(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111373040 B
(45)授权公告日 2024.07.12

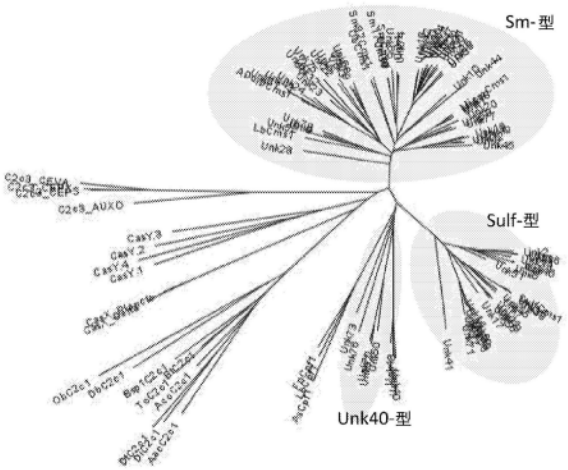
(21)申请号 201880065634.1
(22)申请日 2018.08.08
(65)同一申请的已公布的文献号
申请公布号 CN 111373040 A
(43)申请公布日 2020.07.03
(30)优先权数据
62/542,983 2017.08.09 US
62/551,958 2017.08.30 US
62/565,255 2017.09.29 US
62/599,226 2017.12.15 US
(85)PCT国际申请进入国家阶段日
2020.04.08
(86)PCT国际申请的申请数据
PCT/IB2018/055972 2018.08.08
(87)PCT国际申请的公布数据
WO2019/030695 EN 2019.02.14
(73)专利权人 本森希尔股份有限公司
地址 美国密苏里州
(72)发明人 M·贝其曼 B·N·格雷
(74)专利代理机构 上海专利商标事务所有限公司 31100
专利代理师 钱文宇 陈扬扬
(51)Int.Cl.
C12N 15/10(2006.01)
C12N 15/63(2006.01)
C12N 15/82(2006.01)
(56)对比文件
CN 105142669 A,2015.12.09
CN 109312316 A,2019.02.05
US 2011165679 A1,2011.07.07
US 2017114351 A1,2017.04.27
Michael A. Estrella等.RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex.Genes & Dev. .2016,第460-470页.
审查员 林海生
权利要求书1页 说明书53页
序列表（电子公布） 附图4页

(54)发明名称

修饰基因组的组合物和方法

(57)摘要

提供了用于修饰基因组DNA序列的组合物和方法。该方法在基因组DNA序列中的预定靶位点生成双链断裂(DSB),在基因组中的靶位点导致DNA序列的突变、插入和/或缺失。组合物包括DNA构建体,其包括编码Csm1蛋白的核苷酸序列,其操作性连接至在感兴趣细胞中可操作的启动子。DNA构建体可以用于指导在预定基因组基因座的基因组DNA修饰。本文描述了使用这些DNA构建体来修饰基因组DNA序列的方法。此外,提供了用于调节基因表达的组合物和方法。组合物包括DNA构建体,其包括在感兴趣细胞中操作性的启动子,其操作性地连接至编码消除了生成DSB能力的突变型Cms1蛋白的核苷酸序列,任选地连接至调节转录活性的结构域。该方法可以用于上调或下调预定基因组基因座处的基因表达。



CN 111373040 B

1.一种修饰植物细胞的靶位点的核苷酸序列的方法,其包括:

向所述植物细胞中引入

(i)引导RNA(gRNA),或编码gRNA的DNA多核苷酸,其中所述gRNA包括:(a)第一区段,其包含与靶位点处核苷酸序列互补的核苷酸序列;和(b)第二区段,其与Cms1多肽相互作用;和

(ii)Cms1多肽或编码Cms1多肽的多核苷酸,其中所述Cms1多肽包含:(a)结合RNA的部分,该结合RNA的部分与gRNA相互作用;和(b)活性部分,其显示定点酶促活性,

其中,所述Cms1多肽具有SEQ ID NO:62所示的氨基酸序列,其中所述方法在所述靶位点处修饰所述核苷酸序列,并且其中所述方法用于非治疗目的。

2.如权利要求1所述的方法,其还包括:

在表达所述Cms1多肽并在所述靶位点处切割核苷酸序列以生成经修饰的核苷酸序列的条件下培养所述植物细胞,以产生植物;和

选择包含所述修饰的核苷酸序列的植物。

3.如权利要求1所述的方法,其中,所述修饰的核苷酸序列包括细胞基因组中异源性DNA的插入,细胞基因组中核苷酸序列的缺失,或细胞基因组中至少一个核苷酸的突变。

4.如权利要求1所述的方法,其中,所述修饰的核苷酸序列包括多核苷酸的插入,所述多核苷酸编码向转化的细胞赋予抗生素或除草剂耐受性的蛋白质。

5.如权利要求4所述的方法,其中,所述编码赋予抗生素或除草剂耐受性的蛋白质的多核苷酸包括SEQ ID NO:7,或编码包括SEQ ID NO:8的蛋白质。

6.一种核酸分子,其包含编码Cms1多肽的多核苷酸序列,其中所述多核苷酸序列具有SEQ ID NO:142所示的核苷酸序列,或者其中,所述多核苷酸序列编码具有SEQ ID NO:62所示氨基酸序列的Cms1多肽,并且其中,编码Cms1多肽的所述多核苷酸序列操作性地连接至启动子,所述启动子相对于编码Cms1多肽的多核苷酸序列而言是异源性的。

7.由权利要求6所述的核酸分子编码的Cms1多肽。

8.包括权利要求6所述的核酸分子的真核细胞或原核细胞,其中所述真核细胞不是植物细胞、胚胎干细胞、生殖细胞或受精卵。

9.如权利要求6所述的核酸分子,其中,编码Cms1多肽的所述多核苷酸序列经密码子优化以在植物细胞中表达。

10.如权利要求1所述的方法,其中,编码Cms1多肽的所述多核苷酸经密码子优化以在植物细胞中表达。

11.如权利要求1所述的方法,其中所述gRNA是靶向DNA的RNA。

12.如权利要求1所述的方法,其中所述靶位点在植物细胞基因组中,并且其中所述基因组是核、质体或线粒体基因组。

修饰基因组的组合物和方法

发明领域

[0001] 本发明涉及用于在预选位置编辑基因组序列和用于调节基因表达的组合物和方法。

[0002] 关于通过EFS-WEB以文本文件形式提交的序列表

[0003] 序列表的正式文本通过EFS-Web以按照美国信息交换标准码(ASCII)的文本文件与说明书同时提交,文件名为BHP017P5 Sequence Listing_ST25.txt,创建日期为2018年8月3日,大小为1,848Kb。通过EFS-Web提交的该序列表是本说明书的一部分且通过引用全文纳入本文。

背景技术

[0004] 基因组DNA的修饰对于基础和应用研究是极其重要的。基因组修饰有可能说明并且在一些情况中疗愈病因,以及在包括此类修饰的个体和/或细胞中提供所需特性。基因组修饰可以包括例如植物、动物、真菌的修饰,和/或原核基因组修饰。修饰基因组DNA的最常用方法趋向于在基因组内的随机位点修饰DNA,但是最近的发现使位点特异性基因组修饰成为可能。此类技术依赖于在所需位点上产生DSB。该DSB导致将宿主细胞的原生DNA修复机制募集到DSB。可以控制DNA修复机制,以在预定位点插入异源性DNA,以使原生植物基因组DNA缺失,或以在所需位点生成点突变、插入或缺失。对于位点特异性基因组修饰特别感兴趣的是成簇规律间隔短回文重复序列(CRISPR)核酸酶。CRISPR核酸酶使用引导分子,通常是引导RNA分子,它与核酸酶相互作用并且与靶向的DNA碱基配对,从而允许核酸酶在所需位点产生双链断裂(DSB)。DSB的产生要求原型间隔子-邻近基序(PAM)序列的存在;在PAM序列的识别之后,CRISPR核酸酶能够产生所需DSB。Cms1 CRISPR核酸酶是一类CRISPR核酸酶,相对于其它CRISPR核酸酶(例如Cas9核酸酶)具有某些所需性质。

[0005] 实施基因组修饰的一个领域是植物基因组DNA的修饰。植物基因组DNA的修饰对于基础和应用植物学研究是极其重要的。具有稳定修饰的基因组DNA的转基因植物可以具有新的性状,如除草剂耐受,抗虫性,和/或积累有价值蛋白质,包括它们提供的药用蛋白质和工业酶。原生植物基因的表达可能会被上调或下调或以其他方式改变(例如,通过改变表达原生植物基因的组织),它们的表达可能会被完全消除,DNA序列可能会被改变(例如,通过点突变、插入或缺失),或新的非原生基因可能会被插入植物基因组,从而将新的性状赋予植物。

发明内容

[0006] 提供了使用Cms1 CRISPR系统进行基因组DNA序列修饰的组合物和方法。本文所用基因组DNA表示线性和/或染色体DNA和/或感兴趣的一种或多种细胞中存在的质粒或其它染色体外DNA序列。该方法在基因组DNA序列中的预定靶位点生成双链断裂(DSB),在基因组中的靶位点导致DNA序列的突变、插入和/或缺失。组合物包括DNA构建体,其包括编码Csm1蛋白的核苷酸序列,其操作性连接至在感兴趣细胞中可操作的启动子。在一些实施方式中,

Cms1蛋白包含选自下组的至少一个氨基酸基序:SEQ ID NO:177-186。在其它实施方式中,Cms1蛋白包含选自SEQ ID NO:288-289和187-201的至少一个氨基酸基序。在其它实施方式中,Cms1蛋白包含选自下组的至少一个氨基酸基序:SEQ ID NO:290-296。在某些优选的实施方式中,Cms1蛋白包含选自下组的多于一个氨基酸基序:SEQ ID NO:177-186。在某些优选的实施方式中,Cms1蛋白包含选自下组的多于一个氨基酸基序:SEQ ID NO:288-289和187-201。在某些优选的实施方式中,Cms1蛋白包含选自下组的多于一个氨基酸基序:SEQ ID NO:290-296。具体的Cms1蛋白序列列于SEQ ID NO:10、11、20-23、30-69、154-156、208-211和222-254;具体的Cms1蛋白编码多核苷酸序列列于SEQ ID NO:16-19、24-27、70-146、174-176、212-215和255-287。在某些优选的实施方式中,Cms1蛋白与选自下组的序列具有至少80%的相同性:SEQ ID NO:16-19、24-27、70-146、174-176、212-215和255-287。包含编码本发明的Cms1蛋白的多核苷酸序列的DNA构建体或本发明的Cms1蛋白本身可用于在预定的基因组基因座上指导基因组DNA的修饰。本文描述了使用这些DNA构建体来修饰基因组DNA序列的方法。本文还涵盖经修饰的真核生物和真核细胞,包括酵母,变形虫,昆虫,真菌,哺乳动物,植物,植物细胞,植物部分和种子,以及经修饰的原核生物,包括细菌和古细菌。还提供了用于调节基因表达的组合物和方法。该方法靶向蛋白质至基因组中预定位点以实现上调或下调一种或多种基因,其表达由基因组中靶向的位点调节。组合物包括含有核苷酸序列的DNA构建体,所述核苷酸序列编码具有减弱或消失的核酸酶活性的修饰的Csm1蛋白,任选地融合转录激活或抑制结构域。本文描述了使用这些DNA构建体来修饰基因表达的方法。

附图说明

[0007] 图1显示了从指示的V型核酸酶氨基酸序列的RuvC锚着的MUSCLE比对得出的系统发育树。显示了Sm型,Sulf型和Unk40型Cms1核酸酶。

[0008] 图2显示了Sm型Cms1蛋白之间共有的氨基酸基序概述。框1-10中的weblogo图分别对应于SEQ ID NO:177-186,并显示了它们在SmCms1蛋白(SEQ ID NO:10)上的位置。

[0009] 图3显示了Sulf型Cms1蛋白之间共有的氨基酸基序概述。框1-17中的weblogo图分别对应于SEQ ID NO:288-289和SEQ ID NO:187-201,并显示了它们在SulfCms1蛋白(SEQ ID NO:11)上的位置。

[0010] 图4显示了Unk40型Cms1蛋白之间共有的氨基酸基序的概述。框1-7中的weblogo图分别对应于SEQ ID NO:290-296,并显示了它们在Unk40Cms1蛋白(SEQ ID NO:68)上的位置。

[0011] 发明详述

[0012] 本文提供了用于控制基因表达的方法和组合物,涉及与CRISPR-Cms系统及其组件有关的序列靶向(例如基因组干扰或基因编辑)。本发明的CRISPR酶选自Cms酶,例如,Cms1直系同源物或突变的Cms1酶。Cms1是源自小基因组菌(Microgenomates)和史密斯氏菌(Smithella)的CRISPR的缩写,之所以这样命名,是因为这些组中的某些细菌物种编码Cms1核酸酶;术语Csm1和Cms1在本文中可互换使用。Cms1核酸酶也可以称为Cas12f核酸酶。该方法和组合物包括核酸以结合靶DNA序列。这是有利的,因为生产核酸相比生产(例如)肽要容易且成本低得多,并且特异性可根据所需同源性的延伸段(stretch)的长度而不同。例如,

不要求具有复杂的多指3D定位。

[0013] 还提供编码Cms1多肽的核酸,以及使用Cms1多肽来修饰宿主细胞(包括植物细胞)染色体(即,基因组)或细胞器DNA序列的方法。Cms1多肽与特定的引导RNA(gRNA)相互作用,其将Cms1内切核酸酶引导至特定的靶位点,Cms1内切核酸酶在此处引入双链断裂,该双链断裂可通过DNA修复过程修复,从而修饰DNA序列。因为特异性由引导RNA提供,所以Cms1多肽是通用的,并且可与不同引导RNA联用以靶向不同的基因组序列。相较于CRISPR阵列常规使用的Cas核酸酶(例如,Cas9)而言,Cms1内切核酸酶具有某些优势。例如,Cms1相关的CRISPR阵列能被加工为成熟的crRNA而无需其它反式活化crRNA(tracrRNA)。此外,Cms1-crRNA复合物能够切割前方具有通常富含T的短原型间隔子(protospacer)-邻近基序(PAM)的靶DNA,这与许多Cas9系统中在靶DNA之后具有富含G的PAM形成对比。此外,Cms1核酸酶可引入交错的DNA双链断裂。本文公开的方法可以用于靶向和修饰特定染色体序列和/或在真核和原核细胞基因组中的靶位置处引入外源序列。所述方法还可用于引入序列或修饰细胞器(例如,叶绿体和/或线粒体)中的区域。此外,靶向是特异性的,脱靶效应有限。

[0014] I.Cms1内切核酸酶

[0015] 本文提供了用于修饰基因组(包括植物基因组)的Cms1内切核酸酶及其片段和变体。本文中所用术语Cms1内切核酸酶或Cms1多肽指SEQ ID NO:10、11、20-23、30-69、154-156、208-211和222-254中所示的Cms1多肽序列的同源物、直系同源物和变体。通常,Cms1内切核酸酶可在不使用tracrRNA的情况下起作用,并且可引入交错的DNA双链断裂。通常,Cms1多肽包含至少一个RNA识别和/或RNA结合结构域。RNA识别和/或RNA结合结构域与引导RNA相互作用。通常,引导RNA包含具有与Cms1多肽相互作用的茎环结构的区域。该茎环通常包含序列UCUACN$_{3-5}$GUAGAU(SEQ ID NO:312-314,由SEQ ID NO:315-317编码),带有“UCUAC”和“GUAGA”碱基配对以形成茎-环的茎。N$_{3-5}$表示在该位置可存在任何碱基,并且在该位置可包含3、4或5个核苷酸。Cms1多肽还可包括核酸酶结构域(即,DNA酶或RNA酶结构域),DNA结合结构域,解旋酶结构域,RNA酶结构域,蛋白质-蛋白质相互作用结构域,二聚化结构域,以及其它结构域。在特定的实施方式中,Cms1多肽或编码Cms1多肽的多核苷酸包含:与靶向DNA的RNA相互作用的RNA结合部分,和显示定点酶促活性的活性部分,例如RuvC内切核酸酶结构域。

[0016] Cms1多肽可以是野生型Cms1多肽,修饰的Cms1多肽或野生型或修饰的Cms1多肽的片段。Cms1多肽可经修饰以增加核酸结合亲和性和/或特异性,改变酶活性,和/或改变该蛋白质的另一性质。例如,可对Cms1多肽的核酸酶(即,DNA酶,RNA酶)结构域进行修饰、使之缺失或失活。或者,可将Cms1多肽截短以去除对蛋白质功能非必需的结构域。

[0017] 在一些实施方式中,Cms1多肽可衍生自野生型Cms1多肽或其片段。在其它实施方式中,Cms1多肽可衍生自经修饰的Cms1多肽。例如,Cms1多肽的氨基酸序列可经修饰以改变该蛋白质的一种或多种性质(例如,核酸酶活性,亲和性,稳定性等)。或者,可消除该蛋白质中不参与RNA引导的切割的Cms1多肽的结构域,从而使经修饰的Cms1多肽小于野生型Cms1多肽。

[0018] 通常,Cms1多肽包含至少一个核酸酶(即DNA酶)结构域,但不需要包含HNH结构域,例如Cas9蛋白中存在的一个。例如,Cms1多肽可包含RuvC或RuvC样核酸酶结构域。在一些实施方式中,Cms1多肽可经修饰以使核酸酶结构域失活,从而使其不再起作用。在其中核酸酶

结构域之一是失活的一些实施方式中,Cms1多肽不切割双链DNA。在特定实施方式中,当以使核酸酶活性减小或消除的最大相同性比对时,突变的Cms1多肽在对应于SmCms1(SEQ ID NO:10)的701或922位或SulfCms1(SEQ ID NO:11)的848和1213位的位置包含一个或多个突变。可以使用众所周知的方法,例如定点诱变,PCR介导的诱变和总基因合成,以及本领域已知的其它方法来修饰核酸酶结构域。具有失活的核酸酶结构域的Cms1蛋白(dCms1蛋白)可用于调节基因表达而无需修饰DNA序列。在某些实施方式中,可以通过使用合适的gRNA将dCms1蛋白靶向基因组的特定区域,例如感兴趣的一个或多个基因的启动子。dCms1蛋白可结合至所需DNA区域,并可干扰RNA聚合酶与DNA的该区域结合和/或干扰转录因子与DNA该区域结合。该技术可用于上调或下调一个或多个感兴趣的基因的表达。在某些其它实施方式中,dCms1蛋白可与阻抑物结构域融合,以进一步下调一种或多种基因的表达,所述一种或多种基因的表达被RNA聚合酶、转录因子或其它转录调节物与gRNA靶向的染色体DNA区域间的相互作用所调节。在某些其它实施方式中,dCms1蛋白可与活化结构域融合以上调一种或多种基因的表达,所述一种或多种基因的表达被RNA聚合酶、转录因子或其它转录调节物与gRNA靶向的染色体DNA区域间的相互作用所调节。

[0019] 本文所公开的Cms1多肽还可包含至少一个核定位信号(NLS)。NLS通常包含一段碱性氨基酸。本领域已知核定位信号(参见,例如,Lange等,J.Biol.Chem.(2007)282:5101-5105)。NLS可以定位于Cms1多肽的N末端,C末端,或内部位置。在一些实施方式中,Cms1多肽还可包含至少一个细胞穿透性结构域。细胞穿透性结构域可定位于该蛋白质的N末端,C末端,或内部位置。

[0020] 本文所公开的Cms1多肽还可包含至少一种质体靶向信号肽,至少一种线粒体靶向信号肽,或使Cms1多肽靶向质体和线粒体两者的信号肽。本领域已知质体、线粒体和双靶向信号肽定位信号(参见,例如,Nassoury和Morse(2005)Biochim Biophys Acta 1743:5-19;Kunze和Berger(2015)Front Physiol 6:259;Herrmann和Neupert(2003)IUBMB Life 55:219-225;Soll(2002)Curr Opin Plant Biol 5:529-535;Carrie和Small(2013)Biochim Biophys Acta 1833:253-259;Carrie等(2009)FEBS J 276:1187-1195;Silva-Filho(2003)Curr Opin Plant Biol6:589-595;Peeters和Small(2001)Biochim Biophys Acta 1541:54-63;Murcha等(2014)J Exp Bot 65:6301-6335;Mackenzie(2005)Trends Cell Biol 15:548-554;Glaser等(1998)Plant Mol Biol 38:311-338)。质体、线粒体或双靶向信号肽可以定位于Cms1多肽的N末端,C末端,或内部位置。

[0021] 在其他实施方式中,Cms1多肽还可以还包括至少一个标志物结构域。标志物结构域的非限制性示例包括荧光蛋白,纯化标签和表位标签。在某些实施方式中,标志物结构域可以是荧光蛋白。合适的荧光蛋白的非限制性示例包括绿色荧光蛋白(例如GFP,GFP-2,tagGFP,turboGFP,EGFP,Emerald,Azami Green,单体型Azami Green,CopGFP,AceGFP,ZsGreen1),黄色荧光蛋白(例如YFP,EYFP,Citrine,Venus,YPet,PhiYFP,ZsYellow1),蓝色荧光蛋白(例如EBFP,EBFP2,Azurite,mKalama1,GFPuv,Sapphire,T-sapphire),青色荧光蛋白(例如ECFP,Cerulean,CyPet,AmCyan1,Midoriishi-Cyan),红色荧光蛋白(mKate,mKate2,mPlum,DsRed单体,mCherry,mRFP1,DsRed-Express,DsRed2,DsRed-单体,HcRed-Tandem,HcRed1,AsRed2,eqFP611,mRasberry,mStrawberry,Jred),和橙色荧光蛋白(mOrange,mKO,Kusabira-Orange,Monomeric Kusabira-Orange,mTangerine,tdTomato),

或任何其它合适的荧光蛋白。在其他实施方式中,标志物结构域可以是纯化标签和/或表位标签。示例性的标签包括但不限于,谷胱甘肽S-转移酶(GST)、甲壳素结合蛋白(CBP)、麦芽糖结合蛋白质、硫氧还蛋白(TRX)、多聚(NANP)、串联亲和纯化(TAP)标签、myc、AcV5、AU1、AU5、E、ECS、E2、FLAG、HA、nus、Softag 1、Softag3、Strep、SBP、Glu-Glu、HSV、KT3、S、S1、T7、V5、VSV-G、6xHis、生物素羧基载体蛋白(BCCP)和钙调蛋白。

[0022] 在某些实施方式中,Cms1多肽可以是含有引导RNA的蛋白质-RNA复合物的一部分。引导RNA与Cms1多肽相互作用将Cms1多肽引导至特定靶位点,其中引导RNA的5'端可与植物基因组中感兴趣的核苷酸序列的特定原型间隔子序列碱基配对,可以是核、质体和/或线粒体基因组的任何部分。本文所用术语"靶向DNA的RNA"表示这样的引导RNA,其与植物细胞基因组中感兴趣的核苷酸序列靶位点以及Cms1多肽相互作用。靶向DNA的RNA,或编码靶向DNA的RNA的DNA多核苷酸,可包括:包含与靶DNA中序列互补的核苷酸序列的第一区段,以及与Cms1多肽相互作用的第二区段。

[0023] 本文公开的编码Cms1多肽的多核苷酸可用于从其它原核或真核生物或从原生宿主生物不明或未知的宏基因组来源的序列分离相应的序列。由此,PCR、杂交等方法可用于根据此类序列与本文所示序列的序列同源性或相同性来鉴定该此类序列。本发明涵盖基于与本文所述的整个Cms1序列或其变体和片段的序列同一性而分离的序列。此类序列包括本文公开的Cms1序列的直向同源物的序列。"直向同源物"是指源自共同祖先基因且由于物种形成而在不同物种中发现的基因。当在不同物种中发现的基因的核苷酸序列和/或它们的编码蛋白质序列具有至少约75%、约80%、约85%、约90%、约91%、约92%、约93%、约94%、约95%、约96%、约97%、约98%、约99%或更大的序列同一性时,它们被认为是直向同源物。直向同源物的功能通常在物种之间高度保守。因此,本发明涵盖分离的多核苷酸,其编码具有Cms1内切核酸酶活性的多肽,并且与本文公开的序列具有至少约75%或更大的序列同一性。如本文所用,Cms1内切核酸酶活性指CRISPR内切核酸酶活性,其中,与Cms1多肽关联的引导RNA(gRNA)引起Cms1-gRNA复合物结合至预定的核苷酸序列,该核苷酸序列与gRNA互补;并且其中,Cms1活性可在gRNA靶向的位点处或附近引入双链断裂。在某些实施方式中,该双链断裂可以是交错的DNA双链断裂。本文所用"交错的DNA双链断裂"可以使双链断裂在切割后在3'或5'端上具有约1个、约2个、约3个、约4个、约5个、约6个、约7个、约8个、约9个或约10个核苷酸的突出端。在特定的实施方式中,Cms1多肽引入具有5'突出端的交错的DNA双链断裂。该双链断裂可以发生在靶向DNA靶向的RNA(例如,引导RNA)序列所靶向的序列处或其附近。

