饰为缺乏核酸酶活性,以及iii.指导多核苷酸,其与所述靶多核苷酸的序列具有互补性,其中所述Cas内切核酸酶和所述指导RNA形成识别并结合所述靶多核苷酸的复合物;并且(b)检测在DNA靶位点处的至少一个修饰。

[0587] 方面177:一种编辑靶多核苷酸的多个碱基的方法,所述方法包括:(a)使所述靶多核苷酸与以下接触:i.至少一种脱氨酶,ii多种Cas内切核酸酶,每种能够与表4-83中列出的PAM共有序列选择性杂交,其中所述Cas内切核酸酶已被修饰为缺乏核酸酶活性,以及iii.指导多核苷酸,其与所述靶多核苷酸的序列具有互补性,其中所述Cas内切核酸酶和所述指导RNA形成识别并结合所述靶多核苷酸的复合物;并且(b)检测在DNA靶位点处的至少一个修饰。

[0588] 方面178:一种优化Cas分子的活性的方法,所述方法包括将至少一个核苷酸修饰引入包含至少一种选自下组的氨基酸特征的序列,该组由以下组成:(a)位置13处的异亮氨酸(I),(b)位置21处的异亮氨酸(I),(c)位置71处的亮氨酸(L),(d)位置149处的亮氨酸(L),(e)位置150处的丝氨酸(S),(f)位置444处的亮氨酸(L),(g)位置445处的苏氨酸(T),(h)位置503处的脯氨酸(P),(i)位置587处的F(苯丙氨酸),(j)位置620处的A(丙氨酸),(k)位置623处的L(亮氨酸),(l)位置624处的T(苏氨酸),(m)位置632处的I(异亮氨酸),(n)位置692处的Q(谷氨酰胺),(o)位置702处的L(亮氨酸),(p)位置781处的I(异亮氨酸),(q)位

置810处的K(赖氨酸),(r)位置908处的L(亮氨酸),(s)位置931处的V(缬氨酸),(t)位置933处的N/Q(天冬酰胺或谷氨酰胺),(u)位置954处的K(赖氨酸),(v)位置955处的V(缬氨酸),(w)位置1000处的K(赖氨酸),(x)位置1100处的V(缬氨酸),(y)位置1232处的Y(酪氨酸),以及(z)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQ ID NO:1125的序列比对确定的;并且与核苷酸修饰之前的分子相比,鉴定至少一种改善的特征。

[0589] 方面179:一种通过以下来优化Cas9分子的活性的方法:使亲本Cas9分子经历至少一轮随机蛋白改组,并选择具有至少一种不存在于所述亲本Cas9分子中的特征的所得分子;其中所述亲本Cas9分子包含至少一种选自下组的氨基酸特征,该组由以下组成:(a)位置13处的异亮氨酸(I),(b)位置21处的异亮氨酸(I),(c)位置71处的亮氨酸(L),(d)位置149处的亮氨酸(L),(e)位置150处的丝氨酸(S),(f)位置444处的亮氨酸(L),(g)位置445处的苏氨酸(T),(h)位置503处的脯氨酸(P),(i)位置587处的F(苯丙氨酸),(j)位置620处的A(丙氨酸),(k)位置623处的L(亮氨酸),(l)位置624处的T(苏氨酸),(m)位置632处的I(异亮氨酸),(n)位置692处的Q(谷氨酰胺),(o)位置702处的L(亮氨酸),(p)位置781处的I(异亮氨酸),(q)位置810处的K(赖氨酸),(r)位置908处的L(亮氨酸),(s)位置931处的V(缬氨酸),(t)位置933处的N/Q(天冬酰胺或谷氨酰胺),(u)位置954处的K(赖氨酸),(v)位置955处的V(缬氨酸),(w)位置1000处的K(赖氨酸),(x)位置1100处的V(缬氨酸),(y)位置1232处的Y(酪氨酸),以及(z)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQID NO:1125的序列比对确定的。

[0590] 方面180:一种通过以下来优化Cas9分子的活性的方法:使亲本Cas9分子经历至少一轮非随机蛋白改组,并选择具有至少一种不存在于所述亲本Cas9分子中的特征的所得分子;其中所述亲本Cas9分子包含基序,所述基序选自由以下组成的组:包含至少一种选自下组的氨基酸特征,该组由以下组成:(a)位置13处的异亮氨酸(I),(b)位置21处的异亮氨酸(I),(c)位置71处的亮氨酸(L),(d)位置149处的亮氨酸(L),(e)位置150处的丝氨酸(S),(f)位置444处的亮氨酸(L),(g)位置445处的苏氨酸(T),(h)位置503处的脯氨酸(P),(i)位置587处的F(苯丙氨酸),(j)位置620处的A(丙氨酸),(k)位置623处的L(亮氨酸),(l)位置624处的T(苏氨酸),(m)位置632处的I(异亮氨酸),(n)位置692处的Q(谷氨酰胺),(o)位置702处的L(亮氨酸),(p)位置781处的I(异亮氨酸),(q)位置810处的K(赖氨酸),(r)位置908处的L(亮氨酸),(s)位置931处的V(缬氨酸),(t)位置933处的N/Q(天冬酰胺或谷氨酰胺),(u)位置954处的K(赖氨酸),(v)位置955处的V(缬氨酸),(w)位置1000处的K(赖氨酸),(x)位置1100处的V(缬氨酸),(y)位置1232处的Y(酪氨酸),以及(z)位置1236处的I(异亮氨酸);其中位置编号是通过针对SEQID NO:1125的序列比对确定的。

[0591] 尽管已经参照优选实施例和各种替代实施例明确展示和描述了本发明,但是本领域技术人员应理解,在不脱离本发明的精神和范围的情况下,可以对其在形式和细节上进行各种改变。例如,尽管下面的特定实例可以阐述本文中使用特定植物来描述的方法和实施例,但是这些实例中的原理可以应用于任何植物。因此,应当理解,本发明的范围被本文和说明书中记载的本发明的实施例所涵盖,而不是由以下示例的具体实例所涵盖。出于所有目的,在本申请中提到的所有引用的专利和出版物通过引用以其整体并入本文,其程度如同它们各自单独和特别地通过引用并入。

[0592] 实例

[0593] 以下是本发明一些方面的具体实施例的实例。提供这些实例仅出于说明目的，而无意以任何方式限制本发明的范围。就使用的数字(例如量、温度等)而言，已努力确保其准确性，但仍应允许有一些实验误差和偏差。

[0594] 缩写的含义如下："sec"意指秒、"min"意指分钟、"h"意指小时、"d"意指天、"μL"或"uL"或"μl"或"ul"意指微升、"mL"意指毫升、"L"意指升、"μM"意指微摩尔、"mM"意指毫摩尔、"M"意指摩尔、"mmol"意指毫摩尔、"μmole"或"umole"微摩尔、"g"意指克、"μg"或"ug"意指微克、"ng"意指纳克、"U"意指单位、"bp"意指碱基对、以及"kB"意指千碱基。

[0595] 实例1:Cas9直系同源物及其指导RNA的鉴定

[0596] 在该实例中，描述了从II型CRISPR(成簇的规律间隔的短回文重复序列)-Cas(CRISPR相关的)基因座鉴定Cas9蛋白及其相关的指导RNA的方法。

[0597] Cas9鉴定

[0598] II型Cas9内切核酸酶是通过首先使用PILER-CR(Edgar,R.C.(2007)BMC Bioinformatics.[BMC生物信息学]8:18)搜索公共序列库中指示细菌和古细菌的基于CRISPR-Cas核酸的自适应免疫系统的成簇的规律间隔的短回文重复序列(CRISPR)(Bhaya,D.等人(2011)Annu.Rev.Genet.[遗传学年度综述]45:273-97)的存在来鉴定的。鉴定CRISPR阵列后，检查CRISPR阵列周围的DNA区域(CRISPR阵列5′和3'的约20kb)是否存在编码大于750个氨基酸的蛋白的可读框(ORF)。接下来，为了鉴定与Cas9同源的CRISPR相关基因，使用MUSCLE(Edgar,R.C.(2004)Nucleic Acids Res.[核酸研究]32:1792-97)对来自不同Cas9内切核酸酶库的蛋白序列进行多序列比对，并如先前所述(Fonfara,I.等人(2014)Nucleic Acids Res.[核酸研究]42:2577-2590)使用HMMER(Eddy,S.R.(1998)Bioinformatics.[生物信息学]14:755-63和Eddy,S.R.(2011)PLoS Comput.Biol.[PLoS计算生物学]7:e1002195)将其用于为Cas9子家族建立谱隐马氏模型(HMM)。然后将所得的HMM用于针对与Cas9同源的cas基因的存在来搜索从CRISPR相关的ORF翻译的蛋白序列。仅包含关键的HNH和RuvC核酸裂解结构域以及定义II型Cas9蛋白的催化残基的蛋白(Nishimasu,H.等人(2014)Cell.[细胞]156:935-49)。通过比较分析，将Cas9蛋白解析为不同的家族，并且每个家族的代表性成员用于在MEGA7(Kumar,S.等人(2016)Mol.Biol.Evol.[分子生物学与进化]33:1870-74)情况下(利用邻近连接(Neighbor-Joining)(Saitou,N.等人(1987)Mol.Biol.Evol.[分子生物学与进化]4:406-25)和泊松校正(Zuckerkandl,E.等人(1965)Evol.genes proteins.[进化的基因和蛋白]97:97-166))方法构建系统发生树以计算进化史。

[0599] 根据系统发生距离，将代表675种II型Cas9序列(SEQ ID NO:86-170和511-1135)的系统发生树分为12个进化枝。然后选择蛋白以捕获Cas9直系同源物呈现的多样性(图1)。以约20％的速率挖掘产生具有阳性属性(例如，在真核细胞中的活性或目的原间隔子邻近基序(PAM)识别)的先前表征的Cas9蛋白的进化枝，而对其他所有进化枝以约10％的调查。总共选择了85个Cas9蛋白进行进一步表征(表1)。

[0600] 接下来，进行结构分析以进一步确认候选蛋白为Cas9直系同源物。首先，使用Ssearch36(Smith,T.F.和Waterman,M.S.(1981)J.Mol.Biol.[分子生物学杂志]147:195-97和Pearson,W.R.(1991)Genomics[基因组学]11:635-50)将整个序列与来自蛋白数据银行(Protein Data Bank)(PDB,蛋白数据银行H.M.Berman,J.Westbrook,Z.Feng,

G.Gilliland,T.N.Bhat,H.Weissig,I.N.Shindyalov,P.E.Bourne(2000)Nucleic Acids Research[核酸研究],28:235-242)的已知Cas9结构进行比对。然后,将最佳匹配结构用作模板,以根据已知Cas9中定义的结构性结构域分配功能结构域边界。基于与REC子结构域处变化最大的建模模板的相似性,所得到的结构比对产生了六个不同的组。

[0601] 将REC组I Cas9直系同源物(SEQ ID NO:93、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、136、137、138、139、140、141、143、144、145、146、148、158、160、161、162、142、168和169)与金黄色葡萄球菌Cas9结构PDB ID 5CZZ_A(“Crystal structure of Staphylococcus aureus Cas9[金黄色葡萄球菌Cas9的晶体结构]”,Nishimasu,H.,Cong,L.,Yan,W.X.,Ran,F.A.,Zetsche,B.,Li,Y.,Kurabayashi,A.,Ishitani,R.,Zhang,F.,Nureki,O.,(2015)Cell[细胞]162:1113-1126)比对。共有序列如图4所示,其中保守残基用黑体加下划线的文本(X)描绘。

[0602] REC组II(由单Cas9直向同源物表示,SEQ ID NO:96)在全长上与PDB:5czz比对,但是在RuvCIII结构域特征螺旋之前包含约312个氨基酸残基的新插入。这是该组的独特特点。

[0603] 将REC组III Cas9直系同源物(86、87、88、89、90、91、92、94、119、120、121、122、123、124、125、126、127、128、129、130、131、132、147、149、150、151、152、153、154、155、156、157、159、163、164、165、166、167和170)与酿脓链球菌血清型M1结构PDB ID 4UN3_B(“Structural Basis of Pam-Dependent Target DNA Recognition by the Cas9 Endonuclease[Cas9内切核酸酶对Pam依赖性靶DNA识别的结构基础]”,Anders,C.,Niewoehner,O.,Duerst,A.,Jinek,M.,(2014)Nature[自然]513:569-73)比对。共有序列如图5所示,其中保守残基用黑体加下划线的文本(X)描绘。

[0604] 将REC组IV Cas9直系同源物(SEQ ID NO:133和134)与内氏放线菌结构PDB ID 4OGE_A(“Structures of Cas9 endonucleases reveal RNA-mediated conformational activation[Cas9内切核酸酶的结构揭示了RNA介导的构象激活]”,Jinek,M.,Jiang,F.,Taylor,D.W.,Sternberg,S.H.,Kaya,E.,Ma,E.,Anders,C.,Hauer,M.,Zhou,K.,Lin,S.,Kaplan,M.,Iavarone,A.T.,Charpentier,E.,Nogales,E.,Doudna,J.A.,(2014)Science[科学]343:1247997)比对。共有序列如图6所示,其中保守残基用黑体加下划线的文本(X)描绘。组IV的共有序列特征是具有多个色氨酸残基,这是所检查的Cas9中的独特特征。

[0605] SEQ ID NO:95、96和135仅与已知结构模板部分地比对。因此,使用HHsearch(Soding,J.(2005)Bioinformatics[生物信息学].21:951-60)(一种谱-谱搜索程序)来扩展候选者-模板比对。SEQ ID NO:95(REC组V)与PDB:4oge完全对齐,并且SEQ ID NO:135(REC组VI)从头到尾与新凶手弗朗西丝菌(Francisella novicida)Cas9(PDB:5b2o)对齐。

[0606] 总而言之,序列属于Cas9家族,并按此顺序包含所有主要功能结构域:RuvCI、桥螺旋、REC、RuvCII、HNH、RuvCIII、WED和PI(表2A)。像其他已知的Cas9蛋白一样,序列长度变化,范围从约1,000到约1600个残基。表2B列出了每个Cas9直系同源物的每个结构域的SEQ ID。

[0607] 与系统发生分析相比,基于模板的方法将序列聚簇进入与其长度一致的组:例如,组I具有约1,100aa,组III具有约1,350aa。主要的序列长度变化发生在负责核苷酸链结合的REC结构域处。一致地,REC结构域也是Cas9蛋白超家族中最保守的序列区段。进化枝I-X

和组I-II-III-V彼此非常相似,形成一个家族,而对应于组IV的进化枝XI和对应于组VI的进化枝XII表现出更大的差异。

[0608] 指导RNA鉴定

[0609] 接下来,预测了能够与本文所述(表1)的Cas9直系同源物复合并对其进行指导以识别与合适的PAM(原间隔子邻近基序)相邻的DNA靶序列的一种或多种小RNA。首先,通过搜索cas9基因附近的区域(反重复序列,其与CRISPR重复序列碱基配对并且与一个或多个CRISPR阵列不同),鉴定了II型系统(Jinek,M.等人(2012)Science[科学].337:816-21和Karvelis,T.等人(2013)RNA Biol.[RNA生物学]10:20-19)中CRISPR RNA(crRNA)成熟(Deltcheva,E.等人(2011)Nature.[自然]471:602-7)和Cas9定向靶位点切割必不可少的反式激活RNA(tracrRNA)。一旦被鉴定,则通过检查二级结构(使用UNAfold(Markham,N.R.等人(2008)Methods Mol.Biol.[分子生物学方法]453:3-31))和RNA版本中存在的与反重复序列周围的有义和反义转录场景相对应的可能终止信号(如描述于Karvelis,T.等人(2015)Genome Biology.[基因组生物学]16:253中)来确定每个新系统的推定的一种或多种tracrRNA的可能转录方向。一旦预测到tracrRNA,就可以推导出crRNA的转录方向(因为tracrRNA必须以5′至3′方向与crRNA杂交)。根据指导RNA的预测,设计了代表crRNA和tracrRNA的非天然人工连接的单指导RNA(sgRNA)(Jinek,M.等人(2012)Science.[科学]337:816-21),并列在表中3。

[0610] 本研究中使用的所有sgRNA分子均使用TranscriptAid T7高产量转录试剂盒(赛默飞世尔科技公司(Thermo Fisher Scientific))通过体外转录合成,或直接在体外翻译(IVT)反应中转录。sgRNA转录的模板是通过PCR扩增合成的片段(IDT和金斯瑞公司(Genscript))生成的。

[0611] 实例2:确定Cas9直系同源物的原间隔子邻近基序要求和靶切割模式

[0612] 在该实例中,描述了快速表征原间隔子邻近基序(PAM)要求以及直系同源Cas9蛋白进行双链DNA靶切割的位置和类型(例如平端、5′突出端或3′突出端)的方法。

[0613] 为了确定支持DNA靶识别和切割的PAM序列,按照制造商推荐的方案,使用连续交换的1步人偶联IVT试剂盒(赛默飞世尔科技公司)或PURExpress细菌IVT试剂盒(新英格兰实验室公司(New England Biolabs))产生Cas9蛋白。这是通过首先产生编码Cas9直系同源物的质粒DNA来实现的。对于人偶联试剂盒,基因经过人密码子优化并合成(金斯瑞公司和推斯特生物科学公司(Twist Bioscience))到pT7-N-His-GST(赛默飞世尔科技公司)中。对于细菌IVT试剂盒,基因经过大肠杆菌密码子优化,合成(金斯瑞公司和推斯特生物科学公司)并克隆到pET28a(新英格兰生物实验室)表达盒中。

[0614] 在体外表达后,产生了Cas9核糖核蛋白(RNP)复合物。这通过首先在4℃下14,000g离心30分钟从反应清除碎片来进行。接下来,在1μl(40U)RiboLock RNA酶抑制剂(赛默飞世尔科技公司,美国)存在下,将20μl含可溶性Cas9蛋白的上清液立即与2μg T7转录的一种或多种指导RNA组合,并在室温下孵育15分钟。在某些情况下,通过提供包含T7启动子和编码相应sgRNA的序列的DNA模板,可在IVT反应中直接转录sgRNA。在这种情况下,Cas9-指导RNA核糖蛋白(RNP)复合物不再进行进一步处理,并且直接用于下一步。

[0615] 接下来,通过将10μl Cas9-指导RNA裂解混合物与90μl反应缓冲液(10mM Tris-HCl,在37℃下pH7.5,100mM NaCl和1mM DTT,10mM MgCl2)和1μg来自Karvelis等人2015的

包含T1靶序列的7bp随机PAM文库轻轻组合，进行随机PAM文库的消化。在37℃下1小时后，通过将反应与1μl(5U)的T4 DNA聚合酶和1μl的10mM dNTP混合物(赛默飞世尔科技公司，USA)在11℃下孵育20分钟，来使反应经受DNA末端修复。然后通过将其加热至75℃10分钟使反应失活。为了通过衔接子连接来有效捕获游离的DNA末端，通过将反应混合物与1μl(5U)DreamTaq聚合酶(赛默飞世尔科技公司，EP0701)在72℃下孵育30分钟来添加3′-dA突出端。然后通过在37℃下孵育1μl RNA酶A/T1(赛默飞世尔科技公司，美国)30分钟，从反应中去除过量的RNA。然后使用Monarch PCR&DNA Cleanup纯化柱(新英格兰实验室公司，美国)纯化所得的DNA。

[0616] 消化和末端修复后，随后通过衔接子连接来捕获支持切割的PAM序列。这通过以下完成：首先通过以下来制备具有3′-dT突出的衔接子：在95℃加热A1(5′-CGGCATTCCTGCTGAACCGCTCTTCCGATCT-3′(SEQ ID NO:1731))和磷酸化的A2(5′-GATCGGAAGAGCGGTTCAGCAGGAATGCCG-3′(SEQ ID NO:1732)寡核苷酸的等摩尔混合物5分钟并且在退火(A)缓冲液(10mM Tris-HCl，37℃下pH 7.5，50mM NaCl)中缓慢冷却(约0.1℃/s)至室温，使两者退火。然后通过以下将衔接子连接至末端修复的3′-dA突出切割产物：在25μl连接缓冲液(40mM Tris-HCl，在25℃pH 7.8，10mM MgCl2，10mM DTT，0.5mM ATP，5%(w/v)PEG 4000))中，将100ng的所述产物和衔接子与5U的T4连接酶(美国赛默飞世尔科技公司，美国)组合，并使反应在室温下进行1小时。

[0617] 接下来，分别使用R0(5′-GCCAGGGTTTTCCCAGTCACGA-3’(SEQ ID NO:1733))和特异于7bp PAM文库的A1寡核苷酸和衔接子富集含有PAM序列的切割的产物。使用具有高保真(HF)缓冲液(赛默飞世尔科技公司，美国)或Q5 DNA聚合酶(新英格兰实验室公司，美国)的Phusion高保真度PCR预混液，使用10μl的连接反应作为模板进行PCR。使用两步扩增方案(98℃-30s初始变性，98℃-15s，72℃-30s变性、退火和合成，15个循环，以及72℃-5分钟的最终延伸)。对于在不存在Cas9的情况下组装的样品，使用R0和C0引物(5′-GAAATTCTAAACGCTAAAGAGGAAGAGG-3’(SEQ ID NO:1734))对进行PCR，其中C0与原间隔子序列互补。接下来，使用Monarch PCR&DNA Cleanup纯化柱(新英格兰实验室公司，美国)纯化扩增产物(对于A1/R0和C0/R0引物对分别为148bp和145bp)。

[0618] 接着，将依诺米那(Illumina)深度测序所需的序列和索引掺入Cas9切割的DNA片段的末端，并对所得产物进行深度测序。这通过以下来完成：根据制造商的说明，使用HF缓冲液(新英格兰实验室公司，美国)中的Phusion高保真PCR预混液进行两轮PCR。使用20ng Cas9切割的衔接子连接的PAM侧模板组装一级PCR，并进行10个循环。该反应使用可以与衔接子杂交的正向引物F1(5′-CTACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGGCGGCATTCCTGCTGAAC-3’(SEQ ID NO:1735))和与PAM随机区域的3′位点结合的反向引物R1(5′-CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTCGGCGACGTTGGGTC-3′(SEQ ID NO:1736))。除了与衔接子连接的PAM片段杂交外，引物还包含从其5′末端延伸的依诺米那序列。对于正向引物，额外序列包括桥扩增所需的序列的一部分(5′-CTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3’(SEQ ID NO:1737))，之后是可互换的独特索引序列(5′-AAGG-3′)(如果同时测序，其允许对多个扩增子进行去卷积)。对于反向引物，另外的序列仅包含在扩增子的3′末端进行桥扩增所需的序列(5′-CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT-3’(SEQ ID NO:1738))。使用以下PCR循环条件：95℃-30s初始变性，95℃-10s，60℃-15s，72℃-5s变性、退火和合成，10个循环，以

及72℃-5min进行最终延伸。初次PCR后，使用2μl(总体积为50μl)的第一轮PCR作为模板进行第二轮PCR扩增。二级PCR中使用的正向引物F2(5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG-3’(SEQ ID NO:1739))与F1的5′区杂交，进一步延伸了依诺米那深度测序所需的序列。二级PCR中使用的反向引物R2(5′-CAAGCAGAAGACGGCATA-3’(SEQ ID NO:1740))仅与初级PCR扩增子的3′末端结合。使用以下PCR循环条件:95℃-30s初始变性，95℃-10s，58℃-15s，72℃-5s变性、退火和合成，10个循环，以及72℃-5min进行最终延伸。建立文库后，按照制造商的说明，使用QIAquick PCR纯化试剂盒(凯杰公司(Qiagen)，美国)纯化扩增产物，并以等摩尔浓度组合成单样品。接下来，将文库在MiSeq个人测序仪(依诺米那公司，美国)上进行单读深度测序，其中掺入为25％(v/v)的PhiX对照v3(依诺米那公司，美国)，并按照制造商的说明进行序列后处理和反卷积。请注意，初始PAM文库也已作为对照进行测序，以考虑会影响下游PAM分析的固有偏差。这如上所述进行，除了在初级PCR中使用正向引物C1(5′-CTACACTCTTTCCCTACACGACGCTCTTCCGATCTGGAATAAACGCTAAAGAGGAAGAGG-3’(SEQ ID NO:1741))代替F1，因为它直接与未剪切的PAM文库中的原间隔子区域杂交。

[0619] 接下来，评估PAM识别。这是通过以下来完成:首先生成代表靶区域内双链DNA切割和衔接子连接的所有可能结果的序列集合。例如，紧接在靶的第三位置之后的切割和衔接子连接将产生以下序列(5′-CTTCCGATCTACA-3’(SEQ ID NO:1742))，其中衔接子和靶序列分别包含5′-CTTCCGATCT-3′(SEQ ID NO:1743)和5′-ACA-3’。接下来，在序列数据集中搜索这些序列以及7bp PAM区域5’的10bp序列(5′-AGTTGACCCA-3’(SEQ ID NO:1744))。将其中过量回收Illumina序列(导致相比阴性对照的读段覆盖的峰或尖)的原间隔子-衔接子连接位置表示为切割位置(图9)。那些在前间隔子位置而不是紧接3之后产生显性切割的Cas9蛋白然后通过捕获由切割、末端修复、3’腺嘌呤添加和切割的文库靶的前间隔子侧的衔接子连接产生的切割产物来重新检查(图10A)。最后，然后将所得频率针对切割的原间隔子和PAM侧进行比较，并在考虑T4 DNA聚合酶末端修复情况下确定切割的位置和类型(图10B)。

[0620] 接下来，检查包含主要切割点的序列的PAM偏好。这是通过从这些读段中分离PAM序列并修剪掉5′和3′侧翼序列来完成的。接下来，将提取的PAM序列的频率归一化为初始PAM文库，以说明初始文库固有的偏差。首先，枚举相同的PAM序列，并计算相比于数据集中的总读段的频率。然后，使用以下方程式对每个PAM进行归一化，以说明初始文库中代表不足或代表过量的PAM序列:

[0621] 归一化的频率＝(处理频率)/(((对照频率)/(平均对照频率)))

[0622] 归一化后，计算位置频率矩阵(PFM)。这是通过根据与每个PAM相关的频率(归一化)对每个位置的每个核苷酸加权来完成的。例如，如果5′-CGGTAGC-3’的PAM的归一化频率为0.15％，则在确定第一PAM位置的核苷酸频率时，第一位置的C的频率将为0.15％。接下来，将数据集中每个位置处的每个核苷酸的总体贡献相加并整理成表，其中最丰富的核苷酸表明Cas9 PAM偏好(表4-83，其中:A＝腺嘌呤，C＝胞嘧啶，G＝鸟嘌呤，T＝胸腺嘧啶，R＝A或G，Y＝C或T，S＝G或C，W＝A或T，D＝A或G或T，H＝A或C或T，K＝G或T，M＝A或C，N＝任何碱基，B＝C或G或T，V＝A或C或G)并显示为WebLogo(图3)。

[0623] 用纯化的核糖核蛋白(RNP)以几种不同的浓度证实了IVT方法结果。选择的Cas9直系同源物的WebLogo比较如图8所示。

[0624] 总之，获得了不同的范围的PAM序列偏好。这些包括新颖的富含G、富含C、富含A和

富含T的PAM识别。此外，与其他Cas9典型的平端DNA靶切割模式相反，大约10％的Cas9直系同源物显示5′交错突出切割(1-3nt)。综上所述，Cas9直系同源物呈现的这种多样性提供了丰富的DNA靶识别和生物物理特性，其可用于基因组编辑应用。

[0625] 实例3：在大肠杆菌细胞中的表达分析

[0626] 在确定PAM要求和功能性sgRNA序列后，选择目的候选基因用于在大肠杆菌细胞中分析以及从大肠杆菌细胞中纯化。主要选择标准包括期望的或其他目的PAM、基因组编辑活性、异常切割模式和蛋白大小。将候选Cas9核酸酶编码基因亚克隆到大肠杆菌表达载体中，以产生编码包含C末端6-His标签的融合蛋白的构建体。在一些情况下，还将编码核定位序列(SV40起源)的序列掺入Cas9基因的5′和3′末端。表达分析可以在不同的大肠杆菌菌株中在各种生长条件下(培养基、温度、诱导)进行，并通过SDS-PAGE和蛋白印迹分析进行检测。当在大肠杆菌中表达时至少一些Cas9蛋白是可溶的，并且在纯化时是可溶且稳定的。可以选择优化的条件进行纯化。使用标准IMAC和离子交换色谱法从细胞裂解物中纯化蛋白。

[0627] 在烧瓶规模成功纯化的Cas9蛋白在高密度生物反应器中进行了表达试验。确定适用于GMP(良好生产规范)生产的可扩展纯化方案。使用纳米差示扫描荧光测定法(nanoDSF)和体外DNA内切核酸酶测定相结合，确定最佳的储存条件和纯化蛋白的稳定性。在经荧光末端标记的DNA片段上进行DNA内切核酸酶测定，并在96孔板中使用毛细管电泳进行检测和定量。

[0628] 实例4：用Cas9直系同源核酸酶修饰靶多核苷酸的体外方法

[0629] 本文公开的组合物可以在典型的细胞环境之外用于体外修饰一种或多种靶多核苷酸。在一些方面，从基因组来源分离并纯化靶多核苷酸。在一些方面，靶多核苷酸在环化或线性化质粒上。在一些方面，靶多核苷酸是PCR产物。在一些方面，靶多核苷酸是合成的寡核苷酸。

[0630] 在一些方面，所述修饰包括结合、切口或切割靶多核苷酸。

[0631] 材料

[0632] 使用了以下材料：

[0633] a.Cas9直系同源物多肽；cas9直系同源物多核苷酸；功能性Cas9直系同源物变体；功能性Cas9直系同源物片段；包含活性或失活的Cas9直系同源物的融合蛋白；Cas9直系同源物，其在C末端上或在N末端上或在N和C末端两者上进一步包含一个或多个核定位序列(NLS)；生物素化的Cas9直系同源物；Cas9直系同源物切口酶；Cas9直系同源物内切核酸酶；进一步包含组氨酸标签的Cas9直系同源物；具有不同PAM特异性的Cas9直系同源物的混合物；和上述任何两者或更多的混合物。

[0634] b.pH 6.5的10X反应缓冲液：200mM HEPES、50mM MgCl2、1M NaCl、1mM EDTA或支持活性的等效缓冲液

[0635] c.蛋白酶(例如，蛋白酶K，分子生物学等级，新英格兰生物实验室产品#P8107S)

[0636] d.无核酸酶的水

[0637] e.sgRNA或其他包含靶向靶(底物)多核苷酸的目标区域中的序列的指导多核苷酸，其中所述靶向序列与靶(底物)多核苷酸的靶序列的片段基本互补

[0638] f.靶(底物)多核苷酸，其包含靶序列

[0639] g.优选将每个靶位点的Cas9和sgRNA/指导多核苷酸摩尔比保持在1:1:1或更高，

以获得最佳切割效率。

[0640] 方法

[0641] 每个30ul反应在室温下组装：

[0642] 1．20ul无核酸酶的水

[0643] 2．3ul 10X反应缓冲液

[0644] 3.sgRNA或其他多肽

[0645] 4.材料章节中部分a中所述的Cas9直系同源物或其他分子。

[0646] 将混合物在25摄氏度(或支持核糖核蛋白复合物形成的其他温度)下孵育1分钟或更长时间。添加底物多核苷酸。将混合物充分混合并在微量离心机中脉冲旋转。将样品在37摄氏度(或支持最佳活性的其他温度)下孵育5分钟或更长时间。向每个样品中添加1ul蛋白酶，然后将其充分混合并在微量离心机中脉冲旋转。样品在室温下孵育10分钟，并准备进行后续分析。

[0647] 实例5:纯化的蛋白的体外表征

[0648] 适于制造的纯化的Cas9蛋白(包括期望的稳定性、溶解性和/或其他特性的蛋白)在体外进一步表征。首先，通过标准质粒DNA切割确认通过前述测定确定的PAM序列(Karvelis等人，2015)。使用具有最佳PAM和至少三个不同靶(不同CG含量)的质粒测试了每个Cas9的切割模式。使用体外DNA内切核酸酶测定和基于细胞的基因组编辑测定来确定下一个切割条件和最佳sgRNA结构。

[0649] 用两种不同长度的间隔子(20个核苷酸和24个核苷酸)测试的Cas9直系同源物中的一些的数据显示在图11中。

[0650] 选择相比于SpCas9表现出相似或更好的体外切割效率的变体进行另外的测试。表84总结了针对代表性数目的Cas9直系同源物获得的体外和体内切割数据。

[0651] 实例6:同源定向修复(HDR)活性评估

[0652] 确定了新颖Cas9直系同源物在体外在培养的人细胞和植物细胞中对某一种或多种靶的切割活性。基于细胞系的功能获得性荧光报告系统经工程改造用于评估由Cas9蛋白诱导的HDR效率。简而言之，eGFP基因通过插入含有针对各种新颖Cas9的多个终止密码子和PAM的区域而失活。可以测试两种方法(图7)：i)用于修复的同源臂(约500bp)在eGFP基因中重复；ii)将修复模板与Cas9一起引入细胞中。为了直接比较不同的Cas9蛋白，将转染效率和Cas9表达归一化。

