Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2985506 A1 2016/11/24

(21) 2 985 506

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2016/05/12
(87) Date publication PCT/PCT Publication Date: 2016/11/24
(85) Entrée phase nationale/National Entry: 2017/11/08
(86) N° demande PCT/PCT Application No.: US 2016/032073
(87) N° publication PCT/PCT Publication No.: 2016/186953
(30) Priorités/Priorities: 2015/05/15 (US62/162,353); 2015/05/15 (US62/162,377); 2015/07/24 (US62/196,535)

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

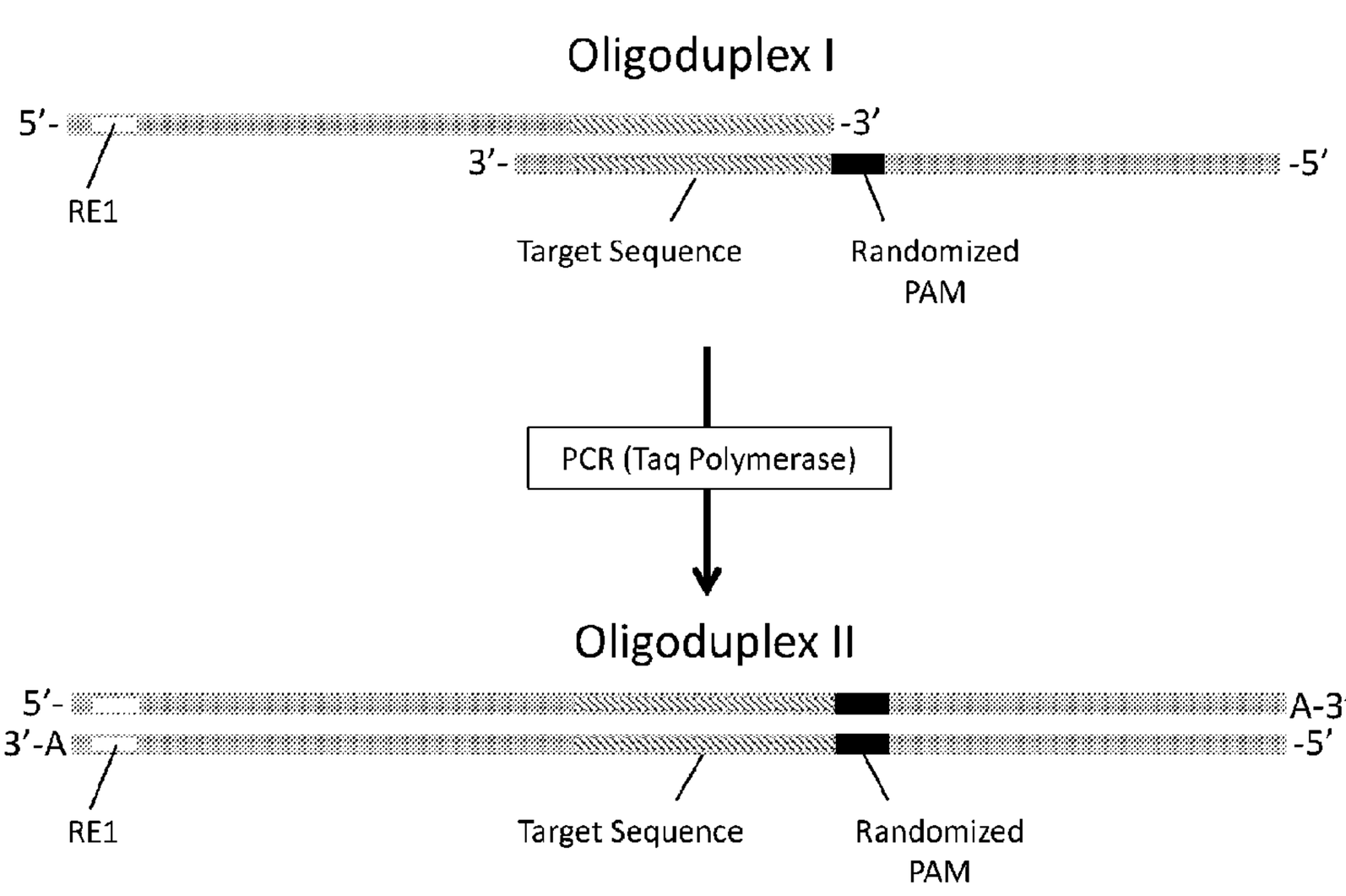
(71) Demandeur/Applicant:
PIONEER HI-BRED INTERNATIONAL, INC., US

(72) Inventeurs/Inventors:
CIGAN, ANDREW MARK, US;
GASIUNAS, GIEDRIUS, LT;
KARVELIS, TAUTVYDAS, LT;
SIKSNYS, VIRGINIJUS, LT;
YOUNG, JOSHUA K., US

(74) Agent: TORYS LLP

(54) Titre : SYSTEMES ARN GUIDE/ENDONUCLEASE CAS
(54) Title: GUIDE RNA/CAS ENDONUCLEASE SYSTEMS

Figure 1



(57) Abrégé/Abstract:
Compositions and methods are provided for novel guide RNA/ Cas endonuclease systems. Type II Cas9 endonuclease systems originating from Brevibacillus laterosporus, Lactobacillus reuteri Mlc3, Lactobacillus rossiae DSM 15814, Pediococcus pentosaceus SL4, Lactobacillus nodensis JCM 14932, Sulfurospirillum sp. SCADC, Bifidobacterium thermophilum DSM 20210, Loktanella vestfoldensis, Sphingomonas sanxanigenens NX02, Epilithonimonas tenax DSM 16811, Sporocytophaga myxococcoides are described herein. The present disclosure also describes methods for genome modification of a target sequence in the genome of a cell, for gene editing, and for inserting a polynucleotide of interest into the genome of a cell.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 November 2016 (24.11.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/186953 A1

(51) International Patent Classification:
C12N 9/22 (2006.01) *C12N 15/90* (2006.01)
C12N 15/113 (2010.01)

(21) International Application Number:
PCT/US2016/032073

(22) International Filing Date:
12 May 2016 (12.05.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

62/162,377	15 May 2015 (15.05.2015)	US
62/162,353	15 May 2015 (15.05.2015)	US
62/196,535	24 July 2015 (24.07.2015)	US

(71) Applicant: **PIONEER HI BRED INTERNATIONAL INC** [US/US]; 7100 N.W. 62nd Avenue, Johnston, Iowa 50131-1014 (US).

(72) Inventors: **CIGAN, Andrew Mark**; 5764 Chatham Circle, Johnston, Iowa 50131 (US). **GASIUNAS, Giedrius**; Institute Of Biotechnology Vilnius University, Graiciuno 8, LT-02241 Vilnius (LT). **KARVELIS, Tautvydas**; Institute Of Biotechnology Vilnius University, Graiciuno 8, LT-02241 Vilnius (LT). **SIKSNYS, Virginijus**; Institute Of Biotechnology Vilnius University, Graiciuno 8, LT-02241 Vilnius (LT). **YOUNG, Joshua K.**; 5981 Somerset Place, Johnston, Iowa 50131 (US).

(74) Agent: **STOOP, Johan M.**; E. I. du Pont de Nemours and Company, Legal Patent Records Center, Chestnut Run Plaza 721/2340, 974 Centre Road, PO Box 2915 Wilmington, Delaware 19805 (US).

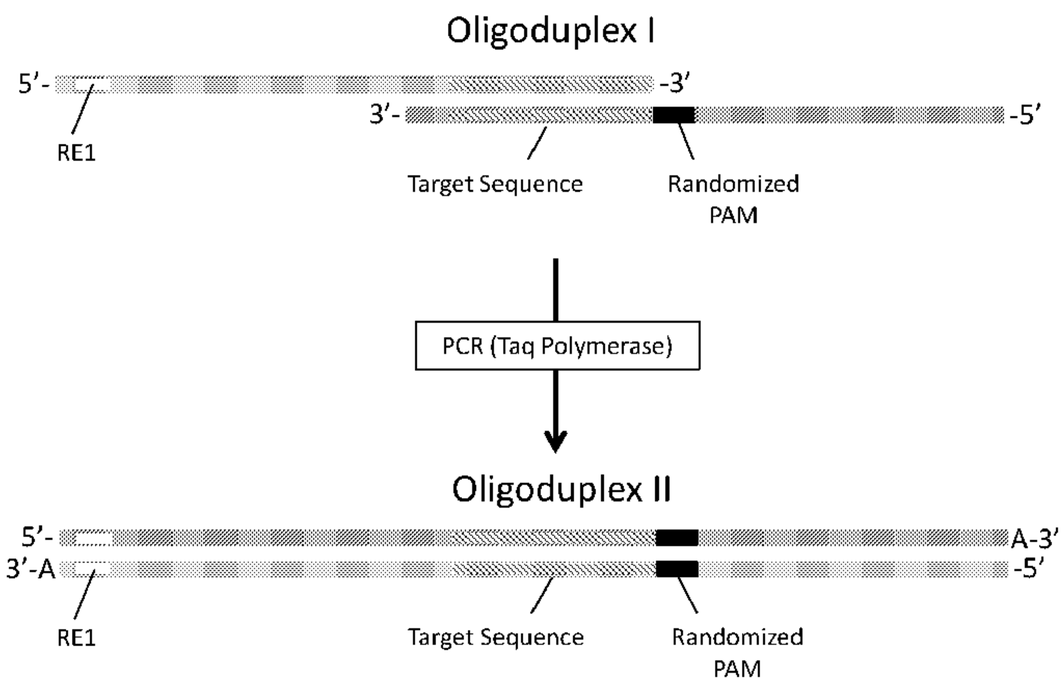
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,

[Continued on next page]

(54) Title: GUIDE RNA/CAS ENDONUCLEASE SYSTEMS

Figure 1



(57) Abstract: Compositions and methods are provided for novel guide RNA/ Cas endonuclease systems. Type II Cas9 endonuclease systems originating from *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis, Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* are described herein. The present disclosure also describes methods for genome modification of a target sequence in the genome of a cell, for gene editing, and for inserting a polynucleotide of interest into the genome of a cell.

WO 2016/186953 A1

WO 2016/186953 A1

SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with sequence listing part of description (Rule 5.2(a))*

GUIDE RNA/CAS ENDONUCLEASE SYSTEMS

This application claims the benefit of U.S. Provisional Application No. 62/162,377, filed May 15, 2015, U.S. Provisional Application No. 62/162,353, filed May 15, 2015 and U.S. Provisional Application No. 62/196,535, filed July 24, 2015, which are incorporated herein in their entirety by reference.

FIELD

The disclosure relates to the field of plant molecular biology, in particular, to compositions for novel guide RNA/Cas endonuclease systems and compositions and methods for altering the genome of a cell.

REFERENCE TO SEQUENCE LISTING SUBMITTED ELECTRONICALLY

The official copy of the sequence listing is submitted electronically via EFS-Web as an ASCII formatted sequence listing with a file named 20160502_BB2539PCT_SequenceListing.txt created on May 2, 2016 and having a size 236 kilobytes and is filed concurrently with the specification. The sequence listing contained in this ASCII formatted document is part of the specification and is herein incorporated by reference in its entirety.

BACKGROUND

Recombinant DNA technology has made it possible to insert DNA sequences at targeted genomic locations and/or modify (edit) specific endogenous chromosomal sequences, thus altering the organism's phenotype. Site-specific integration techniques, which employ site-specific recombination systems, as well as other types of recombination technologies, have been used to generate targeted insertions of genes of interest in a variety of organism. Genome-editing techniques such as designer zinc finger nucleases (ZFNs) or transcription activator-like effector nucleases (TALENs), or homing meganucleases, are available for producing targeted genome perturbations, but these systems tends to have a low specificity and employ designed nucleases that need to be redesigned for each target site, which renders them costly and time-consuming to prepare.

Although several approaches have been developed to target a specific site for modification in the genome of an organism, there still remains a need for new genome engineering technologies that are affordable, easy to set up, scalable, and amenable to targeting multiple positions within the genome of an organism

BRIEF SUMMARY

Compositions and methods are provided for rapid characterization of novel Cas endonuclease systems and the elements comprising such a systems, including, but not limiting to, rapid characterization of PAM sequences, guide RNA elements and CAS endonucleases.

In one embodiment of the disclosure, the guide RNA is a guide RNAs capable of forming a guide RNA/Cas endonuclease complex, wherein said guide RNA/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said guide RNA/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence, wherein said tracrRNA is originated from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755.

In another embodiment of the disclosure, the guide RNA is a guide RNA capable of forming a guide RNA/Cas endonuclease complex, wherein said guide RNA/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a single molecule comprising a chimeric non-naturally occurring crRNA linked to a tracrRNA originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, wherein said chimeric

non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

In another embodiment of the disclosure, the guide RNA is a guide RNA capable of forming a guide RNA/Cas endonuclease complex, wherein said guide RNA/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said chimeric non-naturally occurring crRNA comprises at least a fragment of a crRNA originating from an organism selected from the group consisting of *Brevibacillus laterosporus,* *Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

In another embodiment of the disclosure, the guide RNA is a guide RNA capable of forming a guide RNA/Cas endonuclease complex, wherein said guide RNA/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a single molecule comprising a tracrRNA linked to a chimeric non-naturally occurring crRNA comprising at least a fragment of a crRNA originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

Also provided are nucleic acid constructs, plants, plant cells, explants, seeds and grain having an altered target site or altered polynucleotide of interest produced

by the methods described herein. Additional embodiments of the methods and compositions of the present disclosure are shown herein.

BRIEF DESCRIPTION OF THE DRAWINGS AND THE SEQUENCE LISTING

The disclosure can be more fully understood from the following detailed description and the accompanying drawings and Sequence Listing, which form a part of this application. The sequence descriptions and sequence listing attached hereto comply with the rules governing nucleotide and amino acid sequence disclosures in patent applications as set forth in 37 C.F.R. §§1.821-1.825. The sequence descriptions contain the three letter codes for amino acids as defined in 37 C.F.R. §§ 1.821-1.825, which are incorporated herein by reference.

Figures

Figure 1 shows a diagram of the formation of a full Oligoduplex II comprising a restriction enzyme recognition site (RE1), a target sequence and a randomized Protospacer-Adjacent-Motif (PAM) sequence.

Figure 2 show a diagram of the design and construction of a 5 nucleotide (5N) randomized Protospacer-Adjacent-Motif (PAM) plasmid library and host cell library. RE1= restriction endonuclease 1, RE2 = restriction endonuclease 2.

Figure 3 shows a diagram of the production of enriched PAM sided products for deep sequencing and identification of PAM preferences.

Figure 4 depicts the PAM sequence distribution from a 5 nucleotide (5N) randomized Protospacer-Adjacent-Motif (PAM) plasmid library.

Figure 5 shows the PAM preferences (NGGNG) for *Streptococcus thermophilus* CRISPR3 (Sth3) Cas9 endonuclease in both 50 nM and 100 nM digests.

Figure 6 shows the PAM preferences (NGG) for *Streptococcus pyogenes* (Spy) Cas9 endonuclease in both 50 nM and 100 nM digests.

Figure 7 shows the effect of decreasing Sth3 and Spy Cas9-crRNA-tracrRNA complex concentration and digestion time to determine the minimal Sth3 and Spy Cas9 concentration and shortest digestion time where PCR amplified cleavage products may still be obtained from the randomized PAM plasmid library.

Figure 8 shows the PAM preferences (NGGNG) for *Streptococcus thermophilus* CRISPR3 (Sth3) Cas9 endonuclease positive controls in both 50 nM and 100 nM digests.

Figure 9 shows the PAM preferences (NGG) for *Streptococcus pyogenes* (Spy) Cas9 endonuclease positive controls in both 50 nM and 100 nM digests.

Figure 10 shown the PAM preferences (NGGNG) observed in the minimally *Streptococcus thermophilus* Sth3 digested libraries (0.5 nM-60 min and 50 nM-1 min) compared to that exhibited by the respective 50 nM-60 minute positive control .

Figure 11 shown the PAM preferences (NGGNG) observed in the minimally *Streptococcus pyogenes* Spy digested libraries (0.5 nM-60 min and 50 nM-1 min) compared to that exhibited by the respective 50 nM-60 minute positive control .

Figure 12 shows the PAM preferences for *Streptococcus pyogenes* (Spy) Cas9 endonuclease guided by a single guide RNA (sgRNA) or guided by a crRNA:tracrRNA duplex. The NGGNG PAM preference is nearly identical regardless of the type of guide RNA used

Figure 13 shows the PAM preferences (NGG) for *Streptococcus pyogenes* (Spy) Cas9 endonuclease guided by a single guide RNA (sgRNA) or guided by a crRNA:tracrRNA duplex. The NGG PAM preference is nearly identical regardless of the type of guide RNA used

Figure 14 shows the PAM preferences for *Streptococcus thermophilus* CRISPR3 (Sth3) Cas9 endonuclease positive controls for comparing of a 5N randomized PAM plasmid DNA library and a 7N randomized PAM plasmid DNA library.

Figure 15 shows the PAM preferences (NGG) for *Streptococcus pyogenes* (Spy) Cas9 endonuclease positive controls for comparing of a 5N randomized PAM plasmid DNA library and a 7N randomized PAM plasmid DNA library.

Figure 16 shows the PAM preferences (NNAGAAW) for *Streptococcus thermophilus* CRISPR1 (Sth1) Cas9 endonuclease in both 50 nM and 0.5nM nM digests.

Figure 17-A shows a genomic DNA region from, *Brevibacillus laterosporus* representing the Type II CRISPR-Cas system described herein. Figure 17-B list 8 repeat sequences (SEQ ID NOs:37-44) of the genomic DNA region from the *Brevibacillus laterosporus.*

Figure 18 shows a diagram of the "direct" scenario and the "reverse " scenario of the tracrRNA and CRISPR array to determine a guide RNA for the Cas9 protein identified from the *Brevibacillus laterosporus* (Blat).

Figure 19 shows the secondary structure of the "direct" tracrRNA region downstream of the anti-repeat (SEQ ID NO: 68) from, *Brevibacillus laterosporus.*

Figure 20 shows the secondary structure of the "reverse" tracrRNA region downstream of the anti-repeat (SEQ ID NO: 69) from, *Brevibacillus laterosporus.*

Figure 21 shown an agarose gel with reaction products, indicating that only the "direct" sgRNA (dirsgRNA), but not the "reverse" sgRNA (revsgRNA) supported plasmid library cleavage in combination with a Cas9 endonuclease originating from *Brevibacillus laterosporus*.(BlatCas9).

Figure 22 shows the effect of decreasing BlatCas9 concentration and digestion time to determine the minimal Blast Cas9 concentration and shortest digestion time where PCR amplified cleavage products may still be obtained from the randomized PAM plasmid library.

Figure 23 shows the PAM preferences (NNNNCND) for *Brevibacillus laterosporus* (Blat) Cas9 endonuclease in both 50 nM and 0.5 nM digests.

Figure 24 depict sequencing results indicating that plasmid DNA cleavage occurred in the protospacer 3 bp away from the PAM sequence.

Figure 25 shows a genomic DNA region from *Lactobacillus reuteri Mlc3* representing an example of a Type II CRISPR-Cas system described herein.

Figure 26 shows a genomic DNA region from *Lactobacillus rossiae DSM 15814* representing an example of a Type II CRISPR-Cas system described herein.

Figure 27 shows a genomic DNA region from *Pediococcus pentosaceus SL4* representing an example of a Type II CRISPR-Cas system described herein.

Figure 28 shows a genomic DNA region from *Lactobacillus nodensis JCM 14932* representing an example of a Type II CRISPR-Cas system described herein.

Figure 29 shows a genomic DNA region from *Sulfurospirillum sp. SCADC* representing an example of a Type II CRISPR-Cas system described herein.

Figure 30 shows a genomic DNA region from *Bifidobacterium thermophilum DSM 20210* representing an example of a Type II CRISPR-Cas system described herein.

Figure 31 shows a genomic DNA region from *Loktanella vestfoldensis* representing an example of a Type II CRISPR-Cas system described herein.

Figure 32 shows a genomic DNA region from *Sphingomonas sanxanigenens NX02* representing an example of a Type II CRISPR-Cas system described herein.

Figure 33 shows a genomic DNA region from *Epilithonimonas tenax DSM 16811* representing an example of a Type II CRISPR-Cas system described herein.

Figure 34 shows a genomic DNA region from *Sporocytophaga myxococcoides* representing an example of a Type II CRISPR-Cas system described herein.

Figure 35 shows a genomic DNA region from *Psychroflexus torquis ATCC 700755* representing an example of a Type II CRISPR-Cas system described herein.

Figure 36 *Bifidobacterium thermophilum* (Bthe) Cas9 non-homologous end-joining (NHEJ) mutation frequencies with different single guide RNA (sgRNA) variable targeting domain (spacer) lengths (20 nt, 25 nt and 29 nt) at 2 maize target sites. NHEJ mutations were detected by deep sequencing 2 days after transformation.

Sequences

Table 1. Summary of Nucleic Acid and Protein SEQ ID Numbers

Description	Nucleic acid SEQ ID NO.	Protein SEQ ID NO.
Target sequence T1	1 (80 bases)	
Single oligonucleotide GG-821N	2 (47 bases)	
Oligonucleotide GG-820	3 (44 bases)	
TK-119 primer	4 (22 bases)	
pUC-dir primer	5 (22 bases)	
JKYS800.1 forward primer	6 (59 bases)	
JKYS803 reverse primer	7 (53 bases)	
Universal Forward primer	8 (43 bases)	
Universal Reverse primer	9 (18 bases)	
Sth1-dir primer	10 (34 bases)	
Sth1-rev primer	11 (27 bases)	
Sth3-dir primer	12 (26 bases)	
Sth3-rev primer	13 (30 bases)	
Spy-dir primer	14 (38 bases)	
Spy-rev primer	15 (32 bases)	
Streptococcus thermophilus (Sth3) crRNA	16 (42 bases)	
Streptococcus thermophilus (Sth3) tracrRNA	17 (78 bases)	
Streptococcus pyogenes (Spy) crRNA	18 (42 bases)	

Streptococcus pyogenes (Spy) tracrRNA	19 (78 bases)	
TK-117	20 (31 bases)	
TK-111	21 (30 bases)	
JKYS807.1 primer	22 (56 bases)	
JKYS807.2 primer	23 (56 bases)	
JKYS807.3 primer	24 (56 bases)	
JKYS807.4 primer	25 (56 bases)	
Sth3 sgRNA	26 (123 bases)	
Spy sgRNA	27 (105 bases)	
GG-940-G oligonucleotide	28 (59 bases)	
GG-940-C oligonucleotide	29 (59 bases)	
GG-940-A oligonucleotide	30 (59 bases)	
GG-940-T oligonucleotide	31 (59 bases)	
JKYS812	32 (49 bases)	
Streptococcus thermophilus CRISPR1 (Sth1) crRNA	33 (42 bases)	
Streptococcus thermophilus CRISPR1 Sth1 tracrRNA	34 (80 bases)	
Streptococcus thermophilus CRISPR3 (Sth3) Cas9		35 (1388 aa)
Cas9 single long open-reading-frame from the *Brevibacillus laterosporus* bacterial strain SSP360D4	36 (3279 bases)	
Repeat 1, *Brevibacillus laterosporus* SSP360D4	37 (36 bases)	

Repeat 2, *Brevibacillus laterosporus* SSP360D4	38 (36 bases)	
Repeat 3, *Brevibacillus laterosporus* SSP360D4	39 (36 bases)	
Repeat 4, *Brevibacillus laterosporus* SSP360D4	40 (36 bases)	
Repeat 5, *Brevibacillus laterosporus* SSP360D4	41 (36 bases)	
Repeat 6, *Brevibacillus laterosporus* SSP360D4	42 (36 bases)	
Repeat 7, *Brevibacillus laterosporus* SSP360D4	43 (36 bases)	
Repeat 8, *Brevibacillus laterosporus* SSP360D4	44 (36 bases)	
Blat-Cas9-dir	45 (29 bases)	
Blat-Cas9-rev	46 35 bases)	
Blat sgRNA Direct	47 (177 bases)	
Blat sgRNA Reverse	48 (118 bases)	
GG-969 oligonucleotide	49 (68 bases)	
GG-839 oligonucleotide	50 (62 bases)	
TK-149	51 55 bases)	
TK-150	52 (62 bases)	
GG-840	53 (71 bases)	
GG-841	54 (75 bases)	
TK-124	55 (37 bases)	
TK-151	56 (26 bases)	
TK-126;	57 (32 bases)	
GG-935	58 (37 bases)	
GG-936	59 (45 bases)	

pUC-EheD primer	60 (21 bases)	
pUC-LguR primer	61 (22 bases)	
Sense DNA Strand of Cleaved Sequencing Template	62 (21 bases)	
Anti-Sense DNA Strand Sequencing Read	63 (11 bases)	
Anti-Sense DNA Strand of Cleaved Sequencing Template	64 (21 bases)	
Sense DNA Strand of DNA Sequencing Read	65 (11 bases)	
Sense DNA Strand of Target and PAM	66 (27 bases)	
Anti-Sense DNA Strand of Target and PAM	67 (27 bases)	
"Direct" tracrRNA region downstream of the anti-repeat	68 (118 bases)	
"Reverse" tracrRNA region downstream of the anti-repeat	69 (58 bases)	
Lactobacillus reuteri Mlc3 (Lreu) Cas9 Open Reading Frame	70 (4107 bases)	
Lactobacillus rossiae DSM 15814 (Lros) Cas9 Open Reading Frame	71 (4110 bases)	
Pediococcus pentosaceus SL4 (Ppen) Cas9 Open Reading Frame	72 (4041 bases)	
Lactobacillus nodensis JCM 14932 (Lnod) Cas9 Open Reading Frame	73 (3393 bases)	
Sulfurospirillum sp. SCADC (Sspe) Cas9 Open Reading Frame	74 (4086 bases)	
Bifidobacterium thermophilum DSM 20210 (Bthe) Cas9 Open Reading Frame	75 (3444 bases)	
Loktanella vestfoldensis (Lves) Cas9 Open Reading Frame	76 (3216 bases)	

Sphingomonas sanxanigenens NX02 (Ssan) Cas9 Open Reading Frame	77 (3318 bases)	
Epilithonimonas tenax DSM 16811 (Eten) Cas9 Open Reading Frame	78 (4200 bases)	
Sporocytophaga myxococcoides (Smyx) Cas9 Open Reading Frame	79 (4362 bases)	
Psychroflexus torquis ATCC 700755 (Ptor) Cas9 Open Reading Frame	80 (4530 bases)	
Lreu Cas9 Endonuclease		81 (1368 aa)
Lros Cas9 Endonuclease		82 (1369 aa)
Ppen Cas9 Endonuclease		83 (1346 aa)
Lnod Cas9 Endonuclease		84 (1130 aa)
Sspe Cas9 Endonuclease		85 (1361 aa)
Bthe Cas9 Endonuclease		86 (1147 aa)
Lves Cas9 Endonuclease		87 (1071 aa)
Ssan Cas9 Endonuclease		88 (1105 aa)
Eten Cas9 Endonuclease		89 (1399 aa)
Smyx Cas9 Endonuclease		90 (1453 aa)
Ptor Cas9 Endonuclease		91 (1509 aa)
Lreu CRISPR Repeat Consensus	92 (36 bases)	
Lros CRISPR Repeat Consensus	93 (36 bases)	
Ppen CRISPR Repeat Consensus	94 (36 bases)	
Lnod CRISPR Repeat Consensus	95 (36 bases)	
Sspe CRISPR Repeat Consensus	96 (36 bases)	

Bthe CRISPR Repeat Consensus	97 (36 bases)	
Lves CRISPR Repeat Consensus	98 (36 bases)	
Ssan CRISPR Repeat Consensus	99 (36 bases)	
Eten CRISPR Repeat Consensus	100 (47 bases)	
Smyx CRISPR Repeat Consensus	101 (47 bases)	
Ptor CRISPR Repeat Consensus	102 46 bases)	
Lreu Anti-Repeat	103 (36 bases)	
Lros Anti-Repeat	104 (37 bases)	
Ppen Anti-Repeat	105 (37 bases)	
Lnod Anti-Repeat	106 (38 bases)	
Sspe Anti-Repeat	107 (39 bases)	
Bthe Anti-Repeat	108 (36 bases)	
Lves Anti-Repeat	109 (36 bases)	
Ssan Anti-Repeat	110 (36 bases)	
Eten Anti-Repeat	111 (47 bases)	
Smyx Anti-Repeat	112 (47 bases)	
Ptor Anti-Repeat	113 (46 bases)	
Lreu Single guide RNA	114 (169 bases)	
Lros Single guide RNA	115 (166 bases)	
Ppen Single guide RNA	116 (168 bases)	
Lnod Single guide RNA	117 (114 bases)	
Sspe Single guide RNA	118 (180 bases)	

Sspe Single guide RNA	119 (117 bases)	
Bthe Single guide RNA	120 (254 bases)	
Lves Single guide RNA	121 (200 bases)	
Ssan Single guide RNA	122 (195 bases)	
Eten Single guide RNA	123 (155 bases)	
Smyx Single guide RNA	124 (149 bases)	
Ptor Single guide RNA	125 (155 bases)	
GG-939	126 (57 bases)	
Single guide RNA	127 (174 bases)	
Lreu Single guide RNA	128 (166 bases)	
Lros Single guide RNA	129 (163 bases)	
Ppen Single guide RNA	130 (165 bases)	
Lnod Single guide RNA	131 (111 bases)	
Sspe Single guide RNA	132 (177 bases)	
Sspe Single guide RNA	133 (114 bases)	
Bthe Single guide RNA	134 (251 bases)	
Lves Single guide RNA	135 (197 bases)	
Ssan Single guide RNA	136 (192 bases)	
Eten Single guide RNA	137 (152 bases)	
Smyx Single guide RNA	138 (146 bases)	
Ptor Single guide RNA	139 (152 bases)	
Cas9 endonuclease *Brevibacillus laterosporus*		140 (1092 aa)

bacterial strain SSP360D4		
Variable Targeting domain-direct	141	
Variable Targeting domain-reverse	142	
16 nt loop of the repeat-direct	143	
16 nt loop of the repeat-reverse	144	
anti-repeat region-direct	145	
anti-repeat region-reverse	146	
Putative 3' tracrRNA Sequence - direct	147	
Putative 3' tracrRNA Sequence - reverse	148	
Lactobacillus reuteri Mlc3 (Lreu) crRNA repeat region	149	
Lactobacillus rossiae DSM 15814 (Lros) crRNA repeat region	150	
Pediococcus pentosaceus SL4 (Ppen) crRNA repeat region	151	
Lactobacillus nodensis JCM 14932 (Lnod) crRNA repeat region	152	
Sulfurospirillum sp. SCADC (Sspe) crRNA repeat region	153-154	
Bifidobacterium thermophilum DSM 20210 (Bthe) crRNA repeat region	155	
Loktanella vestfoldensis (Lves) crRNA repeat region	156	
Sphingomonas sanxanigenens NX02 (Ssan) crRNA repeat region	157	
Epilithonimonas tenax DSM 16811 (Eten) crRNA repeat region	158	

Sporocytophaga myxococcoides (Smyx) crRNA repeat region	159	
Psychroflexus torquis ATCC 700755 (Ptor) crRNA repeat region	160	
Lactobacillus reuteri Mlc3 (Lreu) tracrRNA anti-repeat	161	
Lactobacillus rossiae DSM 15814 (Lros) tracrRNA anti-repeat	162	
Pediococcus pentosaceus SL4 (Ppen) tracrRNA anti-repeat	163	
Lactobacillus nodensis JCM 14932 (Lnod) tracrRNA anti-repeat	164	
Sulfurospirillum sp. SCADC (Sspe) tracrRNA anti-repeat	165-166	
Bifidobacterium thermophilum DSM 20210 (Bthe) tracrRNA anti-repeat	167	
Loktanella vestfoldensis (Lves) tracrRNA anti-repeat	168	
Sphingomonas sanxanigenens NX02 (Ssan) tracrRNA anti-repeat	169	
Epilithonimonas tenax DSM 16811 (Eten) tracrRNA anti-repeat	170	
Sporocytophaga myxococcoides (Smyx) tracrRNA anti-repeat	171	
Psychroflexus torquis ATCC 700755 (Ptor) tracrRNA anti-repeat	172	
Lactobacillus reuteri Mlc3 (Lreu) 3' tracrRNA	173	
Lactobacillus rossiae DSM 15814 (Lros) 3' tracrRNA	174	

Pediococcus pentosaceus SL4 (Ppen) 3' tracrRNA	175	
Lactobacillus nodensis JCM 14932 (Lnod) 3' tracrRNA	176	
Sulfurospirillum sp. SCADC (Sspe) 3' tracrRNA	177-178	
Bifidobacterium thermophilum DSM 20210 (Bthe) 3' tracrRNA	179	
Loktanella vestfoldensis (Lves) 3' tracrRNA	180	
Sphingomonas sanxanigenens NX02 (Ssan) 3' tracrRNA	181	
Epilithonimonas tenax DSM 16811 (Eten) 3' tracrRNA	182	
Sporocytophaga myxococcoides (Smyx) 3' tracrRNA	183	
Psychroflexus torquis ATCC 700755 (Ptor) 3' tracrRNA	184	

DETAILED DESCRIPTION

Compositions and methods are provided for rapid characterization of Cas endonuclease systems and the elements comprising such a systems, including, but not limiting to, rapid characterization of PAM sequences, guide RNA elements and Cas endonucleases. Cas9 endonuclease systems originating from *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* are described herein.

The present disclosure also describes methods for genome modification of a target sequence in the genome of a cell, for gene editing, and for inserting a polynucleotide of interest into the genome of a cell.

CRISPR (clustered regularly interspaced short palindromic repeats) loci refers to certain genetic loci encoding factors of DNA cleavage systems, for example, used by bacterial and archaeal cells to destroy foreign DNA (Horvath and Barrangou, 2010, Science 327:167-170). A CRISPR locus can consist of a CRISPR array, comprising short direct repeats separated by short variable DNA sequences (called 'spacers'), which can be flanked by diverse Cas (CRISPR-associated) genes. Multiple CRISPR-Cas systems have been described including Class 1 systems, with multisubunit effector complexes, and Class 2 systems , with single protein effectors (such as but not limiting to Cas9, Cpf1 ,C2c1,C2c2, C2c3). (Zetsche et al., 2015, Cell 163,1-13; Shmakov et al., 2015, Molecular_Cell 60,1-13; Makarova et al. 2015, Nature Reviews Microbiology Vol. 13 :1-15; WO 2013/176772 A1 published on November 23, 2013 and incorporated by its entirety by reference herein).

The type II CRISPR/Cas system from bacteria employs a crRNA (CRISPR RNA) and tracrRNA (trans-activating CRISPR RNA) to guide a Cas9 endonuclease (encoded by a *cas9* gene) to its DNA target. The crRNA contains a spacer region complementary to one strand of the double strand DNA target and a region that base pairs with the tracrRNA (trans-activating CRISPR RNA) forming a RNA duplex that directs the Cas9 endonuclease to cleave the DNA target. Spacers are acquired through a not fully understood process involving Cas1 and Cas2 proteins. All type II CRISPR-Cas loci contain cas1 and cas2 genes in addition to the cas9 gene (Makarova et al. 2015, Nature Reviews Microbiology Vol. 13:1-15). Type II CRISR-Cas loci can encode a tracrRNA, which is partially complementary to the repeats within the respective CRISPR array, and can comprise other proteins such as Csn1 and Csn2. The presence of cas9 in the vicinity of *cas1* and *cas2* genes is the hallmark of type II loci (Makarova et al. 2015, Nature Reviews Microbiology Vol. 13:1-15).

The number of CRISPR-associated genes at a given CRISPR locus can vary between species (Haft et al., 2005, Computational Biology, PLoS Comput Biol 1(6): e60. doi:10.1371/journal.pcbi.0010060; Makarova et al. 2015, Nature Reviews Microbiology Vol. 13 :1-15; WO 2013/176772 A1 published on November 23, 2013 and incorporated by its entirety by reference herein).

The term “Cas gene” herein refers to a gene that is generally coupled, associated or close to, or in the vicinity of flanking CRISPR loci. The terms “Cas gene”, “CRISPR-associated (Cas) gene” are used interchangeably herein.

The term “Cas endonuclease” herein refers to a protein encoded by a Cas gene. A Cas endonuclease herein, when in complex with a suitable polynucleotide component, is capable of recognizing, binding to, and optionally nicking or cleaving all or part of a specific DNA target sequence. A Cas endonuclease described herein comprises one or more nuclease domains. Cas endonucleases of the disclosure includes those having a HNH or HNH-like nuclease domain and / or a RuvC or RuvC-like nuclease domain. A Cas endonuclease of the disclosure includes a Cas9 protein, a Cpf1 protein, a C2c1 protein, a C2c2 protein, a C2c3 protein, Cas3, Cas 5, Cas7, Cas8, Cas10, or complexes of these.

As used herein, the terms “guide polynucleotide/Cas endonuclease complex”, “guide polynucleotide/Cas endonuclease system”, “ guide polynucleotide/Cas complex”, “guide polynucleotide/Cas system” are used interchangeably herein and refer to at least one guide polynucleotide and at least one Cas endonuclease that are capable of forming a complex, wherein said guide polynucleotide/Cas endonuclease complex can direct the Cas endonuclease to a DNA target site, enabling the Cas endonuclease to recognize, bind to, and optionally nick or cleave (introduce a single or double strand break) into the DNA target site. A guide polynucleotide/Cas endonuclease complex herein can comprise Cas protein(s) and suitable polynucleotide component(s) of any of the four known CRISPR systems (Horvath and Barrangou, Science 327:167-170) such as a type I, II, or III CRISPR system. A Cas endonuclease unwinds the DNA duplex at the target sequence and optionally cleaves at least one DNA strand, as mediated by recognition of the target sequence by a polynucleotide (such as, but not limited to, a crRNA or guide RNA) that is in complex with the Cas protein. Such recognition and cutting of a target sequence by a Cas endonuclease typically occurs if the correct protospacer-adjacent motif (PAM) is located at or adjacent to the 3' end of the DNA target sequence. Alternatively, a Cas protein herein may lack DNA cleavage or nicking activity, but can still specifically bind to a DNA target sequence when complexed with a suitable RNA component. (See also U.S. Patent Application US 2015-

0082478 A1, published on March 19, 2015 and US 2015-0059010 A1, published on February 26, 2015, both are hereby incorporated in its entirety by reference).

A guide polynucleotide/Cas endonuclease complex can cleave one or both strands of a DNA target sequence. A guide polynucleotide/Cas endonuclease complex that can cleave both strands of a DNA target sequence typically comprises a Cas protein that has all of its endonuclease domains in a functional state (e.g., wild type endonuclease domains or variants thereof retaining some or all activity in each endonuclease domain). Thus, a wild type Cas protein (e.g., a Cas9 protein disclosed herein), or a variant thereof retaining some or all activity in each endonuclease domain of the Cas protein, is a suitable example of a Cas endonuclease that can cleave both strands of a DNA target sequence. A Cas9 protein comprising functional RuvC and HNH nuclease domains is an example of a Cas protein that can cleave both strands of a DNA target sequence. A guide polynucleotide/Cas endonuclease complex that can cleave one strand of a DNA target sequence can be characterized herein as having nickase activity (e.g., partial cleaving capability). A Cas nickase typically comprises one functional endonuclease domain that allows the Cas to cleave only one strand (i.e., make a nick) of a DNA target sequence. For example, a Cas9 nickase may comprise (i) a mutant, dysfunctional RuvC domain and (ii) a functional HNH domain (e.g., wild type HNH domain). As another example, a Cas9 nickase may comprise (i) a functional RuvC domain (e.g., wild type RuvC domain) and (ii) a mutant, dysfunctional HNH domain. Non-limiting examples of Cas9 nickases suitable for use herein are disclosed by Gasiunas et al. (Proc. Natl. Acad. Sci. U.S.A. 109:E2579-E2586), Jinek et al. (Science 337:816-821), Sapranauskas et al. (Nucleic Acids Res. 39:9275-9282) and in U.S. Patent Appl. Publ. No. 2014/0189896, which are incorporated herein by reference.

A pair of Cas9 nickases can be used to increase the specificity of DNA targeting. In general, this can be done by providing two Cas9 nickases that, by virtue of being associated with RNA components with different guide sequences, target and nick nearby DNA sequences on opposite strands in the region for desired targeting. Such nearby cleavage of each DNA strand creates a double strand break (i.e., a DSB with single-stranded overhangs), which is then recognized as a

substrate for non-homologous-end-joining, NHEJ (leading to indel formation) or homologous recombination, HR. Each nick in these embodiments can be at least about 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, or 100 (or any integer between 5 and 100) bases apart from each other, for example. One or two Cas9 nickase proteins herein can be used in a Cas9 nickase pair. For example, a Cas9 nickase with a mutant RuvC domain, but functioning HNH domain (i.e., Cas9 HNH+/RuvC-), could be used (e.g., *Streptococcus pyogenes* Cas9 HNH+/RuvC-). Each Cas9 nickase (e.g., Cas9 HNH+/RuvC-) would be directed to specific DNA sites nearby each other (up to 100 base pairs apart) by using suitable RNA components herein with guide RNA sequences targeting each nickase to each specific DNA site.

A Cas protein can be part of a fusion protein comprising one or more heterologous protein domains (e.g., 1, 2, 3, or more domains in addition to the Cas protein). Such a fusion protein may comprise any additional protein sequence, and optionally a linker sequence between any two domains, such as between Cas and a first heterologous domain. Examples of protein domains that may be fused to a Cas protein herein include, without limitation, epitope tags (e.g., histidine [His], V5, FLAG, influenza hemagglutinin [HA], myc, VSV-G, thioredoxin [Trx]), reporters (e.g., glutathione-5-transferase [GST], horseradish peroxidase [HRP], chloramphenicol acetyltransferase [CAT], beta-galactosidase, beta-glucuronidase [GUS], luciferase, green fluorescent protein [GFP], HcRed, DsRed, cyan fluorescent protein [CFP], yellow fluorescent protein [YFP], blue fluorescent protein [BFP]), and domains having one or more of the following activities: methylase activity, demethylase activity, transcription activation activity (e.g., VP16 or VP64), transcription repression activity, transcription release factor activity, histone modification activity, RNA cleavage activity and nucleic acid binding activity. A Cas protein can also be in fusion with a protein that binds DNA molecules or other molecules, such as maltose binding protein (MBP), S-tag, Lex A DNA binding domain (DBD), GAL4A DNA binding domain, and herpes simplex virus (HSV) VP16.

A guide polynucleotide/Cas endonuclease complex in certain embodiments can bind to a DNA target site sequence, but does not cleave any strand at the target site sequence. Such a complex may comprise a Cas protein in which all of its nuclease domains are mutant, dysfunctional. For example, a Cas9 protein herein

that can bind to a DNA target site sequence, but does not cleave any strand at the target site sequence, may comprise both a mutant, dysfunctional RuvC domain and a mutant, dysfunctional HNH domain. A Cas protein herein that binds, but does not cleave, a target DNA sequence can be used to modulate gene expression, for example, in which case the Cas protein could be fused with a transcription factor (or portion thereof) (e.g., a repressor or activator, such as any of those disclosed herein).

In one embodiment, the Cas endonuclease gene is a Type II Cas9 endonuclease , such as but not limited to, Cas9 genes listed in SEQ ID NOs: 462, 474, 489, 494, 499, 505, and 518 of WO2007/025097published March 1, 2007, and incorporated herein by reference. In another embodiment, the Cas endonuclease gene is a plant, maize or soybean optimized Cas9 endonuclease gene. The Cas endonuclease gene can be operably linked to a SV40 nuclear targeting signal upstream of the Cas codon region and a bipartite VirD2 nuclear localization signal (Tinland et al. (1992) Proc. Natl. Acad. Sci. USA 89:7442-6) downstream of the Cas codon region.

"Cas9" (formerly referred to as Cas5, Csn1, or Csx12) herein refers to a Cas endonuclease of a type II CRISPR system that forms a complex with a crNucleotide and a tracrNucleotide, or with a single guide polynucleotide, for specifically recognizing and cleaving all or part of a DNA target sequence. Cas9 protein comprises a RuvC nuclease domain and an HNH (H-N-H) nuclease domain, each of which can cleave a single DNA strand at a target sequence (the concerted action of both domains leads to DNA double-strand cleavage, whereas activity of one domain leads to a nick). In general, the RuvC domain comprises subdomains I, II and III, where domain I is located near the N-terminus of Cas9 and subdomains II and III are located in the middle of the protein, flanking the HNH domain (Hsu et al, Cell 157:1262-1278).

Cas9 endonculeases are typically derived from a type II CRISPR system, which includes a DNA cleavage system utilizing a Cas9 endonuclease in complex with at least one polynucleotide component. For example, a Cas9 can be in complex with a CRISPR RNA (crRNA) and a trans-activating CRISPR RNA (tracrRNA). In another example, a Cas9 can be in complex with a single guide RNA

In one embodiment of the disclosure, the composition comprises at least one Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81-91, or a functional fragment thereof.

In one embodiment of the disclosure, the composition comprises at least one recombinant DNA vector encoding the Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81-91 (such as the DNA sequences form SEQ ID NO: 70-80), or mRNA encoding Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81-91. The Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81-91 can form a (Ribonucleotide Protein –RNP) complex with at least one guide RNA, wherein said complex is capable of recognizing, binding to, and optionally nicking or cleaving all or part of a target site.

Recombinant DNA expressing the Cas9 endonucleases described herein (including functional fragments thereof, plant or microbe codon optimized Cas9 endonuclease) can be stably integrated into the genome of an organism. For example, plants can be produced that comprise a cas9 gene stably integrated in the plant's genome. Plants expressing a stably integrated Cas endonuclease can be exposed to at least one guide RNA and/or a polynucleotide modification templates and/or donor DNAs to enable genome modications such as gene knockout, gene editing or DNA insertions.

A variant of a Cas9 protein sequence may be used, but should have specific binding activity, and optionally endonucleolytic activity, toward DNA when associated with an RNA component herein. Such a variant may comprise an amino acid sequence that is at least about 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identical to the amino acid sequence of the reference Cas9. Alternatively, a Cas9 protein may comprise an amino acid sequence that is at least about 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identical to any of the foregoing amino acid sequences, for example. Such a variant Cas9 protein should have specific binding activity, and optionally cleavage or nicking activity, toward DNA when associated with an RNA component herein.

The Cas endonuclease can comprise a modified form of the Cas9 polypeptide. The modified form of the Cas9 polypeptide can include an amino acid

change (e.g., deletion, insertion, or substitution) that reduces the naturally-occurring nuclease activity of the Cas9 protein. For example, in some instances, the modified form of the Cas9 protein has less than 50%, less than 40%, less than 30%, less than 20%, less than 10%, less than 5%, or less than 1% of the nuclease activity of the corresponding wild-type Cas9 polypeptide (US patent application US20140068797 A1, published on March 6, 2014). In some cases, the modified form of the Cas9 polypeptide has no substantial nuclease activity and is referred to as catalytically "inactivated Cas9" or "deactivated cas9 (dCas9)." Catalytically inactivated Cas9 variants include Cas9 variants that contain mutations in the HNH and RuvC nuclease domains. These catalytically inactivated Cas9 variants are capable of interacting with sgRNA and binding to the target site in vivo but cannot cleave either strand of the target DNA.

A catalytically inactive Cas9 can be fused to a heterologous sequence (US patent application US20140068797 A1, published on March 6, 2014). Suitable fusion partners include, but are not limited to, a polypeptide that provides an activity that indirectly increases transcription by acting directly on the target DNA or on a polypeptide (e.g., a histone or other DNA-binding protein) associated with the target DNA. Additional suitable fusion partners include, but are not limited to, a polypeptide that provides for methyltransferase activity, demethylase activity, acetyltransferase activity, deacetylase activity, kinase activity, phosphatase activity, ubiquitin ligase activity, deubiquitinating activity, adenylation activity, deadenylation activity, SUMOylating activity, deSUMOylating activity, ribosylation activity, deribosylation activity, myristoylation activity, or demyristoylation activity. Further suitable fusion partners include, but are not limited to, a polypeptide that directly provides for increased transcription of the target nucleic acid (e.g., a transcription activator or a fragment thereof, a protein or fragment thereof that recruits a transcription activator, a small molecule/drug-responsive transcription regulator, etc.). A catalytically inactive Cas9 can also be fused to a FokI nuclease to generate double strand breaks (Guilinger et al. Nature biotechnology, volume 32, number 6, June 2014).

A Cas protein herein such as a Cas9 endonuclease protein can comprise a heterologous nuclear localization sequence (NLS). A heterologous NLS amino acid sequence herein may be of sufficient strength to drive accumulation of a Cas protein

in a detectable amount in the nucleus of a yeast cell herein, for example. An NLS may comprise one (monopartite) or more (e.g., bipartite) short sequences (e.g., 2 to 20 residues) of basic, positively charged residues (e.g., lysine and/or arginine), and can be located anywhere in a Cas amino acid sequence but such that it is exposed on the protein surface. An NLS may be operably linked to the N-terminus or C-terminus of a Cas protein herein, for example. Two or more NLS sequences can be linked to a Cas protein, for example, such as on both the N- and C-termini of a Cas protein. The Cas endonuclease gene can be operably linked to a SV40 nuclear targeting signal upstream of the Cas codon region and a bipartite VirD2 nuclear localization signal (Tinland et al. (1992) Proc. Natl. Acad. Sci. USA 89:7442-6) downstream of the Cas codon region. Non-limiting examples of suitable NLS sequences herein include those disclosed in U.S. Patent No. 7309576, which is incorporated herein by reference.

The terms “functional fragment “, “fragment that is functionally equivalent” and “functionally equivalent fragment” of a Cas endonuclease are used interchangeably herein, and refer to a portion or subsequence of the Cas endonuclease sequence of the present disclosure in which the ability to recognize, bind to, and optionally nick or cleave (introduce a single or double strand break in) the target site is retained.

The terms “functional variant “, “Variant that is functionally equivalent” and “functionally equivalent variant” of a Cas endonuclease are used interchangeably herein, and refer to a variant of the Cas endonuclease of the present disclosure in which the ability to recognize, bind to, and optionally nick or cleave (introduce a single or double strand break in) the target site is retained. Fragments and variants can be obtained via methods such as site-directed mutagenesis and synthetic construction.

In one embodiment, the Cas endonuclease gene is a plant codon optimized *Streptococcus pyogenes* Cas9 gene that can recognize any genomic sequence of the form N(12-30)NGG can in principle be targeted.

In one embodiment, the Cas endonuclease is a Cas9 endonuclease originated from organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814,

Pediococcus pentosaceus SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, wherein said Cas9 endonuclease can form a guide RNA/Cas endonuclease complex capable of recognizing, binding to, and optionally nicking or cleaving all or part of a DNA target sequence.

The Cas endonuclease can be introduced directly into a cell by any method known in the art, for example, but not limited to transient introduction methods, transfection and/or topical application.

The guide polynucleotides and guide polynucleotide/Cas endonuclease systems described herein include guide polynucleotides comprising a crRNA (comprising a variable targeting (VT) domain linked to tracr-mate sequence that can hybridized to the tracr nucleotide) wherein said guide polynucleotide directs sequence-specific binding of the guide polynucleotide/Cas endonuclease complex to a target sequence in a eukaryotic cell. In an aspect, the guide polynucleotide targets a target sequence in a non-human eukaryotic organism preferably a multicellular eukaryotic organism, comprising a eukaryotic host cell. In one aspect, the guide polynucleotide is a non-naturally occurring guide polynucleotide or a guide polynucleotide targeting a target sequence that is not natural to bacteria. The disclosed guide polynucleotides can be reprogrammed to target nucleotide sequences in non-bacterial cells such as, but not limiting to changing the VT domain to target non-bacterial target sequences and sequences not naturally acquired by the system from which the crRNA was obtained. Alternatively, the VT domain can be programmed to guide the crRNA to a target sequence in a eukaryotic genome. Any sequence in a eukaryotic genome can be targeted using the disclosed guide polynucleotides, such as, mammalian (e.g. human, mouse, etc.), yeast, insect, animal, and plant sequences. In other embodiments, the VT domain can be programmed to guide the crRNA to a target sequence in a prokaryotic genome or bacterial plasmid sequence that is not naturally targeted by the native system.

In some embodiments, the guide polynucleotide/Cas endonuclease complex comprises one or more nuclear localization sequences of sufficient strength to drive

accumulation of said complex in a detectable amount in the nucleus of a eukaryotic cell. For example, nuclear localization signals can be added to the N- or C- or both the N- and C- terminus of the Cas protein. In other embodiments, one or more cellular localization signals can be included in the complex to provide for accumulation of the complex in a detectable amount in cellular organelles in which a desired target sequence is contained. For example, chloroplast targeting sequences can be added to the Cas protein to provide accumulation in a chloroplast organelle in a plant cell where the desired target sequence is found in the plant chloroplast genome.

The guide polynucleotide/Cas endonuclease system described herein can be provided to eukaryotic cells and reprogrammed to facilitate cleavage of endogenous eukaryotic target polynucleotides.

Endonucleases are enzymes that cleave the phosphodiester bond within a polynucleotide chain, and include restriction endonucleases that cleave DNA at specific sites without damaging the bases. Restriction endonucleases include Type I, Type II, Type III, and Type IV endonucleases, which further include subtypes. In the Type I and Type III systems, both the methylase and restriction activities are contained in a single complex. Endonucleases also include meganucleases, also known as homing endonucleases (HEases), which like restriction endonucleases, bind and cut at a specific recognition site, however the recognition sites for meganucleases are typically longer, about 18 bp or more (patent application WO-PCT PCT/US12/30061 filed on March 22, 2012). Meganucleases have been classified into four families based on conserved sequence motifs, the families are the LAGLIDADG, GIY-YIG, H-N-H, and His-Cys box families. These motifs participate in the coordination of metal ions and hydrolysis of phosphodiester bonds. HEases are notable for their long recognition sites, and for tolerating some sequence polymorphisms in their DNA substrates. The naming convention for meganuclease is similar to the convention for other restriction endonuclease. Meganucleases are also characterized by prefix F-, I-, or PI- for enzymes encoded by free-standing ORFs, introns, and inteins, respectively. One step in the recombination process involves polynucleotide cleavage at or near the recognition site. This cleaving activity can be used to produce a double-strand break. For

reviews of site-specific recombinases and their recognition sites, see, Sauer (1994) Curr Op Biotechnol 5:521-7; and Sadowski (1993) FASEB 7:760-7. In some examples the recombinase is from the Integrase or Resolvase families.