[0024] 本文涵盖Cms1多核苷酸和由此编码的Cms1氨基酸序列(其保留Cms1核酸酶活性)的片段和变体。"Cms1核酸酶活性"意在表示由引导RNA介导的预定DNA序列的结合。在其中Cms1核酸酶保留功能性RuvC结构域的实施方式中,Cms1核酸酶活性还可包括双链断裂诱导。"片段"是指多核苷酸的部分或氨基酸序列的部分。"变体"是指基本相似的序列。对于多核苷酸,变体包括具有以下的多核苷酸:在5'和/或3'端处的缺失(即,截短);在原生多核苷酸中一个或多个内部位点处一个或多个核苷酸的缺失和/或添加;和/或在原生多核苷酸中一个或多个位点处一个或多个核苷酸的取代。本文所用的"原生"多核苷酸或多肽分别包含天然产生的核苷酸序列或氨基酸序列。一般而言,本发明的特定多核苷酸的变体将与该特定多核苷酸有至少约75%、80%、85%、90%、91%、92%、93%、94%、95%、96%、97%、

98％、99％或更大的序列相同性,如由本文他处所述的序列比对程序和参数所确定。

[0025] “变体”氨基酸或蛋白质是指通过下述过程衍生自原生氨基酸或蛋白质的氨基酸或蛋白质:在原生蛋白质的N-末端和/或C-末端处缺失(也称为截短)一个或多个氨基酸,在原生蛋白质的一个或多个内部位点处缺失和/或添加一个或多个氨基酸,或在原生蛋白质的一个或多个位点处取代一个或多个氨基酸。本发明包括的变体蛋白质有生物活性,即它们继续具有原生蛋白质的所需生物活性。原生多肽的生物活性变体将与由本文所述的序列比对程序和参数确定的原始序列的氨基酸序列具有至少约80％、85％、90％、91％、92％、93％、94％、95％、96％、97％、98％、99％或更大的序列相同性。本发明的蛋白质的生物活性变体与该蛋白质可相差少至1-15个氨基酸残基,少至1-10个,如6-10个,少至5个,少至4、3、2或甚至1个氨基酸残基。

[0026] 也可通过分析经测序的基因组的现有数据库来鉴定变体序列。在这种方式中,可鉴定相应序列并用于本发明的方法中。

[0027] 比对序列用于比较的方法是本领域熟知的。因此,可采用数学算法确定任意两个序列的序列相同性百分数。该数学算法的非限制性示例是Myers和Miller(1988)CABIOS 4:11-17的算法;Smith等.(1981)Adv.Appl.Math.2:482的局部比对算法;Needleman和Wunsch(1970)J.Mol.Biol.48:443-453的全局比对算法;Pearson和Lipman(1988)Proc.Natl.Acad.Sci.85:2444-2448的搜索局部比对方法;Karlin和Altschul(1990)Proc.Natl.Acad.Sci.USA 87:2264-2268的算法,由Karlin和Altschul(1993)Proc.Natl.Acad.Sci.USA 90:5873-5877改良。

[0028] 这些数学算法的计算机实施手段可用于比较序列来确定序列相同性。这类实施手段包括但不限于:PC/Gene程序中的CLUSTAL(购自美国加利福尼亚州芒廷维尤的智慧遗传公司(Intelligenetics,Mountain View,California);ALIGN程序(2.0版)和GCG Wisconsin遗传软件包中的GAP,BESTFIT、BLAST、FASTA和TFASTA,第10版(购自阿克赛勒里公司(Accelrys Inc.),美国加利福尼亚州圣地亚哥Scranton路9685号)。可用默认参数来进行使用这些程序的比对。CLUSTAL程序由以下详细描述:Higgins等,(1988)Gene 73:237-244;Higgins等,(1989)CABIOS 5:151-153;Corpet等,(1988)Nucleic Acids Res.16:10881-90;Huang等,(1992)CABIOS 8:155-65;和Pearson等,(1994)Meth.Mol.Biol.24:307-331。ALIGN程序基于Myers和Miller(1988)(同上)的算法。比较氨基酸序列时,PAM120权重残基表、缺口长度罚分12和缺口罚分4可与ALIGN程序联用。用于多重序列比对的MUSCLE算法可用于多个核酸或蛋白质序列的比较(Edgar(2004)Nucleic Acids Research 32:1792-1797)。Altschul等,(1990)J.Mol.Biol.215:403的BLAST程序基于Karlin和Altschul(1990)(同上)的算法。可利用BLASTN程序进行BLAST核苷酸搜索(评分＝100,字长＝12),以获得与编码本发明蛋白质的核苷酸序列同源的核苷酸序列。可利用BLASTX程序进行BLAST蛋白质搜索(评分＝50,字长＝3),以获得与本发明蛋白质或多肽同源的氨基酸序列。为了获得缺口比对(出于比较目的),可如Altschul等,(1997)Nucleic Acids Res.,25:33893402所述利用缺口BLAST(在BLAST 2.0中)。或者,可利用PSI-BLAST(在BLAST 2.0中)进行迭代搜索,其用来检测分子之间的远近关系。参见Altschul等,(1997)同上。利用BLAST、缺口BLAST和PSI-BLAST程序时,可使用各程序(例如针对蛋白质的BLASTX,针对核苷酸序列的BLASTN)的默认参数。参见网站www.ncbi.nlm.nih.gov。也可通过检查来人工进行

比对。

[0029] 编码Cms1多肽或其片段或变体的核酸分子可以经密码子优化,用于在感兴趣的植物或感兴趣的其它细胞或生物体中表达。"密码子优化的基因"是这样的基因,其密码子使用频率经设计以模拟宿主细胞的偏好密码子使用频率。核酸分子可以是完全或部分优化的密码子。因为任一氨基酸(除了甲硫氨酸和色氨酸)均由多种密码子编码,所述核酸分子的序列可变化但不改变编码的氨基酸。密码子优化是在核酸水平上改变一种或多种密码子时,致使氨基酸不变,但在具体的宿主生物体中的表达增加。本领域普通技术人员将知晓密码子表,并且,提供关于广泛生物体的偏好信息的其它参考文献是本领域中可得的(参见例如,Zhang等.(1991)Gene 105:61-72;Murray等.(1989)Nucl.Acids Res.17:477-508)。就植物中表达优化核苷酸序列的方法提供于例如美国专利号6,015,891和其中引用的参考文献。用于在植物中表达的密码子优化的多核苷酸的示例示于:SEQ ID NO:16-19、110-120和174-176。

[0030] II.融合蛋白

[0031] 本文提供了融合蛋白,其包括Cms1多肽或其片段或变体以及效应物结构域。通过引导RNA可以将Cms1多肽引导至靶位点,在该位点效应物结构域可以修饰或影响靶向的核酸序列。效应物结构域可以是切割结构域,表观遗传修饰结构域,转录活化结构域或转录阻遏物结构域。融合蛋白还可包含选自以下的至少一个其它结构域:核定位信号、质体信号肽、线粒体信号肽、能够运输蛋白质至多个亚细胞位置的信号肽、细胞穿透结构域或标志物结构域,这些中的任何一种都可以定位于融合蛋白的N末端、C末端或内部位置。Cms1多肽可以定位于融合蛋白的N末端,C末端,或内部位置。Cms1多肽可直接融合至效应物结构域,或可通过接头融合。在特定实施方式中,将Cms1多肽与效应物结构域融合的接头序列长度可以是至少1、2、3、4、5、6、7、8、9、10、15、20、25、30、40或50个氨基酸。例如,接头的长度可以在1-5、1-10、1-20、1-50、2-3、3-10、3-20、5-20或10-50个氨基酸之间。

[0032] 在一些实施方式中,融合蛋白的Cms1多肽可源自野生型Cms1蛋白。Cms1源性的蛋白质可以是经修饰的变体或片段。在一些实施方式中,Cms1多肽可以经修饰以含有核酸酶活性减弱或消除的核酸酶结构域(例如,RuvC或RuvC样结构域)。例如,Cms1源性多肽可以经修饰,从而使得核酸酶结构域缺失或突变,进而使其不再具有功能性(即,不存在核酸酶活性)。特别地,当为了最大同一性进行比对时,Cms1多肽可在对应于SmCms1(SEQ ID NO:10)的701或922位或对应于SulfCms1(SEQ ID NO:11)的848和1213位的位置处具有突变。可以使用已知的方法通过一个或多个缺失突变、插入突变和/或取代突变使核酸酶结构域失活,如定点诱变,PCR介导的诱变,和全基因合成,以及本领域已知的任何其它方法。在示例性的实施方式中,融合蛋白的Cms1多肽通过使RuvC样结构域突变来修饰,从而使得Cms1多肽不具有核酸酶活性。

[0033] 融合蛋白还包括效应物结构域,其定位于该融合蛋白的N末端,C末端,或内部位置。在一些实施方式中,效应物结构域是切割结构域。本文所用"切割结构域"表示切割DNA的结构域。切割结构域可获自任何内切核酸酶或外切核酸酶。可衍生出切割结构域的内切核酸酶的非限制性示例包括但不限于限制性内切核酸酶和寻靶内切核酸酶。参见例如,新英格兰生物实验室公司(New England Biolabs)产品目录或Belfort等(1997)Nucleic Acids Res.25:3379-3388。切割DNA的其它酶是已知的(例如,S1核酸酶;绿豆核酸酶;胰DNA

酶I;微球菌核酸酶;酵母HO内切核酸酶)。也参见Linn等.(编)《核酸酶》(Nucleases),冷泉港实验室出版社(Cold Spring Harbor Laboratory Press),1993。可将一种或多种这些酶(或其功能性片段)用作切割结构域的来源。

[0034] 在一些实施方式中,切割结构域可以源自II-S型内切核酸酶。II-S型内切核酸酶在通常距识别位点数个碱基对的位点切割DNA,因此具有可分离的识别和切割结构域。这些酶通常是这样的单体,其瞬时地组合在一起形成二聚体以在交错位置切割DNA的各链。合适的II-S内切核酸酶的非限制性示例包括BfiI、BpmI、BsaI、BsgI、BsmBI、BsmI、BspMI、FokI、MboⅡI和SapI。

[0035] 在某些实施方式中,II-S型切割可经修饰以促进两个不同的切割结构域的二聚化(其各自连接Cms1多肽或其片段)。在其中效应物结构域是切割结构域的实施方式中,可以如本文讨论的那样修饰Cms1多肽,从而消除其内切核酸酶活性。例如,Cms1多肽可以通过使RuvC样结构域突变来修饰,从而使得多肽不再展现出内切核酸酶活性。

[0036] 在其他实施方式中,融合蛋白的效应物结构域可以是表观遗传修饰结构域。通常,表观遗传修饰结构域在不改变DNA序列的情况下改变组蛋白结构和/或染色体结构。组蛋白和/或染色质结构的改变可以导致基因表达的改变。表观遗传修饰的示例包括但不限于,组蛋白中赖氨酸残基的乙酰化作用或甲基化作用,和DNA中胞嘧啶残基的甲基化。合适的表观遗传修饰结构域的非限制性示例包括,组蛋白乙酰基转移酶(acetyltansferase)结构域,组蛋白脱乙酰酶结构域,组蛋白甲基转移酶结构,组蛋白脱甲基酶结构,DNA甲基转移酶结构域和DNA脱甲基酶结构域。

[0037] 在效应物结构域是组蛋白乙酰基转移酶(HAT)结构域的实施方式中,HAT结构域可以源自EP300(即E1A结合蛋白p300)、CREBBP(即CREB结合蛋白)、CDY1、CDY2、CDYL1、CLOCK、ELP3、ESA1、GCN5(KAT2A)、HAT1、KAT2B、KAT5、MYST1、MYST2、MYST3、MYST4、NCOA1、NCOA2、NCOA3、NCOAT、P/CAF、Tip60、TAFII250或TF3C4。在其中效应物结构域是表观遗传修饰结构域的实施方式中,可以如本文讨论的那样修饰Cms1多肽,从而消除其内切核酸酶活性。例如,Cms1多肽可以通过使RuvC样结构域突变来修饰,从而使得多肽不再具有核酸酶活性。

[0038] 在一些实施方式中,融合蛋白的效应物结构域可以是转录活化结构域。通常,转录活化结构域与转录控制元件和/或转录调节蛋白(即,转录因子,RNA聚合酶等)相互作用以增强和/或活化一种或多种基因的转录。在一些实施方式中,转录活化结构域可以是但不限于单纯性疱疹病毒VP16活化结构域,VP64(其是VP16的四聚物衍生物),NFκB p65活化结构域,p53活化结构域1和2,CREB(cAMP响应元件结合蛋白)活化结构域,E2A活化结构域,和NFAT(活化的T细胞的核因子)活化结构域。在其他实施方式中,转录活化结构域可以是Gal4、Gcn4、MLL、Rtg3、Gln3、Oaf1、Pip2、Pdr1、Pdr3、Pho4和Leu3。转录活化域可以是野生型,也可以是原始转录活化域的修饰形式。在一些实施方式中,融合蛋白的效应物结构域是VP16或VP64转录活化结构域。在其中效应物结构域是转录活化结构域的实施方式中,可以如本文讨论的那样修饰Cms1多肽,从而消除其内切核酸酶活性。例如,Cms1多肽可以通过使RuvC样结构域突变来修饰,从而使得多肽不再具有核酸酶活性。

[0039] 在其他实施方式中,融合蛋白的效应物结构域可以是转录阻遏物结构域。通常,转录阻遏物结构域与转录控制元件和/或转录调节蛋白(即,转录因子,RNA聚合酶等)相互作用以降低和/或终止一种或多种基因的转录。合适的转录阻遏物结构域的非限制性示例包

括诱导性cAMP早期阻遏物(ICER)结构域,Kruppel-相关盒A(KRAB-A)阻遏物结构域,YY1富甘氨酸阻遏物结构域,Sp1样阻遏物,E(spl)阻遏物,I.κ.B阻遏物和MeCP2。在其中效应物结构域是转录阻遏结构域的实施方式中,可以如本文讨论的那样修饰Cms1多肽,从而消除其内切核酸酶活性。例如,Cms1多肽可以通过使RuvC样结构域突变来修饰,从而使得多肽不再具有核酸酶活性。

[0040] 在一些实施方式中,融合蛋白还包括至少一个其它结构域。合适的其它结构域的非限制性示例包括核定位信号、细胞穿透性结构域或易位结构域,和标志物结构域。

[0041] 当融合蛋白的效应物结构域是切割结构域时,可以形成包括至少一个融合蛋白的二聚体。二聚体可以是同二聚体或异二聚体。在一些实施方式中,异二聚体包含两种不同的融合蛋白。在其他实施方式中,异二聚体包括一种融合蛋白和一种其它蛋白。

[0042] 二聚体可以是同二聚体,其中两个融合蛋白单体的一级氨基酸序列是相同的。在二聚体是同二聚体的一个实施方式中,Cms1多肽可经修饰,从而消除内切核酸酶活性。在某些实施方式中,Cms1多肽经修饰,从而使得内切酶活性被消除,各融合蛋白单体可包括相同的Cms1多肽以及相同的切割结构域。切割结构域可以是任何结构域,如本文所提供的各种示例性切割结构域中的任一种。在这样的实施方式中,特定的引导RNA会将融合蛋白单体引导至不同但非常邻近的位点,从而在二聚体形成后使两个单体的核酸酶结构域在靶DNA中产生双链断裂。

[0043] 二聚体也可以是两种不同融合蛋白的异二聚体。例如,每个融合蛋白的Cms1多肽可衍生自不同的Cms1多肽或直系同源Cms1多肽。例如,各融合蛋白可包含衍生自不同来源的Cms1多肽。在这些实施方式中,各融合蛋白将识别不同的靶位点(即,由原型间隔子和/或PAM序列确定)。例如,引导RNA可以将异二聚体定位于不同但非常邻近的位点,从而使其核酸酶结构域在靶DNA中产生有效的双链断裂。

[0044] 或者,异二聚体的两个融合蛋白可以具有不同的效应物结构域。在效应物结构域是切割结构域的实施方式中,各融合蛋白可包含不同的经修饰的切割结构域。在这些实施方式中,Cms1多肽可经修饰,从而使它们的内切核酸酶活性被消除。形成异二聚体的两个融合蛋白的Cms1多肽结构域和效应物结构域可以不同。

[0045] 在上述任一所述实施方式中,同二聚体或异二聚体可以包括选自下述的至少一个其它结构域:核定位信号(NLS),质体信号肽,线粒体信号肽,能够运输蛋白质至多个亚细胞位置的信号肽,细胞穿透,易位结构域和标志物结构域(如上所述)。在上述任一所述实施方式中,可以修饰其中Cms1多肽之一或两个,从而消除或修饰多肽的内切核酸酶活性。

[0046] 异二聚体还可包含一种融合蛋白和其它蛋白质。例如,其它蛋白质可以是核酸酶。在一个实施方式中,核酸酶是锌指核酸酶。锌指核酸酶包含锌指DNA结合结构域和切割结构域。锌指识别并结合三个(3)核苷酸。锌指DNA结合结构域可包含约三个锌指至约七个锌指。锌指DNA结合结构域可以源自天然产生的蛋白质或者其可以经工程改造。参见例如,Beerli等(2002)Nat.Biotechnol.20:135-141;Pabo等(2001)Ann.Rev.Biochem.70:313-340;Isalan等(2001)Nat.Biotechnol.19:656-660;Segal等(2001)Curr.Opin.Biotechnol.12:632-637;Choo等(2000)Curr.Opin.Struct.Biol.10:411-416;Zhang等(2000)J.Biol.Chem.275(43):33850-33860;Doyon等(2008)Nat.Biotechnol.26:702-708;和Santiago等(2008)Proc.Natl.Acad.Sci.USA 105:5809-5814。锌指核酸酶的切割结构域可

以是本文所详述任何切割结构域。在一些实施方式中,锌指核酸酶可以包括选自下述的至少一个其它结构域:核定位信号(NLS),质体信号肽,线粒体信号肽,能够运输蛋白质至多个亚细胞位置的信号肽,细胞穿透或易位结构域(本文对其进行详述)。

[0047] 在某些实施方式中,以上详述的任一融合蛋白或包括至少一种融合蛋白的二聚体可以是包括至少一个引导RNA的蛋白质-RNA复合物的部分。引导RNA与融合蛋白的Cms1多肽相互作用以将融合蛋白引导至特定靶位点,其中引导RNA的5'端与特定原型间隔子序列碱基配对。

[0048] III.编码Cms1多肽或融合蛋白的核酸

[0049] 提供了编码本文所述任一Cms1多肽或融合蛋白的核酸。核酸可以是RNA或DNA。编码Cms1多肽的多核苷酸的示例示于SEQ ID NO:16-19、24-27、70-146、174-176、212-215和255-287。在一个实施方式中,编码Cms1多肽或融合蛋白的核酸是mRNA。该mRNA可以是5′-加帽和/或3′-多腺苷酸化。在另一个实施方式中,编码Cms1多肽或融合蛋白的核酸是DNA。DNA可以存在于载体中。

[0050] 编码Cms1多肽或融合蛋白的核酸可以经密码子优化,用于在感兴趣的植物细胞中高效翻译成蛋白质。本领域已知用于密码子优化的程序(例如,位于genomes.urv.es/OPTIMIZER的OPTIMIZER;OptimumGene.TM.来自GenScript,网址:www.genscript.com/codon_opt.html)。

[0051] 在某些实施方式中,编码Cms1多肽或融合蛋白的DNA可以操作性地连接至少一个启动子序列。该DNA编码序列可被操作性地连接至启动子控制序列以在感兴趣的宿主细胞中表达。在一些实施方式中,宿主细胞是植物细胞。"操作性地连接"是指2个或更多个元件之间的功能性连接。例如,启动子和感兴趣的编码区域(例如,编码Cms1多肽或引导RNA的区域)之间的操作性连接是能够表达感兴趣的编码区域的功能性连接。操作性地连接的元件可以是邻近的或非邻近的。当用于两个蛋白质编码区域之间的接合时,述及操作性地连接意在表示这些编码区域处于同一阅读框中。

[0052] 启动子序列可以是组成型,调控型,生长期特异性或组织特异性的。认识到通过在核酸分子中使用不同的启动子来调节Cms1多肽和/或引导RNA表达时间、位置和/或水平可以增强不同应用。这样的核酸分子还可以含有(如果需要)启动子调节区(例如,产生诱导型、组成型,环境或发育调节的,或细胞或组织特异性/选择性表达),转录起始起始位点,核糖体结合位点,RNA处理信号,转录终止位点,和/或多聚腺苷酸化信号。

[0053] 在一些实施方式中,本文所提供的核酸分子可与组成型、组织优先型(tissue-preferred)、发育优先型或其它启动子组合用于在植物中表达。植物细胞中组成型启动子的示例包括花椰菜花叶病病毒(CaMV)35S转录起始区域,源自根癌农杆菌(Agrobacterium tumafaciens)T-DNA的1'-或2'-启动子,泛素1启动子,Smas启动子,肉桂醇脱氢酶启动子(美国专利号5,683,439),Nos启动子,pEmu启动子,rubisco启动子,GRP1-8启动子和来自本领域技术人员已知的多种植物基因的其它转录起始区域。如果需要低水平的表达,可以使用弱启动子。弱组成型启动子包括例如Rsyn7启动子的核心启动子(WO 99/43838和美国专利号6,072,050),核心35S CaMV启动子等。其它组成型启动子包括,例如,美国专利号5,608,149;5,608,144;5,604,121;5,569,597;5,466,785;5,399,680;5,268,463和5,608,142。参见美国专利号6,177,611,其通过引用纳入本文。

[0054] 诱导型启动子的示例是可通过缺氧或冷应激诱导的Adh1启动子,可通过热应激诱导的Hsp70启动子,可通过光诱导的PPDK启动子和PEP羧化酶(pepcarboxylase)启动子。同样可用的是化学诱导的启动子,如安全剂诱导的In2-2启动子(美国专利号5,364,780),雄性激素诱导的ERE启动子,和Axig1启动子,其经植物生长素诱导并且是绒毡层特异性,但是同样在愈伤组织具有活性(PCT US01/22169)。

[0055] 植物中受发育控制的启动子的示例包括在某些组织诸如叶、根、果实、种子或花中优先启动转录的启动子。“组织特异性”启动子是仅在某些组织中起始转录的启动子。与基因的组成型表达不同,组织特异性表达是基因调控的几个相互作用水平的结果因此,同源性或密切相关的植物物质的启动子可以优先用于实现特定组织中高效和可靠的转基因表达。在一些实施方式中,表达包括组织优选启动子。“组织优先型”启动子是这样的启动子,其在某些组织中优先启动转录,但并不必需完全或仅在某些组织中启动。

[0056] 在一些实施方式中,编码Cms1多肽和/或引导RNA的核酸分子包括细胞类型特异性启动子。“细胞类型特异性”启动子是主要驱动一个或多个器官中某些细胞类型表达的启动子。细胞类型特异性启动子在植物中的功能性可以被首先活化的植物细胞的一些示例包括例如,BETL细胞,根、叶、茎细胞中的维管细胞,和干细胞。核酸分子还可以包括细胞类型优先型启动子。“细胞类型优先型”启动子是这样一种启动子,在一种或多种器官中的某些细胞类型中主要驱动表达,但并不必需完全或仅在某些细胞类型中。细胞类型优先型启动子在植物中的功能性可以被优先活化的植物细胞的一些示例包括例如,BETL细胞,根、叶、茎细胞中的维管细胞,和干细胞。本文所述核酸分子还可以包括种子优先型启动子。在一些实施方式中,种子优先型启动子在胚囊、早期胚胎、早期胚乳、糊粉和/或基底胚乳转移细胞层(BETL)中表达。

[0057] 种子优先型启动子的示例包括但不限于,27kDγ玉米蛋白启动子和糯性基因启动子(waxy promoter),Boronat,A.等(1986)Plant Sci.47:95-102;Reina,M.等Nucl.Acids Res.18(21):6426;和Kloesgen,R.B.等(1986)Mol.Gen.Genet.203:237-244。胚,果皮和胚乳中表达的启动子公开于美国专利号6,225,529和PCT公开WO 00/12733中。这些引用文献各自的公开内容通过引用其全文的方式纳入本文。

[0058] 可以驱动基因表达以植物种子优先方式在胚囊、早期胚胎、早期胚乳、糊粉和/或基底胚乳转移细胞层(BETL)中表达的启动子可以用于本文所公开的组合物和方法。这样的启动子包括但不限于这样的启动子,其天然地连接玉米(Zea mays)早期胚乳5基因,玉米早期胚乳1基因,玉米早期胚乳2基因,GRMZM2G124663,GRMZM2G006585,GRMZM2G120008,GRMZM2G157806,GRMZM2G176390,GRMZM2G472234,GRMZM2G138727,玉米CLAVATA1,玉米MRP1,水稻(Oryza sativa)PR602,水稻PR9a,玉米BET1,玉米BETL-2,玉米BETL-3,玉米BETL-4,玉米BETL-9,玉米BETL-10,玉米MEG1,玉米TCCR1,玉米ASP1,水稻ASP1,硬粒小麦(Triticum durum)PR60,硬粒小麦PR91,硬粒小麦GL7,AT3G10590,AT4G18870,AT4G21080,AT5G23650,AT3G05860,AT5G42910,AT2G26320,AT3G03260,AT5G26630,AtIPT4,AtIPT8,AtLEC2,LFAH12。其它这类启动子述于美国专利号7803990,8049000,7745697,7119251,7964770,7847160,7700836,美国专利申请公开号20100313301,20090049571,20090089897,20100281569,20100281570,20120066795,20040003427;PCT公开号WO/1999/050427,WO/2010/129999,WO/2009/094704,WO/2010/019996和WO/2010/147825,其各自通