[0653] 对绿色细胞的直接计数可以对HDR频率进行评分，而随后进行的T7内切核酸酶测定(或深度测序)可以评估同一细胞中的切割效率和NHEJ效率。这些实验导致选择新颖Cas9蛋白，其中切割修复输出转移到HDR。该系统的优点是可以直接比较Cas9核酸酶系统之间的HDR效率。评估Cas9直系同源物的生物物理特性，包括：平末端或粘性突出端DNA切割，靶位点释放和复发靶位点切割的频率。HDR分析结合体外DNA切割的详细表征有助于将Cas9核酸酶的生物物理特性与期望的HDR结果联系起来。

[0654] 实例7:用Cas9直系同源核酸酶对植物细胞靶多核苷酸进行体内修饰

[0655] 在一些方面，本文公开的组合物可用于修饰细胞的基因组中的靶多核苷酸。在一些方面，所述细胞是真核细胞。在真核细胞的一个实例中，使用植物细胞。用Cas9直系同源物转化真核细胞以实现基因组多核苷酸编辑可以通过已知在植物中有效的各种方法来完

成，所述方法包括粒子介导的递送、农杆菌介导的转化、PEG介导的递送和电穿孔。应当理解，可以利用本领域中已知的任何方法。实例方法如下所述。

[0656] 为了赋予有效表达，将新颖Cas9内切核酸酶基因按本领域已知的标准技术进行优化，并且引入马铃薯ST-LS1内含子2以便于消除该基因在大肠杆菌和农杆菌中的表达。为了促进在玉蜀黍细胞中的核定位，将编码两个版本的猿猴病毒40(SV40)单份核定位信号的核苷酸序列添加到5引发、3引发、或5引发和3引发两者的末端。然后将所得的编码不同的经优化的Cas9内切核酸酶基因和核定位信号变体的序列通过标准分子生物学技术可操作地连接至启动子，例如玉蜀黍遍在蛋白启动子、玉蜀黍遍在蛋白5′非翻译区(UTR)、玉蜀黍遍在蛋白内含子1和合适的终止子。

[0657] 用小RNA(本文中称为指导RNA)引导Cas9内切核酸酶，从而切割双链DNA。这些指导RNA包括辅助Cas9识别的序列(称为Cas9识别结构域)和用于通过与DNA靶位点的一条链碱基配对引导Cas9切割的序列(Cas9可变靶向结构域)。为了在玉蜀黍细胞中转录对于引导Cas9内切核酸酶切割活性必需的小RNA，将U6聚合酶III启动子和终止子从玉蜀黍分离，并且与在转录后将生成对于Cas9内切核酸酶而言适合的指导RNA的DNA序列的末端可操作地融合。为了促进指导RNA从玉蜀黍U6聚合酶III启动子的最佳转录，将一个G核苷酸添加至待转录的序列的5’末端。

[0658] 粒子介导的递送

[0659] 如下进行使用粒子递送转化玉蜀黍未成熟胚。培养基配方如下。

[0660] 将穗剥皮并在30％Clorox漂白剂加上0.5％微量洗涤剂中表面消毒20分钟，并用无菌水冲洗两次。将未成熟胚分离，并以每个平板25个胚将胚轴侧向下(盾片侧向上)放置于560Y培养基上持续4小时，并且然后排列在2.5-em靶区内准备进行轰击。可替代地，将分离的胚置于560L(起始培养基)上并在范围从26℃至37℃的温度下在黑暗中放置8至24小时，之后在26℃下放置于560Y中4小时，之后如上所述进行轰击。

[0661] 使用标准分子生物学技术构建包含Cas9直系同源物和供体DNA的质粒，并用含有发育基因ODP2(AP2结构域转录因子ODP2(胚珠发育蛋白2);US 20090328252 A1)和Wushel(US 2011/0167516)的质粒进行共同轰击。

[0662] 如下使用水溶性阳离子脂质转染试剂，将质粒和目的DNA沉淀到0.6微米(平均直径)金球粒上。使用1ug的质粒DNA和任选地用于共轰击的其他构建体(例如50ng(0.5ul)的包含发育基因ODP2(AP2结构域转录因子ODP2(胚珠发育蛋白2);US 20090328252 A1)和Wushel的各个质粒)，在冰上制备DNA溶液。向预混的DNA中添加在水中的20ul的制备的金粒子(15mg/ml)和5ul的水溶性阳离子脂质转染试剂并小心混合。将金粒子在微型离心机中以10,000rpm沉淀1分钟并去除上清液。用100ml的100％EtOH小心冲洗所得球粒，而不重悬球粒，并且小心去除EtOH冲洗剂。添加105ul的100％EtOH，并通过简短的超声处理将粒子进行重悬。然后，将10ul点在每个巨载剂的中心上，并在轰击前允许其干燥约2分钟。

[0663] 可替代地，使用氯化钙(CaCl2)沉淀程序，通过混合在水中的100ul的制备的钨粒子、Tris EDTA缓冲液中的10ul(1ug)DNA(1ug总DNA)、100ul 2.5M CaCl2、和10ul 0.1M亚精胺，将质粒和目的DNA沉淀到1.1um(平均直径)钨球粒上。混合下，将每种试剂顺序地添加至钨粒子悬浮液中。将最终混合物短暂超声处理，并且允许在恒定涡旋下温育10分钟。在沉淀期后，将管短暂离心，去除液体，并且用500ml 100％乙醇洗涤粒子，随后是30秒离心。再次

去除液体，并且添加105ul的100％乙醇到最终钨粒子球粒中。为了粒子枪轰击，将钨/DNA粒子短暂超声处理。将10ul的钨/DNA粒子点在每个巨载剂的中心上，此后在轰击前允许点的粒子干燥约2分钟。

[0664] 用Biorad氦气枪，在水平#4轰击样品板。所有样品接受在450PSI的单次射击，其中从每个制备的粒子/DNA的管中取总共十个等分试样。

[0665] 轰击后，将胚在26℃至37℃的温度范围下在560P(维持培养基)上孵育12至48小时，并且然后置于26℃。在5至7天后，将胚胎转移至含有3mg/升双丙氨膦的560R选择培养基上，并在26℃下每2周继代培养。在约10周的选择之后，将选择抗性愈伤组织克隆转移到288J培养基中以开始植物再生。在体细胞胚成熟(2-4周)后，将发育良好的体细胞胚转移到培养基上进行萌芽，并且转移到有光照的培养室中。在约7-10天后，将发育的小植物转移到试管中的272V不含激素的培养基中7-10天，直到小植物良好地生长。然后将植物转移到包含盆栽土壤的平托花盆(inserts in flats)(相当于2.5″盆)中，并在生长室中生长1周，随后在温室中再生长1-2周，然后转移到经典的600盆(1.6加仑)中并生长至成熟。对植物进行监测并对转化效率和/或再生能力的改变进行评分。

[0666] 起始培养基(560L)包含4.0g/l N6基础盐(西格玛公司(SIGMA)C-1416)、1.0ml/l埃里克松(Eriksson’s)维生素混合液(1000X西格玛公司(SIGMA)-1511)、0.5mg/l硫胺素HCl、20.0g/l蔗糖、1.0mg/l 2,4-D、以及2.88g/l L-脯氨酸(用D-I H2O定容，之后用KOH调节至pH 5.8)；2.0g/l结冷胶(在用D-I H2O定容之后添加)和8.5mg/l硝酸银(在将培养基灭菌并且冷却至室温后添加)。

[0667] 维持培养基(560P)包含4.0g/l N6基础盐(西格玛公司(SIGMA)C-1416)、1.0ml/l埃里克松(Eriksson’s)维生素混合液(1000X西格玛公司(SIGMA)-1511)、0.5mg/l硫胺素HCl、30.0g/l蔗糖、2.0mg/l2,4-D、以及0.69g/l L-脯氨酸(用D-I H2O定容，之后用KOH调节至pH 5.8)；3.0g/l结冷胶(在用D-I H2O定容之后添加)和0.85mg/l硝酸银(在将培养基灭菌并且冷却至室温后添加)。

[0668] 轰击培养基(560Y)包含4.0g/l N6基础盐(西格玛公司(SIGMA)C-1416)、1.0ml/l埃里克松(Eriksson’s)维生素混合液(1000X西格玛公司(SIGMA)-1511)、0.5mg/l硫胺素HCl、120.0g/l蔗糖、1.0mg/l 2,4-D、以及2.88g/l L-脯氨酸(用D-I H2O定容，之后用KOH调节至pH 5.8)；2.0g/l结冷胶(在用D-I H2O定容之后添加)和8.5mg/l硝酸银(在将培养基灭菌并且冷却至室温后添加)。

[0669] 选择培养基(560R)包含4.0g/l N6基础盐(西格玛公司(SIGMA)C-1416)、1.0ml/l埃里克松(Eriksson’s)维生素混合液(1000X西格玛公司(SIGMA)-1511)、0.5mg/l硫胺素HCl、30.0g/l蔗糖、以及2.0mg/l 2,4-D(用D-I H2O定容，之后用KOH调节至pH 5.8)；3.0g/l结冷胶(在用D-I H2O定容之后添加)和0.85mg/l硝酸银和3.0mg/l双丙氨膦(在对培养基进行灭菌并冷却至室温之后添加这两者)。

[0670] 植物再生培养基(288J)包含4.3g/l MS盐(GIBCO 11117-074)、5.0ml/l MS维生素储液(0.100g烟酸、0.02g/l硫胺素HCL、0.10g/l吡哆醇HCL、和0.40g/l甘氨酸，用精制的D-I H2O定容)(Murashige和Skoog(1962)Physiol.Plant.[植物生理学]15:473)、100mg/l肌醇、0.5mg/l玉米素、60g/l蔗糖、以及1.0ml/l的0.1mM脱落酸(用精制的D-I H2O定容，之后调节至pH 5.6)；3.0g/l结冷胶(在用D-I H2O定容之后添加)和1.0mg/l吲哚乙酸以及3.0mg/l双

丙氨膦(在将培养基灭菌并且冷却至60℃后添加)。

[0671] 无激素培养基(272V)包含4.3g/l MS盐(GIBCO 11117-074)、5.0ml/l MS维生素储液(0.100g/l烟酸、0.02g/l硫胺素HCL、0.10g/l吡哆醇HCL、和0.40g/l甘氨酸,用精制的D-I H2O定容)、0.1g/l肌醇、以及40.0g/l蔗糖(用精制的D-I H2O定容,之后调节pH至5.6);以及6g/l细菌用琼脂(在用精制的D-I H2O定容之后添加),灭菌并冷却至60℃。

[0672] 与质粒或RNA相比,将核糖核蛋白(ribonucleoprotein,RNP)递送至细胞(包括植物或动物细胞)具有几个优势。当完整的复合物被递送到细胞中时,DNA可以更快、更高效地被修饰。此外,在这种情况下,可以更严格地控制Cas9的浓度,从而有可能降低脱靶率。

[0673] 为了进行玉蜀黍转化,类似于先前描述(Svitashev等人2015和Karvelis等人2015),将Hi-Type II 8的粒子枪转化进入10天大的未成熟胚(IE)中。简而言之,利用TransIT-2020将DNA表达盒共沉淀在0.6μM(平均大小)的金粒子上,通过离心沉淀,用无水乙醇洗涤,然后通过超声重新分散。超声处理后,将10μl包被有DNA的金粒子装载到巨载剂上并风干。接下来,使用具有425lb/平方英寸破裂片的PDS-1000/He枪(伯乐公司(Bio-Rad))进行生物射弹转化。由于粒子枪转化会是高度可变的,所以也将编码青色荧光蛋白(CFP)的可视标志物DNA表达盒共递送,从而有助于均匀转化的IE的选择,并且一式三份进行每个处理。

[0674] 农杆菌介导的转化

[0675] 基本上如在Djukanovic等人(2006)Plant Biotech J[植物生物技术杂志]4:345-57中所描述地进行农杆菌介导的转化。简言之,将10-12日龄的未成熟胚(尺寸为0.8-2.5mm)从灭菌的仁切下并放置于液体培养基(4.0g/L N6基础盐(西格玛公司(Sigma)C-1416)、1.0ml/L埃里克松(Eriksson's)维生素混合液(西格玛公司(Sigma)E-1511)、1.0mg/L硫胺素HCl、1.5mg/L 2,4-D、0.690g/L L-脯氨酸、68.5g/L蔗糖、36.0g/L葡萄糖,pH 5.2)中。收集胚后,用1ml浓度为0.35-0.45OD550的农杆菌代替培养基。将玉蜀黍胚与农杆菌在室温下一起孵育5分钟,然后将混合物倾倒在培养基平板上,该培养基平板包含4.0g/LN6基础盐(西格玛公司(Sigma)C-1416)、1.0ml/L埃里克松(Eriksson's)维生素混合液(西格玛公司(Sigma)E-1511)、1.0mg/L硫胺素HCl、1.5mg/L 2,4-D、0.690g/L L-脯氨酸、30.0g/L蔗糖、0.85mg/L硝酸银、0.1nM乙酰丁香酮、以及3.0g/L结冷胶,pH 5.8。将胚在20℃在黑暗中轴向下地孵育3天,然后在28℃在黑暗中孵育4天,然后转移到新的培养基平板上,该培养基平板包含4.0g/L N6基础盐(西格玛公司(Sigma)C-1416)、1.0ml/L埃里克松(Eriksson's)维生素混合液(西格玛公司(Sigma)E-1511)、1.0mg/L硫胺素HCl、1.5mg/L 2,4-D、0.69g/L L-脯氨酸、30.0g/L蔗糖、0.5g/L MES缓冲液、0.85mg/L硝酸银、3.0mg/L双丙氨膦、100mg/L羧苄青霉素、以及6.0g/L琼脂,pH 5.8。将胚每三周进行继代培养,直到鉴定到转基因事件。通过将少量组织转移到再生培养基(4.3g/L MS盐(Gibco 11117)、5.0ml/L MS维生素储液、100mg/L肌醇、0.1μM ABA、1mg/LIAA、0.5mg/L玉蜀黍素、60.0g/L蔗糖、1.5mg/L双丙氨膦、100mg/L羧苄青霉素、3.0g/L结冷胶,pH 5.6)上来诱导体细胞胚发生,并在28℃下在黑暗中孵育两周。将所有具有可视芽和根的物质都转移到以下培养基上,该培养基包含4.3g/L MS盐(Gibco 11117)、5.0ml/L MS维生素储液、100mg/L肌醇、40.0g/L蔗糖、1.5g/L结冷胶(pH 5.6),并在28℃下在人造光下孵育。一周后,将小植物移入包含相同培养基的玻璃管中并生长直到它们被取样和/或移植到土壤中。

[0676] 核糖核蛋白转化

[0677] 可以重组表达和纯化Cas9和一种或多种相关的指导多核苷酸核糖核蛋白(RNP)复合物。RNP复合物装配可在重组表达组分的细胞中直接进行或在体外进行。纯化后,可以如Svitashev,S.等人(2016)Nat.Commun.[自然通讯]7:13274中所述通过粒子枪转化来递送一种或多种RNP复合物。简而言之,使用水溶性阳离子脂质TransIT-2020(米卢斯公司(Mirus),美国)将RNP(以及任选的DNA表达)沉淀到0.6mm(平均直径)的金粒子(伯乐公司,美国)上,如下:将50ml金粒子(10mg/ml的水悬浮液)和2ml的TransIT-2020水溶液添加到预混合的RNP(以及任选的DNA表达载体)中,轻轻混合,并在冰上孵育10分钟。然后将包被有RNP/DNA的金粒子在微型离心机中以8,000g沉淀30s,并除去上清液。然后通过短暂的超声处理将沉淀物重悬于50ml无菌水中。超声处理后,立即将包被的金粒子装载到微载剂(每个10ml)上并风干。授粉后8-10天,使用具有425磅/平方英寸的破裂压力的PDS-1000/He枪(伯乐公司,美国)轰击未成熟的玉蜀黍胚。如上所述,进行轰击后培养、选择和植物再生。

[0678] 递送方面的不同

[0679] Cas9和指导多核苷酸可以作为DNA表达盒、RNA、信使RNA(5′-带帽的和聚腺苷酸化的)或蛋白或其组合进行递送。还可以建立细胞系或转化体,以稳定地表达形成功能性指导多核苷酸/Cas复合物所需的全部组分中缺少的一种或多种组分,使得在递送所述一种或多种缺少的组分后,可以形成功能性指导多核苷酸/Cas复合物。

[0680] 基因组多核苷酸修饰的序列验证

[0681] 通过本领域已知的任何方法获得转化植物的样品并进行测序,并将其与未用Cas9和/或指导多核苷酸转化的同系植物的基因组序列进行比较。由DNA修复引起的非同源末端连接(NHEJ)插入和/或缺失(插入/缺失)突变的存在也可以用作检测切割活性的标志。

[0682] 这可以在转化后2天或更长时间进行。多种组织可以是样品,包括但不限于愈伤组织和叶组织。可以提取总基因组DNA,并可以用 Phusion® 高保真PCR预混合物(新英格兰生物实验室公司,M0531L)加上对于扩增子-特异性条形码以及依诺米那测序(使用“加尾的”引物)必要的序列通过两轮PCR对预期靶位点周围的区域进行PCR扩增并且进行深度测序。然后可以通过与其中从转录中省略小RNA转录盒的对照实验相比,检测所得的读段预期切割位点处是否存在突变。

[0683] 基因组多核苷酸修饰的序列验证

[0684] 如前所述(Svitashev等人2015和Karvelis等人2015),使用快速瞬时测定在玉蜀黍中评估了Cas9直系同源物的细胞切割活性。简要地,2天后,基于其荧光,收获20-30个最均匀转化的IE。提取总基因组DNA,并用 Phusion® 高保真PCR预混合物(新英格兰生物实验室公司,M0531L)加上对于扩增子-特异性条形码以及依诺米那测序(使用“加尾的”引物)必要的序列通过两轮PCR对预期靶位点周围的区域进行PCR扩增并且进行深度测序。然后通过与其中从转录中省略小RNA转录盒的对照实验相比,检测所得的读段预期切割位点处是否存在突变。

[0685] 图16显示了与用酿脓链球菌Cas9修饰的对照植物相比,玉蜀黍T0植物中跨三个不同靶位点(MS45、MS26和LIG)的两种不同Cas9直系同源物(ID33和ID64)的结果。图15和19显示了在玉蜀黍细胞中Cas9直系同源物ID33(图15A)、ID64(图15B)、ID46(图19A)和ID56(图

19B)的突变读段结果。

[0686] 实例8:用Cas9直系同源核酸酶对人细胞靶多核苷酸进行体内修饰

[0687] 在人模型细胞系HEK293中测量了选择的Cas9蛋白的基因组编辑活性。用编码Cas9候选物的质粒和编码其关联sgRNA的U6驱动的dsDNA共转染细胞。该方法不需要纯化的蛋白,并且一旦确定了支持切割活性的PAM偏好和sgRNA,即可启动该方法。靶向内源基因允许评估选择的Cas9在染色体DNA上的活性。使用T7内切核酸酶测定测试内源人基因的靶向频率,并且然后通过跨靶区域的深度PCR扩增子进行评估。对野生型和突变型扩增子计数以得出编辑得分。组合每个靶的编辑得分以获得总计得分。针对每种Cas9蛋白测试了三到五个不同的靶。在平行转染中,将选择的Cas9候选物的基因组编辑活性与SpCas9的活性进行比较。对于候选Cas9核酸酶,靶向附近或重叠(如果可能)的靶位置,使靶GC含量尽可能与SpCas9靶匹配。

[0688] 深度测序不仅可以允许比较所研究的Cas9蛋白的切割效率,而且还可以提供有关由每个新颖Cas9直系同源物产生的dsDNA断裂的主要NHEJ修复结果的有价值信息。RNP(核糖核蛋白,ribonucleoprotein)向细胞(包括植物或动物细胞)中的递送与质粒或RNA相比具有几个优势。当完整的复合物被递送到细胞中时,DNA可以更快、更高效地被修饰。此外,在这种情况下,可以更严格地控制Cas9的浓度,从而有可能降低脱靶率。为了验证新颖Cas9核酸酶在人细胞中的功能活性,使用纯化的蛋白和体外转录的sgRNA组装RNP复合物。通过电穿孔将RNP引入HEK293细胞。如上所述,使用T7内切核酸酶I测定和对应于基因组靶的扩增子的深度测序来评估基因组编辑活性。比较了新颖Cas9变体与SpCas9的基因组编辑效率。选择相比于带有相同NLS和His标签序列的SpCas9显示出相似或更好的基因组编辑效率的变体。这种方法可以预测当作为RNP引入模型细胞时新颖Cas9核酸酶的功能活性,这对于开发用于递送基因编辑工具的新方法很有用。

[0689] 细胞培养物电穿孔

[0690] 使用龙沙公司4D-Nucleofector系统和SF细胞系4D-Nucleofector®X试剂盒(龙沙公司)将Cas9 RNP电穿孔进入HEK293(ATCC目录号CRL-1573)细胞中。对于每次电穿孔,通过在室温下将100pmol sgRNA与50pmol Cas9蛋白在17μL体积的核转染溶液中孵育20分钟来形成RNP。将HEK293细胞使用TrypLE™Express Enzyme 1X(赛默飞世尔公司(ThermoFisher))从培养容器中释放,用不含Ca++或Mg++的1X PBS(赛默飞世尔公司)洗涤并使用XXX LUNA™自动细胞计数器(罗格斯生物系统公司(LogosBiosystems))XXX进行计数。对于每次电穿孔,将1x 10^5个活细胞重悬浮于9μL电穿孔溶液中。将细胞和RNP混合并转移到16孔带的一个孔中,并使用CM-130程序进行电穿孔。将75μL预热的培养物添加到每个孔中,并将10μL的得到的重悬浮的细胞分配到含有125μL预热培养基的96孔培养容器的孔中。在分析基因组编辑之前,将电穿孔的细胞在潮湿培养箱中在37℃、5%CO2孵育48小时。

[0691] 细胞培养物脂质转染

[0692] 人胚胎肾(HEK)细胞293(ATCC-CRL-1573)细胞在37℃和5%CO2孵育的情况下维持在具有GlutaMAX(赛默飞世尔科技公司)的杜尔贝科(Dulbecco)改良伊戈尔(Eagle)培养基(DMEM)中,所述培养基补充有10%胎牛血清(赛默飞世尔科技公司)和10,000单位/mL青霉素和10,000μg/mL链霉素(赛默飞世尔科技公司)。

[0693] 转染前一天，将HEK293细胞以每孔18,000个细胞的密度接种到96孔板(赛默飞世尔科技公司)中。按照制造商推荐的方案，使用Lipofectamine 3000(赛默飞世尔科技公司)转染细胞。对于96孔板的每个孔，总共使用200ng DNA，包含30fmol的质粒Cas9编码质粒和27fmol的具有适当的U6-gRNA模板的PCR片段。

[0694] 转染后，在基因组DNA提取之前，将细胞在5％CO2中于37℃孵育48小时。将细胞用200μl 1X DPBS(赛默飞世尔科技公司)洗涤两次并重悬浮于25μl 50mM Tris-HCl、150mM NaCl、0.05％Tween 20，pH 7.6(西格玛奥德里奇公司(Sigma Aldrich))和0.2mg/ml蛋白酶K(赛默飞世尔科技公司)裂解缓冲液中。将重悬浮的细胞在55℃孵育30分钟并且在98℃孵育20分钟。如上所述，使用引物X和Y对每个Cas9靶位点周围的基因组区域进行PCR扩增，并用T7内切核酸酶进行分析。

[0695] 基因组多核苷酸修饰的序列验证

[0696] 为了进行基因组编辑分析，根据制造商的建议，对于96孔培养容器的每个孔，使用50μL Epicenter QuickExtract™ DNA提取液在电穿孔后48小时提取基因组DNA。根据制造商的建议，使用Q5®热启动高保真2X预混液(NEB)，并在25μL反应中使用2μL基因组DNA(在水中以1:5稀释)，对预期靶位点周围的区域进行PCR扩增。

[0697] 使用T7内切核酸酶I测定评估基因组编辑。将每个PCR反应中的5μL与2μL NEBuffer 2(NEB)和12μL水混合，然后在95℃变性5分钟，并且然后通过以下进行重新退火：以-2℃/s从95℃-85℃温度斜变，然后以-0.1℃/s从85℃-25℃斜变。向每个重新退火的样品中添加1μL T7内切核酸酶I(NEB)，并将切割反应在37℃下孵育15分钟。通过在每个样品中添加1μL蛋白酶K(NEB)并在25℃下孵育5分钟来终止反应。使用CRISPR Discovery凝胶试剂盒试剂(AATI)在AATI片段分析仪(AATI)上分析片段。

[0698] 基因组编辑结果通过对来自靶基因座的PCR扩增子进行深度测序来进行表征。根据制造商的建议，使用对于Illumina®的NEBNext® Ultra™ II DNA文库制备试剂盒和对于Illumina®的NEBNext®多重寡核苷酸(96种索引引物)(NEB)构建依诺米那测序文库。测序后，通过与RNP靶向基因组的不同区域的对照实验相比较，针对在预期的切割位点处突变的存在检查了读段。

[0699] 图17显示了与酿脓链球菌Cas9的活性相比，用重组构建体(所述重组构建体包含编码相应Cas9直系同源物的DNA序列)转化的细胞中，选择的Cas9直系同源物在HEK细胞WTAP基因座处的结果。

[0700] 图18显示了与酿脓链球菌Cas9的活性相比，用重组构建体(所述重组构建体包含编码相应Cas9直系同源物的DNA序列)转化的细胞中，选择的Cas9直系同源物在HEK细胞RunX1基因座处的结果。

[0701] 图20显示了与酿脓链球菌Cas9的活性相比，用核糖核蛋白(所述核糖核蛋白包含各自Cas9直系同源物及其适当指导RNA)转化的细胞中，选择的Cas9直系同源物在HEK细胞WTAP基因座处的结果。

[0702] 实例9:分析Cas9直系同源物以鉴定关键残基、预测直系同源物活性，以及设计变体的方法

[0703] 确定了在活性Cas9中保守且在非活性Cas9中代表不足的氨基酸残基。这是通过使

用MUSCLE(默认参数)首先比对直系同源物来完成的。接下来,解析每个位置并评估每个位置处每个氨基酸的频率。接下来,分别通过求和并且除以每个数据集中的总数来定义活性和非活性数据集中每个位置处的每个氨基酸的总分率。然后,从活性数据(其中的正值表示活性Cas9中的在非活性集合中代表不足的保守氨基酸)中减去非活性数据集。最后,通过仅选择得分大于或等于+0.4的那些位置(其中7个活性Cas9中至少有5个展现出保守的和代表不足的氨基酸)来手动组织定义活性Cas9的关键位置(图21和表86A)。

[0704] 在定义了活性Cas9的一组结构特征(“指纹”)(表86B中列出的所有已鉴定的指纹位置)后,对Cas9直系同源物进行评分,作为位置得分的总和。本文所述方法的最高得分为12.52,最低得分为0。在评估了不同的Cas9集合后,评分范围从11.64到0.4。实验确定许多在真核细胞中有活性的Cas9在活性得分的前8%-10%。所有活性Cas9直系同源物都具有已鉴定的结构特征中的至少一种。表86C显示了本文公开的Cas9直系同源物中每个的计算的活性类别(通过SEQID)。得分大于中值(3.14)的直系同源物预计在真核细胞中具有阳性切割活性。其他直系同源物也可能具有活性。

[0705] 使用本文描述的方法,可以确定任何Cas9直系同源物的活性得分、结构指纹和类别。这些或类似方法可用于预测Cas9直系同源物的活性,定义活性Cas9所需的关键氨基酸和结构特征,定义负责粘性或平端切割活性的残基,并提供残基和区域以产生工程改造的变体。

[0706] 可以通过分析本文所述的Cas9直系同源物的序列-结构-功能关系来工程改造具有不同期望特性(例如但不限于:改变的PAM识别序列、经修饰的特异性和/或改变的切割活性)的Cas9直系同源物变体。在一些方面,分析了功能上重要的结构域(例如,PI结构域)的进化。在一些方面,关于保守和非保守氨基酸或氨基酸基序的信息用于预测Cas9直系同源物的活性并设计Cas9蛋白中可能调节活性或分子特性的可能的突变。在一些方面,使用合理的设计。在一些方面,使用随机诱变。在一些方面,使用定向进化。在一些方面,使用合理设计、随机诱变和定向进化的组合。

[0707] 在产生变体之后,选择并测试Cas9直系同源物变体以确定PAM序列、在培养的细胞(例如人或植物)中的活性,进行纯化和/或进一步表征。

[0708] 表

[0709] 表1:选择的用于表征的Cas9直系同源物

[0710] 列出了基因ORF和翻译后的编码蛋白的SEQ ID,整个Cas9蛋白系统发生进化枝,唯一ID#和源生物体。

[0711]

NT SEQID	PRT SEQID	直系同源物 ID#	进化枝	源生物体
1	86	**2**	1	栖组织普雷沃菌（*Prevotella histicola*）
2	87	**3**	1	禽金黄杆菌（*Chryseobacterium gallinarum*）
3	88	**4**	1	副拟杆菌属物种（*Parabacteroides sp.*）
4	89	**5**	1	犬碳酸噬胞菌（*Capnocytophaga canis*）
5	90	**6**	1	鼻气管炎鸟细菌（*Ornithobacterium rhinotracheale*）
6	91	**8**	1	有毒威克斯菌（*Weeksella virosa*）
7	92	**9**	1	冷黄杆菌（*Flavobacterium frigidarium*）
8	93	**12**	2	理研菌科物种（*Rikenellaceae sp.*）
9	94	**13**	2	*Jejuia pallidilutea*
10	95	**16**	3	*Caenispirillum salinarum*
11	96	**17**	3	*Salinispira pacifica*
12	97	**18**	3	东克拉亚硫酸杆菌（*Sulfitobacter donghicola*）
13	98	**19**	3	*Mucispirillum schaedleri*
14	99	**21**	3	中慢生根瘤菌属物种（*Mesorhizobium sp.*）
15	100	**27**	5	脑膜炎奈瑟氏菌（*Neisseria meningitidis*）
16	101	**28**	5	土芽孢杆菌属物种（*Geobacillus sp.*）
17	102	**29**	5	欧肯斯芽孢杆菌（*Bacillus okhensis*）
18	103	**30**	5	运动替斯崔纳菌（*Tistrella mobilis*）
19	104	**32**	5	金格金氏杆菌（*Kingella kingae*）
20	105	**33**	5	产气荚膜梭菌（*Clostridium perfringens*）
21	106	**35**	5	奈瑟氏菌属物种（*Neisseria sp.*）
22	107	**41**	5	结肠弯曲菌（*Campylobacter coli*）
23	108	**43**	5	硫磺单胞菌属物种（*Sulfurospirillum sp.*）
24	109	**44**	5	脱氮脱氯菌（*Dechloromonas denitrificans*）
25	110	**46**	6	*Nitratifractor salsuginis*
26	111	**47**	7	盲肠肠球菌（*Enterococcus cecorum*）
27	112	**48**	7	人费克蓝姆菌（*Facklamia hominis*）
28	113	**50**	7	中华链球菌（*Streptococcus sinensis*）

[0712]

29	114	**51**	7	*细长真杆菌（Eubacterium dolichum）*
30	115	**52**	7	*马克顿链球菌（Streptococcus macedonicus）*
31	116	**56**	7	*Turicibacter 属物种*
32	117	**60**	7	*尼亚美芽孢杆菌（Bacillus niameyensis）*
33	118	**61**	7	*Massilibacterium senegalense*
34	119	**63**	8	*Kurthia huakuii*
35	120	**64**	9	*马肠链球菌（Streptococcus equinus）*
36	121	**65**	9	*马链球菌（Streptococcus equi）*
37	122	**66**	9	*屎肠球菌（Enterococcus faecium）*
38	123	**67**	9	*意大利肠球菌（Enterococcus italicus）*
39	124	**68**	9	*无乳链球菌（Streptococcus agalactiae）*
40	125	**70**	9	*鼠链球菌（Streptococcus ratti）*
41	126	**71**	9	*单核细胞增多性李斯特氏菌（Listeria monocytogenes）*
42	127	**77**	10	*乳杆菌属物种（Lactobacillus sp.）*
43	128	**78**	10	*乳酸片球菌（Pediococcus acidilactici）*
44	129	**79**	10	*氨基酸球菌属物种（Acidaminococcus sp.）*
45	130	**80**	10	*乳杆菌属物种（Lactobacillus sp.）*
46	131	**81**	10	*恶臭螺旋体（Treponema putidum）*
47	132	**87**	10	*真杆菌属物种（Eubacterium sp.）*
48	133	**94**	11	*邦比双岐杆菌（Bifidobacterium bombi）*
49	134	**97**	11	*卡泼西斯棒状杆菌（Corynebacterium camporealensis）*
50	135	**102**	12	*嗜肺军团菌（Legionella pneumophila）*
51	136	**83**	1	*环境宏基因组（Environmental metagenome）*
52	137	**84**	1	*环境宏基因组、*
53	138	**85**	5	*环境宏基因组、*
54	139	**88**	5	*环境宏基因组、*
55	140	**91**	3	*环境宏基因组、*
56	141	**93**	3	*环境宏基因组、*
57	142	**139**	3	*环境宏基因组、*
58	143	**96**	5	*环境宏基因组、*
59	144	**98**	3	*环境宏基因组、*
60	145	**101**	3	*环境宏基因组、*
61	146	**103**	2	*环境宏基因组、*
62	147	**104**	1	*环境宏基因组、*
63	148	**105**	2	*环境宏基因组、*