TAL effector nucleases are a new class of sequence-specific nucleases that can be used to make double-strand breaks at specific target sequences in the genome of a plant or other organism. (Miller *et al.* (2011) *Nature Biotechnology* 29:143–148). Zinc finger nucleases (ZFNs) are engineered double-strand break inducing agents comprised of a zinc finger DNA binding domain and a double-strand-break-inducing agent domain. Recognition site specificity is conferred by the zinc finger domain, which typically comprising two, three, or four zinc fingers, for example having a C2H2 structure, however other zinc finger structures are known and have been engineered. Zinc finger domains are amenable for designing polypeptides which specifically bind a selected polynucleotide recognition sequence. ZFNs include an engineered DNA-binding zinc finger domain linked to a non-specific endonuclease domain, for example nuclease domain from a Type IIs endonuclease such as FokI. Additional functionalities can be fused to the zinc-finger binding domain, including transcriptional activator domains, transcription repressor domains, and methylases. In some examples, dimerization of nuclease domain is required for cleavage activity. Each zinc finger recognizes three consecutive base pairs in the target DNA. For example, a 3 finger domain recognized a sequence of 9 contiguous nucleotides, with a dimerization requirement of the nuclease, two sets of zinc finger triplets are used to bind an 18 nucleotide recognition sequence.

As used herein, the term “guide polynucleotide”, relates to a polynucleotide sequence that can form a complex with a Cas endonuclease and enables the Cas endonuclease to recognize, bind to, and optionally cleave a DNA target site. The guide polynucleotide can be a single molecule or a double molecule. The guide polynucleotide sequence can be a RNA sequence, a DNA sequence, or a combination thereof (a RNA-DNA combination sequence). Optionally, the guide polynucleotide can comprise at least one nucleotide, phosphodiester bond or linkage modification such as, but not limited, to Locked Nucleic Acid (LNA), 5-methyl dC, 2,6-Diaminopurine, 2’-Fluoro A, 2’-Fluoro U, 2'-O-Methyl RNA, phosphorothioate

bond, linkage to a cholesterol molecule, linkage to a polyethylene glycol molecule, linkage to a spacer 18 (hexaethylene glycol chain) molecule, or 5' to 3' covalent linkage resulting in circularization. A guide polynucleotide that solely comprises ribonucleic acids is also referred to as a "guide RNA" or "gRNA" (See also U.S. Patent Application US 2015-0082478 A1, published on March 19, 2015 and US 2015-0059010 A1, published on February 26, 2015, both are hereby incorporated in its entirety by reference).

In one embodiment of the disclosure, the guide polynucleotide is a single guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said single guide RNA is selected from the group consisting of SEQ ID NOs: 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138 and 139.

In one embodiment of the disclosure, the guide polynucleotide is a single guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said single guide RNA comprises a chimeric non-naturally occurring crRNA linked to a tracrRNA, wherein said tracrRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183 and 184., wherein said chimeric non-naturally occurring crRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159 and 160.

In one embodiment of the disclosure, the guide polynucleotide is a guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said tracrRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183 and 184, wherein said chimeric non-naturally occurring crRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 149, 150, 151, 152,

153, 154, 155, 156, 157, 158, 159 and 160, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

The guide polynucleotide can be a double molecule (also referred to as duplex guide polynucleotide) comprising a crNucleotide sequence and a tracrNucleotide sequence. The crNucleotide includes a first nucleotide sequence domain (referred to as Variable Targeting domain or VT domain) that can hybridize to a nucleotide sequence in a target DNA and a second nucleotide sequence (also referred to as a tracr mate sequence) that is part of a Cas endonuclease recognition (CER) domain. The tracr mate sequence can hybridized to a tracrNucleotide along a region of complementarity and together form the Cas endonuclease recognition domain or CER domain. The CER domain is capable of interacting with a Cas endonuclease polypeptide. The crNucleotide and the tracrNucleotide of the duplex guide polynucleotide can be RNA, DNA, and/or RNA-DNA- combination sequences. In some embodiments, the crNucleotide molecule of the duplex guide polynucleotide is referred to as “crDNA” (when composed of a contiguous stretch of DNA nucleotides) or “crRNA” (when composed of a contiguous stretch of RNA nucleotides), or “crDNA-RNA” (when composed of a combination of DNA and RNA nucleotides). The crNucleotide can comprise a fragment of the crRNA naturally occurring in Bacteria and Archaea. The size of the fragment of the crRNA naturally occurring in Bacteria and Archaea that can be present in a crNucleotide disclosed herein can range from, but is not limited to, 2, 3, 4, 5, 6, 7, 8, 9,10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more nucleotides. In some embodiments the tracrNucleotide is referred to as “tracrRNA” (when composed of a contiguous stretch of RNA nucleotides) or “tracrDNA” (when composed of a contiguous stretch of DNA nucleotides) or “tracrDNA-RNA” (when composed of a combination of DNA and RNA nucleotides. In one embodiment, the RNA that guides the RNA/ Cas9 endonuclease complex is a duplexed RNA comprising a duplex crRNA-tracrRNA. The tracrRNA (trans-activating CRISPR RNA) contains, in the 5’-to-3’ direction, (i) a sequence that anneals with the repeat region of CRISPR type II crRNA and (ii) a stem loop-containing portion (Deltcheva et al., Nature 471:602-607). The duplex guide polynucleotide can form a complex with a Cas endonuclease, wherein said

guide polynucleotide/Cas endonuclease complex (also referred to as a guide polynucleotide/Cas endonuclease system) can direct the Cas endonuclease to a genomic target site, enabling the Cas endonuclease to recognize, bind to, and optionally nick or cleave (introduce a single or double strand break) into the target site. (See also U.S. Patent Application US 2015-0082478 A1, published on March 19, 2015 and US 2015-0059010 A1, published on February 26, 2015, both are hereby incorporated in its entirety by reference.)

The guide polynucleotide can also be a single molecule (also referred to as single guide polynucleotide) comprising a crNucleotide sequence linked to a tracrNucleotide sequence. The single guide polynucleotide comprises a first nucleotide sequence domain (referred to as Variable Targeting domain or VT domain) that can hybridize to a nucleotide sequence in a target DNA and a Cas endonuclease recognition domain (CER domain), that interacts with a Cas endonuclease polypeptide. By "domain" it is meant a contiguous stretch of nucleotides that can be RNA, DNA, and/or RNA-DNA-combination sequence. The VT domain and /or the CER domain of a single guide polynucleotide can comprise a RNA sequence, a DNA sequence, or a RNA-DNA-combination sequence. The single guide polynucleotide being comprised of sequences from the crNucleotide and the tracrNucleotide may be referred to as "single guide RNA" (when composed of a contiguous stretch of RNA nucleotides) or "single guide DNA" (when composed of a contiguous stretch of DNA nucleotides) or "single guide RNA-DNA" (when composed of a combination of RNA and DNA nucleotides). The single guide polynucleotide can form a complex with a Cas endonuclease, wherein said guide polynucleotide/Cas endonuclease complex (also referred to as a guide polynucleotide/Cas endonuclease system) can direct the Cas endonuclease to a genomic target site, enabling the Cas endonuclease to recognize, bind to, and optionally nick or cleave (introduce a single or double strand break) the target site. (See also U.S. Patent Application US 2015-0082478 A1, published on March 19, 2015 and US 2015-0059010 A1, published on February 26, 2015, both are hereby incorporated in its entirety by reference.)

The term "variable targeting domain" or "VT domain" is used interchangeably herein and includes a nucleotide sequence that can hybridize (is complementary) to

one strand (nucleotide sequence) of a double strand DNA target site. The % complementation between the first nucleotide sequence domain (VT domain) and the target sequence can be at least 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% or 100%. The variable target domain can be at least 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 or 30 nucleotides in length. In some embodiments, the variable targeting domain comprises a contiguous stretch of 12 to 30 nucleotides. The variable targeting domain can be composed of a DNA sequence, a RNA sequence, a modified DNA sequence, a modified RNA sequence, or any combination thereof.

The term “Cas endonuclease recognition domain” or “CER domain” (of a guide polynucleotide) is used interchangeably herein and includes a nucleotide sequence that interacts with a Cas endonuclease polypeptide. A CER domain comprises a tracrNucleotide mate sequence followed by a tracrNucleotide sequence. The CER domain can be composed of a DNA sequence, a RNA sequence, a modified DNA sequence, a modified RNA sequence (see for example US 2015-0059010 A1, published on February 26, 2015, incorporated in its entirety by reference herein), or any combination thereof.

The nucleotide sequence linking the crNucleotide and the tracrNucleotide of a single guide polynucleotide can comprise a RNA sequence, a DNA sequence, or a RNA-DNA combination sequence. In one embodiment, the nucleotide sequence linking the crNucleotide and the tracrNucleotide of a single guide polynucleotide can be at least 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 or 100 nucleotides in length. In another embodiment, the nucleotide sequence linking the crNucleotide and the tracrNucleotide of a single guide polynucleotide can comprise a tetraloop sequence, such as, but not limiting to a GAAA tetraloop sequence.

Nucleotide sequence modification of the guide polynucleotide, VT domain and/or CER domain can be selected from, but not limited to , the group consisting of a 5' cap, a 3' polyadenylated tail, a riboswitch sequence, a stability control sequence, a sequence that forms a dsRNA duplex, a modification or sequence that targets the guide poly nucleotide to a subcellular location, a modification or sequence that provides for tracking , a modification or sequence that provides a binding site for proteins , a Locked Nucleic Acid (LNA), a 5-methyl dC nucleotide, a 2,6-Diaminopurine nucleotide, a 2'-Fluoro A nucleotide, a 2'-Fluoro U nucleotide; a 2'-O-Methyl RNA nucleotide, a phosphorothioate bond, linkage to a cholesterol molecule, linkage to a polyethylene glycol molecule, linkage to a spacer 18 molecule, a 5' to 3' covalent linkage, or any combination thereof. These modifications can result in at least one additional beneficial feature, wherein the additional beneficial feature is selected from the group of a modified or regulated stability, a subcellular targeting, tracking, a fluorescent label, a binding site for a protein or protein complex, modified binding affinity to complementary target sequence, modified resistance to cellular degradation, and increased cellular permeability.

The terms "a polynucleotide originating from organism", "a polynucleotide derived from organism" are used interchangeably herein and refer to a polynucleotide (such as but not limited to crRNA and tracrRNA) that is naturally occurring in said organism (native to said organism) or is isolated from said organism, or is a synthetic oligonucleotide that is identical to the polynucleotide isolated from said organism). For example, a tracrRNA originating from *Brevibacillus laterosporus* refers to a tracrRNA that occurs in *Brevibacillus laterosporus*, or is isolated from *Brevibacillus laterosporus*, or is a synthetic oligonucleotide that is identical to the tracrRNA isolated from *Brevibacillus laterosporus*.

The terms "functional fragment ", "fragment that is functionally equivalent" and "functionally equivalent fragment" of a guide RNA, crRNA or tracrRNA are used interchangeably herein, and refer to a portion or subsequence of the guide RNA, crRNA or tracrRNA , respectively, of the present disclosure in which the ability to function as a guide RNA, crRNA or tracrRNA, respectively, is retained.

The terms “functional variant “, “Variant that is functionally equivalent” and “functionally equivalent variant” of a guide RNA, crRNA or tracrRNA (respectively) are used interchangeably herein, and refer to a variant of the guide RNA, crRNA or tracrRNA, respectively, of the present disclosure in which the ability to function as a guide RNA, crRNA or tracrRNA, respectively, is retained.

As used herein, the terms “single guide RNA" and “sgRNA” are used interchangeably herein and relate to a synthetic fusion of two RNA molecules, a crRNA (CRISPR RNA) comprising a variable targeting domain (linked to a tracr mate sequence that hybridizes to a tracrRNA), fused to a tracrRNA (trans-activating CRISPR RNA). The single guide RNA can comprise a crRNA or crRNA fragment and a tracrRNA or tracrRNA fragment of the type II CRISPR/Cas system that can form a complex with a type II Cas endonuclease, wherein said guide RNA/Cas endonuclease complex can direct the Cas endonuclease to a genomic target site, enabling the Cas endonuclease to t recognize, bind to, and optionally nick or cleave (introduce a single or double strand break) into a genomic target site.

The components of the single or dual guide polynucleotides described herein (such as but no limiting to the crRNA, tracrRNA, variable targeting domain, crRNA repeat, tracr-mate domain, loop, tracrRNA anti-repeat, 3’tracrRNA sequence) can be modified to create functional variants of these components such that these functional variants can be combined to create a functional single or dual guide polynucleotide. Examples of guide polynucleotide component modifications are described herein and include nucleotide extensions at the 3’ end , 5’ end, or both end of any of components of the guide polynucleotide, and/or nucleotide sequence modifications (substitutions, insertions, deletions), and /or chemical modifications, and/or linkage modifications, or any combinations thereof.

Extensions at 3’ end , 5’ end, or both ends of any of components of the guide polynucleotide can be can be at least 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 or 100 nucleotides in length.

Nucleotide sequence modification of the guide polynucleotide components include a 5' cap, a 3' polyadenylated tail, a riboswitch sequence, a stability control sequence, a sequence that forms a dsRNA duplex, a modification or sequence that targets the guide polynucleotide to a subcellular location, a modification or sequence that provides for tracking , a modification or sequence that provides a binding site for proteins , a Locked Nucleic Acid (LNA), a 5-methyl dC nucleotide, a 2,6-Diaminopurine nucleotide, a 2'-Fluoro A nucleotide, a 2'-Fluoro U nucleotide; a 2'-O-Methyl RNA nucleotide, a phosphorothioate bond, linkage to a cholesterol molecule, linkage to a polyethylene glycol molecule, linkage to a spacer 18 molecule, a 5' to 3' covalent linkage, or any combination thereof.

In one aspect, the functional variant single or dual guide polynucleotide has a similar activity than the guide polynucleotides of SEQ ID NOs : 127-139. In another aspect, the functional variant single or dual guide polynucleotide has an increased activity when compared to the guide polynucleotides of SEQ ID NOs : 127-139. The guide activity includes guide polynucleotide/Cas endonuclease ability to recognize, bind to and cleave a double strand break and/or RGEN mutation frequency.

The terms "guide RNA/Cas endonuclease complex", "guide RNA/Cas endonuclease system", " guide RNA/Cas complex", "guide RNA/Cas system", "gRNA/Cas complex", "gRNA/Cas system", "RNA-guided endonuclease" , "RGEN" are used interchangeably herein and refer to at least one RNA component and at least one Cas endonuclease that are capable of forming a complex , wherein said guide RNA/Cas endonuclease complex can direct the Cas endonuclease to a DNA target site, enabling the Cas endonuclease to recognize, bind to, and optionally nick or cleave (introduce a single or double strand break) the DNA target site. A guide RNA/Cas endonuclease complex herein can comprise Cas protein(s) and suitable RNA component(s) of any of the four known CRISPR systems (Horvath and Barrangou, Science 327:167-170) such as a type I, II, or III CRISPR system. A guide RNA/Cas endonuclease complex can comprise a Type II Cas9 endonuclease and at least one RNA component (e.g., a crRNA and tracrRNA, or a gRNA). (See also U.S. Patent Application US 2015-0082478 A1, published on March 19, 2015 and US 2015-0059010 A1, published on February 26, 2015, both are hereby incorporated in its entirety by reference).

The guide polynucleotide can be introduced into a cell transiently, as single stranded polynucleotide or a double stranded polynucleotide, using any method known in the art such as, but not limited to, particle bombardment, *Agrobacterium transformation* or topical applications. The guide polynucleotide can also be introduced indirectly into a cell by introducing a recombinant DNA molecule (via methods such as, but not limited to, particle bombardment or *Agrobacterium transformation*) comprising a heterologous nucleic acid fragment encoding a guide polynucleotide, operably linked to a specific promoter that is capable of transcribing the guide RNA in said cell. The specific promoter can be, but is not limited to, a RNA polymerase III promoter, which allow for transcription of RNA with precisely defined, unmodified, 5'- and 3'-ends (DiCarlo et al., Nucleic Acids Res. 41: 4336-4343; Ma et al., Mol. Ther. Nucleic Acids 3:e161).

The terms "target site", "target sequence", "target site sequence, "target DNA", "target locus", "genomic target site", "genomic target sequence", "genomic target locus" and "protospacer", are used interchangeably herein and refer to a polynucleotide sequence such as, but not limited to, a nucleotide sequence on a chromosome, episome, or any other DNA molecule in the genome (including chromosomal, choloroplastic, mitochondrial DNA, plasmid DNA) of a cell, at which a guide polynucleotide/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave . The target site can be an endogenous site in the genome of a cell, or alternatively, the target site can be heterologous to the cell and thereby not be naturally occurring in the genome of the cell, or the target site can be found in a heterologous genomic location compared to where it occurs in nature. As used herein, terms "endogenous target sequence" and "native target sequence" are used interchangeable herein to refer to a target sequence that is endogenous or native to the genome of a cell and is at the endogenous or native position of that target sequence in the genome of the cell. Cells include, but are not limited to, human, non-human, animal, bacterial, fungal, insect, yeast, non-conventional yeast, and plant cells as well as plants and seeds produced by the methods described herein.

An "artificial target site" or "artificial target sequence" are used interchangeably herein and refer to a target sequence that has been introduced into the genome of a cell. Such an artificial target sequence can be identical in sequence to an

endogenous or native target sequence in the genome of a cell but be located in a different position (*i.e.*, a non-endogenous or non-native position) in the genome of a cell.

An "altered target site", "altered target sequence", "modified target site", "modified target sequence" are used interchangeably herein and refer to a target sequence as disclosed herein that comprises at least one alteration when compared to non-altered target sequence. Such "alterations" include, for example: (i) replacement of at least one nucleotide, (ii) a deletion of at least one nucleotide, (iii) an insertion of at least one nucleotide, or (iv) any combination of (i) – (iii).

The length of the target DNA sequence (target site) can vary, and includes, for example, target sites that are at least 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more nucleotides in length. It is further possible that the target site can be palindromic, that is, the sequence on one strand reads the same in the opposite direction on the complementary strand. The nick/cleavage site can be within the target sequence or the nick/cleavage site could be outside of the target sequence. In another variation, the cleavage could occur at nucleotide positions immediately opposite each other to produce a blunt end cut or, in other Cases, the incisions could be staggered to produce single-stranded overhangs, also called "sticky ends", which can be either 5' overhangs, or 3' overhangs. Active variants of genomic target sites can also be used. Such active variants can comprise at least 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99% or more sequence identity to the given target site, wherein the active variants retain biological activity and hence are capable of being recognized and cleaved by an Cas endonuclease. Assays to measure the single or double-strand break of a target site by an endonuclease are known in the art and generally measure the overall activity and specificity of the agent on DNA substrates containing recognition sites.

A "protospacer adjacent motif" (PAM) herein refers to a short nucleotide sequence adjacent to a target sequence (protospacer) that is recognized (targeted) by a guide polynucleotide/Cas endonuclease system described herein. The Cas endonuclease may not successfully recognize a target DNA sequence if the target DNA sequence is not followed by a PAM sequence. The sequence and length of a

PAM herein can differ depending on the Cas protein or Cas protein complex used. The PAM sequence can be of any length but is typically 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 nucleotides long.

A "randomized PAM" and "randomized protospacer adjacent motif" are used interchangeably herein, and refer to a random DNA sequence adjacent to a target sequence (protospacer) that is recognized (targeted) by a guide polynucleotide/Cas endonuclease system described herein. The randomized PAM sequence can be of any length but is typically 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 nucleotides long. A randomized nucleotide includes anyone of the nucleotides A, C, G or T.

The PAM sequence plays a key role in target recognition by licensing crRNA-guided base pairing to the protospacer sequence (Szczelkun et al, 2014, Proc. Natl. Acad. Sci. U. S. A. 111: 9798–803). A strict PAM requirement constrains DNA target selection and poses a limit to Cas9 genome editing applications. Target site selection may be further confined if unique genomic sites are required especially in large complex plant genomes like maize (Xie et al, 2014, Mol. Plant 7: 923–6). These constraints imposed by the PAM and the specificity of the Spy Cas9 can be overcome by systematically redesigning the PAM specificity of a single Cas9 protein (Kleinstiver et al, 2015, Nature 523, 481–485. Described herein is a different method to overcome constraints imposed by the PAM and the specificity of the Cas9, namely by exploring the natural diversity of Cas9 proteins. The method described herein can also be combined with the method of systematically redesigning the PAM specificity to overcome constraints imposed by the PAM and the specificity of the Cas endonucleases.

Cas9 proteins from different bacteria recognize different PAM sequences (Horvath et al, 2008, J. Bacteriol. 190: 1401–12; Jinek et al, 2012, Science 337: 816–21; Gasiunas et al, 2012, Cell 154: 442 – 451; Zhang et al, 2013, Cell 50: 488–503; Fonfara et al, 2014, Nucleic Acids Res. 42: 2577–2590). Typically, the PAM sequences of new Cas9 proteins are identified by computational analysis of sequences immediately flanking putative protospacers in bacteriophage genomes (Shah et al, 2013, RNA Biol. 10: 1–9). Currently, with >1000 Cas9 protein orthologues available (Chylinski et al, 2014 Nucleic Acids Res. 42: 6091–6105; Hsu

et al, 2014, Cell 157: 1262–1278), most spacers in Type II CRISPR arrays show only a few if any matches to the phage sequences present in databases, indicating that the vast majority of the phage universe is still unexplored. This constrains computational PAM identification methods and hinders the exploration of Cas9 protein diversity for genome editing applications.

As described herein, to address this problem a method was developed to empirically examine the PAM sequence requirements for any Cas9 protein. The method is based on the analysis of the in vitro cleavage products of a plasmid DNA library which contains a fixed protospacer target sequence and a stretch of 5 or 7 randomized base pairs in the putative PAM region. Based on the methods described herein, the a stretch of randomized base pairs in the putative PAM region can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 , 19, or 20 base pairs.

Using the method described herein, the canonical PAM preferences for Cas9 proteins of S. pyogenes and S. thermophilus CRISPR1 and CRISPR3 systems were first confirmed. Next, the method described herein was applied to identify the PAM and guide RNA requirements for a novel Cas9 protein from the Type II CRISPR-Cas system of B*. laterosporus* SPP360D4. In the Type II system of *B. laterosporus*, the transcriptional direction of the tracrRNA and CRISPR region could not be reliably predicted by computational approaches. Therefore, two single guide RNA (sgRNA) variants for both possible sense and anti-sense expression scenarios of the tracrRNA and CRISPR array (Examples 5-8, 10-12 described herein) were synthesized and only one of the designed sgRNAs supported cleavage of the randomized PAM plasmid library by *B. laterosporus* Cas9. Deep sequencing analysis of the cleavage products revealed a novel PAM requirement for the B. laterosporus Cas9. One that requires a strong preference for a C residue at position 5 of the PAM sequence followed by moderate preferences for A residues at positions 7 and 8 with an overall PAM consensus of NNNNCNDD (N=G,C, A or T; D=A, G or T). With a strong preference for just a single nucleotide, B. laterosporus Cas9 provides a useful addition to the Cas9 genome editing toolbox.

To examine the genome editing potential of a novel Cas9 and sgRNA characterized with the method described herein, the *B. laterosporus* SPP360D4

Cas9 and sgRNA were tested in maize (Examples 5-8, 10-12, described herein). As a result of cleavage, imperfect DNA repair resulted in INDEL mutations at all 3 chromosomal sites tested with robust INDEL frequencies observed at 2 of the 3 sites. Interestingly, at one of the sites, a ~30% enhancement in the recovery of INDEL mutations was observed for the B. laterosporus Cas9 over the S. pyogenes Cas9 (Example 12).

In one embodiment described herein it is shown that cleavage of permissive PAMs is dependent on Cas9 concentration. For all Cas9 proteins analyzed, PAM sequences licensing plasmid DNA cleavage at higher (50 nM) Cas9 concentrations were more relaxed than PAM sequences identified at low (0.5 nM) Cas9 concentrations. This finding corroborates previous studies which demonstrated that lowering Cas9 concentration and shortening cleavage time prevents off-target cleavage by S. pyogenes Cas9 (Pattanayak et al, 2013, Nat. Biotechnol.: 1–7; Lin et al, 2014, Elife 3: e04766. doi: 10.7554/eLife.04766.). Since most other PAM determination methods have been performed in cells or cell extracts by expressing Cas9 at undefined concentrations (Ran et al, 2015, 2015 Apr 9;520(7546):186-91. doi: 10.1038/nature14299; Jiang et al, 2013, Nat. Biotechnol. 31: 233–9; Esvelt et al, 2013, Nov;10(11):1116-21. doi: 10.1038/nmeth.2681; Kleinstiver et al, 2015), our method further refines PAM specificity assessments by the dose-dependent control of recombinant Cas9 protein in vitro. This allows the careful detailed examination of Cas9 PAM specificity as a function of Cas9 guide RNA complex concentration.

In one embodiment, the method describes herein further refines Cas9 PAM discovery efforts by the use of recombinant Cas9 protein and reframes PAM specificity as being non-static and dependent on Cas9-guide RNA complex concentration.

Described herein are novel Cas endonucleases derived from diverse organisms capable or forming guide polynucleotide/Cas endonuclease complexes with guide polynucleotides comprising crRNA and tracrRNA sequences fragments derived from their respective organisms. In one example, a Cas endonuclease derived from *Brevibacillus laterosporus* (SEQ ID NO: 140) was able to from a RGEN complex with a guide polynucleotide comprising a crRNA and a tracrRNA fragment derived from *Brevibacillus laterosporus* (such as SEQ ID NO: 47 or 127).

The Cas endonucleases described herein can also be used in complexes with guide polynucleotides derived from other Cas systems. In one example, the crRNA and/or tracrRNA domains of a guide polynucleotide capable of forming a complex with a Cas endonuclease from organism 1 (such that said RGEN complex is capable of recognizing, binding to, and optionally nicking or cleaving all or part of a specific DNA target sequence), can be exchanged with a crRNA and /or tracrRNA domain, or fragment thereof, derived from a different organism (organism 2), thereby forming a chimeric guide, and still be able to form a functional complex with the Cas endonuclease derived from organism 1.

Similarities in guide RNA s between different Cas systems can be determined based on sequence composition and secondary structures of the guide RNAs. In one example, the secondary structure and sequence similarity of the sgRNAs from *Lactobacillus reuteri* Mlc3 (Lreu) (SEQ ID NO: 114), *Lactobacillus rossiae* DSM 15814 (Lros) SEQ ID NO: 115) and *Pediococcus pentosaceus* SL4 (Ppen) SEQ ID NO: 116) were determined and revealed that these three sgRNAs have very similar secondary structures. It is anticipated that fragments from Lreu, Lros and PPen guide RNAs, such as but not limited to repeat structures or anti-repeat structures or any-one guide RNA domain, can be exchanged and/or mixed with one another to create chimeric guides capable of forming a RGEN with any one of the Lrue, Lros or Ppen Cas endonuclease (SEQ ID NOs: 81. 82 and 93, respectively). In another example, the secondary structure and sequence similarity of the sgRNAs from *Lactobacillus nodensis* JCM 14932 (Lnod) (SEQ ID NO:117), *Loktanella vestfoldensis* (Lves) (SEQ ID NO:121) and *Sphingomonas sanxanigenens* NX02 (Ssan) (SEQ ID NO: 122) was determined to be very similar, indicating that fragments from Lnod, Lves and Ssan guide RNAs, such as but not limited to repeat structures or anti-repeat structures or any-one guide RNA domain, can be exchanged and/or mixed with one another to create chimeric guides capable of forming a RGEN with any one of the Lnod, Lves or Ssan Cas endonuclease (SEQ ID NOs: 84, 87 and 88, respectively).

In another example, the secondary structure and sequence similarity of the sgRNAs from *Epilithonimonas tenax* DSM 16811 (Eten) (SEQ ID NO:123), *Sporocytophaga myxococcoides* (Smyx) (SEQ ID NO:138) and *Psychroflexus*

torquis ATCC 700755 (Ptor) (SEQ ID NO: 139) was determined to be very similar, indicating that fragments from Eten, Smyx and Ptor guide RNAs, such as but not limited to repeat structures or anti-repeat structures or any-one guide RNA domain, can be exchanged and/or mixed with one another to create chimeric guides capable of forming a RGEN with any one of the Eten, Smyx or Ptor Cas endonuclease (SEQ ID NOs: 89, 90 and 91, respectively).

In one aspect, the Cas endonuclease and the crRNA and/or tracrRNA (or sgRNA) capable of forming a functional complex are derived or obtained from phylogenetically related groups. (See, for example, Fonfara et al Nucleic acid research 2014 Vol 42, No 4 pg. 2577-2590). It is understood that, based on the components of the novel Cas endonuclease systems described herein (crRNAs, tracrRNAs, Cas endonucleases, PAM sequences) one skilled in the art can exchange and/or mix any one component derived from one organism with any one component derived from another organism to make a functional guide polynucleotide/Cas endonuclease complex.

Guide polynucleotides can be modified to contain different sequence or structure yet be functionally equivalent or possess superior activity (binding, cutting, specificity). In one aspect, the chimeric guide polynucleotide can comprise at least one nucleotide, phosphodiester bond or linkage modification, or chemical modification such as, but not limited, to Locked Nucleic Acid (LNA), 5-methyl dC, 2,6-Diaminopurine, 2’-Fluoro A, 2’-Fluoro U, 2'-O-Methyl RNA, 2'-O-Methyl (M) modification, 2'-O-Methyl 3’phosphorothioate (MS) modification, 2'-O-Methyl 3’thioPACE (MSP) modification, phosphorothioate bond, linkage to a cholesterol molecule, linkage to a polyethylene glycol molecule, linkage to a spacer 18 (hexaethylene glycol chain) molecule, or 5’ to 3’ covalent linkage resulting in circularization (Hendel et al. 2015 Nature Biotechnology Vol. 33 pg. 985-991). Chimeric guide polynucleotides can be generated chemically, with or without sugar or backbone modifications. Chimeric guide polynucleotides can also be generated by in vitro transcription or delivered by DNA molecules containing promoters for expression

The PAM interacting domain, HNH or HNH-like nuclease domain, and / or RuvC or RuvC-like nuclease domains from the Cas endonuclease proteins

described herein find use for creating Cas scaffolds (US2016/0102324 entitled "New compact scaffold of Cas9 in the type II CRISPR system, published April 14, 2016 and incorporated herein by reference). The boundaries of the PAM interacting domain, RuvC and HNH domains of the Cas endonuclease described herein can be determined and new shorter Cas endonucleases derived from the Cas endonucleases described herein (or any one functional combination /fusion protein thereof) can be designed,

The terms "targeting", "gene targeting" and "DNA targeting" are used interchangeably herein. DNA targeting herein may be the specific introduction of a knock-out, edit, or knock-in at a particular DNA sequence, such as in a chromosome or plasmid of a cell. In general, DNA targeting can be performed herein by cleaving one or both strands at a specific DNA sequence in a cell with a Cas protein associated with a suitable polynucleotide component. Such DNA cleavage, if a double-strand break (DSB), can prompt NHEJ or HDR processes which can lead to modifications at the target site.

The terms "knock-out", "gene knock-out" and "genetic knock-out" are used interchangeably herein. A knock-out represents a DNA sequence of a cell that has been rendered partially or completely inoperative by targeting with a Cas protein; such a DNA sequence prior to knock-out could have encoded an amino acid sequence, or could have had a regulatory function (e.g., promoter), for example. A knock-out may be produced by an indel (insertion or deletion of nucleotide bases in a target DNA sequence through NHEJ), or by specific removal of sequence that reduces or completely destroys the function of sequence at or near the targeting site.

In one embodiment of the disclosure, the method comprises a method for modifying a target site in the genome of a cell, the method comprising providing to said cell at least one Cas9 endonuclease originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and

Psychroflexus torquis ATCC 700755, and at least one guide RNA, wherein said guide RNA and Cas endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site. The embodiment can further comprise identifying at least one cell that has a modification at said target, wherein the modification at said target site is selected from the group consisting of (i) a replacement of at least one nucleotide, (ii) a deletion of at least one nucleotide, (iii) an insertion of at least one nucleotide, and (iv) any combination of (i) – (iii).

The guide polynucleotide/Cas endonuclease system can be used in combination with a co-delivered polynucleotide modification template to allow for editing (modification) of a genomic nucleotide sequence of interest. (See also U.S. Patent Application US 2015-0082478 A1, published on March 19, 2015 and WO2015/026886 A1, published on February 26, 2015, both are hereby incorporated in its entirety by reference.)

A “modified nucleotide” or “edited nucleotide” refers to a nucleotide sequence of interest that comprises at least one alteration when compared to its non-modified nucleotide sequence. Such “alterations” include, for example: (i) replacement of at least one nucleotide, (ii) a deletion of at least one nucleotide, (iii) an insertion of at least one nucleotide, or (iv) any combination of (i) – (iii).

The term “polynucleotide modification template” includes a polynucleotide that comprises at least one nucleotide modification when compared to the nucleotide sequence to be edited. A nucleotide modification can be at least one nucleotide substitution, addition or deletion. Optionally, the polynucleotide modification template can further comprise homologous nucleotide sequences flanking the at least one nucleotide modification, wherein the flanking homologous nucleotide sequences provide sufficient homology to the desired nucleotide sequence to be edited.

In one embodiment, the disclosure describes a method for editing a nucleotide sequence in the genome of a cell, the method comprising providing a guide polynucleotide, a polynucleotide modification template, and at least one Cas endonuclease to a cell, wherein the Cas endonuclease is capable of introducing a single or double-strand break at a target sequence in the genome of said cell,

wherein said polynucleotide modification template includes at least one nucleotide modification of said nucleotide sequence. Cells include, but are not limited to, human, non-human, animal, bacterial, fungal, insect, yeast, and plant cells as well as plants and seeds produced by the methods described herein. Plant cells include cells selected from the group consisting of maize, rice, sorghum, rye, barley, wheat, millet, oats, sugarcane, turfgrass, or switchgrass, soybean, canola, alfalfa, sunflower, cotton, tobacco, peanut, potato, tobacco, *Arabidopsis*, and safflower cells. The nucleotide to be edited can be located within or outside a target site recognized and cleaved by a Cas endonuclease. In one embodiment, the at least one nucleotide modification is not a modification at a target site recognized and cleaved by a Cas endonuclease. In another embodiment, there are at least 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 or 1000 nucleotides between the at least one nucleotide to be edited and the genomic target site.

In one embodiment of the disclosure, the method comprises a method for editing a nucleotide sequence in the genome of a cell, the method comprising providing to said cell at least one Cas9 endonuclease originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, a polynucleotide modification template, and at least one guide RNA, wherein said polynucleotide modification template comprises at least one nucleotide modification of said nucleotide sequence, wherein said guide RNA and Cas endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site. Cells include, but are not limited to, human, non-human, animal, bacterial, fungal, insect, yeast, and plant cells as well as plants and seeds produced by the methods described herein.

Genome editing can be accomplished using any method of gene editing available. For example, gene editing can be accomplished through the introduction

into a host cell of a polynucleotide modification template (sometimes also referred to as a gene repair oligonucleotide) containing a targeted modification to a gene within the genome of the host cell. The polynucleotide modification template for use in such methods can be either single-stranded or double-stranded. Examples of such methods are generally described, for example, in US Publication No. 2013/0019349.

In some embodiments, gene editing may be facilitated through the induction of a double-stranded break (DSB) in a defined position in the genome near the desired alteration. DSBs can be induced using any DSB-inducing agent available, including, but not limited to, TALENs, meganucleases, zinc finger nucleases, Cas9-gRNA systems (based on bacterial CRISPR-Cas systems), and the like. In some embodiments, the introduction of a DSB can be combined with the introduction of a polynucleotide modification template.

The process for editing a genomic sequence combining DSB and modification templates generally comprises: providing to a host cell, a DSB-inducing agent, or a nucleic acid encoding a DSB-inducing agent, that recognizes a target sequence in the chromosomal sequence and is able to induce a DSB in the genomic sequence, and at least one polynucleotide modification template comprising at least one nucleotide alteration when compared to the nucleotide sequence to be edited. The polynucleotide modification template can further comprise nucleotide sequences flanking the at least one nucleotide alteration, in which the flanking sequences are substantially homologous to the chromosomal region flanking the DSB. Genome editing using DSB-inducing agents, such as Cas9-gRNA complexes, has been described, for example in U.S. Patent Application US 2015-0082478 A1, published on March 19, 2015, WO2015/026886 A1, published on February 26, 2015, US application 62/023246, filed on July 07, 2014, and US application 62/036,652, filed on August 13, 2014, all of which are incorporated by reference herein.

The terms “knock-in”, “gene knock-in , “gene insertion” and “genetic knock-in” are used interchangeably herein. A knock-in represents the replacement or insertion of a DNA sequence at a specific DNA sequence in cell by targeting with a Cas protein (by HR, wherein a suitable donor DNA polynucleotide is also used). Examples of knock-ins are a specific insertion of a heterologous amino acid coding

sequence in a coding region of a gene, or a specific insertion of a transcriptional regulatory element in a genetic locus.

Various methods and compositions can be employed to obtain a cell or organism having a polynucleotide of interest inserted in a target site for a Cas endonuclease. Such methods can employ homologous recombination to provide integration of the polynucleotide of Interest at the target site. In one method provided, a polynucleotide of interest is provided to the organism cell in a donor DNA construct. As used herein, "donor DNA" is a DNA construct that comprises a polynucleotide of Interest to be inserted into the target site of a Cas endonuclease. The donor DNA construct further comprises a first and a second region of homology that flank the polynucleotide of Interest. The first and second regions of homology of the donor DNA share homology to a first and a second genomic region, respectively, present in or flanking the target site of the cell or organism genome. By "homology" is meant DNA sequences that are similar. For example, a "region of homology to a genomic region" that is found on the donor DNA is a region of DNA that has a similar sequence to a given "genomic region" in the cell or organism genome. A region of homology can be of any length that is sufficient to promote homologous recombination at the cleaved target site. For example, the region of homology can comprise at least 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 or more bases in length such that the region of homology has sufficient homology to undergo homologous recombination with the corresponding genomic region. "Sufficient homology" indicates that two polynucleotide sequences have sufficient structural similarity to act as substrates for a homologous recombination reaction. The structural similarity includes overall length of each polynucleotide fragment, as well as the sequence similarity of the polynucleotides. Sequence similarity can be described by the percent sequence identity over the whole length of the sequences, and/or by conserved regions comprising localized similarities such as contiguous nucleotides

having 100% sequence identity, and percent sequence identity over a portion of the length of the sequences.

The amount of homology or sequence identity shared by a target and a donor polynucleotide can vary and includes total lengths and/or regions having unit integral values in the ranges of about 1-20 bp, 20-50 bp, 50-100 bp, 75-150 bp, 100-250 bp, 150-300 bp, 200-400 bp, 250-500 bp, 300-600 bp, 350-750 bp, 400-800 bp, 450-900 bp, 500-1000 bp, 600-1250 bp, 700-1500 bp, 800-1750 bp, 900-2000 bp, 1-2.5 kb, 1.5–3 kb, 2-4 kb, 2.5-5 kb, 3-6 kb, 3.5-7 kb, 4-8 kb, 5-10 kb, or up to and including the total length of the target site. These ranges include every integer within the range, for example, the range of 1-20 bp includes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20 bps. The amount of homology can also described by percent sequence identity over the full aligned length of the two polynucleotides which includes percent sequence identity of about at least 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% or 100%. Sufficient homology includes any combination of polynucleotide length, global percent sequence identity, and optionally conserved regions of contiguous nucleotides or local percent sequence identity, for example sufficient homology can be described as a region of 75-150 bp having at least 80% sequence identity to a region of the target locus. Sufficient homology can also be described by the predicted ability of two polynucleotides to specifically hybridize under high stringency conditions, see, for example, Sambrook *et al.*, (1989) *Molecular Cloning: A Laboratory Manual*, (Cold Spring Harbor Laboratory Press, NY); *Current Protocols in Molecular Biology*, Ausubel *et al.*, Eds (1994) Current Protocols, (Greene Publishing Associates, Inc. and John Wiley & Sons, Inc.); and, Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes*, (Elsevier, New York).

In one embodiment of the disclosure, the method comprises a method for modifying a target site in the genome of a cell, the method comprising providing to said cell at least one guide RNA, at least one donor DNA, and at least one Cas9 endonuclease originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* MIc3, *Lactobacillus rossiae* DSM

15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, wherein said at least one guide RNA and at least one Cas endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site, wherein said donor DNA comprises a polynucleotide of interest. Cells include, but are not limited to, human, non-human, animal, bacterial, fungal, insect, yeast, and plant cells as well as plants and seeds produced by the methods described herein. The embodiment can further comprise, identifying at least one cell that said polynucleotide of interest integrated in or near said target site.

As used herein, a “genomic region” is a segment of a chromosome in the genome of a cell that is present on either side of the target site or, alternatively, also comprises a portion of the target site. The genomic region can comprise at least 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 or more bases such that the genomic region has sufficient homology to undergo homologous recombination with the corresponding region of homology.

Polynucleotides of interest and/or traits can be stacked together in a complex trait locus as described in US-2013-0263324-A1, published 03 Oct 2013 and in PCT/US13/22891, published January 24, 2013, both applications are hereby incorporated by reference. The guide polynucleotide/Cas9 endonuclease system described herein provides for an efficient system to generate double strand breaks and allows for traits to be stacked in a complex trait locus.

The guide polynucleotide/Cas endonuclease system can be used for introducing one or more polynucleotides of interest or one or more traits of interest into one or more target sites by providing one or more guide polynucleotides, one Cas endonuclease, and optionally one or more donor DNAs to a plant cell. ((as described in US patent application No. 14/463,687, file August 20, 2014,

incorporated by reference herein). A fertile plant can be produced from that plant cell that comprises an alteration at said one or more target sites, wherein the alteration is selected from the group consisting of (i) replacement of at least one nucleotide, (ii) a deletion of at least one nucleotide, (iii) an insertion of at least one nucleotide, and (iv) any combination of (i) – (iii). Plants comprising these altered target sites can be crossed with plants comprising at least one gene or trait of interest in the same complex trait locus, thereby further stacking traits in said complex trait locus. (see also US-2013-0263324-A1, published 03 Oct 2013 and in PCT/US13/22891, published January 24, 2013).

The structural similarity between a given genomic region and the corresponding region of homology found on the donor DNA can be any degree of sequence identity that allows for homologous recombination to occur. For example, the amount of homology or sequence identity shared by the "region of homology" of the donor DNA and the "genomic region" of the organism genome can be at least 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% or 100% sequence identity, such that the sequences undergo homologous recombination

The region of homology on the donor DNA can have homology to any sequence flanking the target site. While in some embodiments the regions of homology share significant sequence homology to the genomic sequence immediately flanking the target site, it is recognized that the regions of homology can be designed to have sufficient homology to regions that may be further 5' or 3' to the target site. In still other embodiments, the regions of homology can also have homology with a fragment of the target site along with downstream genomic regions. In one embodiment, the first region of homology further comprises a first fragment of the target site and the second region of homology comprises a second fragment of the target site, wherein the first and second fragments are dissimilar.

As used herein, "homologous recombination" includes the exchange of DNA fragments between two DNA molecules at the sites of homology. The frequency of homologous recombination is influenced by a number of factors. Different organisms vary with respect to the amount of homologous recombination and the relative proportion of homologous to non-homologous recombination. Generally, the

length of the region of homology affects the frequency of homologous recombination events: the longer the region of homology, the greater the frequency. The length of the homology region needed to observe homologous recombination is also species-variable. In many cases, at least 5 kb of homology has been utilized, but homologous recombination has been observed with as little as 25-50 bp of homology. See, for example, Singer et al., (1982) Cell 31:25-33; Shen and Huang, (1986) Genetics 112:441-57; Watt et al., (1985) Proc. Natl. Acad. Sci. USA 82:4768-72, Sugawara and Haber, (1992) Mol Cell Biol 12:563-75, Rubnitz and Subramani, (1984) Mol Cell Biol 4:2253-8; Ayares et al., (1986) Proc. Natl. Acad. Sci. USA 83:5199-203; Liskay et al., (1987) Genetics 115:161-7.

Homology-directed repair (HDR) is a mechanism in cells to repair double-stranded and single stranded DNA breaks. Homology-directed repair includes homologous recombination (HR) and single-strand annealing (SSA) (Lieber. 2010 Annu. Rev. Biochem . 79:181-211). The most common form of HDR is called homologous recombination (HR), which has the longest sequence homology requirements between the donor and acceptor DNA. Other forms of HDR include single-stranded annealing (SSA) and breakage-induced replication, and these require shorter sequence homology relative to HR. Homology-directed repair at nicks (single-stranded breaks) can occur via a mechanism distinct from HDR at double-strand breaks (Davis and Maizels. PNAS (0027-8424), 111 (10), p. E924-E932.

Alteration of the genome of a plant cell, for example, through homologous recombination (HR), is a powerful tool for genetic engineering. Despite the low frequency of homologous recombination in higher plants, there are a few examples of successful homologous recombination of plant endogenous genes. The parameters for homologous recombination in plants have primarily been investigated by rescuing introduced truncated selectable marker genes. In these experiments, the homologous DNA fragments were typically between 0.3 kb to 2 kb. Observed frequencies for homologous recombination were on the order of 10^{-4} to 10^{-5}. See, for example, Halfter et al., (1992) Mol Gen Genet 231:186-93; Offringa et al., (1990) EMBO J 9:3077-84; Offringa et al., (1993) Proc. Natl. Acad. Sci. USA

90:7346-50; Paszkowski et al., (1988) EMBO J 7:4021-6; Hourda and Paszkowski, (1994) Mol Gen Genet 243:106-11; and Risseeuw et al., (1995) Plant J 7:109-19.

Homologous recombination has been demonstrated in insects. In Drosophila, Dray and Gloor found that as little as 3 kb of total template:target homology sufficed to copy a large non-homologous segment of DNA into the target with reasonable efficiency (Dray and Gloor, (1997) Genetics 147:689-99). Using FLP-mediated DNA integration at a target FRT in Drosophila, Golic et al., showed integration was approximately 10-fold more efficient when the donor and target shared 4.1 kb of homology as compared to 1.1 kb of homology (Golic et al., (1997) Nucleic Acids Res 25:3665). Data from Drosophila indicates that 2-4 kb of homology is sufficient for efficient targeting, but there is some evidence that much less homology may suffice, on the order of about 30 bp to about 100 bp (Nassif and Engels, (1993) Proc. Natl. Acad. Sci. USA 90:1262-6; Keeler and Gloor, (1997) Mol Cell Biol 17:627-34).

Homologous recombination has also been accomplished in other organisms. For example, at least 150-200 bp of homology was required for homologous recombination in the parasitic protozoan Leishmania (Papadopoulou and Dumas, (1997) Nucleic Acids Res 25:4278-86). In the filamentous fungus Aspergillus nidulans, gene replacement has been accomplished with as little as 50 bp flanking homology (Chaveroche et al., (2000) Nucleic Acids Res 28:e97). Targeted gene replacement has also been demonstrated in the ciliate Tetrahymena thermophila (Gaertig et al., (1994) Nucleic Acids Res 22:5391-8). In mammals, homologous recombination has been most successful in the mouse using pluripotent embryonic stem cell lines (ES) that can be grown in culture, transformed, selected and introduced into a mouse embryo. Embryos bearing inserted transgenic ES cells develop as genetically offspring. By interbreeding siblings, homozygous mice carrying the selected genes can be obtained. An overview of the process is provided in Watson et al., (1992) Recombinant DNA, 2nd Ed., (Scientific American Books distributed by WH Freeman & Co.); Capecchi, (1989) Trends Genet 5:70-6; and Bronson, (1994) J Biol Chem 269:27155-8. Homologous recombination in mammals other than mouse has been limited by the lack of stem cells capable of being transplanted to oocytes or developing embryos. However, McCreath et al.,

Nature 405:1066-9 (2000) reported successful homologous recombination in sheep by transformation and selection in primary embryo fibroblast cells.

Error-prone DNA repair mechanisms can produce mutations at double-strand break sites. The Non-Homologous-End-Joining (NHEJ) pathways are the most common repair mechanism to bring the broken ends together (Bleuyard et al., (2006) DNA Repair 5:1-12). The structural integrity of chromosomes is typically preserved by the repair, but deletions, insertions, or other rearrangements are possible. The two ends of one double-strand break are the most prevalent substrates of NHEJ (Kirik et al., (2000) EMBO J 19:5562-6), however if two different double-strand breaks occur, the free ends from different breaks can be ligated and result in chromosomal deletions (Siebert and Puchta, (2002) Plant Cell 14:1121-31), or chromosomal translocations between different chromosomes (Pacher et al., (2007) Genetics 175:21-9).

Episomal DNA molecules can also be ligated into the double-strand break, for example, integration of T-DNAs into chromosomal double-strand breaks (Chilton and Que, (2003) Plant Physiol 133:956-65; Salomon and Puchta, (1998) EMBO J 17:6086-95). Once the sequence around the double-strand breaks is altered, for example, by exonuclease activities involved in the maturation of double-strand breaks, gene conversion pathways can restore the original structure if a homologous sequence is available, such as a homologous chromosome in non-dividing somatic cells, or a sister chromatid after DNA replication (Molinier et al., (2004) Plant Cell 16:342-52). Ectopic and/or epigenic DNA sequences may also serve as a DNA repair template for homologous recombination (Puchta, (1999) Genetics 152:1173-81).

Once a double-strand break is induced in the DNA, the cell's DNA repair mechanism is activated to repair the break. Error-prone DNA repair mechanisms can produce mutations at double-strand break sites. The most common repair mechanism to bring the broken ends together is the nonhomologous end-joining (NHEJ) pathway (Bleuyard et al., (2006) DNA Repair 5:1-12). The structural integrity of chromosomes is typically preserved by the repair, but deletions, insertions, or other rearrangements are possible (Siebert and Puchta, (2002) Plant Cell 14:1121-31; Pacher et al., (2007) Genetics 175:21-9).

Alternatively, the double-strand break can be repaired by homologous recombination between homologous DNA sequences. Once the sequence around the double-strand break is altered, for example, by exonuclease activities involved in the maturation of double-strand breaks, gene conversion pathways can restore the original structure if a homologous sequence is available, such as a homologous chromosome in non-dividing somatic cells, or a sister chromatid after DNA replication (Molinier et al., (2004) Plant Cell 16:342-52). Ectopic and/or epigenic DNA sequences may also serve as a DNA repair template for homologous recombination (Puchta, (1999) Genetics 152:1173-81).

DNA double-strand breaks appear to be an effective factor to stimulate homologous recombination pathways (Puchta et al., (1995) Plant Mol Biol 28:281-92; Tzfira and White, (2005) Trends Biotechnol 23:567-9; Puchta, (2005) J Exp Bot 56:1-14). Using DNA-breaking agents, a two- to nine-fold increase of homologous recombination was observed between artificially constructed homologous DNA repeats in plants (Puchta et al., (1995) Plant Mol Biol 28:281-92). In maize protoplasts, experiments with linear DNA molecules demonstrated enhanced homologous recombination between plasmids (Lyznik et al., (1991) Mol Gen Genet 230:209-18).

The donor DNA may be introduced by any means known in the art. For example, a plant having a target site is provided. The donor DNA may be provided by any transformation method known in the art including, for example, Agrobacterium-mediated transformation or biolistic particle bombardment. The donor DNA may be present transiently in the cell or it could be introduced via a viral replicon. In the presence of the Cas endonuclease and the target site, the donor DNA is inserted into the transformed plant's genome.

Further uses for guide RNA/Cas endonuclease systems have been described (See U.S. Patent Application US 2015-0082478 A1, published on March 19, 2015, WO2015/026886 A1, published on February 26, 2015, US 2015-0059010 A1, published on February 26, 2015, US application 62/023246, filed on July 07, 2014, and US application 62/036,652, filed on August 13, 2014, all of which are incorporated by reference herein) and include but are not limited to modifying or replacing nucleotide sequences of interest (such as a regulatory elements),

insertion of polynucleotides of interest, gene knock-out, gene-knock in, modification of splicing sites and/or introducing alternate splicing sites, modifications of nucleotide sequences encoding a protein of interest, amino acid and/or protein fusions, and gene silencing by expressing an inverted repeat into a gene of interest.

Given the diversity of Type II CRISPR-Cas systems (Fonfara et al. (2014) Nucleic Acids Res. 42:2577-2590), it is plausible that many of the Cas9 endonucleases and cognate guide RNAs may have unique sequence recognition and enzymatic properties different from those previously described or characterized. For example, cleavage activity and specificity may be enhanced or proto-spacer adjacent motif (PAM) sequence may be different leading to increased genomic target site density. To tap into this vast unexplored diversity and expand the repertoire of Cas9 endonucleases and cognate guide RNAs available for genome targeting, two components of target site recognition need to be cooperatively characterized for each new system, the PAM sequence and the guide RNA (either duplexed CRISPR RNA (crRNA) and trans-activating CRISPR RNA (tracrRNA) or chimeric fusion of crRNA and tracrRNA (single guide RNA (sgRNA. Rapid in vitro methods described herein have been developed to concertedly characterize both the guide RNA and PAM sequence of Type II Cas9 proteins.

Methods for assaying Cas9 PAM preferences have been described herein (see Example 3, Example 4 and Example 7). In one embodiment, the Cas9 endonuclease PAM preferences was assayed in a dose dependent manner by subjecting the randomized PAM libraries described herein to in vitro digestion with different concentrations of recombinant Cas9 protein preloaded with guide RNA. After digestion with Cas9-guide RNA ribonucleoprotein (RNP) complexes, PAM sequence combinations from the randomized PAM library that supported cleavage were captured by ligating adapters to the free-ends of the plasmid DNA molecules cleaved by the Cas9-guide RNA complex (Figure 3). To promote efficient ligation and capture of the cleaved ends, the typically blunt-ended double-stranded DNA cut generated by Cas9 endonucleases was modified to contain a 3' dA overhang and adapters were modified to contain a complementary 3'dT overhang. To generate sufficient quantities of DNA for sequencing, DNA fragments harboring the PAM sequence supporting cleavage were PCR amplified using a primer in the adapter

and another directly adjacent to the PAM region. The resulting PCR amplified Cas9 PAM libraries were converted into ampli-seq templates and single-read deep sequenced from the adapter-side of the amplicon. To ensure adequate coverage, the Cas9 PAM libraries were sequenced to a depth at least 5 times greater than the diversity in the initial randomized PAM library (5,120 and 81,920 reads for the 5 and 7 bp PAM randomized libraries, respectively). PAM sequences were identified from the resulting sequence data by only selecting those reads containing a 12 nt sequence match flanking either side of the 5 or 7 nt PAM sequence (depending on the randomized PAM library used); capturing only those PAM sequences resulting from perfect Cas9-guide RNA target site recognition and cleavage. To compensate for the inherent bias in the initial randomized PAM libraries, the frequency of each PAM sequence was normalized to its frequency in the starting library. The composition of the resulting PAM sequences can then be examined using a position frequency matrix (PFM) (Stormo, 2013 Quant. Biol. 1: 115–130)

As described herein, to validate the randomness of the PAM library disclosed herein (PAM library validation), PCR fragments spanning the 5 bp and 7 bp randomized PAM regions were generated by Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific) amplification (15 cycles of a 2-step amplification protocol) using the primer pair combinations TK-119/pUC-dir and TK-113/pUC-dir (SEQ ID NO: 175/ SEQ ID NO:5) for the 5 bp and 7bp libraries, respectively. The resulting 145 bp PCR product was purified using GeneJET PCR Purification Kit (Thermo Fisher Scientific) and the sequences necessary for amplicon-specific barcodes and Illumina sequencing were “tailed” on through two rounds of PCR each consisting of 10 cycles. In some examples, the primer pair combinations in the first round of PCR were JKYS800.1/JKYS803 and JKYS921.1 (SEQ ID NO:176)/JKYS812 (SEQ ID NO: 32) for the 5 bp and 7 bp libraries, respectively. A set of primers, JKYS557 (SEQ ID NO: 177) /JKYS558 (SEQ ID NO: 178), universal to all primary PCR reactions were utilized for the secondary PCR amplification. The resulting PCR amplifications were purified with a Qiagen PCR purification spin column, concentration measured with a Hoechst dye-based fluorometric assay, combined in an equimolar ratio and single read 60-100 nucleotide-length deep sequencing was performed on Illumina’s MiSeq Personal

Sequencer with a 5-10% (v/v) spike of PhiX control v3 (Illumina, FC-110-3001) to off-set sequence bias. The PAM sequence for only those reads containing a perfect 12 nt sequence match flanking either side of the randomized PAM sequence were captured and used to examine the frequency and diversity of PAM sequences present in the library.