过引用纳入其全部内容用于所用目的。本文所述启动子的功能变体或功能片段也可与本文公开的核酸操作性地连接。

[0059] 化学调节启动子通过应用外源性化学调节物可以用于调整基因的表达。取决于目标,启动子可以是应用化学物时诱导基因表达的化学诱导型启动子,或是应用化学物时抑制基因表达的化学阻遏型启动子。本领域已知化学诱导型启动子并且包括但不限于,由苯磺酰胺除草安全剂活化的玉米In2-2启动子,由用作芽前除草剂的疏水亲电子化合物活化的玉米GST启动子,以及由水杨酸活化的烟草PR-1a启动子。其它感兴趣的化学调节启动子包括类固醇响应性启动子(参见例如,Schena等.(1991)Proc.Natl.Acad.Sci.USA 88:10421-10425和McNellis等(1998)Plant J.14(2):247-257)中的糖皮质激素诱导型启动子,以及四环素诱导型和四环素阻遏型启动子(参见例如,Gatz等(1991)Mol.Gen.Genet.227:229-237以及美国专利号5,814,618和5,789,156),通过引用纳入本文。

[0060] 组织优先型启动子可以被用于靶向特定组织内表达构建体增强的表达。在某些实施方式中,组织优先型启动子可在植物组织中具有活性。组织优先型启动子是本领域已知的。参见例如,Yamamoto等,(1997)Plant J.12(2):255-265;Kawamata等,(1997)Plant Cell Physiol.38(7):792-803;Hansen等,(1997)Mol.Gen Genet.254(3):337-343;Russell等,(1997)Transgenic Res.6(2):157-168;Rinehart等,(1996)Plant Physiol.112(3):1331-1341;Van Camp等,(1996)Plant Physiol.112(2):525-535;Canevascini等,(1996)Plant Physiol.112(2):513-524;Yamamoto等,(1994)Plant Cell Physiol.35(5):773-778;Lam(1994)Results Probl.Cell Differ.20:181-196;Orozco等,(1993)Plant Mol Biol.23(6):1129-1138;Matsuoka等,(1993)Proc Natl.Acad.Sci.USA 90(20):9586-9590;和Guevara-Garcia等,(1993)Plant J.4(3):495-505。必要时,此类启动子可经修饰以用于弱表达。

[0061] 叶优先型启动子是本领域已知的。参见,例如,Yamamoto等,(1997)Plant J.12(2):255-265;Kwon等,(1994)Plant Physiol.105:357-67;Yamamoto等,(1994)Plant Cell Physiol.35(5):773-778;Gotor等,(1993)Plant J.3:509-18;Orozco等,(1993)Plant Mol.Biol.23(6):1129-1138;和Matsuoka等,(1993)Proc.Natl.Acad.Sci.USA 90(20):9586-9590。此外,也可以使用cab和rubisco启动子。参见例如,Simpson等(1958)EMBO J 4:2723-2729和Timko等(1988)Nature 318:57-58。

[0062] 根优先型启动子是已知的并且可以选自文献中可得的许多或由各种相容物种从头分离。参见例如,Hire等(1992)Plant Mol.Biol.20(2):207-218(大豆根特异性谷氨酰胺合成酶基因);Keller和Baumgartner(1991)Plant Cell3(10):1051-1061(法国豆GRP 1.8基因的根特异性控制元件);Sanger等(1990)Plant Mol.Biol.14(3):433-443(根癌农杆菌(Agrobacterium tumefaciens)甘露碱合酶(MAS)基因的根特异性启动子);和Miao等(1991)Plant Cell 3(1):11-22(编码胞质谷氨酰胺合成酶(GS)的全长cDNA克隆,其在大豆的根和根瘤中表达)。同样参见Bogusz等(1990)Plant Cell 2(7):633-641,其中描述了分离自血红蛋白基因的两种根特异性启动子,该血红蛋白基因来自固氮非豆科植物山黄麻(Parasponia andersonii)以及相关的非固氮非豆科植物山油麻(Trema tomentosa)。这些基因的启动子连接β-葡萄糖醛酸酶报告物基因,并且被引入非豆科植物烟草和豆科植物百

脉根(Lotus corniculatus),并且在两种情况中,根特异性启动子的活性被保留。Leach和Aoyagi(1991)描述了它们对毛根农杆菌(Agrobacterium rhizogenes)高表达roIC和roID根诱导型基因的启动子的分析(参见Plant Science(Limerick)79(1):69-76)。他们总结了增强子和组织优先型DNA决定簇在这些启动子中是分离的。Teeri等(1989)使用与lacZ的基因融合体显示编码章鱼碱合酶的农杆菌T-DNA基因在根尖表皮中活性特别高,TR2'基因在完整植物中具有根特异性,并因叶组织的损伤而被刺激,一种特别理想的特性组合,可与杀虫或杀幼虫基因联用(参见EMBO J.8(2):343-350)。融合至nptII(新霉素磷酸转移酶II)的TR1'基因显示相似的特征。其它根优先型启动子包括VfENOD-GRP3基因启动子(Kuster等(1995)Plant Mol.Biol.29(4):759-772);和roIB启动子(Capana等(1994)Plant Mol.Biol.25(4):681-691。同样参见美国专利号5,837,876;5,750,386;5,633,363;5,459,252;5,401,836;5,110,732和5,023,179。菜豆素基因(Murai等(1983)Science 23:476-482和Sengopta-Gopalen等(1988)PNAS 82:3320-3324。启动子序列可以是野生型的,或其可经修饰以更高效或有效地表达。

[0063] 编码Cms1多肽或融合蛋白的核酸序列可以操作性地连接由噬菌体RNA聚合酶识别的启动子序列用于体外mRNA合成。这样的实施方式中,体外转录的RNA可以经纯化用于本文所述的基因组修饰的方法中。例如,启动子序列可以是T7、T3或SP6启动子序列或T7、T3或SP6启动子序列的变化形式。在一些实施方式中,可将编码Cms1多肽或融合蛋白的序列操作性地连接至启动子序列,以在植物细胞中体外表达Cms1多肽或融合蛋白。这样的实施方式中,表达的蛋白质可以经纯化用于本文所述的基因组修饰的方法中。

[0064] 在某些实施方式中,编码Cms1多肽或融合蛋白的DNA还可以连接聚腺苷酸化信号(例如,SV40多聚A信号和在感兴趣的细胞起作用的其它信号)和/或至少一个转录终止序列。此外,编码Cms1多肽或融合蛋白的序列还可以连接这样的序列,所述序列编码本文他处所述的至少一个核定位信号,至少一个质体信号肽,至少一个线粒体信号肽,能够运输蛋白质至多个亚细胞位置的至少一个信号肽,至少一个细胞穿透结构域,和/或至少一个标志物结构域。

[0065] 编码Cms1多肽或融合蛋白的DNA可以存在于载体中。合适的载体包括质粒载体,噬菌粒,粘粒,人工/微型染色体,转座子和病毒载体(例如慢病毒载体,腺相关病毒载体等)。在一实施方式中,编码Cms1多肽或融合蛋白的DNA可以存在于质粒载体中。合适的质粒载体的非限制性实例包括pUC、pBR322、pET、pBluescript、pCAMBIA以及其变体。载体可以包括其它表达控制序列(例如,增强子序列,Kozak序列,聚腺苷酸化序列,转录终止序列等),可选择标志物序列(例如,抗生素抗性基因),复制的起点等。其它信息可以在《新编分子生物学实验指南(Current Protocols in Molecular Biology)》Ausubel等,约翰韦利森出版社(John Wiley&Sons),纽约,2003或《分子克隆:实验室手册(Molecular Cloning:A Laboratory Manual)》Sambrook和Russell,冷泉港实验室出版社(Cold Spring Harbor Press),纽约州冷泉港,第三版,2001。

[0066] 在一些实施方式中,包括编码Cms1多肽或融合蛋白的序列的表达载体可以还包括编码引导RNA的序列。编码引导RNA的序列可以操作性地连接至少一个转录控制序列,用于在植物中或感兴趣的植物细胞中表达引导RNA。例如,编码引导RNA的DNA可以操作性地连接由RNA聚合酶III(Pol III)识别的启动子序列。合适的Pol III启动子的实例包括但不限

于,哺乳动物U6,U3,H1,和7SL RNA启动子和水稻U6和U3启动子。

[0067] IV.修饰基因组中核苷酸序列的方法

[0068] 本文提供了用于修饰基因组的核苷酸序列的方法。基因组的非限制性示例包括细胞,核,细胞器,质粒和病毒基因组。所述方法包括将一种或多种靶向DNA的多核苷酸引入基因组宿主(例如,细胞或细胞器),所述靶向DNA的多核苷酸例如靶向DNA的RNA(“引导RNA”,“gRNA”,“CRISPR RNA”或“crRNA”)或编码靶向DNA的RNA的DNA多核苷酸,其中,所述靶向DNA的多核苷酸包含:(a)第一区段,其包含与靶DNA中的序列互补的核苷酸序列;和(b)第二区段,其与Cms1多肽相互作用并且还将Cms1多肽或编码Cms1多肽的多核苷酸引入基因组宿主,其中Cms1多肽包含:(a)多核苷酸结合部分,其与gRNA或其它靶向DNA的多核苷酸相互作用;和(b)活性部分,其显示定点酶促活性。然后,可在表达Cms1多肽并切割被gRNA靶向的核苷酸序列的条件下培养基因组宿主。需指出的是,本文所述系统不需要添加外源性Mg^{2+}或任何其他离子。最后,可以选择包含修饰的核苷酸序列的基因组宿主。

[0069] 本文公开的方法包括将至少一种Cms1多肽或编码至少一种Cms1多肽的核酸引入基因组宿主,如本文所述。在一些实施方式中,Cms1多肽可以分离的蛋白质形式引入基因组宿主。在这样的实施方式中,Cms1多肽可以还包括至少一个细胞穿透结构域,其促进蛋白质的细胞摄取。在一些实施方式中,Cms1多肽可以与引导多核苷酸复合的核蛋白形式(例如,以与引导RNA复合的核糖核蛋白形式)引入基因组宿主。在其它实施方式中,Cms1多肽可以编码Cms1多肽的mRNA分子形式引入基因组宿主。在其它实施方式中,Cms1多肽可以DNA分子形式引入基因组宿主中,该DNA分子包含编码Cms1多肽的开放阅读框。编码本文所述的Cms1多肽或融合蛋白的DNA序列一般操作性地连接至将在基因组宿主中起作用的启动子序列。DNA序列可以是线性的,或DNA序列可以是载体的一部分。在其它实施方式中,Cms1多肽或融合蛋白可以包含引导RNA或融合蛋白和引导RNA的RNA-蛋白质复合物形式引入基因组宿主。

[0070] 在某些实施方式中,编码Cms1多肽的mRNA可以靶向细胞器(例如,质体或线粒体)。在某些实施方式中,编码一种或多种引导RNA的mRNA可以靶向细胞器(例如,质体或线粒体)。在某些实施方式中,编码Cms1多肽和一种或多种引导RNA的mRNA可以靶向细胞器(例如,质体或线粒体)。靶向mRNA至细胞器的方法为本领域已知(参见例如,美国专利申请号2011/0296551;美国专利申请号2011/0321187;Gómez和Pallás(2010)PLoS One 5:e12269),并且通过引用纳入本文。

[0071] 在某些实施方式中,编码Cms1多肽的DNA可以还包括编码引导RNA的序列。通常,将编码Cms1多肽和引导RNA的各序列操作性地连接至一个或多个合适的启动子控制序列,所述启动子控制序列允许Cms1多肽和引导RNA在基因组宿主中分别表达。编码Cms1多肽和引导RNA的DNA序列进一步包括其它表达对照、调控、和/或处理序列。编码Cms1多肽和引导RNA的DNA序列可以是线性的或是载体的部分。

[0072] 本文所述的方法还可包括将至少一种引导RNA或编码至少一种多核苷酸(例如引导RNA)的DNA引入基因组宿主。引导RNA与Cms1多肽相互作用,以将Cms1多肽引导至特定的靶位点,在该位点,引导RNA碱基与靶位点中的特定DNA序列配对。引导RNA可以包括三个区域:与靶DNA序列中靶位点互补的第一区域,形成茎环结构的第二区域,和基本保持单链的第三区域。各引导RNA的第一区域是不同的,因此各引导RNA将Cms1多肽导向特定靶位点。各引导RNA的第二和第三区域在所有引导RNA中可以相同。

[0073] 引导RNA的一个区域与靶DNA中靶位点的序列(即原型间隔子序列)互补,从而引导RNA的第一区域可与靶位点碱基配对。在各种实施方式中,引导RNA的第一区域可以包括约8个核苷酸至超过约30个核苷酸。例如,引导RNA的第一区域与核苷酸序列中靶位点之间碱基配对区域的长度可以是约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约22、约23、约24、约25、约27、约30或超过30个核苷酸。在一个示例性实施方式中,引导RNA的第一区域的长度是约23、24或25个核苷酸。引导RNA还可以包括形成二级结构的第二区域。在一些实施方式中,二级结构包括茎或发夹。茎的长度可变。例如,茎的长度可以是约5至约6,约10,至约15,约20至约25个碱基对。茎可以包括1至约10个核苷酸的一个或多个凸起(bulge)。在一些优选的实施方式中,发夹结构包含序列UCUACN$_{3-5}$GUAGAU(SEQ ID NO:312-314,由SEQ ID NO:315-317编码),其用“UCUAC”和“GUAGA”碱基配对以形成茎。“N$_{3-5}$”表示3、4或5个核苷酸。因此,第二区域的总长度可以在约14至约25个核苷酸的范围内。在某些实施方式中,环的长度为约3、4或5个核苷酸,而茎包含约5、6、7、8、9或10个碱基对。

[0074] 引导RNA还可以包括基本上保持单链的第三区域。因此,第三区域与感兴趣的细胞中的任何核苷酸序列都不互补,并且与其余引导RNA没有互补性。第三区域的长度可变。第三区域的长度通常大于约4个核苷酸。例如,第三区域的长度可以在约5至约60个核苷酸。引导RNA的第二和第三区域(也称为通用或支架区域)的合并长度可以在约30至约120个核苷酸的范围内。在一方面,引导RNA的第二和第三区域组合的长度可以在约40至约45个核苷酸。

[0075] 在一些实施方式中,引导RNA包括含有所有三个区域的单个分子。在其他实施方式中,引导RNA可以包括两个不同的分子。第一RNA分子可以包括引导RNA的第一区域以及引导RNA第二区域“茎”的一半。第二RNA分子可以包括引导RNA第二区域“茎”的另一半以及引导RNA的第三区域。因此,在该实施方式中,第一和第二RNA分子各自含有彼此之间相互互补的核苷酸序列。例如,在一实施方式中,第一和第二RNA分子各自包括与其它序列碱基配对的序列(约6至约25个核苷酸)以形成功能性引导RNA。在具体实施方式中,引导RNA是单个分子(即crRNA),其在不需要第二引导RNA(即tracrRNA)的情况下与染色体中的靶位点和Cms1多肽相互作用。

[0076] 在某些实施方式中,引导RNA可以RNA分子形式引入基因组宿主。RNA分子可以体外转录。或者,RNA分子可以化学合成。在其它实施方式中,引导RNA可以DNA分子形式引入基因组宿主。在这种情况下,可将编码引导RNA的DNA操作性地连接至一个或多个启动子序列,以在基因组宿主中表达引导RNA。例如,RNA编码序列可以与RNA聚合酶III(Pol III)识别的启动子序列可操作地连接。

[0077] 编码引导RNA的DNA分子可以是线性或环状的。在一些实施方式中,编码引导RNA的DNA序列可以是载体的部分。合适的载体包括质粒载体,噬菌粒,粘粒,人工/微型染色体,转座子和病毒载体。在一个示例性的实施方式中,编码引导RNA的DNA存在于质粒载体中。合适的质粒载体的非限制性实例包括pUC、pBR322、pET、pBluescript、pCAMBIA以及其变体。载体可以包括其它表达控制序列(例如,增强子序列,Kozak序列,聚腺苷酸化序列,转录终止序列等),可选择标志物序列(例如,抗生素抗性基因),复制的起点等。

[0078] 在Cms1多肽和引导RNA两者以DNA分子形式被引入基因组宿主的实施方式中,其各

自可以是分开的分子的部分(例如,一个载体含有Cms1多肽或融合蛋白编码序列,第二载体含有引导RNA编码序列),或者其可以是同一分子的部分(例如,一个载体含有Cms1多肽或融合蛋白和引导RNA两者的编码(和调节)序列)。

[0079] 与引导RNA联合的Cms1多肽被引导至基因组宿主中的靶位点,其中所述Cms1多肽在靶DNA中引入双链断裂。靶位点没有序列限制,除了该序列紧接共有序列之前(上游)之外。该共有序列也称为原型间隔子邻近基序(protospacer adjacent motif)。PAM序列的示例包括但不限于TTTN,NTTN,TTTV和NTTV(其中N被定义为任何核苷酸,而V被定义为A,G或C)。本领域中众所周知,合适的PAM序列必须位于相对于靶DNA序列的正确位置,以允许Cms1核酸酶产生所需的双链断裂。对于迄今已表征的所有Cms1核酸酶,PAM序列都位于靶DNA序列的5'附近。目前无法通过计算预测给定Cms1核酸酶的PAM位点要求,而必须使用本领域可用的方法通过实验确定(Zetsche等.(2015)Cell 163:759-771;Marshall等.(2018)Mol Cell 69:146-157)。本领域已知对给定核酸酶具有特异性的PAM序列受到酶浓度的影响(Karvelis等(2015)Genome Biol 16:253)。因此,调节递送至感兴趣的细胞或体外系统的Cms1蛋白的浓度体现了改变与该Cms1酶相关的一个或多个PAM位点的一种方式。例如通过改变用于表达Cms1编码基因的启动子,通过改变递送至细胞或体外系统的核糖核蛋白浓度,或通过添加或去除在调节基因表达水平中可能起作用的内含子,可以实现调整感兴趣的系统中的Cms1蛋白浓度。如本文所详述,引导RNA的第一区域与靶序列的原型间隔子互补。通常,引导RNA的第一区域的长度是19-21个核苷酸。

[0080] 靶位点可以在基因的编码区域中,基因的内含子中,基因的控制区域中,基因间的非编码区域等。基因可以是蛋白质编码基因或RNA编码基因。该基因可以是本文所述的任何感兴趣的基因。

[0081] 在一些实施方式中,本文公开的方法还包括将至少一种供体多核苷酸引入基因组宿主。供体多核苷酸包括至少一种供体序列。在一些方面,供体多核苷酸的供体序列对应于靶DNA中存在的内源或天然序列。例如,供体序列可以与靶位点处或附近的DNA序列的部分基本相同,但是包含至少一个核苷酸变化。因此,供体序列可在靶位点处包含野生型序列的修饰形式,从而在与原生序列整合或交换后,靶位置处的序列包含至少一个核苷酸变化。例如,改变可以是一个或多个核苷酸的插入,一个或多个核苷酸的缺失,一个或多个核苷酸的取代或其组合。由于经修饰序列的整合,基因组宿主可从靶染色体序列产生经修饰的基因产物。

[0082] 供体多核苷酸的供体序列可替代地对应于外源序列。如本文所用,"外源"序列是指不原生于基因组宿主的序列,或者其在基因组宿主中的原生位置处于不同位置的序列。例如,外源性序列可以包括蛋白质编码序列,其可以操作性地连接外源性启动子控制序列,因此在整合到基因组后,基因组宿主能够表达该整合序列所编码的蛋白质。例如,供体序列可以是任何感兴趣的基因,例如编码如本文他处所述的农艺学上重要的性状的那些。或者,可将外源序列整合进入靶DNA序列,从而使其表达受内源性启动子控制序列调节。在其他的重复形式中,外源性序列可以是转录控制序列,其它的表达控制序列或RNA编码序列。将外源性序列整合到靶DNA序列被称为"敲入"。供体序列可以具有各种长度,从几个核苷酸到数百个核苷酸到数千个核苷酸。

[0083] 在一些实施方式中,供体多核苷酸中的供体序列侧接上游序列和下游序列,其与

分别位于靶位点上游和下游的序列具有实质上的序列同一性。因为这些序列相似性,供体多核苷酸的上游和下游序列允许供体多核苷酸和靶向的序列之间的同源重组,从而使得供体序列被整合到靶DNA序列(或与之交换)。

[0084] 本文所用上游序列指这样的核酸序列,其与靶位点上游的DNA序列具有实质上的序列同一性。类似地,下游序列指与靶位点下游的DNA序列具有实质上的序列同一性的核酸序列。本文所用短语"实质上的序列同一性"指序列具有至少约75%的序列同一性。因此,供体多核苷酸中的上游和下游序列与靶向的位点上游或下游序列可以具有约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%的序列同一性。在示例性的实施方式中,供体多核苷酸中的上游和下游序列与靶向的位点上游或下游的核苷酸序列可以具有约95%或100%的序列同一性。在一实施方式中,上游序列与位于靶向的位点上游紧邻的(即邻近靶向的位点)核苷酸序列具有实质上的序列同一性。在其它实施方式中,上游序列与位于靶向的位点上游约一百个(100)核苷酸内的核苷酸序列具有实质上的序列同一性。因此例如,上游序列与位于靶向的位点上游约1-约20,约21-约40,约41-约60,约61-约80,或约81-约100核苷酸内的核苷酸序列具有实质上的序列同一性。在一实施方式中,下游序列与位于靶向的位点下游紧邻的(即邻近靶向的位点)核苷酸序列具有实质上的序列同一性。在其它实施方式中,下游序列与位于靶向的位点下游约一百个(100)核苷酸内的核苷酸序列具有实质上的序列同一性。因此例如,下游序列与位于靶向的位点下游约1-约20,约21-约40,约41-约60,约61-约80,或约81-约100核苷酸内的核苷酸序列具有实质上的序列同一性。

[0085] 各上游或下游序列的长度可以在约20个核苷酸至约5000个核苷酸。在一些实施方式中,上游和下游序列可包含约50、100、200、300、400、500、600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900、2000、2100、2200、2300、2400、2500、2600、2800、3000、3200、3400、3600、3800、4000、4200、4400、4600、4800或5000个核苷酸。在示例性的实施方式中,上游或下游序列的长度可以在约50个核苷酸至约1500个核苷酸。

[0086] 包含与靶核苷酸序列具有序列相似性的上游和下游序列的供体多核苷酸可以是线性或环状的。在供体多核苷酸是环状的实施方式中,其可以是载体的一部分。例如,载体可以是质粒载体。

[0087] 在某些实施方式中,供体多核苷酸还可以包括由Cms1多肽识别的至少一个靶向的切割位点。可将添加到供体多核苷酸中的靶向切割位点置于供体序列的上游或下游或上游和下游。例如,供体序列可以由靶向的切割位点侧接,因此在通过Cms1多肽切割后,供体序列由突出端侧接,所述突出端与通过Cms1多肽切割后生成的核苷酸序列中的那些相容。因此,可以用切割的核苷酸序列在通过非同源性修复过程修复双链断裂期间连接供体序列。通常,包括靶向的切割位点的供体多核苷酸是环状的(例如,可以是质粒载体的部分)。

[0088] 供体多核苷酸可以是包括具有任选的短突出端的短供体序列的线性分子,所述任选的短突出端与Cms1多肽生成的突出端相容。在这样的实施方式中,供体序列可在双链断裂的修复过程中与切割的染色体序列直接连接。在一些情况中,供体序列可以少于约1,000,少于约500,少于约250,或少于约100个核苷酸。在某些情况下,供体多核苷酸可以是包含具有钝末端的短供体序列的线性分子。在其它重复情况中,供体多核苷酸可以是线性分子,其包含具有5'和/或3'突出端的短供体序列。该突出端可以包括1、2、3、4或5个核苷酸。

[0089] 在一些实施方式中,供体多核苷酸将是DNA。DNA可以是单链或双链和/或线性或环状。供体多核苷酸可以是DNA质粒、细菌人工染色体(BAC)、酵母人工染色体(YAC)、病毒载体、DNA的线性部分、PCR片段、裸核酸或与递送载剂如脂质体或泊咯沙姆复合的核酸。在具体实施方式中,包括供体序列的供体多核苷酸可以是质粒载体的部分。在任何这些情况下,包含供体序列的供体多核苷酸还可包含至少一个其它序列。