[0713]

64	149	**106**	10	*肠氨基酸球菌（Acidaminococcus_intestini）_RyC-MR95*
65	150	**107**	8	*小球科里氏杆菌（Coriobacterium_glomerans）_PW2*
66	151	**108**	8	*埃格特菌属物种（Eggerthella_sp.）_YY7918*
67	152	**109**	10	*大芬戈尔德菌（Finegoldia_magna）_ATCC_29328*
68	153	**112**	10	*鼠李糖乳杆菌（Lactobacillus_rhamnosus）_LOCK900*
69	154	**116**	7	*鸡毒支原体（Mycoplasma_gallisepticum）_CA06*
70	155	**119**	9	*无乳链球菌（Streptococcus_agalactiae）_NEM316*
71	156	**120**	9	*停乳链球菌似马亚种（Streptococcus_dysgalactiae_subsp._equisimilis）_AC-2713*
72	157	**121**	9	*解没食子酸链球菌解没食子酸亚种（Streptococcus_gallolyticus_subsp._gallolyticus）_ATCC_43143*
73	158	**122**	7	*格氏链球菌卡尔斯株系 CH1 亚株系（Streptococcus_gordonii_str._Challis_substr._CH1）*
74	159	**123**	9	*变形链球菌（Streptococcus_mutans）_GS-5]*
75	160	**124**	7	*唾液链球菌（Streptococcus_salivarius）_JIM8777*
76	161	**125**	7	*猪链球菌（Streptococcus_suis）_D9*
77	162	**126**	7	*嗜热链球菌（Streptococcus_thermophilus）_LMG_18311*
78	163	**127**	10	*齿密螺旋体（Treponema_denticola）_ATCC_35405*
79	164	**131**	9	*动物乳杆菌（Lactobacillus animalis）KCTC 3501*
80	165	**132**	10	*西堤乳杆菌（Lactobacillus ceti）DSM 22408*
81	166	**136**	9	*Tissierellia 细菌 KA00581*
82	167	**138**	10	*小韦荣球菌（Veillonella parvula）ATCC 17745*
83	168	**141**	7	*解没食子酸链球菌（Streptococcus gallolyticus）*
84	169	**142**	7	*巴氏葡萄球菌（Staphylococcus pasteuri）*
85	170	**140**	9	*粪肠球菌（Enterococcus faecalis）OG1RF*

[0714]

表 2A：Cas9 直系同源结构域的氨基酸位置

Cas9 直系同源物通过在 REC 结构域具有最大变异的序列相似性分组。为了确定功能结构域边界，将组 I、II、III、IV、V 和 VI 的 Cas9 候选物序列与已知高分辨率 3D 结构（包括 PDBID：5czz、5czz、4un3、4ogc、4ogc 和 5b2o）的最接近同源序列进行比对。基于这些比对，将每个候选物序列插入到其相应的结构模板中进行建模，并根据相关出版物参考中模板的结构域定义来分配结构域边界。*表示 HNH 和 RuvCIII 结构域之间的未结构化的插入。

ID#	PRT SEQID	RuvCI 起始	RUVCI 结束	BH 起始	BH 结束	REC 起始	REC 结束	RuvCII 起始	RUVCII 结束	HNH 起始	HNH 结束	RUVCIII 起始	RUVCIII 结束	WED 起始	WED 结束	PI 起始	PI 结束
组 I																	
12	**93**	1	41	42	81	82	518	519	622	623	758	759	929	930	1035	1036	1053
18	**97**	1	40	41	78	79	458	459	558	559	681	682	824	825	925	926	1071
19	**98**	1	48	49	86	87	448	449	548	549	680	681	813	814	895	896	1044
21	**99**	1	51	52	89	90	503	504	605	606	743	744	887	888	946	947	1118
27	**100**	1	51	52	89	90	458	459	538	539	660	661	831	832	950	951	1082
28	**101**	1	39	40	77	78	456	457	534	535	656	657	804	805	925	926	1087
29	**102**	1	50	51	88	89	462	463	541	542	670	671	814	815	932	933	1074
30	**103**	1	47	48	85	86	450	451	538	539	662	663	819	820	900	901	1049
32	**104**	1	48	49	86	87	457	458	537	538	659	660	814	815	924	925	1060
33	**105**	1	43	44	81	82	455	456	535	536	655	656	823	824	938	839	1065
35	**106**	1	48	49	86	87	461	462	541	542	666	667	816	817	931	932	1069
41	**107**	1	36	37	74	75	439	440	521	522	638	639	784	785	837	838	1001
43	**108**	1	45	46	82	83	453	454	537	538	657	658	796	797	853	854	1048
44	**109**	1	39	40	77	78	474	475	570	571	697	698	863	864	981	982	1115
46	**110**	1	46	47	85	86	487	488	572	573	689	690	836	837	967	968	1137
47	**111**	1	42	43	76	77	462	463	543	544	683	684	824	825	973	974	1134
48	**112**	1	37	38	71	72	466	467	549	550	681	682	830	831	991	992	1142
50	**113**	1	39	40	73	74	462	463	542	543	677	678	822	823	966	967	1122
51	**114**	1	39	40	73	74	434	435	513	514	646	647	783	784	933	934	1091
52	**115**	1	40	41	74	75	461	462	542	543	677	678	823	824	968	969	1130
56	**116**	1	38	39	72	73	449	450	530	531	667	668	806	807	950	951	1107
60	**117**	1	41	42	75	76	451	452	530	531	662	663	799	800	926	927	1064
61	**118**	1	40	41	73	74	437	438	518	519	643	644	787	788	913	914	1063
83	**136**	1	58	59	100	101	456	457	515	516	679	680	792	793	905	906	1039

[0715]

84	**137**	1	44	45	88	89	622	623	674	675	834	835	978	979	1200	1201	1354
85	**138**	1	42	43	83	84	456	457	515	516	677	678	791	792	830	831	972
88	**139**	1	39	40	77	78	447	448	502	503	662	663	788	789	899	900	1046
91	**140**	1	43	44	87	88	482	483	558	559	715	716	842	843	964	965	1094
93	**141**	1	43	44	81	82	463	464	526	527	688	689	806	807	919	920	1037
139	**142**	1	39	40	82	83	600	601	653	654	822	*1150	1228	1229	1392	1393	1525
96	**143**	1	45	46	83	84	450	451	508	509	670	671	788	789	843	844	978
98	**144**	1	47	48	85	86	472	473	549	550	718	719	831	832	903	904	1037
101	**145**	1	42	43	80	81	448	449	505	506	674	675	789	780	908	909	1028
103	**146**	1	41	42	79	80	451	452	502	503	658	659	770	771	884	885	1008
105	**148**	1	45	46	87	88	511	512	571	572	735	736	846	847	997	998	1124
122	**158**	1	40	41	73	74	459	460	514	515	687	688	814	815	963	964	1136
124	**160**	1	40	41	73	74	466	467	521	522	694	695	819	820	969	970	1127
125	**161**	1	41	42	74	75	460	461	515	516	688	689	816	817	963	964	1122
126	**162**	1	40	41	73	74	460	461	515	516	688	689	813	814	964	965	1122
141	**168**	1	41	42	74	75	460	461	515	516	688	689	816	817	967	968	1130
142	**169**	1	41	42	74	75	430	431	485	486	652	653	774	775	909	910	1054
组 II																	
17	**96**	1	40	41	86	87	538	539	629	630	751	752	1208	1209	1322	1323	1458
组 III																	
2	**86**	1	58	59	94	95	637	638	692	693	852	853	1053	1054	1126	1127	1380
3	**87**	1	59	60	96	97	653	654	707	708	866	867	1014	1015	1147	1148	1403
4	**88**	1	58	59	94	95	669	670	724	725	881	882	1082	1083	1155	1156	1424
5	**89**	1	58	59	94	95	672	673	733	734	893	894	1099	1100	1172	1173	1430
6	**90**	1	59	60	94	95	695	696	755	756	962	963	1190	1191	1268	1269	1535
8	**91**	1	58	59	92	93	703	704	763	764	967	968	1189	1190	1208	1209	1440
9	**92**	1	58	59	93	94	612	613	674	675	829	830	1027	1028	1100	1101	1345
13	**94**	1	47	48	82	83	722	723	783	784	937	938	1104	1105	1167	1168	1459
63	**119**	1	44	45	77	78	719	720	774	775	930	931	1070	1071	1090	1091	1368
64	**120**	1	59	60	94	95	716	717	772	773	930	931	1112	1113	1156	1157	1375
65	**121**	1	59	60	94	95	715	716	771	772	922	923	1083	1084	1120	1121	1348
66	**122**	1	59	60	94	95	728	729	784	785	932	933	1090	1091	1127	1128	1340
67	**123**	1	59	60	94	95	720	721	776	777	924	925	1078	1079	1115	1116	1330

[0716]

68	**124**	1	59	60	94	95	731	732	787	788	942	943	1078	1079	1115	1116	1330
70	**125**	1	59	60	94	95	720	721	776	777	928	929	1101	1102	1138	1139	1370
71	**126**	1	76	77	105	106	730	731	786	787	937	938	1095	1096	1132	1133	1345
77	**127**	1	50	51	85	86	729	730	785	786	939	940	1081	1082	1124	1125	1365
78	**128**	1	48	49	83	84	729	730	784	785	938	939	1088	1089	1125	1126	1366
79	**129**	1	47	48	82	83	725	726	781	782	939	940	1068	1069	1103	1104	1358
80	**130**	1	50	51	85	86	747	748	804	805	967	968	1126	1127	1168	1169	1396
81	**131**	1	50	51	85	86	744	745	800	801	961	962	1096	1097	1159	1160	1395
87	**132**	1	53	54	88	89	727	728	784	785	946	947	1079	1080	1130	1131	1345
104	**147**	1	44	45	88	89	646	647	713	714	881	882	1039	1040	1253	1254	1399
106	**149**	1	46	47	77	78	715	716	777	778	941	942	1062	1063	1104	1105	1358
107	**150**	1	51	52	82	83	757	758	817	818	977	978	1124	1125	1169	1170	1384
108	**151**	1	50	51	81	82	754	755	813	814	970	971	1120	1121	1165	1166	1380
109	**152**	1	48	49	79	80	726	727	786	787	954	955	1079	1080	1129	1130	1348
112	**153**	1	49	50	80	81	720	721	782	783	941	942	1075	1076	1125	1126	1361
116	**154**	1	49	50	78	79	529	530	588	589	766	767	913	914	1102	1103	1269
119	**155**	1	47	48	89	90	707	708	766	767	930	931	1102	1103	1149	1150	1377
120	**156**	1	58	59	89	90	708	709	767	768	924	925	1096	1097	1140	1141	1371
121	**157**	1	59	60	91	92	710	711	769	770	933	934	1102	1103	1149	1150	1371
123	**159**	1	58	59	89	90	709	710	768	769	925	926	1076	1077	1123	1124	1345
127	**163**	1	49	50	80	81	733	734	796	797	963	964	1091	1090	1135	1136	1395
131	**164**	1	63	64	94	95	708	709	767	768	921	922	1065	1066	1109	1110	1318
132	**165**	1	51	52	82	83	743	744	806	807	968	969	1099	1100	1150	1151	1395
136	**166**	1	50	51	81	82	725	726	786	787	952	953	1089	1090	1149	1150	1400
138	**167**	1	63	64	94	95	747	748	809	810	979	980	1105	1106	1158	1159	1398
140	**170**	1	58	59	89	90	720	721	779	780	936	937	1081	1082	1125	1126	1337
									组 IV								
94	**133**	1	49	50	96	97	532	533	579	580	726	727	909	910	1025	1026	1239
97	**134**	1	41	42	88	89	470	471	517	518	672	673	820	821	913	914	1095
									组 V								
16	**95**	1	44	45	96	97	606	607	661	662	844	845	1000	1001	1103	1104	1442
									组 VI								
102	**135**	1	52	53	86	87	626	627	685	686	842	843	954	955	1184	1185	1372

[0717] 表2B:选择的Cas9直系同源物的结构域的SEQ ID

[0718]

Cas9 直系同源物 ID	REC 结构域 SEQID	RUVC1 结构域 SEQID	RUVC2 结构域 SEQID	RUVC3 结构域 SEQID	HNH 结构域 SEQID	WED 结构域 SEQID	PI 结构域 SEQID
2	1136	1221	1306	1391	1476	1561	1646
3	1137	1222	1307	1392	1477	1562	1647
4	1138	1223	1308	1393	1478	1563	1648
5	1139	1224	1309	1394	1479	1564	1649
6	1140	1225	1310	1395	1480	1565	1650
8	1141	1226	1311	1396	1481	1566	1651
9	1142	1227	1312	1397	1482	1567	1652
12	1143	1228	1313	1398	1483	1568	1653
13	1144	1229	1314	1399	1484	1569	1654
16	1145	1230	1315	1400	1485	1570	1655
17	1146	1231	1316	1401	1486	1571	1656
18	1147	1232	1317	1402	1487	1572	1657
19	1148	1233	1318	1403	1488	1573	1658
21	1149	1234	1319	1404	1489	1574	1659
27	1150	1235	1320	1405	1490	1575	1660
28	1151	1236	1321	1406	1491	1576	1661
29	1152	1237	1322	1407	1492	1577	1662
30	1153	1238	1323	1408	1493	1578	1663
32	1154	1239	1324	1409	1494	1579	1664
33	1155	1240	1325	1410	1495	1580	1665
35	1156	1241	1326	1411	1496	1581	1666
41	1157	1242	1327	1412	1497	1582	1667
43	1158	1243	1328	1413	1498	1583	1668
44	1159	1244	1329	1414	1499	1584	1669
46	1160	1245	1330	1415	1500	1585	1670
47	1161	1246	1331	1416	1501	1586	1671
48	1162	1247	1332	1417	1502	1587	1672
50	1163	1248	1333	1418	1503	1588	1673
51	1164	1249	1334	1419	1504	1589	1674
52	1165	1250	1335	1420	1505	1590	1675
56	1166	1251	1336	1421	1506	1591	1676
60	1167	1252	1337	1422	1507	1592	1677
61	1168	1253	1338	1423	1508	1593	1678
63	1169	1254	1339	1424	1509	1594	1679
64	1170	1255	1340	1425	1510	1595	1680

[0719]

65	1171	1256	1341	1426	1511	1596	1681
66	1172	1257	1342	1427	1512	1597	1682
67	1173	1258	1343	1428	1513	1598	1683
68	1174	1259	1344	1429	1514	1599	1684
70	1175	1260	1345	1430	1515	1600	1685
71	1176	1261	1346	1431	1516	1601	1686
77	1177	1262	1347	1432	1517	1602	1687
78	1178	1263	1348	1433	1518	1603	1688
79	1179	1264	1349	1434	1519	1604	1689
80	1180	1265	1350	1435	1520	1605	1690
81	1181	1266	1351	1436	1521	1606	1691
83	1182	1267	1352	1437	1522	1607	1692
84	1183	1268	1353	1438	1523	1608	1693
85	1184	1269	1354	1439	1524	1609	1694
87	1185	1270	1355	1440	1525	1610	1695
88	1186	1271	1356	1441	1526	1611	1696
91	1187	1272	1357	1442	1527	1612	1697
93	1188	1273	1358	1443	1528	1613	1698
94	1189	1274	1359	1444	1529	1614	1699
96	1190	1275	1360	1445	1530	1615	1700
97	1191	1276	1361	1446	1531	1616	1701
98	1192	1277	1362	1447	1532	1617	1702
101	1193	1278	1363	1448	1533	1618	1703
102	1194	1279	1364	1449	1534	1619	1704
103	1195	1280	1365	1450	1535	1620	1705
104	1196	1281	1366	1451	1536	1621	1706
105	1197	1282	1367	1452	1537	1622	1707
106	1198	1283	1368	1453	1538	1623	1708
107	1199	1284	1369	1454	1539	1624	1709
108	1200	1285	1370	1455	1540	1625	1710
109	1201	1286	1371	1456	1541	1626	1711
112	1202	1287	1372	1457	1542	1627	1712
116	1203	1288	1373	1458	1543	1628	1713
119	1204	1289	1374	1459	1544	1629	1714
120	1205	1290	1375	1460	1545	1630	1715
121	1206	1291	1376	1461	1546	1631	1716
122	1207	1292	1377	1462	1547	1632	1717
123	1208	1293	1378	1463	1548	1633	1718
124	1209	1294	1379	1464	1549	1634	1719

[0720]

125	1210	1295	1380	1465	1550	1635	1720
126	1211	1296	1381	1466	1551	1636	1721
127	1212	1297	1382	1467	1552	1637	1722
131	1213	1298	1383	1468	1553	1638	1723
132	1214	1299	1384	1469	1554	1639	1724
136	1215	1300	1385	1470	1555	1640	1725
138	1216	1301	1386	1471	1556	1641	1726
139	1217	1302	1387	1472	1557	1642	1727
140	1218	1303	1388	1473	1558	1643	1728
141	1219	1304	1389	1474	1559	1644	1729
142	1220	1305	1390	1475	1560	1645	1730

[0721] 表3:本文所述的Cas9直系同源物中的一些的sgRNA溶液及其组分(VT、crRNA重复序列、环、反重复序列和3’tracrRNA)的实例

[0722] 如本文所述的,sgRNA的可变靶向结构域可以变化,例如,但不限于从至少12个至30个核苷酸。如本文所述的,crRNA和反重复序列之间的环的长度可以从至少3个核苷酸至100个核苷酸变化。

[0723]

ID#	进化枝	ORF DNA SEQID	PRT SEQID	crRNA 重复序列 SEQID	反重复序列 SEQID	3' tracrRNA SEQID	sgRNA (CER 结构域) SEQID
2	1	1	86	171	256	341	426
3	1	2	87	172	257	342	427
4	1	3	88	173	258	343	428
5	1	4	89	174	259	344	429
6	1	5	90	175	260	345	430
8	1	6	91	176	261	346	431
9	1	7	92	177	262	347	432
12	2	8	93	178	263	348	433
13	2	9	94	179	264	349	434
16	3	10	95	180	265	350	435
17	3	11	96	181	266	351	436
18	3	12	97	182	267	352	437
19	3	13	98	183	268	353	438
21	3	14	99	184	269	354	439
27	5	15	100	185	270	355	440
28	5	16	101	186	271	356	441
29	5	17	102	187	272	357	442
30	5	18	103	188	273	358	443

[0724]

32	5	19	104	189	274	359	444
33	5	20	105	190	275	360	445
35	5	21	106	191	276	361	446
41	5	22	107	192	277	362	447
43	5	23	108	193	278	363	448
44	5	24	109	194	279	364	449
46	6	25	110	195	280	365	450
47	7	26	111	196	281	366	451
48	7	27	112	197	282	367	452
50	7	28	113	198	283	368	453
51	7	29	114	199	284	369	454
52	7	30	115	200	285	370	455
56	7	31	116	201	286	371	456
60	7	32	117	202	287	372	457
61	7	33	118	203	288	373	458
63	8	34	119	204	289	374	459
64	9	35	120	205	290	375	460
65	9	36	121	206	291	376	461
66	9	37	122	207	292	377	462
67	9	38	123	208	293	378	463
68	9	39	124	209	294	379	464
70	9	40	125	210	295	380	465
71	9	41	126	211	296	381	466
77	10	42	127	212	297	382	467
78	10	43	128	213	298	383	468
79	10	44	129	214	299	384	469
80	10	45	130	215	300	385	470
81	10	46	131	216	301	386	471
87	10	47	132	217	302	387	472
94	11	48	133	218	303	388	473
97	11	49	134	219	304	389	474
102	12	50	135	220	305	390	475
83	1	51	136	221	306	391	476
84	1	52	137	222	307	392	477
85	5	53	138	223	308	393	478
88	5	54	139	224	309	394	479
91	3	55	140	225	310	395	480
93	3	56	141	226	311	396	481
139	3	57	142	227	312	397	482

[0725]

96	5	58	143	228	313	398	483
98	3	59	144	229	314	399	484
101	3	60	145	230	315	400	485
103	2	61	146	231	316	401	486
104	1	62	147	232	317	402	487
105	2	63	148	233	318	403	488
106	10	64	149	234	319	404	489
107	8	65	150	235	320	405	490
108	8	66	151	236	321	406	491
109	10	67	152	237	322	407	492
112	10	68	153	238	323	408	493
116	7	69	154	239	324	409	494
119	9	70	155	240	325	410	495
120	9	71	156	241	326	411	496
121	9	72	157	242	327	412	497
122	7	73	158	243	328	413	498
123	9	74	159	244	329	414	499
124	7	75	160	245	330	415	500
125	7	76	161	246	331	416	501
126	7	77	162	247	332	417	502
127	10	78	163	248	333	418	503
131	9	79	164	249	334	419	504
132	10	80	165	250	335	420	505
136	9	81	166	251	336	421	506
138	10	82	167	252	337	422	507
141	7	83	168	253	338	423	508
142	7	84	169	254	339	424	509
140	9	85	170	255	340	425	510

[0726] 表4:ID2进化枝1的前间隔子邻近基序(PAM)偏好

[0727] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0728]

		PAM 位置						
		1	**2**	**3**	**4**	**5**	**6**	**7**
核苷酸	G	36.14%	21.25%	[54.36%]	0.16%	0%	[91.52%]	7.65%
	A	7.44%	[78.48%]	/45.64%/	/46.12%/	/48.14%/	3.33%	6.68%
	T	24.12%	0%	0%	/46.68%/	34.78%	3.08%	28.66%
	C	32.30%	0.27%	0%	7.04%	17.07%	2.07%	/57.01%/
共有序列		**N**	**A**	**R (G>A)**	**W**	**H**	**G**	**N**

[0729]

						(A>T>C)		**(C>T>R)**

[0730] 表5:ID3进化枝1的前间隔子邻近基序(PAM)偏好

[0731] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0732]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	28.58%	23.58%	[55.97%]	1.33%	0.02%	1.83%	16.89%
	A	10.31%	/57.81%/	/40.56%/	11.2%	2.37%	0.26%	24.79%
	T	13.88%	2.88%	0%	[77.09%]	[81.69%]	[85.73%]	/42.4%/
	C	/47.23%/	15.73%	3.47%	10.38%	15.93%	12.18%	15.92%
共有序列		N (C>D)	V (A>S)	R (G>A)	T	T	T	N (T>V)

[0733] 表6:ID4进化枝1的前间隔子邻近基序(PAM)偏好

[0734] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0735]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	30.63%	33.91%	9.17%	0.12%	0.19%	0.08%	8.43%
	A	15.52%	/53.21%/	20.43%	5.77%	4.39%	0.43%	6.52%
	T	22.02%	3.04%	[60.65%]	[85.47%]	[72.35%]	[90.08%]	[73.38%]
	C	31.83%	9.84%	9.75%	8.64%	23.07%	9.4%	11.67%
共有序列		N	V (A>G>C)	T	T	T	T	T

[0736] 表7:ID5进化枝1的前间隔子邻近基序(PAM)偏好

[0737] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0738]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	30.31%	31.67%	7.44%	0.01%	0.01%	0%	4.94%
	A	17.59%	[60.32%]	19.98%	2.08%	1.74%	0.09%	4.29%
	T	28.33%	1.01%	[63.72%]	[93.23%]	[90.31%]	[97.29%]	[83.28%]
	C	23.77%	7%	8.86%	4.68%	7.94%	2.62%	7.48%
共有序列		N	A	T	T	T	T	T

[0739] 表8:ID6进化枝1的前间隔子邻近基序(PAM)偏好

[0740] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0741]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	24.08%	8.85%	9.57%	6.63%	10.8%	/52.38%/	26.21%
	A	20.44%	33.32%	[89.83%]	[82.42%]	[61.84%]	35.19%	25.1%
	T	18.01%	26.95%	0.56%	0%	8.44%	5.22%	22.01%
	C	37.48%	30.88%	0.05%	10.95%	18.91%	7.21%	26.68%
共有序列		N	N (H>G)	A	A	A	N (G>A>Y)	N

[0742] 表9:ID8进化枝1的前间隔子邻近基序(PAM)偏好

[0743] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0744]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	10.17%	6.73%	0.89%	1.22%	2.56%	3.05%	22.15%
	A	23.01%	27.71%	[99.11%]	[98.51%]	[94.16%]	4.91%	37.94%
	T	/42.68%/	33.86%	0%	0.24%	0.13%	[86.66%]	26.05%
	C	24.14%	31.70%	0%	0.03%	3.15%	5.37%	13.85%
共有序列		N (T>V)	N	A	A	A	T	N

[0745] 表10:ID9进化枝1的前间隔子邻近基序(PAM)偏好

[0746] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0747]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	27.23%	12.35%	35.91%	5.65%	0%	29.72%	31.39%
	A	9.6%	[83.2%]	/48.04%/	19.98%	0%	21.22%	9.29%
	T	24.91%	0.73%	4.92%	[70.58%]	0%	12.79%	30.15%
	C	38.26%	3.72%	11.13%	3.79%	[100%]	36.27%	29.17%
共有序列		N	A	V (A>G>C)	T	C	N	N

[0748] 表11:ID12进化枝2的前间隔子邻近基序(PAM)偏好

[0749] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0750]

		PAM 位置

[0751]

		1	2	3	4	5	6	7
核苷酸	G	21.92%	20.6%	14.54%	21.79%	0%	0%	6.48%
	A	21.26%	/46.96%/	26.87%	38.08%	0%	0%	8.92%
	T	23.77%	8.06%	27.05%	34.31%	0%	0%	/44.69%/
	C	33.04%	24.38%	31.54%	5.82%	[100%]	[100%]	39.92%
共有序列		N	N (A>S>T)	N	N (W>G>C)	C	C	N (Y>R)

[0752] 表12:ID13进化枝2的前间隔子邻近基序(PAM)偏好

[0753] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0754]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	25.31%	23.72%	2.93%	3.87%	0%	0%	25.89%
	A	15.05%	37.23%	[97.02%]	24.57%	[93.86%]	0%	28.74%
	T	30.05%	12.64%	0%	/45.21%/	3.67%	12.01%	23.85%
	C	29.59%	26.41%	0.05%	26.35%	2.48%	[87.99%]	21.52%
共有序列		N	N	A	H (T>M)	A	C	N

[0755] 表13:ID16进化枝3的前间隔子邻近基序(PAM)偏好

[0756] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0757]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	14.16%	[93.5%]	1.83%	[85.98%]	0.16%	33.41%	26.87%
	A	26.12%	3.56%	13.32%	11.24%	[86.61%]	11.29%	23.92%
	T	24.65%	0.3%	[64.11%]	2.68%	2.69%	33.07%	30.21%
	C	35.07%	2.65%	20.73%	0.1%	10.54%	22.23%	19.01%
共有序列		N	G	T	G	A	N	N

[0758] 表14:ID17进化枝3的前间隔子邻近基序(PAM)偏好

[0759] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0760]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	31.01%	1.81%	[48.09%]	20.51%	0.22%	1.27%	24.04%
	A	10.3%	[97.24%]	[51.62%]	/41.94%/	[96.02%]	1.54%	35.49%
	T	37.06%	0.42%	0%	29.98%	0.04%	[92.67%]	16.87%

[0761]

	C	21.62%	0.54%	0.29%	7.58%	3.73%	4.52%	23.59%
共有序列		N	A	R	N (A>K>C)	A	T	N

[0762] 表15:ID18进化枝3的前间隔子邻近基序(PAM)偏好

[0763] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/

表示弱PAM偏好。

[0764]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	22.25%	/53.26%/	[53.02%]	22.86%	0%	7.32%	23%
	A	18.57%	35.41%	[46.92%]	28.78%	0.45%	0.12%	34.66%
	T	26.14%	0%	0	25.08%	[98.68%]	[92.53%]	27.46%
	C	33.04%	11.33%	0.06	23.27%	0.87%	0.03%	14.88%
共有序列		N	V (G>A>C)	R	N	T	T	N

[0765] 表16:ID19进化枝3的前间隔子邻近基序(PAM)偏好

[0766] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0767]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	24.51%	6.95%	/42.48%/	34.06%	0%	0%	35.8%
	A	14.06%	/50.32%/	/48.28%/	/43.01%/	6.8%	0%	31.95%
	T	29.38%	17.63%	1%	16.44%	0%	3.89%	16.29%
	C	32.06%	25.1%	8.24%	6.5%	[93.2%]	[96.11%]	15.95%
共有序列		N	N (A>B)	R	N (A>G>T >C)	C	C	N

[0768] 表17:ID27进化枝5的前间隔子邻近基序(PAM)偏好

[0769] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0770]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	27.54%	12.25%	24.63%	11.40%	0%	0%	3.11%
	A	19.03%	/41.8%/	37.36%	19.92%	0%	0%	/55.4%/
	T	20.49%	27.98%	24.88%	/54.55%/	0%	0.30%	23.50%
	C	32.95%	17.97%	13.13%	14.13%	[100%]	[99.7%]	18%

[0771]

共有序列		N	N (A>B)	N	N (T>V)	C	C	H (A>Y)

[0772] 表18:ID28进化枝5的前间隔子邻近基序(PAM)偏好

[0773] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0774]

		PAM 位置							
		1	2	3	4	5	6	7	8
核苷酸	G	20.1%	13.69%	8.1%	10.23%	0.5%	27.01%	0.38%	0.52%
	A	24.09%	26.66%	25.49%	29.16%	0.1%	32.22%	[95.74%]	[99.03%]
	T	24.69%	26.9%	32.15%	26.02%	0%	39.55%	0.44%	0.39%
	C	31.12%	32.76%	34.25%	34.59%	[99.39%]	1.22%	3.44%	0.07%
共有序列		N	N	N (H>G)	N	C	D	A	A

[0775] 表19:ID29进化枝5的前间隔子邻近基序(PAM)偏好

[0776] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0777]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	20.24%	6.48%	32.16%	[91.37%]	[93.46%]	24.58%	15.75%
	A	16.76%	26.36%	/40.8%/	5.83%	6.54%	30.98%	/48.29%/
	T	24.40%	31.57%	25.32%	2.70%	0%	39.92%	24.16%
	C	38.60%	35.58%	1.71%	0.09%	0%	4.52%	11.80%
共有序列		N	N (H>G)	D (A>K)	G	G	D	N (A>B)

[0778] 表20:ID30进化枝5的前间隔子邻近基序(PAM)偏好

[0779] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0780]

		PAM 位置							
		1	2	3	4	5	6	7	8
核苷酸	G	17.53%	11.24%	16.65%	15.25%	0.00%	0.00%	0.00%	[97.99%]
	A	21.12%	26.13%	29.25%	29.16%	30.95%	2.88%	[100.00%]	0.84%
	T	28.26%	30.76%	36.33%	33.24%	0.00%	3.18%	0.00%	0.35%
	C	33.09%	31.88%	17.77%	22.36%	[69.05%]	[93.94%]	0.00%	0.82%
共有序列		N	N	N	N	C	C	A	G

[0781] 表21:ID32进化枝5的前间隔子邻近基序(PAM)偏好

[0782] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0783]

		PAM 位置							
		1	2	3	4	5	6	7	8
核苷酸	G	21.46%	5.68%	11.12%	13.79%	0.00%	0.93%	1.59%	5.92%
	A	14.73%	36.25%	29.20%	26.40%	0.00%	2.40%	[64.92%]	[80.85%]
	T	25.36%	27.28%	34.96%	28.56%	0.00%	[60.92%]	33.49%	5.07%
	C	38.45%	30.79%	24.71%	31.25%	[100.00%]	35.76%	0.00%	8.16%
共有序列		N	N	N	N	C	T	A	A

[0784] 表22:ID33进化枝5的前间隔子邻近基序(PAM)偏好

[0785] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/

表示弱PAM偏好。

[0786]

		PAM 位置									
		1	2	3	4	5	6	7	8	9	10
核苷酸	G	22.09%	7.14%	14.13%	11.46%	0.00%	29.62%	[98.54%]	8.83%	14.01%	19.37%
	A	5.88%	31.83%	30.44%	34.78%	0.00%	39.89%	1.32%	[72.61%]	/51.42%/	31.58%
	T	29.82%	32.90%	29.77%	22.67%	0.00%	0.02%	0.14%	13.59%	16.89%	26.71%
	C	/42.21%/	28.12%	25.67%	31.08%	[100.00%]	30.47%	0.00%	4.96%	17.68%	22.34%
共有序列		N	N	N	N	C	V	G	A	N	N

[0787] 表23:ID35进化枝5的前间隔子邻近基序(PAM)偏好

[0788] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0789]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	22.03%	9.34%	25.15%	17.12%	0%	0%	22.47%
	A	14.56%	39.21%	35.63%	9.50%	0%	0%	25.37%
	T	22.33%	24.30%	21.03%	[71.71%]	0%	0%	36.60%
	C	/41.08%/	27.15%	18.19%	1.66%	[100%]	[100%]	15.57%
共有序列		N (C>D)	N	N	T	C	C	N

[0790] 表24:ID41进化枝5的前间隔子邻近基序(PAM)偏好

[0791] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0792]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	19.6%	16.88%	11.98%	35.91%	0.2%	0.23%	0.8%
	A	26.01%	25.05%	30.09%	23.09%	1.17%	0.01%	[97.57%]
	T	25.84%	26.95%	35.06%	9.22%	0%	[97.83%]	0.23%
	C	28.54%	31.12%	22.86%	31.78%	[98.63%]	1.93%	1.4%
共有序列		N	N	N	N	C	T	A