In one embodiment of the disclosure, the method comprises a method for producing a plasmid DNA library containing a randomized Protospacer-Adjacent-Motif (PAM) sequence, the method comprising: a) providing a first single stranded oligonucleotide comprising a target sequence that can be recognized by a guide RNA/Cas endonuclease complex; b) providing a second single stranded oligonucleotide comprising a randomized PAM sequence adjacent to a nucleotide sequence capable of hybridizing with the target sequence of (a);c) producing an oligoduplex comprising said randomized PAM sequence by combining the first single stranded oligonucleotide of (a) and the second single stranded oligonucleotide of (b);d) producing a ligation product by ligating the oligoduplex from (c) with a linearized plasmid; and, e) transforming host cells with the ligation product of (e) and recovering multiple host cell colonies representing the plasmid library.

Host cells include, but are not limited to, human, non-human, animal, bacterial, fungal, insect, yeast, non-conventional yeast, and plant cells. One skilled in the art can ligate the oligoduplex of (c) directly into a linearized vector without restriction enzyme digestion, or can use two restriction enzyme sites, one upstream (5') and one downstream (3') of the target site. The first single stranded oligonucleotide can comprise a restriction endonuclease recognition site located upstream of a target sequence and the ligation product of (d) is produced by first cleaving the oligoduplex with a restriction endonuclease that recognizes the restriction endonuclease recognition site of (a) followed by ligating the cleaved oligoduplex from (d) with a linearized plasmid.

In one embodiment of the disclosure, the method comprises a method for producing a plasmid DNA library containing a randomized Protospacer-Adjacent-Motif (PAM) sequence, the method comprising transforming at least one host cell with a ligation product and recovering multiple host cell colonies representing the plasmid library, wherein said ligation product was generated by contacting a library

of linear oligoduplexes with a linearized plasmid, wherein each oligoduplex member of said library of oligoduplexes comprises a first single stranded oligonucleotide comprising a-target sequence , and a second single stranded oligonucleotide comprising a randomized PAM sequence adjacent to a nucleotide sequence capable of hybridizing with said target sequence. One skilled in the art can ligate the oligoduplex of (c) directly into a linearized vector without restriction enzyme digestion, or can use two restriction enzyme sites, one upstream (5') and one downstream (3') of the target site.

In one embodiment of the disclosure, the method comprises a method for producing a ligation product containing a randomized Protospacer-Adjacent-Motif (PAM) sequence, the method comprising: a) providing a first single stranded oligonucleotide comprising restriction endonuclease recognition site located upstream of a target sequence that can be recognized by a guide RNA/Cas endonuclease complex; b) providing a second single stranded oligonucleotide comprising a randomized PAM sequence adjacent a nucleotide sequence capable of hybridizing with the target sequence of (a); c) producing an oligoduplex comprising said randomized PAM sequence by combining the first single stranded oligonucleotide of (a) and the second single stranded oligonucleotide of (b); and, d) producing a ligation product by ligating the oligoduplex from (c) with a linearized plasmid.

In one embodiment of the disclosure, the method comprises a method for identification of a Protospacer-Adjacent-Motif (PAM) sequence, the method comprising: a)providing a library of plasmid DNAs, wherein each one of said plasmid DNAs comprises a randomized Protospacer-Adjacent-Motif sequence integrated adjacent to a target sequence that can be recognized by a guide RNA/Cas endonuclease complex; b) providing to said library of plasmids a guide RNA and a Cas endonuclease protein, wherein said guide RNA and Cas endonuclease protein can form a complex that is capable of introducing a double strand break into the said target sequence, thereby creating a library of cleaved targets; c) ligating adaptors to the library of cleaved targets of (b) allowing for the library of cleaved targets to be amplified; d) amplifying the library of cleaved targets such that cleaved products containing the randomized PAM sequence are enriched, thereby producing

a library of enriched PAM-sided targets; e) sequencing the library of (a) and the library of enriched PAM-sided targets of (d) and identifying the nucleotide sequence adjacent to the cleaved targets of (b) on either strand of the plasmid DNA, wherein said nucleotide sequence represents a putative Protospacer-Adjacent-Motif sequences; and, f) determining the fold enrichment of each nucleotide within the putative Protospacer-Adjacent-Motif sequence relative to the plasmid DNA library of (a).

The randomized PAM libraries described herein can also be used in combination with immunoprecipitation then sequencing approach using dCAS9 for further PAM discovery. The randomized PAM libraries can also be put on a microchip followed by cleaving the chip-array library. The randomized PAM libraries described herein can also be used in combination with Phage-display as a method to identify PAMs. (Isalan,M., Klug,A. and Choo,Y. (2001) A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 120 promoter. Nat. Biotechnol., 19, 656–660.; Dreier,B., Fuller,R.P., Segal,D.J., Lund,C., Blancafort,P., Huber,A., Koksch,B. and Barbas,C.F.,III (2005) Development of zinc finger domains for recognition of the 50 -CNN-30 family DNA sequences and their use in the construction of artificial transcription factors. J. Biol. 125 Chem., 280, 35588–35597).

In one embodiment of the disclosure, the method comprises a method for identification of a tracrRNA of an organism, the method comprising: a) providing a first single guide RNA candidate comprising a chimeric non-naturally occurring crRNA comprising a variable targeting domain capable of hybridizing to a target sequence in the genome of a cell, linked to a first nucleotide sequence representing the sense expression of a candidate tracrRNA naturally occurring in said organism ; b) providing a second single guide RNA candidate comprising a chimeric non-naturally occurring crRNA comprising a variable targeting domain capable of hybridizing to a target sequence in the genome of said cell, linked to a second nucleotide sequence representing the sense expression of a candidate tracrRNA naturally occurring in said organism; c) providing to the first and second single guide RNA candidates a Cas endonuclease protein, wherein said Cas endonuclease protein can form a complex with either the first single guide RNA

candidate or the second single guide RNA candidate, wherein said complex is capable of introducing a double strand break into said target sequence; and d) identification of the first or second guide RNA candidate and its tracrRNA component that complexes to the Cas endonuclease of (c) and results in cleavage of the target sequence in the genome of said cell.

In one embodiment of the disclosure, the method comprises a method for identification of a tracrRNA of an organism, the method comprising: a)identifying a CRISPR array repeat sequence in a genomic locus of said organism; b) aligning the CRISPR array repeat sequence of (a) with the sequence of the genomic locus of (a) and identifying an antirepeat sequence that encodes a tracrRNA; and, c) determining the transcriptional direction of the tracrRNA.

In one embodiment of the disclosure, the method comprises a method for designing a single guide RNA, the method comprising: a) aligning a tracrRNA sequence with a CRISPR array repeat sequence from a genomic locus of an organism, wherein said CRISPR array repeat sequence comprises a crRNA sequence ;b) deducing the transcriptional direction of the CRISPR array, thereby also deducing the crRNA sequence; and, c) designing a single guide RNA comprising said tracrRNA and crRNA sequences.

In one embodiment of the disclosure, the method comprises a method for producing target sequences, the method comprising: a) identifying a polynucleotides of interest ; b) introducing a Protospacer-Adjacent-Motif (PAM) sequence adjacent to said polynucleotide of interest, wherein said PAM sequence comprises the nucleotide sequence NNNNCND, thereby creating a thereby creating a target site for a guide RNA/Cas9 endonuclease complex; and , c) identifying a polynucleotides of interest .

Polynucleotides of interest are further described herein and include polynucleotides reflective of the commercial markets and interests of those involved in the development of the crop. Crops and markets of interest change, and as developing nations open up world markets, new crops and technologies will emerge also. In addition, as our understanding of agronomic traits and characteristics such as yield and heterosis increase, the choice of genes for genetic engineering will change accordingly.

Further provided are methods for identifying at least one plant cell, comprising in its genome, a polynucleotide of interest integrated at the target site. A variety of methods are available for identifying those plant cells with insertion into the genome at or near to the target site without using a screenable marker phenotype. Such methods can be viewed as directly analyzing a target sequence to detect any change in the target sequence, including but not limited to PCR methods, sequencing methods, nuclease digestion, Southern blots, and any combination thereof. See, for example, US Patent Application 12/147,834, herein incorporated by reference to the extent necessary for the methods described herein. The method also comprises recovering a plant from the plant cell comprising a polynucleotide of Interest integrated into its genome. The plant may be sterile or fertile. It is recognized that any polynucleotide of interest can be provided, integrated into the plant genome at the target site, and expressed in a plant.

Polynucleotides/polypeptides of interest include, but are not limited to, herbicide-resistance coding sequences, insecticidal coding sequences, nematicidal coding sequences, antimicrobial coding sequences, antifungal coding sequences, antiviral coding sequences, abiotic and biotic stress tolerance coding sequences, or sequences modifying plant traits such as yield, grain quality, nutrient content, starch quality and quantity, nitrogen fixation and/or utilization, fatty acids, and oil content and/or composition. More specific polynucleotides of interest include, but are not limited to, genes that improve crop yield, polypeptides that improve desirability of crops, genes encoding proteins conferring resistance to abiotic stress, such as drought, nitrogen, temperature, salinity, toxic metals or trace elements, or those conferring resistance to toxins such as pesticides and herbicides, or to biotic stress, such as attacks by fungi, viruses, bacteria, insects, and nematodes, and development of diseases associated with these organisms. General categories of genes of interest include, for example, those genes involved in information, such as zinc fingers, those involved in communication, such as kinases, and those involved in housekeeping, such as heat shock proteins. More specific categories of transgenes, for example, include genes encoding important traits for agronomics, insect resistance, disease resistance, herbicide resistance, fertility or sterility, grain characteristics, and commercial products. Genes of interest include, generally,

those involved in oil, starch, carbohydrate, or nutrient metabolism as well as those affecting kernel size, sucrose loading, and the like that can be stacked or used in combination with other traits, such as but not limited to herbicide resistance, described herein.

Agronomically important traits such as oil, starch, and protein content can be genetically altered in addition to using traditional breeding methods. Modifications include increasing content of oleic acid, saturated and unsaturated oils, increasing levels of lysine and sulfur, providing essential amino acids, and also modification of starch. Hordothionin protein modifications are described in U.S. Patent Nos. 5,703,049, 5,885,801, 5,885,802, and 5,990,389, herein incorporated by reference.

Polynucleotide sequences of interest may encode proteins involved in providing disease or pest resistance. By "disease resistance" or "pest resistance" is intended that the plants avoid the harmful symptoms that are the outcome of the plant-pathogen interactions. Pest resistance genes may encode resistance to pests that have great yield drag such as rootworm, cutworm, European Corn Borer, and the like. Disease resistance and insect resistance genes such as lysozymes or cecropins for antibacterial protection, or proteins such as defensins, glucanases or chitinases for antifungal protection, or Bacillus thuringiensis endotoxins, protease inhibitors, collagenases, lectins, or glycosidases for controlling nematodes or insects are all examples of useful gene products. Genes encoding disease resistance traits include detoxification genes, such as against fumonisin (U.S. Patent No. 5,792,931); avirulence (avr) and disease resistance (R) genes (Jones et al. (1994) Science 266:789; Martin et al. (1993) Science 262:1432; and Mindrinos et al. (1994) Cell 78:1089); and the like. Insect resistance genes may encode resistance to pests that have great yield drag such as rootworm, cutworm, European Corn Borer, and the like. Such genes include, for example, *Bacillus thuringiensis* toxic protein genes (U.S. Patent Nos. 5,366,892; 5,747,450; 5,736,514; 5,723,756; 5,593,881; and Geiser *et al.* (1986) *Gene* 48:109); and the like.

An "herbicide resistance protein" or a protein resulting from expression of an "herbicide resistance-encoding nucleic acid molecule" includes proteins that confer upon a cell the ability to tolerate a higher concentration of an herbicide than cells that do not express the protein, or to tolerate a certain concentration of an herbicide

for a longer period of time than cells that do not express the protein. Herbicide resistance traits may be introduced into plants by genes coding for resistance to herbicides that act to inhibit the action of acetolactate synthase (ALS), in particular the sulfonylurea-type herbicides, genes coding for resistance to herbicides that act to inhibit the action of glutamine synthase, such as phosphinothricin or basta (e.g., the bar gene), glyphosate (e.g., the EPSP synthase gene and the GAT gene), HPPD inhibitors (e.g, the HPPD gene) or other such genes known in the art. See, for example, US Patent Nos. 7,626,077, 5,310,667, 5,866,775, 6,225,114, 6,248,876, 7,169,970, 6,867,293, and US Provisional Application No. 61/401,456, each of which is herein incorporated by reference. The *bar* gene encodes resistance to the herbicide basta, the *nptII* gene encodes resistance to the antibiotics kanamycin and geneticin, and the ALS-gene mutants encode resistance to the herbicide chlorsulfuron.

Sterility genes can also be encoded in an expression cassette and provide an alternative to physical detasseling. Examples of genes used in such ways include male fertility genes such as MS26 (see for example U.S. Patents 7,098,388, 7,517,975, 7,612,251), MS45 (see for example U.S. Patents 5,478,369, 6,265,640) or MSCA1 (see for example U.S. Patent 7,919,676). Maize plants (Zea mays L.) can be bred by both self-pollination and cross-pollination techniques. Maize has male flowers, located on the tassel, and female flowers, located on the ear, on the same plant. It can self-pollinate ("selfing") or cross pollinate. Natural pollination occurs in maize when wind blows pollen from the tassels to the silks that protrude from the tops of the incipient ears. Pollination may be readily controlled by techniques known to those of skill in the art. The development of maize hybrids requires the development of homozygous inbred lines, the crossing of these lines, and the evaluation of the crosses. Pedigree breeding and recurrent selections are two of the breeding methods used to develop inbred lines from populations. Breeding programs combine desirable traits from two or more inbred lines or various broad-based sources into breeding pools from which new inbred lines are developed by selfing and selection of desired phenotypes. A hybrid maize variety is the cross of two such inbred lines, each of which may have one or more desirable characteristics lacked by the other or which complement the other. The new inbreds are crossed

with other inbred lines and the hybrids from these crosses are evaluated to determine which have commercial potential. The hybrid progeny of the first generation is designated F1. The F1 hybrid is more vigorous than its inbred parents. This hybrid vigor, or heterosis, can be manifested in many ways, including increased vegetative growth and increased yield.

Hybrid maize seed can be produced by a male sterility system incorporating manual detasseling. To produce hybrid seed, the male tassel is removed from the growing female inbred parent, which can be planted in various alternating row patterns with the male inbred parent. Consequently, providing that there is sufficient isolation from sources of foreign maize pollen, the ears of the female inbred will be fertilized only with pollen from the male inbred. The resulting seed is therefore hybrid (F1) and will form hybrid plants.

Field variation impacting plant development can result in plants tasseling after manual detasseling of the female parent is completed. Or, a female inbred plant tassel may not be completely removed during the detasseling process. In any event, the result is that the female plant will successfully shed pollen and some female plants will be self-pollinated. This will result in seed of the female inbred being harvested along with the hybrid seed which is normally produced. Female inbred seed does not exhibit heterosis and therefore is not as productive as F1 seed. In addition, the presence of female inbred seed can represent a germplasm security risk for the company producing the hybrid.

Alternatively, the female inbred can be mechanically detasseled by machine. Mechanical detasseling is approximately as reliable as hand detasseling, but is faster and less costly. However, most detasseling machines produce more damage to the plants than hand detasseling. Thus, no form of detasseling is presently entirely satisfactory, and a need continues to exist for alternatives which further reduce production costs and to eliminate self-pollination of the female parent in the production of hybrid seed.

Mutations that cause male sterility in plants have the potential to be useful in methods for hybrid seed production for crop plants such as maize and can lower production costs by eliminating the need for the labor-intensive removal of male flowers (also known as de-tasseling) from the maternal parent plants used as a

hybrid parent. Mutations that cause male sterility in maize have been produced by a variety of methods such as X-rays or UV-irradiations, chemical treatments, or transposable element insertions (ms23, ms25, ms26, ms32) (Chaubal et al. (2000) Am J Bot 87:1193-1201). Conditional regulation of fertility genes through fertility/sterility "molecular switches" could enhance the options for designing new male-sterility systems for crop improvement (Unger et al. (2002) Transgenic Res 11:455-465).

Besides identification of novel genes impacting male fertility, there remains a need to provide a reliable system of producing genetic male sterility.

In U.S. Patent No. 5,478,369, a method is described by which the Ms45 male fertility gene was tagged and cloned on maize chromosome 9. Previously, there had been described a male fertility gene on chromosome 9, ms2, which had never been cloned and sequenced. It is not allelic to the gene referred to in the '369 patent. See Albertsen, M. and Phillips, R.L., "Developmental Cytology of 13 Genetic Male Sterile Loci in Maize" Canadian Journal of Genetics & Cytology 23:195-208 (Jan. 1981). The only fertility gene cloned before that had been the Arabidopsis gene described at Aarts, et al., supra.

Furthermore, it is recognized that the polynucleotide of interest may also comprise antisense sequences complementary to at least a portion of the messenger RNA (mRNA) for a targeted gene sequence of interest. Antisense nucleotides are constructed to hybridize with the corresponding mRNA. Modifications of the antisense sequences may be made as long as the sequences hybridize to and interfere with expression of the corresponding mRNA. In this manner, antisense constructions having 70%, 80%, or 85% sequence identity to the corresponding antisense sequences may be used. Furthermore, portions of the antisense nucleotides may be used to disrupt the expression of the target gene. Generally, sequences of at least 50 nucleotides, 100 nucleotides, 200 nucleotides, or greater may be used.

In addition, the polynucleotide of interest may also be used in the sense orientation to suppress the expression of endogenous genes in plants. Methods for suppressing gene expression in plants using polynucleotides in the sense orientation are known in the art. The methods generally involve transforming plants

with a DNA construct comprising a promoter that drives expression in a plant operably linked to at least a portion of a nucleotide sequence that corresponds to the transcript of the endogenous gene. Typically, such a nucleotide sequence has substantial sequence identity to the sequence of the transcript of the endogenous gene, generally greater than about 65% sequence identity, about 85% sequence identity, or greater than about 95% sequence identity. See, U.S. Patent Nos. 5,283,184 and 5,034,323; herein incorporated by reference.

The polynucleotide of interest can also be a phenotypic marker. A phenotypic marker is screenable or a selectable marker that includes visual markers and selectable markers whether it is a positive or negative selectable marker. Any phenotypic marker can be used. Specifically, a selectable or screenable marker comprises a DNA segment that allows one to identify, or select for or against a molecule or a cell that contains it, often under particular conditions. These markers can encode an activity, such as, but not limited to, production of RNA, peptide, or protein, or can provide a binding site for RNA, peptides, proteins, inorganic and organic compounds or compositions and the like.

Examples of selectable markers include, but are not limited to, DNA segments that comprise restriction enzyme sites; DNA segments that encode products which provide resistance against otherwise toxic compounds including antibiotics, such as, spectinomycin, ampicillin, kanamycin, tetracycline, Basta, neomycin phosphotransferase II (NEO) and hygromycin phosphotransferase (HPT)); DNA segments that encode products which are otherwise lacking in the recipient cell (e.g., tRNA genes, auxotrophic markers); DNA segments that encode products which can be readily identified (e.g., phenotypic markers such as β-galactosidase, GUS; fluorescent proteins such as green fluorescent protein (GFP), cyan (CFP), yellow (YFP), red (RFP), and cell surface proteins); the generation of new primer sites for PCR (e.g., the juxtaposition of two DNA sequence not previously juxtaposed), the inclusion of DNA sequences not acted upon or acted upon by a restriction endonuclease or other DNA modifying enzyme, chemical, etc.; and, the inclusion of a DNA sequences required for a specific modification (e.g., methylation) that allows its identification.

Additional selectable markers include genes that confer resistance to herbicidal compounds, such as glufosinate ammonium, bromoxynil, imidazolinones, and 2,4-dichlorophenoxyacetate (2,4-D). See for example, Yarranton, (1992) Curr Opin Biotech 3:506-11; Christopherson et al., (1992) Proc. Natl. Acad. Sci. USA 89:6314-8; Yao et al., (1992) Cell 71:63-72; Reznikoff, (1992) Mol Microbiol 6:2419-22; Hu et al., (1987) Cell 48:555-66; Brown et al., (1987) Cell 49:603-12; Figge et al., (1988) Cell 52:713-22; Deuschle et al., (1989) Proc. Natl. Acad. Sci. USA 86:5400-4; Fuerst et al., (1989) Proc. Natl. Acad. Sci. USA 86:2549-53; Deuschle et al., (1990) Science 248:480-3; Gossen, (1993) Ph.D. Thesis, University of Heidelberg; Reines et al., (1993) Proc. Natl. Acad. Sci. USA 90:1917-21; Labow et al., (1990) Mol Cell Biol 10:3343-56; Zambretti et al., (1992) Proc. Natl. Acad. Sci. USA 89:3952-6; Baim et al., (1991) Proc. Natl. Acad. Sci. USA 88:5072-6; Wyborski et al., (1991) Nucleic Acids Res 19:4647-53; Hillen and Wissman, (1989) Topics Mol Struc Biol 10:143-62; Degenkolb et al., (1991) Antimicrob Agents Chemother 35:1591-5; Kleinschnidt et al., (1988) Biochemistry 27:1094-104; Bonin, (1993) Ph.D. Thesis, University of Heidelberg; Gossen et al., (1992) Proc. Natl. Acad. Sci. USA 89:5547-51; Oliva et al., (1992) Antimicrob Agents Chemother 36:913-9; Hlavka et al., (1985) Handbook of Experimental Pharmacology, Vol. 78 (Springer-Verlag, Berlin); Gill et al., (1988) Nature 334:721-4. Commercial traits can also be encoded on a gene or genes that could increase for example, starch for ethanol production, or provide expression of proteins. Another important commercial use of transformed plants is the production of polymers and bioplastics such as described in U.S. Patent No. 5,602,321. Genes such as β-Ketothiolase, PHBase (polyhydroxyburyrate synthase), and acetoacetyl-CoA reductase (see Schubert *et al.* (1988) *J. Bacteriol.* 170:5837-5847) facilitate expression of polyhyroxyalkanoates (PHAs).

Exogenous products include plant enzymes and products as well as those from other sources including prokaryotes and other eukaryotes. Such products include enzymes, cofactors, hormones, and the like. The level of proteins, particularly modified proteins having improved amino acid distribution to improve the nutrient value of the plant, can be increased. This is achieved by the expression of such proteins having enhanced amino acid content.

The transgenes, recombinant DNA molecules, DNA sequences of interest, and polynucleotides of interest can be comprise one or more DNA sequences for gene silencing. Methods for gene silencing involving the expression of DNA sequences in plant are known in the art include, but are not limited to, cosuppression, antisense suppression, double-stranded RNA (dsRNA) interference, hairpin RNA (hpRNA) interference, intron-containing hairpin RNA (ihpRNA) interference, transcriptional gene silencing, and micro RNA (miRNA) interference

As used herein, “nucleic acid” means a polynucleotide and includes a single or a double-stranded polymer of deoxyribonucleotide or ribonucleotide bases. Nucleic acids may also include fragments and modified nucleotides. Thus, the terms “polynucleotide”, “nucleic acid sequence”, “nucleotide sequence” and “nucleic acid fragment” are used interchangeably to denote a polymer of RNA and/or DNA that is single- or double-stranded, optionally containing synthetic, non-natural, or altered nucleotide bases. Nucleotides (usually found in their 5’-monophosphate form) are referred to by their single letter designation as follows: “A” for adenosine or deoxyadenosine (for RNA or DNA, respectively), “C” for cytosine or deoxycytosine, “G” for guanosine or deoxyguanosine, “U” for uridine, “T” for deoxythymidine, “R” for purines (A or G), “Y” for pyrimidines (C or T), “K” for G or T, “H” for A or C or T, “I” for inosine, and “N” for any nucleotide.

“Open reading frame” is abbreviated ORF.

The terms “subfragment that is functionally equivalent” and “functionally equivalent subfragment” are used interchangeably herein. These terms refer to a portion or subsequence of an isolated nucleic acid fragment in which the ability to alter gene expression or produce a certain phenotype is retained whether or not the fragment or subfragment encodes an active enzyme. For example, the fragment or subfragment can be used in the design of genes to produce the desired phenotype in a transformed plant. Genes can be designed for use in suppression by linking a nucleic acid fragment or subfragment thereof, whether or not it encodes an active enzyme, in the sense or antisense orientation relative to a plant promoter sequence.

The term “conserved domain” or “motif” means a set of amino acids conserved at specific positions along an aligned sequence of evolutionarily related proteins. While amino acids at other positions can vary between homologous

proteins, amino acids that are highly conserved at specific positions indicate amino acids that are essential to the structure, the stability, or the activity of a protein. Because they are identified by their high degree of conservation in aligned sequences of a family of protein homologues, they can be used as identifiers, or "signatures", to determine if a protein with a newly determined sequence belongs to a previously identified protein family.

Polynucleotide and polypeptide sequences, variants thereof, and the structural relationships of these sequences can be described by the terms "homology", "homologous", "substantially identical", "substantially similar" and "corresponding substantially" which are used interchangeably herein. These refer to polypeptide or nucleic acid fragments wherein changes in one or more amino acids or nucleotide bases do not affect the function of the molecule, such as the ability to mediate gene expression or to produce a certain phenotype. These terms also refer to modification(s) of nucleic acid fragments that do not substantially alter the functional properties of the resulting nucleic acid fragment relative to the initial, unmodified fragment. These modifications include deletion, substitution, and/or insertion of one or more nucleotides in the nucleic acid fragment.

Substantially similar nucleic acid sequences encompassed may be defined by their ability to hybridize (under moderately stringent conditions, e.g., 0.5X SSC, 0.1% SDS, 60°C) with the sequences exemplified herein, or to any portion of the nucleotide sequences disclosed herein and which are functionally equivalent to any of the nucleic acid sequences disclosed herein. Stringency conditions can be adjusted to screen for moderately similar fragments, such as homologous sequences from distantly related organisms, to highly similar fragments, such as genes that duplicate functional enzymes from closely related organisms. Post-hybridization washes determine stringency conditions.

The term "selectively hybridizes" includes reference to hybridization, under stringent hybridization conditions, of a nucleic acid sequence to a specified nucleic acid target sequence to a detectably greater degree (e.g., at least 2-fold over background) than its hybridization to non-target nucleic acid sequences and to the substantial exclusion of non-target nucleic acids. Selectively hybridizing sequences

typically have about at least 80% sequence identity, or 90% sequence identity, up to and including 100% sequence identity (i.e., fully complementary) with each other.

The term "stringent conditions" or “stringent hybridization conditions” includes reference to conditions under which a probe will selectively hybridize to its target sequence in an *in vitro* hybridization assay. Stringent conditions are sequence-dependent and will be different in different circumstances. By controlling the stringency of the hybridization and/or washing conditions, target sequences can be identified which are 100% complementary to the probe (homologous probing). Alternatively, stringency conditions can be adjusted to allow some mismatching in sequences so that lower degrees of similarity are detected (heterologous probing). Generally, a probe is less than about 1000 nucleotides in length, optionally less than 500 nucleotides in length.

Typically, stringent conditions will be those in which the salt concentration is less than about 1.5 M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salt(s)) at pH 7.0 to 8.3, and at least about 30°C for short probes (e.g., 10 to 50 nucleotides) and at least about 60°C for long probes (e.g., greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide. Exemplary low stringency conditions include hybridization with a buffer solution of 30 to 35% formamide, 1 M NaCl, 1% SDS (sodium dodecyl sulphate) at 37°C, and a wash in 1X to 2X SSC (20X SSC = 3.0 M NaCl/0.3 M trisodium citrate) at 50 to 55°C. Exemplary moderate stringency conditions include hybridization in 40 to 45% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 0.5X to 1X SSC at 55 to 60°C. Exemplary high stringency conditions include hybridization in 50% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 0.1X SSC at 60 to 65°C.

“Sequence identity” or “identity” in the context of nucleic acid or polypeptide sequences refers to the nucleic acid bases or amino acid residues in two sequences that are the same when aligned for maximum correspondence over a specified comparison window.

The term “percentage of sequence identity” refers to the value determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide or polypeptide sequence in the comparison window

may comprise additions or deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the results by 100 to yield the percentage of sequence identity. Useful examples of percent sequence identities include, but are not limited to, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90% or 95%, or any integer percentage from 50% to 100%. These identities can be determined using any of the programs described herein.

Sequence alignments and percent identity or similarity calculations may be determined using a variety of comparison methods designed to detect homologous sequences including, but not limited to, the MegAlign™ program of the LASERGENE bioinformatics computing suite (DNASTAR Inc., Madison, WI). Within the context of this application it will be understood that where sequence analysis software is used for analysis, that the results of the analysis will be based on the "default values" of the program referenced, unless otherwise specified. As used herein "default values" will mean any set of values or parameters that originally load with the software when first initialized.

The "Clustal V method of alignment" corresponds to the alignment method labeled Clustal V (described by Higgins and Sharp, (1989) *CABIOS* 5:151-153; Higgins *et al.*, (1992) *Comput Appl Biosci* 8:189-191) and found in the MegAlign™ program of the LASERGENE bioinformatics computing suite (DNASTAR Inc., Madison, WI). For multiple alignments, the default values correspond to GAP PENALTY=10 and GAP LENGTH PENALTY=10. Default parameters for pairwise alignments and calculation of percent identity of protein sequences using the Clustal method are KTUPLE=1, GAP PENALTY=3, WINDOW=5 and DIAGONALS SAVED=5. For nucleic acids these parameters are KTUPLE=2, GAP PENALTY=5, WINDOW=4 and DIAGONALS SAVED=4. After alignment of the sequences using the Clustal V program, it is possible to obtain a "percent identity" by viewing the "sequence distances" table in the same program.

The "Clustal W method of alignment" corresponds to the alignment method labeled Clustal W (described by Higgins and Sharp, (1989) *CABIOS* 5:151-153; Higgins *et al.*, (1992) *Comput Appl Biosci* 8:189-191) and found in the MegAlign™ v6.1 program of the LASERGENE bioinformatics computing suite (DNASTAR Inc., Madison, WI). Default parameters for multiple alignment (GAP PENALTY=10, GAP LENGTH PENALTY=0.2, Delay Divergen Seqs (%)=30, DNA Transition Weight=0.5, Protein Weight Matrix=Gonnet Series, DNA Weight Matrix=IUB). After alignment of the sequences using the Clustal W program, it is possible to obtain a "percent identity" by viewing the "sequence distances" table in the same program.

Unless otherwise stated, sequence identity/similarity values provided herein refer to the value obtained using GAP Version 10 (GCG, Accelrys, San Diego, CA) using the following parameters: % identity and % similarity for a nucleotide sequence using a gap creation penalty weight of 50 and a gap length extension penalty weight of 3, and the nwsgapdna.cmp scoring matrix; % identity and % similarity for an amino acid sequence using a GAP creation penalty weight of 8 and a gap length extension penalty of 2, and the BLOSUM62 scoring matrix (Henikoff and Henikoff, (1989) *Proc. Natl. Acad. Sci. USA* 89:10915). GAP uses the algorithm of Needleman and Wunsch, (1970) *J Mol Biol* 48:443-53, to find an alignment of two complete sequences that maximizes the number of matches and minimizes the number of gaps. GAP considers all possible alignments and gap positions and creates the alignment with the largest number of matched bases and the fewest gaps, using a gap creation penalty and a gap extension penalty in units of matched bases.

"BLAST" is a searching algorithm provided by the National Center for Biotechnology Information (NCBI) used to find regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches to identify sequences having sufficient similarity to a query sequence such that the similarity would not be predicted to have occurred randomly. BLAST reports the identified sequences and their local alignment to the query sequence.

It is well understood by one skilled in the art that many levels of sequence identity are useful in identifying polypeptides from other species or modified

naturally or synthetically wherein such polypeptides have the same or similar function or activity. Useful examples of percent identities include, but are not limited to, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90% or 95%, or any integer percentage from 50% to 100%. Indeed, any integer amino acid identity from 50% to 100% may be useful in describing the present disclosure, such as 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% or 99%.

"Gene" includes a nucleic acid fragment that expresses a functional molecule such as, but not limited to, a specific protein, including regulatory sequences preceding (5' non-coding sequences) and following (3' non-coding sequences) the coding sequence. "Native gene" refers to a gene as found in nature with its own regulatory sequences.

A "mutated gene" is a gene that has been altered through human intervention. Such a "mutated gene" has a sequence that differs from the sequence of the corresponding non-mutated gene by at least one nucleotide addition, deletion, or substitution. In certain embodiments of the disclosure, the mutated gene comprises an alteration that results from a guide polynucleotide/Cas endonuclease system as disclosed herein. A mutated plant is a plant comprising a mutated gene.

As used herein, a "targeted mutation" is a mutation in a native gene that was made by altering a target sequence within the native gene using a method involving a double-strand-break-inducing agent that is capable of inducing a double-strand break in the DNA of the target sequence as disclosed herein or known in the art.

The guide RNA/Cas endonuclease induced targeted mutation can occur in a nucleotide sequence that is located within or outside a genomic target site that is recognized and cleaved by a Cas endonuclease.

The term "genome" as it applies to a plant cells encompasses not only chromosomal DNA found within the nucleus, but organelle DNA found within subcellular components (e.g., mitochondria, or plastid) of the cell.

A “codon-modified gene” or “codon-preferred gene” or “codon-optimized gene” is a gene having its frequency of codon usage designed to mimic the frequency of preferred codon usage of the host cell.

An “allele” is one of several alternative forms of a gene occupying a given locus on a chromosome. When all the alleles present at a given locus on a chromosome are the same, that plant is homozygous at that locus. If the alleles present at a given locus on a chromosome differ, that plant is heterozygous at that locus.

“Coding sequence” refers to a polynucleotide sequence which codes for a specific amino acid sequence. “Regulatory sequences” refer to nucleotide sequences located upstream (5’ non-coding sequences), within, or downstream (3’ non-coding sequences) of a coding sequence, and which influence the transcription, RNA processing or stability, or translation of the associated coding sequence. Regulatory sequences may include, but are not limited to: promoters, translation leader sequences, 5’ untranslated sequences, 3’ untranslated sequences, introns, polyadenylation target sequences, RNA processing sites, effector binding sites, and stem-loop structures.

"A plant-optimized nucleotide sequence" is nucleotide sequence that has been optimized for increased expression in plants, particularly for increased expression in plants or in one or more plants of interest. For example, a plant-optimized nucleotide sequence can be synthesized by modifying a nucleotide sequence encoding a protein such as, for example, double-strand-break-inducing agent (*e.g.*, an endonuclease) as disclosed herein, using one or more plant-preferred codons for improved expression. *See*, for example, Campbell and Gowri (1990) *Plant Physiol.* 92:1-11 for a discussion of host-preferred codon usage.

Methods are available in the art for synthesizing plant-preferred genes. See, for example, U.S. Patent Nos. 5,380,831, and 5,436,391, and Murray *et al.* (1989) *Nucleic Acids Res.* 17:477-498, herein incorporated by reference. Additional sequence modifications are known to enhance gene expression in a plant host. These include, for example, elimination of: one or more sequences encoding spurious polyadenylation signals, one or more exon-intron splice site signals, one or more transposon-like repeats, and other such well-characterized sequences that

may be deleterious to gene expression. The G-C content of the sequence may be adjusted to levels average for a given plant host, as calculated by reference to known genes expressed in the host plant cell. When possible, the sequence is modified to avoid one or more predicted hairpin secondary mRNA structures. Thus, "a plant-optimized nucleotide sequence" of the present disclosure comprises one or more of such sequence modifications.

A promoter is a region of DNA involved in recognition and binding of RNA polymerase and other proteins to initiate transcription. The promoter sequence consists of proximal and more distal upstream elements, the latter elements often referred to as enhancers. An “enhancer” is a DNA sequence that can stimulate promoter activity, and may be an innate element of the promoter or a heterologous element inserted to enhance the level or tissue-specificity of a promoter. Promoters may be derived in their entirety from a native gene, or be composed of different elements derived from different promoters found in nature, and/or comprise synthetic DNA segments. It is understood by those skilled in the art that different promoters may direct the expression of a gene in different tissues or cell types, or at different stages of development, or in response to different environmental conditions. It is further recognized that since in most cases the exact boundaries of regulatory sequences have not been completely defined, DNA fragments of some variation may have identical promoter activity. Promoters that cause a gene to be expressed in most cell types at most times are commonly referred to as “constitutive promoters”.

It has been shown that certain promoters are able to direct RNA synthesis at a higher rate than others. These are called “strong promoters”. Certain other promoters have been shown to direct RNA synthesis at higher levels only in particular types of cells or tissues and are often referred to as “tissue specific promoters”, or “tissue-preferred promoters” if the promoters direct RNA synthesis preferably in certain tissues but also in other tissues at reduced levels. Since patterns of expression of a chimeric gene (or genes) introduced into a plant are controlled using promoters, there is an ongoing interest in the isolation of novel promoters which are capable of controlling the expression of a chimeric gene or

(genes) at certain levels in specific tissue types or at specific plant developmental stages.

A plant promoter can include a promoter capable of initiating transcription in a plant cell, for a review of plant promoters, see, Potenza *et al.*, (2004) *In Vitro Cell Dev Biol* 40:1-22. Constitutive promoters include, for example, the core promoter of the Rsyn7 promoter and other constitutive promoters disclosed in WO99/43838 and U.S. Patent No. 6,072,050; the core CaMV 35S promoter (Odell *et al.*, (1985) *Nature* 313:810-2); rice actin (McElroy *et al.*, (1990) *Plant Cell* 2:163-71); ubiquitin (Christensen *et al.*, (1989) *Plant Mol Biol* 12:619-32; Christensen *et al.*, (1992) *Plant Mol Biol* 18:675-89); pEMU (Last *et al.*, (1991) *Theor Appl Genet* 81:581-8); MAS (Velten *et al.*, (1984) *EMBO J* 3:2723-30); ALS promoter (U.S. Patent No. 5,659,026), and the like. Other constitutive promoters are described in, for example, U.S. Patent Nos. 5,608,149; 5,608,144; 5,604,121; 5,569,597; 5,466,785; 5,399,680; 5,268,463; 5,608,142 and 6,177,611. In some examples an inducible promoter may be used. Pathogen-inducible promoters induced following infection by a pathogen include, but are not limited to those regulating expression of PR proteins, SAR proteins, beta-1,3-glucanase, chitinase, *etc*.

Chemical-regulated promoters can be used to modulate the expression of a gene in a plant through the application of an exogenous chemical regulator. The promoter may be a chemical-inducible promoter, where application of the chemical induces gene expression, or a chemical-repressible promoter, where application of the chemical represses gene expression. Chemical-inducible promoters include, but are not limited to, the maize In2-2 promoter, activated by benzene sulfonamide herbicide safeners (De Veylder *et al.*, (1997) *Plant Cell Physiol* 38:568-77), the maize GST promoter (GST-II-27, WO93/01294), activated by hydrophobic electrophilic compounds used as pre-emergent herbicides, and the tobacco PR-1a promoter (Ono *et al.*, (2004) *Biosci Biotechnol Biochem* 68:803-7) activated by salicylic acid. Other chemical-regulated promoters include steroid-responsive promoters (see, for example, the glucocorticoid-inducible promoter (Schena *et al.*, (1991) *Proc. Natl. Acad. Sci. USA* 88:10421-5; McNellis *et al.*, (1998) *Plant J* 14:247-257); tetracycline-inducible and tetracycline-repressible promoters (Gatz *et al.*, (1991) *Mol Gen Genet* 227:229-37; U.S. Patent Nos. 5,814,618 and 5,789,156).

Tissue-preferred promoters can be utilized to target enhanced expression within a particular plant tissue. Tissue-preferred promoters include, for example, Kawamata *et al.*, (1997) *Plant Cell Physiol* 38:792-803; Hansen *et al.*, (1997) *Mol Gen Genet* 254:337-43; Russell *et al.*, (1997) *Transgenic Res* 6:157-68; Rinehart *et al.*, (1996) *Plant Physiol* 112:1331-41; Van Camp *et al.*, (1996) *Plant Physiol* 112:525-35; Canevascini *et al.*, (1996) *Plant Physiol* 112:513-524; Lam, (1994) *Results Probl Cell Differ* 20:181-96; and Guevara-Garcia *et al.*, (1993) *Plant J* 4:495-505. Leaf-preferred promoters include, for example, Yamamoto *et al.*, (1997) *Plant J* 12:255-65; Kwon *et al.*, (1994) *Plant Physiol* 105:357-67; Yamamoto *et al.*, (1994) *Plant Cell Physiol* 35:773-8; Gotor *et al.*, (1993) *Plant J* 3:509-18; Orozco *et al.*, (1993) *Plant Mol Biol* 23:1129-38; Matsuoka *et al.*, (1993) *Proc. Natl. Acad. Sci. USA* 90:9586-90; Simpson *et al.*, (1958) *EMBO J* 4:2723-9; Timko *et al.*, (1988) *Nature* 318:57-8. Root-preferred promoters include, for example, Hire *et al.*, (1992) *Plant Mol Biol* 20:207-18 (soybean root-specific glutamine synthase gene); Miao *et al.*, (1991) *Plant Cell* 3:11-22 (cytosolic glutamine synthase (GS)); Keller and Baumgartner, (1991) *Plant Cell* 3:1051-61 (root-specific control element in the GRP 1.8 gene of French bean); Sanger *et al.*, (1990) *Plant Mol Biol* 14:433-43 (root-specific promoter of *A. tumefaciens* mannopine synthase (MAS)); Bogusz *et al.*, (1990) *Plant Cell* 2:633-41 (root-specific promoters isolated from *Parasponia andersonii* and *Trema tomentosa*); Leach and Aoyagi, (1991) *Plant Sci* 79:69-76 (*A. rhizogenes* rolC and rolD root-inducing genes); Teeri *et al.*, (1989) *EMBO J* 8:343-50 (*Agrobacterium* wound-induced TR1' and TR2' genes); VfENOD-GRP3 gene promoter (Kuster *et al.*, (1995) *Plant Mol Biol* 29:759-72); and rolB promoter (Capana *et al.*, (1994) *Plant Mol Biol* 25:681-91; phaseolin gene (Murai *et al.*, (1983) *Science* 23:476-82; Sengopta-Gopalen *et al.*, (1988) *Proc. Natl. Acad. Sci. USA* 82:3320-4). See also, U.S. Patent Nos. 5,837,876; 5,750,386; 5,633,363; 5,459,252; 5,401,836; 5,110,732 and 5,023,179.

Seed-preferred promoters include both seed-specific promoters active during seed development, as well as seed-germinating promoters active during seed germination. See, Thompson *et al.*, (1989) *BioEssays* 10:108. Seed-preferred promoters include, but are not limited to, Cim1 (cytokinin-induced message); cZ19B1 (maize 19 kDa zein); and milps (myo-inositol-1-phosphate synthase);

(WO00/11177; and U.S. Patent 6,225,529). For dicots, seed-preferred promoters include, but are not limited to, bean β-phaseolin, napin, β-conglycinin, soybean lectin, cruciferin, and the like. For monocots, seed-preferred promoters include, but are not limited to, maize 15 kDa zein, 22 kDa zein, 27 kDa gamma zein, waxy, shrunken 1, shrunken 2, globulin 1, oleosin, and nuc1. See also, WO00/12733, where seed-preferred promoters from *END1* and *END2* genes are disclosed.

The term "inducible promoter" refers to promoters that selectively express a coding sequence or functional RNA in response to the presence of an endogenous or exogenous stimulus, for example by chemical compounds (chemical inducers) or in response to environmental, hormonal, chemical, and/or developmental signals. Inducible or regulated promoters include, for example, promoters induced or regulated by light, heat, stress, flooding or drought, salt stress, osmotic stress, phytohormones, wounding, or chemicals such as ethanol, abscisic acid (ABA), jasmonate, salicylic acid, or safeners.

An example of a stress-inducible is RD29A promoter (Kasuga et al. (1999) Nature Biotechnol. 17:287-91). One of ordinary skill in the art is familiar with protocols for simulating drought conditions and for evaluating drought tolerance of plants that have been subjected to simulated or naturally-occurring drought conditions. For example, one can simulate drought conditions by giving plants less water than normally required or no water over a period of time, and one can evaluate drought tolerance by looking for differences in physiological and/or physical condition, including (but not limited to) vigor, growth, size, or root length, or in particular, leaf color or leaf area size. Other techniques for evaluating drought tolerance include measuring chlorophyll fluorescence, photosynthetic rates and gas exchange rates. Also, one of ordinary skill in the art is familiar with protocols for simulating stress conditions such as osmotic stress, salt stress and temperature stress and for evaluating stress tolerance of plants that have been subjected to simulated or naturally-occurring stress conditions.

Another example of an inducible promoter useful in plant cells has been described in US patent application, US 2013-0312137A1, published on November 21, 2013, incorporated by reference herein. US patent application US 2013-0312137A1 describes a ZmCAS1 promoter from a CBSU-Anther_Subtraction library

(CAS1) gene encoding a mannitol dehydrogenase from maize, and functional fragments thereof. The ZmCAS1 promoter (also refered to as “CAS1 promoter”, “mannitol dehydrogenase promoter “ , “mdh promoter”) can be induced by a chemical or stress treatment. The chemical can be a safener such as, but not limited to, N-(aminocarbonyl)-2-chlorobenzenesulfonamide (2-CBSU). The stress treatment can be a heat treatment such as, but not limited to, a heat shock treatment (see also US provisional patent application,62/120421, filed on February 25, 2015, incorporated by reference herein.

New promoters of various types useful in plant cells are constantly being discovered; numerous examples may be found in the compilation by Okamuro and Goldberg, (1989) In *The Biochemistry of Plants*, Vol. 115, Stumpf and Conn, eds (New York, NY: Academic Press), pp. 1-82.

“Translation leader sequence” refers to a polynucleotide sequence located between the promoter sequence of a gene and the coding sequence. The translation leader sequence is present in the mRNA upstream of the translation start sequence. The translation leader sequence may affect processing of the primary transcript to mRNA, mRNA stability or translation efficiency. Examples of translation leader sequences have been described (e.g., Turner and Foster, (1995) *Mol Biotechnol* 3:225-236).

“3’ non-coding sequences”, “transcription terminator” or “termination sequences” refer to DNA sequences located downstream of a coding sequence and include polyadenylation recognition sequences and other sequences encoding regulatory signals capable of affecting mRNA processing or gene expression. The polyadenylation signal is usually characterized by affecting the addition of polyadenylic acid tracts to the 3’ end of the mRNA precursor. The use of different 3’ non-coding sequences is exemplified by Ingelbrecht *et al.*, (1989) *Plant Cell* 1:671-680.

“RNA transcript” refers to the product resulting from RNA polymerase-catalyzed transcription of a DNA sequence. When the RNA transcript is a perfect complimentary copy of the DNA sequence, it is referred to as the primary transcript or pre-mRNA. A RNA transcript is referred to as the mature RNA or mRNA when it is a RNA sequence derived from post-transcriptional processing of the primary

transcript pre mRNAt. "Messenger RNA" or "mRNA" refers to the RNA that is without introns and that can be translated into protein by the cell. "cDNA" refers to a DNA that is complementary to, and synthesized from, a mRNA template using the enzyme reverse transcriptase. The cDNA can be single-stranded or converted into double-stranded form using the Klenow fragment of DNA polymerase I. "Sense" RNA refers to RNA transcript that includes the mRNA and can be translated into protein within a cell or *in vitro*. "Antisense RNA" refers to an RNA transcript that is complementary to all or part of a target primary transcript or mRNA, and that blocks the expression of a target gene (see, e.g., U.S. Patent No. 5,107,065). The complementarity of an antisense RNA may be with any part of the specific gene transcript, i.e., at the 5' non-coding sequence, 3' non-coding sequence, introns, or the coding sequence. "Functional RNA" refers to antisense RNA, ribozyme RNA, or other RNA that may not be translated but yet has an effect on cellular processes. The terms "complement" and "reverse complement" are used interchangeably herein with respect to mRNA transcripts, and are meant to define the antisense RNA of the message.

The term "operably linked" refers to the association of nucleic acid sequences on a single nucleic acid fragment so that the function of one is regulated by the other. For example, a promoter is operably linked with a coding sequence when it is capable of regulating the expression of that coding sequence (i.e., the coding sequence is under the transcriptional control of the promoter). Coding sequences can be operably linked to regulatory sequences in a sense or antisense orientation. In another example, the complementary RNA regions can be operably linked, either directly or indirectly, 5' to the target mRNA, or 3' to the target mRNA, or within the target mRNA, or a first complementary region is 5' and its complement is 3' to the target mRNA.

Standard recombinant DNA and molecular cloning techniques used herein are well known in the art and are described more fully in Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual; Cold Spring Harbor Laboratory*: Cold Spring Harbor, NY (1989). Transformation methods are well known to those skilled in the art and are described *infra*.

"PCR" or "polymerase chain reaction" is a technique for the synthesis of specific DNA segments and consists of a series of repetitive denaturation, annealing, and extension cycles. Typically, a double-stranded DNA is heat denatured, and two primers complementary to the 3' boundaries of the target segment are annealed to the DNA at low temperature, and then extended at an intermediate temperature. One set of these three consecutive steps is referred to as a "cycle".

The term "recombinant" refers to an artificial combination of two otherwise separated segments of sequence, e.g., by chemical synthesis, or manipulation of isolated segments of nucleic acids by genetic engineering techniques.

The terms "plasmid", "vector" and "cassette" refer to an extra chromosomal element often carrying genes that are not part of the central metabolism of the cell, and usually in the form of double-stranded DNA. Such elements may be autonomously replicating sequences, genome integrating sequences, phage, or nucleotide sequences, in linear or circular form, of a single- or double-stranded DNA or RNA, derived from any source, in which a number of nucleotide sequences have been joined or recombined into a unique construction which is capable of introducing a polynucleotide of interest into a cell. "Transformation cassette" refers to a specific vector containing a gene and having elements in addition to the gene that facilitates transformation of a particular host cell. "Expression cassette" refers to a specific vector containing a gene and having elements in addition to the gene that allow for expression of that gene in a host.

The terms "recombinant DNA molecule", "recombinant construct", "expression construct", " construct", "construct", and "recombinant DNA construct" are used interchangeably herein. A recombinant construct comprises an artificial combination of nucleic acid fragments, e.g., regulatory and coding sequences that are not all found together in nature. For example, a construct may comprise regulatory sequences and coding sequences that are derived from different sources, or regulatory sequences and coding sequences derived from the same source, but arranged in a manner different than that found in nature. Such a construct may be used by itself or may be used in conjunction with a vector. If a vector is used, then the choice of vector is dependent upon the method that will be used to transform

host cells as is well known to those skilled in the art. For example, a plasmid vector can be used. The skilled artisan is well aware of the genetic elements that must be present on the vector in order to successfully transform, select and propagate host cells. The skilled artisan will also recognize that different independent transformation events may result in different levels and patterns of expression (Jones *et al.*, (1985) *EMBO J* 4:2411-2418; De Almeida *et al.*, (1989) *Mol Gen Genetics* 218:78-86), and thus that multiple events are typically screened in order to obtain lines displaying the desired expression level and pattern. Such screening may be accomplished standard molecular biological, biochemical, and other assays including Southern analysis of DNA, Northern analysis of mRNA expression, PCR, real time quantitative PCR (qPCR), reverse transcription PCR (RT-PCR), immunoblotting analysis of protein expression, enzyme or activity assays, and/or phenotypic analysis.

The term “expression”, as used herein, refers to the production of a functional end-product (e.g., an mRNA, guide RNA, or a protein) in either precursor or mature form.

The term “providing” includes providing a nucleic acid (e.g., expression construct) or protein into a cell. Providing includes reference to the incorporation of a nucleic acid into a eukaryotic or prokaryotic cell where the nucleic acid may be incorporated into the genome of the cell, and includes reference to the transient provision of a nucleic acid or protein to the cell. Introduced includes reference to stable or transient transformation methods, as well as sexually crossing. Thus, “providing” in the context of inserting a nucleic acid fragment (e.g., a recombinant DNA construct/expression construct) into a cell, means “transfection” or “transformation” or “transduction” and includes reference to the incorporation of a nucleic acid fragment into a eukaryotic or prokaryotic cell where the nucleic acid fragment may be incorporated into the genome of the cell (e.g., chromosome, plasmid, plastid, or mitochondrial DNA), converted into an autonomous replicon, or transiently expressed (e.g., transfected mRNA).

“Mature” protein refers to a post-translationally processed polypeptide (i.e., one from which any pre- or propeptides present in the primary translation product have been removed). “Precursor” protein refers to the primary product of translation

of mRNA (i.e., with pre- and propeptides still present). Pre- and propeptides may be but are not limited to intracellular localization signals.

"Stable transformation" refers to the transfer of a nucleic acid fragment into a genome of a host organism, including both nuclear and organellar genomes, resulting in genetically stable inheritance. In contrast, "transient transformation" refers to the transfer of a nucleic acid fragment into the nucleus, or other DNA-containing organelle, of a host organism resulting in gene expression without integration or stable inheritance. Host organisms containing the transformed nucleic acid fragments are referred to as "transgenic" organisms.

The commercial development of genetically improved germplasm has also advanced to the stage of introducing multiple traits into crop plants, often referred to as a gene stacking approach. In this approach, multiple genes conferring different characteristics of interest can be introduced into a plant. Gene stacking can be accomplished by many means including but not limited to co-transformation, retransformation, and crossing lines with different genes of interest.