[0090] 在一些实施方式中,该方法可包括将一种Cms1多肽(或编码核酸)和一种引导RNA(或编码DNA)引入基因组宿主,其中所述Cms1多肽在靶DNA中引入一个双链断裂。在不存在任选供体多核苷酸实施方式中,核苷酸序列中的双链锻炼可以通过非同源末端连接(NHEJ)修复过程进行修复。因为NHEJ是易错的,缺失至少一个核苷酸、插入至少一个核苷酸、取代至少一个核苷酸或其组合可能会出现在修复断裂期间。因此,靶向的核苷酸序列可以经修饰或被失活。例如,单核苷酸改变(SNP)可以产生改变的蛋白质产物,或者编码序列阅读框的移动可以灭活或"敲除"序列,从而不再产生蛋白质产物。在存在任选供体多核苷酸的实施方式中,供体多核苷酸中的供体序列在修复双链断裂期间可与靶位点的核苷酸序列交换或整合至其中。例如,在供体序列侧接这样上游和下游序列的实施方式中,所述上游和下游序列具有分别与核苷酸序列中靶位点的上游和下游序列实质上的序列同一性,供体序列在通过同源性导向的修复过程介导的修复期间可以与靶位点的核苷酸序列交换或整合至其中。或者,在供体序列侧接相容突出端(或者该相容突出端由Cms1多肽原位生成)的实施方式中,供体序列在双链断裂修复期间通过非同源性修复过程可以直接连接切割的核苷酸序列。将供体序列交换或整合至核苷酸序列修饰靶核苷酸序列,或者将外源性序列引入靶核苷酸序列。

[0091] 本文公开的方法还可包括,将一种或多种Cms1多肽(或编码核酸)和两个引导多核苷酸(或编码DNA)引入基因组宿主,其中Cms1多肽在靶核苷酸序列中引入两个双链断裂。这两个断裂可以在几个碱基对之内,在几十个碱基对之内,或者可以相隔成千上万个碱基对。在不存在任选供体多核苷酸的实施方式中,得到的双链断裂可以通过非同源性修复过程修复,这样的话两个切割位点之间的序列丢失和/或在修复断裂期间可能会出现缺失至少一个核苷酸,插入至少一个核苷酸,取代至少一个核苷酸或其组合。在存在任选的供体多核苷酸的实施方式中,在通过基于同源性的修复过程(例如,在供体序列侧接这样上游和下游序列的实施方式中,所述上游和下游序列具有分别与核苷酸序列中靶位点的上游和下游序列实质上的序列同一性中)或非同源性的修复过程(例如,在供体序列侧接相容突出端的实施方式中)的双链断裂修复期间,供体多核苷酸中的供体序列可以与靶核苷酸序列交换或整合至靶核苷酸序列中。

[0092] A.修饰植物基因组中的核苷酸序列的方法

[0093] 植物细胞具有核,质体和线粒体基因组。本发明的组合物和方法可以用于修饰核、质体和/或线粒体基因组的序列,或者可以用于调节由核、质体和/或线粒体基因组编码的一种或多种基因的表达。因此,"染色体"或"染色体的"指核、质体或线粒体基因组DNA。当"基因组"适用于植物细胞时,其不但包括存在于细胞核的染色体DNA,也包括存在于细胞亚细胞组分(例如,线粒体或质体)中的细胞器DNA。可以使用本文所述的方法修饰植物细胞,细胞器或胚胎中的任何感兴趣的核苷酸序列。在具体实施方式中,本文所公开的方法被用于修饰编码农艺学重要性状的核苷酸序列,如植物激素,植物防御蛋白,营养转运蛋白,生

物结合蛋白,所需输入性状,所需输出性状,应激抗性基因,疾病/病原体抗性基因,雄性不育,发育基因,调节基因,参与光合作用的基因,DNA修复基因,转录调节基因或任何其他感兴趣的多核苷酸和/或多肽。也可以修饰农艺学重要性状如油脂、淀粉和蛋白质含量。修饰包括增加油酸、饱和和不饱和油脂的含量,增加赖氨酸和硫的水平,提供必需氨基酸,以及淀粉的改性。硫堇蛋白(hordothionin)蛋白质修饰描述于美国专利号5,703,049、5,885,801、5,885,802和5,990,389中,其通过引用纳入本文。另一实例是富赖氨酸和/或硫种子蛋白,其由美国专利号5,850,016中所述大豆2S白蛋白所编码,以及来自大麦的糜蛋白酶阻遏物,述于Williamson等(1987)Eur.J.Biochem.165:99-106,其公开通过引用纳入本文。

[0094] Cms1多肽(或编码核酸)、引导RNA(或编码DNA)和任选的供体多核苷酸可以通过包括转化的各种方法引入植物细胞、细胞器或植物胚胎。转化方案以及向植物中引入多肽或多核苷酸序列的方案可根据转化靶向的植物或植物细胞的类型(即,单子叶或双子叶)而变化。向植物细胞中引入多肽和多核苷酸的合适方法包括微注射(Crossway等,(1986)Biotechniques 4:320-334)、电穿孔(Riggs等,(1986)Proc.Natl.Acad.Sci.USA 83:5602-5606),农杆菌-介导的转化(美国专利号5,563,055和美国专利号5,981,840),直接基因转化(Paszkowski等,(1984)EMBO J.3:2717-2722),和弹道颗粒加速(参见例如,美国专利号4,945,050;美国专利号5,879,918;美国专利号5,886,244;和5,932,782;Tomes等,(1995)《植物细胞、组织和器官培养中的基础方法》(Plant Cell,Tissue,and Organ Culture:Fundamental Methods),Gamborg和Phillips编(Springer-Verlag,Berlin);McCabe等,(1988)Biotechnology 6:923-926);和Lec1转化(WO00/28058)。还参见Weissinger等,(1988)Ann.Rev.Genet.22:421-477;Sanford等,(1987)Particulate Science and Technology 5:27-37(洋葱);Christou等,(1988)Plant Physiol.87:671-674(大豆);McCabe等,(1988)Bio/Technology 6:923-926(大豆);Finer和McMullen(1991)In Vitro Cell Dev.Biol.27P:175-182(大豆);Singh等,(1998)Theor.Appl.Genet.96:319-324(大豆);Datta等,(1990)Biotechnology 8:736(水稻);Klein等,(1988)Proc.Natl.Acad.Sci.USA85:4305-4309(玉米);Klein等,(1988)Biotechnology 6:559-563(玉米);美国专利号5,240,855;5,322,783;和5,324,646;Klein等,(1988)Plant Physiol.91:440-444(玉米);Fromm等,(1990)Biotechnology 8:833-839(玉米);Hooykaas-Van Slogteren等,(1984)Nature(伦敦)311:763-764;美国专利号5,736,369(谷类);Bytebier等,(1987)Proc.Natl.Acad.Sci.USA 84:5345-5349(百合);De Wet等,(1985)《胚珠组织实验操作》(The Experimental Manipulation of Ovule Tissues),Chapman等编,(纽约朗文出版社(Longman,New York),第197-209页(花粉);Kaeppler等,(1990)Plant Cell Reports 9:415-418和Kaeppler等,(1992)Theor.Appl.Genet.84:560-566(须-介导的转化);D'Halluin等,(1992)Plant Cell 4:1495-1505(电穿孔);Li等,(1993)Plant Cell Reports 12:250-255以及Christou和Ford(1995)Annals of Botany 75:407-413(水稻);Osjoda等,(1996)Nature Biotechnology 14:745-750(玉米,通过根癌农杆菌);其全部通过引用纳入本文。已经证明了通过生物弹射引入包括核酸酶以及合适的引导RNA的核糖核蛋白进行的对植物细胞的位点特异性基因组编辑(Svitashev等(2016)Nat Commun 7:13274);这些方法通过引用纳入本文。"稳定转化"是指引入植物的核苷酸构建体整合到植物的基因组中并且能够被其后代遗传。核苷酸构建体可以整合进入植物的

核,质体或线粒体基因组中。用于质体转化的方法为本领域已知(参见例如,《叶绿体生物技术:方法和方案(Chloroplast Biotechnology:Methods and Protocols)》(2014)Pal Maliga编著和美国专利申请号2011/0321187),并且本领域已经描述了用于植物线粒体转化的方法(参见例如美国专利申请号2011/0296551),通过引用纳入本文。

[0095] 按照常规方式,已经转化的细胞可长成植物(即培养)。参见,例如,McCormick等,(1986)Plant Cell Reports 5:81-84。由此,本发明提供了具有稳定整合到其基因组中核酸修饰的转化的种子(也称为"转基因种子")。

[0096] "引入"在将核酸片段(例如重组DNA构建体)插入细胞的上下文中表示"转染"或"转化"或"转导"并且包括将核酸片段纳入植物细胞,其中核酸片段可以被纳入细胞的基因组中(例如,核染色体、质粒、质体染色体或线粒体染色体),转化成独立复制的复制子,或瞬时表达(例如,转染的mRNA)。

[0097] 本发明可用于任何植物物种的转化,包括但不限于单子叶和双子叶(即单子叶植物和双子叶植物)。感兴趣植物物种的示例包括但不限于:玉米(Zea mays)、油菜种(例如甘蓝型油菜(B.napus)、白菜型油菜(B.rapa)、芥菜型油菜(B.juncea))、尤其是用作菜籽油来源的那些油菜物种、苜蓿(Medicago sativa)、水稻(Oryza sativa)、黑麦(Secale cereale)、高粱(Sorghum bicolor,Sorghum vulgare)、荠蓝(Camelina sativa)、粟(例如珍珠粟(Pennisetum glaucum)、黍(Panicum miliaceum)、小米(Setaria italica)、穆子(Eleusine coracana))、向日葵(Helianthus annuus)、藜(Chenopodium quinoa)、菊苣(Cichorium intybus)、莴苣(Lactuca sativa),红花(Carthamus tinctorius)、小麦(Triticum aestivum)、大豆(Glycine max)、烟草(Nicotiana tabacum)、马铃薯(Solanum tuberosum)、花生(Arachis hypogaea)、棉花(Gossypium barbadense,Gossypium hirsutum)、甘薯(Ipomoea batatas)、木薯(Manihot esculenta)、咖啡(Coffea spp.)、椰子(Cocos nucifera)、菠萝(Ananas comosus)、柠檬树(Citrus spp.)、可可(Theobroma cacao)、茶(Camellia sinensis)、香蕉(Musa spp.)、鳄梨(Persea americana)、无花果(Ficus casica)、番石榴(Psidium guajava)、芒果(Mangifera indica)、橄榄(Olea europaea)、番木瓜(Carica papaya)、腰果(Anacardium occidentale)、澳洲坚果(Macadamia integrifolia)、杏(Prunus amygdalus)、甜菜(Beta vulgaris)、甘蔗(Saccharum spp.)、油棕榈(Elaeis guineensis)、白杨(杨树(Populus spp.))、桉树(Eucalyptus spp.)、燕麦(Avena sativa)、大麦(Hordeum vulgare)、蔬菜、观赏植物和针叶树。

[0098] Cms1多肽(或编码核酸)、引导RNA(或编码引导RNA的DNA)和任选的供体多核苷酸可以同时或依次引入植物细胞、细胞器或植物胚胎。Cms1多肽(或编码核酸)与引导RNA(或编码DNA)的比例通常约为化学计量的,从而这两个组分可与靶DNA形成RNA-蛋白质复合物。在一个实施方式中,编码Cms1多肽的DNA以及编码引导RNA的DNA在质粒载体中一起递送。

[0099] 本发明的组合物和方法可以用于改变植物中感兴趣基因的表达,如参与光合作用的基因的表达。因此,可与对照植物相比调节编码光合作用中涉及的蛋白质的基因的表达。"对象植物或植物细胞"是其中已经实现感兴趣基因的遗传改变如突变,或者是源自如此改变的植物或细胞并包含改变的植物或植物细胞。"对照"或"对照植物"或"对照植物细胞"提供了测量对象植物或植物细胞的表型变化的参照点。因此,根据本发明的方法,表达水平高

于或低于对照植物中的表达水平。

[0100] 一种对照植物或植物细胞可包含,例如:(a)野生型植物或细胞,即具有与用于产生对象植物或细胞的遗传改变的起始材料相同的基因型;(b)与起始材料有相同基因型但已经用无效构建体(即,用对感兴趣性状没有已知影响的构建体,如包含标记基因的构建体)转化的植物或植物细胞;(c)植物或植物细胞,其是对象植物或植物细胞的后代中的非转化分离体;(d)与对象植物或植物细胞在遗传上相同但没有接触会诱导感兴趣基因表达的条件或刺激的植物或植物细胞;或(e)在不表达感兴趣基因的条件下的对象植物或植物细胞本身。

[0101] 虽然本发明以转化的植物描述,应认识到本发明的转化的生物体可包括植物细胞、植物原生质体、可再生出植物的植物组织培养物、植物愈伤组织、植物块和在植物或植物部分中完整的植物细胞如胚胎、花粉、胚珠、种子、叶、花、枝条、果实、仁、穗、穗轴、外壳、柄、根、根尖、花粉囊等。种粒是指由商业种植者出于生长或繁殖物种以外的目的产生的成熟种子。再生植物的后代、变体和突变体也包括在本发明的范围内,只要这些部分包含引入的多核苷酸。

[0102] 可以使用本文所公开的方法制备编码序列的衍生物,从而在编码的多肽中增加预选氨基酸的水平。例如,编码大麦高赖氨酸多肽(BHL)的基因源自1996年11月1日提交的美国专利申请序列号08/740,682和WO 98/20133的大麦糜蛋白酶阻遏物,其公开通过引用纳入本文。其它蛋白质包括富蛋氨酸植物蛋白,如来自向日葵籽(Lilley等(1989)关于人类食品和动物饲料中植物蛋白利用的世界大会报告(Proceedings of the World Congress on Vegetable Protein Utilization in Human Foods and Animal Feedstuffs),Applewhite编著(伊利诺伊州香槟市美国油脂化学会(American Oil Chemists Society)),第497-502页;通过引用纳入本文);玉米(Pedersen等(1986)J.Biol.Chem.261:6279;Kirihara等(1988)Gene71:359;两者通过引用纳入本文);和水稻(Musumura等(1989)Plant Mol.Biol.12:123通过引用纳入本文)。其它农艺学重要的基因编码乳胶、Floury 2、生长因子、种子储存因子和转录因子。

[0103] 本文所公开的方法可以用于修饰除草剂抗性特性,包括编码除草剂抗性的基因,其能够抑制乙酰乳酸合酶(ALS)的作用,尤其是磺酰脲类除草剂(例如,含有导致这类抗性的突变的乙酰乳酸合酶(ALS)基因,尤其是S4和/或Hra突变),编码除草剂抗性的基因,其能够抑制谷氨酰胺合成酶的作用,如草丁膦或巴斯达(basta)(例如,bar基因);草甘膦(例如,EPSPS基因和GAT基因;参见例如美国公开号20040082770和WO 03/092360);其它为本领域已知的这类基因。bar基因编码对除草剂Basta的抗性,nptII基因编码对卡那霉素和遗传霉素的抗性,而ALS基因突变体编码对除草剂氯磺隆的抗性。例如,美国专利申请2016/0208243中描述了其它除草剂抗性性状,其通过引用纳入本文。

[0104] 还可以修饰不育基因,并为物理去雄提供替代方法。以这样方式使用的基因的实例包括雄性组织优选基因以及具有雄性不育表型的基因如QM,述于美国专利号5,583,210中。其它基因包括激酶和编码对雄或雌配子体发育有毒的化合物的那些。其它不育性状述于例如美国专利申请2016/0208243中,其通过引用纳入本文。

[0105] 谷物的质量可以通过修饰编码性状的基因来改变,如油脂的类型和水平,饱和和未饱和,必需氨基酸的数量和质量,以及纤维素的水平。在玉米中,经修饰的大麦硫堇蛋白

述于美国专利号5,703,049、5,885,801、5,885,802和5,990,389。

[0106] 商业性状也可以通过修饰基因来改变,或者其将可以例如增加用于乙醇生产的淀粉,或提供蛋白质的表达。经修饰植物的另一重要的商业用途是聚合物和生物塑料的生产,见述于例如美国专利号5,602,321。基因(例如β-酮硫醇酶,聚羟基牛酸酯合成酶(PHBase)和乙酰乙酰基-CoA还原酶)能促进聚羟基烷酸酯(PHA)的表达(参见Schubert等.(1988)J.Bacteriol.170:5837-5847)。

[0107] 外源性产物包括植物酶和产物,以及来自包括原核生物或其它真核生物的那些。这样的产物包括酶,辅因子,激素等。可以增加蛋白质的水平,特别是具有改善的氨基酸分布的经修饰的蛋白质以改善植物的营养价值。这通过表达具有增强的氨基酸含量的蛋白质来实现。

[0108] 本文所公开的方法还可以用于插入异源性基因和/或修饰天然植物基因表达以实现所需的植物性状。这些性状包括例如抗病性,除草剂耐受性,抗旱性,耐盐性,昆虫抗性,对寄生杂草的抗性,改善的植物营养价值,改善的草料消化率,增加的谷物产量,胞质雄性不育,改变的果实成熟度,增加的植物或植物部分的储存寿命,减少的变应原产生,和,增加或减少的木质素含量。美国专利申请2016/0208243中公开了能够赋予这些所需性状的基因,其通过引用纳入本文。

[0109] B.修饰非植物真核基因组中的核苷酸序列的方法

[0110] 本文提供了用于修饰非植物真核细胞或非植物真核细胞器的核苷酸序列的方法。在一些实施方式中,非植物真核细胞是哺乳动物细胞。在具体实施方式中,非植物真核细胞是非人哺乳动物细胞。该方法包括向靶细胞或细胞器引入靶向DNA的RNA或编码靶向DNA的RNA的DNA多核苷酸,其中靶向DNA的RNA包括:(a)第一区段,其包含与靶DNA中序列互补的核苷酸序列;和(b)第二区段,其与Cms1多肽相互作用;和,向靶细胞或细胞器引入Cms1多肽或编码Cms1多肽的多核苷酸,其中Cms1多肽包括:(a)结合RNA的部分,其与靶向DNA的RNA相互作用;和(b)活性部分,其显示定点酶促活性。然后可以在嵌合的核酸酶多肽表达并且切割核苷酸序列的条件下培养靶细胞或细胞器。需指出的是,本文所述系统不需要添加外源性Mg^{2+}或任何其他离子。最后,可选择包含经修饰的核苷酸序列的非植物真核细胞或细胞器。

[0111] 在一些实施方式中,该方法可以包括向非植物真核细胞或细胞器中引入一个Cms1多肽(或编码核酸)和一个引导RNA(或编码DNA),其中Cms1多肽在核或细胞器染色体DNA的靶核苷酸序列中引入一个双链断裂。在一些实施方式中,该方法可以包括向非植物真核细胞或细胞器中引入一个Cms1多肽(或编码核酸)和至少一个引导RNA(或编码DNA),其中Cms1多肽在核或细胞器染色体DNA的靶核苷酸序列中引入超过一个(即2、3或超过3个双链断裂)双链断裂。在不存在任选供体多核苷酸实施方式中,核苷酸序列中的双链锻炼可以通过非同源末端连接(NHEJ)修复过程进行修复。因为NHEJ是易错的,缺失至少一个核苷酸、插入至少一个核苷酸、取代至少一个核苷酸或其组合可能会出现在修复断裂期间。因此,靶向的核苷酸序列可以经修饰或灭活。例如,单核苷酸改变(SNP)可以产生改变的蛋白质产物,或者编码序列阅读框的移动可以灭活或"敲除"序列,从而不再产生蛋白质产物。在存在任选供体多核苷酸的实施方式中,供体多核苷酸中的供体序列在修复双链断裂期间可与靶位点的核苷酸序列交换或整合至其中。例如,在供体序列侧接这样上游和下游序列的实施方式中,所述上游和下游序列具有分别与非植物真核细胞或细胞器核苷酸序列中靶位点的上游和

下游序列实质上的序列同一性,供体序列在通过同源性导向的修复过程介导的修复期间可以与靶位点的核苷酸序列交换或整合至其中。或者,在供体序列侧接相容突出端(或者该相容突出端由Cms1多肽原位生成)的实施方式中,供体序列在双链断裂修复期间通过非同源性修复过程可以直接连接切割的核苷酸序列。将供体序列交换或整合至核苷酸序列修饰靶向的核苷酸序列,或者将外源性序列引入非植物真核细胞或细胞器靶向的核苷酸序列。

[0112] 在一些实施方式中,由一种或多种Cms1核酸酶作用所导致的双链断裂以这样的方式修复,所述方式使DNA从非植物真核细胞或细胞器染色体中缺失。在一些实施方式中,一个碱基、数个碱基(即2、3、4、5、6、7、8、9或10个碱基)或大部分的DNA(即,超过10、超过50、超过100、或超过500个碱基)从非植物真核细胞或细胞器中缺失。

[0113] 在一些实施方式中,作为由一种或多种Cms1核酸酶所导致的双链断裂的结果,非植物真核基因的表达可能会被调节。在一些实施方式中,非植物真核基因的表达可能会被变体Cms1酶所调节,所述变体Cms1酶包括使Cms1核酸酶无法生成双链断裂的突变。在一些优选实施方式中,包括使Cms1核酸酶不可以生成双链断裂的突变的变体Cms1核酸酶可以融合转录活化或转录抑制结构域。

[0114] 在一些实施方式中,培养这样的真核细胞以生成真核生物,所述真核细胞在其核和/或细胞器染色体DNA包括由一种或多种Cms1核酸酶作用所导致的突变。在一些实施方式中,培养这样的真核细胞以生成真核生物,所述真核细胞中的基因表达因为一种或多种Cms1核酸酶或一种或多种变体Cms1核酸酶而被调节。培养非植物真核细胞以生成真核生物的方法为本领域已知,例如美国专利申请号2016/0208243和2016/0138008,其各自通过引用纳入本文。

[0115] 本发明可用于任何真核物种的转化,包括但不限于动物(包括但不限于哺乳动物、昆虫、鱼类、鸟类和爬行动物)、真菌、变形虫和酵母。

[0116] 向非植物真核细胞或细胞器引入核酸酶蛋白质、编码核酸酶蛋白质的DNA或RNA分子、引导RNA或编码引导RNA的DNA分子、和任选的供体序列DNA分子的方法为本领域已知,例如美国专利申请号2016/0208243,通过引用纳入本文。对工业应用特别具有价值的非植物真核细胞或细胞器的示例性遗传修饰也为本领域已知,例如美国专利申请号2016/0208243,通过引用纳入本文。

[0117] C.修饰原核基因组中核苷酸序列的方法

[0118] 本文提供了用于修饰原核(例如,细菌或古细菌)细胞核苷酸序列的方法。该方法包括向靶细胞引入靶向DNA的RNA或编码靶向DNA的RNA的DNA多核苷酸,其中靶向DNA的RNA包括:(a)第一区段,其包含与靶DNA中序列互补的核苷酸序列;和(b)第二区段,其与Cms1多肽相互作用;和,向靶细胞引入Cms1多肽或编码Cms1多肽的多核苷酸,其中Cms1多肽包括:(a)结合RNA的部分,其与靶向DNA的RNA相互作用;和(b)活性部分,其显示定点酶促活性。然后可以在Cms1多肽表达并且切割核苷酸序列的条件下培养靶细胞。需指出的是,本文所述系统不需要添加外源性Mg^{2+}或任何其他离子。最后,可选择包含经修饰核苷酸序列的原核细胞。还应注意,包含经修饰的一个或多个核苷酸序列的原核细胞不是编码感兴趣的Cms1多肽的多核苷酸的原生宿主细胞,并且,利用非天然产生的引导RNA来实现一个或多个原核核苷酸序列中的所需变化。需要进一步指出的是靶向的DNA可能作为原核染色体的部分存在或者存在于原核细胞中的一个或多个质粒或其它非染色体DNA分子。

[0119] 在一些实施方式中,该方法可以包括向原核细胞中引入一个Cms1多肽(或编码核酸)和一个引导RNA(或编码DNA),其中Cms1多肽在原核细胞DNA的靶核苷酸序列中引入一个双链断裂。在一些实施方式中,该方法可以包括向原核细胞中引入一个Cms1多肽(或编码核酸)和至少一个引导RNA(或编码DNA),其中Cms1多肽在原核细胞DNA的靶核苷酸序列中引入超过一个双链断裂(即2、3或超过3个双链断裂)。在不存在任选供体多核苷酸实施方式中,核苷酸序列中的双链锻炼可以通过非同源末端连接(NHEJ)修复过程进行修复。因为NHEJ是易错的,缺失至少一个核苷酸、插入至少一个核苷酸、取代至少一个核苷酸或其组合可能会出现在修复断裂期间。因此,靶向的核苷酸序列可以经修饰或灭活。例如,单核苷酸改变(SNP)可以产生改变的蛋白质产物,或者编码序列阅读框的移动可以灭活或"敲除"序列,从而不再产生蛋白质产物。在存在任选供体多核苷酸的实施方式中,供体多核苷酸中的供体序列在修复双链断裂期间可与靶位点的核苷酸序列交换或整合至其中。例如,在供体序列侧接这样上游和下游序列的实施方式中,所述上游和下游序列具有分别与原核细胞核苷酸序列中靶位点的上游和下游序列实质上的序列同一性,供体序列在通过同源性导向的修复过程介导的修复期间可以与靶位点的核苷酸序列交换或整合至其中。或者,在供体序列侧接相容突出端(或者该相容突出端由Cms1多肽原位生成)的实施方式中,供体序列在双链断裂修复期间通过非同源性修复过程可以直接连接切割的核苷酸序列。将供体序列交换或整合至核苷酸序列修饰靶向的核苷酸序列,或者将外源性序列引入原核细胞DNA的靶向的核苷酸序列。