[0793] 表25:ID44进化枝5的前间隔子邻近基序(PAM)偏好

[0794] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0795]

		PAM 位置							
		1	2	3	4	5	6	7	8
核苷酸	G	19.80%	7.57%	11.08%	15.61%	[98.54%]	0.00%	0.00%	0.16%
	A	17.69%	38.78%	29.27%	22.89%	1.46%	0.00%	[93.02%]	[98.91%]
	T	23.27%	23.76%	27.37%	30.29%	0.00%	[45.31%]	6.98%	0.83%
	C	39.24%	29.90%	32.27%	31.22%	0.00%	[54.69%]	0.00%	0.10%
共有序列		N	N	N	N	C	Y	A	A

[0796] 表26:ID46进化枝6的前间隔子邻近基序(PAM)偏好

[0797] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0798]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	26.51%	25.76%	[97.21%]	[37.66%]	[73.44%]	28.66%	8.28%
	A	16.02%	[70.60%]	2.08%	[44.79%]	16.96%	24.92%	2.22%
	T	12.28%	0.00%	0.01%	0.60%	8.66%	31.22%	[47.73%]
	C	/45.19%/	3.64%	0.70%	16.96%	0.94%	15.20%	[41.77%]
共有序列		N	A	G	R	G	N	Y

[0799] 表27:ID47进化枝7的前间隔子邻近基序(PAM)偏好

[0800] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0801]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	21.09%	14.51%	[96.97%]	1.68%	0.47%	1.22%	6.41%
	A	21.36%	31.40%	2.71%	/46.42%/	[91.5%]	[98.06%]	[80.67%]

[0802]

	T	25.16%	29.52%	0.13%	/39.18%/	0.91%	0.56%	7.71%
	C	32.39%	24.57%	0.19%	12.72%	7.12%	0.16%	5.21%
共有序列		N	N	G	H (W>C)	A	A	A

[0803] 表28:ID48进化枝7的前间隔子邻近基序(PAM)偏好

[0804] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0805]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	25.12%	13.23%	[96.52%]	2.72%	1.12%	2.51%	27.13%
	A	19.76%	37.09%	1.57%	[95.9%]	[90.8%]	[95.87%]	31.21%
	T	27.23%	32.68%	1.52%	0.02%	0.04%	0.52%	22.90%
	C	27.89%	17%	0.39%	1.36%	8.04%	1.11%	18.75%
共有序列		N	N	G	A	A	A	N

[0806] 表29:ID50进化枝7的前间隔子邻近基序(PAM)偏好

[0807] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0808]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	18.16%	9.71%	2.12%	1.86%	0.48%	0.87%	19.56%
	A	15.19%	25.57%	[97.47%]	[97.38%]	[98.98%]	[98.68%]	[61.85%]
	T	36.44%	35.35%	0.03%	0%	0%	0.13%	11.97%
	C	30.21%	29.37%	0.38%	0.76%	0.54%	0.32%	6.62%
共有序列		N	N	A	A	A	A	A

[0809] 表30:ID51进化枝7的前间隔子邻近基序(PAM)偏好

[0810] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0811]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	23.52%	1.72%	[99.37%]	9.50%	/39.07%/	5.89%	7.91%
	A	21.33%	6.50%	0.58%	[89.72%]	[59.06%]	/45.26%/	9.79%
	T	25.10%	[65.77%]	0.01%	0%	1.05%	23.46%	/39.29%/
	C	30.05%	26.02%	0.04%	0.78%	0.82%	25.40%	/43.01%/
共有序列		N	T	G	A	R (G>A)	N (A>Y>G)	N (Y>R)

[0812] 表31:ID52进化枝7的前间隔子邻近基序(PAM)偏好

[0813] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0814]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	18.00%	4.92%	[87.99%]	1.61%	18.62%	13.60%	12.07%
	A	20.27%	34.84%	11.02%	6.15%	/53.71%/	[69.84%]	/52.19%/
	T	18.20%	20.00%	0.00%	/55.44%/	13.96%	12.71%	21.31%
	C	/43.53%/	/40.24%/	0.99%	36.80%	13.72%	3.85%	14.44%
共有序列		N (C>D)	H (C>W)	G	H (Y>A)	N(A>B)	A	N(A>T>S)

[0815] 表32:ID56进化枝7的前间隔子邻近基序(PAM)偏好

[0816] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0817]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	18.78%	15.33%	4.88%	11.14%	18.77%	0.21%	20.14%
	A	23.55%	25.44%	[91.9%]	[82.72%]	[76.54%]	8.37%	33.96%
	T	27.99%	29.19%	0.46%	0.26%	0%	2.49%	24.76%
	C	29.68%	30.04%	2.77%	5.89%	4.69%	[88.93%]	21.15%
共有序列		N	N	A	A	A	C	N

[0818] 表33:ID60进化枝7的前间隔子邻近基序(PAM)偏好

[0819] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0820]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	24.17%	15.28%	[97.1%]	0.41%	0.09%	0.18%	4.03%
	A	29.63%	27.87%	2.34%	7.16%	[96.54%]	[55.4%]	3.18%
	T	19.14%	31.83%	0.31%	[80.64%]	0.09%	2.32%	[47.41%]
	C	27.07%	25.02%	0.25%	11.79%	3.28%	/42.09%/	[45.38%]
共有序列		N	N	G	T	A	M (A>C)	Y

[0821] 表34:ID61进化枝7的前间隔子邻近基序(PAM)偏好

[0822] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0823]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	16.33%	2.30%	10.45%	[49.71%]	10.27%	5.21%	15.67%
	A	22.71%	/40.64%/	[82.63%]	[48.82%]	31.37%	24.51%	24.47%
	T	24.79%	27.85%	1.16%	0.10%	20.68%	18.23%	26.59%
	C	36.17%	29.22%	5.76%	1.37%	37.68%	/52.04%/	33.27%
共有序列		N	H (A>Y)	A	R	N	N (C>W> G)	N

[0824] 表35:ID63进化枝8的前间隔子邻近基序(PAM)偏好

[0825] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0826]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	18.02%	[100.00%]	[100.00%]	5.80%	13.04%	11.96%	23.28%
	A	1.58%	0.00%	0.00%	/44.96%/	33.20%	37.33%	28.59%
	T	16.39%	0.00%	0.00%	26.50%	/42.62%/	23.30%	26.37%
	C	[64.01%]	0.00%	0.00%	22.73%	11.14%	27.41%	21.77%
共有序列		B (C>K)	G	G	N(A>Y> G)	N	N	N

[0827] 表36:ID64进化枝9的前间隔子邻近基序(PAM)偏好

[0828] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0829]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	12.01%	0%	[100%]	0.07%	19.95%	26.33%	24.20%
	A	8.86%	[99.63%]	0%	[94.81%]	/50.21%/	29.24%	25.36%
	T	/48.83%/	0.37%	0%	3.02%	24.39%	34.46%	24.57%
	C	30.30%	0%	0%	2.11%	5.45%	9.97%	25.87%
共有序列		N (T>C>R)	A	G	A	N (A>K>C)	N	N

[0830] 表37:ID65进化枝9的前间隔子邻近基序(PAM)偏好

[0831] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0832]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	29.95%	[98.81%]	[100%]	20.33%	11.57%	20.52%	21.23%
	A	22.13%	1.11%	0%	/40.36%/	28.8%	25.49%	21.63%
	T	23.24%	0%	0%	32.01%	39.99%	27.35%	28.24%
	C	24.68%	0.08%	0%	7.31%	19.64%	26.64%	28.91%
共有序列		N	G	G	N (A>T>G>C)	N	N	N

[0833] 表38:ID66进化枝9的前间隔子邻近基序(PAM)偏好

[0834] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0835]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	29.95%	[100.00%]	[100.00%]	8.51%	3.40%	24.99%	26.27%
	A	9.78%	0.00%	0.00%	/50.57%/	20.08%	30.56%	20.09%
	T	/42.89%/	0.00%	0.00%	38.92%	[62.19%]	20.92%	25.07%
	C	17.38%	0.00%	0.00%	2.01%	14.32%	23.53%	28.56%
共有序列		N	G	G	D (A>T>G)	T	N	N

[0836] 表39:ID67进化枝9的前间隔子邻近基序(PAM)偏好

[0837] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0838]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	/42.62%/	[100.00%]	[100.00%]	4.86%	5.70%	18.40%	25.58%
	A	9.95%	0.00%	0.00%	[60.99%]	25.61%	/40.20%/	26.75%
	T	30.10%	0.00%	0.00%	30.95%	/54.61%/	19.59%	22.24%
	C	17.33%	0.00%	0.00%	3.20%	14.08%	21.81%	25.42%
共有序列		N	G	G	A	N(T>A>C>G)	N	N

[0839] 表40:ID68进化枝9的前间隔子邻近基序(PAM)偏好

[0840] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0841]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	14.54%	[100.00%]	[100.00%]	4.29%	28.33%	26.60%	19.02%
	A	[74.70%]	0.00%	0.00%	[41.25%]	22.57%	18.82%	23.93%
	T	5.28%	0.00%	0.00%	[50.74%]	/42.19%/	26.56%	33.25%
	C	5.47%	0.00%	0.00%	3.72%	6.91%	28.02%	23.80%
共有序列		C	G	G	W	N (T>R>C)	N	N

[0842] 表41:ID70进化枝9的前间隔子邻近基序(PAM)偏好

[0843] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0844]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	24.91%	[99.98%]	[100.00%]	5.34%	[94.33%]	19.93%	29.84%
	A	26.13%	0.02%	0.00%	[46.68%]	1.55%	23.48%	30.32%
	T	18.33%	0.00%	0.00%	[40.21%]	4.09%	37.54%	28.07%
	C	30.63%	0.00%	0.00%	7.78%	0.04%	19.05%	11.76%
共有序列		N	G	G	W	G	N	N

[0845] 表42:ID71进化枝9的前间隔子邻近基序(PAM)偏好

[0846] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0847]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	33.94%	[96.51%]	[100%]	21.22%	10.39%	17.04%	21.07%
	A	8.38%	3.38%	0%	38.2%	21.19%	25.41%	19.39%
	T	24.58%	0.02%	0%	30%	/45.92%/	28.63%	27.51%
	C	33.09%	0.09%	0%	10.57%	22.5%	28.92%	32.03%
共有序列		N (B>A)	G	G	N	N (T>V)	N	N

[0848] 表43:ID77进化枝10的前间隔子邻近基序(PAM)偏好

[0849] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0850]

		PAM 位置

[0851]

		1	2	3	4	5	6	7
核苷酸	G	20.44%	16.02%	[100%]	5.88%	0.49%	0.4%	34.54%
	A	22.94%	33.83%	0%	/50.41%/	[97.92%]	0.01%	16.29%
	T	17.07%	16.73%	0%	/39.08%/	1.45%	[58.62%]	33.89%
	C	39.56%	33.41%	0%	4.63%	0.14%	/40.98%/	15.27%
共有序列		N	N	G	D (A>T>G)	A	Y (T>C)	N

[0852] 表44:ID78进化枝10的前间隔子邻近基序(PAM)偏好

[0853] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0854]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	10.68%	2.39%	15.41%	0%	3.57%	9.44%	22.67%
	A	23.8%	16.85%	[84.22%]	[99.64%]	[93.98%]	[70.52%]	29.29%
	T	/44.87%/	/51.64%/	0.03%	0%	0.99%	14.92%	29.54%
	C	20.65%	29.11%	0.34%	0.36%	1.46%	5.12%	18.5%
共有序列		N (T>V)	H(T>C>A)	A	A	A	A	N

[0855] 表45:ID79进化枝10的前间隔子邻近基序(PAM)偏好

[0856] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0857]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	17.96%	[49.6%]	0.33%	0.12%	23.37%	14.15%	25.08%
	A	19.51%	[50.11%]	0%	[99.66%]	[67.06%]	30.69%	24.04%
	T	39.37%	0.03%	[99.45%]	0%	0.49%	39.64%	32.45%
	C	23.16%	0.26%	0.22%	0.22%	9.08%	15.51%	18.43%
共有序列		N	R	T	A	A	N	N

[0858] 表46:ID80进化枝10的前间隔子邻近基序(PAM)偏好

[0859] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0860]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	8.42%	0.03%	0.44%	0.06%	4.62%	15.47%	29.89%
	A	33.01%	0.61%	[99.2%]	[98.11%]	17.78%	6.43%	23.57%
	T	30.66%	8.58%	0%	0.26%	35.06%	38.25%	24.99%
	C	27.91%	[90.78%]	0.35%	1.57%	/42.53%/	39.84%	21.55%
共有序列		N (H>G)	C	A	A	H (Y>A)	N (Y>R)	N

[0862] 表47:ID81进化枝10的前间隔子邻近基序(PAM)偏好

[0863] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0864]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	29.96%	27.29%	0.34%	1.38%	2.24%	11.3%	22.57%
	A	14.59%	[65.08%]	1.88%	[97.76%]	[67.48%]	/48.92%/	35.93%
	T	27.33%	0%	[88.08%]	0%	28.63%	30.55%	23.15%
	C	28.12%	7.63%	9.7%	0.86%	1.66%	9.23%	18.35%
共有序列		N	A	T	A	A	N (A>T>S)	N

[0865] 表48:ID87进化枝10的前间隔子邻近基序(PAM)偏好

[0866] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0867]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	25.83%	31.80%	38.79%	12.23%	0.08%	0%	20.01%
	A	25.90%	/50.88%/	/55.74%/	[87.6%]	2.01%	3.30%	30.63%
	T	25.64%	4.20%	3.18%	0%	6.79%	25.75%	26.88%
	C	22.64%	13.12%	2.29%	0.18%	[91.12%]	[70.96%]	22.49%
共有序列		N	V (A>G>C)	R (A>G)	A	C	C	N

[0868] 表49:ID94进化枝11的前间隔子邻近基序(PAM)偏好

[0869] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0870]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	13.46%	6.70%	13.39%	28.71%	[99.1%]	25.66%	0%
	A	3.38%	24.93%	[59.5%]	/48.9%/	0.90%	[69.36%]	0%
	T	22.26%	25.44%	16.06%	4.46%	0%	2.51%	33.08%
	C	[60.9%]	/42.94%/	11.05%	17.93%	0%	2.46%	[66.92%]
共有序列		C	N (C>W> G)	A	V (A>S)	G	A	C

[0871]

[0872] 表50:ID97进化枝11的前间隔子邻近基序(PAM)偏好

[0873] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0874]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	19.77%	7.13%	/49.18%/	19.04%	0%	0.55%	0.51%
	A	15.06%	31.96%	/50.58%/	39.67%	0.51%	[82.96%]	0.16%
	T	29.42%	26.91%	0.04%	23.74%	14.81%	3.03%	38.27%
	C	35.75%	33.99%	0.20%	17.55%	[84.68%]	13.46%	[61.06%]
共有序列		N	N	R	N	C	A	C

[0875] 表51:ID102进化枝12的前间隔子邻近基序(PAM)偏好

[0876] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0877]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	16.73%	[99.91%]	[100.00%]	13.17%	/43.25%/	23.63%	18.92%
	A	/55.36%/	0.09%	0.00%	[36.82%]	23.17%	28.78%	33.64%
	T	16.66%	0.00%	0.00%	[46.75%]	29.00%	23.22%	29.38%
	C	11.26%	0.00%	0.00%	3.26%	4.58%	24.37%	18.06%
共有序列		N(A>B)	G	G	D (W>G)	D (G>W)	N	N

[0878] 表52:ID83进化枝1的前间隔子邻近基序(PAM)偏好

[0879] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0880]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	21.29%	[69.99%]	/55.33%/	[96.57%]	3.91%	0.03%	27.00%
	A	4.07%	30.01%	26.95%	3.43%	11.82%	0.09%	/42.82%/
	T	36.48%	0.00%	16.30%	0.00%	[78.79%]	0.36%	24.52%
	C	38.16%	0.00%	1.42%	0.00%	5.47%	[99.52%]	5.66%
共有序列		B	G	D	G	T	C	N(A>K>

[0881]

				(G>W)				C)

[0882] 表53:ID84进化枝1的前间隔子邻近基序(PAM)偏好

[0883] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0884]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	26.80%	23.57%	28.47%	9.76%	29.69%	0.00%	22.61%
	A	17.55%	[68.75%]	[71.16%]	/46.84%/	[70.25%]	0.00%	36.36%
	T	25.16%	0.05%	0.00%	30.92%	0.00%	0.00%	17.25%
	C	30.49%	7.63%	0.36%	12.47%	0.06%	[100.00%]	23.78%
共有序列		N	A	A	N	A	C	N

[0885] 表54:ID85进化枝5的前间隔子邻近基序(PAM)偏好

[0886] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0887]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	17.42%	[53.62%]	4.73%	0.45%	2.01%	18.15%	15.99%
	A	30.45%	[43.97%]	0.06%	/49.54%/	[92.82%]	/53.05%/	36.07%
	T	30.96%	1.11%	[92.25%]	31.86%	4.44%	16.94%	29.85%
	C	21.16%	1.30%	2.96%	18.15%	0.73%	11.85%	18.09%
共有序列		N	R	T	H	A	N(A>B)	N

[0888] 表55:ID88进化枝5的前间隔子邻近基序(PAM)偏好

[0889] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0890]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	15.23%	3.65%	6.39%	28.21%	[100.00%]	10.46%	3.68%
	A	1.33%	35.26%	34.29%	22.17%	0.00%	19.54%	16.19%
	T	31.94%	23.85%	24.91%	35.52%	0.00%	/48.96%/	37.01%
	C	/51.50%/	37.23%	34.40%	14.10%	0.00%	21.04%	/43.11%/
共有序列		B	H	N (H>G)	N	G	N(T>M)	H(Y>A)

[0891] 表56:ID91进化枝3的前间隔子邻近基序(PAM)偏好

[0892] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0893]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	17.45%	9.96%	[47.08%]	12.89%	0.08%	3.20%	15.05%
	A	18.82%	[48.84%]	[48.45%]	/42.75%/	[90.63%]	10.35%	28.37%
	T	23.00%	1.78%	0.00%	21.97%	1.91%	33.16%	28.30%
	C	/40.72%/	[39.42%]	4.47%	22.39%	7.38%	/53.30%/	28.28%
共有序列		N	M	R	N(A>Y>G)	A	H(C>T>A)	N

[0894] 表57:ID93进化枝3的前间隔子邻近基序(PAM)偏好

[0895] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0896]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	25.77%	12.84%	17.81%	0.00%	0.01%	0.00%	32.43%
	A	13.74%	33.00%	26.81%	5.22%	[96.69%]	0.01%	28.00%
	T	23.55%	27.15%	31.60%	7.76%	2.97%	0.00%	21.13%
	C	36.95%	27.01%	23.78%	[87.03%]	0.33%	[99.99%]	18.44%
共有序列		N	N	N	C	A	C	N

[0897] 表58:ID94进化枝3的前间隔子邻近基序(PAM)偏好

[0898] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0899]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	33.90%	3.24%	[40.55%]	10.77%	0.40%	0.01%	35.20%
	A	24.40%	[96.24%]	[56.77%]	32.74%	[92.08%]	1.03%	24.78%
	T	19.50%	0.30%	0.10%	/47.78%/	0.33%	0.13%	17.92%
	C	22.20%	0.22%	2.59%	8.71%	7.19%	[98.83%]	22.10%
共有序列		N	A	R	N(T>A>S)	A	C	N

[0900] 表59:ID96进化枝5的前间隔子邻近基序(PAM)偏好

[0901] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0902]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	24.38%	17.48%	26.35%	30.52%	0.04%	0.00%	0.29%

[0903]

	A	22.39%	27.59%	34.39%	23.04%	[99.96%]	0.00%	[55.85%]
	T	30.35%	32.34%	21.12%	32.84%	0.00%	[89.28%]	[43.70%]
	C	22.89%	22.59%	18.14%	13.60%	0.00%	10.72%	0.17%
共有序列		N	N	N	N	A	T	W

[0904] 表60:ID98进化枝3的前间隔子邻近基序(PAM)偏好

[0905] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0906]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	8.87%	[89.17%]	1.36%	21.49%	[84.56%]	0.17%	32.45%
	A	21.23%	7.29%	1.95%	24.66%	3.76%	3.87%	/40.20%/
	T	28.78%	0.01%	9.16%	15.83%	9.76%	7.63%	12.82%
	C	41.12%	3.53%	[87.53%]	38.01%	1.92%	[88.33%]	14.54%
共有序列		N	G	C	N	G	C	N

[0907] 表61:ID101进化枝3的前间隔子邻近基序(PAM)偏好

[0908] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0909]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	20.01%	11.34%	23.82%	0.00%	0.00%	0.00%	20.18%
	A	20.55%	26.03%	24.66%	12.82%	[98.81%]	8.54%	35.07%
	T	19.48%	23.24%	32.59%	0.45%	1.00%	[91.33%]	26.49%
	C	39.96%	39.39%	18.94%	[86.73%]	0.19%	0.13%	18.26%
共有序列		N	N	N	C	A	T	N

[0910] 表62:ID103进化枝2的前间隔子邻近基序(PAM)偏好

[0911] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0912]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	16.15%	18.90%	[65.15%]	31.12%	[75.29%]	0.00%	14.10%
	A	32.93%	[74.24%]	34.60%	35.43%	24.71%	2.54%	26.89%
	T	22.28%	0.00%	0.00%	17.78%	0.00%	0.00%	32.85%
	C	28.64%	6.86%	0.25%	15.67%	0.00%	[97.46%]	26.16%
共有序列		N	A	G	N	G	C	N

[0913] 表63:ID104进化枝1的前间隔子邻近基序(PAM)偏好

[0914] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0915]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	26.47%	23.57%	[49.94%]	5.78%	0.00%	0.00%	32.11%
	A	21.51%	[64.48%]	[47.51%]	19.40%	1.31%	1.41%	28.90%
	T	20.60%	0.07%	1.15%	/43.06%/	0.00%	1.64%	20.22%
	C	31.41%	11.88%	1.39%	31.76%	[98.69%]	[96.95%]	18.77%
共有序列		N	A	R	N(T>M>G)	C	C	N

[0916] 表64:ID105进化枝2的前间隔子邻近基序(PAM)偏好

[0917] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0918]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	26.70%	11.24%	3.62%	0.48%	0.00%	5.98%	25.19%
	A	25.30%	[60.72%]	14.33%	10.21%	2.18%	0.15%	22.86%
	T	23.50%	22.59%	[64.96%]	8.66%	0.00%	[81.78%]	16.31%
	C	24.51%	5.45%	17.09%	[80.65%]	[97.82%]	12.09%	35.64%
共有序列		N	A	T	C	C	T	N

[0919] 表65:ID106进化枝6的前间隔子邻近基序(PAM)偏好

[0920] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0921]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	19.69%	[46.70%]	0.00%	0.00%	24.29%	11.63%	24.06%
	A	16.38%	[53.30%]	0.00%	[100.00%]	[71.30%]	29.64%	23.15%
	T	38.91%	0.00%	[100.00%]	0.00%	0.00%	/46.72%/	33.44%
	C	25.02%	0.00%	0.00%	0.00%	4.41%	12.01%	19.35%
共有序列		N	R	T	A	A	N (T>A>S)	N

[0922] 表66:ID107进化枝8的前间隔子邻近基序(PAM)偏好

[0923] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0924]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	19.44%	0.54%	1.45%	32.87%	0.50%	13.62%	20.21%
	A	7.49%	[98.74%]	[98.05%]	/58.52%/	14.45%	33.94%	20.81%
	T	32.50%	0.18%	0.00%	8.48%	3.05%	31.30%	34.38%
	C	/40.56%/	0.54%	0.50%	0.13%	[81.99%]	21.14%	24.59%
共有序列		N(C>T>G>A)	A	A	D (A>G>T)	C	N	N

[0925] 表67:ID108进化枝8的前间隔子邻近基序(PAM)偏好

[0926] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0927]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	12.41%	5.35%	0.46%	21.00%	[75.85%]	28.22%	19.32%
	A	13.28%	[87.06%]	[99.54%]	[79.00%]	20.68%	29.04%	30.07%
	T	37.38%	1.04%	0.00%	0.00%	2.60%	29.23%	33.21%
	C	36.93%	6.54%	0.00%	0.00%	0.87%	13.51%	17.40%
共有序列		N	A	A	A	G	N	N

[0928] 表68:ID109进化枝10的前间隔子邻近基序(PAM)偏好

[0929] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0930]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	19.23%	24.61%	[99.54%]	0.00%	0.00%	16.90%	32.02%
	A	24.52%	38.19%	0.46%	[91.30%]	2.36%	28.36%	27.48%
	T	25.09%	23.78%	0.00%	0.00%	6.53%	35.06%	24.12%
	C	31.16%	13.42%	0.00%	8.70%	[91.11%]	19.68%	16.37%
共有序列		N	N	G	A	C	N	N

[0931] 表69:ID112进化枝10的前间隔子邻近基序(PAM)偏好

[0932] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0933]

		PAM 位置

[0934]

		1	2	3	4	5	6	7
核苷酸	G	16.33%	14.11%	0.00%	0.00%	2.17%	8.84%	25.79%
	A	19.13%	25.38%	6.25%	[100.00%]	[97.22%]	/54.51%/	23.51%
	T	/42.09%/	38.68%	[93.65%]	0.00%	0.61%	34.03%	34.56%
	C	22.44%	21.83%	0.09%	0.00%	0.00%	2.61%	16.13%
共有序列		N(T>V)	N	T	A	A	D (A>T>G)	N

[0935] 表70:ID116进化枝7的前间隔子邻近基序(PAM)偏好

[0936] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0937]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	28.10%	11.55%	[100.00%]	16.07%	/41.26%/	28.27%	25.86%
	A	21.76%	32.90%	0.00%	[83.93%]	29.91%	23.75%	24.55%
	T	12.65%	37.58%	0.00%	0.00%	27.98%	29.78%	28.58%
	C	37.49%	17.98%	0.00%	0.00%	0.85%	18.21%	21.01%
共有序列		N	N	G	A	D (G>W)	N	N

[0938] 表71:ID119进化枝9的前间隔子邻近基序(PAM)偏好

[0939] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0940]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	28.50%	[99.98%]	[99.96%]	8.85%	15.49%	22.77%	23.35%
	A	32.03%	0.01%	0.02%	34.59%	30.04%	26.82%	22.36%
	T	17.95%	0.02%	0.00%	/42.56%/	33.76%	25.64%	27.82%
	C	21.52%	0.00%	0.02%	14.00%	20.72%	24.77%	26.48%
共有序列		N	G	G	N(W>S)	N	N	N

[0941] 表72:ID120进化枝9的前间隔子邻近基序(PAM)偏好

[0942] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0943]

		PAM 位置						
		1	2	3	4	5	6	7

[0944]

核苷酸	G	24.68%	[97.47%]	[100.00%]	15.48%	[80.56%]	34.49%	27.40%
	A	20.40%	1.49%	0.00%	/46.59%/	2.72%	19.17%	33.53%
	T	/40.19%/	0.40%	0.00%	36.15%	16.69%	36.64%	29.09%
	C	14.72%	0.65%	0.00%	1.79%	0.03%	9.70%	9.98%
共有序列		N(T>V)	G	G	D(W>G)	G	N	N

[0945] 表73:ID121进化枝9的前间隔子邻近基序(PAM)偏好

[0946] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0947]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	21.64%	[99.88%]	[100.00%]	18.79%	8.85%	18.07%	23.75%
	A	23.84%	0.12%	0.00%	/46.56%/	23.04%	25.24%	18.83%
	T	29.96%	0.00%	0.00%	30.30%	/50.13%/	30.47%	30.07%
	C	24.56%	0.00%	0.00%	4.35%	17.98%	26.23%	27.36%
共有序列		N	G	G	D(A>T>G)	N(T>M>G)	N	N

[0948] 表74:ID122进化枝7的前间隔子邻近基序(PAM)偏好

[0949] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0950]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	24.43%	19.65%	0.02%	1.20%	5.14%	[98.14%]	29.37%
	A	20.98%	28.14%	[99.98%]	[98.35%]	[94.63%]	1.63%	25.64%
	T	35.18%	31.89%	0.00%	0.00%	0.00%	0.23%	24.28%
	C	19.40%	20.32%	0.00%	0.44%	0.23%	0.00%	20.70%
共有序列		N	N	A	A	A	G	N

[0951] 表75:ID123进化枝9的前间隔子邻近基序(PAM)偏好

[0952] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0953]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	39.61%	[99.95%]	[100.00%]	17.78%	6.15%	16.45%	24.04%

[0954]

	A	19.53%	0.05%	0.00%	/41.69%/	20.33%	29.55%	23.86%
	T	23.46%	0.00%	0.00%	36.41%	/56.54%/	26.96%	28.20%
	C	17.40%	0.00%	0.00%	4.12%	16.99%	27.04%	23.91%
共有序列		N(G>H)	G	G	D	N(T>M>G)	N	N

[0955] 表76:ID124进化枝7的前间隔子邻近基序(PAM)偏好

[0956] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0957]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	16.86%	16.49%	20.40%	[88.50%]	2.70%	1.11%	13.86%
	A	32.84%	30.59%	[78.18%]	2.07%	[94.36%]	[95.77%]	[67.88%]
	T	22.62%	26.74%	0.08%	8.07%	0.54%	2.89%	10.86%
	C	27.68%	26.18%	1.34%	1.36%	2.40%	0.23%	7.40%
共有序列		N	N	A	G	A	A	A

[0958] 表77:ID125进化枝7的前间隔子邻近基序(PAM)偏好

[0959] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0960]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	16.14%	6.88%	0.92%	0.35%	0.31%	0.81%	21.38%
	A	23.49%	21.94%	[98.19%]	[99.39%]	[99.27%]	[97.69%]	[64.31%]
	T	32.76%	/43.27%/	0.17%	0.09%	0.09%	0.89%	9.87%
	C	27.61%	27.90%	0.72%	0.17%	0.33%	0.61%	4.43%
共有序列		N	N(T>M>G)	A	A	A	A	A

[0961] 表78:ID126进化枝7的前间隔子邻近基序(PAM)偏好

[0962] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0963]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	19.49%	7.57%	[95.84%]	2.21%	27.10%	8.72%	10.45%
	A	20.71%	23.44%	4.12%	8.27%	/56.45%/	[83.93%]	[68.27%]
	T	16.90%	25.97%	0.00%	[57.84%]	12.18%	6.62%	14.06%
	C	/42.90%/	/43.02%/	0.04%	31.68%	4.27%	0.73%	7.22%

[0964]

共有序列		N(C>D)	N(C>W>G)	G	W(T>C)	D(A>G>T)	A	A

[0965] 表79:ID127进化枝10的前间隔子邻近基序(PAM)偏好

[0966] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0967]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	25.25%	26.55%	8.71%	3.98%	1.52%	1.10%	22.96%
	A	12.91%	[69.11%]	[82.08%]	[95.92%]	[77.80%]	0.09%	27.92%
	T	34.16%	0.04%	2.68%	0.00%	1.28%	[50.31%]	24.96%
	C	27.68%	4.30%	6.54%	0.10%	19.39%	[48.50%]	24.16%
共有序列		N	A	A	A	A	Y	N

[0968] 表80:ID131进化枝9的前间隔子邻近基序(PAM)偏好

[0969] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0970]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	32.38%	[45.99%]	[94.38%]	10.89%	11.50%	22.57%	17.71%
	A	28.22%	[52.41%]	4.24%	33.95%	26.70%	26.58%	27.18%
	T	11.53%	0.96%	0.52%	34.44%	/45.50%/	24.63%	26.27%
	C	27.87%	0.64%	0.86%	20.73%	16.30%	26.22%	28.85%
共有序列		N	R	G	N	N	N	N

[0971] 表81:ID132进化枝10的前间隔子邻近基序(PAM)偏好

[0972] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0973]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	17.77%	6.33%	0.80%	[65.36%]	5.70%	11.21%	14.11%
	A	14.33%	[71.50%]	6.90%	26.81%	33.68%	4.99%	37.97%
	T	32.59%	3.73%	[63.88%]	0.00%	34.29%	[68.57%]	29.70%
	C	35.31%	18.44%	28.42%	7.83%	26.34%	15.22%	18.21%
共有序列		N	A	T	G	N (H>G)	T	N

[0974] 表82:ID136进化枝9的前间隔子邻近基序(PAM)偏好

[0975] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0976]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	17.84%	25.26%	38.64%	3.86%	0.06%	0.44%	15.50%
	A	31.37%	37.18%	39.78%	[95.11%]	0.49%	0.16%	12.36%
	T	34.07%	28.69%	19.79%	0.00%	[98.97%]	[98.40%]	[65.48%]
	C	16.73%	8.86%	1.79%	1.02%	0.48%	1.01%	6.66%
共有序列		N	N	D	A	T	T	T

[0977] 表83:ID138进化枝10的前间隔子邻近基序(PAM)偏好

[0978] 展示为位置频率矩阵(PFM)。括号[x]中的数字表示强PAM偏好,斜线中的数字/x/表示弱PAM偏好。

[0979]

		PAM 位置						
		1	2	3	4	5	6	7
核苷酸	G	22.46%	20.19%	0.68%	8.49%	[43.74%]	0.00%	9.78%
	A	18.76%	[78.12%]	10.48%	[91.44%]	[53.85%]	18.32%	19.57%
	T	34.94%	0.00%	[83.47%]	0.00%	1.09%	11.13%	30.01%
	C	23.84%	1.69%	5.38%	0.07%	1.31%	[70.54%]	/40.64%/
共有序列		N	A	T	A	R	C	N(C>T>A>G)