The term "plant" refers to whole plants, plant organs, plant tissues, seeds, plant cells, seeds and progeny of the same. Plant cells include, without limitation, cells from seeds, suspension cultures, embryos, meristematic regions, callus tissue, leaves, roots, shoots, gametophytes, sporophytes, pollen and microspores. Plant parts include differentiated and undifferentiated tissues including, but not limited to roots, stems, shoots, leaves, pollens, seeds, tumor tissue and various forms of cells and culture (e.g., single cells, protoplasts, embryos, and callus tissue). The plant tissue may be in plant or in a plant organ, tissue or cell culture. The term "plant organ" refers to plant tissue or a group of tissues that constitute a morphologically and functionally distinct part of a plant. The term "genome" refers to the entire complement of genetic material (genes and non-coding sequences) that is present in each cell of an organism, or virus or organelle; and/or a complete set of chromosomes inherited as a (haploid) unit from one parent. "Progeny" comprises any subsequent generation of a plant.

A transgenic plant includes, for example, a plant which comprises within its genome a heterologous polynucleotide introduced by a transformation step. The heterologous polynucleotide can be stably integrated within the genome such that

the polynucleotide is passed on to successive generations. The heterologous polynucleotide may be integrated into the genome alone or as part of a recombinant DNA construct. A transgenic plant can also comprise more than one heterologous polynucleotide within its genome. Each heterologous polynucleotide may confer a different trait to the transgenic plant. A heterologous polynucleotide can include a sequence that originates from a foreign species, or, if from the same species, can be substantially modified from its native form. Transgenic can include any cell, cell line, callus, tissue, plant part or plant, the genotype of which has been altered by the presence of heterologous nucleic acid including those transgenics initially so altered as well as those created by sexual crosses or asexual propagation from the initial transgenic. The alterations of the genome (chromosomal or extra-chromosomal) by conventional plant breeding methods, by the genome editing procedure described herein that does not result in an insertion of a foreign polynucleotide, or by naturally occurring events such as random cross-fertilization, non-recombinant viral infection, non-recombinant bacterial transformation, non-recombinant transposition, or spontaneous mutation are not intended to be regarded as transgenic.

In certain embodiments of the disclosure, a fertile plant is a plant that produces viable male and female gametes and is self-fertile. Such a self-fertile plant can produce a progeny plant without the contribution from any other plant of a gamete and the genetic material contained therein. Other embodiments of the disclosure can involve the use of a plant that is not self-fertile because the plant does not produce male gametes, or female gametes, or both, that are viable or otherwise capable of fertilization. As used herein, a "male sterile plant" is a plant that does not produce male gametes that are viable or otherwise capable of fertilization. As used herein, a "female sterile plant" is a plant that does not produce female gametes that are viable or otherwise capable of fertilization. It is recognized that male-sterile and female-sterile plants can be female-fertile and male- fertile, respectively. It is further recognized that a male fertile (but female sterile) plant can produce viable progeny when crossed with a female fertile plant and that a female fertile (but male sterile) plant can produce viable progeny when crossed with a male fertile plant.

The term “non-conventional yeast” herein refers to any yeast that is not a *Saccharomyces* (e.g., *S. cerevisiae*) or *Schizosaccharomyces* yeast species. Non-conventional yeast are described in Non-Conventional Yeasts in Genetics, Biochemistry and Biotechnology: Practical Protocols (K. Wolf, K.D. Breunig, G. Barth, Eds., Springer-Verlag, Berlin, Germany, 2003), which is incorporated herein by reference. Non-conventional yeast in certain embodiments may additionally (or alternatively) be yeast that favor non-homologous end-joining (NHEJ) DNA repair processes over repair processes mediated by homologous recombination (HR). Definition of a non-conventional yeast along these lines – preference of NHEJ over HR – is further disclosed by Chen et al. (*PLoS ONE* 8:e57952), which is incorporated herein by reference. Preferred non-conventional yeast herein are those of the genus *Yarrowia* (e.g., *Yarrowia lipolytica*). The term “yeast” herein refers to fungal species that predominantly exist in unicellular form. Yeast can alternative be referred to as “yeast cells” herein. (see also US provisional application 62/036,652, filed on August 13, 2014, which is incorporated by reference herein.

A "centimorgan" (cM) or "map unit" is the distance between two linked genes, markers, target sites, loci, or any pair thereof, wherein 1% of the products of meiosis are recombinant. Thus, a centimorgan is equivalent to a distance equal to a 1% average recombination frequency between the two linked genes, markers, target sites, loci, or any pair thereof.

The present disclosure finds use in the breeding of plants comprising one or more transgenic traits. Most commonly, transgenic traits are randomly inserted throughout the plant genome as a consequence of transformation systems based on *Agrobacterium*, biolistics, or other commonly used procedures. More recently, gene targeting protocols have been developed that enable directed transgene insertion. One important technology, site-specific integration (SSI) enables the targeting of a transgene to the same chromosomal location as a previously inserted transgene. Custom-designed meganucleases and custom-designed zinc finger meganucleases allow researchers to design nucleases to target specific chromosomal locations, and these reagents allow the targeting of transgenes at the chromosomal site cleaved by these nucleases.

The currently used systems for precision genetic engineering of eukaryotic genomes, e.g. plant genomes, rely upon homing endonucleases, meganucleases, zinc finger nucleases, and transcription activator–like effector nucleases (TALENs), which require de novo protein engineering for every new target locus. The highly specific, RNA-directed DNA nuclease, guide RNA/ Cas9 endonuclease system described herein, is more easily customizable and therefore more useful when modification of many different target sequences is the goal. This disclosure takes further advantage of the two component nature of the guide RNA/ Cas system, with its constant protein component, the Cas endonuclease, and its variable and easily reprogrammable targeting component, the guide RNA or the crRNA.

The guide RNA/Cas system described herein is especially useful for genome engineering, especially plant genome engineering, in circumstances where nuclease off-target cutting can be toxic to the targeted cells. In one embodiment of the guide RNA/Cas system described herein, the constant component, in the form of an expression-optimized Cas9 gene, is stably integrated into the target genome, e.g. plant genome. Expression of the Cas9 gene is under control of a promoter, e.g. plant promoter, which can be a constitutive promoter, tissue-specific promoter or inducible promoter, e.g. temperature-inducible, stress-inducible, developmental stage inducible, or chemically inducible promoter. In the absence of the variable component, i.e. the guide RNA or crRNA, the Cas9 protein is not able to cut DNA and therefore its presence in the plant cell should have little or no consequence. Hence a key advantage of the guide RNA/Cas system described herein is the ability to create and maintain a cell line or transgenic organism capable of efficient expression of the Cas9 protein with little or no consequence to cell viability. In order to induce cutting at desired genomic sites to achieve targeted genetic modifications, guide RNAs or crRNAs can be introduced by a variety of methods into cells containing the stably-integrated and expressed cas9 gene. For example, guide RNAs or crRNAs can be chemically or enzymatically synthesized, and introduced into the Cas9 expressing cells via direct delivery methods such a particle bombardment or electroporation.

Alternatively, genes capable of efficiently expressing guide RNAs or crRNAs in the target cells can be synthesized chemically, enzymatically or in a biological

system, and these genes can be introduced into the Cas9 expressing cells via direct delivery methods such a particle bombardment, electroporation or biological delivery methods such as Agrobacterium mediated DNA delivery.

A guide RNA/Cas system mediating gene targeting can be used in methods for directing transgene insertion and / or for producing complex transgenic trait loci comprising multiple transgenes in a fashion similar as disclosed in WO2013/0198888 (published August 1, 2013) where instead of using a double strand break inducing agent to introduce a gene of interest, a guide RNA/Cas system as disclosed herein is used. In one embodiment, a complex transgenic trait locus is a genomic locus that has multiple transgenes genetically linked to each other. By inserting independent transgenes within 0.1, 0.2, 0.3, 0.4, 0.5 , 1.0, 2, or even 5 centimorgans (cM) from each other, the transgenes can be bred as a single genetic locus (see, for example, U.S. patent application 13/427,138) or PCT application PCT/US2012/030061. After selecting a plant comprising a transgene, plants containing (at least) one transgenes can be crossed to form an F1 that contains both transgenes. In progeny from these F1 (F2 or BC1) 1/500 progeny would have the two different transgenes recombined onto the same chromosome. The complex locus can then be bred as single genetic locus with both transgene traits. This process can be repeated to stack as many traits as desired.

Chromosomal intervals that correlate with a phenotype or trait of interest can be identified. A variety of methods well known in the art are available for identifying chromosomal intervals. The boundaries of such chromosomal intervals are drawn to encompass markers that will be linked to the gene controlling the trait of interest. In other words, the chromosomal interval is drawn such that any marker that lies within that interval (including the terminal markers that define the boundaries of the interval) can be used as a marker for northern leaf blight resistance. In one embodiment, the chromosomal interval comprises at least one QTL, and furthermore, may indeed comprise more than one QTL. Close proximity of multiple QTLs in the same interval may obfuscate the correlation of a particular marker with a particular QTL, as one marker may demonstrate linkage to more than one QTL. Conversely, e.g., if two markers in close proximity show co-segregation with the desired phenotypic trait, it is sometimes unclear if each of those markers identifies

the same QTL or two different QTL. The term "quantitative trait locus" or "QTL" refers to a region of DNA that is associated with the differential expression of a quantitative phenotypic trait in at least one genetic background, e.g., in at least one breeding population. The region of the QTL encompasses or is closely linked to the gene or genes that affect the trait in question. An "allele of a QTL" can comprise multiple genes or other genetic factors within a contiguous genomic region or linkage group, such as a haplotype. An allele of a QTL can denote a haplotype within a specified window wherein said window is a contiguous genomic region that can be defined, and tracked, with a set of one or more polymorphic markers. A haplotype can be defined by the unique fingerprint of alleles at each marker within the specified window.

A variety of methods are available to identify those cells having an altered genome at or near a target site without using a screenable marker phenotype. Such methods can be viewed as directly analyzing a target sequence to detect any change in the target sequence, including but not limited to PCR methods, sequencing methods, nuclease digestion, Southern blots, and any combination thereof.

Proteins may be altered in various ways including amino acid substitutions, deletions, truncations, and insertions. Methods for such manipulations are generally known. For example, amino acid sequence variants of the protein(s) can be prepared by mutations in the DNA. Methods for mutagenesis and nucleotide sequence alterations include, for example, Kunkel, (1985) *Proc. Natl. Acad. Sci. USA* 82:488-92; Kunkel *et al.*, (1987) *Meth Enzymol* 154:367-82; U.S. Patent No. 4,873,192; Walker and Gaastra, eds. (1983) *Techniques in Molecular Biology* (MacMillan Publishing Company, New York) and the references cited therein. Guidance regarding amino acid substitutions not likely to affect biological activity of the protein is found, for example, in the model of Dayhoff *et al.*, (1978) *Atlas of Protein Sequence and Structure* (Natl Biomed Res Found, Washington, D.C.). Conservative substitutions, such as exchanging one amino acid with another having similar properties, may be preferable. Conservative deletions, insertions, and amino acid substitutions are not expected to produce radical changes in the characteristics of the protein, and the effect of any substitution, deletion, insertion, or combination

thereof can be evaluated by routine screening assays. Assays for double-strand-break-inducing activity are known and generally measure the overall activity and specificity of the agent on DNA substrates containing target sites.

A variety of methods are known for the introduction of nucleotide sequences and polypeptides into an organism, including, for example, transformation, sexual crossing, and the introduction of the polypeptide, DNA, or mRNA into the cell.

Methods for contacting, providing, and/or introducing a composition into various organisms are known and include but are not limited to, stable transformation methods, transient transformation methods, virus-mediated methods, and sexual breeding. Stable transformation indicates that the introduced polynucleotide integrates into the genome of the organism and is capable of being inherited by progeny thereof. Transient transformation indicates that the introduced composition is only temporarily expressed or present in the organism.

Protocols for introducing polynucleotides and polypeptides into plants may vary depending on the type of plant or plant cell targeted for transformation, such as monocot or dicot. Suitable methods of introducing polynucleotides and polypeptides into plant cells and subsequent insertion into the plant genome include microinjection (Crossway *et al.*, (1986) *Biotechniques* 4:320-34 and U.S. Patent No. 6,300,543), meristem transformation (U.S. Patent No. 5,736,369), electroporation (Riggs *et al.*, (1986) *Proc. Natl. Acad. Sci. USA* 83:5602-6, *Agrobacterium*-mediated transformation (U.S. Patent Nos. 5,563,055 and 5,981,840), direct gene transfer (Paszkowski *et al.*, (1984) *EMBO J* 3:2717-22), and ballistic particle acceleration (U.S. Patent Nos. 4,945,050; 5,879,918; 5,886,244; 5,932,782; Tomes *et al.*, (1995) "Direct DNA Transfer into Intact Plant Cells via Microprojectile Bombardment" in *Plant Cell, Tissue, and Organ Culture: Fundamental Methods,* ed. Gamborg & Phillips (Springer-Verlag, Berlin); McCabe *et al.*, (1988) *Biotechnology* 6:923-6; Weissinger *et al.*, (1988) *Ann Rev Genet* 22:421-77; Sanford *et al.*, (1987) *Particulate Science and Technology* 5:27-37 (onion); Christou *et al.*, (1988) *Plant Physiol* 87:671-4 (soybean); Finer and McMullen, (1991) *In Vitro Cell Dev Biol* 27P:175-82 (soybean); Singh *et al.*, (1998) *Theor Appl Genet* 96:319-24 (soybean); Datta *et al.*, (1990) *Biotechnology* 8:736-40 (rice); Klein *et al.*, (1988) *Proc. Natl. Acad. Sci. USA* 85:4305-9 (maize); Klein *et al.*, (1988) *Biotechnology* 6:559-63

(maize); U.S. Patent Nos. 5,240,855; 5,322,783 and 5,324,646; Klein *et al.*, (1988) *Plant Physiol* 91:440-4 (maize); Fromm *et al.*, (1990) *Biotechnology* 8:833-9 (maize); Hooykaas-Van Slogteren *et al.*, (1984) *Nature* 311:763-4; U.S. Patent No. 5,736,369 (cereals); Bytebier *et al.*, (1987) *Proc. Natl. Acad. Sci. USA* 84:5345-9 (*Liliaceae*); De Wet *et al.*, (1985) in *The Experimental Manipulation of Ovule Tissues,* ed. Chapman *et al.*, (Longman, New York), pp. 197-209 (pollen); Kaeppler *et al.*, (1990) *Plant Cell Rep* 9:415-8) and Kaeppler *et al.*, (1992) *Theor Appl Genet* 84:560-6 (whisker-mediated transformation); D'Halluin *et al.*, (1992) *Plant Cell* 4:1495-505 (electroporation); Li *et al.*, (1993) *Plant Cell Rep* 12:250-5; Christou and Ford (1995) *Annals Botany* 75:407-13 (rice) and Osjoda *et al.*, (1996) *Nat Biotechnol* 14:745-50 (maize via *Agrobacterium tumefaciens*).

Alternatively, polynucleotides may be introduced into plants by contacting plants with a virus or viral nucleic acids. Generally, such methods involve incorporating a polynucleotide within a viral DNA or RNA molecule. In some examples a polypeptide of interest may be initially synthesized as part of a viral polyprotein, which is later processed by proteolysis *in vivo* or *in vitro* to produce the desired recombinant protein. Methods for introducing polynucleotides into plants and expressing a protein encoded therein, involving viral DNA or RNA molecules, are known, see, for example, U.S. Patent Nos. 5,889,191, 5,889,190, 5,866,785, 5,589,367 and 5,316,931. Transient transformation methods include, but are not limited to, the introduction of polypeptides, such as a double-strand break inducing agent, directly into the organism, the introduction of polynucleotides such as DNA and/or RNA polynucleotides, and the introduction of the RNA transcript, such as an mRNA encoding a double-strand break inducing agent, into the organism. Such methods include, for example, microinjection or particle bombardment. See, for example Crossway *et al.*, (1986) *Mol Gen Genet* 202:179-85; Nomura *et al.*, (1986) *Plant Sci* 44:53-8; Hepler *et al.*, (1994) *Proc. Natl. Acad. Sci. USA* 91:2176-80; and, Hush *et al.*, (1994) *J Cell Sci* 107:775-84.

The term "dicot" refers to the subclass of angiosperm plants also knows as "dicotyledoneae" and includes reference to whole plants, plant organs (e.g., leaves, stems, roots, etc.), seeds, plant cells, and progeny of the same. Plant cell, as used herein includes, without limitation, seeds, suspension cultures, embryos,

meristematic regions, callus tissue, leaves, roots, shoots, gametophytes, sporophytes, pollen, and microspores.

The term "crossed" or "cross" or "crossing" in the context of this disclosure means the fusion of gametes via pollination to produce progeny (i.e., cells, seeds, or plants). The term encompasses both sexual crosses (the pollination of one plant by another) and selfing (self-pollination, i.e., when the pollen and ovule (or microspores and megaspores) are from the same plant or genetically identical plants).

The term "introgression" refers to the transmission of a desired allele of a genetic locus from one genetic background to another. For example, introgression of a desired allele at a specified locus can be transmitted to at least one progeny plant via a sexual cross between two parent plants, where at least one of the parent plants has the desired allele within its genome. Alternatively, for example, transmission of an allele can occur by recombination between two donor genomes, e.g., in a fused protoplast, where at least one of the donor protoplasts has the desired allele in its genome. The desired allele can be, e.g., a transgene, a modified (mutated or edited) native allele, or a selected allele of a marker or QTL.

Standard DNA isolation, purification, molecular cloning, vector construction, and verification/characterization methods are well established, see, for example Sambrook *et al.*, (1989) *Molecular Cloning: A Laboratory Manual*, (Cold Spring Harbor Laboratory Press, NY). Vectors and constructs include circular plasmids, and linear polynucleotides, comprising a polynucleotide of interest and optionally other components including linkers, adapters, regulatory or analysis. In some examples a recognition site and/or target site can be contained within an intron, coding sequence, 5' UTRs, 3' UTRs, and/or regulatory regions.

The present disclosure further provides expression constructs for expressing in a plant, plant cell, or plant part a guide RNA/Cas system that is capable of binding to and creating a double strand break in a target site. In one embodiment, the expression constructs of the disclosure comprise a promoter operably linked to a nucleotide sequence encoding a Cas gene and a promoter operably linked to a guide RNA of the present disclosure. The promoter is capable of driving expression of an operably linked nucleotide sequence in a plant cell.

A phenotypic marker is a screenable or selectable marker that includes visual markers and selectable markers whether it is a positive or negative selectable marker. Any phenotypic marker can be used. Specifically, a selectable or screenable marker comprises a DNA segment that allows one to identify, or select for or against a molecule or a cell that contains it, often under particular conditions. These markers can encode an activity, such as, but not limited to, production of RNA, peptide, or protein, or can provide a binding site for RNA, peptides, proteins, inorganic and organic compounds or compositions and the like.

Examples of selectable markers include, but are not limited to, DNA segments that comprise restriction enzyme sites; DNA segments that encode products which provide resistance against otherwise toxic compounds including antibiotics, such as, spectinomycin, ampicillin, kanamycin, tetracycline, Basta, neomycin phosphotransferase II (NEO) and hygromycin phosphotransferase (HPT)); DNA segments that encode products which are otherwise lacking in the recipient cell (e.g., tRNA genes, auxotrophic markers); DNA segments that encode products which can be readily identified (e.g., phenotypic markers such as β-galactosidase, GUS; fluorescent proteins such as green fluorescent protein (GFP), cyan (CFP), yellow (YFP), red (RFP), and cell surface proteins); the generation of new primer sites for PCR (e.g., the juxtaposition of two DNA sequence not previously juxtaposed), the inclusion of DNA sequences not acted upon or acted upon by a restriction endonuclease or other DNA modifying enzyme, chemical, etc.; and, the inclusion of a DNA sequences required for a specific modification (e.g., methylation) that allows its identification.

Additional selectable markers include genes that confer resistance to herbicidal compounds, such as glufosinate ammonium, bromoxynil, imidazolinones, and 2,4-dichlorophenoxyacetate (2,4-D). See for example, Yarranton, (1992) *Curr Opin Biotech* 3:506-11; Christopherson *et al.*, (1992) *Proc. Natl. Acad. Sci. USA* 89:6314-8; Yao *et al.*, (1992) *Cell* 71:63-72; Reznikoff, (1992) *Mol Microbiol* 6:2419-22; Hu *et al.*, (1987) *Cell* 48:555-66; Brown *et al.*, (1987) *Cell* 49:603-12; Figge *et al.*, (1988) *Cell* 52:713-22; Deuschle *et al.*, (1989) *Proc. Natl. Acad. Sci. USA* 86:5400-4; Fuerst *et al.*, (1989) *Proc. Natl. Acad. Sci. USA* 86:2549-53; Deuschle *et al.*, (1990) *Science* 248:480-3; Gossen, (1993) Ph.D. Thesis, University of

Heidelberg; Reines *et al.*, (1993) *Proc. Natl. Acad. Sci. USA* 90:1917-21; Labow *et al.*, (1990) *Mol Cell Biol* 10:3343-56; Zambretti *et al.*, (1992) *Proc. Natl. Acad. Sci. USA* 89:3952-6; Baim *et al.*, (1991) *Proc. Natl. Acad. Sci. USA* 88:5072-6; Wyborski *et al.*, (1991) *Nucleic Acids Res* 19:4647-53; Hillen and Wissman, (1989) *Topics Mol Struc Biol* 10:143-62; Degenkolb *et al.*, (1991) *Antimicrob Agents Chemother* 35:1591-5; Kleinschnidt *et al.*, (1988) *Biochemistry* 27:1094-104; Bonin, (1993) Ph.D. Thesis, University of Heidelberg; Gossen *et al.*, (1992) *Proc. Natl. Acad. Sci. USA* 89:5547-51; Oliva *et al.*, (1992) *Antimicrob Agents Chemother* 36:913-9; Hlavka *et al.*, (1985) *Handbook of Experimental Pharmacology,* Vol. 78 (Springer-Verlag, Berlin); Gill *et al.*, (1988) *Nature* 334:721-4.

The cells having the introduced sequence may be grown or regenerated into plants using conventional conditions, see for example, McCormick *et al.*, (1986) *Plant Cell Rep* 5:81-4. These plants may then be grown, and either pollinated with the same transformed strain or with a different transformed or untransformed strain, and the resulting progeny having the desired characteristic and/or comprising the introduced polynucleotide or polypeptide identified. Two or more generations may be grown to ensure that the polynucleotide is stably maintained and inherited, and seeds harvested.

Any plant can be used, including monocot and dicot plants. Examples of monocot plants that can be used include, but are not limited to, corn (*Zea mays*), rice (*Oryza sativa*), rye (*Secale cereale*), sorghum (*Sorghum bicolor*, *Sorghum vulgare*), millet (e.g., pearl millet (*Pennisetum glaucum*), proso millet (*Panicum miliaceum*), foxtail millet (*Setaria italica*), finger millet (*Eleusine coracana*)), wheat (*Triticum aestivum*), sugarcane (*Saccharum spp.*), oats (*Avena*), barley (*Hordeum*), switchgrass (*Panicum virgatum*), pineapple (*Ananas comosus*), banana (*Musa spp.*), palm, ornamentals, turfgrasses, and other grasses. Examples of dicot plants that can be used include, but are not limited to, soybean (*Glycine max*), canola (*Brassica napus* and *B. campestris*), alfalfa (*Medicago sativa*), tobacco (*Nicotiana tabacum*), *Arabidopsis* (*Arabidopsis thaliana*), sunflower (*Helianthus annuus*), cotton (*Gossypium arboreum*), and peanut (*Arachis hypogaea*), tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*) *etc*.

The transgenes, recombinant DNA molecules, DNA sequences of interest, and polynucleotides of interest can comprise one or more genes of interest. Such genes of interest can encode, for example, a protein that provides agronomic advantage to the plant.

Also provided are kits for performing any of the above methods described herein. The kits typically contain polynucleotides encoding one or more Cas endonuclease, or Cas endonuclease protein wherein the Cas endonuclease protein is provided as a purified protein, a cell lysate comprising said Cas endonuclease, a dilution of a cell lysate comprising said Cas endonuclease, an in-vitro translation mixture or an dilution of an in-vitro translation mixture, and/or single or dual guide polynucleotides, and/or template polynucleotides for gene editing and/or donor polynucleotides for inserting polynucleotides of interest into a genome of interest, as described herein. The kit can further contain instructions for administering all these components into the cells. The kits can also contain cells, buffers for transformation of cells, culture media for cells, and/or buffers for performing assays. The kits can further contain one or more inhibitors of proteins involved in NHEJ, or components which promote or increase homology-dependent repair (HDR) and instructions for introducing the Cas endonucleases and inhibitors into the cells such that Cas endonuclease-mediated gene disruption and/or targeted integration is enhanced. Optionally, cells containing the target site(s) of the Cas endonuclease may also be included in the kits described herein.

Inhibitors of non-homologous end joining (NHEJ) are known in the art and include molecules, such as but not limited to small molecules that inhibits (decrease) the binding or activity of a DNA-dependent- protein kinase catalytic subunit (DNA-PKcs), a Poly(ADP-ribose) polymerase 1/2 (PARPI/2), a PARPI, Ku70/80, a DNA-PKcs, a XRCC4/XLF, a Ligase IV, a Ligase III, a XRCCI, an Artemis Polynucleotide Kinase (PNK), SCR7, and any one combinations thereof (Sfeir et al. 2015, TIBS Vol 40 (11), pp701-713; Srivastava, M. *et al.* An inhibitor of nonhomologous end-joining abrogates double-strand break repair and impedes cancer progression. *Cell* **151**, 1474–1487 (2012); US patent application US2014/0242702, published on August 28, 2014, herein incorporated in its entirety by reference). Other molecules that decrease the activity of the non-homologous

end joining (NHEJ) DNA repair complex are known in the art and include RNAi-molecules, antisense nucleic acid molecules, ribozymes, compounds inhibiting the formation of a functional DNA Ligase IV (LIG4) complex and compounds enhancing proteolytic degradation of a functional DNA Ligase IV complex (US patent application 2014/0304847, published on Oct 9, 2014, herein incorporated in its entirety by reference.

Activators of HDR are known in the art and include molecules, such as but not limited to RS1, RAD51 and RAD51B (Song et al. 2016 “RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency” Nature communications 7, Article number:10548; Takaku, M. *et al* 2009. Recombination activator function of the novel RAD51- and RAD51B-binding protein, human EVL. *J. Biol. Chem.* **284**, 14326–14336 (2009).

In certain embodiments, the kits comprise at least one construct with a target gene and a Cas endonuclease described herein capable of cleaving within or in close proximity to the target gene. Such kits are useful for optimization of cleavage conditions in a variety of varying host cell types. In one aspect, the kit is a kit useful for increasing gene disruption, gene editing and/or targeted integration following Cas endonuclease mediated cleavage of a cell's genome.

In one embodiment, the kit includes a Cas endonuclease described herein capable of cleaving within a known target locus within a genome, and may additionally comprise a template DNA for gene editing and/or a donor nucleic acid for introducing a polynucleotide of interest into the cell’s genome. Such kits are useful for optimization of conditions for template recognition, donor integration or for the construction of specifically modified cells, cell lines, and transgenic plants and animals containing gene disruptions , gene edits or targeted insertions. These and other aspects will be readily apparent to the skilled artisan in light of disclosure as a whole.

Also provided are kits containing any one or more of the elements disclosed in compositions described herein. In one aspect, the kits comprise a single guide polynucleotide comprising a crRNA, as described herein linked to a tracrRNA, wherein the crRNA comprises a variable targeting domain operably linked to a tracr mate sequence and/or one or more insertion sites for inserting or exchanging the

variable targeting domain upstream of the tracr mate sequence, wherein when expressed, the single guide polynucleotide directs sequence-specific binding of a guide polynucleotide/Cas endonuclease complex to a target sequence in a eukaryotic cell. In another aspect, the kits comprise a dual guide polynucleotide comprising a crRNA molecule and a tracrRNA molecule, as described herein, wherein the crRNA molecule comprises a variable targeting domain operably linked to a tracr mate sequence and/or one or more insertion sites for inserting or exchanging the variable targeting domain upstream of the tracr mate sequence, wherein when expressed, the dual guide polynucleotide directs sequence-specific binding of a guide polynucleotide/Cas endonuclease complex to a target sequence in a eukaryotic cell.

The kits can contain one or more vectors encoding the guide polynucleotides, Cas endonucleases and /or template DNAs and /or donor DNAs described herein, and or the kits can contain the elements (guide polynucleotides, DNA templates, DNA donors and/or Cas endonucleases in purified or non-purified forms).

In one aspect, the kit comprises a Cas endonuclease as described herein, and/or a polynucleotide modification template and /or a donor DNA for inserting a polynucleotide of interest as described herein.

Components may be provide individually or in combinations, and may be provided in any suitable container, such as a vial, a bottle, or a tube. . For example, a kit may provide one or more reaction or storage buffers. Reagents may be provided in a form that is usable in a particular assay, or in a form that requires addition of one or more other components before use (e.g. in concentrate or lyophilized form). A buffer can be any buffer, including but not limited to a sodium carbonate buffer, a sodium bicarbonate buffer, a borate buffer, a Tris buffer, a MOPS buffer, a HEPES buffer, and combinations thereof. In some embodiments, the buffer is alkaline. In some embodiments, the buffer has a pH from about 7 to about 10. In some aspects, the kit includes instructions in one or more languages, for example in more than one language.

The meaning of abbreviations is as follows: “sec” means second(s), “min” means minute(s), “h” means hour(s), “d” means day(s), “µL” means microliter(s), “mL” means milliliter(s), “L” means liter(s), “µM” means micromolar, “mM” means

millimolar, “M” means molar, “mmol” means millimole(s), “µmole” mean micromole(s), “g” means gram(s), “µg” means microgram(s), “ng” means nanogram(s), “U” means unit(s), “bp” means base pair(s) and “kb” means kilobase(s).

Non-limiting examples of compositions and methods disclosed herein are as follows:

1. A method for producing a plasmid DNA library containing a randomized Protospacer-Adjacent-Motif (PAM) sequence, the method comprising:
   a) providing a first single stranded oligonucleotide comprising a target sequence that can be recognized by a guide RNA/Cas endonuclease complex;
   b) providing a second single stranded oligonucleotide comprising a randomized PAM sequence adjacent to a nucleotide sequence capable of hybridizing with the target sequence of (a);
   c) producing an oligoduplex comprising said randomized PAM sequence by combining the first single stranded oligonucleotide of (a) and the second single stranded oligonucleotide of (b);
   d) producing a ligation product by ligating the oligoduplex from (c) with a linearized plasmid; and,
   e) transforming host cells with the ligation product of (e) and recovering multiple host cell colonies representing the plasmid library.
2. A method for producing a ligation product containing a randomized Protospacer-Adjacent-Motif (PAM) sequence, the method comprising:
   a) providing a first single stranded oligonucleotide comprising restriction endonuclease recognition site located upstream of a target sequence that can be recognized by a guide RNA/Cas endonuclease complex;
   b) providing a second single stranded oligonucleotide comprising a randomized PAM sequence adjacent a nucleotide sequence capable of hybridizing with the target sequence of (a);
   c) producing an oligoduplex comprising said randomized PAM sequence by combining the first single stranded oligonucleotide of (a) and the second single stranded oligonucleotide of (b); and,

d) producing a ligation product by ligating the oligoduplex from (c) with a linearized plasmid;

3. The method of embodiment 1, wherein the host cells of (e) are *E. coli* cells.

4. A ligation product produced by the method of anyone of embodiments 1-2.

5. A library of host cells produced by the method of embodiment 1.

6. The method of anyone of embodiments 1-2, wherein the first single stranded oligonucleotide comprises a restriction endonuclease recognition site located upstream of a target sequence and wherein the ligation product of (d) is produced by first cleaving the oligoduplex with a restriction endonuclease that recognizes the restriction endonuclease recognition site of (a) followed by ligating the cleaved oligoduplex from (d) with a linearized plasmid.

7. The method of anyone of embodiments 1-2, wherein the second single stranded oligonucleotide comprises a randomized PAM of at least 5 randomized nucleotides (5Ns).

8. The method of anyone of embodiments 1-2, wherein the second single stranded oligonucleotide comprises a randomized PAM of at least 7 randomized nucleotides (7Ns).

9. A method for identification of a Protospacer-Adjacent-Motif (PAM) sequence, the method comprising:

a) providing a library of plasmid DNAs, wherein each one of said plasmid DNAs comprises a randomized Protospacer-Adjacent-Motif sequence integrated adjacent to a target sequence that can be recognized by a guide RNA/Cas endonuclease complex;

b) providing to said library of plasmids a guide RNA and a Cas endonuclease protein, wherein said guide RNA and Cas endonuclease protein can form a complex that is capable of introducing a double strand break into the said target sequence, thereby creating a library of cleaved targets;

c) ligating adaptors to the library of cleaved targets of (b) allowing for the library of cleaved targets to be amplified;

d) amplifying the library of cleaved targets such that cleaved products containing the randomized PAM sequence are enriched, thereby producing a library of enriched PAM-sided targets;

e) sequencing the library of (a) and the library of enriched PAM-sided targets of (d) and identifying the nucleotide sequence adjacent to the cleaved targets of (b) on either strand of the plasmid DNA, wherein said nucleotide sequence represents a putative Protospacer-Adjacent-Motif sequences; and,

f) determining the fold enrichment of each nucleotide within the putative Protospacer-Adjacent-Motif sequence relative to the plasmid DNA library of (a).

10. The method of anyone of embodiments 1-2 and 9, wherein the randomized PAM sequence comprises at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 randomized nucleotides.

11. The method of anyone of anyone of embodiments 1-2 and 9, wherein the target sequence is at least 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 or 30 nucleotides in length.

12. The method of embodiment 9, wherein the Cas endonuclease is a Cas9 endonuclease from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755.

13. The method of embodiment 9, wherein the guide RNA comprises a single molecule of a chimeric non-naturally occurring crRNA linked to a tracrRNA.

14. The method of embodiment 9, wherein the guide RNA comprises a duplex molecule of a chimeric non-naturally occurring crRNA and a tracrRNA.

15. The method of embodiment 9, wherein the chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to a target sequence in the genome of an organism, wherein said crRNA is linked a tracrRNA originating from organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932,

Sulfurospirillum sp. SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755.

16. The method of embodiment 9, wherein the chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to a target sequence in the genome of an organism, wherein said crRNA can form a duplex with a tracrRNA originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755.

17. The method of embodiment 9, wherein the chimeric non-naturally occurring crRNA comprises at least a fragment of a crRNA originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755.

18. A recombinant construct comprising at least one of the Protospacer-Adjacent-Motif (PAM) sequence identified by the method of embodiment 9.

19. A method for identification of a tracrRNA of an organism, the method comprising:

a) providing a first single guide RNA candidate comprising a chimeric non-naturally occurring crRNA comprising a variable targeting domain capable of hybridizing to a target sequence in the genome of a cell, linked to a first nucleotide sequence representing the sense expression of a candidate tracrRNA naturally occurring in said organism ;

b) providing a second single guide RNA candidate comprising a chimeric non-naturally occurring crRNA comprising a variable targeting domain capable of hybridizing to a target sequence in the genome of said cell, linked to a second nucleotide sequence representing the sense expression of a candidate tracrRNA naturally occurring in said organism;

c) providing to the first and second single guide RNA candidates a Cas endonuclease protein, wherein said Cas endonuclease protein can form a complex with either the first single guide RNA candidate or the second single guide RNA candidate, wherein said complex is capable of introducing a double strand break into said target sequence; and,

d) identification of the first or second guide RNA candidate and its tracrRNA component that complexes to the Cas endonuclease of (c) and results in cleavage of the target sequence in the genome of said cell.

20. A method for identification of a tracrRNA of an organism, the method comprising:

a) identifying a CRISPR array repeat sequence in a genomic locus of said organism;

b) aligning the CRISPR array repeat sequence of (a) with the sequence of the genomic locus of (a) and identifying an antirepeat sequence that encodes a tracrRNA; and,

c) determining the transcriptional direction of the tracrRNA.

21. A guide RNA capable of forming a guide RNA/Cas endonuclease complex, wherein said guide RNA/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence, wherein said tracrRNA is originated from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* MIc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas*

tenax DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755.

22. A guide RNA capable of forming a guide RNA/Cas endonuclease complex, wherein said guide RNA/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a single molecule comprising a chimeric non-naturally occurring crRNA linked to a tracrRNA originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

23. A guide RNA capable of forming a guide RNA/Cas endonuclease complex, wherein said guide RNA/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said chimeric non-naturally occurring crRNA comprises at least a fragment of a crRNA originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

24. A guide RNA capable of forming a guide RNA/Cas endonuclease complex, wherein said guide RNA/Cas endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a single molecule comprising a tracrRNA linked to a chimeric non-naturally occurring

crRNA comprising at least a fragment of a crRNA originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC , *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

25. A guide RNA/Cas endonuclease complex comprising a Cas9 endonuclease originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, and at least one guide RNA, wherein said guide RNA/Cas9 endonuclease complex is capable of recognizing, binding to, and optionally nicking or cleaving all or part of a target sequence.

26. The guide RNA/Cas endonuclease complex of embodiment 25 comprising at least one guide RNA of any one of embodiments 21-24.

27. The guide RNA/Cas endonuclease complex of embodiment 25, wherein said target sequence is located in the genome of a cell.

28. The guide RNA/Cas endonuclease complex of embodiment 25, wherein said Cas endonuclease is a Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 35 and 81-91, or a functional fragment thereof, wherein said guide RNA/Cas9 endonuclease capable of recognizing, binding to, and optionally nicking or cleaving all or part of a specific DNA target sequence.

29. A method for modifying a target site in the genome of a cell, the method comprising providing to said cell at least one Cas9 endonuclease originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814,

Pediococcus pentosaceus SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, and at least one guide RNA, wherein said guide RNA and Cas endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site.

30. The method of embodiment 29, further comprising identifying at least one cell that has a modification at said target, wherein the modification at said target site is selected from the group consisting of (i) a replacement of at least one nucleotide, (ii) a deletion of at least one nucleotide, (iii) an insertion of at least one nucleotide, and (iv) any combination of (i) – (iii).

31. A method for editing a nucleotide sequence in the genome of a cell, the method comprising providing to said cell at least one Cas9 endonuclease originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, a polynucleotide modification template, and at least one guide RNA, wherein said polynucleotide modification template comprises at least one nucleotide modification of said nucleotide sequence, wherein said guide RNA and Cas endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site.

32. A method for modifying a target site in the genome of a cell, the method comprising providing to said cell at least one guide RNA, at least one donor DNA, and at least one Cas9 endonuclease originating from an organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*,

Sphingomonas sanxanigenens NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, wherein said at least one guide RNA and at least one Cas endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site, wherein said donor DNA comprises a polynucleotide of interest.

33. The method of embodiment 32, further comprising identifying at least one cell that said polynucleotide of interest integrated in or near said target site.

34. The method of any one of embodiments 29-33, wherein the cell is selected from the group consisting of a human, non-human, animal, bacterial, fungal, insect, yeast, non-conventional yeast, and plant cell.

35. The method of embodiment 34, wherein the plant cell is selected from the group consisting of a monocot and dicot cell.

36. The method of embodiment 35, wherein the plant cell is selected from the group consisting of maize, rice, sorghum, rye, barley, wheat, millet, oats, sugarcane, turfgrass, or switchgrass, soybean, canola, alfalfa, sunflower, cotton, tobacco, peanut, potato, tobacco, Arabidopsis, and safflower cell.

37. A plant comprising a modified target site, wherein said plant originates from a plant cell comprising a modified target site produced by the method of any of embodiments 29-36.

38. A plant comprising an edited nucleotide, wherein said plant originates from a plant cell comprising an edited nucleotide produced by the method of embodiment 31.

39. A method for designing a single guide RNA, the method comprising:

a) aligning a tracrRNA sequence with a CRISPR array repeat sequence from a genomic locus of an organism, wherein said CRISPR array repeat sequence comprises a crRNA sequence ;

b) deducing the transcriptional direction of the CRISPR array, thereby also deducing the crRNA sequence; and,

c) designing a single guide RNA comprising said tracrRNA and crRNA sequences

40. A method for producing target sequences, the method comprising:

a) identifying a polynucleotides of interest ;

b) introducing a Protospacer-Adjacent-Motif (PAM) sequence adjacent to said polynucleotide of interest, wherein said PAM sequence comprises the nucleotide sequence NNNNCND, thereby creating a thereby creating a target site for a guide RNA/Cas9 endonuclease complex; and,

c) identifying a polynucleotides of interest ;

41. The method for embodiment 40, wherein the guide RNA/Cas9 endonuclease complex, comprises at least one Cas9 endonuclease originated from organism selected from the group consisting of *Brevibacillus laterosporus, Lactobacillus reuteri* Mlc3, *Lactobacillus rossiae* DSM 15814, *Pediococcus pentosaceus* SL4, *Lactobacillus nodensis* JCM 14932, *Sulfurospirillum sp.* SCADC, *Bifidobacterium thermophilum* DSM 20210, *Loktanella vestfoldensis*, *Sphingomonas sanxanigenens* NX02, *Epilithonimonas tenax* DSM 16811, *Sporocytophaga myxococcoides* and *Psychroflexus torquis* ATCC 700755, wherein said guide RNA/Cas9 endonuclease complex is capable of recognizing, binding to, and optionally nicking or cleaving all or part of a target sequence

42. A method for producing a plasmid DNA library containing a randomized Protospacer-Adjacent-Motif (PAM) sequence, the method comprising transforming at least one host cell with a ligation product and recovering multiple host cell colonies representing the plasmid library, wherein said ligation product was generated by contacting a library of linear oligoduplexes with a linearized plasmid, wherein each oligoduplex member of said library of oligoduplexes comprises a first single stranded oligonucleotide comprising a-target sequence , and a second single stranded oligonucleotide comprising a randomized PAM sequence adjacent to a nucleotide sequence capable of hybridizing with said target sequence.

43. A method for identification of a Protospacer-Adjacent-Motif (PAM), the method comprising:

a) providing a library of plasmids, wherein each one of said plasmids comprise a randomized Protospacer-Adjacent-Motif sequence integrated adjacent to a target sequence that can be recognized by a guide RNA/Cas endonuclease complex;;

b) producing a 3 prime (3') or 5 prime (5') overhang into the target sequence of (a) by providing to the plasmids of (a) a 3 prime deoxy-adenine, a guide RNA and a Cas endonuclease protein, wherein said guide RNA and Cas endonuclease can form a complex that is capable of introducing a double strand break into said target sequence;

c) ligating adapters to the 3 prime or 5 prime overhang of (c), thereby creating a library of cleaved targets that can be amplified;

d) amplifying the library of cleaved targets such that cleaved products containing the randomized PAM sequence are enriched;

e) sequencing the library of (a) and the library of enriched PAM-sided targets of (d) and identifying the nucleotide sequence adjacent to the cleaved targets of (b) on either strand of the plasmid DNA, wherein said nucleotide sequence represents a putative Protospacer-Adjacent-Motif sequences; and,

f) determining the fold enrichment of each nucleotide within the putative Protospacer-Adjacent-Motif sequence relative to the plasmid DNA library of (a).

44. A single guide RNA selected from the group consisting of SEQ ID NOs: 47, 127, 114-125, and 128-139.

45. A single guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said single guide RNA is selected from the group consisting of SEQ ID NOs: 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138 and 139.

46. A single guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said single guide RNA comprises a chimeric non-naturally occurring crRNA linked to a tracrRNA, wherein said tracrRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183 and 184.

47. A single guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said single guide RNA comprises a chimeric non-naturally occurring crRNA linked to a tracrRNA, wherein said chimeric non-naturally occurring crRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159 and 160.

48. A guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence, wherein said tracrRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183 and 184, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

49. A guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said chimeric non-naturally occurring crRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159 and 160, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

50. A guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said tracrRNA comprises a

nucleotide sequence selected from the group consisting of SEQ ID NOs: 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183 and 184, wherein said chimeric non-naturally occurring crRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159 and 160, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

51. A guide RNA/Cas9 endonuclease complex comprising a Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81, 82, 83, 84, 85, 86, 87, 88, 89, 90 and 91, or a functional fragment thereof, and at least one guide RNA, wherein said guide RNA/Cas9 endonuclease complex is capable of recognizing, binding to, and optionally nicking or cleaving all or part of a target sequence.

52. A guide RNA/Cas9 endonuclease complex comprising at least one guide RNA and a Cas9 endonuclease, wherein said Cas9 endonuclease is encoded by a DNA sequence selected from the group consisting of SEQ ID NOs: 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, and 80, wherein said guide RNA/Cas9 endonuclease complex is capable of recognizing, binding to, and optionally nicking or cleaving all or part of a target sequence.

53. The guide RNA/Cas9 endonuclease complex of embodiment 7, wherein said guide RNA is selected from the group consisting of SEQ ID NOs: 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138 and 139.

54. The guide RNA/Cas9 endonuclease complex of embodiments 7, wherein said target sequence is located in the genome of a cell.

55. A method for modifying a target site in the genome of a cell, the method comprising providing to said cell at least one Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81, 82, 83, 84, 85, 86, 87, 88, 89, 90 and 91, or a functional fragment thereof, and at least one guide RNA, wherein said guide RNA and Cas9 endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site.

56. The method of embodiment 10, further comprising identifying at least one cell that has a modification at said target, wherein the modification at said target site is selected from the group consisting of (i) a replacement of at least one nucleotide, (ii) a deletion of at least one nucleotide, (iii) an insertion of at least one nucleotide, and (iv) any combination of (i) – (iii).

57. A method for editing a nucleotide sequence in the genome of a cell, the method comprising providing to said cell at least one Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81, 82, 83, 84, 85, 86, 87, 88, 89, 90 and 91, or a functional fragment thereof, a polynucleotide modification template, and at least one guide RNA, wherein said polynucleotide modification template comprises at least one nucleotide modification of said nucleotide sequence, wherein said guide RNA and Cas9 endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site.

58. A method for modifying a target site in the genome of a cell, the method comprising providing to said cell at least one guide RNA, at least one donor DNA, and at least one Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81, 82, 83, 84, 85, 86, 87, 88, 89, 90 and 91, or a functional fragment thereof, wherein said at least one guide RNA and at least one Cas9 endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site, wherein said donor DNA comprises a polynucleotide of interest.

59. The method of embodiments 11, 13 or 14 , wherein said guide RNA is selected from the group consisting of SEQ ID NOs: 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138 and 139.

60. The method of embodiment 13, further comprising identifying at least one cell that said polynucleotide of interest integrated in or near said target site.

61. The method of any one of embodiments 10-14, wherein the cell is selected from the group consisting of a human, non-human, animal, bacterial, fungal, insect, yeast, non-conventional yeast, and plant cell.

62. A single guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can

recognize, bind to, and optionally nick or cleave a target sequence, wherein said single guide RNA comprises a chimeric non-naturally occurring crRNA linked to a tracrRNA, wherein said tracrRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183 and 184., wherein said chimeric non-naturally occurring crRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159 and 160.

63. A kit for binding, cleaving or nicking a target sequence in eukaryotic cells or organisms comprising a guide RNA specific for said target DNA, and a Cas endonuclease protein selected from the group consisting of SEQ ID NOs: 81, 82, 83, 84, 85, 86, 87, 88, 89, 90 and 91.

64. A kit for cleaving a target sequence in eukaryotic cells or organisms comprising a guide RNA specific for said target DNA, and a Cas endonuclease protein, wherein said guide RNA is capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave said target sequence, wherein said guide RNA is selected from the group consisting of 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138 and 139.

65. A kit for targeted mutagenesis in eukaryotic cells or organisms comprising a guide RNA specific for said target DNA, a polynucleotide modification template, and a Cas endonuclease protein, wherein said guide RNA is capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave said target sequence, wherein said guide RNA is selected from the group consisting of SEQ ID NOs: 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138 and 139, wherein said Cas endonuclease protein is selected from the group consisting of SEQ ID NOs: 81, 82, 83, 84, 85, 86, 87, 88, 89, 90 and 91.

66. The kit of any one of embodiments 63-65, further comprising a molecule selected from the group consisting of an inhibitors of NHEJ, an activator of HDR or MMEJ repair pathways, an exogenous sequence, a homologous recombination DNA, a donor DNA, and any one combination thereof.

EXAMPLES

In the following Examples, unless otherwise stated, parts and percentages are by weight and degrees are Celsius. It should be understood that these Examples, while indicating embodiments of the disclosure, are given by way of illustration only. From the above discussion and these Examples, one skilled in the art can make various changes and modifications of the disclosure to adapt it to various usages and conditions. Such modifications are also intended to fall within the scope of the appended claims.

Example 1

Design and construction of 5N randomized Protospacer-Adjacent-Motif (PAM) library for assaying Cas9 PAM preferences.

To characterize the Protospacer-Adjacent-Motif (PAM) specificity of Cas9 proteins from Type II CRISPR (clustered, regularly interspaced, short palindromic repeats)-Cas (CRISPR-associated) nucleic acid-based adaptive immune systems found in most archaea and some bacteria, a plasmid DNA library containing a section of 5 random base pairs immediately adjacent to a 20 base pair target sequence, T1 (CGCTAAAGAGGAAGAGGACA (SEQ ID NO: 1), was developed. Randomization of the PAM sequence was generated through the synthesis of a single oligonucleotide, GG-821N (TGACCATGATTACGAATTCNNNNNTGTCCTCTTCCTCTTTAGCGAGC (SEQ ID NO: 2), with hand-mixing used to create a random incorporation of nucleotides across the 5 random residues (represented as N in the sequence of GG-821N). To convert the single stranded template of GG-821N into a double-stranded DNA template for cloning into the plasmid vector, a second oligonucleotide, GG-820 (AAGGATCCCCGGGTACCGAGCTGCTCGCTAAAGAGGAAGAGGAC (SEQ ID NO: 3), was synthesized with complementation to the 3' end of GG-821N to form a partial oligonucleotide duplex (oligoduplex I) as depicted in Figure 1. The partial duplex was then extended by PCR using DreamTaq polymerase (Thermo Fisher Scientific) to generate a full duplex containing the target sequence, 5 NNNNN randomized base pairs downstream of the target sequence and cleavage site for the BamHI restriction enzyme (oligoduplex II in Figure 1). To generate the plasmid library, the oligoduplex, purified using GeneJET PCR Purification Kit (Thermo Fisher

Scientific), was digested with BamHI and ligated into pTZ57R/T vector (Thermo Fisher Scientific) pre-cleaved with BamHI. Linear pTZ57R/T vector contains protruding ddT nucleotide at the 3' ends, whereas PCR fragments generated with DreamTaq polymerase contains dA at the 3' ends. Therefore one end of the PCR fragment is ligated into the vector through BamHI sticky ends, while another through A/T ends (Figure 2). The *E. coli* DH5α strain was transformed (Ca^{2+} transformation) with the ligated plasmid library and plated onto Luria Broth (LB) containing agar. The transformation efficiency was estimated from plated dilutions. Overall, ~12,000 colonies were recovered. The colonies were harvested from the plate by gently resuspending them in liquid LB media and plasmid DNA was purified using GeneJET Plasmid Miniprep kit (Thermo Fisher Scientific).

To validate the randomness of the resulting PAM library, PCR fragments spanning the 5 bp randomized PAM region were generated by Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific) amplification (15 cycles of a 2-step amplification protocol) using a TK-119 (GAGCTCGCTAAAGAGGAAGAGG (SEQ ID NO: 4) and pUC-dir (GCCAGGGTTTTCCCAGTCACGA (SEQ ID NO: 5) primer pair and 50 ng of plasmid DNA library as template. The resulting 122 bp PCR product was purified using GeneJET PCR Purification Kit (Thermo Fisher Scientific). 40 ng of the resulting PCR product was then amplified with Phusion® High Fidelity PCR Master Mix (New England Biolabs, M0531L) adding on the sequences necessary for amplicon-specific barcodes and Illumnia sequencing using "tailed" primers through two rounds of PCR each consisting of 10 cycles. The primers used in the primary PCR reaction are shown in Table 2 and a set of primers (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG (Universal Forward, SEQ ID NO: 8) and CAAGCAGAAGACGGCATA (Universal Reverse, SEQ ID NO: 9) universal to all primary PCR reactions were utilized for the secondary PCR amplification. The resulting PCR amplifications were purified with a Qiagen PCR purification spin column, concentration measured with a Hoechst dye-based fluorometric assay, combined in an equimolar ratio, and single read 60-100 nucleotide-length deep sequencing was performed on Illumina's MiSeq Personal Sequencer with a 5-10% (v/v) spike of PhiX control v3 (Illumina, FC-110-3001) to off-set sequence bias. The PAM sequence for only those reads containing a perfect

12 nt sequence match flanking either side of the 5 nucleotide randomized PAM sequence were captured and used to examine the frequency and diversity of PAM sequences present in the library. The frequency of each PAM sequence was calculated by dividing the number of reads with a given PAM by the total number of reads. The PAM sequence distribution was visualized by ordering the frequency of each PAM from greatest to least and displaying them graphically and by calculating the standard deviation of the resulting PAM frequencies relative to the average. As shown in Figure 4, all 1,024 possible PAM sequences were present at an average frequency of 0.10% with a coefficient of variation of 40.86%.

Table 2. Primary PCR primer sequences for tailing on the sequences needed for Illumina deep sequencing of initial uncut 5 bp randomized PAM pTZ57R/T library.

Primer Name	Primer Orientation	Primary PCR Primer Sequence	SEQ ID NO.
JKYS800.1	Forward	CTACACTCTTTCCCTACACGACGCTCTTCCG ATCTAAGTGAGCTCGCTAAAGAGGAAGA	6
JKYS803	Reverse	CAAGCAGAAGACGGCATACGAGCTCTTCCG ATCTGAATTCGAGCTCGGTACCT	7

Example 2

Protein expression and purification of *Streptococcus pyogenes*, *Streptococcus thermophilus* CRISPR1 and *Streptococcus thermophilus* CRISPR3 Cas9 proteins.