[0120] 在一些实施方式中,由一种或多种Cms1核酸酶作用所导致的双链断裂以这样的方式修复,所述方式使DNA从原核细胞DNA中缺失。在一些实施方式中,一个碱基、数个碱基(即2、3、4、5、6、7、8、9或10个碱基)或大部分的DNA(即,超过10、超过50、超过100、或超过500个碱基)从原核细胞DNA中缺失。

[0121] 在一些实施方式中,作为一种或多种Cms1核酸酶所导致的双链断裂的结果,原核基因的表达可能会被调节。在一些实施方式中,原核基因的表达可能会被变体Cms1核酸酶所调节,所述变体Cms1核酸酶包括使Cms1核酸酶无法生成双链断裂的突变。在一些优选实施方式中,包括使Cms1核酸酶不可以生成双链断裂的突变的变体Cms1核酸酶可以融合转录活化或转录抑制结构域。

[0122] 本发明可以用于转化任何原核生物,包括但不限于蓝藻细菌,棒状杆菌(Corynebacterium sp.),双歧杆菌(Bifidobacterium sp.),分枝杆菌(Mycobacterium sp.),链霉菌(Streptomyces sp.),温双岐菌(Thermobifida sp.),衣原体(Chlamydia sp.),原绿球藻(Prochlorococcus sp.),聚球藻(Synechococcus sp.),热聚球藻(Thermosynechococcus sp.),泉栖热菌(Thermus sp.),芽孢杆菌(Bacillus sp.),梭菌(Clostridium sp.),土芽孢杆菌(Geobacillus sp.),乳杆菌(Lactobacillus sp.),李斯特菌(Listeria sp.),葡萄球菌(Staphylococcus sp.),链球菌(Streptococcus sp.),梭菌(Fusobacterium sp.),农杆菌(Agrobacterium sp.),慢生根瘤菌(Bradyrhizobium sp.),埃立克体(Ehrlichia sp.),中慢生根瘤菌(Mesorhizobium sp.),硝酸菌(Nitrobacter sp.),立克次体(Rickettsia sp.),沃尔巴克氏体(Wolbachia sp.),单胞发酵菌(Zymomonas sp.),伯克霍尔德菌(Burkholderia sp.),奈瑟氏菌(Neisseria sp.),罗尔斯通菌(Ralstonia sp.),不动杆菌(Acinetobacter sp.),欧文氏菌(Erwinia sp.),埃

希氏杆菌(Escherichia sp.),嗜血杆菌(Haemophilus sp.),军团杆菌(Legionella sp.),巴斯德菌(Pasteurella sp.),假单胞菌(Pseudomonas sp.),嗜冷杆菌(Psychrobacter sp.),沙门氏菌(Salmonella sp.),希瓦氏菌(Shewanella sp.),志贺氏杆菌(Shigella sp.),弧菌(Vibrio sp.),黄单胞菌(Xanthomonas sp.),木杆菌(Xylella sp.),耶尔森菌(Yersinia sp.),弯曲杆菌(Campylobacter sp.),脱硫弧菌(Desulfovibrio sp.),螺杆菌(Helicobacter sp.),地杆菌(Geobacter sp.),细螺旋体(Leptospira sp.),密螺旋体(Treponema sp.),支原菌(Mycoplasma sp.)和热袍菌(Thermotoga sp.)。

[0123] 向原核细胞或细胞器引入核酸酶蛋白质、编码核酸酶蛋白质的DNA或RNA分子、引导RNA或编码引导RNA的DNA分子、和任选的供体序列DNA分子的方法为本领域已知,例如美国专利申请号2016/0208243,通过引用纳入本文。对工业应用特别具有价值的原核细胞或细胞器的示例性遗传修饰也为本领域已知,例如美国专利申请号2016/0208243,通过引用纳入本文。

[0124] D.修饰病毒基因组中核苷酸序列的方法

[0125] 本文提供了用于修饰病毒基因组的核苷酸序列的方法。该方法包括向包含感兴趣的病毒的细胞引入靶向DNA的RNA或编码靶向DNA的RNA的DNA多核苷酸,其中靶向DNA的RNA包括:(a)第一区段,其包含与靶DNA中序列互补的核苷酸序列;和(b)第二区段,其与Cms1多肽相互作用;和,向靶细胞引入Cms1多肽或编码Cms1多肽的多核苷酸,其中Cms1多肽包括:(a)结合RNA的部分,其与靶向DNA的RNA相互作用;和(b)活性部分,其显示定点酶促活性。然后可在表达Cms1多肽并切割病毒核苷酸序列的条件下培养包含感兴趣的病毒的靶细胞。或者,可以在体外操作病毒基因组,其中将引导多核苷酸,Cms1多肽和任选的供体多核苷酸与感兴趣的病毒DNA序列在细胞宿主外部一起孵育。

[0126] V.调节基因表达的方法

[0127] 本文公开的方法还包括基因组宿主中核苷酸序列的修饰或核苷酸序列的表达的调节。该方法可包括向基因组宿主中引入编码至少一种融合蛋白或编码至少一种融合蛋白的核酸,其中融合蛋白包括Cms1多肽或其片段或变体和效应物结构域,和(b)至少一种引导RNA或编码引导RNA的DNA,其中引导RNA将融合蛋白的Cms1多肽引导至靶DNA中的靶位点,并且融合蛋白的效应物结构域修饰染色体序列或调节靶DNA序列处或附近的一种或多种基因的表达。

[0128] 本文描述了融合蛋白,其包括Cms1多肽或其片段或变体以及效应物结构域。通常,本文所公开的融合蛋白可以还包括至少一种核定位信号、质体信号肽、线粒体信号肽或能够运输蛋白质至多个亚细胞位置的信号肽。本文描述了编码融合蛋白的核酸。在一些实施方式中,融合蛋白可以分离的蛋白质(其还可包含细胞穿透域)的形式引入基因组宿主。此外,分离的融合蛋白可以是包括引导RNA的蛋白质-RNA复合物的部分。在其它实施方式中,融合蛋白可以RNA分子(可以被加帽和/或聚腺苷酸化)形式引入基因组宿主中。在其它实施方式中,融合蛋白可以DNA分子形式引入基因组宿主。例如,融合蛋白和引导RNA可以离散的DNA分子形式或以同一DNA分子的部分形式引入基因组宿主。这类DNA分子可以是质粒载体。

[0129] 在一些实施方式中,该方法还包括向基因组宿主引入本文所述的至少一种供体多核苷酸。本文描述了将分子引入基因组宿主(例如细胞)中的手段,以及用于培养细胞(包括含细胞器的细胞)的手段。

[0130] 在其中融合蛋白效应物结构域是切割结构域的具体实施方式中,该方法可以包括向基因组宿主引入一种融合蛋白(或编码一种融合蛋白的核酸)和两种引导RNA(或编码两种引导RNA的DNA)。两种引导RNA将融合蛋白引导至染色体序列中的两个不同靶位点,其中融合蛋白二聚化(例如,形成同二聚体),因此两个切割结构域可以将双链断裂引入靶DNA序列。在不存在任选供体多核苷酸的实施方式中,靶DNA序列中的双链断裂可以通过非同源末端连接(NHEJ)修复过程进行修复。因为NHEJ是易错的,缺失至少一个核苷酸、插入至少一个核苷酸、取代至少一个核苷酸或其组合可能会出现在修复断裂期间。因此,靶染色体序列可经修饰或灭活。例如,单核苷酸改变(SNP)可以产生改变的蛋白质产物,或者编码序列阅读框的移动可以灭活或敲除摄序列,从而不再产生蛋白质产物。在存在任选供体多核苷酸的实施方式中,供体多核苷酸中的供体序列在修复双链断裂期间可与靶位点的靶DNA序列交换或整合至其中。例如,在供体序列侧接这样上游和下游序列的实施方式中,所述上游和下游序列具有分别与靶DNA序列中靶位点的上游和下游序列实质上的序列同一性,供体序列在通过同源性导向的修复过程介导的修复期间可以与靶位点的靶DNA序列交换或整合至其中。或者,在供体序列侧接相容突出端(或者该相容突出端由Cms1多肽原位生成)的实施方式中,供体序列在双链断裂修复期间通过非同源性修复过程可以直接连接切割的靶DNA序列。将供体序列交换或整合至靶DNA序列修饰靶DNA序列,或者将外源性序列引入靶DNA序列。

[0131] 在其中融合蛋白效应物结构域是切割结构域的其它实施方式中,该方法可包括向基因组宿主引入两种不同的融合蛋白(或编码两种不同的融合蛋白的核酸)和两种引导RNA(或编码两种引导RNA的DNA)。融合蛋白可以不同,如本文他处详述。各引导RNA将融合蛋白引导至靶DNA序列中的特定靶位点,其中融合蛋白可以二聚化(例如,形成同二聚体),从而两个切割结构域可以将双链断裂引入靶DNA序列。在不存在任选供体多核苷酸的实施方式中,得到的双链断裂可以通过非同源性修复过程修复,这样的话在断裂修复期间可能会出现缺失至少一个核苷酸,插入至少一个核苷酸,取代至少一个核苷酸或其组合。在任选供体多核苷酸存在的实施方式中,在通过基于同源性的修复过程(例如,在供体序列侧接这样上游和下游序列的实施方式中,所述上游和下游序列具有分别与染色体序列中靶位点的上游和下游序列实质上的序列同一性中)或非同源性的修复过程(例如,在供体序列侧接相容突出端的实施方式中)的双链断裂修复期间,供体多核苷酸中的供体序列可以与染色体序列交换或整合至其中。

[0132] 在其中融合蛋白效应物结构域是转录活化结构域或转录阻遏物结构域的某些实施方式中,该方法可以包括向基因组宿主引入一种融合蛋白(或编码一种融合蛋白的核酸)和一种引导RNA(或编码一种引导RNA的DNA)。引导RNA将融合蛋白导向特定靶DNA序列,其中转录活化结构域或转录阻遏物结构域分别活化或抑制位于靶DNA序列附近的一个或多个基因的表达。即,转录可能会受到与靶DNA序列非常接近的基因的影响,或者可能受到与靶DNA序列相距更远的基因的影响。本领域已知可以通过远距离序列(distantly located sequence)调控基因转录,所述远距离序列可能离转录起始位点数千碱基远的位置或者甚至在不同的染色体上(Harmston和Lenhard(2013)Nucleic Acids Res41:7185-7199)。

[0133] 在其中融合蛋白效应物结构域是表观遗传修饰结构域的其它实施方式中,该方法可以包括向基因组宿主引入一种融合蛋白(或编码一种融合蛋白的核酸)和一种引导RNA

(或编码一种引导RNA的DNA)。该引导RNA将融合蛋白导向至特定靶DNA序列,其中表观遗传修饰结构域修饰靶DNA序列的结构。表观遗传修饰包括乙酰化,组蛋白的甲基化和/或核苷酸甲基化。在一些情况下,染色体序列的结构修饰导致染色体序列表达的变化。

[0134] VI.包含遗传修饰的生物体

[0135] A.真核生物

[0136] 本文提供了真核生物、真核细胞、细胞器和植物胚胎,其包括已经使用本文所述的Cms1多肽介导的或融合蛋白介导的方法修饰的至少一种核苷酸序列。还提供了真核生物、真核细胞、细胞器和植物胚胎,其包括至少一种DNA或RNA分子,其编码Cms1多肽或融合蛋白,其靶向感兴趣的染色体序列或融合蛋白,至少一种引导RNA,以及任选的一种或多种供体多核苷酸。本文公开的经遗传修饰的真核生物对于修饰的核苷酸序列可以是杂合的,或对于修饰的核苷酸序列可以是纯合的。在细胞器DNA中包括一种或多种基因修饰的真核细胞可以是异质的或同质的。

[0137] 可以对真核生物、真核细胞、细胞器和植物胚胎的经修饰的染色体序列进行修饰从而使其灭活,具有上调的或下调的表达,或生成改变的蛋白产物,或包括整合的序列。可以将修饰的染色体序列灭活,从而使序列不再转录和/或功能性蛋白质产物不再生成。因此,包括灭活的染色体序列的经遗传修饰的真核生物可以被称为“敲除”或“条件性敲除”。失活的染色体序列可包括缺失突变(即,一个或多个核苷酸的缺失),插入突变(即,一个或多个核苷酸的插入)或无义突变(即,用单核苷酸取代另一核苷酸从而引入终止密码子)。突变的结果是,靶染色体序列失活,从而不产生功能蛋白。失活的染色体序列不包含外源引入的序列。本文还包括遗传修饰的真核生物,其中2、3、4、5、6、7、8、9或10个或更多个染色体序列被灭活。

[0138] 修饰的染色体序列还可以被改变,从而使其编码变体蛋白产物。例如,包含修饰的染色体序列的经遗传修饰的真核生物可包含靶点突变或其它修饰,从而产生改变的蛋白质产物。在一个实施方式中,可以修饰染色体序列,从而改变至少一个核苷酸,并且表达的蛋白质包含一个改变的氨基酸残基(错义突变)。在另一个实施方式中,可以修饰染色体序列以包含多于一个的错义突变,从而改变多于一个的氨基酸。另外,可以修饰染色体序列以具有三个核苷酸的缺失或插入,从而表达的蛋白质包括单个氨基酸的缺失或插入。与野生型蛋白质相比,改变或变异的蛋白质可具有改变的特性或活性,例如改变的底物特异性,改变的酶活性,改变的动力学速率等。

[0139] 在一些实施方式中,遗传修饰的真核生物可以包括至少一个染色体整合的核苷酸序列。包括整合序列的遗传修饰的真核生物可以被称为“敲入”或“条件性敲入”。作为整合序列的核苷酸序列可以例如编码直系同源蛋白质,内源性蛋白质或两者的组合。在一个实施方式中,可将编码直系同源蛋白质或内源性蛋白质的序列整合到编码蛋白质的核或细胞染色体序列中,从而使染色体序列失活,但是表达外源序列。在这样的情况中,编码直向同源蛋白或内源性蛋白的序列可以操作性地连接启动子控制序列。或者,可将编码直系同源蛋白质或内源性蛋白质的序列整合到核或细胞染色体序列中而不影响染色体序列的表达。例如,编码蛋白质的序列可以被整合到“安全港”基因座中。本公开还包括遗传修饰的真核生物,其中2、3、4、5、6、7、8、9或10个或更多个序列(包括编码蛋白质的序列)被整合到基因组中。本文公开的任何感兴趣的基因均可被引入整合进入真核核或细胞器的染色体序列。

在特定实施方式中,将增加植物生长或产量的基因整合到染色体中。

[0140] 编码蛋白质的染色体整合的序列可以编码感兴趣的蛋白质的野生型或者可以编码包括至少一种修饰的蛋白质,从而生成蛋白质的改变形式。例如,编码疾病或病症相关蛋白质的染色体整合序列可包含至少一种修饰,从而产生的蛋白质的变化形式能引起或增强相关的病症。或者,编码疾病或病症相关蛋白质的染色体整合序列可包含至少一种修饰,从而该蛋白质的改变形式保护真核生物或真核细胞免受相关疾病或病症的发展。

[0141] 在某些实施方式中,遗传修饰的真核生物可以包括编码蛋白质的至少一种修饰的染色体序列,从而改变蛋白质的表达模式。例如,控制蛋白质表达的调控区域如启动子或转录因子结合位点可以经改变,从而使蛋白质过表达,或者改变蛋白质的组织特异性或时序性表达或其组合。或者,可以使用条件性敲除系统改变蛋白质的表达模式。条件性敲除系统的非限制示例包括Cre-lox重组系统。Cre-lox重组系统包含Cre重组酶,这是一种位点特异性DNA重组酶,其可以催化核酸分子中特定位点(lox位点)之间的核酸序列重组。使用该系统产生时间和组织特异性表达的方法是本领域已知的。

[0142] B.原核生物

[0143] 本文提供了原核生物和原核细胞,其包括已经使用本文所述的Cms1多肽介导的或融合蛋白介导的方法修饰的至少一种核苷酸序列。还提供了原核生物和原核细胞,其包括至少一种DNA或RNA分子,其编码Cms1多肽或融合蛋白,其靶向感兴趣的DNA序列或融合蛋白,至少一种引导RNA,以及任选的一种或多种供体多核苷酸。

[0144] 可以对原核生物和原核细胞的经修饰的DNA序列进行修饰从而使其灭活,具有上调的或下调的表达,或生成改变的蛋白产物,或包括整合的序列。可以将修饰的DNA序列灭活,从而使序列不再转录和/或功能性蛋白质产物不再生成。因此,包括灭活的染色体序列的经遗传修饰的原核生物可以被称为“敲除”或“条件性敲除”。失活的DNA序列可包括缺失突变(即,一个或多个核苷酸的缺失),插入突变(即,一个或多个核苷酸的插入)或无义突变(即,用单核苷酸取代另一核苷酸从而引入终止密码子)。突变的结果是,靶DNA序列失活,从而不产生功能蛋白。失活的DNA序列不包含外源引入的序列。本文还包括遗传修饰的原核生物,其中2、3、4、5、6、7、8、9或10个或更多个DNA序列被灭活。

[0145] 经修饰的DNA序列还可以被改变,从而使其编码变体蛋白质产物。例如,包含经修饰的DNA序列的经遗传修饰的原核生物可包含靶点突变或其它修饰,从而产生改变的蛋白质产物。在一个实施方式中,可以修饰DNA序列,从而改变至少一个核苷酸,并且表达的蛋白质包含一个改变的氨基酸残基(错义突变)。在另一个实施方式中,可以修饰DNA序列以包含多于一个的错义突变,从而改变多于一个的氨基酸。另外,可以修饰DNA序列以具有三个核苷酸的缺失或插入,从而表达的蛋白质包括单个氨基酸的缺失或插入。与野生型蛋白质相比,改变或变异的蛋白质可具有改变的特性或活性,例如改变的底物特异性,改变的酶活性,改变的动力学速率等。

[0146] 在一些实施方式中,经遗传修饰的原核生物可以包括至少一个整合的核苷酸序列。包括整合序列的遗传修饰的原核生物可以被称为“敲入”或“条件性敲入”。作为整合序列的核苷酸序列可以例如编码直系同源蛋白质,内源性蛋白质或两者的组合。在一个实施方式中,可将编码直系同源蛋白质或内源性蛋白质的序列整合到编码蛋白质的原核DNA序列中,从而使该原核序列失活,但是表达外源序列。在这样的情况中,编码直向同源蛋白或

内源性蛋白的序列可以操作性地连接启动子控制序列。或者,可将编码直系同源蛋白质或内源性蛋白质的序列整合进入原核DNA序列,而不影响原生原核序列的表达。例如,编码蛋白质的序列可以被整合到“安全港”基因座中。本公开还包括经遗传修饰的原核生物,其中2、3、4、5、6、7、8、9或10个或更多个序列(包括编码蛋白质的序列)被整合进入原核基因组或原核生物所含质粒中。如本文所公开的任何感兴趣的基因都可被整合进入原核染色体、质粒或其它染色体外DNA的DNA序列中。

[0147] 编码蛋白质的整合的序列可以编码感兴趣的蛋白质的野生型或者可以编码包括至少一种修饰的蛋白质,从而生成蛋白质的改变形式。例如,编码疾病或病症相关蛋白质的整合序列可包含至少一种修饰,从而产生的蛋白质的变化形式能引起或增强相关的病症。或者,编码疾病或病症相关蛋白质的整合序列可包含至少一种修饰,从而蛋白质的改变形式能降低原核生物的感染性。

[0148] 在某些实施方式中,经遗传修饰的原核生物可以包括编码蛋白质的至少一种修饰的DNA序列,从而改变蛋白质的表达模式。例如,控制蛋白质表达的调控区域如启动子或转录因子结合位点可以经改变,从而使蛋白质过表达,或者改变蛋白质的时序性表达或其组合。或者,可以使用条件敲除系统改变蛋白质的表达模式。条件性敲除系统的非限制示例包括Cre-lox重组系统。Cre-lox重组系统包含Cre重组酶,这是一种位点特异性DNA重组酶,其可以催化核酸分子中特定位点(lox位点)之间的核酸序列重组。使用该系统产生时序表达的方法是本领域已知的。

[0149] C.病毒

[0150] 本文提供了病毒和病毒基因组,其包括已经使用本文所述的Cms1多肽介导的或融合蛋白介导的方法修饰的至少一种核苷酸序列。还提供了病毒和病毒基因组,其包括至少一种DNA或RNA分子,其编码Cms1多肽或融合蛋白,其靶向感兴趣的DNA序列或融合蛋白,至少一种引导RNA,以及任选的一种或多种供体多核苷酸。

[0151] 可以对病毒和病毒基因组的经修饰的DNA序列进行修饰从而使其灭活,具有上调的或下调的表达,或生成改变的蛋白产物,或包括整合的序列。可以将修饰的DNA序列灭活,从而使序列不再转录和/或功能性蛋白质产物不再生成。因此,包括灭活的染色体序列的经遗传修饰的病毒可以被称为“敲除”或“条件性敲除”。失活的DNA序列可包括缺失突变(即,一个或多个核苷酸的缺失),插入突变(即,一个或多个核苷酸的插入)或无义突变(即,用单核苷酸取代另一核苷酸从而引入终止密码子)。突变的结果是,靶DNA序列失活,从而不产生功能蛋白。失活的DNA序列不包含外源引入的序列。本文还包括遗传修饰的病毒,其中2、3、4、5、6、7、8、9或10个或更多个病毒序列被灭活。

[0152] 经修饰的DNA序列还可以被改变,从而使其编码变体蛋白产物。例如,包含经修饰的DNA序列的经遗传修饰的病毒可包含靶点突变或其它修饰,从而产生改变的蛋白质产物。在一个实施方式中,可以修饰DNA序列,从而改变至少一个核苷酸,并且表达的蛋白质包含一个改变的氨基酸残基(错义突变)。在另一个实施方式中,可以修饰DNA序列以包含多于一个的错义突变,从而改变多于一个的氨基酸。另外,可以修饰DNA序列以具有三个核苷酸的缺失或插入,从而表达的蛋白质包括单个氨基酸的缺失或插入。与野生型蛋白质相比,改变或变异的蛋白质可具有改变的特性或活性,例如改变的底物特异性,改变的酶活性,改变的动力学速率等。

[0153] 在一些实施方式中,经遗传修饰的病毒可以包括至少一个整合的核苷酸序列。包括整合序列的遗传修饰的病毒可以被称为“敲入”或“条件性敲入”。作为整合序列的核苷酸序列可以例如编码直系同源蛋白质,内源性蛋白质或两者的组合。在一个实施方式中,可将编码直系同源蛋白质或内源性蛋白质的序列整合到编码蛋白质的病毒DNA序列中,从而使该病毒序列失活,但是表达外源序列。在这样的情况中,编码直向同源蛋白或内源性蛋白的序列可以操作性地连接启动子控制序列。或者,可将编码直系同源蛋白质或内源性蛋白质的序列整合进入病毒DNA序列,而不影响原生病毒序列的表达。例如,编码蛋白质的序列可以被整合到“安全港”基因座中。本公开还包括遗传修饰的病毒,其中2、3、4、5、6、7、8、9或10个或更多个序列(包括编码蛋白质的序列)被整合到病毒基因组中。本文公开的任何感兴趣的基因都可以被整合到病毒基因组的DNA序列中。

[0154] 编码蛋白质的整合的序列可以编码感兴趣的蛋白质的野生型或者可以编码包括至少一种修饰的蛋白质,从而生成蛋白质的改变形式。例如,编码疾病或病症相关蛋白质的整合序列可包含至少一种修饰,从而产生的蛋白质的变化形式能引起或增强相关的病症。或者,编码疾病或病症相关蛋白质的整合序列可包含至少一种修饰,从而蛋白质的改变形式能降低病毒的感染性。

[0155] 在某些实施方式中,经遗传修饰的病毒可以包括编码蛋白质的至少一种修饰的DNA序列,从而改变蛋白质的表达模式。例如,控制蛋白质表达的调控区域如启动子或转录因子结合位点可以经改变,从而使蛋白质过表达,或者改变蛋白质的时序性表达或其组合。或者,可以使用条件敲除系统改变蛋白质的表达模式。条件性敲除系统的非限制示例包括Cre-lox重组系统。Cre-lox重组系统包含Cre重组酶,这是一种位点特异性DNA重组酶,其可以催化核酸分子中特定位点(lox位点)之间的核酸序列重组。使用该系统产生时序表达的方法是本领域已知的。

[0156] 本说明书中涉及的所有专利申请和出版物指示本发明涉及领域技术人员的水平。所有发表物和专利申请通过引用纳入本文,就好像将各篇单独的发表物或专利申请具体和单独地通过引用纳入本文那样。

[0157] 虽然出于方便理解的目的,通过阐述和举例的方式详细描述了上述发明,但可明显看出,某些改变和修改应属于所附权利要求书的范围。

[0158] 本发明的实施方式包括:

[0159] 1.一种修饰真核细胞基因组中靶位点的核苷酸序列的方法,其包括:

[0160] 向所述真核细胞中引入

[0161] (i)靶向DNA的RNA,或编码靶向DNA的RNA的DNA多核苷酸,其中所述靶向DNA的RNA包括:(a)第一区段,其包含与靶DNA中序列互补的核苷酸序列;和(b)第二区段,其与Cms1多肽相互作用;和

[0162] (ii)Cms1多肽或编码Cms1多肽的多核苷酸,其中所述Cms1多肽包含:(a)结合RNA的部分,其与靶向DNA的RNA相互作用;和(b)活性部分,其显示定点酶促活性。

[0163] 2.一种修饰原核细胞基因组中靶位点的核苷酸序列的方法,其包括:

[0164] 向所述原核细胞中引入

[0165] (i)靶向DNA的RNA,或编码靶向DNA的RNA的DNA多核苷酸,其中所述靶向DNA的RNA包括:(a)第一区段,其包含与靶DNA中序列互补的核苷酸序列;和(b)第二区段,其与Cms1多

肽相互作用;和

[0166] (ii)Cms1多肽或编码Cms1多肽的多核苷酸,其中所述Cms1多肽包含:(a)结合RNA的部分,其与靶向DNA的RNA相互作用;和(b)活性部分,其显示定点酶促活性,

[0167] 其中,所述原核细胞不是编码所述Cms1多肽的基因的原生宿主。

[0168] 3.一种修饰植物细胞基因组中靶位点的核苷酸序列的方法,其包括:

[0169] 向所述植物细胞中引入

[0170] (i)靶向DNA的RNA,或编码靶向DNA的RNA的DNA多核苷酸,其中所述靶向DNA的RNA包括:(a)第一区段,其包含与靶DNA中序列互补的核苷酸序列;和(b)第二区段,其与Cms1多肽相互作用;和

[0171] (ii)Cms1多肽或编码Cms1多肽的多核苷酸,其中所述Cms1多肽包含:(a)结合RNA的部分,其与靶向DNA的RNA相互作用;和(b)活性部分,其显示定点酶促活性。

[0172] 4.如实施方式3所述的方法,其还包括:

[0173] 在表达所述Cms1多肽并在所述靶位点处切割核苷酸序列以生成经修饰的核苷酸序列的条件下培养所述植物;和

[0174] 选择包含所述经修饰的核苷酸序列的植物。

[0175] 5.如实施方式1-4中任一项所述的方法,其中切割靶位点的核苷酸序列包括双链断裂,所述双链断裂位于或邻近靶向DNA的RNA序列所靶向的序列。

[0176] 6.如实施方式5所述的方法,其中所述双链断裂是交错的双链断裂。

[0177] 7.如实施方式6所述的方法,其中所述交错的双链断裂产生3-6个核苷酸的5'突出端。

[0178] 8.如实施方式1-7中任一项所述的方法,其中所述靶向DNA的RNA是引导RNA(gRNA)。

[0179] 9.如实施方式1-8中任一项所述的方法,其中所述经修饰的核苷酸序列包括细胞基因组中异源性DNA的插入,细胞基因组中核苷酸序列的缺失,或细胞基因组中至少一个核苷酸的突变。

[0180] 10.如实施方式1-9中任一项所述的方法,其中,所述Cms1多肽选自下组:SEQ ID NO:20-23、30-69、208-211和222-254。

[0181] 11.如实施方式1-10中任一项所述的方法,其中,编码Cms1多肽的所述多核苷酸选自下组:SEQ ID NO:16-19、24-27、70-146、174-176、212-215和255-287。

[0182] 12.如实施方式1-11中任一项所述的方法,其中,所述Cms1多肽与选自下组的一个或多个多肽序列具有至少80%同一性:SEQ ID NO:20-23、30-69、208-211和222-254。

[0183] 13.如实施方式1-12中任一项所述的方法,其中,编码Cms1多肽的所述多核苷酸与选自下组的一个或多个核酸序列具有至少70%同一性:SEQ ID NO:16-19、24-27、70-146、174-176、212-215和255-287。

[0184] 14.如实施方式1-13中任一项所述的方法,其中,所述Cms1多肽形成同二聚体或异二聚体。

[0185] 15.如实施方式3所述的方法,其中,所述植物细胞来自单子叶植物。

[0186] 16.如实施方式3所述的方法,其中,所述植物细胞来自双子叶植物。

[0187] 17.如实施方式1-16中任一项所述的方法,其中Cms1多肽的表达在诱导型或组成

型启动子的控制下。

[0188] 18.如实施方式1-17中任一项所述的方法,其中Cms1多肽的表达在细胞类型特异性或发育优先型启动子的控制下。

[0189] 19.如实施方式1-18中任一项所述的方法,其中PAM序列包括5'-TTN,其中N可以是任何核苷酸。

[0190] 20.如实施方式3所述的方法,其中位于植物细胞基因组靶位点的所述核苷酸序列编码SBP酶,FBP酶,FBP醛缩酶,AGP酶大亚基,AGP酶小亚基,蔗糖磷酸合成酶,淀粉合成酶,PEP羧化酶,丙酮酸磷酸二激酶,转酮醇酶,rubisco小亚基,或rubisco活化酶蛋白,或编码调节一个或多个基因表达的转录因子,所述基因编码SBP酶,FBP酶,FBP醛缩酶,AGP酶大亚基,AGP酶小亚基,蔗糖磷酸合成酶,淀粉合成酶,PEP羧化酶,丙酮酸磷酸二激酶,转酮醇酶,rubisco小亚基或rubisco活化酶蛋白。

[0191] 21.如实施方式1-20中任一项所述的方法,所述方法还包括将靶位点与供体多核苷酸接触,其中供体多核苷酸、供体多核苷酸的部分、供体多核苷酸的拷贝或供体多核苷酸拷贝的部分整合至靶DNA中。

[0192] 22.如实施方式1-21中任一项所述的方法,其中所述靶DNA经修饰,从而使靶DNA内的核苷酸缺失。

[0193] 23.如实施方式1-22中任一项所述的方法,其中编码Cms1多肽的所述多核苷酸经密码子优化以在植物细胞中表达。

[0194] 24.如实施方式1-23中任一项所述的方法,其中,所述核苷酸序列的表达增加或降低。

[0195] 25.如实施方式1-24中任一项所述的方法,其中,编码Cms1多肽的多核苷酸操作性连接至启动子,所述启动子是组成型、细胞特异型、诱导型或被自杀外显子的可变剪接活化的启动子。

[0196] 26.如实施方式1-25中任一项所述的方法,其中,所述Cms1多肽包括一个或多个突变,所述突变减弱或消除所述Cms1多肽的核酸酶活性。

[0197] 27.如实施方式26所述的方法,其中,所述突变的Cms1多肽包含突变,当经比对以实现最大同一性时,所述突变处在对应于SmCms1(SEQ ID NO:10)的701或922位的位置处,或对应于SulfCms1(SEQ ID NO:11)的848或1213位的位置处。

[0198] 28.如实施方式27所述的方法,其中,处于对应于SmCms1(SEQ ID NO:10)的701或922位的位置处的所述突变分别是D701A和E922A,或处于对应于SulfCms1(SEQ ID NO:11)的848和1213位的位置处的所述突变分别是D848A和D1213A。

[0199] 29.如实施方式26-28中任一项所述的方法,其中,突变的Cms1多肽与转录活化结构域融合。

[0200] 30.如实施方式29所述的方法,其中,突变的Cms1多肽直接融合至转录活化结构域或通过接头融合至转录活化结构域。

[0201] 31.如实施方式26-28中任一项所述的方法,其中,突变的Cms1多肽与转录阻遏物结构域融合。

[0202] 32.如实施方式31所述的方法,其中,突变的Cms1多肽通过接头与转录阻遏物结构域融合。

[0203] 33.如实施方式1-32中任一项所述的方法,其中,所述Cms1多肽还包括核定位信号。

[0204] 34.如实施方式33所述的方法,其中所述核定位信号包括SEQ ID NO:1,或其由SEQ ID NO:2编码。

[0205] 35.如实施方式1-32中任一项所述的方法,其中,所述Cms1多肽还包括叶绿体信号肽。

[0206] 36.如实施方式1-32中任一项所述的方法,其中,所述Cms1多肽还包含线粒体信号肽。

[0207] 37.如实施方式1-32中任一项所述的方法,其中,所述Cms1多肽还包含将所述Cms1多肽靶向至多个亚细胞位置的信号肽。

[0208] 38.一种核酸分子,其包含编码Cms1多肽的多核苷酸序列,其中所述多核苷酸序列经密码子优化以在植物细胞中表达。

[0209] 39.一种核酸分子,其包含编码Cms1多肽的多核苷酸序列,其中所述多核苷酸序列经密码子优化以在真核细胞中表达。

[0210] 40.一种核酸分子,其包含编码Cms1多肽的多核苷酸序列,其中所述多核苷酸序列已经密码子优化以在原核细胞中表达,其中所述原核细胞不是所述Cms1多肽的原生宿主。

[0211] 41.如实施方式38-40中任一项所述的核酸分子,其中,所述多核苷酸序列选自下组:SEQ ID NO:16-19、24-27、70-146、174-176、212-215和255-287,或其片段或变体,或其中所述多核苷酸序列编码选自下组的Cms1多肽:SEQ ID NO:20-23、30-69、208-211和222-254,并且其中编码Cms1多肽的所述多核苷酸序列操作性地连接至启动子,所述启动子对于编码Cms1多肽的多核苷酸序列而言是异源的。

[0212] 42.如实施方式38-40中任一项所述的核酸分子,其中,所述变体多核苷酸序列与选自下组的多核苷酸序列具有至少70%的序列同一性:SEQ ID NO:16-19、24-27、70-146、174-176、212-215和255-287,或其中所述多核苷酸序列编码与选自下组的多肽具有至少80%序列同一性的Cms1多肽:SEQ ID NO:20-23、30-69、208-211和222-254,并且其中编码Cms1多肽的所述多核苷酸序列操作性地连接至启动子,所述启动子对于编码Cms1多肽的多核苷酸序列而言是异源的。

[0213] 43.如实施方式38-40中任一项所述的核酸分子,其中,所述Cms1多肽包含选自下组的氨基酸序列:SEQ ID NO:20-23、30-69、208-211和222-254,或其片段或变体。

[0214] 44.如实施方式43所述的核酸分子,其中所述变体多肽序列与选自下组的多肽序列具有至少70%的序列同一性:SEQ ID NO:20-23、30-69、208-211和222-254。

[0215] 45.如实施方式38-44中任一项所述的核酸分子,其中,编码Cms1多肽的所述多核苷酸序列操作性地连接至植物细胞中有活性的启动子。

[0216] 46.如实施方式38-44中任一项所述的核酸分子,其中,编码Cms1多肽的所述多核苷酸序列操作性地连接至真核细胞中有活性的启动子。

[0217] 47.如实施方式38-44中任一项所述的核酸分子,其中,编码Cms1多肽的所述多核苷酸序列操作性地连接至原核细胞中有活性的启动子。

[0218] 48.如实施方式38-44中任一项所述的核酸分子,其中,编码Cms1多肽的所述多核苷酸序列操作性地连接至组成型启动子、诱导型启动子、细胞类型特异性启动子或发育优

先型启动子。

[0219] 49.如实施方式38-44中任一项所述的核酸分子,其中,所述核酸分子编码包含所述Cms1多肽和效应物结构域的融合蛋白。

[0220] 50.如实施方式49所述的核酸分子,其中,所述效应物结构域选自下组:转录活化因子、转录阻遏物、核定位信号和细胞穿透信号。

[0221] 51.如实施方式50所述的核酸分子,其中,所述Cms1多肽经突变以使核酸酶活性降低或消除。

[0222] 52.如实施方式51所述的核酸分子,其中,所述突变的Cms1多肽包含突变,当经比对以实现最大同一性时,所述突变处在对应于SmCms1(SEQ ID NO:10)的701或922位的位置处,或对应于SulfCms1(SEQ ID NO:11)的848和1213位的位置处。

[0223] 53.如实施方式49-52中任一项所述的核酸分子,其中,所述Cms1多肽通过接头融合至所述效应物结构域。

[0224] 54.如实施方式38-53中任一项所述的核酸分子,其中,所述Cms1多肽形成二聚体。

[0225] 55.由实施方式49-54中任一项所述的核酸分子编码的融合蛋白。

[0226] 56.由实施方式38-44中任一项所述的核酸分子编码的Cms1多肽。

[0227] 57.经突变以减小或消除核酸酶活性的Cms1多肽。

[0228] 58.如实施方式57所述的Cms1多肽,其中,所述突变的Cms1多肽包含突变,当经比对以实现最大同一性时,所述突变处在对应于SmCms1(SEQ ID NO:10)的701或922位的位置处,或对应于SulfCms1(SEQ ID NO:11)的848和1213位的位置处。

[0229] 59.包括实施方式38-54中任一项所述的核酸分子的植物细胞、真核细胞或原核细胞。

[0230] 60.包括实施方式55-58中任一项所述的融合蛋白或多肽的植物细胞、真核细胞或原核细胞。

[0231] 61.通过实施方式1和3-37中任一项所述的方法产生的植物细胞。

[0232] 62.包括实施方式38-54中任一项所述的核酸分子的植物。

[0233] 63.包括实施方式55-58中任一项所述的融合蛋白或多肽的植物。

[0234] 64.通过实施方式1和3-37中任一项所述方法产生的植物。

[0235] 65.如实施方式62-64中任一项所述的植物的种子。

[0236] 66.如实施方式1和3-37中任一项所述的方法,其中,所述经修饰的核苷酸序列包含多核苷酸的插入,所述多核苷酸编码向转化的细胞赋予抗生素或除草剂耐受性的蛋白质。

[0237] 67.如实施方式66所述的方法,其中,编码赋予抗生素或除草剂耐受性的蛋白质的所述多核苷酸包含SEQ ID NO:7,或编码包含SEQ ID NO:8的蛋白质。

[0238] 68.如实施方式3-37中任一项所述的方法,其中植物细胞的基因组中的所述靶位点包含SEQ ID NO:12,或与SEQ ID NO:12的部分或片段具有至少80%同一性。

[0239] 69.如实施方式1-37中任一项所述的方法,其中,编码靶向DNA的RNA的所述DNA多核苷酸包含SEQ ID NO:15。

[0240] 70.如实施方式38-54中任一项所述的核酸分子,其中,编码Cms1多肽的所述多核苷酸序列还包含编码核定位信号的多核苷酸序列。

[0241] 71.如实施方式70所述的核酸分子,其中,所述核定位信号包含SEQ ID NO:1,或由SEQ ID NO:2编码。

[0242] 72.如实施方式38-54中任一项所述的核酸分子,其中,编码Cms1多肽的所述多核苷酸序列还包含编码叶绿体信号肽的多核苷酸序列。

[0243] 73.如实施方式38-54中任一项所述的核酸分子,其中,编码Cms1多肽的所述多核苷酸序列还包含编码线粒体信号肽的多核苷酸序列。

[0244] 74.如实施方式38-54中任一项所述的核酸分子,其中,编码Cms1多肽的所述多核苷酸序列还包含编码信号肽的多核苷酸序列,所述信号肽将所述Cms1多肽靶向至多个亚细胞位置。

[0245] 75.如实施方式55所述的融合蛋白,其中,所述融合蛋白还包含核定位信号,叶绿体信号肽,线粒体信号肽,或将所述Cms1多肽靶向至多个亚细胞位置的信号肽。

[0246] 76.如实施方式56-58中任一项所述的Cms1多肽,其中,所述Cms1多肽还包含核定位信号,叶绿体信号肽,线粒体信号肽,或将所述Cms1多肽靶向至多个亚细胞位置的信号肽。

[0247] 77.如实施方式1-37中任一项所述的方法,其中,所述Cms1多肽包含选自下组的一个或多个序列基序:SEQ ID NO:177-186。

[0248] 78.如实施方式1-37中任一项所述的方法,其中,所述Cms1多肽包含选自下组的一个或多个序列基序:SEQ ID NO:288-289和187-201。

[0249] 79.如实施方式1-37中任一项所述的方法,其中,所述Cms1多肽包含选自下组的一个或多个序列基序:SEQ ID NO:290-296。

[0250] 80.如实施方式38-54中任一项所述的核酸分子,其中,所述Cms1多肽包含选自下组的一个或多个序列基序:SEQ ID NO:177-186。

[0251] 81.如实施方式38-54中任一项所述的核酸分子,其中,所述Cms1多肽包含选自下组的一个或多个序列基序:SEQ ID NO:288-289和187-201。

[0252] 82.如实施方式38-54中任一项所述的核酸分子,其中,所述Cms1多肽包含选自下组的一个或多个序列基序:SEQ ID NO:290-296。

[0253] 通过说明的方式,而非限制性方式提供以下实施例。

[0254] 实验部分

[0255] 实施例1-克隆植物转化构建体

[0256] 含Cms1的构建体汇总于表1。简言之,Cms1基因经植物密码子优化,通过GenScript(新泽西州的皮斯卡塔韦(Piscataway))从头合成,并且通过PCR扩增以向感兴趣的Cms1编码序列添加N-末端SV40核定位标签(SEQ ID NO:2)(框内),以及限制性酶位点,用于克隆。利用合适的限制性酶位点,将各个个体Cms1基因克隆至2x35s启动子(SEQ ID NO:3)的下游。注意到,编码ADurb.160Cms1蛋白(SEQ ID NO:20)的SEQ ID NO:16源自一种生物体,该生物体似乎使用TGA密码子编码甘氨酸,而不是大多数生物体使用的通用遗传编码中的终止密码子。因此,编码ADurb.160Cms1蛋白的原生基因(SEQ ID NO:24)包括(其似乎是)多个过早终止密码子;然而,用编码甘氨酸的TGA对该基因进行的分析发现了全长的开放阅读框。类似地,SEQ ID NO:82、91、92、100、105、213、255、259、266、267、268、270、271、272、273、275、276、277、279、280、284、285和286也似乎使用非通用遗传编码,具有编码甘氨酸的TGA

密码子。

[0257] 合成编码引导RNA的质粒,采用在其5'端侧接稻U6(OsU6)启动子(SEQ ID NO:5)且在其3'端侧接OsU6终止子(SEQ ID NO:6)的引导RNA,其靶向至稻(水稻(Oryza sativa cv.Kitaake))CAO1基因(SEQ ID NO:12)的区域。引导RNA具有SEQ ID NO:15的序列。引导RNA质粒汇总于表2。

[0258] 包含修复供体盒(SEQ ID NO:13)的质粒131632经设计以在OsCAO1基因中靶位点上游和下游具有约1,000个碱基对同源性。修复供体盒包括玉米遍在蛋白启动子(SEQ ID NO:9),该启动子与潮霉素抗性基因(SEQ ID NO:7,编码SEQ ID NO:8)操作性连接,其在其3'端侧接花椰菜花叶病毒35S多聚A序列(SEQ ID NO:4)。质粒131592的设计与质粒131632类似,但在潮霉素盒的上游或下游没有任何同源臂。这样,质粒131592含有来自SEQ ID NO:13的核苷酸1,001-4,302,包括玉米遍在蛋白启动子(SEQ ID NO:9),其与潮霉素抗性基因(SEQ ID NO:7,编码SEQ ID NO:8)操作性地连接,其在其3'端侧接花椰菜花叶病毒35S多聚A序列(SEQ ID NO:4)。

[0259] 表1:Cms1载体

[0260]

构建体编号	启动子	Cms1 基因[1]	终止子
132363	2X 35S (SEQ ID NO: 3)	ADurb.160Cms1 (SEQ ID NO:16，编码 SEQ ID NO:20)	35S 多聚 A (SEQ ID NO: 4)
132388	2X 35S (SEQ ID NO: 3)	AuxCms1 (SEQ ID NO:17，编码 SEQ ID NO:21)	35S 多聚 A (SEQ ID NO: 4)
132389	2X 35S (SEQ ID NO: 3)	LAHSCms1 (SEQ ID NO:18，编码 SEQ ID NO:22)	35S 多聚 A (SEQ ID NO: 4)
132390	2X 35S (SEQ ID NO: 3)	Sm82Cms1 (SEQ ID NO:19，编码 SEQ ID NO:23)	35S 多聚 A (SEQ ID NO: 4)
132437	2X 35S (SEQ ID NO: 3)	Unk1Cms1 (SEQ ID NO:110，编码 SEQ ID NO:30)	35S 多聚 A (SEQ ID NO: 4)
132438	2X 35S (SEQ ID NO: 3)	Unk2Cms1 (SEQ ID NO:111，编码 SEQ ID NO:31)	35S 多聚 A (SEQ ID NO: 4)
132439	2X 35S (SEQ ID NO: 3)	Unk3Cms1 (SEQ ID NO:112，编码 SEQ ID NO:32)	35S 多聚 A (SEQ ID NO: 4)
132455	2X 35S (SEQ ID NO: 3)	Unk4Cms1 (SEQ ID NO:113，编码 SEQ ID NO:33)	35S 多聚 A (SEQ ID NO: 4)
132463	2X 35S (SEQ ID NO: 3)	Unk5Cms1 (SEQ ID NO:114，编码 SEQ ID NO:34)	35S 多聚 A (SEQ ID NO: 4)
132470	2X 35S (SEQ ID NO: 3)	Unk6Cms1 (SEQ ID NO:115，编码 SEQ ID NO:35)	35S 多聚 A (SEQ ID NO: 4)
132456	2X 35S (SEQ ID NO: 3)	Unk7Cms1 (SEQ ID NO:116，编码 SEQ ID NO:36)	35S 多聚 A (SEQ ID NO: 4)
132464	2X 35S (SEQ ID NO: 3)	Unk8Cms1 (SEQ ID NO:117，编码 SEQ ID NO:37)	35S 多聚 A (SEQ ID NO: 4)
132465	2X 35S (SEQ ID NO: 3)	Unk9Cms1 (SEQ ID NO:118，编码 SEQ ID NO:38)	35S 多聚 A (SEQ ID NO: 4)
132457	2X 35S (SEQ ID NO: 3)	Unk10Cms1 (SEQ ID NO:119，编码 SEQ ID NO:39)	35S 多聚 A (SEQ ID NO: 4)
132466	2X 35S (SEQ ID NO: 3)	Unk11Cms1 (SEQ ID NO:120，编码 SEQ ID NO:40)	35S 多聚 A (SEQ ID NO: 4)
132502	2X 35S (SEQ ID NO: 3)	Unk4Cms1 (SEQ ID NO:221，编码 SEQ ID NO:33)	35S 多聚 A (SEQ ID NO: 4)
132504	2X 35S (SEQ ID NO: 3)	Unk14Cms1 (SEQ ID NO:122，编码 SEQ ID NO:42)	35S 多聚 A (SEQ ID NO: 4)
132505	2X 35S (SEQ ID NO: 3)	Unk15Cms1 (SEQ ID NO:123，编码 SEQ ID NO:43)	35S 多聚 A (SEQ ID NO: 4)
132506	2X 35S (SEQ ID NO: 3)	Unk16Cms1 (SEQ ID NO:124，编码 SEQ ID NO:44)	35S 多聚 A (SEQ ID NO: 4)
132507	2X 35S (SEQ ID NO: 3)	Unk17Cms1 (SEQ ID NO:125，编码 SEQ ID NO:45)	35S 多聚 A (SEQ ID NO: 4)
132508	2X 35S (SEQ ID NO: 3)	Unk18Cms1 (SEQ ID NO:126，编码 SEQ ID NO:46)	35S 多聚 A (SEQ ID NO: 4)

[0261]