[0980] 表84:一些Cas9直系同源物的剪切数据的汇总

[0981]

Cas9 直系同源物 ID#	NT SEQID	PRT SEQ ID	平末端剪切	粘性末端剪切	体外	植物细胞	HEK 细胞
2	1	86	X		X		
3	2	87	X		X		
4	3	88	X		X		
5	4	89	X		X		
6	5	90	X		X		X
8	6	91	X		X		X
9	7	92	X		X		
12	8	93	X		X		
13	9	94	X		X		
16	10	95	X		X		
17	11	96	X		X		X
18	12	97	X		X		
19	13	98	X		X		

[0982]

21	14	99					
27	15	100	X		X		X
28	16	101	X		X		
29	17	102	X		X		
30	18	103	X		X		
32	19	104	X		X		
33	20	105	X		X	X	X
35	21	106	X		X		
41	22	107	X		X		
43	23	108					
44	24	109	X		X		
46	25	110		X	X		X
47	26	111	X		X		
48	27	112	X		X	X	X
50	28	113	X		X	X	
51	29	114	X		X		
52	30	115	X		X		
56	31	116	X		X		X
60	32	117	X		X		
61	33	118	X		X	X	
63	34	119		X	X	X	
64	35	120	X		X	X	X
65	36	121	X		X		
66	37	122	X		X		
67	38	123	X		X		
68	39	124		X	X		X
70	40	125		X	X	X	
71	41	126	X		X		
77	42	127	X		X		
78	43	128	X		X		X
79	44	129	X		X		X
80	45	130	X		X	X	
81	46	131	X		X		
83	51	136	X		X		
84	52	137	X		X		
85	53	138	X		X		
87	47	132	X		X		
88	54	139	X		X		
91	55	140	X		X		

[0983]

93	56	141	X		X		
94	48	133	X		X		
96	58	143	X		X		
97	49	134	X		X		
98	59	144	X		X		
101	60	145	X		X		
102	50	135		X	X		
103	61	146	X		X		
104	62	147	X		X		
105	63	148	X		X		
106	64	149	X		X		
107	65	150	X		X		
108	66	151		X	X		
109	67	152	X		X		
112	68	153	X		X		
116	69	154	X		X		
119	70	155		X	X		
120	71	156	X		X		
121	72	157	X		X		
122	73	158	X		X		
123	74	159	X		X		
124	75	160	X		X		
125	76	161	X		X		
126	77	162	X		X		
127	78	163	X		X		
131	79	164		X	X		
132	80	165	X		X		
136	81	166	X		X		
138	82	167	X		X		
139	57	142	X		X		

[0984] 表85:一些Cas9直系同源物的真核细胞数据的汇总

[0985] %NHEJ突变等位基因(针对瞬时和稳定转化的植物(平均跨一到三个基因座:MS26、MS45和Lig)),用DNA表达盒转化的HEK293细胞(平均跨两个基因座:WTAP和RunX1)和用RNP(包含Cas9蛋白和sgRNA多核糖核苷酸的核糖蛋白)转化的HEK293细胞(针对一个基因座(WTAP))。酿脓链球菌Cas9作为比较剂平行进行测试。*表示对于在植物中的最佳活性而言可能需要热激。

[0986]

% NHEJ 突变等位基因

[0987]

	五蜀黍		HEK293	
Cas9 直系同源物 ID#	瞬时	稳定	表达盒	RNP
3	0.00%		0.06%	0.00%
4	0.00%		0.00%	0.00%
5	0.00%		0.00%	0.00%
6	0.00%		0.29%	3.02%
8	0.00%		3.32%	0.00%
12	0.00%		0.00%	0.00%
13	0.00%		0.00%	0.00%
17	0.00%		1.52%	0.00%
18	0.00%		0.00%	0.00%
19	0.00%		0.07%	0.00%
27	0.00%		1.34%	0.62%
30	0.00%		0.00%	0.00%
33	1.20%	43.75%	5.32%	28.40%
35	0.00%		0.30%	0.00%
41	0.00%		0.00%	0.00%
46	*		30.36%	9.22%
48	0.30%		4.05%	0.00%
50	0.22%		0.88%	0.00%
56	0.00%		17.13%	0.00%
61	0.18%		0.20%	0.00%
63	0.23%		0.00%	0.00%
64	0.43%	50.39%	4.00%	6.45%
67	0.00%		0.00%	0.33%
68	0.00%		2.67%	0.85%
70	0.24%		0.00%	0.00%
77	0.00%		0.26%	0.00%
78	0.00%		1.27%	0.00%
79	0.00%		3.34%	0.92%
80	0.07%		0.00%	0.00%
81	0.00%		0.00%	0.00%
87	0.00%		0.00%	0.00%
94	0.00%		0.00%	0.00%
SpCas9	0.58%	41.13%	21.57%	87.45%

[0988]

表 86A：Cas9 直系同源物氨基酸位置评分

各个 Cas9 直系同源物的特定氨基酸位置的评分（相比于 SpyCas9 序列 SEQ ID NO: 1125 中的位置进行参考）。分别通过求和并且除以每个数据集中的总数来定义活性和非活性数据集中每个位置处的每个氨基酸的总分率。然后，从活性数据（其中的正值表示活性 Cas9 中的在非活性集合中代表不足的保守氨基酸）中减去非活性数据集。最终得分>=0.25 用●（圆形）符号表示，并用于创建“指纹”以鉴定活性 Cas9 直系同源物。

SpCas9 位置	氨基酸																				
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.03	0.00	● 0.51	-0.03	0.00	0.00	0.00	0.00	-0.14	-0.23	0.00	0.00	-0.07
21	0.00	0.00	-0.03	0.00	0.00	0.00	0.00	0.00	0.00	● 0.47	0.18	0.00	0.00	0.00	0.00	-0.03	-0.14	0.00	-0.03	-0.41
71	0.00	-0.16	-0.17	0.00	0.00	-0.03	0.00	0.00	-0.03	0.00	● 0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.03
149	0.15	0.00	0.00	0.00	0.00	0.00	-0.03	0.00	-0.03	-0.24	● 0.40	0.00	0.00	-0.10	0.00	0.00	0.00	0.00	0.00	-0.07
150	-0.07	0.00	0.00	-0.09	0.00	0.00	-0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	● 0.51	0.00	-0.21	0.00	0.00	0.00
444	-0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.24	● 0.44	-0.03	-0.03	0.04	0.00	0.00	0.00	0.00	0.00	-0.07
445	-0.03	-0.03	0.11	0.00	0.00	0.00	0.11	0.00	-0.03	-0.03	0.00	-0.07	-0.07	-0.34	0.00	-0.03	● 0.51	0.00	-0.07	0.00
503	0.00	0.00	0.00	-0.03	0.00	-0.03	-0.08	0.00	0.00	-0.07	-0.07	0.11	0.00	-0.03	● 0.40	-0.03	0.00	0.00	0.00	0.00
587	0.00	0.00	0.00	0.00	-0.03	0.00	0.00	0.00	0.00	0.07	-0.17	0.00	-0.03	● 0.41	0.00	-0.03	0.00	-0.07	-0.07	-0.03
620	● 0.54	-0.03	0.00	-0.10	0.00	-0.07	0.00	-0.07	-0.03	-0.24	0.00	0.00	0.00	0.00	0.00	-0.07	0.00	0.00	0.00	0.08
623	-0.07	0.00	-0.03	0.00	0.00	0.00	0.00	0.00	-0.21	-0.10	● 0.69	0.00	-0.07	-0.07	0.00	-0.14	0.00	0.00	0.00	0.00
624	0.00	-0.03	0.00	-0.03	0.00	0.14	0.00	0.00	0.00	-0.07	-0.14	-0.03	0.00	0.00	0.00	-0.17	● 0.44	0.00	-0.03	-0.07
632	0.11	0.00	0.00	0.00	0.00	0.00	0.00	-0.07	0.00	● 0.55	-0.24	-0.03	0.00	-0.14	0.00	0.00	-0.05	0.00	0.00	-0.03
692	0.00	-0.17	-0.09	-0.07	0.00	● 0.50	0.00	0.00	-0.05	0.00	-0.10	-0.07	0.00	0.00	-0.03	0.00	0.00	0.00	-0.07	0.00
702	-0.10	0.00	0.00	-0.03	0.00	0.00	0.00	-0.07	0.00	0.00	● 0.62	0.00	-0.03	-0.03	0.00	-0.03	0.00	0.00	-0.10	0.00
781	0.00	-0.07	0.00	0.00	0.00	-0.03	-0.03	-0.03	0.00	● 0.44	-0.10	-0.03	0.07	0.00	0.00	0.00	0.00	-0.03	-0.10	-0.07
810	-0.03	-0.24	-0.03	-0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	● 0.41	-0.03	0.00	0.00	0.00	0.00	0.00	0.00	-0.03
908	0.00	0.00	0.11	0.00	0.00	-0.03	-0.03	0.00	0.00	-0.17	● 0.48	0.00	-0.07	-0.14	0.00	0.00	-0.05	0.00	0.00	0.00
931	-0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	-0.06	0.00	0.00	0.00	0.00	0.00	-0.07	0.00	0.00	● 0.40
933	-0.03	-0.07	● 0.40	-0.03	0.00	● 0.36	-0.17	-0.03	-0.03	-0.07	-0.03	-0.10	0.00	0.00	0.00	-0.03	-0.07	0.00	0.00	-0.07
954	0.00	-0.24	-0.10	0.00	0.00	-0.07	-0.14	0.14	-0.10	0.00	0.00	● 0.47	0.00	0.00	0.00	-0.03	0.11	0.00	0.00	0.00
955	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	● 0.48
1000	-0.03	0.04	0.00	0.00	-0.03	-0.03	0.07	-0.03	0.00	-0.03	-0.07	● 0.44	0.00	-0.03	0.00	-0.03	0.00	0.00	-0.10	-0.14
1100	-0.17	0.00	0.00	0.00	0.00	-0.10	0.00	0.00	0.00	0.00	-0.03	-0.07	0.00	0.11	-0.21	-0.03	-0.10	0.00	0.00	● 0.62
1232	0.00	0.00	0.11	-0.03	0.00	0.00	-0.03	-0.03	-0.14	0.00	0.00	-0.06	0.00	-0.10	-0.03	-0.07	0.00	-0.03	● 0.44	0.00
1236	-0.03	0.00	-0.03	0.00	-0.03	0.00	0.00	0.00	0.00	● 0.51	0.11	0.00	0.00	-0.21	0.00	0.00	0.00	0.00	0.00	-0.31

[0989] 表86B:活性Cas9直系同源物指纹

[0990] 在真核细胞中具有较高活性可能性的直系同源物的特征氨基酸残基。位置编号是关于酿脓链球菌Cas9(SEQ ID NO:1125)的类似氨基酸位置编号。在真核细胞中具有阳性剪切活性的直系同源物包含这些结构特征中的一个或多个。

[0991]

相对位置	氨基酸

13	I
21	I
71	L
149	L
150	S
444	L
445	T
503	P
587	F
620	A
623	L
624	T
632	I
692	Q
702	L
781	I
810	K
908	L
931	V
933	N or Q
954	K
955	V
1000	K
1100	V
1232	Y
1236	I

[0992] 表86C:Cas9直系同源氨基酸位置总得分(总和)

[0993] Cas9直系同源物评分,以总得分/直系同源PRT SEQID,基于表86A中鉴定的位置的得分的总和。

[0994]

PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分
527	11.64	**1005**	8.46	**1126**	6.97	**708**	5.85	**1078**	5.05	**998**	4.17	**772**	3.72
116	11.43	**885**	8.39	**841**	6.90	**855**	5.84	**564**	5.02	**152**	4.16	**892**	3.71
860	11.19	**110**	8.37	**546**	6.89	**138**	5.74	**1133**	5.02	**656**	4.16	**723**	3.70

[0995]

868	11.09	**125**	8.36	**1059**	6.86	**512**	5.70	**103**	5.01	**888**	4.13	**789**	3.69
115	10.69	**157**	8.32	**1028**	6.83	**932**	5.67	**1016**	5.01	**1036**	4.13	**515**	3.69
160	10.66	**691**	8.32	**981**	6.80	**143**	5.66	**1047**	4.98	**839**	4.10	**853**	3.69
162	10.65	**697**	8.32	**1042**	6.76	**980**	5.64	**1114**	4.96	**1082**	4.10	**926**	3.69
666	10.25	**801**	8.32	**939**	6.76	**648**	5.64	**970**	4.89	**543**	4.09	**1130**	3.69
821	10.25	**1121**	8.32	**678**	6.75	**856**	5.62	**684**	4.87	**513**	4.08	**551**	3.68
633	10.21	**1122**	8.32	**754**	6.71	**680**	5.61	**108**	4.85	**706**	4.07	**907**	3.68
514	10.18	**1123**	8.28	**913**	6.68	**661**	5.56	**1015**	4.82	**1041**	4.07	**530**	3.66
105	10.15	**953**	8.25	**999**	6.67	**664**	5.53	**102**	4.79	**1077**	4.07	**1010**	3.66
922	10.07	**793**	8.11	**751**	6.59	**735**	5.51	**1068**	4.78	**1106**	4.07	**852**	3.64
169	9.85	**877**	7.97	**159**	6.53	**727**	5.48	**123**	4.71	**1061**	4.06	**1131**	3.64
526	9.85	**1076**	7.89	**570**	6.53	**679**	5.47	**750**	4.69	**532**	4.06	**711**	3.63
168	9.73	**111**	7.88	**571**	6.52	**993**	5.47	**715**	4.66	**1014**	4.02	**845**	3.62
660	9.70	**911**	7.87	**531**	6.51	**802**	5.45	**518**	4.62	**990**	4.01	**641**	3.59
1102	9.64	**669**	7.84	**985**	6.51	**936**	5.44	**806**	4.61	**1128**	4.01	**997**	3.58
756	9.59	**630**	7.84	**948**	6.50	**826**	5.42	**1004**	4.60	**90**	3.98	**613**	3.57
978	9.43	**799**	7.84	**949**	6.50	**126**	5.40	**659**	4.59	**1092**	3.96	**910**	3.57
589	9.30	**1032**	7.84	**792**	6.34	**559**	5.40	**580**	4.58	**807**	3.94	**718**	3.57
726	9.30	**1039**	7.84	**849**	6.34	**590**	5.40	**884**	4.57	**119**	3.94	**923**	3.56
1038	9.30	**1048**	7.80	**759**	6.33	**592**	5.40	**552**	4.54	**614**	3.91	**720**	3.55
942	9.26	**741**	7.77	**716**	6.33	**1117**	5.40	**987**	4.53	**572**	3.88	**873**	3.55
113	9.26	**121**	7.75	**941**	6.32	**539**	5.39	**947**	4.50	**815**	3.87	**134**	3.54
161	9.26	**624**	7.71	**848**	6.32	**729**	5.38	**603**	4.50	**850**	3.87	**902**	3.53
681	9.26	**112**	7.58	**117**	6.31	**797**	5.37	**693**	4.42	**602**	3.85	**1085**	3.49
1049	9.16	**101**	7.50	**553**	6.29	**780**	5.37	**765**	4.42	**745**	3.85	**876**	3.48
938	9.09	**114**	7.50	**835**	6.28	**654**	5.34	**668**	4.41	**757**	3.85	**810**	3.47
898	8.98	**966**	7.48	**1045**	6.27	**104**	5.32	**794**	4.39	**634**	3.85	**989**	3.46
158	8.90	**586**	7.44	**808**	6.27	**927**	5.30	**882**	4.35	**579**	3.84	**955**	3.45
777	8.90	**124**	7.44	**118**	6.23	**139**	5.30	**1099**	4.33	**804**	3.84	**961**	3.45
891	8.86	**155**	7.44	**598**	6.17	**918**	5.30	**820**	4.32	**895**	3.84	**107**	3.44
120	8.83	**690**	7.44	**604**	6.17	**1050**	5.23	**881**	4.31	**109**	3.83	**145**	3.44
946	8.83	**636**	7.40	**1134**	6.10	**619**	5.20	**653**	4.30	**695**	3.83	**976**	3.44
937	8.79	**623**	7.36	**790**	6.09	**1074**	5.17	**1056**	4.28	**1008**	3.83	**612**	3.43
944	8.79	**1072**	7.31	**519**	6.08	**812**	5.16	**924**	4.27	**696**	3.81	**701**	3.39
1031	8.79	**713**	7.29	**140**	6.06	**764**	5.16	**811**	4.26	**1001**	3.81	**167**	3.37
865	8.75	**722**	7.22	**774**	6.06	**1043**	5.16	**903**	4.26	**637**	3.81	**582**	3.37
156	8.73	**1064**	7.18	**795**	6.02	**164**	5.15	**788**	4.24	**1115**	3.81	**640**	3.36
762	8.73	**916**	7.18	**972**	5.95	**883**	5.15	**901**	4.24	**890**	3.80	**781**	3.36
833	8.71	**688**	7.16	**106**	5.93	**1044**	5.13	**958**	4.24	**854**	3.80	**1080**	3.34
747	8.71	**725**	7.14	**587**	5.92	**904**	5.13	**1135**	4.24	**520**	3.79	**599**	3.33
842	8.71	**934**	7.00	**731**	5.91	**782**	5.12	**674**	4.22	**542**	3.79	**871**	3.33
732	8.60	**628**	6.99	**100**	5.90	**851**	5.09	**540**	4.21	**710**	3.79	**1058**	3.33
935	8.57	**1120**	6.99	**1023**	5.90	**683**	5.09	**1086**	4.21	**749**	3.76	**896**	3.32
967	8.54	**861**	6.97	**592**	5.90	**1026**	5.05	**1025**	4.19	**1067**	3.74	**98**	3.32
893	8.47	**862**	6.97	**122**	5.88	**1037**	5.05	**658**	4.17	**737**	3.72	**151**	3.30

[0996]

PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分	PRT SEQID	总得分
714	3.29	**574**	3.09	**671**	2.76	**738**	2.58	**763**	2.23	**973**	1.86	**798**	1.64	**585**	0.99
859	3.29	**694**	3.09	**771**	2.75	**549**	2.51	**894**	2.23	**149**	1.85	**140**	1.57	**597**	0.99
1081	3.29	**662**	3.07	**889**	2.74	**921**	2.48	**1095**	2.23	**778**	1.84	**675**	1.54	**670**	0.99
712	3.28	**535**	3.07	**920**	2.74	**917**	2.44	**886**	2.22	**1109**	1.84	**739**	1.53	**700**	0.99
736	3.28	**561**	3.07	**1098**	2.74	**135**	2.42	**146**	2.20	**547**	1.84	**704**	1.51	**746**	0.99

[0997]

1009	3.28	**621**	3.07	**131**	2.74	**154**	2.42	**643**	2.20	**761**	1.84	**1104**	1.49	**783**	0.99
1097	3.27	**629**	3.06	**1046**	2.74	**524**	2.42	**838**	2.20	**857**	1.84	**595**	1.47	**615**	0.96
525	3.26	**900**	3.06	**545**	2.71	**677**	2.42	**130**	2.20	**1110**	1.84	**92**	1.46	**642**	0.96
717	3.25	**1017**	3.05	**550**	2.71	**136**	2.41	**974**	2.19	**94**	1.83	**529**	1.46	**733**	0.96
837	3.25	**1073**	3.05	**805**	2.71	**914**	2.41	**992**	2.19	**584**	1.83	**959**	1.46	**956**	0.96
730	3.25	**994**	3.03	**984**	2.71	**968**	2.41	**1012**	2.18	**1052**	1.83	**768**	1.45	**652**	0.95
803	3.25	**563**	3.03	**166**	2.71	**988**	2.41	**1096**	2.17	**672**	1.83	**91**	1.43	**766**	0.95
899	3.24	**1003**	3.02	**915**	2.70	**743**	2.37	**573**	2.16	**523**	1.82	**1018**	1.42	**925**	0.95
607	3.23	**905**	3.00	**1040**	2.70	**825**	2.37	**625**	2.16	**682**	1.82	**86**	1.40	**1084**	0.95
645	3.22	**635**	3.00	**875**	2.69	**950**	2.37	**647**	2.16	**844**	1.81	**88**	1.40	**1111**	0.95
631	3.21	**1087**	2.98	**796**	2.69	**1030**	2.37	**709**	2.16	**740**	1.80	**516**	1.40	**836**	0.94
719	3.21	**544**	2.95	**665**	2.68	**129**	2.36	**866**	2.16	**823**	1.80	**609**	1.40	**1029**	0.92
840	3.21	**558**	2.95	**755**	2.68	**1034**	2.36	**897**	2.16	**610**	1.80	**689**	1.40	**1075**	0.92
1002	3.21	**626**	2.95	**770**	2.67	**646**	2.35	**1007**	2.16	**620**	1.80	**1132**	1.40	**933**	0.91
1105	3.21	**651**	2.89	**969**	2.67	**1093**	2.35	**93**	2.14	**1088**	1.80	**537**	1.39	**809**	0.88
769	3.20	**773**	2.89	**1089**	2.67	**1107**	2.34	**748**	2.14	**1101**	1.80	**560**	1.39	**919**	0.88
1066	3.19	**685**	2.88	**1116**	2.67	**616**	2.31	**957**	2.14	**611**	1.79	**567**	1.39	**1054**	0.88
141	3.19	**1006**	2.88	**555**	2.66	**1053**	2.31	**827**	2.13	**847**	1.79	**818**	1.39	**1079**	0.88
707	3.18	**618**	2.88	**639**	2.66	**816**	2.31	**870**	2.13	**843**	1.77	**830**	1.39	**87**	0.48
878	3.18	**699**	2.86	**724**	2.66	**1070**	2.31	**686**	2.09	**676**	1.76	**1108**	1.38	**554**	0.48
127	3.17	**1112**	2.86	**931**	2.66	**557**	2.30	**702**	2.09	**1083**	1.76	**594**	1.36	**627**	0.48
153	3.17	**822**	2.85	**977**	2.66	**979**	2.30	**533**	2.05	**863**	1.76	**622**	1.36	**775**	0.48
600	3.17	**912**	2.84	**1033**	2.66	**601**	2.27	**99**	2.01	**142**	1.75	**753**	1.36	**817**	0.48
644	3.17	**144**	2.83	**785**	2.66	**632**	2.27	**565**	1.98	**906**	1.75	**1113**	1.36	**824**	0.48
657	3.17	**1055**	2.83	**872**	2.66	**846**	2.27	**828**	1.98	**1069**	1.75	**1129**	1.36	**1063**	0.48
945	3.17	**150**	2.82	**1090**	2.66	**538**	2.27	**1065**	1.98	**964**	1.74	**577**	1.33	**1100**	0.48
874	3.15	**528**	2.82	**869**	2.65	**596**	2.27	**534**	1.97	**760**	1.73	**1119**	1.33	**1103**	0.48
569	3.15	**591**	2.82	**963**	2.65	**578**	2.26	**1020**	1.96	**940**	1.73	**95**	1.32	**928**	0.40
132	3.14	**703**	2.82	**1091**	2.65	**951**	2.26	**971**	1.94	**975**	1.73	**97**	1.32	**960**	0.40
606	3.14	**779**	2.82	**929**	2.64	**986**	2.26	**133**	1.91	**1000**	1.73	**752**	1.32		
767	3.14	**1051**	2.82	**541**	2.64	**96**	2.25	**568**	1.91	**1057**	1.73	**786**	1.32		
784	3.14	**649**	2.81	**617**	2.64	**522**	2.25	**834**	1.91	**583**	1.72	**1071**	1.32		
148	3.13	**880**	2.81	**787**	2.64	**965**	2.25	**588**	1.90	**721**	1.72	**650**	1.31		
791	3.12	**954**	2.81	**1118**	2.64	**995**	2.25	**991**	1.90	**673**	1.69	**734**	1.31		
1024	3.12	**165**	2.81	**687**	2.63	**137**	2.24	**1022**	1.90	**1094**	1.68	**983**	1.31		
705	3.12	**1011**	2.81	**908**	2.63	**831**	2.24	**887**	1.89	**698**	1.66	**147**	1.29		
996	3.11	**909**	2.80	**930**	2.63	**605**	2.24	**562**	1.87	**536**	1.66	**517**	1.29		
879	3.11	**829**	2.79	**1027**	2.63	**608**	2.24	**576**	1.87	**943**	1.66	**814**	1.29		
521	3.10	**692**	2.78	**744**	2.61	**728**	2.24	**581**	1.87	**961**	1.66	**867**	1.28		
758	3.10	**1060**	2.78	**128**	2.60	**1021**	2.24	**654**	1.87	**1062**	1.66	**864**	1.10		
1019	3.10	**858**	2.78	**800**	2.60	**566**	2.23	**1013**	1.87	**819**	1.65	**982**	1.06		
575	3.10	**163**	2.77	**832**	2.60	**638**	2.23	**1124**	1.87	**556**	1.64	**89**	0.99		
813	3.10	**663**	2.77	**952**	2.58	**667**	2.23	**776**	1.86	**742**	1.64	**548**	0.99		

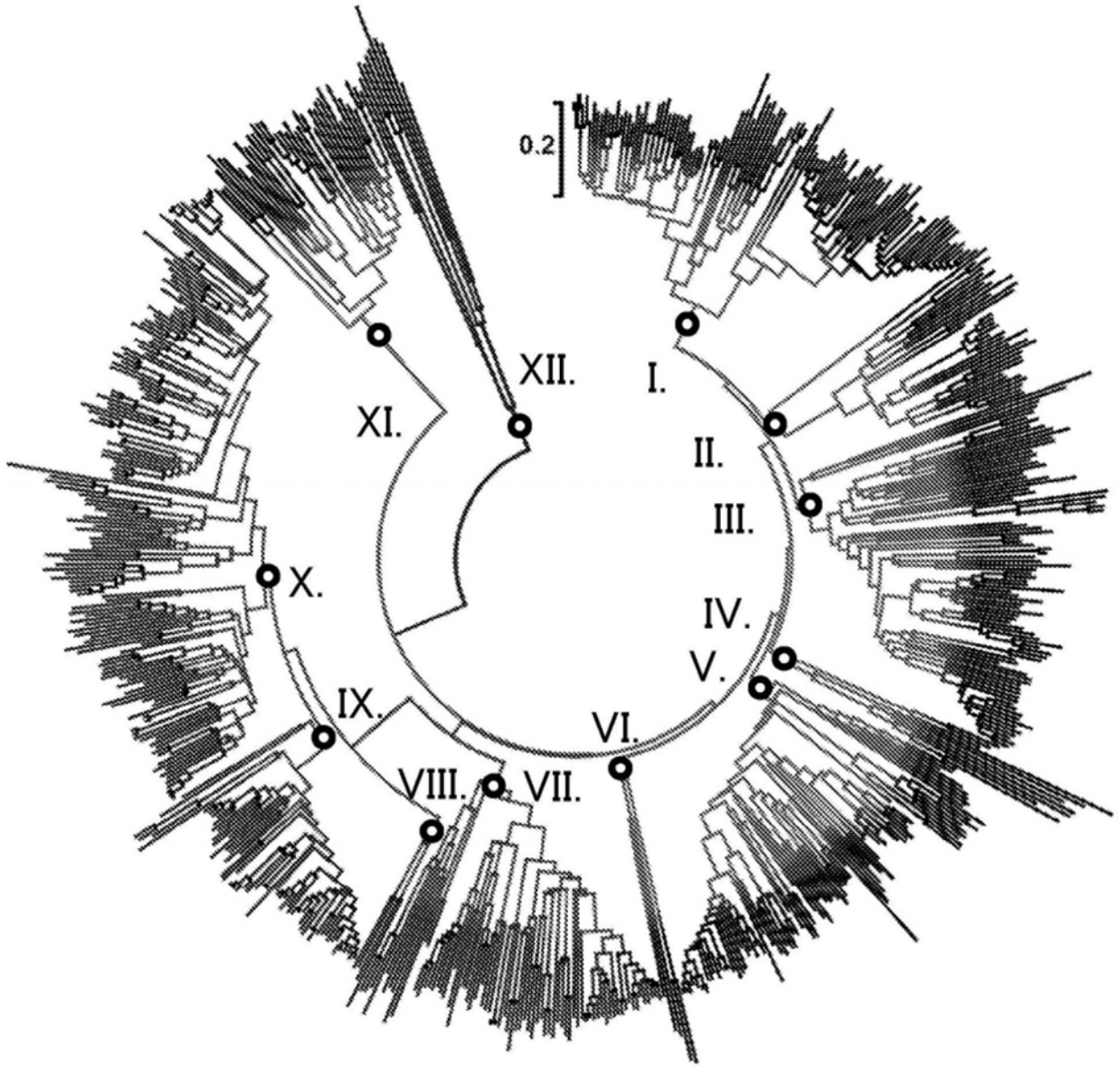


图1

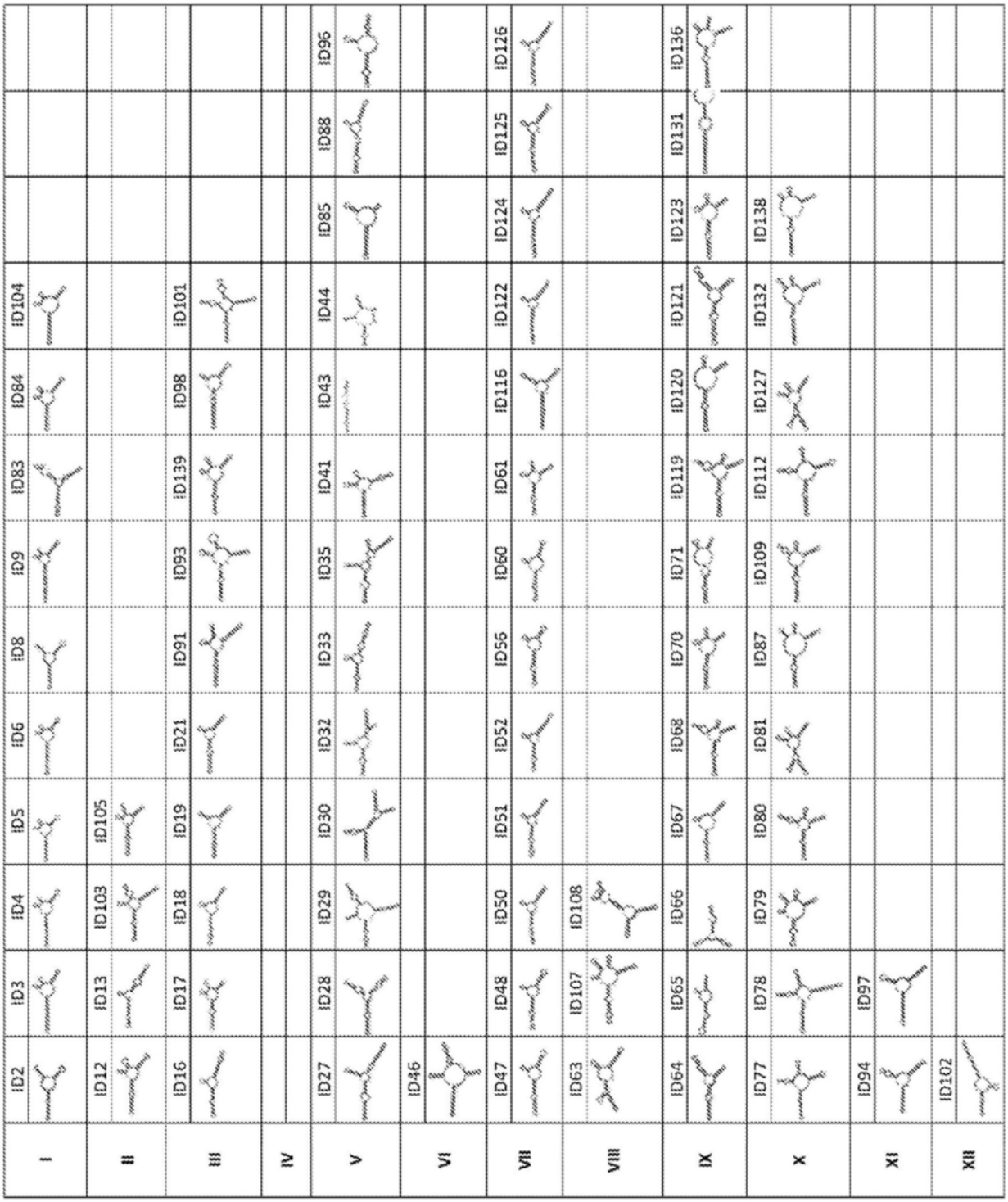

I
ID2 ID3 ID4 ID5 ID6 ID8 ID9 ID83 ID84 ID104
II
ID12 ID13 ID103 ID105
III
ID16 ID17 ID18 ID19 ID21 ID91 ID93 ID139 ID98 ID101
IV
V
ID27 ID28 ID29 ID30 ID32 ID33 ID35 ID41 ID43 ID44 ID85 ID88 ID95
VI
ID46
VII
ID47 ID48 ID50 ID51 ID52 ID56 ID60 ID61 ID116 ID122 ID124 ID125 ID126
VIII
ID63 ID107 ID108
IX
ID64 ID65 ID66 ID67 ID68 ID70 ID71 ID119 ID120 ID121 ID123 ID131 ID136
X
ID77 ID78 ID79 ID80 ID81 ID87 ID109 ID112 ID127 ID132 ID138
XI
ID94 ID97
XII
ID102