To examine the PAM specificity of the Cas9 proteins from the *Streptococcus pyogenes* (Spy) (Jinek *et al*. (2012) *Science* 337:816-21), *Streptococcus thermophilus* CRISPR1 (Sth1) (Horvath *et al*. (2008) *Journal of Bacteriology* 190:1401–12) and *Streptococcus thermophilus* CRISPR3 (Sth3) (Horvath *et al*. (2008) *Journal of Bacteriology* 190:1401–12) Type II CRISPR-Cas systems, Spy, Sth1 and Sth3 Cas9 proteins were *E. coli* expressed and purified. Briefly, the *cas9* genes of the CRISPR1-Cas and CRISPR3-Cas systems of *Streptococcus thermophilus* (Sth1 and Sth3) were amplified from a genomic DNA sample, while the *cas9* gene of *Streptococcus pyogenes* (Spy) was amplified from a plasmid, pMJ806 (Addgene plasmid # 39312)). DNA fragments encoding Sth1, Sth3 and Spy Cas9

were PCR amplified using Sth1-dir/Sth1-rev (ACGTCTCACATGACTAAGCCATACTCAATTGGAC (SEQ ID NO: 10); ACTCGAGACCCTCTCCTAGTTTGGCAA (SEQ ID NO: 11), Sth3-dir/Sth3-rev (GGGGGGTCTCACATGAGTGACTTAGT (SEQ ID NO: 12); AATTACTCGAGAAAATCTAGCTTAGGCTTA (SEQ ID NO: 13) and Spy-dir/Spy-rev (AAGGTCTCCCATGGATAAGAAATACTCAATAGGCTTAG (SEQ ID NO: 14); TTCTCGAGGTCACCTCCTAGCTGACTCAAATC (SEQ ID NO: 15) primer pairs, accordingly, and ligated into a pBAD24-CHis expression vector digested over NcoI and XhoI sites.

Sth3 and Spy Cas9 proteins were expressed in *E. coli* DH10HB strain grown in LB broth supplemented with ampicillin (100mg/ml). Cells were grown at 37°C to an OD 600 of 0.5 at which time the growth temperature was decreased to 16°C and expression induced with 0.2% (w/v) arabinose for 20 h. Cells were pelleted and resuspended in loading buffer (20 mM KH_2PO_4 pH7.0, 0.5 M NaCl, 10 mM imidazole, 5% glycerol) and disrupted by sonication. Cell debris was removed by centrifugation. The supernatant was loaded onto the Ni^{2+}-charged 5ml HiTrap chelating HP column (GE Healthcare) and eluted with a linear gradient of increasing imidazole concentration. The fractions containing Cas9 were pooled and subsequently loaded onto HiTrap heparin HP column (GE Healthcare) for elution using a linear gradient of increasing NaCl concentration (from 0.5 to 1 M NaCl). The fractions containing Cas9 were pooled and dialyzed against 10 mM Bis-Tris-HCl pH 7.0, 300 mM KCl, 1mM EDTA,1mM DTT, 50% (v/v) glycerol and stored at -20°C.

Example 3

Identification of PAM preferences for *Streptococcus pyogenes* and *Streptococcus thermophilus* CRISPR3 Cas9 proteins.

To empirically examine the PAM preferences for *Streptococcus pyogenes* (Spy) and *Streptococcus thermophilus* CRISPR3 (Sth3) Cas9 proteins, the randomized PAM library described in Example 1 was subject to digestion with purified Sth3 and Spy Cas9 proteins and guide RNA containing a variable targeting domain that hybridizes with, i.e., is complementary to, a sequence in the target DNA molecule (referred herein as target sequence), T1 (SEQ ID NO: 1). Sth3 and Spy Cas9-crRNA-tracrRNA complexes were assembled by mixing Cas9 protein with pre-

annealed crRNA and tracrRNA duplex (Table 3) at 1:1 molar ratio followed by incubation in a complex assembly buffer (10 mM Tris-HCl pH 7.5 at 37°C, 100 mM NaCl, 1mM EDTA, 1 mM DTT) at 37°C for 1 h. 1 µg of plasmid DNA library with randomized 5 bp NNNNN PAM was cleaved with 50 nM and 100 nM of Cas9 complex in a reaction buffer (10 mM Tris-HCl pH 7.5 at 37°C, 100 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT) for 60 min. at 37°C in a 100 µl reaction volume (Figure 3).

Table 3. RNA molecules used for Sth3 and Spy Cas9-crRNA-tracrRNA complex assembly.

Name	Sequence (5'-3')	Origin	SEQ ID NO.
Sth3 crRNA	CGCUAAAGAGGAAGAGGACAGUUUUAGAGCUGUGUUGUUUCG	Synthetic oligonucleotide	16
Sth3 tracrRNA	GGGCGAAACAACACAGCGAGUUAAAAUAAGGCUUAGUCCGUACUCAACUUGAAAAGGUGGCACCGAUUCGGUGUUUUU	*In vitro* transcription	17
Spy crRNA	CGCUAAAGAGGAAGAGGACAGUUUUAGAGCUAUGCUGUUUUG	Synthetic oligonucleotide	18
Spy tracrRNA	GGGAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUU	*In vitro* transcription	19

To efficiently capture the blunt-ends of the plasmid library generated by Sth3 or Spy cleavage, a 3' dA was added by incubating the completed digestion reactions with 2.5 U of DreamTaq DNA Polymerase (Thermo Fisher Scientific) and 0.5 µl of 10 mM dATP (or dNTP) for an additional 30 min. at 72°C (Figure 3). Reaction products were purified using GeneJET PCR Purification Kit (Thermo Fisher Scientific). Next adapters with a 3' dT overhang were generated by annealing TK-117 (CGGCATTCCTGCTGAACCGCTCTTCCGATCT (SEQ ID NO: 20) and

phosphorylated TK-111 (GATCGGAAGAGCGGTTCAGCAGGAATGCCG (SEQ ID NO: 21) oligonucleotides. 100 ng of the resulting adapter was ligated to an equal concentration of the purified 3' dA overhanging cleavage products for 1 hour at 22°C in a 25 µl reaction volume in ligation buffer (40 mM Tris-HCl pH 7.8 at 25°C, 10 mM $MgCl_2$, 10 mM DTT, 0.5 mM ATP, 5% (w/v) PEG 4000, 0.5 U T4 Ligase; Thermo Fisher Scientific) (Figure 3). Next, to selectively enrich for cleaved products containing the PAM sequence, PCR amplification was performed with a forward primer, pUC-dir (SEQ ID NO: 5), specific to the PAM-side of the cleaved pTZ57R/T plasmid vector and with a reverse primer, TK-117 (SEQ ID NO: 20), specific to the ligated TK-117/TK-111 adapter sequence (Figure 3). PCR fragments were generated by Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific) amplification (15 cycles of a 2-step amplification protocol) with 10 µl of ligation reaction mixtures as a template (in 100 µl total volume). The resulting 131 bp PCR products amplified from the Cas9 pre-cleaved plasmid libraries were purified with GeneJET PCR Purification Kit (Thermo Fisher Scientific) and prepared for Illumina deep sequencing as described in Example 1 except the barcode containing forward primers used in the primary reaction were specific to the TK-117/TK-111 adapter sequence and are shown in Table 4 (Figure 3).

Table 4. Primary PCR primer sequences for tailing on the sequences needed for Illumina deep sequencing of cleaved and adapter ligated 5 bp randomized PAM pTZ57R/T library

Primer Name	Digestion Experiment	Primer Orientation	Primary PCR Primer Sequence	SEQ ID NO.
JKYS807.1	50 nM Sth3	Forward	CTACACTCTTTCCCTACACGAC GCTCTTCCGATCTAAGGCGGCA TTCCTGCTGAAC	22
JKYS807.2	100 nM Sth3	Forward	CTACACTCTTTCCCTACACGAC GCTCTTCCGATCTTTCCCGGCA	23

			TTCCTGCTGAAC	
JKYS807.3	50 nM Spy	Forward	CTACACTCTTTCCCTACACGAC GCTCTTCCGATCTGGAACGGCA TTCCTGCTGAAC	24
JKYS807.4	100 nM Spy	Forward	CTACACTCTTTCCCTACACGAC GCTCTTCCGATCTCCTTCGGCA TTCCTGCTGAAC	25

The resulting Illumina compatible libraries were then sequenced as described in Example 1. The PAM sequence for only those reads containing a perfect 12 nt sequence match flanking either side of the 5 nucleotide randomized PAM sequence were captured and used to examine the frequency and diversity of PAM sequences present in the Sth3 and Spy Cas9–guide RNA cleaved libraries. Given the inherent bias in the uncut library observed in Figure 4 and described in Example 1, PAM preferences were calculated relative to the uncut library by dividing the frequency of a given PAM from the Sth3 or Spy Cas9-guide RNA digested library by the frequency of the same PAM sequence in the uncut library with the resulting value being represented as a fold enrichment correlative to the uncut control. To examine the PAM preferences of Sth3 and Spy Cas9 proteins, the percent nucleotide composition of the PAM sequences with ≥2 fold enrichment relative to the uncut control were examined. As shown in Figure 5 and Figure 6, the canonical PAM preferences for both Sth3 and Spy Cas9 proteins, NGGNG and NGG, respectively, are observed in both the 50 nM and 100 nM digests. For Sth3 Cas9 protein, a slight preference (not previously reported) for a C or T bp at position 1 is also evident. Next, the effect of decreasing Sth3 and Spy Cas9-crRNA-tracrRNA complex concentration and digestion time on PAM preferences was examined. To this end, the minimal Cas9 concentration and shortest time where PCR amplified cleavage products may still be obtained from the randomized PAM plasmid library were determined. First, the reaction time was held constant at 60 minutes while the Cas9-crRNA-tracrRNA complex concentration was varied between 0.5-100 nM. Next, the Cas9-crRNA-tracrRNA complex concentration was fixed at 50 nM and the

reaction time was varied between 1-60 minutes. Optimization of the cleavage reaction conditions revealed that the concentration and cleavage time for Sth3 and Spy Cas9 complexes could be reduced to 0.5 nM (at a 60 min. incubation time) or 1 min. (at a 50 nM concentration of Cas9 complex), respectively (Figure 7).

To examine the PAM sequences present in the minimally digested Sth3 and Spy Cas9-guide RNA libraries, 0.5 nM-60 minute and 50 nM-1 minute PCR amplified cleavage products were purified with the GeneJET PCR Purification Kit (Thermo Fisher Scientific) and subjected to Illumina deep sequencing as described above for the 50 nM and 100 nM-60 minute Sth3 and Spy digests. As a positive control and to demonstrate the reproducibility of PAM preferences derived from our assay, the 50 nM-60 minute digests for Sth3 and Spy were repeated and Illumina deep sequenced again. PAM preference analysis was carried-out as described above for the Sth3 and Spy (50 nM and 100 nM-60 minute digests) examining the percent nucleotide composition of the PAM sequences with ≥2 fold enrichment relative to the uncut library. As shown in Figure 8 and Figure 9, the positive controls (Sth3 and Spy 50 nM-60 minute digests) demonstrated very similar trends in PAM preferences compared to that observed previously indicating a high degree of assay reproducibility. The PAM preferences observed in the minimally Sth3 and Spy digested libraries compared to that exhibited by the respective 50 nM-60 minute positive control are shown in Figure 10 and Figure 11. When the concentration of Sth3 Cas9-crRNA-tracrRNA complex is lowered to 0.5 nM, the percentage of uncanonical PAM residues cleaved by Sth3 decreases; resulting in a tightening of specificity (Figure 10). This is most evident at positions 2 and 3 where on-nucleotide preferences for a G increase and off-nucleotide preferences decrease. A similar shift in PAM preference towards the reported PAM sequence for Spy (NGG) is observed when the Spy Cas9-crRNA-tracrRNA complex is lowered to 0.5 nM. Here the percentage of PAMs with an uncanonical A residue at position 2 declines from over 20% in the 50 nM-60 minute and 50 nM-1 minute digests to almost zero in the 0.5 nM-60 minute digest (Figure 11).

Next, the effect of using a chimeric fusion of crRNA and tracrRNA (single guide RNA (sgRNA)) (Jinek *et al*. (2012) *Science* 337:816-21 and Gasiunas *et al*. (2012) *Proc. Natl acad. Sci. USA* 109: E2579-E2586) on Sth3 and Spy Cas9 PAM

preferences was assayed. Digestion, enrichment, Illumina deep sequencing and PAM preference analysis was carried-out as described above against the randomized 5 bp PAM plasmid DNA library except a sgRNA (Table 5) was used in place of the crRNA-tracrRNA duplex and digests were only performed with 0.5 nM of sgRNA-Cas9 complex for 60 min.

Table 5. RNA molecules used for Cas9-sgRNA complex assembly.

Name	Sequence (5'-3')	Origin	SEQ ID NO.
Sth3 sgRNA	GGGCGCUAAAGAGGAAGAGGACAGUUUUAGAGCUGUGUUGUUUCGGUUAAAACAACACAGCGAGUUAAAAUAAGGCUUAGUCCGUACUCAACUUGAAAAGGUGGCACCGAUUCGGUGUUUUUU	*In vitro* transcription	26
Spy sgRNA	GGGCGCUAAAGAGGAAGAGGACAGUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUU	*In vitro* transcription	27

As shown in Figure 12 and Figure 13, the PAM preferences for Sth3 and Spy Cas9 proteins (NGGNG and NGG respectively) are nearly identical regardless of the type of guide RNA used; either a crRNA-tracrRNA duplex or sgRNA.

Example 4

Identification of PAM preferences for *Streptococcus thermophilus* CRISPR1 Cas9 protein.

To empirically examine the PAM preferences for *Streptococcus thermophilus* CRISPR1 (Sth1) Cas9 protein with a reported PAM sequence of 7 nucleotides, NNAGAAW (Horvath *et al*. (2008) *Journal of Bacteriology* 190:1401–12), a randomized 7 bp PAM plasmid DNA library was generated as described for the 5 bp randomized PAM library in Example 1 with the following modifications.

Randomization of the PAM sequence was generated through the synthesis of four oligonucleotides, GG-940-G (GTGCACGCCGGCGACGTTGGGTCAACTNNGNNNNTGTCCTCTTCCTCTTTAG CGTTTAG (SEQ ID NO: 28), GG-940-C (GTGCACGCCGGCGACGTTGGGTCAACTNNCNNNNTGTCCTCTTCCTCTTTAG CGTTTAG (SEQ ID NO: 29), GG-940-A (GTGCACGCCGGCGACGTTGGGTCAACTNNANNNNTGTCCTCTTCCTCTTTAG CGTTTAG (SEQ ID NO: 30) and GG-940-T (GTGCACGCCGGCGACGTTGGGTCAACTNNTNNNNTGTCCTCTTCCTCTTTAG CGTTTAG (SEQ ID NO: 31), with hand-mixing used to create a random incorporation of nucleotides across the random residues (represented as N). The randomized single stranded oligonucleotides were each separately converted into double-stranded DNA templates for cloning into the plasmid vector using a second oligonucleotide, GG-939 (GACTAGACCTGCAGGGGATCCCGTCGACAAATTCTAAACGCTAAAGAGGAAG AGGAC (SEQ ID NO: 126), with complementation to the 3' end of GG-940-G, GG-940-C, GG-940-A and GG-940-T and by PCR extension with DreamTaq polymerase (Thermo Fisher Scientific) (oligoduplexes I & II Figure 1). To avoid cleavage of some species of the randomized positions, the resulting double-stranded templates were each digested with an 8 bp cutting restriction endonuclease, SdaI, so that overhangs were present at each end; a PstI compatible overhang and a Taq added single 3' A overhang. The resulting overhangs were used to directionally ligate the 4 double-stranded templates into pTZ57R/T (Thermo Fisher Scientific) pre-cleaved with PstI. The ligations were Ca^{2+} transformed into DH5α *E. coli* cells, plasmid DNA was recovered and combined from each of the 4 transformants derived from GG-940-G, GG-940-C, GG-940-A and GG-940-T to generate the randomized 7 bp NNNNNNN PAM plasmid DNA library.

PAM preference experiments with Sth1 Cas9 protein on the resulting 7 bp randomized PAM plasmid DNA library were carried-out similarly to that described in Example 3 for the *Streptococcus thermophilus* CRISPR3 (Sth3) and *Streptococcus pyogenes* (Spy) Cas9 proteins (against the 5 bp randomized PAM library). Briefly, Sth1 Cas9-crRNA-tracrRNA complexes were assembled by mixing Cas9 protein

with pre-annealed crRNA and tracrRNA duplex (Table 6) at 1:1 molar ratio followed by incubation in a complex assembly buffer (10 mM Tris-HCl pH 7.5 at 37°C, 100 mM NaCl, 1 mM EDTA, 1 mM DTT) at 37°C for 1 h. Digests were performed using 1 µg of randomized 7 bp PAM library with 50 nM Sth1 crRNA-tracrRNA-Cas9 complexes at 37°C for 60 min., 50 nM Sth1 crRNA-tracrRNA-Cas9 complexes at 37°C for 1 min. and 0.5 nM Sth1 crRNA-tracrRNA-Cas9 complexes at 37°C for 60 min. (Figure 3). As a positive control, 1 µg of the randomized 7 bp PAM library was also digested with Sth3 and Spy Cas9-sgRNA complexes (0.5 nM at 37°C for 60 min.). A 3' dA was added to the blunt-ends of the cleaved fragments (Figure 3). Next, duplexed adapter TK-117/TK-111 with a 3' dT overhang was ligated to the A overhang (Figure 3). Then, PCR was assembled using primers pUC-dir (SEQ ID NO: 5) and TK-117 (SEQ ID NO: 20) to enrich for PAM sequences that supported cleavage (Figure 3). 40 ng of the resulting PCR product was then amplified with Phusion® High Fidelity PCR Master Mix (New England Biolabs, M0531L) adding on the sequences necessary for amplicon-specific barcodes and Illumina sequencing using "tailed" primers through two rounds of PCR each consisting of 10 cycles (Figure 3). The sequences of the barcode specific forward primers used in the primary PCR reaction were similar to those listed in Table 3 and the reverse primer, JKYS812 (CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTCGGCGACGTTGGGTC (SEQ ID NO: 32)), was paired with each of the forward primers. A set of primers, AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG (Universal Forward, SEQ ID NO: 8) and CAAGCAGAAGACGGCATA (Universal Reverse, SEQ ID NO: 9), universal to all primary PCR reactions were utilized for the secondary PCR amplification.

Table 6. RNA molecules used for Sth1 Cas9-crRNA-tracrRNA complex assembly.

Name	Sequence (5'-3')	Origin	SEQ ID NO.
Sth1 crRNA	CGCUAAAGAGGAAGAGGACAGU UUUUGUACUCUCAAGAUUUA	Synthetic oligonucleotide	33

Sth1 tracrRNA	GGGUAAAUCUUGCAGAAGCUACA AAGAUAAGGCUUCAUGCCGAAAU CAACACCCUGUCAUUUUAUGGCA GGGUGUUUUCG	*In vitro* transcription	34

The resulting PCR amplifications were prepared and Illumina deep sequenced as described in Example 1 and PAM preference analysis was carried-out as described in Example 3 for the Sth3 and Spy (50 nM and 100 nM-60 minute digests) examining the percent nucleotide composition of the PAM sequences with ≥2 fold enrichment relative to the uncut library. As shown in Figure 14 and Figure 15, the PAM preferences for the positive controls, Sth3 and Spy Cas9 proteins, are nearly identical regardless of the length of the randomized PAM plasmid DNA library used; either 5 bp or 7 bp. The PAM preferences observed for the Sth1 Cas9 protein are shown in Figure 16 and match those previously reported (NNAGAAW). Just as observed for Sth3 and Spy Cas9 proteins, the PAM specificity of Sth1 is more relaxed at higher concentrations of guide RNA-Cas9 complex. This is most evident at position 5 where an off-preference for a C nucleotide is less prevalent at the lower 0.5 nM complex concentration.

Since the canonical PAM sequence preferences for Spy, Sth1 and Sth3 Cas9 proteins may be recapitulated with our assay regardless of the type of guide RNA used (either crRNA-tracrRNA or sgRNA) or length of the randomized PAM sequence, suggests that the *in vitro* PAM library assay described herein or derivations of it may be used to directly interrogate PAM specificity from any Cas9 assuming the guide RNA sequences, either crRNA-tracrRNA or sgRNA, may be successfully deduced. Additionally, our assay grants precise control over the amount of Cas9 protein used in the *in vitro* digestion assays described herein allowing a detailed examination of Cas9 PAM specificity as a function of Cas9-guide RNA complex concentration as evident by the apparent broadening in PAM specificity as Cas9-guide RNA complex concentration was increased.

Example 5

Identification of *Brevibacillus laterosporus* crRNA, tracrRNA and Cas9 endonuclease

To empirically examine the PAM preferences for a Cas9 protein whose PAM was undefined, a *cas9* gene from an uncharacterized Type II CRISPR-Cas system was identified by searching internal Pioneer-DuPont databases consisting of microbial genomes with the amino acid sequence of *S. thermophilus* CRISPR3 (Sth3) Cas9 (SEQ ID NO: 35). Amino acid alignment of Sth3 revealed 12.9% identity and 24.4% similarity at the protein level with a protein derived from a single long open-reading-frame of 3279 nucleotides (SEQ ID NO: 36) from the *Brevibacillus laterosporus* bacterial strain SSP360D4. Translation of the open-reading-frame encodes a protein of 1092 amino acids (not including the stop codon). Based on PFAM database searches the protein contained HNH endonuclease and CRISPR-associated domains all hallmarks of a Cas9 protein. The *cas9* gene of SSP360D4 was also located upstream of a CRISPR array comprised of 7 repeat-spacer units (Figure 17A). The repeat and spacer length (36 and 30 bp, accordingly) is similar to other Type II CRISPR-Cas systems. However, 5 of 8 repeats contain 1 or 2 bp mutations (Figure 17B). Sequences of Repeat1 (SEQ ID NO: 37), Repeat4 (SEQ ID NO: 40) and Repeat5 (SEQ ID NO: 41) are conserved; therefore this sequence was selected as a template for designing single guide RNAs (sgRNAs). A region upstream of the *cas9* gene is partially complementary (anti-repeat) to the 5'-terminus of the repeat suggesting a putative tracrRNA (Figure 17A). The possible transcriptional directions of the putative tracrRNA were considered by examining the secondary structures and possible termination signals present in a RNA version of the sense and anti-sense genomic DNA sequences surrounding the anti-repeat. However, the transcriptional direction of the tracrRNA and CRISPR region could not be reliably determined bioinformatically, so a method described in Example 7 was designed to empirically determine the appropriate directions of transcription. . Other genes typically found in a Type II CRISPR-Cas locus were either truncated, as was the case for cas1, or missing (Figure 17A).

Example 6

Protein expression and purification of *Brevibacillus laterosporus* Cas9 protein.

To examine the PAM specificity and guide RNA of *Brevibacillus laterosporus* (Blat) Cas9 protein with the *in vitro* cleavage assays described in Examples 7& 8, Blat Cas9 protein was *E. coli* expressed and purified. Briefly, a DNA fragment encoding the *Brevibacillus laterosporus* Cas9 protein was PCR amplified directly from the Pioneer-DuPont strain, SSP360D4, using Blat-Cas9-dir (TACCATGGCATACACAATGGGAATAGATG (SEQ ID NO: 45) and Blat-Cas9-rev (TTCTCGAGACGACTAGTTGATTTAATCGAATTGAC (SEQ ID NO: 46) primer pair and cloned into a pBAD24-CHis expression vector pre-cleaved over NcoI and XhoI sites. To establish optimal expression conditions three different *E. coli* strains, BL21 (DE3), DH10B and Rosetta (DE3), were analyzed. Highest expression yield of soluble Blat Cas9 protein was obtained in the BL21 (DE3) strain.

For purification, Blat Cas9 protein was expressed in *E. coli* BL21 (DE3) strain grown in LB broth supplemented with ampicillin (100 mg/ml). Cells were grown at 37°C to an OD 600 of 0.5 at which time the growth temperature was decreased to 16°C and expression induced with 0.2% (w/v) arabinose for 20 h. Cells were pelleted and resuspended in loading buffer (20 mM KH_2PO_4 pH7.0, 0.5 M NaCl, 10 mM imidazole, 5% glycerol) and disrupted by sonication. Cell debris was removed by centrifugation. The supernatant was loaded onto the Ni^{2+}-charged 5ml HiTrap chelating HP column (GE Healthcare) and eluted with a linear gradient of increasing imidazole concentration. The fractions containing Cas9 were pooled and subsequently loaded onto heparin column for elution using a linear gradient of increasing NaCl concentration (from 0.5 to 1 M NaCl). The fractions containing Cas9 were pooled and dialyzed against 10mMBis-Tris-HCl pH 7.0, 300 mM KCl, 1mM EDTA, 1 mM DTT, 50% (v/v) glycerol and stored at -20°C.

Example 7

Determination of guide RNAs for the Cas9 of *Brevibacillus laterosporus.*

To determine a guide RNA for the Cas9 protein identified in the *Brevibacillus laterosporus* (Blat) Type II CRISPR-Cas system, we designed two single guide RNA (sgRNA) variants to account for both possible expression scenarios of the tracrRNA and CRISPR array (Figure 18 & Table 7) and used them to probe which expression

scenario supported cleavage activity of Blat Cas9 in the 7 bp randomized PAM plasmid DNA library from Example 4.

sgRNAs were designed by first identifying the boundaries of the putative tracrRNA molecules by analyzing regions which were partially complementary to the 22 nt 5' terminus of the repeat (anti-repeat). Next, to determine the 3' end of the tracrRNA, possible secondary structures and terminators were used to predict the region of termination in the downstream fragment (Figures 19 and 20) using Mini-fold (Markham *et al*. (2008) *Methods in Molecular Biology* 453: 3-31). The sgRNAs contained a T7 polymerase transcription initiation recognition signal at the 5' end followed by a 20 nt target recognition sequence, 16 nt of crRNA repeat, 4 nt self-folding hairpin loop and anti-repeat sequence complementary to the repeat region of the crRNA followed by the remaining 3' part of the putative tracrRNA (Table 7). The sgRNA variant which contains a putative tracrRNA transcribed in the same direction as the *cas9* gene is termed "direct" sgRNA, while the sgRNA containing the tracrRNA transcribed in the opposite direction a "reverse" sgRNA (Figure 18).

The "direct" sgRNA encoding gene was obtained in two PCR steps. First two fragments were generated by PCR using GG-969/GG-839 and TK-149/TK-150 oligonucleotide primer pairs (Table 8). The fragments were purified with the GeneJET PCR Purification Kit (Thermo Fisher Scientific) and the full length sgRNA gene was assembled from these fragments by overlapping PCR using GG-969/TK-150 primer pairs. The "reverse" sgRNA encoding gene was amplified by PCR using GG-840/GG-841 oligonucleotide primer pairs (Table 8). To generate the sgRNA encoding plasmids pUC-Blat-dir-sgRNA and pUC-Blat-rev-sgRNA, the PCR fragments were cloned into pUC18 vector digested with SacI.

Table 7. “Direct” and “reverse” Blat sgRNAs used to deduce transcriptional direction of crRNA and tracrRNA loci.

Blat sgRNA	T7 Transcription Initiation	Variable Targeting domain (SEQ ID NO:)	16 nt of the repeat	Loop	Anti-Repeat	Remaining Putative 3’ tracrRNA Sequence	SEQ ID NO:
Direct	GGG	193	195	GAAA	197	199	47
Reverse	GGG	194	196	GAAA	198	200	48

Table 8. Oligonucleotides used for Blat sgRNA gene construction and sgRNA production.

Name	Sequence (5’-3’)	SEQ ID NO.
GG-969	GGGCGCTAAAGAGGAAGAGGACAGCTATAGTTCCTTACT GAAAGGTAAGTTGCTATAGTAAGGGCAAC	49
GG-839	CTAAAAACGGGCTAGGCGATCCCCAACGCCTCGGGTCTG TTGCCCTTACTATAGCAACTTAC	50
-149	GATCGCCTAGCCCGTTTTTACGGGCTCTCCCCATATTCAA AATAATGACAGACGA	
TK-150	AAAAAAAAGCACCTCGGAAATAAATGCTCCAAGGTGCTCG TCTGTCATTATTTTGAATATGG	
GG-840	GGGCGCTAAAGAGGAAGAGGACAATCATATCATATCGAG GAAACTTGATATGATATGATACTTTCATTTTA	53
GG-841	CATAAAATAGACAGATAAATGAGATTGACTTCGATGATATA TGGATATAAAATGAAAGTATCATATCATATCAAG	54
TK-124	TAATACGACTCACTATAGGGCGCTAAAGAGGAAGAGG	55
TK-151	AAAAAAAAGCACCTCGGAAATAAATG	56
TK-126	ATAAAATAGACAGATAAATGAGATTGACTTCG	57

“Direct” and “reverse” Blat sgRNAs were obtained by *in vitro* transcription using TranscriptAid T7 High Yield Transcription Kit (Thermo Fisher Scientific) from the

PCR *fragments* containing a T7 promoter at the proximal end of the RNA coding sequence. The "direct" sgRNA encoding fragment (177 nt) was generated using the TK-124/TK-151primer pair (Table 8) with pUC-Blat-dir-sgRNA plasmid DNA as template, whereas the "reverse" sgRNA encoding fragment (118 nt) was generated using the TK-124/TK-126 primer pair with pUC-Blat-rev-sgRNA plasmid as template (Table 7). The resulting sgRNAs were purified using GeneJET RNA Cleanup and Concentration Micro Kit (Thermo Fisher Scientific) and used for complex assembly. Blat Cas9-sgRNA complexes were assembled by mixing Cas9 protein with sgRNA at 1:1 molar ratio followed by incubation in a complex assembly buffer (10 mM Tris-HCl pH 7.5, 100 mM NaCl, 1mM EDTA, 1 mM DTT) at 37°C for 1 h. Blat Cas9 cleavage of the 7 bp randomized PAM plasmid DNA library was performed similarly as described above for Spy and Sth3 Cas9 proteins (Example 3). Briefly, 50 nM of Blat Cas9 complexes, assembled using "direct" or "reverse" sgRNAs, respectively, were incubated with 1 µg plasmid DNA for 60 min at 37°C. After library digestion and addition of 3‘ dA overhangs, adapters were ligated and cleavage products were PCR amplified (Figure 3). Analysis of reaction products by agarose gel electrophoresis revealed that the "direct" sgRNA, but not the "reverse" sgRNA supported plasmid library cleavage (Figure 21). Single guide RNAs targeting a target site in the genome of an organism can be designed by changing the targeting sequence of SEQ ID NO: 47 with any random nucleotide that can hybridize to any desired target sequence (guide RNA as shown in SEQ ID NO: 127).

Example 8:

Identification of PAM preferences for *Brevibacillus laterosporus* Cas9 protein

After determining a guide RNA for *Brevibacillus laterosporus* (Blat) Cas9, PAM identification was performed similarly to that described in Example 3 for the Spy and Sth3 Cas9 proteins. Briefly, 1 µg of 7 bp randomized PAM plasmid library was digested with various concentrations of Blat Cas9-"direct" sgRNA complex, ranging between 0.5-50 nM, and at various reaction times, ranging from 1 to 60 minutes. Next, 3‘ dA overhangs were added to the cleavage products, adapters ligated and adapter-ligated cleavage products were PCR amplified. PCR reactions were then electrophoresed on a 1% agarose gel and visualized. As shown in Figure 22 and similarly to that described for Sth3 and Spy Cas9 proteins, the minimal

concentration and cleavage time needed to support visualization after PCR amplification were 0.5 nM (at a 60 min. incubation time) or 1 min. (at a 50 nM concentration of Cas9 complex). Next, the amplifications for the 50 nM-60 min., 50 nM-1 min. and 0.5 nM-60 min. digests were purified with the GeneJET PCR Purification Kit (Thermo Fisher Scientific) and Illumina sequencing anchors added by two rounds of PCR as described in Example 3 for the Sth3 and Spy Cas9 proteins when examining their PAM preferences with the 7 bp randomized PAM library. The resulting Illumina compatible libraries were then sequenced as described in Example 1 and PAM preference analysis was carried-out as described in Example 3 for the Sth3 and Spy (50 nM and 100 nM-60 minute digests) examining the percent nucleotide composition of the PAM sequences with ≥2 fold enrichment relative to the uncut library. When the composition of the PAM sequences with ≥2 fold enrichment for the 50 nM-60 minute, 50 nM-1 minute and 0.5nM-60 minute digests were analyzed, the consensus PAM sequence for the Blat Cas9 protein was NNNNCND (N=G, C, A or T; D=A, G or T) with a strong preference for a C at position 5 of the PAM sequence (Figure 23). A moderate preference for an A was observed at position 7 and slight preferences for a C or T at position 4 and G, C or A over T at position 6 was also noted. Similarly to Sth1, Sth3 and Spy Cas9 proteins, the PAM specificity broadens as the Cas9-sgRNA complex concentration increases. This is most evident at position 5 where a larger proportion of PAM sequences containing an A residue support cleavage at 50 nM compared with 0.5 nM Cas9-sgRNA complexes.

To confirm the cleavage positions for the Blat Cas9 protein, we engineered the pUC18-T1-GTCCCGT-PAM plasmid containing a 20 base pair region matching the spacer T1 (SEQ ID NO: 1) followed by a PAM sequence, GTCCCGT, falling within the PAM consensus for Blat. To generate the plasmid, first the synthetic oligoduplex containing T1 and GTCCCGT PAM sequences was assembled by annealing complementary oligonucleotides GG-935 (CAAATTCTAAACGCTAAAGAGGAAGAGGACAGTCCCG (SEQ ID NO: 58) and GG-936 (AATTCGGGACTGTCCTCTTCCTCTTTAGCGTTTAGAATTTGAGCT (SEQ ID NO: 59) and ligated into pUC18 vector pre-cleaved with ScaI and EcoRI. 2.5 µg of the resulting plasmid was then digested with 100 nM of the Blat Cas9–

sgRNA complex in the 500 µl of reaction buffer at 37°C for 60 min., purified using GeneJET PCR Purification Kit (Thermo Fisher Scientific) and electrophoresed on an agarose gel. Linear digestion products were then purified from the agarose gel using the GeneJET Gel Extraction Kit (Thermo Fisher Scientific). The cleaved region in Blat Cas9 linearized pUC18-T1-GTCCCGT-PAM plasmid was then directly sequenced with the pUC-EheD (CCGCATCAGGCGCCATTCGCC (SEQ ID NO: 60) and pUC-LguR (GCGAGGAAGCGGAAGAGCGCCC (SEQ ID NO: 61) primers. The sequence results confirmed that plasmid DNA cleavage occurred in the protospacer 3 bp away from the PAM sequence (Figure 24) similar to that observed for Sth3 and Spy Cas9 proteins.

The NNNNCND PAM sequence identified herein, can be introduced adjacent to any polynucleotide of interest, thereby creating a target site that can be recognized by a guide RNA/Cas9 endonuclease complex described herein, wherein the guide RNA/Cas9 endonuclease system is capable of recognizing, binding to, and optionally nicking or cleaving all or part of the target sequence adjacent to the NNNNCND PAM sequence.

Example 9

Characterization of Cas9 endonucleases and their PAM preferences, and cognate guide RNAs from diverse organisms.

The rapid *in vitro* methods described herein (Examples 1-8) can be used to identify and characterize Cas endonucleases from any organism and their related PAM preferences and guide RNAs elements.

Cas9 proteins of Type II-A, II-B and II-C subtypes were identified from the NCBI NR database using the PSI-BLAST program (Altschul SF, *et al.* (1997) *Nucleic Acids Res.* 25:3389-3402). A phylogenetic relationship of each Cas9 protein was visualized with CLANs software (Frickey T, Lupas A. (2004) *Bioinformatics* 20:3702-3704) and putative Cas9 endonucleases from different groupings were selected. Genomic DNA regions derived from non-pathogenic sources and those containing a clustered-regularly-interspace-short-palindromic repeat (CRISPR) array and a putative trans-activating CRISPR RNA (tracrRNA) coding region (defined by homology to the CRISPR repeat and termed the anti-

repeat) in the vicinity of the Cas9 were chosen. In total, 11 diverse genomic DNA regions were selected for further analysis (Table 9)

A schematic of the genomic locus for each system is depicted in Figures 25-35. The *cas9* gene open-reading-frame (ORF), CRISPR array, anti-repeat (the genomic DNA region demonstrating partial homology to the repeat consensus that indicates the location of the encoded tracrRNA) and other CRISPR-Cas genes are indicated for each system. The genomic DNA sequence and length of each *cas9* gene ORF and *cas9* gene translation (not including the stop codon) are referenced in Table 10 for each system. Table 10 lists the consensus sequence of the CRISPR array repeats from the genomic DNA locus of each system and the sequences of the anti-repeat for each system (as genomic DNA sequence on the same strand as the *cas9* gene ORF).

As was done for the *Brevibacillus laterosporus* (BLAT) Type II CRISPR/Cas system (described in Example 6), the possible transcriptional directions of the putative tracrRNAs for each new system were considered by examining the secondary structures and possible termination signals present in a RNA version of the sense and anti-sense genomic DNA sequences surrounding the anti-repeat. Based on the hairpin-like secondary structures present for each system, the transcriptional direction of the tracrRNA was deduced for 10 of the 11 diverse Type II CRISPR-Cas systems. Because the anti-repeat in the tracrRNA can hybridize to the crRNA derived from the CRISPR array to form a duplexed RNA capable of guiding the Cas9 endonuclease to cleave invading DNA the transcriptional direction of the CRISPR array may also be determined based of the direction of tracrRNA transcription (since double-stranded RNA hybridizes with 5' to 3' directionality). The deduced transcriptional directions of both the tracrRNA and CRISPR array for each system are listed in Table 10 and are depicted in Figures 25-35. Based on the likely transcriptional direction of the tracrRNA and CRISPR array, single guide RNAs (sgRNAs) were also designed and are shown in Table 12. For the system, *Sulfurospirillum* sp. SCADC, where the transcriptional direction of the tracrRNA and CRISPR array could not be deduced two sgRNAs were designed (as described in Example 7 for the Blat Type II CRISPR-Cas system); one for each possible direction of tracrRNA transcription (Table 12).

Next the sgRNAs, will be complexed with the respective purified Cas9 protein and assayed for their ability to support cleavage of the 7 bp randomized PAM plasmid DNA library (as described in Example 7 for the Blat Type II CRISPR-Cas system). If the sgRNA does not support cleavage activity, new guide RNA designs (either sgRNA or duplexed crRNA and tracrRNA; in both possible transcriptional directions of the CRISPR array and anti-repeat region) will be tested for their ability to support cleavage.

Once a guide RNA that supports Cas9 cleavage has been established, the PAM specificity of each Cas9 endonuclease can be assayed (as described in Example 7 for the Blat Type II CRISPR-Cas system). After PAM preferences have been determined, the sgRNAs may be further refined for maximal activity or cellular transcription by either increasing or decreasing the tracrRNA 3' end tail length, increasing or decreasing crRNA repeat and tracrRNA anti-repeat length, modifying the 4 nt self-folding loop or altering the sequence composition.

Table 9. List of 11 organisms selected for the identification of diverse Type II CRISPR-Cas systems described herein.

Bacterial Origin	Abbreviation	CRISPR-Cas System Subtype	Isolated from
Lactobacillus reuteri Mlc3	Lreu	II-A	Sourdough
Lactobacillus rossiae DSM 15814	Lros	II-A	Sourdough
Pediococcus pentosaceus SL4	Ppen	II-A	Meat
Lactobacillus nodensis JCM 14932	Lnod	II-A	Dairy
Sulfurospirillum sp. SCADC	Sspe	II-B	Oil sands tailings pond
Bifidobacterium thermophilum DSM 20210	Bthe	II-C	Dairy
Loktanella vestfoldensis	Lves	II-C	Lakes Ace and Pendant, Vestfold Hills, Antarctica

Sphingomonas sanxanigenens NX02	Ssan	II-C	Isolated from soil
Epilithonimonas tenax DSM 16811	Eten	II-C	River epilthon
Sporocytophaga myxococcoides	Smyx	II-C	From soil, cellulose decomposing organism
Psychroflexus torquis ATCC 700755	Ptor	II-C	Prydz Bay, Antarctica

Table 10. Sequence and length of the *cas9* gene ORF and *cas9* gene translation from each Type II CRISPR-Cas system identified by the methods described herein.

Bacterial Origin	*cas9* Gene ORF (SEQ ID NO)	Length of *cas9* Gene ORF (bp)	Translation of *cas9* Gene ORF (not including the stop codon) (SEQ ID NO)	Length of *cas9* Gene Translation (No. of Amino Acids)
Lreu	70	4107	81	1368
Lros	71	4110	82	1369
Ppen	72	4041	83	1346
Lnod	73	3393	84	1130
Sspe	74	4086	85	1361
Bthe	75	3444	86	1147
Lves	76	3216	87	1071
Ssan	77	3318	88	1105
Eten	78	4200	89	1399
Smyx	79	4362	90	1453
Ptor	80	4530	91	1509

Table 11. CRISPR repeat consensus, anti-repeat (putative tracrRNA coding region) and deduced transcriptional directions of tracrRNA and CRISPR array relative to the *cas9* gene ORF for 11 diverse Type II CRISPR-Cas systems.

Bacterial Origin	CRISPR Repeat Consensus (SEQ ID NO)	Anti-Repeat (SEQ ID NO)	tracrRNA Transcriptional Direction (Relative to the *cas9* Gene ORF)	CRISPR Array Transcriptional Direction (Relative to the *cas9* Gene ORF)
Lreu	92	103	Antisense	Sense
Lros	93	104	Antisense	Sense
Ppen	94	105	Antisense	Sense
Lnod	95	106	Sense	Sense
Sspe	96	107	Sense/Antisense	Sense/Antisense
Bthe	97	108	Sense	Antisense
Lves	98	109	Antisense	Antisense
Ssan	99	110	Antisense	Antisense
Eten	100	111	Antisense	Antisense
Smyx	101	112	Antisense	Sense
Ptor	102	113	Antisense	Antisense

Table 12. Examples of sgRNAs components for each new diverse Type II CRISPR-Cas system described herein.

Bacterial Origin	T7 Transcription Initiation	Variable Targeting domain (VT)	crRNA Repeat	Loop	tracrRNA Anti-Repeat	Remaining Putative 3' tracrRNA Sequence	SEQ ID NO:
Lreu	GGG	N_{20} (*)	149	N_4 (**)	161	173	128
Lros	GGG	N_{20} (*)	150	N_4 (**)	162	174	129
Ppen	GGG	N_{20} (*)	151	N_4 (**)	163	175	130
Lnod	GGG	N_{20} (*)	152	N_4 (**)	164	176	131
Sspe	GGG	N_{20} (*)	153	N_4 (**)	165	177	132

(tracrRNA Sense-crRNA Sense)							
Sspe (tracrRNA Antisense-crRNA Antisense)	GGG	N_{20} (*)	154	N_4 (**)	166	178	133
Bthe	GGG	N_{20} (*)	155	N_4 (**)	167	179	134
Lves	GGG	N_{20} (*)	156	N_4 (**)	168	180	135
Ssan	GGG	N_{20} (*)	157	N_4 (**)	169	181	136
Eten	GGG	N_{20} (*)	158	N_4 (**)	170	182	137
Smyx	GGG	N_{20} (*)	159	N_4 (**)	171	183	138
Ptor	GGG	N_{20} (*)	160	N_4 (**)	172	184	139

N_{20} (*) indicates a series of 20 nucleotides as one example of a sgRNA variable targeting domain. As described herein, the variable targeting domain of a sgRNA can vary for example, but not limiting from at least 12 to 30 nucleotides. N_4 (**) indicates a loop of 4 nucleotides such as but not limiting to GAAA. As described herein, the length of the loop can vary from at least 3 nucleotides to 100 nucleotides.

Single guide RNAs targeting a target site in the genome of an organism can be designed by changing the targeting sequence of any one of SEQ ID NOs: 114-125 with any random nucleotide that can hybridize to any desired target sequence (such as, but not limiting to, guide RNAs as shown in SEQ ID NO: 128-139).

Example 10

PAM specificity is not greatly influenced by the type or composition of the guide RNA.

As described in Example 3 and 4, to empirically examine the PAM preferences for *Streptococcus pyogenes* (Spy), *Streptococcus thermophilus* CRISPR3 (Sth3) and *Streptococcus thermophilus* CRISPR1 Cas9 proteins, two randomized PAM libraries (described in Example 1 and 4) were generated. The two libraries increased in size and complexity from 5 randomized base pairs (1,024 potential PAM combinations) to 7 randomized base pairs (16,384 potential PAM combinations). These randomized libraries were subject to digestion with purified Sth3 and Spy Cas9 proteins (5N library, Example 3) and Sth1 (Example 4, 7N library) and guide RNA containing a variable targeting domain T1 that hybridizes with, i.e., is complementary to, a sequence in the target DNA molecule (referred herein as target sequence), T1 (SEQ ID NO: 1).

To confirm that PAM specificity is independent of the type of guide RNA, duplexed crRNA: tracrRNA or single guide RNA (sgRNA), Spy, Sth3 and Sth1 Cas9 PAM preferences were examined using Cas9 sgRNA RNP complexes instead of Cas9 and crRNA:tracrRNA RNP complexes. Digestion was carried-out at a single RNP complex concentration of 0.5 nM and PAM preference analysis was performed as described herein. PAM preferences were nearly identical regardless of the type of guide RNA used; either a crRNA:tracrRNA duplex or sgRNA (US patent application 62/196535, filed July 24. 2015, which is incorporated herein in its entirety by reference) .

To confirm that PAM specificity is not greatly influenced by the composition of the target DNA or spacer sequence, the sequence on the opposite side of the 5 or 7 bp randomized library was targeted for cleavage with a different variable targeting domain, T2-5 for the 5 bp library or T2-7 for the 7 bp library. Spy and Sth3 Cas9 proteins preloaded with sgRNAs targeting the T2 sequence were used to interrogate the 5 bp randomized PAM library while the Sth1 Cas9-T2 sgRNA complexes were used to digest the 7 bp randomized PAM library. The library was digested with Spy, Sth3 and Sth1 Cas9 proteins preloaded with sgRNAs targeting the T2 sequence and PAM preferences were assayed as described above. The PAM preferences for

all 3 Cas9 proteins were nearly identical regardless of spacer and target DNA sequence (US patent application 62/196535, filed July 24. 2015).

Example 11

Identification of extended PAM sequences.

As shown in Figure 23 (Example 8) , the PAM consensus for the Blat Cas9 protein under the 0.5 nM digest conditions was NNNNCND (N=G,C, A or T; D=A, G or T) with a strong preference for a C at position 5 of the PAM sequence. A moderate preference for an A was observed at position 7 and slight preferences for a C or T at position 4 and G, C or A over T at position 6 were also noted when closely examining Figure 23. Similarly to Spy, Sth3 and Sth1 Cas9 proteins, the PAM specificity broadens as the Cas9-sgRNA complex concentration increases. This was most evident at position 5 where a larger proportion of PAM sequences containing an A residue support cleavage at 50 nM compared with 0.5 nM digest conditions (Figure 23).

Since Blat Cas9 may accept any base in the first 3 positions of its PAM sequence (Figure 23), the spacer domain T1 (and corresponding variable targeting domain in the guide RNA) was shifted by 3 nucleotides to allow PAM identification to be extended from 7 to 10 bp. The shifted T1 variable targeting domain, T1-3, was incorporated into the Blat “direct” sgRNA resulting in a sgRNA referred to as Blat sgRNA (T1-3) and PAM identification was performed as described previously for Spy, Sth3, Sth1 and Blat Cas9 proteins. PAM preference analysis revealed the PAM specificity for Blat Cas9 can be extended out to position 8 where there is a moderate preference for an additional A (US patent application 62/196535, filed July 24, 2015).

To validate the PAM specificity for Blat Cas9, plasmids were engineered to contain mutations (GTCCCGAA (reference), GTC**A**CGAA , GTCC**T**GAA , GTCCCG**C**A , GTCCCGA**C** , GTCCCG**CC** with mutations shown in bold and underlined , US patent application 62/196535, filed July 24, 2015) in the most conserved residues of the PAM immediately downstream of a 20 base pair region matching the variable targeting domain T1. *In vitro* cleavage reactions with the various PAM sequences were initiated by mixing supercoiled plasmid DNA with pre-assembled Blat Cas9-sgRNA complex (1:1 v/v ratio) at 15°C. The final reaction

mixture contained 3 nM plasmid, 50 nM Cas9, 10 mMTris-HCl (pH 7.5 at 37°C), 100 mM NaCl, 1 mM DTT and 10 mM MgCl2 in a 100 µl reaction volume. Aliquots were removed at timed intervals and quenched with phenol/chloroform. The aqueous phase was mixed with 3× loading dye solution (0.01% (w/v) bromophenol blue and 75 mM EDTA in 50% (v/v) glycerol) and reaction products analyzed by agarose gel electrophoresis. The amount of supercoiled (SC) form was evaluated by densitometric analysis of ethidium bromide stained gels using the software ImageJ. Values of reaction rate constants were obtained as described by Szczelkun *et al*, 2014, *Proc. Natl. Acad. Sci. U. S. A.* 111: 9798–803). Replacement of the C nucleotide at the 5th position abolished plasmid DNA cleavage confirming its key role in Blat Cas9 PAM recognition. Replacement of A nucleotides at the 7th and 8th positions significantly reduced (43x and 12x, respectively) the cleavage rate of supercoiled plasmid also indicating the importance of these nucleotides in Blat Cas9 PAM recognition.

To confirm the cleavage positions for the Blat Cas9 protein with an optimal PAM sequence, a plasmid was engineered that contained a 20 base pair region matching the variable targeting domain T1 followed by a PAM sequence, GTCCCGAA, falling within the PAM consensus for Blat Cas9, NNNNCNDD. We used direct sequencing to determine the ends of the linear DNA molecule generated by the Blat Cas9 RNP complex. The sequence results confirmed that plasmid DNA cleavage occurred in the protospacer 3 nucleotides away from the PAM sequence (similar to that observed for Spy, Sth3 and Sth1 Cas9 proteins (Garneau et al, 2010, Nature 468: 67–71; Gasiunas et al, 2012, Proc. Natl. Acad. Sci. U. S. A. 109: E2579–2586; Jinek et al, 2012, Science 337: 816–21).

Example 12

In planta genome editing using Blat cas9 and sgRNA.

Following elucidation of the sgRNA and PAM preferences for Blat Cas9, maize optimized Cas9 and sgRNA expression cassettes were generated for in planta testing. The Blat cas9 gene was maize codon optimized and intron 2 of the potato ST-LSI gene was inserted to disrupt expression in *E. coli* and facilitate optimal splicing (Libiakova et al, 2001. Plant Cell Rep. 20: 610–615). To facilitate nuclear localization of the Blat Cas9 protein in maize cells, Simian virus 40 (SV40)

monopartite and *Agrobacterium tumefaciens* bipartite VirD2 T-DNA border endonuclease nuclear localization signals were incorporated at the amino and carboxyl-termini of the Cas9 open reading frame, respectively (US patent application 62/196535, filed July 24, 2015). To express the resulting maize optimized Blat cas9 gene in a robust constitutive manner, it was operably linked to a maize Ubiquitin promoter, 5' UTR and intron (Christensen et al, 1992, Plant Mol. Biol. 18: 675–689) and pinII terminator (An et al, 1989, Plant Cell 1: 115–122) in a plasmid DNA vector. To confer efficient sgRNA expression in maize cells, a maize U6 polymerase III promoter region isolated from *Zea mays* cultivar B73 residing on chromosome 8 at position 165,535,024-165,536,023 (B73 RefGen_v3) and terminator (TTTTTTTT) were isolated and operably fused to the 5' and 3' ends of a modified Blat sgRNA encoding DNA sequence. The modified Blat sgRNA contained two modifications from the sgRNA that was used in the in vitro studies (see Blat sgRNA (T1) direct; SEQ ID NO: 151), a T to G alteration at position 101 and a T to C modification at 159. The changes were introduced to remove potential premature U6 polymerase III signals in the Blat sgRNA. Alterations were introduced to have minimal impact on the secondary structure of the sgRNA compared to the version used in the in vitro studies. For a direct comparison with the Blat Cas9 sgRNA system, equivalent Cas9 and sgRNA DNA expression vectors were also prepared for the Spy Cas9 sgRNA system.

To carefully compare the mutational efficiency resulting from the imperfect non-homologous end-joing (NHEJ) repair of DNA double-strand breaks (DSBs) resulting from Spy and Blat Cas9 cleavage, protospacer identical genomic target sites were selected by identifying targets with Spy and Blat Cas9 compatible PAMs, NGGYCVAA. Since Blat and Spy Cas9 both cleave between the 3 and 4 bp upstream of their respective PAM, genomic targets will be cleaved at the exact same position allowing a tighter correlation between NHEJ mutation frequency and cleavage activity. Identical variable targeting domain sequences were selected for Blat and Spy Cas9 by capturing the 18 to 21 nt sequence immediately upstream of the PAM. To ensure optimal U6 polymerase III expression and not introduce a mismatch within the sgRNA variable targeting domain (spacer), all target sequences were selected to naturally terminate in a G at their 5' end. Targets were selected in

exon 1 and 4 of the maize fertility gene Ms45 (referred to as MS45 Exon1 and MS45 Exon 4; see also U.S. Patent No. 5,478,369 incorporated herein by reference) and within the promoter region of the maize liguleless-1 gene (refered to as LIG34 Promoter target herein; Moreno et al. 1997. Genes and Development 11:616-628).

To rapidly examine the mutational activity of Blat Cas9 with the PAM and sgRNA identified herein, Blat and the equivalent Spy Cas9 and sgRNA DNA expression vectors were independently introduced into maize Hi-II (Armstrong & Green, 1985, Planta 164: 207–214) immature embryos (IEs) by particle gun transformation similar to that described in (Ananiev et al, 2009 , Chromosoma 118: 157–177). Since particle gun transformation can be highly variable, a visual marker DNA expression cassette, Ds-Red, was also co-delivered with the Cas9 and sgRNA expression vectors to aid in the selection of evenly transformed IEs. In total, 3 transformation replicates were performed on 60-90 IEs and 20-30 of the most evenly transformed IEs from each replicate were harvested 3 days after transformation. Total genomic DNA was extracted and the region surrounding the target site was PCR amplified and deep sequenced to a read depth in excess of 300,000. The resulting reads were examined for the presence of mutations at the expected site of cleavage by comparison to control experiments where only the Cas9 DNA expression cassette was transformed. Mutations arising at the expected site of cleavage for Blat Cas9 were detected with the most prevalent types of mutations being single base pair insertions or deletions. This pattern of imprecise mutagenic repair of the double-stranded DNA cut introduced by Blat Cas9 was also observed for the Spy Cas9 (US patent application 62/196535, filed July 24, 2015) and at other Cas9 sites in maize (data not shown). The mutational activity for Blat Cas9 was robust at 2 of the 3 sites tested and exceeded that of the Spy Cas9 at the Ms45 Exon 4 target site by ~30%.