132509	2X 35S (SEQ ID NO: 3)	Unk19Cms1 (SEQ ID NO:127，编码 SEQ ID NO:47)	35S 多聚 A (SEQ ID NO: 4)
132510	2X 35S (SEQ ID NO: 3)	Unk20Cms1 (SEQ ID NO:128，编码 SEQ ID NO:48)	35S 多聚 A (SEQ ID NO: 4)
132511	2X 35S (SEQ ID NO: 3)	Unk21Cms1 (SEQ ID NO:129，编码 SEQ ID NO:49)	35S 多聚 A (SEQ ID NO: 4)
132512	2X 35S (SEQ ID NO: 3)	Unk22Cms1 (SEQ ID NO:130，编码 SEQ ID NO:50)	35S 多聚 A (SEQ ID NO: 4)
132513	2X 35S (SEQ ID NO: 3)	Unk23Cms1 (SEQ ID NO:131，编码 SEQ ID NO:51)	35S 多聚 A (SEQ ID NO: 4)
132514	2X 35S (SEQ ID NO: 3)	Unk24Cms1 (SEQ ID NO:132，编码 SEQ ID NO:52)	35S 多聚 A (SEQ ID NO: 4)
132515	2X 35S (SEQ ID NO: 3)	Unk25Cms1 (SEQ ID NO:133，编码 SEQ ID NO:53)	35S 多聚 A (SEQ ID NO: 4)
132516	2X 35S (SEQ ID NO: 3)	Unk26Cms1 (SEQ ID NO:134，编码 SEQ ID NO:54)	35S 多聚 A (SEQ ID NO: 4)
132517	2X 35S (SEQ ID NO: 3)	Unk27Cms1 (SEQ ID NO:135，编码 SEQ ID NO:55)	35S 多聚 A (SEQ ID NO: 4)
132518	2X 35S (SEQ ID NO: 3)	Unk28Cms1 (SEQ ID NO:136，编码 SEQ ID NO:56)	35S 多聚 A (SEQ ID NO: 4)
132519	2X 35S (SEQ ID NO: 3)	Unk29Cms1 (SEQ ID NO:137，编码 SEQ ID NO:57)	35S 多聚 A (SEQ ID NO: 4)
132520	2X 35S (SEQ ID NO: 3)	Unk30Cms1 (SEQ ID NO:138，编码 SEQ ID NO:58)	35S 多聚 A (SEQ ID NO: 4)
132521	2X 35S (SEQ ID NO: 3)	Unk31Cms1 (SEQ ID NO:139，编码 SEQ ID NO:59)	35S 多聚 A (SEQ ID NO: 4)
132522	2X 35S (SEQ ID NO: 3)	Unk32Cms1 (SEQ ID NO:140，编码 SEQ ID NO:60)	35S 多聚 A (SEQ ID NO: 4)
132523	2X 35S (SEQ ID NO: 3)	Unk33Cms1 (SEQ ID NO:141，编码 SEQ ID NO:61)	35S 多聚 A (SEQ ID NO: 4)
132524	2X 35S (SEQ ID NO: 3)	Unk34Cms1 (SEQ ID NO:142，编码 SEQ ID NO:62)	35S 多聚 A (SEQ ID NO: 4)
132525	2X 35S (SEQ ID NO: 3)	Unk35Cms1 (SEQ ID NO:143，编码 SEQ ID NO:63)	35S 多聚 A (SEQ ID NO: 4)
132526	2X 35S (SEQ ID NO: 3)	Unk36Cms1 (SEQ ID NO:144，编码 SEQ ID NO:64)	35S 多聚 A (SEQ ID NO: 4)
132527	2X 35S (SEQ ID NO: 3)	Unk37Cms1 (SEQ ID NO:145，编码 SEQ ID NO:65)	35S 多聚 A (SEQ ID NO: 4)
132528	2X 35S (SEQ ID NO: 3)	Unk38Cms1 (SEQ ID NO:146，编码 SEQ ID NO:66)	35S 多聚 A (SEQ ID NO: 4)
132529	2X 35S (SEQ ID NO: 3)	Unk39Cms1 (SEQ ID NO:174，编码 SEQ ID NO:67)	35S 多聚 A (SEQ ID NO: 4)

[0262]

132530	2X 35S (SEQ ID NO: 3)	Unk40Cms1 (SEQ ID NO:175，编码 SEQ ID NO:68)	35S 多聚 A (SEQ ID NO: 4)
132531	2X 35S (SEQ ID NO: 3)	Unk41Cms1 (SEQ ID NO:176，编码 SEQ ID NO:69)	35S 多聚 A (SEQ ID NO: 4)

[0263] [1]-各Cms1基因与SV40核定位信号(SEQ ID NO:2,编码氨基酸序列SEQ ID NO:1)在其5'端框内融合。

[0264] 表2:引导RNA载体

[0265]

构建体编号	启动子	gRNA 序列	终止子
131608	OsU6 (SEQ ID NO:5)	AATTTCTACTGTTGTAGATTGGAGCAACACCTGAAG GAAGGCT (SEQ ID NO:15)	OsU6 (SEQ ID NO: 6)

[0266] 实施例2-水稻转化

[0267] 为了将Cms1盒、含gRNA的质粒和修复供体盒引入水稻细胞,使用了颗粒轰击。对于轰击,称取2mg的0.6μm金颗粒,并将其转移至无菌1.5-mL试管。添加500mL的100%乙醇,然后管用超声处理10-15秒。离心后,移除乙醇。然后将1毫升灭菌双蒸水加入含有金珠的试管。将珠沉淀短暂涡旋,然后通过离心重整(re-formed),然后从管中除去水。在无菌层流罩中,将DNA被覆到珠子上。表3示出添加到珠上的DNA的量。将含有Cms1盒的质粒、含有gRNA的质粒和修复供体盒添加到珠,并且添加灭菌双蒸水以使总体积达到50μL。为此,添加20μL的亚精胺(1M),然后是50μL的$CaCl_2$(2.5M)。通过重力使金颗粒沉淀几分钟,然后通过离心将其沉淀。移除上清液液体,并且添加800μL的100%乙醇。短暂超声处理后,使金颗粒通过重力沉淀3-5分钟,然后将试管离心以形成沉淀。移除上清液,并且向试管添加30μL的100%乙醇。DNA涂覆的金颗粒通过涡旋重悬于该乙醇中,并将10μL重悬的金颗粒各自添加到3个大型运载体(加利福尼亚州州赫尔克里斯的生物辐射公司(Bio-Rad))。允许大型运载体在层流罩中风干5-10分钟以允许乙醇蒸发。

[0268] 表3:用于颗粒轰击实验的DNA的量(所有量为每2mg金颗粒)

[0269]

Cms1 质粒	1.5 μg
含 gRNA 的质粒	1.5 μg

[0270]

修复供体盒质粒	3-15 μg
无菌、双蒸水	添加至总体积 50 μL

[0271] 水稻愈伤组织用于轰击。在轰击之前,将水稻愈伤组织在愈伤组织诱导培养基(CIM;3.99g/L N6盐和维生素,0.3g/L酪蛋白水解物,30g/L蔗糖,2.8g/L L-脯氨酶,2mg/L 2,4-D,8g/L琼脂,调整至pH 5.8)以28℃在黑暗中维持4-7天。颗粒轰击之前,将各自大小为0.2-0.3cm且重量为总计1-1.5g的大约80-100个愈伤组织块置于含有渗透固体培养基(补充有0.4M山梨醇和0.4M甘露醇的CIM)的皮氏培养皿中心进行4小时渗透预处理。对于轰击,含有涂覆DNA的金颗粒的大型运载体被组装成大型运载体支持物(holder)。按照生产商的说明组装防爆片(1,100psi)、停止屏(stopping screen)和大型运载体支持物。将含有待轰击的水稻愈伤组织的平板置于停止屏下6cm,并且在真空室达到25-28汞柱后轰击愈伤组织块。轰击后,将愈伤组织置于渗透培养基16-20小时,然后将愈伤组织块转移至选择培养基(补充有50mg/L潮霉素和100mg/L特美汀的CIM)。将平板转移至孵育器,并维持在28筷黑暗中以开始转化细胞的恢复。每两周,将愈伤组织在新鲜选择培养基上继代培养。在大约5-6

周后的选择培养基上出现潮霉素抗性愈伤组织块。将个别潮霉素抗性愈伤组织块转移至新的选择平板以允许细胞分裂并且生长,从而生成足够多的组织用于分子分析取样。表4总结了用于这些水稻轰击实验的DNA载体的组合。

[0272] 表4:水稻粒子轰击实验总结

[0273]

实验	Cms1质粒	gRNA质粒	修复供体质粒
166	132363	131608	131632
187	132388	131608	131632
188	132389	131608	131632
189	132390	131608	131632
201	132437	131608	131632
202	132438	131608	131632
211	132439	131608	131632
212	132455	131608	131632
217	132456	131608	131632
218	132457	131608	131632

[0274]

220	132463	131608	131632
221	132464	131608	131632
222	132465	131608	131632
223	132466	131608	131632
224	132470	131608	131632
231	132502	131608	131632
233	132504	131608	131632
234	132505	131608	131632
238	132506	131608	131632
239	132507	131608	131632
240	132508	131608	131632
241	132509	131608	131632
247	132510	131608	131632
248	132511	131608	131632
249	132512	131608	131632
251	132513	131608	131632
252	132514	131608	131632
253	132515	131608	131632
254	132516	131608	131632
255	132517	131608	131632
256	132518	131608	131632
257	132519	131608	131632
258	132520	131608	131632
259	132521	131608	131632
260	132522	131608	131632
261	132523	131608	131632
262	132524	131608	131632
264	132525	131608	131632
265	132526	131608	131632
266	132527	131608	131632
270	132522	131608	131592
271	132523	131608	131592
272	132524	131608	131592
273	132525	131608	131592
278	132526	131608	131592
279	132527	131608	131592

[0275]

280	132528	131608	131592
283	132456	131608	131592
284	132463	131608	131592
293	132529	131608	131592
294	132530	131608	131592
295	132531	131608	131592
300	132464	131608	131592

[0276] 实施例3-水稻分子分析

[0277] 在将来自各转化实验的潮霉素抗性愈伤组织块转移至新平板后,他们生长至足够进行采样的大小。从每一片抗潮霉素的水稻愈伤组织中收获少量组织,并从这些组织样品中提取DNA用于PCR,DNA测序和T7内切核酸酶(T7EI)分析。PCR分析使用引物设计,这些引物既不会从野生型水稻DNA产生扩增子也不会单从修复供体质粒产生扩增子,而是在同源臂之外的水稻基因组中具有一个引物结合位点,而在插入盒中具有另一个引物结合位点,因此指示在水稻CAO1基因座处的插入事件。

[0278] 对由上述PCR分析产生的PCR扩增子进行桑格测序和/或下一代测序,以确认PCR扩增子实际指示在预期基因组位点的插入,而不仅仅是实验假象。表5汇总了这些测序分析的结果。

[0279] 表5-水稻愈伤组织基因组编辑实验结果汇总

[0280]

核酸酶	实验数	CAO1 基因组编辑
ADurb.160Cms1 (SEQ ID NO:16，编码 SEQ ID NO:20)	166	-186/+90 (SEQ ID NO:14)
AuxCms1 (SEQ ID NO:17，编码 SEQ ID NO:21)	187	-344/+104 (SEQ ID NO:28)
LAHSCms1 (SEQ ID NO:18，编码 SEQ ID NO:22)	188	-431 (SEQ ID NO:29)
Unk1Cms1 (SEQ ID NO:110，编码 SEQ ID NO:30)	201	-431 (SEQ ID NO:202)
Unk2Cms1 (SEQ ID NO:111，编码 SEQ ID NO:31)	202	-314/+116 (SEQ ID NO:203)
Unk3Cms1 (SEQ ID NO:112，编码 SEQ ID NO:32)	211	-63 (SEQ ID NO:204)
Unk3Cms1 (SEQ ID NO:112，编码 SEQ ID NO:32)	211	-42 (SEQ ID NO:205)
Unk4Cms1 (SEQ ID NO:113，编码 SEQ ID NO:33)	212	-22 (SEQ ID NO:13)
Unk7Cms1 (SEQ ID NO:116，编码 SEQ ID NO:36)	217	-26 (SEQ ID NO:318)
Unk10Cms1 (SEQ ID NO:119，编码 SEQ ID NO:39)	218	-38/+257 (SEQ ID

[0281]

		NO:214)
Unk5Cms1 (SEQ ID NO:114，编码 SEQ ID NO:34)	220	-4 (SEQ ID NO:319)
Unk8Cms1 (SEQ ID NO:117，编码 SEQ ID NO:37)	221	-22 (SEQ ID NO:320)
Unk9Cms1 (SEQ ID NO:118，编码 SEQ ID NO:38)	222	-244 (SEQ ID NO:208)
Unk11Cms1 (SEQ ID NO:120，编码 SEQ ID NO:40)	223	-216 (SEQ ID NO:209)
Unk6Cms1 (SEQ ID NO:115，编码 SEQ ID NO:35)	224	-216 (SEQ ID NO:210)
Unk4Cms1 (SEQ ID NO:221，编码 SEQ ID NO:33)	231	-24 (SEQ ID NO:211)
Unk14Cms1 (SEQ ID NO:122，编码 SEQ ID NO:42)	233	-293 (SEQ ID NO:207)
Unk15Cms1 (SEQ ID NO:123，编码 SEQ ID NO:43)	234	-124 (SEQ ID NO:321)
Unk16Cms1 (SEQ ID NO:124，编码 SEQ ID NO:44)	238	-8 (SEQ ID NO:322)
Unk17Cms1 (SEQ ID NO:125，编码 SEQ ID NO:45)	239	-392/+349 (SEQ ID NO:213)
Unk18Cms1 (SEQ ID NO:126，编码 SEQ ID NO:46)	240	-16 (SEQ ID NO:323)
Unk19Cms1 (SEQ ID NO:127，编码 SEQ ID NO:47)	241	-397/+356 (SEQ ID NO:215)
Unk20Cms1 (SEQ ID NO:128，编码 SEQ ID NO:48)	247	-26 (SEQ ID NO:324)
Unk21Cms1 (SEQ ID NO:129，编码 SEQ ID NO:49)	248	-305/+402 (SEQ ID NO:216)
Unk22Cms1 (SEQ ID NO:130，编码 SEQ ID NO:50)	249	-26 (SEQ ID NO:324)
Unk23Cms1 (SEQ ID NO:131，编码 SEQ ID NO:51)	251	-26 (SEQ ID NO:324)
Unk24Cms1 (SEQ ID NO:132，编码 SEQ ID NO:52)	252	-364/+95 (SEQ ID NO:217)
Unk25Cms1 (SEQ ID NO:133，编码 SEQ ID NO:53)	253	-304 (SEQ ID NO:219)
Unk27Cms1 (SEQ ID NO:135，编码 SEQ ID NO:55)	255	-284/+1 (SEQ ID NO:220)
Unk28Cms1 (SEQ ID NO:136，编码 SEQ ID NO:56)	256	-470/+238 (SEQ ID NO:218)
Unk29Cms1 (SEQ ID NO:137，编码 SEQ ID NO:57)	257	-26 (SEQ ID NO:324)
Unk30Cms1 (SEQ ID NO:138，编码 SEQ ID NO:58)	258	-26 (SEQ ID NO:324)
Unk31Cms1 (SEQ ID NO:139，编码 SEQ ID NO:59)	259	-4 (SEQ ID NO:319)
Unk32Cms1 (SEQ ID NO:140，编码 SEQ ID NO:60)	270	-26 (SEQ ID NO:324)
Unk33Cms1 (SEQ ID NO:141，编码 SEQ ID NO:61)	271	-26 (SEQ ID NO:324)
Unk34Cms1 (SEQ ID NO:142，编码 SEQ ID NO:62)	272	-26 (SEQ ID NO:324)
Unk35Cms1 (SEQ ID NO:143，编码 SEQ ID NO:63)	273	-26 (SEQ ID NO:324)
Unk36Cms1 (SEQ ID NO:144，编码 SEQ ID NO:64)	278	-26 (SEQ ID NO:324)
Unk37Cms1 (SEQ ID NO:145，编码 SEQ ID NO:65)	279	-16 (SEQ ID NO:323)

[0282]

Unk38Cms1 (SEQ ID NO:146，编码 SEQ ID NO:66)	280	-29 (SEQ ID NO:325)

[0283] 除PCR和DNA测序分析外,还进行了T7EI分析,以检测CAO1基因座上是否存在小插

入和/或缺失。如前所述进行T7EI分析(Begemann等.(2017)Sci Reports 7:11606)。对于T7EI分析指示潜在插入或缺失的愈伤组织样品,进行DNA测序分析以检测CAO1基因座是否存在插入和/或缺失。

[0284] 实施例4-在CAO1基因座处有遗传修饰的水稻植株的再生

[0285] 将如上所述转化的水稻愈伤组织在组织培养基上培养以产生芽。随后将这些芽转移至生根培养基,然后将生根的植物转移至土壤以在温室里栽培。从生根植物提取DNA进行PCR和DNA测序分析。使T0代植物生长到成熟并自花授粉以产生T1代种子。种植这些T1代种子,并对所得的T1代植物进行基因分型,以鉴定纯合子,半合子和无效分离子植物。将植物表型分型以检测与CAO1基因的纯合敲除相关的黄叶表型(Lee等.(2005)Plant Mol Biol 57:805-818)。

[0286] 实施例5-编辑玉米(Zea mays)中的预定基因组基因座

[0287] 设计一种或多种gRNA以在玉米基因组中的所需位点退火,并且允许与一个或多个Cms1蛋白的相互作用。将这些gRNA克隆至载体,从而使其操作性地连接在植物细胞中可操作的启动子(“gRNA盒”)。将编码Cms1蛋白的一种或多种基因克隆到载体,从而使其操作性地连接在植物细胞中可操作的启动子(“Cms1盒”)。将gRNA盒和Cms1盒克隆到单个载体中,或者克隆到适合植物转化的两个独立载体中,然后将该载体或这些载体转化至农杆菌(Agrobacterium)细胞中。将这些细胞与适合转化的玉米组织接触。在与农杆菌细胞孵育后,在适合再生完整植物的组织培养基上培养玉米细胞。玉米植物由与农杆菌细胞接触的细胞再生,所述农杆菌细胞具有包含Cms1盒和gRNA盒的载体。在玉米植物再生后,收获植物组织并且从组织提取DNA。酌情进行T7EI试验、PCR试验和/或测序试验,以确定DNA序列中的改变是否发生在所需基因组位置。

[0288] 或者,使用颗粒轰击将Cms1盒和gRNA盒引入玉米细胞。将包含Cms1盒和gRNA盒的单个载体,或分别包含Cms1盒和gRNA盒的独立载体被覆在金珠或钛珠上,然后用它们轰击适合再生的玉米组织。轰击后,将玉米组织转移至用于再生玉米植物的组织培养基。在玉米植物再生后,收获植物组织并且从组织提取DNA。酌情进行T7EI试验、PCR试验和/或测序试验,以确定DNA序列中的改变是否发生在所需基因组位置。

[0289] 实施例6-Cms1核酸酶和其它V型核酸酶的计算分析

[0290] CRISPR核酸酶通常按类型分类,例如被分类为II型核酸酶的Cas9核酸酶和被分类为V型的Cpf1核酸酶(Koonin等.(2017)Curr Opin Microbiol 37:67-78)。对于Cms1核酸酶蛋白序列的研究表明,部分基于RuvC结构域的存在和HNH结构域的缺失,应将这些核酸酶归为V型核酸酶。迄今,科学文献中已经描述了多组V型核酸酶,包括Cpf1(也称为VA型),C2c1(也称为VB型),C2c3(也称为VC型),CasY(也称为VD型)和CasX(也称为VE型)。

[0291] V型氨基酸序列的MUSCLE比对通常无法正确比对这些蛋白质中RuvCI,RuvCII和RuvCIII域的催化残基。鉴于这些结构域在蛋白质功能中的核心重要性,必须对这些残基进行正确的比对。针对本文和美国专利号9,896,696中公开的Cms1核酸酶(SEQ ID NO:10、11、20-23、30-69和154-156),三种Cpf1核酸酶(SEQ ID NO:147-149),C2c1核酸酶(SEQ ID NO:150和157-164),C2c3核酸酶(SEQ ID NO:152和166-168)(Shmakov等.(2016)Mol Cell 60:385-397),CasX核酸酶(SEQ ID NO:151和165)和CasY核酸酶(SEQ ID NO:153和169-173)(Burstein等(2017)Nature 542:237-241)的氨基酸序列鉴定RuvCI,RuvCII和RuvCIII催化

残基。表6显示各结构域的催化残基以及紧接催化残基之前的三个氨基酸和紧接催化残基之后的三个氨基酸。

[0292] 表6:V型核酸酶的RuvCI、RuvCII和RuvCIII催化残基的汇总

[0293]

蛋白质	RuvCI	RuvCII	RuvCIII
MicroCms1 (SEQ ID NO:154)	YGIDRGL	IALENLD	HNSDDVA
ObCms1 (SEQ ID NO:155)	FGIDRGN	VALENLA	NSPDTVA
Sm17Cms1 (SEQ ID NO:156)	YGIDAGE	ISIEDLK	DSNDKVA
SmCms1 (SEQ ID NO:10)	YGIDAGE	ISIEDLK	NDPDKVA
ADurb.160Cms1 (SEQ ID NO:20)	YGIDKGT	ICYETLN	ESGDDLA
Sm82Cms1 (SEQ ID NO:23)	FGIDVGN	IVLEYLT	DGPDKVA

[0294]

Unk1Cms1 (SEQ ID NO:30)	YGIDRGL	IALENLD	NNSDEVA
Unk3Cms1 (SEQ ID NO:32)	YGLDRGQ	IVFEGLD	DNSDSVA
Unk4Cms1 (SEQ ID NO:33)	FGVDTGE	IAIENLA	HSNDAVA
Unk5Cms1 (SEQ ID NO:34)	YGLDRGE	ISLENLE	NSSDDIA
Unk8Cms1 (SEQ ID NO:37)	YGIDRGQ	INLENLI	KNSDEVA
Unk9Cms1 (SEQ ID NO:38)	YGIDRGN	VVLEDLN	NDPDKIA
Unk10Cms1 (SEQ ID NO:39)	FGIDVGT	VVLENLK	DTNDKIA
Unk15Cms1 (SEQ ID NO:43)	LGIDNGE	IIKEGFD	HSNDGIA
Unk16Cms1 (SEQ ID NO:44)	YGIDRGQ	INLENLH	KNSDDVA
Unk18Cms1 (SEQ ID NO:46)	YGIDRGL	IALENLD	HNSDDVA
Unk19Cms1 (SEQ ID NO:47)	YGIDAGE	ISIEDLK	DSNDKVA
Unk20Cms1 (SEQ ID NO:48)	YGIDRGL	IAFEDMD	DDSDKVA
Unk21Cms1 (SEQ ID NO:49)	LGIDNNE	IVKEGFD	HSNDGIA
Unk22Cms1 (SEQ ID NO:50)	YGIDRGQ	ITLEDLD	KNSDDVA
Unk23Cms1 (SEQ ID NO:51)	YGIDRGE	IYFEEKN	NSGDDLA
Unk24Cms1 (SEQ ID NO:52)	YGLDKGT	ICFETLD	KSGDDLA
Unk25Cms1 (SEQ ID NO:53)	LGIDNGE	VVKEGFG	HSNDGIA
Unk26Cms1 (SEQ ID NO:54)	FGIDNGE	IIKEGFD	HSNDGIA
Unk27Cms1 (SEQ ID NO:55)	FGIDNGE	IVKEGFG	HSNDEIA
Unk28Cms1 (SEQ ID NO:56)	CGIDIGE	VVLENIP	KSCDIVA

[0295]

Unk29Cms1 (SEQ ID NO:57)	FGIDSGE	IAKEGFD	HSNDGVA
Unk30Cms1 (SEQ ID NO:58)	FGIDNGE	IVKEGFD	HSNDGIA
Unk31Cms1 (SEQ ID NO:59)	LGIDNGE	VVKEAFD	HRNDGIA
Unk32Cms1 (SEQ ID NO:60)	YGIDRGD	MFLENKK	KSGDDLA
Unk39Cms1 (SEQ ID NO:67)	FGIDNGE	IAKEGFG	HSNDGIA
Unk42Cms1 (SEQ ID NO:208)	LGIDNGE	IVKEGFD	HSNDGVA
Unk43Cms1 (SEQ ID NO:209)	YGLDKGT	IVREGLG	KSGDDLA
Unk44Cms1 (SEQ ID NO:210)	IGIDTGT	IAFEGFD	DCNDKVA
Unk45Cms1 (SEQ ID NO:211)	FGIDRGN	INLENLH	DNSDSVA
Unk46Cms1 (SEQ ID NO:222)	YFIDIWE	IIISNFI	不清楚
Unk47Cms1 (SEQ ID NO:223)	FGIDNGE	IIKEGFG	HSNDGIA
Unk49Cms1 (SEQ ID NO:225)	YGIDRGD	INLENLH	KNSDDVA
Unk52Cms1 (SEQ ID NO:228)	YGIDRGS	VVLENLK	SDPDKIA
Unk54Cms1 (SEQ ID NO:229)	FGLDNGE	IVKEGFD	HSNDGIA
LbCms1Cms1 (SEQ ID NO:232)	YGIDVGQ	IFLEDLK	DNPDSLA
Unk58Cms1 (SEQ ID NO:234)	YGIDRGI	IYLENLE	INYDSIA
Unk60Cms1 (SEQ ID NO:236)	YGLDRGK	MCFENLN	DNSDSVA
Unk61Cms1 (SEQ ID NO:237)	YWIDKWT	ICYETLD	KSWDDLA
Unk65Cms1 (SEQ ID NO:241)	YGIDTGI	ITIEYLD	DSNDKVA
Unk67Cms1 (SEQ ID NO:243)	YWIDKWD	MFLENKK	KSWDDLA

[0296]