图2

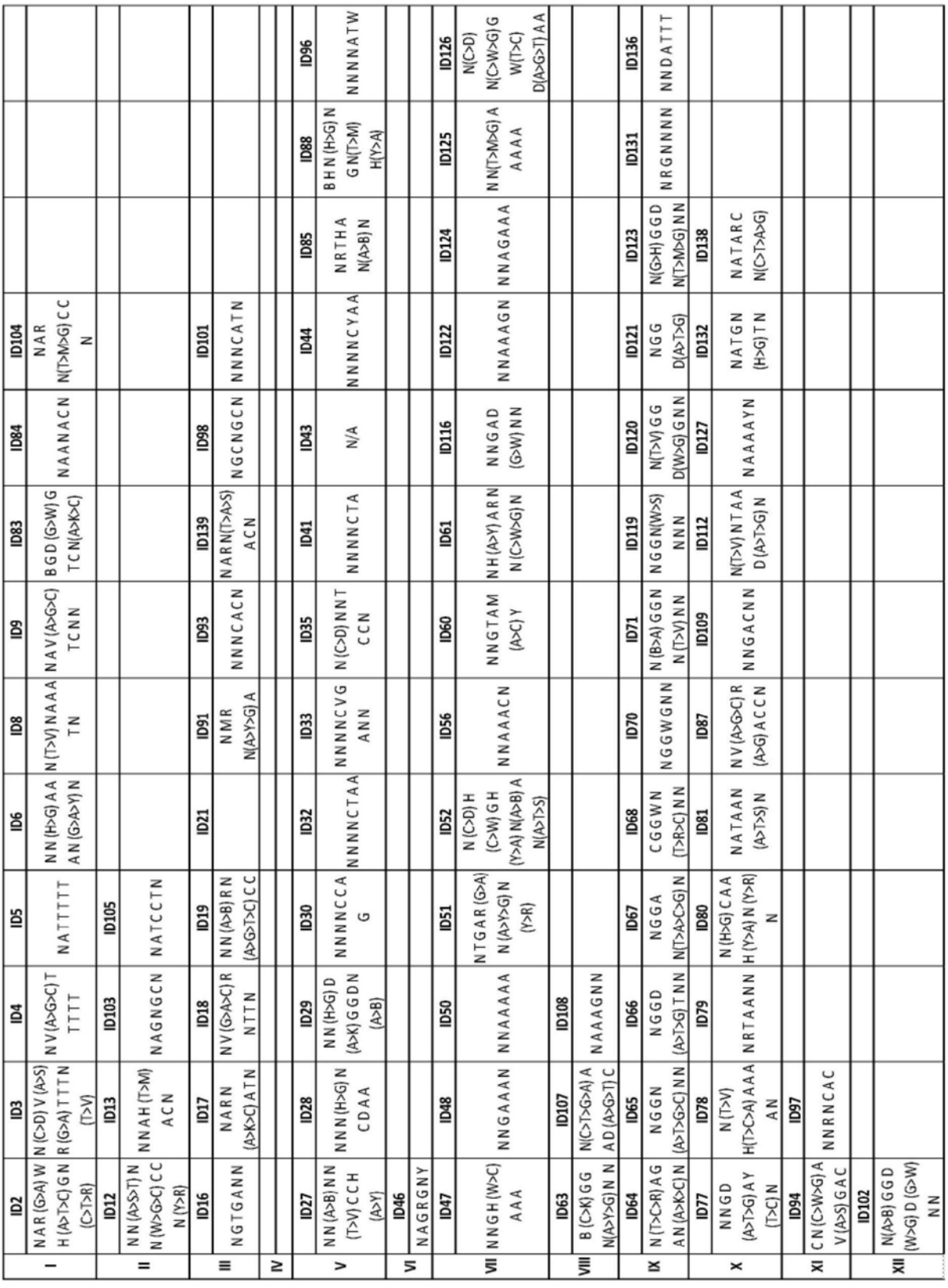

I
ID2 NAR (G>A) W H (A>T>C) G N (C>T>R)
ID3 N (C>D) V (A>S) R (G>A) T T T N (T>V)
ID4 N V (A>G>C) T T T T
ID5 N A T T T T T
ID6 N N (H>G) A A A N (G>A>Y) N
ID8 N (T>V) N A A A T N
ID9 N A V (A>G>C) T C N N
ID83 B G D (G>W) G T C N (A>K>C)
ID84 N A A N A C N
ID104 N A R N (T>M>G) C C N
II
ID12 N N (A>S>T) N N (W>G>C) C C N (Y>R)
ID13 N N A H (T>M) A C N
ID103 N A G N G C N
ID105 N A T C C T N
III
ID16 N G T G A N N
ID17 N A R N (A>K>C) A T N
ID18 N V (G>A>C) R N T T N
ID19 N N (A>B) R N (A>G>T>C) C C
ID21
ID91 N M R N (A>Y>G) A
ID93 N N N C A C N
ID139 N A R N (T>A>S) A C N
ID98 N G C N G C N
ID101 N N N C A T N
IV
V
ID27 N N (A>B) N N (T>V) C C H (A>Y)
ID28 N N N (H>G) N C D A A
ID29 N N (H>G) D (A>K) G G D N (A>B)
ID30 N N N N C C A G
ID32 N N N N C T A A
ID33 N N N N C V G A N N
ID35 N (C>D) N N T C C N
ID41 N N N N C T A
ID43 N/A
ID44 N N N N C Y A A
ID85 N R T H A N (A>B) N
ID88 B H N (H>G) N G N (T>M) H (Y>A)
ID96 N N N N A T W
VI
ID46
N A G R G N Y
VII
ID47 N N G H (W>C) A A A
ID48 N N G A A A N
ID50 N N A A A A A
ID51 N T G A R (G>A) N (A>Y>G) N (Y>R)
ID52 N (C>D) H (C>W) G H (Y>A) N (A>B) A N (A>T>S)
ID56 N N A A A C N
ID60 N N G T A M (A>C) Y
ID61 N H (A>Y) A R N N (C>W>G) N
ID116 N N G A D (G>W) N N
ID122 N N A A A G N
ID124 N N A G A A A
ID125 N N (T>M>G) A A A A A
ID126 N (C>D) N (C>W>G) G W (T>C) D (A>G>T) A A
VIII
ID63 B (C>K) G G N (A>Y>G) N N
ID107 N (C>T>G>A) A A D (A>G>T) C
ID108 N A A A G N N
IX
ID64 N (T>C>R) A G A N (A>K>C) N
ID65 N G G N (A>T>G>C) N N
ID66 N G G D (A>T>G) T N N
ID67 N G G A N (T>A>C>G) N
ID68 C G G W N (T>R>C) N N
ID70 N G G W G N N
ID71 N (B>A) G G N N (T>V) N N
ID119 N G G N (W>S) N N N
ID120 N (T>V) G G D (W>G) G N N
ID121 N G G D (A>T>G)
ID123 N (G>H) G G D N (T>M>G) N N
ID131 N R G N N N N
ID136 N N D A T T T
X
ID77 N N G D (A>T>G) A Y (T>C) N
ID78 N (T>V) H (T>C>A) A A A N
ID79 N R T A A N N
ID80 N (H>G) C A A H (Y>A) N (Y>R) N
ID81 N A T A A N (A>T>S) N
ID87 N V (A>G>C) R (A>G) A C C N
ID109 N N G A C N N
ID112 N (T>V) N T A A D (A>T>G) N
ID127 N A A A A Y N
ID132 N A T G N (H>G) T N
ID138 N A T A R C N (C>T>A>G)
XI
ID94 C N (C>W>G) A V (A>S) G A C
ID97 N N R N C A C
ID102
XII
N (A>B) G G D (W>G) D (G>W) N N

图3

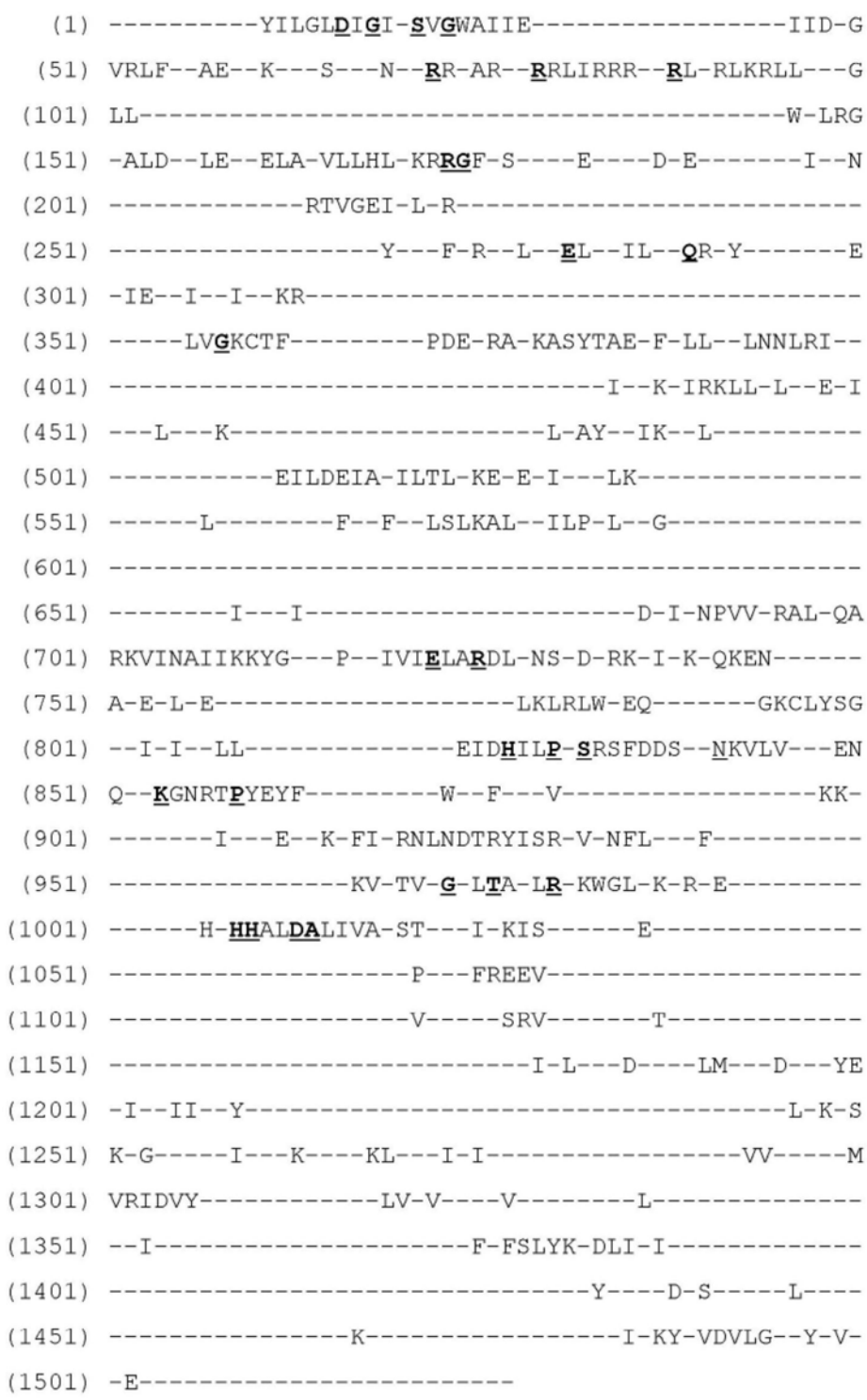

(1) ----------YILGLDIGI-SVGWAIIE----------IID-G
(51) VRLF--AE--K---S---N--RR-AR--RRLIRRR--RL-RLKRLL---G
(101) LL----------W-LRG
(151) -ALD--LE--ELA-VLLHL-KRRGF-S----E----D-E-------I--N
(201) ----------RTVGEI-L-R----------
(251) ----------Y---F-R--L--EL--IL--QR-Y-------E
(301) -IE--I--I--KR----------
(351) -----LVGKCTF---------PDE-RA-KASYTAE-F-LL--LNNLRI--
(401) ----------I--K-IRKLL-L--E-I
(451) ---L---K----------L-AY--IK--L----------
(501) ----------EILDEIA-ILTL-KE-E-I---LK----------
(551) ------L--------F--F--LSLKAL--ILP-L--G----------
(601) ----------
(651) --------I---I----------D-I-NPVV-RAL-QA
(701) RKVINAIIKKYG---P--IVIELARDL-NS-D-RK-I-K-QKEN------
(751) A-E-L-E----------LKLRLW-EQ-------GKCLYSG
(801) --I-I--LL----------EIDHILP-SRSFDDS--NKVLV---EN
(851) Q--KGNRTPYEYF---------W--F---V----------KK-
(901) -------I---E--K-FI-RNLNDTRYISR-V-NFL---F----------
(951) ----------KV-TV-G-LTA-LR-KWGL-K-R-E---------
(1001) ------H-HHALDALIVA-ST---I-KIS------E----------
(1051) ----------P---FREEV----------
(1101) ----------V-----SRV-------T----------
(1151) ----------I-L---D----LM---D---YE
(1201) -I--II--Y----------L-K-S
(1251) K-G-----I---K----KL---I-I----------VV-----M
(1301) VRIDVY----------LV-V----V--------L----------
(1351) --I----------F-FSLYK-DLI-I----------
(1401) ----------Y----D-S-----L----
(1451) ----------K----------I-KY-VDVLG--Y-V-
(1501) -E----------

图4

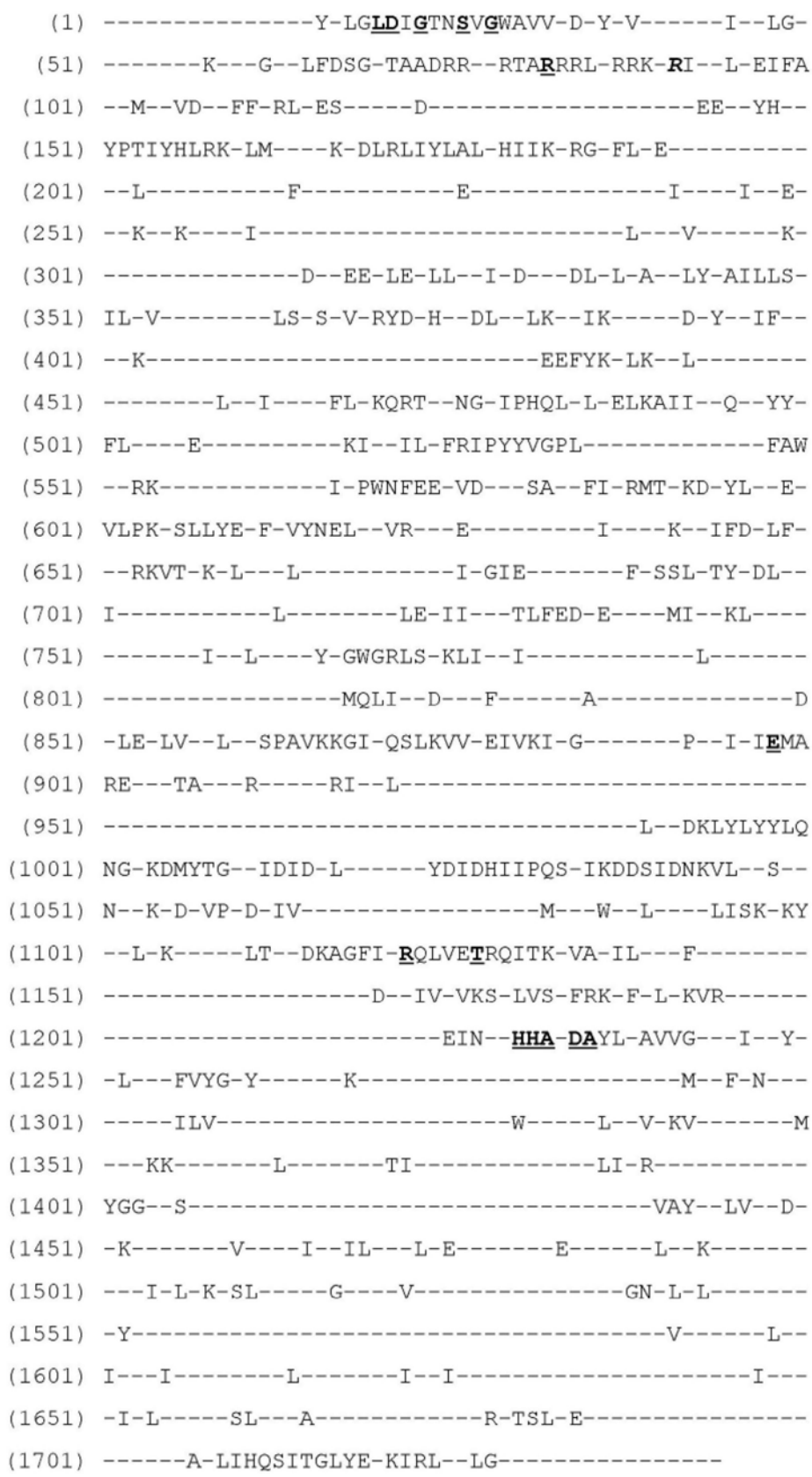
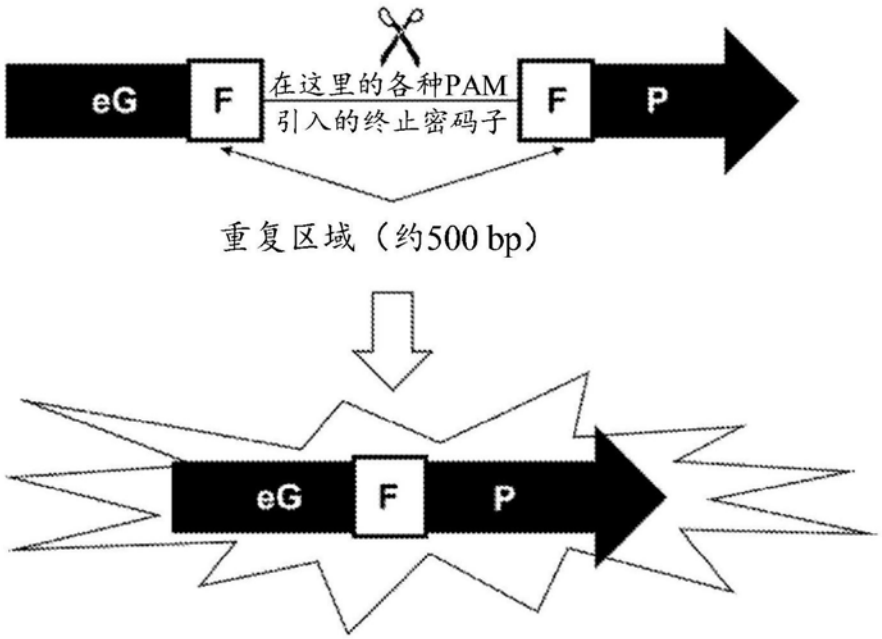

(1) ----------Y-LGLDIGTNSVGWAVV-D-Y-V------I--LG-
(51) -------K---G--LFDSG-TAADRR--RTARRRL-RRK-RI--L-EIFA
(101) --M--VD--FF-RL-ES-----D----------EE--YH--
(151) YPTIYHLRK-LM----K-DLRLIYLAL-HIIK-RG-FL-E----------
(201) --L----------F----------E----------I----I--E-
(251) --K--K----I----------L---V------K-
(301) ----------D--EE-LE-LL--I-D---DL-L-A--LY-AILLS-
(351) IL-V--------LS-S-V-RYD-H--DL--LK--IK-----D-Y--IF--
(401) --K----------EEFYK-LK--L--------
(451) --------L--I----FL-KQRT--NG-IPHQL-L-ELKAII--Q--YY-
(501) FL----E----------KI--IL-FRIPYYVGPL----------FAW
(551) --RK----------I-PWNFEE-VD---SA--FI-RMT-KD-YL--E-
(601) VLPK-SLLYE-F-VYNEL--VR---E---------I----K--IFD-LF-
(651) --RKVT-K-L---L----------I-GIE-------F-SSL-TY-DL--
(701) I----------L--------LE-II---TLFED-E----MI--KL----
(751) -------I--L----Y-GWGRLS-KLI--I----------L-------
(801) ----------MQLI--D---F------A----------D
(851) -LE-LV--L--SPAVKKGI-QSLKVV-EIVKI-G-------P--I-IEMA
(901) RE---TA---R-----RI--L----------
(951) ----------L--DKLYLYYLQ
(1001) NG-KDMYTG--IDID-L------YDIDHIIPQS-IKDDSIDNKVL--S--
(1051) N--K-D-VP-D-IV----------M---W--L----LISK-KY
(1101) --L-K-----LT--DKAGFI-RQLVETRQITK-VA-IL---F--------
(1151) ----------D--IV-VKS-LVS-FRK-F-L-KVR------
(1201) ----------EIN--HHA-DAYL-AVVG---I--Y-
(1251) -L---FVYG-Y------K----------M--F-N---
(1301) -----ILV----------W-----L--V-KV-------M
(1351) ---KK-------L-------TI----------LI-R----------
(1401) YGG--S----------VAY--LV--D-
(1451) -K-------V----I--IL---L-E-------E------L--K-------
(1501) ---I-L-K-SL-----G----V----------GN-L-L-------
(1551) -Y----------V------L--
(1601) I---I--------L-------I--I----------I---
(1651) -I-L-----SL---A----------R-TSL-E----------
(1701) ------A-LIHQSITGLYE-KIRL--LG----------

图5

```
(1) -----L---K-YRVGIDVGTHSVGLAAIEVDDH--------PI-ILSALS
(51) LIHDSGVDPD--K-A-TRKA-SGVARRTRRLHK-RR-RL-KLDEVLNDLG
(101) FPI-----F-D----SDPYI-WNVRAKLVE-FIPDD--RG--ISIAIRHI
(151) ARHRGWRNPYSKV-SL-SPA--S--------------------------
(201) --------------E-IR--II---GD-L--GITIGQLI--A-I---KIR
(251) RD------------IISAKLHQSDHA-EI--I--RQ-VD-DL-KQLLDAV
(301) F-ADSP--KGAAL-RVGKDPL-----F-RA-KATPAFQRYRIIAIIANLR
(351) IRET-GE-RLTTDDRRKIFD-IL-LPS--D--------LTWLDVAE-LGI
(401) -R-DLRGTASLTDDGERSAAKPPV-DTNR-ILQSKI-PL--WW--ANSDE
(451) R-AMIKFLSNA----D--D-D-P-DAEIA-IIAEL-E-D-DKLDSLHLPA
(501) GRAAYS-DTL--LTDHML-T--DLHEAR--LF-VAK-WAPPAP-I-EPVG
(551) NPSVDRTLKIIARWL-AM---WG-PESI-IEHVRDGFSSEA-A-E-DRDN
(601) -RRYNDN-ELL-KIQ---G-EG--SRADI-RI-ALQRQNC-CIYCG-TIT
(651) F-TCQMDHIVPRAGPGS-NKRDNLVAVC-RCNKSKSNTPFAVWAK---IP
(701) -V-LKEAL-RIR-W-KDT--MSSKDF-RFK--VIARLKRT--DEPLDNRS
(751) MESVAWMANELR-RIAA-YGEH----------KV-VYRGSITAAAR---
(801) ----------AAGIDSKL-FIDG-G-KSRLDRRHHAVDASVIALM---VA
(851) KILAERSSIR-E-------------L-KK-D-WRNFTGSTDA-RE-F--W
(901) -A---M--LTDLLN-KLAEDKI-VT-NIRLRLGNG-AH-DTI--LMS-RV
(951) GDALSVT-IDRA-T-ALWCALTRD-DFD-K-GLPANP-RRIRVHG-WFDA
(1001) DDHI-VF--A----------------------GAI-VRGGFAEIG-SI
(1051) HH-RFYKI-GKKP-------IYAMLRVFT-DL-A--------R--DLFSL
(1101) -IPPQSISMR-AEPKLRKAI-DGNAEYLGWIVVDDELEI---SF------
(1151) -------IARLL-DFP-T-RWRI-GF-SNSKL-LRPIQLAAEGL---ASA
(1201) --R------IVD--GWR-AIN-LLSALHLTVIRR-ALG-LR--SNSNLPT
(1251) SWKID----
```

图6

7A


eG
F
在这里的各种PAM
引入的终止密码子
F
P
重复区域（约500 bp）
eG
F
P

7B

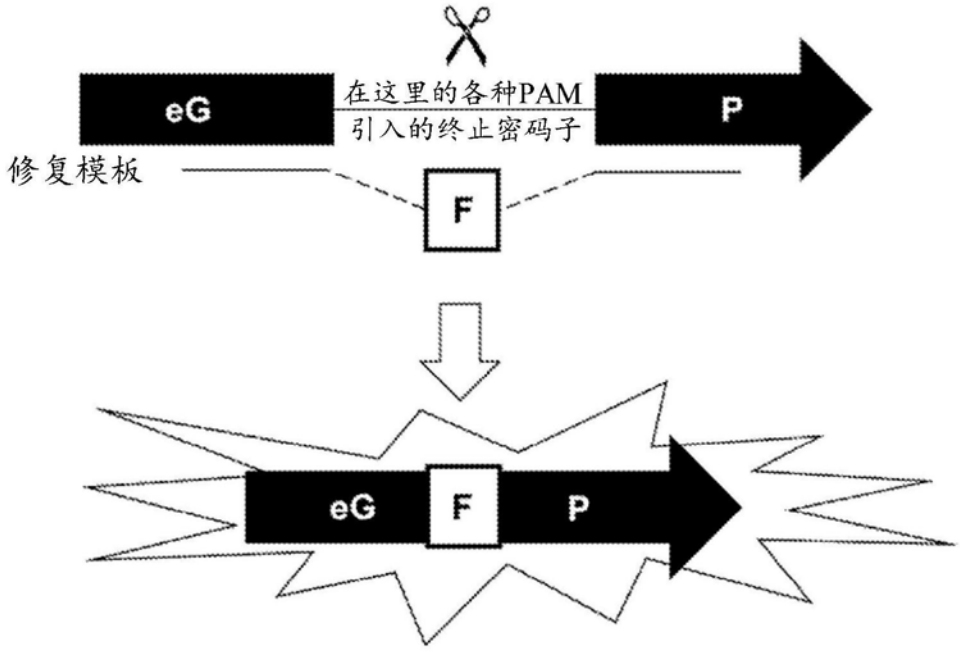
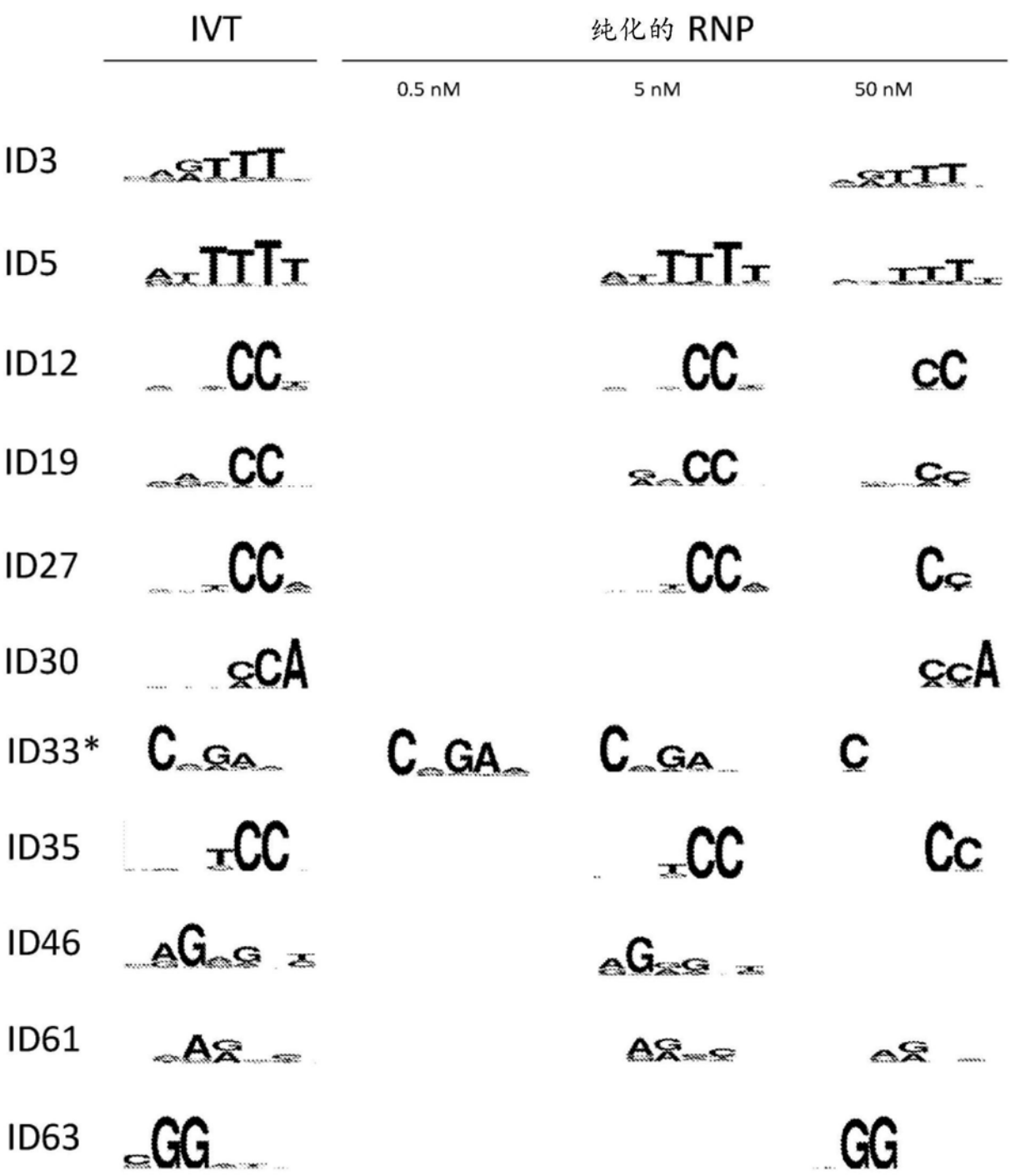

eG
在这里的各种PAM
引入的终止密码子
P
修复模板
F
eG
F
P

图7



*间隔子和PAM偏移3 nt

图8

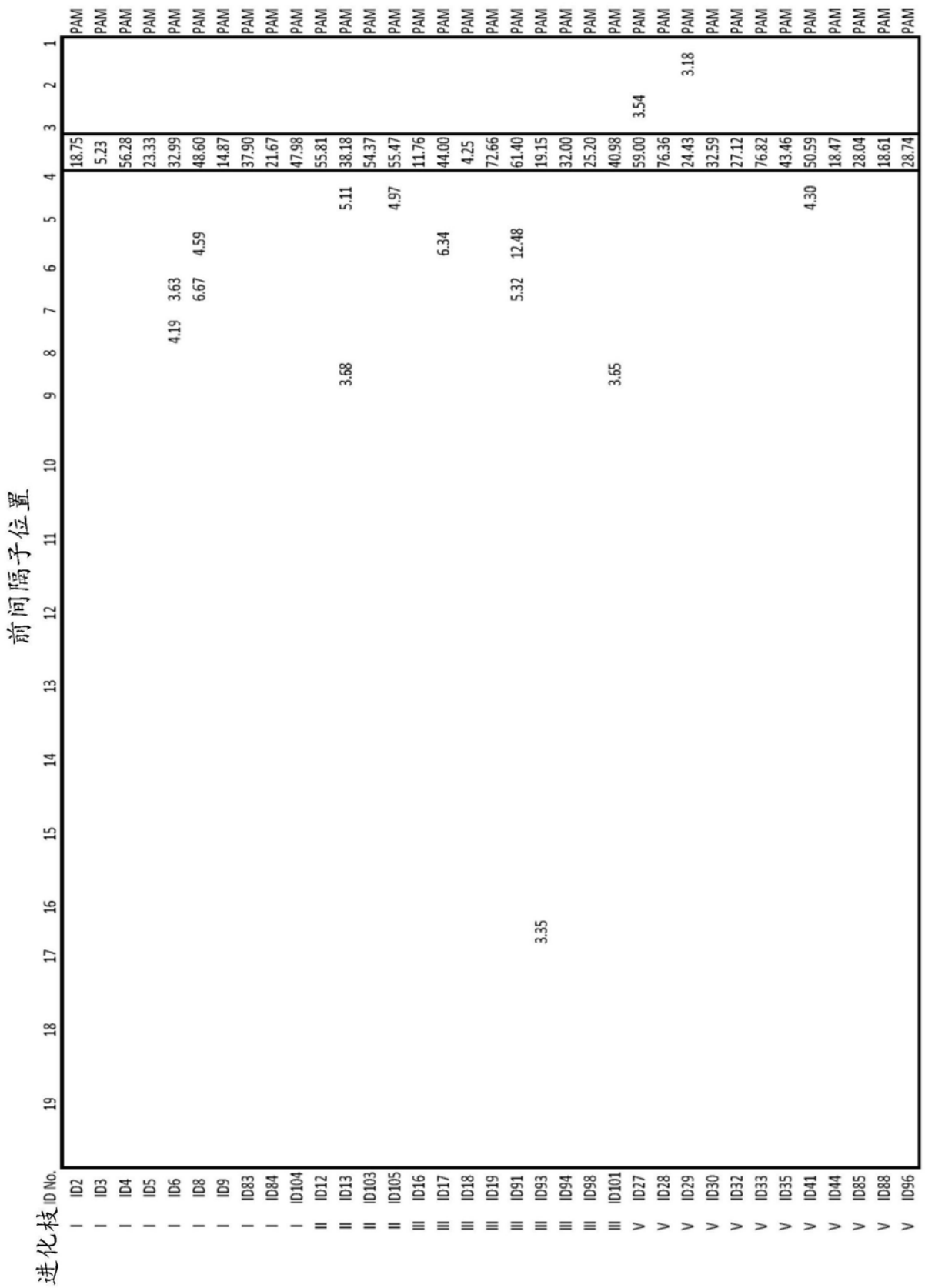

前间隔子位置
进化枝
ID No.
PAM
ID2 ID3 ID4 ID5 ID6 ID8 ID9 ID83 ID84 ID104 ID12 ID13 ID103 ID105 ID16 ID17 ID18 ID19 ID91 ID93 ID94 ID98 ID101 ID27 ID28 ID29 ID30 ID32 ID33 ID35 ID41 ID44 ID85 ID88 ID96
I I I I I I I I I I II II II II III III III III III III III III III V V V V V V V V V V V
18.75 5.23 56.28 23.33 32.99 48.60 14.87 37.90 21.67 47.98 55.81 38.18 54.37 55.47 11.76 44.00 4.25 72.66 61.40 19.15 32.00 25.20 40.98 59.00 76.36 24.43 32.59 27.12 76.82 43.46 50.59 18.47 28.04 18.61 28.74
4.19 3.63 6.67 4.59 3.68 5.11 4.97 6.34 5.32 12.48 3.35 3.65 3.54 3.18 4.30