In planta mutation detection

The DNA region surrounding the expected site of cleavage for each Cas9-guide RNA was amplified by PCR using Phusion® High Fidelity PCR Master Mix (NEB,USA) “tailing” on the sequences necessary for amplicon-specific barcodes and Illumina sequences through two rounds of PCR each consisting of 20 cycles. The primer pairs used in the primary PCR were primer pairs corresponding to the

Ms45 exon 1, Ms45 exon 4 and Lig34 promoter regions, respectively. A set of primers universal to the products from the primary reactions, were used in the secondary PCR reaction (US patent application 62/196535, filed July 24, 2015). The resulting PCR amplifications were purified with a Qiagen PCR purification spin column (Qiagen, Germany), concentration measured with a Hoechst dye-based fluorometric assay, combined in an equimolar ratio, and single read 100 nucleotide-length amplicon sequencing was performed on Illumina's MiSeq Personal Sequencer with a 5-10% (v/v) spike of PhiX control v3 (Illumina, FC-110-3001) to off-set sequence bias. Only those reads with a ≥1 nucleotide INDEL arising within the 10 nt window centered over the expected site of cleavage and not found in the negative controls were classified as mutations. Mutant reads with an identical mutation were counted and collapsed into a single read and the top 10 most prevalent mutations were visually confirmed as arising within the expected site of cleavage. The total numbers of visually confirmed mutations were then used to calculate the percentage of mutant reads based on the total number of reads of an appropriate length containing a perfect match to the barcode and forward primer.

Example 13

Simplified construction of randomized Protospacer-Adjacent-Motif (PAM) libraries for assaying Cas endonuclease PAM preferences.

To simplify construction for randomized PAM libraries, a fully double-stranded DNA oligoduplex as described in Example 1 (oligoduplex II) containing a region of randomization immediately adjacent to a DNA target sequence may be used directly as template for Cas endonuclease digestion. This would eliminate the cloning of the oligoduplex II fragment into a plasmid DNA vector allowing randomized PAM libraries to be constructed without the downstream *E. coli* transformation and plasmid DNA isolation steps. PAM sequences supporting Cas endonuclease cleavage in these linearized double-stranded DNA libraries would be captured and deep sequenced as described in Examples 3, 4 and 8 for Spy, Sth3, Sth1 and Blat Cas9 proteins. To identify those sequences that have truly been cleaved by a Cas endonuclease and not just the result of adaptor ligation to the end of an un-cleaved oligoduplex, an *in silico* enrichment step may be applied to the resulting deep sequencing reads by selecting for only those reads that contain an appropriate

sequence junction resulting for Cas endonuclease cleavage and adapter ligation. Once reads harboring a PAM sequence that supported cleavage have been identified, their nucleotide composition may be analyzed similar to that described for Spy, Sth3, Sth1 and Blat Cas9 proteins in Examples 3, 4 and 8.

Example 14

Cas endonuclease Proto-Spacer Adjacent Motifs (PAMs) may be assayed directly in *E. coli* cell lysate.

Cas endonuclease protein produced in *E. coli* may be directly (without subsequent purification steps) used to assay proto-spacer adjacent motif (PAM) recognition and single guide RNA (sgRNA) requirements upon cell lysis.

Streptococcus thermophilus CRISPR1 (Sth1) and *Streptococcus thermophilus* CRISPR3 (Sth3) Cas9 protein was produced in *E. coli* cells as described in Example 2 but without the purification steps. In brief, after cultures were grown, induced and allowed to express Cas9 protein, cell lysis was performed via sonication and cell debri was pelleted by centrifugation resulting in a cell lysate containing soluble Cas9 protein. Cas9-guide RNA complexes were assembled by combining 20 µl of resulting cell lysate with RiboLock RNase Inhibitor (40 U; Thermo Fisher Scientific) and 2 µg of T7 *in vitro* transcribed sgRNA (generated as described in Example 7) and incubated at room temperature for 15 min. To examine PAM preferences at different Cas9 concentrations, 1 µg of the 7 bp randomized PAM library (Example 4) was incubated with 10 µl of various dilutions (1-fold (undiluted), 10-fold and 100-fold) of cell lysate containing assembled Cas9 complexes in a 100 µl reaction buffer (10 mM Tris-HCl pH 7.5 at 37°C, 100 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT) so that *E. coli* lysate was diluted to a final concentration of either 10-fold, 100-fold or 1000-fold, respectively. Reactions mixtures were incubated for 60 min. at 37°, DNA end repaired with 2.5 U T4 DNA polymerase (Thermo Fisher Scientific), RNA digested with 1 µl RNase A/T1 Mix (Thermo Fisher Scientific) and 3' dA added with 2.5 U of DreamTaq DNA Polymerase (Thermo Fisher Scientific). Finally, DNA was recovered using a GeneJET PCR Purification Kit (Thermo Fisher Scientific). DNA fragments resulting from cleavage by Cas9 were tagged with adapters, captured and prepared for Illumina deep sequencing as described in Example 3 (Figure 3). The resulting libraries were deep sequenced as described in Example 1.

PAM sequences were identified from the resulting sequence data as described in Example 3 by only selecting those reads containing a perfect 12 nt sequence match flanking either side of the 7 nt PAM sequence capturing only those PAM sequences resulting from perfect Cas9-guide RNA target site recognition, cleavage and adapter ligation. The collection of resulting PAM sequences were then collapsed into like sequences, counted, and frequency of each PAM supporting cleavage calculated. To compensate for inherent bias in the initial randomized PAM libraries, the frequency of each PAM sequence was next normalized to its frequency in the starting library. Next, a PAM consensus was calculated using a position frequency matrix (PFM). This was accomplished by first aligning the collapsed PAM sequences. Then, each nucleotide (G, C, A, or T) at each position of the PAM was weighted based on the frequency of the PAM sequence with which it was associated. Finally, the total contribution of each nucleotide (G, C, A, or T) at each PAM position was summed to generate the overall probability of identifying a given nucleotide at each PAM position within the dataset.

Tables 13-18 represent the position frequency matrix (PFM) and resulting PAM consensus at each position of the 7 bp randomized PAM library for the *Streptococcus thermophilus* CRISPR1 (Sth1) and *Streptococcus thermophilus* CRISPR3 (Sth3) Cas9 proteins when assayed at different concentrations of *E.coli* cell lysate. The nucleotide positions of the 7 bp randomized PAM library are indicated by 1, 2, 3, 4, 5, 6, and 7 in a 5' to 3' direction with 1 being the closest to the DNA sequence involved in spacer target site recognition. The frequency of each nucleotide (G, C, A, T) at a respective position is indicated as a %. The consensus PAM preference is listed at the bottom of the table (consensus). The grayed highlighted cells indicate the nucleotide preference(s) at each position of the protospacer adjacent motif (PAM). The percentages in the position frequency matrix (PFM) tables represent the probability of finding the corresponding nucleotide at each position of the PAM sequence and can be used to infer the strength of PAM recognition at each position.

Table 13. Position frequency matrix (PFM) and PAM consensus for *Streptococcus thermophilus* CRISPR1 Cas9 with Cas9 protein provided via 10 fold dilution of *E. coli* cell lysate.

	1	2	3	4	5	6	7
G	17.69%	14.97%	22.16%	41.47%	9.34%	8.56%	21.79%
C	27.64%	29.63%	5.67%	17.96%	28.97%	10.45%	13.89%
A	26.54%	25.79%	70.38%	16.85%	55.56%	64.09%	26.22%
T	28.13%	29.61%	1.79%	23.72%	6.13%	16.90%	38.10%
Consensus	N	N	A	G	A	A	W

Table 14. Position frequency matrix (PFM) and PAM consensus for *Streptococcus thermophilus* CRISPR1 Cas9 with Cas9 protein provided via 100 fold dilution of *E. coli* cell lysate.

	1	2	3	4	5	6	7
G	19.80%	16.70%	27.37%	43.52%	11.01%	7.87%	20.20%
C	25.74%	27.47%	6.01%	16.02%	24.04%	8.77%	12.49%
A	29.40%	25.80%	64.19%	18.73%	59.60%	69.09%	27.66%
T	25.06%	30.03%	2.43%	21.73%	5.36%	14.27%	39.65%
Consensus	N	N	A	G	A	A	W

Table 15. Position frequency matrix (PFM) and PAM consensus for *Streptococcus thermophilus* CRISPR1 Cas9 with Cas9 protein provided via 1000 fold dilution of *E. coli* cell lysate.

	1	2	3	4	5	6	7
G	19.72%	16.25%	24.92%	53.70%	10.39%	3.79%	18.40%
C	26.89%	30.09%	4.08%	13.55%	22.65%	3.32%	10.18%
A	27.92%	26.35%	70.37%	15.20%	64.60%	86.15%	33.19%
T	25.46%	27.30%	0.64%	17.55%	2.37%	6.73%	38.23%
Consensus	N	N	A	G	A	A	W

Table 16. Position frequency matrix (PFM) and PAM consensus for *Streptococcus thermophilus* CRISPR3 Cas9 with Cas9 protein provided via 10 fold dilution of *E. coli* cell lysate.

	1	2	3	4	5	6	7
G	12.46%	49.67%	80.76%	21.03%	49.94%	23.46%	21.96%
C	26.60%	9.72%	5.67%	15.73%	10.22%	20.97%	24.97%
A	16.71%	22.42%	8.85%	35.35%	19.75%	27.10%	25.69%
T	44.23%	18.18%	4.72%	27.89%	20.10%	28.46%	27.39%
Consensus	N	G	G	N	G	N	N

Table 17. Position frequency matrix (PFM) and PAM consensus for *Streptococcus thermophilus* CRISPR3 Cas9 with Cas9 protein provided via 100 fold dilution of *E. coli* cell lysate.

	1	2	3	4	5	6	7
G	12.06%	55.16%	82.16%	23.38%	53.61%	23.02%	22.39%
C	28.81%	11.09%	5.10%	17.36%	10.19%	21.26%	24.06%
A	22.84%	17.33%	9.02%	31.55%	18.80%	25.87%	25.64%
T	36.28%	16.42%	3.72%	27.71%	17.40%	29.84%	27.91%
Consensus	N	G	G	N	G	N	N

Table 18. Position frequency matrix (PFM) and PAM consensus for *Streptococcus thermophilus* CRISPR3 Cas9 with Cas9 protein provided via 1000 fold dilution of *E. coli* cell lysate.

	1	2	3	4	5	6	7
G	12.26%	63.66%	89.19%	27.07%	54.77%	26.19%	23.09%
C	30.31%	7.86%	2.78%	17.23%	9.70%	19.39%	22.85%
A	21.26%	15.31%	6.18%	29.16%	17.45%	26.56%	26.21%
T	36.17%	13.17%	1.86%	26.55%	18.08%	27.87%	27.86%
Consensus	N	G	G	N	G	N	N

As shown in Tables 13-18, all lysate dilutions yielded the canonical PAM preferences for Sth1 and Sth3 Cas9 proteins, NNAGAAW and NGGNG, respectively. Similar to the results with purified protein in Examples 3, 4 and 8, higher concentrations of lysate and consequentially Cas9 protein resulted in a relaxation of PAM specificity. This was most notable for the Sth3 Cas9 protein at PAM position 2 where the preference for a G residue is reduced from approximately 64% in the PFM in the 1000-fold dilution (final concentration) reaction to around 50% in the 10-fold dilution (final concentration) experiment (Tables 16-18). For Sth1 Cas9 protein, PAM positions 4, 5 and 6 were most particularly affected by different concentrations of Cas9 protein in the lysate dilution experiments.

This data indicates that the *in vitro* PAM library assay described herein obtained the same results for the PAM preferences for Sth1 and Sth3 Cas9 proteins when compared to assays where the Sth1 and Sth3 Cas9 proteins are stably expressed (in-vivo expressed). Hence, the *in vitro* PAM library assay described herein, or derivations of it, may be used to assay PAM specificity from any Cas endonuclease using unpurified Cas protein coming directly from *E. coli* lysate. Additionally by diluting *E. coli* lysate containing Cas9 protein, the *in vitro* PAM library

assay permits the measurement of PAM specificity to be examined as a function of Cas endonuclease concentration as is evident by the apparent broadening in PAM specificity as *E. coli* lysate containing Cas9 protein was increased.

Example 15

Cas endonuclease Proto-Spacer Adjacent Motifs (PAMs) may be assayed directly with *in vitro* translated protein.

Cas endonuclease protein produced by *in vitro* translation may be used to directly (without subsequent purification steps) assay proto-spacer adjacent motifs (PAM) and single guide RNA (sgRNA) requirements.

The *Streptococcus pyogenes* (Spy) cas9 gene was codon optimized for expression in eukaryotes (maize) with standard methods known in the art and operably linked to the *in vitro* translation (IVT) vector pT7CFE1-NHIS-GST-CHA (Thermo Fisher Scientific). To eliminate expression of the HA tag, a stop codon was included between the Spy *cas9* gene and C-terminal tag. The resulting plasmid was purified by phenol:chloroform extraction to remove residual RNases and further purified by precipitation with 2 volumes of ethanol in the presence of sodium acetate. Next, Spy protein was produced *in vitro* using a 1-Step Human Coupled IVT Kit (Thermo Fisher Scientific) per the manufacturer's instruction allowing the reaction to proceed overnight at 30°C. Following the incubation, the reactions were centrifuged at 10,000 rpm for 5 min. 20 µl of supernatant containing soluble Cas9 protein was mixed with 2 µg of T7 *in vitro* transcribed sgRNA (generated as described in Example 7) and incubated for 15 min. at room temperature. To examine PAM preferences at different Cas9 concentrations, 1 µg of the 7 bp randomized PAM library (Example 4) was incubated with 10 µl of various dilutions (1-fold (undiluted), 10-fold and 100-fold) of *in vitro* translation mixtures containing assembled Cas9 complexes in a 100 µl reaction buffer (10 mM Tris-HCl pH 7.5 at 37°C, 100 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT) so that IVT supernatant was diluted to a final concentration of either 10-fold, 100-fold or 1000-fold. Reactions mixtures were incubated for 60 min at 37°, DNA end repaired with 2.5 U T4 DNA polymerase (Thermo Fisher Scientific), RNA digested with 1 µl RNase A/T1 Mix (Thermo Fisher Scientific) and 3' dA added with 2.5 U of DreamTaq DNA Polymerase (Thermo Fisher Scientific). Finally, DNA was recovered using a GeneJET PCR Purification

Kit (Thermo Fisher Scientific). PAM sequences supporting cleavage were captured by adapter ligation and enriched for as described in Example 3 (Figure 3). The resulting libraries were deep sequenced as described in Example 1. PAM sequences were identified from the resulting sequence data as described in Example 3 by only selecting those reads containing a perfect 12 nt sequence match flanking either side of the 5 or 7 nt PAM sequence capturing only those PAM sequences resulting from perfect Cas9-guide RNA target site recognition, cleavage and adapter ligation. To compensate for inherent bias in the initial randomized PAM library, the frequency of each PAM sequence was normalized to its frequency in the starting library and a PAM consensus was then calculated with a position frequency matrix (PFM) as described in Example 14.

Tables 19-21 represent the position frequency matrix (PFM) and resulting PAM consensus at each position of the 7 bp randomized PAM library for the *Streptococcus pyogenes* Cas9 protein when assayed at different concentrations of in vitro translated (IVT) supernatant. The nucleotide positions of the 7 bp randomized PAM library are indicated by 1, 2, 3, 4, 5, 6, and 7 in a 5' to 3' direction with 1 being the closest to the DNA sequence involved in spacer target site recognition. The frequency of each nucleotide (G, C, A, T) at a respective position is indicated as a %. The consensus PAM preference is listed at the bottom of the table (consensus). The grayed highlighted cells indicate the nucleotide preference(s) at each position of the protospacer adjacent motif (PAM). The percentages in the position frequency matrix (PFM) tables represent the probability of finding the corresponding nucleotide at each position of the PAM sequence and can be used to infer the strength of PAM recognition at each position.

Table 19. Position frequency matrix (PFM) and PAM consensus for *Streptococcus pyogenes* Cas9 with Cas9 protein provided via 10 fold dilution of in-vitro translated solution (IVT).

	1	2	3	4	5	6	7
G	24.18%	53.04%	72.63%	19.30%	14.19%	19.97%	23.65%
C	25.97%	7.16%	8.52%	24.26%	25.67%	28.52%	27.44%
A	25.21%	28.71%	14.69%	22.57%	23.80%	19.66%	20.39%
T	24.64%	11.09%	4.16%	33.87%	36.34%	31.85%	28.52%
Consensus	N	G	G	N	N	N	N

Table 20. Position frequency matrix (PFM) and PAM consensus for *Streptococcus pyogenes* Cas9 with Cas9 protein provided via 100 fold dilution of in-vitro translated solution (IVT).

	1	2	3	4	5	6	7
G	23.84%	52.07%	78.60%	21.17%	14.72%	19.66%	22.39%
C	24.16%	6.26%	4.34%	21.69%	23.72%	28.60%	27.09%
A	26.64%	34.55%	14.85%	25.33%	25.90%	20.48%	21.29%
T	25.36%	7.12%	2.21%	31.82%	35.66%	31.26%	29.23%
Consensus	N	G	G	N	N	N	N

Table 21. Position frequency matrix (PFM) and PAM consensus for *Streptococcus pyogenes* Cas9 with Cas9 protein provided via 1000 fold dilution of in-vitro translated solution (IVT).

	1	2	3	4	5	6	7
G	23.39%	81.14%	95.35%	27.51%	15.79%	19.98%	22.92%
C	22.34%	2.54%	0.80%	14.69%	23.08%	26.85%	25.30%
A	29.08%	12.52%	3.07%	26.65%	25.51%	22.87%	22.57%
T	25.19%	3.80%	0.78%	31.15%	35.63%	30.29%	29.22%
Consensus	N	G	G	N	N	N	N

As illustrated in Tables 19-21, the PAM requirement preferences reported for the Spy Cas9 protein (NGG) may be recapitulated under all IVT dilutions. Similar to the results with purified protein in Examples 3, 4 and 8, higher concentrations of IVT supernatant and consequentially Cas9 protein resulted in a broadening of PAM specificity. This was most notable for Spy Cas9 at PAM position 2 where the frequency for an uncanonical A residue increases from approximately 13% in the PFM with the 1000-fold dilution (final concentration) reaction to around 29% in the 10-fold dilution (final concentration) experiment.

This data indicates that the *in vitro* translation (IVT) assay described herein obtained the same results for the PAM preferences for Spy Cas9 protein when compared to assays where the Spy Cas9 protein is stably expressed (*in-vivo* expressed). Hence, the *in vitro* translation (IVT) assay described herein, or derivations of it, may be used to assay PAM specificity from any Cas endonuclease. Additionally by diluting IVT products containing Cas9 protein, our assay permits the measurement of PAM specificity to be examined as a function of Cas endonuclease concentration as evident by the apparent broadening in PAM specificity as IVT supernatant containing Cas9 protein was increased.

Example 16

Guide RNA and PAM requirements for novel Cas endonucleases.

The single guide RNA (sgRNA) and PAM requirements of the Cas9 endonucleases from *Lactobacillus reuteri* Mlc3 (Lreu), *Lactobacillus nodensis* JCM 14932 (Lnod), *Sulfurospirillum* sp. SCADC (Sspe), *Bifidobacterium thermophilum* DSM 20210 (Bthe), *Loktanella vestfoldensis* (Lves), *Epilithonimonas tenax* DSM 16811 (Eten) and *Sporocytophaga myxococcoides* (Smyx) (Example 9) were determined with the methods described herein.

If purified protein could not be easily obtained as described in Example 2, Cas9 protein from E. coli cell lysate as described in Example 14 or *in vitro* translated (IVT) Cas9 protein as described in Example 15 was utilized. Once a source of Cas9 protein was established, 1 µg of the 7 bp randomized PAM plasmid DNA library (Example 4) was subject to Cas9-guide RNA digestion at various concentrations of either purified protein, lysate, or IVT protein. DNA fragments resulting from cleavage by Cas9 were ligated to adapters, captured and prepared for Illumina deep sequencing as described in Example 3 (Figure 3). The resulting libraries were deep sequenced as described in Example 1. Since the position of cleavage within target sites for novel Cas9 proteins is unknown, reads were 1st examined for the most predominant cleavage location by examining the junction resulting from cleavage and adapter ligation. After properly defining the position of cleavage, PAM sequences were identified from the resulting sequence data as described in Example 3 by only selecting those reads containing a perfect 12 nt sequence match flanking either side of the 5 or 7 nt PAM sequence. To compensate for inherent bias in the initial randomized PAM library, the frequency of each PAM sequence was normalized to its frequency in the starting library and a PAM consensus was then calculated with a position frequency matrix (PFM) as described in Example 14. To obtain the most accurate read-out on PAM specificity and avoid conditions that are conducive to promiscuous PAM recognition (Examples 3, 4, 8, 14 and 15), the lowest concentration of Cas9 (purified, E. coli lysate or IVT supernatant) that supported cleavage was used to ascertain the PAM recognition of each Cas9 protein.

Tables 22-28 represent the position frequency matrix (PFM) and resulting PAM consensus at each position of the 7 bp randomized PAM library for several previously uncharacterized Cas9 proteins. Results derived from the lowest concentration of Cas9 coming from either purified, *E. coli* lysate or in vitro translation (IVT) supernatant that supported cleavage are shown. The nucleotide positions of the 7 bp randomized PAM library are indicated by 1, 2, 3, 4, 5, 6, and 7 in a 5' to 3' direction with 1 being the closest to the DNA sequence involved in spacer target site recognition. The frequency of each nucleotide (G, C, A, T) at a respective position is indicated as a %. The consensus PAM preference is listed at the bottom of the table (consensus). The grayed highlighted cells indicate the nucleotide preference(s) at each position of the protospacer adjacent motif (PAM). The percentages in the position frequency matrix (PFM) tables represent the probability of finding the corresponding nucleotide at each position of the PAM sequence and can be used to infer the strength of PAM recognition at each position.

Table 22. Table 22. Position frequency matrix (PFM) and PAM consensus for *Lactobacillus reuteri* Cas9 when purified Cas9 protein was used (0.5 nM Cas9-guide RNA complex and 60 minute digestion time).

	1	2	3	4	5	6	7
G	15.57%	83.27%	98.90%	31.64%	39.04%	25.51%	15.86%
C	15.96%	2.44%	0.12%	17.94%	24.13%	26.77%	34.32%
A	17.74%	11.81%	0.66%	14.84%	11.30%	22.13%	18.37%
T	50.73%	2.48%	0.32%	35.58%	25.53%	25.58%	31.44%
Consensus	N (T>V)	G	G	N	N (G>H)	N	N

Table 23. Position frequency matrix (PFM) and PAM consensus for *Lactobacillus nodensis* Cas9 when purified Cas9 protein was used (50 nM Cas9-guide RNA complex and 60 minute digestion time).

	1	2	3	4	5	6	7
G	21.47%	13.95%	2.62%	7.92%	4.07%	5.67%	24.14%
C	25.74%	23.76%	2.07%	1.53%	1.68%	1.29%	16.67%
A	22.41%	19.73%	94.31%	89.34%	93.77%	91.48%	33.13%
T	30.38%	42.56%	0.99%	1.22%	0.48%	1.55%	26.07%
Consensus	N	N (T>V)	A	A	A	A	N

Table 24. Position frequency matrix (PFM) and PAM consensus for *Sulfurospirillum sp.* SCADC Cas9 with Cas9 protein provided via 1000 fold dilution of in vitro translated solution (IVT).

	1	2	3	4	5	6	7
G	16.26%	97.32%	97.67%	18.52%	22.18%	18.86%	23.20%
C	24.43%	0.95%	0.85%	20.37%	20.90%	25.19%	22.14%
A	35.19%	1.11%	0.74%	31.97%	22.61%	26.12%	23.94%
T	24.13%	0.61%	0.74%	29.13%	34.31%	29.82%	30.72%
Consensus	N	G	G	N	N	N	N

Table 25. Position frequency matrix (PFM) and PAM consensus for *Bifidobacterium thermophilum* Cas9 when purified Cas9 protein was used (0.5 nM Cas9-guide RNA complex and 60 minute digestion time).

	1	2	3	4	5	6	7
G	18.93%	16.16%	20.28%	0.10%	0.03%	2.53%	3.19%
C	34.69%	31.11%	27.80%	99.55%	99.05%	5.34%	47.56%
A	23.13%	28.52%	28.76%	0.13%	0.40%	91.44%	1.17%
T	23.24%	24.20%	23.17%	0.21%	0.52%	0.69%	48.08%
Consensus	N	N	N	C	C	A	Y

Table 26. Position frequency matrix (PFM) and PAM consensus for *Loktanella vestfoldensis* Cas9 with Cas9 protein provided via 1000 fold dilution of *E. coli* cell lysate.

	1	2	3	4	5	6	7
G	21.74%	62.30%	51.21%	13.71%	17.79%	32.03%	23.70%
C	29.99%	8.00%	5.94%	10.17%	5.73%	14.72%	23.82%
A	16.37%	21.66%	37.33%	63.65%	64.49%	13.01%	25.97%
T	31.89%	8.03%	5.51%	12.47%	11.99%	40.24%	26.51%
Consensus	N	G	R (G>A)	A	A	K	N

Table 27. Position frequency matrix (PFM) and PAM consensus for *Epilithonimonas tenax* Cas9 with Cas9 protein provided via 10 fold dilution of *E. coli* cell lysate.

	1	2	3	4	5	6	7
G	30.87%	25.83%	39.60%	18.03%	14.19%	87.26%	91.40%
C	30.34%	7.27%	3.14%	7.13%	11.68%	2.31%	2.27%
A	15.84%	63.47%	54.64%	71.53%	31.83%	3.18%	2.88%
T	22.95%	3.43%	2.61%	3.30%	42.29%	7.25%	3.45%
Consensus	N (S>W)	A	R	A	N (W>S)	G	G

Table 28. Position frequency matrix (PFM) and PAM consensus for *Sporocytophaga myxococcoides* Cas9 when purified Cas9 protein was used (50 nM Cas9-guide RNA complex and 60 minute digestion time).

	1	2	3	4	5	6	7
G	10.48%	19.15%	2.54%	4.72%	1.02%	7.00%	23.48%
C	26.01%	14.45%	0.56%	23.74%	0.80%	3.23%	17.02%
A	19.94%	59.05%	96.61%	7.74%	97.97%	79.56%	28.21%
T	43.58%	7.35%	0.29%	63.80%	0.21%	10.21%	31.28%
Consensus	N (T>V)	A	A	T	A	A	N

Table 29. Summary of sgRNA and PAM requirement for novel Cas endonucleases.

Bacterial Origin	Abbreviation	PAM consensus	sgRNA SEQ ID NO:
Lactobacillus reuteri Mlc3	Lreu	Table 22	114
Lactobacillus nodensis JCM 14932	Lnod	Table 23	117
Sulfurospirillum sp. SCADC	Sspe	Table 24	119
Bifidobacterium thermophilum DSM 20210	Bthe	Table 25	120
Loktanella vestfoldensis	Lves	Table 26	121
Epilithonimonas tenax DSM 16811	Eten	Table 27	123
Sporocytophaga myxococcoides	Smyx	Table 28	124

Among the Cas9 proteins examined, both the length and composition of PAM recognition was diverse. Two of the Cas9 proteins, Lreu and Sspe (Tables 22-23), exhibited PAM recognition similar to the *Streptococcus pyogenes* (Spy) Cas9 protein which predominantly recognizes a NGG PAM while others exhibited very C-rich (Bthe, Table 25) or A-rich (Lnod and Smyx; Tables 23 and 28) PAM recognition. Additionally, a couple of the Cas9 proteins, Eten and Lves (Tables 26 and 27), yielded characteristics of both G-rich and A-rich PAM recognition.

Unlike the diversity observed for PAM recognition, the position of target site cleavage did not differ greatly and was determined to be between the 3rd and 4th bp

upstream (5 prime) of the PAM for all Cas9 proteins except for one, the Cas9 protein from *Sulfurospirillum* sp. SCADC. Interestingly, the predominant cleavage location by examining the junction resulting from cleavage and adapter ligation was around the 7th bp upstream (5 prime) of the PAM sequence.

Taken together, these data further suggest that the methods described herein can be used to characterize novel Cas endonuclease PAM and guide RNA requirements.

Example 17

In planta genome editing with novel Cas9 endonucleases.

After determining the proto-spacer adjacent motif (PAM) and guide RNA requirement as described herein, Cas9 proteins with novel PAM recognition were selected and tested for their ability to cleave and mutagenize maize chromosomal DNA as described in Example 12.

To expand the number and diversity of sites available for genome editing, Cas9 proteins with diverse PAM recognition were selected for evaluation in corn by preferentially choosing systems with either A, T or C-rich PAM recognition to best complement the G-rich PAM of the *Streptococcus pyogenes* (Spy) Cas9 protein. Once systems were selected, DNA target sites adjacent to the appropriate PAM sequence were chosen and maize optimized *cas9* gene and single guide RNA (sgRNA) expression vectors were constructed and delivered into maize immature embryos as described in Example 12. Embryos were harvested two days after transformation and chromosomal DNA was analyzed for the presence of mutations resulting from DNA target site cleavage and repair as described in Example 12. The frequency of mutations identified at each target site for each Cas9 is listed in Table 30.

Interestingly, the *Bifidobacterium thermophilum* (Bthe) Cas9 protein failed to effectively mutagenize its target sites. However when different spacer lengths were tested for Bthe, the frequency of mutagenesis improved dramatically with a spacer length around 25 nt being the most optimal (Figure 36). Since the minimal spacer length for the *Streptococcus pyogenes* (Spy) Cas9 sgRNA is approximately 17 nt in length, it seems that the sgRNA spacer DNA target interactions for Bthe Cas9 may provide enhanced specificity relative to the Spy Cas9 protein.

Table 30. Maize chromosomal target DNA mutation frequencies two days after transformation by particle gun.

Origin of cas9 gene	DNA Target Location	sgRNA Spacer Length	Mutation Frequency
Bifidobacterium thermophilum DSM 20210	Chr1: 51.81 cM	25	0.29%
	Chr9:119.15 cM	25	0.05%
Lactobacillus nodensis JCM 14932	Chr1: 51.81 cM	21	0.06%
	Chr9:119.15 cM	22	0.28%

Taken together, these results indicate that the methods described herein to characterize Cas endonuclease PAM recognition and guide RNA requirements are robust. Ultimately, allowing new Cas endonuclease systems to be characterized for genome editing applications.

THAT WHICH IS CLAIMED:

1. A single guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said single guide RNA is selected from the group consisting of SEQ ID NOs: 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138 and 139.

2. A single guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said single guide RNA comprises a chimeric non-naturally occurring crRNA linked to a tracrRNA, wherein said tracrRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183 and 184.

3. A single guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said single guide RNA comprises a chimeric non-naturally occurring crRNA linked to a tracrRNA, wherein said chimeric non-naturally occurring crRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159 and 160.

4. A guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence, wherein said tracrRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183 and 184, wherein said chimeric non-naturally occurring crRNA

comprises a variable targeting domain capable of hybridizing to said target sequence.

5. A guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said chimeric non-naturally occurring crRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159 and 160, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

6. A guide RNA capable of forming a guide RNA/Cas9 endonuclease complex, wherein said guide RNA/Cas9 endonuclease complex can recognize, bind to, and optionally nick or cleave a target sequence, wherein said guide RNA is a duplex molecule comprising a chimeric non-naturally occurring crRNA and a tracrRNA, wherein said tracrRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183 and 184, wherein said chimeric non-naturally occurring crRNA comprises a nucleotide sequence selected from the group consisting of SEQ ID NOs: 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159 and 160, wherein said chimeric non-naturally occurring crRNA comprises a variable targeting domain capable of hybridizing to said target sequence.

7. A guide RNA/Cas9 endonuclease complex comprising a Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81, 82, 83, 84, 85, 86, 87, 88, 89, 90 and 91, or a functional fragment thereof, and at least one guide RNA, wherein said guide RNA/Cas9 endonuclease complex is capable of recognizing, binding to, and optionally nicking or cleaving all or part of a target sequence.

8. A guide RNA/Cas9 endonuclease complex comprising at least one guide RNA and a Cas9 endonuclease, wherein said Cas9 endonuclease is encoded by

a DNA sequence selected from the group consisting of SEQ ID NOs: 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, and 80, wherein said guide RNA/Cas9 endonuclease complex is capable of recognizing, binding to, and optionally nicking or cleaving all or part of a target sequence.

9. The guide RNA/Cas9 endonuclease complex of claim 7, wherein said guide RNA is selected from the group consisting of SEQ ID NOs: 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138 and 139.

10. The guide RNA/Cas9 endonuclease complex of claims 7, wherein said target sequence is located in the genome of a cell.

11. A method for modifying a target site in the genome of a cell, the method comprising providing to said cell at least one Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81, 82, 83, 84, 85, 86, 87, 88, 89, 90 and 91, or a functional fragment thereof, and at least one guide RNA, wherein said guide RNA and Cas9 endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site.

12. The method of claim 10, further comprising identifying at least one cell that has a modification at said target, wherein the modification at said target site is selected from the group consisting of (i) a replacement of at least one nucleotide, (ii) a deletion of at least one nucleotide, (iii) an insertion of at least one nucleotide, and (iv) any combination of (i) – (iii).

13. A method for editing a nucleotide sequence in the genome of a cell, the method comprising providing to said cell at least one Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81, 82, 83, 84, 85, 86, 87, 88, 89, 90 and 91, or a functional fragment thereof, a polynucleotide modification template, and at least one guide RNA, wherein said polynucleotide modification template comprises at least one nucleotide modification of said nucleotide

sequence, wherein said guide RNA and Cas9 endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site.

14. A method for modifying a target site in the genome of a cell, the method comprising providing to said cell at least one guide RNA, at least one donor DNA, and at least one Cas9 endonuclease selected from the group consisting of SEQ ID NOs: 81, 82, 83, 84, 85, 86, 87, 88, 89, 90 and 91, or a functional fragment thereof, wherein said at least one guide RNA and at least one Cas9 endonuclease can form a complex that is capable of recognizing, binding to, and optionally nicking or cleaving all or part of said target site, wherein said donor DNA comprises a polynucleotide of interest.

15. The method of claims 11, 13 or 14 , wherein said guide RNA is selected from the group consisting of SEQ ID NOs: 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138 and 139.

16. The method of claim 13, further comprising identifying at least one cell that said polynucleotide of interest integrated in or near said target site.

17. The method of any one of claims 10-14, wherein the cell is selected from the group consisting of a human, non-human, animal, bacterial, fungal, insect, yeast, non-conventional yeast, and plant cell.

Figure 1

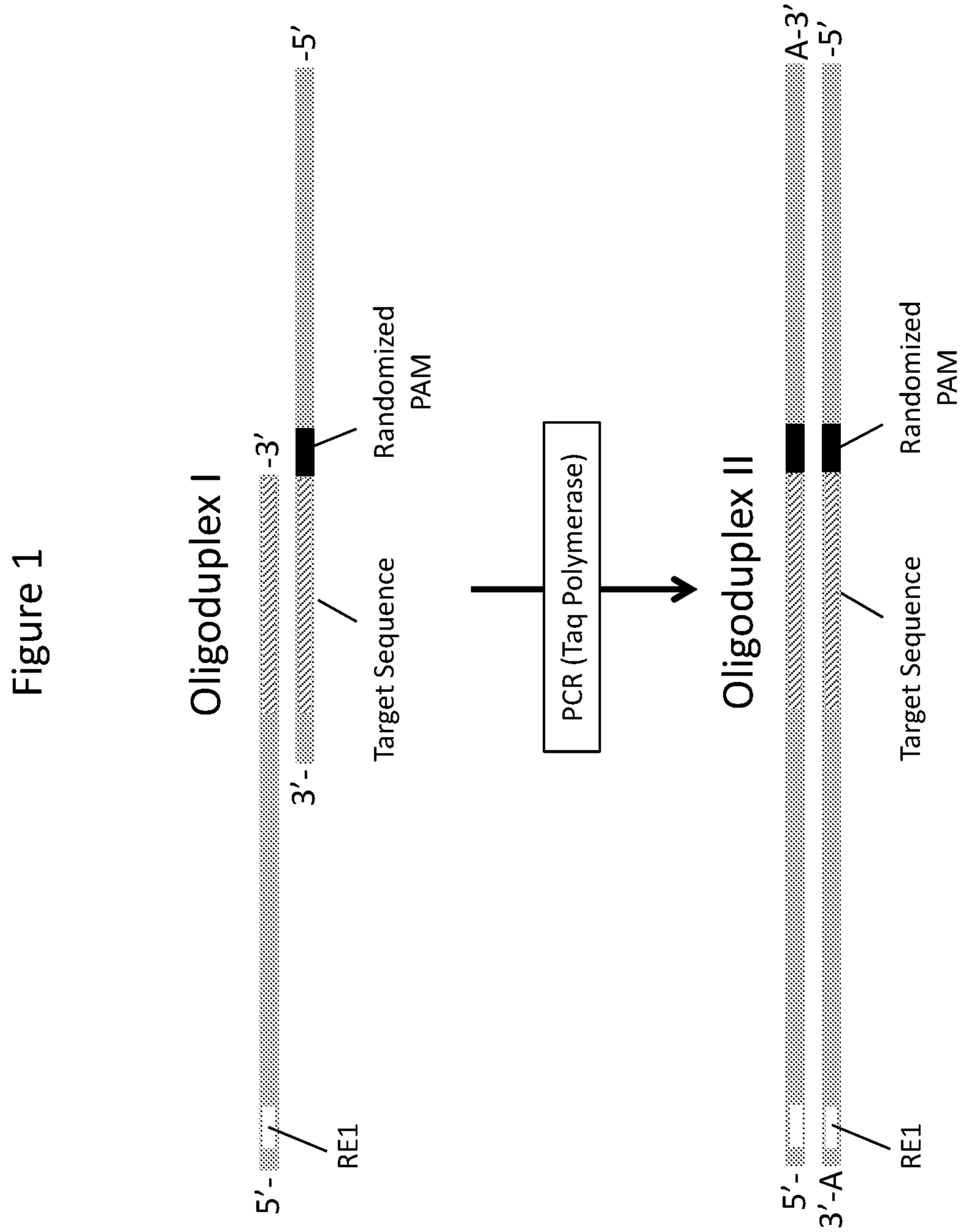

Oligoduplex I
5'-
3'-
RE1
Target Sequence
Randomized PAM
PCR (Taq Polymerase)
Oligoduplex II
5'-
3'-A
A-3'
-5'
RE1
Target Sequence
Randomized PAM

Figure 2

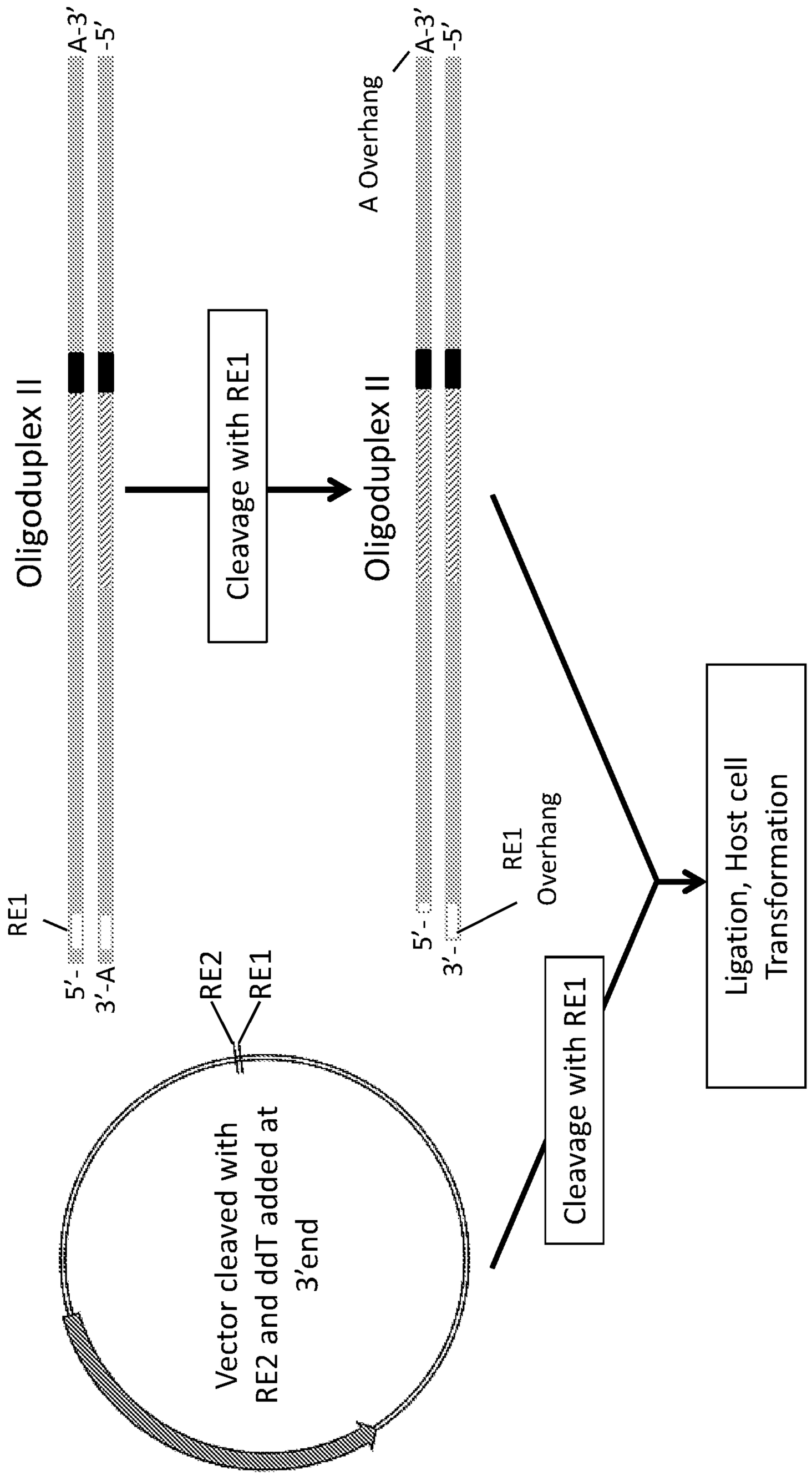

Oligoduplex II
Cleavage with RE1
Oligoduplex II
A Overhang
A-3'
-5'
RE1
5'-
3'-A
RE2
RE1
Vector cleaved with RE2 and ddT added at 3'end
5'-
3'-
RE1
Overhang
Cleavage with RE1
Ligation, Host cell Transformation

Figure 3

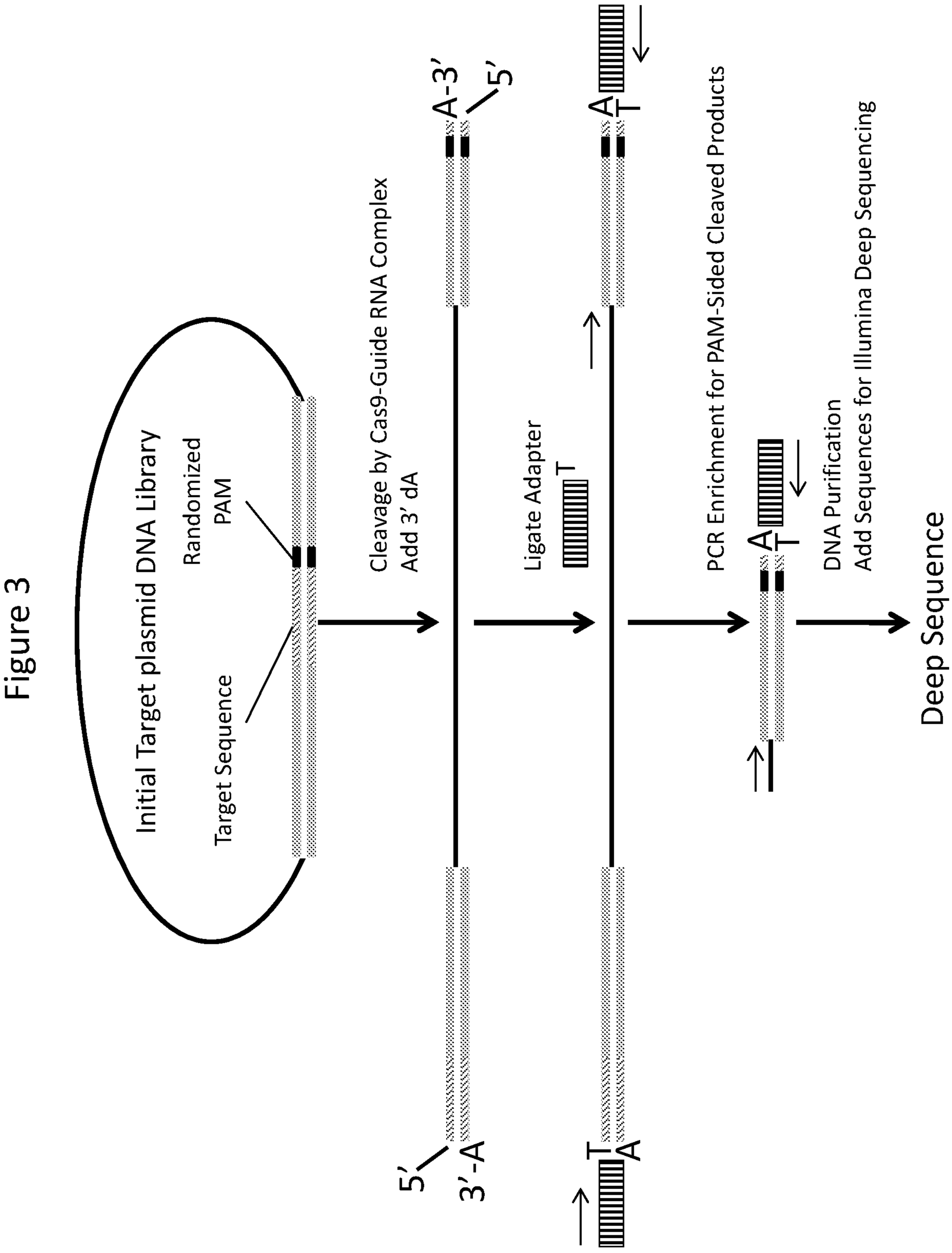

Initial Target plasmid DNA Library
Randomized PAM
Target Sequence
Cleavage by Cas9-Guide RNA Complex
Add 3' dA
Ligate Adapter
PCR Enrichment for PAM-Sided Cleaved Products
DNA Purification
Add Sequences for Illumina Deep Sequencing
Deep Sequence
5'
3'-A
T
A
A-3'
5'
A
T

Figure 4

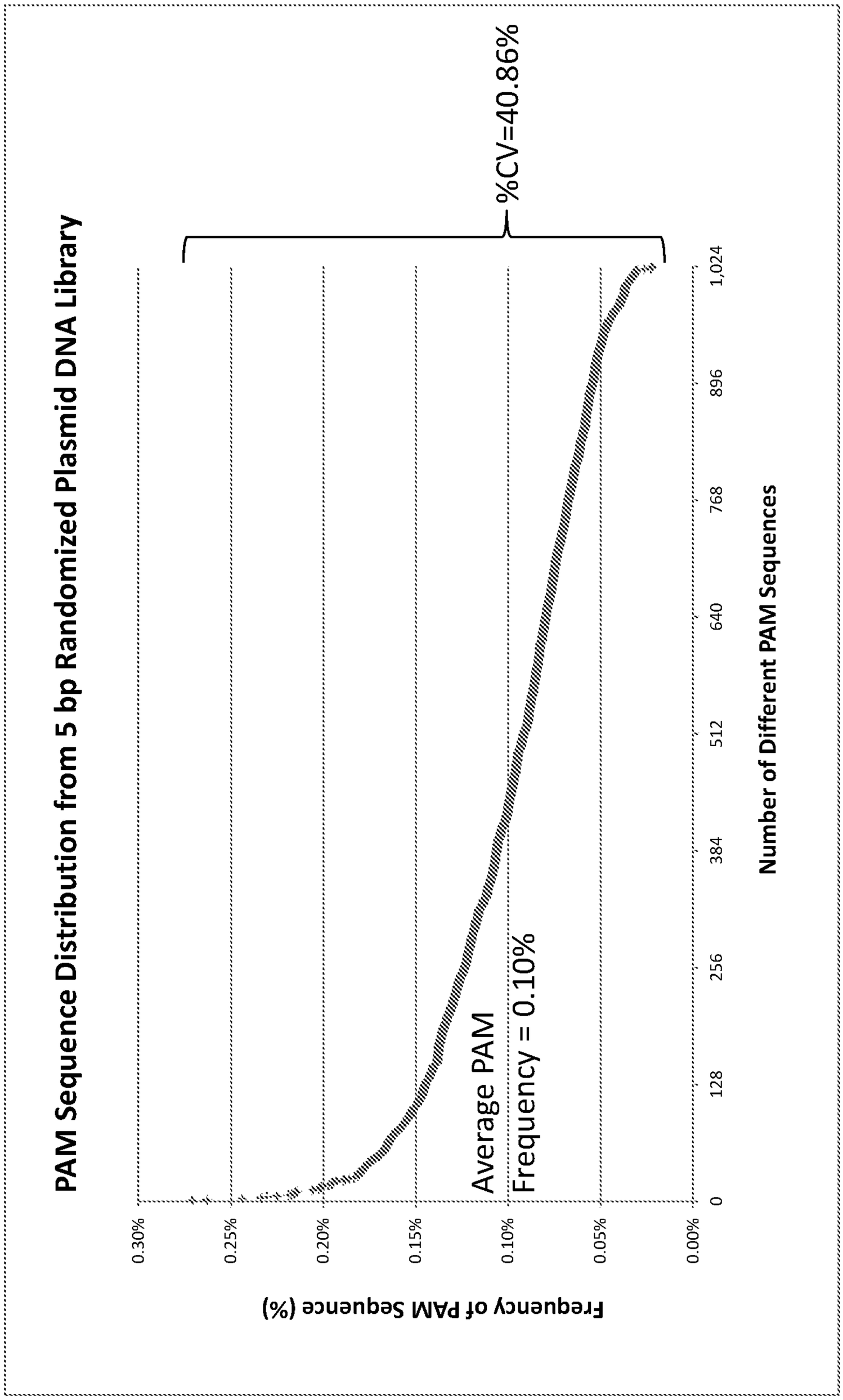

PAM Sequence Distribution from 5 bp Randomized Plasmid DNA Library
%CV=40.86%
Average PAM Frequency = 0.10%
Frequency of PAM Sequence (%)
0.30%
0.25%
0.20%
0.15%
0.10%
0.05%
0.00%
0
128
256
384
512
640
768
896
1,024
Number of Different PAM Sequences

Figure 5

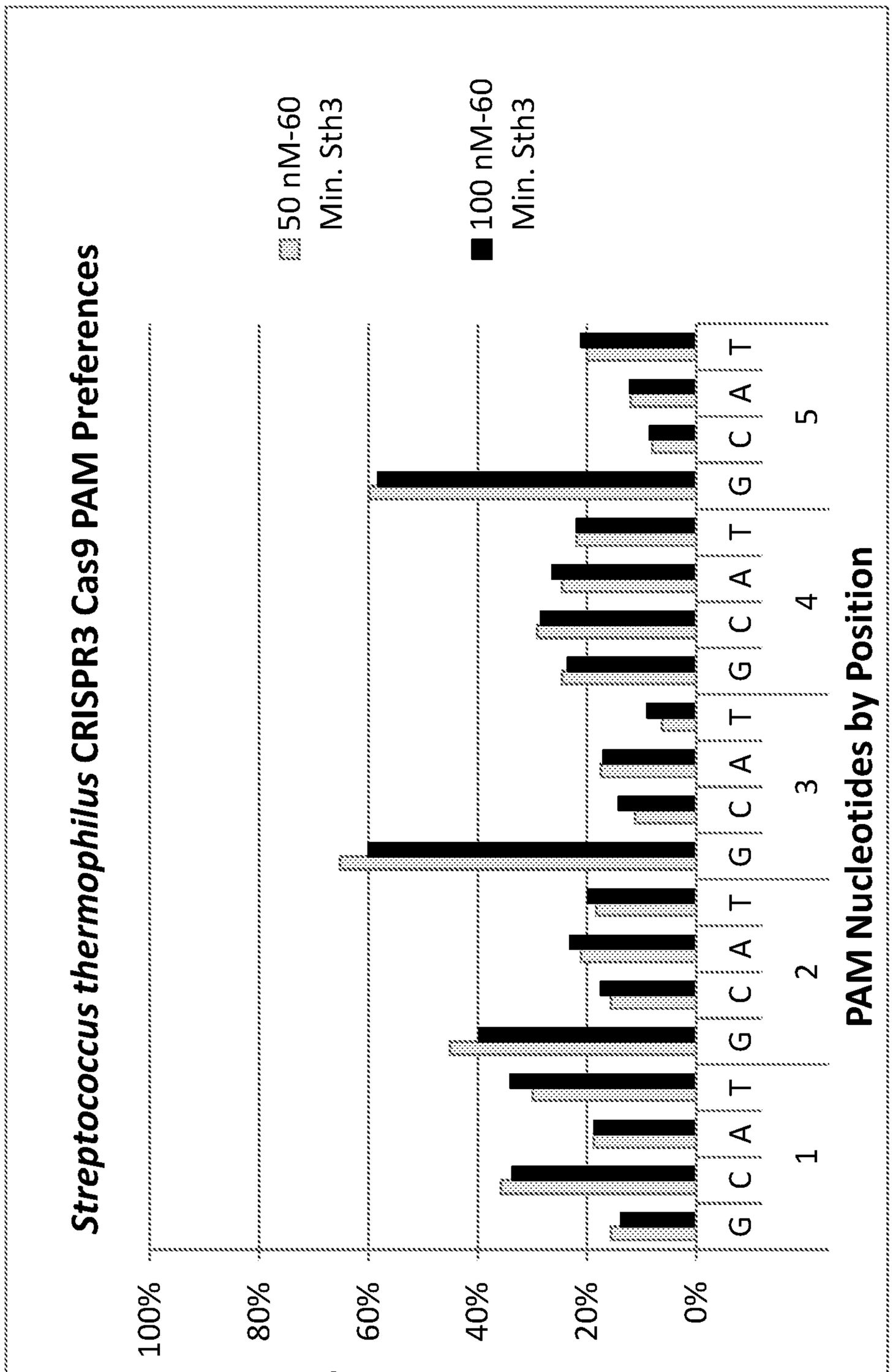

Streptococcus thermophilus CRISPR3 Cas9 PAM Preferences
50 nM-60 Min. Sth3
100 nM-60 Min. Sth3
PAM Nucleotides by Position
1
2
3
4
5
G C A T G C A T G C A T G C A T G C A T
% Nucleotide Composition
100%
80%
60%
40%
20%
0%