Unk69Cms1 (SEQ ID NO:245)	LGIDNGE	IVKEGFD	HSNNGVA
Unk72Cms1 (SEQ ID NO:248)	YGIDRGQ	INLENLT	KNSDEVA
Unk74Cms1 (SEQ ID NO:250)	FGIDTGE	IAIENLA	HSNDAVA
Unk75Cms1 (SEQ ID NO:251)	YWFDKWE	FVFEDKT	HSWDDLA
Unk77Cms1 (SEQ ID NO:253)	YGIDRGI	IFLENLD	LNYDSIA
Unk78Cms1 (SEQ ID NO:254)	YGIDRGE	IILEDIE	DDPDKVA
Unk79Cms1 (SEQ ID NO:41)	YGLDRGK	VAFENLD	DNSDKVA
SulfCms1 (SEQ ID NO:11)	IGIDRGL	ISLEDLS	HNGDDNG
AuxCms1 (SEQ ID NO:21)	IGIDRGQ	ISLEDLS	KSGDDNA
LAHSCms1 (SEQ ID NO:22)	FGIDRGQ	ISLEDLS	KSGDDNA
Unk2Cms1 (SEQ ID NO:31)	FGIDRGQ	ISLEDLS	KSGDDNA
Unk6Cms1 (SEQ ID NO:35)	FGIDRGQ	ISLEDLT	KSGDDNA
Unk7Cms1 (SEQ ID NO:36)	IGIDRGL	ISIENLT	SNGDENG
Unk11Cms1 (SEQ ID NO:40)	IGIDRGI	IALEDLT	TDGDQNG
Unk14Cms1 (SEQ ID NO:42)	FGIDRGI	ISLENLS	KNGDDNA
Unk17Cms1 (SEQ ID NO:45)	FGIDRGL	ISLEDLT	QNGDENG
Unk33Cms1 (SEQ ID NO:61)	IGIDRGI	IALEDLT	TDGDQNG
Unk34Cms1 (SEQ ID NO:62)	FGIDRGQ	IALEDLT	KSGDDNA
Unk35Cms1 (SEQ ID NO:63)	IGIDRGL	VSLEDLS	HNGDDNG
Unk36Cms1 (SEQ ID NO:64)	FGIDRG	ISLEDLS	KSGDDN

[0297]

	Q		A
Unk37Cms1 (SEQ ID NO:65)	FGIDRGI	ITLENLN	KNGDDNA
Unk38Cms1 (SEQ ID NO:66)	IGIDRGL	VSLEDLS	HNGDDNG
Unk41Cms1 (SEQ ID NO:69)	YGIDRGI	IVLENIA	RSGDQSA
Unk51Cms1 (SEQ ID NO:227)	FGIDRGQ	IALEDLT	KNGDDNA
Unk55Cms1 (SEQ ID NO:230)	FGIDRGI	ISFEDLT	TNGDDNG
Unk56Cms1 (SEQ ID NO:231)	IGIDRGI	IALEDLT	TDGDQNG
Unk59Cms1 (SEQ ID NO:235)	FGIDSWI	ISLEDLS	KNWDDNG
Unk63Cms1 (SEQ ID NO:239)	FGIDSWI	ISLENLS	KNGDDNA
Unk64Cms1 (SEQ ID NO:240)	FGIDSWI	ISLENLS	NNYKKQC
Unk66Cms1 (SEQ ID NO:242)	FGIDSWI	ISLEDLS	KNWDDNG
Unk68Cms1 (SEQ ID NO:244)	FGIDSWI	ISLEDLS	KNGDDNG
Unk71Cms1 (SEQ ID NO:247)	FGIDSWI	IVLENLS	KNWDDNG
Unk40Cms1 (SEQ ID NO:68)	VGLDRGE	VSLENLN	NGGDVLA
Unk48Cms1 (SEQ ID NO:224)	IGLDRGE	VSLENLN	TGGDTLA
Unk50Cms1 (SEQ ID NO:226)	VGIDLGE	IVFENLD	KSCDEIA
Unk57Cms1 (SEQ ID NO:233)	IGLDRGE	VSFENLN	NGGDVLA
Unk62Cms1 (SEQ ID NO:238)	IGIDLGE	IVFENLD	KSCDEIA
Unk70Cms1 (SEQ ID NO:246)	IGIDLWE	IVFENLD	KSCDEIA
Unk73Cms1 (SEQ ID NO:249)	LGMDRGE	IVLEDLD	KTGDDLA

[0298]

Unk76Cms1 (SEQ ID NO:252)	IGLDRGE	FIFENQT	KSGDNLA
AsCpf1 (SEQ ID NO:148)	IGIDRGE	VVLENLN	MDADANG
FnCpf1 (SEQ ID NO:147)	LSIDRGE	VVFEDLN	QDADANG
LbCpf1 (SEQ ID NO:149)	IGIDRGE	IALEDLN	KNADANG
CasY.1 (SEQ ID NO:153)	LGLDVGE	IIYEISI	TDADIQA
CasY.2 (SEQ ID NO:172)	MGIDIGE	PVYEFEI	SDADIQA
CasY.3 (SEQ ID NO:173)	IGIDIGE	LSFEYEV	SHADKQA
CasY.4 (SEQ ID NO:169)	LGIDIGE	IVYELEV	ADADIQA
CasY.5 (SEQ ID NO:171)	AVVDVLD	AANELHR	不清楚
CasY.6 (SEQ ID NO:170)	LGLDAGE	VVHEESV	不清楚
CasX_δ(SEQ ID NO:151)	IGVDRGE	LVFENLS	VHADEQA
CasX_Plancto (SEQ ID NO:165)	IGIDRGE	LIFENLS	THADEQA
C2c3_AUXO (SEQ ID NO:152)	VSIDQGE	PILEKQV	QHADVNA
C2c3_CEVA (SEQ ID NO:167)	VAIDLGE	PVLESSV	CHADENA
C2c3_CEPX (SEQ ID NO:166)	VAIDLGE	PVLEFQI	GHADENA
C2c3_CEPS (SEQ ID NO:168)	LAIDLGE	PVLESSV	GHADENA
AcoC2c1 (SEQ ID NO:157)	MSVDLGV	ILFEDLS	VHADINA
ObC2c1 (SEQ ID NO:160)	LGVDLG	VVIENL	MQADL

[0299]

	T	S	NA
DbC2c1 (SEQ ID NO:164)	LSVDLGH	VVIENLA	IHADLNA
DiC2c1 (SEQ ID NO:158)	LSVDLGM	ILFEDLA	IHADMNA
DtC2c1 (SEQ ID NO:159)	LSVDLGV	ILFEDLA	IHADINA
AacC2c1 (SEQ ID NO:150)	MSVDLGL	ILLEELS	IHADLNA
Bsp1C2c1 (SEQ ID NO:163)	MSIDLGL	ILFENLS	LQADINA
TcC2c1 (SEQ ID NO:161)	MSVDLGQ	VLFEDLS	THADINA
BtC2c1 (SEQ ID NO:162)	MSIDLGQ	ILFEDLS	THADINA

[0300] 序列比对和其它计算分析未显示CasY.5或CasY.6的清楚RuvCIII催化残基。Unk64和Unk69中假定的催化残基分别是赖氨酸和天冬酰胺,而其它的均在该位置具有不变的天冬氨酸残基。对于其余的V型核酸酶,将表6中汇总的RuvC催化残基用于产生RuvC锚着的序列比对,其中催化残基用作固定锚点,使用先前描述的方法进行(Begemann等.(2017) BioRxiv doi:10.1101/192799)。所得的RuvC锚着氨基酸比对用于构建系统发育树,如图1所示。如图所示,Cms1核酸酶与其它V型核酸酶位于不同的进化枝上。此外,在此分析中,至少有三个独立组的Cms1核酸酶簇集在一起(在表6中,这些组分别包含MicroCms1至Unk78Cms1,SulfCms1至Unk71Cms1和Unk40Cms1至Unk76Cms1),表明在该较大分组中至少存在三组Cms1核酸酶。对于包括SmCms1(SEQ ID NO:10),SulfCms1(SEQ ID NO:11)和Unk40Cms1(SEQ ID NO:68)的核酸酶的组,这三个组分别标记为“Sm型”,“Sulf型”和“Unk40型”。

[0301] 研究了Cms1核酸酶的氨基酸序列比对,以鉴定这些核酸酶之间保守性良好的蛋白质序列内的基序。观察到,Cms1核酸酶存在于图1所示的系统发育树上的三个良好分离的进化枝中。这些进化枝中的一个包括SmCms1(SEQ ID NO:10),另一个包括SulfCms1(SEQ ID NO:11),另一个包括Unk40Cms1(SEQ ID NO:68)。因此,将每个进化枝的成员分开排列,以鉴定这些核酸酶中的部分和/或完全保守的氨基酸基序。对于SmCms1样核酸酶的比对,SEQ ID NO:10、20、23、30、32-34、37-39、41、43、44、46-60、67、154-156、208-211、222、223、225、228、229、232、234、236、237、241、243、245、248、250、251、253和254是比对上的。对于SulfCms1样核酸酶的比对,SEQ ID NO:11、21、22、31、35、36、40、42、45、61-66、69、227、230、231、235、239、240、242、244和247是比对上的。对于Unk40样核酸酶的比对,SEQ ID NO:68、224、226、233、238、246、249和252是比对上的。使用MUSCLE进行这些比对,并手动检查所得的比对,以鉴定在所有比对的蛋白质中显示出保守性的区域。从SmCms1样核酸酶的比对中鉴定出SEQ ID NO:177-186所示的氨基酸基序;从SulfCms1样核酸酶的比对中鉴定出SEQ ID NO:288-289和187-201所示的氨基酸基序;从Unk40Cms1样核酸酶的比对中鉴定出SEQ ID NO:290-296所示的氨基酸基序。Weblogos使用序列比对创建,并以图形示于图2-4(分别为SmCms1

样,SulfCms1样和Unk40Cms1样序列基序;weblogo.berkeley.edu)以及示意图中,其显示这些保守基序在SmCms1、SulfCms1和Unk40Cms1蛋白序列上的位置。

[0302] 如本文所述用Cms1核酸酶编辑植物基因组表明,与V型核酸酶的一些其它描述一致,许多(即使不是全部)Cms1核酸酶也可接近TTTN或TTN PAM位点。进行计算分析以鉴定BLAST命中位点,其对应于编码Cms1核酸酶的重叠群上存在的CRISPR间隔子。使用CRISPRfinder在线(crispr.i2bc.paris-saclay.fr/Server/)鉴定CRISPR间隔子;这些间隔子被用作针对基因组的BLAST搜索的种子。针对来自编码AuxCms1,Unk15Cms1,Unk19Cms1和Unk40Cms1(分别为SEQ ID NO:297-300)的重叠群的CRISPR间隔子鉴定出BLAST命中位点。这些BLAST命中示于SEQ ID NO:301-307,并连同在BLAST命中之前和之后的核苷酸汇总于表7。

[0303] 表7:来自编码的Cms1重叠群的CRISPR间隔子的BLAST命中汇总

[0304]

有 CRISPR 间隔子的重叠群	BLAST 命中与周围核苷酸	SEQ ID NO	BLAST 命中的核苷酸位置
Aux (SEQ ID NO:297)	CTCTTATGGTACAGACGGGTCATGAATGTAACGCTGTCCAG	301	2896-2923
Unk15 (SEQ ID NO:298)	CTTTTATTGCGGATTTGCTCAATGCAACGTTCTCTAATAAA	302	5486-5513

[0305]

Unk15 (SEQ ID NO:298)	CATTTAGAGGAAATC TATAGTCATGTTTTGT TAAGAGATTT	303	1971-1999
Unk19 (SEQ ID NO:299)	TCTTTACCAAGTCCC CCCGCAACATCATAA AACATTTTAGA	304	4823-4850
Unk19 (SEQ ID NO:299)	TATTTCTAGCAACCC ACTCAGCATAATCGT TTTCCGGAACG	304	5831-5859
Unk19 (SEQ ID NO:299)	CCATTAACCTGGCGG AGGCTAACCCTCCGC CTATAAACAAA	305	1487-1514
Unk19 (SEQ ID NO:299)	ACTTTAGAATACTTA TCAATAACCTGCTCT TCGGTTTGGTT	306	725-752
Unk40 (SEQ ID NO:300)	CGTTTATATTCGGTTG CCACTCCTCGAAGTA TTGCTTATAG	307	209-236
Unk54 (SEQ ID NO:308)	CTTTTAATCCACGCG CCGCCCACTATGATA ACTTGCCGGAA	309	6064-6092
Unk54 (SEQ ID NO:308)	TGGTTAATAATTCAT TGTTTATTTTTGGGTT AAAAATTTCG	310	4102-4131
Unk54 (SEQ ID NO:308)	TCGTTAATAATTGGT GAATATGATTTACAA CAAATGGCTGC	311	17-44

[0306] 表7中,带下划线的碱基表示CRISPR间隔子BLAST命中。值得注意的是,此表中BLAST命中5'紧接碱基均显示TTA或TTC,并且此表中11个BLAST命中的7个显示TTTA或TTTC。这些数据,结合上述植物基因组编辑数据,有力地表明,至少这些Cms1核酸酶(以及可能的大多数或所有Cms1核酸酶)可以触及至少TTM PAM位点下游的靶位点,并且TTTM PAM位点优先。值得注意的是,这些类型的经计算鉴定的PAM位点不仅考虑了核酸酶PAM的要求,还考虑了CRISPR间隔子获取机制的要求,因此,核酸酶可能比能够触及比本文中经计算鉴定的那些PAM位点更多的PAM位点。

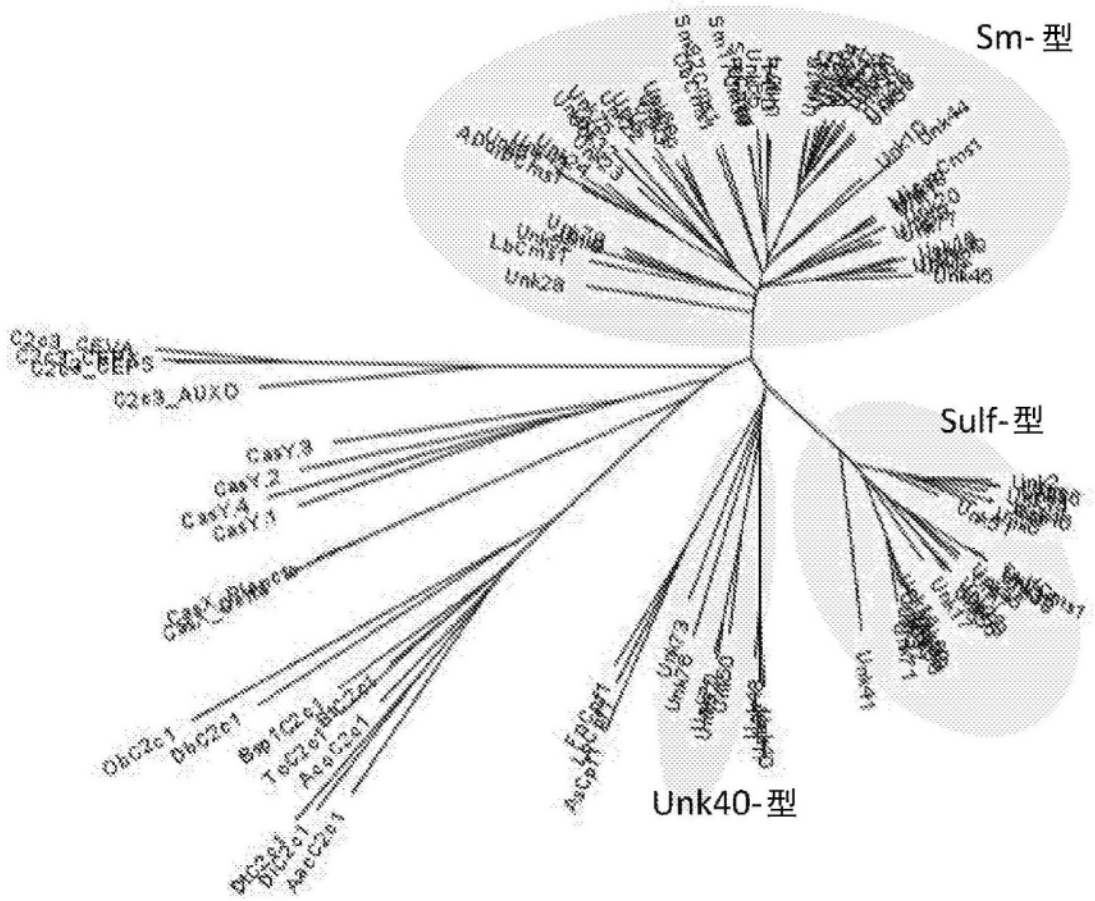

Sm-型
Sulf-型
Unk40-型
C2c3_AUXO
CasY.3
CasY.2
CasY.4
CasY.1
Unk28
Unk48
Unk41

图1

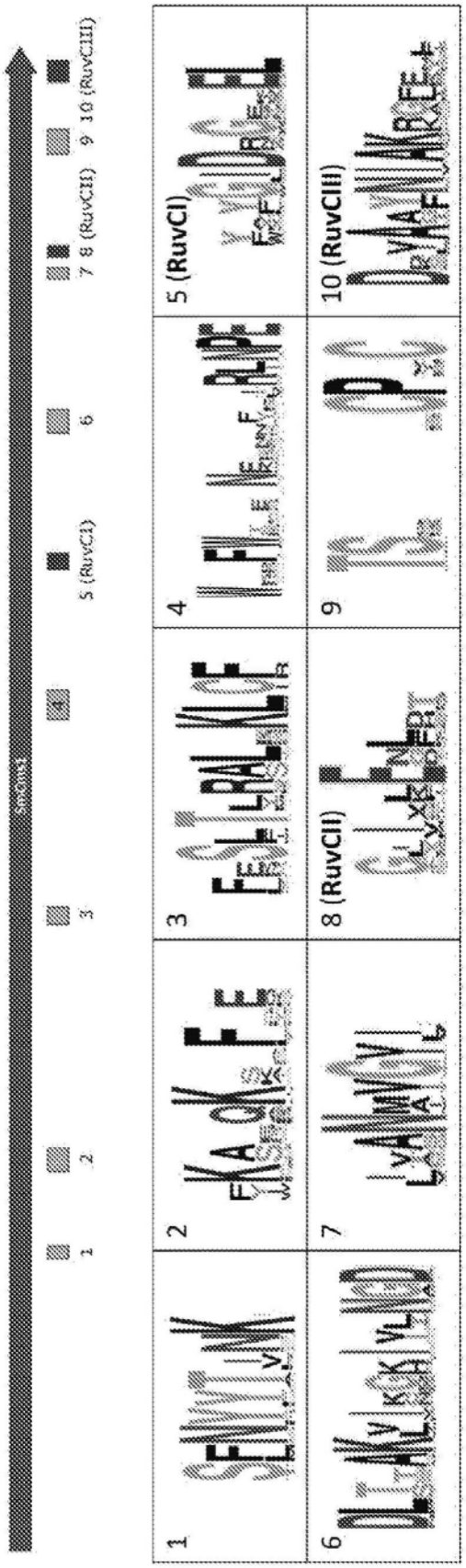


图2

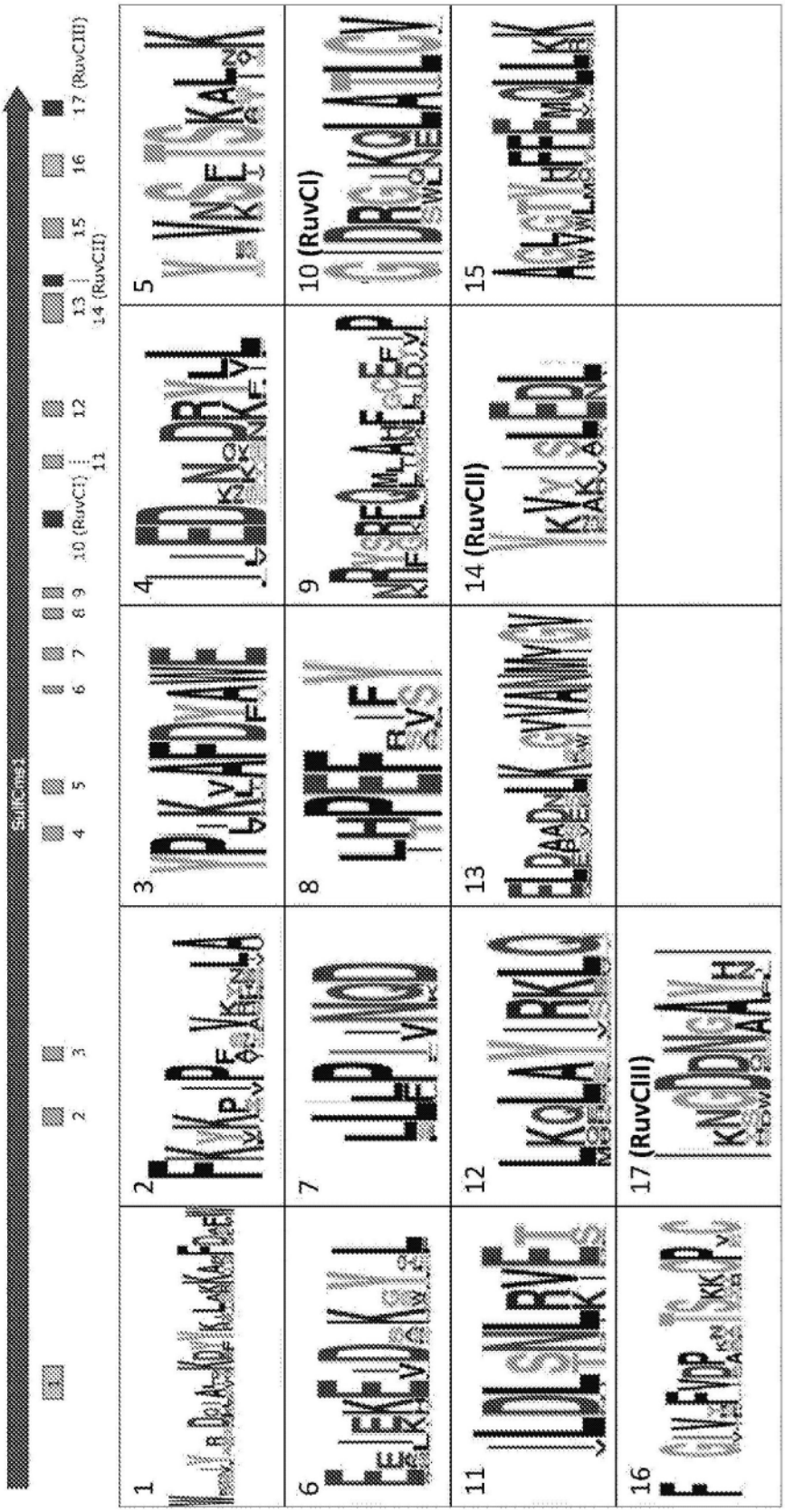

SunCmk1
2 3 4 5 6 7 8 9 10 (RuvCI) 11 12 13 14 (RuvCII) 15 16 17 (RuvCIII)
1
2
3
4
5
6
7
8
9
10 (RuvCI)
11
12
13
14 (RuvCII)
15
16
17 (RuvCIII)

图3

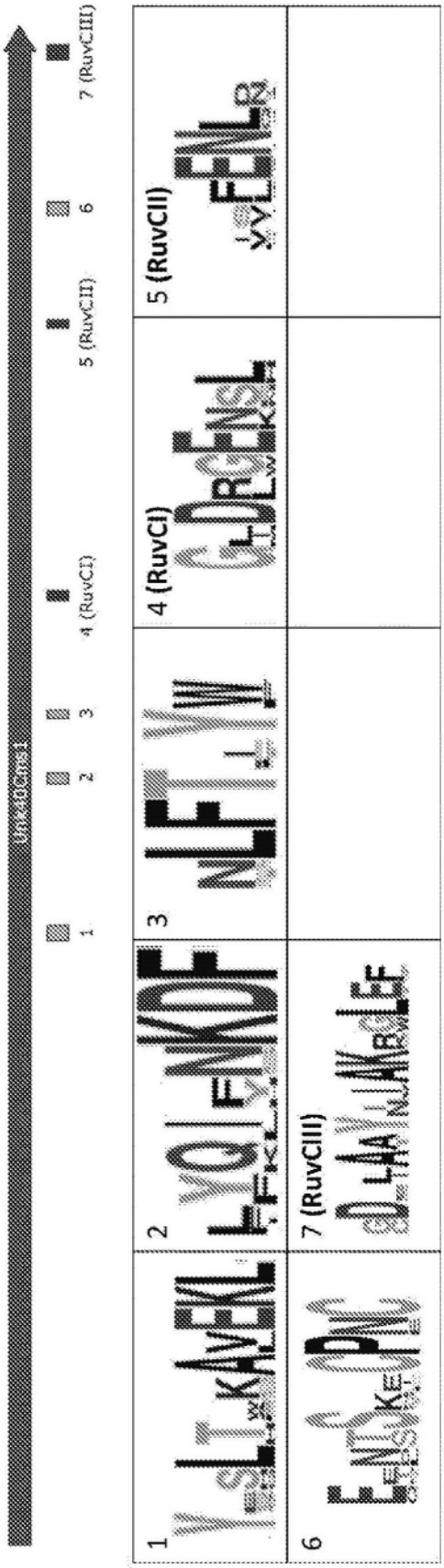

Unit10.ms1
1
2
3
4 (RuvCI)
5 (RuvCII)
6
7 (RuvCIII)
1
2
3
4 (RuvCI)
5 (RuvCII)
6
7 (RuvCIII)

图4