图9A

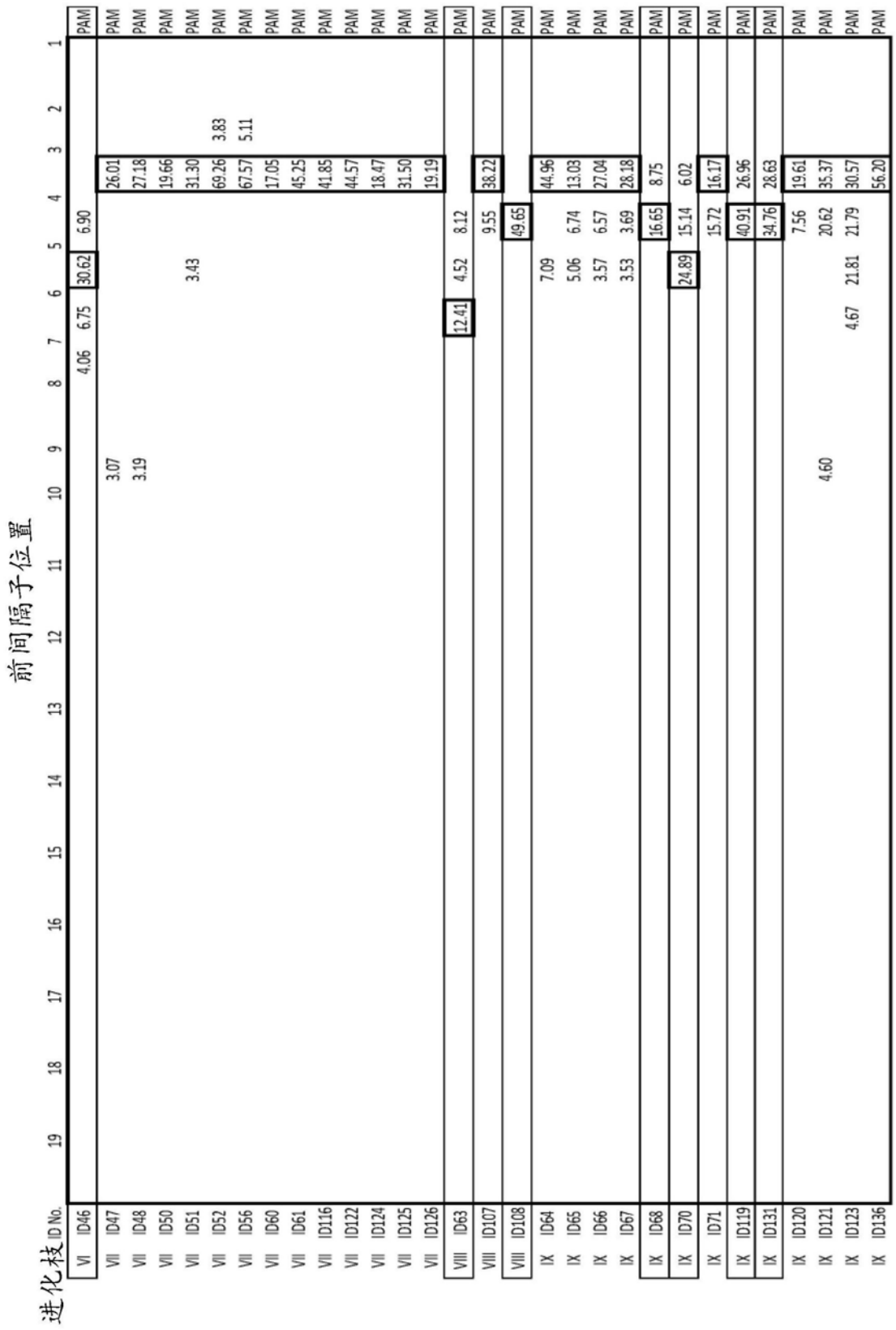


图9B

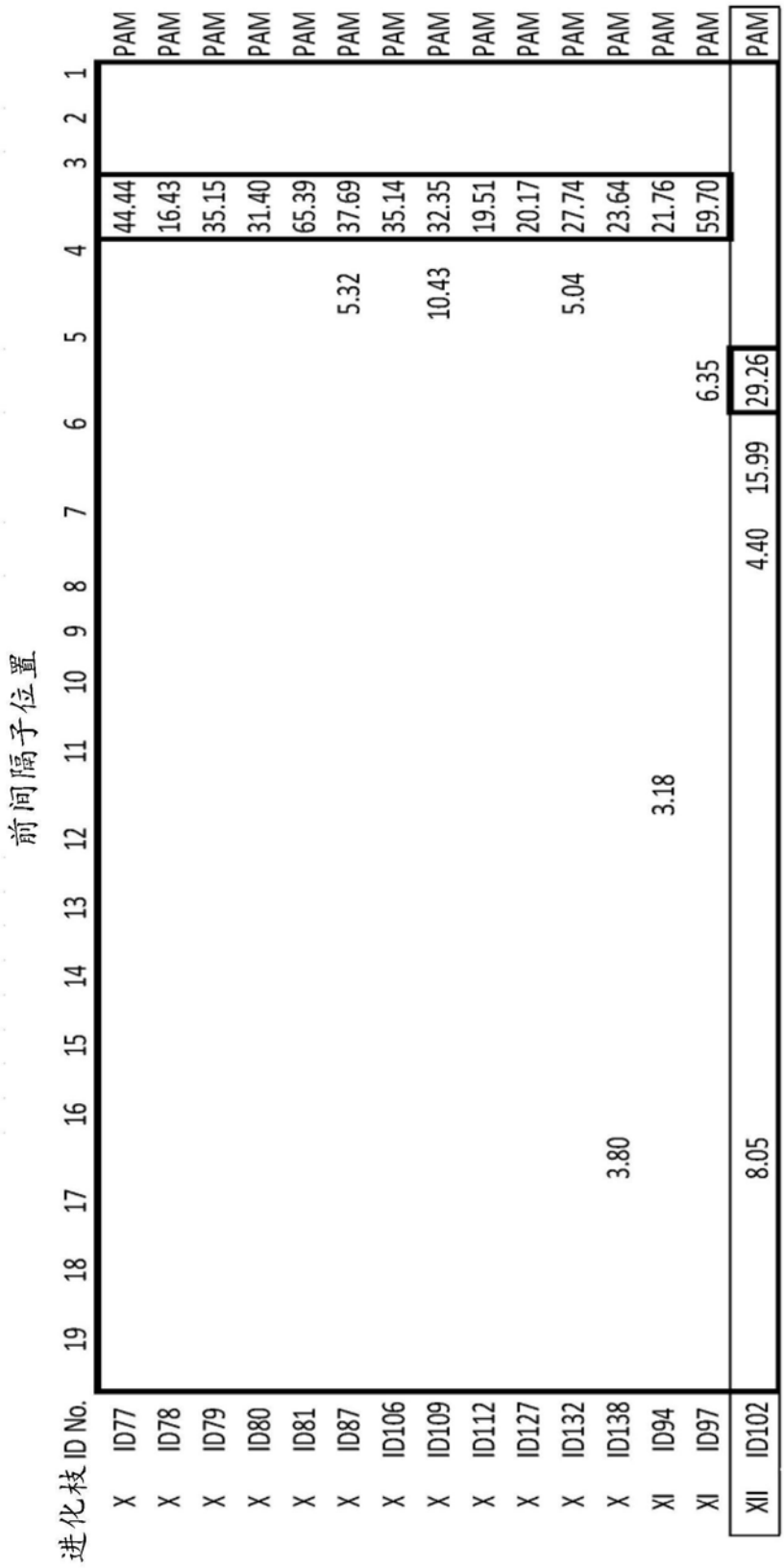

前间隔子位置
进化枝 ID No.
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
X ID77 44.44 PAM
X ID78 16.43 PAM
X ID79 35.15 PAM
X ID80 31.40 PAM
X ID81 65.39 PAM
X ID87 37.69 5.32 PAM
X ID106 35.14 PAM
X ID109 32.35 10.43 PAM
X ID112 19.51 PAM
X ID127 20.17 PAM
X ID132 27.74 5.04 PAM
X ID138 23.64 3.80 PAM
XI ID94 21.76 3.18 PAM
XI ID97 59.70 6.35 PAM
XII ID102 29.26 4.40 15.99 8.05 PAM

图9C

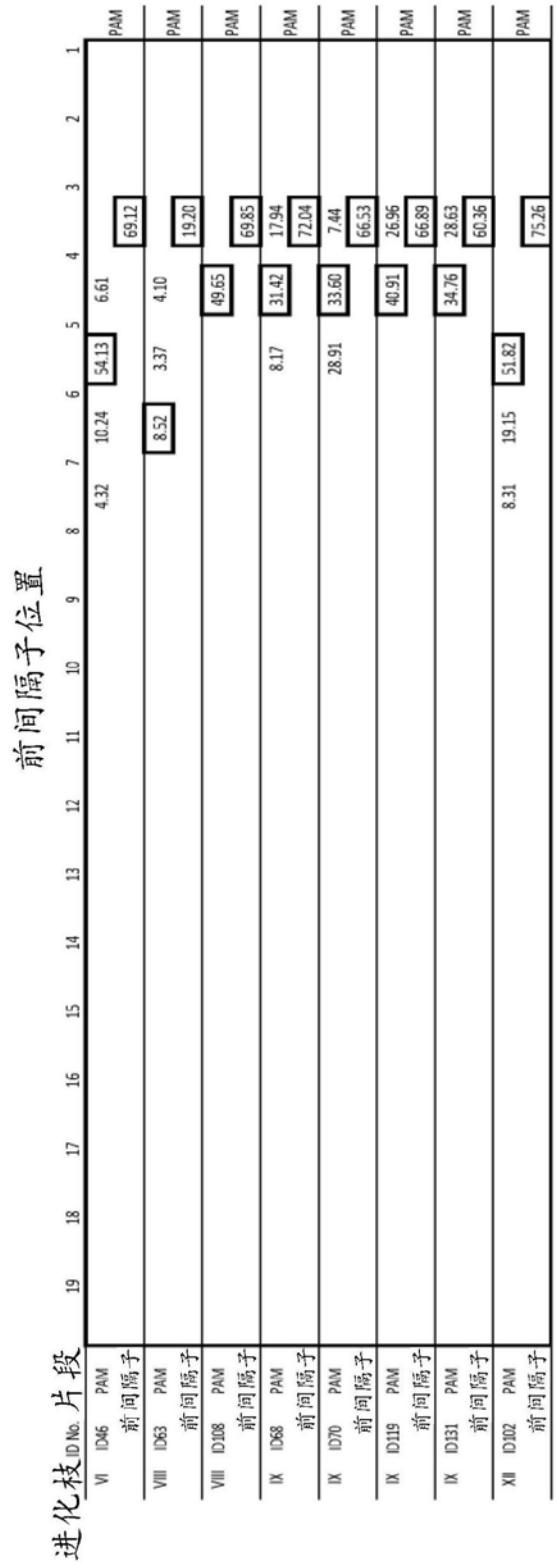

前间隔子位置
进化枝
ID No.
片段
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
VI ID46 PAM 前间隔子
4.32 10.24 54.13 6.61 69.12
VIII ID63 PAM 前间隔子
8.52 3.37 4.10 19.20
VIII ID108 PAM 前间隔子
49.65 69.85
IX ID68 PAM 前间隔子
8.17 31.42 17.94 72.04
IX ID70 PAM 前间隔子
28.91 33.60 7.44 66.53
IX ID119 PAM 前间隔子
40.91 26.96 66.89
IX ID131 PAM 前间隔子
34.76 28.63 60.36
XII ID102 PAM 前间隔子
8.31 19.15 51.82 75.26

图10A

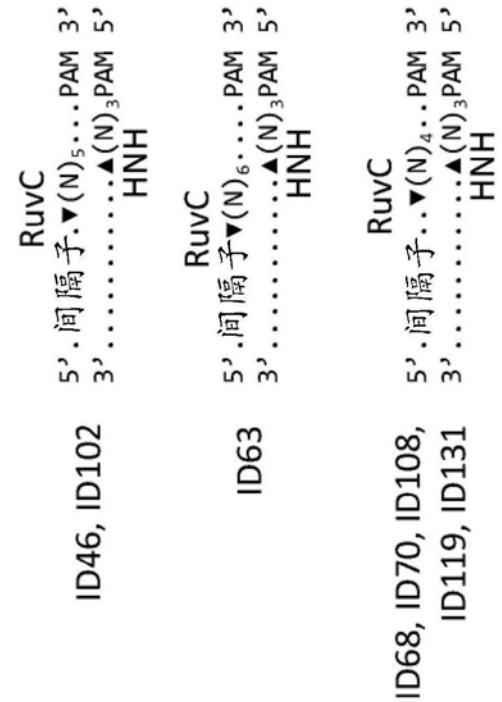

RuvC
5’.间隔子.▼(N)5....PAM 3’
3’.....▲(N)3PAM 5’
HNH
ID46, ID102
RuvC
5’.间隔子▼(N)6....PAM 3’
3’.....▲(N)3PAM 5’
HNH
ID63
RuvC
5’.间隔子..▼(N)4..PAM 3’
3’.....▲(N)3PAM 5’
HNH
ID68, ID70, ID108,
ID119, ID131

图10B

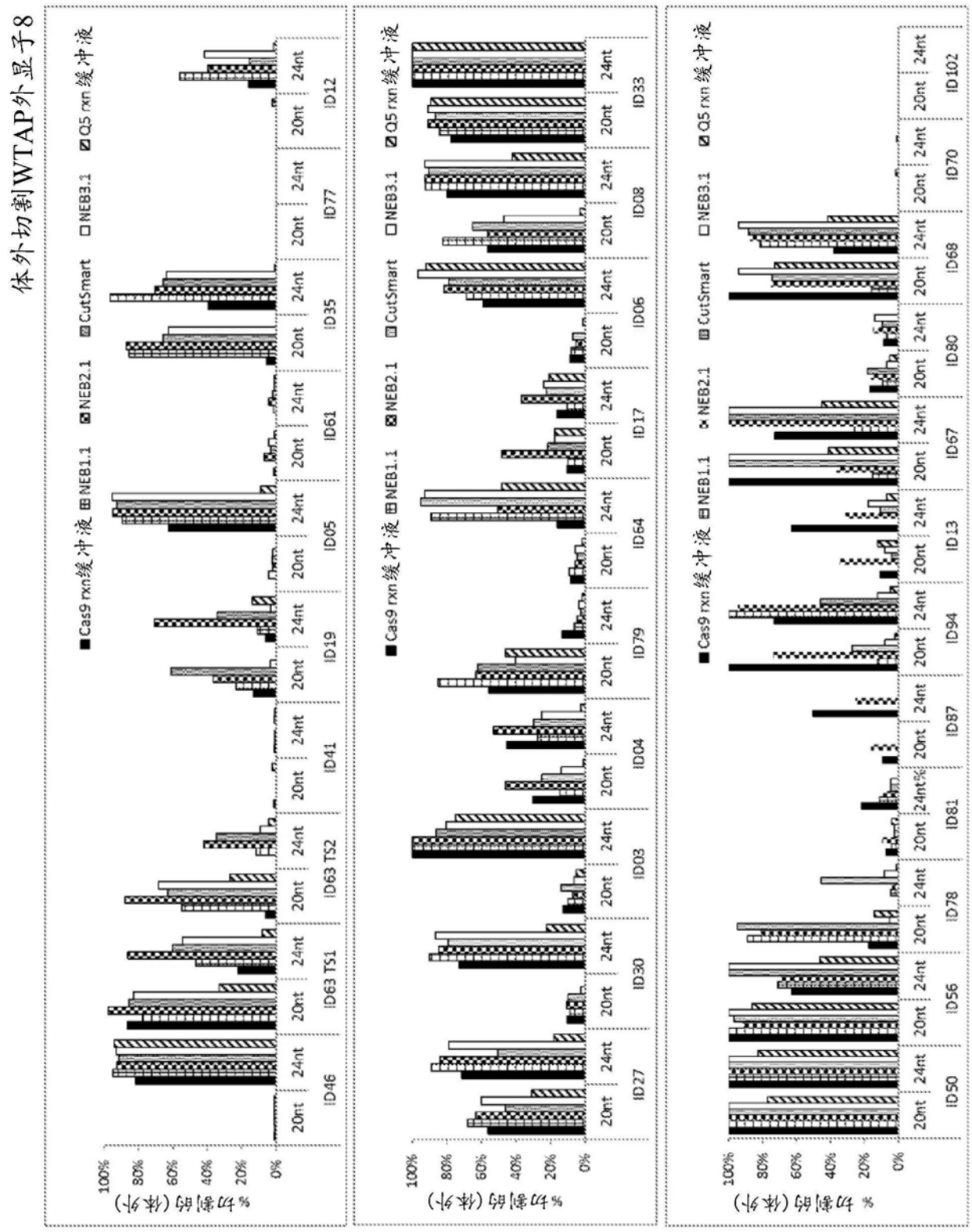

体外切割WTAP外显子8
Cas9 rxn缓冲液
NEB1.1
NEB2.1
CutSmart
NEB3.1
Q5 rxn缓冲液
% 切割的 (体外)
100%
80%
60%
40%
20%
0%
20nt
24nt
ID46
ID63 TS1
ID63 TS2
ID41
ID19
ID05
ID61
ID35
ID77
ID12
ID27
ID30
ID03
ID04
ID79
ID64
ID17
ID06
ID08
ID33
ID50
ID56
ID78
ID81
ID87
ID94
ID13
ID67
ID80
ID68
ID70
ID102

图11

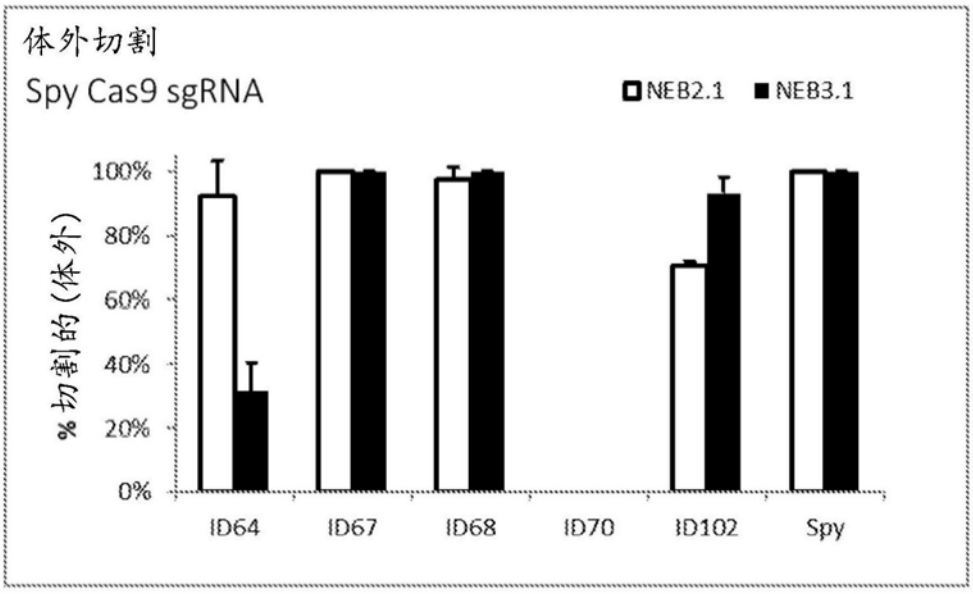

体外切割
Spy Cas9 sgRNA
NEB2.1
NEB3.1
% 切割的(体外)
100%
80%
60%
40%
20%
0%
ID64
ID67
ID68
ID70
ID102
Spy

图12A

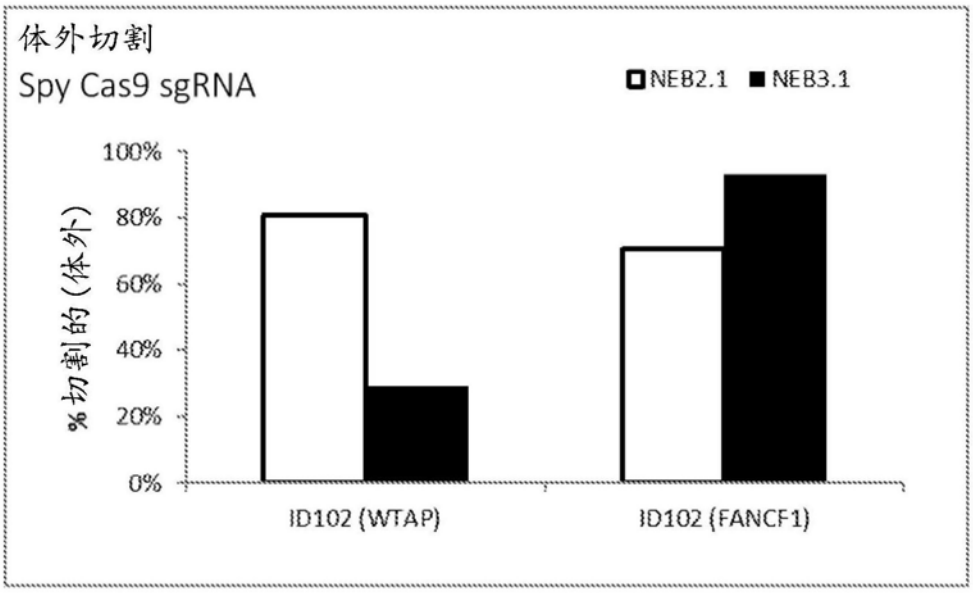

体外切割
Spy Cas9 sgRNA
NEB2.1
NEB3.1
% 切割的(体外)
100%
80%
60%
40%
20%
0%
ID102 (WTAP)
ID102 (FANCF1)

图12B

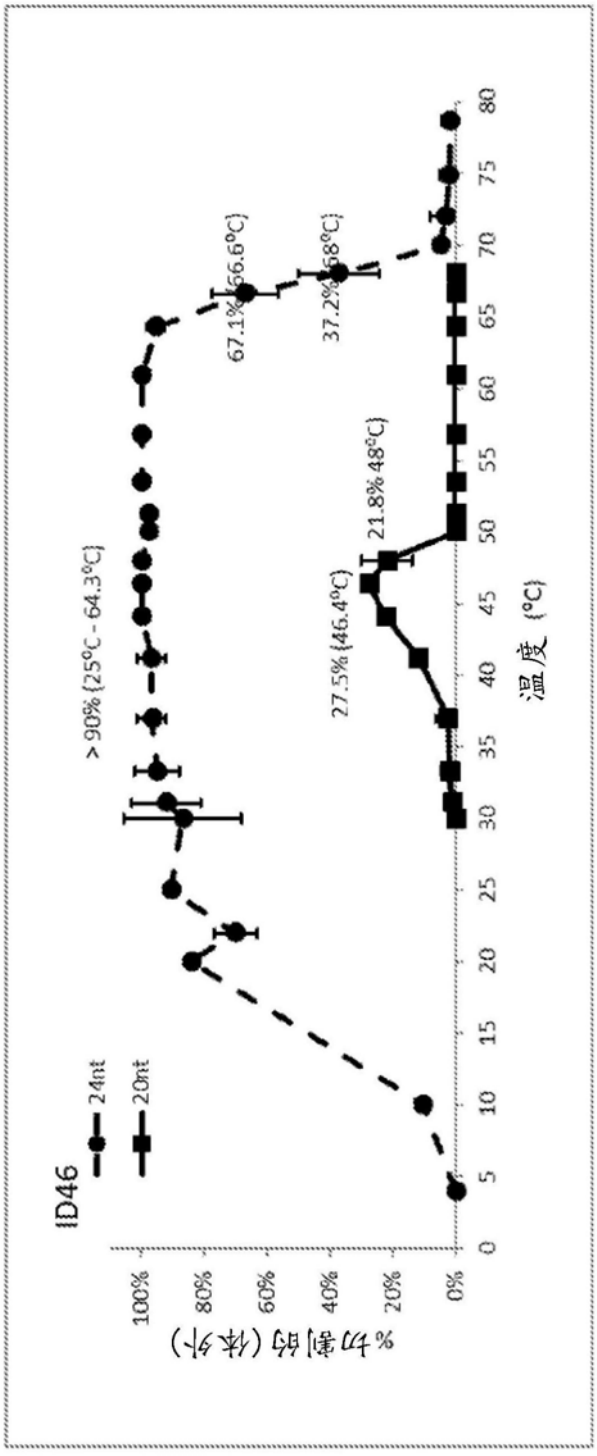

ID46
24nt
20nt
% 切割的 (体外)
100%
80%
60%
40%
20%
0%
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
温度 (°C)
> 90% (25°C - 64.3°C)
67.1% (66.6°C)
37.2% (68°C)
21.8% (48°C)
27.5% (46.4°C)

图13

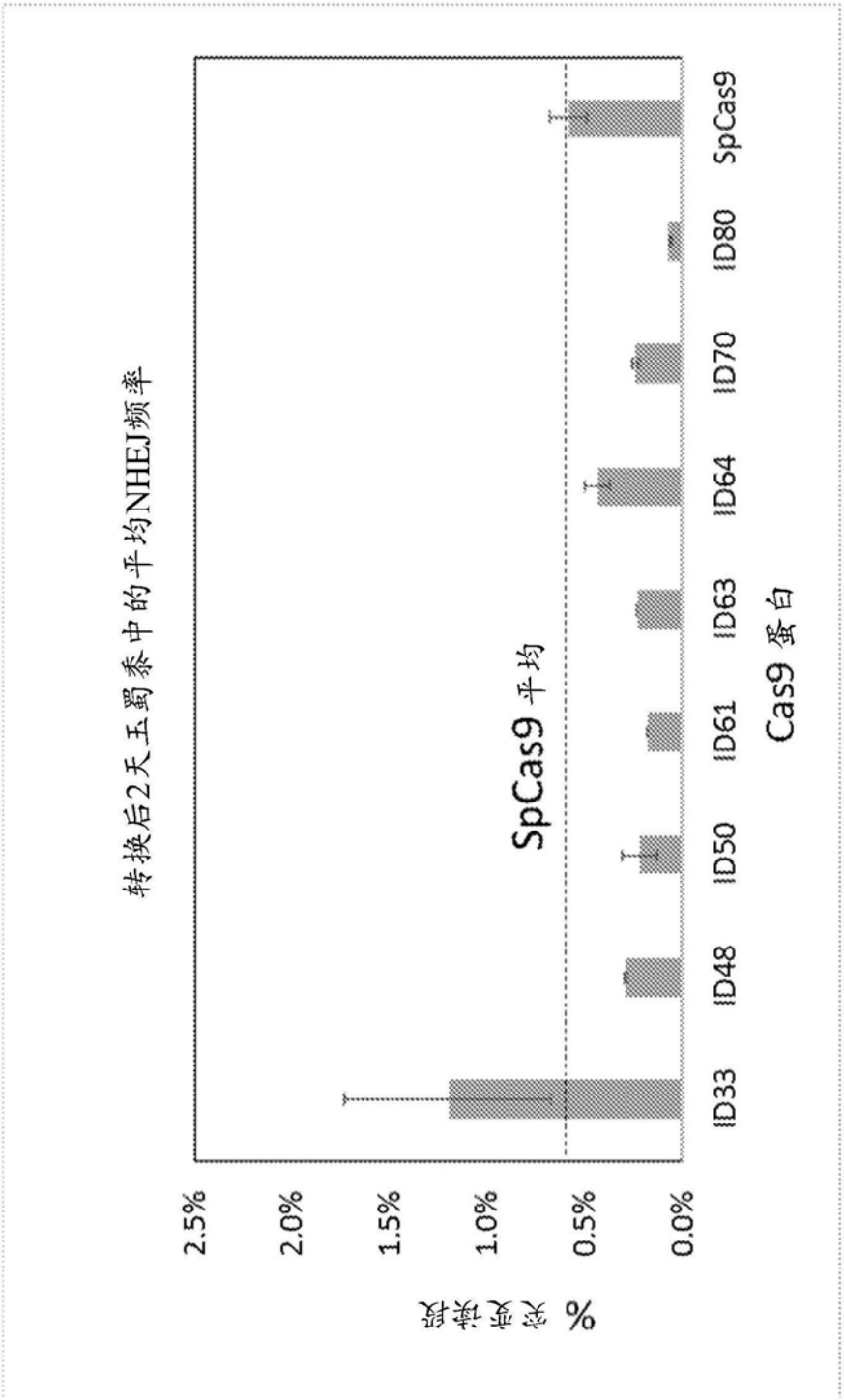

转换后2天玉蜀黍中的平均NHEJ频率
% 突变频率
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
SpCas9 平均
ID33
ID48
ID50
ID61
ID63
ID64
ID70
ID80
SpCas9
Cas9 蛋白

图14

预期剪切位点
gRNA 靶
PAM
SEQ ID NO:
读段计数
频率比Wt对照中高30X
Wt 参考
突变体 1
突变体 2
突变体 3
突变体 4
突变体 5
突变体 6
突变体 7
突变体 8
突变体 9
突变体 10
突变体 11
突变体 12
突变体 13
突变体 14
突变体 15
突变体 16
突变体 17
突变体 18
突变体 19
突变体 20
1746 1347355
1747 3267 是
1748 2320 是
1749 1859 是
1750 753 是
1751 169 是
1752 163 是
1753 156 是
1754 131 是
1755 98 是
1756 75 是
1757 70 是
1758 65 是
1759 63 是
1760 61 是
1761 54 是
1762 50 是
1763 49 是
1764 42 是
1765 41 是
1766 41 是