Figure 6

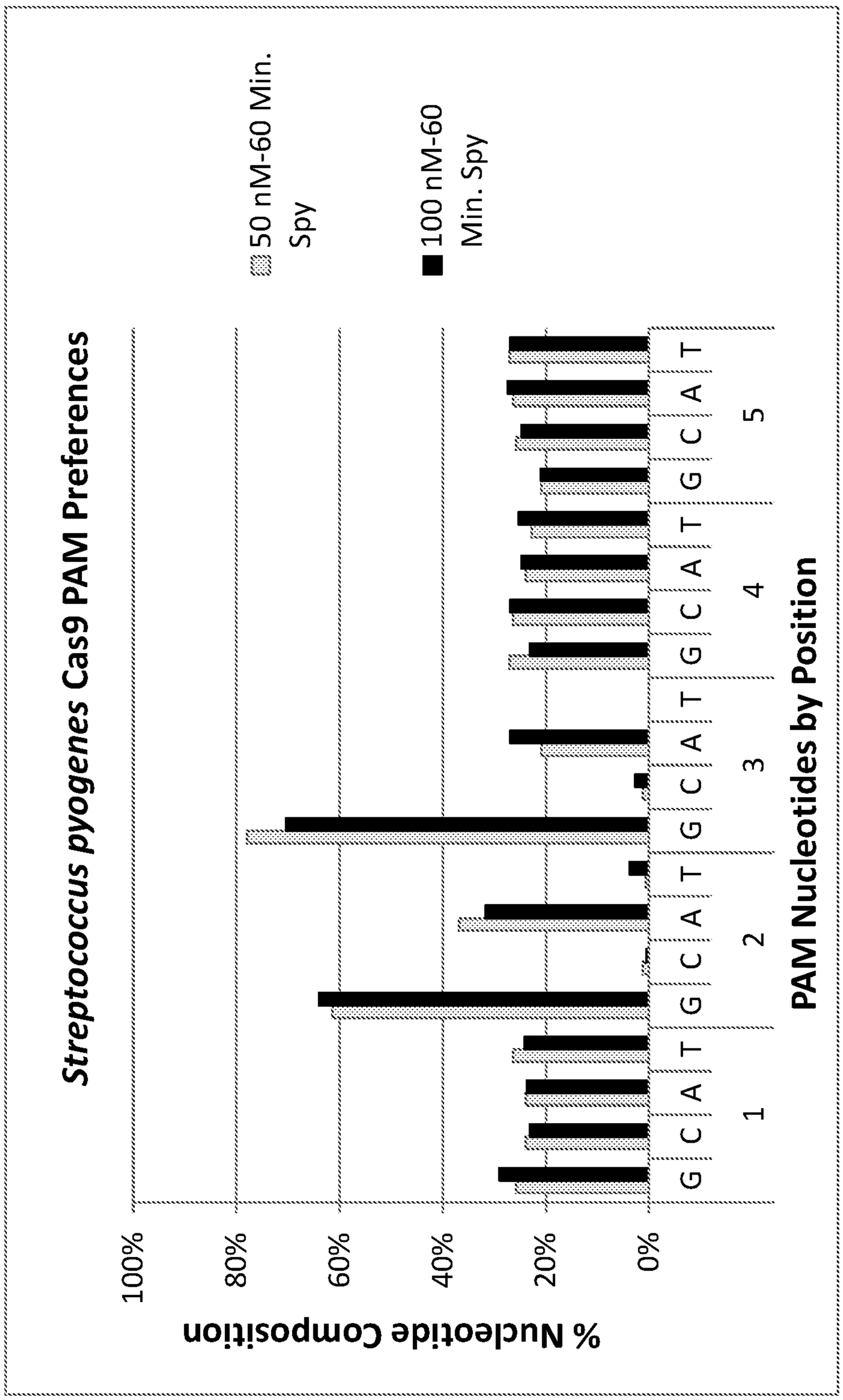

Streptococcus pyogenes Cas9 PAM Preferences
50 nM-60 Min. Spy
100 nM-60 Min. Spy
PAM Nucleotides by Position
1
2
3
4
5
G C A T
% Nucleotide Composition
100%
80%
60%
40%
20%
0%

Figure 7

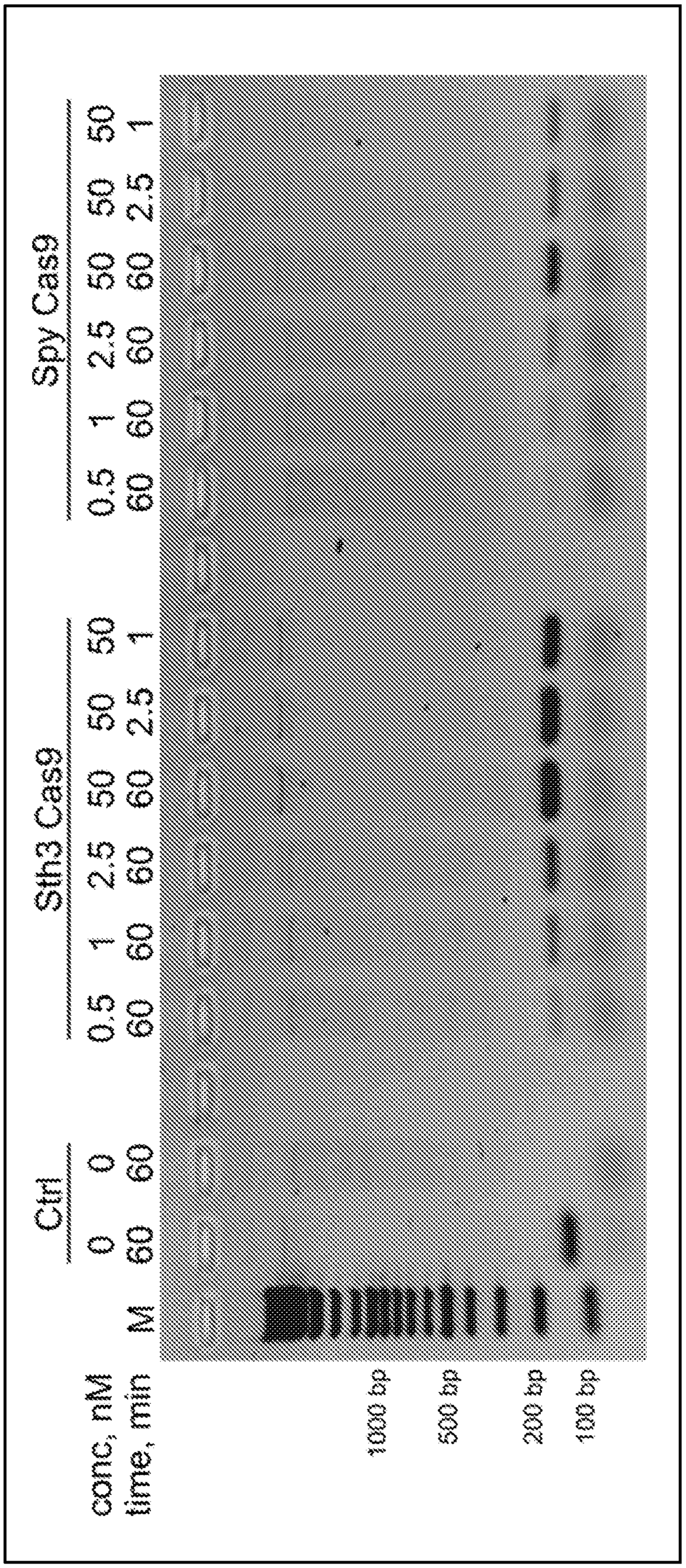

conc, nM
time, min
M
Ctrl
0 0
60 60
Sth3 Cas9
0.5 1 2.5 50
60 60 60 60
Spy Cas9
0.5 1 2.5 50
60 60 60 60
Sth3 Cas9
50 50 50 50
60 2.5 1
Spy Cas9
50 50
2.5 1
9
1000 bp
500 bp
200 bp
100 bp

Figure 8

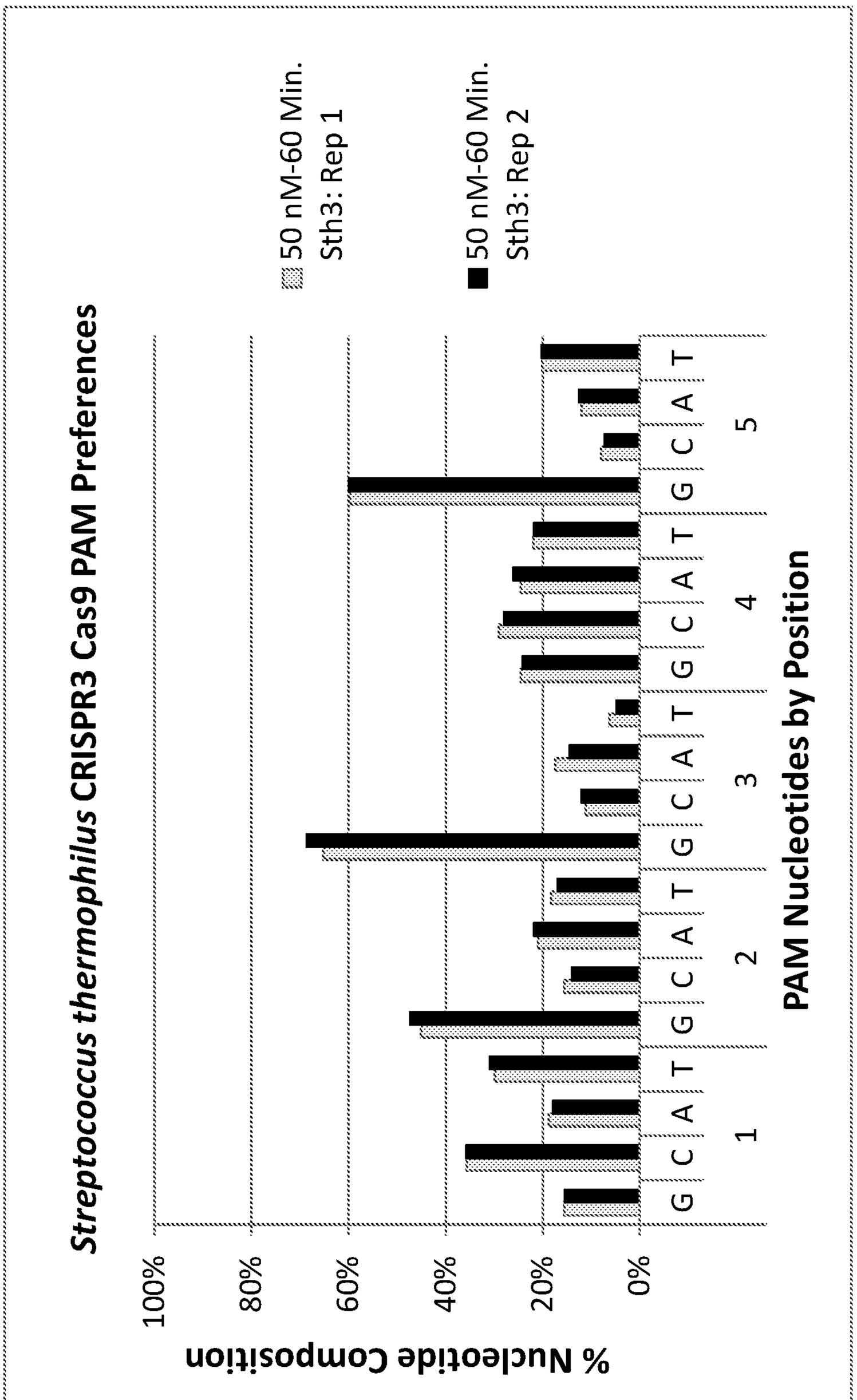

Streptococcus thermophilus CRISPR3 Cas9 PAM Preferences
% Nucleotide Composition
100%
80%
60%
40%
20%
0%
G C A T
1
G C A T
2
G C A T
3
G C A T
4
G C A T
5
PAM Nucleotides by Position
50 nM-60 min.
Sth3: Rep 1
50 nM-60 min.
Sth3: Rep 2

Figure 9

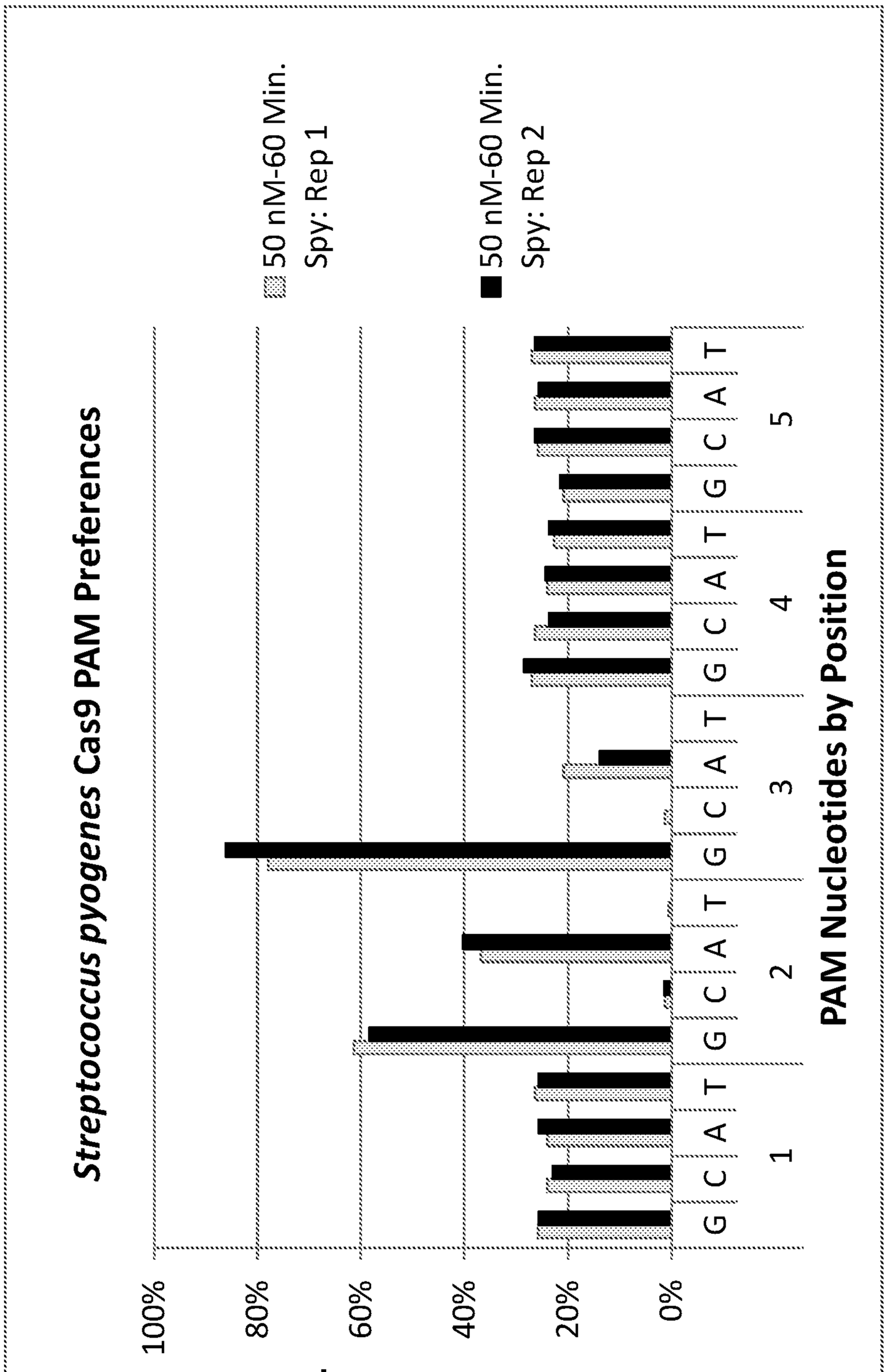

Streptococcus pyogenes Cas9 PAM Preferences
50 nM-60 Min. Rep 1
Spy: NGG
50 nM-60 Min. Rep 2
Spy: NGG
PAM Nucleotides by Position
1
2
3
4
5
G C A T G C A T G C A T G C A T G C A T
% Nucleotide Composition
100%
80%
60%
40%
20%
0%

Figure 10

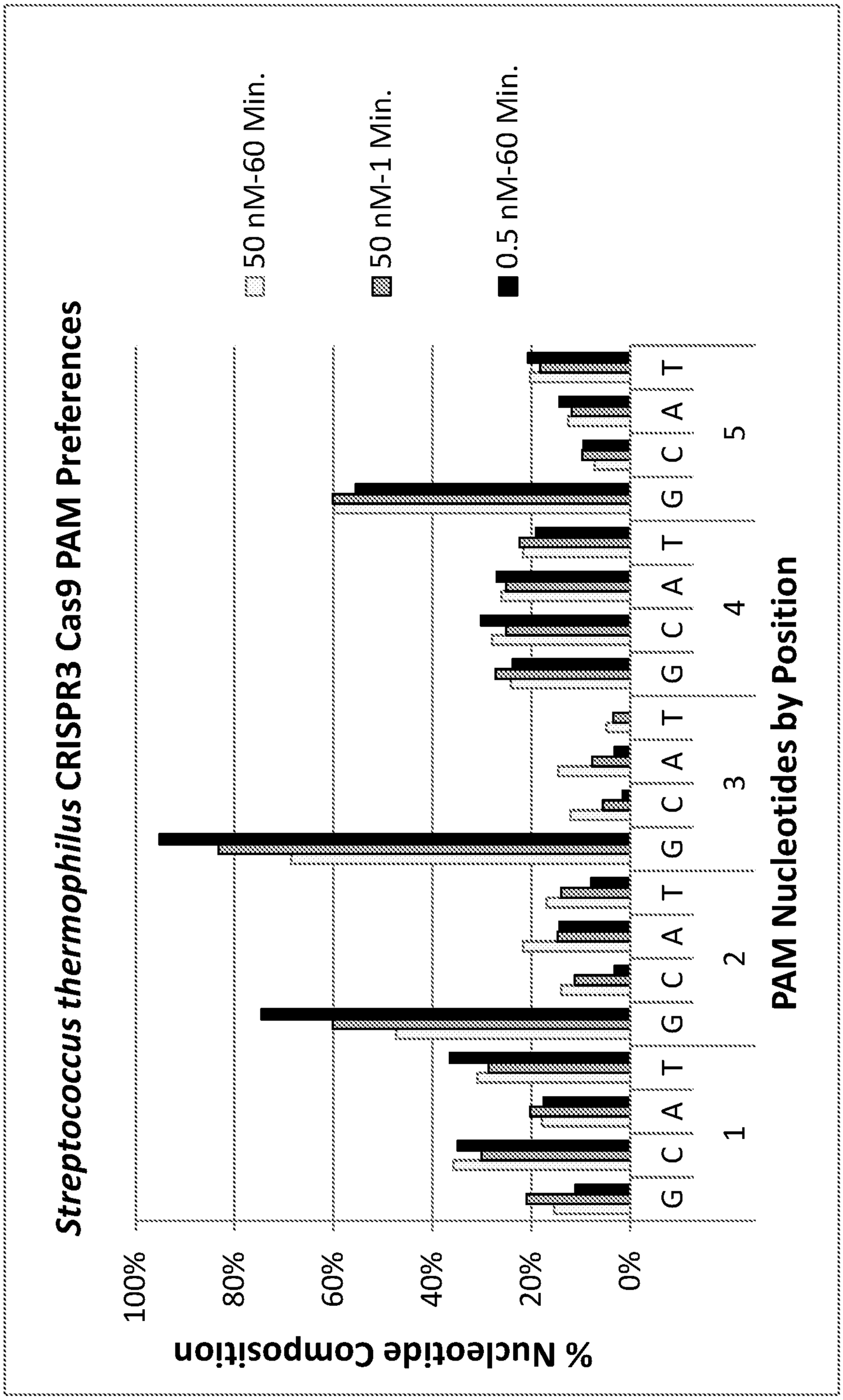

St1CRISPR3 Cas9 PAM Preferences
50nM-60 Min.
50nM-1 Min.
0.5nM-60 Min.
% Nucleotide Composition
100%
80%
60%
40%
20%
0%
PAM Nucleotides by Position
1
2
3
4
5
G C A T G C A T G C A T G C A T T

Figure 11

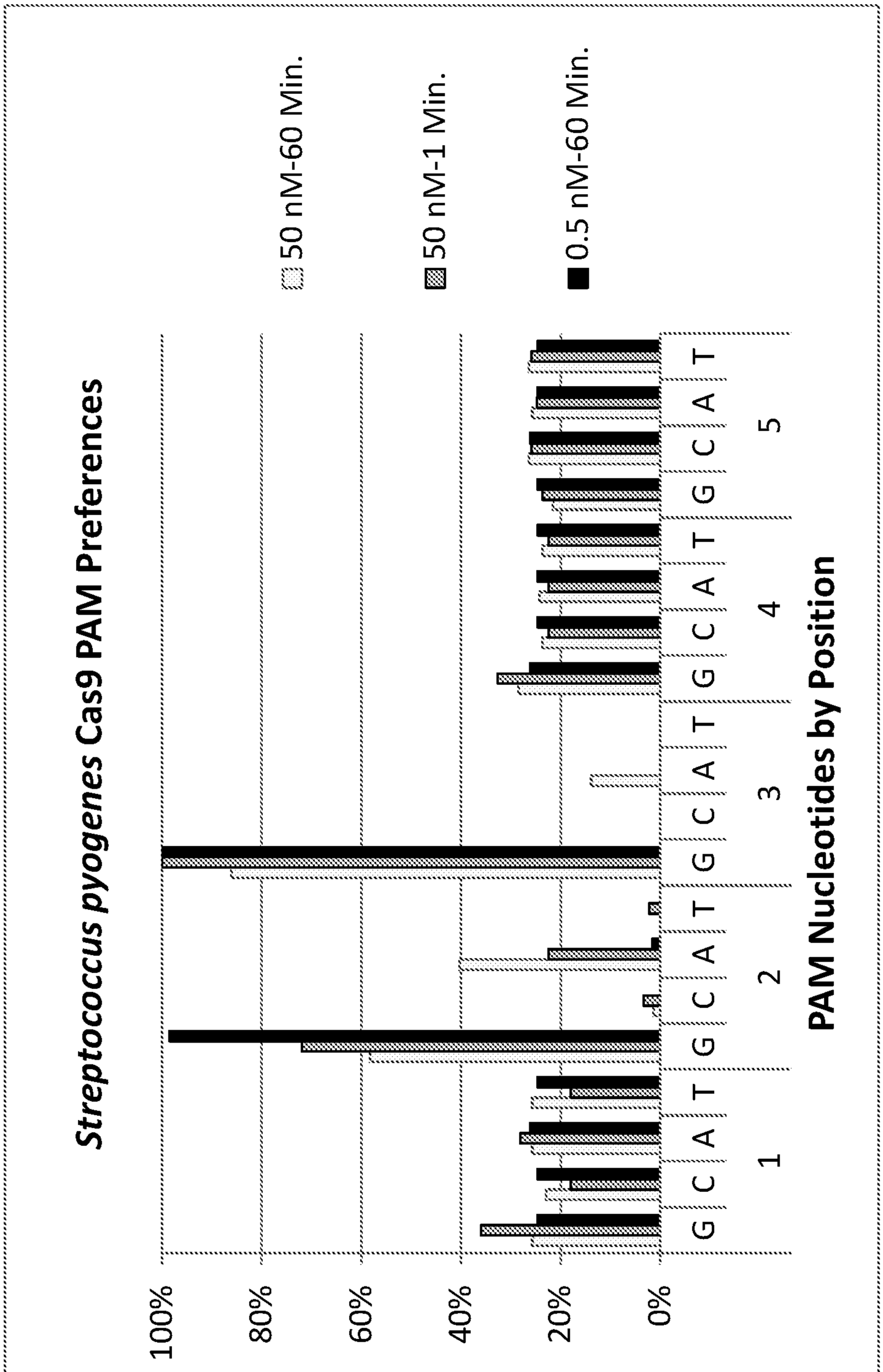

Streptococcus pyogenes Cas9 PAM Preferences
50 nM-60 Min.
50 nM-1 Min.
0.5 nM-60 Min.
PAM Nucleotides by Position
1
2
3
4
5
G C A T G C A T G C A T G C A T G C A T
% Nucleotide Composition
100%
80%
60%
40%
20%
0%

Figure 12

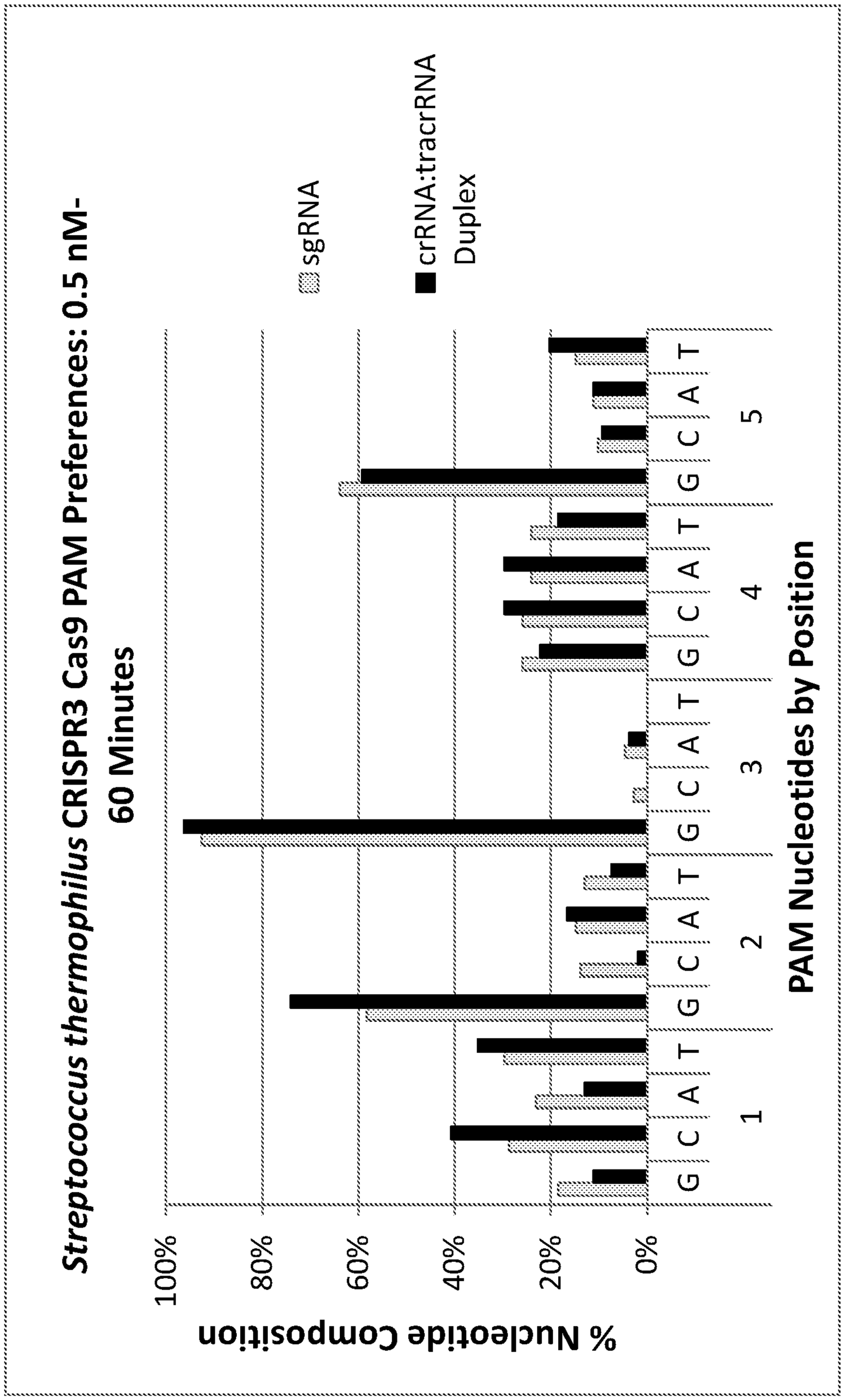
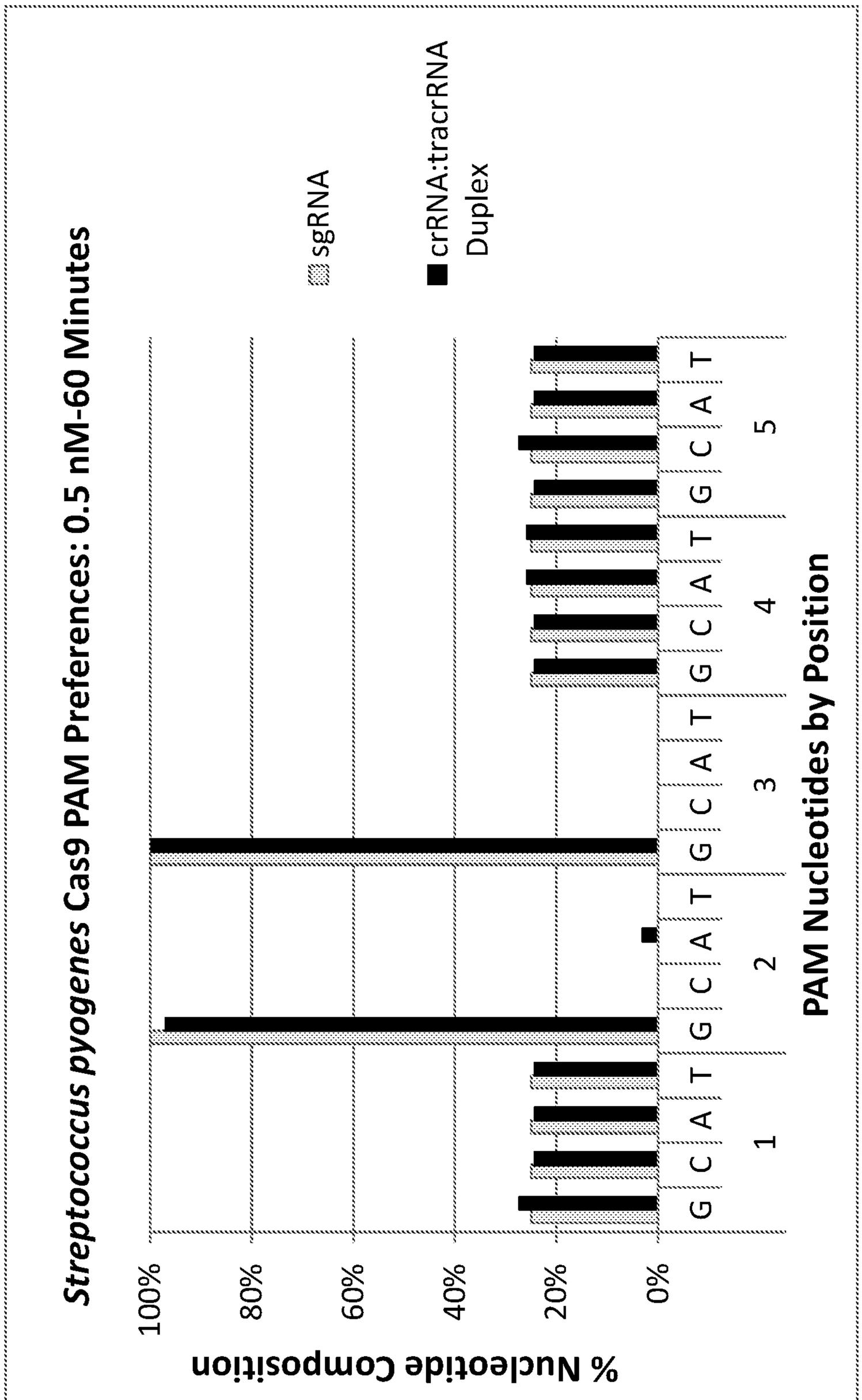

Streptococcus thermophilus CRISPR3 Cas9 PAM Preferences: 0.5 nM-60 Minutes
sgRNA
crRNA:tracrRNA Duplex
PAM Nucleotides by Position
1
2
3
4
5
G C A T
G C A T
G C A T
G C A T
T G C A T
0%
20%
40%
60%
80%
100%
% Nucleotide Composition

Figure 13


Streptococcus pyogenes Cas9 PAM Preferences: 0.5 nM-60 Minutes
% Nucleotide Composition
100%
80%
60%
40%
20%
0%
PAM Nucleotides by Position
1
2
3
4
5
G C A T G C A T G C A T G C A T
sgRNA
crRNA:tracrRNA Duplex

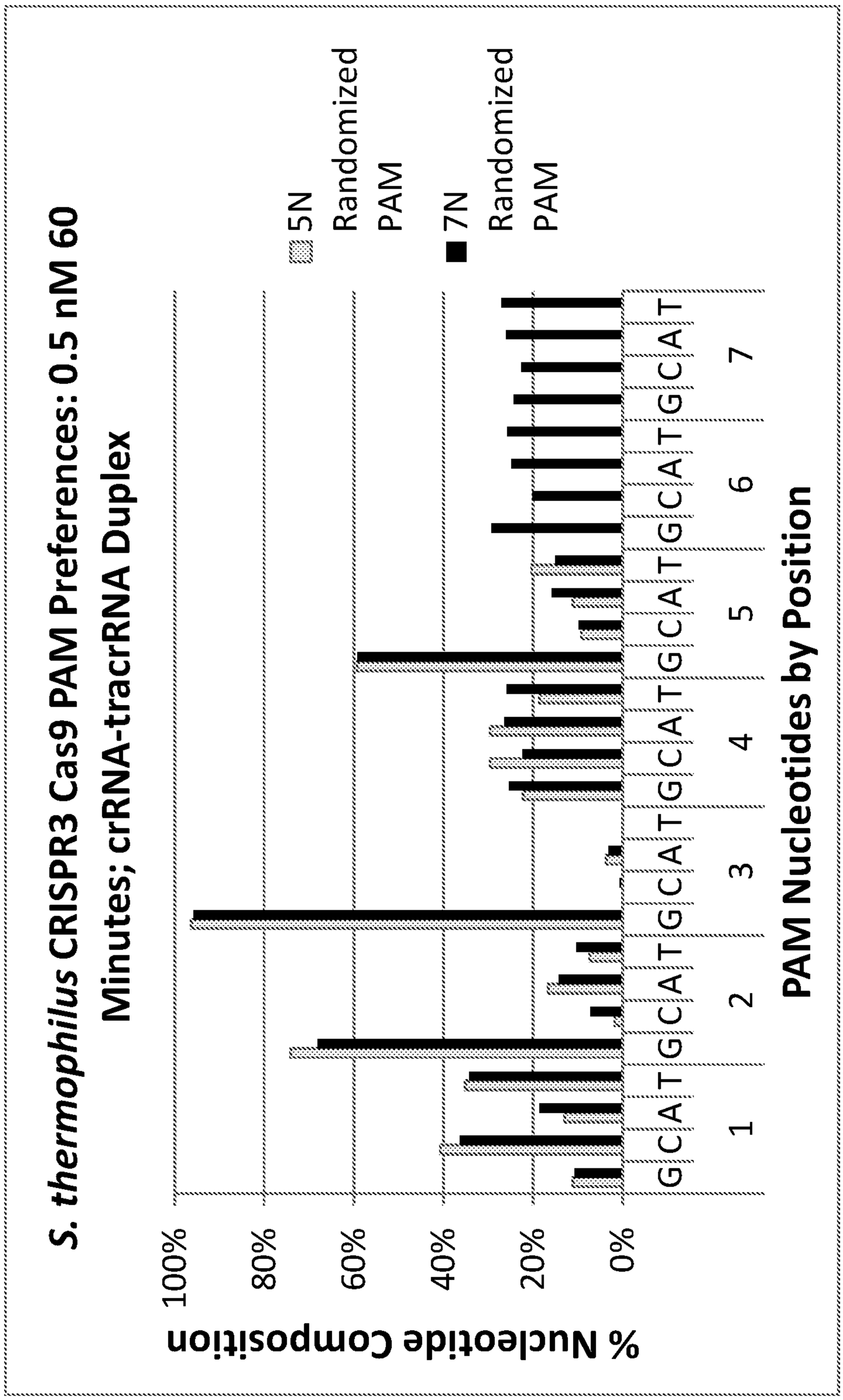

S. thermophilus CRISPR3 Cas9 PAM Preferences: 0.5 nM 60 Minutes; crRNA-tracrRNA Duplex
5N Randomized PAM
7N Randomized PAM
PAM Nucleotides by Position
1
2
3
4
5
6
7
G C A T G C A T G C A T G C A T G C A T G C A T G C A T
100%
80%
60%
40%
20%
0%
% Nucleotide Composition

Figure 14

Figure 15

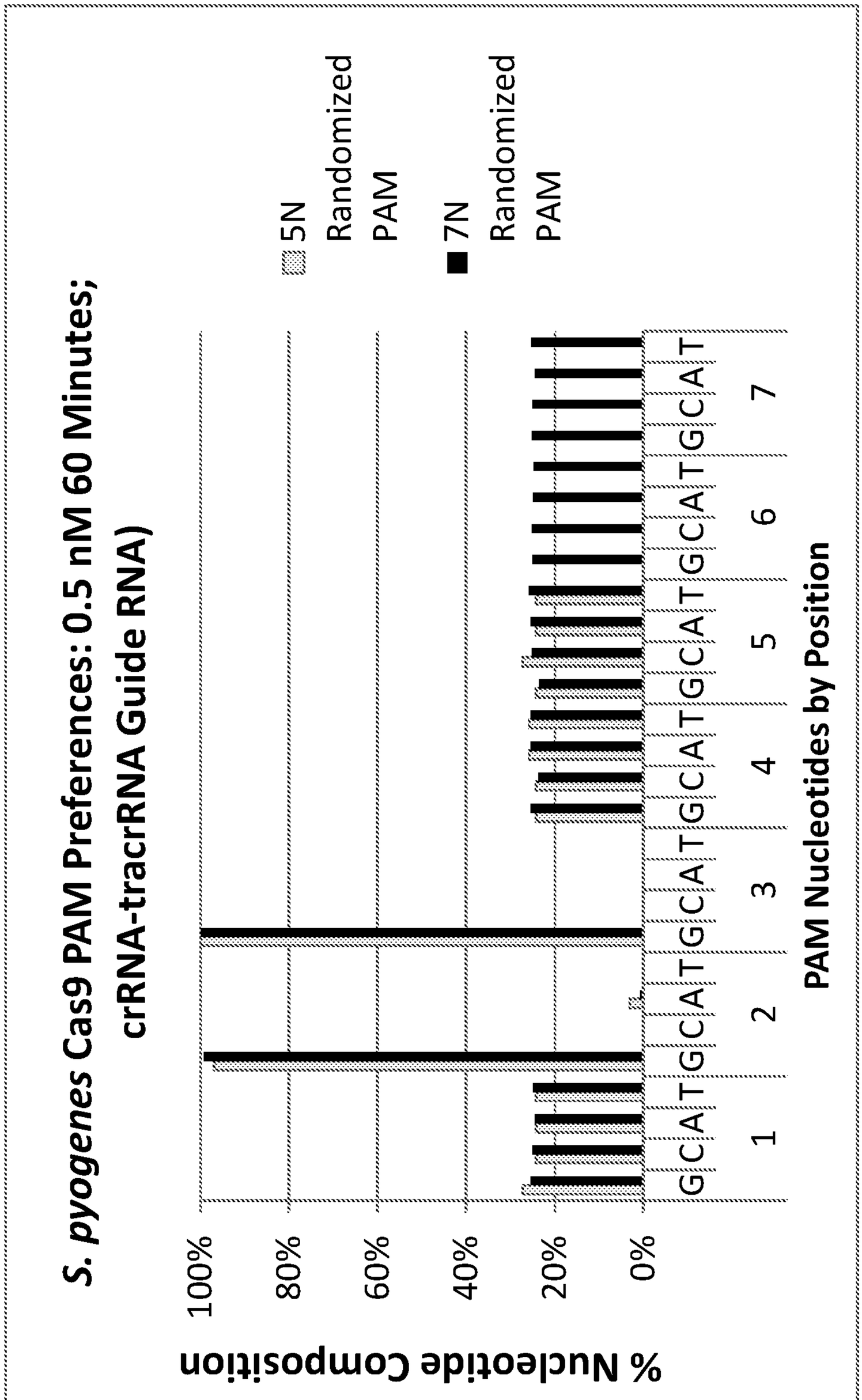

S. pyogenes Cas9 PAM Preferences: 0.5 nM 60 Minutes;
crRNA-tracrRNA Guide RNA)
5N
Random PAM
7N
Random PAM
100%
80%
60%
40%
20%
0%
% Nucleotide Composition
G C A T G C A T G C A T G C A T
1 2 3 4 5 6 7
PAM Nucleotides by Position

Figure 16

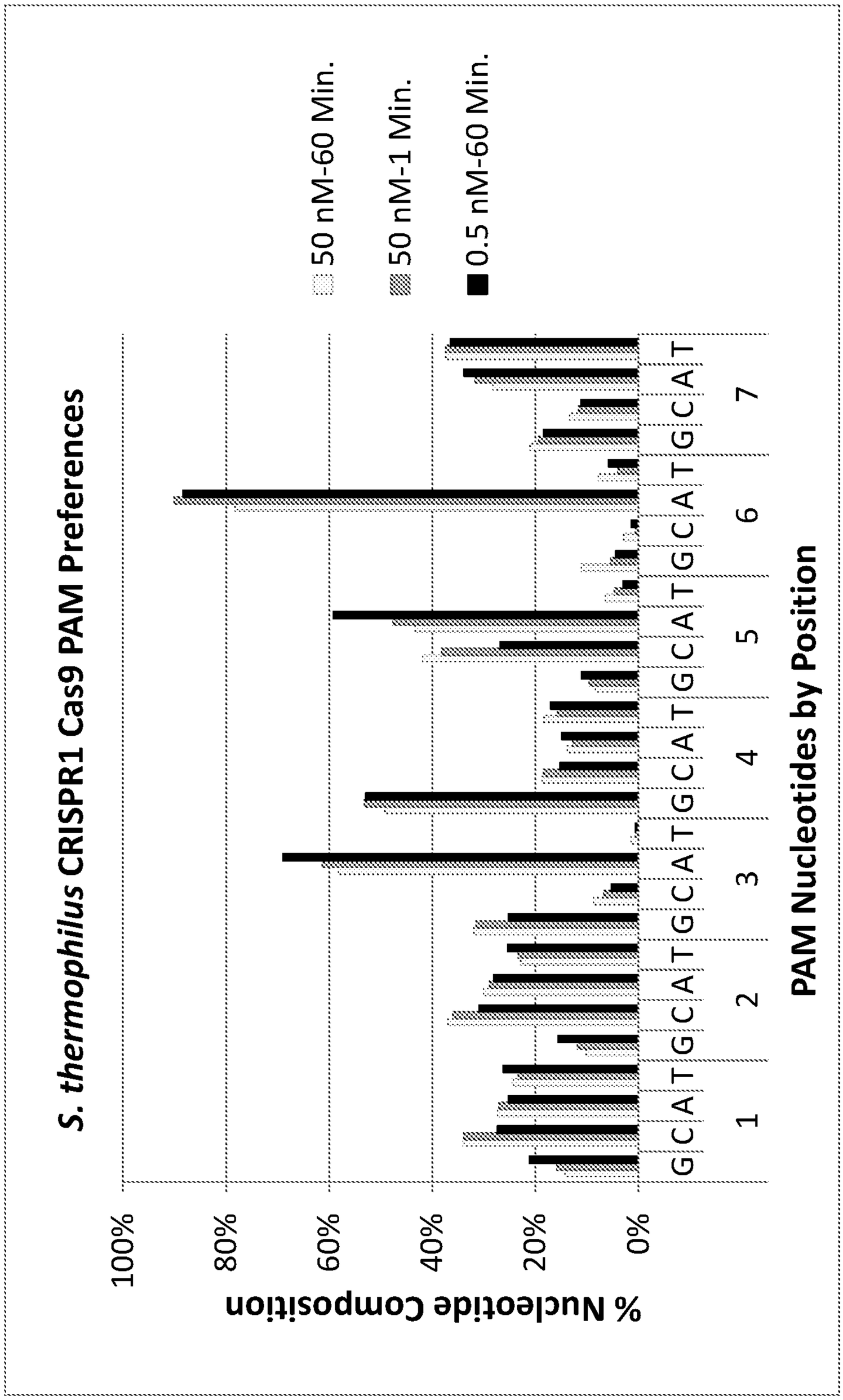

S. thermophilus CRISPR1 Cas9 PAM Preferences
50nM-60 Min.
50nM-1 Min.
0.5nM-60 Min.
PAM Nucleotides by Position
1 2 3 4 5 6 7
G C A T G C A T G C A T G C A T G C A T G C A T G C A T
% Nucleotide Composition
100%
80%
60%
40%
20%
0%

Figure 17-A

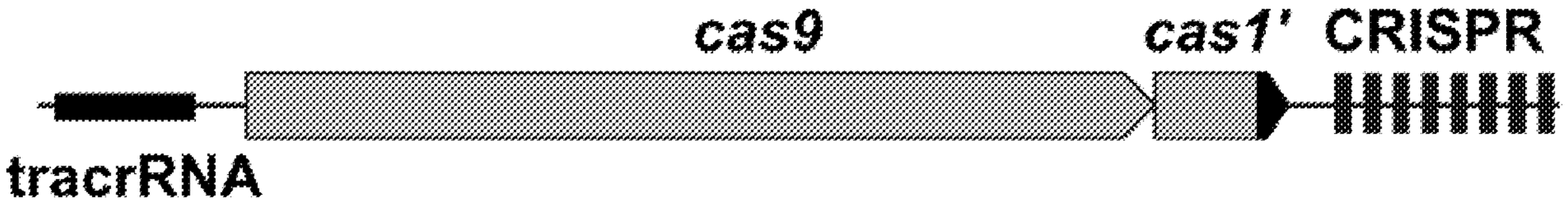


Figure 17-B

Repeat1	ATCATATCATATCGAGTTTTAGTAAGGAACTATAGC	SEQ ID NO: 37
Repeat2	ATCATATCATATCGAG**C**TTTAGTAAGGAACTATAGC	SEQ ID NO: 38
Repeat3	ATCATATCATATCGAGTTTTAGTAAGGAAC**C**ATAGC	SEQ ID NO: 39
Repeat4	ATCATATCATATCGAGTTTTAGTAAGGAACTATAGC	SEQ ID NO: 40
Repeat5	ATCATATCATATCGAGTTTTAGTAAGGAACTATAGC	SEQ ID NO: 41
Repeat6	ATCATATCATATCGAG**C**TT**C**AGTAAGGAACTATAGC	SEQ ID NO: 42
Repeat7	ATCATATCATATC**A**AG**C**TTTAGTAAGGAACTATAGC	SEQ ID NO: 43
Repeat8	ATCATATCATATCGAGTTTTAGTAAGGAACTATAG**T**	SEQ ID NO: 44

WO 2016/186953
PCT/US2016/032073
CA 02985506 2017-11-08

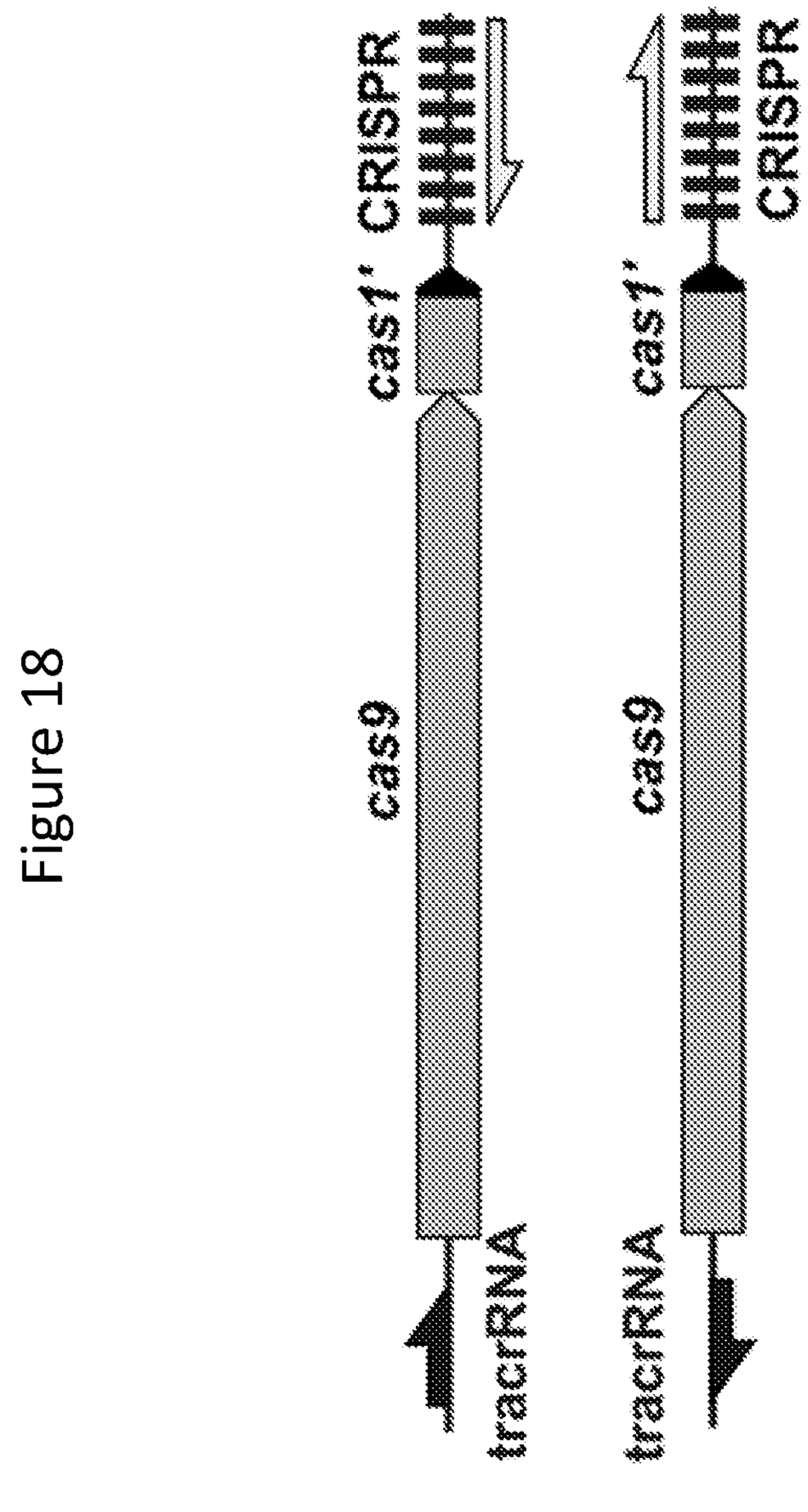


Figure 18

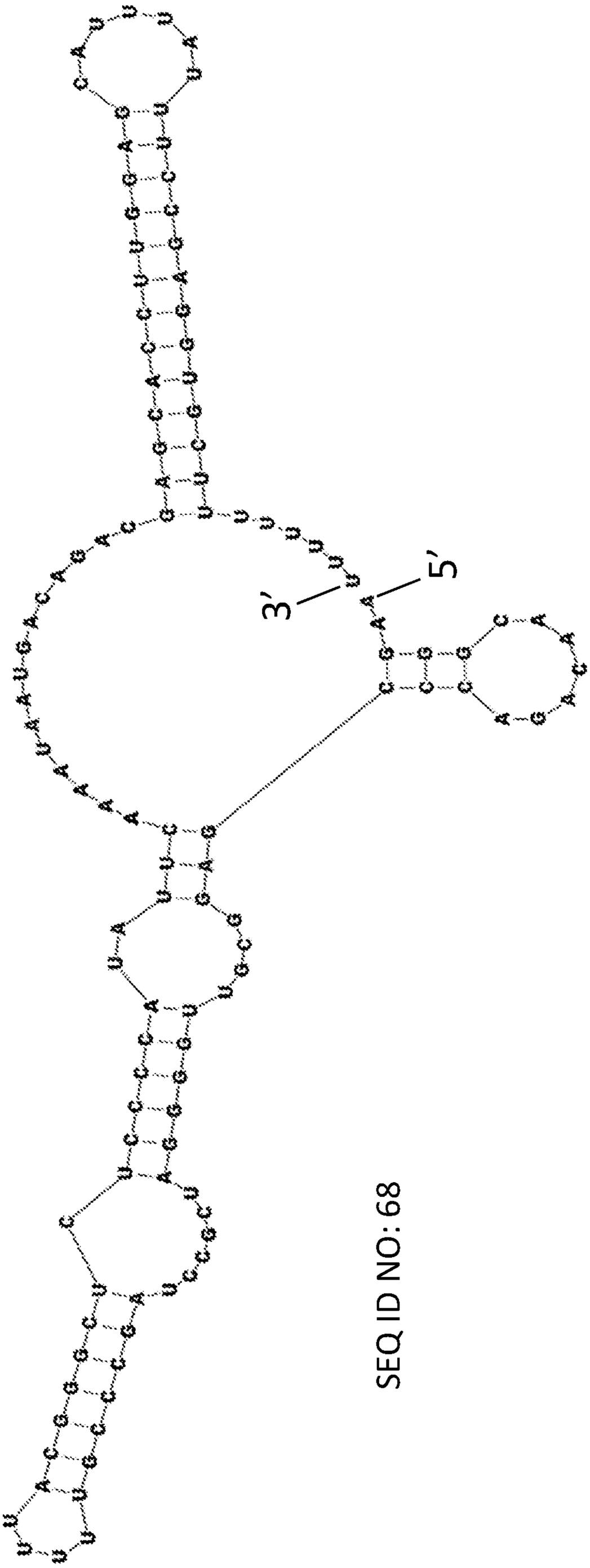


Figure 19

SEQ ID NO: 68

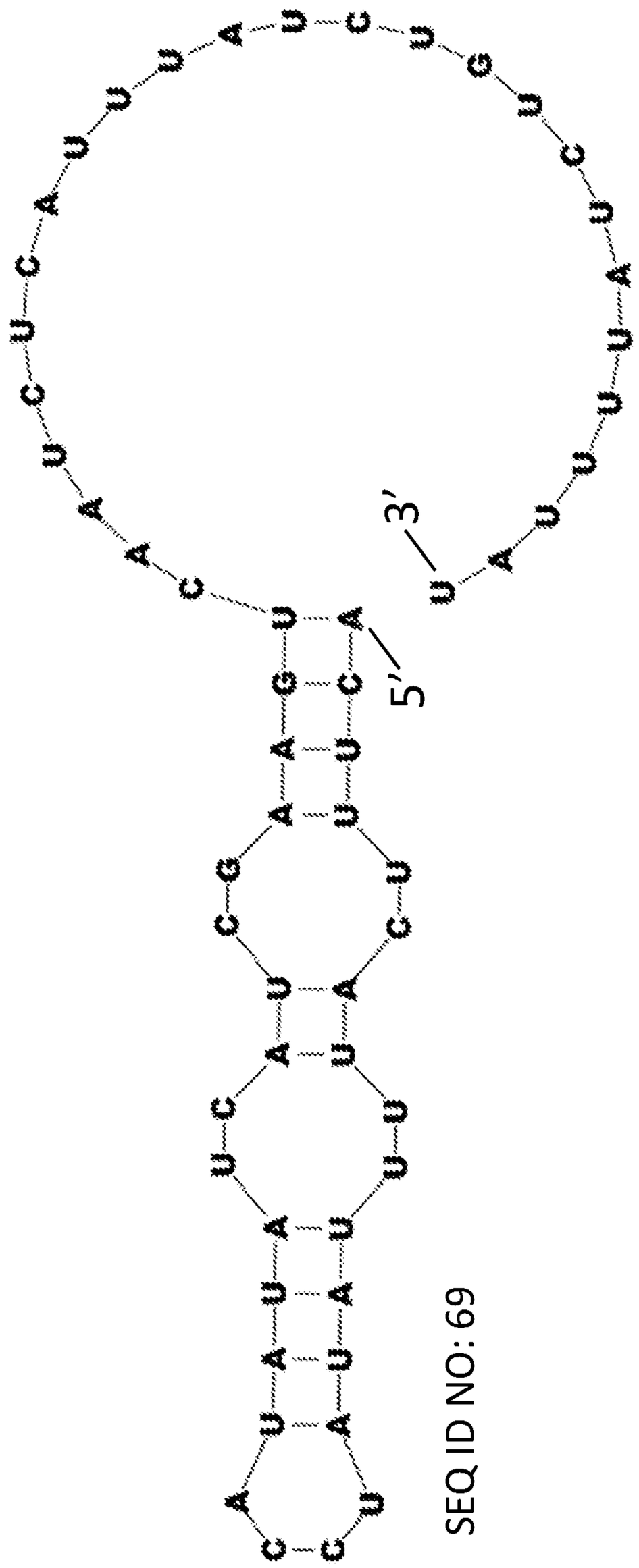
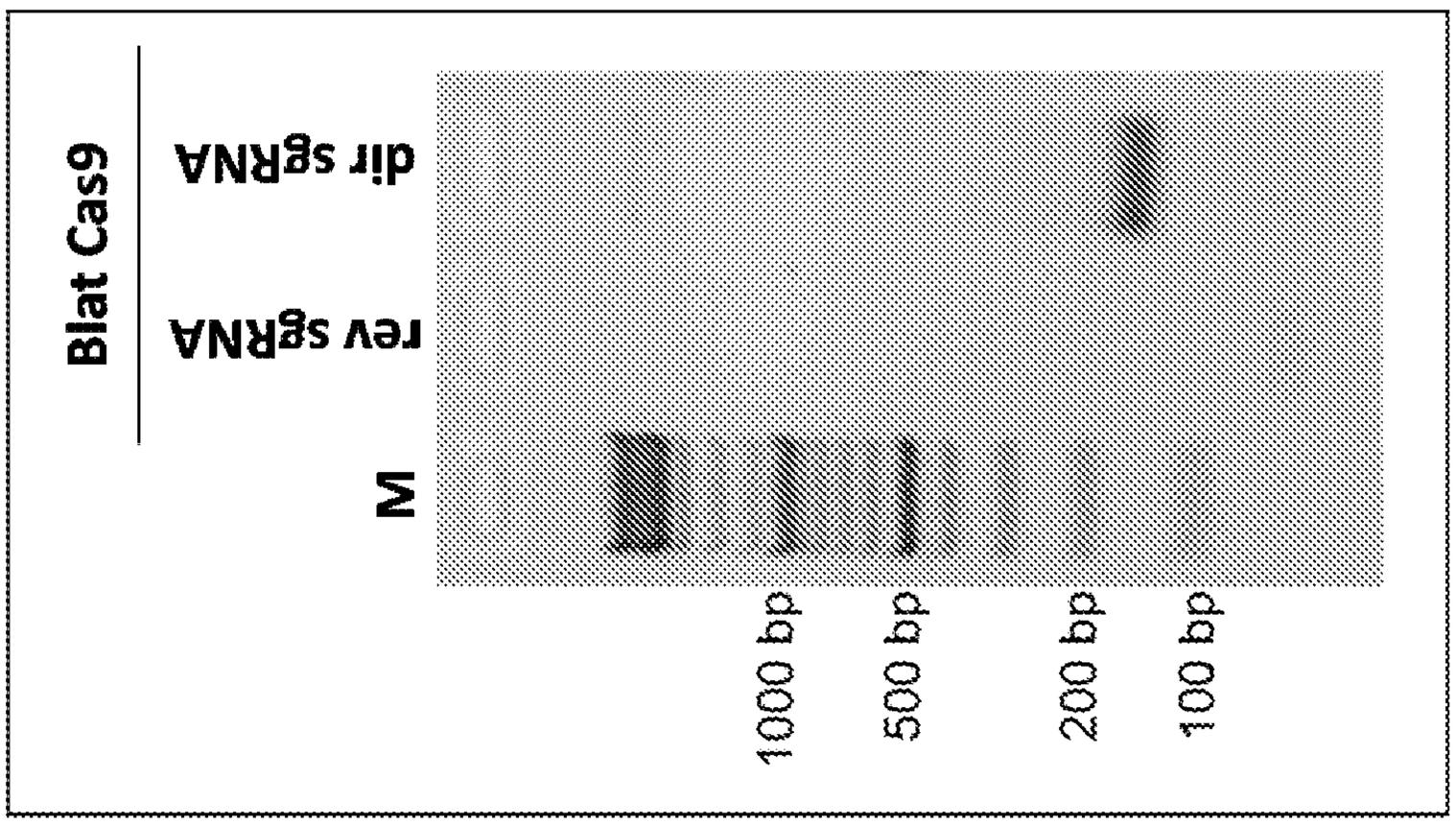