图15A

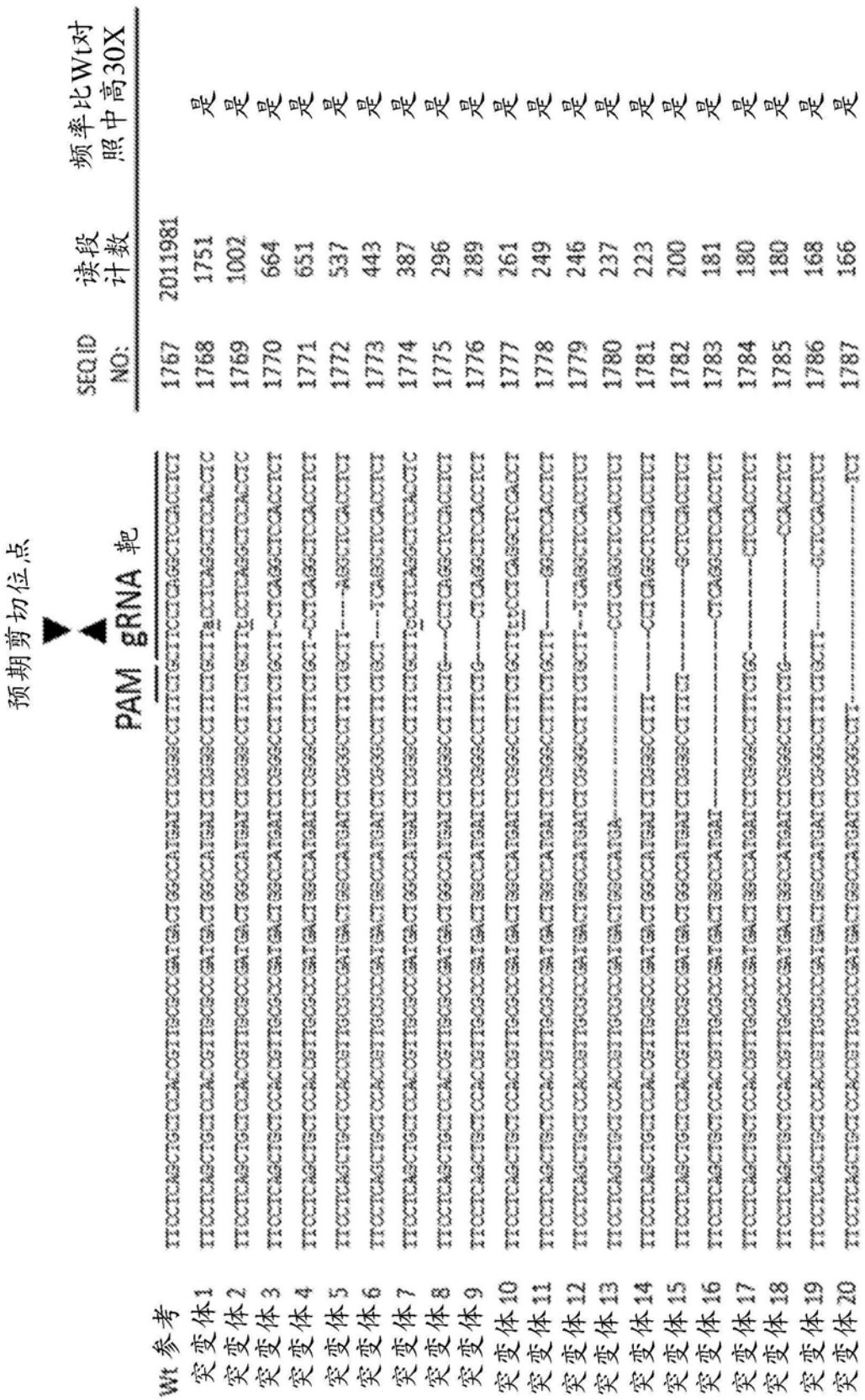

预期剪切位点
PAM gRNA 靶
SEQ ID NO:
读段计数
频率比Wt对照中高30X
Wt 参考 1767 2011981
突变体1 1768 1751 是
突变体2 1769 1002 是
突变体3 1770 664 是
突变体4 1771 651 是
突变体5 1772 537 是
突变体6 1773 443 是
突变体7 1774 387 是
突变体8 1775 296 是
突变体9 1776 289 是
突变体10 1777 261 是
突变体11 1778 249 是
突变体12 1779 246 是
突变体13 1780 237 是
突变体14 1781 223 是
突变体15 1782 200 是
突变体16 1783 181 是
突变体17 1784 180 是
突变体18 1785 180 是
突变体19 1786 168 是
突变体20 1787 166 是

图15B

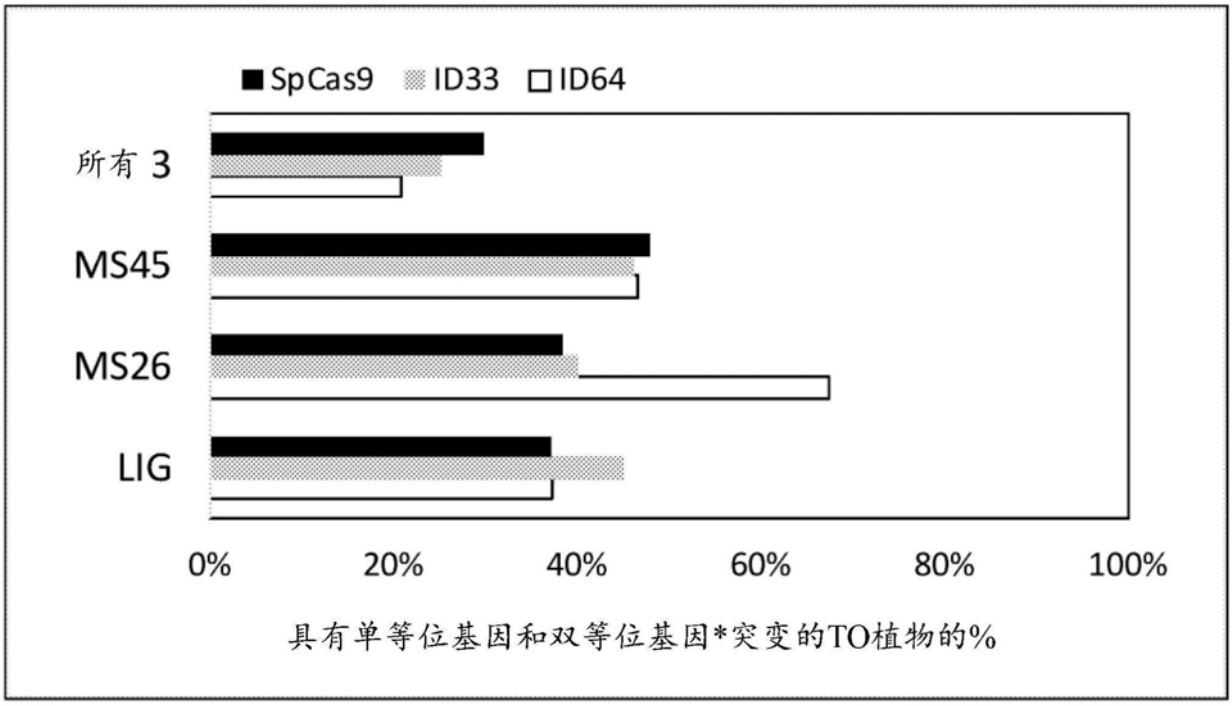
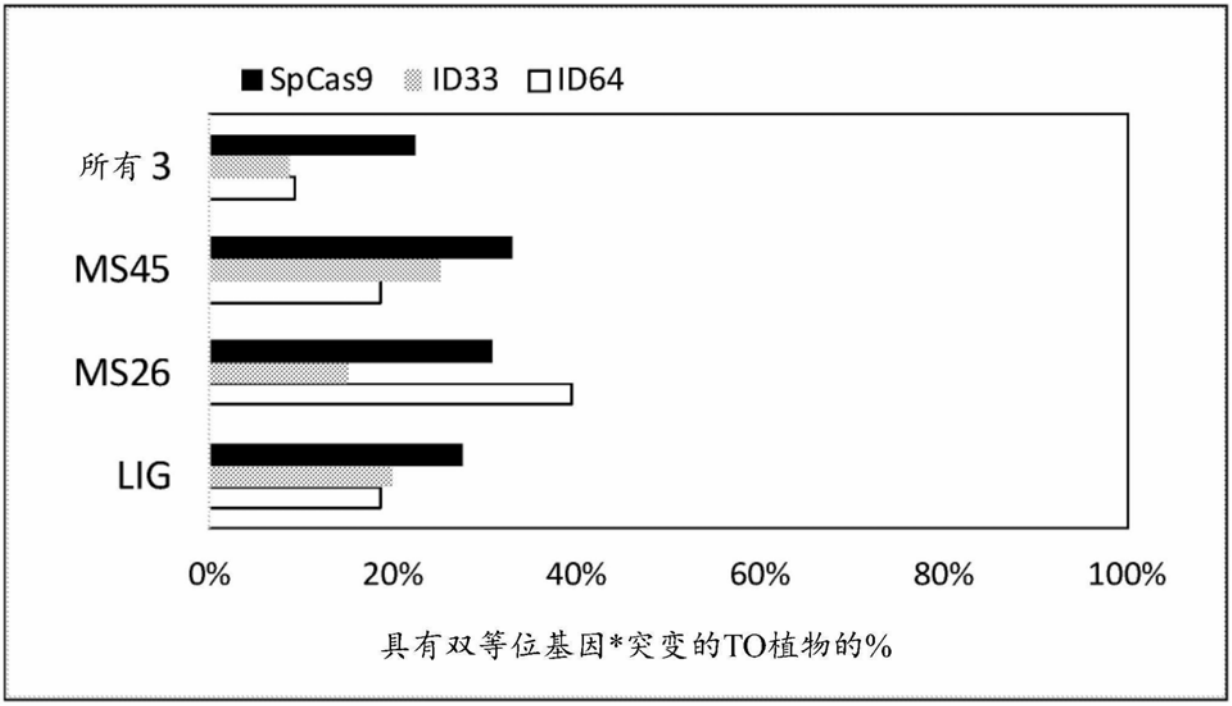


图16A



*基于不存在野生型读段

图16B

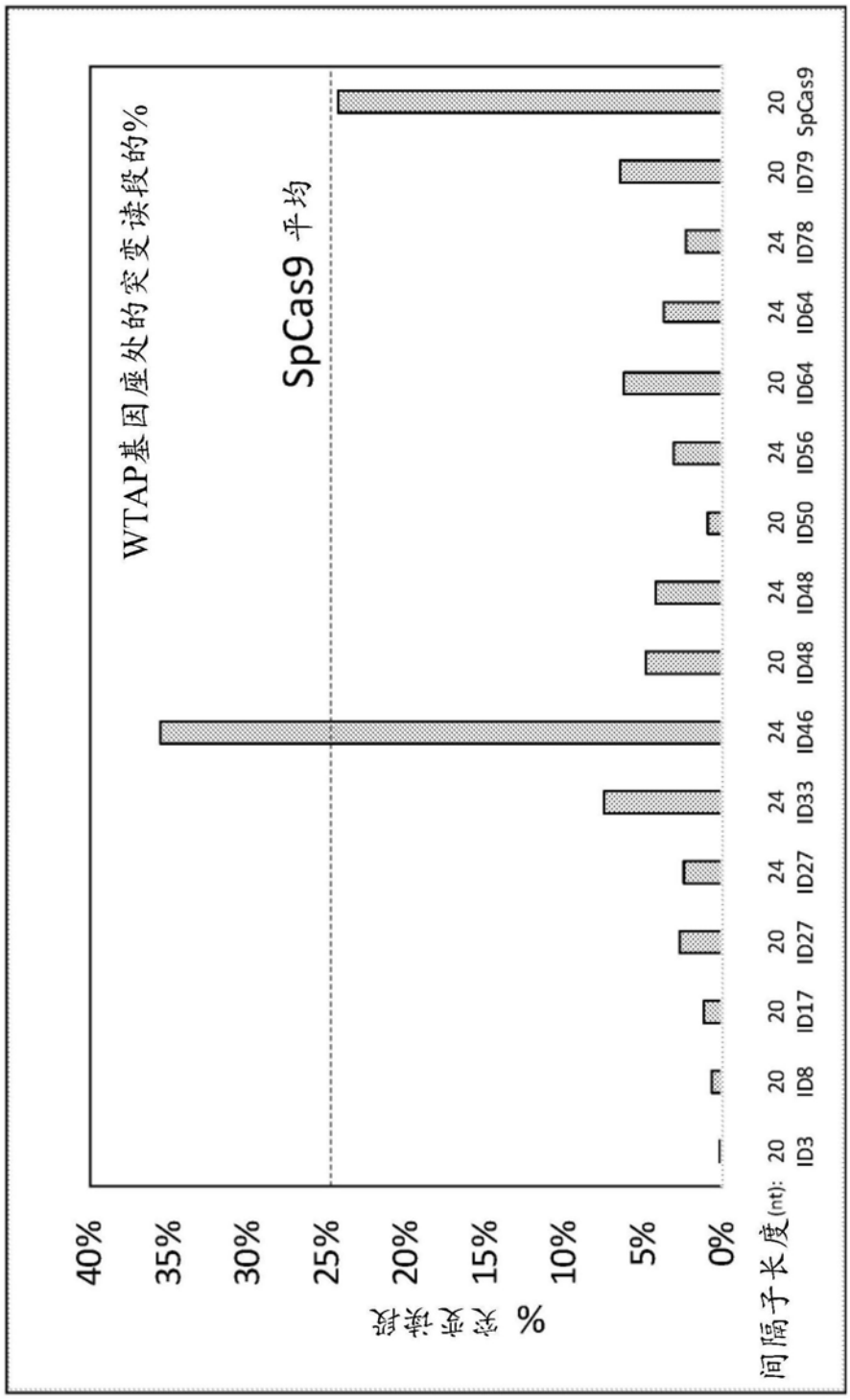

WTAP基因座处的突变读段的%
SpCas9 平均
% 突变读段
40%
35%
30%
25%
20%
15%
10%
5%
0%
间隔子长度(nt):
ID3 20
ID8 20
ID17 20
ID27 20
ID27 24
ID33 24
ID46 24
ID48 20
ID48 24
ID50 20
ID56 24
ID64 20
ID64 24
ID78 24
ID79 20
SpCas9 20

图17

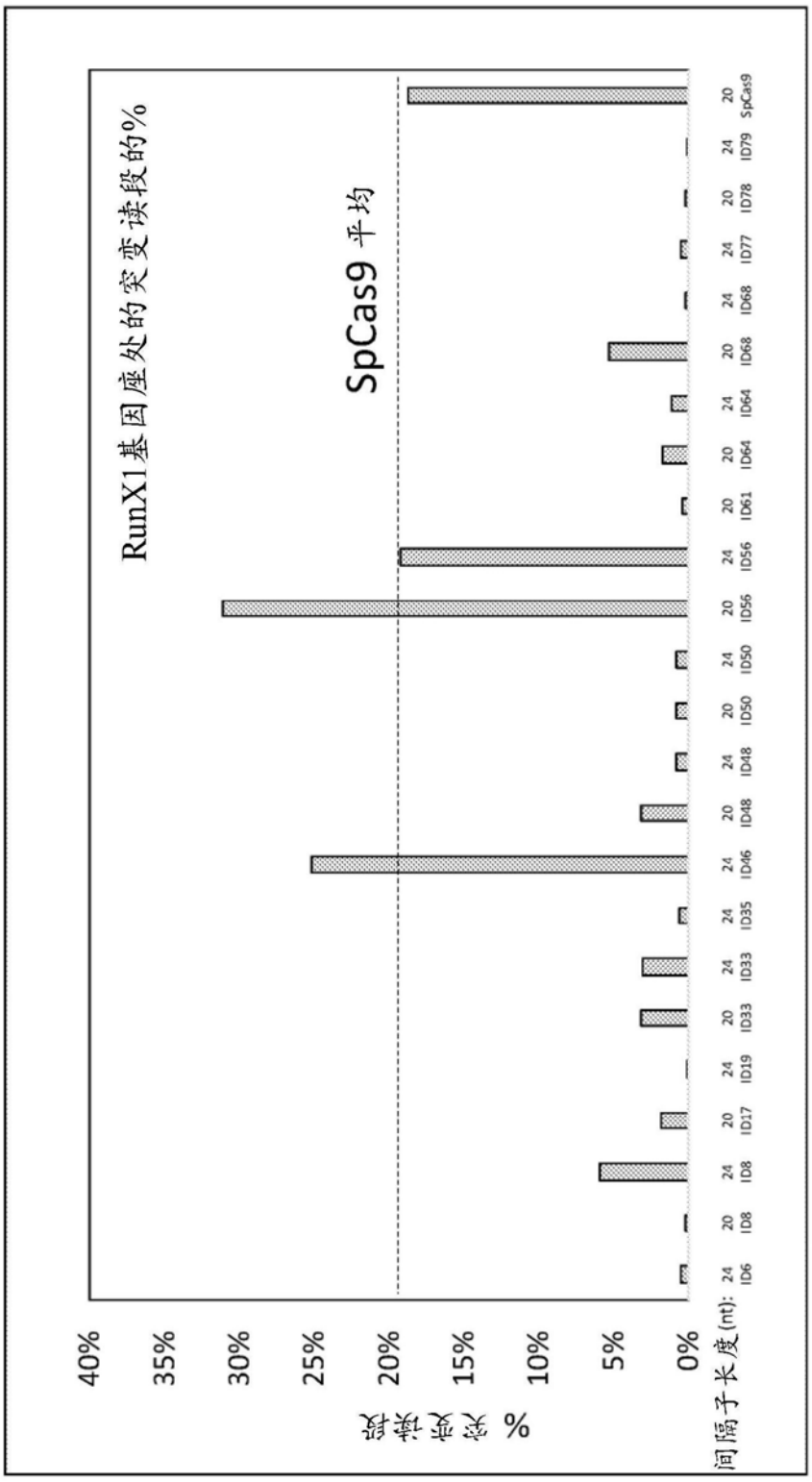

RunX1基因座处的突变读段的%
SpCas9 平均
40%
35%
30%
25%
20%
15%
10%
5%
0%
% 突变读段
间隔子长度(nt):
ID6 ID8 ID8 ID17 ID19 ID33 ID33 ID35 ID46 ID48 ID48 ID50 ID50 ID56 ID56 ID61 ID64 ID64 ID68 ID68 ID77 ID78 ID79 SpCas9

图18

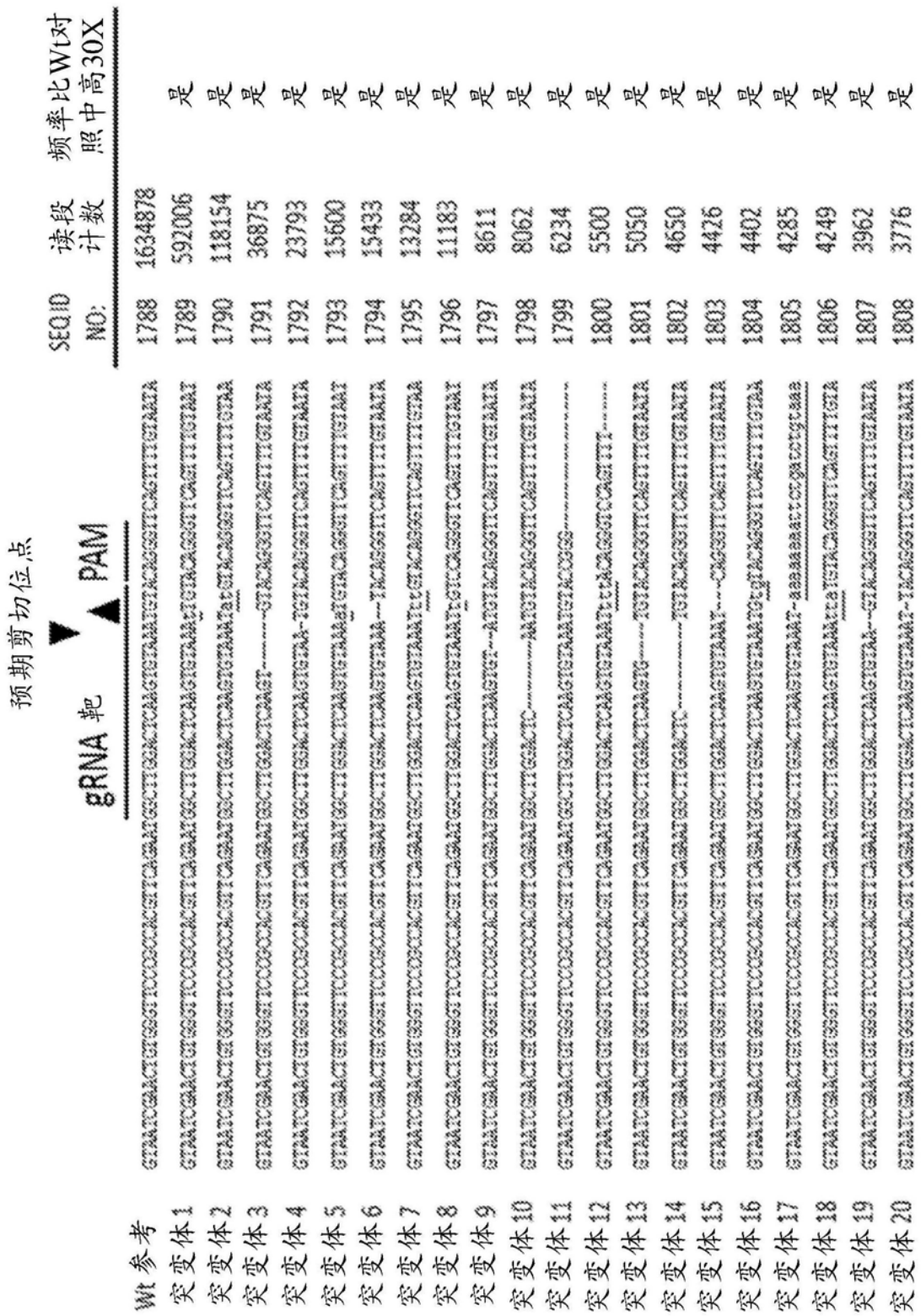

预期剪切位点
gRNA 靶
PAM
SEQ ID NO:
读段计数
频率比Wt对照中高30X
Wt 参考
突变体1
突变体2
突变体3
突变体4
突变体5
突变体6
突变体7
突变体8
突变体9
突变体10
突变体11
突变体12
突变体13
突变体14
突变体15
突变体16
突变体17
突变体18
突变体19
突变体20
1788 1634878
1789 592006 是
1790 118154 是
1791 36875 是
1792 23793 是
1793 15600 是
1794 15433 是
1795 13284 是
1796 11183 是
1797 8611 是
1798 8062 是
1799 6234 是
1800 5500 是
1801 5050 是
1802 4650 是
1803 4426 是
1804 4402 是
1805 4285 是
1806 4249 是
1807 3962 是
1808 3776 是

图19A

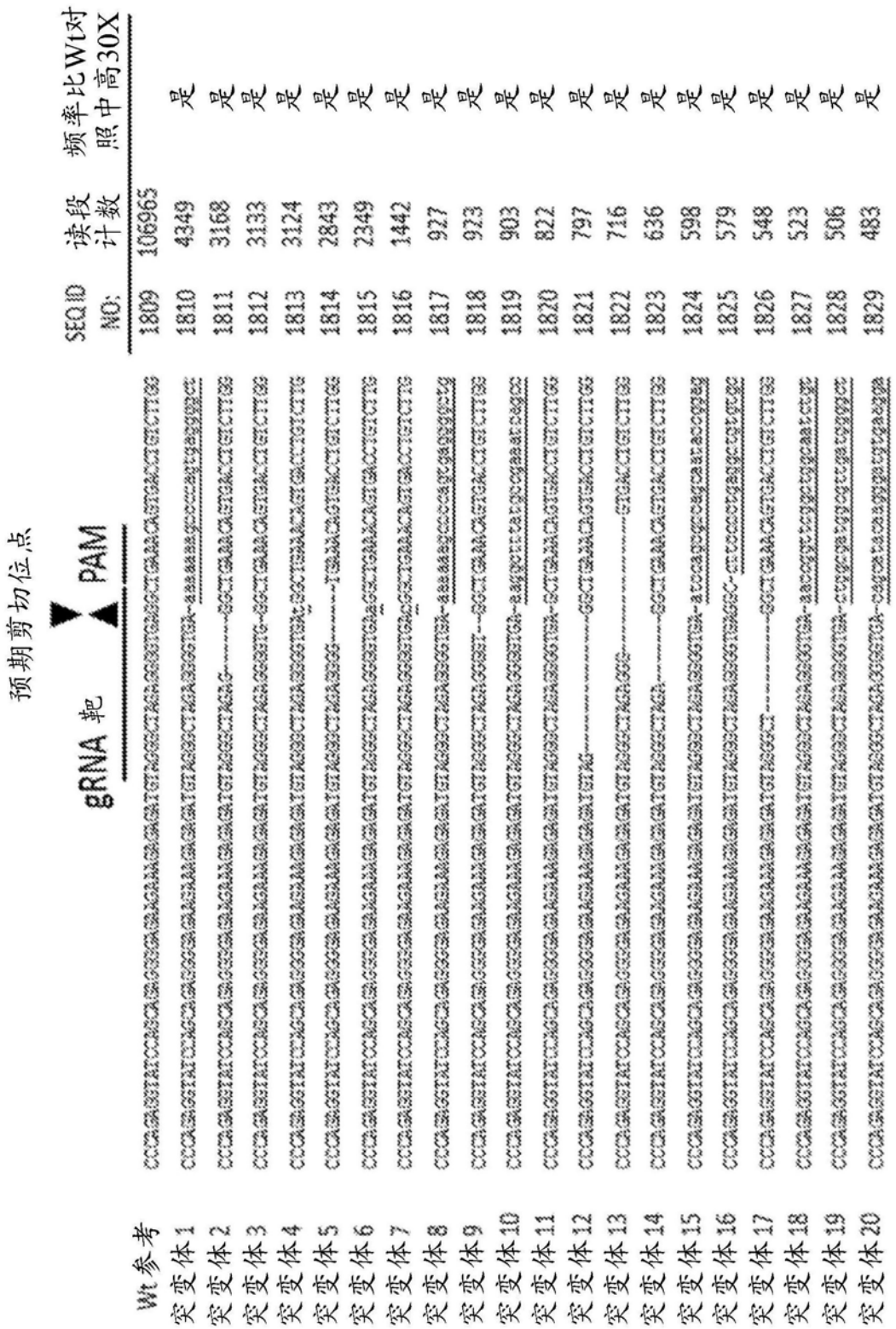

预期剪切位点
gRNA 靶
PAM
SEQ ID NO:
读段计数
频率比Wt对照中高30X
Wt 参考
突变体 1
突变体 2
突变体 3
突变体 4
突变体 5
突变体 6
突变体 7
突变体 8
突变体 9
突变体 10
突变体 11
突变体 12
突变体 13
突变体 14
突变体 15
突变体 16
突变体 17
突变体 18
突变体 19
突变体 20
1809 106965
1810 4349 是
1811 3168 是
1812 3133 是
1813 3124 是
1814 2843 是
1815 2349 是
1816 1442 是
1817 927 是
1818 923 是
1819 903 是
1820 822 是
1821 797 是
1822 716 是
1823 636 是
1824 598 是
1825 579 是
1826 548 是
1827 523 是
1828 506 是
1829 483 是

图19B

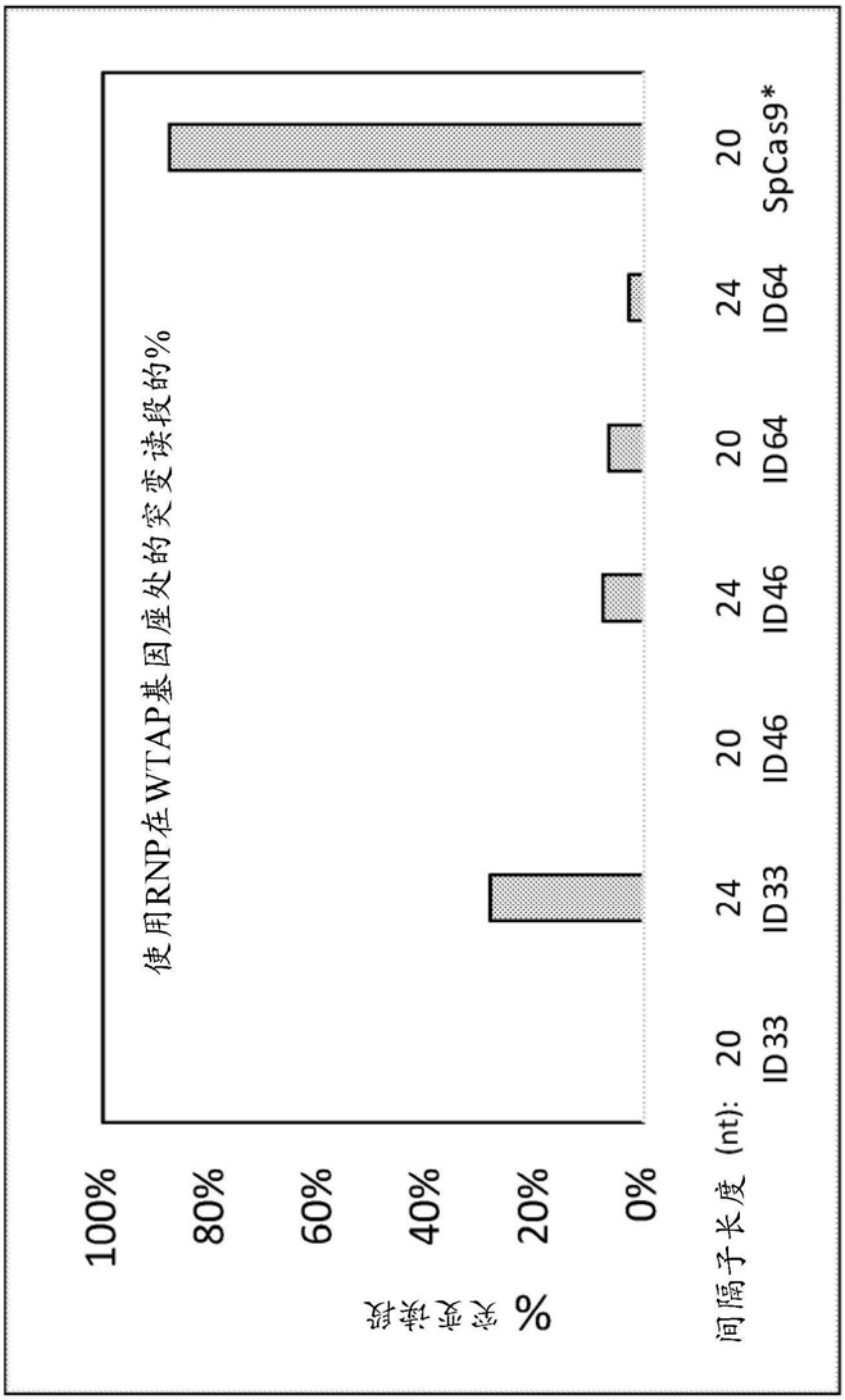

使用RNP在WTAP基因座处的突变读段的%
%突变读段
100%
80%
60%
40%
20%
0%
间隔子长度 (nt):
20 24 20 24 20 24 20
ID33 ID33 ID46 ID46 ID64 ID64 SpCas9*
*仅SpCas9蛋白包含NLS

图20

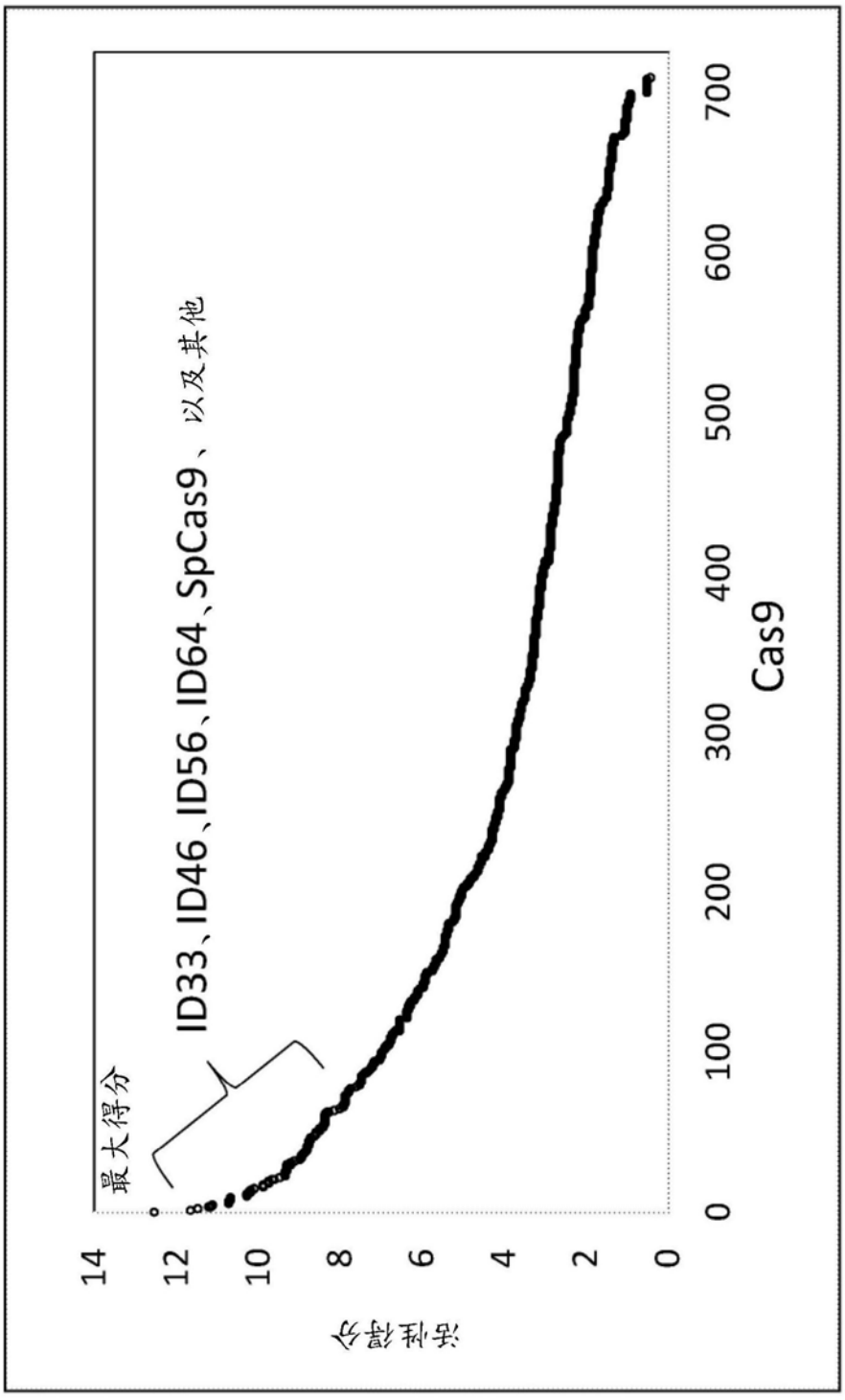

活性得分
14
12
10
8
6
4
2
0
0
100
200
300
400
500
600
700
Cas9
最大得分
ID33、ID46、ID56、ID64、SpCas9、以及其他

图21