Figure 20

Figure 21



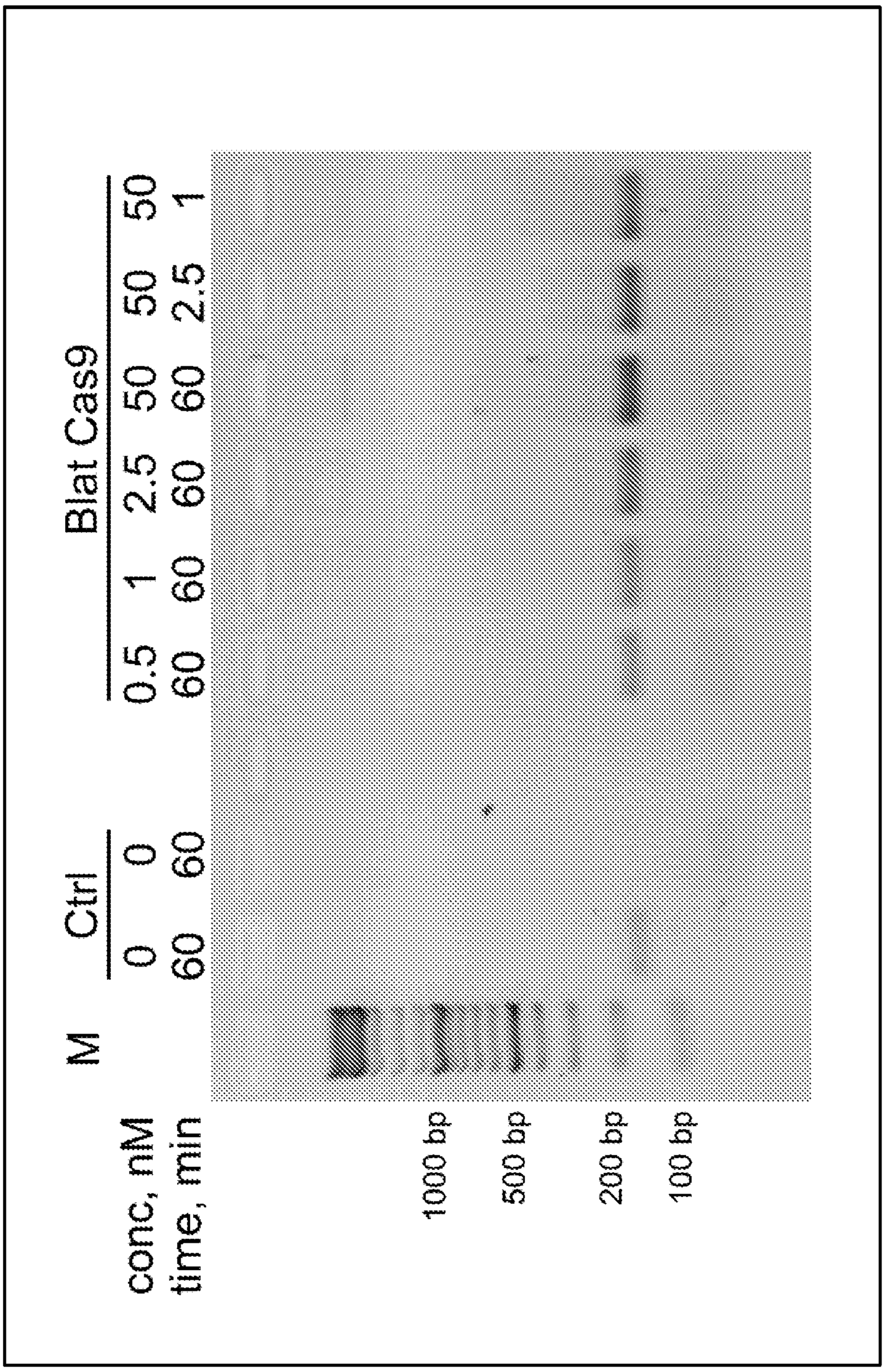

Blat Cas9
0.5 1 2.5 50 50 50
60 60 60 60 2.5 1
M Ctrl
0 0
60 60
conc, nM
time, min
1000 bp
500 bp
200 bp
100 bp

Figure 22

Figure 23

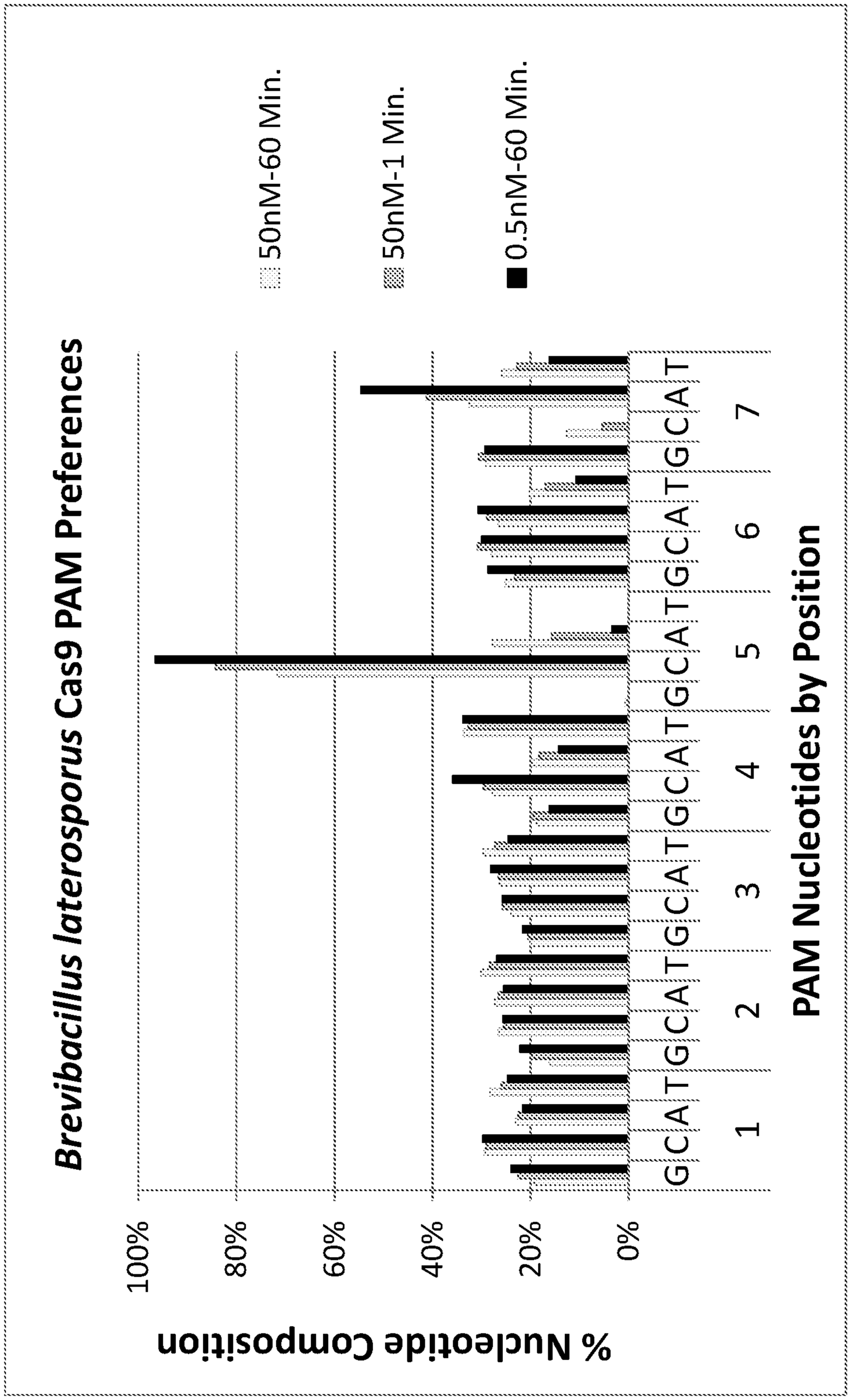
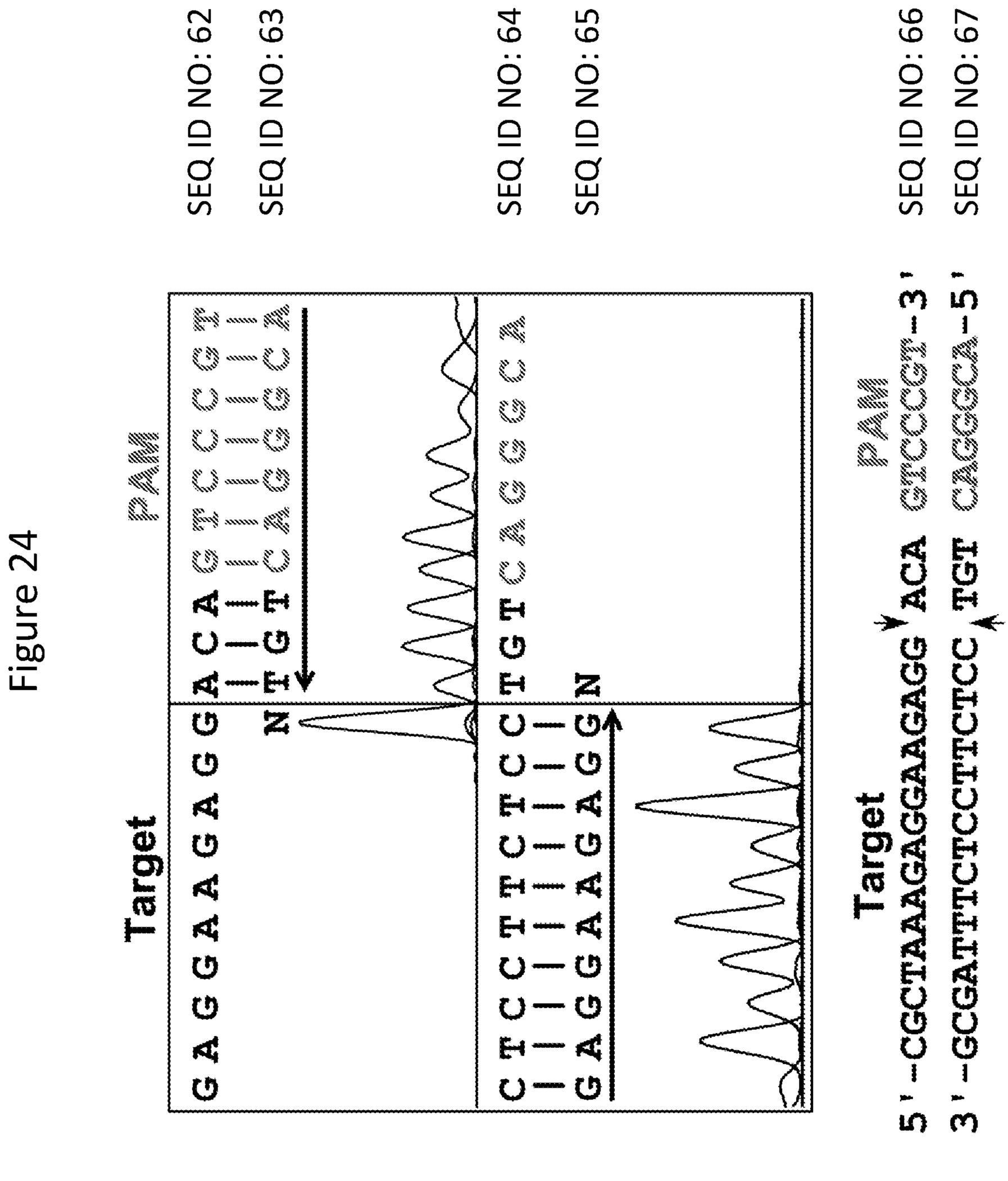

Brevibacillus laterosporus Cas9 PAM Preferences
50nM-60 Min.
50nM-1 Min.
0.5nM-60 Min.
100%
80%
60%
40%
20%
0%
% Nucleotide Composition
PAM Nucleotides by Position
1
2
3
4
5
6
7
G
C
A
T

Figure 24


PAM
Target
G A G G A A G A G G A
N T C
C T C C T T C T C C C
— — — — — — — — — —
G A G G A A G A G G N
PAM
5'-CGCTAAAGAGGAAGAGGG ACA GTCCCGT-3'
3'-GCGATTTCTCCTTCTCCC TGT CAGGCA-5'
Target
SEQ ID NO:62
SEQ ID NO:63
SEQ ID NO:64
SEQ ID NO:65
SEQ ID NO:66
SEQ ID NO:67

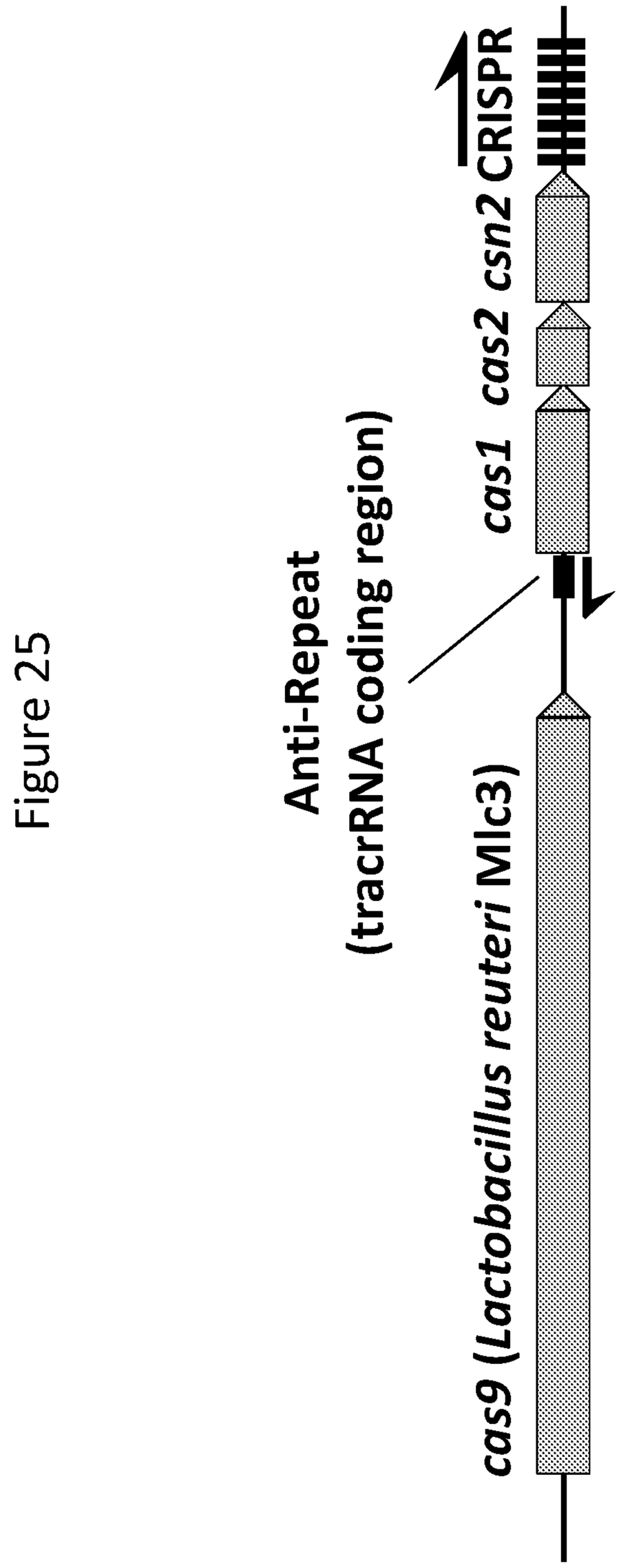

cas1 cas2 csn2 CRISPR
Anti-Repeat
(tracrRNA coding region)
cas9 (Lactobacillus reuteri Mlc3)

Figure 25

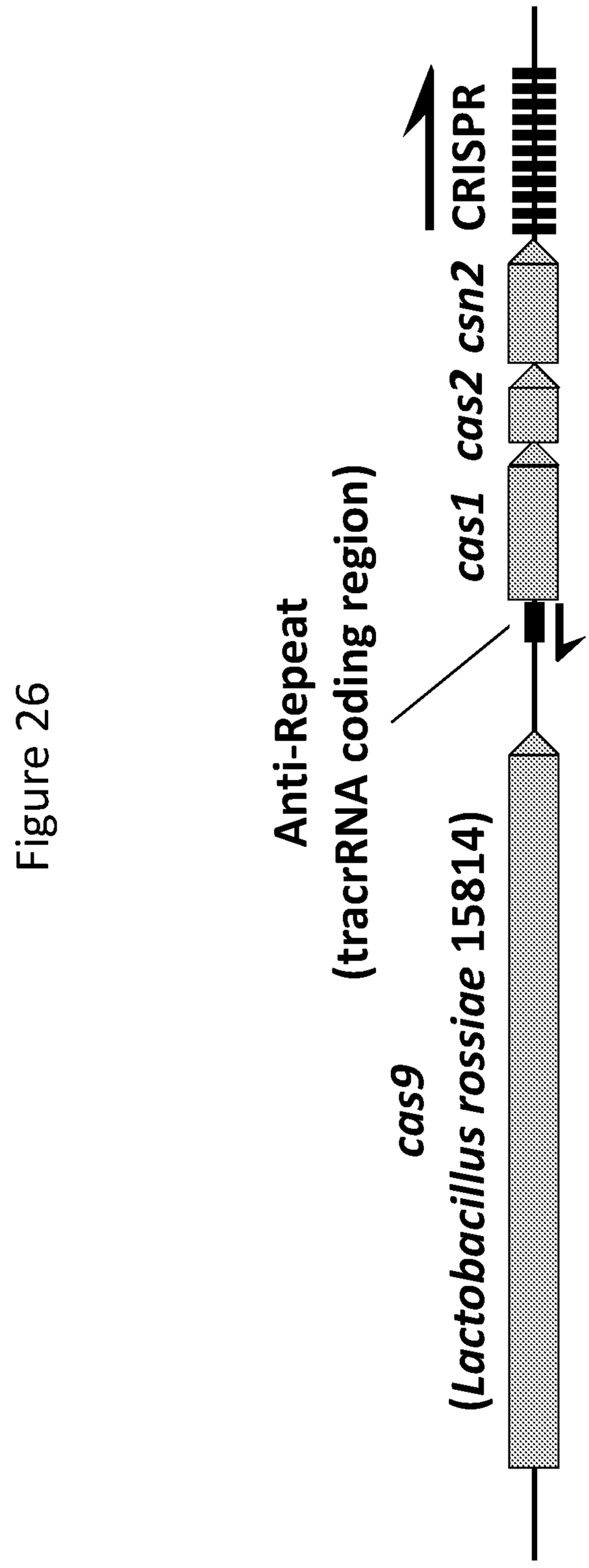

CRISPR
cas1 cas2 csn2
Anti-Repeat
(tracrRNA coding region)
cas9
(Lactobacillus rossiae 15814)

Figure 26

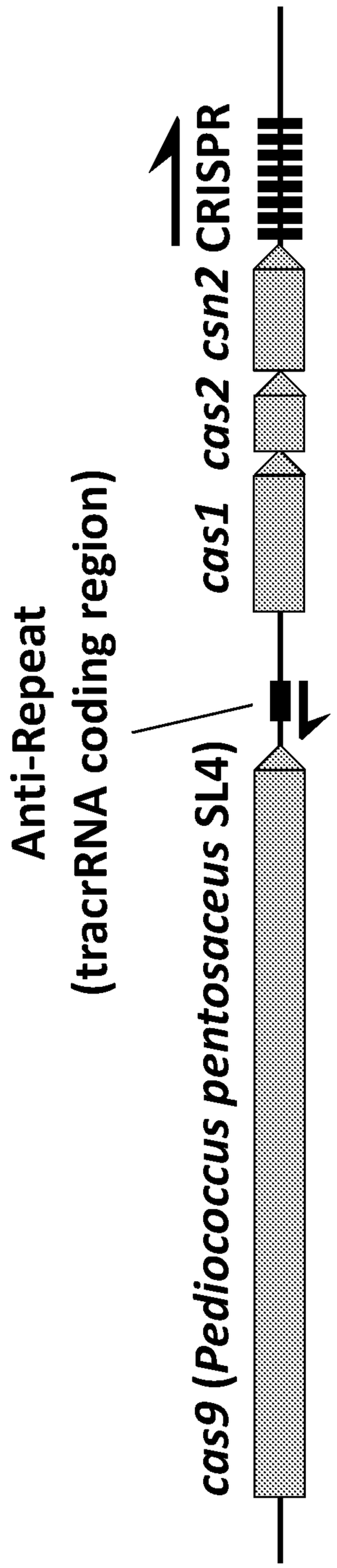

Anti-Repeat
(tracrRNA coding region)
cas1 cas2 csn2 CRISPR
cas9 (Pediococcus pentosaceus SL4)
cas9

Figure 27

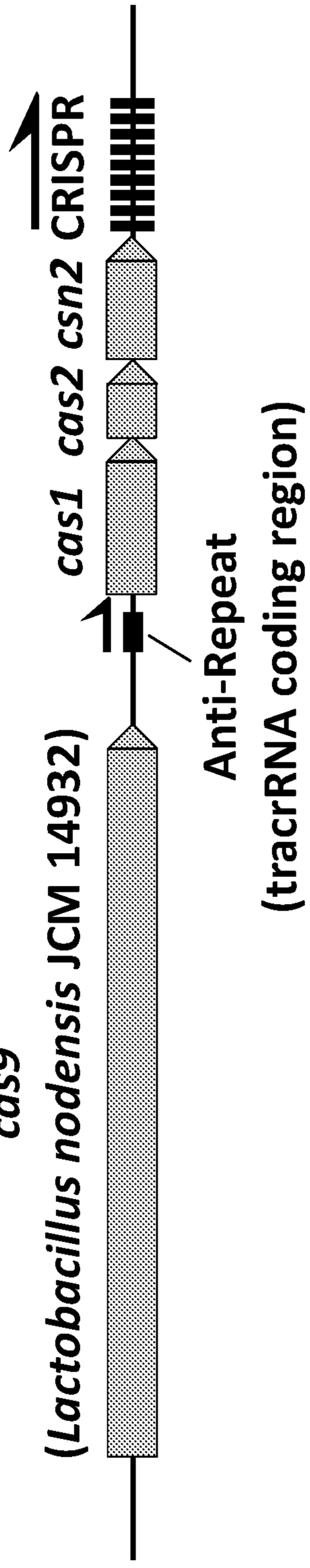

cas1 cas2 csn2 CRISPR
Anti-Repeat
(tracrRNA coding region)
cas9
(Lactobacillus nodensis JCM 14932)

Figure 28

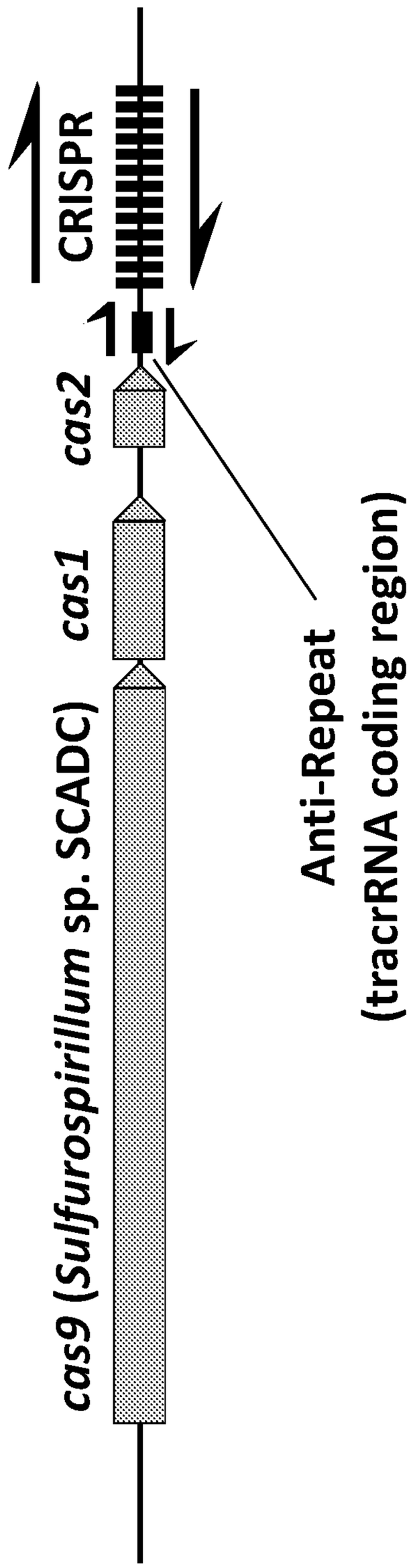

CRISPR
cas2
cas1
Anti-Repeat
(tracrRNA coding region)
cas9 (Sulfurospirillum sp. SCADC)

Figure 29

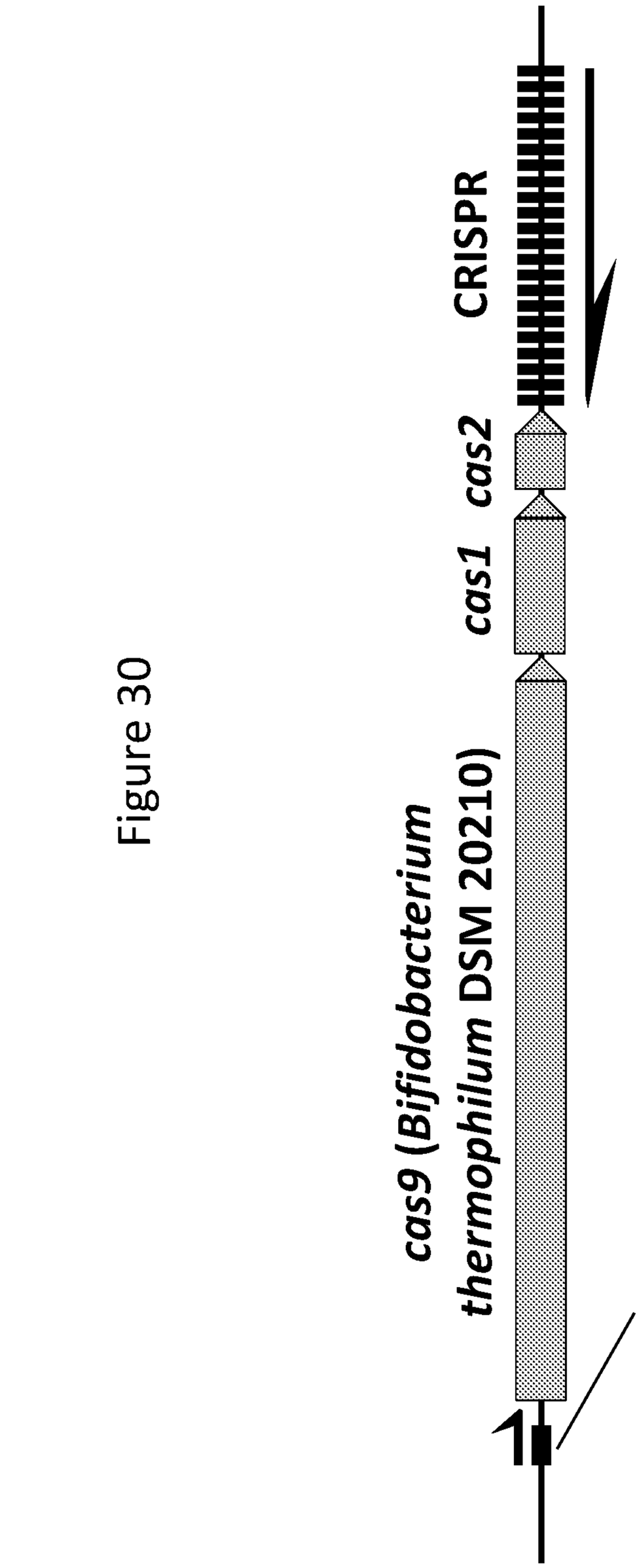

CRISPR
cas2
cas1
cas9 (Bifidobacterium thermophilum DSM 20210)
Anti-Repeat (tracrRNA coding region)

Figure 30

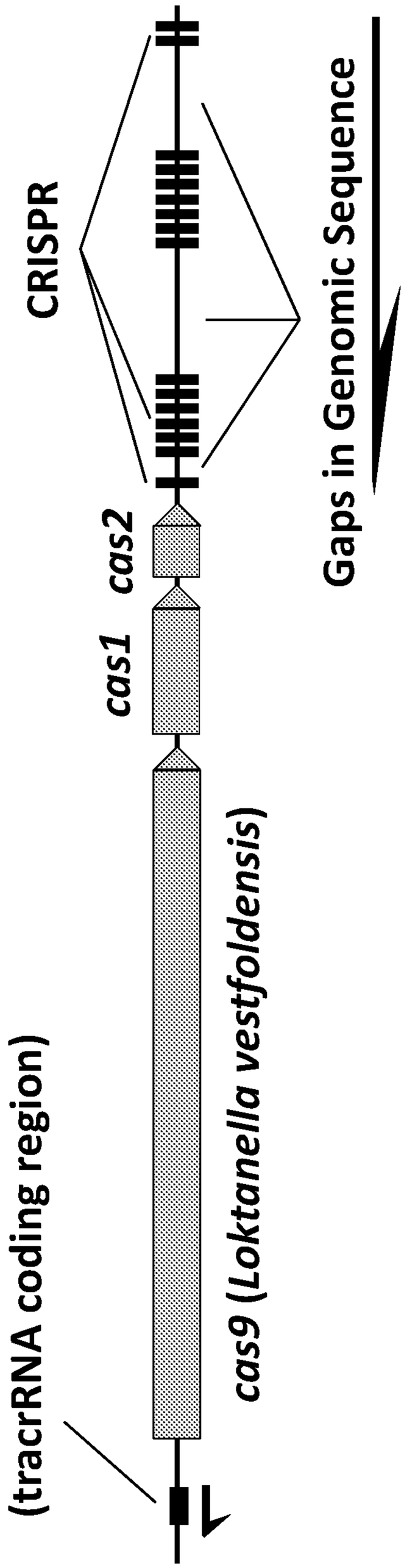

Anti-Repeat
(tracrRNA coding region)
CRISPR
Gaps in Genomic Sequence
cas1
cas2
cas9 (Loktanella vestfoldensis)

Figure 31

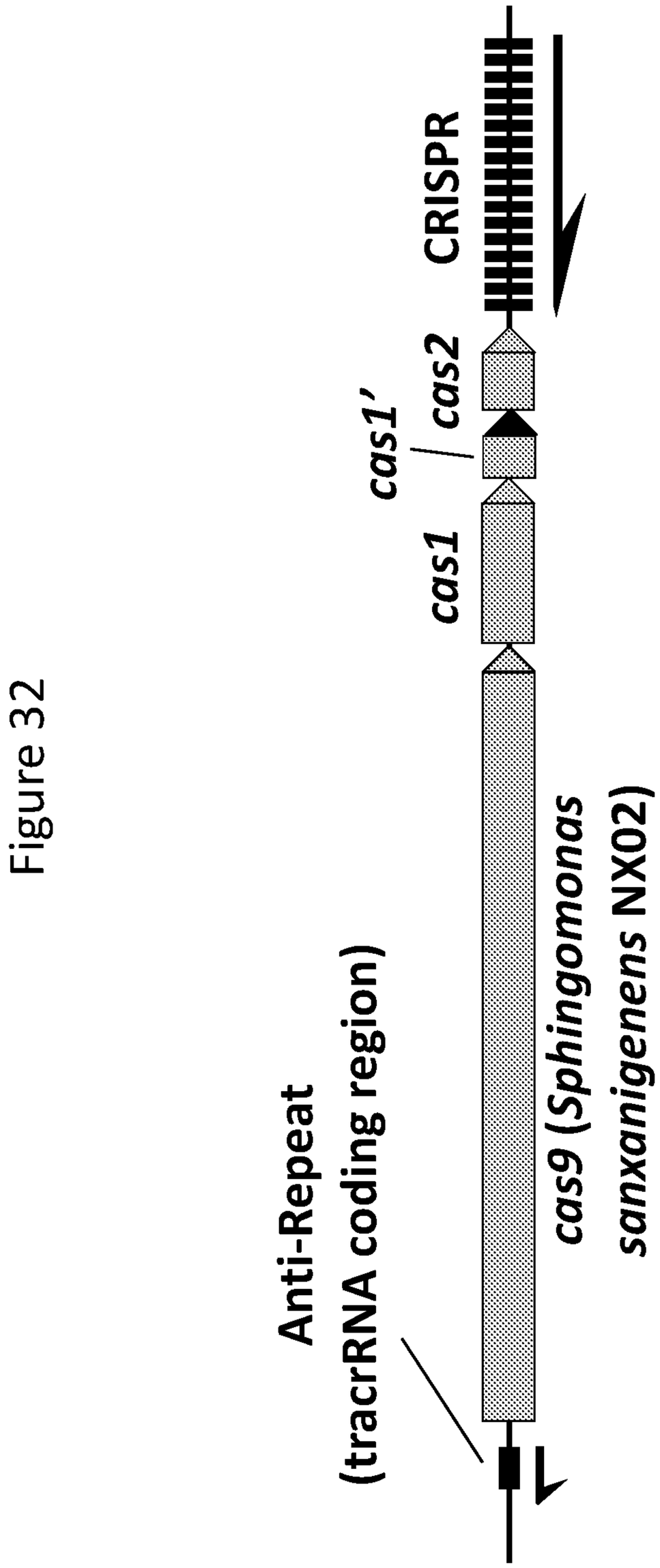

CRISPR
cas1'
cas2
cas1
Anti-Repeat
(tracrRNA coding region)
cas9 (Sphingomonas
sanxanigenens NX02)

Figure 32

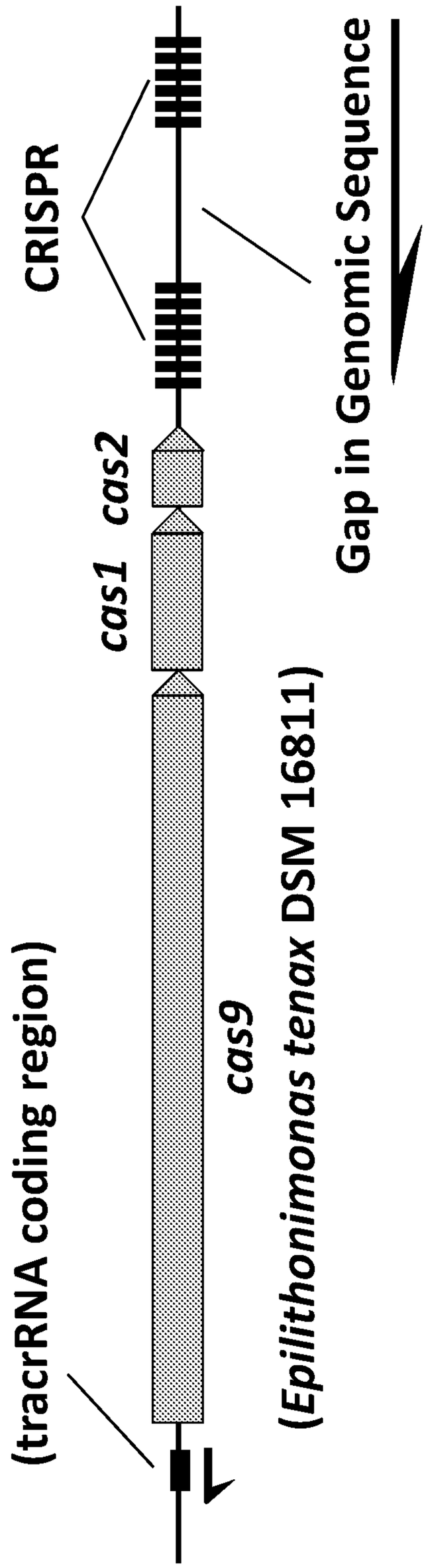

Anti-Repeat
(tracrRNA coding region)
cas9
(Epilithonimonas tenax DSM 16811)
cas1 cas2
CRISPR
Gap in Genomic Sequence

Figure 33

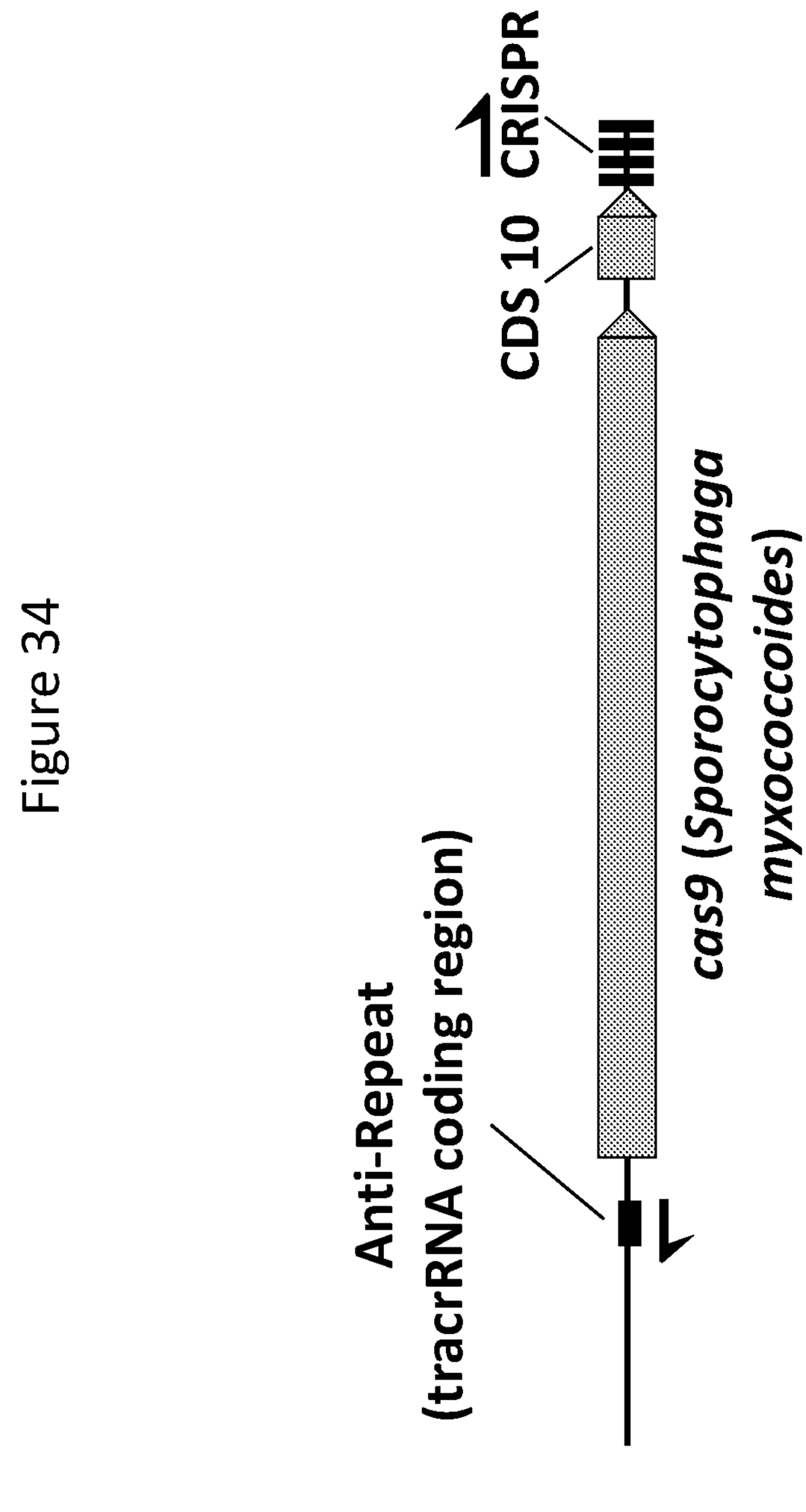

tracrRNA
CRISPR
CDS 10
cas9 (Sporocytophaga myxococcoides)
Anti-Repeat
(tracrRNA coding region)

Figure 34

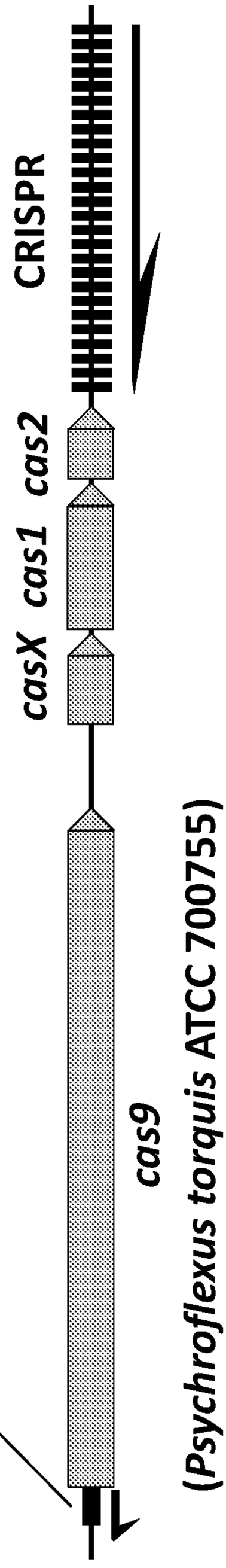

CRISPR
casX cas1 cas2
cas9
(Psychroflexus torquis ATCC 700755)
Anti-Repeat
(tracrRNA coding region)

Figure 35

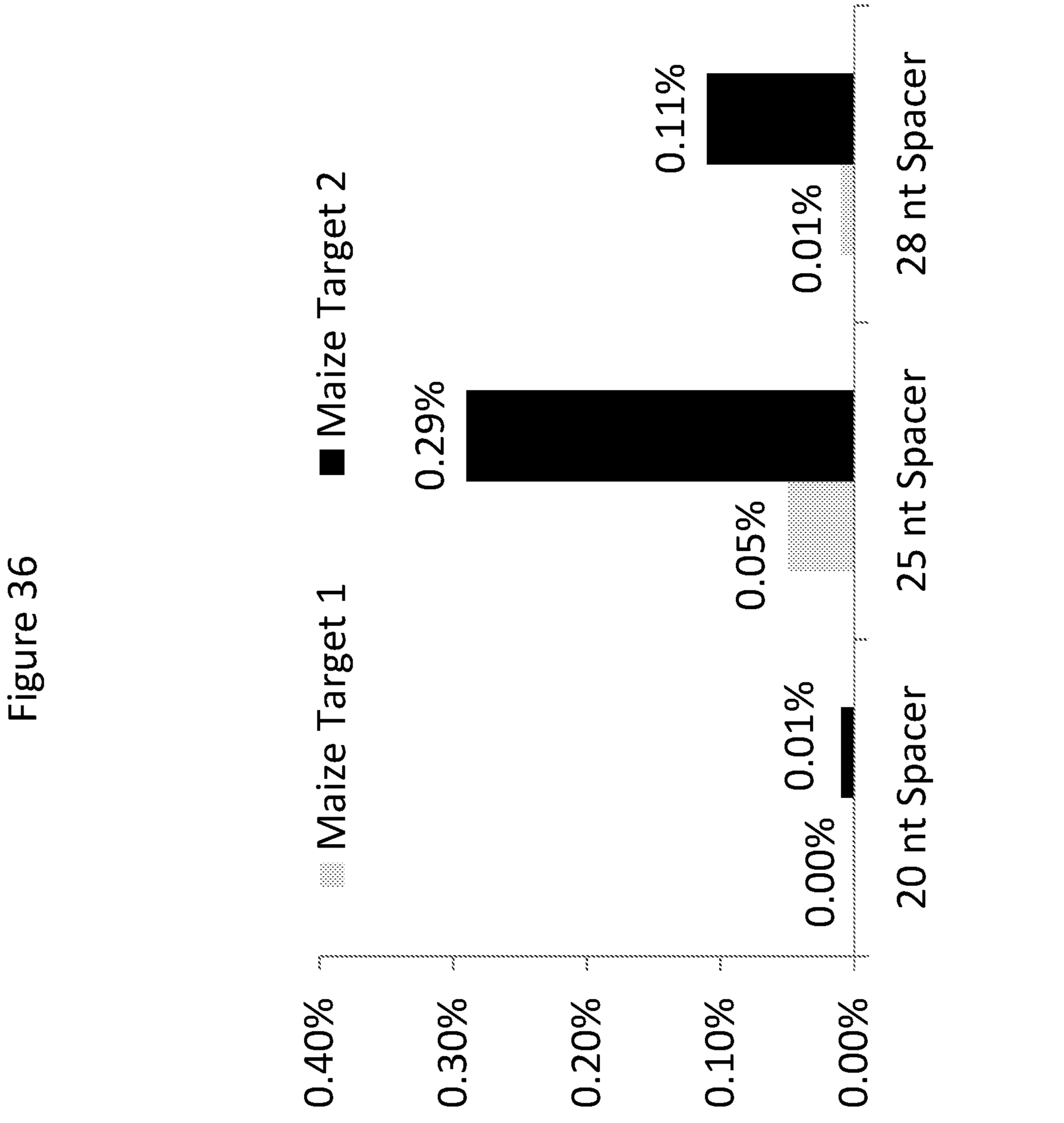
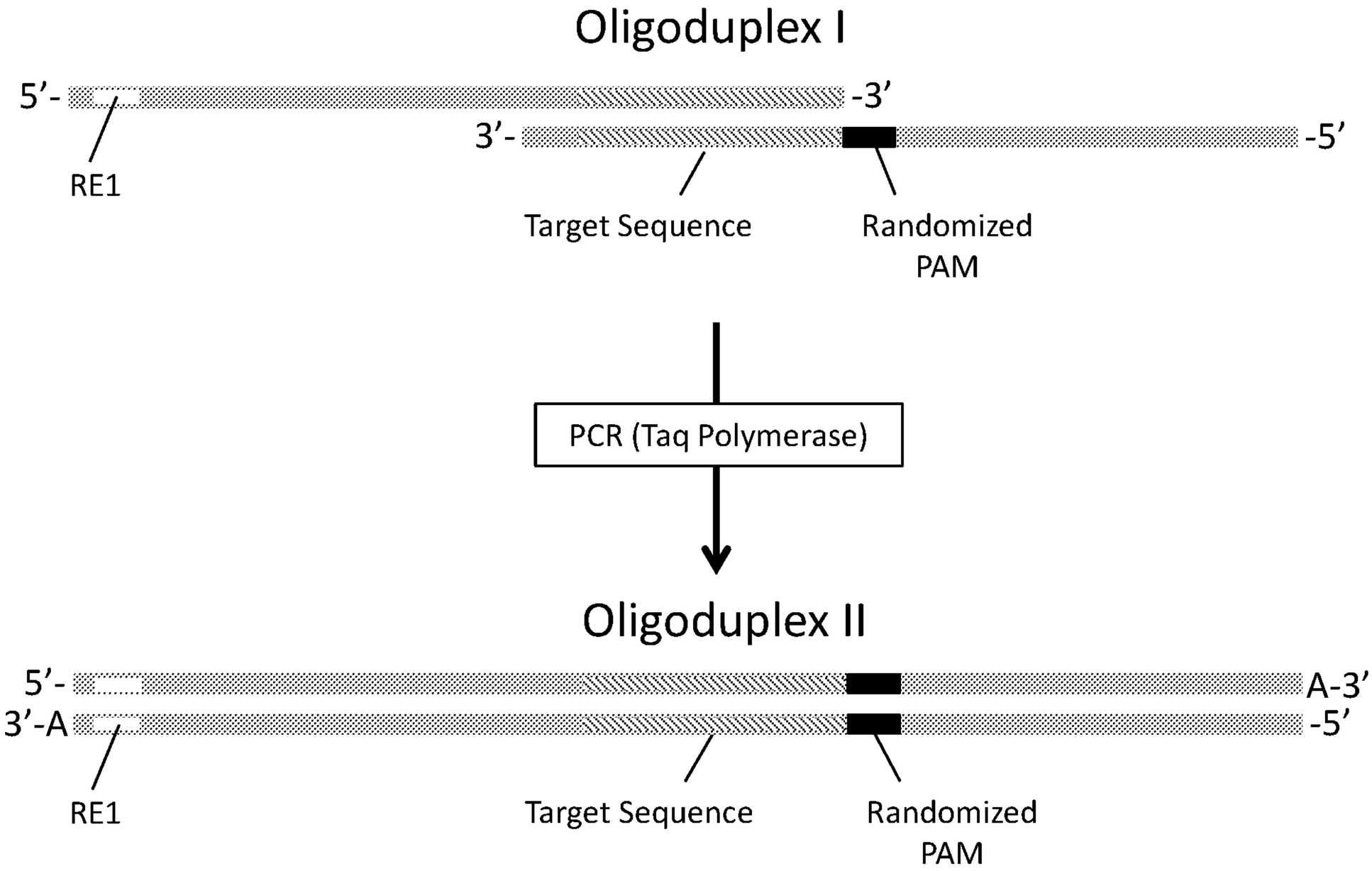

% Mutation Frequency
0.40%
0.30%
0.20%
0.10%
0.00%
Maize Target 1
Maize Target 2
0.29%
0.05%
0.11%
0.01%
0.00%
0.01%
25 nt Spacer
28 nt Spacer
20 nt Spacer

Figure 36

Figure 1


Oligoduplex I
5'-
-3'
3'-
-5'
RE1
Target Sequence
Randomized
PAM
PCR (Taq Polymerase)
Oligoduplex II
5'-
A-3'
3'-A
-5'
RE1
Target Sequence
Randomized
PAM